The Adaptive Lasso and Its Oracle Properties

Hui Zou

The lasso is a popular technique for simultaneous estimation and variable selection. Lasso variable selection has been shown to be consistent under certain conditions. In this work we derive a necessary condition for the lasso variable selection to be consistent. Consequently, there exist certain scenarios where the lasso is inconsistent for variable selection. We then propose a new version of the lasso, called the adaptive lasso, where adaptive weights are used for penalizing different coefficients in the ℓ_1 penalty. We show that the adaptive lasso enjoys the oracle properties; namely, it performs as well as if the true underlying model were given in advance. Similar to the lasso, the adaptive lasso is shown to be near-minimax optimal. Furthermore, the adaptive lasso can be solved by the same efficient algorithm for solving the lasso. We also discuss the extension of the adaptive lasso in generalized linear models and show that the oracle properties still hold under mild regularity conditions. As a byproduct of our theory, the nonnegative garotte is shown to be consistent for variable selection.

KEY WORDS: Asymptotic normality; Lasso; Minimax; Oracle inequality; Oracle procedure; Variable selection.

1. INTRODUCTION

There are two fundamental goals in statistical learning: ensuring high prediction accuracy and discovering relevant predictive variables. Variable selection is particularly important when the true underlying model has a sparse representation. Identifying significant predictors will enhance the prediction performance of the fitted model. Fan and Li (2006) gave a comprehensive overview of feature selection and proposed a unified penalized likelihood framework to approach the problem of variable selection.

Let us consider model estimation and variable selection in linear regression models. Suppose that $\mathbf{y} = (y_1, \ldots, y_n)^T$ is the response vector and $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$, $j = 1, \ldots, p$, are the linearly independent predictors. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ be the predictor matrix. We assume that $E[y|\mathbf{x}] = \beta_1^* x_1 + \cdots + \beta_p^* x_p$. Without loss of generality, we assume that the data are centered, so the intercept is not included in the regression function. Let $\mathcal{A} = \{j : \beta_j^* \neq 0\}$ and further assume that $|\mathcal{A}| = p_0 < p$. Thus the true model depends only on a subset of the predictors. Denote by $\hat{\boldsymbol{\beta}}(\delta)$ the coefficient estimator produced by a fitting procedure δ. Using the language of Fan and Li (2001), we call δ an *oracle* procedure if $\hat{\boldsymbol{\beta}}(\delta)$ (asymptotically) has the following oracle properties:

- Identifies the right subset model, $\{j : \hat{\beta}_j \neq 0\} = \mathcal{A}$
- Has the optimal estimation rate, $\sqrt{n}(\hat{\boldsymbol{\beta}}(\delta)_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^*) \rightarrow_d \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}^*)$, where $\boldsymbol{\Sigma}^*$ is the covariance matrix knowing the true subset model.

It has been argued (Fan and Li 2001; Fan and Peng 2004) that a good procedure should have these oracle properties. However, some extra conditions besides the oracle properties, such as continuous shrinkage, are also required in an optimal procedure.

Ordinary least squares (OLS) gives nonzero estimates to all coefficients. Traditionally, statisticians use best-subset selection to select significant variables, but this procedure has two fundamental limitations. First, when the number of predictors is large, it is computationally infeasible to do subset selection. Second, subset selection is extremely variable because of its inherent discreteness (Breiman 1995; Fan and Li 2001). Stepwise selection is often used as a computational surrogate to subset selection; nevertheless, stepwise selection still suffers from the high variability and in addition is often trapped into a local optimal solution rather than the global optimal solution. Furthermore, these selection procedures ignore the stochastic errors or uncertainty in the variable selection stage (Fan and Li 2001; Shen and Ye 2002).

The lasso is a regularization technique for simultaneous estimation and variable selection (Tibshirani 1996). The lasso estimates are defined as

$$\hat{\boldsymbol{\beta}}(\text{lasso}) = \arg\min_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \qquad (1)$$

where λ is a nonnegative regularization parameter. The second term in (1) is the so-called "ℓ_1 penalty," which is crucial for the success of the lasso. The ℓ_1 penalization approach is also called *basis pursuit* in signal processing (Chen, Donoho, and Saunders 2001). The lasso continuously shrinks the coefficients toward 0 as λ increases, and some coefficients are shrunk to exact 0 if λ is sufficiently large. Moreover, continuous shrinkage often improves the prediction accuracy due to the bias–variance trade-off. The lasso is supported by much theoretical work. Donoho, Johnstone, Kerkyacharian, and Picard (1995) proved the near-minimax optimality of soft thresholding (the lasso shrinkage with orthogonal predictors). It also has been shown that the ℓ_1 approach is able to discover the "right" sparse representation of the model under certain conditions (Donoho and Huo 2002; Donoho and Elad 2002; Donoho 2004). Meinshausen and Bühlmann (2004) showed that variable selection with the lasso can be consistent if the underlying model satisfies some conditions.

It seems safe to conclude that the lasso is an oracle procedure for simultaneously achieving consistent variable selection and optimal estimation (prediction). However, there are also solid arguments against the lasso oracle statement. Fan and Li (2001) studied a class of penalization methods including the lasso. They showed that the lasso can perform automatic variable selection because the ℓ_1 penalty is singular at the origin. On the other hand, the lasso shrinkage produces biased estimates for the large coefficients, and thus it could be suboptimal in terms of estimation risk. Fan and Li conjectured that the oracle properties do not hold for the lasso. They also proposed a smoothly clipped absolute deviation (SCAD) penalty for variable selection and proved its oracle properties.

Hui Zou is Assistant Professor of Statistics, School of Statistics, University of Minnesota, Minneapolis, MN 55455 (E-mail: *hzou@stat.umn.edu*). The author thanks an associate editor and three referees for their helpful comments and suggestions. Sincere thanks also go to a co-editor for his encouragement.

Journal of the American Statistical Association
December 2006, Vol. 101, No. 476, Theory and Methods
DOI 10.1198/016214506000000735

Meinshausen and Bühlmann (2004) also showed the conflict of optimal prediction and consistent variable selection in the lasso. They proved that the optimal λ for prediction gives inconsistent variable selection results; in fact, many noise features are included in the predictive model. This conflict can be easily understood by considering an orthogonal design model (Leng, Lin, and Wahba 2004).

Whether the lasso is an oracle procedure is an important question demanding a definite answer, because the lasso has been used widely in practice. In this article we attempt to provide an answer. In particular, we are interested in whether the ℓ_1 penalty could produce an oracle procedure and, if so, how. We consider the asymptotic setup where λ in (1) varies with n (the sample size). We first show that the underlying model must satisfy a nontrivial condition if the lasso variable selection is consistent. Consequently, there are scenarios in which the lasso selection cannot be consistent. To fix this problem, we then propose a new version of the lasso, the adaptive lasso, in which adaptive weights are used for penalizing different coefficients in the ℓ_1 penalty. We show that the adaptive lasso enjoys the oracle properties. We also prove the near-minimax optimality of the adaptive lasso shrinkage using the language of Donoho and Johnstone (1994). The adaptive lasso is essentially a convex optimization problem with an ℓ_1 constraint. Therefore, the adaptive lasso can be solved by the same efficient algorithm for solving the lasso. Our results show that the ℓ_1 penalty is at least as competitive as other concave oracle penalties and also is computationally more attractive. We consider this article to provide positive evidence supporting the use of the ℓ_1 penalty in statistical learning and modeling.

The nonnegative garotte (Breiman 1995) is another popular variable selection method. We establish a close relation between the nonnegative garotte and a special case of the adaptive lasso, which we use to prove consistency of the nonnegative garotte selection.

The rest of the article is organized as follows. In Section 2 we derive the necessary condition for the consistency of the lasso variable selection. We give concrete examples to show when the lasso fails to be consistent in variable selection. We define the adaptive lasso in Section 3, and then prove its statistical properties. We also show that the nonnegative garotte is consistent for variable selection. We apply the LARS algorithm (Efron, Hastie, Johnstone, and Tibshirani 2004) to solve the entire solution path of the adaptive lasso. We use a simulation study to compare the adaptive lasso with several popular sparse modeling techniques. We discuss some applications of the adaptive lasso in generalized liner models in Section 4, and give concluding remarks in Section 5. We relegate technical proofs to the Appendix.

2. THE LASSO VARIABLE SELECTION COULD BE INCONSISTENT

We adopt the setup of Knight and Fu (2000) for the asymptotic analysis. We assume two conditions:

(a) $y_i = \mathbf{x}_i\boldsymbol{\beta}^* + \epsilon_i$, where $\epsilon_1, \ldots, \epsilon_n$ are independent identically distributed (iid) random variables with mean 0 and variance σ^2

(b) $\frac{1}{n}\mathbf{X}^T\mathbf{X} \to \mathbf{C}$, where $\mathbf{C}$ is a positive definite matrix.

Without loss of generality, assume that $\mathcal{A} = \{1, 2, \ldots, p_0\}$. Let

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix},$$

where $\mathbf{C}_{11}$ is a $p_0 \times p_0$ matrix.

We consider the lasso estimates, $\hat{\boldsymbol{\beta}}^{(n)}$,

$$\hat{\boldsymbol{\beta}}^{(n)} = \arg\min_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j\beta_j \right\|^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|, \tag{2}$$

where λ_n varies with n. Let $\mathcal{A}_n = \{j : \hat{\beta}_j^{(n)} \neq 0\}$. The lasso variable selection is consistent if and only if $\lim_n P(\mathcal{A}_n = \mathcal{A}) = 1$.

Lemma 1. If $\lambda_n/n \to \lambda_0 \geq 0$, then $\hat{\boldsymbol{\beta}}^{(n)} \to_p \arg\min V_1$, where

$$V_1(\mathbf{u}) = (\mathbf{u} - \boldsymbol{\beta}^*)^T\mathbf{C}(\mathbf{u} - \boldsymbol{\beta}^*) + \lambda_0 \sum_{j=1}^{p} |u_j|.$$

Lemma 2. If $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta}^*) \to_d \arg\min(V_2)$, where

$$V_2(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T\mathbf{C}\mathbf{u} + \lambda_0 \sum_{j=1}^{p} \big[u_j \operatorname{sgn}(\beta_j^*) I(\beta_j^* \neq 0) + |u_j| I(\beta_j^* = 0)\big],$$

and $\mathbf{W}$ has a $\mathrm{N}(\mathbf{0}, \sigma^2\mathbf{C})$ distribution.

These two lemmas are quoted from Knight and Fu (2000). From an estimation standpoint, Lemma 2 is more interesting, because it shows that the lasso estimate is root-n consistent. In Lemma 1, only $\lambda_0 = 0$ guarantees estimation consistency. However, when considering the asymptotic behavior of variable selection, Lemma 2 actually implies that when $\lambda_n = O(\sqrt{n})$, $\mathcal{A}_n$ basically cannot be $\mathcal{A}$ with a positive probability.

Proposition 1. If $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$, then $\limsup_n P(\mathcal{A}_n = \mathcal{A}) \leq c < 1$, where c is a constant depending on the true model.

Based on Proposition 1, it seems interesting to study the asymptotic behavior of $\hat{\boldsymbol{\beta}}^{(n)}$ when $\lambda_0 = \infty$, which amounts to considering the case where $\lambda_n/n \to 0$ and $\lambda_n/\sqrt{n} \to \infty$. We provide the following asymptotic result.

Lemma 3. If $\lambda_n/n \to 0$ and $\lambda_n/\sqrt{n} \to \infty$, then $\frac{n}{\lambda_n}(\hat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta}^*) \to_p \arg\min(V_3)$, where

$$V_3(\mathbf{u}) = \mathbf{u}^T\mathbf{C}\mathbf{u} + \sum_{j=1}^{p} \big[u_j \operatorname{sgn}(\beta_j^*) I(\beta_j^* \neq 0) + |u_j| I(\beta_j^* = 0)\big].$$

Two observations can be made from Lemma 3. The convergence rate of $\hat{\boldsymbol{\beta}}^{(n)}$ is slower than $\sqrt{n}$. The limiting quantity is nonrandom. Note that the optimal estimation rate is available only when $\lambda_n = O(\sqrt{n})$, but it leads to inconsistent variable selection.

