Adaptive L_1 Regularized Second-order Least Squares Method for Model Selection

by

Lin Xue

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfilment of the requirements of the degree of

Master of Science

Department of Statistics
University of Manitoba
Winnipeg

Abstract

The second-order least squares (SLS) method in regression model proposed by Wang (2003, 2004) is based on the first two conditional moments of the response variable given the observed predictor variables. Wang and Leblanc (2008) show that the SLS estimator (SLSE) is asymptotically more efficient than the ordinary least squares estimator (OLSE) if the third moment of the random error is nonzero. We apply the SLS method to variable selection problems and propose the adaptively weighted L_1 regularized SLSE (L_1-SLSE). The L_1-SLSE is robust against the shape of error distributions in variable selection problems. Finite sample simulation studies show that the L_1-SLSE is more efficient than L_1-OLSE in the case of asymmetric error distributions. A real data application with L_1-SLSE is presented to demonstrate the usage of this method.

Acknowledgments

I would like to express my sincere gratitude to my advisor Dr. Liqun Wang, for his guidance and constant supervision during my graduate study. Without his inspiration and encouragement this thesis would be impossible. I also thank the Department of Statistics and Dr. Liqun Wang for financial support through their research grants during my Master's program.

My sincere thanks also go to my co-advisor Dr. Depeng Jiang, for providing training opportunities in biostatistical consulting and for offering insightful comments in this research. I also appreciate the experience working with him at the Department of Community Health Sciences.

I also wish to thank my thesis committee members: Dr. James Fu and Dr. Mahmoud Torabi for their encouragement. This thesis is improved with the knowledge they shared and the suggestions they provided. I would like to extend my thanks to Dr. Alexandre Leblanc, Dr. Brad Johnson and Dr. Saman Muthukumarana for being really helpful during my master study.

Last but not least, I would like to thank my parents for their love, support and encouragement.

Contents

List of Tables

Chapter 1

Introduction

1.1 Motivation and Structure of the Thesis

The ordinary least squares estimator is the most efficient when the error distribution is normal. However, it is not necessarily so when the error distribution is non-normal (e.g., asymmetric gamma distribution). In addition, the distribution of data encountered in real world is often unknown to us. Hence the normally distributed assumption for the random error is not always appropriate. Wang and Leblanc (2008) proposed the second-order least squares method. By exploiting more information from higher order moments, the second-order least squares estimator is shown to be asymptotically more efficient than the ordinary least squares estimator for asymmetric error distributions, and the variance-covariance matrix is asymptotically equivalent for both estimators when the error distribution is symmetric. In addition, the second-order least squares method does not require the error distributions to be known. Hence it is robust against the shape of the error distributions.

For variable selection problems, most of the existing regularization methods are designed for either least squares or likelihood model fit, which requires the error

distribution to be either symmetric or known. For example, the ordinary least squares method is not robust against the shape of the error distributions. And the likelihood method is not applicable when the error distribution is unknown to us. These limitations motivate us to propose an estimation method for variable selection problems. The proposed estimator should be able to select those significant predictors and drop the redundant ones. At the same time, it is robust against the shape of error distributions. It should work well even for situations where the error distributions are unknown. Based on the considerations above, we apply the second-order least squares method (Wang and Leblanc (2008)) in variable selection problems and propose L_1-SLSE. We show that the L_1-SLSE is robust against the shape of the error distribution. And the error distribution is not required to be known. Specifically, when the distribution of random error is symmetric (e.g., normal), the L_1-SLSE and L_1-OLSE are asymptotically equivalent. When the random error is asymmetrically distributed, the L_1-SLSE is asymptotically more efficient than the L_1-OLSE. Finite sample size simulation studies show that the L_1-SLSE performs well in many situations.

The thesis is organized as follows. The least squares methods, subset selection and shrinkage methods are introduced in Chapter 1. The L_1-SLSE is defined in Chapter 2, followed by its propositions in Section 2.1. Since the existing algorithm cannot be applied to L_1-SLSE directly, an computational algorithm is presented in Section 2.2, where the selection of the regularization parameter is also discussed. In Chapter 3, we present the simulation results of different numerical examples, followed by a real data application. A summary of the thesis is given in Chapter 4. The appendix includes some model assumptions, lemmas, proofs and sample

simulation codes.

1.2 Linear Least Squares Estimation

Let $\{\boldsymbol{x}_i', y_i\}, i = 1, \cdots, n$, be a random sample from the linear regression model

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \epsilon, \tag{1.1}$$

where $y \in \mathcal{R}$ is a response variable, $\boldsymbol{x} \in \mathcal{R}^p$ is a vector of covariates, $\boldsymbol{\beta}$ is a p-dimensional regression coefficient vector, and ϵ is a random error with mean 0 and finite variance σ^2. Linear regression models as defined in (2.1) are commonly fitted with the least squares approach. The ordinary least squares method minimizes the residual sum of squares

$$\text{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2, \tag{1.2}$$

with respect to the unknown parameter $\boldsymbol{\beta}$. We can write RSS in matrix notation as

$$RSS(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta}), \tag{1.3}$$

where $\boldsymbol{y}_{n\times 1} = (y_1, y_2, \cdots, y_n)'$ and $X_{n\times p} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)'$. The first order condition is

$$\frac{\partial \text{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2X'(\boldsymbol{y} - X\boldsymbol{\beta}) = 0. \tag{1.4}$$

If the design matrix X is of full rank, we obtain the ordinary least squares estimator

$$\hat{\boldsymbol{\beta}}^{ols} = (X'X)^{-1}X'\boldsymbol{y}. \tag{1.5}$$

It is easy to show that the expectation of $\hat{\boldsymbol{\beta}}^{ols}$ is

$$\mathrm{E}(\hat{\boldsymbol{\beta}}^{ols}) = \beta, \tag{1.6}$$

and the variance of $\hat{\boldsymbol{\beta}}^{ols}$ is

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}^{ols}) = \sigma^2(X'X)^{-1}. \tag{1.7}$$

The OLSE is known to be the most efficient estimator when the error distribution is normal. However, it is not necessarily so when the error distribution is non-normal (e.g. asymmetric gamma distribution). In addition, the distributions of data encountered in real world are often unknown to us. Hence the assumption of normally distributed random error is not always appropriate. Wang and Leblanc (2008) proposed a SLS method for regression models. Specifically, they showed that the SLSE is asymptotically more efficient than the OLSE if the distribution of random error is asymmetric, and both estimators (OLSE and SLSE) have the same asymptotic variance-covariance matrix if the distribution of random error is symmetric. Hence the SLSE is robust against the shape of the error distribution. It turns out that the SLS method applies to both linear and nonlinear regression models, where the error distributions are not required to be known. The SLS method is introduced in the next section.

1.3 Second-order Least Squares Estimation

The SLS method was first introduced by Wang (2003) for estimation of parameters in general nonlinear regression models with Berkson type measurement errors in

predictor variables. Wang (2004) generalized the results to the nonlinear models with multivariate predictor variables, where the distribution of measurement error is of general parametric form (not necessarily normal) and the distribution of the random error is allowed to be nonparametric. Wang and Leblanc (2008) further developed the second-order nonlinear least squares estimator. They proposed a second-order least squares (SLS) estimation procedure for a class of general regression models. The resultant SLSE achieves efficiency gains over the OLSE when the distribution of random error is asymmetric, since it utilizes more information contained in the higher order moments of the data. Wang and Leblanc (2008) also showed that the variance-covariance matrix of OLSE and SLSE are asymptotically equivalent if the distribution of random error is symmetric.

Specifically, consider the general regression model

$$y = g(\boldsymbol{x}; \boldsymbol{\beta}) + \epsilon, \tag{1.8}$$

where $y \in \mathcal{R}$ is the response variable, $\boldsymbol{x} \in \mathcal{R}^k$ is the predictor variable, $\boldsymbol{\beta} \in \mathcal{R}^p$ is the regression parameter and ϵ is the random error satisfying $E(\epsilon|\boldsymbol{x}) = 0$ and $E(\epsilon^2|\boldsymbol{x}) = \sigma^2$. The regression function $g(\boldsymbol{x}; \boldsymbol{\beta})$ can be linear or nonlinear in either $\boldsymbol{x}$ or $\boldsymbol{\beta}$. It is easily seen that under this model

$$E(y|\boldsymbol{x}) = g(\boldsymbol{x}; \boldsymbol{\beta}),$$

and

$$E(y^2|\boldsymbol{x}) = g^2(\boldsymbol{x}; \boldsymbol{\beta}) + \sigma^2.$$

Define the parameter vector as $\boldsymbol{\gamma} = (\boldsymbol{\beta}', \sigma^2)'$ and the parameter space as $\Gamma = \mathcal{B} \times \Sigma \subset \mathcal{R}^{p+1}$. The goal is to estimate the true parameter $\boldsymbol{\gamma}_0 = (\boldsymbol{\beta}_0', \sigma_0^2)' \in \Gamma$.

The SLSE of $\boldsymbol{\gamma}$, say $\hat{\boldsymbol{\gamma}}^{\text{sls}}$, is obtained by minimizing the distances of the response variable and the squared response variable to their respective conditional mean simultaneously, that is

$$\hat{\boldsymbol{\gamma}}^{\text{sls}} = \underset{\boldsymbol{\gamma}\in\mathcal{B}\times\Sigma}{\arg\min}\, Q_n(\boldsymbol{\gamma}), \tag{1.9}$$

where

$$Q_n(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \rho_i'(\boldsymbol{\gamma}) W_i \rho_i(\boldsymbol{\gamma}),$$

$$\rho_i(\boldsymbol{\gamma}) = \left(y_i - g(\boldsymbol{x}_i;\boldsymbol{\beta}), y_i^2 - g^2(\boldsymbol{x}_i;\boldsymbol{\beta}) - \sigma^2\right)'$$

and $W_i = W(\boldsymbol{x}_i)$ is a 2×2 nonnegative definite matrix which may depend on $\boldsymbol{x}_i$. The SLSE is asymptotically consistent and normal under some regularity conditions which are provided in the appendix. Wang and Leblanc (2008) showed that with the optimal weighting matrix W, the SLSE is asymptotically more efficient than the OLSE if the third moment of the random error is nonzero, and both SLSE and OLSE have the same asymptotic variance-covariance matrix if the third moment of the random error is zero. Kim and Ma (2012) extended the method to the heteroscedastic error model. They proposed the semiparametric efficient (SE) estimator, which is equivalent to the SLSE in terms of achieving the optimal semiparametric efficiency bound. Then the SE estimator is extended to more general cases where the second moment can be an arbitrary function of the covariates.

