Ann Inst Stat Math (2016) 68:155–180
DOI 10.1007/s10463-014-0497-2

The EBIC and a sequential procedure for feature selection in interactive linear models with high-dimensional data

Yawei He · Zehua Chen

Received: 14 December 2013 / Revised: 8 August 2014 / Published online: 3 December 2014

Abstract High-dimensional data arises in many important scientific fields. The analysis of high-dimensional data poses great challenges to statisticians. In high-dimensional data, the relationship among the variables is complex. It involves main effects as well as interaction effects of the covariates. The effect of some covariates is only realized through their interaction with the others. This makes the consideration of interactive models imperative in the analysis of high-dimensional data. Because of the existence of high spurious correlation among the covariates in high-dimensional data, conventional tools for dealing with interactive models become inappropriate. In this paper, we develop specific tools for feature selection in high-dimensional data with interactive models, including a version of the extended BIC (EBIC) for interactive models and a sequential feature selection procedure. Main-effect and interaction features are treated differently in the EBIC for interactive models and the sequential procedure due to their different natures. The selection consistency of the EBIC for interactive models and the sequential procedure is established. Simulation studies are carried out to vindicate the asymptotic property in finite samples as well as to compare with non-sequential procedures. The approach developed in this paper is also applied to a real data set.

Keywords High-dimensional data · EBIC · Feature selection · Interactive model · Sequential procedure · Selection consistency

Y. He
Department of Applied Statistics, Chongqing Jiaotong University,
Chongqing 400074, China
e-mail: yaweih11@gmail.com

Z. Chen (✉)
Department of Statistics and Applied Probability, National University of Singapore,
Singapore 117546, Singapore
e-mail: stachenz@nus.edu.sg

1 Introduction

High-dimensional data arises from many conventional research fields such as genetic research, financial studies, web information analysis, etc. A common nature of high-dimensional data is the so-called small-n-large-p structure, that is, the number of observations is much smaller than the number of variables. In the studies mentioned above, one usually needs to establish a relationship between a particular variable called response variable and some other variables called covariates. For this purpose, one needs to select the covariates among the huge number of variables under consideration. However, most traditional methods for variable selection are no longer applicable due to the small-n-large-p structure of the high-dimensional data. The small-n-large-p structure poses both computational and theoretical problems, which makes the variable selection a challenging task, especially, when both the main and interaction effects of the variables are considered. We refer to both main effects and interaction effects as features. In this article, we deal with feature selection in interactive linear models with high-dimensional data. We consider both feature selection criteria and selection procedures.

Various classical criteria have been used in variable selection problems with a small and fixed p. These criteria include the Akaike's information criterion (AIC) (Akaike 1973), the Bayes information criterion (BIC) (Schwarz 1978), the cross-validation (CV) (Stone 1974) and generalized cross-validation (GCV) (Craven and Wahba 1978). However, in the case of high-dimensional data, these classical criteria are no longer appropriate. They are generally too liberal in the sense that they choose too many variables which are not the true covariates. This phenomenon was first observed in genetic studies by a few researchers, see Broman and Speed (2002), Siegmund (2004) and Bogdan et al. (2004), and has now become a common knowledge.

Many attempts have been made to modify the classical criteria so that they become appropriate for model selection with high-dimensional data. A few examples follow. In the context of density estimation and non-parametric regression, Yang and Barron (1998), Barron et al. (1999) and Yang (1999) considered a modification of AIC by adding an additional penalty term for model complexity. Baraud (2000) considered another modification which replaces the factor 2 in AIC by a theoretically determined constant $c(> 1)$. These authors concentrated on the prediction error and derived the risk bounds of the model selected by the modified AIC. These modified AICs could be used for feature selection. It is worthy to mention that Yang (1999) dealt with multivariate non-parametric regression and interactions of unknown orders were also considered. However, the selection consistency of the criteria was not established. The BIC has been modified in different aspects. In essence, the BIC for a model s is negative 2 times the log posterior probability of s given below:

$$p(s|\boldsymbol{Y}) = \frac{m(\boldsymbol{Y}|s)p(s)}{\sum_{\tilde{s}\in\mathcal{S}} m(\boldsymbol{Y}|\tilde{s})p(\tilde{s})},$$

where $p(s)$ is the prior probability of s and $m(\boldsymbol{Y}|s)$ is the probability density function of data $\boldsymbol{Y}$ given model s. In the derivation of the original BIC, the prior $p(s)$ is taken as a constant and $m(\boldsymbol{Y}|s)$ is approximated by a Laplace approximation. Clyde et al. (2007) and an unpublished work of Berger have focused on $m(Y|s)$ and rectified the problems caused by the Laplace approximation. However, the rectification on the Laplace approximation does not target the problems caused by the small-n-large-p structure. Bogdan et al. (2004) and Chen and Chen (2008) have focused on the modification of the prior $p(s)$. The modification obtained by Bogdan et al. (2004) is called the modified BIC (mBIC). The modification resulting from Chen and Chen (2008) is referred to as the extended BIC (EBIC). The mBIC is a single criterion, while the EBIC is indeed a family of criteria which includes the original BIC and mBIC as special cases. The properties of EBIC for feature selection in a variety of models have been comprehensively investigated, see Chen and Chen (2008), Foygel and Drton (2010), Chen and Chen (2012), Luo and Chen (2013) and Luo et al. (2014). In these papers, the selection consistency of EBIC has been established for linear, generalized linear, graphical and survival models. However, all these models are confined to main effects. The property of EBIC for feature selection in interactive models has not been touched yet.

The interaction between covariates cannot be ignored in practical problems, especially when the number of covariates is large. For example, in genetic studies, it has been found that many diseases are affected by the interaction effects of genes, see, e.g., Storey et al. (2005) and Zou and Zeng (2009). More crucially, in certain situations, the effect of covariates is realized only through their interaction. Without the consideration of interaction, the covariates are not detectable. Thus, one is obliged to consider the interactive models. An interactive model involves both main-effect and interaction features. The number of interaction features is in the square order of the number of main-effect features. The small-n-large-p structure of high-dimensional data results in high spurious correlations among the variables in the data. Due to the high spurious correlations, the effect of main-effect features might be masked by the false (or spurious) effects of certain interaction features. To explain, consider the interactive model

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{j<k} \theta_{jk} x_{ij} x_{ik} + \epsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

Suppose that covariates j and l have an interaction effect but the main-effect coefficient β_j is much smaller than the interaction coefficient θ_{jl}. If there are a lot of x_k's that are correlated spuriously with x_l in the data, then, by fitting the above interactive model, all the coefficients θ_{jk}'s corresponding to those x_k's will appear more significant than β_j. It has been reported in Zhao and Chen (2011) that, in an application to the analysis of a prostate cancer data, the approach using SCAD penalized logistic model together with the original EBIC, which treats the main-effect and interaction features equally, only selects the interaction features. Though those selected interaction features involve the SNPs selected by using main-effect models, the main-effect features corresponding to those SNPs are not selected. This motivated us to explore a specific version of EBIC suitable for interactive models. We call this specific version the EBIC for interactive models. The EBIC for interactive models imposes different penalties on the number of main-effect features and the number of interaction features. Its selection consistency for feature selection in interactive models is established in this article.

A selection criterion is usually applied together with a model selection procedure. The latter produces a sequence of candidate models and the criterion is used to select the optimal one. The most popular model selection method in high-dimensional data analysis is the penalized likelihood approach. In the context of main-effect linear models, the penalized likelihood approach amounts to:

$$\text{Minimizing}_{\beta_j, 0\le j\le p} \left\{ \left\| \boldsymbol{y} - \beta_0 \mathbf{1} - \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j \right\|_2^2 + \sum_{j>0} p_\lambda(|\beta_j|) \right\}, \tag{2}$$

where $\boldsymbol{y}$ is the vector of response values, $\boldsymbol{x}_j$ the vector of the values of the jth covariate, λ a regulating parameter, p_λ a penalty function, and $\|\cdot\|_2$ is the L_2-norm. For a fixed value of λ, the minimization of (2) yields only a certain number of non-zero β_j's which correspond to a particular model. Thus, by setting λ to a sequence of values, the minimization of (2) produces a sequence of candidate models. Various penalty functions have been used in the penalized likelihood approach. They include Lasso (Tibshirani 1996), SCAD (Fan and Li 2001), Adaptive Lasso (Zou 2006), MCP (Zhang 2010) and so on. A so-called oracle property is of major concern for penalized likelihood approaches. The property consists of two aspects: (1) selection consistency—the true features can be exactly selected with probability converging to 1, and (2) estimation consistency—the parameters can be consistently estimated the same as they would be were the true features known in advance. Lasso does not have the oracle property in general. For Lasso to possess the oracle property, a condition called irrepresentability must be satisfied. The irrepresentability condition was discovered, though by different names, in Zou (2006), Meinshausen and Bühlmann (2006) and Zhao and Yu (2006). However, the irrepresentability condition is too strong to be satisfied in practical problems. The Adaptive Lasso, which puts a weight for each term $|\beta_j|$ in the Lasso penalty, has the oracle property for fixed p when the inverse of the absolute ordinary least squares estimate of β_j is used as the weight of $|\beta_j|$. This result was established in Zou (2006). For diverging p, Huang et al. (2008) showed that Adaptive Lasso using marginal least squares estimates as the weights has the oracle property under a partial orthogonality condition. The properties of SCAD were studied in Fan and Li (2001, 2004), Fan and Peng (2004) and Xie and Huang (2009). In these papers, the oracle property of SCAD was established for various models when p is fixed or diverging to infinity. The MCP penalty has similar properties to the SCAD penalty. The asymptotic properties of the MCP penalty were studied in Zhang (2010).

The penalized likelihood approach has also been extended to the interactive models. In all the extensions, a certain hierarchical structure is imposed on the interactive model. The hierarchical structure requires that if an interaction feature is included in the model, then either at least one of or both constituent main-effect features should be also included in the model. When only at least one of the main-effect features is required, it is referred to as weak hierarchy; otherwise, it is referred to as strong hierarchy. Commonly, a hierarchical structure is imposed through certain group-Lasso penalties [Yuan and Lin (2006)], see, e.g., Zhao (2009), Yuan et al. (2009), Choi et al. (2010) and Radchenko and James (2010), etc. A different approach called hierarchical Lasso was considered in Bien et al. (2013). The hierarchical lasso imposes the hierarchical structures by adding a set of convex constraints to a slightly different version of Lasso, instead of using a group-lasso penalty. We refer to the above extensions as lasso-type methods. A common nature of the lasso-type methods is that, by imposing hierarchical structures through either the group-lasso penalties or the convex constraints, interaction features are made harder to be selected than main-effect features, see e.g., section 3 of Bien et al. (2013). However, hierarchical structures are not necessarily conditions which must be imposed so that a selection procedure has the above nature.

