Journal of Machine Learning Research 13 (2012) 1037-1057 Submitted 6/11; Revised 1/12; Published 4/12

Consistent Model Selection Criteria on High Dimensions

Yongdai Kim YDKIM@STATS.SNU.AC.KR
Department of Statistics
Seoul National University
Seoul 151-742, Korea

Sunghoon Kwon SHKWON0522@GMAIL.COM
School of Statistics
University of Minnesota
Minneapolis, MN 55455, USA

Hosik Choi CHOI.HOSIK@GMAIL.COM
Department of Informational Statistics
Hoseo University
Chungnam 336-795, Korea

Editor: Xiaotong Shen

Abstract

Asymptotic properties of model selection criteria for high-dimensional regression models are studied where the dimension of covariates is much larger than the sample size. Several sufficient conditions for model selection consistency are provided. Non-Gaussian error distributions are considered and it is shown that the maximal number of covariates for model selection consistency depends on the tail behavior of the error distribution. Also, sufficient conditions for model selection consistency are given when the variance of the noise is neither known nor estimated consistently. Results of simulation studies as well as real data analysis are given to illustrate that finite sample performances of consistent model selection criteria can be quite different.

Keywords: model selection consistency, general information criteria, high dimension, regression

1. Introduction

Model selection is a fundamental task for high-dimensional statistical modeling where the number of covariates can be much larger than the sample size. In such cases, classical model selection criteria such as the Akaike information criterion or AIC (Akaike, 1973), the Bayesian information criterion or BIC (Schwarz, 1978) and cross validations or generalized cross validation (Craven and Wahba, 1979; Stone, 1974) tend to select more variables than necessary. See, for example, Broman and Speed (2002) and Casella et al. (2009). Also, Yang and Barron (1998) discussed severe selection bias of AIC which damages predictive performance for high-dimensional models.

Recently, various model selection criteria for high-dimensional models have been introduced. Wang et al. (2009) proposed a modified BIC which is consistent when the dimension of covariates is diverging slower than the sample size. Here, the consistency of a model selection criterion means that the probability of the selected model being equal to the true model converges to 1. See Section 2 for a rigorous definition. The extended BIC by Chen and Chen (2008) and corrected RIC by Zhang and Shen (2010) are shown to be consistent even when the dimension of covariates is larger than the

sample size. Some sparse penalized approaches including the LASSO (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996) and SCAD (Smoothly Clipped Absolute Deviation) (Fan and Li, 2001) are proven to be consistent for high-dimensional models. See Zhao and Yu (2006) for the LASSO and Kim et al. (2008) for the SCAD.

In this paper, we study asymptotic properties of a large class of model selection criteria based on the generalized information criterion (GIC) considered by Shao (1997). The class of GICs is large enough to include many well known model selection criteria such as the AIC, BIC, modified BIC by Wang et al. (2009), risk inflation criterion (RIC) by Foster and George (1994), modified risk inflation criterion (MRIC) by Foster and George (1994), corrected RIC by Zhang and Shen (2010). Also, as we will show, the extended BIC by Chen and Chen (2008) is asymptotically equivalent to a GIC.

We give sufficient conditions for a given GIC to be consistent. Our sufficient conditions are general enough to include cases where the error distribution can be other than Gaussian and the variance of the error distribution is not consistently estimated. For a case of the Gaussian error distribution with consistent estimator of the variance, our sufficient conditions include most of the previously proposed consistent model selection criteria such as the modified BIC (Wang et al., 2009), extended BIC (Chen and Chen, 2008) and corrected RIC (Zhang and Shen, 2010).

For high-dimensional models, it is not practically feasible to find the best model among all possible submodels since the number of submodels are too large. A simple remedy is to find a sequence of submodels with increasing complexities (e.g., increasing number of covariates) and find the best model among them using a given model selection criterion. Examples of constructing a sequence of submodels are the forward selection procedure and solution paths of penalized regression approaches. Our sufficient conditions are still valid as long as the sequence of submodels includes the true model with probability converging to 1. We discuss more on these issues in Section 4.1.

The paper is organized as follows. In Section 2, the GIC is introduced. In Section 3, sufficient conditions for the consistency of GICs are given. Various remarks about application of GICs to real data analysis are given in Section 4. In Section 5, results of simulations as well as a real data analysis are presented, and concluding remarks follow in Section 6.

2. Generalized Information Criterion

Let $\mathcal{L} = \{(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)\}$ be a given data set of independent pairs of response and covariates, where $y_i \in R$ and $\mathbf{x}_i \in R^{p_n}$. Suppose the true regression model for $(y, \mathbf{x})$ is given as

$$y = \mathbf{x}'\beta^* + \varepsilon,$$

where $\beta^* \in R^{p_n}, \mathrm{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$. For simplicity, we assume that σ^2 is known. For unknown σ^2, see Section 4.2.

Let $Y_n = (y_1, \ldots, y_n)'$ and $\mathbf{X}_n$ be the $n \times p_n$ dimensional design matrix whose jth column is $X_n^j = (x_{1j}, \ldots, x_{nj})'$. For given $\beta \in R^{p_n}$, let

$$R_n(\beta) = \|Y_n - \mathbf{X}_n\beta\|^2,$$

where $\|\cdot\|$ is the Euclidean norm. For a given subset $\pi \subset \{1, \ldots, p_n\}$, let

$$\hat{\beta}_\pi = \mathrm{argmin}_{\beta:\beta_j=0,\, j\in\pi^c} R_n(\beta).$$

For a given sequence of positive numbers $\{\lambda_n\}$, the GIC indexed by $\{\lambda_n\}$, denoted by GIC_{λ_n}, gives a sequence of random subsets $\hat{\pi}_{\lambda_n}$ of $\{1,\ldots,p_n\}$ defined as

$$\hat{\pi}_{\lambda_n} = \mathrm{argmin}_{\pi\subset\{1,\ldots,p_n\}} R_n(\hat{\beta}_\pi) + \lambda_n|\pi|\sigma^2,$$

where $|\pi|$ is the cardinality of π. The AIC corresponds to $\lambda_n = 2$, the BIC to $\lambda_n = \log n$, the RIC of Foster and George (1994) to $\lambda_n = 2\log p_n$, the RIC of Zhang and Shen (2010) to $\lambda_n = 2(\log p_n + \log\log p_n)$. Shao (1997) studied the asymptotic properties of the GIC focusing on the AIC and BIC.

When p_n is large, it would not be wise to search all possible subsets of $\{1,\ldots,p_n\}$. Instead, we set an upper bound on the cardinality of π, say s_n and search the optimal model among submodels whose cardinalities are smaller than s_n. Chen and Chen (2008) considered a similar model selection procedure. Let $\mathcal{M}^{s_n} = \{\pi \subset \{1,\ldots,p_n\} : |\pi| \le s_n\}$. We define the restricted GIC_{λ_n} as

$$\hat{\pi}_{\lambda_n} = \mathrm{argmin}_{\pi\in\mathcal{M}^{s_n}} R_n(\hat{\beta}_\pi) + \lambda_n|\pi|\sigma^2. \tag{1}$$

The restricted GIC is the same as the GIC if $s_n = p_n$. In the following, we will only consider the restricted GIC and suppress the term "restricted" unless there is any confusion.

3. Consistency of GIC on High Dimensions

Let $\pi_n^* = \{j : |\beta_j^*| \neq 0\}$. We say that the GIC_{λ_n} is consistent if

$$\Pr(\hat{\pi}_{\lambda_n} = \pi_n^*) \to 1$$

as $n\to\infty$. In this section, we prove the consistency of the GIC_{λ_n} under regularity conditions.

For a given subset π of $\{1,\ldots,p_n\}$, let $\mathbf{X}_\pi = (X_n^j, j\in\pi)$ be the $n\times|\pi|$ matrix whose columns consist of $X_n^j, j\in\pi$. For a given symmetric matrix $\mathbf{A}$, let $\xi(\mathbf{A})$ be the smallest eigenvalue of $\mathbf{A}$.

3.1 Regularity Conditions

We assume the following regularity conditions.

