Asymptotics and Characterizations of Nonconvex Penalized Least Squares Estimators*

Runze Li
Department of Statistics
Pennsylvania State University
University Park, PA 16802-2111

September 18, 2001

Abstract

In this paper, nonconvex penalized least squares approaches are proposed to select significant predictors in linear and nonlinear regression models. The proposed approaches achieve the purpose of variable selection via automatically estimating the coefficients of insignificant variables to be zero rather than subset selection. Compared with best subset variable selection, they are very efficient in terms of computation, moreover, they reduce model complexity as effectively as the best subset variable selection. We establish characterizations of the penalized least squares. These characterizations provide us a guideline for choosing an appropriate penalty function for the purpose of variable selection. We also establish rates of convergence of the resulting estimator, and show that with proper choice of the regularization parameter and certain choices of penalty function, the resulting estimator possesses an oracle property in language similar to Donoho and Johnstone (1994), namely, they work as well as when the correct model is known in advance. Furthermore, we investigate the local asymptotic properties of the resulting estimator in order to understand finite sample performance of the proposed approaches. Using local quadratic approximation (Fan and Li, 2001), an iterative ridge regression is employed for finding the solution of the penalized least squares. A connection between the local quadratic approximation and majorize-minimize (MM) algorithm (Lange, Hunter and Yang, 2000) is established. This connection enables us to analyze the local and global convergence properties of the proposed algorithm by using the techniques related to the MM algorithm. Some simulation studies are conducted.

Abbreviated Title: Asymptotic and Characterization for Penalized LS.
AMS 2000 subject classifications: 62F12, 62J05.
Key Words and phrases: EM algorithm; LASSO; MM algorithm, Penalized Least Squares; SCAD; Wavelets.

This research was supported by a National Science Foundation (NSF) grant DMS-0102505. Parts of this research were conducted while the author received summer finance support from National Institute on Drug Abuse (NIDA) Grant 1-P50-DA10075. The author is grateful to Dr. Linda Collins and Dr. Joseph L. Schafer for their generous support.

1 Introduction

Regression is one of the most useful techniques in statistics. Variable selection is an important topic in the context of linear regression. There is a large body of literature on this subject. The monograph of Miller (1990) gives a comprehensive summary of various variable selection procedures as well as extensive bibliography of this field. Many of the variable selection procedures in use are stepwise selection procedures. These procedures usually consist of two stages: Firstly choose a criterion of variable selection, for instance, adjusted R^2 (Theil, 1961), C_p (Mallows, 1973), PRESS (Allen, 1974), AIC (Akaike, 1970, 1974), BIC (Schwarz, 1978), and among others. Then select an algorithm for finding subsets which fit well among algorithms: forward selection, stepwise regression, backward elimination, sequential replacement algorithms and the algorithm of exhaustive search over all possible subsets (referred to as best subset) and among others. These procedures are practically useful. However, most of them ignore stochastic errors inherited in the stages of variable selection. Hence, their theoretic properties are somewhat hard to understand. Furthermore, it is known that the best subset variable selection suffers from several drawbacks, the most severe of which is its lack of stability as analyzed, for example, by Breiman (1996).

Consider the linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_d x_{id} + \varepsilon_i = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \cdots, n,$$

where ε_i's are independent and identically distributed with mean 0 and finite variance σ^2. Without loss of generality, we assume in this paper that the design matrix is standardized so that each column has mean zero and variance one. Furthermore, we assume that y has zero mean and exclude the intercept β_0 from the linear model. Thus, throughout this paper, we focus on the standardized linear regression model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \cdots, n. \tag{1.1}$$

Many variable selection criteria are closely related to the penalized least squares, denoted by PLS for short,

$$\frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \sum_{j=1}^{d} \lambda_j p_j(|\beta_j|),$$

where $p_j(\cdot)$ is a penalty function and λ_j is a tuning parameter, which can be chosen by using a data-driven method, such as, cross validation (CV) and generalized cross validation (GCV, Craven and Wahba, 1979). The penalty functions $p_j(\cdot)$ and λ_j are not necessarily the same for all j.

This allows us to incorporate hierarchical prior information for the unknown coefficients by using different penalty functions or taking different values of λ_j for the different regression coefficients. For instance, we may wish to keep important predictors in linear regression models and hence do not want to penalize their corresponding coefficients. For ease of presentation, we assume that the penalty functions and the tuning parameters for all coefficients are the same. Moreover, we denote $\lambda_j p_j(\cdot)$ by $p_{\lambda_n}(\cdot)$, so the penalty function may be allowed to depend on λ, and the λ may be depend on n. Extensions to the case with different penalty functions and different tuning parameters do not involve any extra difficulties. Therefore, in this paper, we study properties and behaviors of the following PLS:

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|). \tag{1.2}$$

Many penalty functions have been used in the PLS. Here we give a brief discussion. Take the penalty function to be the entropy penalty, namely, $p_{\lambda_n}(|\theta|) = \frac{1}{2}\lambda_n^2 I(|\theta| \neq 0)$, which is also referred to as L_0-penalty in the literature, where $I(\cdot)$ is an indicator function. Note that the dimension or the size of a model equals to the number of nonzero regression coefficients in the model. This actually equals to $\sum_j I(|\beta_j| \neq 0)$. In other words, the PLS (1.2) with the entropy penalty can be rewritten as

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \frac{1}{2}\lambda_n^2|M|, \tag{1.3}$$

where $|M| = \sum_j I(|\beta_j| \neq 0)$, the size of the underlying candidate model. Hence, many popular variable selection criteria can be derived from the PLS (1.3) by choosing different values of λ_n. For instance, the AIC (or C_p), BIC, ϕ-criterion (Hannan and Quinn, 1979 and Shibata, 1984) and RIC (Foster and George, 1994) correspond to $\lambda_n = \sqrt{2}(\sigma/\sqrt{n})$, $\sqrt{\log n}(\sigma/\sqrt{n})$, $\sqrt{\log\log n}(\sigma/\sqrt{n})$ and $\sqrt{2\log(d)}(\sigma/\sqrt{n})$, respectively, although these criteria were motivated from different principles. Other proposals, including Zheng and Loh (1995), covariance inflation criterion (denoted by CIC, Tibshirani and Knight, 1999), generalized AIC such as GIC method (Nishii, 1984, Rao and Wu, 1989) and its analogous (Pötscher, 1989), and Shen and Ye (2001), also can be written as the form of (1.3). Thus, the best subset variable selection can be regarded as a solution of PLS in some sense. Asymptotic behaviors of various variable selection criteria derived from the PLS with the entropy penalty have been systematically studied by Shao (1997). Since the entropy penalty function is discontinuous, it requires searching over all possible subsets for finding the solution of this PLS. This is very expensive in terms of computational cost. An interesting fact is that, when the design matrix is column-orthogonal, the minimization problem (1.3) is equivalent to minimizing the PLS function componentwise with respect to β_j, therefore, the resulting estimator with the entropy penalty coincides with a hard thresholding rule (Figure 1 (a)), and the universal thresholding corresponds to $\lambda_n = \sqrt{2\log(d)}(\sigma/\sqrt{n})$. (Donoho and Johnstone, 1994).

Many authors have been working on the PLS with the L_p penalty $p_{\lambda_n}(|\theta|) = \lambda_n p^{-1}|\theta|^p$. It is well known that the L_2 penalty results in a ridge regression estimator. The quadratic penalty has been used widely in the context of smooth spline (Wahba, 1990). The L_p $(0 < p < 1)$ penalty yields the bridge regression (Frank and Friedman, 1993). The non-negative garrote (Breiman, 1995) depicted in Figure 1 (c) is closely related to the penalized least squares with the L_p penalty. See Section 2 for more arguments. With the L_1 penalty, the PLS estimator is the LASSO (Tibshirani, 1996, 1997), which becomes a soft thresholding rule displayed in Figure 1 (b) (Bickel, 1983, Donoho and Johnstone, 1994), when the design matrix is column-orthogonal. Fu (1998) proposed an algorithm to find the solution of LASSO. Knight and Fu (2000) has studied asymptotic properties of the PLS estimators with the L_p-penalty, but their mathematical formulation is different from ours.

With the aid of recent advances in computer technology, it is easy to collect data with many variables and with complicated structures. In practice, automatic and simultaneous variable selectors are desirable. Fan and Li (2001) proposed a class of variable selection approaches via nonconcave penalized likelihood. Their approaches achieve the purpose of variable selection by estimating the coefficients of insignificant variables to be 0. Under some regularity conditions and with proper rate of the regularization parameter and certain choices of the penalty function, they show that the resulting estimator via the nonconcave penalized likelihood approaches possesses an oracle property in the terminology of Donoho and Johnstone (1994). Namely, the proposed approaches work as well as when the correct submodel is known in advance. The LASSO and ridge regression do not possess this oracle property. Furthermore, compared with the best subset variable selection, their proposed approaches need much less computation. This encourages us to pursue further research on this topic. The main contributions of this paper are summarized as follows.

In this article, we first study characterizations of nonconvex penalized least squares. The characterizations have been shown by Antoniadis and Fan (2001) in wavelets settings, in which design matrices are orthonormal, and therefore the corresponding minimization problem reduces to a componentwise minimization problem. We show that the characterizations are valid for general linear regression models without the assumption of orthonormality on the design matrices. These characterizations provide us a guideline for choosing an appropriate penalty function to achieve the purpose of selecting significant variables.

To understand large sample properties of nonconvex PLS estimators, we establish their rates of convergence and show that with proper choice of the regularization parameter, they possess an oracle property. Our conditions on the regularization parameter and the penalty function are weaker than those in Fan and Li (2001). We also study the sampling properties of the PLS estimators with the entropy penalty. This allows us to theoretically compare the nonconvex penalized least square and the best subset variable selection by using the same mathematical formulation. We show theoretically that with proper choice of the regularization parameter, the proposed estimators reduce model complexity as effectively as the best subset variable selection. This has been empirically observed in Fan and Li's Monte Carlo simulation. Furthermore, the nonconvex PLS estimators require much less computation and is more stable for model prediction than the best subset variable selection.

Minimizing a nonconvex PLS function is challenging, because the target function is high-dimensional nonconvex function with singularities. We employed an iterative ridge regression algorithm to obtain solution of the nonconvex PLS by using local quadratic approximation (Fan and Li, 2001). To study the convergence of the iterative ridge regression algorithm, we establish a connection between the local quadratic approximation and the majorize-minimize (MM) algorithm (Lange, Hunter and Yang, 2000). The MM algorithm indeed is an extension of the EM algorithm (Demspter, Laird and Rubin, 1977). Also see a series of works by Meng and his coauthors (See Meng, 1993 and Meng and van Dyk, 1997 and references therein) for other extensions of the EM algorithm. The proposed algorithm with the local quadratic approximation is not an EM algorithm. However, using this connection, the local convergence and global convergence properties of the iterative ridge regression algorithm can be analyzed by using techniques employed in the EM algorithm (Wu, 1983 and Lange, 1995).

In practice, we are interested in finite sample performance. In other words, it is of interest to investigate the behaviors of the resulting PLS estimators when a model contains small coefficients in some sense. Hence, we study local asymptotic properties of the nonconvex PLS estimators in the presence of small regression coefficients. We first establish rates of convergence for the resulting PLS estimators, and then derive their limiting distribution. Our results show that with proper choice of regularization parameter, the nonconvex PLS estimators may significantly reduce model complexity by deleting the small coefficients on the order $O_p(n^{-1/2})$. This is similar to the thresholding rules, including the hard and soft thresholding rules, with the universal thresholding parameter in the context of wavelets (Donoho and Johnstone, 1994).

The rest of paper is organized as follows. In next section, we establish characterizations of the nonconvex PLS for general linear regression models. In Section 3, we study the sampling properties of the penalized least squares estimators. Their rates of convergence and oracle properties are established. In this section, we also study asymptotic properties of the PLS with the entropy penalty. In Section 4, an iterative ridge regression is employed to find the solution of the PLS by using local quadratic approximation. A connection between the local quadratic approximation and the MM algorithm is also established. Section 5 is devoted to studying local asymptotic properties of the PLS estimators. We extend the ideas for linear regression models to nonlinear regression models in Section 6. Some Monte Carlo simulation studies are presented in Section 7.

2 Characterizations for Penalized Least Squares Estimators

Minimizing (1.2) with respect to $\boldsymbol{\beta}$ results in a penalized least squares estimator of $\boldsymbol{\beta}$. As discussed by Fan and Li (2001), to achieve the purpose of variable selection for linear models, a good penalty function should yield an estimator with three properties: (a) **sparsity**: the resulting estimator may automatically set small estimated coefficients to be zero in order to reduce model complexity. In other words, the resulting estimator should be a thresholding rule. (b) **unbiasedness**: the resulting estimator is nearly unbiased when the true unknown coefficient is large in order to avoid unnecessary modeling bias; (c) **continuity**: the resulting estimator should be continuous in some sense in order to avoid instability in model prediction.

When the design matrix for model (1.1) is orthogonal, such as the setting of wavelets, minimizing (1.2) with respect to $\boldsymbol{\beta}$ reduces to a componentwise minimization problem, which is equivalent to minimizing the following simple PLS

$$\frac{1}{2}(z-\theta)^2 + p_{\lambda_n}(|\theta|),$$

where both z and θ are scalar variables. Based on this simple form of PLS, Antoniadis and Fan (2001) established characterizations for these three properties when the design matrix is orthogonal. As shown in the following proposition, those characterizations are still valid for general linear regression models without the assumption of orthogonality on the design matrix.

