Statistica Sinica Preprint No: SS-2018-0170	
Title	Penalized linear regression with high-dimensional pairwise screening
Manuscript ID	SS-2018-0170
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0170
Complete List of Authors	Siliang Gong Kai Zhang and Yufeng Liu
Corresponding Author	Siliang Gong
E-mail	siliang@pennmedicine.upenn.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

Penalized linear regression with high-dimensional pairwise screening

Siliang Gong, Kai Zhang and Yufeng Liu

University of Pennsylvania

and The University of North Carolina at Chapel Hill

Abstract: In relation to variable selection, most existing screening methods focus on marginal effects and ignore the dependence between covariates. To improve the performance of variable selection, we incorporate pairwise effects in covariates for screening and penalization. We achieve this by studying the asymptotic distribution of the maximal absolute pairwise sample correlation between independent covariates. The novelty of the theory is that the convergence is related to the dimensionality p, and is uniform with respect to the sample size n. Moreover, we obtain an upper bound for the maximal pairwise R squared when regressing the response onto two covariates. Based on these extreme-value results, we propose a screening procedure to detect covariates pairs that are potentially correlated and associated with the response. We further combine the pairwise screening with sure independence screening and develop a new regularized variable selection procedure. Numerical studies show that our method is competitive in terms of both prediction accuracy and variable selection accuracy.

Key words and phrases:Pairwise Screening, Penalized Regression, Sure Independence Screening, Variable Selection

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

1. Introduction

With the growing prevalence of big data, high-dimensional problems are becoming increasingly commonplace 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 microarray data sets based on only a few hundred patients, and in neuroscience, fMRI images may contain millions of voxels.

Many recent studies have focused on how to handle high-dimensional data analyses. Of the methods proposed, the penalized least squares plays an important role. One of the most well-known methods is the LASSO, proposed by Tibshirani [1996], which is the solution to the following penalized problem:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + \lambda P(\boldsymbol{\beta}), \tag{1.1}$$

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 zero. Despite its good theoretical properties and practical performance, the LASSO has two major drawbacks. First, it may over-shrink the estimates, causing significant bias. Second, in the case of a group of highly correlated variables, the LASSO tends to select only one of them. To address these issues, Zou and Hastie [2005] introduced the elastic net method, which uses $\lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2$ as the regularization term in (1.1), thus encouraging a grouping effect. Furthermore, various other penalized variable selection methods have been proposed as extensions to the LASSO, including the Dantzig selector [Candès and

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

Tao, 2007] and the smoothly clipped absolute deviation (SCAD) penalty [Fan and Li, 2001], among many others; see Hastie et al. [2003] and Fan and Lv [2010] for a comprehensive overview.

In high-dimensional variable selection, it is crucial that we account for the dependency structure of the covariates. Such information improves the accuracy of selection and provides practical insights. For instance, in gene expression data, rather than working independently, genes usually function as biological pathways. However, classical penalized variable selection methods usually do not explicitly consider the relationships between covariates. To address this problem, Yuan and Lin [2006] proposed the group LASSO method, which takes advantage of the grouping of the covariates. Extension to the group LASSO include, but are not limited to Breheny and Huang [2015]. Other methods use the structure information as a predictor graph (see Li and Li [2008], Pan et al. [2010], Zhu et al. [2013], Yu and Liu [2016], among others).

A common assumption in the aforementioned methods is that the underlying predictor graph is given, which may not hold in practice. When prior information is not available, clustering can be used to improve regression performance. Specifically, Park et al. [2007] proposed performing hierarchical clustering on the covariates, and then using the cluster averages as new predictors for the regression. Other methods use 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 method aims to make correlated covariates have similar regression coefficients [She, 2010]. Nevertheless, a large sample correlation between two variables does not necessarily indicate that they are

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

dependent, in the population sense. When the dimensionality continues to increase, the maximal pairwise correlation between p independent covariates can be close to one [Fan and Lv, 2010]. Therefore, it is important to identify covariates that are truly correlated, and to incorporate such information into the variable selection procedures.

In this study, we examine the limiting behavior of the maximal absolute pairwise sample correlation between covariates when they are independent Gaussian random variables. In contrast to prior works, we investigate the limiting distribution as the dimensionality p diverges. Therefore, the proposed asymptotic results can potentially be applied to data sets with arbitrarily large dimensionality. We also discuss the extreme behavior of the maximal absolute Spearman rho statistic for covariates with general distributions, and obtain the upper bound of the maximal pairwise R squared when regressing the response onto pairs of covariates. Using the extreme-value results, we formulate a screening procedure to identify covariate pairs that are potentially dependent and associated with the response. We further combine the pairwise screening with 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 paper is organized as follows. We first investigate the limiting distribution of the maximal pairwise sample correlation between covariates in Section 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 squared in Section

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

2.2. In Section 3.1, we formulate our proposed variable selection approach as a penalized maximum likelihood problem, and discuss potential extensions of our method in Section 3.2. Theoretical properties are discussed in Section 4. In Section 5, we use simulated experiments and two real data sets to show that the proposed method exhibits improved performance when important variables are highly correlated. Finally, we conclude this paper and discuss possible future work in Section 6. Proofs of the theoretical results are provided in the Appendix.