- A1 : There exists a positive constant M_1 such that $X_n^{j'}X_n^j/n \le M_1$ for all $j = 1,\ldots,p_n$ and all n.
- A2 : There is a positive constant M_2 such that $\xi(\mathbf{X}'_{\pi_n^*}\mathbf{X}_{\pi_n^*}/n) \ge M_2$ for all n.
- A3 : There exist positive constants c_1 and M_3 such that $0 \le c_1 < 1/2$ and $\rho_n \ge M_3 n^{-c_1}$, where

$$\rho_n = \inf_{\pi:|\pi|\le s_n} \xi(\mathbf{X}'_\pi\mathbf{X}_\pi/n).$$

- A4 : There exist positive constants c_2 and M_4 such that $2c_1 < c_2 \le 1$ and

$$n^{(1-c_2)/2}\min_{j\in\pi_n^*}|\beta_j^*| \ge M_4.$$

- A5 : $q_n = O(n^{c_3})$ for some $0 \le c_3 < c_2$, and $q_n \le s_n$, where $q_n = |\pi_n^*|$.

Condition A1 assumes that the covariates are bounded. Condition A2 means that the design matrix of the true model is well posed. Condition A3 is called the sparse Riesz condition and used in Chen and Chen (2008), Zhang (2010) and Kim and Kwon (2012). Condition A4 and A5 allow the nonzero regression coefficients to converge to 0 and the number of signal variables to diverge, respectively.

Remark 1 *Condition A3 implies that* $s_n \leq n$.

3.2 The Main Theorem

The following theorem proves consistency of the GIC_{λ_n}. The proofs are deferred to Appendix.

Theorem 2 *Suppose* $\text{E}(\varepsilon^{2k}) < \infty$ *for some integer* $k > 0$. *If* $\lambda_n = o(n^{c_2-c_1})$ *and* $p_n/(\lambda_n \rho_n)^k \to 0$, *then the* GIC_{λ_n} *is consistent.*

In Theorem 2, p_n can diverge only polynomially fast in n since $p_n = o(\lambda_n^k) = o(n^{kc_2})$. Since k can be considered as a degree of tail lightness of the error distribution, we can conclude that the lighter the tail of the error distribution is, the more covariates the GIC_{λ_n} is consistent with. When ε is Gaussian, the following theorem proves that the GIC_{λ_n} can be consistent when p_n diverges exponentially fast.

Theorem 3 *Suppose* $\varepsilon \sim N(0, \sigma^2)$. *If* $\lambda_n = o(n^{c_2-c_1}), s_n \log p_n = o(n^{c_2-c_1})$ *and* $\lambda_n - 2\log p_n - \log\log p_n \to \infty$, *then the* GIC_{λ_n} *is consistent.*

In the following, we give three examples for (i) fixed p_n, (ii) polynomially diverging p_n and (iii) exponentially diverging p_n. For simplicity, we let $c_1 = 0$ (i.e., $\rho_n \geq M_3 > 0$), $c_2 = 1$ (i.e., $\min_{j \in \pi_n^*} |\beta_j^*| > 0$) and $c_3 = 0$ (i.e., q_n is fixed). In addition, we let s_n be fixed.

Example 1 *Consider a standard case where* p_n *is fixed and n goes to infinity. Theorem 2 implies that the* GIC_{λ_n} *is consistent if* $\lambda_n/n \to 0$ *and* $\lambda_n \to \infty$ *regardless of the tail lightness (i.e., k) of the error distribution, provided the variance exists. The BIC, which is the GIC with* $\lambda_n = \log n$, *satisfies these conditions and hence is consistent. Note that the AIC does not satisfy the conditions in Theorem 2. Any GIC with* $\lambda_n = n^c, 0 < c < 1$ *is consistent, which suggests that the class of consistent model selection criteria is quite large. See Shao (1997) for more discussions.*

Example 2 *Consider a case of* $p_n = n^\gamma, \gamma > 0$. *The GIC with* $\lambda_n = n^\xi, 0 < \xi < 1$ *and* $\gamma < k\xi$ *is consistent. That is, for larger* p_n, *we need larger* λ_n *for consistency, which is reasonable because we need to be more careful not to overfit when* p_n *is large. When the error distribution is Gaussian, Theorem 3 can be compared with other previous results of consistency. First, the BIC (i.e., the GIC with* $\lambda_n = \log n$*) is consistent when* $\gamma < 1/2$. *For* $0 < \gamma < 1$, *Theorem 3 implies that the modified BIC of Wang et al. (2009), which is a GIC with* $\lambda_n = \log\log p_n \log n$, *is consistent. Chen and Chen (2008) proposed a model selection criterion called the extended BIC given by*

$$\hat{\pi}^{eBIC} = \text{argmin}_{\pi \subset \{1,\ldots,p_n\}, |\pi| \leq K} R_n(\hat{\beta}_\pi) + |\pi|\sigma^2 \log n + 2\kappa\sigma^2 \log \binom{p_n}{|\pi|}$$

for some $K > 0$ *and* $0 \leq \kappa \leq 1$, *and proved that the extended BIC is consistent when* $\kappa > 1 - 1/(2\gamma)$. *Since* $\log \binom{p_n}{|\pi|} \asymp |\pi| \log p_n$ *for* $|\pi| \leq K$, *we have*

$$|\pi|\sigma^2 \log n + 2\gamma\sigma^2 \log \binom{p_n}{|\pi|} \asymp (\log n + 2\kappa \log p_n)|\pi|\sigma^2.$$

Hence, Theorem 3 confirms the result of Chen and Chen (2008).

Example 3 *When the error distribution is Gaussian, the GIC can be consistent for exponentially increasing p_n (i.e., ultra-high dimensional cases). The GIC with $\lambda_n = n^{\xi}, 0 < \xi < 1$ is consistent when $p_n = O(\exp(\alpha n^{\gamma}))$ for $0 < \gamma < \xi$ and $\alpha > 0$. Also, it can be shown by Theorem 3 that the extended BIC with $\gamma = 1$ is consistent with $p_n = O(\exp(\alpha n^{\gamma}))$ for $0 < \gamma < 1/2$. The consistency of the corrected RIC of Zhang and Shen (2010) can be confirmed by Theorem 3, but the regularity conditions for Theorem 3 are more general than those of Zhang and Shen (2010).*

4. Remarks

Remarks regarding to applications of the GIC to real data analysis are given.

4.1 Construction of Sub-Models

For high-dimensional models, it is computationally infeasible to search the optimal model among all possible submodels. A simple remedy is to construct a sequence of submodels and select the optimal model among the sequence of submodels. Examples of constructing a sequence of submodels are the forward selection (Wang, 2009) and the solution path of a sparse penalized estimator obtained by, for example, the Lars algorithm (Efron et al., 2004) or the PLUS algorithm (Zhang, 2010). The following algorithm exemplifies the model selection procedure with the GIC and a sparse penalized regression approach.

- For a given sparse penalty $J_\eta(t)$ indexed by $\eta \geq 0$, find the solution path of a penalized estimator $\{\hat{\beta}(\eta) : \eta > 0\}$, where

$$\hat{\beta}(\eta) = \text{argmin}_\beta \left(R_n(\beta) + \sum_{j=1}^{p} J_\eta(|\beta_j|) \right).$$

  The LASSO corresponds to $J_\eta(t) = \eta t$ and the SCAD penalty corresponds to

$$\begin{aligned} J_\eta(t) &= \eta t I(0 \leq 0 < \eta) \\ &+ \left\{ \frac{a\eta(t-\eta) - (t^2-\eta^2)/2}{a-1} + \eta^2 \right\} I(\eta \leq t < a\eta) \\ &+ \left\{ \frac{(a-1)\eta^2}{2} + \eta^2 \right\} I(t \geq a\eta) \end{aligned}$$

  for some $a > 2$.

- Let $S(\eta) = \{ j : \hat{\beta}(\eta)_j \neq 0 \}$ and $\Upsilon = \{\eta : S(\eta) \neq S(\eta-), |S(\eta)| \leq s_n\}$.

