Statistica Sinica **21** (2011), 391-419

NONCONCAVE PENALIZED M-ESTIMATION WITH A DIVERGING NUMBER OF PARAMETERS

Gaorong Li[1], Heng Peng[2] and Lixing Zhu[2]

[1]*Beijing University of Technology and* [2]*Hong Kong Baptist University*

Abstract: M-estimation is a widely used technique for robust statistical inference. In this paper, we investigate the asymptotic properties of a nonconcave penalized M-estimator in sparse, high-dimensional, linear regression models. Compared with classic M-estimation, the nonconcave penalized M-estimation method can perform parameter estimation and variable selection simultaneously. The proposed method is resistant to heavy-tailed errors or outliers in the response. We show that, under certain appropriate conditions, the nonconcave penalized M-estimator has the so-called "Oracle Property"; it is able to select variables consistently, and the estimators of nonzero coefficients have the same asymptotic distribution as they would if the zero coefficients were known in advance. We obtain consistency and asymptotic normality of the estimators when the dimension p_n of the predictors satisfies the conditions $p_n \log n/n \to 0$ and $p_n^2/n \to 0$, respectively, where n is the sample size. Based on the idea of sure independence screening (SIS) and rank correlation, a robust rank SIS (RSIS) is introduced to deal with ultra-high dimensional data. Simulation studies were carried out to assess the performance of the proposed method for finite-sample cases, and a dataset was analyzed for illustration.

Key words and phrases: Linear model, oracle property, rank correlation, robust estimation, SIS, variable selection.

1. Introduction

The modern technologies employed in many scientific fields allow production and storage of large datasets with ever-increasing sample sizes and dimensions, and that often include superfluous variables or information. Effective variable selection procedures are thus required to improve both the accuracy and interpretability of learning techniques. Selecting too small a subset leads to misspecification, whereas choosing too many variables aggravates the "curse of dimensionality"; selecting the right subset of variables and excluding unimportant variables when modeling high-dimensional data is of considerable importance.

In high-dimensional modeling, the classical L_0 penalized variable selection methods, AIC, BIC, C_P and so on, all suffer from a heavy computational burden, and the statistical properties of the estimators are difficult to analyze. To overcome these insufficiencies, various shrinkage or L_1 penalized variable selection methods, such as Bridge Regression (Frank and Friedman (1993)), LASSO

(Tibshirani (1996)), Elastic-Net (Zou and Hastie (2005)) and Adaptive LASSO (Zou (2006)), have been extensively investigated for linear and generalized linear models. Fan and Li (2001) proposed a nonconcave penalized likelihood method for variable selection in likelihood-based models, and showed that this method has some good properties compared to other penalized methods. The nonconcave likelihood approach has been further extended to Cox's model for survival data (Fan and Li (2002)), partially linear models for longitudinal data (Fan and Li (2004)) and varying coefficient partially linear models (Li and Liang (2008)). These shrinkage methods are much more computationally efficient than the classical variable selection methods and the properties of the estimators for most shrinkage methods are easier to study. It is easy to show that if the tuning parameters can be appropriately selected, then the true model can be consistently identified. For sparse high-dimensional models, Candés and Tao (2007) proposed the Dantzig selector, which is solution to an L_1-regularization problem. They showed that this selector has certain oracle properties in sparse high-dimensional linear regression models. Fan and Peng (2004) investigated the properties of nonconcave penalized estimators using the high-dimension likelihood approach. In addition, the asymptotic properties of penalized least squares estimators for sparse high-dimensional linear regression models have been widely investigated, see Huang, Horowitz and Ma (2008) for Bridge Regression, Zhang and Huang (2008) for LASSO, Zou and Zhang (2009) for Elastic-Net, Huang and Xie (2007) for SCAD.

To avoid model misspecification and increase the robustness of the estimation, a number of model-free dimension reduction methods have been considered. Zhou and He (2008), for example, proposed a constrained dimension reduction method based on canonical correlation and the L_1 constraint. Zhong et al. (2005) proposed a novel procedure called regularized sliced inverse regression (RSIR) to directly identify the linear combinations and further the functional TFBS, while avoiding estimation of the link function. Although these model-free dimension reduction methods are not only able to reduce the model dimensions, but also to shrink some of the coefficients of the selected linear combinations to zero, without information on the model structure the estimation efficiency is somewhat difficult to study and compare. The robust methods of some of the specified models have been widely investigated, but robust variable/model selection has received little attention. The seminal papers that do address this issue include those of Ronchetti (1985) and Ronchetti and Staudte (1994) introducing robust versions of the AIC and Mallows' C_p selection criteria, respectively. Ronchetti, Field, and Blanchard (1997) proposed robust model selection by cross-validation. Ronchetti (1997) discussed the robust model selection variants of the classical model selection criteria. Agostinelli (2002) used weighted likelihood to increase

the robustness of model selection. Wu and Zen (1999) proposed a linear model selection procedure based on M-estimation that includes many classical model selection criteria as special cases. Zheng, Freidlin, and Gastwirth (2004) suggested two measures based on Kullback-Leibler information for choosing a model and demonstrated the robustness of their proposed methods. Müller and Welsh (2005) proposed the selection of a regression model based on combining a robust penalized criterion and a robust conditional expected prediction loss function that is estimated using a stratified bootstrap. Khan, Van Aelst, and Zamar (2007) suggested a robust linear model selection method based on Least Angle Regression. Recently, Salibian-Barrera and Van Aelst (2008) proposed a robust model selection method that uses a fast and robust bootstrap. Unfortunately, most of the aforementioned robust model selection methods are based on classical model selection criteria, and hence are computationally expensive. Although a number of fast robust algorithms have been proposed, estimation properties remain little known, and it is thus difficult to investigate them in high-dimensional situations.

When using a high-dimensional statistical regression model to fit data, there are several problems that cannot be avoided. First, because of the high-dimension nature of the model and the data, it is difficult to determine outlying observations from the data by simple techniques or criteria. High-dimensionality also increases the likelihood of extreme covariates in the dataset. Second, as Fan and Lv (2008) discuss, strong correlation always exists between the covariates when the model dimensions are ultra-high. Thus, even when the model dimensions are smaller than the sample size, the design matrix is close to a singular matrix. Third, most of the theoretical results on penalized least squares in a high-dimensional regression model setting are based on the assumption of normality or the sub-Gaussian distribution of white noise. This assumption seems too restrictive. The white noise distribution is difficult to substantiate, and too many superfluous variables in a model affect the estimation and the final distribution of the residuals. Accordingly, the simple and direct use of penalized least squares is not a good choice because it is not a robust estimation method, and there is a large gap between the theoretical analysis and practical application.

The study of robust methods in a high-dimensional model setting is a necessary one. Due to the robust property of M-estimation and the good properties of the nonconcave penalty, we investigate nonconcave penalized M-estimation for high-dimension models and show that it has the so-called "Oracle property", and that it retains its robustness properties. We relax the assumption concerning white noise and assume only the existence of moments. Of course, there is a cost to doing so; we cannot obtain theoretical results that support us directly in applying the nonconcave penalized M-estimation to a case in which the model

dimensions are ultra-high relative to the sample size. To deal with this issue, we draw on the sure independence screening (SIS) concept presented in Fan and Lv (2008), and on rank correlation, to propose a robust rank SIS method, the RSIS. We first use the RSIS method to reduce the model dimensions to below the sample size; then, nonconcave penalized M-estimation is used to obtain the final estimation. Our proposed two-step procedure should retain some of the robustness properties supported by our numerical studies.

The remainder of the paper is organized as follows. In Section 2, we define the nonconcave penalized M-estimator and provide its asymptotic properties. The RSIS method is introduced in Section 3. In Section 4, we describe the algorithm used to compute the nonconcave penalized M-estimator and the criterion used to choose the regularization parameter. A two-step robust procedure based on the RSIS and penalized M-estimation is also discussed in this section to deal with ultra high-dimension cases. In Section 5, we apply a number of simulations to assess the finite sample performance of our penalized M-estimation method, and to compare it with other variable selection procedures. A simulation study and application are also presented in this section to compare the performance of the the RSIS with that of the SIS. The proofs of the main results are relegated to the Appendix.

2. Nonconcave Penalized M-estimation

2.1. M-estimation

Consider the linear regression model

$$y_i = \mathbf{x}_i^T \beta_n + e_i, \quad 1 \le i \le n, \tag{2.1}$$

where β_n is a $p_n \times 1$ unknown regression coefficient vector, and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip_n})^T$ are $p_n \times 1$ known predictors. Here, the subscript is used to make it explicit that both the covariates and the parameters may change with n. We assume that $e_1, \ldots, e_n$ are i.i.d. variables with common distribution function F throughout, unless otherwise stated. Without loss of generality, we assume the data are centered, and so the intercept is not included in the regression model.

As is well known, least squares (LS) is not a robust method, because it is sensitive to outliers and is much less efficient if the error distribution has heavier tails than the normal distribution. A robust method provides a useful and stable alternative that is not sensitive to outliers. Huber (1964, 1973, 1981) introduced M-estimation of β_n, which is defined as any value of $\hat{\beta}_n$ that minimizes

$$\sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \beta_n) \tag{2.2}$$

with a suitable choice of function ρ; or as any value of β_n that satisfies the estimating equation

$$\sum_{i=1}^{n} \psi(y_i - \mathbf{x}_i^T \beta_n)\mathbf{x}_i = 0 \tag{2.3}$$

for a suitable choice of function ψ. A natural method of obtaining (2.3) is to take the derivative of (2.2) with respect to β_n when ρ is continuously differentiable, and to equate it with the null vector. In general, ρ is a convex function. Important examples include Huber's estimate with $\rho(x) = (x^2\mathbf{1}_{|x|\le c})/2 + (c|x| - c^2/2)\mathbf{1}_{|x|>c}, c > 0$, the L_q regression estimate with $\rho(x) = |x|^q, 1 \le q \le 2$, and the regression quantiles with $\rho(x) = \rho_\alpha(x) = \alpha x^+ + (1-\alpha)(-x)^+, 0 < \alpha < 1$, where $x^+ = \max(x, 0)$. If $q = 1$ or $\alpha = 1/2$, then the minimizer of (2.2) is called the least absolute deviation (LAD).

