Extended BIC for linear regression models with diverging number of relevant features and high or ultra-high feature spaces

By SHAN LUO[1] and ZEHUA CHEN[*2]

[1,2]Department of Statistics and Applied Probability

National University of Singapore

3 Science Drive 2

Singapore 117543

Republic of Singapore

Email: [1]luoshan08@nus.edu.sg. [2]stachenz@nus.edu.sg,

Running title: EBIC for linear models with diverging parameters

1

Summary

In many conventional scientific investigations with high or ultra-high dimensional feature spaces, the relevant features, though sparse, are large in number compared with classical statistical problems, and the magnitude of their effects tapers off. It is reasonable to model the number of relevant features as a diverging sequence when sample size increases. In this article, we investigate the properties of the extended Bayes information criterion (EBIC) (Chen and Chen, 2008) for feature selection in linear regression models with diverging number of relevant features in high or ultra-high dimensional feature spaces. The selection consistency of the EBIC in this situation is established. The application of EBIC to feature selection is considered in a two-stage feature selection procedure. Simulation studies are conducted to demonstrate the performance of the EBIC together with the two-stage feature selection procedure in finite sample cases.

Keywords: Diverging number of parameters, Feature selection, Extended Bayes information criterion, High dimensional feature space, Penalized likelihood, Selection consistency.

1   Introduction

In the setting of a regression model, if the number of features (covariates) p is of the polynomial order or exponential order of the sample size n, i.e., $p = O(n^\kappa)$ or $p = O(\exp(n^\kappa))$, the feature space is referred to as a high-dimensional or ultra-high dimensional feature space. Regression problems with high or ultra-high dimensional feature spaces arise in many important fields of scientific research such as genomics study, medical study, risk management, machine learning, etc.. Such problems are generally referred to as small-n-large-p problems. In many small-n-large-p problems the relevant (or causal, true, as referred by some other authors) features, though sparse, are relatively large in number compared with classical statistical problems, and their effects usually taper off to zero from the largest to the smallest. To reflect the estimability of the feature effects, it is reasonable to model the number of relevant features as a diverging sequence depending on the sample size. [7] and [11] are among the earliest papers dealing with diverging number of relevant features. In this article, we consider model selection criteria for linear regression models with high or ultra-high feature space and diverging number of relevant features.

In general, there are two goals in model selection. The first one is to select models to do prediction and the focus is on prediction accuracy. The second one is to identify relevant features and the focus is on selection consistency. In traditional model selection problems where the number of features under study is small, these two goals might be addressed at the same time. But, in small-n-large-p problems, the two goals need to be treated separately. We concentrate on the second goal in this article and refer to the problem as feature selection. A model selection criterion is crucial for feature selection. The traditional model selection criteria such as Akaike's information criterion (AIC) [1], cross-validation (CV) [16], generalized cross-validation (GCV) [6]

and the Bayes information criterion (BIC) [14] are not suitable for feature selection in small-n-large-p problems. The CV or GCV, which aims to minimize prediction errors, does not address the issue of selection consistency. The AIC and BIC are overly liberal; that is, the criteria select far more features than the relevant ones, see, [3, 15, 2]. [2] proposed a modified BIC (mBIC) for the study of genetic QTL mapping to address problems caused by too many features. [4] developed a family of extended Bayes information criteria (EBIC) for feature selection in small-n-large-p problems. The family of EBIC is indexed by a parameter γ in the range $[0, 1]$. The original BIC is a special case of EBIC with $\gamma = 0$. The mBIC is also a special case of EBIC in an asymptotic sense; that is, it is asymptotically equivalent to the EBIC with $\gamma = 1$. [4] considered the case of high dimensional feature space with fixed number of relevant features. They established the selection-consistency of EBIC when $p = O(n^\kappa)$ and $\gamma > 1 - \frac{1}{2\kappa}$ for any $\kappa > 0$.

Model selection criterion for diverging number of relevant features in high or ultra-high dimensional feature space is still almost a void. [19] considered a BIC type criterion for diverging number of relevant features but their criterion applies only when the dimension of the feature space is smaller than n, in fact, they require $p/n^\xi < 1$ for some $0 < \xi < 1$. In this paper, we investigate the property of the EBIC when the number of relevant features diverges at the order $O(n^c)$ for some $0 < c < 1$ and $p = O(n^\kappa)$ for any κ or $p = O(\exp(n^\kappa))$ for some $0 < \kappa < 1$. We identify the conditions under which the EBIC remains selection consistent and provide the theoretical proof (Theorem 1). Since the seminal paper on LASSO [17], penalized likelihood methods with various penalty functions have been largely used for model selection, see, e.g., [9, 8, 22]. It has been shown that if the penalty parameter in the penalized likelihood is properly chosen the penalized likelihood methods are

4

selection consistent under certain conditions, see [9, 11, 23, 8, 18, 12]. However, in practice, without a proper criterion for the selection of the penalty parameter (which corresponds to the selection of model), the selection consistency cannot be realized. The commonly used criterion in the penalized likelihood methods, the CV, cannot be selection consistent in small-n-large-p problems, as we have already pointed out in the previous paragraph. In this paper, we also consider the application of the EBIC for the selection of the penalty parameter in penalized likelihood methods. Simulation studies are conducted to demonstrate the finite sample properties of the EBIC and the selection procedures.

The remainder of the paper is arranged as follows. In §2, the selection consistency of EBIC with diverging number of relevant features are established. In §3, a two-stage feature selection procedure with the application of the EBIC is described and discussed. In §4, simulation results are reported. Technical details and proofs are provided in the Appendix.

2 Selection consistency of EBIC with diverging number of relevant features

We denote by p_n the number of features under investigation to make its dependence on n explicit. Let $(y_i, x_{i1}, \ldots, x_{ip_n}), i = 1, \ldots, n$, be independent observations. We consider the following linear model

$$y_i = \sum_{j=1}^{p_n} \beta_{nj} x_{ij} + \epsilon_i, \ \ i = 1, \ldots, n, \tag{1}$$

where ϵ_i's are i.i.d. with mean zero and variance σ^2. In matrix notation, (1) is expressed as

$$\boldsymbol{y}_n = X_n \boldsymbol{\beta}_n + \boldsymbol{\epsilon}_n,$$

where $\boldsymbol{\beta}_n = (\beta_{n1}, \ldots, \beta_{np_n})^T$, $\boldsymbol{y}_n = (y_1, \ldots, y_n)^T$ and $X_n = (x_{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,p_n}}$. Here p_n is either of a polynomial order or an exponential order of n, and $\boldsymbol{\beta}_n$ is sparse, meaning that only a few of its components are non-zero.

We first introduce some notations. Let $s_{0n} = \{j : \beta_{nj} \neq 0, j \in \{1, \ldots, p_n\}\}$. Let s be any subset of $\{1, \ldots, p_n\}$. For convenience, we also refer to s as a submodel. We denote by $X_n(s)$ the matrix composed of the columns of X_n with indices in s. Similarly, $\boldsymbol{\beta}_n(s)$ denotes the vector consisting of components of $\boldsymbol{\beta}_n$ with indices in s. Let $\nu(s)$ denote the number of components in s. Let $p_{0n} = \nu(s_{0n})$. Let $H_n(s)$ be the projection matrix of $X_n(s)$, i.e., $H_n(s) = X_n(s)[X_n(s)^T X_n(s)]^{-1} X_n(s)^T$. Define

$$\Delta_n(s) = \|\mu_n - H_n(s)\mu_n\|_2^2,$$

where $\mu_n = E\boldsymbol{y}_n = X_n(s_{0n})\boldsymbol{\beta}_n(s_{0n})$ and $\|\cdot\|_2$ is the L_2 norm.

