Grouped Variable Screening for Ultrahigh Dimensional Data under Linear Model

by

Debin Qiu

(Under the direction of Jeongyoun Ahn)

Abstract

High or ultrahigh dimensional data set with group structure emerge in a wide range of scientific research and applications nowadays. However, sparsity may exist in this high or ultrahigh dimensional data with such group form. In such case, our primary goal is to select the important groups that are significantly correlated with outcome. In particular, grouped variable selection plays a critical role in selecting groups and estimating the nonzero coefficients for these covariates within these important groups. Nevertheless, in the presence of ultra-high dimensional data consisting of grouped variables, many algorithms for grouped variable selection may fail to converge or yield insensible results. Even if the algorithm works, it will suffer from a rather intensive computation load.

In this dissertation, we propose a two-stage procedure, grouped variable screening and selection, to solve those challenging issues. At the first stage, grouped variable screening is applied to reduce the dimensionality of data by filtering out the unimportant groups that have no contribution to outcome. A sure screening property is established to ensure an overwhelming probability of retaining all important groups after the screening procedure under suitable conditions. This work will mainly focus on four grouped variable screening criteria.

At the second stage, since the data have been reduced from ultra-high dimensionality to the moderate one or even lower than sample size, grouped variable selection methods are able to select the important groups effectively and estimate the nonzero coefficients accurately. Meanwhile, the computation can be decreased dramatically in terms of running time and complexity when executing the grouped variable selection. The performance of the proposed two-stage procedure is evaluated by various simulated examples and a real data set in genetic analysis. An R package called `grpss` is developed to incorporate the two-stage procedure into real applications.

INDEX WORDS: grouped variables, grouped variable selection, grouped variable screening, marginal correlation learning, penalized regression, random permutation, sure screening property

Grouped Variable Screening for Ultrahigh Dimensional Data Under Linear Model

by

Debin Qiu

B.S., Soochow University, 2010

M.S., Soochow University, 2012

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2016

Grouped Variable Screening for Ultrahigh Dimensional Data Under Linear Model

by

Debin Qiu

Approved:

Major Professors: Jeongyoun Ahn

Committee: Lily Wang
Pengsheng Ji
William McCormick
Wenxuan Zhong

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2016

Dedication

To my beloved wife (Yicheng Wang) , my parents and sister.

Acknowledgments

First and foremost, I would like to thank my PhD advisor Jeongyoun Ahn who gave me this exciting and interesting research topic. I really want to appreciate her for the tremendous patience and valuable support in answering my questions and doubts, revising my manuscript, giving me constructive advice for my research and career. Without her strong guidance and valuable suggestions, I would never finish such a wonderful dissertation.

I would also thank my committee members Dr. William McCormick, Dr. Pengsheng Ji, Dr. Wenxuan Zhong and Dr. Lily Wang who gave me their scientific insights and suggestions for my dissertation. I am truly grateful for their contribution on my research and for the amount of time they spent on reading my dissertation. I must acknowledge the professors and faculties in the department to help me lay a solid foundation in methodology and theory in statistics. I want to give my special thanks to Dr. Lijian Yang, my former master advisor, who gave me great encouragements in pursuing doctoral study in statistics and wrote a very supportive recommendation letter to bring me into this great statistics department in 2012.

I also want to thank my graduate coordinator Dr. Lynne Seymour for recruiting me as a graduate student, giving me tremendous assistance and bringing me to the get-together that made my life enjoyable in the U.S.

I also want to express my gratitude to my officemates Adam Jaeger, Haileab Hilafu, Xinlian Zhang, Chul Moon, Chris Helms and my colleagues Yuanwen Wang, Fei Liu, Yan Zhen, Ionan Alexei, Lina Liao, Xianyan Chen, Ye Wang, Xijue Tan, Guannan Wang, Rui Xie and many others who shared their interesting stories and gave me a lot of fun in my life of studying and working in the Department of Statistics at University of Georgia.

Last but not least, I would like to thank my wife Yicheng Wang and my parents for giving their perpetual and absolute love and trust.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the rapid development and great aid of modern technology, large and complex data sets with tens of thousands of variables are easily accessible and collected nowadays with fairly low or even no costs. In a wide range of scientific research and applications, such as micro-arrays, genomics and brain images, quantitative measurements are used to study the connections between outcomes and explanatory variables. In many cases, the explanatory variables are naturally grouped because they are similar in some behaviors, or have similar effects, or have a high correlation between each other within groups. Groups can be even specified from the experts' knowledge. Another scenario is that researchers wish to accelerate the analysis process or to improve the analysis accuracy under the consideration of group structure. The groups of explanatory variables can be formed by an ad-hoc k-means clustering method or some other clustering methods in the case there is no any prior information of the groups. The subsequent analysis can take advantage of this prior knowledge of group structure and gives rise to a more meaningful and interpretable result. Common examples of using grouped variables can be seen in multifactor analysis of variance and nonparametric additive regression. In ANOVA problem, a multiple-level factor can be expressed into a group of dummy variables. In nonparametric additive model, each additive component can

be represented by a linear combination of a series of basis functions such as polynomial basis, B-spline basis or Fourier basis. Another concrete example includes the statistical applications in biological study, where genes can form groups based on their pathway they belong to or some other biological characteristics. In these modeling problems, we can essentially improve the interpretability and accuracy of the models by taking the group structure into account.

Consider a linear regression model with p predictors. Suppose the predictors can be naturally divided into J non-overlapping groups. The linear model can be then written as

$$Y = \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j + \varepsilon = \sum_{j=1}^{J} \sum_{k=1}^{p_j} X_{jk} \beta_{jk} + \varepsilon, \tag{1.1}$$

where $Y = (y_1, \cdots, y_n)^T$ is an $n \times 1$ vector of response, $\mathbf{X}_j = (X_{j1}, \cdots, X_{jp_j})$ is an $n \times p_j$ design matrix of the p_j predictors in the j-th group, $\boldsymbol{\beta}_j = (\beta_{j1}, \cdots, \beta_{jp_j})^T \in \mathbb{R}^{p_j}$ is the $p_j \times 1$ vector of regression coefficients of the j-th group and $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^T$ is the independent and identically distributed (i.i.d) random errors with mean 0 and variance σ^2, i.e., $E(\varepsilon) = 0, E(\varepsilon^2) = \sigma^2 I_n$. Note that the total number of variables is the sum of the number of variables in each group, i.e., $p = \sum_{j=1}^{J} p_j$. Without the imposed group structure, model (1.1) can be expressed as $Y = \mathbf{X}\boldsymbol{\beta} + \varepsilon$ which is a classical multiple linear regression, where $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_J)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_J^T)$. In effect, the linear model with p individual predictors can be regarded as a special case of the model with J groups with single predictor if $p_1 = p_2 = \cdots = p_J = 1$, resulting in $J = p$. The predictors $\mathbf{X}$ can be either categorical values including their interactions as in ANOVA or continuous values as in nonparametric additive model, or even the mixture of categorical and continuous values as well. If Y is a categorical or discrete response, (1.1) becomes a classification model or logistic regression model with appropriate transformations on Y, which will be addressed elsewhere.

In the statistical research and practice, we often encounter ultrahigh dimensional data sets

in which the number of variables p far exceeds the number of observations n. By ultrahigh dimensionality, we loosely refer to the definition $\log(p) = O(n^a)$ for some positive constant $a \in (0, 1/2)$, which is also called non polynomial (NP) dimensionality, compared to polynomial dimensionality equivalent to high dimensionality defined as $p = O(n^a)$, see Fan and Song (2010). The dimensionality can be ultrahigh because researchers often like to obtain a large number of potential variables in order to retain the possibly important and necessary connections between predictive factors $\mathbf{X}$ and the outcome Y. However, some variables may be useless or redundant, and hence extracting useful and necessary information from these ultrahigh dimensional data is a critical issue and indispensable task in many research fields. Without considering the group structure, this problem is called "large p, small n" and has brought a lot of challenges in the statistical analysis and applications. In most typical cases, nevertheless, only a small number of predictors are supposed to be truly relevant to the response Y, which corresponds to the assumption of sparse structure of the coefficient vector $\boldsymbol{\beta}$. The sparse structure means that only a small portion of the coefficients are nonzero in model (1.1). That is, we have $\beta_{jk} = 0$ for which the predictor has no any connections with Y, which amounts to eliminating the redundant variable $X_{jk}, j = 1, \ldots, J$, $k = 1, \ldots p_j$, from the model (1.1). In the presence of high dimensional data sets where $p > n$, the classical least squares method is not applicable due to the rank deficiency of the matrix $\mathbf{X}^T\mathbf{X}$ and the lack of degrees of freedom in the fitting process. In such cases, dimension reduction or a selection of significant variables with nonzero coefficients is an essential task and central theme of high dimensional data analysis. Dropping the useless variables enables us to improve the accuracy of predictions as well as reduce the variability of models by using some well-developed methods for low dimensional data. Statisticians usually use stepwise deletion and subset selection that are practically useful and intuitively simple in conjunction with the information criteria such as AIC or BIC, but they all suffer from several severe drawbacks such as ignoring stochastic errors, and lack of stability as discussed by Breiman (1995), and

their computation time is huge for large problems. In particular, they are infeasible when the computational cost is another primary concern, especially in the case of high or even ultrahigh dimensional data sets since the number of candidate models grows exponentially as the number of variables increases. Also, their solutions are locally but not globally optimal because they do not consider all variables at the same time. To handle these issues, under the assumption of sparsity, a large number of variable selection approaches have been proposed to select the significant variables automatically and estimate the sparse model simultaneously by imposing a continuous penalty on the coefficients in model (1.1), see, for example, the nonnegative garrote in Breiman (1995); Lasso in Tibshirani (1996); the smoothly clipped absolute deviation (SCAD) in Fan and Li (2001); the least angle regression algorithm (LARS) in Efron, *et al.*, (2003); adaptive Lasso in Zou (2006); minimax concave penalty (MCP) in Zhang (2007), etc. Those methods mentioned above can reduce the dimensionality effectively as well as maintain the sparsity of the model efficiently under the constraint that the dimension p is smaller than or even fairly but not much larger than the sample size n.

In the ultrahigh dimensional problems where $p \gg n$, however, the aforementioned individual variable selection methods are not consistent for model selection under a general condition discussed in Leng *et al.*, (2006) and Zhao and Yu (2006). They also suffer from the computational complexity in these high or ultrahigh dimensional problems, which brings us the simultaneous challenges of statistical accuracy, algorithmic stability and computational expediency as discussed by Fan, et al., (2009). To address these challenges, Fan and Lv (2008) proposed a marginal correlation learning method called sure independence screening (SIS) and iterated sure independence screening (ISIS, a refined version of SIS) to reduce the dimensionality prior to applying the individual variable selection methods in the context of linear model with sparse structure. They used the marginal Pearson correlation as a criterion that measures the strength of relationship between each predictor and response in order to screen out the important predictors, equivalently, remove the unimportant predictors. An

remarkable virtue is that SIS can reduce the dimensionality significantly while preserving the true model with large probability under suitable assumptions. Also, the SIS is computationally simple, practically useful and theoretically appealing. Motivated by these ideas, various screening methods and their extensions have been developed to efficiently rule out the unimportant variables under the different settings of true model and assumptions thereafter, see, for example, Wang (2009, 2012); Zhu, *et al.*, (2011); Fan and Fan (2008); Fan, *et al.*, (2009); Fan, *et al.*, (2011); Wang and Leng (2013). All of these methods were used for the case of individual variables, without considering their group structure. Most of them were established based on marginal correlation learning that studies the relationship between each predictor and response separately. However, marginal correlation learning is unable to tackle the case where predictors are marginally uncorrelated but jointly correlated with response. This means the marginal correlation may be a misleading measurement in some situations. To this end, Wang and Leng (2013) proposed a method called the high-dimensional ordinary least-squares projector (HOLP) for screening individual variables. The main advantage of HOLP is to accommodate the drawbacks of SIS or other marginal correlation learning methods by relaxing the assumption of marginal correlation condition as it is easily violated in the ultrahigh dimensional data sets where predictors are often correlated. Also, the screening criterion of HOLP can be easily calculated and HOLP has the sure screening property, which makes HOLP attractive by its theoretical support and simple computation for the case of ungrouped variables.

Similar challenges and concerns discussed above can arise in the context of high dimensional data set that is composed of grouped variables, where the number of groups J can be also much larger than the sample size n. In the meanwhile, there also exists the sparse structure of the linear model for the case in which the predictors are grouped. Consequently, a selection of significant groups or even the significant members within these groups becomes more and more important for high dimensional data analysis. By significant group we

mean that there are at least one significant variable with nonzero coefficients within the group. In this sense, setting $\boldsymbol{\beta}_j = \mathbf{0}$ is equivalent to removing the entire variables in the j-th group $\mathbf{X}_j, j = 1, \ldots, J$. Compared with the individual variable selection, grouped variable selection has a completely different framework. The individual variable selection methods may perform inefficiently and the subsequent analysis may be inappropriate by ignoring the group information, especially under the strong group sparsity and a group sparse eigenvalue condition, see Huang and Zhang (2010). Therefore, to solve these problems, there are a fair amount of grouped variable selection methods that arise from individual variables selection and yield the sparse solution at the group level, or even at the within-group level. In greater details, Yuan and Lin (2006) proposed the group Lasso, in which the penalty function is comprised of L_2-norm of the coefficients with respect to the grouped variables. This is a natural extension of the Lasso (Tibshirani, 1996) for individual variable selection. Meier, *et al.*, (2008) studied the group Lasso for logistic regression and presented an efficient algorithm. Kim, *et al.*, (2006) extended the group Lasso by using the same L_1 penalty but more general loss functions. Zhao, *et al.*, (2006) considered a generalization of the group Lasso by utilizing composite absolute penalty. Although the Lasso penalty function performs very well in selecting the significant variables, it is not consistent in terms of variable selection and tends to over-shrink large coefficients producing large bias as discussed in Fan and Li (2001). Such inconsistency and over-shrinkage are thus inherited by the group Lasso regardless of the types of the model. To overcome those drawbacks, Wang, *et al.*, (2007) proposed group SCAD and Breheny, *et al.*, (2009) proposed group MCP respectively, both of which possess the oracle property referring to that the probability of selecting the right set of variables with nonzero coefficients converges to one. On the other hand, we are also interested in selecting the variables within the groups in many situations. Zhou and Zhu (2010) proposed a group hierarchical Lasso and Huang, *et al.*, (2009) proposed a group bridge for the grouped variable selection. Both of them along with group MCP are called

bi-level selection, meaning that they are able to select the groups as well as the individual members within a group simultaneously. The bi-level selection is achieved by combining penalties at the group and individual variables levels. However, group bridge suffers from several computational shortcomings that limit its applicability in practice due to the fact that the bridge penalty is not everywhere differentiable. To solve this problem, Breheny (2014) proposed a new method called the group exponential lasso that performs better than group bridge in terms of the computation speed as well as the estimation accuracy. The various group selection methods stated above have desirable performances for "small J, large n" problems under suitable assumptions, but they are not ideal for the large problems in terms of the computation and accuracy. We will provide more technical details of group Lasso, group SCAD and group MCP that are commonly used in practice among others in Chapter 3.

Likewise, similar challenges and difficulties for grouped variable selection can emerge in the case of ultrahigh dimensional data sets consisting of grouped variables. In this work, we consider the ultrahigh dimensional data with group structure, which is referred to the definition $\log(J) = O(n^a)$ for some positive constant $a \in (0, 1/2)$, which certainly indicates $\log(p) = O(n^a)$. That is, we focus on the number of total groups instead of total variables. When the number of groups J grows much faster than the sample size n, the algorithm of grouped variable selection may fail to converge, especially at low value of regularization parameter where the model is nonidentifiable or nearly singular. Even if the algorithm does converge in the setting of "large J, small n", the estimated coefficients may be the ones that are not globally optimal solutions which are not interesting for us at all. For these reasons, we feel that there is a need for new screening methods that can reduce the dimensionality of data significantly before selecting the important groups and variables within these groups. This idea is motivated by the scheme used for the linear model without considering the group structure in which we reduce the number of predictors, whereas we reduce the number of

groups in the linear model with group structure. To our best knowledge, almost none have been working on the grouped variable screening so far. Although Li, *et al.*, (2012) proposed a feature screening via distance correlation learning and pointed out that the distance correlation can be used for the grouped variable screening, they only showed a rather simple simulated example without giving much details. Also, their assumptions and sure screening property were all established for the individual explanatory variable cases. In addition, the distance correlation learning is exactly the same as SIS under the framework of linear regression with normally distributed predictors and random error. In this case, the distance correlation learning still strongly relies on the marginal correlation between individual predictors and response. Lastly, since we have to doubly centralize the corresponding matrices in the calculation of distance correlation, the computational load will be overwhelmingly heavy for large sample and large group sizes as shown in the simulation study in Chapter 4. Nevertheless, the SIS and HOLP are primarily designed to screen out the important individual variables. When directly applied to the variables with group structure, they tend to make selection based on the strength of individual variables rather than that of grouped variables. This often results in selecting less groups than necessary for a given threshold due to using the replicated groups that contains several top correlation values. In this dissertation work, we extend the individual variable screening methods SIS and HOLP to the grouped variable screening by taking one simple step further with the incorporation of the grouped structure. More specially, we take the L_1-norm of the vector of screening criterion values preliminarily estimated by SIS or HOLP with the adjustment of group size for each group and use them as a new screening criterion. This motivation is similar to the case where group Lasso is extended from standard Lasso. We also re-explore the group version of distance correlation learning from Li, *et al.*, (2012) and then compare their performances in practice, where we observe the above weaknesses. To use the possible marginal information of groups, we investigate another screening method by fitting a group-wise regression and

making use of the adjusted R^2 that naturally measures the strength of correlation between each group and response. The adjusted R^2 can be used as a new screening criterion and is somewhat parallel to SIS method in the sense of marginal correlation learning. We introduce a framework and conduct simulations that shed light on the behavior of these methods. As the data set is reduced accurately from ultrahigh to moderate or even low dimension, we can apply the sophisticated grouped variable selection methods to achieve good estimation of the model without the influence of computational complexity. As a result, we not only reduce the computational burden, but also gain the algorithmic stability and accurate estimation by implementing this two-stage procedure.

In this dissertation, we will first introduce four screening methods for the case of grouped variables under the assumption of linear model in Chapter 2, and briefly outline three widely used grouped variable selection methods for the estimation at the second stage in Chapter 3. To examine performance of the proposed methods, we conduct intensive numerical studies in Chapter 4 and apply the proposed methods to analyze a real data set in Chapter 5. An R package `grpss` will be described in Chapter 6. The technical proofs of sure screening property of four screening approcahes will be given in the Appendix.

Chapter 2

Grouped Variable Screening

In this chapter, we describe four grouped variable screening techniques in great details. Without loss of generality, we assume that the random error ε in linear model (1.1) follows a normal distribution $\mathcal{N}(0, \sigma^2)$. We also centralize the response such that we can ignore the intercept and express the linear model as equation (1.1) and standardize all predictors such that each predictor has mean 0 and standard deviation 1, i.e., $E(X_{jk}) = 0, \text{Var}(X_{jk}) = 1, j = 1, \ldots, J, k = 1, \ldots, p_j$. The main goal of grouped variable screening is to screen out the important groups, each of which has at least one nonzero coefficients of predictors in model (1.1). Since the screening procedure aims to seek the important groups, we do not need to estimate the coefficients of predictors within groups accurately but maintain the ranks of importance of groups ideally. Thus, we wish to recover the strength of underlying relationship between each group and response as much as possible. The issue of accurate estimation of coefficients will be taken care of at the second stage of analysis by the methods described in Chapter 3. In the subsequent sections, we will discuss four grouped variable screening methods: group SIS, group HOLP, group-wise adjusted R^2 and group-wise distance correlation (gSIS, gHOLP, gAR2, gDC for short, respectively). The first three methods are constructed under the assumption of linear model in (1.1) but the last one does not require

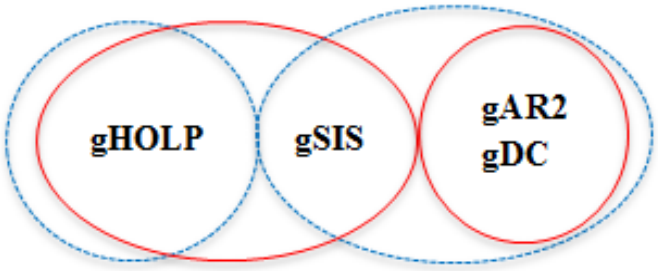


Figure 2.1: The similarities and differences among gSIS, gHOLP, gAR2 and gDC from two perspectives. Red solid circle focuses on the estimation of criteria (i.e., estimate coefficients [left solid circle] vs. correlation [right solid circle]), and the blue dash circle on the used information (i.e., use joint [left dash circle] vs. marginal information [right dash circle]).

this assumption. More specifically, gSIS and gHOLP are extensions from SIS and HOLP to the settings of grouped variables under linear model and will be described in section 2.1 and 2.2, respectively. The gAR2 is a newly proposed screening method and will be introduced in section 2.3. The gDC was originally used in Li, *et al.*, (2012) and will be reinvestigated for the settings of grouped variables in section 2.4. Figure 2.1 shows the similarities and differences among gSIS, gHOLP, gAR2 and gDC. That is, the commonality of the gSIS, gAR2, gDC is that they all strongly rely on the marginal information of correlation between predictors of groups and response, but gHOLP is built on the joint relationship between response and all predictors. In one word, gHOLP is the joint correlation learning while the others are the marginal correlation learning. This is visualized in Figure 2.1 by dashed blue circles in terms of the used information. From the point of view on the criterion marked by solid red circles in Figure 2.1, gSIS and gHOLP focus on the magnitude of estimated regression coefficients, while the gAR2 and gDC emphasize the strength of correlation.

The gSIS and gHOLP will outperform the individual variable screening methods SIS and HOLP due to the fact that they consider all predictors in each group simultaneously. When variables are grouped, ignoring the group structure and directly applying SIS in (2.1) or

HOLP in (2.3) may be suboptimal or even yield an insensible final model. Additionally, the gSIS and gHOLP do not add too much computational complexity contrast to SIS and HOLP. The gAR2 and gDC are very similar under the linear model settings, but gDC can also handle the situation where the relationship between predictor and response is nonlinear. However, the gDC is not recommended to use for the large problems because of the complicate or even exceedingly computation of distance correlation.

A dimensionality reduction method is desirable if it possesses the sure screening property that all the important groups survive with probability tending to 1 after applying grouped variable screening as the sample size becomes large enough. This means the screening procedure can retain all important grouped variables with overwhelming probability. The sure screening property is very important and indispensable to all screening methods because otherwise it is nonsense to filter out the useful information that can be used for later analysis. We let $\mathcal{M}_S^g = \{\mathbf{X}_j : \boldsymbol{\beta}_j \neq \mathbf{0}, 1 \leq j \leq J\}$ or $\mathcal{S} = \{1 \leq j \leq J : \boldsymbol{\beta}_j \neq \mathbf{0}\}$ be the set of true sparse model with non-sparsity group size $s = |\mathcal{S}| = |\mathcal{M}_S^g| \ll J$ and the selected submodel by a grouped variable screening be $\mathcal{M}_\mathcal{D}^g$ with non-sparsity group size $d = |\mathcal{M}_\mathcal{D}^g| < J$, in which d is a threshold and can be manually chosen as an integer from $[1, J]$ depending on the problem of interest, and satisfies $s < d$, or automatically selected by random permutation method discussed in section 2.5. Without ambiguity, we use $\mathcal{M}_\mathcal{D}^g$ as a notation of submodel with group size d for grouped variable case and $\mathcal{M}_\mathcal{D}$ for individual variable case in the subsequent sections. By sure screening property, we have the probability $\Pr(\mathcal{M}_S^g \subseteq \mathcal{M}_\mathcal{D}^g)$ going to 1 as $n \to \infty$ for some given threshold $d \geq s$. A sure screening property of gSIS, gHOLP, gAR2 and gDC will be established in Theorem 2.1, 2.2, 2.3, 2.4, respectively. It is obvious that larger d implies larger probability of including the true model $\mathcal{M}_S^g$ in the final model $\mathcal{M}_\mathcal{D}^g$ but may bring more complex computations at the second stage of analysis. In some sense, the threshold controls the tradeoff between the complexity of computations and the accuracy of estimation. Without giving too much computational complexity, we set the threshold to

be the sample size in the numerical simulation for the purpose of comparison. The issue of choosing data-driven threshold d will be addressed in section 2.5. In particular, we can adopt the random permutation idea in Zhao and Li (2010) to determine a data-driven threshold. It is worthwhile noting that in the feature screening process, there is a situation where some groups are known to be truly related to the response from experts' knowledge or experience in advance. Therefore, we want to preserve them in the final model and do not wish to include them in the variable screening as well as selection procedure. In other words, we let the predictors from these groups participate in the estimation procedure directly.

2.1 Group Sure Independence Screening (gSIS)

Fan and Lv (2008) introduced a new framework of variable screening via marginal correlation learning and suggested to fit a component-wise regression between predictors and response using ordinary least squares to obtain the SIS screening criterion $\widehat{\boldsymbol{\omega}} = (\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_J)$ defined as

$$\widehat{\boldsymbol{\omega}} = \mathbf{X}^T Y, \tag{2.1}$$

where $\boldsymbol{\omega}_j = (\omega_{j1}, \cdots, \omega_{jp_j}), j = 1, \ldots, J$, is the marginal coefficient vector for the j-th group. Each element of $\widehat{\boldsymbol{\omega}}$ is also equivalent to the marginal pearson correlation between each predictor and response, rescaled by the standard deviation of the response. To see this, recalling the definition of correlation $\text{corr}(X_{jk}, Y) = \text{cov}(X_{jk}, Y)/(s_{X_{jk}} \times s_Y)$, one has

$$\omega_{jk} \;=\; \sum_{i=1}^{n} x_{ijk} y_i = \frac{n^{-1}\sum_{i=1}^{n} x_{ijk} y_i}{n^{-1}\sum_{i=1}^{n} x_{ijk}^2} = \frac{\text{cov}(X_{jk}, Y) s_Y}{s_{X_{jk}} s_Y} = \text{corr}(X_{jk}, Y) s_Y,$$

following by $\text{corr}(X_{jk}, Y) \propto \omega_{jk}$, where x_{ijk} is the i-th observation of the k-th variable in the j-th group, $\bar{X}_{jk} = n^{-1}\sum_{i=1}^{n} x_{ijk} = 0, \bar{Y} = n^{-1}\sum_{i=1}^{n} y_i = 0$ and $s_{X_{jk}}^2 = n^{-1}\sum_{i=1}^{n} x_{ijk}^2 = 1$ by the assumptions of the predictors and response, $s_Y^2 = n^{-1}\sum_{i=1}^{n} y_i^2$. We then sort the

magnitude of $\widehat{\boldsymbol{\omega}}$ in a decreasing order and select the submodel $\mathcal{M}_{\mathcal{D},SIS}$ using the following rule.

$$\mathcal{M}_{\mathcal{D},SIS} = \{X_{jk} : |\omega_{jk}| \text{ are among the largest } d \text{ of all } \omega_{jk}\text{'s}\},$$

which is a straightforward way to shrink the full model $\mathcal{M} := \{X_{jk} : j = 1, \ldots, J; k = 1, \ldots, p_j\}$ with ultrahigh dimensionality p down to a submodel $\mathcal{M}_{\mathcal{D},SIS}$ with moderate or small dimensionality d that is much smaller than p, i.e., $d \ll p$.

To incorporate the group information, we only need to move a step forward based on the SIS by taking the L_1-norm of the $\widehat{\boldsymbol{\omega}}$ defined in (2.1) for each $\boldsymbol{\omega}_j$ of group coefficient vector, $j = 1, 2, \ldots, J$. Specially, we define a new criterion $\widehat{\boldsymbol{\omega}}^g = (\omega_1^g, \cdots, \omega_J^g)$ for grouped variable screening, where

$$\omega_j^g = p_j^{-1} \|\boldsymbol{\omega}_j\|_1 := p_j^{-1} \left(|\omega_{j1}| + \cdots + |\omega_{jp_j}|\right), \tag{2.2}$$

in which $\|\cdot\|_1$ is the L_1-norm of a vector. Certainly, we can use L_2 or L_∞-norm on the vector $\boldsymbol{\omega}_j$ but L_1-norm has theoretically attractive advantages as shown in the proofs. Note that p_j is the number of variables in the j-th group and used to compensate for the size of each group. Thus, the submodel can be chosen as

$$\mathcal{M}_{\mathcal{D},SIS}^g = \{\mathbf{X}_j : \omega_j^g \text{ are among the largest } d \text{ of all } \omega_j^g\text{'s}\},\ d \leq J.$$

We call this new screening method group SIS (gSIS) which represents the sure independence screening modified for grouped variables.

The following Theorem 2.1 states the sure screening property of gSIS that we can retain the truly important groups with overwhelming probability under some conditions.

Theorem 2.1 (**Sure screening property of gSIS**). *Assume that (A1)-(A4) that are stated in Appendix A hold. If we choose γ_n such that*

$$\frac{\gamma_n}{n^{1-\kappa}} \to 0, \text{and } \frac{\gamma_n \sqrt{\log(n)}}{n^{1-\kappa}} \to \infty,$$

then for some constant C, we have

$$\Pr\left(\max_{j\notin S}\omega_j^g < \gamma_n < \min_{j\in S}\omega_j^g\right) = 1 - O\left\{\exp\left(-C\frac{n^{1-2\kappa-2\tau-\nu-\gamma}}{2\log(n)}\right)\right\}.$$

This means if we choose a submodel $\mathcal{M}_{\mathcal{D},SIS}^g$ with $d > s$, we have

$$P\left(\mathcal{M}_S^g \subseteq \mathcal{M}_{\mathcal{D},SIS}^g\right) = 1 - O\left\{\exp\left(-C\frac{n^{1-2\kappa-2\tau-\nu-\gamma}}{2\log(n)}\right)\right\}.$$

Theorem 2.1 implies that there is a clear separation of the strength of ω_j^g between the important groups $j \in S$ and unimportant groups $j \notin S$. These two sets of groups can be easily identified when the separation is sufficiently large. A more useful of implication is that a model selection consistency result can be derived under the assumptions (A1) - (A4). This theorem also indicates the dimensionality of groups can be exponentially high. A similar and general conclusion can be obtained from the following sure screening properties of gHOLP, gAR2 and gDC and will be omitted thereby.

2.2 Group High-dimensional Ordinary Least-squares Projection (gHOLP)

Although gSIS is computationally simple, it adopted the marginal correlation learning meaning that the screening result will be affected by the predictors that are marginally uncorrelated but jointly correlated with the response. In theoretical proofs of Fan and Lv (2008), a condition was imposed to rule out the aforementioned situation. Another disadvantage of SIS is that the unimportant predictors with high correlation with important predictors have higher probability to be selected than the other important predictors with relatively weak correlation with the response. We should expect that the gSIS would perform poorly

or fail completely when these situations occur as shown in simulation studies. To this end, Wang and Leng (2015) proposed a new screening criterion called High-dimensional Ordinary Least-squares Projector (HOLP) that can well address the case where the predictors are marginally uncorrelated but jointly correlated with response.

The HOLP in Wang and Leng (2015) was primarily proposed to screen out the important individual variables before the second stage of refined analysis such as the regression analysis. They showed that the performance of HOLP was at least competitive or even superior to SIS due to taking the joint correlation structure of all predictors into account, rather than the marginal correlation information. The submodel $\mathcal{M}_{\mathcal{D},HOLP}$ is chosen according to the magnitude of estimated coefficients defined as

$$\widehat{\boldsymbol{\beta}} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}Y, \tag{2.3}$$

where $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1, \cdots, \widehat{\boldsymbol{\beta}}_J)$, in which $\widehat{\boldsymbol{\beta}}_j = (\widehat{\beta}_{j1}, \cdots, \widehat{\beta}_{jp_j}), j = 1, \ldots, J$. In fact, the estimated coefficients (2.3) can be written as $\widehat{\boldsymbol{\beta}} = \mathbf{X}^+Y$, where $\mathbf{X}^+ \equiv \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}$ and is termed as Penrose-Moore inverse. This leads to the estimator $\widehat{\boldsymbol{\beta}}$ being the same calculation form as $\widehat{\boldsymbol{\omega}}$ defined in (2.1). We then use the same strategy as SIS, so the submodel $\mathcal{M}_{\mathcal{D},HOLP}$ can be chosen as

$$\mathcal{M}_{\mathcal{D},HOLP} = \{X_{jk} : |\widehat{\beta}_{jk}| \text{ are among the largest } d \text{ of all } \widehat{\beta}_{jk}\text{'s}\},\ d \leq p,$$

in which $j = 1, \ldots, J, k = 1, \ldots, p_j$. Similar to the procedure of gSIS from SIS, the grouped version of HOLP can proceed in the same fashion from HOLP. Denote a vector $\widehat{\boldsymbol{\beta}}^g = (\widehat{\beta}_1^g, \cdots, \widehat{\beta}_J^g)$ with J elements representing the new criterion values, which can be computed by

$$\widehat{\beta}_j^g = p_j^{-1}\|\widehat{\boldsymbol{\beta}}_j\|_1 := p_j^{-1}\left(|\widehat{\beta}_{j1}| + \cdots + |\widehat{\beta}_{jp_j}|\right), \tag{2.4}$$

and then the submodel can be chosen as

$$\mathcal{M}^g_{\mathcal{D},HOLP} = \{\mathbf{X}_j : \widehat{\beta}^g_j \text{ are among the largest } d \text{ of all } \widehat{\beta}^g_j\text{'s}\},\ d \leq J,$$

which is named as group HOLP (gHOLP), the HOLP for grouped variables.

