Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties *

Jianqing Fan Runze Li

Abstract

Variable selection is fundamental to high-dimensional statistical modeling, including nonparametric regression. Many approaches in use are stepwise selection procedures, which can be expensive in computation and ignore stochastic errors in the variable selection process. In this paper, penalized likelihood approaches are proposed to handle these kinds of problems. The proposed methods select variables and estimate coefficients simultaneously. Hence they enable us to construct confidence intervals for estimated parameters. The proposed approaches distinguish from others in that the penalty functions are symmetric, nonconcave on $(0, \infty)$, and have singularities at the origin in order to produce sparse solutions. Further, the penalty functions should be bounded by a constant in order to reduce bias and satisfy certain conditions in order to yield continuous solutions. A new algorithm is proposed for optimizing penalized likelihood functions. The proposed ideas are widely applicable. They are readily applied to a variety of parametric models such as generalized linear models and robust regression models. They can also be easily applied to nonparametric modeling by using wavelets and splines. Rates of convergence of the proposed penalized likelihood estimators are established. Further, with proper choice of regularization parameters, we have shown that the proposed estimators perform as well as the oracle procedure in variable selection; namely, they work as well as if the correct submodel were known. Our simulation shows that the newly proposed methods compare favorably with other variable selection techniques. Furthermore, the standard error formulas are tested to be accurate enough for practical applications.

Key Words: Hard thresholding, LASSO, nonnegative garrote, penalized likelihood, oracle estimator, SCAD, soft thresholding.

Abbreviated title: Nonconcave Penalized Likelihood

Jianqing Fan is Professor of Statistics, Department of Statistics, Chinese University of Hong Kong and professor, Department of Statistics, University of California, Los Angeles, CA 90095. Runze Li is assistant professor, Department of Statistics, Pennsylvania State University, University Park, PA 16802-2111. Fan's research was partially supported by NSF grants DMS-0196041 and DMS-9977096 and a grant from University of California at Los Angeles. Li's research was supported by a NSF grant DMS-0102505. The authors would like to thanks the associate editor and referees for constructive comments that led to improve substantially an earlier draft of the paper.

1 Introduction

Variable selection is an important topic in linear regression analysis. In practice, a large number of predictors are usually introduced at the initial stage of modeling in order to attenuate possible modeling biases. On the other hand, to enhance predictability and to select significant variables, statisticians usually use stepwise deletion and subset selection. While they are practically useful, these selection procedures ignore stochastic errors inherited in the stages of variable selections. Hence, their theoretical properties are somewhat hard to understand. Further, the best subset variable selection suffers from several drawbacks, and the most severe one of these is its lack of stability as analyzed for instance by Breiman (1996). In an attempt to automatically and simultaneously select variables, we propose a unified approach via penalized least squares, retaining good features of both subset selection and the ridge regression. The penalty functions have to be singular at the origin in order to produce sparse solutions (many estimated coefficients are zero), to satisfy certain conditions to produce continuous models (for stability of model selection) and to be bounded by a constant to produce nearly unbiased estimates for large coefficients. The bridge regression proposed in Frank and Friedman (1993) and the LASSO proposed by Tibshirani (1996, 1997) are members of the penalized least squares, though their associated L_q penalty functions do not satisfy all of the three required properties above.

The penalized least squares idea can naturally be extended to likelihood based models in various statistical contexts. Our approaches distinguish from traditional methods (usually quadratic penalty) in that the penalty functions are symmetric, convex on $(0, \infty)$ (rather than concave for the negative quadratic penalty in the penalized likelihood situation) and possess singularities at the origin. A few penalty functions are discussed. These allow statisticians to select a penalty function to enhance the predictive power of a model and engineers to sharpen noisy images. Optimizing a penalized likelihood is challenging, since the target function is a high-dimensional nonconcave function with singularities. A new and generic algorithm is proposed. This yields a unified variable selection procedure. A standard error formula for estimated coefficients is obtained by using a sandwich formula. The formula is tested accurately enough for practical purpose, even when the sample size is very moderate. The proposed procedures are compared with various other variable selection approaches. The results indicate the favorable performance of the newly proposed procedures.

Unlike the traditional variable selection procedures, the sampling properties on the penalized likelihood can be established. We will demonstrate how the rates of convergence for the penalized likelihood estimators depend on the regularization parameter. We will further show that the penalized likelihood

estimators perform as well as the oracle procedure in terms of selecting the correct model, when the regularization parameter is appropriately chosen. In other words, when the true parameters have some zero components, they are estimated as 0 with probability tending to one, and the non-zero components are estimated as well as if the correct submodel were known. This improves the accuracy for estimating not only the null components, but also the non-null components. In short, the penalized likelihood estimators work as well as if the correct submodel were known in advance. The significance of this is that the proposed procedures outperform the maximum likelihood estimator and perform as well as we hope. This is very analogous to the super-efficiency phenomenon in the Hodges example (see page 405 of Lehmann 1983).

The proposed penalized likelihood method can be applied readily to high-dimensional nonparametric modeling. After approximating regression functions using splines or wavelets, it remains very critical to select significant variables (terms in the expansion) to efficiently represent unknown functions. In a series of work by Stone and his collaborators (see Stone *et al.* 1997), they modify traditional variable selection approaches to select useful spline subbasis. It remains very challenging to understand the sampling properties of these data-driven variable selection techniques. Penalized likelihood approaches, outlined in Wahba (1990), Green and Silverman (1994) and references therein, are based on a quadratic penalty. They reduce the variability of estimators via the ridge regression. In wavelet approximations, Donoho and Johnstone (1994a) select significant subbases (terms in the wavelet expansion) via thresholding. Our penalized likelihood approach can be directly applied to these problems (see Fan and Antoniadis, 1999). Because we select variables and estimate parameters simultaneously, the sampling properties of such a data-driven variable selection method can be established.

In Section 2, we discuss the relationship between the penalized least squares and the subset selection when design matrices are orthonormal. In Section 3 we then extend the penalized likelihood approach discussed in Section 2 to various parametric regression models, including traditional linear regression models, robust linear regression models and generalized linear models. The asymptotic properties of the penalized likelihood estimators are established in §3.2. Based on local quadratic approximations, a unified iterative algorithm for finding penalized likelihood estimators is proposed in §3.3. The formulas for covariance matrices of the estimated coefficients are also derived in this section. Two data-driven methods for finding unknown thresholding parameters are discussed in Section 4. Numerical comparisons and simulation studies are given in this section. Some discussion is given in Section 5. Technical proofs are relegated to the Appendix.

2 Penalized least squares and variable selection

Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\mathbf{y}$ is an $n \times 1$ vector and $\mathbf{X}$ is an $n \times d$ matrix. As in the traditional linear regression model, we assume that y_i's are conditionally independent given the design matrix. There are strong connections between the penalized least squares and the variable selection in the linear regression model. To gain more insights about various variable selection procedures, in this section we assume that the columns of $\mathbf{X}$ in (2.1) are orthonormal. The least squares estimate is obtained via minimizing $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$, which is equivalent to $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$, where $\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$ is the ordinary least squares estimate.

Denote by $\mathbf{z} = \mathbf{X}^T\mathbf{y}$ and let $\widehat{\mathbf{y}} = \mathbf{X}\mathbf{X}^T\mathbf{y}$. A form of the penalized least squares is

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{d} p_j(|\beta_j|) = \frac{1}{2}\|\mathbf{y} - \widehat{\mathbf{y}}\|^2 + \frac{1}{2}\sum_{j=1}^{d}(z_j - \beta_j)^2 + \lambda \sum_{j=1}^{d} p_j(|\beta_j|). \tag{2.2}$$

The penalty functions $p_j(\cdot)$ in (2.2) are not necessarily the same for all j. For example, one may wish to keep important predictors in a parametric model and hence not be willing to penalize their corresponding parameters. For simplicity of presentation, we will assume that the penalty functions for all coefficients are the same, denoted by $p(|\cdot|)$. Furthermore, we denote $\lambda p(|\cdot|)$ by $p_\lambda(|\cdot|)$, so $p(|\cdot|)$ may be allowed to depend on λ. Extensions to the case with different thresholding functions do not involve any extra difficulties.

The minimization problem of (2.2) is equivalent to minimizing componentwise. This leads us to considering the penalized least squares problem:

$$\frac{1}{2}(z - \theta)^2 + p_\lambda(|\theta|). \tag{2.3}$$

By taking the following hard thresholding penalty function [see Figure 5(a)]

$$p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda), \tag{2.4}$$

one obtains the hard thresholding rule (see Antoniadis 1997 and Fan 1997)

$$\widehat{\theta} = zI(|z| > \lambda). \tag{2.5}$$

See Figure 2(a). In other words, the solution to (2.2) is simply $z_j I(|z_j| > \lambda)$, which coincides with the best subset selection and stepwise deletion and addition for orthonormal designs. Note that the hard thresholding penalty function is a smoother penalty function than the entropy penalty $p_\lambda(|\theta|) =$

$\frac{\lambda^2}{2} I(|\theta| \neq 0)$, which also results in (2.5). The former facilitates computational expedience in other settings.

A good penalty function should result in an estimator with three properties: (a) **unbiasedness**: the resulting estimator is nearly unbiased when the true unknown parameter is large in order to avoid unnecessary modeling bias; (b) **sparsity**: the resulting estimator is a thresholding rule, which automatically sets small estimated coefficients to zero in order to reduce model complexity; (c) **continuity**: the resulting estimator is continuous in data z in order to avoid instability in model prediction. We now provide some insights on these requirements.