Let $\mathcal{S}_j$ be the set of all combinations of j indices in $\{1, \ldots, p_n\}$. Interchangeably we also call $\mathcal{S}_j$ the class of submodels consisting of j features. Let $\tau(\mathcal{S}_j)$ be the size of $\mathcal{S}_j$; that is, $\tau(\mathcal{S}_j) = \binom{p_n}{j}$.

The family of EBIC proposed in [4] under model (1) is defined as

$$\text{EBIC}_\gamma(s) = n \ln \left(\frac{\|\boldsymbol{y}_n - H_n(s)\boldsymbol{y}_n\|_2^2}{n} \right) + \nu(s) \ln n + 2\gamma \ln \tau(\mathcal{S}_j), \quad s \in \mathcal{S}_j, \gamma \geq 0.$$

The family of EBIC is motivated from a Bayesian framework which gives rise to the BIC. The BIC of a model s is an approximation to the minus 2 log-transform of the posterior probability of s while the prior probability on each model is equal. With the equal prior probabilities, the prior probability on the submodel class $\mathcal{S}_j$ is proportional to its size $\tau(\mathcal{S}_j)$. This makes BIC favor models with larger number of features in small-n-large-p problems. Instead of imposing an equal prior probability on each model, the EBIC imposes different prior probabilities on models in different submodel classes such that the prior probability on $\mathcal{S}_j$ is proportional to $\tau(\mathcal{S}_j)^{-\gamma}$.

6

The parameter γ is determined such that the resultant EBIC is selection consistent. In the case of high dimensional feature space, i.e., $p_n = O(n^\kappa)$ for any $\kappa > 0$, and a fixed number of relevant features, [4] showed that if $\gamma > 1 - 1/(2\kappa)$ the EBIC is selection consistent. In the following, we deal with the case that the number of relevant features diverges and the feature space is high or ultra-high dimensional. First we consider the following condition:

Consistency Condition:

$$\lim_{n \to \infty} \min\{\frac{\Delta_n(s)}{p_{0n} \ln p_n} : s_0 \not\subset s, \nu(s) \le k_n\} = \infty.$$

where $k_n = kp_{0n}$ for any fixed $k > 1$.

This condition is slightly different from what is called the asymptotic identifiability condition in [4]. The restriction $\nu(s) \le k_n$ is imposed because in practice only the models with size comparable with and smaller than the true model will be considered. Implicitly, the consistency condition requires that

$$\sqrt{\frac{n}{p_{0n} \ln p_n}} \min\{|\beta_{nj}| : j \in s_{0n}\} \to \infty. \tag{2}$$

We now discuss a relationship between the consistency condition above and the well known sparse Reisz condition which is given as follows:

$$
\begin{aligned}
0 < c_{\min} &\le \min\{\lambda_{\min}(\frac{1}{n}X_n(s)^T X_n(s)) : \nu(s) \le k_n\} \\
&\le \max\{\lambda_{\max}(\frac{1}{n}X_n(s)^T X_n(s)) : \nu(s) \le k_n\} \le c_{\max} < \infty,
\end{aligned}
$$

where $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and the largest eigenvalues respectively. If p_{0n} is fixed and hence so is $\{\beta_{nj} : j \in s_{0n}\}$ then the sparse Reisz condition implies the consistency condition as shown in [4]. If p_{0n} diverges then the sparse Reisz condition together with (2) imply the consistency condition. When the number of relevant features diverges, conditions of the type (2) are always imposed for selection consistency

7

in penalized likelihood procedures, see [23, 18, 12]. As the following proposition implies, the sparse Reisz condition together with (2) are stronger assumptions than the consistency condition.

Proposition 1. *Assume $s_{0n} = \{1, 2, \ldots, p_{0n}\}$. Let s_{-k} be the set with the kth element of s_{0n} removed. Let $k(s) = s_{-k} \cup s$. If (2) is satisfied and*

$$\lim_{n \to \infty} \min_{s:\nu(s) \leq k_n, s_0 \not\subset s} \frac{\max_k\{\|[I - H_n(k(s))]X_n(\{k\})\|\}}{p_{0n} \ln p_n} = \infty \qquad (3)$$

then the consistency condition holds.

The above proposition is similar to a result in [4] which deals with a high dimensional feature space and a fixed number of relevant features. The same as in [4], examples can be constructed such that (3) holds but the sparse Reisz condition does not hold.

Condition (2) determines the divergence pattern of (n, p_{0n}, p_n) and the constraint on β_{nj}. Now consider the high and ultra-high dimensional feature spaces separately. If $p_n = O(n^\kappa)$ for any fixed $\kappa > 0$ and $p_{0n} = n^c$ for some $0 < c < \kappa$, (2) reduces to $\frac{n^{1-c}}{\ln n} \min\{|\beta_{nj}^2| : j \in s_{0n}\} \to \infty$. The induced constraint on β_{nj} is that $\min\{|\beta_{nj}^2| : j \in s_{0n}\}$ must have a magnitude larger than $O(n^{-(1-c)})$. Let b be any number bigger than c. Then the following provides a consistency pattern: $(n, p_{0n}, p_n) = (n, O(n^c), O(n^\kappa))$, $\min\{|\beta_{nj}| : j \in s_{0n}\} = O(n^{-(1-b)/2})$, $0 < c < \kappa$, $c < b < 1$. If $p_n = O(\exp(n^\kappa))$ and $p_{0n} = n^c$ then, by the same argument, (2) induces the following consistency pattern: $(n, p_{0n}, p_n) = (n, O(n^c), O(\exp(n^\kappa)))$, $\min\{|\beta_{nj}| : j \in s_{0n}\} = O(n^{-(1-b)/2})$, $0 < c, \kappa < 1$, $c + \kappa < b < 1$.

We now state the main result on the selection consistency of the EBIC with diverging number of relevant features in high or ultra-high dimensional feature spaces.

Theorem 1. *Assume model (1) and the consistency condition. In addition, assume that $p_{0n} \ln p_n = o(n)$, $\ln p_{0n} / \ln p_n \to \delta \geq 0$. Let $k_n = k p_{0n}$ for any constant $k > 1$.*

Then

$$P\{\min_{s:\nu(s)\leq k_n} EBIC_\gamma(s) > EBIC_\gamma(s_{0n})\} \to 1,$$

if $\gamma > \frac{1+\delta}{1-\delta} - \frac{\ln n}{2(1-\delta)\ln p_n}$.

The following are immediate corollaries of Theorem 1.

Corollary 1. *If* $p_n = O(n^\kappa)$ *for any constant* $\kappa > 0$, $p_{0n} = p_0$ *is fixed, the EBIC is selection consistent with* $\gamma > 1 - \frac{\ln n}{2\ln p_n} = 1 - \frac{1}{2\kappa}$ *among all models* s *with* $\nu(s) \leq k_n$.

