STUDY ON CORRELATIONS IN HIGH DIMENSIONAL DATA

Siliang Gong

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2018

Approved by:

Yufeng Liu

Kai Zhang

Shankar Bhamidi

J. S. Marron

Donglin Zeng

ABSTRACT

Siliang Gong: Study on Correlations in High Dimensional Data
(Under the direction of Yufeng Liu and Kai Zhang)

With the prevalence of high dimensional data, variable selection is crucial in many real applications. Although various methods have been investigated in the past decades, challenges still remain when tens of thousands of predictor variables are available for modeling. One difficulty arises from the spurious correlation, referring to the phenomenon that the sample correlation between two variables can be large when the dimension is relatively high even if they are independent. While many classical variable selection methods choose a variable based upon its marginal correlation with the response, the existence of spurious correlation may result in a high false discovery rate. On the other hand, when important variables are highly correlated, it is desirable to include all of them into the model. However, there is no such guarantee in many existing methods. Another challenge is in most variable selection approaches one needs to implement model selection to control the model complexity. While cross-validation is commonly used, it is computationally expensive and lacks statistical interpretation. In this proposal, we introduce some novel variable selection approaches to address the challenges mentioned above. Our proposed methods are based upon the investigations on the limiting distribution of the spurious correlation. For the first project, we study the maximal absolute sample partial correlation between the covariates and the response, and introduce a testing-based variable selection procedure. In the second project, we take advantage of the asymptotic results of the maximal absolute sample correlation among covariates and incorporate them into a penalized variable selection approach. The third project considers applications of the asymptotic results in multiple-response regression. Numerical studies demonstrate the effectiveness of our proposed methods.

ACKNOWLEDGEMENTS

I would like to express deep gratitude to my Ph.D. advisors, Dr. Yufeng Liu and Dr. Kai Zhang. Without their support and encouragement, I would not have finished my dissertation work successfully. They are also wonderful researchers that gave me lots of suggestions for my future career development. Their passionate for research has always been an inspiration for me.

Thanks are also due to Dr. Di Wu, for her kind mentorship when I was working as a research assistant in her group. She gave me an opportunity to understand the role of statistics and statisticians in many other fields.

Besides that, I am grateful to my committee members, Dr. Steve Marron, Dr. Shankar Bhamidi and Dr. Donglin Zeng, for their helpful comments and suggestions on my dissertation.

I have enjoyed the time with my friends at UNC Chapel Hill. Some of them have graduated, while the others are still fighting for their degrees. It is their support that makes my Ph.D. studies colorful and memorable.

Last but not least, I would like to thank my family. Their love always encourages me to keep going after my dreams.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Variable selection in high dimensional problems

Due to the advance of technology, high dimensional data are prevalent in many different scientific disciplines, where the number of variables may be much larger than the sample size. For example, in biological studies, microarray data often contain the expressions of thousands of genes, where only tens of them are responsible for the phenotype. In financial markets, high-frequency financial data often include numerous variables that are influential to assets pricing. High-dimensional data are also frequently involved in health studies, nueroscience, economics, and many more.

For high dimensional data, linear model is one of the most commonly used models in practice. Consider the following linear model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\mathbf{y}$ is an $n \times 1$ response vector, X is an $n \times p$ design matrix, $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$ is a vector of unknown coefficients and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^T$ is a vector of random errors. Here $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ are i.i.d. mean zero random variables independent of the covariates.

Under the high dimensional setting where $p \gg n$, traditional modeling methods such as ordinary least squares (OLS) cannot be applied directly to (1.1). On the other hand, for datasets with a large number of candidate predictors, sometimes only a few of them are truly relevant to the response (Fan et al., 2014). In other words, the underlying true model is sparse, i.e., the number of non-zero parameters in $\boldsymbol{\beta}$, denoted by s, cannot be too large. Therefore, it is important to identify such variables, and this consideration makes variable selection crucial in high dimensional problems.

In the context of linear regression model (1.1), various variable selection procedures have been intensively investigated in the past decades. One well known example is best subset selection (c.f., Furnival and Wilson (2000)), which can be viewed as the solution to l_0-penalized least squares,

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_0, \tag{1.2}$$

where $\|\cdot\|_2$ and $\|\cdot\|_0$ denote the l_2 and l_0 norms respectively. Although best subset selection admits good sampling properties (Barron et al., 1999), the optimization problem in (1.2) is NP-hard, thus is computationally infeasible for high dimensional problems.

Variants of best subset selection methods include the forward stepwise regression (FSR), backward stepwise regression, stagewise regression (Hocking, 1976), ect. FSR starts with a null model, and adds one variable at a time such that the residual sum of squares is minimized. Backward stepwise selection, on the contrary, starts with the full model and deletes a variable that is least statistically significant. In contrast with best subset selection, these methods are more computationally efficient. However, those algorithms can be too greedy. For instance, at each step, FSR fixes the variables already in the model, and selects an additional one among those remaining that minimizes residual sum of squares. Due to the much reduced search space, FSR often misses important variables. Taking a sequential selection strategy similar to stepwise algorithms, Efron et al. (2004) proposed the Least Angle Regression (LARS), which is less greedy by construction. At the first step, LARS chooses a variable if it has the largest absolute sample correlation with the response. For successive steps, the estimate is obtained such that all active variables have equal sample correlation with the current residuals. Here active variables refer to those already included in the model. Although LARS is discrete in terms of variable selection, it is closely connected with continuous selection methods, which will be discussed later.

Another class of extensions of best subset selection is to use shrinkage techniques. In particular, Problem (1.2) can be generalized to penalized regression as follows:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + \lambda P(\boldsymbol{\beta}), \tag{1.3}$$

where $\lambda P(\boldsymbol{\beta})$ is the regularization term with $P(\boldsymbol{\beta})$ being a function of $\boldsymbol{\beta}$. One of the most well known penalized variable selection methods is the least absolute shrinkage and selection operator (LASSO) introduced by Tibshirani (1996). Tibshirani (1996) proposed to impose an l_1 penalty $P(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} \beta_j$, leading to a sparse estimator that shrinks the OLS solution and sets some of the estimated coefficients exactly to zeros. In the literature, a lot of work has been done on developing both the computational and theoretical properties of the LASSO estimator (Knight and Fu, 2000; Greenshtein et al., 2006; Wu and Lange, 2008; Friedman et al., 2010). Efron et al. (2004) demonstrated that LARS can be used to compute the entire solution path of LASSO efficiently. Moreover, the optimization problem in LASSO is convex, hence many convex programming methods can be applied to solve the LASSO. Theoretical studies of LASSO mainly focus on two aspects: prediction accuracy and variable selection consistency, where the latter is more related to our thesis. Meinshausen and Bühlmann (2006) provided a neighborhood stability condition that is necessary and sufficient for LASSO to recover the support set of $\boldsymbol{\beta}$, which is equivalent to the irrepresentablility condition introduced by Zhao and Yu (2006).

Due to the shrinkage nature, LASSO may over-shrink the estimates and cause significant bias. An alternative approach to address such a problem is the adaptive LASSO (Zou, 2006). Zou (2006) suggested a reweighted version of the l_1 penalty $P(\boldsymbol{\beta}) = \sum_{j=1}^{p} w_j \beta_j$ with $w_j = |\beta_j^{init}|^{-\gamma}$ for some $\gamma > 0$, where β_j^{init} is an initial estimate for β_j. Zou (2006) also demonstrated that the adaptive LASSO is consistent in variable selection under less stringent conditions than LASSO. There are other modifications of the LASSO in the literature. For example, Zou and Hastie (2005) pointed out some limitations of LASSO as a variable selection procedure: first, in the high dimensional setting where $p \gg n$, LASSO can only select up to n variables; secondly, if there is a group of variables that are highly correlated, LASSO tends to select only one of them. To address these problems, Zou and Hastie (2005) introduced the elastic net method, using $\lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2$ as the regularization term in (3.1) and thus encouraging a grouping effect. Besides the elastic net, various penalized variable selection methods have been proposed as extensions to LASSO, including the Dantzig selector (Candes and Tao, 2007), the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001), among many others.

Despite good theoretical properties and numerical performance in practice, those penalized variable selection procedures mentioned above may not work well when the dimensionality p is very large. When p diverges with an exponential rate with respect to n, this situation is referred to as the ultrahigh dimensional setting. With the development in mass storage technology, ultrahigh dimensional data have become prevalent in different applications. As a result, ultrahigh dimensional variable selection has gained much attention. A well known example is the Sure Independence Screening (SIS) proposed by Fan and Lv (2008), which screens out covariates based upon their marginal sample correlation with the response. More specifically, let $\mathbf{w} = (\mathbf{w_1}, \mathbf{w_2}, \cdots, \mathbf{w_p})$ be a vector such that $w_j = |\widehat{\text{Corr}}(X_j, Y)|$ and γ is a constant between $(0, 1)$, then a sub-model is defined as

$$\mathcal{M}_\gamma = \{j : w_j \text{ is amongst the largest } [\gamma n] \text{ of all the correlations}\}, \tag{1.4}$$

where $[\gamma n]$ denotes the integer part of γn. Fan and Lv (2008) further demonstrated that SIS is screening consistent under some conditions. This guarantees that all those X_j's with $\beta_j \neq 0$ is included in the subset of covariates. Other examples of variable screening include, but are not limited to Wang (2009); Li et al. (2012); Zhu et al. (2012).

So far we have discussed variable selection approaches under a high dimensional setting. Although these methods are useful in many practical applications, they can be sub-optimal in the ultrahigh dimensional problems due to spurious correlation. In the following section, we will discuss some potential issues caused by correlation.

1.2 Spurious correlation

One challenge for high dimensional variable selection arises from the spurious correlation among variables. To be more specific, the sample correlation between two variables (here the two variables might be a covariate and a response, or two covariates) can be large even if they are uncorrelated in the population sense. To better understand the impact of spurious correlation, we use a simple toy example to demonstrate the issue. Let $\mathbf{x} = (X_1, X_2, \cdots, X_p)$ be i.i.d. standard normal random variables, and 50 independent observations are generated from the distribution of $\mathbf{x}$. Figure 1.1 displays the maximal absolute sample correlation between X_1

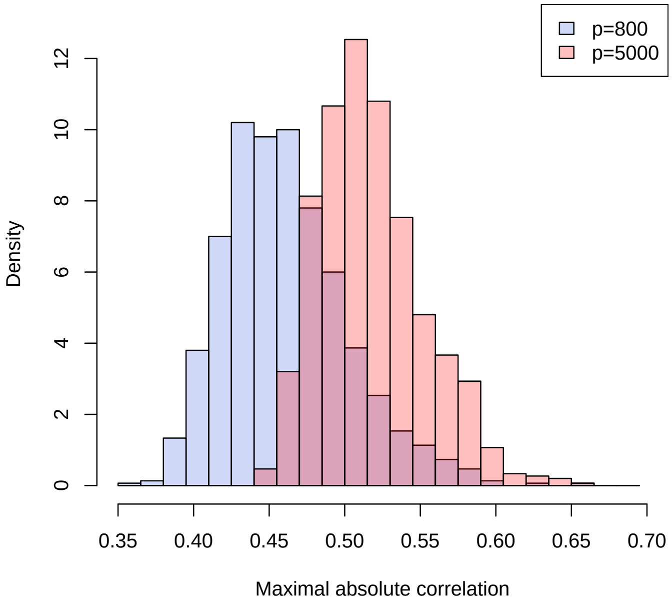


Figure 1.1: Distribution of maximal absolute sample correlation between X_1 and $\{X_j : j \neq 1\}$.

and the remaining variables out of 1000 simulations with $p = 800$ and $p = 5000$ respectively. One can see that even if X_i's are independent of each other, the magnitude of maximal sample correlation is much greater than zero.

The phenomenon of spurious correlation may challenge the aforementioned variable selection approaches and can result in failures to identify important variables. Note that LARS and FSR add a covariate into the model when it maximizes the absolute sample correlation with the response or residuals, but such a large sample correlation can be spurious, especially when the number of candidate variables increases. In other words, with the presence of spurious correlation, it becomes harder to distinguish important variables from unimportant ones. Moreover, in many applications where important variables tend to be highly correlated, it will be desirable to include all of those covariates into the model when implementing variable selection. However, due to the spurious correlation issue, it is difficult to identify covariates that are truly correlated based on their pairwise sample correlations. Although spurious correlation has been recognized in the literature (Fan et al., 2014, 2015), few of them are devoted to high dimensional variable selection, which is of our primary interest.

1.3 New contributions and outline

In this proposal, we introduce some novel variable selection methods for high dimensional linear models. We first propose a testing-based variable selection procedure that utilizes the limiting distribution of the maximal absolute partial sample correlation between covariates and the response. Moreover, we study the extreme distributions of the absolute pairwise sample correlation among covariates and incorporate the results to a penalized variable selection approach. We further consider utilizing the asymptotic results for multiple-response regression problems. The main outline is as follows:

- In Chapter 2, we introduce a flexible and efficient test-based variable selection approach that could be incorporated with any sequential selection procedure. The test is on the overall signal in the remaining inactive variables using the maximal absolute partial correlation among the inactive variables with the response given active variables. We develop the asymptotic null distribution of the proposed test statistic as the dimension grows towards infinity uniformly in the sample size. We also prove the consistency of the test procedure. With this test, at each step of the selection, we include a new variable if and only if the p-value is below some pre-defined level. Numerical studies show that the proposed method delivers very competitive performance in terms of both variable selection accuracy and computational complexity compared to cross-validation.

- We study in Chapter 3 the asymptotic distribution of the maximal absolute correlation among p independent covariates as the dimensionality p goes to infinity. Using the limiting distribution, we propose a threshold to screen covariates pairs. We further combine the pair screening with marginal screening using Sure Independence Screening (SIS) (Fan and Lv, 2008) and establish a penalized variable selection procedure that can penalize covariates selectively according to the screening results. Numerical studies demonstrate that our method is competitive in both variable selection and model prediction.

- In Chapter 4, we proposal a multiple-response regression approach using weighted L_2 penalty. We first consider applications of the asymptotic distribution of the maximal absolute sample correlation between each of the covariates and all responses. Based on the extreme value theory results, we propose a weight function using the exponential of p-values for each covariate and construct a weighted simultaneous variable selection estimator for the parameter matrix. We also introduce a blockwise descent algorithm to compute the estimator. We show that our method performs well in practice by several numerical examples.

There are some potential future work that are related to the topics discussed in this thesis. For instance, we only investigated the limiting distribution of the maximal absolute sample correlation between the i.i.d. Gaussian covariates and the response. A possible extension is to expand the result to covariates with other correlation structures. Another potential work is to develop a decision rule for the sequential testing framework in Section 2.3.1 with multiple false discovery rate control (Benjamini and Hochberg, 1995). We shall try to solve these problems in the future work.

CHAPTER 2
Test-based Variable Selection in High-dimensional linear models

2.1 Introduction

Thanks to technological advancement, high-dimensional data are now prevalent in science. Unfortunately, traditional techniques such as ordinary least squares cannot be applied directly to these high-dimensional settings, where the number of variables is typically much larger than the sample size. Furthermore, it is often the case that only a few candidate predictors are truly relevant to the response (Fan et al., 2014). In other words, the inherent high-dimensional model is sparse. It is then crucial to identify such variables, whence the important problem of variable selection arises.

In the context of linear regression, various variable selection procedures have been intensively investigated in the past decades. One type of methods are stepwise regression, including forward stepwise regression (FSR), backward stepwise regression, etc. Another well-known example is the least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996), which imposes the L_1 penalty on regression coefficients. Efron et al. (2004) proposed the least angle regression (LARS) method, which can compute efficiently the entire solution path of the LASSO with respect to the tuning parameter. Details of FSR, LASSO and LARS can be found in Chapter 1.

The variable selection methods discussed above usually involve a penalty parameter which controls the complexity of the resulting model. In practice, cross-validation (CV) is a commonly used technique for selecting the penalty parameter. However, CV is computationally inefficient. Moreover, it is based on minimizing in-sample prediction errors, and thus does not have a clear inferential meaning. Besides CV, another class of model selection approaches is based on hypothesis testing. For example, Goeman et al. (2006) and Zhong and Chen (2011) focused on testing the regression coefficients globally.

Other testing schemes have been implemented adaptively in sequential selection procedures. For example, Lockhart et al. (2014) proposed the covariance test statistic for the LASSO. Another example is the truncated Gaussian (TG) test (Tibshirani et al., 2016) developed for LARS, FSR and LASSO. While these methods are specifically designed for particular variable selection procedures, Fithian et al. (2015) introduced a general framework for testing the goodness of fit that applies to FSR, LARS and LASSO. However, their tests are developed separately for FSR and LARS (LASSO). In addition, the method of Fithian et al. (2015) requires MCMC sampling for the null distribution, which can be time consuming.

For LARS, FSR and LASSO, test-based approaches are applicable because these procedures are sequential in nature: typically, only one variable is added into the model at each step (though the LASSO can sometimes include steps in which variables are dropped). Therefore, tests can be conducted at each step of the selection procedure. One can further develop some stopping criterion based on the p-values associated with these hypothesis tests.

Another common feature of these procedures is that at each step, a variable is selected if, among all unselected (or inactive) variables, it has the largest absolute sample correlation with the current residuals, i.e., the difference between the response and its estimates from the previous step. However, such a large sample correlation can be spurious. Indeed in situations where the number of predictors is large compared to the sample size, it may happen that the response is theoretically independent from all of them and yet some of these predictors appear to be highly correlated with the response simply by chance. This phenomenon can be particularly severe in high-dimensional problems. As proved by Fan et al. (2014), the maximal correlation observed in a sample of fixed size n between a response and independent covariates can be close to 1 if the number p of such covariates is sufficiently large.

In this thesis, we introduce an efficient high-dimensional test-based variable selection method. We focus on the variable selection problem under the sparse linear model setting. Motivated by the spurious correlation issue discussed above, we construct a test statistic based on the maximal absolute sample partial correlation between the inactive covariates and the response conditioning on the active covariates at each step of the procedure. Our null hypothesis assumes that the remaining variables are conditionally independent of the response given the active variables. Based on the null distribution of the test statistic, we can detect whether there exist important covariates for the response in the inactive set. We further develop a stopping criterion from the p-values.

There are three key advantages to the proposed method, namely:

(i) The method is flexible: the proposed tests and stopping criterion can be incorporated into any sequential selection procedure, such as the aforementioned LARS, LASSO and FSR.

(ii) The method is much more computationally efficient than CV, especially when p is large.

(iii) The method can accommodate arbitrarily large p, since the asymptotic null distribution of the test statistic is developed as $p \to \infty$ uniformly in n.

This chapter is organized as follows. In Section 2.2.1, we formulate the null hypothesis and introduce the corresponding test statistic for the proposed method. In Sections 2.2.2 and 2.2.3, we discuss the asymptotic null distribution and power of our test statistic with independent covariates, respectively, and we extend the results for equally correlated covariates in Section 2.2.4. In Section 2.2.5, we introduce the permutation test for covariates with arbitrary correlation structure. In Section 2.3, we incorporate our hypothesis testing approach into sequential variable selection procedures. In Section 2.4, we demonstrate the performance of the new method through three simulation studies and a microarray data study. We present several proofs in Section 2.7.

2.2 Global test to control spurious correlation

2.2.1 Global null for testing significant variables

Consider the linear model

$$Y = \mathbf{X}^\top \boldsymbol{\beta} + \varepsilon, \tag{2.1}$$

where Y is the response variable, $\mathbf{X} = (X_1, \ldots, X_p)^\top$ is a p-dimensional covariate vector, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is the unknown coefficient vector which may be sparse, and ε is a random noise from $\mathcal{N}(0, \sigma^2)$ with σ^2 unknown. For now we assume that $\mathbf{X}$ is from a p-dimensional Gaussian distribution with some unknown covariance matrix Σ. We will discuss the non-Gaussian case in the numerical studies. Let $\mathbf{y} = (y_1, \ldots, y_n)^\top$ and $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^\top$ respectively stand for the vectors of independent observations from Y and $\mathbf{X}$.

For variable selection problems, the primary goal is to recover the support set of $\boldsymbol{\beta}$, which is the index set of non-zero components of the coefficient, denoted by $\mathcal{M}^*$. Suppose we are given a candidate set $\mathcal{M}$, which includes the indices of all selected variables, and that we want to know whether there are remaining important covariates in $\mathcal{M}^\complement$. We then need to test

$$\mathcal{H}_0 : \mathcal{M}^* \subseteq \mathcal{M}. \tag{2.2}$$

The following proposition demonstrates that under (2.1) and the Gaussian assumption, we can convert the above hypothesis into the problem of testing the conditional independence between Y and the X_j's with $j \in \mathcal{M}^\complement$.

Proposition 1. *Suppose that $\mathbf{X} = (X_1, \ldots, X_p)^\top$ has a multivariate Gaussian distribution and the response Y is generated from the linear model* (2.1). *If $\mathcal{M}$ is a subset of $\{1, \ldots, p\}$, then $\mathcal{M}^* \subseteq \mathcal{M}$ if and only if Y is independent of all $X_j s$ for $j \in \mathcal{M}^\complement$ conditional on $X_\mathcal{M}$.*

Proof of Proposition 1. The proof consists of two parts: if and only if. We first prove the if part by contradiction.

Given Y is independent of all X_j's for $j \in \mathcal{M}^c$ conditioning on $\mathcal{M}$, assume that $\mathcal{M}^* \not\subset \mathcal{M}$. Denote $\mathcal{M}_0 = \mathcal{M}^* \cap \mathcal{M}^c$, then $\mathcal{M}_0 \neq \emptyset$. Let $\mathbf{X}_{\mathcal{M}_0}$ be a sub-vector of $\mathbf{X}$ indexed by $\mathcal{M}_0$. Similarly we have the notations $\mathbf{X}_{\mathcal{M}^*}$, $\mathbf{X}_{\mathcal{M}^* \cap \mathcal{M}}$, $\boldsymbol{\beta}_{\mathcal{M}_0}$, $\boldsymbol{\beta}_{\mathcal{M}^*}$ and $\boldsymbol{\beta}_{\mathcal{M}^* \cap \mathcal{M}}$. Then we have

$$\begin{aligned}
\mathrm{Cov}(\mathrm{Y}, \mathbf{X}_{\mathcal{M}_0} | \mathcal{M}) &= \mathrm{Cov}(\mathbf{X}_{\mathcal{M}^*}^\top \boldsymbol{\beta}_{\mathcal{M}^*} + \varepsilon, \mathbf{X}_{\mathcal{M}_0} | \mathcal{M}) \\
&= \mathrm{Cov}(\mathbf{X}_{\mathcal{M}_0}^\top \boldsymbol{\beta}_{\mathcal{M}_0}, \mathbf{X}_{\mathcal{M}_0} | \mathcal{M}) + \mathrm{Cov}(\mathbf{X}_{\mathcal{M}^* \cap \mathcal{M}}^\top \boldsymbol{\beta}_{\mathcal{M}^* \cap \mathcal{M}} + \varepsilon, \mathbf{X}_{\mathcal{M}_0} | \mathcal{M}) \\
&= \mathrm{Cov}(\mathbf{X}_{\mathcal{M}_0}^\top \boldsymbol{\beta}_{\mathcal{M}_0}, \mathbf{X}_{\mathcal{M}_0} | \mathcal{M}) \\
&\quad \text{(as } \mathcal{M}^* \cap \mathcal{M} \subset \mathcal{M}, \text{ and } \varepsilon \text{ is independent of any } X_j) \\
&= \mathrm{Var}(\mathbf{X}_{\mathcal{M}_0} | \mathcal{M}) \boldsymbol{\beta}_{\mathcal{M}_0},
\end{aligned}$$

where $\mathrm{Var}(\mathbf{X}_{\mathcal{M}_0} | \mathcal{M})$ is the conditional covariance matrix given $\mathcal{M}$. Since $\mathcal{M}_0 \cap \mathcal{M} = \emptyset$, $\mathrm{Var}(\mathbf{X}_{\mathcal{M}_0} | \mathcal{M})$ is positive definite. On the other hand, $\mathcal{M}_0 \subset \mathcal{M}^*$ implies $\boldsymbol{\beta}_{\mathcal{M}_0} \neq \mathbf{0}$. Therefore, $\mathrm{Var}(\mathbf{X}_{\mathcal{M}_0} | \mathcal{M}) \boldsymbol{\beta}_{\mathcal{M}_0} \neq \mathbf{0}$, leading to contradiction. As a result, we can conclude $\mathcal{M}^* \subset \mathcal{M}$.

Now we prove the only if part. Suppose $\mathcal{M}^* \subset \mathcal{M}$, then for any $j \in \mathcal{M}^c$, we have

$$\begin{aligned}\mathrm{Cov}(\mathrm{Y}, \mathrm{X_j}|\mathcal{M}) &= \mathrm{Cov}(\mathbf{X}^\top \boldsymbol{\beta} + \varepsilon, \mathrm{X_j}|\mathcal{M}) = \mathrm{Cov}(\mathbf{X}_{\mathcal{M}^*}^\top \boldsymbol{\beta}_{\mathcal{M}^*} + \varepsilon, \mathrm{X_j}|\mathcal{M}) \\ &= \mathbf{X}_{\mathcal{M}^*}^\top \mathrm{Cov}(\boldsymbol{\beta}_{\mathcal{M}^*}, \mathrm{X_j}|\mathcal{M}) + \mathrm{Cov}(\varepsilon, \mathrm{X_j}|\mathcal{M}) \quad (\text{ as } \mathcal{M}^* \subset \mathcal{M}) \\ &= 0. \quad (\text{as } \varepsilon \text{ is independent of any } X_j)\end{aligned}$$

Since Y and X_j are normally distributed, it follows that Y is independent of X_j for $j \in \mathcal{M}^c$ conditioning on $\mathcal{M}$. □

Proposition 1 guarantees that testing (2.2) is equivalent to the following null hypothesis:

$$\mathcal{H}_0^{\mathcal{M}} : \text{Given } X_{\mathcal{M}}, Y \text{ is independent of all } X_j\text{s for } j \in \mathcal{M}^{\complement}. \tag{2.3}$$

Unless the noise level σ is large, the correlation between an important covariate and the response should be stronger than the maximal spurious correlation. In fact, many existing variable selection methods, such as the LASSO and FSR, select variables that maximize the absolute marginal correlation between the covariates and the response or the current residuals. Moreover, it is easy and efficient to obtain the maximal absolute correlation, even if the dimension p is high. Therefore, studying the distribution of the maximal absolute correlation under the null hypothesis (2.3) can help discover true important covariates among the candidate predictor variables.

We cannot directly test (2.3) based on the correlation between Y and X_j because they can be both correlated with X_i for some $i \in \mathcal{M}$. In classical regression, the partial correlation is commonly used to test conditional independence given a controlling variable. Motivated by that observation, we develop our test statistic based on the sample partial correlation between $\{X_j : j \in \mathcal{M}^{\complement}\}$ and Y conditioning on $X_{\mathcal{M}}$. We first regress $\{X_j : j \in \mathcal{M}^{\complement}\}$ and Y onto $X_{\mathcal{M}}$, respectively; we then obtain the regression residual vectors

$$\mathbf{r}_j = (I - P_{\mathcal{M}})\mathbf{x}_j, \quad j \in \mathcal{M}^{\complement}, \qquad \mathbf{r} = (I - P_{\mathcal{M}})\mathbf{y}, \tag{2.4}$$

where $P_{\mathcal{M}} = X_{\mathcal{M}}(X_{\mathcal{M}}^\top X_{\mathcal{M}})^\dagger X_{\mathcal{M}}^\top$ is the projection onto the column space of $X_{\mathcal{M}}$. Here $X_{\mathcal{M}}$ consists of the columns of X indexed by $\mathcal{M}$ and a vector column of 1s, so that all residual vectors have zero mean, and $A^\dagger$ denotes the Moore–Penrose pseudo-inverse of a matrix A. We then compute the maximal absolute sample correlation between $\{\mathbf{r}_j : j \in \mathcal{M}^\complement\}$ and $\mathbf{r}$. In this way, we define our test statistic as

$$R_{\mathcal{M}} = \max_{j:j\in\mathcal{M}^c} |\widehat{\mathrm{Corr}}(\mathbf{r}_j, \mathbf{r})|, \tag{2.5}$$

where $\widehat{\mathrm{Corr}}(\mathbf{r}_j, \mathbf{r})$ is the Pearson sample correlation between $\mathbf{r}_j$ and $\mathbf{r}$. Note that the distribution of $R_{\mathcal{M}}$ depends on n, p and s, but for simplicity we omit them in the notation for $R_{\mathcal{M}}$. Since both $\mathbf{r}_j$ and $\mathbf{r}$ have zero mean, we can write

$$R_{\mathcal{M}} = \max_{j:j\in\mathcal{M}^c} \frac{|\langle \mathbf{r}_j, \mathbf{r}\rangle|}{\|\mathbf{r}_j\|\|\mathbf{r}\|},$$

where $\langle\,\cdot\,,\,\cdot\,\rangle$ is the inner product of two vectors and $\|\cdot\|$ represents the L_2 norm. Moreover, note that our test statistic does not depend on the mean and variance of the covariates or the response.

To gain insight into the proposed test statistic, we start from a special case where $\mathcal{M} = \emptyset$. The properties of the Pearson sample correlation have been intensively studied under the classical setting $n > p$. In particular, it has been shown that when X_j and Y are independent Gaussian random variables, $|\widehat{\mathrm{Corr}}(X_j, Y)|^2 \sim \mathcal{B}[1/2, (n-2)/2]$; see, e.g., Muirhead (2009). Here $\mathcal{B}(s, t)$ represents a Beta distribution with parameters s, t. Therefore, the magnitude of each $\widehat{\mathrm{Corr}}(X_j, Y)$ cannot be too large. However, by taking maxima, $R_{\mathcal{M}}$ will be larger as p increases. In fact, for a fixed sample size n, under (2.3), $R_{\mathcal{M}}$ can get close to 1 as $p \to \infty$; see, e.g., Fan et al. (2014). The phenomenon of irrelevant covariates being highly correlated with the response is referred to as "spurious correlation", which challenges variable selection and may lead to false scientific discoveries. Thus it is important to study the distribution of $R_{\mathcal{M}}$, especially for high-dimensional problems.

In what follows, we discuss the asymptotic null distribution (Section 2.2.2) and power (Section 2.2.3) of $R_{\mathcal{M}}$ respectively for the situation where the X_js are independent random variables. We discuss the situation where the covariates are dependent in Section 2.2.4.

2.2.2 Null distribution of the test statistic with independent covariates

The limiting distribution of the maximal absolute sample correlation has been investigated recently under the assumption of independent Gaussian covariates; see Theorem II.4 in (Zhang, 2017). The latter paper focuses on the global null hypothesis that Y is independent of the X_js, which is a special case of (2.3) with $\mathcal{M} = \emptyset$. We expand the results to a more general setting and derive the exact asymptotic distribution of the proposed test statistic under (2.3), as described in the following theorem

Theorem 1. *Suppose we observe a random sample of size n from the linear model* (2.1) *and we further assume that the X_js are independent. Let $\mathcal{M}$ be a candidate set with cardinality $|\mathcal{M}| = s < n-2$ and $R_{\mathcal{M}}$ be defined as in* (2.5). *Define*

$$a(p,n,s) = 1 - (p-s)^{-2/(n-s-2)}c(p,n,s), \quad b(p,n,s) = \frac{2}{n-s-2}(p-s)^{-2/(n-s-2)}c(p,n,s),$$

where

$$c(p,n,s) = \{\frac{n-s-2}{2}\mathcal{B}(\frac{1}{2}, \frac{n-s-2}{2})\sqrt{1-(p-s)^{-2/(n-s-2)}}\}^{2/(n-s-2)}$$

is a correction factor with $\mathcal{B}(s,t)$ being the Beta function. Then under the null hypothesis (2.3), *for all $x \in \mathbb{R}$,*

$$\lim_{p\to\infty} \sup_{n\geq s+3} \left| \Pr\left\{ \frac{R^2_{\mathcal{M}} - a(p,n,s)}{b(p,n,s)} < x \right\} - F_{n,s}(x) \right| = 0,$$

where

$$F_{n,s}(x) = \exp\left\{ -\left(1 - \frac{2}{n-s-2}x\right)^{(n-s-2)/2} \right\} \mathbf{1}\left(x \leq \frac{n-s-2}{2}\right) + \mathbf{1}\left(x > \frac{n-s-2}{2}\right). \tag{2.6}$$

Remark 2.2.1. *The convergence in Theorem 1 is with respect to p instead of n, making it possible to test models where $p \gg n$. Therefore, the proposed test statistic is applicable to high-*

dimensional or ultra-high-dimensional problems. In addition, the convergence is uniform for any $n \geq s+3$, and thus ensures finite-sample performance.

With the results in Theorem 1, we can further compute the p-value associated with the null hypothesis (2.3). Let $r_{\mathcal{M}}$ denote the observed value of $R_{\mathcal{M}}$. Then the p-value of $R_{\mathcal{M}}$ for (2.3) is approximated by

$$p(r_{\mathcal{M}}) = 1 - F_{n,s}\left\{\frac{r_{\mathcal{M}}^2 - a(p,n,s)}{b(p,n,s)}\right\}, \tag{2.7}$$

with $F_{n,s}$ as specified in Theorem 1. If the p-value is small, it is likely that at least one variable from $\{X_j : j \in \mathcal{M}^{\complement}\}$ is correlated with the response. Therefore we can construct a stopping criterion based on p-values in sequential selection procedures. We will provide a detailed discussion in Section 2.3.

Our test statistic can be connected to the conventional t-test for testing whether the population correlation is zero. The t-statistic is defined as $t = r\sqrt{(n-2)/(1-r^2)}$, where r is the Pearson sample correlation between two Gaussian random variables. Motivated by that connection, we also develop a maximal t-statistic corresponding to the proposed test statistic $R_{\mathcal{M}}$. The maximal t-statistic is

$$T_{\mathcal{M}} = \sqrt{\frac{(n-s-2)R_{\mathcal{M}}^2}{1-R_{\mathcal{M}}^2}}. \tag{2.8}$$

Analogous to the results in Theorem 1, we derive next the asymptotic null distribution of $T_{\mathcal{M}}$.