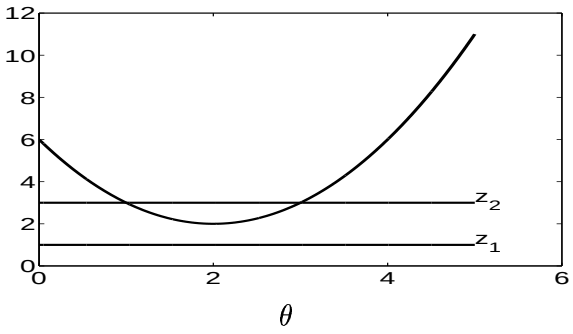

Figure 1: *A plot of* $\theta + p'_\lambda(\theta)$ *against* θ *(*$\theta > 0$*).*

The first order derivative of (2.3) with respect to θ is $\text{sgn}(\theta)\{|\theta| + p'_\lambda(|\theta|)\} - z$. It is easy to see that when $p'_\lambda(|\theta|) = 0$ for large $|\theta|$, the resulting estimator is z when $|z|$ is sufficiently large. Thus, when the true parameter $|\theta|$ is large, the observed value $|z|$ is large with high probability. Hence the penalized least squares will simply be $\widehat{\theta} = z$, which is approximately unbiased. Thus, the condition that $p'_\lambda(|\theta|) = 0$ **for large** $|\theta|$ is a sufficient condition for unbiasedness for a large true parameter. It corresponds to an improper prior distribution in the Bayesian model selection setting. A sufficient condition for the resulting estimator being a thresholding rule is that **the minimum of the function** $|\theta| + p'_\lambda(|\theta|)$ **is positive**. Figure 1 provides further insights into this statement. When $|z| < \min_{\theta \neq 0}\{|\theta| + p'_\lambda(|\theta|)\}$, the derivative of (2.3) is positive for all positive θ's, (and is negative for all negative θ's). Therefore the penalized least squares estimator is 0 in this situation, namely $\widehat{\theta} = 0$ for $|z| < \min_{\theta \neq 0}\{|\theta| + p'_\lambda(|\theta|)\}$. When $|z| > \min_{\theta \neq 0} |\theta| + p'_\lambda(|\theta|)$, there may exist two crossings as shown in Figure 1, and the larger one is a penalized least squares estimator. This implies that a sufficient and necessary condition for continuity is that **the minimum of the function** $|\theta| + p'_\lambda(|\theta|)$ **is attained at 0**. From the above discussion, a penalty function satisfying the conditions of sparsity and continuity must be singular at the origin.

It is well known that the L_2 penalty $p_\lambda(|\theta|) = \lambda|\theta|^2$ results in a ridge regression. The L_1 penalty $p_\lambda(|\theta|) = \lambda|\theta|$ yields a soft thresholding rule

$$\widehat{\theta}_j = \text{sgn}(z_j)(|z_j| - \lambda)_+, \tag{2.6}$$

which was proposed by Donoho and Johnstone (1994a). LASSO, proposed by Tibshirani (1996, 1997), is the penalized least squares estimate with the L_1 penalty in the general least squares and likelihood settings. The L_q penalty $p_\lambda(|\theta|) = \lambda|\theta|^q$ leads to a bridge regression (Frank and Friedman, 1993 and Fu, 1998). The solution is continuous only when $q \geq 1$. However, when $q > 1$, the minimum of $|\theta| + p'_\lambda(|\theta|)$ is zero and hence it does not produce a sparse solution (See Figure 3 (a)). The only continuous solution with a thresholding rule in this family is the L_1 penalty, but this comes at the price of shifting the resulting estimator by a constant λ (see Figure 2 (b)).

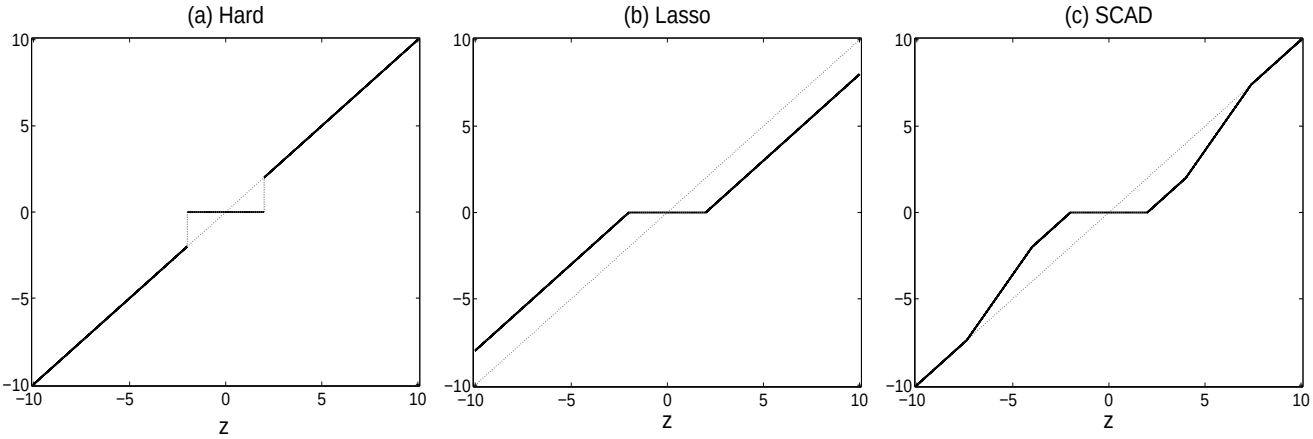

Figure 2: *Plot of thresholding functions. (a), (b) and (c) are the hard, soft and SCAD thresholding functions with* $\lambda = 2$ *and* $a = 3.7$ *for SCAD, respectively.*

2.1 Smoothly clipped absolute deviation penalty

The L_q and the hard thresholding penalty functions do not simultaneously satisfy the mathematical conditions for unbiasedness, sparsity and continuity. The continuous differentiable penalty function defined by

$$p'_\lambda(\theta) = \lambda\{I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\} \text{ for some } a > 2 \quad \text{and } \theta > 0, \tag{2.7}$$

improves the properties of the L_1 penalty and the hard thresholding penalty function given by (2.4) (see Figure 5(c) and discussion below) . We will call this penalty function the smoothly clipped absolute deviation (SCAD) penalty. This corresponds to a quadratic spline function with knots at λ

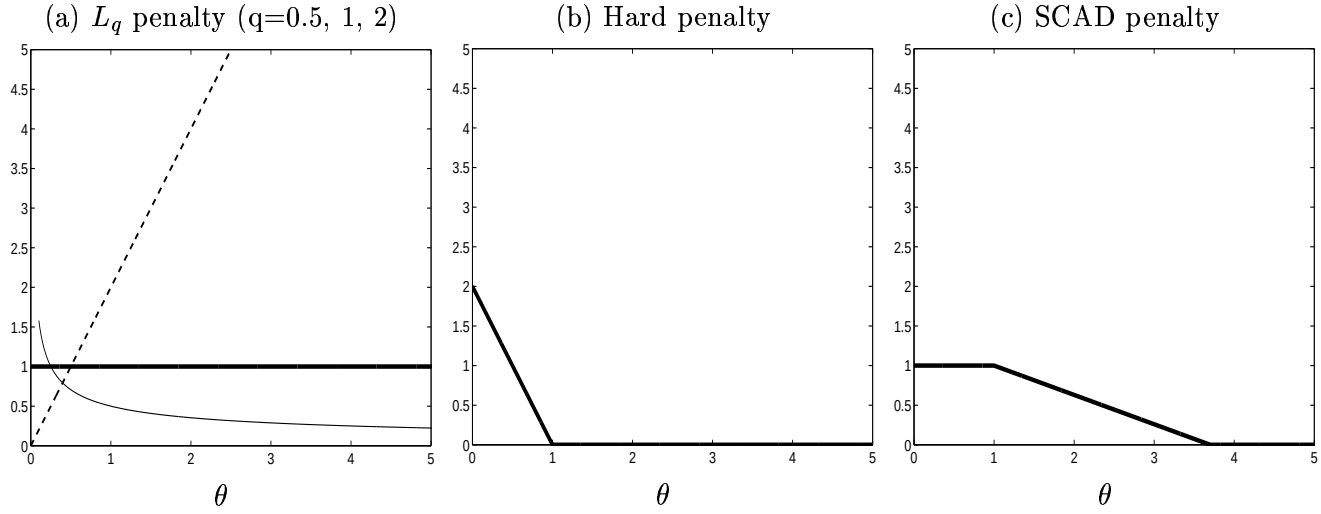

Figure 3: *Plot of $p'_\lambda(\theta)$ functions over $\theta > 0$. (a) stands for L_q penalties, the thicker solid, dashdot and thin solid curves corresponds to the L_1, $L_{0.5}$ and L_2 penalty, respectively; (b) stands for the hard thresholding penalty and (c) for the SCAD penalty.*

and $a\lambda$. This penalty function leaves large values of θ not excessively penalized and makes the solution continuous. The resulting solution is given by

$$\widehat{\theta} = \begin{cases} \text{sgn}(z)(|z| - \lambda)_+, & \text{when } |z| \le 2\lambda; \\ \{(a-1)z - \text{sgn}(z)a\lambda\}/(a-2), & \text{when } 2\lambda < |z| \le a\lambda; \\ z, & \text{when } |z| > a\lambda. \end{cases} \tag{2.8}$$

See Figure 2(c). This solution is due to Fan (1997), where a brief discussion in the settings of wavelets is given. In this paper, we will use it to develop an effective variable selection procedure for a broad class of models, including linear regression models and generalized linear models. For simplicity of presentation, we will use the name SCAD for all procedures using the SCAD penalty. The performance of SCAD is similar to that of firm shrinkage proposed by Bruce and Gao (1997) when design matrices are orthonormal.

