Smoothly Clipped Absolute Deviation on High Dimensions

Yongdai KIM, Hosik CHOI, and Hee-Seok OH

The smoothly clipped absolute deviation (SCAD) estimator, proposed by Fan and Li, has many desirable properties, including continuity, sparsity, and unbiasedness. The SCAD estimator also has the (asymptotically) oracle property when the dimension of covariates is fixed or diverges more slowly than the sample size. In this article we study the SCAD estimator in high-dimensional settings where the dimension of covariates can be much larger than the sample size. First, we develop an efficient optimization algorithm that is fast and always converges to a local minimum. Second, we prove that the SCAD estimator still has the oracle property on high-dimensional problems. We perform numerical studies to compare the SCAD estimator with the LASSO and SIS–SCAD estimators in terms of prediction accuracy and variable selectivity when the true model is sparse. Through the simulation, we show that the variance estimator of Fan and Li still works well for some limited high-dimensional cases where the true nonzero coefficients are not too small and the sample size is moderately large. We apply the proposed algorithm to analyze a high-dimensional microarray data set.

KEY WORDS: High dimension; Oracle property; Regression; Regularization; Smoothly clipped absoluetly deviation penalty.

1. INTRODUCTION

Variable selection is a fundamental task for high-dimensional statistical modeling. Traditional approaches follow stepwise and subset selection procedures, which are computationally intensive, unstable, and difficult to draw sampling properties from (see, e.g., Breiman 1996). Alternative variable selection methods are sparse penalized approaches, including bridge regression (Frank and Friedman 1993), least absolute shrinkage and selection operator (LASSO; Tibshirani 1996), and the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001). Among these, the SCAD estimator has all of the desirable properties, including unbiasedness, sparsity, and continuity.

The SCAD estimator has an oracle property when the dimension of predictive variables are not large compared with the sample size (Fan and Li 2001; Fan and Peng 2004). Here the oracle property means that the SCAD estimator is asymptotically equivalent to the oracle estimator, which is obtained by deleting all irrelevant predictive variables (i.e., variables whose true regression coefficients are 0) in advance. But such results are not applicable to the case where the dimension of covariates is larger than the sample size, which is encountered in many situations, including microarray data analysis. For LASSO, Meinshausen and Bühlmann (2006) and Zhao and Yu (2006) recently proved the model selection consistency in high-dimensional cases; that is, one can select only the relevant covariates by choosing the regularization parameter appropriately in LASSO. They also showed that the regularization parameter for model selection consistency is not optimal for prediction accuracy, however.

In this article we study the SCAD estimator in high-dimensional cases where the dimension of covariates is much larger than the sample size. First, we propose an efficient computational algorithm, that is a coupling of the concave convex procedure (CCCP) (An and Tao 1997; Yuille and Rangarajan 2003) with the LASSO algorithm. Advantages of the proposed algorithm over the algorithm of Fan and Li (2001) are that it is less sensitive to the initial solution, faster, more stable, and always guaranteed to converge to a local minimum. Second, we prove the oracle property in the case where the dimension of covariates is allowed to grow at a certain polynomial rate that depends on the moment condition of the noise provided that the true model is sparse. Moreover, for Gaussian noise, we show that the dimension of covariates can grow exponentially fast. This result implies that the SCAD can achieve model selection consistency and optimal prediction simultaneously in high-dimensional cases, which is impossible for LASSO (Meinshausen and Bühlmann 2006; Zou 2006).

The article is organized as follows. In Section 2 we briefly review the SCAD penalty. In Section 3 we describe the proposed computational algorithm, and in Section 4 we investigate the oracle property. In Section 5 we compare the performance of the proposed method with the SIS–SCAD (Fan and Lv 2008) and LASSO estimators by simulation studies as well as a real data analysis. We provide concluding remarks in Section 6.

2. REVIEW OF SCAD

Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ be n predictive-response variable pairs that are assumed to be a random sample where $\mathbf{x}_i \in R^p$ and $y_i \in R$. We consider estimating the regression coefficient $\boldsymbol{\beta}$ by minimizing the penalized least square (PLS),

$$C(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \sum_{j=1}^{p} J_\lambda(|\beta_j|). \tag{1}$$

for a penalty function $J_\lambda(\cdot)$. $J_\lambda(|\beta_j|) = \lambda|\beta_j|^\gamma$, $\gamma \geq 0$ leads to the bridge estimator (Frank and Friedman 1993), in particular, when $\gamma = 1$, the penalty yields the LASSO estimator (Tibshirani 1996). Fan and Li (2001) suggested the SCAD estimator, which corresponds to the SCAD penalty as

$$\begin{aligned} J_\lambda(|\beta|) = {} & \lambda|\beta| I(0 \leq |\beta| < \lambda) \\ & + \left(\frac{a\lambda(|\beta| - \lambda) - (|\beta|^2 - \lambda^2)/2}{(a-1)} + \lambda^2\right) \\ & \times I(\lambda \leq |\beta| \leq a\lambda) \\ & + \left(\frac{(a-1)\lambda^2}{2} + \lambda^2\right) I(|\beta| \geq a\lambda). \end{aligned}$$

Yongdai Kim is Associate Professor (E-mail: *ydkim0903@gmail.com*) and Hee-Seok Oh is Associate Professor (E-mail: *heeseok@stats.snu.ac.kr*), Department of Statistics, Seoul National University, Seoul, Korea. Hosik Choi is Full-Time Lecturer, Department of Informational Statistics, Hoseo University, Asan, Chungnam, Korea (E-mail: *choi.hosik@gmail.com*). This work was supported by the Korea Research Foundation grants funded by the Korean Government (KRF-2005-070-C00021) and (KRF-2008-314-C00046).

Journal of the American Statistical Association
December 2008, Vol. 103, No. 484, Theory and Methods
DOI 10.1198/016214508000001066

If $J_\lambda(|\beta|) = \lambda^2 - (|\beta| - \lambda)^2 I(|\beta| < \lambda)$, then we obtain the hard thresholding estimator (Fan and Li 2001).

Antoniadis and Fan (2001) and Fan and Li (2001) discussed three desirable properties of penalized estimators: unbiasedness, sparsity, and continuity. For the bridge estimator, with $\gamma > 1$, only the continuity holds; with $\gamma = 1$, the sparsity and continuity hold; and with $\gamma < 1$, the unbiasedness and sparsity hold (see, e.g., Knight and Fu 2000). For the hard thresholding estimator, the sparsity and unbiasedness hold. In contrast, the SCAD estimator satisfies all three properties (Fan and Li 2001).

More recently, Zou (2006) proposed an adaptive LASSO estimator that also satisfies the three desirable properties. But this procedure requires an initial $\sqrt{n}$ consistent estimator, which is difficult to construct in high-dimensional cases. Huang, Ma, and Zhang (2006) proposed a sufficient condition for constructing an initial $\sqrt{n}$-consistent estimator for high-dimensional cases, but their sufficient condition is difficult to check in practice. The SCAD estimator does not require any initial $\sqrt{n}$ consistent estimator, and, as we prove later, it still has the oracle property in high-dimensional cases.