1.4 Subset Selection

High dimensional data are encountered more often nowadays. For example, the DNA microarrays data has thousands of features with only a few tens to hundreds of

samples. These datasets contain a large number of predictors but with a relatively small sample size. The least squares method cannot be applied directly to these datasets, where the number of predictors exceeds the sample size. When faced with many possible explanatory variables, how do we find a simple model with a small number of relevant regressors that explains the dependent variable well? Which explanatory variables are important and should be included in the model and which of these are not useful and redundant? A common method for variable selection is called subset selection. The subset selection method yields sparse models which retain a smaller subset of predictors with the strongest effects on the response variable.

Best Subset Selection

Best subset selection compares all possible models with a fixed number of regressors. The scheme is as follows:

For each subset of size k, $0 \leq k \leq p$, fit all $\binom{p}{k}$ models with exact k predictors, then choose the best model with the smallest residual sum of squares. Denote the best model $\mathcal{M}_k$ for each k. Among the candidate models $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$, choose the 'best' model based on information criteria (AIC, BIC) or cross validation (CV).

Since the best subset selection fits all possible models (2^p), we can select the optimal one based on some model selection criteria. However, the computation is heavy and sometimes infeasible when p is too large. Next we introduce the stepwise selection, which requires less computation work.

Stepwise Selection

There are three main approaches of stepwise selection: forward selection, backward selection and the mixture of the two.

Forward stepwise selection starts with the intercept only. Among the variables that are not selected, add one variable into the model each time which contributes the most to the regression sum of squares (often takes the form of F test). Then we use AIC, BIC or CV to determine the number of predictors of the final model as mentioned before.

Backward stepwise selection starts with all predictors, then delete one variable at a time which contributes least to the model fit (conducted by z test with the smallest z score). Compared with forward stepwise selection, it can only be applied in cases where $N > p$.

While the stepwise selection has computational advantage over the best subset selection, there is no guarantee that it yields the optimal model compared with best subset selection. In the next section we introduce some shrinkage regression methods, which impose different constraints on the size of the regression coefficients.

1.5 Shrinkage Methods

In this section, we introduce some shrinkage methods in regression models. The ridge regression and lasso method are introduced first, followed by some recent developments on regularized regression methods.

1.5.1 Ridge Regression

From 1.5 we know that $\hat{\boldsymbol{\beta}}^{ols} = (X'X)^{-1}X'\boldsymbol{y}$. When $X'X$ is not of full rank, it cannot be inversed. Hence the $\hat{\boldsymbol{\beta}}^{ols}$ does not exist in this case. In order to overcome the

problem where $X'X$ cannot be inversed, Hoerl and Kennard (1970) proposed the ridge estimator $\hat{\beta}^{ridge}$, which is defined as the value of β that minimizes

$$\sum_{i=1}^{n}(y_i - \beta_0 - x_i'\beta_1)^2 + \lambda\sum_{j=1}^{p}\beta_j^2, \tag{1.10}$$

where $\lambda > 0$. The ridge estimators are not invariant to the scales of x_i variables. Hence it is suggested that the predictors ($x_j s$) be standardized before minimizing (1.10). Also, the intercept β_0 is not included in the penalty term, which is estimated by $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. By centering the response variable by $y_i - \bar{y}$, the intercept β_0 is incorporated into y_is. Similarly, the objective function can be written (1.10) as

$$(\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{\beta}. \tag{1.11}$$

From the first order condition $X'(\boldsymbol{y} - X\boldsymbol{\beta}) - \lambda\boldsymbol{\beta} = 0$, the solution is given by

$$\hat{\boldsymbol{\beta}}^{ridge} = (X'X + \lambda I)^{-1}X'\boldsymbol{y}. \tag{1.12}$$

Compared with $\hat{\boldsymbol{\beta}}^{ols}$ in (1.5), the ridge regression introduces a positive constant λ to the diagonal of $X'X$ before the inversion, which makes the problem nonsingular. At the same time it introduces some bias through the tuning parameter λ. From (1.12) we observe that

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{ridge} &\to \hat{\boldsymbol{\beta}}^{ols}, \text{as } \lambda \to\ 0, \\ \hat{\boldsymbol{\beta}}^{ridge} &\to 0, \quad \text{as } \lambda \to \infty, \end{aligned} \tag{1.13}$$

Information criteria and cross validation methods can be used for the selection of tuning parameter λ, which will be discussed in later section.

For both ordinary least squares estimator and ridge estimator, we estimate the coefficients of all regressors, which increase the difficulty of interpretability, especially when the number of predictors is large. As discussed in Tibshirani (1996), we often would like to select a smaller subset of the regressors that have the strongest effect on the response variable. Next we introduce lasso estimator, which retains good features of both subset selection and ridge regression method.

1.5.2 Lasso Estimator

As discussed above, the subset selection is a variable selection procedure which yields interpretable models. However, it can be unstable because of its discrete feature (the predictors are either dropped or retained in the model). The ridge regression is a continuous shrinkage method which is more stable. But it does not drop any of the explanatory variables out of the model. Tibshirani (1996) introduced the least absolute shrinkage and selection operator (lasso), which retains good features of both subset selection and ridge regression method. The lasso estimator is defined as the value of β that minimizes the objective function

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - x_i'\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|, \tag{1.14}$$

where $\lambda > 0$. Similar as the ridge regression, we normally standardize the predictors and centralize the response variable. We can write the objective function (1.14) as

$$\frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p}|\beta_j|. \tag{1.15}$$

The unique feature of L_1 regularization produces sparse solutions. In order to see why the lasso estimator works for variable selection problems, let's assume design matrix X is orthonormal for the moment. Then the ordinary least squares estimator becomes

$$\hat{\boldsymbol{\beta}}^{ols} = (X'X)^{-1}X'\boldsymbol{y} = X'\boldsymbol{y}.$$

The first order condition is

$$X'(\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda \text{ sign}(\boldsymbol{\beta}) = 0, \tag{1.16}$$

where $\text{sign}(\boldsymbol{\beta}) = \Big(\text{sign}(\beta_1), \text{sign}(\beta_2), \ldots, \text{sign}(\beta_p)\Big)'$. Solving equation (1.16) for β (see appendix) we obtain the lasso estimator:

$$\hat{\beta}_j^{lasso} = \text{sign}(\hat{\beta}_j^{ols})(|\hat{\beta}_j^{ols}| - \lambda)_+, \tag{1.17}$$

where a_+ denotes the positive part of a. That is, when $|\hat{\beta}_j^{ols}|$ is less than λ for some j (with small effects), lasso shrinks it to zero. Hence it achieves the goals of parameter estimation and variable selection simultaneously.

1.5.3 Recent Work on Regularized Regression

In this section we introduce some recent work on the regularized regression methods. From the objective function of ridge regression and lasso method we see that, in general, the shrinkage method for variable selection can be decomposed as:

$$\text{measure of model fit} + \lambda \times \text{penaly on the size of parameter.} \tag{1.18}$$

There are different penalties on the size of the coefficients. For example, the ridge regression penalty (Hoerl and Kennard (1970)) is

$$\lambda \sum_{j=1}^{p} \beta_j^2.$$

As we discussed, the ridge regression method keeps all the predictors in the model. Hence it cannot produce a sparse model. The L_1 absolute value penalty (Tibshirani (1996)) is

$$\lambda \sum_{j=1}^{p} |\beta_j|,$$

which can produce the sparse model owing to the nature of the L_1 penalty. The elastic net (Zou and Hastie (2005)) penalty is defined as

$$\lambda \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1-\alpha)|\beta_j|\right),$$

where $0 < \alpha < 1$. The elastic net method is a hybrid of the ridge regression and lasso method, which encourages the grouping effects among the predictors. As discussed in Zou and Hastie (2005), for typical microarray data with thousands of features (genes), the sample size is less than hundreds. For the genes sharing similar functional features and biological pathways, it is reasonable to group those genes together when the correlations among them are high. For those situations with grouping effects where $p \gg n$, the elastic net method performs better than the lasso estimator. Because at most n out of p predictors can be selected for the lasso method (Efron et al. (2004) and Zou and Hastie (2005)). Tibshirani et al. (2005) proposed fused lasso with the penalty function

$$\lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|.$$

With the additional penalization on the successive differences, the fused lasso is shown to perform well for similar situations where $p \gg n$. Fan and Li (2001) proposed smoothly clipped absolute deviation (SCAD) penalty, which penalize large coefficients less severely. The SCAD penalty function is defined as:

$$p_\lambda(|\beta|) = \begin{cases} \lambda|\beta|, & |\beta| \leq \lambda \\ \dfrac{\lambda a|\beta|}{a-1} - \dfrac{\beta^2+\lambda^2}{2(a-1)}, & \lambda < |\beta| \leq a\lambda \\ \dfrac{(a+1)\lambda^2}{2}, & a\lambda < |\beta| \end{cases}$$

where a is some positive number with $a > 2$. Fan and Li (2001) showed that the SCAD estimator enjoys the oracle properties which lasso does not have. Zou (2006) proposed the adaptive lasso method. Zhang and Lu (2007) proposed the adaptive lasso for Cox proportional hazards model. The penalty function is defined as

$$\lambda \sum_j \hat{w}_j |\beta_j|.$$

By introducing the adaptive weights $\hat{w}_j$, the adaptive lasso is shown to have the oracle properties. With the adaptive weights, larger penalty is imposed on the coefficients with small effects. Hence the important variables are more likely to be retained in the model, and the unimportant ones are more likely to be dropped.

Similar as the penalty functions, there are different measures of model fit. For example, least squares (Tibshirani (1996), Zou (2006)), likelihood (Fan and Li

(2001)), quantile regression (Belloni et al. (2011), Zou and Yuan (2008) and Li and Zhu (2008)), and generalized method of moments (Caner (2009), Caner et al. (2013) and Liao (2013)). In particular, Fan et al. (2014) proposed the weighted robust lasso (WR-lasso) for heavy-tailed high-dimensional problems. WR-lasso is a penalized quantile regression with the adaptively weighted L_1 penalty on the size of the regression coefficients. They showed that WR-lasso only requires mild conditions of the error distribution hence is robust against heavy-tailed problems. However, there is not much literature focusing on the variable selection problems with asymmetric error distributions. We propose the L_1-SLSE, which is robust against the shape of the error distribution for variable selection problems. Specifically, we apply the SLS method in variable selection problems, and we show that the L_1-SLSE is asymptotically more efficient than L_1-OLSE when the distribution of random error is asymmetric. And the variance-covariance matrix is asymptotically equivalent for both estimators when the distribution of random error is symmetric.

Chapter 2

Weighted L_1 Regularized Second-Order Least Squares Method

We introduce the L_1-SLSE in this chapter. With the regularization parameter properly chosen, the L_1-SLSE is asymptotically more efficient than the L_1-OLSE when the error distribution is asymmetric. And the two estimators are asymptotically equivalent when the error distribution is symmetric (e.g., normal). The existing algorithms for lasso type estimators are designed for either least squares or likelihood model fit, which cannot be applied to the L_1-SLSE directly. We provide an algorithm for L_1-SLSE in this chapter. The methods of selecting the regularization parameter is discussed at the end of this chapter.

