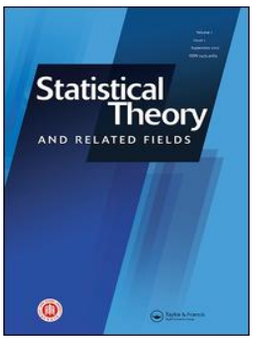

Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Group screening for ultra-high-dimensional feature under linear model

Yong Niu, Riquan Zhang, Jicai Liu & Huapeng Li

To cite this article: Yong Niu, Riquan Zhang, Jicai Liu & Huapeng Li (2019): Group screening for ultra-high-dimensional feature under linear model, Statistical Theory and Related Fields, DOI: 10.1080/24754269.2019.1633763

To link to this article: https://doi.org/10.1080/24754269.2019.1633763

Published online: 04 Jul 2019.

Submit your article to this journal

Article views: 7

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tstf20

STATISTICAL THEORY AND RELATED FIELDS
https://doi.org/10.1080/24754269.2019.1633763

Taylor & Francis
Taylor & Francis Group

Check for updates

Group screening for ultra-high-dimensional feature under linear model

Yong Niu[a,b], Riquan Zhang[a], Jicai Liu[c] and Huapeng Li[a]

[a]School of Statistics, East China Normal University, Shanghai, People's Republic of China; [b]Department of Mathematics and Physics, Hefei University, Hefei, People's Republic of China; [c]Department of Mathematics, Shanghai Normal University, Shanghai, People's Republic of China

ABSTRACT

Ultra-high-dimensional data with grouping structures arise naturally in many contemporary statistical problems, such as gene-wide association studies and the multi-factor analysis-of-variance (ANOVA). To address this issue, we proposed a group screening method to do variables selection on groups of variables in linear models. This group screening method is based on a working independence, and sure screening property is also established for our approach. To enhance the finite sample performance, a data-driven thresholding and a two-stage iterative procedure are developed. To the best of our knowledge, screening for grouped variables rarely appeared in the literature, and this method can be regarded as an important and non-trivial extension of screening for individual variables. An extensive simulation study and a real data analysis demonstrate its finite sample performance.

ARTICLE HISTORY
Received 18 July 2018
Revised 14 April 2019
Accepted 17 June 2019

KEYWORDS
Ultra-high-dimensional; group screening; linear model; sure screening property

AMS 2000 SUBJECT CLASSIFICATIONS
62G05; 62E20

1. Introduction

Nowadays, grouping predictors arise naturally in many regression problems. It means that we are interested in finding relevant predictors in modelling the response variable, where each predictors may be represented by a group of indictor variables or a set of basis functions. Grouping structures can be introduced into a regression model naturally in hoping that the prior knowledge about predictors may be used to the full. Thus, grouping structure problems become increasingly important in various research fields. One common example is the representation of multi-level analysis-of-variance (ANOVA) in a regression model with a group of derived input variables. The aim of ANOVA is often to select relevant main factors and interactions, that is the selection of groups of derived input variables. Another example is the additive model with nonparametric components, where each component can be expressed a linear combination of a set of basis functions of the original predictors. Thus, in both cases, variable selection amounts to the selection of groups of variables rather than individual derived variables.

Using the penalised method, many researchers have considered the group selection problems in various parametric or nonparametric regression models. These articles include, but are not limited to the following. First, Bakin (1999) proposed the group LASSO in his doctoral dissertation. Yuan and Lin (2006) further studied the group LASSO and related group selection methods, such as the group LARS and the group Garrote, and proposed the corresponding algorithms. But they did not give any asymptotic properties of the group LASSO. Wei and Huang (2010) showed that, under a generalised sparsity condition and the sparse Riesz condition proposed by Zhang and Huang (2008), together with some regularity conditions, the group LASSO can select a model with the same order as the underlying model. They also established the asymptotic properties of the adaptive group LASSO, which can correctly select groups with probability tending to one. Under the assumption of generalised linear models, Breheny and Huang (2009) established a general framework for simultaneous group and individual variable selection, or bi-level selection and the corresponding local coordinate descent algorithm. In addition to the group LASSO, many authors also proposed other methods for various parametric models. For example, Huang, Ma, Xie, and Zhang (2009) showed that simultaneous group and individual variable selection can be conducted by a group bridge method. They showed that it can correctly selected relevant groups with probability tending to one. Moreover, Zhao, Rocha, and Yu (2009) introduced a quite general composite penalty for groups selection by combing different norms to form an intelligent penalty. All these methods are very useful for moderate number of predictors to be smaller than the sample size or comparable with it. However, with rapid progress of computing power and modern technology

CONTACT Riquan Zhang Zhangriquan@163.com, rqzhang@stat.ecnu.edu.cn School of Statistics, East China Normal University, 200062 Shanghai, People's Republic of China

© East China Normal University 2019

for data collection, massive amounts of ultra-high-dimensional data are frequently seen in diverse fields of scientific research. Due to the "curse of dimensionality" in terms of simultaneous challenges to computational expediency, statistical accuracy and algorithm stability, the above methods are limited in handling ultra-high-dimensional problems.

In the seminal work of Fan and Lv (2008), a new framework for sure independence screening (SIS) was established. They showed that the method based on Pearson correlation learning possess a sure screening property for linear regressions. That is, all relevant predictors can be selected with probability tending to one even if the number of predictors p can grow much faster than the number of observations n with $\log p = O(n^{\alpha})$ for some $\alpha \in \left(0, \frac{1}{2}\right)$. Following Fan and Lv (2008), we call this non-polynomial dimensionality or ultra-high dimensionality. From this on, various screening methods based on model assumption or model free have been developed (Fan & Lv, 2008; Fan, Samworth, & Wu, 2009; Fan & Song, 2010; He, Wang, & Hong, 2013; Li, Zhong, & Zhu, 2012; Shao & Zhang, 2014; Wang, 2009; Zhao & Li, 2012). However, all these screening methods deal with the individual variables rather than grouped predictors. To the best of our knowledge, screening methods for grouped predictors are quite limited in existing literatures. Thus it is very important to propose a new screening method based on grouped predictors.

Motivated by the theory of SIS, we consider how to deal with these ultra-high-dimensional grouped predictors in the assumption of linear regression model. Considering grouping structures in linear regression model, we have

$$Y = \sum_{j=1}^{J} \boldsymbol{X}_j^{\mathrm{T}} \boldsymbol{\beta}_j + \varepsilon, \tag{1}$$

where Y is the response variable, $\boldsymbol{X}_j = (X_{j1}, \ldots, X_{jp_j})^{\mathrm{T}}$ is a $p_j \times 1$ random vector representing the jth group, $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp_j})^{\mathrm{T}}$ is the $p_j \times 1$ parameter vector corresponding to the jth group predictors, and ε is the random error with mean 0.

Our method is a two-stage approach. First, an efficient screening procedure is employed to reduce the number of group predictors to a moderate order under sample size, and then the existing group selection methods can be used to recover the final sparse model. Due to fastness and efficiency of the screening method for group predictors, we consider a independence screening method by ranking the magnitude of marginal estimators based on each grouped predictor. That is, we fit p marginal linear regressions of the response Y against the variables of the jth group respectively, and the select the relevant group predictors by a measure of the goodness of fit in its marginal linear regression model. Under some mild conditions, We show that there is a significant difference between relevant group predictors and irrelevant ones, according to the strength of these marginal utility. Thus we can distinguish active group predictors from much more inactive ones. Next, the existing group selection methods, such as the group LASSO, the group SCAD (Breheny & Huang, 2015) and the group MCP (Breheny & Huang, 2015), can be used to obtain the final sparse model. We refer to our screening procedure as the Group-SIS, and theoretically establish the sure screening property of our approach. In order to further reduce the false-positive rate, we propose a iterative version of algorithm, named ISIS-Group-Lasso. To enhance performance and speed up the computation of ISIS-Group-Lasso, a greedy modification to the above iterative algorithm, named g-ISIS-Group-Lasso is also developed. Our simulation studies indicate that ISIS-Group-Lasso and g-ISIS-Group-Lasso significantly outperform the competitive group selection, such as distance correlation-based screening method and the group LASSO, especially when the dimensionality is ultra-high.

The rest of the article is organised as follows. In Section 2, we introduce a marginal group SIS in linear regression models. Under some mild conditions, the sure screening property and model selection consistency of the Group-SIS will be established in Section 3. In Section 4, simulation studies and a real data analysis are carried out to assess the performance of our method. Concluding remarks are given in Section 5. All technical proofs for the main theoretical results are given in the Appendix.

