Robust regression modeling via L_1 regularization

Heewon Park

Department of Mathematics

Graduate School of Science and Engineering

March 2013

Chuo University

Acknowledgements

Standing at this delightful moment and looking back on the past, many things going through my head. Most of all, I am overwhelmed with gratitude to many people.

First and foremost I offer my sincerest gratitude to my advisor, Professor Sadanori Konishi who has supported me throughout my Ph.D study with his encouragement and valuable advice. Without his guidance and persistent help, this dissertation would not have been possible. Although it would be hard to describe how much I appreciate all of his support, meeting him is the best luckiest thing in my lifetime.

I am heartily thankful to Professor Fumitake Sakaori for his lots of help and encouragement through the whole of my research. I would like to thank Professors Toshinari Kamakura, Yoshikazu Kobayashi and Masaaki Taguri in Chuo University for their helpful comments and suggestions.

I also would like to express many thanks to friends, seniors, juniors and staff members in the department of Mathematics, Chuo University for their kind assistance.

I wish to express my deep appreciation to Professor Jong-Hyup Lee in Sungshin women's University, Korea. Without his valuable advice and encouragement, I cannot finish studying in Japan. I am also grateful to Professor Seong-Keon Lee in Sungshin women's University, Korea for his great deal of encouragement and advice.

I wish to thank my friends, seniors and juniors in Korea for their encouragement. Finally, I would like to thank my parents and younger sister from the bottom of my heart for their endless support and boundless trust.

I will never forget what many of you have done for me as long as l live.

Contents

Acknowledgements **ii**

1 Introduction **1**

2 Sparse regression modeling **10**

2.1 Motivation of L_1-type regularization 11

2.2 L_1-type regularization 14

2.2.1 Lasso 14

2.2.2 Adaptive lasso 16

2.2.3 Elastic net 17

2.2.4 Smoothly clipped absolute deviation: SCAD 18

2.3 Estimation of sparse regression model 19

2.3.1 Local quadratic approximation 20

2.3.2 LARS algorithm 21

2.3.3 Coordinate descent algorithm 23

2.4 Selection of regularization parameters 25

2.4.1 Cross-validation 26

3 Robust regression modeling via L_1 regularization **30**

3.1 Robust L_1-type regularization 32

3.1.1 Literature review: existing robust L_1-type regularization . . . 33

3.1.2 Least trimmed squares elastic net 34

3.2 Selection of tuning parameters in robust L_1-type regularization 37

3.2.1 Efficient bootstrap information criterion 39

3.2.2 Generalized information criterion 44

3.3 Robust estimation for sparse regression model 49

3.3.1 Robust LARS 50

3.3.2 Robust coordinate descent procedure 51

3.4 Robust selection of the tuning parameters 60

3.4.1 Literature review: robust model selection criteria 61

3.4.2 Robust efficient bootstrap information criterion 63

3.5 Simulation studies 67

3.6 Real-world examples 81

4 Lag weighted lasso for time series model **85**

4.1 Time series model 86

4.2 Lag weighted lasso for time series model 87

4.3 Simulation studies 91

4.4 Real-world example: Cerebrovascular Mortality data 95

5 Symbolic candle chart-valued time series **98**

5.1 Symbolic data 100

5.2 Approaches for symbolic interval-valued data 102

5.2.1 Centre and Range method 102
5.2.2 NCRM1 and NCRM2 method 103
5.3 Approach for symbolic candle chart-valued time series (CTS) 105
5.3.1 Time series model for forecasting CTS 105
5.3.2 Parameter constraint and estimation 108
5.4 Applications: Stock market indices of five major Asian countries . . . 109

6 Summary and concluding remarks **114**

Bibliography **118**

Chapter 1

Introduction

With the development of computer and data collection technologies, the database sizes continue to grow and various statistical methodologies have been developed over the past several decades, such as regularization method, novel robust modeling strategies, symbolic data analysis, etc. There is currently much discussion about the L_1-type regularized regression in various fields. By imposing an L_1-type penalty term to the least squares loss function, the L_1-type regularization can perform simultaneous parameter estimation and variable selection.

In this thesis, the following issues are discussed,

1. L_1-type regularized regression,
2. Robust regression modeling via L_1-type regularization,
3. Lag weighted lasso for time series model,
4. New type of Symbolic Data Analysis: Candle chart-valued time series.

These topics will be discussed in detail below.

L_1-type regularized regression

The first topic is the L_1-type regularization method. As data size (i.e., number of predictor variables p) continues to increase, not only estimation but also variable selection is becoming a difficult problem in regression modeling. The traditional model selection procedures, such as forward, backward, stepwise selection and all subset regression, however, are not feasible in a large number of predictor variables. Furthermore, the least squares method, which is the most widely used for regression modeling, has a demerit on multicollinearity under a large number of predictor variables, and thus suffers from high variances of the estimated parameters.

To overcome the aforementioned drawbacks, several regularization methods have been proposed. Hoerl and Kennard (1970) proposed a ridge regression imposing an L_2 norm penalty to the least squares loss function. Although the ridge regression can overcome the high variance of estimator by imposing the L_2 norm penalty, it cannot perform variable selection simultaneously, and hence the traditional model selection procedures have to be used for selecting an optimal model. This implies that the ridge regression also has the demerit in a large number of predictor variables.

Tibshirani (1996) proposed a lasso (least absolute shrinkage and selection operator), which minimizes residual sum of squares subject to the L_1 norm constraint. Unlike the ridge regression, the lasso shrinks some coefficients to exactly zero. This implies that the lasso performs estimation and variable selection simultaneously, depending on a regularization parameter. Furthermore, numerous L_1-type regularization methods have been proposed to improve modeling procedure, such as adaptive lasso (Zou, 2006), elastic net (Zou and Hastie, 2005), smoothly clipped absolute devia-

tion (Fan and Li, 2001), etc. The estimates of regression coefficients by the lasso-type approaches, however, cannot be analytically derived due to indifferentiability of the L_1-type penalty term. To settle on this issue, several effective algorithms were proposed. We briefly introduce the various L_1-type regularization methods and effective algorithms for the L_1-type regularization (e.g., Local quadratic approximation, LARS and coordinate descent algorithm).

A crucial issue in the L_1-type regularized regression is a proper choice of the regularization parameters. This issue can be viewed as a model selection and evaluation problem. The regularization parameters were often selected by the cross-validation. We discuss the methods for choosing the regularization parameters by the K-fold cross-validation and generalized cross-validation. The problem of choosing the regularization parameters is essential in regression modeling, and needs to be taken significantly. Thus we will discuss this issue in various perspectives in the next topic.

Robust regression modeling via L_1 regularization

We consider the robust regression modeling via L_1-type regularization in various aspects: methodology, model estimation and evaluation. This topic is a main part of the present thesis.

We first discuss a robust L_1-type regularization. Although the L_1-type regularization showed the superiority in regression modeling, its performance takes a sudden turn for the worst in the presence of outliers, since it is based on the penalized least squares method. To overcome the demerit, robust L_1-type approaches were proposed by replacing the least squares loss function with robust loss function: least absolute lasso (Wang et al., 2007a), M-lasso (Zhang et al., 2009) and M-adaptive lasso

(Lambert-Lacroix and Zwald, 2010). The M-estimate and LAD estimate, however, do not have a high breakdown point even though they are better than the least squares estimate (LSE). To improve robustness, we consider the least trimmed squares (LTS) estimation procedure having a high breakdown point, and propose a least trimmed squares elastic. By replace the least squares loss function with robust loss function, the robust L_1-type regularization performs well in sparse regression modeling under the properly selected regularization parameters and tuning constant, even in the presence of outliers.

In the robust sparse modeling, the selection of the regularization parameters and also a tuning constant in outlier detection is a critical issue. Although the performance of the robust sparse regression strongly depends on a proper choice of these tuning parameters, relatively little attention was paid to this issue, particularly in the presence of outliers. We propose novel methods for choosing an optimal set of the regularization parameters and tuning constant in line with the information-theoretic view point. We first introduce the use of the efficient bootstrap information criteria (Konishi and Kitagawa, 1996) for choosing an optimal set of the tuning parameters. By using the variance reduction method, the variance due to the bootstrap resampling can be significantly reduced, and thus we can expect to efficient modeling. We also consider the generalized information criterion (Konishi and Kitagawa, 1996), which can be applied to evaluate statistical models constructed by various types of estimation procedure. The calculation of an influence function is crucial in deriving an information criterion. However, it is difficult to obtain the influence function corresponding to the L_1-type regularization methods, because the L_1-type penalty functions are not differentiable in origin. To settle on the problem, we use the local

quadratic approximation of the lasso-type penalty functions (Fan and Li, 2001), and then derive an information criterion for evaluating robust sparse regression models in line with the generalized information criteria.

We then consider the robust estimation of the L_1-type regularized regression model via an outlier-resistant algorithm. The algorithms, which showed the remarkable performance for sparse regression modeling, are based on the sample mean, standard deviation and correlation or inner product estimated in a non-robust manner, and thus their procedures suffer from outliers. To overcome the drawback, Khan et al. (2007) proposed robust model selection techniques by robustify the LARS. We consider robust regression modeling via the coordinate descent procedure, which is competitive with the well known LARS for the L_1-type regularization. In order to robust sparse regression modeling, we robustify the coordinate descent procedure based on outlier-resistant inner product and pre-treatment techniques. By using the proposed robust coordinate descent procedure, we can efficiently perform robust regression modeling without much additional calculation.

Finally, we consider the robust model evaluation problem. For robust regression modeling procedures, numerous studies on the robust estimation have been conducted. However, relatively few studies have been devoted to robust model evaluation. Ronchetti et al. (1997) introduced a robust cross-validation based on a robust loss function. Jung (2009) proposed a robust generalized cross-validation by replacing the least squares loss function with median, trimmed squares and mean absolute loss functions. Ronchetti and Staudte (1994) proposed a robust version of Mallow's C_p

by using the weight based on the residual of observations. We consider the robust model evaluation criterion in line with the efficient bootstrap information criterion. Although the bootstrap information criterion has several advantages on its flexibility and weak assumptions, a bootstrap sample may contain more outliers compared with those in the original sample, since bootstrap sample is drawn randomly. This implies that the resulting criterion from the highly contaminated bootstrap sample may produce biased results. To overcome the drawback, we propose a robust efficient bootstrap information criterion via the Winsorizing technique (Srivastava et al., 2010). By using the proposed robust efficient bootstrap information criterion, we can perform effective and stable model evaluation even in the presence of outliers.

Lag weighted lasso for time series model

A time series model is usually constructed by current and past values of predictor variables and past values of response variable. In other words, a response variable is explained by a parametric function of the present and past values of predictor variables and past values of response variable. It implies that one of the important factors in the time series modeling is a length of lag. We consider the time series modeling in line with the adaptive lasso (Zou, 2006), which assigns different penalties to each coefficient based on a weight. Although the adaptive lasso showed an exceptional performance for regression modeling by imposing different weights to each coefficient, it may not give proper and interpretable results for time series model with lagged variables, since its weight does not take account of the length of lag.

We propose a lag weighted lasso, which additionally considers the effect of lag length, for time series modeling. The proposed method shrinks the coefficient based

on weights reflecting not only coefficients size but also the lag length, unlike the adaptive lasso. In other words, the coefficient of variable in the distant past with a small effect is estimated as small, or this variable is deleted from the model. In short, the proposed lag weighted lasso can reflect the properties of the time series data, and thus we can expect to improve the forecasting accuracy of time series model.

New type of Symbolic Data Analysis: Candle chart-valued time series

Database has continued to grow, and thus summarization and visualization of enormous amounts of data are increasingly important. To address this issue, symbolic data analysis (SDA), such as interval-valued data, histogram valued data, multimodal data, was introduced (Bock and Diday, 2000). The symbolic data analysis takes into account the information that cannot be represented within the classical data model, and can perform effective summarization and visualization of huge databases.

We introduce a new type of symbolic data, a candle chart-valued time series (CTS) constructed with the four stock indices (i.e., open, close, highest and lowest indices), and propose forecasting methods for CTS by using a statistical model in the viewpoint of symbolic data analysis. In order to modeling CTS, we first propose a method based on the original four stock indices consisting of the candle chart. We also propose a method based on the two mid-point time series and two half-range time series of intervals between the open and close indices, and between the highest and lowest indices respectively. By using the proposed approaches, we can forecast the direction of future stock index more accurately.

The rest of this thesis is organized as follows.

In Chapter 2, we briefly present a motivation of the regularization method, and

introduce various L_1-type regularization methods. Then, we review several algorithms for the L_1-type regularization and methods for choosing the regularization parameters.

In Chapter 3, we consider robust regression modeling via L_1-type regularization in aspects of methodology, estimation and evaluation. We first discuss about the robust L_1-type regularization, and then introduce novel methods for choosing the optimal set of the regularization parameters and tuning constant via the efficient bootstrap information criterion and generalized information criterion. We also present the robust sparse regression modeling via the outlier-resistant algorithm for L_1-type regularization, and propose the robust coordinate descent procedures based on Winsorization and trimming techniques. Finally, we introduce robust model evaluation criteria, and propose a novel robust criterion for choosing the tuning parameters, called a robust efficient bootstrap information criterion. Monte Carlo simulations and real data analysis were conducted to investigate the effectiveness of the proposed robust regression modeling strategies. We observed that the proposed robust modeling strategies perform well even in the presence of outliers.

In Chapter 4, we propose a novel L_1-type regularization method for time series model, called a lag weighted lasso. To reflect the property of the time series model constructed by lagged variables, we consider three types of weights which reflect not only coefficient size but also length of lag. We illustrate the performance of the proposed lag weighed lasso using simulation studies and real data analysis through cerebrovascular disease data.

In Chapter 5, we introduce a new type of symbolic data, a candle chart-valued time series (CTS) constructed with four stock indices (open, close, highest and lowest indices). To modeling the CTS, we propose new forecasting methods based on the four stock indices, and based on two mid-point time series and two half-range time series of intervals between open and close indices, and between the highest and lowest indices, respectively. We investigate the effectiveness of the proposed approaches through the analysis of the stock indices of five major Asian countries.

In Chapter 6, we present summary and concluding remarks.

Chapter 2

Sparse regression modeling

The regression analysis is the most widely used technique for investigating and modeling relationship between interested variable and predictor variables. Efron mentioned that *"the most important problem in statistics is a single problem: variable selection in regression. This entails selecting variables from candidate variables, estimation of parameter for those variables and inference."* (Hesterberg et al., 2008). As data size and number of predictor variables increase, not only parameter estimation but also variable selection has become increasingly important in the regression modeling to achieve the following goals (Hesterberg et al., 2008):

- Prediction accuracy,
- Interpretation,
- Stability,
- Avoiding bias hypothesis test.

Traditional variable selection methods (e.g., stepwise selection, all subset regression, etc.), however, have a demerit on unstable results, and thus we cannot expect prediction accuracy. Furthermore, the procedures are not feasible in the large number of predictor variables. In recent year, the L_1-type regularization has drawn a large amount of attention for regression modeling. By imposing an L_1-type penalty to a loss function, the L_1-type regularization methods can perform not only variable selection and estimation simultaneously, but also stable regression modeling by preventing high variances of estimates.

This chapter provides the overall procedures of the L_1-type regularized regression modeling. We first introduce a motivation of the regularization method, and briefly review the various L_1-type approaches constructed by least squares loss function with L_1-type of norm penalty. Then, we present the effective algorithms for the L_1-type regularization. Finally, we introduce a method for choosing the regularization parameters, which is a crucial issue in the L_1-type regularized regression, since it can be viewed as a model selection and estimation problem.

The rest of this chapter is organized as follows. In Section 2.1, we introduce a motivation of the regularization method. We review the L_1-type regularization in Section 2.2. Section 2.3 presents the several algorithms for the L_1-type regularization. We introduce methods for choosing the regularization parameters in Section 2.4.

2.1 Motivation of L_1-type regularization

Suppose we have n independent observations $\{(y_i, \mathbf{x}_i); i = 1, ..., n\}$, where y_i are random response variables and $\mathbf{x}_i$ are p-dimensional vectors of the predictor variables.

Consider the linear regression model,

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, ..., n, \tag{2.1}$$

where $\boldsymbol{\beta}$ is an unknown p-dimensional vector of regression coefficients and ε_i are the random errors which are assumed to be independently and identically distributed with mean 0 and variance σ^2. In the present thesis, we assume that y_i are centered and x_{ij} are standardized by their mean and standard deviation: $\sum_i^n y_i/n = 0$, $\sum_i^n x_{ij}/n = 0$ and $\sum_i^n x_{ij}^2/n = 1$.

The most widely useed method for estimating the linear regression model in (2.1) is a least squares (LS) procedure that minimizes,

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2. \tag{2.2}$$

A matrix form of LS procedure is given by,

$$RSS(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{2.3}$$

where $\mathbf{X}$ is $n \times p$ matrix with an input vector in each row and $\mathbf{y}$ is the n-vector of output. By differentiating (2.3) with respect to $\boldsymbol{\beta}$, we obtain a least squares estimator (LSE),

$$\hat{\boldsymbol{\beta}}^{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{2.4}$$

The least squares estimator is the best linear unbiased estimator. Although the LSE is an unbiased estimator having minimum variance, it may be revealed a multicollinearity in the highly correlated predictor variables and overfitting problem. In practical, we cannot avoid increasing a correlation between predictors (i.e., multicollinearity) in the large number of predictor variables. This implies that the matrix $\mathbf{X}^T\mathbf{X}$ has

similar values in off-diagonal and diagonal elements, and thus a variance of the least square estimator

$$Var(\hat{\boldsymbol{\beta}}^{LS}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \tag{2.5}$$

increases. Consequentially, this leads to unstable results as shown in Breiman (1996), because a small change in data can cause large change in modeling results. If worse comes the worst, we cannot find the least squares estimator, since the matrix $\mathbf{X}^T\mathbf{X}$ is not full rank under the exact multicollinearity.

To overcome the demerit, Hoerl and Kennard (1970) proposed a ridge regression,

$$\hat{\boldsymbol{\beta}}^{ridge} = \arg\min_{\boldsymbol{\beta}}\{RSS(\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p}\beta_j^2\}, \tag{2.6}$$

where $\lambda > 0$ is a regularization parameter controlling model complexity. By differentiating (2.6), we can find the solution of the ridge regression as follows,

$$\hat{\boldsymbol{\beta}}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}, \tag{2.7}$$

where $\mathbf{I}$ is the $p \times p$ identity matrix. The ridge regression adds the positive λ to the diagonal of matrix $\mathbf{X}^T\mathbf{X}$, and thus it prevents that the matrix $\mathbf{X}^T\mathbf{X}$ has similar values in diagonal and off-diagonal elements. This implies that the ridge regression can overcome the large variances of coefficients estimates, and hence it can perform stable regression modeling and improve prediction accuracy. Although the ridge regression is a very attractive methodology in linear regression modeling, it is a technique for estimation, and thus variable selection is conducted by using the traditional methods, such as subset selection, stepwise selection, etc.

To settle on the issue, a L_1-type regularization was proposed by imposing the L_1-type of norm penalty not L_2 norm.

2.2 L_1-type regularization

The L_1-type regularization method has been received much attention for regression modeling in various fields. By imposing the L_1-type penalty term, the following L_1-type regularization method can perform simultaneous variable selection and estimation,

$$\hat{\boldsymbol{\beta}}^{L_1.type} = \arg\min_{\boldsymbol{\beta}}\{RSS(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta|)\}, \tag{2.8}$$

where $\sum_{j=1}^{p} p_\lambda(|\beta|)$ is an L_1-type penalty with a regularization parameter λ (> 0) controlling the amount of shrinkage on the parameters. We briefly introduce various L_1-type regularization methods in this section.

2.2.1 Lasso

The lasso (least absolute shrinkage and selection operator), proposed by Tibshirani (1996), is a regularization method imposing an L_1 norm penalty on regression coefficients,

$$\hat{\boldsymbol{\beta}}^{lasso} = \arg\min_{\boldsymbol{\beta}}\{RSS(\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p} |\beta_j|\}, \tag{2.9}$$

and, an alternative formulation is given by

$$\hat{\boldsymbol{\beta}}^{lasso} = \arg\min_{\boldsymbol{\beta}}\{RSS(\boldsymbol{\beta})\}, \quad \text{subject to} \sum_{j=1}^{p} |\beta_j| \leq t. \tag{2.10}$$

The lasso is a shrinkage method similar to the ridge, but only difference being that it shrinks some coefficients to exactly zero. It means that the lasso can perform variable selection and estimation simultaneously. We briefly explain that how can the lasso estimate some coefficients to exactly zero in one dimensional situation (Sohail, 2011).

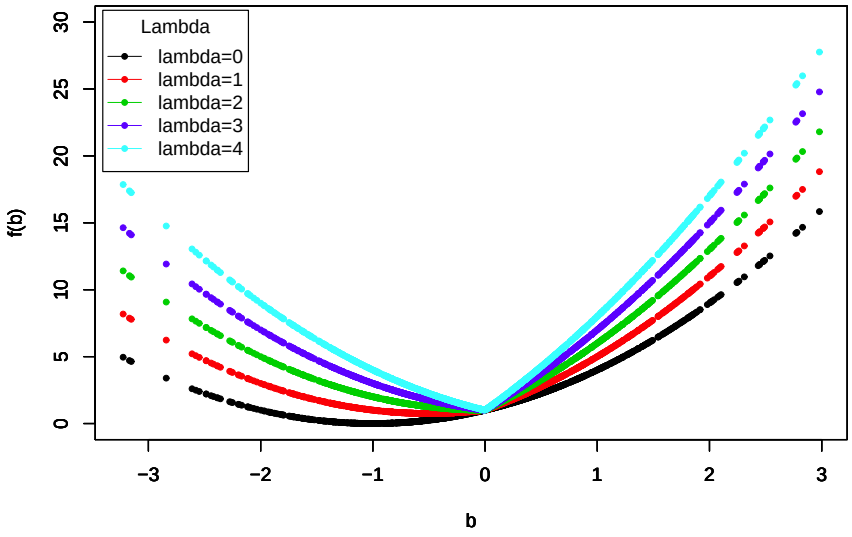


Figure 2.1: Plot of $f(b)$ with various λ

Consider the model,

$$f(b) = (b+1)^2 + \lambda|b|. \tag{2.11}$$

The first derivative of (2.11) is given by,

$$f^{'}(b) = 2(b+1) + \lambda \text{sign}(b), \tag{2.12}$$

where $\text{sign}(b) = -1, 0, 1$ for $b < 0, b = 0$ and $b > 0$ respectively, and $\lambda > 0$. Figure 2.1 shows the plot of $f(b)$ with various λ. As shown in Figure 2.1, the lasso sets b to zero when the signs of $f^{'}(b)$ and b are changed at the same time. Because of $f^{'}(b) \geq 0$ when $b \geq 0$, b becomes zero when $f^{'}(b) < 0$ and b passes through zero simultaneously.

Let $L(\beta)$ be the lasso problem in (2.9), then

$$M = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} \tag{2.13}$$
$$= -2\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \text{sign}(\beta_j).$$

Like to one dimensional problem, if a sign of M is changed when β_j passes through zero, then the lasso procedure estimates $\hat{\beta}_j$ to exactly zero, if not $\hat{\beta}_j \neq 0$. Let $A = \{j : \hat{\beta}_j \neq 0\}$ be an active set, and $A^c = \{j : \hat{\beta}_j = 0\}$ be a non-active set, and thus at $\hat{\beta}_j$

$$j \in A \quad \text{if} \quad -2\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \lambda \text{sign}(\hat{\beta}_j) = 0. \tag{2.14}$$
$$j \in A^c \quad \text{if} \quad |-2\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})| < \lambda.$$

As shown above, the lasso estimates some coefficients to exactly zero. This implies that the lasso performs not only estimation and but also variable selection simultaneously depending on the regularization parameter λ.

2.2.2 Adaptive lasso

The recent studies exposed that the lasso estimator may be inefficient and true model selection result could be inconsistent (Fan and Li, 2001; Yuan and Lin, 2007; Zou, 2006). To overcome the problem, Zou (2006) proposed an adaptive lasso imposing different penalties to each coefficient based on weight,

$$\hat{\boldsymbol{\beta}}^{adlasso} = \arg\min_{\boldsymbol{\beta}} \{RSS(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \hat{w}_j |\beta_j|\}, \tag{2.15}$$

where $\hat{\mathbf{w}} = 1/|\hat{\boldsymbol{\beta}}|^\gamma$, $\gamma > 0$ and the least square estimator or the ridge estimator can be used as $\hat{\boldsymbol{\beta}}$. In the adaptive lasso, the amount of shrinkage is controlled by $\hat{\boldsymbol{\beta}}$, i.e.,

the coefficients of variables with large effect are shrunk slightly, whereas coefficients of variables with small effect are shrunk significantly. Thus, the adaptive lasso is able to identify the true model consistently and estimator is efficient as shown in oracle properties (Fan and Li, 2001; Zou, 2006).

2.2.3 Elastic net

The recent work suggests that the lasso may have some limitations as follows (Zou and Hastie, 2005):

- In the $p > n$ case, the lasso can select at most n variables, because of the convex optimization problem.
- The lasso cannot take account the group effect of predictor variables.
- For usual $n > p$ case, if there are high correlations between predictors, the lasso is inconsistent and dominated by the ridge regression in the viewpoint of the prediction performance.

To overcome the drawbacks, Zou and Hastie (2005) proposed a new regularized technique called an elastic net as follows,

$$\hat{\boldsymbol{\beta}}^{\text{elastic net}} = \arg\min_{\boldsymbol{\beta}} \{RSS(\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2\}, \tag{2.16}$$

where λ_1 and λ_2 are regularization parameters controlling model complexity. Let $\varrho = \lambda_2/(\lambda_1 + \lambda_2)$, then (2.16) can be seen the following problem,

$$\hat{\boldsymbol{\beta}}^{\text{elastic net}} = \arg\min_{\boldsymbol{\beta}} \{RSS(\boldsymbol{\beta}) + (1-\varrho) \sum_{j=1}^{p} |\beta_j| + \varrho \sum_{j=1}^{p} \beta_j^2\}, \tag{2.17}$$

where $\varrho \in [0, 1]$. The penalty term of the elastic net is a convex combination of the ridge and lasso penalties. When $\varrho = 1$, the elastic net becomes the ridge regression, whereas when $\varrho = 0$, it becomes the lasso. And, when $0 < \varrho < 1$, the elastic net performs variable selection and estimation along with the characteristics of both lasso and ridge regression. The elastic net having the two properties is a useful technique, particularly in the $p > n$ case, and grouped variable situation.