In this article, we develop a sequential procedure which does not impose the hierarchical structures. However, it has the same nature that an interaction feature is harder to be selected than a main-effect feature. At each step, the procedure first selects the main-effect feature most correlated with the current residual among all the main-effect features as well as the interaction feature most correlated with the current residual among all the interaction features. Then the two selected features are evaluated by the EBIC for interactive models, the one that reduces the EBIC more is finally selected at the step. As will be seen, it is the EBIC for interactive models that imposes implicitly different thresholds for main-effect and interaction features such that the procedure has the nature mentioned above. The sequential procedure is selection consistent. In addition, the procedure is computationally much simpler than the lasso-type methods.

The remainder of the article is arranged as follows. In Sect. 2, we present the EBIC for interactive models and its asymptotic properties. In Sect. 3, we describe the sequential procedure and establish its selection consistency. In Sect. 4, we report the simulation studies which demonstrate the validity of the sequential procedure and provide the analysis of a real data using the sequential procedure. Technical details and proofs are given in an appendix.

2 EBIC for interactive models and its selection consistency

Let $\{(y_i, x_{ij}) : i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, p\}$ be the observations where y_i and x_{ij} are, respectively, the value of the response variable and the value of the jth covariate observed on the ith individual. Consider the model

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{0<j<k\le p} \theta_{jk} x_{ij} x_{ik} + \epsilon_i, \quad i = 1, \ldots, n. \tag{3}$$

Assume that p is of the order $O(\exp\{n^{\kappa}\})$ for some κ in between 0 and 1, and that the ϵ_i's are i.i.d. random errors distributed as the normal distribution $N(0, \sigma^2)$. First, we introduce some notation. Let

$$\begin{aligned} s_{0\mathrm{M}} &= \{j : \beta_j \neq 0, j = 1, \ldots, p\}, \\ s_{0\mathrm{I}} &= \{(jk) : \theta_{jk} \neq 0, j = 1, \ldots, p-1, k = j+1, \ldots, p\}, \\ s_0 &= s_{0\mathrm{M}} \cup s_{0\mathrm{I}}. \end{aligned}$$

Denote by $\nu(s)$ the cardinality of a set s, i.e., the number of elements in s. Let $p_0 = \nu(s_0)$ and assume that $p_0 = O(n^c)$ for some constant c such that $0 < c < 1-\kappa$.

Implicitly, the sets $s_{0\mathrm{M}}$, $s_{0\mathrm{I}}$ and hence s_0 as well as p_0 depend on n. But, for the sake of clarity, we suppress the dependence on n in the notation. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^\tau$ and Z be the matrix whose columns are indexed by $j, j = 1, \ldots, p$, and (jk), $j = 1, \ldots, p-1, k = j+1, \ldots, p$. The column indexed by j is $(x_{1j}, \ldots, x_{nj})^\tau$ and the column indexed by (jk) is $(x_{1j}x_{1k}, \ldots, x_{nj}x_{nk})^\tau$. Let s be any subset of $\mathcal{S} = \{j : j = 1, \ldots, p\} \cup \{(jk) : j = 1, \ldots, p-1; k = j+1, \ldots, p\}$. When it is necessary, we decompose s into $s = s_\mathrm{M} \cup s_\mathrm{I}$ where s_M consists of the elements with single indices and s_I consists of the elements with double indices. By an abuse of terminology, we also refer to s as the model consisting of the features indexed by s. Denote by $Z(s)$ the sub matrix of Z consisting of the columns with indices in s. Thus, we refer to s_0 as the true model. Let $\boldsymbol{\xi}$ denote the vector of coefficients β_j's and θ_{jk}'s and $\boldsymbol{\xi}(s)$ the sub vector consists of components with indices in s. In matrix form, model (3) is expressed as

$$\boldsymbol{y} = Z\boldsymbol{\xi} + \boldsymbol{\epsilon}, \tag{4}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\tau$. Let $H(s)$ denote the projection matrix of $Z(s)$, i.e.,

$$H(s) = Z(s)[Z(s)^\tau Z(s)]^{-1} Z(s)^\tau.$$

Now, we turn to the development of the EBIC for interactive models. As mentioned in Sect. 1, the EBIC is obtained by modifying the prior probability $p(s)$ of model s in the Bayesian framework which leads to BIC. In the derivation of BIC, $p(s)$ is taken as a constant. In the development of EBIC, Chen and Chen (2008) classified the models according to the number of features they involve. Let $\mathcal{S}_j$ be the class of models consisting of j features. In the derivation of EBIC, the prior probability $p(s)$ is specified as follows. Let $\tau(\mathcal{S}_j) = \binom{p}{j}$, the size of $\mathcal{S}_j$. For $s \in \mathcal{S}_j$, $p(s) = \Pr(s|\mathcal{S}_j)\Pr(\mathcal{S}_j)$ where $\Pr(s|\mathcal{S}_j) = 1/\tau(\mathcal{S}_j)$ and $\Pr(\mathcal{S}_j)$ is proportional to $\tau^\xi(\mathcal{S}_j)$ for some ξ between 0 and 1. Thus, for $s \in \mathcal{S}_j$, $p(s) = c\tau^{-\gamma}(\mathcal{S}_j)$ where $\gamma = 1 - \xi$ and c is the normalizing constant. This prior leads to the EBIC for linear models given by

$$\mathrm{EBIC}_\gamma(s) = n \ln\left(\frac{\|[I - H_n(s)]\boldsymbol{y}\|_2^2}{n}\right) + \nu(s)\ln n + 2\gamma \ln\binom{p}{\nu(s)} \quad 0 \le \gamma \le 1.$$

Taking into account different natures of main-effect and interaction features, we classify the interactive models according to the number of main-effect features and the number of interaction features they contain. Let $\mathcal{S}_{jk}$ be the class of models that consists of j main-effect features and k interaction features. Note that $\tau(\mathcal{S}_{jk}) = \binom{p}{j}\binom{p(p-1)/2}{k}$. For $s \in \mathcal{S}_{jk}$, we modify the prior to $p(s) = c\binom{p}{j}^{-\gamma_\mathrm{M}}\binom{p(p-1)/2}{k}^{-\gamma_\mathrm{I}}$. Thus, for an interactive model $s = s_\mathrm{M} \cup s_\mathrm{I}$, we arrive at the EBIC

$$\begin{aligned}\mathrm{EBIC}_{\gamma_\mathrm{M},\gamma_\mathrm{I}}(s) = n \ln\left(\frac{\|[I - H_n(s)]\boldsymbol{y}\|_2^2}{n}\right) + \nu(s)\ln n \\ +2\gamma_\mathrm{M} \ln\binom{p}{\nu(s_\mathrm{M})} + 2\gamma_\mathrm{I} \ln\binom{p(p-1)/2}{\nu(s_\mathrm{I})}.\end{aligned} \tag{5}$$

Let $\Delta_n(s) = \|[I - H_n(s)]\boldsymbol{\mu}\|_2^2$, where $\boldsymbol{\mu} = Z(s_0)\boldsymbol{\xi}(s_0)$. The following theorem establishes the selection consistency of the EBIC for interactive model.

Theorem 1 *Assume model* (3). *Suppose that*

$$\min_s \left\{ \frac{\Delta_n(s)}{p_0 \ln p} : s_0 \nsubseteq s,\, \nu(s) \le r p_0 \right\} \to \infty, \quad \text{for some } r > 1. \tag{6}$$

In addition, assume that $p_0 \ln p = o(n)$, $\ln p_0 / \ln p \to 0$. *Then, when* $n \to \infty$,

$$P\left(\min_{s : s \neq s_0, \nu(s) \le r p_0} \mathrm{EBIC}_{\gamma_\mathrm{M}, \gamma_\mathrm{I}}(s) > \mathrm{EBIC}_{\gamma_\mathrm{M}, \gamma_\mathrm{I}}(s_0) \right) \to 1, \tag{7}$$

if $\gamma_\mathrm{M} > 1 - \frac{\ln n}{2 \ln p}$, $\gamma_\mathrm{I} > 1 - \frac{\ln n}{4 \ln p}$.

The proof of the theorem is given in the appendix. We provide some remarks on the theorem to end this section.

Theorem 1 confines the range of models to those whose size has a bound of order $O(n^c)$, the same as that of the true model. This is reasonable since, by the sparsity assumption of high-dimensional models, these are the only models of one's concern.

The following caution should be mentioned. In the case of high-dimensional data, we can always find a model s with $\nu(s) > n$ such that $\|[I - H_n(s)]\mathbf{y}\|_2^2 = 0$ and hence $\mathrm{EBIC}_{\gamma_\mathrm{M}, \gamma_\mathrm{I}}(s) = -\infty$. This implies that the EBIC always attains its minimum at a false model, if models with size larger than n are considered. This suggests that the EBIC cannot be used to assess models with size close to or larger than n. It is usually the case that, when a sequence of models is formed according to the sequence of features yielded by the selection procedures mentioned above, the EBIC values of the models will first decrease then increase and decrease again when the size of the model gets close to n. Therefore, as a rule of thumb, in consistence with the sparsity assumption, the models with size close to or larger than n should not be selected even if they have smaller EBIC values.

Compared with the original BIC, in the EBIC for interactive models, there are two additional penalty terms, $2\gamma_\mathrm{M} \ln \binom{p}{\nu(s_\mathrm{M})}$ and $2\gamma_\mathrm{I} \ln \binom{p(p-1)/2}{\nu(s_\mathrm{I})}$. The penalty increases approximately by $2\gamma_\mathrm{M} \ln p$ when the number of main-effect features increases by 1 while the penalty increases approximately by $4\gamma_\mathrm{I} \ln p$ when the number of interaction features increases by 1. If $\gamma_\mathrm{M} = \gamma_\mathrm{I}$, an additional interaction feature incurs a penalty as twice as that incurred by an additional main-effect feature. As given in Theorem 1, in the range of consistency values of γ_M and γ_I, the lower bound of γ_I is larger than that of γ_M. This suggests that we should take γ_I larger than γ_M. Theorem 1 indicates that to achieve selection consistency, a higher penalty should be imposed for selecting an interaction feature than a main-effect feature, which makes an interaction feature harder to be selected than a main-effect feature. This is spiritually in line with the lasso-type methods discussed in Sect. 1.