Comparing equations (2.1) and (2.3), we can observe that the latter has one more term $(\mathbf{X}\mathbf{X}^T)^{-1}$ which accounts for the joint information among predictors. If the predictors are orthonormal, i.e., $\mathbf{X}\mathbf{X}^T = I_n$, the HOLP (resp., gHOLP) coincides with the SIS (resp., gSIS). Therefore, SIS (resp., gSIS) agrees with HOLP (resp., gHOLP) when, for instance, predictors are mutually independent or uncorrelated. On the other hand, HOLP (resp., gHOLP) is invariant to the scale of the signal defined by $\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\boldsymbol{\beta}$, while SIS (resp., gSIS) is not scale-invariant to the signal defined by $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$. The implication of this property is that the gSIS may be severely affected by the way of scaling variables but the gHOLP is robust to it. Therefore, it is recommended to scale the predictors before using gSIS. Another significant insight can be seen that $\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}$ is more diagonally dominating than $\mathbf{X}^T\mathbf{X}$ for some cases. Accordingly, gHOLP is more likely to maintain the true order of important groups than gSIS.

Similar to Theorem 2.1, we can also establish the sure screening property of gHOLP present in Theorem 2.2 as follows.

Theorem 2.2 (**Sure screening property of gHOLP**). *Assume that (B1)-(B4) that are stated in Appendix B hold. If we choose γ_n such that*

$$\frac{p\gamma_n}{n^{1-\tau'-\kappa}} \to 0, \frac{p\gamma_n\sqrt{\log(n)}}{n^{1-\tau'-\kappa}} \to \infty,$$

then for some C_1 specified in Assumption (B1), we have

$$\Pr\left(\min_{j\in S}\hat{\beta}^g_j > \gamma_n > \max_{j\notin S}\hat{\beta}^g_j\right) = 1 - O\left\{\exp\left(-C_1\frac{n^{1-2\kappa-5\tau'-\nu-\gamma}}{2\log(n)}\right)\right\}.$$

This means if we choose a submodel $\mathcal{M}^g_{\mathcal{D},HOLP}$ with $d > s$ we have

$$P\left(\mathcal{M}^g_S \subseteq \mathcal{M}^g_{\mathcal{D},HOLP}\right) = 1 - O\left\{\exp\left(-C_1 \frac{n^{1-2\kappa-5\tau'-\nu-\gamma}}{2\log(n)}\right)\right\}.$$

2.3 Groupwise adjusted R-squared (gAR2)

We now introduce a new screening criterion which is complementary to gSIS but is more sophisticated and can better take advantage of the information between groups and response. That is, we propose another screening criterion called group-wise adjusted R^2 (gAR2) which utilizes the marginal information like gSIS but also naturally incorporates the whole contribution of group information that L_1-norm of $\boldsymbol{\omega}_j$ may lose in gSIS. In greater details, we fit a multiple linear model between the response $\mathbf{y}$ and the predictors $\mathbf{X}_j = (X_{j1}, \cdots, X_{jp_j})$ of the j-th group for each $j = 1, \ldots, J$ separately. The adjusted R^2 can be interpreted as a multiple correlation that measures the relationship between the response and multiple predictors in each group. For notation at ease, the adjusted R^2 is denoted to be $\bar{R}^2_j$ calibrating the relationship between the j-th group $\mathbf{X}_j$ and response Y, $j = 1, \ldots, J$, and can be calculated by

$$\bar{R}^2_j = \frac{n-1}{n-p_j-1} R^2_j - \frac{p_j}{n-p_j-1}, \tag{2.5}$$

where R^2_j is the multiple correlation and defined as $R^2_j = 1 - \sum_{i=1}^n (\hat{y}_{i,j} - y_i)^2 / \sum_{i=1}^n (y_i - \bar{y})^2$, in which $\hat{y}_{i,j}$ is the fitted value for the i-th observation in the j-th group by the group-wise linear regression, $i = 1, \ldots, n; j = 1, \ldots, J$. The new criterion is defined as $\mathbf{R} = (\bar{R}^2_1, \cdots, \bar{R}^2_J)$ and the submodel can be selected as

$$\mathcal{M}^g_{\mathcal{D},AR2} = \{\mathbf{X}_j : \bar{R}^2_j \text{ are among the largest } d \text{ of all } \bar{R}^2_j\text{'s}\},\ d \leq J,$$

and this screening procedure is referred to gAR2. It is worthwhile pointing out that the group size p_j should be smaller than the sample size n for all $j = 1, \ldots, J$ so that the $\bar{R}^2_j$ is

meaningful for multiple linear regression. In effect, there is another simple way to calculate the multiple correlation R_j^2 instead of fitting a multiple linear regression. The multiple correlation R_j^2 can be computed by

$$R_j^2 = \widehat{\boldsymbol{\gamma}}_j^T \widehat{\boldsymbol{R}}_{\mathbf{X}_j}^{-1} \widehat{\boldsymbol{\gamma}}_j,$$

where $\widehat{\boldsymbol{\gamma}}_j^T = (\widehat{\gamma}_{j1}, \cdots, \widehat{\gamma}_{jp_j}), \widehat{\gamma}_{jk} = \widehat{\text{cor}}(X_{jk}, Y)$, and $\widehat{\boldsymbol{R}}_{\mathbf{X}_j}$ is the estimator of inner-correlation matrix of $\mathbf{X}_j$. By the assumptions of predictors and response, one can easily obtain that

$$R_j^2 = s_{\mathbf{y}}^{-2} \widehat{\boldsymbol{\rho}}_j^T \widehat{\boldsymbol{S}}_{\mathbf{X}_j}^{-1} \widehat{\boldsymbol{\rho}}_j \propto \widehat{\boldsymbol{\rho}}_j^T \widehat{\boldsymbol{S}}_{\mathbf{X}_j}^{-1} \widehat{\boldsymbol{\rho}}_j,$$

where now $\widehat{\boldsymbol{\rho}}_j^T = (\widehat{\rho}_{j1}, \cdots, \widehat{\rho}_{jp_j}), \widehat{\rho}_{jk} = \widehat{\text{cov}}(X_{jk}, Y)$, and $\widehat{\boldsymbol{S}}_{\mathbf{X}_j}$ is the estimator of inner-covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}_j}$ of $\mathbf{X}_j$. Under the constrain $p_j \leq n$ and $\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}_j})$ is bounded away from infinity, $\widehat{\boldsymbol{S}}_{\mathbf{X}_j}$ is always positive definite and thus invertible.

Theorem 2.3 states the sure screening property of gAR2. Although we impose an assumption of maximum eigen values of covariance as in Assumption (A4) to rule out the situation of strong collinearity, adjusted R^2 can still accurately quantifies the linear relationship between response and predictors in practice because these global statistics do not depend on individual parameters and their standard errors. In this sense, the proposed gAR2 should be comparable with or even better than gSIS due to making better use of the entire group information and also being able to overcome the effect of strong collinearity within groups. Furthermore, as $\bar{R}_j^2$ is not affected by the scale of the data, the gAR2 is also invariant to the scale of the predictors and the results will be robust to the way of scaling.

Theorem 2.3 (**Sure screening property of gAR2**). *Assume that (A1)-(A4) that are*

stated in Appendix A hold. If we choose γ_n *such that*

$$\frac{\gamma_n}{n^{2-2\kappa-\tau+\gamma''}} \to 0 \text{ and } \frac{\gamma_n \log(n)}{n^{2-2\kappa-\tau+\gamma''}} \to \infty.$$

then for some constant C*, we have*

$$\Pr\left\{\max_{j\notin S} R_j^2 < \gamma_n < \min_{j\in S} R_j^2\right\} = 1 - O\left\{\exp\left(-C\frac{n^{1-2\kappa-3\tau-\nu-\gamma''}}{\log(n)}\right)\right\},$$

This means if we choose a submodel $\mathcal{M}^g_{\mathcal{D},AR2}$ *with* $d > s$*, we have*

$$P\left(\mathcal{M}^g_S \subseteq \mathcal{M}^g_{\mathcal{D},AR2}\right) = 1 - O\left\{\exp\left(-C\frac{n^{1-2\kappa-3\tau-\nu-\gamma''}}{\log(n)}\right)\right\}.$$

2.4 Groupwise distance correlation (gDC)

The screening criteria mentioned above are all used to measure the magnitude of linear relationship between groups and response. A possible problem comes to the situation where the relationship is nonlinear. Fortunately, Sźekely, *et al.*, (2007) proposed distance correlation (DC) to measure the dependence between two random vectors. The important property of DC is that the DC of two random vectors is zero if and only if these two random vectors are independent. The advantage of DC is that it is not only capable of capturing the linear relationship as the pearson correlation or multiple correlation does, but also the nonlinear dependence between two random vectors. Based on these aspects, Li, *et al.*, (2012) advocated using the DC for measuring the strength of correlation between response and predictors without assuming the linear structure defined in (1.1). To be precise, the DC between the j-th grouped predictor $\mathbf{X}_j$ and response Y is defined as follows. We first compute the pairwise distances

$$a_{i,h} = \|\mathbf{X}_{ij} - \mathbf{X}_{hj}\|_2, b_{i,h} = |y_i - y_h|^2,$$

where $\|\cdot\|_2$ denotes the L_2-norm of a vector and $\mathbf{X}_{ij}, \mathbf{X}_{hj}$ are the i-th and k-th observations of $\mathbf{X}_j$, $i, h = 1, \ldots, n, j = 1, \ldots, J$. Let $A = \{a_{i,h}\}_{i,h=1}^n$ and $B = \{b_{i,h}\}_{i,h=1}^n$, both of which are $n \times n$ matrices. Note that A and B are both symmetric matrices and are called the distance matrices. Secondly, we doubly center all columns and rows of matrices A and B. That is, let

$$A_{i,h} = a_{i,h} - \bar{a}_{i.} - \bar{a}_{.h} + \bar{a}_{..}, B_{i,h} = b_{i,h} - \bar{b}_{i.} - \bar{b}_{.h} + \bar{b}_{..},$$

where $\bar{a}_{i.}, \bar{b}_{i.}$ are the i-th row means, $\bar{a}_{h.}, \bar{b}_{h.}$ are the k-th column means, and $\bar{a}_{..}, \bar{b}_{..}$ are the grand means of the distance matrices A and B. The matrices A and B are then updated by $A = \{A_{i,h}\}_{i,h=1}^n$ and $B = \{B_{i,h}\}_{i,h=1}^n$. The distance correlation between $\mathbf{X}_j$ and $\mathbf{y}$ is defined as

$$\mathrm{dCor}(\mathbf{X}_j, Y) = \frac{\mathrm{dCov}(\mathbf{X}_j, Y)}{\sqrt{\mathrm{dVar}(\mathbf{X}_j)\,\mathrm{dVar}(Y)}}, \tag{2.6}$$

where $\mathrm{dCov}(\mathbf{X}_j, Y)$ is the distance covariance of $\mathbf{X}_j$ and Y, simply defined as the arithmetic average of the products $A_{i,h}$ and $B_{i,h}$, i.e., $\mathrm{dCov}^2(\mathbf{X}_j, Y) = n^{-2}\sum_{i,h=1}^n A_{i,h}B_{i,h}$. Similar definitions can be applied for the distance variances of $\mathrm{dVar}(\mathbf{X}_j)$ and $\mathrm{dVar}(Y)$, respectively. Namely, $\mathrm{dVar}^2(\mathbf{X}_j) = n^{-2}\sum_{i,h=1}^n A_{i,h}^2, \mathrm{dVar}^2(Y) = n^{-2}\sum_{i,h=1}^n B_{i,h}^2$. For simplicity, we let $\mathcal{D}_j = \mathrm{dCor}^2(\mathbf{X}_j, Y), j = 1, \ldots, J$. Note that $\mathcal{D}_j \in [0, 1]$. Thus, the screening criterion by DC is defined as $\mathbf{D} = (\mathcal{D}_1, \cdots, \mathcal{D}_J)$ and the submodel can be chosen as

$$\mathcal{M}_{\mathcal{D},DC}^g = \{\mathbf{X}_j : \mathcal{D}_j \text{ are among the largest } d \text{ of all } \mathcal{D}_j\text{'s}\},\ d \leq J.$$

As pointed out by Sźekely, et al., (2007), the DC of two univariate normal random variables is a strictly increasing function of their absolute value of Pearson correlation. In this case, the DC agrees with the SIS for individual variable screening under the assumption of linear model with normal covariates. This is also the special case of grouped variable screening by using groupwise distance correlation when every group has only one variable of normal distribution. Since we calculate the distance correlation for each group separately, the assumption of strongly marginal correlation is still needed and shown in the assumption (D2) of Appendix.

On the other hand, another major downside of distance screening criterion is that since we apply a three-step procedure, it may require intensive computation to obtain the distance correlation, especially in the case where the sample size and number of groups are large. Thus, it would make the distance correlation screening impractical for large problems. Simulation results in Chapter 4 show that the computation load increases rather quickly as the sample size and number of groups increase.

Theorem 2.4 (**Sure screening property for gDC**). *Under assumptions (D1) and (D2) stated in the Appendix D, if the size of selected submodel $d \geq s$, we have*

$$\Pr(\mathcal{M}_S^g \subseteq \mathcal{M}_{\mathcal{D},DC}^g) = 1 - O(s[\exp(-c_5' n^{1-2(\kappa'+\gamma')}) + n\exp(-c_6' n^{\gamma'})])$$

for the screening criteria of groupwise distance correlation, where c_5' and c_6' are positive constants.

Note that if we choose $\gamma' = (1-2\kappa')/3$, Theorem 2.4 becomes

$$\begin{aligned}\Pr\left(\mathcal{M}_S^g \subseteq \mathcal{M}_{\mathcal{D},DC}^g\right) &= 1 - O\left(s(n+1)\exp\left(-c_5' n^{(1-2\kappa')/3}\right)\right) \\ &= 1 - O\left(s\exp\left\{-c_5' n^{(1-2\kappa')/3} + \log(n+1)\right\}\right).\end{aligned}$$

2.5 Data-driven threshold

To achieve the sure screening property, Fan and Lv (2008) suggested taking the threshold $d = n-1$ or $d = \lfloor n/\log(n) \rfloor$ which is below the sample size n, where $\lfloor a \rfloor$ denotes the integer part of a number a. However, the choices are subjective and usually conservative in the sense that many unimportant variables would be included in the screened submodel. Although larger threshold d implies larger probability of containing all important variables, more unimportant variables will also be contained. This leads to a larger false positive rate, the

proportion of unimportant variables incorrectly included in submodel $\mathcal{M}_d^g$ and increase the computation load in the further analysis. To solve this issue, Zhao and Li (2010) proposed a practical method to choose a threshold while achieving the sure screening property as well as controlling the false positive rate and computation complexity. Fan, *et al.* (2011) implemented this idea to determine a data-driven threshold in nonparametric independence screening for sparse ultra-high dimensional additive models and called it random permutation approach.

The idea of random permutation works as follows. Let $\mathbf{X}_{i\cdot} = (\mathbf{X}_{i1}, ..., \mathbf{X}_{iJ})$ be the i-th observation of $\mathbf{X}, i = 1, ..., n$. Given data of the form $\{(\mathbf{X}_{i\cdot}, y_i), i = 1, ..., n\}$, we decouple $\mathbf{X}_{i\cdot}$ and y_i by a random permutation of index $1, ..., n$ such that we obtain new data of the form $\{(\mathbf{X}_{\pi(i)\cdot}, y_i), i = 1, ..., n\}$, where $\pi(1), ..., \pi(n)$ are random permutation of the index $1, ..., n$. The resulting data $\{(\mathbf{X}_{\pi(i)\cdot}, y_i), i = 1, ..., n\}$ then follows a null model in which the predictors have no correlation with response. As we know that under a null model, a set of genetic variants has absolutely no effect on the outcome. For the newly permuted data, we recompute the values of grouped screening criterion. For simplicity, we outline the algorithm of random permutation for gSIS as follows. The other grouped screening criteria can proceed in the same way.

(i). For every $j \in (1, ..., J)$ and original data $(\mathbf{X}, Y)$ we find the values of grouped screening criterion $\widehat{\boldsymbol{\omega}}^g = (\omega_1^g, \cdots, \omega_J^g)$ calculated by equations (2.1) and (2.2). Randomly permute the rows of $\mathbf{X}$, yielding $\mathbf{X}^* = (\mathbf{X}_1^*, ..., \mathbf{X}_J^*)$ and let $\alpha_{(q)}$ be the q-th quantile of $\widehat{\boldsymbol{\omega}}^{g*} = (\widehat{\omega}_1^{g*}, ..., \widehat{\omega}_J^{g*})$, where

$$\widehat{\omega}_j^{g*} = p_j^{-1} \|\boldsymbol{\omega}_j^*\|_1$$

and

$$\boldsymbol{\omega}_j^* = \mathbf{X}_j^{*T} \mathbf{y}.$$

We select the groups in the submodel

$$\mathcal{M}_{\mathcal{D}}^{g} = \{\mathbf{X}_j : \omega_j^g \geq \alpha_{(q)}, j = 1, ..., J\},$$

where $d = |\mathcal{D}|$ is not a pre-specified constant, but is determined by $\alpha_{(q)}$ such that $d = \#\{1 \leq j \leq J : \omega_j^g \geq \alpha_{(q)}\}$.

(ii). Apply penalized likelihood estimation described in Chapter 3 on the screened submodel $\mathcal{M}_{\mathcal{D}}^{g}$ to obtain a final model that contains the selected grouped variables with nonzero coefficients.

The quantile q in step (i) controls the number of unimportant groups that enter the submodel $\mathcal{M}_{\mathcal{D}}^{g}$. As suggested by Fan, *et al.* (2011), we take somewhat aggressive $q = 1$ which means to taking the maximum value of the empirical criterion of the permuted estimates. This choice also implies we do not allow any unimportant groups to enter the submodel $\mathcal{M}_{\mathcal{D}}^{g}$.
To see the rationale behind the random permutation idea, we let $\bar{\mathcal{M}}_{S,SIS}^{g}$ be the complement of the true model $\mathcal{M}_{S,SIS}^{g}$. Note that the false positive rate can be written as $\left|\bar{\mathcal{M}}_{S,SIS}^{g} \cap \mathcal{M}_{d,SIS}^{g}\right| / \left|\bar{\mathcal{M}}_{S,SIS}^{g}\right|$. Then the expected false positive rate can be expressed as

$$E\left(\frac{\left|\bar{\mathcal{M}}_{S,SIS}^{g} \cap \mathcal{M}_{d,SIS}^{g}\right|}{\left|\bar{\mathcal{M}}_{S,SIS}^{g}\right|}\right) = \frac{1}{J-s} \sum_{j \in \bar{\mathcal{M}}_{S,SIS}^{g}} \Pr(\omega_j^g \geq \alpha_n),$$

where α_n is a pre-specified constant. Suppose the values of screening criterion ω_j^g has the cumulative distribution function F, we can see that the α_n controls the expected false positive rate at $2\{1 - F(\alpha_n)\}$. Thus, if we choose $\alpha_n = \alpha_{(q)}$ with $q = 1$, the expected false positive rate would be zero as $F(\alpha_{(q)}) = 1$. More importantly, using such strategy can also maintain the sure screening property, see Theorem 5 in Zhao and Li (2010). Similar explanation can be given to the other screening criteria.

Chapter 3

Grouped Variable Selection

As the grouped variable screening is able to accurately reduce the original dimensionality J of dataset to a relatively moderate or small dimension d lower than the sample size, the grouped variable selection method can be then applied to estimate the sparse model without the inherently computational complexity any more. Furthermore, the grouped variable screening can be used iteratively until the reduced dimension d is desirable if necessary. This strategy has been used in Fan and Lv (2008). The scheme of screening and selection is shown in Figure 3.1 and is typically a two-stage procedure. The first stage is to reduce the dimensionality from J to d by grouped variable screening, and the second stage is to perform the grouped variable selection that further reduces dimensionality from d to s. After the first stage, the

Figure 3.1: The schematic diagram of feature screening. The J, d, s are dimensions satisfying $J \gg d > s$.

original problem of estimating large J $\boldsymbol{\beta}_j$'s in model (1.1) simplifies to estimating smaller d

$\boldsymbol{\beta}_j$'s, leading to the reduced linear model

$$Y = \sum_{j \in \mathcal{D}} \mathbf{X}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}, \tag{3.1}$$

where $\mathcal{D}$ is the indices set of the obtained submodel $\mathcal{M}_{\mathcal{D}}^{g}$ with size d. It is obvious that the process of grouped variable selection can be speeded up dramatically as the dimensionality has been reduced significantly while significant variables are all retained, especially when the original dimension J is reasonably high or even ultrahigh. At the second stage, the problem of interest generally involves estimating a series of vectors of coefficients $\boldsymbol{\beta}_j, j \in \mathcal{D}$ by minimizing an objective function that consists of a loss function and a penalty function. In the following we briefly review several commonly used grouped variable selection techniques that can estimate the sparse model in (3.1). These approaches include the group Lasso in Yuan and Lin (2006), the group smoothly clipped absolute deviation (group SCAD) in Wang, *et al.*, (2007) and the group minimax concave penalty (group MCP) of Breheny and Huang (2009), among others. Note that the parameters estimated by group Lasso do not achieve the consistency resulting from using the convex penalized loss function while group SCAD and group MCP gain the unbiased estimators by using the concave penalized loss function. Furthermore, group MCP is able to select the targets at both the group and within-group individual levels. For this reason, group MCP is always called bi-level selection method. Whether we want to select the within-group individual variables depends on the situation of interest. For example, as the individual variables are represented by a set of basis functions that are artificially constructed in nonparametric additive model, selecting the important members in a group is not necessary. On the contrary, it is important to select both the important genes and within-group individual SNPs in gene expression study.

To obtain the estimates of the coefficients, we implement the group descent algorithms in Breheny and Huang (2011) that is much faster and more stable than the local linear

or quadratic approximation, even for the large datasets, and have been implemented in R package `grpreg`. Whereas the group descent algorithms can estimate the sparse model rapidly for high dimensional data sets, the simulation shows the feature screening prior to selection improves the computational expediency drastically, and is not an redundant procedure at all.

3.1 Group Lasso

For a column vector $\mathbf{u} \in R^k, k \geq 1$, and a positive definite matrix C, we define $\|\mathbf{u}\|_C = (\mathbf{u}'C\mathbf{u})^{1/2}$. The solution of parameters using regularization method is generally achieved by minimizing the penalized loss function, which follows the solution $\widehat{\boldsymbol{\beta}}_j(\lambda)$'s of group Lasso in Yuan and Lin (2006) are denoted to be a minimizer of an objective function $Q(\boldsymbol{\beta})$ defined as

$$Q(\boldsymbol{\beta}) = \frac{1}{2n}\left\|Y - \sum_{j\in\mathcal{D}} \mathbf{X}_j\boldsymbol{\beta}_j\right\|_2^2 + \lambda\sum_{j\in\mathcal{D}}\sqrt{p_j}\|\boldsymbol{\beta}_j\|_{C_j}, \tag{3.2}$$

where λ is the regularization parameter and C_j's are $p_j \times p_j$ positive definite matrices, and $\sqrt{p_j}$ attempts to adjust for the group size. An critical issue in (3.2) is the choice of the positive definite matrices C_j's, $j \in \mathcal{D}$. Yuan and Lin (2006) originally suggested using $C_j = I_{p_j}$ for orthonormal $\mathbf{X}_j$ with $\mathbf{X}_j'\mathbf{X}_j/n = I_{p_j}, j \in \mathcal{D}$. However, this is not always the case since the scales of the predictors may not be the same. Huang, et al., (2012) recommended taking $C_j = \mathbf{X}_j'\mathbf{X}_j/n$ regardless of the scales of predictors. This amounts to standardizing the predictors at the group level that we assumed at the beginning of Section 2.1. The standardization also ensures that the penalty is invariant to the scale. Thus, group Lasso imposed the Lasso penalty on the L_2-norm of coefficients of each group, leading to the sparsity and variable selection at the group level. Due to the nature of L_1 penalty on the group norm of coefficients, group Lasso tends to over-shrink the large coefficients, yielding more important groups than necessary in order to compensate this over-shrinkage.

This also leads to relatively high false positive rates. Thus, the group Lasso tends to select a larger model than the true one, causing the caution to use group Lasso. Some popular tuning methods can be applied to choose the optimal λ, such as AIC, BIC, and generalized cross-validation.

3.2 Group SCAD

Fan and Li (2001) defined a more general penalty function called smoothly clipped absolute deviation (SCAD) that was singular at the origin resulting in the sparse coefficient estimators. It also produces continuous estimators and possessed the oracle property under certain reasonable conditions and a proper choice of the regularization parameter. Wang, *et al.*, (2007) extended SCAD to fit a linear regression for microarray time course gene expression data. More specially, they proposed the group SCAD whose solutions to $\boldsymbol{\beta}_j$'s are obtained by minimizing the penalized loss function $Q(\boldsymbol{\beta})$ as follows.

$$Q(\boldsymbol{\beta}) = \frac{1}{2n} \left\| Y - \sum_{j \in \mathcal{D}} \mathbf{X}_j \boldsymbol{\beta}_j \right\|_2^2 + \sum_{j \in \mathcal{D}} \rho_{\lambda,a}(\|\boldsymbol{\beta}_j\|_2), \tag{3.3}$$

where $\rho_{\lambda,a}(\cdot)$ is the SCAD penalty with regularization parameters λ, a and is defined as

$$\rho_{\lambda,a}(|x|) = \begin{cases} \lambda|x|, \text{if } |x| \leq \lambda \\ -(|x|^2 - 2a\lambda|x| + \lambda^2)/(2(a-1)), \text{if } \lambda < |x| < a\lambda \\ (a+1)\lambda^2/2, \text{if } |x| > a\lambda. \end{cases}$$

The penalty function $\rho_{\lambda,a}(|x|)$ is a quadratic spline function with two knots at λ and $a\lambda$. Following the suggestion in Fan and Li (2001), we also take the extra regularization parameter $a = 3.7$ for group variable selection case.

3.3 Group MCP

As group Lasso can only select the variables at the group level, Breheny and Huang (2009) proposed another method called group minimax concave penalty (grMCP) which can also select the individual variables within a group at the same time. The original MCP in Zhang (2007) is a nonconcave penalty that has the same motivation with SCAD in Fan and Li (2001) but faster rate of penalization for some situations. Like SCAD penalty function, MCP is also a piecewise function and defined as

$$f_{\lambda,a}(x) = \begin{cases} \lambda x - x^2/2a, \text{if } x \leq a\lambda \\ a\lambda^2/2, \text{if } x > a\lambda, \end{cases}$$

for $\lambda > 0$, in which a, λ are two regularization parameters. Observe that the first derivative of MCP function is $f'_{\lambda,a}(x) = \lambda - x/a$ if $x \leq a\lambda$ and 0 otherwise with respect to x. This implies that the rate of penalization drops to 0 when $x > a\lambda$, obtaining the unbiased estimation of the large coefficients that is greater than $a\lambda$. To accomplish the bi-level selection, the group MCP in Breheny and Huang (2009) places an outer MCP on a sum of inner MCP for each group, which yields the following objective function

$$Q(\boldsymbol{\beta}) = \frac{1}{2n} \left\| Y - \sum_{j \in \mathcal{D}} \mathbf{X}_j \boldsymbol{\beta}_j \right\|_2^2 + \sum_{j \in \mathcal{D}} f_{\lambda,a} \left(\sum_{k=1}^{p_j} f_{\lambda,b}(|\beta_{jk}|) \right),$$

where b is the regularization parameter for inner penalty function and is typically chosen to be $p_j a\lambda/2$ to ensure that the group level penalty attains its maximum. This choice also simplifies three regularization parameters a, b and λ into two, which makes the selection procedure simpler. Breheny and Huang (2009) recommended $a = 3$ that works well in practice if the variables have been initially standardized. We ultimately have only one regularization parameter λ to choose, similar to group Lasso and group SCAD.

Chapter 4

Numerical Studies

In this chapter, we carry out the intensive simulations to evaluate the finite sample performance of gSIS, gHOLP, gAR2 and gDC. We first examine the screening accuracy at the first stage in section 4.1, and compare the performance of grouped variable selection at the second stage after screening in section 4.2. The simulation settings of the first three models are very similar to those in Yuan and Lin (2006), except that we generate much larger number of p (or equivalently, J) predictors. That is, Yuan and Lin (2006) only considered the "small J, large n" case but we focus on the reverse situation now. Also, we consider two special models: Model IV and VII which are sparse within groups and a nonlinear model, respectively. Finally, Model VI and V are used to check the effect of severe collinearity among groups and the weakly marginal but strongly joint correlation between groups and response, respectively. We expect the screening methods by marginal correlation learning would fail on these two models. We examine the performance of these methods mentioned in section 2.1 for the cases where $(n, J) = (200, 2000)$ and $(800, 5000)$, respectively. More specifically, the simulation settings are described as follows. The error term ε for all models follows a normal distribution $\mathcal{N}(0, \sigma^2)$, where σ^2 is set such that $\mathcal{R}^2 = \mathrm{Var}(\mathbf{X}^T\boldsymbol{\beta})/\mathrm{Var}(Y)$ is equal to $0.3, 0.5, 0.9$ for checking the performance of screening criteria in the presence of

low, moderate, and high signal-to-noise ratio, respectively.

- *Model I:* J latent variables $Z_1, \ldots, Z_J$ are first generated from the multivariate normal distribution with zero mean vector and covariance between Z_i and Z_j being $0.5^{|i-j|}$. Then the covariates $Z_1, \ldots, Z_J$ are discretized to $0, 1, 2$ by $\Phi^{-1}(1/3)$ and $\Phi^{-1}(2/3)$ leading to $\mathbf{X}_j = (I(Z_j = 0), I(Z_j = 1), I(Z_j = 2)), j = 1, \ldots, J$. The response Y is obtained from the model

$$Y = \sum_{j \in \mathcal{S}} \boldsymbol{\beta}_j \mathbf{X}_j + \varepsilon,$$

  where $\mathcal{S}$ is the indices set randomly selected from $1, \ldots, J$ with $s = 4$ different indices in it, e.g., $\mathcal{S} = \{3, 21, 34, 59\}$, and let $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, 0)^T$. We set $\beta_{j3} = 0$ here to avoid the perfect collinearity. Following Fan and Lv (2008), β_{j1}, β_{j2} are simulated independently from $\beta = (-1)^U(a + |z|)$ for $j \in \mathcal{S}$, where $a = 4\log(n)/\sqrt{n}, U \sim Bernoulli(0.4)$ and $z \sim \mathcal{N}(0, 1)$, leading to the model

$$Y = \sum_{j \in \mathcal{S}} [\beta_{j_1} I(Z_j = 0) + \beta_{j_2} I(Z_j = 1)] + \varepsilon.$$

- *Model II*: Random variables $Z_1, \ldots, Z_J$ are generated from the same way as in Model I. The new covariates X_j are defined as $X_j = (Z_j + W)/\sqrt{2}$, where W independent of Z_j is generated from standard normal distribution. Each of $X_1, \ldots, X_J$ are expanded through a third order orthogonal polynomial, i.e., $\mathbf{X}_j = (X_j, X_j^2, X_j^3)$, and only main effects of them are considered. The response and the index $\mathcal{S}$ of true groups are generated in the same way as in Model I, except that β_{j3} is also generated from $\beta = (-1)^U(a + |z|)$, leading to $\beta_{j1}, \ldots, \beta_{j3} \neq 0, j \in \mathcal{S}$. Specifically, the model is written as

$$Y = \sum_{j \in \mathcal{S}} \left(\beta_{j1} X_j + \beta_{j2} X_j^2 + \beta_{j3} X_j^3\right) + \varepsilon.$$

Note that the number of true important groups is $s = 4$.

- *Model III*: This model contains both continuous and categorical group variables. For simplicity, we generate $\lfloor J/2 \rfloor$ continuous group covariates in the same fashion as in Model II, and $J - \lfloor J/2 \rfloor$ categorical group variables in the same fashion as in Model I, so that the total number of groups is still J. However, the number of true important groups for continuous covariates is still $s_1 = 4$ with indices set $\mathcal{S}_1$ chosen from $\{1, \ldots, \lfloor J/2 \rfloor\}$ and the $s_2 = 1$ with index set $\mathcal{S}_2$ chosen from $\{J - \lfloor J/2 \rfloor + 1, \cdots, J\}$ for categorical covariates case such that the total number of true important groups is $s = s_1 + s_2 = 5$ and $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$. Note that $\mathcal{S}_1 \cap \mathcal{S}_2 = \varnothing$. The coefficients for continuous covariates are generated exactly the same as in Model II. The coefficient vector for categorical covariates in j_2-th group is set to be $(2, 1, 0)^T, j_2 \in \mathcal{S}_2$. Thus, the response Y is generated by

$$Y = \sum_{j_1 \in \mathcal{S}_1} \left(\beta_{j_1 1} X_{j_1} + \beta_{j_1 2} X_{j_1}^2 + \beta_{j_1 3} X_{j_1}^3 \right) + 2I(X_{j_2 1} = 0) + I(X_{j_2 2} = 0) + \varepsilon.$$

- *Model IV*: We consider a linear model that is sparse within groups. The covariates $\mathbf{X}_j$ and response Y are generated in the same manner as in Model III, except that the coefficients vector for group $j_1 \in \mathcal{S}_1$ with continuous covariates are set to be 0 in the second entry, i.e., $\beta_{j_1 2} = 0$, and the coefficients for group $j_2 \in \mathcal{S}_2$ with categorical covariates are also set to be 0 in the second entry, i.e, $(2, 0, 0)$. Then the model is expressed as

$$Y = \sum_{j_1 \in \mathcal{S}_1} \left(\beta_{j_1 1} X_{j_1} + \beta_{j_1 3} X_{j_1}^3 \right) + 2I(X_{j_2 1} = 0) + \varepsilon.$$

  Note that the number of true important groups is $s = 5$.