Throughout this paper, we assume that ρ is a nonmonotonic convex function, ψ is a non-trivial nondecreasing function, and $p_n \to \infty$ as $n \to \infty$. Readers are referred to Huber (1973), Portnoy (1984, 1985), and Welsh (1989) for proofs of consistency and asymptotic normality for a class of the M-estimators of regression parameters under regularity conditions. The consistency of the M-estimates in high-dimensional regression models was considered by Huber (1973) for $p_n^2/n \to 0$, Yohai and Maronna (1979) for $p_n^2/n \to 0$, and Portnoy (1984) for $p_n \log p_n/n \to 0$, and their asymptotic normality by Huber (1973) for $p_n^3/n \to 0$, Yohai and Maronna (1979) for $p_n^{5/2}/n \to 0$, Portnoy (1985) for $(p_n \log n)^{3/2}/n \to 0$, and Mammen (1989) for $p_n^{3/2} \log n/n \to 0$. Welsh (1989) considered this problem under weaker conditions on functions ψ and F and stronger conditions on the ratio p_n/n. Bai and Wu (1994) further pointed out that the condition on p_n can be viewed as an integrated part of the design conditions. He and Shao (2000) further considered the asymptotic behavior of M-estimators with increasing dimensions for the more general parametric models.

However, these papers did not consider variable selection. To address this gap, we investigate penalized M-estimation for high-dimensional linear models. More specifically, we determine situations in which the nonconcave penalized M-estimator can correctly distinguish between nonzero and zero coefficients in sparse high-dimensional settings. We also investigate conditions under which the estimators of the nonzero coefficients have the same asymptotic distributions that they would have if the zero coefficients were known with certainty. Thus, in fact, we show that the nonconcave penalized M-estimators have the so-called oracle property in the sense discussed by Fan and Li (2001) and Fan and Peng (2004). The asymptotic properties of these estimators are investigated with $p_n \log n/n \to 0$ for consistency, and $p_n^2/n \to 0$ for asymptotic normality. Our investigation addresses the gap between SIS (Fan and Lv (2008)) and nonconcave penalized

least squares. Fan and Lv (2008) introduced the concept of sure screening and proposed the SIS method to reduce ultra-high dimensions to a relatively large scale that is smaller than or equal to the sample size for linear models. Because this relatively large scale is normally of the order $n/\log n$, nonconcave penalized M-estimation procedures can then be used to estimate the coefficients and select the variables simultaneously.

2.2. Penalized M-estimation

Let $\{y_i, \mathbf{x}_i^T\}, i = 1, \ldots, n$, be a random sample that satisfies

$$y_i = \mathbf{x}_i^T \beta_n + e_i. \tag{2.4}$$

Suppose that the p_n covariates can be classified into two categories: important covariates, whose corresponding coefficients are nonzero, and trivial ones, whose coefficients are zero. Throughout, let the true parameter value be β_{n0}. Let β_{n0} can be partitioned such that $\beta_{n0} = (\beta_{I0}^T, \beta_{II0}^T)^T$, where β_{I0} is a $k_n \times 1$ vector and β_{II0} is a $m_n \times 1$ vector, and $k_n + m_n = p_n$. Suppose that $\beta_{I0} \neq \mathbf{0}$ and $\beta_{II0} = \mathbf{0}$, where $\mathbf{0}$ is the vector with all components zero, so k_n is the number of nonzero coefficients and m_n is the number of zero coefficients. Which coefficients are nonzero and which are zero is unknown to us, but we partition β_{n0} in this way to facilitate the statement of the assumptions. Let $\mathbf{y} = (y_1, \ldots, y_n)^T$, and let $\mathbf{x} = (x_{ij}, 1 \leq i \leq n, 1 \leq j \leq p_n)$ be the $n \times p_n$ design matrix. According to the partition of β_{n0}, we write $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, where $\mathbf{x}_1$ and $\mathbf{x}_2$ are the $n \times k_n$ and $n \times m_n$ matrices, respectively. Let $S_n = \mathbf{x}^T\mathbf{x}$ and $S_{1n} = \mathbf{x}_1^T\mathbf{x}_1$.

We estimate the unknown parameter vector β_{n0} by minimizing

$$\sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \beta_n) + n \sum_{j=1}^{p_n} p_\lambda(|\beta_{nj}|), \tag{2.5}$$

where p_n is the dimension of β_n, $p_\lambda(\cdot)$ is a penalty function, and λ is a regularization parameter that can be chosen by a data-driven criterion such as cross-validation (CV), generalized cross-validation (GCV) (Craven and Wahba (1979); Tibshirani (1996)), or the BIC-type tuning parameter selector (Wang, Li, and Tsai (2007)). In practice, we may allow different parameters to have penalty functions with different regularization parameters. Various penalty functions have been used in the variable selection literature for linear regression models. Frank and Friedman (1993) considered the L_q penalty, $p_\lambda(|\beta_n|) = \lambda|\beta_n|^q, (0 < q < 1)$, which yields a "Bridge Regression". Tibshirani (1996) proposed the LASSO, which can be viewed as a solution to the penalized least squares with the L_1 penalty $p_\lambda(|\beta_n|) = \lambda|\beta_n|$. Fan and Li (2001) suggested that a good penalty function should have three properties: sparsity, unbiasedness, and continuity; more

details on the characterization of these three properties can be found in Fan and Li (2001) and Antoniadis and Fan (2001). These authors showed that singularity at the origin is a necessary condition to generate sparsity for penalty functions and that nonconvexity is required to reduce estimation bias. It is well known that the hard penalty function cannot satisfy the continuity condition, and that the L_q-penalty with $q > 1$ cannot satisfy the sparsity condition. The L_1-penalty (LASSO) possesses sparsity and continuity, but generates estimation bias.

Fan and Li (2001) showed that nonconcave penalty functions such as $L_q, 0 < q < 1$, can have these three properties. Fan (1997) proposed a special nonconcave penalty function called the Smoothing Clipped Absolute Deviation (SCAD) penalty function, which is defined by

$$p'_\lambda(|\beta_n|) = \lambda \left\{ I(|\beta_n| \le \lambda) + \frac{(a\lambda - |\beta_n|)_+}{(a-1)\lambda} I(|\beta_n| > \lambda) \right\} \quad \text{for some} \quad a > 2, \tag{2.6}$$

where the notation z_+ stands for the positive part of z. The SCAD penalty is continuously differentiable on $(-\infty, 0) \cup (0, \infty)$, but not differentiable at 0, and its derivative vanishes outside $[-a\lambda, a\lambda]$. As a consequence, SCAD penalized regression can produce sparse solutions and unbiased estimates for large parameters. To simplify the tuning parameter selection, Fan and Li (2001) suggested using $a = 3.7$ for the SCAD penalty function, drawing on a Bayesian point of view.

The nonconcave penalized M-estimator of β_{n0} is obtained by minimizing the objective function with nonconcave penalty

$$Q_n(\mathbf{b}; \lambda, a) = \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \mathbf{b}) + n \sum_{j=1}^{p_n} p_\lambda(|b_{nj}|). \tag{2.7}$$

If ρ is convex with derivative ψ, then (2.7) is equivalent to

$$\sum_{i=1}^{n} \psi(y_i - \mathbf{x}_i^T \hat{\beta}_n)\mathbf{x}_i - nP'_\lambda(|\hat{\beta}_n|) = 0, \tag{2.8}$$

where $P'_\lambda(|\hat{\beta}_n|)$ is a $p_n \times 1$ vector whose jth element is $p'_\lambda(|\hat{\beta}_{nj}|)\text{sgn}(\hat{\beta}_{nj})$. Here, the function $\text{sgn}(x)$ is equal to 1 for $x > 0$, 0 for $x = 0$, and -1 for $x < 0$. When $p_\lambda(\cdot) \equiv 0$, the solution of (2.7) is the M-estimate (Huber (1973)). For a given penalty parameter λ, the nonconcave penalized M-estimator of β_{n0} is

$$\hat{\beta}_n \equiv \hat{\beta}_n(\lambda) = \arg\min_{\mathbf{b}} \; Q_n(\mathbf{b}; \lambda). \tag{2.9}$$

2.3. Asymptotic properties

Here are the regularity conditions on $\rho, \mathbf{x}_i$, and the penalty functions that we employ:

(C1) ρ is a convex function on $\mathbb{R}^1$ with right and left derivatives $\psi_+(\cdot)$ and $\psi_-(\cdot)$, is any choice of the subgradient of $\rho(\cdot)$,

$$\psi_-(t) \le \psi(t) \le \psi_+(t) \quad \text{for all} \quad t \in \mathbb{R}^1, \tag{2.10}$$

and $\mathcal{S}$ is the set of discontinuity points of ψ.

(C2) The common distribution function F of e_i satisfies $F(\mathcal{S})$=0. $E[\psi(e_1)] = 0$, $0 < E[\psi^2(e_1)] = \sigma^2 < \infty$, and

$$G(t) \equiv E[\psi(e_1 + t)] = \gamma t + o(|t|), \quad \text{as} \quad t \to 0, \tag{2.11}$$

where γ is a positive constant. Furthermore,

$$\lim_{t\to 0} E[\psi(e_1 + t) - \psi(e_1)]^2 = 0. \tag{2.12}$$

Throughout the paper, $\rho_1(A) \le \cdots \le \rho_{p_n}(A)$ are the eigenvalues and $\mathrm{tr}(A)$ is the trace operator of a matrix A.

(C3) There are N_0 and constants b and B such that, for $n \ge N_0$,

$$0 < bn \le \rho_1(S_n) \le \rho_{p_n}(S_n) \le Bn. \tag{2.13}$$

(C4) There exists a sequence of fixed vectors $\{u\}$ in $\mathbb{R}^{p_n}$, with $\|u\|$ bounded, such that

$$\max\{|\mathbf{x}_i^T u| : i = 1, \ldots, n\} = O(\sqrt{\log n}). \tag{2.14}$$

(C5) $d_n^2 \equiv \max_{1\le i\le n} \mathbf{x}_{1i}^T S_{1n}^{-1} \mathbf{x}_{1i}$, where $\mathbf{x}_{1i}$ is an $k_n \times 1$ vector. When n is large enough, there exists a constant $s > 0$ such that $d_n \le sn^{-1/2}$.

