The Pennsylvania State University

The Graduate School

Eberly College of Science

VARIABLE SELECTION

IN ROBUST LINEAR MODELS

A Thesis in

Statistics

by

Bo Kai

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2008

The thesis of Bo Kai was reviewed and approved* by the following:

Runze Li
Associate Professor of Statistics
Thesis Adviser

Thomas P. Hettmansperger
Professor of Statistics

Bruce G. Lindsay
Willaman Professor of Statistics
Head of the Department of Statistics

*Signatures are on file in the Graduate School.

Abstract

Variable selection plays very important roles in statistical learning. Traditional stepwise subset selection methods are widely used in practice, but they are difficult to implement when the number of predictors is large. Modern penalized likelihood estimation methods were introduced to overcome the deficiency of classical ones and to select variables and estimate their coefficients simultaneously. A number of authors have systematically studied the theoretical properties of various penalized least squares estimators. However, it is well known that the least squares estimators are badly affected by the existence of outliers in datasets. Robust penalized estimators were proposed to resist the influence of outliers. In this thesis, we consider the penalized M-estimation in linear models to achieve robustness and variable selection simultaneously. We establish the theoretical results for robust penalized estimators under general settings, i.e. general robust loss function and general penalty function. We show that the oracle property still hold for penalized M-estimators. Our finite simulation studies demonstrate satisfactory performances of the penalized M-estimators.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to express my gratitude to my advisor, Dr. Runze Li. I appreciate his broad knowledge and skills in a number of areas, his helpful ideas in our discussions and his assistance in writing this thesis. I would also like to thank Dr. Bruce G. Lindsay and Dr. Thomas P. Hettmansperger for their assistance and valuable suggestions on my work.

Finally, I would like to thank my family for the support they provided to me and in particular, my girlfriend, Jingwen Zhang, without whose love and encouragement, I would not have finished this thesis.

This thesis research has been supported by National Institute on Drug Abuse grants R21 DA024260 and P50 DA10075, and National Science Foundation grants DMS 0348869 and DMS 0722351.

Chapter 1

Introduction

Variable selection is fundamental to select important features in the high dimensional data analysis. Traditional variable selection procedures, such as the best subset selection procedures, are difficult to implement for high dimensional data due to the heavy computational burden. Fortunately, there are some modern variable selection procedures developed in the recent literature. Frank and Friedman (1993) proposed the bridge regression via the L_q penalty functions and Tibshirani (1996) proposed the Least Absolute Shrinkage and Selection Operator (LASSO) to select significant variables. Fan and Li (2001) proposed a unified variable selection framework via nonconcave penalized likelihood. All these methods are distinguished from the traditional variable selection procedures in that the methods select significant variables and estimate their coefficient simultaneously. Numerical algorithms, such as linear programming and the MM algorithm (Hunter and Li 2005) can be used to select significant features. So the computational cost can be dramatically reduced. This makes feature selection for high dimensional data feasible.

In the presence of outliers or contamination, the ordinary least squares method or likelihood based method result in biased estimates, and may lead to a

misleading conclusion. To our knowledge, only a few researches have been done for variable selection for robust linear models. Fan and Li (2001) already pointed out that the least squares estimate is not robust and an outlier-resistant loss function (such as L_1 loss or Huber's ψ-function) can be used to replace the L_2 loss to obtain robust estimators for $\boldsymbol{\beta}$. Their simulation studies showed that the proposed procedure also works well in robust regression. Wu and Liu (2007) recently demonstrated the oracle properties for the SCAD and adaptive LASSO penalized quantile regressions. Li and Zhu (2005) found the solution path of the L_1 penalized quantile regression. The results to date are only limited to specific loss function or specific penalty function. In this thesis, general penalized M-estimation are considered. Under certain regularity conditions, we show that consistency, normality and oracle property still hold for these types of estimators with general loss functions and general penalty functions. By choosing a proper loss function ρ and a suitable penalty function $p_\lambda(\cdot)$, we can achieve robustness and variable selection simultaneously.

The thesis is organized as follows. Chapter 2 presents the literature review of this thesis research. Chapter 3 discusses the methods for variable selection in robust linear regression via penalized M-estimator. We establish the asymptotic results for the penalized robust estimators under general settings. In Chapter 4, Some simulation studies are investigated to evaluate the finite sample performance

of penalized M-estimators for different penalty functions. In simulation, the proposed estimator demonstrates satisfactory performance. Finally, all the proofs are given in Chapter 5.

Chapter 2

Literature Review

This chapter presents a brief literature review of this thesis research. The proposal uses research findings from the following two topics: traditional and modern variable selection methods for regression models and robust estimators. Both are classical but active topics in modern statistics.

2.1 Variable Selection for Regression Models

Variable selection plays very important roles in statistical learning, especially in high-dimensional cases. At the initial stage of the statistical modeling, we may include a large number of prediction variables to reduce possible model biases because we do not know among which of them will have effect on the response variable. However, many of them may have little effect on the response. Therefore, a major task is to find a simple model, which is a model with as few predictors as possible while still with a good fit. Typically, simple models are desirable because they will significantly improve the prediction accuracy of the fitted model. Even when we are not sure about the complexity of the true underlying model, selecting significant variables can also improve the interpretability of a model and speed up learning process.

Suppose that a dataset contains n observations $\{(\mathbf{x}_i, y_i),\ i = 1, \cdots n\}$, where $\mathbf{x}_i = (x_{i1}, \cdots, x_{id})^T$ consists of d prediction variables for the i^{th} observation. Consider the linear regression model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \cdots n, \tag{2.1}$$

where ϵ_i are independent and identically distributed (i.i.d.) random errors with mean zero.

Denote $\mathbf{y} = (y_1, \cdots, y_n)^T$, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$ and $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^T$, then model (2.1) above can be expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2.2}$$

Here $\mathbf{X}$ is usually called the design matrix for model (2.1).

2.1.1 Classical Variable Selection Criteria

There are various variable selection criteria in the literature. Detailed reviews can be found in Breiman (1996), Shao (1997) and Miller (2002).

A selection criterion is a statistic calculated from the fitted model. In least squares settings, most of them are built on the residual sum of squares (RSS), which is defined by

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2, \tag{2.3}$$

where $\hat{y}_i$ is the predicted value for the i^{th} observation and $\hat{\mathbf{y}} = (\hat{y}_1, \cdots, \hat{y}_n)^T$. Denote RSS_p to be the residual sum of squares when there are p $(0 \leq p \leq d)$ predictors in the model.

Based on different statistical perspectives, the selection criteria include two main classes, which are prediction criteria and information (or likelihood) criteria.

Prediction sum of squares (PRESS) is a prediction based criterion proposed by Allen (1974). For a given subset of p predictors, each observation is predicted in turn from the model fitted by the other $n-1$ observations. Let $\hat{y}_{ip}$ be the predicted value for y_i, then the PRESS statistic for a particular subset of p predictors is defined as

$$PRESS_p = \sum_{i=1}^{n} (y_i - \hat{y}_{ip})^2. \tag{2.4}$$

In calculating (2.4), a different set of regression coefficients is calculated for each case with the same subset of p predictors. So it involves a large amount of computation.

In theory, it can be shown that when n is much larger than p, the PRESS statistic has an asymptotic approximation

$$PRESS_p = RSS_p \frac{n^2}{(n-p)^2}. \tag{2.5}$$

The PRESS statistic is closely related to the cross-validation (CV) approach. The idea of cross-validation is that we set a small part of the data aside

and then use the model fitted from the remainder to predict the data aside. This is done repeatedly by setting aside a different part of the data till a rotation of all the observations. If we set aside one observation each time, it is called leave-one-out cross-validation, which is exactly the PRESS. If we equally divide the whole dataset divided into K parts and leave out one part at a time, it is called the K-folded CV. Usually K is chosen to be 5 or 10. In these cases, the computation costs are much cheaper compared to the leave-one-out cross-validation, especially when the sample size n is large. These cross-validation approaches provide us a good way to estimate the prediction error of models, which will be introduced in the next subsection.

Craven and Wahba (1979) proposed the generalized cross-validation statistic, which is defined by

$$GCV = \frac{\frac{1}{n}\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{(1 - df/n)^2}, \tag{2.6}$$

where $\hat{\mathbf{y}}$ is a linear estimator in terms of $\mathbf{y}$, that is, there exists a matrix A such that $\hat{\mathbf{y}} = A\mathbf{y}$. The df in (2.6) is defined to be trace(A). When $\hat{\mathbf{y}}$ is the least squares estimator with p predictors, it is easy to see that $df = p$. Now nGCV is equal to $\frac{RSS_p}{(1 - p/n)^2}$, which is asymptotically equals to the PRESS statistic.