Then we would like to ask whether the consistency in variable selection could be achieved if we were willing to sacrifice the rate of convergence in estimation. Unfortunately, this is not always guaranteed either. The next theorem presents a necessary condition for consistency of the lasso variable selection. [After finishing this work, we learned, through personal com-

munication that Yuan and Lin (2005) and Zhao and Yu (2006) also obtained a similar conclusion about the consistency of the lasso selection.]

Theorem 1 (Necessary condition). Suppose that $\lim_n P(\mathcal{A}_n = \mathcal{A}) = 1$. Then there exists some sign vector $\mathbf{s} = (s_1, \ldots, s_{p_0})^T$, $s_j = 1$ or -1, such that

$$|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{s}| \le 1. \tag{3}$$

The foregoing inequality is understood componentwise.

Immediately, we conclude that if condition (3) fails, then the lasso variable selection is inconsistent. The necessary condition (3) is nontrivial. We can easily construct an interesting example as follows.

Corollary 1. Suppose that $p_0 = 2m + 1 \ge 3$ and $p = p_0 + 1$, so there is one irrelevant predictor. Let $\mathbf{C}_{11} = (1 - \rho_1)\mathbf{I} + \rho_1\mathbf{J}_1$, where $\mathbf{J}_1$ is the matrix of 1's and $\mathbf{C}_{12} = \rho_2\vec{1}$ and $\mathbf{C}_{22} = 1$. If $-\frac{1}{p_0-1} < \rho_1 < -\frac{1}{p_0}$ and $1 + (p_0 - 1)\rho_1 < |\rho_2| < \sqrt{(1 + (p_0 - 1)\rho_1)/p_0}$, then condition (3) cannot be satisfied. Thus the lasso variable selection is inconsistent.

In many problems the lasso has shown its good performance in variable selection. Meinshausen and Bühlmann (2004) showed that the lasso selection can be consistent if the underlying model satisfies some conditions. A referee pointed out that assumption (6) of Meinshausen and Bühlmann (2004) is related to the necessary condition (3), and it cannot be further relaxed, as indicated in their proposition 3. There are some simple settings in which the lasso selection is consistent and the necessary condition (3) is satisfied. For instance, orthogonal design guarantees the necessary condition and consistency of the lasso selection. It is also interesting to note that when $p = 2$, the necessary condition is always satisfied, because $|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\operatorname{sgn}(\boldsymbol{\beta}^*_{\mathcal{A}})|$ reduces to $|\rho|$, the correlation between the two predictors. Moreover, by taking advantage of the closed-form solution of the lasso when $p = 2$ (Tibshirani 1996), we can show that when $p = 2$, the lasso selection is consistent with a proper choice of λ_n.

Given the facts shown in Theorem 1 and Corollary 1, a more important question for lasso enthusiasts is that whether the lasso could be fixed to enjoy the oracle properties. In the next section we propose a simple and effective remedy.

3. ADAPTIVE LASSO

3.1 Definition

We have shown that the lasso cannot be an oracle procedure. However, the asymptotic setup is somewhat unfair, because it forces the coefficients to be equally penalized in the ℓ_1 penalty. We can certainly assign different weights to different coefficients. Let us consider the weighted lasso,

$$\arg\min_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j\beta_j \right\|^2 + \lambda \sum_{j=1}^{p} w_j|\beta_j|,$$

where $\mathbf{w}$ is a known weights vector. We show that if the weights are data-dependent and cleverly chosen, then the weighted lasso can have the oracle properties. The new methodology is called the adaptive lasso.

We now define the adaptive lasso. Suppose that $\hat{\boldsymbol{\beta}}$ is a root-n–consistent estimator to $\boldsymbol{\beta}^*$; for example, we can use $\hat{\boldsymbol{\beta}}(\text{ols})$. Pick a $\gamma > 0$, and define the weight vector $\hat{\mathbf{w}} = 1/|\hat{\boldsymbol{\beta}}|^\gamma$. The adaptive lasso estimates $\hat{\boldsymbol{\beta}}^{*(n)}$ are given by

$$\hat{\boldsymbol{\beta}}^{*(n)} = \arg\min_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j\beta_j \right\|^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j|\beta_j|. \tag{4}$$

Similarly, let $\mathcal{A}^*_n = \{j : \hat{\beta}^{*(n)}_j \ne 0\}$.

It is worth emphasizing that (4) is a convex optimization problem, and thus it does not suffer from the multiple local minimal issue, and its global minimizer can be efficiently solved. This is very different from concave oracle penalties. The adaptive lasso is essentially an ℓ_1 penalization method. We can use the current efficient algorithms for solving the lasso to compute the adaptive lasso estimates. The computation details are presented in Section 3.5.

3.2 Oracle Properties

In this section we show that with a proper choice of λ_n, the adaptive lasso enjoys the oracle properties.

Theorem 2 (Oracle properties). Suppose that $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{(\gamma-1)/2} \to \infty$. Then the adaptive lasso estimates must satisfy the following:

1. Consistency in variable selection: $\lim_n P(\mathcal{A}^*_n = \mathcal{A}) = 1$
2. Asymptotic normality: $\sqrt{n}(\hat{\boldsymbol{\beta}}^{*(n)}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}) \to_d \mathrm{N}(\mathbf{0}, \sigma^2 \times \mathbf{C}_{11}^{-1})$.

Theorem 2 shows that the ℓ_1 penalty is at least as good as any other "oracle" penalty. We have several remarks.

Remark 1. $\hat{\boldsymbol{\beta}}$ is not required to be root-n consistent for the adaptive lasso. The condition can be greatly weakened. Suppose that there is a sequence of $\{a_n\}$ such that $a_n \to \infty$ and $a_n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = O_p(1)$. Then the foregoing oracle properties still hold if we let $\lambda_n = o(\sqrt{n})$ and $a_n^\gamma\lambda_n/\sqrt{n} \to \infty$.

Remark 2. The data-dependent $\hat{w}$ is the key in Theorem 2. As the sample size grows, the weights for zero-coefficient predictors get inflated (to infinity), whereas the weights for nonzero-coefficient predictors converge to a finite constant. Thus we can simultaneously unbiasedly (asymptotically) estimate large coefficient and small threshold estimates. This, in some sense, is the same rationale behind the SCAD. As pointed out by Fan and Li (2001), the oracle properties are closely related to the super-efficiency phenomenon (Lehmann and Casella 1998).

Remark 3. From its definition, we know that the adaptive lasso solution is continuous. This is a nontrivial property. Without continuity, an oracle procedure can be suboptimal. For example, bridge regression (Frank and Friedman 1993) uses the L_q penalty. It has been shown that the bridge with $0 < q < 1$ has the oracle properties (Knight and Fu 2000). But the bridge with $q < 1$ solution is not continuous. Because the discontinuity results in instability in model prediction, the L_q ($q < 1$) penalty is considered less favorable than the SCAD and the ℓ_1 penalty (see Fan and Li 2001 for a more detailed explanation). The discontinuity of the bridge with $0 < q < 1$ can be best demonstrated with orthogonal predictors, as shown in Figure 1.

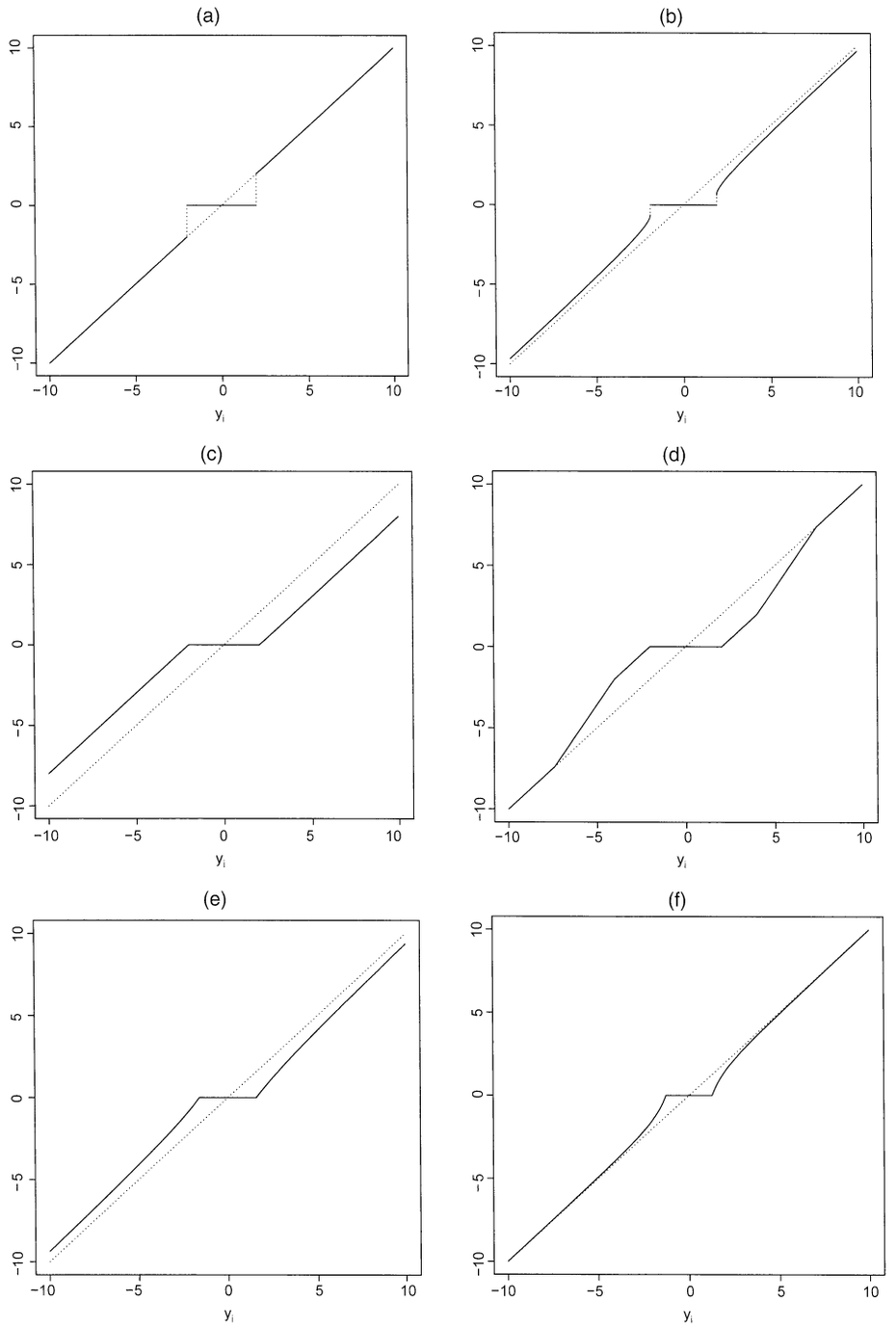

Figure 1. Plot of Thresholding Functions With $\lambda = 2$ for (a) the Hard; (b) Bridge $L_{.5}$; (c) the Lasso; (d) the SCAD; (e) the Adaptive Lasso $\gamma = .5$; and (f) the Adaptive Lasso, $\gamma = 2$.

3.3 Oracle Inequality and Near-Minimax Optimality

As shown by Donoho and Johnstone (1994), the ℓ_1 shrinkage leads to the near–minimax-optimal procedure for estimating nonparametric regression functions. Because the adaptive lasso is a modified version of the lasso with subtle and important differences, it would be interesting to see whether the modification affects the minimax optimality of the lasso. In this section we derive a new oracle inequality to show that the adaptive lasso shrinkage is near-minimax optimal.

