Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2009)
71, *Part* 3, *pp.* 671–683

Shrinkage tuning parameter selection with a diverging number of parameters

Hansheng Wang,

Peking University, Beijing, People's Republic of China

Bo Li

Tsinghua University, Beijing, People's Republic of China

and Chenlei Leng

National University of Singapore, Singapore

[Received March 2008. Revised September 2008]

Summary. Contemporary statistical research frequently deals with problems involving a diverging number of parameters. For those problems, various shrinkage methods (e.g. the lasso and smoothly clipped absolute deviation) are found to be particularly useful for variable selection. Nevertheless, the desirable performances of those shrinkage methods heavily hinge on an appropriate selection of the tuning parameters. With a fixed predictor dimension, Wang and co-worker have demonstrated that the tuning parameters selected by a Bayesian information criterion 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. Consequently, our theoretical results further enlarge not only the scope of applicability of the traditional Bayesian information criterion type criteria but also that of those shrinkage estimation methods.

Keywords: Bayesian information criterion; Diverging number of parameters; Lasso; Smoothly clipped absolute deviation

1. Introduction

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

For a typical linear regression model, it is well understood that the traditional best subset selection method with the Bayesion information criterion (BIC) (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. the lasso and 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.*, 2007). Recently, similar results

Address for correspondence: Hansheng Wang, Guanghua School of Management, Peking University, Beijing 100871, People's Republic of China.
E-mail: hansheng@gsm.pku.edu.cn

 1369–7412/09/71671

have also been extended to the situation with a diverging number of parameters (Fan and Peng, 2004; Huang *et al.*, 2007, 2008). Such an effort substantially enlarges the scope of applicability of those shrinkage methods, from a traditional fixed dimensional to a more challenging high dimensional setting. For an excellent discussion of the challenging issues that are encountered in high dimensional settings, we refer to Fan and Peng (2004) and Fan and Li (2006).

Obviously, the consistency of selection of those shrinkage methods relies on an appropriate choice of the tuning parameters, and the method of generalized cross-validation (GCV) has been widely used in the past literature. However, in the traditional model selection literature, it has been well understood that the asymptotic behaviour of GCV is similar to that of Akaike's information criterion, 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, we can reasonably conjecture that the shrinkage parameter that is 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 the adaptive lasso by Wang and Leng (2007). Unfortunately, their theoretical results were developed under the assumption of a fixed predictor dimension and thus are not directly applicable with a diverging number of parameters. This immediately raises one interesting question: how should we select the tuning parameters with a diverging number of parameters?

Note that the traditional BIC can identify the true model consistently, as long as the predictor dimension is fixed. Thus, it is natural to conjecture that such a BIC 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. the lasso and SCAD). Nevertheless, how to prove this conclusion theoretically is quite challenging. In a traditional fixed dimension setting, the number of candidate models is fixed. Thus, as long as the corresponding BIC can consistently differentiate the true model from an arbitrary candidate model, we know immediately that the true model can be identified with probability tending to 1. Nevertheless, if the predictor dimension also goes to ∞, 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) and Wang *et al.* (2007b)) are no longer applicable.

To overcome such a challenging difficulty, we propose here a slightly modified BIC and then develop a set of novel probabilistic inequalities (see for example expression (B.3) in Appendix B). Those inequalities can bound the overfitting effect elegantly, and thus enable us to study the asymptotic behaviour of the modified BIC rigorously. In particular, we show theoretically that the modified BIC 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. consistency of selection) about a shrinkage estimator (e.g. the lasso and 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 enlarge not only the scope of applicability of the traditional BIC-type criteria but also that of those shrinkage estimation methods (Tibshirani, 1996; Huang *et al.*, 2007; Fan and Li, 2001; Fan and Peng, 2004).

The rest of the paper 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 Appendices A and B.

2. Bayesian information criterion with unpenalized estimators

2.1. The Bayesian information criterion

Let $(Y_i, \mathbf{X}_i), i = 1, \ldots, n$, be n independent and identically distributed observations, where $Y_i \in \mathbb{R}^1$ is the response of interest and $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^{\mathrm{T}} \in \mathbb{R}^d$ 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):

$$Y_i = \mathbf{X}_i^{\mathrm{T}} \beta + \varepsilon_i, \tag{2.1}$$

where ε_i is some random error with mean 0 and variance σ^2 and $\beta = (\beta_1, \ldots, \beta_d)^{\mathrm{T}} \in \mathbb{R}^d$ is the regression coefficient. The true regression coefficient is denoted as $\beta_0 = (\beta_{01}, \ldots, \beta_{0d})^{\mathrm{T}}$. Without loss of generality, we assume that $\beta_{0j} \neq 0$ for every $1 \leqslant j \leqslant d_0$ 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 predictors. Here d_0 is allowed to be either fixed or diverging to ∞ as $n \to \infty$.

