Shrinkage Tuning Parameter Selection with a Diverging Number of Parameters

Hansheng Wang, Bo Li, and Chenlei Leng

Peking University, Tsinghua University, and National University of Singapore

This Version: September 28, 2008

Abstract

Contemporary statistical research frequently deals with problems involving a diverging number of parameters. For those problems, various shrinkage methods (e.g., LASSO, SCAD, etc) are found particularly useful for the purpose of variable selection (Fan and Peng, 2004; Huang et al., 2007b). Nevertheless, the desirable performances of those shrinkage methods heavily hinge on an appropriate selection of the tuning parameters. With a fixed predictor dimension, Wang et al. (2007b) and Wang and Leng (2007) demonstrated that the tuning parameters selected by a BIC-type criterion can identify the true model consistently. In this work, similar results are further extended to the situation with a diverging number of parameters for both unpenalized and penalized estimators (Fan and Peng, 2004; Huang et al., 2007b). Consequently, our theoretical results further enlarge not only the applicable scope of the traditional BIC-type criteria but also that of those shrinkage estimation methods (Tibshirani, 1996; Huang et al., 2007b; Fan and Li, 2001; Fan and Peng, 2004).

KEY WORDS: BIC; Diverging Number of Parameters; LASSO; SCAD

[†]Hansheng Wang is from Guanghua School of Management, Peking University, Beijing, P. R. China (*hansheng@gsm.pku.edu.c*). Bo Li is from School of Economics and Management, Tsinghua University, Beijing, P. R. China (*libo@sem.tsinghua.edu.cn*). Chenlei Leng is from Department of Statistics and Applied Probability, National University of Singapore, Singapore (*stalc@nus.edu.sg*).

1. INTRODUCTION

Contemporary research frequently deals with problems involving a diverging number of parameters (Fan and Li, 2006). For the sake of variable selection, various shrinkage methods have been developed. Those methods include but are not limited to: *least absolute shrinkage and selection operator* (Tibshirani, 1996, LASSO) and *smoothly clipped absolute deviation* (Fan and Li, 2001, SCAD).

For a typical linear regression model, it has been well understood that the traditional best subset selection method with the BIC criterion (Schwarz, 1978) can identify the true model consistently (Shao, 1997; Shi and Tsai, 2002). Unfortunately, such a method is computationally expensive, particularly in high dimensional situations. Thus, various shrinkage methods (e.g., LASSO, SCAD) have been proposed, which are computationally much more efficient. For those shrinkage methods, it has been shown that, if the tuning parameters can be selected appropriately, the true model can be identified consistently (Fan and Li, 2001; Fan and Peng, 2004; Zou, 2006; Wang et al., 2007a; Huang et al., 2007b). Recently, similar results are also extended to the situation with a diverging number of parameters (Fan and Peng, 2004; Huang et al., 2007a,b). Such an effort substantially enlarges the applicable scope of those shrinkage methods, from a traditional fixed-dimensional setting to a more challenging high-dimensional one. For an excellent discussion about the challenging issues encountered in high dimensional settings, we refer to Fan and Peng (2004) and Fan and Li (2006).

Obviously, the selection consistency of those shrinkage methods relies on an appropriate choice of the tuning parameters, and the method of GCV has been widely used in the past literature. However, in the traditional model selection literature, it has been well understood that the asymptotic behavior of GCV is similar to that of AIC, which is a well known *loss efficient* but *selection inconsistent* variable selection criterion. For a formal definition of *loss efficiency* and *selection consistency*, we refer to Shao (1997) and Yang (2005). Thus, one can reasonably conjecture that the shrinkage parameter selected by GCV might not be able to identify the true model consistently (just like its performance with unpenalized estimators). Such a conjecture has been formally verified by Wang et al. (2007b) for the SCAD method. In addition to that, Wang et al. (2007b) also confirmed that the SCAD estimator, with the tuning parameter chosen by a BIC-type criterion, can identify the true model consistently. Similar work has been done for adaptive LASSO by Wang and Leng (2007). Unfortunately, their theoretical results are developed under the assumption of a fixed predictor dimension, thus are not directly applicable with a diverging number of parameters. This immediately raises one interesting question: how should one select the tuning parameters with a diverging number of parameters?

Note that the traditional BIC criterion can identify the true model consistently, as long as the predictor dimension is fixed. Thus, it is natural to conjecture that such a BIC criterion or its slightly modified version can still find the true model consistently with a diverging number of parameters. We may further conjecture that this conclusion is even correct for penalized estimators (e.g., LASSO, SCAD, etc). Nevertheless, how to prove this conclusion theoretically is rather challenging. In a traditional fixed dimension setting, the number of candidate models is fixed. Thus, as long as the corresponding BIC criterion can consistently differentiate the true model from an arbitrary candidate one, we know immediately that the true model can be identified with probability tending to one. Nevertheless, if the predictor dimension also goes to infinity, the number of candidate models increases at an extremely fast speed. Even if the predictor dimension is not too large, the number of candidate models can exceed the sample size drastically. Thus, the traditional theoretical arguments (e.g., Shao, 1997; Shi and Tsai, 2002; Wang et al., 2007b) are no longer applicable.

To overcome such a challenging difficulty, we propose here a slightly modified BIC criterion and then develop in this article a set of novel probabilistic inequalities (see for example (A.7) in Appendix B). Those inequalities can bound the overfitting effect elegantly, and thus enable us to study the asymptotic behavior of the modified BIC criterion rigorously. In particular, we show theoretically that the modified BIC criterion is consistent in model selection even with a diverging number of parameters for both unpenalized and penalized estimators. This conclusion is correct regardless of whether the dimension of the true model is finite or diverging. We remark that many attractive properties (e.g., selection consistency) about a shrinkage estimator (e.g., LASSO, SCAD) cannot be realized in real practice, if a consistent tuning parameter selector (e.g., a BIC-type criterion) does not exist (Wang et al., 2007b). Thus, our theoretical results further enlarges not only the applicable scope of the traditional BIC-type criteria but also that of those shrinkage estimation methods (Tibshirani, 1996; Huang et al., 2007b; Fan and Li, 2001; Fan and Peng, 2004).

The rest of the article is organized as follows. The main theoretical results are given in Section 2 and numerical studies are reported in Section 3. A short discussion is provided in Section 4. All technical details are deferred to the Appendix.

2. BIC with Unpenalized Estimators

2.1. The BIC Criterion

Let $(Y_i, \boldsymbol{X}_i)$, $i = 1, \cdots, n$, be n independent and identically distributed observations, where $Y_i \in \mathbb{R}^1$ is the response of interest, $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{id})^\top \in \mathbb{R}^d$ is the associated d-dimensional predictor. In this paper, d is allowed to diverge to ∞ as

$n \to \infty$. We assume that the data are generated according to the following linear regression model (Shi and Tsai, 2002; Fan and Peng, 2004)

$$Y_i = \boldsymbol{X}_i^\top \beta + \varepsilon_i, \tag{2.1}$$

where ε_i is some random error with mean 0 and variance σ^2, $\beta = (\beta_1, \cdots, \beta_d)^\top \in \mathbb{R}^d$ is the regression coefficient. The true regression coefficient is denoted as $\beta_0 = (\beta_{01}, \cdots, \beta_{0d})^\top$. Without loss of generality, we assume that $\beta_{0j} \neq 0$ for every $1 \leq j \leq d_0$ but $\beta_{0j} = 0$ for every $j > d_0$. Simply speaking, we assume that the true model contains only the first d_0 predictors as relevant ones. Here d_0 is allowed to be either fixed or diverging to ∞ as $n \to \infty$.