3. OPTIMIZATION ALGORITHM

In this section we propose an efficient computational algorithm for finding a local minimum of the PLS (1) with the SCAD penalty. The main idea of the proposed algorithm is to decompose the PLS (1) as the sum of the convex and concave functions. The CCCP algorithm (An and Tao 1997; Yuille and Rangarajan 2003), one of the powerful optimization algorithms for nonconvex problems, is then applied. The CCCP algorithm has been used in many learning problems, including those of Shen, Tseng, Zhang, and Wong (2003) and Collobert, Sinz, Weston, and Bottou (2006). The key idea of the CCCP algorithm is to update the solution with the minimizer of the tight convex upper bound of the objective function obtained at the current solution. To be specific, suppose that there are convex and concave functions, $C_{vex}(\boldsymbol{\beta})$ and $C_{cav}(\boldsymbol{\beta})$, such that $C(\boldsymbol{\beta}) = C_{vex}(\boldsymbol{\beta}) + C_{cav}(\boldsymbol{\beta})$. Given a current solution $\boldsymbol{\beta}^c$, the tight convex upper bound is defined by $Q(\boldsymbol{\beta}) = C_{vex}(\boldsymbol{\beta}) + \nabla C_{cav}(\boldsymbol{\beta}^c)'\boldsymbol{\beta}$, where $\nabla C_{cav}(\boldsymbol{\beta}) = \partial C_{cav}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$. We then update the solution by the minimizer of $Q(\boldsymbol{\beta})$. Note that $Q(\boldsymbol{\beta})$ is a convex function, so it can be easily minimized. An important property of the CCCP algorithm is that after each iteration, the objective function always decreases. Thus eventually, the solution converges to a local minimum. We also note that the SCAD penalty can be decomposed by the sum of the concave and convex functions,

$$J_\lambda(|\beta_j|) = \tilde{J}_\lambda(|\beta_j|) + \lambda|\beta_j|,$$

where $\tilde{J}_\lambda(|\beta_j|)$ is a differentiable concave function and $|\beta_j|$ is a convex function (see Fig. 1). Thus the PLS (1) with the SCAD penalty can be rewritten as

$$C(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \sum_{j=1}^{p}\tilde{J}_\lambda(|\beta_j|) + \lambda\sum_{j=1}^{p}|\beta_j|, \quad (2)$$

which is the sum of convex and concave functions. We apply the CCCP algorithm as follows. Given a current solution $\boldsymbol{\beta}^c$, the tight convex upper bound is given as

$$Q(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \sum_{j=1}^{p}\nabla\tilde{J}_\lambda(|\beta_j^c|)\beta_j + \lambda\sum_{j=1}^{p}|\beta_j|. \quad (3)$$

We then update the current solution of $\boldsymbol{\beta}^c$ with the minimizer of (3). To minimize (3), we use the algorithm of Rosset and Zhu (2007), because $Q(\boldsymbol{\beta})$ is a piecewise quadratic function. The foregoing derivation leads to the following iterative algorithm

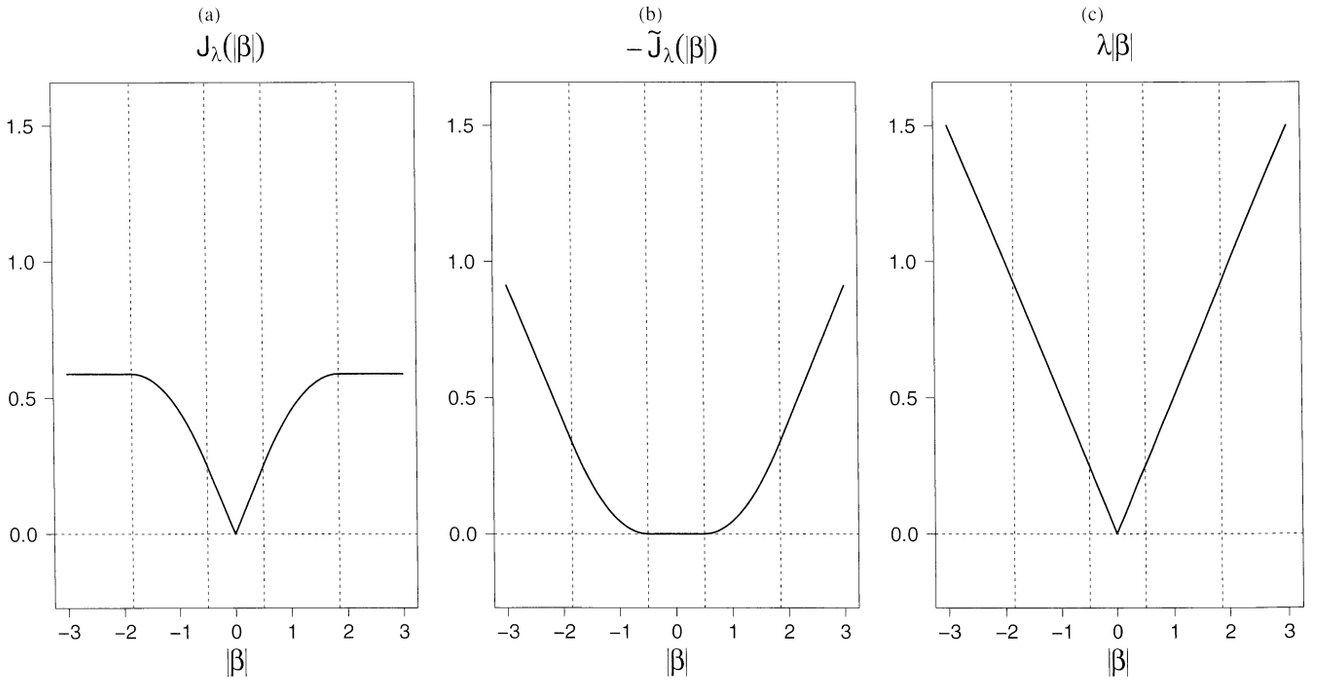

Figure 1. SCAD penalty when $a = 3.7$ and $\lambda = .5$.

Table 1. Comparison of the CCCP–SCAD and Fan and Li algorithms

	Algorithm	$C(\boldsymbol{\beta})$	CPU time (seconds)
Case 1	Fan and Li	$.505_{(.004)}$	$77.456_{(2.625)}$
	CCCP–SCAD	$.507_{(.004)}$	$84.650_{(5.159)}$
Case 2	Fan and Li	$.553_{(.004)}$	$47.765_{(1.653)}$
	CCCP–SCAD	$.507_{(.004)}$	$86.840_{(5.625)}$

NOTE: The numbers in the parentheses are standard errors.

for high-dimensional SCAD, which we call the CCCP–SCAD algorithm.

The CCCP–SCAD algorithm

- Initialization: Let $\boldsymbol{\beta}^c = 0$.
- Do until convergence:
  1. Find the minimizer $\boldsymbol{\beta}$ of $Q(\boldsymbol{\beta})$ in (3).
  2. $\boldsymbol{\beta}^c = \boldsymbol{\beta}$.

To compare the CCCP–SCAD algorithm with the algorithm of Fan and Li (2001), we performed a small simulation. Simulated data with size $n = 100$ were generated from model (4) in Section 5.1, with $p = 2{,}000$, $q = 5$, and $r = .3$. For the initial solution, we considered two cases. In case 1, we used the LASSO estimate, and in case 2, we used the estimate obtained by the forward selection. Then we calculated $C(\boldsymbol{\beta})$ and CPU time over 100 simulations. The results are summarized in Table 1. When we started with the initial LASSO estimate, the CCCP–SCAD and Fan and Li's algorithms were competitive; however, when we started with the initial estimate from the forward selection, the CCCP–SCAD algorithm gave smaller values of $C(\beta)$. This is in part because when a coefficient is 0 in the initial solution, it stays 0 forever in the algorithm of Fan and Li. In contrast, the coefficients that are initially 0 can be nonzero in the CCCP–SCAD algorithm; that is, the CCCP–SCAD algorithm is less sensitive to the choice of initial solutions.

Also, we can see from Table 1 that computing time of the Fan and Li algorithm depends heavily on the initial solution, whereas the CCCP–SCAD algorithm does not. This is because the Fan and Li's algorithm updates only nonzero coefficients in the initial solution. On the other hand, the CCCP–SCAD algorithm treats all coefficients equally, and so computing time is not seriously affected by the initial solution.

4. THE ORACLE PROPERTY

In this section we study asymptotic properties of the SCAD estimator. First, we prove that the oracle estimator becomes a local minimum of (1) with the SCAD penalty asymptotically, which is known as to be the oracle property (Fan and Li 2001). In particular, we consider the case where $p = O(n^\alpha)$ for $\alpha > 0$, whereas Fan and Li (2001) and Fan and Peng (2004) considered the case where $p = o(n)$. This result implies that we can find a good estimator among the local minima of (1) with the SCAD penalty; however, we do not know which local minimum is a good estimator. To clarify this problem, we give sufficient conditions under which the SCAD estimator—the global minimum of (1) with the SCAD penalty—is asymptotically equivalent to the oracle estimator. Finally, we discuss the choice of the initial solution.