For the minimax arguments, we consider the same multiple estimation problem discussed by Donoho and Johnstone (1994). Suppose that we are given n independent observations $\{y_i\}$ generated from

$$y_i = \mu_i + z_i, \qquad i = 1, 2, \ldots, n,$$

where the z_i's are iid normal random variables with mean 0 and known variance σ^2. For simplicity, let us assume that $\sigma = 1$. The objective is to estimate the mean vector (μ_i) by some estimator $(\hat{\mu}_i)$, and the quality of the estimator is mea-

sured by ℓ_2 loss, $R(\hat{\mu}) = E[\sum_i^n (\hat{\mu}_i - \mu_i)^2]$. The ideal risk is $R(\text{ideal}) = \sum_i^n \min(\mu_i^2, 1)$ (Donoho and Johnstone 1994). The soft-thresholding estimates $\hat{\mu}_i(\text{soft})$ are obtained by solving the lasso problems:

$$\hat{\mu}_i(\text{soft}) = \arg\min_u \left(\frac{1}{2}(y_i - u)^2 + \lambda |u| \right) \quad \text{for } i = 1, 2, \ldots, n.$$

The solution is $\hat{\mu}_i(\text{soft}) = (|y_i| - \lambda)_+ \operatorname{sgn}(y_i)$, where z_+ denotes the positive part of z; it equals z for $z > 0$ and 0 otherwise. Donoho and Johnstone (1994) proved that the soft-thresholding estimator achieves performance differing from the ideal performance by at most a $2 \log n$ factor, and that the $2 \log n$ factor is a sharp minimax bound.

The adaptive lasso concept suggests using different weights for penalizing the naive estimates (y_i). In the foregoing setting, we have only one observation for each μ_i. It is reasonable to define the weight vectors as $|y_i|^{-\gamma}$, $\gamma > 0$. Thus the adaptive lasso estimates for (μ_i) are obtained by

$$\hat{\mu}_i^* = \arg\min_u \left(\frac{1}{2}(y_i - u)^2 + \lambda \frac{1}{|y_i|^\gamma} |u| \right) \quad \text{for } i = 1, 2, \ldots, n. \tag{5}$$

Therefore, $\hat{\mu}_i^*(\lambda) = (|y_i| - \frac{\lambda}{|y_i|^\gamma})_+ \operatorname{sgn}(y_i)$.

Antoniadis and Fan (2001) considered thresholding estimates from penalized linear squares,

$$\arg\min_u \left(\frac{1}{2}(y_i - u)^2 + \lambda J(|u|) \right) \quad \text{for } i = 1, 2, \ldots, n,$$

where J is the penalty function. The L_0 penalty leads to the hard-thresholding rule, whereas the lasso penalty yields the soft-thresholding rule. Bridge thresholding comes from $J(|u|) = |u|^q$ $(0 < q < 1)$. Figure 1 compares the adaptive lasso shrinkage with other popular thresholding rules. Note that the shrinkage rules are discontinuous in hard thresholding and the bridge with an $L_{.5}$ penalty. The lasso shrinkage is continuous but pays a price in estimating the large coefficients due to a constant shift. As continuous thresholding rules, the SCAD and the adaptive lasso are able to eliminate the bias in the lasso. We establish an oracle inequality on the risk bound of the adaptive lasso.

Theorem 3 (Oracle inequality). Let $\lambda = (2 \log n)^{(1+\gamma)/2}$, then

$$R(\hat{\mu}^*(\lambda)) \le \left(2 \log n + 5 + \frac{4}{\gamma} \right) \times \left(R(\text{ideal}) + \frac{1}{2\sqrt{\pi}} (\log n)^{-1/2} \right). \tag{6}$$

By theorem 3 of Donoho and Johnstone (1994, sec. 2.1), we know that

$$\inf_{\hat{\mu}} \sup_{\mu} \frac{R(\hat{\mu})}{1 + R(\text{ideal})} \sim 2 \log n.$$

Therefore, the oracle inequality (6) indicates that the adaptive lasso shrinkage estimator attains the near-minimax risk.

3.4 $\gamma = 1$ and the Nonnegative Garotte

The nonnegative garotte (Breiman 1995) finds a set of nonnegative scaling factors $\{c_j\}$ to minimize

$$\left\| \mathbf{y} - \sum_{j=1}^p \mathbf{x}_j \hat{\beta}_j(\text{ols}) c_j \right\|^2 + \lambda_n \sum_{j=1}^p c_j \quad \text{subject to} \quad c_j \ge 0 \qquad \forall j. \tag{7}$$

The garotte estimates are $\hat{\beta}_j(\text{garotte}) = c_j \hat{\beta}_j(\text{ols})$. A sufficiently large λ_n shrinks some c_j to exact 0. Because of this nice property, the nonnegative garotte is often used for sparse modeling. Studying the consistency of the garotte selection is of interest. To the best of our knowledge, no answer has yet been reported in the literature. In this section we show that variable selection by the nonnegative garotte is indeed consistent. [Yuan and Lin (2005) independently proved the consistency of the nonnegative garotte selection.]

We first show that the nonnegative garotte is closely related to a special case of the adaptive lasso. Suppose that $\gamma = 1$ and we choose the adaptive weights $\hat{\mathbf{w}} = 1/|\hat{\boldsymbol{\beta}}(\text{ols})|$; then the adaptive lasso solves

$$\arg\min_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{j=1}^p \mathbf{x}_j \beta_j \right\|^2 + \lambda_n \sum_{j=1}^p \frac{|\beta_j|}{|\hat{\beta}_j(\text{ols})|}. \tag{8}$$

We see that (7) looks very similar to (8). In fact, they are almost identical. Because $c_j = \hat{\beta}_j(\text{garotte}) / \hat{\beta}_j(\text{ols})$, we can equivalently reformulate the garotte as

$$\arg\min_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{j=1}^p \mathbf{x}_j \beta_j \right\|^2 + \lambda_n \sum_{j=1}^p \frac{|\beta_j|}{|\hat{\beta}_j(\text{ols})|}, \quad \text{subject to} \quad \beta_j \hat{\beta}_j(\text{ols}) \ge 0 \qquad \forall j. \tag{9}$$

Hence the nonnegative garotte can be considered the adaptive lasso ($\gamma = 1$) with additional sign constraints in (9). Furthermore, we can show that with slight modifications (see the App.), the proof of Theorem 2 implies consistency of the nonnegative garotte selection.

Corollary 2. In (7), if we choose a λ_n such that $\lambda_n / \sqrt{n} \to 0$ and $\lambda_n \to \infty$, then the nonnegative garotte is consistent for variable selection.

3.5 Computations

In this section we discuss the computational issues. First, the adaptive lasso estimates in (4) can be solved by the LARS algorithm (Efron et al. 2004). The computation details are given in Algorithm 1, the proof of which is very simple and so is omitted.

Algorithm 1 (The LARS algorithm for the adaptive lasso).

1. Define $\mathbf{x}_j^{**} = \mathbf{x}_j / \hat{w}_j$, $j = 1, 2, \ldots, p$.
2. Solve the lasso problem for all λ_n,
$$\hat{\boldsymbol{\beta}}^{**} = \arg\min_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{j=1}^p \mathbf{x}_j^{**} \beta_j \right\|^2 + \lambda_n \sum_{j=1}^p |\beta_j|.$$
3. Output $\hat{\beta}_j^{*(n)} = \hat{\beta}_j^{**} / \hat{w}_j$, $j = 1, 2, \ldots, p$.

The LARS algorithm is used to compute the entire solution path of the lasso in step (b). The computational cost is of order $O(np^2)$, which is the same order of computation of a single OLS fit. The efficient path algorithm makes the adaptive lasso an attractive method for real applications.

Tuning is an important issue in practice. Suppose that we use $\hat{\boldsymbol{\beta}}(\text{ols})$ to construct the adaptive weights in the adaptive lasso; we then want to find an optimal pair of (γ, λ_n). We can use two-dimensional cross-validation to tune the adaptive lasso. Note that for a given γ, we can use cross-validation along with the LARS algorithm to exclusively search for the optimal λ_n. In principle, we can also replace $\hat{\boldsymbol{\beta}}(\text{ols})$ with other consistent estimators. Hence we can treat it as the third tuning parameter and perform three-dimensional cross-validation to find an optimal triple $(\hat{\boldsymbol{\beta}}, \gamma, \lambda_n)$. We suggest using $\hat{\boldsymbol{\beta}}(\text{ols})$ unless collinearity is a concern, in which case we can try $\hat{\boldsymbol{\beta}}(\text{ridge})$ from the best ridge regression fit, because it is more stable than $\hat{\boldsymbol{\beta}}(\text{ols})$.

3.6 Standard Error Formula

We briefly discuss computing the standard errors of the adaptive lasso estimates. Tibshirani (1996) presented a standard error formula for the lasso. Fan and Li (2001) showed that local quadratic approximation (LQA) can provide a sandwich formula for computing the covariance of the penalized estimates of the nonzero components. The LQA sandwich formula has been proven to be consistent (Fan and Peng 2004).

We follow the LQA approach to derive a sandwich formula for the adaptive lasso. For a nonzero β_j, consider the LQA of the adaptive lasso penalty,

$$|\beta_j|\hat{w}_j \approx |\beta_{j0}|\hat{w}_j + \frac{1}{2}\frac{\hat{w}_j}{|\beta_{j0}|}(\beta_j^2 - \beta_{j0}^2).$$

Suppose that the first d components of $\boldsymbol{\beta}$ are nonzero. Then let $\boldsymbol{\Sigma}(\boldsymbol{\beta}) = \text{diag}(\frac{\hat{w}_1}{|\beta_1|}, \ldots, \frac{\hat{w}_d}{|\beta_d|})$. Let $\mathbf{X}_d$ denote the first d columns of $\mathbf{X}$. By the arguments of Fan and Li (2001), the adaptive lasso estimates can be solved by iteratively computing the ridge regression,

$$(\beta_1, \ldots, \beta_d)^T = \left(\mathbf{X}_d^T\mathbf{X}_d + \lambda_n\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\right)^{-1}\mathbf{X}_d^T\mathbf{y},$$

which leads to the estimated covariance matrix for the nonzero components of the adaptive lasso estimates $\hat{\boldsymbol{\beta}}^{*(n)}$,

$$\widehat{\text{cov}}\left(\hat{\boldsymbol{\beta}}^{*(n)}_{\mathcal{A}_n^*}\right) = \sigma^2\left(\mathbf{X}^T_{\mathcal{A}_n^*}\mathbf{X}_{\mathcal{A}_n^*} + \lambda_n\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\beta}}^{*(n)}_{\mathcal{A}_n^*}\right)\right)^{-1} \times \mathbf{X}^T_{\mathcal{A}_n^*}\mathbf{X}_{\mathcal{A}_n^*}\left(\mathbf{X}^T_{\mathcal{A}_n^*}\mathbf{X}_{\mathcal{A}_n^*} + \lambda_n\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\beta}}^{*(n)}_{\mathcal{A}_n^*}\right)\right)^{-1}.$$

If σ^2 is unknown, then we can replace σ^2 with its estimates from the full model. For variables with $\hat{\beta}_j^{*(n)} = 0$, the estimated standard errors are 0 (Tibshirani 1996; Fan and Li 2001).

