Statistica Sinica Preprint No: SS-2018-0253	
Title	Sequential interaction group selection by the principle of correlation search for high-dimensional interaction models
Manuscript ID	SS-2018-0253
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0253
Complete List of Authors	Shan Luo and Zehua Chen
Corresponding Author	Zehua Chen
E-mail	stachenz@nus.edu.sg
Notice: Accepted version subject to English editing.	

Sequential interaction group selection by the principle of correlation search for high-dimensional interaction models

Shan Luo[1] AND Zehua Chen[2]

[1]Shanghai Jiao Tong University

[2]National University of Singapore

Email: [1]sluomath@sjtu.edu.cn, [2]stachenz@nus.edu.sg.

Abstract

High-dimensional interaction models have important applications in many scientific fields, especially, in genetic research and medical studies. Like in high-dimensional main-effect models, feature selection is unavoidable in high-dimensional interaction models. However, feature selection methods for high-dimensional main-effect models cannot be directly applied to high-dimensional interaction models because of imbalanced spurious correlations among main-effect features and interaction features. The major idea which has dominated the methods for high-dimensional interaction models in the literature is to impose a so-called hierarchy principle through various mechanisms. However, the imposition of the hierarchy principle is questionable, as we argue in this article. In this paper, we propose a sequential interaction group selection (SIGS) method based on the principle of correlation search. The SIGS method avoids the drawbacks due to the imposition of the hierarchy principle and has desirable properties. The selection consistency of the SIGS method is established. Simulation studies demonstrate that the SIGS method has an edge over those methods which impose the hierarchy principle.

Key Words: feature selection, group search, hierarchy principle, high-dimensional interaction models, principle of correlation search, sequential procedure.

1 Introduction

High-dimensional models arise from many conventional scientific fields such as genetic research, medical studies, financial analysis, web information analysis, etc. The problem of feature selection is an indispensable part in the analysis of high-dimensional models. There have been volumes of research papers contributed to this important problem. For some seminal ones, see Tibshirani (1996), Fan and Li (2001), Zou (2006), Zhang (2010), Yuan and Lin (2006), and so on. However, the major focus in the literature has been on the so-called main-effect models. High-dimensional models with interactions have drawn relatively less attentions. Interaction effect is quite a common phenomena in medical and genetic studies. For example, 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). In genetics, the effects of many genes are realized only through their interaction with other genes, see Evans et al. (2006), Manolio and Collins (2007), Kooperberg and LeBlanc (2008), Cordell (2009), etc.. Therefore, people are obliged to consider interaction models. There have been a few papers devoted to high-dimensional interaction models, which we will review later, but there are still many issues to be addressed.

An interaction model with pairwise interaction effects can be formulated as:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \sum_{1 \le j < k \le p} \theta_{jk} X_j X_k + \epsilon, \tag{1}$$

where Y is the response variable and $X_1, \ldots, X_p$ are p covariates. We refer to each X_j as a main-effect feature and to each product $X_j X_k$ as an interaction feature. The interaction model cannot be treated as an augmentation of a main-effect model by considering the interaction features simply as additional covariates. The spurious correlations among the interaction features are much higher than those among the main-effect features since the number, $p(p-1)/2$, of the interaction features is much larger than the number, p, of the main-effect features. Because of the higher spurious correlations, much more false interaction features can be easily selected by a feature selection procedure designed

for main-effect models. The effect of relevant main-effect features might be masked by the false selection of irrelevant interaction features. In the analysis of the CGEMS prostate cancer data which is available at `http://cgems.cancer.gov`, a SCAD penalized likelihood approach, which treats the main-effect features and the interaction features on the same footing, is applied in Zhao and Chen (2011) but only certain interaction features are discovered, though significant main-effect features are identified in other studies, see, e.g., Yeager et al. (2007). Special considerations must be taken into account in dealing with high-dimensional interaction models.

The consideration of a so-called hierarchy principle has dominated the methods developed in the literature for the analysis of high-dimensional interaction models. The hierarchy principle requires that if an interaction feature, X_kX_j, is included in a model then either at least one of or both its parent main-effect features, X_j and X_k, must also be included in the model. If only the inclusion of at least one parent main-effect feature is required, it is referred to as the weak hierarchy principle, otherwise, it is referred to as the strong hierarchy principle. Different methods developed so far are essentially different ways for enforcing either the strong or weak hierarchy in the procedure of feature selection. One methodology for enforcing the hierarchy principle is through certain group-Lasso penalties in penalized likelihood approaches, see, e.g., Zhao et al. (2009), Yuan et al. (2009), Choi et al. (2010), Radchenko and James (2010), Bien et al. (2013), She et al. (2016), etc. Two typical such penalties are given as follows. The first one is considered in Zhao et al. (2009):

$$p_1(\Theta, \boldsymbol{\beta}) = \lambda_1 \sum_{j \neq k} |\theta_{jk}| + \lambda_2 \sum_{j \neq k} \|(\beta_j, \beta_k, \theta_{jk})\|_{\gamma_{jk}},$$

where γ_{jk} are some positive constants, in particular, they can be the same and equal to 2. The other one is considered in She et al. (2016):

$$p_2(\Theta, \boldsymbol{\beta}) = \lambda_1 \sum_{j \neq k} |\theta_{jk}| + \lambda_2 \sum_{j=1}^{p} \|(\beta_j, \theta_{j1}, \ldots, \theta_{jp})\|_2.$$

A slightly different approach called hierarchical Lasso was considered in Bien et al. (2013), which is

a relaxed convex version of the following non-convex problem:

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=1}^{n}\left[y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}-\frac{1}{2}\sum_{j\neq k}\theta_{jk}x_{ij}x_{ik}\right]^2+\lambda\left[\sum_{j=1}^{p}|\beta_j|+\frac{1}{2}\sum_{j\neq k}|\theta_{jk}|\right], \\ \text{subject to} \quad & \|\boldsymbol{\theta}_j\|_1 \le |\beta_j|, \end{aligned}$$

where $\boldsymbol{\theta}_j = (\theta_{j1},\ldots,\theta_{jp})^\top$.

A different methodology for enforcing the hierarchy principle is through multi-step selections. Two schemes of a forward regression approach with this nature are considered in Hao and Zhang (2014). In the first scheme, the main-effect features are selected first, then only those interaction features whose parent main-effect features are selected are subjected to further selection. In the second scheme, at the beginning, the candidate set is confined to the collection of all main-effect features, at the subsequent steps, once the parents of an interaction feature are selected, the interaction feature is added to the candidate set. A method called the Regularization Algorithm under Marginality Principle (RAMP) is considered in Hao et al. (2018). The RAMP consists of steps determined by a sequence of values of the penalty parameter in decreasing order, at each step, the main-effect features already selected are not penalized and all the remaining main-effect features and those interaction features whose parents are among the selected main-effect features are included in the model and penalized with the penalty parameter value at that step. Then the set of selected main-effect features is augmented by newly selected main-effect features and the parents of newly selected interaction features. We refer to the above methods as the search-main-effect-first approaches.

By enforcing the hierarchy principle, the above approaches address the issue of imbalanced spurious correlations mentioned in the second paragraph, i.e., the much higher spurious correlations among the interaction features make interaction features easier to be selected. In fact, these methods either impose heavier penalties on interaction features or make the selection of the parent main-effect features a premise for the selection of an interaction feature. In effect, this makes the selection of interaction features harder than that of main-effect features. However, there are some

unwanted natures of these approaches, which we will discuss in the next section.

A method which does not impose the hierarchy principle is considered in He and Chen (2014). The method is a sequential search procedure. At each step of the procedure, the main-effect feature which is most correlated with the current residual among all un-selected main-effect features and the interaction feature which is most correlated with the current residual among all un-selected interaction features are identified first, then a version of EBIC (Chen and Chen, 2008) for interaction models is used to select between the two identified features. The EBIC version for interaction models imposes a heavier penalty on interaction features than that on main-effect features, which is an alternative way to address the issue discussed above. But this method separates the main-effect features and the interaction features and ignores the intrinsic connections between an interaction feature and its parent main-effect features, which, as we will see in our simulation studies, has a potential to reduce the accuracy of feature selection.

In this article, we develop a sequential interaction group selection (SIGS) method. In what follows, a single main-effect or interaction feature is referred to as a simple feature. A group of simple features is referred to as a composite feature. In our SIGS method, we consider composite features of the form $\{X_k, X_j, X_jX_k\}$. Our method consists of two sequential procedures. The first procedure selects composite features sequentially by the principle of correlation search which will be discussed later. The second procedure select sequentially simple features among those contained in the selected composite features. In both procedures, the EBIC is used as the stopping rule. The SIGS method overcomes the unwanted natures of the methods reviewed above, which will be elaborated later. Furthermore, under certain mild conditions, the method is selection consistent. Simulation studies demonstrate that the SIGS method has an edge over the other existing methods in terms of the accuracy of feature selection. Another advantage of our method is that its applicability is not limited by the dimension of the data.

