Journal of Multivariate Analysis 102 (2011) 1141–1151

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Consistent tuning parameter selection in high dimensional sparse linear regression

Tao Wang, Lixing Zhu *

Hong Kong Baptist University, Hong Kong, China

ARTICLE INFO

Article history:
Received 1 June 2010
Available online 28 March 2011

AMS 2000 subject classifications:
62H15
62G20

Keywords:
Adaptive Elastic Net
Bayesian information criterion
High dimensionality
Sure independence screening
Tuning parameter selection
Variable selection

ABSTRACT

An exhaustive search as required for traditional variable selection methods is impractical in high dimensional statistical modeling. Thus, to conduct variable selection, various forms of penalized estimators with good statistical and computational properties, have been proposed during the past two decades. The attractive properties of these shrinkage and selection estimators, however, depend critically on the size of regularization which controls model complexity. In this paper, we consider the problem of consistent tuning parameter selection in high dimensional sparse linear regression where the dimension of the predictor vector is larger than the size of the sample. First, we propose a family of high dimensional Bayesian Information Criteria (HBIC), and then investigate the selection consistency, extending the results of the extended Bayesian Information Criterion (EBIC), in Chen and Chen (2008) to ultra-high dimensional situations. Second, we develop a two-step procedure, the SIS + AENET, to conduct variable selection in $p > n$ situations. The consistency of tuning parameter selection is established under fairly mild technical conditions. Simulation studies are presented to confirm theoretical findings, and an empirical example is given to illustrate the use in the internet advertising data.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

High dimensional statistical problems arise frequently in diverse fields of contemporary scientific research and knowledge discovery, such as genomics and machine learning, where researchers very often encounter massive and high-throughput data with tens or even hundreds of thousands of variables [7]. Typically, the predictor dimension p is comparable to, or possibly even larger than the sample size n. Hence, p should be modeled as $p = p_n$, which diverges as n gets large. The ultra-high dimensionality refers to situations where p is of a non-polynomial order of n. To make high dimensional statistical inference possible, it is commonly assumed that the p-vector of regression coefficients is sparse in the sense that most of its components are zero, and nonzero components indicate important or active variables. With the sparsity, variable selection can improve the estimation accuracy by effectively identifying the subset of active variables and enhance the model interpretability with parsimonious representation. As a result, modern statistical applications call for high dimensional variable selection procedures which do play a pivotal role. In a recent work, Fan and Lv [9] have given an excellent review of variable selection in high dimensional feature space.

In regression analysis, the Bayesian Information Criterion (BIC, [13]) and its variants are among the most representable and widely applied criteria for variable selection. It is well-known that, under some fairly mild conditions, the BIC is consistent for model selection in the sense that, when there is a finite-dimensional correct model, it chooses the smallest correct model with probability approaching 1. However, it has been recognized in the high dimensional literature that

* Corresponding author.
E-mail address: lzhu@hkbu.edu.hk (L. Zhu).

0047-259X/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2011.03.007

traditional variable selection criteria, such as the BIC, are too liberal in a sense that they tend to over-select many spurious predictors, see [1,14,2] and references therein. Therefore, the asymptotic behaviors, such as selection consistency with increasing size of sample, of classical selection criteria cannot be guaranteed without further deep investigation. Nevertheless, substantial progress in this regard has been made recently in the high dimensional linear regression which then deserves a brief review. On one hand, [17] demonstrated that, with a slightly modified version of the traditional BIC, the MBIC, it is possible to identify the true model consistently in situations with a diverging number of parameters but $p < n$. On the other hand, with sound Bayesian motivation, Chen and Chen [2] developed a family of extended Bayesian Information Criteria, the EBIC, for model selection which are applicable to "large p and small n" paradigms. Specifically, the predictor dimension p is allowed to grow at a polynomial rate as the sample size n increases, or more formally $p = O(n^{\alpha})$ for some positive constant $\alpha > 0$. Their theoretical results further imply that the traditional BIC may no longer be consistent when p is of a larger order than $n^{1/2}$. These appealing results therefore enlarge the scope of applicability of classical BIC-like criteria. Then, it will be interesting to see whether some BIC-like criterion, especially the EBIC, can be extended to situations with non-polynomial dimensionality, that is, $\log(p) = O(n^{\kappa})$ for some constant $\kappa > 0$.

In practice, however, the best subset selection is combinatorial and hence computationally too expensive to implement. Thus, to conduct variable selection in high dimensional situations, various forms of penalized estimators with good statistical as well as computational properties, have been proposed during the past two decades. These popular methods include but are not limited to the LASSO [15], the SCAD [6], the Elastic Net [21] and the adaptive LASSO [20]. Recently, Zou and Zhang [22] introduced the adaptive Elastic Net, the AENET, of which the adaptive LASSO and the Elastic Net are two special cases. Under appropriate regularity conditions, they further established the so-called oracle property [6,10] of the AENET when p diverges with n but $p < n$. For those shrinkage and selection methods, however, the true model can be identified consistently only if the tuning parameters can be chosen appropriately. Therefore, consistent tuning parameter selection becomes the main challenge. In a recent paper, Wang et al. [18] have discovered that, with a fixed predictor dimension and for the SCAD estimator, the tuning parameters selected by the BIC can identify the true model consistently. When the MBIC is applied to choose the tuning parameters, similar results were further extended by Wang et al. [17] for general penalized estimators, to situations with a diverging number of parameters but $p < n$.

In $p > n$ situations, the story changes drastically because it remains unknown whether the oracle property still holds for those oracle-like estimators, and the performance of those penalized estimators is unsatisfactory possibly due to the simultaneous challenges of computational burden, statistical accuracy and algorithm stability. To bypass these difficulties, Fan and Lv [8] introduced the property of sure screening and developed the Sure Independence Screening method, the SIS, to reduce the ultra-high dimensionality p to a relatively large scale d with $d < n$. Given that the reduced model contains all the variables in the true model with an overwhelming probability, the AENET or other lower dimensional penalized methods can be applied directly. Then, how to consistently select the tuning parameters in $p > n$ situations remains an interesting open problem for discussion.

In this paper, we consider the variable selection problem in the high dimensional sparse linear regression. First, we propose a family of high dimensional Bayesian Information Criteria, the HBIC, and then investigate the selection consistency, extending the results for the EBIC to ultra-high dimensional situations. Second, we develop a two-step variable selection procedure, the Sure Independence Screening method followed by the adaptive Elastic Net, the SIS + AENET, and apply the EBIC or the HBIC to practically select the tuning parameters. The consistency of tuning parameter selection is established under some weak conditions. Finally, simulation experiments are studied to illustrate the finite sample performance. All technical proofs are relegated to the Appendix.

2. Main results

2.1. Basic notations

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed observations from the linear regression model

$$Y_i = X_i^T \boldsymbol{\beta} + \epsilon_i, \tag{2.1}$$

where $X_i = (X_{i1}, \ldots, X_{ip})^T \in \mathbb{R}^p$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ and $\epsilon_i \sim N(0, \sigma^2)$. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ be the response vector, and let $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ be the design matrix with jth column $\boldsymbol{X}_j = (X_{1j}, \ldots, X_{nj})^T$. Without loss of generality, we assume that the data are centered, so the intercept is not included in the regression function. We also note that most of the quantities and data objects in our discussion are functions of n, such as $p = p_n$, but this dependence on n is often left implicit, especially for n-vectors and matrices with n rows. For any subset $M \subseteq F = \{1, \ldots, p\}$, let $\boldsymbol{X}_M$ be the design matrix with columns $\{\boldsymbol{X}_j, j \in M\}$, and let $\boldsymbol{\beta}_M$ be the coefficient vector with components $\{\beta_j, j \in M\}$. Throughout, we call M a sub-model and denote $|M|$ its model size or dimension. Thus, F is the full model with dimension $|F| = p > n$. Let M_0 be the unique true model with true dimension $|M_0| = q < n$.

