ROBUST VARIABLE SELECTION IN LINEAR REGRESSION MODELS

ALSHQAQ, SHOKRYA SALEH A

FACULTY OF SCIENCE

UNIVERSITY OF MALAYA

KUALA LUMPUR

2015

ROBUST VARIABLE SELECTION IN LINEAR REGRESSION MODELS

ALSHQAQ, SHOKRYA SALEH A

THESIS SUBMITTED FOR THE FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

INSTITUTE OF MATHEMATICAL SCIENCES
FACULTY OF SCIENCE

UNIVERSITY OF MALAYA

KUALA LUMPUR

2015

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name: **ALSHQAQ, SHOKRYA SALEH A**

Registration/Matric No: SHB100022

Name of Degree: DOCTOR OF PHILOSOPHY

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

ROBUST VARIABLE SELECTION IN LINEAR REGRESSION MODELS

Field of Study: STATISTICS

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date

Subscribed and solemnly declared before,

Witness's Signature Date

Name:
Designation:

ABSTRACT

This study looks at two problems related to the robust variable selection in linear regression models with six objectives in mind. The first three objectives are concerned with the problem of selection variables in small data sets in a linear regression model. The first is the investigation of the robustness of various best variable selection criteria in the presence of outliers and leverage points in the data set. The second derives the influence function of AIC, Cp, and SIC criteria and discussed the properties of these functions. The third is to explore the role of two robust methods for selecting the best variable in the linear regression.

The first approach considered is a modified version of AIC, Cp, and SIC statistics by utilizing the high breakdown point estimators of the regression model. The other methods are based on diagnostic regression approach using outliers and leverage diagnostics in regression model procedures. For each method, the power of performance is compared with classical non-robust criteria and the existing criteria, based on M-estimation. In general, our findings show that these criteria are capable of selecting the appropriate models in the presence of outliers.

The following three objectives look at the development of $LASSO$ variable selection regression to solve the problem of multicollinearity and large data in variable selection procedure. The fourth is to investigate the sensitivity of non-robust $LASSO$ ($LASSO$ and $adaptive$-$LASSO$) and robust $LASSO$ (LAD-$LASSO$ and Huber-$LASSO$) toward the existence of outliers and leverage points in the data. The fifth looks at extending the Huber-$LASSO$ to include more robust estimators. We present the GM-$LASSO$ and MM-$LASSO$ methods. If the multicollinearity does exist, we use the idea of the

$LASSO$ regression analysis to find the best variable in the model. The performance of these methods has also been compared with classical non-robust $LASSO$, and the existing robust LAD-$LASSO$ and Huber-$LASSO$ are generally good. The final objective is to prepare a new $LASSO$ method based on diagnostic regression approach.

ABSTRAK

Kajian ini menyelidiki dua masalah berkaitan dengan pemilihan pembolehubah teguh dalam model regresi yang berdasarkan enam objektif. Tiga objektif pertama melibatkan masalah pemilihan pembolehubah dalam data set kecil bagi suatu model regresi linear. Yang pertama adalah menyelidiki keteguhan berbagai kriteria pemilihan pembolehubah terbaik dalam kehadiran nilai-nilai terpencil dan titik-titik tuasan dalam sesuatu data set. Yang kedua adalah mendapatkan fungsi pengaruh bagi kriteria AIC, Cp, dan SIC dan membincangkan ciri-ciri fungsi tersebut. Yang ketiga bertujuan untuk meninjau peranan dua kaedah keteguhan untuk memilih pembolehubah terbaik dalam regresi tersebut.

Pendekatan pertama yang dipertimbangkan adalah versi statistik AIC, Cp dan SIC yang diubahsuai dengan menggunakan penganggar musnah tinggi model regresi yang terpilih. Kaedah lain berdasarkan pendekatan diagnosis regresi menggunakan diagnosis nilai terpencil dan titik tuasan dalam prosedur model regresi. Untuk setiap kaedah, kuasa prestasinya dibandingkan dengan kriteria klasikal tidak teguh dan kriteria sedia ada, berasaskan M-estimation. Secara keseluruhan, hasil kajian ini menunjukkan kriteria tersebut mempunyai keupayaan memilih model yang sesuai apabila terdapat nilai-nilai terpencil.

Objektif ketiga yang berikut melihat pembangunan pemilihan pembolehubah regresi $LASSO$ untuk menyelesaikan masalah kekolinearan berganda dan data berdimensi besar dalam prosedur pemilihan pembolehubah. Yang ke empat adalah untuk menyelidiki sensitiviti $LASSO$ tidak teguh ($LASSO$ dan $adaptive$-$LASSO$) dan $LASSO$ teguh (LAD-$LASSO$ dan Huber-$LASSO$) terhadap kehadiran nilai terpencil dan titik tuasan dalam data. Yang ke lima melihat cara untuk memperluaskan kaedah Huber-$LASSO$ un-

tuk merangkumi penganggar-penganggar yang lebih teguh. Penyelidik membentangkan kaedah GM-$LASSO$ dan MM-$LASSO$. Jika terdapat kekolinearan berganda, penyelidik menggunakan idea analisis regresi $LASSO$ untuk mendapatkan pembolehubah paling baik dalam model tersebut. Prestasi kaedah-kaedah tersebut telah dibandingkan dengan $LASSO$ klasikal tidak teguh, LAD-$LASSO$ teguh dan Huber-$LASSO$ yang sedia ada, dan didapati memuaskan secara keseluruhan. Objektif terakhir adalah untuk menghasilkan kaedah $LASSO$ baru yang berdasarkan pendekatan diagnosis regresi.

ACKNOWLEDGEMENT

Firstly, I would like to express my sincere gratitude to my supervisors, Prof. Dr. Nor Aishah Binti Hamzah and Dr. Rossita Binti Mohamad Yunus, for their guidance, support, invaluable help, encouragement and supervision throughout my research. Their understanding, patience and valuable advice have been the keys to the success of this study.

With deep sense of gratitude, I would like to thank, my husband, Fahdel Salman , my son, Salman, my daughter Logyn for their love, understanding, support and encouragement in all stages of this study. Also to my parents and family members who have always prayed for me throughout the period. Without their tremendous love and encouragement, I would not have been able to concentrate on my study and endure some difficult times through all these years.

I am thankful to the Ministry of Higher Education Saudi Arabia especially King Abdullah Program for Foreign Scholarships for providing the scholarship. I also would like to gratefully acknowledge the support given by the staff members of the Institute of Mathematical Sciences, University of Malaya especially Puan Budiyah. And Alhamdulillah, Praise to Allah for His Blessings and without His Will, this survey will never be finished.

Contents

List of Figures

List of Tables

LIST OF SYMBOLS AND ABBREVIATION

y Dependent variable (response variable)

X Independent variable (observed regressor)

p Number of variable (column) in the dataset

n Number of observation (row) in the dataset

$\boldsymbol{\epsilon}$ Errors

σ^2 Residual scale

$N(0, \sigma^2)$ Normally distributed with mean zero and constant variance σ^2.

$\boldsymbol{\beta}$ Parameter

$E(.)$ Mean

MSE Mean-Squared Error

LS Least Squares

R^2 The Squared Multiple Correlation Coefficient

SSE Sum of Square Error

SST Total sum of square

Cp Mallow's Cp

FPE The Final Prediction Error

AIC Aikaike Information Criteria

BIC Bayesian Information Criteria

SIC Schwarz Criterion

HQ Hanaan and Quinn Criteria

$\overline{y}$ Average of the dependent variables

$X^{\star}$ The contaminated values on X

$y^{\star}$ The contaminated values on y

$LASSO$ Least Absolute Shrinkage and Selection Operator

LAD Least absolute deviations

BS Biweight S-estimate

LTS Least Trimmed Squares

IF Influence function

$MRPE$ The Median Relative Prediction Errors

GM Generalized M-estimation

RCp Robust Mallow's Cp

$RFPE$ Robust Final Prediction Error

$RAIC$ Robust Aikaike Information Criteria

$RSIC$ Robust Schwarz Criterion

MAD Median absolute devation

MLE Maximum likelihood estimator

BP Break down point

LMS Least median of squares

CV Cross-Validation

iid independent and identically distributed

$Eqn.$ Equation

$Eqns.$ Equations

CHAPTER 1

OVERVIEW

1.1 Introduction

1.1.1 Background of the Study

Linear regression model analysis is the most widely used statistical technique that deals with linear and additive relationships between variables. This model usually applies under an assumption of independently, identically, and normally distributed errors. In many situations, the main purpose of fitting a regression equation is to predict the response variable. If the number of predictor variables is large and the number of observations is relatively small, fitting the model using all the predictors will yield poorly estimated coefficients, especially when predictors are highly correlated. More precisely, the variances of the estimated coefficients will be high and therefore the forecasts made with the estimated model will have a large variance, too. A common practice to overcome this difficulty is to fit a model using only a subset of variables selected based on statistical criteria.

In order to deal with this key issue, various variable selection techniques have been proposed that are able to select important variable in regression data analysis. Among those are spawning methods such as F tests for nested models, Akaike information criterion (AIC), Mallows Cp, exhaustive search, stepwise, backward, forward selection procedures, cross-validation, and Bayesian information criterion (BIC). This study focuses especially on penalized-likelihood criteria like AIC and BIC, or Cp as rather common

techniques. Although these methods perform well only in small sample, for large data sets $LASSO$ regression method are considered.

1.1.2 The Linear Regression Model

Let y_i denote the response variable and $\mathbf{X}_i$ the p explanatory variables. Suppose that, we have data set $\{(\mathbf{X}_i, y_i)\}$, where, $\mathbf{X}_i = (\mathbf{x}_{i1}, ..., \mathbf{x}_{ip})^T$, $i = 1, ..., n$ consists of p explanatory variables for the i^{th} observation vector of explanatory values and y_i the i^{th} response value. The linear regression model makes the following assumptions:

Assumption 1: a linear relationship between $\mathbf{X}_i$ and y_i is:

$$y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \epsilon_i, \tag{1.1}$$

where, $\epsilon_i = (\epsilon_1, ..., \epsilon_n)$ are the errors.

Assumption 2: the errors are independent and normally distributed with mean zero and constant variance σ^2, $\epsilon_i \sim N(0, \sigma^2)$.

Assumption 3: $y_i \sim N(\mathbf{X}_i^T \boldsymbol{\beta}, \sigma^2)$.

The most popular way of estimating $\boldsymbol{\beta}$ in Eqn. (1.1) is to minimize the ordinary least squares (LS) criterion,

$$\sum_{i=1}^{n} (y_i - \mathbf{X}_i^T \boldsymbol{\beta})^2, \tag{1.2}$$

which yield the estimator, $\hat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T y$, where $\mathbf{X}$ is as $n \times p$ matrix whose ith row is $\mathbf{X}_i$ with full rank p, and y is the response vector as follows,

$$\mathbf{X} = \begin{pmatrix} x_{12} & \cdots & x_{1n} \\ x_{22} & \cdots & x_{2n} \\ \vdots & \ddots & \vdots \\ x_{p2} & \cdots & x_{pn} \end{pmatrix}, y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Uses of Regression Equation

The general goal of regression analysis is to identify the relationship between observed variables based on numerical data. All of the available variables cannot be used automatically, because the magnitude of $var(\hat{y}_i)$ is influenced by the number of applied regressors. In addition, the $var(\hat{\boldsymbol{\beta}})$ will increase, as the number of regressor increases. Consequently, the actual subset of regressors that should be used in the model needs to be determined for making proper influence on the data.

Lacking selection of an appropriate subset of regressors for the model, causes the following problems: (1) exclusion of important variable, which impacts the misspecification of the model (the least square estimate is the biased estimate, the large variance is probably a biased estimate, and the variance of predicted value is large) and the coefficient of multiple determination gives very small value. (2) inclusion of unnecessary variables that, consequently increases the mean sum of square error, the coefficient of multiple determination gives very small value, and lead to multicollinearity problem. In view of these, it is necessary to show the effect of under or over fit on LS estimation.

1.1.3 The Effect of Under Fitting on LS Estimation

Assume that, the true model for n-vector of observations is of the form:

$$y = \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + \epsilon, \tag{1.3}$$

where, $\mathbf{x}_1$ and $\mathbf{x}_2$ are two matrices of size $n \times p$ and $n \times q$, respectively, and β_1 and β_2 are vectors of size p and q, respectively. However, let the model fit to the data takes in the reduced form: $y = \mathbf{x}_1\beta_1 + \epsilon$. Then for the reduced model, the LS estimators for β_1, and σ^2 are: $\hat{\beta}_{1R} = (\mathbf{x}_1^T\mathbf{x}_1)^{-1}\mathbf{x}_1^T y$, and $\hat{\sigma}_R^2 = y^T(I-H)y/(n-p)$, where, $H = \mathbf{x}_1(\mathbf{x}_1^T\mathbf{x}_1)^{-1}\,\mathbf{x}_1^T$.

The expected value of the estimated parameter vector is: $E(\hat{\beta}_{1R}) = (\mathbf{x}_1^T\mathbf{x}_1)^{-1}\mathbf{x}_1^T E(y) = \beta_1 + (\mathbf{x}_1^T\mathbf{x}_1)^{-1}\mathbf{x}_1^T\mathbf{x}_2\beta_2 \neq \beta_1$, where, $E(y) = x_1\beta_1 + x_2\beta_2$. And the expected value of the estimated error variance is: $E(\hat{\sigma}_R^2) = \sigma^2 + \beta_2^T\mathbf{x}_2^T(I-H)\mathbf{x}_2^T\beta_2/(n-p) \geq \sigma^2$.

Finally, the mean squared error for β_{1R} is: $MSE(\hat{\beta}_{1R}) = var(\hat{\beta}_{1R}) + bias(\hat{\beta}_{1R})^2 = \sigma^2(\mathbf{x}_1^T\mathbf{x}_1)^{-1} + A\beta_2\beta_2^T A^T$, where, $A = (\mathbf{x}_1^T\mathbf{x}_1)^{-1}\mathbf{x}_1^T\mathbf{x}_2 > 0$.

Thus, the following properties are summarized:

- $\hat{\beta}_{1R}$ is a biased estimate of β_1 and $\hat{y}_R$ is a biased estimate of $E(y)$ unless the true regression coefficient for each deleted variable is zero ($\beta_2 = 0$) or in the case of $\hat{\beta}_{1R}$, each deleted variable is orthogonal to the other retained variables ($\mathbf{x}_1$ is orthogonal to $\mathbf{x}_2$ or both).

- $\hat{\sigma}_R^2$ is biased positively, unless $\beta_2 = 0$.

- $MSE(\hat{\beta}_{1R})$ is a positively biased estimate of σ^2, unless the true regression coeffi-

cients for all deleted variables are zero.

1.1.4 The Effect of Over Fitting on LS Estimation

Suppose the true model for the n-vector of observations is of the following reduced form:

$$y = \mathbf{x}_1\beta_1 + \epsilon. \tag{1.4}$$

However, suppose the model fits to the data in the following full model form: $y = \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + \epsilon$. Thus, there is an over fit model. Let the LS estimators of the parameter vectors for the full model be: $\hat{\beta}_{1F}$ and $\hat{\beta}_{2F}$; and $\hat{\beta}_{1R}$ is the estimator for the reduced model, then the relationship between the estimates are: $\hat{\beta}_{1F} = \hat{\beta}_{1R} - A\hat{\beta}_{2F}$ and $\hat{\beta}_{2F} = (\mathbf{x}_2^T(I - H_1)\mathbf{x}_2)^{-1}\mathbf{x}_2^T(I - H_1)y$, where, $E(\hat{\beta}_{1F}) = \beta_1$ and $E(\hat{\beta}_{2F}) = 0$, (for reduced model true) $E(\hat{\sigma}_F^2) = \sigma^2$. Then, the consequences are:

- Unbiased estimators of model parameters,
- Loss of precision, due to fitting the wrong model: $var(\hat{\beta}_{1F}) \geq var(\hat{\beta}_{1R})$.

Finding an appropriate subset of variables for the model is called "variable selection problem". This problem fits a model using only a subset of variables selected according to some statistical criteria. The details of the selection criteria are given in the next section.

1.1.5 Subset Selection Criteria

Why is Subset of Variable Selection?

Variable selection is an important topic in linear regression analysis. At the initial stage of most applications of regression, one may be uncertain about the exact structure of the model. It may be unknown that whether all of the explanatory variables are really necessary and which of them affects the response variable.

However, many of the variables may have little effect on the response. Therefore, variable selection aims to build a regression model with appropriate set of regressors. This model has as few predictors as possible and with a good fit. Typically, simple models are desirable regarding their potential to improve the prediction accuracy of the fitted model. Furthermore it is easy to enhance interpretability of a simple model and accelerate its learning process.

This research aims to study the variable selection problems in linear regression model for two types of data sets;

(i) small data sets (when p is small and $p \leq n$)

(ii) large data sets (when p is large and/or $p > n$)

Various classical variable selection criteria are suggested when the number p of variables is small. For example, Hocking and Leslie (1967), Miller (1984) elaborated the computational algorithms for selecting subset of regression variable in linear regression models and investigated the LS criterion.

In most of the classical selection criterion computing for possible subset P of the predictors, which is the number of P for each model is equal to $2^p - 1$. Therefore, due to the inversion of $\mathbf{X}_i^T\mathbf{X}_i$, these classical variable selection criteria require $p \leq n$.

There are large data set with large number of p or a number of variables that far outstrip the number of observations. For example, DNA microarray data is time-consuming to collect per subject, but often yield thousands of variables (genes). An example of a

typical study is the well-known analysis of prostate cancer patients by Singh et al. (2002).

On the other hand, if the number of p is large, the number of P dramatically increases. For example, if there are 10 independent variables available for selection, then there are $2^{10} - 1 = 1,023$ possible models to be evaluated. This makes the best subset selection methods to be computationally complicated.

In sparse high-dimensional modeling, it is assumed that most parameters are exactly zero, that is only a few predictors contribute to the response. In this study, the objective of variable selection is to identify important predictors with nonzero regression coefficient that give accurate estimates of those parameters.

Hoerl and Kennard (1970) were the pioneers to investigate the ridge regression problem with correlated coefficients in regression models. In fact, ridge regression modifies LS estimators into:

$$\hat{\boldsymbol{\beta}}_{ridge} = \arg\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} (y_i - \mathbf{X}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right], \lambda \geq 0. \tag{1.5}$$

Ridge regression shrinks the coefficients, but does not select variables because it does not force coefficients to be zero. That is why it is not considered as a method for variable selection.

Tibshirani (1996) proposed the '$LASSO$ penalty', a regularization technique for simultaneous estimation and variable selection for large data sets. The $LASSO$ estimators

are defined by:

$$\hat{\boldsymbol{\beta}}_{LASSO} = \arg\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} (y_i - \mathbf{X}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \mid \beta_j \mid \right], \lambda \geq 0, \tag{1.6}$$

where, $\lambda \in [0, \infty]$ is the $LASSO$ tuning parameters.

1.1.6 Outliers in Linear Regression

The outlier is a common problem in the statistical analysis. It is defined as an observation that is very different to the other observations in a set of data. Beckman and Cook (1983) and Barnett and Lewis (1994) defined an outlier in a set of data to be an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data. It is a prevalent practice to classify between two types of outliers. Outliers in response, often referred to as vertical outliers and outliers with respect to the covariates, **X** are called leverage points.

However, it is important to note the disparity between two types of leverage points: bad leverage and good leverage point. A bad leverage point refers to a point that lies far from the center of the covariates $\bar{\mathbf{X}}_i = (\bar{\mathbf{x}}_1, ..., \bar{\mathbf{x}}_p)$. A good leverage points are consistent with the majority of the data. That is, both good leverage and bad leverage point promote and reduce the precision of the regression coefficients, respectively.

Effect of Outliers on Regression Model

The presence of outlier has potentially serious effects on the LS estimation of the regression coefficients. Maronna et al. (2006) introduced the effect of outliers on LS estimation in two ways: First, if y_i replaced by $y_i^\star$ where, $y_i^\star = Ay_i$, which implies $y_i = A^{-1}y_i^\star$, then the linear regression in Eqn. (1.1) can be rewritten as follows: $A^{-1}y_i^\star = \mathbf{X}_i^T\boldsymbol{\beta} + \epsilon_i$,

thus, $\hat{\boldsymbol{\beta}}(\mathbf{X}_i, y_i^\star) = A^{-1}\hat{\boldsymbol{\beta}}(\mathbf{X}_i, y_i)$. Second, if $\mathbf{X}_i$ replaced by $\mathbf{X}_i^\star$, where $\mathbf{X}_i^\star = A\mathbf{X}_i$. Then Eqn. (1.1) can be rewritten as follows: $\hat{\boldsymbol{\beta}}(\mathbf{X}_i^\star, y_i) = A^{-1}\hat{\boldsymbol{\beta}}(\mathbf{X}_i, y_i)$.

1.1.7 Multicollinearity in Multiple Linear Regression

In some cases, the independent variables in a model might be near-linear dependence, leading to a problem of multicollinearity. This problem will cause difficulty to assess the relative importance of individual predictors from the estimated coefficients of the regression equation. In some extreme cases, may fail to obtain the estimates; because the matrix is close to being singular. Perfect multicollinearity occurs when correlation between two independent variables is equal to 1 or -1. Mansfield and Helms (1982) presented several indications of the multicollinearity problem including:

1. High correlation between pairs of independent variables,
2. Statistically nonsignificant regression coefficients on important predictors,
3. The extreme effect of the changes of sign or magnitude of regression coefficients when an independent variable is included or excluded.

Effect of Multicollinearity

In studying the effect of multicollinearity on regression modeling, Hoerl and Kennard (1970), and Swindel (1976) considered the unbiased linear estimation with minimum variance or maximum likelihood estimation when the random vector, ε, is normally distributed giving the estimator in Eqn. (1.2) as the estimate of $\boldsymbol{\beta}$. This gives the minimum sum of squares of the residuals

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{1.7}$$

The properties of $\hat{\boldsymbol{\beta}}$ can be found in Scheffe (1999) for the case $\mathbf{X}^T\mathbf{X}$ is not nearly a unit matrix. Hoerl and Kennard (1970) demonstrated the effects of the multicollinearity on the estimation of $\boldsymbol{\beta}$ by considering the variance-covariance matrix

$$COV(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

and the distance of $\boldsymbol{\beta}$. From its expected value, say, $L_1 \equiv \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ giving

$$L_1^2 = \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right), \tag{1.8}$$

with $E\left[L_1^2\right] = \sigma^2 tr\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right]$, or equivalently $E\left[\hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\beta}}\right] = \hat{\boldsymbol{\beta}}^T\boldsymbol{\beta} + \sigma^2 tr\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$. Using these properties, attempt to show the uncertainty in $\hat{\boldsymbol{\beta}}$ when $\mathbf{X}^T\mathbf{X}$ moves from a unit matrix to an ill-conditioned one. If the eigenvalues of $\mathbf{X}^T\mathbf{X}$ are denoted by

$$\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p = \lambda_{\min} > 0, \tag{1.9}$$

then

$$E\left[L_1^2\right] = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}, \tag{1.10}$$

and the variance when the error is normally distributed is given by

$$Var\left[L_1\right]^2 = 2\sigma^4 \sum_{j=1}^{p} \left(\frac{1}{\lambda_j}\right)^2. \tag{1.11}$$

Note that when the matrix $\mathbf{X}^T\mathbf{X}$ is ill-conditioned due to multicollinearity, then some of the λ_j will be small. Hence, from Eqn. (1.10), the least squares estimates $\hat{\boldsymbol{\beta}}$ is farther away from true parameter $\boldsymbol{\beta}$ and, from Eqn. (1.11), the variances of the least squares estimator of the regression coefficient have larger values. Hence, proper handling of mul-

ticollinearity problem is greatly needed.

To overcome these problems, Hoerl and Kennard (1970) and Breiman and Leo (1996) proposed regression modeling by regularization techniques. The regularization methods are based on penalty terms and should yield unique estimates of the parameter vector $\boldsymbol{\beta}$. Furthermore, an improvement of the prediction accuracy can be achieved by shrinking the coefficients or setting some of them to zero. Thereby, regression models are obtained that should contain only the strongest effects and those which are easier to interpret. In the following section, an overview of some already established regularization techniques is given.

1.2 Statement of the Problem

Data subjected to outliers are commonly encountered in applications, which may appear either in response variables or in the predictors. In this case, the model selection method based on LS estimator is reputed to be not efficient. It is due to the sensitivity of this estimator to outliers and other departures from the normality assumption on the error distribution. With regard to this problem, a special variable selection method in analyzing regression for contamination, small and large data sets is needed.

The robust variable selection methods have attracted the interest of both statisticians and researchers. As a result, new robust variable selection methods have been developed to overcome the problem with vertical outliers including robust versions of R^2, RCp, $RFPE$, $RAIC$ and $RSIC$ for data with small number of p. On the other hand, LAD-$LASSO$ and Huber-$LASSO$ have been proposed for large data. All these robust criteria based on objective functions define M-estimators for a parametric model.

However, the challenge to reduce the effect of leverage points in variable selection criterion has not received enough attention yet. In this regard, very few studies (such as Tharmaratnam and Claeskens, 2013; Arslan, 2012) focused on the effect of leverage points in variable selection criterion.

Tharmaratnam and Claeskens (2013) used S and MM-estimators on AIC criteria. Alfons et al. (2013) suggested a noble method by introducing sparse least trimmed squares (LTS) regression for analyzing large data sets. They believed that this model is capable to control the impact of leverage point problem.

1.3 Contribution of Thesis

The objectives of this study are listed as follows:

- **Variables Selection for Small Data Sets**

    1. To investigate the sensitivity of classical variable selections and robust variable selection based on M-estimation in the presence of outliers and leverage points in the data.
    2. To propose a new procedure for robust variable selection, which may contain high leverage points in the data set, and to study the influence function of proposed method.
    3. To suggest using a more robust estimate of scale in variable selection.

- **Variable Selection for Large Data Sets**

    1. To investigate the sensitivity of non-robust $LASSO$ ($LASSO$ and $adaptive$-$LASSO$) and robust $LASSO$ (LAD-$LASSO$ and Huber-$LASSO$) toward

the existence of outliers and leverage points in the data.

2. To propose a new procedure to address leverage points problem and to study its theoretical properties.

- To develop a new variable selection based on the diagnostic regression approach.
- In both cases, small & large data cases, we compare the performances of the methods using generated data. The experiences gained is the employed to real data sets.

1.4 Significance of the Study

The findings from this study will be beneficial in the following ways:

1. Contribute to the body of knowledge regarding the variable selection methods of linear regression and detection of the effect of outliers.
2. Optimize the variable selection methods in linear regression models by derive of influence function.
3. Contribute to the new methods that deal with leverage point and multicollinearity problem and optimize the methods.

1.5 Research Outline

The present research is outlined as follows:

Chapter 2 reviews related literature on variable selection, on both small and large regression data sets and robust procedures (estimation and diagnostic).

Chapter 3 establish the idea, the influence of outlier on variable selection criteria in small sample is illustrated through a small experiment and real data sets. Finally, the the-

ory of the $LASSO$ and $adaptive\text{-}LASSO$ variable selection for regression model based on consistency and oracle properties has been discussed. As such, the advantage of proposed methods over the breakdown point was discussed. Simulation study has also been carried out to see the effect of leverage points on the robust variable selections methods in large data set.

Chapter 4 presents the development of the new proposed robust variable selection, for AIC, Cp and SIC criteria using high breakdown point estimate of scale, instead of the classical scale, handling the leverage points problem in variable selection. The robustness of the proposed methods is studied through its influence function and gross-error sensitivity. The performance of proposed criteria are illustrated through simulation and real data set.

Chapter 5 presents the development of the robust $LASSO$ regression model and the treatment of leverage points problem in the linear regression models using GM and MM regression approach. Simulation study and real data are used to illustrate the performance of proposed methods.

Chapter 6 presents a variable selection methods statistic based on the idea of diagnostic tool statistic in linear regression that can be used to detect possible outliers in the regression models. Via simulation, the cut-off points are obtained and the power of performance is investigated. The statistic is then applied on simulation and real data sets.

Chapter 7 presents the summary of the study and the suggestion for further research.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter presents the main subjects related to the topic of this study by reviewing current literature. Since mid 90's, studying robust variable selection in regression data analysis became a mainstream focus area, particularly for the data sets with the smaller number of independent candidate covariates, and for high-dimensional data sets. Both these topics have been extensively studied in modern statistics. This chapter gives an overview of variable selection methods, on both small and large regression data sets and robust procedures.

2.2 Classical Variable Selection

2.2.1 Variable Selection Methods in Small Samples

The classical variable selection criteria such as R^2, Mallow's Cp and AIC are the popular methods for selecting the best model and has widely been used in linear regression model. However these criteria unfavourable when the number of explanatory variable is too high due to multicollinearity that may exists in the data. Alternative to these classical variable selection, criteria for large data sets are as discussed in Section 2.2.2. The classical variable selection defined in Table 2.1 are in terms of the sum of squares of residuals for full model (SSE) for the least squares squares estimate which are not robust against outliers. Section 2.3 gives the robust variable selection methods for low and large dimensional

data sets. Note, $SSE_p = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the sum of squares error for sub model with

Table 2.1: Classical variable selection criteria

Methods	Formulas	References
R^2 statistic	$1 - \frac{SSE_p}{SST}$	(Hahn (1973); Kvålseth (1985); Willett and Singer (1988)
Mallow's(Cp)	$\frac{SSE_p}{\hat{\sigma}^2_{full}} - n + 2p$	Mallows (1973b)
The Final Prediction Error (FPE)	$\frac{SSE_p}{\hat{\sigma}^2_{full}} + 2p$	Akaike (1969)
Akaike Information Criteria (AIC)	$\log(\frac{SSE}{n}) + 2p$	Akaike (1973); Bhansali and Downham (1977)
Schwarz Information Criterion (SIC)	$\log(\frac{SSE}{n}) + \frac{p\log(n)}{n}$	Schwarz (1978)
Hanaan and Quinn Criteria (HQ)	$\log(\frac{SSE}{n}) + \frac{2p\log(\log(n))}{n}$	Hannan and Quinn (1979)

p variables, and $SST = \Sigma_{i=1}^{n}(y_i - \bar{y_i})^2$, with $\bar{y_i}$ is the sample average of the dependent variable. $\hat{\sigma}_{full} = \sqrt{SSE/(n-p)}$ is an estimate of the error variance σ^2 that is usually computed in the full model.

2.2.2 Variable Selection Methods in Large Data Sets

Penalized Least Squares

Regularization approaches for normal regression problems are based on penalized least squares

$$PLS(\lambda, \boldsymbol{\beta}) = \sum_i \left(y_i - \mathbf{X}_i^T \boldsymbol{\beta}\right)^2 + P(\lambda, \boldsymbol{\beta}), \tag{2.1}$$

and estimates of the parameter vector $\boldsymbol{\beta}$ are obtained by minimizing this equation, i.e.

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} PLS(\lambda, \boldsymbol{\beta}). \tag{2.2}$$

The penalty term $PLS(\lambda, \boldsymbol{\beta})$ depends on the tuning parameter λ that controls the shrinkage intensity. For the tuning parameter $\lambda = 0$ the ordinary least squares solution are

obtained. On the contrary, for large values of λ the influence of the penalty term on the coefficient estimates increases. Hence, the penalty region determines the properties of the estimated parameter vector; whereas, desirable features are variable selection and a grouping effect. An estimator shows the grouping property if it tends to estimate the absolute value of coefficients (nearly) equal if the corresponding predictors are highly correlated.

Fan and Li (2001) proposed a unified approach based on non-concave penalized likelihood estimators that performs as oracle estimator in variable selection. Fan and Peng (2004) then, suggested the following oracle properties in an ideal technique: (1) consistency in variable selection, (2) asymptotic normality.

Ridge Regression

Hoerl and Kennard (1970) was the pioneer to investigate the ridge regression problem with correlated coefficients in regression models. In fact, ridge regression modifies LS estimators into:

$$\hat{\boldsymbol{\beta}}_{ridge} = \arg\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} (y_i - \mathbf{X}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right], \lambda \geq 0. \tag{2.3}$$

Ridge regression shrinks the coefficients, but does not select variables because it does not force coefficients to be zero. So it is not considered as the method for variable selection.

***LASSO* Regression**

Tibshirani (1996) has proposed the $LASSO$ penalty, a regularization technique for simultaneous estimation and variable selection for large data sets. The $LASSO$ estimators are

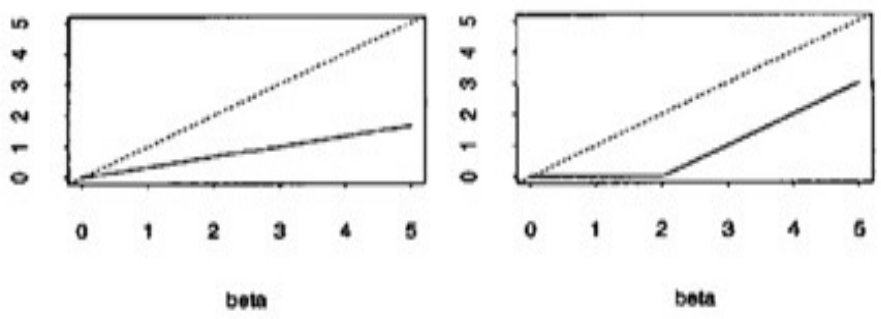


Figure 2.1: (Tibshirani, 1996) The left one is ridge and the right one is $LASSO$ regression

defined by:

$$\hat{\boldsymbol{\beta}}_{LASSO} = \arg\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} (y_i - \mathbf{X}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \mid \beta_j \mid \right], \lambda \geq 0, \tag{2.4}$$

where, $\lambda \in [0, \infty]$ is the $LASSO$ tuning parameters.

$LASSO$ minimizing the residual sum of squares, subjected to the sum of the absolute value of the coefficients, being less than a constant. Thus, this constraint tend to produce some coefficients that are exactly 0; and hence gives interpretable models. There are a number of theoretical support for the $LASSO$ method (Donoho and Elad (2003); Candes and Tao (2007); Bickel et al. (2009)). Figure (2.1) gives an idea of how the ridge regression and $LASSO$ work in the orthogonal case. It shows that, ridge regression only shrinks the coefficient, and does not set any coefficients to 0. However, $LASSO$ sets some of the coefficient 0, and shrinks the other ones. For the two-dimensional case, Figure (2.2) shows why the $LASSO$ exhibits the ability to select predictors. The contours of the residual sum of squares are ellipses, centered at the ordinary least squares estimate. The constraint region for the $LASSO$ is the rotated square $\mid \beta_1 \mid + \mid \beta_2 \mid \leq t$, whereas that for ridge regression is the disk $\beta_1^2 + \beta_2^2 \leq t$. The first point where the elliptical contours touch the constraint region corresponds to the $LASSO$ and ridge solution, respectively.

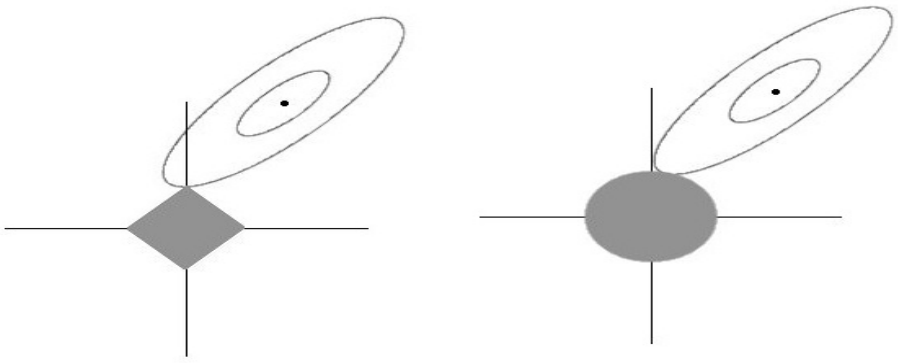


Figure 2.2: (Tibshirani, 1996) Estimation picture for $LASSO$ (left) and ridge regression (right)

Since the first osculation point of the ellipses can be a vertex of the square, the $LASSO$ solution can have one coefficient β_j equal to zero. In contrast, ridge regression cannot produce zero solutions because there are no vertices in the constraint region that can be touched.

Thus, $LASSO$ regression has been extensively studied in the literatures (see, [Bickel et al. (2009); Bunea et al. (2007); Khan et al. (2007); Leng et al. (2006); Osborne et al. (2000); Zhao and Yu (2006), Lounici (2008); Wainwright (2009); Efron, Hastie, Johnstone, and Tibshirani (2004); Khan, Van Aelst, and Zamar (2007)]).

Meinshausen and Bühlmann (2006) and Leng et al. (2006) stated that $LASSO$ is reliable in selecting variables, provided that the fundamental model fulfils few conditions and the variable selection is not consistent when accuracy is used as the criterion for choosing the penalty.

Adaptive-*LASSO* (*ada-LASSO*)

Besides the advantage of variable selection, the $LASSO$ also has some limitations. As discussed by Tibshirani (1996) ridge regression dominates $LASSO$ with regard to prediction accuracy in common case of $n > p$ case if there are high correlations among the

variables. Another drawback of the $LASSO$ solution is the fact that in $p > n$ situations, it selects at most n variables. Moreover, $LASSO$ cannot be an oracle procedure as pointed out by Zou (2006).

An alternative method was proposed by Zou (2006) to improve $LASSO$ in terms of achieving consistency of variable selection and prediction accuracy in large data sets. This model is based on the weighted $LASSO$ and has oracle property that, is a major improvement for $LASSO$. Zou (2006) proposed $ada\text{-}LASSO$ which assigns different weights to different coefficients and illustrates an oracle procedure. The $ada\text{-}LASSO$ criterion is defined by

$$\hat{\boldsymbol{\beta}}^{(n)}_{ada-LASSO} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \mathbf{X}_i^T \boldsymbol{\beta}\right)^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j \mid \beta_j \mid \right\}, \tag{2.5}$$

where w_j known weight vector. To define $ada\text{-}LASSO$, suppose that $\hat{\boldsymbol{\beta}}$ is a root n-consistent estimator to $\boldsymbol{\beta}$; for example, $\hat{\boldsymbol{\beta}}_{LS}$ can be used to pick a $\gamma > 0$, and define the weight vector $\hat{w}_j = \frac{1}{|\hat{\beta}_j|^\gamma}$. Similarly, let $\mathbf{A} = \left\{ j : \hat{\beta}_j^{(n)} \neq 0 \right\}$. With a proper choice of λ_n, for $w_j = 1$, $LASSO$ regression are obtained, whereas for $w_j = 1/\hat{\boldsymbol{\beta}}_{LS}$; then the adaptive solves

$$\hat{\boldsymbol{\beta}}^{(n)}_{ada-LASSO} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{X}_i^T \boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p} \frac{\mid \beta_j \mid}{\mid \hat{\beta}_{j_L S} \mid} \right\}, s.t. \beta_j \hat{\beta}_{j_L S} \geq 0 \forall j. \tag{2.6}$$

In the following sections 2.3 to 2.5, some general knowledge on robust statistics are written.

2.3 A Review of Robust Procedures

There are two approaches to reduce the effect of outliers in regression models. First, robust regression estimators, which tries to obtain estimators that are not so strongly affected by outliers. The second approach is based on regression diagnostics, where certain quantities are computed from the data with the purpose of pinpointing influential points, after which these outliers can be removed.

2.3.1 Robust Regression Estimation Techniques

Several robust techniques have been proposed to obtain the estimates that are not influenced by outliers and have high efficiencies, relative to LS estimates under the assumption of normally distribution errors. Some of the robust techniques are resistant to vertical outliers, that is, their breakdown points are near 0.5, but have breakdown with leverage points. On the other hand, others achieve both, verticals and leverage points. However, in this section some of these robust estimator techniques used for estimating the regression coefficients are reviewed.

M-estimators

In Eqn. (1.1), if $(y_i - \mathbf{X}_i^T\boldsymbol{\beta})^2$ has been replaced by $\rho(y_i - \mathbf{X}_i^T\boldsymbol{\beta})$, where $\rho(\cdot)$ is a function less intensively demonstrates the dimension of the residual. This concept results in the idea of M-estimation as described by Huber (2011), Hampel et al. (2011), and Birkes and Dodge (2011).

In case of a linear model, suppose that the observed responses y_i are independent, but

not identically distributed and have density functions as follows:

$$f_i(y_i) = \frac{1}{\sigma} f\left(\frac{y_i - \mathbf{X}_i^T \boldsymbol{\beta}}{\sigma}\right), \tag{2.7}$$

where, the log likelihood is given by:

$$\ell(\boldsymbol{\beta}) = -n \log \sigma + \sum_{i=1}^{n} \log\left(f\left(\frac{y_i - \mathbf{X}_i^T \boldsymbol{\beta}}{\sigma}\right)\right). \tag{2.8}$$

M-estimates on regression are defined as the value $\boldsymbol{\beta}$ that, minimizes the following criterion:

$$\sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right), \tag{2.9}$$

where, ρ a known function and $\hat{\sigma}$ is a preliminary robust error scale such as median absolute deviation (MAD) scale given by: $\hat{\sigma} = c \times median\,|r_i - median(r_i)|$ where, $r_i = y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}$ and the tuning constant $c = 1.4825$.

For each $i = 1, 2, ..., n$, for the M-estimator, the ρ-function in Eqn. (2.9) is a filter function constructed subject to the following properties: $\rho\left(\frac{r_i}{\hat{\sigma}}\right) \geq 0$, $\rho(0) = 0$, $-\rho\left(\frac{r_i}{\hat{\sigma}}\right) = \rho\left(-\frac{r_i}{\hat{\sigma}}\right)$ and $\rho(\infty)$=1, if ρ is bounded. Differentiating Eqn. (2.9) with respect to $\boldsymbol{\beta}$ yields the normal Equation

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right) = 0, \tag{2.10}$$

where, $\psi = \rho'$. In particular, if $\rho = -\log(f(\mathbf{X}_i))$, then the solution of the normal equation becomes the maximum likelihood estimation (MLE) of $\boldsymbol{\beta}$. Whereas, if $\rho = \frac{1}{2}(y_i - \mathbf{X}_i^T \boldsymbol{\beta})^2$, then the solution of the normal equation becomes the LS estimate. Generally, Biweight or Huber functions have been widely used as ρ.

Biweight Function: Define the weight matrix = diag(w_i), with $w_i = \frac{\psi\left(\frac{r_i}{\hat{\sigma}}\right)}{\left(\frac{r_i}{\hat{\sigma}}\right)}$, then Eqn.

(2.10) can be written as:

$$\sum_{i=1}^{n} w_i \rho \left(\frac{y_i - \mathbf{X}_i^T \boldsymbol{\beta}}{\hat{\sigma}} \right) \mathbf{X}_i = 0, \tag{2.11}$$

this equation can be combined into the following single matrix equation $\mathbf{X}_i^T W \mathbf{X}_i \boldsymbol{\beta} = \mathbf{X}_i^T W y_i$. Therefore, the estimator is $\hat{\boldsymbol{\beta}} = (\mathbf{X}_i^T W \mathbf{X}_i)^{-1} \mathbf{X}_i^T W y_i$, where W is $n \times n$ matrix of weight matrix.

In practice, the weighted matrix W involves $\boldsymbol{\beta}$ and is unknown. Therefore, iterative algorithm to solve this problem should be used, that is, use the estimator of $\boldsymbol{\beta}$ in the last iteration to calculate W, then use it to obtain the estimator of $\boldsymbol{\beta}$ in the current iteration. The algorithm stops when the estimator converges. This is the so-called iteratively reweighted least-squares (IRLS) algorithm.

Huber Function: Huber function is given by the following equation:

$$\rho_H(r) = \begin{cases} r_i^2, & \text{if } |r_i| \leq M, \\ 2M|r_i| - M^2, & \text{elsewhere,} \end{cases} \tag{2.12}$$

this function is quadratic in small values of r, but grows linearly for large values of r. Huber (2011) have proposed to fix $M = 1.345$ to increase the robustness as much as possible, while being efficient for normal distributed data. Then Eqn. (2.9) can be written as: $\sum_{i=1}^{n} \rho_H \left(\frac{r_i}{\hat{\sigma}} \right)$, or Huber's criterion with concomitant scale with respect to $\boldsymbol{\beta}$ and σ (see, (Mallows, 1973a, 1975)),

$$n\sigma + \sum_{i=1}^{n} \rho_H(\frac{r_i}{\hat{\sigma}})\sigma, \sigma > 0.$$

Least Absolute Deviations (*LAD*) Regression or (L_1): If $\rho = \mid \mathbf{X}_i \mid$, then the LAD estimates are achieved by minimizing the sum of the absolute values of the residuals:

$$\hat{\boldsymbol{\beta}}_{LAD} = \min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} \left| y_i - \mathbf{X}_i^T \boldsymbol{\beta} \right| \right], \tag{2.13}$$

where the random errors ε_i have median zero.

High Breakdown Point and Bounded Influence Estimators

Regression LMS-Estimates

A lot of approaches rely on decreasing a more effective scale estimate against the sum of squared residuals. For example, Rousseeuw and Yohai (1984) have presented a high breakdown approach known as 'least median of squares' (LMS), which is defined by minimizing the median of squared residuals as opposed to their total,

$$Med \left(\frac{y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}}{\hat{\sigma}} \right) = \min . \tag{2.14}$$

The LMS achieves the highest breakdown point value, $BP = ([(n-p)/2] + 1)/n$. This means that LMS fit stays in a bounded region whenever $[(n-p)/2]$ or fewer observations are outliers (Rousseeuw and Van Driessen, 2006).

Regression LTS-Estimates

According to Rousseeuw (1984), the least trimmed squares (LTS) is the other high breakdown and bounded influence estimator that minimize,

$$\hat{\boldsymbol{\beta}}_{(LTS,H,N)} = \arg\min \sum_{i=1}^{H} r_{[i]}^2(\boldsymbol{\beta}), \tag{2.15}$$

where, $H \in 1, ..., n$, and $\left|r_{[1]}\right| \leq \left|r_{[2]}\right| \leq ... \leq \left|r_{[n]}\right|$ denote the ordered absolute residuals. When $H = n/2$ is equivalent to finds the estimates corresponding to the half samples

having the smallest sum of squares of residuals. As such, breakdown point is 50%. When $H = [(n + p + 1)/2]$ is equivalent to LMS and when $H = n$, LTS and LS coincide: $\hat{\boldsymbol{\beta}}_{(LTS,n,N)} = \hat{\boldsymbol{\beta}}_{(LS,N)}$.

Regression Biweight S-Estimate

According to Rousseeuw and Yohai (1984), S-estimates are defined by:

$$\hat{\sigma}\left(r_1(\hat{\boldsymbol{\beta}}), ..., r_n(\hat{\boldsymbol{\beta}})\right) = \min, \tag{2.16}$$

where, $\hat{\sigma}\left(r_1(\hat{\boldsymbol{\beta}}), ..., r_n(\hat{\boldsymbol{\beta}})\right)$ is the scale M-estimate which is defined as the solution to

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{r_1(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right) = \delta, \tag{2.17}$$

where, δ is taken to be $E_\Phi[\rho_0(r)]$ and Φ is the standard normal distribution. A commonly used family of loss function ρ_0 is given by Tukey's Biweight function (Beaton and Tukey, 1974),

$$\rho(r; d) = \begin{cases} 3(r/d)^2 - 3(r/d)^4 + (r/d)^6, & \text{if } |r| \leq d, \\ 1, & \text{elsewhere}, \end{cases} \tag{2.18}$$

where, $d = 1.5476$ yields $b = E_\Phi[\rho(Z; d)] = 2(1 - F_0(d))$, with Φ the standard normal cumulative distribution function and $Z \sim N(0, 1)$. Maronna et al. (2006) stated that associated BS-estimator has maximal asymptotic breakdown point 50%.

2.3.2 Outliers Diagnostics in Regression Model

Single Outliers Diagnostics

The ordinary residual vector is defined as $e_i = y_i - \hat{y}_i = (1 - H)y_i$, where $\hat{y}_i$ is the vector of the fitted values and H is the hat or leverage matrix which is a symmetric and idempotent matrix. The matrix H contains the information on the influence of the response value

y_i on the corresponding fitted value $\hat{y}_i = H_i^T y_i$, where H_i^T is the ith row of matrix H. The h_{ii} is the diagonal elements of the hat matrix. Huber (2011) suggested that h_{ii} with values less than 0.2 appearing to be safe, values between 0.2 and 0.5 as being risky and values greater than 0.5, if possible, be avoided by the control of the design matrix. Belsey et al. (1980) suggested an approximation cut-off value at 0.05 level of significance to be $2p/n$, where p is the number of model coefficients.

Well-known Mahalanobis (MD_i) distances is suggested to use as measures of leverage points in the literature (see Leroy and Rousseeuw (1987)), it's defined as: $MD_i = \sqrt{(\mathbf{X}_i - \bar{\mathbf{X}})\mathbf{C}^{-1}(\mathbf{X}_i - \bar{\mathbf{X}})}$, where $\bar{\mathbf{X}}_i = 1/n \sum_{i=1}^{n} \mathbf{X}_i$ is the mean vector and $\mathbf{C} = 1/(n-1) \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})^T (\mathbf{X}_i - \bar{\mathbf{X}})$ is variance-covariance matrix. The values of MD_i^2 are computed with $\chi^2_{p,0.95}$ and observations exceeding that cut-off value are considered as high leverage points.

Another technique is proposed by Rousseeuw and Van Zomeren (1990), they suggest using the least median of squares (LMS) estimated to detect regression outliers. This method begins by computing the residuals associated with LMS regression

$$s = 1.4826 \left(1 + \frac{5}{(n-p-1)}\right) \sqrt{M_r}, \tag{2.19}$$

where, M_r is the median of $r_1^2, ..., r_n^2$, the squared residuals, p is the number of predictors. However, a regression outlier is ith vector that satisfy, $(\mid r_i \mid /s) > 2.5$.

The effect of deleting one row on the estimation of parameters and their covariance, residual sum of squares and fitted values can be used to identify outliers in the data set. First, we look at the effect of outliers on the parameter estimation of $\boldsymbol{\beta}$. Let $\hat{\boldsymbol{\beta}}_{(-i)}$

be the least square estimate of $\boldsymbol{\beta}$ when the ith observation is deleted. Then $\hat{\boldsymbol{\beta}}_{(-i)} = (\mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)})^{-1} \mathbf{X}_{(-i)}^T \mathbf{y}_{(-i)}$ where $\mathbf{X}_{(-i)}$ and $y_{(-i)}$ are obtained by removing the ith row in $\mathbf{X}$ and y, respectively.

The change in the estimate of the parameter vector $\boldsymbol{\beta}$ when the ith observation is deleted is given by

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)} = \frac{(\mathbf{X}_i^T \mathbf{X}_i)^{-1} \mathbf{X}_i^T e_i}{1 - h_{ii}}. \tag{2.20}$$

Hadi (1992) introduced potentials as a single leverage deleted measure define as

$$p_{ii} = \mathbf{x}_i^T (\mathbf{X}_{(i)}^T \mathbf{X}_{(i)})^{-1} \mathbf{x}_i. \tag{2.21}$$

But the problem with this cut-off point is that both mean and variance of p_{ii} may be non-robust in the presence of a single extreme value yielding a high cut-off point. To avoid such a problem Hadi (1992) suggested replacing the mean and the standard deviation in Eqn. (2.21) by the median and the median absolute deviation (MAD) respectively.
A cut-off point for p_{ii} is $Median(p_{ii}) + 3.MAD(p_{ii})$, where MAD is median absolute deviation. Ryan (2008) reviewed a different types of residuals for the diagnostic purpose, the commonly used is Studentized residuals define as

$$t_i = \frac{y_i - \mathbf{X}_i^T \boldsymbol{\beta}^{(-i)}}{\hat{\sigma}_{(i)} \sqrt{1 - h_{ii}}}. \tag{2.22}$$

And observation i is termed as outlier if $|t_i| > c$, where c is a constant value $2 \leq c \leq 3$.

Belsey et al. (1980) introduced $DFFITS$ defined as

$$DFFITS_i = \sqrt{\frac{h_{ii}}{1 - h_{ii}}} t_i \tag{2.23}$$

and Belsey et al. (1980) recommended considering observations as influential if $|DFFITS_i| \geq 2\sqrt{p/n}$. However, the quantity DFFITS is closely related to the well-known Cook's distance proposed by Cook (2000, 1979), which considered a statistic based on the confidence ellipsoids for investigating the contribution of each data point i to the least squares estimate of the parameter, $\boldsymbol{\beta}$, which is given by

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbf{X}_i^T \mathbf{X}_i (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{ps^2} \sim F_{p,n-p}. \tag{2.24}$$

In order to determine the degree of influence of the ith data point on the estimated parameter vector, $\boldsymbol{\beta}$, Cook suggested the measure of the critical nature of each data point to be

$$\begin{aligned} D_{-i} &= \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{(-i)})^T \mathbf{X}_i^T \mathbf{X}_i (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{(-i)})}{ps^2} \sim F_{p,n-p}. \\ &= \frac{e_i^2}{ps^2} \left\{ \frac{h_{ii}}{(1-h_{ii})^2} \right\}. \end{aligned} \tag{2.25}$$

A large value of D_{-i} indicates that the associated observation has a strong influence on the estimate of parameter vector $\hat{\boldsymbol{\beta}}$.

Another technique is to compare the estimated covariance matrix of $\boldsymbol{\beta}$ using all available data, $\sigma^2(\mathbf{X}_i^T \mathbf{X}_i)^{-1}$, with the estimated covariance matrix when the ith observation is deleted, $\sigma^2(\mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)})^{-1}$. The relationship between CD_i and $DFFITS_i$ is given by

$$CD_i = \frac{\hat{\sigma}_{(i)}}{p\hat{\sigma}^2} DFFITS_i^2. \tag{2.26}$$

Finally, Belsey et al. (1980) suggested to compare the two matrices using a determinant

ratio, which is given by

$$COVRATIO_{(-i)} = \frac{det\{s^2_{(-i)}[\mathbf{X}^T_{(-i)}\mathbf{X}_{(-i)}]^{-1}\}}{det\{s^2(\mathbf{X}^T\mathbf{X})^{-1}\}}$$

$$= \left(\frac{s_{(-i)}}{s}\right)^{2p} \frac{1}{1-h_{ii}}. \tag{2.27}$$

A value of $COVRATIO_{(-i)}$ which is not near unity indicates that the ith observation is possibly influential. They further proposed that any data point with $COVRATIO_{(-i)} - 1$ close to or larger than $(3p/n)$ is identified as an outlier.

Imon (2002) proposed a method to identify the suspected outliers and high leverage points using some diagnostic measure. Baum et al. (2003) discussed instrumental variables (IV) estimation in the broader context of the generalized method of moments (GMM), and describe an extended IV estimation routine that provides GMM estimates as well as additional diagnostic tests.

Group Outliers Diagnostics

In large data, the single case deleted measure may be ineffective for identification of multiple influential observations. Suppose that a regression data contains K outliers. Most diagnostic tools seem to be successful to separate the data into a clean subset without outliers and a complementary subset that contains all prospect outliers. However, when delete the subset of K observations, produces the largest reduction in the residual sum of squares. Let the clean subset of R observation remaining in the analysis, hence a clean set contains $(n-k)$ cases after $k < (n-p)$ cases deleted. When a group of observation K is omitted, then the ith diagonal element of the $\mathbf{X}(\mathbf{X}^T_R\mathbf{X}_R)^{-1}\mathbf{X}^T$ matrix define as $h_{ii(R)} = \mathbf{x}^T_i(\mathbf{X}^T_R\mathbf{X}_R)^{-1}\mathbf{x}_i$.

Imon (1996) considered the generalized potentials for all members in the data set that are defined as

$$p_{ii} = \begin{cases} \frac{h_{ii(R)}}{1-h_{ii(R)}}, & \text{for } i \in R, \\ h_{ii(R)}, & \text{for } i \in K. \end{cases} \tag{2.28}$$

Thus, one could consider p_{ii} to be large if $p_{ii} > Median(p_{ii}) + 3MAD(p_{ii})$. Various numbers of diagnostic methods of regression model have been developed in the past (see Leroy and Rousseeuw (1987), Chatterjee and Hadi (2009), and Barnett and Lewis (1994)), for the identification of multiple outliers. Most of these methods attempt to separate the data into a clean subset without outliers and a complementary subset that contain all the potential outliers. The ith external Studentized residual, for the observations remaining in the data set indexed by R, as

$$t_i^\star = \frac{y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}_R}{\hat{\sigma}_{R-i}\sqrt{1 - h(ii)R}}. \tag{2.29}$$

The generalized Studentized residuals (GS) and the generalized cook's distance (GCD) are defined respectively as

$$t_i^\star = \begin{cases} \frac{\hat{\epsilon}_{i(R)}}{\hat{\sigma}_{R-i}\sqrt{1-h_{ii(R)}}}, & \text{for } i \in R, \\ \frac{\hat{\epsilon}_{i(R)}}{\hat{\sigma}_R\sqrt{1+h(ii(R))}}, & \text{for } i \in K. \end{cases} \tag{2.30}$$

Rahmatullah Imon (2005) defined the generalized $DFFITS$ as

$$GDFFITS_i = \begin{cases} \frac{\hat{y}_{i(R)} - \hat{y}_{i(R-i)}}{\hat{\sigma}_{R-i}\sqrt{h_{ii(R)}}}, & \text{for } i \in R, \\ \frac{\hat{y}_{i(R+i)} - \hat{y}_{i(R)}}{\hat{\sigma}_R\sqrt{h_i i(R+i)}}, & \text{for } i \in K, \end{cases} \tag{2.31}$$

and the author considered observations as influential if $|GDFFITS_i| \geq 3\sqrt{p/(n-k)}$.

2.4 Definition of Statistical Functional

Let $X_1, ..., X_n$ be a sample from a population with distribution function F and let $T_n = T_n(X_1, ..., X_n)$ be a statistic. When T_n can be written as a functional T of the empirical distribution function F_n, $T_n = T(F_n)$ where T does not depend on n, then we call T a statistical functional. The domain of T is assumed to contain the empirical distribution functions F_n for all $n > 1$ and the population distribution function F. The range of T is assumed to be $\mathbb{R}$.

2.5 Measuring of Robustness

Several measures of robustness are used in different studies the to explore good properties of the estimates (e.g. (Wilcox, 2012)). The most common measures are breakdown point and influence function. The following sections elaborate some measures.

2.5.1 Influence Function

Influence function (IF) describes the effect of an infinitesimal contamination at x on the estimator T. It is defined as:

$$IF(x_0; T, F) = \lim_{\varepsilon \downarrow 0} \frac{T\left((1-\varepsilon)F + \varepsilon \Delta_{x_0}\right) - T(F)}{\varepsilon}, \tag{2.32}$$

where Δ_{x_0} is the point-mass at x_0.

2.5.2 Hampel's Empirical Influence Function Hampel et al. (2011)

Given a sample of n observations, replace one, say x_n, by an arbitrary x and define the empirical influence function as

$$EIF_n(x; T_n) = T_n(x_1, ..., x_{n-1}, x). \tag{2.33}$$

2.5.3 Tukey's Sensitivity Curve

The sensitivity curve is a tool for evaluating the effect on an estimate of perturbing an observation at a finite sample. Tukey defined a version for addition of an observation as follows (see, Hoaglin et al. (1983)). Given an estimator T_n and a sample $x_1, ..., x_{n-1}$, define the sensitivity curve as a function of an additional observation x scaled by the sample size n. Formally, we have

$$SC_n(x; T_n) = n[T_n(x_1, ..., x_{n-1}, x) - T_{n-1}(x_1, ..., x_{n-1})]. \tag{2.34}$$

2.5.4 Gross-Error Sensitivity (Hampel, 1968)

The influence function given by Eqn. (2.32) may be used to study several robustness properties. One of the simplest and most revealing is the gross-error sensitivity of an estimator T at a distribution F. It is defined by

$$\gamma^\star = \sup_x |IF(x; T, F)|. \tag{2.35}$$

By taking the supremum over all x for which the $IF(x; T, F)$ exists, gross-error sensitivity measures the worst possible influence on an estimator by an arbitrary infinitesimal contaminant. If the gross-error sensitivity is unbounded, $\gamma^\star = \infty$, then the estimator is completely intolerant of outliers; a single outlier can ruin the estimator.

2.5.5 Breakdown Point

The breakdown point of an estimate $\hat{\boldsymbol{\beta}}$ of the parameters $\boldsymbol{\beta}$ is the largest amount of contamination that the data may contain such that $\hat{\boldsymbol{\beta}}$ even turns over some information about $\boldsymbol{\beta}$. In other words, breakdown point of an estimate $\hat{\boldsymbol{\beta}}$ shows the effects of replacing several

data values by outliers (Maronna et al. (2006)). The breakdown point for the regression estimator $\hat{\boldsymbol{\beta}}$ of the sample $Z = (\mathbf{X}, y)$ is defined as

$$\varepsilon^{\star}(\hat{\boldsymbol{\beta}}; Z) = \min \left\{ \frac{m}{n} \sup_{\tilde{Z}} \|\hat{\boldsymbol{\beta}}(Z^T)\|_2 = \infty \right\}, \tag{2.36}$$

where $\tilde{Z}$ are contaminated data obtained from Z by replacing m of the original n data by outliers.

2.5.6 The Properties of LS and M Estimators

LS-estimates have a breakdown point equals $1/n$, which tends to zero when the sample size n gets large. Therefore, one single unusual observation can cause LS-estimates breakdown. That is, the IF of LS-estimate given by:

$$IF(\mathbf{X}_i, y_i; T, F) = V^{-1}(\psi, F)\mathbf{X}_i\mathbf{X}_i^T\psi\left(y_i - \mathbf{X}_i^T T(F)\right), \tag{2.37}$$

where $\psi = \rho' = 2\mathbf{X}_i(y_i - \mathbf{X}_i^T\boldsymbol{\beta})$ and V is a certain $p \times p$ matrix given by:

$$V(\psi, F) = \int \psi'(y_i - \mathbf{X}_i^T T(F))\mathbf{X}_i\mathbf{X}_i^T dF(\mathbf{X}_i, y_i) = E\left[\rho'(\varepsilon)\varepsilon\right]. \tag{2.38}$$

However, that is, the influence function (IF) of M-estimator given by:

$$IF(\mathbf{X}_i, y_i; T, F) = V^{-1}(\psi, F)\mathbf{X}_i\mathbf{X}_i^T\psi\left(y_i - \mathbf{X}_i^T T(F)\right), \tag{2.39}$$

where $\psi = \rho'$ bounded function and V is as in Eqn. (2.38). It is remarkable that LS-estimator is sensitive to outliers and M-estimator is sensitive to leverage point.

2.6 Robust Variable Selection

The classical algorithm namely $LASSO$ and weighted $LASSO$ are much affected by outliers and often fails to select the correct linear prediction model that would have been chosen if there were no outliers. Additionally, a number of studies have proposed different approaches to deal with the outliers in response variables for the large data sets. However, seminal papers addressed the robust $LASSO$, Least Absolute Deviation (LAD-$LASSO$) Wang et al. (2007), Huber M-estimation function (Huber-$LASSO$) Lambert-Lacroix and Zwald (2011), and least trimmed squared (sparse LTS) Alfons et al. (2013). Several classical selection variable methods in large data are discussed in this section.

2.6.1 Robust Variable Selection Methods in Small Samples

Robust model selection procedures received a great deal of interest in the mid 90's, mainly to improve existing selection criteria. These procedures limit the influence of outliers on the chosen models. In this regard, Rousseeuw (1985) have proposed a robust version of the selection criteria AIC. Later, Anderson-Sprecher (1994) and Croux and Dehon (2003) have introduced the robust R^2 that, relies on the effective scale S and the class of M-estimator of residual scale selected, correspondingly. On the other hand, Ronchetti and Staudte (1994) studied the robust version of Cp. Similarly a robust version of SIC has been proposed by Machado (1993). In another study, a robust analogue to the classical FPE criterion was proposed by Ronchetti et al. (1997), in which ideal requirements rely upon the objective functions that, determine the M-estimators for a parametric model.

It is worth to mention that the influence function of M-estimator with respect to y_i can be bounded, but it is unbounded with respect to X- direction. A number of studies have used an alternative to robust variable selection based on M-estimation. For example,

Tharmaratnam and Claeskens (2013) improved the performance of the method proposed by Rousseeuw (1985) by using S and MM-estimators. Table 2.2 shows the most applicable methods with regard to the related formulas and references. Here, $\hat{\sigma}$ is some robust

Table 2.2: Robust variable selection criteria

Methods	Formulas	References
The Robust R^2	$R^2_{L_1} = 1 - \left(\frac{\sum_{i=1}^{n} \lvert y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{L_1} - \hat{\mu}_{L_1} \rvert}{\sum_{i=1}^{n} \lvert y_i - median_i(y_i) \rvert} \right)$	Anderson-Sprecher (1994)
The Robust Version of AIC	$RAIC = \sum_i \rho \left(\frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}}{\hat{\sigma}} \right) + \alpha p$	Rousseeuw (1985)
The Robust Version of Cp	$RCp = \frac{W_p}{\hat{\sigma}^2} - (U_p - V_p)$	Ronchetti and Staudte (1994)
The Robust Version of SIC	$RSIC = \sum_{i=1}^{n} \rho(\frac{r_i}{\sigma}) + \frac{p \log(n)}{n}$	Machado (1993)
The Robust FPE	$RFPE = \sum_{i=1}^{n} E \left[\rho \left(\frac{y_i - \mathbf{x}_i^T \hat{\beta}_p}{\sigma} \right) \right]$	Ronchetti, Field, & Blanchard (1997)

estimate of σ, $\hat{\boldsymbol{\beta}}$ is the M-estimator of $\boldsymbol{\beta}$, and

$$\alpha = 2E \left[\psi^2 \left((y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}) / \hat{\sigma} \right) \right] / (E \left[\psi' \left((y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}) / \hat{\sigma} \right) \right].$$

However, $W_p = \sum \hat{w}_i^2 r_i^2$ is the weighted residual sum of squares, $\hat{\sigma}^2$ is a robust and consistent estimate of σ^2 from the full model, and U_p and V_p are constant depending on the weight function and the number of parameters p given by:

$$V_p = tr(RM^{-1}QM^{-1}),$$

and

$$U_p - V_p = E||\rho||^2 - 2tr(NM^{-1}) + tr(LM^{-1}QM^{-1}).$$

In addition, $M = E[\rho'(\mathbf{X}_i, \varepsilon)\mathbf{X}_i\mathbf{X}_i^T]$ with ρ' denoting the derivative of ρ with respect to its second argument.

$Q = E \left[\rho^2(\mathbf{X}_i, \varepsilon)\mathbf{X}_i\mathbf{X}_i^T \right]$, $||\rho||^2 = \sum_{1 \leq i \leq n} \rho^2(\mathbf{X}_i, \varepsilon_i)$, $N = E[\rho^2 \rho' \mathbf{X}\mathbf{X}_i^T]$, $L = E[\hat{w}\varepsilon((w'\varepsilon) + 4w)\mathbf{X}\mathbf{X}_i^T] = E[(\hat{\rho})^2 + 2\rho' w - 3w^2)\mathbf{X}\mathbf{X}_i^T]$, and, $R = E[w^2 \mathbf{X}_i \mathbf{X}_i^T]$. If sub model holds, $\hat{\sigma}^2 \approx w_p/U_p$, and $RCp \approx V_p$. Therefore, models with values of RCp which are close

to V_p or smaller than V_p will be preferred to others and a plot of RCp versus V_p will aid in this selection. The RCp values were calculated with the S-Plus routine RCp using Huber's function with M equal to 1.345. When the weight are identically 1, RCp reduces to Mallows C_p.