Assume that the data are generated from a linear regression model, $Y_n = \mathbf{X}_n\boldsymbol{\beta}^* + \boldsymbol{\epsilon}_n$, where $\boldsymbol{\epsilon}_n = (\epsilon_1, \ldots, \epsilon_n)'$ is a vector of iid random variables with mean 0 and variance σ^2, $Y_n = (y_1, \ldots, y_n)'$, and $\mathbf{X}_n = (X_n^1, \ldots, X_n^p)$ with $X_n^j = (x_{j1}, \ldots, x_{jn})'$, $j = 1, \ldots, p$. For the sparse model, we consider a situation where most of the regression coefficients $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)'$ are exactly 0. Without loss of generality, we assume that the first q regression coefficients are nonzero and the remaining $p - q$ regression coefficients are 0. We let $\mathbf{X}_n = (\mathbf{X}_n^{(1)}, \mathbf{X}_n^{(2)})$, where $\mathbf{X}_n^{(1)}$ is the first $n \times q$ submatrix and $\mathbf{X}_n^{(2)}$ is the last $n \times (p - q)$ submatrix of $\mathbf{X}_n$. Similarly, we write $\boldsymbol{\beta}^* = (\boldsymbol{\beta}^{(1)*}, \boldsymbol{\beta}^{(2)*})$. Let $\mathbf{C}_n = \mathbf{X}_n'\mathbf{X}_n/n$ and $\mathbf{C}_n^{(i,j)} = \mathbf{X}_n^{(i)'}\mathbf{X}_n^{(j)}/n$ for $i, j = 1, 2$. We express p_n and q_n to emphasize that p and q vary with n.

4.1 Asymptotic Properties of the Oracle Estimator

In this section we prove that the oracle estimator is asymptotically a local minimum of $C(\boldsymbol{\beta})$. Here the oracle estimator is defined by $\hat{\boldsymbol{\beta}}^o = (\hat{\boldsymbol{\beta}}^{(1)o}, \mathbf{0}^{(2)})$, where $\hat{\boldsymbol{\beta}}^{(1)o}$ is the minimizer of $\|Y_n - \mathbf{X}_n^{(1)}\boldsymbol{\beta}^{(1)}\|_2^2$, $\mathbf{0}^{(2)}$ is the $(p_n - q_n)$-dimensional 0 vector, and $\|\cdot\|_2$ is the standard Euclidean norm on R^n.

We assume the following regularity conditions:

A1. There exists a positive constant M_1 such that

$$\frac{1}{n}(X_n^j)'X_n^j \le M_1 \quad \text{for all } j = 1, \ldots, p_n \text{ and all } n.$$

A2. There exists a positive constant M_2 such that $\alpha'\mathbf{C}_n^{(1,1)} \times \alpha \ge M_2$ for all $\|\alpha\|_2^2 = 1$.

A3. $q_n = O(n^{c_1})$ for some $0 < c_1 < 1$.

A4. There exist positive constants c_2 and M_3 such that $c_1 < c_2 \le 1$ and

$$n^{(1-c_2)/2} \min_{j=1,\ldots,q_n} |\beta_j^*| \ge M_3.$$

The foregoing regularity conditions were used by Zhao and Yu (2006) to prove the model selection consistency of the LASSO estimator. The following theorem is the main result, the proof of which is deferred to the Appendix.

Theorem 1. Assume that $\mathrm{E}(\epsilon_i)^{2k} < \infty$ for an integer $k > 0$. Let $\mathcal{A}_n(\lambda)$ be the set of local minima of (1) with the SCAD penalty and a regularization parameter λ_n. Then

$$\Pr(\hat{\boldsymbol{\beta}}^o \in \mathcal{A}_n(\lambda_n)) \to 1$$

as $n \to \infty$ provided that $\lambda_n = o(n^{-(1-(c_2-c_1))/2})$ and $p_n/(\lambda_n \times \sqrt{n})^{2k} \to 0$.

Theorem 1 can include the most of the previous results, such as those of Fan and Li (2001) and Fan and Peng (2004). When ϵ_i has the all moments, the oracle property holds when $p_n = O(n^\alpha)$ for any $\alpha > 0$, as $\mathrm{E}(\epsilon_i)^{2k} < \infty$ for all $k > 0$. For the Gaussian noise, the following theorem proves that the oracle property holds when $p_n = O(\exp(c_3 n))$ for some $c_3 > 0$.

Theorem 2. Assume that the ϵ_i's are iid Gaussian random variables. Then

$$\Pr(\hat{\boldsymbol{\beta}}^o \in \mathcal{A}_n(\lambda_n)) \to 1$$

as $n \to \infty$, provided that $p_n = O(\exp(c_3 n))$ and $\lambda_n = O(n^{-(1-c_4)/2})$, where $0 < c_3 < c_4 < c_2 - c_1$.

4.2 Asymptotic Equivalence of the SCAD and Oracle Estimators

Here we provide sufficient conditions under which the SCAD estimator becomes the oracle estimator asymptotically. We assume the following condition together with A1–A4:

A5. Suppose that $p_n \le n$ and $\mathbf{C}_n$ is nonsingular, with the smallest eigenvalue $\rho > 0$ and the largest eigenvalue bounded by M.

Theorem 3. Let $\hat{\boldsymbol{\beta}}$ be the global minimum of (1) with the SCAD penalty and a regularization parameter, λ_n. Then, under A1–A5,

$$\Pr(\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^o) \to 1$$

as $n \to \infty$, provided that $\lambda_n = o(n^{-(1-(c_2-c_1))/2})$ and $p_n/(\lambda_n \times \sqrt{n})^{2k} \to 0$.

Remark. In the high-dimensional setting, where $p_n > n$, A5 cannot be satisfied. But if we find a subset $\mathcal{G}$ of $\{1, \ldots, p\}$ such that $\{1, \ldots, q\} \in \mathcal{G}$ and the design matrix $\mathbf{X}_{\mathcal{G}} = (X^{(j)}, j \in \mathcal{G})$ satisfies A5, then we can use $\mathbf{X}_{\mathcal{G}}$ to find the oracle estimator. We discuss how to find such a $\mathcal{G}$ using the LASSO estimate in the next section.

4.3 Choice of the Initial Estimator

The key question is how to find such a $\mathcal{G}$. Zhao and Yu (2006) proved that under the same assumptions as those for the high-dimensional SCAD, a $\mathcal{G}$ can be obtained using the LASSO estimate with the λ_n as in Theorem 1 when the original matrix satisfies the so-called *strong irrepresentable condition*. This condition assumes that there exists a $(p_n - q_n)$-dimensional positive constant vector $\boldsymbol{\eta}$ such that

$$\left|\mathbf{C}^{(2,1)}\left(\mathbf{C}^{(1,1)}\right)^{-1} \operatorname{sign}\left(\boldsymbol{\beta}^{(1)*}\right)\right| \le \mathbf{1} - \boldsymbol{\eta},$$

where $\mathbf{1}$ is a $(p_n - q_n)$-dimensional vector of 1's and the inequality holds element-wise; that is, if we let $\mathcal{G} = \{j : \hat{\beta}_j^{lasso} \ne 0\}$, then $\mathbf{X}_{\mathcal{G}}$ satisfies A5.

The CCCP–SCAD algorithm gives only a local minimum even if we start with the LASSO estimate. To investigate the performance of the LASSO estimate as an initial solution, we performed a toy simulation. A set of data with size $n = 100$ was generated from model (4); $p = 2{,}000$, $q = 5$, and $r = .3$ were used for the simulation. Over 100 sets of samples, we compared solutions based on randomly generated initial solutions with solutions obtained from the LASSO estimate as the initial solution. Based on the simulation results, the solution from the initial LASSO estimate always had the smallest PLS value among the PLS values of 100 solutions by initial random solutions. Thus the LASSO estimate is a good initial solution that leads to the global optimum with high probability.

Remark. One interesting feature of the CCCP–SCAD algorithm is that it gives the LASSO estimate after one iteration of the CCCP when we start with the zero estimate (all coefficients being 0) as the initial solution. Thus we recommend using the zero estimate as the initial solution for the CCCP–SCAD algorithm.