The remainder of the article is arranged as follows. In section 2, we develop the sequential inter-

action group selection method and provide its computation algorithms and theoretical properties. In section 3, we report our simulation studies for the comparison of our method with other available methods. In section 4, we consider a real example. Technical proof of theoretical results is provided in a supplementary document.

2 Sequential interaction group selection by the principle of correlation search

In this section, some issues on the hierarchy principle are addressed first, which provides us with a motivation. Then the sequential interaction group selection (SIGS) method is developed and its theoretical properties are provided.

2.1 Issues on the hierarchy principle

First we make a clarification on the concepts of main effect and marginal effect of a covariate in a linear model. In general, these two concepts are not the same. The marginal effect is the effect of the covariate averaging over the other covariates. In a model contains only main-effect terms, the main effect and marginal effect are identical. However, in an interaction model like (1), the so-called main effect is not the marginal effect. In fact, the meaning of the main effect is ambiguous and dependent of the centers of the covariates. A similar clarification has been made in Hao and Zhang (2017). Here, we illustrate this by a simple interaction model consisting of two binary covariates which represent two factors A and B, each of two levels. Let $x_1 = 1$ if A is at level 2; 0 otherwise. Similarly, let $x_2 = 1$ if B is at level 2; 0 otherwise. Consider the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \theta_{12} x_1 x_2 + \epsilon. \tag{2}$$

At different values of (x_1, x_2), the model can be expressed as follows.

$$y = \begin{cases} \beta_0 + \epsilon, & \text{if both A and B are at level 1;} \\ \beta_0 + \beta_1 + \epsilon, & \text{if A is at level 2 and B is at level 1;} \\ \beta_0 + \beta_2 + \epsilon, & \text{if A is at level 1 and B is at level 2;} \\ \beta_0 + \beta_1 + \beta_2 + \theta_{12} + \epsilon, & \text{if both A and B are at level 2.} \end{cases}$$

Assume that the number of observations at each of the four level combinations is the same. Then we have the following. The effect of A within level 1 of B is $(\beta_0+\beta_1)-\beta_0=\beta_1$; the effect of A within level 2 of B is $(\beta_0+\beta_1+\beta_2+\theta_{12})-(\beta_0+\beta_2)=\beta_1+\theta_{12}$, the marginal effect (the average effect) of A is $\frac{1}{2}[\beta_1+(\beta_1+\theta_{12})]=\beta_1+\frac{1}{2}\theta_{12}$, and the interaction effect between A and B is $[(\beta_1+\theta_{12})-\beta_1]=\theta_{12}$. In fact, the so-called main effect of x_1, i.e. β_1, is the effect of A within level 1 of B which is not the same as the marginal effect of A. Furthermore, if we make linear transformations: $z_1=x_1+c$ and $z_2=x_2+d$, $c,d\neq 0$, and consider the model

$$y=\beta_0+\beta_1 z_1+\beta_2 z_2+\theta_{12} z_1 z_2+\epsilon, \tag{3}$$

then it is easy to see that β_1 is no longer the effect of A within level 1 of B, instead, this effect is given by $\beta_1+\theta_{12}d$. After all, the so-called main effect in an interaction model does not reflect the importance of the corresponding covariate.

It is because of the problem caused by an arbitrary centering that the hierarchy principle was imposed in interaction modeling by early authors, see, Nelder (1977) and McCullagh and Nelder (1983). However, the indefiniteness of the so-called main-effect in an interaction model is not a major issue. It can be solved by a standardization of the covariates. Furthermore, this issue is indeed irrelevant in feature selection, since, in a particular problem, the features (the covariates with fixed scales or the centers) are given, we only care about whether or not the main-effect terms and interaction terms of the given features should be selected.

The drawback of the hierarchy principle is that by enforcing main-effect features into a model when they are indeed irrelevant it unnecessarily causes higher variation in the fitted model, which has a detrimental impact on feature selection. The hierarchy principle in the search-main-effect-first approaches is even more problematic. At the first step of a search-main-effect-first approach, it actually selects the features based on the marginal effects of the covariates, since the main effect and marginal effect are the same when interaction features are not involved. As illustrated by model (2), the marginal effect of A, $\beta_1+\frac{1}{2}\theta_{12}$, is the average of the effect of A at the two levels of B. If the effect

of A at the two levels are in different directions, then β_1 and θ_{12} have different signs, the marginal effect could be weaker than the interaction effect. In the extreme case, the marginal effect could be zero while the interaction effect is in fact substantial. Because of a weaker or zero marginal effect, the covariate might not be selected no matter how strong its interaction with another covariate is. Thus the interaction feature involving such covariates will have no chance to be selected at the second step. This is a more serious problem than that caused by enforcing irrelevant features into the model. The search-main-effect-first approaches implicitly take the marginal (or main) effect as an indication of the importance of a covariate. But, as we have argued above, a covariate having negligible marginal effect could have substantial interaction effect with other covariates. To address this issue, recently, Hao et al. (2018) proposed the notion of important predictor. By this notion, an important predictor is a covariate with either a nonzero main-effect or any nonzero interaction effect with other covariates. A reasonable feature selection approach should be able to select at least the important predictors.

The major issue with high-dimensional interaction models is what we mentioned in the introduction, i.e., the interaction features can be selected more easily than main-effect features because of their imbalanced spurious correlations. How can we address the issue caused by the imbalanced spurious correlations and at the same time avoid the problems discussed above? This is what our SIGS procedure aims to answer.

2.2 The principle of correlation search and the SIGS method

The SIGS method which we propose is based on what we called the principle of correlation search. The principle of correlation search is statistically desirable and appealing. The intrinsic mechanism for feature selection is in fact the correlation of the features with the response variable. In penalized least squares approaches, at each fixed value of the penalty parameter, the active set of the features is indeed the set of features whose correlation with the response exceed a certain threshold. In sequential procedures such as the least angle regression (LAR), Efron et al. (2004), the orthogonal

matching pursuit (OMP), Cai and Wang (2011), the forward stepwise regression (FSR), Wang (2009), and the sequential LASSO, Luo and Chen (2014), at each step, the next feature is selected according to the Pearson's correlation coefficients of the unselected features with the residual of the current model. The differences among the various sequential procedures lie only in how the current model is fitted. The problem of feature selection is essentially to select features to which the unexplained variation of the response can be attributed. The capacity of a feature to explain the variation of a variable can be measured by its correlation with that variable. The principle of correlation search is that, whenever there is unexplained variation of the response, features should be selected according to their correlation with the unexplained part of the response. There are two requirements for the application of the principle of correlation search. First, the candidate features must have the same status except their unknown relations to the response. Second, there must be an appropriate measure of correlation.

In high-dimensional interaction models, the principle of correlation search cannot be applied directly to simple main-effect and interaction features because of their imbalanced spurious correlations. In other words, the main-effect features and the interaction features do not have the same status. However, if we consider the composite features $\{X_j, X_k, X_jX_k\}$ instead of simple main-effect and interaction features, the imbalance in spurious correlations no longer matters, since all these composite features have the same status. What left is an appropriate correlation measure. In the current context, a ready choice of correlation measure is the multiple correlation coefficient which measures the correlation between a scaler random variable and a random vector. Let Y be the scaler random variable and $\boldsymbol{z}$ the random vector. The multiple correlation coefficient between Y and $\boldsymbol{z}$ is given by

$$\rho^2 = \max_{\boldsymbol{a}}[\mathrm{corr}(Y, \boldsymbol{a}^\top \boldsymbol{z})]^2 = \frac{\Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy}}{\sigma_y^2},$$

where Σ_{yz} is the covariance vector between Y and $\boldsymbol{z}$, $\Sigma_{zy} = \Sigma_{yz}^\top$, Σ_{zz} is the covariance matrix of $\boldsymbol{z}$, and σ_y^2 is the variance of Y. When the selection of $\boldsymbol{z}$ is of concern, we can ignore the factor σ_y^2.

Therefore we can take $R(y, \boldsymbol{z}) = \Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy}$ as the correlation measure.