Corollary 2. *If* $p_n = O(n^\kappa)$ *for any constant* $\kappa > 0$, $p_{0n} = O(n^c)$, $\min\{|\beta_{nj}| : j \in s_{0n}\} = O(n^{-(1-b)/2})$, $0 < c < \kappa$, $c < b < 1$, *then the EBIC is selection consistent with* $\gamma > \frac{\kappa+c-0.5}{\kappa-c}$ *among all models* s *with* $\nu(s) \leq k_n$.

Corollary 3. *If* $p_n = O(\exp(n^\kappa))$ *for* $0 < \kappa < 1$, $p_{0n} = O(n^c)$, $\min\{|\beta_{nj}| : j \in s_{0n}\} = O(n^{-(1-b)/2})$, $0 < c, \kappa < 1$, $c + \kappa < b < 1$, *the EBIC is selection consistent with* $\gamma > 1 - \frac{\ln n}{2\ln p_n}$ *among all models* s *with* $\nu(s) \leq k_n$.

The following lemmas are needed in the proof of Theorem 1.

Lemma 1. *If* $\dfrac{\ln j}{\ln p} \to \delta$ *as* $p \to +\infty$, *we have*

$$\ln\left(\frac{p!}{j!(p-j)!}\right) = j\ln p(1-\delta)(1+o(1)).$$

Lemma 2. *Let* χ_k^2 *denote a* χ^2 *random variable with degrees of freedom* k. *If* $m \to +\infty$ *and* $\dfrac{K}{m} \to 0$ *then*

$$P(\chi_k^2 \geq m) = \frac{1}{\Gamma(k/2)}(m/2)^{k/2-1}e^{-m/2}(1+o(1)),$$

uniformly for all $k \leq K$.

The proofs of Lemmas 1 and 2 and Theorem 1 are given in the Appendix.

9

3   Application of EBIC in feature selection procedures

In this section. we consider the application of EBIC for choosing tuning parameters in feature selection procedures using penalized likelihood methods. When the dimension of the feature space is high or ultra-high, a natural first step in feature selection is to reduce the dimensionality of the feature space by some screening procedure and then to apply the penalized likelihood method with the reduced feature space. This has become a well-accepted strategy for feature selection with high or ultra-high feature space, see, e.g., [10, 20, 5]. In the following, we describe a general feature selection procedure of this nature where EBIC is used to choose the penalty parameter in the penalized likelihood.

Screening stage: Let $\mathcal{F}_n$ denote the set of all the features. This stage screens out obviously irrelevant features by a screening procedure and reduces $\mathcal{F}_n$ to a set $\mathcal{S}_n^*$ with dimension smaller than n. The screening procedure we recommend is as follows. First using the sure independence screening (SIS) advocated in [10] to reduce the dimension of $\mathcal{F}_n$ to a low power order of n, say $n^{3/2}$, then using LASSO by choosing an appropriate penalty parameter to further reduce the dimension below n.

Selection stage: Select features by optimizing a penalized log likelihood of the form

$$l_{n,\lambda}(X(\mathcal{S}_n^*), \boldsymbol{\beta}(\mathcal{S}_n^*)) = -2\ln L(X(\mathcal{S}_n^*), \boldsymbol{\beta}(\mathcal{S}_n^*)) + \sum_{j \in \mathcal{S}_n^*} p_\lambda(|\beta_j|),$$

where $L(X(\mathcal{S}_n^*), \boldsymbol{\beta}(\mathcal{S}_n^*))$ is the likelihood function of the model with all features in $\mathcal{S}_n^*$, $p_\lambda(\cdot)$ is a penalty function and λ is the penalty parameter. An appropriate penalty function to use is the SCAD penalty proposed in [9]. The λ is chosen

by EBIC as follows. For each λ, let $s_{n\lambda}$ be the set of features with non-zero coefficient when $l_{n,\lambda}(X(\mathcal{S}_n^*), \boldsymbol{\beta}(\mathcal{S}_n^*))$ is minimized. Compute

$$\text{EBIC}_\gamma(\lambda) = -2 \ln L(X(s_{n\lambda}), \hat{\boldsymbol{\beta}}(s_{n\lambda})) + \nu(s_{n\lambda}) \ln n + 2\gamma \ln \begin{pmatrix} p_n \\ \nu(s_{n\lambda}) \end{pmatrix},$$

where $\hat{\boldsymbol{\beta}}(s_{n\lambda})$ is the maximum likelihood estimate (without penalty) of $\boldsymbol{\beta}(s_{n\lambda})$ and γ is taken to be $1 - \frac{\ln n}{C \ln p_n}$ for some $C > 2$. Let λ^* be the one which attains the minimum EBIC_γ. The set $s_{n\lambda^*}$ is taken as the set of selected features.

We shortly discuss the properties of the above feature selection procedure in the following. For a screening procedure, if $P(\mathcal{S}_n^* \subset \mathcal{F}_n) \to 1$, as n goes to infinity, the screening procedure is said to have the property of sure screening, see [10]. For a penalized likelihood function of the above type, if there is λ_n such that $P(s_{n\lambda_n} = s_{0n}) \to 1$, the penalized likelihood is said to have an oracle property (in a narrower sense). If the screening procedure in the screening stage has the property of sure screening, the reduced feature space $\mathcal{S}_n^*$ will contain all the relevant features with probability converging to 1 as n goes to infinity. If the penalized likelihood has the oracle property with the reduced feature space, there will be a λ value such that its corresponding set $s_{n\lambda}$ is the same as s_{0n}, the true set of relevant features, in the selection stage when $\mathcal{S}_n^*$ contains all the relevant features. Then the selection consistency of EBIC will guarantee that the true set of relevant feature is selected. Thus the feature selection procedure will be selection consistent if the conditions required by the sure screening property of the screening procedure, the oracle property of the penalized likelihood and the selection consistency of EBIC are met simultaneously.

Fan and Lv [10] showed that, under certain conditions (conditions 1-4 in section 5 of their paper), the SIS has the sure screening property if the dimension of the feature space is reduced to an order $O(n^{1-\theta})$ for some $\theta > 0$. If the tuning parameter in

11

LASSO is chosen such that the number of non-zero coefficients is large enough (smaller than n), the LASSO procedure can retain all the true features almost surely as n goes to infinity, see [5]. Kim *et al.* [12] considered the SCAD with diverging number of relevant features under the following conditions. C1: There are $0 \leq c < b \leq 1$ and $M_1 > 0$ such that $p_{0n} = O(n^c)$ and $n^{(1-b)/2}\min\{|\beta_{nj}| : j \in s_{0n}\} \geq M_1$. C2: There exists $M_2 > 0$ such that $n^{-1}(X_n(\{j\})^T X_n(\{j\}) \leq M_2$, for any j. C3: There exists $M_3 > 0$ such that $\lambda_{\min}(n^{-1}X_n(s_{n0})^T X_n(s_{n0})) \geq M_3$, where $\lambda_{\min}$ denotes the smallest eigenvalue. C4: $p_n \leq n$ and the eigenvalues of $n^{-1}X_n^T X_n$ are uniformly bounded from both below and above. They showed that under the above conditions the oracle property of the SCAD holds. The condition $n^{(1-b)/2}\min\{|\beta_j| : j \in s_{0n}\} \geq M_1$ implies that $n^{(1-c-\kappa)/2}\min\{|\beta_j| : j \in s_{0n}\} \to \infty$ for some $\kappa < b - c$. If C4 is replaced by C4$'$: $\tau(\mathcal{S}_n^*) < n$ and the eigenvalues of $n^{-1}X_n(s_n)^T X_n(s_n)$ for any $s_n \subset \mathcal{S}_n^*$ are uniformly bounded from both below and above, then together with C1-C3 the oracle property of the SCAD penalized likelihood in the selection stage will be guaranteed. Therefore, suppose that conditions 1-4 in [10], C1-C3, C4$'$ and the consistency condition hold, then the two-stage procedure described above is selection consistent. The reason we recommend a two-step screening procedure is that if only SIS is used to reduce the dimensionality below n condition C4$'$ might not hold because SIS does not reduce the level of the spurious correlations in the original feature space. On the other hand, LASSO does reduce the level of the spurious correlations since it tends to select only one of the highly correlated features, see [24], but due to the capacity of the computing facilities it might not be able to handle ultra-high dimensional feature space. When the two steps are combined it is more likely that C4$'$ will be satisfied while the sure screening property is retained. In fact, the conditions in [10] for the sure screening property can be much relaxed when the dimensionality is only reduced to a power

order of n higher than n. The performance of the feature selection procedure described in this section is investigated in simulation studies which are presented in the next section.