2. Group SIS and iterative algorithm

2.1. Marginal linear regression based on grouped predictors

Suppose that we have n random sample from model (1) of the form

$$y_i = \sum_{j=1}^{J} \boldsymbol{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_j + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{2}$$

in which $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijp_j})^{\mathrm{T}}$, y_i and ε_i are scalar. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\boldsymbol{x}_j = (\boldsymbol{x}_{1j}, \ldots, \boldsymbol{x}_{nj})^{\mathrm{T}}$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$, where $\boldsymbol{x}_j$ is the $n \times p_j$ design matrix corresponding to the jth group for each $j = 1, 2, \ldots, p$. Then the model (2) can be rewritten as

$$\boldsymbol{y} = \sum_{j=1}^{J} \boldsymbol{x}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}. \tag{3}$$

For simplicity, the number of variables in each group is uniformly bounded. That is, there exists a positive constant K such that $p_j \leq K$ for $j = 1, 2, \ldots, J$. To rapidly select the relevant grouped predictors, we consider the following J marginal linear regressions against

the grouped predictors:

$$\min_{\boldsymbol{\gamma}_j \in \mathbb{R}^{p_j}} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_{ij}^{\mathrm{T}} \boldsymbol{\gamma}_j)^2, \tag{4}$$

where $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jp_j})'$ is a p_j-dimensional vector for each $j = 1, 2, \ldots, J$.

It is easy to see that the minimiser of (4) is given by

$$\hat{\boldsymbol{\gamma}}_j = (\boldsymbol{x}_j^{\mathrm{T}} \boldsymbol{x}_j)^{-1} \boldsymbol{x}_j^{\mathrm{T}} \boldsymbol{y}.$$

Then we define the marginal utility of the jth grouped predictors as

$$\|\hat{\upsilon}_{nj}\|_n^2 \equiv \frac{1}{n} \sum_{i=1}^{n} (\hat{\boldsymbol{\gamma}}_j^{\mathrm{T}} \boldsymbol{x}_{ij})^2 = \frac{1}{n} \boldsymbol{y}^{\mathrm{T}} \boldsymbol{x}_j (\boldsymbol{x}_j^{\mathrm{T}} \boldsymbol{x}_j)^{-1} \boldsymbol{x}_j^{\mathrm{T}} \boldsymbol{y}. \tag{5}$$

We now select a set of relevant grouped predictors as follows:

$$\hat{\mathcal{M}}_\kappa = \left\{ 1 \le j \le J : \frac{1}{p_j} \|\hat{\upsilon}_{nj}\|_n^2 \ge \pi_n \right\},$$

where κ is a positive constant and π_n is a pre-specified threshold value which will be given later.

Equivalently, we can also define a screening criterion by ranking the residual sum of squares of the corresponding marginal linear regression. These two ways can reduce the group dimensionality from J to a moderate size $|\hat{\mathcal{M}}_\kappa|$. Here, the pre-specified threshold value is crucial in the screening procedure. If we choose it too small, we may select many irrelevant grouped predictors in the final model. On the contrary, we have the risk of losing some important variables. In a word, we should select all of the relevant grouped predictors and control the selected model size simultaneously. In Section 3, we will theoretically show that this group screening approach possesses a sure screening property and the final model size is only of polynomial order. Noted that, our method is the same as traditional feature screening when each group has one variable. In this sense, our method can be regarded as a non-trivial extension of feature screening under the context of single feature screening.

2.2. Iterative group-SIS algorithm

For these ultra-high-dimensional group variable selection problems, we propose a two-stage procedure. That is, we first apply a sure screening method such as Group-SIS to reduce the number of groups from J to a relatively large scale d, where the dimensionality of the selected d group is below sample size n. Then we can use a lower dimensional group-wise variable selection procedure, such as group Lasso, group SCAD or group MCP. In this article, we use group lasso penalty as our group selection strategy. In fact, other group variable selection methods would also work.

However, as Fan and Lv (2008) point out, this marginal independence screening method would still suffer from false negative (i.e., miss some important group predictors that are marginally uncorrelated, but jointly correlated with response), and false positive (i.e., select some unimportant group predictors which have higher marginal correlation than some important group variables). Therefore, we propose an iterative framework to enhance the finite performance of this screening method. That is, we can iteratively use a large-scale group screening and moderate-scale group variable selection strategy.

To obtain a data-driven thresholding for independence group screening, we extend the random permutation idea of Zhao and Li (2012), which select a small proportion of inactive variables to enter the model in each screening step. Let $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_J)$, randomly permute the row of $\boldsymbol{x}$ to get the decouple data $\tilde{\boldsymbol{x}}$ and $\boldsymbol{y}$. Based on the randomly decoupled data $(\tilde{\boldsymbol{x}}, \boldsymbol{y})$, which has no relationship between group variables and response, we compute the value of $\|\hat{\upsilon}_{nj}^*\|_n^2$ similar to $\|\hat{\upsilon}_{nj}\|_n^2$ for $j = 1, 2, \ldots, J$. These values serve as the baseline of the marginal group screening utilities under the null model (no relationship between group variables and response). To obtain the screening threshold, we choose ω_q as the q-ranked magnitude of $\{\|\hat{\upsilon}_{nj}^*\|_n^2, j = 1, 2, \ldots, J\}$. In our simulation, we uses $q = 1$, namely, the largest marginal group screening utilities under the null model. For the sake of completeness, our ISIS-Group-Lasso algorithm proceeds as follows.

Step 1. Compute J marginal utility $\|\hat{\upsilon}_{nj}\|_n^2$, and the initial index subset is chosen as

$$\mathcal{A}_1 = \left\{ 1 \le j \le J : \frac{1}{p_j} \|\hat{\upsilon}_{nj}\|_n^2 \ge \omega_q \right\}.$$

Step 2. Apply the group Lasso (Breheny & Huang, 2009) on the index subset $\mathcal{A}_1$ to obtain a subset $\mathcal{M}_1$. In this step, we choose the regularisation parameter by the Bayesian Information Criterion (BIC) method.

Step 3. Conditioning on $\mathcal{M}_1$, compute the marginal regression

$$\hat{\upsilon}_{nj} = \min_{\boldsymbol{\gamma}_j \in \mathbb{R}^{p_j}} \sum_{i=1}^{n} \left(y_i - \sum_{\tilde{k} \in \mathcal{M}_1} \boldsymbol{x}_{i\tilde{k}}^{\mathrm{T}} \boldsymbol{\gamma}_{\tilde{k}} - \boldsymbol{x}_{ij}^{\mathrm{T}} \boldsymbol{\gamma}_j \right)^2$$

for each $j \in \mathcal{M}_1^c$. By randomly permuting only the groups not in $\mathcal{M}_1$, we obtain a new index subset $\mathcal{A}_2$ similar to step 1. Apply the group LASSO on the index subset $\mathcal{A}_2 \cup \mathcal{M}_1$ to obtain a new subset $\mathcal{M}_2$.

Step 4. Repeat the process until we have the final index set $\mathcal{A}_k$ such that $|\mathcal{A}_k| \ge k_o$ or $\mathcal{A}_k = \mathcal{A}_l$ for some $l < k$.

In order to further reduce false positive and speed up computation, we propose a greedy modification to enhance the finite performance of the above algorithm. Specifically, we restrict the number of the selected

groups in the iterative screening steps to be at most J_0, a small positive integer, and the procedure stops when none of the group predictors is recruited. In our simulation, we set $J_0 = 1$. This greedy version of ISIS-Group-Lasso algorithm is called g-ISIS-Group-Lasso. When $J_0 = 1$, this method is connected with forward regression screening (Wang, 2009), which select at most one new group predictor into the model at a time. However, there is a great difference between the two methods, that is our method includes a deletion step via group selection that can remove multiple group predictors. This makes our procedure more effective, because it is more flexible in terms of recruiting and deleting group predictors. Based on our simulation results in Section 4, g-ISIS-Group-Lasso outperforms other methods in terms of lower false-positive rate, higher percentage of selected the corrected model and small model error.