For the interaction model (1) with n observations, it is expressed as:

$$\boldsymbol{y} = \beta_0\mathbf{1} + \sum_{j=1}^{p}\beta_j\boldsymbol{x}_j + \sum_{1\le j<k\le p}\theta_{jk}\boldsymbol{x}_j \circ \boldsymbol{x}_k + \boldsymbol{\epsilon},$$

where $\boldsymbol{x}_j \circ \boldsymbol{x}_k$ denotes the component-wise product, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ is a random vector of i.i.d. components with mean zero. Denote by $\mathcal{Z}_{jk}$ the composite feature $\{X_j, X_k, X_jX_k\}$. We will also consider $\mathcal{Z}_{jk}$ as the set consisting of X_j, X_k and X_jX_k. Let Z_{jk} denote the matrix $(\boldsymbol{x}_j, \boldsymbol{x}_k, \boldsymbol{x}_j \circ \boldsymbol{x}_k)$. Let $\mathcal{Z}$ be any set of simple features. The cardinality of $\mathcal{Z}$ is denoted by $|\mathcal{Z}|$. $H(\mathcal{Z})$ denotes the projection matrix of the space spanned by the observed vectors of the simple features in $\mathcal{Z}$. For example, $H(\mathcal{Z}_{jk}) = Z_{jk}(Z_{jk}^\top Z_{jk})^{-1}Z_{jk}^\top$. Let $\tilde{\boldsymbol{y}}$ be a generic notation for the residual vector of $\boldsymbol{y}$ fitted to a linear model. Suppose that all the simple main-effect and interaction feature vectors are centered. The correlation measure between the residual and $\mathcal{Z}_{jk}$ is given by

$$r(\tilde{\boldsymbol{y}}, Z_{jk}) = \frac{1}{n}\tilde{\boldsymbol{y}}^\top Z_{jk}\left(Z_{jk}^\top Z_{jk}\right)^{-1}Z_{jk}^\top\tilde{\boldsymbol{y}} = \frac{1}{n}\tilde{\boldsymbol{y}}^\top H(\mathcal{Z}_{jk})\tilde{\boldsymbol{y}}.$$

The correlation measure has another geometric interpretation: it is indeed the squared norm of the residual vector projected onto the space spanned by the columns of Z_{jk} scaled by $1/n$.

As briefly mentioned in the introduction, the SIGS method consists of two sequential procedures. The first one selects composite features and the second one selects simple features contained in the composite features selected in the first procedure. In the following, we describe the detailed algorithm for the SIGS method.

SIGS Algorithm

Selection of Composite Features:

Let $\tilde{\boldsymbol{y}} = \boldsymbol{y}$, $\mathcal{Z}^* = \emptyset$. Repeat

(i) For un-selected Z_{jk}, find j^* and k^* such that $r(\tilde{\boldsymbol{y}}, Z_{j^*k^*}) = \max_{jk} r(\tilde{\boldsymbol{y}}, Z_{jk})$.

(ii) If $\text{EBIC}(\mathcal{Z}^* \cup \mathcal{Z}_{j^*k^*}) < \text{EBIC}(\mathcal{Z}^*)$, update $\mathcal{Z}^*$ to $\mathcal{Z}^* = \mathcal{Z}^* \cup \mathcal{Z}_{j^*k^*}$ and $\tilde{\boldsymbol{y}} = [I - H(Z^*)]\boldsymbol{y}$; otherwise, stop.

Selection of Simple Features:

Denote the simple feature vectors in $\mathcal{Z}^*$ by $\{\boldsymbol{z}_i, i = 1, \ldots, |\mathcal{Z}^*|\}$. Let $\tilde{\boldsymbol{y}} = \boldsymbol{y}$, $\mathcal{Z}^{**} = \emptyset$. Repeat

(i) For un-selected $\boldsymbol{z}_i$ in $\mathcal{Z}^*$, find i^* such that $|\text{corr}(\tilde{\boldsymbol{y}}, \boldsymbol{z}_{i^*})| = \max_i |\text{corr}(\tilde{\boldsymbol{y}}, \boldsymbol{z}_{i^*})|$.

(ii) If $\text{EBIC}(\mathcal{Z}^{**} \cup \boldsymbol{z}_{i^*}) < \text{EBIC}(\mathcal{Z}^{**})$, update $\mathcal{Z}^{**}$ to $\mathcal{Z}^{**} = \mathcal{Z}^{**} \cup \boldsymbol{z}_{i^*}$ and $\tilde{\boldsymbol{y}} = [I - H(\mathcal{Z}^{**})]\boldsymbol{y}$; otherwise, stop.

The EBIC in the above algorithm has slightly different forms for the composite feature selection and the simple feature selection. For the composite feature selection, suppose $\mathcal{Z}^*$ is the union of m selected composite features, it is given by

$$\text{EBIC}(\mathcal{Z}^*) = n \ln \frac{1}{n} \|[I - H(\mathcal{Z}^*)]\boldsymbol{y}\|_2^2 + |\mathcal{Z}^*| \ln n + 2\gamma \binom{N}{m},$$

where $N = p(p-1)/2$ and $\gamma = 1 - \frac{\ln n}{2 \ln N}$. For the simple feature selection, it is given by

$$\text{EBIC}(\mathcal{Z}^{**}) = n \ln \frac{1}{n} \|[I - H(\mathcal{Z}^{**})]\boldsymbol{y}\|_2^2 + |\mathcal{Z}^{**}| \ln n + 2\gamma \binom{|\mathcal{Z}^*|}{|\mathcal{Z}^{**}|},$$

where $\gamma = 1 - \frac{\ln n}{2 \ln |\mathcal{Z}^*|}$.

In the composite feature selection procedure, a composite feature is selected according to the joint effect of its constituent simple features, the selection is neither based on the marginal effects alone nor on the interaction alone. A composite feature can be selected in either of the following possible cases: either or both the two marginal effects are substantial but there is no interaction, the marginal effects are negligible but the interaction is substantial, and both the marginal effects and the interaction are substantial. In the simple feature selection procedure following the composite feature selection, irrelevant simple features, either main-effect or interaction features, are eliminated, which is not restricted by the hierarchy principle. Therefore, the SIGS method avoids the drawbacks of those methods which impose the hierarchy principle. Intuitively, the SIGS method would have a potential edge over those methods. This is indeed vindicated in our simulation studies reported in Section 3.

Because of its sequential nature, the implementation of the SIGS method is not limited by the dimension of the data, i.e., the number of covariates p. The only concern is that the computation might take a long time when p is very large. However, this issue can be solved by a pre-screening procedure with sure screening property. In the following, we propose two such screening methods. In both methods, the main-effect features and interaction features are screened separately. The final screened-out set of covariates consists of the screened-out main-effect features and the parents of the screened-out interaction features. The sure independence screening (SIS) of Fan and Lv (2008) is used for main-effect feature screening in both methods. In one method, the interaction features are screened by their direct correlations with the response which is referred to as direct interaction screening (DIS). In the other method, the interaction features are screened by their partial correlations with the response adjusting for the effects of their parents which is referred to as interaction screening by partial correlation (ISPC). The ISPC approach is proposed in Niu et al. (2018). The two screening methods are dubbed as SIS+DIS and SIS+ISPC.

SIS+DIS Method:

(i) Compute $r_j = |\operatorname{Corr}(\boldsymbol{y}, \boldsymbol{x}_j)|$ for $j = 1, \ldots, p$. Denote the $[c_{\mathrm{M}} n/\ln n]$'s largest r_j by $r_{([c_{\mathrm{M}} n/\ln n])}$. Keep the set $S_{\mathrm{M}} = \{X_j : r_j \geq r_{([c_{\mathrm{M}} n/\ln n])}\}$.

(ii) Compute $\rho_{ij} = |\operatorname{Corr}(\boldsymbol{y}, \boldsymbol{x}_i \circ \boldsymbol{x}_j)|$ for $1 \leq i < j \leq p$. Denote the $[c_{\mathrm{I}} n/\ln n]$'s largest ρ_{ij} by $r_{([c_m n \ln n])}$. Keep the set $S_{\mathrm{I}} = \{\{X_i, X_j\} : \rho_{ij} \geq r_{([c_{\mathrm{I}} n/\ln n])}\}$.

(iii) Take the set of retained covariates as $S = S_{\mathrm{M}} \cup S_{\mathrm{I}}$.

SIS+ISPC Method:

The same as SIS+DIS except (ii) is replaced by

(ii)$'$ Compute $\rho_{ij} = |\operatorname{Corr}((I - H_{ij})\boldsymbol{y}, (I - H_{ij})\boldsymbol{x}_i \circ \boldsymbol{x}_j)|$ for $1 \leq i < j \leq p$, where H_{ij} is the projection matrix of $(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Denote the $[c_{\mathrm{I}} n/\ln n]$'s largest ρ_{ij} by $r_{([c_m n \ln n])}$. Keep the set $S_{\mathrm{I}} = \{\{X_i, X_j\} : \rho_{ij} \geq r_{([c_{\mathrm{I}} n/\ln n])}\}$.