3.7 Some Numerical Experiments

In this section we report a simulation study done to compare the adaptive lasso with the lasso, the SCAD, and the nonnegative garotte. In the simulation we considered various linear models, $y = \mathbf{x}^T\boldsymbol{\beta} + \text{N}(0, \sigma^2)$. In all examples, we computed the adaptive weights using OLS coefficients. We used the LARS algorithm to compute the lasso and the adaptive lasso. We implemented the LQA algorithm of Fan and Li (2001) to compute the SCAD estimates and used quadratic programming to solve the nonnegative garotte. For each competitor, we selected its tuning parameter by fivefold cross-validation. In the adaptive lasso, we used two-dimensional cross-validation and selected γ from $\{.5, 1, 2\}$; thus the difference between the lasso and the adaptive lasso must be contributed by the adaptive weights.

We first show a numerical demonstration of Corollary 1.

Model 0 (Inconsistent lasso path). We let $y = \mathbf{x}^T\boldsymbol{\beta} + \text{N}(0, \sigma^2)$, where the true regression coefficients are $\boldsymbol{\beta} = (5.6, 5.6, 5.6, 0)$. The predictors $\mathbf{x}_i$ $(i = 1, \ldots, n)$ are iid $\text{N}(\mathbf{0}, \mathbf{C})$, where $\mathbf{C}$ is the $\mathbf{C}$ matrix in Corollary 1 with $\rho_1 = -.39$ and $\rho_2 = .23$.

In this model we chose $\rho_1 = -.39$ and $\rho_2 = .23$ such that the conditions in Corollary 1 are satisfied. Thus the design matrix $\mathbf{C}$ does not allow consistent lasso selection. To show this numerically, we simulated 100 datasets from the foregoing model for three different combinations of sample size (n) and error variance (σ^2). On each dataset, we computed the entire solution path of the lasso, then estimated the probability of the lasso solution path containing the true model. We repeated the same procedure for the adaptive lasso. As n increases and σ decreases, the variable selection problem is expected to become easier. However, as shown in Table 1, the lasso has about a 50% chance of missing the true model regardless of the choice of (n, σ). In contrast, the adaptive lasso is consistent in variable selection.

We now compare the prediction accuracy of the lasso, the adaptive lasso, the SCAD, and the nonnegative garotte. Note that $E[(\hat{y} - y_{\text{test}})^2] = E[(\hat{y} - \mathbf{x}^T\boldsymbol{\beta})^2] + \sigma^2$. The second term is the inherent prediction error due to the noise. Thus for comparison, we report the relative prediction error (RPE), $E[(\hat{y} - \mathbf{x}^T\boldsymbol{\beta})^2]/\sigma^2$.

Model 1 (A few large effects). In this example, we let $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)$. The predictors $\mathbf{x}_i$ $(i = 1, \ldots, n)$ were iid normal vectors. We set the pairwise correlation between $\mathbf{x}_{j_1}$ and $\mathbf{x}_{j_2}$ to be $\text{cor}(j_1, j_2) = (.5)^{|j_1 - j_2|}$. We also set $\sigma = 1, 3, 6$ such that the corresponding signal-to-noise ratio (SNR) was about 21.25, 2.35, and .59. We let n be 20 and 60.

Model 2 (Many small effects). We used the same model as in model 1 but with $\beta_j = .85$ for all j. We set $\sigma = 1, 3, 6$, and the corresponding SNR is 14.46, 1.61, and .40. We let $n = 40$ and $n = 80$.

In both models we simulated 100 training datasets for each combination of (n, σ^2). All of the training and tuning were done on the training set. We also collected independent test datasets of 10,000 observations to compute the RPE. To estimate the standard error of the RPE, we generated a bootstrapped sample from the 100 RPEs, then calculated the bootstrapped

Table 1. Simulation Model 0: The Probability of Containing the True Model in the Solution Path

	$n = 60, \sigma = 9$	$n = 120, \sigma = 5$	$n = 300, \sigma = 3$
lasso	.55	.51	.53
adalasso($\gamma = .5$)	.59	.68	.93
adalasso($\gamma = 1$)	.67	.89	1
adalasso($\gamma = 2$)	.73	.97	1
adalasso(γ by cv)	.67	.91	1

NOTE: In this table "adalasso" is the adaptive lasso, and "γ by cv" means that γ was selected by five-fold cross-validation from three choices: $\gamma = .5$, $\gamma = 1$, and $\gamma = 2$.

Table 2. Simulation Models 1 and 2, Comparing the Median RPE Based on 100 Replications

	$\sigma = 1$	$\sigma = 3$	$\sigma = 6$
Model 1 ($n = 20$)			
Lasso	$.414_{(.046)}$	$.395_{(.039)}$	$.275_{(.026)}$
Adaptive lasso	$.261_{(.023)}$	$.369_{(.029)}$	$.336_{(.031)}$
SCAD	$.218_{(.029)}$	$.508_{(.044)}$	$.428_{(.019)}$
Garotte	$.227_{(.007)}$	$.488_{(.043)}$	$.385_{(.030)}$
Model 1 ($n = 60$)			
Lasso	$.103_{(.008)}$	$.102_{(.008)}$	$.107_{(.012)}$
Adaptive lasso	$.073_{(.004)}$	$.094_{(.012)}$	$.117_{(.008)}$
SCAD	$.053_{(.008)}$	$.104_{(.016)}$	$.119_{(.014)}$
Garotte	$.069_{(.006)}$	$.102_{(.008)}$	$.118_{(.009)}$
Model 2 ($n = 40$)			
Lasso	$.205_{(.015)}$	$.214_{(.014)}$	$.161_{(.009)}$
Adaptive lasso	$.203_{(.015)}$	$.237_{(.016)}$	$.190_{(.008)}$
SCAD	$.223_{(.018)}$	$.297_{(.028)}$	$.230_{(.009)}$
Garotte	$.199_{(.018)}$	$.273_{(.024)}$	$.219_{(.019)}$
Model 2 ($n = 80$)			
Lasso	$.094_{(.008)}$	$.096_{(.008)}$	$.091_{(.008)}$
Adaptive lasso	$.093_{(.007)}$	$.094_{(.007)}$	$.104_{(.009)}$
SCAD	$.096_{(.104)}$	$.099_{(.012)}$	$.138_{(.014)}$
Garotte	$.095_{(.006)}$	$.111_{(.007)}$	$.119_{(.006)}$

NOTE: The numbers in parentheses are the corresponding standard errors (of RPE).

sample median. We repeated this process 500 times. The estimated standard error was the standard deviation of the 500 bootstrapped sample medians.

Table 2 summarizes the simulation results. Several observations can be made from this table. First, we see that the lasso performs best when the SNR is low but the oracle methods tend to be more accurate when the SNR is high. This phenomenon is more evident in model 1. Second, the adaptive lasso seems to be able to adaptively combine the strength of the lasso and the SCAD. With a medium or low level of SNR, the adaptive lasso often outperforms the SCAD and the garotte. With a high SNR, the adaptive lasso significantly dominates the lasso by a good margin. The adaptive lasso tends to be more stable than the SCAD. The overall performance of the adaptive lasso appears to be the best. Finally, our simulations also show that each method has its unique merits; none of the four methods can universally dominate the other three competitors. We also considered model 2 with $\beta_j^* = .25$ and obtained the same conclusions.

Table 3 presents the performance of the four methods in variable selection. All the four methods can correctly identify the three significant variables (1, 2, and 5). The lasso tends to select two noise variables into the final model. The other three methods select less noise variables. It is worth noting that these good variable selection results are achieved with very moderate sample sizes.

Table 3. Median Number of Selected Variables for Model 1 With n = 60

	$\sigma = 1$		$\sigma = 3$	
	C	I	C	I
Truth	3	0	3	0
Lasso	3	2	3	2
Adaptive lasso	3	1	3	1
SCAD	3	0	3	1
Garotte	3	1	3	1.5

NOTE: The column labeled "C" gives the number of selected nonzero components, and the column labeled "I" presents the number of zero components incorrectly selected into the final model.

Table 4. Standard Errors of the Adaptive Lasso Estimates for Model 1 With n = 60 and $\sigma = 1$

$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
SD_{true}	SD_{est}	SD_{true}	SD_{est}	SD_{true}	SD_{est}
.153	$.152_{(.016)}$	.167	$.154_{(.019)}$	.159	$.135_{(.018)}$

We now test the accuracy of the standard error formula. Table 4 presents the results for nonzero coefficients when $n = 60$ and $\sigma = 1$. We computed the true standard errors, denoted by SD_{true}, by the 100 simulated coefficients. We denoted by SD_{est} the average of estimated standard errors in the 100 simulations; the simulation results indicate that the standard error formula works quite well for the adaptive lasso.

4. FURTHER EXTENSIONS

4.1 The Exponential Family and Generalized Linear Models

Having shown the oracle properties of the adaptive lasso in linear regression models, we would like to further extend the theory and methodology to generalized linear models (GLMs). We consider the penalized log-likelihood estimation using the adaptively weighted ℓ_1 penalty, where the likelihood belongs to the exponential family with canonical parameter θ. The generic density form can be written as (McCullagh and Nelder 1989)

$$f(y|\mathbf{x}, \theta) = h(y)\exp(y\theta - \phi(\theta)). \tag{10}$$

Generalized linear models assume that $\theta = \mathbf{x}^T\boldsymbol{\beta}^*$.

Suppose that $\hat{\boldsymbol{\beta}}(\text{mle})$ is the maximum likelihood estimates in the GLM. We construct the weight vector $\hat{\mathbf{w}} = 1/|\hat{\boldsymbol{\beta}}(\text{mle})|^\gamma$ for some $\gamma > 0$. The adaptive lasso estimates $\hat{\boldsymbol{\beta}}^{*(n)}(\text{glm})$ are given by

$$\hat{\boldsymbol{\beta}}^{*(n)}(\text{glm}) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}\bigl(-y_i(\mathbf{x}_i^T\boldsymbol{\beta}) + \phi(\mathbf{x}_i^T\boldsymbol{\beta})\bigr) + \lambda_n\sum_{j=1}^{p}\hat{w}_j|\beta_j|. \tag{11}$$

For logistic regression, the foregoing equation becomes

$$\hat{\boldsymbol{\beta}}^{*(n)}(\text{logistic}) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}\bigl(-y_i(\mathbf{x}_i^T\boldsymbol{\beta}) + \log\bigl(1 + e^{\mathbf{x}_i^T\boldsymbol{\beta}}\bigr)\bigr) + \lambda_n\sum_{j=1}^{p}\hat{w}_j|\beta_j|. \tag{12}$$

In Poisson log-linear regression models, (11) can be written as

$$\hat{\boldsymbol{\beta}}^{*(n)}(\text{poisson}) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}\bigl(-y_i(\mathbf{x}_i^T\boldsymbol{\beta}) + \exp(\mathbf{x}_i^T\boldsymbol{\beta})\bigr) + \lambda_n\sum_{j=1}^{p}\hat{w}_j|\beta_j|. \tag{13}$$

Assume that the true model has a sparse representation. Without loss of generality, let $\mathcal{A} = \{j : \beta_j^* \neq 0\} = \{1, 2, \ldots, p_0\}$ and $p_0 < p$. We write the Fisher information matrix

$$\mathbf{I}(\boldsymbol{\beta}^*) = \begin{bmatrix} \mathbf{I}_{11} & \mathbf{I}_{12} \\ \mathbf{I}_{21} & \mathbf{I}_{22} \end{bmatrix},$$

where $\mathbf{I}_{11}$ is a $p_0 \times p_0$ matrix. Then $\mathbf{I}_{11}$ is the Fisher information with the true submodel known. We show that under some mild regularity conditions (see the App.), the adaptive lasso estimates $\hat{\boldsymbol{\beta}}^{*(n)}(\text{glm})$ enjoys the oracle properties if λ_n is chosen appropriately.