- *Model V*: In this model we examine the case where two groups are marginally uncorrelated but jointly correlated with the response. Specifically, we first generate J random

variables $Z_1, \ldots, Z_J$ from multivariate normal distribution with zero mean vector and covariance $\mathbf{\Sigma} = (\sigma_{ij})_{J\times J}$, where $\sigma_{ii} = 1, i = 1, \ldots, J$ and $\sigma_{ij} = 0.5, i \neq j$. This implies that all Z_j's have the same correlation strength $\rho = 0.5$ between each other. In each group with size $p_j = 4$, each predictor $\mathbf{x}_{jp_j}$ is generated from $Z_j + \epsilon_{jp_j}$, where $\epsilon_{jp_j} \sim \mathcal{N}(0, 1), j = 1, \ldots, J$. The response is then generated from

$$Y = \mathbf{5}\mathbf{X}_1 + \mathbf{5}\mathbf{X}_2 - \frac{\mathbf{10}}{\mathbf{3}}\mathbf{X}_3 - \frac{\mathbf{10}}{\mathbf{3}}\mathbf{X}_4 + \varepsilon,$$

in which $\mathbf{5} = (5, 5, 5, 5), \mathbf{10/3} = (10/3, 10/3, 10/3, 10/3)$. Note that the true important group $\mathcal{S} = \{1, 2, 3, 4\}$ with size $s = 4$. Clearly, $\mathbf{X}_3, \mathbf{X}_4$ are marginally uncorrelated (i.e., checking $\text{cov}(Y, \mathbf{X}_3) = 0, \text{cov}(Y, \mathbf{X}_4) = 0$) but jointly correlated with response Y due to the nonzero coefficients in the model.

- *Model VI*: To check the effect of strong correlation between groups, we now generate all predictors $\mathbf{X}$ from multivariate normal distribution with mean vector $\mathbf{1}$ and covariance $\mathbf{\Sigma} = (\sigma_{ij})_{4J\times 4J}$, where $\sigma_{ii} = 1, i = 1, \ldots, 4J$ and $\sigma_{ij} = 0.8, i \neq j$, i.e., $\mathbf{X} \sim \mathcal{N}(\mathbf{1}, \mathbf{\Sigma})$, in which $\mathbf{1}$ is a $1 \times 4J$ vectors with all entries 1. The true important groups index and the coefficients are taken the same as in Model V and Model I, respectively. Consequently, the model is

$$Y = \boldsymbol{\beta}_1\mathbf{X}_1 + \boldsymbol{\beta}_2\mathbf{X}_2 + \boldsymbol{\beta}_3\mathbf{X}_3 + \boldsymbol{\beta}_4\mathbf{X}_4 + \varepsilon,$$

where $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{j4})$ are generated from $\beta = (-1)^U(a + |z|)$ independently as in Model I, and $j \in \mathcal{S} = \{1, 2, 3, 4\}$.

- *Model VII*: We finally examine the performance of the proposed group screening methods for nonlinear model. As gDC is model-free screening criterion, we expect it would stand out in this example in terms of the screening accuracy. The grouped covariates $\mathbf{X}_j \sim \mathcal{N}(0.5v\mathbf{1}, \mathbf{\Sigma}_j), j = 1, \ldots, J$ where v is a random number from $[0, 1]$ and $\mathbf{1}$ is a

$1 \times p_j$ vector, and $\mathbf{\Sigma}_j = (\sigma_{ik})_{p_j \times p_j}$ is a $p_j \times p_j$ covariance matrix that have the same values as in Model VI, i.e., $\sigma_{ii} = 1, i = 1, \ldots, p_j$ and $\sigma_{ik} = 0.8, i \neq k, k = 1, \ldots, p_j$. The coefficients are also generated from the same settings as Model VI, except that the response Y now becomes

$$Y = \sum_{j \in \mathcal{S}} \sum_{k=1}^{p_j} 2\beta_{jk} \sin^2(X_{jk}) + \varepsilon,$$

in which $\mathcal{S} = [1, 2, 3, 4]$ and group size $p_j = 4$ for all $j = 1, \ldots, J$.

4.1 Screening accuracy

For each model, the threshold d is chosen as the sample size n for the purpose of comparison. We record the proportion of recovering the number of true important groups from the selected submodels $\mathcal{M}_{\mathcal{D}}^{g}$. By proportion of recovering the number of true important groups we mean a proportion that equals the number of true important groups covered by $\mathcal{M}_{\mathcal{D}}^{g}$ divided by the number of true important groups s, the true size of $\mathcal{M}_{S}^{g}$. For example, suppose the true important group indices in $\mathcal{M}_{S}^{g}$ is $\{1, 2, 3, 4\}$ and the selected group indices by screening in $\mathcal{M}_{\mathcal{D}}^{g}$ is $\{1, 3, 4, 5, 7, 8\}$. Because the selected groups recover three true important groups $\{1, 3, 4\}$, the proportion of coverage of true important groups is $3/4 = 0.75$. Table 4.1 and Table 4.2 show the mean proportion of recovering the number of true important groups (**Proportion Coverage**) for each screening criterion based on 100 repetitions for $(n, J) = (200, 2000)$ and $(n, J) = (800, 5000)$, respectively. Meanwhile, we also report the mean proportion of recovering the exact number of true important groups (**Exact Coverage**) base on the 100 replications in Table 4.1 for $(n, J) = (200, 2000)$ and 4.2 for $(n, J) = (800, 5000)$. For the example given above, proportion 0.75 means the selected submodels recover all true important groups $\{1, 2, 3, 4\}$ exactly 75 times out of 100 simulations. In other words, there

are 75 screened submodels $\mathcal{M}_d^g$ containing all the true important groups $\{1, 2, 3, 4\}$ in these 100 selected submodels.

The common phenomena from Table 4.1 and 4.2 is that the performance is getting better as the signal-to-noise ratio is increasing. It is sensible because higher signal-to-noise ratio leads to more contribution of predictors to response or outcome. We can also observe that gSIS, gHOLP, gAR2 and gDC are competitive in terms of the coverage rate for the linear model with independent predictors within groups and weak correlation between groups (e.g., Model I, II, III, IV), but gAR2 may be slightly better than the others due to the fact that the $\bar{R}^2$ is able to measure the relationship between multiple predictors and response more naturally, and is much more stable than the estimation of coefficients that gSIS needs to do, regardless of the within-group correlation of predictors. In a word, gAR2 better takes into account for the group structure than gSIS does. In Model V, as the $\mathbf{X}_3$ and $\mathbf{X}_4$ are marginally uncorrelated but jointly correlated with response, the gSIS, gAR2 and gDC that rely on the assumption of strongly marginal correlation will fail to screen out these two groups modulo the random guess, and are only able to discover the first two true groups, resulting in only 50% coverage as expected showing in Table 4.1 and 4.2 even if the signal-to-noise ratio is high, but gHOLP performs rather well and stands out in this case. When there exists severe collinearity between groups or even within groups as in Model VI, gSIS, gAR2 and gDC will still suffer from this problem, leading to poor performance in the screening, but gAR2 is still superior to the gSIS and gDC because it is more robust to the collinearity. Meanwhile, gHOLP still performs the best. Finally, for the nonlinear model in Model VII, gDC is the only method that can successfully screen out the important groups. Note that it is flexible to adjust the submodel size d to increase the coverage depending on the situation of interest. For example, we can enlarge d to the twice sample size $2n$, which can certainly improve the screening accuracy but also increase the computation load as well as the false positive rate. We can use random permutation idea discussed in section 2.5 to determine d if we only focus

on one specific screening criterion and do not intent to compare the performance of different methods.

Table 4.1: The proportion of recovering the number of true important groups when $(n, J) = (200, 2000)$ based on 100 replications.

Model	$\mathcal{R}^2$	**Proportion Coverage**				**Exact Coverage**			
		gSIS	gHOLP	gAR2	gDC	gSIS	gHOLP	gAR2	gDC
Model I	0.3	0.903	0.900	0.903	0.878	0.690	0.660	0.680	0.600
	0.5	0.988	0.870	0.983	0.968	0.950	0.880	0.930	0.880
	0.9	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Model II	0.3	0.945	0.938	0.948	0.863	0.790	0.760	0.790	0.540
	0.5	0.988	0.990	0.990	0.980	0.950	0.960	0.960	0.920
	0.9	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Model III	0.3	0.608	0.610	0.650	0.636	0.090	0.070	0.090	0.090
	0.5	0.800	0.814	0.846	0.792	0.320	0.320	0.390	0.270
	0.9	0.934	0.954	0.950	0.906	0.680	0.770	0.760	0.580
Model IV	0.3	0.440	0.446	0.518	0.510	0.000	0.000	0.000	0.030
	0.5	0.570	0.568	0.656	0.624	0.000	0.020	0.060	0.100
	0.9	0.780	0.804	0.856	0.808	0.210	0.260	0.390	0.270
Model V	0.3	0.578	0.935	0.558	0.558	0.070	0.750	0.010	0.020
	0.5	0.553	0.988	0.543	0.553	0.030	0.950	0.000	0.020
	0.9	0.535	1.000	0.523	0.533	0.010	1.000	0.000	0.020
Model VI	0.3	0.285	0.465	0.450	0.273	0.010	0.120	0.080	0.000
	0.5	0.365	0.645	0.640	0.370	0.020	0.310	0.190	0.000
	0.9	0.498	0.985	0.860	0.590	0.020	0.950	0.510	0.020
Model VII	0.3	0.205	0.205	0.163	0.443	0.000	0.000	0.000	0.030
	0.5	0.240	0.248	0.190	0.575	0.010	0.000	0.000	0.060
	0.9	0.293	0.298	0.273	0.685	0.010	0.000	0.000	0.180

4.2 Selection accuracy

The scheme of screening in Figure 3.1 reveals the two-stage procedure that variable screening is followed by variable selection. In this section, we investigate this two-stage procedure strategy by comparing the performance of grouped variable selection without and with applying screening at the first stage. Precisely, "None" screening is defined as a one-stage

Table 4.2: The proportion of recovering the number of true important groups when $(n, J) = (800, 5000)$ based on 100 replications.

Model	$\mathcal{R}^2$	**Proportion Coverage**				**Exact Coverage**			
		gSIS	gHOLP	gAR2	gDC	gSIS	gHOLP	gAR2	gDC
Model I	0.3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.9	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Model II	0.3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.9	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Model III	0.3	0.664	0.674	0.698	0.704	0.300	0.300	0.300	0.400
	0.5	0.792	0.794	0.832	0.804	0.500	0.300	0.500	0.400
	0.9	0.934	0.954	0.952	0.904	0.700	0.700	0.800	0.700
Model IV	0.3	0.488	0.486	0.548	0.558	0.200	0.100	0.200	0.200
	0.5	0.616	0.606	0.672	0.642	0.400	0.300	0.400	0.400
	0.9	0.742	0.782	0.792	0.854	0.400	0.500	0.700	0.800
Model V	0.3	0.540	1.000	0.540	0.533	0.100	1.000	0.100	0.000
	0.5	0.528	1.000	0.525	0.523	0.030	1.000	0.000	0.000
	0.9	0.508	1.000	0.505	0.508	0.010	1.000	0.000	0.000
Model VI	0.3	0.490	0.855	0.790	0.535	0.000	0.700	0.600	0.100
	0.5	0.558	0.948	0.860	0.588	0.100	0.900	0.600	0.100
	0.9	0.615	1.000	0.930	0.663	0.200	1.000	0.800	0.200
Model VII	0.3	0.413	0.380	0.358	0.788	0.000	0.000	0.000	0.400
	0.5	0.478	0.468	0.450	0.858	0.000	0.000	0.100	0.500
	0.9	0.575	0.563	0.555	0.923	0.000	0.000	0.100	0.900

procedure and means that the grouped variable selection is used directly for the original datasets, without applying any screening methods. The others are two-stage procedure with utilizing different screening methods. As the grouped variable selection can be only used for linear model, we only consider the first six linear models with the fixed $\mathcal{R}^2 = 0.9$ for simplicity. To choose the regularization parameter λ in group Lasso, group SCAD and group MCP, we make use of 10-fold cross-validation method at certain grid points. To accomplish the aim of selection, we employ the R package `grpreg` with `cv.grpreg` function to choose

the optimal regularization parameter λ and use `grpreg` to conduct the grouped variable selection as well as the parameter estimation. The default method of `cv.grpreg` is 10-fold cross-validation. To compare the performance of these methods, we record the following measurements:

- #FNG: the average number of false negative groups.
- #FPG: the average number of false positive groups.
- Coverage: the average proportion of including the true models from selected models.
- Exact: the average proportion of selecting models being equal to the true models exactly.
- Error: the average estimation error defined as $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$.
- Size: the average size (nonzero groups) of the selected model.
- Time: the average computation time in seconds.

Note that all measurements are calculated from 100 replications (i.e., 100 datasets for each model). Since gSIS, gAR2 and gDC are all marginal correlation learning and have very similar performances in terms of the measurements above, we only display the results of gHOLP and gAR2 in Tables 4.3 and 4.4 to save space.

Table 4.3 and Table 4.4 show the commonly used measurements of accuracy for $(n, J) = (200, 2000)$ and $(n, J) = (800, 5000)$ based on 100 replications. One can observe that the two-stage procedure is fairly competitive or even superior to the one-stage procedure in which the grouped variable selection is applied directly, in terms of the accurate measurements of estimations. Even if in some cases the two-stage procedure is slightly worse than one-stage procedure, it does not lose much efficiency. However, if we compare the computation time on an ordinary PC with Intel Core i5 1.60 GHz processor and 6.0 GB RAM, the two-stage

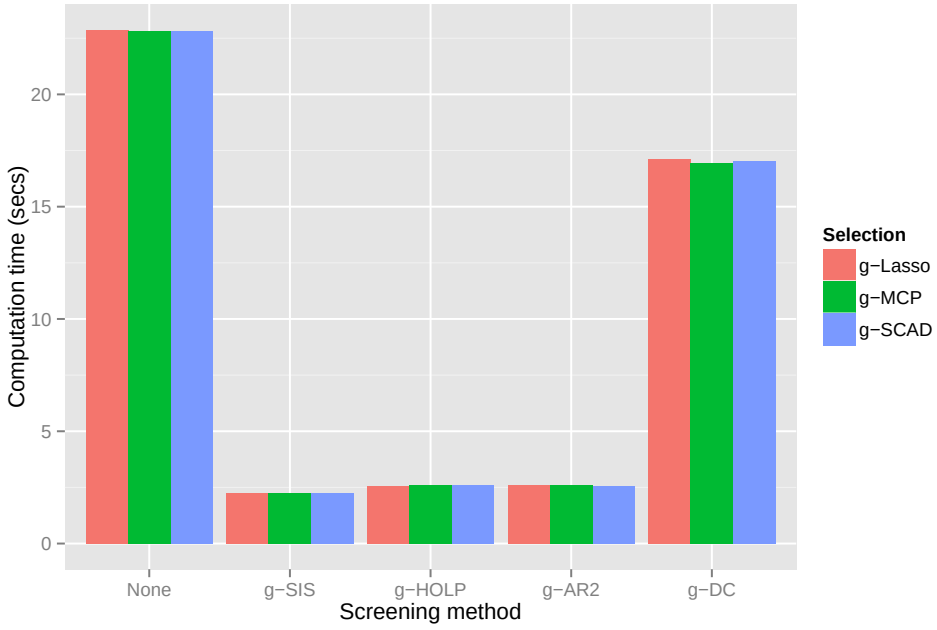
procedures with incorporation of gSIS, gHOLP, gAR2 are remarkably faster than the one-stage procedure, which can be seen in Figures 4.1 and 4.2. Since the computation time for six simulated models are very similar, we only display the bar chart of computation time for the first two models in Figures 4.1 and 4.2. We can see that the only exception is that the two-procedure by gDC screening is not superior or even worse than the one-stage procedure in terms of the computation time due to the fact that the gDC costs a lot of time to compute the distance correlation with three complex steps, especially in situations where the sample size and group size are large. It is concluded that screening procedures by gSIS, gHOLP and gAR2 not only maintain or even improve the estimation accuracy of grouped variable selection, but also boost the computation speed. We need to keep in mind that the gDC may make the situation worse if the sample size or group size is large. The simulation results also show that the performance of grLasso is slightly worse than the other grouped variable selection methods in terms of the false positive groups due to its biased estimation.

Table 4.3: Accuracy of grouped variable selection for linear models I - VI with $(n, J) = (200, 2000)$ based on 100 replications.

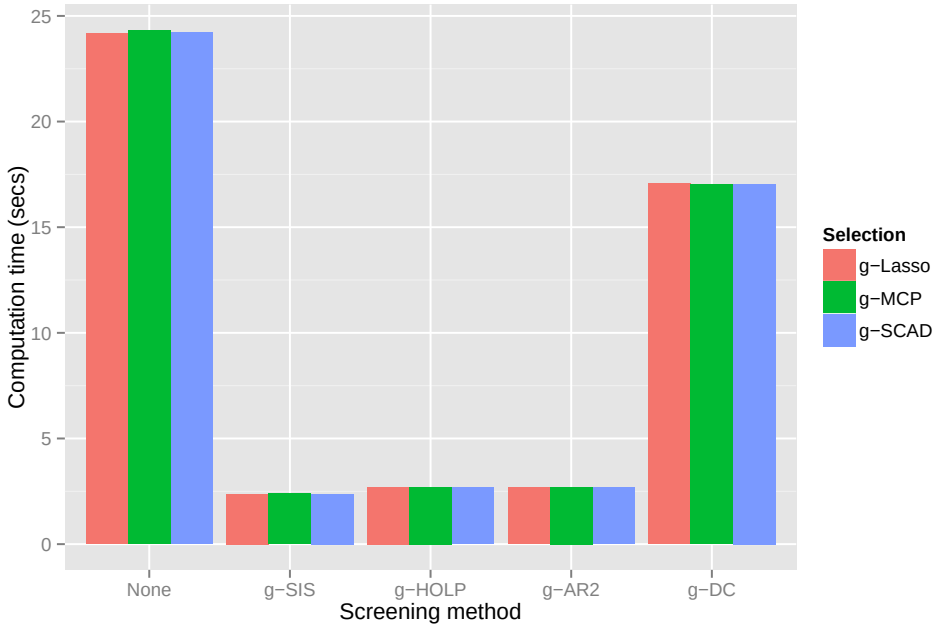
Model	Screening	Selection	#FNG	#FPG	Coverage	Exact	Size	Error
Model I	None	grLasso	1.91	8.07	0.52	0.07	10.16	5.24
	gHOLP		1.90	15.07	0.53	0.07	17.17	5.26
	gAR2		1.92	14.52	0.52	0.07	16.60	5.26
	None	grSCAD	1.92	5.67	0.52	0.07	7.75	5.03
	gHOLP		1.92	9.19	0.52	0.07	11.27	5.11
	gAR2		1.93	9.23	0.52	0.07	11.30	5.11
	None	grMCP	1.94	5.50	0.52	0.07	7.56	5.10
	gHOLP		1.89	9.04	0.53	0.07	11.15	5.13
	gAR2		1.92	8.38	0.52	0.07	10.46	5.11
Model II	None	grLasso	1.45	14.93	0.64	0.17	17.48	7.71
	gHOLP		1.46	23.73	0.64	0.17	26.27	7.71
	gAR2		1.46	22.75	0.64	0.17	25.29	7.71
	None	grSCAD	1.53	10.42	0.62	0.18	12.89	7.64
	gHOLP		1.64	16.05	0.59	0.17	18.41	7.69
	gAR2		1.67	15.15	0.58	0.17	17.48	7.69
	None	grMCP	1.62	7.67	0.60	0.17	10.05	7.66
	gHOLP		1.77	12.18	0.56	0.16	14.41	7.74
	gAR2		1.90	12.91	0.53	0.14	15.01	7.75
Model III	None	grLasso	1.83	15.36	0.63	0.06	18.53	7.95
	gHOLP		1.76	15.83	0.60	0.06	18.81	7.74
	gAR2		1.79	18.55	0.60	0.05	21.53	7.76
	None	grSCAD	1.86	10.70	0.63	0.06	13.84	7.79
	gHOLP		1.75	10.11	0.60	0.06	13.10	7.61
	gAR2		1.81	12.19	0.59	0.05	15.15	7.63
	None	grMCP	1.96	7.13	0.61	0.06	10.17	7.82
	gHOLP		1.79	8.83	0.59	0.06	11.78	7.63
	gAR2		1.85	9.19	0.58	0.05	12.11	7.66
Model IV	None	grLasso	2.08	13.36	0.58	0.06	16.28	6.52
	gHOLP		1.58	14.74	0.48	0.01	17.13	5.76
	gAR2		1.68	17.22	0.49	0.00	19.65	5.86
	None	grSCAD	2.12	9.05	0.58	0.06	11.93	6.36
	gHOLP		1.62	11.07	0.47	0.00	13.42	5.63
	gAR2		1.70	11.80	0.48	0.00	14.21	5.73
	None	grMCP	2.13	6.75	0.57	0.06	9.62	6.35
	gHOLP		1.61	8.10	0.47	0.01	10.46	5.64
	gAR2		1.71	8.88	0.48	0.01	11.28	5.75
Model V	None	grLasso	2.00	27.49	0.50	0.00	29.49	6.24
	gHOLP		2.00	30.47	0.50	0.00	32.47	7.11
	gAR2		0.09	29.91	0.50	0.00	31.91	14.44
	None	grSCAD	2.00	18.33	0.50	0.00	20.33	9.11
	gHOLP		2.00	19.05	0.50	0.00	21.05	9.88
	gAR2		0.09	12.20	0.50	0.00	14.20	26.71
	None	grMCP	2.00	8.74	0.50	0.00	10.74	10.17
	gHOLP		2.00	10.45	0.50	0.00	12.45	13.06
	gAR2		0.09	10.82	0.50	0.00	12.82	33.07
Model VI	None	grLasso	1.31	30.64	0.67	0.50	33.33	6.15
	gHOLP		1.32	50.17	0.66	0.44	52.79	6.67
	gAR2		1.09	37.73	0.59	0.20	40.08	6.01
	None	grSCAD	1.94	14.10	0.52	0.10	16.16	6.27
	gHOLP		1.97	23.93	0.49	0.09	25.90	6.91
	gAR2		1.49	17.86	0.49	0.08	19.81	6.52
	None	grMCP	2.12	7.08	0.47	0.04	8.96	6.15
	gHOLP		2.11	13.52	0.46	0.06	15.35	8.53
	gAR2		1.66	10.85	0.45	0.05	12.63	7.24

Table 4.4: Accuracy of grouped variable selection for linear models I - VI with $(n, J) = (800, 5000)$ based on 100 replications.

Model	Screening	Selection	#FNG	#FPG	Coverage	Exact	Size	Error
Model I	None	grLasso	2.06	1.08	0.49	0.03	3.02	3.84
	gHOLP		2.05	1.58	0.49	0.04	3.53	3.84
	gAR2		2.05	1.35	0.49	0.04	3.30	3.84
	None	grSCAD	2.03	0.81	0.49	0.04	2.78	3.75
	gHOLP		2.04	0.92	0.49	0.04	2.88	53.75
	gAR2		2.04	0.75	0.49	0.04	2.71	3.75
	None	grMCP	2.02	1.30	0.50	0.04	3.28	3.75
	gHOLP		2.04	1.68	0.49	0.04	3.64	3.75
	gAR2		2.04	1.46	0.49	0.04	3.42	3.75
Model II	None	grLasso	1.35	12.65	0.66	0.22	15.30	6.08
	gHOLP		1.35	19.47	0.66	0.22	22.12	6.08
	gAR2		1.35	19.38	0.66	0.22	22.03	6.08
	None	grSCAD	1.35	9.51	0.66	0.22	12.16	6.06
	gHOLP		1.35	14.05	0.66	0.22	16.70	6.06
	gAR2		1.35	13.94	0.66	0.22	16.59	6.06
	None	grMCP	1.35	9.90	0.66	0.22	12.55	6.06
	gHOLP		1.35	14.01	0.66	0.22	16.66	6.07
	gAR2		1.35	13.21	0.66	0.22	15.86	6.07
Model III	None	grLasso	1.99	1.57	0.60	0.11	4.58	6.45
	gHOLP		1.84	1.09	0.57	0.05	3.92	6.27
	gAR2		1.89	1.20	0.57	0.05	4.03	6.30
	None	grSCAD	2.01	0.84	0.60	0.11	3.83	6.40
	gHOLP		1.86	0.45	0.56	0.05	3.26	6.21
	gAR2		1.91	0.48	0.56	0.05	3.29	6.24
	None	grMCP	2.01	0.75	0.60	0.11	3.74	6.39
	gHOLP		1.86	0.38	0.56	0.05	3.19	6.20
	gAR2		1.91	0.42	0.56	0.05	3.23	6.23
Model IV	None	grLasso	2.37	1.25	0.53	0.04	3.88	5.30
	gHOLP		1.70	0.86	0.42	0.00	2.98	4.71
	gAR2		1.77	1.01	0.43	0.00	3.14	4.75
	None	grSCAD	2.35	0.96	0.53	0.04	3.61	5.27
	gHOLP		1.73	0.48	0.42	0.00	2.57	4.68
	gAR2		1.77	0.56	0.43	0.00	2.69	4.72
	None	grMCP	2.35	0.79	0.53	0.04	3.44	5.25
	gHOLP		1.71	0.46	0.42	0.00	2.57	4.66
	gAR2		1.78	0.50	0.42	0.00	2.62	4.70
Model V	None	grLasso	2.00	30.15	0.50	0.00	32.15	5.02
	gHOLP		2.00	22.16	0.50	0.00	24.16	7.47
	gAR2		0.02	9.94	0.50	0.00	11.94	9.56
	None	grSCAD	2.00	22.16	0.50	0.00	24.16	7.47
	gHOLP		2.00	23.50	0.50	0.00	25.50	7.40
	gAR2		0.02	9.94	0.50	0.00	11.94	9.56
	None	grMCP	2.00	17.29	0.50	0.00	19.29	7.60
	gHOLP		2.00	19.04	0.50	0.00	21.04	7.58
	gAR2		0.02	10.10	0.50	0.00	12.10	13.01
Model VI	None	grLasso	1.14	39.31	0.72	0.51	42.17	2.60
	gHOLP		1.17	46.46	0.71	0.46	49.29	2.63
	gAR2		1.06	36.14	0.67	0.37	38.80	2.39
	None	grSCAD	1.60	5.18	0.60	0.10	7.58	1.86
	gHOLP		1.56	5.71	0.61	0.11	8.15	1.52
	gAR2		1.38	7.48	0.59	0.13	9.82	1.66
	None	grMCP	1.62	4.55	0.60	0.10	6.93	1.53
	gHOLP		1.59	5.44	0.60	0.10	7.85	1.54
	gAR2		1.47	6.33	0.56	0.07	8.58	1.62

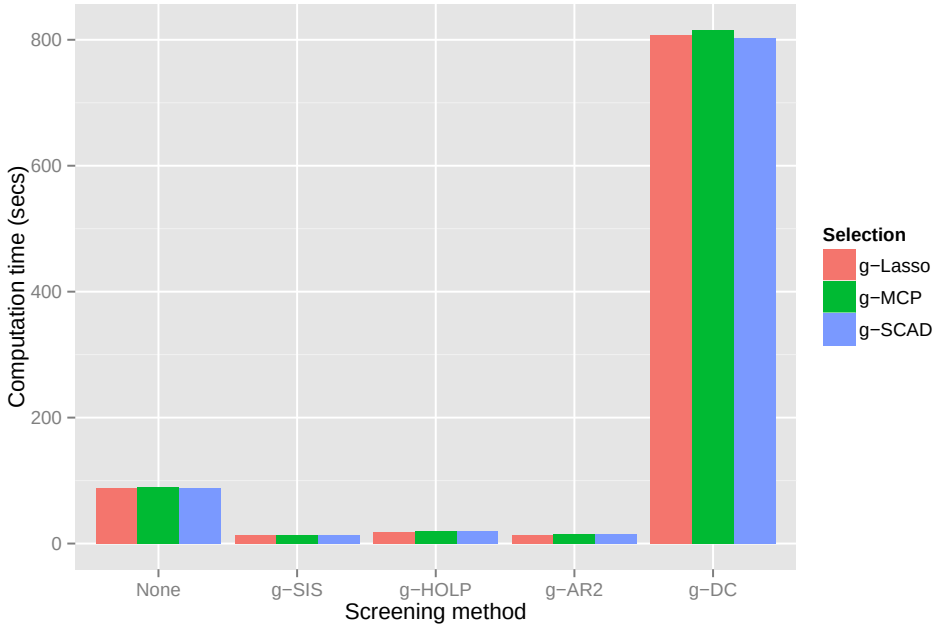


(a) Model I

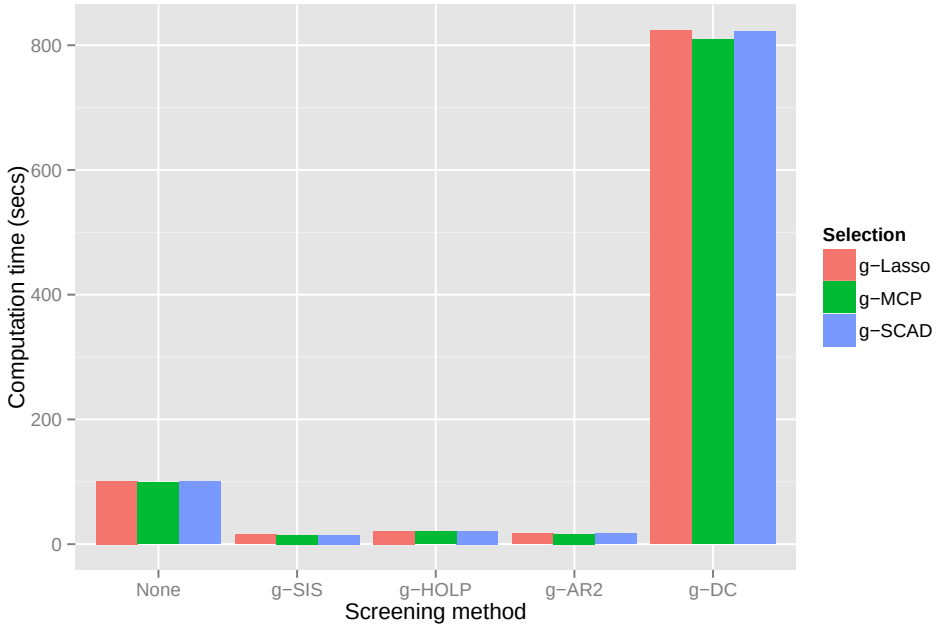


(b) Model II

Figure 4.1: Bar charts of average computation time for the combinations of group screening and selection procedure with $(n, J) = (200, 2000)$ based on 100 simulations.



(a) Model I



(b) Model II

Figure 4.2: Bar charts of average computation time for the combinations of group screening and selection procedure with $(n, J) = (800, 5000)$ based on 100 simulations.

Chapter 5

Application for GAW17 Dataset

To test the proposed approaches on a realistic situation, we analyze a real dataset that is a hybrid of simulated and real data from the 2010 Genetic Analysis Workshop 17 (GAW17). GAW17 dataset contains real exome sequencing data from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010, http://www.1000genomes.org) that is designed to survey genetic variation at the sequence level across multiple human population groups.The data consist of 697 unrelated individuals and 24487 autosomal single-nucleotide polymorphisms (SNPs) that are assigned to 3205 genes based on the first intersection found of the marker location and the basepair coordinates of all the genes. In other words, we collected 24487 SNPs and grouped into 3205 genes for each sample. Two hundred independent quantitative risk factors Q1 were simulated by the organizers of the workshop according to a plausible phenotype model. Almasy *et al.*, (2011) described greater details of dataset and simulation settings. In the notation of this dissertation, $n = 697, J = 3205, p = 24487$. To simplify the process, we take the average of two hundred independent phenotype Q1 as the response for each individual sample. According to the phenotype model, we know that the quantitative risk factor Q1 was influenced by 39 SNPs in 9 genes (see Table 1 in Almasy *et al.*, 2011). There were 1-11 functional SNPs per gene, which means the number of co-

variates varies among the groups. To select the significant genes and/or SNPs, we apply group Lasso (grLasso), group SCAD (grSCAD), group MCP (grMCP) and group exponential Lasso (GEL) to perform the grouped variable selection on the original dataset without screening procedure and the screened dataset with screening step, respectively. Note that only grLasso does not perform bi-level selection. In particular, the threshold of screening procedure is set to be 1000 groups and we use gSIS, gHOLP and gAR2, respectively at the first stage, combined with grLasso, grSCAD, grMCP and GEL at the second stage for the two-stage procedure. Thus, we reduce almost two thirds of the total groups (i.e., decreases $J = 3205$ to $d = 1000$) at the first stage but retain most of the useful information by screening approach, such that the computation load is decreased and the accuracy of estimation is maintained or even improved at the second stage. To compare their performances, we report the running time in seconds (Time), mean squared error (MSE $= n^{-1}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$), number of selected SNPs (SNPs) and genes (genes) that have nonzero coefficients, number of truly selected SNPs (SNPs_true) from 39 true SNPs and genes (genes_true) from 9 true genes.

Table 5.1 shows that the proposed two-stage procedures are superior to the one-stage procedures in terms of almost all measurements, except for grLasso. However, even one stage of grLasso is somewhat better than the two-stage procedure, its computation load is much more intensive and false discovery rate is larger than the others in the sense that the number of selected SNPs and genes are much larger than those selected by other methods. We can also see that gAR2 performs better than other screening methods due to the severe collinearity of the covariates in GAW17 dataset. The real data analysis confirms again that the proposed grouped variable screening methods can improve the results of grouped variable selection in terms of the computation time and accuracy of estimation.

	Time	MSE	SNPs	genes	SNPs_true	genes_true
grLasso	106.44	0.07	297.00	159.00	27.00	5.00
gSIS-grLasso	17.31	0.08	252.00	122.00	23.00	3.00
gHOLP-grLasso	56.36	0.07	295.00	139.00	22.00	3.00
gAR2-grLasso	58.64	0.10	290.00	101.00	27.00	5.00
grSCAD	106.28	0.11	106.00	59.00	15.00	3.00
gSIS-grSCAD	20.29	0.12	60.00	29.00	10.00	1.00
gHOLP-grSCAD	60.35	0.11	72.00	37.00	10.00	1.00
gAR2-grSCAD	61.84	0.11	100.00	35.00	23.00	4.00
grMCP	101.56	0.12	48.00	26.00	10.00	1.00
gSIS-grMCP	18.42	0.12	40.00	18.00	10.00	1.00
gHOLP-grMCP	61.30	0.10	51.00	26.00	10.00	1.00
gAR2-grMCP	60.82	0.12	50.00	19.00	11.00	2.00
GEL	57.75	0.11	49.00	11.00	23.00	4.00
gSIS-GEL	11.14	0.12	42.00	5.00	21.00	2.00
gHOLP-GEL	52.06	0.12	43.00	8.00	21.00	2.00
gAR2-GEL	39.62	0.11	44.00	7.00	23.00	4.00

Table 5.1: Grouped variable selection results for GAW17 dataset.