Under certain mild conditions, both the SIS+DIS and SIS+ISPC methods possess the sure screening property, that is, with probability tending to 1, the retained set S contains all important covariates in the sense of Hao et al. (2018). The SIS+ISPC takes more computation time than SIS+DIS. In our simulation studies, both methods are comparable in terms of final results of the feature selection procedures. In some simulation settings, SIS+ISPC is slightly better than SIS+DIS.

2.3 The asymptotic properties of the SIGS method

Denote by $\mathcal{Z}_0$ the set of all relevant composite features and $\mathcal{Z}_0^*$ any subset of $\mathcal{Z}_0$. Let $\Sigma_{\mathcal{Z}_{jk}}$ be the covariance matrix of $\mathcal{Z}_{jk}$. Define the residual of Y adjusting for the effects of $\mathcal{Z}_0^*$ as $\tilde{Y}(\mathcal{Z}_0^*) = Y - \alpha - \boldsymbol{\eta}_0^\top \mathcal{Z}_0^*$ where α and $\boldsymbol{\eta}_0$ minimize $E(Y - \alpha - \boldsymbol{\eta}_0^\top \mathcal{Z}_0^*)^2$, that is, the residual is the difference between Y and its best linear predictor in terms of $\mathcal{Z}_0^*$. It turns out that $\boldsymbol{\eta}_0 = \Sigma_{Z_0^* Z_0^*}^{-1} \Sigma_{Z_0^* y}$ where $\Sigma_{Z_0^* Z_0^*}$ is the variance matrix of $\mathcal{Z}_0^*$ and $\Sigma_{Z_0^* y}$ is the covariance vector between $\mathcal{Z}_0^*$ and Y. Let $\Sigma_{\tilde{y} z_{jk}}(\mathcal{Z}_0^*)$ be the covariance vector between $\mathcal{Z}_{jk}$ and the residual $\tilde{Y}(\mathcal{Z}_0^*)$. The multiple correlation coefficient between Z_{jk} and $\tilde{Y}(\mathcal{Z}_0^*)$ is given by

$$R(\tilde{Y}(\mathcal{Z}_0^*), Z_{jk}) = \Sigma_{\tilde{y} Z_{jk}}(\mathcal{Z}_0^*) \Sigma_{\mathcal{Z}_{jk}}^{-1} \Sigma_{Z_{jk} \tilde{y}}(\mathcal{Z}_0^*),$$

where $\Sigma_{Z_{jk} \tilde{y}}(\mathcal{Z}_0^*) = \Sigma_{\tilde{y} Z_{jk}}^\top(\mathcal{Z}_0^*)$. By an abuse of notation, we also denote the index sets of the composite features in $\mathcal{Z}_0$ and $\mathcal{Z}_0^*$, respectively, by $\mathcal{Z}_0$ and $\mathcal{Z}_0^*$. Furthermore, denote by s any set of simple features, s_0 the set of all relevant simple features, and s_0^* the set of relevant simple features contained in $\mathcal{Z}_0^*$.

The following lemma establishes the property of the correlation measure.

Lemma 1. Assume the following conditions:

A1. $|s_0|^3 \ln p / n \to 0$.

A2. The eigenvalues of $\{\Sigma_{s,s} : |s| \leq 3|s_0|\}$ are bounded from below and above.

A3. Denote by Z_j's all the simple features. $\max_{j,l}\{\mathrm{E}\exp(t(Z_j-\mathrm{E}Z_j)(Z_l-\mathrm{E}Z_l), \mathrm{E}\exp(t\epsilon^2)\} \leq C$ for all $|t| \leq \eta$ for some constants η and C.

Suppose that $\mathcal{Z}_{j^*k^*}$ is the composite feature such that $R(\tilde{Y}(\mathcal{Z}_0^*), \mathcal{Z}_{j^*k^*}) = \max_{(j,k)\in(\mathcal{Z}_0^*)^c} R(\tilde{Y}(\mathcal{Z}_0^*), \mathcal{Z}_{jk})$. Then, as $n \to \infty$, uniformly for all $\mathcal{Z}_0^* \subset \mathcal{Z}_0$ with $|\mathcal{Z}_0^*| \leq 3|s_0|$, we have

$$P\left(r(\tilde{\boldsymbol{y}}(Z_0^*), Z_{j^*k^*}) = \max_{(j,k)\in(Z_0^*)^c} r(\tilde{\boldsymbol{y}}(Z_0^*), Z_{jk})\right) \to 1,$$

where $r(\tilde{\boldsymbol{y}}(Z_0^*), Z_{jk}) = (1/n)\tilde{\boldsymbol{y}}^\top(Z_0^*)H(Z_{jk})\tilde{\boldsymbol{y}}(Z_0^*)$ with $\tilde{\boldsymbol{y}}(Z_0^*) = [I - H(Z_0^*)]\boldsymbol{y}$ is the sample version of $R(\tilde{Y}(\mathcal{Z}_0^*), \mathcal{Z}_{j^*k^*})$.

In our SIGS method, the feature selection mechanism is the intrinsic correlations between the response and the features. The above lemma implies that this mechanism can be effected through the sample version of the correlation measure. In what follows, we establish the property of selection consistency for the SIGS method. We assume the following conditions.

B1. As $n \to +\infty$,

$$\sqrt{n}\min_{j\in s_0}|\xi_j|/\sqrt{|s_0|\ln p} \to +\infty,$$

where ξ_j's are the coefficients of the simple features in s_0.

B2. For all $\mathcal{Z}_0^* \subset \mathcal{Z}_0$ with $|\mathcal{Z}_0^*| \leq 3|s_0|$, denote $s_z^\star$ as the set of relevant simple features contained in $\mathcal{Z}_0^*$ and $s_z^{\star-} = s_0 \backslash s_z^\star$. Define $\tilde{\mathcal{Z}}_0^\star = \{Z_{jk} : Z_{jk} \notin \mathcal{Z}_0^*,\ Z_{jk} \cap s_z^{\star-} \neq \emptyset\}$. There exists a $0 < q_1 < 1$ such that

$$\max_{(j,k):\mathcal{Z}_{jk}\notin\tilde{\mathcal{Z}}_0^\star} R(\tilde{Y}(\mathcal{Z}_0^*), Z_{jk}) < q_1 \max_{(j,k):\mathcal{Z}_{jk}\in\tilde{\mathcal{Z}}_0^\star} R(\tilde{Y}(\mathcal{Z}_0^*), Z_{jk}).$$

B3. There exists a $0 < q_2 < 1$, such that for any $s \subset s_0$,

$$\max_{j\in s_0^c}|(\Sigma_{js_0} - \Sigma_{js}\Sigma_{ss}^{-1}\Sigma_{ss_0})\boldsymbol{\xi}| < q_2 \max_{j\in s^-}|(\Sigma_{js_0} - \Sigma_{js}\Sigma_{ss}^{-1}\Sigma_{ss_0})\boldsymbol{\xi}|,$$

where $\boldsymbol{\xi}$ is the coefficient vector of all the simple features in s_0.

Theorem 1. Assume conditions A1–A3 and B1 – B3. Let s^* be the selected set of simple features by the procedures of SIGS. Then, we have, as $n \to \infty$, $P(s^* = s_0) \to 1$.