Denote

$$\mathbf{V}_n = \frac{1}{n}\mathbf{X}^T\mathbf{X} = (v_{ij}),$$

where $\mathbf{X}$ is the design matrix of model (1.2). Let $\mathbf{z}$ be the least squares estimator of $\boldsymbol{\beta}$

$$\mathbf{z} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = (z_1, \cdots, z_d)^T, \text{ and } z_j^* = z_j - \sum_{k\neq j} v_{jk}(\beta_k - z_k).$$

Proposition 2.1 *Let $p_\lambda(\cdot)$ be a nonnegative, nondecreasing and differentiable function on $(0, \infty)$. Furthermore, assume that the function $-\beta - p'_\lambda(\beta)$ is strictly unimodal on $(0, \infty)$, and the design matrix is full rank and standardized so that $\frac{1}{n}\sum_{i=1}^n x_{ij} = 0$, and $\frac{1}{n}\sum_{i=1}^n x_{ij}^2 = 1$. Then we have the following results.*

(a) Solutions to the minimization problem (1.2) exist.

(b) The solution $\widehat{\boldsymbol{\beta}}$ satisfies

$$\widehat{\beta}_j(\widehat{z}_j^*) = \begin{cases} 0, & \text{if } |\widehat{z}_j^*| \leq p_0, \\ \widehat{z}_j^* - sgn(\widehat{z}_j^*)p'_\lambda(|\widehat{\beta}_j|), & \text{if } |\widehat{z}_j^*| > p_0, \end{cases}$$

where $\widehat{z}_j^* = z_j - \sum_{k\neq j} v_{jk}(\widehat{\beta}_k - z_k)$ and $p_0 = \min_{\beta\geq 0}\{\beta + p'_\lambda(\beta)\}$. Moreover, $|\widehat{\beta}_j| \leq |\widehat{z}_j^*|$.

(c) If $p'_\lambda(\cdot)$ is nonincreasing, then for $|\widehat{z}_j^*| > p_0$, we have

$$|\widehat{z}_j^*| - p_0 \leq |\widehat{\beta}_j| \leq |\widehat{z}_j^*| - p_\lambda(|\widehat{z}_j^*|).$$

(d) When $p'_\lambda(\beta)$ is continuous on $(0, +\infty)$, the solution $\widehat{\beta}_j(z_j^*)$ is continuous in z_j^* if and only if the minimum of $|\beta| + p'_\lambda(|\beta|)$ is attained at the origin zero.

(e) When $p'_\lambda(\beta)$ is continuous on $(0, +\infty)$, and if $p'_\lambda(|\widehat{\beta}_j|) \to 0$ as $|\widehat{\beta}_j| \to +\infty$, then

$$\widehat{\beta}_j(\widehat{z}_j^*) = \widehat{z}_j^* - p'_\lambda(|\widehat{z}_j^*|)sgn(\widehat{z}_j^*) + o\{p'_\lambda(|\widehat{z}_j^*|)\}.$$

In particular, if the columns of $\mathbf{X}$ are orthogonal, then

$$\widehat{\beta}_j(z_j) = z_j - p'_\lambda(|z_j|)sgn(z_j) + o\{p'_\lambda(|z_j|)\}.$$

Proof. Denote $\mathbf{e} = \mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, the residual vector, and $Q(\boldsymbol{\beta})$ the penalized least squares function in (1.2). Decomposing the total sum of squares, we have

$$Q(\boldsymbol{\beta}) = \frac{1}{2n}\|\mathbf{e}\|^2 + \frac{1}{2}(\boldsymbol{\beta} - \mathbf{z})^T\mathbf{V}_n(\boldsymbol{\beta} - \mathbf{z}) + \sum_{j=1}^d p_\lambda(|\beta_j|) \geq 0. \tag{2.1}$$

As $\mathbf{V}_n$ is finite and positive definite, the PLS function $Q(\boldsymbol{\beta})$ tends to infinity as $\|\boldsymbol{\beta}\| \to \infty$. Thus, solutions of the minimization problem exist. This completes the proof of Part (a).

Note that the design matrix is standardized, the diagonal elements of $\mathbf{V}_n$ are equal to 1. Thus for $j = 1, \cdots, d$,

$$\begin{aligned} Q_j'(\boldsymbol{\beta}) &\equiv \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} \\ &= (\beta_j - z_j) + \sum_{k \neq j} v_{jk}(\beta_k - z_k) + p_\lambda'(|\beta_j|)\mathrm{sgn}(\beta_j) \\ &= \{|\beta_j| + p_\lambda'(|\beta_j|)\}\mathrm{sgn}(\beta_j) - z_j^* \end{aligned}$$

for $\beta_j \neq 0$. Let $\widehat{\boldsymbol{\beta}}$ be a minimizer of the PLS function $Q(\boldsymbol{\beta})$. Take $\beta_k = \widehat{\beta}_k$, $k \neq j$, and regard $Q_j'(\boldsymbol{\beta})$ as a function of β_j. Thus, by the definition of p_0, if $0 \leq |\widehat{z}_j^*| < p_0$, then the sign of $Q_j'(\beta_j)$ is the same as that of β_j. Therefore $\widehat{\beta}_j = 0$. On the other hand, if $|\widehat{z}_j^*| > p_0$, $\widehat{\beta}_j$ should satisfy

$$Q_j'(\widehat{\boldsymbol{\beta}}) = 0.$$

This implies that

$$\widehat{\beta}_j = \widehat{z}_j^* - p_\lambda'(|\widehat{\beta}_j|)\mathrm{sgn}(\widehat{\beta}_j).$$

Note that the $\widehat{\beta}_j$ and $\widehat{z}_j^*$ have the same sign because $\{|\widehat{\beta}_j| + p_\lambda(|\widehat{\beta}_j|)\}\mathrm{sgn}(\widehat{\beta}_j) = \widehat{z}_j^*$. Hence

$$\widehat{\beta}_j = \widehat{z}_j^* - p_\lambda'(|\widehat{\beta}_j|)\mathrm{sgn}(\widehat{z}_j^*).$$

This completes the proof of Part (b).

Now we prove Part (c). Since $\widehat{\beta}_j$ and $\widehat{z}_j^*$ have the same sign, without loss of generality, assume that $\widehat{z}_j^* > 0$. Denote $\beta_0 = \mathrm{argmin}_{\beta \geq 0}\{\beta + p_\lambda'(\beta)\}$. We first show that $\widehat{\beta}_j \geq \beta_0$. This is trivial when the function $\beta + p_\lambda'(\beta)$ is strictly increasing on $(0, +\infty)$. When the function $\beta + p_\lambda'(\beta)$ has a valley on $(0, \infty)$ and $\widehat{z}_j^* \geq p_0$, the equation

$$\beta_j + p_\lambda'(\beta_j)\mathrm{sgn}(\beta_j) - z_j^* = 0$$

has two possible roots on $(0, +\infty)$. The larger one is the minimizer because the derivative function at that point is increasing. Then

$$p_\lambda'(\widehat{\beta}_j) \leq p_\lambda'(\beta_0) \leq \beta_0 + p_\lambda'(\beta_0) = p_0.$$

Thus, the first equality in Part (c) holds. Again notice that $|\widehat{\beta}_j| \leq |\widehat{z}_j^*|$, the second inequality in Part (c) also holds.

As for Part (d), it is clear that continuity of the solution $\widehat{\beta}_j$ at the point at $\widehat{z}_j^*$ if and only if the minimum of the function $|\beta| + p'_\lambda(|\beta|)$ is attained at zero. The continuity at other location follows directly from the monotonicity and continuity of the function $\beta + p'_\lambda(\beta)$ in the interval $(0, \infty)$.

Note that

$$\widehat{\beta}_j(\widehat{z}_j^*) = \widehat{z}_j^* - p'_\lambda\{|\widehat{\beta}_j(\widehat{z}_j^*)|\}\text{sgn}(\widehat{z}_j^*),$$

and the continuity of $p'_\lambda(\cdot)$, the last conclusion follows.

These characterizations provide us a guideline on how to choose an appropriate penalty function for the PLS. Figure 2 depicts the plot of $\beta + p'_\lambda(\beta)$ over $\beta > 0$ for various penalty functions, which gives us more insights into how to choose the penalty function. Note that z_j^* is continuous in $\mathbf{z}$ and $\boldsymbol{\beta}$. Part (d) of Proposition 2.1 indicates that in order for the resulting PLS estimators to be continuous in data, it is needed to choose a penalty function so that $|\beta| + p'_\lambda(|\beta|)$ attains its minimum at the origin. The L_p penalty satisfies this condition only when $p \geq 1$. From Part (b) of Proposition 2.1, in order to automatically delete small coefficients, the minimum of $|\beta| + p'_\lambda(|\beta|)$ should be positive. The L_p penalty satisfies this condition only when $0 < p \leq 1$. On the other hand, from Part (e) of Proposition 2.1, to reduce unnecessary modeling bias, $p'_\lambda(|\beta|)$ should tend to zero as $|\beta| \to \infty$. The L_1 penalty does not satisfy this condition. Therefore, none of the L_p penalties simultaneously satisfy these three mathematical requirements for continuity, sparsity and unbiasedness. See also Figure 2 (b).

The hard thresholding penalty, defined by

$$p_\lambda(|\beta|) = \lambda^2 - (|\beta| - \lambda)^2 I(|\beta| < \lambda) \tag{2.2}$$

is a smooth version of the entropy penalty. This penalty was proposed by Fan (1997) and improved by Antoniadis (1997). We will use the acronym HARD for all procedures using the hard thresholding penalty. When the design matrix is column-orthogonal, the solutions of the penalized least squares with the entropy penalty and with the hard thresholding penalty are the same. The solution actually is a hard-thresholding rule. It is clear that the hard thresholding rule is not continuous in $\mathbf{z}$ because the hard-thresholding penalty does not satisfy the condition in Proposition 2.1 (d). See also Figure 2 (a)

The smoothly clipped absolute deviation (SCAD) penalty, proposed by Fan (1997),

$$p'_\lambda(\beta) = \lambda I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{a-1} I(\beta > \lambda) \text{ for some } a > 2 \quad \text{and} \quad \beta > 0 \tag{2.3}$$

with $p_\lambda(0) = 0$ satisfies all the three mathematical conditions in Proposition 2.1 (b), (d) and (e). See Figure 2 (a). We use the acronym SCAD for all procedures using the SCAD penalty. When the design matrix is column-orthogonal, the SCAD has a closed form solution (Figure 1 (d)). The SCAD involves two unknown parameters λ and a. Fan and Li (2001) suggested using $a = 3.7$ based on a Bayesian argument. They also found that the performances of the SCAD with $a = 3.7$ is as good as those of the SCAD with the value of a chosen by the GCV. Hence, this value will be used throughout the whole paper.

There are many penalty functions satisfying the three mathematical conditions. For instance, the following modified L_p $(0 < p < 1)$ penalty, defined by

$$p'_\lambda(|\beta|) = \lambda\beta_0^{p-1}I(|\beta| \le \beta_0) + \lambda|\beta|^{p-1}I(|\beta| > \beta_0) \tag{2.4}$$

with $p_\lambda(0) = 0$, where $\beta_0 = \{\lambda(1-p)\}^{1/(2-p)}$. This modification makes the solution of the penalized least squares continuous. See Figure 2 (c). With this modification, the resulting estimator is similar to nonnegative garrote (Breiman, 1995), when the design matrix is column-orthogonal.

Now we introduce a new family of penalty functions, namely, log-transformed L_p $(0 < p \le 1)$ penalty, defined by

$$p_\lambda(|\beta|) = p^{-1}\log\left\{\frac{1 + \lambda(1 + |\beta|)^p}{1 + \lambda}\right\}. \tag{2.5}$$

It can be verified that this family satisfies the conditions imposed in Proposition 2.1 (b), (d) and (e). See Figure 2 (c). When $p = 1$, there is an analytic expression for the solution of the penalized least squares when design matrix is column-orthogonal.

In summary, a good penalty function should be symmetric, nonconvex on $(0, \infty)$, and be singular at the origin in order to yield a sparse solution. The penalty function should be bounded by a constant in order to reduce possible estimation bias, and satisfy certain conditions in order to produce continuous solutions. Many penalty functions may satisfy the three mathematical conditions. A challenging question arising here is which penalty results in the best solution under some criteria. Further research is needed. We give some empirical comparisons in Section 7. Here we give a brief note. As shown in next section, the resulting estimators with the L_p $(p \ge 1)$, or the log-transformed L_p $(0 < p < 1)$ penalty do not possess an oracle property. However, the hard-thresholding penalty, SCAD, L_p $(0 < p < 1)$ or modified L_p $(0 < p < 1)$ yields a solution with the oracle property.