2.1 Weighted L_1 Regularized SLSE

We apply the SLS method to variable selection problems. In particular, an adaptively weighted penalty term on the L_1 norm of the regression coefficients is introduced

to the objective function of SLSE. The resultant estimator L_1-SLSE has the oracle properties.

Consider the liner regression model

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \epsilon, \tag{2.1}$$

where $y \in \mathcal{R}$ is a response variable, $\boldsymbol{x} \in \mathcal{R}^p$ is a vector of covariates, $\boldsymbol{\beta}$ is a p-dimensional regression coefficient vector, and ϵ is the random error satisfying $E(\epsilon|\boldsymbol{x}) = 0$ and $E(\epsilon^2|\boldsymbol{x}) = \sigma^2$. It is easily seen that

$$E(y|\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}$$

and

$$E(y^2|\boldsymbol{x}) = (\boldsymbol{x}'\boldsymbol{\beta})^2 + \sigma^2$$

Define the parameter vector as $\boldsymbol{\gamma} = (\boldsymbol{\beta}', \sigma^2)'$ and the parameter space as $\Gamma = \mathcal{B} \times \Sigma \subset \mathcal{R}^{p+1}$. The goal is to estimate the true parameter $\boldsymbol{\gamma}_0 = (\boldsymbol{\beta}_0', \sigma_0^2)' \in \Gamma$. The proposed estimator L_1-SLSE $\hat{\boldsymbol{\gamma}}_n$ is defined as

$$\hat{\boldsymbol{\gamma}}_n = \operatorname{argmin} \left(Q_n(\boldsymbol{\gamma}) + n\lambda_n \sum_{j=1}^{p} w_j |\beta_j| \right), \tag{2.2}$$

where

$$Q_n(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \rho_i'(\boldsymbol{\gamma}) W_i \rho_i(\boldsymbol{\gamma}),$$

$$\rho_i(\boldsymbol{\gamma}) = \left(y_i - \boldsymbol{x}'\boldsymbol{\beta}, y_i^2 - (\boldsymbol{x}'\boldsymbol{\beta})^2 - \sigma^2 \right)',$$

$W_i = W(\boldsymbol{x}_i)$ is a 2×2 nonnegative definite matrix which may depend on $\boldsymbol{x}_i$, λ_n is the regularization parameter and w_js are some weights. The model fit part $Q_n(\gamma)$

exploits the information from both the first and the second moment, which makes the resultant estimator asymptotically more efficient than the L_1-OLSE for the asymmetrically distributed random error. In addition, the two estimators L_1-OLSE and L_1-SLSE are asymptotically equivalent for the symmetrically distributed random error. The adaptive weights w_js are chosen in the way such that the unimportant predictors (with small effects on the response variable) receive larger penalization than that of the important predictors. Hence the important predictors are more likely to be selected into the model and the redundant predictors will be dropped.

In particular we take $w_j = 1/|\hat{\beta}_j^{sls}|$ and λ_n is determined by 5-fold cross validation. We will use the optimal weight matrix W for all the L_1-SLSE discussed in the rest of the paper without extra mention. Under the assumptions given in the appendix, the proposed estimator L_1-SLSE has the following propositions.

Proposition 1 *Define the L_1-SLSE (as in 2.2)*

$$\hat{\boldsymbol{\gamma}}_n = argmin \left(Q_n(\boldsymbol{\gamma}) + n\lambda_n \sum_{j=1}^{p} w_j |\beta_j| \right).$$

If $\sqrt{n}\lambda_n = O_p(1)$, then $||\hat{\gamma}_n - \gamma_0|| = O_p(n^{-1/2})$.

Proposition 1 shows that $\hat{\boldsymbol{\gamma}}_n$ is root-n consistent if $\lambda_n \to 0$ at a proper rate. Let the true parameter be $\boldsymbol{\gamma}_0 = (\boldsymbol{\beta}'_{10}, \boldsymbol{\beta}'_{20}, \sigma_0^2)'$ and without loss of generality assume $\boldsymbol{\beta}_{20} \equiv 0$. Denote the nonzero part of the true parameter $\boldsymbol{\gamma}_{10} = (\boldsymbol{\beta}'_{10}, \sigma_0^2)'$ and its estimator $\hat{\boldsymbol{\gamma}}_{1n} = (\hat{\boldsymbol{\beta}}'_{1n}, \hat{\sigma}_n^2)'$. As discussed in Fan and Li (2001), a good estimator for variable selection problems should have the oracle properties. Using the words of

Fan and Li (2001), the oracle properties is explained as follows. When there are zero coefficients in the true model, they are estimated as 0 with probability approaching to 1. The nonzero coefficients are estimated as well as when the correct submodel is known in advance. Next we show that, with appropriately chosen λ_n, $\hat{\boldsymbol{\gamma}}_n$ enjoys the oracle properties.

Proposition 2 *If $\sqrt{n}\lambda_n \to 0$ and $n\lambda_n \to \infty$, then the L_1-SLSE has the following properties:*

(i) $P(\hat{\boldsymbol{\beta}}_{2n} = 0) \to 1$;

(ii) $\sqrt{n}(\hat{\boldsymbol{\gamma}}_{1n} - \boldsymbol{\gamma}_{10}) \xrightarrow{d} N\Big(0, \Sigma\Big)$,

where

$$\Sigma = (A^{-1})' B A^{-1},$$

$$A = E\Big[\frac{\partial \rho'(\boldsymbol{\gamma}_{10})}{\partial \boldsymbol{\gamma}} W(X) \frac{\partial \rho(\boldsymbol{\gamma}_{10})}{\partial \boldsymbol{\gamma}}\Big],$$

$$B = E\Big[\frac{\partial \rho'(\boldsymbol{\gamma}_{10})}{\partial \boldsymbol{\gamma}} W(X) \rho(\boldsymbol{\gamma}_{10}) \rho'(\boldsymbol{\gamma}_{10}) W(X) \frac{\partial \rho(\boldsymbol{\gamma}_{10})}{\partial \boldsymbol{\gamma}}\Big].$$

As discussed in Zou (2006), any consistent estimator of $\hat{\beta}^{sls}$ can be used for the adaptive weights w_i. Also, the estimation of the parameters involves the weighting matrix W, which also requires to be estimated. We suggest using the identity matrix $\hat{W}_1 = I$ for the first stage estimation. For the second stage, calculate the optimal $\hat{W}_2$ based on the estimated parameter from the first step.

The propositions indicate that the L_1-SLSE is asymptotically more efficient than the L_1-OLSE, when the distribution of random error is asymmetric. In fact the adaptive L_1-SLSE does not make assumptions on the distribution of random error. In situations where the random error follows normal or other symmetric distributions, the two estimators yield the same variance-covariance matrix asymptotically.

2.2 Computational Algorithms

The computation of the lasso solution (1.14) is a convex optimization problem, which can be solved by standard numerical algorithms. Efron et al. (2004) proposed least angle regression algorithm, which is computationally efficient for lasso type estimators. However, it cannot be applied on the L_1-SLSE directly. We present the algorithm for L_1-SLSE and methods for tuning parameter selection in this section.

2.2.1 Algorithm for L_1-SLSE

Since the solution of L_1-SLSE is a convex optimization problem, and the objective function is complicated which involves the matrix multiplication, we can approximate the object function by its quadratic form. Then the problem can be solved by standard Newton's method provided the gradient and Hessian matrix are calculated. Also, the quadratic optimization problem can be solved by several standard statistical optimization packages as mentioned before. Here we use CVX (Grant and Boyd (2014)) to solve the optimization problem, which is a Matlab-based modeling system for convex optimization. The algorithm is as follows

1. Calculate the gradient vector G and Hessian matrix H of the object function $Q_n(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$.

2. Calculate the Cholesky decomposition of the Hessian matrix as $H = L'L$. Define a pseudo response vector $Y = (L')^{-1}\{H\boldsymbol{\gamma} - G\}$ and approximate the object function by

$$\tilde{Q}(\boldsymbol{\gamma}) = \frac{1}{2}(Y - L\boldsymbol{\gamma})'(Y - L\boldsymbol{\gamma}).$$

3. Output

$$\hat{\boldsymbol{\gamma}} = \text{argmin } \tilde{Q}(\boldsymbol{\gamma}) + n\lambda_n \sum_{j=1}^{p} w_j|\beta_j|.$$

   The CVX package in Matlab software is used at this step for the optimization.

2.2.2 Selection of Regularization Parameter

In the variable selection literature, there are two commonly used methods for the selection of the tuning parameter: information criterion and cross validation.

Information Criterion

Akaike information criterion (Akaike (1973)) and Bayesian information criterion (Schwarz (1978)) are usually written in the form $-2ln(L) + kp$, with $k = 2$ for AIC and $k = ln(n)$ for BIC, where L is the likelihood and n is the sample size. The information criterion gives a trade-off between the model fit and model complexity. For linear models, the degree of freedom p is the number of predictors. However, for nonlinear problems, we need to estimate the effective degrees of freedom, which

is discussed in details by Zou et al. (2007). Wang et al. (2007) proposed BIC based tuning parameter selector for SCAD estimator, which is shown to be able to identify the true model consistently. Fan and Tang (2013) proposed to select the tuning parameter in high dimensional penalized likelihood by generalized information criterion, which is a generalization of AIC and BIC. Information-criterion based model selection is fast in computation, but it relies on a proper estimation of the degrees of freedom and relative large sample size as discussed above.

Cross Validation

Usually good model fit of a sample does not guarantee a satisfactory prediction. For example, a polynomial regression with high enough order will fit the data perfectly. But it usually fails in the prediction. Cross validation offers a way of estimating the prediction error, which splits the data into training set and validation set. Hence we can estimate the tuning parameter by minimizing the sum of prediction error. Cross validation is simple and intuitive but requires heavy computation. K-fold cross validation (with $K = 5$ or 10) is usually used in practice. We briefly describe the procedure of K-fold cross validation (see Tibshirani (2013)).

1. Divide the set $\{1, 2, \ldots, n\}$ into K subsets (K folds) with roughly equal size $F_1, F_2, \ldots, F_K$.

2. For $k = 1, 2, \ldots, K$

   Consider training on $(x_i, y_i),\ i \notin F_k$, and validating on $(x_i, y_i),\ i \in F_k$

   For each tuning parameter $\lambda \in \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$, compute $\hat{f}_\lambda^{-k}$ on the training

set, and record the total error on the validation set:

$$e_k(\lambda) = \sum_{i \in F_k} \left(y_i - \hat{f}_\lambda^{-k}(x_i) \right)^2 .$$

3. Compute the average error over all folds,

$$\mathrm{CV}(\lambda) = \frac{1}{n} \sum_{k=1}^{K} e_k(\lambda).$$

The estimation of optimal tuning parameter λ is given by

$$\hat{\lambda} = \underset{\lambda \in \{\lambda_1, \lambda_2, \ldots, \lambda_m\}}{\operatorname{argmin}} \mathrm{CV}(\lambda).$$

The minimal theoretical value of the tuning parameter (λ_{min}) such that all lasso coefficient are zero for elastic net estimator is suggested by Friedman et al. (2007). For the adaptive weighted L_1 regularized second-order least squares estimator, we determine λ_{min} by the results of several pilot simulations, where we assign a large enough value to λ which shrinks all coefficients to zero. In our simulation study, both BIC and CV yield similar results for the tuning parameter selection. We report the results based on the 5-fold cross validation.