3. Theoretical properties

Before we establish the sure screening property of our method for linear models, let us introduce some notations first. Denote the Euclidean and the sup norm of a vector α by $\|\alpha\|$ and $\|\alpha\|_\infty$, respectively. For any symmetric matrix A, let $\|A\|_\infty = \max_{i,j} A_{ij}$ be the infinity norm and $\|A\|$ the operator norm. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the minimum and maximum eigenvalue of the matrix A. Let $\boldsymbol{X} = (\boldsymbol{X}_1^{\mathrm{T}}, \boldsymbol{X}_2^{\mathrm{T}}, \ldots, \boldsymbol{X}_J^{\mathrm{T}})^{\mathrm{T}}$ and $E(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}) = \boldsymbol{\Sigma}$. For each $j = 1, 2, \ldots, J$ and $k = 1, 2, \ldots, p_j$, let $[a, b]$ be the support of X_{jk}. Define the index set of the truly model $\mathcal{M}_*$ by

$$\mathcal{M}_* = \{1 \le j \le J : \boldsymbol{\beta}_j \neq 0\}.$$

To gain theoretical insights into the Group-SIS, we need to define υ_j, which is the population version of (4), by minimising

$$\min E(Y - \upsilon_j)^2 \equiv \min_{\boldsymbol{\gamma}_j \in \mathbb{R}^{p_j}} E(Y - \boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{\gamma}_j)^2$$

with respect to $\boldsymbol{\gamma}_j \in \mathbb{R}^{p_j}$. Then we have

$$\upsilon_j = \boldsymbol{X}_j^{\mathrm{T}}(E\boldsymbol{X}_j\boldsymbol{X}_j^{\mathrm{T}})^{-1}E\boldsymbol{X}_j Y.$$

Similar to (5), we define

$$\begin{aligned}\|\upsilon_j\|^2 &\equiv E(\boldsymbol{X}_j^{\mathrm{T}}(E\boldsymbol{X}_j\boldsymbol{X}_j^{\mathrm{T}})^{-1}E\boldsymbol{X}_j Y)^2 \\ &= (E\boldsymbol{X}_j Y)^{\mathrm{T}}(E\boldsymbol{X}_j\boldsymbol{X}_j^{\mathrm{T}})^{-1}(E\boldsymbol{X}_j Y).\end{aligned}$$

Next we collect the technical assumptions to establish the sure screening property of our group screening method.

(i) $\frac{1}{p_j}\min_{j\in\mathcal{M}_*}\|\upsilon_j\|^2 \ge cn^{-\kappa}$, for some $0 < \kappa < \frac{1}{2}$ and $c > 0$.
(ii) $\|\sum_{j=1}^{J} X_j'\beta_j\|_\infty < M_1$ for $M_1 > 0$.
(iii) For any $B > 0$ and $i = 1, 2, \ldots, n$, there is a positive constant M_2 such that $E[\exp\{B|\varepsilon_i|\}] < M_2$.
(iv) For $j = 1, 2, \ldots, J$, the eigenvalues of $\Sigma_j = E\boldsymbol{X}_j\boldsymbol{X}_j^{\mathrm{T}}$ are bounded away from zero and infinity. That is, there are some positive constants τ_1 and τ_2 such that $0 < \tau_1 \le \lambda_{\min}(\Sigma_j) \le \lambda_{\max}(\Sigma_j) \le \tau_2 < \infty$.

Under condition (i), we obtain the minimum signal of the relevant grouped predictors. That is, the magnitude of these marginal utilities of the grouped predictors can preserve the non-sparsity signal of the real model with the fastest convergence rate. This condition is often seen in screening literatures, which is important as it guarantees that marginal utilities carry information about the relevant covariates in the active set. And, conditions (ii) and (iii) are two mild conditions needed in using Berstein's inequality (Van der Vaart & Wellner, 1996). Because we allow $|\mathcal{M}_\kappa|$ increase with n, condition (ii) ensures the convergence of $\sum_{j=1}^{J}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{\beta}_j$. Condition (iv) is also easy to be satisfied, for the small number of variables in each grouped predictor.

Remark 3.1: The above assumptions only serve to help us to further understand the new group screening methodology. Thus these conditions are imposed to facilitate the technical proofs, and the weaker condition may be an interesting topic for future research.

The following Theorem 3.1 provides the sure screening property of our group screening method.

Theorem 3.1: *Suppose that conditions* (i)–(iv) *hold. There exists a constant* c_1, *such that we have*

$$P\{\mathcal{M}_* \subset \hat{\mathcal{M}}_\kappa\} \ge 1 - 4\sum_{j=1}^{J}(p_j + p_j^2)\exp\{-c_1 n^{1-2\kappa}\}.$$

Theorem 3.1 indicates that we can select all the relevant grouped predictors with probability tending to 1, and the key to the theorem's proof is on how to obtain the uniform consistence of $\|\hat{\upsilon}_{jn}\|^2$ to $\|\upsilon_{jn}\|^2$. Denote $\sum_{j=1}^{J} p_j = p$. Note that $p_j \le K$ uniformly, the dimensionality can be handled as high as

$$\log p = o(n^{1-2\kappa}).$$

That is, under some mild conditions, our method has the sure screening property and can reduce from the exponentially growing dimension p to a relatively moderate scale which will be applied in the next group variable selection. Especially, we should point out that κ is very important to our screening procedure. The greater κ is, the higher number of groups that our method can deal with.

On the other hand, although we can select the relevant grouped predictors with probability tending to 1, the cardinality of the $\hat{\mathcal{M}}_\kappa$ may be relatively large

compared with the sample size. That is, there are many unimportant grouped predictors in the final model. Thus controlling the false-positive rate is also necessary for our method. From the perspective in simulation, we have proposed an iterative algorithm in Section 2.2 to enhance the performance of our method in terms of the false selection rates. Under the same conditions as in Theorem 3.1, we will show theoretically that the size of the final model is as large as $n^{\kappa}\lambda_{\max}(\boldsymbol{\Sigma})$ with probability tending to one exponentially.

Theorem 3.2: *Suppose that conditions* (i)–(iv) *hold, there exists some positive constant* c_0,

$$P\{|\hat{\mathcal{M}}_{\kappa}| \le c_0 n^{\kappa}\lambda_{\max}(\boldsymbol{\Sigma})\} \ge 1 - 4\sum_{j=1}^{J}(p_j + p_j^2)\exp\{-c_1 n^{1-2\kappa}\}.$$

Theorem 3.2 shows that, with probability tending to 1, the size of the selected model by our procedure is as large as polynomial order when $\lambda_{\max}(\boldsymbol{\Sigma})$ is of polynomial order. It is crucial to the next group selection stage, which make our two-stage approach much better than traditional group variable selection methods. Because there is no guarantee that existing group selection methods can select the relevant grouped predictors consistently if there are too irrelevant grouped predictors. Thus this theorem, together with Theorem 3.1, implies we can select a model which includes all relevant grouped predictors and a small number of irrelevant ones with high probability.

4. Numerical studies

4.1. Simulation results

In this section, we carry out some simulation studies to demonstrate the finite sample performance of our group screen methods described in Section 2. We consider two group size scenarios of simulation models. In the first scenario, the group sizes are equal. In the second, the group sizes vary. In our simulation, we set all groups with size 5 or 3. We set the sample size $n = 200$, and the following three configurations with $J = 200$, 400, 1000 groups are considered for generating the covariates $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_J)$. For example, when $J = 200$ and group size is 5, the final predictor matrix has the number of variables $p = 1000$. To gauge the difficulties of the simulation models, different scenarios of signal-to-noise ratio (SNR) are given in Examples 4.1–4.4, where $\text{SNR} = \text{Var}(\sum_{j=1}^{J}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{\beta}_j)/\text{Var}(\epsilon)$. It is obviously that the larger the value of SNR, the higher probability that our group screening method can select the relevant groups. In all examples, the simulation results are based on 200 replications for each parameter setup.

To further explore the finite sample performance of our methods, we create some unimportant group variables highly correlated with the response due to the presence of the important group variables associated with the spurious group variables. The correlation between group variables can be specified as follows.

The group vector is $\boldsymbol{X} = (\boldsymbol{X}_1^{\mathrm{T}}, \ldots, \boldsymbol{X}_J^{\mathrm{T}})^{\mathrm{T}}$ where $\boldsymbol{X}_j = (X_{j1}, \ldots, X_{jp_j})^{\mathrm{T}}$, $j = 1, 2, \ldots, J$. In the following four examples, we set two different group size and three configurations of the number of groups are considered. For example, we set $p_j = 5$ for each $j = 1, \ldots, J$. To generate $\boldsymbol{X}$, we first simulate $5J$ random variables $T_1, \ldots, T_{5J}$ independently from $N(0, 1)$. Then $Z_1, \ldots, Z_{50}$ are simulated from a multivariate normal distribution with mean 0 and $Cov(Z_{j_1}, Z_{j_2}) = 0.6^{|j_1 - j_2|}$. For $k = 1, \ldots, 5$, the group variable X_{jk} are generated as

$$X_{jk} = \begin{cases} \dfrac{Z_j + T_{5(j-1)+k}}{\sqrt{2}}, & j = 1, \ldots, 50, \\ T_{5(j-1)+k}, & j = 51, \ldots, J. \end{cases}$$

Here, the group components were the relevant variables, while most of components were spurious variables not used in the model but correlated to the relevant group variables. The random error ε_i was generated from a standard normal distribution. To ensure that the theoretical value of SNR was not too weak or strong, ε_i can be multiplied by a constant σ.