- Apply the GIC_{λ_n} to $S(\eta), \eta \in \Upsilon$ to select the optimal model. That is, let $\hat{\pi}_{\lambda_n} = S(\eta^*)$ where

$$\eta^* = \text{argmin}_{\eta \in \Upsilon} \left(R_n(\hat{\beta}_\eta) + \lambda_n |S(\eta)| \right)$$

  and

$$\hat{\beta}_\eta = \text{argmin}_{\beta : \beta_j = 0, j \in S(\eta)^c} R_n(\beta).$$

It is easy to see that a consistent GIC is still consistent with a sequence of sub-models as long as the sequence of submodels includes the true model with probability converging to 1. For the LASSO solution path, Zhao and Yu (2006) proved the selection consistency under the irrepresentable condition, which is almost necessary (Zou, 2006). However, the irrepresentable condition is hardly satisfied for high-dimensional models. The consistency of the solution path of a nonconvex penalized estimator with either the SCAD penalty or minimax concave penalty is proved by Zhang (2010) and Kim and Kwon (2012). By combining Theorem 4 of Kim and Kwon (2012) and Theorem 2 of the current paper, we can prove the consistency of the GIC with the solution path of the SCAD penalty or minimax concave penalty, which is formally stated in the following theorem.

Theorem 4 *Condition A3 is replaced by A3', where*

- *A3': There exist positive constants c_1 and M_3 such that $0 \le c_1 < 1/2$ and $\rho_n \ge M_3 n^{c_1/2}$.*

Suppose $\mathrm{E}(\varepsilon^{2k}) < \infty$ for some integer $k > 0$. If $p_n = o(n^{k(c_2/2-c_1)})$, the under the regularity conditions A1 to A5 with A3 being replaced by A3', the solution path of the SCAD or minimax concave penalty included the true model with probability converging to 1, and hence the GIC_{λ_n} with $\lambda_n = o(n^{c_2-c_1})$ is consistent with the solution path of the SCAD or minimax concave penalty.

Remark 5 *Condition A3' is a technical modification needed for Theorem 4 of Kim and Kwon (2012). Note that A3 is weaker than A3', which is an advantage of using the l_0 penalty rather than nonconvex penalties which are linear around 0.*

Remark 6 *Theorem 3 can be modified similarly for the GIC with the solution path of the SCAD or minimax concave penalty, since Theorem 4 of Kim and Kwon (2012) can be modified accordingly for the Gaussian error distribution.*

4.2 Estimation of the Variance

To use the GIC in practice, we need to know σ^2. If σ^2 is unknown, we can replace it by its estimate. Theorems 2 and 3 are still valid as long as σ^2 is estimated consistently. When p_n is fixed, we can estimate σ^2 consistently by the mean squared error of the full model. For high-dimensional data, it is not obvious how to estimate σ^2. However, a weaker condition can be put on an estimator $\hat{\sigma}^2$ of σ^2 for the GIC to be consistent. Suppose that

$$0 < r_{inf} = \liminf \frac{\hat{\sigma}^2}{\sigma^2} \le \limsup \frac{\hat{\sigma}^2}{\sigma^2} = r_{sup} < \infty \tag{2}$$

with probability 1. This condition essentially assumes that $\hat{\sigma}^2$ is neither too small nor too large. It is not difficult to show that Theorem 2 is still valid with $\hat{\sigma}^2$ satisfying (2). This, however, is not true for Theorem 3. A slightly weak version of Theorem 3 which only requires (2) is given in the following theorem.

Theorem 7 *Suppose $\varepsilon \sim N(0,\sigma^2)$. Let $\hat{\sigma}^2$ be an estimator of σ^2 satisfying (2). If $\lambda_n = o(n^{c_2-c_1})$ and $\lambda_n - 2M_1 \log p_n/\rho_n r_{inf} \to \infty$, then the GIC_{λ_n} with the estimated variance is consistent.*

The corrected RIC, the GIC with $\lambda_n = 2(\log p_n + \log\log p_n)$, does not satisfy the condition in Theorem 7, and hence may not be consistent with an estimated variance. On the other hand, the GIC with $\lambda_n = \alpha_n \log p_n$ is consistent as long as $\alpha_n \to \infty$.

4.3 The Size of s_n

For condition A5, s_n should be large enough so that $q_n \leq s_n$. In many cases, s_n can be sufficiently large for practical purposes. For example, suppose $\{\mathbf{x}_i, i \leq n\}$ are independent and identically distributed p_n dimensional random vectors such that $\mathrm{E}(\mathbf{x}_1) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{x}_1) = \Sigma = [\sigma_{jk}]$. For a given $\rho > 0$, let s^* be the largest integer such that the smallest eigenvalue of $\Sigma_\eta = [\sigma_{jk}, j,k \in \eta]$ is greater than ρ for any $\eta \subset \{1,\ldots,p_n\}$ with $|\eta| \leq s^*$. For example, when Σ is compound symmetry, that is $\sigma_{jj} = 1$ and $\sigma_{jk} = \nu$ for $j \neq k$ and $\nu \in [0,1)$, the smallest eigenvalue of Σ_η is $1-\nu$ for all $\eta \subset \{1,\ldots,p_n\}$ and hence $s^* = p_n$ if $1-\nu > \rho$. Let $\mathbf{A} = \Sigma_\eta - \mathbf{X}_\eta^{'}\mathbf{X}_\eta/n$. By the inequality (2) in Greenshtein and Ritov (2004), we have

$$\sup_{j,k}\left|\sum_{i=1}^{n} x_{ij}x_{ik}/n - \sigma_{jk}\right| = O_p\left(\sqrt{\frac{\log n}{n}}\right),$$

and hence $\sup_{jk}|a_{jk}| = O_p(\sqrt{\log n/n})$, where a_{jk} is the (j,k) entry of $\mathbf{A}$. Since the largest eigenvalue of $\mathbf{A}$ is bounded by $|\eta|O_p(\sqrt{\log n/n})$, the smallest eigenvalue of $\mathbf{X}_\eta{}'\mathbf{X}_\eta/n$ is greater than $\rho - |\eta|O_p(\sqrt{\log n/n})$ if $|\eta| \leq s^*$. So, we can let $s_n = \min\{n^c, s^*\}$ for $c < 1/2$.

5. Numerical Analysis

In this section, we investigate finite sample performance of various GICs by simulation experiments as well as real data analysis. We consider the five GICs whose corresponding λ_ns are given as

- GIC_1(=BIC) : $\lambda_n^{(1)} = \log n$,
- GIC_2 : $\lambda_n^{(2)} = p_n^{1/3}$,
- GIC_3 : $\lambda_n^{(3)} = 2\log p_n$,
- GIC_4 : $\lambda_n^{(4)} = 2(\log p_n + \log\log p_n)$,
- GIC_5 : $\lambda_n^{(5)} = \log\log n \log p_n$,
- GIC_6 : $\lambda_n^{(6)} = \log n \log p_n$.

The GIC_1 is the BIC. By Theorem 2, the GIC_2 can be consistent when $\mathrm{E}(\varepsilon^8) < \infty$. That is, the GIC_2 can be consistent when the tail of the error distribution is heavier than that of the Gaussian distribution. The GIC_3 and GIC_4 are the RIC of Foster and George (1994) and the corrected RIC of Zhang and Shen (2010). The GIC_5 and GIC_6 are consistent when the error distribution is Gaussian.

5.1 Simulation 1

The first simulation model is

$$y = \mathbf{x}^{'}\beta^* + \varepsilon$$

where $\mathbf{x} = (x_1,\ldots,x_p)^{'}$ is a multivariate Gaussian random vector with mean 0 and covariances of x_k and x_l being $0.5^{|k-l|}$. The ε is a random variable with mean 0 and $\sigma^2 = 4$. For $\beta^* = (3, 1.5, 0, 0, 2, 0_{p-5}^{'})^{'}$ with 0_k denoting a $k-$dimensional vector of zeros. This simulation setup was considered in Fan and

Li (2001). We consider two distributions for ε : the Gaussian distribution and the t-distribution with 3 degrees of freedom multiplied by a positive constant to make the variance be 4.

First, we compare performances of the GICs applied to all possible submodels with those applied to submodels constructed by the solution path of a sparse penalized approach. For a sparse penalized approach, we use the SCAD penalty with the PLUS algorithm (Zhang, 2010). Table 1 summarizes the results when $p = 10$ and $n = 100$ based on 300 repetitions of the simulation. In the table, 'Signal', 'Noise', 'PTM' and 'Error (s.e.)' represent the average number of variables included in the selected model among the signal variables, the average number of variables included in the selected model among noisy variables, the proportion of the true model being exactly identified, and the average of the squared Euclidean distance of $\hat{\beta}_{\hat{\pi}_{\lambda_n}}$ form β^* with the standard error in the parenthesis, respectively. From Table 1, we can see that the results based on the SCAD solution path are almost identical to those based on the all possible search, which suggests that the model selection with the SCAD solution path is a promising alternative to all possible search.