Chapter 3

Numerical Studies

We report several simulation studies comparing the performance of L_1-OLSE and L_1-SLSE. We simulated datasets from linear model $y = \boldsymbol{x}'\boldsymbol{\beta} + \epsilon$. The tuning parameter λ_n are determined by 5-fold cross validation throughout all the simulations. We approximate the SLS objective function by its quadratic form (see appendix). We estimate the model error defined as $(\hat{\beta} - \beta)'V(\hat{\beta} - \beta)$ as discussed in Tibshirani (1996), where V is the population covariance matrix of X.

In order to illustrate the asymptotic efficiency of L_1-SLSE, we need to simulate different error distributions. In particular, normal, student t, chi square and gamma distribution are considered. The asymptotic equivalence between L_1-SLSE and L_1-OLSE are tested through normal and student t error distributions. The asymptotic efficiency of L_1-SLSE over L_1-OLSE is illustrated with chi-square and gamma error distributions.

The error distributions tested have wide applications in real world. For example, chi-square distribution is often encountered in magnetic resonance imaging (MRI). As discussed in den Dekker and Sijbers (2014), MRI is used as diagnostic tool

for biomedicine. The noise distribution is non-central chi-square for multiple-coil magnitude images. Gamma distribution is used for modelling the size of insurance claims (Boland (2007)), the description of rainfalls in environmental science (Aksoy (2000)) and inter-spike intervals in neuroscience (Robson and Troy (1987)).

The statistics reported in the table are as follows.

- Corr. is the average of 0 coefficients restricted to the true zero coefficients.
- Incorr. is the average of coefficients erroneously set to 0.
- MSE is the model error, which measures the performance of different model selection procedures. The MSE is calculated by $(\hat{\beta} - \beta)'V(\hat{\beta} - \beta)$.
- Mean($\hat{\beta}_i$) is the mean of nonzero coefficients.
- The numbers in parentheses are standard errors.

3.1 Simulation Example 1

We simulate 100 datasets from the linear model:

$$y = \mathbf{X}\beta + \epsilon,$$

where $\mathbf{X}$ is multivariate normal with mean 0, the correlation between x_i and x_j is $\rho^{|i-j|}$ with $\rho = 0.5$ and $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)'$. This model setting is adopted in several papers on regularization regression. For example Tibshirani (1996), Zou and Hastie (2005) and Fan and Li (2001).

The error term ϵ follows $N(0,1)$, $t(5)$ and normalized $\chi^2(3)$ respectively. Specifically, the standard normal distribution $N(0,1)$ serves as a reference group. $t(5)$ is a symmetric distribution but with thicker tails compared with the standard normal one. In order to illustrate the relative efficiency of adaptive L_1-SLSE, we include the asymmetric distribution (normalized $\chi^2(3)$) which is used in Wang and Leblanc (2008).

Simulation results of different sample sizes ($n = 100, 200, 300$) are given in the table (6.1), (6.2) and (6.3). Within each table, the average number of 0 coefficients restricted to the true zero coefficients of both L_1-OLSE and L_1-SLSE is close to the number of true zero coefficients (5) for different error distributions ($N(0,1)$, $t(5)$ and $\chi^2(3)$). The average number of coefficients erroneously set to 0 restricted to the true nonzero coefficients is close to zero for all different sample sizes, which means that both the methods (L_1-OLS and L_1-SLS) do not penalize too much on the size of the coefficients. The mean of the each nonzero estimator is close to the true coefficients (3, 1.5 and 2) for both L_1-OLSE and L_1-SLSE. The results observed

above agree with the propositions since both L_1-OLSE and L_1-SLSE are consistent estimators. When the error distribution is normal (N(0,1)), the MSE is smaller for L_1-OLSE than that of L_1-SLSE. In addition, the standard error of each nonzero estimator ($\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_5$) is smaller for L_1-OLSE than that of L_1-SLSE. These results are as expected since we know that the OLSE is the most efficient when the random error is normally distributed. For the symmetric error distribution $t(5)$, the two estimators L_1-OLSE and L_1-SLSE perform similarly as that of N(0,1) error distribution. The results are consistent with the propositions which state that the L_1-SLSE is asymptotically equivalent as the L_1-OLSE when the error distribution is symmetric. For the $\chi^2(3)$ error distribution, the MSE is smaller for L_1-SLSE than that of L_1-OLSE. In addition, the standard errors of nonzero estimators are smaller than that of L_1-OLSE. These results are expected from the propositions, which state that the L_1-SLSE is asymptotically more efficient than the L_1-OLSE when the distribution of random error is asymmetric.

Throughout the three tables with different sample sizes (6.1), (6.2) and (6.3), both estimators L_1-OLSE and L_1-SLSE perform better with larger sample size. Since the propositions give the asymptotic properties of the L_1-SLSE, we expect its better performance with larger sample size. Specifically, the average number of 0 coefficients restricted to the true zero coefficients for both L_1-OLSE and L_1-SLSE gets closer to the number of true zero coefficients (5). The MSE and standard errors of nonzero estimators decrease with the increase of sample size. In addition, the L_1-SLSE remains more efficient than L_1-OLSE with smaller standard errors throughout different sample sizes.

Table 3.1: Simulation example 1 with sample size n = 100

	N(0,1)		t(5)		$\chi^2(3)$	
Statistic	L_1-OLS	L_1-SLS	L_1-OLS	L_1-SLS	L_1-OLS	L_1-SLS
Corr. (5)	4.706	4.647	4.608	4.667	4.412	4.490
Incorr. (0)	0.000	0.000	0.000	0.000	0.000	0.000
MSE	0.105	0.106	0.178	0.116	0.171	0.134
se	(0.104)	(0.090)	(0.195)	(0.100)	(0.147)	(0.127)
mean($\hat{\beta}_1$)	2.991	3.007	3.017	2.961	2.982	2.924
se($\hat{\beta}_1$)	(0.211)	(0.219)	(0.278)	(0.259)	(0.331)	(0.202)
mean($\hat{\beta}_2$)	1.456	1.484	1.410	1.519	1.469	1.408
se($\hat{\beta}_2$)	(0.177)	(0.186)	(0.219)	(0.196)	(0.238)	(0.173)
mean($\hat{\beta}_5$)	2.011	1.964	1.941	1.948	1.953	1.966
se($\hat{\beta}_5$)	(0.173)	(0.182)	(0.182)	(0.150)	(0.185)	(0.200)

Table 3.2: Simulation example 1 with sample size n = 200

	N(0,1)		t(5)		$\chi^2(3)$	
Statistic	L_1-OLS	L_1-SLS	L_1-OLS	L_1-SLS	L_1-OLS	L_1-SLS
Corr. (5)	4.902	4.647	4.824	4.726	4.745	4.706
Incorr. (0)	0.000	0.000	0.000	0.000	0.000	0.000
MSE	0.066	0.056	0.086	0.063	0.077	0.046
se	(0.063)	(0.049)	(0.064)	(0.051)	(0.077)	(0.029)
mean($\hat{\beta}_1$)	2.998	3.013	2.996	2.997	2.953	3.001
se($\hat{\beta}_1$)	(0.193)	(0.182)	(0.217)	(0.158)	(0.183)	(0.155)
mean($\hat{\beta}_2$)	1.468	1.503	1.460	1.471	1.458	1.465
se($\hat{\beta}_2$)	(0.105)	(0.106)	(0.134)	(0.141)	(0.133)	(0.107)
mean($\hat{\beta}_5$)	1.984	1.960	1.977	1.988	2.007	1.998
se($\hat{\beta}_5$)	(0.124)	(0.111)	(0.142)	(0.127)	(0.149)	(0.117)

Table 3.3: Simulation example 1 with sample size n = 300

	N(0,1)		t(5)		$\chi^2(3)$	
Statistic	L_1-OLS	L_1-SLS	L_1-OLS	L_1-SLS	L_1-OLS	L_1-SLS
Corr. (5)	4.882	4.686	4.863	4.745	4.843	4.706
Incorr. (0)	0.000	0.000	0.000	0.000	0.000	0.000
MSE	0.035	0.039	0.047	0.044	0.047	0.042
se	(0.032)	(0.033)	(0.047)	(0.037)	(0.035)	(0.046)
mean($\hat{\beta}_1$)	3.004	3.014	2.986	3.014	3.040	3.015
se($\hat{\beta}_1$)	(0.132)	(0.153)	(0.136)	(0.152)	(0.151)	(0.145)
mean($\hat{\beta}_2$)	1.477	1.479	1.467	1.472	1.454	1.492
se($\hat{\beta}_2$)	(0.090)	(0.087)	(0.105)	(0.097)	(0.112)	(0.099)
mean($\hat{\beta}_5$)	1.981	1.994	1.969	1.994	1.981	2.000
se($\hat{\beta}_5$)	(0.110)	(0.095)	(0.131)	(0.110)	(0.116)	(0.099)

3.2 Simulation Example 2

We simulate the data from the linear model:

$$y = \mathbf{X}\beta + \epsilon,$$

where $\mathbf{X}$ is normal with mean 0, the correlation between x_i and x_j is $\rho^{|i-j|}$ with $\rho = 0.5$ and $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)'$. The error distribution follows $N(0, \sigma^2)$, with $\sigma = 1, 3, 9$ respectively.

The model setting is the same as the model in simulation example 1 except the distributions of random errors. As σ grows, the distribution gets less concentrated. We examine the performance of adaptive L_1-SLSE when the variance of the error distribution gets larger. From the propositions we expect the L_1-SLSE to perform similarly as L_1-OLSE when the error distribution is symmetric.

Simulation results of different sample size ($n = 100, 200, 300$) are given in the table (6.4), (6.5) and (6.6). Within each table, the error distributions are normal with increasing variances. It is known that the OLSE is the most efficient for the normally distributed random errors. However, it requires large sample size when the variance is large. With the finite sample size simulation results, several observations can be made. The average number of 0 coefficients restricted to the true zero coefficients for both L_1-OLSE and L_1-SLSE decreases with the increase of variances. The average number of coefficients erroneously set to 0 restricted to the true nonzero coefficients increases with the increasing variances, which means that the variable selection procedure gets worse for both L_1-OLS and L_1-SLS when the variance of random error gets larger. The MSE and standard errors of nonzero estimators increase as the variances of error terms become larger for both L_1-OLSE and L_1-SLSE.