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

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

where $C_n > 0$ is some positive constant to be discussed more carefully; see remark 1 in Section 2.4. As we can see, if $C_n = 1$, the modified BIC (2.2) reduces to the traditional BIC. With $C_n = 1$, Shao (1997) and Shi and Tsai (2002) have demonstrated that the above BIC can 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 the best of our knowledge).

2.2. The main challenge

Since the BIC (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 the 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 (2.2) is consistent with a fixed number of parameters. An important step to prove this conclusion is to show that the BIC (2.2) can differentiate the true model $\mathcal{S}_T$ from an arbitrary overfitted model (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 equation (2.2), we have

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

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 equation (2.3) is guaranteed to be positive as long as the sample size is sufficiently large. Thus, we have

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

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

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

As a result, we know that the BIC (2.2) can differentiate the true model $\mathcal{S}_T$ from *every* overfitted model consistently. Nevertheless, establishing result (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, inequality (2.4) no longer implies the desired conclusion (2.5). Thus, special techniques must be developed to overcome this issue; for details see Appendices A and B.

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 $\mathbf{X}_i$. To study the asymptotic behaviour of the modified BIC (2.2), the following technical conditions are needed.

Condition 1. $\mathbf{X}_i$ has componentwise finite fourth-order moment, i.e. $\max_{1 \leqslant j \leqslant d}\{E(X_{ij}^4)\} < \infty$.

Condition 2. There is a positive number κ such that $\tau_{\min}(\Sigma) \geqslant \kappa$ for every $d > 0$.

Condition 3. The predictor dimension satisfies that $\limsup(d/n^{\kappa^*}) < 1$ for some $\kappa^* < 1$.

Condition 4. $\sqrt{[n/\{C_n d \log(n)\}]} \liminf_{n\to\infty}(\min_{j \in \mathcal{S}_T} |\beta_{0,j}|) \to \infty$, and $C_n d \log(n)/n \to 0$.

Condition 1 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 2 is also a reasonable condition that is widely assumed in the literature (Fan and Peng, 2004; Huang *et al.*, 2007). Otherwise, the predictors become linearly dependent on each other asymptotically. Condition 3 characterizes the speed at which the predictor dimension is allowed to diverge to ∞. Condition 4 puts a requirement on the size of the non-zero coefficients. Intuitively, if some non-zero coefficients converge to 0 too fast, those non-zero coefficients can hardly be estimated accurately; see Fan and Peng (2004) and Huang *et al.* (2007). Lastly, condition 4 also constrains that the value of the diverging constant C_n cannot be too large. Intuitively, a C_n-value that is too large will lead to seriously underfitted models.

2.4. Bayesian information criterion 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 about ε's tail heaviness; see Huang *et al.* (2007).

Theorem 1. Assume technical conditions 1–4, $C_n \to \infty$, and that ε is normally distributed; we then have

$$P\{\min_{\mathcal{S} \not\supset \mathcal{S}_T} (\mathrm{BIC}_{\mathcal{S}}) > (\mathrm{BIC}_{\mathcal{S}_F})\} \to 1.$$

By theorem 1 we know that the minimal BIC value that is 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 1, any underfitted model cannot be selected by the BIC (2.2) because it is not even as favourable 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 rate of divergence can be arbitrarily slow. For example, $C_n = \log\{\log(d)\}$ is used for all our numerical experiments and the simulation results are quite encouraging.

Theorem 2. Assume technical conditions 1–4, $C_n \to \infty$, and that ε is normally distributed; we then have

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

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

2.5. Bayesian information criterion 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 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 $\hat{\beta}_\lambda = (\hat{\beta}_{\lambda,1}, \ldots, \hat{\beta}_{\lambda,d})^{\mathrm{T}}$. For example, $\hat{\beta}_\lambda$ becomes the SCAD estimator, if $p_{\lambda,j}(\cdot)$ is a function with its first-order derivative given by $\dot{p}_{\lambda,j}(t) = \lambda\{I(t \leqslant \lambda) + I(t > \lambda)(a\lambda - t)_+\}/\{(a-1)\lambda\}$ with $a = 3.7$ and $(t)_+ = t\, I(t > 0)$; 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.*, 2008).