Conditions (C1)–(C2) are standardly it imposed in the M-estimation theory of linear models; for examples, see Bai, Rao, and Wu (1992) and Wu (2007). Condition (C3) is a classical condition that has been assumed in the linear model literature. Condition (C4) can be found in Portnoy (1985). Condition (C5) is basically the Lindeberg-Feller type condition, it is required in the proof of the asymptotic normality of the estimators of nonzero coefficients, and can be found in Wu (2007). With condition (C5), the diagonal elements of the hat matrix $\mathbf{x}_1 S_{1n}^{-1} \mathbf{x}_1^T$ are uniformly negligible. If $\mathbf{x}_{1i_1}, \ldots, \mathbf{x}_{1i_{k_n}}$ are linearly independent, $1 \le i_1 \le \cdots \le i_{k_n}$, and $Q = (\mathbf{x}_{1i_1}, \ldots, \mathbf{x}_{1i_{k_n}})$, then Q is nonsingular, $Q^T S_{1n}^{-1} Q \to 0$, and consequently $S_{1n}^{-1} \to 0$ as $n \to \infty$. This implies that the minimum eigenvalue of S_{1n} diverges to ∞. It is a classical condition for weak consistency of the least squares estimators.

Let

$$a_n = \max\{|p'_{\lambda_n}(|\beta_{n0j}|)| : \beta_{n0j} \ne 0\}, \tag{2.15}$$

$$b_n = \max\{|p''_{\lambda_n}(|\beta_{n0j}|)| : \beta_{n0j} \ne 0\}, \tag{2.16}$$

where we write λ as λ_n to emphasize that λ_n depends the sample size n.

(C6) $\liminf_{n\to\infty} \liminf_{\theta\to 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0$.

(C7) $a_n = O(n^{-1/2})$.

(C8) $b_n \to 0$ as $n \to \infty$.

(C9) There are constants C and D such that, when $\theta_1, \theta_2 > C\lambda_n$, $|p''_{\lambda_n}(\theta_1) - p''_{\lambda_n}(\theta_2)| \le D|\theta_1 - \theta_2|$.

The following theorem shows how the convergence rates for the nonconcave-penalized M-estimators depend on the regularization parameter.

Theorem 1. *Suppose that ρ and $\mathbf{x}_i$ satisfy conditions* (C1–C4) *and the penalty function $p_{\lambda_n}(\cdot)$ satisfies conditions* (C7)–(C9). *If $p_n \log n/n \to 0$ as $n \to \infty$, then there exists a nonconcave penalized M-estimator $\hat{\beta}_n$ such that $\|\hat{\beta}_n - \beta_{n0}\| = O_P(p_n^{1/2}(n^{-1/2} + a_n))$.*

Theorem 1 has the nonconcave penalized M-estimator of β_{n0} root-(n/p_n) consistent if $a_n = O(n^{-1/2})$. For the SCAD penalty, if the nonzero coefficients are larger than $a\lambda_n$, then it can be easily shown that $a_n = 0$ when n is large enough, and hence the estimate of nonconcave penalized M-estimator β_n is root-(n/p_n) consistent.

Let

$$\mathbf{b}_n = (p'_{\lambda_n}(|\beta_{n01}|)\mathrm{sgn}(\beta_{n01}), \ldots, p'_{\lambda_n}(|\beta_{n0k_n}|)\mathrm{sgn}(\beta_{n0k_n}))^T, \tag{2.17}$$

$$\Sigma_{\lambda_n} = \mathrm{diag}\{p''_{\lambda_n}(|\beta_{n01}|), \ldots, p''_{\lambda_n}(|\beta_{n0k_n}|))\}, \tag{2.18}$$

where k_n is the number of components in β_{I0}.

Theorem 2.(Oracle property) *Under conditions* (C1)–(C9), *if $\lambda_n \to 0$, $\sqrt{n/p_n}\lambda_n \to \infty$, and $p_n^2/n \to 0$ as $n \to \infty$, then with probability tending to 1, the root-(n/p_n) consistent nonconcave penalized M-estimator $\hat{\beta}_n = (\hat{\beta}_I^T, \hat{\beta}_{II}^T)^T$ of* (2.9) *satisfies the following.*

(i) (Sparsity) $\hat{\beta}_{II} = \mathbf{0}$.

(ii) (Asymptotic normality) *If there exists a σ_4 such that $E[\psi^4(e_i)] \le \sigma_4 < \infty$, then*

$$A_n S_{1n}^{-1/2}\{\gamma S_{1n} + n\Sigma_{\lambda_n}\}[(\hat{\beta}_I - \beta_{I0}) + n\{\gamma S_{1n} + n\Sigma_{\lambda_n}\}^{-1}\mathbf{b}_n] \xrightarrow{L} N(\mathbf{0}, \sigma^2 G), \tag{2.19}$$

where "$\xrightarrow{L}$" stands for the convergence in distribution, and A_n is a $q \times k_n$ matrix such that $A_n A_n^T \to G$.

Theorem 2 implies that the nonconcave penalized M-estimators of the zero coefficients are exactly zero with strong probability when n is large. When n is large enough, and the nonzero-valued coefficients are larger than $a\lambda_n$, $\Sigma_{\lambda_n} = 0$ and $\mathbf{b}_n = 0$ for the SCAD penalty. Hence, the asymptotic normality (ii) of Theorem 2 becomes

$$A_n S_{1n}^{1/2}(\hat{\beta}_I - \beta_{I0}) \xrightarrow{L} N(\mathbf{0}, \gamma^{-2}\sigma^2 G), \tag{2.20}$$

which has the same efficiency as the M-estimator of β_{I0}, based on the submodel with β_{II0} known in advance. This has the nonconcave penalized M-estimator as efficient as the oracle estimate even when the number of parameters diverges. Fan and Peng (2004) considered maximum penalized likelihood estimation and required that the number of parameters, p_n, satisfy $p_n^4/n \to 0$ for consistency, and $p_n^5/n \to 0$ for asymptotic normality. It is easy to see from Theorem 1 and Theorem 2 that the consistency and the asymptotic normality of the nonconcave penalized M-estimator are true as long as $p_n \log n/n \to 0$ and $p_n^2/n \to 0$, respectively. This improves the order in some of the literature without requiring strong orthogonality between the y_i's, as in Portnoy (1985). It also addresses the gap between the SIS (Fan and Lv (2008)) and nonconcave penalized methods.

3. Rank Sure Independence Screening (RSIS)

In this section, we discuss how the method proposed in Section 2 can be applied to ultra-high dimensional data with p_n much larger than n. Candés and Tao (2007) suggested using the Dantzig selector, which is able to achieve the ideal estimation risk up to a $\log(p_n)$ factor under the uniform uncertainty condition. However, Fan and Lv (2008) showed that this condition may easily fail, and that the $\log(p_n)$ factor becomes too large when p_n is exponentially large. Moreover, the computational cost of the Dantzig selector becomes very high when p_n is large. To overcome these difficulties, Fan and Lv (2008) proposed a two-stage procedure. First, SIS is used as a fast, but crude, method of reducing the ultra-high dimensionality to a relatively large scale that is still smaller than or equal to sample size n; then, a more sophisticated technique can perform the final variable selection and parameter estimation simultaneously. This relatively large scale is normally of the order $n/\log n$.

Based on the SIS concept, let $\omega = (\omega_1, \ldots, \omega_{p_n})^T$ be a p_n-vector that is obtained by computing

$$\omega_k = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} I(x_{ik} < x_{jk}) I(y_i < y_j) - \frac{1}{4}, \quad k = 1, \ldots, p_n. \tag{3.1}$$

We sort the p_n magnitudes of the vector $|\omega|$ in a decreasing order and define a submodel

$$\mathcal{M} = \{1 \le i \le p_n : |\omega_i| \text{ is among the first } [\frac{cn}{\log n}] \text{ largest of all}\}, \qquad (3.2)$$

where c is a positive constant and $[cn/\log n] < n$. This is a straightforward way of shrinking the full model, $\{1, \ldots, p_n\}$, to a submodel $\mathcal{M}$ with size $d_n = [cn/\log n] < n$. Similar to SIS, we can reduce the model size from ultra-high dimensionality to a relatively large scale if the first $[cn/\log n]$ of $|\omega_i|, i = 1, \ldots, p_n$ has a high probability of including all of the effective variables. We call such a method Rank SIS (RSIS).

The SIS concept is based on correlation learning, but the Pearson correlation it uses is sensitive to the outlying or influence points. Moreover, Pearson correlation is unable to identify the nonlinear relationship between the response variables and predictor variables. RSIS thus makes use of rank correlation rather than Pearson correlation and we expect the proposed RSIS not only to reduce the model size, similarly to SIS, but also to be more robust than it.

Using RSIS or SIS, model dimensions are reduced to a relatively large scale. Lower-dimensional model selection methods, such as the SCAD, LASSO, adaptive Elastic-Net, and hard thresholding, can then be used to estimate the model. We have many choices, but to obtain a robust and efficient estimation of the model, we prefer RSIS with penalized M-estimation.

4. Practical Issues Surrounding Penalized M-estimation

Finding the estimator of β_n that minimizes the objective function (2.7) poses a number of interesting challenges because the penalized functions are nondifferentiable at the origin and nonconcave with respect to β_n. Fan and Li (2001) suggest iterative, local approximation of the penalty function by a quadratic function, referring to such approximation as local quadratic approximation (LQA). With the aid of LQA, the optimization of the penalized objective function can be carried out using a modified Newton-Raphson algorithm. However, as pointed out in Fan and Li (2001) and Hunter and Li (2005), the LQA algorithm shares the drawback of backward stepwise variable selection: a covariate deleted at any step in the LQA algorithm is necessarily excluded from the final model. To overcome this computational difficulty, Hunter and Li (2005) proposed an MM algorithm that optimizes a slightly perturbed version of the LQA. Although the MM algorithm addresses the drawback of the LQA, the perturbation size is difficult to determine. To overcome this weakness of the LQA algorithm, Zou and Li (2008) proposed a unified algorithm based on local linear approximation (LLA). In the present paper, we use both the original LQA algorithm (Fan and Li (2001)) and

the perturbed LAQ algorithm (Hunter and Li (2005)) to compute the nonconcave penalized M-estimators for a given λ_n and a. The algorithms are presented in Subsection 4.1.