Corollary 1. *Consider the same setting as in Theorem 1, and let $T_{\mathcal{M}}$ be defined as in* (2.8)*. Then, for all $x \in \mathbb{R}$, uniformly for any $n \geq s+3$,*

$$\lim_{p\to\infty} \Pr\left\{\frac{T_{\mathcal{M}} - \widetilde{a}(p,n,s)}{\widetilde{b}(p,n,s)} < x\right\} = F_{n,s}(x),$$

where $\widetilde{a}(p,n,s) = \sqrt{\{(n-s-2)a(p,n,s)\}/\{1-a(p,n,s)\}}$, $\widetilde{b}(p,n,s) = [(n-s-2)a(p,n,s)\{1-a(p,n,s)\}]^{-1/2}$ with $a(p,n,s)$ given in Theorem 1, and $F_{n,s}(x)$ as in (2.6)*.*

Proof of Corollary 1. Write $g(x) = \sqrt{\frac{(n-s-2)x}{1-x}}$, then we have $T_{\mathcal{M}} = g(R_{\mathcal{M}}^2)$ and $\widetilde{a}(p,n,s) = g(a(p,n,s))$.

By the Taylor expansion,

$$T_{\mathcal{M}} - g(a(p,n,s)) \approx g'\big(a(p,n,s)\big)b(p,n,s) \cdot \frac{R^2_{\mathcal{M}} - a(p,n,s)}{b(p,n,s)}.$$

Then we have

$$\begin{aligned} g^{'}(a(p,n,s))b(p,n,s) &= \frac{1}{2\sqrt{(n-s-2)}}\sqrt{\frac{1}{a(p,n,s)(1-a(p,n,s))^{3/2}}} \cdot b(p,n,s) \\ &= \Big((n-s-2)a(p,n,s)\big(1-a(p,n,s)\big)\Big)^{-1/2} \\ &= \widetilde{b}(p,n,s). \end{aligned}$$

By the delta method, it follows that

$$\begin{aligned} &\lim_{p\to\infty} \sup_{n\geq s+3} \left| \Pr\left(\frac{T_{\mathcal{M}} - \widetilde{a}(p,n,s)}{\widetilde{b}(p,n,s)} < x \right) - F_{n,s}(x) \right| \\ =&\lim_{p\to\infty} \sup_{n\geq s+3} \left| \Pr\left(\frac{R^2_{\mathcal{M}} - a(p,n,s)}{b(p,n,s)} < x \right) - F_{n,s}(x) \right| \\ =&\ 0. \end{aligned}$$

□

Our simulation results show that the difference between p-values obtained from $R_{\mathcal{M}}$ and $T_{\mathcal{M}}$ is negligible. Moreover, when the covariates are correlated, the null distribution of $R_{\mathcal{M}}$ is easier to approximate, which will be discussed in Section 2.2.4. Therefore we develop our test-based procedure with $R_{\mathcal{M}}$ instead of $T_{\mathcal{M}}$.

2.2.3 Asymptotic power with independent covariates

In this section, we still focus on independent Gaussian covariates. We analyze the asymptotic power of $R_{\mathcal{M}}$ by considering the following alternative hypothesis:

$$\begin{aligned} \mathcal{H}_1 :&\text{Conditionally on } X_{\mathcal{M}}, \text{ there exists at least one } j \in \mathcal{M}^{\complement} \text{ such that } Y \\ &\text{is correlated with } X_j. \end{aligned} \tag{2.9}$$

In the following theorem we show that under (2.9), the asymptotic power of the proposed test statistic $R_{\mathcal{M}}$ is 1.

Theorem 2. *Suppose we have the linear model* (2.1) *and assume that the X_js are independent Gaussian variables. Then under the alternative hypothesis* (2.9)*, as* $\ln p/n \to 0$ *and* $n \to \infty$, $\Pr\{R_{\mathcal{M}} \geq x_\alpha(p, n, s) | \mathcal{H}_1\} \longrightarrow 1$*, where* $x_\alpha(p, n, s)$ *is the critical value of* $\mathcal{H}_0^{\mathcal{M}}$ *at significance level* α*.*

Theorem 2 shows the consistency of our dependency test based on the proposed test statistic when at least one covariate is correlated with the response under the linear model setting.

2.2.4 Null distribution of the test statistic with equally correlated covariates

In Theorem 1 we have derived the exact asymptotic distribution of $R_{\mathcal{M}}$ under (2.3) when the covariates are independent Gaussian variables. When the X_j's have an arbitrary correlation structure, it is difficult to obtain similar results. We can point to some results in classical extreme-value theory; see, e.g., Chapter 3.8 in (Galambos, 1978). In particular, if $U_1, \ldots, U_n$ is a stationary Gaussian sequence with zero expectation and unit variance, then the limiting distribution of $W_n = \max(U_1, \ldots, U_n)$ only depends on the limiting behavior of $r_m/\ln(m)$, where $r_m = \mathrm{E}(U_i U_{i+m})$ is the correlation between U_i and U_{i+m}. Note that due to the stationarity assumption, r_m does not change with respect to i. More specifically, if there is another zero-mean, unit-variance stationary Gaussian sequence $U_1', \ldots, U_n'$ that has equal pairwise correlation $r = r(n)$, and $r(n)/\ln(n)$ has the same limiting form as $r_m/\ln(m)$, then $W_n' = \max(U_1', \ldots, U_n')$ has the same asymptotic distribution as W_n when $n \to \infty$. Inspired by that result, we focus on analyzing the null distribution of $R_{\mathcal{M}}$ when $X_1, \ldots, X_p$ are equally correlated, i.e., $\mathrm{Corr(X_i, X_j)} = \rho$ with $-1/(p-1) \leq \rho \leq 1$ for all $i \neq j$.

Without loss of generality, we assume that each of the X_js has zero mean and unit variance. Under the equal correlation assumption, it is well known that we can decompose X_j into a linear combination of iid standard Gaussian random variables $Z_1, \ldots, Z_p$, i.e.,

$$X_j = \sqrt{1-\rho}\, Z_j + h_\rho \frac{1}{\sqrt{p}} \sum_{i=1}^{p} Z_i, \tag{2.10}$$

where $h_\rho = \{\sqrt{1+(p-1)\rho} - \sqrt{1-\rho}\}/\sqrt{p}$. In fact, we can also replace p by $p-s$ in (2.10) such that each of $\{X_j : j \in \mathcal{M}^{\complement}\}$ is decomposed into a linear combination of $p-s$ i.i.d. Gaussian random variables. However, under high-dimensional sparse model settings, $p \gg s$. Hence the two decompositions are almost the same. For notational simplicity, we consider using p instead of $p-s$.

Let $\mathbf{z}_j = (z_{j1}, \ldots, z_{jn})^\top$ be n independent samples of Z_j and $\tilde{\mathbf{r}}_j = (I - P_{\mathcal{M}})\mathbf{z}_j$ be the residuals from projecting $\mathbf{z}_j$ onto the column space of $X_{\mathcal{M}}$. It follows from (2.10) that

$$\mathbf{r}_j = \sqrt{1-\rho}\,\tilde{\mathbf{r}}_j + h_\rho \frac{1}{\sqrt{p}} \sum_{i=1}^{p} \tilde{\mathbf{r}}_i.$$

Hence we have

$$\langle \mathbf{r}_j, \mathbf{r} \rangle = \sqrt{1-\rho}\,\langle \tilde{\mathbf{r}}_j, \mathbf{r} \rangle + h_\rho \left\langle p^{-1/2} \sum_{i=1}^{p} \tilde{\mathbf{r}}_i, \mathbf{r} \right\rangle,$$

where $\mathbf{r}_j$ and $\mathbf{r}$ are defined as in (2.4).

Recall that by assumption, $\mathrm{var}(Z_j) = \mathrm{var}(X_j) = 1$. Thus conditioning on $X_{\mathcal{M}}$, we have

$$\|\mathbf{r}_j\|^2 \overset{\mathrm{d}}{\sim} \chi^2_{n-s-1}, \quad \|\tilde{\mathbf{r}}_j\|^2 \overset{\mathrm{d}}{\sim} \chi^2_{n-s-1}, \quad \|p^{-1/2} \sum_{i=1}^{p} \tilde{\mathbf{r}}_i\|^2 \overset{\mathrm{d}}{\sim} \chi^2_{n-s-1}.$$

For moderately large n, we can approximate $\widehat{\mathrm{Corr}}(\mathbf{r}_j, \mathbf{r}) = \langle \mathbf{r}_j, \mathbf{r} \rangle / (\|\mathbf{r}_j\| \|\mathbf{r}\|)$ by

$$\widehat{\mathrm{Corr}}(\mathbf{r}_j, \mathbf{r}) \approx \sqrt{1-\rho}\,\widehat{\mathrm{Corr}}(\tilde{\mathbf{r}}_j, \mathbf{r}) + h_\rho \widehat{\mathrm{Corr}}\left(p^{-1/2} \sum_{i=1}^{p} \tilde{\mathbf{r}}_i, \mathbf{r} \right).$$

Taking the maximum on both sides, we find

$$\max_{j: j \in \mathcal{M}^c} \widehat{\mathrm{Corr}}(\mathbf{r}_j, \mathbf{r}) \approx \sqrt{1-\rho} \max_{j: j \in \mathcal{M}^c} \widehat{\mathrm{Corr}}(\tilde{\mathbf{r}}_j, \mathbf{r}) + h_\rho \widehat{\mathrm{Corr}}\left(p^{-1/2} \sum_{i=1}^{p} \tilde{\mathbf{r}}_i, \mathbf{r} \right). \tag{2.11}$$

Under the null hypothesis (2.3), note that $\mathbf{r} = (I - P_{\mathcal{M}})\mathbf{y} = (I - P_{\mathcal{M}})\boldsymbol{\varepsilon}$ and thus $\tilde{\mathbf{r}}_j = (I - P_{\mathcal{M}})\mathbf{z}_j$ is conditionally independent of $\mathbf{r}$ given $X_{\mathcal{M}}$ for all $j \in \{1, \ldots, p\}$. Hence the variables $\{|\widehat{\mathrm{Corr}}(\tilde{\mathbf{r}}_j, \mathbf{r})|^2 : j \in \mathcal{M}^{\complement}\}$ are independently distributed as $\mathcal{B}(\frac{1}{2}, \frac{n-s-2}{2})$ conditioning on $X_{\mathcal{M}}$. Furthermore, from a property of the normal distribution,

$$p^{-1/2}\sum_{i=1}^{p} Z_i \overset{\mathrm{d}}{\sim} \mathcal{N}(0,1).$$

Thus the conditional distribution of $|\widehat{\mathrm{Corr}}(p^{-1/2}\sum_{i=1}^{p}\tilde{\mathbf{r}}_i,\mathbf{r})|^2$ given $X_{\mathcal{M}}$ is also $\mathcal{B}(\frac{1}{2},\frac{n-s-2}{2})$. Therefore, the two terms on the right-hand side of (2.11) have corresponding exact distributions. Letting f_1, f_2 be the densities of $\max\limits_{j:j\in\mathcal{M}^c}\widehat{\mathrm{Corr}}(\tilde{\mathbf{r}}_j,\mathbf{r})$ and $\widehat{\mathrm{Corr}}(p^{-1/2}\sum_{i=1}^{p}\tilde{\mathbf{r}}_i,\mathbf{r})$, respectively, we have

$$f_1(x;p,n,s) = p|x|f_B(x^2;n,s)\left\{\frac{1+\mathrm{sign(x)}F_B(x^2;n,s)}{2}\right\}^{p-s-1},$$
$$f_2(x;n,s) = |x|f_B(x^2;n,s),$$

where $f_B(x;n,s)$ and $F_B(x;n,s)$ are the density and the cumulative distribution function of $\mathcal{B}(\frac{1}{2},\frac{n-s-2}{2})$, respectively.

It is known that when $p\to\infty$, $\max(Z_1,\ldots,Z_p)$ and $Z_1+\cdots+Z_p$ are independent; see, e.g., (James et al., 2007). With asymptotic independence, the density $f_3(x;p,n,s)$ of $\max\limits_{j:j\in\mathcal{M}^c}\widehat{\mathrm{Corr}}(\tilde{\mathbf{r}}_j,\mathbf{r})$ can be approximated, for all $z\in[0,1]$, by

$$f_3(z;p,n,s) \approx \int_{-\infty}^{\infty}\tilde{f}_1(z-x)\tilde{f}_2(x)dx, \tag{2.12}$$

with $\tilde{f}_1(x) = \rho^{-1/2}f_1(\rho^{-1/2}x;p,n,s)$ and $\tilde{f}_2(x) = f_2(x/h_\rho;n,s)/h_\rho$. In practice, ρ can be estimated by the average of pairwise correlations among the covariates. Let

$$U_{\mathcal{M}} = \max_{j:j\in\mathcal{M}^c}\widehat{\mathrm{Corr}}(\mathbf{r}_j,\mathbf{r}),\quad V_{\mathcal{M}} = -\min_{j:j\in\mathcal{M}^c}\widehat{\mathrm{Corr}}(\mathbf{r}_j,\mathbf{r}).$$

Note that $R_{\mathcal{M}} = \max(U_{\mathcal{M}},V_{\mathcal{M}})$, where $U_{\mathcal{M}}$ and $V_{\mathcal{M}}$ have identical distributions, but are not independent.

Due to the dependence between $U_{\mathcal{M}}$ and $V_{\mathcal{M}}$, it is difficult to derive the distribution of $R_{\mathcal{M}}$ and the corresponding p-value when we use $R_{\mathcal{M}}$ as the test statistic. One possible way to tackle this problem is to take $U_{\mathcal{M}}$ or $V_{\mathcal{M}}$ as the test statistic instead. However, the resulting test might not be powerful enough. For example, when the true model is $Y=-X_1+\varepsilon$, a larger value of $V_{\mathcal{M}}$ it is difficult to reject the null hypothesis (2.3) based on the null distribution of $U_{\mathcal{M}}$. Similarly, if the true model is $Y = X_1 + \varepsilon$, then using $V_{\mathcal{M}}$ as the test statistic might be unable to detect X_1. However, note that if the null hypothesis does not hold, i.e., there are important variables remaining in $\mathcal{M}^{\complement}$, it can be expected that the tail probability of $R_{\mathcal{M}}$ will be very small. It can then be approximated by

$$\Pr\left(R_{\mathcal{M}} \geq x\right) \approx \Pr\left(U_{\mathcal{M}} \geq x\right) + \Pr\left(V_{\mathcal{M}} \geq x\right) = 2\Pr\left(U_{\mathcal{M}} \geq x\right). \tag{2.13}$$

Since $\Pr\left(R_{\mathcal{M}} \geq x\right) \in \left[\Pr\left(U_{\mathcal{M}} \geq x\right), 2\Pr\left(U_{\mathcal{M}} \geq x\right)\right]$ always holds, if $2\Pr\left(U_{\mathcal{M}} \geq x\right)$ is small, $\Pr\left(R_{\mathcal{M}} \geq x\right)$ will also be very small, which implies that the null hypothesis may be rejected. Therefore we can compare $2\Pr\left(U_{\mathcal{M}} \geq x\right)$ with a pre-specified constant c to determine which test statistic, $R_{\mathcal{M}}$ or $U_{\mathcal{M}}$ to use. In general, we propose to compute the p-value corresponding to (2.3) in the following way:

$$p = \begin{cases} \Pr\left(R_{\mathcal{M}} \geq x\right) \approx 2\Pr\left(U_{\mathcal{M}} \geq x\right) & \text{if } 2\Pr\left(U_{\mathcal{M}} \geq x\right) \leq c, \\ \Pr\left(U_{\mathcal{M}} \geq x_1\right) & \text{otherwise,} \end{cases} \tag{2.14}$$

where x and x_1 represent the observed value of $R_{\mathcal{M}}$ and $U_{\mathcal{M}}$, respectively, and

$$\Pr\left(U_{\mathcal{M}} \geq t\right) \approx \int_t^\infty f_3(z; p, n, s) dz.$$

The constant c is essentially a parameter balancing the accuracy and conservatism of the resulting p-value. Specifically, if c is too small, the p-value is then computed from $U_{\mathcal{M}}$, which can be too conservative; if c is too large, the approximation in (2.13) will be invalid. Our numerical studies indicate that so long as c is relatively small, the performance of our method won't be affected much. Thus we set $c = 0.01$ throughout the numerical studies in Section 2.4.

To investigate the accuracy of the proposed asymptotic null distributions in Sections 2.2.2 and 2.2.4, we compare the p-values obtained from the proposed asymptotic null distributions with the uniform distribution on $[0, 1]$ respectively. We simulate 1000 independent datasets with 200 samples from the linear model $Y = \mathbf{X}^\top \boldsymbol{\beta} + \varepsilon$, where $\boldsymbol{\beta}$ is a vector of dimensionality 2000 with non-zero components $(\beta_1, \beta_2, \beta_3) = (3, -1.5, 2)$. We generate X_j's from i.i.d. normal random variables for the independence case; while for the equally-correlated covariates, the correlation is $\rho = 0.3$. The random noise is from $\varepsilon \sim \mathcal{N}(0, 9)$. We first project the response and unimportant covariates onto the space spanned by X_1, X_2, X_3 and calculate $R_{\mathcal{M}}$ with $\mathcal{M} = \{1, 2, 3\}$, then obtain the p-value using the proposed asymptotic distributions. The two Q-Q plots are presented in Figures 2.1a and 2.1b respectively. It can be seen from the plots that in both scenarios, the distribution of the empirical p-values is close to a uniform distribution on $[0, 1]$, indicating that our theoretical results are valid.

2.2.5 Permutation test

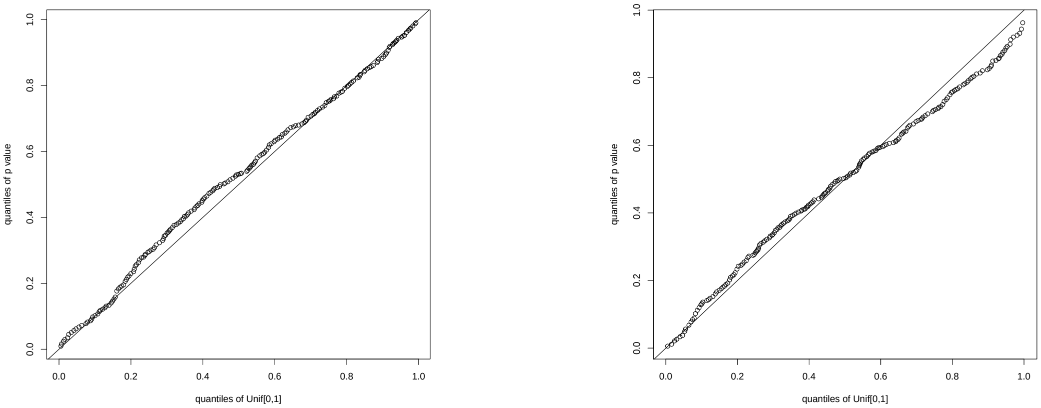
In the previous subsection, we mentioned when the correlation structure of the covariates is unknown, we can still obtain the p-value approximately using the proposed asymptotic distributions. In fact, the p-value can also be computed using the permutation test, which is a well-known resampling procedure that has many applications. A permutation test is applicable if the samples are exchangeable when the null hypothesis holds. In fact, under certain assumptions, the exchangeability condition can be satisfied.

Remark 2.2.2. Suppose $\tilde{\mathbf{y}}$ is a random permuted sample from $\mathbf{y}$ and we obtain the test statistic as $R_{\mathcal{M}}(\tilde{\mathbf{y}}, \mathbf{X})$. If Y is independent of the covariates, i.e., $\boldsymbol{\beta} = \mathbf{0}$, then $R_{\mathcal{M}}(\tilde{\mathbf{y}}, \mathbf{X})$ has the same distribution as $R_{\mathcal{M}}$.

To conduct the permutation test, at each step of the sequential selection, we randomly permute the observations of Y and obtain a new sample. Then we can compute the test statistic based on the new sample. The permutations are implemented repeatedly, and the p-value is obtained by the ranking of the original test statistic among the permuted test statistics over the total number of permutations. We further illustrate the permutation test step by step as below:

1. At Step k, we shuffle the observations of Y at random Q times and obtain the permuted sample $Y^{(q)} = (y_1^q, \ldots, y_n^q)$ for $q \in \{1, \ldots, Q\}$.

2. Compute the corresponding test statistic $R_{\mathcal{M}}^q$ for each $Y^{(q)}$, and compare the test statistic $R_{\mathcal{M}}$ obtained from the original Y.



(a) Independent covariates

(b) Equally correlated covariates

Figure 2.1: Q-Q plots of p-values against U[0,1] when the covariates are (a) independent; (b) equally-correlated with correlation $\rho = 0.3$.

3. Suppose the rank of $R_{\mathcal{M}}$ among $R^1_{\mathcal{M}}, \ldots, R^Q_{\mathcal{M}}$ is r_k. Then the p-value of the permutation test can be written as $p_k = r_k/Q$.

Recall that our goal is to use the distribution information to provide guidance for sequential selection procedures. In what follows, we introduce a test-based variable selection procedure by applying the results obtained in Section 2.2.

2.3 Sequential testing for variable selection

2.3.1 Testing-based variable selection procedure

For sequential selection procedures, it is crucial to find a stopping criterion. In other words, at each step of a particular selection procedure, we want to know whether there are remaining important covariates in the inactive set. Therefore, we propose to conduct the dependence test introduced in the previous section correspondingly at each step and stop the procedure once we accept the null hypothesis. This leads to a test-based variable selection approach.

Suppose we are at Step k ($k \geq 1$) of a sequential selection procedure, and let $\mathcal{A}_{k-1}$ denote the active set that includes the indices of selected variables from the previous step. We want to emphasize that here $\mathcal{A}_{k-1}$ is fixed given the data. In contrast, we use the notation $\hat{\mathcal{A}}_{k-1}(\mathbf{X}, Y)$ to denote the index set for sampling from the data, which is random. Then one needs to know whether the remaining inactive covariates are all uncorrelated with the response, which is equivalent to testing (2.3) with $\mathcal{M} = \mathcal{A}_{k-1}$ under the Gaussian assumption. Note that $\mathcal{A}_0 = \emptyset$ when $k = 1$. More specifically, we consider the following null hypothesis at Step k:

$$\mathcal{H}_0^{(k)} : \text{ Conditioning on } X_{\mathcal{A}_{k-1}}, Y \text{ and } X_j \text{ are independent for } \forall j \notin \mathcal{A}_{k-1}. \tag{2.15}$$

We note here that the proposed testing in Section 2.2 conditions on $X_{\mathcal{A}_{k-1}}$, where $\mathcal{A}_{k-1}$ is non-random, rather than on both $X_{\mathcal{A}_{k-1}}$ and $\hat{\mathcal{A}}_{k-1}(\mathbf{X}, Y) = \mathcal{A}_{k-1}$. However, below are a few justifications for using the proposed test in the model selection procedure.

1. The main purpose of using the test in Section 2.2 is to control the entry of variables with spurious partial correlation in the selection process. The ultimate goal is to assist the selected model in having good properties on FP, FN and MSE. In this regard, the problem is essentially different from post-selection inference (Fithian et al., 2015; Tibshirani et al., 2016), where the aim is to obtain valid conclusions for scientific discoveries. The simulation and real data studies in Section 2.4 demonstrate the good model selection properties of the proposed procedure.

2. In Section 2.3.2 we compare the empirical distributions of the unconditional test statistic in Section 2.2 and the conditional ones through extensive simulations. We find that the difference is very small.

3. The unconditional test provides a valid p-value at the first step of model selection to prevent any spurious variables from entering the model when $\boldsymbol{\beta} = \mathbf{0}$. For later steps, our test provides a good approximation of spurious correlation control.

Based on the above considerations, we propose to incorporate the test in Section 2.2 in the sequential selection procedure. The procedure is detailed below. Under (2.15), the corresponding test statistic can be written as

$$R^{(k)} = \max_{j: j \in \mathcal{A}_{k-1}^c} |\widehat{\mathrm{Corr}}(\mathbf{r}_j^{(k)}, \mathbf{r}^{(k)})|, \tag{2.16}$$

where

$$\mathbf{r}_j^{(k)} = (I - P_{\mathcal{A}_{k-1}})\mathbf{x}_j, \quad \mathbf{r}^{(k)} = (I - P_{\mathcal{A}_{k-1}})\mathbf{y}$$

with $P_{\mathcal{A}_{k-1}}$ defined in the similar way as in Section 2.2.1. Note that when $k = 1$, we have $\mathbf{r}_j^{(1)} = \mathbf{x}_j - \bar{x}_j \mathbf{1}_n$, where $\bar{x}_j$ is the mean of $\mathbf{x}_j$ and $\mathbf{1}_n$ is an n-dimensional vector of 1s since $P_{\mathcal{A}_0} = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$. Similarly $\mathbf{r}^{(1)}$ reduces to $\mathbf{y} - \bar{y}\,\mathbf{1}_n$.

From Theorem 1, we can see that when the covariates are independent, the p-value of $R^{(k)}$ converges to a uniform distribution on the unit interval, $\mathcal{U}(0,1)$, under null hypothesis (2.15). This conclusion is formally stated below.

Corollary 2. *Suppose we have a linear model as in* (2.1) *and we assume that the covariates are independent Gaussian variables. Let $x^{(k)}$ be the observed value of the test statistic $R^{(k)}$ as defined in* (2.16)*. Then the p-value can be obtained from $p(x^{(k)}) = 1 - F_{n,k-1}(x^{(k)})$. Under the null hypothesis* (2.15)*, we have $p(x^{(k)}) \xrightarrow{d} \mathcal{U}(0,1)$ as $p \to \infty$.*

We omit the proof because it follows directly from Theorem 1. Corollary 2 suggests that it is possible and reasonable to use the proposed test statistic $R^{(k)}$ when the covariates are independent Gaussian variables. For dependent covariates, although we do not have similar theoretical results for the distribution of the p-value, we can use the approximation described in Section 2.2.4 to obtain the p-value. Our numerical studies demonstrate that such an approximation can work well.

Thus far we have discussed how to construct our dependency tests sequentially. Now we introduce our test-based variable selection method. In each step of the selection procedure, we compute the current test statistic and the corresponding p-value, and stop the selection when the p-value exceeds a pre-defined level γ. More specifically, our method is implemented in the following way.

1. Set the active set to be $\mathcal{A}_0 = \emptyset$.

2. (a) In the kth step ($k \geq 1$), compute the residuals $\mathbf{r}_j^{(k)} = (I - P_{\mathcal{A}_{k-1}})\mathbf{x}_j$ and $\mathbf{r}^{(k)} = (I - P_{\mathcal{A}_{k-1}})\mathbf{y}$ for each inactive covariate X_j and the response, respectively. Then derive the test statistic $R^{(k)}$ as in (2.16).

(b) Compute the p-value p_k as in (2.7) for independent covariates and (2.14) for dependent covariates.

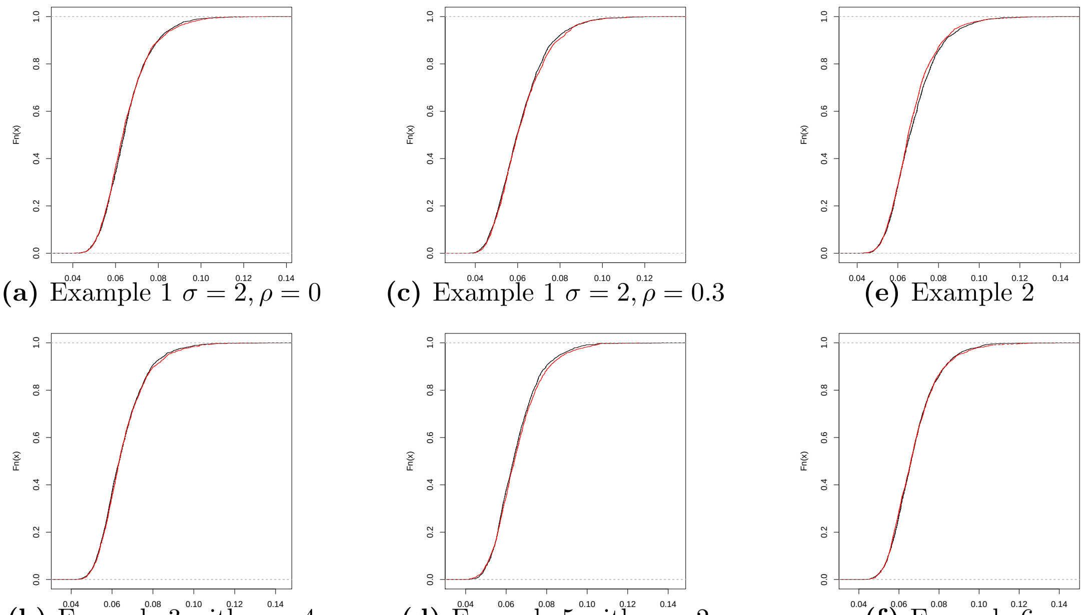
3. If $p_k \leq \gamma$ and $k \leq n-2$, update the active set $\mathcal{A}_k$ and get the estimates of $\boldsymbol{\beta}$ using the same approach as the original selection procedure; otherwise, terminate the procedure.

In the above procedure, the stopping criterion in Step 3 involves a constant level γ. Here we do not provide a specific value of γ, because the choice of an appropriate γ should depend on the goal of the selection, which might vary in different contexts. More specifically, if we aim to detect important variables other than losing any information, we could set a large γ. However, if we want to avoid false discoveries, we should choose a small γ. We will illustrate the effect of γ by simulation examples in Section 2.4. In practice, we also need to determine which null distribution to use in order to obtain the p-value. As mentioned in Section 2.2.4, we first compute the average of the pairwise sample correlation among the covariates, say $\hat{\rho}$, to estimate ρ. If $|\hat{\rho}| < 0.01$, we use (2.7) to compute the p-value; otherwise we apply (2.14) instead.

Our method conducts a sequence of hypothesis tests adaptively until the null hypothesis (2.15) is accepted. Moreover, at each step we perform the dependency test before adding the next variable into the active set, which stands alone from the original variable selection procedure. Hence the proposed method essentially adds (or drops in the LASSO path) the variables one by one in the same order as in the original sequential selection approach. This property makes our method very flexible because it can be incorporated into any sequential selection procedure.

2.3.2 Comparison of empirical distributions

In this subsection, we compare the empirical cumulative distribution functions (cdf) of $R_{\mathcal{M}}$ and $R_{\mathcal{M}}|\hat{\mathcal{M}} = \mathcal{M}$ under the null hypothesis for all simulated experiments introduced in Section 2.4.1. In each example we consider one scenario. To obtain the empirical cdfs, we generate 5000 datasets from the linear model for each scenario. Let $\mathcal{M}^*$ denote the index set of all important variables of size s_0. We implement LARS-Corr $s_0 + 1$ steps and obtain the set of selected variables $\hat{\mathcal{M}}$. Among the 5000 datasets, we find all with $\hat{\mathcal{M}} = \mathcal{M}^*$ and record



(a) Example 1 $\sigma = 2, \rho = 0$ **(c)** Example 1 $\sigma = 2, \rho = 0.3$ **(e)** Example 2

(b) Example 3 with $\sigma = 4$ **(d)** Example 5 with $\sigma = 2$ **(f)** Example 6

Figure 2.2: Empirical cdfs for $R_{\mathcal{M}}$ and $R_{\mathcal{M}}|\hat{\mathcal{M}} = \mathcal{M}$. Each panel compares the empirical cdfs of $R_{\mathcal{M}}$ (the red curve) and $R_{\mathcal{M}}|\hat{\mathcal{M}} = \mathcal{M}$ (the black curve).

the corresponding test statistic at step $s_0 + 1$. Then we use such test statistics to obtain the empirical cdfs of $R_{\mathcal{M}}|\hat{\mathcal{M}} = \mathcal{M}$. For each of the datasets such that $\hat{\mathcal{M}} = \mathcal{M}^*$, we simulate an independent dataset of the same size and directly calculate $R_{\mathcal{M}}$ with $\mathcal{M} = \mathcal{M}^*$ to generate the empirical cdfs of $R_{\mathcal{M}}$.

We illustrate the empirical cdfs of $R_{\mathcal{M}}$, $R_{\mathcal{M}}|\hat{\mathcal{M}} = \mathcal{M}$ in Figure 2.2. One can see that the differences between the two empirical cdfs are very small. It implies that our proposed testing scheme is not much affected by $\hat{\mathcal{M}}$ for later steps during the sequential selection procedures.

2.3.3 Prostate cancer data example

In Section 2.3.1, we have discussed how to implement our test-based variable selection approach in sequential selection procedures. To better illustrate how our method works, we apply it to the prostate cancer data, which has been well studied in the literature (Tibshirani, 1996). This dataset contains 97 observations and eight predictor variables, of which 67 are

Table 2.1: Testing-based LARS procedure applied to the prostate cancer data. For each step, we report the variable selected by LARS, the active set $\mathcal{A}_{k-1}$ in null hypothesis (2.15) and the p-value obtained from our testing approach. The stepwise p-value is calculated before the selected variable enters the candidate model.

Step	Variable Selected	Active Set $\mathcal{A}_k$	p-value
0		$\emptyset$	0.0000
1	lcavol	1	0.0010
2	lweight	1, 2	0.0791
3	svi	1, 2, 5	0.0645
4	lbph	1, 2, 5, 4	0.2996
5	pgg45	1, 2, 5, 4, 8	0.9482
6	age	1, 2, 5, 4, 8, 3	0.7591
7	lcp	1, 2, 5, 4, 8, 3, 6	0.5681
8	gleason	1, 2, 5, 4, 8, 3, 6, 7	

training samples. The goal of the study is to predict the logarithm of prostate-specific antigen level (*lpsa*) of men who were about to receive a radical prostatectomy.