Throughout the three tables (6.1), (6.2) and (6.3) with the increase of the sample size, the MSE and standard errors decrease for both L_1-OLSE and L_1-SLSE. The average number of 0 coefficients restricted to the true zero coefficients for L_1-SLSE gets closer to the number of true zero coefficients (5) than that of L_1-OLSE, which is a good property in terms of producing a sparse model. However, the MSE and its standard error are also large for L_1-SLSE. Hence a two step estimation procedure is suggested for the L_1-SLSE in order to reduce the variance. The L_1-SLS estimation procedure is performed in the first step. Then the regular SLS estimation procedure is conducted in the second step.

Table 3.4: Simulation example 2 with sample size n = 100

	N(0,1)		N(0,3^2)		N(0,9^2)	
Statistic	L_1-OLS	L_1-SLS	L_1-OLS	L_1-SLS	L_1-OLS	L_1-SLS
Corr. (5)	4.706	4.647	4.157	4.686	4.020	4.490
Incorr. (0)	0.000	0.000	0.020	0.020	0.863	1.157
MSE	0.105	0.106	0.458	0.591	5.448	6.342
se	(0.104)	(0.090)	(0.437)	(0.463)	(4.628)	(3.534)
mean($\hat{\beta}_1$)	2.991	3.007	3.018	2.957	2.771	2.481
se($\hat{\beta}_1$)	(0.211)	(0.219)	(0.477)	(0.453)	(1.535)	(1.461)
mean($\hat{\beta}_2$)	1.456	1.484	1.312	1.304	0.861	0.681
se($\hat{\beta}_2$)	(0.177)	(0.186)	(0.437)	(0.521)	(1.246)	(1.080)
mean($\hat{\beta}_5$)	2.011	1.964	1.951	1.801	1.311	0.752
se($\hat{\beta}_5$)	(0.173)	(0.182)	(0.335)	(0.382)	(1.181)	(0.932)

Table 3.5: Simulation example 2 with sample size n = 200

	N(0,1)		N(0,3^2)		N(0,9^2)	
Statistic	L_1-OLS	L_1-SLS	L_1-OLS	L_1-SLS	L_1-OLS	L_1-SLS
Corr. (5)	4.902	4.647	4.392	4.647	4.000	4.569
Incorr. (0)	0.000	0.000	0.000	0.000	0.353	0.706
MSE	0.066	0.056	0.211	0.232	2.513	4.590
se	(0.063)	(0.049)	(0.180)	(0.190)	(2.528)	(3.614)
mean($\hat{\beta}_1$)	2.998	3.013	2.984	3.018	2.654	2.800
se($\hat{\beta}_1$)	(0.193)	(0.182)	(0.306)	(0.347)	(0.895)	(1.177)
mean($\hat{\beta}_2$)	1.468	1.503	1.403	1.368	1.331	0.687
se($\hat{\beta}_2$)	(0.105)	(0.106)	(0.247)	(0.285)	(1.013)	(0.867)
mean($\hat{\beta}_5$)	1.984	1.960	1.956	1.877	1.547	1.076
se($\hat{\beta}_5$)	(0.124)	(0.111)	(0.268)	(0.230)	(0.822)	(1.010)

Table 3.6: Simulation example 2 with sample size n = 300

	N(0,1)		N(0,3^2)		N(0,9^2)	
Statistic	L_1-OLS	L_1-SLS	L_1-OLS	L_1-SLS	L_1-OLS	L_1-SLS
Corr. (5)	4.882	4.686	4.529	4.726	3.706	4.667
Incorr. (0)	0.000	0.000	0.000	0.000	0.078	0.451
MSE	0.035	0.039	0.118	0.114	0.979	2.984
se	(0.032)	(0.033)	(0.091)	(0.082)	(0.797)	(2.497)
mean($\hat{\beta}_1$)	3.004	3.014	3.006	2.996	3.024	2.821
se($\hat{\beta}_1$)	(0.132)	(0.153)	(0.232)	(0.223)	(0.624)	(0.776)
mean($\hat{\beta}_2$)	1.477	1.479	1.422	1.414	1.210	0.849
se($\hat{\beta}_2$)	(0.090)	(0.087)	(0.201)	(0.210)	(0.645)	(0.828)
mean($\hat{\beta}_5$)	1.981	1.994	1.943	1.934	1.825	1.155
se($\hat{\beta}_5$)	(0.110)	(0.095)	(0.192)	(0.173)	(0.539)	(0.770)

3.3 Simulation Example 3

We simulate the data from the linear model

$$y = \mathbf{X}\beta + \epsilon,$$

where $\mathbf{X}$ is normal with mean 0, the correlation between x_i and x_j is $\rho^{|i-j|}$ with $\rho = 0.5$ and $\beta = (0.7, 0.7, 0, 0, 0.7, 0, 0, 0)'$.

The error distribution follows Gamma(0.5,1) and Gamma(1,2) respectively. The $N(0,1)$ distribution is provided as reference. The performance is tested in this simulation example with different parameter settings and error distributions (Gamma).

Simulation results of different sample sizes ($n = 100, 200, 300$) are given in tables (6.7), (6.8) and (6.9). Within each table, for different error distributions (N(0,1), G(0.5,1) and G(1,2)), the average number of 0 coefficients restricted to the true zero coefficients of both L_1-OLSE and L_1-SLSE is close to the number of true zero coefficients (5). The average number of coefficients erroneously set to 0 restricted to the true nonzero coefficients is close to zero for all different sample sizes, which means that both the methods (L_1-OLS and L_1-SLS) do not penalize too much on the size of the coefficients. The mean of the each nonzero estimator is close to the true coefficients (0.7, 0.7 and 0.7) for both L_1-OLSE and L_1-SLSE. The results above agree with the propositions since both L_1-OLSE and L_1-SLSE are consistent estimators. When the error distribution is normal (N(0,1)), the MSE is smaller for L_1-OLSE than that of L_1-SLSE. In addition, the standard error of each nonzero estimator ($\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_5$) is smaller for L_1-OLSE than that of L_1-SLSE. These results are as expected since we know that the OLSE is the most efficient when

the random error is normally distributed. For the asymmetric error distributions $G(0.5,1)$ and $G(1,2)$, the MSE is smaller for L_1-SLSE than that of L_1-OLSE. In addition, the standard errors of nonzero estimators are smaller than that of L_1-OLSE. These results are expected from the propositions, which states that the L_1-SLSE is asymptotically more efficient than the L_1-OLSE when the distributions of random errors are asymmetric.

Throughout the three tables (6.7), (6.8) and (6.9) with different sample sizes, both estimators L_1-OLSE and L_1-SLSE perform better with increasing sample size. Since the propositions give the asymptotic properties of the L_1-SLSE, we expect its better performace with larger sample size. Specifically, the average number of 0 coefficients restricted to the true zero coefficients for both L_1-OLSE and L_1-SLSE gets closer to the number of true zero coefficients (5). The MSE and standard errors of nonzero estimators decrease with the increase of sample size. In addition, the L_1-SLSE remains more efficient than L_1-OLSE with smaller standard errors throughout the simulations with different sample sizes.

Table 3.7: Simulation example 3 with sample size n = 100

	N(0,1)		G(0.5,1)		G(1,2)	
Statistic	L_1-OLS	L_1-SLS	L_1-OLS	L_1-SLS	L_1-OLS	L_1-SLS
Corr. (5)	4.294	4.686	4.431	4.608	3.627	4.647
Incorr. (0)	0.000	0.000	0.000	0.000	0.255	0.000
MSE	0.051	0.077	0.029	0.037	0.291	0.040
se	(0.049)	(0.070)	(0.020)	(0.033)	(0.280)	(0.030)
mean($\hat{\beta}_1$)	0.669	0.659	0.715	0.701	0.607	0.675
se($\hat{\beta}_1$)	(0.131)	(0.157)	(0.098)	(0.135)	(0.290)	(0.117)
mean($\hat{\beta}_2$)	0.664	0.635	0.669	0.698	0.653	0.686
se($\hat{\beta}_2$)	(0.149)	(0.162)	(0.076)	(0.131)	(0.331)	(0.115)
mean($\hat{\beta}_5$)	0.687	0.612	0.657	0.672	0.539	0.635
se($\hat{\beta}_5$)	(0.129)	(0.146)	(0.112)	(0.091)	(0.327)	(0.115)

Table 3.8: Simulation example 3 with sample size n = 200

	N(0,1)		G(0.5,1)		G(1,2)	
Statistic	L_1-OLS	L_1-SLS	L_1-OLS	L_1-SLS	L_1-OLS	L_1-SLS
Corr. (5)	4.471	4.667	4.686	4.608	4.157	4.647
Incorr. (0)	0.000	0.000	0.000	0.000	0.020	0.000
MSE	0.022	0.028	0.018	0.013	0.089	0.014
se	(0.019)	(0.022)	(0.013)	(0.013)	(0.094)	(0.010)
mean($\hat{\beta}_1$)	0.680	0.680	0.680	0.689	0.705	0.694
se($\hat{\beta}_1$)	(0.095)	(0.107)	(0.092)	(0.072)	(0.170)	(0.072)
mean($\hat{\beta}_2$)	0.682	0.691	0.704	0.695	0.663	0.699
se($\hat{\beta}_2$)	(0.084)	(0.104)	(0.088)	(0.066)	(0.210)	(0.075)
mean($\hat{\beta}_5$)	0.685	0.648	0.677	0.681	0.664	0.687
se($\hat{\beta}_5$)	(0.088)	(0.089)	(0.075)	(0.054)	(0.193)	(0.071)

Table 3.9: Simulation example 3 with sample size n = 300

	N(0,1)		G(0.5,1)		G(1,2)	
Statistic	L_1-OLS	L_1-SLS	L_1-OLS	L_1-SLS	L_1-OLS	L_1-SLS
Corr. (5)	4.510	4.686	4.627	4.706	3.980	4.628
Incorr. (0)	0.000	0.000	0.000	0.000	0.000	0.000
MSE	0.013	0.015	0.007	0.007	0.053	0.008
se	(0.010)	(0.012)	(0.005)	(0.006)	(0.039)	(0.006)
mean($\hat{\beta}_1$)	0.690	0.691	0.701	0.690	0.690	0.687
se($\hat{\beta}_1$)	(0.072)	(0.075)	(0.045)	(0.050)	(0.161)	(0.061)
mean($\hat{\beta}_2$)	0.687	0.698	0.681	0.699	0.686	0.685
se($\hat{\beta}_2$)	(0.070)	(0.077)	(0.057)	(0.057)	(0.160)	(0.047)
mean($\hat{\beta}_5$)	0.681	0.673	0.680	0.689	0.661	0.687
se($\hat{\beta}_5$)	(0.064)	(0.069)	(0.037)	(0.048)	(0.132)	(0.049)

3.4 Real Data Example

We conduct a real data application in this chapter. The data is collected through the Work Limitations Questionnaire (WLQ) from a study of injured workers attending Shoulder & Elbow Speciality clinic, which is managed by Workplace Safety & Insurance Board (WSIB) of Ontario, Canada. The work disability is an important issue in public health, caused by which the productivity loss at work can exceed the direct medical cost. The WLQ developed by Lerner et al. (2001) and Lerner et al. (2012) offers a way of measuring how the health problems affect the job performance and the productivity loss at work. The WLQ has shown its good criterion validity and is adopted by several research institutes such as Ida et al. (2012) and Tang et al. (2011).