2.2.4 Smoothly clipped absolute deviation: SCAD

Fan and Li (2001) introduced three properties of good penalty function for the L_1-type regularized regression,

1. Unbiasedness: The resulting estimator is nearly unbiased when the true unknown parameter is large to avoid modeling bias,
2. Sparsity: Small estimated coefficients become zero to reduce model complexity,
3. Continuity: The resulting estimator is continuous to avoid instability in model prediction,

and they proposed the smoothly clipped absolute deviation (SCAD) satisfying above three properties,

$$\hat{\boldsymbol{\beta}}^{\text{SCAD net}} = \arg\min_{\boldsymbol{\beta}} \{\frac{1}{2} RSS(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|)\}, \tag{2.18}$$

where

$$p_\lambda(|\beta_j|) = \begin{cases} \lambda|\beta_j|, & \text{if } |\beta_j| \leq \lambda, \\ -(\frac{|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)}), & \text{if } \lambda < |\beta_j| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2}, & \text{if } |\beta_j| > a\lambda, \end{cases}$$

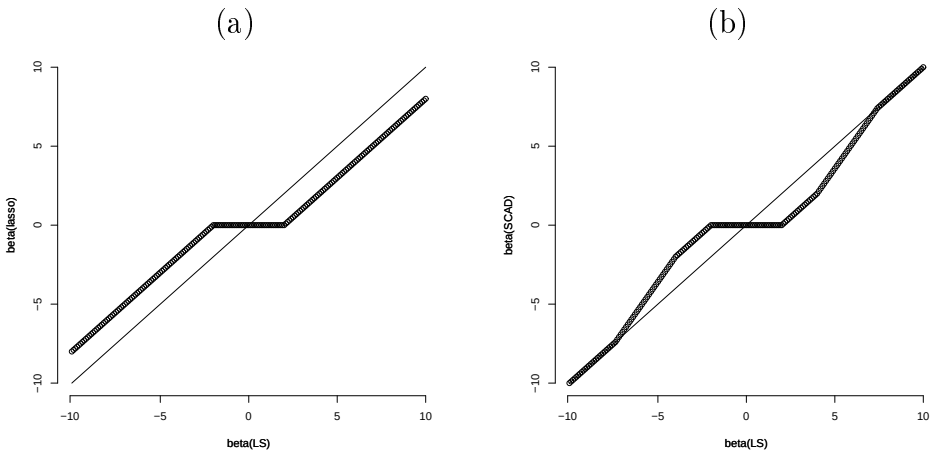


Figure 2.2: Thresholding function with $\lambda = 2$ for (a) the lasso and (b) the SCAD (a=3.7).

where $a > 2$ and $\lambda > 0$. The solution of the SCAD is given by,

$$\hat{\beta}_j^{SCAD} = \begin{cases} (\hat{\beta}_j - \lambda)_+ \text{sign}(\hat{\beta}_j) & \text{if } |\beta_j| \leq 2\lambda, \\ \{(a-1)\hat{\beta}_j - \text{sign}(\hat{\beta})_j \lambda\} & \text{if } 2\lambda < |\beta_j| \leq a\lambda, \\ \hat{\beta}_j & \text{if } |\beta_j| > a\lambda. \end{cases}$$

As shown above, the SCAD has the property "sparsity" like to the lasso for $|\beta_j| \leq 2\lambda$, and can achieve the unbiasedness for large $|\beta_j| > a\lambda$. Figure 2.2 (a) and (b) show the estimator of the lasso and SCAD, respectively. From Figure 2.2, it can be seen that the SCAD produces unbiased estimation results for large $|\beta_j|$ unlike to the lasso.

2.3 Estimation of sparse regression model

The lasso-type approaches provide a useful tool for the sparse regression modeling. Although they have shown a remarkable performance in the regression modeling, their estimates cannot be analytically derived due to indifferentiability of the L_1-type

penalty term. To settle on the issue, several algorithms have been proposed. In this section, we introduce the algorithms to implement the L_1-type penalty, such as local quadratic approximation (Fan and Li, 2001), LARS (Efron et al., 2004) and coordinate descent algorithm (Friedman et al., 2007).

2.3.1 Local quadratic approximation

Fan and Li (2001) presented the following local quadratic approximation of L_1-type penalty for estimation of the sparse regression model.

Let $p_\lambda(|\cdot|)$ be the L_1-type penalty term. Suppose that we give an initial value $\boldsymbol{\beta}_0$ that is close to the minimizer of (2.8). If β_{j0} is very close to 0, then set $\hat{\beta}_j = 0$. Otherwise the penalty term is locally approximated by a quadratic function as follows,

$$[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta_j|)\text{sign}(\beta_j) \approx \{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\}\beta_j, \tag{2.19}$$

where $\beta_j \neq 0$. It means that,

$$p_\lambda(|\beta_j|) = p_\lambda(|\beta_{j0}|) + \frac{1}{2}\{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\}(\beta_j^2 - \beta_{j0}^2), \tag{2.20}$$

for $\beta_j \approx \beta_{j0}$.

In case of the lasso, the L_1 norm penalty term can be locally approximated as follows,

$$\lambda|\beta_j| \approx \lambda|\beta_{j0}| + \frac{\lambda}{2}\{\frac{\beta_j^2}{|\beta_{j0}|} - |\beta_{j0}|\}. \tag{2.21}$$

Figure 2.3 shows the local quadratic approximation of lasso penalty term in (2.21). As shown in Figure 2.3, the approximated lasso penalty term is differentiable, and thus we can settle on the derivation problem. This implies that, we can find a solution

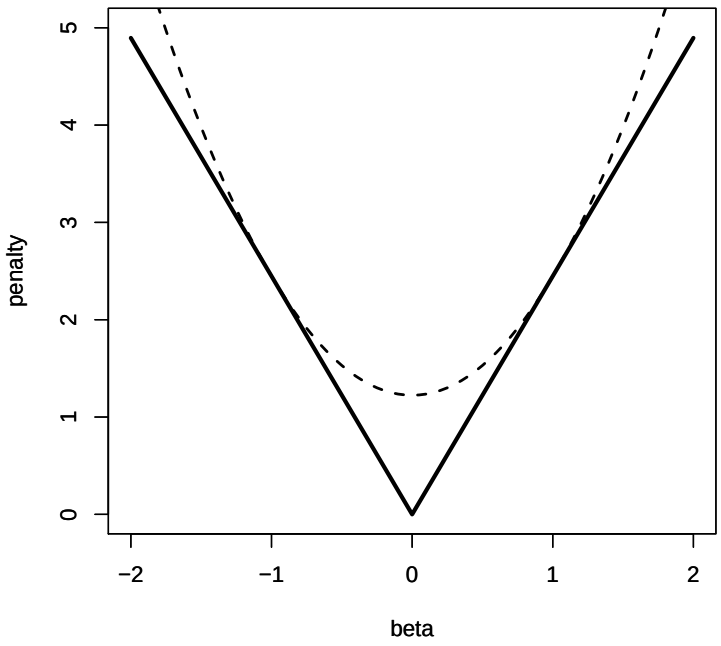


Figure 2.3: Local quadratic approximation of lasso penalty $\lambda|\beta_j|$

of the L_1-type approaches by derivative of (2.8) with approximated penalty in (2.20).

2.3.2 LARS algorithm

The least angle regression (LAR) is a method for variable selection and estimation, similar to the forward stepwise regression. Efron et al. (2004) modified the LAR for the lasso. In this present thesis, we call the procedure to LARS. The computational cost of the LARS for entire p step is same with the usual least squares estimate procedure for full model. Hastie et al. (2007) described the LARS algorithm for the lasso as follows,

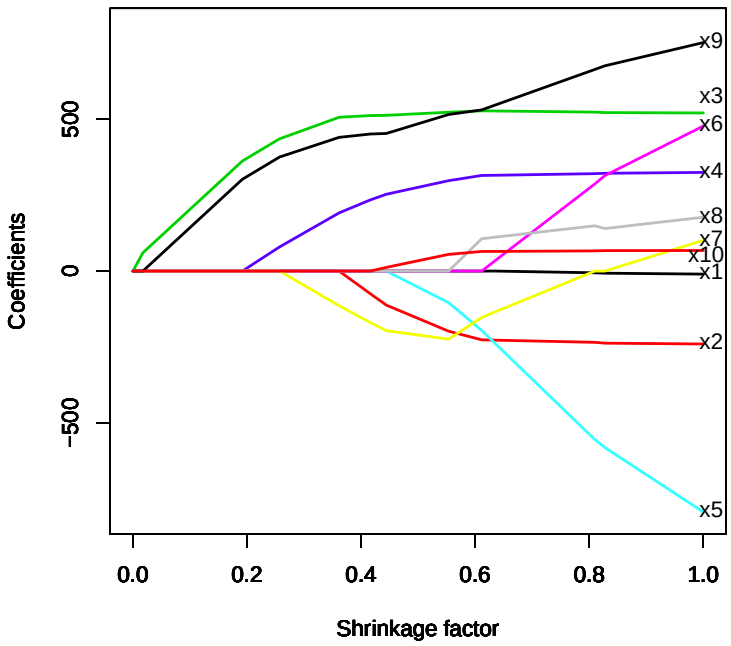


Figure 2.4: LARS procedure for Diabetes dataset

1. Standardize the predictors to mean zero and unit variance. Start with $\mathbf{r} = \mathbf{y} - \bar{y}$, $\beta_1, \beta_2, ..., \beta_p$=0.

2. Find the predictor $\mathbf{x}_j$ the most correlated with $\mathbf{r}$.

3. Move β_j from 0 towards $\hat{\beta}^{LS}$, until some other competitor $\mathbf{x}_k$ has as much correlation with the current residual as does $\mathbf{x}_j$.

4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on $(\mathbf{x}_j, \mathbf{x}_k)$, until some other competitor $\mathbf{x}_l$ has as much correlation with the current residual.

4a. If a non-zero coefficient hits zero, drop its variable from the active set of

variables and recompute the current joint least squares direction.

5. Continue in this way until all p predictors have been entered. After $\min(N-1, p)$ steps, we arrive at the full least squares solution.

Figure 2.4 shows the LARS procedure for diabetes data (Efron et al., 2007) by using *lars* package in R, which well describes the properties of the lasso. As shown in Figure 2.4, the variables are joined into the active set at each step as increasing the shrinkage factor. It implies that the lasso performs variable selection and estimation simultaneously along with a pathwise solution by choosing the shrinkage factor. The small value of shrinkage factor corresponding large value λ shrinks coefficients significantly. This implies that the regularization parameter plays a key role in the L_1-type regularized regression modeling.

2.3.3 Coordinate descent algorithm

Coordinate descent algorithm is very competitive with well known LARS procedure for the lasso problem (Friedman et al., 2007). Furthermore, it can be applied to various L_1-type regularization, such as the elastic net and garrote. The coordinate descent algorithm has an advantage that the coordinate minimization can be done quickly, and hence it is suitable method for regression modeling with a large number of predictor variables.

We introduce the algorithm for the lasso. The lasso in (2.10) can be seen the following problem,

$$f(\beta) = \arg\min_{\boldsymbol{\beta}} \{\frac{1}{2}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\}. \tag{2.22}$$

In order to better understand the procedure, we first introduce the soft-thresholding version of the lasso solution (Donoho and Johnstone, 1995) with single predictor variable. The problem of (2.22) with single standardized x and coefficient β is given by

$$f_s(\beta) = \arg\min_{\beta}\{\frac{1}{2}(\beta - \hat{\beta})^2 + \lambda|\beta|\Big\}, \tag{2.23}$$

where $\hat{\beta} = \sum_i^n x_i y_i$ is the least squares estimator, since the predictor x is standardized. If $\beta > 0$, (2.23) can be differentiated as follows,

$$\frac{\partial f_s(\beta)}{\partial \beta} = \beta - \hat{\beta} + \lambda = 0, \tag{2.24}$$

and if $\beta < 0$ then $\beta = \hat{\beta} + \lambda$, thus the solution is given by

$$\hat{\beta}(\lambda) = S(\hat{\beta}, \lambda) \equiv \text{sign}(\hat{\beta})(|\hat{\beta}| - \lambda)_+ \tag{2.25}$$

$$= \begin{cases} \hat{\beta} - \lambda, & \text{if } \hat{\beta} > 0 \text{ and } \lambda < |\hat{\beta}|, \\ \hat{\beta} + \lambda, & \text{if } \hat{\beta} < 0 \text{ and } \lambda < |\hat{\beta}|, \\ 0, & \text{if } \lambda \geq |\hat{\beta}|. \end{cases}$$

For the p predictor variables, (2.22) can be expressed with partial residual as follows,

$$f(\tilde{\beta}) = \arg\min_{\beta}\{\frac{1}{2}\sum_{i=1}^{n}(y_i - \sum_{k\neq j}^{p} x_{ik}\tilde{\beta}_k - x_{ij}\beta_j)^2 + \lambda\sum_{k\neq j}^{p}|\tilde{\beta}_k| + \lambda|\beta_j|\}, \tag{2.26}$$

where all values of $\tilde{\beta}_k$ for $k \neq j$ are fixed, and thus (2.26) can be considered as the soft thresholding with j^{th} predictor with partial residual. The differentiation of (2.26) by β_j is given,

$$\begin{aligned} \frac{\partial f(\tilde{\beta})}{\partial \beta_j} &= -\sum_{i=1}^{n}(y_i - \sum_{k\neq j}^{p} x_{ik}\tilde{\beta}_k - x_{ij}\beta_j)x_{ij} + \lambda\,\text{sign}(\beta_j) \\ &= -\sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_i^{(j)}) + \sum_{i=1}^{n} x_{ij}^2\beta_j + \lambda\,\text{sign}(\beta_j), \end{aligned} \tag{2.27}$$

where $\tilde{y}_i^{(j)} = \sum_{k \neq j}^{p} x_{ik}\tilde{\beta}_k$, and thus the coordinate update has following form,

$$\tilde{\beta}_j \leftarrow S\Big(\sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_i^{(j)}), \lambda\Big) \tag{2.28}$$

as a minimizing (2.26) with respect to β_j. The (2.28) can be viewed as the univariate regression coefficient of the partial residual $(y_i - \tilde{y}_i)$ on the j^{th} predictor variable, and hence it is updated for $j = 1, 2, ..., p, 1, 2, ...$ until convergence for the regression model with p predictors.

The coordinate update in (2.28) can be expressed by,

$$\tilde{\beta}_j \leftarrow S\Big(\tilde{\beta}_j(\lambda) + \sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_i), \lambda\Big), \qquad j = 1, 2, ..., p, 1, 2, ... \tag{2.29}$$

By compute the simple least squares coefficient on the partial residual $(y_i - \tilde{y}_i)$, we can find the lasso solution. This implies that the coordinate descent algorithm is an effective method for the lasso. Although we focus on the procedure for the lasso, it is a useful tool for various L_1-type regularization methods, such as the elastic net, garrote and ridge regression.

2.4 Selection of regularization parameters

As shown above, an appropriate choice of the regularization parameters is a vital matter in the L_1-type regularization, since variable selection and estimation heavily depend on the adjusted parameters. The large value of the regularization parameter shrinks largely, and thus the coefficients are estimated in small. Furthermore, some coefficients are estimated to exactly zero.

The traditional model evaluation criteria, such as AIC and BIC, however, cannot be directly applied for choosing the regularization parameters, since they were derived

under the assumptions that the model is estimated by the maximum likelihood, and they carried out in a parametric family of distributions including the true model (Konishi and Kitagawa, 2008).

In practice, the most usually used method for choosing the regularization parameters is a cross-validation. In this section, we introduce the cross-validation for regularization parameter selection.

2.4.1 Cross-validation

The regularization parameter was often selected by the cross-validation, which is the simplest and useful method for modeling based on the predictive point of view. The cross-validation procedure is conducted by estimating a predictive mean squares error (PSE) from the separated dataset as a training data for model estimation and test data for model evaluation.

- **K-fold cross-validation**:

The K-fold cross-validation is executed by the following step,

1. Data set is randomly divided into K-parts.
2. Remove the k^{th} part of data.
3. The model is estimated based on the remaining $K-1$ parts of data: $\hat{f}^{(-k)}(\lambda, \mathbf{x})$.
4. For the k^{th} part removed in step 3, the PSE is calculated: $\{y_k - \hat{f}^{(-k)}(\lambda, \mathbf{x}_k)\}^2$

5. Do step 2 to 4 for $k = 1, 2, ..., K$, and calculate

$$\mathrm{CV}(\lambda) = \frac{1}{K}\sum_{k=1}^{K}\{y_k - \hat{f}^{(-k)}(\lambda, \mathbf{x}_k)\}^2 \tag{2.30}$$

as an estimate of the PSE.

In the L_1-type regularized regression, we select the regularization parameter λ that minimizes the $\mathrm{CV}(\lambda)$.

The choice of K, number of data partition, is also crucial in practice. The case of $K = n$ (i.e., n-fold cross-validation) is known as *leave-one-out* cross-validation. Although the *leave-one-out* cross-validation can perform stable selection of tuning parameters, it is time consuming. Generally, the 10-fold cross-validation is widely used in various fields.

- **Generalized cross-validation**:

In large data set, the K-fold cross-validation has a computational difficulty. To overcome the drawback, a generalized cross-validation was proposed (Craven and Wahba, 1979). The generalized cross-validation (GCV) focuses on the predicted value $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = H\mathbf{y}$, where the hat matrix H does not depends on the data $\mathbf{y}$. This implies that K times repetition by removing observations in k^{th} part does not required, and thus amount of the computation is significantly reduced.

Lemma 2.4.1 *The generalized cross-validation based on the hat matrix H is given,*

$$GCV = \frac{1}{n}\frac{\sum_{\alpha=1}^{n}\{y_\alpha - \hat{f}(\lambda, \mathbf{x}_\alpha)\}^2}{\{1 - \frac{1}{n}\mathrm{trH}\}^2}. \tag{2.31}$$

Proof. Let consider the *leave-one-out* cross-validation. Removing the α^{th} data point $(y_\alpha, \mathbf{x}_\alpha)$ and estimate the regression function $\hat{f}^{(-\alpha)}(\lambda, \mathbf{x}) = \mathbf{x}^T \hat{\beta}^{(-\alpha)}$ by the lasso. And then, we set $z_i = y_i$, and replace the α^{th} data point y_α with $\hat{f}^{(-\alpha)}(\lambda, \mathbf{x}_\alpha)$,

$$\mathbf{z} = (y_1, y_2, ..., \hat{f}^{(-\alpha)}(\lambda, \mathbf{x}_\alpha), ..., y_n)^T. \tag{2.32}$$

The regression function $\hat{f}^{(-\alpha)}(\lambda, \mathbf{x})$ is estimated without the α^{th} data point by minimize,

$$\begin{aligned}
&\sum_{i=1}^{n} \{z_i - \mathbf{x}_i^T \boldsymbol{\beta}\}^2 + \lambda \sum_{j=1}^{p} |\beta_j| \qquad (2.33)\\
&\geq \sum_{i \neq \alpha}^{n} \{z_i - \mathbf{x}_i^T \boldsymbol{\beta}\}^2 + \lambda \sum_{j=1}^{p} |\beta_j| \\
&\geq \sum_{i \neq \alpha}^{n} \{z_i - \hat{f}^{(-\alpha)}(\lambda, \mathbf{x}_i)\}^2 + \lambda \sum_{j=1}^{p} |\hat{\beta}_j^{(-\alpha)}| \\
&= \sum_{i=1}^{n} \{z_i - \hat{f}^{(-\alpha)}(\lambda, \mathbf{x}_i)\}^2 + \lambda \sum_{j=1}^{p} |\hat{\beta}_j^{(-\alpha)}|.
\end{aligned}$$

Note that $z_\alpha - \hat{f}^{(-\alpha)}(\lambda, \mathbf{x}_\alpha) = 0$, and thus $\hat{f}^{(-\alpha)}(\lambda, \mathbf{x})$ can be seen as a minimizer of the first line in (2.33).

Let $h_{\alpha,i}$ is $(\alpha, i)^{th}$ component of the hat matrix H, then

$$\begin{aligned}
\hat{f}^{(-\alpha)}(\lambda, \mathbf{x}_\alpha) - y_\alpha &= \sum_{i=1}^{n} h_{\alpha,i} z_i - y_\alpha \qquad (2.34)\\
&= \sum_{i \neq \alpha}^{n} h_{\alpha,i} y_i + h_{\alpha\alpha} \hat{f}^{(-\alpha)}(\lambda, \mathbf{x}_\alpha) - y_\alpha \\
&= \sum_{i=\alpha}^{n} h_{\alpha,i} y_i - y_\alpha + h_{\alpha\alpha} \{\hat{f}^{(-\alpha)}(\lambda, \mathbf{x}_\alpha) - y_\alpha\} \\
&= \hat{f}(\lambda, \mathbf{x}_\alpha) - y_\alpha + h_{\alpha\alpha} \{\hat{f}^{(-\alpha)}(\lambda, \mathbf{x}_\alpha) - y_\alpha\},
\end{aligned}$$

and thus,

$$y_\alpha - \hat{f}^{(-\alpha)}(\lambda, \mathbf{x}_\alpha) = \frac{y_\alpha - \hat{f}(\lambda, \mathbf{x}_\alpha)}{1 - h_{\alpha\alpha}}. \tag{2.35}$$

By substituting (2.35) into the *leave-one-out* cross-validation version of (2.30), the CV(λ) is given by

$$\mathrm{CV}(\lambda) = \frac{1}{n}\sum_{\alpha=1}^{n}\{\frac{y_\alpha - \hat{f}(\lambda, \mathbf{x}_\alpha)}{1 - h_{\alpha\alpha}}\}^2. \tag{2.36}$$

The generalized cross-validation (2.31) can be obtained by replacing $1 - h_{\alpha\alpha}$ in (2.36) with its average $1 - \frac{1}{n}\mathrm{tr}H$.

In practically, it is difficult to calculate the hat matrix H of lasso-type regularization, since the L_1-type penalty functions are not differentiable in origin. Tibshirani (1997) proposed an approximate generalized cross-validation by using the local quadratic approximation of lasso-type penalty (Fan and Li, 2001).

Chapter 3

Robust regression modeling via L_1 regularization

We discuss a robust regression modeling via the L_1 regularization in various aspects. Although the L_1-type approaches showed the exceptional performances in regression modeling, existing studies on

- Methodology
- Model estimation
- Model evaluation

for L_1-type regularization were conducted under the assumption of absence of outliers. Recent studies (Khan et al., 2007; Zhang et al., 2009; Lambert-Lacroix and Zwald, 2010) exposed that their performances take a sudden turn for the worst in the presence of outliers, since the procedures are based on non-robust methodologies. We introduce robust modeling strategies via the L_1-type regularization in the aspects of

methodology, model estimation and evaluation,

- Robust L_1-type regularization,
- Robust algorithm for estimation of the L_1-type regularized regression model,
- Robust model evaluation for tuning parameter selection.

First, we introduce the robust L_1-type regularization methods. Although the L_1-type regularization provides an efficient tool for regression modeling, its performance suffers from outliers, since it is constructed by penalized least squares method. To overcome the drawback, several robust L_1-type regularization methods were proposed by replace the least squares loss function with robust loss function. We briefly introduce the existing robust methodologies and propose a least trimmed squares elastic net having a high breakdown point. In the robust sparse regression modeling, an appropriate choice of the regularization parameters and tuning constant in outlier detection is a crucial issue, because the modeling procedure rely on the adjusted tuning parameters. We propose novel methods for choosing the tuning parameters in line with the information-theoretic approach.

Then, we discuss a robust estimation for L_1-type regularized regression model via an outlier-resistant algorithm. The existing algorithms for L_1-type regularization as shown in Section 2.3 cannot perform well in the presence of outliers, since the procedures are based on sample mean, standard deviation and correlation or inner product obtained from a non-robust manner. To settle on the problem, Khan et al. (2007) proposed a robust LARS procedure based on the robust correlation. We consider the robust regression modeling via the coordinate descent algorithm, which is competitive with the LARS, by winzorization and trimming techniques.

Finally, we discuss a robust model evaluation for choosing the tuning parameters of the robust lasso-type approaches. In the robust L_1-type approaches, the robust selection of the tuning parameters is a vital matter, since robust modeling procedure heavily depends on appropriately selected tuning parameters. We propose a robust efficient bootstrap information criterion via the Winsorization technique for choosing an optimal set of regularization parameters and tuning constant.

We show through Monte Carlo simulations and real-world examples the effectiveness of the proposed robust strategies.

The rest of this chapter is organized as follows. In Section 3.1, we review the robust lasso-type approaches, and then propose a new robust L_1-type regularization, called a least trimmed squares elastic net. We present methods for choosing an optimal set of the regularization parameters and tuning constant in Section 3.2. We discuss the outlier-resistant algorithm for the sparse regression modeling in Section 3.3. In Section 3.4, we discuss the robust model evaluation for choosing the tuning parameters. Monte Carlo simulations are conducted to investigate the efficiency of the proposed methods in Section 3.5. The real world examples are shown in Section 3.6.

3.1 Robust L_1-type regularization

The L_1-type regularization as shown in Section 2.2 is a useful tool for regression modeling. Although the lasso-type approaches have shown the exceptional performance in various fields of research, their performance takes a sudden turn for the worst in the presence of outliers, since they are constructed by least squares loss function

and L_1-type penalty. To overcome the problem, several studies were conducted for outlier-resistant lasso-type approaches by replace the least squares loss function with the robust loss function.

3.1.1 Literature review: existing robust L_1-type regularization

We briefly introduce the existing studies on the robust lasso-type approaches. To overcome non-robustness of lasso-type estimator, the least squares loss function is replaced with other loss function (e.g., least squares trimmed loss function, least absolute loss function, M-estimation function, etc.) as follows:

- Least absolute lasso (Wang et al., 2007a):

$$\hat{\boldsymbol{\beta}}^{\text{LAD-lasso}} = \arg\min_{\boldsymbol{\beta}} \{\sum_{i=1}^{n} |y_i - \mathbf{x}_i^T \boldsymbol{\beta}| + \lambda_j \sum_{j=1}^{p} |\beta_j|\}, \tag{3.1}$$

  where λ_j is a regularization parameter for β_j, and thus the least absolute lasso allows for a different regularization parameter for each coefficient.

- Least trimmed squares lasso (Mateos and Giannakis, 2010):

$$\hat{\boldsymbol{\beta}}^{\text{LTS-lasso}} = \arg\min_{\boldsymbol{\beta}} \{\sum_{i=1}^{s} r_{[i]}^2 + \lambda \sum_{j=1}^{p} |\beta_j|\}, \tag{3.2}$$

  where s is a tuning constant, $r_{[i]}^2$ is the i^{th} order statistic of squared residuals and $r_i = y_i - \mathbf{x}_i^T \boldsymbol{\beta}$.

- M-lasso (Zhang et al., 2009):

$$\hat{\boldsymbol{\beta}}^{\text{M-lasso}} = \arg\min_{\boldsymbol{\beta}} \{\sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|\}, \tag{3.3}$$

where $\rho(\cdot)$ is M-estimation function,

-Huber M-function,

$$\begin{aligned} \rho(\cdot) &= r^2/2, && \text{if } |r| < k \\ &= k(|r| - k/2), && \text{if } |r| \geq k. \end{aligned} \tag{3.4}$$

-Tukey function :

$$\begin{aligned} \rho(\cdot) &= (k^2/6)(1 - [1 - (r/k)^2]^3), && \text{if } |r| < k \\ &= k^2/6, && \text{if } |r| \geq k. \end{aligned} \tag{3.5}$$

where k is a tuning constant.

- M-adaptive lasso (Lambert-Lacroix and Zwald, 2010):

$$\hat{\boldsymbol{\beta}}^{\text{M-adlasso}} = \arg\min_{\boldsymbol{\beta}} \{\sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \hat{w}_j |\beta_j|\}, \tag{3.6}$$

where $\hat{\mathbf{w}} = 1/|\hat{\boldsymbol{\beta}}|^{\gamma}$, $\gamma > 0$ and the least squares estimator or the ridge estimator can be used as $\hat{\boldsymbol{\beta}}$.

The robust lasso-type regularization methods are composed of the robust loss functions and the penalty term of the lasso or adaptive lasso. By replacing the least squares loss function, the robust lasso-type regularization methods are able to perform variable selection and estimation effectively, even in the presence of outliers.