3 Oracle Properties

Consider the linear regression model

$$\mathbf{y} = \mathbf{X}_1^T\boldsymbol{\beta}_1 + \mathbf{X}_2^T\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon},$$

where $E(\boldsymbol{\varepsilon}|\mathbf{x}) = 0$ and $\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$. To formulate the oracle property, without loss of generality, assume that $\boldsymbol{\beta}_2 = \mathbf{0}$, and all components of $\boldsymbol{\beta}_1$ are not equal to 0. This implies that the first part in the model is significant, while the second part is insignificant and should be deleted from the model. An ideal estimator for $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_2 = 0,$$

and

$$\widehat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1^T\mathbf{X})^{-1}\mathbf{X}_1^T\mathbf{y}.$$

This estimator correctly identifies the true model and efficiently estimates the nonzero components as if the true model were known. Therefore, it is referred to as an oracle estimator, and this property is termed an oracle property, a terminology from Donoho and Johnstone (1994). The oracle estimator deletes an insignificant variable via estimating its coefficient to be zero. In other words, the oracle estimator selects a model and estimates the regression coefficients of significant variables simultaneously. The oracle estimator not only identifies the zero components correctly, but also estimates the nonzero components more efficiently than the ordinary least squares estimator. Therefore, when the error is normally distributed, it is more efficient than the maximum likelihood estimator for $\boldsymbol{\beta}$. This is very analogous to the super-efficiency phenomenon in the Hodges example (see page 405 of Lehmann, 1983). Fan and Li (2001) has established the oracle property for their penalized likelihood estimators with certain choices of the penalty function and under a proper convergent rate of the regularization parameter. In this section, we will establish rates of convergence for the PLS estimators, and show that they possess the oracle property. The conditions on penalty function and the regularization parameter in Theorem 3.2 below is weaker than those in Theorem 2 of Fan and Li (2001). Using our conditions, it is easy to verify that the resulting estimator with the L_p $(0 < p < 1)$ penalty or the modified L_p penalty possesses the oracle property under condition (3.4) and $\sqrt{n}\lambda_n \to 0$. However, the conditions in Theorem 2 of Fan and Li (2001) require that $\sqrt{n}\lambda_n \to \infty$. This implies that one cannot directly apply their results to the L_p penalty with $0 < p < 1$. Thus, our results may be applied for more general situations than those of Fan and Li (2001).

We assume hereafter that $p_\lambda(\cdot)$ is a nonnegative, nondecreasing function with $p_\lambda(0) = 0$. Further assume that the penalty function $p_\lambda(\cdot)$ has a second order continuous derivative at nonzero components of $\boldsymbol{\beta}_0$, the true value of $\boldsymbol{\beta}$. Let

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

where $\boldsymbol{\beta}_{10}$ consists of the first s components of $\boldsymbol{\beta}_0$. Without loss of generality, assume that $\boldsymbol{\beta}_{20} = \mathbf{0}$ and all components of $\boldsymbol{\beta}_{10}$ are not equal to zero. Denote

$$a_n = \max\{|p'_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}, \qquad \text{and} \qquad b_n = \max\{|p''_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}. \tag{3.1}$$

Theorem 3.1 *Suppose that the observations $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$ are independent and identically distributed from the model (1.1), and assume that the matrix $E\mathbf{x}\mathbf{x}^T$, denoted by $\mathbf{V}$, is finite and positive definite, and the random error has mean zero and finite positive variance σ^2. If $a_n \to 0$ and $b_n \to 0$, then with probability tending to one, there exists a local minimizer $\widehat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| = O_P(n^{-1/2} + a_n)$.*

Proof: Let $\alpha_n = n^{-1/2} + a_n$. It is sufficient to show that with probability tending to one, there exists a local minimum in the ball $\{\boldsymbol{\beta}_0 + \alpha_n \mathbf{u} : \|\mathbf{u}\| \leq C\}$ for a large constant C.

To this end, denote

$$D_n(\mathbf{u}) = Q(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - Q(\boldsymbol{\beta}_0),$$

and we shall show that $\inf_{||\mathbf{u}||=C} D_n(\mathbf{u}) > 0$ with probability tending to one. It follows by the definition of $D_n(\mathbf{u})$ that

$$D_n(\mathbf{u}) = \frac{1}{2n}\sum_{i=1}^n [\{y_i - \mathbf{x}_i^T(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u})\}^2 - (y_i - \mathbf{x}_i^T\boldsymbol{\beta}_0)^2] + \sum_{j=1}^d \{p_{\lambda_n}(|\beta_{j0} + \alpha_n u_j|) - p_{\lambda_n}(|\beta_{j0}|)\}. \tag{3.2}$$

The first term in the right-hand side of (3.2) equals to

$$\frac{\alpha_n^2}{2}\mathbf{u}^T\left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{u} - \alpha_n\left\{\frac{1}{n}\sum_{i=1}^n \mathbf{u}^T\mathbf{x}_i(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_0)\right\}.$$

By the Strong Law of Large Numbers (SLLN), $\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T = \mathbf{V}\{1 + o_p(1)\}$. Note that $\frac{1}{n}\sum_{i=1}^n \mathbf{u}^T\mathbf{x}_i(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_0) = \sqrt{\sigma^2\mathbf{u}^T\mathbf{V}\mathbf{u}}O_P(n^{-1/2})$ by using $R = E(R) + O_P\{\sqrt{\mathrm{Var}(R)}\}$ for any random variable R with a finite second moment.

Using $p_{\lambda_n}(0) = 0$, it follows that

$$\begin{aligned} D_n(\mathbf{u}) &\geq \frac{1}{2}\alpha_n^2 \mathbf{u}^T\mathbf{V}\mathbf{u}\{1 + o_P(1)\} + \sqrt{\sigma^2 \mathbf{u}^T\mathbf{V}\mathbf{u}}O_P(n^{-1/2}\alpha_n) \\ &\quad + \sum_{j=1}^{s}\{p_{\lambda_n}(|\beta_{j0} + \alpha_n u_j|) - p_{\lambda_n}(|\beta_{j0}|)\}. \end{aligned} \tag{3.3}$$

Note that $O_P(n^{-1/2}\alpha_n) = O_P(\alpha_n^2)$, and that $\mathbf{V}$ is finite and positive definite. By choosing a sufficiently large C, the second term in (3.3) will be dominated by the first term, uniformly in $\|\mathbf{u}\| = C$. By Taylor's expansion, the third term on the right hand side of (3.3) becomes

$$\sum_{j=1}^{s}\left[\alpha_n p'_{\lambda_n}(|\beta_{j0}|)\text{sgn}(\beta_{j0})u_j + \alpha_n^2 p''_{\lambda_n}(|\beta_{j0}|)u_j^2\{1 + o(1)\}\right],$$

which is bounded by

$$\sqrt{s}\alpha_n a_n\|\mathbf{u}\| + \alpha_n^2 b_n\|\mathbf{u}\|^2.$$

This is also dominated by the first term of (3.3) as $b_n \to 0$. Hence, by choosing sufficiently large C, $D_n(\mathbf{u}) > 0$ with probability tending to one. This completes the proof of the theorem.

Under the regularity conditions in Theorem 3.1, the resulting estimator is root n consistent if $a_n = O_P(n^{-1/2})$ and $b_n \to 0$, which requires that $\lambda_n \to 0$ for the HARD and SCAD penalty, and that $\lambda_n = O_P(n^{-1/2})$ for the L_p $(p > 0)$ penalty and penalties defined by (2.4) and (2.5).

Theorem 3.2 *Under the conditions of Theorem 3.1, assume that*

$$\liminf_{n\to\infty}\liminf_{\beta\to 0^+}\sqrt{n}p'_{\lambda_n}(\beta) = +\infty, \tag{3.4}$$

then with probability tending to 1, the root n consistent local minimizer $\widehat{\boldsymbol{\beta}} = \begin{pmatrix}\widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2\end{pmatrix}$ *in Theorem 3.1 must satisfy:*

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

*(ii) (***Asymptotic normality***)*

$$\sqrt{n}(\mathbf{V}_{11} + \Sigma)\left\{\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\mathbf{V}_{11} + \Sigma)^{-1}\mathbf{b}\right\} \to N\left(\mathbf{0}, \sigma^2\mathbf{V}_{11}\right)$$

in distribution, where $\mathbf{V}_{11}$ *consists of the first s rows and columns of* $\mathbf{V}$ *and*

$$\Sigma = diag\left\{p''_{\lambda_n}(|\beta_{10}|), \cdots, p''_{\lambda_n}(|\beta_{s0}|)\right\}, \tag{3.5}$$

$$\mathbf{b} = \left(p'_{\lambda_n}(|\beta_{10}|)sgn(\beta_{10}), \cdots, p'_{\lambda_n}(|\beta_{s0}|)sgn(\beta_{s0})\right)^T. \tag{3.6}$$

Proof: To prove Part (i), we show that with probability tending to one, for any given $\boldsymbol{\beta}_1$ satisfying that $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_P(n^{-1/2})$ and any constant $C > 0$,

$$Q\left\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \mathbf{0}\end{pmatrix}\right\} = \min_{\|\boldsymbol{\beta}_2\| \le Cn^{-1/2}} Q\left\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}\right\}. \tag{3.7}$$

To this end, we prove that with probability tending to 1, for any $\boldsymbol{\beta}_1$ satisfying $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10} = O_P(n^{-1/2})$ and for some small $\varepsilon_n = Cn^{-1/2}$ and $j = s+1, \cdots, d$, $\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}$ has the same sign as that of β_j when $0 < |\beta_j| < \varepsilon_n$.

By some straightforward algebraic calculation, it follows that when $\beta_j \neq 0$, for $j = s+1, \cdots, d$,

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} = -\frac{1}{n}\mathbf{x}_{(j)}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0) + \frac{1}{n}\mathbf{x}_{(j)}^T\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + p'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j),$$

where $\mathbf{x}_{(j)}$ is the j-th column of $\mathbf{X}$.

Note that by the CLT and the SLLN,

$$\frac{1}{n}\mathbf{x}_{(j)}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0) = O_P(n^{-1/2}), \quad \text{and} \quad \frac{1}{n}\mathbf{x}_{(j)}^T\mathbf{X} = EX_j\mathbf{x}^T + o_P(1),$$

where X_j is the j-th component of $\mathbf{x}$.

When $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_P(n^{-1/2})$ and $\|\boldsymbol{\beta}_2\| \le Cn^{-1/2}$, which implies that $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2})$, it follows that

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

Since $\liminf_{\beta \to 0^+} \sqrt{n}p'_{\lambda_n}(\beta) \to \infty$ as $n \to \infty$, the sign of the derivative is the same as that of β_j. Thus we complete the proof of Part (i).

Now we prove Part (ii). Following the proof of Theorem 3.1, it can be shown that there exists a $\widehat{\boldsymbol{\beta}}_1$ that is a root n consistent local minimizer of $Q\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \mathbf{0}\end{pmatrix}\}$, regarded as a function of $\boldsymbol{\beta}_1$, and satisfying the following equations:

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

Denote $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \mathbf{0})^T$. Note that $\widehat{\boldsymbol{\beta}}_1$ is a consistent estimator,

$$\begin{aligned}\frac{\partial Q(\widehat{\boldsymbol{\beta}})}{\partial \beta_j} &= -\frac{1}{n}\mathbf{x}_{(j)}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0) + \frac{1}{n}\mathbf{x}_{(j)}^T\mathbf{X}_{(1)}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) + p'_{\lambda_n}(|\widehat{\beta}_j|)\mathrm{sgn}(\widehat{\beta}_j)\\ &= -\frac{1}{n}\mathbf{x}_{(j)}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0) + \mathbf{V}_{(j)}^T(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})\{1 + o_P(1)\}\\ &\quad + \left[p'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0}) + \{p''_{\lambda_n}(|\beta_{j0}|) + o_P(1)\}(\widehat{\beta}_j - \beta_{j0})\right].\end{aligned}$$

where $X_{(1)}$ consists of the first s columns of $\mathbf{X}$ and $\mathbf{V}_{(j)}$ is the jth-column of V_{11}. It follows by Slutsky's Theorem and the CLT that

$$\sqrt{n}(\mathbf{V}_{11} + \Sigma)\left\{\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\mathbf{V}_{11} + \Sigma)^{-1}\mathbf{b}\right\} \to N\left(\mathbf{0}, \sigma^2 \mathbf{V}_{11}\right)$$

in distribution.

Finally we show that $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \mathbf{0})^T$ is a root n consistent local minimizer of $Q(\boldsymbol{\beta})$. It is clear that $\widehat{\boldsymbol{\beta}}$ is root n consistent. Assume that $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1^T, \tilde{\boldsymbol{\beta}}_2^T)^T$ is a local minimizer of $Q(\boldsymbol{\beta})$ satisfying $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2})$. Then $Q(\tilde{\boldsymbol{\beta}}) \leq Q(\widehat{\boldsymbol{\beta}})$. By (3.7), we have $\tilde{\boldsymbol{\beta}}_2 = \mathbf{0}$. Thus by the definition of $\widehat{\boldsymbol{\beta}}_1$, we have $Q(\tilde{\boldsymbol{\beta}}) \geq Q(\widehat{\boldsymbol{\beta}})$. Thus, $Q(\tilde{\boldsymbol{\beta}}) = Q(\widehat{\boldsymbol{\beta}})$. Therefore, $\widehat{\boldsymbol{\beta}}$ is a root n consistent local minimizer of $Q(\boldsymbol{\beta})$. This completes the proof of the theorem.