2.6.2 Robust Variable Selection Methods in Large Data Sets

LASSO Least Absolute Deviation (*LAD-LASSO*)

The robust $LASSO$ regression estimator proposed by Wang et al. (2007), used the penalty least absolute deviation (LAD), written in Eqn. (2.13), the LAD-$LASSO$ estimator is computed using the following criterion:

$$\hat{\boldsymbol{\beta}}_{LAD-LASSO} = \min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} \left| y_i - \mathbf{X}_i^T \boldsymbol{\beta} \right| + n \sum_{j=1}^{p} \lambda_j |\beta_j| \right], \tag{2.40}$$

where λ is the tuning parameter determined as in $LASSO$ estimator. LAD-$LASSO$ can simultaneously estimate parameters, and perform variable selection. In addition, it is resistant to heavy-tailed errors or outliers in the response. With proper choice of tuning parameters, the LAD-$LASSO$ estimator also enjoys the oracle property (Wang et al., 2007).

Note that, the squared loss has been replaced by the L_1 loss. Unfortunately, this loss is not adapted for small errors: it strongly penalizes the small residuals. In particular, when the error has no heavy tail and do not suffer from outliers, this estimator is expected to be less efficient than the ada-$LASSO$. As a result, Lambert-Lacroix and Zwald (2011),

preferred to consider the criterion like M-estimation as a loss function. Thus,

$$\hat{\boldsymbol{\beta}}_{M-LASSO} = \min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} \rho \left(y_i - \mathbf{X}_i^T \boldsymbol{\beta} \right) + n \sum_{j=1}^{p} \lambda_j \mid \beta_j \mid \right]. \tag{2.41}$$

***LASSO* Through The *M*-estimators (Huber-*ada-LASSO*)**

Lambert-Lacroix and Zwald (2011) suggested robust $LASSO$ regression by combining the Huber's criterion and $ada\text{-}LASSO$ penalty ($\lambda_n \sum_{j=1}^{p} \hat{w}_j \mid \beta_j \mid$). The Huber criteria given in Eqn. (2.3.1), is quadratic in small values of r, but linearly grows for large values of r. When $M \to \infty$, then $\hat{\boldsymbol{\beta}}_{M-LASSO}$ is $\hat{\boldsymbol{\beta}}_{LASSO}$. Likewise, when $M \to 0$, then $\hat{\boldsymbol{\beta}}_{M-LASSO}$ is $\hat{\boldsymbol{\beta}}_{LAD-LASSO}$. Thus, by choosing an appropriate cut-off M, $\hat{\boldsymbol{\beta}}_M$ is robust and efficient at the normal distribution.

In order to allow for oracle properties, Lambert-Lacroix and Zwald (2011) consider Huber's criterion with concomitant scale defined by,

$$L_\rho(\mu, \boldsymbol{\beta}, s) = ns + \sum_{i=1}^{n} \rho \left(\frac{(y_i - \mathbf{X}_i^T \boldsymbol{\beta})}{s} \right) s, s > 0. \tag{2.42}$$

2.7 Summary

We have reviewed the variable selection methods, classical and robust variable selection criteria of small and large data sets in this chapter. We have also looked at the robust procedure measuring of robustness in linear regression. We intend to propose other robust criteria to the both cases low and large data sets in the subsequent chapters.

CHAPTER 3

EFFECT OF OUTLIERS ON DIFFERENT VARIABLE SELECTION CRITERIA

3.1 Introduction

This chapter is aimed to illustrated the problem of outliers and leverage points in existing robust model selection for small data set. Then, the theory of $LASSO$ and $ada\text{-}LASSO$ variable selection for regression model based on consistency and oracle properties discussed, too. The simulation study has also been carried out to compare ridge, $LASSO$ and $ada\text{-}LASSO$ procedures. The application of the Ozone data set was presented. A simulation study has also been carried out to see the effect of leverage points on the robust variable selection methods based on M-estimation.

3.2 The Effect of Outliers in Different Variable Selection Criterion

In statistical analysis, the existence of outlying values in the data set should raise concern. The existence of outliers in linear data sets and linear regression has been investigated extensively (c.f. Beckman and Cook (1983); Barnett and Lewis (1994); Belsey et al. (1980); Montgomery et al. (2012)). The effect of outliers on variable selection methods is known to be severe.

It is essential to study the behaviour of the classical and robust variable selection, in the presence of vertical and leverage point before dealing with the presence of outliers in different variable selections methods.

Experiment 1 aims to study the behavior of the classical variable selection criterion (AIC, R^2, Cp, SIC, FPE, HQ) in the presence of vertical outlier and or leverage point. Furthermore, experiment 2, study the performance of robust variable selection methods (R^2_M, $RAIC$, RCp, $RSIC$, $RFPE$) in the presence of vertical and leverage point. The design of experiment 1 is briefly discussed in subsection 3.2.1. In experiment 1, both types of influence points affect the classical criterion; also, the robust criterion based on M-estimators are affected by leverage points.

3.2.1 Experiment 1

For simplicity, a set of independent random uniform variable **X** on [-2,2] was generated according to the simple regression model given as follows:

$$y_i = \mathbf{X}_i + \varepsilon_i, i = 1, ..., 19 \tag{3.1}$$

where, the ε_i are iid, normally distributed with expectation 0 and variance (0.1^2). The data has been presented in Table 3.1 and Figure (3.1). For the purpose of the present study, only the problem that appears in more-complex situations is highlighted without a model selection procedure.

In the first part of this experiment, the coordinate (0,y_{10}) is added, then the value of y ranges between (-1.5, 3). Figure (3.2) shows the situation. A similar approach used for leverage points, by replacing the value **X** with (0, x_{10}), then the value of x_{10} ranges between [2.5, 4.5] and [-2.5,-4.5] (Figure (3.3)).

For each of the 10 values of y and 10 values of $\mathbf{X}_i$, different classical variable selection

criteria (AIC, R^2, Cp, SIC, FPE, HQ) were recomputed.

Table 3.1: The data set

$\mathbf{X}_i$	y_i
-1.2	1.2
-1.15	1.35
-1.1	1.02
-1.05	1.16
-1	0.95
-0.95	1.05
-0.9	0.73
-0.85	0.91
-0.8	0.85
x_{10}	y_{10}
0.8	-0.88
0.85	-0.61
0.9	-0.81
0.95	-0.97
1	-1.18
1.05	-1.08
1.1	-0.99
1.15	-1.11
1.2	-1.14

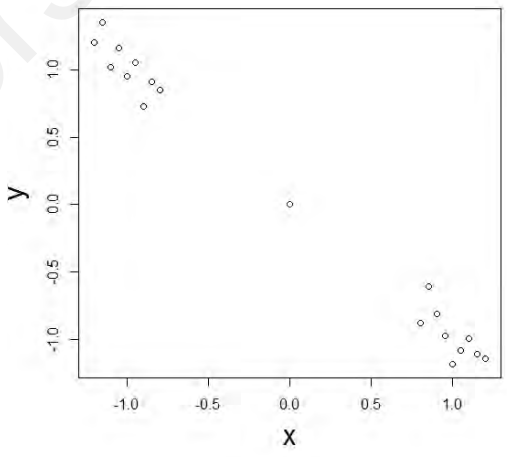


Figure 3.1: The data set

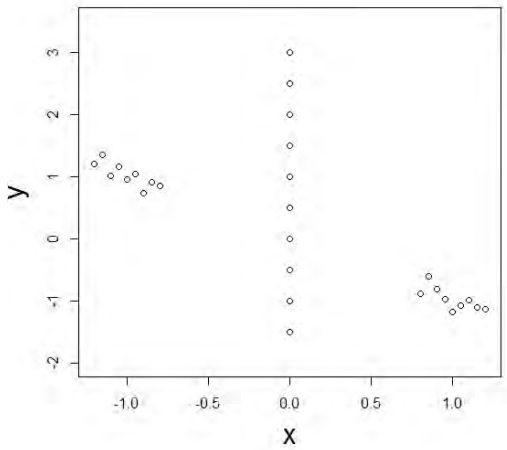


Figure 3.2: Data and positions for y_{10}

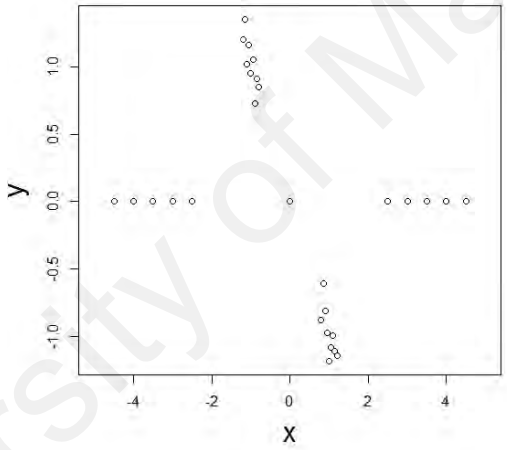


Figure 3.3: Data and positions for x_{10}

In the same manner, different robust classical variable selection criteria ($RAIC$, R^2_M, RCp, $RSIC$, $RFPE$, and RHQ) were recomputed.

3.2.2 Result- Experiment 1

Tables 3.2 and 3.3, reported the performance results of different classical variable selection criterion for each value of y_{10} and x_{10}. When $(x_{10}, y_{10}) = (0, 0)$, the lowest values of AIC, Cp, SIC, FPE, HQ and highest values of R^2 were obtained. In the presence of the y_{10} or x_{10}, the value of AIC, Cp, SIC, FPE, HQ increased and the value of R^2

decreased.

Figures (3.4) to (3.11) (left panel) confirmed that the classical variable selection criteria are highly sensitive to single vertical outlier. Furthermore, by letting y_{10} or x_{10} tend to infinity, AIC, Cp, SIC, FPE, HQ will even tend to infinity and R^2 will tend to zero.

Table 3.2: Different classical variable selection criterion for different value of y_{10} (vertical outliers)

y_{10}	AIC	R^2	Cp	SIC	FPE	HQ
3	3.2	0.680	12.746	-0.482	-6.254	-0.565
2.5	2.9	0.753	12.392	-0.839	-6.608	-0.922
2	2.4	0.824	11.821	-1.270	-7.179	-1.353
1.5	1.9	0.889	10.805	-1.811	-8.195	-1.893
1	1.2	0.942	8.768	-2.523	-10.231	-2.605
0.5	0.2	0.977	4.511	-3.475	-14.489	-3.557
0	**-0.5**	**0.988**	**1.165**	**-4.156**	**-17.835**	**-4.238**
-0.5	0.3	0.975	7.560	-3.371	-11.439	-3.453
-1	1.2	0.938	11.796	-2.441	-7.204	-2.524
-1.5	1.9	0.883	13.269	-1.750	-5.731	-1.833

Table 3.3: Different classical variable selection criterion for different value of x_{10} (leverage point)

x_{10}	AIC	R^2	Cp	SIC	FPE	HQ
4.5	3.3	0.488	565.78	-0.288	546.78	-0.467
4	3.2	0.546	469.76	-0.384	450.765	-0.587
3.5	3.0	0.610	366.08	-0.504	347.080	-0.739
3	2.8	0.679	258.889	-0.657	239.889	-0.935
2.5	2.6	0.751	156.079	-0.852	137.079	-1.189
0	**-0.5**	**0.988**	**1.165**	**-1.107**	**-17.835**	**-4.238**
-2.5	2.6	0.742	131.244	-4.156	112.244	-1.151
-3	2.9	0.669	209.494	-1.069	200.054	-0.902
-3.5	3.1	0.599	215.316	-0.819	293.631	-0.712
-4	3.2	0.535	220.160	-0.629	386.285	-0.563

Tables 3.4 and 3.5, reported the performance results of different robust variable selection criteria for each value of y_{10} and x_{10}. When $(x_{10}, y_{10}) = (0, 0)$, the lowest value of $RAIC$, RCp,$RSIC$, and the highest value of R^2 obtained. In the presence of y_{10} (Table 3.4), all robust criteria became constant when the outliers moved further a way from the

origin. Figures (3.4) to (3.11) (right panel) confirmed that the robust variable selection criteria based on M-estimator are stable to single vertical.

The performance of robust criteria in the presence of leverage point x_{10} were reported in Table 3.5. The values of $RAIC$, RCp, and $RSIC$, increased, and the value of R^2 decreased. Figures (3.4) to (3.11) (right panel) confirmed that the robust variable selection criteria, based on M-estimators are highly affected to single leverage point.

Figures (3.4) to (3.11) summarized the results of the variable selection in regression model. Robust criteria, based on M-estimators were much more robust than classical methods (based on LS), and suffered from leverage points than from vertical outliers.

Table 3.4: Different robust variable selection criterion for different value of y_{10} (vertical outliers)

y_{10}	$RAIC$	R^2_M	RCp	$RSIC$	$RFPE$	RHQ
3	3.7	0.922	-4.954	-4.023	-23.954	-4.350
2.5	3.7	0.920	-4.954	-4.023	-23.954	-4.350
2	3.7	0.920	-4.954	-4.023	-23.954	-4.350
1.5	3.7	0.922	-4.954	-4.022	-23.954	-4.350
1	3.7	0.916	-4.954	-4.022	-23.954	-4.350
0.5	3.7	0.918	-4.953	-4.022	-23.954	-4.350
0	**3.6**	**0.920**	**1.612**	**-4.094**	**-17.388**	**-4.289**
-0.5	3.7	0.917	-1.295	-4.022	-20.295	-4.039
-1	3.7	0.916	-1.294	-4.022	-20.294	-4.039
-1.5	3.7	0.920	-1.285	-4.022	-20.285	-4.036

Table 3.5: Different robust variable selection criterion for different value of x_{10} (leverage point)

x_{10}	$RAIC$	R^2_M	RCp	$RSIC$	$RFPE$	RHQ
4.5	3.9	0.909	30.472	-3.817	5.677	-3.419
4	3.7	0.916	15.392	-3.968	-0.095	-3.579
3.5	3.6	0.921	2.801	-4.108	-5.415	-3.747
3	3.6	0.923	0.692	-4.139	-9.576	-3.904
2.5	3.7	0.917	-0.363	-4.021	-11.219	-3.974
0	**3.5**	**0.921**	**1.612**	**-4.094**	**-17.388**	**-4.289**
-2.5	3.7	0.917	-5.727	-4.021	-19.569	-4.430
-3	3.7	0.922	-6.373	-4.139	-20.573	-4.503
-3.5	3.9	0.921	-4.110	-3.817	-17.052	-4.269
-4	4	0.915	-2.329	-3.662	-14.279	-4.118

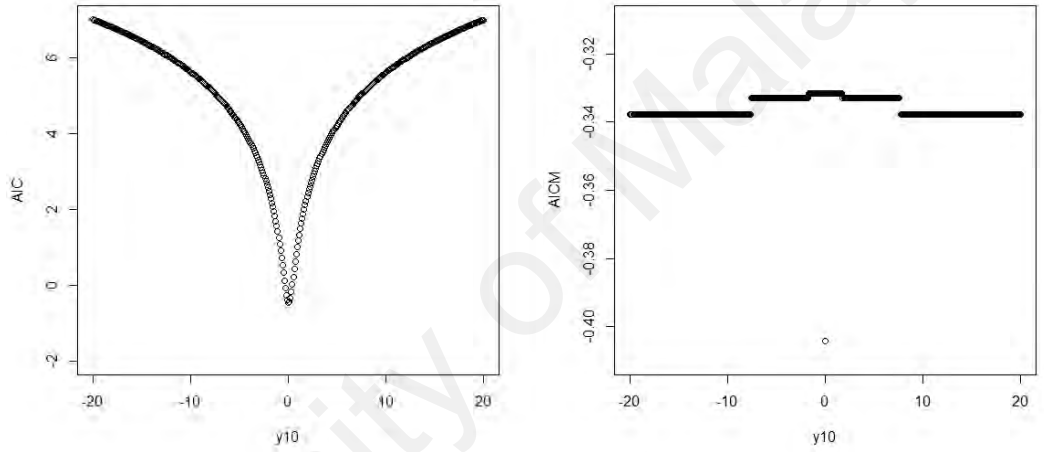


Figure 3.4: Effect of adding one observation (0,y_{10}) on the values of AIC and $RAIC$

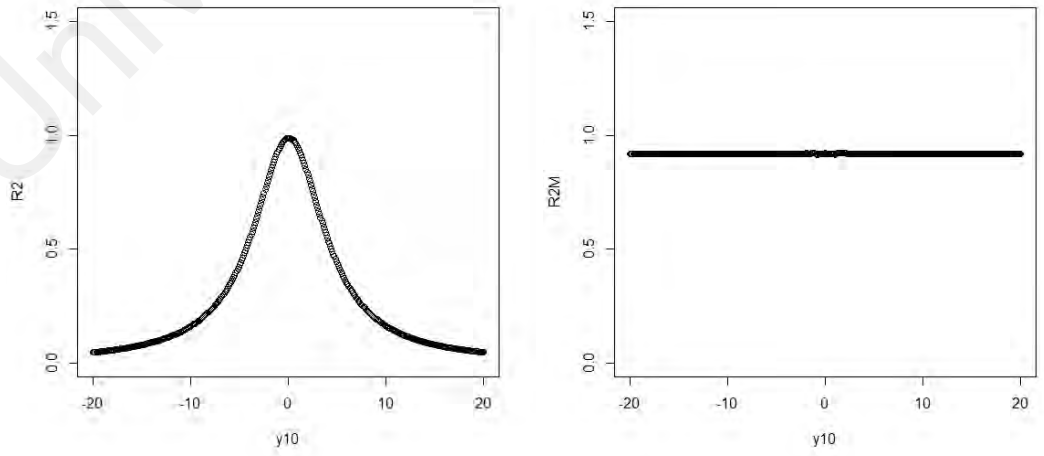


Figure 3.5: Effect of adding one observation (0,y_{10}) on the values of R^2 and R^2_M

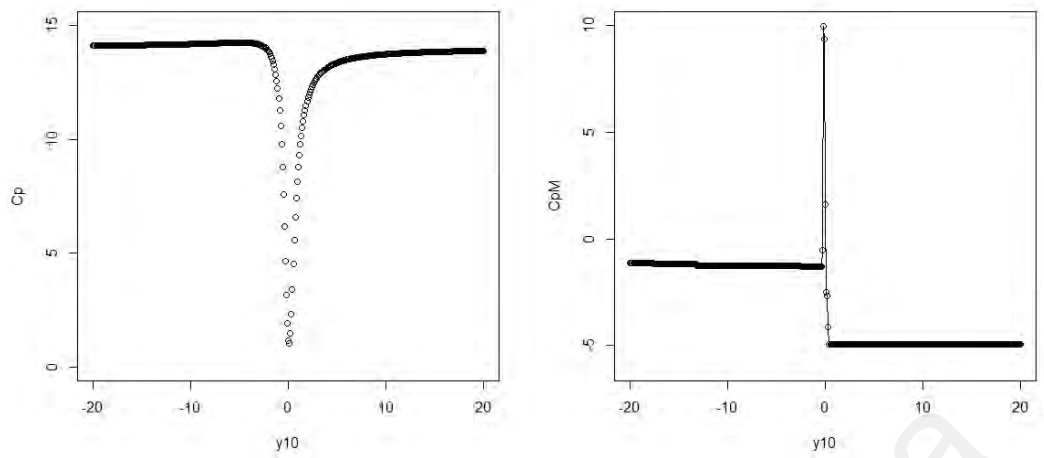


Figure 3.6: Effect of adding one observation (0,y_{10}) on the values of Cp and RCp

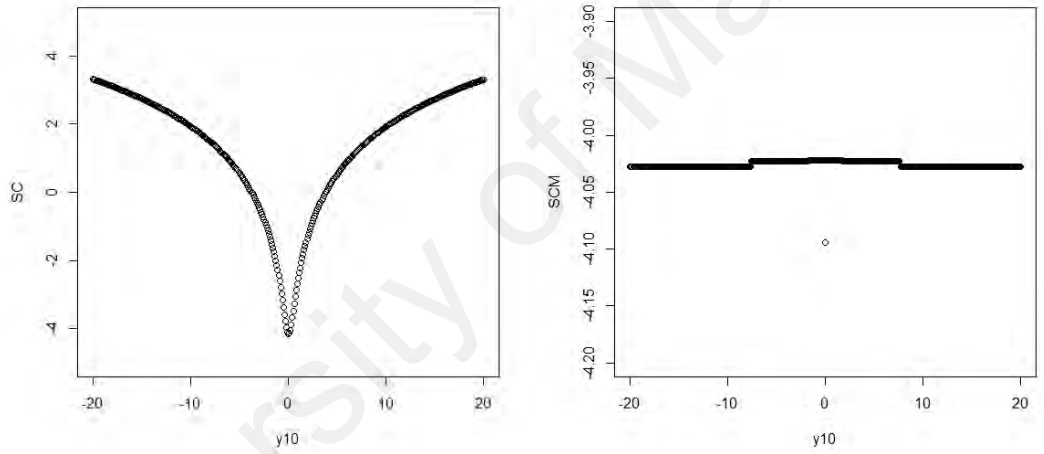


Figure 3.7: Effect of adding one observation (0,y_{10}) on the values of SIC and $RSIC$

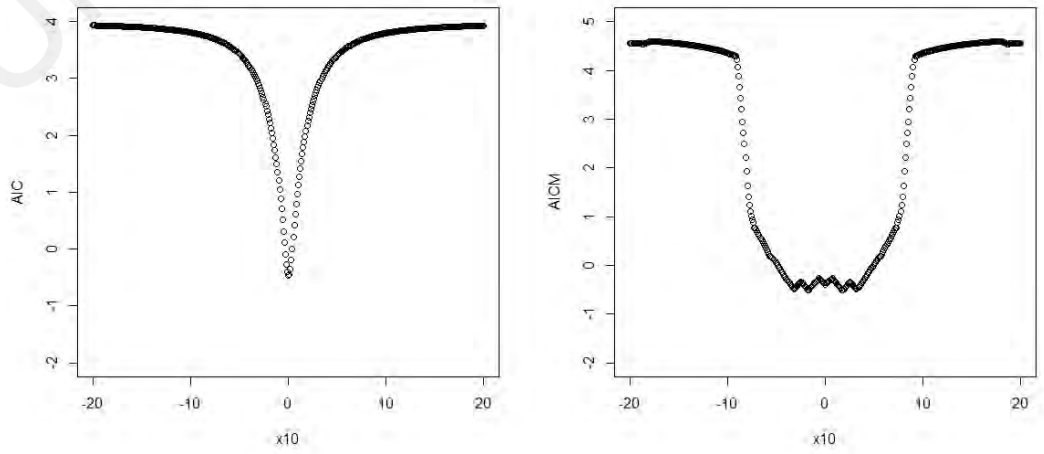


Figure 3.8: Effect of adding one observation (x_{10},0) on the values of AIC and $RAIC$

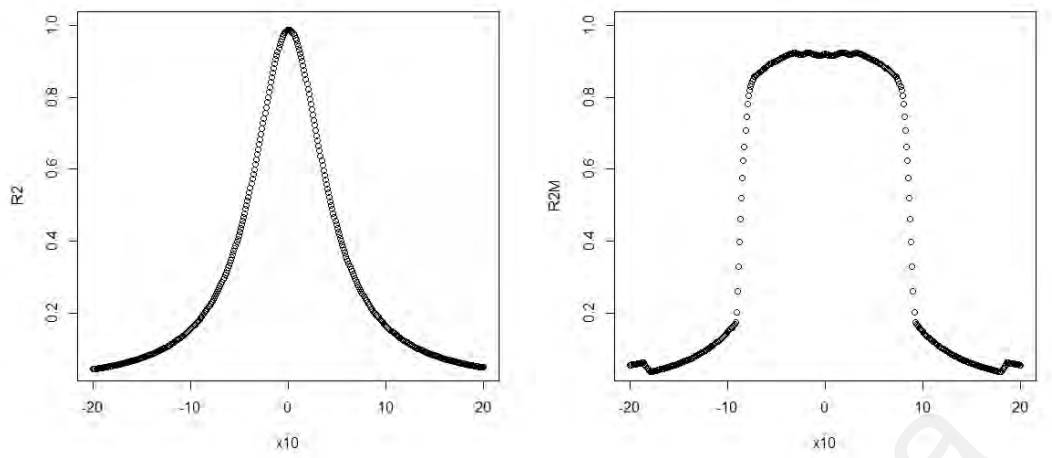


Figure 3.9: Effect of adding one observation (x_{10},0) on the values of R^2 and R^2_M

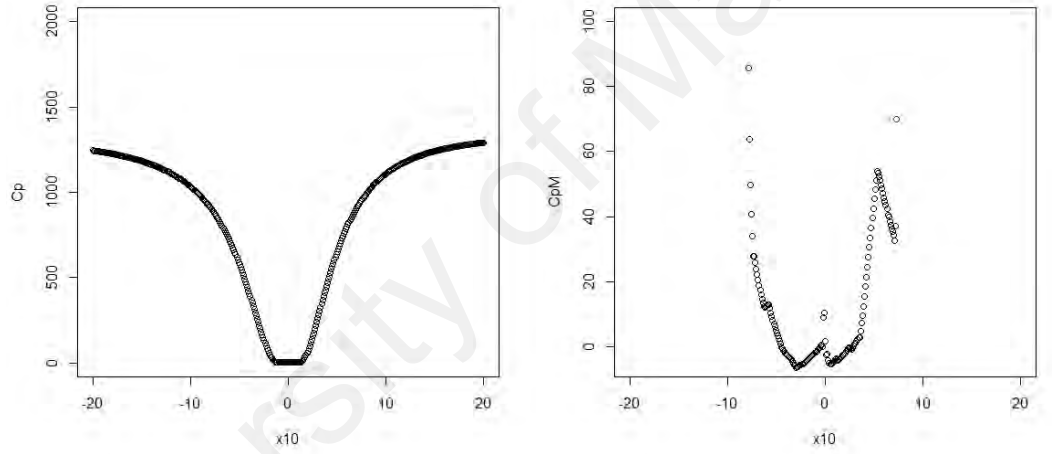


Figure 3.10: Effect of adding one observation (x_{10},0) on the values of Cp and RCp

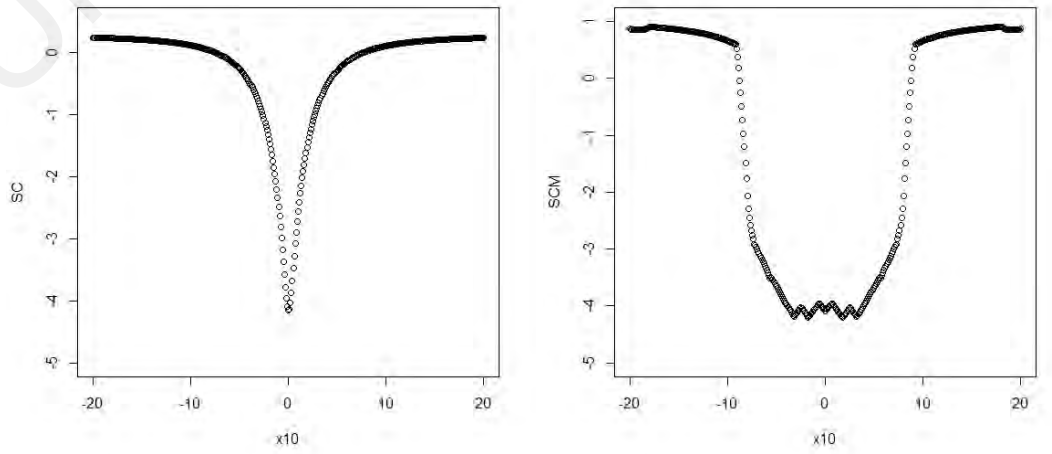


Figure 3.11: Effect of adding one observation (x_{10},0) on the values of SIC and $RSIC$

3.3 Practical Example

In this section, the effect of outliers and leverage point on classical and robust variable selection methods is discussed regarding two real data sets: Belgian Telephone data Leroy and Rousseeuw (1987) and Hawkins-Bradu-Kass data (Hawkins et al. (1984)).

3.3.1 Belgian Telephone Data

There is one variable in data Belgian Statistical Survey. The data set comprising the total number (in tens of millions) of international phone calls made between the years 1950 and 1973, is presented in Appendix 1 and the scatter plot of phone via cell is depicted in Figure (3.12). This time series of data contains heavy contamination from 1964 to 1969.

In the presence of outliers, Table 3.6 shows the low value of R^2 and high values of AIC, SIC, FPE, HQ and this suggested that there may be a scope to improve the fitting of the model. When all outliers were omitted from data, then the high value of R^2 and the low values of AIC, SIC, FPE and HQ obtained. Figure (3.13) shows all situations, the horizon axis of Figure (3.13) denotes different classical criteria, in which "1" indicates R^2, "2" indicates to AIC, "3" indicates to Cp, "4" indicates to SIC, "5" indicates to FPE, and "6" indicates to HQ. The vertical axis represents the value of the corresponding criterion with (blue line) and without outliers (red line). It can be seen that the blue line obtained higher values, that means the value of criteria are affected by outliers. Accordingly, classical method criteria are not foolproof methods of variable selection.

Table 3.6: Different classical variable selection criteria for Belgian Telephone data with and without outliers

	Set of variables	AIC	R^2	Cp	SC	FPE	HQ
With outliers	(y,**X**)	3.53	0.296	2.00	3.63	34.25	3.56
Without outliers	(y,**X**)	-1.5	0.837	2.00	-1.4	0.21	-1.5

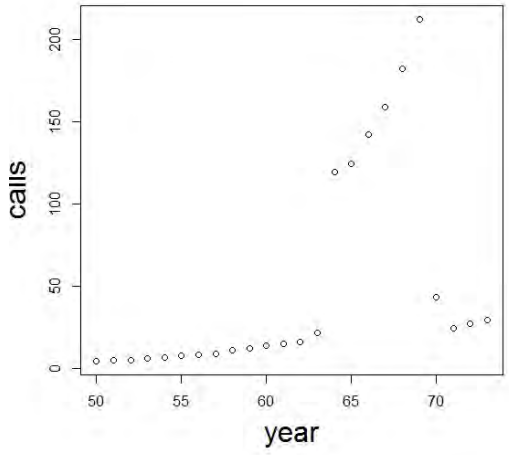


Figure 3.12: Scatter plot of phone cell via year

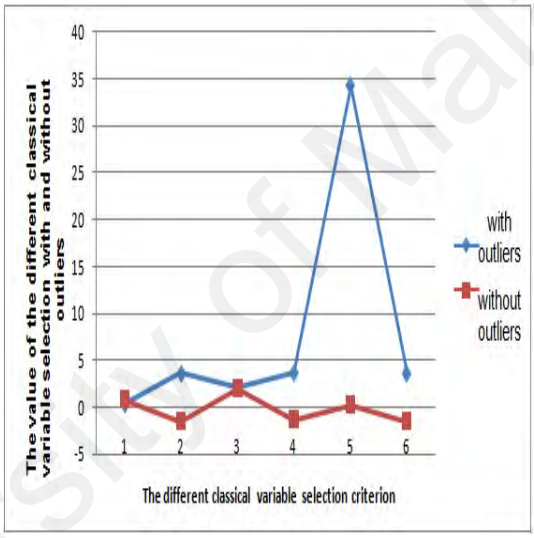


Figure 3.13: Values of different variable selection with and without outliers for Belgian Telephone data ($1 \rightarrow R^2$, $2 \rightarrow AIC$, $3 \rightarrow Cp$, $4 \rightarrow SIC$, $5 \rightarrow FPE$, $6 \rightarrow HQ$)

3.3.2 Hawkins-Bradu-Kass Data

This data has been generated by Hawkins et al. (1984) for illustrating some of the merits of robust technique (the full data set is given in Appendix 2). They pointed out that the first 10 observations are bad leverage points; i.e. the first 10 observations are outliers and the next 4 observations are good leverage points (see Imon, 2005). Figures (3.14) to (3.16) showed the regression plot of y_i via different variables (Hawkins, Bradu, Kass, 1984).

Table 3.7 shows all robust criteria agree on the importance of one variable, Kass, which appears in high value of R^2_M and low values of AIC, RCP, and $RSIC$. Table 3.8 shows the result when all outliers and leverage points were omitted; the value of R^2_M is larger than that of the value with outliers, and the values of the other criteria are smaller than those values with outliers.

Figures (3.17) to (3.20) compare the values of robust criterion for different cases versus the number of set of variables in both situations, with and without outliers. The small values of criteria are considered to show the best model.

Considering both examples, it can be concluded that the presence of outliers affect the classical variable selection. Whereas, the robust variable selection methods based on M-estimator have been affected by the presence of leverage points in data.

Table 3.7: The values of different robust variable selection criterion for Hawkins-Bradu-Kass data for different cases with contamination points

Set of variables	$RAIC$	R^2_M	RCp	$RSIC$
(y,Hawkins)	4.81	0.987	130.77	0.9
(y,Bradu)	4.14	0.988	32.55	0.26
(y,Kass)	**3.62**	**0.991**	**-9.72**	**-0.26**
(y,Hawkins,Bradu)	5.62	0.988	-7.53	-0.20
(y,Hawkins,Kass)	5.79	0.991	4.13	-0.03
(y,Bradu,Kass)	5.67	0.990	-4.38	-0.15
(y,Hawkins,Bradu,Kass)	7.76	0.991	4.00	-0.15

Table 3.8: The values of different robust variable selection criterion for different cases without contamination points

Set of variables	$RAIC$	R^2_M	RCp	$RSIC$
(y,Hawkins)	**3.39**	0.976	15.47	-0.47
(y,Bradu)	**3.25**	0.975	5.93	-0.61
(y,Kass)	**3.38**	0.976	14.19	-0.48
(y,Hawkins,Bradu)	5.37	0.976	15.98	-0.42
(y,Hawkins,Kass)	5.22	0.976	5.89	-0.57
(y,Bradu,Kass)	5.33	0.976	13.27	-0.46
(y,Hawkins,Bradu,Kass)	7.15	**0.977**	**4.00**	**-0.75**

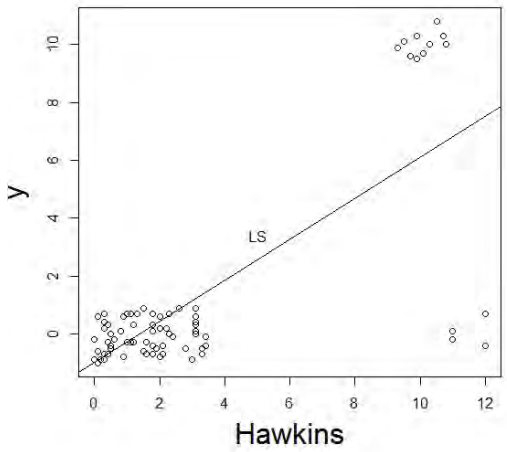


Figure 3.14: The regression plot of y via Hawkins

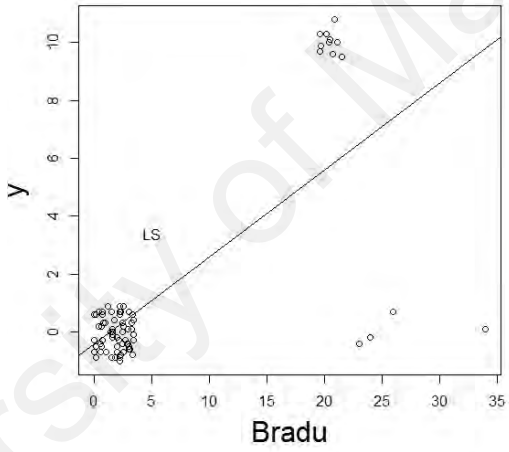


Figure 3.15: The regression plot of y via Bradu

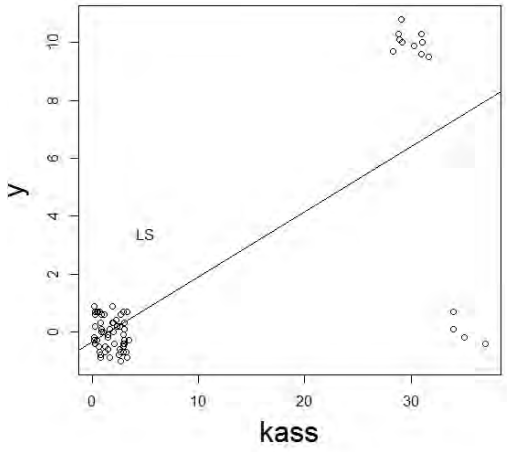


Figure 3.16: The regression plot of y via Kass

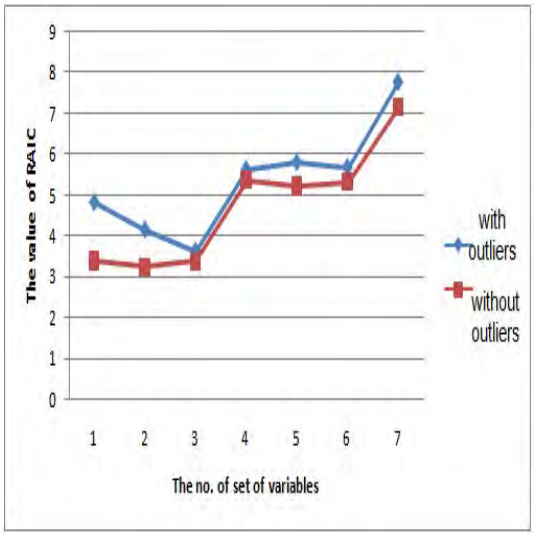


Figure 3.17: The value of $RAIC$ for different cases versus the no.of set of variables (1→Hawkins, 2→Bradu, 3→Kass, 4→(Hawkins,Bradu), 5→(Hawkins,Kass), 6→(Bradu,Kass), 7→(Hawkins,Bradu,Kass))

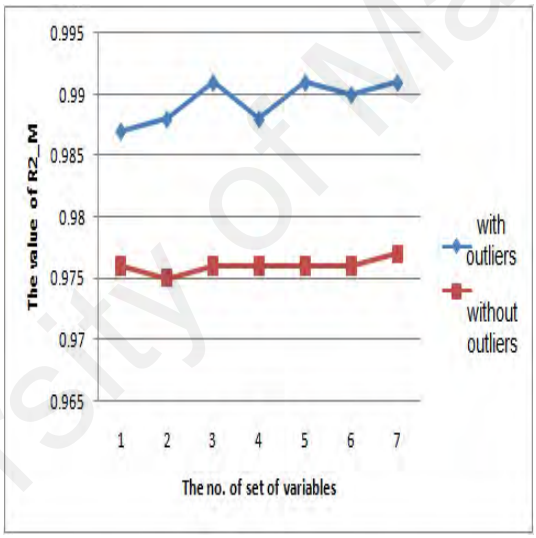


Figure 3.18: The value of R^2_M for different cases versus the no.of set of variables (1→Hawkins, 2→Bradu, 3→Kass, 4→(Hawkins,Bradu), 5→(Hawkins,Kass), 6→(Bradu,Kass), 7→(Hawkins,Bradu,Kass))

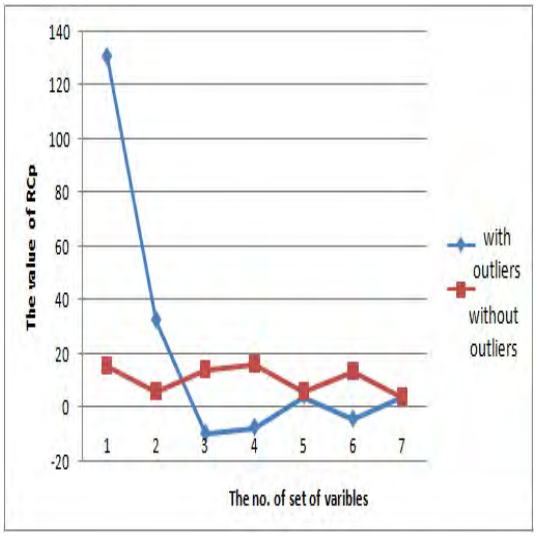


Figure 3.19: The value of RCp for different cases versus the no.of set of variables ((1→Hawkins, 2→Bradu, 3→Kass, 4→(Hawkins,Bradu), 5→(Hawkins,Kass), 6→(Bradu,Kass), 7→(Hawkins,Bradu,Kass))

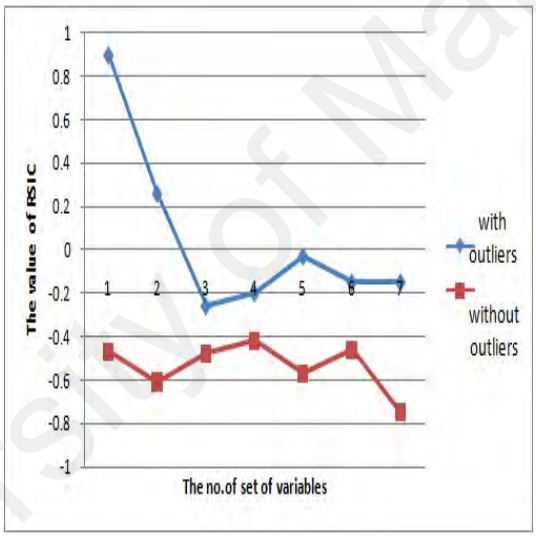


Figure 3.20: The value of $RSIC$ for different cases versus the no.of set of variables (1→Hawkins, 2→Bradu, 3→Kass, 4→(Hawkins,Bradu) , 5→(Hawkins,Kass), 6→(Bradu,Kass), 7→(Hawkins,Bradu,Kass))

3.4 The $LASSO$ Variable Selection and Consistency

Although $LASSO$ has shown successes in many situations, it has some limitations. One limitation of $LASSO$ is that, in the case of $p > n$, $LASSO$ can only, select at most n variables (Osborne et al., 2000). Another limitation is inconsistency in variable selection; Donoho and Huo (2001), Donoho and Elad (2003), and Donoho (2006) have showed that, L_1 approach is able to discover the sparse representation of the model, under certain con-

ditions. It has also been shown that, variable selection with $LASSO$ can be consistent, if the underlying model satisfies some conditions Meinshausen and Bühlmann (2006).

On the other hand, Fan and Li (2001) conjectured that, the oracle properties do not hold for $LASSO$. Zou (2006) derive the necessary condition of the $LASSO$ variable selection, as follows:

Consider the linear regression model in Eqn. (4.21). Let $\mathbf{x}_{i(1)}$ and $\mathbf{x}_{i(2)}$ the first q and last $p-q$ columns of $\mathbf{x}_i$ respectively, and let $\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i$ where $\mathbf{C}$ is a positive definite matrix. By setting $\mathbf{C}_{(11)} = \frac{1}{n}\mathbf{x}_{n(1)}^T \mathbf{x}_{n(1)i}$, $\mathbf{C}_{(22)} = \frac{1}{n}\mathbf{x}_{n(2)}^T \mathbf{x}_{n(2)}$, $\mathbf{C}_{(12)} = \frac{1}{n}\mathbf{x}_{n(1)}^T \mathbf{x}_{n(2)}$, and $\mathbf{C}_{(21)} = \frac{1}{n}\mathbf{x}_{n(2)}^T \mathbf{x}_{n(1)}$, the matrix $\mathbf{C}$ can then expressed in a block-wise form as follows:

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{(11)} & \mathbf{C}_{(12)} \\ \mathbf{C}_{(21)} & \mathbf{C}_{(22)} \end{pmatrix},$$

where $\mathbf{C}_{(11)}$ is $q \times q$ matrix and $\mathbf{C}_{(22)}$ is $(p-q) \times (p-q)$ matrix. The $LASSO$ estimates are considerd, $\hat{\boldsymbol{\beta}}^{(n)}$,

$$\hat{\boldsymbol{\beta}}_{LASSO}^{(n)} = \arg\min \left[\sum_{i=1}^{n} (y_i - \mathbf{X}^T \boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p} \mid \beta_j \mid \right], \tag{3.2}$$

where λ_n varies with n. Let $\mathbf{A}_n = \left\{ j : \hat{\beta}_j^{(n)} \neq 0 \right\}$.

Definition 3.1. (Zou, 2006)

The $LASSO$ variable selection is consistent if and only if

$$\lim_n P\left(\mathbf{A}_n = \mathbf{A}\right) = 1,$$

where $A = \{j : \beta_j \neq 0\}$. The definition of the consistency is subject to the condition given in Eqn. (3.3). To define this condition the following theorem are needed:

Theorem 3.2. *(Zou, 2006)*

Suppose that

$$\lim_{n} P(\mathbf{A}_n = \mathbf{A}) = 1.$$

Then there exists some vector $\boldsymbol{s} = (s_1, ..., s_q)^T, s_j = 1$ *or* -1*, such that*

$$\mid \boldsymbol{C}_{(21)}\, \boldsymbol{C}_{(11)}^{-1}\boldsymbol{s} \mid \leq 1. \tag{3.3}$$

This theorem presents a necessary condition for consistency of the $LASSO$ variable selection. Proof of this theorem is based on Zou (2006). Note that when $p = 2$ the necessary condition Eqn. (3.3) is always satisfied, because $\mid \mathbf{C}_{(21)}\mathbf{C}_{(11)}^{-1} sgn(\boldsymbol{\beta}_{\mathbf{A}}) \mid$ reduces to $\mid \rho \mid$, the correlation between two predictors.

Lemma 3.3. (Zou, 2006)

If $\frac{\lambda_n}{\sqrt{n}} \to \lambda_0 \geq 0$, then $\sqrt{n}(\hat{\beta}_j - \beta_j) \to_d \arg\min(V_2)$, where

$$V_2(u) = u'\mathbf{C}u - 2u'\mathbf{W} + \lambda_0 \sum_{j=1}^{p} \left[u_j sgn(\beta_j) I(\beta_j \neq 0) + \mid u_j \mid I(\beta_j = 0)\right],$$

and, $\mathbf{W} \sim N\left(0, \sigma^2\mathbf{C}^{-1}\right)$.

Proof of this lemma is given in Knight and Fu (2000).

Lemma 3.3 shows that the $LASSO$ estimate is root-n consistent. However, based on the asymptotic behaviour of variable selection, Lemma 3.3 actually implies that $\mathbf{A}_n$ basically cannot be $\mathbf{A}$ with a positive probability, when $\lambda_n = O(\sqrt{n})$. The inconsistency of $LASSO$ estimation in the general case is proved through the following proposition.

Proposition 3.4. (Zou, 2006)

If $\frac{\lambda_n}{\sqrt{n}} \to \lambda_0 \geq 0$, then

$$\limsup_{n} P\left(\mathbf{A}_n = \mathbf{A}\right) < 1.$$

The proof of this proposition is on Appendix 4.

It is concluded that if the condition in Eqn. (3.3) fails, the $LASSO$ variable selection is inconsistent. However, the asymptotic setup is somewhat unfair, because it forces the coefficients to be equally penalized in the L_1 penalty. So conclude that the $LASSO$ cannot be an oracle procedure.

Note that when $p = 2$ the necessary condition in Eqn. (3.3) is always satisfied, because

$$\mid \mathbf{C}_{21}\mathbf{C}_{11}^{-1} sgn(\boldsymbol{\beta}_{\mathbf{A}}) \mid$$

reduces to $\mid \rho \mid$, the correlation between two predictors (Zou, 2006).

Theorem 3.5. ***(Oracle Properties** (Zou, 2006))*

Suppose that $\lambda_n/n \to 0$ *and* $\lambda_n n^{(\gamma-1))/2} \to \infty$. *Then the* $ada\text{-}LASSO$ *estimates must satisfy the following:*

1. *Consistency in variable selection:*

$$\lim_n P(\mathbf{A}_n = \mathbf{A}) = 1.$$

2. *Asymptotic normality:* $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{\mathbf{A}}^{(n)} - \boldsymbol{\beta}_{\mathbf{A}}\right) \to_d N(0, \sigma^2 \times \boldsymbol{C}_{11}^{-1})$.

The proof of this theorem is in Appendix 5. Thus, the $ada\text{-}LASSO$ procedure will be more efficient than the traditional $LASSO$ procedure to estimate the parameters and to select the significant variables.

3.5 Simulation Studies

3.5.1 Simulation Studies: Example 1

This section presents simulation examples as generated by Zhao and Yu (2006), to compare ridge, $LASSO$, and $ada\text{-}LASSO$ regularization methods in the simple case $p = 3$, $q = 1$. The aim is to show some practical sense of the $LASSO$ and $ada\text{-}LASSO$ algorithm behaviors when condition in Eqn. (3.3) holds and fails. First, the response y were generated by

$$y_i = \mathbf{x}_{i1}\beta_1 + \mathbf{x}_{i2}\beta_2 + \mathbf{x}_{i3}\beta_3 + \varepsilon_i,$$

where the true regression coefficients are in two settings:

(a) $\boldsymbol{\beta} = \{2, 3, 0\}$

(b) $\boldsymbol{\beta} = \{-2, 3, 0\}$.

In both settings the $\mathbf{x}_1$ and $\mathbf{x}_2$ are i.i.d with mean 0 and variance 1 for $i = 1, ..., 100$. The third predictor $\mathbf{x}_3$ is correlated with $\mathbf{x}_1$ and $\mathbf{x}_2$ by

$$\mathbf{x}_3 = \frac{2}{3}\mathbf{x}_1 + \frac{2}{3}\mathbf{x}_2 + \frac{1}{3}\epsilon,$$

where $\epsilon \sim N(0, 1)$. Here $\mathbf{x}_{n1} = (\mathbf{x}_1, \mathbf{x}_2)$, $\mathbf{x}_{n2} = \mathbf{x}_3$, and

$$\mathbf{C}_{11} = \begin{pmatrix} 1 & -0.05 \\ -0.05 & 1 \end{pmatrix},$$

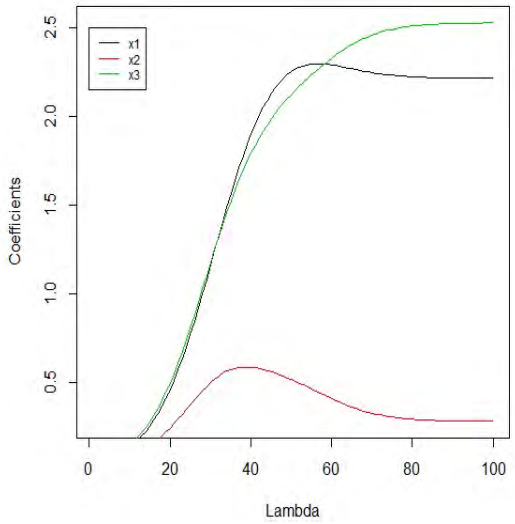
and

$$\mathbf{C}_{21} = \begin{pmatrix} 0.65 & 0.65 \end{pmatrix},$$

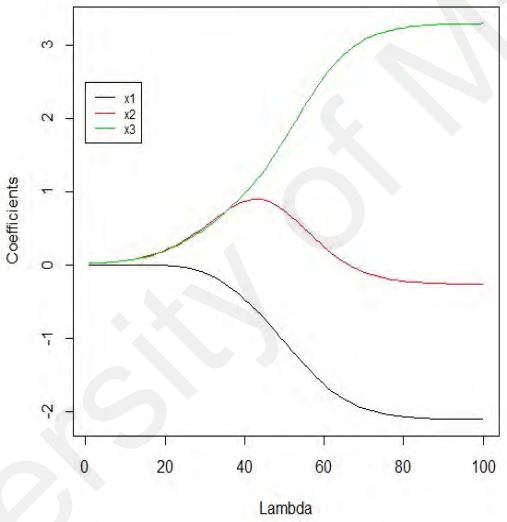
obtain

$$\mathbf{C}_{21}\mathbf{C}_{11}^{-1} = (\ 2/3 \quad 2/3\ ).$$

However, the necessary condition in Eqn.(3.3) fails for setting (a) and holds for setting (b). $Ridge$, $LASSO$ and $ada\text{-}LASSO$ were applied by moving the parameter λ from $\lambda = 0$ to $\lambda = 100$. Figure (3.21) shows the ridge path solution for both setting (a) and (b), which indicates that ridge does not set any coefficients to 0. Different $LASSO$ solutions are obtained which form the $LASSO$ path as illustrated in Figure (3.22). Note that in setting (a) $LASSO$ did not shrink β_3 to 0. For setting (b), with a proper amount of regularization, $LASSO$ correctly shrinks β_3 to 0. Figure (3.23) shows that, the adaptive situation path is consistent with variable selection, and the regularization seemed to prefers $\mathbf{x}_1$ and $\mathbf{x}_2$ and ignored $\mathbf{x}_3$.

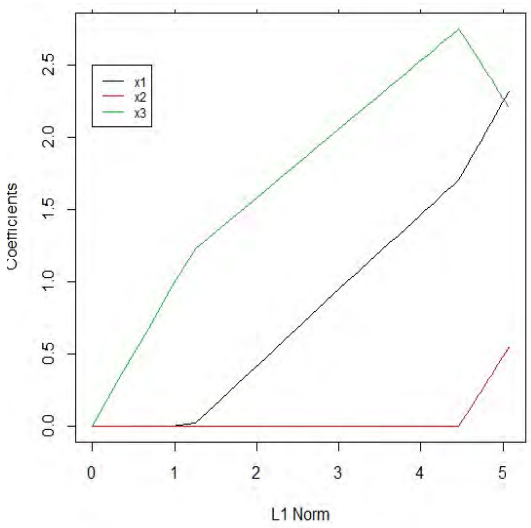


(a) The ridge path for setting(a)

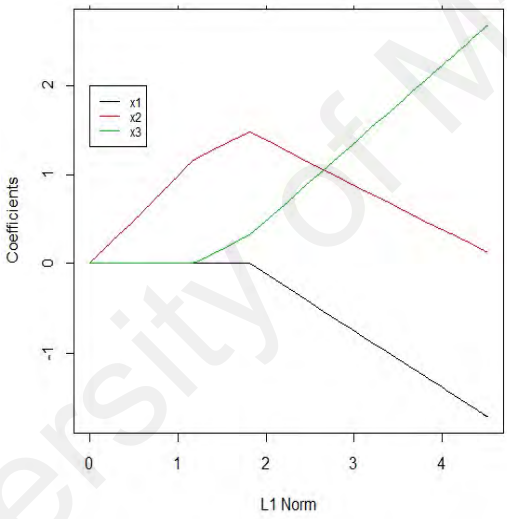


(b) The ridge path for setting(b)

Figure 3.21: The ridge path

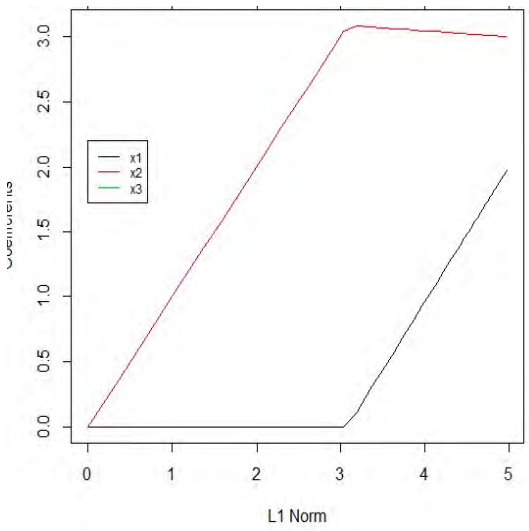


(a) The $LASSO$ path for setting(a)

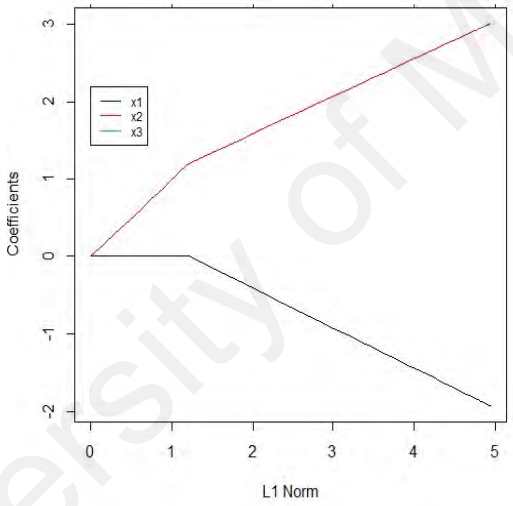


(b) The $LASSO$ path for setting(b)

Figure 3.22: The $LASSO$ path



(a) The $ada\text{-}LASSO$ path for setting(a)



(b) The $ada\text{-}LASSO$ path for setting(b)

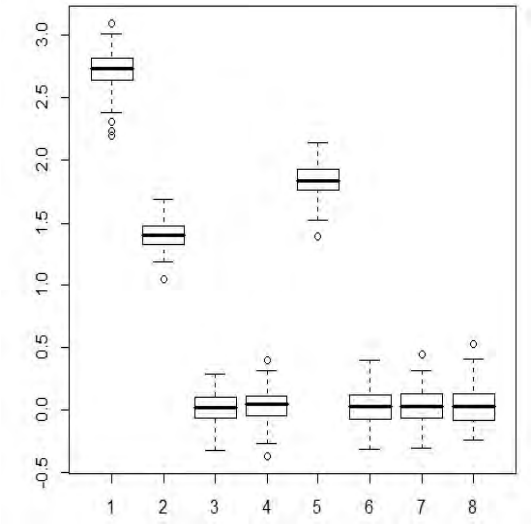
Figure 3.23: The $ada\text{-}LASSO$ path

3.5.2 Simulation Studies: Example 2

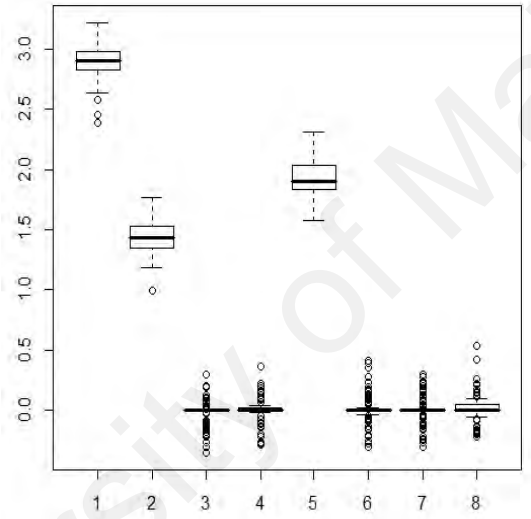
This section compares the performance of selection ability of the ridge, $LASSO$ and $ada\text{-}LASSO$. Here, the penalty parameters (λ) for all versions are chosen using five fold cross-validation of Breiman and Spector (1992) who proposed using fivefold or ten fold in practice. The adaptive weight $\hat{w} = 1/ \mid \hat{\boldsymbol{\beta}}_{LS} \mid^{\gamma}$ with $\gamma = 1$.

In this simulation, the linear regression model in Eqn. (1.1) are considered, where the true regression coefficients are $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)$, and $\varepsilon_i \sim N(0, \sigma^2)$. $\mathbf{x}_i, (i = 1, ..., n)$ are Gaussian vector $N_8(0.5, \Sigma_r)$, where, Σ_r is the covariance matrix with different level of correlation, $(r = 0.3, 0.5,$ and $0.95)$. 200 data were simulated for three different combinations of size $(n = 60, 100, 300)$. On each data set, boxplots of eight coefficients are provided.

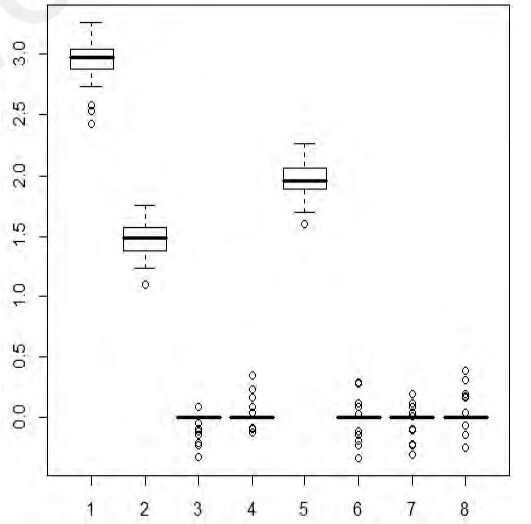
However, as shown in Figures (3.24) to (3.26), n and r increased, the estimation of $ada\text{-}LASSO$ parameter seemed to perform better and produces sparse solution more effectively than $LASSO$. $LASSO$ selected nearly the correct number of zero coefficients, but, suffers from variability as shown in the boxplots. Thus, the $ada\text{-}LASSO$ method more efficient than traditional $LASSO$ method to estimate the parameters and select the significant variables. In contrast, the ridge regression, performed poorly in selection variables.



(a) Ridge

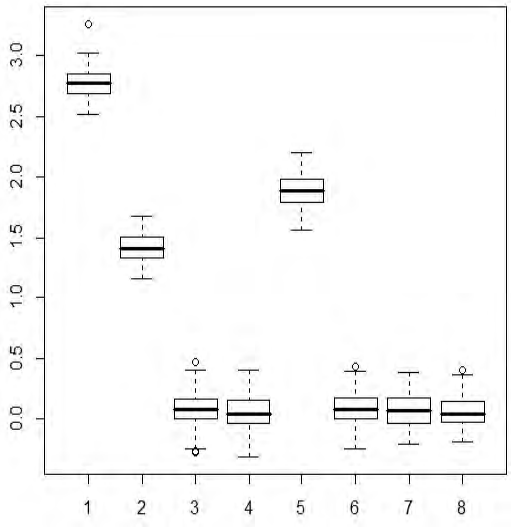


(b) $LASSO$

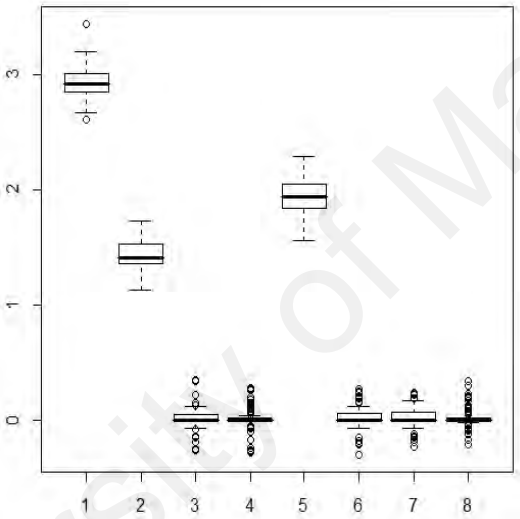


(c) ada-$LASSO$

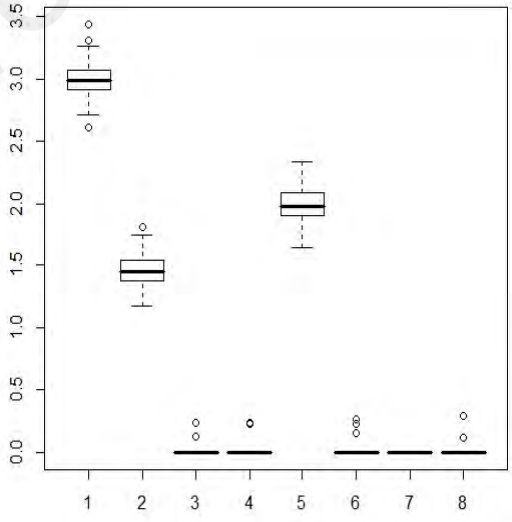
Figure 3.24: The ridge, $LASSO$, and ada-$LASSO$ estimates for eight coefficients via 200 simulation with $r = 0.1$, $n = 60$



(a) Ridge

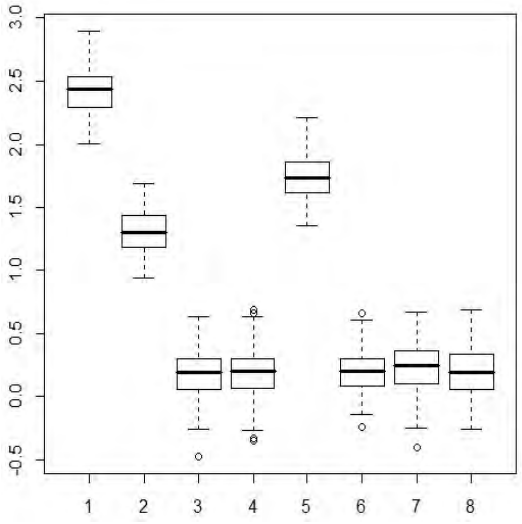


(b) $LASSO$

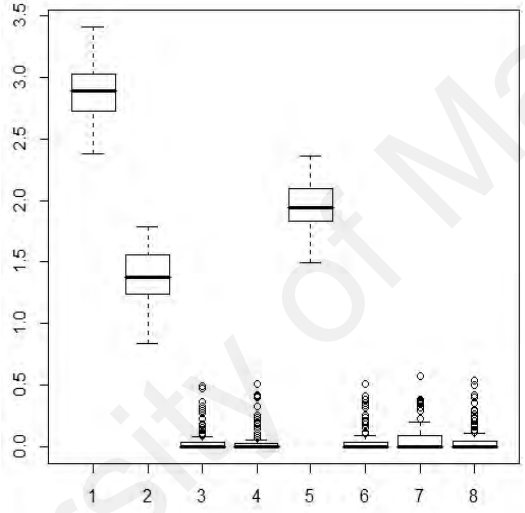


(c) $ada\text{-}LASSO$

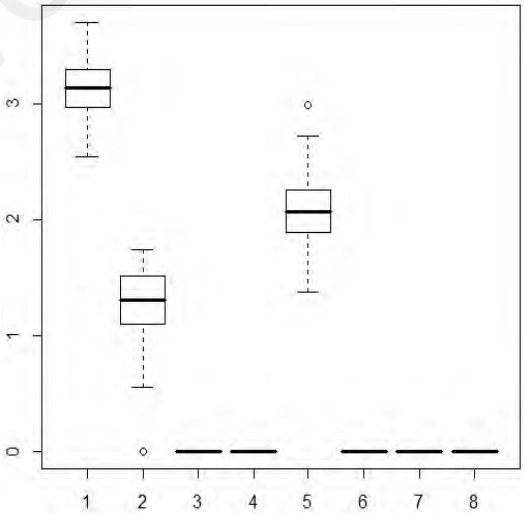
Figure 3.25: The ridge, $LASSO$, and $ada\text{-}LASSO$ estimates for eight coefficients via 200 simulation with $r = 0.5$, $n = 100$



(a) Ridge



(b) $LASSO$



(c) $ada\text{-}LASSO$

Figure 3.26: The ridge, $LASSO$, and $ada\text{-}LASSO$ estimates for eight coefficients via 200 simulation with $r = 0.95$, $n = 300$

3.6 Practical Example (Ozone Data)

In this stage, the regularization path and selection ability of ridge, $LASSO$ and ada-$LASSO$ are compared on Ozone data set which is available in R with a package **cosso**. The Ozone data initially was used in Breiman and Friedman (1985)'s study. This data set contains 330 observations. Each observation is a daily measurement, and 8 variables.

Table 3.9 shows an overview of the variables included in the data. The output variable is Ozone reading. The interested here is to compared the trace of $\boldsymbol{\beta}(\lambda)$ for ridge, $LASSO$ and ada-$LASSO$ criteria on all 8 variables. In each path the coefficient vectors start at $\lambda = 100$ and $\boldsymbol{\beta}$ grows when λ goes from $100 \rightarrow 0$.

Figure (3.27) shows the path of three criteria result. For ridge regression, the coefficients jump away from zero and there is no sparse solution. $LASSO$ behaved differently and gave the sparse solution (zero solution) when the coefficients are small and behaves like ridge once the coefficients are large. Comparing the ada-$LASSO$ regression, found the path more stable than $LASSO$. with some differences between the coefficients. The most obvious ones are "invTemp" which get zero in the ada-$LASSO$ and non-zero in $LASSO$ path. It may be because the ada-$LASSO$ method dampens down the large correlation between variables.

Table 3.10 demonstrates the correlation results among $(i, j)^{th}$ coefficients among which the largest correlation between is "invTemp" , "milpress", and "temp".

Figure (3.28) shows boxplots of 100 λ values of different criterion estimates. For ridge, there are no predictors to give estimated coefficients 0. As a result, ada-$LASSO$

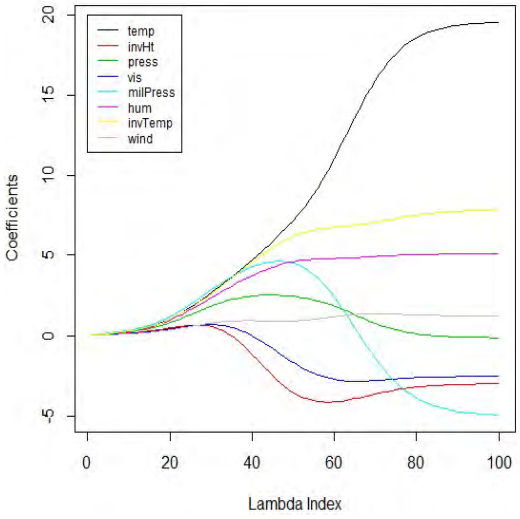
was more stable than $LASSO$.