The outcome variable y is the Work Limitations Questionnaire (WLQ) index, which evaluates the proportion of time where difficulty is experienced in the following four different domains: time management (TM), physical demands (PD), mental-interpersonal (MI) and output demands (OD). The index quantifies the productivity loss at-work as a result of health disorders. The predictors (prognostic factors) are in table 3.10.

The covariates are standardized and the response variable is centered around its mean. We report the results of L_1-OLSE, L_1-SLSE, as well as the ordinary least squares estimator (OLSE) and stepwise regression. Several observations can be made from the table.

First, the directions of covariates obtained from each regression method are consistent. For example, the predictors x_2 and x_3 are retained for all the four

Table 3.10: Predictors of WSIB data

x_1	Age
x_2	Lower quick Disabilities of the Arm, Shoulder and Hand (DASH) score
x_3	Better mental health factor score
x_4	Better physical health factor score
x_5	Supervisor support
x_6	Organization support
x_7	Higher work flexibility
x_8	Lower shoulder pain and disability index (SPADI) pain
x_9	Lower von Korff pain disability score
x_{10}	Better general health
x_{11}	Skill discretion
x_{12}	Decision authority
x_{13}	Psychological pace
x_{14}	Higher job satisfactory
x_{15}	Coworker support

methods (L_1-SLS, L_1-OLS, OLS and stepwise regression). The covariates x_2 and x_3 have negative effects on the response variable, which means that with the lower quick disabilities score and better mental health, the expected productivity loss at-work will be less. For the methods L_1-SLS, L_1-OLS and OLS, the variable x_4 is retained in the model. The negative sign of x_4 means that with better physical health factor score, there will be lower expected productivity loss at-work. The covariates x_6, x_7, x_{10}, x_{11} and x_{12} are selected into the model for both methods L_1-OLS and OLS. The better organization support and the decision authority are, the lower productivity loss will be.

Second, although the data are standardized for all the regression methods, the scale of the resulting regression coefficient varies. The shrinkage method L_1-OLS and L_1-SLS impose a constraint on the L_1 norm of the coefficients. The scales of the resulting coefficients (of x_2, x_3 and x_4) given by the L_1-OLS and L_1-SLS methods

are smaller than the scales given by the OLS method. In particular, the coefficient of x_2 (DASH score) is -1.69 for the stepwise regression method and -1.30 for the OLS method. In contrast, the coefficients of x_2 for L_1-SLS and L_1-OLS are -0.74 and -1.17 respectively.

Third, the standard error varies among different regression methods. As we discussed before, the L_1-SLSE is asymptotically more efficient than the L_1-OLSE when the distribution of the random error is asymmetric. And the two estimators are asymptotically equivalent when the random error is symmetrically distributed. Hence the L_1-SLSE is more robust against the shape of the error distribution than the L_1-OLSE. For example, the standard error for the predictor x_4 is 0.30 for L_1-SLS, 0.41 for L_1-OLS and 0.68 for OLS. For the OLS regression method, the significance test for the coefficients are conducted through the z test. The test statistic is calculated by $\hat{\beta}_j/se(\beta_j)$, then compared with the critical z or t value. For the lasso type estimator, a significance test method is suggested by Lockhart et al. (2014), which will not be discussed here.

Fourth, the number of covariates selected varies among different regression methods. As discussed above, there is no guarantee that the stepwise method will yield the optimal model. In this real data example, the covariates selected by the stepwise method are x_2(DASH score) and x_3(better mental health factor score). Besides x_2 and x_3, the L_1-SLS keeps another covariate x_4(better physical health factor score). For the L_1-OLS method, besides the predictors selected by L_1-SLS, 5 more variables are retained in the model. In terms of the model interpretation, the L_1-SLS yields a model which is easy to interpret while retaining the important predictors.

Table 3.11: Results of WSIB data analysis

	L_1 SLS		L_1 OLS		OLS		Stepwise	
	coef	se	coef	se	coef	se	coef	se
x_1			0.23	(0.16)	0.38	(0.43)		
x_2	-0.74	(0.38)	-1.17	(0.40)	-1.30	(0.74)	-1.69	(0.46)
x_3	-1.46	(0.24)	-2.05	(0.30)	-2.11	(0.60)	-1.61	(0.46)
x_4	-0.45	(0.30)	-0.88	(0.41)	-1.07	(0.68)		
x_5					0.34	(0.54)		
x_6			-0.01	(0.20)	-0.36	(0.56)		
x_7			0.35	(0.25)	0.51	(0.55)		
x_8					0.25	(0.59)		
x_9					0.07	(0.52)		
x_{10}			0.69	(0.28)	0.88	(0.57)		
x_{11}			0.06	(0.27)	0.41	(0.62)		
x_{12}			-0.18	(0.35)	-0.63	(0.64)		
x_{13}					-0.03	(0.51)		
x_{14}					0.07	(0.58)		
x_{15}					0.03	(0.56)		

Chapter 4

Summary and Future Work

The ordinary least squares estimator (OLSE) is the most efficient estimator when the error distribution is normal. However, when the error term is not normally distributed, the ordinary least squares estimator is not necessarily the most efficient. In addition, the data collected in real-world does not always follow normal distribution, and often unknown to us. Hence it is not always reasonable to make assumptions that the error terms are normally distributed. Wang and Leblanc (2008) proposed the second-order least squares estimator (SLSE), which is robust against the shape of the error distribution. They showed that by exploiting more information form the higher order moments, the SLSE is asymptotically more efficient than the OLSE when the error distribution is asymmetric. And the two estimators are asymptotically equivalent when the error distribution is symmetric. The second-order least squares method does not require assumptions on the error distribution, which is robust against the shape of the error distribution compared with the ordinary least squares method.

There are two reasons that we are not satisfied with the ordinary least squares

method, as discussed in Hastie et al. (2001). The first reason is the prediction accuracy. The OLSE often has small bias but high variance. The prediction accuracy can be improved by the shrinkage of the coefficients or setting some coefficients to zero, which is a trade-off between the bias and the variance. The resultant estimator often has smaller mean square error.

The subset selection method is a discrete process of selecting the variables. The best subset selection method yields the best model but requires heavy computation when the number of predictors p is large. The stepwise selection method is fast in computation but not stable, where the best model is not guaranteed. The shrinkage method is a continuous process of the parameter estimation and variable selection. There are different shrinkage methods. The ridge regression shrinks all the regression coefficients by the same amount, which keeps all the predictors. The least absolute shrinkage and selection operator (lasso) keeps good features of both the subset selection ridge regression methods, which achieves the parameter estimation and variable selection simultaneously. The lasso type penalized regression methods can be decomposed of the model fit part and the penalization part. There are different model fit parts, for example, least squares, likelihood, quantile regression and generalized method of moments. There are also different penalizations on the size of the regression coefficients. For example, the L_1 (lasso), L_2 (ridge) penalization, the mixture of the two (elastic net), the general L_q ($q > 0$) norm penalization, adaptive lasso and smoothly clipped absolute deviation (SCAD) penalization. These penalization functions have their own advantages for different problems. For example, the elastic net performs well for simulations where the grouping effects exists. The SCAD and adaptive lasso estimator have the oracle properties. The oracle properties

are as follows. When there are zero coefficients in the true model, they are estimated as 0 with probability approaching to 1. The nonzero coefficients are estimated as well as when the correct submodel is known in advance.

Fan et al. (2014) proposed the weighted robust lasso (WR lasso) estimator, which is penalized quantile regression. The WR lasso is robust against the heavy-tail error distribution in high-dimensional settings. However, there are not much literature focusing on the estimators that are robust against asymmetric error distributions for high-dimensional problems. In addition, most of existing regularization methods are designed for least squares or likelihood, which requires the error distribution to be either symmetric or known. We apply the second-order least squares method in high dimensional settings and propose adaptively weighted L_1 regularized SLSE (L_1-SLSE) for high dimensional problems, which is robust against the shape of the error distribution. Also, the error distribution is not required to be known. The L_1-SLSE is asymptotically more efficient than L_1-OLSE when the error distribution is asymmetric. And the two estimators are asymptotically equivalent when the error distribution is symmetric. From the simulation results, the L_1-SLSE is shown to perform well in many situations. The comparison among different methods in the real data application shows that L_1-SLSE yields the estimators with smaller standard errors and a model which is easier to interpret compared with other methods.

The following aspects of future research on L_1-SLSE are of interest. First, different penalty functions (e.g., adaptive lasso, SCAD and elastic net) have their own advantages in different situations. It is worthwhile exploring the performance of these different penalty functions when combined with the second-order least squares model fit. Second, the model being tested throughout the thesis is simple linear.

It can be observed from the propositions that the model can be either linear or nonlinear (e.g., logit and probit models). It is worth extending the estimator to the nonlinear models. Third, the existing algorithms for the lasso type problems cannot be applied to the L_1-SLSE directly. In addition, the optimization part of the algorithm of the L_1-SLSE is based on the CVX package, which is a general optimaztion package for various problems. Hence a more efficient algorithm for the L_1-SLSE is of interest.

Chapter 5

Appendix

5.1 Regularity Conditions and Lemmas

Regularity conditions for adaptive L_1 regularized second-order least squares estimator:

1. $g(\boldsymbol{x};\boldsymbol{\beta})$ is a measurable function of $\boldsymbol{x}$ for every $\boldsymbol{\beta} \in \Theta$, and is continuous in $\boldsymbol{\beta} \in \Theta$ for μ-almost all x.

2. $E||W(\boldsymbol{x})||(sup_{\Theta} g^4(\boldsymbol{x};\boldsymbol{\beta}) + 1) < \infty$.