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

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

with $\hat{\sigma}_\lambda^2 = \mathrm{SSE}_\lambda / n$ and $\mathrm{SSE}_\lambda = \|\mathbb{Y} - \mathbb{X}\hat{\beta}_\lambda\|^2$. Let $\mathcal{S}_\lambda = \{j : \hat{\beta}_{\lambda,j} \neq 0\}$ be the model that is identified by $\hat{\beta}_\lambda$. Here we should carefully differentiate two notations, i.e. SSE_λ and $\mathrm{SSE}_{\mathcal{S}_\lambda}$. Specifically, SSE_λ is the residual sum of squares that is associated with the shrinkage estimate $\hat{\beta}_\lambda$ and $\mathrm{SSE}_{\mathcal{S}_\lambda}$ is the residual sum of squares that is associated with the unpenalized estimator based on $\mathcal{S}_\lambda$. By

definition, we know immediately that $\mathrm{SSE}_\lambda \geqslant \mathrm{SSE}_{\mathcal{S}_\lambda}$. Thus, we have $\mathrm{BIC}_\lambda \geqslant \mathrm{BIC}_{\mathcal{S}_\lambda}$. Then, the optimal tuning parameter is given by $\hat{\lambda} = \arg\min_\lambda(\mathrm{BIC}_\lambda)$, which identifies the model $\mathcal{S}_{\hat{\lambda}}$.

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

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

where $\hat{\beta}_{\mathcal{S}_T} = (n^{-1}\mathbb{X}_{\mathcal{S}_T}\mathbb{X}_{\mathcal{S}_T}^{\mathrm{T}})^{-1}(n^{-1}\mathbb{X}_{\mathcal{S}_T}^{\mathrm{T}}\mathbb{Y})$, $\dot{p}_\lambda(|\hat{\beta}_{\lambda_n,a}|) = \{\dot{p}(|\hat{\beta}_{\lambda_n,j}|)\}_{j=1}^{d_0}$ 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

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

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

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

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 similar techniques to those in lemma 3 of Wang *et al.* (2007b) and verify that $\|\dot{p}_\lambda(\hat{\beta}_{\lambda,a})\| = 0$ with probability tending to 1. Thus, 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 $\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^{\mathrm{T}}$ 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 $\lambda_n\sqrt{n} \to 0$ 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 that $(n/d)\lambda_n = (n/d)^\delta \to \infty$ under condition 3. If we can further assume (for example) that the κ^* in condition 3 is no larger than $\frac{1}{4}$ (see for example theorem 1 in Fan and Peng (2004)), we then have $n^{1/2}\lambda_n < n^{1/2}n^{3\delta/4-3/4} = n^{3\delta/4-1/4} \to 0$ 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, assumption (2.10) is also reasonable for the adaptive lasso estimator under appropriate conditions.

Remark 4. Requiring $\kappa^* \leqslant \frac{1}{4}$ 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 1–4, $C_n \to \infty$, that ε is normally distributed and also condition (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 (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 and a diverging number of parameters.

3. Numerical studies

Simulation experiments are conducted to confirm our theoretical findings. Here we report only 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 $\mathbf{X}_i$ 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 set-up, a total of 500 simulation replications are conducted.

3.1. Example 1

The first example is from Fan and Peng (2004). More specifically, we take $d = [4n^{1/4}] - 5$, $\beta = (11/4, -23/6, 37/12, -13/9, 1/3, 0, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^d$ 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 non-zero 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 of MRME, MS and CM, we refer to Fan and Li (2001) and Wang and Leng (2007). For comparison, both SCAD (Fan and Li, 2001) and the adaptive lasso (Zou, 2006) 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 Figs 1(a)–1(c). 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 because of its overfitting effect. In contrast, that of the BIC approaches 100% quickly, which clearly confirms the consistency of the BIC proposed. As a direct consequence, the MRME- and MS-values of BIC are consistently smaller than that of GCV for both SCAD and the adaptive lasso.

3.2. 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 $1 \leqslant j \leqslant d_0$ from a uniform distribution on $[0.5, 1.5]$. The detailed results are summarized in Figs 1(d)–1(f). The findings are similar to those in example 1. This example further confirms that the BIC (2.2) is still consistent even if d_0 is diverging.