Table 3.9: Variables of the Ozone data

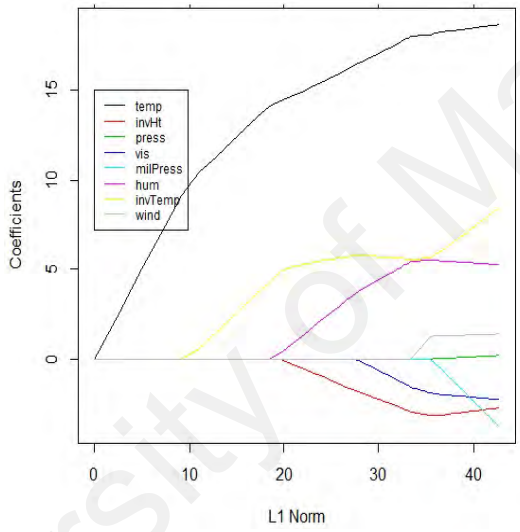
Name	Description
temp	Temperature (degree C). Minimum 25 and maximum 93 in original scale.
invHt	Inversion base height (feet). Minimum 111 and maximum 5000 in original scale.
press	Pressure gradient (mm Hg). Minimum -69 and maximum 107 in original scale.
vis	Visibility (miles). Minimum 0 and maximum 350 in the original scale.
milPress	500 millibar pressure height (m). Minimum 5320 and maximum 5950 in original scale.
hum	Humidity (percent). Minimum 19 and maximum 93.
invTemp	Inversion base temperature (degrees F). Minimum -25 and maximum 332 in original scale.
wind	Wind speed (mph). Minimum 0 and maximum 21 in original scale.

Table 3.10: The correlation results among the $(i, j)^{th}$ of the Ozone data

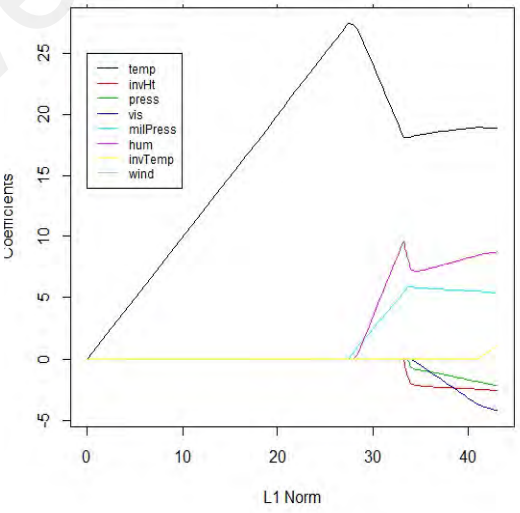
(i, j)	**temp**	**invHt**	**press**	**vis**	**milPress**	**hum**	**invTemp**	**wind**
temp	1							
invHt	-0.5326	1						
press	0.1892	0.0370	1					
vis	-0.3877	0.3866	-0.1258	1				
milPress	0.8080	-0.5048	-0.1480	-0.3600	1			
hum	0.3404	-0.2423	0.6477	-0.4010	0.0744	1		
invTemp	0.8647	-0.7769	-0.0950	-0.4223	0.8520	0.2036	1	
wind	-0.0320	0.2065	0.3357	0.14722	-0.24366	0.2102	-0.1795	1



(a) Ridge

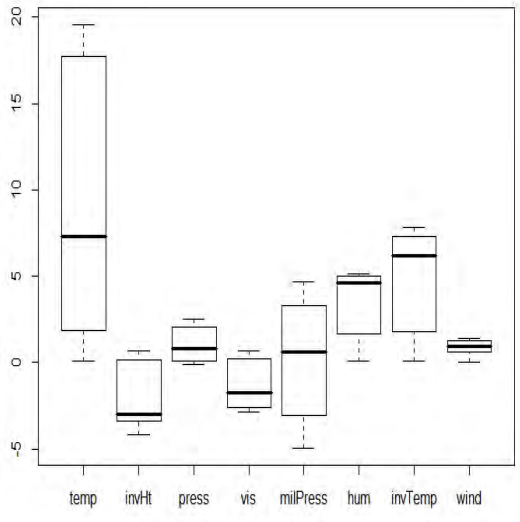


(b) $LASSO$

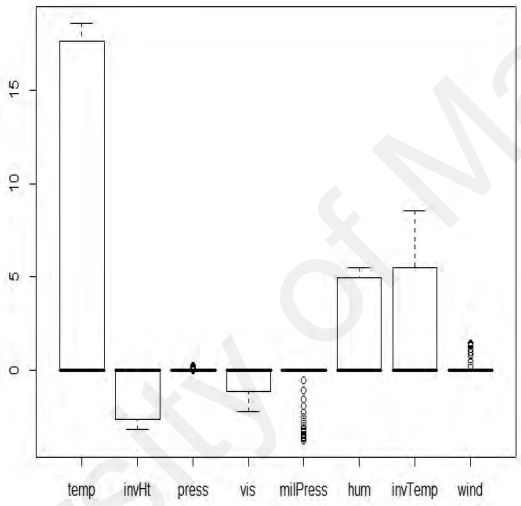


(c) $ada\text{-}LASSO$

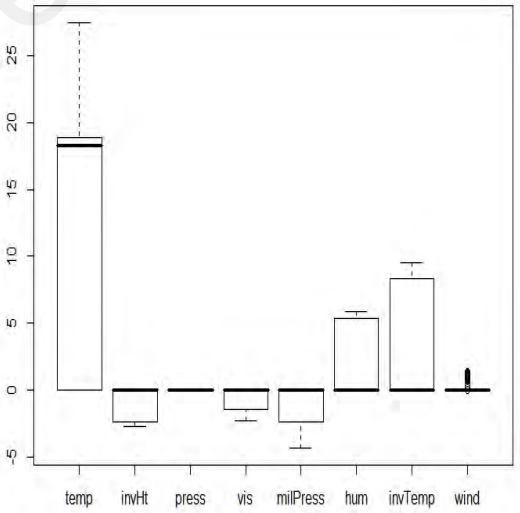
Figure 3.27: The ridge, $LASSO$, and $ada\text{-}LASSO$ estimate for Ozone data



(a) Ridge



(b) $LASSO$



(c) $ada\text{-}LASSO$

Figure 3.28: Boxplots of 100 λ values of the ridge, $LASSO$, and $ada\text{-}LASSO$ coefficients estimates for the eight predictors in the Ozone data

3.7 Effect of Leverage Points on Robust $LASSO$ Regression Methods (LAD-$LASSO$ and Huber-$LASSO$)

In statistical analysis, the existence of high leverage points in the data set should raise some concern. The effect of leverage points on the variable selection method is known to be severe. Here, it is useful to investigate the effect of leverage point in existing robust methods (LAD-$LASSO$ and Huber-$LASSO$) by introducing leverage points in the data set.

3.7.1 Simulation Procedure

A simulation study was carried out to investigate the effect of leverage points on the robust variable selection (LAD-$LASSO$ and Huber-$LASSO$) of regression models. For simplicity, the case when $\lambda = log(n)$ were considered. The statistical software applied in this stage was CVX that is a package for specifying and solving convex programs in Matlab (Grant et al., 2008). The following set of parameters were estimated: $\beta_1 = 0.5$, $\beta_2 = 1$, $\beta_3 = 1.5$, $\beta_4 = 2$ and $\beta_j = 0$ for $5 \leq j \leq p$. The variables $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)$ were generated from n independent Gaussian vectors $N_{20}(0.5, \Sigma_r)$, $r = 0.5$ $p = 20$ and $n = 200$.

The response variable y was generated according to the regression model in Eqn. (1.1), where, ϵ, the error terms follow a standard normal distribution. In order to investigate the robustness of the methods against outliers, three situations were considered:

1. No ncontamination
2. Good leverage points
3. Bad leverage points

such that the percentages of contamination used are $c\% = 10\%, 20\%$, from the sample size n. For case (good leverage points), we considered the different percentages of outliers ($c\% = 10\%, 20\%$) on the variables X_1, X_2 and X_4, that were generated from a $N(100, 0.5^2)$ distribution, then obtained Y from the liner regression model to get the good leverage point. For case (bad leverage points), different percentages of outliers ($c\% = 10\%, 20\%$) on the variables X_1, X_2 and X_4 were generated from a $N(100, 0.5^2)$ distribution. Below are the steps of the simulation:

(i) Generating the design matrix $\mathbf{X} \sim N_{20}(0.5, \Sigma_{0.5})$

(ii) Generating ϵ of size $n = 200$ from $N_{200}(0, 1)$

(iii) The true values of the coefficients were $\beta_1 = 0.5$, $\beta_2 = 1$, $\beta_3 = 1.5$, $\beta_4 = 2$ and $\beta_j = 0$ for $5 \leq j \leq p$. Then y using Eqn. (1.1) were obtained.

(iv) Obtaining the unpenalized estimator LAD (betaLAD=$\hat{\boldsymbol{\beta}}_{LAD}$) and Huber (betaHU=$\hat{\boldsymbol{\beta}}_{H}$). And the tuning parameter λ was fixed and denoted by lambda.

(v) For uncontaminated model, the coefficients $(\beta_1, ..., \beta_{20})$ were fitted to the regression model using the following CVX program to give the parameter LAD-$LASSO$ and Huber-$LASSO$ estimates $\hat{\boldsymbol{\beta}}_{LAD-LASSO}$ and $\hat{\boldsymbol{\beta}}_{Huber-LASSO}$.

(vi) For $c\%$ contaminated data, the first $c \times n/100$ predictors $\mathbf{X}$ in (iv) were replaced by the newly generated value $\mathbf{X}^\star$. Then, the generated contaminated regression data were fitted to give the parameter estimated $\hat{\boldsymbol{\beta}}^\star_{LAD-LASSO}$ and $\hat{\boldsymbol{\beta}}^\star_{Huber-LASSO}$ using the CVX program .

```
%%   LAD-LASSO
cvx begin
variable beta(p)
minimize norm(y-x*beta,1)+lambada*norm(beta./betaLAD,1);
beta;
cvx end
```

```
%% Huber-LASSO
cvx begin
variables beta(p) s v(n)
minimize (n*s+quad_over_lin(y-x*beta-v,s)+2*k*norm(v,1)+lambada*norm(beta./betaHU1)
subject to
's > 0';
beta;
cvx end
```

(vii) Finally, the steps (i)-(v) above repeated for $simu = 100$ times. For each parameter $(\beta_1, ..., \beta_{20}) = (\hat{\beta}_1, ..., \hat{\beta}_{20})$, the median of Relative Prediction Errors (MRPE) were calculated using the following formula:

$$MRPE = median(E(y - \mathbf{X}\hat{\boldsymbol{\beta}}_m)^2), \tag{3.4}$$

for $m = 1, 2, ..., 100$ simulation, to evaluate the selection ability, we plot the box-plots of parameters estimation.

Discussion

The results were tabulated in Table 3.11 and Figures (3.29) to (3.35) for each situation. Several observed results are as follows:

- For leverage point-free data set, the relative prediction errors and the model selection ability of LAD-$LASSO$ and Huber-$LASSO$ close to the ones of true model.
- When the data were contaminated with bad leverage point, the value for relative prediction errors of each methods increased when the percentages of contamination

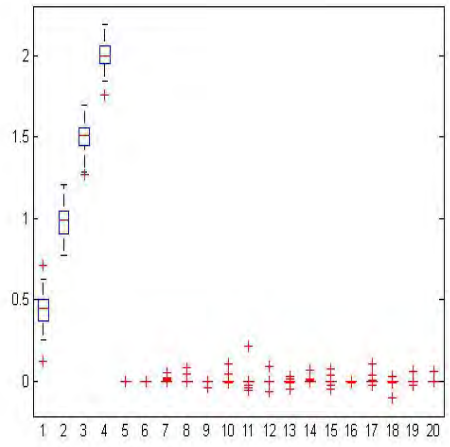
increased. The bias and standard deviation was generally larger than the uncontaminated data set. In this case, both methods led to a poor estimation of coefficients (see Figures (3.30) to (3.32)) and selected overfit models.

- In case of good leverage points, relative prediction errors for both methods generally got larger as the percentages of contamination increased, but still better than the values with bad leverage point. LAD-$LASSO$ and Huber-$LASSO$ have better model selection ability than the previous situation (see Figures (3.33) to (3.35)).

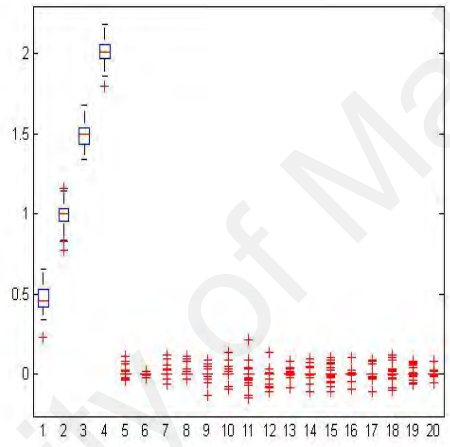
By looking the above mentioned results for uncontaminated data, robust $LASSO$ methods performed well in the selection of the parameters of the regression model. However, the methods were affected by the presence of leverage points in the data. The effect was worse with higher percentage of bad leverage points in the data.

Table 3.11: Simulation result, the MRPE based on 100 replications

Methods	LAD-$LASSO$	Huber-$LASSO$
Uncontaminated	0.0300	0.0258
5% bad leverage	3.0857	6.3125
10% bad leverage	58.7081	58.0209
20% bad leverage	122.8706	121.7094
5% good leverage	0.0305	0.0321
10% good leverage	0.0330	0.0243
20% good leverage	0.0311	0.0272

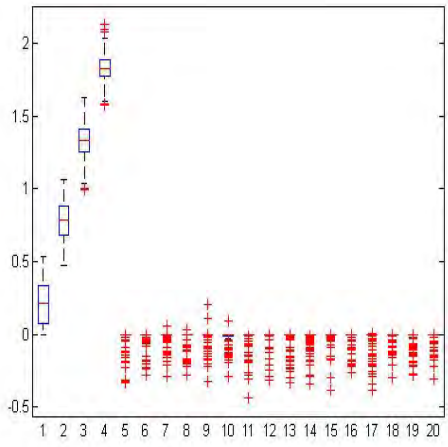


(a) $LAD\text{-}LASSO$

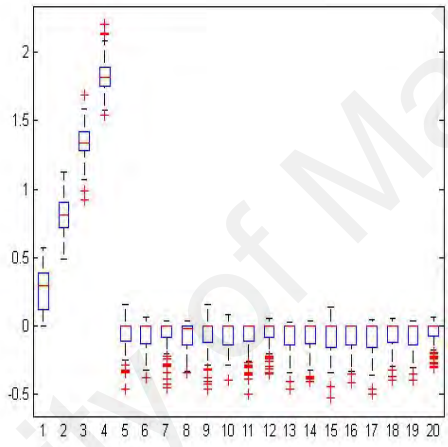


(b) Huber-$LASSO$

Figure 3.29: Boxplots of estimates for the 20 coefficients from 100 simulated data sets, uncontaminated data

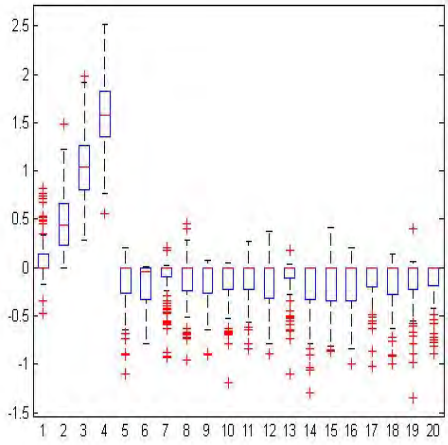


(a) $LAD\text{-}LASSO$

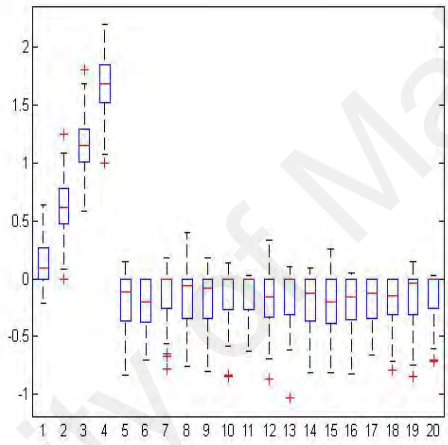


(b) Huber-$LASSO$

Figure 3.30: Boxplots of estimates for the 20 coefficients from 100 simulated data sets, 5% bad leverage

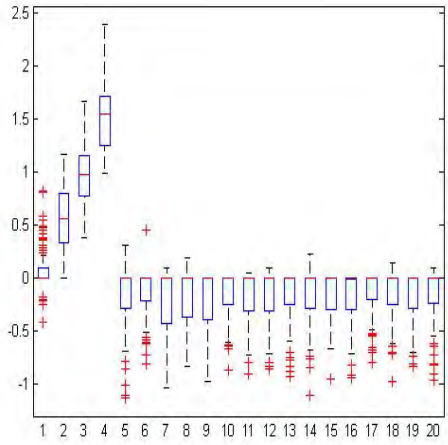


(a) $LAD\text{-}LASSO$

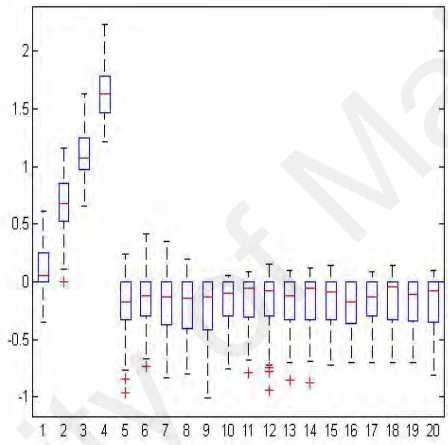


(b) Huber-$LASSO$

Figure 3.31: Boxplots of estimates for the 20 coefficients from 100 simulated data sets, 10% bad leverage

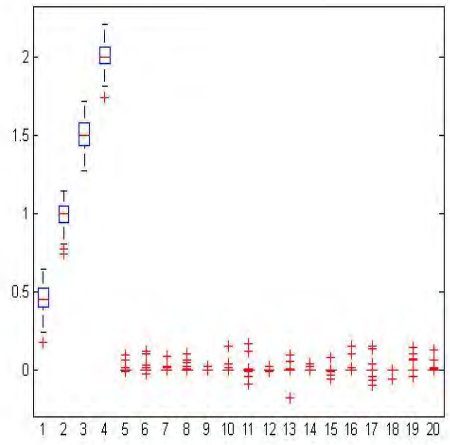


(a) $LAD\text{-}LASSO$

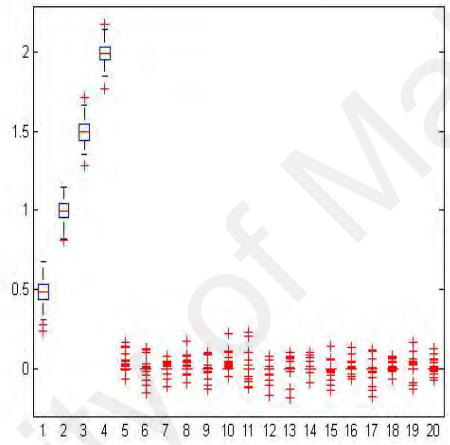


(b) Huber-$LASSO$

Figure 3.32: Boxplots of estimates for the 20 coefficients from 100 simulated data sets, 20% bad leverage

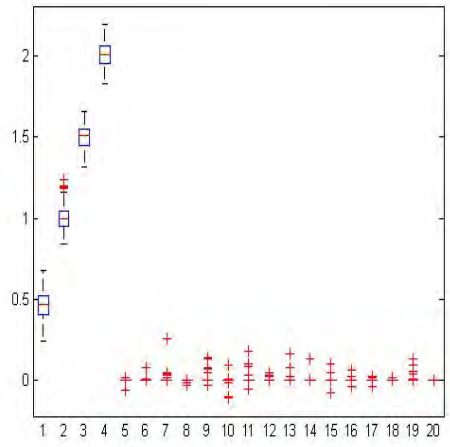


(a) $LAD\text{-}LASSO$

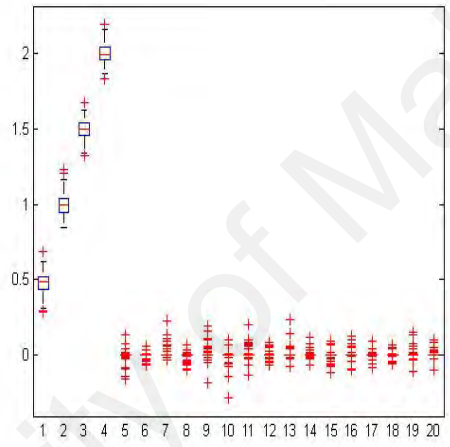


(b) Huber-$LASSO$

Figure 3.33: Boxplots of estimates for the 20 coefficients from 100 simulated data sets, 5% good leverage

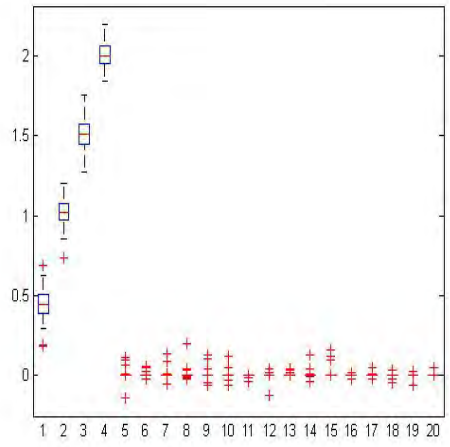


(a) $LAD\text{-}LASSO$

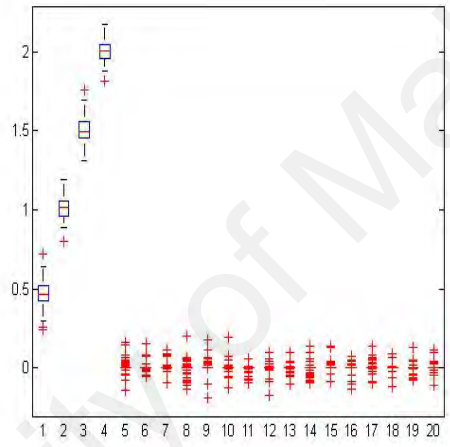


(b) Huber-$LASSO$

Figure 3.34: Boxplots of estimates for the 20 coefficients from 100 simulated data sets, 10% good leverage



(a) $LAD\text{-}LASSO$



(b) Huber-$LASSO$

Figure 3.35: Boxplots of estimates for the 20 coefficients from 100 simulated data sets, 10% good leverage

3.8 Summary

Throughout the work in this chapter, the experiment study has been carried out to see the effect of verticals and leverage points on the variable selection methods. We illustrate the sensitivity of classical method to vertical and also identified certain limitation of robust variable selection methods based on M-estimators. However, when leverage point are presence in the data set, the robust variable selection based on M-estimator tends to

select wrong fit model.

On the other hands, robust $LASSO$ methods based on LAD estimator and Huber function are performed well in the selection of the parameters of the regression model. Also, the methods were affected by the presence of leverage points in the data. The effect was worse with higher percentage of bad leverage points in the data.

It can be concluded that the presence of outliers has an impact the classical variable selection. Whereas, the robust variable selection methods based on M-estimation have been affected by the presence of leverage points in data.

CHAPTER 4

VARIABLE SELECTION BASED ON HIGH BREAKDOWN SCALE ESTIMATOR

4.1 Introduction

Chapter two reviews a variety of robust variable selection criteria in regression model which use M-estimator of coefficients in the procedures, namely the $RAIC$ (Ronchetti, 1985), R^2_M (Anderson-Sprecher, 1994), RCp (Ronchetti and Staudte, 1994), and $RSIC$ (Machado, 1993) statistics. M-estimators are efficient and highly robust to unusual values of y, but leverage point can break them down completely. This is illustrated in Figures (3.8) to (3.11) that the value of $RAIC$, R^2_M, RCp, and $RSIC$ increase drastically even if a one single leverage point is added to the data.

This chapter discusses the possibility of extending the idea of using robust estimators in the model selection criteria by using high breakdown point estimators in the procedure. It is known that almost all different model selection methods are expressed in terms of the variance, which are computed in LS or M-estimation methods. In this study, instead of working with the classical scale or M-estimators scale, high breakdown point estimators for variable selection was used. This in turn reduces the effect of outliers and leverage point. Subsequently modified robust model selection statistics are defined that is based on high breakdown point estimators. The proposed methods were evaluated using influence function. Simulation and real data were applied to investigate the performance power of

the modified robust selection statistic and compared them with the classical methods as well as the existing robust M-estimation methods.

4.2 Existing Approaches of Variable Selection

Consider the full model which is a standard linear regression model defined in Eqn. (4.21). For each set $K \subset \{1, ..., p\}$, the most common methods used to selection variables in standard linear regression model are the Akaike Information Criteria (AIC), Mallow's (Cp), and Schwarz Information Criterion (SIC). These criteria are given as:

$$AIC = \log\left(\frac{SSE_K}{n}\right) + 2k, \tag{4.1}$$

$$C_p = \frac{(n-k)SSE_K}{SSE} - n + 2k, \tag{4.2}$$

$$SIC = \log\left(\frac{SSE_K}{n}\right) + \frac{k\log(n)}{n}, \tag{4.3}$$

where, $k = \#(K)$, $SSE_K = \sum_{i=1}^{n} r_{iK}^2$ and $r_{iK} = y_i - \hat{\mu} - \mathbf{x}_{iK}^T\hat{\boldsymbol{\beta}}_{LS_K}$, the residuals from the LS of reduced model based on the set K. Therefore, $SSE_p = \sum_{i=1}^{n} r_{ip}^2$ and $r_{ip} = y_i - \hat{\mu} - \mathbf{x}_{ip}^T\hat{\boldsymbol{\beta}}_{LS_p}$, the residuals from the LS of full model. The best subset K is chosen as the one minimizing AIC, Cp, and SIC.

It is clear, that a few outliers may harm the values of AIC, Cp, and SIC. It is noteworthy that all these criteria are computed by using the variance of the residuals defined by:

$$Var(r_{ip}) = \frac{SSE_p}{n-p}. \tag{4.4}$$

Thus, the analogous to Eqns. (4.1), (4.2), and (4.3) are

$$AIC = \log\left(\frac{(n-k)Var(r_{ik})}{n}\right) + 2k, \tag{4.5}$$

$$C_p = \frac{(n-k)Var(r_{ik})}{Var(r_i)} - n + 2k, \tag{4.6}$$

$$SIC = \log\left(\frac{(n-k)Var(r_{ik})}{n}\right) + \frac{k\log(n)}{n}. \tag{4.7}$$

So, to robustify these criteria, only the variance was considered to be robust. The robust variance of the residual process in Eqn. (4.4) is commonly computed as,

$$\hat{\sigma}_p = \left[\frac{\sum_{i=1}^{n}\rho_i(\hat{\boldsymbol{\beta}})}{n-p}\right]^{1/2}, \tag{4.8}$$

where, $\rho_i(\hat{\boldsymbol{\beta}})$ is symmetric and non decreasing on $[0, \infty[$. Furthermore, $\rho(0) = 0$ and ρ is almost everywhere continuously differentiable, this the same as in M-estimate procedure. In the presence of leverage points, the M-residuals are affected resulted in a large value of $\hat{\sigma}_p$. As $\hat{\sigma}_p$ may be over estimated in the presence of influence point, an alternative estimate of the variance is needed.

If we take $\rho(u) = u^2$, then $\hat{\sigma}_p$ is a sum of squares and we get the classical variable selection for Eqns. (4.5), (4.6), and (4.7). Instead of a squared loss function one could take $\rho(u) = |u|$, which leads to the minimization of the sum of absolute values of the residuals and yields the L_1 regression estimator. the associated variable selection is given by

$$AIC = \log\left(\frac{(n-k)\sum_{i=1}^{n}|r_{ik}|}{n}\right) + 2k, \tag{4.9}$$

$$C_p = \frac{(n-k)Var(r_{ik})}{\sum_{i=1}^{n}|r_i|} - n + 2k, \tag{4.10}$$

$$SIC = \log\left(\frac{(n-k)\sum_{i=1}^{n}|r_{ik}|}{n}\right) + \frac{k\log(n)}{n}. \tag{4.11}$$

The above measure, and a variant of it, have affected in the presence of leverage point.

A smooth and bounded ρ function is Tukey Biweight given in Eqn. (2.18). The resulting estimator is then an S-estimator (Rousseeuw& Yohai, 1984), which we call the Biweight S-estimator (BS). The constant d in Eqn. (2.18) determines the breakdown point of the estimator, which is the maximal fraction of contamination that an estimator can withstand. The following scale alternative is proposed. In new methods a high breakdown and bounded estimate is used. Therefore, AIC, Cp, and SIC criteria must be robustified along with a high breakdown and bounded influence error scale $\hat{\sigma}_p$.

4.3 A High Breakdown and Bounded Influence Variance (Scale)

Consider a high breakdown regression estimator defined by:

$$(\hat{\mu}, \hat{\boldsymbol{\beta}}) = \arg\min_{(\mu,\boldsymbol{\beta})} S_n\left(r_1(\mu,\boldsymbol{\beta}), ..., r_n(\mu,\boldsymbol{\beta})\right), \tag{4.12}$$

where, S_n is a residual scale estimator verifying $S_n = (ae_1, ..., ae_n) = |a|(e_1, ..., e_n)$ for all $e_1, ..., e_n$, and $a \in \mathbb{R}$.

Now a high breakdown and bounded influence scale is defined by analogy with the classical formula in Eqn. (4.4) as:

$$\hat{\sigma}_{S_p} = \left[\frac{S(r_i)}{n-p}\right]^{1/2}, \tag{4.13}$$

where $\hat{\sigma}_{S_p}$ is the high breakdown scale estimator for sub model. There is another residual

scale $\hat{\sigma}_S$ for every high breakdown regression estimator, where $\hat{\sigma}_S$ is the high breakdown scale estimator for full model. Thus, $\hat{\sigma}_{LMS}$, $\hat{\sigma}_{LTS}$, and $\hat{\sigma}_{BS}$ are defined by fitting LMS, LTS, and Biwieght S-estimation, respectively.

4.4 Different Variable Selection Criteria Based on High Breakdown and Bounded Influence Scale Estimate

The proposed methods called 'High Breakdown and Bounded Influence Variable Selection Criteria' are defined by using high breakdown and bounded influence scale estimator $\hat{\sigma}_S$ as follows:

$$AIC_S = \log\left(\frac{(n-k)\hat{\sigma}^2_{S_k}}{n}\right) + 2k, \tag{4.14}$$

$$C_{p_S} = \frac{(n-k)\hat{\sigma}^2_{S_k}}{\hat{\sigma}^2_S} - n + 2k, \tag{4.15}$$

$$SIC_S = \log\left(\frac{(n-k)\hat{\sigma}^2_{S_k}}{n}\right) + \frac{k\log(n)}{n}. \tag{4.16}$$

There is an additional value of AIC_S, C_{p_S}, and SIC_S for each regression estimate and residual scale estimator S. Small value of AIC_S, C_{p_S}, and SIC_S reveals that the explanatory variables adequately explain the distribution of y. If the estimated scale is from LTS, thus,

$$\hat{\sigma}^2_{LTS} = 1/H \sum_{i=1}^{H} r_i^2(\hat{\boldsymbol{\beta}}_{LTS}), \tag{4.17}$$

where, $H \in 1, ..., n$, and $\left|r_{[1]}\right| \leq \left|r_{[2]}\right| \leq ... \leq \left|r_{[n]}\right|$ denote the ordered absolute residuals. When $H = n/2$ is equivalent to finds the estimates corresponding to the half samples having the smallest sum of squares of residuals. As such, breakdown point is 50%. When $H = [(n+p+1)/2]$ is equivalent to LMS and when $H = n$, LTS and LS coincide: $\hat{\boldsymbol{\beta}}_{(LTS,n,N)} = \hat{\boldsymbol{\beta}}_{(LS,N)}$.

Then, the robust variable selection based on LTS estimators are define as:

$$AIC_{LTS} = \log\left(\frac{(n-k)\hat{\sigma}^2_{LTS_K}}{n}\right) + 2k, \tag{4.18}$$

$$C_{p_{LTS}} = \frac{(n-k)\hat{\sigma}^2_{LTS_K}}{\hat{\sigma}^2_{LTS}} - n + 2k, \tag{4.19}$$

$$SIC_{LTS} = \log\left(\frac{(n-k)\hat{\sigma}^2_{LTS_K}}{n}\right) + \frac{k\log(n)}{n}. \tag{4.20}$$

4.4.1 Experiment 2

In this experiment, experiment 1 (described in Chapter three) was repeated, for the AIC, R^2, Cp and SIC using high breakdown scale estimators including LMS, LTS and BS-estimator.

4.4.2 Discussion

Figures (4.1) to (4.4) showed the results: The LMS, LTS, and BS-estimator show a very robust behavior; there is only a slight loss in criteria, becoming constant when the outlier moves further away from the origin.

Based on the results, it is evident that the variable selection methods based on high breakdown point estimators show robust behavior in the presence of verticals or leverage point.

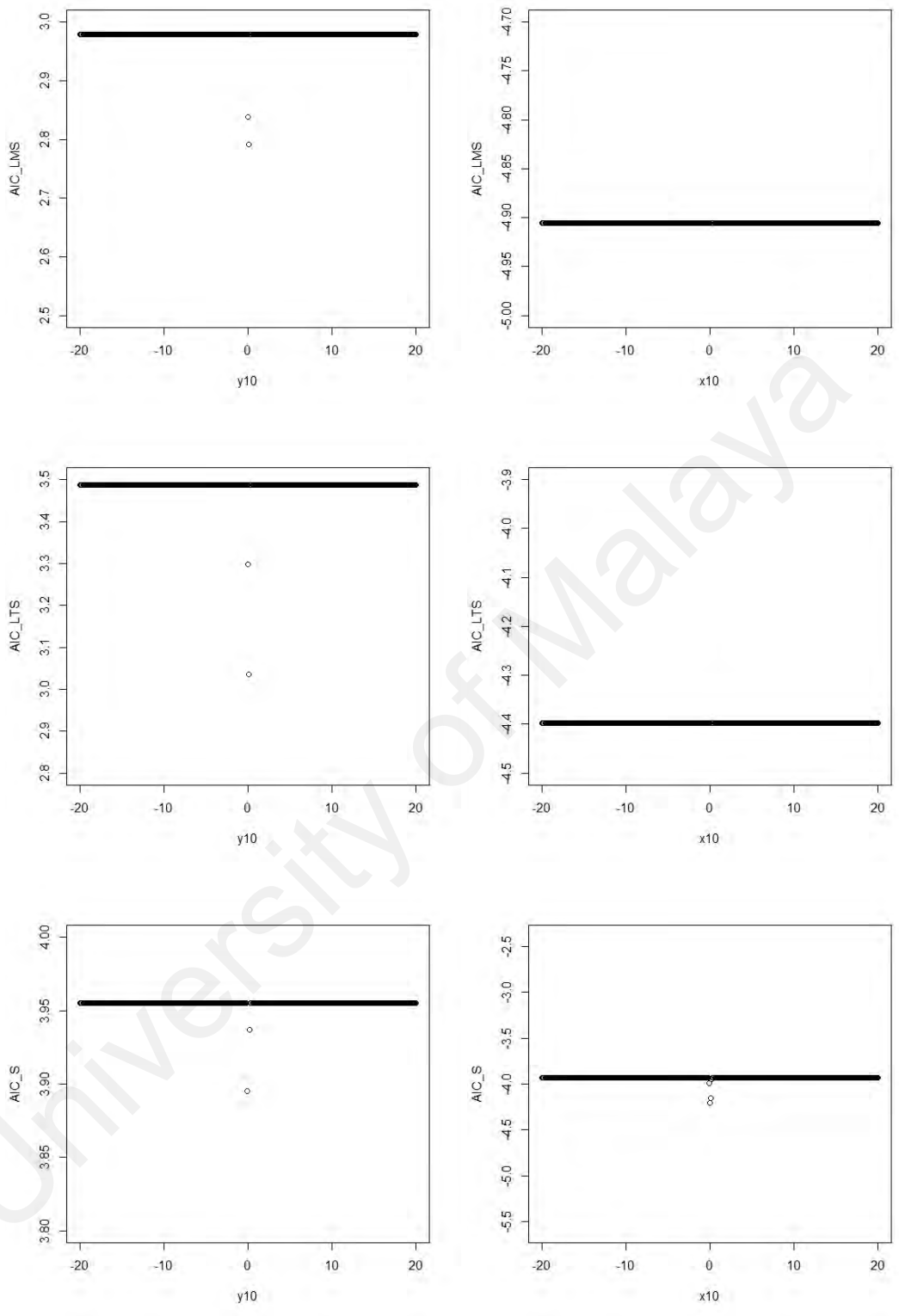


Figure 4.1: Effect of adding on observation (0,y_{10}) on the values of AIC_{LMS}, AIC_{LTS} and AIC_{BS} (left figures) and effect of adding on observation (x_{10},0) (right figures)

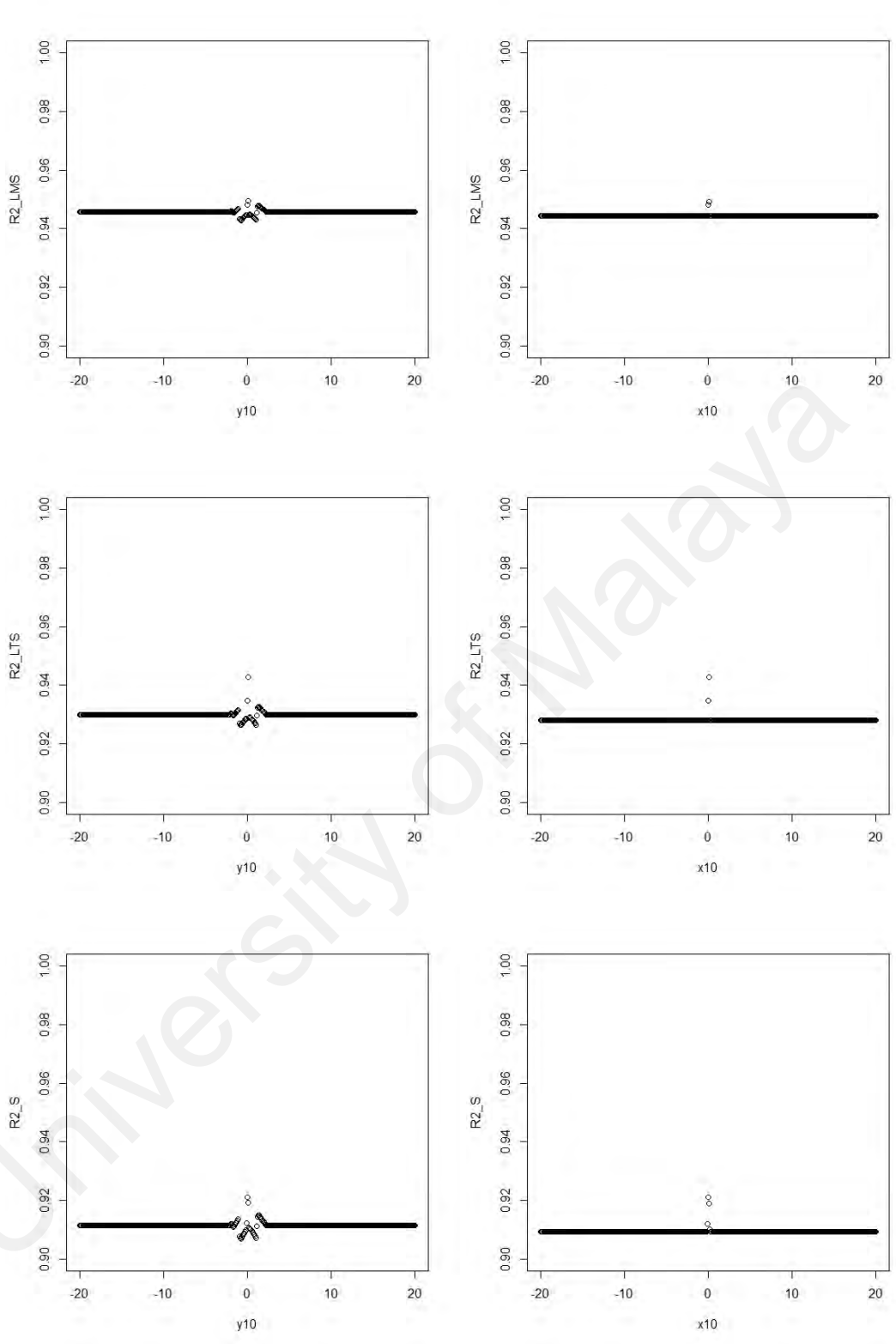


Figure 4.2: Effect of adding on observation (0,y_{10}) on the values of R^2_{LMS}, R^2_{LTS} and R^2_{BS} (left figures) and effect of adding on observation (x_{10},0) (right figures)

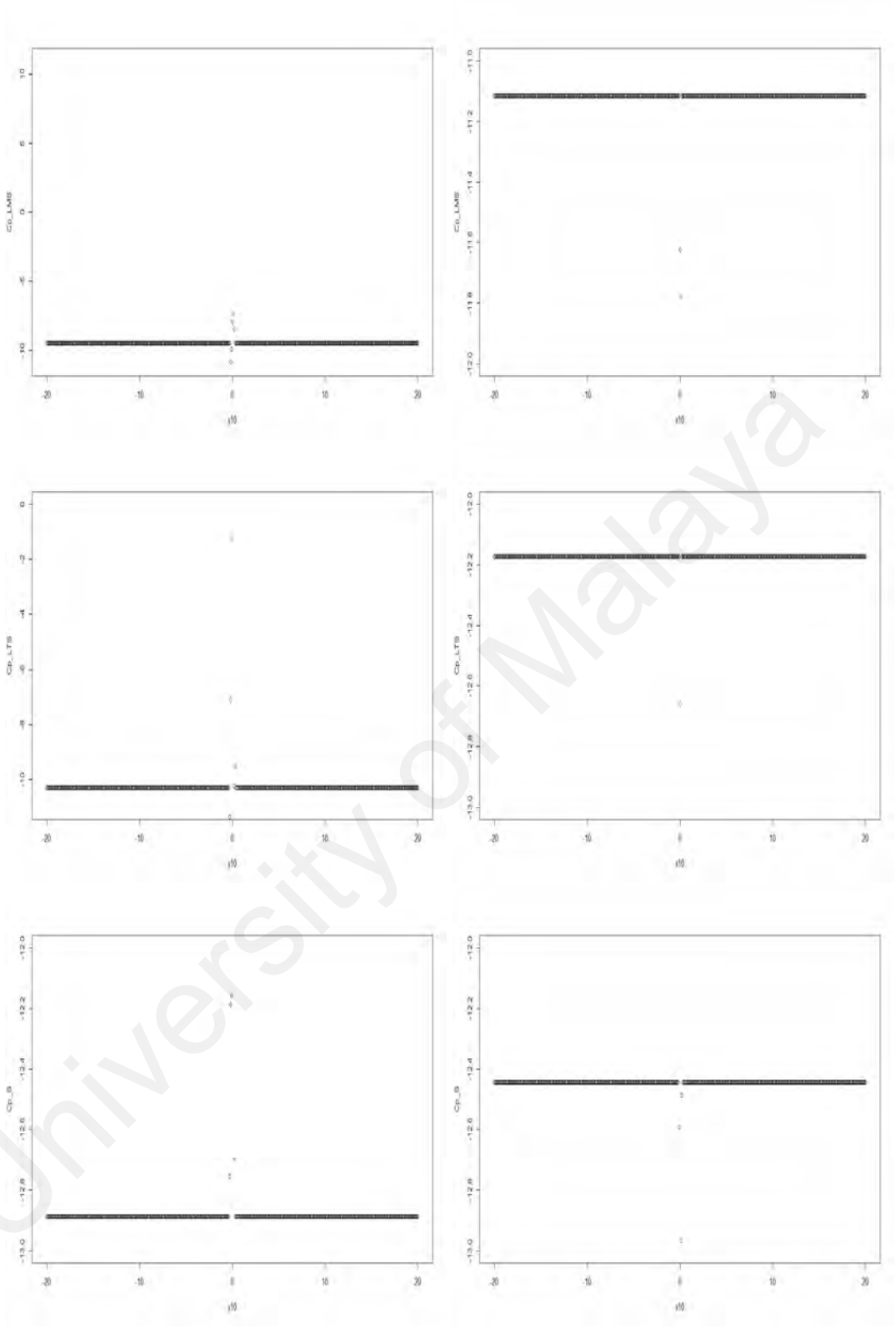


Figure 4.3: Effect of adding on observation (0,y_{10}) on the values of Cp_{LMS}, Cp_{LTS} and Cp_{BS} (left figures) and effect of adding on observation (x_{10},0) (right figures)

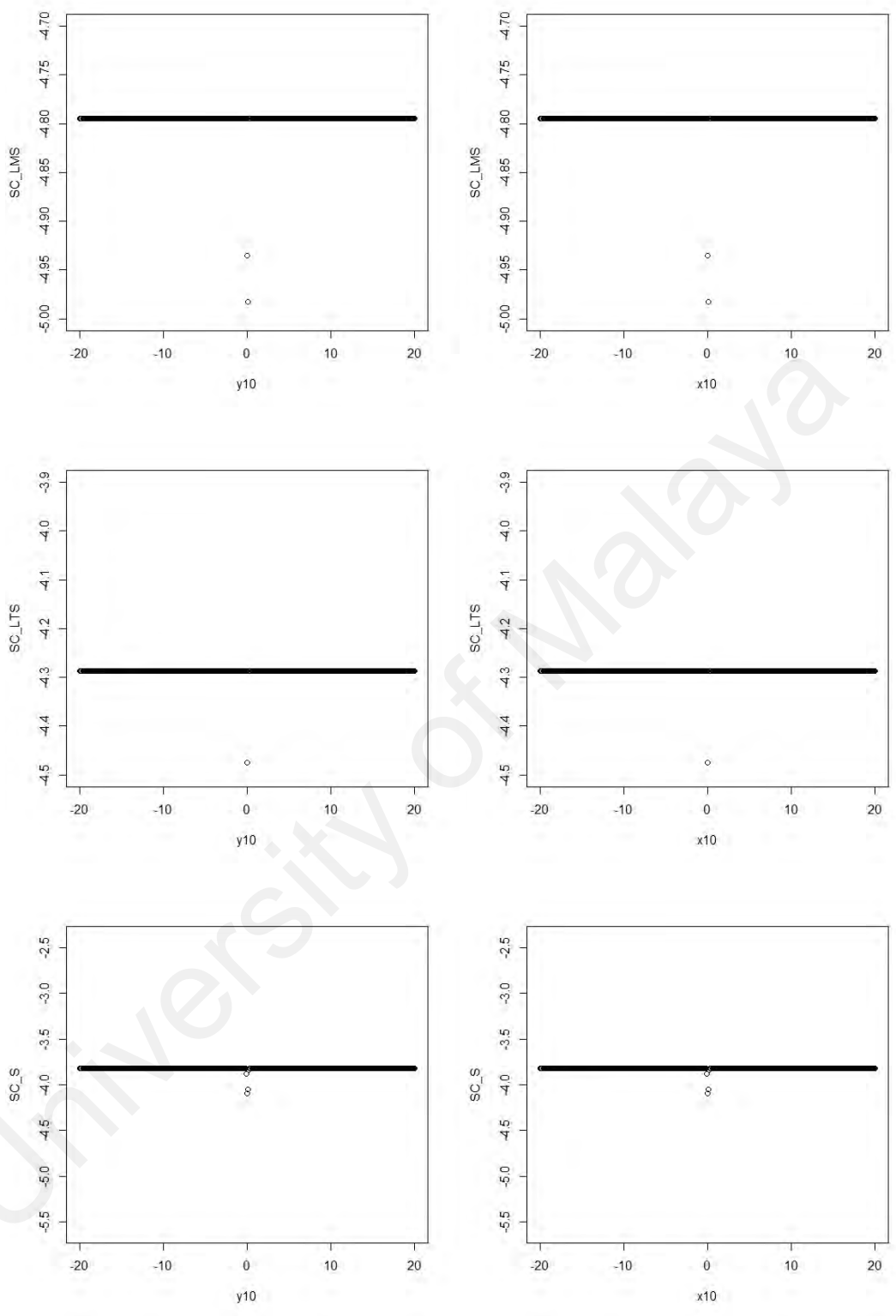


Figure 4.4: Effect of adding on observation (0,y_{10}) on the values of SIC_{LMS}, SIC_{LTS} and SIC_{BS} (left figures) and effect of adding on observation (x_{10},0) (right figures)

4.5 Properties of the Proposed Robust Selection Criteria

In this section, properties of variable selection criteria through its influence function are derived. The gross-error sensitivity is constructed of the proposed variable selection procedures, AIC_S, Cp_S, and SIC_S. Other studies on influence function of R^2 can be found in Croux and Dehon (2003).

4.5.1 Influence Functions of the Proposed Criteria

Consider the standard linear regression model with intercept μ given by:

$$y_i = \mu + \mathbf{X}_i^T \boldsymbol{\beta} + \varepsilon_i. \tag{4.21}$$

Assume that the distribution of errors satisfying $F_\sigma(\mathbf{X}) = F_0(\mathbf{X}/\sigma)$, where σ is the residual scale parameter and F_0 is symmetric with certain probability density function.

Suppose that $\epsilon_i \sim F_\sigma(\mathbf{X})$, let $\mathbf{X}$ and y be independent stochastic variables with distribution H. The functional T is Fisher-consistent for the parameters $(\mu, \boldsymbol{\beta})$ at the model distribution H as follows:

$$T(H) = \begin{bmatrix} a(H) \\ \boldsymbol{b}(H) \end{bmatrix} = \begin{bmatrix} \mu \\ \boldsymbol{\beta} \end{bmatrix}. \tag{4.22}$$

For a Fisher-consistent scale estimator, $F_\sigma(\mathbf{X}) = F_0(\mathbf{X}/\sigma)$ for all $\sigma > 0$. In general, the influence function of T at the distribution F is defined as:

$$IF((\mathbf{X}, y), T, H) = \lim_{\epsilon \longrightarrow 0} \frac{T\left((1-\epsilon)H + \epsilon\Delta_{(\mathbf{X},y)}\right) - T(H)}{\epsilon} = \frac{\partial}{\partial\epsilon}(T(\Delta_{(\mathbf{X},y)})), \tag{4.23}$$

where $T(H)$ is the functional defined as the solution of the objective model and $\Delta_{(\mathbf{X},y)}$ is the distribution which contains outliers. The following theorem gives the influence function of variable selection criteria, AIC, Cp, and SIC with any scale S.

4.5.2 Theorem 1

Let H be some distribution other than F. Take $(\mathbf{X}, y) \sim H$ and denote ϵ the error term of the model. Assume that S has the property that is differentiable with partial derivatives equal to zero at the origin (0, 0). Then:

1.

$$IF((\mathbf{X}, y), AIC_S, H) = \frac{2n}{(n-k)} IF\left(r_{ik}/\sigma_k, S, F_0\right), \tag{4.24}$$

2.

$$IF((\mathbf{X}, y), Cp_S, H) = \frac{2(n-k)S_k^2}{S_p^2}\left(IF(r_{ik}/\sigma_k, S, F_0) - IF(r_p/\sigma_p, S, F_0)\right), \tag{4.25}$$

3.

$$IF((\mathbf{X}, y), SIC_S, H) = \frac{2n}{(n-k)} IF\left(r_{ik}/\sigma_k, S, F_0\right), \tag{4.26}$$

where $r_{ik} = y_i - \widehat{\mu} - \mathbf{x}_{ik}^T \boldsymbol{\beta}_k$, $\sigma_k^2 = SSE/(n-k)$, S_k are computed from sub model K, furthermore, $\sigma_p^2 = SSE/(n-p)$ and S_p are computed from full model.

4.5.3 Proof of Theorem 1

Proof of Theorem 1 (1.)

By using the definition of influence function in Eqn. (4.23),

$$IF((\mathbf{X}, y), AIC, H) = \frac{\partial}{\partial \epsilon}(AIC(\Delta_{(\mathbf{X},y)}))|_{\epsilon=0}$$

$$= \frac{\partial}{\partial \epsilon} \log(\tfrac{n-k}{n}.S_k^2(H_\epsilon)) + 2k|_{\epsilon=0}$$

$$= \frac{n}{(n-k).S_k^2(H)}.\frac{\partial}{\partial\epsilon}(S_k^2(H_\epsilon))|_{\epsilon=0} \tag{4.27}$$

$$= \frac{n}{(n-k)\sigma_k^2}.2\sigma_k\frac{\partial}{\partial\epsilon}(S_k(H_\epsilon)|_{\epsilon=0} \tag{4.28}$$

where, $AIC(\Delta_{(\mathbf{X},y)})) = \log(\frac{n-k}{n}.S_k^2(H_\epsilon)) + 2k$, $AIC(H) = \log(\frac{n-k}{n}.S_k^2(H)) + 2k$, and $S_k^2(H) = \sigma_k^2$.

In the regression status there is a stochastic variable $(\mathbf{X}, y)$ at $(p+1)$ dimensional distribution H. Therefore, the regression functional identical to Eqn. (4.12) is given by:

$$(a(H), \mathbf{b}(H)) = \arg\min_{(\mu,\boldsymbol{\beta})} S\left(y_i - \mu - \mathbf{X}_i^T\boldsymbol{\beta}\right) \tag{4.29}$$

eventually obtain,

$$S_k^2(H) = S^2\left(y_i - a(H) - \mathbf{X}_i^T\mathbf{b}(H)\right) \tag{4.30}$$

and,

$$S_k(H_\varepsilon) = S\left(y_i - a(H_\varepsilon) - \mathbf{X}_i^T\mathbf{b}(H_\varepsilon)\right). \tag{4.31}$$

Suppose B_ε is Bernoulli variable, which takes value 1 with success probability $(1-\varepsilon)$ and value 0 with failure probability ε, $\varepsilon > 0$, $a \in \mathbb{R}$, and $b \in \mathbb{R}^k$, then,

$$S_k(H_\varepsilon) = S_k\left(B_\varepsilon(y_i - a(H_\varepsilon) - \mathbf{X}_i^T\mathbf{b}(H)_\varepsilon) + (1-B_\varepsilon)(y_i - a(H_\varepsilon) - \mathbf{X}_i^T\mathbf{b}(H)_\varepsilon)\right). \tag{4.32}$$

Equivalently

$$S_k(H_\varepsilon) = S_k\left(B_\varepsilon(y_i - a - \mathbf{X}_i^T\mathbf{b}) + (1-B_\varepsilon)(y_i - a - \mathbf{X}_i^T\mathbf{b})\right)$$

$$+S\left(\varepsilon + \mu - a(H_\varepsilon)\right) + S\left(\varepsilon + \boldsymbol{\beta}\mathbf{X}^T - \mathbf{b}(H)_\varepsilon\mathbf{X}^T\right). \tag{4.33}$$

If the derivative are computed at $\varepsilon = 0$ of $S_k(H_\varepsilon)$, ε indicate now for the error term, obtain,

$$\frac{\partial}{\partial \varepsilon}\left(S\left(\varepsilon + \mu - a(H_\varepsilon)\right)\right)|_{\epsilon=0} = 0,$$

and

$$\frac{\partial}{\partial \varepsilon}\left(S\left(\varepsilon + \boldsymbol{\beta}\mathbf{X}^T - \mathbf{b}(H)_\varepsilon \mathbf{X}^T\right)\right)|_{\epsilon=0} = 0,$$

then,

$$\frac{\partial}{\partial \epsilon} S_k(H_\epsilon)|_{\epsilon=0} = \frac{\partial}{\partial \epsilon} S_k\left(B_\varepsilon(y_i - a - \mathbf{X}^T\mathbf{b}) + (1 - B_\varepsilon)(y_i - a - \mathbf{X}^T\mathbf{b})\right)|_{\epsilon=0}$$

$$= \frac{\partial}{\partial \epsilon} S((1-\epsilon)F_{\sigma_k} + \epsilon\Delta_{r_{ik}})|_{\epsilon=0}$$

$$= \sigma_k \frac{\partial}{\partial \epsilon} S(\big(1-\epsilon)F_0 + \epsilon\Delta_{r_{ik}/\sigma_k}\big)|_{\epsilon=0}$$

$$= \sigma_k IF(r_{ik}/\sigma_k, S, F_\sigma) \tag{4.34}$$

where $r_{ik} = y_i - \widehat{\mu} - \mathbf{x}_{ik}^T\widehat{\boldsymbol{\beta}}_k$. Inserting Eqn. (4.34) into Eqn. (4.28) yields,

$$IF((\mathbf{X}, y), AIC, H) = \frac{2n}{(n-k)} IF\left(r_{ik}/\sigma_k, S, F_0\right) \tag{4.35}$$

Proof of Theorem 1 (2.)

$$IF((\mathbf{X}, y), Cp, H) = \frac{\partial}{\partial \epsilon}(Cp(\Delta_{(\mathbf{x},y)}))|_{\epsilon=0}$$

$$= \frac{\partial}{\partial \epsilon}\left(\frac{(n-k)S_k^2(H_\epsilon)}{S_p^2(H_\epsilon)} - n + 2k\right)|_{\epsilon=0}$$

$$= \frac{2(n-k)S_k(H)S_p^2(H)\frac{\partial}{\partial \epsilon}S_k(H_\epsilon)}{\sigma_p^4} - \frac{(n-k)S_k^2.2S_p(H)\frac{\partial}{\partial \epsilon}S_p(H_\epsilon)}{\sigma_p^4}|_{\epsilon=0}$$

$$= \frac{2(n-k)\sigma_k\sigma_p^2\frac{\partial}{\partial \epsilon}S_k(H_\epsilon)}{\sigma_p^4} - \frac{2(n-k)\sigma_k^2\sigma_p^2\frac{\partial}{\partial \epsilon}S_p(H_\epsilon)}{\sigma_p^4}|_{\epsilon=0}$$

$$= \frac{2(n-k)\sigma_k\sigma_p[\sigma_p\frac{\partial}{\partial\epsilon}S_k(H_\epsilon) - \sigma_k\frac{\partial}{\partial\epsilon}S_p(H_\epsilon)]}{\sigma_p^2}|_{\epsilon=0} \tag{4.36}$$

where, $Cp(\Delta_{(\mathbf{X},y)}) = \frac{(n-k)S_k^2(H_\epsilon)}{S_p^2(H_\epsilon)} - n + 2k$, $Cp(H) = \frac{(n-k)S_k^2(H)}{S_p^2(H)} - n + 2k$, $S(H)_k^2 = \sigma_k^2$, and $S(H)_p^2(H) = \sigma_p^2$. In an analogous way to Eqn. (4.36) it can be shown that,

$$\frac{\partial}{\partial\epsilon}S_k(H_\epsilon) = \sigma_k IF(r_{ik}/\sigma_k, S, F_0), \tag{4.37}$$

and

$$\frac{\partial}{\partial\epsilon}S_p(H_\epsilon) = \sigma_p IF(r_{ip}/\sigma_p, S, F_0), \tag{4.38}$$

Inserting Eqn. (4.36) and Eqn. (4.37) into Eqn. (4.38) yields,

$$IF((\mathbf{X}, y), Cp, H) = \frac{2(n-k)\sigma_k}{\sigma_p}\left(\sigma_k\sigma_p IF(r_{ik}/\sigma_k, S, F_0) - \sigma_k\sigma_p IF(r_{ip}/\sigma_p, S, F_0)\right)$$

$$= \frac{2(n-k)\sigma_k}{\sigma_p^2}\left(IF(r_{ik}/\sigma_k, S, F_0) - IF(r_{ip}/\sigma_p, S, F_0)\right). \tag{4.39}$$

Proof of Theorem 1 (3.)

$$IF((\mathbf{X}, y), SIC, H) = \tfrac{\partial}{\partial\epsilon}(SIC(\Delta_{(\mathbf{X},y)}))|_{\epsilon=0}$$

$$= \tfrac{\partial}{\partial\epsilon}\log\left(\tfrac{n-k}{n}.S_k^2(H_\epsilon) + p\tfrac{\log(n)}{n}\right)|_{\epsilon=0}$$

$$= \frac{n}{(n-k).S_k^2(H)}.\frac{\partial}{\partial\epsilon}(S_k^2(H_\epsilon))|_{\epsilon=0} \tag{4.40}$$

$$= \frac{n}{(n-k)\sigma_k^2}.2\sigma_k\frac{\partial}{\partial\epsilon}(S_k(H_\epsilon)|_{\epsilon=0} \tag{4.41}$$

where, $SIC(\Delta_{(\mathbf{X},y)}) = \log\left(\frac{n-k}{n}.S_k^2(H_\epsilon) + p\frac{\log(n)}{n}\right)$, $SIC(H) = \log\left(\frac{n-k}{n}.S_k^2(H) + p\frac{\log(n)}{n}\right)$, and $S_k^2(H) = \sigma_k^2$.

In a similar way to Eqn. (4.36) can be shown, as

$$IF((\mathbf{X}, y), SIC, H) = \frac{2n}{(n-k)}IF(r_{ik}/\sigma_k, S, F_0). \tag{4.42}$$

End of the proof of Theorem 1

Eqns. (4.35), (4.39), and (4.42) give the influence function of different variable selection criteria using any scale S. Subsequently, the influence function of different classical variable selection now follow Theorem 1 by using a well-known expression of influence function of estimator of scale.

4.5.4 Proposition 1

Let $(\mathbf{X}, y) \sim H$ where the distribution H verifies (H). Then

1.

$$IF((\mathbf{X}, y), AIC_S, H) = \frac{2n}{(n-k)E_{F_0}\left[\rho'(\varepsilon)\varepsilon\right]} \cdot \rho\left(\frac{y_i - \mu - \boldsymbol{\beta}_k \mathbf{x}_{ik}^T}{\sigma_k}\right), \qquad (4.43)$$

2.

$$IF((\mathbf{X}, y), Cp_S, H) = \frac{2(n-k)\sigma_k}{\sigma_p^2 E_{F_0}\left[\rho'(\varepsilon)\varepsilon\right]} \cdot \left(\rho\left(\frac{y_i - \mu - \boldsymbol{\beta}_k \mathbf{x}_{ik}^T}{\sigma_k}\right) - \rho\left(\frac{y_i - \mu - \boldsymbol{\beta}_p \mathbf{x}_{ip}^T}{\sigma_p}\right)\right), \qquad (4.44)$$

3.

$$IF((\mathbf{X}, y), SIC_S, H) = \frac{2n}{(n-k)E_{F_0}\left[\rho'(\varepsilon)\varepsilon\right]} \cdot \rho\left(\frac{y_i - \mu - \boldsymbol{\beta}_k \mathbf{x}_{ik}^T}{\sigma_k}\right). \qquad (4.45)$$

Results and Discussion

The influence function of different variable selection criteria is bounded if the ρ-function is bounded. Figure (4.5) showed the influence function for a bivariate normal distribution H with associated regression parameters $\mu = 0$ and $\boldsymbol{\beta} = 1$ (Anderson-Sprecher, 1994).

The classical criteria AIC, Cp, and SIC are non-robust, since ρ in Proposition 1 equals, $(y_i - \mu - \mathbf{X}^T\boldsymbol{\beta})^2$. Thus, $IF((\mathbf{X}, y),\ AIC_{LS}, H)$, $IF((\mathbf{X}, y),\ Cp_{LS}, H)$, and $IF((\mathbf{X}, y),\ SIC_{LS}, H)$ are unbounded in two directions, X and Y (see Figure (4.5) (a)). For M-estimators, $\rho_M = \mathbf{X}.\rho(y_i - \mu - \mathbf{X}^T\boldsymbol{\beta})$, therefore $IF((\mathbf{X}, y),\ AIC_M, H)$, $IF((\mathbf{X}, y),\ Cp_M, H)$, and $IF((\mathbf{X}, y),\ SIC_M, H)$ are unbounded influence function with respect to the X direction. Figure (4.5) (b) shows that M-estimator is robust with respect to vertical outliers, but breaks down in the presence of large leverage points. For good leverage points, the $IF((\mathbf{X}, y),\ AIC_M, H)$, $IF((\mathbf{X}, y),\ Cp_M, H)$, and $IF((\mathbf{X}, y),\ SIC_M, H)$ tend to infinity, while bad leverage point with enormous $\mathbf{X}$ values can have unbounded negative influence on robust variable selection criteria for M-estimator.

Whereas, the influence function for proposed criteria, $IF((\mathbf{X}, y), AIC_{LMS}, H)$, $IF((\mathbf{X}, y),\ Cp_{LMS}, H)$, and $IF((\mathbf{X}, y), SIC_{LMS}, H)$ are bounded and discontinuous (see Figure (4.5) (c) and 4.5 (d)). Note that a large zone outliers have zero influence, even when they are bad leverage points. This is because they have a similar influence on the spread of y as on the spread of the residuals. Bad leverage points with an outlying y value are harmless, and give a zero value for influence function of high breakdown scale estimate. On the other hand, bad leverage points with non outlying y values have a negative, but bounded influence.