The selection consistency is an asymptotic property. Though the selection consistency holds for any values of γ_M and γ_I in their consistency range, in the case of finite samples, different choices of γ_M and γ_I result in a not inconsiderable difference in the selection. In a practical feature selection problem, one is usually concerned with the positive discovery rate (PDR) and the false discovery rate (FDR) defined in Sect. 4. One wishes to have a high PDR and a low FDR. However, larger values of γ_{M} and γ_{I} give rise to a lower FDR but also a lower PDR, and smaller values of γ_{M} and γ_{I} give rise to a higher PDR but also a higher FDR. A reasonable strategy for the choice of γ_{M} and γ_{I} is to maximize the PDR, while the FDR is controlled. Following this strategy, one should choose γ_{M} and γ_{I} as close to their respective lower bounds as possible. Essentially, we can take γ_{M} and γ_{I} as their lower bounds. When p is smaller than n, the lower bounds could be negative. Hence, we suggest to take $(\gamma_{\mathrm{M}}, \gamma_{\mathrm{I}})$ as $\left(\max\{0, 1-\frac{\ln n}{2\ln p}\}, \max\left\{0, 1-\frac{\ln n}{4\ln p}\right\}\right)$. When both γ_{M} and γ_{I} are 0, the EBIC reduces to the original BIC.

3 A sequential procedure for feature selection with interactive models

In this section, we describe the sequential procedure for feature selection with interactive models. The procedure is essentially the approach of orthogonal matching pursuit (OMP) (or correlation pursuit by another terminology). Suppose that we have a current estimate $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu} = E\boldsymbol{y}$ and the current residual $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \hat{\boldsymbol{\mu}}$. The OMP selects the next feature which has the largest absolute correlation with the current residual. The OMP is slightly modified in our procedure. Let $\boldsymbol{x}_j, j = 1, \ldots, p$, denote the vectors of main-effect features and $\boldsymbol{z}_{jk}, j = 1, \ldots, p-1, k = j + 1, \ldots, p$, denote the vectors of interaction features. First, we select the main-effect feature that attains $\max_{1 \le j \le p} |\mathrm{corr}(\boldsymbol{x}_j, \tilde{\boldsymbol{y}})|$ and the interaction feature that attains $\max_{1 \le j \le p-1, j+1 \le k \le p} |\mathrm{corr}(\boldsymbol{z}_{jk}, \tilde{\boldsymbol{y}})|$, then select between the main-effect feature and the interaction feature by the EBIC for interactive models. In the following, we describe the algorithm of this sequential procedure.

Let all the $\boldsymbol{x}_j$'s and $\boldsymbol{z}_{jk}$'s be standardized such that each of the vectors has an L_2-norm n and is orthogonal to the vector with all elements 1. Let $\mathcal{S}_{\mathrm{M}}$ and $\mathcal{S}_{\mathrm{I}}$ be the index set of all the main-effect features and all the interaction features, respectively, i.e., $\mathcal{S}_{\mathrm{M}} = \{1, \ldots, p\}$ and $\mathcal{S}_{\mathrm{I}} = \{(j, k) : j = 1, \ldots, p-1, k = j + 1, \ldots, p\}$. The algorithm goes as follows:

Initialization: Set $s^* = \emptyset$ and $\tilde{\boldsymbol{y}} = \boldsymbol{y}$.
Iteration:

- Compute $|\boldsymbol{x}_j^{\tau}\tilde{\boldsymbol{y}}|$ for $j \in \mathcal{S}_{\mathrm{M}} \backslash s^*$ and identify j^* such that

$$|\boldsymbol{x}_{j^*}^{\tau}\tilde{\boldsymbol{y}}| = \max_{j \in \mathcal{S}_{\mathrm{M}} \backslash s^*} |\boldsymbol{x}_j^{\tau}\tilde{\boldsymbol{y}}|,$$

  and let $s^*_{+\mathrm{M}} = s^* \cup \{j^*\}$.
- Compute $|\boldsymbol{z}_{jk}^{\tau}\tilde{\boldsymbol{y}}|$ for $(jk) \in \mathcal{S}_{\mathrm{I}} \backslash s^*$ and identify (j^*k^*) such that

$$|\boldsymbol{z}_{j^*k^*}^{\tau}\tilde{\boldsymbol{y}}| = \max_{(j^*k^*) \in \mathcal{S}_{\mathrm{I}} \backslash s^*} |\boldsymbol{z}_{jk}^{\tau}\tilde{\boldsymbol{y}}|,$$

  and let $s^*_{+\mathrm{I}} = s^* \cup \{(j^*k^*)\}$.

- If $\mathrm{EBIC}_{\gamma_{\mathrm{M}},\gamma_{\mathrm{I}}}(s^*_{+\mathrm{M}}) < \mathrm{EBIC}_{\gamma_{\mathrm{M}},\gamma_{\mathrm{I}}}(s^*_{+\mathrm{I}})$, let $s^*_{\mathrm{NEW}} = s^*_{+\mathrm{M}}$, otherwise, let $s^*_{\mathrm{NEW}} = s^*_{+\mathrm{I}}$.
- If $\mathrm{EBIC}_{\gamma_{\mathrm{M}},\gamma_{\mathrm{I}}}(s^*_{\mathrm{NEW}}) < \mathrm{EBIC}_{\gamma_{\mathrm{M}},\gamma_{\mathrm{I}}}(s^*)$, let

$$s^* = s^*_{\mathrm{NEW}},\ \ \tilde{\boldsymbol{y}} = [I - H(s^*)]\boldsymbol{y},$$

  and continue; otherwise, stop.
- In the EBIC, $(\gamma_{\mathrm{M}}, \gamma_{\mathrm{I}})$ is taken as $(1 - \frac{\ln n}{2\ln p}, 1 - \frac{\ln n}{4\ln p})$.

Output: The s^* obtained when the iteration stops is the index set of the selected features. The estimate of $\boldsymbol{\beta}(s^*)$ is given by the ordinary least squares estimate, i.e., $\hat{\boldsymbol{\xi}}(s^*) = [Z^\tau(s^*)Z(s^*)]^{-1}Z^\tau(s^*)\boldsymbol{y}$.

The above procedure differs from the traditional forward stepwise selection. To explain, consider the selection of the main-effect features. After the set s^* is selected, the above procedure selects the next main-effect feature by maximizing the correlation $|\boldsymbol{x}_j^\tau[I - H(s^*)]\boldsymbol{y}|$. The traditional forward stepwise selection selects the next main-effect feature by minimizing the residual sum of squares $\boldsymbol{y}^\tau[I - H(s^* \cup \{j\})]\boldsymbol{y}$, which is equivalent to maximizing $|\boldsymbol{x}_j^\tau[I - H(s^*)]\boldsymbol{y}|/\sqrt{\boldsymbol{x}_j^\tau[I - H(s^*)]\boldsymbol{x}_j}$, an inflated version of the correlation. This inflated correlation favors the features that have higher correlation with the features already selected. This is a disadvantage for the identification of true features when high spurious correlations are present. For more detailed interpretation, see section 6 of Luo and Chen (2014).

To get more insight into the sequential procedure, let us take a closer look at the EBIC for interactive models given in (5). When p is large and j is relatively small, we have $\binom{p}{j} \approx p^j$. Thus, the last two terms of the EBIC in (5) are approximately $2\gamma_{\mathrm{M}}\nu(s_{\mathrm{M}}) \ln p$ and $4\gamma_{\mathrm{I}}\nu(s_{\mathrm{I}}) \ln p$. When the number of main-effect features increase from $\nu(s_{\mathrm{M}})$ to $\nu(s_{\mathrm{M}}) + 1$, for a new main-effect feature to reduce the EBIC, its contribution to the reduction of residual sum of squares must be larger than $\ln n + 2\gamma_{\mathrm{M}} \ln p$. When the number of interaction features increases from $\nu(s_{\mathrm{I}})$ to $\nu(s_{\mathrm{I}}) + 1$, for a new interaction feature to reduce the EBIC, its contribution to the reduction of residual sum of squares must be larger than $\ln n + 4\gamma_{\mathrm{I}} \ln p$. Since $4\gamma_{\mathrm{I}} > 2\gamma_{\mathrm{M}}$, an interaction feature needs to have larger effect than a main-effect feature to be selected by the sequential procedure.

The sequential procedure mimics the sequential Lasso (SLasso) cum EBIC procedure developed in Luo and Chen (2014). The SLasso cum EBIC procedure selects features by sequentially solving partially penalized likelihood problems where the coefficients of the features already selected are not penalized. Consider the following partially penalized likelihood function:

$$\ell_p(\boldsymbol{\xi}(s^* \cup \mathcal{S}_{\mathrm{M}})) = \|\boldsymbol{y} - Z(s^* \cup \mathcal{S}_{\mathrm{M}})\boldsymbol{\xi}(s^* \cup \mathcal{S}_{\mathrm{M}})\|_2^2 + \lambda \sum_{j \in \mathcal{S}_{\mathrm{M}} \backslash s^*} |\xi_j|,$$

where, among the components of $\boldsymbol{\xi}(s^* \cup \mathcal{S}_{\mathrm{M}})$, those with indices in s^* are not penalized. The minimization of $\ell_p(\boldsymbol{\xi}(s^* \cup \mathcal{S}_{\mathrm{M}}))$ is equivalent to the minimization of

$$\ell_p(\boldsymbol{\xi}(\mathcal{S}_{\mathrm{M}} \backslash s^*)) = \|\tilde{\boldsymbol{y}} - \tilde{Z}(\mathcal{S}_{\mathrm{M}} \backslash s^*)\boldsymbol{\xi}(\mathcal{S}_{\mathrm{M}} \backslash s^*)\|_2^2 + \lambda \sum_{j \in \mathcal{S}_{\mathrm{M}} \backslash s^*} |\xi_j|,$$

where $\tilde{\boldsymbol{y}} = [I - H(s^*)]\boldsymbol{y}$ and $\tilde{Z}(\mathcal{S}_{\mathrm{M}} \backslash s^*) = [I - H(s^*)]Z(\mathcal{S}_{\mathrm{M}} \backslash s^*)$. If we set λ at the largest value that allows at least one component of $\boldsymbol{\xi}(\mathcal{S}_{\mathrm{M}} \backslash s^*)$ to be estimated non-zero, then, in the minimization of $\ell_p(\boldsymbol{\xi}(\mathcal{S}_{\mathrm{M}} \backslash s^*))$, the non-zero component corresponds to the $\boldsymbol{x}_{j^*}$ which achieves the maximum absolute correlation $\max_{j \in \mathcal{S}_{\mathrm{M}} \backslash s^*} |\boldsymbol{x}_j^\tau \tilde{\boldsymbol{y}}|$. Replacing $\mathcal{S}_{\mathrm{M}}$ by $\mathcal{S}_{\mathrm{I}}$ in the above argument, we have the same result for $z_{j^*k^*}$. Therefore, the sequential procedure described above can be considered as a modification of the SLasso cum EBIC procedure. It has been shown in Luo and Chen (2014) that the SLasso cum EBIC procedure is selection consistent; that is, as $n \to \infty$, $P(s^* = s_0) \to 1$, where s^* is the index set selected by the SLasso cum EBIC procedure and s_0 is the index set of the true model. The selection consistency of the sequential procedure described in this section can also be established. We state the result as follows.