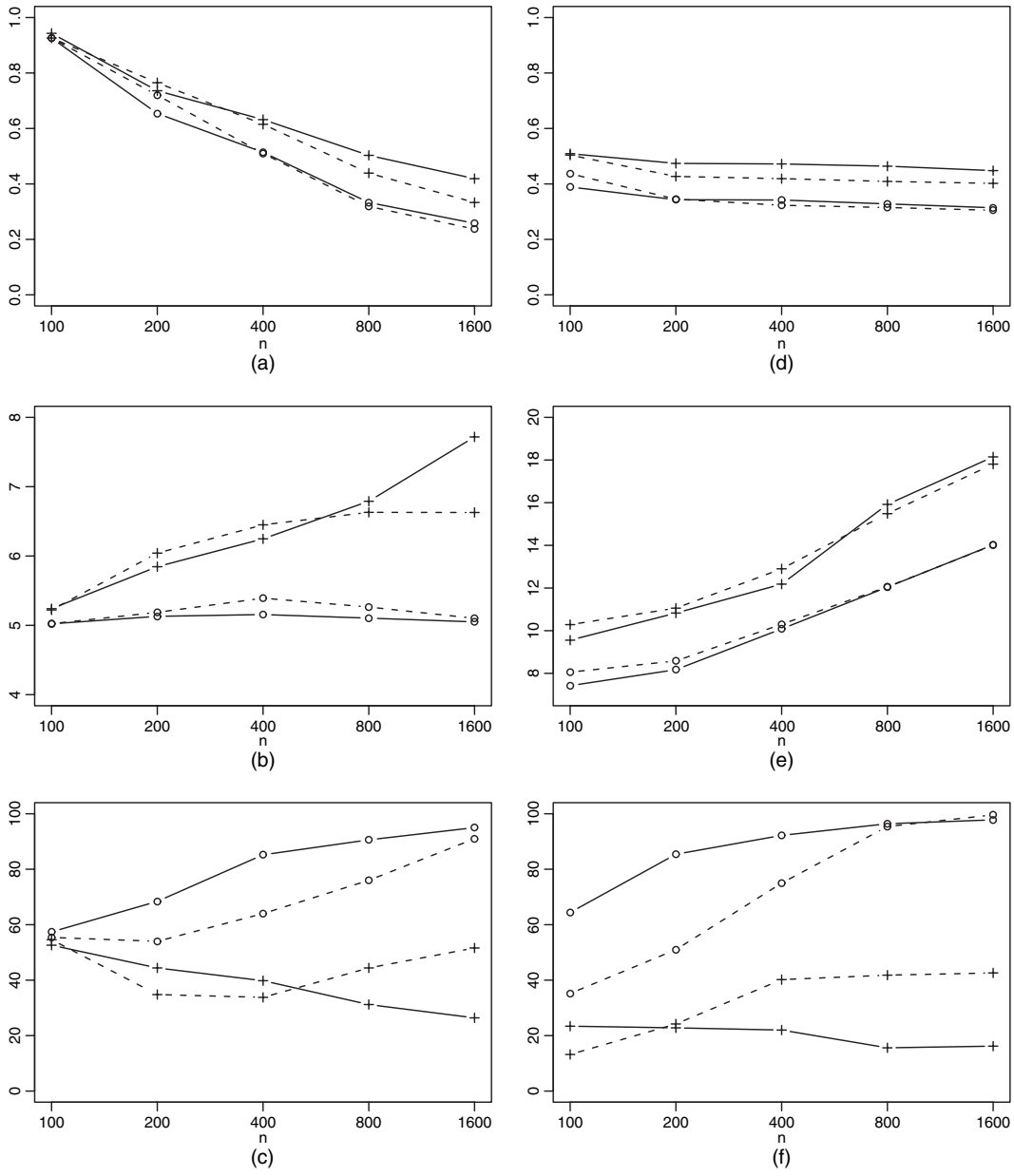

Fig. 1. Detailed simulation results with normal ε (– ∘ –, adaptive lasso plus BIC; – + –, adaptive lasso plus GCV; - - ∘ - -, SCAD plus BIC; - - + - -, SCAD plus GCV): (a) example 1, MRME; (b) example 1, MS; (c) example 1, CM; (d) example 2, MRME; (e) example 2, MS; (f) example 2, CM

3.3. Example 3

As our concluding example, we reanalyse the gender discrimination data as studied by Fan and Peng (2004), where detailed information about the data set 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 repre-

Table 1. Analysis results of the gender discrimination data set

Variable	*Ordinary least squares*	*Results from the adaptive lasso*		*Results from SCAD*	
		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

sent 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 SCAD and the adaptive lasso to the data set with both GCV and BIC as tuning parameter selectors. The detailed results are given in Table 1. As we can see, regardless of the estimation method (i.e. the adaptive lasso or SCAD), the BIC method typically yields more sparse solutions than the GCV method does, which is a pattern that is consistent with our simulation experience. In addition to that, except for the adaptive lasso method 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 remark that the normality assumption that is used for ε is mainly for simplification of the proof. In fact, our numerical experience indicates that the theorem results are reasonably insensitive to this assumption. For example, if we replace the normally distributed ε in the simulation study by a double-exponentially distributed ε, the final simulation results are almost identical. Secondly, the model set-up that is 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, 2008) 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 (2.2) is no longer well defined (owing to the $\log(\hat{\sigma}^2_{\mathcal{S}_F})$ term). Thus, how to define a sensible BIC with $p > n$ by itself is still an interesting question that is open for discussion.