5. NUMERICAL STUDIES

In this section, we investigate the finite-sample performance of the SCAD estimator through simulation experiments, as well as real data analysis. In particular, we compare the SCAD estimator with the SIS–SCAD (Fan and Lv 2008) and LASSO estimators in terms of prediction accuracy and variable selectivity. The SIS–SCAD estimator is obtained through a two-stage procedure, first reducing the number of covariates using the marginal correlations between the covariates and response variable and then applying the SCAD penalty. We also study the validity of the covariance estimation of nonzero coefficients of the SCAD estimator for high-dimensional cases.

5.1 Prediction Accuracy and Variable Selectivity

The simulation model is

$$y = \sum_{k=1}^{p} \beta_k x_k + \epsilon, \tag{4}$$

where $\mathbf{x} = (x_1, \ldots, x_p)'$ is a multivariate Gaussian random vector with mean 0 and covariances of x_k and x_l, being $r^{|k-l|}$ for some $r \in [0, 1)$, and ϵ is a Gaussian random variable with mean 0 and variance 1. For the values of the β's, we consider two scenarios. In the first scenario, we set all of the β's except the first q β's to be 0 and $\beta_k = c/k$ for $k = 1, \ldots, q$ for some c, which is selected so that the signal-to-noise ratio is approximately 3. In the second scenario, we set $\beta_{rk} = c/k$ for $k = 1, \ldots, q$, where $r = p/q$, and set the other β's to be 0. Note that the signal covariates (covariates with coefficients not 0) are correlated for the first scenario when $r > 0$, whereas they are almost independent for the second scenario. When $r = 0$, the two scenarios are the same. The number of dimension is fixed as $p = 2{,}000$, and the sample size is set as $n = 100$. We then investigate the performance for $q = (5, 20, 200, 1{,}000)$ and $r = (.0, .3, .6)$. For $q = 1{,}000$, we consider only Scenario 2, because Scenarios 1 and 2 are similar. Also, for $q = 200$ and $q = 1{,}000$, we do not include the results of the oracle estimate, because it is not defined well.

Figure 2 shows the solution paths of the LASSO and SCAD estimates of the first five regression coefficients from the model (4) with $q = 5$ and $r = 0$ according to various values of λ's. The coefficient values of the SCAD estimates are larger than those of the LASSO estimates, which implies that the SCAD estimator is less biased than the LASSO estimator.

The results on prediction errors and variable selectivity are presented in Tables 2 and 3. The values are the averages based on 100 repetitions of the simulation. For the SCAD and SIS–SCAD, we use the zero estimate as an initial solution. For the SIS–SCAD, we select $n/\log n$ many covariates in the first stage, because it is known to perform well (Fan and Lv 2008). The regularization parameters are selected by the five-fold cross-validation. The prediction errors in Table 2 are measured on independent test samples of size 10,000. The variable selectivity in Table 3 is the frequencies of appearance of the first five signal variables (Signal variable) and other noisy variables (Others) in the models chosen in 100 repetitions when $q = 5$.

From Table 2, we can see that the SCAD outperforms the LASSO and SIS–SCAD in most cases when the true model is

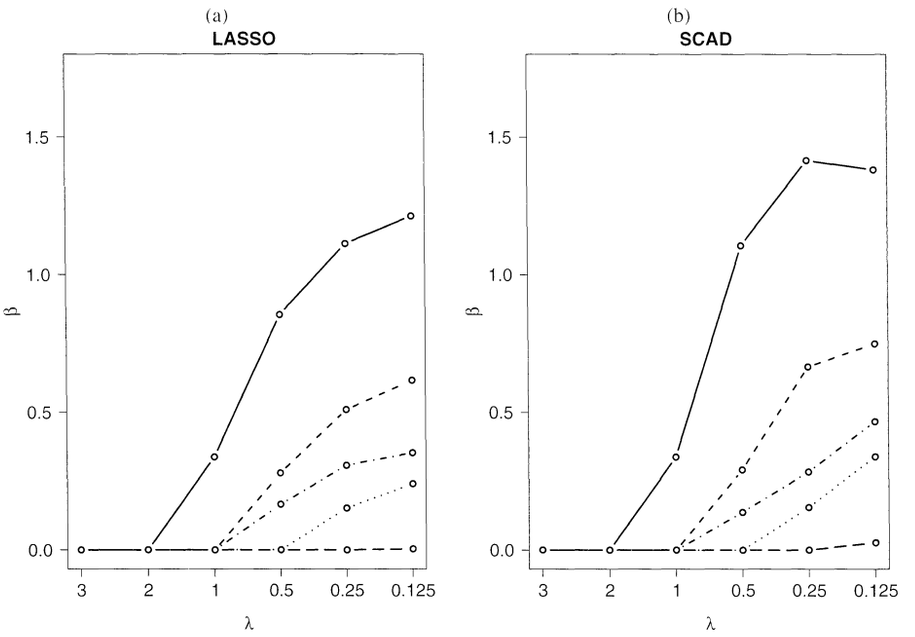

Figure 2. The solution paths of the LASSO (a) and SCAD (b) estimates of the first five regression coefficients from the model (4) with $q = 5$ and $r = 0$ (——, β_1; - - -, β_2; ·····, β_3; - · - ·, β_4; – –, β_5).

moderately sparse ($q = 200$) or nonsparse ($q = 1{,}000$). A reason for the inferior performance of the SIS–SCAD is that it deletes too many covariates in the first stage.

For the sparse true model (i.e., $q = 5$ and $q = 20$), the performance depends on the scenario and the value of r. When $r = 0$, the SCAD outperforms the LASSO and SIS–SCAD. When $r > 0$, the performance depends on the scenario. The SCAD is worst when $r = .6$ for Scenario 1, whereas it is the best for Scenario 2 regardless of the value of r. This can be explained in part as follows. First, sparse methods are not always good even when the true model is sparse. In particular, when signal variables are highly correlated (see Friedman and Popescu 2004; Zou and Hastie 2005), less sparse methods work better. Thus for Scenario 1 with $r = .6$, the LASSO is expected to outperform the SCAD. But even if r is large, the SCAD works better for Scenario 2, because the signal variables are not correlated. Second, the performance of the SIS–SCAD depends on the correlation of the signal variables, because it uses the marginal correlation between covariates and response variable. If the signal variables are highly correlated as in Scenario 1, then the likelihood that all signal variables are highly correlated with the response variable is high, and obviously all signal variables are

Table 2. Comparison of prediction errors

Scenario	r	q	Oracle	LASSO	SIS–SCAD	SCAD
1 and 2	0	5	$1.067_{(.006)}$	$1.628_{(.023)}$	$1.678_{(.037)}$	$1.425_{(.017)}$
		20	$1.277_{(.010)}$	$1.938_{(.026)}$	$1.968_{(.032)}$	$1.761_{(.021)}$
		200		$2.048_{(.029)}$	$2.377_{(.049)}$	$1.855_{(.023)}$
		1,000		$2.055_{(.027)}$	$2.507_{(.097)}$	$1.830_{(.021)}$
1	.3	5	$1.067_{(.005)}$	$1.378_{(.015)}$	$1.330_{(.017)}$	$1.341_{(.014)}$
		20	$1.267_{(.009)}$	$1.731_{(.023)}$	$1.752_{(.026)}$	$1.662_{(.018)}$
		200		$1.849_{(.022)}$	$2.142_{(.036)}$	$1.810_{(.018)}$
	.6	5	$1.066_{(.005)}$	$1.272_{(.014)}$	$1.257_{(.010)}$	$1.330_{(.014)}$
		20	$1.285_{(.011)}$	$1.520_{(.017)}$	$1.509_{(.015)}$	$1.550_{(.017)}$
		200		$1.683_{(.020)}$	$1.854_{(.026)}$	$1.724_{(.019)}$
2	.3	5	$1.062_{(.005)}$	$1.554_{(.020)}$	$1.596_{(.026)}$	$1.424_{(.013)}$
		20	$1.275_{(.012)}$	$1.759_{(.021)}$	$1.816_{(.027)}$	$1.623_{(.017)}$
		200		$1.786_{(.018)}$	$2.114_{(.039)}$	$1.648_{(.016)}$
		1,000		$1.826_{(.022)}$	$2.118_{(.037)}$	$1.706_{(.017)}$
	.6	5	$1.066_{(.006)}$	$1.496_{(.015)}$	$1.484_{(.020)}$	$1.383_{(.013)}$
		20	$1.288_{(.011)}$	$1.588_{(.016)}$	$1.597_{(.019)}$	$1.495_{(.014)}$
		200		$1.614_{(.019)}$	$1.892_{(.040)}$	$1.527_{(.020)}$
		1,000		$1.688_{(.021)}$	$1.861_{(.023)}$	$1.656_{(.020)}$

NOTE: Standard errors are in parentheses.