Submodels	Criterion	Signal	Noise	PTM	Error (s.e.)
All	GIC_1	3	0.22	0.80	0.220 (0.013)
	GIC_2	3	0.92	0.39	0.371 (0.018)
	GIC_3	3	0.22	0.80	0.220 (0.013)
	GIC_4	3	0.09	0.91	0.190 (0.016)
	GIC_5	3	0.39	0.67	0.267 (0.016)
	GIC_6	3	0.02	0.98	0.158 (0.015)
SCAD	GIC_1	3	0.21	0.80	0.218 (0.013)
	GIC_2	3	0.93	0.40	0.367 (0.018)
	GIC_3	3	0.21	0.80	0.218 (0.013)
	GIC_4	3	0.10	0.90	0.191 (0.016)
	GIC_5	3	0.39	0.67	0.266 (0.016)
	GIC_6	3	0.03	0.97	0.163 (0.015)

Table 1: Comparison of the 6 GICs with the all possible search and SCAD solution path when $p = 10$ and $n = 100$.

For simulation with high-dimensional models, we consider $p = 500$ and $p = 3000$. The results of prediction accuracy and variable selectivity for $n = 100$ and $n = 300$ with the error distribution being the Gaussian and t-distributions are presented in Tables 2 and 3, respectively. We use the SCAD solution path to construct a sequence of submodels. The values are the averages based on 300 repetitions of the simulation.

First of all, the GIC_1 (the BIC) is the worst in terms of prediction accuracy for $p = 500$ and $p = 3000$. This is mainly because the GIC_1 selects too many noisy variables compared to the other selection criteria even though it detects signal variables well. The GIC_4 is the best in terms of both the prediction accuracy and variable selectivity for $n = 100$, and the GIC_6 is the best for $n = 300$. The GIC_2, GIC_3 and GIC_5 perform reasonably well but tend to select variables more necessary. By comparing the results of the Gaussian and t distributions, we have found that less signal and more noisy variables are selected when the tail of the error distribution is heavier. However, the relative performances of the model selection criteria are similar. That is, the GIC_1 is the worst, the GIC_4 and GIC_6 are the best and so on. Based on these observations, we conclude that (i) model

n	p	Criterion	Signal	Noise	PTM	Error (s.e.)
100	500	GIC_1	2.99	4.35	0.00	1.369 (0.039)
		GIC_2	2.98	1.25	0.26	0.706 (0.037)
		GIC_3	2.96	0.20	0.80	0.351 (0.036)
		GIC_4	2.95	0.05	0.90	0.289 (0.036)
		GIC_5	2.98	0.67	0.52	0.509 (0.035)
		GIC_6	2.81	0.00	0.81	0.620 (0.061)
	3000	GIC_1	2.99	5.69	0.00	1.667 (0.036)
		GIC_2	2.94	0.26	0.76	0.444 (0.047)
		GIC_3	2.92	0.14	0.82	0.431 (0.049)
		GIC_4	2.89	0.05	0.87	0.445 (0.053)
		GIC_5	2.95	0.58	0.55	0.569 (0.046)
		GIC_6	2.63	0.00	0.63	1.092 (0.075)
300	500	GIC_1	3	4.89	0.00	0.561 (0.015)
		GIC_2	3	1.69	0.15	0.280 (0.010)
		GIC_3	3	0.17	0.84	0.083 (0.005)
		GIC_4	3	0.03	0.97	0.057 (0.004)
		GIC_5	3	0.40	0.66	0.119 (0.007)
		GIC_6	3	0.00	1.00	0.049 (0.002)
	3000	GIC_1	3	9.80	0.00	1.045 (0.018)
		GIC_2	3	0.38	0.67	0.136 (0.008)
		GIC_3	3	0.20	0.83	0.099 (0.007)
		GIC_4	3	0.02	0.98	0.057 (0.004)
		GIC_5	3	0.47	0.60	0.154 (0.009)
		GIC_6	3	0.00	1.00	0.050 (0.002)

Table 2: Comparison of the 6 GICs with Simulation 1 when the error follows the Gaussian distribution.

selection criteria specialized for high-dimensional models are necessary for optimal prediction and variable selection, (ii) finite sample performances of consistent GICs are quite different, and (iii) the tail lightness of the error distribution does not affect seriously to relative performances of model selection criteria.

5.2 Simulation 2

We consider a more challenging case by modifying the model for Simulation 1. We divide the p components of β^* into continuous blocks of size 20. We randomly select 5 blocks and assign the value $(3, 1.5, 0, 0, 2, 0'_{15})/1.5$ to each block. The entries in other blocks are set to be zero.

The results are summarized in Tables 4 and 5. We observe similar phenomena as in Simulation 1: the GIC_1 is the worst, the GIC_4 and GIC_6 are the best and etc. However, when $n = 100$, the GIC_1 is better in terms of prediction accuracy than some other GICs which are selection consistent, which is an example of the conflict between selection consistency and prediction optimality.

n	p	Criterion	Signal	Noise	PTM	Error (s.e.)
100	500	GIC_1	2.98	4.27	0.09	2.236 (0.702)
		GIC_2	2.97	1.24	0.51	1.478 (0.696)
		GIC_3	2.96	0.48	0.81	1.224 (0.696)
		GIC_4	2.94	0.35	0.86	1.198 (0.695)
		GIC_5	2.97	0.82	0.68	1.348 (0.696)
		GIC_6	2.84	0.12	0.83	1.271 (0.692)
	3000	GIC_1	2.96	5.45	0.01	1.683 (0.106)
		GIC_2	2.92	0.51	0.74	0.701 (0.094)
		GIC_3	2.91	0.40	0.78	0.673 (0.088)
		GIC_4	2.88	0.22	0.82	0.619 (0.086)
		GIC_5	2.94	0.69	0.68	0.729 (0.093)
		GIC_6	2.59	0.03	0.59	1.273 (0.086)
300	500	GIC_1	3	4.26	0.06	0.501 (0.034)
		GIC_2	3	1.52	0.38	0.261 (0.022)
		GIC_3	3	0.28	0.84	0.100 (0.013)
		GIC_4	3	0.08	0.95	0.063 (0.008)
		GIC_5	3	0.49	0.75	0.133 (0.016)
		GIC_6	3	0.00	1.00	0.044 (0.003)
	3000	GIC_1	3	9.58	0.00	1.057 (0.061)
		GIC_2	3	0.83	0.71	0.248 (0.043)
		GIC_3	3	0.59	0.81	0.205 (0.042)
		GIC_4	3	0.24	0.91	0.131 (0.029)
		GIC_5	3	0.90	0.68	0.262 (0.044)
		GIC_6	3	0.02	0.99	0.062 (0.019)

Table 3: Comparison of the 6 GICs with Simulation 1 when the error follows the t-distribution.

5.3 Real Data Analysis

We analyze the data set used in Scheetz et al. (2006), which consists of gene expression levels of 18,975 genes obtained from 120 rats. The main objective of the analysis is to find genes that are correlated with gene TRIM32 known to cause Bardet-Biedl syndromes. As was done by Huang et al. (2008), we first select 3000 genes with the largest variance in expression level, and then choose the top p genes that have the largest absolute correlation with gene TRIM32 among the selected 3000 genes.

We compare prediction accuracies of the 6 GICs with the submodels obtained from the SCAD solution path. Each data set was divided into two parts, training and test data sets, by randomly selecting 2/3 observations and 1/3 observations, respectively. We use the training data set to select the model and estimate the regression coefficients, and use the test data set to evaluate the prediction performance.