For the HARD penalty and the SCAD penalty, if $\lambda_n \to 0$, and $\sqrt{n}\lambda_n \to \infty$, then $a_n = 0$. From Theorem 3.1, there exists a root n consistent local minimizer. Furthermore, from Theorem 3.2, this root n consistent local minimizer satisfies that $\widehat{\boldsymbol{\beta}}_2 = 0$ and $\sqrt{n}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})$ has an asymptotic normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{V}_{11}^{-1}$ as $b_n = 0$. Therefore the resulting estimators possess the oracle property.

For the L_p penalty with $0 < p < 1$, if $\lambda_n = O_P(n^{-1/2})$, then by Theorems 3.1 and 3.2, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})$ has an asymptotic normal distribution with mean $\sqrt{n}\mathbf{b}$ and covariance matrix $(\mathbf{V}_{11} + \Sigma)^{-1}\mathbf{V}_{11}(\mathbf{V}_{11} + \Sigma)^{-1}$. To reduce the bias, one should take $\lambda_n = o_P(n^{-1/2})$ so that $\sqrt{n}\mathbf{b} \to 0$ as $n \to \infty$, and then the resulting estimator possesses the oracle property.

Now we discuss the modified L_p penalty with $0 < p < 1$. Similar to the L_p penalty, one should take $\lambda_n = o_P(n^{-1/2})$. The condition (3.4) is satisfied when $\sqrt{n}\lambda_n^{1/(2-p)} \to \infty$. This can be done, for instance, by taking λ_n such that $\lambda_n = O_p(n^{-1/2}(\log n)^{-1})$. Therefore, the resulting estimator with the modified L_p penalty possesses the oracle property.

For the L_p penalty, the condition (3.4) cannot be satisfied when $p > 1$. While for L_1 penalty, the transformed L_1 penalty, and the log-transformed L_p penalty $(0 < p \leq 1)$, the root n consistency requires that $\lambda_n = O_P(n^{-1/2})$. On the other hand, the condition (3.4) requires that $\sqrt{n}\lambda_n \to \infty$. These two conditions cannot be satisfied simultaneously. We conjecture that the oracle property does not hold for the resulting estimator with the L_1 and the log-transformed L_p penalty.

The results in Theorems 3.1 and 3.2 cannot be directly applied for the PLS estimator with the entropy penalty $p_{\lambda_n}(|\theta|) = \frac{1}{2}\lambda_n^2 I(|\theta| \neq 0)$. As discussed in Section 1, using this penalty, many existing variable selection criteria can be derived by taking different values of the regularization parameter λ_n. Thus, it is of interest to study the asymptotic properties of the resulting estimator with the entropy penalty using our mathematical formulation. This will provide us a comparison between the best subset variable selection and the PLS with nonconvex penalty under the same mathematical formulation.

Theorem 3.3 *Suppose that the observation $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$ are independent and identically distributed from the model (1.1), and assume that the matrix $\mathbf{V} = E\mathbf{x}\mathbf{x}^T$ is finite and positive definite, and the random error has mean zero and finite variance σ^2. For the entropy penalty $p_{\lambda_n}(|\beta|) = \frac{1}{2}\lambda_n^2 I(|\beta| \neq 0)$, if $\lambda_n \to 0$, then with probability tending to one, there exists a local minimizer $\widehat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2})$.*

Proof. It is sufficient to show that for any given $\varepsilon > 0$, there exists a large constant C such that

$$P\left\{\inf_{\|\mathbf{u}\|=C} Q(\boldsymbol{\beta} + n^{-1/2}\mathbf{u}) > Q(\boldsymbol{\beta}_0)\right\} \geq 1 - \varepsilon. \tag{3.8}$$

Denote

$$D_n(\mathbf{u}) = Q(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) - Q(\boldsymbol{\beta}_0),$$

and

$$\tilde{p}_{\lambda_n}(|\beta|) = \frac{1}{2}\{\lambda_n^2 - (|\beta| - \lambda_n)^2 I(|\beta| < \lambda_n)\}.$$

Using $p_{\lambda_n}(0) = 0$ and $\frac{1}{2}\lambda_n^2 I(|\beta| \neq 0) \geq \tilde{p}_{\lambda_n}(|\beta|)$, it follows along the same arguments in the proof of Theorem 3.1 that

$$\begin{aligned} D_n(\mathbf{u}) \geq\ & \frac{1}{2n}\mathbf{u}^T\mathbf{V}\mathbf{u}\{1 + o_P(1)\} - \frac{1}{n}\frac{1}{\sqrt{n}}\mathbf{u}^T\mathbf{X}^T(\mathbf{y} - \mathbf{X}^T\boldsymbol{\beta}_0) \\ & + \sum_{j=1}^{s}\{\tilde{p}_{\lambda_n}(|\beta_{j0} + n^{-1/2}u_j|) - \frac{1}{2}\lambda_n^2\}. \end{aligned} \tag{3.9}$$

The third term in the right hand side of (3.9) equals to

$$\sum_{j=1}^{s}\{\tilde{p}_{\lambda_n}(|\beta_{j0} + n^{-1/2}u_j|) - \tilde{p}_{\lambda_n}(|\beta_{j0}|)\} - \frac{1}{2}\sum_{j=1}^{s}\{(|\beta_{j0}| - \lambda_n)^2 I(|\beta_{j0}| < \lambda_n)\}. \tag{3.10}$$

Taking n large enough, the second term in (3.10) equals to 0 as $\lambda_n \to 0$. The first term in (3.10) is bounded by

$$\sqrt{s}n^{-1/2}\tilde{a}_n\|\mathbf{u}\| + n^{-1}\tilde{b}_n\|\mathbf{u}\|^2.$$

where

$$\tilde{a}_n = \max\{\tilde{p}'_{\lambda_n}(|\beta_{j0}|) : \beta_{j0} \neq 0\}, \qquad \text{and} \qquad \tilde{b}_n = \max\{|\tilde{p}''_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\},$$

which are equal to 0 provided that λ_n is small enough. Thus, by choosing a sufficiently large C, the first term in the right hand side of (3.9) will dominate the other two terms in the right hand side of (3.9). Therefore (3.8) holds. This completes the proof of the theorem.

Lemma 3.1 *Under the conditions of Theorem 3.3, for the entropy penalty, if* $\lambda_n \to 0$ *and* $\sqrt{n}\lambda_n \to \infty$*, then with probability tending to one, for any given* $\boldsymbol{\beta}_1$ *satisfying* $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_P(n^{-1/2})$ *and constant C*

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

Proof. We want to show that for any $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_P(n^{-1/2})$ and $\|\boldsymbol{\beta}_2\| \leq Cn^{-1/2}$, with probability tending to one, the following inequality holds:

$$Q\left\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \mathbf{0}\end{pmatrix}\right\} - Q\left\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}\right\} \leq 0. \tag{3.11}$$

If $\boldsymbol{\beta}_2 = \mathbf{0}$, then the inequality holds. Now consider $\boldsymbol{\beta}_2 \neq \mathbf{0}$, which implies that $\sum_{j=s+1}^d I(\beta_j \neq 0) \geq 1$. By the definition of $Q(\boldsymbol{\beta})$,

$$\begin{aligned} & Q\left\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \mathbf{0}\end{pmatrix}\right\} - Q\left\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}\right\} \\ = & \frac{1}{2n}\|\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1\|^2 - \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 - \frac{\lambda_n^2}{2}\sum_{j=s+1}^d I(\beta_j \neq 0) \\ = & \frac{1}{n}(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_{10})^T\mathbf{X}_2\boldsymbol{\beta}_2 + \frac{1}{n}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T\mathbf{X}_1^T\mathbf{X}_2\boldsymbol{\beta}_2 \\ & -\frac{1}{2n}\boldsymbol{\beta}_2^T(\mathbf{X}_2^T\mathbf{X}_2)\boldsymbol{\beta}_2 - \frac{\lambda_n^2}{2}\sum_{j=s+1}^d I(\beta_j \neq 0). \end{aligned} \tag{3.12}$$

Note that $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_P(n^{-1/2})$ and $\|\boldsymbol{\beta}_2\| = O_P(n^{-1/2})$. By standard arguments, the first three terms in the right hand side of (3.12) are of the order $O_P(n^{-1})$.

$$Q\left\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \mathbf{0}\end{pmatrix}\right\} - Q\left\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}\right\} = -\frac{1}{n}\left\{\frac{n\lambda_n^2}{2}\sum_{j=s+1}^d I(\beta_j \neq 0) + O_P(1)\right\}.$$

Note that $\sqrt{n}\lambda_n \to \infty$ and $\sum_{j=s+1}^d I(\beta_j \neq 0) \geq 1$. With probability tending to one, (3.11) holds. This completes the proof of the lemma.

Theorem 3.4 *Suppose that the conditions in Theorem 3.3 hold. For the entropy penalty, if* $\lambda_n \to 0$ *and* $\sqrt{n}\lambda_n \to \infty$*, then, with probability tending to one, the root n consistent local minimizer* $\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{pmatrix}$ *in Theorem 3.3 must satisfy:*

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

*(ii) (***Asymptotic normality***)*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \to N\left(\mathbf{0}, \sigma^2 \mathbf{V}_{11}^{-1}\right)$$

in distribution, where $\mathbf{V}_{11}$ *consists of the first s rows and columns of* $\mathbf{V}$*.*

Proof. Part (i) follows from Lemma 3.1. The proof of Part (ii) is similar to that of Part (ii) in Theorem 3.2. To save space, we omit it here.

From Theorem 3.4, the BIC and ϕ-criterion results in a solution possessing the oracle property. On the other hand, the AIC and RIC do not satisfy the condition $\sqrt{n}\lambda_n \to \infty$. We conjecture that the corresponding resulting estimators do not possess the oracle property when the dimension d is finite. From Theorems 3.2 and 3.4, we find that the SCAD, HARD and the PLS estimators with the entropy penalty have the same limiting distribution provided that $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$. Therefore, the SCAD and HARD reduce model complexity as effectively as the best subset variable selection. On the other hand, note that the entropy penalty is discontinuous. To find the solution of the PLS solution, it is required to search over all possible subsets, like the best subset variable selection. This is very expensive in terms of computational cost, and impossible, even when the number of covariates is moderate. Furthermore, the resulting estimator of the PLS with the entropy penalty, for instance, the hard thresholding rule, is discontinuous. Thus the resulting solution suffers the drawback of instability.

4 Iterative ridge regression and MM algorithm

Note that the penalty function satisfying the three mathematical requirements derived in Section 2 should be symmetric and nonconvex over $(0, +\infty)$. Moreover, the minimization problem of the penalized least squares is high-dimensional. Therefore, it is very challenging in finding solution of the nonconvex PLS. The singularity at the origin imposes more challenging in minimizing the PLS. Furthermore, some good penalty functions, such as the HARD and the SCAD, may not have the second derivative at some points. Several authors have proposed algorithms for finding solution of the PLS with the L_1 penalty. Tibshirani (1996) proposed an algorithm for solving constrained least squares problems of the LASSO, while Fu (1998) provided a "shooting algorithm" for the LASSO, See also LASSO2 submitted by Berwin Turlach at Statlib (http://lib.stat.cmu.edu/S/). Fan and Li (2001) suggested the use of local quadratic approximation for the general nonconvex penalty function. But the convergence properties of their algorithm have not been studied yet. In this section, we employ an iterative ridge regression for finding the solution of the PLS by using the local quadratic approximation. Furthermore, we establish a connection between the local quadratic approximation and the majorize-minimize or minorize-maximize (MM) algorithm, proposed by Lange, Hunter and Yang (2000). The MM algorithm is an extension of the EM algorithm (Dempster, Laird and Rubin, 1977) and shares the principle of the EM algorithm. This connection enables us to analyze local and global convergence of the iterative ridge regression algorithm by employing the techniques used for the EM algorithm (Wu 1983, and Lange, 1995). This connection indicates that the local quadratic approximation provides a general mechanism for constructing a majorization function in the MM algorithm for a nonconvex function in a minimization problem.

4.1 Iterative ridge regression

In order to apply the Newton-Raphson algorithm to their penalized likelihood function, Fan and Li suggested to locally approximate the nonconvex penalty function by a quadratic function as follows. Given an initial value $\boldsymbol{\beta}^0$ that is close to the true value of $\boldsymbol{\beta}$. When β_j^0 is very close to zero, for instance, its absolute value is less than half of its standard deviation, set $\widehat{\beta}_j = 0$; otherwise, the penalty $p_\lambda(|\beta_j|)$ can be locally approximated by the quadratic function as

$$[p_\lambda(|\beta_j|)]' = p_\lambda'(|\beta_j|)\text{sgn}(\beta_j) \approx \left\{p_\lambda'(|\beta_j^0|)/|\beta_j^0|\right\}\beta_j.$$

In other words,

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^0|) + \frac{1}{2}\left\{p_\lambda'(|\beta_j^0|)/|\beta_j^0|\right\}(\beta_j^2 - \{\beta_j^0\}^2). \tag{4.1}$$

The top panel of Figure 3 depicts this approximation for SCAD, $L_{0.5}$ and L_1 penalties at two different locations. With the local quadratic approximation, the solution for the PLS can be found by iteratively computing the following ridge regression with an initial value $\boldsymbol{\beta}^0$:

$$\boldsymbol{\beta}^{(1)} = \{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\boldsymbol{\beta}^0)\}^{-1}\mathbf{X}^T\mathbf{y},$$

where

$$\Sigma_\lambda(\boldsymbol{\beta}^0) = \mathrm{diag}\{p'_\lambda(|\beta_1^0|)/|\beta_1^0|, \cdots, p'_\lambda(|\beta_d^0|)/|\beta_d^0|\}.$$

Note that we update the initial value $\boldsymbol{\beta}^0$ in each step during iteration and take the unpenalized least squares estimator of $\boldsymbol{\beta}$ as the initial value for the first step. When the algorithm converges, the estimator satisfies the condition

$$\mathbf{x}_{(j)}^T(\mathbf{X}\boldsymbol{\beta}^0 - \mathbf{y}) + np'_\lambda(|\widehat{\beta}_j^0|)\mathrm{sgn}(\widehat{\beta}_j^0) = 0, \tag{4.2}$$

the penalized least squares equation for nonzero components, where $\mathbf{x}_{(j)}$ is the j-th column of $\mathbf{X}$ for $j = 1, \cdots, s$.