3.1.2 Least trimmed squares elastic net

Limitation : We considered that the M-estimator and LAD estimator do not have a high breakdown point, and hence the existing robust L_1-type regularization con-

sisting of M-function and LAD-loss function also do not have a high breakdown point.

Here, a breakdown point is a main aim of the robust statistics and measures the smallest percentage of outliers in data which can effect to the estimation procedure as following (Rousseeuw and Leroy, 1987),

Definition 1 : Breakdown point.

Let sample of n data points,

$$\mathbf{Z} = \{(x_{11}, ..., x_{1p}, y_1), ..., (x_{n1}, ..., x_{np}, y_n)\}, \tag{3.7}$$

and $\mathbf{T}(\mathbf{Z})$ be a regression estimator. By apply $\mathbf{T}$ to such a sample $\mathbf{Z}$, we get a vector of regression coefficient,

$$\mathbf{T}(\mathbf{Z}) = \hat{\boldsymbol{\beta}}. \tag{3.8}$$

Consider contaminated data $\mathbf{Z}'$ obtained by replacing any m of the original data points. Let us denote by bias$(m; \mathbf{T}, \mathbf{Z})$, the maximum bias that can be caused by such a contamination:

$$bias(m; \mathbf{T}, \mathbf{Z}) = sup_{\mathbf{Z}'} \|\mathbf{T}(\mathbf{Z}) - \mathbf{T}(\mathbf{Z}')\|. \tag{3.9}$$

If bias$(m; \mathbf{T}, \mathbf{Z})$ is infinite, it means that m outliers have considerable effect on $\mathbf{T}$, which is expressed that "the estimator breaks down". The breakdown point of the estimator $\mathbf{T}$ at the sample $\mathbf{Z}$ is defined,

$$\varepsilon_n^*(\mathbf{T}, \mathbf{Z}) = min\{\frac{m}{n}; bias(m; \mathbf{T}, \mathbf{Z}) \text{ is infinite}\}. \tag{3.10}$$

It can be also expressed by "masking effect" of outliers.

Remark 3.1.1 *The breakdown point of the least absolute deviation (LAD) and M estimators is* 0%. *It can be intuitively understood from the loss function of LAD in (3.1) and M-function in (3.4) and (3.5).*

For the robust lasso-type approaches having a high breakdown point, we consider the following least trimmed squares (LTS) estimator (Rousseeuw and Leroy 1987) having the maximum breakdown point $\{[(n-p)/2]+1\}/n$, which is asymptotically equal 50%, for $s = [(n+p+1)/2]$ (see Rousseeuw and Leroy (1987), Theorem 6),

$$\hat{\boldsymbol{\beta}}^{\text{LTS}} = \arg\min_{\boldsymbol{\beta}} \{\sum_{i=1}^{s} r_{[i]}^2\}, \tag{3.11}$$

where $r_{[i]}^2$ is the i^{th} order statistic of squared residuals, $r_i = y_i - \mathbf{x}_i^T\boldsymbol{\beta}$. It is similar to least squares estimator but only difference being that observations having the large squared residuals $r_{[s+1]}^2$,...,$r_{[n]}^2$ are not used (Rousseeuw and Leroy 1987) for estimation procedure, and thus we can reduce the effect of outliers in regression modeling.

In the LTS procedure, the sample size used to estimate is decreased from n to s. In other words, there is a possibility that sample size s, is smaller than the number of predictor variables p. Hence, although the LTS has a high breakdown point, it is unsuitable for using with the lasso because of the limitation of lasso as a variable selection method in $p > n$ situation (see Section 2.2.3). As mentioned in Section 2.2.3, the elastic net was proposed to settle the problem of the lasso in $p > n$ situation. Therefore, we propose a robust elastic net, called a least trimmed square-elastic net (LTS-Ela),

$$\hat{\boldsymbol{\beta}}^{\text{LTS-Ela}} = \arg\min_{\boldsymbol{\beta}} \{\sum_{i=1}^{s} r_{[i]}^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2\}, \tag{3.12}$$

where s is a tuning constant. The proposed LTS-Ela is composed of the LTS loss function and the elastic net penalty term. The LTS-Ela is similar to the original elastic net but with a key difference for the robustness. Unlike the original elastic net, the LTS-Ela is based on only s observations having small residuals to reduce an effect of outliers, and thus outliers having large residual are not used for regression modeling procedure. This implies that choosing the not only regularization parameters but also tuning constant s is a crucial matter, since the selected s plays a key role for robustness in the LTS-Ela. We will present methods for choosing the optimal set of these tuning parameters in following Section 3.2.

The proposed LTS-Ela shows the outstanding performance for the robust sparse regression modeling in the viewpoint of forecasting accuracy and sparsity (see Part 1 in Section 3.5. Simulation studies).

3.2 Selection of tuning parameters in robust L_1-type regularization

The robust lasso-type approach is an effective tool for regression modeling in the presence of outliers. By replacing the least squares loss function with robust loss function, the robust L_1-type regularization can perform simultaneous parameter estimation and variable selection robustly. Crucial issues in the robust sparse modeling include the selection of regularization parameters and also a tuning constant in outlier detection, because the features of the modeling procedure rely on the proper choice of the adjusted parameters. However, relatively little attention was paid for

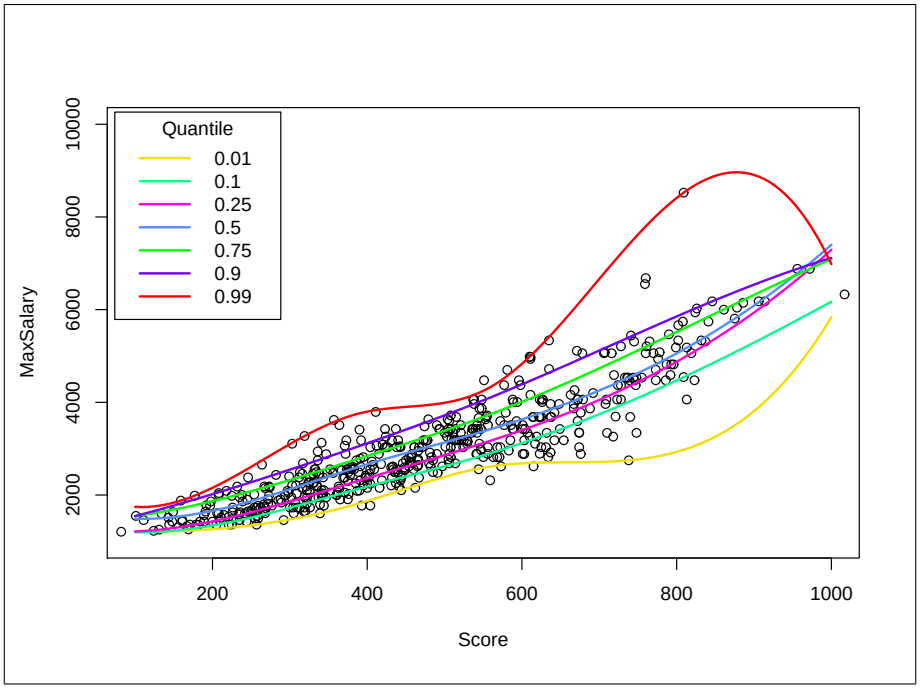


Figure 3.1: Quantile regression with various quantiles

this issue, particularly in the presence of outliers. In fact, existing studies on the robust L_1-type regularization, such as the M-lasso and M-adaptive lasso (Zhang et al., 2009; Lambert-Lacroix and Zwald, 2010), did not consider the selection of the tuning constant. They selected only the regularization parameters controlling the model complexity by the cross-validation under the fixed tuning constant $k = 1.34$, which is a value for general M-estimation without lasso-type penalty. However, we should consider the issue as an appropriate choice of set of regularization parameters and tuning constant, since the tuning constant also plays a key role in the robust sparse regression modeling. Figure 3.1 shows the quantile regression for salary data (Weisberg, 2005), which is one of the robust regression modeling, with various quantiles. In the quantile regression, quantile can be seen as a tuning constant for controlling the effect of outliers. As shown in Figure 3.1, the regression fitting line is

significant changed as increasing the quantile. This implies that choosing the tuning constant is crucial in the robust regression modeling. As we mentioned in Section 2.4, the traditional criteria (i.e., AIC and BIC) are not suitable for choosing the tuning parameters of L_1-type regularization because of their assumption. Furthermore, the usually used cross-validation for regularization parameter selection also has some demerits on instability and over-fitting effect (Wang et al., 2007b).

In this section, we discuss about an appropriate choice of the regularization parameters and tuning constant in line with the information-theoretic viewpoint. We first introduce to use the efficient bootstrap information criteria for choosing the tuning parameters (Konishi and Kitagawa, 1996; Park et al., 2012a). We also present a model selection criterion in line with the generalized information criterion (Konishi and Kitagawa, 1996; Park et al., 2012b).

3.2.1 Efficient bootstrap information criterion

We introduce the efficient bootstrap information criteria for choosing the optimal set of the tuning parameters of the robust lasso-type approaches.

Consider the case in which a model is given in the form of a probability distribution $\{f(y|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta \subset R^p\}$ having p-dimensional parameters. We assume that the data $\mathbf{y}_n = \{y_1, ..., y_n\}$ are generated from the true distribution function $G(y)$. Our task is to evaluate the expected goodness or badness of the estimated model $f(z|\hat{\boldsymbol{\theta}})$ when it is used to predict the independent future data $Z = z$ generated from the unknown true

distribution. The general form of an information criterion is constructed as follows:

$$\begin{aligned} IC(\mathbf{y}_n; \hat{G}) &= -2(\text{log likelihood of statistical model} - \text{bias estimator}) \\ &= -2\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) + 2\{\text{estimator for } b(G)\}, \end{aligned} \tag{3.13}$$

where $b(G)$ is a bias of the log-likelihood as an estimator of the expected log-likelihood depending on the unknown probability distribution G. That is, the bias $b(G)$ is given by

$$b(G) = E_{G(\mathbf{y}_n)}\left[\log f(\mathbf{y}_n|\hat{\boldsymbol{\theta}}(\mathbf{y}_n)) - nE_{G(z)}\left[\log f(Z|\hat{\boldsymbol{\theta}}(\mathbf{y}_n))\right]\right], \tag{3.14}$$

where $\log f(\mathbf{y}_n|\hat{\boldsymbol{\theta}}(\mathbf{y}_n)) = \sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}(\mathbf{y}_n))$ and the expectation $E_{G(\mathbf{y}_n)}$ is taken with respect to the joint distribution, $\prod_{i=1}^{n} G(y_i) = G(\mathbf{y}_n)$ of the sample $\mathbf{y}_n$ (Konishi and Kitagawa, 2008). In general, the bias $b(G)$ can take various forms depending on the method employed to construct a statistical model.

Konishi and Kitagawa (1996) showed that the difference between the log-likelihood of the model and n times the expected log-likelihood

$$D(\mathbf{y}_n; G) = \log f(\mathbf{y}_n|\hat{\boldsymbol{\theta}}) - n\int \log f(z|\hat{\boldsymbol{\theta}}) dG(z), \tag{3.15}$$

can be decomposed into three terms

$$D(\mathbf{y}_n; G) = D_1(\mathbf{y}_n; G) + D_2(\mathbf{y}_n; G) + D_3(\mathbf{y}_n; G), \tag{3.16}$$

where

$$\begin{aligned} D_1(\mathbf{y}_n; G) &= \log f(\mathbf{y}_n|\hat{\boldsymbol{\theta}}) - \log f(\mathbf{y}_n|\boldsymbol{\theta}), \\ D_2(\mathbf{y}_n; G) &= \log f(\mathbf{y}_n|\boldsymbol{\theta}) - n\int \log f(z|\boldsymbol{\theta}) dG(z), \\ D_3(\mathbf{y}_n; G) &= n\int \log f(z|\boldsymbol{\theta}) dG(z) - n\int \log f(z|\hat{\boldsymbol{\theta}}) dG(z). \end{aligned} \tag{3.17}$$

By taking the expectation term by term on (3.16), the second term is

$$\begin{aligned} E_G[D_2(\mathbf{y}_n; G)] &= E_G\left[\log f(\mathbf{y}_n|\boldsymbol{\theta}) - n\int \log f(z|\boldsymbol{\theta}) dG(z)\right] \\ &= \sum_{i=1}^{n} E_G\left[\log f(y_i|\boldsymbol{\theta}) - nE_G[\log f(Z|\boldsymbol{\theta})]\right. \\ &= 0. \end{aligned} \tag{3.18}$$

Thus, the expectation of bias correction term in (3.15) can be expressed without $D_2(\mathbf{y}_n; G)$ term as follows,

$$E_G[D(\mathbf{y}_n; G)] = E_G[D_1(\mathbf{y}_n; G) + D_3(\mathbf{y}_n; G)]. \tag{3.19}$$

In the bootstrap information criteria, the true distribution $G(y)$ is replaced with an empirical distribution function $\hat{G}(y)$. With this replacement, the random variable and estimator in (3.14) are substituted as follows:

$$\begin{gathered} G(y) \longrightarrow \hat{G}(y), \\ y_i \sim G(y) \longrightarrow y_i^* \sim \hat{G}(y), \\ Z \sim G(z) \longrightarrow Z^* \sim \hat{G}(z), \\ E_{G(\mathbf{y})}, E_{G(z)} \longrightarrow E_{\hat{G}(\mathbf{y}^*)}, E_{\hat{G}(z^*)}, \\ \hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y}) \longrightarrow \hat{\boldsymbol{\theta}}^* = \hat{\boldsymbol{\theta}}(\mathbf{y}^*). \end{gathered}$$

Therefore, the bootstrap bias estimate of (3.14) is given by

$$b^*(\hat{G}) = E_{\hat{G}(\mathbf{y}^*)}\left[\sum_{i=1}^{n} \log f(y_i^*|\hat{\boldsymbol{\theta}}(\mathbf{y}_n^*)) - nE_{\hat{G}(z^*)}\left[\log f(Z^*|\hat{\boldsymbol{\theta}}(\mathbf{y}_n^*))\right]\right]. \tag{3.20}$$

Let us B sets of bootstrap samples of size n and write the b^{th} bootstrap sample as $\mathbf{y}_n^*(b) = \{y_1^*(b), ..., y_n^*(b)\}$. In the bootstrap estimate, (3.19) is replaced by

$$E_{\hat{G}}[D(\mathbf{y}_n^*; \hat{G})] = E_{\hat{G}}[D_1(\mathbf{y}_n^*; \hat{G}) + D_3(\mathbf{y}_n^*; \hat{G})]. \tag{3.21}$$

Therefore, we can use

$$b_B(\hat{G}) = \frac{1}{B}\sum_{b=1}^{B}\{D_1(\mathbf{y}_n^*(b);\hat{G}) + D_3(\mathbf{y}_n^*(b);\hat{G})\} \tag{3.22}$$

as a bootstrap bias estimate.

Conditional on the observed data, Konishi and Kitagawa (1996) showed that the orders of asymptotic conditional variances of two bootstrap estimates are

$$\begin{aligned} &\mathrm{Var}\left[\frac{1}{B}\sum_{b=1}^{B}\{D(\mathbf{y}_n^*;\hat{G})\}\right] = \frac{1}{B}O(n), \\ &\mathrm{Var}\left[\frac{1}{B}\sum_{b=1}^{B}\{D_1(\mathbf{y}_n^*;\hat{G}) + D_3(\mathbf{y}_n^*;\hat{G})\}\right] = \frac{1}{B}O(1). \end{aligned} \tag{3.23}$$

Figure 3.2 shows the box plots of the bootstrap estimates of D, $D_1 + D_3$, D_1, D_2, and D_3 for n=25, 100, 400, and 1600. From Figure 3.2, we can see that D and $D_1 + D_3$ fluctuate in a different manner because of the spreading of the distribution D_2. As shown in Figure 3.2, for the small n, such as n=25, the fluctuations of D_1 and D_3 are slightly large compared with of D_2. On the other hand, when n increases, the fluctuation of D_2 becomes dominant and that of $D_1 + D_3$ becomes significantly smaller than that of D. It implies that the variance due to the bootstrap resampling can be reduced significantly, and thus we can expect to efficient modeling.

Consequently, the efficient bootstrap information criterion based on variance reduction method is defined as follows

$$\begin{aligned} \mathrm{EIC}_{\mathrm{eff}} &= -2\sum_{i=1}^{n}\log f(y_i|\hat{\boldsymbol{\theta}}) + 2\{b_B(\hat{G})\} \\ &= -2\sum_{i=1}^{n}\log f(y_i|\hat{\boldsymbol{\theta}}) + \frac{2}{B}\sum_{b=1}^{B}\{D_1(\mathbf{y}_n^*(b);\hat{G}) + D_3(\mathbf{y}_n^*(b);\hat{G})\}. \end{aligned} \tag{3.24}$$

For details on the theoretical justification for sample variance reduction technique, see Konishi and Kitagawa (1996; 2008).

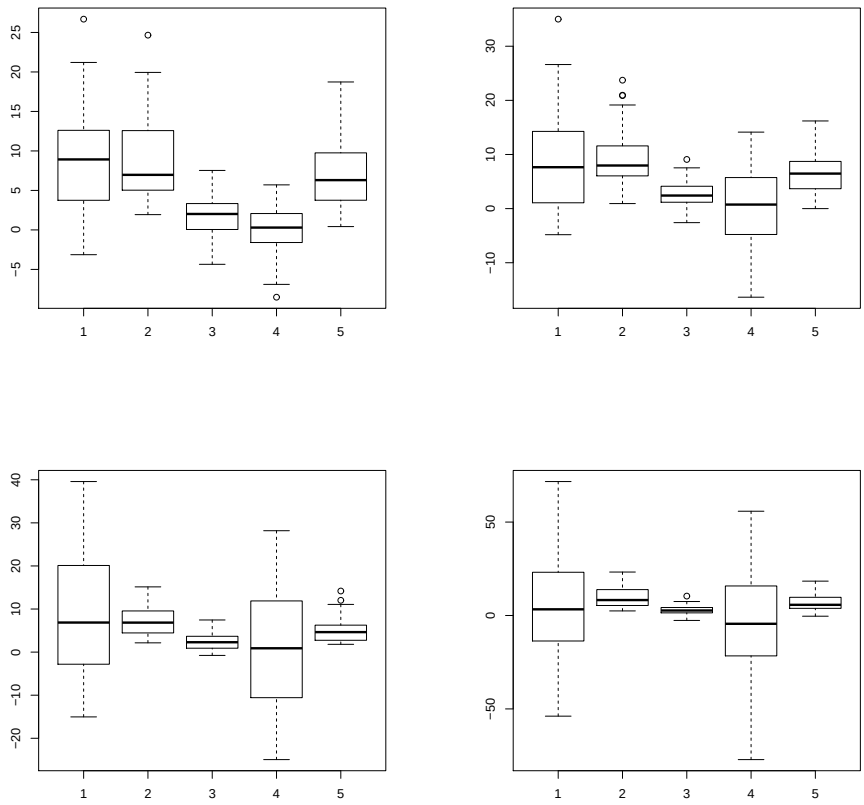


Figure 3.2: Box plots of the bootstrap estimates of D, $D_1 + D_3$, D_1, D_2, and D_3 for n=25 (top left), 100 (top right), 400 (bottom left), and 1600 (bottom right).

We choose an optimal set of the tuning parameters in the robust lasso-type approaches based on the $\mathrm{EIC}_{\mathrm{eff}}$. Under the assumption that ε_i in (2.1) are the random errors from $N(0, \sigma^2)$, the linear regression model is expressed as follows,

$$f(y_i|\mathbf{x}_i;\boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\{y_i - \mathbf{x}_i^T\boldsymbol{\beta}\}^2}{2\sigma^2}\right]. \tag{3.25}$$

To calculate the $\mathrm{EIC}_{\mathrm{eff}}$ for the linear regression model, we generate bootstrap samples denoted as $\mathbf{y}_n^* = \{y_1^*, ..., y_n^*\}$ using a x-fixing method. In the x-fixing method, we consider predictor variables $\mathbf{x}_n$ not random variables and $\mathbf{y}_n^* = \mathbf{x}_n^T\hat{\boldsymbol{\beta}} + \mathbf{e}_n^*$, where

$\mathbf{e}_n^*$ are generated from $\mathbf{e}_n(= \mathbf{y}_n - \mathbf{x}_n^T\hat{\boldsymbol{\beta}})$. And then, we calculate the EIC_{eff} based on the estimated $\hat{\boldsymbol{\beta}}$ by the robust lasso-type approaches at the each set of tuning parameters, and then we perform model selection and estimation by choosing the optimal set of tuning parameters that minimize the EIC_{eff}. As shown above, by using the EIC_{eff}, we can effectively perform the robust sparse regression modeling, since the replications number of bootstrap can be reduced due to the variance reduction method. This implies that the EIC_{eff} is suitable for regularization method with many tuning parameters and large number of predictors. The EIC_{eff} showed the superiority for tuning parameters selection of the proposed LTS-Ela in Section 3.1 (see Part 1 in Section 3.5. Simulation studies).

3.2.2 Generalized information criterion

We derive an information criterion for choosing an optimal set of the regularization parameters and a tuning constant in line with the generalized information criteria (GIC), which can be applied to evaluate statistical models constructed by various types of estimation procedure (Konish and Kitagawa, 1996; Park et al., 2012b). In deriving an information criterion, the calculation of an influence function is essential. However, we cannot derive the influence function corresponding to the L_1-type regularization methods, because the L_1-type penalty functions are not differentiable in origin. To settle on the issue, we use the local quadratic approximation of the lasso-type penalty functions (Fan and Li, 1996), and then derive an information criterion for evaluating robust sparse regression models. We first introduce the generalized information criteria (Konish and Kitagawa, 1996).

Generalized information criteria

Akaike's (Akaike, 1973) information criterion (AIC) was often used for model selection and evaluation in various fields. With the development of the modeling techniques, it has been necessary that an information criterion for models estimated by various techniques, not only maximum likelihood method. Konishi and Kitagawa (1996) proposed generalized information criteria (GIC) by relaxing the following assumptions imposed on the AIC,

- Estimation is by maximum likelihood.
- It carried out in a parametric family of distributions including the true model.

For the GIC, which can evaluate to various modeling techniques, Konishi and Kitagawa (1996) employed a functional estimator, $\hat{\boldsymbol{\theta}} = \mathbf{T}(\hat{G})$, which is second order compact differentiable at G. The p-dimensional functional estimator can be expressed as

$$\hat{\boldsymbol{\theta}} = \boldsymbol{T}(\hat{G}) = (T_1(\hat{G}), ..., T_p(\hat{G}))^T. \tag{3.26}$$

Given a functional $T_j(G)$ (j=1,...,p), the influence function, which plays an essential role in the information criteria, is defined by

$$T_j^{(1)}(y; G) = \lim_{\epsilon \to 0} \frac{T_j((1-\epsilon)G + \epsilon\delta_y) - T_j(G)}{\epsilon}, \tag{3.27}$$

where δ_y is a distribution function having a probability of 1 at point y. They define the p-dimensional vector of influence function having $T_j^{(1)}(y; G)$ as the j^{th} element as follow

$$\boldsymbol{T}^{(1)}(y; G) = (T_1^{(1)}(y; G), ..., T_p^{(1)}(y; G))^T. \tag{3.28}$$

The bias of the log-likelihood for the model $f(y|\hat{\boldsymbol{\theta}})$ in estimating the expected log-likelihood is given by

$$b(G) = E_G[\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) - n\int \log f(z|\hat{\boldsymbol{\theta}})dG(y)] \tag{3.29}$$
$$= \mathrm{tr}\{\int \boldsymbol{T}^{(1)}(y;G)\frac{\partial \log f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(y)\} + O(n^{-1}).$$

The asymptotic bias, $b(G)$, can be estimated by replacing the unknown distribution G with an empirical distribution $\hat{G}$. Thus, the generalized information criteria for evaluating the statistical model $f(y|\hat{\boldsymbol{\theta}})$ with a functional estimator, $\hat{\boldsymbol{\theta}} = \boldsymbol{T}(\hat{G})$, is given by

$$GIC = -2\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) + \frac{2}{n}\sum_{i=1}^{n} \mathrm{tr}\{\boldsymbol{T}^{(1)}(y_i;\hat{G})\frac{\partial \log f(y_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\mathbf{T}(\hat{G})}\}. \tag{3.30}$$

Proposed criterion for robust sparse regression modeling

We present an information criterion for evaluating the robust sparse regression models in line with the GIC. To derive an information criterion, the second order differentiable functional estimator, $\hat{\boldsymbol{\theta}} = \boldsymbol{T}(\hat{G})$, is required. In the case of the lasso-type approaches, however, the functional estimator is not differentiable because of the L_1-type penalty function. To settle on the problem, we use the local quadratic approximation as shown in Section 2.3.1. We calculate the influence function corresponding the lasso-type estimator based on the approximated L_1-type penalty function, and then derive an information criterion for choosing the regularization parameters and a tuning constant.

For the linear regression model, we use robust lasso-type estimates (e.g., with Huber function in (3.4)) of the regression coefficients $\boldsymbol{\beta}$ given as the following solution.