3. The parameter space $\Gamma \subset \mathbb{R}^{p+1}$ is compact.

4. For any $\boldsymbol{\gamma} \in \Gamma$,

$$E[\rho(\boldsymbol{\gamma}) - \rho(\boldsymbol{\gamma}_0)]' W(\boldsymbol{x})[\rho(\boldsymbol{\gamma}) - \rho(\boldsymbol{\gamma}_0)] = 0$$

   if and only if $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$, where

$$\rho(\gamma) = (\boldsymbol{y} - g(\boldsymbol{x};\boldsymbol{\beta}), \boldsymbol{y}^2 - g^2(\boldsymbol{x};\boldsymbol{\beta}) - \sigma^2)'.$$

5. For any $\boldsymbol{\beta}_0$ is an interior point of $\boldsymbol{\beta}$ and, for μ-almost all $\boldsymbol{x}$, $g(\boldsymbol{x};\boldsymbol{\beta})$ is twice continuously differentiable in $\boldsymbol{\beta}$. Furthermore, the first two derivatives satisfy

$$E||W(\boldsymbol{x})||sup_{\boldsymbol{\beta}}\Big|\Big|\frac{\partial g(\boldsymbol{x};\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|\Big|^4 < \infty,$$

$$E||W(\boldsymbol{x})\Big|\Big|sup_{\boldsymbol{\beta}}\Big|\Big|\frac{\partial^2 g(\boldsymbol{x};\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\Big|\Big|^4 < \infty.$$

6. The matrix

$$A = E\Big[\frac{\partial \rho'(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}} W(X) \frac{\partial \rho(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}}\Big]$$

is positive definite, where

$$\frac{\partial \rho'(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}} = -\begin{pmatrix} \frac{\partial g(\boldsymbol{x};\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} & 2g(\boldsymbol{x};\boldsymbol{\beta}_0)\frac{\partial g(\boldsymbol{x};\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \\ 0 & 1 \end{pmatrix}.$$

In this thesis, our simulation study focuses on the linear model as in (2.1). The model assumptions for adaptive weighted regularized L_1 second-order least squares estimator can be simplified as follows.

1.* $E\big(||W(\boldsymbol{x})|| \cdot (||\boldsymbol{x}||^4 + 1)\big) \leq \infty$

2.* The parameter space $\Gamma \subset \mathbb{R}^{p+1}$ is compact.

3.* For any $\boldsymbol{\gamma} \in \Gamma$,

$$E[\rho(\boldsymbol{\gamma}) - \rho(\boldsymbol{\gamma}_0)]' W(\boldsymbol{x})[\rho(\boldsymbol{\gamma}) - \rho(\boldsymbol{\gamma}_0)] = 0$$

if and only if $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$, where

$$\rho(\boldsymbol{\gamma}) = (\boldsymbol{y} - g(\boldsymbol{x};\boldsymbol{\beta}), \boldsymbol{y}^2 - g^2(\boldsymbol{x};\boldsymbol{\beta}) - \sigma^2)'.$$

4.* The matrix

$$A = E\Big[\frac{\partial \rho'(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}} W(\boldsymbol{x}) \frac{\partial \rho(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}}\Big]$$

is positive definite, where

$$\frac{\partial \rho'(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}} = -\begin{pmatrix} \boldsymbol{x} & 2(\boldsymbol{x}\boldsymbol{\beta}_0)\boldsymbol{x} \\ 0 & 1 \end{pmatrix}.$$

Below are the lemmas from Wang and Leblanc (2008), which are used in the construction of the propositions in this thesis.

Lemma 1 *Under Assumptions 1-4,*

$$\frac{1}{n} Q_n(\boldsymbol{\gamma}) \xrightarrow{a.s.} Q(\boldsymbol{\gamma}) = E\rho_1'(\boldsymbol{\gamma}) W \rho_1(\boldsymbol{\gamma})$$

uniformly for all $\boldsymbol{\gamma} \in \Gamma$.

Lemma 2 *Under Assumptions 1-6,*

i)

$$\frac{1}{\sqrt{n}} \frac{\partial Q_n(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}} \xrightarrow{d} N(0, 4B),$$

where

$$B = E\Big[\frac{\partial \rho'(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}} W(\boldsymbol{x}) \rho(\boldsymbol{\gamma}_0) \rho'(\boldsymbol{\gamma}_0) W(\boldsymbol{x}) \frac{\partial \rho(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}}\Big].$$

ii)

$$\frac{1}{n} \frac{\partial^2 Q(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} = 2A,$$

where

$$A = E\Big[\frac{\partial \rho'(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}} W(X) \frac{\partial \rho(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}}\Big].$$

5.2 Quadratic Approximation Proof

Next we illustrate the quadratic approximation of $Q_n(\beta)$. Define $Q_n(\beta)$ a function of parameter β of interest. Let $G = \nabla Q_n(\beta) = \partial Q_n(\beta)/\partial\beta$ and $H = \nabla^2 Q_n(\beta)/\partial\beta\partial\beta'$. Let $L'L$ be the Cholesky decomposition of H. Define pseudo response vector $Y = (L')^{-1}\{H\beta - G\}$. Using Taylor expansion for $Q_n(\beta)$ at $\beta = \beta_0$, we have

$$\begin{aligned}
Q_n(\beta) &\approx Q_n(\beta_0) + G_0'(\beta - \beta_0) + \frac{1}{2}(\beta - \beta_0)' H_0 (\beta - \beta_0) \\
&= Q_n(\beta_0) + (\beta - \beta_0)' G_0 + \frac{1}{2}(\beta - \beta_0)' H_0 (\beta - \beta_0) \\
&= Q_n(\beta_0) - \beta_0' G_0 + \frac{1}{2}\beta_0' H_0 \beta_0 + \frac{1}{2}\beta' H_0 \beta - \beta'(H_0\beta_0 - G_0) \\
&= c_1 + \frac{1}{2}\{(L_0\beta)'(L_0\beta) - 2\beta' L_0' L_0'^{-1}[H_0\beta_0 - G_0]\} \\
&= c_2 + \frac{1}{2}\{L_0'^{-1}[H_0\beta_0 - G_0] - L_0\beta\}'\{L_0'^{-1}[H_0\beta_0 - G_0] - L_0\beta\} \\
&= c_2 + \frac{1}{2}(Y_0 - L_0\beta)'(Y_0 - L_0\beta)
\end{aligned}$$

Thus, by second-order Taylor expansion, $Q_n(\beta)$ can be approximated by the quadratic form $\frac{1}{2}(Y_0 - L_0\beta)'(Y_0 - L_0\beta)$.