Let $\mathbb{Y} = (Y_1, \cdots, Y_n)^\top \in \mathbb{R}^n$ be the response vector, and $\mathbb{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^\top \in \mathbb{R}^{n \times d}$ be the design matrix. We assume that the data have been standardized so that $E(X_{ij}) = 0$ and $\text{var}(X_{ij}) = 1$. We use the generic notation $\mathcal{S} = \{j_1, \cdots, j_{d^*}\}$ to denote an arbitrary candidate model, which includes $X_{j_1}, \cdots, X_{j_{d^*}}$ as relevant predictors. We use $|\mathcal{S}|$ to denote the size of the model $\mathcal{S}$ (i.e., $|\mathcal{S}| = d^*$). Next, define $X_{\mathcal{S}} = (X_{j_1}, \cdots, X_{j_{d^*}})^\top$, $\beta_{\mathcal{S}} = (\beta_{j_1}, \cdots, \beta_{j_{d^*}})^\top$, and $\mathbb{X}_{\mathcal{S}} = (\mathbb{X}_{j_1}, \cdots, \mathbb{X}_{j_{d^*}}) \in \mathbb{R}^{n \times d^*}$, where $\mathbb{X}_j \in \mathbb{R}^n$ stands for the jth column of $\mathbb{X}$. Furthermore, we use $\mathcal{S}_F = \{1, \cdots, d\}$ to represent the full model and $\mathcal{S}_T = \{1, \cdots, d_0\}$ to represent the true model. Finally, let $\hat{\sigma}_{\mathcal{S}}^2 = \text{SSE}_{\mathcal{S}}/n = \inf_{\beta_{\mathcal{S}}} \|\mathbb{Y} - \mathbb{X}_{\mathcal{S}}\beta_{\mathcal{S}}\|^2/n$. Based on the above notations, we define a modified BIC criterion as

$$\text{BIC}_{\mathcal{S}} = \log(\hat{\sigma}_{\mathcal{S}}^2) + |\mathcal{S}| \times \frac{\log(n)}{n} \times C_n, \tag{2.2}$$

where $C_n > 0$ is some positive constant to be discussed more carefully; see Remark 1 in Section 2.4. As one can see, if $C_n = 1$, the modified BIC criterion (2.2) reduces to the traditional one. With $C_n = 1$, Shao (1997) and Shi and Tsai (2002) have demonstrated that the above BIC criterion is able to identify the true model consistently, if a finite dimension true model truly exists and the predictor dimension is fixed. Similar results have been extended to shrinkage methods by Wang et al. (2007b) and Wang and Leng (2007). Nevertheless, whether such a BIC-type criterion can still identify the true model consistently with a diverging number of parameters (i.e., $d \to \infty$) is largely unknown (to our best knowledge).

2.2. The Main Challenge

Since the BIC criterion (2.2) is a consistent model selection criterion with a fixed predictor dimension (Shao, 1997; Shi and Tsai, 2002; Zhao and Kulasekera, 2006), one might wonder whether we can apply similar proof techniques with a diverging number of parameters. In fact, proving BIC's consistency with a diverging number of parameters is much more difficult. To appreciate this fact, we need to know firstly why the BIC criterion (2.2) is consistent with a fixed number of parameters. An important step to prove this conclusion is to show that the BIC criterion (2.2) is able to differentiate the true model $\mathcal{S}_T$ from an arbitrary overfitted one (i.e., $\mathcal{S} \supset \mathcal{S}_T$ but $\mathcal{S} \neq \mathcal{S}_T$). For example, let $\mathcal{S}$ denote an arbitrary overfitted model (i.e., $\mathcal{S} \supset \mathcal{S}_T$ but $\mathcal{S} \neq \mathcal{S}_T$). We then must have $|\mathcal{S}| > |\mathcal{S}_T|$. By (2.2), we have

$$\text{BIC}_{\mathcal{S}} - \text{BIC}_{\mathcal{S}_T} = \log\left(\frac{\hat{\sigma}^2_{\mathcal{S}}}{\hat{\sigma}^2_{\mathcal{S}_T}}\right) + \Big(|\mathcal{S}| - |\mathcal{S}_T|\Big) \times \frac{\log n}{n} \times C_n. \qquad (2.3)$$

Under the assumption that d is fixed, one can easily show that $\log(\hat{\sigma}^2_{\mathcal{S}}/\hat{\sigma}^2_{\mathcal{S}_T}) = O_p(n^{-1})$. This quantity is asymptotically dominated by the second term $C_n(|\mathcal{S}| - |\mathcal{S}_T|)\log n/n > C_n \log n/n$ as long as $0 < C_n = O_p(1)$; see Shao (1997), Shi and Tsai (2002), Wang et al. (2007b), and Wang and Leng (2007). Consequently, one knows immediately that the right hand side of (2.3) is guaranteed to be positive as long as the sample size is sufficiently large. Thus, we have

$$P\Big(\mathrm{BIC}_{\mathcal{S}} > \mathrm{BIC}_{\mathcal{S}_T}\Big) \to 1 \tag{2.4}$$

for any overfitted candidate model $\mathcal{S}$. If the predictor dimension is fixed, one can have only a finite number of overfitted models. Consequently, (2.4) also implies that

$$P\bigg(\min_{\mathcal{S}\neq\mathcal{S}_T, \mathcal{S}\supset\mathcal{S}_T} \mathrm{BIC}_{\mathcal{S}} > \mathrm{BIC}_{\mathcal{S}_T}\bigg) \to 1. \tag{2.5}$$

As a result, we know that the BIC criterion (2.2) is able to differentiate the true model $\mathcal{S}_T$ from *every* overfitted model consistently. Nevertheless, establishing (2.5) with a diverging number of parameters is much more difficult. The reason is that, with a diverging number of parameters, the total number of all possible overfitted models is no longer a fixed number, and in fact it increases at an extremely fast speed as the sample size increases. Consequently, the inequality (2.4) no longer implies the desired conclusion (2.5). Thus, special techniques have to be developed to overcome this issue; see for details in the Appendix.

2.3. Technical Conditions

Let $\tau_{\min}(A)$ be the minimal eigenvalues of an arbitrary positive definite matrix A. Let Σ denote the covariance matrix of $\boldsymbol{X}_i$. To study the asymptotic behavior of the modified BIC criterion (2.2), the following technical conditions are needed.

(C1) $\boldsymbol{X}_i$ has component-wise finite fourth order moment, i.e., $\max_{1\leq j\leq d} EX_{ij}^4 < \infty$.

(C2) There exists a positive number κ such that $\tau_{\min}(\Sigma) \geq \kappa$ for every $d > 0$.

(C3) The predictor dimension satisfies that $\limsup d/n^{\kappa^*} < 1$ for some $\kappa^* < 1$.

(C4) $\sqrt{n/(C_n d \log n)} \times \liminf_{n\to\infty}\{\min_{j\in\mathcal{S}_T} |\beta_{0,j}|\} \to \infty$, and $C_n d \log n/n \to 0$.

Note that condition (C1) is a standard moment condition, which is routinely needed even in the fixed predictor dimension setting (Shi and Tsai, 2002; Wang et al., 2007b). Condition (C2) is also a reasonable condition widely assumed in the literature (Fan and Peng, 2004; Huang et al., 2007b). Otherwise, the predictors become linearly dependent with each other asymptotically. Condition (C3) characterizes the speed at which the predictor dimension is allowed to diverge to infinity. Condition (C4) puts a requirement on the size of the nonzero coefficients. Intuitively, if some nonzero coefficients converge to 0 too fast, those nonzero coefficients can hardly be estimated accurately; see Fan and Peng (2004) and Huang et al. (2007b). Lastly, (C4) also constraints that the value of the diverging constant C_n cannot be too large. Intuitively, a too large C_n value will lead to seriously underfitted models.

2.4. BIC with Unpenalized Estimators

For simplicity, we assume that ε is normally distributed. This assumption can be relaxed but at the cost of more complicated technical proofs and certain assumptions for ε's tail heaviness; see Huang et al. (2007b).

Theorem 1. *Assume technical conditions (C1)–(C4), $C_n \to \infty$, and ε is normally distributed, we then have*

$$P\left(\min_{\mathcal{S}\not\supset\mathcal{S}_T} BIC_{\mathcal{S}} > BIC_{\mathcal{S}_F}\right) \to 1.$$

By Theorem 1 we know that the minimal BIC value associated with underfitted models (i.e., $\mathcal{S} \not\supset \mathcal{S}_T$) is guaranteed to be larger than that of the full model as long as the sample size is sufficiently large. Thus, we know that, with probability tending to one, any underfitted model cannot be selected by the BIC criterion (2.2) because it is not even as favorable as that of the full model, i.e., $\mathrm{BIC}_{\mathcal{S}_F}$.