4.6 The Gross-Error Sensitivity of Variable Selection Criteria

The gross-error sensitivity of the different variable selection criteria is defined as the supreme influence that an observation can have. If $\boldsymbol{\beta} = 0$, then $IF = 0$, so it is assumed that $\boldsymbol{\beta} \neq 0$ and the observation $(\mathbf{X}, y)$ follows the regression line $y_i = \mu + \mathbf{X}^T\boldsymbol{\beta}$, where S is LTS-scale estimate. Then, if $\mathbf{X}$ tend to ∞, the gross-error sensitivity of AIC_{LTS},

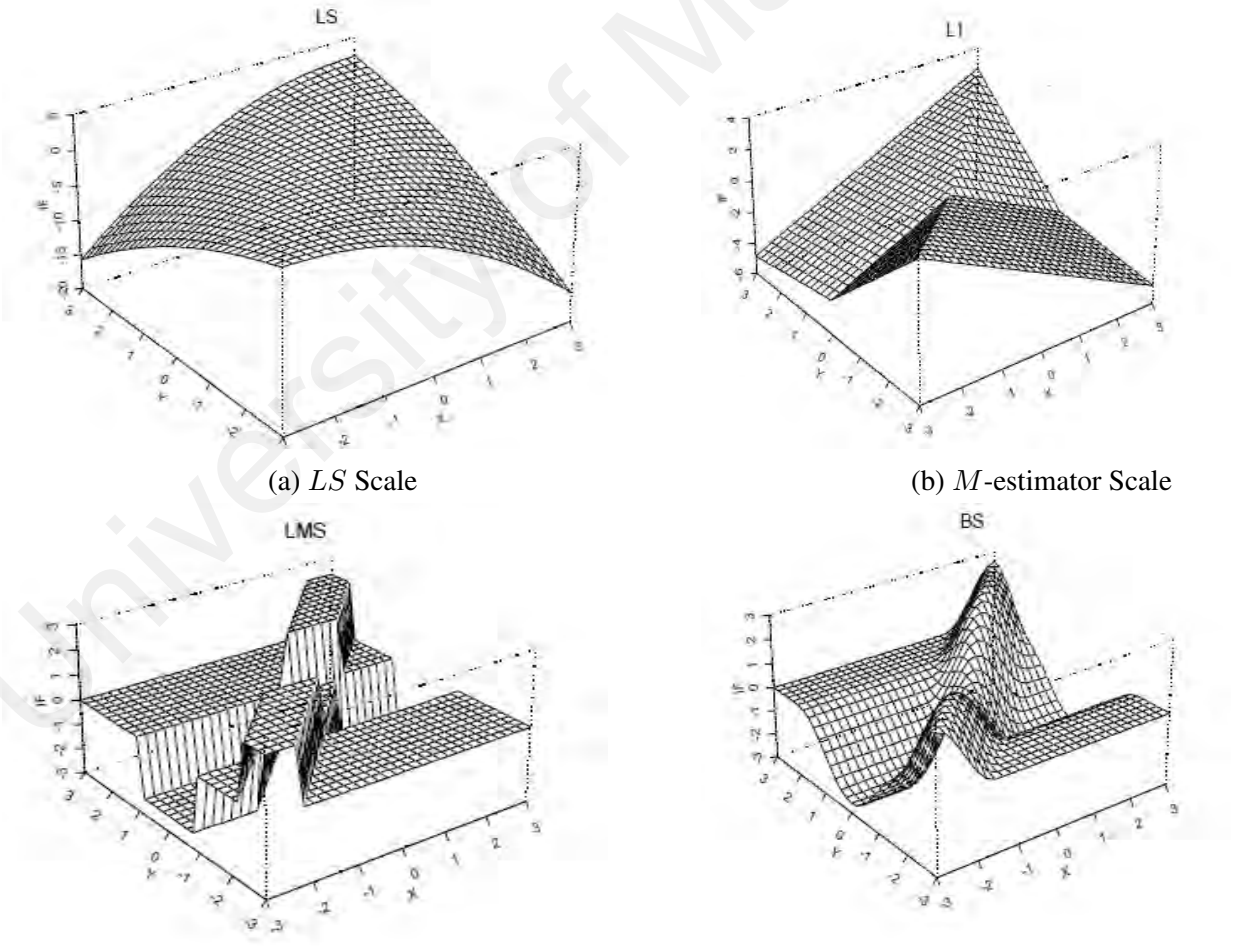
Cp_{LTS}, and SIC_{LTS} will turn into:

$$\gamma^*(AIC_{LTS}, F) = \sup_{(\mathbf{X},y)} IF((\mathbf{X}, y), AIC_{LTS}, H) = \frac{2n}{(n-k)E_{F_0}\left[\rho'(\varepsilon)\varepsilon\right]} \cdot \rho(\infty), \quad (4.46)$$

$$\gamma^*(Cp_{LTS}, F) = \sup_{(\mathbf{X},y)} IF((\mathbf{X}, y), Cp_{LTS}, H) = \frac{2(n-k)\sigma_k}{\sigma_p^2 E_{F_0}\left[\rho'(\varepsilon)\varepsilon\right]} \cdot \rho(\infty), \quad (4.47)$$

$$\gamma^*(SIC_{LTS}, F) = \sup_{(\mathbf{X},y)} IF((\mathbf{X}, y), SIC_{LTS}, H) = \frac{2n}{(n-k)E_{F_0}\left[\rho'(\varepsilon)\varepsilon\right]} \cdot \rho(\infty). \quad (4.48)$$

Briefly, if $\mathbf{X}$ tends to infinity, both LS and M-estimators gain $\rho-$function yields high gross-error sensitivity. On the other hand, high breakdown and bounded influence estimator compute with $\rho-$function which yield the lowest γ^*.



(a) LS Scale (b) M-estimator Scale

(c) LMS Scale (d) BS Scale

Figure 4.5: Plot of influence function of LS, M, LMS, and BS estimators of scales

4.7 Simulations

In this section, several simulation studies are discussed that are conducted with the aim to investigate the performance of the proposed criteria statistic for detecting best variable in the linear regression model Eqn. (1.1) based on Eqns. (4.14), (4.15), and (4.16). In this study, 50 independent replicates of 3 independent uniform random variables on [-1,1] of $\mathbf{x}_{i1}$, $\mathbf{x}_{i2}$ and $\mathbf{x}_{i3}$, and 50 independent normally distributed errors $\varepsilon_i \sim N(0, 9)$ were generated. The true model is given by $y_i = \mathbf{x}_{i1} + \mathbf{x}_{i2} + \varepsilon_i$, for $i = 1, ..., 50$ using two variables $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$. In order to illustrate the robustness to outliers, the following cases are considered:

- Vertical outliers (outliers in the y only)
- Bad leverage points (outliers in some **X** only)
- good leverage points (outliers in **X** follow the pattern of the majority of the data)

For vertical outliers case, we randomly generated different percentages of outliers (0%, 5%, 10%, 20%, 30%, and 40%) from $N(50, 0.1^2)$ for each of the simulated cases. For good leverage case, we considered the different percentages of outliers (0%, 5%, 10%, 20%, 30%, and 40%) on the variables X_1 and X_2 were generated from a $N(100, 0.5^2)$ distribution, then generated y to get good leverage points. For bad leverage case, different percentages of outliers (0%, 5%, 10%, 20%, 30%, and 40%) on variables X_1 and X_2 were generated from a $N(100, 0.5^2)$ distribution. For each of these setting 1000 samples were simulated.

4.7.1 Performance of Simulations

The purpose of this simulation is based on the following aims:

- AIC_{LTS}, AIC_{LMS} and AIC_{BS} are compared with AIC_{LS} and AIC_M

- Cp_{LTS}, Cp_{LMS} and Cp_{BS} are compared with Cp_{LS} and Cp_M
- SIC_{LTS}, SIC_{LMS} and SIC_{BS} are compared with SIC_{LS} and SIC_M

4.7.2 Simulation Result

The resulting fit to the data is classified as one of the following:

- True model (correct fit)
- Models that contain all the variables in the true model plus other variables that are redundant (Over fit)
- Models that contain only a few of the variables in the true model (Under fit)
- Models that fit none of the above (Wrong fit)

Tables 4.1 to 4.3 provided the results of different versions of AIC with and without the presence of outliers and leverage point. AIC_{LS} performed better compared to robust AIC with high percentage (84.6%) in uncontaminated data. However, as the percentage of outliers increased to 5%, AIC_{LS} selected a large proportion of wrong fit models. The AIC_M continues to yielded higher percentage compared to AIC_{LS} and these results hold as the percentage of vertical outliers increased to 10%, then it tends to under fit. Thus, AIC_M method ignored some of the important variables in the model. As expected, the percentage of true model in all cases of AIC_S with high breakdown scale estimate was always large in the presence of less than 20% vertical outliers. Then, the same behaviour was obtained for AIC_M.

Table 4.2 reported the performance results of classical and robust AIC in the presence of outliers in **X** variables (bad leverage point). With 5% bad leverage points in the data, AIC_{LS} tended to produce over fit model, AIC_M tended to produce either an under

fit or wrong fit model, and the proposed criteria performed better. As the percentage of bad leverage point increases to more than 20%, the AIC_{LS} and AIC_M tended to produce wrong fit. All AIC_S with high breakdown scale estimate criteria, however tended to produce either correct fit or under fit the model. In the presence of outliers in $\mathbf{X}$ and y_i (good leverage points), AIC_{LS} tended to produce over fit. On the other hand, the robust AIC tend to produce either correct fit or under fit model.

Tables 4.4 to 4.6 presented the results of the simulation study for different versions of Mallows's Cp methods. For data without outliers, the classical Cp worked better than robust Cp_S. Cp_S selected a large proportion of correct fit model or over fit in this case. In the presence of 5% contamination, the Cp performed worse than the Cp_S criteria. Likewise, when the percentage of outliers increased at most 10%, the classical Cp selected a large proportion of under fit or wrong fit models, while robust Cp_S still selects higher percentage of correct fit model. In case of good leverage point, Table 4.5 illustrated that both Cp and RCp statistics pick up the right model with low percentage and preferred the over fit model in all level of contamination. However, Cp_{LTS} and Cp_{LMS} selected the correct fit and over fit models until 20% contamination. Next, Cp_{LTS} and Cp_{LMS} selected the correct fit model with a high percentage, while, Cp_{BS} gave a good solution here. Table 4.6 shows that the Cp and RCp would choose over fit model when 5% of data are bad leverage points. Both of these methods worked worse for the case with a high contamination level of bad leverage. Otherwise, the proposed methods picked the true model with good percentage.

Tables 4.7 to 4.9 showed the detail of simulation result for different versions of SIC methods. For uncontaminated data, the classical SIC_{LS} performed best compared to ro-

bust SIC, whereas it selected a large proportion of under fit or wrong fit models for the data with vertical outliers. As expected, the robust SIC will usually (that is, with higher proportion) select the correct model. For bad leverage points, it was observed that SIC_{LS} and SIC_M tended to produce either an over fit or wrong fit the model. However, the robust estimate produced comparable power in the presence of bad leverage points.

In general, robust variable selection criteria with M-estimation are robust in the presence of outliers in response variable (Y-direction). In the presence of high leverage point (X-direction), the value of these criteria will be affected and differs significantly from the true fit as the percentage of leverage point increases. But, the robust variable criteria with high breakdown scale estimate less affected in all cases in the presence of outliers in X- and Y-directions.

Table 4.1: Percentage of selecting correct models from classical AIC_{LS}, robust AIC_M, AIC_{LTS}, AIC_{LMS}, AIC_{BS}, with vertical outliers

ϵ %		AIC_{LS}	AIC_M	AIC_{LTS}	AIC_{LMS}	AIC_{BS}
0	Correct fit	84.6%	54.4%	57.6%	65.2%	45.2%
	Over fit	15.4%	0%	0%	0%	0%
	Under fit	0%	43.6%	41.2%	34%	54.6%
	Wrong fit	0%	2.0%	1.2%	0.8%	0.2%
5	Correct fit	2.8%	49.6%	56.8%	62.6%	45.4%
	Over fit	2%	0%	0%	0%	0%
	Under fit	22%	51%	42.8%	36.6%	54.6%
	Wrong fit	73.2%	1.4%	0.4%	0.8%	0%
10	Correct fit	3.2%	45.0%	51%	56%	39.8%
	Over fit	0.8%	0%	0%	0%	0%
	Under fit	20.2%	52.4%	48%	43.8%	55.2%
	Wrong fit	75.8%	2.6%	1.0%	0.2%	0%
20	Correct fit	4.2%	30.8%	49.4%	58%	34.2%
	Over fit	0.2%	0%	0%	0%	0%
	Under fit	22.8%	67.8%	49.6%	41.4%	65.8%
	Wrong fit	72.8%	1.4%	1.0%	0.6%	0%
30	Correct fit	1.6%	0%	44.0%	51.2%	23.2%
	Over fit	0.2%	0%	0%	0%	0%
	Under fit	22.6%	73.8%	55.4%	48.2%	76.8%
	Wrong fit	72.8%	26.2%	0.6%	0.6%	0%
40	Correct fit	2.2%	0%	37.6%	41.2%	12.4%
	Over fit	0.6%	0%	0%	0%	0%
	Under fit	22.8%	70.8%	62.4%	58.4%	87.4%
	Wrong fit	74.4%	29.2%	0%	0.4%	0.2%

Table 4.2: Percentage of selecting correct models from classical AIC_{LS} , robust AIC_M and robust AIC_{LTS}, AIC_{LMS}, AIC_{BS} with bad leverage points

ϵ %		AIC_{LS}	AIC_M	AIC_{LTS}	AIC_{LMS}	AIC_{BS}
5	Correct fit	0%	0%	54.6%	60.8%	43.8%
	Over fit	70.4%	0%	0%	0%	0%
	Under fit	0%	64.4%	44%	38.4%	55.4%
	Wrong fit	29.6%	35.6%	1.4%	0.8%	0.8%
10	Correct fit	0%	0%	63.8%	67.8%	51.0%
	Over fit	63%	0%	0%	0%	0%
	Under fit	0%	54.6%	34.6%	31.4%	48.8%
	Wrong fit	37%	42.4%	1.6%	0.8%	0.2%
20	Correct fit	0%	0%	56.6%	63.4%	49.2%
	Over fit	54.8%	0%	0%	0%	0%
	Under fit	0.2%	60.8%	42.4%	35.8%	50.6%
	Wrong fit	44.8%	39.2%	1.0%	0.8%	0.2%
30	Correct fit	0%	0%	56.6%	61.4%	46%
	Over fit	37.6%	0%	0%	0%	0%
	Under fit	0.25%	29.6%	42.6%	36%	53.8%
	Wrong fit	60.4%	42.6%	0.8%	2.4%	0.2%
40	Correct fit	1.0%	0%	55.2%	64.6%	51.4%
	Over fit	13.8%	0%	0%	0%	0%
	Under fit	1.2%	54.4%	43.8%	33.8%	48.6%
	Wrong fit	81%	45.4%	1.0%	1.6%	0%

Table 4.3: Percentage of selecting correct models from classical AIC_{LS}, robust AIC_M and robust AIC_{LTS} ,AIC_{LMS}, AIC_{BS}, with good leverage points

ϵ %		AIC_{LS}	AIC_M	AIC_{LTS}	AIC_{LMS}	AIC_{BS}
5	Correct fit	0.2%	47.0%	53.6%	58.4%	42.6%
	Over fit	99.8%	0%	0%	0%	0%
	Under fit	0%	50.4%	46%	41.2%	57.4%
	Wrong fit	0%	2.6%	0.4%	0.4%	0%
10	Correct fit	0%	44.2%	54.6%	59.4%	40.4%
	Over fit	99.6%	0%	0%	0%	0%
	Under fit	0.2%	53.8%	44.6%	39.8%	59.6%
	Wrong fit	0.2%	2.0%	0.8%	0.8%	0%
20	Correct fit	0.8%	32.4%	50.4%	56.2%	33.2%
	Over fit	97.6%	0%	0%	0%	0%
	Under fit	0.8%	66.6%	48.4%	43%	66.4%
	Wrong fit	0.8%	1.0%	1.2%	0.8%	0.2%
30	Correct fit	1.8%	0%	46.0%	50.6%	27.0%
	Over fit	97.8%	96.8%	0%	0%	0%
	Under fit	2.8%	2.8%	53.6%	49.2%	72.8%
	Wrong fit	0.6%	0.4%	0.4%	0.2%	0.2%
40	Correct fit	0.2%	0%	37.4%	37%	13.8%
	Over fit	97.4%	100%	0%	0%	0%
	Under fit	2.2%	0%	62.6%	62.6%	86.2%
	Wrong fit	0.2%	0%	0%	0.4%	0%

Table 4.4: Percentage of selecting correct models from classical Cp_{LS}, robust Cp_M, Cp_{LTS}, Cp_{LMS}, Cp_S, with vertical outliers

ϵ %		Cp_{LS}	Cp_M	Cp_{LTS}	Cp_{LMS}	Cp_S
0	Correct fit	83.8	56.8%	51.4%	54.6%	74.6%
	Over fit	16.2%	42%	48.6%	45.4%	25.4%
	Under fit	0%	0%	0%	0%	0%
	Wrong fit	0%	0%	0%	0%	0%
5	Correct fit	2.0%	51.4%	54.2%	58%	76.6%
	Over fit	0.6%	48.6%	45.8%	42%	23.4%
	Under fit	61.4%	0%	0%	0%	0%
	Wrong fit	36%	0%	0%	0%	0%
10	Correct fit	2.8%	55.2%	59.8%	62.6%	85.8%
	Over fit	0.2%	44.8%	40.2%	37.4%	14.2%
	Under fit	62.2%	0%	0%	0%	0%
	Wrong fit	34.8%	0%	0%	0%	0%
20	Correct fit	3.4%	59.8%	71.2%	70.8%	91.4%
	Over fit	1.0%	40.2%	28.8%	29.2%	8.6%
	Under fit	61.2%	0%	0%	0%	0%
	Wrong fit	4.4%	0%	0%	0%	0%
30	Correct fit	2.0%	19.6%	74.2%	75.8%	93.6%
	Over fit	1.0%	75.4%	25.8%	24.2%	6.4%
	Under fit	59.8%	0.8%	0%	0%	0%
	Wrong fit	37.2%	4.2%	0%	0%	0%
40	Correct fit	3.2%	14%	88.2%	84.2%	99.2%
	Over fit	0.4%	41.4%	11.8%	15.8%	0.8%
	Under fit	60.2%	14.6%	0%	0%	0%
	Wrong fit	36.2%	30.0%	0%	0%	0%

Table 4.5: Percentage of selecting correct models from classical Cp_{LS}, robust Cp_M, Cp_{LTS}, Cp_{LMS}, Cp_S with good leverage points

ϵ %		Cp_{LS}	Cp_M	Cp_{LTS}	Cp_{LMS}	Cp_S
5	Correct fit	1.8%	7.6%	32.8%	34.2%	70.4%
	Over fit	98.2%	92%	67.2%	65.8%	29.6%
	Under fit	0%	0%	0%	0%	0%
	Wrong fit	0%	0%	0%	0%	0%
10	Correct fit	2.6%	8%	33.4%	30%	69.4%
	Over fit	97.4%	92%	66.6%	70%	30.6%
	Under fit	0%	0%	0%	0%	0%
	Wrong fit	0%	0%	0%	0%	0%
20	Correct fit	6.6%	9.8%	34.2%	38.2%	78%
	Over fit	93.4%	90.2%	65.8%	61.2%	22%
	Under fit	0%	0%	0%	0%	0%
	Wrong fit	0%	0%	0%	0%	0%
30	Correct fit	0.2%	0.6%	13%	25%	85.4%
	Over fit	99%	95.4%	87%	75%	14.6%
	Under fit	0%	1.6%	0%	0%	0%
	Wrong fit	0.8%	2.4%	0%	0%	0%
40	Correct fit	0%	0.4%	0.8%	2.8%	79%
	Over fit	98%	99.2%	99.2%	97.2%	21%
	Under fit	2.0%	0.4%	0%	0%	0%
	Wrong fit	0.8%	2.4%	0%	0%	0%

Table 4.6: Percentage of selecting correct models from classical Cp_{LS}, robust Cp_M, Cp_{LTS}, Cp_{LMS}, Cp_S with bad leverage points

ϵ %		Cp_{LS}	Cp_M	Cp_{LTS}	Cp_{LMS}	Cp_S
5	Correct fit	0%	0.8%	70.4%	71.6%	92.6%
	Over fit	66.2%	69.4%	29.6%	28.4%	7.4%
	Under fit	0%	0.8%	0%	0%	0%
	Wrong fit	33.8%	29%	0%	0%	0%
10	Correct fit	0%	2.0%	87.6%	86.8%	98.8%
	Over fit	64.4%	70.2%	12.4%	13.2%	1.2%
	Under fit	0%	0.6%	0%	0%	0%
	Wrong fit	35%	27.2%	0%	0%	0%
20	Correct fit	0%	2.0%	98.8%	98.4%	100%
	Over fit	51.6%	65.6%	1.2%	1.6%	0%
	Under fit	0.2%	0.6%	0%	0%	0%
	Wrong fit	48.2%	31.8%	0%	0%	0%
30	Correct fit	0%	2.8%	100%	100%	100%
	Over fit	16.2%	42%	0%	0%	0%
	Under fit	1.6%	3.8%	0%	0%	0%
	Wrong fit	82.2%	51.4%	0%	0%	0%
40	Correct fit	0.2%	2.6%	100%	100%	100%
	Over fit	13.4%	36.4%	0%	0%	0%
	Under fit	4.2%	5.6%	0%	0%	0%
	Wrong fit	82.2%	55.4%	0%	0%	0%

Table 4.7: Percentage of selecting correct models from classical SIC_{LS}, robust SIC_M, SIC_{LTS}, SIC_{LMS}, SIC_{BS}, with vertical outliers

ϵ %		SIC_{LS}	SIC_M	SIC_{LTS}	SIC_{LMS}	SIC_{BS}
0	Correct fit	94.0%	63.2%	38.0%	38.0%	74.2%
	Over fit	6.0%	36.8%	62.0%	62.0%	25.8%
	Under fit	0%	0%	0%	0%	0%
	Wrong fit	0%	0%	0%	0%	0%
5	Correct fit	0%	66.2%	36.8%	36.8%	74.6%
	Over fit	0.2%	33.8%	63.2%	63.2%	25.4%
	Under fit	64.4%	0%	0%	0%	0%
	Wrong fit	35.4%	0%	0%	0%	0%
10	Correct fit	0.6%	66.8%	46.8%	46.8%	83.4%
	Over fit	0.2%	33.2%	53.2%	53.2%	16.6%
	Under fit	67.4%	0%	0%	0%	0%
	Wrong fit	31.8%	0%	0%	0%	0%
20	Correct fit	0.6%	68.2%	50.4%	50.4%	90.2%
	Over fit	0%	31.8%	49.6%	49.6%	9.8%
	Under fit	67.6.8%	0%	0%	0%	0%
	Wrong fit	31.8%	0%	0%	0%	0%
30	Correct fit	0%	71.8%	62.6%	62.6%	93.6%
	Over fit	0%	27.2%	37.4%	0%	6.4%
	Under fit	68%	1%	0%	0%	0%
	Wrong fit	32%	0%	0%	0%	0%
40	Correct fit	0.4%	65.8%	37.2%	37.2%	98.4%
	Over fit	0%	16.4%	26.8%	26.8%	1.6%
	Under fit	63.4%	13.2%	0%	0%	0%
	Wrong fit	36.2%	4.6%	0%	0%	0%

Table 4.8: Percentage of selecting correct models from classical SIC_{LS}, robust SIC_M and robust SIC_{LTS}, SIC_{LMS}, SIC_{BS} with good leverage points

ϵ %		SIC_{LS}	SIC_M	SIC_{LTS}	SIC_{LMS}	SIC_{BS}
5	Correct fit	6.8%	18.6%	41.6%	41.6%	78.4%
	Over fit	93.2%	80.6%	58.4%	58.4%	21.6%
	Under fit	0%	0.6%	0%	0%	0%
	Wrong fit	0%	0.2%	0%	0%	0%
10	Correct fit	16.4%	22.0%	42.2%	42.2%	80.4%
	Over fit	83.6%	77.2%	57.8%	57.8%	19.6%
	Under fit	0.8%	0%	0%	0%	0%
	Wrong fit	0%	0%	0%	0%	0%
20	Correct fit	34.0%	24.4%	48.4%	48.4%	85.4%
	Over fit	65.8%	74.8%	51.6%	51.6%	14.6%
	Under fit	0.2%	0.4%	0%	0%	0%
	Wrong fit	0%	0.4%	0%	0%	0%
30	Correct fit	42.2%	23.8%	50.2%	50.2%	74.8%
	Over fit	57.4%	74.0%	46.6%	46.6%	24.8%
	Under fit	0.2%	1.4%	0.2%	0%	0.2%
	Wrong fit	0.2%	0.6%	0%	0%	0.2%
40	Correct fit	51.8%	23.4%	42.8%	42.8%	31.6%
	Over fit	42.4%	72.8%	54.6%	54.6%	65.4%
	Under fit	5.4%	2.6%	0.4%	0.4%	2%
	Wrong fit	0.4%	1.2%	2.2%	2.2%	1%

Table 4.9: Percentage of selecting correct models from classical SIC_{LS}, robust SIC_M and robust SIC_{LTS}, SIC_{LMS}, SIC_{BS}, with bad leverage points

ϵ %		SIC_{LS}	SIC_M	SIC_{LTS}	SIC_{LMS}	SIC_{BS}
5	Correct fit	0%	3.8%	38.6%	38.6%	76.6%
	Over fit	35.2%	47.2%	61.4%	61.4%	23.4%
	Under fit	0.4%	4.4%	0%	0%	0%
	Wrong fit	64.2%	44.6%	0%	0%	0%
10	Correct fit	0%	3.8%	43.8%	43.8%	81.8%
	Over fit	38.0%	44.4%	56.2%	56.2%	18.2%
	Under fit	0.6%	6%	0%	0%	0%
	Wrong fit	61.4%	45.8%	0%	0%	0%
20	Correct fit	0%	3.4%	52.2%	52.2%	87.6%
	Over fit	32.2%	39.8%	47.8%	47.8%	12.4%
	Under fit	2.2%	7%	0%	0%	0%
	Wrong fit	65.6%	49.8%	0%	0%	0%
30	Correct fit	0%	4.6%	62.0%	62.0%	94.2%
	Over fit	31.2%	39.8%	37.8%	37.8%	5.4%
	Under fit	4%	6.6%	0%	0%	0%
	Wrong fit	64.8%	49%	0.2%	1.6%	0.4%
40	Correct fit	0%	6.0%	63.0%	63.0%	61.6%
	Over fit	28.8%	38.4%	32.6%	32.6%	13.8%
	Under fit	6.8%	6.8%	0%	0%	3.2%
	Wrong fit	64.4%	48.8%	4.4%	4.4%	21.4%

4.8 Practical Example: (Stack Loss Data)

Stack Loss data appear in by Brownlee (1965). The data set consists of 21 observations on three independent variables ((1) $\mathbf{x}_{i1}$ =**Air.Flow**: represents operation rate of plant; (2) $\mathbf{x}_{i2}$ =**Water.Temp**: temperature of cooling water; and (3) $\mathbf{x}_{i3}$ =**Acid.Conc**) and contains four outliers (cases 1, 3, 4, and 21) and high leverage points (cases 1, 2, 3 and 21). These observations (10% of the data) are considered as leverage outliers. The data are given in Appendix 3, and the following model is considered: model:

$$Stack.Loss = \beta_0 + \beta_1 \textbf{Air.Flow} + \beta_2 \textbf{Water.Temp} + \beta_3 \textbf{Acid.Conc}.$$

The least squares estimates are $\hat{\beta}_0 = -39.920$, $\hat{\beta}_1 = 0.716$, $\hat{\beta}_2 = 1.295$, and $\hat{\beta}_3 = -0.152$. Thus, the regression line is

$$\widehat{Stack.Loss} = -39.920 + 0.716\textbf{Air.Flow} + 1.295\textbf{Water.Temp} - 0.152\textbf{Acid.Conc}. \tag{4.49}$$

In addition, the LMS estimator for this data given by

$$\widehat{Stack.Loss} = -34.5 + 0.714\textbf{Air.Flow} + 0.357\textbf{Water.Temp} + 0.000\textbf{Acid.Conc}.$$

However, the robust LMS estimator suggested that the true model contains x_1 and x_2. the Q-Q plots (Figure (4.6)) of the residuals associated with the fitted model and suggest the occurrence of outliers in the data set. Hence, the best variable selection procedures proposed in this chapter on the data set.

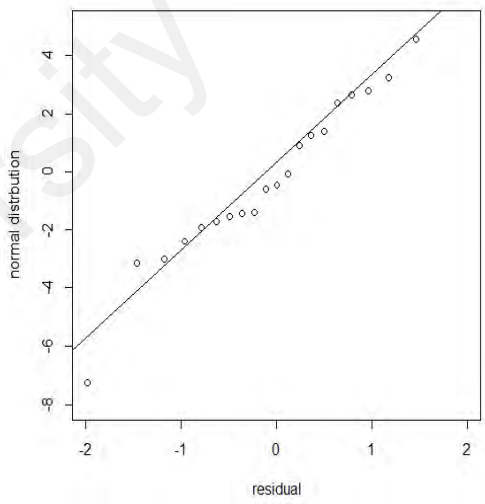


Figure 4.6: Q-Q plot for regression residuals of Stack Loss data

Best Variable Selection of Stack Loss Data

All 2^3 possible models are fitted with a combination of any of these covariant and computed several criteria AIC, Cp, and SIC values for each model. The results are comparable to the simulations part. The classical AIC and SIC select a model for all three

explanatory variables. Tables 4.10 to 4.12, however shows that AIC with M-estimation selects either under fit ($\mathbf{x}_{i1}$) or wrong fit ($\mathbf{x}_{i3}$). A relatively high value of classical Cp has suggested that this may be a scope to improve the fitting to the model based on the outcome of robust RCp with the full model. The value of RCp is close to Vp value given in section 2.6, and for the true model $\mathbf{x}_{i1}, \mathbf{x}_{i2}$ ($\mathbf{x}_{i1}, \mathbf{x}_{i2}$ have the positive LS value estimation), the value of V_p is smaller than RCp. On the other hand, all robust criteria with high breakdown point estimators methods the importance of two variables $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$, because the breakdown point of the estimators conciliates about 10% of outliers in the data.

Table 4.10: The different version of AIC selection variable criterion of Stack Loss data

Selected variables	AIC	$RAIC$	AIC_{LTS}	AIC_{LMS}	AIC_{BS}
x_1	6.7	8.0	**4.7**	**4.3**	**5.5**
x_2	7.1	**6.5**	5.9	5.4	7.0
x_3	8.4	**7.3**	7.0	6.4	7.3
x_1, x_2	8.2	9.0	**5.5**	**4.7**	**6.9**
x_1, x_3	8.7	8.9	6.9	6.3	7.6
x_2, x_3	9.1	9.0	8.1	7.3	8.8
x_1, x_2, x_3	**4.7**	10.6	7.6	6.7	9.1

Table 4.11: The different version of Cp selection variable criterion of Stack Loss data

Selected variables	Cp	RCp (V_p)	Cp_{LTS}	Cp_{LMS}	Cp_S
x_1	13.3	-2.1(1.96)	93.7	158.2	14.8
x_2	28.9	34.9(1.96)	404.9	565.6	221.9
x_3	148.9	62.4(1.96)	1260.3	1557	309.9
x_1, x_2	**2.95**	4.47(2.95)	**-1.22**	**-2.92**	**-8.15**
x_1, x_3	14.3	7.2(2.95)	124.9	155.9	23.24
x_2, x_3	30.1	61.6(2.95)	515.2	554.9	182.3
x_1, x_2, x_3	4.0	**3.9(3.93)**	4.0	4.0	4.0

Table 4.12: The different version of SIC selection variable criterion of Stack Loss data

Selected variables	SIC	SIC_M	SIC_{LTS}	SIC_{LMS}	SIC_S
x_1	3.010	4.273	1.023	0.581	1.796
x_2	3.425	2.848	2.194	1.678	3.278
x_3	4.706	3.556	3.258	2.642	3.562
x_1, x_2	2.721	3.392	**-0.048**	**-0.909**	**1.378**
x_1, x_3	3.124	**2.787**	1.325	0.672	2.007
x_2, x_3	3.553	3.480	2.522	1.769	3.221
x_1, x_2, x_3	**2.631**	3.172	0.133	-0.821	1.664

4.9 Summary

In this chapter, different variable selection criteria (AIC, Cp, and SIC) were considered to be used with high breakdown and bounded influence scale estimators. The influence function of the variable selection criteria for linear regression model based on the generalized scale approach was derived and discussed. The simulation study was carried out to examine the effect of vertical outliers and leverage points on the variable selection methods. The application on real data set presented, too. In general, robust variable selection criteria with M-estimation are robust in the presence of outliers in response variable (Y-direction). In the presence of high leverage point (X-direction), the value of these criteria will be affected and differs significantly from the true fit as the percentage of leverage point increases. But, the robust variable criteria with high breakdown scale estimate are less affected in all cases in the presence of outliers in X- and Y-directions.

CHAPTER 5

LASSO REGRESSION THROUGH GM- AND MM- LOSS FUNCTION

5.1 Introduction

Tibshirani (1996) suggested the application of $LASSO$ regression for variable selection and estimation in regression equation. According to the author, this method can be used when numerous correlated variables in a linear model. Applying this method also remarkably reduces computational cost for large data. To fix the problem of consistent, Zou (2006) modified $LASSO$ estimator to adaptive-$LASSO$, $\hat{\boldsymbol{\beta}}_{ada-LASSO}$ that is given in Eqn. (2.5).

Since the $LASSO$ estimate is a non-linear and non-differentiable function of the response values, it is difficult to obtain an accurate estimate of parameters. We may approximate the solution by a ridge regression of the form $\boldsymbol{\beta}^{\star} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{W}^T)^{-1}\mathbf{X}^T y$, where $\mathbf{W}$, is a diagonal matrix with diagonal elements $|\hat{\beta}_j|$, $\mathbf{W}^T$ denotes the generalized inverse of $\mathbf{W}$, and λ is chosen.

Among all unbiased linear estimators, LS method gives unbiased and minimum variance in the case that the errors are independent and distributed identically and normally with clean data sets. Nevertheless, LS and subsequently ada-$LASSO$ can produce very poor variable selection in the presence of outliers. That is why, a robust variable selection techniques suitable for high-dimensional data sets, such as $LASSO$, have been consid-

ered remarkably.

In this regard, Fan and Li (2001), Efron et al. (2004), and Owen (2007) proposed modification of $LASSO$ in order to be less sensitive to the presence of outliers. Moreover, the $LASSO$ is addressed through the least absolute deviation (LAD-$LASSO$) estimator, that is robust to outlier in response direction. This technique initially was proposed by Wang and Jiang (2007) and later Lambert-Lacroix & Zwald (2011) extended it with M-estimator. The author suggested to compute $LASSO$ regression using Huber function and named this as Huber-$LASSO$.

However, none of these robust $LASSO$ achieves 'boundedness' in the X-direction or high breakdown point estimators. Furthermore, it is widely known that leverage points resulted from bad data points in X-direction; have key effects on the M-estimators. At present the robust $LASSO$ with respect to leverage points receive less attention in literature. Recently, Alfons, Croux, & Gelper (2013) suggested combining $LASSO$ and the well known least trimmed squares (LTS) estimator known as $sparseLTS$ to obtained reliable estimates and variable selection especially in the presence of the laverage point. The LTS has shown good performances in robust regression estimation.

This chapter combines $LASSO$ regression with GM-estimator for errors in variables regression and MM-estimator for high efficiency and high breakdown point estimators. We called these modified methods of the robust $LASSO$ regression based on GM- and MM-estimators as GM-$LASSO$ and MM-$LASSO$, respectively. It was expected that the modified methods is less sensitive to outliers and possess a high breakdown point since the influence of outliers is insensitive to highly robust and efficient GM-and MM-

estimators. A simulation study is carried out to perform subset selection of the variables and to investigate the power performance of the GM-$LASSO$ and MM-$LASSO$.

5.2 Generalized M-estimators (GM) for Linear Regression Model

The general $\hat{\boldsymbol{\beta}}_{GM}$ Krasker and Welsch (1982) is defined as the value $\hat{\boldsymbol{\beta}}$, which solves:

$$\sum w_i(\mathbf{X}_i)\psi\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{v_i(\mathbf{X}_i)\hat{\sigma}}\right)\mathbf{X}_i = 0, \tag{5.1}$$

where, $w_i(\mathbf{X}_i)$ is a weight function, $v_i(\mathbf{X}_i)$ initially depends on the model matrix $\mathbf{X}$ from the initial LS regression to the data, and ψ is a bounded function as in the case of M-estimation.

The w_i and v_i are calculated from the hat values ($H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$) . Because hat values range from 0 to 1, a weights of $w_i = v_i = \sqrt{1-h_{ii}}$ which ensure that observations with high leverage receive less weight than observations with small leverage. Also, v_i is adjusted according to the size of the residual. However, the breakdown point for GM-estimators, $BP(\hat{\boldsymbol{\beta}}_{GM}) = 1/(p+1)$, is better than M-estimators, $BP(\hat{\boldsymbol{\beta}}_{M}) = 1/p$ which ignores leverage points. In other words, good leverage points that fall in line with the pattern in the bulk of the data by down-weights, results in a loss efficiency [see, Leroy and Rousseeuw (1987)].

If both, w and v are equal 1, then the M-estimation has been obtained, which have unbounded influence functions. With $v = 1$, and $w_i = \sqrt{1-h_{ii}}$, Mallows-type GM-estimate (Mallows, 1975) is obtained, see (Hill, 1977).

5.3 MM-estimators for Linear Regression Model

MM-estimators have been proposed by Yohai (1987) to improve the efficiency of the high breakdown estimators and this a three-stage procedure. Starting with $\hat{\boldsymbol{\beta}}_0$, a high-breakdown point "initial" estimate for $\boldsymbol{\beta}$, the second stage, a robust M-estimate of scale $\hat{\sigma}$ of the residuals, is then computed based on $\hat{\boldsymbol{\beta}}_0$ satisfying:

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{y_i-\mathbf{X}_i^T\hat{\boldsymbol{\beta}}_0}{\hat{\sigma}}\right)=b. \tag{5.2}$$

In the third stage, an M-estimate $\hat{\boldsymbol{\beta}}$ using IRLS procedure (see Eqn. (2.11)) has been identified starting at $\hat{\boldsymbol{\beta}}_0$. Then, the regression MM-estimation is defined as the solution of:

$$\sum_{i=1}^{n}\rho_1'\left(\frac{y_i-\mathbf{X}_i^T\hat{\boldsymbol{\beta}}_0}{\hat{\sigma}}\right)\mathbf{X}_i=0. \tag{5.3}$$

The ρ_1' is a function satisfying the M-estimation assumptions (Huber (1973)), such that $\rho_1(u)\leq\rho_0(u)$, and $\sup\rho_1(u)=\sup\rho_0(u)=a$. The MM- estimators posses simultaneous properties of high efficiency when the errors are normal distributed and that BP is 0.5. Next we provide a brief description of the influence functions of GM and MM-estimators.

5.4 The Influence Function of GM and MM Estimates

5.4.1 Definition the Influence Function of GM-Estimate

The influence function of the GM-estimate T at the distribution F and at the distribution $\Delta_{(\mathbf{X}_i,y_i)}$, which $(\mathbf{X}_i,y_i)$ contains outliers, has the following form,

$$IF(\mathbf{X}_i,y_i,T,F)=V^{-1}\left(\psi,F\right)\mathbf{X}_i w_i\mathbf{X}_i^T\psi\left(y_i-\mathbf{X}_i^T T(F)\right), \tag{5.4}$$

where $\psi = \rho'$ bounded function and V is as in Eqn. (2.38). The right side in Eqn. (5.4) is similar to the influence function for the M-estimator, but, $\mathbf{X}_i\mathbf{X}_i^T$ is replaced by the weighted matrix $\mathbf{X}_i w_i \mathbf{X}_i^T$, which may down weight large leverage points.

5.4.2 Definition the Influence Function of MM-Estimate

The influence function of the MM-estimate T at the distribution F and at the contaminated distribution $\Delta_{(\mathbf{X}_i,y_i)}$, which $(\mathbf{X}_i, y_i)$ contains outliers, has the following form,

$$IF(\mathbf{X}_i, y_i, T, F) = V^{-1}(\psi, F)\,\mathbf{X}_i\mathbf{X}_i^T\psi_1\left(y_i - \mathbf{X}_i^T T(F)\right), \tag{5.5}$$

where $\psi = \rho'$ bounded function and V is as in Eqn. (2.38). However, the MM-estimator T is defined as any solution of $\sum_{i=1}^{n}\psi_1\left[\frac{r_i(\theta)}{\sigma_n}\right]\mathbf{X}_i = 0$, and which must also satisfy, $S(T_1) \leq S(T_0)$ where, $S(\theta) = \sum_{i=1}^{n}\rho\left(\frac{r_i(\theta)}{\sigma_n}\right)$ and $\rho_1(\frac{0}{0}) = 0$.

In the present study, a slight modification of the robust $LASSO$ regression technique based on GM and MM estimators is proposed. The modified method is expected to be more robust than the LAD-$LASSO$ and the Huber-$LASSO$. The next section elaborates this idea.

5.5 $LASSO$ Regression Through GM- and MM- Loss Functions

In the present study, a slight modification of the Huber-$LASSO$ defined in Eqn. (2.41) is proposed. The GM- estimator of $\boldsymbol{\beta}$ is used instead of M-estimator in computing the loss function in order to reduce the effect of outliers in both the y and X-direction. Likewise, the MM-estimator of $\boldsymbol{\beta}$ is used instead of M-estimator in Eqn. (2.41) to compute the loss

function. However, the GM-$LASSO$ and MM-$LASSO$ estimator are given by:

$$\hat{\boldsymbol{\beta}}_{GM-LASSO} = \arg\min \left[\sum_{i=1}^{n} w_i(\mathbf{X}_i)\rho \left(\frac{y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}}{\hat{\sigma}} \right) + \lambda \sum_{j=1}^{p} \hat{W}_j |\beta_j| \right]. \tag{5.6}$$

$$\hat{\boldsymbol{\beta}}_{MM-LASSO} = \arg\min \left[\boldsymbol{\beta}_{MM} + \lambda \sum_{j=1}^{p} \hat{W}_j |\beta_j| \right]. \tag{5.7}$$

It can be see that, GM-$LASSO$ and MM-$LASSO$ combine the GM and MM, and $LASSO$ penalty, and hence the resulting estimators are expected to be robust against leverage points and also enjoy sparse representation (variable selection).

5.6 The Estimation Procedure for The GM-$LASSO$

The procedure for the GM-$LASSO$ regression method is as follows:

1. First, compute the weights $w_i(\mathbf{X}_i)$. For this step, we can use the definition of the Mallows-type GM estimate provided in the literature (Mallows, 1975), that is $w_i = \sqrt{1 - h_{ii}}$. The observations for which $h_{ii} > 2p/n$ can be identified as the leverage points. We can then compute positive weights w_i through the relation $w_i = 1$ or, $w_i = \sqrt{1 - h_{ii}}$ given that w_i decrease as h_{ii} increases; the leverage points are then assigned smaller weights.

2. Find the initial $\boldsymbol{\beta}^{(0)}_{GM-LASSO}$, such as the Huber-$LASSO$ estimates.

3. At each iteration t, find $\tilde{\mathbf{X}}^{(t-1)} = w_i \mathbf{X}^{(t-1)}$ from the previous iteration.

4. For a fixed penalty parameter λ, establish that following:

$$\hat{\boldsymbol{\beta}}^{(t)}_{GM-LASSO} = \sum_{i=1}^{n} w_i(\mathbf{X})\rho_M \left(\frac{y_i - \mathbf{X}_i^T \boldsymbol{\beta}}{\sigma} \right) + \lambda \sum_{j=1}^{p} \hat{W}_j |\beta_j| =$$

$$\sum_{i=1}^{n} \rho_M \left(\frac{y_i - \tilde{\mathbf{X}}_i^{\,T} \boldsymbol{\beta}}{\sigma} \right) + \lambda \sum_{j=1}^{p} \hat{W}_j |\beta_j|. \tag{5.8}$$

This equation is the Huber-$LASSO$ criterion for the data set $(y_i, \tilde{\mathbf{X}}_i)$, which can be solved by using either the available Huber-$LASSO$ program (Lambert-Lacroix and Zwald, 2011) or the MATLAB code *CVX* (Grant et al., 2008). We can also use nodewise Huber-$ada\ LASSO$, the program performed in R by Saharon Rosset, Ji Zhu (2003), which is available from

http://people.brunel.ac.uk/ mastvvv/Software/AdaptiveNodewiseGraph.R,

and

http://dept.stat.lsa.umich.edu/ jizhu/code/piecewise/robust/huber.r

5. Steps 2 and 3 above are repeated until the estimated coefficients converge.

5.7 Theoretical Discussion

5.7.1 Asymptotic Normality of GM-$LASSO$ and MM-$LASSO$ Regression

Arslan (2012) proposed the weighted LAD-$LASSO$ estimators ($WLAD$-$LASSO$), which is considered as an a special case of GM-$LASSO$ estimator. In addition MM-$LASSO$ estimator are M-$LASSO$ when $v_i = 1$, $w_i = 1$. Following Arslan (2012) and Lambert-Lacroix and Zwald (2011) we can show that:

Result: Any minimized $(\hat{\mu}, \hat{\boldsymbol{\beta}})$ of Eqns. (5.6) and (5.7) then satisfies the following oracle properties:

1. Consistency in variable selection.

2. Asymptotic normality.

The proof of the above result is adopted with the proof of Lambert-Lacroix and Zwald (2011), Theorem 3.2 .

5.7.2 The Sensitivity Curve (Measuring the Effect of an Outliers)

Consider good data set: $x_1, ..., x_{n-1}$, the particular estimator: $T_{n-1} = T(x_1, ..., x_{n-1})$. Let the contaminated data set: $x_1, ..., x_{n-1}, \mathbf{x}$, and $T_n = T(x_1, ..., x_{n-1}, \mathbf{x})$. The sensitivity curve is defined as follows:

$$SC(\mathbf{x}) = n(T_n - T_{n-1}). \tag{5.9}$$

Since T is the $GM\text{-}LASSO$ or $MM\text{-}LASSO$ estimators with bounded function, then theirs SC is bounded.

5.8 Choice of the Tuning Parameter

The suitable value of the shrinkage tuning parameter λ is not known in advance. Tibshirani (1996) proposed to select λ by estimating prediction performance through cross-validation. The procedure for this is as follows:

1. Divide the data into roughly k equal parts, $K_0, K_1, ..., K_k$.
2. For each K, find the best subsets with the tuning parameter λ, to the other $K - 1$ parts, giving $\hat{\boldsymbol{\beta}}^{-k}(\lambda)$.
3. Compute the prediction error, $PE(\lambda) = \sum_{i \in K} \left(y_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}^{-k}(\lambda) \right)^2$.
4. This gives the cross-validation error, $CV(\lambda) = \frac{1}{K} \sum_{k=1}^{K} PE(\lambda)$.
5. Do this for several λ and choose the value of λ that makes $CV(\lambda)$ the smallest.

Here we propose to select λ by optimizing the Bayes Information Criterion (BIC). The BIC of a given model estimated with tuning parameter λ is given by

$$BIC(\lambda) = log(\hat{\sigma}) + p \cdot df(\lambda)\frac{log(n)}{n}, \tag{5.10}$$

where $\hat{\sigma}$ is the corresponding residual scale estimate, and $df(\lambda)$ is the degree of freedom of the model, given by the number of non-zero estimated parameters in $\hat{\boldsymbol{\beta}}$. The selecting λ then minimizes $BIC(\lambda)$ or $CV(\lambda)$ over a grid of values in the interval $[0, \lambda_0]$.

5.9 Simulation Study

This simulation was conducted to show the performance of the proposed estimators of $GM\text{-}LASSO$ and $MM\text{-}LASSO$. Furthermore, we compared these methods with already established ones. The following estimators were used in this study: (1) $LASSO$, (2) $ada\text{-}LASSO$, (3) $LAD\text{-}LASSO$, (4) Huber-$LASSO$, (5) $GM\text{-}LASSO$, and (6) $MM\text{-}LASSO$.

The simulation based on the linear model is given in Eqn. (1.1). For $LAD\text{-}LASSO$, Huber-$LASSO$, $GM\text{-}LASSO$ and $MM\text{-}LASSO$, the penalty parameters were chosen by applying the BIC. For $LASSO$ and $ada\ LASSO$, we estimated λ using 10-fold cross validation. The correspondent data set $n = 200$ observations. The parameters were $\boldsymbol{\beta}_{true} = (3, 1.5, 0, 2, 0, 0, 0, 0)$, and $\varepsilon_i \sim N(0, 1)$. The correlation between the ith and jth vector was demonstrated as follows:

$$corr(i, j) = 0.5^{|i-j|}, \forall i, j \in \{1, 2, ..., 8\}. \tag{5.11}$$

To investigate the robustness of the methods against outliers and leverage points, the fol-

lowing points were considered:

1. No contamination,

2. Vertical contamination (outliers on the response variables),

3. Bad Leverage points (outliers on the covariates),

4. Good Leverage points (outliers in **X** follow the pattern of the majority of the data).

For the vertical outliers, different percentages ($c\%$= 5%, 10%, and 20%) of the error terms in the regression model follow the normal $N(20, 1)$, instead of a $N(0, 1)$. However, for bad leverage points with the same $c\%$, contaminated variables X_1 and X_2 by generating the predictor variables from a $N(50, 1)$ distribution, instead of a $N(0.5, 1)$. For good leverage case, we considered the different percentages of outliers c$\%$ on the variables X_1 and X_2, that were generated from a $N(50, 1)$ distribution, then generated y to get good leverage points. The simulations were performed in R. To run the simulations, the package **parcor** (Kraemer and Schaefer, 2010) was used for $LASSO$ and ada-$LASSO$, and the package **quantreg** (Koenker, 2007) was used for LAD-$LASSO$.

The performance of the proposed methods was then determined by assessing summary statistics based on 100 Monte Carlo trials. The statistics computed the sample mean (Mean), $\bar{\boldsymbol{\beta}}$ of the parameters, $\hat{\boldsymbol{\beta}}$ as follows:

$$\bar{\beta}_j = \frac{\sum_{j=1}^{m} \hat{\beta}_j}{m}, j = 1, 2, ..., 8, \tag{5.12}$$

where $\hat{\beta}_j$ is the parameter obtained for $j = 1, ..., m$; with m as the number of simulations.

The median of Standard error (MSE) of the parameters, $\hat{\boldsymbol{\beta}}$ given by

$$MSE(\hat{\beta}_j) = median\left(\sqrt{\frac{\sum_{j=1}^{m}(\hat{\beta}_j - \bar{\beta}_j)^2}{m-1}}\right). \tag{5.13}$$

The median of Relative Prediction Errors $MRPE$ of parameter $\hat{\boldsymbol{\beta}}_j$ is given by

$$MRPE(\hat{\beta}_j) = median\left((y_i - \mathbf{X}_i^T \hat{\beta}_{j_m})^2\right). \tag{5.14}$$

The sample standard deviation, std of parameter $\hat{\boldsymbol{\beta}}_j$ given by

$$std(\hat{\beta}_j) = \sqrt{\frac{\sum_{j=1}^{m}(\hat{\beta}_j - \bar{\beta})^2}{m}}. \tag{5.15}$$

In addition, the median number of zero coefficients is reported. To evaluate the accuracy of the coefficient estimation, a boxplots of estimated parameters over the simulations is presented.

Discussion and Result

The simulation results are represented in Tables 5.1 to 5.10 for each situation. In order to provide the indicators defined below, a coefficient is considered to be zero if it absolute value is strictly less than 0.01. Furthermore, the MSE, $MRPE$, and std should be relatively small. The following results were observed in this study.

1. For the outlier-free data set, as shown in Table 5.1, both non-robust and robust methods performed well in model selection (Figure (5.1)). Although LAD-$LASSO$ showed excellent results with small values of MSE and $MRPE$, the median number of the zero coefficients (3 zero coefficients) revealed that we could over fit the subset selection. ada-$LASSO$, Huber-$LASSO$, GM-$LASSO$ and MM-$LASSO$

selected the correct number of zero coefficients (5 zero coefficients). Indeed, the model selection ability of Huber-$LASSO$ was close to that of GM-$LASSO$. GM-$LASSO$ was more stable than MM-$LASSO$ in this procedure, for example, the $std = 0$ of GM-$LASSO$ for $\hat{\beta}_3$, and the $std = 0.0271$ of GM-$LASSO$ for $\hat{\beta}_3$.

2. By introducing vertical outliers, the non-robust methods ($LASSO$ and ada-$LASSO$) showed poor results compared with the robust methods. In this case, the values of MSE, $RMSE$, and std for parameters were generally large than other. From the perspective of error and model selection ability, LAD-$LASSO$ obtained better results than non-robust methods (Tables 5.2 to 5.4 and Figures (5.2) to (5.4)). Furthermore, the Huber-$LASSO$ and GM-$LASSO$ performed similar in terms of summary statistics and model selection ability (both of them choose 5 zero coefficients). The results also verified that MM-$LASSO$ could correctly identify the three significant variables (1,2, and 4) and zero variables (3,5,6,7, and 8). Remarkably, by increasing the percentage of contamination to 20%, MM-$LASSO$ gave better results, whereas Huber-$LASSO$ and GM-$LASSO$ selected fewer variables in the final model.

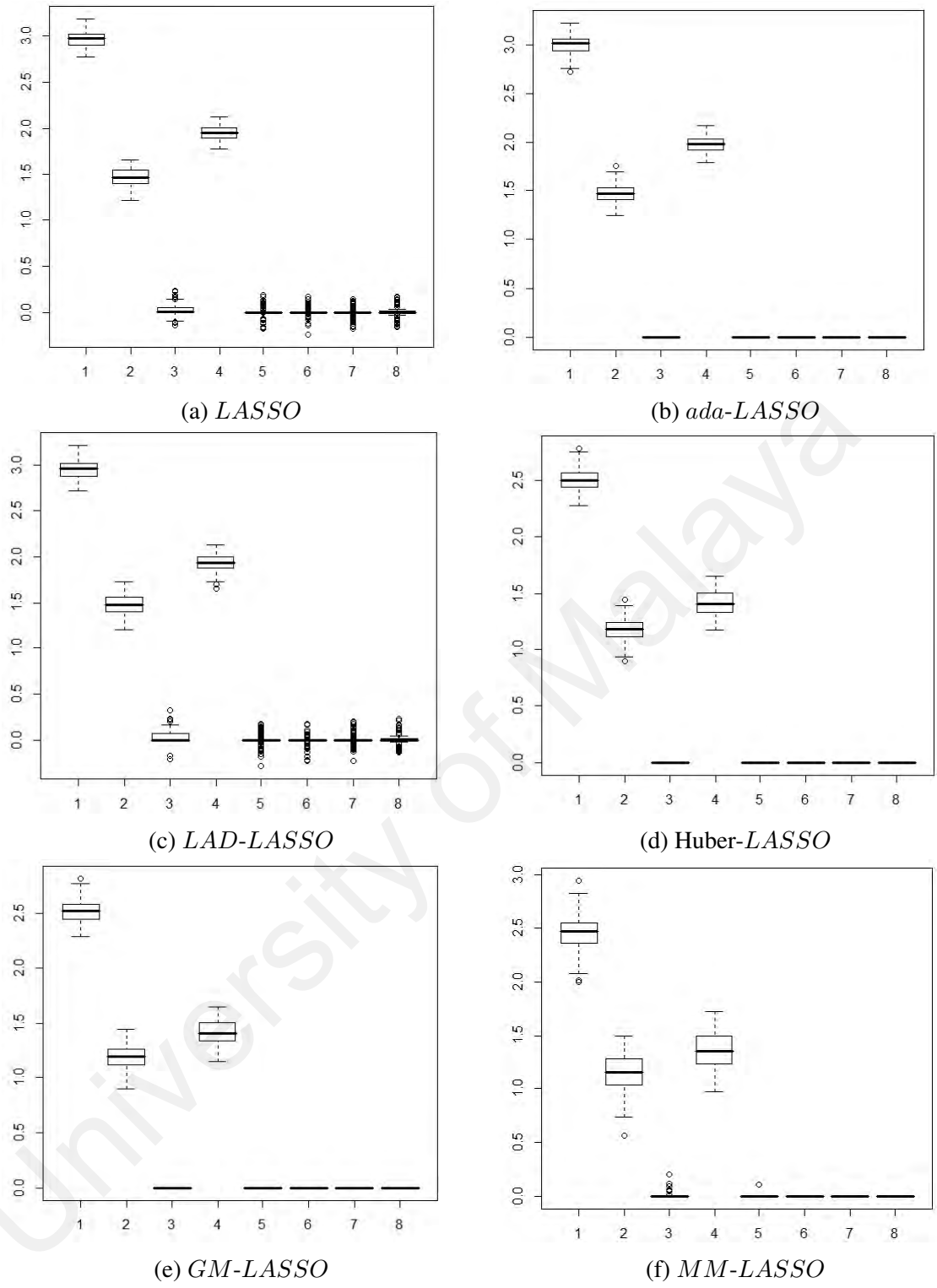
3. With a low percentage of leverage points, GM-$LASSO$ and MM-$LASSO$ obtained better results than Huber-$LASSO$ and LAD-$LASSO$ from MSE and $RMSE$ (Tables 5.5 to 5.10 and Figures (5.5) to (5.10)). GM-$LASSO$ selected numerous noise variables as the percentage of contamination increased, whereas $LASSO$, ada-$LASSO$, and LAD-$LASSO$ performed unsatisfactorily in selection variables when data suffered from leverage points.

In conclusion, the results revealed that the classical (non-robust) $LASSO$ worked well for data sets with only a few outlying observations. However, applying robust $LASSO$ based on M-estimators is not suggested for data sets with high contamination levels of

outliers because the method is affected by the outliers presence in the data. Such an effect worsens when a higher percentage of contaminated observations is evident in the data. Therefore, in this situation, using $LAD\text{-}LASSO$ and Huber- $LASSO$ was suitable compared with $LASSO$ or $ada\text{-}LASSO$. Moreover, when leverage points were introduced, $GM\text{-}LASSO$ and $MM\text{-}LASSO$ showed the best overall performance in model selection ability.

Table 5.1: The estimation of parameters for simulated data sets when no contaminated data

Coefficients	True Values	Mean	Median (MSE)	Median ($MRPE$)	std	Mean	Median (MSE)	Median ($MRPE$)	std
		$LASSO$				ada-$LASSO$			
$\hat{\beta}_0$	0	0.0668	0.0910	0.0105	0.0720	0.0646	0.0104	0.0014	0.0718
$\hat{\beta}_1$	3	2.9818	0.02382	0.0090	0.0813	2.9973	0.0087	0.0012	0.0810
$\hat{\beta}_2$	1.5	1.5798	0.0456	0.0156	0.0957	1.5752	0.0140	0.0041	0.0871
$\hat{\beta}_3$	0	0.0000	0.0985	0.0079	0.0663	0.0000	0.0016	0.0039	0.0151
$\hat{\beta}_4$	2	1.8910	0.0345	0.0098	0.0698	1.8970	0.0115	0.0034	0.0640
$\hat{\beta}_5$	0	0.0000	0.0291	0.0066	0.0619	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_6$	0	0.0000	0.02391	0.0058	0.0538	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_7$	0	0.0000	0.0341	0.0058	0.0541	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_8$	0	0.0363	0.00344	0.0074	0.0627	0.0000	0.0000	0.0000	0.0000
median NO. of Zero coefficients		**3**					**5**		
		LAD-$LASSO$				Huber-$LASSO$			
Coefficients	True Values	Mean	Median (MSE)	Median ($MRPE$)	std	Mean	Median (MSE)	Median ($MRPE$)	std
$\hat{\beta}_0$	0	0.1069	0.0151	0.0015	0.0882	0.2453	0.0282	0.0082	0.1039
$\hat{\beta}_1$	3	2.8970	0.0125	0.0025	0.1059	2.4220	0.0141	0.0041	0.1057
$\hat{\beta}_2$	1.5	1.5389	0.0137	0.0089	0.1125	1.2182	0.0123	0.0021	0.1078
$\hat{\beta}_3$	0	0.0251	0.0075	0.0012	0.0692	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_4$	2	1.9250	0.0100	0.0025	0.0934	1.2067	0.0254	0.0012	0.1160
$\hat{\beta}_5$	0	0.0000	0.0072	0.0014	0.0673	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_6$	0	0.0435	0.0078	0.0024	0.0548	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_7$	0	0.0000	0.0074	0.0012	0.0684	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_8$	0	0.0478	0.0084	0.0023	0.0675	0.0000	0.0000	0.00000	0.0000
median NO. of Zero coefficients		**3**					**5**		
		GM-$LASSO$				MM-$LASSO$			
Coefficients	True Values	Mean	Median (MSE)	Median ($MRPE$)	std	Mean	Median (MSE)	Median ($MRPE$)	std
$\hat{\beta}_0$	0	0.2420	0.0278	0.1030	0.0010	0.1429	0.0176	0.0017	0.0861
$\hat{\beta}_1$	3	2.4326	0.0147	0.1099	0.0010	2.2069	0.0325	0.0032	0.1629
$\hat{\beta}_2$	1.5	1.2232	0.0125	0.1112	0.0012	1.3597	0.0290	0.0029	0.1803
$\hat{\beta}_3$	0	0.0000	0.0000	0.0000	0.0000	0.00000	0.0030	0.0030	0.0271
$\hat{\beta}_4$	2	1.2102	0.0258	0.1173	0.0017	1.2140	0.0239	0.0023	0.1687
$\hat{\beta}_5$	0	0.0000	0.0000	0.0000	0.0000	0.00000	0.0011	0.0001	0.0104
$\hat{\beta}_6$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.00000	0.0000
$\hat{\beta}_7$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.00000	0.0000
$\hat{\beta}_8$	0	0.0000	0.0000	0.0000	0.0000	0.00000	0.0000	0.00000	0.0000
median NO. of Zero coefficients		**5**					**5**		



(a) $LASSO$ (b) $ada\text{-}LASSO$

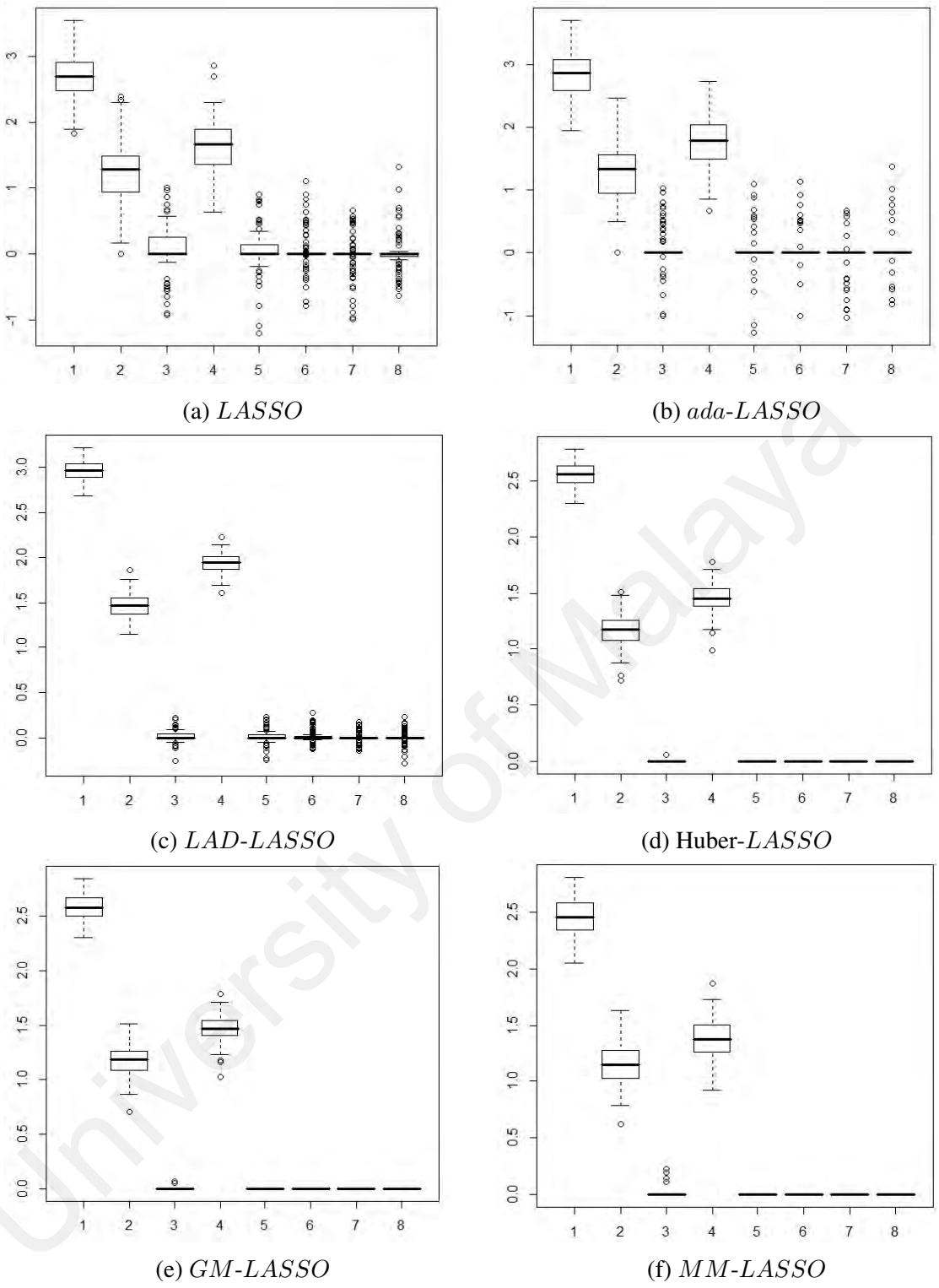
(c) $LAD\text{-}LASSO$ (d) Huber-$LASSO$

(e) $GM\text{-}LASSO$ (f) $MM\text{-}LASSO$

Figure 5.1: Boxplots of estimates for the eight coefficients from 100 simulated data sets, no contaminated data

Table 5.2: The estimation of parameters for simulated data sets when vertical

Coefficients	True Values	$LASSO$ Mean	$LASSO$ Median (MSE)	$LASSO$ Median ($MRPE$)	$LASSO$ std	ada-$LASSO$ Mean	ada-$LASSO$ Median (MSE)	ada-$LASSO$ Median ($MRPE$)	ada-$LASSO$ std
$\hat{\beta}_0$	0	0.9710	0.0136	0.0097	0.1242	0.9130	0.0164	0.0091	0.1219
$\hat{\beta}_1$	3	2.5939	0.0396	0.0041	0.3553	2.6539	0.0448	0.0035	0.3714
$\hat{\beta}_2$	1.5	1.9158	0.0859	0.0042	0.4707	1.7502	0.0762	0.0025	0.5273
$\hat{\beta}_3$	0	-0.5229	0.0745	0.0052	0.3348	-0.0688	0.0344	0.0007	0.2996
$\hat{\beta}_4$	2	1.7960	0.0458	0.0020	0.4014	1.6091	0.0465	0.0039	0.4032
$\hat{\beta}_5$	0	0.0136	0.0342	0.0001	0.3173	0.0000	0.0322	0.0000	0.2991
$\hat{\beta}_6$	0	0.0000	0.0294	0.0000	0.2707	0.0000	0.0293	0.0000	0.2705
$\hat{\beta}_7$	0	0.0000	0.0297	0.0000	0.2755	0.0000	0.0279	0.0000	0.2577
$\hat{\beta}_8$	0	-0.1656	0.0344	0.0017	0.2689	0.0000	0.0288	0.0000	0.2686
median NO. of Zero coefficients		**3**				**5**			
Coefficients	True Values	LAD-$LASSO$ Mean	LAD-$LASSO$ Median (MSE)	LAD-$LASSO$ Median ($MRPE$)	LAD-$LASSO$ std	Huber-$LASSO$ Mean	Huber-$LASSO$ Median (MSE)	Huber-$LASSO$ Median ($MRPE$)	Huber-$LASSO$ std
$\hat{\beta}_0$	0	-0.0765	0.0182	0.0008	0.0893	0.0847	0.0114	0.0008	0.1043
$\hat{\beta}_1$	3	2.8406	0.0177	0.0016	0.1166	2.5249	0.0119	0.0048	0.1064
$\hat{\beta}_2$	1.5	1.6024	0.0202	0.0010	0.1325	1.4046	0.0296	0.0010	0.1447
$\hat{\beta}_3$	0	0.0000	0.0071	0.0000	0.0638	0.0000	0.0010	0.0000	0.0095
$\hat{\beta}_4$	2	2.0169	0.0143	0.0002	0.1117	1.5156	0.0159	0.0048	0.1354
$\hat{\beta}_5$	0	0.0000	0.0079	0.0000	0.0722	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_6$	0	-0.0056	0.0068	0.0001	0.0618	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_7$	0	-0.0336	0.0070	0.0003	0.0531	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_8$	0	0.0000	0.0070	0.0000	0.0655	0.0000	0.0000	0.0000	0.0000
median NO. of Zero coefficients		**3**				**5**			
Coefficients	True Values	GM-$LASSO$ Mean	GM-$LASSO$ Median (MSE)	GM-$LASSO$ Median ($MRPE$)	GM-$LASSO$ std	MM-$LASSO$ Mean	MM-$LASSO$ Median (MSE)	MM-$LASSO$ Median ($MRPE$)	MM-$LASSO$ std
$\hat{\beta}_0$	0	0.0909	0.0113	0.0009	0.1041	0.1349	0.0165	0.0013	0.0828
$\hat{\beta}_1$	3	2.5172	0.0138	0.0048	0.1155	2.3412	0.0222	0.0066	0.1685
$\hat{\beta}_2$	1.5	1.4470	0.0331	0.0005	0.1478	1.1695	0.0203	0.0033	0.1882
$\hat{\beta}_3$	0	0.0000	0.0011	0.0000	0.0105	0.0000	0.0037	0.0000	0.0339
$\hat{\beta}_4$	2	1.5192	0.0158	0.0048	0.1360	1.5267	0.0250	0.0047	0.1784
$\hat{\beta}_5$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_6$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_7$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_8$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
median NO. of Zero coefficients		**5**				**5**			