4.1. Local quadratic approximation (LQA)

The nonconcave penalty function is singular at the origin and has no continuous second-order derivative. Suppose that we assign an initial value β_n that is close to the true value β_{n0}. If β_{nj} is very close to 0, then we set $\hat{\beta}_{nj} = 0$; otherwise, the penalty function is locally approximated by a quadratic function using

$$[p_{\lambda_n}(|\beta_{nj}|)]' = p'_{\lambda_n}(|\beta_{nj}|)\mathrm{sgn}(\beta_{nj}) \approx \Big\{\frac{p'_{\lambda_n}(|\beta_{n0j}|)}{|\beta_{n0j}|}|\Big\}\beta_{nj}.$$

In other words,

$$p_{\lambda_n}(|\beta_{nj}|) \approx p_{\lambda_n}(|\beta_{n0j}|) + \frac{1}{2}\Big\{\frac{p'_{\lambda_n}(|\beta_{n0j}|)}{|\beta_{n0j}|}\Big\}(\beta_{nj}^2 - \beta_{n0j}^2), \quad \text{for } \beta_{nj} \approx \beta_{n0j}. \tag{4.1}$$

Then, the Newton-Raphson algorithm can be modified to find the minimum of the nonconcave penalized M-estimation objective function (2.7). More specifically, we take the unpenalized M-estimate to be the initial value β_n^0. We then update the estimate of β_n repeatedly until convergence with

$$\beta_n^{(k+1)} = \arg\min_{\beta}\left\{\sum_{i=1}^{n}\rho(y_i - \mathbf{x}_i^T\beta_n) + n\sum_{j=1}^{p_n}\frac{p'_{\lambda_n}(|\beta_{nj}^{(k)}|)}{2|\beta_{nj}^{(k)}|}\beta_{nj}^2\right\}, \quad k = 1, 2, \ldots. \tag{4.2}$$

Fan and Li (2001) suggested that to avoid numerical instability if $\beta_{nj}^{(k)}$ in (4.2) is smaller than or equal to the predefined small cutoff value ϵ_0, then set $\hat{\beta}_{nj} = 0$ and delete the jth component of the covariate from the iteration. There is no criterion for such a cutoff value. In our numerical study, this value is set to $\epsilon_{0j} = \tau \cdot \mathrm{std}(\hat{\beta}_{nj}), j = 1, \ldots, p_n$, where $\tau = 0.6$ and $\mathrm{std}(\hat{\beta}_{nj})$ is the estimate of the standard deviation of the nonpenalized M-estimate of β_{nj}. The use of $\mathrm{std}(\hat{\beta}_{nj})$ is designed to remove the effect of scale, and $\tau = 0.6$ is an empirical selection. In our numerical study, we also tried τ as $0.1, \ldots, 0.5$, to update the penalized M-estimate. The results exhibited little difference from those with $\tau = 0.6$, but demanded more computational time due to the additional number of iterative steps that sometimes renders the numerical results unstable. We recommend $\tau = 0.6$ for practical use.

We also note that when using cutoff values to set some coefficients to zero in every iterative step, the LQA algorithm becomes a backward stepwise algorithm

in selecting the variables and estimating the coefficients. To avoid this drawback, Hunter and Li (2005) suggested optimizing a slightly perturbed version of (4.2) with the denominator bounded away from zero. More specifically, they recursively solved

$$\beta_n^{(k+1)} = \arg\min \left\{ \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \beta_n) + n \sum_{j=1}^{p_n} \frac{p'_{\lambda_n}(|\beta_{nj}^{(k)}|)}{2\{|\beta_{nj}^{(k)}| + \tau_0\}} \beta_{nj}^2 \right\} \qquad k = 1, 2, \ldots, \tag{4.3}$$

for a prespecified size perturbation τ_0, and then performed the iteration until the sequence of $\{\beta_n^{(k)}\}$ converged. Again, it is difficult to choose the size of the perturbation in implementation. Furthermore, the size of τ_0 potentially affects the solution's degree of sparsity and the speed of convergence. Hunter and Li (2005) suggested using

$$\tau_0 = \frac{\xi}{2n\lambda_n} \min\{|\beta_{n0j}| : \beta_{n0j} \neq 0\} \tag{4.4}$$

for a given tolerance ξ and give a more detailed discussions. As in their algorithm, we are involved in selecting a tuning parameter ξ, but do not consider Hunter and Li's algorithm in our comparisons.

4.2. Selection of λ_n

To implement the procedures described in Section 2, we need to choose the regularization parameter λ_n. One can select λ_n by minimizing the generalized cross validation criterion. Wang, Li, and Tsai (2007) pointed out that this criterion has a nonignorable overfitting effect even as the sample size goes to infinity. They further proposed a BIC-based tuning parameter selector that they showed to be able to identify the true model consistently. This motivated us to select the optimal λ_n by minimizing the BIC (Schwarz (1978)) information criterion

$$\mathrm{BIC}(\lambda_n) = n \log \left(\sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \hat{\beta}_n) \right) + \mathrm{DF}_{\lambda_n} \log(n). \tag{4.5}$$

Here DF_{λ_n} is the generalized degrees of freedom (Fan and Li (2001)) $\mathrm{DF}_{\lambda_n} = \mathrm{tr}\{\mathbf{x}(\mathbf{x}^T\mathbf{x} + nD(\hat{\beta}_n; \lambda_n))^{-1}\mathbf{x}^T\}$, where $D(\hat{\beta}_n; \lambda_n)$ is the diagonal matrix whose diagonal elements are $(1/2)p'_{\lambda_n}(|\hat{\beta}_{nj}|)/|\hat{\beta}_{nj}|, j = 1, \ldots, p_n$.

5. Numerical Studies

5.1. Penalized M-estimation

In this subsection, we report on two simulation studies to illustrate the finite sample properties of the nonconcave penalized M-estimator with heavy-tailed

errors. We investigated in Table 1 two features: (i) variable selection and (ii) prediction performance. For (i), we report in Table 1 the average numbers of the correct and incorrect zero coefficients in the final models. For (ii), we compute the *model error* $\mathrm{ME} \equiv (\hat{\beta}_n - \beta_n)^T E(\mathbf{x}\mathbf{x}^T)(\hat{\beta}_n - \beta_n)$.

Example I. We simulated covariates $\mathbf{x}_i, i = 1, \ldots, n$ from the multivariate normal distributions with mean 0 and

$$\mathrm{Cov}(x_{ij}, x_{il}) = \rho^{|j-l|}, \quad 1 \le j, l \le p_n. \tag{5.1}$$

The response variables were generated according to the model

$$y_i = \mathbf{x}_i^T \beta_n + \sigma e_i, \tag{5.2}$$

where $\beta_n = (2, 1.5, 0.8, -1.5, 0.4, 0, \ldots, 0)^T$. Thus the first $k_n = 5$ regression variables were significant, but the rest were not, and the dimensionality of the parametric component was taken to be $p_n = \lfloor 1.8n^{1/2} \rfloor$. Noise e_i was generated from four different distributions: the standard normal, the mixture normal $0.9N(0,1) + 0.1N(0,9)$, the standard t with three degrees of freedom, and the standard t with five degrees of freedom. Two different values, $\sigma = 0.5$ and 1.0, which represent strong and weak signal-to-noise ratios, were considered. For comparison, three ρ functions were employed as loss functions: $\rho_1(t) = t^2$ (LS); $\rho_2(t) = |t|$ (LAD); and $\rho_3(t) = 0.5t^2$ if $|t| \le 1.345$ and $\rho_3(t) = 1.345|t| - 0.5(1.345)^2$ otherwise (Huber ρ). We abbreviate the estimators obtained by minimizing these three loss functions with the SCAD penalty as the LS-SCAD, LAD-SCAD, and Huber-SCAD estimators, respectively. Sample size n was taken to be 500 and 1,200 in this simulation, and the corresponding dimensions of parameter vector β_n were 40 and 62, respectively. In this simulation example, we applied the BIC criterion of (4.5) to select the tuning parameters, and took the unpenalized M-estimate to be the initial estimate by using the aforementioned three loss functions, respectively. For each case, we repeated the experiment 100 times and adopted only the SCAD as the penalty function to compare its performance with the original LQA algorithm (4.2) and the perturbed LAQ algorithm (4.3) with $\xi = 10^{-8}$ in (4.4). The simulation results of these two algorithms were similar, and so we report only the simulation results of the LQA algorithm here.

Table 1 reports the average number of zero coefficients for the linear model (5.2) with the different error distributions and signal-to-noise ratios, and Figure 1 gives the box plots for these average model errors.

As can be seen from Table 1, all of the variable selection procedures were able to correctly identify the true submodel, but the LAD-SCAD and Huber-SCAD procedures performed significantly better than did the LS-SCAD procedure and fit the true submodel very well. It can be seen from Figure 1 that the average