We incorporate our approach into LARS and perform the variable selection on the training data. At each LARS step, we obtain the variable that enters into the model, the corresponding active set as well as the p-value. As the average of pairwise correlation is about 0.3, we use (2.14) to compute the p-value. The results are reported in Table 2.1. It must be pointed out that the p-value is not associated with each variable, but the inactive set $\mathcal{A}_{k-1}^c$ at each selection step. For example, the p-value 0.0010 at Step 1 means that given the selected variable *lcavol*, there is strong evidence that there is at least one important variable in the inactive set $\mathcal{A}_1^c$. If one sets the constant level γ described in Section 2.3.1 to be 0.1, the selected variables are *lcavol*, *lweight* and *svi*; if γ is increased to 0.5, there is one more variable *lbph* added into the final model.

2.4 Numerical studies

In this section, we explore the performance of our method in terms of both simulation and real data studies. We incorporate the proposed approach into sequential selection procedures and compare the results with that using 10-fold CV to conduct model selection for each particular procedure.

2.4.1 Simulation study

In our simulation experiments, we consider three sequential selection procedures: LARS, LASSO and FSR. When our test-based approach is incorporated into a particular procedure, we denote the corresponding variable selection method as LARS-Corr. Similarly we use the notations LASSO-Corr and FSR-Corr to represent our methods integrated with LASSO and FSR, respectively. In addition, we perform permutation tests in each of these three variable selection procedures and denote the corresponding methods by LARS-Perm, LASSO-Perm and FSR-Perm, respectively. For comparison, we use 10-fold CV in LARS, LASSO and FSR to implement model selection. We represent these three CV-based methods by LARS-CV, LASSO-CV and FSR-CV. We also perform the truncated Gaussian tests (Fithian et al., 2015) introduced in Section 2.1 in the sequential selection procedures LARS and FSR, denoted as LARS-TG, FSR-TG, respectively. For permutation tests, we implement 500 permutations.

Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^\top$ denote the estimated coefficient vector. We evaluate the variable selection accuracy by two quantities: False Negatives (FN) and False Positives (FP), respectively defined as

$$FN = \sum_{j=1}^{p} \mathbf{1}(\hat{\beta}_j = 0) \times \mathbf{1}(\beta_j \neq 0) \quad \text{and} \quad FP = \sum_{j=1}^{p} \mathbf{1}(\hat{\beta}_j \neq 0) \times \mathbf{1}(\beta_j = 0),$$

where $\mathbf{1}$ denotes an indicator function.

We consider three simulated examples to generate the response variable. For the first two examples, the covariate vector $\mathbf{X}$ is generated from a p-dimensional Gaussian distribution $\mathcal{N}(0, \Sigma)$ with correlation matrix $\Sigma = (\rho_{i,j})$. For the third example, we aim to assess the robustness of our procedure, and therefore we generate independent covariates and random noise from a central Student t distribution with 5 degrees of freedom to distinguish from Gaussian noise. Throughout the simulation experiments, we fix $p = 2000$. We generate 100 simulated datasets with $n = 200$ observations from each model. In each replication, given a set of selected variables, we refit a linear model on these variables and calculate the out-of-sample mean squared errors (MSE) using an independent test dataset with 500 observations. More

Table 2.2: Results for simulated Example 1 with $\rho = 0$ and $\sigma = 2$. For each method, we report the average MSE, FN, FP and computational time (in seconds) over 100 replications (with standard errors given in parentheses). For our approaches, we show the results with $\gamma = 0.01, 0.05, 0.2, 0.5$ in the stopping criterion described in Section 2.3.1. For each sequential selection procedure, we highlight the smallest MSE and run time (in seconds) in bold font. One can see that the performance of the proposed method is competitive to CV and is more computationally efficient.

Methods	γ	MSE	FN	FP	Time
LARS-CV		4.35 (0.05)	0.00 (0.00)	1.85 (0.32)	28.37 (0.15)
LARS-Perm	0.01	**4.05 (0.03)**	0.00 (0.00)	0.01 (0.01)	5.70 (0.04)
LARS-Perm	0.05	4.08 (0.03)	0.00 (0.00)	0.10 (0.04)	5.82 (0.07)
LARS-Perm	0.2	4.15 (0.04)	0.00 (0.00)	0.40 (0.08)	6.29 (0.13)
LARS-Perm	0.5	4.36 (0.05)	0.00 (0.00)	2.04 (0.41)	8.70 (0.58)
LARS-Corr	0.01	**4.05 (0.03)**	0.00 (0.00)	0.00 (0.00)	0.76 (0.02)
LARS-Corr	0.05	4.07 (0.03)	0.00 (0.00)	0.08 (0.03)	**0.70 (0.01)**
LARS-Corr	0.2	4.13 (0.04)	0.00 (0.00)	0.32 (0.08)	0.83 (0.02)
LARS-Corr	0.5	4.33 (0.05)	0.00 (0.00)	1.44 (0.22)	1.03 (0.05)
LARS-TG	0.01	10.22 (0.08)	1.99 (0.01)	0.00 (0.00)	12.16 (0.12)
LARS-TG	0.05	9.89 (0.13)	1.90 (0.03)	0.00 (0.00)	12.17 (0.12)
LARS-TG	0.2	8.89 (0.23)	1.62 (0.06)	0.01 (0.01)	12.41 (0.14)
LARS-TG	0.5	6.85 (0.26)	0.97 (0.08)	0.23 (0.06)	12.31 (0.13)
LASSO-CV		4.70 (0.06)	0.00 (0.00)	4.78 (0.70)	39.74 (0.58)
LASSO-Perm	0.01	4.07 (0.03)	0.00 (0.00)	0.00 (0.00)	5.67 (0.04)
LASSO-Perm	0.05	4.08 (0.03)	0.00 (0.00)	0.03 (0.02)	5.72 (0.06)
LASSO-Perm	0.2	4.17 (0.04)	0.00 (0.00)	0.40 (0.09)	6.31 (0.14)
LASSO-Perm	0.5	4.36 (0.05)	0.00 (0.00)	1.76 (0.32)	8.35 (0.45)
LASSO-Corr	0.01	**4.07 (0.03)**	0.00 (0.00)	0.00 (0.00)	0.70 (0.01)
LASSO-Corr	0.05	4.08 (0.03)	0.00 (0.00)	0.02 (0.01)	**0.70 (0.00)**
LASSO-Corr	0.2	4.13 (0.03)	0.00 (0.00)	0.25 (0.06)	0.83 (0.02)
LASSO-Corr	0.5	4.34 (0.04)	0.00 (0.00)	1.46 (0.24)	1.07 (0.07)
FSR-CV		**4.05 (0.03)**	0.00 (0.00)	0.01 (0.01)	12.24 (0.12)
FSR-Perm	0.01	**4.05 (0.03)**	0.00 (0.00)	0.00 (0.00)	5.55 (0.04)
FSR-Perm	0.05	4.07 (0.03)	0.00 (0.00)	0.07 (0.03)	5.66 (0.06)
FSR-Perm	0.2	4.12 (0.03)	0.00 (0.00)	0.26 (0.05)	5.88 (0.08)
FSR-Perm	0.5	4.31 (0.04)	0.00 (0.00)	0.97 (0.12)	6.76 (0.16)
FSR-Corr	0.01	**4.05 (0.03)**	0.00 (0.00)	0.00 (0.00)	0.74 (0.02)
FSR-Corr	0.05	4.07 (0.03)	0.00 (0.00)	0.06 (0.02)	**0.73 (0.02)**
FSR-Corr	0.2	4.12 (0.03)	0.00 (0.00)	0.23 (0.05)	0.76 (0.02)
FSR-Corr	0.5	4.27 (0.04)	0.00 (0.00)	0.80 (0.09)	0.89 (0.03)
FSR-TG	0.01	10.13 (0.10)	1.96 (0.02)	0.00 (0.00)	11.30 (0.08)
FSR-TG	0.05	10.00 (0.12)	1.93 (0.03)	0.00 (0.00)	11.30 (0.07)
FSR-TG	0.2	9.26 (0.20)	1.73 (0.05)	0.00 (0.00)	11.46 (0.10)
FSR-TG	0.5	7.20 (0.26)	1.08 (0.08)	0.24 (0.07)	11.48 (0.08)

specifically, we fit a regression model using the selected variables as the covariates on the test dataset and calculate $\frac{1}{500}\sum_{i=1}^{500}(y_i - \hat{y_i})^2$. The details of the simulation examples are as follows.

Example 1. We generate the response from the following sparse linear model $Y = 3X_1 - 1.5X_2 + 2X_3 + \varepsilon$, where the covariates have equal pairwise correlation, i.e., $\rho_{i,j} = \mathrm{Corr}(\mathrm{X_i}, \mathrm{X_j}) = \rho$ for all $i \neq j$. We set $\rho = 0$ for independent covariates and $\rho = 0.3$ for dependent covariates. We also consider $\sigma = 2$ for strong signal and $\sigma = 6$ for weak signal.

Table 2.3: Results for simulated Example 1 with $\rho = 0$ and $\sigma = 6$. The format of the table is the same as Table 2.2.

Methods	γ	MSE	FN	FP	Time
LARS-CV		41.33 (0.48)	1.24 (0.09)	0.82 (0.23)	27.31 (0.14)
LARS-Perm	0.01	40.33 (0.39)	1.40 (0.07)	0.03 (0.02)	3.88 (0.11)
LARS-Perm	0.05	40.05 (0.40)	1.25 (0.07)	0.14 (0.04)	4.23 (0.13)
LARS-Perm	0.2	**39.88 (0.41)**	1.01 (0.06)	0.48 (0.08)	5.06 (0.17)
LARS-Perm	0.5	41.31 (0.49)	0.75 (0.06)	1.99 (0.33)	7.66 (0.50)
LARS-Corr	0.01	40.37 (0.38)	1.42 (0.07)	0.02 (0.01)	**0.44 (0.02)**
LARS-Corr	0.05	40.10 (0.39)	1.27 (0.07)	0.13 (0.04)	0.47 (0.02)
LARS-Corr	0.2	39.90 (0.41)	1.02 (0.06)	0.47 (0.09)	0.61 (0.03)
LARS-Corr	0.5	41.26 (0.48)	0.78 (0.06)	1.60 (0.20)	0.97 (0.06)
LARS-TG	0.01	43.57 (0.36)	2.11 (0.03)	0.01 (0.01)	12.86 (0.22)
LARS-TG	0.05	43.26 (0.34)	2.06 (0.03)	0.02 (0.02)	13.07 (0.23)
LARS-TG	0.2	42.48 (0.32)	1.91 (0.04)	0.05 (0.03)	13.03 (0.22)
LARS-TG	0.5	41.70 (0.36)	1.53 (0.06)	0.43 (0.08)	13.11 (0.22)
LASSO-CV		42.33 (0.54)	1.16 (0.08)	1.94 (0.60)	35.52 (0.24)
LASSO-Perm	0.01	41.21 (0.38)	1.56 (0.06)	0.01 (0.01)	3.58 (0.11)
LASSO-Perm	0.05	40.34 (0.36)	1.28 (0.06)	0.09 (0.04)	4.14 (0.13)
LASSO-Perm	0.2	40.10 (0.38)	1.02 (0.06)	0.43 (0.08)	4.97 (0.17)
LASSO-Perm	0.5	41.46 (0.45)	0.76 (0.07)	1.73 (0.24)	7.23 (0.39)
LASSO-Corr	0.01	41.30 (0.38)	1.58 (0.06)	0.01 (0.01)	**0.39 (0.01)**
LASSO-Corr	0.05	40.36 (0.36)	1.30 (0.06)	0.06 (0.03)	0.47 (0.02)
LASSO-Corr	0.2	**40.00 (0.38)**	1.02 (0.06)	0.39 (0.08)	0.61 (0.02)
LASSO-Corr	0.5	41.41 (0.44)	0.80 (0.06)	1.49 (0.18)	0.88 (0.04)
FSR-CV		40.72 (0.40)	1.46 (0.07)	0.04 (0.02)	11.63 (0.15)
FSR-Perm	0.01	40.72 (0.38)	1.51 (0.06)	0.01 (0.01)	3.55 (0.09)
FSR-Perm	0.05	39.75 (0.34)	1.17 (0.06)	0.07 (0.03)	4.12 (0.10)
FSR-Perm	0.2	39.66 (0.40)	0.93 (0.06)	0.31 (0.06)	4.74 (0.14)
FSR-Perm	0.5	40.72 (0.44)	0.75 (0.06)	0.98 (0.12)	5.93 (0.22)
FSR-Corr	0.01	40.80 (0.37)	1.54 (0.06)	0.00 (0.00)	**0.39 (0.01)**
FSR-Corr	0.05	39.77 (0.34)	1.19 (0.06)	0.05 (0.02)	0.46 (0.01)
FSR-Corr	0.2	**39.59 (0.38)**	0.95 (0.06)	0.26 (0.05)	0.54 (0.02)
FSR-Corr	0.5	40.38 (0.43)	0.75 (0.06)	0.83 (0.10)	0.68 (0.02)
FSR-TG	0.01	43.32 (0.38)	2.10 (0.03)	0.00 (0.00)	12.03 (0.21)
FSR-TG	0.05	43.12 (0.37)	2.06 (0.03)	0.01 (0.01)	12.17 (0.23)
FSR-TG	0.2	42.58 (0.38)	1.88 (0.05)	0.09 (0.04)	12.13 (0.22)
FSR-TG	0.5	41.93 (0.39)	1.55 (0.07)	0.37 (0.09)	12.23 (0.22)

Example 2. We demonstrate that when the covariates do not have equal pairwise correlations, we can still apply our approach using the approximated null distribution discussed in Section 2.2.4. We simulate data from $Y = 2X_1 + \cdots + 2X_{10} + \varepsilon$, where $\rho_{i,j} = 0.5^{|i-j|}$ for $i \neq j$ and $\sigma = 3$.

Example 3. We demonstrate that our method performs well when the Gaussian assumption is not satisfied. To this end, we consider the same linear relationship as in Example 1, i.e., $Y = 3X_1 - 1.5X_2 + 2X_3 + \sigma\,\varepsilon$, but the X_js and ε are generated independently from the Student t distribution with 5 degrees of freedom. We take 5 degrees of freedom such that the tail is

Table 2.4: Results for simulated Example 1 with $\rho = 0.3$ and $\sigma = 2$. The format of the table is the same as Table 2.2.

Methods	γ	MSE	FN	FP	Time
LARS-CV		4.52 (0.06)	0.00 (0.00)	4.72 (0.86)	28.31 (0.24)
LARS-Perm	0.01	**4.06 (0.03)**	0.00 (0.00)	0.04 (0.02)	5.23 (0.04)
LARS-Perm	0.05	4.07 (0.03)	0.00 (0.00)	0.07 (0.03)	5.26 (0.05)
LARS-Perm	0.2	4.13 (0.04)	0.00 (0.00)	0.33 (0.09)	5.62 (0.12)
LARS-Perm	0.5	4.51 (0.09)	0.00 (0.00)	7.64 (2.81)	15.61 (3.85)
LARS-Corr	0.01	4.08 (0.03)	0.00 (0.00)	0.10 (0.03)	**0.54 (0.01)**
LARS-Corr	0.05	4.09 (0.03)	0.00 (0.00)	0.16 (0.04)	0.55 (0.01)
LARS-Corr	0.2	4.14 (0.03)	0.00 (0.00)	0.47 (0.08)	0.57 (0.01)
LARS-Corr	0.5	4.29 (0.04)	0.00 (0.00)	1.84 (0.36)	0.76 (0.04)
LARS-TG	0.01	8.46 (0.05)	2.00 (0.00)	0.00 (0.00)	10.97 (0.03)
LARS-TG	0.05	8.33 (0.07)	1.95 (0.02)	0.00 (0.00)	11.02 (0.04)
LARS-TG	0.2	7.79 (0.12)	1.72 (0.05)	0.00 (0.00)	11.12 (0.05)
LARS-TG	0.5	6.95 (0.16)	1.35 (0.07)	0.05 (0.03)	11.09 (0.04)
LASSO-CV		4.73 (0.06)	0.00 (0.00)	6.69 (0.83)	38.05 (0.53)
LASSO-Perm	0.01	**4.07 (0.03)**	0.00 (0.00)	0.00 (0.00)	5.67 (0.04)
LASSO-Perm	0.05	4.08 (0.03)	0.00 (0.00)	0.03 (0.02)	5.72 (0.06)
LASSO-Perm	0.2	4.17 (0.04)	0.00 (0.00)	0.40 (0.09)	6.31 (0.14)
LASSO-Perm	0.5	4.36 (0.05)	0.00 (0.00)	1.76 (0.32)	8.35 (0.45)
LASSO-Corr	0.01	4.11 (0.03)	0.00 (0.00)	0.09 (0.03)	**0.55 (0.01)**
LASSO-Corr	0.05	4.12 (0.03)	0.00 (0.00)	0.15 (0.04)	0.56 (0.01)
LASSO-Corr	0.2	4.17 (0.03)	0.00 (0.00)	0.37 (0.06)	0.59 (0.01)
LASSO-Corr	0.5	4.30 (0.04)	0.00 (0.00)	1.58 (0.32)	0.76 (0.03)
FSR-CV		4.07 (0.03)	0.00 (0.00)	0.01 (0.01)	12.45 (0.12)
FSR-Perm	0.01	**4.05 (0.03)**	0.00 (0.00)	0.00 (0.00)	5.55 (0.04)
FSR-Perm	0.05	4.07 (0.03)	0.00 (0.00)	0.07 (0.03)	5.66 (0.06)
FSR-Perm	0.2	4.12 (0.03)	0.00 (0.00)	0.26 (0.05)	5.88 (0.08)
FSR-Perm	0.5	4.31 (0.04)	0.00 (0.00)	0.97 (0.12)	6.76 (0.16)
FSR-Corr	0.01	4.08 (0.03)	0.00 (0.00)	0.03 (0.02)	**0.52 (0.01)**
FSR-Corr	0.05	4.11 (0.03)	0.00 (0.00)	0.12 (0.03)	0.53 (0.01)
FSR-Corr	0.2	4.16 (0.03)	0.00 (0.00)	0.30 (0.05)	0.56 (0.01)
FSR-Corr	0.5	4.63 (0.07)	0.00 (0.00)	2.44 (0.32)	0.94 (0.07)
FSR-TG	0.01	8.49 (0.06)	2.00 (0.00)	0.00 (0.00)	10.62 (0.03)
FSR-TG	0.05	8.37 (0.08)	1.95 (0.02)	0.00 (0.00)	10.69 (0.04)
FSR-TG	0.2	7.92 (0.13)	1.76 (0.05)	0.00 (0.00)	10.73 (0.05)
FSR-TG	0.5	6.85 (0.16)	1.30 (0.07)	0.09 (0.05)	10.77 (0.04)

distinguished from that of Gaussian random variables but is not too heavy. We set $\sigma = 4$ and $\sigma = 8$ to make the signal to noise ratio comparable with Example 1.

The results for the three simulated examples are summarized in Tables 2.2–2.8. In LARS-Corr, LASSO-Corr, permutation and truncated Gaussian tests-based methods, we take $\gamma \in \{0.01, 0.05, 0.2, 0.5\}$. Based on the simulation results, we can draw the following conclusions.

First, the test-based methods LARS-Corr, LASSO-Corr and FSR-Corr outperform the corresponding CV-based methods respectively for all scenarios, and the improvement of performance for our methods is more substantial when the signal is strong. Second, when the covariates are not equally correlated, our approach can still work well using (2.12) as an approximation

Table 2.5: Results for simulated Example 1 with $\rho = 0.3$ and $\sigma = 6$. The format of the table is the same as Table 2.2.

Methods	γ	MSE	FN	FP	Time
LARS-CV		41.89 (0.51)	1.38 (0.07)	2.40 (0.76)	29.13 (0.24)
LARS-Perm	0.01	40.66 (0.31)	1.88 (0.04)	0.02 (0.01)	2.65 (0.06)
LARS-Perm	0.05	40.59 (0.37)	1.70 (0.06)	0.31 (0.13)	3.25 (0.21)
LARS-Perm	0.2	42.83 (0.57)	1.33 (0.08)	5.82 (2.05)	10.89 (2.67)
LARS-Perm	0.5	48.02 (0.94)	0.90 (0.08)	25.78 (5.17)	37.59 (6.84)
LARS-Corr	0.01	40.61 (0.31)	1.77 (0.05)	0.01 (0.01)	**0.26 (0.02)**
LARS-Corr	0.05	40.34 (0.30)	1.62 (0.06)	0.10 (0.03)	0.32 (0.03)
LARS-Corr	0.2	**39.95 (0.33)**	1.41 (0.06)	0.23 (0.05)	0.44 (0.05)
LARS-Corr	0.5	40.39 (0.37)	1.25 (0.06)	1.03 (0.28)	0.56 (0.05)
LARS-TG	0.01	42.21 (0.42)	2.11 (0.03)	0.00 (0.00)	11.47 (0.06)
LARS-TG	0.05	41.73 (0.37)	2.03 (0.03)	0.09 (0.06)	11.46 (0.06)
LARS-TG	0.2	41.46 (0.37)	1.91 (0.04)	0.29 (0.09)	11.46 (0.06)
LARS-TG	0.5	41.45 (0.35)	1.66 (0.05)	1.22 (0.20)	11.82 (0.08)
LASSO-CV		42.26 (0.54)	1.36 (0.07)	2.58 (0.73)	39.07 (0.56)
LASSO-Perm	0.01	41.21 (0.38)	1.56 (0.06)	0.01 (0.01)	3.58 (0.11)
LASSO-Perm	0.05	40.34 (0.36)	1.28 (0.06)	0.09 (0.04)	4.14 (0.13)
LASSO-Perm	0.2	**40.10 (0.38)**	1.02 (0.06)	0.43 (0.08)	4.97 (0.17)
LASSO-Perm	0.5	41.46 (0.45)	0.76 (0.07)	1.73 (0.24)	7.23 (0.39)
LASSO-Corr	0.01	41.38 (0.39)	1.82 (0.05)	0.22 (0.15)	**0.23 (0.01)**
LASSO-Corr	0.05	40.90 (0.40)	1.60 (0.06)	0.36 (0.17)	0.28 (0.02)
LASSO-Corr	0.2	40.62 (0.40)	1.45 (0.06)	0.46 (0.17)	0.43 (0.04)
LASSO-Corr	0.5	40.75 (0.43)	1.25 (0.06)	1.18 (0.32)	0.55 (0.05)
FSR-CV		41.18 (0.42)	1.80 (0.06)	0.05 (0.02)	11.38 (0.19)
FSR-Perm	0.01	40.72 (0.38)	1.51 (0.06)	0.01 (0.01)	3.55 (0.09)
FSR-Perm	0.05	39.75 (0.34)	1.17 (0.06)	0.07 (0.03)	4.12 (0.10)
FSR-Perm	0.2	**39.66 (0.40)**	0.93 (0.06)	0.31 (0.06)	4.74 (0.14)
FSR-Perm	0.5	40.72 (0.44)	0.75 (0.06)	0.98 (0.12)	5.93 (0.22)
FSR-Corr	0.01	40.38 (0.34)	1.78 (0.05)	0.04 (0.02)	**0.23 (0.01)**
FSR-Corr	0.05	40.02 (0.34)	1.58 (0.06)	0.12 (0.04)	0.27 (0.01)
FSR-Corr	0.2	39.91 (0.36)	1.43 (0.06)	0.24 (0.05)	0.32 (0.01)
FSR-Corr	0.5	42.67 (0.92)	1.12 (0.06)	3.79 (1.97)	0.95 (0.33)
FSR-TG	0.01	42.33 (0.43)	2.14 (0.04)	0.02 (0.01)	11.16 (0.06)
FSR-TG	0.05	42.01 (0.40)	2.09 (0.03)	0.04 (0.02)	11.17 (0.06)
FSR-TG	0.2	41.54 (0.41)	1.96 (0.04)	0.10 (0.04)	11.05 (0.04)
FSR-TG	0.5	41.29 (0.42)	1.68 (0.06)	0.41 (0.08)	11.46 (0.08)

for the null distribution. Third, although LARS-Perm, LASSO-Perm and FSR-Perm have comparable performance to LARS-Corr, LASSO-Corr and FSR-Perm, respectively, they carry more computational costs. In addition, note that the permutation test can have much larger FP in some scenarios (e.g., LARS-Perm in Tables 2.4–2.5). Fourth, although the truncated Gaussian tests have smaller false positives, their power is not very large. Therefore, the false negatives are still quite large even when $\gamma = 0.5$. As a result, the prediction errors are not well controlled. Finally, throughout the simulation experiments, the computational time of our methods drops dramatically compared with CV and permutation test.

From Examples 1–3, one can see that our methods can control FN and FP by choosing a proper value of γ. We illustrate how the performance changes as the value of γ varies for two

Table 2.6: Results for simulated Example 2. The format of the table is the same as Table 2.2.

Methods	γ	MSE	FN	FP	Time
LARS-CV		10.78 (0.14)	0.00 (0.00)	4.04 (0.50)	27.25 (0.19)
LARS-Perm	0.01	9.57 (0.09)	0.04 (0.02)	0.04 (0.02)	14.76 (0.11)
LARS-Perm	0.05	9.61 (0.09)	0.02 (0.01)	0.18 (0.07)	14.95 (0.15)
LARS-Perm	0.2	9.85 (0.11)	0.01 (0.01)	0.59 (0.12)	15.69 (0.19)
LARS-Perm	0.5	10.51 (0.13)	0.01 (0.01)	2.60 (0.36)	18.24 (0.52)
LARS-Corr	0.01	9.56 (0.09)	0.02 (0.01)	0.11 (0.06)	**1.73 (0.01)**
LARS-Corr	0.05	**9.55 (0.08)**	0.01 (0.01)	0.12 (0.07)	1.74 (0.02)
LARS-Corr	0.2	9.77 (0.09)	0.01 (0.01)	0.52 (0.12)	1.80 (0.03)
LARS-Corr	0.5	10.25 (0.13)	0.01 (0.01)	3.67 (1.88)	2.40 (0.37)
LARS-TG	0.01	12.36 (0.14)	1.98 (0.02)	0.00 (0.00)	12.38 (0.15)
LARS-TG	0.05	12.32 (0.14)	1.95 (0.03)	0.02 (0.01)	12.24 (0.13)
LARS-TG	0.2	11.83 (0.10)	1.75 (0.04)	0.10 (0.04)	12.41 (0.13)
LARS-TG	0.5	11.51 (0.11)	1.48 (0.05)	0.45 (0.10)	12.46 (0.14)
LASSO-CV		12.02 (0.12)	0.00 (0.00)	10.41 (0.84)	40.09 (0.30)
LASSO-Perm	0.01	**9.57 (0.08)**	0.02 (0.01)	0.03 (0.02)	14.68 (0.09)
LASSO-Perm	0.05	9.61 (0.08)	0.01 (0.01)	0.14 (0.07)	14.86 (0.12)
LASSO-Perm	0.2	10.02 (0.13)	0.01 (0.01)	1.25 (0.45)	16.60 (0.76)
LASSO-Perm	0.5	10.75 (0.15)	0.01 (0.01)	3.59 (0.61)	19.33 (0.78)
LASSO-Corr	0.01	**9.57 (0.08)**	0.01 (0.01)	0.09 (0.06)	**1.72 (0.01)**
LASSO-Corr	0.05	9.65 (0.09)	0.01 (0.01)	0.33 (0.17)	1.76 (0.03)
LASSO-Corr	0.2	9.90 (0.11)	0.01 (0.01)	1.01 (0.43)	1.87 (0.08)
LASSO-Corr	0.5	10.4 (0.14)	0.01 (0.01)	2.56 (0.51)	2.17 (0.09)
FSR-CV		12.04 (0.34)	0.89 (0.12)	0.48 (0.09)	11.87 (0.11)
FSR-Perm	0.01	10.84 (0.28)	0.40 (0.09)	0.55 (0.12)	14.87 (0.22)
FSR-Perm	0.05	**10.37 (0.19)**	0.16 (0.05)	0.92 (0.19)	15.64 (0.25)
FSR-Perm	0.2	10.60 (0.21)	0.12 (0.04)	1.47 (0.26)	16.59 (0.36)
FSR-Perm	0.5	11.16 (0.25)	0.09 (0.03)	2.56 (0.36)	17.84 (0.47)
FSR-Corr	0.01	10.46 (0.20)	0.22 (0.05)	0.72 (0.15)	**1.72 (0.02)**
FSR-Corr	0.05	10.39 (0.20)	0.14 (0.04)	1.04 (0.21)	1.81 (0.03)
FSR-Corr	0.2	10.69 (0.25)	0.09 (0.03)	1.77 (0.35)	1.91 (0.06)
FSR-Corr	0.5	11.26 (0.30)	0.13 (0.05)	2.87 (0.46)	2.11 (0.07)
FSR-TG	0.01	12.24 (0.13)	1.96 (0.02)	0.00 (0.00)	11.58 (0.13)
FSR-TG	0.05	12.21 (0.13)	1.94 (0.03)	0.00 (0.00)	11.53 (0.12)
FSR-TG	0.2	11.67 (0.13)	1.67 (0.06)	0.03 (0.02)	11.67 (0.12)
FSR-TG	0.5	11.04 (0.15)	1.19 (0.08)	0.24 (0.08)	11.77 (0.13)

scenarios in Figure 2.3. This figure shows that as γ increases, the FP of our methods has an increasing trend while the FN will decrease. Furthermore, our approach always outperforms CV in terms of MSE and computational time as γ varies.

For independent cases, we also evaluate the performance of the proposed method using the maximal t-statistic described in (2.8). We find that the performance of our method with the maximal t-statistic is only slightly better than that with the maximal absolute correlation as the test statistic. Hence we do not include the detailed simulation results for the maximal t-statistic in this thesis.

Table 2.7: Results for simulated Example 3 with $\sigma = 4$. The format of the table is the same as Table 2.2.

Methods	γ	MSE	FN	FP	Time
LARS-CV		17.65 (0.21)	0.00 (0.00)	2.15 (0.37)	26.57 (0.23)
LARS-Perm	0.01	16.25 (0.13)	0.02 (0.01)	0.00 (0.00)	5.74 (0.05)
LARS-Perm	0.05	16.28 (0.13)	0.01 (0.01)	0.04 (0.03)	5.81 (0.07)
LARS-Perm	0.2	16.64 (0.15)	0.00 (0.00)	0.43 (0.10)	6.26 (0.16)
LARS-Perm	0.5	17.49 (0.24)	0.00 (0.00)	3.58 (1.96)	10.65 (2.68)
LARS-Corr	0.01	**16.25 (0.13)**	0.02 (0.01)	0.00 (0.00)	0.93 (0.02)
LARS-Corr	0.05	16.28 (0.13)	0.01 (0.01)	0.04 (0.03)	**0.88 (0.01)**
LARS-Corr	0.2	16.61 (0.15)	0.00 (0.00)	0.39 (0.09)	1.04 (0.04)
LARS-Corr	0.5	17.26 (0.19)	0.00 (0.00)	1.27 (0.21)	1.21 (0.06)
LARS-TG	0.01	38.07 (0.57)	2.00 (0.03)	0.00 (0.00)	12.00 (0.06)
LARS-TG	0.05	37.71 (0.56)	1.95 (0.04)	0.02 (0.02)	12.02 (0.06)
LARS-TG	0.2	36.69 (0.50)	1.80 (0.05)	0.04 (0.03)	11.95 (0.05)
LARS-TG	0.5	34.24 (0.57)	1.24 (0.08)	0.51 (0.14)	12.30 (0.09)
LASSO-CV		18.16 (0.24)	0.00 (0.00)	3.27 (0.49)	44.60 (0.51)
LASSO-Perm	0.01	16.44 (0.13)	0.03 (0.02)	0.02 (0.01)	5.70 (0.05)
LASSO-Perm	0.05	16.41 (0.11)	0.01 (0.01)	0.06 (0.02)	5.77 (0.06)
LASSO-Perm	0.2	16.65 (0.14)	0.00 (0.00)	0.34 (0.09)	6.10 (0.13)
LASSO-Perm	0.5	17.42 (0.21)	0.00 (0.00)	3.21 (1.94)	11.66 (4.35)
LASSO-Corr	0.01	**16.37 (0.12)**	0.02 (0.01)	0.00 (0.00)	1.07 (0.02)
LASSO-Corr	0.05	16.41 (0.11)	0.01 (0.01)	0.05 (0.02)	**1.00 (0.02)**
LASSO-Corr	0.2	16.64 (0.14)	0.00 (0.00)	0.33 (0.09)	1.03 (0.03)
LASSO-Corr	0.5	17.22 (0.17)	0.00 (0.00)	1.08 (0.19)	1.08 (0.04)
FSR-CV		16.54 (0.17)	0.07 (0.03)	0.05 (0.03)	13.29 (0.19)
FSR-Perm	0.01	16.42 (0.15)	0.06 (0.02)	0.01 (0.01)	5.56 (0.06)
FSR-Perm	0.05	16.37 (0.13)	0.02 (0.01)	0.07 (0.03)	5.69 (0.07)
FSR-Perm	0.2	16.54 (0.14)	0.00 (0.00)	0.24 (0.06)	5.85 (0.08)
FSR-Perm	0.5	17.13 (0.17)	0.00 (0.00)	0.79 (0.10)	6.47 (0.13)
FSR-Corr	0.01	16.41 (0.15)	0.06 (0.02)	0.00 (0.00)	0.96 (0.02)
FSR-Corr	0.05	**16.34 (0.12)**	0.02 (0.01)	0.05 (0.02)	**0.85 (0.01)**
FSR-Corr	0.2	16.50 (0.13)	0.00 (0.00)	0.21 (0.05)	1.00 (0.03)
FSR-Corr	0.5	17.00 (0.17)	0.00 (0.00)	0.68 (0.09)	1.16 (0.04)
FSR-TG	0.01	37.40 (0.46)	1.96 (0.03)	0.00 (0.00)	10.67 (0.05)
FSR-TG	0.05	37.28 (0.46)	1.94 (0.03)	0.00 (0.00)	10.68 (0.05)
FSR-TG	0.2	35.92 (0.47)	1.71 (0.06)	0.02 (0.01)	10.67 (0.04)
FSR-TG	0.5	34.17 (0.54)	1.33 (0.08)	0.17 (0.05)	11.00 (0.06)

2.5 Robustness of the proposed method

In this subsection, we consider several additional simulation experiments for LARS, LASSO and FSR, especially when the assumptions of independence or equal correlation among covariates are not satisfied.