Table 3. Comparison of variable selectivity when $q = 5$

Scenario	r	Method	Signal variables					Others
1	0	LASSO	100	98	84	55	39	1.174
		SIS–SCAD	100	86	39	19	13	.485
		SCAD	100	98	87	58	39	1.078
	.3	LASSO	100	100	93	86	60	.959
		SIS–SCAD	100	100	91	62	34	.397
		SCAD	100	100	91	79	58	1.154
	.6	LASSO	100	97	94	84	53	.779
		SIS–SCAD	100	87	66	72	49	.363
		SCAD	100	73	74	69	43	1.185
2	.3	LASSO	100	93	66	45	29	1.087
		SIS–SCAD	100	78	30	15	12	.554
		SCAD	100	97	70	48	28	1.056
	.6	LASSO	100	87	41	23	10	.956
		SIS–SCAD	100	70	27	16	6	.567
		SCAD	100	90	49	28	10	.902

selected in the first stage with high probability. Thus the final estimator performs well. But if the correlations of the signal variables are not high as in Scenario 2, then some signal variables can be missed in the first stage, and so the performance of the final SIS–SCAD estimator can be poor. We note that the SIS–SCAD is even worse than the LASSO for Scenario 2. This is also conformed by Table 3, which shows that the SCAD selects the signal variables well for Scenario 2 regardless of the value of r but does not perform well for Scenario 1 with large value of r.

The simulation results suggest that the SCAD is a promising method for high-dimensional data when either the true model is not sparse or the signal variables are not strongly correlated. If the signal variables are strongly correlated and the true model is believed to be sparse, then LASSO or SIS–SCAD is recommended.

5.2 Estimation of Variance

We now investigate the performance of the variance estimator of the SCAD estimator for high-dimensional cases. We use model (4) with Scenario 1, $q = 5$, and $r = .3$. As was done by Fan and Li (2001), we consider the median absolute deviation divided by .6745 (denoted by SD) of 100 simulated coefficients in the 100 simulations as the true standard error. The median of the 100 estimated SDs, denoted by SD_m, and the median absolute deviation error of the 100 estimated standard errors divided by .6745, denoted by $\mathrm{SD}_{\mathrm{mad}}$, gauge the overall performance of the variance estimation. Table 4 presents the results for the first five coefficients when the sample sizes $n = 100$ and $n = 300$. When $n = 100$, the variances are seriously underestimated. The situation becomes better when $n = 300$; in that case the variance estimations of SIS–SCAD and SCAD perform well for the first two largest coefficients, β_1 and β_2, but they still underestimate the last three. The results for Scenario 2 are similar and thus are omitted. Based on these observations, we conclude that the variance estimation proposed by Fan and Li (2001) is valid only when the sample size is moderately large and the true nonzero coefficients are not too small.

5.3 Real Data Analysis

We apply SCAD to a regression problem of gene microarrays in high-dimensional settings. We use the data set used by Scheetz et al. (2006), which consists of gene expression levels of 18,975 genes obtained from 120 rats. The main objective of the analysis is to find genes that are correlated with the *TRIM32* gene, known to cause Bardet–Biedl syndrome. As was done by Huang et al. (2006), we first select 3,000 genes with the largest variance in expression level and then choose the top p genes that have the largest absolute correlation with *TRIM32* among the selected 3,000 genes.

We compare the prediction accuracy of SCAD with those of LASSO and SIS–SCAD. Each data set is divided into two parts, training and test data sets, by randomly selecting $2/3$ observations and $1/3$ observations. The optimal values of the regularization parameters are chosen by cross-validation. The other parameter, a, of the SCAD penalty is fixed at 3.7 by simply following the suggestion of Fan and Li (2001). Results of 100 replicated experiments are summarized in Table 5 according to the number of covariates, $p = 500$, 1,000, and 3,000. In Table 5, "nonzero" denotes the number of predictive variables in the final model and "MSE" is the mean squared error. All values are arithmetic means of 100 replicated experiments.

As shown in Table 5, the SCAD performs best in terms of MSE when $p = 1{,}000$ and 3,000, whereas the LASSO is the best when $p = 500$. SIS–SCAD always performs worst. Note

Table 4. Comparison of the variance estimation

		SIS–SCAD			SCAD			Oracle		
n		SD	$\mathrm{SD_m}$	$\mathrm{SD_{mad}}$	SD	$\mathrm{SD_m}$	$\mathrm{SD_{mad}}$	SD	$\mathrm{SD_m}$	$\mathrm{SD_{mad}}$
100	$\hat{\beta}_1$	.154	.093	.012	.166	.083	.023	.085	.100	.015
	$\hat{\beta}_2$	.147	.092	.013	.245	.062	.017	.116	.107	.014
	$\hat{\beta}_3$	.181	.081	.018	.188	.038	.025	.111	.109	.017
	$\hat{\beta}_4$	.301	.067	.050	.181	.031	.030	.107	.107	.017
	$\hat{\beta}_5$	.000	.000	.000	.021	.005	.007	.106	.101	.012
300	$\hat{\beta}_1$	.061	.059	.005	.069	.056	.005	.056	.060	.005
	$\hat{\beta}_2$	.059	.062	.004	.074	.055	.005	.060	.063	.005
	$\hat{\beta}_3$	.106	.057	.007	.125	.050	.010	.055	.063	.005
	$\hat{\beta}_4$	.143	.041	.015	.124	.032	.016	.053	.063	.004
	$\hat{\beta}_5$	.096	.027	.014	.092	.020	.016	.066	.061	.005

Table 5. Performance results of 100 random partitions of the data

	LASSO		SIS–SCAD		SCAD	
p	Nonzero	MSE	Nonzero	MSE	Nonzero	MSE
500	$12.55_{(.206)}$	$.405_{(.010)}$	$5.91_{(.243)}$	$.428_{(.011)}$	$12.04_{(.392)}$	$.416_{(.010)}$
1,000	$13.26_{(.263)}$	$.416_{(.009)}$	$6.00_{(.262)}$	$.438_{(.011)}$	$15.73_{(.338)}$	$.404_{(.010)}$
3,000	$13.77_{(.295)}$	$.422_{(.010)}$	$6.04_{(.251)}$	$.439_{(.011)}$	$18.29_{(.437)}$	$.418_{(.010)}$

NOTE: Standard errors are in parentheses.

that the number of nonzero coefficients of SIS–SCAD is much smaller than those of LASSO and SCAD. This suggests that the poor performance of SIS–SCAD might be because some signal variables are not selected in the first stage.

6. CONCLUDING REMARKS

In this article we have developed an efficient optimization algorithm for SCAD on high dimensions and have proved the oracle property. The numerical results given in Section 5 show that the SCAD estimator is inferior to the oracle estimator in terms of prediction accuracy. In part, this might be because the selected regularization parameter is not optimal. An important direction of future research is to develop a better way to select the regularization parameter.

The oracle property of SCAD on high dimensions would hold for generalized linear models including the logistic regression. The proof is similar to that for the linear regression cases. We will report the results in the near future.

In addition, we have given sufficient conditions under which the global optimum of the SCAD penalty is asymptotically equivalent to that of the oracle estimator. This result is new and interesting in itself; however, our sufficient conditions are rather strong, because it assumes that $p_n \le n$. We believe that the result can be extended to the case where $p_n > n$.

We have considered only the situation where the true model is sparse. When the true model is not sparse, the oracle property will not hold. For LASSO, Greenshtein and Ritov (2004), Van De Geer (2006), Greenshtein (2006), and Wang and Shen (2007) proved persistency, which means that the prediction error converges to its minimum, for various problems including generalized linear model and support vector machines. The persistency also would holds for SCAD when the true model is not sparse, which needs to be proved.