For $s = s_{\mathrm{M}} \cup s_{\mathrm{I}} \subset \mathcal{S}$, let s_{M}^c and s_{I}^c be the complements of s_{M} and s_{I} in $\mathcal{S}_{\mathrm{M}}$ and $\mathcal{S}_{\mathrm{I}}$, respectively. Let $s_{\mathrm{M}}^- = s_{\mathrm{M}}^c \cap s_{0\mathrm{M}}$ and $s_{\mathrm{I}}^- = s_{\mathrm{I}}^c \cap s_{0\mathrm{I}}$. For $s \subset s_0$, define

$$\Gamma_{\mathrm{M}n}(j, s, \boldsymbol{\xi}) = \frac{1}{n}\boldsymbol{x}_j^\tau [I - H(s)]Z\boldsymbol{\xi}, \quad j \in \mathcal{S}_{\mathrm{M}},$$
$$\Gamma_{\mathrm{I}n}((jk), s, \boldsymbol{\xi}) = \frac{1}{n}\boldsymbol{z}_{jk}^\tau [I - H(s)]Z\boldsymbol{\xi}, \quad (jk) \in \mathcal{S}_{\mathrm{I}}.$$

We have the following theorem:

Theorem 2 *Assume that*

(i) $\ln p = O(n^\kappa)$, $\kappa < 1/3$, $p_0 = O(n^c)$, $c < 1/6$.
(ii) *There is a constant* q, $0 < q < 1$, *such that*

$$\max_{j \in s_{0\mathrm{M}}^c} |\Gamma_{\mathrm{M}n}(j, s, \boldsymbol{\xi}) < q \max_{j \in s_{\mathrm{M}}^-} |\Gamma_{\mathrm{M}n}(j, s, \boldsymbol{\xi})|,$$
$$\max_{(jk) \in s_{0\mathrm{I}}^c} |\Gamma_{\mathrm{I}n}((jk), s, \boldsymbol{\xi}) < q \max_{(jk) \in s_{\mathrm{I}}^-} |\Gamma_{\mathrm{I}n}((jk), s, \boldsymbol{\xi})|.$$

(iii) *There is a constant* C *such that* $\lambda_{\min}(\frac{1}{n}Z(s_0)^\tau Z(s_0)) \min_{j \in s_0} |\beta_j| \geq Cn^{-1/6+\delta}$, *where* δ *is an arbitrarily small positive number.*

Let s^* *be the index set selected by the sequential procedure described in this section. Then,*

$$P(s^* = s_0) \to 1, \quad \textit{as} \quad n \to \infty.$$

Some remarks on the conditions of Theorem 2 are in order. Condition (i) specifies the diverging pattern of (n, p_0, p) with which the selection consistency of the sequential procedure holds. Condition (ii) essentially requires that the spurious correlations of irrelevant features are less than the true correlations of the true features. Condition (iii) essentially imposes a lower bound for the magnitude of effects to be detectable. If $\lambda_{\min}(\frac{1}{n}Z(s_0)^\tau Z(s_0))$ is bounded away from zero, which is a common assumption for high-dimensional data, condition (iii) simply requires that for an effect to be detectable it must have a magnitude larger than $Cn^{-1/6}$.

Denote the sets of features selected by the procedure from the first step onwards as $s_1^*, s_2^*, \ldots, s_{k^*}^*$, where $s_{k^*}^* = s^*$. Theorem 2 implies that (a) for $k \leq k^*$, $P(s_k^* \subset s_0) \to 1$ uniformly in k, (b) for $k < k^*$, $P(\text{EBIC}(s_k^*) > \text{EBIC}(s_{k+1}^*)) \to 1$ uniformly in k, and (c) $P(s^* = s_0, \min_{s:s_0 \subset s} \text{EBIC}(s) > \text{EBIC}(s^*)) \to 1$. In words, it implies that, asymptotically with probability 1, at each step of the procedure, only true features can be selected, and the procedure stops only when all the true features have been selected.

The proof of Theorem 2 is similar to that of Theorem 3.3 in Luo and Chen (2014) and is omitted here.

4 Numerical studies

In this section, we report the results of numerical studies including two sets of simulations and a real data analysis. The purpose of the first set of simulations is to demonstrate the selection consistency of the EBIC for interactive models through the trend of positive discovery rate (PDR) and false discovery rate (FDR) in finite samples and to compare the sequential procedure with non-sequential procedures. The purpose of the second set of simulations is to demonstrate the advantage of interactive models over main-effect models for the identification of causal variables.

The PDR and FDR are defined below.

$$\text{PDR} = \frac{\nu(s^* \cap s_0)}{\nu(s_0)}, \quad \text{FDR} = \frac{\nu(s^* \backslash s_0)}{\nu(s^*)}, \tag{8}$$

where s^* is the selected model and s_0 is the true model. The PDR and FDR are closely related to selection consistency. The selection consistency implies that the PDR and FDR converge to 1 and 0, respectively, in probability. The PDR and FDR are used as the criteria for the comparison between different procedures in the simulation studies.

4.1 Simulation study I

In this simulation study, we take the settings of (n, p_0, p) given in Table 1. For each setting, the p covariates are generated by three different correlation structures given below:

S1: The covariates are generated in independent blocks of size 50. The covariates within each block are generated from a multivariate normal distribution with zero mean vector and covariance matrix

Table 1 The settings of (n, p_0, p) in simulation study I

n	p_0	p
100	11	107
200	14	365
400	17	1706

$$\Sigma = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \cdots & \cdots & \cdots & \cdots \\ \rho & \cdots & \rho & 1 \end{pmatrix}.$$

S2: The covariates are generated by the time series:

$$X_j = \rho X_{j-1} + \sqrt{1-\rho^2} Z_j, \quad j = 1, 2, \ldots, p,$$

where X_0 and Z_j, $j = 1, \ldots, p$, are i.i.d. standard normal variables.

S3: First, Z_j, $j = 0, 1, \ldots, p$, are generated as standard normal variables. The X_j's are then generated as

$$X_j = \frac{1}{\sqrt{5}} Z_0 + \frac{\sqrt{2}}{5} Z_j, \quad 1 \le j \le p_0,$$
$$X_j = \rho X_{j-k} + \sqrt{1-\rho^2} Z_j, \quad p_0 + 1 \le j \le p.$$

Three values of ρ, i.e., $\rho = 0.3, 0.5$ and 0.7 are considered in the simulation. The response variable y is generated according to the following models:

$$\text{Model A:} \quad y = \sum_{j=1}^{k} \beta_j X_j + \sum_{j=1}^{p_0-k} \beta_{k+j} X_{2j-1} X_{2j} + \epsilon, \;\; k = [p_0/2] + 1,$$

$$\text{Model B:} \quad y = \sum_{j=1}^{5} \beta_j X_j + \beta_6 X_1 X_2 + \beta_7 X_1 X_3 + \beta_8 X_1 X_6 + \beta_9 X_5 X_6$$
$$+ \sum_{j=10}^{p_0} \beta_j X_{j-1} X_j + \epsilon,$$

where ϵ is generated as a normal variable with mean zero and variance σ^2. Two values of σ, i.e., $\sigma = 1$ and 1.5 are considered. The β_j's are generated independently as $2n^{-0.175} + |z|/10$, where $z \sim N(0, 1)$.

By the combination of model, correlation structure and the sample size, we have all together 72 simulation settings. For each setting, 200 replicates of data are generated. For the demonstration of selection consistency, besides EBIC, the original BIC is also used as a criterion in order to see the difference between EBIC and BIC. In EBIC, $(\gamma_{\mathrm{M}}, \gamma_{\mathrm{I}})$ is taken as $(1 - \frac{\ln n}{2 \ln p}, 1 - \frac{\ln n}{4 \ln p})$. In the comparison of the sequential procedure with non-sequential procedures, we choose the penalized likelihood approach with SCAD penalty (we simply refer to this approach as SCAD) as the representative of the non-sequential procedures, because, in simulation studies carried out elsewhere, e.g., Luo et al. (2014), the SCAD routinely performs well in comparison of penalized likelihood methods in terms of feature selection. The same selection criterion is used in both methods. The R package NCVREG [Breheny and Huang (2011)] is used for the computation of SCAD.

Table 2 The trend of PDR and FDR of the sequential procedure under the simulation settings with correlation structure S1 and model A (the PDR and FDR are averaged over 200 replicates for each setting, the numbers in parentheses are standard deviations)

σ	n	ρ	PDR		FDR	
			BIC	EBIC	BIC	EBIC
1	100	0.3	0.776 (0.191)	0.751 (0.229)	0.929 (0.017)	0.283 (0.187)
		0.5	0.649 (0.201)	0.542 (0.237)	0.906 (0.090)	0.418 (0.180)
		0.7	0.441 (0.164)	0.313 (0.161)	0.828 (0.132)	0.565 (0.182)
	200	0.3	0.916 (0.081)	0.920 (0.071)	0.945 (0.005)	0.128 (0.097)
		0.5	0.848 (0.118)	0.833 (0.142)	0.940 (0.049)	0.214 (0.128)
		0.7	0.620 (0.177)	0.519 (0.200)	0.715 (0.143)	0.415 (0.172)
	400	0.3	0.953 (0.047)	0.953 (0.047)	0.938 (0.007)	0.074 (0.080)
		0.5	0.942 (0.056)	0.942 (0.056)	0.934 (0.028)	0.091 (0.083)
		0.7	0.879 (0.098)	0.840 (0.126)	0.676 (0.112)	0.195 (0.109)
1.5	100	0.3	0.618 (0.196)	0.577 (0.228)	0.944 (0.018)	0.353 (0.205)
		0.5	0.541 (0.163)	0.376 (0.189)	0.933 (0.060)	0.489 (0.204)
		0.7	0.360 (0.169)	0.231 (0.131)	0.841 (0.128)	0.617 (0.198)
	200	0.3	0.801 (0.196)	0.852 (0.155)	0.952 (0.012)	0.156 (0.108)
		0.5	0.715 (0.167)	0.685 (0.197)	0.947 (0.046)	0.271 (0.148)
		0.7	0.464 (0.162)	0.378 (0.178)	0.753 (0.126)	0.508 (0.184)
	400	0.3	0.939 (0.075)	0.938 (0.059)	0.951 (0.021)	0.092 (0.073)
		0.5	0.930 (0.052)	0.924 (0.067)	0.948 (0.033)	0.102 (0.081)
		0.7	0.783 (0.086)	0.700 (0.164)	0.692 (0.104)	0.289 (0.141)

A part of the results for demonstrating the selection consistency of EBIC is presented in Tables 2 and 3. These tables provide the PDR and FDR averaged over 200 replicates for each of the settings when correlation structure S1 is combined with model A and correlation structure S2 is combined with model B. Since the results under other settings are similar, for the sake of clarity, the results under other settings are omitted. The comparison between the sequential procedure and the SCAD is also made by comparing their PDR and FDR averaged over 200 replicates under all the simulation settings. The results under model A and model B are quite similar. Therefore, only the results under model B given in Table 4 are reported.