Example 4. In this example, all the setups are the same as Example 2 in Section 2.4.1, except that the correlation structure is $\mathrm{Corr}(\mathrm{X_i}, \mathrm{X_j}) = 0.9^{|\mathrm{i}-\mathrm{j}|}$, and we consider $\sigma = 1$ and 3 respectively.

Example 5. In this example, all the other setups are the same as Example 1 in Section 2.4.1, except that the covariance matrix of X_j's is generated by LL', where $L = (l_{ij})_{p \times p}$ is a

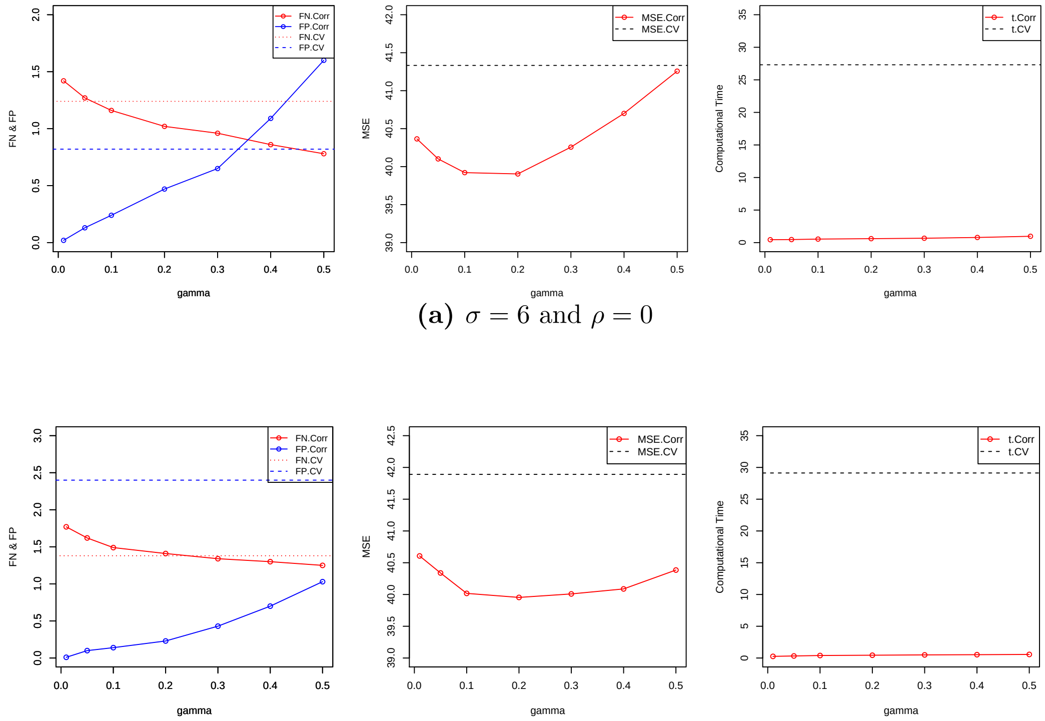
Table 2.8: Results for simulated Example 3 with $\sigma = 8$. The format of the table is the same as Table 2.2.

Methods	γ	MSE	FN	FP	Time
LARS-CV		76.59 (0.97)	1.62 (0.10)	0.97 (0.30)	25.94 (0.08)
LARS-Perm	0.01	73.71 (0.71)	1.78 (0.06)	0.00 (0.00)	3.03 (0.09)
LARS-Perm	0.05	72.04 (0.74)	1.43 (0.07)	0.07 (0.04)	3.62 (0.12)
LARS-Perm	0.2	**71.75 (0.77)**	1.15 (0.07)	0.41 (0.10)	4.47 (0.18)
LARS-Perm	0.5	74.19 (0.94)	0.93 (0.07)	3.39 (1.95)	8.68 (2.54)
LARS-Corr	0.01	73.72 (0.76)	1.76 (0.06)	0.01 (0.01)	**0.43 (0.02)**
LARS-Corr	0.05	72.23 (0.74)	1.50 (0.07)	0.03 (0.02)	0.55 (0.02)
LARS-Corr	0.2	71.99 (0.78)	1.19 (0.07)	0.38 (0.09)	0.71 (0.03)
LARS-Corr	0.5	73.61 (0.80)	0.93 (0.07)	1.22 (0.15)	1.07 (0.05)
LARS-TG	0.01	125.93 (1.84)	2.47 (0.05)	0.01 (0.01)	12.11 (0.08)
LARS-TG	0.05	124.84 (1.82)	2.35 (0.05)	0.04 (0.02)	12.26 (0.10)
LARS-TG	0.2	124.42 (1.78)	2.25 (0.05)	0.15 (0.05)	12.26 (0.10)
LARS-TG	0.5	124.51 (1.72)	2.00 (0.06)	0.55 (0.10)	12.29 (0.09)
LASSO-CV		72.70 (1.26)	1.49 (0.11)	1.06 (0.42)	40.59 (0.30)
LASSO-Perm	0.01	73.83 (0.76)	1.71 (0.07)	0.01 (0.01)	3.05 (0.10)
LASSO-Perm	0.05	72.33 (0.72)	1.43 (0.08)	0.07 (0.03)	3.53 (0.12)
LASSO-Perm	0.2	72.21 (0.72)	1.19 (0.08)	0.34 (0.08)	4.22 (0.18)
LASSO-Perm	0.5	74.24 (0.85)	0.95 (0.07)	3.14 (1.93)	10.46 (4.80)
LASSO-Corr	0.01	73.91 (0.79)	1.71 (0.07)	0.02 (0.01)	**0.51 (0.02)**
LASSO-Corr	0.05	72.60 (0.74)	1.49 (0.07)	0.03 (0.02)	0.53 (0.02)
LASSO-Corr	0.2	**72.09 (0.72)**	1.19 (0.08)	0.31 (0.08)	0.71 (0.03)
LASSO-Corr	0.5	73.83 (0.77)	0.98 (0.07)	1.18 (0.16)	1.01 (0.05)
FSR-CV		73.01 (0.71)	1.59 (0.07)	0.07 (0.03)	13.25 (0.19)
FSR-Perm	0.01	72.76 (0.72)	1.63 (0.06)	0.01 (0.01)	3.18 (0.09)
FSR-Perm	0.05	71.90 (0.71)	1.41 (0.07)	0.06 (0.02)	3.53 (0.10)
FSR-Perm	0.2	**71.18 (0.75)**	1.12 (0.07)	0.23 (0.05)	4.08 (0.12)
FSR-Perm	0.5	73.63 (0.98)	0.95 (0.07)	0.97 (0.12)	5.26 (0.17)
FSR-Corr	0.01	72.73 (0.73)	1.63 (0.06)	0.02 (0.01)	**0.49 (0.02)**
FSR-Corr	0.05	71.80 (0.71)	1.41 (0.07)	0.05 (0.02)	0.55 (0.02)
FSR-Corr	0.2	71.43 (0.75)	1.17 (0.07)	0.22 (0.05)	0.66 (0.03)
FSR-Corr	0.5	73.30 (0.92)	0.96 (0.07)	0.90 (0.11)	0.87 (0.03)
FSR-TG	0.01	125.42 (1.52)	2.45 (0.05)	0.00 (0.00)	10.89 (0.07)
FSR-TG	0.05	124.59 (1.55)	2.36 (0.05)	0.03 (0.02)	10.94 (0.08)
FSR-TG	0.2	123.48 (1.52)	2.24 (0.06)	0.09 (0.03)	10.93 (0.08)
FSR-TG	0.5	123.87 (1.57)	1.93 (0.07)	0.45 (0.09)	11.01 (0.08)

lower triangular matrix with $l_{ij} = (2B_{ij} - 1) \cdot U_{ij}$ for $i \geq j$ and 0 otherwise. Here B_{ij}'s are i.i.d. Bernoulli random variables with success probability 0.5, and U_{ij}'s follow a uniform distribution on $[0, 1]$. As in Example 1, we consider $\sigma = 2$ for strong signal and $\sigma = 6$ for weak signal.

Example 6. In this example, all the other setups are the same as Example 2 in Section 2.4.1, except that we consider different covariance structure for the covariates. Specifically, we set $\mathrm{Corr}(\mathrm{X_i}, \mathrm{X_j}) = 0.6$ for $1 \leq i < j \leq 5$ and $6 \leq i < j \leq 10$ and 0 for the other $i \neq j$. In other words, important variables are highly correlated within several groups.

We show the results for Example 4 in Tables 2.9 and 2.10. One can see that in terms of prediction error, LASSO-Corr and LASSO-Perm outperform LASSO-CV; while LARS-Corr



(a) $\sigma = 6$ and $\rho = 0$

(b) $\sigma = 6$ and $\rho = 0.3$

Figure 2.3: Performance of LARS-Corr and LARS-CV in simulated example 1 with (a) $\sigma = 6$ and $\rho = 0$ and (b) $\sigma = 6$ and $\rho = 0.3$. In LARS-Corr, and $\gamma \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$. For all three panels, the solid curve corresponds to LARS-Corr and the dashed curve corresponds to LARS-CV. In the first panel of (a) and (b), the red curves represent FN while the blue ones represent FP.

and LARS-Perm have comparable performance to LARS-CV. Moreover, Corr-type methods are much more computationally efficient. Note that FSR-Corr and FSR-Perm may not work well compared with FSR-CV. This is, however, resulted from the failure of the FSR procedure. With the existence of strong multicollinearity, and FSR implements full regression at each step of selection, unimportant variables tend to enter the model before active covariates along the solution path, adding more difficulty to testing-based approaches (Derksen and Keselman, 1992). Due to the same reason, although with smaller MSE than FSR-Corr and FSR-Perm, FSR-CV performs much worse than procedures using CV-based LARS or LASSO for selection.

We display the simulation outputs for Example 5 in Tables 2.11 and 2.12. In Corr-type methods, since the average of off-diagonal elements in the correlation matrix of the covariates is close to 0, we use (2.7) to obtain the p-values. The improvement of performance for our methods is significant when the signal is strong. As the noise level increase, Corr-type methods are still

Table 2.9: Results for simulated Example 4 with $\sigma = 1$. The format of the table is the same as Table 2.2.

Methods	γ	MSE	FN	FP	Time
LARS-CV		1.07 (0.01)	0.00 (0.00)	0.64 (0.17)	24.31 (0.22)
LARS-Perm	0.01	**1.05 (0.01)**	0.00 (0.00)	0.00 (0.00)	15.78 (0.10)
LARS-Perm	0.05	1.06 (0.01)	0.00 (0.00)	0.06 (0.03)	16.04 (0.16)
LARS-Perm	0.2	1.08 (0.01)	0.00 (0.00)	0.43 (0.08)	16.50 (0.18)
LARS-Perm	0.5	1.12 (0.01)	0.00 (0.00)	1.51 (0.35)	18.00 (0.49)
LARS-Corr	0.01	**1.05 (0.01)**	0.00 (0.00)	0.01 (0.01)	**1.82 (0.01)**
LARS-Corr	0.05	1.06 (0.01)	0.00 (0.00)	0.07 (0.04)	1.83 (0.01)
LARS-Corr	0.2	1.07 (0.01)	0.00 (0.00)	0.31 (0.07)	1.88 (0.02)
LARS-Corr	0.5	1.10 (0.01)	0.00 (0.00)	0.85 (0.12)	1.95 (0.02)
LARS-TG	0.01	38.78 (1.27)	8.18 (0.08)	0.00 (0.00)	11.99 (0.10)
LARS-TG	0.05	34.92 (1.37)	7.83 (0.10)	0.00 (0.00)	11.99 (0.10)
LARS-TG	0.2	27.23 (1.45)	7.08 (0.14)	0.00 (0.00)	12.17 (0.10)
LARS-TG	0.5	18.33 (1.48)	5.59 (0.22)	0.05 (0.04)	12.17 (0.09)
LASSO-CV		1.49 (0.01)	0.00 (0.00)	25.42 (0.50)	48.19 (0.45)
LASSO-Perm	0.01	**1.06 (0.01)**	0.00 (0.00)	0.00 (0.00)	15.83 (0.10)
LASSO-Perm	0.05	**1.06 (0.01)**	0.00 (0.00)	0.03 (0.02)	16.02 (0.15)
LASSO-Perm	0.2	1.08 (0.01)	0.00 (0.00)	0.33 (0.08)	16.37 (0.19)
LASSO-Perm	0.5	1.14 (0.02)	0.00 (0.00)	3.68 (1.89)	24.17 (5.60)
LASSO-Corr	0.01	**1.06 (0.01)**	0.00 (0.00)	0.01 (0.01)	1.82 (0.01)
LASSO-Corr	0.05	**1.06 (0.01)**	0.00 (0.00)	0.02 (0.01)	**1.81 (0.01)**
LASSO-Corr	0.2	1.07 (0.01)	0.00 (0.00)	0.20 (0.06)	1.86 (0.02)
LASSO-Corr	0.5	1.10 (0.01)	0.00 (0.00)	0.88 (0.13)	1.95 (0.02)
FSR-CV		**4.80 (0.14)**	5.38 (0.06)	0.80 (0.10)	9.61 (0.06)
FSR-Perm	0.01	8.37 (0.30)	5.04 (0.07)	32.57 (1.26)	55.11 (1.77)
FSR-Perm	0.05	9.09 (0.31)	5.04 (0.07)	39.95 (1.42)	65.88 (2.04)
FSR-Perm	0.2	9.82 (0.31)	5.04 (0.07)	52.97 (2.85)	83.82 (3.92)
FSR-Perm	0.5	10.94 (0.38)	5.04 (0.07)	98.50 (6.82)	146.89 (9.50)
FSR-Corr	0.01	8.58 (0.30)	5.04 (0.07)	34.37 (1.24)	**6.85 (0.22)**
FSR-Corr	0.05	9.25 (0.31)	5.04 (0.07)	42.15 (1.52)	8.21 (0.27)
FSR-Corr	0.2	9.85 (0.31)	5.04 (0.07)	51.68 (2.11)	9.91 (0.38)
FSR-Corr	0.5	10.77 (0.37)	5.04 (0.07)	84.36 (6.02)	16.17 (1.18)
FSR-TG	0.01	34.25 (1.39)	8.00 (0.08)	0.00 (0.00)	11.31 (0.09)
FSR-TG	0.05	31.14 (1.47)	7.71 (0.10)	0.00 (0.00)	11.32 (0.09)
FSR-TG	0.2	23.83 (1.34)	7.04 (0.12)	0.00 (0.00)	11.45 (0.09)
FSR-TG	0.5	15.39 (1.23)	5.79 (0.17)	0.00 (0.00)	11.40 (0.07)

competitive compared with CV in terms of MSE, and still enjoy computational efficiency as in other examples.

The results for Example 6 are displayed in Table 2.13. This setting is more difficult for our approach due to the nature of the proposed conditioning test. However, our method still outperforms all the other competitors in terms of prediction accuracy.

2.5.1 A microarray data study

We use a cardiomyopathy microarray dataset to demonstrate the performance of our method for high-dimensional problems. These data were previously analyzed in Segal et al. (2003); Hall

Table 2.10: Results for simulated Example 4 with $\sigma = 3$. The format of the table is the same as Table 2.2.

Methods	γ	MSE	FN	FP	Time
LARS-CV		**9.63 (0.09)**	0.02 (0.01)	0.63 (0.20)	24.72 (0.18)
LARS-Perm	0.01	10.76 (0.09)	2.03 (0.10)	0.00 (0.00)	12.02 (0.14)
LARS-Perm	0.05	10.49 (0.11)	1.64 (0.12)	0.03 (0.02)	12.57 (0.18)
LARS-Perm	0.2	10.27 (0.10)	1.02 (0.11)	0.32 (0.08)	13.90 (0.25)
LARS-Perm	0.5	10.32 (0.10)	0.38 (0.07)	1.79 (0.36)	16.77 (0.52)
LARS-Corr	0.01	10.72 (0.10)	1.97 (0.10)	0.00 (0.00)	**1.38 (0.02)**
LARS-Corr	0.05	10.46 (0.11)	1.58 (0.12)	0.03 (0.02)	1.46 (0.02)
LARS-Corr	0.2	10.24 (0.11)	1.08 (0.11)	0.21 (0.07)	1.57 (0.03)
LARS-Corr	0.5	10.14 (0.10)	0.49 (0.08)	0.97 (0.19)	1.79 (0.04)
LARS-TG	0.01	58.64 (0.67)	8.95 (0.02)	0.00 (0.00)	11.71 (0.10)
LARS-TG	0.05	56.67 (0.86)	8.85 (0.04)	0.00 (0.00)	11.77 (0.10)
LARS-TG	0.2	50.38 (1.24)	8.46 (0.07)	0.00 (0.00)	11.59 (0.09)
LARS-TG	0.5	41.47 (1.59)	7.55 (0.15)	0.00 (0.00)	12.00 (0.14)
LASSO-CV		12.26 (0.11)	0.02 (0.01)	12.44 (0.44)	50.47 (0.41)
LASSO-Perm	0.01	10.75 (0.10)	1.87 (0.11)	0.00 (0.00)	12.26 (0.16)
LASSO-Perm	0.05	10.43 (0.11)	1.38 (0.12)	0.06 (0.03)	13.00 (0.20)
LASSO-Perm	0.2	10.31 (0.10)	0.96 (0.11)	0.36 (0.09)	14.09 (0.24)
LASSO-Perm	0.5	10.50 (0.12)	0.33 (0.07)	2.47 (0.50)	17.86 (0.72)
LASSO-Corr	0.01	10.66 (0.11)	1.74 (0.11)	0.00 (0.00)	**1.41 (0.02)**
LASSO-Corr	0.05	10.40 (0.11)	1.37 (0.12)	0.05 (0.03)	1.48 (0.02)
LASSO-Corr	0.2	10.25 (0.10)	1.01 (0.11)	0.21 (0.08)	1.57 (0.03)
LASSO-Corr	0.5	**10.13 (0.10)**	0.46 (0.08)	0.91 (0.20)	1.81 (0.04)
FSR-CV		**16.59 (0.29)**	6.45 (0.05)	0.29 (0.07)	8.68 (0.05)
FSR-Perm	0.01	21.22 (0.88)	6.00 (0.07)	11.76 (1.35)	22.04 (1.70)
FSR-Perm	0.05	23.63 (0.99)	6.00 (0.07)	16.93 (1.56)	28.64 (1.96)
FSR-Perm	0.2	26.29 (1.06)	6.00 (0.07)	25.93 (2.87)	40.45 (3.78)
FSR-Perm	0.5	29.00 (1.11)	6.00 (0.07)	39.31 (4.54)	57.63 (5.89)
FSR-Corr	0.01	21.88 (0.91)	6.00 (0.07)	13.18 (1.40)	**2.81 (0.23)**
FSR-Corr	0.05	23.93 (1.00)	6.00 (0.07)	17.80 (1.58)	3.52 (0.25)
FSR-Corr	0.2	26.68 (1.08)	6.00 (0.07)	27.05 (2.88)	5.08 (0.49)
FSR-Corr	0.5	29.14 (1.11)	6.00 (0.07)	39.64 (4.52)	7.27 (0.80)
FSR-TG	0.01	58.15 (0.86)	8.91 (0.03)	0.00 (0.00)	11.16 (0.08)
FSR-TG	0.05	54.86 (1.11)	8.78 (0.04)	0.00 (0.00)	11.27 (0.09)
FSR-TG	0.2	47.96 (1.38)	8.40 (0.07)	0.00 (0.00)	11.09 (0.09)
FSR-TG	0.5	37.69 (1.57)	7.54 (0.13)	0.00 (0.00)	11.38 (0.12)

and Miller (2012); Li et al. (2012). The aim of this study is to determine the most influential genes for a G protein-coupled receptor (Ro1) in mice. The dataset contains gene expression levels of 6320 genes on 30 specimens, in which the response variable is the expression level of Ro1 and the covariates X_j are the expression levels of the remaining $p = 6319$ genes.

As in simulation studies, we perform all the methods, i.e., LARS-Corr, LASSO-Corr, FSR-Corr, LARS-Perm, LASSO-Perm, FSR-Perm, LARS-TG, FSR-TG, LARS-CV, LASSO-CV and FSR-CV on the dataset. For CV-based methods, we use 5-fold CV to implement model selection. As the average of pairwise correlations among covariates is close to 0 (less than 0.003), we use the null distribution for independent covariates in our test-based approaches.

Table 2.11: Results for simulated Example 5 with $\sigma = 2$. The format of the table is the same as Table 2.2.

Methods	γ	MSE	FN	FP	Time
LARS-CV		4.32 (0.05)	0.00 (0.00)	2.09 (0.25)	25.52 (0.35)
LARS-Perm	0.01	**4.08 (0.03)**	0.00 (0.00)	0.24 (0.05)	5.89 (0.07)
LARS-Perm	0.05	4.11 (0.03)	0.00 (0.00)	0.36 (0.07)	6.04 (0.10)
LARS-Perm	0.2	4.17 (0.04)	0.00 (0.00)	0.64 (0.10)	6.43 (0.14)
LARS-Perm	0.5	4.38 (0.04)	0.00 (0.00)	2.08 (0.22)	8.42 (0.30)
LARS-Corr	0.01	**4.08 (0.03)**	0.00 (0.00)	0.23 (0.04)	**0.64 (0.01)**
LARS-Corr	0.05	4.11 (0.03)	0.00 (0.00)	0.34 (0.07)	0.66 (0.01)
LARS-Corr	0.2	4.16 (0.04)	0.00 (0.00)	0.57 (0.09)	0.69 (0.02)
LARS-Corr	0.5	4.33 (0.04)	0.00 (0.00)	1.63 (0.16)	0.89 (0.03)
LARS-TG	0.01	11.43 (0.37)	1.96 (0.02)	0.00 (0.00)	11.35 (0.08)
LARS-TG	0.05	11.20 (0.39)	1.89 (0.03)	0.00 (0.00)	11.33 (0.08)
LARS-TG	0.2	9.84 (0.37)	1.56 (0.06)	0.01 (0.01)	11.48 (0.09)
LARS-TG	0.5	7.92 (0.29)	1.04 (0.08)	0.24 (0.06)	11.54 (0.08)
LASSO-CV		4.74 (0.05)	0.00 (0.00)	5.14 (0.34)	39.63 (0.56)
LASSO-Perm	0.01	**4.11 (0.03)**	0.00 (0.00)	0.20 (0.04)	5.85 (0.07)
LASSO-Perm	0.05	4.14 (0.03)	0.00 (0.00)	0.33 (0.07)	6.01 (0.11)
LASSO-Perm	0.2	4.24 (0.04)	0.00 (0.00)	0.84 (0.15)	6.76 (0.25)
LASSO-Perm	0.5	4.46 (0.05)	0.00 (0.00)	2.41 (0.26)	8.98 (0.39)
LASSO-Corr	0.01	**4.11 (0.03)**	0.00 (0.00)	0.20 (0.04)	**0.66 (0.01)**
LASSO-Corr	0.05	4.14 (0.03)	0.00 (0.00)	0.33 (0.07)	0.69 (0.01)
LASSO-Corr	0.2	4.21 (0.04)	0.00 (0.00)	0.67 (0.12)	0.74 (0.02)
LASSO-Corr	0.5	4.41 (0.05)	0.00 (0.00)	1.98 (0.23)	0.98 (0.04)
FSR-CV		4.45 (0.08)	0.47 (0.08)	0.60 (0.11)	9.89 (0.15)
FSR-Perm	0.01	4.37 (0.07)	0.37 (0.07)	0.81 (0.15)	6.08 (0.16)
FSR-Perm	0.05	4.41 (0.07)	0.37 (0.07)	1.06 (0.19)	6.38 (0.21)
FSR-Perm	0.2	4.53 (0.08)	0.37 (0.07)	1.49 (0.23)	6.95 (0.24)
FSR-Perm	0.5	4.78 (0.10)	0.37 (0.07)	2.43 (0.29)	8.17 (0.33)
FSR-Corr	0.01	**4.35 (0.07)**	0.38 (0.07)	0.70 (0.13)	**0.64 (0.01)**
FSR-Corr	0.05	4.40 (0.07)	0.37 (0.07)	1.02 (0.19)	0.70 (0.03)
FSR-Corr	0.2	4.51 (0.08)	0.37 (0.07)	1.41 (0.22)	0.76 (0.03)
FSR-Corr	0.5	4.74 (0.09)	0.37 (0.07)	2.28 (0.28)	0.90 (0.04)
FSR-TG	0.01	11.54 (0.38)	2.00 (0.00)	0.00 (0.00)	10.91 (0.09)
FSR-TG	0.05	11.11 (0.40)	1.88 (0.04)	0.00 (0.00)	10.87 (0.08)
FSR-TG	0.2	9.95 (0.40)	1.57 (0.06)	0.01 (0.01)	10.94 (0.08)
FSR-TG	0.5	8.25 (0.38)	1.06 (0.08)	0.21 (0.07)	10.96 (0.08)

Since the correlation structure of the covariates in the gene expression data is different from iid Gaussian random variables, we also implement the permutation tests incorporated into LARS, LASSO and FSR correspondingly. In addition, we consider $\gamma \in \{0.05, 0.1, 0.2\}$ for LARS-Corr, LASSO-Corr, FSR-Corr, LARS-Perm, LASSO-Perm, FSR-Perm, LARS-TG and FSR-TG. In the experiment, 100 replications are conducted. For each replication, we randomly select 20 samples as the training data, and the remaining 10 as test data to obtain out-of-sample MSE.

We report the average of MSE and computational time with standard errors in Table 2.14. One can see that our test-based methods using theoretical distribution have better prediction accuracy than CV-based ones. While permutation test has competitive performance for MSE,

Table 2.12: Results for simulated Example 5 with $\sigma = 6$. The format of the table is the same as Table 2.2 in the thesis.

Methods	γ	MSE	FN	FP	Time
LARS-CV		**39.67 (0.39)**	0.83 (0.06)	1.14 (0.24)	23.99 (0.18)
LARS-Perm	0.01	40.57 (0.33)	1.59 (0.05)	0.03 (0.02)	3.39 (0.08)
LARS-Perm	0.05	40.01 (0.35)	1.32 (0.06)	0.19 (0.05)	4.03 (0.12)
LARS-Perm	0.2	39.97 (0.35)	1.10 (0.06)	0.58 (0.09)	4.88 (0.18)
LARS-Perm	0.5	40.71 (0.38)	0.74 (0.06)	2.05 (0.29)	7.56 (0.44)
LARS-Corr	0.01	40.77 (0.37)	1.78 (0.07)	0.15 (0.04)	**0.31 (0.01)**
LARS-Corr	0.05	40.16 (0.34)	1.39 (0.06)	0.15 (0.04)	0.37 (0.01)
LARS-Corr	0.2	39.90 (0.34)	1.10 (0.06)	0.55 (0.09)	0.49 (0.02)
LARS-Corr	0.5	40.24 (0.37)	0.76 (0.06)	1.44 (0.16)	0.70 (0.03)
LARS-TG	0.01	42.45 (0.53)	2.01 (0.02)	0.01 (0.01)	11.82 (0.08)
LARS-TG	0.05	42.40 (0.53)	1.99 (0.03)	0.02 (0.01)	11.68 (0.07)
LARS-TG	0.2	41.78 (0.52)	1.81 (0.05)	0.03 (0.02)	11.70 (0.09)
LARS-TG	0.5	41.10 (0.52)	1.44 (0.07)	0.52 (0.12)	11.85 (0.10)
LASSO-CV		40.38 (0.38)	0.76 (0.06)	1.87 (0.30)	38.13 (0.64)
LASSO-Perm	0.01	41.00 (0.33)	1.66 (0.05)	0.03 (0.02)	3.31 (0.08)
LASSO-Perm	0.05	40.49 (0.35)	1.40 (0.07)	0.20 (0.05)	3.94 (0.14)
LASSO-Perm	0.2	40.51 (0.35)	1.09 (0.07)	0.76 (0.14)	5.18 (0.26)
LASSO-Perm	0.5	41.35 (0.42)	0.75 (0.06)	2.32 (0.32)	7.97 (0.51)
LASSO-Corr	0.01	40.55 (0.37)	1.77 (0.07)	0.19 (0.04)	**0.32 (0.01)**
LASSO-Corr	0.05	40.44 (0.35)	1.42 (0.06)	0.17 (0.05)	0.37 (0.01)
LASSO-Corr	0.2	**40.28 (0.35)**	1.11 (0.07)	0.61 (0.11)	0.51 (0.02)
LASSO-Corr	0.5	40.86 (0.40)	0.77 (0.06)	1.67 (0.21)	0.77 (0.04)
FSR-CV		40.89 (0.32)	1.85 (0.04)	0.17 (0.04)	9.37 (0.05)
FSR-Perm	0.01	40.57 (0.31)	1.77 (0.05)	0.20 (0.04)	3.33 (0.08)
FSR-Perm	0.05	40.53 (0.33)	1.70 (0.06)	0.44 (0.06)	3.73 (0.10)
FSR-Perm	0.2	41.21 (0.41)	1.64 (0.06)	0.88 (0.09)	4.39 (0.12)
FSR-Perm	0.5	42.91 (0.51)	1.61 (0.06)	1.71 (0.13)	5.53 (0.16)
FSR-Corr	0.01	41.01 (0.35)	1.88 (0.07)	0.13 (0.04)	**0.28 (0.01)**
FSR-Corr	0.05	**40.51 (0.33)**	1.70 (0.06)	0.42 (0.06)	0.36 (0.01)
FSR-Corr	0.2	41.10 (0.40)	1.63 (0.06)	0.86 (0.09)	0.45 (0.01)
FSR-Corr	0.5	42.60 (0.49)	1.61 (0.06)	1.58 (0.12)	0.57 (0.02)
FSR-TG	0.01	43.03 (0.65)	2.06 (0.03)	0.00 (0.00)	11.36 (0.08)
FSR-TG	0.05	43.01 (0.65)	2.05 (0.03)	0.00 (0.00)	11.28 (0.08)
FSR-TG	0.2	42.22 (0.56)	1.89 (0.04)	0.11 (0.04)	11.21 (0.08)
FSR-TG	0.5	41.76 (0.53)	1.64 (0.06)	0.58 (0.10)	11.30 (0.09)

it has the most expensive computational cost among all methods. On the contrary, compared with CV as well as permutation test, the computational expenses of our test-based approaches are reduced for all three sequential selection procedures.

To better demonstrate the performance of our test-based approach, we show a stepwise plot and an overall MSE plot for LARS-Corr as in Figure 2.4. Figure 2.4a illustrates the stepwise p-value and MSE for the first 15 steps of LARS-Corr. Here the out-of-sample MSE at Step k is with respect to the model containing variables selected by the first k LARS steps. Note that such models might vary through 100 replications, resulting in relatively large standard errors for MSE. By the one standard error rule, Figure 2.4a implies that a candidate model of size

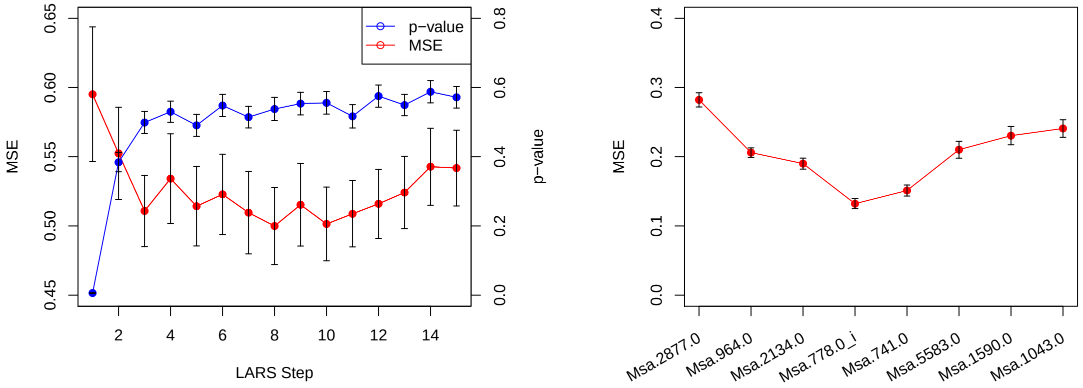
Table 2.13: Results for simulated Example 6. The format of the table is the same as Table 2.2 in the thesis.