Acknowledgements

The authors are very grateful to the Joint Editor, the Associate Editor and two referees for their careful reading and insightful comments, which led to a substantially improved manuscript. Hansheng Wang is supported in part by National Natural Science Foundation of China grant

10771006. Bo Li is supported in part by National Natural Science Foundation of China grants 10801086, 70621061 and 70831003. Chenlei Leng is supported by National University of Singapore academic research grants and a grant from the National University of Singapore Risk Management Institute.

Appendix A: Proof of theorem 1

Recall that $\tilde{\beta}$ is the unpenalized full model estimator. Under conditions 1–3, and by the results of Bai and Silverstein, (2006), we know immediately that

$$E(\|\tilde{\beta}-\beta_0\|^2)=\mathrm{tr}\{\mathrm{cov}(\tilde{\beta})\}=\sigma^2\,\mathrm{tr}\{(\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}\}\leqslant dn^{-1}\sigma^2\tau_{\min}^{-1}(n^{-1}\mathbb{X}^{\mathrm{T}}\mathbb{X})=O_p(d/n).$$

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

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

By technical condition 4, we know that the right-hand side of inequality (A.1) is guaranteed to be positive with probability tending to 1. Next, we follow the basic idea of Wang *et al.* (2007b) and consider

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

The right-hand side of this equation can be written as

$$\begin{aligned}\min_{\mathcal{S}\not\supset\mathcal{S}_T}\left[\log\left\{1+\frac{(\hat{\beta}^{(\mathcal{S})}-\tilde{\beta})^{\mathrm{T}}(n^{-1}\mathbb{X}^{\mathrm{T}}\mathbb{X})(\hat{\beta}^{(\mathcal{S})}-\tilde{\beta})}{\hat{\sigma}_{\mathcal{S}_F}^2}\right\}\right]-\frac{d\log(n)}{n}C_n.&\\ \geqslant\min_{\mathcal{S}\not\supset\mathcal{S}_T}\left\{\log\left(1+\frac{\hat{\tau}_{\min}\|\hat{\beta}^{(\mathcal{S})}-\tilde{\beta}\|^2}{\hat{\sigma}_{\mathcal{S}_F}^2}\right)\right\}-\frac{d\log(n)}{n}C_n,&\end{aligned}\tag{A.2}$$

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

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

Because $d\log(n)/n\to0$ under condition 4, thus with probability tending to 1 we must have $\log(2)-C_nd\log(n)/n>0$. Consequently, as long as we can show that, with probability tending to 1,

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

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

$$\begin{aligned}&\geqslant\frac{\tau_{\min}}{\sigma^2}\left\{\min_{j\in\mathcal{S}_T}(\beta_{0,j}^2)-O_p\left(\frac{d}{n}\right)\right\}\{1+o_p(1)\}-\frac{d\log(n)}{n}C_n\\ &=\frac{C_nd\log(n)}{n}\frac{\tau_{\min}}{\sigma^2}\left\{\frac{n}{C_nd\log(n)}\min_{j\in\mathcal{S}_T}(\beta_{0j}^2)\right\}\{1+o_p(1)\}-\frac{d\log(n)}{n}C_n,\end{aligned}$$

which is guaranteed to be positive asymptotically under condition 4. This proves that, with probability tending to 1, the right-hand side of expression (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

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

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

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

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

$$\begin{aligned}
\mathrm{SSE}_{\mathcal{S}_T} - \mathrm{SSE}_{\mathcal{S}} &= (n^{-1/2} \hat{\mathcal{E}}^{\mathrm{T}} \tilde{\mathbb{X}}_{\mathcal{S}^c})(n^{-1} \tilde{\mathbb{X}}_{\mathcal{S}^c}^{\mathrm{T}} \tilde{\mathbb{X}}_{\mathcal{S}^c})^{-1} (n^{-1/2} \tilde{\mathbb{X}}_{\mathcal{S}^c} \hat{\mathcal{E}}) \\
&\leqslant \hbar_{\max}^{\mathcal{S}^c} \|n^{-1/2} \hat{\mathcal{E}}^{\mathrm{T}} \tilde{\mathbb{X}}_{\mathcal{S}^c}\|^2 = \hbar_{\max}^{\mathcal{S}^c} \sum_{j \in \mathcal{S}^c} (n^{-1/2} \hat{\mathcal{E}}^{\mathrm{T}} \tilde{\mathbb{X}}_j)^2 \\
&\leqslant \hbar_{\max}^{\mathcal{S}^c} \max_{j \in \mathcal{S}^c} (n^{-1/2} \hat{\mathcal{E}}^{\mathrm{T}} \tilde{\mathbb{X}}_j)^2 |\mathcal{S}^c| \\
&\leqslant \hbar_{\max}^{\mathcal{S}^c} \max_{j \in \mathcal{S}_F \backslash \mathcal{S}_T} \{(n^{-1/2} \hat{\mathcal{E}}^{\mathrm{T}} \tilde{\mathbb{X}}_j)^2\} |\mathcal{S}^c|,
\end{aligned}$$