The thresholding rule in (2.8) involves two unknown parameters λ and a. In practice, we could search the best pair (λ, a) over the two dimensional grids using some criteria, such as the cross-validation and the generalized cross-validation (Craven and Wahba, 1977). Such an implementation can be computationally expensive. To implement tools in the Bayesian risk analysis, we assume that for given a and λ, the prior distribution for θ is a normal distribution with zero mean and variance $a\lambda$. We computed the Bayes risk via numerical integration. Figure 4(a) depicts the Bayes risk as a function of a under the squared loss, for the universal thresholding $\lambda = \sqrt{2\log(d)}$ (see Donoho and Johnstone, 1994a) with $d = 20,\ 40,\ 60$ and 100, and Figure 4(b) is for $d = 512$, 1024, 2048 and 4096. From Figures

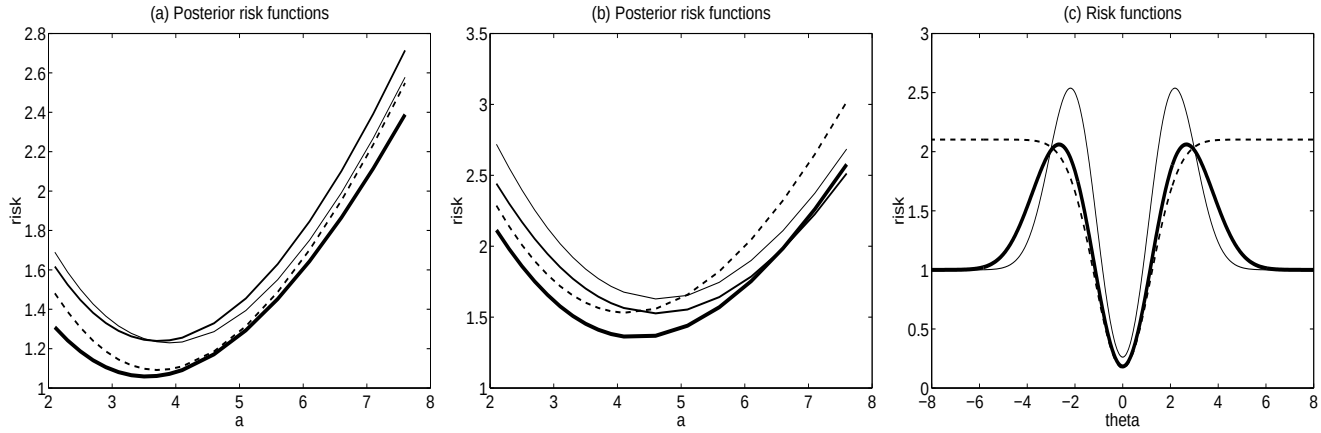

Figure 4: *Risk functions of proposed procedures under the quadratic loss. (a) and (b) are posterior risk functions of the SCAD under the prior* $\theta \sim N(0, a\lambda)$ *using the universal thresholding* $\lambda = \sqrt{2\log(d)}$ *for 4 different values d; the thicker solid, dashed, solid and thinner solid lines are for* $d = 20,\ 40,\ 60$ *and* 100*, respectively. (b) is similar to those for (a) with the thicker solid, dashed, solid, thinner solid lines for* $d = 512,\ 1024,\ 2048$ *and* 4096*, separately. (c) Risk functions of the four different thresholding rules. The thicker solid, dashed, and solid lines are for minimum SCAD, hard and soft thresholding rules.*

4(a) and 4(b), it can be seen from these two figures that the Bayesian risks are not very sensitive to the values of a. It can be seen from Figure 4(a) that the Bayes risks achieve their minimums at $a \approx 3.7$ when the value of d is less than 100. This choice gives pretty good practical performance for various variable selection problems. Indeed, based on the simulations in Section 4.3, the choice of $a = 3.7$ works similarly to that chosen by the GCV method.

2.2 Performance of thresholding rules

We now compare the performance of the above four thresholding rules. Marron *et al.* (1998) applied the tool of risk analysis to understand the small sample behavior of the hard and soft thresholding rules. The closed forms for the L_2 risk functions $R(\widehat{\theta}, \theta) = E(\widehat{\theta} - \theta)^2$ have been derived under the Gaussian model $Z \sim N(\theta, \sigma^2)$ for the hard thresholding and soft thresholding rules by Donoho and Johnstone (1994b). The risk function of the SCAD thresholding rule can be found in Li (2000). To gauge the performance of the four thresholding rules, Figure 4(c) depicts their L_2 risk functions under the Gaussian model $Z \sim N(\theta, 1)$. To make the scale of the thresholding parameters roughly comparable, we took $\lambda = 2$ for the hard thresholding rule, and adjusted the values of λ for the other thresholding rules so that their estimated values are the same when $\theta = 3$. The SCAD performs favorably compared with the other two thresholding rules. This can also be understood via their corresponding penalty

functions plotted in Figure 5. It is clear that the SCAD retains the good mathematical properties of the other two thresholding penalty functions. Hence, it is expected to perform the best. For general σ^2, the picture is the same, except scaled vertically by σ^2, and the θ axis should be replaced by θ/σ.

3 Variable selection via penalized likelihood

The methodology in the previous section can be directly applied to many other statistical contexts. In this section we consider linear regression models, robust linear models and likelihood-based generalized linear models. From now on, we assume that the design matrix $\mathbf{X} = (x_{ij})$ is standardized so that each column has mean zero and variance one.

3.1 Penalized least squares and likelihood

In the classical linear regression model, the least squares estimate is obtained via minimizing the sum of squared residual errors. Therefore (2.2) can be naturally extended to the situation in which design matrices are not orthonormal. Similar to (2.2), a form of penalized least squares is

$$\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{3.1}$$

Minimizing (3.1) with respect to $\boldsymbol{\beta}$ leads to a penalized least squares estimator of $\boldsymbol{\beta}$.

It is well known that the least squares estimate is not robust. One can consider the outlier-resistant loss functions such as the L_1 loss or more generally Huber's ψ-function (see Huber (1981)). Therefore, instead of minimizing (3.1), we minimize

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

with respect to $\boldsymbol{\beta}$. This results in a penalized robust estimator for $\boldsymbol{\beta}$.

For generalized linear models, statistical inferences are based on underlying likelihood functions. The penalized maximum likelihood estimator can be used to select significant variables. Assume that the data $\{(\mathbf{x}_i, Y_i)\}$ are collected independently. Conditioning on $\mathbf{x}_i$, Y_i has a density $f_i(g(\mathbf{x}_i^T\boldsymbol{\beta}), y_i)$, where g is a known link function. Let $\ell_i = \log f_i$ denote the conditional log-likelihood of Y_i. A form of the penalized likelihood is

$$\sum_{i=1}^{n} \ell_i(g(\mathbf{x}_i^T\boldsymbol{\beta}), y_i) - n\sum_{j=1}^{d} p_\lambda(|\beta_j|).$$

Maximizing the penalized likelihood function is equivalent to minimizing the following function with respect to $\boldsymbol{\beta}$:

$$-\sum_{i=1}^{n} \ell_i(g(\mathbf{x}_i^T\boldsymbol{\beta}), y_i) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{3.3}$$

To obtain a penalized maximum likelihood estimator of $\boldsymbol{\beta}$, we minimize (3.3) with respect to $\boldsymbol{\beta}$ for some thresholding parameter λ.

3.2 Sampling properties and oracle properties

In this section, we establish the asymptotic theory for our nonconcave penalized likelihood estimator. Let

$$\boldsymbol{\beta}_0 = (\beta_{10}, \cdots, \beta_{d0})^T = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T.$$

Without loss of generality, assume that $\boldsymbol{\beta}_{20} = \mathbf{0}$. Let $I(\boldsymbol{\beta}_0)$ be the Fisher information matrix and $I_1(\boldsymbol{\beta}_{10}, \mathbf{0})$ be the Fisher information knowing $\boldsymbol{\beta}_{20} = \mathbf{0}$. We first show that there exists a penalized likelihood estimator that converges at the rate

$$O_P(n^{-1/2} + a_n), \tag{3.4}$$

where $a_n = \max\{p'_{\lambda_n}(|\beta_{j0}|) : \beta_{j0} \neq 0\}$. This implies that for the hard thresholding and SCAD penalty functions, the penalized likelihood estimator is root-n consistent if $\lambda_n \to 0$. Further, we will demonstrate that such a root-n consistent estimator must satisfy $\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}$ and $\widehat{\boldsymbol{\beta}}_1$ is asymptotic normal with covariance matrix I_1^{-1}, if $n^{1/2}\lambda_n \to \infty$. This implies that the penalized likelihood estimator performs as well as if $\boldsymbol{\beta}_{20} = \mathbf{0}$ were known. In the similar language of Donoho and Johnstone (1994a), the resulting estimator performs as well as the oracle estimator, which knows in advance $\boldsymbol{\beta}_{20} = \mathbf{0}$.

The above oracle performance is closely related to the super-efficiency phenomenon. Consider the simplest linear regression model $\mathbf{y} = \mathbf{1}_n\mu + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, I_n)$. A superefficient estimate for μ is

$$\delta_n = \begin{cases} \bar{Y} & \text{if } |\bar{Y}| \geq n^{-1/4} \\ c\bar{Y} & \text{if } |\bar{Y}| < n^{-1/4} \end{cases}$$

due to Hodges (see page 405 of Lehmann, 1983). If we set c to 0, then δ_n coincides with the hard thresholding estimator with the thresholding parameter $\lambda_n = n^{-1/4}$. This estimator correctly estimates the parameter at point 0 without paying any price for estimating the parameter elsewhere.

We now state the result in a fairly general setting. To facilitate the presentation, we assume that the penalization is applied to every component of $\boldsymbol{\beta}$. However, there is no extra difficulty to extend it to the case where some components (e.g. variance in the linear models) are not penalized.