(a) $LASSO$ (b) $ada\text{-}LASSO$

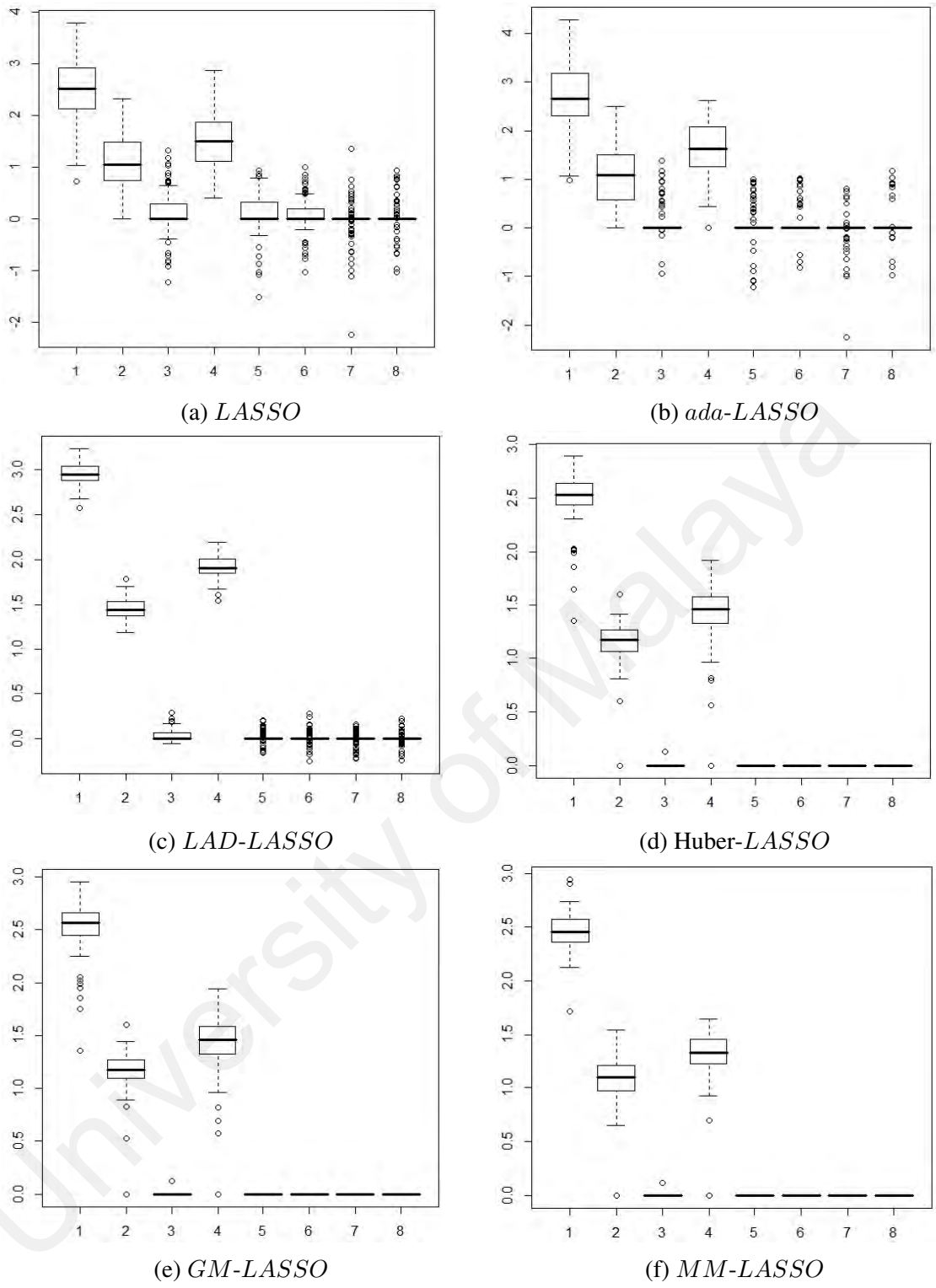
(c) $LAD\text{-}LASSO$ (d) Huber-$LASSO$

(e) $GM\text{-}LASSO$ (f) $MM\text{-}LASSO$

Figure 5.2: Boxplots of estimates for the eight coefficients from 100 simulated data sets, 5% vertical

Table 5.3: The estimation of parameters for simulated data sets when 10% vertical

Coefficients	True Values	$LASSO$ Mean	$LASSO$ Median (MSE)	$LASSO$ Median $(MRPE)$	$LASSO$ std	$ada\text{-}LASSO$ Mean	$ada\text{-}LASSO$ Median (MSE)	$ada\text{-}LASSO$ Median $(MRPE)$	$ada\text{-}LASSO$ std
$\hat{\beta}_0$	0	2.4561	0.0504	0.0246	0.1536	2.4762	0.0524	0.0248	0.1518
$\hat{\beta}_1$	3	1.9339	0.0861	0.0107	0.5756	2.2113	0.0829	0.0079	0.6090
$\hat{\beta}_2$	1.5	0.0000	0.1348	0.0150	0.5518	0.0000	0.1346	0.0150	0.6683
$\hat{\beta}_3$	0	0.0000	0.0451	0.0000	0.4051	0.0000	0.0381	0.0000	0.3390
$\hat{\beta}_4$	2	1.1166	0.0721	0.0088	0.5578	1.3094	0.0710	0.0069	0.5846
$\hat{\beta}_5$	0	0.8321	0.0896	0.0083	0.4049	1.0126	0.1112	0.0101	0.3593
$\hat{\beta}_6$	0	0.0000	0.0364	0.0000	0.3322	0.0000	0.0319	0.0000	0.2917
$\hat{\beta}_7$	0	0.4973	0.0703	0.0050	0.3948	0.7640	0.0938	0.0076	0.3353
$\hat{\beta}_8$	0	0.0000	0.0354	0.0000	0.3306	0.0000	0.0327	0.0000	0.3034
median NO. of Zero coefficients		**3**				**4**			
Coefficients	True Values	$LAD\text{-}LASSO$ Mean	$LAD\text{-}LASSO$ Median (MSE)	$LAD\text{-}LASSO$ Median $(MRPE)$	$LAD\text{-}LASSO$ std	Huber-$LASSO$ Mean	Huber-$LASSO$ Median (MSE)	Huber-$LASSO$ Median $(MRPE)$	Huber-$LASSO$ std
$\hat{\beta}_0$	0	0.2569	0.0159	0.0026	0.1036	0.6988	0.0507	0.0070	0.1278
$\hat{\beta}_1$	3	2.9656 0	.0138	0.0003	0.1281	1.9885	0.0603	0.0101	0.2340
$\hat{\beta}_2$	1.5	1.2152	0.0284	0.0028	0.1234	0.0000	0.1246	0.0150	0.2485
$\hat{\beta}_3$	0	0.1317	0.0122	0.0013	0.0673	0.0000	0.0014	0.0000	0.0130
$\hat{\beta}_4$	2	1.9445	0.0137	0.0006	0.1257	0.8185	0.0701	0.0118	0.2967
$\hat{\beta}_5$	0	0.0000	0.0069	0.0000	0.0638	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_6$	0	0.0000	0.0072	0.0000	0.0667	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_7$	0	0.0000	0.0068	0.0000	0.0634	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_8$	0	0.0000	0.0068	0.0000	0.0638	0.0000	0.0000	0.0000	0.0000
median NO. of Zero coefficients		**3**				**5**			
Coefficients	True Values	$GM\text{-}LASSO$ Mean	$GM\text{-}LASSO$ Median (MSE)	$GM\text{-}LASSO$ Median $(MRPE)$	$GM\text{-}LASSO$ std	$MM\text{-}LASSO$ Mean	$MM\text{-}LASSO$ Median (MSE)	$MM\text{-}LASSO$ Median $(MRPE)$	$MM\text{-}LASSO$ std
$\hat{\beta}_0$	0	0.6980	0.0507	0.0070	0.1294	0.0961	0.0143	0.0010	0.0950
$\hat{\beta}_1$	3	1.9943	0.0619	0.0101	0.2365	2.4587	0.0180	0.0054	0.1687
$\hat{\beta}_2$	1.5	0.0000	0.1259	0.0150	0.2523	0.7262	0.0451	0.0077	0.2152
$\hat{\beta}_3$	0	0.0000	0.0013	0.0000	0.0122	0.0000	0.0012	0.0000	0.0116
$\hat{\beta}_4$	2	0.8146	0.0718	0.0119	0.3027	1.1514	0.0296	0.0085	0.2182
$\hat{\beta}_5$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_6$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_7$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_8$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
median NO. of Zero coefficients		**5**				**5**			



(a) $LASSO$ (b) $ada\text{-}LASSO$

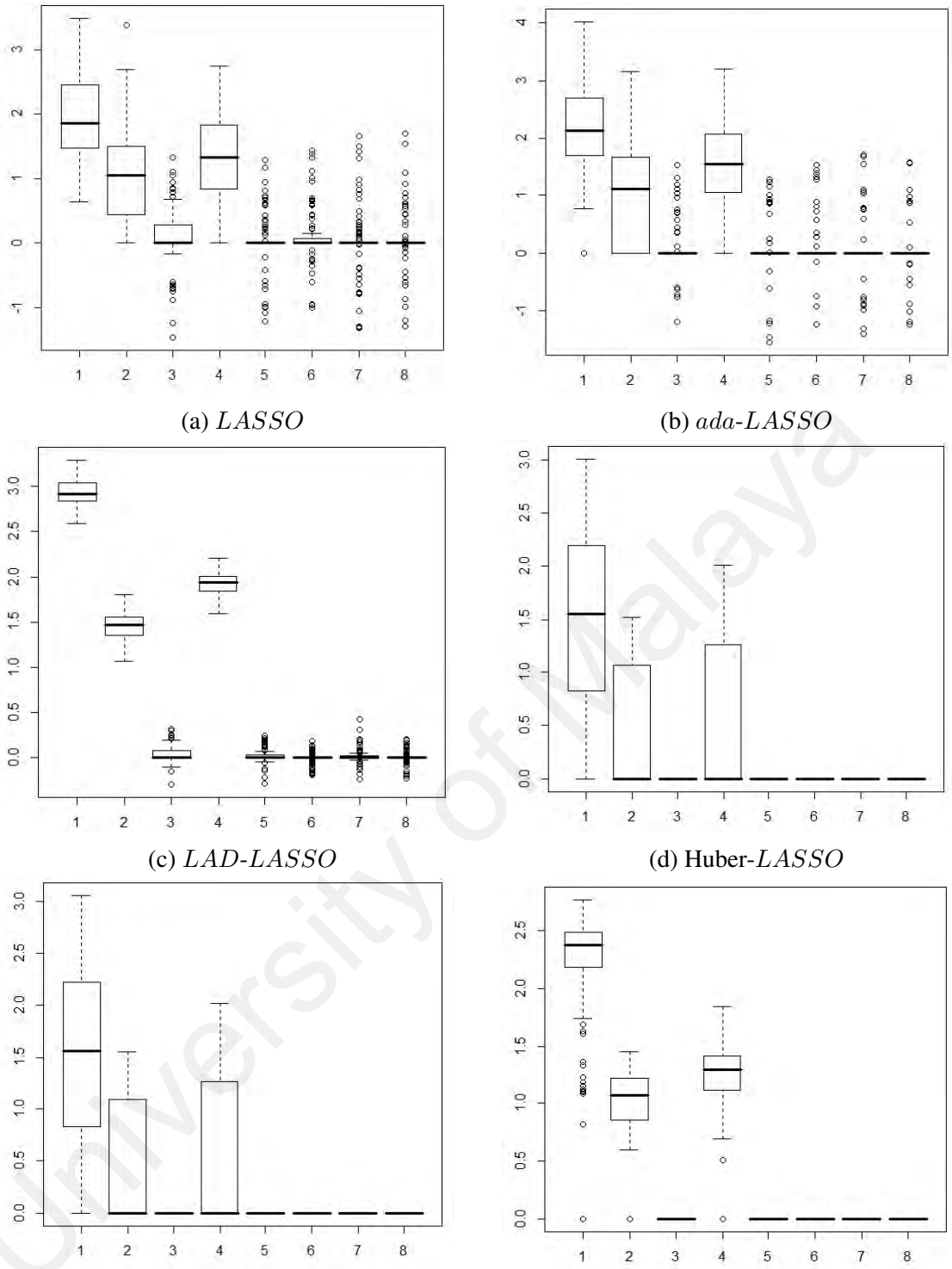
(c) $LAD\text{-}LASSO$ (d) Huber-$LASSO$

(e) $GM\text{-}LASSO$ (f) $MM\text{-}LASSO$

Figure 5.3: Boxplots of estimates for the eight coefficients from 100 simulated data sets, 10% vertical

Table 5.4: The estimation of parameters for simulated data sets when 20% vertical

Coefficients	True Values	$LASSO$ Mean	$LASSO$ Median (MSE)	$LASSO$ Median ($MRPE$)	$LASSO$ std	ada-$LASSO$ Mean	ada-$LASSO$ Median (MSE)	ada-$LASSO$ Median ($MRPE$)	ada-$LASSO$ std
$\hat{\beta}_0$	0	3.8845	0.0231	0.0388	0.1886	3.9024	0.0226	0.0390	0.1877
$\hat{\beta}_1$	3	1.7530	0.0711	0.0125	0.6390	2.2517	0.0835	0.0075	0.7752
$\hat{\beta}_2$	1.5	1.3426	0.0819	0.0016	0.7067	1.7045	0.1134	0.0020	0.8240
$\hat{\beta}_3$	0	0.0000	0.0467	0.0000	0.4272	0.0000	0.0409	0.0000	0.3750
$\hat{\beta}_4$	2	0.7641	0.0957	0.0124	0.6967	1.2475	0.0918	0.0075	0.8390
$\hat{\beta}_5$	0	0.0000	0.0416	0.0000	0.3877	0.0000	0.0456	0.0000	0.4242
$\hat{\beta}_6$	0	0.0000	0.0443	0.0000	0.4021	0.0000	0.0422	0.0000	0.3863
$\hat{\beta}_7$	0	0.0000	0.0519	0.0000	0.4841	0.0000	0.0507	0.0000	0.4714
$\hat{\beta}_8$	0	0.0000	0.0442	0.0000	0.4107	0.0000	0.0420	0.0000	0.3915
median NO. of Zero coefficients		**3**				**5**			
Coefficients	True Values	LAD-$LASSO$ Mean	LAD-$LASSO$ Median (MSE)	LAD-$LASSO$ Median ($MRPE$)	LAD-$LASSO$ std	Huber-$LASSO$ Mean	Huber-$LASSO$ Median (MSE)	Huber-$LASSO$ Median ($MRPE$)	Huber-$LASSO$ std
$\hat{\beta}_0$	0	0.3998	0.0121	0.0040	0.0991	0.5892	0.0783	0.0059	0.5495
$\hat{\beta}_1$	3	2.6947	0.0298	0.0031	0.1435	2.1719	0.1290	0.0083	0.9489
$\hat{\beta}_2$	1.5	1.5708	0.0197	0.0007	0.1515	1.3070	0.1062	0.0019	0.5723
$\hat{\beta}_3$	0	0.0000	0.0107	0.0000	0.0913	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_4$	2	1.9352	0.0137	0.0006	0.1278	1.3871	0.1199	0.0061	0.7037
$\hat{\beta}_5$	0	0.0108	0.0087	0.0001	0.0805	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_6$	0	0.0238	0.0071	0.0002	0.0606	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_7$	0	0.0000	0.0088	0.0000	0.0808	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_8$	0	0.0000	0.0068	0.0000	0.0632	0.0000	0.0000	0.0000	0.0000
median NO. of Zero coefficients		**3**				**7**			
Coefficients	True Values	GM-$LASSO$ Mean	GM-$LASSO$ Median (MSE)	GM-$LASSO$ Median ($MRPE$)	GM-$LASSO$ std	MM-$LASSO$ Mean	MM-$LASSO$ Median (MSE)	MM-$LASSO$ Median ($MRPE$)	MM-$LASSO$ std
$\hat{\beta}_0$	0	0.5925	0.0780	0.0059	0.5496	-0.0531	0.0255	0.0005	0.2265
$\hat{\beta}_1$	3	2.1694	0.1289	0.0083	0.9603	2.2441	0.0714	0.0076	0.6572
$\hat{\beta}_2$	1.5	1.3125	0.1066	0.0019	0.5791	1.2991	0.0645	0.0020	0.4602
$\hat{\beta}_3$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_4$	2	1.3833	0.1197	70.0062	0.7139	1.5651	0.0776	0.0043	0.5437
$\hat{\beta}_5$	0	0.0000	0.0000	70.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_6$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_7$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_8$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
median NO. of Zero coefficients		**7**				**5**			



(a) $LASSO$ (b) $ada\text{-}LASSO$

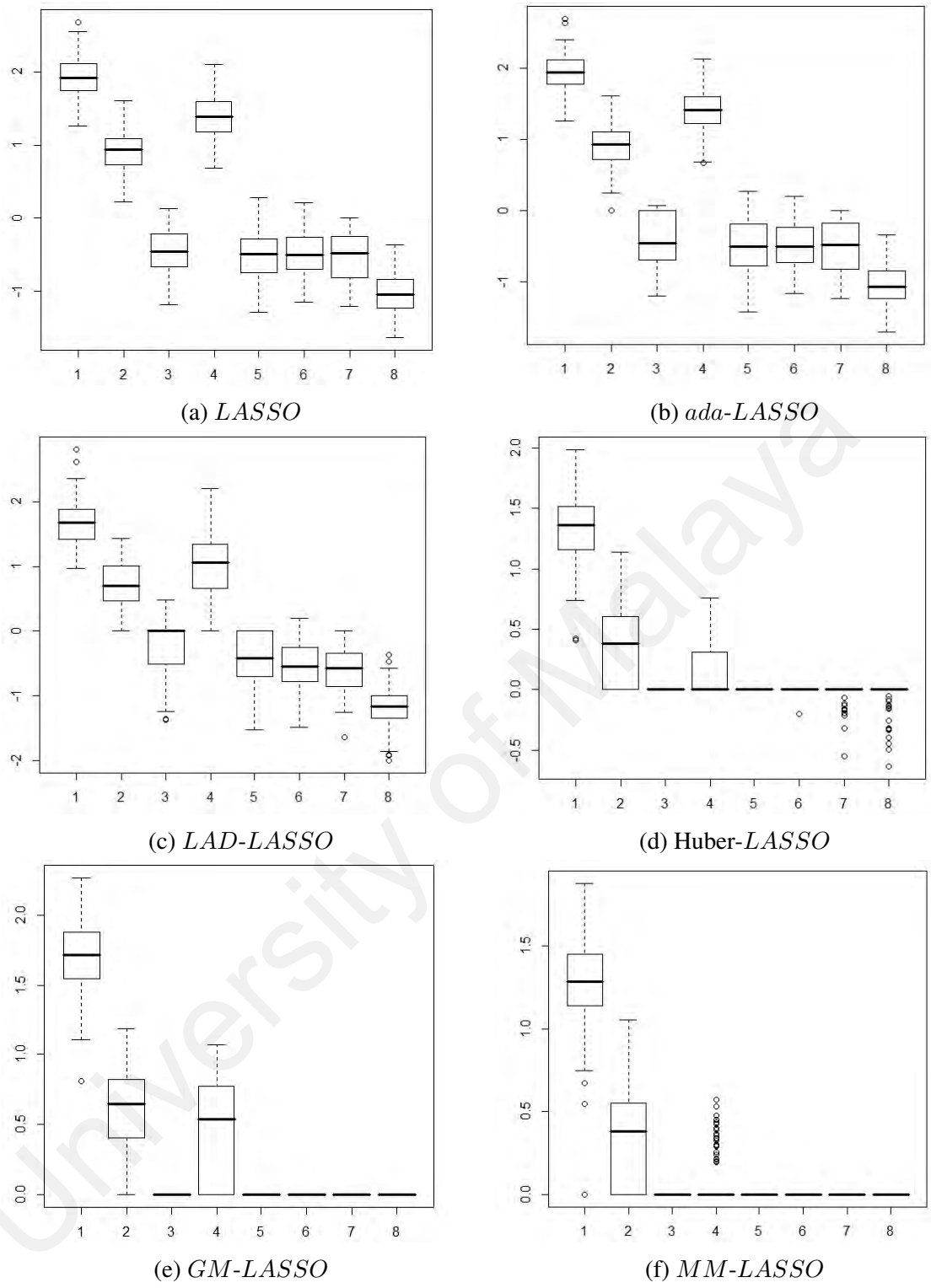
(c) $LAD\text{-}LASSO$ (d) Huber-$LASSO$

(e) $GM\text{-}LASSO$ (f) $MM\text{-}LASSO$

Figure 5.4: Boxplots of estimates for the eight coefficients from 100 simulated data sets, 20% vertical

Table 5.5: The estimation of parameters for simulated data sets when 5% bad leverage point

Coefficients	True Values	$LASSO$ Mean	$LASSO$ Median (MSE)	$LASSO$ Median ($MRPE$)	$LASSO$ std	ada-$LASSO$ Mean	ada-$LASSO$ Median (MSE)	ada-$LASSO$ Median ($MRPE$)	ada-$LASSO$ std
$\hat{\beta}_0$	0	-0.8225	0.0561	0.0082	0.1935	-0.8217	0.0557	0.0082	0.1935
$\hat{\beta}_1$	3	1.6722	0.0393	0.0133	0.2645	1.7192	0.0378	0.0128	0.2677
$\hat{\beta}_2$	1.5	1.1000	0.0350	0.0040	0.2724	1.1976	0.0467	0.0030	0.3146
$\hat{\beta}_3$	0	-0.5362	0.0341	0.0054	0.3030	-0.7132	0.0484	0.0071	0.3415
$\hat{\beta}_4$	2	1.0953	0.0481	0.0090	0.3115	1.2163	0.0419	0.0078	0.3341
$\hat{\beta}_5$	0	0.0000	0.0649	0.0000	0.3323	0.0000	0.0672	0.0000	0.3698
$\hat{\beta}_6$	0	-0.9945	0.0641	0.0099	0.3026	-1.0506	0.0703	0.0105	0.3285
$\hat{\beta}_7$	0	0.0000	0.0677	0.0000	0.3379	0.0000	0.0687	0.0000	0.3781
$\hat{\beta}_8$	0	-0.8386	0.0360	0.0084	0.2809	-0.8820	0.0362	0.0088	0.2941
median NO. of Zero coefficients		**0**				**0**			
Coefficients	True Values	LAD-$LASSO$ Mean	LAD-$LASSO$ Median (MSE)	LAD-$LASSO$ Median ($MRPE$)	LAD-$LASSO$ std	Huber-$LASSO$ Mean	Huber-$LASSO$ Median (MSE)	Huber-$LASSO$ Median ($MRPE$)	Huber-$LASSO$ std
$\hat{\beta}_0$	0	-0.2616	0.0222	0.0026	0.1165	-0.4791	0.0384	0.0048	0.2583
$\hat{\beta}_1$	3	2.4431	0.0238	0.0056	0.1653	1.1661	0.0373	0.0183	0.3028
$\hat{\beta}_2$	1.5	1.2639	0.0172	0.0024	0.1572	0.7773	0.0530	0.0072	0.3000
$\hat{\beta}_3$	0	0.0000	0.0105	0.0000	0.0962	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_4$	2	1.6360	0.0193	0.0036	0.1791	0.0000	0.0265	0.0200	0.2085
$\hat{\beta}_5$	0	0.0000	0.0127	70.0000	0.1066	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_6$	0	-0.1156	0.0126	0.0012	0.1129	0.0000	0.0021	0.0000	0.0196
$\hat{\beta}_7$	0	0.0000	0.0220	0.0000	0.1615	0.0000	0.0083	0.0000	0.0751
$\hat{\beta}_8$	0	0.0000	0.0304	0.0000	0.1473	0.0000	0.0145	0.0000	0.1262
median NO. of Zero coefficients		**3**				**6**			
Coefficients	True Values	GM-$LASSO$ Mean	GM-$LASSO$ Median (MSE)	GM-$LASSO$ Median ($MRPE$)	GM-$LASSO$ std	MM-$LASSO$ Mean	MM-$LASSO$ Median (MSE)	MM-$LASSO$ Median ($MRPE$)	MM-$LASSO$ std
$\hat{\beta}_0$	0	-0.3631	0.0291	0.0036	0.2202	0.0995	0.0325	0.0010	0.2705
$\hat{\beta}_1$	3	1.8314	0.0311	0.0117	0.2590	1.2155	0.0311	0.0178	0.2836
$\hat{\beta}_2$	1.5	0.9163	0.0492	0.0058	0.3240	0.9088	0.0689	0.0059	0.2906
$\hat{\beta}_3$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_4$	2	0.7789	0.0568	0.0122	0.3888	0.0000	0.0190	0.0200	0.1565
$\hat{\beta}_5$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_6$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_7$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_8$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
median NO. of Zero coefficients		**5**				**6**			



(a) $LASSO$ (b) $ada\text{-}LASSO$

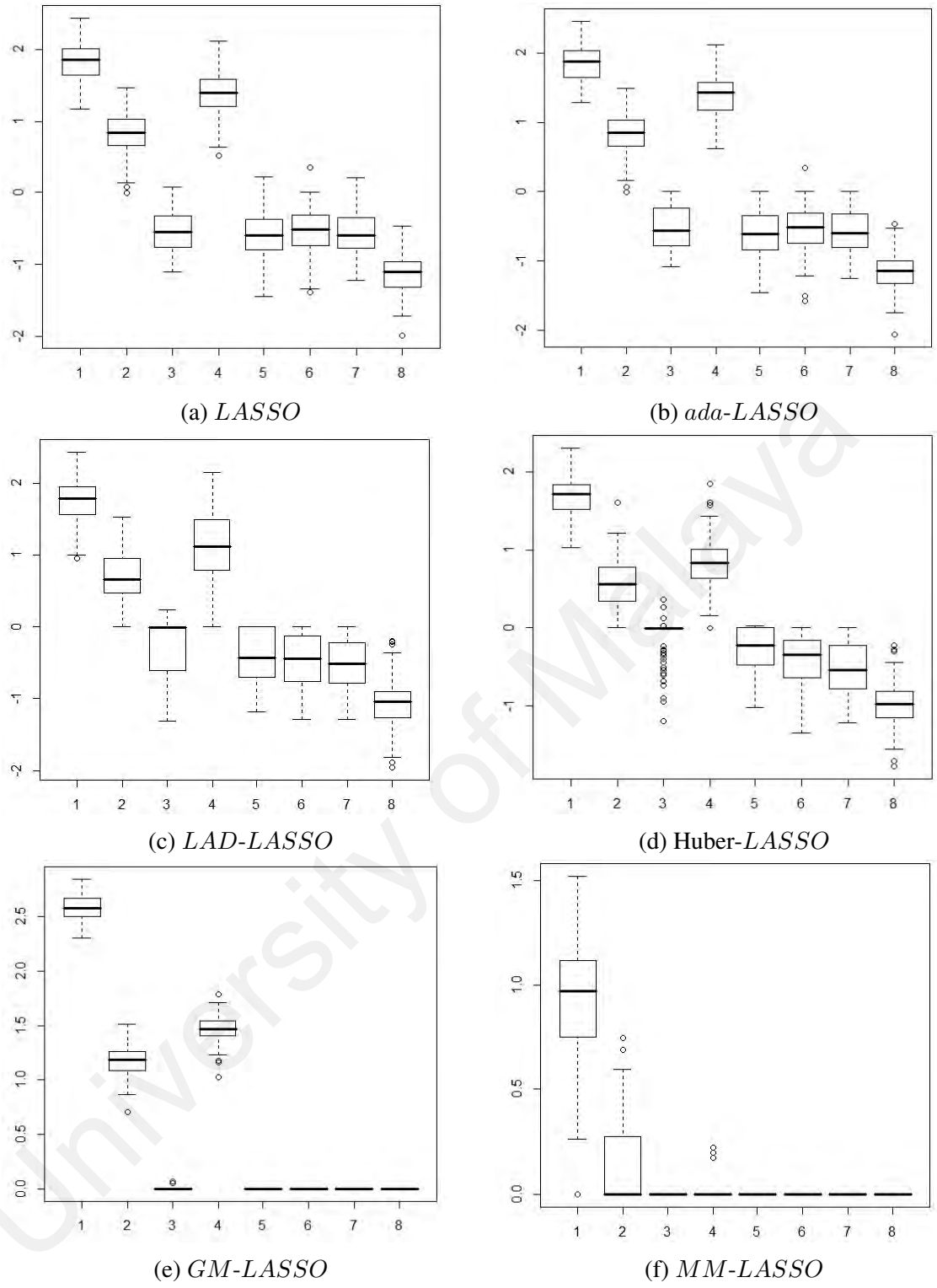
(c) $LAD\text{-}LASSO$ (d) Huber-$LASSO$

(e) $GM\text{-}LASSO$ (f) $MM\text{-}LASSO$

Figure 5.5: Boxplots of estimates for the eight coefficients from 100 simulated data sets, 5% bad leverage point

Table 5.6: The estimation of parameters for simulated data sets when 10% bad leverage point

Coefficients	True Values	$LASSO$ Mean	$LASSO$ Median (MSE)	$LASSO$ Median ($MRPE$)	$LASSO$ std	ada-$LASSO$ Mean	ada-$LASSO$ Median (MSE)	ada-$LASSO$ Median ($MRPE$)	ada-$LASSO$ std
$\hat{\beta}_0$	0	-0.4997	0.0279	0.0050	0.2330	-0.4985	0.0277	0.0050	0.2320
$\hat{\beta}_1$	3	2.0660	0.0369	0.0093	0.2594	2.0742	0.0359	0.0093	0.2597
$\hat{\beta}_2$	1.5	0.5368	0.0449	0.0096	0.3046	0.5268	0.0463	0.0097	0.3259
$\hat{\beta}_3$	0	-0.3248	0.0384	0.0032	0.3076	-0.3165	0.0408	0.0032	0.3338
$\hat{\beta}_4$	2	1.3192	0.0347	0.0068	0.3177	1.3213	0.0335	0.0068	0.3023
$\hat{\beta}_5$	0	-0.5414	0.0357	0.0054	0.3318	-0.5434	0.0376	0.0054	0.3487
$\hat{\beta}_6$	0	-0.5485	0.0335	0.0055	0.3116	-0.5557	0.0370	0.0056	0.3440
$\hat{\beta}_7$	0	-0.3358	0.0405	0.0034	0.2992	-0.3213	0.0436	0.0032	0.3270
$\hat{\beta}_8$	0	-1.5307	0.0519	0.0153	0.2668	-1.5434	0.0510	0.0154	0.2722
median NO. of Zero coefficients		**0**				**0**			
Coefficients	True Values	LAD-$LASSO$ Mean	LAD-$LASSO$ Median (MSE)	LAD-$LASSO$ Median ($MRPE$)	LAD-$LASSO$ std	Huber-$LASSO$ Mean	Huber-$LASSO$ Median (MSE)	Huber-$LASSO$ Median ($MRPE$)	Huber-$LASSO$ std
$\hat{\beta}_0$	0	-0.7671	0.0474	0.0077	0.2708	-0.4928	0.0288	0.0049	0.2449
$\hat{\beta}_1$	3	2.1116	0.0497	0.0089	0.2996	2.1301	0.0555	0.0087	0.2723
$\hat{\beta}_2$	1.5	0.4956	0.0450	0.0100	0.3603	0.3003	0.0445	0.0120	0.3057
$\hat{\beta}_3$	0	0.0000	0.0510	0.0000	0.3845	0.0000	0.0285	0.0000	0.2484
$\hat{\beta}_4$	2	0.9479	0.0520	0.0105	0.4561	0.6494	0.0429	0.0135	0.3532
$\hat{\beta}_5$	0	-0.4631	0.0405	0.0046	0.3752	-0.3026	0.0302	0.0030	0.2792
$\hat{\beta}_6$	0	-0.3914	0.0408	0.0039	0.3711	-0.1773	0.0465	0.0018	0.3537
$\hat{\beta}_7$	0	0.0000	0.0673	0.0000	0.3544	-0.0377	0.0605	0.0004	0.3284
$\hat{\beta}_8$	0	-1.8870	0.0964	0.0189	0.3445	-1.7661	0.0895	0.0177	0.3114
median NO. of Zero coefficients		**1**				**2**			
Coefficients	True Values	GM-$LASSO$ Mean	GM-$LASSO$ Median (MSE)	GM-$LASSO$ Median ($MRPE$)	GM-$LASSO$ std	MM-$LASSO$ Mean	MM-$LASSO$ Median (MSE)	MM-$LASSO$ Median ($MRPE$)	MM-$LASSO$ std
$\hat{\beta}_0$	0	-0.4939	0.0282	0.0049	0.2414	0.4182	0.0352	0.0042	0.3143
$\hat{\beta}_1$	3	2.1664	0.0541	0.0083	0.2757	1.0428	0.0315	0.0196	0.2674
$\hat{\beta}_2$	1.5	0.2834	0.0469	0.0122	0.3108	0.0000	0.0250	0.0150	0.2017
$\hat{\beta}_3$	0	0.0000	0.0266	0.0000	0.2324	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_4$	2	0.6764	0.0416	0.0132	0.3436	0.0000	0.0037	0.0200	0.0339
$\hat{\beta}_5$	0	-0.3129	0.0305	0.0031	0.2757	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_6$	0	-0.1895	0.0454	0.0019	0.3554	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_7$	0	-0.0313	0.0599	0.0003	0.3267	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_8$	0	-1.7766	0.0925	0.0178	0.3082	0.0000	0.0000	0.0000	0.0000
median NO. of Zero coefficients		**2**				**7**			



(a) $LASSO$ (b) $ada\text{-}LASSO$

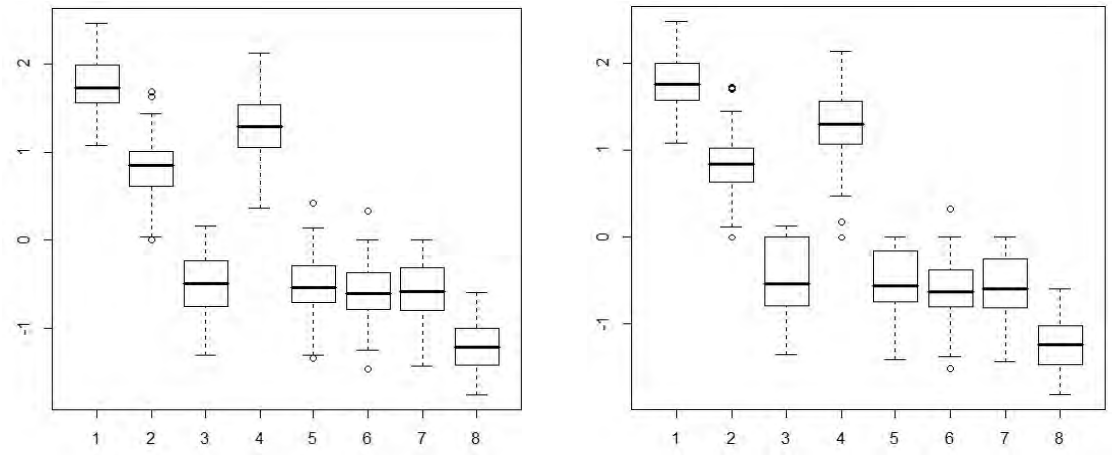
(c) $LAD\text{-}LASSO$ (d) Huber-$LASSO$

(e) $GM\text{-}LASSO$ (f) $MM\text{-}LASSO$

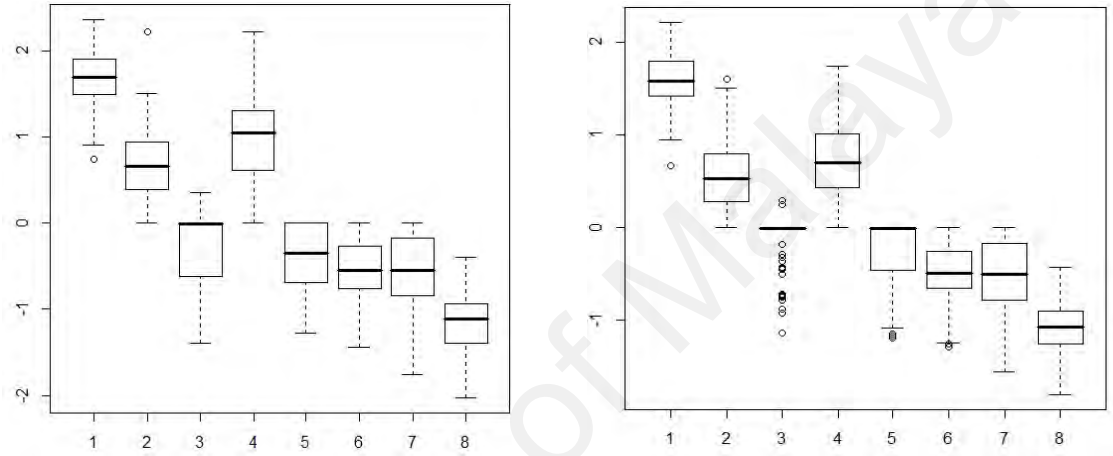
Figure 5.6: Boxplots of estimates for the eight coefficients from 100 simulated data sets, 10% bad leverage point

Table 5.7: The estimation of parameters for simulated data sets when 20% bad leverage point

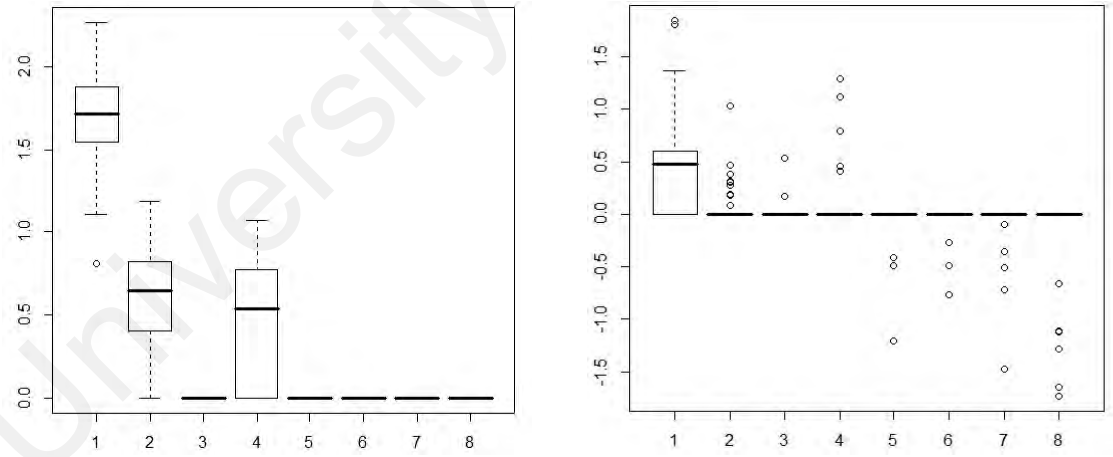
Coefficients	True Values	$LASSO$ Mean	$LASSO$ Median (MSE)	$LASSO$ Median ($MRPE$)	$LASSO$ std	ada-$LASSO$ Mean	ada-$LASSO$ Median (MSE)	ada-$LASSO$ Median ($MRPE$)	ada-$LASSO$ std
$\hat{\beta}_0$	0	-0.6083	0.0325	0.0061	0.2533	-0.6099	0.0328	0.0061	0.2572
$\hat{\beta}_1$	0	1.6951	0.0335	0.0130	0.3080	1.7420	0.0336	0.0126	0.3118
$\hat{\beta}_2$	3	0.9233	0.0393	0.0058	0.3503	0.9494	0.0441	0.0055	0.3842
$\hat{\beta}_3$	1.5	0.0000	0.0648	0.0000	0.3562	0.0000	0.0669	0.0000	0.3939
$\hat{\beta}_4$	0	0.4671	0.0952	0.0153	0.3820	0.5163	0.0936	0.0148	0.4079
$\hat{\beta}_5$	2	-0.3736	0.0408	0.0037	0.3576	-0.4336	0.0412	0.0043	0.3744
$\hat{\beta}_6$	0	-0.4047	0.0398	0.0040	0.3187	-0.4230	0.0421	0.0042	0.3507
$\hat{\beta}_7$	0	-0.3460	0.0455	0.0035	0.3587	-0.3132	0.0492	0.0031	0.3897
$\hat{\beta}_8$	0	-1.4170	0.0362	0.0142	0.2712	-1.4902	0.0402	0.0149	0.2832
median NO. of Zero coefficients		**0**				**1**			
Coefficients	True Values	LAD-$LASSO$ Mean	LAD-$LASSO$ Median (MSE)	LAD-$LASSO$ Median ($MRPE$)	LAD-$LASSO$ std	Huber-$LASSO$ Mean	Huber-$LASSO$ Median (MSE)	Huber-$LASSO$ Median ($MRPE$)	Huber-$LASSO$ std
$\hat{\beta}_0$	0	-0.5621	0.0365	0.0056	0.3104	-0.5468	0.0317	0.0055	0.2772
$\hat{\beta}_1$	3	1.4144	0.0440	0.0159	0.3203	1.4100	0.0373	0.0159	0.2959
$\hat{\beta}_2$	1.5	0.9872	0.0559	0.0051	0.4230	0.9707	0.0602	0.0053	0.3766
$\hat{\beta}_3$	0	0.0000	0.0531	0.0000	0.4064	0.0000	0.0298	0.0000	0.2594
$\hat{\beta}_4$	2	0.0000	0.1197	0.0200	0.4932	0.0000	0.0907	0.0200	0.4452
$\hat{\beta}_5$	0	-0.0389	0.0550	0.0004	0.3778	0.0000	0.0453	0.0000	0.3376
$\hat{\beta}_6$	0	-0.3139	0.0448	0.0031	0.3559	-0.2582	0.0409	0.0026	0.3166
$\hat{\beta}_7$	0	0.0000	0.0720	0.0000	0.4078	-0.0848	0.0592	0.0008	0.3741
$\hat{\beta}_8$	0	-1.5588	0.0555	0.0156	0.3329	-1.5139	0.0565	0.0151	0.3053
median NO. of Zero coefficients		**1**				**2**			
Coefficients	True Values	GM-$LASSO$ Mean	GM-$LASSO$ Median (MSE)	GM-$LASSO$ Median ($MRPE$)	GM-$LASSO$ std	MM-$LASSO$ Mean	MM-$LASSO$ Median (MSE)	MM-$LASSO$ Median ($MRPE$)	MM-$LASSO$ std
$\hat{\beta}_0$	0	-0.5087	0.0304	0.0051	0.2759	1.5053	0.0630	0.0151	0.5730
$\hat{\beta}_1$	3	1.4279	0.0373	0.0157	0.2960	0.0000	0.0628	0.0300	0.3771
$\hat{\beta}_2$	1.5	0.9728	0.0602	0.0053	0.3834	0.0000	0.0151	0.0150	0.1360
$\hat{\beta}_3$	0	0.0000	0.0294	0.0000	0.2551	0.0000	0.0061	0.0000	0.0564
$\hat{\beta}_4$	2	0.0000	0.0923	0.0200	0.4468	0.0000	0.0218	0.0200	0.1986
$\hat{\beta}_5$	0	0.0000	0.0455	0.0000	0.3343	0.0000	0.0148	0.0000	0.1365
$\hat{\beta}_6$	0	-0.2082	0.0449	0.0021	0.3206	0.0000	0.0103	0.0000	0.0947
$\hat{\beta}_7$	0	-0.1530	0.0545	0.0015	0.3766	0.0000	0.0189	0.0000	0.1734
$\hat{\beta}_8$	0	-1.5091	0.0551	0.0151	0.3154	0.0000	0.0345	0.0000	0.3130
median NO. of Zero coefficients		**2**				**7**			



(a) $LASSO$ (b) $ada\text{-}LASSO$



(c) $LAD\text{-}LASSO$ (d) Huber-$LASSO$

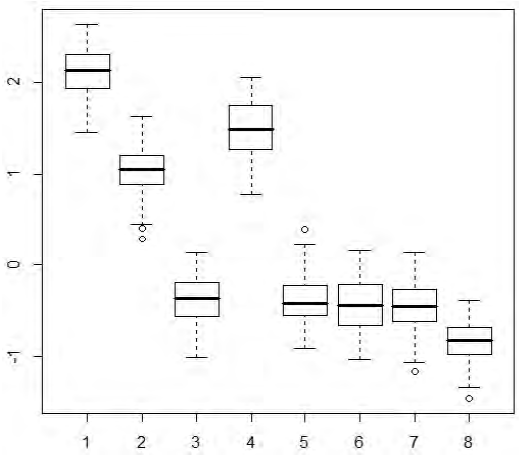


(e) $GM\text{-}LASSO$ (f) $MM\text{-}LASSO$

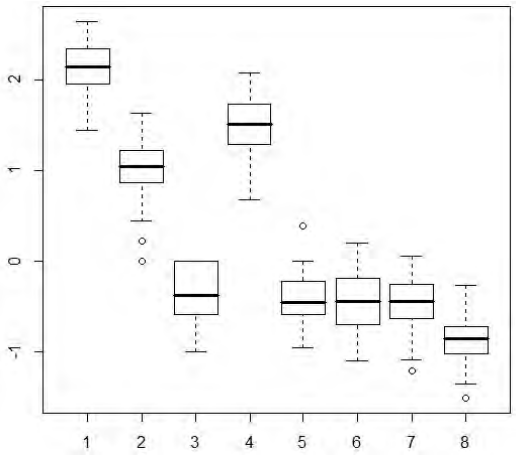
Figure 5.7: Boxplots of estimates for the eight coefficients from 100 simulated data sets, 20% bad leverage point

Table 5.8: The estimation of parameters for simulated data sets when 5% good leverage point

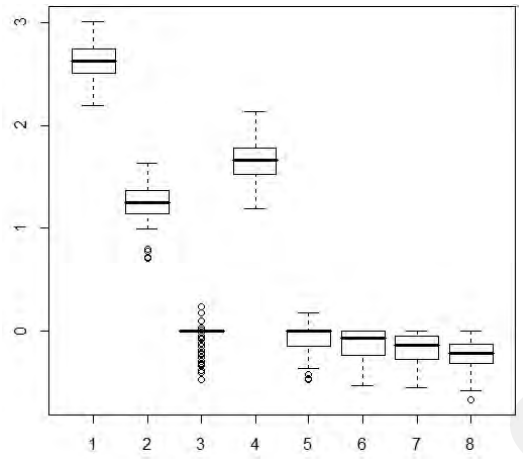
Coefficients	True Values	$LASSO$ Mean	$LASSO$ Median (MSE)	$LASSO$ Median ($MRPE$)	$LASSO$ std	ada-$LASSO$ Mean	ada-$LASSO$ Median (MSE)	ada-$LASSO$ Median ($MRPE$)	ada-$LASSO$ std
$\hat{\beta}_0$	0	-0.3285	0.0236	0.0033	0.2141	-0.3282	0.0236	0.0033	0.2147
$\hat{\beta}_1$	3	2.4893	0.0475	0.0051	0.2423	2.4940	0.0466	0.0051	0.2448
$\hat{\beta}_2$	1.5	0.9051	0.0321	0.0059	0.2707	0.9011	0.0338	0.0060	0.2906
$\hat{\beta}_3$	0	-0.6937	0.0448	0.0069	0.2700	-0.6954	0.0466	0.0070	0.2866
$\hat{\beta}_4$	2	1.8251	0.0480	0.0017	0.3067	1.8270	0.0478	0.0017	0.3083
$\hat{\beta}_5$	0	-0.4915	0.0280	0.0049	0.2464	-0.4909	0.0311	0.0049	0.2759
$\hat{\beta}_6$	0	-0.4465	0.0303	0.0045	0.2832	-0.4443	0.0345	0.0044	0.3226
$\hat{\beta}_7$	0	-0.4934	0.0270	0.0049	0.2510	-0.4882	0.0310	0.0049	0.2873
$\hat{\beta}_8$	0	-0.9743	0.0259	0.0097	0.2022	-0.9813	0.0264	0.0098	0.2165
median NO. of Zero coefficients		**0**				**0**			
Coefficients	True Values	LAD-$LASSO$ Mean	LAD-$LASSO$ Median (MSE)	LAD-$LASSO$ Median ($MRPE$)	LAD-$LASSO$ std	Huber-$LASSO$ Mean	Huber-$LASSO$ Median (MSE)	Huber-$LASSO$ Median ($MRPE$)	Huber-$LASSO$ std
$\hat{\beta}_0$	0	-0.0973	0.0136	0.0010	0.1272	-0.0411	0.0327	0.0004	0.2584
$\hat{\beta}_1$	3	2.8095	0.0277	0.0019	0.1792	1.7599	0.0308	0.0124	0.2794
$\hat{\beta}_2$	1.5	1.3343	0.0213	0.0017	0.1787	0.3416	0.0418	0.0116	0.3194
$\hat{\beta}_3$	0	-0.1443	0.0170	0.0014	0.1133	0.0000	0.0006	0.0000	0.0051
$\hat{\beta}_4$	2	1.9028	0.0345	0.0010	0.1988	0.4634	0.0300	0.0154	0.2760
$\hat{\beta}_5$	0	-0.1358	0.0153	0.0014	0.1306	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_6$	0	-0.1997	0.0172	0.0020	0.1402	0.0000	0.0036	0.0000	0.0332
$\hat{\beta}_7$	0	-0.3392	0.0238	0.0034	0.1309	0.0000	0.0106	0.0000	0.0940
$\hat{\beta}_8$	0	-0.2253	0.0153	0.0023	0.1426	0.0000	0.0144	0.0000	0.1207
median NO. of Zero coefficients		**3**				**5**			
Coefficients	True Values	GM-$LASSO$ Mean	GM-$LASSO$ Median (MSE)	GM-$LASSO$ Median ($MRPE$)	GM-$LASSO$ std	MM-$LASSO$ Mean	MM-$LASSO$ Median (MSE)	MM-$LASSO$ Median ($MRPE$)	MM-$LASSO$ std
$\hat{\beta}_0$	0	-0.1419	0.0190	0.0014	0.1772	0.4967	0.0427	0.0050	0.2556
$\hat{\beta}_1$	3	2.1574	0.0203	0.0084	0.1887	1.8192	0.0319	0.0118	0.2535
$\hat{\beta}_2$	1.5	0.7503	0.0345	0.0075	0.2913	0.0000	0.0627	0.0150	0.2778
$\hat{\beta}_3$	0	0.0000	0.0004	0.0000	0.0036	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_4$	2	0.9779	0.0294	0.0102	0.2742	0.3293	0.0247	0.0167	0.2295
$\hat{\beta}_5$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_6$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_7$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_8$	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
median NO. of Zero coefficients		**5**				**5**			



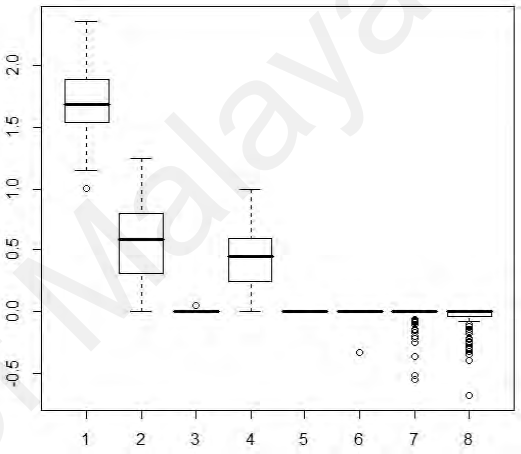
(a) $LASSO$



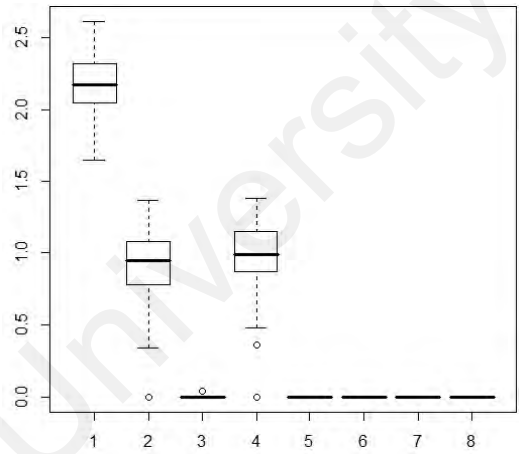
(b) $ada\text{-}LASSO$



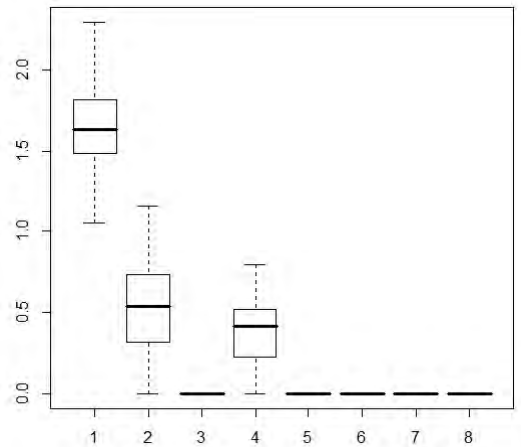
(c) $LAD\text{-}LASSO$



(d) Huber-$LASSO$



(e) $GM\text{-}LASSO$

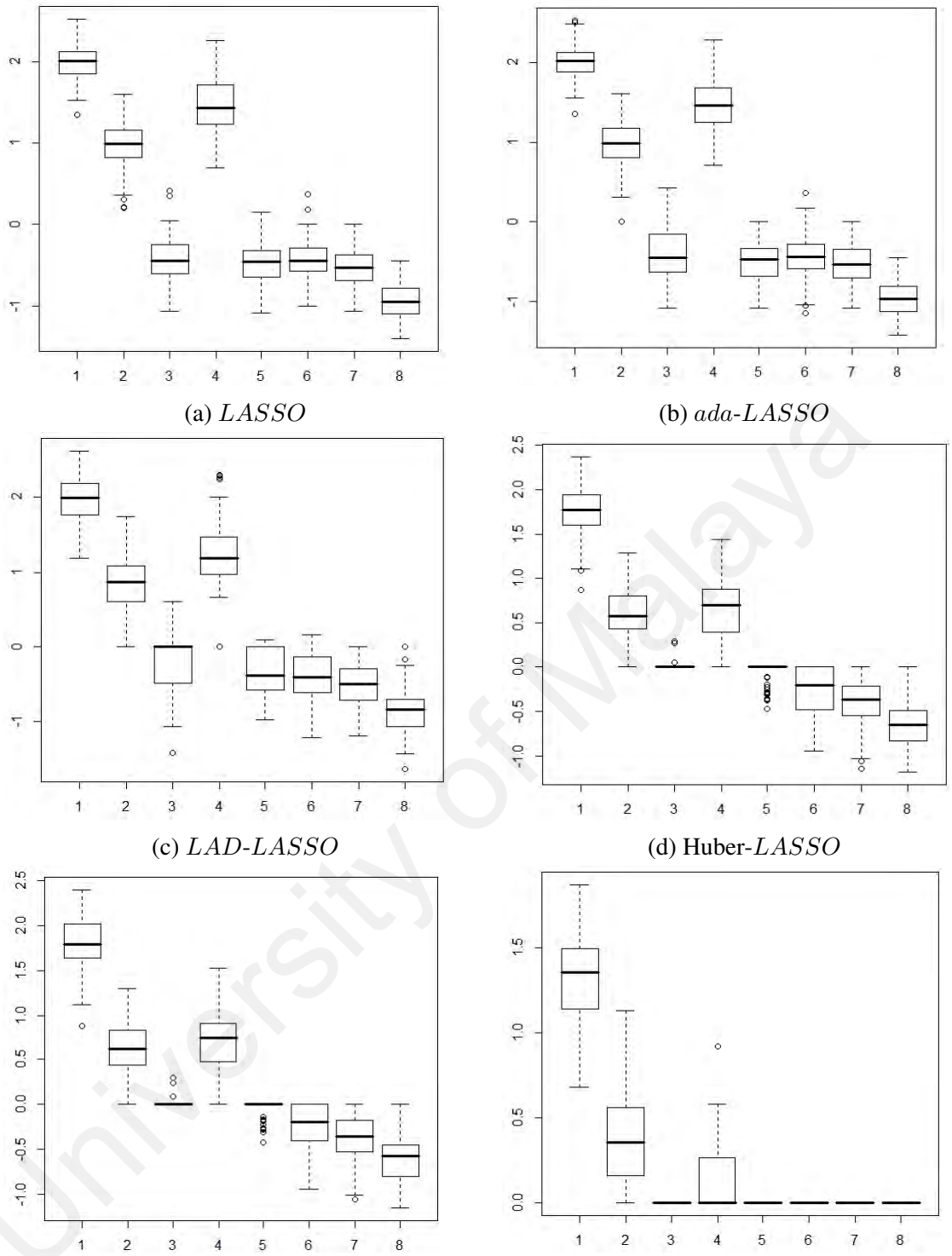


(f) $MM\text{-}LASSO$

Figure 5.8: Boxplots of estimates for the eight coefficients from 100 simulated data sets, 5% good leverage point

Table 5.9: The estimation of parameters for simulated data sets when 10% good leverage point

Coefficients	True Values	$LASSO$ Mean	$LASSO$ Median (MSE)	$LASSO$ Median $(MRPE)$	$LASSO$ std	$ada\text{-}LASSO$ Mean	$ada\text{-}LASSO$ Median (MSE)	$ada\text{-}LASSO$ Median $(MRPE)$	$ada\text{-}LASSO$ std
$\hat{\beta}_0$	0	-0.27337	0.0253	0.0027	0.2321	-0.2712	0.0253	0.0027	0.2317
$\hat{\beta}_1$	3	2.1172	0.0264	0.0088	0.2102	2.1213	0.0258	0.0088	0.2118
$\hat{\beta}_2$	1.5	1.0244	0.0323	0.0048	0.2978	1.0323	0.0354	0.0047	0.3244
$\hat{\beta}_3$	0	-0.6766	0.0428	0.0068	0.3070	-0.6944	0.0457	0.0069	0.3272
$\hat{\beta}_4$	2	1.3258	0.0401	0.0067	0.3481	1.3362	0.0388	0.0066	0.3305
$\hat{\beta}_5$	0	-0.3442	0.0306	0.0034	0.2521	-0.3384	0.0321	0.0034	0.2582
$\hat{\beta}_6$	0	-0.8450	0.0526	0.0084	0.2623	-0.8694	0.0558	0.0087	0.2876
$\hat{\beta}_7$	0	-0.1941	0.0431	0.0019	0.2395	-0.1623	0.0466	0.0016	0.2622
$\hat{\beta}_8$	0	-0.5894	0.0454	0.0059	0.2268	-0.6038	0.0464	0.0060	0.2322
median NO. of Zero coefficients		**0**				**0**			
Coefficients	True Values	$LAD\text{-}LASSO$ Mean	$LAD\text{-}LASSO$ Median (MSE)	$LAD\text{-}LASSO$ Median $(MRPE)$	$LAD\text{-}LASSO$ std	Huber-$LASSO$ Mean	Huber-$LASSO$ Median (MSE)	Huber-$LASSO$ Median $(MRPE)$	Huber-$LASSO$ std
$\hat{\beta}_0$	0	-0.3988	0.0284	0.0040	0.2516	-0.4104	0.0304	0.0041	0.2694
$\hat{\beta}_1$	3	2.3244	0.0481	0.0068	0.2866	2.1826	0.0555	0.0082	0.2768
$\hat{\beta}_2$	1.5	0.9556	0.0421	0.0054	0.3747	0.4630	0.0359	0.0104	0.3041
$\hat{\beta}_3$	0	-0.4962	0.0481	0.0050	0.3565	0.0000	0.0042	0.0000	0.0388
$\hat{\beta}_4$	2	0.9093	0.0563	0.0109	0.3921	0.3542	0.0491	0.0165	0.3235
$\hat{\beta}_5$	0	-0.1933	0.0361	0.0019	0.2926	0.0000	0.0122	0.0000	0.1056
$\hat{\beta}_6$	0	-0.9840	0.0702	0.0098	0.3286	-0.3883	0.0306	0.0039	0.2565
$\hat{\beta}_7$	0	0.0000	0.0620	0.0000	0.2961	-0.2243	0.0334	0.0022	0.2656
$\hat{\beta}_8$	0	-0.6297	0.0395	0.0063	0.2779	-0.4122	0.0370	0.0041	0.2516
median NO. of Zero coefficients		**3**				**5**			
Coefficients	True Values	$GM\text{-}LASSO$ Mean	$GM\text{-}LASSO$ Median (MSE)	$GM\text{-}LASSO$ Median $(MRPE)$	$GM\text{-}LASSO$ std	$MM\text{-}LASSO$ Mean	$MM\text{-}LASSO$ Median (MSE)	$MM\text{-}LASSO$ Median $(MRPE)$	$MM\text{-}LASSO$ std
$\hat{\beta}_0$	0	-0.4153	0.0300	0.0042	0.2602	0.2076	0.0427	0.0021	0.3178
$\hat{\beta}_1$	3	2.2379	0.0548	0.0076	0.2800	1.6459	0.0444	0.0135	0.2501
$\hat{\beta}_2$	1.5	0.4466	0.0380	0.0105	0.2962	0.3405	0.0315	0.0116	0.2929
$\hat{\beta}_3$	0	0.0000	0.0042	0.0000	0.0389	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_4$	2	0.3838	0.0503	0.0162	0.3153	0.0000	0.0239	0.0200	0.1887
$\hat{\beta}_5$	0	0.0000	0.0110	0.0000	0.0969	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_6$	0	-0.2965	0.0268	0.0030	0.2435	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_7$	0	-0.1941	0.0336	0.0019	0.2600	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_8$	0	-0.3268	0.0409	0.0033	0.2551	0.0000	0.0000	0.0000	0.0000
median NO. of Zero coefficients		**5**				**5**			



(a) $LASSO$ (b) $ada\text{-}LASSO$

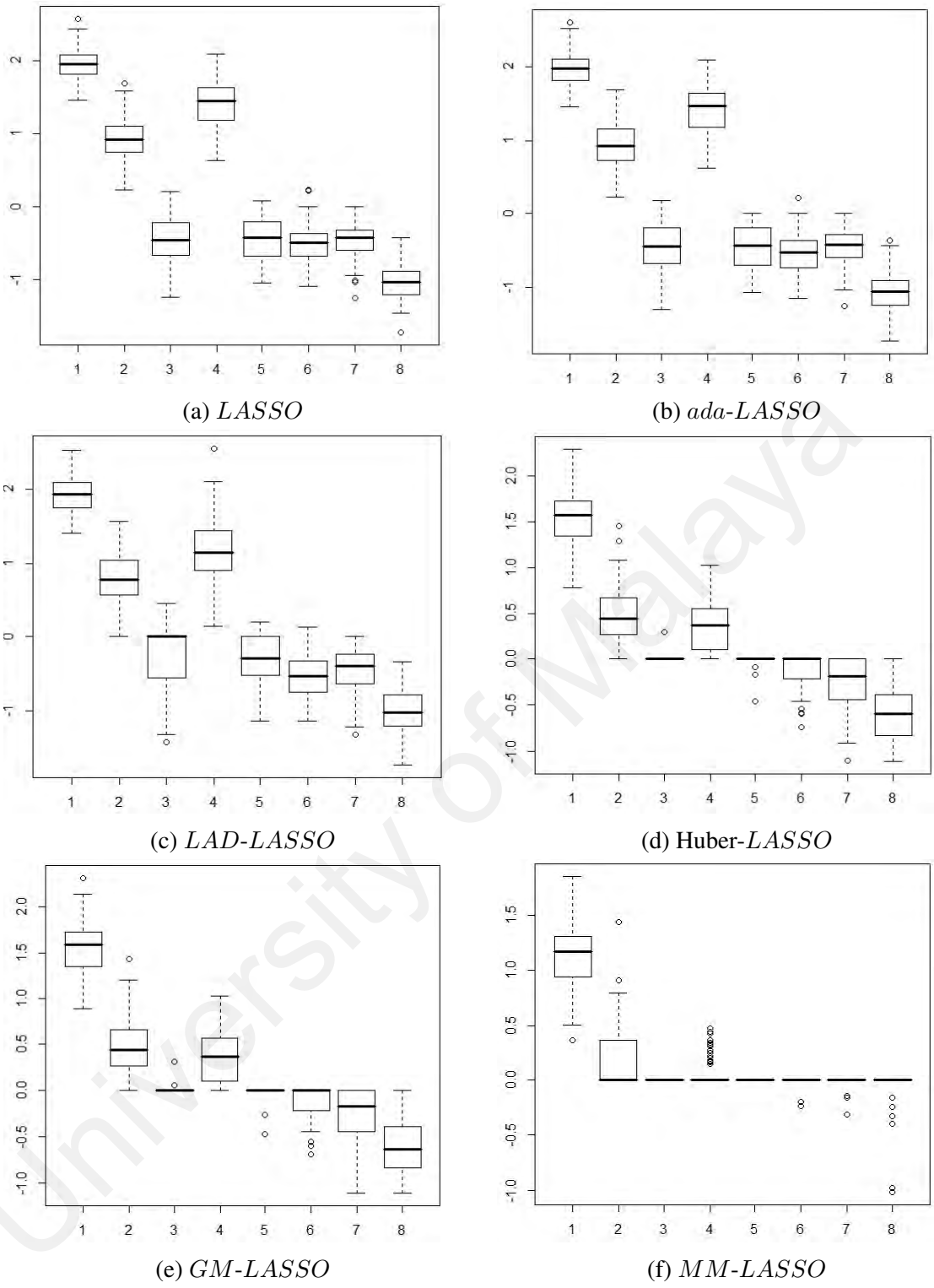
(c) $LAD\text{-}LASSO$ (d) Huber-$LASSO$

(e) $GM\text{-}LASSO$ (f) $MM\text{-}LASSO$

Figure 5.9: Boxplots of estimates for the eight coefficients from 100 simulated data sets, 10% good leverage point

Table 5.10: The estimation of parameters for simulated data sets when 20% good leverage point

Coefficients	True Values	$LASSO$ Mean	$LASSO$ Median (MSE)	$LASSO$ Median ($MRPE$)	$LASSO$ std	ada-$LASSO$ Mean	ada-$LASSO$ Median (MSE)	ada-$LASSO$ Median ($MRPE$)	ada-$LASSO$ std
$\hat{\beta}_0$	0	-0.7574	0.0484	0.0076	0.2321	-0.7607	0.0485	0.0076	0.2319
$\hat{\beta}_1$	3	1.6672	0.0381	0.0133	0.2081	1.6755	0.0393	0.0132	0.2112
$\hat{\beta}_2$	1.5	1.2399	0.0442	0.0026	0.2794	1.2612	0.0474	0.0024	0.2957
$\hat{\beta}_3$	0	-0.6435	0.0396	0.0064	0.3123	-0.6808	0.0432	0.0068	0.3295
$\hat{\beta}_4$	2	1.5983	0.0392	0.0040	0.3131	1.6254	0.0396	0.0037	0.3115
$\hat{\beta}_5$	0	-0.2634	0.0375	0.0026	0.3002	-0.2574	0.0397	0.0026	0.3204
$\hat{\beta}_6$	0	-0.7498	0.0373	0.0075	0.2592	-0.7691	0.0391	0.0077	0.2825
$\hat{\beta}_7$	0	-0.4088	0.0269	0.0041	0.2443	-0.3972	0.0302	0.0040	0.2770
$\hat{\beta}_8$	0	-0.8958	0.0294	0.0090	0.2384	-0.9121	0.0312	0.0091	0.2520
median NO. of Zero coefficients		**0**				**0**			
Coefficients	True Values	LAD-$LASSO$ Mean	LAD-$LASSO$ Median (MSE)	LAD-$LASSO$ Median ($MRPE$)	LAD-$LASSO$ std	Huber-$LASSO$ Mean	Huber-$LASSO$ Median (MSE)	Huber-$LASSO$ Median ($MRPE$)	Huber-$LASSO$ std
$\hat{\beta}_0$	0	-0.9754	0.0702	0.0098	0.3022	-0.8700	0.0592	0.0087	0.2713
$\hat{\beta}_1$	3	1.9581	0.0274	0.0104	0.2525	1.2475	0.0435	0.0175	0.2660
$\hat{\beta}_2$	1.5	1.0113	0.0434	0.0049	0.3457	0.4937	0.0316	0.0101	0.2940
$\hat{\beta}_3$	0	-0.5256	0.0497	0.0053	0.3729	0.0000	0.0032	0.0000	0.0295
$\hat{\beta}_4$	2	1.3267	0.0464	0.0067	0.4060	0.3493	0.0291	0.0165	0.2721
$\hat{\beta}_5$	0	0.0000	0.0502	0.0000	0.3452	0.0000	0.0053	0.0000	0.0494
$\hat{\beta}_6$	0	-1.1348	0.0749	0.0113	0.3059	-0.1593	0.0181	0.0016	0.1651
$\hat{\beta}_7$	0	-0.3978	0.0310	0.0040	0.2890	-0.1850	0.0276	0.0018	0.2503
$\hat{\beta}_8$	0	-0.7693	0.0385	0.0077	0.2766	-0.1607	0.0562	0.0016	0.2922
median NO. of Zero coefficients		**3**				**5**			
Coefficients	True Values	GM-$LASSO$ Mean	GM-$LASSO$ Median (MSE)	GM-$LASSO$ Median ($MRPE$)	GM-$LASSO$ std	MM-$LASSO$ Mean	MM-$LASSO$ Median (MSE)	MM-$LASSO$ Median ($MRPE$)	MM-$LASSO$ std
$\hat{\beta}_0$	0	-0.8426	0.0572	0.0084	0.2741	0.6611	0.0735	0.0066	0.4912
$\hat{\beta}_1$	3	1.2892	0.0414	0.0171	0.2685	1.5515	0.0551	0.0145	0.2827
$\hat{\beta}_2$	1.5	0.4654	0.0311	0.0103	0.2903	0.0000	0.0365	0.0150	0.2747
$\hat{\beta}_3$	0	0.0000	0.0033	0.0000	0.0309	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_4$	2	0.3607	0.0294	0.0164	0.2752	0.0000	0.0136	0.0200	0.1165
$\hat{\beta}_5$	0	0.0000	0.0058	0.0000	0.0539	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_6$	0	-0.1456	0.0178	0.0015	0.1649	0.0000	0.0033	0.0000	0.0306
$\hat{\beta}_7$	0	-0.1597	0.0288	0.0016	0.2538	0.0000	0.0041	0.0000	0.0379
$\hat{\beta}_8$	0	-0.2062	0.0531	0.0021	0.2985	0.0000	0.0164	0.0000	0.1505
median NO. of Zero coefficients		**5**				**5**			



(a) $LASSO$ (b) $ada\text{-}LASSO$

(c) $LAD\text{-}LASSO$ (d) Huber-$LASSO$

(e) $GM\text{-}LASSO$ (f) $MM\text{-}LASSO$

Figure 5.10: Boxplots of estimates for the eight coefficients from 100 simulated data sets, 20% good leverage point

5.9.1 Simulation Study (Multicollinearity)

In this section, we performed a Monte Carlo simulation study to demonstrate the efficiency of the proposed estimators, $GM\text{-}LASSO$ and $MM\text{-}LASSO$ in comparison with

several existing estimators. We allowed various degrees of multicollinearity and non-normal disturbance distributions to be present simultaneously in this simulation. There are four estimators in the study, (i) $ada\text{-}LASSO$, (ii) $LAD\text{-}LASSO$, (iii) $GM\text{-}LASSO$ and (iv) $MM\text{-}LASSO$. The following model was used in this simulation study;

$$y_i = \boldsymbol{\beta}_1 x_{i1} + \boldsymbol{\beta}_2 x_{i2} + \boldsymbol{\beta}_3 x_{i3} + \boldsymbol{\beta}_4 x_{i4} + \boldsymbol{\beta}_5 x_{i5} + \boldsymbol{\beta}_6 x_{i6} + \epsilon_i,$$

where $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3 = 1$, and $\boldsymbol{\beta}_4 = \boldsymbol{\beta}_5 = \boldsymbol{\beta}_6 = 0$. The sample size used was $n = 50$. The value of ρ represents the correlation between the six explanatory variables. The chosen values were 0.0, 0.5, 0.8. And, ϵ were generated from two different distributions,

i Standard normal distribution

ii Student-t distribution with degrees of freedom three.