where $\hbar_{\max}^{\mathcal{S}^c} = \tau_{\min}^{-1}(n^{-1} \tilde{\mathbb{X}}_{\mathcal{S}^c}^{\mathrm{T}} \tilde{\mathbb{X}}_{\mathcal{S}^c})$; recall that $\mathbb{X}_j$ 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 $\tau_{\min}(n^{-1} \tilde{\mathbb{X}}_{\mathcal{S}^c}^{\mathrm{T}} \tilde{\mathbb{X}}_{\mathcal{S}^c}) \geqslant \tau_{\min}(n^{-1} \tilde{\mathbb{X}}_{\mathcal{S}_T^c}^{\mathrm{T}} \tilde{\mathbb{X}}_{\mathcal{S}_T^c}) \doteq (\hbar_{\max}^{\mathcal{S}_T^c})^{-1}$. We then must have

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

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

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

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 $\gamma^* = -(\mathbb{X}_{\mathcal{S}_T}^{\mathrm{T}} \mathbb{X}_{\mathcal{S}_T})^{-1} \mathbb{X}_{\mathcal{S}_T}^{\mathrm{T}} \mathbb{X}_{\mathcal{S}_T^c} \gamma$. Thus, we know that

$$\begin{aligned}
\tau_{\min}(n^{-1} \tilde{\mathbb{X}}_{\mathcal{S}_T^c}^{\mathrm{T}} \tilde{\mathbb{X}}_{\mathcal{S}_T^c}) &= n^{-1} \|\mathbb{X}_{\mathcal{S}_T^c} \gamma + \mathbb{X}_{\mathcal{S}_T} \gamma^*\|^2 = n^{-1} \|\mathbb{X} \alpha\|^2 = \alpha^{\mathrm{T}} (n^{-1} \mathbb{X}^{\mathrm{T}} \mathbb{X}) \alpha \\
&\geqslant \tau_{\min}(n^{-1} \mathbb{X}^{\mathrm{T}} \mathbb{X}) \|\alpha\|^2 \geqslant \tau_{\min}(n^{-1} \mathbb{X}^{\mathrm{T}} \mathbb{X}) \geqslant \kappa
\end{aligned} \tag{B.2}$$

with probability tending to 1. Here $\alpha = (\gamma^{*\mathrm{T}}, \gamma^{\mathrm{T}})^{\mathrm{T}}$ satisfies that $\|\alpha\| \geq 1$. This implies that $\hbar_{\max}^{\mathcal{S}^c} \leqslant \kappa^{-1}$. 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

$$\begin{aligned}
\sigma_j^2 &= n^{-1} \operatorname{tr}\{E(\mathcal{E}^{\mathrm{T}} Q_{\mathcal{S}_T} \tilde{\mathbb{X}}_j \tilde{\mathbb{X}}_j^{\mathrm{T}} Q_{\mathcal{S}_T} \mathcal{E})\} = n^{-1} \operatorname{tr}\{E(\mathcal{E}^{\mathrm{T}} \tilde{\mathbb{X}}_j \tilde{\mathbb{X}}_j^{\mathrm{T}} \mathcal{E})\} \\
&= n^{-1} \operatorname{tr}\{E(\mathcal{E}\mathcal{E}^{\mathrm{T}}) \tilde{\mathbb{X}}_j \tilde{\mathbb{X}}_j^{\mathrm{T}}\} = n^{-1} \sigma^2 \|\tilde{\mathbb{X}}_j\|^2 \leqslant n^{-1} \sigma^2 \|\mathbb{X}_j\|^2 < (1 + \varphi) \sigma^2
\end{aligned}$$

with probability tending to 1 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 $\mathrm{var}(X_j) = 1$. Then, the right-hand side of inequality (B.1) can be further bounded by