Actually, there is a question of non-identifiability for $\boldsymbol{\beta}$ in $p > n$ situations. Therefore, we introduce the widely-used notion of sparse eigenvalues [3,12] and then discuss the identifiability condition that is important to our theory development. If A is any square matrix, let $\phi(A)$ and $\Phi(A)$ denote the smallest and largest eigenvalues of A.

Definition 2.1. For some positive integer m, the m-sparse minimal eigenvalue and m-sparse maximal eigenvalue of $\boldsymbol{X}^T\boldsymbol{X}/n$ are defined as

$$\phi_n(m) = \min_{M:|M|=m} \phi\left(\frac{1}{n}\boldsymbol{X}_M^T\boldsymbol{X}_M\right), \quad \text{and} \quad \Phi_n(m) = \max_{M:|M|=m} \Phi\left(\frac{1}{n}\boldsymbol{X}_M^T\boldsymbol{X}_M\right).$$

Under the technical conditions (A)–(C) below, Lemma A.1 in the Appendix implies that, for any m satisfying $m\log(p) = o(n)$, model (2.1) is identifiable to order m in the sense that $c_1 < \phi_n(m) < \Phi_n(m) < c_2$ holds with probability approaching 1, for some constants $0 < c_1 < c_2 < \infty$. Given the m-order identifiability condition, it follows that the ordinary least squares estimator $\hat{\boldsymbol{\beta}}_M = (\boldsymbol{X}_M^T\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^T\boldsymbol{Y}$ is well-defined for any sub-model M such that $|M| \le m$. Let $\mathrm{RSS}_M = \|\boldsymbol{Y} - \boldsymbol{X}_M\hat{\boldsymbol{\beta}}_M\|_2^2$ be the corresponding residual sum of squares with $\|\cdot\|_2$ being the Euclidean norm.

2.2. Technical conditions

The following standard regularity conditions are imposed to facilitate the technical proofs. We have no intent to make our assumptions the weakest possible. There are some universe constants $0 < c_1 < c_2 < \infty$, $0 < \kappa < 1$ such that

(A) $X \sim N(0, \Sigma)$,
(B) $c_1 < \phi(\Sigma) < \Phi(\Sigma) < c_2$, and
(C) $p > n$, $\log(p) = O(n^\kappa)$.

Note that the normality condition (A) has been popularly assumed in the high dimensional literature to facilitate theory development, and similar conditions were considered in [8,16]. Condition (B) is the population analog of the sparse Riesz condition defined in [19]. Assume conditions (A)–(C), [19, Section 4] further show that, for any m satisfying $m\log(p) = o(n)$, the sparse Riesz condition with rank m, or in this paper the m-order identifiability condition holds with probability approaching 1, see Lemma A.1 and Corollary A.1 in the Appendix for a more detailed discussion.

2.3. High dimensional Bayesian information criterion

Based on the above notations, we review the following class of EBICs,

$$\mathrm{EBIC}_\gamma(M) = n\log\left(\frac{1}{n}\mathrm{RSS}_M\right) + \{\log(n) + 2\gamma\log(p)\}|M|, \quad 0 \le \gamma \le 1. \tag{2.2}$$

The EBIC is motivated from a sound Bayesian perspective, see [2] for details. Clearly, setting $\gamma = 0$ reverses the EBIC to the traditional BIC. Assume that $p = O(n^\alpha)$ for some constant $\alpha > 0$, it has been shown that, under the asymptotic identifiability condition [2, Condition 1, page 762], the EBIC is selection consistent as long as $\gamma \ge 1 - 1/(2\alpha)$. Therefore, the version EBIC_1 is consistent in situations where p grows at a polynomial rate, and the EBIC_0 or the traditional BIC, is likely to be inconsistent when p is of a larger order than $n^{1/2}$. However, whether the EBIC is still consistent in ultra-high dimensional situations, with $\log(p) = O(n^\kappa)$ for some $0 < \kappa < 1$, remains unknown. Thus, we propose a new family of BIC-like criteria for high dimensional variable selection as

$$\mathrm{HBIC}_\gamma(M) = n\log\left(\frac{1}{n}\mathrm{RSS}_M\right) + 2\gamma\log(p)|M|, \gamma \ge 1. \tag{2.3}$$

We have the following proposition on the selection consistency of the HBIC. The assumptions can be weakened at the expense of more complicated proofs. In particular, we can let $|M_0|$ increase with n and $\min_{j\in M_0}|\beta_j|$ decrease with n.

Proposition 2.1. *Assume the technical conditions* (A)–(C). *If* $\gamma > 1$, *then for any integer* K *satisfying* $K \ge |M_0|$ *and* $K\log(p) = o(n)$, *we have*

$$\lim_{n\to\infty} P\Big(\min_{M\neq M_0, |M|\le K} \mathrm{HBIC}_\gamma(M) > \mathrm{HBIC}_\gamma(M_0)\Big) = 1.$$

It follows from Proposition 2.1 that, if $\gamma > 1$, then subset selection with the EBIC_γ applied may identify the true model consistently, even if the predictor dimension p grows at a non-polynomial rate as the sample size n increases.

Remark 2.1. If we force $\gamma = 1$, then the HBIC reduces to the risk inflation criterion, the RIC, proposed in [11], which is minimax when the design matrix X is orthogonal in a sense that it asymptotically minimizes the maximum predictive risk inflation due to selection. The choice $\gamma = 1$ and the related minimax property were also discovered by Donoho and Johnstone [4] in the wavelet regression context. Since our main focus here is on consistent variable selection, it turns out $\gamma > 1$ is necessary to the theory development.

Remark 2.2. Assume a slightly weak identifiability condition, Chen and Chen [2] have investigated the selection consistency of the EBIC (2.2) in the case of deterministic designs. Major differences between our results and theirs are that they assumed implicitly that K is fixed, unlike us, and we require $\gamma > 1$ to handle the non-polynomial dimensionality, unlike them.

As an upper bound for the true dimension $|M_0|$, K relieves our searching endeavor by focusing exclusively on the class of sub-models $\{M : |M| \leq K\}$. However, the condition that K be fixed seems too stringent both in theory as well as in practice, especially in $p > n$ situations. Under fairly mild conditions, we are able to extend their results to situations with ultra-high dimensionality p and then for general K such that $K \log(p) = o(n)$. We also point out that, in theory, the EBIC_1 may not be selection consistent when p grows exponentially as n increases. This is because the EBIC_1 is asymptotically equivalent to the HBIC_1 as long as $\log(n)$ is of a lower order than $\log(p)$.

Remark 2.3. At first sight, the potential value of the result seems limited since the search is restricted to the class of sub-models $\{M : |M| \leq K\}$. This is not necessarily the case. On one hand, the result itself is interesting because it sheds some light on the difficulty of the problem of ultra-high dimensional variable selection. On the other hand, the result here can be used to further investigate the consistency of tuning parameter selection which will be discussed in the next section. This latter important and practical point, however, is not well-studied in [2].

2.4. Consistency of tuning parameter selection

In high dimensional situations, an exhaustive search as required for traditional variable selection methods is impractical. Adopting penalized estimators, such as the LASSO or the SCAD, makes it possible to deal with high dimensionality. The attractive properties of these shrinkage and selection estimators, however, depend critically on the size of regularization which controls the model's complexity. Thus, the appropriate choice of the tuning parameter is crucial to ensure consistent model selection. Throughout, we focus mainly on the recently proposed adaptive Elastic Net estimator (AENET, [22]) defined as

$$\hat{\boldsymbol{\beta}}(\lambda) = \left(1 + \frac{\lambda_2}{n}\right) \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_1 \sum_{j=1}^{p} \omega_j |\beta_j|. \tag{2.4}$$

Here $\lambda = (\lambda_1, \lambda_2)^T$ is the pair of tuning parameters, and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)^T$ is the adaptive data-driven weight vector. A simple solution is to construct the weights by $\omega_j = |\tilde{\beta}_j|^{-\tau}$, where τ is a pre-specified positive constant and $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^T$ is some initial consistent estimator of $\boldsymbol{\beta}$. If we force $\lambda_2 = 0$, then the AENET reduces to the adaptive LASSO. By setting $\omega_j = 1$ for all j, we get the Elastic Net estimator. Under weak regularity conditions, the AENET has the desired oracle property when $p = O(n^\alpha)$ for some $\alpha < 1$, and can properly handle the collinearity problem. Finally, fast algorithms are available for solving the AENET efficiently.