The aim of this simulation study is to see the effect of combined problems of multicollinearity and outliers on the $ada\text{-}LASSO$, $LAD\text{-}LASSO$, $GM\text{-}LASSO$ and $MM\text{-}LASSO$ estimators. The performances of the aforementioned estimators were assessed by looking at parameter estimate and $MRPE$s on 1000 simulation runs.

Result and Discussion

Case 1 (Error distribution following the normal distribution):

Table 5.11 presented the parameter estimates and the respective Median of Relative Predictor Errors ($MRPE$) for simulated data sets with normal distribution with mean 0 and variance 1. It is obvious that the $MRPE$ of the $ada\text{-}LASSO$ is relatively smaller than the other estimators when the errors are normally distributed and multicollinearity is not present. As expected, the $ada\text{-}LASSO$ gave the best variable selection for the normal case as shown in Figure (5.11).

However, for normal error distribution and when a moderate correlation i.e. $\rho = 0.5$, is present in the data, the $GM\text{-}LASSO$ and $MM\text{-}LASSO$ give smaller $MRPE$s for all parameters estimated compared to the other two methods as shown in Table 5.12. We can see in Figure (5.12) that all estimators correctly fit the zero and non-zero variables. The $GM\text{-}LASSO$ portrays the lowest variability among the three estimators.

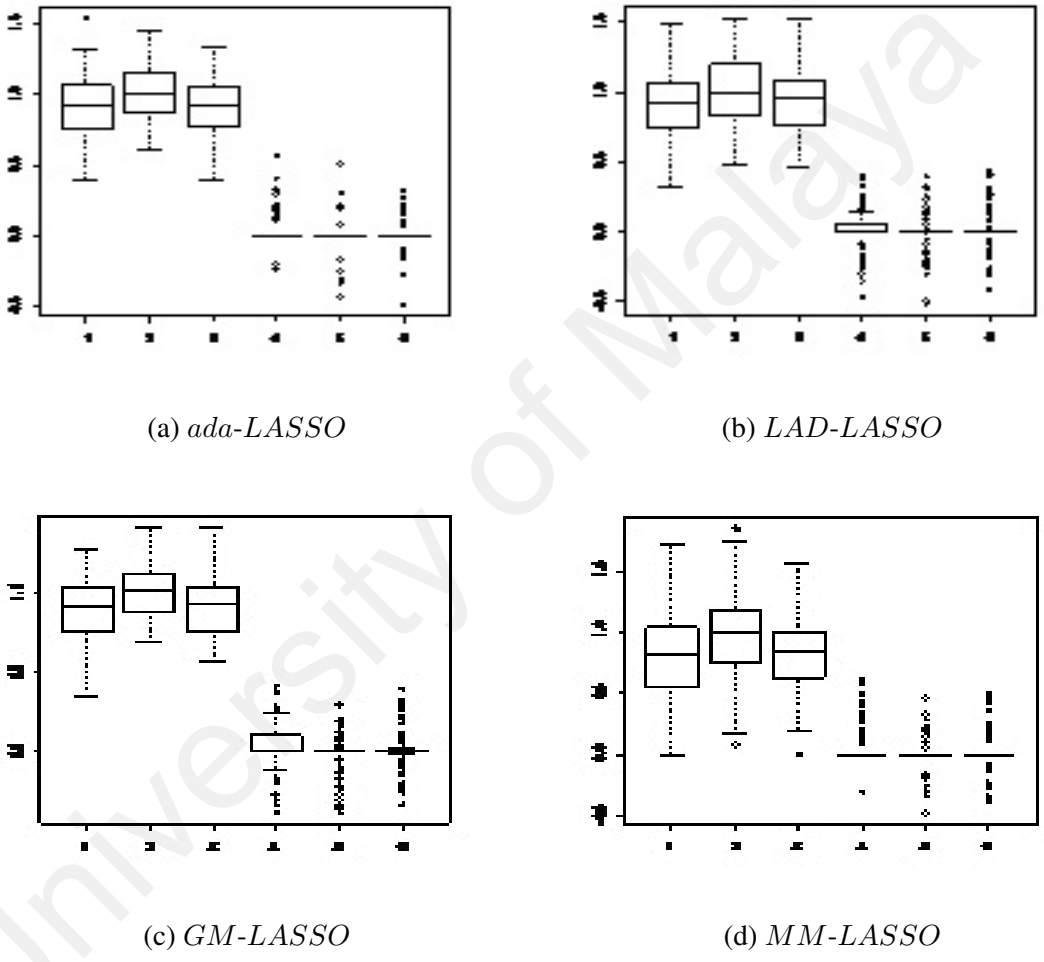
For normal error distributions when the correlation is high in the data where $\rho = 0.8$, the $GM\text{-}LASSO$ and $MM\text{-}LASSO$ outperform the other two estimators in variable selection. This can be clearly seen in Table 5.13 based on the $MRPE$ values. According to Figure (5.13), the $GM\text{-}LASSO$ and $MM\text{-}LASSO$ correctly fits the zero and non-zero coefficients while the $ada\text{-}LASSO$ and $LAD\text{-}LASSO$ tend to slightly over fit as there are four non-zero coefficients instead.

Case 2 (Error distribution following Student-t distribution):

Table 5.14 shows the parameters estimates when the error distribution follows the t-distribution with three degrees of freedom. Unlike the case discussed in Table 5.11 and Figure (5.11), the $ada\text{-}LASSO$ estimator no longer correctly fits the variables. Here, when the error distribution has heavier tails, the $MM\text{-}LASSO$ appears to be more superior compared to the other three estimators. This can be seen in Figure (5.14). When the correlation is increased to 0.5 and 0.8, the $GM\text{-}LASSO$ and $MM\text{-}LASSO$ are still seen to outperform the other two estimators, in which $MM\text{-}LASSO$ is superior, this is clearly shown in Tables 5.15 and 5.16 and Figures (5.15) and (5.16).

Table 5.11: Parameter estimates and their MRPEs (bracketed) for simulated data sets with normal distribution errors with mean 0 and variance 1, where $\beta_1 = \beta_2 = \beta_3 = 1$, and $\beta_4 = \beta_5 = \beta_6 = 0$

	Values of $\rho = 0$			
Parameter estimates	*Ada-LASSO*	*LAD-LASSO*	*GM-LASSO*	*MM-LASSO*
$\hat{\beta}_1$	1.1900 (0.1836)	1.0972(0.1886)	1.2821(0.2274)	1.0645(0.2160)
$\hat{\beta}_2$	0.8821 (0.1909)	0.8218(0.2162)	0.8927 (0.4991)	0.7934(0.3026)
$\hat{\beta}_3$	0.8135 (0.2058)	0.9525 (0.2302)	0.9048(0.3478)	0.6012(0.3599)
$\hat{\beta}_4$	0.2216 (0.2695)	0.2437 (0.6258)	0.3192 (0.05329)	0.5845(0.4577)
$\hat{\beta}_5$	0.0000 (0.2352)	0.0000 (0.4722)	0.0000 (0.3773)	0.0000(0.3846)
$\hat{\beta}_6$	0.0000 (0.1486)	0.0000 (0.1171)	0.0000 (0.0000)	0.0000(0.0000)



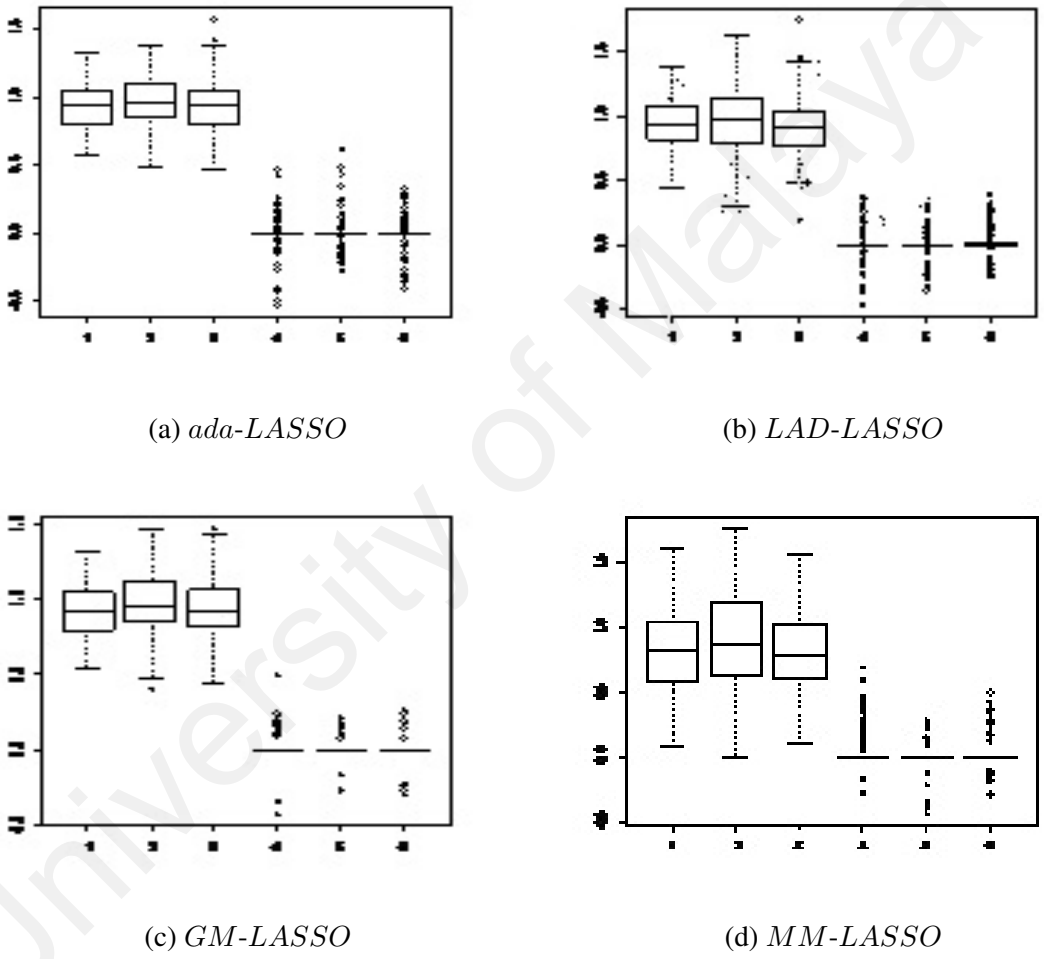
(a) *ada-LASSO* (b) *LAD-LASSO*

(c) *GM-LASSO* (d) *MM-LASSO*

Figure 5.11: Boxplots of estimates for six coefficients from 1000 simulated data sets, with normal distribution errors with mean 0 and variance 1 and , $\rho = 0$

Table 5.12: Parameter estimates and their MRPEs (bracketed) for simulated data sets with normal distribution errors with mean 0 and variance 1, where $\beta_1 = \beta_2 = \beta_3 = 1$, and $\beta_4 = \beta_5 = \beta_6 = 0$

Values of $\rho = 0.5$				
Parameter estimates	*Ada-LASSO*	*LAD-LASSO*	*GM-LASSO*	*MM-LASSO*
$\hat{\beta_1}$	0.9475(0.3026)	1.6560 (0.4991)	0.9543 (0.2162)	0.9809 (0.1909)
$\hat{\beta_2}$	0.9391(0.3599)	1.4631 (0.3478)	0.9750(0.2302)	1.0467(0.2758)
$\hat{\beta_3}$	0.7104 (0.4577)	0.4706(0.6258)	0.6979 (0.0532)	0.6979 (0.2695)
$\hat{\beta_4}$	0.1112(0.3846)	0.3144 (0.4722)	0.1791 (0.3773)	0.0735 (0.2352)
$\hat{\beta_5}$	0.0000(0.1486)	-0.2696 (0.1171)	0.0000 (0.0000)	0.0000(0.0000)
$\hat{\beta_6}$	-0.3065 (0.1506)	-0.2441(0.1292)	-0.2650(0.0000)	-0.1800(0.0000)



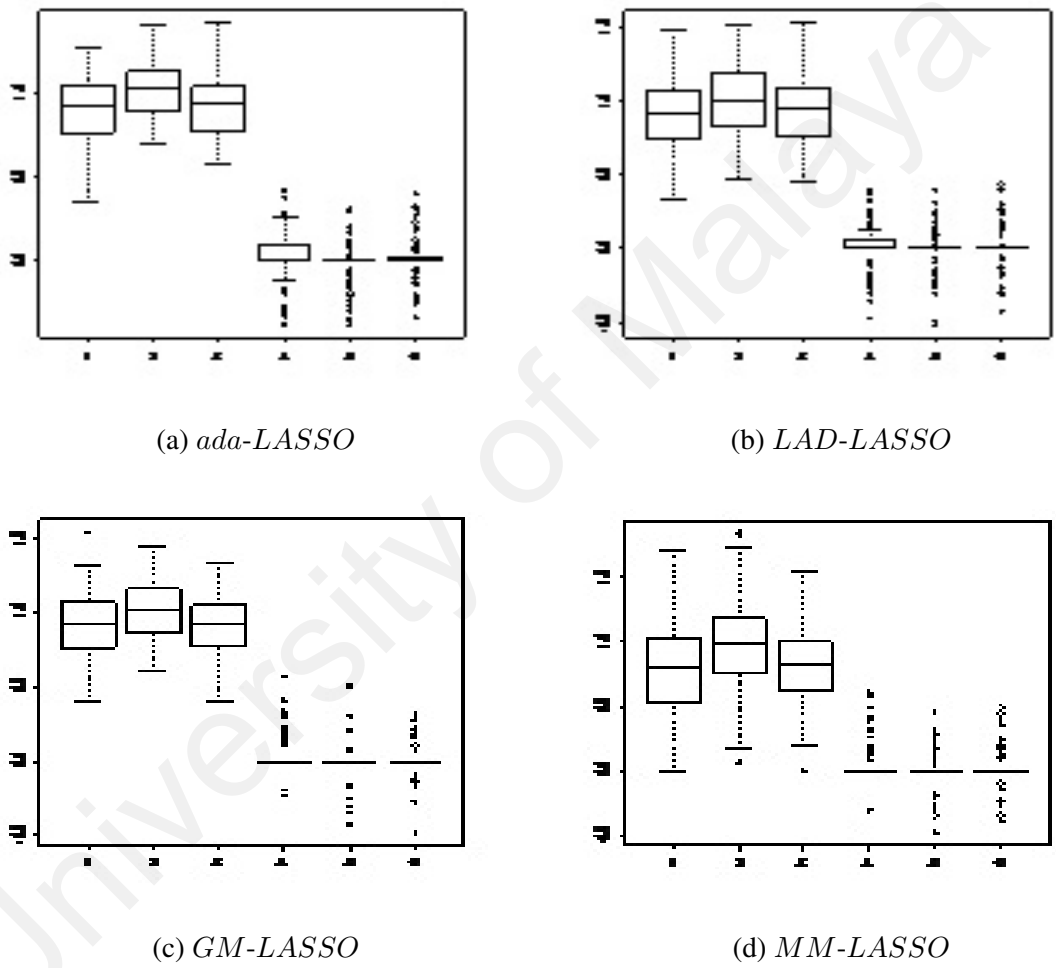
(a) *ada-LASSO* (b) *LAD-LASSO*

(c) *GM-LASSO* (d) *MM-LASSO*

Figure 5.12: Boxplots of estimates for six coefficients from 1000 simulated data sets, with normal distribution errors with mean 0 and variance 1 and , $\rho = 0.5$

Table 5.13: Parameter estimates and their MRPEs (bracketed) for simulated data sets with normal distribution errors with mean 0 and variance 1, where $\beta_1 = \beta_2 = \beta_3 = 1$, and $\beta_4 = \beta_5 = \beta_6 = 0$

	Values of $\rho = 0.8$			
Parameter estimates	*Ada-LASSO*	*LAD-LASSO*	*GM-LASSO*	*MM-LASSO*
$\hat{\beta}_1$	1.1900 (0.3026)	1.69721(0.4991)	1.2821(0.2162)	1.0645 (0.1909)
$\hat{\beta}_2$	0.8821 (0.3599)	0.8018(0.3478)	0.8927(0.2302)	0.7934(0.2758)
$\hat{\beta}_3$	0.8135(0.4577)	0.4525(0.6258)	0.9048(0.0532)	0.6012 (0.2695)
$\hat{\beta}_4$	0.2216 (0.3846)	0.2437(0.3773)	0.3192(0.4722)	0.5845 (0.2352)
$\hat{\beta}_5$	0.0000(0.1486)	0.0000 (0.1171)	0.0000(0.0000)	0.0000(0.0000)
$\hat{\beta}_6$	0.0000(0.1506)	0.0000(0.1292)	0.0000(0.0000)	0.0000(0.0000)



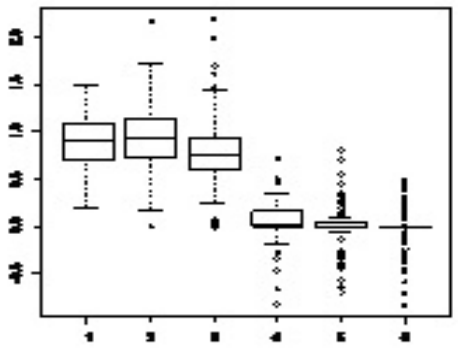
(a) *ada-LASSO* (b) *LAD-LASSO*

(c) *GM-LASSO* (d) *MM-LASSO*

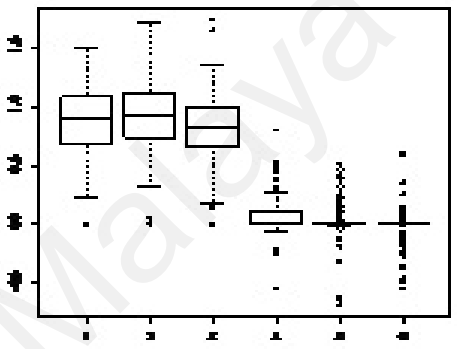
Figure 5.13: Boxplots of estimates for six coefficients from 1000 simulated data sets, with normal distribution errors with mean 0 and variance 1 and , $\rho = 0.8$

Table 5.14: Parameter estimates and their MRPEs (bracketed) for simulated data sets with Student's t-distribution errors with 3 degrees of freedom, where $\beta_1 = \beta_2 = \beta_3 = 1$, and $\beta_4 = \beta_5 = \beta_6 = 0$

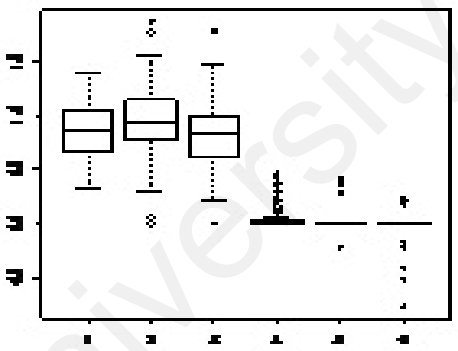
Values of $\rho = 0.0$				
Parameter estimates	$Ada\text{-}LASSO$	$LAD\text{-}LASSO$	$GM\text{-}LASSO$	$MM\text{-}LASSO$
$\hat{\beta_1}$	0.6788(0.1909)	0.3467(0.2162)	0.8699(0.1991)	0.9917(0.0026)
$\hat{\beta_2}$	0.5772(0.2758)	0.5763(0.2302)	0.8938(0.1478)	1.1246(0.0599)
$\hat{\beta_3}$	0.7029(0.2695)	0.6513(0.6258)	0.9649(0.0532)	1.0066(0.1577)
$\hat{\beta_4}$	0.2028(0.2352)	0.4769 (0.4722)	0.3590(0.1773)	0.0000(0.0846)
$\hat{\beta_5}$	0.2457(0.1486)	0.0000(0.1171)	0.0000(0.0000)	0.0000(0.0000)
$\hat{\beta_6}$	0.0000(0.1506)	0.0000(0.1292)	0.0000(0.0000)	0.0000(0.0000)



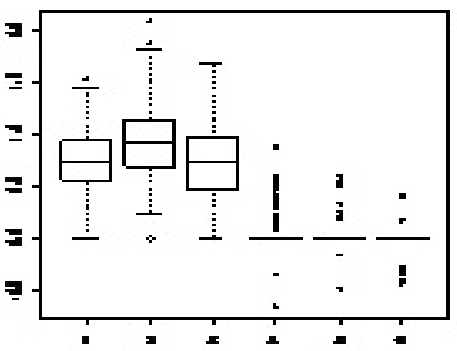
(a) $ada\text{-}LASSO$



(b) $LAD\text{-}LASSO$



(c) $GM\text{-}LASSO$

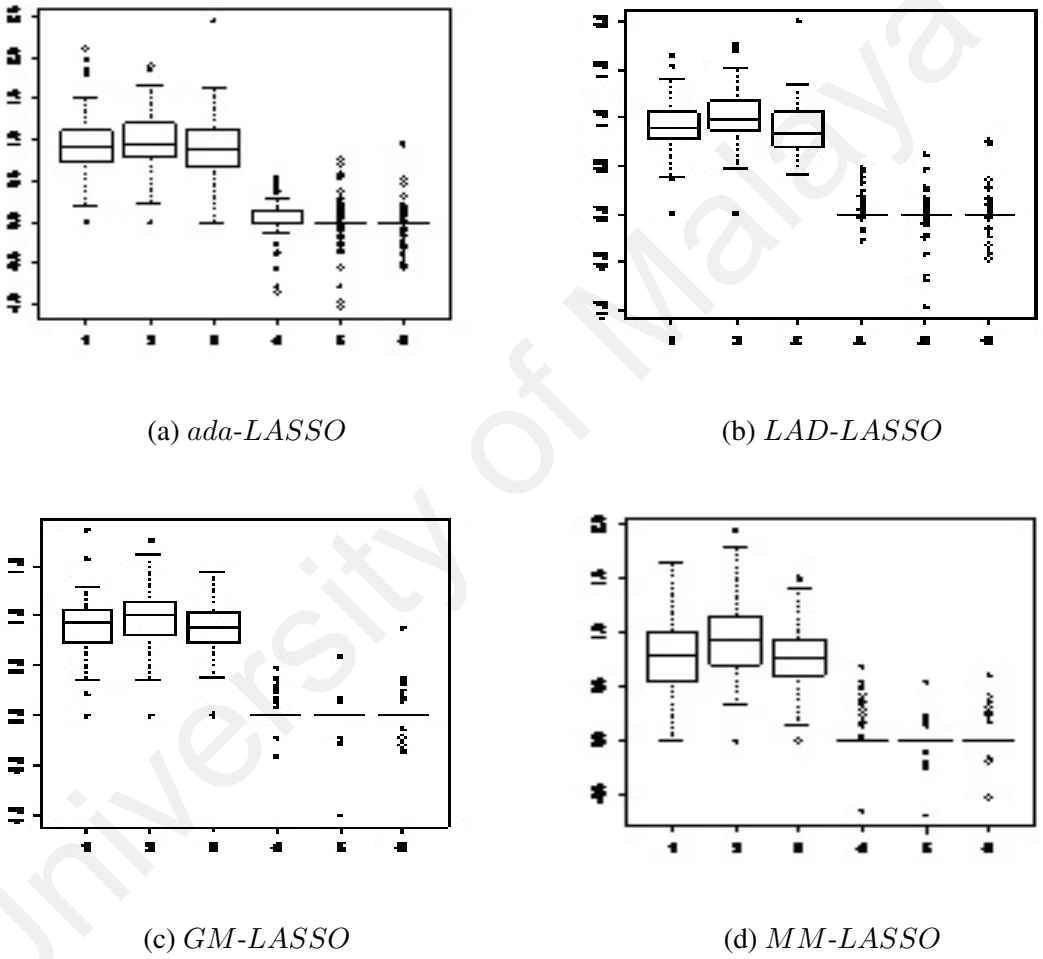


(d) $MM\text{-}LASSO$

Figure 5.14: Boxplots of estimates for six coefficients from 1000 simulated data sets, with Student's t-distribution errors with 3 degrees of freedom, $\rho = 0.0$

Table 5.15: Parameter estimates and their MRPEs (bracketed) for simulated data sets with Student's t-distribution errors with 3 degrees of freedom, where $\beta_1 = \beta_2 = \beta_3 = 1$, and $\beta_4 = \beta_5 = \beta_6 = 0$

Values of $\rho = 0.5$				
Parameter estimates	*Ada-LASSO*	*LAD-LASSO*	*GM-LASSO*	*MM-LASSO*
$\hat{\beta_1}$	0.7000 (0.6909)	0.6751(0.9162)	0.98333 (0.1991)	0.9164(0.1026)
$\hat{\beta_2}$	1.2811 (0.2758)	1.19134 (0.6302)	1.3450 (0.2078)	1.0649(0.1599)
$\hat{\beta_3}$	0.3098 (0.2695)	0.3238(0.6258)	0.9917 (0.0532)	0.9980(0.0577)
$\hat{\beta_4}$	0.2431 (0.2352)	0.1630 (0.4722)	0.2041 (0.1773)	0.0000(0.3846)
$\hat{\beta_5}$	0.0000 (0.1486)	0.0000 (0.1171)	0.0000 (0.0000)	0.0000(0.0000)
$\hat{\beta_6}$	0.1498 (0.1506)	0.0593(0.12929)	0.1547 (0.0000)	0.0000(0.0000)



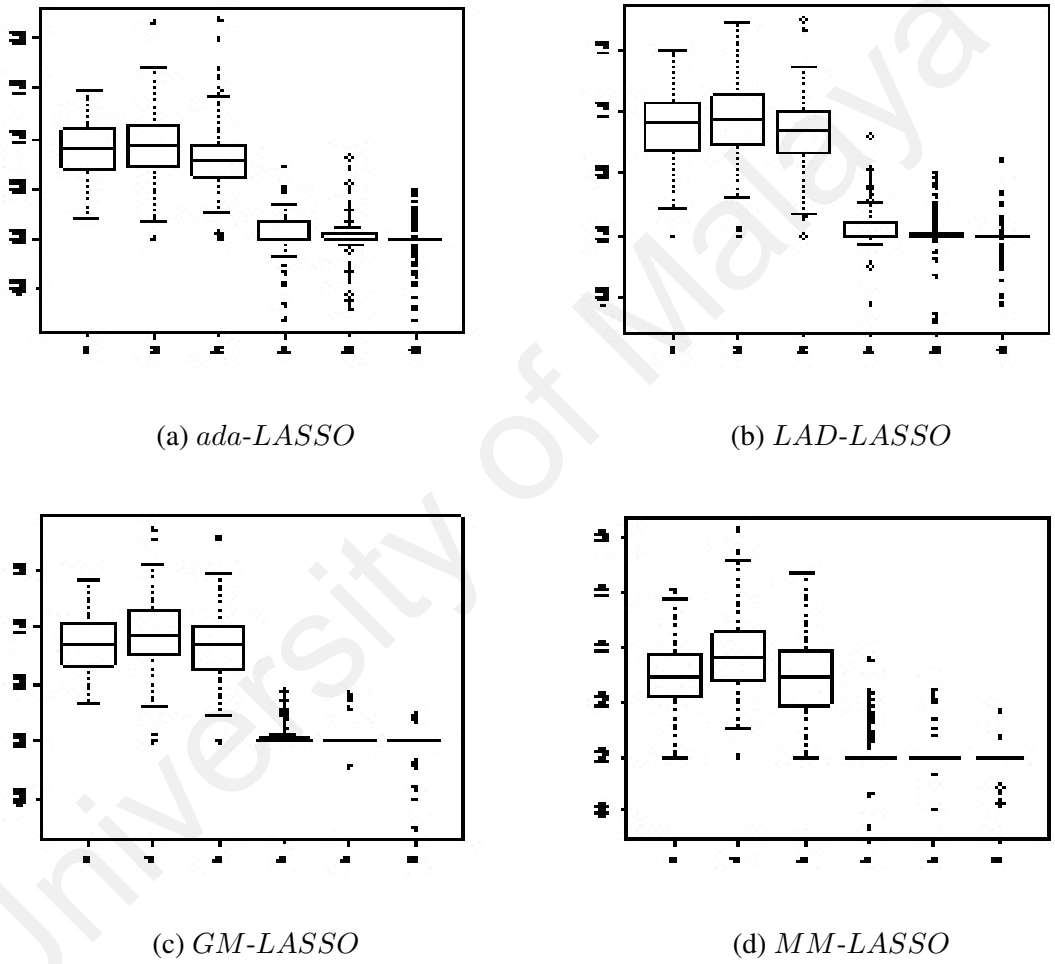
(a) *ada-LASSO* (b) *LAD-LASSO*

(c) *GM-LASSO* (d) *MM-LASSO*

Figure 5.15: Boxplots of estimates for six coefficients from 1000 simulated data sets, with Student's t-distribution errors with 3 degrees of freedom, $\rho = 0.5$

Table 5.16: Parameter estimates and their MRPEs (bracketed) for simulated data sets with Student's t-distribution errors with 3 degrees of freedom, where $\beta_1 = \beta_2 = \beta_3 = 1$, and $\beta_4 = \beta_5 = \beta_6 = 0$

Values of $\rho = 0.8$				
Parameter estimates	*Ada-LASSO*	*LAD-LASSO*	*GM-LASSO*	*MM-LASSO*
$\hat{\beta_1}$	0.3788 (0.6909)	0.2467(0.5162)	0.8699 (0.1991)	1.0517(0.1026)
$\hat{\beta_2}$	0.4772 (0.7758)	0.4763 (0.4302)	0.9638(0.1478)	1.1246(0.0599)
$\hat{\beta_3}$	0.4529 (0.6695)	0.6813(0.4258)	0.9649 (0.0532)	1.0066(0.1577)
$\hat{\beta_4}$	0.2028 (0.2352)	0.4769 (0.6722)	0.0590 (0.3773)	0.0000(0.0846)
$\hat{\beta_5}$	0.2457 (0.1486)	0.7900 (0.6171)	0.0000 (0.0000)	0.0000(0.0000)
$\hat{\beta_6}$	0.0000 (0.1506)	0.0000 (0.1292)	0.0000 (0.0000)	0.0000(0.0000)



(a) $ada\text{-}LASSO$ (b) $LAD\text{-}LASSO$

(c) $GM\text{-}LASSO$ (d) $MM\text{-}LASSO$

Figure 5.16: Boxplots of estimates for six coefficients from 1000 simulated data sets, with Student's t-distribution errors with 3 degrees of freedom, $\rho = 0.8$

5.9.2 Simulation Study ($p > n$)

In this simulation we examine the performance of the $GM\text{-}LASSO$ and $MM\text{-}LASSO$ for $p > n$ model data set. We simulated 1000 data sets each having $n = 15$ observations and $p = 20$ variables. The linear regression was used where $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3 = 1$ and

$\beta_j = 0$ for $j = 4, 5, ..., 20$. The X and ϵ come from standard normal distribution. The correlation between variables is $\rho = 0.5$. We considered three situations of data set. First, data with no outliers, second, data with 5% and 10% vertical, finally, data with 5% and 10% bad leverage. The median of mean square error (MSE) over 1000 simulated data set are summarized and the median number of zero coefficients are also reported. Moreover, we reported the boxplots of each situation.

Result and discussion

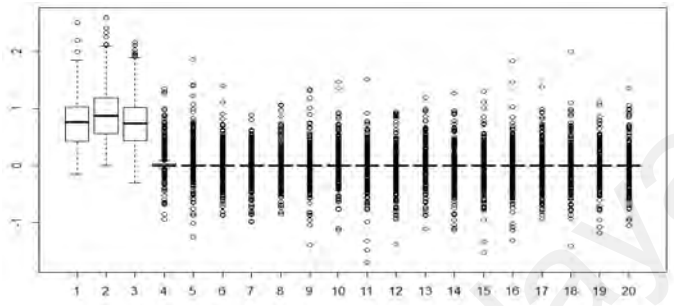
From Table 5.17 and Figure (5.17), we can see that the median (MSE) of the ada-$LASSO$ is 0.0636, where it is relatively smaller than the other estimators when the data is uncontaminated. As expected, the ada-$LASSO$ gave the best results here.

Tables 5.18 (data with verticals=5%) and 5.19 (data with verticals=10%) and Figures (5.18) (data with verticals=5%) and (5.19) (data with verticals=10%) show the result for data with verticals, where the MSE of ada-$LASSO$ in both 5% and 10% verticals were, 1.1012 and 3.2740, and estimated only 4 zero coefficients. The robust methods were better than the ada-$LASSO$. The Huber-$LASSO$ and GM-$LASSO$'s performance are almost as good as the MM-$LASSO$.

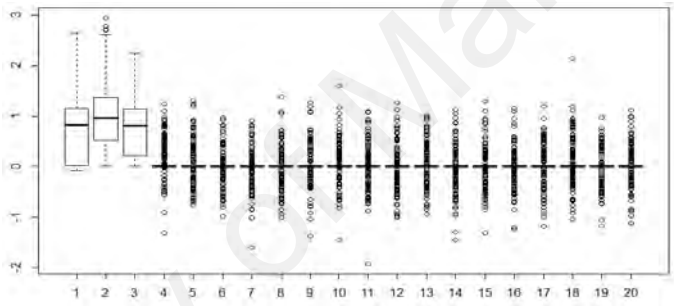
According to Tables 5.20 (data with bad leverage=5%) and 5.21 (data with bad leverage=10%), GM-$LASSO$ and MM-$LASSO$ perform better than all the other estimators. They selected approximately the correct number of zero coefficients which is 17, but suffer from too much variability as shown in the boxplots (Figures (5.20) and (5.21)). The ada-$LASSO$ estimator does poorly and has higher median mean squared error than other estimators.

Table 5.17: Result simulation when $p > n$ for no contaminated data sets

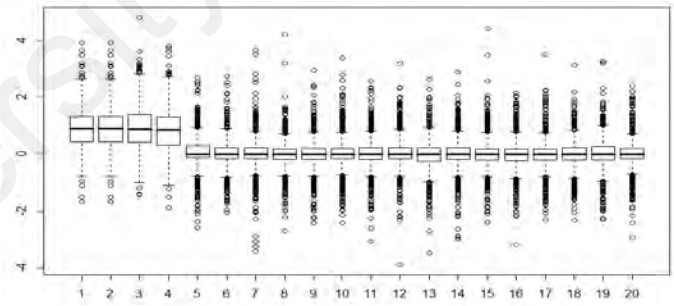
Methods	Median (MSE)	median no. of zero coefficients
ada-$LASSO$	0.0636	16
Huber-$LASSO$	0.4423	13
GM-$LASSO$	0.7805	5
MM-$LASSO$	0.7252	5



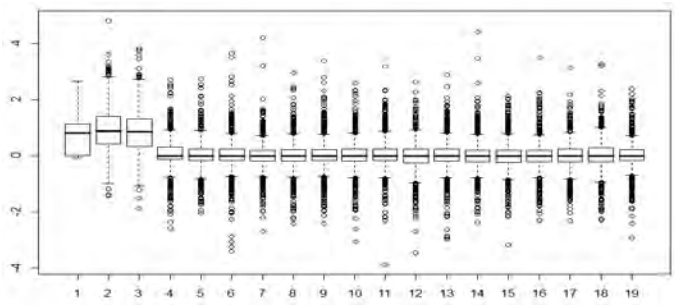
(a) ada-$LASSO$



(b) Huber-$LASSO$



(c) GM-$LASSO$

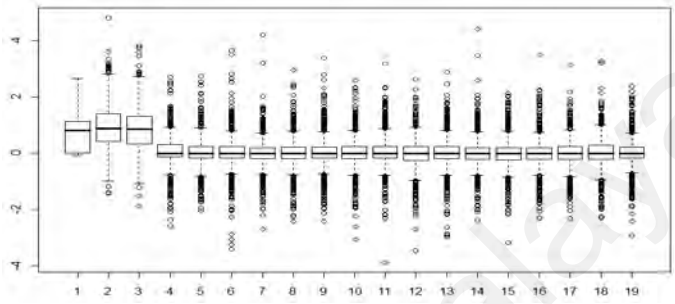


(d) MM-$LASSO$

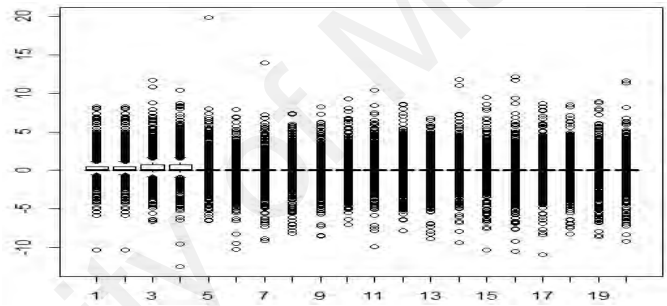
Figure 5.17: Boxplots of estimates for 20 coefficients with no contaminated simulated data sets

Table 5.18: Result simulation when $p > n$ for data set with 5% verticals

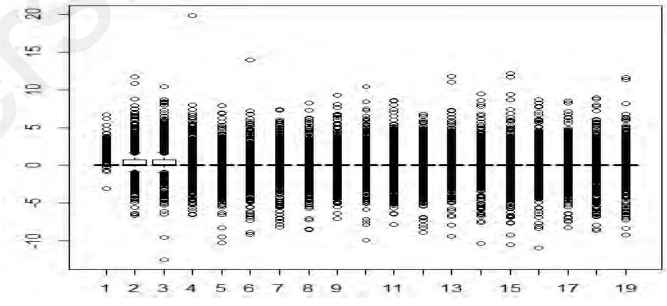
Methods	Median (MSE)	median no. of zero coefficients
ada-$LASSO$	1.1012	4
Huber-$LASSO$	0.2114	16
GM-$LASSO$	0.5281	12
MM-$LASSO$	0.0028	17



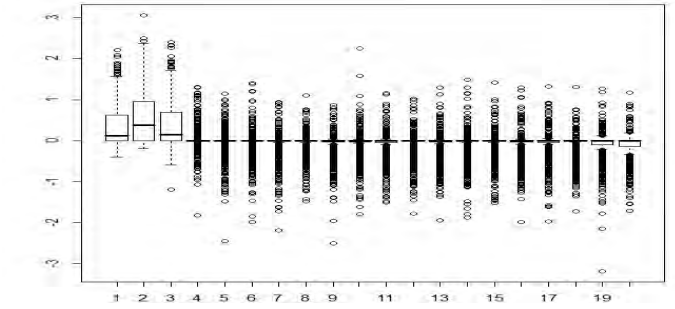
(a) ada-$LASSO$



(b) Huber-$LASSO$



(c) GM-$LASSO$

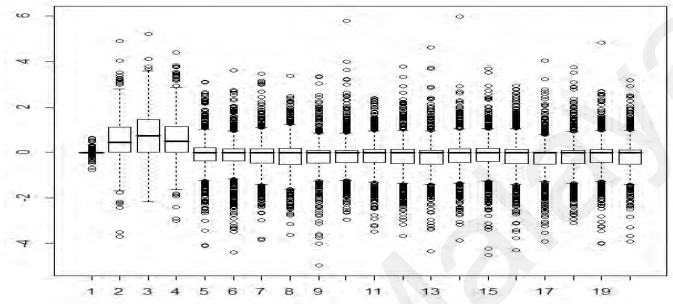


(d) MM-$LASSO$

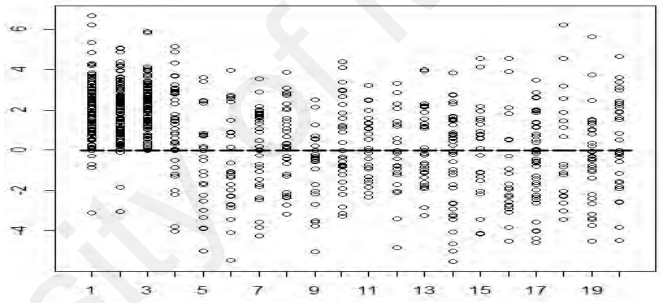
Figure 5.18: Boxplots of estimates for 20 coefficients with 5% verticals simulated data sets

Table 5.19: Result simulation when $p > n$ for data set with 10% verticals

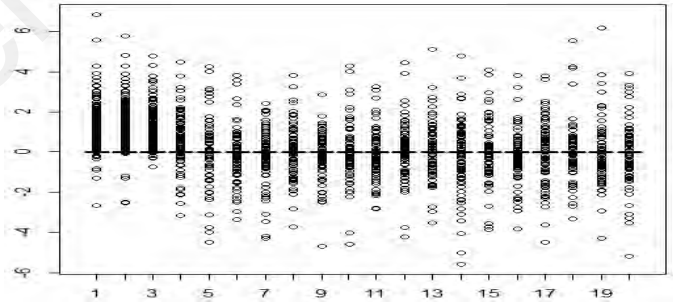
Methods	Median (MSE)	median no. of zero coefficients
$ada\text{-}LASSO$	3.2740	4
Huber-$LASSO$	0.3081	12
$GM\text{-}LASSO$	0.5245	12
$MM\text{-}LASSO$	0.0247	15



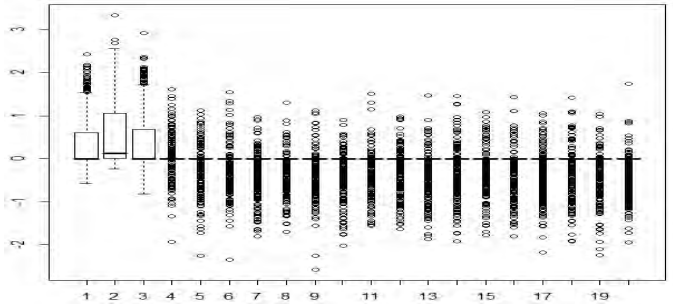
(a) $ada\text{-}LASSO$



(b) Huber-$LASSO$



(c) $GM\text{-}LASSO$

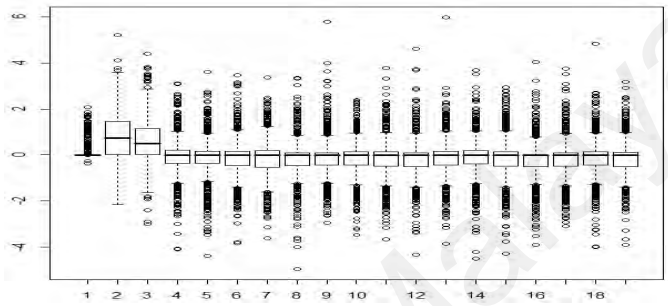


(d) $MM\text{-}LASSO$

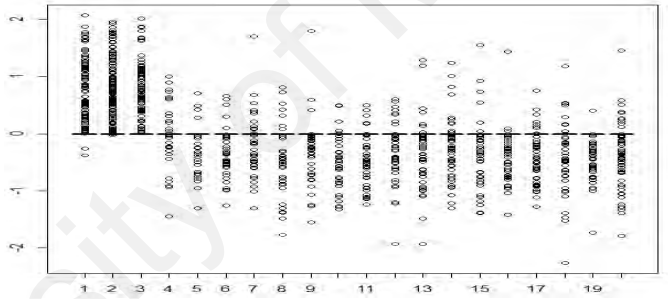
Figure 5.19: Boxplots of estimates for 20 coefficients with 10% verticals simulated data sets

Table 5.20: Result simulation when $p > n$ for data set with 5% bad leverage

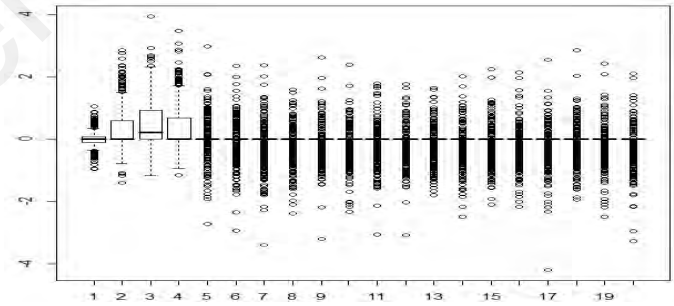
Methods	Median (MSE)	median no. of zero coefficients
$ada\text{-}LASSO$	4.278	4
Huber-$LASSO$	2.4127	19
$GM\text{-}LASSO$	o.6089	15
$MM\text{-}LASSO$	0.1090	17



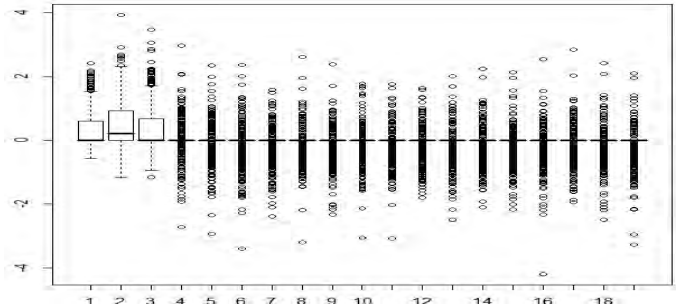
(a) $ada\text{-}LASSO$



(b) Huber-$LASSO$



(c) $GM\text{-}LASSO$

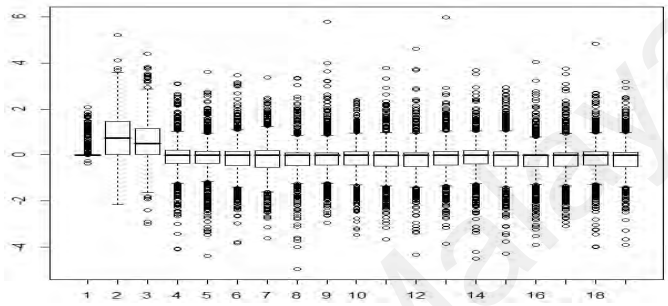


(d) $MM\text{-}LASSO$

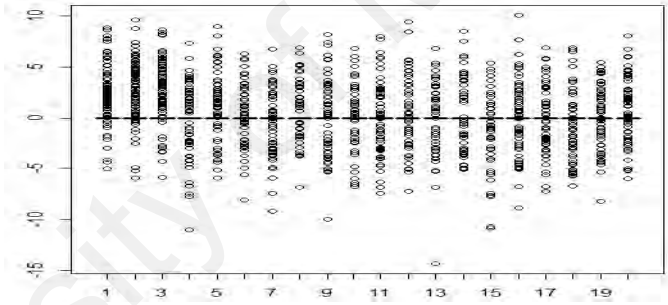
Figure 5.20: Boxplots of estimates for 20 coefficients with 5% bad leverage simulated data sets

Table 5.21: Result simulation when $p > n$ for data set with 10% bad leverage

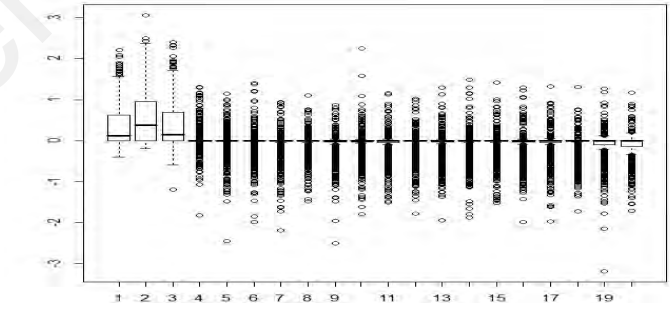
Methods	Median (MSE)	median no. of zero coefficients
$ada\text{-}LASSO$	4.7210	3
Huber-$LASSO$	4.4278	19
$GM\text{-}LASSO$	0.6424	15
$MM\text{-}LASSO$	0.1127	16



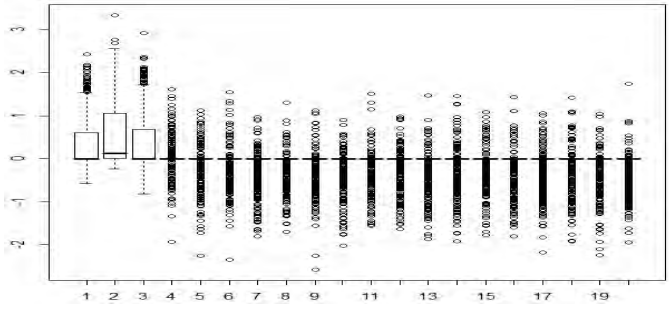
(a) $ada\text{-}LASSO$



(b) Huber-$LASSO$



(c) $GM\text{-}LASSO$



(d) $MM\text{-}LASSO$

Figure 5.21: Boxplots of estimates for 20 coefficients with 10% bad leverage simulated data sets

5.10 Practical Example

This section applying the proposed model selection methods on two data sets; the Ozone data, and Prostate Cancer data.

5.10.1 Ozone data

This data have been described in Section 3.6. The correlation matrix in Table 3.10 suggested that, certain correlation is present between the covariances. For example, the pairwise coefficient is 0.808 between Temperature (temp) and millibar pressure height (milPress), 0.864 between temp and Inversion base temperature (invTemp), and 0.647 between Pressure gradient (press) and Humidity (hum), and so on.

We fit the following model with 8 candidate predictors:

$$Ozone = \beta_0 + \beta_1 temp + \beta_2 invHt + \beta_3 press + \beta_4 vis + \beta_5 milPress$$

$$+\beta_6 hum + \beta_7 invTemp + \beta_8 wind.$$

The following methods were applied for comparison: $LASSO$, $ada\text{-}LASSO$, $LAD\text{-}LASSO$, Huber-$LASSO$, $GM\text{-}LASSO$, and $MM\text{-}LASSO$. The prediction accuracy of these methods were measured by compute the root mean squared prediction error ($RMSPE$) given by $RMSPE(\hat{\boldsymbol{\beta}}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y - \mathbf{X}\hat{\boldsymbol{\beta}})^2}$. The optimal value λ is selected using cross-validation for $LASSO$ and $ada\text{-}LASSO$ methods, whereas λ is selected using BIC for other methods.

For comparison purposes, the results of the model based on the LS and MM estimators are also reported; these are summarized in Table 5.22.

Discussion

Table 3.10 demonstrates the correlation results among $(i, j)^{th}$ coefficients among which the largest correlation is between "invTemp" , "milpress", and "temp".

Table 5.22 shows that all parameters of the LS and MM models are nonzero and their $RMSPE$ are large ($RMSPE_{LS} = 5.9123$ and $RMSPE_{MM} = 5.1248$) and substantially worse than all other variable selection methods. Furthermore, both non-robust $LASSO$ ($LASSO$ and ada-$LASSO$) select a model with six explanatory variables and three zero variables.

For robust Huber-$LASSO$ and GM-$LASSO$ methods, the number of selected variables is 4 variables, lower than for the LAD-$LASSO$ criteria which selected 6 variables. For MM-$LASSO$ method, the number of selected variables is 3, lower than for the other criteria with small value of $RMSPE$ (3.7297). Based on the smallest value of $RMSPE$, the MM-$LASSO$ is the best method here.

Table 5.22: Estimation results of Ozone data

Variable	LS	MM	$LASSO$	ada-$LASSO$	LAD-$LASSO$	Huber-$LASSO$	GM-$LASSO$	MM-$LASSO$
intercept	-1.1681	-0.5965	-1.6758	-1.8422	-0.0124	1.8786	3.3786	-0.9643
temp	18.6244	18.7043	17.9711	18.5883	19.8520	16.8129	15.2957	18.3841
invHt	-2.5980	-2.9994	-2.9138	-3.2466	-4.3513	-2.8686	-3.4073	0
press	0.2766	0.1918	0	0	0	0	0	0
vis	-2.2896	-2.2520	-1.5399	-1.7336	-1.4379	0	0	0
milPress	-4.3264	-3.7909	0	0	0	0	0	0
hum	5.2074	5.1369	5.4193	5.8341	5.0834	2.5183	1.7916	3.0894
invTemp	9.0848	7.1938	5.5373	5.2498	0	0	0	0
wind	1.4159	1.9789	0	0	2.6972	0	0	0
RMSPE	**5.9123**	**5.1248**	**4.6744**	**4.4923**	**4.5277**	**4.4334**	**4.4578**	**3.7297**

5.10.2 Prostate Cancer Data

The prostate cancer data come from a study by Stamey et al. (1989), the study had a total of 97 observations of male patients aged from 41 to 79 years. Table 5.23 gives an overview of the variables included in the data.

The response variable is the log(prostate specific antigen) (denoted by lpsa). The explanatory variables are log(cancer volume) (lcavol), log(prostate weight) (lweight), age, log(benign prostatic hyperplasia amount) (lbph), seminal vesicle invasion (svi), log(capsular penetration) (lcp), gleason score (gleason), percentage gleason scores 4 or 5 (pgg45), and log(prostate specific antigen) (lpsa).

Discussion

Table 5.24 demonstrates correlation results among $(i, j)^{th}$ coefficients which the largest correlation is between "icp" , "svi", and "pgg45".

Tibshirani (1996) applied $LASSO$ for this data set, whereas in this example, robust $LASSO$ selection methods (such as LAD, Huber, GM, and MM based methods) applied to the Prostate Cancer data.

The results based on non-robust and robust $LASSO$ are reported (see Table 5.25). $RMSE$ for non sparse methods (LS and MM) is larger ($RMSE_{LS} = 0.6747$ and $RMSE_{MM} = 0.6906$) and worse than other variable selection methods. A similar results as in Tibshirani (1996) are obtained by both Huber-$LASSO$ and MM-$LASSO$ (selected, (icavol), (iweight), (ibph), (svi), and (pgg45)).

For $GM\text{-}LASSO$, we obtained $RMSE_{=}0.7116$, with three zeros variables, so $GM\text{-}LASSO$ is superior here.

Table 5.23: Variables of the Prostate Cancer data

Name	**Description**
lcavol	log(cancer volume).
lweight	log(prostate weight).
age	age.
lbph	log(benign prostatic hyperplasia amount).
svi	seminal vesicle invasion.
lcp	log(capsular penetration).
gleason	Gleason score.
pgg45	percentage Gleason scores 4 or 5.
lpsa	log(prostate specific antigen).

Table 5.24: The correlation results among the $(i, j)^{th}$ of the Prostate Cancer data

	icavol	iweight	age	ibph	svi	icp	gleason	pgg45
icavol	1.0000							
iweight	0.1941	1.0000						
age	0.2250	0.3075	1.0000					
ibph	0.0273	0.4349	0.3502	1.0000				
svi	0.5388	0.1088	0.1177	-0.0858	1.0000			
icp	0.6753	0.1002	0.1277	-0.0070	0.6731	1.0000		
gleason	0.4324	-0.0013	0.2689	0.0778	0.3204	0.5148	1.0000	
pgg45	0.4337	0.0508	0.2761	0.0785	0.4576	0.6315	0.7519	1.0000

Table 5.25: Estimation Results of Prostate Cancer Data

Variable	LS	MM	$LASSO$	ada-$LASSO$	LAD-$LASSO$	Huber-$LASSO$	GM-$LASSO$	MM-$LASSO$
intercept	0.6694	0.7088	0.4339	0.0677	-0.0565	0.3280	2.4472	0.5809
icavol	0.5870	0.5870	0.5113	0.5628	0.5482	0.5131	0.5568	0.3418
iweight	0.4545	0.4403	0.3292	0.4161	0.4772	0.3531	0	0.3388
age	-0.0196	-0.0204	0	0	-0.0228	0	-0.0138	0
ibph	0.1071	0.1309	0.0421	0.0139	0.1604	0.0525	0.1303	0.0767
svi	0.7662	0.7919	0.5436	0.5993	0.7777	0.5571	0.6551	0.6361
icp	-0.1055	-0.1301	0	0	-0.0938	0	0	0
gleason	0.0451	0.0486	0	0	0.1826	0	0	0
pgg45	0.0045	0.0056	0.0012	0	0.0022	0.0016	0.0014	0.0020
RMSE	0.6747	0.6761	0.6906	0.6926	0.6827	0.6963	0.7116	0.7318

5.11 Summary

None- sparse estimators like GM- and MM- estimation are widely used in robust regression models. However, these estimators do not allow sparse model estimates and cannot be applied to data when $p > n$. In this chapter, we present our proposed GM-$LASSO$ and MM-$LASSO$ methods for improving the robustness of the adaptive and Huber-based $LASSO$ methods. The GM-$LASSO$ and MM-$LASSO$ combine the properties of the GM and MM regression method and the adaptive $LASSO$ penalty. The simulation results and the application in real data clearly show that the MM-$LASSO$ performs better than the other methods mentioned earlier. Moreover, the proposed methods perform similarly to the Huber-$LASSO$ given a data set with outliers and perform better than the Huber-$LASSO$ given a data set with high leverage points. However, the use of the Huber-$LASSO$ is discouraged for data sets that are highly contaminated with tailed errors. Considering the ease in the computation of the GM-$LASSO$ method using weighting data and Huber-$LASSO$ regression algorithms, the GM-$LASSO$ and MM-$LASSO$ methods offer several advantages.

CHAPTER 6

A DIAGNOSTIC-ROBUST MODEL SELECTION PROCEDURES

6.1 Introduction

Section 3.2 pointed out that most of the commonly used variable selection techniques for model building are affected in the presence of vertical and high leverage points, and often could produce very misleading conclusions. A robust version of this estimator is produced by replacing the ordinary squared residuals (LS) by a function $\rho(.)$, of residuals. Hence, distinguishing outliers and high leverage points is important in variable selection procedures analysis. This chapter aims to propose robust variable selection methods where the suspected outliers and high leverage points are identified by regression diagnostics tools; the best variables are then selected after performing diagnostic checks. The usefulness of our newly proposed methods is compared with the classical non-robust criteria and the existing criteria, based on M-estimators through simulations and real data sets.

6.2 Diagnostic-Regression Variable Selection Procedures

6.2.1 Variable Selection Methods in Small Samples with Diagnostic Tool

Akaike information criterion (AIC) (Akaike, 1998), Mallows' Cp Mallows (1973b), and Schwartz criterion (SIC) (Schwarz et al., 1978), powerful criteria for variables selection are defined as follows:

$$Z = G(SSE) + c. \tag{6.1}$$

Here the $G(SSE)$ is a function in terms of the sum of square error, $SSE = \sum_{i=1}^{n} r_i^2$, with residual $r_i = y_i - \mathbf{x}_i^T\boldsymbol{\beta}$ and c is a constant.

The $G(SSE)$ value equals $\log(SSE/n)$, SSE/σ^2, and $\log(SSE/n)$ for the AIC, Cp and SIC respectively, where n is the sample size. In the classical criteria, the $\hat{\boldsymbol{\beta}}$ is the LS-estimator corresponding to the traditional square function.

A general idea to outlier diagnostic (see, Section 2.3.2) is to form a clean subset of data that is free of outliers, . Let R be the set of indexes of the observations in the clean subset, y_R and $\mathbf{x}_R$ be the subsets of observations indexed by R, $\hat{\boldsymbol{\beta}}_R$ are the estimated regression coefficients computed from fitting the model to the set R. And let SSE_R be the corresponding sum of squares residual that finds the estimates corresponding to the clean samples having the smallest sum of squares of residuals. This study suggests using SSE_R in different model selection criteria.

The diagnostic version of model selection criteria

Consider the diagnostic sum of squares error SSE_R, by replacing the value of SSE in Eqn. (6.1) in terms of SSE_R , the criteria in Eqn. (6.1), can be expressed as follows:

$$Z_R = G(SSE_R) + c, \tag{6.2}$$

where SSE_R is compute from the diagnostic-LS (LS_R) estimator defined as:

$$\hat{\boldsymbol{\beta}}_{LS_R} = \arg\min \sum_{i=1}^{R} (r_R^2(\boldsymbol{\beta}_R))_i. \tag{6.3}$$

Therefore, LS_R corresponds to find the clean subset of R observations whose least squares fit produces the lowest sum of squared residuals, and has a high breakdown point. It is resistant to outliers, including leverage points. In Eqn. (6.2), the estimates corresponding to the R samples are having the smallest sum of residuals. This would be the most direct implementation of the idea that one wants to find the model which fits best for the majority of the data. However, the distributional properties of LS residuals are much better understood.

6.2.2 Variable Selection Methods in Large Data Sets Through Diagnostic ada-$LASSO$

A variable selection and regularized version of the diagnostic-LS is obtained by adding an L_1 penalty with penalty parameter λ to Eqn. (6.3), leading to the diagnostic-$LASSO$ ($LASSO_R$) estimator

$$\hat{\boldsymbol{\beta}}_{LASSO_R} = \arg\min_{\boldsymbol{\beta}_R} \sum_{i=1}^{R} (y_i - \mathbf{X}^T \boldsymbol{\beta}_R)^2 + \lambda |\boldsymbol{\beta}_R|, \tag{6.4}$$

define weight vector $\hat{w}_R = 1/|\hat{\boldsymbol{\beta}}_R|$. The ada-$LASSO_R$ estimates $\hat{\boldsymbol{\beta}}_{ada-LASSO_R}$ are given by

$$\hat{\boldsymbol{\beta}}_{ada-LASSO_R} = \arg\min \sum_{i=1}^{R} (r_R^2(\boldsymbol{\beta}_R))_i + \lambda \sum_{j=1}^{p} \hat{w}_{jR} |\beta_{jR}|. \tag{6.5}$$

The ada-$LASSO_R$ has a high breakdown point. It is resistant to outliers, including leverage points.

6.2.3 Breakdown Point of Diagnostic Variable Selection Methods

The breakdown point of the diagnostic model selection with subset size $n_R \leq n$ is given by

$$\varepsilon^{\star}(\hat{\boldsymbol{\beta}}_R; Z_R) = (n - n_R + 1)/n. \tag{6.6}$$

We suggest to take a value of R equal to a fraction α of the sample size, with $\alpha = 0.75$, such that the final estimate is based on a sufficiently large number of observations. This guarantees a sufficiently high statistical efficiency, The resulting breakdown point is then about $(1 - \alpha)100\% = 25\%$. Notice that the breakdown point does not depend on the dimension p. Even if the number of predictor variables is larger than the sample size, a high breakdown point is guaranteed.

Applying Eqn. (6.6) to the LS and to the $ada\text{-}LASSO$ ($n - n_R = 0$) yields a finite sample breakdown point of

$$\varepsilon^{\star}(\hat{\boldsymbol{\beta}}_{LS}; Z) = 1/n, \tag{6.7}$$

and

$$\varepsilon^{\star}(\hat{\boldsymbol{\beta}}_{ada-LASSO}; Z) = 1/n. \tag{6.8}$$

However, only one outlier can already send the variable selection and $ada\text{-}LASSO$ values to infinity, this non-robust variable selection comes from the use of squared residuals. Using other convex loss functions, as done in the robust variable selection using M-estimators and $LAD\text{-}LASSO$, does not solve the problem and results in a breakdown point of $1/n$ as well. The theoretical results on robustness are also reflected in the application to the generated data in Section 6.3, where the classical variable selection are much more influenced by the outliers than the diagnostic model selection methods.

6.3 Simulation

6.3.1 Simulation Example 1 (Small Data Set)

A simulation study was carried out to investigate the performance of the AIC_R, Cp_R, and SIC_R statistic for detecting best variables in the regression model based on Equations

$$AIC_R = \log(SSE_R/n_R) + 2p, \tag{6.9}$$

$$Cp_R = SSE_{R_p}/\hat{\sigma}^2_{full} - n_R + 2p, \tag{6.10}$$

$$SIC_R = \log(SSE_{R_p}/n_R) + (p\log(n_R))/n_R. \tag{6.11}$$

The simulation was based on three following aims:

(a) AIC_R are compared with non-robust AIC and robust based on M-estimation $RAIC$.

(b) Cp_R are compared with non-robust Cp and robust based on M-estimation RCp.

(c) SIC_R are compared with non-robust SIC and robust based on M-estimation $RSIC$.