There is another well-known prediction based criterion named Mallows' C_p (Mallows 1973). It is defined as

$$C_p = \frac{RSS_p}{\sigma^2} - (n - 2p). \tag{2.7}$$

In practice, we use the unbiased estimate

$$\hat{\sigma}^2 = \frac{RSS_d}{n-d}, \tag{2.8}$$

for the full model to substitute σ^2 in (2.7).

Among the information criteria, the most famous two are the Akaikes information criterion (AIC, Akaike 1973, 1974) and the Bayesian information criterion (BIC, Schwarz 1978).

The AIC was developed by considering the Kullback-Leibler distance of a model from the true likelihood function. In the general case, the AIC is defined to be

$$AIC = -2\log L + 2p, \tag{2.9}$$

where L is the likelihood function. In the linear regression model with normal errors, it becomes

$$AIC = n\log(RSS_p) + 2p. \tag{2.10}$$

The BIC is defined in general to be

$$BIC = -2\log L + \log(n)p. \tag{2.11}$$

In the linear regression model with normal errors, it becomes

$$BIC = n\log(RSS_p) + \log(n)p. \tag{2.12}$$

It was shown that the BIC is a consistent criterion in the sense that if the true model exists and contains only finite parameters, the BIC criterion can determine the true model as sample size goes to infinity. On the contrary, the AIC tends to overfit the model.

Many other classical variable selection criteria is of the form:

$$-2\log L + cp, \tag{2.13}$$

where c is a regularization parameter. For example, ψ-criterion (Hannan and Quinn 1979):

$$\psi_p = -2\log L + c\log(\log(n))p$$

for some constant c. And risk inflation criterion (RIC, Foster and George 1994):

$$RIC_p = -2\log L + 2\log(d)p.$$

and many among others.

2.1.2 Variable Selection via Penalized Likelihood

Classical stepwise subset selection methods are widely used in practice, but actually they suffer from several drawbacks. First, their theoretical properties are hard to understand because they ignore stochastic errors in the variable selection process. Second, the best subset selection may become infeasible for high-dimensional due to the expensive computational cost. Third, as analyzed in Breiman (1996), subset selection methods lack of stability in terms that a small change in data could lead to a large change in the fitted equation. To overcome these deficiencies, modern penalized likelihood estimation methods were introduced gradually from the 1990s. By adding a continuous penalty term to the likelihood and then maximizing the penalized likelihood, we can achieve selecting variables and obtaining estimates simultaneously. Hence, it enables us to study theoretical properties and make statistical inferences on the model.

Penalized least squares method is a special case of the penalized likelihood, in which our aim is to minimize the least squares with some penalty. The penalized least squares function is defined to be

$$\frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2+n\sum_{j=1}^{d}p_{\lambda_j}(|\beta_j|). \tag{2.14}$$

Note that the penalty functions $p_{\lambda_j}(\cdot)$ in (2.14) are not necessarily the same for all j. For the sake of simplicity, we assume that the penalty functions are the same for all coefficients and denote it by $p_\lambda(|\cdot|)$.

The well-known ridge regression (Hoerl and Kennard 1970) is just a solution of penalized least squares. The penalty term used in ridge regression is the L_2 penalty, namely $p_\lambda(|\theta|) = \frac{\lambda}{2}|\theta|^2$. Ridge regression shrinks coefficients but does not select variables because it does not force coefficients to zeros. So actually it is not a proper method for variable selection. Frank and Friedman (1993) proposed the bridge regression via the L_q penalty functions, namely, $p_\lambda(|\theta|) = \frac{\lambda}{q}|\theta|^q$. And Tibshirani (1996) proposed the Least Absolute Shrinkage and Selection Operator (LASSO), which is equivalent to penalized least squares with L_1 penalty, to select significant variables. More recently, Fan and Li (2001) proposed a unified approach via nonconcave penalized likelihood and first introduced the oracle property. They showed that the nonconcave penalized likelihood estimators may perform as well as the oracle estimator in variable selection, that is, they work as well as if we knew the true underlying submodel in advance. About the choice of penalty functions, they pointed out that a good penalty function should result in an estimator with three nice properties:

1. Unbiasedness: The penalized estimator should be unbiased or nearly unbiased when the true parameter is large.

2. Sparsity: The penalized estimator should be a thresholding rule and set small estimates to zero.

3. Continuity: The penalized estimator should be a continuous function in data, that is, a small change in data will not result in a large change in estimates.

And they introduced a family of penalty functions that satisfy all the three properties above. The smoothly clipped absolute deviation (SCAD) penalty function is a representative among them with simple form but good performance.

Fan and Li (2001) also mentioned that there are close connections between classical stepwise subset selection and the penalized least squares methods. The classical stepwise selection methods may be viewed as special cases of penalized least squares with the so-called L_0 penalty, which is zero at point 0 and a positive constant everywhere else. Furthermore, when the design matrix is orthonormal, the penalized least squares estimators with the hard thresholding penalty function and a proper tuning parameter λ is equivalent to ones obtained by best subsets selection.

To see clearly the variable selection effect for penalized least squares, we first consider the situation where the columns in design matrix $\mathbf{X}$ are orthonormal, i.e.

$\mathbf{X}^T\mathbf{X} = I_{p\times p}$. Denote $\mathbf{z} = \mathbf{X}^T\mathbf{y}$, we have

$$\begin{aligned} & \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{d} p_\lambda(|\beta_j|) \\ = \; & \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{z}\|^2 + \frac{1}{2}\|\mathbf{z} - \boldsymbol{\beta}\|^2 + \sum_{j=1}^{d} p_\lambda(|\beta_j|) \\ = \; & \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{z}\|^2 + \sum_{j=1}^{d}\left[\frac{1}{2}(z_j - \beta_j)^2 + p_\lambda(|\beta_j|)\right]. \end{aligned} \tag{2.15}$$

The first term in (2.15) does not involve $\boldsymbol{\beta}$. So the minimization problem (2.15) is equivalent to minimizing the second term componentwise. Thus we only need to consider the following equivalent minimization problem

$$\min_\theta\{\frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|)\}. \tag{2.16}$$

Fan and Li (2001) thoroughly studied the conditions for penalties satisfying the three properties under orthonormal case.

The choice of penalty function will straightly influence the properties and performance of the resulting estimates. For example, L_2 penalty leads to ridge regression directly (Frank and Friedman 1993; Fu 1998); L_1 penalty is equivalent to LASSO, which was proposed by Tibshirani (1996); L_q penalty will lead to a bridge regression introduced by Frank and Friedman (1993); and L_0 penalty will lead to a best subset selection with AIC, BIC, etc. (which implies traditional subset selection methods are actually a special type of penalized least squares with L_0

penalty). Now let us look at some newly proposed penalty functions and their individual properties.

- SCAD penalty:

$$p_{\lambda}^{'}(|\theta|) = \lambda\{I(|\theta| \leq \lambda) + \frac{(a\lambda - |\theta|)_{+}}{(a-1)\lambda} I(|\theta| > \lambda)\}, \tag{2.17}$$

  where a is a constant that is greater than 2. SCAD penalty was proposed by Fan and Li (2001), which possesses all the three nice properties: unbiasedness, sparsity, and continuity. Another important fact argued by Fan and Li (2001) is that the SCAD enjoys the oracle property, that is it works as well as if the true underlying model is known in asymptotic sense. Actually, the SCAD is only a representative among a large family of penalties with all the three properties above. For detailed discussion, readers are referred to Fan and Li (2001).

- Adaptive LASSO penalty: The adaptive LASSO is a new penalized likelihood method newly proposed by Zou (2006). It starts from the weighted LASSO

$$\arg\min_{\boldsymbol{\beta}} \|y - X\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{d} w_j |\beta_j|, \tag{2.18}$$

and define the adaptive LASSO as

$$\arg\min_{\boldsymbol{\beta}} \|y - X\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{d} \frac{1}{|\hat{\beta}_j|^{\gamma}} |\beta_j|, \tag{2.19}$$

where $\gamma > 0$ and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \cdots, \hat{\beta}_d)^T$ is a root-n-consistent estimator of $\boldsymbol{\beta}_0$. A possible choice for $\hat{\boldsymbol{\beta}}$ is the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$. Zou (2006) showed that the adaptive LASSO estimator also has the oracle property, which is a major improvement for LASSO. Meanwhile, the adaptive LASSO estimator can be solved by the same efficient algorithm (LARS) used for solving LASSO. So it could become a favorable alternative for LASSO. In Zou (2006), he also showed that the nonnegative garotte (Breiman 1995) is close related to a special case of adaptive LASSO, and hence the nonnegative garotte is also consistent for variable selection.