$$\max_{\mathcal{S}^c \subset \mathcal{S}_F \backslash \mathcal{S}_T} \left(\frac{\mathrm{SSE}_{\mathcal{S}_T} - \mathrm{SSE}_{\mathcal{S}}}{|\mathcal{S}^c|} \right) \leqslant (1 + \varphi) \sigma^2 \kappa^{-1} \max_{j \in \mathcal{S}_F \backslash \mathcal{S}_T} \{\chi_j^2(1)\}, \tag{B.3}$$

where $\chi_j^2(1) = \sigma_j^{-2}(n^{-1}\hat{\mathcal{E}}^{\mathrm{T}}\tilde{\mathbb{X}}_j)^2$ follows a χ^2-distribution with 1 degree of freedom. We should note that these χ^2-variables $\chi_j^2(1)$, $j \in \mathcal{S}_F \backslash \mathcal{S}_T$, may be dependent. Nevertheless, we can proceed by using Bonferroni's inequality to obtain

$$\begin{aligned} P[\max_{j \in \mathcal{S}_F \backslash \mathcal{S}_T} \{\chi_j^2(1)\} > 2\log(d)] &\leqslant d\, P\{\chi_1^2(1) > 2\log(d)\} \\ &= (2\pi)^{-1/2} d \int_{2\log(d)}^{\infty} x^{-1/2} \exp\left(-\frac{x}{2}\right) \mathrm{d}x \\ &\leqslant \frac{(2\pi)^{-1/2} d}{\sqrt{\{2\log(d)\}}} \int_{2\log(d)}^{\infty} \exp\left(-\frac{x}{2}\right) \mathrm{d}x = \frac{(2\pi)^{-1/2}}{\sqrt{\{2\log(d)\}}}, \end{aligned}$$

which implies that $\max_{j \in \mathcal{S}_F \backslash \mathcal{S}_T} \{\chi_j^2(1)\} \leqslant 2\log(d)$ with probability tending to 1 as $d \to \infty$. In conjunction with inequality (B.3), we see that

$$\max_{\mathcal{S}^{\mathrm{c}} \subset \mathcal{S}_F \backslash \mathcal{S}_T} \left(\frac{\mathrm{SSE}_{\mathcal{S}_T} - \mathrm{SSE}_{\mathcal{S}}}{|\mathcal{S}^{\mathrm{c}}|} \right) \leqslant 2(1+\varphi)\sigma^2 \kappa^{-1} \log(d) \tag{B.4}$$

with probability tending to 1. Consequently, we know that $\max_{\mathcal{S}^{\mathrm{c}} \subset \mathcal{S}_F \backslash \mathcal{S}_T} (\mathrm{SSE}_{\mathcal{S}_T} - \mathrm{SSE}_{\mathcal{S}}) = O_p\{d\log(d)\} = o(n)$ by condition 3. Thus, the following Taylor series expansion holds:

$$\begin{aligned} n(\mathrm{BIC}_{\mathcal{S}} - \mathrm{BIC}_{\mathcal{S}_T}) &= n \log\left\{ 1 + \frac{n^{-1}(\mathrm{SSE}_{\mathcal{S}} - \mathrm{SSE}_{\mathcal{S}_T})}{\hat{\sigma}^2_{\mathcal{S}_T}} \right\} + |\mathcal{S}^{\mathrm{c}}| \log(n) C_n \\ &= \frac{1}{\hat{\sigma}^2_{\mathcal{S}_T}} (\mathrm{SSE}_{\mathcal{S}} - \mathrm{SSE}_{\mathcal{S}_T}) + |\mathcal{S}^{\mathrm{c}}| \log(n) C_n + o_p(1) \\ &= \frac{1}{\sigma^2_{\mathcal{S}_T}} (\mathrm{SSE}_{\mathcal{S}} - \mathrm{SSE}_{\mathcal{S}_T})\{1 + o_p(1)\} + |\mathcal{S}^{\mathrm{c}}| \log(n) C_n + o_p(1). \end{aligned}$$

Then, by inequality (B.4), we know that the right-hand side of this equation can be uniformly bounded by

$$\geqslant |\mathcal{S}^{\mathrm{c}}| (C_n \log(n) - 2(1+\varphi)\kappa^{-1} \log[d\{1 + o_p(1)\}]).$$