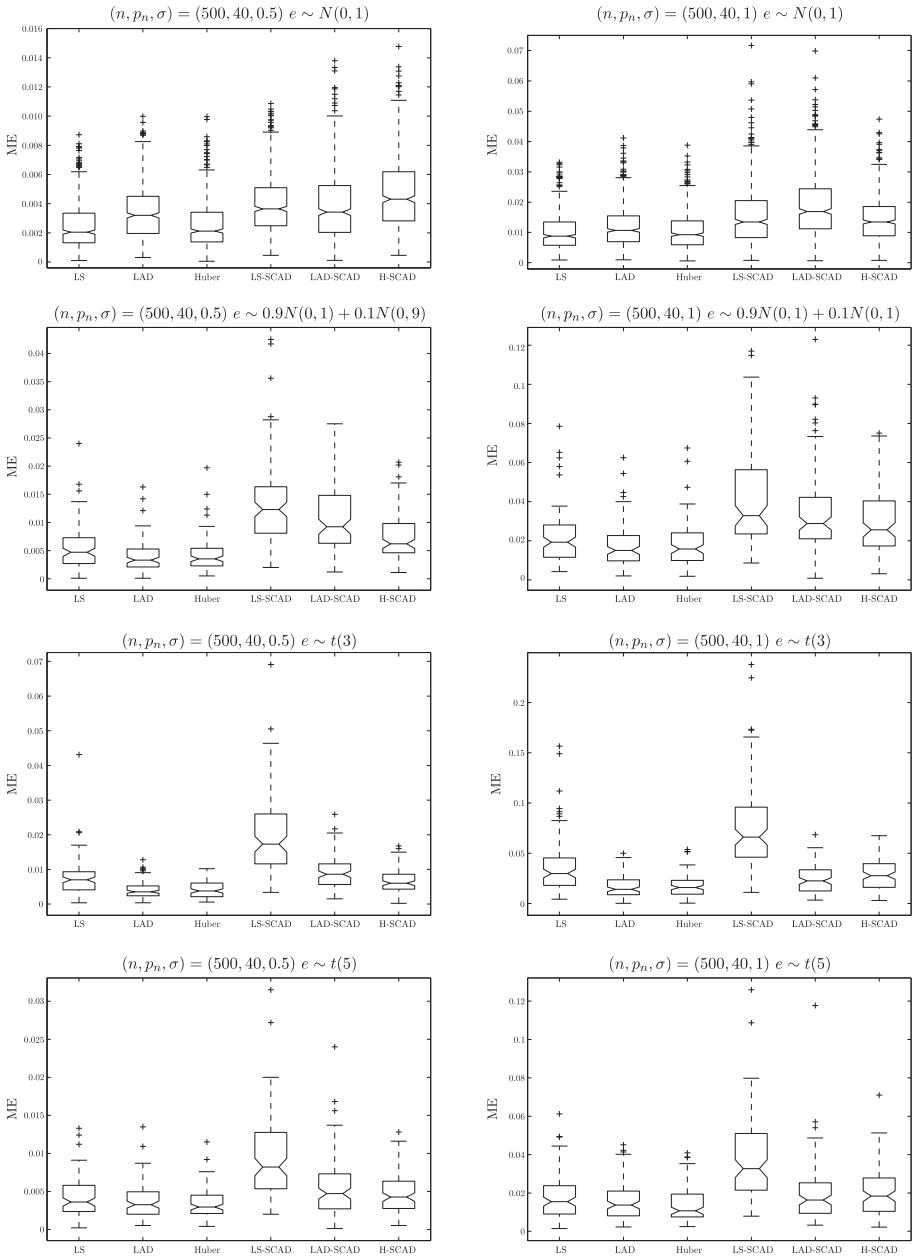

Figure 1. Box plots of the average model errors for the ρ functions LS, LAD, and Huber, Here without the SCAD penalty for the oracle linear model. LS-SCAD, LAD-SCAD and H-SCAD represent the three methods with the SCAD penalty, using the LQA algorithm for the original high-dimensional linear model.

Table 1. Average numbers of zero coefficients.

	$e \sim$	$N(0,1)$		$N_{\text{mix}}{}^a$		$t(3)$		$t(5)$	
(n, p_n, σ)	Method	C	IC	C	IC	C	IC	C	IC
(500,40,0.5)	Truth	35.00	0.00	35.00	0.00	35.00	0.00	35.00	0.00
	LS-SCAD	28.98	0.00	29.04	0.00	29.50	0.00	29.43	0.00
	LAD-SCAD	31.07	0.00	31.74	0.00	32.72	0.00	32.31	0.00
	Huber-SCAD	33.10	0.00	33.00	0.00	33.36	0.00	33.35	0.00
(500,40,1.0)	LS-SCAD	28.86	0.00	29.34	0.00	29.44	0.01	29.43	0.00
	LAD-SCAD	33.32	0.00	34.09	0.00	34.00	0.00	33.99	0.00
	Huber-SCAD	32.90	0.00	33.16	0.00	32.85	0.00	33.16	0.00
(1,200,62,0.5)	Truth	57.00	0.00	57.00	0.00	57.00	0.00	57.00	0.00
	LS-SCAD	47.70	0.00	47.65	0.00	47.95	0.00	47.29	0.00
	LAD-SCAD	51.44	0.00	52.22	0.00	53.40	0.00	52.51	0.00
	Huber-SCAD	54.19	0.00	53.77	0.00	53.92	0.00	53.99	0.00
(1,200,62,1.0)	LS-SCAD	47.50	0.00	47.67	0.00	47.95	0.00	47.92	0.00
	LAD-SCAD	54.98	0.00	55.70	0.00	55.80	0.00	55.50	0.00
	Huber-SCAD	53.88	0.00	54.02	0.00	53.85	0.00	54.02	0.00

[a] N_{mix} denotes the mixture normal distribution $0.9N(0,1) + 0.1N(0,9)$.
"C" presents the average numbers of zero coefficients correctly estimated to be zero;
"IC" presents the average numbers of nonzero coefficients erroneously set to zero.

model errors tended to be more diffuse as σ increased. The oracle estimators performed the best. The LAD-SCAD and Huber-SCAD performed comparably to the oracle estimators, and had smaller average model errors and were more stable than the LS-SCAD under heavy-tailed error distributions.

For each estimator $\hat{\beta}_I$ of the nonzero coefficients, estimation accuracy was measured by the bias and the median absolute deviation divided by 0.6745 (MAD) among 100 simulations. The bias was computed by the difference between the median of the estimated coefficients based on 100 simulations and the true value. MAD is a robust measure of variability, and a more robust estimator than the variance or standard deviation. Table 2 presents the results for the nonzero coefficients when the error distribution was the standard t-distribution with three degrees of freedom. As the results for the other cases were similar, we do not report them here. From the simulation results in Table 2, we can see that the LAD-SCAD estimator exhibited similar performance to that of the Huber-SCAD estimator in terms of biasedness and the median absolute deviation. In the situation in which the data were generated from the $t(3)$ distribution, the LAD-SCAD and Huber-SCAD estimates had smaller biases and median absolute deviations (MAD) relative to the LS-SCAD estimates. They performed better and were more stable than the LS-SCAD even as σ increased and the dimension of the parameters grew with sample size n. It is worth mentioning that the

Table 2. Bias and MAD (multiplied by 1,000) of the estimators with error $e \sim t(3)$.

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_3$		$\hat{\beta}_4$		$\hat{\beta}_5$	
Method	Bias	MAD	Bias	MAD	Bias	MAD	Bias	MAD	Bias	MAD
$n = 500,\ p_n = 40,\ \sigma = 0.5$										
LS-SCAD	-7.0	36.5	-4.0	48.0	-3.1	52.1	6.2	51.8	-1.0	50.5
LAD-SCAD	1.9	30.3	-2.9	42.3	1.2	31.6	3.2	40.7	-0.7	34.9
Huber-SCAD	-2.3	28.9	-0.8	36.1	-2.6	34.6	1.7	38.5	-0.7	34.1
$n = 500,\ p_n = 40,\ \sigma = 1$										
LS-SCAD	15.9	76.5	-6.3	87.6	-6.0	97.0	-24.7	115	-8.8	88.5
LAD-SCAD	4.9	62.4	-7.6	79.3	3.5	61.2	-8.4	60.1	-1.1	72.7
Huber-SCAD	-2.7	73.3	-1.2	84.0	3.4	74.5	-11.0	80.7	1.1	69.1
$n = 1,200,\ p_n = 62,\ \sigma = 0.5$										
LS-SCAD	1.8	30.2	0.8	31.8	-1.5	35.9	7.0	28.2	-1.8	28.3
LAD-SCAD	1.1	24.5	1.4	27.0	2.4	25.9	1.8	22.3	1.0	24.8
Huber-SCAD	1.1	26.6	1.6	22.7	-0.2	25.4	1.2	22.5	-0.7	27.1
$n = 1,200,\ p_n = 62,\ \sigma = 1$										
LS-SCAD	-13.7	60.5	3.4	63.7	-7.2	71.8	13.9	56.4	-4.1	56.6
LAD-SCAD	-4.0	44.8	-0.2	50.4	-3.0	52.7	4.6	45.4	-0.9	54.2
Huber-SCAD	-5.3	55.9	-0.9	46.1	-3.8	52.1	4.2	46.6	-0.9	54.8

LAD-SCAD and Huber-SCAD estimators of the nonzero coefficients had trivial biases in various situations.

Example II. For comparison with the LASSO (Tibshirani (1996)), the adaptive Elastic-net (AEnet) (Zou and Hastie (2005); Zou and Zhang (2009)), and the hard thresholding rule (Antoniadis (1997); and Fan and Li (2001)), we considered model (5.2) with noise levels $\sigma = 1.5$ and 3, and took the (moderate) sample sizes $n = 200$ and 400 and the corresponding dimensions of the parameter vector $p_n = 25$ and 36, respectively. In this example, the BIC was applied to estimate the tuning parameter for each variable selection procedure by using the corresponding loss functions. To compare the performance of the methods, the mean and standard deviation (SD) of the model errors, and the average number of zero coefficients of 500 simulated datasets are summarized in Table 3 based on the error distribution $t(3)$ (the results for the other cases are similar). It is worth mentioning that we used only the LAD with $\rho(t) = |t|$ as the loss function for the Oracle estimates in Table 3.

From Table 3, it can be seen that even when the noise level was high and the sample size was small, the LAD-SCAD and Huber-SCAD performed best and significantly reduced both model error and complexity, whereas the AEnet performed better than LASSO. The other variable selection procedures also reduced model error and model complexity. However, when the noise level increased, the LS-SCAD and the hard thresholding rule performed the worst.

Table 3. Model selection and fitting results based on error $e \sim t(3)$.

Method	ME	No. of Zeros		ME	No. of Zeros	
	mean(SD)	C	IC	mean(SD)	C	IC
	$(n, p_n, \sigma) = (200, 25, 1.5)$			$(n, p_n, \sigma) = (200, 25, 3)$		
Oracle	0.0882 (0.0839)	20.00	0.00	0.3762 (0.2426)	20.00	0.00
LS-SCAD	0.3785 (0.2838)	16.90	0.04	2.1495 (1.0999)	16.92	0.13
LAD-SCAD	0.2307 (0.1463)	19.79	0.07	0.9975 (0.6298)	19.99	0.30
Huber-SCAD	0.2027 (0.1163)	18.71	0.03	0.9834 (0.5318)	18.67	0.13
LASSO	0.2690 (0.1343)	18.34	0.04	1.3451 (0.5829)	19.68	0.19
AEnet	0.2402 (0.1082)	18.82	0.04	1.2570 (0.4505)	19.33	0.13
Hard	0.2449 (0.1708)	19.11	0.04	1.6119 (1.0691)	19.10	0.12
	$(n, p_n, \sigma) = (400, 36, 1.5)$			$(n, p_n, \sigma) = (400, 36, 3)$		
Oracle	0.0436 (0.0280)	31.00	0.00	0.1941 (0.1354)	31.00	0.00
LS-SCAD	0.2164 (0.1435)	26.00	0.01	0.8612 (0.4823)	26.01	0.06
LAD-SCAD	0.0954 (0.0576)	30.68	0.02	0.5502 (0.3488)	31.00	0.20
Huber-SCAD	0.0982 (0.0512)	28.87	0.00	0.4448 (0.2280)	29.14	0.06
LASSO	0.1469 (0.0677)	28.62	0.00	0.7564 (0.3311)	30.60	0.13
AEnet	0.1329 (0.0631)	28.56	0.00	0.5401 (0.3993)	29.42	0.10
Hard	0.1370 (0.0934)	29.47	0.01	0.9396 (0.5142)	29.69	0.07

5.2. Ultra-high dimensional case

To examine the performance of the proposed method in the ultra-high dimensional case, we first applied RSIS and SIS to reduce the dimensions down to the order of $n/\log n$, and then fit the data by using the procedure proposed in Section 2 and compared it with existing approaches. For comparison, the Dantzig selector procedure proposed by Candés and Tao (2007) was used in the following example. The Matlab codes for the algorithm are available at the website `http://www.acm.caltech.edu/l1magic/`.