Set $\mathbf{V}_i = (\mathbf{X}_i, Y_i)$, $i = 1, \cdots, n$. Let $L(\boldsymbol{\beta})$ be the log-likelihood function of the observations $\mathbf{V}_1, \cdots, \mathbf{V}_n$ and $Q(\boldsymbol{\beta})$ be the penalized likelihood function $L(\boldsymbol{\beta}) - n\sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|)$. We state our theorems here but their proofs are relegated to the Appendix where the conditions for the theorems can also be found.

Theorem 1. Let $\mathbf{V}_1, \cdots, \mathbf{V}_n$ be independent and identically distributed, each with a density $f(\mathbf{V}, \boldsymbol{\beta})$ (with respect to a measure μ) which satisfies Conditions (A)$-$(C) in the Appendix. If $\max\{|p''_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\} \to 0$, then there exists a local maximizer $\widehat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2} + a_n)$, where a_n is given by (3.4).

It is clear from Theorem 1 that by choosing a proper λ_n, there exists a root-n consistent penalized likelihood estimator. We now show that this estimator must possess the sparsity property $\widehat{\boldsymbol{\beta}}_2 = 0$, which is stated as follows.

Lemma 1. Let $\mathbf{V}_1, \cdots, \mathbf{V}_n$ be independent and identically distributed, each with a density $f(\mathbf{V}, \boldsymbol{\beta})$ which satisfies Conditions (A)$-$(C) in the Appendix. Assume that

$$\liminf_{n\to\infty} \liminf_{\theta\to 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0. \tag{3.5}$$

If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$, then with probability tending to 1, for any given $\boldsymbol{\beta}_1$ satisfying that $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_P(n^{-1/2})$ and any constant C,

$$Q\{\begin{pmatrix} \boldsymbol{\beta}_1 \\ \mathbf{0} \end{pmatrix}\} = \max_{\|\boldsymbol{\beta}_2\| \leq Cn^{-1/2}} Q\{\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}\}.$$

Denote

$$\Sigma = \text{diag}\,\{p''_{\lambda_n}(|\beta_{10}|), \cdots, p''_{\lambda_n}(|\beta_{s0}|)\}$$

and

$$\mathbf{b} = (p'_{\lambda_n}(|\beta_{10}|)\text{sgn}(\beta_{10}), \cdots, p'_{\lambda_n}(|\beta_{s0}|)\text{sgn}(\beta_{s0}))^T,$$

where s is the number of components of $\boldsymbol{\beta}_{10}$.

Theorem 2. (**Oracle property**) Let $\mathbf{V}_1, \cdots, \mathbf{V}_n$ be independent and identically distributed, each with a density $f(\mathbf{V}, \boldsymbol{\beta})$ satisfying Conditions (A)$-$(C) in Appendix. Assume that the penalty function $p_{\lambda_n}(|\theta|)$ satisfies condition (3.5). If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$, then with probability tending to 1, the root n consistent local maximizers $\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{pmatrix}$ in Theorem 1 must satisfy:

(i) (**Sparsity**) $\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}$;

(ii) (**Asymptotic normality**)

$$\sqrt{n}(I_1(\boldsymbol{\beta}_{10}) + \Sigma)\left\{\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (I_1(\boldsymbol{\beta}_{10}) + \Sigma)^{-1}\mathbf{b}\right\} \to N\left\{\mathbf{0}, I_1(\boldsymbol{\beta}_{10})\right\}$$

in distribution, where $I_1(\boldsymbol{\beta}_{10}) = I_1(\boldsymbol{\beta}_{10}, \mathbf{0})$, the Fisher information knowing $\boldsymbol{\beta}_2 = \mathbf{0}$.

As a consequence, the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}_1$ is

$$\frac{1}{n}\left\{I_1(\boldsymbol{\beta}_{10}) + \Sigma\right\}^{-1} I_1(\boldsymbol{\beta}_{10})\left\{I_1(\boldsymbol{\beta}_{10}) + \Sigma\right\}^{-1},$$

which approximately equals $\frac{1}{n}I_1^{-1}(\boldsymbol{\beta}_{10})$ for the thresholding penalties discussed in Section 2 if λ_n tends to 0.

Remark 1: For the hard and SCAD thresholding penalty functions, if $\lambda_n \to 0$, $a_n = 0$. Hence, by Theorem 2, when $\sqrt{n}\lambda_n \to \infty$, their corresponding penalized likelihood estimators possess the oracle property and perform as well as the maximum likelihood estimates for estimating $\boldsymbol{\beta}_1$ knowing $\boldsymbol{\beta}_2 = \mathbf{0}$. However, for the L_1 penalty, $a_n = \lambda_n$. Hence, the root-n consistency requires that $\lambda_n = O_P(n^{-1/2})$. On the other hand, the oracle property in Theorem 2 requires that $\sqrt{n}\lambda_n \to \infty$. These two conditions for LASSO cannot be satisfied simultaneously. Indeed, for the L_1 penalty, we conjecture that the oracle property does not hold. But for L_q penalty with $q < 1$, the oracle property continues to hold with suitable choice of λ_n.

Now we briefly discuss the regularity conditions (A)-(C) for the generalized linear models (see McCullagh and Nelder, 1989). With a canonical link, the condition distribution of Y given $\mathbf{X} = \mathbf{x}$ belongs to the canonical exponential family, i.e., with a density function

$$f(y; \mathbf{x}, \boldsymbol{\beta}) = c(y)\exp\left\{\frac{y\mathbf{x}^T\boldsymbol{\beta} - b(\mathbf{x}^T\boldsymbol{\beta})}{a(\phi)}\right\}.$$

Clearly, the regularity conditions (A) are satisfied. The Fisher information matrix is

$$I(\boldsymbol{\beta}) = E\{b''(\mathbf{x}^T\boldsymbol{\beta})\mathbf{x}\mathbf{x}^T\}/a(\phi).$$

Therefore if $E\{b''(\mathbf{x}^T\boldsymbol{\beta})\mathbf{x}\mathbf{x}^T\}$ is finite and positive definite, then Condition (B) holds. If for all $\boldsymbol{\beta}$ in some neighborhood of $\boldsymbol{\beta}_0$, $|b^{(3)}(\mathbf{x}^T\boldsymbol{\beta})| \le M_0(\mathbf{x})$ for some function $M_0(\mathbf{x})$ satisfying $E_{\boldsymbol{\beta}_0}\{M_0(\mathbf{x})X_jX_kX_l\} < \infty$ for all j, k, l, then Condition (C) holds. For general link functions, similar conditions need to guarantee

Conditions (B) and (C). The mathematical derivation of those conditions does not involve any extra difficulty except more tedious notation. Results in Theorems 1 and 2 can also be established for the penalized least squares (3.1) and the penalized robust linear regression (3.2) under some mild regularity conditions. See Li (2000) for details.

3.3 A new unified algorithm

Tibshirani (1996) proposed an algorithm for solving constrained least squares problems of LASSO, while Fu (1998) provided a "shooting algorithm" for LASSO. See also LASSO2 submitted by Berwin Turlach at Statlib (http://lib.stat.cmu.edu/S/). In this section we propose a new unified algorithm for the minimization problems (3.1), (3.2) and (3.3) via local quadratic approximations. The first term in (3.1), (3.2) and (3.3) may be regarded as a loss function of $\boldsymbol{\beta}$. Denote it by $\ell(\boldsymbol{\beta})$. Then, the expressions (3.1), (3.2) and (3.3) can be written in a unified form as

$$\ell(\boldsymbol{\beta}) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{3.6}$$

The L_1 , hard thresholding and SCAD penalty functions are singular at the origin, and they do not have continuous second order derivatives. However they can be locally approximated by a quadratic function as follows. Suppose that we are given an initial value $\boldsymbol{\beta}_0$ that is close to the minimizer of (3.6). If β_{j0} is very close to 0, then set $\widehat{\beta}_j = 0$. Otherwise they can be locally approximated by a quadratic function as

$$[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta_j|)\mathrm{sgn}(\beta_j) \approx \{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\}\beta_j,$$

when $\beta_j \neq 0$. In other words,

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{j0}|) + \frac{1}{2}\{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\}(\beta_j^2 - \beta_{j0}^2), \qquad \text{for } \beta_j \approx \beta_{j0}. \tag{3.7}$$

Figure 5 shows the L_1, hard thresholding and SCAD penalty functions and their approximations on the right hand side of (3.7) at two different values of β_{j0}. A drawback of this approximation is that once a coefficient is shrunken to zero, it will stay at zero. However, this method reduces significantly the computational burden.

If $\ell(\boldsymbol{\beta})$ is the L_1 loss as in (3.2), then it does not have continuous second order partial derivatives with respect to $\boldsymbol{\beta}$. However, $\psi(|y - \mathbf{x}^T\boldsymbol{\beta}|)$ in (3.2) can be analogously approximated by $\{\psi(y - \mathbf{x}^T\boldsymbol{\beta}_0)/(y - \mathbf{x}^T\boldsymbol{\beta}_0)^2\}(y - \mathbf{x}^T\boldsymbol{\beta})^2$, as long as the initial value $\boldsymbol{\beta}_0$ of $\boldsymbol{\beta}$ is close to the minimizer. When some of the residuals $|y - \mathbf{x}^T\boldsymbol{\beta}_0|$ are small, this approximation is not very good. See Section 3.4 for some slight modifications of this approximation.