If $\mathbf{T}_{R.la,0}(G)$ is very close to 0, then set $\mathbf{T}_{R.la}(G) = 0$. Otherwise it is assumed that the functional $\mathbf{T}_{R.la}(G)$ is given as a solution of the implicit equation

$$\int \{\psi(y - \mathbf{x}^T\mathbf{T}_{R.la}(G))\mathbf{x} - p_\lambda'(|\boldsymbol{T}_{R.la,0}|)\text{sign}(\boldsymbol{T}_{R.la})\}\ dG(y) = \mathbf{0}. \tag{3.31}$$

By using the local quadratic approximations, we can rewrite (3.31) as

$$\approx \int \left[\psi(y - \mathbf{x}^T\mathbf{T}_{R.la}(G))\mathbf{x} - \{p_\lambda'(|\boldsymbol{T}_{R.la,0}|)/|\boldsymbol{T}_{R.la,0}|\}\boldsymbol{T}_{R.la}(G)\right]\ dG(y) = \mathbf{0}. \tag{3.32}$$

To derive the influence function $\mathbf{T}^{(1)}_{R.la}(G)$, G in (3.32) is substituted with $(1-\varepsilon)G+\varepsilon\delta_y$ as follows,

$$\begin{aligned}\int \Big[&\psi(y - \mathbf{x}^T\mathbf{T}_{R.la}((1-\varepsilon)G + \varepsilon\delta_y))\mathbf{x} \\ &-\{p_\lambda'(|\boldsymbol{T}_{R.la,0}|)/|\boldsymbol{T}_{R.la,0}|\}\boldsymbol{T}_{R.la}((1-\varepsilon)G + \varepsilon\delta_y)\Big]\ d((1-\varepsilon)G + \varepsilon\delta_y) = \mathbf{0}.\end{aligned} \tag{3.33}$$

Differentiating both sides of the equation with respect to ε and setting ε=0 yields

$$\begin{aligned}&\int \left[\psi(y - \mathbf{x}^T\mathbf{T}_{R.la}(G))\mathbf{x} - \{p_\lambda'(|\boldsymbol{T}_{R.la,0}|)/|\boldsymbol{T}_{R.la,0}|\}\boldsymbol{T}_{R.la}(G)\right]\ d(\delta_y(y) - G(y)) \\ &+\Big[\int -\frac{\partial}{\partial\boldsymbol{\beta}}\psi(y - \mathbf{x}^T\mathbf{T}_{R.la}(G))\mathbf{x}\mathbf{x}^T - \{p_\lambda'(|\boldsymbol{T}_{R.la,0}|)/|\boldsymbol{T}_{R.la,0}|\}\Big|_{\boldsymbol{\beta}=\boldsymbol{T}_{R.la}} dG(y) \\ &\cdot\frac{\partial}{\partial\varepsilon}\mathbf{T}_{R.la}((1-\varepsilon)G + \varepsilon\delta_y)\Big] = \mathbf{0}.\end{aligned} \tag{3.34}$$

Consequently, the influence function, $\boldsymbol{T}^{(1)}_{R.la}(G)$, of the functional that defines the robust lasso-type estimator is given by

$$\begin{aligned}\mathbf{T}^{(1)}_{R.la}(G) &\equiv \frac{\partial}{\partial\varepsilon}\mathbf{T}_{R.la}((1-\varepsilon)G + \varepsilon\delta_y)\Big|_{\varepsilon=0} \\ &= \left[\int \frac{\partial}{\partial\boldsymbol{\beta}}\psi(y - \mathbf{x}^T\mathbf{T}_{R.la}(G))\mathbf{x}\mathbf{x}^T + \{p_\lambda'(|\boldsymbol{T}_{R.la,0}|)/|\boldsymbol{T}_{R.la,0}|\}dG(y)\right]^{-1} \\ &\quad\cdot\left[\psi(y - \mathbf{x}^T\mathbf{T}_{R.la}(G))\mathbf{x} - \{p_\lambda'(|\boldsymbol{T}_{R.la,0}|)/|\boldsymbol{T}_{R.la,0}|\}\boldsymbol{T}_{R.la}(G)\right].\end{aligned} \tag{3.35}$$

Thus, the asymptotic bias of the log-likelihood of $f(y|\mathbf{x}, \hat{\boldsymbol{\beta}}_{R.la}(\lambda))$ is given by

$$b^{(1)}_{R.la} = \mathrm{tr}\Big(\Big[\int \psi^{'}(y - \mathbf{x}^T \mathbf{T}_{R.la}(G)) \mathbf{x}\mathbf{x}^T + \{p^{'}_{\lambda}(|\boldsymbol{T}_{R.la,0}|)/|\boldsymbol{T}_{R.la,0}|\} dG \Big]^{-1} \tag{3.36}$$
$$\times [\int \Big[\psi(y - \mathbf{x}^T \mathbf{T}_{R.la}(G)) \mathbf{x} - [p^{'}_{\lambda}(|\boldsymbol{T}_{R.la,0}|)/|\boldsymbol{T}_{R.la,0}|] \boldsymbol{T}_{R.la}(G) \Big]$$
$$\cdot \frac{\partial \log f(y|\mathbf{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \Big|_{\boldsymbol{\beta} = \mathbf{T}_{R.la}(G)} dG] \Big) + O(n^{-1}).$$

By replacing the unknown distribution G by the empirical distribution $\hat{G}$, and subtracting the asymptotic bias estimate from the log-likelihood, we have an information criterion for the statistical model $f(y|\mathbf{x}, \hat{\boldsymbol{\beta}}_{R.la}(\lambda))$ with the functional estimator, $\hat{\boldsymbol{\beta}}_{R.la}(\lambda) = \mathbf{T}(\hat{G})$, in the following

$$GIC_{R.la} = -2 \sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{R.la}(\lambda)) + 2\{R(\psi, \hat{G})^{-1} Q(\psi, \hat{G})\}, \tag{3.37}$$

where

$$R(\psi, \hat{G}) = -\frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n} \Big[\frac{\partial}{\partial \beta_j} \psi(y_i - \mathbf{x}_i^T \mathbf{T}_{R.la}) x_{i,j}^2 \Big] + \{p^{'}_{\lambda}(|T_{R.la,j0}|)/|T_{R.la,j0}|\}, \tag{3.38}$$

$$Q(\psi, \hat{G}) = \frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n} \Big[\Big(\psi(y_i - \mathbf{x}_i^T \mathbf{T}_{R.la}) x_{i,j} - \{p^{'}_{\lambda}(|T_{R.la,j0}|)/|T_{R.la,j0}|\} T_{R.la,j} \Big) \tag{3.39}$$
$$\cdot \frac{\partial \log f(y_i|\mathbf{x}_i, \boldsymbol{\beta})}{\partial \beta_j} \Big|_{\boldsymbol{\beta} = \mathbf{T}_{R.la}} \Big].$$

For the robust sparse regression modeling, we choose the optimal set of regularization parameters and a tuning constant minimizing the information criterion $GIC_{R.la}$, using the grid search. The proposed $GIC_{R.la}$ showed outstanding performance for the robust sparse regression modeling in the viewpoint of sparsity (see Part 2 in Section 3.5. Simulation studies).

3.3 Robust estimation for sparse regression model

We present a robust estimation via outlier-resistance algorithms for L_1-type regularized regression modeling. Although the existing algorithms effectively perform sparse regression modeling as shown in Section 2.3, they suffer from outliers, since the procedures are based on sample mean, standard deviation and correlation or inner product obtained by a non-robust manner. It implies that we cannot expect stable modeling results by using the existing algorithms in the presence of outliers. To overcome the problem, Khan et al. (2007) proposed robust model selection techniques by modify the LARS. In order to robustify the LARS procedure, Khan et al. (2007) used standardized data by median and median absolute deviation, and robust correlation.

We consider the robust sparse regression modeling via the coordinate descent algorithm, which is competitive with the well known LARS for the L_1-type regularization. Although the coordinate descent algorithm effectively performs sparse regression modeling, it also suffers from outliers, since the solution path is delivered based on standardized data by mean, standard deviation, and inner product of predictor and partial residual obtained by a non-robust manner. In order to robust regression modeling, we first standardize data by median and median absolute deviation instead of mean and standard deviation like Khan et al. (2007), and propose robust coordinate descent algorithms based on an outlier-resistant inner product. We also introduce pre-treatment techniques for data cleaning by using the Winsorization and trimming methods based on the robust Mahalanobis distance.

3.3.1 Robust LARS

We briefly introduce the robust linear model selection method via the LARS algorithm proposed Khan et al. (2007). They demonstrated that the LARS algorithm sensitive to outliers because it is based on the mean, variance and correlation. In order to robustify the LARS algorithm, they replaced the mean, variance and correlation estimated in a non-robust manner with median, median absolute deviation (MAD) and robust correlation via the Winsorization technique, respectively.

For robust LARS procedure, they first robustly standardize data by using the median and MAD, and then proposed the bivariate wisorization for robust correlation,

$$\mathbf{u} = \min(\sqrt{(k/D(\mathbf{x}))}, 1), \tag{3.40}$$

where $\mathbf{x} = (x_1, x_2)^T$, $k = 5.99$ as a 95% quantile of the $\chi^2(\text{df} = 2)$ distribution and $D(\mathbf{x})$ is the Mahalanobis distance based on some initial correlation matrix $\mathbf{R}_0$,

$$D(\mathbf{x}_j^{\mathbf{R}_0}) = (\mathbf{x}_j - \bar{\mathbf{X}}_j)^T \mathbf{S}_j^{-1(\mathbf{R}_0)} (\mathbf{x}_j - \bar{\mathbf{X}}_j), \tag{3.41}$$

where superscript $\mathbf{R}_0$ means ones based on the adjusted Winsorized data, and the initial correlation matrix is obtained by adjusted Winsorized data.

Remark 3.3.1 *The adjusted Winsorization controls the outliers by using two tuning constants. The large tuning constant c_1 is used to winsorize the majority of standardized data, and a smaller tuning constant c_2 is used to winsorize the remaining data. They used $c_1 = 2$ and $c_2 = \sqrt{q}c_1$, where $q = n_2/n_1$, n_1 is the number of observations detected as a non-outlier by tuning constant c_1, and $n_2 = n - n_1$.*

Figure 3.3 shows the adjusted Winsorization for initial correlation matrix $\mathbf{R}_0$. From

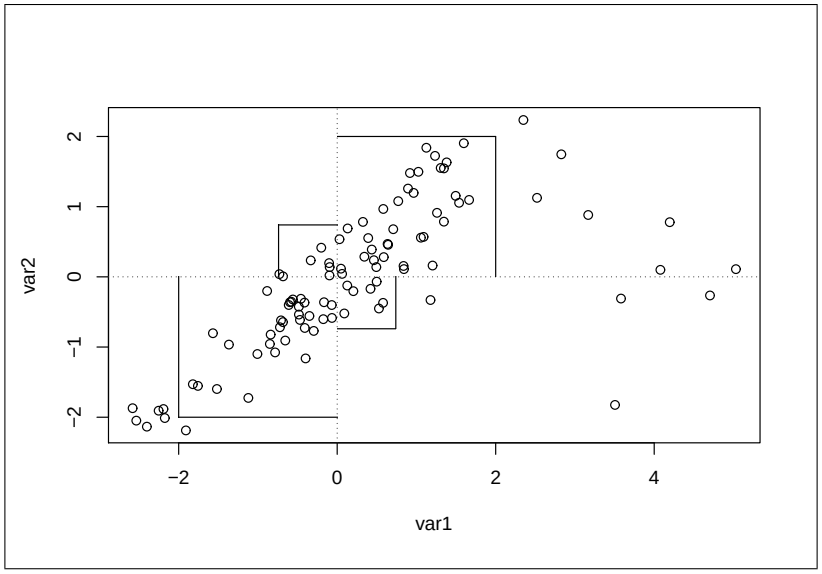


Figure 3.3: Adjusted Winsorization with $c_1 = 2$ for initial correlation proposed by Khan et al. (2007)

the Figure 3.3, it can be seen that using the adjusted Winsorization, the Mahalanobis distance in (3.41) can detect outliers exquisitely.

They propose a robust LARS based on the robust correlation via the bivariate Winsorized data in (3.41).

3.3.2 Robust coordinate descent procedure

It is well known that the coordinate descent algorithm is competitive with the LARS for the lasso-type approaches. Table 3.1 shows the running timings for regression modeling by the coordinate descent algorithm and LARS procedure. The algorithms are implemented as *R* functions (i.e., *glmnet* and *lars*, respectively). We fit the regression model in (2.1) with each p and n based on the lasso. As shown in Table 3.1, the coordinate descent algorithm is faster then LARS procedure, especially in

large dataset. Furthermore, the coordinate descent algorithm shows the outstanding performance for L_1-type regularized regression modeling compared with LARS (see columns forecasting RMSE and True negative (T.N) which means a sparsity).

We consider the robust sparse regression modeling via coordinate descent procedure. As shown in Section 2.3.3, the coordinate descent procedure is based on the data standardized by mean and standard deviation and inner product of x_{ij} with

Table 3.1: Comparison with Coordinate descent and LARS procedures

	model	procedure	timing	RMSE	T.N	F.N
σ=1	p=8,n=40	LARS	24.29	1.10	0.46	0.00
		Coordi	**15.4**	**1.07**	**0.58**	0.00
	$p = 40, n = 40$	LARS	83.3	6.19	0.10	0.00
		Coordi	**20.14**	**2.32**	**0.73**	0.00
	$p = 100, n = 40$	LARS	121.01	-	-	-
		Coordi	**36.53**	-	-	-
	$p = 100, n = 40$	LARS	325.88	-	-	-
		Coordi	**74.85**	-	-	-
σ=5	$p = 8, n = 40$	LARS	23.63	5.49	0.53	0.15
		Coordi	**16.94**	**5.15**	**0.60**	0.14
	$p = 40, n = 40$	LARS	93.55	22.59	0.12	0.04
		Coordi	**26.27**	**8.74**	**0.58**	0.10
	$p = 100, n = 40$	LARS	110.78	-	-	-
		Coordi	**43.00**	-	-	-
	$p = 100, n = 40$	LARS	315.88	-	-	-
		Coordi	**67.95**	-	-	-

partial residual $(y_i - \tilde{y}_i)$, and hence the presence of outliers in either predictors or partial residual may cause distorted results in estimation and variable selection.

To illustrate the drawback, we apply the algorithm to the diabetes dataset (Efron et al., 2004). The dataset consists of the 10 predictor variables, age, sex, BMI, BP and six serum measurements, and a quantitative measure of disease progression as a response variable for $n = 442$ patients. We label the 10 predictor variables as x1 to x10. The 10 predictor variables are standardized: $\sum_i x_{ij}/n = 0$, $\sum_i x_{ij}^2/n = 1$. The coordinate descent algorithm is applied to the diabetes dataset with selected λ by the 10-fold cross-validation. Figure 3.4 (a) shows iterates for each coefficients $\hat{\beta}^{\text{lasso}}(\lambda_{cv})$ by the algorithm. The coefficients $\hat{\beta}^{\text{lasso}}(\lambda_{cv})$ are converged after 26 steps.

To show the seriousness of outlier effect on the procedure, we contaminate the dataset by replacing 10% observations in x1 and x9 with outliers for $N(5, 3^2)$, and then apply the algorithm to the contaminated dataset. In the presence of outliers, the

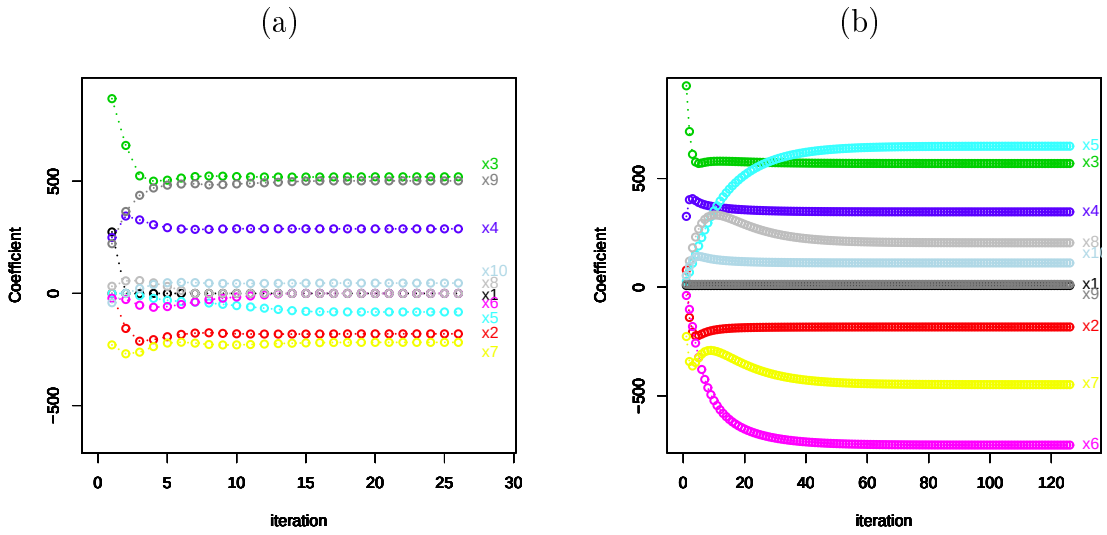


Figure 3.4: Iterates for each coefficients in the coordinate descent procedure

procedure shows unstable process and the iteration number for convergence increases as shown in Figure 3.4 (b). It implies that outliers significantly change the process of the coordinate descent algorithm, which demonstrates the sensitivity of the algorithm against outliers, and thus it makes distorted modeling results.

To settle on the drawback, we consider the robust coordinate descent algorithm based on the standardized data by median and median absolute deviation like Khan et al. (2007), and outlier-resistant inner product via the Winsorization and trimming technique by the Mahalanobis distance, which is basis for multivariate outlier detection. The Mahalanobis distance, however, can be also influenced by outliers, since it is based on location and covariance estimated in a non-robust manner. Rousseeuw and Zomeren (1990) proposed the robust Mahalanobis distance based on the minimum covariance determinant (MCD),

Definition 2 : Minimum Covariance Determinant (MCD).

The MCD is the mean and covariance matrix based on the sample size of h ($h \leq n$) that minimizes the determinant of covariance matrix (John and David, 2004),

$$MCD = (\bar{\mathbf{X}}^*_{jM}, \mathbf{S}^*_{jM}), \tag{3.42}$$

where

$$M=\{\textit{set of } h \textit{ points: } |\mathbf{S}^*_{jM}| \leq |\mathbf{S}^*_{jK}| \forall \textit{ sets } K \textit{ s.t. } \#|K| = h\},$$

where $\#|\omega|$ defines the number of elements set ω

$$\bar{\mathbf{X}}^*_{jM} = \frac{1}{h}\sum_{i \in M} \mathbf{x}_{ij}, \tag{3.43}$$

$$\mathbf{S}^*_{jM} = \frac{1}{h}\sum_{i \in M} (\mathbf{x}_{ij} - \bar{\mathbf{X}}^*_{jM})^T (\mathbf{x}_{ij} - \bar{\mathbf{X}}^*_{jM}). \tag{3.44}$$

The MCD has the highest breakdown point: $h = (n + p + 1)/2$.

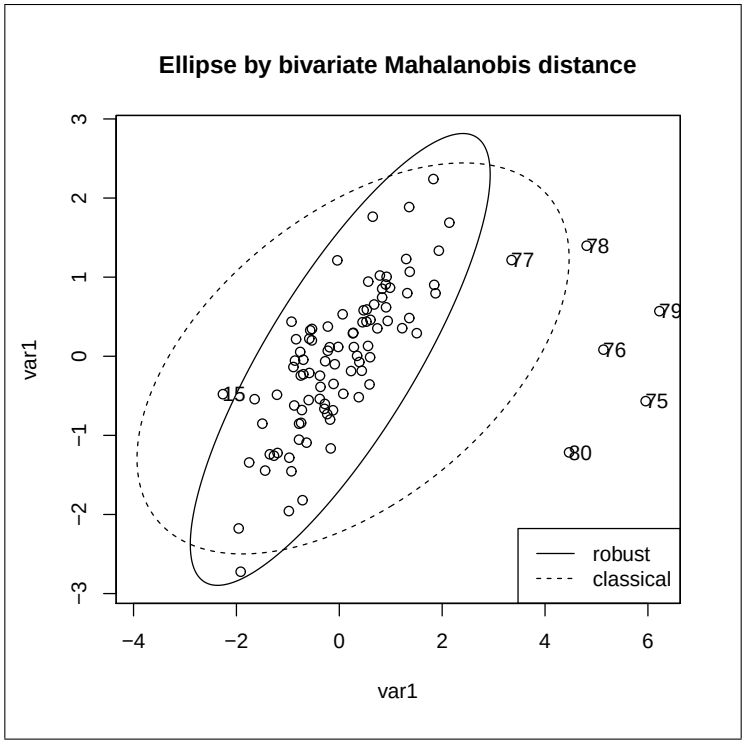


Figure 3.5: Bivariate ellipse plot

By replacing location and covariance matrix with those in MCD, the robust Mahalanobis distance was proposed (Rousseeuw and Zomeren, 1990),

$$RD(\mathbf{x}_j) = (\mathbf{x}_j - \bar{\mathbf{X}}^*_{jM})^T \mathbf{S}^{*-1}_{jM} (\mathbf{x}_j - \bar{\mathbf{X}}^*_{jM}). \tag{3.45}$$

Figure 3.5 shows the ellipse plot based on the Mahalanobis distance. The dashed line shows an ellipse based on the classical $D(\mathbf{x})$ and the normal line shows a robust ellipse based on the $RD(\mathbf{x})$. By using the robust Mahalanobis distance, the observations placed outside of robust ellipse are detected as outliers, and thus they are controlled by some techniques. For instance, the observations 15 and 77 are detected as outliers by using the $RD(\mathbf{x})$, while they are not considered as outliers by using $D(\mathbf{x})$. It implies that the $RD(\mathbf{x})$ more effectively performs outlier detection compared with

$D(\mathbf{x})$, and thus we can expect robust modeling procedure by using the $RD(\mathbf{x})$. We robustify the coordinate descent procedure by using the robust Mahalanobis distance.

(A) W.coordinate descent algorithm

The coordinate descent procedure is progressed by calculate the inner product with j^{th} predictor and partial residual, and thus we should control outliers in both j^{th} predictor and partial residual. In order to robustify the coordinate descent procedure, we proposed the robust bivariate Winsorization via the robust Mahalanobis distance,

$$\text{RD.W.ob}_j = \min(\sqrt{k/RD(\mathbf{x}_j)}, 1)\, \mathbf{x}_j, \tag{3.46}$$

where $k = 5.99$ as the 95% quantile of the $\chi^2(\text{df} = 2)$ distribution like Khan et al. (2007). The coordinate descent procedure is conducted based on a robust inner product by using the Winsorized data RD.W.ob_j which is updated in every iteration $j = 1, 2, ..., p, 1, 2...$, for robust sparse regression modeling. We call this procedure as a **W.coordinate descent algorithm** in Algorithm 1.

Algorithm 1: W.coordinate descent algorithm

Step 1. Set $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_p)$=0.

Step 2. Calculate partial residual: $y_i - \tilde{y}_i (= \sum_{j=1}^{p} \mathbf{x}_i \tilde{\beta}_j)$.

Step 3. Robust bivariate Winsorizing of $\mathbf{x}_j$: $\mathbf{x}_j^W$=$(x_j^W, (y_i - \tilde{y}_i)^W)$ is based on

$$\text{RD.W.ob}_j = \min(\sqrt{k/RD(\mathbf{x}_j)}, 1)\, \boldsymbol{x}_j.$$

Step 4. Apply the coordinate descent algorithm to Winsorized data x_j^W,

$$\tilde{\beta}_j \leftarrow S\Big(\sum_{i=1}^{n} x_{ij}^W (y_i - \tilde{y}_i)^W, \lambda\Big).$$

Step 5. Iterate steps 2 to 4 until convergence.

(B) T.coordinate descent algorithm

We also consider a trimming method, which is widely used to reduce effect of outliers. The trimming technique controls an effect of outliers by eliminating extreme observations at each tails. We propose the robust bivariate trimming technique similar to robust bivariate Winsorization,

$$\text{RD.T.ob}_j = \mathbf{x}_j\{I(\sqrt{k/RD(\mathbf{x}_j)} \geq 1)\}. \tag{3.47}$$

For the robust bivariate trimming technique, we also use the robust Mahalanobis distance $RD(\mathbf{x}_j)$ and $k = 5.99$ as the 95% quantile of the $\chi^2(\text{df} = 2)$ distribution. RD.T.ob_j is updated in each iterations step, and then the variable selection and estimation are conducted base on the robust inner product by the trimmed data RD.T.ob_j in the coordinate descent procedure. By using the bivariate trimming technique, we use the observations placed only in the robust ellipse as shown in Figure 3.5 to calculate the inner product. This implies that the robustified coordinate descent procedure may not be affected by outliers, and thus can performs robust

---

Algorithm 2: T.coordinate descent algorithm

Step 1. Set $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_p)$=0.

Step 2. Calculate partial residual: $y_i - \tilde{y}_i (= \sum_{j=1}^{p} \mathbf{x}_i \tilde{\beta}_j)$.

Step 3. Robust bivariate trimming of $\mathbf{x}_j$: $\mathbf{x}_j^{TR}$=(x_j^{TR},$(y_i - \tilde{y}_i)^{TR}$) is based on

$$\text{RD.T.ob}_j = \mathbf{x}_j\{I(\sqrt{k/RD(\mathbf{x}_j)} \geq 1)\}.$$

Step 4. Apply the coordinate descent algorithm to trimmed data $\boldsymbol{x}_j^{TR}$,

$$\tilde{\beta}_j \leftarrow S\Big(\sum_{i=1}^{n} x_{ij}^{TR}(y_i - \tilde{y}_i)^{TR}, \lambda\Big).$$

Step 5. Iterate steps 2 to 4 until convergence.

---

variable selection and estimation. We call this procedure as a **T.coordinate descent algorithm** in Algorithm 2.

(C) Pre-treatment techniques for outliers

We introduce alternative approaches for robust coordinate descent procedure via pre-treatment techniques for data cleaning. In order to clean the whole dataset, we use the multivariate Winsorization and trimming techniques based on the robust Mahalanobis distance. Figure 3.6 shows the elliptical ball based on the multivariate Mahalanobis distance ($p = 3$). The thick gray elliptical ball shows the robust multivariate Mahalanobis distance and light one shows the classical one in 3-dimension. We control the outliers placed outside of thick elliptical ball by Winsorizing and trimming techniques. In the pre-treatment technique, we control the outliers in p predictor variables

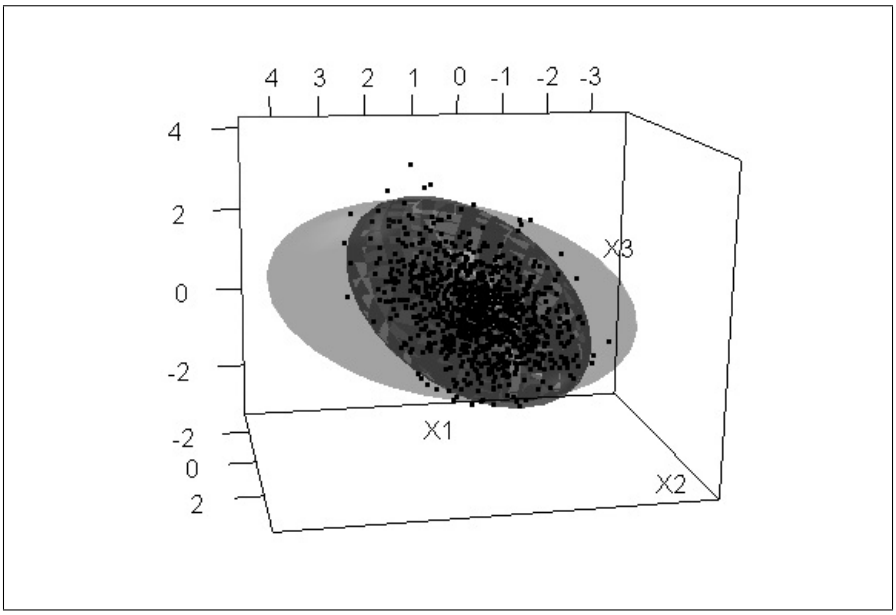


Figure 3.6: 3-dimensional elliptical ball based on multivariate Mahalanobis distance (thick ball: ordinary Mahalanobis distance, light ball: Robust Mahalanobis distance)

and response variable y at once by using multivariate Winsorization and trimming techniques.

- ***Data cleaning via the multivariate Winsorization technique.*** We first introduce the pre-treatment technique by using the multivariate Winsorization. The multivariate Winsorization technique based on the robust multivariate Mahalanobis distance is given,

$$\text{RD.W.x} = \min(\sqrt{k/RD(\mathbf{x})}, 1)\mathbf{x}, \tag{3.48}$$

where $\mathbf{x} = (x_1, x_2, ..., x_p, y)^T$ and $k = \chi^2(\text{df} = p + 1)$ as a 95% quantile of the $\chi^2(\text{df} = p + 1)$ distribution (Khan et al., 2007).

After Winsorizing the dataset, the ordinary coordinate descent procedure performs variable selection and estimation using the Winsorized data RD.W.x. We call this procedure as a **W.pre-treatment**.

- ***Data cleaning via the multivariate trimming technique.*** We also introduce the data cleaning method using the multivariate trimming technique. Similar to Winsorization techinique, robust multivariate Mahalanobis distance is used for detecting outliers, and then we discard the observations detected as outliers by using following multivariate trimming technique,

$$\text{RD.T.x} = \mathbf{x}\{I(\sqrt{k/RD(\mathbf{x})} > 1)\}. \tag{3.49}$$

We also use $k = \chi^2(\text{df} = p + 1)$ as a 95% quantile of the $\chi^2(\text{df} = p + 1)$ distribution.