- Sine-Shape (SS) penalty:

$$p_{\lambda}(|\theta|) = c\lambda^{\gamma} \left\{ \sin(\frac{\pi|\theta|}{2\lambda}) I(|\theta| \leq \lambda) + I(|\theta| > \lambda) \right\} \tag{2.20}$$

$$p'_{\lambda}(|\theta|) = c'\lambda^{\gamma-1} \cos(\frac{\pi|\theta|}{2\lambda}) I(|\theta| \leq \lambda) \tag{2.21}$$

where c and c' are constants and γ is a shape parameter [see Figure 2.1]. *This is a new penalty function proposed in this thesis.* The asymptotic properties of this sine shape penalty is similar to those of the SCAD penalty, and it

will be shown that it also enjoys the oracle property with proper chosen λ. Furthermore, simulation results show that it has very good performance in finite sample study, which will be given in Chapter 4.

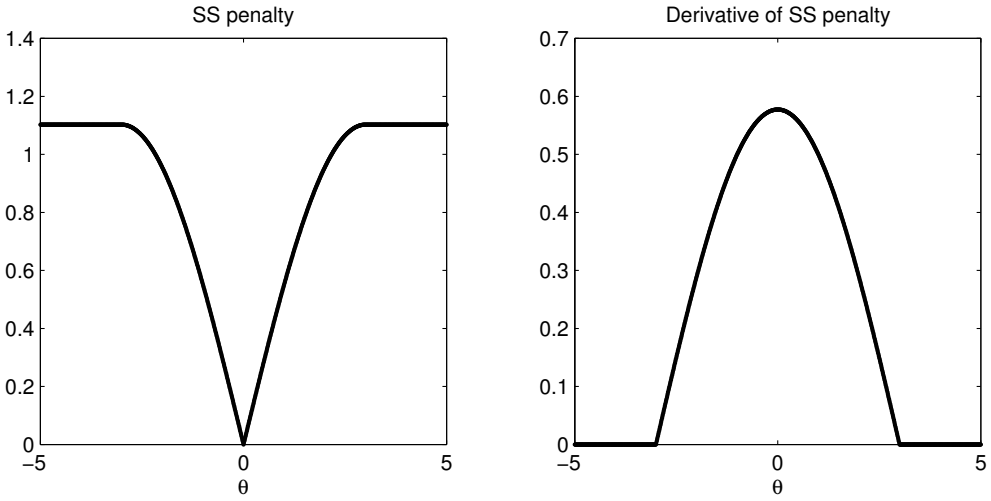

Fig. 2.1. Plots of SS penalty and its derivative with $\lambda = 3$ and $\gamma = 0.5$.

2.2 Robust Regression

It is well known that the ordinary least squares estimation (OLS) for regression is sensitive to outliers or deviation of model assumptions. Instead of OLS, we should consider robust regression if there are strong suspicions of heteroscedasticity or presence of outliers in the data. Outliers can be generated by simple operational mistakes or including a small portion of sample from a different population. The presence of outliers may make serious effect in statistical inference.

A popular alternative estimating method in a regression model that are less sensitive to outliers is to use the least absolute deviation (LAD) regression. The LAD estimator is defined by minimizing the sum of the absolute values of the residuals.

The primary purpose of robust analysis is to provide methods that are competitive with classical methods, but are not seriously affected by outliers or other small departures from model assumptions. As described in (Huber 1981, page 5), a good robust statistical procedure should possess the following desirable features:

1. It should have a reasonably good (optimal or nearly optimal) efficiency at the assumed model.

2. It should be robust in the sense that small deviation from the model assumptions should impair the performance only slightly, that is, the latter (described, say, in terms of the asymptotic variance of an estimate, or of the level and power of a test) should be close to the nominal value calculated at the model.

3. Somewhat larger deviations from the model should not cause a catastrophe.

Good reference books on robust statistics include those by Huber (1981), Hampel et al. (1986) and Rousseeuw and Leroy (1987).

2.2.1 Huber's M-estimator

Robust regression estimators were first introduced by Huber (1973, 1981) and they are well known as M-type (Maximum likelihood type) estimators. There are three major types of estimators. Except for M-type estimators, the other two are R-type (Rank tests based type) and L-type (Linear combination of order statistics) estimators. However, M-type estimators is the most popular one because of their generality, high breakdown point, and their efficiency (Huber 1981).

M-estimators are a kind of generalization of maximum likelihood estimators (MLEs). We know that MLE is to maximize $\prod_{i=1}^{n} f(\theta; x_i)$ or, equivalently, minimize $\sum_{i=1}^{n} -\log f(\theta; x_i)$. Huber proposed to generalize this to the minimization of $\sum_{i=1}^{n} \rho(\theta; x_i)$, where ρ is a function with certain properties. Thus, MLEs are a special case of M-estimators with $\rho = -\log f$.

In linear regression context, the M-estimator is defined by

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \boldsymbol{\beta}). \tag{2.22}$$

If ρ is differentiable, minimizing $\sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \boldsymbol{\beta})$ is equivalent to solve

$$\sum_{i=1}^{n} \psi(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) \mathbf{x}_i = 0, \tag{2.23}$$

where $\psi(x) = \dfrac{d\rho(x)}{dx}$. This can be done based on the following argument. Define the weight matrix $W = \text{diag}(w_i)$ with $w_i = \dfrac{\psi(y_i - \mathbf{x}_i^T\boldsymbol{\beta})}{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})}$, then (2.23) can be written as

$$\sum_{i=1}^{n} w_i(y_i - \mathbf{x}_i^T\boldsymbol{\beta})\mathbf{x}_i = 0, \tag{2.24}$$

The above equations can be combined into the following single matrix equation

$$X^T W X\boldsymbol{\beta} = X^T W\mathbf{y}. \tag{2.25}$$

Therefore, the estimator is

$$\hat{\boldsymbol{\beta}} = (X^T W X)^{-1} X^T W\mathbf{y}. \tag{2.26}$$

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

2.2.2 Alternatives Ways

In the 1980s, several alternatives to M-estimator were proposed. Rousseeuw (1984) introduced the least median of squares (LMS) and the least trimmed squares

(LTS) estimators. These estimators minimize the median and the trimmed mean of the squared residuals respectively. They are very high breakdown point estimator. However, both of these methods are inefficient, producing parameter estimates with high variability. Moreover, computing any of these estimators exactly is impractical except for small data sets. They are based on resampling techniques and their solutions are determined randomly (Rousseeuw and Leroy 1987), and then they can be even inconsistent. Another proposed solution was S-estimation (Rousseeuw 1984). This method finds a line that minimizes a robust estimate of the scale of the residuals, which is highly resistant to leverage points, and is robust to outliers in the response. But unfortunately, this method was also found to be inefficient. So in this thesis, we will mainly focus on the M-type estimators.

Chapter 3

Variable Selection in Robust Linear Regression

This chapter first discusses the methods for variable selection in robust linear regression via penalized M-estimation. And then we will establish the asymptotic results for the penalized robust estimators in general settings.

3.1 Variable Selection via Penalized M-Estimation

So far there are only limited publications in the literature related to the topic of variable selection in robust regression models. Fan and Li (2001) pointed out that the least squares estimate is not robust and an outlier-resistant loss function (such as L_1 loss or Huber's ψ-function) can be used to replace the L_2 loss to obtain robust estimators for $\boldsymbol{\beta}$. Their simulation studies showed that the proposed procedure also works well in robust regression. Koenker (2004) applied the L_1 penalty to the mixed-effect quantile regression model for longitudinal data to shrink estimates of the random effects. Li and Zhu (2005) found the solution path of the L_1 penalized quantile regression. Wu and Liu (2007) demonstrated the oracle properties for the SCAD and adaptive LASSO penalized quantile regressions. Zou and Yuan (2007) proposed a new quantile regression method called composite quantile regression (CQR) by considering the summation of multiple

different quantile regression models. They showed that the relative efficiency of CQR compared to the least squares is always greater than 70% regardless the error distribution and is equal to 95.5% under Gaussian model. They also proved that the penalized CQR estimator by using the adaptive LASSO also enjoys the oracle property.