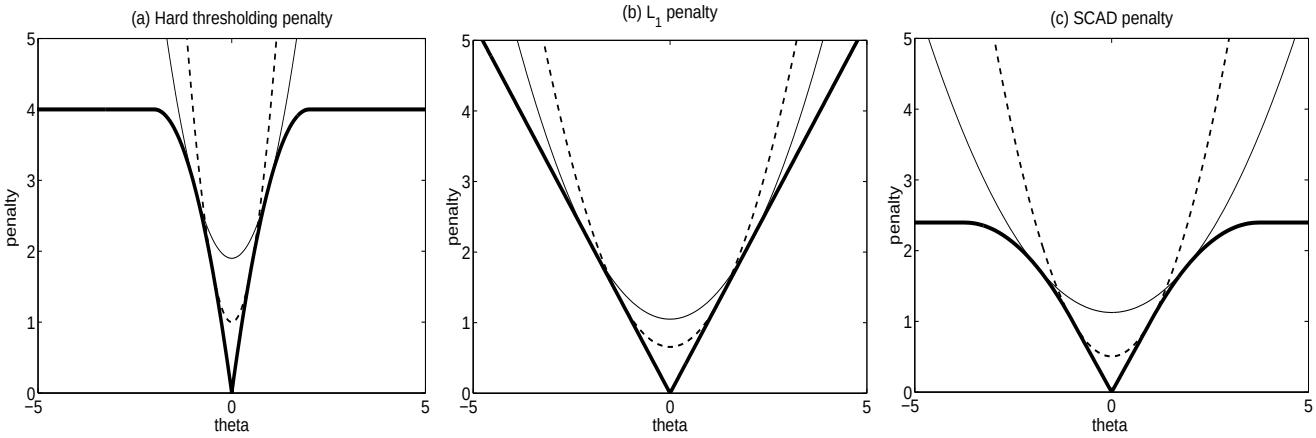

Figure 5: *Three penalty functions $p_\lambda(\theta)$ and their quadratic approximations. The values of λ are the same as those in Figure 4(c).*

Now assume that the log-likelihood function is smooth with respect to $\boldsymbol{\beta}$ so that its first two partial derivatives are continuous. Thus the first term in (3.6) can be locally approximated by a quadratic function. Therefore the minimization problem (3.6) can be reduced to a quadratic minimization problem and the Newton-Raphson algorithm can be used. Indeed, (3.6) can be locally approximated (except for a constant term) by

$$\ell(\boldsymbol{\beta}_0) + \nabla\ell(\boldsymbol{\beta}_0)^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\nabla^2\ell(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \frac{1}{2}n\boldsymbol{\beta}^T\Sigma_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta}, \tag{3.8}$$

where

$$\nabla\ell(\boldsymbol{\beta}_0) = \frac{\partial\ell(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}, \quad \nabla^2\ell(\boldsymbol{\beta}_0) = \frac{\partial^2\ell(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}, \quad \Sigma_\lambda(\boldsymbol{\beta}_0) = \mathrm{diag}\{p'_\lambda(|\beta_{10}|)/|\beta_{10}|, \cdots, p'_\lambda(|\beta_{d0}|)/|\beta_{d0}|\}.$$

The quadratic minimization problem (3.8) yields the solution

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 - \{\nabla^2\ell(\boldsymbol{\beta}_0) + n\Sigma_\lambda(\boldsymbol{\beta}_0)\}^{-1}\{\nabla\ell(\boldsymbol{\beta}_0) + n\mathbf{U}_\lambda(\boldsymbol{\beta}_0)\}, \tag{3.9}$$

where $\mathbf{U}_\lambda(\boldsymbol{\beta}_0) = \Sigma_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta}_0$. When the algorithm converges, the estimator satisfies the condition

$$\frac{\partial\ell(\widehat{\boldsymbol{\beta}}_0)}{\partial\beta_j} + np'_\lambda(|\widehat{\beta}_{j0}|)\mathrm{sgn}(\widehat{\beta}_{j0}) = 0,$$

the penalized likelihood equation, for non-zero elements of $\widehat{\boldsymbol{\beta}}_0$. Specifically, for the penalized least squares problem (3.1), the solution can be found by iteratively computing the following ridge regression:

$$\boldsymbol{\beta}_1 = \{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\boldsymbol{\beta}_0)\}^{-1}\mathbf{X}^T\mathbf{y}.$$

Similarly we obtain the solution for (3.2) by iterating

$$\boldsymbol{\beta}_1 = \{\mathbf{X}^T\mathbf{W}\mathbf{X} + \frac{1}{2}n\Sigma_\lambda(\boldsymbol{\beta}_0)\}^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y},$$

where $\mathbf{W} = \text{diag}\{\psi(|y_1 - \mathbf{x}_1^T\boldsymbol{\beta}_0|)/(y_1 - \mathbf{x}_1^T\boldsymbol{\beta}_0)^2, \cdots, \psi(|y_n - \mathbf{x}_n^T\boldsymbol{\beta}_0|)/(y_n - \mathbf{x}_n^T\boldsymbol{\beta}_0)^2\}$.

As in the maximum likelihood estimation (MLE) setting, with the good initial value $\boldsymbol{\beta}_0$, the one-step procedure can be as efficient as the fully iterative procedure, namely, the penalized maximum likelihood estimator, when one uses the Newton-Raphson algorithm (See Bickel (1975)). Now regarding $\boldsymbol{\beta}^{(k-1)}$ as a good initial value at the k-th step, the next iteration can also be regarded as a one-step procedure and hence the resulting estimator can still be as efficient as the fully iterative method. See Robinson (1988) for the theory on the difference between the MLE and the k-step estimators. Therefore estimators obtained by the aforementioned algorithm with a few iterations can always be regarded as a one-step estimator, which is as efficient as the fully iterative method. In this sense, one does not have to iterate the above algorithm until convergence as long as the initial estimators are good enough. The estimators from the full models can be used as initial estimators, as long as they are not overly parameterized.

3.4 Standard error formula

The standard errors for the estimated parameters can be directly obtained because we are estimating parameters and selecting variables at the same time. Following the conventional technique in the likelihood setting, the corresponding sandwich formula can be used as an estimator for the covariance of the estimates $\widehat{\boldsymbol{\beta}}_1$, the non-vanishing component of $\widehat{\boldsymbol{\beta}}$. That is,

$$\widehat{\text{cov}}(\widehat{\boldsymbol{\beta}}_1) = \{\nabla^2\ell(\widehat{\boldsymbol{\beta}}_1) + n\Sigma_\lambda(\widehat{\boldsymbol{\beta}}_1)\}^{-1}\widehat{\text{cov}}\{\nabla\ell(\widehat{\boldsymbol{\beta}}_1)\}\{\nabla^2\ell(\widehat{\boldsymbol{\beta}}_1) + n\Sigma_\lambda(\widehat{\boldsymbol{\beta}}_1)\}^{-1}. \tag{3.10}$$

Compare with Theorem 2(ii). This formula will be shown to have good accuracy for moderate sample sizes.

When the L_1 loss is used in the robust regression, some slight modifications are needed in the aforementioned algorithm and its corresponding sandwich formula. For $\psi(x) = |x|$, the diagonal elements of $\mathbf{W}$ are $\{|r_i|^{-1}\}$ with $r_i = y_i - \mathbf{x}_i^T\boldsymbol{\beta}_0$. Thus, for a given current value of $\boldsymbol{\beta}_0$, when some of the residuals $\{r_i\}$ are close to 0, these points receive too much weight. Hence, we replace the weight by $(a_n + |r_i|)^{-1}$. In our implementations, we took a_n as the $2n^{-1/2}$ quantile of the absolute residuals $\{|r_i|, i = 1, \cdots, n\}$. Thus, the constant a_n is changing from iteration to iteration.

3.5 Testing convergence of the algorithm

We now demonstrate that our algorithm converges to the right solution. To this end, we took a 100-dimensional vector $\boldsymbol{\beta}$ consisting of 50 zeros and other nonzero elements being generated from $N(0, 5^2)$ and used a 100×100 orthonormal design matrix $\mathbf{X}$. We then generated a response vector $\mathbf{y}$ from the linear model (2.1). We chose an orthonormal design matrix for our testing case, because the penalized least squares has a closed form mathematical solution so that we can compare our output with the mathematical solution. Our experiment did show that the proposed algorithm converged to the right solution. It took MATLAB 0.27, 0.39 and 0.16 seconds for the penalized least squares with the SCAD, L_1 and hard thresholding penalties to converge. The numbers of iterations are 30, 30 and 5 respectively for the penalized least squares with the SCAD, L_1 and the hard thresholding penalty. In fact, after 10 iterations, the penalized least squares estimators are already very close to the true one.

4 Numerical Comparisons

The purpose of this section is to compare the performance of the proposed approaches and existing ones and to test the accuracy of the standard error formula. We also illustrate our penalized likelihood approaches by a real data example. In all examples in this section, we computed the penalized likelihood estimate with the L_1 penalty, referred as to LASSO, by our algorithm rather than those of Tibshirani (1996) and Fu (1998).

4.1 Prediction and model error

The prediction error is defined as the average error in the prediction of Y given $\mathbf{x}$ for future cases not used in the construction of a prediction equation. There are two regression situations, *X-random* and *X-controlled*. In the case that X is random, both Y and $\mathbf{x}$ are randomly selected. In the controlled situation, design matrices are selected by experimenters and only y is random. For ease of presentation, we consider only the X-random case.

In *X-random* situations, the data $(\mathbf{x}_i, Y_i)$ are assumed to be a random sample from their parent distribution $(\mathbf{x}, Y)$. Then, if $\widehat{\mu}(\mathbf{x})$ is a prediction procedure constructed using the present data, the prediction error is defined as

$$\mathrm{PE}(\widehat{\mu}) = E\{Y - \widehat{\mu}(\mathbf{x})\}^2,$$

where the expectation is only taken with respect to the new observation $(\mathbf{x}, Y)$. The prediction error

can be decomposed as

$$\mathrm{PE}(\widehat{\mu}) = E\{Y - E(Y|\mathbf{x})\}^2 + E\{E(Y|\mathbf{x}) - \widehat{\mu}(\mathbf{x})\}^2.$$

The first component is the inherent prediction error due to the noise. The second component is due to lack of fit to an underlying model. This component is called *model error* and is denoted by $\mathrm{ME}(\widehat{\mu})$. The size of the model error reflects performances of different model selection procedures. If $Y = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon$, where $E(\varepsilon|\mathbf{x}) = 0$, then $\mathrm{ME}(\widehat{\mu}) = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T E(\mathbf{x}\mathbf{x}^T)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

4.2 Selection of thresholding parameters

To implement the methods described in Sections 2 and 3, we need to estimate the thresholding parameters λ and a (for the SCAD). Denote by $\boldsymbol{\theta}$ the tuning parameters to be estimated, i.e., $\boldsymbol{\theta} = (\lambda, a)$ for the SCAD, while $\boldsymbol{\theta} = \lambda$ for the other penalty functions. Here we discuss two methods of estimating $\boldsymbol{\theta}$: fivefold cross-validation and generalized cross-validation, as suggested by Breiman (1995), Tibshirani (1996) and Fu (1998).