APPENDIX A: PROOF OF THEOREM 1

Let $L_\lambda(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^n (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \sum_{j=1}^p \tilde{J}_\lambda(|\beta_j|)$. Note that $\partial L_\lambda(\boldsymbol{\beta})/\partial\beta_j$ is

$$\begin{cases} -\sum_{i=1}^n x_{ij}(y_i - \mathbf{x}_i'\boldsymbol{\beta})/n - \lambda\,\mathrm{sign}(\beta_j) & \text{if } |\beta_j| \ge a\lambda \\ -\sum_{i=1}^n x_{ij}(y_i - \mathbf{x}_i'\boldsymbol{\beta})/n & \\ \quad + \dfrac{a\lambda - |\beta_j|}{(a-1)}\,\mathrm{sign}(\beta_j) - \lambda\,\mathrm{sign}(\beta_j) & \text{if } \lambda \le |\beta_j| < a\lambda \\ -\sum_{i=1}^n x_{ij}(y_i - \mathbf{x}_i'\boldsymbol{\beta})/n & \text{if } 0 \le |\beta_j| < \lambda. \end{cases}$$

By the second-order sufficiency of the Karush–Kuhn–Tucker condition (see, e.g., Bertsekas 1999, p. 320), any $\boldsymbol{\beta}$ to satisfy

$$S_j(\boldsymbol{\beta}) = 0 \quad \text{and} \quad |\beta_j| \ge a\lambda \quad \text{for } j = 1, \ldots, q_n \tag{A.1}$$

and

$$|S_j(\boldsymbol{\beta})| \le \lambda \quad \text{and} \quad |\beta_j| < \lambda \quad \text{for } j = q_n + 1, \ldots, p_n \tag{A.2}$$

is an element of $\mathcal{A}(\lambda)$. Here $S_j(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^n x_{ij}(y_i - \mathbf{x}_i'\boldsymbol{\beta})$. Thus it suffices to show that $\hat{\boldsymbol{\beta}}^o$ satisfies (A.1) and (A.2) with $\lambda = \lambda_n$.

For $j \le q_n$, $S_j(\hat{\boldsymbol{\beta}}^o) = 0$ holds trivially by the definition of the oracle estimator. So for proving (A.1), it suffices to prove that as $n \to \infty$,

$$\Pr(|\hat{\boldsymbol{\beta}}_j^o| \ge a\lambda_n \text{ for } j = 1, \ldots, q_n) \to 1. \tag{A.3}$$

Note that

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(1)o} &= \frac{1}{n}\big(\mathbf{C}_n^{(1,1)}\big)^{-1}\mathbf{X}_n^{(1)'}Y_n \\ &= \frac{1}{n}\big(\mathbf{C}_n^{(1,1)}\big)^{-1}\mathbf{X}_n^{(1)'}\big(\boldsymbol{\epsilon}_n + \mathbf{X}_n^{(1)}\boldsymbol{\beta}^{*(1)}\big) \\ &= \frac{1}{n}\big(\mathbf{C}_n^{(1,1)}\big)^{-1}\mathbf{X}_n^{(1)'}\boldsymbol{\epsilon}_n + \boldsymbol{\beta}^{*(1)}. \end{aligned}$$

Note that $|\hat{\beta}_j^o| \ge |\beta_j^*| - |\hat{\beta}_j^o - \beta_j^*|$. Because $\min_{j\le q_n}|\beta_j^*| = O(n^{-(1-c_2)/2})$ and $\lambda_n = o(n^{-(1-c_2+c_1)/2})$, it suffices to show that

$$\max_{j\le q_n}|\hat{\beta}_j^o - \beta_j^*| = o_p\big(n^{-(1-c_2)/2}\big). \tag{A.4}$$

Let $z_j = \sqrt{n}(\hat{\beta}_j^o - \beta_j^*)$. For (A.4), we show that

$$\max_{j\le q_n}|z_j| = o_p\big(n^{c_2/2}\big). \tag{A.5}$$

Write

$$\mathbf{z} = \big(\mathbf{C}_n^{(1,1)}\big)^{-1}\frac{\mathbf{X}_n^{(1)'}\boldsymbol{\epsilon}_n}{\sqrt{n}} = \mathbf{H}^{(1)'}\boldsymbol{\epsilon}_n,$$

where $\mathbf{z} = (z_1, \ldots, z_{q_n})'$ and $\mathbf{H}^{(1)'} = (\mathbf{h}_1^{(1)}, \ldots, \mathbf{h}_{q_n}^{(1)})' = (\mathbf{C}_n^{(1,1)})^{-1} \times \mathbf{X}_n^{(1)'}/\sqrt{n}$. Because $\mathbf{H}^{(1)'}\mathbf{H}^{(1)} = (\mathbf{C}_n^{(1,1)})^{-1}$, regularity condition A2 implies that $\|\mathbf{h}_j^{(1)}\|_2^2 \le 1/M_2$ for all $j \le q_n$. Thus $\mathrm{E}(z_j)^{2k} < \infty$ for all $j \le q_n$ because $\mathrm{E}(\epsilon_i)^{2k} < \infty$. Therefore,

$$\Pr(|z_j| > t) = O(t^{-2k}). \tag{A.6}$$

For any $\eta > 0$, we can write

$$\begin{aligned} &\Pr\big(|z_j| > \eta n^{c_2/2} \text{ for some } j = 1, \ldots, q_n\big) \\ &\quad \le \sum_{j=1}^{q_n}\Pr\big(|z_j| > \eta n^{c_2/2}\big) \\ &\quad \le \sum_{j=1}^{q_n}\frac{1}{\eta}n^{-c_2 k} \\ &\quad = \frac{1}{\eta}q_n n^{-c_2 k} \le \frac{1}{\eta}n^{-(c_2-c_1)k} \to 0, \end{aligned}$$

which completes the proof of (A.1) for the oracle estimator.

To show (A.2), we have $|\hat{\beta}_j^o| \le \lambda_n$ for $j > q_n$, because $\hat{\beta}_j^o = 0$ by the definition. Thus it suffices to show that

$$\Pr\bigl(|S_j(\hat{\boldsymbol{\beta}}^O)| > \lambda_n \text{ for some } j = q_n + 1, \ldots, p_n\bigr) \to 0. \tag{A.7}$$

Note that

$$\begin{aligned}
&(S_j(\hat{\boldsymbol{\beta}}^O),\ j = q_n + 1, \ldots, p_n)\\
&\quad = -\frac{1}{n}\mathbf{X}_n^{(2)'}\bigl(Y_n - \mathbf{X}_n^{(1)}\hat{\boldsymbol{\beta}}^{O(1)}\bigr)\\
&\quad = -\frac{1}{n}\mathbf{X}_n^{(2)'}\left(Y_n - \mathbf{X}_n^{(1)}\frac{1}{n}\bigl(\mathbf{C}_n^{(1,1)}\bigr)^{-1}\mathbf{X}_n^{(1)'}Y_n\right)\\
&\quad = -\frac{1}{n}\mathbf{X}_n^{(2)'}\left(\mathbf{X}_n^{(1)}\boldsymbol{\beta}^{*(1)} + \boldsymbol{\epsilon}_n\right.\\
&\qquad\qquad \left. - \mathbf{X}_n^{(1)}\frac{1}{n}\bigl(\mathbf{C}_n^{(1,1)}\bigr)^{-1}\mathbf{X}_n^{(1)'}\bigl(\mathbf{X}_n^{(1)}\boldsymbol{\beta}^{*(1)} + \boldsymbol{\epsilon}_n\bigr)\right)\\
&\quad = -\frac{1}{n}\mathbf{X}_n^{(2)'}\left(\mathbf{I} - \frac{1}{n}\mathbf{X}_n^{(1)}\bigl(\mathbf{C}_n^{(1,1)}\bigr)^{-1}\mathbf{X}_n^{(1)'}\right)\boldsymbol{\epsilon}_n.
\end{aligned}$$