In this thesis, we consider general penalized M-estimation. Suppose we have n bivariate observations $\{(\mathbf{x}_i, y_i),\ i = 1, \cdots n\}$, where $\mathbf{x}_i = (x_{i1}, \cdots, x_{id})^T$ are d prediction variables for the i^{th} observation, the penalized form of M-estimation is defined to be

$$\sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + n \sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{3.1}$$

Here $\rho(\cdot)$ is a general robust loss function, for example, L_1 loss or Huber's loss, etc. And $p_\lambda(\cdot)$ is a penalty function, which could be chosen from L_1 penalty, SCAD penalty, SS penalty, etc.

By minimizing the penalized M-estimation (3.1) with respect to $\boldsymbol{\beta}$, we will obtain the penalized M-estimator of $\boldsymbol{\beta}$, which performs much better than the penalized least squares estimator in the presence of outliers in the data.

3.2 Algorithm for Solving Penalized M-estimation

Assume that the robust loss function ρ is differentiable everywhere except possibly at finitely many points. And let $\psi(r) = \rho'(r)$ be the derivative of ρ,

wherever it exists. For solving the penalized M-estimation problem

$$\sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + n \sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{3.2}$$

we firstly use LQA (Fan and Li 2001) to locally approximate $p_\lambda(|\theta|)$ by a quadratic function

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{j0}|) + \frac{1}{2} \frac{p_\lambda'(|\beta_{j0}|)}{|\beta_{j0}|} (\beta_j^2 - \beta_{j0}^2), \tag{3.3}$$

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

Differentiating the objective function with respect to $\boldsymbol{\beta}$ and setting the partial derivatives to 0, produces a system of d estimating equations for the coefficients:

$$\sum_{i=1}^{n} \psi(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) x_{ij} + n \frac{p_\lambda'(|\beta_{j0}|)}{|\beta_{j0}|} \beta_j = 0, \quad j = 1, \cdots d. \tag{3.4}$$

Then we apply the reweighted least squares algorithm to solve (3.4). Let $\boldsymbol{\beta}^{(s)}$ be the parameter value at the s^{th} iteration. Let $r_i^{(s)} = y_i - \mathbf{x}_i^T \boldsymbol{\beta}^{(s)}$, $w_i^{(s)} = \psi(r_i^{(s)})/r_i^{(s)}$ and $v_j^{(s)}) = \dfrac{p_\lambda'(|\beta_j^{(s)}|)}{|\beta_j^{(s)}|}$, then the next iteration gives

$$\boldsymbol{\beta}^{(s+1)} = \left(\mathbf{X}^T \mathbf{W}^{(s)} \mathbf{X} + n \boldsymbol{\Sigma}_\lambda^{(s)} \right)^{-1} \mathbf{X}^T \mathbf{W}^{(s)} \mathbf{y}, \tag{3.5}$$

where $\mathbf{W}^{(s)} = \text{diag}\{w_i^{(s)}\}$ and $\boldsymbol{\Sigma}_\lambda^{(s)} = \text{diag}\{v_i^{(s)}\}$. The reweighted least squares algorithm usually converges quickly. But when some of the residuals are close to 0, these points receive too much weight. We adopt the modification (Fan and Li

2001) to replace the weight by $\psi(r_i^{(s)})/(r_i^{(s)} + a_n)$, where a_n is the $2n^{-1/2}$ quantile of the absolute residuals $\{|r_i|, i = 1, \cdots n\}$. A generalization via an MM algorithm proposed by Hunter and Lange (2000) can be used for solving (3.1) for more general choice of ρ functions.

3.3 Theoretical Results

In this section, we will establish asymptotic properties (including consistency, normality and oracle property) for penalized quantile regression.

Consider the linear regression model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \cdots, n. \tag{3.6}$$

Here $\mathbf{x}_i$ and $\boldsymbol{\beta}$ are d-dim vectors and have the partition $\mathbf{x}_i = (\mathbf{x}_{i1}^T, \mathbf{x}_{i2}^T)^T$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, respectively, where $\boldsymbol{\beta}_1$ has dimension $s \times 1$ and $\boldsymbol{\beta}_2$ has dimension $(d-s) \times 1$. Denote the true regression coefficients as $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$ and without loss of generality assume $\boldsymbol{\beta}_{20} = \mathbf{0}$ and $\boldsymbol{\beta}_{10}$ contains all nonzero components of $\boldsymbol{\beta}_0$. In this chapter, we only consider the situation that the dimension of the parameter space $\boldsymbol{\Omega}$ for $\boldsymbol{\beta}$ is large but finite.

First let's introduce some assumptions and notations on the penalty function $p_\lambda(\theta)$. we assume that the penalty function $p_\lambda(\theta)$ has a second order derivative at $\theta = \beta_{j0}, j = 1, \cdots, s$. And $p_\lambda(\theta)$ is nonnegative, singular at point 0 with

$p_\lambda(0) = 0$. Denote

$$
\begin{aligned}
a_n &= \max\{|p'_{\lambda_n}(|\beta_{j0}|)| : j = 1, \cdots, s\}, && (3.7)\\
b_n &= \max\{|p''_{\lambda_n}(|\beta_{j0}|)| : j = 1, \cdots, s\}, && (3.8)\\
\mathbf{b} &= \Big(p'_{\lambda_n}(|\beta_{10}|)\mathrm{sgn}(\beta_{10}), \cdots, p'_{\lambda_n}(|\beta_{s0}|)\mathrm{sgn}(\beta_{s0})\Big)^T && (3.9)\\
\Sigma &= \mathrm{diag}\{p''_{\lambda_n}(|\beta_{10}|), \cdots, p''_{\lambda_n}(|\beta_{s0}|)\}. && (3.10)
\end{aligned}
$$

Before we move on to the theoretical results , we need the following regularity conditions:

1. $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are deterministic design points. And there exists a finite and positive definite matrix V such that $\lim_{n\to\infty} V_n = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T = V$. We partition $V_{d\times d}$ into the block matrix $\begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}$ with V_{11} an $s \times s$ submatrix.

2. The regression errors ϵ_i are independent and identically distributed with a possible location shift such that $E[\psi(\epsilon_i)] = 0$ and $\sigma^2 = Var[\psi(\epsilon_i)] = E[\psi(\epsilon_i)]^2 < \infty$.

3. The robust loss function ρ is convex and differentiable everywhere except possibly finite non-differentiable points. And there exist a constant $\delta > 0$ such that

$$E[\rho(\epsilon_i - t) - \rho(\epsilon_i)] = \frac{\delta}{2}t^2\Big(1 + o(1)\Big).$$

4. There exists a constant $c_1 > 0$ such that

$$E|\psi(\epsilon_i - t) - \psi(\epsilon_i)| = c_1 t\Big(1 + o(1)\Big).$$

Furthermore, there exists constants $0 < c_2, C_1 < \infty$ such that

$$|\psi(t - s) - \psi(t)| \leq C_1$$

for any $|s| \leq c_2$.

Now we are ready to establish the theoretical results for the penalized robust estimators.

Theorem 3.1 ($\sqrt{n}$-Consistency)**.** *Assume the sample* $\{(\mathbf{x}_i, y_i), i = 1, \cdots, n\}$ *is from model (3.6) that satisfies conditions (1)—(4). Let* $L_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \boldsymbol{\beta})$ *and* $Q_n(\boldsymbol{\beta})$ *be the penalized quantile function*

$$L_n(\boldsymbol{\beta}) + n \sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|).$$

If $a_n = O(n^{-1/2})$ *and* $b_n \to 0$*,then there exists a local minimizer* $\hat{\boldsymbol{\beta}}$ *of* $Q_n(\boldsymbol{\beta})$ *such that* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$.

Theorem 3.2 (Sparsity). *Suppose the sample* $\{(\mathbf{x}_i, y_i), i = 1, \cdots, n\}$ *is from model (3.6) that satisfies conditions (1)—(4). Assume that*

$$\liminf_{n\to\infty} \liminf_{\theta\to 0^+} p'_{\lambda_n}(\theta)/\lambda_n > 0. \tag{3.11}$$

If $\lambda_n \to 0$ *and* $\sqrt{n}\lambda_n \to \infty$*, then with probability tending to 1, any* $\sqrt{n}$*-consistent estimator* $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1^T, \tilde{\boldsymbol{\beta}}_2^T)^T$ *must satisfy* $\tilde{\boldsymbol{\beta}}_1 = \mathbf{0}$.