The results presented in Tables 2 and 3 are summarized as follows. Under all the settings, the PDR for EBIC has an strong upward trend towards 1 and the FDR for EBIC has a strong downward trend towards 0. For example, for $\sigma = 1$ and $\rho = 0.3$, as n varies from 100 to 200 and 400, in Table 2, the PDR for EBIC varies from 0.751 to 0.909 and 0.953, the FDR for EBIC varies from 0.283 to 0.128 and 0.074, and, in Table 3, the PDR for EBIC varies from 0.881 to 0.929 and 0.949, the FDR for EBIC varies from 0.212 to 0.117 and 0.068. For other values of σ and ρ, the trends of PDR and FDR for EBIC are similar. On the other hand, though the PDR for BIC has a upward trend, the FDR for BIC stays at high levels and does not show any trend

Table 3 The trend of PDR and FDR of the sequential procedure under the simulation settings with correlation structure S2 and model B (the PDR and FDR are averaged over 200 replicates for each setting, the numbers in parentheses are standard deviations)

σ	n	ρ	PDR		FDR	
			BIC	EBIC	BIC	EBIC
1	100	0.3	0.798 (0.244)	0.881 (0.185)	0.927 (0.022)	0.212 (0.158)
		0.5	0.783 (0.257)	0.859 (0.193)	0.929 (0.023)	0.234 (0.171)
		0.7	0.727 (0.237)	0.742 (0.234)	0.934 (0.022)	0.305 (0.190)
	200	0.3	0.918 (0.119)	0.929 (0.092)	0.945 (0.007)	0.117 (0.106)
		0.5	0.904 (0.150)	0.925 (0.091)	0.946 (0.009)	0.127 (0.096)
		0.7	0.848 (0.163)	0.867 (0.130)	0.949 (0.010)	0.186 (0.125)
	400	0.3	0.950 (0.049)	0.949 (0.049)	0.946 (0.005)	0.068 (0.069)
		0.5	0.947 (0.055)	0.947 (0.055)	0.948 (0.007)	0.075 (0.070)
		0.7	0.945 (0.057)	0.943 (0.059)	0.952 (0.008)	0.087 (0.073)
1.5	100	0.3	0.579 (0.258)	0.684 (0.251)	0.947 (0.023)	0.275 (0.181)
		0.5	0.553 (0.241)	0.634 (0.235)	0.950 (0.022)	0.286 (0.196)
		0.7	0.518 (0.208)	0.527 (0.227)	0.953 (0.019)	0.361 (0.210)
	200	0.3	0.745 (0.241)	0.823 (0.177)	0.955 (0.015)	0.160 (0.125)
		0.5	0.748 (0.242)	0.801 (0.204)	0.955 (0.015)	0.173 (0.146)
		0.7	0.675 (0.233)	0.718 (0.217)	0.959 (0.014)	0.237 (0.158)
	400	0.3	0.930 (0.064)	0.929 (0.065)	0.965 (0.002)	0.075 (0.071)
		0.5	0.925 (0.061)	0.923 (0.063)	0.967 (0.004)	0.088 (0.076)
		0.7	0.918 (0.074)	0.908 (0.081)	0.968 (0.004)	0.106 (0.086)

towards 0. Even when $n = 400$, the smallest FDR for BIC is 0.676 in Table 2 and 0.946 in Table 3. The findings demonstrate that, for EBIC, PDR $\to 1$ and FDR $\to 0$, which is the evidence for the selection consistency of EBIC, and that, for BIC, FDR does not converge to 0, which is an indication that BIC is not selection consistent.

For the comparison between the sequential procedure and the SCAD, we report the results under model B in Table 4. The results under model A are similar and, hence, are omitted. We can see from Table 4 that, except a few settings, the sequential procedure has higher PDRs and lower FDRs than the SCAD. Table 5 provides the average PDR and FDR of the two procedures over all the settings for each of the sample sizes. It is clear from Table 5 that, overall, the sequential procedure has a much higher PDR and lower FDR than the SCAD. This justifies the edge of the sequential procedure over non-sequential procedures for feature selection in interactive models.

4.2 Simulation study II

In certain practical problems, the major concern is to identify the covariates which affect the response variable no matter whether the effects of the covariates are in the form of main effect or interaction. For example, the goal of a QTL mapping study

Table 4 Results of simulation study I: comparison of PDR and FDR between the sequential interactive procedure (SIP) and the non-sequential procedure SCAD (the numbers without parentheses are PDR's and those within parentheses are FDR's

	ρ	PDR (FDR)		
		$n = 100$	$n = 200$	$n = 400$
$\sigma = 1$				
S1				
SIP	0.3	0.788 (0.268)	0.891 (0.139)	0.945 (0.062)
	0.5	0.577 (0.393)	0.698 (0.253)	0.914 (0.093)
	0.7	0.356 (0.543)	0.336 (0.482)	0.704 (0.272)
SCAD	0.3	0.799 (0.240)	0.813 (0.207)	0.939 (0.097)
	0.5	0.629 (0.336)	0.635 (0.313)	0.898 (0.171)
	0.7	0.432 (0.490)	0.385 (0.502)	0.725 (0.348)
S2				
SIP	0.3	0.881 (0.212)	0.929 (0.117)	0.949 (0.068)
	0.5	0.859 (0.234)	0.925 (0.127)	0.947 (0.075)
	0.7	0.742 (0.305)	0.867 (0.186)	0.943 (0.087)
SCAD	0.3	0.787 (0.270)	0.863 (0.196)	0.945 (0.103)
	0.5	0.804 (0.289)	0.835 (0.208)	0.941 (0.110)
	0.7	0.714 (0.278)	0.785 (0.239)	0.934 (0.111)
S3				
SIP	0.3	0.860 (0.206)	0.917 (0.120)	0.947 (0.067)
	0.5	0.848 (0.222)	0.920 (0.122)	0.950 (0.068)
	0.7	0.757 (0.294)	0.884 (0.165)	0.940 (0.082)
SCAD	0.3	0.775 (0.289)	0.808 (0.198)	0.933 (0.103)
	0.5	0.771 (0.283)	0.829 (0.199)	0.939 (0.102)
	0.7	0.726 (0.323)	0.836 (0.229)	0.933 (0.110)
$\sigma = 1.5$				
S1				
SIP	0.3	0.584 (0.333)	0.769 (0.203)	0.918 (0.096)
	0.5	0.420 (0.460)	0.547 (0.312)	0.865 (0.125)
	0.7	0.285 (0.538)	0.289 (0.499)	0.534 (0.361)
SCAD	0.3	0.509 (0.356)	0.500 (0.239)	0.762 (0.153)
	0.5	0.432 (0.450)	0.399 (0.351)	0.589 (0.259)
	0.7	0.353 (0.517)	0.302 (0.534)	0.371 (0.415)
S2				
SIP	0.3	0.684 (0.275)	0.823 (0.160)	0.929 (0.075)
	0.5	0.634 (0.286)	0.801 (0.173)	0.923 (0.088)
	0.7	0.527 (0.361)	0.718 (0.237)	0.908 (0.106)
SCAD	0.3	0.404 (0.387)	0.402 (0.167)	0.672 (0.203)
	0.5	0.393 (0.332)	0.379 (0.188)	0.676 (0.202)
	0.7	0.361 (0.344)	0.386 (0.286)	0.626 (0.231)

Table 4 continued

	ρ	PDR (FDR)		
		$n = 100$	$n = 200$	$n = 400$
S3				
SIP	0.3	0.674 (0.264)	0.794 (0.180)	0.922 (0.078)
	0.5	0.632 (0.277)	0.823 (0.162)	0.921 (0.083)
	0.7	0.532 (0.360)	0.744 (0.216)	0.898 (0.102)
SCAD	0.3	0.392 (0.311)	0.394 (0.142)	0.652 (0.179)
	0.5	0.394 (0.327)	0.418 (0.205)	0.655 (0.191)
	0.7	0.362 (0.339)	0.413 (0.278)	0.633 (0.246)

Table 5 The average PDR and FDR of the sequential interactive procedure (SIP) and the non-sequential procedure SCAD

	n		
	100	200	400
PDR			
SIP	0.647	0.760	0.892
SCAD	0.558	0.577	0.768
FDR			
SIP	0.324	0.214	0.110
SCAD	0.342	0.260	0.185

is to discover the QTLs. In general, a quantitative trait is affected by many QTLs. Some QTLs have an effect only through their interaction with other QTLs. For those QTLs, their main effect usually appears non-significant. Such QTLs cannot be detected by using only main-effect models. However, it is possible for them to be detected by using interactive models. In this sub section, we demonstrate this through a specifically designed simulation study. We consider the following true model:

$$\boldsymbol{y} = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_5 X_5 + \beta_6 X_1 X_2 + \beta_7 X_3 X_4 + \boldsymbol{\epsilon}.$$

The covariates are generated in the same way as in simulation study I except that, under all correlation structures, the mean of the X_j's is 1 instead of 0. The regression coefficients are specified as $\beta_j = -\beta_6 = -\beta_7 = 0.8$, $j = 1, \ldots, 4$ and $\beta_5 = 0.8 + |z|/10$, where $z \sim N(0, 1)$. With this specification, the effect of the first four covariates on the response variable is only through their interaction terms. If a main-effect model was used to fit the data, the coefficients of the first four covariates would be almost zero. We consider the same settings as those in simulation study I. At each setting, we apply two sequential procedures to the generated data: the one with main-effect models and the other with interactive models. For convenience, we refer the first one as the additive procedure and the second one as the interactive procedure.