When p is much larger than n, as pointed out in [8], the performance of those penalized estimators is still unsatisfactory. To this end, Fan and Lv [8] introduced the Sure Independence Screening method, the SIS, to conduct variable selection in ultra-high dimensional situations. Motivated by their stimulating idea of screening consistency, [22] then proposed a two-step variable selection procedure as follows:

(S1) Apply the SIS to reduce the ultra-high dimensionality p to a moderate large scale d with $d < n$. Denote S the model generated by the SIS.
(S2) Fit the data by implementing the AENET to the reduced variables in S.

We refer to this procedure as the SIS + AENET. In step (S1), it is expected that the SIS can successfully screen unimportant variables with an overwhelming probability. Then, in step (S2), some lower dimensional penalized methods, such as the AENET, can be applied directly. However, the practical problem of tuning parameter selection was not thoroughly pursued in [22]. To this end, we propose to use the HBIC to select the tuning parameters. To be specific, for any $\lambda > 0$, let $M_\lambda = \{j : \hat{\beta}_j(\lambda) \neq 0\}$ be the model identified by the SIS + AENET procedure and let $\text{RSS}_\lambda = \|\boldsymbol{Y} - \boldsymbol{X}_{M_\lambda}\hat{\boldsymbol{\beta}}(\lambda)\|_2^2$. Following Wang et al. [17], we define the BIC-like criterion for tuning parameter selection as

$$\text{HBIC}_\gamma(M_\lambda) = n \log\left(\frac{1}{n}\text{RSS}_\lambda\right) + 2\gamma \log(p)|M_\lambda|, \quad \gamma \geq 1. \tag{2.5}$$

Let $\hat{\lambda}_\gamma = \arg\min_\lambda \text{HBIC}_\gamma(M_\lambda)$ be the optimal tuning parameter. The consistency of tuning parameter selection of the SIS + AENET procedure, with the criterion (2.5) applied, is stated in the following theorem.

Theorem 2.1. *Assume the technical conditions* (A)–(C). *Further assume that*
(D) $|\text{cov}(\beta_j^{-1}Y, X_j)| > 0$ *for any* $j \in M_0$,
(E) $d = O(n^\alpha)$ *for some constant* $0 < \alpha < 1$ *such that* $d \log(p) = o(n)$,
(E) $\lambda_2 = o(n)$, $\lambda_1 = o(\sqrt{n})$, *and* $\lambda_1\sqrt{d} = o(n)$.
If $\gamma > 1$, *then we have*

$$\lim_{n\to\infty} P(M_{\hat{\lambda}_\gamma} = M_0) = 1.$$

Remark 2.4. Condition (D) can be found in [8]. Intuitively, it rules out situations where an important variable is marginally uncorrelated with Y, but jointly correlated with Y. Note that Theorem 2.1 is a direct application of Proposition 2.1 on the selection consistency of the HBIC and Theorem 5.1 in [22] on the oracle property of the SIS+AENET procedure. The conditions

in the above theorem are rather mild. When the HBIC_γ, with $\gamma > 1$, is practically applied to select the tuning parameter, the AENET can select the true model consistently in ultra-high dimensional situations, given that the SIS can successfully screen unimportant variables with probability approaching 1.

Remark 2.5. In situations where $p = O(n^\alpha)$ for some positive constant $\alpha > 0$, similar results also hold when the EBIC_γ is applied to choose the tuning parameters, as long as $\gamma > 1 - 1/(2\alpha)$. The proof is almost the same and is omitted here. Here, we further note that the factor $\log(n)$, which is then of the same order as $\log(p)$, is of key importance to selection consistency. This point will be further confirmed through the simulation studies below.

3. Simulation studies

In this section, numerical studies are conducted to augment our theoretical findings. The proposed SIS+AENET procedure is applied to handle the high dimensionality, and then the BIC-like criterion (2.5) is used to select the tuning parameters. As suggested in [22], we construct the adaptive weights in the AENET estimator (2.4) by $\omega_j = \tilde{\beta}_j^{-1}$ for $j \in S$, where $\tilde{\boldsymbol{\beta}}$ is the Elastic Net estimator and S is the reduced model generated by the SIS. Here, we should point out that the results are not very sensitive to the choice of the initial estimator $\tilde{\boldsymbol{\beta}}$ as long as it is consistent. To evaluate the finite sample performance, five associated features of parameter estimation and variable selection are investigated: (1) the average of the estimated mean squared error, $\text{EMSE} = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$, (2) the average model size of the identified model $\hat{M} = \{j : \hat{\beta}_j \neq 0\}$, $\text{MS} = |\hat{M}|$, (3) the average true positive rate, $\text{TPR} = |\hat{M} \cap M_0|/|M_0|$, (4) the average false discovery rate, $\text{FDR} = |\hat{M} \cap M_0^c|/|\hat{M}|$, and (5) the percentage of correct identified models, CM. Ideally, we wish to have CM close to 1. Low EMSE values also indicate good performance. All summary statistics are computed based on 200 data replications.

Throughout, we simulate data from the linear model

$$Y = X^T\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N(0, \sigma^2). \tag{3.1}$$

For each simulation setup, the p-vector X of predictors is generated from a multivariate normal distribution $N(0, \Sigma)$ whose covariance matrix $\Sigma = (\Sigma_{ij})_{p\times p}$ has entries $\Sigma_{ij} = \rho^{|i-j|}$ for some constant $0 \leq \rho < 1$, for $i, j = 1, \ldots, p$. The maximum and minimum eigenvalues of Σ can be bounded using standard asymptotic results on Toeplitz matrix families.

Example 1. In this example, let $\rho = 0$, and $q = 6$. Thus, different predictors are independent and identically distributed according to $N(0, 1)$ and the true dimension is 6. We consider five such models with $n = 200$, and $p = 400, 1000, 2000, 5000, 10\,000$, respectively. For each model, the first 6 components of the p-vectors $\boldsymbol{\beta}$ are nonzero and are randomly generated as follows. We set $a = 4\log(n)/\sqrt{n}$ and pick nonzero coefficients of the form $(-1)^u(a + |z|)$, where u is randomly drawn from a Bernoulli distribution with parameter 0.4 and z is randomly generated from $N(0, 1)$. We then arrive at $\boldsymbol{\beta} = (1.738, -2.053, -1.653, 2.376, -2.107, 1.793, 0, \ldots, 0)^T$. After having these, we regard them as fixed values. As for the normal error ϵ, we set $\sigma = 2.785$ so that the theoretical signal-to-noise ratio, $\text{SNR} = \text{var}(X^T\boldsymbol{\beta})/\sigma^2$, of these five linear models is given by 3. We first implement the SIS to reduce the high dimensionality from p to a relatively large scale $d = 5n^{2/3} = 171$. Then, the AENET is applied to the reduced model, and the HBIC_γ (2.5) is used to select the best possible candidate model. Four values of γ are explored, 1, 1.25, 1.50 and 1.75. The whole solution path of the AENET, for each given λ_2, can be computed fast by a simple modification of the efficient LARS-EN algorithm [5,21], and a publicly available R package, *elasticnet*, is available at http://www.r-project.org. Details of the algorithm are omitted here. For simplicity, we first pick a relatively small grid of values for λ_2, say (0, 0.0001, 0.01, 0.1, 0.2, 0.5, 1), and then run the algorithm for each λ_2, as done in [21]. For comparison, we evaluate the performance of the EBIC_γ with $\gamma = 0, 0.25, 0.5, 0.75$, and 1. We also report the results of the modified BIC, the MBIC, proposed in [17] although it is in theory valid when $p < n$. We note that another BIC-like criterion, the delta-BIC method of Broman and Speed [1] can also be used in high dimensional situations. However, its performance is very similar to that of the MBIC and thus is not reported here. The detailed numerical results are summarized in Table 1.