2. Pair screening for covariates

Suppose we have the following linear model:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.2}$$

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 independent and identically distributed (i.i.d.) random variables with mean zero and standard deviation σ, and is independent of the covariate vector $\mathbf{x}$.

For the linear model given in (2.2), variable selection methods aim to identify the nonzero components of $\boldsymbol{\beta}$, in other words, the important variables among all candidate predictors. In particular, if two covariates have a large pairwise correlation, we may want to include or

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

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 between the covariates in order to screen the covariate pairs. In the following subsection, we discuss the asymptotic results that generate the screening rule.

2.1 Extreme laws of pairwise sample correlation between covariates

We propose choosing a bound based on the extreme laws of pairwise sample correlations when the p covariates are independent. Our investigations are under two settings: (a) the covariates are normally distributed; (b) the covariates are nonGaussian random variables.

2.1.1 Gaussian covariates

A recent study shows 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 this result, we find that the maximal absolute pairwise sample correlation between p independent covariates also has a limiting distribution, as stated in the following theorem.

Theorem 1. *Suppose $X_1, X_2, \cdots, X_p$ are p independent Gaussian variables, and we observe n independent samples from each X_j. Let $W_{pn} = \max_{1\le i<j\le p} |\rho_{i,j}|$, where $\rho_{i,j} = \widehat{Corr}(X_i, X_j)$*

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

is the Pearson sample correlation between X_i and X_j. Then, as $p \to \infty$,

$$\lim_{p\to\infty} |P(\frac{W_{pn}^2 - a_{p,n}}{b_{p,n}} \le x) - I(x \le \frac{n-2}{2}) \exp\big\{ -\frac{1}{2}\big(1 - \frac{2}{n-2}x\big)^{\frac{n-2}{2}}\big\} - I(x > \frac{n-2}{2})| = 0, \tag{2.3}$$

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 $c_{p,n} = \big(\frac{n-2}{2}B(\frac{1}{2}, \frac{n-2}{2})\sqrt{1-p^{-4/(n-2)}}\big)^{2/(n-2)}$ are the normalizing constants.

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 et al. [2011] studied the asymptotic distribution under certain regularity conditions, and applied the results to test a covariance matrix. Cai and Jiang [2012] obtained the limiting laws of the coherence for different divergence rates of p with respect to n, and summarized the results as phase-transition phenomena. Our result unifies the convergence in terms of the sample size, and includes the results of Cai and Jiang [2012] as special cases, as described in the following corollary.

Corollary 1. *Let W_{pn} be defined as in Theorem 1, where X_j 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 $n \to \infty$,*

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

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

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

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

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

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

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

Compared with those of previous works, our asymptotic distribution is novel in two respects. First, the convergence in Theorem 1 is with respect to p, not n, making it applicable to high-dimensional data, or even ultrahigh-dimensional problems. Moreover, our convergence result is uniform for any $n \ge 3$; thus, finite-sample performance is guaranteed.

2.1.2 NonGaussian covariates

When the covariates are nonGaussian random variables, it is more desirable to choose a distribution-free statistic for the screening rule. Therefore, instead of using Pearson's sample correlation, we study the extreme behavior of the Spearman 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

Then, Spearman rho is defined as

$$\rho_{ij} = \frac{\sum_{i=1}^{n}(Q_{ni}^{j} - \bar{Q}_{n}^{j})(Q_{ni}^{k} - \bar{Q}_{n}^{k})}{\sqrt{\sum_{i=1}^{n}(Q_{ni}^{j} - \bar{Q}_{n}^{j})^2 \sum_{i=1}^{n}(Q_{ni}^{k} - \bar{Q}_{n}^{k})^2}}, \tag{2.4}$$

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

Similarly 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 states that as n increases, S_{pn}^2 converges to a Gumble-type distribution.