Remark 1. Although in theory we require $C_n \to \infty$, its divergence rate can be arbitrarily slow. For example, $C_n = \log\log d$ is used for all our numerical experiments and the simulation results are rather encouraging.

Theorem 2. *Assume technical conditions (C1)–(C4), $C_n \to \infty$, and ε is normally distributed, we then have*

$$P\left(\min_{\mathcal{S}\neq\mathcal{S}_T,\mathcal{S}\supset\mathcal{S}_T} BIC_{\mathcal{S}} > BIC_{\mathcal{S}_T}\right) \to 1.$$

By Theorem 2, we know that, with probability tending to one, any overfitted model cannot be selected by BIC either, because its BIC value is not as favorable as that of the true model (i.e., $\mathrm{BIC}_{\mathcal{S}_T}$). Combining Theorems 1 and 2 shows that the modified BIC criterion is able to identify the true model consistently.

2.5. BIC with Shrinkage Estimators

Because the traditional method of best subset selection is computationally too expensive in high dimensional situations (Fan and Peng, 2004), various shrinkage estimators have been proposed. Those estimators are obtained by optimizing the following penalized least squares objective function

$$Q_\lambda(\beta) = n^{-1}\|\mathbb{Y} - \mathbb{X}\beta\|^2 + \sum_{j=1}^{d} p_{\lambda,j}(|\beta_j|) \tag{2.6}$$

with various penalty function $p_{\lambda,j}(\cdot)$. We denote the resulting estimator by $\hat{\beta}_\lambda = (\hat{\beta}_{\lambda,1}, \cdots, \hat{\beta}_{\lambda,d})^\top$. For example, $\hat{\beta}_\lambda$ becomes the SCAD estimator, if $p_{\lambda,j}(\cdot)$ is a function with its first order derivative given by $\dot{p}_{\lambda,j}(t) = \lambda\{I(t \leq \lambda) + I(t > \lambda)(a\lambda - t)_+/\{(a -$

$1)\lambda\}$ with $a = 3.7$ and $(t)_+ = tI\{t > 0\}$; see Fan and Li (2001). In another situation, $\hat{\beta}_\lambda$ becomes the adaptive LASSO estimator, if $p_{\lambda,j}(t) = \lambda w_j t$ with some appropriately specified weights w_j (Zou, 2006; Zhang and Lu, 2007; Wang et al., 2007a). Furthermore, if we define $p_{\lambda,j}(t) = t^q$ with some $0 < q < 1$, then $\hat{\beta}_\lambda$ becomes the bridge estimator (Fu, 1998; Huang et al., 2007a).

Following Wang et al. (2007b) and Wang and Leng (2007), we define the modified BIC criterion for a shrinkage estimator as

$$\text{BIC}_\lambda = \log(\hat{\sigma}^2_\lambda) + |\mathcal{S}_\lambda| \times \frac{\log n}{n} \times C_n \tag{2.7}$$

with $\hat{\sigma}^2_\lambda = \text{SSE}_\lambda/n$ and $\text{SSE}_\lambda = \|\mathbb{Y} - \mathbb{X}\hat{\beta}_\lambda\|^2$. Let $\mathcal{S}_\lambda = \{j : \hat{\beta}_{\lambda,j} \neq 0\}$ be the model identified by $\hat{\beta}_\lambda$. Here we should carefully differentiate two notations, i.e., SSE_λ and $\text{SSE}_{\mathcal{S}_\lambda}$. Specifically, SSE_λ is the residual sum of squares associated with the shrinkage estimate $\hat{\beta}_\lambda$ and $\text{SSE}_{\mathcal{S}_\lambda}$ is the residual sum of squares associated with the unpenalized estimator based on $\mathcal{S}_\lambda$. By definition, we know immediately $\text{SSE}_\lambda \geq \text{SSE}_{\mathcal{S}_\lambda}$. Thus, we have $\text{BIC}_\lambda \geq \text{BIC}_{\mathcal{S}_\lambda}$. Then, the optimal tuning parameter is given by $\hat{\lambda} = \text{argmin}_\lambda \text{BIC}_\lambda$, which identifies the model $\mathcal{S}_{\hat{\lambda}}$.

For convenience purposes, we write $\hat{\beta}_\lambda = (\hat{\beta}^\top_{\lambda,a}, \hat{\beta}^\top_{\lambda,b})^\top$ with $\hat{\beta}_{\lambda,a} = (\hat{\beta}_{\lambda,1}, \cdots, \hat{\beta}_{\lambda,d_0})^\top$ and $\hat{\beta}_{\lambda,a} = (\hat{\beta}_{\lambda,d_0+1}, \cdots, \hat{\beta}_{\lambda,d})^\top$. Simply speaking, $\hat{\beta}_{\lambda,a}$ is the shrinkage estimator corresponding to the nonzero coefficients while $\hat{\beta}_{\lambda,b}$ is the one corresponding to zero coefficients. Many researchers have demonstrated that there exist a tuning parameter sequence $\lambda_n \to 0$, such that with probability tending to one $\hat{\beta}_{\lambda_n,b} = 0$ and $\hat{\beta}_{\lambda_n,a}$ can be as efficient as the oracle estimator, i.e., the unpenalized estimator obtained under the true model. Because, $\hat{\beta}_{\lambda_n,b} = 0$ with probability tending to one, asymptotically we must have $\hat{\beta}_{\lambda_n,a}$ being the minimizer of the following objective function

$$Q_\lambda^*(\beta_{\mathcal{S}_T}) = n^{-1}\|\mathbb{Y} - \mathbb{X}_{\mathcal{S}_T}\beta_{\mathcal{S}_T}\|^2 + \sum_{j=1}^{d_0} p_{\lambda_n,j}(|\beta_j|).$$

Simple algebra shows that, with probability tending to one, we must have

$$\begin{aligned}\hat{\beta}_{\lambda_n,a} &= \Big\{n^{-1}\mathbb{X}_{\mathcal{S}_T}^\top \mathbb{X}_{\mathcal{S}_T}\Big\}^{-1}\Big\{n^{-1}\mathbb{X}_{\mathcal{S}_T}^\top \mathbb{Y} + 2^{-1}\mathrm{sgn}\{\hat{\beta}_{\lambda_n,a}\}\dot{p}_\lambda\big(|\hat{\beta}_{\lambda_n,a}|\big)\Big\} \\ &= \hat{\beta}_{\mathcal{S}_T} + 2^{-1}\Big\{n^{-1}\mathbb{X}_{\mathcal{S}_T}^\top \mathbb{X}_{\mathcal{S}_T}\Big\}^{-1}\mathrm{sgn}\{\hat{\beta}_{\lambda_n,a}\}\dot{p}_\lambda\big(|\hat{\beta}_{\lambda_n,a}|\big), \qquad (2.8)\end{aligned}$$

where $\hat{\beta}_{\mathcal{S}_T} = \{n^{-1}\mathbb{X}_{\mathcal{S}_T}\mathbb{X}_{\mathcal{S}_T}^\top\}^{-1}\{n^{-1}\mathbb{X}_{\mathcal{S}_T}^\top\mathbb{Y}\}$, $\dot{p}_\lambda(|\hat{\beta}_{\lambda_n,a}|) = \{\dot{p}(|\hat{\beta}_{\lambda_n,j}|)\}_{j=1}^{d_0}$, and $\mathrm{sgn}(\hat{\beta}_{\lambda_n,a})$ is a diagonal matrix with the jth diagonal component given by $\mathrm{sgn}(\hat{\beta}_{\lambda_n,j})$.

To extend the results in least squares estimation to the shrinkage setting, we need to demonstrate that BIC_{λ_n} and $\mathrm{BIC}_{\mathcal{S}_{\lambda_n}}$ are sufficiently similar, or equivalently, SSE_{λ_n} and $\mathrm{SSE}_{\mathcal{S}_{\lambda_n}}$ are very close in some sense. Specifically, it suffices to show that

$$\mathrm{SSE}_{\lambda_n} = \mathrm{SSE}_{\mathcal{S}_{\lambda_n}} + o_p(\log n) \qquad (2.9)$$

In light of (2.8) and Bai and Silverstein (2006), we see that (2.9) boils down to

$$\|\dot{p}_\lambda(\hat{\beta}_{\lambda_n,a})\|^2 = o_p(\log n/n), \qquad (2.10)$$

which, as will be discussed below, is a very reasonable assumption.