In this simulation study, we consider PDR and FDR in terms of covariates rather than features. In other words, we are concerned with whether or not a covariate is correctly

Table 6 Results of simulation study II: Comparison of the sequential interactive procedure and the sequential additive procedure (the numbers without parentheses are PDR's and those within parentheses are FDR's, PDR $_{1234}$ is the positive discovery rate for the detection of the first four covariates)

	ρ	PDR (FDR)		PDR_{1234}	
		$n = 200$	$n = 400$	$n = 200$	$n = 400$
$\sigma = 0.5$					
S1					
Additive	0.3	0.156 (0.161)	0.147 (0.144)	0.023	0.008
	0.5	0.146 (0.159)	0.144 (0.148)	0.006	0.003
	0.7	0.146 (0.118)	0.144 (0.088)	0.006	0.001
Interactive	0.3	0.788 (0.164)	0.887 (0.113)	0.739	0.861
	0.5	0.565 (0.199)	0.610 (0.140)	0.463	0.513
	0.7	0.422 (0.249)	0.426 (0.109)	0.281	0.302
S2					
Additive	0.3	0.150 (0.177)	0.146 (0.139)	0.013	0.005
	0.5	0.152 (0.123)	0.146 (0.125)	0.016	0.005
	0.7	0.151 (0.178)	0.145 (0.113)	0.015	0.004
Interactive	0.3	0.904 (0.152)	0.970 (0.109)	0.885	0.964
	0.5	0.889 (0.146)	0.891 (0.128)	0.864	0.865
	0.7	0.725 (0.175)	0.795 (0.124)	0.656	0.748
S3					
Additive	0.3	0.149 (0.189)	0.146 (0.145)	0.010	0.005
	0.5	0.149 (0.171)	0.149 (0.168)	0.010	0.011
	0.7	0.147 (0.148)	0.145 (0.133)	0.008	0.004
Interactive	0.3	0.842 (0.176)	0.965 (0.121)	0.809	0.958
	0.5	0.879 (0.163)	0.909 (0.120)	0.854	0.889
	0.7	0.752 (0.181)	0.779 (0.124)	0.711	0.759
$\sigma = 1$					
S1					
Additive	0.3	0.151 (0.144)	0.146 (0.116)	0.014	0.005
	0.5	0.145 (0.133)	0.144 (0.150)	0.004	0.003
	0.7	0.142 (0.136)	0.144 (0.087)	0.006	0.003
Interactive	0.3	0.495 (0.184)	0.782 (0.128)	0.371	0.730
	0.5	0.449 (0.245)	0.580 (0.163)	0.323	0.475
	0.7	0.326 (0.301)	0.393 (0.165)	0.203	0.245
S2					
Additive	0.3	0.146 (0.148)	0.144 (0.151)	0.006	0.003
	0.5	0.147 (0.126)	0.145 (0.130)	0.008	0.004
	0.7	0.149 (0.182)	0.144 (0.135)	0.010	0.001
Interactive	0.3	0.654 (0.192)	0.874 (0.153)	0.461	0.844
	0.5	0.549 (0.209)	0.840 (0.144)	0.441	0.803
	0.7	0.515 (0.229)	0.686 (0.146)	0.406	0.608

Table 6 continued

	ρ	PDR (FDR)		PDR_{1234}	
		$n = 200$	$n = 400$	$n = 200$	$n = 400$
S3					
Additive	0.3	0.146 (0.174)	0.144 (0.127)	0.006	.003
	0.5	0.146 (0.159)	0.144 (0.177)	0.005	0.003
	0.7	0.145 (0.135)	0.145 (0.125)	0.004	0.004
Interactive	0.3	0.568 (0.230)	0.900 (0.129)	0.466	0.875
	0.5	0.562 (0.190)	0.850 (0.130)	0.463	0.815
	0.7	0.527 (0.199)	0.664 (0.133)	0.415	0.580

selected regardless of the nature of its effect. Let s_0 be the set of true covariates and s^* the set of selected covariates. In particular, for the model above, s_0 consists of the five covariates X_j, $j = 1, \ldots, 5$, rather than the seven features. If we have selected the features $\{X_1X_2, X_5\}$ then s^* consists of the three covariates X_1, X_2 and X_5 rather than the two features. The PDR and FDR are still defined by (8). The simulation results are reported in Table 6. Again, for each setting, the PDR and FDR are averaged over the 200 replicates. In addition to the overall PDR and FDR, we also include in Table 6 the PDR concerning only the first four covariates, i.e., the proportion of the first four covariates which have been detected. It is denoted by PDR_{1234} in Table 6.

The findings from Table 6 are listed as follows. (1) The additive and interactive procedures have about the same capacity to control the level of FDR. It can be seen from the table that, across all the settings, there does not exist too much difference in FDR between the two procedures. (2) The significant difference between the two procedures is in the PDR. Across all the settings, the PDRs of the additive procedure are extremely low but the PDRs of the interactive procedure are uniformly quite high. (3) The low PDR of the additive procedure is because of its inability to discover the first four covariates. The PDR_{1234}s of the additive procedure are less than or equal to 1 % except only a few settings. This indicates that the additive procedure has no power at all for the discovery of the first four covariates. (4) The interactive procedure has about the same efficiency to discover all the covariates regardless of the nature of their effects. This can be seen by comparing the PDR and PDR_{1234} of the interactive procedure. These two values are quite close across all the settings. In summary, simulation study II provides us with strong evidence to prefer the interactive procedure to the additive procedure in the detection of true covariates when interactions exist.

4.3 A real data example

The sequential interactive procedure is applied to a mouse data set for mapping QTL of locomotor activation and anxiety considered in Bailey et al. (2008). The data set was obtained from an open-field assay test for 196 female and 166 male mice which are F_2 progeny of two phenotypically similar inbred mouse strains: `C57BL/6J` and `C58/J`. Six measures were considered in the open-field assay test: (1) total distance traveled (in cm), (2) ambulatory episodes (number of times animal breaks user defined

Table 7 Results of the real data analysis: the locations of the detected SNPs and the nature of their effects on the behavioral measures

	Chr	Location (Mb)	Effect	Interactive SNPs
Percent time in center	13	89.444	Main	
	2	178.315	Interaction	Chr13:22.251
	13	22.251	Interaction	Chr2:178.315
Total distance	8	57.724	Main	
	17	56.801	Main	
	6	102.455	Interaction	Chr12:20.058
	12	20.058	Interaction	Chr6:102.445
Total rearing	2	153.094	Main	
Ambulatory episodes	8	68.129	Main	
	17	56.801	Main	
	6	102.455	Interaction	Chr12:20.058
	12	20.058	Interaction	Chr6:102.455
Average velocity	8	89.447	Main	
Percent resting	2	97.379	Main	
	7	63.356	Main	
	8	89.447	Main	

number of beams before coming to rest), (3) percent time resting, (4) average velocity (in centimeters per second), (5) number of rearings and (6) percent time spent in center of arena. The data consist of the measurements on the six measures for each of the 362 mice together with their genotypes at 211 SNP markers. There are some missing values in the original data. We dropped the individuals that have more than 30 missing values and imputed the remaining missing values by the R package `Imputation` Wong (2013).

In our analysis, we take each of the six behavioral traits as a response variable. For each trait, the features (either main-effect features or interactive features formed by the 211 SNP markers) affecting the trait are identified by applying the sequential interactive procedure. The identified features are given in Table 7. The SNPs on the same chromosome which are located not far away from each other are usually considered as a single locus. Thus, in Table 7, the identified SNPs on the same chromosome can be considered as a single locus. For convenience, we simply refer to each of the chromosomes as a locus. All together, we have identified seven loci, i.e, 2, 6, 7, 8, 12, 13, 17. Five of the loci, i.e., 2, 6, 8, 12 and 17, affect multiple traits, the other two, i.e., 7 and 13, affect only one trait. In particular, locus 8 affects four traits.

It is interesting to compare our findings with those obtained in Bailey et al. (2008). Bailey et al. (2008) used the multiple-test approach for testing the significance of the main-effect features of the SNPs and used the threshold value 3.2, which is determined by the method of permutation test, for the significance of the LOD scores of the tests. They discovered the following 11 loci: 1, 2, 3, 5, 6, 7, 8, 11, 13, 16, 17. Strikingly, all the loci but one which we discovered are contained in these eleven loci. The exceptional one, locus 12, is discovered through its interaction effect with locus 6 in our analysis.

Further, in Bailey et al. (2008), locus 8 was also found to affect significantly the same four traits as in our analysis. Therefore, we can comfortably claim that the loci we discovered are statistically true QTLs which might be further confirmed by biological experiments.

Appendix: Proof of Theorem 1

The following results are to be used in the proof:

$$P(\chi_j^2 \geq m) = \frac{1}{\Gamma(j/2)}(m/2)^{j/2-1}e^{-m/2}(1+o(1)), \quad \text{if } m \to \infty, \frac{j}{m} \to 0, \tag{9}$$

$$\ln\left(\frac{p!}{j!(p-j)!}\right) = j\ln p(1+o(1)), \quad \text{if } p \to \infty, \frac{\ln j}{\ln p} \to 0, \tag{10}$$

where χ_j^2 is a χ^2 random variable with degrees of freedom j. The proof of these results can be found in Luo and Chen (2013).

Proof Let s be any sub-model. We can express $\text{EBIC}_{\gamma_\text{M},\gamma_\text{I}}(s) - \text{EBIC}_{\gamma_\text{M},\gamma_\text{I}}(s_0)$ as $T_1(s) + T_2(s)$, where

$$\begin{aligned} T_1(s) &= n\ln\frac{\boldsymbol{y}^\tau[I - H(s)]\boldsymbol{y}}{\boldsymbol{y}^\tau[I - H(s_0)]\boldsymbol{y}} = n\ln\frac{\boldsymbol{y}^\tau[I - H(s)]\boldsymbol{y}}{\boldsymbol{\epsilon}^\tau[I - H(s_0)]\boldsymbol{\epsilon}} \\ &= n\ln\left\{1 + \frac{\boldsymbol{y}^\tau[I - H(s)]\boldsymbol{y} - \boldsymbol{\epsilon}^\tau[I - H(s_0)]\boldsymbol{\epsilon}}{\boldsymbol{\epsilon}^\tau[I - H(s_0)]\boldsymbol{\epsilon}}\right\} \end{aligned}$$

and

$$\begin{aligned} T_2(s) = [\nu(s) - \nu(s_0)]\ln n &+ 2\gamma_\text{M}\left[\ln\binom{p}{\nu(s_\text{M})} - \ln\binom{p}{\nu(s_{0\text{M}})}\right] \\ &+ 2\gamma_\text{I}\left[\ln\binom{p(p-1)/2}{\nu(s_\text{I})} - \ln\binom{p(p-1)/2}{\nu(s_{0\text{I}})}\right]. \end{aligned}$$

Under the assumption of Theorem 1, it follows from (9) that

$$\begin{aligned} T_2(s) = [\nu(s) - \nu(s_0)]\ln n &+ 2\gamma_\text{M}[\nu(s_\text{M}) - \nu(s_{0\text{M}})]\ln p(1 + o_p(1)) \\ &+ 4\gamma_\text{I}[\nu(s_\text{I}) - \nu(s_{0\text{I}})]\ln p(1 + o_p(1)). \end{aligned} \tag{11}$$

We are going to show that

$$P\left(\min_{s:\nu(s)\leq rp_0}\{T_1(s) + T_2(s)\} > 0\right) \to 1. \tag{12}$$

Without loss of generality, we assume that $\sigma^2 = 1$ in what follows. Let $A_1 = \{s : s_0 \not\subset s\}$ and $A_2 = \{s : s_0 \subset s\}$. (12) will be established separately for $s \in A_1$ and $s \in A_2$.