With the local quadratic approximation, the iterative ridge regression acts as the Newton-Raphson algorithm. Thus, a robust empirical standard error formula for the estimated coefficients can be derived from the iterative ridge regression. We use the following sandwich formula to estimate the covariance of the PLS estimate

$$\widehat{\mathrm{cov}}(\widehat{\boldsymbol{\beta}}) = \widehat{\sigma}^2\{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\boldsymbol{\beta}^0)\}^{-1}(\mathbf{X}^T\mathbf{X})\{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\boldsymbol{\beta}^0)\}^{-1}, \tag{4.3}$$

where $\widehat{\sigma}^2$ is an estimate of σ^2, for example, the mean squared errors. Here the formula is applied only to components which do not vanish. This formula will be tested in our simulation later on. When sample size is very small, this formula may result in underestimate of the true standard errors. This is a common phenomenon. Hence, for constructing confidence intervals for regression coefficients, the corresponding critical values may need some modifications when the sample size is small. See Kauermann and Carroll (2001) for detailed arguments.

4.2 Connection between local quadratic approximation and MM algorithm

The EM algorithm has been successfully and widely used for finding solution of a complicated likelihood function, namely, $L(\boldsymbol{\beta})$, particularly in missing data perspectives. The principle of the EM algorithm is to construct a surrogate function, denoted by $\tilde{Q}(\boldsymbol{\beta}|\boldsymbol{\beta}^0)$ in the E-step, where $\boldsymbol{\beta}^0$ stands for the current value of $\boldsymbol{\beta}$ at each step, then to maximize $\tilde{Q}(\boldsymbol{\beta}|\boldsymbol{\beta}^0)$ in the M-step instead of maximizing the complicated likelihood function $L(\boldsymbol{\beta})$. The surrogate function $\tilde{Q}(\boldsymbol{\beta}|\boldsymbol{\beta}^0)$ satisfies the following two critical conditions:

$$L(\boldsymbol{\beta}^0) = \tilde{Q}(\boldsymbol{\beta}^0|\boldsymbol{\beta}^0), \tag{4.4}$$

and

$$L(\boldsymbol{\beta}) \geq \tilde{Q}(\boldsymbol{\beta}|\boldsymbol{\beta}^0) \tag{4.5}$$

to guarantee that the resulting solution converges correctly. These two conditions implies that the EM algorithm is an ascent algorithm. The EM algorithm constructs the surrogate function $\tilde{Q}(\boldsymbol{\beta}|\boldsymbol{\beta}^0)$ by taking expectation on the observed data in the E-step. The condition (4.5) is satisfied due to the entropy inequality. Following the principle of the EM algorithm, the minorize-maximize algorithm constructs a surrogate function $\tilde{Q}(\boldsymbol{\beta}|\boldsymbol{\beta}^0)$ using other techniques rather than taking the expectation on the observed data. Provided that the conditions (4.4) and (4.5) are satisfied, under some mild conditions, the minorize-maximize algorithm shares the same local and global convergence properties as those of the EM algorithm (Lange, Hunter and Yang, 2000).

To accelerate the EM algorithm, we may update the value of $\boldsymbol{\beta}^0$ at each step rather than carry out the M-step exactly. Lange (1995) has shown that with this acceleration, the EM algorithm converges faster and preserves the same local and global convergence as those of the EM algorithm.

In summary, the conditions (4.4) and (4.5) are critical for both the EM and the MM algorithms. Now we establish a connection between the local quadratic approximation and the MM algorithm through these two conditions. Define

$$\tilde{Q}(\boldsymbol{\beta}|\boldsymbol{\beta}^0) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \mathbf{x}^T\boldsymbol{\beta})^2 + \sum_{j=1}^{d} q_\lambda(\beta_j|\beta_j^0),$$

where

$$q_\lambda(\beta|\beta^0) = p_\lambda(|\beta^0|) + \frac{1}{2}\left\{p'_\lambda(|\beta^0|)/|\beta^0|\right\}(\beta^2 - \{\beta^0\}^2).$$

for $\beta^0 \neq 0$.

Note that our task is to minimize the PLS function

$$Q(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \sum_{j=1}^{d} p_\lambda(|\beta_j|).$$

Thus, provided that we can verify the following two conditions:

$$Q(\boldsymbol{\beta}^0) = \tilde{Q}(\boldsymbol{\beta}^0|\boldsymbol{\beta}^0), \tag{4.6}$$

and

$$Q(\boldsymbol{\beta}) \leq \tilde{Q}(\boldsymbol{\beta}|\boldsymbol{\beta}^0), \tag{4.7}$$

we may analyze the local and global convergence of the iterative ridge regression with the local quadratic approximation by using techniques related to the EM algorithm.

Theorem 4.1 *Suppose that $p'_\lambda(\cdot)$ is non-increasing over $(0, +\infty)$. Then (4.6) and (4.7) hold.*

Proof. It is obvious that (4.6) holds. Note that

$$\tilde{Q}(\boldsymbol{\beta}|\boldsymbol{\beta}^0) - Q(\boldsymbol{\beta}) = \sum_{j=1}^{d} q_\lambda(\beta_j|\beta_j^0) - p_\lambda(|\beta_j|).$$

Thus, it is sufficient to show that for any β

$$\delta(\beta) \equiv q_\lambda(\beta|\beta^0) - p_\lambda(\beta) \geq 0. \tag{4.8}$$

Note that for $\beta \neq 0$

$$\delta'(\beta) = \frac{p'_\lambda(|\beta^0|)}{|\beta^0|}\beta - p'_\lambda(|\beta|)\text{sgn}(\beta) = \left\{\frac{p'_\lambda(|\beta^0|)}{|\beta^0|} - \frac{p'_\lambda(|\beta|)}{|\beta|}\right\}\beta.$$

Since $p'_\lambda(\cdot)$ is non-increasing, so when $|\beta^0| < |\beta|$, then

$$\frac{p'_\lambda(|\beta^0|)}{|\beta^0|} \geq \frac{p'_\lambda(|\beta|)}{|\beta|}.$$

Thus, $\delta'(\beta)$ either equals to 0 or has the same sign as that of β. On the other hand, if $|\beta^0| > |\beta|$, then $\delta'(\beta)$ either equals to 0 or has different sign from that of β. Therefore, (4.8) holds. This completes the proof of the theorem.

The condition is Theorem 4.1 is very mild. All of the SCAD, HARD, the L_p $(0 < p \leq 1)$ and the modified L_p $(0 < p \leq 1)$ penalties satisfy this condition.

Remark: In the earlier version of Fan and Li (2001), a naive quadratic approximation for the penalty function:

$$p_\lambda(\beta) \approx \{p_\lambda(\beta^0)/|\beta_0|^2\}\beta^2 \tag{4.9}$$

has been considered. The resulting estimator of this approximation does not satisfy the penalized least squares equation (4.2). Furthermore, this approximation does not satisfy condition (4.7). See the bottom panel of Figure 3 for more insights. Hence, we recommend the use of the approximation (4.1).

5 Local Asymptotic properties

To understand finite sample performance of the nonconvex PLS, it is of interest to investigate local asymptotic properties of the resulting estimator. To this end, suppose that we have a triangular array of observations:

$$y_{ni} = \mathbf{x}_{ni}^T\boldsymbol{\beta} + \varepsilon_{ni}, \quad \text{for} \quad i = 1, \cdots, n, \tag{5.1}$$

where $E(\varepsilon_{ni}|\mathbf{x}_{ni}) = 0$ and $\mathrm{Var}(\varepsilon_{ni}|\mathbf{x}_{ni}) = \sigma^2$. Following Knight and Fu (2000) closely, in this section, we allow the true value of $\boldsymbol{\beta}$ depending on n, denoted by $\boldsymbol{\beta}_n$, and let $\boldsymbol{\beta}_n = \boldsymbol{\beta}_0 + c_n\mathbf{t}$, where $c_n > 0$ and $c_n \to 0$, $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$ is the same as that in Section 3, i.e., $\boldsymbol{\beta}_{20} = \mathbf{0}$ and all components of $\boldsymbol{\beta}_{10}$, consisting of the first s components of $\boldsymbol{\beta}_0$, are not equal to 0., and $\mathbf{t} = (\mathbf{t}_1^T, \mathbf{t}_2^T)^T$ is a fixed finite vector. Moreover, $\mathbf{t}_1$ and $\boldsymbol{\beta}_{10}$ have the same dimension. Let $\boldsymbol{\beta}_{1n} = \boldsymbol{\beta}_{10} + c_n\mathbf{t}_1$ and $\boldsymbol{\beta}_0^* = (\boldsymbol{\beta}_{1n}^T, \boldsymbol{\beta}_{20}^T)^T$.

We show in the following theorem that rates of convergence of $\widehat{\boldsymbol{\beta}}$ is $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0^*\| = O_P(n^{-1/2} + a_n + c_n)$. In Theorem 5.2, we investigate its limiting distribution.

Theorem 5.1 *Suppose that the observations $(\mathbf{x}_{ni}, y_{ni})$, $i = 1, \cdots, n$, are independent and identically distributed from the model (5.1), and assume that the matrix $E(\mathbf{x}_{n1}\mathbf{x}_{n1}^T)$, $\mathbf{V}$ say, is finite and positive definite, and the random error has mean zero and finite positive variance σ^2. If both a_n and b_n, defined in (3.1), tend to 0 as $n \to \infty$, then with probability tending to one, there exists a local minimizer $\widehat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0^*\| = O_P(n^{-1/2} + a_n + c_n)$.*

Proof. Following the proof of Theorem 3.1, it follows that with probability tending to one, there exists a local minimizer $\widehat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_n\| = O_P(n^{-1/2} + a_n)$. Note that $\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_0^*\| = c_n\|\mathbf{t}_2\|$. Since $\mathbf{t}_2$ is a finite fixed vector, the proof is completed by the triangular inequality.

From Theorem 5.1, if $c_n = O(n^{-1/2} + a_n)$, then the rate of convergence of the resulting estimator is $O_P(n^{-1/2} + a_n)$. Choosing a proper tuning parameter λ_n, the resulting estimator possesses root n consistency.

Lemma 5.1 *Under conditions of Theorem 5.1, if $c_n = O_P(n^{-1/2})$, and condition (3.4) holds, then with probability tending to one, for any given $\boldsymbol{\beta}_1$ satisfying $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1n}\| = O_P(n^{-1/2})$ and any constant C*

$$Q\left\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \mathbf{0}\end{pmatrix}\right\} = \min_{\|\boldsymbol{\beta}_2\| \le Cn^{-1/2}} Q\left\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}\right\}.$$

Proof. For $j = s+1, \cdots, d$, and when $\beta_j \neq 0$,

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} &= -\frac{1}{n}\mathbf{x}_{(nj)}^T(\mathbf{y}_n - \mathbf{X}_n\boldsymbol{\beta}) + p'_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j) \\ &= -\frac{1}{n}\mathbf{x}_{(nj)}^T(\mathbf{y}_n - \mathbf{X}_n\boldsymbol{\beta}_n) - \frac{1}{n}\mathbf{x}_{(nj)}^T\mathbf{X}_n(\boldsymbol{\beta}_n - \beta_0^*) + \frac{1}{n}\mathbf{x}_{(nj)}^T\mathbf{X}_n(\boldsymbol{\beta} - \boldsymbol{\beta}_0^*) + p'_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j), \end{aligned}$$

where $\mathbf{X}_n$ is the corresponding design matrix and $\mathbf{x}_{(nj)}$ is the j-th column of $\mathbf{X}_n$. Taking $\boldsymbol{\beta}$ be in a neighborhood of $\boldsymbol{\beta}_0^*$ such that $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1n}\| = O_P(n^{-1/2})$ and $\|\boldsymbol{\beta}_2\| \leq Cn^{-1/2}$. Thus $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0^*\| = O_P(n^{-1/2})$, so $\frac{1}{n}\mathbf{x}_{(nj)}^T\mathbf{X}_n(\boldsymbol{\beta} - \boldsymbol{\beta}_0^*) = O_P(n^{-1/2})$ by the SLLN. By the CLT, $\frac{1}{n}\mathbf{x}_{(nj)}^T(\mathbf{y}_n - \mathbf{X}_n\boldsymbol{\beta}_n) = O_P(n^{-1/2})$, and by the SLLN, $\frac{1}{n}\mathbf{x}_{(nj)}^T\mathbf{X}_n(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0^*) = O_P(c_n) = O_P(n^{-1/2})$. Thus

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

Since $\liminf_{n\to\infty} \liminf_{\beta\to 0^+} \sqrt{n}p'_{\lambda_n}(\beta) = +\infty$, the sign of the derivative is completely determined by that of β_j. This completes the proof of the lemma.