Remark 2. For the SCAD estimator (Fan and Li, 2001; Fan and Peng, 2004), if the reference tuning parameter is set to be $\lambda_n = (\log n)^\gamma\sqrt{d/n}$ for some $\gamma > 0$. One can follow the similar techniques as in the Lemma 3 of Wang et al. (2007b) and verify that $\|\dot{p}_\lambda(\hat{\beta}_{\lambda,a})\| = 0$ with probability tending to 1. Thus, the assumption (2.10) is satisfied for the SCAD estimator.

Remark 3. For the adaptive lasso estimator, we can define (for example) $w_j = 1/|\tilde{\beta}_j|$, where $\tilde{\beta} = (\tilde{\beta}_1, \cdots, \tilde{\beta}_p)^\top$ is the unpenalized full model estimator. Then with a fixed d_0 (for example), a sufficient condition for $\hat{\beta}_{\lambda_n}$ to be both $\sqrt{n/d}$- and selection consistent is that $\sqrt{n}\lambda_n \to 0$ but $(n/d)\lambda_n \to \infty$. Under these constraints, we can set $\lambda_n = (d/n)^{1-\delta}$ for some $0 < \delta < 1/6$. We know immediately $(n/d)\lambda_n = (n/d)^\delta \to \infty$ under condition (C3). If we can further assume (for example) the κ^* in condition (C3) is no larger than 1/4; see for example Theorem 1 in Fan and Peng (2004), we then have $\sqrt{n}\lambda_n < n^{1/2}n^{3\delta/4-3/4} = n^{3\delta/4-1/4} \to 0$ asymptotically because $\delta < 1/6$. Furthermore, we can also verify that $\|\dot{p}_\lambda(\hat{\beta}_{\lambda_n,a})\|^2 = O_p(d\lambda_n^2) = O_p(n^{1/4+3\delta/2-3/2}) = O_p(n^{3\delta/2-5/4}) = o_p(\log n/n)$ asymptotically because $\delta < 1/6$. Thus, the assumption (2.10) is also reasonable for the adaptive LASSO estimator under appropriate conditions.

Remark 4. Requiring $\kappa^* \leq 1/4$ in the previous remark for the adaptive LASSO estimator is not necessary. If we define the adaptive weights as $w_j = 1/|\tilde{\beta}_j|^\omega$ with some sufficiently large $\omega > 1$, the value of κ^* can be further improved.

Theorem 3. *Assume technical conditions (C1)–(C4), $C_n \to \infty$, ε is normally distributed, and also (2.10), we then have $P(\mathcal{S}_{\hat{\lambda}} = \mathcal{S}_T) \to 1$ as $n \to \infty$.*

By Theorem 3, we know that the BIC criterion (2.2) is consistent in model selection. Thus, the results of Wang et al. (2007b) and Wang and Leng (2007) are still valid with the modified BIC criterion and a diverging number of parameters.

3. NUMERICAL STUDIES

Simulation experiments are conducted to confirm our theoretical findings. We only report here two representative cases with normally distributed random noise ε. To save computational time, the one-step sparse estimator of Zou and Li (2008) is used for SCAD. The covariate $\boldsymbol{X}_i$ is generated from a multivariate normal distribution with mean 0 and covariance matrix $\Sigma = (\sigma_{j_1 j_2})$ with $\sigma_{j_1 j_2} = 1$ if $j_1 = j_2$ and $\sigma_{j_1 j_2} = 0.5$ if $j_1 \neq j_2$. As we mentioned before, we fix $C_n = \log\log d$ for all our numerical experiments. Lastly, for each simulation setup, a total of 500 simulation replications are conducted.

Example 1. The first example is from Fan and Peng (2004). More specifically, we take $d = [4n^{1/4}] - 5$, $\beta = (11/4, -23/6, 37/12, -13/9, 1/3, 0, 0, \cdots, 0)^\top \in \mathbb{R}^d$, and $[t]$ stands for the largest integer no larger than t. For this example, the predictor dimension d is diverging but the dimension of the true model is fixed to be 5. To summarize the simulation results, we computed the median of the relative model error (Fan and Li, 2001, MRME), the average model size (i.e., the number of nonzero coefficients, MS), and also the percentage of the correctly identified true models (CM). Intuitively, a better model selection procedure should produce more accurate prediction results (i.e., smaller MRME value), more correct model sizes (i.e., CM$\approx$ d_0), and better model selection capability (i.e., larger CM value). For a more detailed explanation about MRME, MS, and CM, we refer to Fan and Li (2001) and Wang and Leng (2007). For the purpose of comparison, both the methods of the smoothly clipped absolute deviation (Fan and Li, 2001, SCAD) and the adaptive LASSO (Zou, 2006, ALASSO) are evaluated. Furthermore, the widely used GCV method (Fan and Li, 2001; Fan and Peng, 2004; Zou, 2006) is also considered. The detailed results are reported in the left panel of Figure 1. As one can see, the GCV method fails to identify the true model consistently because, regardless of the sample size, its CM value is far below 100%, which is mainly due to its overfitting effect. On the other hand, that of the BIC criterion approaches 100% quickly, which clearly confirms the consistency of the proposed BIC criterion. As a direct consequence, the MRME and MS values of BIC are consistently smaller than that of the GCV for both SCAD and ALASSO.

Example 2. In Example 1, although the dimension of the full model is diverging, the dimension of the true model is fixed. In this example, we consider the situation, where the dimension of the true model is also diverging. More specifically, we take $d = [7n^{1/4}]$ and the dimension of the true model to be $|\mathcal{S}_T| = d_0 = [d/3]$. Next, we generate β_{0j} for $1 \leq j \leq d_0$ from a uniform distribution on $[0.5, 1.5]$. The detailed results are summarized in the right panel of Figure 1. The findings are similar to those in Example 1. This example further confirms that the BIC criterion (2.2) is still consistent even if d_0 is diverging.

Table 1: Analysis Results of the Gender Discrimination Dataset

		ALASSO		SCAD	
Method	OLS	GCV	BIC	GCV	BIC
Female	-0.940	-0.808	0	0	0
PcJob	3.685	3.689	3.272	3.170	3.124
Ed1	-1.750	-1.331	0	0	0
Ed2	-3.134	-2.703	-1.272	-1.762	-1.696
Ed3	-2.277	-1.972	-0.756	-1.198	-1.134
Ed4	-2.112	-1.485	0	-0.026	0
Job1	-22.910	-23.101	-23.020	-25.118	-25.149
Job2	-21.084	-21.276	-20.947	-23.038	-23.053
Job3	-17.197	-17.389	-17.032	-19.075	-19.098
Job4	-12.837	-12.829	-11.927	-14.044	-14.057
Job5	-7.604	-7.399	-5.638	-8.215	-8.219

Example 3. As our concluding example, we re-analyze the gender discrimination data as studied by Fan and Peng (2004), where the detailed information about the dataset can be found. We focus on the 199 observations with 14 covariates as suggested by Fan and Peng (2004). Furthermore, following their semiparametric approach, we model the two continuous variables by splines and represent the categorical variables

by dummy variables. This produces a total of 26 predictors. The response of interest here is the annual salary in the year of 1995. We then apply both the method of SCAD and ALASSO to the dataset with both GCV and BIC as tuning parameter selectors. The detailed results are given in Table 3. As one can see, regardless of the estimation method (i.e., ALASSO or SCAD), the BIC method typically yields more sparse solutions than the GCV method does, which is a pattern consistent with our simulation experience. In addition to that, except for the method of ALASSO with GCV, all other methods identify gender (i.e., Female) as one irrelevant predictor, thus suggesting no gender discrimination. We remark that the same conclusion was also obtained by Fan and Peng (2004) but via a likelihood ratio test.