5.3 Sample MATLAB Codes

```
function [ y ] = central( x )
%Centralization
y = x - mean(x);
end

function [ y ] = standardize( x, n )
%Standardization
y = (x-repmat(mean(x),n,1))./repmat(std(x),n,1);
end

function [ G ] = gradient_sls(x,y,p,par,g,w11,w12,w22)
%Gradient SLS
G = [sum(repmat(-2*w11.*(y-g)-2*w12.*(y.^2-g.^2-par(p+1)) ...
    -4*w12.*(y-g).*g-4*w22*(y.^2-g.^2-par(p+1)).*g,1,p).*x),
    ...
    sum(-2*w12.*(y-g)-2*w22*(y.^2-g.^2-par(p+1)))];
end

function [ H ] = hessian_sls( x,y,n,p,par,g,w11,w12,w22)
%Hessian
H = [x'*(repmat(2*w11+12*w12.*g-4*w12.*y+8*w22*g.^2- ...
    4*w22*(y.^2-g.^2-par(p+1)),1,p).*x)
    sum(repmat(2*w12+4*w22*g,1,p).*x)' ; ...
    sum(repmat(2*w12+4*w22*g,1,p).*x) 2*w22*n];
```

```
end

function [ err ] =
alasso_tuning_parm(lower,step,upper,p,matX,vecY,xtest,ytest,wt)
n = (upper - lower)/step + 1;
err = zeros(1,n);

for lambda=lower:step:upper
    cvx_begin quiet
    variable par(p+1);
    minimize(sum((vecY - matX*par).^2) +
    lambda*sum(abs(par(1:p)./wt)));
    cvx_end
    cnt = int64((lambda-lower)/step+1);
    err(cnt) = sum((ytest-xtest*par(1:p)).^2);
end
end

n = 100
simu = 50
beta = (3,1.5,0,0,2,0,0,0)'

lambda_l =
lambda_u =
lambda_int =
lambda_n = (lambda_u - lambda_l)/lambda_int + 1;
```

```
beta_nonzero_pos = find(beta);

r = sum(beta ~= 0);

beta_zero_pos = find(beta==0);


p = length(beta);

MU = zeros(1,p);

NU = 3;


beta_lasso = zeros(p,simu);

sigma_lasso = zeros(1,simu);

mse = zeros(1,simu);

lambdas = zeros(simu,1);

cov_est = zeros(simu,r);

SIGMA = zeros(p,p);


for k = 1:p

    for m = 1:p

        SIGMA(k,m) = 0.5^(abs(k-m));

    end

end


sim = 1;

while sim < simu + 1

    x0 = mvnrnd(MU,SIGMA,n);

    %epsilon0 = trnd(5,n,1);
```

```
%epsilon0 = (chi2rnd(NU,n,1) - NU)/sqrt(NU);
epsilon0 = normrnd(0,1,n,1);
%epsilon0 = (gamrnd(0.5,1,n,1)-0.5)/sqrt(0.5);
%epsilon0 = (gamrnd(1,2,n,1)-2)/sqrt(4);
% G(k=,theta=) M=k theta V=k theta^2
y0 = x0*beta + epsilon0;
x = (x0-repmat(mean(x0),n,1))./repmat(std(x0),n,1);
y = y0 - mean(y0);

%lambda_l = 0.1;
%lambda_u = 20.1;
%lambda_int = 0.2;
lambda_n = (lambda_u - lambda_l)/lambda_int + 1;

beta_can = zeros(p, lambda_n);
sigsq_can = zeros(lambda_n,1);
obj_can = zeros(lambda_n,1);

c = cvpartition(n,'kfold',5);
cv_err = zeros(c.NumTestSets,lambda_n);

for i = 1:c.NumTestSets
    trIdx = c.training(i);
    teIdx = c.test(i);
    xtrain = x(find(trIdx),:);
    ytrain = y(find(trIdx),:);
```

```
xtest = x(find(teIdx),:);

ytest = y(find(teIdx));

beta_ini_tr = (xtrain'*xtrain)\xtrain'*ytrain;

g_tr = xtrain*beta_ini_tr;

err_tr = ytrain - g_tr;

sigsq_tr = sum(err_tr.^2)/(n-2);

mu3_tr = mean(err_tr.^3);

mu4_tr = mean(err_tr.^4);

parm_tr = [beta_ini_tr ; sigsq_tr];

coef_w_tr =

parm_tr(end)*(mu4_tr-(parm_tr(end))^2)-mu3_tr^2;

w11_tr =

(mu4_tr+4*mu3_tr*g_tr+4*parm_tr(p+1)*g_tr.^2-parm_tr(p+

1)^2)/coef_w_tr;

w12_tr = (-mu3_tr-2*parm_tr(p+1)*g_tr)/coef_w_tr;

w22_tr = (parm_tr(p+1))/coef_w_tr;

fd_tr =

[sum(repmat(-2*w11_tr.*(ytrain-g_tr)-2*w12_tr.*(ytrain.

^2-g_tr.^2-parm_tr(p+1))-4*w12_tr.*(ytrain-g_tr).*g_tr-

4*w22_tr*(ytrain.^2-g_tr.^2-parm_tr(p+1)).*g_tr,1,p).*x

train),sum(-2*w12_tr.*(ytrain-g_tr)-2*w22_tr*(ytrain.^2

-g_tr.^2-parm_tr(p+1)))];

sd_tr =
```

```
[xtrain'*(repmat(2*w11_tr+12*w12_tr.*g_tr-4*w12_tr.*ytr
ain+8*w22_tr*g_tr.^2-4*w22_tr*(ytrain.^2-g_tr.^2-parm_t
r(p+1)),1,p).*xtrain)

sum(repmat(2*w12_tr+4*w22_tr*g_tr,1,p).*xtrain)' ;
sum(repmat(2*w12_tr+4*w22_tr*g_tr,1,p).*xtrain)
2*w22_tr*n];

if all(eig(sd_tr) > 0)
    matX_tr = chol(sd_tr);
else
    continue
    %i
end
vecY_tr = (transpose(matX_tr))\(sd_tr*parm_tr -
fd_tr');

for lambda=lambda_l:lambda_int:lambda_u
    cvx_begin quiet
    variable para_cv(p+1);
    minimize(sum((vecY_tr - matX_tr*para_cv).^2) +
    lambda*sum(abs(para_cv(1:p)./beta_ini_tr)));
    cvx_end

    beta_lasso_t = para_cv(1:p);
    cnt = int64((lambda-lambda_l)/lambda_int+1);
```

```
        cv_err(i,cnt) = sum((ytest-xtest*beta_lasso_t).^2);
    end
    %i
end

cv_err1 = sum(cv_err);
index = find(cv_err1==min(cv_err1));
lambda_c = (index-1)*lambda_int+lambda_l;
lambdas(sim,:) = lambda_c;

beta_ini = (x'*x)\x'*y;
g = x*beta_ini;
err = y - g;
sigsq_ini = sum(err.^2)/(n-2);
mu3 = mean(err.^3);
mu4 = mean(err.^4);
parm = [beta_ini ; sigsq_ini];
coef_w = parm(end)*(mu4-(parm(end))^2)-mu3^2;
w11 = (mu4+4*mu3*g+4*parm(p+1)*g.^2-parm(p+1)^2)/coef_w;
w12 = (-mu3-2*parm(p+1)*g)/coef_w;
w22 = (parm(p+1))/coef_w;

fd =
[sum(repmat(-2*w11.*(y-g)-2*w12.*(y.^2-g.^2-parm(p+1))-4*w1
2.*(y-g).*g-4*w22*(y.^2-g.^2-parm(p+1)).*g,1,p).*x),sum(-2*
```

```
w12.*(y-g)-2*w22*(y.^2-g.^2-parm(p+1)))];
sd =
[x'*(repmat(2*w11+12*w12.*g-4*w12.*y+8*w22*g.^2-4*w22*(y.^2
-g.^2-parm(p+1)),1,p).*x)
sum(repmat(2*w12+4*w22*g,1,p).*x)' ;
sum(repmat(2*w12+4*w22*g,1,p).*x) 2*w22*n];

if all(eig(sd) > 0)
    matX = chol(sd);
else
    continue
    sim
end
vecY = (transpose(matX))\(sd*parm - fd');

cvx_begin quiet
variable para(p+1);
minimize(sum((vecY - matX*para).^2) +
lambda*sum(abs(para(1:p)./beta_ini)));
cvx_end
%para
beta_lasso(:,sim) = para(1:p);

%fd_3 =
[sum(repmat(-2*w11_3.*(y-g3)-2*w12_3.*(y.^2-g3.^2-parm3(p+1
))-4*w12_3.*(y-g3).*g3-4*w22_3*(y.^2-g3.^2-parm3(p+1)).*g3,
```

```
    1,p).*x),sum(-2*w12_3.*(y-g3)-2*w22_3*(y.^2-g3.^2-parm3(p+1
    )))];
    %second_d_3 =
    [x'*(repmat(2*w11_3+12*w12_3.*g3-4*w12_3.*y+8*w22_3*g3.^2-4
    *w22_3*(y.^2-g3.^2-parm3(p+1)),1,p).*x)
    sum(repmat(2*w12_3+4*w22_3*g3,1,p).*x)' ;
    sum(repmat(2*w12_3+4*w22_3*g3,1,p).*x) 2*w22_3*n];
    %B = fd_3'*fd_3/(4*n);
    %A = second_d_3/(2*n);
    %D = inv(A)*B*inv(A)
    %p886 W&L
    A = [x'*(repmat(w11+4*w12.*g+4*w22*g.^2,1,p).*x)
    sum(repmat(w12+2*w22*g,1,p).*x)';sum(repmat(w12+2*w22*g,1,p
    ).*x) w22]/n;
    D = inv(A);
    cov_est(sim,:) =
    diag(D(beta_nonzero_pos,beta_nonzero_pos))/n;
    sim = sim + 1;
end

%average mean square errors
for s=1:simu
    model_error(s) =
    (beta_lasso(:,s)-beta)'*SIGMA*(beta_lasso(:,s)-beta);
```

```
end

beta_lasso_none_zero = beta_lasso(beta_nonzero_pos,:);

beta_lasso_zero = beta_lasso(beta_zero_pos,:);

Corr =  sum(sum(abs(beta_lasso_zero) < 1e-6))/sim

Incorr =  sum(sum(abs(beta_lasso_none_zero) < 1e-6))/sim

MSE = median(model_error)

median_y = bootstrp(100,@median,model_error);

se = std(median_y)

mean_beta = mean(beta_lasso,2);

mean_beta_nonzero = mean_beta(beta_nonzero_pos);

std_beta = std(beta_lasso,0,2);

std_beta_nonzero = std_beta(beta_nonzero_pos);

beta_1 = mean_beta_nonzero(1)

se_1 = std_beta_nonzero(1)

beta_2 = mean_beta_nonzero(2)

se_2 = std_beta_nonzero(2)

beta_3 = mean_beta_nonzero(3)

se_3 = std_beta_nonzero(3)

end

fprintf('\n rng(1)\n n=100 200 300\n simu=50\n 3 1.5 0 0 2 0 0

0\n N(0,1)\n')
```

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*.

Aksoy, H. (2000). Use of gamma distribution in hydrological analysis. *Turkish Journal of Engineering and Environmental Sciences*, 24(6):419–428.

Belloni, A., Chernozhukov, V., et al. (2011). 1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130.

Boland, P. J. (2007). *Statistical and probabilistic methods in actuarial science*. CRC Press.

Caner, M. (2009). Lasso-type gmm estimator. *Econometric Theory*, 25(01):270–290.

Caner, M., Han, X., and Lee, Y. (2013). Adaptive elastic net gmm estimator with many invalid moment conditions: A simultaneous model and moment selection. *Manuscript*.

Caner, M. and Zhang, H. H. (2014). Adaptive elastic net for generalized methods of moments. *Journal of Business & Economic Statistics*, 32(1):30–47.

den Dekker, A. and Sijbers, J. (2014). Data distributions in magnetic resonance images: A review. *Physica Medica*, 30(7):725–741.

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

Fan, J., Fan, Y., and Barut, E. (2014). Adaptive robust variable selection. *Ann. Statist.*, 42(1):324–351.

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

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):531–552.

Friedman, J., Hastie, T., Hfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416.

Grant, M. and Boyd, S. (2008). Graph implementations for nonsmooth convex programs. In Blondel, V., Boyd, S., and Kimura, H., editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited. http://stanford.edu/~boyd/graph_dcp.html.

Grant, M. and Boyd, S. (2014). CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054.

Hansen, L. P., Heaton, J., and Yaron, A. (1996). Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*, 14(3):262–280.

Hastie, T. and Tibshirani, R. (2014). Statistical learning: Chapter 6, model selection.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Ida, H., Nakagawa, K., Miura, M., Ishikawa, K., and Yakura, N. (2012). Development of the work limitations questionnaire japanese version (wlq-j): Fundamental examination of the reliability and validity of the wlq-j. *SANGYO EISEIGAKU ZASSHI*, 54(3):101–107.

Kim, M. and Ma, Y. (2012). The efficiency of the second-order nonlinear least squares estimator and its extension. *Annals of the Institute of Statistical Mathematics*, 64(4):751–764.

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

Koenker, R. (2005). *Quantile regression*. Number 38. Cambridge university press.

Lerner, D., Adler, D., Hermann, R. C., Chang, H., Ludman, E. J., Greenhill, A., Perch, K., McPeck, W. C., and Rogers, W. H. (2012). Impact of a work-focused intervention on the productivity and symptoms of employees with depression. *Journal of Occupational and Environmental Medicine*, 54(2):128.

Lerner, D., Amick III, B. C., Rogers, W. H., Malspeis, S., Bungay, K., and Cynn, D. (2001). The work limitations questionnaire. *Medical care*, 39(1):72–85.

Li, Y. and Zhu, J. (2008). L1-norm quantile regression. *Journal of Computational and Graphical Statistics*, 17(1).

Liao, Z. (2013). Adaptive gmm shrinkage estimation with consistent moment selection. *Econometric Theory*, 29(05):857–904.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *Ann. Statist.*, 42(2):413–468.

Robson, J. and Troy, J. (1987). Nature of the maintained discharge of q, x, and y retinal ganglion cells of the cat. *JOSA A*, 4(12):2301–2307.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464.

Tang, K., Beaton, D. E., Boonen, A., Gignac, M. A. M., and Bombardier, C. (2011). Measures of work disability and productivity: Rheumatoid arthritis

specific work productivity survey (wps-ra), workplace activity limitations scale (wals), work instability scale for rheumatoid arthritis (ra-wis), work limitations questionnaire (wlq), and work productivity and activity impairment questionnaire (wpai). *Arthritis Care & Research*, 63(S11):S337–S349.

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

Tibshirani, R. (2013). Model selection: Cross validation. http://stat.cmu.edu/~ryantibs/datamining/lectures.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, pages 91–108.

Wang, H., Li, R., and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568.

Wang, L. (2003). Estimation of nonlinear berkson-type measurement error models. *Statistica Sinica*, 13(4):1201–1210.

Wang, L. (2004). Estimation of nonlinear models with berkson measurement errors. *Annals of Statistics*, pages 2559–2579.

Wang, L. and Leblanc, A. (2008). Second-order nonlinear least squares estimation. *Annals of the Institute of Statistical Mathematics*, 60(4):883–900.

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica*, 19(2):801.

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika*, 94(3):691–703.

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

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

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the degrees of freedom of the lasso. *Ann. Statist.*, 35(5):2173–2192.

Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *The Annals of Statistics*, pages 1108–1126.