Theorem 5.2 *Under the conditions of Lemma 5.1, with probability tending to one, the root n consistent local minimizer $\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{pmatrix}$ in Theorem 5.1 must satisfy:*

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

(ii) (Asymptotic normality)

$$\sqrt{n}(\mathbf{V}_{11} + \Sigma_n)\left\{\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{1n} + (\mathbf{V}_{11} + \Sigma_n)^{-1}(\mathbf{b}(n) - c_n\mathbf{V}_{12}\mathbf{t}_2)\right\} \to N\left(\mathbf{0}, \sigma^2\mathbf{V}_{11}\right)$$

in distribution, where $\mathbf{V}_{11}$ consists of the first s rows and columns of $\mathbf{V}$, $\mathbf{V}_{12}$ consists of the first s rows and the last $d-s$ columns of $\mathbf{V}$, and

$$\begin{aligned} \Sigma_n &= diag\left\{p''_{\lambda_n}(|\beta_{1n}|), \cdots, p''_{\lambda_n}(|\beta_{sn}|)\right\}, \\ \mathbf{b}(n) &= \left(p'_{\lambda_n}(|\beta_{1n}|)sgn(\beta_{1n}), \cdots, p'_{\lambda_n}(|\beta_{sn}|)sgn(\beta_{sn})\right)^T. \end{aligned}$$

Proof. Part (i) follows by Lemma 5.1. Now we prove Part (ii). It can be easily shown that there exists a $\widehat{\boldsymbol{\beta}}_1$ in Theorem 5.1 that is a root n consistent local minimizer of $Q\{(\boldsymbol{\beta}_1^T, \mathbf{0})^T\}$, regarded as a function of $\boldsymbol{\beta}_1$, and satisfying the following equations:

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

Note that $\widehat{\boldsymbol{\beta}}_1$ is a consistent estimator, and denote $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1, \mathbf{0}^T)^T$,

$$\begin{aligned}
\frac{\partial Q(\widehat{\boldsymbol{\beta}})}{\partial \beta_j} &= -\frac{1}{n}\mathbf{x}_{(nj)}^T(\mathbf{y}_n - \mathbf{X}_n\boldsymbol{\beta}_n) - \frac{1}{n}\mathbf{x}_{(nj)}^T\mathbf{X}_n(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0^*) + \frac{1}{n}\mathbf{x}_{(nj)}^T\mathbf{X}_n(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0^*) + p'_{\lambda_n}(|\widehat{\beta}_j|)\mathrm{sgn}(\widehat{\beta}_j) \\
&= -\frac{1}{n}\mathbf{x}_{(nj)}^T(\mathbf{y}_n - \mathbf{X}_n\boldsymbol{\beta}_n) - \frac{1}{n}\mathbf{x}_{(nj)}^T\mathbf{X}_{(n2)}c_n\mathbf{t}_2 + \frac{1}{n}\mathbf{x}_{(nj)}^T\mathbf{X}_{(n1)}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{1n}) + p'_{\lambda_n}(|\widehat{\beta}_j|)\mathrm{sgn}(\widehat{\beta}_j) \\
&= -\frac{1}{n}\mathbf{x}_{(nj)}^T(\mathbf{y}_n - \mathbf{X}_n\boldsymbol{\beta}_n) - c_n\left\{\frac{1}{n}\mathbf{x}_{(nj)}^T\mathbf{X}_{(n1)}\right\}\mathbf{t}_2 + \mathbf{V}_{(j)}^T(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{1n})(1 + o_P(1)) \\
&\quad + \left(p'_{\lambda_n}(|\beta_{jn}|)\mathrm{sgn}(\beta_{jn}) + \{p''_{\lambda_n}(|\beta_{jn}|) + o_P(1)\}(\widehat{\beta}_j - \beta_{jn})\right),
\end{aligned}$$

where $\mathbf{X}_{(n1)}$ and $\mathbf{X}_{(n2)}$ consist of the first s and the last $d-s$ columns of $\mathbf{X}_n$, respectively, and $\mathbf{V}_{(j)}$ is the jth-column of V_{11}. It follows by Slutsky's Theorem and the CLT that

$$\sqrt{n}(\mathbf{V}_{11} + \Sigma_n)\left\{\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{1n} + (\mathbf{V}_{11} + \Sigma_n)^{-1}(\mathbf{b}(n) - c_n\mathbf{V}_{12}\mathbf{t}_2)\right\} \rightarrow N\left(\mathbf{0}, \sigma^2\mathbf{V}_{11}\right)$$

in distribution.

In view of Theorems 5.1 and 5.2, under the regularity conditions, $\sqrt{n}\{\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{1n}\}$ has an asymptotic normal distribution with mean $\sqrt{n}\{\mathbf{b}(n) - c_n\mathbf{V}_{12}\mathbf{t}_2\}$ and covariance matrix

$$(\mathbf{V}_{11} + \Sigma_n)^{-1}\mathbf{V}_{11}(\mathbf{V}_{11} + \Sigma_n)^{-1}.$$

If $a_n = o_P(n^{-1/2})$, and $c_n = o_P(n^{-1/2})$, then the bias $\sqrt{n}(\mathbf{b}(n) - c_n\mathbf{V}_{12}\mathbf{t}_2)$ tends to 0. Thus, if all small coefficients are of the order $o_P(n^{-1/2})$, then the bias of the resulting PLS estimator with the HARD, the SCAD, the L_p $(0 < p < 1)$ or the modified L_p $(0 < p \leq 1)$ is small. On the other hand, if some small coefficients are of the order $O_P(n^{-1/2})$, then estimation bias for large coefficient is not negligible. This is similar to the thresholding rules, including the hard and soft thresholding rules with the universal thresholding parameter in the setting of wavelets. Therefore, data-driven methods, such as CV and GCV, are recommended for choosing λ_n in practice.

6 Variable Selection for Nonlinear Regression Models

The ideas of PLS can be extended naturally to the context of nonlinear regression models. Consider the nonlinear regression model

$$y = f(\mathbf{x}^T\boldsymbol{\beta}) + \varepsilon, \tag{6.1}$$

where $E(\varepsilon|\mathbf{x}) = 0$ and $\text{Var}(\varepsilon|\mathbf{x}) = \sigma^2 v(\mathbf{x})$ with a finite and positive σ^2 and a known function $v(\mathbf{x})$. Denote by $\ell(\boldsymbol{\beta})$ the sum of weighted squares,

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i - f(\mathbf{x}_i^T \boldsymbol{\beta})\}^2 / v(\mathbf{x}_i).$$

Minimizing

$$Q(\boldsymbol{\beta}) = \frac{1}{2n} \ell(\boldsymbol{\beta}) + \sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|)$$

results in a nonlinear PLS estimators. With proper choices of the penalty function and the regularization parameter, we may achieve the purpose of selecting significant variables. Note that

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2 \sum_{i=1}^{n} \{y_i - f(\mathbf{x}_i^T \boldsymbol{\beta})\} f'(\mathbf{x}_i^T \boldsymbol{\beta}) \mathbf{x}_i / v(\mathbf{x}_i),$$

and

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = 2 \sum_{i=1}^{n} \{f'(\mathbf{x}_i^T \boldsymbol{\beta})\}^2 \mathbf{x}_i \mathbf{x}_i^T / v(\mathbf{x}_i) - 2 \sum_{i=1}^{n} \{y_i - f(\mathbf{x}_i^T \boldsymbol{\beta})\} f''(\mathbf{x}_i^T \boldsymbol{\beta}) \mathbf{x}_i \mathbf{x}_i^T / v(\mathbf{x}_i).$$

To establish rates of convergence and the oracle property for the resulting estimator, we need the following regularity conditions. Let $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$ be the true value of $\boldsymbol{\beta}$, where $\boldsymbol{\beta}_{20} = 0$ and all components of $\boldsymbol{\beta}_{10}$ are not equal to zero. Denote a neighborhood of $\boldsymbol{\beta}_0$ by

$$B(\delta) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le \delta\}.$$

Conditions:

Assume that the second order derivative of $f(\cdot)$ exists and finite in $B(\delta)$ for some $\delta > 0$.

(a) $Ef'(\mathbf{x}^T \boldsymbol{\beta}_0) \mathbf{x} \mathbf{x}^T / v(\mathbf{x})$, denoted by $\mathbf{V}_f$, is finite and positive definite.

(b) $E\{f''(\mathbf{x}^T \boldsymbol{\beta}_0)\}^2 \{\mathbf{x}^T \mathbf{x}\}^2 / v(\mathbf{x})$ is finite.

(c) As $\delta \downarrow 0$,

$$E \sup_{\boldsymbol{\beta} \in B(\delta)} |\{f'(\mathbf{x}^T \boldsymbol{\beta})\}^2 - \{f'(\mathbf{x}^T \boldsymbol{\beta}_0)\}^2| \mathbf{x}^T \mathbf{x} / v(\mathbf{x}) \to 0,$$
$$E \sup_{\boldsymbol{\beta} \in B(\delta)} |f(\mathbf{x}^T \boldsymbol{\beta}) f''(\mathbf{x}^T \boldsymbol{\beta}) - f(\mathbf{x}^T \boldsymbol{\beta}_0) f''(\mathbf{x}^T \boldsymbol{\beta}_0)| |\mathbf{x}^T \mathbf{x} / v(\mathbf{x}) \to 0,$$
$$E_{\boldsymbol{\beta}_0} \sup_{\boldsymbol{\beta} \in B(\delta)} |y| |f''(\mathbf{x}^T \boldsymbol{\beta}) - f''(\mathbf{x}^T \boldsymbol{\beta}_0)| \mathbf{x}^T \mathbf{x} / v(\mathbf{x}) \to 0.$$

Theorem 6.1 *Suppose that the observations $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$ are independent and identically distributed from the model (6.1), and assume that the regularity conditions (a)—(c) hold. If both a_n and b_n, defined in (3.1), tend to 0 as $n \to \infty$, then with probability tending to 1, there exists a local minimizer $\widehat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that $\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_0\| = O_P(n^{-1/2} + a_n)$.*

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

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

Using the Taylor expansion, it follows that

$$\begin{aligned}\frac{1}{n}\{\ell(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - \ell(\boldsymbol{\beta}_0)\} &= -\frac{2\alpha_n}{n}\sum_{i=1}^n \{y_i - f(\mathbf{x}_i^T \boldsymbol{\beta}_0)\} f'(\mathbf{x}_i^T \boldsymbol{\beta}_0)\mathbf{x}_i^T \mathbf{u}/v(\mathbf{x}) \\ &\quad + \frac{\alpha_n^2}{n}\sum_{i=1}^n \left[\{f'(\mathbf{x}_i^T \xi)\}^2 - \{y_i - f'(\mathbf{x}_i^T \xi)\} f''(\mathbf{x}_i^T \xi)\right](\mathbf{x}_i^T \mathbf{u})^2/v(\mathbf{x}),\end{aligned}$$

where ξ lies between $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}$. Using Condition (C), it can be shown that

$$\begin{aligned}&\frac{\alpha_n^2}{n}\sum_{i=1}^n \left[\{f'(\mathbf{x}_i^T \xi)\}^2 - \{y_i - f'(\mathbf{x}_i^T \xi)\} f''(\mathbf{x}_i^T \xi)\right](\mathbf{x}_i^T \mathbf{u})^2/v(\mathbf{x}) \\ = \;& \frac{\alpha_n^2}{n}\sum_{i=1}^n \left[\{f'(\mathbf{x}_i^T \boldsymbol{\beta}_0)\}^2 - \{y_i - f'(\mathbf{x}_i^T \boldsymbol{\beta}_0)\} f''(\mathbf{x}_i^T \boldsymbol{\beta}_0)\right](\mathbf{x}_i^T \mathbf{u})^2/v(\mathbf{x})\{1 + o_P(1)\}.\end{aligned}$$

By Condition (b) and the Chebyshev inequality, it follows that

$$\frac{1}{n}\sum_{i=1}^n \{y_i - f'(\mathbf{x}_i^T \boldsymbol{\beta}_0)\} f''(\mathbf{x}_i^T \boldsymbol{\beta}_0)(\mathbf{x}_i^T \mathbf{u})^2/v(\mathbf{x}_i) = o_P(1).$$

Thus, by Condition (a), we have

$$\begin{aligned}&\frac{1}{n}\{\ell(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - \ell(\boldsymbol{\beta}_0)\} \\ = \;& -\frac{2\alpha_n}{n}\sum_{i=1}^n \left\{y_i - f(\mathbf{x}_i^T \boldsymbol{\beta}_0)\right\} f'(\mathbf{x}_i^T \boldsymbol{\beta}_0)\mathbf{x}_i^T \mathbf{u}/v(\mathbf{x}_i) + \frac{\alpha_n^2}{n}\mathbf{u}^T \mathbf{V}_f \mathbf{u}\{1 + o_P(1)\}. \qquad (6.2)\end{aligned}$$

Denote

$$D_n(\mathbf{u}) = Q(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - Q(\boldsymbol{\beta}_0).$$