4. CONCLUDING REMARKS

Firstly, we would like to remark that the normality assumption used for ε is mainly for proof simplification. In fact, our numerical experience indicates that the theorem results are reasonably insensitive towards this assumption. For example, if we replace the normally distributed ε in the simulation study by a double exponentially distributed one, the final simulation results are almost identical. Secondly, the model setup considered in this work is a simple linear regression model. How to establish similar results for a semiparametric model (Wang et al., 2007b; Xie and Huang , 2007) and/or a generalized linear model (Wang and Leng, 2007) are both interesting topics for future study. Lastly, our current theoretical results cannot be directly extended to the situation with $p > n$. This is because with $p > n$ the value of $\hat{\sigma}^2_{\mathcal{S}_F}$ (for example) becomes 0. Under that situation, the BIC criterion (2.2) is no longer well defined (due to the $\log \hat{\sigma}^2_{\mathcal{S}_F}$ term). Thus, how to define a sensible BIC criterion with $p > n$ by itself is still an interesting question open for discussion.

APPENDIX

Appendix A. Proof of Theorem 1

Recall that $\tilde{\beta}$ is the unpenalized full model estimator. Under conditions (C1), (C2), and (C3), and by the results of Bai and Silverstein (2006), we know immediately that,

$$E\|\tilde{\beta} - \beta_0\|^2 = trace\big[\mathrm{cov}(\tilde{\beta})\big] = \sigma^2 trace\big[\{\mathbb{X}^\top \mathbb{X}\}^{-1}\big]$$

$$\leq dn^{-1}\sigma^2 \tau_{\min}^{-1}\{n^{-1}\mathbb{X}^\top \mathbb{X}\} = O_p(d/n).$$

This implies that $\|\tilde{\beta} - \beta_0\|^2 = O_p(d/n)$. Next, for an arbitrary model $\mathcal{S}$, define $\hat{\beta}^{(\mathcal{S})} = \mathrm{argmin}_{\{\beta \in \mathbb{R}^d : \beta_j = 0 \forall j \notin \mathcal{S}\}} \|\mathbb{Y} - \mathbb{X}\beta\|^2$. We then have

$$\min_{\mathcal{S} \not\supset \mathcal{S}_T} \|\hat{\beta}^{(\mathcal{S})} - \tilde{\beta}\|^2 \geq \min_{\mathcal{S} \not\supset \mathcal{S}_T} \|\hat{\beta}^{(\mathcal{S})} - \beta_0\|^2 - \|\tilde{\beta} - \beta_0\|^2 \geq \min_{j \in \mathcal{S}_T} \beta_{0,j}^2 - O_p(d/n). \qquad \text{(A.1)}$$

By technical condition (C4), we know that the right hand side of (A.1) is guaranteed to be positive with probability tending to one. Next, we follow the basic idea of Wang et al. (2007b) and consider

$$\min_{\mathcal{S} \not\supset \mathcal{S}_T} \Big\{ \mathrm{BIC}_{\mathcal{S}} - \mathrm{BIC}_{\mathcal{S}_F} \Big\} \geq \min_{\mathcal{S} \not\supset \mathcal{S}_T} \log\left(\frac{\hat{\sigma}_{\mathcal{S}}^2}{\hat{\sigma}_{\mathcal{S}_F}^2}\right) - \frac{d \log n}{n} \times C_n$$

Note that the right hand side of the above equation can be written as

$$= \min_{\mathcal{S} \not\supset \mathcal{S}_T} \log\left(1 + \frac{(\hat{\beta}^{(\mathcal{S})} - \tilde{\beta})^\top \Big\{n^{-1}\mathbb{X}^\top \mathbb{X}\Big\} (\hat{\beta}^{(\mathcal{S})} - \tilde{\beta})}{\hat{\sigma}_{\mathcal{S}_F}^2}\right) - \frac{d \log n}{n} \times C_n$$

$$\geq \min_{\mathcal{S} \not\supset \mathcal{S}_T} \log\left(1 + \frac{\hat{\tau}_{\min} \times \|\hat{\beta}^{(\mathcal{S})} - \tilde{\beta}\|^2}{\hat{\sigma}_{\mathcal{S}_F}^2}\right) - \frac{d \log n}{n} \times C_n, \qquad \text{(A.2)}$$

where $\hat{\tau}_{\min} \doteq \tau_{\min}\{n^{-1}\mathbb{X}^\top\mathbb{X}\}$. One can verify that $\log(1+x) \geq \min\{0.5x, \log 2\}$ for any $x > 0$. Consequently, the right hand side of (A.2) can be further bounded by

$$\geq \min_{\mathcal{S} \not\supset \mathcal{S}_T} \min\left(\log 2, \frac{\hat{\tau}_{\min} \times \|\hat{\beta}^{(\mathcal{S})} - \hat{\beta}_{\mathcal{S}_F}\|^2}{\hat{\sigma}^2_{\mathcal{S}_F}}\right) - \frac{d \log n}{n} \times C_n. \tag{A.3}$$

Because $d \log n/n \to 0$ under condition (C4), thus with probability tending to one we must have $\log 2 - C_n d \log n/n > 0$. Consequently, as long as we can show that, with probability tending to one,

$$\min_{\mathcal{S} \not\supset \mathcal{S}_T} \left(\frac{\hat{\tau}_{\min} \times \|\hat{\beta}^{(\mathcal{S})} - \hat{\beta}_{\mathcal{S}_F}\|^2}{\hat{\sigma}^2_{\mathcal{S}_F}}\right) - \frac{d \log n}{n} \times C_n \tag{A.4}$$

is positive, we know that the right hand side of (A.3) is guaranteed to be positive asymptotically. Under the normality assumption of ε, we can show that $\hat{\sigma}^2_{\mathcal{S}_F} \to_p \sigma^2$. Furthermore, by Bai and Silverstein (2006), we know that, with probability tending to one, $\hat{\tau}_{\min} \to \tau_{\min} = \tau_{\min}(\Sigma)$. Applying the inequality (A.1) to (A.4), we find that the quantity (A.4) can be further bounded by

$$\geq \frac{\tau_{\min}}{\sigma^2}\left\{\min_{j \in \mathcal{S}_T} \beta^2_{0,j} - O_p(d/n)\right\}\{1 + o_p(1)\} - \frac{d \log n}{n} \times C_n$$

$$= \frac{C_n d \log n}{n} \times \frac{\tau_{\min}}{\sigma^2}\left\{\frac{n}{C_n d \log n} \times \min_{j \in \mathcal{S}_T} \beta^2_{0j}\right\}\{1 + o_p(1)\} - \frac{d \log n}{n} \times C_n,$$

which is guaranteed to be positive asymptotically under condition (C4). This proves that, with probability tending to one, the right hand side of (A.2) must be positive. Such a fact further implies that $\min_{\mathcal{S} \not\supset \mathcal{S}_T}\{\mathrm{BIC}_{\mathcal{S}} - \mathrm{BIC}_{\mathcal{S}_F}\} > 0$ asymptotically. This completes the proof.

Appendix B. Proof of Theorem 2

Consider an arbitrary overfitted model $\mathcal{S}$ (i.e., $\mathcal{S} \supset \mathcal{S}_T$ but $\mathcal{S} \neq \mathcal{S}_T$), we must have

$\mathcal{S}^c = \mathcal{S}\backslash\mathcal{S}_T \neq \emptyset$ and $\mathcal{S} = \mathcal{S}_T \cup \mathcal{S}^c$. Note that the residual sum of squares corresponding to the model $\mathcal{S}$ can be written as

$$\mathrm{SSE}_{\mathcal{S}} = \inf_{\beta_{\mathcal{S}}} \Big\|\mathbb{Y} - \mathbb{X}_{\mathcal{S}}\beta_{\mathcal{S}}\Big\|^2 = \inf_{\beta_{\mathcal{S}_T},\beta_{\mathcal{S}^c}} \Big\|\mathbb{Y} - \mathbb{X}_{\mathcal{S}_T}\beta_{\mathcal{S}_T} - \mathbb{X}_{\mathcal{S}^c}\beta_{\mathcal{S}^c}\Big\|^2.$$