For estimation of the error variance, Zou et al. (2007) used the mean squared error of the full model when $p < n$. This approach, however, is not applicable to our data set since $p > n$. A heuristic method is to set p_{max} first, and to select a model among the SCAD solution path whose number of

n	p	Criterion	Signal	Noise	PTM	Error (s.e.)
100	500	GIC_1	14.82	5.11	0.00	3.553 (0.225)
		GIC_2	14.67	2.39	0.14	3.211 (0.242)
		GIC_3	14.40	1.47	0.24	3.654 (0.285)
		GIC_4	14.17	1.16	0.25	4.212 (0.302)
		GIC_5	14.57	1.86	0.21	3.242 (0.254)
		GIC_6	13.04	0.72	0.16	7.758 (0.398)
	3000	GIC_1	12.08	12.19	0.00	20.192 (1.186)
		GIC_2	11.51	5.78	0.01	19.783 (1.061)
		GIC_3	11.36	5.34	0.01	20.051 (1.055)
		GIC_4	11.06	4.37	0.01	20.649 (1.021)
		GIC_5	11.68	6.62	0.01	19.616 (1.103)
		GIC_6	10.11	2.47	0.01	22.755 (0.894)
300	500	GIC_1	15	4.56	0.00	0.795 (0.015)
		GIC_2	15	1.63	0.17	0.516 (0.013)
		GIC_3	15	0.19	0.82	0.311 (0.009)
		GIC_4	15	0.03	0.97	0.278 (0.007)
		GIC_5	15	0.39	0.68	0.345 (0.011)
		GIC_6	15	0.00	1.00	0.270 (0.006)
	3000	GIC_1	15	9.60	0.00	1.322 (0.020)
		GIC_2	15	0.32	0.72	0.340 (0.010)
		GIC_3	15	0.14	0.88	0.300 (0.008)
		GIC_4	15	0.01	0.99	0.267 (0.006)
		GIC_5	15	0.40	0.66	0.358 (0.010)
		GIC_6	15	0.00	1.00	0.264 (0.006)

Table 4: Comparison of the 6 GICs with Simulation 2 when the error follows the Gaussian distribution.

nonzero coefficients is equal to p_{max}, and to estimate the error variance by the mean squared error of the selected model. Following the results of Scheetz et al. (2006), Chiang et al. (2006), Huang et al. (2008), and Kim et al. (2008), we guess that a reasonable model size would be in between 20 and 40. Table 6 compares the 6 GICs with the number of pre-screened genes being $p = 500$ and $p = 3000$, when the error variance is estimated with p_{max} being 20, 30 and 40, respectively. All values are the arithmetic means of the results from 100 replicated random partitions. In the table, 'Nonzero' denotes the number of nonzero coefficients in the selected model and 'Error (s.e.)' is the prediction error on the test data set and the standard error in the parenthesis obtained on the test data. For $p = 500$, the lowest prediction error is achieved by the GIC_2 and the GIC_3, GIC_4 and GIC_5 perform reasonably well with $p_{max} = 20$. For $p = 3000$, the lowest prediction error is achieved by the GIC_5 with $p_{max} = 20$. So, we choose $p_{max} = 20$ for estimation of the error variance.

As argued by Yang (2005), the standard error obtained by random partition could be misleading. As a supplement, we draw the box plots of the 100 prediction errors of the 6 GICs with $p_{max} = 20$ obtained from 100 random partitions in Figure 1. The relative performances of the GICs with the real data are different from those of the simulation studies in the previous subsections. The GIC_2,

n	p	Criterion	Signal	Noise	PTM	Error (s.e.)
100	500	GIC_1	14.65	3.89	0.07	3.974 (0.401)
		GIC_2	14.55	2.07	0.35	3.686 (0.411)
		GIC_3	14.40	1.45	0.41	3.870 (0.421)
		GIC_4	14.17	1.10	0.41	4.378 (0.424)
		GIC_5	14.53	1.85	0.38	3.649 (0.412)
		GIC_6	13.00	0.59	0.25	7.848 (0.471)
	3000	GIC_1	11.99	9.41	0.02	19.768 (1.154)
		GIC_2	11.53	5.23	0.08	19.806 (1.066)
		GIC_3	11.47	4.78	0.08	19.641 (1.029)
		GIC_4	11.19	3.89	0.08	19.968 (0.959)
		GIC_5	11.61	5.92	0.08	19.96 (1.101)
		GIC_6	10.35	2.28	0.03	21.96 (0.899)
300	500	GIC_1	14.99	4.81	0.05	0.990 (0.098)
		GIC_2	14.99	2.33	0.32	0.748 (0.098)
		GIC_3	14.99	0.75	0.78	0.519 (0.094)
		GIC_4	14.99	0.40	0.89	0.451 (0.090)
		GIC_5	14.99	1.00	0.66	0.565 (0.094)
		GIC_6	14.99	0.06	0.98	0.339 (0.053)
	3000	GIC_1	15	8.18	0.00	1.226 (0.051)
		GIC_2	15	0.58	0.73	0.420 (0.040)
		GIC_3	15	0.31	0.86	0.358 (0.037)
		GIC_4	15	0.12	0.95	0.314 (0.032)
		GIC_5	15	0.63	0.70	0.430 (0.041)
		GIC_6	15	0.01	0.99	0.272 (0.015)

Table 5: Comparison of the 6 GICs with Simulation 2 when the error follows the t-distribution.

GIC_3 and GIC_5 have lower prediction errors than the GIC_4 and GIC_6 while the formers tend to select more variables than necessary in the simulation studies. This observation suggests that there might be many signal genes whose impacts on the response variable are relatively small.

6. Concluding Remarks

The range of consistent model selection criteria is rather large, and it is not clear which one is better with finite samples. It would be interesting to rewrite the class of GICs as $\{\lambda_n = \alpha_n \log p_n : \alpha_n > 0\}$. The GIC_3, GIC_5 and GIC_6 correspond to $\alpha_n = 2$, $\alpha_n = \log\log n$ and $\alpha_n = \log n$, respectively. When the rue model is expected to be very sparse, it would be better to let α_n be rather large (e.g., $\alpha_n = \log n$), while a smaller α_n (e.g., $\alpha_n = 2$ or $\alpha_n = \log\log n$) would be better when many signal covariates with small regression coefficients are expected to exist. The relation of the GICs with larger α_n with those with smaller α_n would be similar to the relation between the AIC and BIC for standard fixed dimensional models.

$p_{max} = 20$

	p			
	500		3000	
	Error (s.e.)	Nonzero	Error (s.e.)	Nonzero
GIC_1	0.742 (0.038)	15.91	0.766 (0.036)	18.62
GIC_2	**0.649** (0.028)	10.95	0.686 (0.035)	3.91
GIC_3	0.656 (0.031)	6.99	0.697 (0.035)	3.69
GIC_4	0.677 (0.034)	5.57	0.719 (0.037)	2.78
GIC_5	0.664 (0.030)	9.76	**0.667** (0.032)	4.92
GIC_6	0.732 (0.038)	3.03	0.792 (0.039)	1.82

$p_{max} = 30$

	p			
	500		3000	
	Error (s.e.)	Nonzero	Error (s.e.)	Nonzero
GIC_1	0.890 (0.035)	27.26	0.868 (0.039)	26.07
GIC_2	0.825 (0.038)	21.77	0.698 (0.031)	14.04
GIC_3	0.752 (0.029)	17.53	0.696 (0.031)	13.25
GIC_4	**0.722** (0.029)	15.19	0.691 (0.034)	10.76
GIC_5	0.800 (0.030)	20.29	0.729 (0.032)	15.99
GIC_6	0.688 (0.030)	11.31	**0.683** (0.034)	5.53