Proposition 1. *Suppose $X_1, \cdots, X_p$ are i.i.d. 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 rho statistics. Then, for $\log p = o(n^{1/3})$, we have*

$$\lim_{n\to\infty} |\mathrm{P}\left((n-1)S_{pn}^2 - 4\log p + \log\log p \le x\right) - \exp\left\{-(8\pi)^{-1/2}\exp(-x/2)\right\}| = 0. \tag{2.5}$$

Theorem 1 and Proposition 1 characterize the magnitude of the maximal pairwise correlation and Spearman rho statistic, respectively, when the covariates are independent. If a pair of covariates, say X_1 and X_2, have an absolute sample correlation greater than the 95% quantile of the distribution given in Theorem 1 or Proposition 1, then they tend to be marginally dependent. Because we are only interested in pairs of truly important covariates, we further investigate the extreme behavior of the maximal pairwise R squared under the null model; that is, all β_j are equal to zero.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

2.2 R-squared screening for pairs of covariates

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

Consider a linear regression in which we regress Y onto a pair of covariates X_i and X_j, with $i \neq j$. Here, we can obtain the corresponding R-squared, R^2_{ij}. Under the model setting in (2.2), when all coefficients are zeros, the maximal pairwise R-squared, $\max_{1\leq i<j\leq p} R^2_{ij}$, cannot be too large. In fact, there exists an asymptotic bound for $\max_{1\leq i<j\leq p} R^2_{ij}$, as described in the following theorem.

Theorem 2. *Let $R^2_{pn} = \max_{1\leq i<j\leq p} R^2_{ij}$, where R^2_{ij} is the pairwise R-squared after regressing Y onto X_i and X_j, where $i \neq j$. Suppose $X_1, \cdots, X_p$ and Y are from the model setting in (2.2) and that Y is normally distributed. Then, when β_j are all zero, we have the following, for any fixed $n \geq 4$, $\delta > 0$, as $p \to \infty$: $P(R^2_{pn} \geq 1 - p^{-(4+\delta)/(n-3)}) = O(p^{-\delta/2}) \to 0$.*

Using the bound given by Theorem 2, we can design a screening rule to find pairs of covariates that are potentially associated with the response. In Section 3, we explain how to use the theoretical results for variable selection.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

3. Penalized variable selection using pairwise screening

In this section, we propose a pairwise screening procedure that takes advantage of the asymptotic results in Section 2. Furthermore, we establish a new penalization algorithm for variable selection.

3.1 Screening-based penalization

Given the limiting distribution of the maximal pairwise sample correlation described in Section 2, we propose the following screening rule to identify covariates pairs that are potentially correlated and related to the response:

$$\mathcal{G} = \{(i,j) : i < j, |\widehat{\text{Corr}}(X_i, X_j)| \geq a \text{ and } R^2_{ij} \geq r_0\}, \tag{3.6}$$

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

The group definition in (3.6) is a screening procedure with respect to covariate pairs. Screening is prevalent for high-dimensional data analyses. In particular, for penalized variable selection methods, high dimensionality makes it more difficult to capture the inherent sparsity structure, making dimension reduction necessary. To this end, Fan and Lv [2008] introduced the SIS method, which ranks the covariates based on the magnitude of their sample correlation with the response. Specifically, let $\mathbf{w} = (w_1, w_2, \cdots, w_p)^T$ be a vector,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

such that $w_j = |\widehat{\text{Corr}}(X_j, Y)|$, and let γ be 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 X_j with $\beta_j \neq 0$ are included in the subset of covariates.

To take advantage of the distribution information in 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 from among $X_1, X_2, \cdots, X_p$. Define the set of paired covariates as

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

Our proposed method solves 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 problem, it can be seen that we apply different penalties to the covariates, based on the results of two types of screening. Intuitively, the proposed penalty works as

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

follows:

- For a covariate that is included in both $\mathcal{C}$ and $\mathcal{M}$, we apply the l_2-penalty only 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 apply the l_1-penalty only, because there is no significant multicollinearity between it and other covariates.
- For a covariate that is not included in $\mathcal{M}$, because 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 a certain covariance structure. In particular, when the covariates are all independent, our method reduces to the SIS-LASSO, which performs marginal screening first, and then implements the LASSO on the remaining covariates; and, when the predictors are all highly correlated, such that $\mathcal{G}$ includes all covariate pairs, our method is equivalent to the SIS-Ridge.

Thus far, we have established a new penalized variable selection. Now, we discuss how to solve the optimization problem in (3.9). The penalty part of (3.9) is convex. Therefore, we can solve it efficiently using coordinate descent algorithm [Friedman et al., 2010]. Specifically,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

the updating rule has the following form:

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

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 $S(z) = \text{sign}(z)(|z| - \lambda)_+$ is the soft-thresholding function. In practice, we can first implement SIS to obtain $\mathcal{M}$ when the dimension is high, and then run the algorithm on the covariates X_j, for $j \in \mathcal{M}$.