Consequently, we know that

$$\begin{aligned} \max_{\mathcal{S} \supset \mathcal{S}_T, \mathcal{S} \neq \mathcal{S}_T} \left\{ \frac{n}{|\mathcal{S}^{\mathrm{c}}|} (\mathrm{BIC}_{\mathcal{S}} - \mathrm{BIC}_{\mathcal{S}_T}) \right\} &\geqslant C_n \log(n) - 2(1+\varphi)\kappa^{-1} \log[d\{1 + o_p(1)\}] \\ &\geqslant \log(n)\{C_n - 2(1+\varphi)\kappa^{-1}\kappa^*\}\{1 + o_p(1)\} \end{aligned} \tag{B.5}$$

with probability tending to 1, where the last inequality is due to condition 3. By theorem assumption, we know that $C_n \to \infty$. This implies that, with probability tending to 1, $\max_{\mathcal{S} \supset \mathcal{S}_T, \mathcal{S} \neq \mathcal{S}_T} (\mathrm{BIC}_{\mathcal{S}} - \mathrm{BIC}_{\mathcal{S}_T})$ must be positive. This proves the theorem's 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, $\Omega_0(\Omega_-, \Omega_+)$ collects all λ-values, which produces correctly underfitted or overfitted models. We follow Wang *et al.* (2007b) and Wang and Leng (2007) and establish the statement of the theorem via the following two steps.

C.1. Case 1: underfitted model (i.e. $\lambda \in \Omega_-$)

Firstly, under 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

$$\begin{aligned} \inf_{\lambda \in \Omega_-} (\mathrm{BIC}_\lambda) - \mathrm{BIC}_{\lambda_n} &\geqslant \inf_{\lambda \in \Omega_-} (\mathrm{BIC}_{\mathcal{S}_\lambda}) - \mathrm{BIC}_{\mathcal{S}_T} + o_p\{\log(n)/n\} \\ &\geqslant \min_{\mathcal{S} \not\supset \mathcal{S}_T} (\mathrm{BIC}_{\mathcal{S}}) - \mathrm{BIC}_{\mathcal{S}_T} + o_p\{\log(n)/n\}. \end{aligned} \tag{C.1}$$

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 $P(\mathrm{BIC}_{\mathcal{S}_F} - \mathrm{BIC}_{\mathcal{S}_T} \geqslant 0) \to 1$. Consequently, we know that $P\{\inf_{\lambda\in\Omega_-}(\mathrm{BIC}_\lambda) - \mathrm{BIC}_{\lambda_n} > 0\} \to 1$.

C.2. Case 2 overfitted model (i.e. $\lambda \in \Omega_+$)

We can argue similarly to the overfitting case to obtain the inequality

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

By inequality (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 to be positive asymptotically. As a consequence, $P\{\inf_{\lambda\in\Omega_+}(\mathrm{BIC}_\lambda) - \mathrm{BIC}_{\lambda_n} > 0\} \to 1$. This completes the proof.

References

Bai, Z. D. and Silverstein, J. W. (2006) *Spectral Analysis of Large Dimensional Random Matrices*. Beijing: Science Press.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.

Fan, J. and Li, R. (2006) Statistical challenges with high dimensionality: feature selection in knowledge discovery. In *Proc. Int. Congr. Mathematicians*, vol. III (eds M. Sanz-Sole, J. Soria, J. L. Varona and J. Verdera), pp. 595–622. Zurich: European Mathematical Society.

Fan, J. and Peng, H. (2004) On non-concave penalized likelihood with diverging number of parameters. *Ann. Statist.*, **32**, 928–961.

Fu, W. J. (1998) Penalized regression: the bridge versus the LASSO. *J. Computnl Graph. Statist.*, **7**, 397–416.

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

Huang, J., Ma, S. and Zhang, C. H. (2007) Adaptive LASSO for sparse high-dimensional regression models. *Technical Report 374*. Department of Statistics and Actuarial Science, University of Iowa, Iowa City. (Available from `www.stat.uiowa.edu/techrep/`.)

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

Shao, J. (1997) An asymptotic theory for linear model selection. *Statist. Sin.*, **7**, 221–264.

Shi, P. and Tsai, C. L. (2002) Regression model selection—a residual likelihood approach. *J. R. Statist. Soc.* B, **64**, 237–252.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.

Wang, H. and Leng, C. (2007) Unified LASSO estimation via least squares approximation. *J. Am. Statist. Ass.*, **101**, 1418–1429.

Wang, H., Li, G. and Tsai, C. L. (2007a) Regression coefficient and autoregressive order shrinkage and selection via the lasso. *J. R. Statist. Soc.* 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. (2008) SCAD-penalized regression in high-dimensional partially linear models. *Ann. Statist.*, to be published.

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. *Statist. Probab. Lett.*, **76**, 520–530.

Zou, H. (2006) The adaptive LASSO and its oracle properties. *J. Am. Statist. Ass.*, **101**, 1418–1429.

Zou, H. and Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.*, **36**, 1509–1566.