Theorem 3.3 (Oracle Property). *Suppose the sample* $\{(\mathbf{x}_i, y_i), i = 1, \cdots, n\}$ *is independent and identically distributed from model (3.6) that satisfies conditions (1)—(4). Further assume the conditions of Theorem 3.1 and Theorem 3.2, then with probability tending to 1, the* $\sqrt{n}$*-consistent penalized M-estimator* $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^{\ T}, \hat{\boldsymbol{\beta}}_2^{\ T})^T$ *of (3.1) must satisfy that* $P(\hat{\boldsymbol{\beta}}_2 = \mathbf{0}) \to 1$ *and*

$$\sqrt{n}(\delta V_{11} + \Sigma)(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\delta V_{11} + \Sigma)^{-1}\mathbf{b}) \xrightarrow{\mathscr{D}} N(0, \sigma^2 V_{11}), \tag{3.12}$$

where $\sigma^2 = Var[\psi(\epsilon_i)]$ *and* V_{11} *is the left-top* $s \times s$ *sub-matrix of the matrix* V.

Remark. For the SS thresholding penalty function, it is flat for coefficient of magnitude larger than λ_n. So if $\lambda_n \to 0$, $a_n = 0$ and $b_n = 0$. Thus from Theorem 3.1 we can see that there exists a $\sqrt{n}$-consistent penalized M-estimator. Theorem 3.2 and Theorem 3.3 tell us if $\sqrt{n}\lambda_n \to \infty$, this estimator possesses sparsity and oracle property. Furthermore, as a special example, if the loss is the check function ρ_τ

used in quantile regression, the results in Theorem 3.3 can be simplified as

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \xrightarrow{\mathscr{D}} \mathbf{N}(\mathbf{0}, \frac{\tau(1-\tau)}{f^2(0)} V_{11}^{-1}), \tag{3.13}$$

where τ is the quantile we are interested in. This estimator performs as well as the quantile regression for estimating $\boldsymbol{\beta}_1$ knowing that $\boldsymbol{\beta}_2 = \mathbf{0}$.

Chapter 4

Simulation Studies

In this chapter, we use simulation studies to evaluate the performance of the proposed penalized M-estimator and also the performance of the newly proposed SS penalty function. The performance is measured by the median of relative model error (MRME). Recall that if $y = \mathbf{x}^T\boldsymbol{\beta} + \epsilon$, where $E(\epsilon) = 0$, then the model error is calculated by

$$ME(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\frac{1}{n}X^TX)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{4.1}$$

we report the median of the ratio of the model error of penalized estimator to the model error of un-penalized estimator (e.g. the LSE in least squares settings).

The tuning parameter λ is selected by fivefold cross-validation (CV) procedure, as suggested by Breiman (1995), Tibshirani (1996). we also tried another method: generalized cross-validation (GCV). The GCV needs less computation compared to fivefold CV because it does not refit model. In least squares settings, the GCV statistic is defined by

$$GCV(\lambda) = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}(\lambda))^2}{(1 - df(\lambda)/n)^2}. \tag{4.2}$$

And in robust settings, we use an analogue of GCV, which is defined by

$$GCV(\lambda) = \frac{\frac{1}{n}\sum_{i=1}^{n}\rho(y_i - \mathbf{x}_i^T\boldsymbol{\beta}(\lambda))}{(1 - df(\lambda)/n)^2}. \tag{4.3}$$

The performance of both procedures are similar. Therefore, we only present the results based on the GCV. For the SCAD penalty, we do not tune the parameter a and just follow Fan and Li (2001) by setting $a = 3.7$ to reduce computational cost.

Best subset procedure works here with the BIC selection criterion.

4.1 Performance of the SS Penalty in Least Squares Settings

In this study, we simulated $N = 1000$ Monte Carlo samples consisting of n observations from the model

$$y = \mathbf{x}^T\boldsymbol{\beta} + \sigma\epsilon,$$

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$. The components of $\mathbf{x}$ and ϵ come from standard normal distribution. The correlation between x_i and x_j is $\rho = 0.5$. This model was first used in Tibshirani (1996) and considered by many other authors later, such as in Fan and Li (2001), etc. We considered combinations of (n, σ) in the set $\{(20, 1),\ (40, 1),\ (40, 3),\ (100, 1),\ (100, 3)\}$. The MRME over 1000 simulated datasets are summarized and the mean and median of correct and incorrect number of 0 coefficients are also reported. Moreover, we reported the percentage of samples

that have correctly selected variables and the percentage of samples that have exactly the same estimator as the oracle estimator. The results are summarized in Table 4.1 — Table 4.5, from which, we can see that when the noise level is low, the SCAD always has the best MRME. Our newly proposed SS has the better performance in terms of selecting the correct variables, especially when the tuning parameter γ is set to 0.5. So we decide to set $\gamma = 0.5$ fixed in the section. And from these tables, we also can see that the performances of SCAD and SS are expected to be as good as that of the oracle estimator as the sample size n increases.

Table 4.1. Summaries under the LS setting: $n = 20, \sigma = 1, N = 1000$

	SS($\gamma = -5$)	SS($\gamma = 0.5$)	SCAD	Best Subset	Oracle
MRME(%):	53.699	59.941	49.232	60.403	22.305
Average number of zero coefficients					
Correct:	3.696	4.180	3.565	4.182	5.000
Incorrect:	0.008	0.008	0.009	0.009	0.000
Percentage of estimators that select the correct variables					
Percentage(%):	27.900	49.900	29.300	47.600	100.000

Table 4.2. Summaries under the LS setting: $n = 40, \sigma = 1, N = 1000$

	SS($\gamma = -5$)	SS($\gamma = 0.5$)	SCAD	Best Subset	Oracle
MRME(%):	47.712	47.922	42.573	48.250	31.098
Average number of zero coefficients					
Correct:	4.579	4.629	3.958	4.622	5.000
Incorrect:	0.000	0.000	0.000	0.000	0.000
Percentage of estimators that select the correct variables					
Percentage(%):	66.200	69.600	44.200	68.600	100.000

Table 4.3. Summaries under the LS setting: $n = 40, \sigma = 3, N = 1000$

	SS($\gamma = -5$)	SS($\gamma = 0.5$)	SCAD	Best Subset	Oracle
MRME(%):	63.826	72.437	69.092	64.915	30.782
Average number of zero coefficients					
Correct:	3.800	4.424	3.724	4.569	5.000
Incorrect:	0.173	0.304	0.162	0.341	0.000
Percentage of estimators that select the correct variables					
Percentage(%):	28.100	45.000	22.300	50.700	100.000

Table 4.4. Summaries under the LS setting: $n = 100, \sigma = 1, N = 1000$

	SS($\gamma = -5$)	SS($\gamma = 0.5$)	SCAD	Best Subset	Oracle
MRME(%):	42.410	42.410	40.603	42.577	32.775
Average number of zero coefficients					
Correct:	4.827	4.827	4.247	4.828	5.000
Incorrect:	0.000	0.000	0.000	0.000	0.000
Percentage of estimators that select the correct variables					
Percentage(%):	85.000	85.000	56.400	84.600	100.000

Table 4.5. Summaries under the LS setting: $n = 100, \sigma = 3, N = 1000$

	SS($\gamma = -5$)	SS($\gamma = 0.5$)	SCAD	Best Subset	Oracle
MRME(%):	46.401	43.843	49.665	43.853	34.730
Average number of zero coefficients					
Correct:	4.053	4.830	4.212	4.828	5.000
Incorrect:	0.006	0.037	0.009	0.039	0.000
Percentage of estimators that select the correct variables					
Percentage(%):	42.800	82.800	50.100	82.300	100.000

4.2 Performance of Penalized M-estimator

In this study, we simulated $N = 100$ Monte Carlo samples consisting of n observations from the model

$$y = \mathbf{x}^T\boldsymbol{\beta} + \sigma\epsilon,$$

where $\boldsymbol{\beta}$ and $\mathbf{x}$ are the same as those in the previous study. The error term ϵ is drawn from the mixture of the standard normal distribution and the standard Cauchy distribution with equal percentage. We considered combinations of (n, σ) in the set $\{(40, 1),\ (40, 3),\ (60, 1),\ (60, 3)\}$. Detailed results are summarized in Table 4.6 — Table 4.9, from which, we can see that when the noise level is low, the SS estimator has the best MRME, which is very close to the one from the oracle estimator. And the percentage to select the correct variables are above 90%. When the noise level is high, all of the estimators have competitive performance. From these tables, we also can see that the performances of SCAD and SS have better performance as the sample size n increases.