Extensive simulation studies have been conducted to demonstrate the finite performance of our group screening method. For comparison purpose, the performance of distance correlation screening (DC-SIS) (Li et al., 2012), which is a model-free screening method that uses the distance correlation to replace Pearson correlation in marginal correlation screening, is examined. For the sake of fairness, we propose first to apply DC-SIS to reduce the number of groups to $\frac{n}{\log n}$, and a group-wise variable selection procedure such as group Lasso is conducted to recover the final model. We call DC-SIS followed by group Lasso DC-SIS-Group-Lasso. To enhance the performance of DC-SIS, Zhong and Zhu (2015) propose an iterative version of DC-SIS, named by DC-ISIS, which will also be chosen as a comparison. For the sake of fairness, group Lasso is conducted after DC-ISIS, referring to DC-ISIS-Group-Lasso. At the same time, the performance of group Lasso was also examined. Thus, we have five group screening method (i.e., ISIS-Group-Lasso, g-ISIS-Group-Lasso, Group-Lasso, DC-SIS-Group-Lasso, DC-ISIS-Group-Lasso) under consideration. In the following four examples, we report five performance measures: true positive (TP), false positive (FP), median of the model size (MEDIAN), percentage of occasions on which the exactly correct groups are selected (CORRECT) and model error (Yuan & Lin, 2006).

Example 4.1: Each group consists of five variables, and the number of the relevant groups is 4. And we generate

the response from the following linear model:

$$Y = \boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{\beta}_1 + \boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{\beta}_2 + \boldsymbol{X}_3^{\mathrm{T}}\boldsymbol{\beta}_3 + \boldsymbol{X}_4^{\mathrm{T}}\boldsymbol{\beta}_4 + \sigma\varepsilon,$$

where

$$\begin{aligned}
\boldsymbol{\beta}_1 &= (1.5, 1.5, 1, 1, -0.5)^{\mathrm{T}},\\
\boldsymbol{\beta}_2 &= (1.5, -0.5, 0.5, 2, 0.5)^{\mathrm{T}},\\
\boldsymbol{\beta}_3 &= (2, -1, 1.5, 1.5, 2)^{\mathrm{T}},\\
\boldsymbol{\beta}_4 &= (-1, -2, -2, 0.5, -1)^{\mathrm{T}},\\
\boldsymbol{\beta}_5 &= \boldsymbol{\beta}_6 = \cdots = \boldsymbol{\beta}_J = (0, 0, 0, 0, 0)^{\mathrm{T}}.
\end{aligned}$$

Example 4.2: Similar to Example 4.1, the number of the relevant groups is 8 with group size 3. We generate the response from the following linear model:

$$\begin{aligned}
Y = {} & \boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{\beta}_1 + \boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{\beta}_2 + \boldsymbol{X}_3^{\mathrm{T}}\boldsymbol{\beta}_3 + \boldsymbol{X}_4^{\mathrm{T}}\boldsymbol{\beta}_4 + \boldsymbol{X}_5^{\mathrm{T}}\boldsymbol{\beta}_5\\
& + \boldsymbol{X}_6^{\mathrm{T}}\boldsymbol{\beta}_6 + \boldsymbol{X}_7^{\mathrm{T}}\boldsymbol{\beta}_7 + \boldsymbol{X}_8^{\mathrm{T}}\boldsymbol{\beta}_8 + \sigma\varepsilon,
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\beta}_1 &= (0.5, -2, -2)^{\mathrm{T}}, \quad \boldsymbol{\beta}_2 = (1, 3, 1)^{\mathrm{T}},\\
\boldsymbol{\beta}_3 &= \sqrt{2} * (1.5, -0.5, 2)^{\mathrm{T}},\\
\boldsymbol{\beta}_4 &= \sqrt{2} * (1, -1.5, -2)^{\mathrm{T}},\\
\boldsymbol{\beta}_5 &= (0.5, -2, -2)^{\mathrm{T}}, \quad \boldsymbol{\beta}_6 = (1, 3, 1)^{\mathrm{T}},\\
\boldsymbol{\beta}_7 &= \sqrt{2} * (1.5, -0.5, 2)^{\mathrm{T}},\\
\boldsymbol{\beta}_8 &= \sqrt{2} * (1, -1.5, -2)^{\mathrm{T}},\\
\boldsymbol{\beta}_9 &= \boldsymbol{\beta}_{10} = \cdots = \boldsymbol{\beta}_J = (0, 0, 0)^{\mathrm{T}}.
\end{aligned}$$

Example 4.3: In this example, the group sizes differ across groups. There are half of the groups with size 5 and the other groups with size 3. The group variables are generated the same way as the above examples. The response variable Y is generate from $Y = \sum_{k=1}^{4} \boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{\beta}_k + \sigma\varepsilon$. However, the regression coefficients

$$\begin{aligned}
&\boldsymbol{\beta}_1 = (0.5, 0.5, -0.5, 2, 1)^{\mathrm{T}}, \quad \boldsymbol{\beta}_2 = (2, 0, 1, 1.5, -1)^{\mathrm{T}},\\
&\boldsymbol{\beta}_3 = (0.5, -2, -2)^{\mathrm{T}}, \quad \boldsymbol{\beta}_4 = (1, 3, 1)^{\mathrm{T}},\\
&\boldsymbol{\beta}_5 = \cdots = \boldsymbol{\beta}_{0.5*(J-4)+4} = (0, 0, 0, 0, 0)^{\mathrm{T}},\\
&\boldsymbol{\beta}_{0.5*(J-4)+5} = \cdots = \boldsymbol{\beta}_J = (0, 0, 0)^{\mathrm{T}}.
\end{aligned}$$

Example 4.4: In this example, the group sizes also differ across groups. This example is a more difficult case than Example 4.3, because it has eight groups with more different regression coefficients. There are $0.5 * J$ groups with size 5 and the other groups with size 3. The group variables are generated the same way as the above Example 4.3. The response variable Y is generate from $Y = \sum_{k=1}^{8} \boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{\beta}_k + \sigma\varepsilon$. However, the regression coefficients

$$\begin{aligned}
\boldsymbol{\beta}_1 &= (1.5, 1.5, 1, 1, -0.5)^{\mathrm{T}},\\
\boldsymbol{\beta}_2 &= (1.5, -0.5, 0.5, 1.5, 0.5)^{\mathrm{T}},\\
\boldsymbol{\beta}_3 &= (1.5, -1, 1, 1, 2)^{\mathrm{T}},\\
\boldsymbol{\beta}_4 &= (-1, -1.5, -1.5, 0.5, -1)^{\mathrm{T}},\\
\boldsymbol{\beta}_5 &= (0.5, -2, -2)^{\mathrm{T}}, \quad \boldsymbol{\beta}_6 = (1, 3, 1)^{\mathrm{T}},\\
\boldsymbol{\beta}_7 &= \sqrt{2} * (0.5, -2, -2)^{\mathrm{T}}, \quad \boldsymbol{\beta}_8 = \sqrt{2} * (1, 3, 1)^{\mathrm{T}},\\
\boldsymbol{\beta}_9 &= \cdots = \boldsymbol{\beta}_{0.5*(J-8)+8} = (0, 0, 0, 0, 0)^{\mathrm{T}},\\
\boldsymbol{\beta}_{0.5*(J-8)+9} &= \cdots = \boldsymbol{\beta}_J = (0, 0, 0)^{\mathrm{T}}.
\end{aligned}$$

Detailed simulation results of Examples 4.1–4.4 are given in Tables 1–4, respectively. Especially, the boxplots of average model size are presented in Figure 1. Obviously, in all these four examples, the relevant groups can almost be selected for four methods except for DC-SIS-Group-Lasso, which misses more groups most of the time. In terms of true positives (TP), the iterative version of DC-SIS-Group-Lasso performs well at the cost of increasing the size of the model, which leads to large FP and ME. On the other hand, the number of false-positive groups selected by group Lasso is much larger than the other four methods. Compared with distance correlation-based screening methods, our approaches have better finite performance. This may be due to the reasons for their methods based on model-free framework, while our screening methods take full advantage of the assumptions of the linear model. Especially for the greedy modification, g-ISIS-Group-Lasso, the size of final selected model is much smaller than the other four methods in all examples. Just because of this, g-ISIS-Group-Lasso outperforms its competitors in terms of the percentage of correct selected model. On the other hand, the simulations show that the value of SNR has an important impact on the results of the three methods. In Examples 4.2 and 4.4, there are eight relevant groups, while the number of relevant groups in the other two examples is 4. It means that screening relevant groups in these two examples is difficult than the other two. Thus, if we want achieve sure screening, the value of SNR in these two examples is much larger than the other two.