Theorem 4. Let $\mathcal{A}_n^* = \{j : \hat{\beta}_j^{*(n)}(\text{glm}) \neq 0\}$. Suppose that $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{(\gamma-1)/2} \to \infty$; then, under some mild regularity conditions, the adaptive lasso estimates $\hat{\boldsymbol{\beta}}^{*(n)}(\text{glm})$ must satisfy the following:

1. Consistency in variable selection: $\lim_n P(\mathcal{A}_n^* = \mathcal{A}) = 1$
2. Asymptotic normality: $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{*(n)}(\text{glm}) - \boldsymbol{\beta}_{\mathcal{A}}^*) \to_d \text{N}(\mathbf{0}, \mathbf{I}_{11}^{-1})$.

The solution path of (11) is no longer piecewise linear, because the negative log-likelihood is not piecewise quadratic (Rosset and Zhu 2004). However, we can use the Newton–Raphson method to solve $\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{*(n)}(\text{glm})$. Note that the LQA of the adaptive lasso penalty is given in Section 3.6. Then an iterative LQA algorithm for solving (11) can be constructed by following the generic recipe of Fan and Li (2001). Because the negative log-likelihood in GLM is convex, the convergence analysis of the LQA of Hunter and Li (2005) indicates that the iterative LQA algorithm is able to find the unique minimizer of (11).

We illustrate the methodology using the logistic regression model of Hunter and Li (2005). In this example, we simulated 100 datasets consisting of 200 observations from the model $\mathbf{y} \sim \text{Bernoulli}\{p(\mathbf{x}^T\boldsymbol{\beta})\}$, where $p(u) = \exp(u)/(1+\exp(u))$ and $\boldsymbol{\beta} = (3, 0, 0, 1.5, 0, 0, 7, 0, 0)$. The components of $\mathbf{x}$ are standard normal, where the correlation between x_i and x_j is $\rho = .75$. We compared the lasso, the SCAD, and the adaptive lasso. We computed the misclassification error of each competitor by Monte Carlo using a test dataset consisting of 10,000 observations. Because the Bayes error is the lower bound for the misclassification error, we define the RPE of a classifier δ as $\text{RPE}(\delta) = (\text{misclassification error of } \delta)/(\text{Bayes error}) - 1$. Figure 2 compares the RPEs of the lasso, the SCAD, and the adaptive lasso over 100 simulations. The adaptive lasso does the best, followed by the SCAD and then the lasso. The lasso and the adaptive lasso seem to be more stable than the SCAD. Overall, all three methods have excellent prediction accuracy.

4.2 High-Dimensional Data

We have considered the typical asymptotic setup in which the number of predictors is fixed and the sample size approaches infinity. The asymptotic theory with $p = p_n \to \infty$ seems to be more applicable to problems involving a huge number of predictors, such as microarrays analysis and document/image classification. In Section 3.3 we discussed the near-minimax optimality of the adaptive lasso when $p_n = n$ and the predictors are orthogonal. In other high-dimensional problems (e.g., microarrays), we may want to consider $p_n > n \to \infty$; then it is nontrivial to find a consistent estimate for constructing the weights in the adaptive lasso. A practical solution is to use the ℓ_2 penalized estimator. Thus the adaptive lasso can be well defined. Note that one more tuning parameter—the ℓ_2 regularization parameter—is included in the procedure. It remains to show that the ℓ_2-penalized estimates are consistent and that the corresponding adaptive lasso estimates have the desired asymptotic properties. In ongoing work, we plan to investigate the asymptotic property of the adaptive lasso in the high-dimensional setting. There is some related work in the literature. Donoho (2004) studied the minimum ℓ_1 solution for large underdetermined systems of equations, Meinshausen and Bühlmann (2004) considered the high-dimensional lasso, and Fan and Peng (2004) proved the oracle properties of the SCAD estimator with a diverging number of predictors. These results should prove very useful in the investigation of the asymptotic properties of the adaptive lasso with a diverging number of predictors. It is also worth noting that when considering $p_n \gg n \to \infty$, we can allow the magnitude of the nonzero coefficients to vary with n; however, to keep the oracle properties, we cannot let the magnitude go to 0 too fast. Fan and Peng (2004) discussed this issue [see their condition (H)].

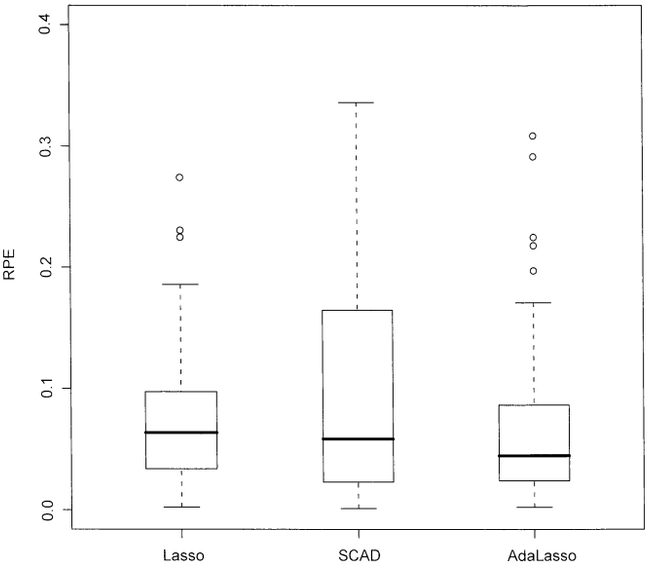

Figure 2. Simulation Example: Logistic Regression Model.

5. CONCLUSION

In this article we have proposed the adaptive lasso for simultaneous estimation and variable selection. We have shown that although the lasso variable selection can be inconsistent in some scenarios, the adaptive lasso enjoys the oracle properties by utilizing the adaptively weighted ℓ_1 penalty. The adaptive lasso shrinkage also leads to a near–minimax-optimal estimator. Owing to the efficient path algorithm, the adaptive lasso enjoys the computational advantage of the lasso. Our simulation has shown that the adaptive lasso compares favorably with other sparse modeling techniques. It is worth emphasizing that the oracle properties do not automatically result in optimal prediction performance. The lasso can be advantageous in difficult prediction problems. Our results offer new insights into the ℓ_1-related methods and support the use of the ℓ_1 penalty in statistical modeling.

APPENDIX: PROOFS

Proof of Proposition 1

Note that $\mathcal{A}_n = \mathcal{A}$ implies that $\hat{\beta}_j = 0$ for all $j \notin \mathcal{A}$. Let $\mathbf{u}^* = \arg\min(V_2(\mathbf{u}))$. Note that $P(\mathcal{A}_n = \mathcal{A}) \leq P(\sqrt{n}\hat{\beta}_j = 0 \ \forall j \notin \mathcal{A})$.

Lemma 2 shows that $\sqrt{n}\hat{\boldsymbol{\beta}}_{\mathcal{A}} \to_d \mathbf{u}^*_{\mathcal{A}}$; thus the weak convergence result indicates that $\limsup_n P(\sqrt{n}\hat{\beta}_j = 0\ \forall j \notin \mathcal{A}) \le P(u^*_j = 0\ \forall j \notin \mathcal{A})$. Therefore, we need only show that $c = P(u^*_j = 0\ \forall j \notin \mathcal{A}) < 1$. There are two cases:

Case 1. $\lambda_0 = 0$. Then it is easy to see that $\mathbf{u}^* = \mathbf{C}^{-1}\mathbf{W} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{C}^{-1})$, and so $c = 0$.

Case 2. $\lambda_0 > 0$. Then $V_2(\mathbf{u})$ is not differentiable at $u_j = 0\ \forall j \in \mathcal{A}$. By the Karush–Kuhn–Tucker (KKT) optimality condition, we have

$$-2W_j + 2(\mathbf{C}\mathbf{u}^*)_j + \lambda_0 \operatorname{sgn}(\beta^*_j) = 0 \qquad \forall j \in \mathcal{A} \tag{A.1}$$

and

$$|-2W_j + 2(\mathbf{C}\mathbf{u}^*)_j| \le \lambda_0 \qquad \forall j \notin \mathcal{A}. \tag{A.2}$$

If $u^*_j = 0$ for all $j \notin \mathcal{A}$, then (A.1) and (A.2) become

$$-2\mathbf{W}_{\mathcal{A}} + 2\mathbf{C}_{11}\mathbf{u}^*_{\mathcal{A}} + \lambda_0 \operatorname{sgn}(\boldsymbol{\beta}^*_{\mathcal{A}}) = \mathbf{0} \tag{A.3}$$

and

$$|-2\mathbf{W}_{\mathcal{A}^c} + 2\mathbf{C}_{21}\mathbf{u}^*_{\mathcal{A}}| \le \lambda_0 \quad \text{componentwise.} \tag{A.4}$$

Combining (A.3) and (A.4) gives

$$\left|-2\mathbf{W}_{\mathcal{A}^c} + \mathbf{C}_{21}\mathbf{C}_{11}^{-1}(2\mathbf{W}_{\mathcal{A}} - \lambda_0 \operatorname{sgn}(\boldsymbol{\beta}^*_{\mathcal{A}})\right| \le \lambda_0 \quad \text{componentwise.}$$

Thus $c \le P(|-2\mathbf{W}_{\mathcal{A}^c} + \mathbf{C}_{21}\mathbf{C}_{11}^{-1}(2\mathbf{W}_{\mathcal{A}} - \lambda_0 \operatorname{sgn}(\boldsymbol{\beta}^*_{\mathcal{A}})| \le \lambda_0) < 1$.

Proof of Lemma 3

Let $\boldsymbol{\beta} = \boldsymbol{\beta}^* + \frac{\lambda_n}{n}\mathbf{u}$. Define

$$\Phi(\mathbf{u}) = \left\|\mathbf{y} - \sum_{j=1}^{p}\mathbf{x}_j\left(\beta^*_j + \frac{\lambda_n}{n}u_j\right)\right\|^2 + \lambda_n\sum_{j=1}^{p}\left|\beta^*_j + \frac{\lambda_n}{n}u_j\right|.$$

Suppose that $\hat{\mathbf{u}}_n = \arg\min \Phi(\mathbf{u})$; then $\hat{\boldsymbol{\beta}}^{(n)} = \boldsymbol{\beta}^* + \frac{\lambda_n}{n}\hat{\mathbf{u}}_n$ or $\hat{\mathbf{u}}_n = \frac{n}{\lambda_n}(\boldsymbol{\beta}^{(n)} - \boldsymbol{\beta}^*)$. Note that $\Phi(\mathbf{u}) - \Phi(\mathbf{0}) = \frac{\lambda_n^2}{n}V_3^{(n)}(\mathbf{u})$, where

$$V_3^{(n)}(\mathbf{u}) \equiv \mathbf{u}^T\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)\mathbf{u} - 2\frac{\boldsymbol{\epsilon}^T\mathbf{X}}{\sqrt{n}}\frac{\sqrt{n}}{\lambda_n}\mathbf{u} + \sum_{j=1}^{p}\frac{n}{\lambda_n}\left(\left|\beta^*_j + \frac{\lambda_n}{n}u_j\right| - |\beta^*_j|\right).$$

Hence $\hat{\mathbf{u}}_n = \arg\min V_3^{(n)}$. Because $\frac{1}{n}\mathbf{X}^T\mathbf{X} \to \mathbf{C}$, $\frac{\boldsymbol{\epsilon}^T\mathbf{X}}{\sqrt{n}}$ is $O_p(1)$. Then $\frac{\lambda_n}{\sqrt{n}} \to \infty$ implies that $\frac{\boldsymbol{\epsilon}^T\mathbf{X}}{\sqrt{n}}\frac{\sqrt{n}}{\lambda_n}\mathbf{u} \to_p 0$ by Slutsky's theorem. If $\beta_j \ne 0$, then $\frac{n}{\lambda_n}(|\beta^*_j + \frac{\lambda_n}{n}u_j| - |\beta^*_j|)$ converges to $u_j\operatorname{sgn}(\beta_j)$; it equals $|u_j|$ otherwise. Therefore, we have $V_3^{(n)}(\mathbf{u}) \to_p V_3(\mathbf{u})$ for every $\mathbf{u}$. Because $\mathbf{C}$ is a positive definite matrix, $V_3(\mathbf{u})$ has a unique minimizer. $V_3^{(n)}$ is a convex function. Then it follows (Geyer 1994) that $\hat{\mathbf{u}}_n = \arg\min(V_3^{(n)}) \to_p \arg\min(V_3)$.