Chapter 6

Implementation in R: grpss Package

6.1 Description

R language has been widely used among statisticians and data miners for developing statistical software, data analysis and visualization. We implement our proposed methods described in Chapter 2 into R language by developing a R package `grpss` which means grouped variable screening and selection. The `grpss` package can be easily accessed and downloaded from R-CRAN website `https://cran.rstudio.com/web/packages/grpss/`. The main functions in this package are described as follows.

- `grp.criValues`: computes values of grouped screening criterion for each group.
- `grpss`: performs grouped variable screening and selection.
- `summary`: summarizes the results of grouped variable screening and selection.
- `predict`: makes a prediction to the fitted penalized regression model.

The most important function `grpss()` implements the two-stage procedure including screening and selection. Certainly it can also perform only the grouped variable screening for the

first stage by setting argument `select = FALSE` without conducting the grouped variable selection. We should point out that at the second stage, the grouped variable selection is accomplished by using the `grpreg` package developed by Patrick Breheny (2015). Thus, the `grpss` package heavily relies on the `grpreg` package that provides functions to conduct many popular grouped variable selection methods at the second stage.

6.2 Usage

In this section, we provide instructions on how to use the functions in `grpss` package, especially on the arguments of four functions described above. The first function `grp.criValues()` is to calculate the values that measure the strength of relationship between each group and response by using formula (2.2), (2.4), (2.5) and (2.6). The second and most important function `grpss()` combines the grouped variable screening and selection through the combination of `grp.criValues()` from `grpss` package and `grpreg()` from `grpreg` package.

```
grp.criValues(X, y, group, criterion = c("gSIS", "gHOLP", "gAR2", "gDC"),
              family = c("gaussian", "binomial", "poisson"),
              scale = c("standardize","normalize", "none"),
              norm = c("L1", "L2", "Linf"))
grpss(X, y, group, threshold = NULL, scale = c("standardize",
      "normalize", "none"), criterion = c("gSIS", "gHOLP", "gAR2", "gDC"),
      family = c("gaussian", "binomial", "poisson"), select = FALSE,
      penalty = c("grSCAD", "grLasso", "grMCP", "gel", "cMCP"),
      cross.validation = FALSE, norm = c("L1", "L2", "Linf"), q = 1,
      perm.seed = 1, nfolds = 10, cv.seed = NULL, parallel = FALSE,
       cl = NULL, cores = NULL, ...)
```

Since `grp.criValues()` is the internal function of `grpss()`, we only introduce the arguments of function `grpss()` as follows.

- `X`: predictors $\mathbf{X}$.
- `y`: response $\mathbf{y}$.
- `group`: the group index for each predictor. Groups labeled 0 or '0' will not participate in the screening procedure and will enter the variable selection directly but without being penalized.
- `threshold`: the threshold d to retain the number of groups at screening procedure. The default `NULL` means a data-driven threshold is determined by random permutation idea described in Section 2.5 of Chapter 2. In case the data-driven threshold is 0, `ncut` will be reset to $\lfloor n/\log(n) \rfloor$.
- `scale`: the type of scaling of predictors $\mathbf{X}$. For example, let x be one of the columns of $\mathbf{X}$. We have

$$standardize = \frac{x - mean(x)}{sd(x)},$$

$$normalize = \frac{x - \min(x)}{\max(x) - \min(x)}.$$

  The default is to standardize each column of $\mathbf{X}$ such that each column has mean 0 and standard deviation 1. Note that gSIS is sensitive to the scale of covariates $\mathbf{X}$, so it is necessary to scale the covariates before conducting screening procedure if the scales of $\mathbf{X}$ are greatly different. The other grouped screening approaches gHOLP, gAR2 and gDC are invariant to the scale of covariates.
- `criterion`: the grouped screening criterion. The default is `gSIS`.
- `family`: a description of the error distribution and link function to be used in the model. The default is `gaussian` which is used for linear regression model.

- `select`: a logical value indicating whether to conduct the grouped variable selection. The default is `FALSE` which means to conduct the grouped variable screening only.

- `penalty`: the penalty to be applied to the screened submodel $\mathcal{M}_d^g$. The default is `grSCAD` with default extra regularization parameter $a = 3.7$.

- `cross.validation`: a logical value indicating whether to perform the k-fold cross-validation when conducting the grouped variable selection. This argument is only valid when setting `select = TRUE`.

- `norm`: the type of norm to incorporate the group structure for gSIS and gHOLP criteria. For example, let $\mathbf{a} = (a_1, ..., a_n)$ be a vector. `"L2"` norm is defined as $\|\mathbf{a}\|_2 = (a_1^2 + \cdots + a_n^2)^{1/2}$ and `"Linf"` norm as $\|\mathbf{a}\|_\infty = \max(a_1, ..., a_n)$. The default is `"L1"` norm, defined as $\|\mathbf{a}\|_1 = |a_1| + \cdots + |a_n|$.

- `q`: a quantile for calculating the data-driven threshold in the permutation-based grouped screening. The default value is 100 percentile (maximum absolute value).

- `perm.seed`: a seed of the random number generator used for the permutation-based screening to obtain the threshold.

- `nfolds`: the number of folds to perform the cross-validation. The default is 10-folds cross-validation which empirically performs rather well in practice.

- `cv.seed`: a seed of the random number generator used for the cross-validation.

- `parallel`: a logical value indicating whether to use the parallel computing. We have to register the parallel backend before using parallel computing. e.g., `library(doParallel)` and then `registerDoParallel(cores = 3)`.

- `cl`: a cluster object as returned by makeCluster, or the number of nodes to be created in the cluster. This is from the argument of `foreach()` function.

- `core`: the number of cores to use for parallel execution. If not specified, the number of core is set to be 3.

- `...`: optional arguments such as `lambda`, `max.iter` and `gamma` passed to `grpreg()` function.

Note that for the case in which `family = "binomial"` and `family = "poisson"`, we calculate the Akaike's Information Criterion (AIC) that marginally measures the correlation between each group $\mathbf{X}_j, j = 1, ..., J$ and response $\mathbf{y}$. AIC characterizes the relationship between grouped variables and response, which is analog to the groupwise adjusted R^2 for linear model at `family = "gaussian"`. Thus, we still use the argument `criterion = "gAR2"` representing the gAIC screening criteria. To improve the computation efficiency, we utilize the parallel computing by setting the argument `parallel = TRUE` and take advantage of the computer cores.

The last two functions `summary()` and `predict()` are the same as usual functions that summarize a fitted model and make predictions based on the fitted model in R language.

```
summary(object, lambda = NULL, digits = 4, ...)
predict(object, newdata, lambda = NULL,
        type = c("response", "class", "probability"), ...)
```

The description of arguments in these two functions are provided as follows.

- `object`: a fitted penalized regression model.

- `lambda`: a regularization parameter at which to summarize or predict.

- `newdata`: a matrix or data frame where to predict. If omits, the fitted predictors are used.

- `type`: the type of prediction: `"response"` gives the fitted values; `"class"` returns the predicted class for the binomial outcome; `"probability"` returns the predicted probabilities for the logistic regression.

- `digits`: number of digits past the decimal point to print out.

The `object` is a fitted penalized regression model obtained from function `grpss()` with arguments `select = TRUE`. It can also be an object that is fitted by function `grpreg()` or `cv.grpreg()` from `grpreg` package. Otherwise, there is no fitted regression model to summarize and predict if we only conduct the grouped variable screening.
In addition, there is a simple function called `importance` which is to arrange and visualize the importance of the groups based on the screening values that are obtained from `grp.criValues()` function.

```
importance(grp.values, n = 10, plot = TRUE)
```

It contains the following arguments:

- `grp.values`: a fitted result from `grp.criValues()` function.

- `n`: the number of top n important groups to display. The default is `10`.

- `plot`: a logical value indicating whether to visualize the importance of top n groups. The default is `TRUE`. A bar plot will be created to show the importance.

6.3 Examples

In this section, we provide several simple examples to demonstrate the usage of functions in `grpss` package. We simply generate three different datasets whose grouped variables are from multivariate normal distribution with different covariance matrices. To be specific, we

let the sample size $n = 100$ and group size $J = 300$. For simplicity, the number of variables within each group is set to be equal, i.e., $p_1 = p_2 = \cdots = p_J = 4$, leading to the total number of predictors $p = 4J = 1200$. The grouped covariates $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{0}$ is a $p \times 1$ mean vector and $\mathbf{\Sigma}$ is a $p \times p$ covariance matrix. We consider three different settings of covariance matrix: (i) Independent: $\mathbf{\Sigma} = \boldsymbol{I}_p$; (ii) Serial correlated: $\mathbf{\Sigma} = \{\sigma_{ij}\}_{i,j=1}^{p} = \{0.6^{|i-j|}\}_{i,j=1}^{p}$; (iii) Compound symmetric: $\mathbf{\Sigma} = \{\sigma_{ij}\}_{i,j=1}^{p} = 0.6$ at $i \neq j$ and 1 at $i = j$. The first three groups are set to be truly correlated with response. That is, the coefficients of the first three groups are nonzero and generated from uniform distribution $U(-2, 5)$. The R code to generate the datasets is given as follows.

```
> set.seed(123)
> n <- 100  # sample size
> p <- 4    # number of predictors within a group
> J <- 300  # entire group size
> group <- rep(1:J,each = p)  # group indices for each predictor
> (betaTrue <- runif(12,-2,5)) # nonzero coefficients
 [1]  0.01304264  3.51813595  0.86283845  4.18112183  4.58327099 -1.68110450
 [7]  1.69673842  4.24693331 1.86004510  1.19630315  4.69783342  1.17333909
> # Case 1: independent predictors
> Sigma1 <- diag(p*J)
> X1 <- MASS:::mvrnorm(n,seq(0,5,length.out = p*J),Sigma1)
> y1 <- X1%*%matrix(c(betaTrue,rep(0,p*J-12)),ncol = 1) + rnorm(n)
>
> # Case 2: serial correlation
> Sigma2 <- 0.6^abs(matrix(1:(p*J),p*J,p*J) - t(matrix(1:(p*J),p*J,p*J)))
> X2 <- MASS:::mvrnorm(n,seq(0,5,length.out = p*J),Sigma2)
```

```
> y2 <- X2%*%matrix(c(betaTrue,rep(0,p*J-12)),ncol = 1) + rnorm(n)
> # Case 3: compund symmetric
> Sigma3 <- matrix(0.6,p*J,p*J)
> diag(Sigma3) <- 1
> X3 <- MASS:::mvrnorm(n,seq(0,5,length.out = p*J),Sigma3)
> y3 <- X3%*%matrix(c(betaTrue,rep(0,p*J-12)),ncol = 1) + rnorm(n)
```

After generating the datasets, we first conduct the grouped variable screening. Here we use the default `threshold = NULL`, but we can also supply the threshold manually, i.e., `threshold = length(y)` or `threshold = floor(length(y)/log(length(y)))`.

```
> # Conduct grouped variable screening
> (gss01 <- grpss(X1,y1,group)) # gSIS for case 1
Call:
grpss.default(X = X1, y = y1, group = group)


Criterion: group SIS
Threshold (ncut):  5
Screened groups:  1 2 3 154 255
> (gss02 <- grpss(X2,y2,group, criterion = "gHOLP"))  # gHOLP for case 2
Call:
grpss.default(X = X2, y = y2, group = group, criterion = "gHOLP")


Criterion: group HOLP
Threshold (ncut):  3
Screened groups:  1 2 3
```

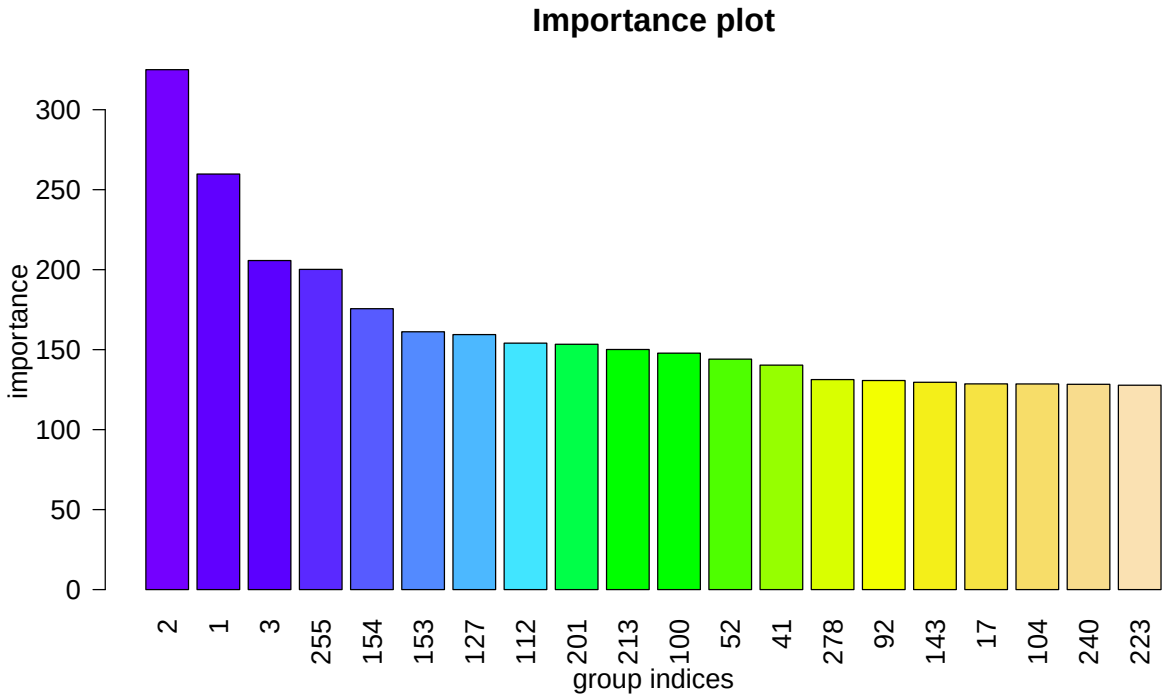
```
 > (gss03 <- grpss(X3,y3,group, criterion = "gAR2")) # gAR2 for case 3
Call:
grpss.default(X = X3, y = y3, group = group, criterion = "gAR2")


Criterion: group AR2
Threshold (ncut):  21
Screened groups:  1 2 3 10 20 37 43 67 73 93 150 167 187 188 193 213 224
 249 253 279 280
```

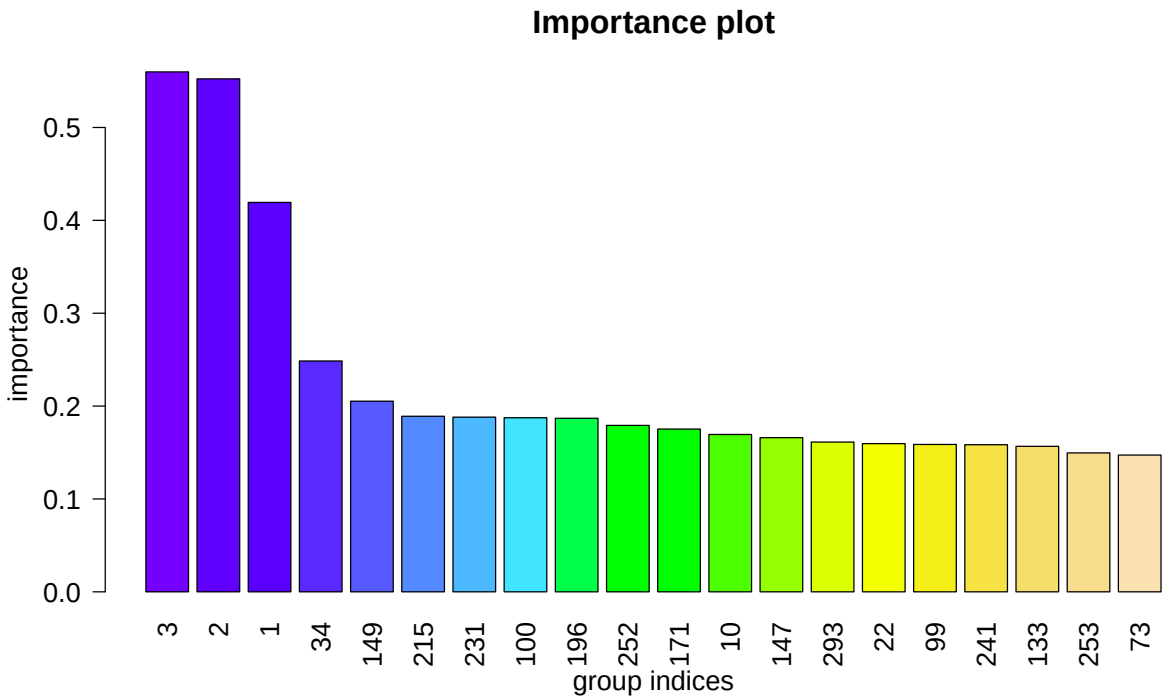
We can also use the `importance` function to visualize the importance of the top `n = 20` groups, though the importance of groups is not a critical issue for the grouped variable screening. The top 20 importance of groups are shown in Figure 6.1. It is obvious that the first three groups 1,2,3 are ranked in top 3 by both grouped screening methods gSIS and gHOLP.

```
> # gSIS for case 1
> grp.valSIS <- grp.criValues(X1,y1,group)
> imp.sis <- importance(grp.valSIS, n = 20) # top 20 important groups
> # gHOLP for case 2
> grp.valHOLP <- grp.criValues(X2,y2,group,criterion = "gHOLP")
> imp.holp <- importance(grp.valHOLP, n = 20)
```

Now we perform both grouped variable screening and selection by setting `select = TRUE`. In this case, we can use the cross-validation to get the optimal regularization parameter λ at which we obtain the minimum cross-validation error. We can see from the results that all methods exactly select the true groups labeled as 1,2,3.



(a) The top 20 important groups by gSIS for case 1 of independent covariance matrix.



(b) The top 20 important groups by gHOLP for case 2 of serial correlated covariance matrix.

Figure 6.1: The top 20 important groups screening by gSIS and gHOLP using `importance()` function in `grpss` package.

```
> # Perform grouped variable screening and selection
> gss11 = grpss(X1,y1,group,select = T,cross.validation = T)
> summary(gss11)
Call:
grpss.default(X = X1, y = y1, group = group, select = T,
              cross.validation = T)

Nonzero coefficients:
(Intercept)      X1          X1          X1          X1          X2
0.05033     0.11295     3.52648     0.86507     4.16351     4.58345
     X2          X2          X2          X3          X3          X3          X3
-1.60723     1.62135     4.26652   1.81505     1.19560     4.71769     1.06263

R-squared:  0.991268 ; Scale estimate (sigma):  0.9352904
Signal-to-noise ratio:  113.5215
----------------------------------
Group SCAD-penalized linear regression with group SIS screening
Optimal model obtained at lambda =  0.1266245
Minimum cross-validation error: 1.179545
>
> gss12 = grpss(X2,y2,group,select = T,cross.validation = T,
                criterion = "gHOLP")
> summary(gss12)
Call:
grpss.default(X = X2, y = y2, group = group, criterion = "gHOLP",
select = T, cross.validation = T)
```

```
Nonzero coefficients:
(Intercept)     X1          X1          X1          X1          X2
-0.05424     0.08309     3.41002     0.98347     3.92783     4.62001
   X2          X2          X2          X3          X3          X3          X3
 -1.47442    1.59385    4.22682     1.92204     1.04885     4.94081     1.19385

R-squared:  0.9945333 ; Scale estimate (sigma):  1.026363
Signal-to-noise ratio:  181.926
----------------------------------
Group SCAD-penalized linear regression with group HOLP screening
Optimal model obtained at lambda =  0.7230969
Minimum cross-validation error: 1.397396
>
> gss13 = grpss(X3,y3,group,select = T,cross.validation = T,
                criterion = "gAR2")
> summary(gss13)
Call:
grpss.default(X = X3, y = y3, group = group, criterion = "gAR2",
select = T, cross.validation = T)

Nonzero coefficients:
(Intercept)     X1          X1          X1          X1          X2
0.06146    -0.03366     3.63686     0.51309     3.67957     4.43205
     X2          X2          X2          X3          X3          X3          X3
-1.59108    1.67209    4.51899    2.09576     1.63608     4.75015     1.17623
```

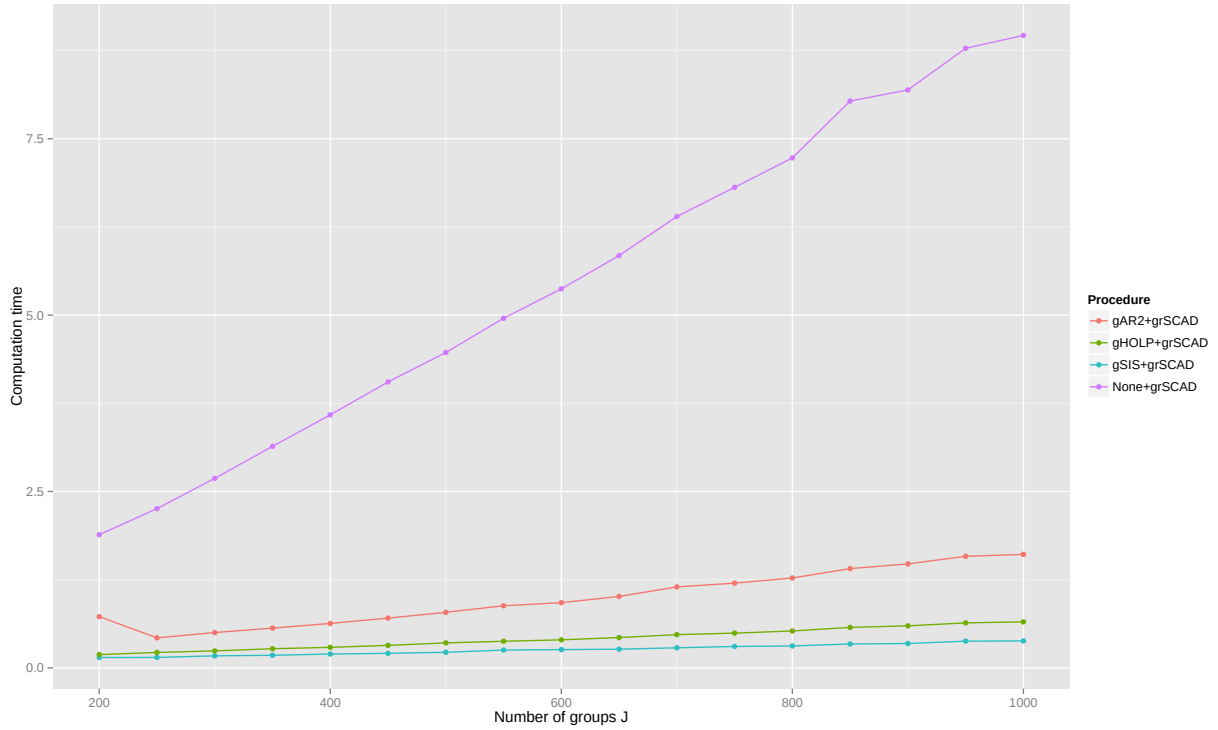
```
R-squared:  0.9979041 ; Scale estimate (sigma):  0.9634959
Signal-to-noise ratio:  476.1455
----------------------------------
Group SCAD-penalized linear regression with group AR2 screening
Optimal model obtained at lambda =  0.6676974
Minimum cross-validation error: 1.246246
```

Finally, we can make predictions based on the fitted model by `grpss()` function. The `predict()` function in `grpss` package is similar to `predict()` function provided in `grpreg` package, but the former can omit the argument `newdata` to get the fitted values.

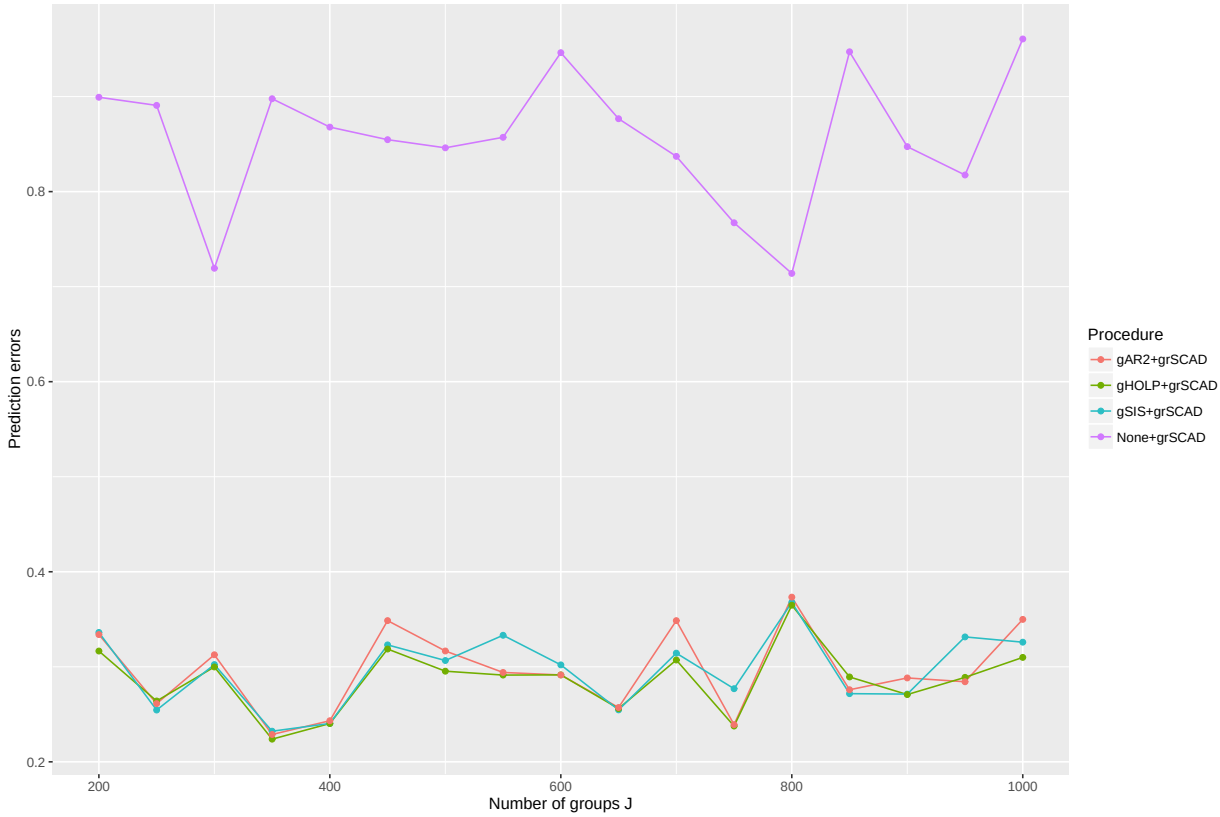
```
> # make predictions
> predict(gss11)[1:10] # fitted values, only print out first 10 values
[1]  -4.296464  22.777814  -9.332836 -14.633269  -3.219178  -3.013977
[7]  -22.712330 -1.242792  -4.437115  1.607212
> predict(gss12, newdata = X2[1:2,]) # predict the values at X2[1:2,]
response
[1,] -17.033098
[2,]  -6.748028
```

We now give a toy example to compare the performances of `grpss` with argument `select = TRUE` and `grpreg` packages in terms of the computation efficiency and estimation accuracy. In Section 4.2 of Chapter 4, we compared the performances of two-stage procedure and one-stage procedure with fixed group size, but now we compare their performances with different group sizes. In greater details, we execute the function `grpss()` and `grpreg()` respectively for the datasets generated from the first two cases of covariance matrix with fixed sample size $n = 100$ but increasing group size from 200 to 1000 with step 50, i.e.,

$J = 200, 250, 300, ..., 1000$. We calculate the average computation time and prediction errors based on 20 simulations. We exclude the gDC screening due to the intensive computation for large datasets. Figure 6.2 and 6.3 shows that the computation time is increasing rapidly as the number of groups is increasing for the selection results obtained from `grpreg()` function. However, with the extra screening procedure, the computation time increases slowly enough without a sharp trend. On average, the prediction errors are also smaller when we apply the extra screening procedure using `grpss()` function. In this toy example, we do not use the parallel computing, otherwise computation time will definitely be much faster if we set `parallel = TRUE` and use more cores of computers.

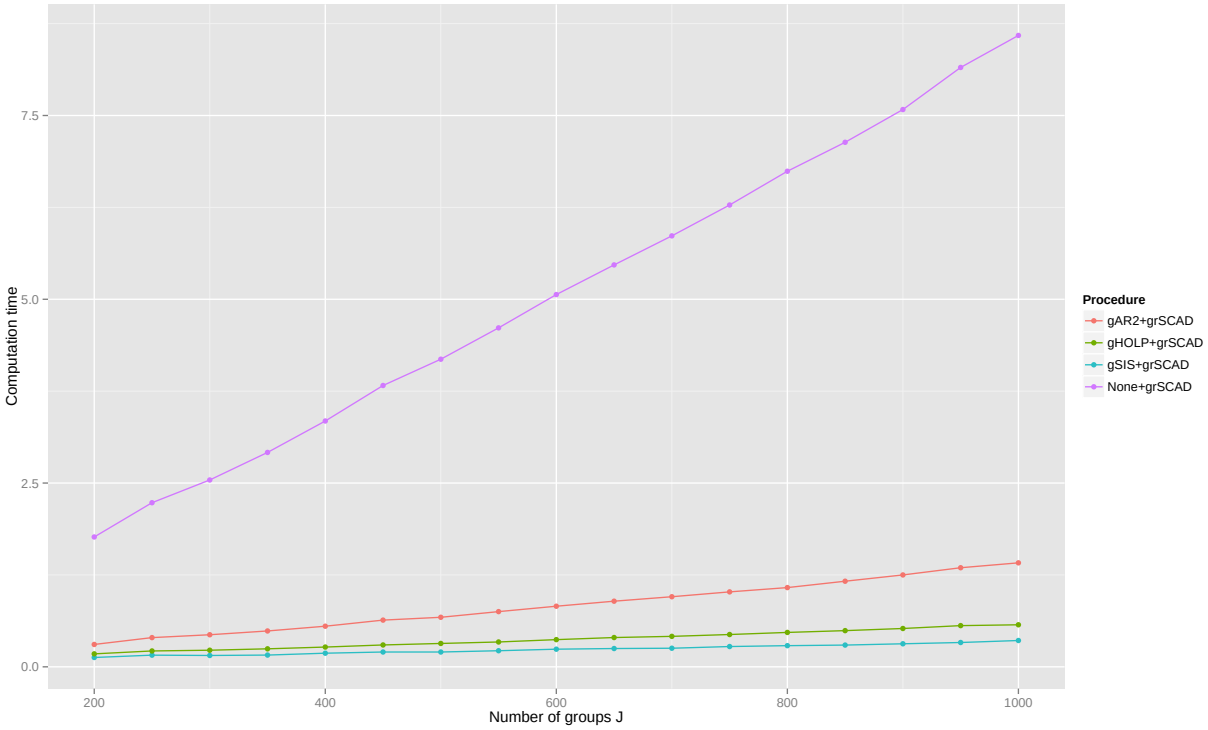


(a) Computation time

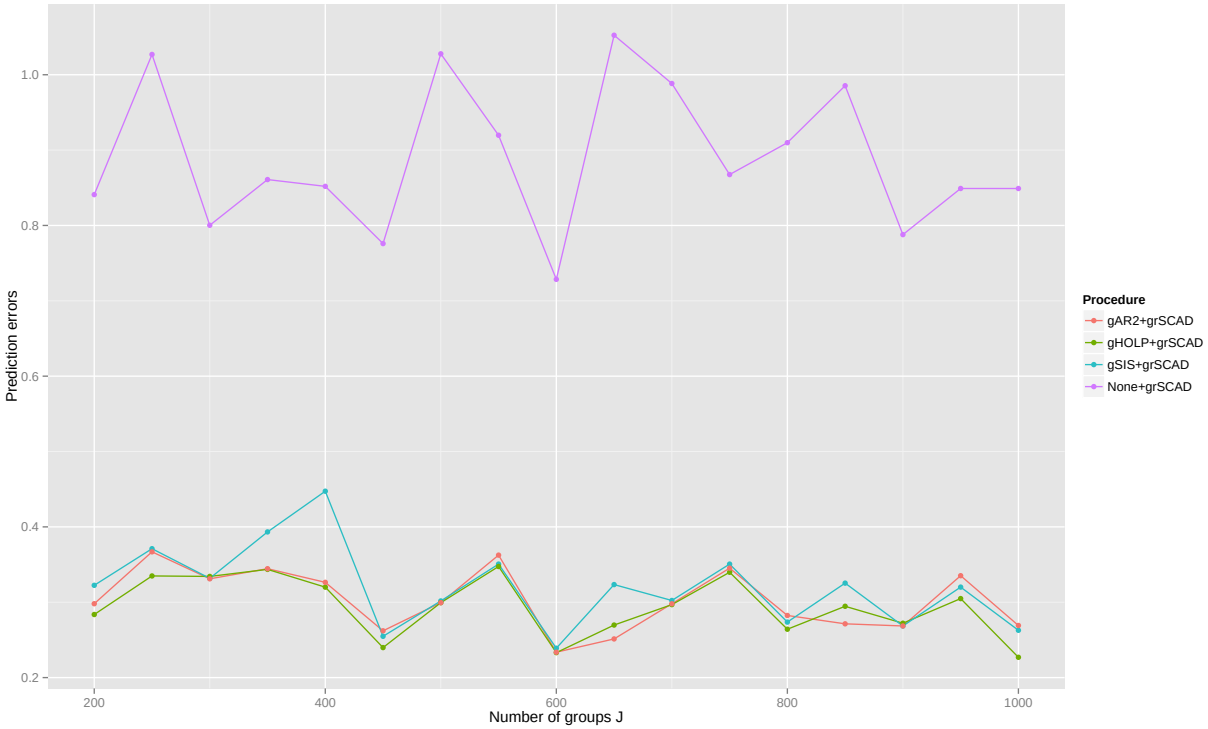


(b) Prediction errors

Figure 6.2: Comparison of `grpss` and `grpreg` packages for the dataset generated from case 1 of independent groups.