4   Simulation studies

The purpose of the simulation studies is to investigate the applicability of EBIC in feature selection procedures and to investigate whether or not the asymptotic property of selection consistency can be realized in finite sample situations. To this end, the two-stage feature selection procedure discussed in §3 is considered in the simulation studies. The R package `plus` [21] is used for the computation. We are mainly concerned about the selection consistency of the EBIC in the consistent range of γ. We take γ slightly bigger than $1 - \frac{\ln n}{2 \ln p_n}$ (in the simulation we take $\gamma = 1 - \frac{\ln n}{4 \ln p_n}$) for demonstrating the performance of the EBIC in finite sample situations. We also consider $\gamma = 0$, which corresponds to the original BIC, and $\gamma = 1$, which corresponds to an asymptotic form of the mBIC proposed in [2]. Throughout the simulation studies, $\tau(\mathcal{S}_n^*)$ is taken to be $0.5n$.

We take the divergence pattern as $(n, p_{0n}, p_n) = (n, c[n^{0.325}], [\exp(n^{0.35})])$ for $n = 100, 200, 500$ and $1,000$, and $c = 1$ and 2, which results in the table below:

n	100	200	500	1,000
p_n	150	595	6,655	74,622
$p_{0n}(c=1)$	4	6	8	9
$p_{0n}(c=2)$	8	12	16	18

For $j \in s_{0n}$ the parameter β_{nj} is independently generated as $\beta_{nj} = (-1)^u (n^{-0.1625} + |z|)$ where $u \sim Bernoulli(0.4)$ and z is a normal random variable with mean 0 and satisfies $P(|z| \geq 0.1) = 0.25$. This ensures, roughly, $\min\{|\beta_{nj}| : j \in s_{0n}\} = O(n^{-0.1625})$. The error variance σ^2 is determined by setting the following ratio to certain values

13

when $n = 100$ and kept unchanged for other n's:

$$h = \frac{E(\boldsymbol{\beta}^{*T}\Sigma\boldsymbol{\beta}^{*})}{E(\boldsymbol{\beta}^{*T}\Sigma\boldsymbol{\beta}^{*}) + \sigma^2},$$

where Σ is the covariance matrix of the predictors and the expectation is with respect to the generating distribution of $\boldsymbol{\beta}^{*}$. This ratio mimics what is called the heritability in broad sense in genetic studies. We considered $h = 0.4, 0.6$ and 0.8. For each simulation setting, 200 data sets are generated and analyzed. The following three correlation structures are considered for the covariates:

Structure I: Power decay correlation. The covariates are generated as a series of normally distributed random variables with mean 0 and correlation coefficient $\rho_{ij} = 0.5^{|i-j|}$.

Structure II: Diagonal block design with equal pairwise correlation. The covariance matrix is a diagonal block matrix. Each block except the last one is of dimension 50×50. The variances in the blocks are all equal to 1 and the off-diagonal correlations are all equal to $\rho = 0.5$.

Structure III: Diagonal block design with uniformly distributed eigenvalues. Unlike the diagonal block matrix in Structure 2, each block is first generated such that its smallest eigenvalue is 1, largest eigenvalue is 50 and other eigenvalues are uniformly distributed over $(1, 50)$, and then it is converted into a correlation matrix.

The finite sample performance of the EBIC is assessed by the positive discovery rate (PDR) and false discovery rate (FDR) defined as follows:

$$\text{PDR}_n = \frac{\nu(s_{n\lambda^*} \cap s_{0n})}{\nu(s_{0n})}, \quad \text{FDR}_n = \frac{\nu(s_{n\lambda^*} \backslash s_{0n})}{\nu(s_{n\lambda^*})},$$

where $s_{n\lambda^*}$ is the set of features selected in the selection stage of the two-stage procedure. The asymptotic property of selection consistency is equivalent to

$$\lim_{n\to\infty} \text{PDR}_n = 1 \quad \text{and} \quad \lim_{n\to\infty} \text{FDR}_n = 0,$$

14

in probability.

The simulated PDR_n and FDR_n averaged over 200 replicates for each setting are reported in Table 1, 2 and 3 respectively for correlation stucture I, II and III. In the tables, $\gamma_{\mathrm{BIC}} = 0$ corresponding to BIC, $\gamma_{\mathrm{SC}} = 1 - \ln n/(4 \ln p_n)$ corresponding to a value in the selection consistent range of γ and $\gamma_{\mathrm{mBIC}} = 1$ corresponding to mBIC.

The following points are manifest in Tables 1, 2 and 3. (i) The finite sample performance of the EBIC closely matches its asymptotic property. That is, under all the three correlation structures, for the procedure with $\mathrm{EBIC}_{\gamma_{\mathrm{SC}}}$, the PDR_n and the FDR_n approach rapidly to 1 and 0 respectively, as n increases from 100 to 1000, at all the three h levels. (ii) The BIC does not appear to be selection consistent. Under all the settings, the FDR_n of the procedure with BIC does not reduce as n increases, it is in fact quite the opposite. (iii) In general, the PDR_n of the procedure with BIC is higher because it always selects much more features. But, as n gets large, the PDR_n of $\mathrm{EBIC}_{\gamma_{\mathrm{SC}}}$ quickly becomes comparable with that of the BIC. (iv) For large n, the mBIC is comparable with $\mathrm{EBIC}_{\gamma_{\mathrm{SC}}}$, which reflects the fact that it is also selection consistent since $\gamma_{\mathrm{mBIC}} = 1$ is in the consistency range of EBIC. But for small n, it loses certain power while overly controlling FDR_n.