Proof of Theorem 1

We first assume that the limits of $\lambda_n/\sqrt{n}$ and λ_n/n exist. In Proposition 1 we showed that if $\lambda_n/\sqrt{n} \to \lambda_0 \ge 0$, then the lasso selection cannot be consistent. If the lasso selection is consistent, then one of three scenarios must occur: (1) $\lambda_n/n \to \infty$; (2) $\lambda_n/n \to \lambda_0$, $0 < \lambda_0 < \infty$; or (3) $\lambda_n/n \to 0$ but $\lambda_n/\sqrt{n} \to \infty$.

If scenario (1) occurs, then it is easy to check that $\hat{\beta}^{(n)}_j \to_p 0$ for all $j = 1, 2, \ldots, p$, which obviously contradicts the consistent selection assumption.

Suppose that scenario (2) occurs. By Lemma 1, $\hat{\boldsymbol{\beta}}^{(n)} \to_p \boldsymbol{\beta}_*$, and $\boldsymbol{\beta}_*$ is a nonrandom vector. Because $P(\mathcal{A}_n = \mathcal{A}) \to 1$, we must have $\beta_{*j} = 0$ for all $j \notin \mathcal{A}$. Pick a $j \in \mathcal{A}$ and consider the event $j \in \mathcal{A}_n$. By the KKT optimality conditions, we have

$$-2\mathbf{x}_j^T\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(n)}\right) + \lambda_n \operatorname{sgn}\left(\hat{\beta}^{(n)}_j\right) = 0.$$

Hence $P(j \in \mathcal{A}_n) \le P(|-2\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(n)})/n| = \lambda_n/n)$. Moreover, note that

$$-2\frac{\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(n)})}{n} = -2\frac{\mathbf{x}_j^T\mathbf{X}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^{(n)})}{n} - 2\frac{\mathbf{x}_j^T\boldsymbol{\epsilon}}{n}$$

$$\to_p -2(\mathbf{C}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_*))_j.$$

Thus $P(j \in \mathcal{A}_n) \to 1$ implies that $|2(\mathbf{C}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_*))_j| = \lambda_0$. Similarly, pick a $j' \notin \mathcal{A}$; then $P(j' \notin \mathcal{A}_n) \to 1$. Consider the event $j' \notin \mathcal{A}_n$. By the KKT conditions, we have

$$\left|-2\mathbf{x}_{j'}^T\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(n)}\right)\right| \le \lambda_n.$$

So $P(j' \notin \mathcal{A}_n) \le P(|-2\mathbf{x}_{j'}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(n)})/n| \le \lambda_n/n)$. Thus $P(j' \notin \mathcal{A}_n) \to 1$ implies that $|2(\mathbf{C}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_*))_{j'}| \le \lambda_0$. Observe that

$$\mathbf{C}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_*) = \begin{bmatrix}\mathbf{C}_{11}(\boldsymbol{\beta}^*_{\mathcal{A}} - \boldsymbol{\beta}_{*\mathcal{A}}) \\ \mathbf{C}_{21}(\boldsymbol{\beta}^*_{\mathcal{A}} - \boldsymbol{\beta}_{*\mathcal{A}})\end{bmatrix}.$$

We have

$$\mathbf{C}_{11}(\boldsymbol{\beta}^*_{\mathcal{A}} - \boldsymbol{\beta}_{*\mathcal{A}}) = \frac{\lambda_0}{2}\mathbf{s}_* \tag{A.5}$$

and

$$|\mathbf{C}_{21}(\boldsymbol{\beta}^*_{\mathcal{A}} - \boldsymbol{\beta}_{*\mathcal{A}})| \le \frac{\lambda_0}{2}, \tag{A.6}$$

where $\mathbf{s}_*$ is the sign vector of $\mathbf{C}_{11}(\boldsymbol{\beta}^*_{\mathcal{A}} - \boldsymbol{\beta}_{*\mathcal{A}})$. Combining (A.5) and (A.6), we have $|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\frac{\lambda_0}{2}\mathbf{s}_*| \le \frac{\lambda_0}{2}$ or, equivalently,

$$|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{s}_*| \le 1. \tag{A.7}$$

If scenario (3) occurs, then, by Lemma 3, $\frac{n}{\lambda_n}(\hat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta}^*) \to_p \mathbf{u}^* = \arg\min(V_3)$, and $\mathbf{u}^*$ is a nonrandom vector. Pick any $j \notin \mathcal{A}$. Because $P(\hat{\beta}^{(n)}_j = 0) \to 1$ and $\frac{n}{\lambda_n}\hat{\beta}^{(n)}_j \to_p u^*_j$, we must have $u^*_j = 0$. On the other hand, note that

$$V_3(\mathbf{u}) = \mathbf{u}^T\mathbf{C}\mathbf{u} + \sum_{j\in\mathcal{A}}[u_j s^*_j] + \sum_{j\notin\mathcal{A}}|u_j|,$$

where $\mathbf{s}^* = \operatorname{sgn}(\boldsymbol{\beta}^*_{\mathcal{A}})$. We get $\mathbf{u}^*_{\mathcal{A}} = -\mathbf{C}_{11}^{-1}(\mathbf{C}_{12}\mathbf{u}^*_{\mathcal{A}^c} + \frac{1}{2}\mathbf{s}^*)$. Then it is straightforward to verify that $\mathbf{u}^*_{\mathcal{A}^c} = \arg\min(Z)$, where

$$Z(\mathbf{r}) = \mathbf{r}^T(\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12})\mathbf{r} - \mathbf{r}^t\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{s}^* + \sum_i |r_i|.$$

But $\mathbf{u}^*_{\mathcal{A}^c} = \mathbf{0}$. By the KKT optimality conditions, we must have

$$|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{s}^*| \le 1. \tag{A.8}$$

Together (A.7) and (A.8) prove (3).

Now we consider the general sequences of $\{\lambda_n/\sqrt{n}\}$ and $\{\lambda_n/n\}$. Note that there is a subsequence $\{n_k\}$ such that the limits of $\{\lambda_{n_k}/\sqrt{n_k}\}$ and $\{\lambda_{n_k}/n_k\}$ exist (with the limits allowed to be infinity). Then we can apply the foregoing proof to the subsequence $\{\hat{\boldsymbol{\beta}}^{*(n_k)}\}$ to obtain the same conclusion.

Proof of Corollary 1

Note that $\mathbf{C}_{11}^{-1} = \frac{1}{1-\rho_1}(\mathbf{I} - \frac{\rho_1}{1+(p_0-1)\rho_1}\mathbf{J}_1)$ and $\mathbf{C}_{21}\mathbf{C}_{11}^{-1} = \frac{\rho_2}{1+(p_0-1)\rho_1}(\vec{1})^T$. Thus $\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{s} = \frac{\rho_2}{1+(p_0-1)\rho_1}(\sum_j^{p_0} s_j)\vec{1}$. Then condition (3) becomes

$$\left|\frac{\rho_2}{1+(p_0-1)\rho_1}\right| \cdot \left|\sum_j^{p_0} s_j\right| \le 1. \tag{A.9}$$

Observe that when p_0 is an odd number, $|\sum_j^{p_0} s_j| \ge 1$. If $|\frac{\rho_2}{1+(p_0-1)\rho_1}| > 1$, then (A.9) cannot be satisfied for any sign vector $\mathbf{s}$. The choice of (ρ_1, ρ_2) in Corollary 1 ensures that $\mathbf{C}$ is a positive definite matrix and $|\frac{\rho_2}{1+(p_0-1)\rho_1}| > 1$.

Proof of Theorem 2

We first prove the asymptotic normality part. Let $\boldsymbol{\beta} = \boldsymbol{\beta}^* + \frac{\mathbf{u}}{\sqrt{n}}$ and

$$\Psi_n(\mathbf{u}) = \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \left(\beta_j^* + \frac{u_j}{\sqrt{n}} \right) \right\|^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j \left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right|.$$

Let $\hat{\mathbf{u}}^{(n)} = \arg\min \Psi_n(\mathbf{u})$; then $\hat{\boldsymbol{\beta}}^{*(n)} = \boldsymbol{\beta}^* + \frac{\hat{\mathbf{u}}^{(n)}}{\sqrt{n}}$ or $\hat{\mathbf{u}}^{(n)} = \sqrt{n} \times (\boldsymbol{\beta}^{*(n)} - \boldsymbol{\beta}^*)$. Note that $\Psi_n(\mathbf{u}) - \Psi_n(\mathbf{0}) = V_4^{(n)}(\mathbf{u})$, where

$$V_4^{(n)}(\mathbf{u}) \equiv \mathbf{u}^T \left(\frac{1}{n} \mathbf{X}^T \mathbf{X} \right) \mathbf{u} - 2 \frac{\boldsymbol{\epsilon}^T \mathbf{X}}{\sqrt{n}} \mathbf{u} + \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} \hat{w}_j \sqrt{n} \left(\left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| - |\beta_j^*| \right).$$

We know that $\frac{1}{n} \mathbf{X}^T \mathbf{X} \to \mathbf{C}$ and $\frac{\boldsymbol{\epsilon}^T \mathbf{X}}{\sqrt{n}} \to_d \mathbf{W} = \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{C})$. Now consider the limiting behavior of the third term. If $\beta_j^* \neq 0$, then $\hat{w}_j \to_p |\beta_j^*|^{-\gamma}$ and $\sqrt{n}(|\beta_j^* + \frac{u_j}{\sqrt{n}}| - |\beta_j^*|) \to u_j \operatorname{sgn}(\beta_j^*)$. By Slutsky's theorem, we have $\frac{\lambda_n}{\sqrt{n}} \hat{w}_j \sqrt{n}(|\beta_j^* + \frac{u_j}{\sqrt{n}}| - |\beta_j^*|) \to_p 0$. If $\beta_j^* = 0$, then $\sqrt{n}(|\beta_j^* + \frac{u_j}{\sqrt{n}}| - |\beta_j^*|) = |u_j|$ and $\frac{\lambda_n}{\sqrt{n}} \hat{w}_j = \frac{\lambda_n}{\sqrt{n}} n^{\gamma/2} (|\sqrt{n} \hat{\beta}_j|)^{-\gamma}$, where $\sqrt{n} \hat{\beta}_j = O_p(1)$. Thus, again by Slutsky's theorem, we see that $V_4^{(n)}(\mathbf{u}) \to_d V_4(\mathbf{u})$ for every $\mathbf{u}$, where

$$V_4(\mathbf{u}) = \begin{cases} \mathbf{u}_{\mathcal{A}}^T \mathbf{C}_{11} \mathbf{u}_{\mathcal{A}} - 2 \mathbf{u}_{\mathcal{A}}^T \mathbf{W}_{\mathcal{A}} & \text{if } u_j = 0 \ \forall j \notin \mathcal{A} \\ \infty & \text{otherwise.} \end{cases}$$

$V_4^{(n)}$ is convex, and the unique minimum of V_4 is $(\mathbf{C}_{11}^{-1} \mathbf{W}_{\mathcal{A}}, 0)^T$. Following the epi-convergence results of Geyer (1994) and Knight and Fu (2000), we have

$$\hat{\mathbf{u}}_{\mathcal{A}}^{(n)} \to_d \mathbf{C}_{11}^{-1} \mathbf{W}_{\mathcal{A}} \qquad \text{and} \qquad \hat{\mathbf{u}}_{\mathcal{A}^c}^{(n)} \to_d \mathbf{0}. \tag{A.10}$$

Finally, we observe that $\mathbf{W}_{\mathcal{A}} = \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{C}_{11})$; then we prove the asymptotic normality part.