After cleaning dataset via multivariate trimming technique, the ordinary coordinate descent procedure performs variable selection and estimation using the cleaned

data RD.T.x. We call this procedure as a **T.pre-treatment**.

The proposed robust coordinate procedures can reduce the effect of outliers, and thus we can expect a robust sparse regression modeling under the appropriately selected tuning parameters λ and h (see Part 3 in Section 3.5. Simulation studies).

3.4 Robust selection of the tuning parameters

In this section, we discuss a robust model evaluation for choosing the regularization parameters and tuning constant. Although numerous studies on the robust estimation have been conducted for the outlier-resistant modeling, relatively little attention was paid for the robust evaluation. Ronchetti et al. (1997) introduced the robust cross-validation by using the robust loss function, and Jung (2009) proposed the three type of generalized cross-validation by replace the least squares loss function. Ronchetti and Staudte (1994) proposed the robust version of Mallow's C_p by using a weight function.

In the robust lasso-type approaches, the robust evaluation for choosing the tuning parameters is a vital matter, because the feature of the robust sparse modeling procedure heavily depends on the proper choice of the adjusted parameters. We introduce a robust information criterion based on the bootstrap technique for choosing the tuning parameters in the robust L_1-type regularization. Although the bootstrap information criterion has several advantages on its flexibility and weak assumptions, the bootstrap information criterion cannot perform robustly in the presence of outliers because of a randomly drawn technique of bootstrap method. In order to overcome the drawbacks, we propose a robust bootstrap information criterion via Winsorizing

technique (Srivastava et al., 2010) in line with the efficient bootstrap information criterion (Konishi and Kitagawa, 1996; Park, 2012).

We first briefly introduce the existing robust model selection criteria, and then present the proposed robust efficient bootstrap information criterion.

3.4.1 Literature review: robust model selection criteria

Robust cross-validation

Ronchetti et al. (1997) introduced a robust linear model selection based on the cross-validation. In order to evaluate models robustly, they proposed the robust criterion by using the robust loss function,

$$\sum_{k \in I_v} \rho(\tilde{e}_i), \tag{3.50}$$

where $\{\tilde{e}_k = y_k - \hat{f}^{-k}(\mathbf{x}_k)\}$ and I_v is a validation data. They use a robust loss function $\rho(t) = \min(t^2, b^2\hat{\sigma}_{\tilde{e}}^2)$, where $b = 1.345$ and $\hat{\sigma}_{\tilde{e}} = 1.483 \cdot \text{median}_{i \in k}|\tilde{e}_i - \text{median}(\tilde{e}_i)|$. Ronchetti et al. (1997) summarized the proposed robust cross-validation as following,

1. Data set is divided by K-parts randomly.
2. Model is estimated by using the data set without k^{th} part of data.
3. For k^{th} part of data, compute the robust loss in (3.50).
4. Repeat step 2 to 3 for K data parts, and compute an average the robust prediction criterion for each model.
5. Choose the model minimizing the average prediction criterion in (3.50).

Robust generalized cross-validation

Jung (2009) proposed the robust generalized cross-validation. Let consider the generalized cross-validation as an expression of the leave-one-out cross-validation version in Section 2.4.1,

$$\text{GCV} = \frac{1}{n}\sum_{\alpha=1}^{n}\{\frac{y_\alpha - \hat{f}(\lambda, \mathbf{x}_\alpha)}{1 - h_{\alpha\alpha}}\}^2. \tag{3.51}$$

The GCV is based on the least squares loss function, and hence it suffers from outliers. To overcome the drawback, Jung (2009) considered three type of robust criteria by modify the least squares loss function as follows,

· median generalized cross-validation:

$$\text{MEDIAN.CV} = \text{median}\big[\{\frac{y_\alpha - \hat{f}(\lambda, \mathbf{x}_\alpha)}{1 - h_{\alpha\alpha}}\}^2\big]. \tag{3.52}$$

· trimmed sum of squares generalized cross-validation:

$$\text{TSS.GCV} = \frac{1}{h}\sum_{\alpha=1}^{h}\{\frac{y_\alpha - \hat{f}(\lambda, \mathbf{x}_\alpha)}{1 - h_{\alpha\alpha}}\}^2. \tag{3.53}$$

· mean absolute generalized cross-validation:

$$\text{MA.GCV} = \frac{1}{n}\sum_{\alpha=1}^{n}|\frac{y_\alpha - \hat{f}(\lambda, \mathbf{x}_\alpha)}{1 - h_{\alpha\alpha}}|. \tag{3.54}$$

Robust Mallow's C_p

The Mallow's C_p is a useful tool for model selection in regression modeling. Mallow considered following criterion based on the mean squares error for model evaluation,

$$\Gamma = \frac{1}{\sigma^2}E[\sum_{i=1}^{n}(\hat{y}_i - E(y_i)^2)], \tag{3.55}$$

and then proposed the following estimate, called a Mallow's C_p,

$$C_p = \hat{\Gamma} = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n. \tag{3.56}$$

Ronchetti and Staudte (1994) demonstrated the ordinary Mallow's C_p is sensitive to outliers because it is based on the RSS_p, and then proposed robust version of Mallow's C_p based on the M-estimator $\hat{\beta}$ with weights $\hat{w}_i = \psi(r_i)/r_i$. They defined the robust version of Γ,

$$\Gamma(w) = \frac{1}{\sigma^2} E[\sum_{i=1}^{n} \hat{w}_i^2 (\hat{y}_i - E(y_i)^2)]. \tag{3.57}$$

And then, Ronchetti and Staudte (1994) proposed a robust version of Mallow's C_p,

$$RC_p = \frac{W_p}{\hat{\sigma}^2} - (U_p - V_p), \tag{3.58}$$

where $W_p = \sum_{i=1}^{n} \hat{w}_i^2 (y_i - \hat{y}_i)^2$, $\hat{\sigma}^2$ is a robust and consistent estimator of σ^2 in the full model, and U_p and V_p are computed based on the weighted function and number of parameters p.

3.4.2 Robust efficient bootstrap information criterion

We consider a robust version of the bootstrap information criterion. Although the bootstrap technique is a practical method, it has a demerit in the presence of outliers that a bootstrap sample may contain more outliers compared with those in the original sample, since the bootstrap sample is drawn randomly (see Figure 3.7). Table 3.2 shows the seriousness of the problem that bootstrap sample contains more outliers than those in the original sample over 100,000 simulated datasets. As shown in Table 3.2, overall more than 35% of bootstrap samples contain more outliers than those

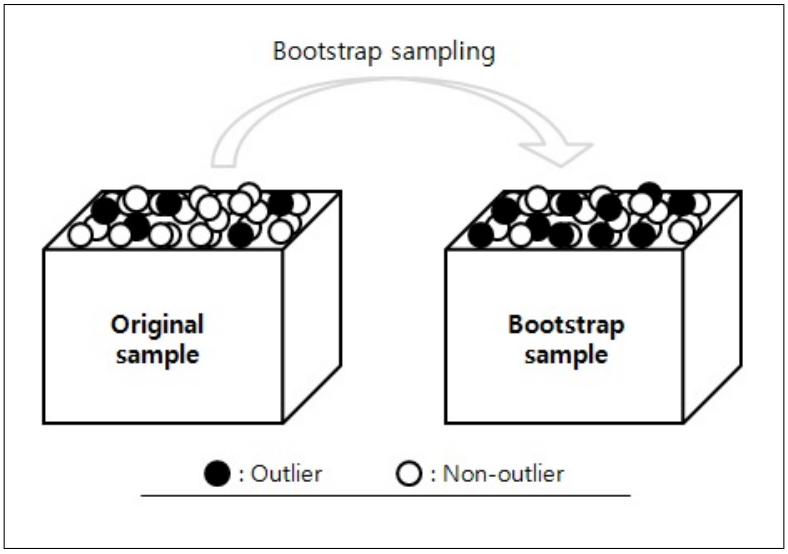


Figure 3.7: The drawback of bootstrap sample in the presence of outliers

in the original samples. This implies that the resulting criterion from the bootstrap sample may produce biased results in the presence of outliers, and hence it does not perform well as a tuning parameter selector. In order to overcome the drawback, we propose a robust bootstrap information criterion via Winsorization technique (Srivastava et al., 2010) in line with the efficient bootstrap information criterion for choosing an optimal set of the regularization parameters and a tuning

Table 3.2: Percentage that bootstrap sample contains more outliers than original sample

	Proportion (%) of outliers in the original sample			
n	1%	5%	10%	15%
100	0.26	0.39	0.42	0.43
500	0.38	0.45	0.46	0.47
1000	0.42	0.46	0.48	0.48

constant (Park, 2012).

A Winsorization is a robust statistical technique that aims to reduce the effect of outliers in the sample (Yale and Forsythe, 1976). First, we introduce a Winsorization bootstrap method (Singh, 1998; Srivastava et al., 2010). Suppose that the order statistics of the original data be denoted by $y_{[1]}, y_{[2]}, ..., y_{[n]}$. A δ-Winsorized sample for $\{y_i\}$ is given by

$$\begin{aligned} y_i^* &= y_{[l+1]}, && \text{if } y_i \leq y_{[l]}, \\ &= y_{[n-l]}, && \text{if } y_i \geq y_{[n-l+1]}, \\ &= y_i, && \text{otherwise}, \end{aligned} \tag{3.59}$$

where $\delta = l/n$, $0 \leq \delta \leq 1/2$ represents a Winsorizing proportion. The Winsorized bootstrap sample $\{y_i^{**}\}$ are randomly drawn from the δ-Winsorized sample $\{y_i^*\}$. This implies that the Winsorized bootstrap sample may not be affected by outliers which are greater than $y_{[l]}$ or smaller than $y_{[n-l+1]}$. Thus, we can reduce the effect of outliers in the bootstrap technique.

Remark 3.4.1 *In the Winsorization technique, choosing the Winsorizing proportion δ is crucial in practice. The simplest way to choose the δ is to specify them in advance (Chen et al., 2001). The δ was determined adaptively from the data in literatures (Welsh, 1987; Dodge and Jurecková, 1997). Chen et al. (2001) mentioned that this issue is a largely philosophical question as to which approaches individual users prefer. Srivastara et al. (2010) showed that the wisorization technique with $\delta \approx$ "proportion of outliers in the original dataset" outperforms in bootstrap regression. Therefore, we use the Winsorizing proportion $\delta =$ "proportion of outliers in the original dataset".*

For outlier-resistant model evaluation, we propose a robust efficient bootstrap information criterion via the Winsorized bootstrap sample. By using the Winsorized bootstrap sample, the bootstrap bias estimate of (3.14) is given by

$$b^{**}(\hat{G}) = E_{\hat{G}(\mathbf{y}^{**})}\left[\sum_{i=1}^{n} \log f(y_i^{**}|\hat{\boldsymbol{\theta}}(\mathbf{y}_n^{**})) - nE_{\hat{G}(z^{**})}\left[\log f(Z^{**}|\hat{\boldsymbol{\theta}}(\mathbf{y}_n^{**}))\right]\right]. \tag{3.60}$$

Let us extract B sets of Winsorized bootstrap samples of size n and write the b^{th} Winsorized bootstrap sample as $\mathbf{y}_n^{**}(b) = \{y_1^{**}(b), ..., y_n^{**}(b)\}$. In the wisorized bootstrap estimate, (3.21) is replaced by

$$E_{\hat{G}}[D(\mathbf{y}_n^{**}; \hat{G})] = E_{\hat{G}}[D_1(\mathbf{y}_n^{**}; \hat{G}) + D_3(\mathbf{y}_n^{**}; \hat{G})]. \tag{3.61}$$

Therefore, the bootstrap bias estimate of (3.14) is substituted by

$$b_B^w(\hat{G}) = \frac{1}{B}\sum_{b=1}^{B}\{D_1(\mathbf{y}_n^{**}(b); \hat{G}) + D_3(\mathbf{y}_n^{**}(b); \hat{G})\}. \tag{3.62}$$

Consequently, the proposed robust efficient bootstrap information criterion is given by

$$\begin{aligned} R.EIC_{\text{eff}} &= -2\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) + 2\{b_B^w(\hat{G})\} \qquad (3.63)\\ &= -2\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) + \frac{2}{B}\sum_{i=1}^{B}\{D_1(\mathbf{y}_n^{**}(b); \hat{G}) + D_3(\mathbf{y}_n^{**}(b); \hat{G})\}. \end{aligned}$$

By using the R.EIC$_{\text{eff}}$, the variance of the bootstrap estimates caused by simulation can be reduced extensively, and then the number of bootstrap replications may be greatly reduced. Furthermore, we can perform accurate and stable model evaluation even in the presence of outliers.

We choose an optimal set of the regularization parameters and a tuning constant in the robust lasso-type regularization by using the R.EIC$_{\text{eff}}$ based on the grid search.

To calculate the R.EIC$_{\text{eff}}$ for the robust sparse regression model, the Winsorized bootstrap samples denoted as $\mathbf{y}_n^{**} = \{y_1^{**}, ..., y_n^{**}\}$ are generated using the x-fixing method. In the x-fixing method, $\mathbf{y}_n^{**} = \mathbf{x}_n^T\hat{\boldsymbol{\beta}} + \mathbf{e}_n^{**}$, where $\mathbf{e}_n^{**}$ are randomly drawn from Winsorized sample $\mathbf{e}_n^*$ of $\mathbf{e}_n (= \mathbf{y}_n - \mathbf{x}_n^T\hat{\boldsymbol{\beta}})$,

$$\begin{aligned} e_i^* &= e_{[l+1]}, \qquad \text{if } e_i \leq e_{[l]}, \\ &= e_{[n-l]}, \qquad \text{if } e_i \geq e_{[n-l+1]}, \\ &= e_i, \qquad \text{otherwise.} \end{aligned} \tag{3.64}$$

Afterwards, we calculate the R.EIC$_{\text{eff}}$ based on the estimate $\hat{\boldsymbol{\beta}}$ by the robust lasso-type approaches at the each set of the regularization parameters and a tuning constant. Finally, we perform model selection and estimation by choosing the optimal set of these tuning parameters that minimizes the R.EIC$_{\text{eff}}$.

The proposed R.EIC$_{\text{eff}}$ showed the robust and efficient performance for the robust sparse regression modeling by using the Winsorization technique and variance reduction method (see Part 4 in Section 3.5. Simulation studies).

3.5 Simulation studies

Monte Carlo simulations are conducted to investigate the effectiveness of the proposed robust modeling strategies. In this chapter, we introduced the robust L_1-type regularization and methods for choosing the tuning parameters in the information-theoretic view point. We also introduced the robust algorithm and model evaluation criterion for robust sparse regression modeling.

In order to show the effectiveness of the proposed methods, we evaluate our meth-

ods comparing with existing ones. This section is composed by the following four parts to evaluate each proposed robust modeling strategies,

Part 1 : LTS-Ela and efficient bootstrap information criterion.

Part 2 : $\text{GIC}_{R.la}$ via local quadratic approximation.

Part 3 : Robust coordinate descent procedure.

Part 4 : Robust efficient bootstrap information criterion.

We simulated N datasets consisting of n observations from the following model

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \sigma \varepsilon_i, \qquad i = 1, ..., n, \tag{3.65}$$

where $\boldsymbol{\beta}$ is p-dimensional vector and ε_i are standard normal. The correlation between x_l and x_m is $\rho^{|l-m|}$ with ρ=0.5.

In order to evaluate the proposed methods, we compare the simulation results for variable selection and forecasting accuracy. The results of variable selection are showed as average percentage of zero coefficients in columns "T.N" and "F.N". The "T.N" means a true negative (i.e., the average percentage of true zero coefficients, that were correctly set to zero), and "F.N" means a false negative (i.e., average number of the true non-zero coefficients, incorrectly set to zero). And forecasting accuracy is measured forecasting root mean squares error (RMSE) by N simulated datasets.

Part 1 : LTS-Ela and efficient bootstrap information criterion.

In Part 1, we evaluate the proposed robust L_1-type regularization, called a LTS-Ela compared with LTS-lasso, and efficient bootstrap information criterion as a tuning parameter selector.

We simulated $N = 50$ datasets in $\sigma = 1$. Two simulations are conducted in the presence of 0%, 10%, 20% and 30% outliers for $\varepsilon_i \sim N(10, 3)$,

1. Simulation 1:

   $p = 10$ as $\boldsymbol{\beta} = (2, 3, 0, 0, 1.5, 0, 0, 1, 0, 0)^T$,

2. Simulation 2:

   $p = 40$ as $\boldsymbol{\beta} = (\underbrace{2, ..., 2}_{5}, \underbrace{0, ..., 0}_{5}, \underbrace{5, ..., 5}_{5}, \underbrace{0, ..., 0}_{5}, \underbrace{2, ..., 2}_{5}, \underbrace{0, ..., 0}_{5}, \underbrace{5, ..., 5}_{5}, \underbrace{0, ..., 0}_{5})^T$.

We first show the stability of the EIC_{eff} compared with the cross-validation (CV). Table 3.3 shows the standard deviation of selected tuning parameters and forecasting root mean squares error (RMSE) in the ordinary elastic net. From Table 3.3, it can be seen that the EIC_{eff} stably performs robust sparse regression modeling compared with the CV.

Table 3.4 and 3.5 show the simulation results for variable selection in Simulation

Table 3.3: Standard deviation of tuning parameters and RMSE in Part 1

Outliers		Simulation 1			Simulation 2		
		λ_1	λ_2	RMSE	λ_1	λ_2	RMSE
0%	CV	0.20	0.07	0.63	0.11	0.31	33.54
	EIC_{eff}	**0.11**	**0.06**	**0.51**	**0.07**	**0.00**	**22.65**
10%	CV	0.30	0.16	**2.56**	0.12	0.25	35.79
	EIC_{eff}	**0.28**	**0.14**	2.94	**0.07**	**0.00**	**18.34**
20%	CV	0.31	0.17	4.76	0.12	0.26	47.78
	EIC_{eff}	**0.28**	**0.16**	**3.88**	**0.10**	**0.00**	**24.08**
30%	CV	0.28	**0.13**	6.78	0.14	0.24	38.30
	EIC_{eff}	0.28	0.16	**5.87**	**0.10**	**0.00**	**22.05**

1 and Simulation 2, respectively. The parts in the shadow of the gray appear the proposed methods. From the columns "T.N" in Table 3.4 and 3.5, it can be seen that the proposed LTS-Ela based on the efficient bootstrap information criterion is outstanding in the viewpoint of the "sparsity", which is a crucial property of the lasso-type approaches. Although the proposed $\mathrm{EIC}_{\mathrm{eff}}$ is not outstanding for "F.N", this is an inevitable result, since there is a trade-off between bias and variance. We focus

Table 3.4: Average number of zero in Part 1 (Simulation 1)

			T.N (%)		F.N (%)	
			LTS-lasso	LTS-Ela	LTS-lasso	LTS-Ela
Outliers	0%	CV	0.35	0.29	0.01	0.01
		AIC	0.05	0.20	0.00	0.00
		BIC	0.05	0.20	0.00	0.00
		$\mathrm{EIC}_{\mathrm{eff}}$	0.29	**0.43**	0.00	0.00
	10%	CV	0.43	0.31	0.01	0.03
		AIC	0.06	0.39	0.00	0.12
		BIC	0.05	0.38	0.00	0.11
		$\mathrm{EIC}_{\mathrm{eff}}$	0.50	**0.61**	0.11	0.15
	20%	CV	0.44	0.37	0.05	0.07
		AIC	0.10	0.27	0.02	0.05
		BIC	0.05	0.25	0.02	0.05
		$\mathrm{EIC}_{\mathrm{eff}}$	0.42	**0.60**	0.11	0.19
	30%	CV	0.46	0.43	0.14	0.17
		AIC	0.11	0.18	0.04	0.07
		BIC	0.10	0.16	0.04	0.06
		$\mathrm{EIC}_{\mathrm{eff}}$	0.50	**0.61**	0.15	0.21

Table 3.5: Average number of zero in Part 1 (Simulation 2)

			T.N (%)		F.N (%)	
			LTS-lasso	LTS-Ela	LTS-lasso	LTS-Ela
Outliers	0%	CV	0.24	0.36	0.01	0.10
		AIC	0.17	0.21	0.00	0.01
		BIC	0.17	0.21	0.00	0.01
		EIC_{eff}	0.13	**0.46**	0.01	0.04
	10%	CV	0.27	0.44	0.05	0.19
		AIC	0.17	0.24	0.04	0.02
		BIC	0.17	0.23	0.04	0.02
		EIC_{eff}	0.10	**0.45**	0.05	0.07
	20%	CV	0.34	0.41	0.10	0.20
		AIC	0.18	0.28	0.04	0.03
		BIC	0.18	0.26	0.04	0.03
		EIC_{eff}	0.12	**0.54**	0.04	0.13
	30%	CV	0.31	0.37	0.10	0.17
		AIC	0.12	0.30	0.05	0.03
		BIC	0.12	0.28	0.05	0.03
		EIC_{eff}	0.11	**0.51**	0.05	0.12

on the "T.N", because a main aim of the lasso-type approaches is "sparsity". We also compare the forecast accuracy of the LTS-lasso and proposed LTS-Ela. Because we confirmed the superiority of the efficient bootstrap information criterion to the other criteria in Table 3.3, 3.4 and 3.5, we will show the forecasting results based on the efficient bootstrap information criterion. The root mean square errors (RMSE) by over the 50 simulated datasets are shown in Table 3.6. From Table 3.6, it can be

Table 3.6: Forecasting root mean square error in Part 1

		0%	10%	20%	30%
Simulation 1	LTS-lasso	**1.09**	2.24	2.87	3.62
	LTS-Ela	1.19	**2.20**	**2.67**	**3.31**
Simulation 2	LTS-lasso	**4.29**	13.15	13.40	28.48
	LTS-Ela	5.74	**7.20**	**8.25**	**8.87**

seen that the proposed LTS-Ela outperforms for the forecasting accuracy compared with the LTS-lasso. In Simulation 2, the LTS-lasso showed poor results, especially under the highly contaminated data. This implies that the lasso cannot perform well in $p > n$ situation.

In summary, the proposed LTS-Ela based on the $\mathrm{EIC}_{\mathrm{eff}}$ outperforms for "sparsity" and forecasting accuracy in the presence of outliers.

Part 2 : $\mathrm{GIC}_{R.la}$ via local quadratic approximation

In part 2, we evaluate the proposed $\mathrm{GIC}_{R.la}$. In order to evaluate the proposed $\mathrm{GIC}_{R.la}$, we compare with results by the BIC which showed the superior performance for choosing the regularization parameters in Wang et al. (2007), and results by the cross-validation.

We evaluate the proposed method as a tuning parameter selector in the M-lasso and M-SCAD with the Huber function,

- M-lasso:

$$\hat{\boldsymbol{\beta}}^{\text{M-la}} = \arg\min_{\boldsymbol{\beta}} \{\sum_{i=1}^{n} \rho(r_i) + \lambda \sum_{j=1}^{p} |\beta_j|\}. \tag{3.66}$$

- M-SCAD:

$$\hat{\boldsymbol{\beta}}^{\text{M-SCAD}} = \arg\min_{\boldsymbol{\beta}} \{\sum_{i=1}^{n} \rho(r_i) + \sum_{j=1}^{p} p_\lambda(|\beta_j|)\}, \tag{3.67}$$

where

$$p_\lambda(|\beta_j|) = \lambda|\beta_j|, \qquad if \ \ |\beta_j| \leq \lambda; \tag{3.68}$$

$$= -(\frac{|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)}), \qquad if \ \ \lambda < |\beta_j| \leq a\lambda; \tag{3.69}$$

$$= \frac{(a+1)\lambda^2}{2}, \qquad if \ \ |\beta_j| > a\lambda. \tag{3.70}$$

For model estimation by the M-lasso and M-SCAD, we use an iterative reweighted least square (IRLS) algorithm (Zhang et al., 2009). Two simulations were

Table 3.7: RMSE and sparsity in Part 2 (Simulation 1)

	Method	Outliers	RMSE	No.of Zeros		Outliers	RMSE	No.of Zeros	
				T.N	F.N			T.N	F.N
M-Lasso	$GIC_{R.la}$	1%	1.10	**5.8**	0.0	5%	**1.12**	**7.0**	0.0
	BIC		1.09	2.8	0.0		**1.12**	4.2	0.0
	CV		**1.07**	5.6	0.0		1.13	5.8	5.8
	$GIC_{R.la}$	15%	**1.52**	**10.4**	0.0	sensible	10.30	**6.0**	0.0
	BIC		**1.52**	8.8	0.0		**10.26**	3.4	0.0
	CV		1.55	5.0	0.0		10.27	0.4	0.0
M-SCAD	$GIC_{R.la}$	1%	1.10	**6.0**	0.0	5%	**1.11**	9.8	0.0
	BIC		1.09	1.8	0.0		**1.11**	**10.0**	0.0
	CV		**1.07**	4.4	0.0		1.14	5.6	0.0
	$GIC_{R.la}$	15%	1.52	**13.0**	0.0	sensible	**10.33**	**10.6**	0.0
	BIC		1.52	9.4	0.0		**10.33**	7.4	0.0
	CV		**1.50**	1.4	0.0		10.52	0.8	0.0

conducted under the cases ε_i are standard normal with 1%, 5% and 15% outliers for $\varepsilon_i \sim N(10,1)$ in $\sigma = 1$, and sensible outliers for $\varepsilon_i \sim D/\sqrt{\text{var}(D)}, \sigma = 9.67$, where D is a standard double exponential distribution,

1. Simulation 1: $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ for $n = 80$,

2. Simulation 2: $p = 40$ as $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0, 3, 1.5, 0, 0, 2, 0, 0, 0, 3, 1.5, 0, 0)^T$ for $n = 40$.

Table 3.7 and 3.8 compare the forecasting root mean square errors (RMSE) and simulation results for variable selection.

From the columns "RMSE", it can be seen that all the model selection procedures

Table 3.8: RMSE and sparsity in Part 2 (Simulation 2)

	Method	Outliers	RMSE	No.of Zeros		Outliers	RMSE	No.of Zeros	
				T.N	F.N			T.N	F.N
M-Lasso	$GIC_{R.la}$	1%	1.44	**0.2**	0.0	5%	1.56	**1.4**	0.0
	BIC		1.43	**0.2**	0.0		1.54	**1.4**	0.0
	CV		**1.41**	0.0	0.0		**1.47**	1.3	5.8
	$GIC_{R.la}$	15%	3.07	**2.8**	0.0	sensible	12.57	**20.8**	0.1
	BIC		3.00	2.3	0.0		**12.55**	5.0	0.0
	CV		**2.70**	2.5	0.0		12.79	3.0	0.0
M-SCAD	$GIC_{R.la}$	1%	1.46	0.4	0.0	5%	1.53	**2.0**	0.0
	BIC		1.46	0.4	0.0		1.52	0.8	0.0
	CV		**1.36**	**1.8**	0.0		**1.47**	**2.0**	0.0
	$GIC_{R.la}$	15%	3.00	**5.0**	0.0	sensible	13.50	**15.0**	0.1
	BIC		2.97	3.0	0.0		13.03	2.1	0.0
	CV		**2.76**	3.2	0.0		**12.68**	1.0	0.0

give the similar forecasting results in Simulation 1, and the cross-validation shows superiority in Simulation 2. The columns "T.N" show that the $\mathrm{GIC}_{R.la}$ is superior for the "sparsity" in all outlier situations and simulation settings. It implies that the proposed $\mathrm{GIC}_{R.la}$ is a useful tool for robust L_1-type regression modeling, especially for the variable selection.