Remark 1. The computational cost of the pairwise screening procedure is $O(p^2)$, which can become very inefficient as p increases. In our proposed procedure, to reduce the computational complexity, we implement the marginal screening first to obtain $\mathcal{M}$. Because the cardinality of $\mathcal{M}$ is $O(n/\log(n))$, the computational cost of applying pairwise screening to $\mathcal{M}$ reduces to $O\big((n/\log(n))^2\big)$.

3.2 Further extensions

As discussed in the previous subsection, we introduce a new penalized method that combines marginal screening with pairwise screening in a linear model setting. Note that the pairwise covariate screening does not involve the response. Therefore, our method can be extended to include generalized linear models (GLM), such as the logistic regression for binary responses, or the 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 $\theta = \mathbf{x}^T\boldsymbol{\beta}$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

for GLMs.

Similarly 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-squared screening for GLMs. This is because for GLMs, it is not reasonable to use the regression R-squared to evaluate the associations between the covariates and the response. We further define the set of paired covariates as follows:

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

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 the logistic regression, we need to solve the following penalized maximum likelihood problem:

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

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 using the 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 $P_{\lambda_1,\lambda_2}(\boldsymbol{\beta})$.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

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 PCS achieves variable selection consistency.

Note that we implemented the marginal screening using SIS on the covariates set. Fan and Lv [2008] demonstrated that, under certain regularity conditions, SIS exhibits screening consistency; that is, the resulting subset of covariates includes all important variables. Owing to space constraints, we present the main result only. The regularity conditions $(A1)$-$(A4)$ are provided in the Appendix.

Proposition 2 (Fan and Lv [2008])**.** *Under* $(A1)$-$(A4)$*, 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* $C > 0$*,*

$$\mathrm{P}\left(\mathcal{M}^* \subset \mathcal{M}_\gamma\right) = 1 - O[\exp\{-C^{1-2\kappa}/\log(n)\}], \tag{4.14}$$

where $\mathcal{M}_\gamma$ *is the subset of covariates obtained from the SIS.*

The above proposition guarantees that all important variables survive the marginal screening with high probability. In order to achieve 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 required such that the event $\mathcal{C} \cap \mathcal{M} \subset \mathcal{M}^*$ occurs with high probability.

Theorem 3. *Suppose the following conditions hold:*

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

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

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

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

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

Here, $Corr(X_i, X_j)$ *denotes the population correlation between covariates* X_i *and* X_j. *Then, under conditions* $(B1)$ *and* $(B2)(a)$ *or conditions* $(B1)$ *and* $(B2)(b)$, *we have that as* $n \to \infty$,

$$\mathrm{P}\left(\mathcal{C} \cap \mathcal{M} \subset \mathcal{M}^*\right) \to 1. \tag{4.15}$$

Given Proposition 2 and Theorem 3, to demonstrate the selection consistency of PCS, we need only show that the l_1-penalty in (3.9) can identify the important variables in $\mathcal{C}^c \cap \mathcal{M}$ exactly. This relates to the selection consistency of the LASSO, which has been studied extensively. In particular, Zhao and Yu [2006] have shown that the Irrepresentable Condition (specified later) is almost necessary and sufficient for the LASSO to select all important variables.

We first introduce some necessary notation. Let $C = \frac{1}{n}X^T X$. Without loss of generality, assume that $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$, where $\beta_j \neq 0$ for $j = 1, \ldots, s$, and $\beta_j = 0$ otherwise. By Theorem 3, we further assume that $\mathcal{C} \cap \mathcal{M} = \{1, \ldots, s_1\}$, where $1 \leq s_1 \leq s$. 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. Similarly, we write $\boldsymbol{\beta}^{(1)}_1 = (\beta_1, \ldots, \beta_{s_1})^T$, $\boldsymbol{\beta}^{(1)}_2 = (\beta_{s_1+1}, \ldots, \beta_s)^T$,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

and $\boldsymbol{\beta}^{(2)} = (\beta_{s+1}, \ldots, \beta_p)^T$.