(a) Computation time



(b) Prediction errors

Figure 6.3: Comparison of `grpss` and `grpreg` packages for the dataset generated from case 2 of moderate serial correlation predictors.

Appendix

Recall that the model we work with is of the form

$$Y = \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j + \varepsilon = \sum_{j=1}^{J} \sum_{k=1}^{p_j} X_{jk} \beta_{jk} + \varepsilon,$$

where $\mathbf{X}_j = \left(X_{j1}, \cdots, X_{jp_j}\right), \boldsymbol{\beta}_j = \left(\beta_{j1}, \cdots, \beta_{jp_j}\right)^T$ are covariates and coefficients for the j-th group, respectively.

We denote $\mathcal{M}_S^g = \left\{\mathbf{X}_j : \left\|\boldsymbol{\beta}_j\right\|_1 \neq 0, 1 \leq j \leq J\right\}$ to be the true sparse model with non-sparsity size $s = |\mathcal{M}_S|$ and also define

$$\begin{aligned} \mathbf{z} &= \Sigma^{-1/2}\mathbf{x}, \\ \mathbf{Z} &= \mathbf{X}\Sigma^{-1/2}, \end{aligned}$$

where $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_J)^T$ and $\boldsymbol{\Sigma} = \operatorname{cov}(\mathbf{x})$, $\mathbf{X} = (\mathbf{X}_1, ..., \mathbf{X}_J)^T$. Clearly, the n rows of the transformed designed matrix $\mathbf{Z}$ are IID copies of $\mathbf{z}$ which now has covariance matrix I_p. Also, we define the total number of individual covariates $p = \sum_{j=1}^{J} p_j$. For simplicity, all the predictors $X_{11}, \cdots, X_{1p_1}, \cdots, X_{J1}, \cdots, X_{Jp_J}$ are standardized such that they have mean 0 and standard deviation 1. This indicates the covariance matrix becomes the correlation matrix, i.e., $\boldsymbol{\Sigma}_{ii} = 1, i = 1, \cdots, p$.

A Proof of Theorem 2.1

As gSIS is based on the SIS, the screening property of gSIS relies on the screening property of SIS. However, we provide a different framework of proofs to show the screening property of SIS, and then proceed to show the screening property of gSIS. To achieve this, we first need two lemmas provided in SIS (Fan and Lv (2008)). Note that the singular value decomposition of $\mathbf{Z}$ is $\mathbf{Z} = \mathbf{V}\mathbf{D}_1\mathbf{U}$, where $\mathbf{V} \in \mathcal{O}(n), \mathbf{U} \in \mathcal{O}(p)$ and $\mathbf{D}_1$ is an $n \times p$ diagonal matrix whose diagonal elements are $\mu_1^{1/2}, \cdots, \mu_n^{1/2}$, in which $\mathcal{O}(n)$ is defined as the orthogonal space with dimension n. Let $\mathbf{S} = \left(\mathbf{Z}^T\mathbf{Z}\right)^- \mathbf{Z}^T\mathbf{Z}$, $\tilde{\mathbf{U}} = (I_n, 0)_{n\times p}\mathbf{U}$, where $\left(\mathbf{Z}^T\mathbf{Z}\right)^-$ denotes the Moore-Penrose generalized inverse of $\mathbf{Z}^T\mathbf{Z}$. By simple linear algebra, we can easily obtain $\mathbf{S} = \tilde{\mathbf{U}}^T\tilde{\mathbf{U}}$. Lemma A.1 and A.2 describe the distribution of $\mathbf{S}$, which only requires the assumptions that $\mathbf{Z}$ has a spherical symmetric distribution and the dimension p is larger than the sample size n. Let $\mathbf{e}_{jk}^T$ be a column vector with the jk-th elements 1 and others 0, i.e., $\mathbf{e}_{jk}^T = (0, \cdots, 1, \cdots, 0)_{p\times 1}$. For ease of notation, we use $\mathbf{e}_1^T = \mathbf{e}_{11}^T, \mathbf{e}_2^T = \mathbf{e}_{12}^T$ without any ambiguity in the subsequent section.

Lemma A.1 (*Lemma 4, Fan and Lv (2008)***):** *For any $C > 0$, there is some constant $c_1 > 1$ such that*

$$\Pr\left(\langle \mathbf{S}\mathbf{e}_1, \mathbf{e}_1\rangle < c_1^{-1}\frac{n}{p} \text{ or } \langle \mathbf{S}\mathbf{e}_1, \mathbf{e}_1\rangle > c_1\frac{n}{p}\right) \leq 4\exp(-Cn).$$

Lemma A.2 (*Lemma 5, Fan and Lv (2008)***):** *Let $\mathbf{S}\mathbf{e}_1 = (V_1, V_2, \cdots, V_p)^T$. Then, given that the first co-ordinate $V_1 = v$, the random vector $(V_2, \cdots, V_p)^T$ is uniformly distributed on the sphere $S^{p-2}\left\{\sqrt{v - v^2}\right\}$. Moreover, for any $C > 0$, there is some $c > 1$ such that*

$$\Pr\left(|V_i| > cn^{1/2}p^{-1}|W|\right) \leq 3\exp(-Cn),$$

where W is an independent $\mathcal{N}(0,1)$-distributed random variable.

Note that the design matrix $\mathbf{X}$ can be transformed into $\mathbf{Z}\mathbf{\Sigma}^{1/2}$. Thus, we make assumptions

on $\mathbf{Z}$ and $\mathbf{\Sigma}$ separately.

Assumptions:

(A1) The transformed z has a spherically symmetric distribution and random matrix $\mathbf{Z}$ has the concentration property such that

$$\Pr\left\{\lambda_{\min}\left(p^{-1}\mathbf{Z}\mathbf{Z}^T\right) < 1/c_1 \text{ or } \lambda_{\max}\left(p^{-1}\mathbf{Z}\mathbf{Z}^T\right) > c_1\right\} \leq \exp\left(-Cn\right)$$

for some $c_1 > 1$ and $C > 0$.

(A2) The random error ε has a normal distribution with mean 0 and standard deviation σ, i.e., $\varepsilon \sim \mathcal{N}\left(0, \sigma^2\right)$, where σ is a constant.

(A3) We assume that $\text{Var}\left(Y\right) = O\left(1\right)$ and

$$\min_{j \in S} \sum\nolimits_{k=1}^{p_j} \left|\beta_{jk}\right| \geq \frac{c_2}{n^{\kappa}}, \lambda_{\min}\left(\mathbf{\Sigma}\right) \geq c_3 \text{ and } \lambda_{\max}\left(\mathbf{\Sigma}\right) \leq c_4 n^{\tau}$$

for some $\kappa \geq 0, \nu \geq 0, \tau \geq 0$ and $c_2, c_4 > 0$, $0 < c_3 < 1$.

(A4) Assume that the number of true important groups $s \leq c_5 n^{\nu}, c_5 > 0$ and the j-th group size $p_j = c_6 n^{\gamma}$ for some constant $c_6 > 0$. Also, assume that $\log\left(J\right) = O\left(n^{\delta}\right)$, $\delta \in (0, 1 - 2\kappa - 2\tau - \nu - \gamma)$, where $\kappa, \tau, \nu, \gamma$ are parameters defined as above.

Assumptions (A1) - (A4) are very similar to those of SIS in Fan an Lv (2008). The key difference is that in Assumption (A3), we make an assumption on the grouped coefficients $\boldsymbol{\beta}_j$ for the important groups, rather than on the individual coefficients $\beta_{jk}, j = 1, ..., J, k = 1, ..., p_j$. Also, note that SIS would fail completely for the case where the predictors are marginally uncorrelated but jointly correlated with Y. That is, Fan and Lv (2008) ruled out this situation by imposing a constriant on the marginal correlation between important variables and response, i.e., $\text{cov}\left(X_{jk}, Y\right) \geq c_2 c_3 / n^{\kappa}$. To rule out the same situation in gSIS,

Assumptions (A3) and (A4) can lead to the similar condition. To see this, we can observe that

$$\begin{aligned}\min_{j\in S}\sum_{k=1}^{p_j}\left|\mathrm{cov}\left(X_{jk},Y\right)\right| &= \min_{j\in S}\sum_{k=1}^{p_j}\left|\mathrm{cov}\left(\mathbf{X}\mathbf{e}_{jk},Y\right)\right| = \min_{j\in S}\sum_{k=1}^{p_j}\left|\mathrm{cov}\left(\mathbf{X}\mathbf{e}_{jk},\mathbf{X}\boldsymbol{\beta}\right)\right| \\ &= \min_{j\in S}\sum_{k=1}^{p_j}\left|\mathbf{e}_{jk}^T\boldsymbol{\Sigma}\boldsymbol{\beta}\right| \geq \min_{j\in S}\sum_{k=1}^{p_j}\left|\mathbf{e}_{jk}^T\lambda_{\min}\left(\boldsymbol{\Sigma}\right)I_p\boldsymbol{\beta}\right| \\ &\geq c_3\min_{j\in S}\sum_{k=1}^{p_j}\left|\beta_{jk}\right| \geq \frac{c_2c_3}{n^{\kappa}}.\end{aligned}$$

Since

$$\mathbf{Z}^T\mathbf{Z} = \mathbf{U}^T\,\mathrm{diag}\left(\mu_1,...\mu_n,...,0\right)\mathbf{U},$$

we can obtain

$$\mathbf{X}^T\mathbf{X} = p\boldsymbol{\Sigma}^{1/2}\tilde{\mathbf{U}}^T\,\mathrm{diag}\left(\mu_1,\cdots,\mu_n\right)\tilde{\mathbf{U}}\boldsymbol{\Sigma}^{1/2},$$

where $\mu_1,...,\mu_n$ are n eigenvalues of $p^{-1}\mathbf{Z}\mathbf{Z}^T$. Recall that the individual SIS is to compute

$$\boldsymbol{\omega} = \mathbf{X}^TY = \mathbf{X}^T\left(\mathbf{X}\boldsymbol{\beta}+\varepsilon\right) = \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}+\mathbf{X}^T\varepsilon := \boldsymbol{\xi}+\boldsymbol{\eta},$$

where $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are signal and noise part, respectively.

For the signal part of the j-th group, $\boldsymbol{\xi}_j = \left(\xi_{j1},\cdots,\xi_{jp_j}\right)$, $j=1,...,J$, where $\xi_{jk} = \mathbf{e}_{jk}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$. We first bound the diagonal and off-diagonal elements of $\mathbf{X}^T\mathbf{X}$. For the diagonal components, using the fact that $\boldsymbol{\Sigma}^{1/2}\mathbf{e}_{jk} = \mathbf{Q}\mathbf{e}_1$ for some $\mathbf{Q}\in\mathcal{O}\left(p\right)$ and $\mathbf{S}\mathbf{Q}\overset{(d)}{=}\mathbf{S}$, we have

$$\begin{aligned}\mathbf{e}_{jk}^T\mathbf{X}^T\mathbf{X}\mathbf{e}_{jk} &= p\mathbf{e}_{jk}^T\boldsymbol{\Sigma}^{1/2}\tilde{\mathbf{U}}^T\,\mathrm{diag}\left(\mu_1,\cdots,\mu_n\right)\tilde{\mathbf{U}}\boldsymbol{\Sigma}^{1/2}\mathbf{e}_{jk} \\ &= p\mathbf{e}_1^T\mathbf{Q}^T\tilde{\mathbf{U}}^T\,\mathrm{diag}\left(\mu_1,\cdots,\mu_n\right)\tilde{\mathbf{U}}\mathbf{Q}\mathbf{e}_1 \\ &\geq p\mathbf{e}_1^TQ^T\tilde{\mathbf{U}}^T\lambda_{\min}\left(p^{-1}\mathbf{Z}\mathbf{Z}^T\right)I_n\tilde{\mathbf{U}}Q\mathbf{e}_1 \\ &\overset{(d)}{=} p\lambda_{\min}\left(p^{-1}\mathbf{Z}\mathbf{Z}^T\right)\left\langle\mathbf{S}\mathbf{e}_1,\mathbf{e}_1\right\rangle,\end{aligned}$$

which implies for some $c' < 1$ with $c' = c'_1 \cdot c'_2$, $c'_1 < 1, c'_2 < 1$,

$$\begin{aligned} \Pr\left(\left|\mathbf{e}_{jk}^T \mathbf{X}^T \mathbf{X} \mathbf{e}_{jk}\right| < c'n\right) &\leq \Pr\left(\lambda_{\min}\left(p^{-1}\mathbf{Z}\mathbf{Z}^T\right)\langle \mathbf{S}\mathbf{e}_1, \mathbf{e}_1\rangle < c'_1 c'_2 \frac{n}{p}\right) \\ &\leq \Pr\left(\lambda_{\min}\left(p^{-1}\mathbf{Z}\mathbf{Z}^T\right) < c'_1\right) + \Pr\left(\langle \mathbf{S}\mathbf{e}_1, \mathbf{e}_1\rangle < c'_2 \frac{n}{p}\right) \\ &\leq \exp\left(-Cn\right). \end{aligned} \tag{A.1}$$

Now for the off-diagonal components, without loss of generality, we only consider $\mathbf{e}_1^T \mathbf{X}^T \mathbf{X} \mathbf{e}_2$ that can be easily extended to other cases. That is, we have

$$\begin{aligned} \mathbf{e}_1^T \mathbf{X}^T \mathbf{X} \mathbf{e}_2 &= p \mathbf{e}_1^T \Sigma^{1/2} \tilde{\mathbf{U}}^T \operatorname{diag}\left(\mu_1, \cdots, \mu_n\right) \tilde{\mathbf{U}} \Sigma^{1/2} \mathbf{e}_2 \\ &\leq p \mathbf{e}_1^T \lambda_{\max}^{1/2}\left(\mathbf{\Sigma}\right) I_p \tilde{\mathbf{U}}^T \lambda_{\max}\left(p^{-1}\mathbf{Z}\mathbf{Z}^T\right) I_n \tilde{\mathbf{U}} \lambda_{\max}^{1/2}\left(\mathbf{\Sigma}\right) I_p \mathbf{e}_2 \\ &\leq p c_4 n^{\tau} \lambda_{\max}\left(p^{-1}\mathbf{Z}\mathbf{Z}^T\right)\langle \mathbf{S}\mathbf{e}_1, \mathbf{e}_2\rangle. \end{aligned}$$

Therefore, by using Lemma A.2 and observing $\langle \mathbf{S}\mathbf{e}_1, \mathbf{e}_2\rangle = V_2$, we can obtain

$$\begin{aligned} \Pr\left(\left|\mathbf{e}_1^T \mathbf{X}^T \mathbf{X} \mathbf{e}_2\right| > c_1 c_4 n^{\tau} c n^{1/2} |W|\right) &\leq \Pr\left(p c_4 n^{\tau} \lambda_{\max}\left(p^{-1}\mathbf{Z}\mathbf{Z}^T\right)\langle \mathbf{S}\mathbf{e}_1, \mathbf{e}_2\rangle > c_1 c_4 n^{\tau} c n^{1/2} |W|\right) \\ &\leq \Pr\left(\lambda_{\max}\left(p^{-1}\mathbf{Z}\mathbf{Z}^T\right) > c_1\right) + \Pr\left(V_2 > c n^{1/2} p^{-1} |W|\right) \\ &\leq O\left\{\exp\left(-Cn\right)\right\}. \end{aligned}$$

Taking $x_n = \sqrt{2} c_1 c_4 c n^{1-\kappa-\nu/2-\gamma/2} / \sqrt{\log(n)}$, we have

$$\begin{aligned} \Pr\left(\left|\mathbf{e}_1^T \mathbf{X}^T \mathbf{X} \mathbf{e}_2\right| > x_n\right) &= \Pr\left(\left|\mathbf{e}_1^T \mathbf{X}^T \mathbf{X} \mathbf{e}_2\right| \times c_1 c_4 n^{\tau} c n^{1/2} |W| > x_n \times c_1 c_4 n^{\tau} c n^{1/2} |W|\right) \\ &\leq \Pr\left(\left|\mathbf{e}_1^T \mathbf{X}^T \mathbf{X} \mathbf{e}_2\right| > c_1 c_4 n^{\tau} c n^{1/2} |W|\right) + \Pr\left(c_1 c_4 n^{\tau} c n^{1/2} |W| > x_n\right) \\ &\leq O\left\{\exp\left(-Cn\right)\right\} + \Pr\left(c_1 c_4 n^{\tau} c n^{1/2} |W| > x_n\right). \end{aligned}$$

For the second term, letting $c_1 c_4 c = C_2$, we have by Gaussian tail bound inequality,

$$\begin{aligned}
\Pr\left(c_1 c_4 n^{\tau} c n^{1/2} |W| > x_n\right) &= \Pr\left\{|W| > x_n / \left(M n^{1/2+\tau}\right)\right\} \\
&\leq \exp\left\{-\frac{x_n^2}{2C_2^2 n^{1+2\tau}}\right\}\left(\frac{\sqrt{2\pi} x_n}{C_2 n^{1/2+\tau}}\right)^{-1} \\
&= \exp\left\{-C\frac{n^{1-2\kappa-2\tau-\nu-\gamma}}{\log(n)}\right\}\left(\frac{2\sqrt{\pi} n^{1/2-\kappa-\tau-\nu/2-\gamma/2}}{\sqrt{\log(n)}}\right)^{-1} \\
&\leq \exp\left\{-C\frac{n^{1-2\kappa-2\tau-\nu-\gamma}}{\log(n)}\right\}.
\end{aligned}$$

Thus, we obtain the bound for the off-diagonal components of $\mathbf{X}^T\mathbf{X}$, $jk \neq j'k'$ for $j, j' = 1, ..., J, k, k' = 1, ..., p_j$,

$$\Pr\left(\left|\mathbf{e}_{jk}^T \mathbf{X}^T \mathbf{X} \mathbf{e}_{j'k'}\right| > C_2 \frac{n^{1-\kappa-\nu/2-\gamma/2}}{\sqrt{\log(n)}}\right) \leq O\left\{\exp\left(-C\frac{n^{1-2\kappa-2\tau-\nu-\gamma}}{\sqrt{\log(n)}}\right)\right\}.$$

Next, we bound the signal part of the j-th group, $p_j^{-1}\left\|\boldsymbol{\xi}_j\right\|_1 = p_j^{-1}\sum_{k=1}^{p_j}\left|\mathbf{e}_{jk}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\right|$ for $j \notin S$ and $j \in S$, respectively. For $j \notin S$, using Cauchy Schwartz inequality, we have

$$\begin{aligned}
\left\|\boldsymbol{\xi}_j\right\|_1 &= \sum_{k=1}^{p_j}\left|\mathbf{e}_{jk}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\right| = \sum_{k=1}^{p_j}\left|\mathbf{e}_{jk}\mathbf{X}^T\mathbf{X}\sum_{j'\in S}\sum_{k'=1}^{p_j}\mathbf{e}_{j'k'}\beta_{j'k'}\right| \\
&= \sum_{k=1}^{p_j}\left|\sum_{j'\in S}\sum_{k'=1}^{p_j}\mathbf{e}_{jk}\mathbf{X}^T\mathbf{X}\mathbf{e}_{j'k'}\beta_{j'k'}\right| \leq \sum_{k=1}^{p_j}\sum_{j'\in S}\sum_{k'=1}^{p_j}\left|\mathbf{e}_{jk}\mathbf{X}^T\mathbf{X}\mathbf{e}_{j'k'}\right|\left|\beta_{j'k'}\right| \\
&\leq \sum_{k=1}^{p_j}\sqrt{\sum_{j'\in S}\sum_{k'=1}^{p_j}\left|\mathbf{e}_{jk}\mathbf{X}^T\mathbf{X}\mathbf{e}_{j'k'}\right|^2}\sqrt{\sum_{j'\in S}\sum_{k'=1}^{p_j}\left|\beta_{j'k'}\right|^2} \\
&\leq \sum_{k=1}^{p_j}\sqrt{\sum_{j'\in S}\sum_{k'=1}^{p_j}\left|\mathbf{e}_{jk}\mathbf{X}^T\mathbf{X}\mathbf{e}_{j'k'}\right|^2}\left\|\boldsymbol{\beta}\right\|_2.
\end{aligned}$$

Note that $c_3\left\|\boldsymbol{\beta}\right\|_2^2 \leq \lambda_{\min}(\boldsymbol{\Sigma})\left\|\boldsymbol{\beta}\right\|_2^2 \leq \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} = \mathrm{Var}(Y) - \sigma^2 < c'$ for some constant c',

resulting in $\|\boldsymbol{\beta}\|_2^2 \leq c'/c_3$. It follows that

$$\begin{aligned}
& \Pr\left\{p_j^{-1}\left\|\boldsymbol{\xi}_j\right\|_1 > \frac{C_2c'\sqrt{c_5c_6}n^{1-\kappa}}{c_3\sqrt{\log(n)}}\right\} \qquad (A.2)\\
= \;& \Pr\left\{p_j^{-1}\sum_{k=1}^{p_j}\sqrt{\sum_{j'\in S}\sum_{k'=1}^{p_j}\left|\mathbf{e}_{jk}^T\mathbf{X}^T\mathbf{X}\mathbf{e}_{j'k'}\right|^2}\|\boldsymbol{\beta}\|_2 > \frac{C_2c'\sqrt{c_5c_6}n^{1-\kappa}}{c_3\sqrt{\log(n)}}\right\}\\
\leq \;& \Pr\left\{p_j^{-1}\sum_{k=1}^{p_j}\sqrt{\sum_{j'\in S}\sum_{k'=1}^{p_j}\left|\mathbf{e}_{jk}^T\mathbf{X}^T\mathbf{X}\mathbf{e}_{j'k'}\right|^2} > \frac{C_2\sqrt{c_5c_6}n^{1-\kappa}}{\sqrt{\log(n)}}\right\}\\
\leq \;& c_5c_6n^{\gamma+\nu}\cdot\Pr\left\{\left|\mathbf{e}_{jk}^T\mathbf{X}^T\mathbf{X}\mathbf{e}_{j'k'}\right| > C_2\frac{n^{1-\kappa-\nu/2-\gamma/2}}{\sqrt{\log(n)}}\right\} \leq O\left\{\exp\left(-C\frac{n^{1-2\kappa-2\tau-\nu-\gamma}}{\log(n)}\right)\right\}.
\end{aligned}$$

For $j \in S$, we have

$$\begin{aligned}
\left\|\boldsymbol{\xi}_j\right\|_1 &= \sum_{k=1}^{p_j}\left|\mathbf{e}_{jk}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\right| = \sum_{k=1}^{p_j}\left|\mathbf{e}_{jk}^T\mathbf{X}^T\mathbf{X}\mathbf{e}_{jk}\beta_{jk}\right| + \sum_{k=1}^{p_j}\left|\mathbf{e}_{jk}^T\mathbf{X}^T\mathbf{X}\sum_{j'\in S}\sum_{k'=1}^{p_j}\mathbf{e}_{j'k'}\beta_{j'k'}\right|\\
&\geq \sum_{k=1}^{p_j}\left|\mathbf{e}_{jk}^T\mathbf{X}^T\mathbf{X}\mathbf{e}_{jk}\right|\left|\beta_{jk}\right| - \sum_{k=1}^{p_j}\left|\mathbf{e}_{jk}^T\mathbf{X}^T\mathbf{X}\sum_{j'\in S}\sum_{k'=1}^{p_j}\mathbf{e}_{j'k'}\beta_{j'k'}\right|,
\end{aligned}$$

which implies for some constant c_5 by combining (A.1) and (A.2),

$$\begin{aligned}
& \Pr\left\{p_j^{-1}\left\|\boldsymbol{\xi}_j\right\|_1 < \left(\frac{c_2c'}{n^\gamma} + \frac{C_2c'\sqrt{c_5c_6}}{c_3\sqrt{\log(n)}}\right)n^{1-\kappa}\right\}\\
\leq \;& \Pr\left(p_j^{-1}\sum_{k=1}^{p_j}\left|\mathbf{e}_{jk}^T\mathbf{X}^T\mathbf{X}\mathbf{e}_{jk}\right|\left|\beta_{jk}\right| < c_2c'n^{1-\kappa-\gamma}\right)\\
& + \Pr\left(p_j^{-1}\sum_{k=1}^{p_j}\left|\mathbf{e}_{jk}^T\mathbf{X}^T\mathbf{X}\sum_{j'\in S}\sum_{k'=1}^{p_j}\mathbf{e}_{j'k'}\beta_{j'k'}\right| > \frac{C_2c'\sqrt{c_5c_6}n^{1-\kappa}}{c_3\sqrt{\log(n)}}\right)\\
\leq \;& \Pr\left(\min_k\left|\mathbf{e}_{jk}^T\mathbf{X}^T\mathbf{X}\mathbf{e}_{jk}\right|\sum_{k=1}^{p_j}\left|\beta_{jk}\right| < c_2c'n^{1-\kappa}\right)\\
& + \Pr\left(p_j^{-1}\sum_{k=1}^{p_j}\left|\mathbf{e}_{jk}^T\mathbf{X}^T\mathbf{X}\sum_{j'\in S}\sum_{k'=1}^{p_j}\mathbf{e}_{j'k'}\beta_{j'k'}\right| > \frac{C_2c'\sqrt{c_5c_6}n^{1-\kappa}}{c_3\sqrt{\log(n)}}\right)\\
\leq \;& c_6n^\gamma\Pr\left(\left|\mathbf{e}_{jk}^T\mathbf{X}^T\mathbf{X}\mathbf{e}_{jk}\right| < c'n\right) + O\left\{\exp\left(-C\frac{n^{1-2\kappa-2\tau-\nu-\gamma}}{\log(n)}\right)\right\}\\
\leq \;& O\left\{\exp(-Cn) + \exp\left(-C\frac{n^{1-2\kappa-2\tau-\nu-\gamma}}{\log(n)}\right)\right\} = O\left\{\exp\left(-C\frac{n^{1-2\kappa-2\tau-\nu-\gamma}}{\log(n)}\right)\right\}.
\end{aligned}$$

Now we turn to bound the noise part $p_j^{-1}\left\|\boldsymbol{\eta}_j\right\|_1 = p_j^{-1}\sum_{k=1}^{p_j}\left|\mathbf{e}_{jk}^T\mathbf{X}^T\varepsilon\right|$.

To bound the noise part, $p_j^{-1}\left\|\boldsymbol{\eta}_j\right\|_1$, we first consider $|\eta_{jk}| = \left|\mathbf{e}_{jk}^T\mathbf{X}^T\varepsilon\right|$. The noise η_{jk} can be decomposed as

$$\eta_{jk} = \left\|\mathbf{e}_{jk}^T\mathbf{X}^T\right\|_2 \times \frac{\mathbf{e}_{jk}^T\mathbf{X}^T}{\left\|\mathbf{e}_{jk}^T\mathbf{X}^T\right\|_2} \times \sigma \times \frac{\varepsilon}{\sigma} = \sigma\left\|\mathbf{e}_{jk}^T\mathbf{X}^T\right\|_2 \times \mathbf{a} \times \frac{\varepsilon}{\sigma},$$

where $\mathbf{a} = (a_1, \cdots, a_n) = \mathbf{e}_{jk}^T\mathbf{X}^T/\left\|\mathbf{e}_{jk}^T\mathbf{X}^T\right\|_2$. Since $\mathbf{X}$ is independent of ε, $\mathbf{a}$ is also independent of ε. Define $W = \mathbf{a}\cdot\varepsilon/\sigma$, we have $\eta_{jk} = \sigma\left\|\mathbf{e}_{jk}^T\mathbf{X}^T\right\|_2 \cdot W$. For the norm term, using the fact $\boldsymbol{\Sigma}^{1/2}\mathbf{e}_{jk} = \mathbf{Q}\mathbf{e}_1$ again, we have

$$\begin{aligned}\left\|\mathbf{e}_{jk}^T\mathbf{X}^T\right\|_2^2 &= \mathbf{e}_{jk}^T\mathbf{X}^T\mathbf{X}\mathbf{e}_{jk} = p\mathbf{e}_{jk}^T\boldsymbol{\Sigma}^{1/2}\tilde{\mathbf{U}}^T diag\left(\mu_1, \cdots, \mu_n\right)\tilde{\mathbf{U}}\boldsymbol{\Sigma}^{1/2}\mathbf{e}_{jk} \\ &\leq p\mathbf{e}_1^T\mathbf{Q}^T\tilde{\mathbf{U}}^T\lambda_{\max}\left(p^{-1}\mathbf{Z}\mathbf{Z}^T\right)I_n\tilde{\mathbf{U}}\mathbf{Q}\mathbf{e}_1 \overset{(d)}{=} p\lambda_{\max}\left(p^{-1}ZZ^T\right)\left\langle\mathbf{S}\mathbf{e}_1, \mathbf{e}_1\right\rangle.\end{aligned}$$

Thus, we have for some $c > 1$,

$$\begin{aligned}\Pr\left(\left\|\mathbf{e}_{jk}^T\mathbf{X}^T\right\|_2^2 > cn\right) &\leq \Pr\left(p\lambda_{\max}\left(p^{-1}\mathbf{Z}\mathbf{Z}^T\right)\left\langle\mathbf{S}\mathbf{e}_1, \mathbf{e}_1\right\rangle > cn\right) \\ &= \Pr\left(p\lambda_{\max}\left(p^{-1}\mathbf{Z}\mathbf{Z}^T\right) > cp\right) + \Pr\left(\left\langle\mathbf{S}\mathbf{e}_1, \mathbf{e}_1\right\rangle > c\frac{n}{p}\right) \\ &= O\left\{\exp\left(-Cn\right)\right\}.\end{aligned} \tag{A.3}$$

Now for the second term W, we can bound it using Gaussian tail bound again. Since $\varepsilon_1, \cdots, \varepsilon_n$ are i.i.d normally distributed $\mathcal{N}\left(0, \sigma^2\right)$ and $\|\mathbf{a}\| = 1$, we know that $W \sim \mathcal{N}\left(0, 1\right)$. Taking $x_n' = Cn^{1/2-\kappa}/\sqrt{\log\left(n\right)}$, we have

$$\begin{aligned}\Pr\left(|W| > x_n'\right) &= \Pr\left(|W| > C\frac{n^{1/2-\kappa}}{\sqrt{\log\left(n\right)}}\right) \leq \exp\left(-C\frac{n^{1-2\kappa}}{\log\left(n\right)}\right)\left(\sqrt{2\pi}C\frac{n^{1/2-\kappa}}{\sqrt{\log\left(n\right)}}\right)^{-1} \\ &= O\left\{\exp\left(-C\frac{n^{1-2\kappa}}{\log\left(n\right)}\right)\right\}.\end{aligned} \tag{A.4}$$

Therefore, combining (A.3) and (A.4), we have for $C_1 > C$,

$$\begin{aligned}\Pr\left(|\eta_{jk}| > C_1\frac{n^{1-\kappa}}{\sqrt{\log(n)}}\right) &\leq \Pr\left(\left\|\mathbf{e}_{jk}^T\mathbf{X}^T\right\|_2 > cn^{1/2}\right) + \Pr\left(|W| > C\frac{n^{1/2-\kappa}}{\sqrt{\log(n)}}\right) \\ &\leq O\left\{\exp\left(-C\frac{n^{1-2\kappa}}{\log(n)}\right)\right\}.\end{aligned}$$

Now we can easily obtain

$$\begin{aligned}\Pr\left(p_j^{-1}\left\|\boldsymbol{\eta}_j\right\|_1 > C_1\frac{n^{1-\kappa}}{\sqrt{\log(n)}}\right) &= \Pr\left(p_j^{-1}\textstyle\sum_{k=1}^{p_j}|\eta_{jk}| > C_1\frac{n^{1-\kappa}}{\sqrt{\log(n)}}\right) \\ &\leq O\left\{n^{\gamma}\exp\left(-C\frac{n^{1-2\kappa}}{\log(n)}\right)\right\} = O\left\{\exp\left(-C\frac{n^{1-2\kappa}}{\log(n)}\right)\right\}.\end{aligned}$$

Finally, we prove the screening property by combining the results above. Note that for any $j = 1, ..., J$,

$$\omega_j^g = p_j^{-1}\left\|\boldsymbol{\omega}_j\right\|_1 = p_j^{-1}\left\|\boldsymbol{\xi}_j + \boldsymbol{\eta}_j\right\|_1 \leq p_j^{-1}\left\|\boldsymbol{\xi}_j\right\|_1 + p_j^{-1}\left\|\boldsymbol{\eta}_j\right\|_1.$$

For $j \in S$, we have

$$\begin{aligned}\Pr\left(\min_{j\in S} p_j^{-1}\left\|\boldsymbol{\xi}_j\right\|_1 < cn^{1-\kappa}\right) &\leq O\left\{s\cdot\exp\left(-C\frac{n^{1-2\kappa-2\tau-\nu-\gamma}}{\log(n)}\right)\right\} \\ &\leq O\left\{\exp\left(-C\frac{n^{1-2\kappa-2\tau-\nu-\gamma}}{\log(n)}\right)\right\},\end{aligned}$$

and

$$\Pr\left(\max_{j\in S} p_j^{-1}\left\|\boldsymbol{\eta}_j\right\|_1 > C_1\frac{n^{1-\kappa}}{\sqrt{\log(n)}}\right) \leq O\left\{s\cdot\exp\left(-C\frac{n^{1-2\kappa}}{\log(n)}\right)\right\} = O\left\{\exp\left(-C\frac{n^{1-2\kappa}}{\log(n)}\right)\right\}.$$

So if we choose a threshold γ_n satisfying

$$\frac{\gamma_n}{n^{1-\kappa}} \to 0, \text{and } \frac{\gamma_n\sqrt{\log(n)}}{n^{1-\kappa}} \to \infty,$$

then we have

$$\begin{aligned}\Pr\left(\min_{j\in S}\omega_j^g<\gamma_n\right) &\leq \Pr\left(\min_{j\in S}p_j^{-1}\left\|\boldsymbol{\xi}_j\right\|_1<cn^{1-\kappa}\right)+\Pr\left(\max_{j\in S}p_j^{-1}\left\|\boldsymbol{\eta}_j\right\|_1>C_1\frac{n^{1-\kappa}}{\sqrt{\log(n)}}\right)\\ &\leq O\left\{\exp\left(-C\frac{n^{1-2\kappa-2\tau-\nu-\gamma}}{\log(n)}\right)\right\}. \qquad (A.5)\end{aligned}$$

Similarly, for $j\notin S$, we have

$$\begin{aligned}\Pr\left\{\max_{j\notin S}p_j^{-1}\left\|\boldsymbol{\xi}_j\right\|_1>O\left(\frac{n^{1-\kappa}}{\sqrt{\log(n)}}\right)\right\} &\leq O\left\{n^{\nu}\exp\left(-C\frac{n^{1-2\kappa-2\tau-\nu-\gamma}}{\log(n)}\right)\right\}\\ &= O\left\{\exp\left(-C\frac{n^{1-2\kappa-2\tau-\nu-\gamma}}{\log(n)}\right)\right\}.\end{aligned}$$

This entails

$$\begin{aligned}\Pr\left\{\max_{j\notin S}\omega_j^g>\gamma_n\right\} &\leq \Pr\left\{\max_{j\notin S}p_j^{-1}\left\|\boldsymbol{\xi}_j\right\|_1>O\left(\frac{n^{1-\kappa}}{\sqrt{\log(n)}}\right)\right\}+\Pr\left(\max_{j\in S}p_j^{-1}\left\|\boldsymbol{\eta}_j\right\|_1>O\left(\frac{n^{1-\kappa}}{\sqrt{\log(n)}}\right)\right)\\ &\leq O\left\{J\exp\left(-C\frac{n^{1-2\kappa-2\tau-\nu-\gamma}}{\log(n)}\right)\right\}+O\left\{J\exp\left(-C\frac{n^{1-2\kappa}}{\log(n)}\right)\right\}\\ &= O\left\{\exp\left(-C\frac{n^{1-2\kappa-2\tau-\nu-\gamma}}{\log(n)}\right)\right\}, \qquad (A.6)\end{aligned}$$

where $\log(J)=O\left(n^{\delta}\right)$ for $\delta\in(0,1-2\kappa-2\tau-\nu-\gamma)$ by Assumption (A4). Combining (A.5) and (A.6), we have

$$\Pr\left(\max_{j\notin S}\omega_j^g<\gamma_n<\min_{j\in S}\omega_j^g\right)\leq 1-O\left\{\exp\left(-C\frac{n^{1-2\kappa-2\tau-\nu-\gamma}}{\log(n)}\right)\right\}.$$

This indicates if we choose a submodel with size $d>s$, we will have

$$\Pr\left(\mathcal{M}_s^g\subset\mathcal{M}_{\mathcal{D},SIS}^g\right)=1-O\left\{\exp\left(-C\frac{n^{1-2\kappa-2\tau-\nu-\gamma}}{\log(n)}\right)\right\}.$$

This completes the proof of Theorem 2.1. ■

B Proof of Theorem 2.2

In this section, we will provide the proofs of the screening property for gHOLP. We first give the assumptions that are very similar to those in (A1) - (A4). Specially, assumptions (B1) and (B2) are exactly the same as assumptions (A1) and (A2). Assumption (B4) is similar to assumption (A4), except for the parameter δ' and δ that have different values.