For completeness, we now describe the details of the cross-validation and the generalized cross validation procedures. Here we only discuss these two procedures for linear regression models. Extensions to robust linear models and likelihood-based linear models do not involve extra difficulties. The fivefold cross-validation procedure is as follows: Denote the full data set by T, and cross-validation training and test set by $T - T^\nu$ and T^ν respectively, for $\nu = 1, \cdots, 5$. For each $\boldsymbol{\theta}$ and ν, we find the estimator $\widehat{\boldsymbol{\beta}}^{(\nu)}(\boldsymbol{\theta})$ of $\boldsymbol{\beta}$ using the training set $T - T^\nu$. Form the cross-validation criterion as

$$\mathrm{CV}(\boldsymbol{\theta}) = \sum_{\nu=1}^{5} \sum_{(y_k, \mathbf{x}_k) \in T^\nu} \{y_k - \mathbf{x}_k^T \widehat{\boldsymbol{\beta}}^{(\nu)}(\boldsymbol{\theta})\}^2.$$

We find a $\widehat{\boldsymbol{\theta}}$ that minimizes $\mathrm{CV}(\boldsymbol{\theta})$.

The second method is the generalized cross-validation. For linear regression models, we update the solution by

$$\boldsymbol{\beta}_1(\boldsymbol{\theta}) = \{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\boldsymbol{\beta}_0)\}^{-1}\mathbf{X}^T\mathbf{y}.$$

Thus the fitted value $\widehat{\mathbf{y}}$ of $\mathbf{y}$ is $\mathbf{X}\{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\boldsymbol{\beta}_0)\}^{-1}\mathbf{X}^T\mathbf{y}$, and

$$\mathbf{P}_\mathbf{X}\{\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\} = \mathbf{X}\{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\widehat{\boldsymbol{\beta}})\}^{-1}\mathbf{X}^T$$

can be regarded as a projection matrix. Define the number of effective parameters in the penalized least squares fit as $e(\boldsymbol{\theta}) = \mathrm{tr}[\mathbf{P}_\mathbf{X}\{\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\}]$. Therefore the generalized cross-validation statistic is

$$\mathrm{GCV}(\boldsymbol{\theta}) = \frac{1}{n}\frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}(\boldsymbol{\theta})\|^2}{\{1 - e(\boldsymbol{\theta})/n\}^2}$$

and $\widehat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}}\{\text{GCV}(\boldsymbol{\theta})\}$.

4.3 Simulation study

In the following examples, we numerically compare the proposed variable selection methods with the ordinary least squares, ridge regression, best subset selection and non-negative garrote (see Breiman (1995)). All simulations are conducted using MATLAB codes. We directly used the constraint least squares module in MATLAB for finding the non-negative garrote estimate. As recommended in Breiman (1995), a five-fold cross-validation was used to estimate the tuning parameter for the non-negative garrote. For the other model selection procedures, both five-fold cross-validation and generalized cross-validation were used for estimating thresholding parameters. However, their performance was similar. Therefore we only present the results based on the generalized cross-validation.

Example 4.1.(Linear regression) In this example we simulated 100 data sets consisting of n observations from the model

$$Y = \mathbf{x}^T\boldsymbol{\beta} + \sigma\varepsilon,$$

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and the components of $\mathbf{x}$ and ε are standard normal. The correlation between x_i and x_j is $\rho^{|i-j|}$ with $\rho = 0.5$. This is a model used in Tibshirani (1996). Firstly we chose $n = 40$ and $\sigma = 3$. Then we reduced σ to 1 and finally increased the sample size to 60. The model error of the proposed procedures are compared to that of the least squares estimator. The Median of Relative Model Errors (MRME) over 100 simulated data sets are summarized in Table 1. The average of 0 coefficients is also reported in Table 1, in which the column labeled "correct" presents the average restricted only to the true zero coefficients, while the column labeled "incorrect" depicts the average of coefficients erroneously set to 0.

From Table 1, it can be seen that when the noise level is high and sample size is small, LASSO performs the best and it significantly reduces both model error and model complexity; while ridge regression only reduces model error. The other variable selection procedures also reduce model error and model complexity. However, when the noise level is reduced, the SCAD outperforms the LASSO and the other penalized least squares. Ridge regression performs very poorly. The best subset selection method performs quite similarly to the SCAD. The nonnegative garrote performs quite well in various situations. Comparing the first two rows in Table 1, one can see that the choice of $a = 3.7$ is very reasonable. Therefore we used it for other examples in this paper. Table 1 also depicts the performance

Table 1: Simulation results for the linear regression model

Method	MRME(%)	Aver. no. of 0 Coeff.	
$n = 40,\ \sigma = 3$		correct	incorrect
SCAD^1	72.90	4.20	0.21
SCAD^2	69.03	4.31	0.27
LASSO	63.19	3.53	0.07
Hard	73.82	4.09	0.19
Ridge	83.28	0	0
Best subset	68.26	4.50	0.35
Garrote	76.90	2.80	0.09
Oracle	33.31	5	0
$n = 40,\ \sigma = 1$		correct	incorrect
SCAD^1	54.81	4.29	0
SCAD^2	47.25	4.34	0
LASSO	63.19	3.51	0
Hard	69.72	3.93	0
Ridge	95.21	0	0
Best subset	53.60	4.54	0
Garrote	56.55	3.35	0
Oracle	33.31	5	0
$n = 60,\ \sigma = 1$		correct	incorrect
SCAD^1	47.54	4.37	0
SCAD^2	43.79	4.42	0
LASSO	65.22	3.56	0
Hard	71.11	4.02	0
Ridge	97.36	0	0
Best subset	46.11	4.73	0
Garrote	55.90	3.38	0
Oracle	29.82	5	0

Note that the value of a in SCAD^1 is obtained by generalized cross-validation, while the value of a in SCAD^2 is 3.7.

Table 2: Standard deviations of estimators for the linear regression model ($n = 60$)

	$\widehat{\beta}_1$		$\widehat{\beta}_2$		$\widehat{\beta}_5$	
Method	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$
SCAD^1	0.166	0.161 (0.021)	0.170	0.160 (0.024)	0.148	0.145 (0.022)
SCAD^2	0.161	0.161 (0.021)	0.164	0.161 (0.024)	0.151	0.143 (0.023)
LASSO	0.164	0.154 (0.019)	0.173	0.150 (0.022)	0.153	0.142 (0.021)
Hard	0.169	0.161 (0.022)	0.174	0.162 (0.025)	0.178	0.148 (0.021)
Best subset	0.163	0.155 (0.020)	0.152	0.154 (0.026)	0.152	0.139 (0.020)
Oracle	0.155	0.154 (0.020)	0.147	0.153 (0.024)	0.146	0.137 (0.019)

of an oracle estimator. From Table 1, it also can be seen that the performance of SCAD is expected to work as well as that of the oracle estimator as the sample size n increase. See Tables 5 and 6 for more details.

We now test the accuracy of our standard error formula (3.10). The median absolute deviation divided by 0.6745, denoted by SD in Table 2, of 100 estimated coefficients in the 100 simulations can be regarded as the true standard error. The median of the 100 estimated SDs, denoted by SD_m, and the median absolute deviation error of the 100 estimated standard errors divided by 0.6745, denoted by SD_{mad}, gauge the overall performance of the standard error formula (3.10). Table 2 presents the results for non-zero coefficients when the sample size $n = 60$. The results for the other two cases with $n = 40$ are similar. Table 2 suggests that the sandwich formula performs surprisingly well.

Example 4.2. (Robust regression) In this example, we simulated 100 data sets consisting of 60 observations from the model

$$Y = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon,$$

where $\boldsymbol{\beta}$ and $\mathbf{x}$ are the same as those in Example 1. The ε is drawn from the standard normal distribution with 10% outliers from the standard Cauchy distribution. The simulation results are summarized in Table 3. From Table 3, it can be seen that the SCAD outperforms somewhat the other procedures. The true and estimated standard deviations of estimators via sandwich formula (3.7) are shown in Table 4. It indicates that the performance of the sandwich formula is very good.