Set $C_{11}^{(11)} = \frac{1}{n} {X_{(1)}^1}^T X_{(1)}^1$, $C_{11}^{(12)} = \frac{1}{n} {X_{(1)}^1}^T X_{(1)}^2$, $C_{11}^{(21)} = \frac{1}{n} {X_{(1)}^2}^T X_{(1)}^1$, $C_{11}^{(22)} = \frac{1}{n} {X_{(1)}^2}^T X_{(1)}^2$, $C_{21}^{(1)} = \frac{1}{n} X_{(2)}^T X_{(1)}^1$, $C_{21}^{(2)} = \frac{1}{n} X_{(2)}^T X_{(1)}^2$, $C_{22} = \frac{1}{n} X_{(2)}^T X_{(2)}$, $C_{12}^{(1)} = \frac{1}{n} {X_{(1)}^1}^T X_{(2)}$, and $C_{12}^{(2)} = \frac{1}{n} {X_{(1)}^2}^T X_{(2)}$. Then, C can be expressed in blockwise form, as follows:

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

We impose the following assumption, which is analogous to the Irrepresentable Condition introduced by Zhao and Yu [2006]. Specifically, we assume that there exists a constant $\delta > 0$, such that

$$\|C_{21}^{(2)} (C_{11}^{(22)})^{-1} \text{sign}(\boldsymbol{\beta}_2^{(1)})\|_{\max} \leq 1 - \delta, \tag{4.16}$$

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

In fact, we can show that this condition 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 4. *Assume there exists $\lambda_0 > 0$, such that $\lambda_{min}(C_{11}^{(11)}) \geq \lambda_0$, $\lambda_{min}(C_{11}^{(22)}) \geq \lambda_0$, and that conditions $(B1)$ and $(B2)(b)$ hold. Suppose the Irrepresentable Condition holds; that is, $\exists \xi > 0$, s.t.*

$$\|C_{21} C_{11}^{-1} sign(\boldsymbol{\beta}_1)\|_{\max} \leq 1 - \xi, \tag{4.17}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

where $C_{11} = \begin{pmatrix} C_{11}^{(11)} & C_{11}^{(12)} \\ C_{11}^{(21)} & C_{11}^{(22)} \end{pmatrix}$, $C_{21} = \begin{pmatrix} C_{21}^{(1)} & C_{21}^{(2)} \end{pmatrix}$, $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_s)^T$, *and* ξ *is a positive constant. Then, with probability tending to one, condition* (4.16) *holds.*

The assumptions $\lambda_{min}(C_{11}^{(11)}) \geq \lambda_0$ and $\lambda_{min}(C_{11}^{(22)}) \geq \lambda_0$ in Theorem 4 require that $C_{11}^{(11)}$ and $C_{11}^{(22)}$ have eigenvalues bounded below. Given the Irrepresentable Condition in (4.17), we need additional constraints on the random noise ε_i and the coefficients of the important variables $\beta_1, \cdots, \beta_s$.

$(C1)$ ε_i are i.i.d. random variables with a finite $2k$ moment $E(\varepsilon_i)^{2k} < \infty$, for an integer $k > 0$.

$(C2)$ There exists $0 < \alpha \leq 1$ and $d_0 > 0$, such that $n^{\frac{1-\alpha}{2}} \min_{j=1,\cdots,s} |\beta_j| \geq d_0$.

Thus far, we have discussed the theoretical assumptions required to ensure the selection consistency of the proposed PCS method. We conclude the consistency result in the following theorem.

Theorem 5. *Suppose conditions (A1)–(A4), (C1)–(C2), and inequality* (4.17) *hold, and that the assumptions of Theorem 4 are satisfied. Then, for any* λ_1 *such that* $\frac{\lambda_1}{\sqrt{n}} = o(n^{\alpha}/2)$ *and* $\frac{1}{p}(\frac{\lambda_1}{\sqrt{n}}) \to \infty$, *we have*

$$\mathrm{P}\left(\{j : \hat{\beta}_j \neq 0\} = \mathcal{M}^*\right) \to 1 \text{ as } n \to \infty, \tag{4.18}$$

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

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

The proof follows immediately from Proposition 2 and Theorems 3 and 4, as well as from the selection consistency of the LASSO. Thus, under certain conditions, our proposed method is consistent in terms of variable selection. In Section 5, we use numerical examples to show that our proposed method performs well in practice.

5. Numerical studies

In Section 3, we 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.

5.1 Simulation study

In this section, we use several simulations to show that our method with PCS or pairwise rank-based correlation screening (PRCS) outperforms some existing variable selection procedures. Specifically, PCS denotes our proposed method using the limiting distribution in Theorem 1, and PRCS uses the asymptotic result in Proposition 1.