4.2. Real example

In this section, we compare ISIS-Group-Lasso, g-ISIS-Group-Lasso and Group-Lasso on colon data (Alon et al., 1999). Alon's work reports the application of a two-way clustering method for analysis a data set consisting of the expression patterns of different cell types. For these data, we were interested in finding the genes that are related to colon tumour. In the original colon data, the identity of the 62 samples from colon-cancer

Table 1. Simulation results of MEDIAN, TP, FP, CORRECT and ME for Example 4.1.

p	Method	MEDIAN	TP	FP	CORRECT	ME
$p = 200$	ISIS-Group-Lasso	6.00	4.00 (0.00)	2.88 (2.99)	0.12	2.19 (0.32)
	g-ISIS-Group-Lasso	4.00	4.00 (0.00)	0.76 (0.75)	0.51	1.90 (0.38)
	Group Lasso	16.00	4.00 (0.00)	11.78 (2.24)	0.00	2.82 (0.36)
	DC-SIS-Group-Lasso	9.00	3.13 (0.00)	5.76 (2.24)	0.00	4.05 (0.28)
	DC-ISIS-Group-Lasso	13.00	4.00 (0.00)	8.52 (2.24)	0.00	2.73 (0.35)
$p = 400$	ISIS-Group-Lasso	7.00	4.00 (0.00)	3.43 (2.99)	0.09	2.30 (0.39)
	g-ISIS-Group-Lasso	4.00	4.00 (0.00)	0.76 (0.75)	0.51	1.92 (0.46)
	Group Lasso	20.00	4.00 (0.00)	16.11 (2.24)	0.00	2.93 (0.38)
	DC-SIS-Group-Lasso	11.00	3.09 (0.00)	7.48 (2.24)	0.00	4.15 (0.29)
	DC-ISIS-Group-Lasso	14.00	4.00 (0.00)	10.17 (1.49)	0.00	2.77 (0.33)
$p = 1000$	ISIS-Group-Lasso	7.00	4.00 (0.00)	3.32 (2.99)	0.14	2.34 (0.49)
	g-ISIS-Group-Lasso	4.00	3.99 (0.00)	0.83 (0.75)	0.51	1.97 (0.50)
	Group Lasso	28.00	4.00 (0.00)	24.15 (2.99)	0.00	3.02 (0.32)
	DC-SIS-Group-Lasso	13.00	3.04 (0.00)	9.81 (2.24)	0.00	4.17 (0.24)
	DC-ISIS-Group-Lasso	16.00	4.00 (0.00)	11.80 (2.24)	0.00	2.80 (0.32)

Note: Robust standard deviations are given in parentheses. ($\sigma = 5$, $SNR = 2.92$.)

Table 2. Simulation results of MEDIAN, TP, FP, CORRECT and ME for Example 4.2.

p	Method	MEDIAN	TP	FP	CORRECT	ME
$p = 200$	ISIS-Group-Lasso	10.00	8.00 (0.00)	2.89 (2.24)	0.12	1.81 (0.34)
	g-ISIS-Group-Lasso	8.00	8.00 (0.00)	0.73 (0.75)	0.52	1.58 (0.33)
	Group Lasso	34.00	8.00 (0.00)	26.44 (3.73)	0.00	2.74 (0.33)
	DC-SIS-Group-Lasso	19.00	6.22 (0.75)	13.23 (2.24)	0.00	4.40 (1.04)
	DC-ISIS-Group-Lasso	23.00	7.94 (0.00)	14.85 (2.24)	0.00	3.15 (0.44)
$p = 400$	ISIS-Group-Lasso	11.00	8.00 (0.00)	3.08 (2.99)	0.13	1.87 (0.37)
	g-ISIS-Group-Lasso	9.00	8.00 (0.00)	0.87 (0.75)	0.44	1.66 (0.33)
	Group Lasso	45.00	8.00 (0.00)	36.86 (3.73)	0.00	2.95 (0.31)
	DC-SIS-Group-Lasso	21.00	5.47 (0.75)	15.39 (2.24)	0.00	4.99 (0.79)
	DC-ISIS-Group-Lasso	25.00	7.90 (0.00)	17.33 (2.24)	0.00	3.35 (0.43)
$p = 1000$	ISIS-Group-Lasso	11.00	8.00 (0.00)	3.37 (2.99)	0.12	1.92 (0.43)
	g-ISIS-Group-Lasso	8.00	8.00 (0.00)	0.83 (0.75)	0.51	1.66 (0.43)
	Group Lasso	60.00	7.99 (0.00)	52.18 (3.92)	0.00	3.13 (0.33)
	DC-SIS-Group-Lasso	24.00	4.86 (0.75)	19.02 (2.99)	0.00	5.29 (0.62)
	DC-ISIS-Group-Lasso	27.00	7.78 (0.00)	19.43 (2.24)	0.00	3.50 (0.47)

Note: Robust standard deviations are given in parentheses. ($\sigma = 4$, $SNR = 5.34$.)

Table 3. Simulation results of MEDIAN, TP, FP, CORRECT and ME for Example 4.3.

p	Method	MEDIAN	TP	FP	CORRECT	ME
$p = 200$	ISIS-Group-Lasso	7.00	4.00 (0.00)	2.87 (2.24)	0.12	1.64 (0.32)
	g-ISIS-Group-Lasso	4.00	4.00 (0.00)	0.75 (0.75)	0.56	1.37 (0.35)
	Group Lasso	19.00	4.00 (0.00)	15.42 (2.24)	0.00	2.15 (0.28)
	DC-SIS-Group-Lasso	12.00	3.68 (0.75)	8.33 (2.43)	0.00	2.41 (0.67)
	DC-ISIS-Group-Lasso	13.00	4.00 (0.00)	8.73 (1.49)	0.00	2.03 (0.27)
$p = 400$	ISIS-Group-Lasso	6.00	4.00 (0.00)	2.88 (2.24)	0.14	1.64 (0.43)
	g-ISIS-Group-Lasso	4.00	4.00 (0.00)	0.74 (0.75)	0.56	1.37 (0.40)
	Group Lasso	27.00	3.99 (0.00)	22.86 (2.99)	0.00	2.22 (0.26)
	DC-SIS-Group-Lasso	13.00	3.51 (0.75)	9.25 (2.24)	0.00	2.64 (0.71)
	DC-ISIS-Group-Lasso	14.00	4.00 (0.00)	9.88 (1.49)	0.00	2.07 (0.28)
$p = 1000$	ISIS-Group-Lasso	7.00	4.00 (0.00)	3.20 (2.99)	0.10	1.77 (0.37)
	g-ISIS-Group-Lasso	5.00	3.99 (0.00)	0.79 (0.75)	0.44	1.47 (0.41)
	Group Lasso	40.00	4.00 (0.00)	36.01 (3.17)	0.00	2.33 (0.26)
	DCS-Group-Lasso	15.00	3.43 (0.75)	11.71 (2.24)	0.00	2.71 (0.56)
	DC-ISIS-Group-Lasso	15.00	4.00 (0.00)	11.53 (2.24)	0.00	2.15 (0.26)

Note: Robust standard deviations are given in parentheses. ($\sigma = 4$, $SNR = 2.69$.)

patients were analysed with an Affymetrix oligonucleotide Hum6000 array. And these data contain the expression of the 2000 genes with the highest minimal intensity across the 62 tissues, where the genes are placed in order of descending minimal intensity. That is, the original data have 2000 numerical variables. For each continuous variable in the additive model, we use five B-spline basis functions to represent its effect, which is originally used by Yang and Zou (2015) in solving group-lasso penalise learning problems. Thus we obtain 10,000 predictors in 2000 groups after basis function expansion. Three methods are used to select relevant additive components.

Before the analysis, all data are standardised in advance such that each variable has zero mean and unit sample variance. To valuate the performance of the three methods, we used cross-validation and compared the average model size (AMS) and the prediction mean squared error (PE). We randomly partitioned the data into a training data set of 50 observations and a test set

Table 4. Simulation results of MEDIAN, TP, FP, CORRECT and ME for Example 4.4.

p	Method	MEDIAN	TP	FP	CORRECT	ME
$p = 200$	ISIS-Group-Lasso	11.00	8.00 (0.00)	3.45 (2.99)	0.13	2.08 (0.43)
	g-ISIS-Group-Lasso	9.00	8.00 (0.00)	0.79 (0.75)	0.46	1.80 (0.32)
	Group Lasso	30.00	8.00 (0.00)	22.17 (2.99)	0.00	3.15 (0.30)
	DC-SIS-Group-Lasso	14.00	5.70 (0.75)	8.08 (2.99)	0.00	5.53 (0.91)
	DC-ISIS-Group-Lasso	18.00	7.52 (0.75)	10.23 (2.24)	0.00	3.52 (1.16)
$p = 400$	ISIS-Group-Lasso	11.00	8.00 (0.00)	3.96 (2.24)	0.07	2.19 (0.38)
	g-ISIS-Group-Lasso	9.00	7.98 (0.00)	0.79 (0.75)	0.49	1.86 (0.34)
	Group Lasso	37.00	7.97 (0.00)	29.55 (3.17)	0.00	3.35 (0.41)
	DC-SIS-Group-Lasso	14.00	5.26 (0.75)	8.99 (2.24)	0.00	5.87 (0.82)
	DC-ISIS-Group-Lasso	18.00	7.44 (0.75)	11.25 (2.99)	0.00	3.69 (1.12)
$p = 1000$	ISIS-Group-Lasso	11.00	8.00 (0.00)	3.91 (2.24)	0.07	2.24 (0.39)
	g-ISIS-Group-Lasso	8.00	7.94 (0.00)	0.75 (0.75)	0.50	1.92 (0.34)
	Group Lasso	49.00	7.91 (0.00)	40.75 (4.48)	0.00	3.60 (0.36)
	DC-SIS-Group-Lasso	16.00	4.88 (0.75)	11.19 (2.99)	0.00	6.07 (0.54)
	DC-ISIS-Group-Lasso	20.00	7.40 (0.75)	12.77 (2.24)	0.00	3.77 (1.11)

Note: Robust standard deviations are given in parentheses. ($\sigma = 4$, $SNR = 6.78$.)