Assumptions:

(B1) The transformed $\mathbf{z}$ has a spherically symmetric distribution and there exist some $c_1 > 1$ and $C_1 > 0$ such that

$$\Pr\left(\lambda_{\max}\left(p^{-1}\mathbf{Z}\mathbf{Z}^T\right) > c_1 \text{ or } \lambda_{\min}\left(p^{-1}\mathbf{Z}\mathbf{Z}^T\right) < c_1^{-1}\right) \leq \exp\left(-C_1 n\right),$$

where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ are the largest and smallest eigenvalues of a matrix respectively.

(B2) The random error ε has a standard normal distribution with mean zero and standard deviation σ, i.e., $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, where σ is a constant.

(B3) We assume that $\mathrm{Var}(Y) = O(1)$ and for some $\kappa \geq 0, \tau' \geq 0$ and $c_2, c_4 > 0$,

$$\min_{j \in S} \left\|\boldsymbol{\beta}_j\right\|_1 \geq \frac{c_2}{n^\kappa}, \text{and } \mathrm{cond}(\boldsymbol{\Sigma}) \leq c_4 n^{\tau'},$$

where $\mathrm{cond}(\boldsymbol{\Sigma}) = \lambda_{\max}(\boldsymbol{\Sigma}) / \lambda_{\min}(\boldsymbol{\Sigma})$ is the conditional number of $\boldsymbol{\Sigma}$.

(B4) Assume the number of true important groups $s \leq c_5 n^\nu$, $c_5 > 0$ for some $\nu > 0$ and group size $p_j = O(n^\gamma) = c_6 n^\gamma$ for some constant $c_6 > 0$. Also, assume $\log(J) = O\left(n^{\delta'}\right)$ for some $\delta' \in (0, 1 - 5\tau' - 2\kappa - \nu - \gamma)$, where $\tau', \kappa, \nu, \gamma$ are parameters defined as above.

The tail behavior of the random error ε has a significant impact on the screening performance,

but in this work, we only focus on the gaussian distribution of random error stated in assumption (B2).

Again, most of the assumptions are similar to those in individual HOLP (Wang and Leng (2015)). The key difference lies in the assumption (B3) on the magnitude of grouped coefficients $\boldsymbol{\beta}_j$ which is constrained on the group level, instead of individual level. This means that for the important groups, the individual β_{jk} can be small enough within the j-th group, but the accumulative effect of β_{jk} within the group should not be too small. In other words, we stress on the group effect of $\boldsymbol{\beta}_j$ rather than the individual contribution of β_{jk}. Also, in Assumption (B4), the p_j can be divergent as n increases for all j. Another key difference between assumption (A3) and (B3) is that the marginal covariance between important groups and response for gHOLP can be much smaller than that of gSIS. To see this, note that $\lambda_{\min}(\boldsymbol{\Sigma}) \geq \left(c_4 n^{\tau'}\right)^{-1}$, so we can obtain

$$\min_{j \in S} \left| \sum\nolimits_{k=1}^{p_j} \operatorname{cov}(X_{jk}, Y) \right| \geq |\lambda_{\min}(\boldsymbol{\Sigma})| \times \min_{j \in S} \left\| \boldsymbol{\beta}_j \right\|_1 \geq \frac{c_2 c_4}{n^{\kappa + \tau'}},$$

while the lower bound of marginal covariance between important groups and response for gSIS is $c_2 c_3 / n^{\kappa}$. This difference indicates that gHOLP can tackle better than gSIS the case where the important groups are marginally uncorrelated but jointly correlated with response. Without considering the group structure, the individual HOLP screening estimator is

$$\hat{\boldsymbol{\beta}} = \mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} Y = \mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X}\boldsymbol{\beta} + \mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \boldsymbol{\varepsilon} := \boldsymbol{\xi} + \boldsymbol{\eta}, \tag{B.1}$$

where $\boldsymbol{\xi}$ can be seen as the signal part and $\boldsymbol{\eta}$ the noise part. Note that the singular value decomposition of $\mathbf{Z}$ as $\mathbf{Z} = \mathbf{V}\mathbf{D}\mathbf{U}^T$, where $\mathbf{V} \in \mathcal{O}(n), \mathbf{U} \in V_{n,p}$ and $\mathbf{D}$ is an $n \times n$ diagonal matrix, where $\mathcal{O}(n)$ is an orthogonal space and $V_{n,p}$ is a Stiefel manifold defined as $V_{n,p} = \left\{X \in R^{p \times n} : X^T X = I_n\right\}$. This entails $\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X} = \mathbf{H}\mathbf{H}^T$, where $\mathbf{H} =$

$\mathbf{\Sigma}^{1/2}\mathbf{U}\left(\mathbf{U}^T\mathbf{\Sigma}\mathbf{U}\right)^{-1/2}$. Consequently,

$$\boldsymbol{\xi} = \mathbf{H}\mathbf{H}^T\boldsymbol{\beta}, \xi_{jk} = \mathbf{e}_{jk}^T\mathbf{H}\mathbf{H}^T\boldsymbol{\beta},$$

where $\mathbf{e}_{jk}^T = (0, \cdots, 0, 1, 0, \cdots, 0)$ is a $p \times 1$ vector with jk-th element 1 and other 0, $j = 1, ..., J, k = 1, ..., p_j$. The decomposition (B.1) indicates the equation $\hat{\boldsymbol{\beta}}_j = \boldsymbol{\xi}_j + \boldsymbol{\eta}_j$ for the j-th group, where $\boldsymbol{\xi}_j = \left(\xi_{j1}, \cdots, \xi_{jp_j}\right)'$ and $\boldsymbol{\eta}_j = \left(\eta_{j1}, \cdots, \eta_{jp_j}\right)'$. Thus, we have

$$\left\|\hat{\boldsymbol{\beta}}_j\right\|_1 = \left\|\boldsymbol{\xi}_j + \boldsymbol{\eta}_j\right\|_1 \leq \left\|\boldsymbol{\xi}_j\right\|_1 + \left\|\boldsymbol{\eta}_j\right\|_1.$$

The gHOLP screening criterion is defined by $\hat{\boldsymbol{\beta}}^g = \left(\hat{\beta}_1^g, \cdots, \hat{\beta}_J^g\right)$, where $\hat{\beta}_j^g = p_j^{-1}\left\|\hat{\boldsymbol{\beta}}_j\right\|_1$. Theorem 2.2 states that as the sample size goes to large enough, the probability of retaining the true important groups will be overwhelming large by using the group HOLP. To prove it, we will first bound the $\left\|\boldsymbol{\xi}_j\right\|_1$ and $\left\|\boldsymbol{\eta}_j\right\|_1$ separately and then adjust them by the group size p_j.

Lemma B.1 (bounding $\left\|\boldsymbol{\xi}_j\right\|_1$) *Assume (B1)-(B4) hold, then we have for any $C > 0$, $c, \tilde{c} > 0$ such that for any $j \in S$,*

$$\Pr\left(\left\|\boldsymbol{\xi}_j\right\|_1 < c\frac{n^{1-\tau'-\kappa+\gamma}}{p}\right) \leq O\left\{\exp\left(-\frac{Cn^{1-5\tau'-2\kappa-\nu-\gamma}}{2\log(n)}\right)\right\},$$

and for any $j \notin S$,

$$\Pr\left(\left\|\boldsymbol{\xi}_j\right\|_1 > \frac{\tilde{c}}{\sqrt{\log(n)}}\frac{n^{1-\tau'-\kappa+\gamma}}{p}\right) \leq O\left\{\exp\left(-\frac{Cn^{1-5\tau'-2\kappa-\nu-\gamma}}{2\log(n)}\right)\right\},$$

where $\tau', \kappa, \nu, \gamma$ are parameters defined in (B3), (B4) and satisfy $(5\tau' + 2\kappa + \nu + \gamma) \in (0, 1)$

.

Proof: Using the results in proof of Lemma 5 in Wang and Leng (2015), the diagonal and

off-diagonal components are bounded by

$$\Pr\left(\mathbf{e}_{jk}^T\mathbf{H}\mathbf{H}^T\mathbf{e}_{jk} < c_1'\frac{n^{1-\tau'}}{p}\right) < 2\exp\left(-Cn\right), \tag{B.2}$$

for $0 < c_1' < 1 < c_2'$ and

$$\Pr\left(\left|\mathbf{e}_{jk}^T\mathbf{H}\mathbf{H}^T\mathbf{e}_{j'k'}\right| > \frac{C_1'}{\sqrt{\log(n)}}\frac{n^{1+\tau'-\alpha}}{p}\right) \leq O\left\{\exp\left(-\frac{Cn^{1-2\alpha}}{2\log(n)}\right)\right\}$$

for $C_1' > 0, \alpha \in (0, 1/2)$, where $jk \neq j'k', j, j' = 1, ..., J, k', k = 1, ..., p_j$.

Now for $j \notin S$, we know that $\beta_{jk} = 0$. Thus we have by Cauchy-Schwartz inequality

$$\begin{aligned}
|\xi_{jk}| &= \left|\mathbf{e}_{jk}^T\mathbf{H}\mathbf{H}^T\boldsymbol{\beta}\right| = \left|\mathbf{e}_{jk}^T\mathbf{H}\mathbf{H}^T\sum\nolimits_{j'\in S}\sum\nolimits_{k'=1}^{p_{j'}}\mathbf{e}_{j'k'}\beta_{j'k'} + \mathbf{e}_{jk}^T\mathbf{H}\mathbf{H}^T\sum\nolimits_{j'\notin S}\sum\nolimits_{k'=1}^{p_{j'}}\mathbf{e}_{j'k'}\beta_{j'k'}\right| \\
&= \left|\mathbf{e}_{jk}^T\mathbf{H}\mathbf{H}^T\sum\nolimits_{j'\in S}\sum\nolimits_{k'=1}^{p_{j'}}\mathbf{e}_{j'k'}\beta_{j'k'}\right| = \left|\sum\nolimits_{j'\in S}\sum\nolimits_{k'=1}^{p_{j'}}\mathbf{e}_{jk}^T\mathbf{H}\mathbf{H}^T\mathbf{e}_{j'k'}\beta_{j'k'}\right| \\
&\leq \sqrt{\left|\sum\nolimits_{j'\in S}\sum\nolimits_{k'=1}^{p_{j'}}\mathbf{e}_{jk}^T\mathbf{H}\mathbf{H}^T\mathbf{e}_{j'k'}\right|^2}\sqrt{\sum\nolimits_{j'\in S}\sum\nolimits_{k'=1}^{p_{j'}}\left|\beta_{j'k'}\right|^2} \\
&\leq \sqrt{\sum\nolimits_{j'\in S}\sum\nolimits_{k'=1}^{p_{j'}}\left|\mathbf{e}_{jk}^T\mathbf{H}\mathbf{H}^T\mathbf{e}_{j'k'}\right|^2}\left\|\boldsymbol{\beta}\right\|_2.
\end{aligned}$$

Note that by the standardization on predictors $\mathbf{X}$, we have $\boldsymbol{\Sigma}_{ii} = 1, i = 1, ..., p$, which implies $\lambda_{\min}(\boldsymbol{\Sigma}) \leq 1 \leq \lambda_{\max}(\boldsymbol{\Sigma})$. So by Assumption (B3) we can easily get $\lambda_{\min}(\boldsymbol{\Sigma}) \geq \left(c_4 n^{\tau'}\right)^{-1}$. Thus, we can obtain the bound for $\|\boldsymbol{\beta}\|_2$,

$$\frac{1}{c_4 n^{\tau'}}\left\|\boldsymbol{\beta}\right\|_2^2 \leq \lambda_{\min}(\boldsymbol{\Sigma})\left\|\boldsymbol{\beta}\right\|_2^2 = \boldsymbol{\beta}^T\lambda_{\min}(\boldsymbol{\Sigma})\, I_p\boldsymbol{\beta} \leq \boldsymbol{\beta}^T\,\boldsymbol{\Sigma}\boldsymbol{\beta} = \mathrm{Var}(Y) - \sigma^2 < c',$$

resulting in $\|\boldsymbol{\beta}\|_2 \leq \sqrt{c' c_4 n^{\tau'}}$ for some constant c' and it follows that for $j \notin S$ and some

constant $C_1' > 0$,

$$\begin{aligned}
& \Pr\left(|\xi_{jk}| \leq \frac{\sqrt{c'c_4c_3c_5}C_1'}{\sqrt{\log(n)}}\frac{n^{1+3\tau'/2+\nu/2+\gamma/2-\alpha}}{p}\right) \\
\geq\ & \Pr\left(\sqrt{c'c_4n^{\tau'}}\sqrt{\sum\nolimits_{j'\in S}\sum\nolimits_{k'=1}^{p_{j'}}\left|\mathbf{e}_{jk}^T\mathbf{H}\mathbf{H}^T\mathbf{e}_{j'k'}\right|^2} \leq \frac{\sqrt{c'c_4}C_1'}{\sqrt{\log(n)}}\frac{n^{1+3\tau'/2+\nu/2+\gamma/2-\alpha}}{p}\right) \\
=\ & \Pr\left(\sqrt{\sum\nolimits_{j'\in S}\sum\nolimits_{k'=1}^{p_{j'}}\left|\mathbf{e}_{jk}^T\mathbf{H}\mathbf{H}^T\mathbf{e}_{j'k'}\right|^2} \leq \frac{\sqrt{c_3c_5}C_1'}{\sqrt{\log(n)}}\frac{n^{1+\tau'+\nu/2+\gamma/2-\alpha}}{p}\right) \\
\geq\ & \Pr\left(\sqrt{c_3n^{\nu}c_5n^{\gamma}}\left|\mathbf{e}_{jk}^T\mathbf{H}\mathbf{H}^T\mathbf{e}_{j'k'}\right| \leq \frac{\sqrt{c_3c_5}C_1'}{\sqrt{\log(n)}}\frac{n^{1+\tau'+\nu/2+\gamma/2-\alpha}}{p}\right) \\
=\ & \Pr\left(\left|\mathbf{e}_{jk}^T\mathbf{H}\mathbf{H}^T\mathbf{e}_{j'k'}\right| \leq \frac{C_1'}{\sqrt{\log(n)}}\frac{n^{1+\tau'-\alpha}}{p}\right) = 1 - O\left\{\exp\left(-\frac{Cn^{1-2\alpha}}{2\log(n)}\right)\right\}.
\end{aligned}$$

Taking $\alpha = (5/2)\,\tau' + \kappa + \nu/2 + \gamma/2 \in (0, 1/2)$, we have

$$\Pr\left(|\xi_{jk}| > \frac{C_1'}{\sqrt{\log(n)}}\frac{n^{1-\tau'-\kappa}}{p}\right) \leq O\left\{\exp\left(-\frac{Cn^{1-5\tau'-2\kappa-\gamma-\nu}}{2\log(n)}\right)\right\}. \tag{B.3}$$

This entails that for $j \notin S$, we have

$$\begin{aligned}
\Pr\left(\left\|\boldsymbol{\xi}_j\right\|_1 > \frac{c_5C_1'}{\sqrt{\log(n)}}\frac{n^{1-\tau-\kappa+\gamma}}{p}\right) &\leq \Pr\left(c_5n^{\gamma}\max_k|\xi_{jk}| > \frac{c_5C_1'}{\sqrt{\log(n)}}\frac{n^{1-\tau-\kappa+\gamma}}{p}\right) \\
&\leq O\left\{\exp\left(-\frac{Cn^{1-5\tau-2\kappa-\gamma-\nu}}{2\log(n)}\right)\right\}.
\end{aligned}$$

Taking $\tilde{c} = c_5C_1'/\sqrt{\log(n)}$, we have for $j \notin S$,

$$\Pr\left(\left\|\boldsymbol{\xi}_j\right\|_1 > \tilde{c}\frac{n^{1-\tau'-\kappa+\gamma}}{p}\right) \leq O\left\{\exp\left(-\frac{Cn^{1-5\tau'-2\kappa-\gamma-\nu}}{2\log(n)}\right)\right\}.$$

Next, we bound the $\left\|\boldsymbol{\xi}_j\right\|_1$ for $j \in S.$, we have for $jk \neq j'k'$,

$$\begin{aligned}
\left\|\boldsymbol{\xi}_j\right\|_1 &= \sum_{k=1}^{p_j} |\xi_{jk}| = \sum_{k=1}^{p_j} \left|\mathbf{e}_{jk}^T \mathbf{H}\mathbf{H}^T \boldsymbol{\beta}\right| \\
&= \sum_{k=1}^{p_j} \left|\mathbf{e}_{jk}^T \mathbf{H}\mathbf{H}^T \mathbf{e}_{jk}\beta_{jk}\right| + \sum_{k=1}^{p_j} \left|\mathbf{e}_{jk}^T \mathbf{H}\mathbf{H}^T \sum\nolimits_{j' \in S} \sum\nolimits_{k'=1}^{p_{j'}} \mathbf{e}_{j'k'}\beta_{j'k'}\right| \\
&= \sum_{k=1}^{p_j} \left|\mathbf{e}_{jk}^T \mathbf{H}\mathbf{H}^T \mathbf{e}_{jk}\right| |\beta_{jk}| + \sum_{k=1}^{p_j} \left|\mathbf{e}_{jk}^T \mathbf{H}\mathbf{H}^T \sum\nolimits_{j' \in S} \sum\nolimits_{k'=1}^{p_{j'}} \mathbf{e}_{j'k'}\beta_{j'k'}\right| \\
&\geq \sum_{k=1}^{p_j} \left|\mathbf{e}_{jk}^T \mathbf{H}\mathbf{H}^T \mathbf{e}_{jk}\right| |\beta_{jk}| - \sum_{k=1}^{p_j} \left|\mathbf{e}_{jk}^T \mathbf{H}\mathbf{H}^T \sum\nolimits_{j' \in S} \sum\nolimits_{k'=1}^{p_{j'}} \mathbf{e}_{j'k'}\beta_{j'k'}\right|.
\end{aligned}$$

By Bonferroni's inequality and Assumption (B4), along with combination of (B.2) and (B.3), we can obtain the bound of $\left\|\boldsymbol{\xi}_j\right\|_1$ by using the same value of α. That is,

$$\begin{aligned}
& \Pr\left\{\left\|\boldsymbol{\xi}_j\right\|_1 < \left(c_2 c_1' + \frac{c_6\sqrt{c'c_4}C_1'}{\sqrt{\log(n)}}\right)\frac{n^{1-\tau'-\kappa+\gamma}}{p}\right\} \\
\leq\ & \Pr\left(\sum_{k=1}^{p_j} \left|\mathbf{e}_{jk}^T \mathbf{H}\mathbf{H}^T \mathbf{e}_{jk}\right| |\beta_{jk}| < c_2 c_1' \frac{n^{1-\tau'-\kappa+\gamma}}{p}\right) + \\
& \Pr\left(\sum_{k=1}^{p_j} \left|\mathbf{e}_{jk}^T \mathbf{H}\mathbf{H}^T \sum\nolimits_{j' \in S} \sum\nolimits_{k'=1}^{p_{j'}} \mathbf{e}_{j'k'}\beta_{j'k'}\right| > \frac{c_6\sqrt{c'c_4}C_1'}{\sqrt{\log(n)}}\frac{n^{1-\tau'-\kappa+\gamma}}{p}\right) \\
\leq\ & \Pr\left(\min_k \left|\mathbf{e}_{jk}^T \mathbf{H}\mathbf{H}^T \mathbf{e}_{jk}\right| \sum_{k=1}^{p_j} |\beta_{jk}| < c_2 c_1' \frac{n^{1-\tau'-\kappa+\gamma}}{p}\right) + \\
& + \Pr\left(c_6 n^\gamma \max_k \left|\mathbf{e}_{jk}^T \mathbf{H}\mathbf{H}^T \sum\nolimits_{j' \in S} \sum\nolimits_{k'=1}^{p_{j'}} \mathbf{e}_{j'k'}\beta_{j'k'}\right| > \frac{c_6\sqrt{c'c_4}C_1'}{\sqrt{\log(n)}}\frac{n^{1-\tau'-\kappa+\gamma}}{p}\right) \\
\leq\ & \Pr\left(\min_k \left|\mathbf{e}_{jk}^T \mathbf{H}\mathbf{H}^T \mathbf{e}_{jk}\right| < c_1' \frac{n^{1-\tau'}}{p}\right) + c_6 n^\gamma \Pr\left(\left|\mathbf{e}_{jk}^T \mathbf{H}\mathbf{H}^T \mathbf{e}_{j'k'}\right| > \frac{C_1'}{\sqrt{\log(n)}}\frac{n^{1-\tau'-\kappa}}{p}\right) \\
\leq\ & c_6 n^\gamma \Pr\left(\left|\mathbf{e}_{jk}^T \mathbf{H}\mathbf{H}^T \mathbf{e}_{jk}\right| < c_1' \frac{n^{1-\tau'}}{p}\right) + c_6 n^\gamma \Pr\left(\left|\mathbf{e}_{jk}^T \mathbf{H}\mathbf{H}^T \mathbf{e}_{j'k'}\right| > \frac{C_1'}{\sqrt{\log(n)}}\frac{n^{1-\tau'-\kappa}}{p}\right) \\
=\ & 2c_6 n^\gamma \cdot O\left(\exp\left(-Cn\right)\right) + c_6 n^\gamma \cdot O\left\{\exp\left(-\frac{Cn^{1-5\tau'-2\kappa-\gamma-\nu}}{2\log(n)}\right)\right\} \\
=\ & O\left\{\exp\left(-\frac{Cn^{1-5\tau-2\kappa-\gamma-\nu}}{2\log(n)}\right)\right\},
\end{aligned}$$

which is equivalent to $\Pr\left(\left\|\boldsymbol{\xi}_j\right\|_1 \geq c\frac{n^{1-\tau-\kappa+\gamma}}{p}\right) \geq 1 - O\left\{\exp\left(-\frac{Cn^{1-5\tau-2\kappa-\gamma-\nu}}{2\log(n)}\right)\right\}$. This completes the proof. ■

Lemma B.2 (bounding $|\eta_{jk}|$, Lemma 6 in Wang and Leng (2015)) *Assume (B1)-(B4) hold, we have for any $j \in \{1,2,...,J\}, k \in \{1,...,p_j\}$,*

$$\Pr\left(|\eta_{jk}| > \frac{\sigma\sqrt{C_1 c_1 c_2' c_4}}{\sqrt{\log(n)}}\frac{n^{1-\kappa-\tau'}}{p}\right) < O\left(\exp\left\{-\frac{C_1 n^{1-4\tau'-2\kappa}}{2\log(n)}\right\}\right),$$

where C_1, c_1, c_4 are defined in the assumption, and $c_2' > 1$.

Lemma B.3 (bounding $\left\|\boldsymbol{\eta}_j\right\|_1$) *Assume (B1)-(B4) hold, we have for any $j \in \{1,2,...,J\}$,*

$$\Pr\left(\left\|\boldsymbol{\eta}_j\right\|_1 > \frac{c_6\sigma\sqrt{C_1 c_1 c_2' c_4}}{\sqrt{\log(n)}}\frac{n^{1-\kappa-\tau'+\gamma}}{p}\right) < O\left(\exp\left\{-\frac{C_1 n^{1-4\tau'-2\kappa}}{2\log(n)}\right\}\right). \tag{B.4}$$

Proof: We let $C' = \sigma\sqrt{C_1 c_1 c_2' c_4}$. Note that $\left\|\boldsymbol{\eta}_j\right\|_1 = \sum_{k=1}^{p_j}|\eta_{jk}|$ and $p_j = c_6 n^{\gamma}$, it follows that

$$\begin{aligned}
\Pr\left(\left\|\boldsymbol{\eta}_j\right\|_1 > \frac{c_6 C'}{\sqrt{\log(n)}}\frac{n^{1-\kappa-\tau'+\gamma}}{p}\right) &\leq \Pr\left(p_j \max_k |\eta_{jk}| > \frac{c_6 C'}{\sqrt{\log(n)}}\frac{n^{1-\kappa-\tau'+\gamma}}{p}\right) \\
&\leq c_6 n^{\gamma}\cdot\Pr\left(c_6 n^{\gamma}|\eta_{jk}| > \frac{c_6 C'}{\sqrt{\log(n)}}\frac{n^{1-\kappa-\tau'+\gamma}}{p}\right) \\
&= c_6 n^{\gamma}\cdot\Pr\left(|\eta_{jk}| > \frac{C'}{\sqrt{\log(n)}}\frac{n^{1-\kappa-\tau'}}{p}\right) \\
&= O\left(\exp\left\{-\frac{C_1 n^{1-4\tau'-2\kappa}}{2\log(n)}\right\}\right).
\end{aligned}$$

This completes the proof. ■

Note that $c_6 C' n^{1-\kappa-\tau'+\gamma}/(p\log(n)) = o\left(n^{1-\kappa-\tau'+\gamma}\right)$. Thus, the inequality (B.4) is equivalent to

$$\Pr\left(\left\|\boldsymbol{\eta}_j\right\|_1 \leq o\left(n^{1-\kappa-\tau'+\gamma}\right)\right) \geq 1 - O\left(\exp\left\{-\frac{C_1 n^{1-4\tau'-2\kappa}}{2\log(n)}\right\}\right).$$

This implies that as $n \to \infty$, the probability of $\left\|\boldsymbol{\eta}_j\right\|_1$ being small enough is going to 1.