We end this sub-section by some remarks on the major conditions of the theorem. If s_0 and p are fixed, B1 is automatically true; otherwise, it requires that, for the relevant features to be detectable, their effects must not tapper off too quickly. To explain B2, express $R(\tilde{Y}(\mathcal{Z}_0^*), Z_{jk})$ (with dependence on $\mathcal{Z}_0^*$ suppressed) as

$$\begin{aligned} R(\tilde{Y}, Z_{jk}) &= \Sigma_{\tilde{y}Z_{jk}}\Sigma_{\mathcal{Z}_{jk}}^{-1}\Sigma_{z_{jk}\tilde{y}} = \mathrm{Cov}(\tilde{Y}, Z_{jk}^\top)\Sigma_{\mathcal{Z}_{jk}}^{-1/2}\Sigma_{\mathcal{Z}_{jk}}^{-1/2}\,\mathrm{Cov}(Z_{jk}, \tilde{Y}) \\ &= \mathrm{Cov}(\tilde{Y}, Z_{jk}^\top\Sigma_{\mathcal{Z}_{jk}}^{-1/2})\,\mathrm{Cov}(\Sigma_{\mathcal{Z}_{jk}}^{-1/2}Z_{jk}, \tilde{Y}) = \|\,\mathrm{Cov}(\Sigma_{\mathcal{Z}_{jk}}^{-1/2}Z_{jk}, \tilde{Y})\|_2^2, \end{aligned}$$

where $\Sigma_{\mathcal{Z}_{jk}}^{-1/2}Z_{jk}$ is the standardized Z_{jk}, that is, $R(\tilde{Y}(\mathcal{Z}_0^*), Z_{jk})$ is the sum of squared Pearson's correlation coefficients of the components of the standardized Z_{jk} with $\tilde{Y}$. Therefore, B2 actually requires that, among the un-selected composite features, the maximum correlation of the relevant ones with the current residual should be larger than that of the irrelevant ones. B3 is a similar condition required of simple relevant features. As argued in Luo and Chen (2014), in the selection of simple features, condition B3 is indeed weaker than the irrepresentability condition required of Lasso. Luo and Chen (2014) provided examples where condition B3 holds but the irrepresentability condition does not. Since B2 is a straightforward extension of B3 to the case of composite features, we could reasonably believe that it is also weaker than the irrepresentability condition. Thus, conditions B1 — B3 are all reasonable minimal requirements.

The proof of the results in this sub-section is given in the supplementary document.

3 Simulation studies

We compare the performance of the SIGS method with some representative existing methods by simulation studies in this section. The following methods are considered in the comparison: (i) the sequential interactive EBIC procedure (SIEP) proposed in He and Chen (2014), (ii) the RAMP

methods proposed in Hao et al. (2018). The reason we choose these methods for comparison is that the RAMP is the most recent existing method which imposes the hierarchy principle and the SIEP method is the only existing method which does not impose the hierarchy principle. There are four versions of RAMP: strong and weak hierarchy structure with LASSO penalty which we refer to as RAMP-sL and RAMP-wL respectively, strong and weak hierarchy structure with MCP penalty which we refer to as RAMP-sM and RAMP-wM respectively. The RAMP methods are implemented by using the R package `RAMP`.

The performance of the methods is evaluated by the positive discovery rate (PDR) and false discovery rate (FDR) separately for main-effect features, denoted by MPDR and MFDR, and for interaction features, denoted by IPDR and IFDR. The PDR and FDR are defined as follows. Let s_0 and s^* be, respectively, the set of true features and the set of selected features.

$$\text{PDR} = \frac{|s^* \cap s_0|}{|s_0|}, \quad \text{FDR} = \frac{|s^* \cap s_0^c|}{|s^*|}.$$

The simulation settings are described in the following.

(i) Correlation structure of the covariates. The covariates are generated as random variables with mean zero, variance 1 and different correlation structures as follows.

XS1: For $p = 80$, the covariates are components of two independent random vectors with 50-variate and 30-variate, respectively, normal distributions having equal pairwise correlation 0.5. For $p = 200$, the covariates are components of four independent random vectors with 50-variate normal distribution having equal pairwise correlations 0.5.

XS2: The covariates are components of a p-variate normal distribution with exponentially decaying correlations $\rho_{ij} = 0.5^{|i-j|}$.

XS3: The covariates are generated as

$$\begin{aligned} X_j &= \frac{1}{\sqrt{5}} Z_0 + \frac{2}{\sqrt{5}} Z_j, \ 1 \le j \le p_0 \\ X_j &= 0.5 X_{j-1} + \sqrt{0.75} Z_j, \ p_0 + 1 \le j \le p, \end{aligned}$$

where $Z_j, j = 0, 1, \ldots, p$, are i.i.d. standard normal random variables.

(ii) Hierarchy structures. Let s_{01} and s_{02} be the index sets of main-effect and interaction features. They are determined under various hierarchy structures as follows.

NH (no hierarchy structure): $s_{01} = \{1, \cdots, 5\}, s_{02} = \{(1,2),(1,3),(1,6),(5,6),(j-1,j), j = 10, \cdots, 15\}$.

SH (strong hierarchy structure): s_{01} is randomly selected from $\{1, \cdots, p\}$ with size 7, s_{02} is randomly selected from $\{(i,j) : i < j, i \in s_{01}, j \in s_{01}\}$ with size 8.

WH (weak hierarchy structure): s_{01} is randomly selected from $\{1, \cdots, p\}$ with size 7, s_{02} is randomly selected from $\{(i,j) : i < j, i \in s_{01}, j \in s_{01}^c$ or $j \in s_{01}, i \in s_{01}^c\}$ with size 8.

AH (anti hierarchy structure): s_{01} is randomly selected from $\{1, \cdots, p\}$ with size 7, s_{02} is randomly selected from $\{(i,j) : i < j, i \in s_{01}^c, j \in s_{01}^c\}$ with size 8.

(iii) Generation of the response variable. The coefficients β_j and θ_{jk} are generated in two ways:

Type I : the nonzero coefficients for both main and interaction terms are i.i.d. as $2n^{-0.175} + |z|/10$ where $z \sim N(0,1)$.

Type II : the nonzero coefficients for both main and interaction terms are i.i.d. from a uniform distribution over the intervals $(-2\sqrt{\ln p/n}, -\sqrt{\ln p/n}) \cup (\sqrt{\ln p/n}, 2\sqrt{\ln p/n})$.

The response variable is then generated as

$$y = \sum_{j \in s_{01}} \beta_j X_j + \sum_{(j,k) \in s_{02}} \theta_{jk} X_j X_k + \epsilon,$$

where ϵ is generated from $N(0, \sigma^2)$ with $\sigma^2 = 4^{-1}\mathrm{Var}\left(\sum_{j \in s_{01}} \beta_j X_j + \sum_{(j,k) \in s_{02}} \theta_{jk} X_j X_k\right)$.

Each of the 24 combinations of the correlation structures for the covariates, the hierarchy structures and the types of coefficients are considered. Under each setting, 200 replicates of simulations

are carried out, and the PDR and FDR are averaged over the 200 replicates. Throughout the simulation studies, we fix the sample size at $n = 200$ and the number of true features at $p_0 = 15$, including both main-effect and interaction features. We consider three numbers of total covariates $p = 80(< n)$, $p = 200(= n)$ and $p = 1,000(> n)$. In the case of $p = 80$, all the methods are directly applied to the original simulated data. In the case of $p = 200$, the data are screened for the sequential methods by a marginal composite feature screening procedure, that is, the data is screened according to the marginal joint effects of the triplets (X_j, X_k, X_jX_k). In the case of $p = 1,000$, the data are screened for the sequential methods using both the SIS+DIS and the SIS+ISPC methods described in Section 2.2.

Among the versions of RAMP method, the versions with MCP penalty perform universally better than the versions with LASSO penalty. Hence we only report the results of the RAMP method with MCP penalty. The two screening methods, SIS+DIS and the SIS+ISPC, produce comparable results. In general, the performance of SIS+ISPC is slightly better than that of SIS+DIS. Hence we only report the results when SIS+ISPC is used in the case of $p = 1,000$. The results in terms of comparison are in general consistent in all the simulation settings. To save space, we only report the results for $p = 80$ and 200 with Type I coefficients and the results for $p = 1,000$ with Type II coefficients. The results for $p = 80, 200$ and $1,000$ are reported respectively in Tables 1 – 3, Tables 4 – 6 and Tables 7 – 9.

The findings in the case of $p = 80$ are discussed in details below. (i) Comparison between SIGS and SIEP: SIGS has on average a slightly lower PDR but a more substantially lower FDR than SIEP. The average PDRs of SIGS and SIEP are, respectively, 0.964 and 0.981, the average FDRs of SIGS and SIEP are, respectively, 0.308 and 0.412. Considering the combined measure DR = PDR+(1−FDR), the DR of SIGS is 1.659 which is better than 1.569 of SIEP. (ii) Comparison between SIGS and RAMP methods: for the selection of interaction features, SIGS has significantly much higher PDR than the RAMP methods, on the other hand, SIGS also has lower or comparable

FDR than the RAMP methods, which are obvious in Tables 1 – 3; for the selection of main-effect features, under covariate correlation structure **XS2**, SIGS universally has higher PDR and lower FDR than the RAMP methods, under the other two structures, i.e., **XS1** and **XS3**, SIGS has a higher PDR but also a higher FDR, in terms of the combined measure DR, they are comparable. The following table gives the PDR, FDR and DR of the three methods for the selection of main-effect features averaged over these two settings:

Method	PDR	FDR	DR
RAMP-wM	0.574	0.226	1.348
RAMP-sM	0.628	0.345	1.283
SIGS	0.950	0.650	1.304

(iii) Under the anti-hierarchy structure, i.e., AH, the RAMP methods cannot select the interaction features at all. The average PDR and FDR over all the three covariate correlation structures are, respectively, 0.009 and 0.442. This could have been expected. By the nature of the RAMP methods, the premise for an interaction feature to be selected is that both or either its parent main-effect features are selected first. But in the anti-hierarchy structure, none of the parent main-effect features of the interaction features are present in the true model.