Table 3: Simulation results for the robust linear model

Method	MRME(%)	Aver. no. of 0 Coeff.	
		correct	incorrect
SCAD (a=3.7)	35.52	4.71	0
LASSO	52.80	4.29	0
Hard	47.22	4.70	0
Best subset	41.53	4.85	0.18
Oracle	23.33	5	0

Example 4.3. (Logistic regression) In this example, we simulated 100 data sets consisting of 200 observations from the model $Y \sim \text{Bernoulli}\{p(\mathbf{x}^T\boldsymbol{\beta})\}$, where $p(u) = \exp(u)/(1 + \exp(u))$, and the

Table 4: Standard deviations of estimators for the robust regression model

	$\widehat{\beta}_1$		$\widehat{\beta}_2$		$\widehat{\beta}_5$	
Method	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$
SCAD	0.167	0.171 (0.018)	0.185	0.176 (0.022)	0.165	0.155 (0.020)
LASSO	0.158	0.165 (0.022)	0.159	0.167 (0.020	0.182	0.154 (0.019)
Hard	0.179	0.168 (0.018)	0.176	0.176 (0.025)	0.157	0.154 (0.020)
Best subset	0.198	0.172 (0.023)	0.185	0.175 (0.024)	0.199	0.152 (0.023)
Oracle	0.163	0.199 (0.040)	0.156	0.202 (0.043)	0.166	0.177 (0.037)

first six components of $\mathbf{x}$ and $\boldsymbol{\beta}$ are the same as those in Example 1. The last two components of $\mathbf{x}$ are independently identically distributed as a Bernoulli distribution with probability of success 0.5. All covariates are standardized. Model errors are computed via 1000 Monte Carlo simulations. The summary of simulation results is depicted in Tables 5 and 6. From Table 5, it can be seen that the performance of the SCAD is much better than other two penalized likelihood estimates. Results in Table 6 show that our standard error estimator works well. From Tables 5 and 6, SCAD works as well as the oracle estimator in terms of the MRME and the accuracies of estimated standard errors.

Table 5: Simulation results for the logistic regression

Method	MRME(%)	Aver. no. of 0 Coeff.	
		correct	incorrect
SCAD(a=3.7)	26.48	4.98	0.04
LASSO	53.14	3.76	0
Hard	59.06	4.27	0
Best subset	31.63	4.84	0.01
Oracle	25.71	5	0

Table 6: Standard deviations of estimators for the logistic regression

	$\widehat{\beta}_1$		$\widehat{\beta}_2$		$\widehat{\beta}_5$	
Method	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$
SCAD ($a = 3.7$)	0.571	0.538 (0.107)	0.383	0.372 (0.061)	0.432	0.398 (0.065)
LASSO	0.310	0.379 (0.037)	0.285	0.284 (0.019)	0.244	0.287 (0.019)
Hard	0.675	0.561 (0.126)	0.428	0.400 (0.062)	0.467	0.421 (0.079)
Best subset	0.624	0.547 (0.121)	0.398	0.383 (0.067)	0.468	0.412 (0.077)
Oracle	0.553	0.538 (0.103)	0.374	0.373 (0.060)	0.432	0.398 (0.064)

We would like to remark that the estimated SDs for the L_1 penalized likelihood estimator (LASSO) are consistently smaller than the SCAD, however, its overall MRME is larger than that of the SCAD. This implies that the biases in the L_1 penalized likelihood estimators are large. This remark applies to all of our examples. Indeed, in Table 7, all coefficients were noticeably shrunken by LASSO.

Example 4.4 We in this example apply the proposed penalized likelihood methodology to the *Burns data*, collected by the General Hospital Burn Center at the University of Southern California. The data set consists of 981 observations. The binary response variable Y is 1 for those victims who survived their burns and 0 otherwise. Covariates $X_1 = age$, $X_2 = sex$, X_3 =log(burn area +1) and binary variable $X_4 = Oxygen$ (0 normal, 1 abnormal) were considered. Quadratic terms of X_1 and X_3, and all interaction terms were included. The intercept term was added and the logistic regression model was fitted. The best subset variable selection with AIC and BIC was applied to this data set. The unknown parameter λ was chosen by the generalized cross-validation. They are 0.6932, 0.0015 and 0.8062 for the penalized likelihood estimates with the SCAD, L_1 and hard thresholding penalties respectively. The constant a in the SCAD was taken as 3.7. With the selected λ, the penalized likelihood estimator was obtained at the 6^{th}, 28^{th} and 5^{th} step iterations for the penalized likelihood with the SCAD, L_1 and hard thresholding penalties, respectively. We also computed ten-step estimators, it took us less than 50 seconds for each penalized likelihood estimator, and the differences between the full iteration estimators and the ten-step estimators were less than one percent. The estimated coefficients and standard errors for the transformed data, based on the penalized likelihood estimators, are reported in Table 7.

From Table 7, the best subset procedure via minimizing the BIC scores chooses 5 out of 13 covariates, while the SCAD chooses 4 covariates. The difference between them is that the best subset keeps X_4. Both SCAD and the best subset variable selection (BIC) do not include X_1^2 and X_3^2 in the selected subset, but both LASSO and the best subset variable selection (AIC) do. LASSO chooses the quadratic term of X_1 and X_3 rather than their linear terms. It also selects an interaction term X_2X_3, which may not be statistically significant. LASSO shrinks noticeably large coefficients. In this example, the penalized likelihood with the hard thresholding penalty retains too many predictors. Particularly, it selects variables X_2 and X_2X_3.

Table 7: Estimated coefficients and standard errors for Example 4.4

Method	MLE	best subset (AIC)	best subset (BIC)	SCAD	LASSO	hard
intercept	5.51 (0.75)	4.81 (0.45)	6.12 (0.57)	6.09 (0.29)	3.70 (0.25)	5.88 (0.41)
X_1	-8.83 (2.97)	-6.49 (1.75)	-12.15 (1.81)	-12.24 (0.08)	0 (−)	-11.32 (1.1)
X_2	2.30 (2.00)	0 (−)	0 (−)	0 (−)	0 (−)	2.21 (1.41)
X_3	-2.77 (3.43)	0 (−)	-6.93 (0.79)	-7.00 (0.21)	0 (−)	-4.23 (0.64)
X_4	-1.74 (1.41)	0.30 (0.11)	-0.29 (0.11)	0 (−)	-0.28 (0.09)	-1.16 (1.04)
X_1^2	-0.75 (0.61)	-1.04 (0.54)	0 (−)	0 (−)	-1.71 (0.24)	0 (−)
X_3^2	-2.70 (2.45)	-4.55 (0.55)	0 (−)	0 (−)	-2.67 (0.22)	-1.92 (0.95)
X_1X_2	0.03 (0.34)	0 (−)	0 (−)	0 (−)	0 (−)	0 (−)
X_1X_3	7.46 (2.34)	5.69 (1.29)	9.83 (1.63)	9.84 (0.14)	0.36 (0.22)	9.06 (0.96)
X_1X_4	0.24 (0.32)	0 (−)	0 (−)	0 (−)	0 (−)	0 (−)
X_2X_3	-2.15 (1.61)	0 (−)	0 (−)	0 (−)	-0.10 (0.10)	-2.13 (1.27)
X_2X_4	-0.12 (0.16)	0 (−)	0 (−)	0 (−)	0 (−)	0 (−)
X_3X_4	1.23 (1.21)	0 (−)	0 (−)	0 (−)	0 (−)	0.82 (1.01)

5 Conclusion

We proposed a variable selection method via penalized likelihood approaches. A family of penalty functions were introduced. Rates of convergence of the proposed penalized likelihood estimators were established. With proper choice of regularization parameters, it has been shown that the proposed estimators perform as well as the oracle procedure in the variable selection. The methods were shown to be effective and the standard errors were estimated with good accuracy. A unified algorithm was proposed for minimizing penalized likelihood function, which is usually a sum of convex and concave functions. Our algorithm is backed up by statistical theory and hence gives estimators with good statistical properties. Comparing with the best subset method, which is very time consuming, the newly proposed methods are much faster, more effective and have strong theoretical backup. They select variables simultaneously via optimizing a penalized likelihood, and hence the standard errors of estimated parameters can be estimated accurately. The LASSO proposed by Tibshirani (1996) is a member of this penalized likelihood family with L_1 penalty. It has good performance when noise to signal ratios is large, but the bias created by this approach is noticeably large. See also the remarks in Example 4.3. The penalized likelihood with the Smoothly Clipped Absolute Deviation (SCAD) penalty function gives the best performance in selecting significant variables without creating excessive biases. The approach proposed here can be applied to other statistical contexts without any extra difficulties.

Appendix: Proofs

Before we present the proofs of the theorems, we first state some regularity conditions. Denote by Ω the parameter space for $\boldsymbol{\beta}$.

Regularity Conditions:

(A) The observations $\mathbf{V}_i$ are independent and identically distributed with probability density $f(\mathbf{V}, \boldsymbol{\beta})$ with respect to some measure μ. $f(\mathbf{V}, \boldsymbol{\beta})$ has a common support and the model is identifiable. Furthermore, the first and second logarithmic derivatives of f satisfying the equations

$$E_{\boldsymbol{\beta}}\left[\frac{\partial \log f(\mathbf{V}, \boldsymbol{\beta})}{\partial \beta_j}\right] = 0 \quad \text{for} \quad j = 1, \cdots, d$$

and

$$I_{jk}(\boldsymbol{\beta}) = E_{\boldsymbol{\beta}}\left[\frac{\partial}{\partial \beta_j} \log f(\mathbf{V}, \boldsymbol{\beta}) \frac{\partial}{\partial \beta_k} \log f(\mathbf{V}, \boldsymbol{\beta})\right] = E_{\boldsymbol{\beta}}\left[-\frac{\partial^2}{\partial \beta_j \partial \beta_k} \log f(\mathbf{V}, \boldsymbol{\beta})\right].$$

(B) The Fisher information matrix

$$I(\boldsymbol{\beta}) = E\left\{\left[\frac{\partial}{\partial \boldsymbol{\beta}} \log f(\mathbf{V}, \boldsymbol{\beta})\right]\left[\frac{\partial}{\partial \boldsymbol{\beta}} \log f(\mathbf{V}, \boldsymbol{\beta})\right]^T\right\}$$

is finite and positive definite at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

(C) There exists an open subset ω of Ω containing the true parameter point $\boldsymbol{\beta}_0$ such that for almost all $\mathbf{V}$ the density $f(\mathbf{V}, \boldsymbol{\beta})$ admits all third derivatives $\frac{\partial f(\mathbf{V}, \boldsymbol{\beta})}{\partial \beta_j \partial \beta_k \partial \beta_l}$ for all $\boldsymbol{\beta} \in \omega$. Further there exist functions M_{jkl} such that

$$\left|\frac{\partial^3}{\partial \beta_j \partial \beta_k \partial \beta_l} \log f(\mathbf{V}, \boldsymbol{\beta})\right| \leq M_{jkl}(\mathbf{V}) \quad \text{for all} \quad \boldsymbol{\beta} \in \omega,$$

where $m_{jkl} = E_{\boldsymbol{\beta}_0}[M_{jkl}(\mathbf{V})] < \infty$ for j, k, l.