Proof of Theorem 2.2: Recall $J = O\left\{\exp\left(n^{\delta'}\right)\right\}, \delta' \in (0, 1-5\tau'-2\kappa-\nu-\gamma)$ and

$s = c_5 n^\nu, \nu \geq 0$. By Lemma B.1, we have for $j \in S$,

$$\begin{aligned}\Pr\left(\min_{j\in S} p_j^{-1} \left\|\boldsymbol{\xi}_j\right\|_1 < c\frac{n^{1-\tau'-\kappa}}{p}\right) &\leq s \cdot O\left\{\exp\left(-\frac{Cn^{1-5\tau'-2\kappa-\nu-\gamma}}{2\log(n)}\right)\right\} \\ &= O\left\{\exp\left(-\frac{Cn^{1-5\tau'-2\kappa-\nu-\gamma}}{2\log(n)}\right)\right\},\end{aligned}$$

and for $j \notin S$,

$$\begin{aligned}\Pr\left(\max_{j\notin S} p_j^{-1} \left\|\boldsymbol{\xi}_j\right\|_1 > \frac{\tilde{c}}{\sqrt{\log(n)}}\frac{n^{1-\tau'-\kappa}}{p}\right) &\leq J \cdot O\left\{\exp\left(-\frac{Cn^{1-5\tau'-2\kappa-\nu-\gamma}}{2\log(n)}\right)\right\} \\ &= O\left\{\exp\left(-\frac{Cn^{1-5\tau'-2\kappa-\nu-\gamma}}{2\log(n)}\right)\right\}.\end{aligned}$$

Also, taking $\tilde{c}_1 = c_6\sqrt{C_1 c_1 c_2' c_4}$ and by Lemma B.3, we have

$$\begin{aligned}\Pr\left(\max_{j} p_j^{-1} \left\|\boldsymbol{\eta}_j\right\|_1 > \frac{\tilde{c}_1}{\sqrt{\log(n)}}\frac{n^{1-\kappa-\tau'}}{p}\right) &< J \cdot O\left(\exp\left\{-\frac{C_1 n^{1-4\tau'-2\kappa}}{2\log(n)}\right\}\right) \\ &= O\left(\exp\left\{-\frac{C_1 n^{1-4\tau'-2\kappa}}{2\log(n)}\right\}\right).\end{aligned}$$

If we choose a threshold γ_n such that

$$\frac{p\gamma_n}{n^{1-\tau'-\kappa}} \to 0, \text{and } \frac{p\gamma_n\sqrt{\log(n)}}{n^{1-\tau'-\kappa}} \to \infty,$$

then we have by Bonferroni's inequality,

$$\begin{aligned}&\Pr\left(\min_{j\in S} p_j^{-1}\left\|\hat{\boldsymbol{\beta}}_j\right\|_1 < \gamma_n\right) \\ \leq\ &\Pr\left(\min_{j\in S} p_j^{-1}\left\|\boldsymbol{\xi}_j\right\|_1 < c\frac{n^{1-\tau'-\kappa}}{p}\right) + \Pr\left(\max_{j\in S} p_j^{-1}\left\|\boldsymbol{\eta}_j\right\|_1 > \frac{c_6 C'}{\sqrt{\log(n)}}\frac{n^{1-\kappa-\tau'}}{p}\right) \\ <\ &O\left\{\exp\left(-\frac{Cn^{1-5\tau'-2\kappa-\nu-\gamma}}{2\log(n)}\right)\right\},\end{aligned}$$

and

$$\begin{aligned}
& \Pr\left(\max_{j\notin S} p_j^{-1}\left\|\hat{\boldsymbol{\beta}}_j\right\|_1 > \gamma_n\right) \\
\leq\ & \Pr\left(\max_{j\notin S} p_j^{-1}\left\|\boldsymbol{\xi}_j\right\|_1 > \frac{\tilde{c}}{\sqrt{\log(n)}}\frac{n^{1-\tau'-\kappa}}{p}\right) + \Pr\left(\max_{j\notin S} p_j^{-1}\left\|\boldsymbol{\eta}_j\right\|_1 > \frac{c_6 C'}{\sqrt{\log(n)}}\frac{n^{1-\kappa-\tau'}}{p}\right) \\
<\ & O\left\{\exp\left(-\frac{Cn^{1-5\tau'-2\kappa-\nu-\gamma}}{2\log(n)}\right)\right\}.
\end{aligned}$$

This indicates

$$\Pr\left(\min_{j\in S}\hat{\beta}_j^g > \gamma_n > \max_{j\notin S}\hat{\beta}_j^g\right) \geq 1 - O\left\{\exp\left(-\frac{Cn^{1-5\tau'-2\kappa-\nu-\gamma}}{2\log(n)}\right)\right\}.$$

Obviously, this implies if we choose a submodel $\mathcal{M}_{\mathcal{D},HOLP}^g$ with $d > s$, we have

$$P\left(\mathcal{M}_S^g \subset \mathcal{M}_{\mathcal{D},HOLP}^g\right) = 1 - O\left\{\exp\left(-C_1\frac{n^{1-2\kappa-5\tau'-\nu-\gamma}}{2\log(n)}\right)\right\}.$$

The proof is completed. ■

C Proof of Theorem 2.3

Without loss of generality, we assume all groups have the same number of variables, p_j for simplicity. This indicates that using R_j^2 is the same as using $\bar{R}_j^2$ for the screening purpose. Recall that $Y = (y_1, ..., y_n)^T$, $\mathbf{X}_j = \left(X_{j1}, ..., X_{jp_j}\right)$, where $X_{jk} = (x_{jk1}, ..., x_{jkn})^T$. For the j-th group, R_j^2 is defined as

$$R_j^2 = 1 - \frac{\left(Y - \hat{Y}\right)^T\left(Y - \hat{Y}\right)}{\left(Y - \bar{Y}\right)^T\left(Y - \bar{Y}\right)},$$

where $\hat{Y} = (\hat{y}_1, ..., \hat{y}_n)^T$ and $\bar{Y} = (\bar{y}, ..., \bar{y})^T$. We first observe that $\hat{Y}$ is orthogonal to $Y - \hat{Y}$, which is equivalent to $\hat{Y}^T \left(Y - \hat{Y} \right) = \hat{Y}^T Y - \hat{Y}^T \hat{Y} = 0$, i.e., $\hat{Y}^T Y = \hat{Y}^T \hat{Y}$. By the assumption $E(Y) = 0$, one has

$$\begin{aligned} R_j^2 &= 1 - \frac{Y^T Y - Y^T \hat{Y} - \hat{Y}^T Y + \hat{Y}^T \hat{Y}}{Y^T Y} = 1 - \frac{Y^T Y - 2Y^T \hat{Y} + \hat{Y}^T \hat{Y}}{Y^T Y} \\ &= 1 - \frac{Y^T Y - 2\hat{Y}^T \hat{Y} + \hat{Y}^T \hat{Y}}{Y^T Y} = \frac{\hat{Y}^T \hat{Y}}{Y^T Y}. \end{aligned}$$

The denominator is the same for all groups, so we only study the property of numerator of R_j^2. Since $\hat{Y} = \mathbf{X}_j \left(\mathbf{X}_j^T \mathbf{X}_j \right)^{-1} \mathbf{X}_j^T Y := \mathbf{H}_j Y$ and $\mathbf{H}_j^T = \mathbf{H}_j$, $\mathbf{H}_j^2 = \mathbf{H}_j$, one can obtain

$$\hat{Y}^T \hat{Y} = (\mathbf{H}_j Y)^T \mathbf{H}_j Y = Y^T \mathbf{H}_j Y = Y^T \mathbf{X}_j \left(\mathbf{X}_j^T \mathbf{X}_j \right)^{-1} \mathbf{X}_j^T Y = \left(\mathbf{X}_j^T Y \right)^T \left(\mathbf{X}_j^T \mathbf{X}_j \right)^{-1} \mathbf{X}_j^T Y.$$

Hence, R_j^2 can be written as

$$R_j^2 = \frac{\left(\mathbf{X}_j^T Y \right)^T \left(\mathbf{X}_j^T \mathbf{X}_j \right)^{-1} \mathbf{X}_j^T Y}{Y^T Y}.$$

To gain the sure screening property of gAR2, we need the same assumptions as (A1) - (A4) in Appendix A, except that $p_j = c_6 n^{\gamma''}$ with $\gamma'' \in (0, 1)$ and $0 \leq c_6 \leq 1$ to ensure the condition $p_j < n$. In addition, we also need an assumption of concentration property for the random matrix $\mathbf{Z}_j := \mathbf{X}_j \mathbf{\Sigma}_j^{-1/2}$ such that

$$\Pr \left(\lambda_{\max}^{-1} \left(p_j^{-1} \mathbf{Z}_j \mathbf{Z}_j^T \right) > c_1 \text{ or } \lambda_{\max}^{-1} \left(p_j^{-1} \mathbf{Z}_j \mathbf{Z}_j^T \right) < c_1^{-1} \right) \leq \exp \left(-C_1 n \right).$$

In the proofs, we do not intent to make the assumptions weakest. Note that $\mathbf{X}_j = \mathbf{Z}_j \mathbf{\Sigma}_j^{1/2}$, one has

$$\mathbf{X}_j^T \mathbf{X}_j = p_j \mathbf{\Sigma}_j^{1/2} \left(p_j^{-1} \mathbf{Z}_j^T \mathbf{Z}_j \right) \mathbf{\Sigma}_j^{1/2}.$$

Using the fact that $\mathbf{Z}_j^T\mathbf{Z}_j$ and $\mathbf{Z}_j\mathbf{Z}_j^T$ have the same eigenvalues, one can obtain $\lambda_{\min}\left(p_j^{-1}\mathbf{Z}_j\mathbf{Z}_j^T\right) I_{p_j} \leq p_j^{-1}\mathbf{Z}_j^T\mathbf{Z}_j \leq \lambda_{\max}\left(p_j^{-1}\mathbf{Z}_j\mathbf{Z}_j^T\right) I_{p_j}$. It entails that

$$\mathbf{X}_j^T\mathbf{X}_j \leq p_j\boldsymbol{\Sigma}_j^{1/2}\lambda_{\max}\left(p_j^{-1}\mathbf{Z}_j\mathbf{Z}_j^T\right) I_{p_j}\boldsymbol{\Sigma}_j^{1/2} = p_j\lambda_{\max}\left(p_j^{-1}\mathbf{Z}_j\mathbf{Z}_j^T\right)\boldsymbol{\Sigma}_j \leq p_j\lambda_{\max}\left(p_j^{-1}\mathbf{Z}_j\mathbf{Z}_j^T\right)\lambda_{\max}\left(\boldsymbol{\Sigma}_j\right) I_{p_j}$$

and

$$\mathbf{X}_j^T\mathbf{X}_j \geq p_j\boldsymbol{\Sigma}_j^{1/2}\lambda_{\min}\left(p_j^{-1}\mathbf{Z}_j\mathbf{Z}_j^T\right) I_{p_j}\boldsymbol{\Sigma}_j^{1/2} \geq p_j\lambda_{\min}\left(p_j^{-1}\mathbf{Z}_j\mathbf{Z}_j^T\right)\lambda_{\min}\left(\boldsymbol{\Sigma}_j\right) I_{p_j},$$

which follows

$$p_j^{-1}\lambda_{\max}^{-1}\left(p_j^{-1}\mathbf{Z}_j\mathbf{Z}_j^T\right)\lambda_{\max}^{-1}\left(\boldsymbol{\Sigma}_j\right) I_{p_j} \leq \left(\mathbf{X}_j^T\mathbf{X}_j\right)^{-1} \leq p_j^{-1}\lambda_{\min}^{-1}\left(p_j^{-1}\mathbf{Z}_j\mathbf{Z}_j^T\right)\lambda_{\min}^{-1}\left(\boldsymbol{\Sigma}_j\right) I_{p_j}.$$

Therefore,

$$p_j^{-1}\lambda_{\max}^{-1}\left(p_j^{-1}\mathbf{Z}_j\mathbf{Z}_j^T\right)\lambda_{\max}^{-1}\left(\boldsymbol{\Sigma}_j\right)\left\|\mathbf{X}_j^TY\right\|_2^2 \leq \hat{Y}^T\hat{Y} \leq p_j^{-1}\lambda_{\min}^{-1}\left(p_j^{-1}\mathbf{Z}_j\mathbf{Z}_j^T\right)\lambda_{\min}^{-1}\left(\boldsymbol{\Sigma}_j\right)\left\|\mathbf{X}_j^TY\right\|_2^2.$$

By Assumption (A3), one has

$$\lambda_{\max}^{-1}\left(\boldsymbol{\Sigma}_j\right) \geq \lambda_{\max}^{-1}\left(\boldsymbol{\Sigma}\right) \geq \left(c_4n^{\tau}\right)^{-1}, \text{and } \lambda_{\min}^{-1}\left(\boldsymbol{\Sigma}_j\right) \leq c_3^{-1}.$$

Since $\omega_j^g = p_j^{-1}\left\|\mathbf{X}_j^TY\right\|_2$ for gSIS and we have the sure screening property that for some $c > 0, c' > 0$ under Assumptions (A1) - (A4)

$$\Pr\left(\max_{j\notin S}\omega_j^g > \frac{c'n^{1-\kappa-\tau/2}}{\sqrt{\log\left(n\right)}}\right) = \Pr\left(\max_{j\notin S}\left\|\mathbf{X}_j^TY\right\|_2^2 > p_j^2\gamma_n^2\right) \leq O\left\{\exp\left(-C\frac{n^{1-2\kappa-3\tau-\nu-\gamma''}}{\log\left(n\right)}\right)\right\},$$

and

$$\Pr\left(\min_{j\in S}\omega_j^g < \frac{cn^{1-\kappa}}{2}\right) = \Pr\left(\min_{j\in S}\left\|\mathbf{X}_j^TY\right\|_2^2 < p_j^2\gamma_n^2\right) \leq O\left\{\exp\left(-C\frac{n^{1-2\kappa-3\tau-\nu-\gamma''}}{\log\left(n\right)}\right)\right\}.$$

These two inequalities are slightly different from those in Appendix A, but they can be easily obtained by simply taking $x_n = \sqrt{2}c_1c_4cn^{1-\kappa-\tau/2-\nu/2-\gamma/2}/\sqrt{\log(n)}$ and $x_n' = cn^{1-\kappa-\tau/2}/\sqrt{\log(n)}$ in the proofs of Theorem A.1, whose details are omitted here. Therefore, for $j \in S$, we have by Bonferroni's inequality

$$\begin{aligned}
& \Pr\left\{\min_{j\in S}\left(\mathbf{X}_j^T Y\right)^T\left(\mathbf{X}_j^T\mathbf{X}_j\right)^{-1}\mathbf{X}_j^T Y < \frac{c^2 p_j n^{2-2\kappa}n^{-\tau}}{4c_1c_4}\right\} \\
\leq\ & \Pr\left\{p_j^{-1}\lambda_{\max}^{-1}\left(p_j^{-1}\mathbf{Z}_j\mathbf{Z}_j^T\right)(c_4 n^{\tau})^{-1}\min_{j\in S}\left\|\mathbf{X}_j^T Y\right\|_2^2 \leq \frac{c^2 p_j}{4c_1c_4}n^{2-2\kappa-\tau}\right\} \\
\leq\ & \Pr\left\{\lambda_{\max}^{-1}\left(p_j^{-1}\mathbf{Z}_j\mathbf{Z}_j^T\right) \leq c_1^{-1}\right\} + \Pr\left\{\min_{j\in S}\left\|\mathbf{X}_j^T Y\right\|_2^2 \leq p_j^2\left(\frac{n^{1-\kappa}}{2}\right)^2\right\} \\
\leq\ & O\left\{\exp\left(-C_1 n\right) + \exp\left(-C\frac{n^{1-2\kappa-3\tau-\nu-\gamma''}}{\log(n)}\right)\right\} \leq O\left\{\exp\left(-C\frac{n^{1-2\kappa-3\tau-\nu-\gamma''}}{\log(n)}\right)\right\}.
\end{aligned}$$

By assumption $\mathrm{Var}(Y) = O(1)$, one has

$$\Pr\left\{\min_{j\in S} R_j^2 \leq \frac{c^2 c_6 n^{2-2\kappa-\tau+\gamma''}}{4c_1c_4}\right\} \leq O\left\{\exp\left(-C\frac{n^{1-2\kappa-3\tau-\nu-\gamma''}}{\log(n)}\right)\right\}.$$

Similarly, for $j \notin S$, we have

$$\begin{aligned}
& \Pr\left\{\max_{j\notin S}\left(\mathbf{X}_j^T Y\right)^T\left(\mathbf{X}_j^T\mathbf{X}_j\right)^{-1}\mathbf{X}_j^T Y > p_j c_1 c_3^{-1}\left(\frac{c' n^{1-\kappa-\tau/2}}{\sqrt{\log(n)}}\right)^2\right\} \\
\leq\ & \Pr\left\{p_j^{-1}\lambda_{\min}^{-1}\left(p_j^{-1}\mathbf{Z}_j\mathbf{Z}_j^T\right)c_3^{-1}\max_{j\notin S}\left\|\mathbf{X}_j^T Y\right\|_2^2 > p_j c_1 c_3^{-1}\left(\frac{c' n^{1-\kappa-\tau/2}}{\sqrt{\log(n)}}\right)^2\right\} \\
=\ & \Pr\left\{\lambda_{\min}^{-1}\left(p_j^{-1}\mathbf{Z}_j\mathbf{Z}_j^T\right)\max_{j\notin S}\left\|\mathbf{X}_j^T Y\right\|_2^2 > p_j^2 c_1\left(\frac{c' n^{1-\kappa-\tau/2}}{\sqrt{\log(n)}}\right)^2\right\} \\
\leq\ & \Pr\left\{\lambda_{\min}^{-1}\left(p_j^{-1}\mathbf{Z}_j\mathbf{Z}_j^T\right) > c_1\right\} + \Pr\left\{\max_{j\notin S}\left\|\mathbf{X}_j^T Y\right\|_2^2 > p_j^2 c_1\left(\frac{c' n^{1-\kappa-\tau/2}}{\sqrt{\log(n)}}\right)^2\right\} \\
\leq\ & O\left\{\exp\left(-C_1 n\right) + \exp\left(-C\frac{n^{1-2\kappa-3\tau-\nu-\gamma''}}{\log(n)}\right)\right\} = O\left\{\exp\left(-C\frac{n^{1-2\kappa-3\tau-\nu-\gamma''}}{\log(n)}\right)\right\},
\end{aligned}$$

and then

$$\Pr\left\{\max_{j\notin S} R_j^2 > \frac{c_1 c_6 c'^2 n^{2-2\kappa-\tau+\gamma''}}{c_3 \log(n)}\right\} \leq O\left\{\exp\left(-C\frac{n^{1-2\kappa-3\tau-\nu-\gamma''}}{\log(n)}\right)\right\}.$$

Finally, we obtain the sure screening property of gAR2

$$\Pr\left\{\max_{j\notin S} R_j^2 < \gamma_n < \min_{j\in S} R_j^2\right\} \geq 1 - O\left\{\exp\left(-C\frac{n^{1-2\kappa-3\tau-\nu-\gamma''}}{\log(n)}\right)\right\},$$

where the thrshold γ_n satisfies

$$\frac{\gamma_n}{n^{2-2\kappa-\tau+\gamma''}} \to 0 \text{ and } \frac{\gamma_n \log(n)}{n^{2-2\kappa-\tau+\gamma''}} \to \infty.$$

The proof of Theorem 2.3 is completed.

D Proof of Theorem 2.4

We now study the screening property of screening method for groupwise distance correlation. The assumptions and proofs are very similar to those in Li, et al., (2012), except that now the response Y is univariate and the predictors $\mathbf{X}_j$ are multivariate variables with p_j dimensions, $p_j \geq 1, j = 1, \ldots, J$. To be more specifically, we switch the roles of multiple responses and univariate predictor that were used in the proofs of Li, et al., (2012). The necessary assumptions below are adapted from Li et al., (2012) but with a slight modification.

(D1) The univariate response Y and groups $\mathbf{X}_j$ satisfy the sub exponential tail probability uniformly in J, which follows

$$\sup_J \max_{1\leq j\leq J} E\left\{\exp(\alpha\|\mathbf{X}_j\|_{p_j}^2)\right\} < \infty,$$

and

$$E\left\{\exp(\alpha\|Y\|_1^2)\right\} < \infty,$$

where $2\log(2) < \alpha \leq 2\alpha_0$ and α_0 is a positive constant. Here, $\|\mathbf{X}_j\|_{p_j}$ is the L_2 norm with subscript p_j representing the dimensions of $\mathbf{X}_j$, which is slightly different from the notation in the previous section.

(D2) The minimum distance correlation between response and truly important groups satisfies

$$\min_{j\in S} \mathcal{D}_j \geq 2cn^{-\kappa'},$$

where c is a constant and $0 \leq \kappa' < 1/2$, S is the indices set of the true model $\mathcal{M}_S^g$ with size s.

In practice, assumption (D1) can be immediately satisfied when Y and $\mathbf{X}_j$ have multivariate normal distribution or are uniformly bounded as pointed out by Li, et al., (2012). Assumption (D2) implies the groupwise distance correlation between the response and truly important groups should be greater than some constants, otherwise it is very difficult to identify the truly important groups at the boundary between truly unimportant groups. In assumption (D2), κ' controls the decrease rate for $\mathcal{D}_j$ towards 0.

Székely, et al., (2007) mentioned that the squares of distance covariance can be expressed by an algebraic identity, i.e.,

$$\mathrm{dcov}^2(\mathbf{u},\mathbf{v}) = S_1 + S_2 - 2S_3,$$

where $S_1 = E\{\|\mathbf{u}-\tilde{\mathbf{u}}\|_{d_u}\|\mathbf{v}-\tilde{\mathbf{v}}\|_{d_v}\}, S_2 = E\{\|\mathbf{u}-\tilde{\mathbf{u}}\|_{d_u}\}E\{\|\mathbf{v}-\tilde{\mathbf{v}}\|_{d_v}\}$ and $S_3 = E\{E\{\|\mathbf{u}-\tilde{\mathbf{u}}\|_{d_u}|\mathbf{u}\}E\{\|\mathbf{v}-\tilde{\mathbf{v}}\|_{d_v}|\mathbf{v}\}\}$, in which $(\mathbf{u},\mathbf{v})$ are two random vectors with dimensions d_u, d_v respectively, and $(\tilde{\mathbf{u}},\tilde{\mathbf{v}})$ are independent copy of $(\mathbf{u},\mathbf{v})$. We replace $(\mathbf{u},\mathbf{v})$ with $(Y,\mathbf{X}_j)$, and $(\tilde{\mathbf{u}},\tilde{\mathbf{v}})$ with $(\widetilde{Y},\widetilde{\mathbf{X}}_j)$ in our context to define $\mathrm{dcov}^2(\mathbf{X}_j,Y) = S_{j1} + S_{j2} - 2S_{j3}$, the distance

covariance between response and the j-th groups, where

$$S_{j1} = E(\|\mathbf{X}_j - \widetilde{\mathbf{X}}_j\|_{p_j}\|Y - \widetilde{Y}\|_1),$$

$$S_{j2} = E(\|\mathbf{X}_j - \widetilde{\mathbf{X}}_j\|_{p_j})E(\|Y - \widetilde{Y}\|_1),$$

$$S_{j3} = E\{E(\|\mathbf{X}_j - \widetilde{\mathbf{X}}_j\|_{p_j}|\mathbf{X}_j)E(\|Y - \widetilde{Y}\|_1|Y)\},$$

which immediately follows their respective sample parts

$$\widehat{S}_{j1} = \frac{1}{n^2}\sum\nolimits_{i,h=1}^{n} \|\mathbf{X}_{ij} - \mathbf{X}_{hj}\|_{p_j}\|y_i - y_h\|_1,$$

$$\widehat{S}_{j2} = \frac{1}{n^2}\sum\nolimits_{i,h=1}^{n} \|\mathbf{X}_{ij} - \mathbf{X}_{hj}\|_{p_j}\frac{1}{n^2}\sum_{i,h=1}^{n}\|y_i - y_h\|_1,$$

$$\widehat{S}_{j3} = \frac{1}{n^3}\sum\nolimits_{i,h,l=1}^{n} \|\mathbf{X}_{ij} - \mathbf{X}_{lj}\|_{p_j}\|y_h - y_l\|_1,$$

where $\mathbf{X}_{ij}, \mathbf{X}_{hj}, \mathbf{X}_{lj}$ are the i-th, h-th, l-th observations of the j-th groups $\mathbf{X}_j$, $i, h, l = 1, \ldots, n$. By the alternative representation of distance covariance, we have the following definition in the population sense,

$$\mathrm{dcov}^2(\mathbf{X}_j, Y) = S_{j1} + S_{j2} - 2S_{j3},$$

and its estimation in the sample sense,

$$\widehat{\mathrm{dcov}}^2(\mathbf{X}_j, Y) = \widehat{S}_{j1} + \widehat{S}_{j2} - 2\widehat{S}_{j3}.$$

Similar definitions can be applied for $\mathrm{dVar}^2(\mathbf{X}_j) = \mathrm{dcov}^2(\mathbf{X}_j, \mathbf{X}_j)$ and $\mathrm{dVar}^2(Y) = \mathrm{dcov}^2(Y, Y)$, as well as their sample counterparts $\widehat{\mathrm{dVar}}^2(\mathbf{X}_j), \widehat{\mathrm{dVar}}^2(Y)$. Therefore, the sample groupwise

distance correlation can be estimated by

$$\widehat{\mathrm{dcor}}(\mathbf{X}_j, Y) = \frac{\widehat{\mathrm{dcov}}(\mathbf{X}_j, Y)}{\sqrt{\widehat{\mathrm{dVar}}(\mathbf{X}_j) \times \widehat{\mathrm{dVar}}(Y)}}. \tag{D.1}$$

To show Theorem 2.4, we first prove

$$\Pr\left(\max_{1\le j\le J}\left|\widehat{\mathcal{D}}_j - \mathcal{D}_j\right| \ge cn^{-\kappa'}\right) \le O\left(J\left[\exp\left\{-c_1' n^{1-2(\kappa'+\gamma')}\right\} + n\exp\left(-c_2' n^{\gamma'}\right)\right]\right)$$

for some positive constants c_1' and c_2', where $0 < \gamma' < 1/2 - \kappa'$. To achieve this, we need to show the uniform consistency of the denominator and the numerator of $\widehat{\mathrm{dcor}}(\mathbf{X}_j, Y)$ defined in (D.1), respectively. To complete the proofs, the following three lemmas will be used in the subsequence.

Lemma D.1. *(Lemma 5.6.1.A, Serfling (1980)) Let $\mu = E(Z)$. If $\Pr(a \le Z \le b) = 1$, then*

$$E[\exp\{\alpha(Z-\mu)\}] \le \exp\{\alpha^2(b-a)^2/8\}$$

for any $\alpha > 0$.

Lemma D.2. *(Theorem 5.6.1.A, Serfling (1980)) Let $h(Z_1, \ldots, Z_m)$ be a kernel of the U statistics U_n, and $\theta = E\{h(Z_1, \ldots, Z_m)\}$. If $a \le h(Z_1, \ldots, Z_m) \le b$, then for any $t > 0$ and $n \ge m$,*

$$\Pr(U_n - \theta \ge t) \le \exp\{-2[n/m]t^2/(b-a)^2\},$$

where $[n/m]$ denotes the integer part of n/m.

Lemma D.3. *(Section 5.1.6, Serfling (1980)) Consider a symmetric kernel $u(z_1, \ldots, z_m)$ and a sample $Z_1, \ldots, Z_n$ of size $n \ge m$. Define $r = [n/m]$ and*

$$W(z_1, \ldots, z_n) = \frac{u(z_1, \ldots, z_m) + u(z_{m+1}, \ldots, z_{2m} + \ldots + u(z_{rm-m+1}, \ldots, z_{rm}))}{r}.$$

Letting $\sum_{n!}$ *denote summation over all* $n!$ *permutations* $(i_1, \ldots, i_n)$ *of* $(1, \ldots, n)$*. Then any* U *statistics* U_n *can be represented as an average of* $n!$ *terms, each of which is itself an average of* r *i.i.d random variables. That is,*

$$U_n = \frac{1}{n!} \sum\nolimits_{n!} W(Z_{i_1}, \ldots, Z_{i_n}).$$

We first deal with the numerator of $\widehat{\text{dcor}}^2(\mathbf{X}_j, Y)$ which is $\widehat{\text{dcov}}^2(\mathbf{X}_j, Y)$. More specially, we show the uniform consistency for $\widehat{S}_{j1}, \widehat{S}_{j2}, \widehat{S}_{j3}$ respectively.

Corollary D.1. *(uniform consistency of* $\widehat{S}_{j1}$*) Under assumptions (D1) and (D2), there exists some constants* c *and* C*, such that for any* $\varepsilon > 0$*,*

$$\Pr\left(|\widehat{S}_{j1} - S_{j1}| \geq \varepsilon\right) \leq 2\exp(-\varepsilon^2 n^{1-2\gamma'}) + 2nC\exp(-\alpha c n^{\gamma'}/4),$$

where $0 < \gamma' < \kappa' - 1/2$*,* κ' *and* α *are defined in assumptions (D1) and (D2).*

Proof: Note that by using Cauchy-Schwartz inequality, we have

$$\begin{aligned} S_{j1} &= E(\|\mathbf{X}_{ij} - \mathbf{X}_{hj}\|_{p_j} \|y_i - y_h\|_1) \leq \left\{E(\|\mathbf{X}_{ij} - \mathbf{X}_{hj}\|_{p_j}^2) E(\|y_i - y_h\|_1^2)\right\}^{1/2} \\ &\leq \left\{E(2^2\|\mathbf{X}_j\|_{p_j}^2) E(2^2\|y_i\|^2)\right\}^{1/2} = 4\left\{E(\|\mathbf{X}_j\|_{p_j}^2) E(y_i^2)\right\}^{1/2}. \end{aligned}$$

Since $E(\|\mathbf{X}_j\|_{p_j}^2)E(y_i^2) \leq E\left(\exp\left\{\alpha\|\mathbf{X}_j\|_{p_j}^2\right\}\right) E\left(\exp\left\{\alpha y_i^2\right\}\right)$ for $\alpha > 2\log(2)$, by assumption (D1), we conclude that S_{j1} is uniformly bounded in J, i.e., $\sup_J \max_{1\leq j\leq J} S_{j1} < \infty$. This implies that for any given $\varepsilon > 0$, we have $S_{j1}/n < \varepsilon$ as n large enough. We define a U statistic $\widehat{S}_{j1}^* = \{n(n-1)\}^{-1} \sum_{i\neq h} \|\mathbf{X}_{ij} - \mathbf{X}_{hj}\|_{p_j} \|y_i - y_h\|_1$, such that $\left|\widehat{S}_{j1} - S_{j1}\right| =$

$\left|\widehat{S}^*_{j1}(n-1)/n - S_{j1}(n-1)/n - S_{j1}/n\right|$. Therefore, we have by Bonferroni's inequality

$$\begin{aligned}
\Pr\left(\left|\widehat{S}_{j1} - S_{j1}\right| \geq 2\varepsilon\right) &= \Pr\left(\left|\frac{n-1}{n}\widehat{S}^*_{j1} - \frac{n-1}{n}S_{j1} - \frac{1}{n}S_{j1}\right| \geq 2\varepsilon\right) \\
&\leq \Pr\left(\left|\frac{n-1}{n}\widehat{S}^*_{j1} - \frac{n-1}{n}S_{j1}\right| \geq 2\varepsilon - \frac{1}{n}S_{j1}\right) \\
&\leq \Pr\left(\frac{n-1}{n}\left|\widehat{S}^*_{j1} - S_{j1}\right| \geq \varepsilon\right) \leq \Pr\left(\left|\widehat{S}^*_{j1} - S_{j1}\right| | \geq \varepsilon\right).
\end{aligned}$$

Thus, if $\widehat{S}^*_{j1}$ is uniformly consistent, it is concluded that $\widehat{S}_{j1}$ is also uniformly consistent. Now we define

$$\begin{aligned}
\widehat{S}^*_{j1} &= \frac{1}{n(n-1)}\sum_{i\neq h} u_1(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h) \times \mathbf{1}\{u_1(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h) \leq M\} + \\
&\quad \frac{1}{n(n-1)}\sum_{i\neq h} u_1(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h) \times \mathbf{1}\{u_1(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h) > M\} \\
&=: \quad \widehat{S}^*_{j1,1} + \widehat{S}^*_{j1,2},
\end{aligned}$$

where $u_1(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h) = \|\mathbf{X}_{ij} - \mathbf{X}_{hj}\|_{p_j}\|y_i - y_h\|_1$ that is the kernel of the U statistic $\widehat{S}^*_{j1}$, $\mathbf{1}\{\cdot\}$ is an indicator function. This follows that S_{j1} can be also decomposed as

$$\begin{aligned}
S_{j1} &= E(u_1(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h)\mathbf{1}\{u_1(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h) \leq M\}) + \\
&\quad E(u_1(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h)\mathbf{1}\{u_1(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h) > M\}) \\
&=: \quad S_{j1,1} + S_{j1,2},
\end{aligned}$$

where M is a finite constant and will be specified later.

Let us first focus on the consistency of $\widehat{S}^*_{j1,1}$. By Markov's inequality, we have for any $t > 0$,

$$\begin{aligned}
\Pr\left(\widehat{S}^*_{j1,1} - S_{j1,1} \geq \varepsilon\right)) &= \Pr\left\{\exp\left[t\left(\widehat{S}^*_{j1,1} - S_{j1,1}\right)\right] \geq \exp(t\varepsilon)\right\} \\
&\leq \exp(-t\varepsilon)\exp(-tS_{j1,1})E\left\{\exp\left(t\widehat{S}^*_{j1,1}\right)\right\}.
\end{aligned}$$

By Lemma D.3, the U statistics $\widehat{S}^*_{j1,1}$ can be expressed as

$$\widehat{S}^*_{j1,1} = \frac{1}{n!}\sum_{n!} W_1(\mathbf{X}_{1j}, y_1; \ldots; \mathbf{X}_{nj}, y_n),$$

where each $W_1(\mathbf{X}_{1j}, y_1; \ldots; \mathbf{X}_{nj}, y_n)$ is an average of $m = [n/2]$ i.i.d random variables $u_1^{(1)}\mathbf{1}(u_1^{(1)} \leq M), \ldots, u_1^{(m)}\mathbf{1}(u_1^{(m)} \leq M)$, i.e., $W_1(\mathbf{X}_{1j}, y_1; \ldots; \mathbf{X}_{nj}, y_n) = m^{-1}\sum_{r=1}^m u_1^{(r)}\mathbf{1}\left\{u_1^{(r)} \leq M\right\}$. Therefore, by the fact that exponential function is convex and $(\mathbf{X}_{1j}, y_1), \ldots, (\mathbf{X}_{nj}, y_n)$ are i.i.d random variables, we have by Jensen's inequality (i.e., $\exp(n^{-1}\sum_{i=1}^n x_i) \leq n^{-1}\sum_{i=1}^n \exp(x_i)$)

$$\begin{aligned} E\left\{\exp\left(t\widehat{S}^*_{j1,1}\right)\right\} &= E\left[\exp\left\{t\frac{1}{n!}\sum_{n!} W_1(\mathbf{X}_{1j}, y_1; \ldots; \mathbf{X}_{nj})\right\}\right] \\ &\leq \frac{1}{n!}E[\exp\{tW_1(\mathbf{X}_{1j}, y_1; \ldots; \mathbf{X}_{nj})\}] \\ &= E^m\left\{\exp\left(m^{-1}tu_1^{(r)}\mathbf{1}\left(u_1^{(r)} \leq M\right)\right)\right\}. \end{aligned}$$

By Lemma D.1, $E\left(u_1^{(r)}\mathbf{1}\left(u_1^{(r)} \leq M\right)\right) = S_{j1,1}$ and $\Pr\left\{0 < u_1^{(r)}\mathbf{1}\left(u_1^{(r)} \leq M\right) \leq M\right\} = 1$, we can obtain

$$\begin{aligned} \exp(-tS_{j1,1})E\left\{\exp\left(t\widehat{S}^*_{j1,1}\right)\right\} &\leq \exp(-tS_{j1,1})E^m\left\{\exp\left(m^{-1}tu_1^{(r)}\mathbf{1}\left(u_1^{(r)} \leq M\right)\right)\right\} \\ &\leq E^m\left\{\exp\left[m^{-1}tu_1^{(r)}\mathbf{1}\left(u_1^{(r)} \leq M - S_{j1,1}\right)\right]\right\} \\ &\leq \exp\left\{\frac{t^2(M-0)^2}{8m}\right\} = \exp\left\{\frac{t^2M^2}{8m}\right\}. \end{aligned}$$

Consequently, by setting $t = 4\varepsilon m/M^2$, we have

$$\Pr\left(\widehat{S}^*_{j1,1} - S_{j1,1} \geq \varepsilon\right) \leq \exp(-t\varepsilon)\exp\left(\frac{M^2t^2}{8m}\right) = \exp\left(-t\varepsilon + \frac{M^2t^2}{8m}\right) = \exp\left(-\frac{2\varepsilon^2 m}{M^2}\right).$$

By the symmetric of U statistics, it entails immediately that

$$\Pr\left(\left|\widehat{S}^*_{j1,1} - S_{j1,1}\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2\varepsilon^2 m}{M^2}\right). \tag{D.2}$$

Now let us turn to the consistency of $\widehat{S}^*_{j1,2}$. Applying Cauchy-Schwartz and Markov's inequality again, we have for any $\alpha' > 0$

$$\begin{aligned}
S^2_{j1,2} &= E[u_1(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h)\mathbf{1}\{u_1(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h) \geq M\}] \\
&\leq E\{u_1^2(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h)\}E\{\mathbf{1}\{u_1(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h) \geq M\} \\
&= E\{u_1^2(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h)\}\Pr\{\exp[\alpha' u_1(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h)] \geq \exp(\alpha' M)\} \\
&\leq E\{u_1^2(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h)\}E[\exp\{\alpha' u_1(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h)\}]\exp(-\alpha' M).
\end{aligned}$$