In this simulation, 50 independent replicates of $p = 3$ independent uniform random variables on [-1,1] of $\mathbf{x}_{i1}$, $\mathbf{x}_{i2}$ and $\mathbf{x}_{i3}$, and 50 independent normally distributed errors $\varepsilon_i \sim N(0,1)$ were generated. The true model is given by $y_i = \mathbf{x}_{i1} + \mathbf{x}_{i2} + \varepsilon_i$, for $i = 1, ..., n$, $n = 50$ and $n = 100$ using two variables $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$. In order to illustrate the robustness to outliers, the following cases were considered:

1. Vertical outliers (outliers in the response only),
2. Bad leverage points (outliers on the covariates),
3. Good leverage points (outliers in outliers in **X** follows the the pattern of the majority of the data).

For vertical outliers case, we randomly generated different percentages of outliers (0%, 5%, 10%, 20%, 30%, and 40%) from $N(50, 0.1^2)$ for each of the simulated cases. For good leverage case, we considered the different percentages of outliers (0%, 5%, 10%, 20%, 30%, and 40%) on the variables X_1 and X_2 were generated from a $N(100, 0.5^2)$ distribution, then generated y to get good leverage points. For bad leverage case, different percentages of outliers (0%, 5%, 10%, 20%, 30%, and 40%) on variables X_1 and X_2 are generated from a $N(100, 0.5^2)$ distribution. For each of these setting 1000 samples were simulated.

Simulation results

A summary of the simulation results is provided by reporting the proportions of selected models that are

1. Correct fit, the true model only (x_1 and x_2).

2. Over fit, models containing all the variables in the true model plus some more that are actually redundant.

3. Under fit, models with only a strict subset of the variables in true model.

4. Wrong fit, all models that are not over fit, not a correct fit nor under fit. These are the models where some of the relevant variables might be present (though not all of them) in addition to some of the redundant variables.

We first consider the vertical outliers case with outlying response values. Table 6.1 shows detailed simulation results for one of the simulation setting with all variable selection criteria, AIC, CP, and SIC methods. As expected, the classical criteria work better than the robust criteria for the data without outliers ($AIC = 82.2\%$, $Cp = 81.0\%$, and

$SIC = 88\%$ with true model). The classical criteria select a large proportion of under fit or wrong fit models for the data with outliers, as shown in rows 2 to 5 in Table 6.1.

While a higher proportion of under fit and correct fit models are selected by robust criteria based on M-estimator with at most 20% contamination level, a higher proportion of correct fit models are selected by diagnostic criteria methods (AIC_R, Cp_R, and SIC_R). All of these methods work better for the cases with contamination level of outliers and break down at 40% of outliers in data. Similar results are obtained as the sample size increases, for example see Table 6.4 for $n = 100$.

In the presence of good leverage points, (see Tables 6.3 and 6.6), a high proportion of correct fit models selected classical AIC. The methods based on RCp and $RSIC$ provide good fit estimate with large sample size, but tend to over fit for small sample sizes. The diagnostic criteria methods performing well for any percentage of good leverage points.

In the presence of bad leverage point, the classical model selection criterion based on LS- and robust criteria based on M-estimation often select the high proportion of over fit or wrong fit model in this case. Interestingly, the diagnostic tool based methods tend to correctly fit the true model more often.

The simulation results illustrate that the performance of the proposed method (AIC_R, Cp_R, and SIC_R) yields a comparable power of selection, correct fit of those obtain in classical or $RAIC$, RCp, and $RSIC$ approaches for both cases in presence of vertical and leverage points.

Table 6.1: Percentage of times, a model is selected using (i) M-estimation, (ii) diagnostic method and (ii) vertical outliers, $n = 50$

ϵ (%)	No. of	Set of (Variables)	AIC_{LS}	$RAIC$	AIC_R	Cp_{LS}	RCp	Cp_R	SIC_{LS}	$RSIC$	SIC_R
0	0	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	0.4	21.2	3.4	0.6	8.0	3.8	2.4	10.4	5.6
		x_2	1.2	21.6	3.6	1.2	8.8	0.0	3.4	11.2	7.4
		x_3	0.2	2.0	0.2	0.2	3.2	0.4	0.2	3.8	0.4
		x_1, x_2	**82.2**	**54.4**	**59.0**	**81.0**	**36.6**	**62.4**	**88.2**	**40.4**	**63.8**
		x_1, x_3	0.0	0.0	2.0	0.0	6.6	2.2	0.0	5.8	3.4
		x_2, x_3	0.2	0.0	2.2	0.2	8.0	0.0	0.2	6.6	1.8
		x_1, x_2, x_3	15.8	0.0	29.4	16.8	28.8	31.2	5.6	21.8	17.6
5	2	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	6.6	23.2	1.8	6.8	5.8	2.2	16.6	7.8	6.6
		x_2	7.2	25	2.0	8.2	8.8	0.0	15.4	12.6	6.0
		x_3	0.2	2.0	0.4	0.4	2.8	1.2	0.4	3.6	1.0
		x_1, x_2	**70.8**	**49.8**	**63.2**	**70.4**	**34.8**	**64.4**	**63.2**	**39.6**	**67.4**
		x_1, x_3	1.2	0.0	1.0	1.2	7.6	1.0	0.8	6.8	1.0
		x_2, x_3	0.4	0.0	1.4	0.2	8.6	0.0	0.2	7.4	2.2
		x_1, x_2, x_3	12.4	0.0	29.8	12.8	31.6	31.2	3.4	22.2	15.8
10	5	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	17.8	24.8	3.0	26.0	7.2	3.4	35.2	11.6	7.6
		x_2	20.2	26.2	3.0	30.6	5.4	0.0	37.8	9.0	6.2
		x_3	7.0	1.6	0.2	15.2	0.2	0.6	17.6	0.2	0.6
		x_1, x_2	**15.2**	**47.4**	**67.0**	**15.2**	**70.4**	**69.0**	**6.0**	**69.2**	**69.2**
		x_1, x_3	6.4	0.0	2.0	6.6	1.2	2.0	1.6	0.6	2.0
		x_2, x_3	3.8	0.0	1.6	4.0	1.8	0.0	1.4	1.8	2.4
		x_1, x_2, x_3	2.4	0.0	23.2	2.4	13.8	25.0	0.4	7.6	12.0
20	10	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	16.6	36.6	4.0	27.4	11.2	4.6	32.4	14.4	9.8
		x_2	21.0	34.2	5.0	35.4	11.0	0.0	41.0	14.6	10.0
		x_3	7.4	1	0.0	17.2	0.2	0.2	20.0	0.6	0.4
		x_1, x_2	**8.4**	**28.2**	**68.4**	**8.4**	**70.4**	**73.2**	**3.2**	**67.0**	**67.4**
		x_1, x_3	4.4	0.0	2.0	4.6	0.8	1.8	1.4	0.6	1.6
		x_2, x_3	4.8	0.0	1.8	5.0	0.4	0.0	1.4	0.4	1.4
		x_1, x_2, x_3	2.0	0.0	18.6	2.0	6.0	20.2	0.6	2.4	9.4
30	15	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	17.8	36.0	6.6	30.4	31.6	7.8	35.0	33.4	12.8
		x_2	18.4	41.6	7.4	33.6	35.8	0.0	37.8	37.0	14.8
		x_3	8.0	22.4	0.4	21.6	6.2	0.6	24.0	6.4	0.8
		x_1, x_2	**6.0**	**0.0**	**66.6**	**6.0**	**25.2**	**71.8**	**2.0**	**22.6**	**62.8**
		x_1, x_3	2.8	0.0	1.2	3.0	0.2	1.2	0.6	0.0	1.0
		x_2, x_3	4.2	0.0	2.2	4.2	0.8	0.0	0.6	0.6	1.2
		x_1, x_2, x_3	1.2	0.0	15.2	1.2	0.2	18.6	0.0	0.0	6.6
40	20	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	18.6	34.8	9.6	35.0	31.0	11.6	37.2	32.0	18.8
		x_2	17.6	35.8	7.0	31.2	33.8	0.0	35.6	35.4	15.4
		x_3	8.0	29.4	0.6	19.2	24.8	1.6	23.6	26.6	1.2
		x_1, x_2	**5.8**	**0.0**	**63.8**	**5.8**	**4.2**	**69.8**	**2.0**	**3.2**	**53.4**
		x_1, x_3	3.2	0.0	3.2	3.4	2.2	3.0	1.0	0.8	2.6
		x_2, x_3	4.0	0.0	2.2	4.0	2.8	0.0	0.6	1.4	1.8
		x_1, x_2, x_3	1.4	0.0	12.6	1.4	1.2	14.0	0.0	0.6	6.8

Table 6.2: Percentage of times, a model is selected using (i)classical, (ii) M-estimation and (iii) diagnostic method, with vertical outliers, with bad leverage points, $n = 50$

ϵ (%)	No. of Leverage	Set of Variables	AIC_{LS}	$RAIC$	AIC_R	Cp_{LS}	RCp	Cp_R	SIC_{LS}	$RSIC$	SIC_R
5	2	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	5.4	31.8	2.8	8.2	12.0	3.2	17.4	17.2	6.2
		x_2	3.8	35.2	3.4	9.4	11.6	0.0	19.4	18.4	7.6
		x_3	0.8	33.0	0.2	7.0	14.0	0.4	15.6	20.2	0.4
		x_1, x_2	**0.8**	**0.0**	**60.4**	**1.6**	**2.4**	**63.8**	**0.4**	**1.0**	**65.0**
		x_1, x_3	32.8	0.0	2.0	1.6	2.6	1.6	0.4	1.0	1.4
		x_2, x_3	36.4	0.0	2.4	1.6	2.8	0.0	0.6	1.8	2.4
		x_1, x_2, x_3	15.4	0.0	28.6	70.6	54.6	31.0	46.2	40.4	17.0
10	5	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	2.2	34.4	4.4	6.6	12.2	4.2	15.6	18.6	8.4
		x_2	3.0	31.6	5.0	7.2	13.0	0.0	16.4	18.6	10.0
		x_3	1.4	34.0	0.6	8.0	12.0	1.2	17.4	18.6	1.2
		x_1, x_2	**0.0**	**0.0**	**57.4**	**0.8**	**2.0**	**62.0**	**0.2**	**1.4**	**62.6**
		x_1, x_3	36.6	0.0	2.0	1.4	1.8	2.2	1.0	0.8	1.2
		x_2, x_3	36.6	0.0	3.4	1.4	2.2	0.0	0.0	1.6	3.4
		x_1, x_2, x_3	15.8	0.0	27.2	74.6	56.8	30.4	49.4	40.4	13.2
20	10	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	2.2	33.0	4.4	5.8	11.4	8.6	15.6	17.6	8.8
		x_2	2.4	34.4	3.2	10.6	12.2	0.0	19.6	19.2	8.6
		x_3	1.6	32.6	2.2	8.4	14.6	5.4	17.0	19.8	6.8
		x_1, x_2	**0.2**	**0.0**	**20.0**	**2.0**	**1.4**	**22.6**	**0.8**	**0.8**	**20.4**
		x_1, x_3	38.0	0.0	25.6	1.4	2.4	28.2	0.2	1.2	25.6
		x_2, x_3	35.0	0.0	20.0	1.2	2.4	0.0	0.2	1.2	19.8
		x_1, x_2, x_3	14.6	0.0	18.8	70.6	55.6	35.2	46.6	40.2	10.0

Table 6.3: Percentage of times, a model is selected using (i)classical, (ii) M-estimation and (iii) diagnostic method, with good leverage points, $n = 50$

ϵ %	No. of leverage	Set of variables	AIC_{LS}	$RAIC$	AIC_R	Cp_{LS}	Cp_M	Cp_R	SIC_{LS}	$RSIC$	SIC_R
5	2	$Intercept$	0	2.6	0	0	0	0	0	0	0
		x_1	0.0	25.3	3.4	0	1.0	3.2	0	1.2	6.4
		x_2	0.0	25.1	1.8	0	0.0	0.0	0	1.0	3.8
		x_3	0.0	2.0	0.0	0	0.2	0.2	0	0.4	0.6
		x_1, x_2	**85.6**	**47.0**	**64.2**	**0**	**0.2**	**65.8**	**0**	**0.0**	**70.4**
		x_1, x_3	0.2	0.0	0.8	0	0.2	1.0	0	0.2	1.8
		x_2, x_3	0.2	0.0	1.6	0	0.0	0.0	0	0.0	1.6
		x_1, x_2, x_3	14.0	0.0	28.2	100	98.4	29.8	100	97.2	15.4
10	5	$Intercept$	0	2.0	0	0	0	0	0	0	0
		x_1	0.0	26.6	2.4	0	0.0	2.6	0	0.0	5.0
		x_2	0.0	26.3	2.8	0	0.0	0.0	0	0.2	5.2
		x_3	0.0	0.0	0.0	0	0.2	0.2	0	0.2	0.2
		x_1, x_2	**81.0**	**44.2**	**62.8**	**0**	**0.0**	**67.2**	**0**	**0.0**	**70.2**
		x_1, x_3	0.2	0.0	1.2	0	0.0	1.0	0	0.0	1.4
		x_2, x_3	0.0	0.0	0.4	0	0.0	0.0	0	0.2	1.2
		x_1, x_2, x_3	18.8	0.0	30.2	100	99.8	29.0	100	99.4	16.8
20	10	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	0.0	30.0	2.6	0	0.0	2.6	0	0.0	6.0
		x_2	0.0	36.6	1.8	0	0.2	0.0	0	0.6	5.6
		x_3	0.0	1.0	0.2	0	0.0	0.4	0	0.0	0.6
		x_1, x_2	**85.2**	**32.4**	**66.8**	**0**	**0.0**	**70.2**	**0**	**0.0**	**71.6**
		x_1, x_3	0.0	0.0	1.2	0	0.0	1.2	0	0.0	1.2
		x_2, x_3	0.0	0.0	2.6	0	0.0	0.0	0	0.0	2.4
		x_1, x_2, x_3	14.8	0.0	24.6	100	99.8	25.6	100	99.4	12.6

Table 6.4: Percentage of times, a model is selected using (i)classical, (ii) M-estimation and (iii) diagnostic method, with vertical outliers, $n = 100$

ϵ (%)	No. of Verticals	Set of Variables	AIC_{LS}	$RAIC$	AIC_R	Cp_{LS}	RCp	Cp_R	SIC_{LS}	$RSIC$	SIC_R
0	0	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	0.0	47.2	0.2	0	1.8	0.4	0.0	3.4	1.2
		x_2	0.0	0.0	0.0	0	1.2	0.0	0.0	2.8	0.6
		x_3	0.0	4.2	0.0	0	0.2	0.0	0.0	0.2	0.0
		x_1, x_2	**83.6**	**48.6**	**69.2**	**84**	**75.2**	**68.8**	**96.0**	**84.0**	**85.2**
		x_1, x_3	0.0	0.0	0.4	0	0.6	0.4	0.2	0.4	0.2
		x_2, x_3	0.0	0.0	0.4	0	0.2	0.0	0.0	0.0	0.2
		x_1, x_2, x_3	16.4	0.0	29.8	16	20.8	30.4	3.8	9.2	12.6
5	5	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	13.0	46.2	0.0	15.0	1.4	0.0	31.2	2.8	1.4
		x_2	12.2	0.0	0.4	13.2	0.6	0.0	33.4	1.8	1.0
		x_3	0.8	4.6	0.0	1.6	0.0	0.0	3.4	0.0	0.0
		x_1, x_2	**53.6**	**49.2**	**73.2**	**54.6**	**80.4**	**74.6**	**30.8**	**88.4**	**88.6**
		x_1, x_3	2.2	0.0	0.0	2.2	0.0	0.0	0.2	0.0	0.2
		x_2, x_3	3.0	0.0	0.0	2.8	0.0	0.0	0.2	0.0	0.0
		x_1, x_2, x_3	10.8	0.0	26.4	10.6	17.6	25.4	0.8	7.0	8.8
10	10	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	18.0	44.8	0.0	22.0	1.6	0.2	37.4	2.4	0.8
		x_2	21.4	0.0	0.0	25.0	1.6	0.0	41.8	3.2	0.8
		x_3	1.8	6.4	0.0	3.6	0.0	0.0	5.8	0.2	0.0
		x_1, x_2	**35.6**	**48.8**	**77.2**	**35.6**	**82.4**	**77.2**	**12.4**	**88.4**	**90.6**
		x_1, x_3	3.8	0.0	0.4	3.8	0.0	0.2	1.2	0.0	0.0
		x_2, x_3	3.8	0.0	0.2	3.8	0.2	0.0	1.0	0.2	0.0
		x_1, x_2, x_3	6.2	0.0	22.2	6.2	14.2	22.4	0.4	5.6	7.8
20	20	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	22.6	45.8	0.0	30.8	1.4	0	42.4	3.6	0.8
		x_2	19.4	0.0	0.4	29.2	2.4	0	39.8	4.0	1.6
		x_3	2.4	6.0	0.0	8.8	0.0	0	11.2	0.0	0.0
		x_1, x_2	**19.4**	**48.2**	**85.4**	**19.4**	**89.2**	**85**	**5.2**	**89.4**	**93.0**
		x_1, x_3	4.4	0.0	0.0	4.4	0.0	0	0.2	0.0	0.0
		x_2, x_3	3.6	0.0	0.0	3.6	0.0	0	0.8	0.0	0.0
		x_1, x_2, x_3	3.8	0.0	14.2	3.8	7.0	15	0.4	3.0	4.6
30	30	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	23.2	37.4	0.2	33.0	14.2	0.0	39.8	17.2	1.8
		x_2	24.4	0.0	1.0	33.6	17.6	0.0	42.8	20.4	3.6
		x_3	6.2	22.4	0.0	11.2	3.6	0.0	13.6	4.2	0.2
		x_1, x_2	**13.8**	**40.2**	**81.2**	**13.8**	**62.8**	**83.2**	**3.4**	**57.8**	**89.6**
		x_1, x_3	2.4	0.0	0.0	2.4	0.4	0.0	0.4	0.2	0.2
		x_2, x_3	3.2	0.0	0.0	3.4	0.2	0.0	0.0	0.2	0.2
		x_1, x_2, x_3	2.6	0.0	17.6	2.6	1.2	16.8	0.0	0.0	4.4
40	40	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	20.4	35.0	1.0	33.6	36.6	1.2	42.6	37.6	4.6
		x_2	19.6	0.0	0.8	30.6	31.2	0.0	38.6	32.2	3.2
		x_3	5.0	28.6	0.0	11.2	20.4	0.0	14.6	21.2	0.2
		x_1, x_2	**15.0**	**36.4**	**81.0**	**15.0**	**9.4**	**83.6**	**3.6**	**8.2**	**86.8**
		x_1, x_3	3.4	0.0	0.6	3.4	0.8	0.6	0.4	0.2	0.0
		x_2, x_3	3.2	0.0	0.0	3.2	1.2	0.0	0.2	0.4	0.2
		x_1, x_2, x_3	3.0	0.0	16.6	3.0	0.4	14.6	0.0	0.2	5.0

Table 6.5: Percentage of times, a model is selected using (i)classical, (ii) M-estimation and (iii) diagnostic method, with bad leverage points, $n = 100$

ϵ (%)	No. of Leverage	Set of Variables	AIC_{LS}	$RAIC$	AIC_R	Cp_{LS}	RCp	Cp_R	SIC_{LS}	$RSIC$	SIC_R
5	5	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	0.2	32.8	0.0	1.0	4.6	0	8.0	10.6	0.4
		x_2	0.2	34.0	0.2	1.2	5.8	0	7.0	11.4	0.6
		x_3	0.0	33.2	0.0	1.0	4.8	0	6.6	10.8	0.0
		x_1, x_2	**0.0**	**0.0**	**70.8**	**1.2**	**1.2**	**70**	**0.0**	**0.4**	**87.4**
		x_1, x_3	28.2	0.0	0.0	0.4	1.4	0	0.2	0.6	0.0
		x_2, x_3	21.0	0.0	0.0	0.4	2.0	0	0.2	0.6	0.0
		x_1, x_2, x_3	49.6	0.0	29.0	94.8	80.2	30	78.0	65.6	11.6
10	10	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	0.0	34.6	0.0	1.6	5.2	0.0	6.8	12.2	1.4
		x_2	0.4	32.4	0.6	6.2	0.0	9.2	12.8	1.0	1.0
		x_3	0.4	33.0	0.0	1.2	6.2	0.0	8.2	11.8	0.0
		x_1, x_2	**0.0**	**0.0**	**71.2**	**0.6**	**1.4**	**72.6**	**0.0**	**0.4**	**87.6**
		x_1, x_3	26.4	0.0	0.4	0.6	2.4	0.4	0.0	0.6	0.4
		x_2, x_3	25.0	0.0	0.0	0.4	1.4	0.0	0.0	0.2	0.2
		x_1, x_2, x_3	47.2	0.0	27.8	93.8	77.2	27.0	75.8	62.0	9.4
20	20	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	0.4	33.6	0.4	2.2	5.6	3.4	8.6	8.6	2.8
		x_2	0.0	32.4	1.0	1.0	6.0	0.0	8.0	11.0	4.2
		x_3	0.4	34.0	0.6	2.6	6.0	3.4	8.6	13.4	6.2
		x_1, x_2	**0.0**	**0.0**	**13.4**	**0.0**	**0.6**	**14.0**	**0.0**	**0.6**	**15.8**
		x_1, x_3	24.8	0.0	26.6	0.2	1.4	26.8	0.2	0.6	30.0
		x_2, x_3	24.8	0.0	27.2	0.2	1.8	0.0	0.2	0.6	32.4
		x_1, x_2, x_3	49.2	0.0	28.6	93.8	78.6	52.4	74.4	65.2	8.6

Table 6.6: Percentage of times, a model is selected using (i)classical, (ii) M-estimation and (iii) diagnostic method, with good leverage points, $n = 100$

ϵ %	No. of leverage	Set of variables	AIC_{LS}	$RAIC$	AIC_R	Cp_{LS}	RCp	Cp_R	SIC_{LS}	$RSIC$	SIC_R
5	2	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	0.0	41.2	0.0	0.0	0.2	0.0	0.0	0.4	0.4
		x_2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2
		x_3	0.0	14.0	0.0	0.0	0.2	0.0	0.0	0.2	0.0
		x_1, x_2	**85.2**	**44.8**	**74.6**	**84.6**	**81.8**	**74.8**	**97.2**	**89.8**	**90.2**
		x_1, x_3	0.0	0.0	0.2	0.0	0.2	0.2	0.0	0.0	0.2
		x_2, x_3	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.2
		x_1, x_2, x_3	14.8	0.0	25.0	15.4	17.6	25.0	2.8	9.4	8.8
10	10	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	0.0	46.4	0.0	0.0	0.0	0.0	0	0.2	0.4
		x_2	0.0	0.0	0.0	0.0	0.0	0.0	0	0.2	0.4
		x_3	0.0	8.4	0.0	0.0	0.0	0.0	0	0.0	0.0
		x_1, x_2	**84.4**	**45.2**	**73.4**	**83.4**	**76.8**	**72.6**	**96**	**91.6**	**91.6**
		x_1, x_3	0.0	0.0	0.2	0.0	0.0	0.2	0	0.0	0.0
		x_2, x_3	0.0	0.0	0.2	0.0	0.0	0.0	0	0.0	0.2
		x_1, x_2, x_3	15.6	0.0	26.2	16.6	23.2	27.2	4	8.0	7.4
20	20	$Intercept$	0	0	0	0	0	0	0	0	0
		x_1	0.0	47.4	0.2	0.0	0.0	0.2	0.0	0.2	0.2
		x_2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		x_3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		x_1, x_2	**84.2**	**52.6**	**73.0**	**84.2**	**81.2**	**71.8**	**96.4**	**92.0**	**87.2**
		x_1, x_3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		x_2, x_3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		x_1, x_2, x_3	15.8	0.0	26.8	15.8	18.8	28.0	3.6	7.8	12.6

6.3.2 Simulation Example 2 (Large Data Set Using $GDFFITS$ Measure Diagnostic)

This section presents a simulation study for comparing the performance of various $LASSO$ estimators. In order to identify influential observations, $ada\text{-}LASSO_R$ was evaluated using the $GDFFITS$ measure proposed by Rahmatullah Imon (2005). In this simulation, the data was generated with less than 25% contamination. $ada\text{-}LASSO$ and $LAD\text{-}LASSO$ were also compared.

The simulations were performed in R-package 'parcor' (Kraemer and Schaefer, 2010) which was applied to compute the $ada\text{-}LASSO$ solution based on k-fold cross-validation. The initial weights for $ada\text{-}LASSO$ were computed from a $LASSO$ fit. A suitable value for the shrinkage tuning parameter 'λ' was selected by tenfold cross-validation. Moreover, the package 'quantreg' (Koenker, 2007) was used for $LAD\text{-}LASSO$ and λ was chosen by applying the classical BIC criterion.

Next, the covariates, $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)$ were generated from n independent Gaussian vectors with correlation $r = 0.5$ and $r = 0.8$, $p = 8$ and $n = 200$. The parameter vector $\boldsymbol{\beta}_{true} = (3, 1.5, 0, 2, 0, 0, 0, 0)$. The response variable y, was then produced according to the regression model where the error terms follow a standard normal distribution. In order to investigate the robustness of the methods against outliers, three situations were considered:

1. No contaminated,
2. Vertical contamination (outliers on the response variables),
3. Bad leverage points (outliers on the covariates),

For each simulation, 100 replications were performed and the Relative Prediction Error ($MRPE$) was computed using the following formula:

$$MRPE = median((y_i - \mathbf{X}^T\boldsymbol{\beta}_g)^2),$$

for $g = 1, ..., 100$ simulations.

Result and Discussion

Table 6.7 lists the $MRPE$ for each criterion described above, and Figures (6.1) to (6.7) display the performance of the models selection through boxplots of the coefficient estimation.

When no contamination, all $LASSO$ versions performed well with respect to both $MRPE$ ($MRPE = 0.4350$) and variable selection. This can be observed from the extremely small values of $MRPE$ with perfect selection of non-zero and zero coefficients. The boxplot in Figure (6.1) shows the well selecting ability by all $LASSO$ methods.

When vertical outliers are introduced, the non-robust $ada\text{-}LASSO$ suffers from a strong influence of these outliers. Although the boxplot shows to the correct mean coefficients, the variance is quiet large, as shown in the boxplots (Figures (6.2) to (6.4)). $LAD\text{-}LASSO$ also shows good variable selection behavior, but the $ada\text{-}LASSO_R$ is the best with respect to $MRPE$ performance.

For the case with bad leverage points, $ada\text{-}LASSO_R$ exhibits its strengths and clearly performs best (Figures (6.5) to (6.7)); the lowest values of $MRPE$ were obtained for $ada\text{-}LASSO_R$. $LAD\text{-}LASSO$ was highly influenced by the leverage points, which was

reflected in the large $MRPE$, and was selected over the fit model (selected model contained all $\beta_j, j = 1, ..., 8$). On the other hand, the influence of the leverage was stronger on $ada\text{-}LASSO$ due to the high variability of the selection variable, and it suffered from the largest $MRPE$ among all the investigated methods.

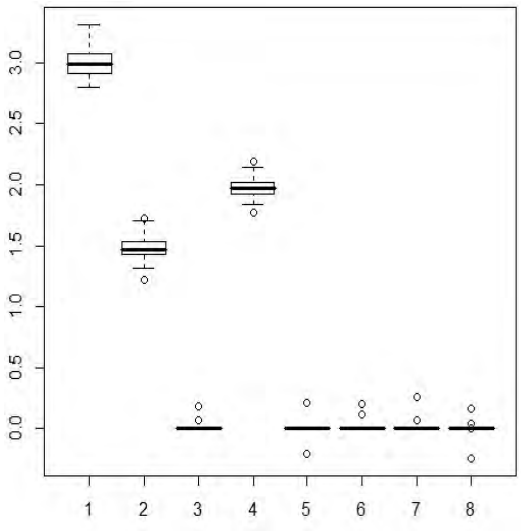
Table 6.8 and Figures (6.8) to (6.14) show the result when disturbances are normal and the correlation is high, $ada\text{-}LASSO_R$ is superior. $LASSO$ outperformed $ada\text{-}LASSO_R$ in the case with no multicollinearity. However, when degree of multicollinearity is high, $ada\text{-}LASSO_R$ is superior to them.

Table 6.7: Relative prediction error ($MRPE$) based on 100 replications, with $r = 0.5$

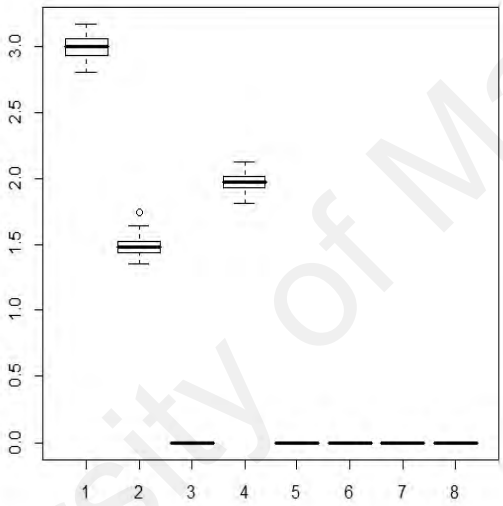
Situations	$ada\text{-}LASSO_R$	$ada\text{-}LASSO$	$LAD\text{-}LASSO$
No contaminated	0.4198	0.4535	0.4350
5% verticals	0.4392	0.7363	0.5481
10% verticals	0.4343	0.6867	0.5587
20% verticals	0.4253	0.6647	0.5471
5% leverage	0.4077	5.7562	5.4415
10% leverage	0.4487	5.9386	5.6834
20% leverage	0.4078	5.8864	5.4639

Table 6.8: Simulation result, relative prediction error ($MRPE$) based on 100 replications, for every method with $r = 0.8$

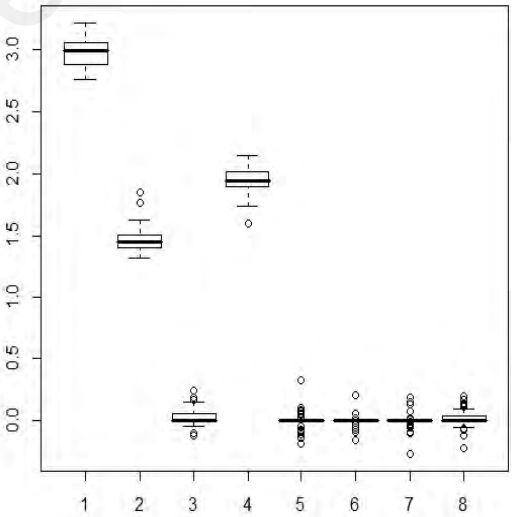
Situations	$ada\text{-}LASSO_R$	$ada\text{-}LASSO$	$LAD\text{-}LASSO$
No contaminated	0.4002	0.4470	0.4333
5% verticals	0.4108	0.5661	0.4938
10% verticals	0.4096	0.6679	0.5418
20% verticals	0.4552	1.689	1.157
5% leverage	0.4120	0.6404	0.6062
10% leverage	0.4423	5.7266	5.7259
20% leverage	0.4668	6.115	5.9698



(a) $ada\text{-}LASSO_R$

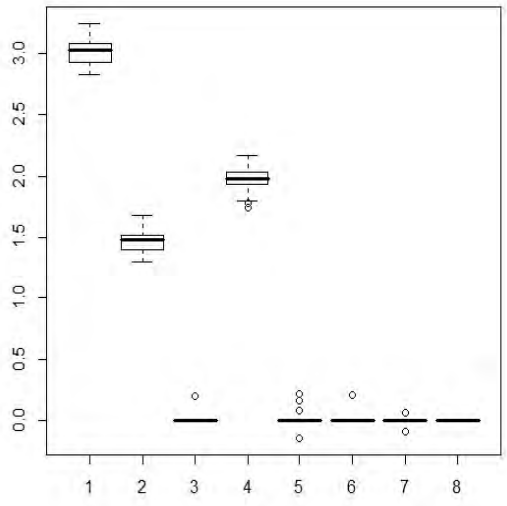


(b) $LASSO$

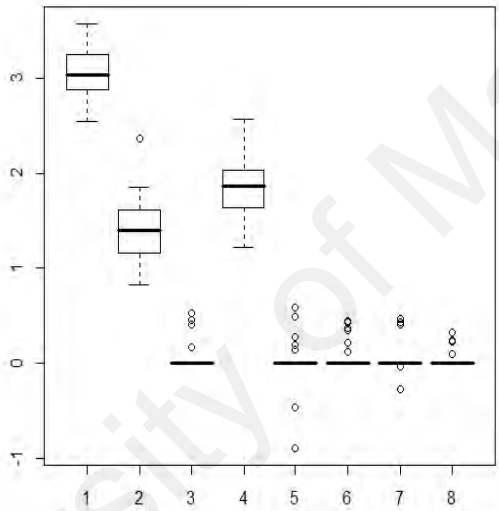


(c) $LAD\text{-}LASSO$

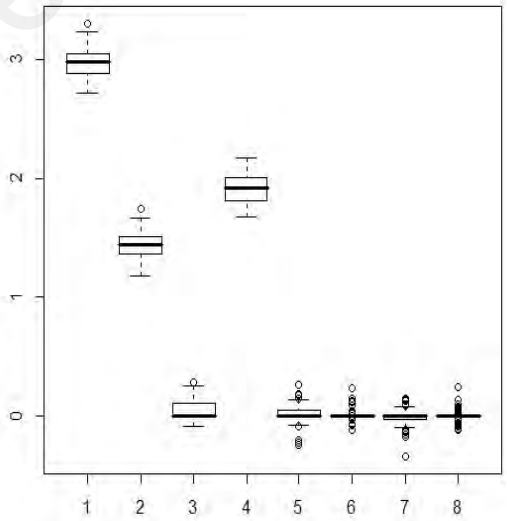
Figure 6.1: Boxplots of estimates for the eight coefficients from 100 simulated data sets with $r = 0.5$, no contaminated data



(a) $ada\text{-}LASSO_R$

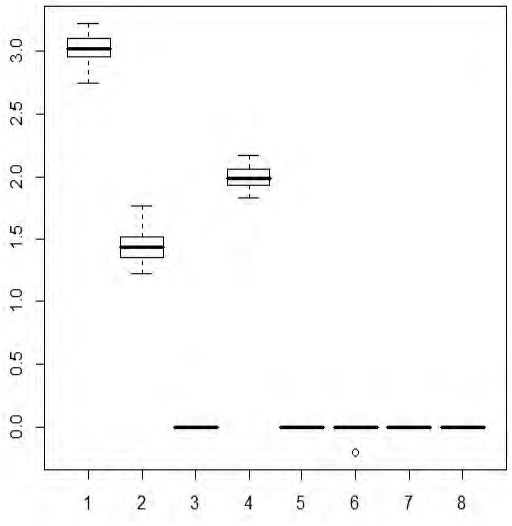


(b) $LASSO$

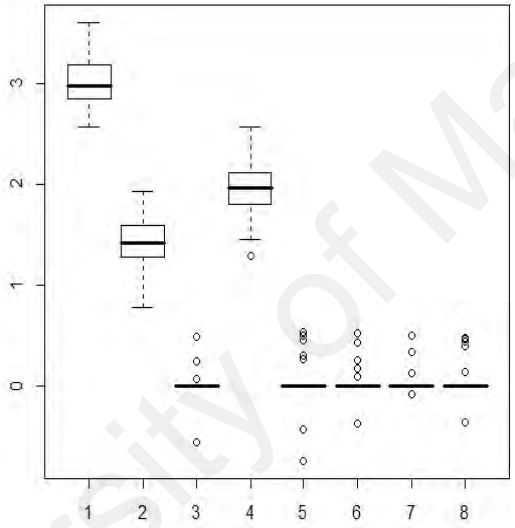


(c) $LAD\text{-}LASSO$

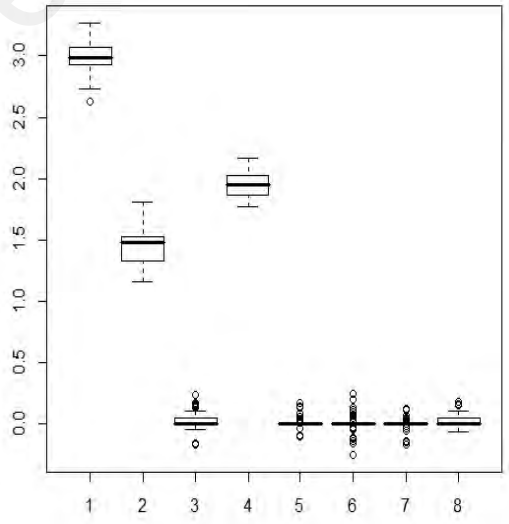
Figure 6.2: Boxplots of estimates for the eight coefficients from 100 simulated data sets with $r = 0.5$, 5% verticals



(a) $ada\text{-}LASSO_R$

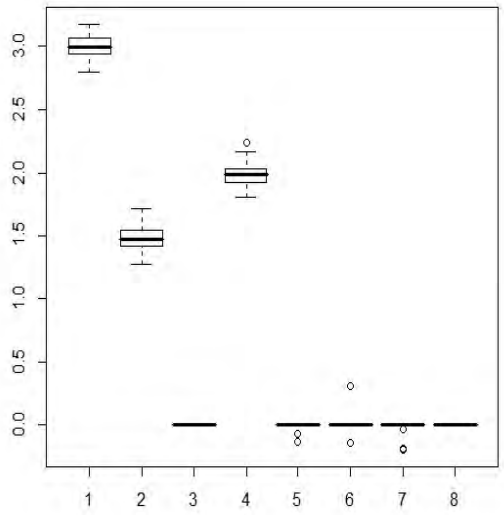


(b) $LASSO$

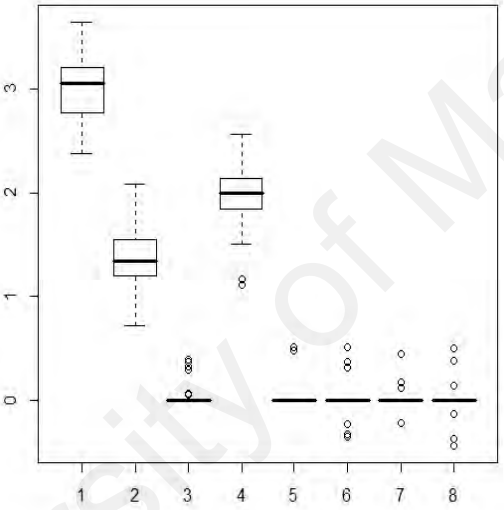


(c) $LAD\text{-}LASSO$

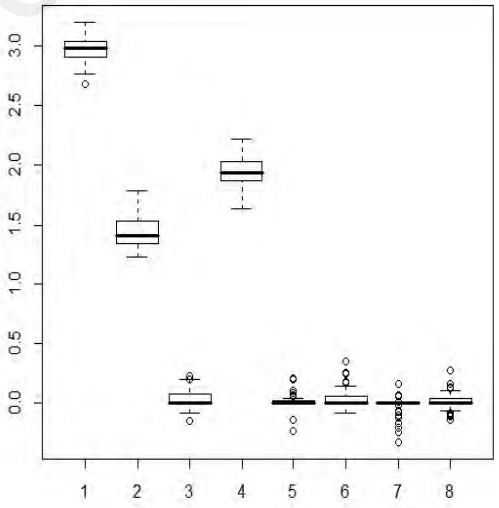
Figure 6.3: Boxplots of estimates for the eight coefficients from 100 simulated data sets with $r = 0.5$, 10% verticals



(a) $ada\text{-}LASSO_R$

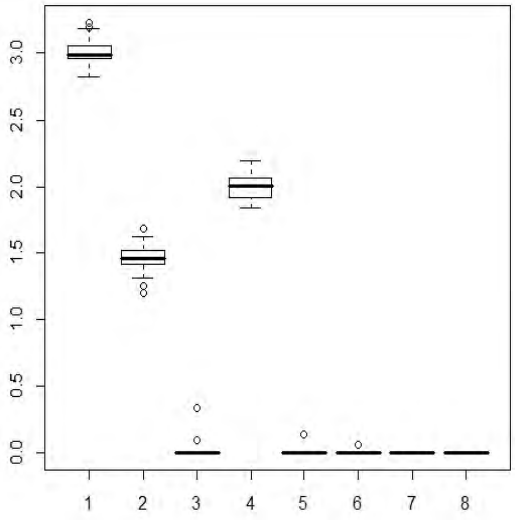


(b) $LASSO$

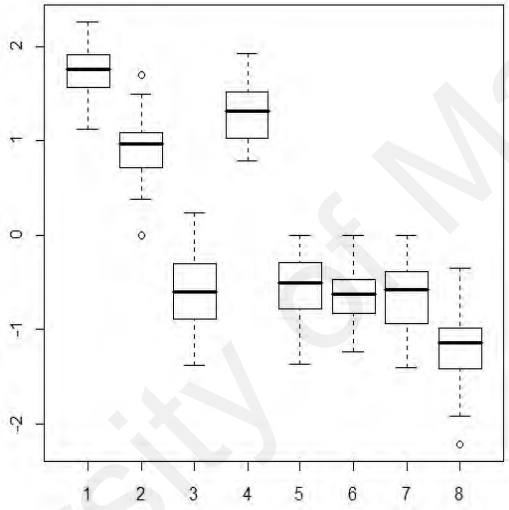


(c) $LAD\text{-}LASSO$

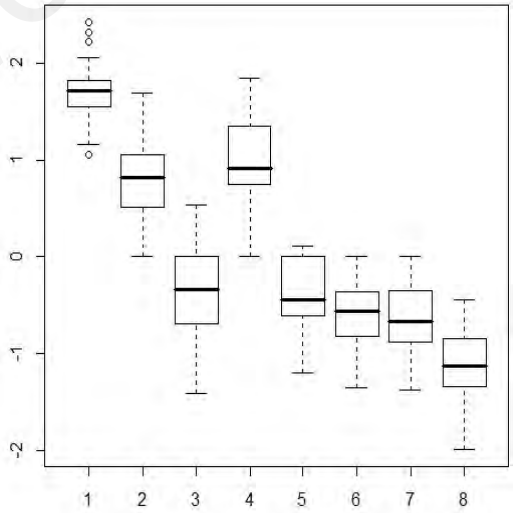
Figure 6.4: Boxplots of estimates for the eight coefficients from 100 simulated data sets with $r = 0.5$, 20% verticals



(a) $ada\text{-}LASSO_R$

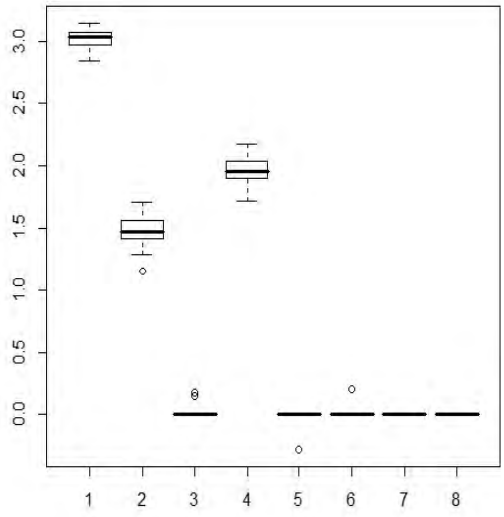


(b) $LASSO$

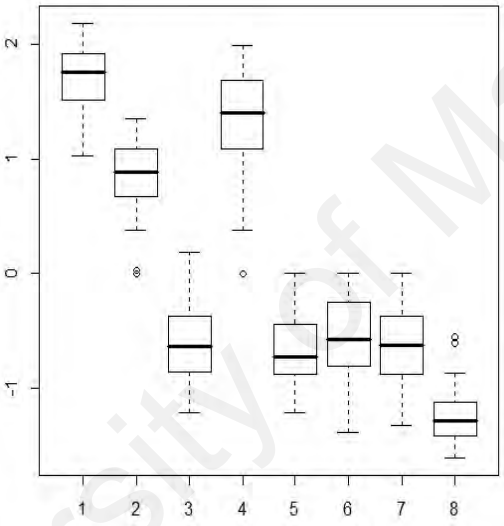


(c) $LAD\text{-}LASSO$

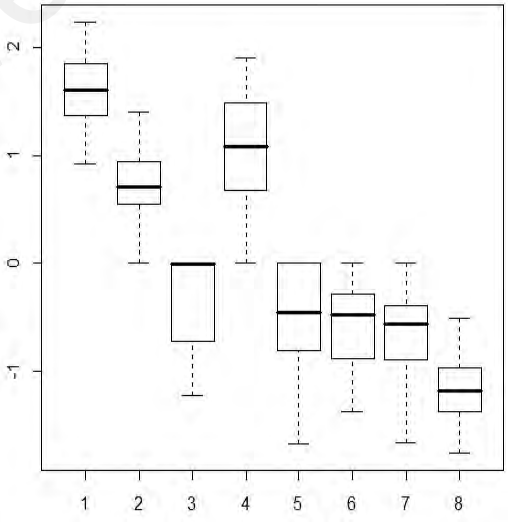
Figure 6.5: Boxplots of estimates for the eight coefficients from 100 simulated data sets with $r = 0.5$, 5% leverage



(a) $ada\text{-}LASSO_R$

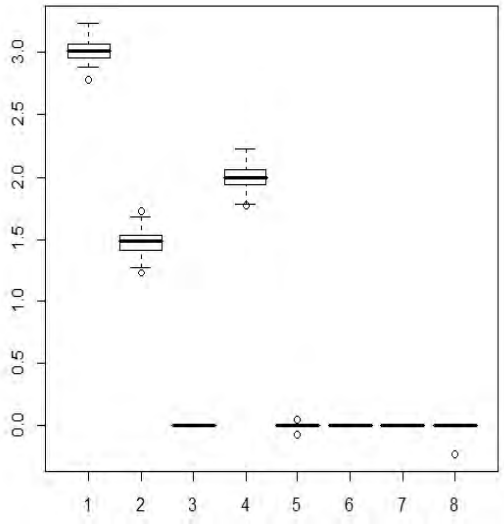


(b) $LASSO$

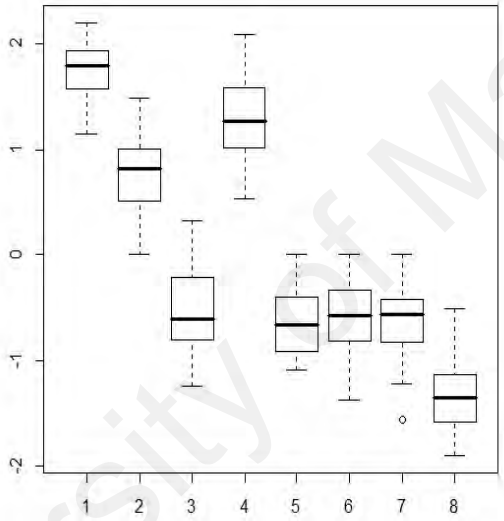


(c) $LAD\text{-}LASSO$

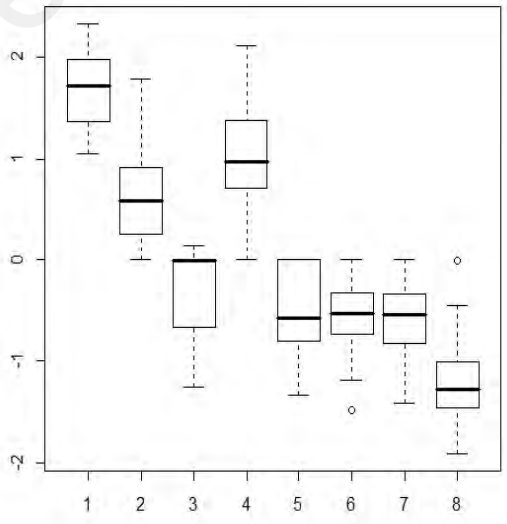
Figure 6.6: Boxplots of estimates for the eight coefficients from 100 simulated data sets with $r = 0.5$, 10% leverage



(a) $ada\text{-}LASSO_R$

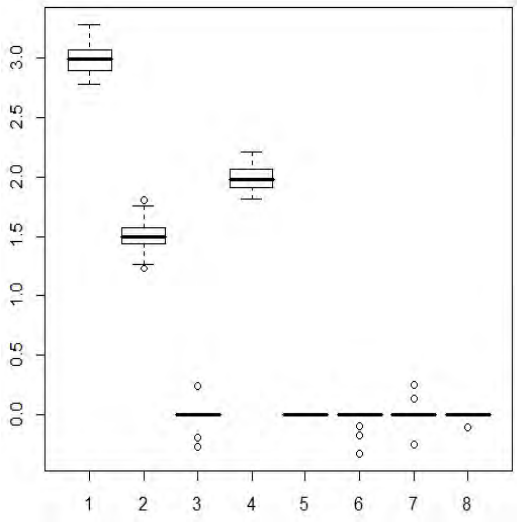


(b) $LASSO$

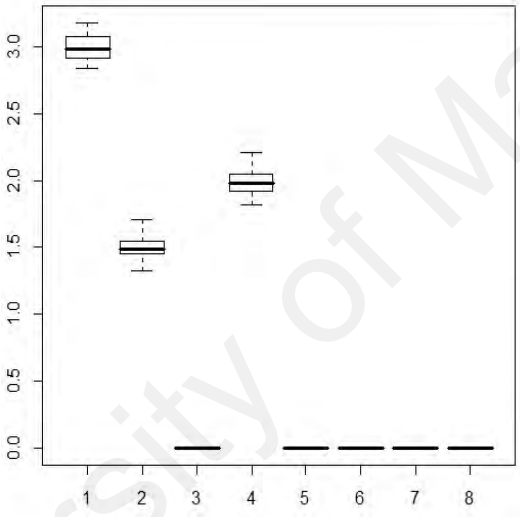


(c) $LAD\text{-}LASSO$

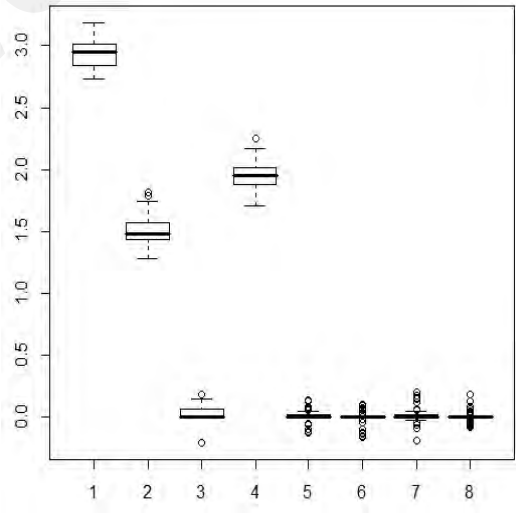
Figure 6.7: Boxplots of estimates for the eight coefficients from 100 simulated data sets with $r = 0.5$, 20% leverage



(a) $ada\text{-}LASSO_R$

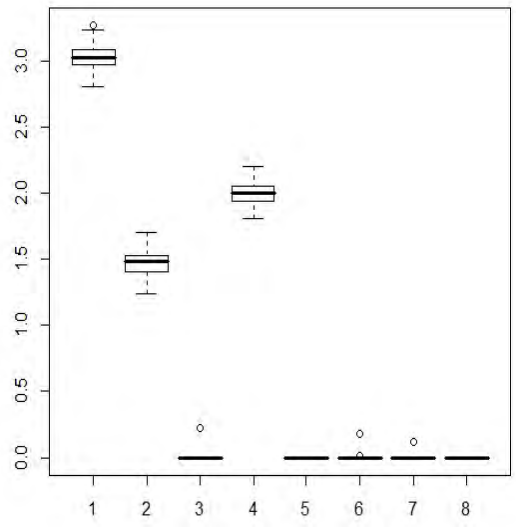


(b) $LASSO$

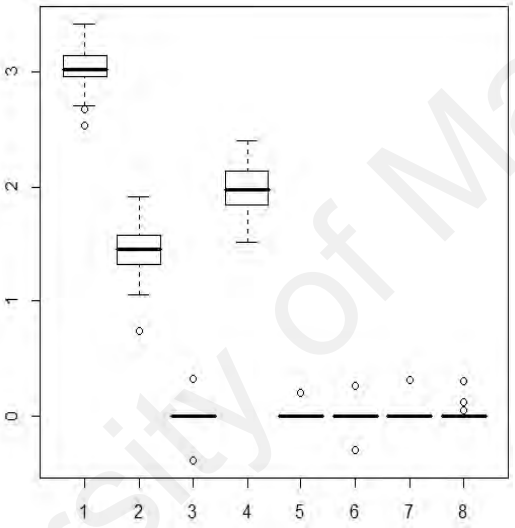


(c) $LAD\text{-}LASSO$

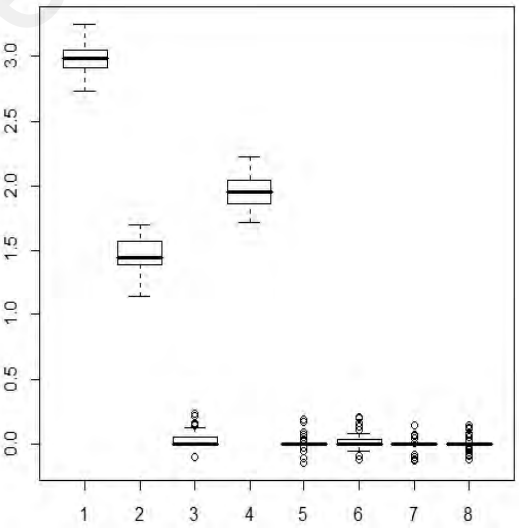
Figure 6.8: Boxplots of estimates for the eight coefficients from 100 simulated data sets with $r = 0.8$, no contaminated data



(a) $ada\text{-}LASSO_R$

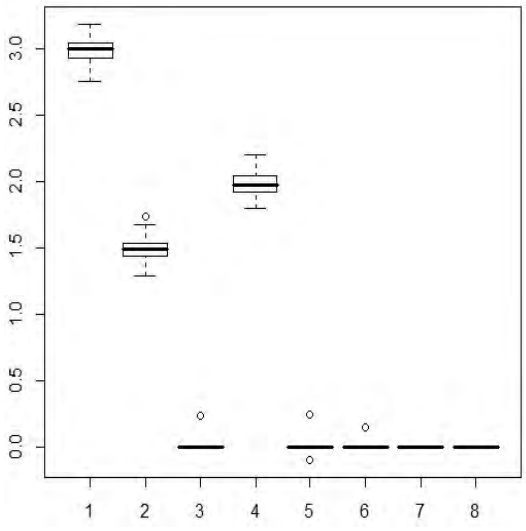


(b) $LASSO$

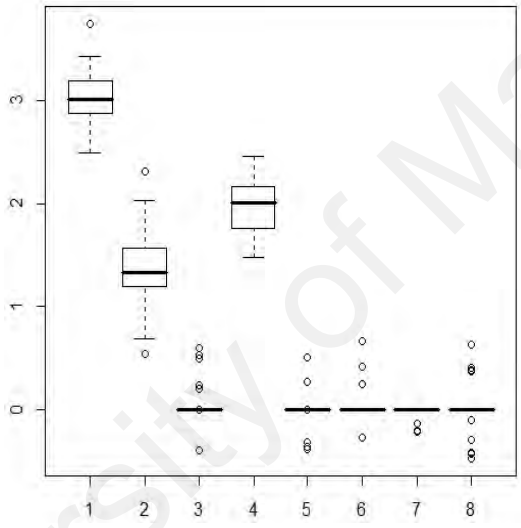


(c) $LAD\text{-}LASSO$

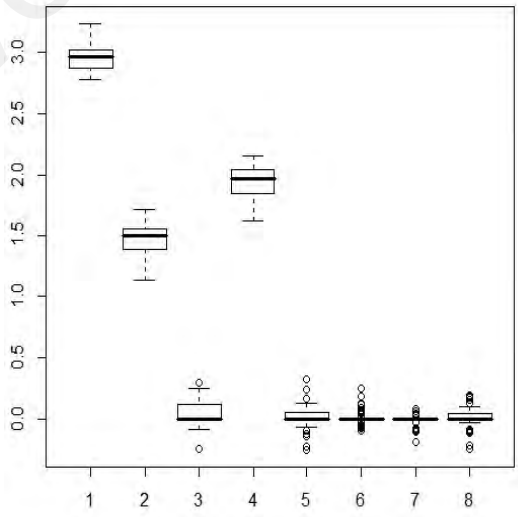
Figure 6.9: Boxplots of estimates for the eight coefficients from 100 simulated data sets with $r = 0.8$, 5% verticals



(a) $ada\text{-}LASSO_R$

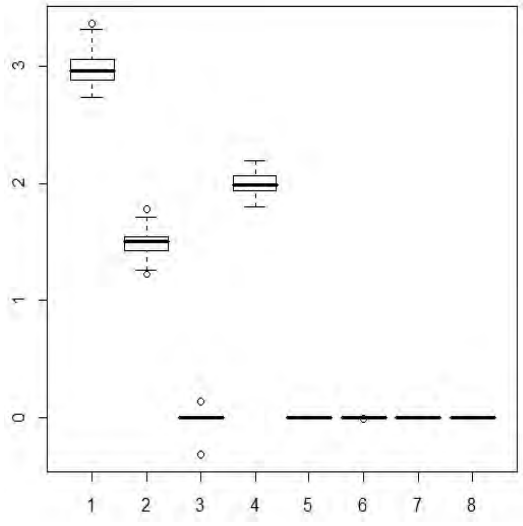


(b) $LASSO$

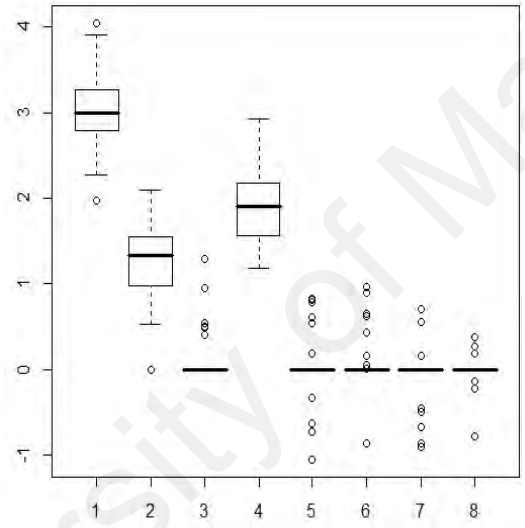


(c) $LAD\text{-}LASSO$

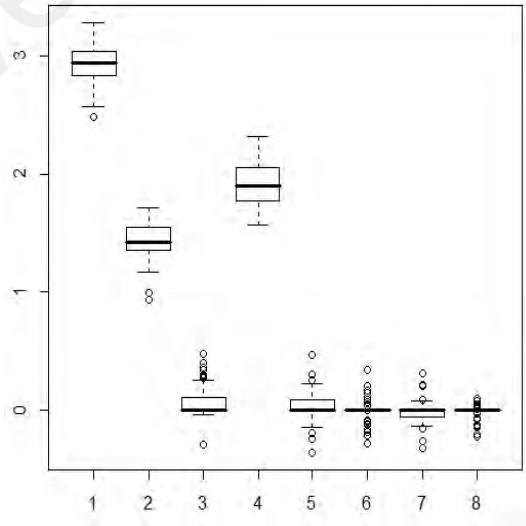
Figure 6.10: Boxplots of estimates for the eight coefficients from 100 simulated data sets with $r = 0.8$, 10% verticals



(a) $ada\text{-}LASSO_R$

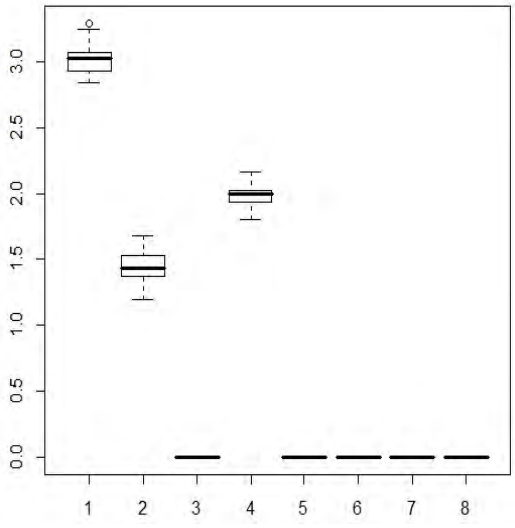


(b) $LASSO$

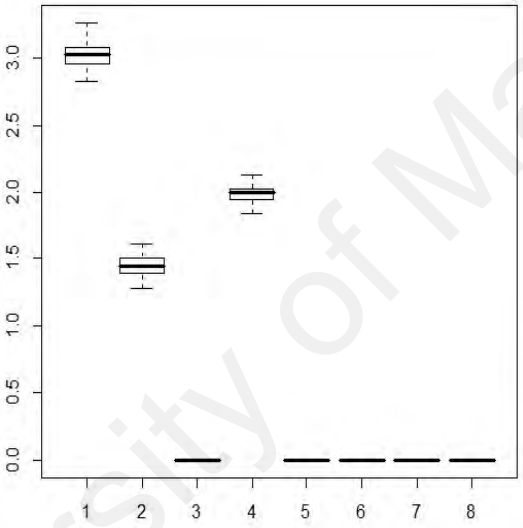


(c) $LAD\text{-}LASSO$

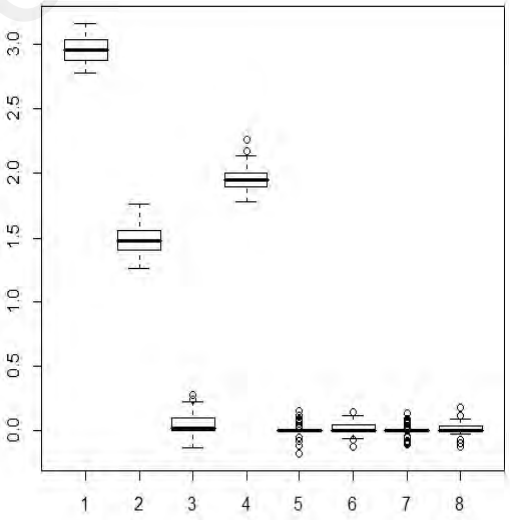
Figure 6.11: Boxplots of estimates for the eight coefficients from 100 simulated data sets with $r = 0.8$, 20% verticals



(a) $ada\text{-}LASSO_R$

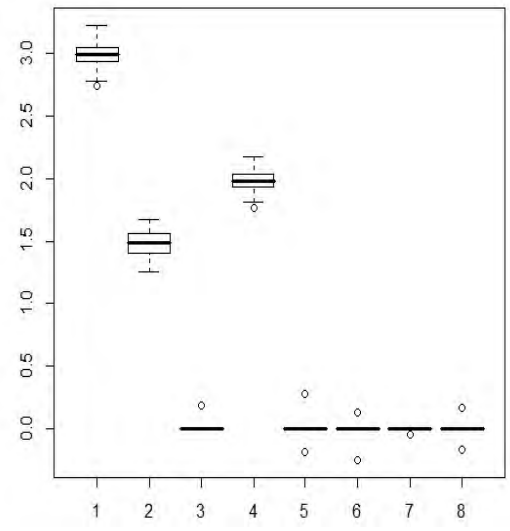


(b) $LASSO$

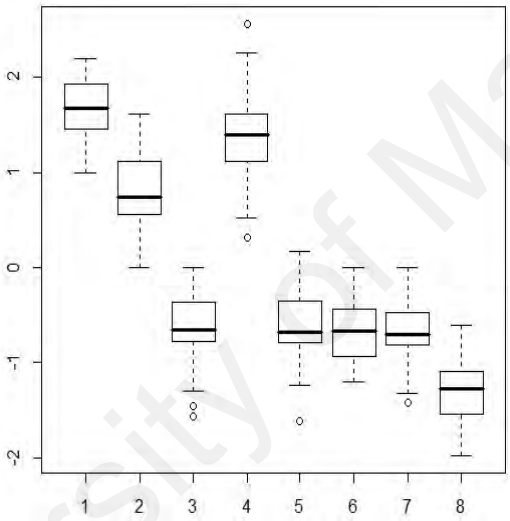


(c) $LAD\text{-}LASSO$

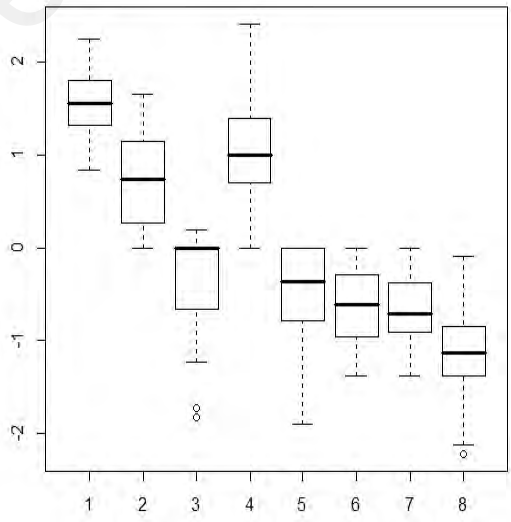
Figure 6.12: Boxplots of estimates for the eight coefficients from 100 simulated data sets with $r = 0.8$, 5% leverage



(a) $ada\text{-}LASSO_R$

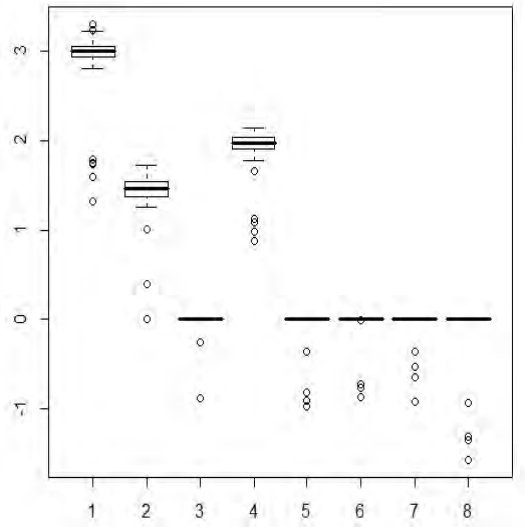


(b) $LASSO$

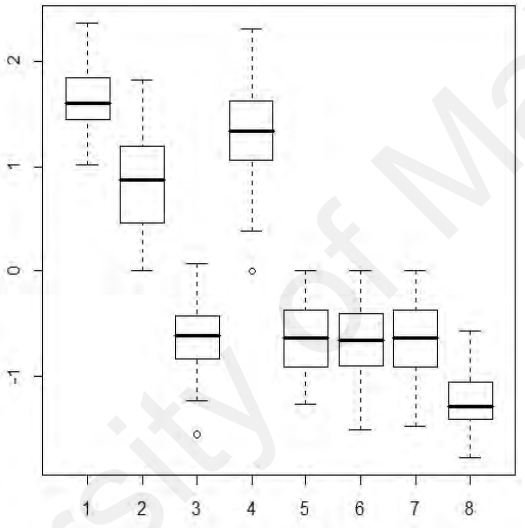


(c) $LAD\text{-}LASSO$

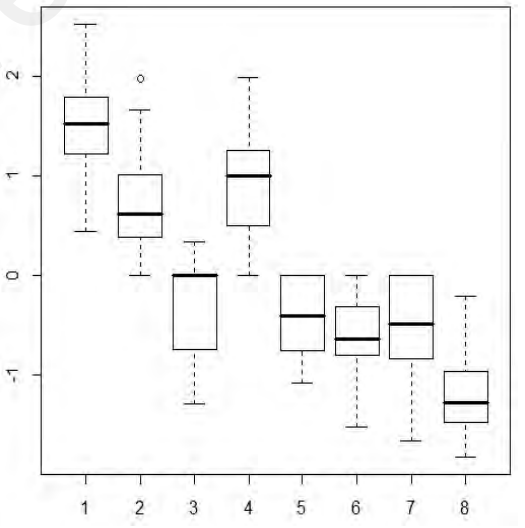
Figure 6.13: Boxplots of estimates for the eight coefficients from 100 simulated data sets with $r = 0.8$, 10% leverage



(a) $ada\text{-}LASSO_R$



(b) $LASSO$



(c) $LAD\text{-}LASSO$

Figure 6.14: Boxplots of estimates for the eight coefficients from 100 simulated data sets with $r = 0.8$, 20% leverage

6.3.3 Simulation Example 3 (Large Data Set)

A simulation study is carried out to investigate the performance of $ada\text{-}LASSO$ based on diagnostic method, namely $ada - LASSO_R$ statistic for selecting model in large data of regression model. Four different situations are considered, uncontaminated data, data with verticals, data with good leverage points and data with bad leverage points. In order to identify influential observations, we use Studentized residuals defined in Eqn. (2.22).