In this simulation study, the model under study was similar to that in Fan and Lv (2008): $Y = \mathbf{x}^T\beta_n + 1.5e$. Here, noise e was drawn from the standard normal and the standard normal with 10% outliers drawn from the standard Cauchy. Covariate $\mathbf{x}$ was generated from the independent standard normal. We considered $(n, p_n) = (200,\ 1{,}000)$ based on 100 datasets, where the number of nonzero coefficients was 8. Each nonzero coefficient was chosen randomly, and generated as $(-1)^U(4\log n/\sqrt{n} + |Z|/4)$ with $Z \sim N(0, 1)$, where U was Bernoulli with parameter 0.5. We first used both RSIS and SIS to reduce the dimensionality from 1,000 to $d_n = [5n/\log n] = 188$. For each method, we report the median of the selected model sizes (SMS), the median of the standard deviation (SD) of model errors (ME), and the median of the estimation errors $\|\hat{\beta}_n - \beta_n\|$ in L_2-norm (EE), see Table 4.

From Table 4, we see the following.

Table 4. Medians of the selected model size (SMS) and the estimation errors (EE) in L_2-norm, and the median and standard deviation (SD) of the model errors (ME).

	$e \sim$	$N(0,1)$			$N(0,1)$ with 10% outliers		
				ME			ME
Method		SMS	EE	median(SD)	SMS	EE	median(SD)
RSIS+							
	LS-SCAD	17.5	0.4762	0.2206 (0.3510)	19	1.7747	1.0114 (0.9151)
	LAD-SCAD	16	0.4401	0.1836 (0.4352)	13	0.4915	0.2464 (0.8663)
	LASSO	45	0.8206	0.6244 (0.4020)	53	1.3505	1.2604 (1.0909)
	AEnet	18	0.4354	0.3647 (0.3634)	23	1.0729	1.0006 (0.7070)
	Hard	53	1.2118	0.9215 (0.5548)	59	2.7174	2.1585 (4.0463)
SIS+							
	LS-SCAD	21	0.4594	0.4078 (0.3896)	33	1.4382	1.5267 (1.1197)
	LAD-SCAD	16	0.4078	0.1977 (0.3491)	17	0.5490	0.3270 (0.9427)
	LASSO	44	0.8289	0.6103 (0.5663)	49.5	1.2593	1.2730 (1.1435)
	AEnet	12	0.4830	0.4176 (0.2264)	30	1.1625	1.0327 (1.0453)
	Hard	54	1.2362	0.9105 (0.5182)	116.5	5.7588	3.4480 (-)
Dantzig		10^3	3.9532	2.2357 (0.2147)	10^3	4.3499	3.2613 (-)

(1) When noise e was drawn from the standard normal, the LAD-SCAD, LS-SCAD, and AEnet based on RSIS and SIS dimensionality reduction outperformed the other variable selection procedures in terms of the selected model sizes, model errors, and estimation errors. The LASSO and hard thresholding rule performed much worse, with larger models and estimation errors. This is not surprising, because the LASSO is not unbiased and the hard thresholding penalty function does not satisfy the continuity condition. The AEnet performed significantly better than the LASSO. The Dantzig selector failed to generate a sparse model and had larger estimation errors.

(2) When the data were contaminated with 10% outliers, the LAD-SCAD method was much more stable and performed much better than did the other variable selection procedures. However, the LS-SCAD performed worse than did the AEnet. The hard thresholding rule and the Dantzig selector had very large standard deviations of model errors.

(3) RSIS outperformed SIS, especially for data with outliers.

Generally speaking, the LAD-SCAD selected a smaller number of important variables and obtained more accurate models than did the other procedures, in view of the estimation errors and model errors. Our proposed methods were not sensitive to outliers or error distributions with heavier tails, and can be considered as robust variable selection and parameter estimation procedures.

Table 5. Results for the ovarian cancer data.

RSIS+	SCAD*	SCAD	LASSO	AEnet	Hard
Number of selected variables	14	18	22	16	33
Test error	2/113	3/113	3/113	2/113	10/113
SIS+	SCAD*	SCAD	LASSO	AEnet	Hard
Number of selected variables	18	22	28	22	35
Test error	2/113	3/113	4/113	3/113	11/113

5.3. A real data example: Ovarian cancer data

The ovarian dataset 8-7-02 was provided by the National Cancer Institute (NCI) and is available at `http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp`. Wu et al. (2003) investigated the classification of ovarian cancer by using several statistical methods, and Yu et al. (2005) proposed a novel method for dimensionality reduction to analyze raw ovarian cancer MS data. The dataset includes 15,154 features and a total of 253 spectra samples: 162 ovarian cancer samples and 91 control samples. We randomly divided this dataset into a training sample with 140 cases (89 ovarian cancer samples and 51 control samples) and a test sample of 113 cases (73 ovarian cancer samples and 40 control samples). Using such a dataset for cancer classification is challenging because the data are of a very high dimension and the sample size is relatively small. Of the large number of features, only a small portion may benefit the correct classification of cancers, with the remainder having little impact. Even worse, some of the features may act as "noise" and undermine pattern recognition. Therefore, feature selection becomes crucial here. By removing features that are irrelevant, prediction accuracy can usually be improved.

The data can be written as $S = \{(\mathbf{x}_i^T, y_i)|\ \mathbf{x}_i \in \mathbb{R}^p, y_i = 0, 1, i = 1, 2, \ldots, n\}$, where $\mathbf{x}_i$ is an intensity vector and y_i denotes the sample cancer status (0 for control, 1 for cancer). The logistic regression model with binary response was used to fit these data. Here, we first applied RSIS and SIS to reduce the dimensionality from $p = 15,154$ to $d_n = [4n/\log n] = 113$, with $n = 140$ the training sample size chosen, and then employed the lower-dimensional model selection methods, the SCAD, LASSO, AEnet, and hard thresholding, to obtain a family of models indexed by regularization parameter λ. The tuning parameter λ was chosen by the BIC of (4.5). In addition, we used an L_1 regression for the SCAD penalty, denoting the method by SCAD* in Table 5. For each method, we fit a model with the training data, and then used it to predict the test data outcomes. The dataset was standardized to zero mean and unit variance across genes, to ensure that different features were comparable.

From Table 5, we can see that RSIS plus SCAD* selected 14 important features and achieved two test errors, whereas SIS plus SCAD* obtained 18

important features and made two test errors. The AEnet obtained fewer variables and performed more stably than did the LASSO. Table 5 suggests that the most parsimonious model was obtained by RSIS plus SCAD*.

6. Concluding Remarks

We have investigated nonconcave penalized M-estimation for relatively high dimensional models and shown that this type of estimation has the so-called "Oracle property." In our numerical studies, nonconcave penalized M-estimation lost little efficiency in comparison with existing penalized least squares methods, and this type of estimation may be more robust than these methods. If outlying or influential observations cannot be cleaned easily, or when it is difficult to determine if the white noise in the model follows a heavy tail distribution, we recommend nonconcave penalized M-estimation.

To handle ultra-high dimension cases, we propose a Rank SIS (RSIS) to first reduce the model size to a relatively large scale, then employ the nonconcave penalized M-estimation to obtain the final model estimation. The RSIS is based on rank correlation and SIS. Compared to SIS, based on Pearson correlation, RSIS inherits the robustness property of rank correlation, as is supported in our numerical studies. Under situations that favor a combination of SIS and the LS-SCAD, our proposed RSIS+LAD based SCAD remains comparable, and is sometimes even better. However, as Fan and Lv (2008) point out, unpredictable situations occur more often with ultra-high dimensional data. Thus, it is difficult to say whether the robust methods or such classical methods as least squares are more efficient and reliable in these situations. The question deserves further study, but is beyond the scope of the current paper.

Acknowledgement

Gaorong Li's research was supported by Funding Project for Academic Human Resources Development in Institutes of Higher Learning Under the Jurisdiction of Beijing Municipality (PHR20110822), Training Programme Foundation for the Beijing Municipal Excellent Talents (2010D005015000002) and National Natural Science Foundation of China (11002005). Heng Peng's research were supported by CERG grants of Hong Kong Research Grant Council (HKBU 201707, HKBU 201809, and HKBU 201610), FRG grants from Hong Kong Baptist University (FRG/06-07/II-14 and FRG/08-09/II-33), and a grant from National Nature Science Foundation of China (NNSF 10871054). Lixing Zhu's research was supported by a grant from the Research Grants Council of Hong Kong. The authors would like to thank the Editor, an associate editor, and the referees for their helpful comments that helped to improve an earlier version of this article.

Appendix

We provide proofs of the results stated in Subsection 2.3.

Proof of Theorem 1. Let $\alpha_n = \sqrt{p_n}(n^{-1/2} + a_n)$ and $\|u\| = C$, where C is a sufficiently large constant. Our aim is to show that for any given ϵ there is a large constant C such that, for a large n, we have

$$P\left\{\inf_{\|u\|=C} Q_n(\beta_{n0} + \alpha_n u) > Q_n(\beta_{n0})\right\} \geq 1 - \epsilon. \tag{A.1}$$

This implies with probability of at least $1-\epsilon$ that there exists a local minimizer in the ball $\{\beta_{n0} + \alpha_n u : \|u\| \leq C\}$. Hence, there exists a local minimizer such that $\|\hat{\beta}_n - \beta_{n0}\| = O_P(\alpha_n)$.