A   Appendix

A.1   Proof of Lemma 1:

Proof. Write

$$\frac{p!}{j!(p-j)!} = \frac{p(p-1)\cdots(p-j+1)}{j!} = \frac{p^j \left(1 - \frac{1}{p}\right)\cdots\left(1 - \frac{j-1}{p}\right)}{j!}.$$

Note that

$$\left(1 - \frac{j-1}{p}\right)^{j-1} < \left(1 - \frac{1}{p}\right)\cdots\left(1 - \frac{j-1}{p}\right) < \left(1 - \frac{1}{p}\right)^{j-1},$$

15

and, see [13], that

$$\sqrt{2\pi}j^{j+1/2}e^{-j+1/(12j+1)} < j! < \sqrt{2\pi}j^{j+1/2}e^{-j+1/(12j)}.$$

We now have

$$
\begin{aligned}
\ln(\frac{p!}{j!(p-j)!}) &\leq j\ln p + (j-1)\ln(1-1/p) - (j+1/2)\ln j + j - \frac{1}{12j+1} - \ln\sqrt{2\pi} \\
&\leq j\ln p - (j+1/2)\ln j + j = j\ln p[1 - \frac{(j+1/2)\ln j}{j\ln p} + \frac{1}{\ln p}] \\
&= j\ln p(1-\delta)(1+o(1)).
\end{aligned}
$$

(4)

and

$$
\begin{aligned}
\ln(\frac{p!}{j!(p-j)!}) &\geq j\ln p + (j-1)\ln(1-\frac{j-1}{p}) \\
&\quad - (j+1/2)\ln j + j - \frac{1}{12j} - \ln\sqrt{2\pi} \\
&\geq j\ln p + (j-1)\ln\left(1 - \frac{j-1}{p}\right) - (j+1/2)\ln j - \ln\sqrt{2\pi} \\
&= j\ln p\left(1 + \frac{(j-1)\ln\left(1-\frac{j-1}{p}\right)}{j\ln p} - \frac{(j+1/2)\ln j}{j\ln p} - \frac{\ln\sqrt{2\pi}}{j\ln p}\right) \\
&= j\ln p(1-\delta)(1+o(1)).
\end{aligned}
$$

(5)

Lemma1 follows from (4) and (5). □

A.2   Proof of Lemma 2

Proof. Denote $\bar{F}_k(m) = P(\chi_k^2 \geq m)$. By integration by parts, we obtain

$$\bar{F}_k(m) = \frac{1}{2^{k/2}\Gamma(k/2)}\int_m^{+\infty} x^{k/2-1}e^{-x/2}dx = \frac{1}{\Gamma(k/2)}(m/2)^{k/2-1}e^{-m/2} + \tilde{F}_{k-2}(m).$$

If k is even,

$$\bar{F}_k(m) = \frac{1}{\Gamma(k/2)}(m/2)^{k/2-1}e^{-m/2}[1 + \sum_{i=1}^{(k-2)/2}(\frac{(k/2-1)\cdots(k/2-i)}{(m/2)^i})].$$

16

If k is odd,

$$\bar{F}_k(m) = \frac{1}{\Gamma(k/2)}(m/2)^{k/2-1}e^{-m/2}[1 + \sum_{i=1}^{(k-3)/2}(\frac{(k/2-1)\cdots(k/2-i)}{(m/2)^i})] + \bar{F}_1(m),$$

where

$$\bar{F}_1(m) = P(\chi_1^2 \geq m) \approx 2\frac{\exp(-m/2)}{\sqrt{2\pi m}} = \frac{1}{\Gamma(k/2)}(m/2)^{k/2-1}e^{-m/2}\frac{2\Gamma(k/2)}{\sqrt{2\pi}(m/2)^{(k-1)/2}}$$

when $m \to +\infty$. We can write

$$\bar{F}_k(m) = \frac{1}{\Gamma(k/2)}(m/2)^{k/2-1}e^{-m/2}[1 + R(k,m)].$$

It is straightforward to see that $R(k,m) \leq R(K,m) \to 0$ when $m \to +\infty$. $\qquad \square$

A.3 Proof of Theorem 1

Proof. Let s be any submodel. Decompose $\text{EBIC}_\gamma(s) - \text{EBIC}_\gamma(s_{0n})$ as follows:

$$
\begin{aligned}
&\text{EBIC}_\gamma(s) - \text{EBIC}_\gamma(s_{0n}) \\
=&n \ln \frac{y_n^T\{I_n - H_n(s)\}y_n}{y_n^T\{I_n - H_n(s_{0n})\}y_n} + (\nu(s) - p_{0n})\ln n + 2\gamma(\ln \tau(\mathcal{S}_j) - \ln \tau(\mathcal{S}_{p_{0n}})) \quad (6) \\
=&T_1 + T_2, \text{ say,}
\end{aligned}
$$

where

$$
\begin{aligned}
T_1 =&n \ln \frac{y_n^T[I_n - H_n(s)]y_n}{y_n^T[I_n - H_n(s_{0n})]y_n} = n \ln \frac{y_n^T[I_n - H_n(s)]y_n}{\epsilon_n^T[I_n - H_n(s_{0n})]\epsilon_n} \\
=&n \ln \left\{1 + \frac{y_n^T[I_n - H_n(s)]y_n - \epsilon_n^T[I_n - H_n(s_{0n})]\epsilon_n}{\epsilon_n^T[I_n - H_n(s_{0n})]\epsilon_n}\right\} \quad (7) \\
T_2 =&(\nu(s) - p_{0n})\ln n + 2\gamma(\ln \tau(\mathcal{S}_{\nu(s)}) - \ln \tau(\mathcal{S}_{p_{0n}})).
\end{aligned}
$$

Case I: $s_{0n} \not\subset s$.

Without loss of generality, assume $\sigma^2 = 1$. We can write

$$\epsilon_n^T\{I_n - H_n(s_{0n})\}\epsilon_n = \sum_{i=1}^{n-p_{0n}} Z_i^2 = (n - p_{0n})(1 + o_p(1)) = n(1 + o_p(1)), \quad (8)$$

17

where Z_i's are i.i.d. standard normal variables, since $H_n(s_{0n})$ is a projection matrix with rank p_{0n}. We have

$$y_n^T[I_n - H_n(s)]y_n - \epsilon_n^T[I_n - H_n(s_{0n})]\epsilon_n$$

$$= \Delta_n(s) + 2\mu_n^T[I_n - H_n(s)]\epsilon_n + \epsilon_n^T H_n(s_{0n})\epsilon_n - \epsilon_n^T H_n(s)\epsilon_n.$$

It is trivial that

$$\epsilon_n^T H_n(s_{0n})\epsilon_n = p_{0n}(1 + o_p(1)). \quad \text{(I)}$$

We will show

$$\max\{\epsilon_n^T H_n(s)\epsilon_n, \nu(s) \le k_n\} = O_p(k_n \ln p_n), \quad \text{(II)}$$

and

$$|\mu_n^T[I_n - H_n(s)]\epsilon_n| = \sqrt{\Delta_n(s)O_p(k_n \ln p_n)}, \quad \text{(III)}$$

uniformly for all s with $\nu(s) \le k_n$. Under the assumption of the theorem, $2k_n \ln p_n = o(n)$. Then, by the asymptotic identifiability condition, (I), (II) and (III) imply that

$$y_n^T[I_n - H_n(s)]y_n - \epsilon_n^T[I_n - H_n(s_{0n})]\epsilon_n = \Delta_n(s)(1 + o_p(1)), \tag{9}$$

uniformly for all s with $\nu(s) \le k_n$. It then follows from (8) and (9) that

$$T_1 = n \ln \left(1 + \frac{\Delta_n(s)}{n}(1 + o_p(1))\right), \tag{10}$$

uniformly for all s with $\nu(s) \le k_n$.