Several conclusions can be drawn as follows. First, as we can see, the TPR values are almost 100% for all the simulation models we considered, which indicates that the SIS + AENET procedure works well. It is safe to use the SIS to reduce the high dimensionality p, that is up to 10 000, to about $d = 171$. It further implies that, the HBIC and its competitors, the EBIC and the MBIC, are applicable because they can select all the active predictors into the working model with high probability. Second, the EBIC_0, or the traditional BIC, fails to select the true model consistently. The FDR values are intolerably high: the lowest is 34.988% when $p = 400$, and the highest is 85.721% when $p = 10\,000$. Accordingly, the EMSE values are very large. Further, the $\text{EBIC}_{0.25}$ performs poorly. The CM values are approximately 40% when the predictor dimension $p = 5000$ and $p = 10\,000$, which indicates inconsistency of selection. Finally, the performance of the other criteria are very competitive. Nevertheless, we can see from the CM values that the $\text{EBIC}_{0.50}$ and the MBIC is less satisfactory. All these results are in agreement with the theory developed in [2] and those in Theorem 2.1 and Remark 2.5. We also vary the SNR by adjusting the error variance σ^2, the conclusions here are very general although the performance deteriorates as the SNR decreases.

Example 2. For the second simulated example, let $\rho = 0.5$ and $q = 3$. Then, the predictors are correlated with each other with a power decay correlation structure. We consider five such models with $n = 70$ and $p = 200, 400, 1000, 2000, 5000,$

Table 1
Example 1: $n = 200$, $q = 6$, $\rho = 0$, SNR $= 3$.

Predictor dimension	Tuning method	EMSE (mean ± sd)	MS (mean)	TPR (%)	FDR (%)	CM (%)
$p = 400$	EBIC_0	1.045 ± 0.518	12.565	100.000	34.988	27.000
	$\text{EBIC}_{0.25}$	0.634 ± 0.264	6.830	100.000	9.154	63.000
	$\text{EBIC}_{0.50}$	0.553 ± 0.197	6.205	100.000	2.785	83.500
	$\text{EBIC}_{0.75}$	0.532 ± 0.183	6.080	100.000	1.107	93.000
	EBIC_1	0.529 ± 0.181	6.030	100.000	0.428	97.000
	HBIC_1	0.545 ± 0.194	6.165	100.000	2.267	86.000
	$\text{HBIC}_{1.25}$	0.532 ± 0.183	6.080	100.000	1.107	93.000
	$\text{HBIC}_{0.50}$	0.529 ± 0.181	6.030	100.000	0.428	97.000
	$\text{HBIC}_{1.75}$	0.527 ± 0.179	6.010	100.000	0.142	99.000
	MBIC	0.592 ± 0.240	6.515	100.000	6.227	71.500
$p = 1000$	EBIC_0	1.634 ± 0.614	24.925	100.000	63.534	8.000
	$\text{EBIC}_{0.25}$	0.695 ± 0.260	7.170	100.000	11.921	54.500
	$\text{EBIC}_{0.50}$	0.630 ± 0.220	6.265	100.000	3.517	80.000
	$\text{EBIC}_{0.75}$	0.619 ± 0.224	6.115	100.000	1.589	90.000
	EBIC_1	0.622 ± 0.230	6.065	100.000	0.910	94.000
	HBIC_1	0.623 ± 0.221	6.185	100.000	2.535	84.500
	$\text{HBIC}_{1.25}$	0.619 ± 0.225	6.095	100.000	1.321	91.500
	$\text{HBIC}_{0.50}$	0.623 ± 0.234	6.055	100.000	0.785	94.500
	$\text{HBIC}_{1.75}$	0.624 ± 0.236	6.040	100.000	0.571	96.000
	MBIC	0.645 ± 0.218	6.530	100.000	6.462	68.500
$p = 2000$	EBIC_0	1.956 ± 0.550	32.230	99.667	75.669	0.500
	$\text{EBIC}_{0.25}$	0.799 ± 0.346	7.570	99.667	13.900	52.500
	$\text{EBIC}_{0.50}$	0.703 ± 0.277	6.275	99.667	4.011	74.500
	$\text{EBIC}_{0.75}$	0.704 ± 0.283	6.100	99.667	1.678	88.000
	EBIC_1	0.706 ± 0.282	6.060	99.667	1.142	90.000
	HBIC_1	0.702 ± 0.280	6.120	99.667	1.964	86.000
	$\text{HBIC}_{1.25}$	0.704 ± 0.280	6.080	99.667	1.392	89.000
	$\text{HBIC}_{0.50}$	0.708 ± 0.289	6.030	99.667	0.714	93.000
	$\text{HBIC}_{1.75}$	0.711 ± 0.292	6.020	99.667	0.571	94.000
	MBIC	0.722 ± 0.278	6.555	99.667	6.999	64.000
$p = 5000$	EBIC_0	2.296 ± 0.505	41.030	99.333	82.746	0.000
	$\text{EBIC}_{0.25}$	0.833 ± 0.393	7.870	99.333	17.036	39.500
	$\text{EBIC}_{0.50}$	0.762 ± 0.312	6.400	99.333	5.520	70.000
	$\text{EBIC}_{0.75}$	0.760 ± 0.319	6.080	99.333	1.592	86.500
	EBIC_1	0.760 ± 0.319	6.010	99.333	0.714	91.000
	HBIC_1	0.761 ± 0.318	6.120	99.333	2.069	84.500
	$\text{HBIC}_{1.25}$	0.759 ± 0.319	6.020	99.333	0.857	90.000
	$\text{HBIC}_{0.50}$	0.763 ± 0.320	5.995	99.333	0.500	92.500
	$\text{HBIC}_{1.75}$	0.762 ± 0.320	5.990	99.333	0.428	93.000
	MBIC	0.773 ± 0.321	6.725	99.333	8.894	57.500
$p = 10\,000$	EBIC_0	2.536 ± 0.514	48.295	99.000	85.721	0.000
	$\text{EBIC}_{0.25}$	0.869 ± 0.374	7.565	99.000	15.276	41.500
	$\text{EBIC}_{0.50}$	0.815 ± 0.347	6.370	99.000	5.528	65.500
	$\text{EBIC}_{0.75}$	0.819 ± 0.350	6.080	99.000	1.946	81.500
	EBIC_1	0.826 ± 0.357	6.000	98.917	0.928	87.000
	HBIC_1	0.818 ± 0.351	6.095	99.000	2.142	80.500
	$\text{HBIC}_{1.25}$	0.823 ± 0.351	6.010	99.000	1.000	87.000
	$\text{HBIC}_{0.50}$	0.829 ± 0.358	5.980	98.917	0.642	89.000
	$\text{HBIC}_{1.75}$	0.850 ± 0.455	5.930	98.417	0.357	90.500
	MBIC	0.824 ± 0.350	6.660	99.000	8.373	57.000

respectively. Let $\boldsymbol{\beta} = (3, 3, 3, 0, \ldots, 0)^T$, that is, the first 3 components of the p-vectors $\boldsymbol{\beta}$ are nonzero and the rest are zero. We set $\sigma = 2.872$ so that the theoretical SNR is about 6. As demonstrated in [8], the SIS may fail to correctly screen unimportant predictors in the presence of collinearity among predictors. To this end, we propose to use the iterated SIS, the ISIS which is an extension of the SIS, to overcome this particular weak point of the SIS. We call this procedure the ISIS+AENET and refer to Fan and Lv [8, Section 4] for a more detailed discussion. The ISIS is used to reduce the high dimensionality from p to a large scale $d = \lceil 4n^{2/3} \rceil = 68$. The numerical results are reported in Table 2.