Part 3 : Evaluation of robust coordinate descent procedure.

In part 3, we evaluate the proposed robust coordinate descent algorithm. We simulated $N = 100$ datasets consisting of $n = 80$ observations from (3.65) with $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and $\sigma = 1$. The tuning parameters λ and h are selected by the generalized information criterion via local quadratic approximation as shown in Section 3.2.2. We consider the following four situations that outliers in only response variable and outliers in both response and predictor variables.

(a) 10% outliers for $N(30, 1)$ in only y_i

(b) 10% outliers for $N(0, 5)$ in only y_i

(c) 5% outliers for $N(30, 1)$ in y_i, and 5% outliers for $N(0, 5)$ in x_1 and x_5

(d) 5% outliers for $N(0, 5)$ in y_i, and 5% outliers for $N(0, 5)$ in x_1 and x_5

We first show the stability of the proposed procedures. Table 3.9 shows the standard deviation of estimated non-zero coefficients $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_5$. From Table 3.9, it can be clearly seen that the proposed methods are more stable than ordinary procedure. The **W.pre-treatment** and **T.pre-treatment**, especially, show the outstanding performance in the all situations. We also evaluate the forecasting accuracy

and variable selection results in Table 3.10. It can be seen from Table 3.10 that the proposed robust procedures produce reliable regression modeling results even in the presence of outliers, especially the **W.coordinate descent algorithm** shows the best performance in overall. In short, the proposed robust coordinate descent

Table 3.9: Standard deviation of estimated $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_5$ in Part 3

		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_5$
(a)	coordi.al	1.256	1.261	1.287
	W.coordi.al	0.566	0.336	0.449
	T.coordi.al	0.526	0.396	0.442
	W.pre-treatment	0.799	0.617	0.730
	T.pre-treatment	0.843	0.538	0.720
(b)	coordi.al	0.888	0.732	0.919
	W.coordi.al	0.509	0.365	0.458
	T.coordi.al	0.503	0.378	0.456
	W.pre-treatment	0.746	0.514	0.674
	T.pre-treatment	0.690	0.559	0.683
(c)	coordi.al	1.237	0.772	1.112
	W.coordi.al	0.436	0.324	0.365
	T.coordi.al	0.467	0.327	0.431
	W.pre-treatment	0.957	0.546	0.876
	T.pre-treatment	1.219	0.554	0.978
(d)	coordi.al	1.211	0.623	1.013
	W.coordi.al	0.458	0.281	0.375
	T.coordi.al	0.443	0.287	0.374
	W.pre-treatment	1.127	0.516	0.980
	T.pre-treatment	1.129	0.462	0.853

Table 3.10: RMSE and sparsity in Part 3

		RMSE	T.N	F.N
	coordi.al	3.73	0.50	0.19
	W.coordi.al	**2.12**	**0.96**	**0.10**
(a)	T.coordi.al	2.26	0.95	**0.10**
	W.pre-treatment	3.29	0.88	0.13
	T.pre-treatment	3.30	0.85	0.11
	coordi.al	2.63	0.79	0.14
	W.coordi.al	**1.98**	**0.95**	**0.08**
(b)	T.coordi.al	2.05	0.94	0.10
	W.pre-treatment	2.32	0.90	0.09
	T.pre-treatment	2.34	0.91	0.13
	coordi.al	2.64	0.68	0.11
	W.coordi.al	**1.76**	**0.95**	**0.05**
(c)	T.coordi.al	**1.76**	0.94	**0.05**
	W.pre-treatment	2.10	0.94	0.10
	T.pre-treatment	2.17	0.93	0.12
	coordi.al	2.11	0.81	0.12
	W.coordi.al	**1.62**	0.94	**0.04**
(d)	T.coordi.al	1.66	0.93	**0.04**
	W.pre-treatment	1.79	**0.95**	0.10
	T.pre-treatment	1.59	**0.95**	0.06

procedures are superior to existing one in the viewpoint of the stability, forecasting accuracy and sparsity in the presence of outliers.

Part 4 : Robust efficient bootstrap information criterion.

In part 4, we evaluate the robust model evaluation criterion, called a robust efficient

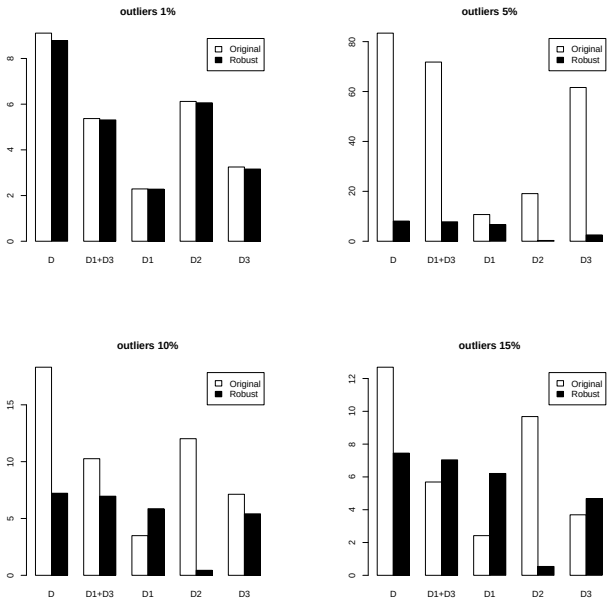


Figure 3.8: Standard deviation of bootstrap estimate of the D, $D_1 + D_3$, D_1, D_2 and D_3

bootstrap information criterion for choosing the tuning parameters. First, we show the stability of the proposed robust efficient bootstrap information criterion in the presence of outliers. Figure 3.8 shows bar plots of the standard deviation of bootstrap estimates D, $D_1 + D_3$, D_1, D_2 and D_3, for sample size $n = 100$. From the Figure 3.8, it can be clearly seen that bootstrap estimates D, $D_1 + D_3$, D_1, D_2, and D_3 in the proposed robust bootstrap information criterion (black bar plots) show smaller standard deviation compared with those in the existing one (white bar plots). It implies that the proposed robust bootstrap information criterion is more efficient and stable against outliers than the existing one, and thus we can expect effective and robust sparse regression modeling by using the proposed method.

We evaluate the proposed robust efficient bootstrap information criterion as a

tuning parameters selector for robust sparse regression modeling. In this part, we simulated $N = 50$ datasets consisting of $n = 80$ observations from (3.65) with $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and $\sigma = 1$. Simulations are conducted in the presence of 5%, 10%, and 15% outliers for $\varepsilon_i \sim N(30, 3)$. To evaluate the proposed method, we choose the regularization parameters and tuning constant by the robust efficient bootstrap information criterion and by the ordinary bootstrap information criterion. We also compare the results by the 10-fold cross-validation. To find the solution of the robust L_1-type regularization, we use an iterative reweighted least square (IRLS) algorithm (Zhang et al., 2009).

We conduct the three simulations for robust sparse regression modeling by the LTS-lasso, M-lasso and M-SCAD with the Huber M-function in (3.4). We show the variable selection results and forecasting accuracy of robust sparse regression modeling under the LTS-lasso, M-lasso and M-SCAD in Table 3.11, 3.12 and 3.13, respectively.

- LTS-lasso

Table 3.11: LTS-lasso in Part 4

Outlier	Method	T.N	F.N	RMSE
5%	CV	0.036	0.000	1.82
	Eff.Boot.IC	0.036	0.000	1.99
	Robust.Eff.Boot.IC	**0.068**	0.000	**1.67**
10%	CV	0.008	0.000	3.17
	Eff.Boot.IC	0.020	0.000	3.33
	Robust.Eff.Boot.IC	**0.048**	0.000	**2.83**
15%	CV	0.004	0.000	4.56
	Eff.Boot.IC	0.008	0.000	5.00
	Robust.Eff.Boot.IC	**0.020**	0.000	**4.07**

- M-lasso

Table 3.12: M-lasso in Part 4

Outlier	Method	T.N	F.N	RMSE
5%	CV	0.076	0.000	1.75
	Eff.Boot.IC	0.076	0.000	1.75
	Robust.Eff.Boot.IC	**0.086**	0.003	**1.74**
10%	CV	0.076	0.007	3.59
	Eff.Boot.IC	0.054	0.000	3.47
	Robust.Eff.Boot.IC	**0.086**	0.003	**3.44**
15%	CV	0.042	0.020	5.25
	Eff.Boot.IC	0.056	0.017	5.24
	Robust.Eff.Boot.IC	**0.060**	0.020	**5.20**

- M-SCAD

Table 3.13: M-SCAD in Part 4

Outlier	Method	T.N	F.N	RMSE
5%	CV	0.076	0.000	1.75
	Eff.Boot.IC	0.084	0.000	1.73
	Robust.Eff.Boot.IC	**0.152**	0.000	**1.72**
10%	CV	0.036	0.000	**3.45**
	Eff.Boot.IC	0.036	0.000	3.58
	Robust.Eff.Boot.IC	**0.064**	0.020	3.58
15%	CV	0.008	0.070	5.39
	Eff.Boot.IC	0.052	0.040	5.29
	Robust.Eff.Boot.IC	**0.084**	0.070	**5.27**

From the column "T.N" in all Tables 3.11, 3.12 and 3.13, it can be seen that the proposed robust efficient bootstrap information criterion is a useful tool as a tuning

parameter selector for "sparsity" in the presence of outliers. It can be also seen that the proposed method is superior to the existing ones for the forecasting accuracy (see column "RMSE"). In short, the proposed robust efficient bootstrap information criterion is effective for robust sparse regression modeling via LTS-lasso, M-lasso and M-SCAD in the viewpoint of the "sparsity" and forecasting accuracy.

3.6 Real-world examples

We illustrate the proposed robust modeling strategies through the real-world data analysis to evaluate the practicality. In this section, we evaluate proposed LTS-Ela with efficient bootstrap information criterion through the McDonald and Schwing dataset (Croux and Ruiz-Gazen, 2005), and robust coordinate descent procedures through the crime data (Agresti and Finlay, 1997).

Part 1: LTS-ela with efficient bootstrap information criterion

In Part 1 of real-world examples, we apply the proposed LTS-Ela and efficient bootstrap information criterion to the McDonald and Schwing dataset (Croux and Ruiz-Gazen, 2005). The McDonald and Schwing dataset consists of p=16 variables

· Y: total Age Adjusted Mortality Rate

· x1: mean annual precipitation in inches

· x2: mean January temperature in degrees Fahrenheit

· x3: mean July temperature in degrees Fahrenheit

· x4: percent of 1960 SMSA population that is 65 years of age or over

· x5: population per household, 1960 SMSA

· x6: median school years completed for those over 25 in 1960 SMSA

· x7: percent of housing units that are found with facilities

· x8: population per square mile in urbanized area in 1960

· x9: percent of 1960 urbanized area population that is non-white

· x10: percent employment in white-collar occupations in 1960 urbanized area

· x11: percent of families with income under 3,000 in 1960 urbanized area

· x12: relative population potential of hydrocarbons, HC

· x13: relative pollution potential of oxides of nitrogen, NOx

· x14: relative pollution potential of sulfur dioxide, SO2

· x15: percent relative humidity, annual average at 1 p.m.

These are socioeconomic and climatological variables measured at each of the n=60 metropolitan statistical areas in the United States. The 15 independent variables and mortality is measured from 1959 to 1961. We estimate the regression model based on the observations 1 to 40, and then we calculate forecasting RMSE based on observations 41 to 60 in Table 3.14.

The variables $x1$, $x3$, $x6$, $x7$, $x8$, $x9$, $x10$, $x12$, $x14$ and variables $x1$, $x2$, $x8$, $x9$, $x10$, $x14$ are selected by LTS-lasso with the cross-validation and with the efficient bootstrap information criterion, respectively. And, the variables $x1$, $x2$, $x3$, $x5$, $x6$, $x7$, $x8$, $x9$, $x10$, $x11$, $x13$, $x14$ and variables $x1$, $x2$, $x3$, $x4$, $x6$, $x7$, $x8$, $x9$, $x10$, $x12$, $x14$ are selected by LTS-Ela with the cross-validation and with the efficient bootstrap

Table 3.14: Real data analysis results in Part 1

	Elastic net	LTS-lasso	LTS-Ela
CV	47.93	47.80	46.12
AIC	49.85	48.27	47.35
BIC	49.85	48.27	47.35
EIC_{eff}	46.82	45.11	**45.02**

information criterion, respectively. The AIC and BIC tend to select all variables in the Elastic, LTS-lasso and LTS-Ela. From Table 3.14, it can be first seen that the LTS-Ela shows the outstanding performance compared with ordinary elastic net and LTS-lasso in the presence of outliers. We can also see that the LTS-Ela, especially, based on the efficient bootstrap information criterion shows the smallest RMSE. This implies that the proposed methods are also effective for the real world data analysis, especially contaminated dataset.

Part 2: Robust coordinate descent procedures

In part 2 of real world examples, we apply the proposed robust coordinate descent procedures to the crime data (Agresti and Finlay, 1997). The crime dataset consists of p=9 variables for $n = 51$ observations as follows,

· crime: violent crimes per 100,000 people

· sid: state id

· state: state name

· murder: murders per 1,000,000 people

· pctmetro: the percent of the population living in metropolitan areas

Table 3.15: Real data analysis results in Part 2

	Coordi.al	W.Coordi.al	T.Coordi.al	W.pre	T.pre
RMSE	126.33	**106.95**	114.58	114.13	107.72

· pctwhite: the percent of the population that is white

· pcths: percent of population with a high school education or above

· poverty: percent of population living under poverty line

· single: percent of population that are single parents

The variable "crime" is considered as a response variable and the variables "murder, pctmetro, pctwhite, pcths, poverty" and "single" are considered as predictor variables (i.e., p=6). The model is estimated by the lasso via the proposed robust coordinate descent procedures. Table 3.15 shows the forecasting RMSE by each algorithm. From Table 3.15, it can be seen that the proposed robust procedures outperform for forecasting accuracy compare with ordinary coordinate descent algorithm. The **W.coordinate descent algorithm**, especially, shows the best performance for the real data analysis.

Chapter 4

Lag weighted lasso for time series model

In this chapter, we introduce a novel L_1-type regularization method for time series model (Park and Sakaori, 2012a). In the real world, many time series occur, such as stock market index, monthly mortality and measurements of environmental factors. The response variable in the time series model is explained by a parametric function of the present and past values of predictor variables and past values of response variable. It implies that the lag length of the past variables is an important factor for time series modeling.

We propose a novel regularization method for time series model in line with the adaptive lasso. The adaptive lasso is able to identify the true model consistently and estimator is efficient by imposing the penalty which reflects the coefficient size of each variable. Although the adaptive lasso provides a useful tool for regression modeling, it is not suitable for time series model, since the adaptive lasso cannot reflect the

effect of lag length, which is a crucial factor in the time series modeling. We propose a lag weighted lasso which considers not only coefficient size but also the lag effects for estimation of the time series model. We observe that the proposed lag weighted lasso improves a forecasting accuracy by reflect the properties of time series.

The rest of this chapter is organized as follows. In Section 4.1, we briefly introduce the general time series models. We propose a lag weighted lasso with three types of weights in Section 4.2. Monte Carlo simulations are conducted to investigate the effectiveness of the proposed lag weighted lasso in Section 4.3. A real world example through cerebrovascular disease data is shown in Section 4.4.

4.1 Time series model

Autoregressive (AR) model

The most widely used and basic time series model is the autoregressive (AR) model. Consider the series $Y_t, Y_{t-1}, ..., Y_{t-q}$. The AR($q$) model is constructed by lagged variables of response variable and an error term,

$$\begin{aligned} Y_t &= \alpha + \sum_{l=0}^{q} \beta_l L^l Y_t + e_t \\ &= \alpha + \beta_1 Y_{t-1} + \cdots + \beta_q Y_{t-q} + e_t, \end{aligned} \tag{4.1}$$

where e_t is a white noise process having a zero mean and a constant variance σ^2, $\text{cov}(e_t, Y_{t-l}) = 0$ for all $l \neq 0$, and L represents the lag operator (i.e $L^0 Y_t = Y_t$, $L^1 Y_t = Y_{t-1}$). As shown in (4.1), the response variable Y_t is explained by only the past values of response variable in AR model.

Autoregressive Distributed Lag (ADL) model

When the present value Y_t cannot be fully explained by the past series Y_{t-l}, we consider additional explanatory variables $X_{j,t-l}$ $(j = 1, ..., p, l = 0, ..., q_j)$ in order to improve forecast accuracy. Autoregressive distributed lag (ADL) model is composed of lagged variables of response variable, and current and lagged variables of explanatory variables. The ADL$(q_0, q_1, q_2, ..., q_p)$ model is given by

$$Y_t = \alpha + \sum_{l=1}^{q_0} \beta_{0,l} L^l Y_t + \sum_{l=0}^{q_1} \beta_{1,l} L^l X_{1,t} + \ldots + \sum_{l=0}^{q_p} \beta_{p,l} L^l X_{p,t} + e_t, \tag{4.2}$$

where e_t is a white noise process having zero mean and constant variance σ^2. We can express (4.2) as follows:

$$Y_t = \alpha + \sum_{j=0}^{p} \sum_{l=0}^{q_j} \beta_{j,l} L^l X_{j,t} + e_t, \tag{4.3}$$

where we assume that $\beta_{0,0} = 0$, $X_{0,t} = Y_t$, and the following three assumptions hold:

1. $E(e_t | Y_{t-1}, Y_{t-2}, ..., X_{1,t}, X_{1,t-1}, ..., X_{p,t-1}, X_{p,t-2}, ...) = 0$.

2. $(Y_t, X_{1,t}, ..., X_{p,t})$ are stationary.

3. The correlation coefficients between $(Y_t, X_{1,t}, ..., X_{p,t})$ and $(Y_{t-l}, X_{1,t-l}, ..., X_{p,t-l})$ decline as l increase.

4.2 Lag weighted lasso for time series model

General time series models are composed of lagged variables, and the effects of the explanatory variables on the response variable decay to zero as time passes by (Ravines et al., 2006). In other words, even for variables having strong effect, the variable effect

reduces as increasing lag length as shown in the assumption 3 of ADL model. This implies that we should consider the lag effects in the time series modeling without seasonality.

In the adaptive lasso, the amount of shrinkage is controlled by $\hat{\boldsymbol{\beta}}$, and hence the coefficients of variable with large effect are shrunk slightly, whereas coefficients of variable with small effects are shrunk significantly. Although the adaptive lasso effectively perform estimation and variable selection for regression modeling by imposing different weights to each coefficient, it may not give proper and interpretable results for time series models with lagged variable, since its weight does not reflect the lag length.

Numerous studies on weight reflecting the lag effects in time series model have been conducted (Matsumoto and Szidarovszly 2010; Shrestha 2007; Tibshirani 2006). We consider the weight in Shrestha (2007) which is in line with our assumption *"The factor effects reduce as lag length increases"*. Shrestha (2007) assumed that the response variable is explained by the cumulative and extended lag effects of explanatory variables. And, they claimed that the t^{th} explanatory variable can be expressed under the following assumption:

$$X_t = \sum_{l=0}^{q} \omega_l X_{t-l}, \tag{4.4}$$

where

$$\omega_l = \alpha(1-\alpha)^l, \tag{4.5}$$

and α is a constant, $0 < \alpha < 1$, that is w_l is a geometrically decreasing weight up to the q^{th} lag. We use the weight $\omega_l = \alpha(1-\alpha)^l$ for reflecting the lag effects, and then

propose a lag weighted lasso with three type of weights for the stationary time series. The three type of weights are composed of two parts, one reflecting the coefficient size and the other one reflecting the lag effects,

- reflecting only the lag effects:

$$w_{j,l}^{(1)} = \frac{1}{[\alpha(1-\alpha)^l]^\gamma}. \tag{4.6}$$

- reflecting coefficient size and lag effects with γ on only the coefficient size:

$$w_{j,l}^{(2)} = \frac{1}{(|\hat{\beta}_{j,l}|)^\gamma} \frac{1}{\alpha(1-\alpha)^l}. \tag{4.7}$$

- reflecting coefficient size and lag effect with γ on both the coefficient size and lag effects:

$$w_{j,l}^{(3)} = \frac{1}{[|\hat{\beta}_{j,l}|\alpha(1-\alpha)^l]^\gamma}. \tag{4.8}$$

The proposed weights in (4.6), (4.7) and (4.8) control the amounts of shrinkage based on the coefficients size and lag length. The lag weighted lasso is a similar manner to the adaptive lasso but with a key difference. The lag weighted lasso has weights which reflect not only coefficients size but also the lag effects unlike the adaptive lasso, and thus the estimators of variable with small $\alpha(1-\alpha)^l$ and $\hat{\boldsymbol{\beta}}$ are considerably shrunk. In other words, coefficient of variable in distant past with small effect is estimated in small, or this variable is deleted from the model. If the time series includes seasonality, we should take a suitable difference of the series before applying the lag weighted lasso.

For ADL model, we first fit the AR model, and then consider additional explanatory variables in order to explain the part of future y_t which cannot be explained by

lagged variables of response variable. To reflect the ADL modeling procedure to the lag weighted lasso, we suggest not only the general weight (i.e., Type 2) imposing both lagged response and explanatory variables, but also Type 1 weight imposing only explanatory variables as follow

- Type 1: the part of lag effect in weights is on only the explanatory variables,
- Type 2: the part of lag effect in weights is on both the explanatory variables and lagged variables of response variable.

It implies that we select additional explanatory variables more strictly than lagged response variable by using the Type 1 weight. Estimates of the lag weighted lasso with Type 1 and Type 2 weights are given in Table 4.1 and 4.2 respectively. We can use either the least square estimator or ridge estimator as $\hat{\boldsymbol{\beta}}$, and choose regularization parameters from $\gamma > 0$, $\lambda > 0$ and $0 < \alpha < 1$. From Tables 4.1 and 4.2, it can be seen that the lag weighted lasso assigns different weights to different coefficient and

Table 4.1: Estimates of the lag weighted lasso with Type 1 weight

weight	lag weighted lasso estimates
$w^{(1)}$	$\hat{\boldsymbol{\beta}}^*_{w^{(1)}} = \arg\min_{\boldsymbol{\beta}} \|y - \sum_{j=0}^{p}\sum_{l=0}^{q_j}\beta_{j,l}L^l x_{j,t}\|^2 + \lambda\sum_{j=0}^{p}\sum_{l=0}^{q_j}\hat{w}^{(1)}_{j,l}\|\beta_{j,l}\|$ $= \arg\min_{\boldsymbol{\beta}} \|y - \sum_{j=0}^{p}\sum_{l=0}^{q_j}\beta_{j,l}L^l x_{j,t}\|^2$ $+\lambda(\sum_{l=1}^{q_0}\frac{\|\beta_{0,l}\|}{\|\hat{\beta}_{0,l}\|^\gamma} + \sum_{j=1}^{p}\sum_{l=0}^{q_j}\frac{\|\beta_{j,l}\|}{[\alpha(1-\alpha)^l]^\gamma})$
$w^{(2)}$	$\hat{\boldsymbol{\beta}}^*_{w^{(2)}} = \arg\min_{\boldsymbol{\beta}} \|y - \sum_{j=0}^{p}\sum_{l=0}^{q_j}\beta_{j,l}L^l x_{j,t}\|^2 + \lambda\sum_{j=0}^{p}\sum_{l=0}^{q_j}\hat{w}^{(2)}_{j,l}\|\beta_{j,l}\|$ $= \arg\min_{\boldsymbol{\beta}} \|y - \sum_{j=0}^{p}\sum_{l=0}^{q_j}\beta_{j,l}L^l x_{j,t}\|^2$ $+\lambda(\sum_{l=1}^{q_0}\frac{\|\beta_{0,l}\|}{\|\hat{\beta}_{0,l}\|^\gamma} + \sum_{j=1}^{p}\sum_{l=0}^{q_j}\frac{\|\beta_{j,l}\|}{(\|\hat{\beta}_{j,l}\|)^\gamma[\alpha(1-\alpha)^l]})$
$w^{(3)}$	$\hat{\boldsymbol{\beta}}^*_{w^{(3)}} = \arg\min_{\boldsymbol{\beta}} \|y - \sum_{j=0}^{p}\sum_{l=0}^{q_j}\beta_{j,l}L^l x_{j,t}\|^2 + \lambda\sum_{j=0}^{p}\sum_{l=0}^{q_j}\hat{w}^{(3)}_{j,l}\|\beta_{j,l}\|$ $= \arg\min_{\boldsymbol{\beta}} \|y - \sum_{j=0}^{p}\sum_{l=0}^{q_j}\beta_{j,l}L^l x_{j,t}\|^2$ $+\lambda(\sum_{l=1}^{q_0}\frac{\|\beta_{0,l}\|}{\|\hat{\beta}_{0,l}\|^\gamma} + \sum_{j=1}^{p}\sum_{l=0}^{q_j}\frac{\|\beta_{j,l}\|}{(\|\hat{\beta}_{j,l}\|[\alpha(1-\alpha)^l])^\gamma})$

Table 4.2: Estimates of the lag weighted lasso with Type 2 weight

weight	lag weighted lasso estimates
$w^{(1)}$	$\hat{\boldsymbol{\beta}}^*_{w^{(1)}} = \arg\min_{\boldsymbol{\beta}} \|y - \sum_{j=0}^{p}\sum_{l=0}^{q_j}\beta_{j,l}L^l x_{j,t}\|^2 + \lambda\sum_{j=0}^{p}\sum_{l=0}^{q_j}\hat{w}^{(1)}_{j,l}\|\beta_{j,l}\|$ $= \arg\min_{\boldsymbol{\beta}} \|y - \sum_{j=0}^{p}\sum_{l=0}^{q_j}\beta_{j,l}L^l x_{j,t}\|^2 + \lambda\sum_{j=0}^{p}\sum_{l=0}^{q_j}\frac{\|\beta_{j,l}\|}{[\alpha(1-\alpha)^l]^\gamma}$
$w^{(2)}$	$\hat{\boldsymbol{\beta}}^*_{w^{(2)}} = \arg\min_{\boldsymbol{\beta}} \|y - \sum_{j=0}^{p}\sum_{l=0}^{q_j}\beta_{j,l}L^l x_{j,t}\|^2 + \lambda\sum_{j=0}^{p}\sum_{l=0}^{q_j}\hat{w}^{(2)}_{j,l}\|\beta_{j,l}\|$ $= \arg\min_{\boldsymbol{\beta}} \|y - \sum_{j=0}^{p}\sum_{l=0}^{q_j}\beta_{j,l}L^l x_{j,t}\|^2 + \lambda\sum_{j=0}^{p}\sum_{l=0}^{q_j}\frac{\|\beta_{j,l}\|}{(\|\hat{\beta}_{j,l}\|)^\gamma[\alpha(1-\alpha)^l]}$
$w^{(3)}$	$\hat{\boldsymbol{\beta}}^*_{w^{(3)}} = \arg\min_{\boldsymbol{\beta}} \|y - \sum_{j=0}^{p}\sum_{l=0}^{q_j}\beta_{j,l}L^l x_{j,t}\|^2 + \lambda\sum_{j=0}^{p}\sum_{l=0}^{q_j}\hat{w}^{(3)}_{j,l}\|\beta_{j,l}\|$ $= \arg\min_{\boldsymbol{\beta}} \|y - \sum_{j=0}^{p}\sum_{l=0}^{q_j}\beta_{j,l}L^l x_{j,t}\|^2 + \lambda\sum_{j=0}^{p}\sum_{l=0}^{q_j}\frac{\|\beta_{j,l}\|}{(\|\hat{\beta}_{j,l}\|[\alpha(1-\alpha)^l])^\gamma}$

lag period. This implies that the lag weighted lasso reflects the properties of time series, and thus we can effectively perform the time series modeling.