$p_{max} = 40$

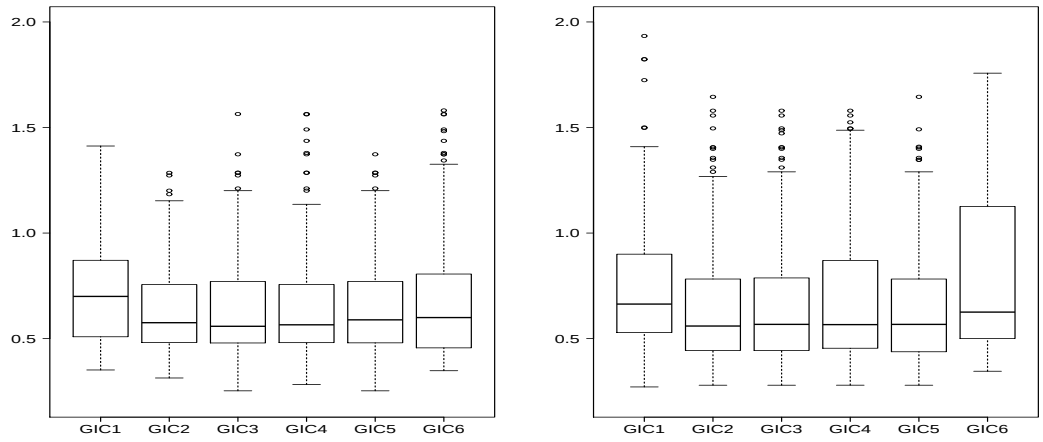
	p			
	500		3000	
	Error (s.e.)	Nonzero	Error (s.e.)	Nonzero
GIC_1	1.040 (0.077)	34.54	0.936 (0.041)	33.80
GIC_2	0.916 (0.036)	29.59	0.892 (0.041)	27.27
GIC_3	0.859 (0.035)	25.10	0.878 (0.040)	26.37
GIC_4	**0.846** (0.039)	23.02	0.846 (0.038)	25.00
GIC_5	0.890 (0.035)	28.20	0.910 (0.040)	28.60
GIC_6	0.763 (0.029)	18.69	**0.800** (0.037)	21.02

Table 6: Comparison of the 6 GICs with the gene expression data. The bold face numbers represent the lowest prediction errors among the 6GICs.

Estimation of σ^2 is an open question. We may use the BIC-like criterion by assuming the Gaussian distribution:

$$\hat{\pi}_{\lambda_n} = \text{argmin}_{\pi \subset \{1,\ldots,p_n\}} \log(R_n(\hat{\beta}_\pi)/n) + \lambda_n |\pi|.$$

If $R_n(\hat{\beta}_\pi)/n$ is bounded above from ∞ and below from 0 in probability (uniformly in π and n), we could derive similar asymptotic properties for the BIC-like criteria as the GICs. We leave this problem as future work.



(a) $p = 500$ (b) $p = 3000$

Figure 1: The boxplot of the prediction errors when (a) $p = 500$ and (b) $p = 3000$ with $p_{max} = 20$.

For consistency, the smallest eigenvalue of the design matrix of the true model is assumed to be sufficiently large (i.e., condition A2). However, it is frequently observed for large dimensional data that some covariates are highly correlated and they affect the output similarly. In this case, selecting some covariates and ignoring the others, which is done by a standard model selection method, is not optimal. See Zou and Hastie (2005) for an example. It would be interesting to develop consistent model selection methods for such cases.

Acknowledgments

This research was supported by the National Research Foundation of Korea grant number 20100012671 funded by the Korea government.

Appendix A. Proof of Theorem 2

Without loss of generality, we let $\pi_n^* = \{1, \ldots, q_n\}$. Let $\hat{\beta}^* = \hat{\beta}_{\pi_n^*}$. Let $\hat{Y}_\pi = \mathbf{X}_n \hat{\beta}_\pi$ and $\hat{Y}_n^* = \mathbf{X}_n \hat{\beta}_{\pi_n^*}$. We let $\beta^* = (\beta^{(1)*}, \beta^{(2)*})$, where $\beta^{(1)*} \in R^{q_n}$ and $\beta^{(2)*} \in R^{p_n - q_n}$. Let $\mathbf{C}_n = \mathbf{X}_n' \mathbf{X}_n / n$ and $\mathbf{C}_n^{(i,j)} = \mathbf{X}_n^{(i)'} \mathbf{X}_n^{(j)} / n$ for $i, j = 1, 2$. We need the following two lemmas.

Lemma 8

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

Proof. Let $z_j = \sqrt{n}(\hat{\beta}_j^* - \beta_j^*)$. For proving Lemma 8, we will show

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

Write

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

where $\mathbf{z} = (z_1,\ldots,z_{q_n})'$, $\varepsilon_n = (\varepsilon_1,\ldots,\varepsilon_n)'$ and $\mathbf{H}^{(1)'} = (\mathbf{h}_1^{(1)},\ldots,\mathbf{h}_{q_n}^{(1)})' = (\mathbf{C}_n^{(1,1)})^{-1}\mathbf{X}_n^{(1)'}/\sqrt{n}$. Since $\mathbf{H}^{(1)'}\mathbf{H}^{(1)} = (\mathbf{C}_n^{(1,1)})^{-1}$, A2 of the regularity conditions implies $\|\mathbf{h}_j^{(1)}\|_2^2 \le 1/M_2$ for all $j \le q_n$. Hence, $\mathrm{E}(z_j)^{2k} < \infty$ for all $j \le q_n$ since $\mathrm{E}(\varepsilon_i)^{2k} < \infty$. Thus

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

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

$$\begin{aligned}
\Pr(|z_j| > \eta n^{c_2/2} \text{ for some } j = 1,\ldots,q_n) &\le \sum_{j=1}^{q_n}\Pr(|z_j| > \eta n^{c_2/2}) \\
&\le \sum_{j=1}^{q_n}\frac{1}{\eta}n^{-c_2k} \\
&= \frac{1}{\eta}q_n n^{-c_2k} \le \frac{1}{\eta}n^{-(c_2-c_3)k} \to 0,
\end{aligned}$$

which completes the proof. ■

Lemma 9

$$\max_{q_n < j \le p_n} |< Y_n - \hat{Y}_n^*, X_n^j >| = o_p(\sqrt{n\lambda_n\rho_n}).$$

Proof. Note that

$$\begin{aligned}
& (< Y_n - \hat{Y}_n^*, X_n^j >, j = q_n+1,\ldots,p_n) \\
= \;& \mathbf{X}_n^{(2)'}\left(Y_n - \mathbf{X}_n^{(1)}\hat{\beta}^{*(1)}\right) \\
= \;& \mathbf{X}_n^{(2)'}\left(Y_n - \mathbf{X}_n^{(1)}\frac{1}{n}(\mathbf{C}_n^{(1,1)})^{-1}\mathbf{X}_n^{(1)'}Y_n\right) \\
= \;& \mathbf{X}_n^{(2)'}\left(\mathbf{X}_n^{(1)}\beta^{*(1)} + \varepsilon_n - \mathbf{X}_n^{(1)}\frac{1}{n}(\mathbf{C}_n^{(1,1)})^{-1}\mathbf{X}_n^{(1)'}(\mathbf{X}_n^{(1)}\beta^{*(1)} + \varepsilon_n)\right) \\
= \;& \mathbf{X}_n^{(2)'}\left(\mathbf{I} - \frac{1}{n}\mathbf{X}_n^{(1)}(\mathbf{C}_n^{(1,1)})^{-1}\mathbf{X}_n^{(1)'}\right)\varepsilon_n.
\end{aligned}$$

Hence, we have

$$< Y_n - \hat{Y}_n^*, X_n^j > /\sqrt{n} = \mathbf{h}_j^{(2)'}\varepsilon_n \quad \text{for } j = q_n+1,\ldots,p_n, \tag{3}$$

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

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

Note that

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

Since the all eigenvalues of $\mathbf{I}-\mathbf{X}_n^{(1)}(\mathbf{X}_n^{(1)'}\mathbf{X}_n^{(1)})^{-1}\mathbf{X}_n^{(1)'}$ are between 0 and 1, we have $\|\mathbf{h}_j^{(2)}\|_2^2 \le M_1$ for all $j = q_n+1,\ldots,p_n$. Hence, $\mathrm{E}(\xi_j)^{2k} < \infty$, where $\xi_j = < Y_n - \hat{Y}_n^*, X_n^j > /\sqrt{n}$, and so

$$Pr\left(|\xi_j| > t\right) = O(t^{-2k}).$$

Finally, for any $\eta > 0$,

$$\begin{aligned}
& \Pr\left(| < Y_n - \hat{Y}_n^*, X_n^j > | > \eta\sqrt{n\lambda_n\rho_n} \text{ for some } j = q_n+1,\ldots,p_n \right) \\
= \;& \Pr\left(|\xi_j| > \eta\sqrt{\lambda_n\rho_n} \text{ for some } j = q_n+1,\ldots,p_n \right) \\
\le \;& \sum_{j=q_n+1}^{p_n} \Pr\left(|\xi_j| > \eta\sqrt{\lambda_n\rho_n} \right) \\
= \;& (p_n - q_n) O\left(\frac{1}{(\lambda_n\rho_n)^k} \right) = O\left(\frac{p_n}{(\lambda_n\rho_n)^k} \right) \to 0,
\end{aligned}$$

which completes the proof. ■

Proof of Theorem 2. For any π, we can write

$$\begin{aligned}
& R_n(\hat{\beta}_\pi) + \lambda_n|\pi|\sigma^2 - R_n(\hat{\beta}^*) - \lambda_n|\pi_n^*|\sigma^2 \\
= \; & -2\textstyle\sum_{j=q_n+1}^{p_n} \hat{\beta}_{\pi,j} < Y_n - \hat{Y}_n^*, X^j > + (\hat{\beta}_\pi - \hat{\beta}^*)'(\mathbf{X}_n'\mathbf{X}_n)(\hat{\beta}_\pi - \hat{\beta}^*) + \lambda_n(|\pi| - |\pi_n^*|)\sigma^2.
\end{aligned}$$