Since the TPR values are very close to 100%, the ISIS + AENET procedure performs well. As we can see, the EBIC_0, the $\text{EBIC}_{0.25}$ and the MBIC fail to select the true model consistently, especially when $p = 400, 1000, 2000, 5000$ where the CM values are almost 0. We can also see that, for the $\text{EBIC}_{0.5}$, the performance is poor in all the simulation models we considered with low CM values and large EMSE values. This is because, according to Remark 2.5, the $\text{EBIC}_{0.5}$ may not be selection consistent in $p > n$ situations. As for the $\text{EBIC}_{0.75}$, the performance becomes better as p increases. Compared with the EBIC_1 and the $\text{HBIC}_{1.25}$, the performance of the HBIC_1 is less satisfactory, especially when $p = 200, 400, 1000$. This corroborates our theoretical findings of the EBIC as well as the HBIC. To ensure selection consistency, the factor $\log(n)$ is of crucial importance for the EBIC, and the requirement that $\gamma > 1$ is necessary for the HBIC. Finally, we point out

Table 2
Example 2: $n = 70$, $q = 3$, $\rho = 0.5$, SNR $= 6$.

Predictor dimension	Tuning method	EMSE (mean ± sd)	MS (mean)	TPR (%)	FDR (%)	CM (%)
$p = 200$	EBIC_0	4.210 ± 2.021	47.630	100.000	91.250	0.500
	$\text{EBIC}_{0.25}$	3.777 ± 2.423	42.065	100.000	73.620	13.500
	$\text{EBIC}_{0.50}$	3.072 ± 2.731	32.850	100.000	50.100	38.000
	$\text{EBIC}_{0.75}$	2.276 ± 2.675	19.795	99.833	28.991	58.500
	EBIC_1	1.482 ± 1.591	9.790	99.333	12.653	77.000
	HBIC_1	2.795 ± 2.762	28.315	100.000	42.281	46.000
	$\text{HBIC}_{1.25}$	1.721 ± 1.851	13.250	99.500	18.276	70.500
	$\text{HBIC}_{0.50}$	1.226 ± 1.249	6.295	99.333	7.027	84.000
	$\text{HBIC}_{1.75}$	1.065 ± 0.843	3.660	98.667	2.330	90.000
	MBIC	3.715 ± 2.471	41.340	100.000	71.245	15.500
$p = 400$	EBIC_0	4.268 ± 1.407	63.265	100.000	95.212	0.000
	$\text{EBIC}_{0.25}$	4.253 ± 1.428	62.985	100.000	94.747	0.500
	$\text{EBIC}_{0.50}$	4.134 ± 1.583	60.715	100.000	90.962	3.500
	$\text{EBIC}_{0.75}$	2.617 ± 2.263	29.825	99.833	41.642	53.000
	EBIC_1	1.337 ± 1.386	6.895	99.167	6.775	88.000
	HBIC_1	3.350 ± 2.100	44.790	100.000	65.071	30.000
	$\text{HBIC}_{1.25}$	1.683 ± 1.931	12.020	99.333	14.666	79.500
	$\text{HBIC}_{0.50}$	1.262 ± 1.051	6.245	99.167	5.445	90.500
	$\text{HBIC}_{1.75}$	1.149 ± 0.828	3.945	98.500	1.975	93.000
	MBIC	4.253 ± 1.428	62.960	100.000	94.744	0.500
$p = 1000$	EBIC_0	3.394 ± 0.810	60.450	100.000	94.999	0.000
	$\text{EBIC}_{0.25}$	3.393 ± 0.810	60.230	100.000	94.977	0.000
	$\text{EBIC}_{0.50}$	3.142 ± 1.117	53.190	99.833	82.663	11.500
	$\text{EBIC}_{0.75}$	1.515 ± 1.173	11.845	99.000	15.136	76.500
	EBIC_1	1.202 ± 0.819	3.945	98.167	2.183	90.500
	HBIC_1	1.862 ± 1.451	20.120	99.167	27.756	64.500
	$\text{HBIC}_{1.25}$	1.256 ± 0.955	4.930	98.333	3.866	88.500
	$\text{HBIC}_{0.50}$	1.237 ± 0.926	3.225	96.333	0.727	90.000
	$\text{HBIC}_{1.75}$	1.304 ± 1.060	3.160	94.500	0.477	88.500
	MBIC	3.393 ± 0.810	60.195	100.000	94.974	0.000
$p = 2000$	EBIC_0	3.313 ± 0.649	59.025	100.000	94.877	0.000
	$\text{EBIC}_{0.25}$	3.313 ± 0.649	58.780	100.000	94.855	0.000
	$\text{EBIC}_{0.50}$	2.941 ± 1.135	49.290	100.000	78.528	16.000
	$\text{EBIC}_{0.75}$	1.333 ± 1.057	7.825	98.333	9.362	80.000
	EBIC_1	1.187 ± 0.843	3.290	97.500	1.477	89.500
	HBIC_1	1.359 ± 1.069	8.765	98.500	11.043	78.000
	$\text{HBIC}_{1.25}$	1.176 ± 0.824	3.300	97.667	1.602	89.500
	$\text{HBIC}_{0.50}$	1.192 ± 0.858	2.940	96.833	0.875	89.000
	$\text{HBIC}_{1.75}$	1.315 ± 1.082	2.815	93.667	0.125	86.500
	MBIC	3.312 ± 0.649	58.745	100.000	94.852	0.000
$p = 5000$	EBIC_0	3.059 ± 0.423	54.155	100.000	94.383	0.000
	$\text{EBIC}_{0.25}$	3.058 ± 0.423	53.590	100.000	94.317	0.000
	$\text{EBIC}_{0.50}$	1.986 ± 1.180	25.915	99.333	42.769	46.500
	$\text{EBIC}_{0.75}$	1.100 ± 0.613	3.365	99.167	2.102	90.500
	EBIC_1	1.125 ± 0.714	2.975	98.167	0.750	92.500
	HBIC_1	1.100 ± 0.613	3.365	99.167	2.102	90.500
	$\text{HBIC}_{1.25}$	1.125 ± 0.714	2.975	98.167	0.750	92.500
	$\text{HBIC}_{0.50}$	1.156 ± 0.774	2.935	97.500	0.250	93.000
	$\text{HBIC}_{1.75}$	1.269 ± 1.023	2.835	94.500	0.000	90.500
	MBIC	3.058 ± 0.423	53.495	100.000	94.305	0.000

that, our extensive simulation studies, including those not reported here, indicates that both the EBIC_1 and the HBIC_γ with $1 < \gamma < 2$ are very competitive. Though the EBIC_1, in theory, may not be applicable to the situations with non-polynomial dimensionality, its performance, in practice, is still encouraging. To conclude, both the EBIC_γ with $0.75 \leq \gamma \leq 1$ and the HBIC_γ with $1 < \gamma < 1.50$ are recommendable to situations where p is much larger than n.

A real data example. We illustrate the performance of our methods on the internet advertising data which are available at the University of California-Irvine machine learning repository. This data set represents a set of possible advertisements on Internet pages. The predictors encode the geometry of the image as well as phrases occurring in the URL, the image's URL and alt text, the anchor text, and words occurring near the anchor text. The task is to predict whether an image is an advertisement (AD) or not (NONAD). After preprocessing the data set contained 2358 observations and 1430 variables. We randomly split the 2358 samples into training and test sets as follows. We set 50% of the observations from class AD and 50% of the observations from the class NONAD as training samples, and the rest as test samples. The SIS + AENET procedure is applied to the training data, with the tuning parameters selected by the EBIC_γ, and their performance is evaluated by the test samples. Four values of γ are explored, 0, 0.25, 0.50, and 0.75. The EBIC_1 and the HBIC_γ with $\gamma > 1$ were not examined

Table 3
Internet advertising data. Classification errors made and the number of variables chosen over 200 random splits of all samples into training and test sets.