Methods	γ	MSE	FN	FP	Time
LARS-CV		**9.63 (0.09)**	0.02 (0.01)	0.63 (0.20)	24.72 (0.18)
LARS-Perm	0.01	9.69 (0.11)	0.15 (0.05)	0.00 (0.00)	15.45 (0.11)
LARS-Perm	0.05	**9.62 (0.10)**	0.07 (0.03)	0.10 (0.04)	15.62 (0.11)
LARS-Perm	0.2	9.85 (0.11)	0.05 (0.03)	0.65 (0.13)	16.24 (0.20)
LARS-Perm	0.5	10.70 (0.15)	0.01 (0.01)	5.71 (1.95)	23.35 (2.73)
LARS-Corr	0.01	9.86 (0.12)	0.24 (0.06)	0.00 (0.00)	**1.78 (0.01)**
LARS-Corr	0.05	9.63 (0.10)	0.10 (0.04)	0.03 (0.02)	1.81 (0.01)
LARS-Corr	0.2	9.72 (0.10)	0.07 (0.03)	0.27 (0.06)	1.86 (0.02)
LARS-Corr	0.5	10.19 (0.12)	0.01 (0.01)	1.65 (0.22)	2.11 (0.04)
LARS-TG	0.01	95.51 (2.09)	8.88 (0.05)	0.00 (0.00)	11.42 (0.06)
LARS-TG	0.05	89.40 (2.49)	8.71 (0.06)	0.00 (0.00)	11.70 (0.12)
LARS-TG	0.2	81.83 (2.74)	8.42 (0.09)	0.00 (0.00)	11.54 (0.09)
LARS-TG	0.5	61.43 (3.27)	7.31 (0.17)	0.00 (0.00)	11.94 (0.14)
LASSO-CV		12.26 (0.11)	0.02 (0.01)	12.44 (0.44)	50.47 (0.41)
LASSO-Perm	0.01	10.75 (0.10)	1.87 (0.11)	0.00 (0.00)	12.26 (0.16)
LASSO-Perm	0.05	10.43 (0.11)	1.38 (0.12)	0.06 (0.03)	13.00 (0.20)
LASSO-Perm	0.2	10.31 (0.10)	0.96 (0.11)	0.36 (0.09)	14.09 (0.24)
LASSO-Perm	0.5	10.50 (0.12)	0.33 (0.07)	2.47 (0.50)	17.86 (0.72)
LASSO-Corr	0.01	10.66 (0.11)	1.74 (0.11)	0.00 (0.00)	**1.41 (0.02)**
LASSO-Corr	0.05	10.40 (0.11)	1.37 (0.12)	0.05 (0.03)	1.48 (0.02)
LASSO-Corr	0.2	10.25 (0.10)	1.01 (0.11)	0.21 (0.08)	1.57 (0.03)
LASSO-Corr	0.5	**10.13 (0.10)**	0.46 (0.08)	0.91 (0.20)	1.81 (0.04)
FSR-CV		15.96 (0.37)	2.64 (0.11)	0.39 (0.07)	9.71 (0.11)
FSR-Perm	0.01	**13.92 (0.36)**	1.43 (0.11)	2.00 (0.22)	16.25 (0.32)
FSR-Perm	0.05	14.54 (0.43)	1.28 (0.11)	3.55 (0.34)	18.46 (0.45)
FSR-Perm	0.2	15.51 (0.52)	1.20 (0.10)	5.51 (0.50)	21.08 (0.63)
FSR-Perm	0.5	16.60 (0.57)	1.18 (0.10)	7.71 (0.60)	24.35 (0.75)
FSR-Corr	0.01	13.97 (0.36)	1.53 (0.11)	1.66 (0.20)	**1.77 (0.04)**
FSR-Corr	0.05	14.30 (0.42)	1.31 (0.11)	2.96 (0.29)	2.04 (0.05)
FSR-Corr	0.2	15.10 (0.47)	1.25 (0.11)	4.60 (0.41)	2.32 (0.06)
FSR-Corr	0.5	16.11 (0.54)	1.20 (0.10)	6.52 (0.51)	2.64 (0.08)
FSR-TG	0.01	91.36 (2.48)	8.81 (0.05)	0.00 (0.00)	11.13 (0.06)
FSR-TG	0.05	83.52 (2.86)	8.63 (0.06)	0.00 (0.00)	11.23 (0.12)
FSR-TG	0.2	69.66 (2.90)	8.20 (0.09)	0.00 (0.00)	11.13 (0.08)
FSR-TG	0.5	52.66 (2.85)	7.21 (0.16)	0.00 (0.00)	11.41 (0.13)

3 would be preferable. Moreover, we also summarize the most frequently identified genes out of 100 replications and sort by frequency from high to low. Figure 2.4b shows the eight most frequently identified genes that are selected at least 10 times over 100 replications, as well as the out-of-sample MSE corresponding to the model containing the first k genes with $k \in \{1, \ldots, 8\}$. Among the eight genes, Msa.2877.0 was also identified in Hall and Miller (2012); Li et al. (2012), and Msa.2134.0 was discovered in Li et al. (2012). Overall, our variable selection method is effective in identifying potential scientific discoveries.

Table 2.14: The average MSE and computational time over 100 replications (with standard errors given in parentheses).

Methods	γ	MSE	Time	Methods	γ	MSE	Time
LARS-CV		0.63 (0.04)	1.48 (0.02)	FSR-CV		0.91 (0.16)	0.78 (0.04)
LARS-Perm	0.05	0.60 (0.05)	1.81 (0.19)	FSR-Perm	0.05	0.62 (0.05)	1.44 (0.04)
LARS-Perm	0.1	0.59 (0.05)	2.62 (0.35)	FSR-Perm	0.1	0.63 (0.05)	1.65 (0.06)
LARS-Perm	0.2	0.59 (0.04)	5.78 (0.62)	FSR-Perm	0.2	0.67 (0.05)	2.22 (0.20)
LARS-Corr	0.05	0.58 (0.05)	**0.44 (0.01)**	FSR-Corr	0.05	0.61 (0.05)	**0.41 (0.01)**
LARS-Corr	0.1	0.55 (0.05)	0.53 (0.02)	FSR-Corr	0.1	**0.60 (0.05)**	0.48 (0.02)
LARS-Corr	0.2	**0.53 (0.04)**	0.58 (0.03)	FSR-Corr	0.2	**0.60 (0.05)**	0.51 (0.02)
LARS-TG	0.05	0.74 (0.05)	3.42 (0.03)	FSR-TG	0.05	0.72 (0.05)	2.94 (0.02)
LARS-TG	0.1	0.72 (0.05)	3.52 (0.03)	FSR-TG	0.1	0.71 (0.05)	3.01 (0.02)
LARS-TG	0.2	0.66 (0.05)	3.56 (0.03)	FSR-TG	0.2	0.65 (0.05)	3.04 (0.02)
LASSO-CV		0.59 (0.04)	1.98 (0.02)				
LASSO-Perm	0.05	0.60 (0.05)	1.47 (0.05)				
LASSO-Perm	0.1	0.57 (0.05)	3.21 (0.60)				
LASSO-Perm	0.2	0.54 (0.04)	7.84 (0.99)				
LASSO-Corr	0.05	0.58 (0.05)	**0.41 (0.01)**				
LASSO-Corr	0.1	0.55 (0.05)	0.49 (0.02)				
LASSO-Corr	0.2	**0.53 (0.04)**	0.55 (0.03)				



(a) Stepwise p-value and MSE

(b) Most frequently identified genes and MSE

Figure 2.4: Performance of LARS-Corr applied to the microarray data. (a) Average p-value and MSE with one standard error bands for the first 15 steps of LARS-Corr over 100 replications. (b) 8 most frequently identified genes by LARS-Corr and the out-of-sample MSE corresponding to the model consists of the first $k \in \{1, \dots, 8\}$ genes.

2.6 Discussion

In this thesis, we propose a test-based variable selection approach in the context of high-dimensional linear regression model with Gaussian covariates. We first formulate the null hypothesis, where we assume that the response is uncorrelated with all of the remaining covariates given a set of selected variables. We also propose the maximal absolute sample partial correlation statistic and discuss its asymptotic null distribution and power. We then incorporate the distribution information with sequential selection procedures. We use three simulated examples and one real data analysis to demonstrate that compared with CV-based procedure, the proposed method can perform variable selection effectively and efficiently.

Our proposed method involves sequential hypothesis testing. Therefore, instead of using a constant test level γ, one can consider multiple testing methods, such as the false discovery rate (FDR) control (Benjamini and Hochberg, 1995), which provides flexible test levels and meaningful probability statements of the selected model. However, due to the adaptive nature of the sequential selection procedures, classical FDR control methods cannot be applied directly. There are some recent papers for sequential testing (Foster and Stine, 2008; Aharoni and Rosset, 2014; G'Sell et al., 2016). However, the approaches in Aharoni and Rosset (2014); Foster and Stine (2008) are known to control the marginal FDR instead of the FDR. In contrast, G'Sell et al. (2016) assumes that the p-values corresponding to the null hypotheses are iid $\mathcal{U}(0, 1)$, which does not usually hold in our setting. We plan to investigate our procedure along this direction in future work.

2.7 Proofs

In this section, we provide proofs to Theorems 1 and 2.

Proof of Theorem 1. Without loss of generality, we assume that X_j's are standard normal variables. Let $\boldsymbol{\beta}^*$ be a vector consisting of all non-zero components of $\boldsymbol{\beta}$. Recall that $\mathcal{M}^*$ is the support set of $\boldsymbol{\beta}$. Thus we can rewrite the regression model in (2.1) as

$$\mathbf{y} = X_{\mathcal{M}^*}\beta^* + \boldsymbol{\varepsilon}.$$

Recall that under the null hypothesis $\mathcal{M}^* \subseteq \mathcal{M}$, $\mathbf{r} = (I - P_{\mathcal{M}})\mathbf{y} = (I - P_{\mathcal{M}})(X_{\mathcal{M}^*}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}) = (I - P_{\mathcal{M}})\boldsymbol{\varepsilon}$. Therefore, for any $j \notin \mathcal{M}$, the sample partial correlation between $\mathbf{r}_j$ and $\mathbf{r}$ given the active set $\mathcal{M}$ can be written as

$$\widehat{\mathrm{Corr}}(\mathbf{r}_j, \mathbf{r}) = \frac{\langle \mathbf{r}_j, \mathbf{r} \rangle}{\|\mathbf{r}_j\| \|\mathbf{r}\|} = \langle \frac{(I - P_{\mathcal{M}})\mathbf{x}_j}{\|(I - P_{\mathcal{M}})\mathbf{x}_j\|}, \frac{(I - P_{\mathcal{M}})\boldsymbol{\varepsilon}}{\|(I - P_{\mathcal{M}})\boldsymbol{\varepsilon}\|} \rangle.$$

Denote $\mathbf{L}_j = \frac{(I-P_{\mathcal{M}})\mathbf{x}_j}{\|(I-P_{\mathcal{M}})\mathbf{x}_j\|}$, $\mathbf{V} = \frac{(I-P_{\mathcal{M}})\boldsymbol{\varepsilon}}{\|(I-P_{\mathcal{M}})\boldsymbol{\varepsilon}\|}$. According to Theorem II.4 in Zhang (2017), to obtain the uniform asymptotic null distribution of $\max_{j:j\in\mathcal{M}^c} |\widehat{\mathrm{Corr}}(\mathbf{r}_j, \mathbf{r})|$, it is enough to show the following two arguments:

(1) Conditioning on $\mathcal{M}$, $\{\mathbf{L}_j; j \in \mathcal{M}^c\}$ are independent, and they have a degenerate uniform distribution on the $(n - s - 2)$-sphere $\mathcal{S}^{(n-s-2)}$ in n-dimensional space.

(2) Conditioning on $\mathcal{M}$, $\mathbf{V}$ is a unit vector independent of $\mathbf{L}_j$ for $\forall j \in \mathcal{M}^c$.

For part (1), as $\mathcal{M}$ is given, $I - P_{\mathcal{M}}$ is deterministic. Since $I - P_{\mathcal{M}}$ is a orthogonal projection and $\mathrm{rank}(I - P_{\mathcal{M}}) = n - s - 1$, there exists an $n \times (n - s - 1)$ matrix U_1 with orthogonal column vectors of unit length such that $I - P_{\mathcal{M}} = U_1 U_1^\top$. In fact, we can take orthonormal basis for the column space of $X_{\mathcal{M}}$ as the column vectors of U_1. Thus we have

$$\mathbf{r}_j = (I - P_{\mathcal{M}})\mathbf{x}_j = U_1 U_1^\top \mathbf{x}_j.$$

Denote $\mathbf{z} = U_1^\top \mathbf{x}_j$. By the Gaussian assumption of the covariates, we have $\mathbf{x}_j \overset{\mathrm{d}}{\sim} \mathcal{N}(\mathbf{0}, I_n)$. Hence $\mathbf{z} \overset{\mathrm{d}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I_{n-s-1}})$. By Lemma A.4.2 of Anderson (2003), it is possible to find an $n \times (s + 1)$ matrix U_2 with $U = \begin{pmatrix} U_1, & U_2 \end{pmatrix}$ being orthogonal, such that $\mathbf{r}_j$ can be rewritten as $\mathbf{r}_j = U\tilde{\mathbf{z}}$. Here $\tilde{\mathbf{z}} = \begin{pmatrix} \mathbf{z} \\ \mathbf{0} \end{pmatrix}$ is an n-vector.

Note that $\frac{\tilde{\mathbf{z}}}{\|\tilde{\mathbf{z}}\|} = (\frac{\mathbf{z}^\top}{\|\mathbf{z}\|}, \mathbf{0})^\top$ is a degenerate uniform distribution on $\mathcal{S}^{(n-s-2)}$ in n-dimensional space. Since U is orthogonal, by the definition of uniform spherical distribution, $L_j = \frac{\mathbf{r}_j}{\|\mathbf{r}_j\|} = \frac{U\tilde{\mathbf{z}}}{\|\tilde{\mathbf{z}}\|} \overset{\mathrm{d}}{\sim} \frac{\tilde{\mathbf{z}}}{\|\tilde{\mathbf{z}}\|}$, thus is also uniformly distributed on $\mathcal{S}^{n-s-2}$. Moreover, $\mathbf{r}_j$'s are independent conditioning on $\mathcal{M}$, which implies that L_j's are also conditionally independent given $\mathcal{M}$.

For part (2), by the model assumption, $\mathbf{x}_j$ is independent of $\boldsymbol{\varepsilon}$ for $\forall j \in \mathcal{M}^c$. Hence given $X_{\mathcal{M}}$, $\{(I - P_{\mathcal{M}})\mathbf{x}_j; j \in \mathcal{M}^c\}$ are also conditionally independent of $(I - P_{\mathcal{M}})\varepsilon$, implying that L_j is conditionally independent of V for all $j \in \mathcal{M}^c$.

Thus far we have justified the two aforementioned arguments. Therefore the proposed test statistic has the asymptotic distribution uniformly for any $n \geq s+3$ as in (1). □

Proof of Theorem 2. Note that $x_\alpha(p,n,s) = F_{n,s}^{-1}(1-\alpha)b(p,n,s) + a(p,n,s)$, with $F_{n,s}(\,\cdot\,)$ being the asymptotic cumulative distribution function of $R_{\mathcal{M}}$ under $H_0^{\mathcal{M}}$ as in (2.6). Before proving the theorem, we first prove the following two lemmas.

Lemma 1. *For $a(p,n,s), b(p,n,s)$ and $c(p,n,s)$, we have the following asymptotic results:*

1. *Uniformly for all $p \geq 2$, $c(p,n,s) \to 1$ as $n \to \infty$.*
2. *As $\frac{\log p}{n} \to 0$ and $n \to \infty$, $a(p,n,s) \to 0, b(p,n,s) \to 0$.*

Proof of Lemma 1. Note that $\sqrt{\frac{2\pi}{n-s-2}} \leq \mathcal{B}(\frac{1}{2}, \frac{n-s-2}{2}) \leq \sqrt{\frac{4\pi}{n-s-2}}$, hence

$$\begin{aligned} c(p,n,s) &\leq (\frac{n-s-2}{2} B(\frac{1}{2}, \frac{n-s-2}{2}))^{2/(n-s-2)} \\ &\leq (\frac{n-s-2}{2}\sqrt{\frac{4\pi}{n-s-2}})^{2/(n-s-2)} \\ &= ((n-s-2)\pi)^{1/(n-s-2)} \\ &\to 1 \text{ as } n \to \infty. \end{aligned}$$

Furthermore, for any $p \geq 2$, $(\frac{n-s-2}{2})(1-p^{-2/(n-s-2)}) \geq 1-1/p$, hence we have

$$\begin{aligned} c(p,n,s) &\geq \left(\frac{n-s-2}{2}\sqrt{\frac{2\pi}{n-s-2}(1-p^{-2/(n-s-2)})}\right)^{2/(n-s-2)} \\ &\geq (\pi(1-1/p))^{1/(n-s-2)} \\ &\geq (\frac{\pi}{2})^{1/(n-s-2)} \\ &\to 1 \text{ as } n \to \infty. \end{aligned}$$

Note that the above two convergence results do not depend on p. Therefore, uniformly for any $p \geq 2$, $c(p,n,s) \to 1$ as $n \to \infty$.

Since $a(p,n,s) = 1-(p-s)^{-2/(n-s-2)}c(p,n,s)$, $b(p,n,s) = \frac{2}{n-s-2}(p-s)^{-2/(n-s-2)}c(p,n,s)$, and $(p-s)^{-2/(n-s-2)} \to 1$ by assumption. It follows directly that as $\frac{\log p}{n} \to 0$ and $n \to \infty$,

$$a(p,n,s) \to 0, \qquad b(p,n,s) \to 0.$$

□

Lemma 2. *Under the alternative hypothesis* (2.9)*, there exists a constant* $c_0 > 0$ *such that uniformly for any* $p \geq s$*,* $\Pr(R_{\mathcal{M}} \geq c_0 | H_1) \to 1$ *as* $n \to \infty$*.*

Proof of Lemma 2. Under (2.9), we can find some $j_0 \in \mathcal{M}^c$ such that $\rho_0 = |\mathrm{Corr}(\mathrm{X_{j_0}}, \mathrm{Y}|\mathrm{X}_{\mathcal{M}})| > 0$, which is the absolute population partial correlation between X_{j_0} and Y given $X_{\mathcal{M}}$. We have that $r_{n,j_0} = |\widehat{\mathrm{Corr}}(X_{j_0}, Y|X_{\mathcal{M}})| \xrightarrow{p} \rho_0$ as $n \to \infty$.

Let $c_0 = \rho_0/2$, then

$$\begin{aligned}\Pr(R_{\mathcal{M}} \geq c_0|H_1) &\geq \Pr(r_{n,j_0} \geq c_0|H_1) \\ &\geq \Pr(|r_{n,j_0} - \rho_0| \leq c_0|H_1) \\ &\to 1 \text{ as } n \to \infty.\end{aligned}$$

As the above convergence does not depend on p, it follows that uniformly for any $p \geq s$, $\Pr(R_{\mathcal{M}} \geq c_0|H_1) \to 1$ as $n \to \infty$. □

Note that $\forall x > 0$, $F_{n,s}^{-1}(x) \to -\log(-\log(x))$ as $n \to \infty$. Thus we have $-1 - \log(-\log(1 - \alpha)) \leq F_{n,s}^{-1}(1-\alpha) \leq 1 - \log(-\log(1-\alpha))$ for large enough n. Together with the results of Lemma 1, we have for any fixed s and α, $x_\alpha(p,n,s) = F_{n,s}^{-1}(1-\alpha)b(p,n,s) + a(p,n,s) \to 0$ as $(\log p)/n \to 0$ and $n \to \infty$.

Therefore, for $\forall 0 < \epsilon < c_0$ with c_0 being the same as in Lemma 6,

$$\begin{aligned}\Pr(R_{\mathcal{M}} \geq x_\alpha(p,n,s)|H_1) =& 1 - \Pr(R_{\mathcal{M}} < x_\alpha(p,n,s)|H_1) \\ =& 1 - \Pr(R_{\mathcal{M}} < x_\alpha(p,n,s), x_\alpha(p,n,s) < \epsilon|H_1) \\ & - \Pr(R_{\mathcal{M}} < x_\alpha(p,n,s), x_\alpha(p,n,s) \geq \epsilon|H_1) \\ \geq& 1 - \Pr(R_{\mathcal{M}} < x_\alpha(p,n,s), x_\alpha(p,n,s) < \epsilon|H_1) \\ & - \Pr(x_\alpha(p,n,s) \geq \epsilon|H_1) \\ \geq& 1 - \Pr(R_{\mathcal{M}} < c_0|H_1) - \Pr(x_\alpha(p,n,s) \geq \epsilon|H_1).\end{aligned}$$

When $(\log p)/n \to 0$ and $n \to \infty$, we have $\Pr(R_{\mathcal{M}} < c_0|H_1) \to 0$, $\Pr(x_\alpha(p,n,s) \geq \epsilon|H_1) \to 0$, thus $\Pr(R_{\mathcal{M}} \geq x_\alpha(p,n,s)|H_1) \to 1$, i.e., the asymptotic power is 1. □

CHAPTER 3

Penalized linear regression with high-dimensional pairwise screening

3.1 Introduction

In the era of big data, high dimensional problems are of interest in many scientific fields, where the number of variables may be comparable to or even much larger than the sample size. For example, in genetic studies, one often has tens of thousands of genes in the microarray datasets with only a few hundreds of patients; in neuroscience, fMRI images may contain millions of voxels.

In recent years, much research effort has been devoted to deal with high dimensional data analysis. Among those methods developed, penalized least squares plays an important role. In particular, one of the most well-known method is the LASSO proposed by Tibshirani (1996), which is the solution to the following penalized problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + \lambda P(\boldsymbol{\beta}), \tag{3.1}$$

where $\lambda P(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\beta_j|$ is the l_1 penalty. Tibshirani (1996) showed that the LASSO leads to a sparse estimator that shrinks the OLS solution and sets some of the estimated coefficients to exact zero. As introduced in Chapter 1, despite with good theoretical properties and practical performance, the LASSO has some drawbacks. To address these issues, Zou and Hastie (2005) introduced the elastic net method, using $\lambda_1\|\boldsymbol{\beta}\|_1+\lambda_2\|\boldsymbol{\beta}\|_2^2$ as the regularization term in (3.1) and thus encouraging a grouping effect. Besides the elastic net, various penalized variable selection methods have been proposed as extensions to LASSO. See Chapter 1 for an overview.

For high dimensional variable selection, it is crucial to account for the dependency structure between covariates. Such structure information not only improves the accuracy of selection, but also have practical meanings. For instance, in gene expression data, genes usually function as biological pathways instead of working independently. Classical penalized variable selection methods, however, usually do not explicitly take into account the relationships between covariates. To address this problem, Yuan and Lin (2006) proposed the group LASSO method, which takes advantage of the grouping information among the covariates. Extensions of group lasso include, but are not limited to the PACS proposed by Breheny and Huang (2015). Other methods use the structure information as predictor graph (see Li and Li (2008); Pan et al. (2010); Zhu et al. (2013); Yu and Liu (2016) among others for reference).

A common assumption for the methods mentioned above is that the underlying predictor graph is given, which may not hold in practice. When the prior information is not available, the idea of clustering can be incorporated to improve regression performance. Specifically, Park et al. (2007) proposed to perform hierarchical clustering on the covariates first and take the cluster average as new predictors for regression. There are also methods using supervised clustering to encourage highly correlated pairs of covariates to be included or excluded simultaneously (Bondell and Reich, 2008; Sharma et al., 2013). Similarly, another type of methods aims to make correlated covariates have similar regression coefficients (She et al., 2010). Nevertheless, a large sample correlation between two variables does not necessarily indicate that they are dependent in the population sense. When the dimensionality continues to increase, the maximal pairwise correlation among p independent covariates can be close to 1 (Fan and Lv, 2010). Therefore, it is important to identify covariates that are truly correlated and incorporate such information into variable selection procedures.

In this chapter, we study the limiting behavior of the maximal absolute pairwise sample correlation among covariates when they are independent Gaussian random variables. Different from existing work, we investigate the limiting distribution as the dimensionality p diverges. Therefore, the proposed asymptotic results potentially can be applied to datasets with arbitrarily large dimensionality. We further discuss the extreme behavior of the maximal absolute Spearman's rho statistic for covariates with general distributions. On the other hand, we obtain the lower bound of maximal pairwise R squares when regressing the response onto pairs of covariates. With the extreme value results, we formulate a screening procedure to identify covariates pairs that are potentially dependent and associated with the response. We further combine the pairwise screening with the Sure Independence Screening (SIS) (Fan and Lv,

2008) and propose a novel penalized variable selection method. More specifically, we assign different penalties to each individual covariate according to the screening results. Numerical experiments show that the performance of our proposed method is competitive compared with existing approaches in terms of both variable selection and prediction accuracy.

The remainder of this chapter is organized as follows: We first investigate the limiting distribution of the maximal pairwise sample correlation among covariates in Section 3.2.1. We also show that our asymptotic results cover that of Cai and Jiang (2012) as a special case. Then we propose an upper bound for the maximal pairwise R squares in Section 3.2.2. In Section 3.3.1 we formulate our proposed variable selection approach as a penalized maximum likelihood problem, and discuss potential extensions of our method in Section 3.3.2. We show with simulated experiments as well as two real datasets in Section 3.5 that the propose method has improved performance when important variables are highly correlated. Finally, we conclude this chapter and discuss possible future work in Section 3.6. Proofs of theoretical results are presented in Section 3.7.

3.2 Pair Screening for covariates

Suppose we have the following linear model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.2}$$

where $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ is the response vector, $X = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p)$ is an $n \times p$ design matrix with $\mathbf{x}_j$ being n independent and identical observations from the covariate X_j. We assume that the covariate vector $\mathbf{x} = (X_1, X_2, \cdots, X_p)^T$ has a multivariate distribution with unknown covariance matrix Σ, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^T$ is a vector of i.i.d. random variables with mean 0 and standard deviation σ, and is independent of the covariate vector $\mathbf{x}$.

For the linear model (3.2), variable selection methods aim to identify the non-zero components of $\boldsymbol{\beta}$, in other words, the important variables among all candidate predictors. Particularly, if two covariates have a large pairwise correlation, we may want to include or exclude these two variables simultaneously when conducting variable selection. However, the sample correlation can be spurious, especially when the number of covariates p is relatively large. Therefore, it is important to identify covariates that are truly correlated. In other words, we need to find a threshold for the pairwise sample correlation among covariates to screen covariates pairs. In the following subsection, we will discuss in details the asymptotic results that generate the screening rule.

3.2.1 Extreme laws of pairwise sample correlation among covariates

We propose to choose a bound based on the extreme laws of the pairwise sample correlation when the p covariates are independent. Our investigations are under two settings: (a) the covariates are normally distributed; (b) the covariates are non-Gaussian random variables.

3.2.1.1 Gaussian covariates

It has been recently studied that the maximal absolute Pearson sample correlation between p i.i.d. Gaussian covariates and an independent response has a Gumble-type limiting distribution as p goes to infinity (Zhang, 2017). Motivated by Zhang (2017)'s work, we find that the maximal absolute pairwise sample correlation among p independent covariates also has a limiting distribution, as stated in the following theorem:

Theorem 3. *Suppose $X_1, X_2, \cdots, X_p$ are p independent Gaussian variables and we observe n independent samples from each of X_j's. Let $W_{pn} = \max_{1\le i<j\le p} |\rho_{i,j}|$, where $\rho_{i,j} = \widehat{\text{Corr}}(X_i, X_j)$ is the Pearson sample correlation between X_i and X_j. Then as $p \to \infty$,*

$$\lim_{p\to\infty} |P(\frac{W_{pn}^2 - a_{p,n}}{b_{p,n}} \le x) - I(x \le \frac{n-2}{2}) \exp\{-\frac{1}{2}(1 - \frac{2}{n-2}x)^{\frac{n-2}{2}}\} - I(x > \frac{n-2}{2})| = 0, \quad (3.3)$$

which is uniform for any $n \ge 3$. Here $a_{p,n} = 1 - p^{-4/(n-2)} c_{p,n}$, $b_{p,n} = \frac{2}{n-2} p^{-4/(n-2)} c_{p,n}$, and $c_{p,n} = \left(\frac{n-2}{2} \mathcal{B}(\frac{1}{2}, \frac{n-2}{2}) \sqrt{1 - p^{-4/(n-2)}}\right)^{2/(n-2)}$ are the normalizing constants.

Remark 3.2.1. The above limiting distribution is similar to 2.6 given in Chapter 2, but differs in the constant within the exponential term. Moreover, the normalizing constant involves $p^{-4/(n-2)}$.

In random matrix theory, W_{pn} is also known as the coherence when the design matrix X is random. Specifically, the coherence is defined as the largest magnitude of the off-diagonal entries of the sample correlation matrix associated with a random matrix. The limiting behavior of the coherence has been well studied when the sample size n goes to infinity. For example, Cai and Jiang (2011) studied the asymptotic distribution under certain regularity conditions with application to the testing of covariance matrix. Cai and Jiang (2012) further obtained the limiting laws of the coherence for different divergence rate of p with respect to n and summarized the results as phase transition phenomena. We can show that our result unifies the convergence in terms of the sample size, and covers Cai and Jiang (2012)'s as special cases, described in the following corollary.

Corollary 3. *Let W_{pn} be defined as in Theorem 3, where we still assume X_j's are independent normal random variables. Let $T_{pn} = \log(1 - W_{pn}^2)$.*

*(a) (**Sub-Exponential Case**) Suppose $p = p_n \to \infty$ as $n \to \infty$ and $(\log p)/n \to 0$, then as $n \to \infty$,*

$$\Pr\left(nT_{pn} + 4\log p - \log\log p \le x\right) \to 1 - e^{-\frac{1}{\sqrt{8\pi}}e^{x/2}}.$$

*(b) (**Exponential Case**) Suppose $p = p_n$ satisfies $(\log p)/n \to \beta \in (0, \infty)$ as $n \to \infty$. Then as $n \to \infty$,*

$$\Pr\left(nT_{pn} + 4\log p - \log\log p \le x\right) \to 1 - \exp\big\{ - K(\beta)e^{(x+8\beta)/2}\big\},$$

where $K(\beta) = \big(\frac{\beta}{2\pi(1-4e^{-4\beta})}\big)^{1/2}$.

*(c) (**Super-Exponential Case**) Suppose $p = p_n$ satisfies $(\log p)/n \to \infty$ as $n \to \infty$. Then as $n \to \infty$,*

$$\Pr\left(nT_{pn} + \frac{4n}{n-2}\log p - \log n \le x\right) \to 1 - e^{-\frac{1}{\sqrt{2\pi}}e^{x/2}}.$$

Compared with previous work, our asymptotic distribution is novel in two aspects. First, the convergence in Theorem 3 is with respect to p instead of n, making it applicable to high dimensional data, or even ultrahigh dimensional problems. Moreover, the convergence result we have discovered is uniform for any $n \ge 3$, thus finite sample performance is guaranteed.

3.2.1.2 Non-Gaussian covariates

When the covariates are non-Gaussian random variables, it is more desirable to choose a distribution-free statistic for the screening rule. Therefore, instead of using the Pearson's sample correlation, we study the extreme behavior of the Spearman's rho statistic (Spearman, 1904). Recall that $\mathbf{x}_j = (X_{1j}, X_{2j}, \cdots, X_{nj})^T$ are n i.i.d. observations from the covariate X_j. Let Q_{ni}^j and Q_{ni}^k be the ranks of X_{ij} and X_{ik} in $\{X_{1j}, \cdots, X_{nj}\}$ and $\{X_{1k}, \cdots, X_{nj}\}$ respectively. Then the Spearman's rho is defined as

$$\rho_{ij} = \frac{\sum_{i=1}^n (Q_{ni}^j - \bar{Q}_n^j)(Q_{ni}^k - \bar{Q}_n^k)}{\sqrt{\sum_{i=1}^n (Q_{ni}^j - \bar{Q}_n^j)^2 \sum_{i=1}^n (Q_{ni}^k - \bar{Q}_n^k)^2}}, \tag{3.4}$$

where $\bar{Q}_n^j = \bar{Q}_n^k = \frac{n+1}{2}$.

Similar to the normal setting, we are particularly interested in the limiting distribution of $S_{pn}^2 = \max_{1 \le i < j \le p} \rho_{ij}^2$ when the covariates are all independent, which has been studied in Han and Liu (2014). The following proposition introduced by Han and Liu (2014) states that as n increases, S_{pn}^2 converges to a Gumble type distribution.

Proposition 2. *[Han and Liu (2014)] Suppose that $X_1, \cdots, X_p$ are independent and identically distributed random variables, and we have n independent samples for each of the covariates. Let $S_{pn}^2 = \max_{1 \le i < j \le p} \rho_{ij}^2$ be the squares of the maximal pairwise Spearman's rho statistics, then for $\log p = o(n^{1/3})$, we have*

$$\lim_{n \to \infty} | \Pr\big((n-1)S_{pn}^2 - 4\log p + \log\log p \le x\big) - \exp\big\{ -(8\pi)^{-1/2}\exp(-y/2)\big\}| = 0. \tag{3.5}$$

Theorem 3 and Proposition 2 characterize the magnitude of the maximal pairwise correlation and Spearman's rho statistic respectively when the covariates are independent. In the following subsection, we further investigate the extreme behavior of the maximal pairwise R squares under the null model, i.e., β_j's are all equal to zero.

3.2.2 R squares screening for pairs of covariates

With the asymptotic distributions introduced in the previous subsections, we can identify covariates pairs that are potentially dependent. However, such screening does not take into account the association between the covariates and the response. It is possible that an important variable has a large sample correlation with unimportant ones; or two highly correlated covariates are both unrelated to the response. To address such an issue, we introduce another screening procedure based on the R squares from regressing the response Y onto the pairs of covariates.

Consider the linear regression where we regress Y onto a pair of covariates X_i and X_j with $i \neq j$, we can obtain the corresponding R squares $R_{ij}^2 = \frac{\sum_{i=1}^n (\hat{y_i} - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})}$ where $\hat{y_i}$ is the regression fit for the ith instance and $\bar{y}$ is the sample mean of y_i's. Under the model setting (3.2), when all the coefficients are zeros, the maximal pairwise R squares $\max\limits_{1 \leq i < j \leq p} R_{ij}^2$ cannot be too large. In fact, there exists an asymptotic bound for $\max\limits_{1 \leq i < j \leq p} R_{ij}^2$, as described in the following theorem.

Theorem 4. *Let $R_{pn}^2 = \max_{1 \leq i < j \leq p} R_{ij}^2$, where R_{ij}^2 is the pairwise R squares from regressing Y onto X_i and X_j where $i \neq j$. Suppose that $X_1, \cdots, X_p$ and Y are from the model setting 3.2 and we further assume that Y is a normally distributed. Then when β_j's are all zeros, we have for any fixed $n \geq 4$, $\delta > 0$, as $p \to \infty$, $P(R_{pn}^2 \geq 1 - p^{-(4+\delta)/(n-3)}) \to 0$.*

Proof of Theorem 4. If Y is normally distributed, then conditioning on X_i and X_j, $R_{ij}^2 | X_i, X_j$ is distributed as $\text{Beta}(1, \frac{n-3}{2})$ (Muirhead, 2009), which is independent of X_i, X_j. Therefore, the unconditional distribution of R_{ij}^2 is also $\text{Beta}(1, \frac{n-3}{2})$.

$$\begin{aligned} P(R_{pn}^2 \geq 1 - p^{-(4+\delta)/(n-3)}) &= P\big(\max_{1 \leq i < j \leq p} R_{ij}^2 \geq 1 - p^{-(4+\delta)/(n-3)}\big) \\ &= P\big(\cup_{1 \leq i < j \leq p} \{R_{ij}^2 \geq 1 - p^{-(4+\delta)/(n-3)}\}\big) \\ &\leq \frac{p(p-1)}{2} P(\{R_{ij}^2 \geq 1 - p^{-(4+\delta)/(n-3)}\}) \\ &= \frac{p(p-1)}{2} \big(p^{-(4+\delta)/(n-3)}\big)^{\frac{(n-3)}{2}} \\ &= O(p^{-\delta/2}) \to 0, \end{aligned}$$

as $p \to \infty$. □

With the bound given by Theorem 4, we can design a screening rule to find pairs of covariates that are potentially associated with the response. In Section 3.3, we introduce how to make use of the theoretical results to benefit variable selection.