Now we show the consistency part. $\forall j \in \mathcal{A}$, the asymptotic normality result indicates that $\hat{\beta}_j^{(n)} \to_p \beta_j^*$; thus $P(j \in \mathcal{A}_n^*) \to 1$. Then it suffices to show that $\forall j' \notin \mathcal{A}$, $P(j' \in \mathcal{A}_n^*) \to 0$. Consider the event $j' \in \mathcal{A}_n^*$. By the KKT optimality conditions, we know that $2 \mathbf{x}_{j'}^T (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}^{*(n)}) = \lambda_n \hat{w}_{j'}$. Note that $\lambda_n \hat{w}_{j'} / \sqrt{n} = \frac{\lambda_n}{\sqrt{n}} n^{\gamma/2} \frac{1}{|\sqrt{n} \hat{\beta}_{j'}|^{\gamma}} \to_p \infty$, whereas

$$2 \frac{\mathbf{x}_{j'}^T (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}^{*(n)})}{\sqrt{n}} = 2 \frac{\mathbf{x}_{j'}^T \mathbf{X} \sqrt{n} (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^{*(n)})}{n} + 2 \frac{\mathbf{x}_{j'}^T \boldsymbol{\epsilon}}{\sqrt{n}}.$$

By (A.10) and Slutsky's theorem, we know that $2 \mathbf{x}_{j'}^T \mathbf{X} \sqrt{n} (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^{*(n)}) / n \to_d$ some normal distribution and $2 \mathbf{x}_{j'}^T \boldsymbol{\epsilon} / \sqrt{n} \to_d \mathrm{N}(\mathbf{0}, 4 \|\mathbf{x}_{j'}\|^2 \sigma^2)$. Thus

$$P(j' \in \mathcal{A}_n^*) \le P\left(2 \mathbf{x}_{j'}^T \left(\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}^{*(n)} \right) = \lambda_n \hat{w}_{j'} \right) \to 0.$$

Proof of Theorem 3

We first show that for all i,

$$E\left[(\hat{\mu}_i^*(\lambda) - \mu_i)^2 \right] \le \left(\lambda^{2/(1+\gamma)} + 5 + \frac{4}{\gamma} \right) \left(\min(\mu_i^2, 1) + q\left(\lambda^{1/(1+\gamma)} \right) \right), \tag{A.11}$$

where $q(t) = \frac{1}{\sqrt{2\pi}\, t} e^{-t^2/2}$. Then, by (A.11), we have

$$R(\hat{\mu}_i^*(\lambda)) \le \left(\lambda^{2/(1+\gamma)} + 5 + \frac{4}{\gamma} \right) \left(R(\text{ideal}) + nq\left(\lambda^{1/(1+\gamma)} \right) \right). \tag{A.12}$$

Observe that when $\lambda = (2 \log n)^{(1+\gamma)/2}$, $\lambda^{2/(1+\gamma)} = 2 \log n$, and $nq(\lambda^{1/(1+\gamma)}) \le \frac{1}{2\sqrt{\pi}} (\log n)^{-1/2}$; thus Theorem 3 is proven.

To show (A.11), consider the decomposition

$$\begin{aligned} E\left[(\hat{\mu}_i^*(\lambda) - \mu_i)^2 \right] &= E\left[(\hat{\mu}_i^*(\lambda) - y_i)^2 \right] + E[(y_i - \mu_i)^2] \\ &\quad + 2E[\hat{\mu}_i^*(\lambda)(y_i - \mu_i)] - 2E[y_i(y_i - \mu_i)] \\ &= E\left[(\hat{\mu}_i^*(\lambda) - y_i)^2 \right] + 1 + E\left[\frac{d \hat{\mu}_i^*(\lambda)}{dy_i} \right] - 2, \end{aligned}$$

where we have applied Stein's lemma (Stein 1981) to $E[\hat{\mu}_i^*(\lambda)(y_i - \mu_i)]$. Note that

$$(\hat{\mu}_i^*(\lambda) - y_i)^2 = \begin{cases} y_i^2 & \text{if } |y_i| < \lambda^{1/(1+\gamma)} \\ \dfrac{\lambda^2}{|y_i|^{2\gamma}} & \text{if } |y_i| > \lambda^{1/(1+\gamma)} \end{cases}$$

and

$$\frac{d \hat{\mu}_i^*(\lambda)}{dy_i} = \begin{cases} 0 & \text{if } |y_i| < \lambda^{1/(1+\gamma)} \\ 1 + \dfrac{\lambda}{\gamma |y_i|^{1+\gamma}} & \text{if } |y_i| > \lambda^{1/(1+\gamma)}. \end{cases}$$

Thus we get

$$\begin{aligned} &E\left[(\hat{\mu}_i^*(\lambda) - \mu_i)^2 \right] \\ &\quad = E\left[y_i^2 I\left(|y_i| < \lambda^{1/(1+\gamma)} \right) \right] \\ &\qquad + E\left[\left(\frac{\lambda^2}{|y_i|^{2\gamma}} + \frac{2\lambda}{\gamma |y_i|^{1+\gamma}} + 2 \right) I\left(|y_i| > \lambda^{1/(1+\gamma)} \right) \right] - 1. \end{aligned} \tag{A.13}$$

So it follows that

$$\begin{aligned} E\left[(\hat{\mu}_i^*(\lambda) - \mu_i)^2 \right] &\le \lambda^{2/(1+\gamma)} P\left(|y_i| < \lambda^{1/(1+\gamma)} \right) \\ &\quad + \left(2 + \frac{2}{\gamma} + \lambda^{2/(1+\gamma)} \right) P\left(|y_i| > \lambda^{1/(1+\gamma)} \right) - 1 \\ &= \lambda^{2/(1+\gamma)} + \left(2 + \frac{2}{\gamma} \right) P\left(|y_i| > \lambda^{1/(1+\gamma)} \right) - 1 \\ &\le \lambda^{2/(1+\gamma)} + 5 + \frac{4}{\gamma}. \end{aligned} \tag{A.14}$$

By the identity (A.13), we also have that

$$\begin{aligned} &E\left[(\hat{\mu}_i^*(\lambda) - \mu_i)^2 \right] \\ &\quad = E[y_i^2] \\ &\qquad + E\left[\left(\frac{\lambda^2}{|y_i|^{2\gamma}} + \frac{2\lambda}{\gamma |y_i|^{1+\gamma}} + 2 - y_i^2 \right) I\left(|y_i| > \lambda^{1/(1+\gamma)} \right) \right] - 1 \\ &\quad = E\left[\left(\frac{\lambda^2}{|y_i|^{2\gamma}} + \frac{2\lambda}{\gamma |y_i|^{1+\gamma}} + 2 - y_i^2 \right) I\left(|y_i| > \lambda^{1/(1+\gamma)} \right) \right] + \mu_i^2 \\ &\quad \le \left(2 + \frac{2}{\gamma} \right) P\left(|y_i| > \lambda^{1/(1+\gamma)} \right) + \mu_i^2. \end{aligned}$$

Following Donoho and Johnstone (1994), we let $g(\mu_i) = P(|y_i| > t)$ and $g(\mu_i) \le g(0) + \frac{1}{2} |\sup g''| \mu_i^2$. Note that $g(0) = 2 \int_t^{\infty} e^{-z^2/2} / \sqrt{2\pi}\, dz \le 2 \int_t^{\infty} z e^{-z^2/2} / (t \sqrt{2\pi})\, dz = 2/(\sqrt{2\pi}\, t) e^{-t^2/2}$ and

$$\begin{aligned} |g''(\mu_i = a)| &= \left| 2 \int_t^{\infty} \frac{(z-a)^2}{\sqrt{2\pi}} e^{-(z-a)^2/2}\, dz - 2 \int_t^{\infty} \frac{e^{-(z-a)^2/2}}{\sqrt{2\pi}} \right| \\ &\le 2 \left| \int_t^{\infty} \frac{(z-a)^2}{\sqrt{2\pi}} e^{-(z-a)^2/2}\, dz \right| + 2 \left| \int_t^{\infty} \frac{e^{-(z-a)^2/2}}{\sqrt{2\pi}} \right| \\ &\le 4. \end{aligned}$$

Thus we have $P(|y_i| > t) \le \frac{2}{\sqrt{2\pi}t} e^{-t^2/2} + 2\mu_i^2$. Then it follows that

$$E\big[(\hat{\mu}_i^*(\lambda) - \mu_i)^2\big] \le \left(4 + \frac{4}{\gamma}\right) q\big(\lambda^{1/(1+\gamma)}\big) + \left(5 + \frac{4}{\gamma}\right)\mu_i^2 \le \left(\lambda^{2/(1+\gamma)} + 5 + \frac{4}{\gamma}\right)\big(\mu_i^2 + q\big(\lambda^{1/(1+\gamma)}\big)\big). \tag{A.15}$$

Combining (A.14) and (A.15), we prove (A.11).

Proof of Corollary 2

Let $\hat{\boldsymbol{\beta}}^{*(n)}(\gamma = 1)$ be the adaptive lasso estimates in (8). By Theorem 2, $\hat{\boldsymbol{\beta}}^{*(n)}(\gamma = 1)$ is an oracle estimator if $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n \to \infty$. To show the consistency of the garotte selection, it suffices to show that $\hat{\boldsymbol{\beta}}^{*(n)}(\gamma = 1)$ satisfies the sign constraint with probability tending to 1. Pick any j. If $j \in \mathcal{A}$, then $\hat{\beta}^{*(n)}(\gamma = 1)_j \hat{\beta}(\text{ols})_j \to_p (\beta_j^*)^2 > 0$; thus $P(\hat{\beta}^{*(n)}(\gamma = 1)_j \hat{\beta}(\text{ols})_j \ge 0) \to 1$. If $j \notin \mathcal{A}$, then $P(\hat{\beta}^{*(n)}(\gamma = 1)_j \hat{\beta}(\text{ols})_j \ge 0) \ge P(\hat{\beta}^{*(n)}(\gamma = 1)_j = 0) \to 1$. In either case, $P(\hat{\beta}^{*(n)}(\gamma = 1)_j \hat{\beta}(\text{ols})_j \ge 0) \to 1$ for any $j = 1, 2, \ldots, p$.

Proof of Theorem 4

For the proof, we assume the following regularity conditions:

1. The Fisher information matrix is finite and positive definite,
$$\mathbf{I}(\boldsymbol{\beta}^*) = E[\phi''(\mathbf{x}^T\boldsymbol{\beta}^*)\mathbf{x}\mathbf{x}^T].$$
2. There is a sufficiently large enough open set $\mathcal{O}$ that contains $\boldsymbol{\beta}^*$ such that $\forall\, \boldsymbol{\beta} \in \mathcal{O}$,
$$|\phi'''(\mathbf{x}^T\boldsymbol{\beta})| \le M(\mathbf{x}) < \infty$$
and
$$E\big[M(\mathbf{x})|x_j x_k x_\ell|\big] < \infty$$
for all $1 \le j, k, \ell \le p$.