4.3 Simulation studies

We examine, through Monte Carlo experiments the effectiveness of the proposed lag weighted lasso for time series model comparing with the lasso and adaptive lasso. We considered the following ADL(5,3,3) model:

$$Y_t = \alpha + \sum_{l=1}^{5}\beta_{0,l}Y_{t-l} + \sum_{l=0}^{3}\beta_{1,l}X_{1,t-l} + \sum_{l=0}^{3}\beta_{2,l}X_{2,t-l} + e_t. \tag{4.9}$$

For estimating the time series model, we used the LARS algorithm (Efron et al., 2004), and the least square estimator as $\hat{\boldsymbol{\beta}}$. The explanatory variables X_1 and X_2 are generated from the bivariate normal distribution $N_2(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, and e_t is a zero mean white noise process with a constant variance σ^2. The regularization parameters are chosen from $0 < \alpha < 1$ and $0 < \gamma \leq 5$ by 10-fold cross-validation.

As mentioned above, the lag weighted lasso is a suitable method when the variable effects decrease with increasing lag length. However, we consider several situations because this assumption is not always true.

Case 1: The variable effects decrease with increasing lag length.

Case 2: The variable effects are not related to the lag period.

For the two cases, we consider two models.

Model 1: Many small effects.

Model 2: A few large effects.

We generated 100 datasets for $n = 100$ and $\sigma = 1, 3, 5$ in each situation. Table 4.3 shows the true model settings for the simulation studies. In practice, the choice of $q_0, ..., q_p$ is a crucial matter. Pesaran (1999) introduced a method for choice the order $q_0, ..., q_p$ by AIC of all model based on the combination $q_0, ..., q_p$. However, it

Table 4.3: Simulation settings

	Y_{t-l}					$X_{1,t-l}$				$X_{2,t-l}$			
	$\beta_{0,1}$	$\beta_{0,2}$	$\beta_{0,3}$	$\beta_{0,4}$	$\beta_{0,5}$	$\beta_{1,0}$	$\beta_{1,1}$	$\beta_{1,2}$	$\beta_{1,3}$	$\beta_{2,0}$	$\beta_{2,1}$	$\beta_{2,2}$	$\beta_{2,3}$
Case 1	Model 1 : many small effects, $\sigma = 1, 3, 5$												
	0.3	−0.2	0.1	0	0	0.9	0.7	0.5	0	1	−0.7	0.5	0
	Model 2 : A few large effects, $\sigma = 1, 3, 5$												
	−0.5	−0.4	0	0	0	3	0	0	0	−3	2	0	0
Case 2	Model 1 : many small effects, $\sigma = 1, 3, 5$												
	0	0.1	−0.2	0	0.3	−0.4	−0.3	0	0.7	−0.7	0	0.3	1
	Model 2 : A few large effects, $\sigma = 1, 3, 5$												
	0	0.4	0	−0.6	0	0	0	2	0	0	−2	0	3

is improper for model with large number of predictor variables in the view point of the cost and time consuming. In general, the order of lagged variable y_{t-l} (i.e., q_0) is selected based on the autocorrelation function (ACF), partial autocorrelation function (PACF) and AIC based on the AR model constructed by only lagged variables of response variable. And then we consider the additional explanatory variables based on the granger causality test (Lee, 2007). The lag order of predictor variables, $q_1, ..., q_p$, are selected by consider a characteristic of data, explanatory power and principle of Parsimony. In this study, we fit the ADL model by using the ACF, PACF and granger causality test.

Figure 4.1 shows the relative prediction error (RPE)

$$\text{RPE} = E[(\hat{y} - \mathbf{x}^{\boldsymbol{T}}\boldsymbol{\beta})^2]/\sigma^2, \tag{4.10}$$

for each method (the horizontal line shows the RPE of the adaptive lasso). From the Figure 4.1, it can be seen that the lag weighted lasso outperforms both the lasso and adaptive lasso for forecast accuracy. Although the optimal weight in the lag weighted lasso is different in the true models, the lag weighted lasso with Type 1 $w^{(2)}$ shows the best performance in overall.

We also compare the accuracy of the true model selection. We compute the probability of the lasso solution path containing the true model in the 100 replications in Table 4.4. The lag weighted lasso shows superiority for true model selection in Case 1. For Case 2, the adaptive lasso shows better results than the lag weighted lasso. This result is comprehensible from common sense, since the lag weighted lasso is a suitable method for situation that the variable effects decrease with increasing lag length (i.e., Case 1).

Figure 4.1: RPE for each method.

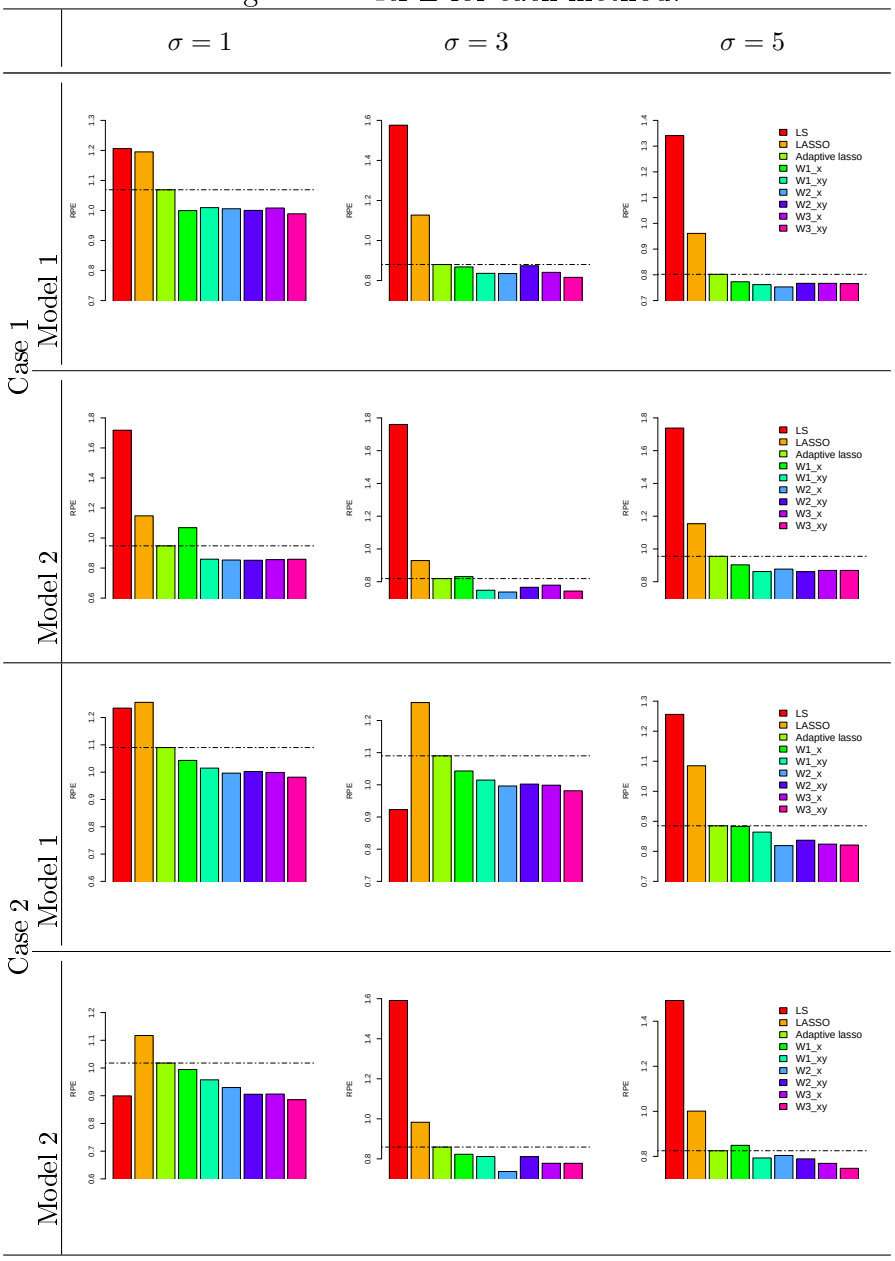


Table 4.4: Probability of containing the true model in solution path

	Case 1						Case 2					
	model 1			model 2			model 1			model 2		
	$\sigma=1$	$\sigma=3$	$\sigma=5$	$\sigma=1$	$\sigma=3$	$\sigma=5$	$\sigma=1$	$\sigma=3$	$\sigma=5$	$\sigma=1$	$\sigma=3$	$\sigma=5$
lasso	0.00	0.01	0.00	0.34	0.43	0.37	0.09	**0.01**	**0.10**	0.87	0.72	0.53
adaptive lasso	0.19	0.04	0.03	**0.90**	**0.82**	0.54	**0.23**	**0.01**	**0.01**	**1.00**	**0.89**	**0.63**
Type1 $w^{(1)}$	0.00	0.00	0.00	0.19	0.00	0.00	0.00	0.00	0.00	0.04	0.02	0.00
Type2 $w^{(1)}$	**0.35**	**0.33**	**0.24**	0.58	0.47	0.48	0.05	0.00	0.00	0.16	0.05	0.00
Type1 $w^{(2)}$	0.19	0.00	0.04	0.80	0.74	0.51	0.20	**0.01**	**0.01**	0.67	0.74	0.46
Type2 $w^{(2)}$	0.30	0.14	0.18	0.84	0.81	**0.65**	0.16	0.00	**0.01**	0.72	0.53	0.43
Type1 $w^{(3)}$	0.13	0.07	0.03	0.57	0.54	0.43	0.10	0.00	0.00	0.50	0.37	0.18
Type2 $w^{(3)}$	0.22	0.08	0.06	0.77	0.70	0.48	0.10	**0.01**	**0.01**	0.71	0.55	0.37

4.4 Real-world example: Cerebrovascular Mortality data

We illustrate the weighted lasso through the analysis of Cerebrovascular Mortality data (Park and Lee, 2009) to evaluate the practicality. The dataset consists of monthly mortality of cerebrovascular diseases and environmental factors from January 1996 to December 2005 in Seoul, Korea as given in Park and Lee (2009). The environmental factors as explanatory variables are air pollutants SO_2, O_3, NO_2, PM_{10}, temperature and humidity obtained from Korea' National Statistics Office as shown in Table 4.5.

This example is suitable for evaluating the lag weighted lasso, since the cerebrovascular disease mortality is affected by cumulative and extended lag effects of

Table 4.5: Environmental factors

variables			unit	sources
Air pollutants	SO_2	X_1	ppm	Korea National Statistical Office (Auto-measured data)
	O_3	X_2		
	NO_2	X_3		
	PM_{10}	X_4	$\mu g/m^3$	
Meteorology	Temperature	X_5	^oC	Korea National Statistical Office (Monthly average)
	Humidity	X_6	%	

the environmental factors. We forecast the monthly mortality in 2005 based on the data from 1996 to 2004 as training data. The model is composed from 1 to 12 lagged variables of mortality and from 0 to 5 lagged variables of environmental factors. Table 4.6 shows the forecast results for each method. From the variable selection results, we identified that the adaptive lasso had tendency that if one variable is selected, all of its lagged variable are also selected. It implies that the adaptive lasso cannot reflect the lag effect. Furthermore, the adaptive lasso did not select y_{t-1} unlike to the lag weighted lasso with Type 2 $w^{(3)}$. In the ADL model, the recent past variables of y_t explain a major part of future y_t, and thus the adaptive lasso might show not good performance for forecasting accuracy as shown in Table 4.6. In short, the lag weighted lasso with Type 2 $w^{(3)}$ outperforms both the lasso and the adaptive lasso in real data analysis.

Table 4.6: Forecast results for cerebrovascular disease mortality in 2005

	LS	lasso	adalasso	Type1$w^{(1)}$	Type2$w^{(1)}$	Type1$w^{(2)}$	Type2$w^{(2)}$	Type1$w^{(3)}$	Type2$w^{(3)}$
RPE	0.7630	0.5849	0.5541	0.6879	0.6270	0.4436	0.4377	0.4078	**0.3934**

In recently years, various studies about effects of environmental factor on mortality have been conducted worldwide. We expect that the lag weighted lasso may helpful to these studies.

Chapter 5

Symbolic candle chart-valued time series

This chapter introduces a new type of symbolic data, called a candle chart valued time series (CTS), and presents novel approaches for forecasting CTS (Park and Sakaori, 2012b).

Along with increasing data size and growth in the use of huge dataset, summarization and visualization of large dataset are increasingly important. To address this issue, symbolic data analysis (SDA) has been introduced by Bock and Diday (2000). The symbolic data analysis is able to effectively summarize and visualize huge databases by represent as formats of lists, intervals and distributions not single values. Furthermore, SDA takes account of the information that cannot be represented within the classical data model.

There is currently much discussion about the interval-valued data analysis which focuses on the interval of variable not single value. Billard and Diday (2000) intro-

duced linear regression modeling approaches for symbolic interval-valued data based on the mid-point of intervals. Lima Neto and De Carvalho (2008) proposed a new approach for the interval-valued data based on the mid-point and half-range of intervals. Lima Neto et al. (2006) also proposed new sum of squares based on correlation between the mid-point and half-range of intervals.

Maia et al. (2008) introduced approaches to interval-valued time series based on autoregressive (AR) model, autoregressive integrated moving average (ARIMA) model, artificial neural network (ANN) model and hybrid ARIMA and ANN model. Arroyo et al. (2009) introduced various forecasting methods for a histogram-valued time series (HTS).

We introduce a new type of symbolic data, a candle chart-valued time series (CTS). The candle chart consisting of open, close, highest, and lowest stock indices (or prices) has been widely used to empirically forecast stock index direction. By using the historical stock indices, we have been establishing a trading strategy. The dataset of candle chart, however, may become extremely large, since the candle chart is constructed by four indices at time t. We consider the candle chart consisting of the four indices as a one symbolic data, called a candle chart valued time series, and then propose the forecasting approaches for future stock index direction based on the CTS. We observe that the information about the stock indices can be effectively summarized and visualized by using the CTS, and forecasting accuracy of stock index direction is improved by using the our approaches.

The rest of this chapter is organized as follows. In Section 5.1, we introduce a typical symbolic data. Section 5.2 presents approaches to the symbolic interval-valued data which are basis of our study. We introduce a candle chart-valued time series, and

propose the forecasting approaches for the CTS in Section 5.3. A real world example through the stock indices of five major Asian countries is presented in Section 5.4.

5.1 Symbolic data

Symbolic data was introduced in order to summarize and visualize information from a large dataset. In the viewpoint of the symbolic data analysis, a huge classical dataset can be organized as a manageable symbolic data by represent as formats of list, distributions, interval, etc. We introduce two typical symbolic data, called interval-valued data and histogram-valued data.

Table 5.1: Classical data: pulse rate

	01Jul12							02Jul12							⋯
	am.0	am.1	⋯	pm.1	pm.2	⋯	pm.11	am.0	am.1	⋯	pm.1	pm.2	⋯	pm.11	⋯
patient 1	1	95	⋯	105	90	⋯	89	1	95	⋯	100	90	⋯	85	⋯
patient 2	2	85	⋯	110	96	⋯	85	1	85	⋯	110	100	⋯	99	⋯
patient 3	4	97	⋯	98	97	⋯	98	1	102	⋯	105	105	⋯	84	⋯
⋮				⋮							⋮				⋮
patient 98	90	95	⋯	103	100	⋯	87	1	83	⋯	89	90	⋯	109	⋯
patient 99	89	98	⋯	112	110	⋯	81	1	97	⋯	99	80	⋯	116	⋯
patient 100	85	95	⋯	101	89	⋯	100	1	101	⋯	103	95	⋯	96	⋯

Suppose we have the dataset consisting of pulse rate measured hourly as shown in Table 5.1. The dataset represented by classical data format may become very huge, and thus it is difficult to effectively figure out information from the huge dataset. In the view point of the symbolic data analysis, however, the pulse rate dataset can

Table 5.2: Interval-valued data: pulse rate

	01Jul12	02Jul12	$\cdots$
patient 1	[85,119]	[88,119]	$\cdots$
patient 2	[80,116]	[85,120]	$\cdots$
patient 3	[81,120]	[83,115]	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
patient 98	[83,113]	[80,121]	$\cdots$
patient 99	[89,109]	[81,120]	$\cdots$
patient 100	[80,123]	[83,116]	$\cdots$

be organized as a daily interval (i.e., [minimum, maximum]) as shown in Table 5.2. These data is an interval-valued symbolic data. This implies that the huge dataset can be effectively expressed by interval-valued data, and thus we can summarize and visualize some information from the huge dataset.

The classical data in Table 5.1 can be also expressed as a format of the histogram. Let consider the patient 1's data in 01Jul12 as follow,

	01Jul12						
	am.0	am.1	$\cdots$	pm.0	pm.1	$\cdots$	pm.12
patient 1	1	95	$\cdots$	105	90	$\cdots$	89

In the viewpoint of the symbolic data analysis, the classical data can be organized by follows,

$$\mathbf{y}_i = \{p_{i1}[a_{i1}, b_{i1}], ..., p_{is_i}[a_{is_i}, b_{is_i}]\}, \tag{5.1}$$

where p_{is_i} is the relative frequency for the sub-interval $[a_{is_i}, b_{is_i}]$, $i = 1, ..., n$, that is, the observed histogram takes values on s_i interval for i^{th} observation (Diday and

Nori, 2008). This is a histogram-valued data.

As shown above, the huge dataset can be summarized by the interval-valued data and histogram-valued data. This implies that we can effectively organize and visualize the large dataset by using the symbolic data analysis.

In order to analyze the information summarized by the symbolic data, numerous studies have been conducted by using statistical models (e.g., regression model, time series model, etc). For details on the symbolic data analysis, see Diday and Nori (2008).

5.2 Approaches for symbolic interval-valued data

We briefly introduce approaches for symbolic interval-valued data based on mid-point and half-range (Billard and Diday, 2000; Lima Neto and De Carvalho, 2006, 2008), which are basis of our study.

5.2.1 Centre and Range method

Centre and Range method (CRM method), proposed by Lima Neto and De Carvalho (2008), considers information about mid-point and half-range of interval between upper bound with lower bound of variable on the linear regression model. The CRM method is constructed by y^c and y^r, x_j^c and x_j^r $(j = 1, ..., p)$ as the mid-point and half-range of the response variables y and predictor variable x_j, respectively. The linear regression model based on the CRM method consists of two vectors, $\mathbf{w}_i = (\mathbf{x}_i^c, y_i^c)$

and $\mathbf{r}_i = (\mathbf{x}_i^r, y_i^r)$, with $\mathbf{x}_i^c = (x_{i1}^c, ..., x_{ip}^c)$ and $\mathbf{x}_i^r = (x_{i1}^r, ..., x_{ip}^r)$ where

$$y_i^c = \frac{y_{L_i} + y_{U_i}}{2}, \quad x_{ij}^c = \frac{x_{L_{ij}} + x_{U_{ij}}}{2}, \tag{5.2}$$
$$y_i^r = \frac{y_{L_i} - y_{U_i}}{2}, \quad x_{ij}^r = \frac{x_{L_{ij}} - x_{U_{ij}}}{2},$$

where y_L and y_U, and x_L and x_U are lower bound and upper bound of y and x, respectively.

The mid-points (y_i^c) and half-range (y_i^r) of response variable are explained by the mid-point (x_{ij}^c) and the half-range (x_{ij}^r) of predictor variables respectively,

$$y_i^c = \beta_0^c + \beta_1^c x_{i1}^c + \cdots + \beta_p^c x_{ip}^c + \varepsilon_i^c, \tag{5.3}$$

$$y_i^r = \beta_0^r + \beta_1^r x_{i1}^r + \cdots + \beta_p^r x_{ip}^r + \varepsilon_i^r, \tag{5.4}$$

and thus the sum of squares of the CRM method is given by,

$$S_{CRM} = \sum_{i=1}^{n} (\varepsilon_i^c)^2 + (\varepsilon_i^r)^2. \tag{5.5}$$

In the CRM method, we assume that the mid-point and half-range of interval are independent, and thus minimizing the sum of square is equivalent to fitting the two independent regression models of the mid-point and half-range of interval, respectively (Lima Neto and De Carvalho, 2008). We estimate $\hat{\boldsymbol{\beta}}^{\boldsymbol{c}} = (\hat{\beta}_0^c, \hat{\beta}_1^c, ..., \hat{\beta}_p^c)$ and $\hat{\boldsymbol{\beta}}^{\boldsymbol{r}} = (\hat{\beta}_0^r, \hat{\beta}_1^r, ..., \hat{\beta}_p^r)$ by minimizing (5.5).

5.2.2 NCRM1 and NCRM2 method

Lima Neto et al. (2006) proposed new sum of squares and new linear regression methods for interval-valued data. They considered a correlation between mid-point and half-range on the regression model with response variables y^c and y^r, and predictor variables x_j^c and x_j^r $(j = 1, ..., p)$.

NCRM1 method

In the NCRM1 method, the two regression models for mid-point and half-range have same regression coefficients,

$$y_i^c = \beta_0 + \beta_1 x_{i1}^c + \cdots + \beta_p x_{ip}^c + \varepsilon_i^c, \tag{5.6}$$

$$y_i^r = \beta_0 + \beta_1 x_{i1}^r + \cdots + \beta_p x_{ip}^r + \varepsilon_i^r. \tag{5.7}$$

The sum of squares of the NCRM1 method is given by,

$$S_{NCRM1} = \sum_{i=1}^{n} [(\varepsilon_i^c + \varepsilon_i^r)^2]. \tag{5.8}$$

We estimate $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)$ by minimizing (5.8), and the lower bound (y_L) and upper bound (y_U) are predicted as follows,

$$\hat{y_L} = \hat{y^c} - \hat{y^r} \text{ and } \hat{y_U} = \hat{y^c} + \hat{y^r}, \tag{5.9}$$

where $\hat{y^c} = \mathbf{x}^c \hat{\boldsymbol{\beta}}$ and $\hat{y^r} = \mathbf{x}^r \hat{\boldsymbol{\beta}}$.

NCRM2 method

The NCRM2 method is similar to the NCRM1 method, but has different regression coefficients of y_i^c and y_i^r,

$$y_i^c = \beta_0 + \beta_1^c x_{i1}^c + \cdots + \beta_p^c x_{ip}^c + \varepsilon_i^c, \tag{5.10}$$

$$y_i^r = \beta_0 + \beta_1^r x_{i1}^r + \cdots + \beta_p^r x_{ip}^r + \varepsilon_i^r. \tag{5.11}$$

The sum of squares of the NCRM2 method is given by

$$S_{NCRM2} = \sum_{i=1}^{n} [(\varepsilon_i^c + \varepsilon_i^r)^2]. \tag{5.12}$$

In the NCRM2 method, we estimate $\hat{\boldsymbol{\beta}}^{\mathbf{c}} = (\hat{\beta}_0, \hat{\beta}_1^c, ..., \hat{\beta}_p^c)$ and $\hat{\boldsymbol{\beta}}^{\mathbf{r}} = (\hat{\beta}_0, \hat{\beta}_1^r, ..., \hat{\beta}_p^r)$ by minimizing (5.12), and the lower bound (y_L) and upper bound (y_U) are predicted like the NCRM1 method.

5.3 Approach for symbolic candle chart-valued time series (CTS)

The candle chart consisting of open, close, highest, and lowest stock indices (or prices) has been widely used to empirically forecast the direction of future stock index. By using the historical stock indices, we have been establishing trading strategies. We introduce a new symbolic data, candle chart-valued time series (CTS) aggregated by the four indices time series. And we propose forecasting approaches for CTS extending the approaches for interval-valued data.

We first briefly introduce the method for symbolic interval-valued time series based on the mid-point and half-range series (Maia et al., 2008). In this method, two time series are considered: mid-point y_t^c and half-range y_t^r of interval of time series,

$$y_t^c = \frac{y_{L_t} + y_{U_t}}{2}, \;\; y_t^r = \frac{y_{L_t} - y_{U_t}}{2} \;\; (t = 1, 2..., T). \tag{5.13}$$

To forecast the interval-valued time series, Maia et al. (2008) applied the AR, ARIMA, ANN, and hybrid models to the mid-point y^c and half-range y^r, respectively. In their study, the hybrid with the ARIMA and ANN model showed the best performance for forecasting interval-valued time series in overall.

5.3.1 Time series model for forecasting CTS

We consider the candle chart consisting of four indices, open, close, highest and lowest as one symbolic variable at time t, called a candle chart-valued time series (CTS), in the viewpoint of the symbolic data analysis. By using the CTS, we can effectively summarize and visualize information about stock indices. In order to forecast the

CTS, we consider the following two mid-points ($y_t^{oc_m}$, $y_t^{hl_m}$) and two half-ranges ($y_t^{oc_r}$, $y_t^{hl_r}$) of intervals between the open (y_t^o) and close (y_t^c) indices, and between the highest (y_t^h) and lowest (y_t^l) indices consisting of candle chart respectively, that is,

$$\begin{aligned} y_t^{oc_m} &= \frac{y_t^c + y_t^o}{2}, \;\; y_t^{oc_r} = \frac{y_t^c - y_t^o}{2}, \\ y_t^{hl_m} &= \frac{y_t^h + y_t^l}{2}, \;\; y_t^{hl_r} = \frac{y_t^h - y_t^l}{2}, \end{aligned} \tag{5.14}$$

where $\hat{y}_t^{oc_m} \geq 0, \hat{y}_t^{hl_m} \geq 0, \hat{y}_t^{hl_r} \geq 0, -\infty < \hat{y}_t^{oc_r} < \infty$ and $\hat{y}_t^l \leq \hat{y}_t^o, \hat{y}_t^c \leq \hat{y}_t^h$.

We consider the hybrid ARIMA and ANN model, which showed the outstanding performance for interval-valued time series (Hansen and Nelson, 2003) to forecast CTS. To explain volatility clustering, we also consider the ARIMA-ARCH model, which is widely used for financial time series having volatility clustering.

Hybrid model

The hybrid model, proposed by Zhang (2003), is composed of linear component and nonlinear component,

$$y_t = L_t + N_t, \tag{5.15}$$

where y_t is a current value of the time series at time t, L_t and N_t denote the linear and nonlinear components, respectively.

The hybrid model is composed of two steps. In the first step, we apply the ARIMA model for the linear component L_t, and then apply the ANN model to the residuals of the ARIMA model,

$$n_t = y_t - \hat{L}_t, \tag{5.16}$$

to capture the nonlinear relation of the series by using p input nodes as follows,

$$n_t = f(n_{t-1}, n_{t-2}, ..., n_{t-p}) + \varepsilon_t. \tag{5.17}$$

Thus, the forecasted time series $\hat{y}_t$ is constructed by

$$\hat{y}_t = \hat{L}_t + \hat{N}_t. \tag{5.18}$$

The hybrid model showed the superiority for forecasting time series and interval-valued time series in various fields of literature (Hansen and Nelson, 2003; Maia *et al.*, 2008). For further details on this method, see Zhang (2003).