3.3 Penalized variable selection using pairwise screening

In this section, we propose a pairwise screening procedure that takes advantages of the asymptotic results in Section 3.2. We further establish a new penalization algorithm for variable selection.

3.3.1 Screening-based penalization

Given the limiting distribution of the maximal pairwise sample correlation described in Section 3.2, we propose the following screening rule to identify covariates pairs that are potentially correlated and related to the response:

$$\mathcal{G} = \{(i,j) : i < j, |\widehat{\text{Corr}}(X_i, X_j)| \geq a \text{ and } R_{ij} \geq r_0\}, \tag{3.6}$$

where a is the $100(1-\alpha)\%$ quantile of the distribution given in Theorem 3 (for Gaussian covariates) or Proposition 2 (for non-Gaussian covariates), and $r_0 = 1 - p^{-(4+\delta)/(n-3)}$. Note that the values of α and δ can affect the size of $\mathcal{G}$. The larger α and δ are, there are fewer pairs included in $\mathcal{G}$. In practice, we suggest to take $\alpha = 0.05$ and $\delta = 0.1$.

The group definition in (3.6) is a screening procedure with respect to covariates pairs. The idea of screening is prevalent for high dimensional data analysis. In particular, for penalized variable selection methods, increasing dimensionality makes it more difficult to capture the inherent sparsity structure. Therefore, dimension reduction is necessary when there are tens of thousands of candidate variables. To this end, Fan and Lv (2008) introduced the Sure Independence Screening (SIS) method, which ranks the covariates based on the magnitude of their sample correlation with the response. Let $\mathbf{w} = (w_1, w_2, \cdots, w_p)^T$ be a vector such that $w_j = |\widehat{\text{Corr}}(X_j, Y)|$ and γ is a constant between $(0,1)$, then a sub-model is defined as

$$\mathcal{M}_\gamma = \{j : w_j \text{ is amongst the largest } [\gamma n] \text{ of all}\}, \tag{3.7}$$

where $[\gamma n]$ denotes the integer part of γn. Fan and Lv (2008) further demonstrated that SIS is screening consistent under some conditions. This guarantees that all those X_j's with $\beta_j \neq 0$ is included in the subset of covariates.

To take advantage of the distribution information while implementing dimension reduction, we propose a new penalized variable selection approach that applies different penalties to each covariate based on the screening results. Let $\mathcal{M}$ be the index set of covariates that have the largest $[n \backslash \log n]$ absolute sample correlation with the response among $X_1, X_2, \cdots, X_p$. We also define the set of paired covariates as

$$\mathcal{C} = \{X_i : \exists j \neq i \text{ such that } (i,j) \in \mathcal{G}\}. \tag{3.8}$$

Our proposed method is established by solving the following optimization problem:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2n} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + \lambda_1 \sum_{j : j \in \mathcal{C}^c \cap \mathcal{M}} |\beta_j| + \lambda_2 \sum_{j : j \in \mathcal{C} \cap \mathcal{M}} \beta_j^2 \tag{3.9}$$

subject to $\beta_j = 0$ for $j \notin \mathcal{M}$. In other words, we ignore the covariates that fail the marginal screening.

From the above penalty, it can be seen that we apply different penalties to covariates based upon the results from two types of screening. The intuition behind the proposed penalty is

- For a covariate that is included in both $\mathcal{C}$ and $\mathcal{M}$, we only apply the l_2 penalty because it tends to be an important variable that we need to include in the final model.

- For a covariate that is included in $\mathcal{M}$ but not in $\mathcal{C}$, we only apply the l_1 penalty since there is no significant multicollinearity between it and other covariates.

- For a covariate that is not included in $\mathcal{M}$, since it does not pass the marginal screening, we no longer consider it in the regression. This is because SIS enjoys screening consistency under certain assumptions, which implies that $\mathcal{M}$ covers all important variables.

Our proposed method is connected with existing penalization approaches when the covariates have certain covariance structure. In particular, when the covariates are all independent, our method reduces to SIS-LASSO, which performs marginal screening first and then implements LASSO on the remaining covariates; when the predictors are all highly correlated such that $\mathcal{G}$ includes all covariates pairs, our method is equivalent to SIS-Ridge.

So far we have proposed a new penalized variable selection. Now we discuss how to solve the optimization problem in (3.9). One can see that the penalty part of (3.9) is convex, so we can efficiently solve it by coordinate descent algorithm (Friedman et al., 2010). Specifically, the updating rule has the following form:

$$\hat{\beta}_j \leftarrow \begin{cases} S(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_i^{(j)}), \lambda_1) & \text{for } j \in \mathcal{C}^c \cap \mathcal{M}, \\ \dfrac{\frac{1}{n}\sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_i^{(j)})}{1+\lambda_2} & \text{for } j \in \mathcal{C} \cap \mathcal{M}, \end{cases} \tag{3.10}$$

where $\tilde{y}_i^{(j)} = \hat{\beta}_0 + \sum_{k \neq j} x_{ik}\hat{\beta}_k$ is the fitted value excluding the effect of x_{ij}, and $S(z, \lambda) = \text{sign}(z)(|z| - \lambda)_+$ is the soft-thresholding function. In practice, we can first implement SIS to obtain $\mathcal{M}$ when the dimension is high, then run the algorithm on the covariates X_j's with $j \in \mathcal{M}$.

3.3.2 Further extensions

As discussed in the previous subsection, we introduce a new penalized method that combines marginal screening with pairwise screening under the linear model setting. Note that the pairwise covariates screening does not involve the response. Therefore, our method can be further extended to generalized linear models (GLM), e.g., logistic regression for binary response, or cox model for survival data. Suppose the response Y is from the following one-parameter exponential family $f(y|\mathbf{x}, \theta) = h(y)\exp\{y\theta - b(\theta)\}$. Moreover, we assume $\theta = \mathbf{x}^T\boldsymbol{\beta}$ for generalized linear models.

Similar to 3.6, we define the pairwise screening as

$$\mathcal{G}_1 = \{(i, j) : i < j, |\widehat{\text{Corr}}(X_i, X_j)| \geq a\}. \tag{3.11}$$

The difference is that we do not consider the R squares screening for GLMs. This is because for GLMs, it is not reasonable to use the regression R squares to evaluate the associations between the covariates and the response. We further define the set of paired covariates as follows

$$\mathcal{C}_1 = \{i : \exists j \text{ such that } (i, j) \in \mathcal{G}_1\}. \tag{3.12}$$

Let

$$P_{\lambda_1,\lambda_2}(\boldsymbol{\beta}) = \lambda_1 \sum_{j:j\in\mathcal{C}_1^c\cap\mathcal{M}} |\beta_j| + \lambda_2 \sum_{j:j\in\mathcal{C}_1\cap\mathcal{M}} \beta_j^2$$

be our proposed screening-based penalty. Then for logistic regression, we need to solve the following penalized maximum likelihood problem

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \big(y_i(\mathbf{x}_i^T\boldsymbol{\beta}) - \log(1 + e^{\mathbf{x}_i^T\boldsymbol{\beta}})\big) + P_{\lambda_1,\lambda_2}(\boldsymbol{\beta}). \tag{3.13}$$

In the above optimization problem, the log likelihood part can be approximated by a quadratic function, which is a weighted least squares term (Friedman et al., 2010). Therefore, it can still be solved by coordinate descent algorithm. Similarly, we can use the algorithm proposed by Simon et al. (2011) to solve the regularized Cox proportional hazard model using the screening based penalty $P_{\lambda_1,\lambda_2}(\boldsymbol{\beta})$.

In Section 3.5, we will show with numerical examples that our proposed method can perform well in practice.

3.4 Theoretical properties

In this section, we study the theoretical properties of the proposed pairwise correlation screening (PCS) method. More specifically, we investigate the conditions under which the PCS achieves the variable selection consistency.

Note that we implemented the marginal screening using SIS to the covariates set. We also introduce the details of the SIS procedure in Section 3.3.1. Fan and Lv (2008) demonstrated that under certain regularity conditions, SIS has the screening consistency, that is, the resulting subset of covariates $\mathcal{M}_\gamma$ contain all important variables. We present this result in the following proposition.

Proposition 3. *[Fan and Lv (2008)] Suppose X and $\mathbf{x}$ are defined the same as in* (3.2). *Define $\mathbf{z} = \Sigma^{-1/2}\mathbf{x}$, $Z = X\Sigma^{-1/2}$. Let $\mathcal{M}^*$ be the index set of covariates with non-zero coefficient. The following assumptions are imposed:*

(A1) $p > n$ and $\log(p) = O(n^{\epsilon})$ for some $\epsilon \in (0, 1 - 2\kappa)$, where κ is given by condition (A3).

(A2) $\mathbf{z}$ has a spherically symmetric distribution, and $\exists c_0, c_1 > 1, C_1 > 0$ such that

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

holds for any $n \times \tilde{p}$ submatrix $\tilde{Z}$ of Z with $c_0 n < \tilde{p} \le p$.

(A3) $Var(Y) = O(1)$, and for some $\kappa \ge 0$ and $c_2, c_3 > 0$,

$$\min_{j \in \mathcal{M}^*} |\beta_j| \ge \frac{c_2}{n^{\kappa}}, \quad \min_{j \in \mathcal{M}^*} Cov(\beta_j^{-1} Y, X_j) \ge c_3$$

(A4) There are some $\tau \ge 0$ and $c_4 > 0$ such that $\lambda_{max}(\Sigma) \le c_4 n^{\tau}$.

Under conditions $(A1) - (A4)$, if $2\kappa + \tau < 1$, then there is some $\theta < 1 - 2\kappa - \tau$ such that , when $\gamma \sim cn^{-\theta}$ with $c > 0$, we have, for some $C > 0$,

$$\Pr\left(\mathcal{M}^* \subset \mathcal{M}_\gamma\right) = 1 - O[\exp\{-C^{1-2\kappa}/\log(n)\}], \tag{3.14}$$

where $\mathcal{M}_\gamma$ is the subset of covariates obtained from the sure independence screening.

The above proposition guarantees that all important variables survive the marginal screening with high probability. In order to achieve the selection consistency, we also need to ensure that only important variables can pass the pairwise screening. In the following theorem, we present the technical conditions that are required such that the event $\mathcal{C} \cap \mathcal{M} \subset \mathcal{M}^*$ occurs with high probability.

Theorem 5. *Suppose the following conditions holds*

(B1) $n/p^2 \to 0$.

(B2) There exists $\eta > 0$ such that either one of the following two conditions holds:

(a) $\lim_{n\to\infty} \log p/n \to \eta_0, \max_{i\in\mathcal{M}^*, j\in\mathcal{M}\setminus\mathcal{M}^*} |\mathrm{Corr}(\mathrm{X_i}, \mathrm{X_j})| < \min\{\eta, 1-\mathrm{e}^{-4\eta_0}\}$

(b) $\lim_{n\to\infty} \log p/n \to 0, \max_{i\in\mathcal{M}^*, j\in\mathcal{M}\setminus\mathcal{M}^*} |\mathrm{Corr}(\mathrm{X_i}, \mathrm{X_j})| < \eta.$

Under conditions $(B1)$ and $(B2)(a)$ or conditions $(B1)$ and $(B2)(b)$, we have that

$$\Pr(\mathcal{C}\cap\mathcal{M}\subset\mathcal{M}^*) \to 1. \tag{3.15}$$

Let $C = \frac{1}{n}X^TX$. Without of loss of generality, assume that $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_s)^T$ where $\beta_j \neq 0$ for $j = 1, \ldots, s$. We further assume that $\{1, \cdots, s_1\} \subset \mathcal{C}\cap\mathcal{M}$. Then the design matrix X can be expressed as $X = (X^1_{(1)}, X^2_{(1)}, X_{(2)})$, where $X^1_{(1)}$ corresponds to the first s_1 columns, $X^2_{(1)}$ corresponds to the (s_1+1)th to the sth columns and $X_{(2)}$ corresponds to the last $p-s$ columns of X respectively. Similarly, we can write $\boldsymbol{\beta}^{(1)}_1 = (\beta_1, \ldots, \beta_{s_1})$, $\boldsymbol{\beta}^{(1)}_2 = (\beta_{s_1+1}, \ldots, \beta_s)$, and $\boldsymbol{\beta}^{(2)} = (\beta_{s+1}, \ldots, \beta_p)$.

Set $C^{(11)}_{11} = \frac{1}{n}{X^1_{(1)}}^T X^1_{(1)}$, $C^{(12)}_{11} = \frac{1}{n}{X^1_{(1)}}^T X^2_{(1)}$, $C^{(21)}_{11} = \frac{1}{n}{X^2_{(1)}}^T X^1_{(1)}$, $C^{(22)}_{11} = \frac{1}{n}{X^2_{(1)}}^T X^2_{(1)}$, $C^{(1)}_{21} = \frac{1}{n}X^T_{(2)}X^1_{(1)}$, $C^{(2)}_{21} = \frac{1}{n}X^T_{(2)}X^2_{(1)}$, $C_{22} = \frac{1}{n}X^T_{(2)}X_{(2)}$, $C^{(1)}_{12} = \frac{1}{n}{X^1_{(1)}}^T X_{(2)}$, $C^{(2)}_{12} = \frac{1}{n}{X^2_{(1)}}^T X_{(2)}$. Then C can be expressed in a block-wise form as follows:

$$\begin{pmatrix} C^{(11)}_{11} & C^{(12)}_{11} & C^{(1)}_{12} \\ C^{(21)}_{11} & C^{(22)}_{11} & C^{(2)}_{12} \\ C^{(1)}_{21} & C^{(2)}_{21} & C_{22} \end{pmatrix}$$

To ensure variable selection consistency, we need to impose the following assumption analogous to the irrepresentability condition proposed by Zhao and Yu (2006). Specifically, we assume that there exists a constant $\delta > 0$, such that

$$\|C^{(1)}_{21}(C^{(22)}_{11})^{-1}\mathrm{sign}(\boldsymbol{\beta}^2_{(1)})\|_{\max} \leq 1-\delta, \tag{3.16}$$

where $\|\cdot\|_{\max}$ is the max norm.

In fact, we can show that the condition mentioned above is implied by the irrepresentable condition on the full covariates set $\mathcal{M}$ under mild assumptions. We illustrate this result in the following theorem:

Theorem 6. *Assume that there exists $\lambda_0 > 0$ so that $\lambda_{min}(C_{11}^{(11)}) \geq \lambda_0$, $\lambda_{min}(C_{11}^{(22)}) \geq \lambda_0$, and conditions $(B1)$ and $(B2)(b)$ holds. Suppose the irrepresentability condition holds, i.e., $\exists \xi > 0$ s.t.*

$$\|\Sigma_{21}\Sigma_{11}^{-1} sign(\boldsymbol{\beta}_1)\|_{\max} \leq 1 - \xi, \tag{3.17}$$

where $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ with Σ_{11} corresponding to $X_{(1)}$ and Σ_{22} corresponding to $X_{(2)}$, $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_s)^T$ and ξ is a positive constant. Then with probability tending to 1, the condition (3.16) *holds.*

So far we have discussed all the theoretical assumptions required to ensure the selection consistency of our PCS method. We conclude the consistency results in the following theorem:

Theorem 7. *Suppose the conditions (A1) to (A4) and inequality* (3.17) *hold, and the assumptions of Theorem 6 are satisfied, then as $n \to \infty$,*

$$\Pr\left(\{j : \hat{\beta}_j \neq 0\} = \mathcal{M}^*\right) \to 1, \tag{3.18}$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ is the solution to (3.9)*.*

The proof follows immediately from Proposition 3 and Theorems 5 and 6. It shows that under certain conditions, the proposed method is consistent in variable selection.

3.5 Numerical Studies

In Section 3, we have established a new regularized variable selection approach for high-dimensional linear models. In this section, we demonstrate the performance of our proposed method using both simulations and real data examples.

3.5.1 Simulation study

In this section, we use several simulated examples to show that our method with pairwise correlation screening (PCS) or pairwise rank-based correlation screening (PRCS) outperforms some existing variable selection procedures. More specifically, PCS denotes our proposed method using the limiting distribution in Theorem 3, and PRCS uses the asymptotic result in Proposition 2.

For comparison, we consider LASSO, elastic net (Enet), SIS-LASSO, SIS-elastic net (SIS-Enet) and SIS-PACS. The SIS-PACS refers to applying the PACS method proposed by Sharma et al. (2013) after implementing the SIS procedure. In SIS-type methods, we first implement SIS and find those covariates with the largest $[n\backslash\log n]$ absolute sample correlations with the response, then perform LASSO, Enet or PACS on these variables. We evaluate the variable selection accuracy using False Negatives (FN) and False Positives (FP). FN is defined as $FN = \sum_{j=1}^{p} I(\hat{\beta}_j = 0) \times I(\beta_j \neq 0)$, where $I(\cdot)$ denotes the indicator function, and FP is defined as $FP = \sum_{j=1}^{p} I(\hat{\beta}_j \neq 0) \times I(\beta_j = 0)$. We use the following quantities to evaluate the prediction accuracy:

- $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2$: the l_2 distance between the estimated coefficient vector and the true coefficients $\boldsymbol{\beta}_0$;
- Out of sample mean squared errors (MSE) on the independent test data;

We generate the simulated data from Model (3.2) and conduct 100 replications. Each simulated dataset includes a training set of size 100, an independent validation set of size 100 and an independent test set of size 400. Here we fix the sample size to be 100 throughout the simulation study. In the next subsection, we also consider varying sample size for sensitivity study. We only fit models on the training data, and we use the validation data to select tuning parameters. Given the fitted model, we can calculate the FN, FP and the estimation error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$, and we make predictions and calculate the out of sample MSEs using the test data. We simulate the covariates from the multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$, with $\Sigma = (\sigma_{ij})_{p\times p}$ being the correlation matrix.

Details of the simulated examples are as follows:

Example 1: We consider $p = 1000$ or 5000, $\sigma = 2$ or 6, and we take $\boldsymbol{\beta} = (2, 2, \cdots, 2, 0, \cdots, 0)^T$ where the first 10 coefficients being non-zero and equal to 2. We set $\sigma_{ij} = 0.8$ for $1 \leq i \neq j \leq 5$, $6 \leq i \neq j \leq 10$ and 0 for all the other $i \neq j$. We also consider $\sigma = 6$ and present the results in the supplementary. In other words, there are two groups in the covariates, where each group has 5 important variables.

Table 3.1: Average MSE, l_2 distance, FN and FP over 100 replications for simulated example 1 (with standard errors given in parentheses).

Method	MSE	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$	FN	FP
		$p = 1000, \quad \sigma = 2$		
Elnet	5.94 (0.07)	1.40 (0.03)	0.00 (0.00)	1.64 (0.24)
SIS-Elnet	5.47 (0.06)	1.30 (0.03)	0.00 (0.00)	1.15 (0.12)
LASSO	5.95 (0.07)	1.50 (0.03)	0.00 (0.00)	1.28 (0.18)
SIS-LASSO	5.47 (0.06)	1.42 (0.03)	0.00 (0.00)	0.85 (0.10)
SIS-Ridge	86.00 (0.76)	4.5 (0.01)	0.00 (0.00)	12.00 (0.00)
PACS	4.69 (0.07)	0.48 (0.02)	0.00 (0.00)	0.01 (0.01)
PCS	4.74 (0.05)	0.76 (0.02)	0.00 (0.00)	0.03 (0.02)
PCRS	4.91 (0.05)	0.93 (0.02)	0.00 (0.00)	2.55 (0.15)
		$p = 5000, \quad \sigma = 2$		
Elnet	6.42 (0.09)	1.57 (0.03)	0.00 (0.00)	2.45 (0.26)
SIS-Elnet	5.64 (0.06)	1.41 (0.03)	0.00 (0.00)	1.28 (0.12)
LASSO	6.41 (0.08)	1.64 (0.04)	0.00 (0.00)	2.06 (0.21)
SIS-LASSO	5.65 (0.06)	1.52 (0.03)	0.00 (0.00)	1.03 (0.10)
SIS-Ridge	88.74 (0.75)	4.59 (0.01)	0.00 (0.00)	12.00 (0.00)
PACS	4.97 (0.08)	0.72 (0.02)	0.00 (0.00)	1.78 (0.43)
PCS	4.77 (0.05)	0.81 (0.03)	0.00 (0.00)	0.02 (0.02)
PCRS	4.85 (0.06)	0.89 (0.03)	0.00 (0.00)	1.21 (0.11)
		$p = 1000, \quad \sigma = 6$		
Elnet	52.11 (0.59)	3.31 (0.07)	0.81 (0.09)	1.85 (0.26)
SIS-Elnet	50.68 (0.53)	3.15 (0.07)	0.63 (0.08)	1.81 (0.20)
LASSO	52.52 (0.57)	3.96 (0.06)	1.50 (0.10)	1.13 (0.17)
SIS-LASSO	50.88 (0.54)	3.91 (0.06)	1.44 (0.10)	1.03 (0.13)
SIS-Ridge	119.9 (1.01)	4.59 (0.01)	0.00 (0.00)	12.00 (0.00)
SIS-PACS	52.50 (0.67)	3.40 (0.06)	0.00 (0.00)	4.86 (0.04)
PCS	41.68 (0.38)	1.67 (0.07)	0.06 (0.04)	0.00 (0.00)
PRCS	43.12 (0.37)	2.04 (0.08)	0.06 (0.04)	2.05 (0.14)
		$p = 5000, \quad \sigma = 6$		
Enet	55.57 (0.64)	3.55 (0.06)	0.99 (0.11)	2.47 (0.29)
SIS-Enet	53.86 (0.60)	3.45 (0.07)	0.99 (0.10)	1.83 (0.19)
LASSO	55.95 (0.64)	4.16 (0.06)	1.77 (0.12)	1.55 (0.17)
SIS-LASSO	53.78 (0.61)	4.02 (0.06)	1.68 (0.10)	1.22 (0.13)
SIS-Ridge	123.29 (1.03)	4.68 (0.01)	0.00 (0.00)	12.00 (0.00)
SIS-PACS	56.45 (0.74)	3.80 (0.04)	0.00 (0.00)	4.94 (0.03)
PCS	42.76 (0.42)	1.96 (0.11)	0.25 (0.07)	0.04 (0.02)
PRCS	43.16 (0.47)	2.11 (0.11)	0.25 (0.07)	0.80 (0.09)

Example 2: We consider $p = 1000$ or 5000, $\sigma = 2$ or 6, $\boldsymbol{\beta}_0 = (3, -1.5, 2, 0, \cdots, 0, \cdots, 0)^T$, where the first 3 coefficients are non-zero ones. We also consider $\sigma = 6$ and present the results in the supplementary. We generated Gaussian covariates with $\sigma_{ij} = 0.5^{|i-j|}$ for $1 \leq i \neq j \leq 1000$.

Table 3.2: Results for simulated example 2. The format of this table is the same as Table 3.1.

Method	MSE	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$	FN	FP
		$p = 1000, \quad \sigma = 2$		
Enet	6.75 (0.08)	2.45 (0.02)	1.00 (0.01)	0.98 (0.25)
SIS-Enet	6.47 (0.10)	2.30 (0.03)	0.76 (0.05)	3.16 (0.41)
LASSO	6.75 (0.08)	2.45 (0.02)	1.00 (0.01)	0.98 (0.25)
SIS-LASSO	6.47 (0.10)	2.30 (0.03)	0.76 (0.05)	3.16 (0.41)
SIS-Ridge	14.14 (0.10)	3.85 (0.00)	0.27 (0.04)	19.27 (0.04)
SIS-PACS	6.53 (0.14)	2.43 (0.04)	1.06 (0.05)	3.39 (0.73)
PCS	5.24 (0.12)	1.41 (0.08)	0.34 (0.05)	1.63 (0.13)
PCRS	5.72 (0.13)	1.75 (0.08)	0.43 (0.05)	1.34 (0.24)
		$p = 5000, \quad \sigma = 2$		
Elnet	7.16 (0.08)	2.55 (0.02)	1.02 (0.01)	0.40 (0.09)
SIS-Elnet	7.02 (0.09)	2.49 (0.03)	0.94 (0.03)	1.31 (0.34)
LASSO	7.16 (0.08)	2.55 (0.02)	1.02 (0.01)	0.36 (0.08)
SIS-LASSO	7.03 (0.09)	2.49 (0.03)	0.94 (0.03)	1.31 (0.34)
SIS-Ridge	14.40 (0.11)	3.87 (0.00)	0.59 (0.05)	19.59 (0.05)
SIS-PACS	7.28 (0.16)	2.83 (0.04)	1.26 (0.07)	2.41 (0.95)
PCS	5.96 (0.14)	1.83 (0.09)	0.63 (0.06)	0.74 (0.08)
PCRS	6.48 (0.13)	2.14 (0.07)	0.68 (0.05)	0.73 (0.24)
		$p = 1000, \quad \sigma = 6$		
Elnet	45.03 (0.35)	3.73 (0.03)	2.28 (0.07)	1.30 (0.59)
SIS-Elnet	45.08 (0.35)	3.75 (0.02)	2.31 (0.07)	1.53 (0.51)
LASSO	45.03 (0.36)	3.74 (0.03)	2.35 (0.06)	0.12 (0.04)
SIS-LASSO	45.09 (0.35)	3.75 (0.02)	2.43 (0.06)	0.12 (0.04)
SIS-Ridge	46.08 (0.30)	3.90 (0.00)	1.07 (0.07)	20.07 (0.07)
SIS-PACS	45.45 (0.34)	3.91 (0.02)	1.07 (0.07)	4.03 (0.06)
PCS	44.01 (0.46)	3.51 (0.06)	2.2 (0.08)	0.24 (0.05)
PRCS	44.98 (0.35)	3.73 (0.03)	2.37 (0.07)	0.14 (0.04)
		$p = 5000, \quad \sigma = 6$		
Elnet	45.78 (0.35)	3.84 (0.01)	2.48 (0.07)	1.09 (0.67)
SIS-Elnet	45.77 (0.35)	3.84 (0.02)	2.47 (0.05)	0.77 (0.36)
LASSO	45.78 (0.35)	3.84 (0.01)	2.57 (0.05)	0.20 (0.04)
SIS-LASSO	45.75 (0.35)	3.83 (0.02)	2.50 (0.05)	0.15 (0.04)
SIS-Ridge	46.14 (0.35)	3.90 (0.00)	1.42 (0.06)	20.42 (0.06)
SIS-PACS	45.76 (0.38)	3.85 (0.02)	2.46 (0.06)	0.76 (0.06)
PCS	45.80 (0.36)	3.85 (0.01)	2.61 (0.05)	0.12 (0.04)
PRCS	45.79 (0.36)	3.84 (0.02)	2.62 (0.05)	0.13 (0.05)

Example 3: The coefficients have the same set up as in Example 1. But we set $\sigma_{ij} = 0.8$ for $1 \leq i \neq j \leq 5$ and 0 for all the other $i \neq j$. Therefore only part of the important variables are highly correlated. We consider $p = 5000$ and $\sigma = 6$ in this Example.

Table 3.3: Results for simulated example 3. The format of this table is the same as Table 3.1.

Method	MSE	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$	FN	FP
Enet	69.71 (0.88)	5.13 (0.03)	4.99 (0.13)	1.57 (0.37)
SIS-Enet	72.54 (0.88)	5.25 (0.03)	5.65 (0.10)	0.23 (0.12)
LASSO	72.78 (0.87)	5.41 (0.03)	6.06 (0.10)	0.09 (0.04)
SIS-LASSO	70.12 (0.86)	5.35 (0.04)	5.69 (0.12)	0.94 (0.19)
SIS-Ridge	109.66 (0.87)	5.74 (0.01)	4.46 (0.06)	16.46 (0.06)
SIS-PACS	71.27 (0.89)	5.58 (0.02)	5.06 (0.02)	3.45 (0.07)
PCS	58.87 (0.50)	4.80 (0.04)	4.95 (0.03)	0.06 (0.06)
PCRS	59.76 (0.56)	4.83 (0.04)	4.97 (0.02)	0.00 (0.00)

Table 3.4: Results for simulated example 4. The format of this table is the same as Table 3.1.

Method	Classification Error	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$	FN	FP
Enet	0.129 (0.003)	5.79 (0.01)	2.16 (0.17)	12.77 (1.54)
SIS-Enet	0.126 (0.003)	5.69 (0.03)	1.37 (0.15)	7.48 (0.39)
LASSO	0.136 (0.003)	5.83 (0.01)	4.19 (0.13)	4.25 (0.49)
SIS-LASSO	0.130 (0.003)	5.75 (0.02)	3.94 (0.12)	3.50 (0.32)
SIS-Ridge	0.311 (0.003)	6.28 (0.01)	0.11 (0.05)	12.11 (0.05)
PCS	0.098 (0.004)	5.39 (0.05)	1.73 (0.14)	2.92 (0.31)
PCRS	0.099 (0.004)	5.34 (0.06)	1.71 (0.13)	3.26 (0.32)

Table 3.5: Results for simulated example 5. The format of this table is the same as Table 3.1.

Method	MSE	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$	FN	FP
Enet	102.47 (1.84)	3.90 (0.08)	1.51 (0.12)	4.88 (0.86)
SIS-Enet	96.60 (2.74)	3.49 (0.09)	1.02 (0.12)	4.20 (0.37)
LASSO	103.11 (1.89)	4.42 (0.08)	2.30 (0.13)	3.74 (0.71)
SIS-LASSO	96.97 (2.78)	4.27 (0.08)	2.05 (0.14)	1.87 (0.20)
SIS-Ridge	226.52 (3.78)	4.95 (0.03)	0.26 (0.08)	12.26 (0.08)
SIS-PACS	84.20 (1.77)	3.24 (0.09)	0.89 (0.09)	7.32 (0.45)
PCS	79.79 (3.16)	2.42 (0.14)	0.42 (0.10)	1.29 (0.33)
PCRS	74.60 (1.24)	2.15 (0.12)	0.31 (0.08)	0.06 (0.03)

Example 4: In this example, we examine the performance of all methods under the logistic regression setting. We simulate the binary response Y from the binomial distribution Binom($1, \frac{\exp\{X^T\boldsymbol{\beta}+\sigma\}}{1+\exp\{X^T\boldsymbol{\beta}+\sigma\}}$), where X, and $\boldsymbol{\beta}$ follow the same set ups as in Example 1.We consider $p = 5000$ and $\sigma = 6$ in this Example. Instead of comparing MSE, we calculate the classification errors on the test data. We did not compare with SIS-PACS in this example since the R program does not support GLM.

Example 5: In this example, we generate the covariates from a multivariate t distribution, where X_j's are t distributed with degrees of freedom 5. The covariance structure of the covariates and the coefficients are set the same as in Example 1. We consider $p = 5000$ and $\sigma = 6$ in this Example.

The results for simulated example 1 is shown in Table 3.1. We see that when there are groups in the covariates, the performance improvement of our approach is significant compared with other penalized methods. While elastic net-based procedures perform better than LASSO-type approaches in terms of FN, as illustrated by Zou and Hastie (2005), they still miss approximately one important covariate on average. In contrast, the model selection results of our method are much closer to the correct model for this example. In addition, although SIS-PACS has competitive performance when σ is small, it tends to include more unimportant variables into the model when the noise level increases, and therefore may not work well.

Table 3.2 displays the performance comparisons for Example 2. Compared with Example 1, this setting is a more difficult one for our method, since correlation exists among all pairs of covariates. Nevertheless, PCS and PCRS perform better than, or as well as all the others in terms of estimation error and prediction accuracy. Moreover, besides SIS-Ridge, our proposed methods are able to identify more important variables than others in this example when the noise level is low.

Table 3.3 shows the results for Example 3, where correlation exists only within part of the important variables. This example is more difficult compared with Example 1 due to the correlation structure of the covariates. One can see that the false negatives are significantly larger for all procedures. Nevertheless our method still outperforms all the others in terms of prediction and variable selection accuracy.

Example 4 considers the logistic regression setting, and the results are provided in Table 3.4. One can see that as the correlations among the covariates vary, the performance of our method is always competitive compared with the others.

Table 3.5 displays the results for all methods under the non-Gaussian covariates setting. Similar to Example 1, our proposed PCS and PCRS achieve much better performance compared with the competitors. Moreover, due to the non-Gaussian set ups, the nonparametric method PCRS outperforms PCS.

As a conclusion, our method can make use of the correlation structure among predictors. Compared with other penalized variable selection procedures, our method performs well, especially when the covariates are highly correlated.