Then using $p_{\lambda_n}(0) = 0$, it follows that

$$\begin{aligned}D_n(\mathbf{u}) \;\geq\; & -\frac{\alpha_n}{n}\sum_{i=1}^n \{y_i - f(\mathbf{x}_i^T \boldsymbol{\beta}_0)\} f'(\mathbf{x}_i^T \boldsymbol{\beta}_0)\mathbf{x}_i^T \mathbf{u}/v(\mathbf{x}_i) + \frac{\alpha_n^2}{2n}\mathbf{u}^T \mathbf{V}_f \mathbf{u}\{1 + o_P(1)\} \\ & + \sum_{j=1}^s \{p_{\lambda_n}(|\boldsymbol{\beta}_{j0} + \alpha_n b_j|) - p_{\lambda_n}(|\beta_{j0}|)\}.\end{aligned}$$

We complete the proof of theorem by employing techniques used in the proof of Theorem 3.1,

Theorem 6.2 *Under the conditions of Theorem 6.1, if the condition (3.4) holds, then with probability tending to one, the root n consistent local minimizer* $\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{pmatrix}$ *in Theorem 6.1 must satisfy:*

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

(ii)

$$\sqrt{n}(\mathbf{V}_{f11} + \Sigma) \left\{ \widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\mathbf{V}_{f11} + \Sigma)^{-1}\mathbf{b} \right\} \to N\left(\mathbf{0}, \sigma^2 \mathbf{V}_{f11}\right)$$

in distribution, where $\mathbf{V}_{11}$ *consists of the first s rows and columns of* $\mathbf{V}$*, and* Σ *and* $\mathbf{b}$ *are given by (3.5) and (3.6).*

Proof. To show Part (i), consider $\boldsymbol{\beta}$ being a neighborhood of $\boldsymbol{\beta}_0$ such that $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_P(n^{-1/2})$ and $\|\boldsymbol{\beta}_2\| \leq Cn^{-1/2}$ for some constant C. This implies that $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2})$. For $j = s+1, \cdots, d$ and $\beta_j \neq 0$, by similar arguments in the proof of Theorems 6.1 and 3.2, it follows that

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

Due to condition (3.4), the sign of the derivative is completely determined by that of β_j. This implies that for any given $\boldsymbol{\beta}_1$ satisfying $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_P(n^{-1/2})$ and any constant C

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

Thus, Part (i) follows.

Using the arguments for (6.2), and note that $\widehat{\boldsymbol{\beta}}$ is root n consistent, Part (ii) can be shown by using techniques related to the proof of Theorem 3.2.

7 Simulations

The objective of this section is to empirically compare the performance of the proposed PLS estimators via Monte Carlo simulations. In this section, we also test the accuracy of the standard error formula.

7.1 Prediction and model error

Suppose that the data $(\mathbf{x}_i, y_i)$ is a random sample from their population $(\mathbf{x}, y)$. If $\widehat{\mu}(\mathbf{x})$ is a prediction procedure constructed based on 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 $(\mathbf{x}, y)$. We can decompose the prediction error 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 due to stochastic errors, while the second one is due to lack of fit to an underlying model. Thus, the second component is called *Model Error*, denoted by $\mathrm{ME}(\widehat{\mu})$. For the linear regression model (1.1), $\mathrm{ME}(\widehat{\mu}) = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T E(\mathbf{x}\mathbf{x}^T)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. We compare performances of the nonconvex PLS estimators in terms of Median of Relative Model Error (MRME), where the relative model error is defined as the ratio of model error of an underlying model to the model error of the full model, which corresponds to the least squares estimator for the full model. Thus, the size of MRME reflects performance of different variable selection procedures.

7.2 Selection of the tuning parameter

The tuning parameter λ in the PLS can be selected via data-driven approaches, such as CV and GCV. Performances of CV and GCV are similar (see Li, 1985 and Wahba, 1985). In our simulations, we only employ the GCV to choose the value of λ because the GCV is less computational. For linear regression model, we update the solution by

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

Thus, the fitted value of $\widehat{\mathbf{y}}$ of $\mathbf{y}$ is $\mathbf{X}\{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda\}^{-1}\mathbf{X}^T\mathbf{y}$. Define the number of effective parameters in the penalized least-squares fit as

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

Thus, the GCV statistic is

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

and

$$\widehat{\lambda} = \mathrm{argmin}_{\lambda} GCV(\lambda).$$

Similarly, the corresponding generalized cross-validation statistic can be defined for the penalized nonlinear least squares estimator.

7.3 Simulation study

This section presents some numerical comparisons of the nonconvex PLS estimators with the SCAD, the HARD, the L_1 and the modified $L_{0.5}$ penalties. We chose $L_{0.5}$ penalty because we expect that its performance is a kind of tradeoff between the entropy (or L_0) penalty and L_1 penalty, corresponding to LASSO. We also compare the performance of the nonconvex penalized least squares with the oracle estimator. The oracle estimate was computed by fitting data to the true model. All simulations were conducted using MATLAB codes.

Example 1. (Linear regression) In this example, we simulated 1000 data sets consisting of 150 observations from the model

$$y = \mathbf{x}^T \boldsymbol{\beta} + \sigma\varepsilon, \tag{7.1}$$

where $\boldsymbol{\beta} = (0,0,0.1,0,0,0.15,0,0,0.2,0,0,0.3,0,0,0.6,0,0,1,0,0,1.5,0,0,3,0,0,6,0,0,9)^T$, which is a 30-dimensional vector only containing 10 nonzero components, and the components of $\mathbf{x}$ and ε are standard normal, where the correlation coefficient between x_i and x_j is $\rho^{|i-j|}$ with $\rho = 0.5$. We first took $\sigma = 1$, which leads to that the standard error for each component of the least squares estimator is around 0.1. In this model, the true value of $\boldsymbol{\beta}$ contains some small coefficients, such as 0.1, 0.15 and some large coefficients, for example, 6, 9. This allows us to examine the finite sample behaviors of the nonconvex PLS in the presence of some small coefficients. This model is challenging for all existing approaches. Since the best subset variable selection is impractical when the dimension of $\mathbf{x}$ is 30 or more, we employ stepwise regression. In the stepwise regression, we set the threshold level equal to 0.05, which implies that a predictor will be deleted from the current model if its estimate is less than twice of its standard error. To understand the effect of noise levels, we also simulated the case in which $\sigma = 0.5$.

Compared with the model in Example 4.1 of Fan and Li (2001), here we focus on a much larger model with some small coefficients. The MRME over 1000 simulation data sets are summarized in Table 1. The average of zero coefficients is also reported in Table 1, in which the column labelled

"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, we can see that the SCAD outperforms the other PLS estimators in terms of MRME. The performance of the PLS with the modified $L_{0.5}$ penalty lies between the SCAD and the LASSO in terms of MRME and reduction of model complexity when noise level is small, corresponding to $\sigma = 0.5$. From Table 1, the noise level does not affect the performance of the SCAD and the HARD. Reducing $\sigma = 1$ to 0.5, all variable selection procedures have better performance in terms of reduction of model complexity, however, the MRME of the LASSO becomes larger. This is undesirable and may be caused by unnecessary modeling bias created by the L_1 penalty. The HARD performs quite poor in this example because the discontinuity of the resulting estimator creates extra variation.

In this example, we also have tested the accuracy of the standard error formula. The results are similar to Example 4.1 in Fan and Li (2001). We opt not to present them to save space. Readers may refer the subsequent example for nonlinear regression models.

Table 1: Simulation Results for Example 1

Method	MRME(%)	Aver. no. of 0 Coeff.		MRME(%)	Aver. no. of 0 Coeff.	
		Correct	Incorrect		Correct	Incorrect
	$\sigma = 1$			$\sigma = 0.5$		
Stepwise	58.95	18.82	2.04	53.10	18.95	0.57
SCAD	51.62	17.60	1.56	48.86	17.84	0.38
HARD	76.40	16.16	1.54	71.97	16.43	0.41
LASSO	60.29	13.72	1.04	62.72	13.52	0.21
$L_{0.5}$	61.19	14.32	1.13	53.52	16.17	0.29
Oracle	27.89	20	0	27.90	20	0

Example 4.2 (Nonlinear regression) In this example, we simulated 1000 data sets each consisting of 30 observations from the model

$$y = f(\mathbf{x}^T\boldsymbol{\beta}) + \varepsilon,$$

where $f(t) = \exp(t)$, $\boldsymbol{\beta} = (1.2, 0.6, 0, 0, 0.8, 0, 0, 0)^T$, and the components of $\mathbf{x}$ and ε are standard normal. The correlation between x_i and x_j is $0.5^{|i-j|}$. The model errors were obtained by 1000 Monte Carlo simulations. Tables 2 shows the simulation results. In this example, the SCAD outperforms the other three penalized least squares estimators, and performs as well as the oracle estimator. The performances of the other three PLS estimators are almost the same in terms of

MRME, although the PLS with the modified $L_{0.5}$ penalty reduce model complexity a little more aggressive.

Table 2: Simulation Results for Example 2

Method	MRME(%)	Aver. no. of 0 Coeff.	
		Correct	Incorrect
SCAD	9.66	4.77	0.08
HARD	21.38	4.17	0.03
LASSO	22.05	4.62	0.13
$L_{0.5}$	24.95	4.91	0.30
Oracle	6.01	5	0

Table 3: Standard Deviations of Estimators in Example 2

	$\widehat{\beta}_1$		$\widehat{\beta}_2$		$\widehat{\beta}_5$	
Method	SD	$SE_m(SE_{mad})$	SD	$SE_m(SE_{mad})$	SD	$SE_m(SE_{mad})$
SCAD	0.0509	0.0420 (0.0266)	0.0532	0.0477 (0.0271)	0.0431	0.0410 (0.0254)
HARD	0.0550	0.0463 (0.0308)	0.0600	0.0543 (0.0336)	0.0580	0.0471 (0.0303)
LASSO	0.0577	0.0417 (0.0257)	0.0652	0.0477 (0.0256)	0.0639	0.0419 (0.0261)
$L_{0.5}$	0.0731	0.0403 (0.0254)	0.0660	0.0438 (0.0231)	0.0593	0.0405 (0.0253)
Oracle	0.0403	0.0427 (0.0289)	0.0472	0.0496 (0.0318)	0.0368	0.0410 (0.0266)

Now we test the accuracy of the standard error formula constructed via using sandwich formula. The median absolute deviation divided by 0.6745, denoted by SD in Table 3, of 1000 estimated coefficients in the 1000 simulations can be regarded as the true standard error. The median of the 1000 estimated standard error, denoted by SE_m, and the median absolute deviation error of the 1000 estimated standard errors divided by 0.6745, denoted by SE_{mad}, gauge the overall performance of the standard error formula. Table 3 displays the results for three nonzero coefficients. From Table 3, we can see that the sandwich formula works well, although the true standard errors are somewhat underestimated. This is a common phenomenon when sample size is small. See Kauermann and Carroll (2001) for detailed arguments.

Acknowledgements: Parts of Section 3 were based on the author's Ph.D. dissertation at University of North Carolina at Chapel Hill. The author is very grateful to his advisors Professors Jianqing Fan and James S. Marron for their inspiring guidance and enthusiastic support. The

author also thanks Professor David R. Hunter for motivating discussions on the MM algorithm.

References

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.*, **22**, 203–217.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, **19**, 716–723.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction, *Technometrics*, **16**, 125-127.

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

Antoniadis, A. and Fan, J. (2001). Regularization of wavelets approximations (with discussions), *Journal of American Statistical Association*, **96**, 939-967.

Bickel, P. J. (1983). Minimax estimation of a normal mean subject to ding well at a point, In *Recent Advances in Statistics* (M. H. Rizvi, J. S. Rustagi, and D. Siegmund, eds), 511-528, Academic Press, New York.

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.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Ser. B.*, **39**, 1-38

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

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

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of American Statistical Association.* In press.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression, *Ann. Statist.*, **22**, 1947-1975.

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.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of autoregression, *Journal of the Royal Statistical Society, Ser. B*, **41**, 190-195.

Kauermann, G. and Carroll, R. J. (2001). A Note on the Efficiency of Sandwich Covariance Matrix Estimation, *Journal of American Statistical Association.* In press.

Knight, K. and Fu, W. (2000). Asymptotic for lasso-type estimators, *Annals of Statistics*, **28**, 1356-1378.

Lange, K. (1995). A gradient algorithm locally equivalent to the EM algorithm, *Journal of Royal Statistical Society*, **57**, 425-437.

Lange, K. Hunter, D. R. and Yang, I. (2000). Optimization transfer using surrogate objective functions, *Journal of Computational and Graphical Statistics*, **9**, 1-59.

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

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

Meng, X. L. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework, *Biometrika*, **80**, 267-278.

Meng, X. L. and Van Dyk, D. A. (1997). The EM algorithm — An old folk song sung to a fast new tune (with discussion), *Journal of the Royal Statistical Society, Series B*, **59**, 511-567.

Miller, A. J. (1990). *Subset Selection in Regression.* Chapman and Hall, London.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**, 758-765.

Pötscher, B. M. (1989). Model selection under nonstationary: autoregressive model and stochastic linear regression models, *Ann. Statist.*, **17**, 1257-1274.

Rao, C. R. and Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem, *Biometrika*, **76**, 369-374.

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

Shao, J. (1997). An asymptotic theory for linear model selection, *Statistica Sinica*, **7**, 221-264.