It can be easily verified that $\mathrm{SSE}_{\mathcal{S}_T} = \|Q_{\mathcal{S}_T}\mathbb{Y}\|^2$ with $Q_{\mathcal{S}_T} = I - \mathbb{X}_{\mathcal{S}_T}(\mathbb{X}_{\mathcal{S}_T}^\top\mathbb{X}_{\mathcal{S}_T})^{-1}\mathbb{X}_{\mathcal{S}_T}^\top$. For an arbitrary matrix A, we use span(A) to denote the linear subspace spanned by the column vectors of A. On can easily verify that span$(\mathbb{X}_{\mathcal{S}_T}, \mathbb{X}_{\mathcal{S}^c})$ = span$(\mathbb{X}_{\mathcal{S}_T}, \tilde{\mathbb{X}}_{\mathcal{S}^c})$, where $\tilde{\mathbb{X}}_{\mathcal{S}^c} = Q_{\mathcal{S}_T}\mathbb{X}_{\mathcal{S}^c}$, the orthogonal complement of $\mathbb{X}_{\mathcal{S}^c}$ with respect to span$(\mathbb{X}_{\mathcal{S}_T})$. This implies immediately that

$$\mathrm{SSE}_{\mathcal{S}} = \inf_{\beta_{\mathcal{S}_T},\beta_{\mathcal{S}^c}} \big\|\mathbb{Y} - \mathbb{X}_{\mathcal{S}_T}\beta_{\mathcal{S}_T} - \tilde{\mathbb{X}}_{\mathcal{S}^c}\beta_{\mathcal{S}^c}\big\|^2.$$

We can verify further that the minimizer of the above optimization problem is given by $\hat{\beta}_{\mathcal{S}_T} = (\mathbb{X}_{\mathcal{S}_T}^\top\mathbb{X}_{\mathcal{S}_T})^{-1}(\mathbb{X}_{\mathcal{S}_T}^\top\mathbb{Y})$ and $\tilde{\beta}_{\mathcal{S}^c} = (\tilde{\mathbb{X}}_{\mathcal{S}^c}^\top\tilde{\mathbb{X}}_{\mathcal{S}^c})^{-1}(\tilde{\mathbb{X}}_{\mathcal{S}^c}^\top\hat{\mathcal{E}})$, where $\hat{\mathcal{E}} = \mathbb{Y} - \mathbb{X}_{\mathcal{S}_T}\hat{\beta}_{\mathcal{S}_T}$ is an estimator of $\mathcal{E} = (\varepsilon_1, \cdots, \varepsilon_n)^\top \in \mathbb{R}^n$. We can the verify the following relationship

$$\mathrm{SSE}_{\mathcal{S}_T} - \mathrm{SSE}_{\mathcal{S}} = \big\{n^{-1/2}\hat{\mathcal{E}}^\top\tilde{\mathbb{X}}_{\mathcal{S}^c}\big\}\big\{n^{-1}\tilde{\mathbb{X}}_{\mathcal{S}^c}^\top\tilde{\mathbb{X}}_{\mathcal{S}^c}\big\}^{-1}\big\{n^{-1/2}\tilde{\mathbb{X}}_{\mathcal{S}^c}\hat{\mathcal{E}}\big\}$$

$$\leq \hbar_{\max}^{\mathcal{S}^c}\big\|n^{-1/2}\hat{\mathcal{E}}^\top\tilde{\mathbb{X}}_{\mathcal{S}^c}\big\|^2 = \hbar_{\max}^{\mathcal{S}^c}\sum_{j\in\mathcal{S}^c}\big(n^{-1/2}\hat{\mathcal{E}}^\top\tilde{\mathbb{X}}_j\big)^2$$

$$\leq \hbar_{\max}^{\mathcal{S}^c}\cdot\max_{j\in\mathcal{S}^c}\big(n^{-1/2}\hat{\mathcal{E}}^\top\tilde{\mathbb{X}}_j\big)^2 \times |\mathcal{S}^c| \leq \hbar_{\max}^{\mathcal{S}^c} \times \max_{j\in\mathcal{S}_F\backslash\mathcal{S}_T}\big(n^{-1/2}\hat{\mathcal{E}}^\top\tilde{\mathbb{X}}_j\big)^2 \times |\mathcal{S}^c|,$$

where $\hbar_{\max}^{\mathcal{S}^c} = \tau_{\min}^{-1}(n^{-1}\tilde{\mathbb{X}}_{\mathcal{S}^c}^\top\tilde{\mathbb{X}}_{\mathcal{S}^c})$, recall $\mathbb{X}_j$ is the jth column of $\mathbb{X}$, and $\tilde{\mathbb{X}}_j = Q_{\mathcal{S}_T}\mathbb{X}_j$. Note that $\mathcal{S}^c \subset \mathcal{S}_T^c = \mathcal{S}_F\backslash\mathcal{S}_T$. Therefore, we have $\tau_{\min}(n^{-1}\tilde{\mathbb{X}}_{\mathcal{S}^c}^\top\tilde{\mathbb{X}}_{\mathcal{S}^c}) \geq \tau_{\min}(n^{-1}\tilde{\mathbb{X}}_{\mathcal{S}_T^c}^\top\tilde{\mathbb{X}}_{\mathcal{S}_T^c}) \doteq \{\hbar_{\max}^{\mathcal{S}_T^c}\}^{-1}$. We then must have

$$\max_{\mathcal{S}^c\subset\mathcal{S}_F\backslash\mathcal{S}_T}\left(\frac{\mathrm{SSE}_{\mathcal{S}_T} - \mathrm{SSE}_{\mathcal{S}}}{|\mathcal{S}^c|}\right) \leq \hbar_{\max}^{\mathcal{S}_T^c} \times \max_{j\in\mathcal{S}_F\backslash\mathcal{S}_T}\big(n^{-1}\hat{\mathcal{E}}^\top\tilde{\mathbb{X}}_j\big)^2. \tag{A.5}$$

We next examine the two terms of the right hand side of (A.5) respectively. Firstly, let γ be the eigenvector associated with $\tau_{\min}(n^{-1}\tilde{\mathbb{X}}_{\mathcal{S}_T^c}^\top \tilde{\mathbb{X}}_{\mathcal{S}_T^c})$, i.e., $\|\gamma\| = 1$ and

$$\tau_{\min}(n^{-1}\tilde{\mathbb{X}}_{\mathcal{S}_T^c}^\top \tilde{\mathbb{X}}_{\mathcal{S}_T^c}) = \gamma^\top(n^{-1}\tilde{\mathbb{X}}_{\mathcal{S}_T^c}^\top \tilde{\mathbb{X}}_{\mathcal{S}_T^c})\gamma = n^{-1}\|\tilde{\mathbb{X}}_{\mathcal{S}_T^c}\gamma\|^2.$$

By definition, we know that $\tilde{\mathbb{X}}_{\mathcal{S}_T^c}\gamma = \mathbb{X}_{\mathcal{S}_T^c}\gamma + \mathbb{X}_{\mathcal{S}_T}\gamma^*$ with $\gamma^* = -(\mathbb{X}_{\mathcal{S}_T}^\top \mathbb{X}_{\mathcal{S}_T})^{-1}\mathbb{X}_{\mathcal{S}_T}^\top \mathbb{X}_{\mathcal{S}_T^c}\gamma$. Thus, we know that

$$\tau_{\min}(n^{-1}\tilde{\mathbb{X}}_{\mathcal{S}_T^c}^\top \tilde{\mathbb{X}}_{\mathcal{S}_T^c}) = n^{-1}\|\mathbb{X}_{\mathcal{S}_T^c}\gamma + \mathbb{X}_{\mathcal{S}_T}\gamma^*\|^2 = n^{-1}\|\mathbb{X}\alpha\|^2 = \alpha^\top(n^{-1}\mathbb{X}^\top\mathbb{X})\alpha$$

$$\geq \tau_{\min}(n^{-1}\mathbb{X}^\top\mathbb{X})\|\alpha\|^2 \geq \tau_{\min}(n^{-1}\mathbb{X}^\top\mathbb{X}) \geq \kappa \qquad \text{(A.6)}$$

with probability tending to one. Here $\alpha = (\gamma^{*\top}, \gamma^\top)^\top$ satisfies that $\|\alpha\| > 1$. This implies that $\hbar_{\max}^{\mathcal{S}^c} \leq \kappa^{-1}$. Secondly, under the normality assumption, one can verify that $n^{-1/2}\hat{\mathcal{E}}\tilde{\mathbb{X}}_j$ follows a normal distribution with mean 0 and variance given by