3.5.2 Sensitivity Study

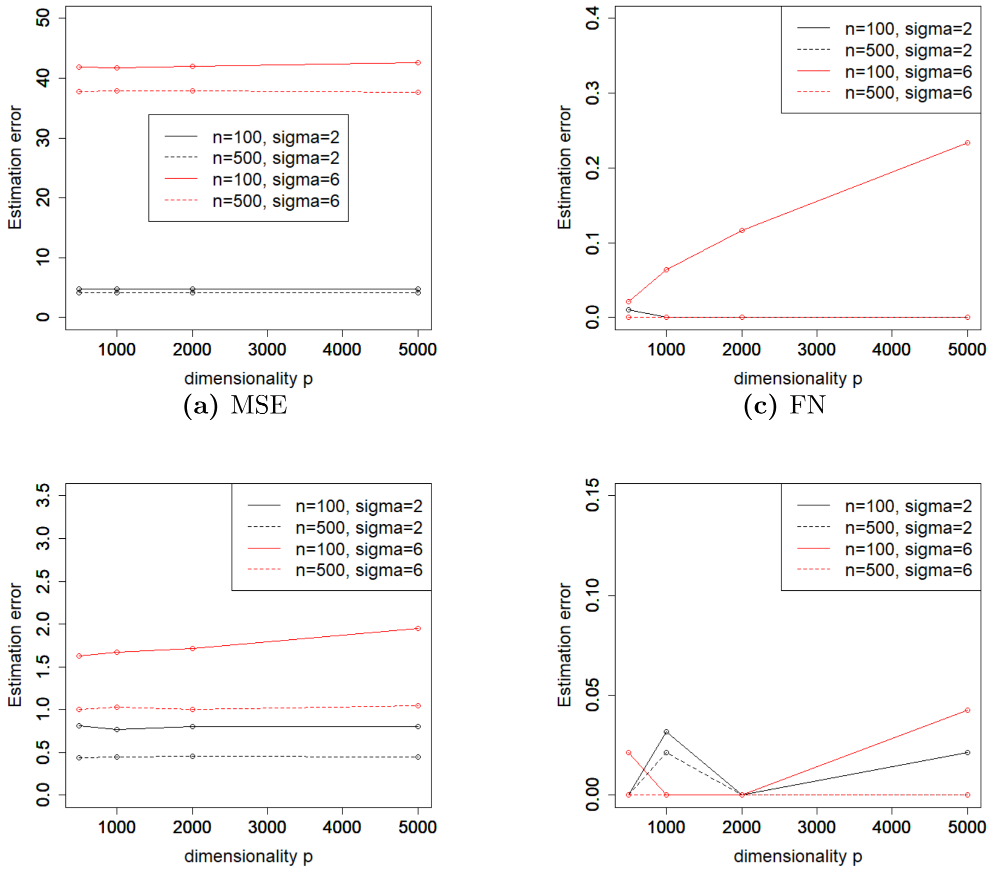
In this subsection, we investigate how the performance of our method depends on the sample size, dimensionality, and noise level. In particular, we consider $n = 100, 500$, $p = 500, 1000, 2000, 5000$ and $\sigma = 2, 6$ in the Simulated Examples 1 and 2 as introduced in Section 3.5.1. We illustrate the MSE, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$, FN and FP against different values of p for each configuration of sample size and noise level in Figures 3.1 and 3.2.

One can see from the plots that the performance of PCS does not change much as the dimensionality p increases from 500 to 5000, especially in terms of MSE and the estimation error of $\boldsymbol{\beta}_0$. Moreover, the performance is better when the sample size and signal to noise ratio (SNR) become larger, which is expected. In general, our proposed PCS method is robust to sample size, dimensionality and SNR.

3.5.3 Soil data

We first demonstrate the performance of our method in real applications using a small dataset. This dataset contains 15 covariates of soil characteristics for 20 plots with the same area in the Appalachian Mountain. The outcome variable is the forest diversity for each plot. More descriptions of the data can be found in Bondell and Reich (2008). To better demonstrate the correlation structure of covariates, we obtain the absolute pairwise correlation matrix and show the heatmap in Figure 3.3. One can see that some predictors are highly correlated. In particular, the magnitude of the pairwise correlations among Sum of Cations (SumCation), calcium, magnesium, Base Saturation (BaseSat), and cation exchange capacity (CEC) are as large as 0.9. The reason is SumCation, BaseSat, CEC are characteristics for cations; while calcium and magnesium are examples of cations (Bondell and Reich, 2008).

We conduct a total of 100 replications. In each replication, 15 samples are randomly chosen as the training set and the remaining as the test set. As in the simulation experiments, we applied LASSO, Enet, Ridge and our proposed PCS, PCRS to the dataset. For each method,



(a) MSE

(c) FN

(b) Estimation error

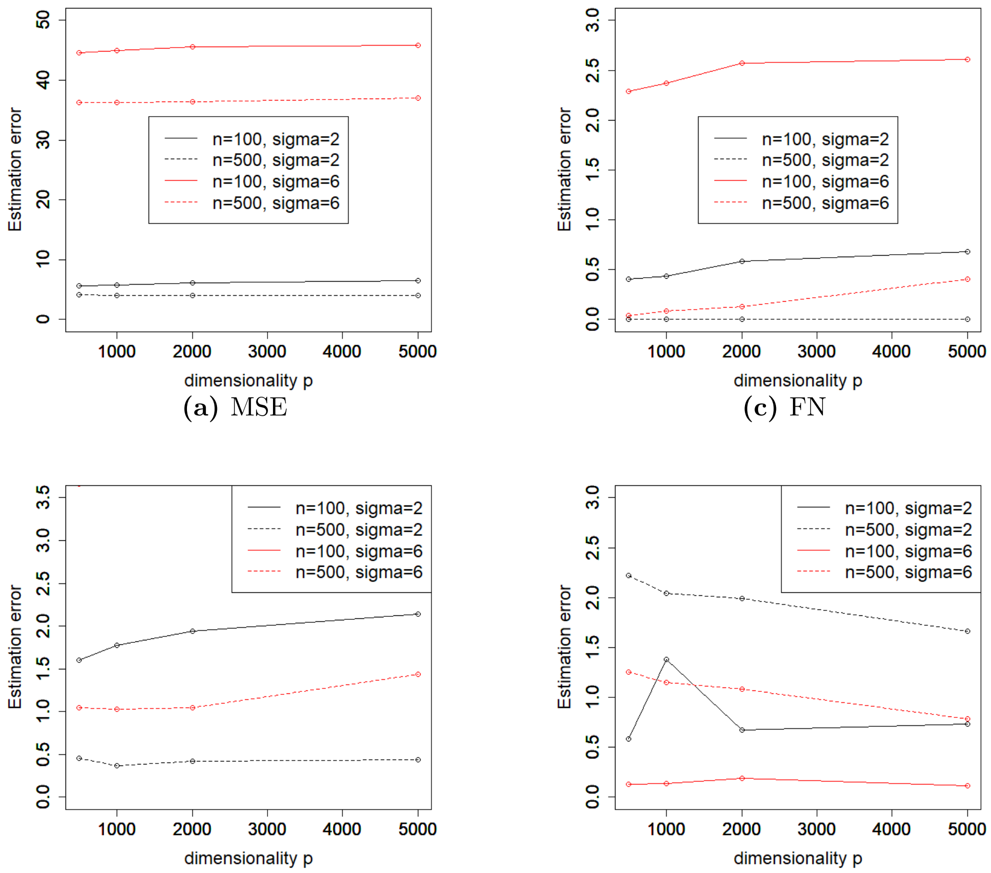
(d) FP

Figure 3.1: Performance of PCS against different dimensionality p.

5-fold cross-validation is used to choose the tuning parameters since the sample size is very small. We report the average prediction errors on the test data and the model size in Table 3.6. One can see that PCS and PCRS outperform all the others in terms of prediction accuracy.

Method	MSE	Model Size
Enet	1.088 (0.047)	3.70 (0.38)
LASSO	1.068 (0.045)	2.08 (0.21)
Ridge	1.113 (0.044)	15.00 (0.00)
PCS	0.996 (0.062)	5.82 (0.37)
PCRS	1.028 (0.063)	5.96 (0.38)

Table 3.6: Average mean squared errors and model size (with standard errors in parenthesis) for Enet, LASSO, Ridge and our method on the soil data.



(a) MSE

(c) FN

(b) Estimation error

(d) FP

Figure 3.2: Performance of PCS against different dimensionality p.

Moreover, PCS and PCRS tend to include more covariates into the model compared with LASSO and Enet.

To further investigate the performance of variable selection, we summarize the frequency that each covariate is selected for LASSO, Enet and our method, which is displayed in Table 3.7. Note that among those variables that are most frequently selected by LASSO and Enet, for instance, CEC, Mn, HumicMatt, they also tend to be included for our method. Moreover, our method can identify covariates that are strongly correlated. For example, potassium, sodium and copper are variables related to cations, and all have a large sample correlation with CEC, which is a potentially important variable. These variables are frequently selected by our method, but not by Enet and LASSO.

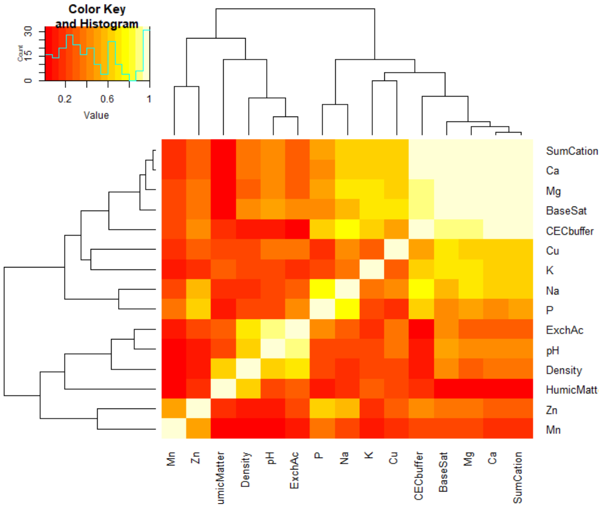


Figure 3.3: Heatmap for the absolute pairwise correlation matrix of the covariates for the soil data.

	PCS	Enet	LASSO
Variables			
BaseSat	16	9	0
SumCaton	32	23	0
CECbuffer	86	62	48
Ca	37	32	11
Mg	6	10	0
K	49	27	12
Na	22	10	6
P	32	15	5
Cu	47	17	9
Zn	29	17	4
Mn	69	43	32
HumicMatt	89	70	69
Density	25	15	4
pH	27	11	4
ExchAc	16	9	4

Table 3.7: Frequency of each variable being selected for PCS, Enet and LASSO out of 100 replications.

3.5.4 Riboflavin data

In this section, we consider a real data set about the riboflavin production in Bacillus subtilis. The data contain $n = 71$ samples, where the response variable is the logarithm of the riboflavin production rate, and the covariates are the logarithm of expression levels of $p = 4081$ genes. More descriptions about the dataset can be found in Bühlmann et al. (2014). Before analysis, all covariates are standardized to have zero means and unit standard deviations.

Method	MSE	Model Size
SIS-Enet	0.358 (0.015)	15.66 (0.46)
SIS-Lasso	0.356 (0.016)	9.12 (0.18)
SIS-Ridge	0.632 (0.024)	26.00 (0.00)
PCS	0.327 (0.014)	15.04 (0.39)
PCRS	0.361 (0.018)	12.77 (0.37)

Table 3.8: Average mean squared errors and model size (with standard errors in parenthesis) for SIS-Enet, SIS-LASSO, SIS-Ridge and PCS applied to the riboflavin data.

For comparison purpose, we apply LASSO, Enet, SIS-LASSO, SIS-Enet, SIS-ridge and our method to the dataset. We conduct 100 replications, and we randomly split the dataset into a training set of size 50 with the remaining as the test data. For all methods, we implement 10-fold cross validation on the training data to select the penalty parameters.

The results are reported in Table 3.8. One can see that PCS has significant improvement in terms of out of sample mean squared errors compared with other competitors. On the other hand, PCRS does not perform well compared with PCS. A possible reason is that in this dataset all the variables have been taken log transformations and are approximated well by Gaussian distribution. Moreover, due to the assumption of Proposition 2 where $\log p = o(n^{1/3})$, PCRS is more sensitive to the dimensionality and the sample size of dataset. As a result, PCRS may not achieve good performance when the dimensionality is too high.

We also examine the gene selection results. There are 8 genes that are selected at least 50 times out of the 100 replications by our method, i.e., XTRA_at, YCKE_at, YDAR_at, YOAB_at, YWFO_at, YXLC_at, YXLD_at and YXLE_at. Besides YXLC_at, all the other genes also appear among the most frequently selected genes by SIS-Enet and SIS-LASSO with a frequency no less than 50. For YXLC_at, we find that the magnitude of the pairwise sample correlations between this gene and two other genes, YXLD_at and YXLE_at, are greater than 0.95. It indicates that our method is capable of identifying potentially important variables that are highly correlated with the others.

3.6 Discussion

In summary, we propose a novel variable selection method that regularizes covariates selectively based on the results from two screening procedures: pairwise screening and marginal screening. The screening process of covariates pairs takes advantage of the distribution information of the maximal absolute pairwise sample correlation among covariates, and is applicable to large scale problems. Simulation experiments and real data study demonstrate that the proposed method performs well when important variables are highly correlated compared with existing approaches. For future research, we can consider other extensions of our proposed method, for example, the Cox model for survival data.

3.7 Proofs

Proof of Theorem 3. To prove Theorem 3, we need to use the following lemma, which is from Arratia et al. (1989).

Lemma 3 (Arratia et al. (1989))**.** *Let I be an index set and $\{B_\alpha, \alpha \in I\}$ be a set of subsets of I, that is, $B_\alpha \subset I$ for each $\alpha \in I$. Let also $\{\eta_\alpha, \alpha \in I\}$ be random variables. For a given $t \in R$, set $\lambda = \sum_{\alpha \in I} \Pr(\eta_\alpha > t)$. Then*

$$|\Pr\left(\max_{\alpha \in I} \eta_\alpha < t\right) - e^{-\lambda}| \leq (1 \wedge \lambda^{-1})(b_1 + b_2 + b_3)$$

where $b_1 = \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} \Pr(\eta_\alpha > t) \Pr(\eta_\beta > t)$, $b_2 = \sum_{\alpha \in I} \sum_{\alpha \neq \beta \in B_\alpha} \Pr(\eta_\alpha > t, \eta_\beta > t)$ and $b_3 = \sum_{\alpha \in I} E|\Pr(\eta_\alpha > t | \sigma(\eta_\beta, \beta \notin B_\alpha)) - \Pr(\eta_\alpha > t)|$, and $\sigma(\eta_\beta, \beta \notin B_\alpha)$ is the σ-algebra generated by $\{\eta_\beta, \beta \notin B_\alpha\}$. In particular, if η_α is independent of $\{\eta_\beta, \beta \notin B_\alpha\}$ for each α, then b_3=0.

In our proof, we take $I = \{(i,j); 1 \leq i \leq j \leq p\}$. Let $\alpha = (i,j) \in I$, we define $B_\alpha = \{(k,l) \in I$; one of k and $l = i$ or j, but $(k,l) \neq \alpha\}$, and $A_\alpha = A_{ij} = \{|\rho_{i,j}|^2 \geq t\}$, where $\rho_{i,j} = |\widehat{\mathrm{Corr}}(X_i, X_j)|$. Let $W_{pn} = \max_{1 \leq i < j \leq p} |\rho_{i,j}|$, by Lemma 3, we have

$$|P(W_{pn}^2 \leq t) - e^{-\lambda_{p,n}}| \leq b_1 + b_2, \tag{3.19}$$

where $\lambda_{p,n} = \sum_{\alpha \in I} P(A_\alpha) = \frac{p(p-1)}{2} P(A_{12})$, and $b_1 = \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} P(A_\alpha) P(A_\beta)$, $b_2 = \sum_{\alpha \in I} \sum_{\alpha \neq \beta \in B_\alpha} P(A_\alpha A_\beta)$.

Moreover, we have $b_1 \leq 2p^3 P(A_{12})^2$ and $b_2 \leq 2p^3 P(A_{12} A_{13})$.

Since $X_1, \cdots, X_p$ are independent, A_{12} and A_{13} are also independent with equal probability. Therefore we have $b_1 \vee b_2 \leq 2p^3 P(A_{12})^2$.

On the other hand, $|\rho_{i,j}|^2 \sim B(\frac{1}{2}, \frac{n-2}{2})$. Take

$$t^* = a_{p,n} + b_{p,n}x,$$

where $(x \leq \frac{n-2}{2})$, $a_{p,n} = 1 - p^{-4/(n-2)}c_{p,n}, b_{p,n} = \frac{2}{n-2}p^{-4/(n-2)}c_{p,n}$, and $c_{p,n} = \big(\frac{n-2}{2}\mathcal{B}(\frac{1}{2}, \frac{n-2}{2})\sqrt{1-p^{-4/(n-2)}}\big)^{2/(n-2)}$. Then

$$\begin{aligned} P(A_{12}^*) &= \frac{2(1-t^*)^{(n-2)/2}}{B(\frac{1}{2}, \frac{n-2}{2})(n-2)\sqrt{t^*}}(1+O(\frac{1}{\log(p)})). \\ &= p^{-2}\big(1-\frac{2}{n-2}x\big)^{\frac{n-2}{2}}\sqrt{\frac{1-p^{-4/(n-2)}}{a_{p,n}}}\big(1+(\frac{b_{p,n}}{a_{p,n}}x)\big)^{-1/2}\big(1+O(\log^{-1}(p))\big). \\ &= p^{-2}\big(1-\frac{2}{n-2}x\big)^{\frac{n-2}{2}}\big(1+O(\frac{\log\log(p)}{\log(p)})\big)\big(1+O(\log^{-1}(p))\big)^2 \\ &= p^{-2}\big(1-\frac{2}{n-2}x\big)^{\frac{n-2}{2}}\big(1+O(\frac{\log\log(p)}{\log(p)})\big) \end{aligned} \tag{3.20}$$

Therefore, uniformly for any $n \geq 3$, $b_1 \vee b_2 = O(1/p)$, and $\lim_{p\to\infty} \lambda_{p,n} = \frac{1}{2}\big(1-\frac{2}{n-2}x\big)^{\frac{n-2}{2}}$ Then it follows from (3.19) that uniformly for any $n \geq 3$ and $x \leq \frac{n-2}{2}$,

$$\lim_{p\to\infty} |P(W_{pn}^2 \leq t^*) - \exp\big\{-\frac{1}{2}\big(1-\frac{2}{n-2}x\big)^{\frac{n-2}{2}}\big\}| = 0. \tag{3.21}$$

When $x \geq \frac{n-2}{2}$, $t^* = 1 + (\frac{2}{n-2}x - 1)p^{-4/(n-2)}c_{p,n} \geq 1$. Therefore, uniformly for any $n \geq 3$,

$$\lim_{p\to\infty} P(W_{pn} \leq t^*) = 1 \tag{3.22}$$

Combining (3.21) and (3.22) we have uniformly for any $n \geq 3$,

$$\lim_{p\to\infty} |P(W_{pn} \leq t^*) - I(x \leq \frac{n-2}{2})\exp\big\{-\frac{1}{2}\big(1-\frac{2}{n-2}x\big)^{\frac{n-2}{2}}\big\} - I(x > \frac{n-2}{2})| = 0. \tag{3.23}$$

Or equivalently,

$$\lim_{p\to\infty} |P(\frac{W_{pn}^2 - a_{p,n}}{b_{p,n}} \le x) - I(x \le \frac{n-2}{2}) \exp\big\{ -\frac{1}{2}\big(1 - \frac{2}{n-2}x\big)^{\frac{n-2}{2}} \big\} - I(x > \frac{n-2}{2})| = 0. \tag{3.24}$$

□

Proof of Corollary 3. First note that $\frac{W_{pn}^2 - a_{p,n}}{b_{p,n}} \ge x$ is equivalent to

$$\log(1 - W_{pn}^2) \le \log(1 - a_{p,n} - b_{p,n}x), \tag{3.25}$$

where $\log(1 - W_{pn}^2) = T_{pn}$. The RHS of (3.25) can be further expressed as

$$\begin{aligned} \log(1 - a_{p,n} - b_{p,n}x) &= \log\big(1 - \frac{2}{n-2}p^{-4/(n-2)}c_{p,n}x - (1 - p^{-4/(n-2)}c_{p,n})\big) \\ &= \log\big(p^{-4/(n-2)}(1 - \frac{2}{n-2}x)c_{p,n}\big) \\ &= -\frac{4\log p}{n-2} + \log(1 - \frac{2}{n-2}x) + \log c_{p,n}. \end{aligned} \tag{3.26}$$

(i) **Sub-Exponential Case**

If $\log(p)/n \to 0$ as $n \to \infty$, then we have

$$\begin{aligned} c_{p,n} &= (\frac{2}{n-2}\mathcal{B}(\frac{1}{2}, \frac{n-2}{2})\sqrt{1 - p^{-4/(n-2)}})^{\frac{2}{n-2}} \\ &= \Bigg(\sqrt{\Big(\frac{(n-2)\pi}{2} + o(1)\Big)(1 - e^{-\frac{4\log p}{n-2}})}\Bigg)^{\frac{2}{n-2}} \\ &= \Bigg(\frac{(n-2)\pi}{2} \cdot \frac{4\log p}{n-2}(1 + o(1))\Bigg)^{\frac{2}{n-2}} \\ &= \exp\left\{\frac{1}{n-2}\Big(\log(2\pi\log p) + o(1)\Big)\right\} \text{ for large enough } n. \end{aligned}$$

Hence for large enough n,

$$\begin{aligned} n\log(1 - a_{p,n} - b_{p,n}x) &= -\frac{4n\log p}{n-2} + n\log(1 - \frac{2}{n-2}x) + \log 2\pi + \log\log p + o(1) \\ &= \log\log p - 4\log p + n\log(1 - \frac{2}{n-2}x) + \log 2\pi + o(1) \end{aligned} \tag{3.27}$$

Let $y = n\log(1-\frac{2}{n-2}x)+\log 2\pi$, then the RHS of (3.26) becomes $\log\log p-4\log p+y+o(1)$. Combing with (3.25) we get

$$\begin{aligned}\lim_{n\to\infty}\Pr\left(\frac{W_{pn}^2-a_{p,n}}{b_{p,n}}\geq x\right) &= \lim_{n\to\infty}\Pr\left(nT_{pn}\leq n\log(1-a_{p,n}-b_{p,n}x)\right)\\ &= \lim_{n\to\infty}\Pr\left(nT_{pn}\leq \log\log p-4\log p+y\right)\end{aligned}\tag{3.28}$$

As $p = p_n \to \infty$ as $n \to \infty$, we have

$$\begin{aligned}&\lim_{n\to\infty}\Pr\left(\frac{W_{pn}^2-a_{p,n}}{b_{p,n}}\geq x\right) = \lim_{n\to\infty,p\to\infty}\Pr\left(\frac{W_{pn}^2-a_{p,n}}{b_{p,n}}\geq x\right)\\ =&\lim_{n\to\infty}\lim_{p\to\infty}\Pr\left(\frac{W_{pn}^2-a_{p,n}}{b_{p,n}}\geq x\right)(\text{ as the convergence is uniform in } n)\\ =&1-\lim_{n\to\infty}G_n(x),\end{aligned}$$

where $G_n(x) = I(x\leq\frac{n-2}{2})\exp\big\{-\frac{1}{2}\big(1-\frac{2}{n-2}x\big)^{\frac{n-2}{2}}\big\}+I(x>\frac{n-2}{2})$.

Note that $1-\frac{2}{n-2}x = \exp\{\frac{1}{n}(y-\log 2\pi)\}$, plugging it into $G_n(x)$ yields

$$\begin{aligned}&\lim_{n\to\infty}G_n(x) = \lim_{n\to\infty}\exp\Big\{-\frac{1}{2}\exp\big\{\frac{n-2}{2n}(y-\log 2\pi)\big\}\Big\}\\ =&\exp\big\{-\frac{1}{\sqrt{8\pi}}\exp(\frac{1}{2}y)\big\}.\end{aligned}$$

Hence part (i) of Corollary 3 follows.

- **Exponential Case**

When $(\log p)/n \to \beta \in (0,\beta)$ as $n\to\infty$, we have

$$\begin{aligned}c_{p,n} &= \big(\frac{2}{n-2}\mathcal{B}(\frac{1}{2},\frac{n-2}{2})\sqrt{1-p^{-4/(n-2)}}\big)^{\frac{2}{n-2}}\\ &= \left(\frac{(n-2)\pi}{2}(1-e^{-4\beta})+o(1)\right)^{\frac{2}{n-2}}\\ &= \exp\left\{\frac{1}{n-2}\log\big(\frac{(n-2)\pi(1-e^{-4\beta})}{2}\big)+o(1)\big)\right\}\text{ for large enough } n.\end{aligned}$$

It follows that for large enough n,

$$
\begin{aligned}
n\log(c_{p,n}) &= \frac{n}{n-2}\log(n-2) + \log\big(\frac{\pi(1-e^{-4\beta})}{2}\big) + o(1) \\
&= \log\log p - \log\beta + \log\big(\frac{\pi(1-e^{-4\beta})}{2}\big) + o(1)
\end{aligned}
$$

Together with (3.26) we have

$$
\begin{aligned}
& n\log(1-a_{p,n}-b_{p,n}x) \\
=& \log\log p - \log\beta + \log\big(\frac{\pi(1-e^{-4\beta})}{2}\big) - \frac{4\log p}{n-2} + n\log(1-\frac{2}{n-2}x) \\
=& \log\log p - 4\log p - 8\beta + n\log(1-\frac{2}{n-2}x) + \log\big(\frac{\pi(1-e^{-4\beta})}{2\beta}\big) + o(1)
\end{aligned}
\tag{3.29}
$$

Let $y = -8\beta + n\log(1-\frac{2}{n-2}x) + \log\big(\frac{\pi(1-e^{-4\beta})}{2\beta}\big)$, then the RHS of (3.29) becomes $\log\log p - 4\log p + y + o(1)$. Again combing with (3.25), we can still get (3.28).

Moreover,

$$
\begin{aligned}
\lim_{n\to\infty} G_n(x) &= \lim_{n\to\infty} \exp\Big\{-\frac{1}{2}\exp\Big\{\frac{n-2}{2n}\Big(y+8\beta-\log\big(\frac{\pi(1-e^{-4\beta})}{2\beta}\big)\Big)\Big\}\Big\} \\
&= \exp\big\{-\big(\frac{\beta}{\pi(1-e^{-4\beta})}\big)^{1/2} e^{(y+8\beta)/2}\big\},
\end{aligned}
$$

which leads to the convergence result in part (ii).

- **Super-Exponential Case**

  If $\log p/n \to \infty$ as $n\to\infty$, then for large enough n,

$$
c_{p,n} = \big(\frac{2}{n-2}\mathcal{B}(\frac{1}{2},\frac{n-2}{2})\sqrt{1-p^{-4/(n-2)}}\big)^{\frac{2}{n-2}} = \exp\Big\{\frac{1}{n-2}\log\big(\frac{(n-2)\pi}{2}\big)\Big\}.
$$

  Combing with (3.26) we obtain

$$
\begin{aligned}
& n\log(1-a_{p,n}-b_{p,n}x) \\
=& -\frac{4n\log p}{n-2} + n\log(1-\frac{2}{n-2}x) + \frac{n}{n-2}\log 2\pi - \frac{n}{n-2}\log(n-2) + o(1) \\
=& -\frac{4n\log p}{n-2} + \log n + n\log(1-\frac{2}{n-2}x) + \log\frac{\pi}{2} + o(1).
\end{aligned}
\tag{3.30}
$$

Let $y = n\log(1-\frac{2}{n-2}x)+\log\frac{\pi}{2}$, then the RHS of (3.29) becomes $-\frac{4n\log p}{n-2}+\log n+y+o(1)$. Moreover,

$$\lim_{n\to\infty} G_n(x) =1 - \lim_{n\to\infty} \exp\Big\{ -\frac{1}{2}\exp\Big\{\frac{n-2}{2n}\big(y-\log\frac{\pi}{2}\big)\Big\}\Big\} = \exp\big\{-\frac{1}{\sqrt{2\pi}}e^{y/2}\big\}.$$

□

Proof of Theorem 3. Let event $A = \{R_{ij}^2 \leq 1 - p^{-(4+\delta)/(n-3)}$ for all $i,j \in \mathcal{M}\backslash\mathcal{M}^*\}$, event $B = \{\hat{\rho}_{ij} \leq f(n,p,\alpha)$ for $i \in \mathcal{M}^*, j \in \mathcal{M}\backslash\mathcal{M}^*\}$ where $\hat{\rho}_{ij} = |\mathrm{Corr}(\mathrm{X_i},\mathrm{X_j})|$, $f(n,p,\alpha)$ is the screening threshold for pairwise correlation screening. Then A implies that no pairs of unimportant variables passed the R squares screening. B implies that important and unimportant variables can not be too highly correlated.

By the definition of $\mathcal{C}$, we have

$$P(\mathcal{C}\cap\mathcal{M}\subset\mathcal{M}^*) \geq P(A\cap B) \geq P(A)+P(B)-1. \tag{3.31}$$

For the event A, we have

$$\begin{aligned}
P(A) &= 1 - P(\bigcup_{i\neq j\in\mathcal{M}\backslash\mathcal{M}^*} R_{ij}^2 \geq 1 - p^{-(4+\delta)/(n-3)})\\
&\geq 1 - \sum_{i\neq j\in\mathcal{M}\backslash\mathcal{M}^*} P(R_{ij}^2 \geq 1 - p^{-(4+\delta)/(n-3)})\\
&= 1 - (n/\log(n))^2 P(\mathrm{Beta}(1,\frac{n-3}{2}) \geq 1 - p^{-(4+\delta)/(n-3)})\\
&= 1 - (n/\log(n))^2 p^{-(4+\delta)/2}
\end{aligned}$$

Under the assumption $(B1)$, $(n/\log(n))^2p^{-(4+\delta)/2} \to 0$ as $n\to\infty$. Therefore we have $P(A)\to 1$.

Next we show that $P(B)\to 1$ as $n\to\infty$. We have

$$
\begin{aligned}
P(B) &= 1 - p(\bigcup_{i\in\mathcal{M}^*, j\in\mathcal{M}\setminus\mathcal{M}^*} \hat{\rho}_{ij} \geq f(n,p,\alpha)) \\
&\geq 1 - \sum_{i\in\mathcal{M}^*, j\in\mathcal{M}\setminus\mathcal{M}^*} P(\hat{\rho}_{ij} \geq f(n,p,\alpha)) \\
&= 1 - (n/\log(n))^2 \Pr\left(\hat{\rho}_{ij} \geq \max\{a_{p,n} + b_{p,n}F_n(\alpha), \eta\}\right) \\
&= 1 - (n/\log(n))^2 \Pr\left(\hat{\rho}_{ij} \geq \delta_{p,n}\right),
\end{aligned}
$$

where $F_n(\alpha)$ is the $100(1-\alpha)$ quantile of the limiting cumulative distribution function of the maximal pairwise correlation statistic, and we denote $\max\{a_{p,n} + b_{p,n}F_n(\alpha), \eta\}$ by $\delta_{p,n}$.

Note that

$$
\begin{aligned}
& a_{p,n} + b_{p,n}F_n(\alpha) = 1 - p^{-4/(n-2)}c_{p,n}(1 - \frac{2}{n-2}F_n(\alpha)) \\
=& 1 - p^{-4/(n-2)}c_{p,n}\{-2\log(1-\alpha)\}^{2/(n-2)} \\
=& 1 - \Big(C_\alpha p^{-2}\frac{n-2}{2}\mathcal{B}(\frac{1}{2}, \frac{n-2}{2})\sqrt{1-p^{-4/(n-2)}}\Big)^{\frac{2}{n-2}} \\
=& 1 - O\Big(\frac{C_\alpha^2(n-2)(1-p^{-4/(n-2)})}{p^4}\Big)^{\frac{1}{n-2}} \quad \text{for large enough } n \\
=& 1 - O\big(e^{-\frac{\log p}{n}}\big) \quad \text{for large enough } n
\end{aligned}
$$

Let ρ_{ij} be the population correlation coefficient between X_i and X_j. Write $z(n) = \frac{1}{2}\log\frac{1+\hat{\rho}_{ij}}{1-\hat{\rho}_{ij}}$, $\xi = \frac{1}{2}\log\frac{1+\rho_{ij}}{1-\rho_{ij}}$. It has been shown that as $n \to \infty$, $n^{1/2}(z(n) - \xi) \to \mathcal{N}(0,1)$.

We have

$$
\begin{aligned}
\Pr\left(\hat{\rho}_{ij} \geq \delta_{p,n}\right) =& \Pr\left(n^{1/2}(z(n)-\xi) \geq n^{1/2}(\frac{1}{2}\log\frac{1+\delta_{p,n}}{1-\delta_{p,n}} - \xi)\right) \\
=& \Pr\left(Z \geq n^{1/2}(\frac{1}{2}\log\frac{1+\delta_{p,n}}{1-\delta_{p,n}} - \xi) + o_n(1)\right) \\
\leq& \frac{e^{-C_{p,n}n}}{\sqrt{2\pi n}C_{p,n}},
\end{aligned}
\tag{3.32}
$$

where $C_{p,n} = \frac{1}{2}\log\frac{1+\delta_{p,n}}{1-\delta_{p,n}} - \xi$.

If $\log(p)/n \to \infty$ as $n \to \infty$, then $a_{p,n} + b_{p,n}F_n(\alpha) \to 1$. Therefore $\delta_{p,n} \to 1$, which yields $C_{p,n} \to \infty$. Then the tail probability in (3.32) goes to zero as $n \to \infty$. It follows that $P(B) \to 1$ as $n \to \infty$.

If $\log(p)/n \to \eta_0$ as $n \to \infty$, then $\delta_{p,n} \to \max\{1 - e^{-4\eta_0}, \eta\}$. Under assumption $(B2)$ that $\rho_{ij} < \max\{1 - e^{-4\eta_0}, \eta\}$, $\lim_{n\to\infty} C_{p,n} = \lim_{n\to\infty} \frac{1}{2}\log\frac{1+\max\{1-e^{-4\eta_0},\eta\}}{1-\max\{1-e^{-4\eta_0},\eta\}} - \xi > 0$. Again the tail probability in (3.32) goes to zero as $n \to \infty$. It follows that $P(B) \to 1$ as $n \to \infty$.

If $\log(p)/n \to 0$ as $n \to \infty$, then $a_{p,n} + b_{p,n}F_n(\alpha) \to 0$. Hence $\delta_{p,n} \to \eta$. Under the assumption $(B2)$, we have $\lim_{n\to\infty} C_{p,n} = \log\frac{1+\eta}{1-\eta} - \xi > 0$. Therefore $P(B) \to 1$ as $n \to \infty$.