Thus we have

$$\sqrt{n}S_j(\hat{\boldsymbol{\beta}}^O) = \mathbf{h}_j^{(2)'}\boldsymbol{\epsilon}_n \quad \text{for } j = q_n + 1, \ldots, p_n,$$

where $\mathbf{h}_j^{(2)}$ is the $j - q_n$ column vector of $\mathbf{H}^{(2)}$ and

$$\mathbf{H}^{(2)'} = \mathbf{C}_n^{(2,1)}\bigl(\mathbf{C}_n^{(1,1)}\bigr)^{-1}\frac{1}{\sqrt{n}}\mathbf{X}_n^{(1)'} - \frac{1}{\sqrt{n}}\mathbf{X}_n^{(2)'}.$$

Note that

$$\mathbf{H}^{(2)'}\mathbf{H}^{(2)} = \frac{1}{n}\mathbf{X}_n^{(2)'}\bigl(\mathbf{I} - \mathbf{X}_n^{(1)}\bigl(\mathbf{X}_n^{(1)'}\mathbf{X}_n^{(1)}\bigr)^{-1}\mathbf{X}_n^{(1)'}\bigr)\mathbf{X}_n^{(2)}.$$

Because all of the eigenvalues of $\mathbf{I} - \mathbf{X}_n^{(1)}(\mathbf{X}_n^{(1)'}\mathbf{X}_n^{(1)})^{-1}\mathbf{X}_n^{(1)'}$ are between 0 and 1, we have that $\|\mathbf{h}_j^{(2)}\|_2^2 \le M_1$ for all $j = q_n + 1, \ldots, p_n$. Thus $\mathrm{E}(\xi_j)^{2k} < \infty$, where $\xi_j = \sqrt{n}S_j(\hat{\boldsymbol{\beta}}^O)$, and so

$$\Pr(|\xi_j| > t) = O(t^{-2k}). \tag{A.8}$$

Finally,

$$\begin{aligned}
&\Pr\bigl(|S_j(\hat{\boldsymbol{\beta}}^O)| > \lambda_n \text{ for some } j = q_n + 1, \ldots, p_n\bigr)\\
&\quad = \Pr(|\xi_j| > \sqrt{n}\lambda_n \text{ for some } j = q_n + 1, \ldots, p_n)\\
&\quad \le \sum_{j=q_n+1}^{p_n}\Pr(|\xi_j| > \sqrt{n}\lambda_n)\\
&\quad = (p_n - q_n)O\left(\frac{1}{(\sqrt{n}\lambda_n)^{2k}}\right) = O\left(\frac{p_n}{(\sqrt{n}\lambda_n)^{2k}}\right) \to 0.
\end{aligned}$$

which completes the proof.

APPENDIX B: PROOF OF THEOREM 2

Note that z_j and ξ_j are Gaussian random variables with mean 0 and bounded second moments. For a Gaussian random variable W with mean 0 and variance σ^2, we have

$$\Pr(|W| > t) \le \sqrt{\frac{2}{\pi}}\exp\left(-\frac{t^2}{2\sigma^2}\right) \tag{B.1}$$

for $t \ge \sigma$. Thus the theorem can be easily proved by replacing the tail bounds of (A.6) and (A.8) with the better exponential bounds of (B.1).

APPENDIX C: PROOF OF THEOREM 3

For notational simplicity, we drop the subscript n. Let $\mathcal{P} = \{1, 2, \ldots, q\}$ and $\mathcal{N} = \{q + 1, \ldots, p\}$. Let $\hat{\mathbf{X}}^{(2)}$ be the columnwise projection of $\mathbf{X}^{(2)}$ onto $\mathbf{X}^{(1)}$; that is, the jth column of $\hat{\mathbf{X}}^{(2)}$ is the projection of the jth column of $\mathbf{X}^{(2)}$ onto the linear space spanned by the column vectors of $\mathbf{X}^{(1)}$. Let $\tilde{\mathbf{X}}^{(2)} = \mathbf{X}^{(2)} - \hat{\mathbf{X}}^{(2)}$. Because $\mathbf{C}$ has the positive smallest eigenvalue ρ, it can be shown that $\tilde{\mathbf{C}}^{(2.2)} = \tilde{\mathbf{X}}^{(2)'}\tilde{\mathbf{X}}^{(2)}/n$ also has the smallest eigenvalue larger than a certain number, $\tilde{\rho} > 0$.

Let $\hat{y}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}^O$. Then we can write

$$\begin{aligned}
C(\boldsymbol{\beta}) = \bigl\|\hat{Y} - \mathbf{X}^{(1)}\boldsymbol{\beta}^{(1)} - \hat{\mathbf{X}}^{(2)}\boldsymbol{\beta}^{(2)}\bigr\|_2^2/2n\\
+ \bigl\|Y - \hat{Y} - \tilde{\mathbf{X}}^{(2)}\boldsymbol{\beta}^{(2)}\bigr\|_2^2/2n + \sum_{j=1}^{p} J_{\lambda_n}(|\beta_j|),
\end{aligned}$$

where $\hat{Y} = (\hat{y}_1, \ldots, \hat{y}_n)'$. Note that

$$\begin{aligned}
\bigl\|Y - \hat{Y} - \tilde{\mathbf{X}}^{(2)}\boldsymbol{\beta}^{(2)}\bigr\|_2^2 = \|Y - \hat{Y}\|_2^2 + n\boldsymbol{\beta}^{(2)'}\tilde{\mathbf{C}}^{(2,2)}\boldsymbol{\beta}^{(2)}\\
- 2\sum_{j\in\mathcal{N}}\beta_j\bigl\langle Y - \hat{Y}, \tilde{X}^{(j)}\bigr\rangle.
\end{aligned} \tag{C.1}$$

Also, by condition A4 and (A.4), we have that $\min_{j\in\mathcal{P}}|\hat{\beta}_j^o| = O_p^+(n^{-(1-c_2)/2}) \ge a\lambda_n$. Here $O_p^+(n^a)$ is defined as a sequence of positive random variables such that there exists a positive number τ with $\Pr(n^{-a}O_p^+(n^a) \ge \tau) \to 1$ as $n \to \infty$. To prove Theorem 3, it suffices to show that

$$\Pr\bigl(C(\boldsymbol{\beta}) \ge C(\hat{\boldsymbol{\beta}}^O) \text{ for all } \boldsymbol{\beta} \in R^p\bigr) \to 1$$

as $n \to \infty$.

Because $S_j(\hat{\boldsymbol{\beta}}^O) = \langle Y - \hat{Y}, X^{(j)}\rangle/n$ for $j \in \mathcal{N}$, (A.7) implies that

$$\max_{j\in\mathcal{N}}\bigl|\bigl\langle Y - \hat{Y}, X^{(j)}\bigr\rangle\bigr|/n = o_p(\lambda_n).$$

Similarly, we can prove that

$$\max_{j\in\mathcal{N}}\bigl|\bigl\langle Y - \hat{Y}, \tilde{X}^{(j)}\bigr\rangle\bigr|/n = o_p(\lambda_n). \tag{C.2}$$

For given $\boldsymbol{\beta} \in R^p$, let $\mathcal{P}^+ = \{j \in \mathcal{P} : |\beta_j| > a\lambda_n\}$, and $\mathcal{N}^+ = \{j \in \mathcal{N} : |\beta_j| > \lambda_n\}$, and let $\mathcal{P}^- = \mathcal{P} - \mathcal{P}^+$ and $\mathcal{N}^- = \mathcal{N} - \mathcal{N}^+$. Consider the linear space $\mathcal{A}$ spanned by $\{\hat{X}^{(j)} : j \in \mathcal{P}^+ \cup \mathcal{N}\}$, where we let $\hat{X}^{(j)} = X^{(j)}$ for $j \in \mathcal{P}$. Let $\hat{Y}_{\mathcal{A}}$ be the projection of $\hat{Y}$ onto $\mathcal{A}$. Because $\min_{j\in\mathcal{P}}|\hat{\beta}_j^o| = O_p^+(n^{-(1-c_2)/2})$ and the smallest eigenvalue of $\mathbf{C}$ is positive, we have

$$\|\hat{Y} - \hat{Y}_{\mathcal{A}}\|_2^2/2n = rO_p^+(n^{-1+c_2}), \tag{C.3}$$

where r is the cardinality of $\mathcal{P}^-$.