By Condition A3,

$$(\hat{\beta}_\pi - \hat{\beta}^*)'(\mathbf{X}_n'\mathbf{X}_n)(\hat{\beta}_\pi - \hat{\beta}^*) \ge \sum_{j\in\pi\cup\pi^*} n\rho_n(\hat{\beta}_{\pi,j} - \hat{\beta}_j^*)^2.$$

Hence, we have for any $\pi \in \mathcal{M}^{s_n}$,

$$R_n(\hat{\beta}_\pi) + \lambda_n|\pi|\sigma^2 - R_n(\hat{\beta}^*) - \lambda_n|\pi_n^*|\sigma^2 \ge \sum_{j\in\pi\cup\pi_n^*} w_j,$$

where

$$w_j = -2\hat{\beta}_{\pi,j} < Y_n - \hat{Y}_n^*, X_n^j > I(j \notin \pi_n^*) + n\rho_n(\hat{\beta}_{\pi,j} - \hat{\beta}_j^*)^2 + \lambda_n(I(j\in\pi - \pi_n^*) - I(j\in\pi_n^* - \pi))\sigma^2.$$

For $j \in \pi_n^* - \pi$, we have $w_j = n\rho_n\hat{\beta}_j^{*2} - \lambda_n\sigma^2$. Let

$$A_n = \{n\rho_n\hat{\beta}_j^{*2} - \lambda_n\sigma^2 > 0, j = 1,\ldots,q_n\}. \tag{4}$$

Then, $\Pr(A_n) \to 1$ by Lemma 8 and Conditions A3 and A4.

For $j \in \pi - \pi_n^*$

$$\begin{aligned}
w_j &= -2\hat{\beta}_{\pi,j} < Y_n - \hat{Y}_n^*, X_n^j > + n\rho_n\hat{\beta}_{\pi,j}^2 + \lambda_n\sigma^2 \\
&\ge - < Y_n - \hat{Y}_n^*, X_n^j >^2 /(n\rho_n) + \lambda_n\sigma^2.
\end{aligned}$$

Let

$$B_n = \{- < Y_n - \hat{Y}_n^*, X_n^j >^2 /(n\rho_n) + \lambda_n\sigma^2 > 0, j = q_n+1,\ldots,p_n\}. \tag{5}$$

Then, $\Pr(B_n) \to 1$ by Lemma 9.

For $j \in \pi \cap \pi_n^*$,

$$w_j = n\rho_n(\hat{\beta}_{\pi,j} - \hat{\beta}_j^*)^2 \geq 0.$$

To sum up, on $A_n \cap B_n$,

$$R_n(\hat{\beta}_\pi) + \lambda_n|\pi|\sigma^2 - R_n(\hat{\beta}^*) - \lambda_n|\pi_n^*|\sigma^2 > 0$$

for all $\pi \neq \pi_n^*$. Since $\Pr(A_n \cap B_n) \to 1$, the proof is done. ■

Appendix B. Proof of Theorem 3

For given $\pi \subset \{1,\ldots,p_n\}$, let $\mathbf{M}_\pi$ be the projection operator onto the space spanned by $(X^{(j)}, j \in \pi)$. That is, $\mathbf{M}_\pi = \mathbf{X}_\pi(\mathbf{X}_\pi^{'}\mathbf{X}_\pi)^{-1}\mathbf{X}_\pi^{'}$ provided $\mathbf{X}_\pi$ is of full rank. Let $\mathbf{X}_n\beta_n^* = \mu_n$ and $\mathbf{I}$ be the $n \times n$ identity matrix. Without loss of generality, we assume $\sigma^2 = 1$.

Lemma 10 *There exists* $\eta > 0$ *such that for any* $\pi \in \mathcal{M}^{s_n}$ *with* $\pi_n^* \not\subseteq \pi$,

$$\mu_n^{'}(\mathbf{I} - \mathbf{M}_\pi)\mu_n \geq \eta|\pi^-|n^{c_2-c_1},$$

where $\pi^- = \pi_n^* - \pi$.

Proof. For given $\pi \in \mathcal{M}^{s_n}$ with $\pi_n^* \not\subseteq \pi$, we have

$$\begin{aligned}
& \mu_n^{'}(\mathbf{I} - \mathbf{M}_\pi)\mu_n \\
= & \inf_{\alpha \in R^{|\pi|}} \|\mathbf{X}_{\pi^-}\beta_{\pi^-}^* - \mathbf{X}_\pi\alpha\|^2 \\
= & \inf_{\alpha \in R^{|\pi|}} (\beta_{\pi^-}^{*'}, \alpha^{'})(\mathbf{X}_{\pi^-}, \mathbf{X}_\pi)^{'}(\mathbf{X}_{\pi^-}, \mathbf{X}_\pi)(\beta_{\pi^-}^{*'}, \alpha^{'})^{'} \\
\geq & n\|\beta_{\pi^-}^*\|^2\rho_n \\
\geq & M_3M_4|\pi^-|n^{c_2-c_1},
\end{aligned}$$

where $\beta_{\pi^-}^* = (\beta_j^*, j \in \pi^-)$ and the last inequality is due to Condition A4. ■

Lemma 11 *For given* $\pi \subset \{1,\ldots,p_n\}$, *let*

$$Z_\pi = \frac{\mu_n^{'}(\mathbf{I} - \mathbf{M}_\pi)\varepsilon_n}{\sqrt{\mu_n^{'}(\mathbf{I} - \mathbf{M}_\pi)\mu_n}}.$$

Then

$$\max_{\pi \in \mathcal{M}^{s_n}} |Z_\pi| = O_p(\sqrt{s_n \log p_n}).$$

Proof. Note that $Z_\pi \sim N(0,1)$ for all $\pi \in \mathcal{M}^{s_n}$. Since

$$\Pr(|Z_\pi| > t) \leq C\exp(-t^2/2) \tag{6}$$

for some $C > 0$, we have

$$\begin{aligned} \Pr\left(\max_{\pi \in \mathcal{M}^{s_n}} |Z_\pi| > t\right) &\leq \sum_{\pi \in \mathcal{M}^{s_n}} C\exp(-t^2/2) \\ &\leq C p_n^{s_n} \exp(-t^2/2). \end{aligned}$$

Hence, if we let $t = \sqrt{w s_n \log p_n}$,

$$\Pr\left(\max_{\pi \in \mathcal{M}^{s_n}} |Z_\pi| > t\right) \leq C\exp((-w/2+1)s_n \log p_n) \to 0$$

as $w \to \infty$. ■

Lemma 12

$$\max_{\pi \in \mathcal{M}^{s_n}} \varepsilon_n^{'} \mathbf{M}_\pi \varepsilon_n = O_p(s_n \log p_n).$$

Proof. For given $\pi \subset \{1, \ldots, p_n\}$, let $r(p)$ be the rank of $\mathbf{X}_\pi$. Note that $\varepsilon_n^{'} \mathbf{M}_\pi \varepsilon_n \sim \chi^2(r(\pi))$ where $\chi^2(k)$ is the chi-square distribution with degree of freedom k. It is easy to see that (see, for example, Yang 1999)

$$\Pr(\varepsilon_n^{'} \mathbf{M}_\pi \varepsilon_n \geq t) \leq \exp\left(-\frac{t - r(\pi)}{2}\right)\left(\frac{t}{r(\pi)}\right)^{r(\pi)/2}. \tag{7}$$