Table 4.6. Summaries under the robust setting: $n = 40, \sigma = 1, N = 100$

	SS($\gamma = 0.5$)	SCAD	Best Subset	Oracle
MRME(%):	36.222	48.138	41.417	28.172
Average number of zero coefficients				
Correct:	4.980	4.640	4.980	5.000
Incorrect:	0.070	0.060	0.330	0.000
Percentage of estimators that select the correct variables				
Percentage(%):	91.000	66.000	76.000	100.000

Table 4.7. Summaries under the robust setting: $n = 40, \sigma = 3, N = 100$

	SS($\gamma = 0.5$)	SCAD	Best Subset	Oracle
MRME(%):	133.178	136.668	127.085	21.480
Average number of zero coefficients				
Correct:	4.640	4.580	4.900	5.000
Incorrect:	1.100	1.100	1.540	0.000
Percentage of estimators that select the correct variables				
Percentage(%):	12.000	10.000	7.000	100.000

Table 4.8. Summaries under the robust setting: $n = 60, \sigma = 1, N = 100$

	SS($\gamma = 0.5$)	SCAD	Best Subset	Oracle
MRME(%):	29.797	43.569	32.964	29.605
Average number of zero coefficients				
Correct:	5.000	4.680	4.990	5.000
Incorrect:	0.010	0.050	0.160	0.000
Percentage of estimators that select the correct variables				
Percentage(%):	99.000	71.000	89.000	100.000

Table 4.9. Summaries under the robust setting: $n = 60, \sigma = 3, N = 100$

	SS($\gamma = 0.5$)	SCAD	Best Subset	Oracle
MRME(%):	159.239	163.487	171.496	25.898
Average number of zero coefficients				
Correct:	4.870	4.720	4.980	5.000
Incorrect:	0.990	0.980	1.450	0.000
Percentage of estimators that select the correct variables				
Percentage(%):	19.000	16.000	9.000	100.000

Chapter 5

Proofs

This chapter will give the proofs in detail for all the results obtained in chapter 3. Before we prove the main results, we need the following lemmas.

Lemma 5.1. *Let* $L_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T\boldsymbol{\beta})$ *and* $G_n(\mathbf{u}) = L_n(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) - L_n(\boldsymbol{\beta}_0)$, *where* $\mathbf{u} \in \mathbb{R}^p$. *Then under regularity conditions (1)—(4), for any fixed* $\mathbf{u}$, *we have*

$$G_n(\mathbf{u}) = -\sqrt{n}W_n^T\mathbf{u} + \frac{\delta}{2}\mathbf{u}^T V\mathbf{u} + o_p(1), \tag{5.1}$$

where $W_n = \frac{1}{n}\sum_{i=1}^{n}\psi(\epsilon_i)\mathbf{x}_i$.

Proof of Lemma 5.1. By the differentiability of $\rho(\cdot)$, the identity

$$\rho(x-t) - \rho(x) = -t\psi(x) + \int_0^t \Big(\psi(x-s) - \psi(x)\Big)ds \tag{5.2}$$

holds at all the differential points of $\rho(\cdot)$. Apply identity (5.2), we have

$$\begin{aligned} G_n(\mathbf{u}) &= L_n(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) - L_n(\boldsymbol{\beta}_0) \\ &= \sum_{i=1}^{n}\Big[\rho(\epsilon_i - n^{-1/2}\mathbf{x}_i^T\mathbf{u}) - \rho(\epsilon_i)\Big] \\ &= \sum_{i=1}^{n}\left[-n^{-1/2}\mathbf{x}_i^T\mathbf{u}\psi(\epsilon_i) + \int_0^{n^{-1/2}\mathbf{x}_i^T\mathbf{u}}\Big(\psi(\epsilon_i - s) - \psi(\epsilon_i)\Big)ds\right] \end{aligned}$$

$$
\begin{aligned}
&= -n^{-1/2}\left(\sum_{i=1}^{n}\psi(\epsilon_i)\mathbf{x}_i\right)^T\mathbf{u} + \sum_{i=1}^{n}\int_0^{n^{-1/2}\mathbf{x}_i^T\mathbf{u}}\Big(\psi(\epsilon_i - s) - \psi(\epsilon_i)\Big)ds \\
&= -\sqrt{n}W_n^T\mathbf{u} + \sum_{i=1}^{n}B_{n,i}, \qquad (5.3)
\end{aligned}
$$

which implies $E[G_n(\mathbf{u})] = \sum_{i=1}^{n}E(B_{n,i})$. Thus, we can rewrite $G_n(\mathbf{u})$ as

$$
\begin{aligned}
G_n(\mathbf{u}) &= -\sqrt{n}W_n^T\mathbf{u} + \sum_{i=1}^{n}B_{n,i} \\
&= -\sqrt{n}W_n^T\mathbf{u} + E[G_n(\mathbf{u})] + \sum_{i=1}^{n}\Big(B_{n,i} - E(B_{n,i})\Big) \\
&= -\sqrt{n}W_n^T\mathbf{u} + E[G_n(\mathbf{u})] + R_n. \qquad (5.4)
\end{aligned}
$$

By assumption 3, we can calculate $E[G_n(\mathbf{u})]$ in the form of

$$
\begin{aligned}
E[G_n(\mathbf{u})] &= \sum_{i=1}^{n}E[\rho(\epsilon_i - n^{-1/2}\mathbf{x}_i^T\mathbf{u}) - \rho(\epsilon_i)] \\
&= \sum_{i=1}^{n}\left[\frac{\delta}{2}(n^{-1/2}\mathbf{x}_i^T\mathbf{u})^2\Big(1 + o(1)\Big)\right] \\
&= \frac{\delta}{2}n^{-1}\mathbf{u}^T(\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{x}_i)\mathbf{u}\Big(1 + o(1)\Big) \\
&= \frac{\delta}{2}\mathbf{u}^T V_n\mathbf{u}\Big(1 + o(1)\Big).
\end{aligned}
$$

Plug it back into (5.4), we have

$$
G_n(\mathbf{u}) = -\sqrt{n}W_n^T\mathbf{u} + \frac{\delta}{2}\mathbf{u}^T V_n\mathbf{u}\Big(1 + o(1)\Big) + R_n. \qquad (5.5)
$$

By the definition of R_n, we know that $E(R_n) = 0$. And for fixed $\mathbf{u}$, we have

$$
\begin{aligned}
& Var\left(R_n\right) = E(R_n^2) \\
= \;& E\left[\sum_{i=1}^{n}\Big(B_{n,i}(\mathbf{u}) - E(B_{n,i}(\mathbf{u}))\Big)\right]^2 \\
= \;& \sum_{i=1}^{n} E\left[\Big(B_{n,i}(\mathbf{u}) - E(B_{n,i}(\mathbf{u}))\Big)\right]^2 \\
= \;& \sum_{i=1}^{n} E\left[\int_0^{n^{-1/2}\mathbf{x}_i^T\mathbf{u}}\Big(\psi(\epsilon_i - s) - \psi(\epsilon_i)\Big)ds - E\int_0^{n^{-1/2}\mathbf{x}_i^T\mathbf{u}}\Big(\psi(\epsilon_i - s) - \psi(\epsilon_i)\Big)ds\right]^2 \\
\leq \;& \sum_{i=1}^{n} E\left[\int_0^{n^{-1/2}\mathbf{x}_i^T\mathbf{u}}\Big(\psi(\epsilon_i - s) - \psi(\epsilon_i)\Big)ds\right]^2 \\
\leq \;& \sum_{i=1}^{n} C_1 n^{-1/2}\mathbf{x}_i^T\mathbf{u}E\Big|\int_0^{n^{-1/2}\mathbf{x}_i^T\mathbf{u}}\Big(\psi(\epsilon_i - s) - \psi(\epsilon_i)\Big)ds\Big| \\
\leq \;& C_1 n^{-1/2}\Big(\max_i\|\mathbf{x}_i\|\Big)\|\mathbf{u}\|\sum_{i=1}^{n}\int_0^{|n^{-1/2}\mathbf{x}_i^T\mathbf{u}|} E\left|\Big(\psi(\epsilon_i - s) - \psi(\epsilon_i)\Big)\right|ds \\
\leq \;& C_1\left(\frac{\max_i\|\mathbf{x}_i\|}{\sqrt{n}}\right)\|\mathbf{u}\|\sum_{i=1}^{n}\int_0^{|n^{-1/2}\mathbf{x}_i^T\mathbf{u}|} c_1 s\Big(1 + o(1)\Big)ds \\
= \;& C_1\left(\frac{\max_i\|\mathbf{x}_i\|}{\sqrt{n}}\right)\|\mathbf{u}\|\left(\frac{c_1}{2}\mathbf{u}^T V_n\mathbf{u}\Big(1 + o(1)\Big)\right) \\
\to \;& 0. \qquad (5.6)
\end{aligned}
$$

Therefore, $R_n = o_p(1)$. Combined with $V_n \to V$, we have

$$
\begin{aligned}
G_n(\mathbf{u}) &= -\sqrt{n}W_n^T\mathbf{u} + \frac{\delta}{2}\mathbf{u}^T V_n\mathbf{u}\Big(1 + o(1)\Big) + R_n \\
&= -\sqrt{n}W_n^T\mathbf{u} + \frac{\delta}{2}\mathbf{u}^T V\mathbf{u} + o_p(1), \qquad (5.7)
\end{aligned}
$$

which completes the proof. □

We also need the well known convexity lemma to strengthen pointwise convergence to uniform convergence.