We now prove (II) and (III) in the following. Let $m = 2k_n[\ln p_n + \ln(k_n \ln p_n)]$. It is obvious that $\dfrac{k_n}{m} \to 0$. Note that we can express $\epsilon_n^T H_n(s)\epsilon_n = \chi_j^2(s)$ where $j = \nu(s)$. By the Bonferroni inequality, we have

$$P(\max\{\epsilon_n^T H_n(s)\epsilon_n : \nu(s) \le k_n\} \ge m)$$

$$= P(\max\{\chi_j^2(s) : s \in \mathcal{S}_j, j \le k_n\} \ge m) \le \sum_{j=1}^{k_n} \tau(\mathcal{S}_j)P(\chi_j^2 \ge m).$$

18

By the fact that $\tau(\mathcal{S}_j) = \binom{p_n}{j} \leq p_n^j$ and Lemma 2, there is some c close to 1, not depending on j for $j \leq k_n$, such that

$$\tau(\mathcal{S}_j)P(\chi_j^2 \geq m) \approx c\frac{1}{2^{j/2-1}\Gamma(j/2)}\frac{\tau(\mathcal{S}_j)}{p_n^{k_n}}(k_n \ln p_n)^{-k_n}m^{j/2-1}$$

$$\leq \frac{c}{m}(k_n \ln p_n)^{-j}m^{j/2} = \frac{c}{m}\left[\sqrt{\frac{m}{(k_n \ln p_n)^2}}\right]^j = \frac{c}{m}q_n^j, \text{ say,}$$

where

$$q_n = \sqrt{\frac{m}{(k_n \ln p_n)^2}} = \sqrt{\frac{2[k_n \ln p_n + k_n \ln(k_n \ln p_n)]}{(k_n \ln p_n)^2}}(1 + o(1)) \leq q,$$

for some q between 0 and 1, when n is large enough, since $q_n \to 0$. Thus

$$P(\max\{\epsilon_n^T H_n(s)\epsilon_n : \nu(s) \leq k_n\} \geq m) \leq \frac{c}{m}\sum_{j=1}^{k_n} q^j \leq \frac{c}{m}\frac{q}{1-q} \to 0; \qquad (11)$$

that is,

$$\max\{\epsilon_n^T H_n(s)\epsilon_n : \nu(s) \leq k_n\} = m(1 + o_p(1)) = O_p(k_n \ln p_n),$$

which establishes (II).

For verifying (III), note that we can express

$$\mu_n^T\{I_n - H_n(s)\}\epsilon_n = \sqrt{\Delta_n(s)}Z(s),$$

where $Z(s) \sim N(0, 1)$. For any s with $\nu(s) \leq k_n$, we have

$$|\mu_n^T\{I_n - H_n(s)\}\epsilon_n| \leq \sqrt{\Delta_n(s)}\max\{|Z(s)| : \nu(s) \leq k_n\}.$$

Let m be the same as above. Consider $P(\max\{|Z(s)| : \nu(s) \leq k_n\} \geq \sqrt{m})$. We have

$$P(\max\{|Z(s)| : \nu(s) \leq k_n\} \geq \sqrt{m}) = P(\max\{|Z(s)| : s \in \mathcal{S}_j, j \leq k_n\} \geq \sqrt{m})$$

$$\leq \sum_{j=1}^{k_n} \tau(\mathcal{S}_j)P(Z(s) \geq \sqrt{m}) = \sum_{j=1}^{k_n} \tau(\mathcal{S}_j)P(\chi_1^2 \geq m)$$

$$\leq \sum_{j=1}^{k_n} \tau(\mathcal{S}_j)P(\chi_j^2 \geq m),$$

19

since $P(\chi_1^2 \geq m) < P(\chi_j^2 \geq m)$ by Lemma 2. We have already shown that the last sum converges to zero. This establishes (III).

Now, putting (6), (7) and (10) together, we have

$$
\begin{aligned}
& \mathrm{EBIC}_\gamma(s) - \mathrm{EBIC}_\gamma(s_{0n}) \\
={} & n \ln\left(1 + \frac{\Delta_n(s)}{n}(1 + o_p(1))\right) + (\nu(s) - p_{0n}) \ln n + 2\gamma(\ln\tau(\mathcal{S}_{\nu(s)}) - \ln\tau(\mathcal{S}_{p_{0n}})) \\
\geq{} & n \ln[1 + C p_{0n} \ln p_n / n(1 + o_p(1))] - p_{0n}(\ln n + 2\gamma \ln p_n),
\end{aligned}
$$

for some positive C, when n is large enough, by the consistency condition. Then by choosing $C > 1 + 2\gamma$, the difference goes to infinity as $n \to \infty$.

Case II: $s_{0n} \subset s$.

When $s_0 \subset s$, $\{I_n - H_n(s)\}X_n(s_0) = 0$. Hence, $y_n^T\{I_n - H_n(s)\}y_n = \epsilon_n^T\{I_n - H_n(s)\}\epsilon_n$ and

$$
\epsilon_n^T\{I_n - H_n(s_0)\}\epsilon_n - \epsilon_n^T\{I_n - H_n(s)\}\epsilon_n = \epsilon_n^T\{H_n(s) - H_n(s_0)\}\epsilon_n = \chi_j^2(s),
$$

where $\chi_j^2(s)$ is a χ^2 random variable depending on s with degrees of freedom j and $j = \nu(s) - p_{0n}$. We obtain that

$$
\begin{aligned}
n\log\left(\frac{\epsilon_n^T\{I - H_n(s_0)\}\epsilon_n}{\epsilon_n^T\{I - H_n(s)\}\epsilon_n}\right) ={} & n\log\left\{1 + \frac{\chi_j^2(s)}{\epsilon_n^T\{I - H_n(s_0)\}\epsilon_n - \chi_j^2(s)}\right\} \\
\leq{} & \frac{n\chi_j^2(s)}{\epsilon_n^T\{I - H_n(s_0)\}\epsilon_n - \chi_j^2(s)}.
\end{aligned}
\tag{12}
$$

As $n \to \infty$, $n^{-1}\epsilon_n^T\{I - H_n(s_0)\}\epsilon_n \to \sigma^2 = 1$, i.e.,

$$
\epsilon_n^T\{I - H_n(s_0)\}\epsilon_n = n(1 + o(1)).
\tag{13}
$$

Let $\tilde{\mathcal{S}}_j = \{s : s \in \mathcal{S}_{j+p_{0n}}, s_0 \subset s\}$. Note that $\tau(\tilde{\mathcal{S}}_j) = \binom{p_n - p_{0n}}{j} \leq p_n^j$. Let $m_j =$

20

$2j[\ln p_n + \ln(j \ln p_n)]$. In the same way as we derive (11), we have

$$P\left(\max_{1 \le j \le k_n - p_{0n}} \frac{\max\{\chi_j^2(s) : s \in \tilde{\mathcal{S}}_j\}}{m_j} \ge 1\right) \le \sum_{j=1}^{k_n - p_{0n}} P(\max\{\chi_j^2(s) : s \in \tilde{\mathcal{S}}_j\} \ge m_j)$$

$$\le \sum_{j=1}^{k_n - p_{0n}} \tau(\tilde{\mathcal{S}}_j) P(\chi_j^2 \ge m_j) \le \frac{1}{\ln p_n} \sum_{j=1}^{k_n - p_{0n}} q_j^j \to 0,$$

where

$$q_j = \sqrt{\frac{2}{j \ln p_n} + \frac{2 \ln(j \ln p_n)}{j(\ln p_n)^2}} \le \sqrt{\frac{2}{\ln p_n}} (1 + o(1)) \to 0.$$

Thus,

$$\max\{\chi_j^2(s) : s \in \mathcal{S}_{j+p_{0n}}, s_0 \subset s\} = m_j\{1 + o_p(1)\}, \tag{14}$$

uniformly for all s with $\nu(s) \le k_n$ and $s_0 \subset s$.