We first prove the asymptotic normality part. Let $\boldsymbol{\beta} = \boldsymbol{\beta}^* + \frac{\mathbf{u}}{\sqrt{n}}$. Define

$$\Gamma_n(\mathbf{u}) = \sum_{i=1}^n \left(-y_i\left(\mathbf{x}_i^T\left(\boldsymbol{\beta}^* + \frac{\mathbf{u}}{\sqrt{n}}\right)\right) + \phi\left(\mathbf{x}_i^T\left(\boldsymbol{\beta}^* + \frac{\mathbf{u}}{\sqrt{n}}\right)\right)\right) + \lambda_n \sum_{j=1}^p \hat{w}_j\left|\beta_j^* + \frac{u_j}{\sqrt{n}}\right|.$$

Let $\hat{\mathbf{u}}^{(n)} = \arg\min_{\mathbf{u}} \Gamma_n(\mathbf{u})$; then $\hat{\mathbf{u}}^{(n)} = \sqrt{n}(\boldsymbol{\beta}^{*(n)}(\text{glm}) - \boldsymbol{\beta}^*)$. Using the Taylor expansion, we have $\Gamma_n(\mathbf{u}) - \Gamma_n(\mathbf{0}) = H^{(n)}(\mathbf{u})$, where

$$H^{(n)}(\mathbf{u}) \equiv A_1^{(n)} + A_2^{(n)} + A_3^{(n)} + A_4^{(n)},$$

with

$$A_1^{(n)} = -\sum_{i=1}^n [y_i - \phi'(\mathbf{x}_i^T\boldsymbol{\beta}^*)]\frac{\mathbf{x}_i^T\mathbf{u}}{\sqrt{n}},$$

$$A_2^{(n)} = \sum_{i=1}^n \frac{1}{2}\phi''(\mathbf{x}_i^T\boldsymbol{\beta}^*)\mathbf{u}^T\frac{(\mathbf{x}_i\mathbf{x}_i^T)}{n}\mathbf{u},$$

$$A_3^{(n)} = \frac{\lambda_n}{\sqrt{n}}\sum_{j=1}^p \hat{w}_j\sqrt{n}\left(\left|\beta_j^* + \frac{u_j}{\sqrt{n}}\right| - |\beta_j^*|\right),$$

and

$$A_4^{(n)} = n^{-3/2}\sum_{i=1}^n \frac{1}{6}\phi'''(\mathbf{x}_i^T\tilde{\boldsymbol{\beta}}_*)(\mathbf{x}_i^T\mathbf{u})^3,$$

where $\tilde{\boldsymbol{\beta}}_*$ is between $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^* + \frac{\mathbf{u}}{\sqrt{n}}$. We analyze the asymptotic limit of each term. By the familiar properties of the exponential family, we observe that

$$E_{y_i,\mathbf{x}_i}\big([y_i - \phi'(\mathbf{x}_i^T\boldsymbol{\beta}^*)](\mathbf{x}_i^T\mathbf{u})\big) = 0$$

and

$$\text{var}_{y_i,\mathbf{x}_i}\big([y_i - \phi'(\mathbf{x}_i^T\boldsymbol{\beta}^*)](\mathbf{x}_i^T\mathbf{u})\big) = E_{\mathbf{x}_i}[\phi''(\mathbf{x}_i^T\boldsymbol{\beta}^*)(\mathbf{x}_i^T\mathbf{u})^2] = \mathbf{u}^T\mathbf{I}(\boldsymbol{\beta}^*)\mathbf{u}.$$

Then the central limit theorem says that $A_1^{(n)} \to_d \mathbf{u}^T\text{N}(\mathbf{0}, \mathbf{I}(\boldsymbol{\beta}^*))$. For the second term $A_2^{(n)}$, we observe that

$$\sum_{i=1}^n \phi''(\mathbf{x}_i^T\boldsymbol{\beta}^*)\frac{(\mathbf{x}_i\mathbf{x}_i^T)}{n} \to_p \mathbf{I}(\boldsymbol{\beta}^*).$$

Thus $A_2^{(n)} \to_p \frac{1}{2}\mathbf{u}^T\mathbf{I}(\boldsymbol{\beta}^*)\mathbf{u}$. The limiting behavior of the third term is discussed in the proof of Theorem 2. We summarize the results as follows:

$$\frac{\lambda_n}{\sqrt{n}}\hat{w}_j\sqrt{n}\left(\left|\beta_j^* + \frac{u_j}{\sqrt{n}}\right| - |\beta_j^*|\right) \to_p \begin{cases} 0 & \text{if } \beta_j^* \ne 0 \\ 0 & \text{if } \beta_j^* = 0 \text{ and } u_j = 0 \\ \infty & \text{if } \beta_j^* = 0 \text{ and } u_j \ne 0. \end{cases}$$

By the regularity condition 2, the fourth term can be bounded as

$$6\sqrt{n}A_4^{(n)} \le \sum_{i=1}^n \frac{1}{n}M(\mathbf{x}_i^T)|\mathbf{x}_i^T\mathbf{u}|^3 \to_p E[M(\mathbf{x})|\mathbf{x}^T\mathbf{u}|^3] < \infty.$$

Thus, by Slutsky's theorem, we see that $H^{(n)}(\mathbf{u}) \to_d H(\mathbf{u})$ for every $\mathbf{u}$, where

$$H(\mathbf{u}) = \begin{cases} \mathbf{u}_{\mathcal{A}}^T\mathbf{I}_{11}\mathbf{u}_{\mathcal{A}} - 2\mathbf{u}_{\mathcal{A}}^T\mathbf{W}_{\mathcal{A}} & \text{if } u_j = 0\ \forall j \notin \mathcal{A} \\ \infty & \text{otherwise,} \end{cases}$$

where $\mathbf{W} = \text{N}(\mathbf{0}, \mathbf{I}(\boldsymbol{\beta}^*))$. $H^{(n)}$ is convex and the unique minimum of H is $(\mathbf{I}_{11}^{-1}\mathbf{W}_{\mathcal{A}}, 0)^T$. Then we have

$$\hat{\mathbf{u}}_{\mathcal{A}}^{(n)} \to_d \mathbf{I}_{11}^{-1}\mathbf{W}_{\mathcal{A}} \qquad \text{and} \qquad \hat{\mathbf{u}}_{\mathcal{A}^c}^{(n)} \to_d \mathbf{0}. \tag{A.16}$$

Because $\mathbf{W}_{\mathcal{A}} = \text{N}(\mathbf{0}, \mathbf{I}_{11})$, the asymptotic normality part is proven.

Now we show the consistency part. $\forall j \in \mathcal{A}$, the asymptotic normality indicates that $P(j \in \mathcal{A}_n^*) \to 1$. Then it suffices to show that $\forall j' \notin \mathcal{A}$, $P(j' \in \mathcal{A}_n^*) \to 0$. Consider the event $j' \in \mathcal{A}_n^*$. By the KKT optimality conditions, we must have

$$\sum_i^n \mathbf{x}_{ij'}\big(y_i - \phi'\big(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{*(n)}(\text{glm})\big)\big) = \lambda_n\hat{w}_{j'};$$

thus $P(j' \in \mathcal{A}_n^*) \le P(\sum_i^n \mathbf{x}_{ij'}(y_i - \phi'(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{*(n)}(\text{glm}))) = \lambda_n\hat{w}_{j'})$. Note that

$$\sum_i^n \mathbf{x}_{ij'}\big(y_i - \phi'\big(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{*(n)}(\text{glm})\big)\big)/\sqrt{n} = \mathbf{B}_1^{(n)} + \mathbf{B}_2^{(n)} + \mathbf{B}_3^{(n)},$$

with

$$\mathbf{B}_1^{(n)} = \sum_i^n \mathbf{x}_{ij'}(y_i - \phi'(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{*(n)}))/\sqrt{n},$$

$$\mathbf{B}_2^{(n)} = \left(\frac{1}{n}\sum_i^n \mathbf{x}_{ij'}\phi''(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{*(n)})\mathbf{x}_i^T\right)\sqrt{n}\big(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^{*(n)}(\text{glm})\big),$$

and

$$\mathbf{B}_3^{(n)} = \frac{1}{n}\sum_i^n \mathbf{x}_{ij'}\phi'''(\mathbf{x}_i^T\tilde{\boldsymbol{\beta}}_{**})\big(\mathbf{x}_i^T\sqrt{n}\big(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^{*(n)}(\text{glm})\big)\big)^2/\sqrt{n},$$

where $\boldsymbol{\beta}_{**}$ is between $\hat{\boldsymbol{\beta}}^{*(n)}(\text{glm})$ and $\boldsymbol{\beta}^*$. By the previous arguments, we know that $\mathbf{B}_1^{(n)} \to_d \text{N}(\mathbf{0}, \mathbf{I}(\boldsymbol{\beta}))$. Observe that $\frac{1}{n}\sum_i^n \mathbf{x}_{ij'}\phi''(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}^*) \times$

$\mathbf{x}_i^T \to_p \mathbf{I}_{j'}$, where $\mathbf{I}_{j'}$ is the j'th row of $\mathbf{I}$. Thus (A.16) implies that $\mathbf{B}_2^{(n)}$ converges to some normal random variable. It follows the regularity condition 2 and (A.16) that $\mathbf{B}_3^{(n)} = O_p(1/\sqrt{n})$. Meanwhile, we have

$$\frac{\lambda_n \hat{w}_{j'}}{\sqrt{n}} = \frac{\lambda_n}{\sqrt{n}} n^{\gamma/2} \frac{1}{|\sqrt{n}\hat{\beta}_{j'}(\text{glm})|^{\gamma}} \to_p \infty.$$

Thus $P(j' \in \mathcal{A}_n^*) \to 0$. This completes the proof.

[Received September 2005. Revised May 2006.]

REFERENCES

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

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

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

Chen, S., Donoho, D., and Saunders, M. (2001), "Atomic Decomposition by Basis Pursuit," *SIAM Review*, 43, 129–159.

Donoho, D. (2004), "For Most Large Underdetermined Systems of Equations, the Minimal ℓ^1-Norm Solution is the Sparsest Solution," technical report, Stanford University, Dept. of Statistics.

Donoho, D., and Elad, M. (2002), "Optimally Sparse Representation in General (Nonorthogonal) Dictionaries via ℓ^1-Norm Minimizations," *Proceedings of the National Academy of Science USA*, 1005, 2197–2202.

Donoho, D., and Huo, X. (2002), "Uncertainty Principles and Ideal Atomic Decompositions," *IEEE Transactions on Information Theory*, 47, 2845–2863.

Donoho, D., and Johnstone, I. (1994), "Ideal Spatial Adaptation via Wavelet Shrinkages," *Biometrika*, 81, 425–455.

Donoho, D., Johnstone, I., Kerkyacharian, G., and Picard, D. (1995), "Wavelet Shrinkage: Asymptopia?" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 57, 301–337.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499.

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.

——— (2006), "Statistical Challenges With High Dimensionality: Feature Selection in Knowledge Discovery," in *Proceedings of the Madrid International Congress of Mathematicians 2006*.

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

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

Geyer, C. (1994), "On the Asymptotics of Constrained M-Estimation," *The Annals of Statistics*, 22, 1993–2010.

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

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

Lehmann, E., and Casella, G. (1998), *Theory of Point Estimation* (2nd ed.), New York: Springer-Verlag.

Leng, C., Lin, Y., and Wahba, G. (2004), "A Note on the Lasso and Related Procedures in Model Selection," technical report, University of Wisconsin Madison, Dept. of Statistics.

McCullagh, P., and Nelder, J. (1989), *Generalized Linear Models* (2nd ed.), New York: Chapman & Hall.

Meinshausen, N., and Bühlmann, P. (2004), "Variable Selection and High-Dimensional Graphs With the Lasso," technical report, ETH Zürich.

Rosset, S., and Zhu, J. (2004), "Piecewise Linear Regularization Solution Paths," technical report, University of Michigan, Department of Statistics.

Shen, X., and Ye, J. (2002), "Adaptive Model Selection," *Journal of the American Statistical Association*, 97, 210–221.

Stein, C. (1981), "Estimation of the Mean of a Multivariate Normal Distribution," *The Annals of Statistics*, 9, 1135–1151.

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

Yuan, M., and Lin, Y. (2005), "On the Nonnegative Garotte Estimator," technical report, Georgia Institute of Technology, School of Industrial and Systems Engineering.

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," technical report, University of California Berkeley, Dept. of Statistics.