The same procedure employed in Section 5.9 is used here to generate the data set. Then fitted to give the parameter estimate is calculated using Eqn. (6.5). Then, the sample mean, the median of standard error (MSE) of the parameters, the median of relative prediction errors $MRPE$ of parameters, the sample standard deviation (SD) of parameters are then calculated.

If the values of $\boldsymbol{\beta}_{ada-LASSO_R}$ are close to the true value, then the procedure has correctly detected the best variables in the data. The process is carried out 100 times. The performance of the procedure is then examined by plot the boxplot of the estimators in the simulation.

Discussion

Tables 6.9 to 6.11 list the $Mean$, median of MSE, and median of $RMSE$ for $ada\text{-}LASSO$ and Figures (6.15) to (6.17) display the models selection abilities through box plots of coefficient estimation.

In the case of no contamination, $ada\text{-}LASSO_R$ performs well with respect to summary statistics and variable selection ability (see Figure (6.15)). This is due to having

extremely small values of median(MSE) (Median(MSE) are 0.0099, 0.0131, ..,0.0028) and median($RMSE$) and almost perfectly selected non-zero and zero coefficients.

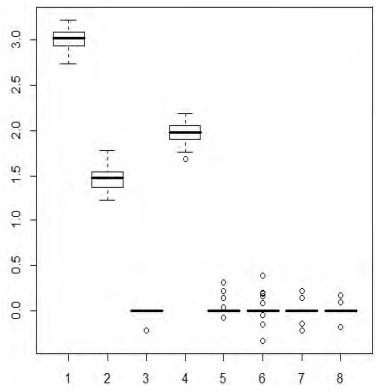
When vertical outliers were introduced, the $ada\text{-}LASSO_R$ is still good with respect to summary statistic. In a scenario with bad leverage points, $ada\text{-}LASSO_R$ exhibited its strength and clearly performed best; the low values of summary statistic were obtained for $ada\text{-}LASSO_R$. The conclusions of this simulation are as the followings:

- The estimated mean for all parameters are consistently close to the true values.
- The MSE for all parameter estimations are generally small.
- The values for $RMSE$ of each parameter are small.
- The standard deviation is consistently small for all parameter estimations

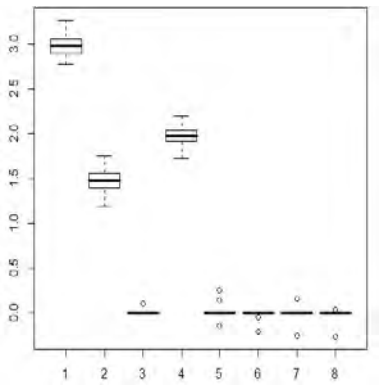
By looking the above results, the robust $ada\text{-}LASSO$ with diagnostic tool estimation method performs well in selecting the variables of the regression models.

Table 6.9: The $ada\text{-}LASSO_R$ estimation of eight estimators for simulated data sets with different level of verticals

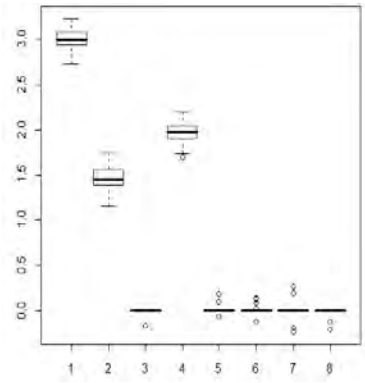
		No contaminated data set			
Coefficients	True Values	Mean	Median(MSE)	Median($RMSE$)	std.dev
$\hat{\beta}_0$	0	-0.0408	0.0099	0.0004	0.0814
$\hat{\beta}_1$	3	3.0626	0.0131	0.0006	0.1124
$\hat{\beta}_2$	1.5	1.3545	0.0174	0.0014	0.1193
$\hat{\beta}_3$	0	0.0000	0.0023	0.0000	0.0220
$\hat{\beta}_4$	2	2.0523	0.0138	0.0005	0.1048
$\hat{\beta}_5$	0	0.0000	0.0044	0.0000	0.0410
$\hat{\beta}_6$	0	0.0000	0.0067	0.0000	0.0630
$\hat{\beta}_7$	0	0.0000	0.0039	0.0000	0.0369
$\hat{\beta}_8$	0	0.0000	0.0028	0.0000	0.0268
		Data set with 5% verticals			
Coefficients	True Values	Mean	Median(MSE)	Median($RMSE$)	std.dev
$\hat{\beta}_0$	0	0.0227	0.0092	0.0002	0.0797
$\hat{\beta}_1$	3	3.0246	0.01208	0.0002	0.1072
$\hat{\beta}_2$	1.5	1.4894	0.0130	0.0001	0.1215
$\hat{\beta}_3$	0	0.0000	0.00113	0.0000	0.0105
$\hat{\beta}_4$	2	1.9581	0.0096	0.0004	0.0883
$\hat{\beta}_5$	0	0.0000	0.0034	0.0000	0.0325
$\hat{\beta}_6$	0	0.0000	0.0022	0.0000	0.0212
$\hat{\beta}_7$	0	0.0000	0.0032	0.0000	0.0300
$\hat{\beta}_8$	0	0.0346	0.0048	0.0003	0.0264
		Data set with 10% verticals			
Coefficients	True Values	Mean	Median(MSE)	Median($RMSE$)	std.dev
$\hat{\beta}_0$	0	-0.0506	0.0121	0.0005	0.0844
$\hat{\beta}_1$	3	3.2226	0.0267	0.0022	0.1054
$\hat{\beta}_2$	1.5	1.1514	0.0371	0.0034	0.1252
$\hat{\beta}_3$	0	0.0000	0.0017	0.0000	0.0163
$\hat{\beta}_4$	2	2.0608	0.0140	0.0006	0.1000
$\hat{\beta}_5$	0	0.0000	0.0023	0.0000	0.0219
$\hat{\beta}_6$	0	0.0000	0.0026	0.0000	0.0246
$\hat{\beta}_7$	0	0.0000	0.0048	0.0000	0.0449
$\hat{\beta}_8$	0	0.0000	0.0033	0.0000	0.0307
		Data set with 20% verticals			
Coefficients	True Values	Mean	Median(MSE)	Median($RMSE$)	std.dev
$\hat{\beta}_0$	0	0.1696	0.0150	0.00169	0.1324
$\hat{\beta}_1$	3	3.0480	0.0167	0.0004	0.1440
$\hat{\beta}_2$	1.5	1.5141	0.0152	0.0001	0.1411
$\hat{\beta}_3$	0	0.0000	0.0076	0.0000	0.0708
$\hat{\beta}_4$	2	1.9537	0.0139	0.0004	0.1275
$\hat{\beta}_5$	0	0.0000	0.0103	0.0000	0.0959
$\hat{\beta}_6$	0	0.0000	0.0079	0.0000	0.0746
$\hat{\beta}_7$	0	0.0000	0.0032	0.0000	0.0299
$\hat{\beta}_8$	0	0.0000	0.0036	0.0000	0.0343
		Data set with 30% verticals			
Coefficients	True Values	Mean	Median(MSE)	Median($RMSE$)	std.dev
$\hat{\beta}_0$	0	0.4940	0.0262	0.0049	0.2351
$\hat{\beta}_1$	3	2.7426	0.0372	0.0025	0.2315
$\hat{\beta}_2$	1.5	1.3284	0.0351	0.0017	0.2861
$\hat{\beta}_3$	0	0.0000	0.0215	0.0000	0.2015
$\hat{\beta}_4$	2	1.6011	0.0504	0.0039	0.2300
$\hat{\beta}_5$	0	0.0000	0.0159	0.0000	0.1485
$\hat{\beta}_6$	0	0.0000	0.0212	0.0000	0.1982
$\hat{\beta}_7$	0	0.0000	0.0158	0.0000	0.1485
$\hat{\beta}_8$	0	0.0000	0.0181	0.0000	0.1699
		Data set with 40% vertical			
Coefficients	True Values	Mean	Median(MSE)	Median($RMSE$)	std.dev
$\hat{\beta}_0$	0	1.2621	0.0600	0.01262	0.1959
$\hat{\beta}_1$	3	3.3253	0.0400	0.0032	0.2614
$\hat{\beta}_2$	1.5	1.208	0.0413	0.0029	0.2841
$\hat{\beta}_3$	0	0.0000	0.0195	0.0000	0.1776
$\hat{\beta}_4$	2	2.5945	0.0750	0.0059	0.2555
$\hat{\beta}_5$	0	0.0000	0.0137	0.0000	0.1282
$\hat{\beta}_6$	0	0.0000	0.0204	0.0000	0.1896
$\hat{\beta}_7$	0	0.0000	0.0223	0.0000	0.2068
$\hat{\beta}_8$	0	0.0000	0.01621	0.0000	0.1512



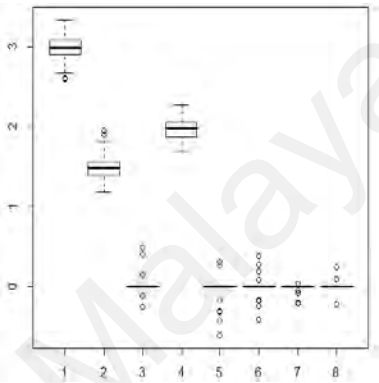
(a) No contaminated



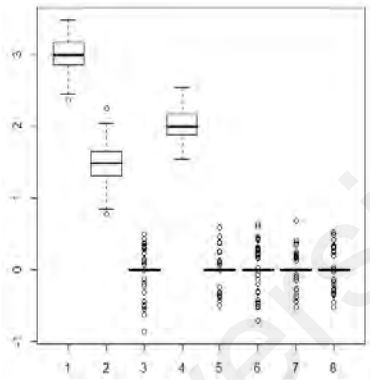
(b) 5% Verticals



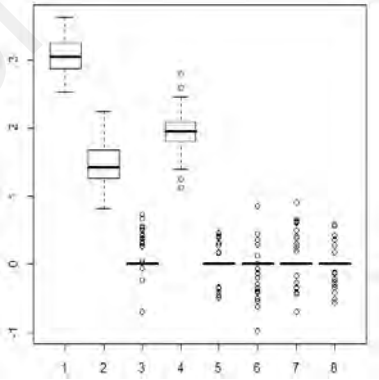
(c) 10% Verticals



(d) 20% Verticals



(e) 30% Verticals

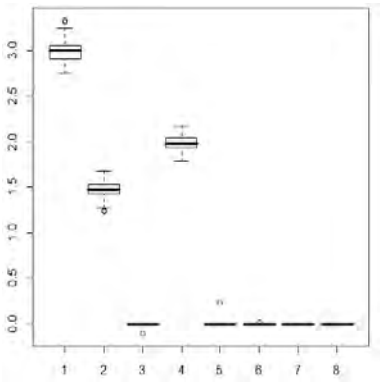


(f) 40% Verticals

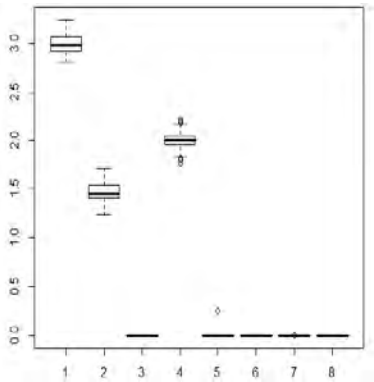
Figure 6.15: Boxplots of $ada\text{-}LASSO_R$ estimates for the eight coefficients from 100 simulated data sets, verticals

Table 6.10: The $adaLASSO_R$ estimation of eight parameters for simulated data sets with different level of bad leverage points

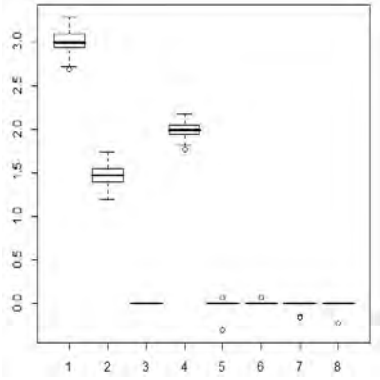
		Data set with 5% bad leverage			
Coefficients	True Values	Mean	Median(MSE)	Median($RMSE$)	std.dev
$\hat{\beta}_0$	0	0.0878	0.0136	0.0008	0.0942
$\hat{\beta}_1$	3	3.0172	0.0121	0.0001	0.1118
$\hat{\beta}_2$	1.5	1.4820	0.0101	0.0001	0.0945
$\hat{\beta}_3$	0	0.0000	0.00103	0.0000	0.0096
$\hat{\beta}_4$	2	1.9761	0.0088	0.0002	0.0812
$\hat{\beta}_5$	0	0.0000	0.0025	0.0000	0.0236
$\hat{\beta}_6$	0	0.0000	0.0001	0.0000	0.0018
$\hat{\beta}_7$	0	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_8$	0	0.0000	0.0000	0.0000	0.0000
		Data set with 10% bad leverage			
Coefficients	True Values	Mean	Median(MSE)	Median($RMSE$)	std.dev
$\hat{\beta}_0$	0	-0.0463	0.0102	5e-04	0.0834
$\hat{\beta}_1$	3	3.0748	0.0137	7e-04	0.1020
$\hat{\beta}_2$	1.5	1.5460	0.0137	5e-04	0.1044
$\hat{\beta}_3$	0	0.0000	0.0000	0e+00	0.0000
$\hat{\beta}_4$	2	1.9988	0.0092	0e+00	0.0863
$\hat{\beta}_5$	0	0.0000	0.0026	0e+00	0.0244
$\hat{\beta}_6$	0	0.0000	0.0000	0e+00	0.0000
$\hat{\beta}_7$	0	0.0000	0.0000	0e+00	0.0001
$\hat{\beta}_8$	0	0.0000	0.0000	0e+00	0.0000
		Data set with 20% bad leverage			
Coefficients	True Values	Mean	Median(MSE)	Median($RMSE$)	std.dev
$\hat{\beta}_0$	0	-0.0129	0.0091	0.0001	0.0844
$\hat{\beta}_1$	3	2.6924	0.0362	0.0031	0.1191
$\hat{\beta}_2$	1.5	1.6491	0.0225	0.0015	0.1124
$\hat{\beta}_3$	0	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_4$	2	1.9156	0.0122	0.0008	0.0830
$\hat{\beta}_5$	0	0.0000	0.0033	0.0000	0.0308
$\hat{\beta}_6$	0	0.0000	0.0010	0.0000	0.0097
$\hat{\beta}_7$	0	0.0000	0.0024	0.0000	0.0221
$\hat{\beta}_8$	0	0.0000	0.0024	0.0000	0.0228
		Data set with 30% bad leverage			
Coefficients	True Values	Mean	Median(MSE)	Median($RMSE$)	std.dev
$\hat{\beta}_0$	0	0.0256	0.0106	3e-04	0.0972
$\hat{\beta}_1$	3	3.0814	0.0147	8e-04	0.1100
$\hat{\beta}_2$	1.5	1.4538	0.0122	5e-04	0.1119
$\hat{\beta}_3$	0	0.0000	0.0000	0e+00	0.0000
$\hat{\beta}_4$	2	1.9916	0.0101	1e-04	0.0934
$\hat{\beta}_5$	0	0.0000	0.0022	0e+00	0.0200
$\hat{\beta}_6$	0	0.0000	0.0030	0e+00	0.0275
$\hat{\beta}_7$	0	0.0000	0.0000	0e+00	0.0000
$\hat{\beta}_8$	0	0.0000	0.0000	0e+00	0.0000
		Data set with 40% bad leverage			
Coefficients	True Values	Mean	Median(MSE)	Median($RMSE$)	std.dev
$\hat{\beta}_0$	0	-0.0825	0.0236	0.0008	0.2203
$\hat{\beta}_1$	3	3.2165	0.0504	0.0022	0.3684
$\hat{\beta}_2$	1.5	1.2125	0.0419	0.0029	0.3584
$\hat{\beta}_3$	0	0.0000	0.0270	0.0000	0.2454
$\hat{\beta}_4$	2	2.1107	0.0287	0.0011	0.2020
$\hat{\beta}_5$	0	0.0000	0.0285	0.0000	0.2582
$\hat{\beta}_6$	0	0.0000	0.0226	0.0000	0.2030
$\hat{\beta}_7$	0	0.0000	0.0252	0.0000	0.2326
$\hat{\beta}_8$	0	0.0000	0.0535	0.0000	0.4749



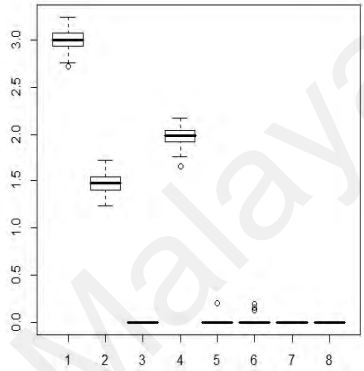
(a) 5% Bad Leverage



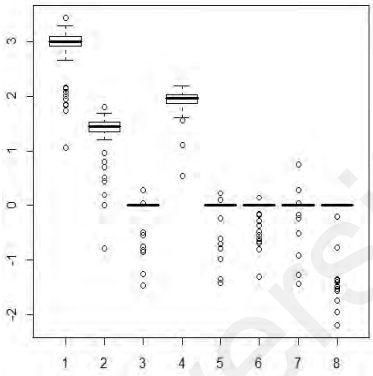
(b) 10% Bad Leverage



(c) 20% Bad Leverage



(d) 30% Bad Leverage

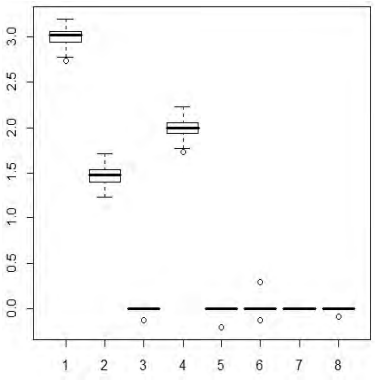


(e) 40% Bad Leverage

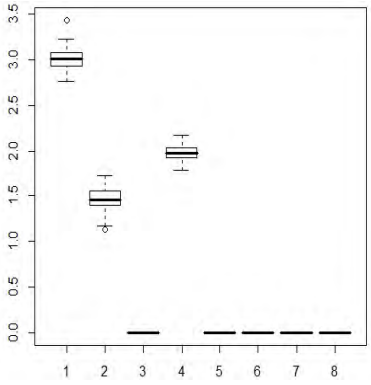
Figure 6.16: Boxplots of $ada\text{-}LASSO_R$ estimates for the eight coefficients from 100 simulated data sets, bad leverage point

Table 6.11: The $ada\text{-}LASSO_R$ estimation of eight parameters for simulated data sets with different level of good leverage points

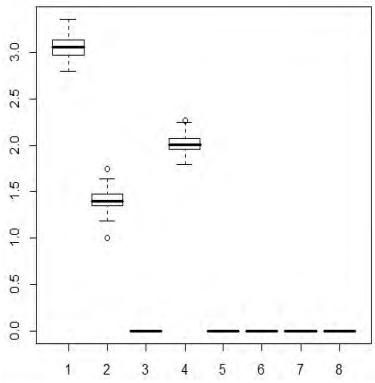
		Data set with 5% good leverage			
Coefficients	True Values	Mean	Median(MSE)	Median($RMSE$)	std.dev
$\hat{\beta}_0$	0	-0.1754	0.0205	0.0018	0.0817
$\hat{\beta}_1$	3	3.1575	0.0197	0.0016	0.0975
$\hat{\beta}_2$	1.5	1.5006	0.0115	0.0000	0.1021
$\hat{\beta}_3$	0	-0.1309	0.0140	0.0013	0.0131
$\hat{\beta}_4$	2	2.0639	0.0129	0.0006	0.0973
$\hat{\beta}_5$	0	0.0000	0.0022	0.0000	0.0203
$\hat{\beta}_6$	0	0.0000	0.0034	0.0000	0.0318
$\hat{\beta}_7$	0	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_8$	0	0.0000	0.0009	0.0000	0.0085
		Data set with 10% good leverage			
Coefficients	True Values	Mean	Median(MSE)	Median($RMSE$)	std.dev
$\hat{\beta}_0$	0	0.1418	0.0179	0.0014	0.0849
$\hat{\beta}_1$	3	2.8559	0.0205	0.0014	0.1087
$\hat{\beta}_2$	1.5	1.6157	0.0203	0.0012	0.1184
$\hat{\beta}_3$	0	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_4$	2	1.9430	0.0094	0.0006	0.0801
$\hat{\beta}_5$	0	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_6$	0	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_7$	0	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_8$	0	0.0000	0.0000	0.0000	0.0000
		Data set with 20% good leverage			
Coefficients	True Values	Mean	Median(MSE)	Median($RMSE$)	std.dev
$\hat{\beta}_0$	0	-0.0114	0.0108	0.0001	0.0859
$\hat{\beta}_1$	3	3.1502	0.0151	0.0015	0.1045
$\hat{\beta}_2$	1.5	1.1850	0.0266	0.0032	0.1145
$\hat{\beta}_3$	0	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_4$	2	2.1633	0.0185	0.0016	0.0896
$\hat{\beta}_5$	0	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_6$	0	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_7$	0	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_8$	0	0.0000	0.0000	0.0000	0.0000
		Data set with 30% good leverage			
Coefficients	True Values	Mean	Median(MSE)	Median($RMSE$)	std.dev
$\hat{\beta}_0$	0	0.1162	0.0125	0.0012	0.0952
$\hat{\beta}_1$	3	2.9537	0.0174	0.0005	0.0952
$\hat{\beta}_2$	1.5	1.5210	0.0185	0.0002	0.1172
$\hat{\beta}_3$	0	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_4$	2	1.9640	0.0103	0.0004	0.0904
$\hat{\beta}_5$	0	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_6$	0	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_7$	0	0.0000	0.0000	0.0000	0.0000
$\hat{\beta}_8$	0	0.0000	0.0000	0.0000	0.0000
		Data set with 40% good leverage			
Coefficients	True Values	Mean	Median(MSE)	Median($RMSE$)	std.dev
$\hat{\beta}_0$	0	0.0881	0.0125	9e-04	0.1024
$\hat{\beta}_1$	3	3.0430	0.0135	4e-04	0.1087
$\hat{\beta}_2$	1.5	1.4059	0.0117	9e-04	0.1071
$\hat{\beta}_3$	0	0.0000	0.0000	0e+00	0.0000
$\hat{\beta}_4$	2	2.0361	0.0097	4e-04	0.0791
$\hat{\beta}_5$	0	0.0000	0.0000	0e+00	0.0000
$\hat{\beta}_6$	0	0.0000	0.0000	0e+00	0.0000
$\hat{\beta}_7$	0	0.0000	0.0000	0e+00	0.0000
$\hat{\beta}_8$	0	0.0000	0.0000	0e+00	0.0000



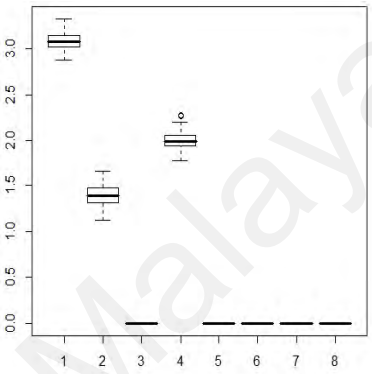
(a) 5% Good Leverage



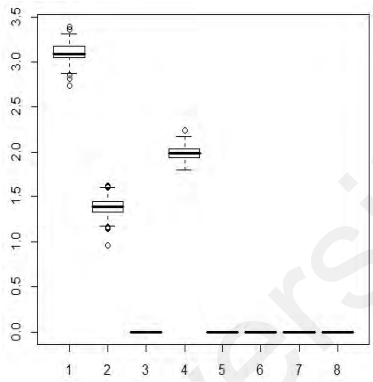
(b) 10% Good Leverage



(c) 20% Good Leverage



(d) 30% Good Leverage



(e) 40% Good Leverage

Figure 6.17: Boxplots of $ada\text{-}LASSO_R$ estimates for the eight coefficients from 100 simulated data sets, good leverage point

6.4 Examples

6.4.1 Example 1 (Small Data Sets)

In this example, the stack-loss dataset is considered. The data have been described in Section 4.8. This data set has been extensively analyzed by many authors (see Leroy and Rousseeuw (1987), Atkinson (1985); Rahmatullah Imon (2005)) and they report that this

three-predictor real data set (air flow, cooling water inlet temperature and acid concentration) contains 21 observations with four outliers (cases 1, 3, 4 and 21) and four high leverage points (cases 1, 2, 3 and 21). The robust regression model was fitted to the data set using M-estimator. The parameter estimates are given by, intercept, $\hat{\beta}_0 = -41.0265$, $\hat{\beta}_1 = 0.8293$, $\hat{\beta}_2 = 0.9261$, and $\hat{\beta}_3 = -0.1278$, which suggest that the best variables of data seem to be $\mathbf{x}_{i1}$, and $\mathbf{x}_{i2}$. Further, the correlation matrix of data is given by

$$\begin{pmatrix} 1 & 0.7818 & 0.5001 \\ 0.7818 & 1 & 0.3909 \\ 0.5001 & 0.3909 & 1 \end{pmatrix}.$$

All 2^3 possible models fitted with a combination of any of these covariance and computed several model selection methods values for each model.

Result and Discussion

Table 6.12 presents the best three selected models based on each criteria. The classical AIC and SIC methods select a model with three explanatory variables (see Tables 6.13 and 6.15). As we see in Tables 6.12 and 6.14, the values of AIC and SIC small with full model, while classical Cp selects model with two variables as showed in Table 6.13.

While AIC, Cp, and SIC based on M-estimation select a model with one variable (under fit), three variables (over fit), and $\mathbf{x}_{i1}, \mathbf{x}_{i3}$ variables (wrong fit), respectively. This is in line with the simulation results where robust $RAIC$ has the tendency to select under fit models in the presence of outliers and bad leverage points. It is observed that RCp has the trend to select over fit models in the presence of bad leverage points.

The proposed methods based on a deletion estimate of scale select the same best model with two variables, Flow of cooling air ($\mathbf{x}_1$) and Cooling Water Inlet Temperature ($\mathbf{x}_2$).

Table 6.12: Stack-Loss data. the selected best variables from best three models based on different classical criteria, robust criteria with M-estimation, and robust criteria using deletion estimate of scale

Criteria	Selected variables		
	Best model	Second best model	Third best model
AIC	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$	$\mathbf{x}_1$	$\mathbf{x}_2$
$RAIC$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_1$
AIC_R	$\mathbf{x}_1, \mathbf{x}_2$	$\mathbf{x}_1$	$\mathbf{x}_2$
Cp	$\mathbf{x}_1, \mathbf{x}_2$	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$	$\mathbf{x}_1$
RCp	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$	$\mathbf{x}_1, \mathbf{x}_2$	$\mathbf{x}_1$
Cp_R	$\mathbf{x}_1, \mathbf{x}_2$	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$	$\mathbf{x}_1$
SIC	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$	$\mathbf{x}_1, \mathbf{x}_2$	$\mathbf{x}_1$
$RSIC$	$\mathbf{x}_1, \mathbf{x}_3$	$\mathbf{x}_2$	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$
SIC_R	$\mathbf{x}_1, \mathbf{x}_2$	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$	$\mathbf{x}_1$

Table 6.13: Values of the classical AIC, and robust $RAIC$, and AIC_R statistics for Stack-Loss data

Selected Variables	AIC	$RAIC$	AIC_R
x_1	6.7	8.0	**5.3233**
x_2	7.1	**6.5**	6.0633
x_3	8.4	**7.3**	6.6277
x_1, x_2	8.2	9.0	**4.2795**
x_1, x_3	8.7	8.9	6.2639
x_2, x_3	9.1	9.0	7.9139
x_1, x_2, x_3	**4.7**	10.6	7.2575

Table 6.14: Values of the classical Cp, and robust RCp, and Cp_R statistics for Stack-Loss data

Selected Variables	Cp	RCp (V_p)	Cp_R
x_1	13.3	-2.1(1.96)	19.9897
x_2	28.9	34.9(1.96)	271.3652
x_3	148.9	62.4(1.96)	420.7398
x_1, x_2	**2.95**	4.47(2.95)	**2.0822**
x_1, x_3	14.3	7.2(2.95)	21.0511
x_2, x_3	30.1	61.6(2.95)	226.0185
x_1, x_2, x_3	4.0	**3.9(3.93)**	4.000

Table 6.15: Values of the classical SIC, and robust $RSIC$, and SIC_R statistics for Stack-Loss data

Selected Variables	SIC	SIC_M	SIC_R
x_1	3.010	4.273	0.6741
x_2	3.425	2.848	2.4579
x_3	4.706	3.556	3.0223
x_1, x_2	2.721	3.392	**-0.0848**
x_1, x_3	3.124	**2.787**	0.8559
x_2, x_3	3.553	3.480	2.5058
x_1, x_2, x_3	**2.631**	3.172	0.0466

6.4.2 Example 2 (Small Data Sets)

In this example, Hawkins-Bradu-Kass dataset is used. The data have been described in Section 3.3.2. This data available from the R library wle as data(artificial). Artificial data set containing 75 observations with 10 outliers (cases 1 to 10) and 14 high leverage points (cases 1 to 14). Scatter plots of Y on each three $\mathbf{X}'s$ as shown in Figures (3.14),

(3.15), and (3.16), clearly separate 10 high leverage outliers, 4 high leverage points and 61 clean observations. The robust regression model based on M-estimator was fitted to the data set. The parameter estimates are given by, intercept = -0.7848, $\boldsymbol{\beta}_{Hawkins} = 0.1791$, $\boldsymbol{\beta}_{Bradu} = 0.0062$, $\boldsymbol{\beta}_{Kass} = 0.2715$. Further, the correlation matrix of data is given by

$$\begin{pmatrix} 1 & 0.9450 & 0.9606 \\ 0.9450 & 1 & 0.9786 \\ 0.9606 & 0.9786 & 1 \end{pmatrix},$$

which suggest that the data seem to be highly concentrated. All 2^3 possible models fitted with a combination of any of these covariates and computed several model selection methods values for each model (see Tables 6.17 to 6.19). The best three selected models based on each version of AIC, Cp, and SIC methods are given in Table 6.16.

We observe from the table that all of the commonly used measures of selection model fail to focus on best variables. Tables 6.17 to 6.19 present the commonly used model selection AIC, Cp, and SIC together with robust $RAIC$, RCp, $RSIC$ methods and AIC_R, Cp_R, and SIC_R. It is clear from the results presented in this table that variable selected by the classical selection methods are not correct enough. Though the robust model selection based on M-estimation is also sensitive to high leverage points, the table shows that they fail to choose the first variable (Hawkins). Robust model selection based on the diagnostic tool suggests that first observation (Hawkins) is best variable. When we apply the diagnostic checking based on LMS and hat matrix cases 1 to 14 return to the contamination subset and thus the AIC_R, and Cp_R finally identify the first variable as best variable. And SIC_R tends to chose Kass as best variable.

Table 6.16: Hawkins-Bradu-Kass, the selected best variables from best three models based on different classical criteria, robust criteria with M-estimation, and robust criteria using a deletion estimate of the scale

Criteria	Selected variables		
	Best model	Second best model	Third best model
AIC	Kass	Hawkins	Bradu
$RAIC$	Kass	Bradu	Hawkins
AIC_R	Hawkins	Kass	Bradu
Cp	Hawkins, Bradu	Hawkins, Bradu, Kass	Hawkins
RCp	Kass	Hawkins, Bradu	Bradu, Kass
Cp_R	Hawkins	Hawkins, Bradu	Bradu
SIC	Kass	Bradu, Kass	Hawkins
$RSIC$	Kass	Hawkins, Bradu	Bradu, Kass
SIC_R	Kass	Hawkins	Bradu

Table 6.17: Values of the classical AIC, and robust $RAIC$, and AIC_R statistics for Hawkins-Bradu-Kass data

Selected Variables	AIC	$RAIC$	AIC_R
(y, Hawkins)	5.68	4.81	**2.74**
(y, Bradu)	5.79	4.14	2.80
(y, Kass)	**5.63**	**3.62**	2.77
(y, Hawkins, Bradu)	7.68	5.62	4.73
(y, Hawkins, Kass)	7.62	5.79	4.69
(y, Bradu, Kass)	7.57	5.67	4.75
(y, Hawkins, Bradu, Kass)	9.56	7.76	6.66

Table 6.18: Values of the classical Cp, and robust RCp, and Cp_R statistics for Hawkins-Bradu-Kass data

Selected Variables	Cp	RCp	Cp_R
(y, Hawkins)	5.30	130.77	**1.19**
(y, Bradu)	8.90	32.55	2.26
(y, Kass)	17.93	**-9.72**	3.16
(y, Hawkins, Bradu)	**2.93**	-7.53	2.03
(y, Hawkins, Kass)	6.68	4.13	3.14
(y, Bradu, Kass)	10.84	-4.38	4.26
(y, Hawkins, Bradu, Kass)	4.00	4.00	4.00

Table 6.19: Values of the classical SIC, and robust $RSIC$, and SIC_R statistics for Hawkins-Bradu-Kass data

Selected Variables	SIC	$RSIC$	SIC_R
(y, Hawkins)	1.79	0.9	-1.04
(y, Bradu)	1.90	0.26	-1.02
(y, Kass)	**1.75**	**-9.72**	**-1.063**
(y, Hawkins, Bradu)	1.85	-0.20	-0.97
(y, Hawkins, Kass)	1.80	-0.03	-1.01
(y, Bradu, Kass)	**1.75**	-0.15	-0.99
(y, Hawkins,Bradu, Kass)	1.79	-0.15	-0.94

6.4.3 Example 3 (Large Data)

We consider the Ozone data which have been described in Section 3.6. We fit the robust regression model to the data set using M-estimation. The parameter estimates are given by, Intercept = -0.6099, temp= 18.6740, invHt = -2.8511, press =0.1824, vis = -2.3249,

milPress= -4.1602, hum= 5.1124, invTemp = 7.8227, and wind =1.8313.

Now, we apply the $ada\text{-}LASSO_R$ statistic to detect best variables in the ozone data. The $ada\text{-}LASSO$ model is also fitted and compared with the LS estimator as in Table 6.20. The root mean squared prediction error ($RMSPE$) for all methods are then computed. The LS model yields a significant effect of the temp and hum, and its $RMSPE$ is 5.9123. In spite of both $ada\text{-}LASSO$ and robust $ada\text{-}LASSO_R$ yield zero coefficients of the milPress and wind, the effect of press becomes non zero and the effect of invTemp becomes zero in the $ada\text{-}LASSO_R$. Three covariates (milPress, invTemp, and wind) vanish in the $ada\text{-}LASSO_R$. According to the reported values of $RMSPE$, the difference between two values remains very small.

Table 6.20: Estimation results of Ozone data

Variable	LS (p-value)	$ada\text{-}LASSO$	$ada\text{-}LASSO_R$
intercept	-1.1681(0.6119)	-1.8422	1.6455
temp	18.6244$(0.0000)^\star$	18.5883	16.1934
invHt	-2.5980(0.0643)	-3.2466	-3.9626
press	0.2766(0.8899)	0	2.2340
vis	-2.2896(0.0847)	-1.7336	-1.7255
milPress	-4.3264(0.2084)	0	0.0000
hum	5.2074$(0.0002)^\star$	5.8341	2.7150
invTemp	9.0848(0.0654)	5.2498	0.0000
wind	1.4159(0.5832)	0	0.0000
RMSPE	**5.9123**	**4.4923**	**4.9983**

Table 6.21: Comparison in model selection of Ozone data

Variable	LS	$ada\text{-}LASSO$	$ada\text{-}LASSO_R$
intercept	N	N	N
temp	N	N	N
invHt	N	N	N
press	N	Z	N
vis	N	N	N
milPress	N	Z	Z
hum	N	N	N
invTemp	N	N	Z
wind	N	Z	Z

N:Non zero variable; Z:Zero variable

6.4.4 Example 4 (Large Data)

In this section, we will consider the prostate cancer data again as given in Section 5.10.2. The robust regression model on the data set is given by, Intercept = 0.7428, icavol = 0.5882, iweight = 0.4648, age = -0.0205, ibph = 0.1304, svi = 0.7923, icp = -0.1416, gleason = 0.0282, and pgg45 = 0.0062. The parameter estimates using LS, $ada\text{-}LASSO$, and $LASSO_R$ for prostate cancer data is given in Table 6.22. Note that the model selection of parameter estimates of the $ada\text{-}LASSO$ model are quite close to the $LASSO_R$ estimates. The root mean squared prediction error ($RMSPE$) of estimators parameters are obtained and are given in the last row of Table 6.22. The differences among the values are reasonably small.

Table 6.22: Estimation results of Prostate Cancer data

Variable	LS	$ada\text{-}LASSO$	$adaLASSO_R$
Intercept	0.6694	0.0677	-0.6145
icavol	0.5870	0.5628	0.5471
iweight	0.4545	0.4161	0.6102
age	-0.0196	0	0
ibph	0.1071	0.0139	0
svi	0.7662	0.5993	0.5470
icp	-0.1055	0	0
gleason	0.0451	0	0
pgg45	0.0045	0	0
$RMSE$	0.6748	0.6926	0.7294

Table 6.23: Comparison in model selection of Prostate Cancer data

Variable	LS	$ada\text{-}LASSO$	$ada\text{-}LASSO_R$
Intercept	N	N	N
temp	N	N	N
invHt	N	N	N
press	N	Z	Z
vis	N	N	Z
milPress	N	N	N
hum	N	Z	Z
invTemp	N	Z	Z
wind	N	Z	Z

N: Non zero variable; Z: Zero variable

6.5 Comparison Between the Proposed Methods

Since our generated data in simulation for variable selection based on high breakdown scale estimate which is presented in Chapter 4, Section 4.7, is the same as the generated data in simulation for diagnostic variable selection which is presented in Chapter 6, Section 6.3, the methods in Chapters 4 and 6 are compared in this section.

For data without outliers, the results for AIC_{LTS}, Cp_{LTS} and SIC_{LTS} methods were quite similar to diagnostic- methods, as we can seen from Tables 4.1, 4.4, and 4.7, the percentage of selecting the true model was $AIC_{LTS} = 57\%$, $Cp_{LTS} = 65\%$, and $SIC_{LTS} = 45.2\%$, respectively, and for the diagnostic- methods presented in Table 6.1 was, $AIC_R = 59\%$, $Cp_R = 62\%$, and $SIC_R = 63.8\%$. In addition, we can see that

the diagnostic methods selected the true model with higher percentage than methods with high break down scale when the verticals present in data. In addition, Tables 4.2, 4.6, and 4.9 show the results when bad leverage points are present in data (when 10% bad leverage in data set, $AIC_{LTS} = 63.8\%$, $Cp_{LTS} = 33.4\%$, and $SIC_{LTS} = 43\%$), which comparable with Table 6.2 (when 10% bad leverage in data set, $AIC_R = 57\%$, $Cp_R = 62\%$, and $SIC_R = 62\%$), however, the diagnostic-methods outperform all the other methods.

It can be concluded that, for the cleaned data, the variable selection methods based on high breakdown scale estimate outperform the diagnostic-methods, otherwise, the diagnostic methods are superior.

Figures (5.1) and (6.1) show that the GM-$LASSO$ is superior to the MM-$LASSO$ and diagnostic ada-$LASSO$ methods when the data are uncontaminated. For contaminated data, Figures (5.2) to (5.10) and Figures (6.2) to (6.7) clearly show that diagnostic-ada-$LASSO$ estimator is superior.

6.6 Summary

A regression diagnostic measure is a robust regression method frequently used in practice. Nevertheless, it has not been applied to variable selection. This chapter introduced the diagnostic- variable selection and diagnostic-ada-$LASSO$ estimator, which combine diagnostic measures and variable selection via selection criteria and ada-$LASSO$ to overcome outliers and variable selection problems. Furthermore, the simulation results illustrated the excellent performance of the diagnostic-variable selection method and showed that it performed similar to or even better than the variable selection methods based on M-estimators and high breakdown point scale estimators. As such, the advantage of pro-

posed methods over the breakdown point was discussed.

CHAPTER 7

CONCLUSIONS

7.1 Summary

This study looks at some problems related to variable selection criteria in the regression model. Few published works can be found on the problem of robust variable selection criteria, and none of the subject area of robust variable selection with resisting to leverage point outliers. In this study, we specifically choose the AIC (Akaike, 1973), Mallow's Cp (Mallows, 1973b), Schwarz information criteria SIC (Akaike, 1998) as variable selection procedures and $LASSO$ regression models proposed by Tibshirani (1996) due to its interesting properties. The first three methods deal with small data sets in regression model and last one deals with multicollinearity and large data set in regression models. We look at three problems associated with the model building methods in regression models.

Firstly, we look at the problem of effect leverage points in the existing robust variable selection methods based on M-estimation. Hence, we first derive the influence function of such measures and consider its properties; then we apply two different robust estimators to select best variables in small data sets based on the high breakdown point scale; LTS, LMS, and BS robust regression methods, which are frequently used in practice. Nevertheless, they are not commonly used in selection models. This research had introduced variable selection criterion based on the LTS, LMS, and BS scale, which are robust against outliers and leverage points. The influence function of the variable selec-

tion criteria for linear regression model based on the generalized scale approach has been derived and discussed. From the cases considered, we conclude that the performance of the variable selection procedures are good, with the high breakdown point scale of LTS, LMS, and BS found to be superior than that methods based on M-estimation for small sample size. For illustration, we apply the procedures on the Stack Loss data set. The second methods based on regression diagnostics, the utility of our newly proposed methods for the detection of regression best variables are studied by Monte Carlo simulations and some well-known data sets.

Secondly, we look at the problem of selection variable in large data set in regression models. The best variable of the data can be obtained using the Least Absolute Shrinkage and Selection Operator $LASSO$ regression method. However, the $LASSO$ estimates are shown to be sensitive to the occurrence of outliers. Hence, we apply three different statistics to robust $LASSO$ based on GM-, MM- loss function and procedure of detecting the problem based on diagnostic tool statistic.

7.2 Contributions

This work has contributed to variable selection methods, analysis in the following ways:

1. We have shown that the classical model selection, such as, Akaike Information, Mallows,s Cp and Schwarz information criteria, of linear regression models are and the existing criteria ($RAIC$, RCp, $RSIC$), based on M-estimators are not robust toward the occurrence of leverage points. Therefore, it is important to develop relevant methods to robustify criteria for further investigation purposes.

2. We derive the influence function of such criteria and study its properties.

3. We have considered two robust methods, procedures to robustify criteria in regres-

sion models using high breakdown point estimators. In simulation, the procedures have been shown to perform well in variable selection in the presence of outliers.

4. We have introduced $LASSO$ variable selection in regression models for accommodating multicollinearity and large data sets in the models. The relevant theory is presented and, via simulation, the method is found to be sensitive to outliers in the data.

5. We have looked at the problem of outliers and leverage points in the $LASSO$ methods. The relevant, robust process of resolving the problem in the model has been presented. We extend the idea of the Huber-$LASSO$ approach in linear regression case to the GM-$LASSO$ and MM-$LASSO$ to give the robust variable selection against leverage pint.

6. We have developed a new robust variable selection and $LASSO$ regression by using the regression diagnostics. The outliers in the regression model were identified using suitable methods. We demonstrate that the diagnostic methods methods perform well when investigated via simulation.

7.3 Further Research

There are various possibilities for further inquiry in this field. Some suggestions are as follows:

(i) While our study has concentrated on the AIC, Cp, SIC, $LASSO$ as variable selection tool, it might be of interest to extend other robust variable selection methods that currently mainly deal with M-estimators, to more advanced robust estimation methods, such as GM or MM-estimators.

(ii) To develop some effective procedures of variable selection as in regression models.

(iii) To extend the idea of $LASSO$ with diagnostic tool statistic in linear regression case to the logistic regression case to give the best variable selection of the model.

(iv) This study have considered regression model with continues variables; however, future studies might consider mixed variables (i.e. continues and dummy) logistic regression model.

We recognize that there are still many problems ready to be explored in variable selection problem for future works.

CHAPTER 8

LIST OF PUBLICATIONS

8.1 Articles

1. Saleh, S., Hamzah, N. A., & Yunus, R. M. (2014, July). A robust version of Schwarz information criterion based on LTS. In PROCEEDINGS OF THE 21ST NATIONAL SYMPOSIUM ON MATHEMATICAL SCIENCES (SKSM21): Germination of Mathematical Sciences Education and Research towards Global Sustainability (Vol. 1605, pp. 967-972). AIP Publishing.

2. Saleh, S. (2014). Robust AIC with High Breakdown Scale Estimate. Journal of Applied Mathematics, Volume 2014, Article ID 286414, 7 pages. (ISI-cited)

3. Saleh, S. (2014). MODEL SELECTION VIA ROBUST VERSION OF R-SQUARED. Journal of Mathematics and Statistics, 10(3), 414-420.

4. SHOKRYA SALEH & Rossita M. Yunus, A note on Robust LASSO Estimators, submitted to Journal, Journal of the Iranian Statistical Society. (JIRSS) (Submitted)

5. SHOKRYA SALEH. Diagnostic Regression Shrinkage and Variable Selection, submitted to Journal, Journal Teknologi (Submitted)

8.2 Conference Attended

1. THE 1st ISM INTERNATIONAL STATISTICAL CONFERENCE 2012, JOHOR, MALAYSIA.

2. INTERNATIONAL CONFERENCE ON MATHEMATICAL SCIENCES AND STATISTICS 2013, KUALA LUMPUR, MALAYSIA.

3. Simposium Kebangsaan Sains Matematik Ke-21, 2013, P.Pinang, Malaysia.

4. Simposium Kebangsaan Sains Matematik Ke-22, 2014, Shah Alam, Malaysia.

Bibliography

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the institute of statistical mathematics 21*(1), 243–247.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory*, pp. 267–281. Akademinai Kiado.

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

Anderson-Sprecher, R. (1994). Model comparisons and R^2. *The American Statistician 48*(2), 113–117.

Arslan, O. (2012). Weighted LAD-lasso method for robust parameter estimation and variable selection in regression. *Computational Statistics & Data Analysis 56*(6), 1952–1965.

Atkinson, A. C. (1985). *Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis*. Clarendon Press Oxford.

Barnett, V. and T. Lewis (1994). *Outliers in statistical data*, Volume 3. Wiley New York.

Baum, C. F., M. E. Schaffer, and S. Stillman (2003). Instrumental variables and GMM: Estimation and testing. *Stata journal 3*(1), 1–31.

Beaton, A. E. and J. W. Tukey (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics 16*(2), 147–185.

Beckman, R. J. and R. D. Cook (1983). Outlier..... s. *Technometrics 25*(2), 119–149.

Belsey, D. A., E. Kuh, and R. E. Welsch (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley.

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

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 1705–1732.

Birkes, D. and Y. Dodge (2011). *Alternative methods of regression*, Volume 190. John Wiley & Sons.

Breiman and Leo (1996). Heuristics of instability and stabilization in model selection. *The annals of statistics 24*(6), 2350–2383.

Breiman, L. and J. H. Friedman (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association 80*(391), 580–598.

Breiman, L. and P. Spector (1992). Submodel selection and evaluation in regression. the x-random case. *International statistical review/revue internationale de Statistique*, 291–319.

Brownlee, K. A. (1965). Statistical theory and methodology in science and engineering. *A Wiley Publication in Applied Statistics, New York: Wiley, 1965, 2nd ed. 1*.

Bunea, F., A. Tsybakov, and M. Wegkamp (2007). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics 1*, 169–194.

Candes, E. and T. Tao (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 2313–2351.

Chatterjee, S. and A. S. Hadi (2009). *Sensitivity analysis in linear regression*, Volume 327. John Wiley & Sons.

Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association 74*(365), 169–174.

Cook, R. D. (2000). Detection of influential observation in linear regression. *Technometrics 42*(1), 65–68.

Croux, C. and C. Dehon (2003). Estimators of the multiple correlation coefficient: Local robustness and confidence intervals. *Statistical Papers 44*(3), 315–334.

Donoho, D. L. (2006). For most large underdetermined systems of linear equations the minimal ℓ_1- norm solution is also the sparsest solution. *Communications on pure and applied mathematics 59*(6), 797–829.

Donoho, D. L. and M. Elad (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences 100*(5), 2197–2202.

Donoho, D. L. and X. Huo (2001). Uncertainty principles and ideal atomic decomposition. *Information Theory, IEEE Transactions on 47*(7), 2845–2862.

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

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

Fan, J. and H. Peng (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics 32*(3), 928–961.

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

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

Hadi, A. S. (1992). A new measure of overall potential influence in linear regression. *Computational Statistics & Data Analysis 14*(1), 1–27.

Hahn, G. J. (1973). Coefficient of determination exposed. *CHEMISCHE TECHNIK* (OCT), 609–612.

Hampel, F. R. (1968). *Contributions to the theory of robust estimation*. University of California.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (2011). *Robust statistics: the approach based on influence functions*, Volume 114. John Wiley & Sons.

Hannan, E. J. and B. G. Quinn (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 190–195.

Hawkins, D. M., D. Bradu, and G. V. Kass (1984). Location of several outliers in multiple-regression data using elemental sets. *Technometrics 26*(3), 197–208.

Hill, R. W. (1977). *Robust regression when there are outliers in the carriers*. Ph. D. thesis, Harvard University.

Hoaglin, D. C., F. Mosteller, and J. W. Tukey (1983). *Understanding robust and exploratory data analysis*, Volume 3. Wiley New York.

Hocking, R. and R. Leslie (1967). Selection of the best subset in regression analysis. *Technometrics 9*(4), 531–540.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics 1*(5), 799–821.

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

Imon, A. (1996). Subsample methods in regression residual prediction and diagnostics. *Unpublished Ph. D. thesis, School of Mathematics and Statistics, University of Birmingham, UK*.

Imon, A. (2002). Identifying multiple high leverage points in linear regression. *Journal of Statistical Studies 3*, 207–218.

Jacod, J. and P. Protter (2000). Convergence of random variables. In *Probability Essentials*, pp. 137–145. Springer.

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

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

Koenker, R. (2007). quantreg: Quantile regression. R package version 4.10.

Kraemer, N. and J. Schaefer (2010). parcor: Regularized estimation of partial correlation matrices. R package version 0.2-2.

Krasker, W. S. and R. E. Welsch (1982). Efficient bounded-influence regression estimation. *Journal of the American statistical Association 77*(379), 595–604.

Kvålseth, T. O. (1985). Cautionary note about R^2. *The American Statistician 39*(4), 279–285.

Lambert-Lacroix, S. and L. Zwald (2011). Robust regression through the Huber's criterion and adaptive lasso penalty. *Electronic Journal of Statistics 5*, 1015–1053.

Leng, C., Y. Lin, and G. Wahba (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica 16*(4), 1273.

Leroy, A. M. and P. J. Rousseeuw (1987). Robust regression and outlier detection. *J. Wiley&Sons, New York*.

Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of statistics 2*, 90–102.

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

Mallows, C. L. (1973a). Influence functions. In *Unpublished paper presented at a conference on robust regression held at Cambridge, Mass., and sponsored by the National Bureau of Economic Research*.

Mallows, C. L. (1973b). Some comments on C_p. *Technometrics 15*(4), 661–675.

Mallows, C. L. (1975). On some topics in robustness. *Unpublished memorandum, Bell Telephone Laboratories, Murray Hill, NJ*.

Mansfield, E. R. and B. P. Helms (1982). Detecting multicollinearity. *The American Statistician 36*(3a), 158–160.

Maronna, R. A., R. D. Martin, and V. J. Yohai (2006). *Robust statistics*. Wiley Chichester.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 1436–1462.

Miller, A. J. (1984). Selection of subsets of regression variables. *Journal of the Royal Statistical Society. Series A (General)*, 389–425.

Montgomery, D. C., E. A. Peck, and G. G. Vining (2012). *Introduction to linear regression analysis*, Volume 821. John Wiley & Sons.

Osborne, M. R., B. Presnell, and B. A. Turlach (2000). On the lasso and its dual. *Journal of Computational and Graphical statistics 9*(2), 319–337.

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

Rahmatullah Imon, A. (2005). Identifying multiple influential observations in linear regression. *Journal of Applied statistics 32*(9), 929–946.

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

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

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. *status: published*.

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

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

Rousseeuw, P. J. and K. Van Driessen (2006). Computing LTS regression for large data sets. *Data mining and knowledge discovery 12*(1), 29–45.

Rousseeuw, P. J. and B. C. Van Zomeren (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association 85*(411), 633–639.

Ryan, T. P. (2008). *Modern regression methods*, Volume 655. John Wiley & Sons.

Scheffe, H. (1999). *The analysis of variance*, Volume 72. John Wiley & Sons.

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

Singh, D., P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer cell 1*(2), 203–209.

Stamey, T. A., J. N. Kabalin, J. E. McNeal, I. M. Johnstone, F. Freiha, E. A. Redwine, and N. Yang (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. radical prostatectomy treated patients. *The Journal of urology 141*(5), 1076–1083.

Swindel, B. F. (1976). Good ridge estimators based on prior information. *Communications in Statistics-Theory and Methods 5*(11), 1065–1075.

Tharmaratnam, K. and G. Claeskens (2013). A comparison of robust versions of the AIC based on M-, S-and MM-estimators. *Statistics 47*(1), 216–235.

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

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on 55*(5), 2183–2202.

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

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing*. Academic Press.

Willett, J. B. and J. D. Singer (1988). Another cautionary note about R^2: Its use in weighted least-squares regression analysis. *The American Statistician 42*(3), 236–238.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics 15*(2), 642–656.

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

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

APPENDIX 1

Belgian Telephone Data

Obs. No.	year	calls
1	50	4.4
2	51	4.7
3	52	4.7
4	53	5.9
5	54	6.6
6	55	7.3
7	56	8.1
8	57	8.8
9	58	10.6
10	59	12.0
11	60	13.5
12	61	14.9
13	62	16.1
14	63	21.2
15	64	119.0
16	65	124.0
17	66	142.0
18	67	159.0
19	68	182.0
20	69	212.0
21	70	43.0
22	71	24.0
23	72	27.0
24	73	29.0

APPENDIX 2

Hawkins-Bradu-Kass Data

Obs. No.	Hawkins	Bradu	kass	y
1	10.1	19.6	28.3	9.7
2	9.5	20.5	28.9	10.1
3	10.7	20.2	31.0	10.3
4	9.9	21.5	31.7	9.5
5	10.3	21.1	31.1	10.0
6	10.8	20.4	29.2	10.0
7	10.5	20.9	29.1	10.8
8	9.9	19.6	28.8	10.3
9	9.7	20.7	31.0	9.6
10	9.3	19.7	30.3	9.9
11	11.0	24.0	35.0	-0.2
12	12.0	23.0	37.0	-0.4
13	12.0	26.0	34.0	0.7
14	11.0	34.0	34.0	0.1
15	3.4	2.9	2.1	-0.4
16	3.1	2.2	0.3	0.6
17	0.0	1.6	0.2	-0.2
18	2.3	1.6	2.0	0.0
19	0.8	2.9	1.6	0.1
20	3.1	3.4	2.2	0.4
21	2.6	2.2	1.9	0.9
22	0.4	3.2	1.9	0.3
23	2.0	2.3	0.8	-0.8
24	1.3	2.3	0.5	0.7
25	1.0	0.0	0.4	-0.3
26	0.9	3.3	2.5	-0.8
27	3.3	2.5	2.9	-0.7
28	1.8	0.8	2.0	0.3
29	1.2	0.9	0.8	0.3
30	1.2	0.7	3.4	-0.3
31	3.1	1.4	1.0	0.0
32	0.5	2.4	0.3	-0.4
33	1.5	3.1	1.5	-0.6
34	0.4	0.0	0.7	-0.7
35	3.1	2.4	3.0	0.3
36	0.1	2.2	2.7	-1.0
37	0.1	3.0	2.6	-0.6

Obs. No.	Hawkins	Bradu	kass	y
38	1.5	1.2	0.2	0.9
39	2.1	0.0	1.2	-0.7
40	0.5	2.0	1.2	-0.5
41	3.4	1.6	2.9	-0.1
42	0.3	1.0	2.7	-0.7
43	0.1	3.3	0.9	0.6
44	1.8	0.5	3.2	-0.7
45	1.9	0.1	0.6	-0.5
46	1.8	0.5	3.0	-0.4
47	3.0	0.1	0.8	-0.9
48	3.1	1.6	3.0	0.1
49	3.1	2.5	1.9	0.9
50	2.1	2.8	2.9	-0.4
51	2.3	1.5	0.4	0.7
52	3.3	0.6	1.2	-0.5
53	0.3	0.4	3.3	0.7
54	1.1	3.0	0.3	0.7
55	0.5	2.4	0.9	0.0
56	1.8	3.2	0.9	0.1
57	1.8	0.7	0.7	0.7
58	2.4	3.4	1.5	-0.1
59	1.6	2.1	3.0	-0.3
60	0.3	1.5	3.3	-0.9
61	0.4	3.4	3.0	-0.3
62	0.9	0.1	0.3	0.6
63	1.1	2.7	0.2	-0.3
64	2.8	3.0	2.9	-0.5
65	2.0	0.7	2.7	0.6
66	0.2	1.8	0.8	-0.9
67	1.6	2.0	1.2	-0.7
68	0.1	0.0	1.1	0.6
69	2.0	0.6	0.3	0.2
70	1.0	2.2	2.9	0.7
71	2.2	2.5	2.3	0.2
72	0.6	2.0	1.5	-0.2
73	0.3	1.7	2.2	0.4
74	0.0	2.2	1.6	-0.9
75	0.3	0.4	2.6	0.2

APPENDIX 3

Stack Loss Data

Obs. No.	Air.Flow	Water.Temp	Acid.Conc.	stack.loss
1	80	27	89	42
2	80	27	88	37
3	75	25	90	37
4	62	24	87	28
5	62	22	87	18
6	62	23	87	18
7	62	24	93	19
8	62	24	93	20
9	58	23	87	15
10	58	18	80	14
11	58	18	89	14
12	58	17	88	13
13	58	18	82	11
14	58	19	93	12
15	50	18	89	8
16	50	18	86	7
17	50	19	72	8
18	50	19	79	8
19	50	20	80	9
20	56	20	82	15
21	70	20	91	15

APPENDIX 4

Proof of Proposition 3.4[1]

Let $\mathbf{A}_n = \mathbf{A}$ implies that $\hat{\beta}_j = 0$ for all $j \in \mathbf{A}$. Let $u^\star = \arg\min(V_2(u))$. Note that

$$P\left(\mathbf{A}_n = \mathbf{A}\right) \leq P\left(\sqrt{n}\hat{\beta}j = 0 \forall j \notin \mathbf{A}\right).$$

Lemma 3.3 show that

$$\sqrt{n}\hat{\boldsymbol{\beta}}_{\mathbf{A}} \rightarrow_d u_A^\star.$$

Thus the weak convergence results that

$$\limsup_n P\left(\sqrt{n}\hat{\beta}_j = 0 \forall j \notin \mathbf{A}\right) \leq P\left(u_A^\star = 0 \forall j \notin \mathbf{A}\right).$$

Therefore, only need to show that

$$c = P\left(u_A^\star = 0 \forall j \notin \mathbf{A}\right) < 1.$$

There are two cases:

Case 1. $\lambda_0 = 0$, then it is easy to see that $u^\star = \mathbf{C}^{-1}\mathbf{W} \sim N(0, \sigma^2\mathbf{C}^{-1})$, and so $c = 0$.

Case 2. $\lambda_0 > 0$, then $V_2(u)$ is not differentiable at $u_j = 0\ \forall j \in \mathbf{A}$.

By the Karush-Kuhn-Tucker (KKT) optimality condition,

$$-2\mathbf{W}_j + 2(\mathbf{C}u^\star)_j + \lambda_0 sgn(\boldsymbol{\beta}_{\mathbf{A}}^\star) = 0, \forall j \in \mathbf{A} \tag{1}$$

and

$$\mid -2\mathbf{W}_j + 2(\mathbf{C}u^\star)_j \mid \leq \lambda_0, \forall j \notin \mathbf{A}. \tag{2}$$

[1]The references of this proof are based on (Zou, 2006)

If $u_j^\star = 0$ for all $j \notin \mathbf{A}$, then Eqn. (1) and Eqn. (2) become,

$$-2\mathbf{W_A} + 2\mathbf{C}_{11}u_{\mathbf{A}}^\star + \lambda_0 sgn(\boldsymbol{\beta}_{\mathbf{A}}^\star) = 0 \tag{3}$$

and,

$$\mid -2\mathbf{W}_{\mathbf{A}^c} + 2\mathbf{C}_{21}u_{\mathbf{A}}^\star \mid \leq \lambda_0. \tag{4}$$

component wise

Combining Eqn. (3) and Eqn. (4) gives

$$\mid -2\mathbf{W}_{\mathbf{A}^c} + 2\mathbf{C}_{21}\mathbf{C}_{11}^{-1}(2\mathbf{W_A} - \lambda_0 sgn(\boldsymbol{\beta}_{\mathbf{A}}^\star)) \mid \leq \lambda_0 \text{ component wise.}$$

Thus,

$$c \leq P\left(\mid -2\mathbf{W}_{\mathbf{A}^c} + 2\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\left(2\mathbf{W_A} - \lambda_0 sgn(\boldsymbol{\beta}_{\mathbf{A}}^\star)\right) \mid \leq \lambda_0\right) < 1. \tag{5}$$

Theorem .1. ***(Slutsky's)***[2] *Let a_n and b_n are A sequence of random variables then : If $a_n \to_d a$ and $b_n \to_d b$ where a is a random variable and b is a constant, then*

- $a_n + b_n \to_d a + b$
- $a_n b_n \to_d ab$
- $a_n/b_n \to_d a/b$ *provided* $P[b = 0] = 0$.

[2]the Proof of this theorem available on Jacod and Protter (2000)

APPENDIX 5

Proof of Theorem 3.5[3]

Let $\boldsymbol{\beta}^{\star} = \boldsymbol{\beta} + \frac{\mathbf{u}}{\sqrt{n}}$, and

$$v_n(\mathbf{u}) = \sum_{i=1}^{n} \left(y_i - \mathbf{X}^T(\beta_j + \frac{u}{\sqrt{n}}) \right)^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j \left| \beta_j + \frac{u}{\sqrt{n}} \right|. \tag{6}$$

Let $\hat{\mathbf{u}}^{(n)} = \arg\min v_n(\mathbf{u})$, then $\hat{\boldsymbol{\beta}}^{(n)} = \boldsymbol{\beta} + \frac{\hat{\mathbf{u}}^{(n)}}{\sqrt{n}}$ or $\hat{\mathbf{u}}^{(n)} = \sqrt{n} \times (\hat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta})$. Note that $v_n(\mathbf{u}) - v_n(\mathbf{0}) = V_n(\mathbf{u})$, where

$$V_n(\mathbf{u}) = \left[u^T \mathbf{C}_n u - 2\mathbf{W}u + \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} \hat{w}_j \sqrt{n} \left(\left| \beta_j + \frac{u}{\sqrt{n}} \right| - \mid \beta_j \mid \right) \right], \tag{7}$$

where $\mathbf{C} \to \frac{1}{n}\mathbf{x}_i^T \mathbf{x}_i$ and $\mathbf{W} \to (\varepsilon_i \mathbf{x}_i)/\sqrt{n} \sim N(0, \sigma^2 \mathbf{C})$.

Now consider the limiting behavior of the third term in Eqn. (7). If $\beta_j \neq 0$, then

$$\hat{w}_j \to_p 1/ \mid \beta_j \mid^{\gamma}$$

and

$$\sqrt{n} \left(\left| \beta_j + \frac{u_j}{\sqrt{n}} \right| - |\beta_j| \right) \to u_j sgn(\beta_j).$$

By Slutsky's theorem,

$$\lambda_n / \sqrt{n} \hat{w}_j \sqrt{n} \left(\left| \beta_j + \frac{u_j}{\sqrt{n}} \right| - \mid \beta_j \mid \right) \to 0.$$

If $\beta_j = 0$, then $\sqrt{n} \left(\left| \beta_j + \frac{u_j}{\sqrt{n}} \right| - \mid \beta_j \mid \right) = \mid u_j \mid$ and $\lambda_n / \sqrt{n} \hat{w}_j = \lambda_n / \sqrt{n} n^{\gamma/2} \left(\left| \sqrt{n} \hat{\beta}_j^{\star} \right| \right)^{-\gamma}$, where $\sqrt{n} \hat{\beta}_j^{\star} = O_p(1)$. thus, again by Slutsky's theorem, see that $V^{(n)}(u) \to_d V(u)$ for

[3]The references of this proof are based on (Zou, 2006)

every u, where

$$V(u) = \begin{cases} -2\mathbf{W}_{\mathbf{A}^\star} u^T_{\mathbf{A}^\star} + u^T_{\mathbf{A}^\star} \mathbf{C}_{(11)} u_{\mathbf{A}^\star}, & \text{If} \quad u_j = 0 \forall j \notin \mathbf{A}^\star, \\ \infty, & \text{elsewhere} \end{cases} \tag{8}$$

$V^{(n)}(u)$ convex, and the unique minimum of $V(u)$ is $\left(\mathbf{C}_{11}^{-1}\mathbf{W}_{\mathbf{A}^\star}, 0\right)^T$. Following the epi-convergence results of (Geyer, 1994) and (Knight and Fu, 2000),

$$\hat{u}^{(n)} = \begin{cases} \mathbf{C}^{-1}_{(11)} \mathbf{W}_{\mathbf{A}^\star}, & \text{If} \quad \beta_j \in \mathbf{A}^\star, \\ 0, & \text{elsewhere.} \end{cases} \tag{9}$$

Finally, $\mathbf{W}_{\mathbf{A}^\star} \sim N(0, \sigma^2 \mathbf{C}_{(11)})$ are observed; then the asymptotic normality part are proved. Now the consistency part $\forall j \in \mathbf{A}^\star$ are showed, the asymptotic normality result indicates that $\hat{\beta}^{(n)}_{j^\star} \to \beta_j$ thus $P(j \in \mathbf{A}_n) \to 1$. Then it suffices to show that

$$\forall j' \notin \mathbf{A}^\star, P(j' \in A_n) \to 0.$$

Consider the event $j' \in \mathbf{A}_n$. By the KKT optimality conditions, know that, $2\mathbf{x}_i^T \left(y - \mathbf{X}\hat{\boldsymbol{\beta}}^{(n)}\right) = \lambda_n \hat{w}_{j'}$. Note that $\lambda_n \hat{w}_{j'}/\sqrt{n} = \lambda_n/\sqrt{n} n^{\gamma/2} \frac{1}{|\sqrt{n}\hat{\boldsymbol{\beta}}^\star_{j'}|} \to_p \infty$, whereas

$$\frac{2\mathbf{x}^T_{j'}\left(y - \mathbf{x}\hat{\boldsymbol{\beta}}^{(n)}\right)}{\sqrt{n}} = 2\frac{\mathbf{x}^T_{j'} X \sqrt{n}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(n)}\right)}{n} + 2\mathbf{x}^T_{j'}\frac{\varepsilon}{\sqrt{n}}. \tag{10}$$

By Eqn. (9) and Slutsky's theorem, $\frac{2\mathbf{x}^T_{j'}\mathbf{X}\sqrt{n}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(n)}\right)}{n} \to_d$ some normal distribution and $\frac{2\mathbf{x}^T_{j'}\varepsilon}{\sqrt{n}} \sim N\left(0, 4 \parallel \mathbf{x}_{j'} \parallel^2 \sigma^2\right)$. Thus

$$P\left(j' \in \mathbf{A}_n\right) \leq P\left[2\mathbf{x}^T_{j'}\left(y - \mathbf{X}\hat{\boldsymbol{\beta}}^{(n)}\right) = \lambda_n \hat{w}_{j'}\right] \to 0. \tag{11}$$