Tuning method	Training error (%, mean ± sd)	Test error (%, mean ± sd)	No. of selected variables (mean ± sd)
$EBIC_0$	3.386 ± 0.689	5.061 ± 0.661	32.135 ± 10.574
$EBIC_0.25$	3.427 ± 0.687	5.024 ± 0.660	27.315 ± 9.246
$EBIC_0.50$	3.462 ± 0.696	5.022 ± 0.662	25.025 ± 8.960
$EBIC_0.75$	3.507 ± 0.711	5.028 ± 0.688	22.780 ± 6.694
MBIC	3.459 ± 0.697	5.021 ± 0.666	25.195 ± 9.004

because here the sample size $n = 1179$ and the predictor dimension $p = 1430$ are very comparable although $p > n$. For comparison purpose, we also examined the performance of the MBIC. To reduce the variability, the splitting into training and test sets was repeated 200 times and the classification errors were then averaged. The results are summarized in Table 3. As we can see, the classification errors are very similar among the tuning methods we considered. Further, the $EBIC_{0.75}$ picked up about 23 variables and then generated the smallest working model. It turns out that the SIS − AENET procedure works pretty well, possibly due to the relatively large size of sample.

Acknowledgments

The research described here was supported by a grant from the Research Council of Hong Kong, and a grant from Hong Kong Baptist University, Hong Kong.

Appendix

For ease of illustration, we assume that $\sigma = 1$ throughout though the extension to unknown σ being straightforward.

Lemma A.1. *Under conditions* (A)–(C), *if* $m\log(p) = o(n)$, *then we have,*

$$\lim_{n\to\infty} P(c_1 < \phi_n(m) < \Phi_n(m) < c_2) = 1.$$

Proof of Lemma A.1. See Proposition 2 in [19, Section 4]. □

Corollary A.1. *Let* $\boldsymbol{\mu} = \boldsymbol{X}_{M_0}\boldsymbol{\beta}_{M_0}$, *and let* $P_M = \boldsymbol{X}_M(\boldsymbol{X}_M^T\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^T$ *be the projection matrix of* $\boldsymbol{X}_M$. *Under conditions* (A)–(C), *if* $m\log(p) = o(n)$, *then,*

$$\lim_{n\to\infty} P(n^{-1}\min_{M\not\supseteq M_0, |M|\le m} \|\boldsymbol{\mu} - P_M\boldsymbol{\mu}\|_2^2 > 0) = 1.$$

Proof of Corollary A.1. For any sub-model $M \subseteq F = \{1, \ldots, p\}$, we have

$$\begin{aligned}
\|\boldsymbol{\mu} - P_M\boldsymbol{\mu}\|_2^2 &= \|\boldsymbol{X}_{M_0\cap M^c}\boldsymbol{\beta}_{M_0\cap M^c} - P_M\boldsymbol{X}_{M_0\cap M^c}\boldsymbol{\beta}_{M_0\cap M^c}\|_2^2 \\
&= \inf_{\boldsymbol{\gamma}} \|\boldsymbol{X}_{M_0\cap M^c}\boldsymbol{\beta}_{M_0\cap M^c} - \boldsymbol{X}_M\boldsymbol{\gamma}\|_2^2 \\
&= \inf_{\boldsymbol{\gamma}_{M_0\cap M^c} = -\boldsymbol{\beta}_{M_0\cap M^c}} \boldsymbol{\gamma}^T\boldsymbol{X}_{M_0\cup M}^T\boldsymbol{X}_{M_0\cup M}\boldsymbol{\gamma} \\
&\ge n\phi_n(|M\cup M_0|)\|\boldsymbol{\beta}_{M_0\cap M^c}\|_2^2,
\end{aligned}$$

where in the third equality we use the fact that the set $M_0\cap M^c$ is the relative complement of the set M in $M_0\cup M$. If $|M| \le m$, then $|M\cup M_0|\log(p) = o(n)$, and Lemma A.1 implies that $\phi_n(|M\cup M_0|) > c_1$ with probability approaching 1. On the other hand, $\|\boldsymbol{\beta}_{M_0\cap M^c}\|_2 > 0$ for any $M \not\supseteq M_0$. We then arrive at

$$n^{-1}\min_{M\not\supseteq M_0, |M|\le m} \|\boldsymbol{\mu} - P_M\boldsymbol{\mu}\|_2^2 \ge c_1 \min_{M\not\supseteq M_0, |M|\le m} \|\boldsymbol{\beta}_{M_0\cap M^c}\|_2^2 > 0,$$

with probability approaching 1. The proof is complete. □

Proof of Proposition 2.1. We first deal with the case $M \not\supseteq M_0$. Under conditions (A)–(C), we obtain from Lemma A.1 that, with probability approaching 1, the rank of P_M is $|M|$ for any sub-model M such that $|M|\log(p) = o(n)$. It follows that

$$\mathrm{RSS}_{M_0} = \boldsymbol{Y}^T(I_n - P_{M_0})\boldsymbol{Y} = \boldsymbol{\epsilon}^T(I_n - P_{M_0})\boldsymbol{\epsilon} = \sum_{j=1}^{n-|M_0|} Z_j^2, \tag{A.1}$$

where Z_j's are i.i.d. according to $N(0, 1)$. By Markov's inequality,

$$\mathrm{RSS}_{M_0} = n\{1 + o_p(1)\}. \tag{A.2}$$

On the other hand, one can show that

$$\mathrm{RSS}_M - \mathrm{RSS}_{M_0} = \boldsymbol{\mu}^T(I_n - P_M)\boldsymbol{\mu} + 2\boldsymbol{\mu}^T(I_n - P_M)\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^T P_M \boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T P_{M_0}\boldsymbol{\epsilon}.$$

Since $K\log(p) = o(n)$, from Corollary A.1, there is some constant $\delta_1 > 0$ such that

$$\min_{M \not\supseteq M_0, |M| \le K} \boldsymbol{\mu}^T(I_n - P_M)\boldsymbol{\mu} \ge \delta_1 n, \tag{A.3}$$

with probability approaching 1. Write

$$\boldsymbol{\mu}^T(I_n - P_M)\boldsymbol{\epsilon} = \|(I_n - P_M)\boldsymbol{\mu}\|_2 Z_M, \tag{A.4}$$

where

$$Z_M = \frac{\boldsymbol{\mu}^T(I_n - P_M)\boldsymbol{\epsilon}}{\|(I_n - P_M)\boldsymbol{\mu}\|_2} \sim N(0, 1).$$

By Bonferroni's inequality, we can arrive at

$$\max_{M \not\supseteq M_0, |M| \le K} |Z_M| = O_p\{\sqrt{K\log(p)}\} = o_p(\sqrt{n}). \tag{A.5}$$

This, together with (A.3) and (A.4) imply that

$$|\boldsymbol{\mu}^T(I_n - P_M)\boldsymbol{\epsilon}| = o_p\{\boldsymbol{\mu}^T(I_n - P_M)\boldsymbol{\mu}\}. \tag{A.6}$$

Since the term $\boldsymbol{\epsilon}^T P_M \boldsymbol{\epsilon}$ is a centralized χ^2-variate with $|M|$ degrees of freedom, by Bonferroni's inequality and tail bounds for χ^2-variables, we get for any $B > 1$

$$\begin{aligned} P\Big(\max_{M \not\supseteq M_0, |M| \le K} \boldsymbol{\epsilon}^T P_M \boldsymbol{\epsilon} \ge 5BK\log(p)\Big) &\le \sum_{M \not\supseteq M_0, |M| \le K} P(\boldsymbol{\epsilon}^T P_M \boldsymbol{\epsilon} \ge 5BK\log(p)) \\ &\le \sum_{M \not\supseteq M_0, |M| \le K} p^{-BK} \le \binom{p}{K} p^{-BK} \le \left(\frac{e}{K}\right)^K p^{-(B-1)K}. \end{aligned}$$