It follows from (12), (13) and (14) that

$$n \log\left(\frac{\epsilon_n^T \{I - H_n(s_0)\}\epsilon_n}{\epsilon_n^T \{I - H_n(s)\}\epsilon_n}\right) \le \frac{nm_j}{[n - m_j(1 + o_p(1))]}$$

$$\le m_j(1 + o_p(1)) \le 2j(1 + \delta) \ln p_n(1 + o_p(1)),$$

uniformly for all s with $\nu(s) \le k_n$ and $s_0 \subset s$, noting that $m_j \le 2j[\ln p_n + \ln((k_n - p_{0n}) \ln p_n)] = 2j(1 + \delta) \ln p_n(1 + o_p(1))$ and $m_j = 2j(1 + \delta) \ln p_n(1 + o_p(1))$ for $j = k_n - p_{0n}$. Thus

$$T_1 \ge -2j(1 + \delta) \ln p_n(1 + o_p(1)).$$

When $p_{0n} \le \nu(s) \le k_n$ we have $\ln \nu(s)/\ln p_n \to \delta$ uniformly, hence, by Lemma 1,

$$T_2 = j \ln n + 2\gamma(1 - \delta)j \ln p_n(1 + o(1)).$$

Finally we have

$$\text{EBIC}_\gamma(s) - \text{EBIC}_\gamma(s_{0n})$$

$$\ge j \ln n + 2\gamma(1 - \delta)j \ln p_n(1 + o(1)) - 2j(1 + \delta) \ln p_n(1 + o_p(1)) > 0,$$

uniformly for all s with $\nu(s) \le k_n$ and $s_0 \subset s$, if n is big enough, when $\gamma > \frac{1+\delta}{1-\delta} - \frac{\ln n}{2(1-\delta) \ln p_n}$.

$\square$

21

References

[1] AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, B.N. Petrox and F. Caski (eds.). Budapest: Akademiai Kiado, page 267.

[2] BOGDAN, M., DOERGE, R., and GHOSH, J. K. (2004). Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics* **167**, 989-99.

[3] BROMAN, K. W. and SPEED, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *J. Roy. Statist. Soc. B* **64**, 641-656.

[4] CHEN, J. and CHEN, Z. (2008). Extended Bayesian information criterion for model selection with large model space. *Biometrika*, **94**, 759-771.

[5] CHEN, Z. and CHEN, J. (2009). Tournament Screening cum EBIC for Feature Selection with High Dimensional Feature Spaces. *Science in China, Series A*, **52**,1327-1341.

[6] CRAVEN, P. and WAHBA, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematika* **31**, 377-403.

[7] DONOHO, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. Aide-Memoire of a Lecture at AMS Conference on Math Challenges of the 21 st Century.

[8] MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37**, 246270.

[9] FAN, J. and LI, R. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96**, 1348-1360.

[10] FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Statist. Soc.*B **70**, 849-911.

[11] FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928-961.

[12] KIM, Y., CHOI, H. and OH, H. (2008). Smoothly clipped absolute deviation on high dimensions. *J. Amer. Statist. Ass.* **103**, 1665-1673.

[13] ROBBINS, H. (1955) A Remark on Stirling's Formula. *The American Mathematical Monthly*, **62**, 26-29.

[14] SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

[15] SIEGMUND, D. (2004). Model selection in irregular problems: Application to mapping quantitative trait loci. *Biometrika* **91**, 785-800.

[16] STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Statist. Soc. B* **39**, 111-147.

[17] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the LASSO. *J. Roy. Statist. Soc. B* **58**, 267-288.

[18] WAINWRIGHT, M. (2006). Sharp thresholds for high-dimensional and noisy recovery of sparsity. Available at arXiv:math. ST/0605740.

[19] WANG, H., LI, B. and LENG, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Statist. Soc. B.* **71**, 671-683.

[20] WASSERMAN, L. and BOEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37**, 2178-2201.

[21] ZHANG, C-H. (2007). Penalize Linear Unbiased Selection. Rutgers University, Department of Statistics and Biostatistics. Technical Report. April 20, 2007.

[22] ZHANG, C-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894-942.

[23] ZHAO, P. and YU, B. (2006). On model selection consistency of LASSO. *J. Machine Learning Research* **7**, 2541-2567.

[24] ZOU, H. and HASTIE, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, **76**,301-320.

[25] ZOU, H. (2006). Adaptive Lasso and its oracle properties. *J. Amer. Statist. Ass.* **101**, 1418-1429.

Table 1: The PDR and FDR of the SIS-SCAD-EBIC procedure with Structure I (power decay correlation) averaged over 200 replicates (the numbers in parentheses are standard deviations)

		$c = 1$					
		PDR			FDR		
n	h	γ_{BIC}	γ_{SC}	γ_{mBIC}	γ_{BIC}	γ_{SC}	γ_{mBIC}
100	.4	.726(.242)	.450(.291)	.384(.288)	.571(.212)	.074(.205)	.050(.181)
	.6	.861(.187)	.700(.271)	.633(.301)	.478(.216)	.080(.170)	.044(.123)
	.8	.973(.090)	.921(.159)	.909(.176)	.363(.204)	.085(147)	.056(.120)
200	.4	.759(.205)	.532(.270)	.467(.270)	.662(.177)	.034(.101)	.017(.071)
	.6	.910(.144)	.758(.256)	.711(.282)	.574(.185)	.080(.145)	.038(.100)
	.8	.989(.056)	.957(.105)	.947(.128)	.389(.200)	.060(.115)	.045(.105)
500	.4	.826(.146)	.640(.212)	.604(.214)	.768(.100)	.037(.090)	.011(.046)
	.6	.943(.100)	.863(.164)	.836(.181)	.660(.133)	.066(.128)	.028(.079)
	.8	.994(.035)	.983(.060)	.980(.067)	.504(.190)	.027(.073)	.019(.065)
1000	.4	1.000(.00)	.999(.008)	.999(.011)	.662(.024)	.019(.041)	.009(.028)
	.6	1.000(.00)	1.000(.00)	1.000(.00)	.531(.037)	.019(.041)	.008(.026)
	.8	1.000(.00)	1.000(.00)	1.000(.00)	.470(.010)	.007(.025)	.002(.014)
		$c = 2$					
		PDR			FDR		
n	h	γ_{BIC}	γ_{SC}	γ_{mBIC}	γ_{BIC}	γ_{SC}	γ_{mBIC}
100	.4	.531(.183)	.243(.169)	.198(.162)	.507(.222)	.069(.204)	.041(.172)
	.6	.680(.166)	.416(.213)	.350(.206)	.447(.187)	.074(.173)	.026(.093)
	.8	.850(.153)	.708(.225)	.628(.248)	.373(.163)	.118(.143)	.068(.118)
200	.4	.613(.162)	.306(.164)	.260(.161)	.619(.162)	.028(.096)	.010(.066)
	.6	.720(.148)	.518(.211)	.456(.207)	.545(.181)	.036(.082)	.018(.061)
	.8	.895(.125)	.745(.199)	.703(.217)	.447(.164)	.086(.117)	.053(.096)
500	.4	.732(.130)	.425(.174)	.371(.166)	.774(.076)	.014(.054)	.004(.025)
	.6	.832(.104)	.635(.176)	.590(.186)	.695(.112)	.028(.064)	.010(.031)
	.8	.956(.067)	.875(.135)	.847(.157)	.535(.159)	.098(.121)	.068(.104)
1000	.4	.758(.108)	.537(.161)	.491(.164)	.825(.055)	.012(.040)	.005(.025)
	.6	.849(.102)	.715(.134)	.689(.144)	.761(.077)	.025(.062)	.010(.035)
	.8	.969(.054)	.925(.084)	.906(.106)	.581(.146)	.095(.110)	.072(.095)