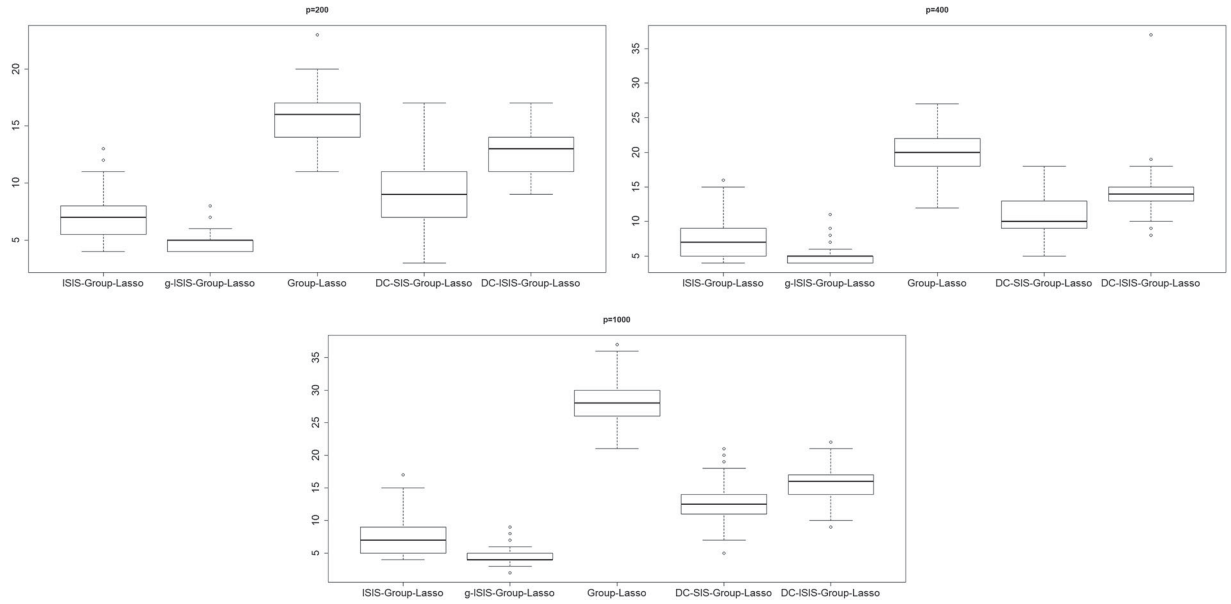

Figure 1. Boxplots of average model sizes for Example 4.1 under different group number.

Table 5. Results of AMS and CORRECT for colon data.

Method	AMS	PE
ISIS-Group-Lasso	6.58 (2.24)	2.08 (0.34)
g-ISIS-Group-Lasso	3.6 (1.49)	2.10 (0.35)
Group Lasso	18.49 (2.24)	3.09 (0.58)

Note: Robust standard deviations are given in parentheses.

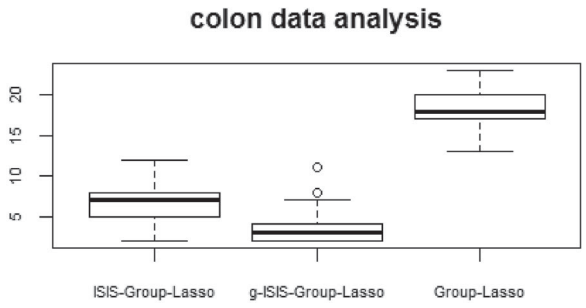

Figure 2. Boxplot of average model sizes for colon data analysis.

of 12 observations. That is, we conduct group screening using 50 observations and the PEs on these 12 test sets. Detailed results based on 100 replications are presented in Table 5. In addition, the boxplot of the average model size is presented in Figure 2. As clearly shown in Table 5, ISIS-Group-Lasso and g-ISIS-Group-Lasso select far fewer genes than Group-Lasso, while the first two methods have a smaller PE. In conclusion, the proposed iterative group screening approach is very useful in high-dimensional scientific studies, which can select a parsimonious model and reveal interesting relationship between group variables.

5. Concluding remarks

In this article, we have proposed the marginal group sure screening method under the context of ultra-high dimensionality. Unlike most existing literatures, we deal with variables, which can be naturally grouped. Our group screening method respects the grouping structure in the data and is based on a working

independence. Theoretically, we establish the sure screening property for this group screening approach. To enhance the finite sample performance, a data-driven thresholding and an iterative procedure, ISIS-Group-Lasso, are developed. A greedy modification to the iterative procedure, g-ISIS-Group-Lasso is also proposed to further reduce the false positive. Simulation results show that these two methods perform well in terms of the five performance measures.

This article leaves the problems of extending the ISIS-Group-Lasso and g-ISIS-Group-Lasso under linear model to the family of generalised linear model and other parametric models. And model-free group screening approach may be appealing for dealing with ultra-high-dimensional data more generally, which avoids the difficult task of specifying the form of a statistical model. These problems are beyond the scopes of this article and are interesting topics for future research.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by the National Natural Science Foundation of China (CN) (11571112), the National Social Science Foundation Key Program (17ZDA091) and Natural Science Fund of Education Department of Anhui Province (KJ2013B233), the 111 Project of China (B14019).

Notes on contributors

Yong Niu is a PhD candidate in the College of Statistics, East China Normal University, Shanghai, China. His research interests include high dimensional data, big data analytics and nonparametric statistics.

Riquan Zhang is a professor and chair of School of Statistics in East China Normal University. His research interests include high dimensional data, big data analytics, functional data analysis, statistical machine learning and nonparametric statistics.

Jicai Liu is an associate professor of statistics in the department of mathematics at Shanghai Normal University, China. His research interests include high dimensional data, lifetime data analysis and nonparametric statistics.

Huapeng Li is an associate professor of statistics in the school of mathematics and statistics at Datong University, China. His research interests include nonparametric and semiparametric statistics based on empirical likelihood, selection biased data and finite mixture models.

References

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, *96*(12), 6745–6750.

Bakin, S. (1999). *Adaptive regression and model selection in data mining problems* (Ph.D. thesis). Australian National University, Canberra.

Breheny, P., & Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and its Interface*, *2*(3), 369.

Breheny, P., & Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, *25*(2), 173–187.

Fan, J., Feng, Y., & Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, *106*(494), 544–557.

Fan, J., & Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *70*(5), 849–911.

Fan, J., Samworth, R., & Wu, Y. (2009). Ultrahigh dimensional variable selection: Beyond the linear model. *Journal of Machine Learning Research*, *10*, 1829–1853.

Fan, J., & Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Annals of Statistics*, *38*(6), 3567–3604.

He, X., Wang, L., & Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Annals of Statistics*, *41*(1), 342–369.

Huang, J., Ma, S., Xie, H., & Zhang, C. H. (2009). A group bridge approach for variable selection. *Biometrika*, *96*(2), 339–355.

Li, R., Zhong, W., & Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, *107*(499), 1129–1139.

Shao, X., & Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *The American Statistical Association*, *109*(507), 1302–1318.

Van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes*. New York: Springer.

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

Wei, F., & Huang, J. (2010). Consistent group selection in high-dimensional linear regression. *Bernoulli*, *16*(4), 1369–1384.

Yang, Y., & Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, *2015*(6), 1129–1141.

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.

Zhang, C. H., & Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, *36*(4), 1567–1594.

Zhao, S. D., & Li, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis*, *105*(1), 397–411.

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

Zhong, W., & Zhu, L.-P. (2015). An iterative approach to distance correlation-based sure independence screening. *Journal of Statistical Computation and Simulation*, *85*(11), 2331–2345.

Appendix 1. Three lemmas

Next, we state some lemmas which will be used in the proof of Theorems 3.1 and 3.2.