Shen, X. and Ye, J. (2001). Adaptive model selection. *Journal of American Statistical Association*, to appear.

Shibata, R. (1984). Approximation efficiency of a selection procedure for the number of regression variables. *Biometrika*, **71**, 43-49.

Theil, H. (1961). *Economic Forecasts and Policy.* North-Holland, Amsterdam.

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.

Tibshirani, R. and Knight, K. (1999). The covariance inflation criterion for adaptive model selection. *J. Royal Statist. Soc. B*, **61**, 529–546.

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

Wu, J. C. F. (1983). On the convergence properties of the EM algorithm, *The Annals of Statistics*, **11**, 95-103.

Zheng, X. and Loh, W.-Y. (1995). Consistent variable selection in linear models, *Journal of American Statistical Association*, **90**, 151-156.

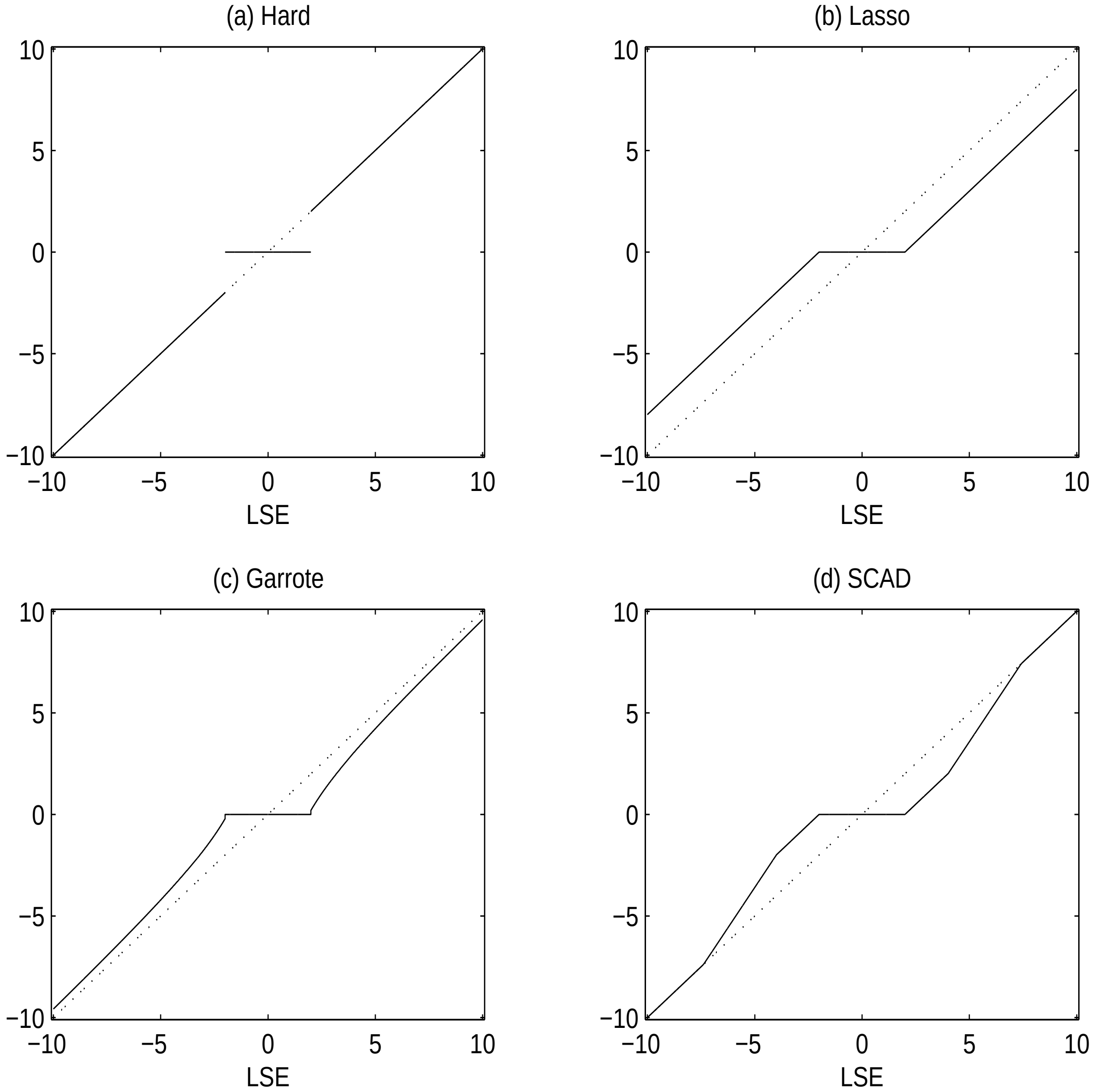

Figure 1: *Thresholding Rules. (a), (b), (c) and (d) are the hard, soft (LASSO), nonnegative garrote and SCAD thresholding rules, respectively. They are the solutions of the penalized least squares when design matrix is orthonormal.*

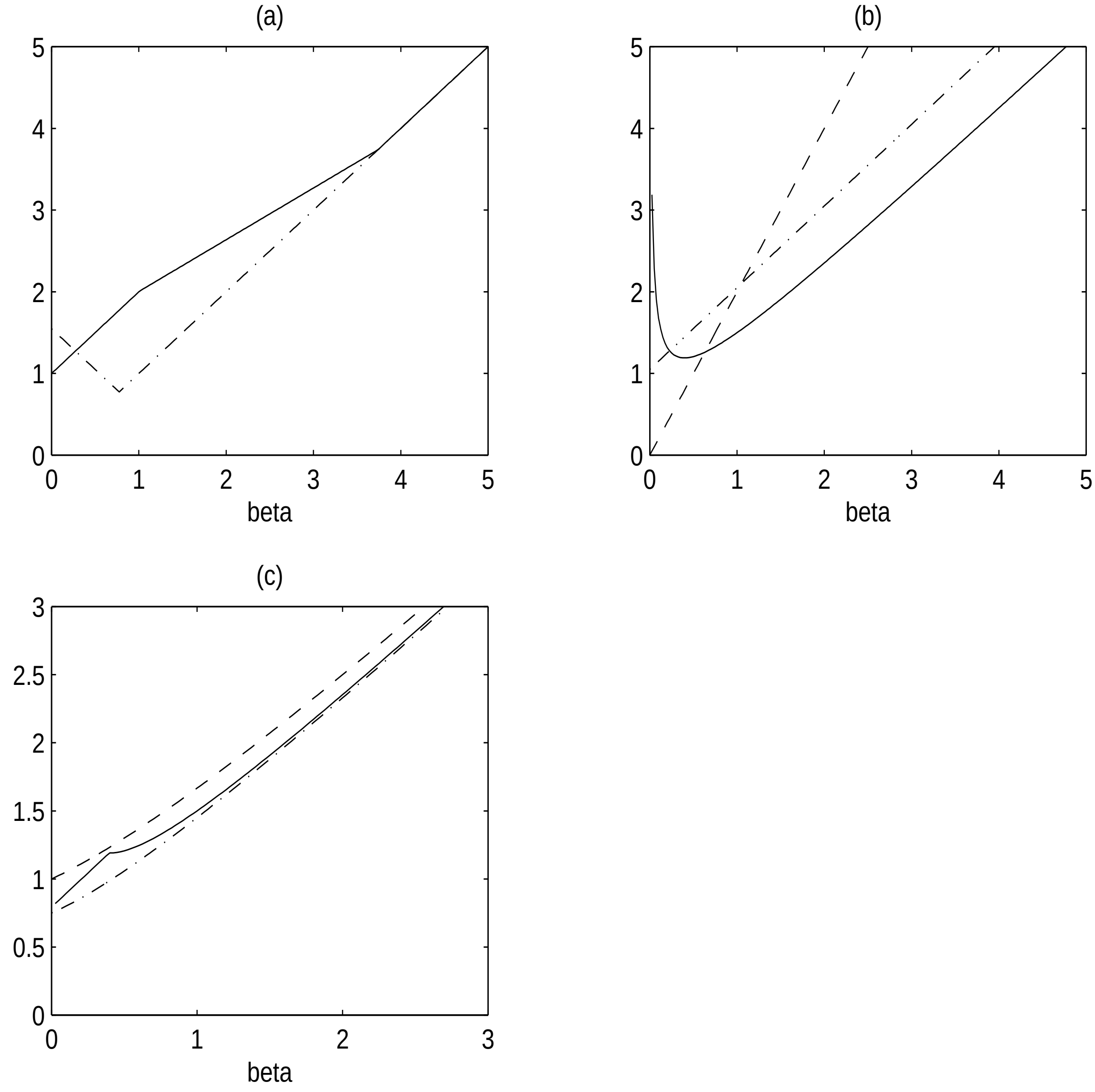

Figure 2: *Plot of $\beta + p'_\lambda(\beta)$ for $\beta > 0$. In (a), the solid line stands for $\beta + p'_\lambda(\beta)$ for SCAD, and dash-dotted line for the hard thresholding penalty function. In (b), the solid, dash-dotted and dashed lines correspond to the $L_{0.5}$, L_1 and L_2 penalty, respectively. In (c), the solid line stands for the modified L_1 penalty, and the dash-dotted and dashed lines are for the log-transformed L_p penalties with $p = 0.5$ and 1, respectively.*

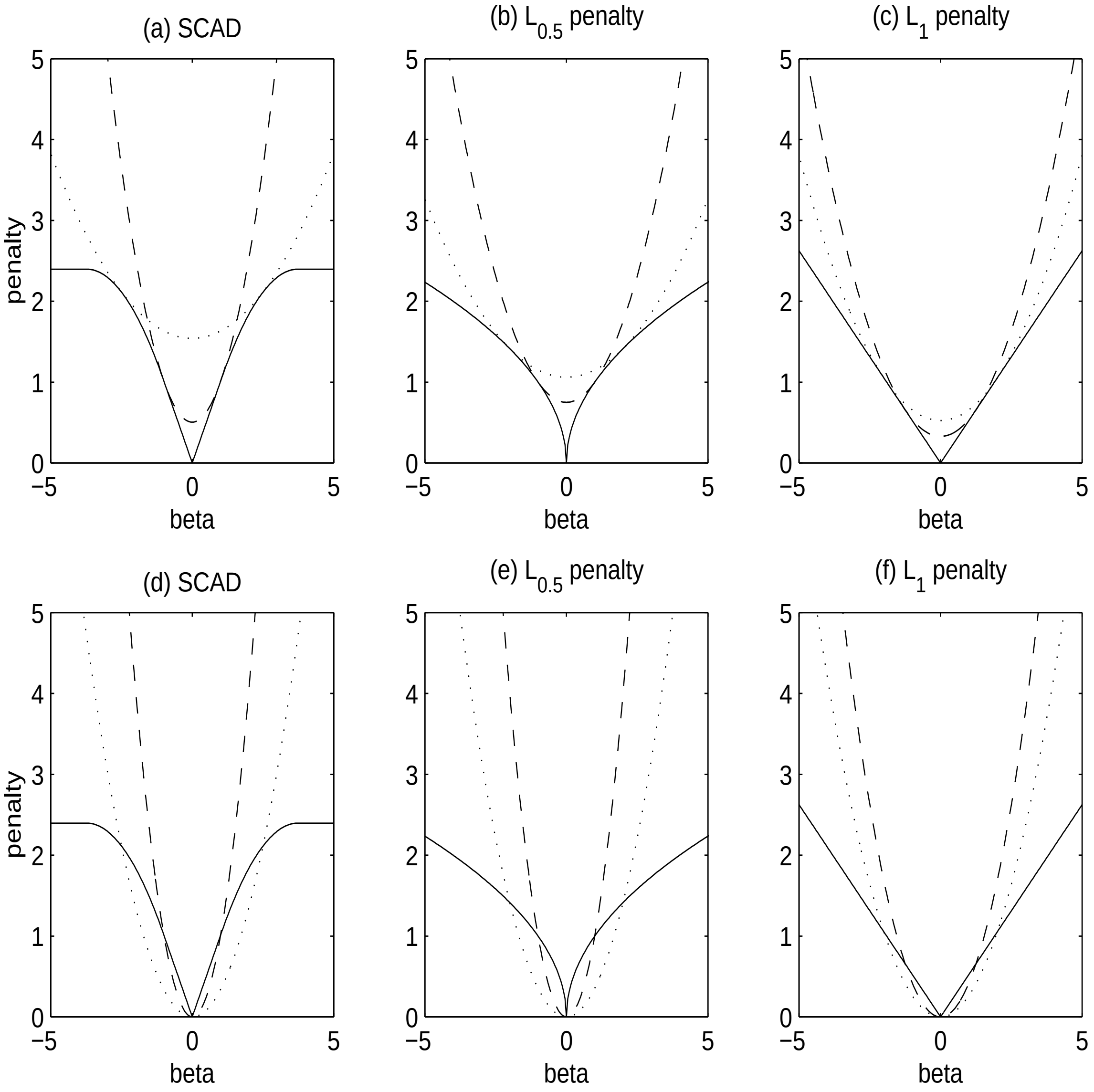

Figure 3: *Local quadratic approximations. (a), (b) and (c) display the local quadratic approximation (4.1) for the SCAD, $L_{0.5}$ and L_1 penalties, respectively. (d), (e) and (f) are the naive local quadratic approximation (4.9) for the SCAD, $L_{0.5}$ and L_1 penalties, respectively.*