Case 1: $s \in A_1$: Let $k_0 = rp_0$. First, we establish that

$$T_1 = n \ln \left(1 + \frac{\Delta_n(s)}{n}(1 + o_p(1))\right), \tag{13}$$

uniformly for all s with $\nu(s) \leq k_0$. Recall that $\Delta_n(s) = \boldsymbol{\mu}^\tau [I - H(s)]\boldsymbol{\mu}$. Let Z_i be i.i.d. standard normal random variables. Since $I - H(s_0)$ is a projection matrix, we can express

$$\boldsymbol{\epsilon}^T \{I - H(s_0)\}\boldsymbol{\epsilon} = \sum_{i=1}^{n-p_0} Z_i^2 = (n - p_0)(1 + o_p(1)) = n(1 + o_p(1)), \tag{14}$$

by the law of large numbers. Thus, (13) follows if

$$\boldsymbol{y}^\tau [I - H(s)]\boldsymbol{y} - \boldsymbol{\epsilon}^\tau [I - H(s_0)]\boldsymbol{\epsilon} = \Delta_n(s)(1 + o_p(1)). \tag{15}$$

We have

$$\begin{aligned}\boldsymbol{y}^\tau [I - H(s)]\boldsymbol{y} - \boldsymbol{\epsilon}^\tau [I - H(s_0)]\boldsymbol{\epsilon} = \Delta_n(s) + 2\boldsymbol{\mu}^\tau [I - H(s)]\boldsymbol{\epsilon} \\ + \boldsymbol{\epsilon}^\tau H(s_0)\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^\tau H(s)\boldsymbol{\epsilon}.\end{aligned}$$

Then, (15) holds under the assumption of Theorem 1, if

$$\boldsymbol{\epsilon}^\tau H(s_0)\boldsymbol{\epsilon} = p_0(1 + o_p(1)); \tag{16}$$

$$\max_{s:\nu(s)\leq k} \boldsymbol{\epsilon}^\tau H(s)\boldsymbol{\epsilon} = O_p(k \ln p); \tag{17}$$

$$|\boldsymbol{\mu}^\tau [I - H(s)]\boldsymbol{\epsilon}| = \sqrt{\Delta_n(s) O_p(k \ln p)}. \tag{18}$$

(16) is a similar result to (14). In the following, we verify (17) and (18).
Verification of (17): Let $a = p(p + 1)/2$ and $m = 2k[\ln a + \ln(k \ln a)]$. Obviously, $\frac{k}{m} \to 0$. Note that $\boldsymbol{\epsilon}^\tau H(s)\boldsymbol{\epsilon} = \chi_j^2(s)$ for $j = \nu(s)$. By the Bonferroni inequality, we get

$$\begin{aligned}&P(\max\{\boldsymbol{\epsilon}^\tau H(s)\boldsymbol{\epsilon} : \nu(s) \leq k\} \geq m) \\ &\quad = P(\max\{\boldsymbol{\epsilon}^\tau H(s)\boldsymbol{\epsilon} : s \in \mathcal{S}_j, j \leq k\} \geq m) \leq \sum_{j=1}^{k} \tau(\mathcal{S}_j) P(\chi_j^2 \geq m),\end{aligned}$$

where $\mathcal{S}_j$ is the set of models consisting of j features (including both main-effect and interaction features). Note that $\tau(\mathcal{S}_j) = \binom{p(p+1)/2}{j}$. Since $j/m \leq k/m \to 0$, by (9), there is some c close to 1 and independent of j such that

$$\tau(\mathcal{S}_j)P(\chi_j^2 \geq m) \approx c\frac{1}{2^{j/2-1}\Gamma(j/2)}\frac{\tau(\mathcal{S}_j)}{a^k}(k\ln a)^{-k}m^{j/2-1}$$
$$\leq \frac{c}{m}(k\ln a)^{-j}m^{j/2} = \frac{c}{m}\left[\sqrt{\frac{m}{(k\ln a)^2}}\right]^j = \frac{c}{m}q_n^j.$$

When n is sufficiently large,

$$q_n = \sqrt{\frac{m}{(k\ln a)^2}} = \sqrt{\frac{2k[\ln a + \ln(k\ln a)]}{(k\ln a)^2}} \leq q$$

for some $0 < q < 1$ since $q_n \to 0$. Thus,

$$P(\max\{\boldsymbol{\epsilon}^\tau H(s)\boldsymbol{\epsilon} : s \in \mathcal{S}_j,\, j \leq k\} \geq m) \leq \frac{c}{m}\sum_{j=1}^{k} q_n^j \leq \frac{c}{m}\frac{q}{1-q} \to 0.$$

Therefore

$$\max_{s:\nu(s)\leq k} \boldsymbol{\epsilon}^\tau H(s)\boldsymbol{\epsilon} = m(1+o_p(1)) = O_p(k\ln p),$$

which verifies (17).

Verification of (18): Note that $\boldsymbol{\mu}^\tau[I - H(s)]\boldsymbol{\epsilon}$ follows a normal distribution with mean 0 and variance $\Delta_n(s)$. Hence, we can express

$$\boldsymbol{\mu}^\tau[I - H(s)]\boldsymbol{\epsilon} = \sqrt{\Delta_n(s)}Z(s),$$

where $Z(s) \sim N(0, 1)$. Thus, we have

$$|\boldsymbol{\mu}^\tau[I - H(s)]\boldsymbol{\epsilon}| \leq \sqrt{\Delta_n(s)}\max\{|Z(s)| : \nu(s) \leq k\}.$$

It is implied by (9) that $P(\chi_1^2 \geq m) \leq P(\chi_j^2 \geq m)$. Thus,

$$\begin{aligned} P(\max\{|Z(s)| : \nu(s) \leq k\} \geq \sqrt{m}) &= P(\max\{|Z(s)| : s \in \mathcal{S}_j,\, j \leq k\} \geq \sqrt{m}) \\ &\leq \sum_{j=1}^{k}\tau(\mathcal{S}_j)P(|Z(s)| \geq \sqrt{m}) \\ &= \sum_{j=1}^{k}\tau(\mathcal{S}_j)P(\chi_1^2 \geq m) \\ &\leq \sum_{j=1}^{k}\tau(\mathcal{S}_j)P(\chi_j^2 \geq m). \end{aligned}$$

It has been shown in the verification of (17) that the last sum above converges to zero. Therefore, $\max\{|Z(s)| : \nu(s) \leq k\} = \sqrt{O_p(k \ln p)}$. This verifies (18).

Now consider two scenarios: $\frac{\Delta_n(s)}{n} \to 0$ and $\frac{\Delta_n(s)}{n} \geq C > 0$. If $\frac{\Delta_n(s)}{n} \to 0$, then $T_1 = n \ln(1 + \frac{\Delta_n(s)}{n}(1 + o_p(1))) \approx \Delta_n(s)(1 + o_p(1))$ when n is sufficiently large. Let $\gamma = \max(\gamma_{\mathrm{M}}, \gamma_{\mathrm{I}})$. Then, it follows from (11) and (13) that

$$\begin{aligned} T_1(s) + T_2(s) &\geq \Delta_n(s)(1 + o_p(1)) - p_0 \ln n - 4\gamma p_0 \ln p \\ &\geq \frac{\Delta_n(s)}{p_0 \ln p}\left(p_0 \ln p - \frac{\ln n}{\ln p} - 4\gamma\right) \to \infty \end{aligned}$$

uniformly for all s with $\nu(s) \leq k$ and any γ under the condition of Theorem 1. If $\frac{\Delta_n(s)}{n} = C > 0$, we have

$$T_1(s) + T_2(s) \geq n \ln(1 + C) - p_0 \ln n - 4\gamma p_0 \ln p \to \infty$$

uniformly for all s with $\nu(s) \leq k$ and any γ. Thus, (12) is established in the case $s \in A_1$.

Case 2: $s \in A_2$. When $s_0 \subset s$, $\{I - H(s)\}X(s_0) = 0$ and hence,

$$\begin{aligned} &\boldsymbol{y}^\tau[I - H(s)]\boldsymbol{y} = \boldsymbol{\epsilon}^\tau[I - H(s)]\boldsymbol{\epsilon}, \\ &\boldsymbol{\epsilon}^\tau[I - H(s_0)]\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^\tau[I - H(s)]\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^\tau[H(s) - H(s_0)]\boldsymbol{\epsilon} = \chi^2_{j(s)}, \end{aligned}$$

where $j(s) = \nu(s) - \nu(s_0)$. Hence, we have

$$\begin{aligned} -T_1(s) &= n \ln \frac{\boldsymbol{\epsilon}^\tau[I - H(s_0)]\boldsymbol{\epsilon}}{\boldsymbol{\epsilon}^\tau[I - H(s)]\boldsymbol{\epsilon}} \\ &= n \ln\left[1 + \frac{\chi^2_{j(s)}}{\boldsymbol{\epsilon}^\tau[I - H(s_0)]\boldsymbol{\epsilon} - \chi^2_{j(s)}}\right] \\ &\leq \frac{n\chi^2_{j(s)}}{\boldsymbol{\epsilon}^\tau[I - H(s_0)]\boldsymbol{\epsilon} - \chi^2_{j(s)}}. \end{aligned}$$