Lemma A.1: *Under conditions* (i)–(iii), *for any* $\delta > 0$, $j = 1, 2, \ldots, J$, *we have*

$$P\left\{\frac{1}{\sqrt{p_j}}\left\|\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{y} - E\boldsymbol{X}_j Y\right\| \geq \frac{\delta}{n}\right\} \leq 4p_j \exp\{-\delta^2/(c_2 n + c_3\delta)\},$$

where $c_2 = \max(8M_0^2 M_1, 16M_2)$ *and* $c_3 = \max\left(\frac{1}{3}M_0 M_1, 1\right)$.

Proof: Using Bonferroni's inequality, we can easily prove that

$$\begin{aligned} &P\left\{\frac{1}{\sqrt{p_j}}\left\|\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{y} - E\boldsymbol{X}_j Y\right\| \geq \frac{\delta}{n}\right\} \\ &\leq P\left\{\bigcup_{k=1}^{p_j}\left(\left[\frac{1}{n}\sum_{i=1}^{n}(x_{ijk}y_i - EX_{jk}Y)\right]^2 \geq \frac{\delta^2}{n^2}\right)\right\} \\ &\leq \sum_{k=1}^{p_j} P\left\{\left|\sum_{i=1}^{n}(x_{ijk}y_i - EX_{jk}Y)\right| \geq \delta\right\}. \end{aligned}$$

Thus we need to show that

$$P\left\{\left|\sum_{i=1}^{n}(x_{ijk}y_i - EX_{jk}Y)\right| \geq \delta\right\} \leq 4\exp\{-\delta^2/(c_2 n + c_3\delta)\},$$

for every $k = 1, 2, \ldots, p_j$. Recall that the support of X_{jk} is $[a, b]$ for $j = 1, 2, \ldots, J$ and $k = 1, 2, \ldots, p_j$, we denote $M_0 = \max(|a|, |b|)$. Because $y_i = \sum_{j=1}^{J}\mathbf{x}_{ij}^{\mathrm{T}}\beta_j + \varepsilon_i$, we can obtain that

$$\begin{aligned} x_{ijk}y_i - EX_{jk}Y &= \left\{x_{ijk}\left(\sum_{j=1}^{J}\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_j\right) - E\left[x_{ijk}\left(\sum_{j=1}^{J}\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_j\right)\right]\right\} \\ &\quad + x_{ijk}\varepsilon_i \\ &\widehat{=} S_{ijk1} + S_{ijk2}. \end{aligned}$$

Next we bound the tails probability of $|S_{ijk1}|$ and $|S_{ijk2}|$ respectively. By condition (ii)–(iii), it is easy to see that

$$\begin{aligned} |S_{ijk1}| &\leq M_0 M_1, \\ \mathrm{Var}(S_{ijk1}) &\leq M_0^2 M_1^2, \\ E|S_{ijk2}|^m &\leq EM_0^m|\varepsilon_i|^m \\ &\leq m!E\exp M_0|\varepsilon_i| \leq M_2 m! \quad (m \geq 2). \end{aligned}$$

Using the Bernstein's inequality (Van der Vaart & Wellner, 1996, lemma 2.2.9 and lemma 2.2.11), we conclude that

$$P\left\{\left|\sum_{i=1}^{n}S_{ijk1}\right| > \frac{\delta}{2}\right\} \leq 2\exp\left\{-\frac{\delta^2}{8}\frac{1}{nM_0^2M_1^2 + M_0M_1\delta/6}\right\}, \tag{A1}$$

$$P\left\{\left|\sum_{i=1}^{n}S_{ijk2}\right| > \frac{\delta}{2}\right\} \leq 2\exp\left\{-\frac{\delta^2}{8}\frac{1}{2nM_2 + \delta/2}\right\}. \tag{A2}$$

Therefore, we combine the results (A1) and (A2) with $c_2 = \max(8M_0^2M_1, 16M_2)$ and $c_3 = \max\left(\frac{1}{3}M_0M_1, 1\right)$, to obtain that

$$P\left\{\left|\sum_{i=1}^{n}(x_{ijk}y_i - EX_{jk}Y)\right| \geq \delta\right\} \leq 4\exp\{-\delta^2/(c_2 n + c_3\delta)\}.$$

This concludes the proof of the lemma. ■

Lemma A.2: *Under conditions* (ii)–(iv), *for any* $\delta > 0$ *and* $j = 1, 2, \ldots, J$, *we have*

$$P\left\{\frac{1}{p_j}\left\|\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j - E\boldsymbol{X}_j\boldsymbol{X}_j^{\mathrm{T}}\right\| \geq \frac{1}{n}\delta\right\} \leq 2p_j^2\exp\left\{-\frac{\delta^2}{c_4 n + c_5\delta}\right\}$$

where $c_4 = 2M_0^4$ *and* $c_5 = \frac{4M_0^2}{3}$.

Proof: For $s = 1, 2, \ldots, p_j$, $t = 1, 2, \ldots, p_j$, let $T_j = \frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\mathbf{x}_j - EX_jX_j^{\mathrm{T}}$ and $T_j^{(s,t)}$ be the entry of T_j. Then we can write $T_j^{(s,t)} = \frac{1}{n}\sum_{i=1}^{n}(x_{ijs}x_{ijt} - EX_{js}X_{jt})$.

By the fact that $\|A\| \leq p\|A\|_\infty$, we have

$$\begin{aligned} &P\left\{\frac{1}{p_j}\left\|\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j - E\boldsymbol{X}_j\boldsymbol{X}_j^{\mathrm{T}}\right\| \geq \frac{1}{n}\delta\right\} \\ &\leq P\left\{\left\|\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j - E\boldsymbol{X}_j\boldsymbol{X}_j^{\mathrm{T}}\right\|_\infty \geq \frac{1}{n}\delta\right\} \\ &\leq \sum_{s=1}^{p_j}\sum_{t=1}^{p_j}P\left\{|T_j^{(s,t)}| \geq \frac{\delta}{n}\right\}. \end{aligned} \tag{A3}$$

Next we also use Bernstein's inequality to bound the tails probability of $T_j^{(s,t)}$. By condition (ii)–(iii), we can obtain easily that

$$\begin{aligned} |x_{ijs}x_{ijt} - EX_{js}X_{jt}| &\leq 2M_0^2, \\ \mathrm{Var}(x_{ijs}x_{ijt}) &\leq M_0^4. \end{aligned}$$

Using Bernstein's inequality, it follows that

$$P\left\{|T_j^{(s,t)}| \geq \frac{\delta}{n}\right\} \leq 2\exp\left\{-\frac{1}{2}\frac{\delta^2}{nM_0^4 + 2M_0^2\delta/3}\right\}. \tag{A4}$$

Thus the desired result is obtained from (A3) and (A4) by taking $c_4 = 2M_0^4$ and $c_5 = \frac{4M_0^2}{3}$. ■

Remark A.1: If A and B are two symmetric matrices of order p, we have the following two results (Fan, Feng, & Song, 2011; He et al., 2013):

$$\begin{aligned} |\lambda_{\min}(A) - \lambda_{\min}(B)| &\leq \max\{|\lambda_{\min}(A - B)|, \lambda_{\min}(B - A)|\}, \\ |\lambda_{\max}(A) - \lambda_{\max}(B)| &\leq \max\{|\lambda_{\max}(A - B)|, \lambda_{\max}(B - A)|\}. \end{aligned}$$

In addition, note that

$$|\lambda_{\min}(A - B)| \leq |\lambda_{\max}(A - B)| \leq p\|A - B\|_\infty.$$

The above results, together with Lemma 2, imply that

$$\begin{aligned} &P\left\{\left|\lambda_{\min}\left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\right) - \lambda_{\min}(E\boldsymbol{X}_j\boldsymbol{X}_j')\right| \geq \frac{p_j}{n}\delta\right\} \\ &\leq 2p_j^2\exp\left\{-\frac{\delta^2}{c_4 n + c_5\delta}\right\}, \end{aligned} \tag{A5}$$

$$\begin{aligned} &P\left\{\left|\lambda_{\max}\left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\right) - \lambda_{\max}(E\boldsymbol{X}_j\boldsymbol{X}_j^{\mathrm{T}})\right| \geq \frac{p_j}{n}\delta\right\} \\ &\leq 2p_j^2\exp\left\{-\frac{\delta^2}{c_4 n + c_5\delta}\right\} \end{aligned} \tag{A6}$$

for $j = 1, 2, \ldots, p_j$.