Then, we have

$$\max_{M \not\supseteq M_0, |M| \le K} \boldsymbol{\epsilon}^T P_M \boldsymbol{\epsilon} = O_p\{K\log(p)\} = o_p(n). \tag{A.7}$$

Similarly,

$$\boldsymbol{\epsilon}^T P_{M_0}\boldsymbol{\epsilon} = o_p(n). \tag{A.8}$$

Therefore, by combining (A.3) and (A.6)–(A.8), we obtain that

$$\min_{M \not\supseteq M_0, |M| \le K} \mathrm{RSS}_M - \mathrm{RSS}_{M_0} \ge \delta_1 n, \tag{A.9}$$

with probability approaching 1. Note that

$$\mathrm{HBIC}_\gamma(M) - \mathrm{HBIC}_\gamma(M_0) = n\log\left(1 + \frac{\mathrm{RSS}_M - \mathrm{RSS}_{M_0}}{\mathrm{RSS}_{M_0}}\right) + \log(p^{2\gamma})(|M| - |M_0|).$$

This, together with (A.2), (A.9) and condition (C), yield that

$$\min_{M \not\supseteq M_0, |M| \le K} \mathrm{HBIC}_\gamma(M) - \mathrm{HBIC}_\gamma(M_0) \ge n\log\left(1 + \frac{\delta_1}{2}\right) - \log(p^{2\gamma})|M_0| > 0, \tag{A.10}$$

with probability tending to 1.

Now consider the case $M \supseteq M_0$ but $M \ne M_0$. By Bonferroni's inequality,

$$\begin{aligned} &P\Big(\min_{M \supseteq M_0, M \ne M_0, |M| \le K} \{\mathrm{RSS}_M + \log(p^{2s})|M|\} \le \mathrm{RSS}_{M_0} + \log(p^{2s})|M_0|\Big) \\ &\le \sum_{M \supseteq M_0, M \ne M_0, |M| \le K} P(\mathrm{RSS}_{M_0} - \mathrm{RSS}_M \ge \log(p^{2s})(|M| - |M_0|)) \\ &= \sum_{M \supseteq M_0, M \ne M_0, |M| \le K} P(\boldsymbol{\epsilon}^T(P_M - P_{M_0})\boldsymbol{\epsilon} \ge \log(p^{2s})(|M| - |M_0|)) \\ &\le \sum_{M \supseteq M_0, M \ne M_0, |M| \le K} [\sqrt{\log(p^{2s})\exp\{1 - \log(p^{2s})\}}]^{|M| - |M_0|} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{|M|=|M_0|+1}^{K} \binom{p-|M_0|}{|M|-|M_0|} [\sqrt{\log(p^{2s})\exp\{1-\log(p^{2s})\}}]^{|M|-|M_0|} \\
&\le \sum_{m=0}^{p-|M_0|} \binom{p-|M_0|}{m} [\sqrt{\log(p^{2s})\exp\{1-\log(p^{2s})\}}]^{m} - 1, \\
&= \left\{1+\sqrt{\frac{e\log(p^{2s})}{p^{2s}}}\right\}^{p-|M_0|} - 1 \to 0, \quad \text{as } n\to\infty.
\end{aligned}
$$

Here $s > 1$ and the second inequality follows from the sharp deviation bound on the chi-square distribution, see [8, Lemma 3]. Thus, for any $s > 1$, with probability approaching 1, we have

$$
\max_{M\supseteq M_0, M\neq M_0, |M|\le K} \frac{\mathrm{RSS}_{M_0} - \mathrm{RSS}_M}{|M|-|M_0|} < \log(p^{2s}). \tag{A.11}
$$

Then, by condition $K\log(p) = o(n)$,

$$
\max_{M\supseteq M_0, M\neq M_0, |M|\le K} (\mathrm{RSS}_{M_0} - \mathrm{RSS}_M) = O_p\{\log(np^2)K\} = o_p(n). \tag{A.12}
$$

Recall that $\mathrm{RSS}_{M_0} = n\{1+o_p(1)\}$. We then have

$$
\max_{M\supseteq M_0, M\neq M_0, |M|\le K} \frac{\mathrm{RSS}_{M_0} - \mathrm{RSS}_M}{\mathrm{RSS}_{M_0}} = o_p(1). \tag{A.13}
$$

By the Taylor's theorem, we get, for n sufficiently large and $\delta_2 > 0$ sufficiently small,

$$
\begin{aligned}
\mathrm{HBIC}_\gamma(M) - \mathrm{HBIC}_\gamma(M_0) - \delta_2\log(p^{2\gamma}) &= n\log\left(1+\frac{\mathrm{RSS}_M - \mathrm{RSS}_{M_0}}{\mathrm{RSS}_{M_0}}\right) + (1-\delta_2)\log(p^{2\gamma})(|M|-|M_0|) \\
&= (\mathrm{RSS}_M - \mathrm{RSS}_{M_0})(1+R_n) + (1-\delta_2)\log(p^{2\gamma})(|M|-|M_0|) \\
&\ge (1+R_n)\left\{\mathrm{RSS}_M - \mathrm{RSS}_{M_0} + \frac{1-\delta_2}{1+|R_n|}\log(p^{2\gamma})(|M|-|M_0|)\right\} \\
&\ge (1+R_n)\left\{\mathrm{RSS}_M - \mathrm{RSS}_{M_0} + \log(p^{2s})(|M|-|M_0|)\right\}.
\end{aligned}
$$

Here $R_n = o_p(1)$ is independent of M due to (A.13), $1 < s < \gamma$, and in the last inequality we invoke the assumption that $\gamma > 1$. This, together with (A.11) imply

$$
\min_{M\supseteq M_0, M\neq M_0, |M|\le K} \mathrm{HBIC}_\gamma(M) - \mathrm{HBIC}_\gamma(M_0) > \delta_2\log(p^{2\gamma}), \tag{A.14}
$$

with probability approaching 1. Therefore, it follows from (A.10) and (A.12) that

$$
\lim_{n\to\infty} P\Big(\min_{M\neq M_0, |M|\le K} \mathrm{HBIC}_\gamma(M) > \mathrm{HBIC}_\gamma(M_0)\Big) = 1. \tag{A.15}
$$

This completes the proof. □

Proof of Theorem 2.1. Under theorem conditions, by invoking Theorem 5.1 in [22, page 1743], we then obtain that the SIS + AENET procedure has the oracle property. Define $\Omega_- = \{\lambda > 0 : M_\lambda \not\supseteq M_0\}$, $\Omega_0 = \{\lambda > 0 : M_\lambda = M_0\}$ and $\Omega_+ = \{\lambda > 0 : M_\lambda \supseteq M_0, M_\lambda \neq M_0\}$, according to whether the model M_λ is under-fitted, correctly fitted or over-fitted. Let $\{\lambda_n\}$ denote a reference tuning parameter sequence such that $\hat{\boldsymbol{\beta}}(\lambda_n)$ has the oracle property. Then, we have

$$
\lim_{n\to\infty} P(M_{\lambda_n} = M_0) = 1, \tag{A.16}
$$

$$
\|\hat{\boldsymbol{\beta}}(\lambda_n) - \boldsymbol{\beta}_{M_0}\|_2^2 = O_p\left(\frac{|M_0|}{n}\right). \tag{A.17}
$$