ARIMA-ARCH model

Financial time series often have a volatility clustering which means that large changes in series tend to cluster together. Numerous studies on the financial time series have been progressed by using the autoregressive conditional heteroskedasticity (ARCH) model to explain the volatility clustering (Chen and Jayaparakash, 2005).

We consider the ARIMA-ARCH model for CTS. The ARIMA(p, d, q)-ARCH(s) model is given by

$$\phi_p(B)(1-B)^d y_t = \theta_q(B)\varepsilon_t + \eta_t, \tag{5.19}$$

$$\eta_t = \sigma_t e_t, \tag{5.20}$$

where $\phi_p(B) = (1 - \phi_1 B - \cdots \phi_p B^p)$ is a stationary autoregressive (AR) operator, $\theta_q(B) = (1 - \theta_1 B - \cdots \theta_q B^q)$ is an invertible moving average (MA) operator, η_t are *i.i.d.* random variables with mean zero and variance one, which is independent of past realizations η_{t-i}, and

$$\sigma_t = \alpha_0 + \sum_{i=1}^{s} \alpha_i \eta_{t-i}^2. \tag{5.21}$$

The differenced series $(1-B)^d y_t$ follow the general stationary ARMA (p, q) process (Wei, 2006).

5.3.2 Parameter constraint and estimation

We propose novel approaches to estimate CTS in the viewpoint of the symbolic data analysis. In order to modeling CTS, we consider a method based on the original four stock indices, open, close, highest and lowest indices, and method based on the mid-point and half-range of stock indices's interval.

The 4-indices method

We first introduce a 4-indices method consisting of original open ($\mathbf{y}^o$), close ($\mathbf{y}^c$), highest ($\mathbf{y}^h$) and lowest ($\mathbf{y}^l$) indices time series. In the 4-indices method, we apply the time series models to the open ($\mathbf{y}^o$), close ($\mathbf{y}^c$), highest ($\mathbf{y}^h$) and lowest ($\mathbf{y}^l$) indexes, respectively. The sum of squares for the 4-indexes method is given by

$$S_{4I} = \sum_{i=1}^{n}(\varepsilon_i^o)^2 + \sum_{i=1}^{n}(\varepsilon_i^c)^2 + \sum_{i=1}^{n}(\varepsilon_i^h)^2 + \sum_{i=1}^{n}(\varepsilon_i^l)^2. \tag{5.22}$$

The 4-indices method minimizing sum of squares S_{4I} is equivalent to fit the four independent time series models for the $\mathbf{y}^o$, $\mathbf{y}^c$, $\mathbf{y}^h$ and $\mathbf{y}^l$, respectively.

MMRR method

We also propose a MMRR method based on three types of sum of squares for modeling CTS that is similar to Lima et al. (2006)' methods for interval-valued data. The MMRR method considers the information about interval between the 4-indices in the viewpoint of the symbolic data analysis. In the MMRR method, we consider the two mid-point time series and two half-range time series in (5.14), and then propose

following three types of sum of squares,

$$S_1 = \sum_{t=1}^{n} (\varepsilon_t^{oc_m})^2 + \sum_{t=1}^{n} (\varepsilon_t^{oc_r})^2 + \sum_{t=1}^{n} (\varepsilon_t^{hl_m})^2 + \sum_{i=1}^{n} (\varepsilon_t^{hl_r})^2, \tag{5.23}$$

$$S_2 = \sum_{t=1}^{n} (\varepsilon_t^{oc_m} + \varepsilon_t^{oc_r})^2 + \sum_{t=1}^{n} (\varepsilon_t^{hl_m} + \varepsilon_t^{hl_r})^2, \tag{5.24}$$

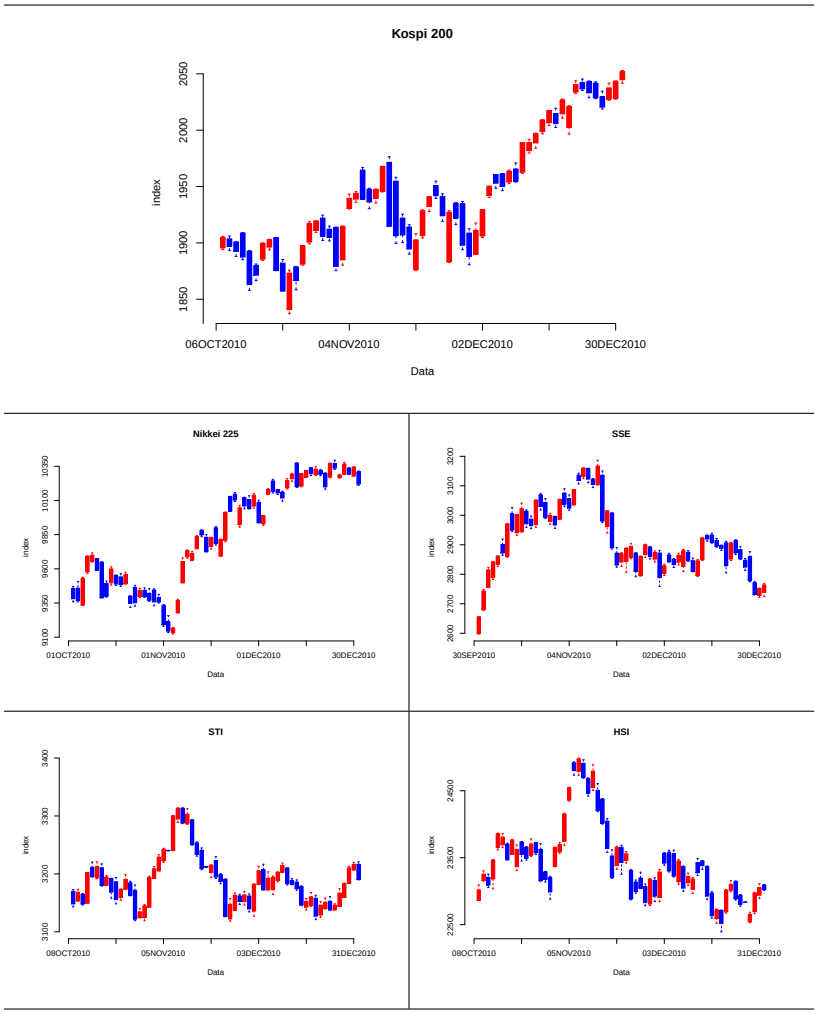
$$S_3 = \sum_{t=1}^{n} (\varepsilon_t^{oc_m} + \varepsilon_t^{oc_r} + \varepsilon_t^{hl_m} + \varepsilon_t^{hl_r})^2. \tag{5.25}$$

In the S_1, the four variables $y_t^{oc_m}$, $y_t^{oc_r}$, $y_t^{hl_m}$, and y_t^{hlr} are independently estimated by the hybrid ARIMA and ANN model or ARIMA-ARCH model. On the other hand, the sum of squares S_2 takes account of correlations between $y_t^{oc_m}$ and $y_t^{oc_r}$, and between $y_t^{hl_m}$ and $y_t^{hl_r}$. In this case, the intercepts of $y_t^{oc_m}$ and $y_t^{oc_r}$ in both the hybrid and ARIMA-ARCH models become the same because of the model identifiability. The intercept of $y_t^{hl_m}$ and $y_t^{hl_r}$ also become the same. The S_3 takes account of correlation between all four variables. In this case, the intercepts in all four models become the same. By considering the information which cannot be presented by classic statistical method, we can analysis the stock indices time series more implicitly. It implies that the proposed approaches based on the CTS is a useful tool for summarization and visualization of huge stock indices data, and thus it is helpful for traders in real stock market.

5.4 Applications: Stock market indices of five major Asian countries

We illustrate the proposed methods for the CTS through the analysis of the stock market indices of five major Asian countries (Japan, Korea, China, Singapore, Hong

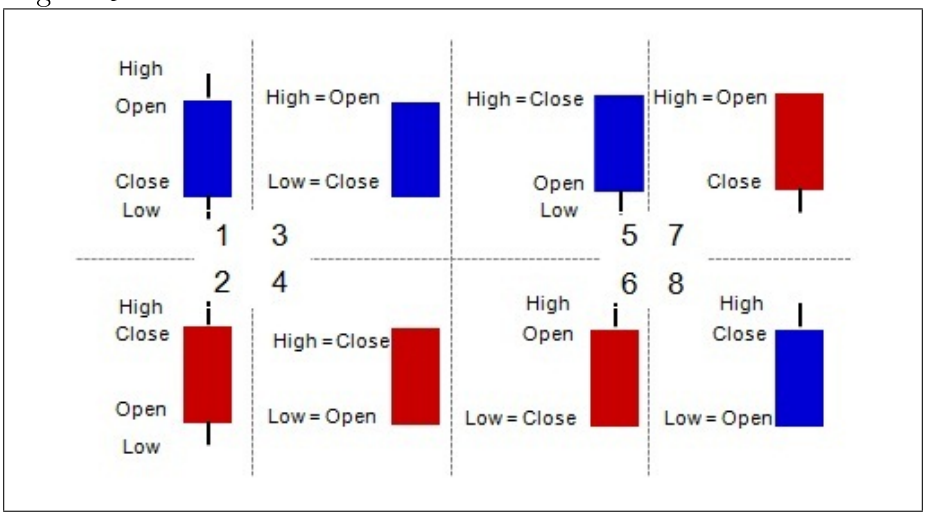
Figure 5.1: Part of the candlestick chart of the stock market index of five major Asia 5 countries



Kong). The databases are composed of the daily open, close, highest, and lowest indices of each of the five countries from January 2009 to April 2011. Figure 5.1 presents the candle chart of the stock indices of five major Asian countries (Japan: Nikkei 225, Korea: Kospi 200, China: SSE, Singapore: STI, Hong Kong: HSI). We estimate the model using the dataset from January 2009 to December 2010, and then forecast future stock indices from January 2011 to April 2011.

We apply the hybrid and ARIMA-ARCH models based on the Akaike informa-

Figure 5.2: Direction of stock index based on the candlestick chart



tion criterion (AIC) (Akaike, 1974). The proposed approaches are compared with a method using only the close index based on the root mean square error (RMSE) and a correctness of the forecasted direction of stock index.

In this study, the stock index direction is forecasted based on following candle chart forms which are widely used in real stock market (see Figure 5.2):

- The stock index will fall : $1, 3, 5$, and 7.
- The stock index will rise : $2, 4, 6$, and 8.

Table 5.3 shows the proportions of correctly forecasted future stock direction. From Table 5.3, it can be seen that the ARIMA-ARCH model with the MMRR method based on S_1 is superior for forecasting stock index direction in the viewpoint of forecast accuracy (i.e., average proportions) and stability. We also compared the root mean square error (RMSE) in Table 5.4. As shown in Table 5.4, the hybrid model shows the smaller RMSE than the ARIMA-ARCH model. However, some indices (i.e., Nikkei 225's y^l, SSE's y^h, y^l, and HSI's y^l) show a large RMSE in the

Table 5.3: Forecasting result of the stock index direction

			Kospi 200		Nikkei 225		SSE		STI		HSI		**Average**
			Up	Down	Up	Down	Up	Down	Up	Down	Up	Down	
ARIMA-ARCH	Close		50.1	50.0	46.4	57.9	31.6	55.6	51.7	55.6	53.6	59.5	51.2
	4INDEX		51.9	44.4	64.5	58.1	63.5	.	57.1	48.5	58.5	50.5	55.2
	MMRR	S_1	91.2	**84.8**	77.4	61.1	89.2	73.1	95.6	91.9	80.0	73.5	**81.8**
		S_2	74.3	73.3	73.5	59.4	89.2	73.1	81.4	75.0	73.8	22.2	69.5
		S_3	82.1	82.8	50.0	15.4	89.2	73.1	85.4	81.8	50.0	42.0	65.2
hybrid	Close		49.0	70.4	73.3	64.7	.	**100.0**	**100.0**	100.0	81.6	72.7	79.1
	4INDEX		96.3	80.8	**80.0**	**64.7**	68.4	39.5	**100.0**	**100.0**	81.6	75.8	78.7
	MMRR	S_1	85.7	75.0	77.4	61.1	80.0	77.8	**100.0**	**100.0**	82.4	73.5	81.3
		S_2	**100.0**	59.5	65.7	53.1	91.4	82.6	**100.0**	**100.0**	85.3	**78.1**	81.6
		S_3	**100.0**	42.3	36.4	20.7	**91.4**	82.6	66.1	**100.0**	**85.7**	**78.1**	70.3

hybrid model. In short, the ARIMA-ARCH model with the MMRR method based on S_2 outperforms others with regard to overall performance and stability as shown blue values in Table 5.4.

We observed through analysis of the stock market indices of five major Asian countries that the proposed approaches outperform for forecasting stock index. Furthermore, the proposed approaches are useful tools for not only specialist but also non-specialist on stock market since it has advantages on summarization and visualization of information about huge stock indices not a theoretical approach.

Table 5.4: Root mean square error of the forecasted stock indices

			Kospi 200				Nikkei 225				SSE				STI				HSI			
			y^o	y^c	y^h	y^l	y^o	y^c	y^h	y^l	y^o	y^c	y^h	y^l	y^o	y^c	y^h	y^l	y^o	y^c	y^h	y^l
	4INDEX		2.49	2.42	2.16	2.79	18.91	23.84	23.84	38.62	4.27	11.70	13.79	9.34	2.88	2.90	3.04	3.33	26.05	28.21	48.57	35.52
ARIMA		S_1	1.97	2.11	3.41	3.18	10.62	17.53	26.93	31.45	2.42	2.18	13.53	8.03	1.17	2.00	7.83	4.03	15.71	18.40	52.07	56.91
ARCH	MMRR	S_2	0.85	0.82	1.75	1.34	6.42	11.54	26.40	13.72	2.42	2.18	10.99	7.32	0.62	2.37	5.78	4.85	8.07	10.36	30.65	17.80
		S_3	2.15	2.07	7.55	7.39	34.79	39.92	236.52	123.48	1.97	1.94	7.45	6.43	1.36	1.67	3.61	3.51	14.47	14.33	53.38	28.12
	4INDEX		1.44	**0.33**	5.08	4.94	**0.42**	**0.40**	**1.27**	96.19	**0.18**	13.31	92.60	73.07	**0.31**	**0.31**	**1.15**	**1.17**	**0.35**	**0.33**	**1.20**	83.14
hybrid		S_1	**0.34**	**0.33**	**0.35**	**0.34**	8.90	18.90	26.83	22.08	**0.18**	0.20	**0.17**	**1.26**	0.34	0.34	5.64	4.47	12.74	12.65	32.49	30.09
	MMRR	S_2	**0.34**	**0.33**	1.40	**0.34**	8.29	8.33	5.04	4.91	**0.18**	**0.20**	1.29	1.95	0.34	0.34	1.25	1.27	3.01	3.16	12.01	**8.42**
		S_3	0.34	0.33	1.40	0.34	29.23	34.45	212.76	110.38	0.18	**0.20**	2.53	6.69	0.34	0.34	0.32	1.27	7.29	7.34	37.32	22.81

Chapter 6

Summary and concluding remarks

In this chapter, the findings of the present thesis are summarized, and some ideas for future study are listed.

We have mainly discussed the robust regression modeling via L_1 regularization in chapter 3. In order to robust regression modeling, we have first proposed the robust L_1-type regularization, called a least trimmed squares elastic net. After the replacement of the least squares loss function with the least trimmed squares loss function, the proposed LTS-Ela performed well variable selection and estimation, even in the presence of outliers. We have also introduced a method for choosing an optimal set of the regularization parameters and tuning constant by using the efficient bootstrap information criterion. The simulation results showed that the proposed LTS-Ela based on the efficient bootstrap information criterion is a useful tool for robust sparse regression modeling in the viewpoint of sparsity and forecasting accuracy. In the real world example through the McDonald and Schwing dataset, the proposed robust sparse regression modeling strategy also showed the outstanding

performance.

We have derived an information criterion ($\mathrm{GIC}_{\mathrm{R.la}}$) for evaluating the robust L_1-type regularized regression models in line with the generalized information criteria. In practice, an information criterion for lasso-type approaches cannot be derived, since the influence function of the lasso-type approaches cannot be calculated due to the L_1-type penalty. To derive the influence function which plays a key role in an information criterion, we used the local quadratic approximation of the L_1-type penalty terms, and then derived an information criterion. From the Monte Carlo experiments, it can be seen that the proposed modeling procedure based on the $\mathrm{GIC}_{\mathrm{R.la}}$ is effective for choosing the tuning parameters.

We have also proposed the robust coordinate descent procedures via robust inner product and pre-treatment technique for the robust L_1-type regularized regression modeling. In order to outlier-resistant procedure, we use the Winsorization and trimming techniques based on the robust Mahalanobis distance. Monte Carlo simulations and a real data analysis were used to investigate the efficiency of the proposed robust procedures. It was observed that the proposed robust coordinate descent procedures are stable and efficient in the viewpoint of forecasting accuracy and sparsity.

We have considered the robust model evaluation for choosing the tuning parameters robustly. Although the efficient bootstrap information criterion showed effectiveness for robust sparse regression modeling, it may produce biased results, since the bootstrap information criterion may be obtained from the contaminated bootstrap sample in the presence of outliers due to the randomly drawn bootstrap technique. To overcome the problem, we have proposed the robust bootstrap information criterion via Winsorizing technique in line with the efficient bootstrap information criterion.

We observed through Monte Carlo experiments that the proposed robust efficient bootstrap information criterion is more efficient and stable against outliers than the existing one.

The second topic of the present thesis is the lag weighted lasso for time series model as shown in chapter 4. In order to consider the properties of time series data, we have proposed the lag weighted lasso based on weights which reflect not only coefficients size but also lag effects, unlike the adaptive lasso. By using these weights, estimators of variables in the distant past and with small effects were considerably shrunk. This implies that the lag weighted lasso can reflect the properties of time series, and thus we can effectively perform time series modeling. Our simulation studies and real world data analysis through the cerebrovascular mortality data conducted herein showed that the lag weighted lasso outperforms the lasso and adaptive lasso for time series modeling.

The final topic of this thesis is the new type of symbolic data as shown in chapter 5. We have introduced the candle chart composed with open, close, highest and lowest stock indices as a new symbolic data, called a candle chart-valued time series (CTS). To forecast the CTS, we proposed novel approaches in the viewpoint of the symbolic data analysis. The proposed approaches were illustrated through the analysis of the stock market indices of five major Asian countries (Japan, Korea, China, Singapore, Hong Kong). We observed that the proposed approaches for CTS provide a useful tool for forecasting future stock indices.

For the future studies,

- The present thesis can be extended to robust non-linear regression modeling.
- To improve the time series modeling procedure, the group lasso version of the lag weighted lasso can be considered, because both lag selection and variable selection should be performed in the time series modeling, especially in the ADL model. Also, the algorithm for group lasso can be considered via the coordinate descent procedure.
- Furthermore, further work remains to be done for constructing a model for high dimensional data analysis (e.g., genomic data analysis).

Bibliography

[1] Agresti, A. and Finlay, B. (1997). *Statistical Methods for the Social Sciences.* Prentice Hall, New Jersey.

[2] Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle.* in Second International Symposium on Information Theory, eds. B. N. Petrov and F. Csaki, Budapest: Akademiai Kiado, pp. 267-281.

[3] Arroyo, J., Gonzalez-Rivera, G., and Maté, C. (2009). Forecasting with interval and histogram data. Some financial applications. *Congress of the European Economic Association and the Econometric Society European meeting (EEA-ESEM).*

[4] Billard, L., and Diday, E. (2000). Regression analysis for interval-valued data. *Data Analysis, Classification and Related Methods. Springer*, 369-374.

[5] Bock, H-H., and Diday, E. (2000). *Analysis of Symbolic Data: Explanatory Methods for Extracting Statistical Information from Complex Data.* Springer, Berlin.

[6] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24, 2350-2383.

[7] Chen, K., C., and Jayaprakash, B., Y. (2005). Conditional probability as a measure of volatility clustering in financial time series. *Quantitative Finance Papers*.

[8] Chen, L.A., Welsh, A.H. and Chan, W. (2001). Estimators for the linear regression model based on Winsorized observations. *Statistica Sinica*, 11, 147-172.

[9] Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Journal of the American Statistical Association*, 31, 377-403.

[10] Croux, C., and Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95, 206-226.

[11] Diday, E., and Nori, M. (2008). *Symbolic Data Analysis and the SODAS Software*. Wiley.

[12] Dodge, Y. and Jurecková, J. (1997). Adaptive choice of trimming proportion in trimmed least-squares estimation. *Statistics & Probability Letters*, 33, 167-176.

[13] Donoho, D.L. and Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90, 1200-1224.

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

[15] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360.

[16] Friedman, J., Hastie, T., Hofling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, 1, 302-332.

[17] Hastie, T., Tibshirani, R. and Friedman, J. (2008). *The Elements of Statistical Learning*. Springer, New York.

[18] Hansen, J., and Nelson, R. (2003). Time-series analysis with neural networks and ARIMA-neural network hybrids. *Journal of Experimental and Theoretical Artificial Intelligence*, 15, 315-330.

[19] Hesterberg, T., Choi, N.H., Meier, L. and Fraley, C. (2008). Least angle and penalized regression: A review. *Statistics Surveys*, 2, 61-93.

[20] Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: biased estimation for non orthogonal problems. *Techonometrics*, 12, 55-67.

[21] John, H. and David, M.R. (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics and Data Analysis*, 44, 625-638.

[22] Jung, K.M. (2009). Robust cross validation in ridge regression. *Journal of applied mathematics & informatics*, 27, 903-908.

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

[24] Konishi, S., and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, 83, 875-890.

[25] Konishi, S., and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. Springer, New York.

[26] Lambert-Lacroix, S. and Zwald, L. (2010). Robust regression through the Huberąś criterion and adaptive lasso penalty. *Electronic Journal of Statistics*, 5, 1015-1053.

[27] Lee, J.H. (2007). *Time Series Analysis and Application*. Freedom Academy, Korea.

[28] Lima Neto, E.A., De Carvalho, F.A.T., and Bezerra, L.X.T. (2006). Linear regression methods to predict interval-valued Data. *Proceedings of the Eighth Conference of the International Federation of Classification Societies*, 281-288.

[29] Lima Neto, E.A., and De Carvalho, F.A.T. (2008). Centre and range method for fitting a linear regression model on symbolic interval data. *Computational Statistics and Data Analysis*, 52, 1500-1515.

[30] Lima Neto, E.A. and De Carvalho, F.A.T. (2010). Constrained linear regression models for symbolic interval-valued variables. *Computational Statistics and Data Analysis*, 54, 333-347.

[31] Maia, A.L.S., De Carvalho, F.d.A.T. and Ludermir, T.B. (2008). Forecasting models for interval-valued time series. *International Journal of Forecasting*, 71, 3344-3352.

[32] Mateos, G. and Giannakis, G.B. (2012). Robust nonparametric regression via

sparsity control with application to load curve data cleansing. *IEEE Trans. Signal Process*, 60, 1571-1584.

[33] Matsumoto, A., Szidarovszly, F. (2010). Dynamic goodwin's business cycles with fixed and continuously distributed time delays, unpublished manuscript avaliable at *http://www2.tamacc.chuo-u.ac.jp/keizaiken/discussno115.pdf*.

[34] Park, H.W. (2012). Novel resampling methods for tuning parameter selection in robust sparse regression modeling. *To appear in Bulletin of Informatics and Cybernetics*.

[35] Park, H.W., Sakaori, F. (2012a). Lag weighted lasso for time series model. *To apprear in Computational statistics*.

[36] Park, H.W., Sakaori, F. (2012b). Forecasting symbolic candle chart-valued time series. *In preparation*.

[37] Park, H.W., Sakaori, F. and Konishi, S. (2012a). Robust sparse regression and tuning parameter selection via the efficient bootstrap information criteria. *To appear in Journal of Statistical Computation and Simulation*.

[38] Park, H.W., Sakaori, F. and Konishi, S. (2012b). Selection of tuning parameters in robust sparse regression modeling. *Proceedings of the COMPSTAT'2012*, pp 713-723.

[39] Park, H.W. and Lee, J.H. (2009). Effects of environmental factors on monthly cerebrovascular mortality in Seoul, Korea. *Journal of the Korean Data Analysis Society*, 11, 687-698.

[40] Pesaran, M.H. (1999). An autoregressive distributed lag modelling approach to cointegration analysis. *Cambridge University*, 134-150.

[41] Ravines, R.R., Schmidt, A.M. and Migon, H.S. (2006). Revisiting distributed lag models through a Bayesian perspective. *Applied Stochastic Models in Business and Industry*, 22, 193-210.

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

[43] Ronchetti, E., Staudte, R.G. (1994). Robust version of Mallows's C_p. *Journal of American Statistical Association*, 89, 550-559.

[44] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

[45] Rousseeuw P.J., Van Zomeren B.C. (1990). Unmasking multivariate outliers leverage points. *Journal of the American Statistical Association*, 85, 633-651.

[46] Shrestha, S.L. (2007). Time series modeling of respiratory hospital admissions and geometrically weighted distributed lag effects from ambient particulate air pollution within kathmandu valley, Nepal. *Environmental Modeling and Assessment*, 12, 239-251.

[47] Singh, K. (1998). Breakdown theory for bootstrap quantiles. *The Annals of Statistics*, 26, 1719-1732.

[48] Sohail, C. (2011). Goodness of fit and lasso variable selection in time series analysis. *phd thesis of University of Nottingham.*

[49] Srivastava, D.K., Pan, J.M., Sarkar, I. and Mudholkar, G.S. (2010). Robust Winsorized regression using bootstrap approach. *Communications in Statistics - Simulation and Computation*, 39, 45-67.

[50] Tharmaratnam, K. and Claeskens, G. (2010). A comparison of robust versions of the AIC based on M, S and MM-estimators. *Technical report*, Katholieke Universiteit Leuven.

[51] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58, 267-288.

[52] Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16, 385-395.

[53] Wang, H., Leng, C. (2008). A note on adaptive group lasso. *Computational Statistics and Data Analysis*, 52, 5277-5286.

[54] Wang, H., Li, G. and Jiang, G. (2007a). Robust regression shrinkage and consistent variable selection through the LAD-lasso. *Journal of Business and Economic Statistics*, 25, 347-355.

[55] Wang, H., Li, R. and Tsai, C.L. (2007b). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94, 553-568.

[56] Wei, W.W.S. (2006). *Time Series Analysis : Univariate and Multivariate Methods*. 2nd Ed. Addison Wesley, New York.

[57] Weisberg, S. (2005). *Applied Linear Regression*. Wiley, New York.

[58] Welsh, A.H. (1987). The trimmed mean in the linear model. *The Annals of Statistics*, 15, 20-36.

[59] Yale, C. and Forsythe, A. B. (1976). Winsorized regression. *Technometrics*, 18, 291-300.

[60] Yuan, M. and Lin, Y. (2007). On the nonnegative garrote estimator. *Journal of the Royal Statistical Society*, 69, 143-161.

[61] Zhang, G. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50, 159-175.

[62] Zhang, Z. G., Chan, S. C., Zhou, Y. and Hu, Y. (2009). Robust linear estimation using M-estimation and weighted L_1 regularization: model selection and recursive implementation. *Proceedings of the 2009 International Symposium on Circuits and Systems*, 1193-1196.

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

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