With the inequality $(a^2 + b^2)/2 \geq (a+b)^2/4 \geq |ab|$, one has

$$\begin{aligned}
u_1(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h) &= \{(\mathbf{X}_{ij} - \mathbf{X}_{hj})^T(\mathbf{X}_{ij} - \mathbf{X}_{hj})(y_i - y_h)^2\}^{1/2} \\
&\leq \{2(\|\mathbf{X}_{ij}\|^2_{p_j} + \|\mathbf{X}_{hj}\|^2_{p_j}) \times 2(y_i^2 + y_h^2)\}^{1/2} \\
&\leq \{\|\mathbf{X}_{ij}\|^2_{p_j} + \|\mathbf{X}_{hj}\|^2_{p_j} + y_i^2 + y_h^2\}^{1/2} \\
&= \|\mathbf{X}_{ij}\|^2_{p_j} + \|\mathbf{X}_{hj}\|^2_{p_j} + y_i^2 + y_h^2,
\end{aligned}$$

which follows by Cauchy-Schwartz inequality, together with assumption (D1), for any $\log(2) \leq \alpha' \leq \alpha_0$, where α_0 is defined in assumption (D1),

$$\begin{aligned}
E[\exp\{\alpha' u_1(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h)\}] &\leq E[\exp\{\alpha'(\|\mathbf{X}_{ij}\|^2_{p_j} + \|\mathbf{X}_{hj}\|^2_{p_j} + y_i^2 + y_h^2)\}] \\
&= E[\exp\{\alpha'(\|\mathbf{X}_{ij}\|^2_{p_j} + \|\mathbf{X}_{hj}\|^2_{p_j}\}]E[\exp\{\alpha'(y_i^2 + y_h^2)\}] \\
&\leq E[\exp\{2\alpha'\|\mathbf{X}_{ij}\|^2_{p_j}\}]E[\exp\{2\alpha' y_i^2\}] < \infty.
\end{aligned}$$

Also, by the inequality $\exp(x) > x^2/2$ for $x > 0$, one can obtain

$$\begin{aligned} E\{u_1^2(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h)\} &\leq E[2\exp\{u_1(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h)\}] \\ &\leq E[\exp\{\alpha' u_1(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h)\}] < \infty. \end{aligned}$$

Therefore, choosing $M = cn^{\gamma'}$ for $0 < \gamma' < 1/2 - \kappa'$ and $\log(2) \leq \alpha' \leq \alpha_0$, it is concluded that $S_{j1,2} \leq \varepsilon/2$ as n is large enough. So one can easily obtain that

$$\Pr\left(\left|\widehat{S}^*_{j1,2} - S_{j1,2}\right| > \varepsilon\right) \leq \Pr\left(\left|\widehat{S}^*_{j1,2}\right| > \varepsilon - S_{j1,2}\right) = \Pr\left(\left|\widehat{S}^*_{j1,2}\right| > \frac{\varepsilon}{2}\right).$$

If we can bound the probability $\Pr\left(\left|\widehat{S}^*_{j1,2}\right| > \varepsilon/2\right)$, then $\widehat{S}^*_{j1,2}$ is uniformly consistent. Assume $\|\mathbf{X}_{ij}\|^2_{p_j} + y_i^2 \leq M/2$, which implies $u_1(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h) \leq \|\mathbf{X}_{ij}\|^2_{p_j} + \|\mathbf{X}_{hj}\|^2_{p_j} + y_i^2 + y_h^2 \leq M$. This entails $\left|\widehat{S}^*_{j1,2}\right| = 0$ by its definition. This assumption implies that if $\left|\widehat{S}^*_{j1,2}\right| > 0$, we have $\|\mathbf{X}_{ij}\|^2_{p_j} + y_i^2 > M/2$, which is equivalent to the events $\left\{\left|\widehat{S}^*_{j1,2}\right| > \varepsilon/2\right\} \subseteq \left\{\|\mathbf{X}_{ij}\|^2_{p_j} + y_i^2 > M/2\right\}$ for any $\varepsilon > 0$ and $j = 1, \ldots, J$. Also, observe that the events satisfy $\{\|\mathbf{X}_{ij}\|^2_{p_j} + y_i^2 > M/2\} \subseteq \{\|\mathbf{X}_{ij}\|^2_{p_j} > M/4\} \cup \{y_i^2 > M/4\}$. Consequently, by Markov's inequality and Bonferroni's inequality, and assumption (D1), there exists a constant C such that

$$\begin{aligned} \max_{1\leq j\leq J} \Pr\left(\left|\widehat{S}^*_{j1,2}\right| > \frac{\varepsilon}{2}\right) &\leq n \max_{1\leq j\leq J} \Pr\left(\|\mathbf{X}_{ij}\|^2_{p_j} + y_i^2 > \frac{M}{2}\right) \\ &\leq n \max_{1\leq j\leq J} \left\{\Pr\left(\|\mathbf{X}_{ij}\|^2_{p_j} > \frac{M}{4}\right) + \Pr\left(y_i^2 > \frac{M}{4}\right)\right\} \\ &\leq n \max_{1\leq j\leq J} \left\{E\{\exp(\alpha\|\mathbf{X}_{ij}\|^2_{p_j}) + E\{\exp(\alpha y_i^2)\}\right\} \times \exp\left(-\frac{\alpha M}{4}\right) \\ &\leq 2nC\exp\left(-\frac{\alpha M}{4}\right) = 2nC\exp\left(-\frac{\alpha c n^{\gamma'}}{4}\right). \end{aligned} \tag{D.3}$$

Therefore, $\Pr\left(\left|\widehat{S}^*_{j1,2} - S_{j1,2}\right| > \varepsilon\right) \leq 2nC\exp(-\alpha c n^{\gamma'}/4)$. Combining (D.2) and (D.3), re-

calling $m = [n/2]$ and $M = cn^{\gamma'}$, one has

$$\Pr\left(\left|\widehat{S}_{j1} - S_{j1}\right| \geq \varepsilon\right) \leq 2\exp\left(-\varepsilon^2 n^{1-2\gamma'}\right) + 2nC\exp\left(-\frac{\alpha c n^{\gamma'}}{4}\right).$$

Thus, the proof of the uniform consistency of $\widehat{S}_{j1}$ is completed. ■

Next we show the uniform consistency of $\widehat{S}_{j2}$.

Corollary D.2. *(uniform consistency of $\widehat{S}_{j2}$) Under assumptions (D1) and (D2), there exists a sufficiently large constant C, such that for any $\varepsilon > 0$*

$$\Pr\left(\left|\widehat{S}_{j2} - S_{j2}\right| > \varepsilon\right) \leq 8\exp\left\{-\frac{\varepsilon^2 n^{1-2\gamma'}}{C^2}\right\} + 8nC\exp\left(-\frac{\alpha n^{\gamma'}}{4}\right),$$

where γ and s are defined in Corollary D.1.

Proof: Define $\widehat{S}_{j2,1} = n^{-2}\sum_{i\neq h}\|\mathbf{X}_{ij} - \mathbf{X}_{hj}\|_{p_j}$ and $\widehat{S}_{j2,2} = n^{-2}\sum_{i\neq h}\|y_i - y_h\|_1$ so that one has $\widehat{S}_{j2} = \widehat{S}_{j2,1}\widehat{S}_{j2,2}$. Also, define $S_{j2,1} = E\{\|\mathbf{X}_{ij} - \mathbf{X}_{hj}\|_{p_j}\}$ and $S_{j2,2} = E\{\|y_i - y_h\|_1\}$, leading to $S_{j2} = S_{j2,1}S_{j2,2}$. Note that

$$\widehat{S}_{j2} - S_{j2} = \left(\widehat{S}_{j2,1} - S_{j2,1}\right)S_{j2,2} + S_{j2,1}\left(\widehat{S}_{j2,2} - S_{j2,2}\right) + \left(\widehat{S}_{j2,1} - S_{j2,1}\right)\left(\widehat{S}_{j2,2} - S_{j2,2}\right).$$

To see this, check the equality $ab - cd = (a-c)d + c(b-d) + (a-c)(b-d)$.

We first prove that $\widehat{S}_{j2,1}$ (resp. $\widehat{S}_{j2,2}$) is a consistent estimator of $S_{j2,1}$ (resp. $S_{j2,2}$). Since $\widehat{S}_{j2,1}$ is a special case of $\widehat{S}_{j1}$ by setting $|y_i - y_h| = 1$ that is indeed uniformly bounded and satisfying assumption (D1). Following the same arguments in Corollary D.1, one can obtain

$$\Pr\left(\left|\widehat{S}_{j2,1} - S_{j2,1}\right| \geq 4\varepsilon\right) \leq 2\exp(-\varepsilon^2 n^{1-2\gamma'}) + 2nC\exp\left(-\frac{\alpha n^{\gamma'}}{4}\right),$$

Similarly, one can also obtain

$$\Pr\left(\left|\widehat{S}_{j2,2} - S_{j2,2}\right| \geq 4\varepsilon\right) \leq 2\exp\left(-\varepsilon^2 n^{1-2\gamma'}\right) + 2nC\exp\left(-\frac{\alpha n^{\gamma'}}{4}\right),$$

which is another special case of Corollary D.1 by setting $\|\mathbf{X}_{ij} - \mathbf{X}_{hj}\|_{p_j} = 1$. Accordingly, by assumption (D1), one has by using inequality $(EX)^2 \leq EX^2$,

$$S_{j2,1} \leq \left\{E(\|\mathbf{X}_{ij} - \mathbf{X}_{hj}\|_{p_j}^2)\right\}^{1/2} \leq \left\{4E(\|\mathbf{X}_{p_j}\|_{p_j}^2)\right\}^{1/2} < \infty \tag{D.4}$$

and

$$S_{j2,2} \leq \{E(|y_i - y_h|^2)\}^{1/2} \leq \{4E(y_i^2)\}^{1/2} < \infty. \tag{D.5}$$

This implies that for some constant C, $S_{j2,1}, S_{j2,2}$ are uniformly bounded in J, i.e.,

$$\max_{1\leq j\leq J}\{\max_{1\leq j\leq J} S_{j2,1}, S_{j2,2}\} \leq C. \tag{D.6}$$

Finally, following by (D.4) and (D.5), one has

$$\Pr\left\{\left|\left(\widehat{S}_{j2,1} - S_{j2,1}\right)\left(\widehat{S}_{j2,2} - S_{j2,2}\right)\right| \geq \varepsilon\right\} \leq 4\exp\left(-\frac{\varepsilon n^{1-2\gamma'}}{16}\right) + 4nC\exp\left(-\frac{\alpha n^{\gamma'}}{4}\right). \tag{D.7}$$

Combining (D.4), (D.5), (D.6) and (D.7), one can obtain

$$\begin{aligned}
\Pr\left(\left|\widehat{S}_{j2} - S_{j2}\right| \geq \varepsilon\right) &= \Pr\left(\left|\widehat{S}_{j2,1}\widehat{S}_{j2,2} - S_{j2,1}S_{j2,2}\right| \geq \varepsilon\right) \\
&\leq \Pr\left\{\left|\left(\widehat{S}_{j2,1} - S_{j2,1}\right)S_{j2,2}\right| \geq \frac{\varepsilon}{3}\right\} + \Pr\left\{\left|S_{j2,1}\left(\widehat{S}_{j2,2} - S_{j2,2}\right)\right| \geq \frac{\varepsilon}{3}\right\} + \\
&\quad \Pr\left\{\left|\left(\widehat{S}_{j2,1} - S_{j2,1}\right)\left(\widehat{S}_{j2,2} - S_{j2,2}\right)\right| \geq \frac{\varepsilon}{3}\right\} \\
&\leq 8\exp\left\{-\frac{\varepsilon^2 n^{1-2\gamma'}}{C^2}\right\} + 8nC\exp\left(-\frac{\alpha n^{\gamma'}}{4}\right).
\end{aligned} \tag{D.8}$$

Thus, the proof of the uniform consistency of $\widehat{S}_{j2}$ is complete. ■

Finally, we show the uniform consistency of $\widehat{S}_{j3}$.

Corollary D.3. *(uniform consistency of $\widehat{S}_{j3}$) Under assumptions (D1) and (D2), there exists a constant C, for any $\varepsilon > 0$,*

$$\Pr\left(\left|\widehat{S}_{j3} - S_{j3}\right| \geq \varepsilon\right) \leq 4\exp\left(-\frac{\varepsilon^2 n^{1-2\gamma'}}{6}\right) + 4nC\exp\left(-\frac{\alpha n^{\gamma'}}{4}\right),$$

where γ', α are defined in Corollary D.1.

Proof: Denote another U statistics

$$\begin{aligned}
\widehat{S}^*_{j3} &= \frac{1}{n(n-1)(n-2)} \sum_{i<h<l} \{\|\mathbf{X}_{ij} - \mathbf{X}_{hj}\|_{p_j}\|y_h - y_l\|_1 + \\
&\quad \|\mathbf{X}_{ij} - \mathbf{X}_{lj}\|_{p_j}\|y_h - y_l\|_1 + \|\mathbf{X}_{ij} - \mathbf{X}_{hj}\|_{p_j}\|y_i - y_l\|_1 + \\
&\quad \|\mathbf{X}_{lj} - \mathbf{X}_{hj}\|_{p_j}\|y_h - y_l\|_1 + ||\mathbf{X}_{ij} - \mathbf{X}_{hj}\|_{p_j}||y_i - y_h\|_1 + \\
&\quad \|\mathbf{X}_{lj} - \mathbf{X}_{ij}\|_{p_j}\|y_i - y_h\|_1\} \\
&=: \quad \frac{6}{n(n-1)(n-2)} \sum_{i<h<l} u_3(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h; \mathbf{X}_{lj}, y_l),
\end{aligned}$$

where $u_3(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h; \mathbf{X}_{lj}, y_l)$ is the kernel of U statistics $\widehat{S}^*_{j3}$. Specifically, $\widehat{S}^*_{j3}$ excludes the following six cases from $\widehat{S}_{j3}$: $i = h, i = l, h = l$ and the symmetrics $h = i, l = i, l = h$. Note that $\widehat{S}^*_{j3}$ has a very similar form as $\widehat{S}^*_{j1}$ that can be divided into two parts bounded by a constant M. That is,

$$\begin{aligned}
\widehat{S}^*_{j3} &= \frac{6}{n(n-1)(n-2)} \sum_{i<h<l} u_3 \mathbf{1}(u_3 \leq M) + \frac{6}{n(n-1)(n-2)} \sum_{i<h<l} u_3 \mathbf{1}(u_3 > M) \\
&=: \quad \widehat{S}^*_{j3,1} + \widehat{S}^*_{j3,2}.
\end{aligned}$$

Accordingly, by the definition of S_{j3}, one has the corresponding decomposition

$$S_{j3} = E\{u_3 \mathbf{1}(u_3 \leq M)\} + E\{u_3 \mathbf{1}(u_3 > M)\} =: S_{j3,1} + S_{j3,2}.$$

Again, $\widehat{S}_{j3,1}, \widehat{S}_{j3,2}$ are unbiased estimator of $S_{j3,1}, S_{j3,2}$ respectively. Observe that $\widehat{S}^*_{j3,1}$ has the similar form as $\widehat{S}^*_{j1,1}$, except that now $\widehat{S}^*_{j3,1}$ is a third-order U statistics. Let $m' = [n/3]$, using the same arguments for proving $\widehat{S}^*_{j1,1}$, together with Lemma D.3, one can easily show that

$$\Pr\left(\left|\widehat{S}^*_{j3,1} - S_{j3,1}\right| \geq \varepsilon\right) \leq 2\exp(-2\varepsilon^2 m'^2). \tag{D.9}$$

Also, for $\widehat{S}^*_{j3,2}$, since

$$u_3(\mathbf{X}_{ij}, y_i; \mathbf{X}_{hj}, y_h; \mathbf{X}_{lj}, y_l) \leq \frac{4}{6}\left(\|\mathbf{X}_{ij}\|^2_{p_j} + \|\mathbf{X}_{hj}\|^2_{p_j} + \|\mathbf{X}_{lj}\|^2_{p_j} + y_i^2 + y_h^2 + y_l^2\right),$$

one has the following result which is similar to that of $\widehat{S}^*_{j1,2}$,

$$\Pr\left(\left|\widehat{S}^*_{j3,2} - S_{j3,2}\right| > \varepsilon\right) \leq \Pr\left(\left|\widehat{S}^*_{j3,2}\right| > \frac{\varepsilon}{2}\right).$$

Additionally, for any $\varepsilon > 0$, one has the similar events as $\widehat{S}^*_{j1,2}$ that satisfy

$$\left\{\left|\widehat{S}^*_{j3,2}\right| > \frac{\varepsilon}{2}\right\} \subseteq \left\{\|\mathbf{X}_{ij}\|^2_{p_j} + y_i^2 > \frac{M}{2}\right\},$$

which implies $\Pr\left(\left|\widehat{S}^*_{j3,2}\right| > \varepsilon/2\right) \leq 2nC\exp(-\alpha M/4)$ by the same arguments in (D.3). Therefore,

$$\Pr\left(\left|\widehat{S}^*_{j3,2} - S_{j3,2}\right| > \varepsilon\right) \leq 2nC\exp\left(-\frac{\alpha M}{4}\right). \tag{D.10}$$

Let $M = cn^{\gamma'}$ for some $0 < \gamma' < 1/2 - \kappa'$ again. One has by (D.9) and (D.10)

$$\begin{aligned}
\Pr\left(\left|\widehat{S}^*_{j3} - S_{j3}\right| \geq \varepsilon\right)) &\leq \Pr\left(\left|\widehat{S}^*_{j3,1} - S_{j3,1}\right| \geq \frac{\varepsilon}{2}\right) + \Pr\left(\left|\widehat{S}^*_{j3,2} - S_{j3,2}\right| \geq \frac{\varepsilon}{2}\right) \\
&\leq 2\exp\left(-\frac{\varepsilon^2 n^{1-2\gamma'}}{2}\right) + 2nC\exp\left(-\frac{\alpha n^{\gamma'}}{4}\right).
\end{aligned} \tag{D.11}$$

By the definition of $\widehat{S}^*_{j3}$ and $\widehat{S}^*_{j1}$, $\widehat{S}_{j3}$ can be written as

$$\widehat{S}_{j3} = \frac{(n-1)(n-2)}{n^2}\left\{\widehat{S}^*_{j3} + \frac{1}{(n-2)}\widehat{S}^*_{j1}\right\}.$$

Thus, one has

$$\widehat{S}_{j3} - S_{j3} = \frac{(n-1)(n-2)}{n^2}\left(\widehat{S}^*_{j3} - S_{j3}\right) - \frac{3n-2}{n^2}S_{j3} + \frac{n-1}{n^2}\left(\widehat{S}^*_{j1} - S_{j1}\right) + \frac{n-1}{n^2}S_{j1}.$$

It is easily to see that $\{(n-1)/n^2\}S_{j1} < S_{j1}/n \leq \varepsilon$. To prove the boundedness of S_{j3}, we use the properties of conditional expectation: $E(E(Z|\mathcal{A})) = E(Z)$ and $[E(Z|\mathcal{A})]^2 \leq E(Z^2|\mathcal{A})$ if $EZ^2 < \infty$, where $\mathcal{A}$ is a sub σ-algebra of a set of events $\mathcal{F}$. By the definition of S_{j3}, with Cauchy-Schwartz inequality, one has

$$\begin{aligned}
S_{j3} &\leq \left\{E\left\{E\left(\|\mathbf{X}_{ij} - \widetilde{\mathbf{X}}_{ij}\|_{p_j}|\mathbf{X}_{ij}\right)\right\}^2 \times E\left\{E\left(\|Y - \widetilde{Y}\|_1|Y\right)\right\}^2\right\}^{1/2} \\
&\leq \left\{E\left[E\left(\|\mathbf{X}_{ij} - \widetilde{\mathbf{X}}_{ij}\|^2_{p_j}|\mathbf{X}_{ij}\right)\right] \times E\left[E\left(\|Y - \widetilde{Y}\|_1|^2Y\right)\right]\right\}^{1/2} \\
&= \left\{E\|\mathbf{X}_{ij} - \widetilde{\mathbf{X}}_{ij}\|^2_{p_j} E\|Y - \widetilde{Y}\|^2_1\right\}^{1/2} \leq 4\left\{E(\|\mathbf{X}_{ij}\|^2_{p_j})E(\|Y\|^2_1)\right\}^{1/2} < \infty,
\end{aligned}$$

leading to $\{(3n-2)/n^2\}S_{j3} < 3S_{j3}/n \leq \varepsilon$ for any given $\varepsilon > 0$ by assumption (D1). Consequently,

$$\begin{aligned}
\Pr\left(\left|\widehat{S}_{j3} - S_{j3}\right| \geq 4\varepsilon\right) &\leq \Pr\left\{\frac{(n-1)(n-2)}{n^2}\left|\widehat{S}^*_{j3} - S_{j3}\right| \geq \varepsilon\right\} + \Pr\left\{\frac{3n-2}{n^2}|S_{j3}| \geq \varepsilon\right\} + \\
&\quad \Pr\left\{\frac{n-2}{n^2}\left|\widehat{S}^*_{j1} - S_{j1}\right| \geq \varepsilon\right\} + \Pr\left\{\frac{n-1}{n^2}|S_{j1}| \geq \varepsilon\right\} \\
&\leq \Pr\left(\left|\widehat{S}^*_{j3} - S_{j3}\right| \geq \varepsilon\right) + \Pr\left(\left|\widehat{S}^*_{j1} - S_{j1}\right| \geq \varepsilon\right) \\
&\leq 4\exp\left(-\frac{\varepsilon^2 n^{1-2\gamma'}}{2}\right) + 4nC\exp\left(-\frac{\alpha n^{\gamma'}}{4}\right).
\end{aligned}$$

Thus, the proof of the uniform consistency of $\widehat{S}_{j3}$ is completed. ■

Now we have the fact that

$$\begin{aligned}
& \Pr\left\{\left|\widehat{\mathrm{dcov}}^2(\mathbf{X}_j, Y) - \mathrm{dcov}^2(\mathbf{X}_j, Y)\right| \geq \varepsilon\right\} \\
= \; & \Pr\left\{\left|\left(\widehat{S}_{j1} + \widehat{S}_{j2} - 2\widehat{S}_{j3}\right) - (S_{j1} + S_{j2} - 2S_{j3})\right| \geq \varepsilon\right\} \\
= \; & \Pr\left\{\left|\left(\widehat{S}_{j1} - S_{j1}\right) + \left(\widehat{S}_{j2} - S_{j2}\right) - 2\left(\widehat{S}_{j3} - S_{j3}\right)\right| \geq \varepsilon\right\} \\
\leq \; & \Pr\left\{\left|\widehat{S}_{j1} - S_{j1}\right| + \left|\widehat{S}_{j2} - S_{j2}\right| + 2\left|\widehat{S}_{j3} - S_{j3}\right| \geq \varepsilon\right\} \\
\leq \; & \Pr\left(\left|\widehat{S}_{j1} - S_{j1}\right| \geq \frac{\varepsilon}{4}\right) + \Pr\left(\left|\widehat{S}_{j2} - S_{j2}\right| \geq \frac{\varepsilon}{4}\right) + \Pr\left(\left|\widehat{S}_{j3} - S_{j3}\right| \geq \frac{\varepsilon}{4}\right).
\end{aligned}$$

Therefore, by Corollary D.1, D.2, and D.3, one can conclude that

$$\Pr\left\{\left|\widehat{\mathrm{dcov}}^2(\mathbf{X}_j, Y) - \mathrm{dcov}^2(\mathbf{X}_j, Y)\right| \geq \varepsilon\right\} = O\left\{\exp\left(-c_1'\varepsilon^2 n^{1-2\gamma'}\right) + n\exp\left(-c_2' n^{\gamma'}\right)\right\}, \tag{D.12}$$

for some positive constants c_1' and c_2'. So we complete the proofs for the convergence rate of the numerator of $\widehat{\mathcal{D}}_j$. Actually, we can use the same arguments to show the convergence rate of $\widehat{\mathrm{dVar}}^2(\mathbf{X}_j) = \widehat{\mathrm{dcov}}^2(\mathbf{X}_j, \mathbf{X}_j)$ and $\widehat{\mathrm{dVar}}^2(Y) = \widehat{\mathrm{dcov}}^2(Y, Y)$ which are special cases of $\widehat{\mathrm{dcov}}^2(\mathbf{X}_j, Y)$ when $\|\mathbf{X}_{ij} - \mathbf{X}_{hj}\|_{p_j}$ (resp. $\|y_i - y_h\|_1$) is replaced by $\|y_i - y_h\|_1$ (resp. $\|\mathbf{X}_{ij} - \mathbf{X}_{hj}\|_{p_j}$) that both are uniformly bounded satisfying the assumption (D1). One can easily show that the distance variances $\widehat{\mathrm{dVar}}^2(\mathbf{X}_j)$ and $\widehat{\mathrm{dVar}}^2(Y)$ have the same convergence rate as $\widehat{\mathrm{dcov}}^2(\mathbf{X}_j, Y)$ in (D.12). It entails immediately that the denominator of $\widehat{\mathcal{D}}_j$ has the same convergence rate as that of the numerator. That is,

$$\Pr\left\{\left|\widehat{\mathrm{dVar}}(\mathbf{X}_j)\widehat{\mathrm{dVar}}(Y) - \mathrm{dVar}(\mathbf{X}_j)\,\mathrm{dVar}(Y)\right| \geq \varepsilon\right\} = O\left\{\exp\left(-c_3'\varepsilon^2 n^{1-2\gamma'}\right) + n\exp\left(-c_4' n^{\gamma'}\right)\right\}, \tag{D.13}$$

for some positive constants c_3' and c_4'.

Observe that $|\left|\mathrm{dcov}^2(\mathbf{X}_j, Y)\right| \leq |S_{j1}| + |S_{j2}| + 2|S_{j3}| < \infty$ by using the results $S_{j1} < \infty, S_{j2} < \infty, S_{j3} < \infty$. So one also has $\mathrm{dVar}^2(\mathbf{X}_j) < \infty$ and $\mathrm{dVar}^2(Y) < \infty$. By using (D.13),

one can conclude that $\widehat{\mathrm{dVar}}(\mathbf{X}_j)\widehat{\mathrm{dVar}}(Y) = \mathrm{dVar}(\mathbf{X}_j)\,\mathrm{dVar}(Y) + O_P(1)$, which implies $1/(\widehat{\mathrm{dVar}}(\mathbf{X}_j)\widehat{\mathrm{dVar}}(Y)) = 1/(\mathrm{dVar}(\mathbf{X}_j)\,\mathrm{dVar}(Y))+O_P(1)$. Therefore, $1/(\widehat{\mathrm{dVar}}(\mathbf{X}_j)\widehat{\mathrm{dVar}}(Y)) = 1/(\mathrm{dVar}(\mathbf{X}_j)\,\mathrm{dVar}(Y)) + O_P(1) \leq C$ for a sufficiently large C. For notation simplicity, we denote $\widehat{X}_j \doteq \widehat{\mathrm{dcov}}^2(\mathbf{X}_j, Y)$, $X_j \doteq \mathrm{dcov}^2(\mathbf{X}_j, Y)$ and $\widehat{Y}_j \doteq \widehat{\mathrm{dVar}}(\mathbf{X}_j) \times \widehat{\mathrm{dVar}}(Y)$, $Y_j \doteq \mathrm{dVar}(\mathbf{X}_j)\,\mathrm{dVar}(Y)$. Note first that

$$\frac{\widehat{X}_j Y_j - \widehat{Y}_j X_j}{\widehat{Y}_j Y_j} = \frac{\left(\widehat{X}_j - X_j\right) Y_j - \left(\widehat{Y}_j - Y_j\right) X_j}{\widehat{Y}_j Y_j} = \frac{\widehat{X}_j - X_j}{\widehat{Y}_j} - \frac{\left(\widehat{Y}_j - Y_j\right) X_j}{\widehat{Y}_j Y_j}$$

and $\mathrm{dcor}^2(\mathbf{X}_j, Y) = X_j/Y_j \leq 1$, $\widehat{Y}_j^{-1} \leq C$. Now for the distance correlation $\mathcal{D}_j$, one has

$$\begin{aligned}
\Pr\left(\left|\widehat{\mathcal{D}}_j - \mathcal{D}_j\right| \geq \varepsilon\right) &= \Pr\left\{\left|\widehat{Y}_j^{-1}\left(\widehat{X}_j - X_j\right) - X_j\left(\widehat{Y}_j Y_j\right)^{-1}\left(\widehat{Y}_j - Y_j\right)\right| \geq \varepsilon\right\} \\
&\leq \Pr\left\{\left|\widehat{Y}_j^{-1}\left(\widehat{X}_j - X_j\right)\right| \geq \frac{\varepsilon}{2}\right\} + \Pr\left\{\left|(X_j/Y_j)\widehat{Y}_j^{-1}\left(\widehat{Y}_j - Y_j\right)\right| \geq \frac{\varepsilon}{2}\right\} \\
&\leq \Pr\left(\left|\widehat{X}_j - X_j\right| \geq \frac{\varepsilon}{2C}\right) + \Pr\left(\left|\widehat{Y}_j - Y_j\right| \geq \frac{\varepsilon}{2C}\right) \\
&= O\{\exp\left(-c_5'\varepsilon^2 n^{1-2\gamma'}\right) + n\exp\left(-c_6' n^{\gamma'}\right)\},
\end{aligned}$$

for some positive constants c_5' and c_6'. Let $\varepsilon = cn^{-\kappa'}$ for $0 < \kappa' < 1/2 - \gamma'$, which follows that

$$\begin{aligned}
\Pr\left\{\max_{1\leq j\leq J}\left|\widehat{\mathcal{D}}_j - \mathcal{D}_j\right| \geq cn^{-\kappa'}\right\} &\leq J\Pr\left\{\left|\widehat{\mathcal{D}}_j - \mathcal{D}_j\right| \geq cn^{-\kappa'}\right\} \\
&= O\left(J\left[\exp(-c_5'\varepsilon^2 n^{1-2\gamma'}) + n\exp(-c_6' n^{\gamma'})\right]\right) \\
&= O\left(J\left[\exp(-c_5' n^{1-2(\kappa'+\gamma')}) + n\exp(-c_6' n^{\gamma'})\right]\right).
\end{aligned}$$

We choose $\mathcal{M}_{\mathcal{D},DC}^g$ such that $\mathcal{M}_{\mathcal{D},DC}^g = \left\{1 \leq j \leq J : \widehat{\mathcal{D}}_j \geq cn^{-\kappa'}\right\}$. This indicates if $\mathcal{M}_S^g \nsubseteq \mathcal{M}_{\mathcal{D},DC}^g$, then for any $j \in S$, one has $\widehat{\mathcal{D}}_j < cn^{-\kappa'}$, following by $\left|\widehat{\mathcal{D}}_j - \mathcal{D}_j\right| \geq cn^{-\kappa'}$ with assumption (D2). So the events satisfy $\left\{\mathcal{M}_S^g \nsubseteq \mathcal{M}_{\mathcal{D},DC}^g\right\} \subseteq \left\{\left|\widehat{\mathcal{D}}_j - \mathcal{D}_j\right| \geq cn^{-\kappa'}\right\}$ for $j \in S$.

Equivalently, one has $\left\{\max_{j\in S}\left|\widehat{\mathcal{D}}_j-\mathcal{D}_j\right|<cn^{-\kappa'}\right\}\subseteq\{\mathcal{M}_S^g\subseteq\mathcal{M}_{\mathcal{D},DC}^g\}$. Therefore,

$$\begin{aligned}\Pr(\mathcal{M}_S^g\subseteq\mathcal{M}_{\mathcal{D},DC}^g) &\geq \Pr\left(\max_{j\in S}\left|\widehat{\mathcal{D}}_j-\mathcal{D}_j\right|<cn^{-\kappa'}\right)=1-\Pr\left(\min_{j\in S}\left|\widehat{\mathcal{D}}_j-\mathcal{D}_j\right|\geq cn^{-\kappa'}\right)\\ &= 1-s\Pr\left(\left|\widehat{\mathcal{D}}_j-\mathcal{D}_j\right|\geq cn^{-\kappa'}\right)\\ &= 1-O\left(s\left[\exp\left(-c_5'\varepsilon^2n^{1-2(\kappa'+\gamma')}\right)+n\exp(-c_6'n^{\gamma'})\right]\right).\end{aligned}$$

Therefore, we finish the proofs of Theorem 2.4.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd Int. Symp. Info. Theory*, Ed. B. N. Petrov and F. Csaki, pp. 267-281. Budapest: Akademia Kiado.

Almasy, L., Dyer, T., Peralta, J., Charlesworth, J., Curran, J. and Blangero, J. (2011). Genetic analysis workshop 17 mini-exome simulation. *BMC Proceedings*, **5** S2.

Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface*, **2**, 369-380.

Breheny, P. (2014). The group exponential lasso for bi-level variable selection. *Manuscript*.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, **37**, 373-384.

Efron, B., Johnstone, I., Hastie, T. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, **32**, 407-499.

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

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society B*, **70**, 849-911.

Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *The Annals of Statistics*, **36**, 2605-2637.

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

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

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

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

Huang, J., Breheny, P., Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistics Science*, **27**, 481-499.

Huang, J. and Zhang, T. (2010). The benefit of group sparsity. *The Annals of Statistics*, **38** 1978-2004. MR2676881.

Leng, C., Lin, Y. and Wahba, G. (2006). A note on lasso and related procedures in model selection. *Statistic Sinica*, **16**, 1273-84.

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

Meier, L., Van de Geer, S. and Bühlmann, P. (2008). The group Lasso for logistic regression. *Journal of the Royal Statistical Society Series B*, **70**, 53-71.

Patrick Breheny (2015), R package `grpreg`: Regularization Paths for Regression Models with Grouped Covariates. R Core Team.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461-4.

Sźekely, G. J., Rizzo,M. L., and Bakirov, N. K. (2007). Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics*, **35**, 2769-2794.

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

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

Wang, H. (2012). Factor profiled sure independence screening. *Biometrika*, **99**, 15-28.

Wang, X. and Leng, C. (2013). High-dimensional Ordinary Least-squares Projector for screening variables. *Manuscript*.

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

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

Zhao, D. and Li, Y. (2010). Principled sure independence for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis*, **105**, 397-411.

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

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, **7**, 2541-67.

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

Wang, L., Chen, G., Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, **23**, 1486.

Zhou, N. and Zhu, J. (2010). Group variable selection via hierarchical lasso and its oracle property. *Statistics and its Interface*, **3**, 557-574.

Zhu, L. P., Li, L., Li, R., and Zhu, L. X. (2011). Model-free feature screening for ultrahigh dimensional data. *Journal of the American Statistical Association*, **696**, 1464-1475.