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

$$\begin{aligned} D_n(u) &\hat{=} Q_n(\beta_{n0} + \alpha_n u) - Q_n(\beta_{n0}) \\ &\geq \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T(\beta_{n0} + \alpha_n u)) - \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \beta_{n0}) \\ &\quad + n\sum_{j=1}^{k_n} \{p_{\lambda_n}(|\beta_{n0j} + \alpha_n u_j|; a) - p_{\lambda_n}(|\beta_{n0j}|; a)\} \\ &\hat{=} I + II, \end{aligned} \tag{A.2}$$

where k_n is the number of components in β_{I0}, and

$$\begin{aligned} II &= \sum_{j=1}^{k_n} [n\alpha_n p'_{\lambda_n}(|\beta_{n0j}|)\mathrm{sgn}(\beta_{n0j})u_j + \frac{1}{2} n\alpha_n^2 p''_{\lambda_n}(|\beta_{n0j}|)u_j^2\{1 + o(1)\}] \\ &\leq \sum_{j=1}^{k_n} [|n\alpha_n p'_{\lambda_n}(|\beta_{n0j}|)\mathrm{sgn}(\beta_{n0j})u_j| + \frac{1}{2} n\alpha_n^2 p''_{\lambda_n}(|\beta_{n0j}|)u_j^2\{1 + o(1)\}] \\ &\leq \sqrt{k_n} n\alpha_n a_n \|u\| + \max_{1 \leq j \leq k_n} p''_{\lambda_n}(|\beta_{n0j}|) n\alpha_n^2 \|u\|^2 \\ &\leq n\alpha_n^2 \|u\| + n b_n \alpha_n^2 \|u\|^2. \end{aligned}$$

Next, we consider I.

$$\begin{aligned} I &= \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T(\beta_{n0} + \alpha_n u)) - \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \beta_{n0}) \\ &= \sum_{i=1}^{n} \int_0^{-\alpha_n \mathbf{x}_i^T u} [\psi(e_i + t) - \psi(e_i)] dt - \alpha_n \sum_{i=1}^{n} \psi(e_i) \mathbf{x}_i^T u \\ &\hat{=} I_1 + I_2. \end{aligned} \tag{A.3}$$

Because $|I_2| \le \alpha_n \|u\| \| \sum_{i=1}^n \psi(e_i)\mathbf{x}_i \|$ and, since it is easy to check that

$$E\left\| \sum_{i=1}^n \psi(e_i)\mathbf{x}_i \right\|^2 = E\left[\sum_{j=1}^{p_n} \sum_{i=1}^n \sum_{l=1}^n x_{ij}x_{lj}\psi(e_i)\psi(e_l) \right] = \sum_{j=1}^{p_n} \sum_{i=1}^n x_{ij}^2 E\psi^2(e_i) \le \sigma^2 n p_n, \tag{A.4}$$

we have $|I_2| \le O_P(\alpha_n\sqrt{np_n})\|u\| = O_P(\alpha_n^2 n)\|u\|$.

Invoking conditions (C1) and (C2), for I_1 we have

$$\begin{aligned} E(I_1) &= \sum_{i=1}^n \int_0^{-\alpha_n \mathrm{x}_i^T u} G(t)dt \\ &= \sum_{i=1}^n \int_0^{-\alpha_n \mathrm{x}_i^T u} \{\gamma t + o(|t|)\}dt \\ &= \frac{1}{2}\alpha_n^2 \gamma u^T S_n u + o_P(1)\frac{1}{2} n\alpha_n^2 \|u\|^2. \end{aligned} \tag{A.5}$$

Because $p_n \log n/n \to 0$ and $\sqrt{p_n \log n}a_n \to 0$ as $n \to \infty$, with condition (C4) we have $\max_{1\le i\le n} |\alpha_n \mathrm{x}_i^T u| \to 0$. With the Schwarz inequality and condition (C2), it is not difficult to show that

$$\begin{aligned} \mathrm{Var}(I_1) &\le \sum_{i=1}^n E\left\{ \int_0^{-\alpha_n \mathbf{x}_i^T u} [\psi(e_i + t) - \psi(e_i)]dt \right\}^2 \\ &\le \sum_{i=1}^n |\alpha_n \mathbf{x}_i^T u| \cdot \left| \int_0^{-\alpha_n \mathbf{x}_i^T u} E[\psi(e_i + t) - \psi(e_i)]^2 dt \right| \\ &= o(1) \cdot \sum_{i=1}^n (\alpha_n \mathbf{x}_i^T u)^2 \to o_p(p_n). \end{aligned} \tag{A.6}$$

From (A.5) and (A.6), I_1 dominates all of the items uniformly in $\|u\| = C$ when a sufficiently large C is chosen. As I_1 is positive, this completes the proof of Theorem 1.

Lemma 1. *Under the conditions of Theorem 1, if* $\lambda_n \to 0$ *and* $\sqrt{n/p_n}\lambda_n \to \infty$ *as* $n \to \infty$, *then the nonconcave penalized M-estimator* $\hat{\beta}_n = (\hat{\beta}_I^T, \hat{\beta}_{II}^T)^T$ *satisfies* $\hat{\beta}_{II}^T = 0$ *with probability tending to* 1.

Proof. From Theorem 1 for a sufficiently large C, $\hat{\beta}_n$ lies in the ball $\{\beta_{n0} + \alpha_n u : \|u\| \le C\}$ with probability converging to 1, where $\alpha_n = \sqrt{p_n}(n^{-1/2} + a_n)$. Taking

the first derivative of $Q_n(\beta_n)$ at any differentiable point $\beta_n = (\beta_{n1}, \ldots, \beta_{np_n})^T$ with respect to $\beta_{nj}, j = k_n + 1, \ldots, p_n$, we have

$$\begin{aligned}\frac{\partial Q_n(\beta_n)}{\partial \beta_{nj}} &= -\sum_{i=1}^{n} \psi(y_i - \mathbf{x}_i^T \beta_n) x_{ij} + n p'_{\lambda_n}(|\beta_{nj}|)\mathrm{sgn}(\beta_{nj}) \\ &= -\sum_{i=1}^{n} \psi(e_i - \mathbf{x}_i^T(\beta_n - \beta_{n0})) x_{ij} + n p'_{\lambda_n}(|\beta_{nj}|)\mathrm{sgn}(\beta_{nj}). \quad \text{(A.7)}\end{aligned}$$

For any $u \in \mathbb{R}^{p_n}$, let

$$\Phi_n(u) = \sum_{i=1}^{n} \psi(e_i - \alpha_n \mathbf{x}_i^T u)\mathbf{x}_i. \quad \text{(A.8)}$$

Note that $E(\Phi_n(\mathbf{0})) = 0$ and $\mathrm{Var}(\Phi_n(\mathbf{0})) = \sigma^2 S_n$. As $\max_{1 \le i \le n} \mathbf{x}_i^T S_n^{-1} \mathbf{x}_i \to 0$, the Lindeberg Theorem yields $\Phi_n(\mathbf{0}) \xrightarrow{L} N(0, \sigma^2 S_n)$. Note that $\max_{1 \le i \le n} |\alpha_n \mathrm{x}_i^T u| \to 0$ and, from the argument of Lemma 3.4 in Bai and Wu (1994), one has

$$\sup_{\|u\| \le C} |\Phi_n(u) - \Phi_n(\mathbf{0}) + \gamma \alpha_n S_n u| = o_P(1). \quad \text{(A.9)}$$

Invoking Theorem 1, for any $\beta_n = (\beta_I^T, \beta_{II}^T)^T$ that satisfies $\beta_I - \beta_{I0} = O_P(\sqrt{p_n/n})$ and $|\beta_{II} - \beta_{II0}| \le \epsilon_n = C(\sqrt{p_n/n})$,

$$\sum_{i=1}^{n} \psi(e_i - \mathrm{x}_i^T(\beta_n - \beta_{n0}))\mathrm{x}_i - \sum_{i=1}^{n} \psi(e_i)\mathrm{x}_i + \gamma S_n(\beta_n - \beta_{n0}) = o_P(1). \quad \text{(A.10)}$$

From (A.4), (A.10), and condition (C3), we have

$$\sum_{i=1}^{n} \psi(y_i - \mathbf{x}_i^T \beta_n)\mathrm{x}_i = O_P(\sqrt{np_n}). \quad \text{(A.11)}$$

Using $\sqrt{p_n/n}/\lambda_n \to 0$ and (C6),

$$\frac{\partial Q_n(\beta_n)}{\partial \beta_{nj}} = n\lambda_n \left\{ -O_P\left(\frac{\sqrt{p_n/n}}{\lambda_n}\right) + \frac{p'_{\lambda_n}(|\beta_{nj}|)}{\lambda_n}\mathrm{sgn}(\beta_{nj}) \right\}. \quad \text{(A.12)}$$

Obviously the sign of β_{nj} determines the sign of $\partial Q_n(\beta_n)/\partial \beta_{nj}$. Hence, (A.12) implies that

$$\frac{\partial Q_n(\beta_n)}{\partial \beta_{nj}} = \begin{cases} > 0, \text{ for } 0 < \beta_{nj} < \epsilon_n \\ < 0, \text{ for } -\epsilon_n < \beta_{nj} < 0, \end{cases}$$

where $j = k_n + 1, \ldots, p_n$. This completes the proof of Lemma 1.

Proof of Theorem 2. Sparsity (i) follows from Lemma 1. Thus we need only prove (ii). As shown in Theorem 1, $\hat{\beta}_n$ is root-(n/p_n) consistent. By Lemma 1, each component of $\hat{\beta}_I$ stays away from zero for a sufficiently large sample size n. At the same time, $\hat{\beta}_{II} = \mathbf{0}_{m_n}$ with probability tending to 1. Thus, with probability tending to 1, the partial derivatives exist for the first k_n components. That is, $\hat{\beta}_I$ satisfies

$$-\sum_{i=1}^{n} \mathbf{x}_{1i}\psi(y_i - \mathbf{x}_{1i}^T\hat{\beta}_I) + nP'_{\lambda_n}(|\hat{\beta}_I|) = 0, \tag{A.13}$$

where $P'_{\lambda_n}(|\hat{\beta}_I|)$ is a $k_n \times 1$ vector whose jth element is $p'_{\lambda_n}(|\hat{\beta}_{nj}|)\mathrm{sgn}(\hat{\beta}_{nj})$. Applying a Taylor expansion to (A.13), we have

$$\{\gamma S_{1n} + n\Sigma_{\lambda_n}\}(\hat{\beta}_I - \beta_{I0}) + n\mathbf{b}_n \hat{=} w_1 - w_2 + \frac{1}{2}w_3, \tag{A.14}$$