For the comparison, we consider the LASSO, elastic net (Enet), SIS-LASSO, SIS-elastic net (SIS-Enet), and SIS-PACS methods. The SIS-PACS applies the PACS method proposed by Sharma et al. [2013] after implementing the SIS procedure. For the SIS-type methods, we first implement SIS to identify those covariates with the largest $[n \backslash \log n]$ absolute sample correlations with the response. Then, we perform the 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

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

Table 1: Results for Example 1. For each method, we report the average MSE, l_2-distance, FN, and FP over 100 replications (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)	4.50 (0.01)	0.00 (0.00)	12.00 (0.00)
SIS-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)
PRCS	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)
SIS-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)
PRCS	4.85 (0.06)	0.89 (0.03)	0.00 (0.00)	1.21 (0.11)

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 (2.2) and conduct 100 replications. Each simulated data set 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 at 100 throughout the simulation study. In the next subsection, we also vary the sample size in our sensitivity

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

Table 2: Results for Example 2. The format of this table is the same as Table 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)
PRCS	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)
PRCS	6.48 (0.13)	2.14 (0.07)	0.68 (0.05)	0.73 (0.24)

Table 3: Results for Example 3. The format of this table is the same as Table 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)
PRCS	59.76 (0.56)	4.83 (0.04)	4.97 (0.02)	0.00 (0.00)

study. We only fit models on the training data, and we use the validation data to select the tuning parameters. Given the fitted model, we can calculate the FN, FP, and estimation error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$, and make predictions and calculate the out-of-sample MSEs using the test

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

Table 4: Results for Example 4. The format of this table is the same as Table 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)
PRCS	0.099 (0.004)	5.34 (0.06)	1.71 (0.13)	3.26 (0.32)

Table 5: Results for Example 5. The format of this table is the same as Table 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	89.82 (2.70)	3.54 (0.15)	0.26 (0.08)	7.32 (0.45)
PCS	79.79 (3.16)	2.42 (0.14)	0.42 (0.10)	1.29 (0.33)
PRCS	74.60 (1.24)	2.15 (0.12)	0.31 (0.08)	0.06 (0.03)

data. We simulate the covariates from the multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$, where $\Sigma = (\sigma_{ij})_{p\times p}$ is the correlation matrix.

Details of the simulated examples are as follows:

Example 1: We consider $p = 1000$ or 5000, $\sigma = 2$, and $\boldsymbol{\beta} = (2, 2, \cdots, 2, 0, \cdots, 0)^T$, where the first 10 coefficients are nonzero and equal to two. We set $\sigma_{ij} = 0.8$ for $1 \le i \ne j \le 5$, $6 \le i \ne j \le 10$ and set it to zero for all the other $i \ne j$. We also consider $\sigma = 6$; see the Supplementary Material. In other words, there are two groups in the covariates, where each group has five important variables.

Example 2: We consider $p = 1000$ or 5000, $\sigma = 2$, and $\boldsymbol{\beta}_0 = (3, -1.5, 2, 0, \cdots, 0, \cdots, 0)^T$,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

where the first three coefficients are nonzero. We also consider $\sigma = 6$; see Supplementary Material. We generated Gaussian covariates with $\sigma_{ij} = 0.5^{|i-j|}$, for $1 \leq i \neq j \leq 1000$.

Example 3: The coefficients have the same setup as those in Example 1. However, we set $\sigma_{ij} = 0.8$ for $1 \leq i \neq j \leq 5$, and to zero for all the other $i \neq j$. Therefore, only some of the important variables are highly correlated. We consider $p = 5000$ and $\sigma = 6$ in this example.

Example 4: Here, 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 setups as those in Example 1. We consider $p = 5000$ and $\sigma = 6$ in this example. Instead of comparing MSEs, we calculate the classification errors on the test data. We do not include SIS-PACS in this example because the R program does not support GLMs.

Example 5: In this example, we generate the covariates from a multivariate t-distribution, where X_j are t-distributed with degrees of freedom five. The covariance structure of the covariates and the coefficients are set up as in Example 1. We consider $p = 5000$ and $\sigma = 6$ in this example.

The results for Example 1 are shown in Table 1. We see that when there are groups in the covariates, the performance improvement of our approach is significant compared with that of other penalized methods. Although the elastic net-based procedures perform better than LASSO-type approaches do 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

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

addition, although SIS-PACS shows competitive performance when σ is small, it tends to include more unimportant variables in the model when the noise level increases, and therefore may not work well.

Table 2 displays the performance comparisons for Example 2. Compared with Example 1, this setting is a more difficult one for our method, because correlations exist between all pairs of covariates. Nevertheless, PCS and PRCS perform better than, or as well as, other methods do in terms of the estimation error and prediction accuracy. Moreover, with the exception of SIS-Ridge, our proposed methods are able to identify more important variables than the other methods do in this example when the noise level is low.

Table 3 shows the results for Example 3, where only some of the important variables are correlated. This example is more difficult than the scenario in Example 1, owing to the correlation structure of the covariates. For example, there are significantly more FNs in all the procedures. Nevertheless, our method still outperforms all the others in terms of prediction and variable selection accuracy.

Example 4 considers a logistic regression setting; see Table 4. The results show that the proposed method performs competitively, even as the correlations between the covariates vary.