Note that

$$\boldsymbol{\beta}^{(2)'}\tilde{\mathbf{C}}^{(2,2)}\boldsymbol{\beta}^{(2)} \ge \tilde{\rho}\sum_{j\in\mathcal{N}}\beta_j^2 \ge \tilde{\rho}\sum_{j\in\mathcal{N}^+}\beta_j^2 \ge \tilde{\rho}\lambda_n\sum_{j\in\mathcal{N}^+}|\beta_j| \tag{C.4}$$

and

$$\begin{aligned}
\Pr\left(\sum_{j=1}^{p}\bigl(J_{\lambda_n}(|\beta_j|) - J_{\lambda_n}(|\hat{\beta}_j^o|)\bigr)\right.\\
\left.\ge \lambda_n\sum_{j\in\mathcal{N}^-}|\beta_j| - rO(\lambda_n^2) \text{ for all } \boldsymbol{\beta} \in R^p\right) \to 1
\end{aligned} \tag{C.5}$$

as $n \to \infty$.

Let $\mathbf{X}_{\mathcal{A}} = (\hat{X}^{(j)}, j \in \mathcal{P}^+ \cup \mathcal{N})$ and $\boldsymbol{\beta}^{\mathcal{A}} = (\beta_j, j \in \mathcal{P}^+ \cup \mathcal{N})$. Similarly, we let $\mathbf{X}_{\mathcal{A}^c} = (\hat{X}^{(j)}, j \in \mathcal{P}^-)$ and $\boldsymbol{\beta}^{\mathcal{A}^c} = (\beta_j, j \in \mathcal{P}^-)$. Then (C.1), (C.2), (C.3), (C.4), and (C.5) imply that the probability of

$$
\begin{aligned}
& C(\boldsymbol{\beta}) - C(\hat{\boldsymbol{\beta}}^o) \\
&= \left\| \hat{Y} - \hat{Y}_{\mathcal{A}} + \hat{Y}_{\mathcal{A}} - \mathbf{X}_{\mathcal{A}}\boldsymbol{\beta}^{\mathcal{A}} - \mathbf{X}_{\mathcal{A}^c}\boldsymbol{\beta}^{\mathcal{A}^c} \right\|_2^2 / 2n \\
&\quad + \boldsymbol{\beta}^{(2)'} \tilde{\mathbf{C}}^{(2,2)} \boldsymbol{\beta}^{(2)} / 2 \\
&\quad - \sum_{j \in \mathcal{N}} \beta_j \langle Y - \hat{Y}, \tilde{X}^{(j)} \rangle / n + \sum_{j=1}^{p} \left(J_{\lambda_n}(|\beta_j|) - J_{\lambda_n}(|\hat{\beta}_j^o|) \right) \\
&\geq \|\hat{Y} - \hat{Y}_{\mathcal{A}}\|_2^2 / 2n + \|\hat{Y}_{\mathcal{A}} - \mathbf{X}_{\mathcal{A}}\boldsymbol{\beta}^{\mathcal{A}}\|_2^2 / 2n - \left\| \hat{\mathbf{X}}_{\mathcal{A}^c}\boldsymbol{\beta}^{\mathcal{A}^c} \right\|_2^2 / 2n \\
&\quad + \lambda_n \sum_{j \in \mathcal{N}^-} |\beta_j| + \frac{\tilde{\rho}}{2}\lambda_n \sum_{j \in \mathcal{N}^+} |\beta_j| - o_p(\lambda_n) \sum_{j \in \mathcal{N}} |\beta_j| \\
&\quad - r\, O(\lambda_n^2)
\end{aligned}
$$

for all $\boldsymbol{\beta} \in R^p$ converges to 1 as $n \to \infty$. Note that the largest eigenvalue of $\mathbf{X}'_{\mathcal{A}^c}\mathbf{X}_{\mathcal{A}^c}/n$ is bounded by a certain positive constant M_1, because that of $\mathbf{C}$ is bounded. So we have

$$
\left\| \hat{\mathbf{X}}_{\mathcal{A}^c}\boldsymbol{\beta}^{\mathcal{A}^c} \right\|_2^2 / 2n \leq \frac{M_1}{2} \sum_{j \in \mathcal{P}^-} \beta_j^2 \leq \frac{M_1}{2} a^2 r \lambda_n^2.
$$

Thus the probability of

$$
\begin{aligned}
C(\boldsymbol{\beta}) - C(\hat{\boldsymbol{\beta}}^o) \geq\ & r\left(O_p^+(n^{-1+c_2}) - O(\lambda_n^2) \right) \\
& + \left(\min\{1, \tilde{\rho}/2\}\lambda_n - o(\lambda_n) \right) \sum_{j \in \mathcal{N}} |\beta_j| \qquad \text{(C.6)}
\end{aligned}
$$

for all $\boldsymbol{\beta} \in R^p$ converges to 1. Because $O(\lambda_n^2) = o(n^{-1+c_2})$, the left side of (C.6) becomes nonnegative in probability, and the proof is complete.

[Received June 2007. Revised July 2008.]

REFERENCES

An, L. T. H., and Tao, P. D. (1997), "Solving a Class of Linearly Constrained Indefinite Quadratic Problems by DC Algorithms," *Journal of Global Optimization*, 11, 253–285.

Antoniadis, A., and Fan, J. (2001), "Regularization of Wavelets Approximations," *Journal of the American Statistical Association*, 96, 939–967.

Bertsekas, D. P. (1999), *Nonlinear Programming* (2nd ed.), Belmount, MA: Athena Scientific.

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

Collobert, R., Sinz, F., Weston, J., and Bottou, L. (2006), "Large-Scale Transductive SVMs," *Journal of Machine Learning Research*, 7, 1687–1712.

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultra–High-Dimensional Feature Space," *Journal of the Royal Statistical Society*, Ser. B, 70, 849–911.

Fan, J., and Peng, H. (2004), "Nonconcave Penalized Likelihood With a Diverging Number of Parameters," *The Annals of Statistics*, 32, 928–961.

Frank, I. E., and Friedman, J. H. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35, 109–148.

Friedman, J. H., and Popescu, B. E. (2004), "Gradient Directed Regularization," technical report, Stanford University, Dept. of Statistics.

Greenshtein, E. (2006), "Best Subset Selection, Persistence in High-Dimensional Statistical Learning and Optimization Under l_1 Constraint," *The Annals of Statistics*, 34, 2367–2386.

Greenshtein, E., and Ritov, Y. (2004), "Persistence in High-Dimensional Linear Predictor Selection and the Virtue of Overparametrization," *Bernoulli*, 10, 971–988.

Huang, J., Ma, S., and Zhang, C.-H. (2006), "Adaptive Lasso for Sparse High-Dimensional Regression Models," Technical Report 374, University of Iowa, Dept. of Statistics and Actuarial Science.

Knight, K., and Fu, W. J. (2000), "Asymptotics for Lasso-Type Estimators," *The Annals of Statistics*, 28, 1356–1378.

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection With the Lasso," *The Annals of Statistics*, 34, 1436–1462.

Rosset, S., and Zhu, J. (2007), Piecewise Linear Regularized Solution Paths," *The Annals of Statistics*, 35, 1012–1030.

Scheetz, T. E., Kim, K. Y., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Sheffield, V. C., and Stone, E. M. (2006), "Regulation of Gene Expression in the Mammalian Eye and Its Relevance to Eye Disease," *Proceedings of the National Academy of Sciences*, 103, 14429–14434.

Shen, X., Tseng, G. C., Zhang, X., and Wong, W. H. (2003), "On ψ-Learning," *Journal of the American Statistical Association*, 98, 724–734.

Tibshirani, R. J. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Ser. B, 58, 267–288.

Van De Geer, S. (2006), "High-Dimensional Generalized Linear Models and Lasso," Research Report 133, Seminar für Statistik, ETH Zürich.

Wang, L., and Shen, X. (2007), "On l_1 Norm Multi-Class Support Vector Machines: Methodology and Theory," *Journal of the American Statistical Association*, 102, 583–594.

Yuille, A., and Rangarajan, A. (2003), "The Concave–Convex Procedure," *Neural Computation*, 15, 915–936.

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563.

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

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society*, Ser. B, 67, 301–320.