The findings in the case of $p = 200$ and $p = 1,000$ are similar in nature to the case of $p = 80$. We only highlight a few points in the case of $p = 1,000$. Comparing with SIEP, the SIGS has comparable PDR but has almost universally lower FDR for the detection of main effect features; except under the anti-hierarchy settings, SIGS has both higher PDR and lower FDR for the detection of interaction features. Comparing with RAMP methods, SIGS has much higher PDR and also a higher FDR, in terms of DR, SIGS performs better than the RAMP methods almost universally. The performance of the RAMP for the detection of interaction features is generally poor, especially, under the anti-hierarchy settings, the RAMP methods can hardly detect the interaction features.

However, comparing with SIGS, the RAMP methods have an advantage in terms of computation time. The average computation times in seconds per simulation replicate with $p = 80, 200$ and $1,000$ are given in the following table. For SIGS, the computation time includes the screening time.

Average computation time per simulation replicate (in second).

p	SIGS	RAMP-wM	RAMP-sM
80	97	7	5
200	43	11	5
1,000	866	44	5

The computation time required of the RAMP methods is much less than the time required of SIGS. In practical problems, if there is a time constraint, or if the user wishes to sacrifice a slight selection accuracy for computational efficiency, the RAMP algorithms can still be good choices.

4 Real examples

In this section, we applied our method on a supermarket data set (Wang, 2009). This data set collects daily sale information of a major supermarket located in northern China. The data consists of observations on the number of customers per day and the daily sale volumes of 6,398 products for 464 days. The supermarket manager is interested in the relationship between the number of customers and the sale volumes of certain products. This data set has been studied in Wang (2009), Hao and Zhang (2014) and Hao et al. (2018). Wang (2009) considered prediction based on main-effect models using different methods. Hao and Zhang (2014) and Hao et al. (2018) analyzed the data based on interaction models and compared the performance of the interaction models and the main-effect models. They found that the interaction models improve the main-effect models substantially in terms of prediction errors.

In our analysis, we focus on the interaction models and apply the four methods in the simulation study to the data. We first screen the 6,398 products by using the SIS approach (Fan and Li, 2001) and retain 200 products. Then, we follow the practice of the papers mentioned above and split the data randomly into a training set with size $n_1 = 400$ and a testing set with size $n_2 = 64$. The model is selected and estimated using the training data. The performance is evaluated using the testing data by the out-of-sample R^2 defined as $100 * (1 - \|Y_{\text{test}} - Z_{\text{test}}\widehat{\boldsymbol{\beta}}\|_2^2/\|Y_{\text{test}} - \bar{Y}_{\text{test}}\mathbf{1}\|_2^2)$, where $\widehat{\boldsymbol{\beta}}$ is estimated based on training data. The average sizes of main-effect and interaction effect terms in

the selected model together with the averaged R^2 are reported below.

Method	MSize	ISize	R^2
SIEP	26(3)	4(1)	89.95(2.69)
SIGS	17(2)	4(2)	88.55(2.84)
RAMP-wM	15(2)	2(1)	87.77(3.37)
RAMP-sM	16(2)	2(1)	88.08(2.94)

The results above are consistent with the findings of the simulation studies. In the simulation studies, we have found that SIGS performs better than the RAMP methods, and that SIEP has slightly a higher PDR than SIGS, which can potentially lead to a slightly better prediction. In the table above, the R^2 of SIEP is slightly higher than that of SIGS, and, in turn, the R^2 of SIGS is higher than those of the two RAMP methods. Since, at the same time, SIEP potentially has a much higher FDR than SIGS, the model selected by SIEP should have a substantially larger size than that by SIGS. It can be seen from the table above that the model size of SIEP is about 1.5 times of that of SIGS. The larger R^2 of SIEP compared with SIGS is probably because that SIEP has selected more relevant products and its selected irrelevant products do not really affect prediction due to their small estimated effects.

One might doubt the advantage of SIGS over the RAMP methods because, while it has a larger R^2, it also has a larger model size. But a larger model size could result from different causes. We can consider three situations: (i) the additional features are all relevant ones, (ii) the additional features are all irrelevant ones, and (iii) some of the additional features are relevant and some of them are irrelevant. In the first situation, we can expect an increase of R^2 which should be proportional to the increase of the model size. In this situation, the larger model size implies higher positive discovery rate (PDR). In the second situation, the R^2 will not necessarily be larger, the larger model size will imply a higher false discovery rate (FDR). In the third situation, we can also expect an increase of R^2 but the increase could not be proportional to the increase of model size. The increase of R^2 and the increase of model size from RAMP-wM to SIGS have a ratio $0.68/4 \approx 0.17$, and those from RAMP-sM to SIGS have a ratio $0.47/3 \approx 0.16$. The increase of R^2 is proportional to the increase

of the model size. Therefore, we might well claim that the SIGS has a higher R^2 as well as a higher PDR than the RAMP methods.

References

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

Cai, T. T. and L. Wang (2011). Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information theory 57*(7), 4680–4688.

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

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

Cordell, H. J. (2009). Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics 10*(6), 392.

Efron, B., T. Hastie, I. Johnstone, R. Tibshirani, et al. (2004). Least angle regression. *The Annals of statistics 32*(2), 407–499.

Evans, D. M., J. Marchini, A. P. Morris, and L. R. Cardon (2006). Two-stage two-locus models in genome-wide association. *PLoS genetics 2*(9), e157.

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

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Methodological) 70*(5), 849–911.

Hao, N., Y. Feng, and H. H. Zhang (2018). Model selection for high-dimensional quadratic regression via regularization. *Journal of the American Statistical Association*, 1–11.

Hao, N. and H. H. Zhang (2014). Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association 109*(507), 1285–1301.

Hao, N. and H. H. Zhang (2017). A note on high-dimensional linear regression with interactions. *The American Statistician 71*(4), 291–297.

He, Y. and Z. Chen (2014). The EBIC and a sequential procedure for feature selection in interactive linear models with high-dimensional data. *Annals of the Institute of Statistical Mathematics*, 1–26.

Kooperberg, C. and M. LeBlanc (2008). Increasing the power of identifying gene× gene interactions in genome-wide association studies. *Genetic epidemiology 32*(3), 255–263.

Luo, S. and Z. Chen (2014). Sequential lasso cum ebic for feature selection with ultra-high dimensional feature space. *Journal of the American Statistical Association 109*(507), 1229–1240.

Manolio, T. A. and F. S. Collins (2007). Genes, environment, health, and disease: facing up to complexity. *Human heredity 63*(2), 63–66.

McCullagh, P. and J. Nelder (1983). *Generalized Linear Models*. Chapman & Hall, London.

Nelder, J. (1977). A reformulation of linear models. *Journal of the Royal Statistical Society. Series A (General)*, 48–77.

Niu, Y. S., N. Hao, and H. H. Zhang (2018). Interaction screening by partial correlation. *Statistics and its Interface 11*(2), 317–325.

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

She, Y., Z. Wang, and H. Jiang (2016). Group regularized estimation under structural hierarchy. *Journal of the American Statistical Association* (just-accepted).

Storey, J. D., J. M. Akey, and L. Kruglyak (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)*, 267–288.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association 104*(488), 1512–1524.

Yeager, M., N. Orr, R. B. Hayes, K. B. Jacobs, P. Kraft, S. Wacholder, M. J. Minichiello, P. Fearnhead, K. Yu, N. Chatterjee, et al. (2007). Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature genetics 39*(5), 645.

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

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

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

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

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

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

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

Table 1: The average PDR and FDR for main-effect and interaction features under covariate correlation structure **XS1** with Type I coefficients and $p = 80$ (the numbers in parentheses are standard errors).