Lemma 5.2 (Convexity Lemma)**.** *Let* $\{g_n(\mathbf{u}) : \mathbf{u} \in \mathbf{U}\}$ *be a sequence of random convex functions defined on a convex, open subset* $\mathbf{U}$ *of* $\mathbb{R}^p$*. Suppose* $g(\cdot)$ *is a real-valued function on* $\mathbf{U}$ *satisfying* $g_n(\mathbf{u}) \xrightarrow{\mathscr{P}} g(\mathbf{u})$ *for each* $\mathbf{u} \in \mathbf{U}$*. Then for each compact subset* $\mathbf{K}$ *of* $\mathbf{U}$*,*

$$\sup_{\mathbf{u}\in\mathbf{K}} |g_n(\mathbf{u}) - g(\mathbf{u})| \xrightarrow{\mathscr{P}} 0. \tag{5.8}$$

The function $g(\cdot)$ *is necessarily convex on* $\mathbf{U}$*.*

There are many versions of proofs for this lemma. To save space, we skip the proof here. Interested readers are referred to Andersen and Gill (1982).

Now we are ready to state the proof of Theorem 3.1.

Proof of Theorem 3.1. Use the same strategy as in Fan and Li (2001). We would like to show that for any given $\epsilon > 0$, there exists a large constant C such that

$$P\{\sup_{\|\mathbf{u}\|=C} Q(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) > Q(\boldsymbol{\beta}_0)\} \geq 1 - \epsilon \tag{5.9}$$

which implies that with probability at least $1 - \epsilon$ there exists a local minimum in the ball $\{\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u} : \|\mathbf{u}\| \leq C\}$. Hence, there exists a local minimizer $\hat{\boldsymbol{\beta}}$ of $Q_n(\boldsymbol{\beta})$ such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2})$.

Because $p_{\lambda_n}(0) = 0$ and $p_{\lambda_n}(|\beta_j|) > 0$ if $\beta_j \neq 0$, so

$$\begin{aligned} D(\mathbf{u}) &= Q(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) - Q(\boldsymbol{\beta}_0) \\ &= L_n(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) - L_n(\boldsymbol{\beta}_0) + n\sum_{j=1}^{d}\{p_{\lambda_n}(|\beta_{j0} + n^{-1/2}\mathbf{u}|) - p_{\lambda_n}(|\beta_{j0}|)\} \\ &\geq G_n(\mathbf{u}) + n\sum_{j=1}^{s}\{p_{\lambda_n}(|\beta_{j0} + n^{-1/2}\mathbf{u}|) - p_{\lambda_n}(|\beta_{j0}|)\}, \end{aligned} \tag{5.10}$$

By Lemma 5.1,

$$G_n(\mathbf{u}) = -\sqrt{n}W_n^T\mathbf{u} + \frac{\delta}{2}\mathbf{u}^T V\mathbf{u} + o_p(1) \tag{5.11}$$

for any fixed $\mathbf{u}$. Note that $\frac{\delta}{2}\mathbf{u}^T V\mathbf{u}$ is a convex function of $\mathbf{u}$, by applying the Convexity Lemma to $G_n(\mathbf{u}) + \sqrt{n}W_n^T\mathbf{u}$, we can strengthens pointwise convergence to uniform convergence on any compact subset of $\mathbb{R}^d$. And by the standard argument on the Taylor expansion of the penalty function, we have

$$\begin{aligned} & n\sum_{j=1}^{s}\{p_{\lambda_n}(|\beta_{j0} + n^{-1/2}\mathbf{u}|) - p_{\lambda_n}(|\beta_{j0}|)\} \\ = & \sum_{j=1}^{s}\left\{p'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0})\sqrt{n}u_j + p''_{\lambda_n}(|\beta_{j0}|)\frac{u_j^2}{2}(1 + o_p(1))\right\} \end{aligned}$$

So

$$\begin{aligned} D(\mathbf{u}) \geq & -\sqrt{n}W_n^T\mathbf{u} + \frac{\delta}{2}\mathbf{u}^T V\mathbf{u} + o_p(1) \\ & + \sum_{j=1}^{s}\left(p'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0})\sqrt{n}u_j + p''_{\lambda_n}(|\beta_{j0}|)\frac{u_j^2}{2}(1 + o_p(1))\right). \end{aligned} \tag{5.12}$$

Note that $\sqrt{n}W_n = O_p(1)$. Thus, the first term on the right hand side of (5.12) is on the order of $O_p(1)$. By convexity lemma and choosing a sufficiently large C, the second term dominates the first term uniformly in $\|\mathbf{u}\| = C$. Note that the third term in (5.12) is bounded by

$$\sqrt{sn}a_n\|\mathbf{u}\| + b_n\|\mathbf{u}\|^2. \tag{5.13}$$

This is also dominated by the second term in (5.12). Thus, by choosing a sufficiently large C, the second term dominates all the other terms in (5.12), which means (5.9) holds. This completes the proof of the theorem. □

Proof of Theorem 3.2. It is sufficient to show that with probability tending to 1, for any $\sqrt{n}$-consistent estimator $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1^T, \tilde{\boldsymbol{\beta}}_2^T)^T$ with $\|\tilde{\boldsymbol{\beta}}_2\| > 0$, we have $Q((\tilde{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T) - Q((\tilde{\boldsymbol{\beta}}_1^T, \tilde{\boldsymbol{\beta}}_2^T)^T) < 0$.

$$\begin{aligned}
& Q((\tilde{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T) - Q((\tilde{\boldsymbol{\beta}}_1^T, \tilde{\boldsymbol{\beta}}_2^T)^T) \\
= \; & [Q((\tilde{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T) - Q((\boldsymbol{\beta}_{10}^T, \mathbf{0}^T)^T)] - [Q((\tilde{\boldsymbol{\beta}}_1^T, \tilde{\boldsymbol{\beta}}_2^T)^T - Q((\boldsymbol{\beta}_{10}^T, \mathbf{0}^T)^T))] \\
= \; & G_n(\sqrt{n}((\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})^T, \mathbf{0}^T)^T) - G_n(\sqrt{n}((\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})^T, \tilde{\boldsymbol{\beta}}_2^T)^T) - n\sum_{j=s+1}^{d} p_{\lambda_n}(|\tilde{\beta}_j|)
\end{aligned}$$

Denote $\mathbf{u}_1 = \sqrt{n}((\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})^T, \mathbf{0}^T)^T$ and $\mathbf{u}_2 = \sqrt{n}((\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})^T, \tilde{\boldsymbol{\beta}}_2^T)^T$. Then by Lemma 5.1,

$$Q((\tilde{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T) - Q((\tilde{\boldsymbol{\beta}}_1^T, \tilde{\boldsymbol{\beta}}_2^T)^T)$$

$$
\begin{aligned}
&= G_n(\mathbf{u}_1) - G_n(\mathbf{u}_2) - n\sum_{j=s+1}^{d} p_{\lambda_n}(|\tilde{\beta}_j|) \\
&= -\sqrt{n}W_n^T\mathbf{u}_1 + \frac{\delta}{2}\mathbf{u}_1^T V\mathbf{u}_1 - (-\sqrt{n}W_n^T\mathbf{u}_2 + \frac{\delta}{2}\mathbf{u}_2^T V\mathbf{u}_2) - n\sum_{j=s+1}^{d} p_{\lambda_n}(|\tilde{\beta}_j|) + o_p(1) \\
&= \sqrt{n}W_n^T\sqrt{n}(\mathbf{0}^T, \tilde{\boldsymbol{\beta}}_2^T)^T + \frac{\delta}{2}\mathbf{u}_1^T V\mathbf{u}_1 - \frac{\delta}{2}\mathbf{u}_2^T V\mathbf{u}_2 - n\sum_{j=s+1}^{d} p_{\lambda_n}(|\tilde{\beta}_j|) + o_p(1)
\end{aligned}
$$