Table 5 displays the results for all methods in a nonGaussian covariate setting. Similarly to Example 1, our proposed PCS and PRCS significantly outperform their competitors. Moreover, owing to the nonGaussian setups, the nonparametric method PRCS outperforms PCS.

In summary, our method is able to take advantage of the correlation structure among

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

the predictors. Compared with other penalized variable selection procedures, our method performs well, especially when the covariates are highly correlated.

5.2 Sensitivity study

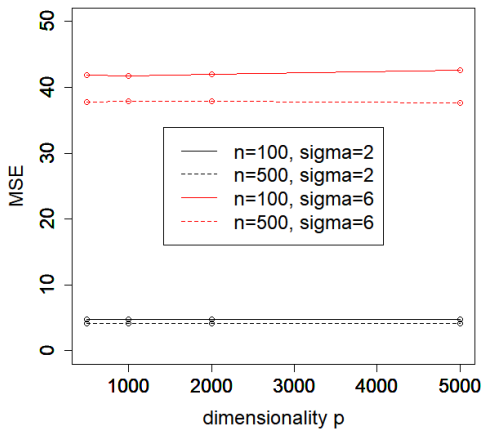
In this subsection, we investigate whether the performance of our method depends on the sample size, dimensionality, and noise level. In particular, we consider $n = 100$ or 500, $p = 500, 1000, 2000$, or 5000, and $\sigma = 2$ or 6 in Example 1 in Section 5.1. We illustrate the MSE, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$, FN, and FP against different values of p for each configuration of sample size and noise level in Figure 1.

The plots show that the performance of PCS does not change much as the dimensionality p increases from 500 to 5000, especially in terms of the 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 the sample size, dimensionality, and SNR.

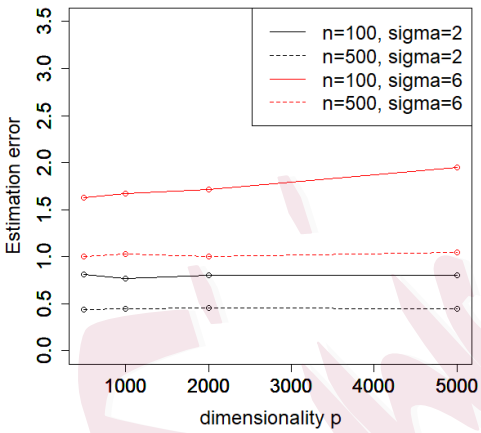
5.3 Soil data

We first demonstrate the performance of our method in real applications using a small data set. This data set contains 15 covariates of soil characteristics for 20 plots within the same area in the Appalachian Mountains. 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 2. One can see that some predictors

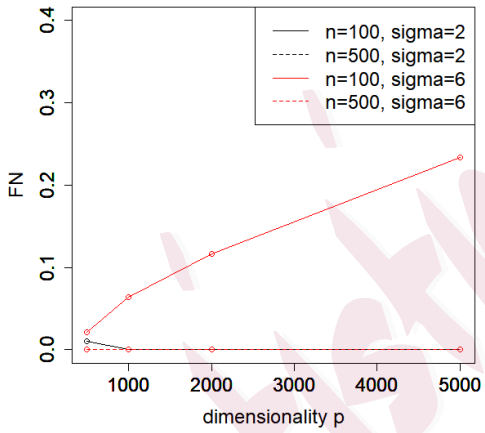
Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170



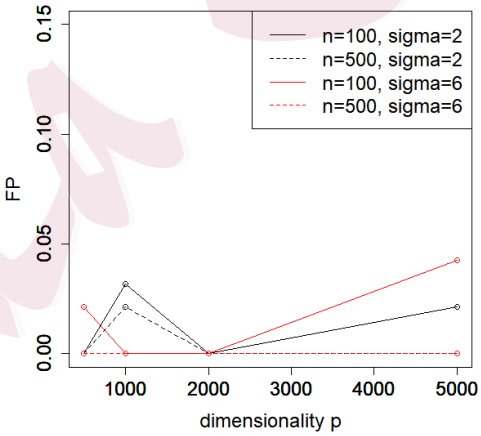
(a) MSE



(b) Estimation error



(c) FN



(d) FP

Figure 1: Performance of PCS for different dimensionality p.

are highly correlated. In particular, the magnitudes of the pairwise correlations between the sum of cations (SumCation), calcium, magnesium, the base saturation (BaseSat), and the cation exchange capacity (CEC) are as large as 0.9. This is because SumCation, BaseSat, and CEC are characteristics of cations, whereas calcium and magnesium are examples of cations [Bondell and Reich, 2008].