$$\sigma_j^2 = n^{-1}trace\big[E(\mathcal{E}^T Q_{\mathcal{S}_T}\tilde{\mathbb{X}}_j\tilde{\mathbb{X}}_j^T Q_{\mathcal{S}_T}\mathcal{E})\big] = n^{-1}trace\big[E(\mathcal{E}^T\tilde{\mathbb{X}}_j\tilde{\mathbb{X}}_j^T\mathcal{E})\big]$$

$$= n^{-1}trace\big[E(\mathcal{E}\mathcal{E}^T)\tilde{\mathbb{X}}_j\tilde{\mathbb{X}}_j^T\big] = n^{-1}\sigma^2\|\tilde{\mathbb{X}}_j\|^2 \leq n^{-1}\sigma^2\|\mathbb{X}_j\|^2 < (1+\varphi)\sigma^2$$

with probability tending to one for an arbitrary but fixed constant $\varphi > 0$. Here we use the fact that $Q_{\mathcal{S}_T}\tilde{\mathbb{X}}_j = \tilde{\mathbb{X}}_j$ and also $\text{var}(X_j) = 1$. Then, the right hand side of (A.5) can be further bounded by

$$\max_{\mathcal{S}^c \subset \mathcal{S}_F \backslash \mathcal{S}_T}\left(\frac{\text{SSE}_{\mathcal{S}_T} - \text{SSE}_{\mathcal{S}}}{|\mathcal{S}^c|}\right) \leq (1+\varphi)\sigma^2\kappa^{-1} \times \max_{j \in \mathcal{S}_F \backslash \mathcal{S}_T} \chi_j^2(1), \qquad \text{(A.7)}$$

where $\chi_j^2(1) = \sigma_j^{-2}(n^{-1}\hat{\mathcal{E}}^\top\tilde{\mathbb{X}}_j)^2$ follows a chi-square distribution with one degree of freedom. We should note that these chi-square variables $\chi_j^2(1), j \in \mathcal{S}_F\backslash\mathcal{S}_T$ may well be dependent. Nevertheless, we can proceed by using Bonferroni's inequality to obtain

$$P\left(\max_{j\in\mathcal{S}_F\setminus\mathcal{S}_T}\chi_j^2(1)>2\log d\right)\le dP\Big(\chi_1^2(1)>2\log d\Big)$$

$$=(2\pi)^{-1/2}d\int_{2\log d}^{\infty}x^{-1/2}\exp(-x/2)dx\le\frac{(2\pi)^{-1/2}d}{\sqrt{2\log d}}\int_{2\log d}^{\infty}\exp(-x/2)dx=\frac{(2\pi)^{-1/2}}{\sqrt{2\log d}},$$

which implies that $\max_{j\in\mathcal{S}_F\setminus\mathcal{S}_T}\chi_j^2(1)\le 2\log d$ with probability tending to one as d tends to infinity. In conjunction with (A.7), we see that

$$\max_{\mathcal{S}^c\subset\mathcal{S}_F\setminus\mathcal{S}_T}\left(\frac{\mathrm{SSE}_{\mathcal{S}_T}-\mathrm{SSE}_{\mathcal{S}}}{|\mathcal{S}^c|}\right)\le 2(1+\varphi)\sigma^2\kappa^{-1}\log d \tag{A.8}$$

with probability tending to one. Consequently, we know that $\max_{\mathcal{S}^c\subset\mathcal{S}_F\setminus\mathcal{S}_T}(\mathrm{SSE}_{\mathcal{S}_T}-\mathrm{SSE}_{\mathcal{S}})=O_p(d\log d)=o(n)$ by (C3). Thus, the following Taylor expansion holds

$$n\Big(\mathrm{BIC}_{\mathcal{S}}-\mathrm{BIC}_{\mathcal{S}_T}\Big)=n\log\left[1+\frac{n^{-1}(\mathrm{SSE}_{\mathcal{S}}-\mathrm{SSE}_{\mathcal{S}_T})}{\hat\sigma^2_{\mathcal{S}_T}}\right]+|\mathcal{S}^c|\times\log n\times C_n$$

$$=\frac{1}{\hat\sigma^2_{\mathcal{S}_T}}(\mathrm{SSE}_{\mathcal{S}}-\mathrm{SSE}_{\mathcal{S}_T})+|\mathcal{S}^c|\times\log n\times C_n+o_p(1)$$

$$=\frac{1}{\sigma^2_{\mathcal{S}_T}}(\mathrm{SSE}_{\mathcal{S}}-\mathrm{SSE}_{\mathcal{S}_T})\{1+o_p(1)\}+|\mathcal{S}^c|\times\log n\times C_n+o_p(1).$$

Then, by the inequality (A.8), we know that the right hand side of the above equation can be uniformly bounded by $\ge|\mathcal{S}^c|(C_n\log n-2(1+\varphi)\kappa^{-1}\log d\{1+o_p(1)\})$. Consequently, we know that

$$\max_{\mathcal{S}\supset\mathcal{S}_T,\mathcal{S}\neq\mathcal{S}_T}\left\{\frac{n}{|\mathcal{S}^c|}\Big(\mathrm{BIC}_{\mathcal{S}}-\mathrm{BIC}_{\mathcal{S}_T}\Big)\right\}\ge C_n\log n-2(1+\varphi)\kappa^{-1}\log d\{1+o_p(1)\}$$

$$\ge\log n\Big\{C_n-2(1+\varphi)\kappa^{-1}\kappa^*\Big\}\{1+o_p(1)\} \tag{A.9}$$

with probability tending to one, where the last inequality is due to condition (C3). By theorem assumption, we know that $C_n \to \infty$. This implies that, with probability tending to one, $\max_{\mathcal{S}\supset\mathcal{S}_T,\mathcal{S}\neq\mathcal{S}_T}(\mathrm{BIC}_{\mathcal{S}} - \mathrm{BIC}_{\mathcal{S}_T})$ must be positive. This proves the theorem conclusion and completes the proof.

Appendix C. Proof of Theorem 3

Define $\Omega_- = \{\lambda > 0 : \mathcal{S}_\lambda \not\supset \mathcal{S}_T\}$, $\Omega_0 = \{\lambda > 0 : \mathcal{S}_\lambda = \mathcal{S}_T\}$, and $\Omega_+ = \{\lambda > 0 : \mathcal{S}_\lambda \supset \mathcal{S}_T, \mathcal{S}_\lambda \neq \mathcal{S}_T\}$. In other words, Ω_0 (Ω_-, Ω_+) collects all λ values which produces correctly (under, over) fitted models. We follow Wang et al. (2007b) and Wang and Leng (2007), and establish the theorem statement via the following two steps.

Case 1. (Underfitted model, i.e., $\lambda \in \Omega_-$). Firstly, under the assumption (2.10), we have $\mathrm{BIC}_{\lambda_n} = \mathrm{BIC}_{\mathcal{S}_{\lambda_n}} + o_p(\log n/n)$. Then, with probability tending to 1, we have

$$\begin{aligned}\inf_{\lambda\in\Omega_-} \mathrm{BIC}_\lambda - \mathrm{BIC}_{\lambda_n} &\geq \inf_{\lambda\in\Omega_-} \mathrm{BIC}_{\mathcal{S}_\lambda} - \mathrm{BIC}_{\mathcal{S}_T} + o_p(\log n/n)\\ &\geq \min_{\mathcal{S}\not\supset\mathcal{S}_T} \mathrm{BIC}_{\mathcal{S}} - \mathrm{BIC}_{\mathcal{S}_T} + o_p(\log n/n) \qquad \text{(A.10)}\end{aligned}$$

By Theorem 1's proof we know that $P(\min_{\mathcal{S}\not\supset\mathcal{S}_T} \mathrm{BIC}_{\mathcal{S}} - \mathrm{BIC}_{\mathcal{S}_F} + o_p(\log n/n) > 0) \to 1$. By Theorem 2, we know that $P(\mathrm{BIC}_{\mathcal{S}_F} - \mathrm{BIC}_{\mathcal{S}_T} \geq 0) \to 1$. Consequently, we know that $P(\inf_{\lambda\in\Omega_-} \mathrm{BIC}_\lambda - \mathrm{BIC}_{\lambda_n} > 0) \to 1$.