where $w_1 = \sum_{i=1}^n \psi(e_i)\mathbf{x}_{1i}, w_2 = \sum_{i=1}^n (\psi'(e_i) - \gamma)\mathbf{x}_{1i}\mathbf{x}_{1i}^T(\hat{\beta}_I - \beta_{I0})$, and $w_3 = \sum_{i=1}^n \psi''(e_i - \mathbf{x}_{1i}^T\beta_n^*)[\mathbf{x}_{1i}^T(\hat{\beta}_I - \beta_{I0})]^2\mathbf{x}_{1i}$. Here, β_n^* is a vector between 0 and $\hat{\beta}_I - \beta_{I0}$. Multiply the two sides of (A.14) by $A_n S_{1n}^{-1/2}$ to obtain

$$A_n S_{1n}^{-1/2}\{\gamma S_{1n} + n\Sigma_{\lambda_n}\}[(\hat{\beta}_I - \beta_{I0}) + n\{\gamma S_{1n} + n\Sigma_{\lambda_n}\}^{-1}\mathbf{b}_n] \hat{=} W_1 - W_2 + \frac{1}{2}W_3, \tag{A.15}$$

where $W_1 = A_n S_{1n}^{-1/2} w_1, W_2 = A_n S_{1n}^{-1/2} w_2$, and $W_3 = A_n S_{1n}^{-1/2} w_3$. Hence, to prove Theorem 2, it suffices to show that W_1 satisfies the conditions of the Lindeberg-Feller Central Limit Theorem and $W_i = o_P(1)$ $(i = 2, 3)$. Invoking Theorem 1 and Lemma 3 of Mammen (1989), (C3), and the Cauchy-Schwarz inequality, we have

$$|W_2| \le \rho_{\max}^{1/2}(A_n A_n^T)\rho_1^{-1/2}(S_{1n})o_P(1)\|\hat{\beta}_I - \beta_{I0}\| = o_P\left(\frac{\sqrt{p_n}}{n}\right) = o_P(1). \tag{A.16}$$

Using $\|W_3\|^2 = \mathrm{tr}(W_3 W_3^T)$, (C4), and $p_n \log n/n \to 0$, we have

$$E\|W_3\|^2 \le (\frac{B}{b})E(\psi''(e_i))^2 \max_{1\le i\le n} |\mathbf{x}_{1i}^T(\hat{\beta}_I - \beta_{I0})|^4 = O_P\left(\frac{p_n^2 \log^2 n}{n^2}\right) = o_P(1). \tag{A.17}$$

From (A.15)–(A.17), we obtain

$$A_n S_{1n}^{-1/2}\{\gamma S_{1n} + n\Sigma_{\lambda_n}\}[(\hat{\beta}_I - \beta_{I0}) + n\{\gamma S_{1n} + n\Sigma_{\lambda_n}\}^{-1}\mathbf{b}_n] = W_1 + o_P(1). \tag{A.18}$$

Next, we verify that the conditions of the Lindeberg-Feller Central Limit Theorem are satisfied by W_1. Let $\omega_{ni} = A_n S_{1n}^{-1/2}\psi(e_i)\mathbf{x}_{1i}$, $i = 1, \ldots, n$. Note first

that $E(\omega_{ni}) = \mathbf{0}$ and

$$\mathrm{Var}\left(\sum_{i=1}^{n}\omega_{ni}\right) = \sigma^2 A_n S_{1n}^{-1}\sum_{i=1}^{n}\mathbf{x}_{1i}\mathbf{x}_{1i}^T A_n^T \to \sigma^2 G \tag{A.19}$$

as $A_n A_n^T \to G$. For any $\varepsilon > 0$,

$$\begin{aligned}\sum_{i=1}^{n} E[\|\omega_{ni}\|^2\mathbf{1}\{\|\omega_{ni}\| > \varepsilon\}] &= nE\|\omega_{ni}\|^2\mathbf{1}\{\|\omega_{ni}\| > \varepsilon\}\\ &\le n\{E\|\omega_{ni}\|^4\}^{1/2}\{P(\|\omega_{ni}\| > \varepsilon)\}^{1/2}. \end{aligned}\tag{A.20}$$

By (C5) and $A_n A_n^T = G$, we have

$$P(\|\omega_{ni}\| > \varepsilon) \le \frac{E\|\omega_{ni}\|^2}{\varepsilon^2} \le \frac{\sigma^2\rho_{\max}(A_n A_n^T)d_n^2}{\varepsilon^2} = O(n^{-1}) \tag{A.21}$$

and, similar to the proof of Theorem 6 in Huang and Xie (2007), we have

$$\begin{aligned}E\{\|\omega_{ni}\|^4\} &= E[\omega_{ni}^T\omega_{ni}]^2\\ &\le \sigma_4\rho_{\max}^2(A_n A_n^T)\rho_1^{-2}(S_{1n})E\Big[\sum_{j=1}^{k_n}x_{1ij}^2\Big]^2\\ &= O\left(\frac{k_n^2}{n^2}\right),\end{aligned}\tag{A.22}$$

where x_{1ij} is the jth component of $\mathbf{x}_{1i}$. Then by (A.20)–(A.22), we have

$$\sum_{i=1}^{n} E[\|\omega_{ni}\|^2\mathbf{1}\{\|\omega_{ni}\| > \varepsilon\}] = O\left(n\frac{p_n}{n}\frac{1}{\sqrt{n}}\right) = o(1). \tag{A.23}$$

From the foregoing argument, and invoking the Lindeberg-Feller Central Limit Theorem, we complete the proof of (ii).

References

Agostinelli, C. (2002). Robust model selection in regression via weighted likelihood methodology. *Statist. Probab. Lett.* **56**, 289-300.

Antoniadis, A. (1997). Wavelets in statistics: A review (with discussion). *J. Italian Statist. Assoc.* **6**, 97-144.

Antoniadis, A. and Fan, J. (2001). Regularization of wavelets approximations (with discussion). *J. Amer. Statist. Assoc.* **96**, 939-967.

Bai, Z. D., Rao, C. R. and Wu, Y. (1992). M-estimation of multivariate linear regression parameters under a convex discrepancy function. *Statist. Sinica* **2**, 237-254.

Bai, Z. D. and Wu, Y. (1994). Limiting behavior of M-estimators of regression coefficients in high dimensional linear models I. scale-dependent case. *J. Multivariate Anal.* **51**, 211-239.

Candés, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. (with discussion). *Ann. Statist.* **35**, 2313-2351.

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**, 337-403.

Fan, J. Q. (1997). Comment on "Wavelets in statistics: a review" by A. Antoniadis. *J. Italian Statist. Assoc.* **6**, 131-138.

Fan, J. Q. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J. Q. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30**, 74-99.

Fan, J. Q. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *J. Amer. Statist. Assoc.* **99**, 710-723.

Fan, J. Q. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. (with discussion). *J. Roy. Statist. Soc. Ser. B* **70**, 849-911.

Fan, J. Q. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928-961.

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

He, X. M. and Shao, Q. M. (2000). On parameters of increasing dimensions. *J. Multivariate Anal.* **73**, 120-135.

Huang, J. and Xie, H. (2007). Asymptotic oracle properties of SCAD-penalized least squares estimators. *Inst. Math. Statist.* **55**, 149-166.

Huang, J., Horowitz, J. L. and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36**, 587-613.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73-101.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1**, 799-821.

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

Hunter, D. and Li, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33**, 1617-1642.

Khan, J. A., Van Aelst, S. and Zamar, R. H. (2007). Robust linear model selection based on least angle regression. *J. Amer. Statist. Assoc.* **102**, 1289-1299.

Li, R. and Liang, H. (2008). Variable selection in semiparametric regression modeling. *Ann. Statist.* **36**, 261-286.

Mammen, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *Ann. Statist.* **17**, 382-400.

Müller, S. and Welsh, A. H. (2005). Outlier robust model selection in linear regression. *J. Amer. Statist. Assoc.* **100**, 1297-1310.

Portnoy, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Statist.* **12**, 1298-1309.

Portnoy, S. (1985). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Ann. Statist.* **13**, 1403-1417. Correction. *Ann. Statist.* **19**, 2282.

Ronchetti, E. (1985). Robust model selection in regression. *Statist. Probab. Lett.* **3**, 21-23.

Ronchetti, E. (1997). Robustness aspects of model choice. *Statist. Sinica* **7**, 327-338.

Ronchetti, E., Field, C. and Blanchard, W. (1997). Robust linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **92**, 1017-1023.

Ronchetti, E. and Staudte, R. G. (1994). A robust version of Mallow's C_P. *J. Amer. Statist. Assoc.* **89**, 550-559.

Salibian-Barrera, M. and Van Aelst, S. (2008). Robust model selection using fast and robust bootstrap. *Comput. Statist. Data Anal.* **52**, 5121-5135.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

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

Welsh, A. H. (1989). On M-processes and M-estimation. *Ann. Statist.* **17**, 337-361.

Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K. and Zhao, H. (2003). Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* **19**, 1636-1643.

Wu, W. B. (2007). M-estimation of linear models with dependent errors. *Ann. Statist.* **35**, 495-521.

Wu, Y. and Zen, M. M. (1999). A strong consistent information criterion for linear model selection based on M-estimation. *Probab. Theory Related Fields* **113**, 599-625.

Yohai, V. J. and Maronna, R. A. (1979). Asymptotic behavior of M-estimators for the linear model. *Ann. Statist.* **7**, 258-268.

Yu, J. S., Ongarello, S., Fiedler, R., Chen, X. W., Toffolo, G., Cobelli, C. and Trajanoski, Z. (2005). Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. *Bioinformatics* **21**, 2200-2209.

Zhang, C. H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* **36**, 1567-1594.

Zheng, G., Freidlin, B. and Gastwirth, J. L. (2004). Using Kullback-Leibler information for model selection when the data-generating model is unknown: Applications to genetic testing problems. *Statist. Sinica* **14**, 1021-1036.

Zhong, W. X., Zeng, P., Ma, P., Liu, J. S. and Zhu, Y. (2005). RSIR: regularized sliced inverse regression for motif discovery. *Bioinformatics* **21**, 4169-4175.

Zhou, J. and He, X. M. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *Ann. Statist.* **36**, 1649-1668.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. B* **67**, 301-320.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. (with discussion). *Ann. Statist.* **36**, 1509-1533.

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.* **37**, 1733-1751.

College of Applied Sciences, Beijing University of Technology, Beijing 100124, P. R. China.
E-mail: ligaorong@gmail.com

Department of Mathematics, Hong Kong Baptist University, Hong Kong, China.
E-mail: hpeng@math.hkbu.edu.hk

Department of Mathematics, Hong Kong Baptist University, Hong Kong, China.
E-mail: lzhu@hkbu.edu.hk

(Received September 2008; accepted August 2009)