Hierarchy Structure	Method	MPDR	MFDR	IPDR	IFDR
NH	SIEP	0.822(0.270)	0.760(0.182)	0.732(0.323)	0.458(0.208)
	SIGS	0.639(0.341)	0.645(0.184)	0.632(0.324)	0.377(0.274)
	RAMP-wM	0.271(0.106)	0.404(0.247)	0.211(0.133)	0.748(0.160)
	RAMP-sM	0.435(0.166)	0.633(0.135)	0.055(0.066)	0.848(0.203)
SH	SIEP	0.997(0.040)	0.868(0.021)	0.996(0.053)	0.264(0.139)
	SIGS	0.989(0.066)	0.779(0.124)	0.978(0.104)	0.113(0.142)
	RAMP-wM	0.640(0.129)	0.212(0.081)	0.832(0.203)	0.155(0.206)
	RAMP-sM	0.575(0.170)	0.388(0.141)	0.252(0.200)	0.638(0.269)
WH	SIEP	1.000(0.000)	0.869(0.002)	1.000(0.000)	0.246(0.140)
	SIGS	0.997(0.040)	0.803(0.078)	0.992(0.079)	0.092(0.122)
	RAMP-wM	0.584(0.133)	0.221(0.067)	0.591(0.214)	0.412(0.218)
	RAMP-sM	0.571(0.152)	0.376(0.122)	0.017(0.048)	0.973(0.083)
AH	SIEP	1.000(0.000)	0.869(0.002)	1.000(0.000)	0.251(0.149)
	SIGS	0.987(0.069)	0.791(0.112)	0.974(0.130)	0.130(0.164)
	RAMP-wM	0.391(0.123)	0.321(0.104)	0.019(0.052)	0.979(0.057)
	RAMP-sM	0.453(0.160)	0.448(0.142)	0.001(0.012)	0.942(0.231)

Table 2: The average PDR and FDR for main-effect and interaction features under covariate correlation structure **XS2** with Type I coefficients and $p = 80$ (the numbers in parentheses are standard errors).

Hierarchy Structure	Method	MPDR	MFDR	IPDR	IFDR
NH	SIEP	0.999(0.014)	0.132(0.138)	0.986(0.055)	0.182(0.135)
	SIGS	0.982(0.070)	0.015(0.049)	0.964(0.098)	0.134(0.117)
	RAMP-wM	0.578(0.137)	0.048(0.121)	0.246(0.168)	0.273(0.288)
	RAMP-sM	0.614(0.131)	0.214(0.201)	0.086(0.052)	0.404(0.367)
SH	SIEP	1.000(0.000)	0.221(0.096)	1.000(0.000)	0.168(0.131)
	SIGS	0.999(0.010)	0.144(0.042)	0.998(0.020)	0.170(0.126)
	RAMP-wM	0.839(0.114)	0.156(0.041)	0.767(0.128)	0.036(0.083)
	RAMP-sM	0.941(0.112)	0.156(0.073)	0.600(0.171)	0.048(0.149)
WH	SIEP	1.000(0.000)	0.216(0.085)	1.000(0.000)	0.164(0.131)
	SIGS	1.000(0.000)	0.139(0.037)	1.000(0.000)	0.150(0.114)
	RAMP-wM	0.905(0.131)	0.152(0.066)	0.701(0.127)	0.060(0.101)
	RAMP-sM	0.834(0.167)	0.203(0.091)	0.017(0.045)	0.578(0.473)
AH	SIEP	1.000(0.000)	0.219(0.093)	1.000(0.000)	0.176(0.154)
	SIGS	1.000(0.000)	0.140(0.037)	0.999(0.009)	0.118(0.109)
	RAMP-wM	0.511(0.263)	0.290(0.168)	0.009(0.041)	0.219(0.404)
	RAMP-sM	0.717(0.230)	0.227(0.109)	0.002(0.015)	0.084(0.272)

Table 3: The average PDR and FDR for main-effect and interaction features under covariate correlation structure **XS3** with Type I coefficients and $p = 80$ (the numbers in parentheses are standard errors).

Hierarchy Structure	Method	MPDR	MFDR	IPDR	IFDR
NH	SIEP	1.000(0.000)	0.637(0.054)	1.000(0.000)	0.245(0.157)
	SIGS	1.000(0.000)	0.502(0.063)	1.000(0.000)	0.105(0.098)
	RAMP-wM	0.477(0.212)	0.028(0.097)	0.283(0.129)	0.447(0.230)
	RAMP-sM	0.565(0.213)	0.212(0.210)	0.104(0.087)	0.333(0.384)
SH	SIEP	1.000(0.000)	0.658(0.037)	1.000(0.000)	0.302(0.183)
	SIGS	1.000(0.000)	0.553(0.038)	1.000(0.000)	0.146(0.122)
	RAMP-wM	0.819(0.112)	0.167(0.061)	0.774(0.135)	0.037(0.089)
	RAMP-sM	0.911(0.117)	0.199(0.095)	0.597(0.177)	0.050(0.152)
WH	SIEP	1.000(0.000)	0.660(0.049)	1.000(0.000)	0.322(0.219)
	SIGS	1.000(0.000)	0.552(0.037)	1.000(0.000)	0.130(0.120)
	RAMP-wM	0.881(0.107)	0.168(0.064)	0.709(0.131)	0.055(0.108)
	RAMP-sM	0.804(0.172)	0.246(0.102)	0.018(0.048)	0.577(0.469)
AH	SIEP	1.000(0.000)	0.663(0.043)	1.000(0.000)	0.322(0.210)
	SIGS	1.000(0.000)	0.552(0.037)	1.000(0.000)	0.110(0.115)
	RAMP-wM	0.526(0.255)	0.288(0.159)	0.023(0.064)	0.332(0.444)
	RAMP-sM	0.711(0.230)	0.258(0.120)	0.002(0.015)	0.093(0.290)

Table 4: The average PDR and FDR for main-effect and interaction features under covariate correlation structure **XS1** with Type I coefficients and $p = 200$ (the numbers in parentheses are standard errors).

Hierarchy Structure	Method	MPDR	MFDR	IPDR	IFDR
NH	SIEP	0.434(0.252)	0.504(0.230)	0.398(0.230)	0.569(0.187)
	SIGS	0.549(0.316)	0.564(0.199)	0.522(0.317)	0.451(0.282)
	RAMP-wM	0.231(0.080)	0.313(0.286)	0.205(0.117)	0.727(0.165)
	RAMP-sM	0.327(0.149)	0.640(0.139)	0.036(0.054)	0.710(0.382)
SH	SIEP	0.987(0.064)	0.863(0.078)	0.949(0.190)	0.421(0.292)
	SIGS	0.983(0.053)	0.557(0.148)	0.868(0.179)	0.227(0.195)
	RAMP-wM	0.667(0.140)	0.297(0.107)	0.607(0.205)	0.275(0.229)
	RAMP-sM	0.681(0.184)	0.384(0.131)	0.365(0.219)	0.384(0.292)
WH	SIEP	0.968(0.110)	0.846(0.106)	0.914(0.241)	0.480(0.273)
	SIGS	0.899(0.152)	0.476(0.148)	0.678(0.282)	0.398(0.238)
	RAMP-wM	0.614(0.177)	0.333(0.128)	0.437(0.244)	0.455(0.296)
	RAMP-sM	0.585(0.188)	0.448(0.145)	0.011(0.038)	0.954(0.170)
AH	SIEP	0.989(0.065)	0.865(0.078)	0.962(0.164)	0.396(0.277)
	SIGS	0.735(0.176)	0.365(0.115)	0.324(0.201)	0.622(0.216)
	RAMP-wM	0.486(0.153)	0.344(0.131)	0.019(0.045)	0.944(0.171)
	RAMP-sM	0.539(0.171)	0.425(0.141)	0.000(0.000)	0.765(0.425)

Table 5: The average PDR and FDR for main-effect and interaction features under covariate correlation structure **XS2** with Type I coefficients and $p = 200$ (the numbers in parentheses are standard errors).

Hierarchy Structure	Method	MPDR	MFDR	IPDR	IFDR
NH	SIEP	0.981(0.071)	0.236(0.227)	0.950(0.128)	0.310(0.258)
	SIGS	0.964(0.094)	0.063(0.099)	0.896(0.144)	0.180(0.137)
	RAMP-wM	0.537(0.137)	0.024(0.087)	0.212(0.139)	0.194(0.273)
	RAMP-sM	0.586(0.128)	0.194(0.205)	0.082(0.047)	0.333(0.365)
SH	SIEP	0.999(0.020)	0.307(0.164)	0.995(0.071)	0.300(0.274)
	SIGS	0.988(0.042)	0.165(0.062)	0.972(0.079)	0.141(0.136)
	RAMP-wM	0.867(0.125)	0.151(0.044)	0.695(0.164)	0.105(0.152)
	RAMP-sM	0.974(0.075)	0.141(0.047)	0.604(0.155)	0.029(0.104)
WH	SIEP	1.000(0.000)	0.312(0.173)	1.000(0.000)	0.313(0.279)
	SIGS	1.000(0.000)	0.160(0.054)	0.992(0.031)	0.147(0.121)
	RAMP-wM	0.878(0.170)	0.157(0.071)	0.636(0.180)	0.099(0.150)
	RAMP-sM	0.786(0.210)	0.222(0.104)	0.016(0.043)	0.414(0.473)
AH	SIEP	0.999(0.010)	0.299(0.157)	0.995(0.071)	0.304(0.263)
	SIGS	0.994(0.032)	0.154(0.056)	0.704(0.165)	0.275(0.167)
	RAMP-wM	0.519(0.262)	0.275(0.173)	0.003(0.020)	0.123(0.324)
	RAMP-sM	0.804(0.204)	0.203(0.099)	0.001(0.009)	0.040(0.196)

Table 6: The average PDR and FDR for main-effect and interaction features under covariate correlation structure **XS3** with Type I coefficients and $p = 200$ (the numbers in parentheses are standard errors).