Given $P(A) \to 1$ and $P(B) \to 1$, we have $P(\mathcal{C} \cap \mathcal{M} \subset \mathcal{M}^*) \to 1$. □

Proof of Theorem 4. By (3.17), as $n \to \infty$, the following inequality holds with probability tending to 1.

$$\|C_{21}C_{11}^{-1}\text{sign}(\boldsymbol{\beta}_1)\|_{\max} \leq 1 - \xi/2. \tag{3.33}$$

where $C_{11} = \begin{pmatrix} C_{11}^{(11)} & C_{11}^{(12)} \\ C_{11}^{(21)} & C_{11}^{(22)} \end{pmatrix}$, $C_{21} = \begin{pmatrix} C_{21}^{(1)} & C_{21}^{(2)} \end{pmatrix}$. It follows from (3.33) directly that

$$\|(C_{21}^{(2)} - C_{21}^{(1)}(C_{11}^{(11)})^{-1}C_{11}^{(12)})(C_{11}^{(22)} - C_{11}^{(21)}(C_{11}^{(11)})^{-1}C_{11}^{(12)})^{-1}\text{sign}(\boldsymbol{\beta}_1^{(2)})\|_{\max} \leq 1 - \xi, \tag{3.34}$$

where $\|\cdot\|_{\max}$ denotes the max norm of a matrix. Based the definition of $\mathcal{C}$, we have the following element wise inequalities $\|C_{11}^{(12)}\|_{\max} \leq c_{n,p,\alpha}$, $\|C_{11}^{(21)}\|_{\max} \leq c_{n,p,\alpha}$. Here $c_{n,p,\alpha}$ is the pairwise correlation screening bound. Since $C_{11}^{(11)}$ is positive definite, there exists an orthogonal matrix Q s.t. $C_{11}^{(11)} = Q\Lambda Q^T$, where Λ is a diagonal matrix consists of the eigenvalues of $C_{11}^{(11)}$. By assumption, we have $\lambda_{min}(C_{11}^{(11)}) \geq \lambda_0$. Therefore $\|C_{11}^{(21)}(C_{11}^{(11)})^{-1}C_{11}^{(12)}\|_{\max} \leq \lambda_0^{-1}c_{n,p,\alpha}^2 s_1^2$. Under the assumption that $\log(p)/n \to 0$, $c_{n,p,\alpha} = o_n(1)$. It follows that $\lambda_0^{-1}c_{n,p,\alpha}^2 s_1^2 = o_n(1)$. By assumption $(B2)$, $\|C_{21}^{(1)}\|_{\max} \leq \eta$. Thus $\|C_{21}^{(1)}(C_{11}^{(11)})^{-1}C_{11}^{(12)}\|_{\max} \leq \lambda_0^{-1}\eta c_{n,p,\alpha}s_1^2$, then

$\|C_{21}^{(1)}(C_{11}^{(11)})^{-1}C_{11}^{(12)}\|_{\max} = o_n(1)$. Therefore

$$\begin{aligned}
&\|(C_{21}^{(2)} - C_{21}^{(1)}(C_{11}^{(11)})^{-1}C_{11}^{(12)})\big(C_{11}^{(22)} - C_{11}^{(21)}(C_{11}^{(11)})^{-1}C_{11}^{(12)}\big)^{-1}\text{sign}(\boldsymbol{\beta}_1^{(2)}) \\
&- C_{21}^{(2)}(C_{11}^{(22)})^{-1}\text{sign}(\boldsymbol{\beta}_1^{(2)})\|_{\max} \\
=&\|\big(C_{21}^{(2)}(C_{11}^{(22)})^{-1}C_{11}^{(21)}(C_{11}^{(11)})^{-1}C_{11}^{(12)} - C_{21}^{(1)}(C_{11}^{(11)})^{-1}C_{11}^{(12)}\big)\big(C_{11}^{(22)} - C_{11}^{(21)}(C_{11}^{(11)})^{-1}C_{11}^{(12)}\big)^{-1}\cdot \\
&\text{sign}(\boldsymbol{\beta}_1^{(2)})\|_{\max}
\end{aligned}$$

Write $A = C_{21}^{(2)}(C_{11}^{(22)})^{-1}C_{11}^{(21)}(C_{11}^{(11)})^{-1}C_{11}^{(12)}, B = C_{21}^{(1)}(C_{11}^{(11)})^{-1}C_{11}^{(12)}, D = C_{11}^{(21)}(C_{11}^{(11)})^{-1}C_{11}^{(12)}$, and $Y = \text{sign}(\boldsymbol{\beta}_1^{(2)})$. Then the above term becomes $\|(A-B)(C_{11}^{(22)} - D)^{-1}Y\|_{\max}$. Moreover, we have

$$\|(A-B)(C_{11}^{(22)} - D)^{-1}Y\|_{\max} \leq (s-s_1)\|A-B\|_{\max}\|(C_{11}^{(22)} - D)^{-1}Y\|_{\max}.$$

Since $\|A\|_{\max} \leq \lambda_0^{-1}(s-s_0)^2\|C_{21}^{(2)}\|_{\max}\|C_{11}^{(21)}(C_{11}^{(11)})^{-1}C_{11}^{(12)}\|_{\max} \leq \lambda_0^{-2}\eta c_{n,p,\alpha}^2 s_1^2(s-s_1)^2$, $\|B\|_{\max} \leq \lambda_0^{-1}\eta c_{n,p,\alpha}s_1^2$, and

$$\|(C_{11}^{(22)} - D)^{-1}Y\|_{\max} \leq (s-s_1)\|(C_{11}^{(22)} - D)^{-1}\|_{\max} \leq (s-s_1)(\lambda_0 - \lambda_0^{-1}c_{n,p,\alpha}^2 s_1^2)^{-1}.$$

Therefore we have

$$\begin{aligned}
&\|(A-B)(C_{11}^{(22)} - D)^{-1}Y\|_{\max} \\
\leq&(s-s_0)^2(\lambda_0^{-2}\eta c_{n,p,\alpha}^2 s_1^2(s-s_1)^2 + \lambda_0^{-1}\eta c_{n,p,\alpha}s_1^2)(\lambda_0 - \lambda_0^{-1}c_{n,p,\alpha}^2 s_1^2)^{-1} \\
\to&0,
\end{aligned}$$

as $n \to \infty$. It follows that $C_{21}^{(2)}(C_{11}^{(22)})^{-1}\text{sign}(\boldsymbol{\beta}_1^{(2)}) \leq 1 - \xi/3$, which concludes the proof if we take $\delta = \xi/3$. □

CHAPTER 4
Multivariate Response regression

4.1 Introduction

In many real applications, it is often to have more than one variables that are associated with a common set of response variables. For instance, in financial econometrics, one may need to predict the asset returns based on the historical data, where multiple-response regression problems arise.

Consider the following multiple-response regression:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{e},$$

where $\mathbf{Y}$ is the $n \times m$ response matrix, $\mathbf{X}$ is the $n \times p$ design matrix, $\mathbf{B}$ is the $p \times m$ unknown parameter matrix, and $\mathbf{e} = [\varepsilon_1, \ldots, \varepsilon_n]^T$ is the error matrix. Our goal is to estimate $\mathbf{B}$ to obtain predictions for the response vector based on a common list of predictors. An intuitive estimation of $\mathbf{B}$ is to solve the following optimization problem:

$$\min_{\mathbf{B}} \operatorname{tr}\{(\mathbf{Y} - \mathbf{X}\mathbf{B})^T(\mathbf{Y} - \mathbf{X}\mathbf{B})\}. \tag{4.1}$$

The solution to (8) is in fact equivalent to solving ordinary least squares separately on each of the response variables. This method, however, does not take into account the underlying correlations among the responses. Suppose the error vectors $\varepsilon_1, \cdots, \varepsilon_n$ are i.i.d. and have a multivariate distribution with an unknown covariance matrix Σ. Denote $\Omega = \Sigma^{-1}$. In order to take advantage of the information of Σ, one can add a penalty term, such as $\lambda(\mathbf{B})$, to (4.1). In the literature, a number of existing papers have been focused on penalized multiple-response regression. For instance, Turlach et al. (2005) proposed the L_∞-penalized simultaneous variable selection method (L_∞-SVS) using $\lambda(\mathbf{B}) = \lambda \sum_{i=1}^{p} \|\mathbf{B}^i\|_\infty$, where $\|\mathbf{B}^i\|_\infty$ is the infinity norm of the ith row of $\mathbf{B}$. The motivation of using L_∞ is that a variable should be selected if it is associated with at least one response variable. Similä and Tikka (2007) proposed a similar method, known as L_2-SVS, where they use the L_2 penalty for $\mathbf{B}^i$, which measures the overall effect of the ith covariate on the responses. Peng et al. (2010) introduced a combined penalty function $\lambda(\mathbf{B}) = \lambda_1 \sum_{i,j} |b_{i,j}| + \lambda_2 \sum_{i=1}^{p} \|\mathbf{B}^i\|_2$, where the first part encourages the sparsity of $\mathbf{B}$, while the second part requires that the rows of $\mathbf{B}$ to be zeros or nonzeros simultaneously.

Instead of using penalization, another type of methods focus on identifying the low-rank structure of the parameter matrix $\mathbf{B}$. Such an idea is closely related to factor analysis. That is, the effects of the covariates are from a few underlying factors. Therefore, a rank constraint is imposed on $\mathbf{B}$. One example is the factor estimation and selection (FES) method introduced by Yuan et al. (2007), where they impose the sparsity constraint on the singular values of $\mathbf{B}$, and thus reduce the rank of $\mathbf{B}$. A similar approach is the sparse reduced-rank regression (SRRR) investigated by Chen and Huang (2012). They decompose the coefficient matrix $\mathbf{B}$ to the multiplication of two low-rank matrices, and add a sparsity constraint to one of the matrices.

In this chapter, we introduce a weighted simultaneous variable selection (WSVS) method for multiple-response regression. Our method is motivated by the extreme behavior of the maximal absolute correlation between each of the covariates and all the responses. Based on the asymptotic distribution, we can calculate p-values for each of the covariates. We further propose a weighted L_2 penalty that utilizes the p-values and design the WSVS estimate for the coefficient matrix.

The rest of this chapter is organized as follows. In Section 4.2, we introduce the modeling framework for multiple-response regression problems and the assumptions required for our method. We investigate the theoretical results in Section 4.2.1 and propose a weighted penalty based screening in Section 4.2.2. Then we demonstrate that our proposed method has competitive performance using numerical examples in Section 4.4. We conclude this Chapter in Section 4.5.

4.2 Weighted simultaneous variable selection

Consider a multivariate-response regression model

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{e}. \tag{4.2}$$

In (4.2), the response matrix is $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T$, where $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})^T$ is the m-dimensional response vector for the ith instance. In addition, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$ is the $n \times p$ design matrix with $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ being the p-dimensional predictors. Let $\mathbf{y}^k = (y_{1k}, \ldots, y_{nk})^T$ be the kth response vector, and $\mathbf{x}^j = (x_{1j}, \ldots, x_{nj})^T$ be the jth predictor. Finally, $\mathbf{e} = [\varepsilon_1, \ldots, \varepsilon_n]^T$ denotes the error matrix and assume that ε_i's are i.i.d. $\mathcal{N}(0, \Sigma)$ with Σ being the unknown covariance matrix.

4.2.1 Extreme value theories

In this section, we investigate the extreme behavior of the pairwise correlations between $\{\mathbf{x}^j : j = 1, \ldots, p\}$ and $\{\mathbf{y}^k, k = 1, \ldots, m\}$. Let $R(\mathbf{x}^j) = \max_{1 \le j \le m} |\widehat{\mathrm{Corr}}(\mathbf{x}^j, \mathbf{y}^k)|$ be the maximal absolute sample correlation between $\mathbf{x}^j$ and $\mathbf{y}^k$. We further define

$$R_{XY} = \min_{1 \le j \le p} R(\mathbf{x}^j). \tag{4.3}$$

Then when Σ is the identity matrix, that is, the random errors are independent, the magnitude of R_{XY} can not be too large. In particular, we study the asymptotic distribution of R_{XY} as m increases and state it in the following theorem.

Theorem 8. *Suppose we observe a random sample of size n from the linear model* (4.2) *and we further assume that ε^j's are independent. Let R_{XY} be defined as in* (4.3). *Define*

$$a_{m,n} = 1 - m^{-2/(n-2)} c_{m,n}, \quad b_{m,n} = \frac{2}{n-2} m^{-2/(n-2)} c_{m,n},$$

where $c_{m,n} = \{2^{-1}(n-2)\mathcal{B}(\frac{1}{2}, \frac{n-2}{2})\sqrt{1-m^{-2/(n-2)}}\}^{2/(n-2)}$ *is a correction factor with* $\mathcal{B}(s,t)$ *being the Beta function. Then under the null hypothesis* (2.3), *for all* $x \in \mathbb{R}$,

$$\lim_{m\to\infty} \sup_{n\geq 3} \left| \Pr\left\{ \frac{R_{XY}^2 - a_{m,n}}{b_{m,n}} < x \right\} - F_{n,p}(x) \right| = 0,$$

where

$$F_{n,p}(x) = \left(1 - \exp\left\{ -\left(1 - \frac{2}{n-2}x\right)^{(n-2)/2} \right\}\right)^p \mathbf{1}\left(x \leq \frac{n-2}{2}\right) + \mathbf{1}\left(x > \frac{n-2}{2}\right). \quad (4.4)$$

The above theorem follows directly from Theorem 1 by the properties of order statistics for i.i.d. random variables. Note that the results hold uniformly for the sample size $n \geq 3$, which means finite sample performance is guaranteed.

To better demonstrate the asymptotic distribution of R_{XY}, we illustrate the limiting distribution function $F_{n,p}$ for $n = 200$, and $p = 10$ or $p = 100$ in Figure 4.1. We also evaluate the accuracy of the extreme value results. In particular, we simulate 1000 independent samples from i.i.d. $\mathbf{x}^j$'s and $\mathbf{y}^k$'s and calculate R_{XY} when $n = 200$, $p = 10$, and $m = 10$ or $m = 100$, and compare $F_{n,p}(\frac{R_{XY}^2 - a_{m,n}}{b_{m,n}})$ with the uniform distribution on $[0,1]$. The Q-Q plots are displayed in Figure 4.2. One can see that the Q-Q plots almost fall onto the straight line $y = x$, which implies that the asymptotic distribution is accurate. Moreover, as expected, the discrepancy becomes smaller as m increases.

Given the covariate vector $\mathbf{x}^j$, we can calculate the maximal absolute correlation between this covariate and the responses, denoted as $R_{\mathbf{x}^j}$. We can further calculate p-value by

$$p(\mathbf{x}^j) = 1 - F_{n,p}(\frac{R_{\mathbf{x}^j}^2 - a_{m,n}}{b_{m,n}}). \quad (4.5)$$

A larger p-value indicates that $\mathbf{x}^j$ has a stronger effect on the responses. Motivated by that, we can assign different weights on the penalty $\lambda(\mathbf{B}^j)$. In the following section, we introduce a weighted simultaneous variable selection (WSVS) approach based on the p-values.

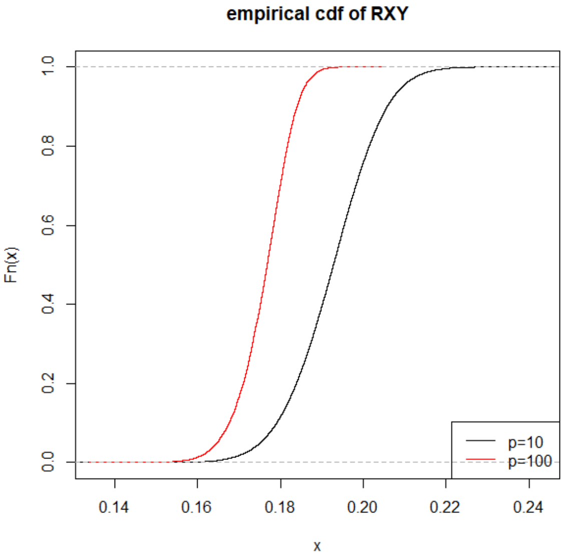
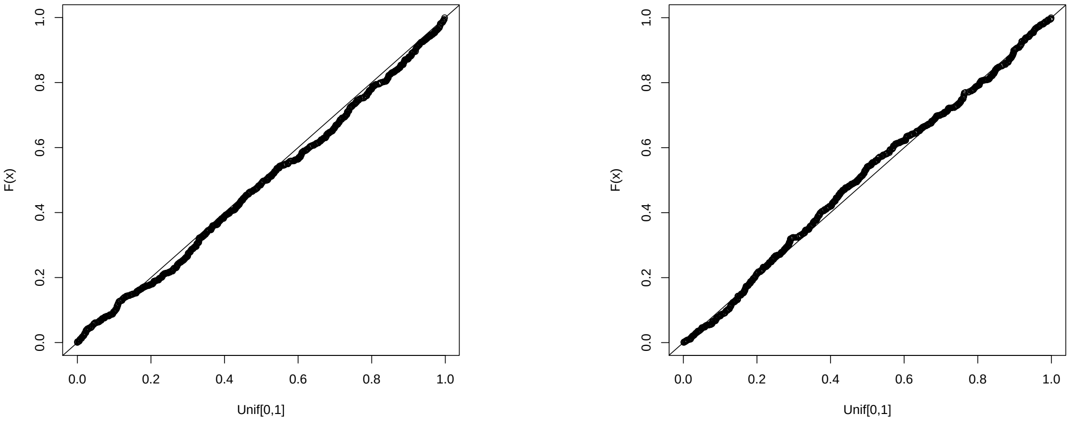


Figure 4.1: Asymptotic cumulative distribution function of R_{XY}



(a) $m = 10$

(b) $m = 100$

Figure 4.2: Q-Q plots of $F_{n,p}(\frac{R_{XY}^2 - a_{m,n}}{b_{m,n}})$ against U[0,1] when (a) $m = 10$; (b) $m = 100$.

4.2.2 Weighted L_2 penalization

In Section 4.1, we introduced the L_2-SVS method proposed by Similä and Tikka (2007), where they suggested an estimator of the coefficient matrix by solving the following optimization problem:

$$\min_{\mathbf{B}} \frac{1}{2}\text{tr}\{(\mathbf{Y}-\mathbf{X}\mathbf{B})^T(\mathbf{Y}-\mathbf{X}\mathbf{B})\} + \lambda \sum_{i=1}^{p} \|\mathbf{B}^i\|_2. \tag{4.6}$$

By introducing the L_2 penalty, the rows of $\mathbf{B}$ will be zeros or non-zeros simultaneously. If the ith row contains all zeros, it implies that the ith covariate has no effect on the responses. One drawback of L_2SVS is that it imposes the same level of penalty to all covariates and does not take into account the individual effect. To address this issue, we propose the following weighted simultaneous variable selection (WSVS) method:

$$\min_{\mathbf{B}} \frac{1}{2}\text{tr}\{(\mathbf{Y}-\mathbf{X}\mathbf{B})^T(\mathbf{Y}-\mathbf{X}\mathbf{B})\} + \lambda \sum_{i=1}^{p} w_i\|\mathbf{B}^i\|_2, \tag{4.7}$$

where $w_i = \exp(p(\mathbf{x}^i))$ with $p(\mathbf{x}^i)$ as defined in (4.5). The rationality of such weighting is that if the p-value $p(\mathbf{x}^i)$ is smaller, it is more likely that the covariate $\mathbf{x}^i$ is an important variable. Therefore, less shrinkage should be assigned to $\mathbf{x}^i$. We take the exponential of the p-values to ensure w_i's are positive and their magnitude falls into a reasonable range.

4.3 Computational algorithm and model selection

In this section, we introduce the computational algorithm to solve the WSVS estimator. We apply the group descent algorithm described in Simon et al. (2013). The optimization problem (4.7) is equivalent to solving

$$\min_{\mathbf{B}} \frac{1}{2} \|\mathbf{Y}-\mathbf{X}\mathbf{B}\|_F^2 + \lambda \sum_{i=1}^{p} w_i \left\|\mathbf{B}^i\right\|_2 \tag{4.8}$$

For problem (4.8), we can decompose $\mathbf{B}$ as $\mathbf{B}^k$, $k = 1, \dots, p$. We can further view the objective function in (4.8) as a function of $\mathbf{B}^k$ with fixed $\mathbf{B}^j$ for all $j \neq k$. Then solving (4.8) is equivalent to

$$\min_{\mathbf{B}^k} \frac{1}{2}\|\mathbf{Y}_{-k} - \mathbf{x}^k\mathbf{B}^k\|_F^2 + \lambda w_k \|\mathbf{B}^k\|_2$$

where $\mathbf{x}^k$ denotes the kth column of $\mathbf{X}$, and $\mathbf{Y}_{-k} = \mathbf{Y} - \sum_{j\neq k}\mathbf{x}^j\mathbf{B}^j$ is the residual of removing the effects of $\mathbf{x}^j, j \neq k$. If we solve the above optimization problem, then we get that the solution $\hat{\mathbf{B}}^k$ satisfies

$$\hat{\mathbf{B}}^k = \frac{1}{\|\mathbf{x}^k\|_2^2}\left(1 - \frac{\lambda w_k}{\|(\mathbf{x}^k)^T\mathbf{Y}_{-k}\|_2}\right)_+ (\mathbf{x}^k)^T\mathbf{Y}_{-k} \tag{4.9}$$

where $(a)_+ = \max(0, a)$. By applying these updates sequentially, we can obtain the global solution to (4.8) due to the convexity of the objective function.

Our proposed WSVS algorithm can be summarized as follows.

WSVS Algorithm:

1. Start with $\mathbf{B}_0^k = \mathbf{0}$

2. Define $\mathbf{Y}_{-k} = \mathbf{Y} - \sum_{j\neq k}\mathbf{x}^j\mathbf{B}^j$, then update $\mathbf{B}^k$ by

$$\mathbf{B}^k \leftarrow \frac{1}{\|\mathbf{x}^k\|_2^2}\left(1 - \frac{\lambda w_k}{\|(\mathbf{x}^k)^T\mathbf{Y}_{-k}\|_2}\right)_+ (\mathbf{x}^k)^T\mathbf{Y}_{-k}$$

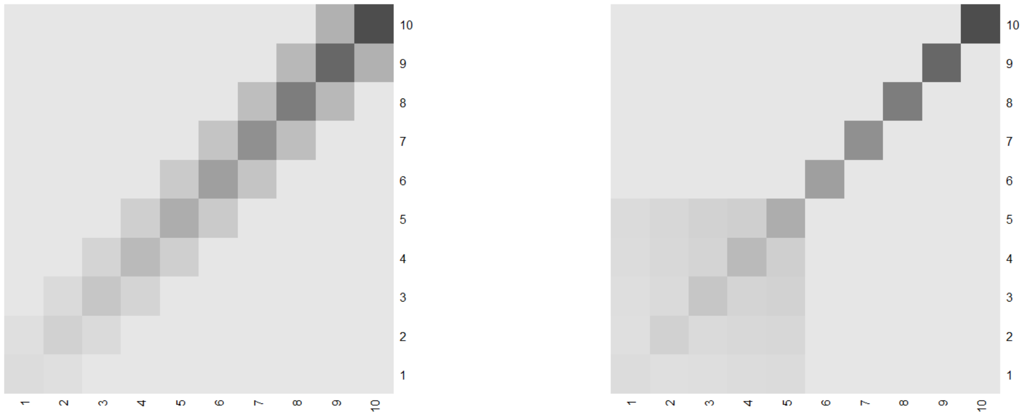
3. Iterate Step (2) over $k = 1, \ldots, p$ until convergence.

For the tuning parameter λ, it can be selected by either using an independent validation set or through K-fold cross-validation (CV). In K-fold CV, we randomly spit the training data into K parts of equal sizes. For each of the K folds, we obtain the estimated $\hat{\mathbf{B}}_\lambda$ using data in all of the remaining folds and evaluate the CV errors. Then we find λ such that the sum of CV errors is minimized.

4.4 Numerical studies

We demonstrate using numerical examples to show that our method has competitive performance compared with several existing methods in this section. We compare with the curds and whey (CW) method introduced by Breiman and Friedman (1997). We use the CW with generalized cross validation (CW-GCV) when $p < n$ and the CW with the ridge regression (CW-RR) when $p \geq n$. We also compare with the separate ridge regression (Ridge) and the separate LASSO.

For all simulated examples, we simulate i.i.d. predictor vector $\mathbf{x}_i$'s from $\mathcal{N}(0, I_p)$ for $i = 1, \ldots, n$, where I_p is the $p \times p$ identify matrix. In other words, the p covariates are all



(a) Example 1

(b) Example 2

Figure 4.3: Plots of the inverse covariance structure for Examples 1 and 2.

independent. For each example, we consider $n = 40, 100, 200$, $p = 20, 40$ and $m = 10$. The coefficient matrix $\mathbf{B}$ is set as $\beta_{j,k} = 0$ for all $j \geq 13$. For $j = 1, \ldots, 10$, $\beta_{j,j} = 3, \beta_{j,j+1} = 4, \beta_{j,j+2} = 3$ and $\beta_{j,k} = 0.5$ otherwise.

We consider two different configurations of the Ω matrix.

- **Example 1**: $\omega_{i,i} = i/5, \omega_{i+1,i} = (i(i+1)/100)^{1/2} (i = 1, \ldots, m-1)$, and $\omega_{i,i} = 0$ otherwise.
- **Example 2**: $\omega_{i,i} = i/5, \omega_{i,j} = (ij/100)^{1/2} (i \neq j, i, j \leq 5)$, and $\omega_{i,i} = 0$ otherwise.

We show the inverse covariance structure for Examples 1 and 2 in Figure 4.3. As shown in the plots, the structure of the inverse covariance in Example 1 is banded, and Example 2 has a nonzero block on the lower left corner.

For model selection, we simulate an independent validation dataset of size n to select the tuning parameter λ. Then we construct the final estimator with the selected λ on the training data. We also generate a testing dataset of size 400 to evaluate the out of sample mean squared prediction errors (MSPE), which is calculated by

$$\text{MSPE} = \text{tr}\{(\mathbf{Y} - \hat{\mathbf{Y}})^T (\mathbf{Y} - \hat{\mathbf{Y}})\}.$$

$p = 20$			
CW	1.43 (0.009)	1.13 (0.003)	1.06 (0.002)
RR	1.71 (0.023)	1.17 (0.005)	1.08 (0.002)
LASSO	1.72 (0.024)	1.21 (0.007)	1.10 (0.003)
WSVS	1.54 (0.021)	1.12 (0.005)	1.06 (0.002)
Oracle	1.00 (0.003)	1.00 (0.003)	1.00 (0.002)
$p = 40$			
CW	4.29 (0.059)	1.25 (0.004)	1.11 (0.002)
RR	3.64 (0.048)	1.45 (0.010)	1.15 (0.003)
LASSO	2.32 (0.035)	1.45 (0.011)	1.18 (0.004)
WSVS	2.39 (0.032)	1.24 (0.010)	1.12 (0.003)
Oracle	1.00 (0.003)	1.00 (0.002)	1.00 (0.001)

Table 4.1: Averages of RMSE and standard erros out of 100 replications (The standard errors are given in the parentheses).

$p = 20$			
CW	1.46 (0.009)	1.13 (0.003)	1.06 (0.002)
RR	1.45 (0.010)	1.13 (0.004)	1.06 (0.002)
LASSO	1.44 (0.010)	1.13 (0.003)	1.06 (0.002)
WSVS	1.53 (0.020)	1.11 (0.005)	1.06 (0.002)
Oracle	1.00 (0.003)	1.00 (0.003)	1.00 (0.002)
$p = 40$			
CW	3.80 (0.069)	1.28 (0.004)	1.12 (0.002)
RR	3.24 (0.062)	1.31 (0.004)	1.12 (0.002)
LASSO	1.96 (0.032)	1.25 (0.005)	1.10 (0.002)
WSVS	2.21 (0.030)	1.21 (0.007)	1.11 (0.003)
Oracle	1.00 (0.003)	0.99 (0.003)	1.00 (0.001)

Table 4.2: Averages of RMSE and standard erros out of 100 replications (The standard errors are given in the parentheses).

The results for the simulated examples 1 and 2 are displayed in Tables 4.1 and 4.2. One can see that our WSVS method has better performance than all the other competitors in most settings. In particular, in Example 2 with $p = 20$, all methods have similar performance. This is possibly due to the fact that the inverse covariance matrix is close to diagonal in this example. Moreover, as expected, when the sample size n increases, all methods have smaller RSPEs.

4.5 Discussion

In this chapter, we propose a new penalized variable selection approach for multiple-response regression. Our proposed method takes advantage of the extreme value theory of the maximal absolute correlation between the covariates and the response vector. We further construct a weighted penalty based on the p-values. Numerical studies demonstrate that our proposed method performs well in practice.

BIBLIOGRAPHY

Ehud Aharoni and Saharon Rosset. Generalized α-investing: definitions, optimality results and application to public databases. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):771–794, 2014.

Theodore Wilbur Anderson. *An introduction to multivariate statistical analysis.* Wiley-Interscience, Hoboken, N.J., 3 edition, 2003.

Richard Arratia, Larry Goldstein, and Louis Gordon. Two moments suffice for poisson approximations: the chen-stein method. *The Annals of Probability*, pages 9–25, 1989.

Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

Howard D Bondell and Brian J Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123, 2008.

Patrick Breheny and Jian Huang. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and computing*, 25(2):173–187, 2015.

Leo Breiman and Jerome H Friedman. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59 (1):3–54, 1997.

Peter Bühlmann, Markus Kalisch, and Lukas Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1:255–278, 2014.

Tony Cai and Tiefeng Jiang. Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *The Annals of Statistics*, 39(3):1496–1525, 2011.

Tony Cai and Tiefeng Jiang. Phase transition in limiting distributions of coherence of high-dimensional random matrices. *Journal of Multivariate Analysis*, 107:24–39, 2012.

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

Lisha Chen and Jianhua Z Huang. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500): 1533–1545, 2012.

Shelley Derksen and HJ Keselman. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45(2):265–282, 1992.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

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

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

Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101, 2010.

Jianqing Fan, Fang Han, and Han Liu. Challenges of big data analysis. *National science review*, 1(2):293–314, 2014.

Jianqing Fan, Qi-Man Shao, and Wen-Xin Zhou. Are discoveries spurious? distributions of maximum spurious correlations and their applications. *arXiv preprint arXiv:1502.04237*, 2015.

William Fithian, Jonathan Taylor, Robert Tibshirani, and Ryan Tibshirani. Selective sequential model selection. *arXiv preprint arXiv:1512.02565*, 2015.

Dean P Foster and Robert A Stine. α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):429–444, 2008.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

George M Furnival and Robert W Wilson. Regressions by leaps and bounds. *Technometrics*, 42(1):69–79, 2000.

János Galambos. *The asymptotic theory of extreme order statistics*. Wiley, New York, 1978.

Jelle J Goeman, Sara A Van De Geer, and Hans C Van Houwelingen. Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):477–493, 2006.

Eitan Greenshtein et al. Best subset selection, persistence in high-dimensional statistical learning and optimization under l1 constraint. *The Annals of Statistics*, 34(5):2367–2386, 2006.

Max Grazier G'Sell, Stefan Wager, Alexandra Chouldechova, and Robert Tibshirani. Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(2):423–444, 2016.

Peter Hall and Hugh Miller. Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, 2012.

Fang Han and Han Liu. Distribution-free tests of independence with applications to testing more structures. *arXiv preprint arXiv:1410.4179*, 2014.

Ronald R Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49, 1976.

Barry James, Kang James, and Yongcheng Qi. Limit distribution of the sum and maximum from multivariate gaussian sequences. *Journal of multivariate analysis*, 98(3):517–532, 2007.

Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378, 2000.

Caiyan Li and Hongzhe Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.

Runze Li, Wei Zhong, and Liping Zhu. Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139, 2012.

Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of statistics*, 42(2):413, 2014.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462, 2006.

Robb J Muirhead. *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons, 2009.

Wei Pan, Benhuai Xie, and Xiaotong Shen. Incorporating predictor network in penalized regression with application to microarray data. *Biometrics*, 66(2):474–484, 2010.

Mee Young Park, Trevor Hastie, and Robert Tibshirani. Averaged gene expressions for regression. *Biostatistics*, 8(2):212–227, 2007.

Jie Peng, Ji Zhu, Anna Bergamaschi, Wonshik Han, Dong-Young Noh, Jonathan R Pollack, and Pei Wang. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The annals of applied statistics*, 4(1):53, 2010.

Mark R Segal, Kam D Dahlquist, and Bruce R Conklin. Regression approaches for microarray data analysis. *Journal of Computational Biology*, 10(6):961–980, 2003.

Dhruv B Sharma, Howard D Bondell, and Hao Helen Zhang. Consistent group identification and variable selection in regression with correlated predictors. *Journal of Computational and Graphical Statistics*, 22(2):319–340, 2013.

Yiyuan She et al. Sparse regression with exact clustering. *Electronic Journal of Statistics*, 4: 1055–1096, 2010.

Timo Similä and Jarkko Tikka. Input selection and shrinkage in multiresponse linear regression. *Computational Statistics & Data Analysis*, 52(1):406–422, 2007.

Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for coxs proportional hazards model via coordinate descent. *Journal of statistical software*, 39 (5):1, 2011.

Noah Simon, Jerome Friedman, and Trevor Hastie. A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. *arXiv preprint arXiv:1311.6529*, 2013.

Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

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

Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.

Berwin A Turlach, William N Venables, and Stephen J Wright. Simultaneous variable selection. *Technometrics*, 47(3):349–363, 2005.

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

Tong Tong Wu and Kenneth Lange. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, pages 224–244, 2008.

Guan Yu and Yufeng Liu. Sparse regression incorporating graphical structure among predictors. *Journal of the American Statistical Association*, 111(514):707–720, 2016.

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

Ming Yuan, Ali Ekici, Zhaosong Lu, and Renato Monteiro. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):329–346, 2007.

Kai Zhang. Spherical cap packing asymptotics and rank-extreme detection. *IEEE Transactions on Information Theory*, 63(7):4572–4584, 2017.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(Nov):2541–2563, 2006.

Ping-Shou Zhong and Song Xi Chen. Tests for high-dimensional regression coefficients with factorial designs. *Journal of the American Statistical Association*, 106(493):260–274, 2011.

Li-Ping Zhu, Lexin Li, Runze Li, and Li-Xing Zhu. Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 2012.

Yunzhang Zhu, Xiaotong Shen, and Wei Pan. Simultaneous grouping pursuit and feature selection over an undirected graph. *Journal of the American Statistical Association*, 108 (502):713–725, 2013.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.