Table 2: The PDR and FDR of the SIS-SCAD-EBIC procedure with Structure II (block with equal pairwise correlation) averaged over 200 replicates (the numbers in parentheses are standard deviations)

| | | | $c = 1$ | | | | |
| | | PDR | | | FDR | | |
n	h	γ_{BIC}	γ_{SC}	γ_{mBIC}	γ_{BIC}	γ_{SC}	γ_{mBIC}
100	.4	.733(.285)	.402(.318)	.343(.291)	.427(.268)	.229(.369)	.198(.362)
	.6	.933(.154)	.772(.297)	.703(.321)	.340(.213)	.117(.197)	.094(.207)
	.8	.996(.042)	.967(.118)	.960(.125)	.293(.203)	.053(.132)	.036(.114)
200	.4	.868(.203)	.534(.303)	.479(.306)	.442(.206)	.133(249)	.110(.246)
	.6	.994(.039)	.931(.168)	.889(.214)	.321(.173)	.107(161)	.078(.143)
	.8	1.000(.00)	.996(.031)	.994(.040)	.292(.165)	.025(.081)	.017(.070)
500	.4	.948(.093)	.754(.178)	.723(.184)	.689(.114)	.056(.107)	.049(.103)
	.6	.993(.035)	.922(.121)	.904(.132)	.626(.127)	.031(.080)	.019(.064)
	.8	1.000(.00)	.997(.024)	.992(.044)	.585(.151)	.060(.110)	.031(.083)
1000	.4	.940(.080)	.813(.158)	.785(.180)	.818(.046)	.073(.113)	.049(.092)
	.6	.995(.025)	.988(.041)	.986(.043)	.739(.066)	.039(.084)	.035(.079)
	.8	.999(.010)	.998(.017)	.996(.024)	.653(.107)	.024(.070)	.017(.061)
			$c = 2$				
		PDR			FDR		
n	h	γ_{BIC}	γ_{SC}	γ_{mBIC}	γ_{BIC}	γ_{SC}	γ_{mBIC}
100	.4	.430(.239)	.193(.174)	.173(.164)	.449(.294)	.310(.411)	.295(.408)
	.6	.684(.234)	.390(.236)	.343(.224)	.343(.220)	.164(.235)	.150(.253)
	.8	.881(.179)	.676(.266)	.603(.284)	.308(.194)	.105(.174)	.096(.175)
200	.4	.489(.206)	.199(.142)	.165(.133)	.416(.235)	.134(.275)	.115(.259)
	.6	.727(.192)	.421(.227)	.356(.214)	.351(.195)	.065(.144)	.055(.132)
	.8	.919(.135)	.718(.254)	.672(.269)	.351(.184)	.055(.099)	.043(.088)
500	.4	.664(.137)	.258(.132)	.238(.132)	.669(.145)	.031(.099)	.020(.076)
	.6	.834).127)	.468(.211)	.407(.209)	.609(.132)	.029(.073)	.014(.047)
	.8	.944(.094)	.804(.244)	.778(.266)	.485(.198)	.084(.108)	.068(.095)
1000	.4	.675(.133)	.311(.158)	.284(.158)	.830(.079)	.017(.055)	.014(.050)
	.6	.882(.134)	.551(.234)	.496(.240)	.744(.115)	.060(.108)	.033(.073)
	.8	.960(.078)	.884(.195)	.877(.202)	.616(.178)	.069(.099)	.061(.087)

Table 3: The PDR and FDR of the SIS-SCAD-EBIC procedure with Structure III (block with uniformly distributed eigenvalues) averaged over 200 replicates (the numbers in parentheses are standard deviations)

		$c = 1$					
		PDR			FDR		
n	h	γ_{BIC}	γ_{SC}	γ_{mBIC}	γ_{BIC}	γ_{SC}	γ_{mBIC}
100	.4	.915(.146)	.667(.302)	.564(.327)	.428(.191)	.041(.102)	.020(.078)
	.6	.996(.031)	.964(.116)	.950(.133)	.360(.181)	.046(.105)	.019(.063)
	.8	1.000(.00)	1.000(.00)	1.000(.00)	.326(.165)	.038(.096)	.011(.051)
200	.4	.993(.037)	.865(.206)	.811(.252)	.575(.162)	.050(.101)	.024(.073)
	.6	1.000(.00)	.999(.014)	.999(.014)	.536(.129)	.032(.081)	.013(.048)
	.8	1.000(.00)	1.000(.00)	1.000(.00)	.457(.138)	.023(.065)	.009(.042)
500	.4	1.000(.00)	.971(.081)	.961(.090)	.768(.042)	.041(.075)	.023(.055)
	.6	1.000(.00)	1.000(.00)	1.000(.00)	.704(.058)	.022(.060)	.010(.043)
	.8	1.000(.00)	1.000(.00)	1.000(.00)	.608(.091)	.016(.050)	.007(.038)
1000	.4	1.000(.00)	.999(.011)	.997(.017)	.790(.040)	.023(.046)	.008(.028)
	.6	1.000(.00)	1.000(.00)	1.000(.00)	.740(.038)	.018(.041)	.005(.021)
	.8	1.000(.00)	1.000(.00)	1.000(.00)	.705(.051)	.005(.022)	.002(.012)
		$c = 2$					
		PDR			FDR		
n	h	γ_{BIC}	γ_{SC}	γ_{mBIC}	γ_{BIC}	γ_{SC}	γ_{mBIC}
100	.4	.643(.218)	.240(.201)	.155(.179)	.409(.206)	.071(.185)	.028(.128)
	.6	.911(.141)	.589(.298)	.461(.302)	.346(.168)	.092(.163)	.045(129)
	.8	.995(.033)	.975(.100)	.964(.135)	.237(.136)	.089(.101)	.069(.092)
200	.4	.801(.147)	.307(.210)	.209(.179)	.536(.136)	.050(.142)	.013(.061)
	.6	.974(.063)	.817(.198)	.742(.236)	.443(.147)	.076(.095)	.045(.073)
	.8	.999(.010)	.993(.041)	.989(.048)	.322(.121)	.046(.074)	.034(.063)
500	.4	.933(.076)	.578(.204)	.451(.215)	.723(.079)	.035(.067)	.009(.036)
	.6	.992(.030)	.946(.073)	.930(.094)	.642(.091)	.062(.078)	.045(.069)
	.8	.999(.005)	.998(.016)	.997(.017)	.498(.105)	.023(.044)	.014(.036)
1000	.4	.970(.049)	.780(.170)	.688(.207)	.809(.051)	.042(.063)	.018(.039)
	.6	.997(.013)	.976(.054)	.973(.058)	.738(.059)	.030(.053)	.024(.042)
	.8	.999(.004)	.998(.011)	.998(.012)	.608(.085)	.011(.031)	.006(.022)