Hierarchy Structure	Method	MPDR	MFDR	IPDR	IFDR
NH	SIEP	1.000(0.000)	0.686(0.111)	1.000(0.000)	0.495(0.328)
	SIGS	1.000(0.000)	0.515(0.081)	0.992(0.029)	0.113(0.103)
	RAMP-wM	0.468(0.208)	0.013(0.062)	0.296(0.100)	0.382(0.226)
	RAMP-sM	0.534(0.212)	0.231(0.213)	0.095(0.084)	0.257(0.355)
SH	SIEP	1.000(0.000)	0.722(0.085)	1.000(0.000)	0.632(0.335)
	SIGS	1.000(0.000)	0.549(0.040)	0.999(0.012)	0.140(0.117)
	RAMP-wM	0.856(0.131)	0.165(0.059)	0.685(0.158)	0.096(0.156)
	RAMP-sM	0.951(0.086)	0.176(0.077)	0.588(0.159)	0.027(0.089)
WH	SIEP	1.000(0.000)	0.715(0.084)	1.000(0.000)	0.610(0.329)
	SIGS	1.000(0.000)	0.530(0.066)	0.987(0.040)	0.120(0.119)
	RAMP-wM	0.862(0.181)	0.178(0.084)	0.629(0.191)	0.087(0.134)
	RAMP-sM	0.801(0.203)	0.251(0.110)	0.013(0.040)	0.374(0.474)
AH	SIEP	1.000(0.000)	0.726(0.083)	1.000(0.000)	0.633(0.343)
	SIGS	0.997(0.020)	0.309(0.104)	0.718(0.157)	0.249(0.165)
	RAMP-wM	0.546(0.272)	0.272(0.175)	0.006(0.027)	0.142(0.343)
	RAMP-sM	0.816(0.202)	0.235(0.095)	0.000(0.000)	0.060(0.238)

Table 7: The average PDR and FDR for main-effect and interaction features under covariate correlation structure **XS1** with Type II coefficients and $p = 1,000$ (the numbers in parentheses are standard errors).

Structure	Method	MPDR	MFDR	IPDR	IFDR
NH	SIEP	0.702(0.296)	0.386(0.346)	0.180(0.162)	0.457(0.351)
	SIGS	0.640(0.300)	0.278(0.302)	0.420(0.316)	0.456(0.273)
	RAMP-wM	0.056(0.096)	0.008(0.079)	0.018(0.043)	0.426(0.470)
	RAMP-sM	0.179(0.211)	0.164(0.290)	0.007(0.028)	0.417(0.483)
SH	SIEP	0.860(0.205)	0.385(0.193)	0.694(0.342)	0.426(0.264)
	SIGS	0.936(0.102)	0.270(0.177)	0.861(0.220)	0.307(0.195)
	RAMP-wM	0.303(0.199)	0.151(0.261)	0.122(0.152)	0.237(0.343)
	RAMP-sM	0.756(0.230)	0.277(0.177)	0.440(0.257)	0.163(0.236)
WH	SIEP	0.702(0.234)	0.422(0.175)	0.408(0.249)	0.571(0.246)
	SIGS	0.714(0.224)	0.308(0.161)	0.451(0.240)	0.567(0.207)
	RAMP-wM	0.225(0.184)	0.183(0.308)	0.081(0.130)	0.184(0.323)
	RAMP-sM	0.281(0.209)	0.248(0.323)	0.004(0.023)	0.087(0.279)
AH	SIEP	0.589(0.216)	0.459(0.192)	0.295(0.186)	0.626(0.225)
	SIGS	0.502(0.219)	0.410(0.204)	0.244(0.169)	0.701(0.210)
	RAMP-wM	0.191(0.142)	0.206(0.314)	0.001(0.009)	0.035(0.184)
	RAMP-sM	0.251(0.181)	0.259(0.309)	0.000(0.000)	0.000(0.000)

Table 8: The average PDR and FDR for main-effect and interaction features under covariate correlation structure **XS2** with Type II coefficients and $p = 1,000$ (the numbers in parentheses are standard errors).

Structure	Method	MPDR	MFDR	IPDR	IFDR
NH	SIEP	0.657(0.239)	0.263(0.258)	0.533(0.177)	0.356(0.234)
	SIGS	0.608(0.246)	0.240(0.211)	0.710(0.272)	0.269(0.184)
	RAMP-wM	0.038(0.088)	0.030(0.165)	0.031(0.068)	0.083(0.222)
	RAMP-sM	0.058(0.118)	0.074(0.230)	0.001(0.010)	0.098(0.295)
SH	SIEP	0.976(0.105)	0.286(0.142)	0.964(0.145)	0.262(0.236)
	SIGS	0.990(0.037)	0.173(0.083)	0.986(0.056)	0.295(0.152)
	RAMP-wM	0.454(0.228)	0.079(0.132)	0.323(0.228)	0.191(0.201)
	RAMP-sM	0.919(0.151)	0.152(0.086)	0.649(0.183)	0.020(0.075)
WH	SIEP	0.926(0.162)	0.335(0.182)	0.582(0.218)	0.507(0.253)
	SIGS	0.927(0.142)	0.199(0.102)	0.588(0.206)	0.441(0.190)
	RAMP-wM	0.285(0.230)	0.098(0.213)	0.140(0.176)	0.085(0.171)
	RAMP-sM	0.374(0.260)	0.156(0.231)	0.001(0.012)	0.012(0.106)
AH	SIEP	0.899(0.143)	0.337(0.176)	0.510(0.176)	0.458(0.264)
	SIGS	0.825(0.199)	0.218(0.128)	0.427(0.215)	0.457(0.262)
	RAMP-wM	0.177(0.180)	0.099(0.244)	0.000(0.000)	0.010(0.100)
	RAMP-sM	0.364(0.271)	0.130(0.205)	0.000(0.000)	0.000(0.000)

Table 9: The average PDR and FDR for main-effect and interaction features under covariate correlation structure **XS3** with Type II coefficients and $p = 1,000$ (the numbers in parentheses are standard errors).

Structure	Method	MPDR	MFDR	IPDR	IFDR
NH	SIEP	0.795(0.166)	0.392(0.258)	0.605(0.120)	0.406(0.250)
	SIGS	0.773(0.189)	0.294(0.227)	0.824(0.241)	0.284(0.191)
	RAMP-wM	0.162(0.160)	0.023(0.129)	0.045(0.074)	0.079(0.211)
	RAMP-sM	0.186(0.180)	0.033(0.147)	0.000(0.000)	0.010(0.100)
SH	SIEP	0.978(0.100)	0.461(0.176)	0.968(0.136)	0.347(0.297)
	SIGS	0.991(0.035)	0.351(0.114)	0.985(0.060)	0.303(0.148)
	RAMP-wM	0.461(0.232)	0.086(0.133)	0.321(0.230)	0.192(0.202)
	RAMP-sM	0.919(0.151)	0.156(0.084)	0.643(0.185)	0.019(0.074)
WH	SIEP	0.924(0.151)	0.371(0.179)	0.589(0.219)	0.542(0.247)
	SIGS	0.934(0.132)	0.212(0.104)	0.601(0.214)	0.439(0.189)
	RAMP-wM	0.281(0.221)	0.117(0.241)	0.143(0.192)	0.100(0.199)
	RAMP-sM	0.397(0.262)	0.174(0.232)	0.002(0.018)	0.018(0.127)
AH	SIEP	0.899(0.146)	0.341(0.156)	0.507(0.182)	0.452(0.256)
	SIGS	0.841(0.196)	0.226(0.124)	0.442(0.213)	0.458(0.251)
	RAMP-wM	0.175(0.178)	0.110(0.257)	0.000(0.000)	0.005(0.071)
	RAMP-sM	0.364(0.272)	0.137(0.207)	0.000(0.000)	0.000(0.000)