These regularity conditions guarantee asymptotic normality of the ordinary maximum likelihood estimates. See for example Lehmann (1983).

Proof of Theorem 1:

Let $\alpha_n = n^{-1/2} + a_n$. We want to show that for any given $\varepsilon > 0$, there exists a large constant C such that

$$P\left\{\sup_{\|\mathbf{u}\|=C} Q(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) < Q(\boldsymbol{\beta}_0)\right\} \geq 1 - \varepsilon. \tag{5.1}$$

This implies with probability at least $1-\varepsilon$ that there exists a local maximum in the ball $\{\boldsymbol{\beta}_0+\alpha_n\mathbf{u}: \|\mathbf{u}\|\le C\}$. Hence, there exists a local maximizer such that $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|=O_P(\alpha_n)$.

Using $p_{\lambda_n}(0)=0$, we have

$$\begin{aligned} D_n(\mathbf{u}) &\equiv Q(\boldsymbol{\beta}_0+\alpha_n\mathbf{u})-Q(\boldsymbol{\beta}_0) \\ &\le L(\boldsymbol{\beta}_0+\alpha_n\mathbf{u})-L(\boldsymbol{\beta}_0)-n\sum_{j=1}^{s}\{p_{\lambda_n}(|\beta_{j0}+\alpha_n u_j|)-p_{\lambda_n}(|\beta_{j0}|)\}, \end{aligned}$$

where s is the number of components of $\boldsymbol{\beta}_{10}$. Let $L'(\boldsymbol{\beta}_0)$ be the gradient vector of L. By the standard argument on the Taylor expansion of the likelihood function, we have

$$\begin{aligned} D_n(\mathbf{u}) &\le \alpha_n L'(\boldsymbol{\beta}_0)^T\mathbf{u}-\frac{1}{2}\mathbf{u}^T I(\boldsymbol{\beta}_0)\mathbf{u}n\alpha_n^2\{1+o_P(1)\} \\ &\quad -\sum_{j=1}^{s}\left[n\alpha_n p'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0})u_j+n\alpha_n^2 p''_{\lambda_n}(|\beta_{j0}|)u_j^2\{1+o(1)\}\right]. \end{aligned} \tag{5.2}$$

Note that $n^{-1/2}L'(\boldsymbol{\beta}_0)=O_P(1)$. Thus, the first term on the right hand side of (5.2) is of the order $O_P(n^{1/2}\alpha_n)=O_P(n\alpha_n^2)$. By a choosing sufficient large C, the second term will dominate the first term, uniformly in $\|\mathbf{u}\|=C$. Note that the third term in (5.2) is bounded by

$$\sqrt{s}n\alpha_n a_n\|\mathbf{u}\|+n\alpha_n^2\max\{|p''_{\lambda_n}(|\beta_{j0}|)|:\beta_{j0}\neq 0\}\|\mathbf{u}\|^2.$$

This is also dominated by the second term of (5.2). Hence, by choosing sufficiently large C, (5.1) holds. This completes the proof of the theorem.

Proof of Lemma 1:

It is sufficient to show that with probability tending to 1 as $n\to\infty$, for any $\boldsymbol{\beta}_1$ satisfying that $\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10}=O_P(n^{-1/2})$ and for some small $\varepsilon_n=Cn^{-1/2}$ and $j=s+1,\cdots,d$,

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial\beta_j}<0 \quad \text{for} \quad 0<\beta_j<\varepsilon_n, \tag{5.3}$$

and

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial\beta_j}>0 \quad \text{for} \quad -\varepsilon_n<\beta_j<0. \tag{5.4}$$

To show (5.3), by Taylor's expansion, we have

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\beta})}{\partial\beta_j} &= \frac{\partial L(\boldsymbol{\beta})}{\partial\beta_j}-np'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j) \\ &= \frac{\partial L(\boldsymbol{\beta}_0)}{\partial\beta_j}+\sum_{l=1}^{d}\frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial\beta_j\partial\beta_l}(\beta_l-\beta_{l0}) \\ &\quad +\sum_{l=1}^{d}\sum_{k=1}^{d}\frac{\partial^3 L(\boldsymbol{\beta}^*)}{\partial\beta_j\partial\beta_l\beta_k}(\beta_l-\beta_{l0})(\beta_k-\beta_{k0})-np'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j), \end{aligned}$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. Note that by the standard arguments

$$n^{-1}\frac{\partial L(\boldsymbol{\beta}_0)}{\partial \beta_j} = O_P(n^{-1/2})$$

and

$$\frac{1}{n}\frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l} = E\left\{\frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l}\right\} + o_P(1).$$

By the assumption that $\boldsymbol{\beta} - \boldsymbol{\beta}_0 = O_P(n^{-1/2})$, we have

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} = n\lambda_n\{-\lambda_n^{-1}p'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j) + O_P(n^{-1/2}/\lambda_n)\}.$$

Since $\liminf_{n\to\infty}\liminf_{\theta\to 0^+}\lambda_n^{-1}p'_{\lambda_n}(\theta) > 0$ and $n^{-1/2}/\lambda_n \to 0$, the sign of the derivative is completely determined by that of β_j. Hence, (5.3) and (5.4) follow. This completes the proof.

Proof of Theorem 2:

It follows by Lemma 1 that Part (i) holds. Now we prove Part (ii). It can be easily shown that there exists a $\widehat{\boldsymbol{\beta}}_1$ in Theorem 1 being a root n consistent local maximizer of $Q\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \mathbf{0}\end{pmatrix}\}$, regarded as a function of $\boldsymbol{\beta}_1$, and satisfying the likelihood equations:

$$\left.\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\begin{pmatrix}\widehat{\boldsymbol{\beta}}_1\\ \mathbf{0}\end{pmatrix}} = 0, \quad \text{for} \quad j = 1, \cdots, s. \tag{5.5}$$

Note that $\widehat{\boldsymbol{\beta}}_1$ is a consistent estimator,

$$\begin{aligned}
&\left.\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\begin{pmatrix}\widehat{\boldsymbol{\beta}}_1\\ \mathbf{0}\end{pmatrix}} - np'_{\lambda_n}(|\widehat{\beta}_j|)\mathrm{sgn}(\widehat{\beta}_j)\\
= \;& \frac{\partial L(\boldsymbol{\beta}_0)}{\partial \beta_j} + \sum_{l=1}^{s}\left\{\frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l} + o_P(1)\right\}(\widehat{\beta}_l - \beta_{l0})\\
& -n\left(p'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0}) + \{p''_{\lambda_n}(|\beta_{j0}|) + o_P(1)\}(\widehat{\beta}_j - \beta_{j0})\right).
\end{aligned}$$

It follows by Slutsky's Theorem and the CLT that

$$\sqrt{n}(I_1(\boldsymbol{\beta}_{10}) + \Sigma)\left\{\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (I_1(\boldsymbol{\beta}_{10}) + \Sigma)^{-1}\mathbf{b}\right\} \to N\{\mathbf{0}, I_1(\boldsymbol{\beta}_{10})\}$$

in distribution.

References

Antoniadis, A. (1997). Wavelets in Statistics: A Review (with discussion). *Journal of Italian Statistical Association*, **6**, 97-144.

Antoniadis, A. and Fan, J. (1999). Regularization of wavelets approximations, *Journal of American Statistical Association*, tentatively accepted.

Bickel, P.J. (1975). One-step Huber estimates in linear models. *Journal of the American Statistical Association*, **70**, 428-433.

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, *Ann. Statist.*, **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.

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

Donoho, D. L. and Johnstone, I. I. (1994b). Minimax risk over ℓ_p balls for ℓ_q error. *Probability Theory and Related Fields*, **99**, 277-303.

Fan, J. (1997). Comments on "Wavelets in statistics: a review" by A. Antoniadis. *Journal of Italian Statistical Association*, **6**, 131-138.

Frank, I.E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109-148.

Fu, W. J. (1998). Penalized regression: the bridge versus the LASSO, *Journal of Computational and Graphical Statistics*, **7**, 397-416.

Gao, H. Y. and Bruce, A. G. (1997). WaveShrink with firm Shrinkage. *Statistica Sinica*, **7**, 855-874.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach.* Chapman and Hall, London.

Huber, P. (1981). *Robust estimation*, Wiley, New York.

Lehmann, E.L. (1983). *Theory of Point Estimation.* Pacific Grove, California: Wadsworth & Brooks/Cole.

Li, R. (2000). High-dimensional modeling via nonconcave penalized likelihood and local likelihood. Ph.D. dissertation, Department of Statistics, University of North Carolina at Chapel Hill.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* 2nd ed. Chapman and Hall, London.

Marron, J. S., Adak, S, Johnstone, I.M., Neumann, M. H. and Patil, P. (1998). Exact risk analysis of wavelet regression. *Journal Computational and Graphical Statistics*, **7**, 278-309.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*, Chapman and Hall, London.

Robinson, P.M. (1988), The stochastic difference between econometric and statistics, *Econometrica*, **56**, 531-547.

Stone, C.J., Hansen, M., Kooperberg, C. and Truong, Y.K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *Ann. Statist.*, **25**, 1371–1470.

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

Tibshirani, R. J. (1997). The LASSO method for variable selection in the Cox model. *Statistics in Medicine*, **16**, 385-395.

Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.