Case 2. (Overfitted model, i.e., $\lambda \in \Omega_+$). We can argue similarly as in the overfitting case to obtain the following inequality

$$\inf_{\lambda\in\Omega_+} \mathrm{BIC}_\lambda - \mathrm{BIC}_{\lambda_n} \geq \min_{\mathcal{S}\supset\mathcal{S}_T,\mathcal{S}\neq\mathcal{S}_T} \mathrm{BIC}_{\mathcal{S}_\lambda} - \mathrm{BIC}_{\mathcal{S}_T} + o_p(\log n/n).$$

By (A.9), we know that we can find a positive number η such that $\min_{\mathcal{S}\supset\mathcal{S}_T,\mathcal{S}\neq\mathcal{S}_T} \mathrm{BIC}_{\mathcal{S}_\lambda} - \mathrm{BIC}_{\mathcal{S}_T} > \eta \log n/n$ with probability tending to 1. Thus we see that the right hand side of the above equation is guaranteed be positive asymptotically. As a consequence,

$P(\inf_{\lambda\in\Omega_+}\mathrm{BIC}_\lambda - \mathrm{BIC}_{\lambda_n} > 0) \rightarrow 1$. This completes the proof.

REFERENCES

Bai, Z. D. and Silverstein, J. W. (2006), *spectral analysis of large dimensional random matrices*, Science Press, Beijing, P. R. China.

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,” *Proceedings of the International Congress of Mathematicians (M. Sanz-Sole, J. Soria, J.L. Varona, J. Verdera, eds.), Vol. III, European Mathematical Society, Zurich*, 595–622.

Fan, J. and Peng, H. (2004), “On non-concave penalized likelihood with diverging number of parameters,” *The Annals of Statistics*, 32, 928–961.

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

Huang, J., Horowitz, J. and Ma, S. (2007a), “Asymptotic properties of bridge estimators in sparse high-dimensional regression models,” *Annals of Statistics*, To appear.

Huang, J., Ma, S., and Zhang, C. H. (2007b), “Adaptive LASSO for sparse high dimensional regression,” *Technical Report No.374, Department of Statistics and Actuarial Science, University of Iowa.*

Schwarz, G. (1978), “Estimating the dimension of a model,” *The Annals of Statistics*, 6, 461–464.

Shi, P. and Tsai, C. L. (2002), “ Regression model selection – a residual likelihood approach,” *Journal of the Royal Statistical Society, Series B*, 64, 237–252.

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

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

Wang, H. and Leng, C. (2007), “Unified LASSO estimation via least squares approximation,” *Journal of the American Statistical Association*, 101, 1418–1429.

Wang, H., Li, G., and Tsai, C. L. (2007a), “Regression coefficient and autoregressive order shrinkage and selection via LASSO,” *Journal of Royal Statistical Society, Series B*, 69, 63–78.

Wang, H., Li, R., and Tsai, C. L. (2007b), “On the Consistency of SCAD Tuning Parameter Selector,” *Biometrika*, 94, 553–558.

Xie, H. and Huang, J. (2007), “SCAD-penalized regression in high-dimensional partially linear models,” *Annals of Statistics*, To appear.

Yang, Y. (2005), “Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation,” *Biometrika*, 92, 937–950.

Zhang, H. H. and Lu, W. (2007), “Adaptive LASSO for Cox’s proportional hazard model,” *Biometrika*, 94, 691-703.

Zhao, M. and Kulasekera, K. B. (2006), "Consistent linear model selection," *Statistics & Probability Letters*, 76, 520–530.

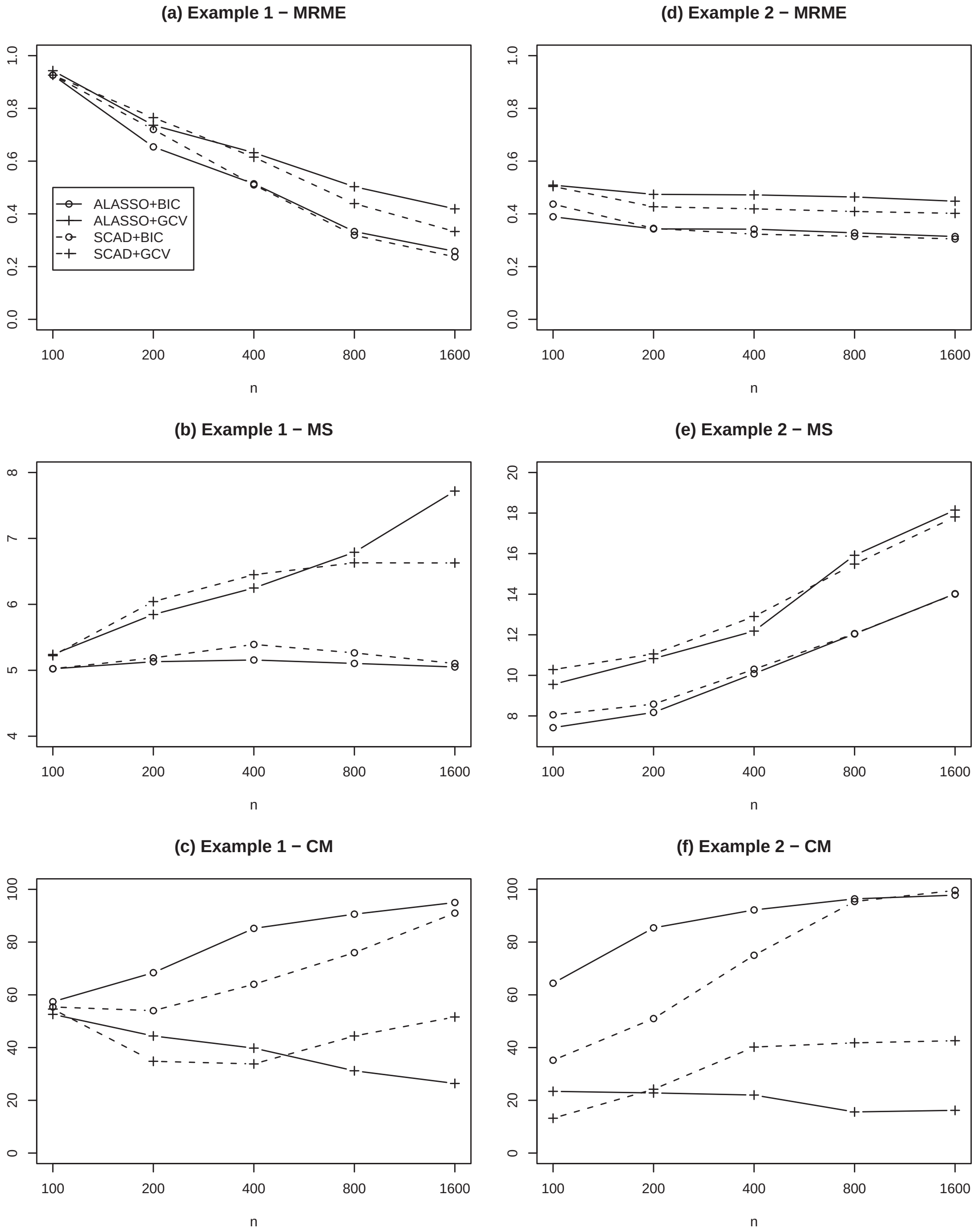
Zou, H. (2006), "The adaptive LASSO and its oracle properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H. and Li, R. (2008), "One-step sparse estimates in nonconcave penalized likelihood models," *The Annals of Statistics*, To appear.

ACKNOWLEDGEMENT

The authors are very grateful to the Editor, the AE, and two referees for their careful reading and insightful comments, which lead to a substantially improved manuscript. Hansheng Wang is supported in part by a NSFC grant (No. 10771006). Bo Li is supported in part by NSFC grants (No. 10801086, No.70621061, and No.70831003). Chenlei Leng is supported by NUS academic research grants and a grant from NUS Risk Management Institute.

Figure 1: The Detailed Simulation Results with Normal ε


(a) Example 1 − MRME
(d) Example 2 − MRME
(b) Example 1 − MS
(e) Example 2 − MS
(c) Example 1 − CM
(f) Example 2 − CM
ALASSO+BIC
ALASSO+GCV
SCAD+BIC
SCAD+GCV
n