It suffices to prove that

$$
\lim_{n\to\infty} P\Big(\inf_{\lambda\in\Omega_-\cup\Omega_+, |M_\lambda|\le d} \mathrm{HBIC}_\gamma(M_\lambda) > \mathrm{HBIC}_\gamma(M_{\lambda_n})\Big) = 1. \tag{A.18}
$$

Since $d\log(p) = o(n)$, it follows from the proof of Proposition 2.1 that, with probability approaching 1, there are some positive constants δ_1, δ_2 such that

$$
\inf_{\lambda\in\Omega_-, |M_\lambda|\le d} \mathrm{HBIC}_\gamma(M_\lambda) - \mathrm{HBIC}_\gamma(M_0) \ge \min_{M\not\supseteq M_0, |M|\le d} \mathrm{HBIC}_\gamma(M) - \mathrm{HBIC}_\gamma(M_0) \ge \delta_1 n,
$$

$$\inf_{\lambda\in\Omega_+,|M_\lambda|\le d}\mathrm{HBIC}_\gamma(M_\lambda)-\mathrm{HBIC}_\gamma(M_0)\ge\min_{M\supseteq M_0,M\ne M_0,|M|\le d}\mathrm{HBIC}_\gamma(M)-\mathrm{HBIC}_\gamma(M_0)\ge\delta_2\log(p^{2\gamma}).$$

It follows that

$$\lim_{n\to\infty}P\Big(\inf_{\lambda\in\Omega_-\cup\Omega_+,|M_\lambda|\le d}\mathrm{HBIC}_\gamma(M_\lambda)\ge\mathrm{HBIC}_\gamma(M_0)+\delta_2\log(p^{2\gamma})\Big)=1. \tag{A.19}$$

Then, it remains to show that

$$\mathrm{HBIC}_\gamma(M_{\lambda_n})-\mathrm{HBIC}_\gamma(M_0)=o_p\{\log(p^{2\gamma})\}. \tag{A.20}$$

By (A.16) and Corollary A.1, we have

$$\begin{aligned}\mathrm{RSS}_{\lambda_n}-\mathrm{RSS}_{M_0}&=\|\boldsymbol{Y}-\boldsymbol{X}_{M_{\lambda_n}}\hat{\boldsymbol{\beta}}(\lambda_n)\|_2^2-\|\boldsymbol{Y}-\boldsymbol{X}_{M_0}\hat{\boldsymbol{\beta}}_{M_0}\|_2^2\\&=\|\boldsymbol{Y}-\boldsymbol{X}_{M_{\lambda_n}}\hat{\boldsymbol{\beta}}(\lambda_n)\|_2^2-\|\boldsymbol{Y}-\boldsymbol{X}_{M_{\lambda_n}}\hat{\boldsymbol{\beta}}_{M_{\lambda_n}}\|_2^2\\&=\{\hat{\boldsymbol{\beta}}(\lambda_n)-\hat{\boldsymbol{\beta}}_{M_{\lambda_n}}\}^T\boldsymbol{X}^T\boldsymbol{X}\{\hat{\boldsymbol{\beta}}(\lambda_n)-\hat{\boldsymbol{\beta}}_{M_{\lambda_n}}\}\\&\le n\Phi_n(2d)\|\hat{\boldsymbol{\beta}}(\lambda_n)-\hat{\boldsymbol{\beta}}_{M_{\lambda_n}}\|_2^2\\&\le 2c_2n\{\|\hat{\boldsymbol{\beta}}(\lambda_n)-\boldsymbol{\beta}_{M_0}\|_2^2+\|\boldsymbol{\beta}_{M_0}-\hat{\boldsymbol{\beta}}_{M_0}\|_2^2\},\end{aligned}$$

with probability tending to 1. Since

$$\|\hat{\boldsymbol{\beta}}_{M_0}-\boldsymbol{\beta}_{M_0}\|_2^2=O_p\left(\frac{|M_0|}{n}\right). \tag{A.21}$$

This, together with (A.17) imply that

$$\mathrm{RSS}_{\lambda_n}-\mathrm{RSS}_{M_0}=O_p(|M_0|)=o_p\{\log(p^{2\gamma})\}. \tag{A.22}$$

Hence, we arrive at (A.18). Consequently,

$$\begin{aligned}&\lim_{n\to\infty}P\Big(\inf_{\lambda\in\Omega_-\cup\Omega_+,|M_\lambda|\le d}\mathrm{HBIC}_\gamma(M_\lambda)>\mathrm{HBIC}_\gamma(M_{\lambda_n})\Big)\\&\quad=\lim_{n\to\infty}P\Big(\inf_{\lambda\in\Omega_-\cup\Omega_+,|M_\lambda|\le d}\mathrm{HBIC}_\gamma(M_\lambda)>\mathrm{HBIC}_\gamma(M_0)+o_p\{\log(p^{2\gamma})\}\Big)=1.\end{aligned}$$

The proof is complete. □

References

[1] K.W. Broman, T.P. Speed, A model selection approach for the identification of quantitative trait loci in experimental crosses, J. Roy. Statist. Soc. Ser. B 64 (2002) 641–656.
[2] J. Chen, Z. Chen, Extended Bayesian information criteria for model selection with large model spaces, Biometrika 95 (2008) 759–771.
[3] D. Donoho, For most large underdetermined systems of linear equations, the minimal l^1-norm near-solution approximates the sparsest near-solution, Comm. Pure Appl. Math. 59 (2006) 797–829.
[4] D. Donoho, I. Johnstone, Ideal spatial adaptation by wavelet shrinkage, Biometrika 81 (1994) 425–455.
[5] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Statist. 32 (2004) 407–499.
[6] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001) 1348–1360.
[7] J. Fan, R. Li, Statistical challenges with high-dimensionality: feature selection in knowledge discovery, in: M. Sanz-Sole, J. Soria, J.L. Varona, J. Verdera (Eds.), Proc. Int. Congr. Mathematicians, vol. III, European Mathimatical Society, Zurich, 2006, pp. 595–622.
[8] J. Fan, J. Lv, Sure independence screening for ultra-high-dimensional feature space, J. Roy. Statist. Soc. Ser. B 70 (2008) 849–911.
[9] J. Fan, J. Lv, A selective overview of variable selection in high dimensional feature space, Statist. Sinica 20 (2010) 101–148.
[10] J. Fan, H. Peng, Non-concave penalized likelihood with a diverging number of parameters, Ann. Statist. 32 (2004) 928–961.
[11] D. Foster, E. George, The risk inflation criterion for multiple regression, Ann. Statist. 22 (1994) 1947–1975.
[12] N. Meinshausen, B. Yu, LASSO-type recovery of sparse representations for high dimensinal data, Ann. Statist. 37 (2009) 246–270.
[13] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.
[14] D. Siegmund, Model selection in irregular problems: application to mapping quantitative loci, Biometrika 91 (2004) 785–800.
[15] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Roy. Statist. Soc. Ser. B 58 (1996) 267–288.
[16] H. Wang, Forward regression for ultra-high dimensional variable screening, J. Amer. Statist. Assoc. 104 (2009) 1512–1524.
[17] H. Wang, B. Li, C. Leng, Shrinkage tuning parameter selection with a diverging number of parameters, J. Roy. Statist. Soc. Ser. B 71 (2009) 671–683.
[18] H. Wang, R. Li, C.L. Tsai, Tuning parameter selectors for the smoothly clipped absolute deviation method, Biometrika 94 (2007) 553–568.
[19] C.H. Zhang, J. Huang, The sparsity and bias of the LASSO selection in high-dimensional linear regression, Ann. Statist. 36 (2008) 1567–1594.
[20] H. Zou, The adaptive LASSO and its oracle properties, J. Amer. Statist. Assoc. 101 (2006) 1418–1429.
[21] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. Roy. Statist. Soc. Ser. B 67 (2005) 301–320.
[22] H. Zou, H.H. Zhang, On the adaptive elastic-net with a diverging number of parameters, Ann. Statist. 37 (2009) 1733–1751.