Hence

$$\Pr\left(\max_{\pi \in \mathcal{M}^{s_n}} \varepsilon_n^{'} \mathbf{M}_\pi \varepsilon_n \geq t\right) \leq \sum_{k=1}^{s_n} \binom{p_n}{k} \Pr(W_k \leq t),$$

where $W_k \sim \chi^2(k)$. Since $\Pr(W_k \geq t) \leq \Pr(W_{s_n} \geq t)$, we have

$$\begin{aligned} \Pr\left(\max_{\pi \in \mathcal{M}^{s_n}} \mathbf{e}_n^{'} \mathbf{M}_\pi \mathbf{e}_n \geq t\right) &\leq \Pr(W_{s_n} \geq t) \sum_{k=1}^{s_n} \binom{p_n}{k} \\ &\leq \Pr(W_{s_n} \geq t) p_n^{s_n}. \end{aligned} \tag{8}$$

The proof is done by applying (7) to (8). ■

Proof of Theorem 3. First, we will show that $\Pr(\pi_n^* \not\subseteq \hat{\pi}_{\lambda_n}) \to 0$. For given $\pi \subset \{1, \ldots, p_n\}$, let $R_n(\pi) = R_n(\hat{\beta}_\pi)$. Note that $R_n(\pi) = Y_n^{'}(\mathbf{I} - \mathbf{M}_\pi)Y_n$. For $\pi \not\supseteq \pi_n^*$, Lemmas 10, 11 and 12 imply

$$\begin{aligned} & R_n(\pi) - R_n(\pi_n^*) + \lambda_n(|\pi| - |\pi_n^*|)\sigma^2 \\ = \;& \mu_n^{'}(\mathbf{I} - \mathbf{M}_\pi)\mu_n + 2\mu_n^{'}(\mathbf{I} - \mathbf{M}_\pi)\varepsilon_n + \varepsilon_n^{'}(\mathbf{M}_{\pi^*} - \mathbf{M}_\pi)\varepsilon_n + \lambda_n(|\pi| - |\pi_n^*|)\sigma^2 \\ \geq \;& \eta|\pi^-|n^{c_2 - c_1} - 2\sqrt{\eta|\pi^-|n^{c_2-c_1}} O_p(\sqrt{s_n \log p_n}) - O_p(s_n \log p_n) - |\pi^-|\lambda_n, \end{aligned}$$

where $\pi^- = \pi_n^* - \pi$. Since $s_n \log p_n \leq o(n^{c_2 - c_1})$ and $\lambda_n = o(n^{c_2 - c_1})$, the proof is done.

It remains to show that the probability of

$$\inf_{\pi \in \mathcal{M}^{s_n}, \pi \not\supseteq \pi_n^*} R_n(\pi) - R_n(\pi_n^*) + \lambda_n(|\pi| - |\pi_n^*|)\sigma^2 > 0 \tag{9}$$

converges to 1. By Theorem 1 of Zhang and Shen (2010), the probability of (9) is larger than

$$2-\left(1+e^{1/2}\exp\left(-\frac{\lambda_n-\log\lambda_n}{2}\right)\right)^{p_n-q_n},$$

which converges to 1 when $2\log p_n-\lambda_n+\log\lambda_n\to-\infty$. The equivalent condition with $2\log p_n-\lambda_n+\log\lambda_n\to-\infty$ is $\lambda_n-2\log p_n-\log\log p_n\to\infty$. ■

Appendix C. Proof of Theorem 4

By Theorem 4 of Kim and Kwon (2012), the solution path of the SCAD or minimax concave penalty include the true model with probability converging to 1. Since condition A3' is stronger than condition A3, the GIC_{λ_n} with $\lambda_n=o(n^{c_2-c_1})$ is consistent, and so is with the solution path of the SCAD or minimax concave penalty.

Appendix D. Proof of Theorem 7

Let $\tilde{A}_n$ and $\tilde{B}_n$ be the sets defined in (4) and (5) except that σ^2 is replaced by $\hat{\sigma}^2$. It suffices to show that $\Pr(\tilde{A}_n\cap\tilde{B}_n)\to 1$. It is not difficult to prove $\Pr(\tilde{A}_n)\to 1$ by Lemma 8 and (2).

For $\tilde{B}_n$, since $\varepsilon_i\sim N(0,\sigma^2)$, (3) implies

$$<Y_n-\hat{Y}_n^*,X_n^j>/\sqrt{n}\sim N(0,\sigma_j^2)$$

where $\sigma_j^2\le\sigma^2M_1$. By (6), we have

$$\begin{aligned}\Pr(\tilde{B}_n^c) &\le \Pr(<Y_n-\hat{Y}_n^*,X_n^j>^2>n\rho_n\lambda_n\hat{\sigma}^2 \text{ for some } j=q_n+1,\ldots,p_n)\\ &\le Cp_n\exp(-\rho_n r_{inf}\lambda_n/2M_1).\end{aligned}$$

Hence, as long as $2M_1\log p_n/(\rho_n r_{inf})-\lambda_n\to-\infty$, $\Pr(\tilde{B}_n^c)\to 0$ and the proof is done. ■

References

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrox and F. Caski, editors, *Second International Symposium on Information Theory*, volume 1, pages 267–281. Budapest: Akademiai Kiado, 1973.

K. W. Broman and T. P. Speed. A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society, Ser. B*, 64:641–656, 2002.

G. Casella, F. J. Giron, M. L. Martinez, and E. Moreno. Consistency of bayesian procedure for variable selection. *The Annals of Statistics*, 37:1207–1228, 2009.

J. Chen and Z. Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 24:759–771, 2008.

A. P. Chiang, J. S. Beck, H.-J. Yen, M. K. Tayeh, T. E. Scheetz, R. Swiderski, D. Nishimura, T. A. Braun, K.-Y. Kim, J. Huang, K. Elbedour, R. Carmi, D. C. Slusarski, T. L. Casavant, E. M. Stone, and V. C. Sheffield. Homozygosity mapping with snp arrays identifies a novel gene for bardet-biedl syndrome (bbs10). *Proc. Nat. Acad. Sci.*, 103:6287–6292, 2006.

P. Craven and G. Wahba. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross validation. *Numer. Math.*, 31:377–403, 1979.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32:407–499, 2004.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

D. P. Foster and E. I. George. The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22:1947–1975, 1994.

E. Greenshtein and Y. Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10:971–988, 2004.

J. Huang, S. Ma, and C-H. Zhang. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18:1603–1618, 2008.

Y. Kim and S. Kwon. The global optimality of the smoothly clipped absolute deviation penalized estimator. *Biometrika*, forthcoming, 2012.

Y. Kim, H. Choi, and H. Oh. Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103:1665–1673, 2008.

T. E. Scheetz, K.-Y. A. Kim, R. E. Swiderski, A. R. Philp1, T. A. Braun, K. L. Knudtson, A. M. Dorrance, G. F. DiBona, J. Huang, T. L. Casavant, V. C. Sheffield, and E. M. Stone. Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103:14429–14434, 2006.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997.

M. Stone. Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Ser. B*, 39:111–147, 1974.

R. J. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Ser. B*, 58:267–288, 1996.

H. Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104:1512–1524, 2009.

H. Wang, B. Li, and C. Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society, Ser. B*, 71:671–683, 2009.

Y. Yang. Model selection for nonparametric regression. *Statistica Sinica*, 9:475–499, 1999.

Y. Yang. Can the strengths of aic and bic be shared? a conflict between model identification and regression estimation. *Biometrika*, 92:937–950, 2005.

Y. Yang and A. R. Barron. An asymptotic property of model selection criteria. *IEEE Transanctions in Information Theory*, 44:95–116, 1998.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38:894–942, 2010.

Y. Zhang and X. Shen. Model selection procedure for high-dimensional data. *Statistical Analysis and Data Mining*, 3:350–358, 2010.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Reserach*, 7:2541–2563, 2006.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Ser. B*, 67:301–320, 2005.

H. Zou, T. Hastie, and R. Tibshirani. On the "degree of freedom" of lasso. *The Annals of Statsitics*, 35:2173–2192, 2007.