Lemma A.3: *Suppose conditions* (ii)–(iv) *hold, there exist some positive constants* τ_3 *and* τ_4 *such that*

$$P\left\{\tau_3 \le \lambda_{\min}\left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\right) \le \lambda_{\max}\left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\right) \le \tau_4\right\} \ge 1 - 2p_j^2 \exp\left\{-\frac{\delta^2}{c_4 n + c_5\delta}\right\}. \tag{A7}$$

That is, with probability approaching 1, *we have*

$$0 < \tau_3 \le \lambda_{\min}\left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\right) \le \lambda_{\max}\left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\right) \le \tau_4 < \infty.$$

Proof: Combing condition (iv) and (A5)–(A6), it is easy to obtain (A7). ■

Appendix 2. Proof of Theorem 3.1

Proof of Theorem 3.1.: The key idea of the proof is to show the uniform consistence of $\|\hat{\upsilon}_{nj}\|_n^2$ under conditions (ii)–(iv). As to the existing literatures, the sure screening property is typically established in this way. Recall that

$$\|\hat{\upsilon}_{nj}\|_n^2 = \left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{y}\right)^{\mathrm{T}}\left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\right)^{-1}\left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{y}\right)$$

and

$$\|\upsilon_j\|^2 = (E\boldsymbol{X}_jY)^{\mathrm{T}}(E\boldsymbol{X}_j\boldsymbol{X}_j^{\mathrm{T}})^{-1}(E\boldsymbol{X}_jY).$$

Thus we need to evaluate

$$\frac{1}{p_j}\|\hat{\upsilon}_{nj}\|_n^2 - \frac{1}{p_j}\|\upsilon_j\|^2 = \frac{1}{p_j}\left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{y}\right)^{\mathrm{T}}\left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\right)^{-1}\left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{y}\right) - \frac{1}{p_j}(E\boldsymbol{X}_jY)^{\mathrm{T}}(E\boldsymbol{X}_j\boldsymbol{X}_j^{\mathrm{T}})^{-1}(E\boldsymbol{X}_jY).$$

By some algebra, we decompose it into three parts

$$\|\hat{\upsilon}_{nj}\|_n^2 - \|\upsilon_j\|^2 = \lambda_1 + \lambda_2 + \lambda_3$$

in which

$$\lambda_1 = \left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{y} - E\boldsymbol{X}_jY\right)^{\mathrm{T}}\left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\right)^{-1}\left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{y} - E\boldsymbol{X}_jY\right),$$

$$\lambda_2 = \left(\frac{1}{n}\boldsymbol{x}_j'\boldsymbol{y} - E\boldsymbol{X}_jY\right)^{\mathrm{T}}\left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\right)^{-1}E\boldsymbol{X}_jY,$$

$$\lambda_3 = (E\boldsymbol{X}_jY)^{\mathrm{T}}\left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\right)^{-1}(E\boldsymbol{X}_j\boldsymbol{X}_j^{\mathrm{T}} - \frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j)(E\boldsymbol{X}_j\boldsymbol{X}_j^{\mathrm{T}})^{-1}E\boldsymbol{X}_jY.$$

Now, we define a event Ω_δ on which we have

$$\frac{1}{\sqrt{p_j}}\left\|\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{y} - E\boldsymbol{X}_jY\right\| \le \frac{\delta}{n},$$

$$\frac{1}{p_j}\left\|\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j - E\boldsymbol{X}_j\boldsymbol{X}_j^{\mathrm{T}}\right\| \le \frac{1}{n}\delta,$$

$$\tau_3 \le \lambda_{\min}\left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\right) \le \lambda_{\max}\left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\right) \le \tau_4$$

for $j = 1, 2, \ldots, J$.

Then the above three lemmas indicate that

$$P(\Omega_\delta) \ge 1 - 4\sum_{j=1}^{J} p_j \exp\{-\delta^2/(c_2 n + c_3\delta)\} - 4\sum_{j=1}^{J} p_j^2 \exp\{-\delta^2/(c_4 n + c_5\delta)\}. \tag{A8}$$

By the fact that $\|AB\| \le \|A\|\|B\|$, we have on Ω_δ,

$$\frac{1}{p_j}|\lambda_1| \le \left\|\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{Y} - E\boldsymbol{X}_jY\right\|^2\left\|\left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\right)^{-1}\right\| \le \frac{\delta^2}{n^2}\frac{1}{\tau_3},$$

$$\frac{1}{p_j}|\lambda_2| \le 2\left\|\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{y} - E\boldsymbol{X}_jY\right\|\left\|\left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\right)^{-1}\right\|$$

$$\|E\boldsymbol{X}_jY\| \le \frac{\delta}{n}\frac{2M_0M_1}{\tau_3},$$

$$\frac{1}{p_j}|\lambda_3| \le \left\|\left(\frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\right)^{-1}\right\|\|(E\boldsymbol{X}_j\boldsymbol{X}_j^{\mathrm{T}})^{-1}\|\left\|E\boldsymbol{X}_j\boldsymbol{X}_j^{\mathrm{T}} - \frac{1}{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\right\|$$

$$\|E\boldsymbol{X}_jY\|^2 \le \frac{\delta}{n}\frac{M_0M_1}{\tau_1\tau_3}.$$

Take $\delta = c_6 n^{1-\kappa}$, there exists a constant c_7 such that

$$\frac{\delta^2}{n^2}\frac{1}{\tau_3} + \frac{\delta}{n}\frac{2M_0M_1}{\tau_3} + \frac{\delta}{n}\frac{M_0M_1}{\tau_1\tau_3} \le c_6c_7n^{-\kappa}.$$

Choosing c_6 such that $c_6c_7 \le c$, we can easily obtain that

$$\left|\left\|\frac{1}{p_j}\hat{\upsilon}_{nj}\right\|_n^2 - \frac{1}{p_j}\|\upsilon_j\|^2\right| \le cn^{-\kappa}.$$

By invoking condition (i), we have on Ω_δ for sufficiently large n

$$\left\|\frac{1}{p_j}\hat{\upsilon}_{nj}\right\|_n^2 \ge 2cn^{-\kappa}.$$

If we choose $\pi_n \le 2cn^{-\kappa}$, it is easy to show that $j \in \mathcal{M}_*$. This, together with (A8), indicates that there exists a constant c_1 such that

$$P\{\mathcal{M}_* \subset \hat{\mathcal{M}}_\kappa\} \ge 1 - 4\sum_{j=1}^{J}(p_j + p_j^2)\exp\{-c_1n^{1-2\kappa}\}.$$ ■

Appendix 3. Proof of Theorem 3.2

Proof of Theorem 3.2.: Following the similar argument of the proof of Theorem 3.1, we have on Ω_δ

$$\left|\left\{1 \le j \le J : \frac{1}{p_j}\|\hat{\upsilon}_{nj}\|_n^2 \ge 2cn^{-\kappa}\right\}\right| \le \left|\left\{1 \le j \le J : \frac{1}{p_j}\|\upsilon_j\|^2 \ge cn^{-\kappa}\right\}\right|,$$

where $|.|$ denotes the size of the set. This implies that

$$\sum_{j\in\hat{\mathcal{M}}_\kappa}\frac{1}{p_j}\|\upsilon_j\|^2 \ge cn^{-\kappa}|\hat{\mathcal{M}}_\kappa|.$$

By some algebra, it follows that $|\hat{\mathcal{M}}_\kappa| \le O(n^\kappa\sum_{j=1}^{J}\|E\boldsymbol{X}_jY\|^2) = O(n^\kappa\|E\boldsymbol{X}Y\|^2)$. That is, we have

$$P\{|\hat{\mathcal{M}}_\kappa| \le O(n^\kappa\|E\boldsymbol{X}Y\|^2)\} \ge 1 - 4\sum_{j=1}^{J}(p_j + p_j^2)\exp\{-c_1n^{1-2\kappa}\}. \tag{A9}$$

Thus the key point is to show that $\|E\boldsymbol{X}Y\|^2 = O(1)$. For this purpose, we consider the following linear regression:

$$\min_{\boldsymbol{\alpha}} E(Y - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\alpha})^2$$

with respect to $\boldsymbol{\alpha} \in \mathbb{R}^{\sum_{j=1}^{J} p_j}$. By least square, we can easily obtain that

$$\|E\boldsymbol{X}Y\|^2 = \hat{\boldsymbol{\alpha}}'[E(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})]^2\hat{\boldsymbol{\alpha}} \leq \lambda_{\max}(\Sigma)\hat{\boldsymbol{\alpha}}'E(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})\hat{\boldsymbol{\alpha}}$$

in which $\hat{\boldsymbol{\alpha}}$ is the least square estimator. On the other hand, the orthogonal decomposition of least square implies that $\mathrm{Var}(Y) = \mathrm{Var}(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{\alpha}}) + \mathrm{Var}(Y - \boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{\alpha}})$. Because $\mathrm{Var}(Y) = O(1)$, we conclude that

$$\|E\boldsymbol{X}Y\|^2 \leq O(1). \tag{A10}$$

Combining (A9) and (A10), the desired result can be easily obtained. ■