Let $b = p(p-1)/2$, $j = j_{\mathrm{M}} + j_{\mathrm{I}}$ and $\widetilde{m}_j = 2j_{\mathrm{M}}(\ln p + \ln(j \ln p)) + 2j_{\mathrm{I}}(\ln b + \ln(j \ln b))$. Let $\tilde{\mathcal{S}}_{j_{\mathrm{M}}j_{\mathrm{I}}}$ denote the collection of sets having j_{M} main-effect indices and j_{I} interaction indices and containing s_0. We have

$$P\left(\frac{\max_{j_{\mathrm{M}}, j_{\mathrm{I}}: j_{\mathrm{M}}+j_{\mathrm{I}}=j} \max_{s \in \tilde{\mathcal{S}}_{j_{\mathrm{M}}j_{\mathrm{I}}}} \chi^2_{j(s)}}{\widetilde{m}_j} \geq 1\right) \leq \sum_{j_{\mathrm{M}}+j_{\mathrm{I}}=j} \tau(\tilde{\mathcal{S}}_{j_{\mathrm{M}}j_{\mathrm{I}}}) P(\chi^2_j \geq \widetilde{m}_j).$$

Note that $\tau(\tilde{\mathcal{S}}_{j_{\mathrm{M}}j_{\mathrm{I}}}) = \binom{p-\nu(s_{0\mathrm{M}})}{j_{\mathrm{M}}}\binom{b-\nu(s_{0\mathrm{I}})}{j_{\mathrm{I}}} \leq p^{j_{\mathrm{M}}} b^{j_{\mathrm{I}}}$. Following the same argument in the verification of (17), we have

$$\tau(\tilde{\mathcal{S}}_{j_{\mathrm{M}}j_{\mathrm{I}}}) P(\chi^2_j \geq \widetilde{m}_j) \leq \frac{c}{\widetilde{m}_j} q_{\mathrm{M}}^{j_{\mathrm{M}}} q_{\mathrm{I}}^{j_{\mathrm{I}}} \leq \frac{c}{j \ln p} q_{\mathrm{M}}^{j_{\mathrm{M}}} q_{\mathrm{I}}^{j_{\mathrm{I}}},$$

where

$$q_{\mathrm{M}} = \sqrt{\frac{\widetilde{m}_j}{(j\ln p)^2}},\ q_{\mathrm{I}} = \sqrt{\frac{\widetilde{m}_j}{(j\ln b)^2}}.$$

When n is sufficiently large,

$$\max\{q_{\mathrm{M}}, q_{\mathrm{I}}\} \le \sqrt{\frac{4}{\ln p}(1+o(1))} \le q,$$

for some $0 < q < 1/2$. Thus,

$$\sum_{j_{\mathrm{M}}+j_{\mathrm{I}}=j} \tau(\tilde{\mathcal{S}}_{j_{\mathrm{M}}j_{\mathrm{I}}})P(\chi_j^2 \ge \widetilde{m}_j) \le \sum_{j_{\mathrm{M}}+j_{\mathrm{I}}=j} \frac{c}{j\ln p} q^j \le c(2q)^j,$$

and hence,

$$P\left(\max_{1\le j\le k-p_0} \frac{\max_{j_{\mathrm{M}},j_{\mathrm{I}}:j_{\mathrm{M}}+j_{\mathrm{I}}=j} \max_{s\in\tilde{\mathcal{S}}_{j_{\mathrm{M}}j_{\mathrm{I}}}} \chi^2_{j(s)}}{\widetilde{m}_j} \ge 1\right) \le \sum_{j=1}^{k-p_0} c(2q)^j < c\frac{2q}{1-2q} \to 0.$$

Thus, uniformly,

$$\max_{s\in\tilde{\mathcal{S}}_{j_{\mathrm{M}}j_{\mathrm{I}}}} \chi^2_{j(s)} = \widetilde{m}_j(1+o_p(1)).$$

Since $\ln b = 2\ln p(1+o(1))$, we have

$$\begin{aligned}\widetilde{m}_j &\le 2j_{\mathrm{M}}(\ln p + \ln((k-p_0)\ln p)) + 2j_{\mathrm{I}}(\ln b + \ln((k-p_0)\ln b))\\ &\le (2j_{\mathrm{M}} + 4j_{\mathrm{I}})\ln p(1+o_p(1)),\end{aligned}$$

because $\frac{\ln((k-p_0)\ln p)}{\ln p} \to 0$.

In addition, when $n\to\infty$, $n^{-1}\boldsymbol{\epsilon}^\tau[I-H(s_0)]\boldsymbol{\epsilon} \to \sigma^2 = 1$, that is, $\boldsymbol{\epsilon}^\tau[I-H(s_0)]\boldsymbol{\epsilon} = n(1+o(1))$, we have

$$\begin{aligned}&\frac{n\chi_j^2(s)}{\boldsymbol{\epsilon}^\tau[I-H(s_0)]\boldsymbol{\epsilon} - \chi_j^2(s)}\\ &\quad\le \frac{n\widetilde{m}_j}{n-\widetilde{m}_j(1+o_p(1))} = \widetilde{m}_j(1+o_p(1))\\ &\quad\le [2j_{\mathrm{M}}\ln p + 4j_{\mathrm{I}}\ln p](1+o_p(1)).\end{aligned}$$

Therefore,

$$T_1 \ge -[2j_{\mathrm{M}}\ln p + 4j_{\mathrm{I}}\ln p](1+o_p(1))$$

It follows from (11) that

$$T_2 = j \ln n + [2\gamma_{\mathrm{M}} j_{\mathrm{M}} \ln p + 4\gamma_{\mathrm{I}} j_{\mathrm{I}} \ln p](1 + o(1)).$$

Finally, we have,

$$\begin{aligned} T_1(s) + T_2(s) \geq\ & (j_{\mathrm{M}} + j_{\mathrm{I}}) \ln n + [2\gamma_{\mathrm{M}} j_{\mathrm{M}} \ln p + 4\gamma_{\mathrm{I}} j_{\mathrm{I}} \ln p](1 + o(1)) \\ & - [2 j_{\mathrm{M}} \ln p + 4 j_{\mathrm{I}} \ln p](1 + o_p(1)) \\ =\ & j_{\mathrm{M}} \ln n + 2\gamma_{\mathrm{M}} j_{\mathrm{M}} \ln p(1 + o(1)) - 2 j_{\mathrm{M}} \ln p(1 + o_p(1)) \\ & + j_{\mathrm{I}} \ln n + 4\gamma_{\mathrm{I}} j_{\mathrm{I}} \ln p(1 + o(1)) - 4 j_{\mathrm{I}} \ln p(1 + o_p(1)) \\ >\ & 0, \end{aligned}$$

if $\gamma_{\mathrm{M}} > 1 - \frac{\ln n}{2 \ln p}$, $\gamma_{\mathrm{I}} > 1 - \frac{\ln n}{4 \ln p}$, uniformly for all s such that $\nu(s) \leq k$, and $s_0 \subset s$, when n is sufficiently large. This verifies (12) in the case $s \in A_2$, and hence Theorem 1 is proved. □

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: B. N. Petrov, F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

Bailey, J., Grabowski-Boase, L., Steffy, B., Wiltshire, T., Churchill, G., Tarantino, L. (2008). Identification of quantitative trait loci for locomotor activation and anxiety using closely related inbred strains. *Genes, Brain and Behavior*, *7*(7), 761–769.

Baraud, Y. (2000). Model selection for regression on a fixed design. *Probability Theory and Related Fields*, *117*(4), 467–493.

Barron, A., Birgé, L., Massart, P. (1999). Risk bounds for model selection via penalization. *Probability theory and related fields*, *113*(3), 301–413.

Bien, J., Taylor, J., Tibshirani, R., et al. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, *41*(3), 1111–1141.

Bogdan, M., Ghosh, J. K., Doerge, R. (2004). Modifying the schwarz bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics*, *167*(2), 989–999.

Breheny, P., Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, *5*(1), 232–253.

Broman, K. W., Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society: Series B (Methodological)*, *64*(4), 641–656.

Chen, J., Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika.*, *95*(3), 759–771.

Chen, J., Chen, Z. (2012). Extended bic for small-n-large-p sparse glm. *Statistica Sinica*, *22*(2), 555.

Choi, N. H., Li, W., Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association.*, *105*(489), 354–364.

Clyde, M., Berger, J., Bullard, F., Ford, E., Jefferys, W., Luo, R., Paulo, R., Loredo, T. (2007). Current challenges in bayesian model choice. In: Astronomical Society of the Pacific Conference Series, ASP (vol. 371, p. 224).

Craven, P., Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, *31*(4), 377–403.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Fan, J., Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*, *99*(467), 710–723.

Fan, J., Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, *32*(3), 928–961.

Foygel, R., Drton, M. (2010). Extended bayesian information criteria for gaussian graphical models. arXiv:1011.6640.

Huang, J., Ma, S., Zhang, C. H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, *18*(4), 1603.

Luo, S., Chen, Z. (2013). Extended bic for linear regression models with diverging number of relevant features and high or ultra-high feature spaces. *Journal of Statistical Planning and Inference*, *143*, 497–504.

Luo, S., Chen, Z. (2014). Sequential lasso for feature selection with ultra-high dimensional feature space. *Journal of the American Statistical Association*, *109*, 1229–1240.

Luo, S., Xu, J., Chen, Z. (2014). Extended bayesian information criterion in the cox model with a high-dimensional feature space. *Annals of the Institute of Statistical Mathematics* (accepted).

Meinshausen, N., Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, *34*(3), 1436–1462.

Radchenko, P., James, G. M. (2010). Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association*, *105*(492), 1541–1553.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Siegmund, D. (2004). Model selection in irregular problems: applications to mapping quantitative trait loci. *Biometrika*, *91*(4), 785–800.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society Series B (Methodological)*, *36*(2), 111–147.

Storey, J. D., Akey, J. M., Kruglyak, L. (2005). Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biology*, *3*(8), e267.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological)*, *58*(1), 267–288.

Wong, J. (2013). Imputation, r version 2.0.1. https://github.com/jeffwong/imputation. Accessed 3 Apr 2012.

Xie, H., Huang, J. (2009). Scad-penalized regression in high-dimensional partially linear models. *The Annals of Statistics*, *37*(2), 673–696.

Yang, Y. (1999). Model selection for nonparametric regression. *Statistica Sinica*, *9*(2), 475–499.

Yang, Y., Barron, A. R. (1998). An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory*, *44*(1), 95–116.

Yuan, M., Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67.

Yuan, M., Joseph, V. R., Zou, H. (2009). Structured variable selection and estimation. *The Annals of Applied Statistics*, *3*(4), 1738–1757.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2), 894–942.

Zhao, J., Chen, Z. (2011). A two-stage penalized logistic regression approach to case-control genome-wide association studies. *Journal of Probability and Statistics, 2012*, Art ID 642403. doi:10.1155/2012/642403.

Zhao, P., Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, *7*, 2541–2563.

Zhao, P., Rocha, G., Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, *37*(6A), 3468–3497.

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

Zou, W., Zeng, Z. (2009). Multiple interval mapping for gene expression qtl analysis. *Genetica*, *137*(2), 125–134.