We know that $\sqrt{n}W_n^T = O_p(1)$. And the condition $\tilde{\boldsymbol{\beta}}$ implies $\mathbf{u}_1 = O_p(1)$, $\mathbf{u}_2 = O_p(1)$ and $\tilde{\boldsymbol{\beta}}_2 = O_p(n^{-1/2})$. Note that

$$
\begin{aligned}
& n\sum_{j=s+1}^{d} p_{\lambda_n}(|\tilde{\beta}_j|) \\
\geq\ & n\sum_{j=s+1}^{d}\left(\lambda_n \liminf_{n\to\infty}\liminf_{\theta\to 0^+}\frac{p'_{\lambda_n}(\theta)}{\lambda_n}|\tilde{\beta}_j| + o_p(|\tilde{\beta}_j|)\right) \\
=\ & \sqrt{n}\lambda_n\left(\liminf_{n\to\infty}\liminf_{\theta\to 0^+}\frac{p'_{\lambda_n}(\theta)}{\lambda_n}\right)\left(\sqrt{n}\sum_{j=s+1}^{d}|\tilde{\beta}_j|\right)\Big(1 + o_p(1)\Big) \\
\to\ & \infty
\end{aligned}
$$

Therefore, we have

$$
Q((\tilde{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T) - Q((\tilde{\boldsymbol{\beta}}_1^T, \tilde{\boldsymbol{\beta}}_2^T)^T) < 0
$$

as $n \to \infty$ with probability tending to 1. This completes the proof. □

Proof of Theorem 3.3. From the proofs of Theorem 3.1 and Theorem 3.2, we know that $\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})$ is the minimizer of

$$Q(((\boldsymbol{\beta}_{10} + \frac{\boldsymbol{\theta}}{\sqrt{n}})^T, \mathbf{0}^T)^T) - Q((\boldsymbol{\beta}_{10}^T, \mathbf{0}^T)^T),$$

which is equivalent to

$$G_n((\boldsymbol{\theta}^T, \mathbf{0}^T)^T) + n\sum_{j=1}^{s}\left(p_{\lambda_n}(|\beta_{j0} + \frac{\theta_j}{\sqrt{n}}|) - p_{\lambda_n}(|\beta_{j0}|)\right),$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_s)^T \in \mathbb{R}^s$.

Lemma 5.1 and the Convexity Lemma imply that

$$G_n((\boldsymbol{\theta}^T, \mathbf{0}^T)^T) = -\sqrt{n}W_{n1}^T\boldsymbol{\theta} + \frac{\delta}{2}\boldsymbol{\theta}^T V_{11}\boldsymbol{\theta} + o_p(1)$$

uniformly in any compact subset of $\mathbb{R}^s$. Here W_{n1} is vector containing the first s elements of W_n. And by Taylor expansion,

$$\begin{aligned} & n\sum_{j=1}^{s}\left(p_{\lambda_n}(|\beta_{j0} + \frac{\theta_j}{\sqrt{n}}|) - p_{\lambda_n}(|\beta_{j0}|)\right) \\ = \; & n\sum_{j=1}^{s}\left(p'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0})\frac{\theta_j}{\sqrt{n}} + p''_{\lambda_n}(|\beta_{j0}|)\frac{\theta_j^2}{2n} + o_p(n^{-1})\right) \\ = \; & \sqrt{n}\mathbf{b}^T\boldsymbol{\theta} + \boldsymbol{\theta}^T\Sigma\boldsymbol{\theta}/2 + o_p(1). \end{aligned}$$

Therefore, $\hat{\boldsymbol{\theta}} = \sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})$ is the minimizer of

$$D(\boldsymbol{\theta}) = -\sqrt{n}W_{n1}^T\boldsymbol{\theta} + \frac{\delta}{2}\boldsymbol{\theta}^T V_{11}\boldsymbol{\theta} + \sqrt{n}\mathbf{b}^T\boldsymbol{\theta} + \frac{1}{2}\boldsymbol{\theta}^T\Sigma\boldsymbol{\theta} + o_p(1),$$

which leads to

$$\hat{\boldsymbol{\theta}} = (\delta V_{11} + \Sigma)^{-1}\sqrt{n}(W_{n1}^T - b) + o_p(1).$$

By the central limit theorem,

$$\sqrt{n}W_{n1} \xrightarrow{\mathscr{D}} N(0, \sigma^2 V_{11})$$

It follows by Slutsky's theorem that

$$\sqrt{n}(\delta V_{11} + \Sigma)(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\delta V_{11} + \Sigma)^{-1}\mathbf{b}) \xrightarrow{\mathscr{D}} N(0, \sigma^2 V_{11}).$$

This completes the proof. □

Chapter 6

Conclusion and Further Research

Fan and Li (2001) introduced the concept of oracle estimator and proposed penalized likelihood methods to select significant variables. By using the SCAD penalty in Fan and Li (2001), the resulting estimators possess the oracle property, which means that they work as well as if the correct submodel were known.

This thesis is a natural extension of Fan and Li (2001). We know that likelihood method depends on the underlying error distribution. Therefore, it is not robust and influenced by outliers or contamination. If we replace the loss function by a robust one in the penalized form, we can achieve robustness and variable selection simultaneously. We have shown that under certain regularity conditions the oracle property still holds for the proposed penalized M-estimator. In this thesis, we also propose a new penalty function (SS), which has satisfactory performance in our simulation studies. However, as we see in the literature, there does not exist a penalty function that can dominate all the others. So how to choose an optimal penalty function might be an open problem for future research.

Bibliography

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *Second International Symposium on Information Theory*, pp. 267–281.

Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

Allen, D. M. (1974), "The Relationship between Variable Selection and Data Agumentation and a Method for Prediction," *Technometrics*, 16, 125–127.

Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, 37, 373–384.

Breiman, L. (1996), "Heuristics of Instability and Stabilization in Model Selection," *Annals of Statistics*, 24, 2350–2383.

Craven, P. and Wahba, G. (1979), "Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numer. Math*, 31, 377–403.

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.

Foster, D. P. and George, E. I. (1994), "The Risk Inflation Criterion for Multiple Regression," *The Annals of Statistics*, 22, 1947–1975.

Frank, I. E. and Friedman, J. H. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35, 109–135.

Fu, W. J. (1998), "Penalized regressions: The bridge versus the lasso." *Jounral of Computational and Graphical Statistics*, 7, 397–416.

Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986), *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York.

Hannan, E. J. and Quinn, B. G. (1979), "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society. Series B (Methodological)*, 41, 190–195.

Hoerl, A. E. and Kennard, R. W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67.

Huber, P. J. (1973), "Robust Regression: Asymptotics, Conjectures and Monte Carlo," *The Annals of Statistics*, 1, 799–821.

Huber, P. J. (1981), *Robust Statistic*, John Wiley & Sons.

Hunter, D. R. and Lange, K. (2000), "Quantile Regression Via an MM Algorithm," *Journal of Computational and Graphical Statistics*, 9, 60–77.

Hunter, D. R. and Li, R. (2005), "Variable Selection Using Mm Algorithms," *The Annals of Statistics*, 33, 1617–1642.

Koenker, R. (2004), "Quantile regression for longitudinal data," *Journal of Multivariate Analysis*, 91, 74–89.

Li, Y. and Zhu, J. (2005), "L_1-norm Quantile Regression," Submitted.

Mallows, C. L. (1973), "Some comments on C_p," *Technometrics*, 15, 661–675.

Miller, A. J. (2002), *Subset Selection in Regression*, Chapman & HALL/CRC, New York, second edn.

Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871–880.

Rousseeuw, P. J. and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, John Wiley & Sons.

Schwarz, G. (1978), "Estimating the Dimension of a Model." *The Annals of Statistics*, 19, 461–464.

Shao, J. (1997), "An Asymptotic Theory for Linear Model Selection." *Statistica Sinica*, 7, 221–264.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the American Statistical Association*, 58, 267–288.

Wu, Y. and Liu, Y. (2007), "Variable Selection in Quantile Regression," *Statistica Sinica*, To appear.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H. and Yuan, M. (2007), "Composite Quantile Regression and the Oracle Model Selection Theory," *The Annals of Statistics*, To appear.