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

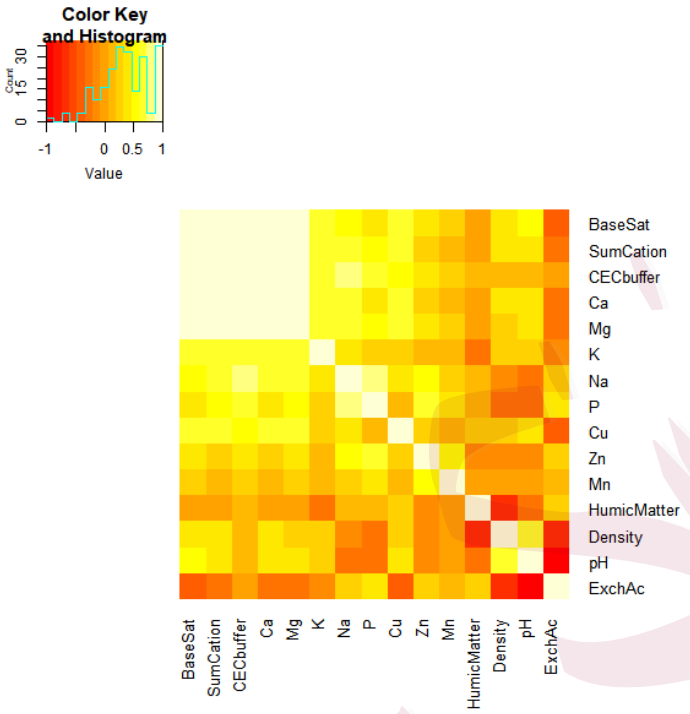


Figure 2: Heatmap for the absolute pairwise correlation matrix of the covariates for soil data.

We conduct a total of 100 replications. In each replication, 15 samples are chosen randomly as the training set, and the remainder form the test set. As in the simulation experiments, we applied the LASSO, Enet, Ridge, and our proposed PCS and PRCS to the data set. For each method, five-fold cross-validation is used to choose the tuning parameters,

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)
PRCS	1.028 (0.063)	5.96 (0.38)

Table 6: Average MSE and model size (with standard errors in parentheses) for Enet, LASSO, Ridge, and the proposed method for soil data.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

because the sample size is very small. We report the average prediction errors on the test data and the model size in Table 6. The results show that PCS and PRCS outperform all other procedures in terms of prediction accuracy. Moreover, PCS and PRCS tend to include more covariates in the model than the LASSO and Enet do.

To further investigate the performance in terms of variable selection, we summarize the frequency with which each covariate is selected for the LASSO, Enet, and the proposed method; see Table 7. The variables that are most frequently selected by the LASSO and Enet, for instance, CEC, Mn, and HumicMatt, also tend to be included by 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 the Enet or LASSO.

5.4 Riboflavin data

In this section, we consider a real data set on 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 logarithms of the expression levels of $p = 4081$ genes. More detail about the data set can be found in Bühlmann et al. [2014]. Before the analysis, all covariates are standardized to have zero means and unit standard deviations.

For the comparison, we apply the LASSO, Enet, SIS-LASSO, SIS-Enet, SIS-ridge, and our method to the data set. We conduct 100 replications, and randomly split the data set into a training set of size 50, with the remainder as the test data. For all methods, we

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

Variables	PCS	Enet	LASSO
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 7: Frequency of each variable being selected for PCS, Enet, and the LASSO out of 100 replications.

implement 10-fold cross-validation on the training data to select the penalty parameters.

The results are reported in Table 8, and show that PCS exhibits significant improvement in terms of the out-of-sample MSE over those of its competitors. On the other hand, PRCS does not perform well compared with PCS. A possible reason is that, in this data set, all variables are log transformations and are approximated well by a Gaussian distribution. Moreover, owing to the assumption of Proposition 1, where $\log p = o(n^{1/3})$, PRCS is more sensitive to the dimensionality and the sample size of the data set. As a result, PRCS may not achieve good performance when the dimensionality is too high.

We also examine the gene selection results. Eight genes are selected at least 50 times in the 100 replications using our method, that is, XTRA_at, YCKE_at, YDAR_at, YOAB_at, YWFO_at, YXLC_at, YXLD_at, and YXLE_at. Apart from YXLC_at, all the other genes

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

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)
PRCS	0.361 (0.018)	12.77 (0.37)

Table 8: Average MSE and model size (with standard errors in parentheses) for SIS-Enet, SIS-LASSO, SIS-Ridge, PCS, and PRCS for riboflavin data.

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. This indicates that our method is capable of identifying potentially important variables that are highly correlated with other variables.

6. Discussion

We have proposed 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 for covariate pairs takes advantage of the distribution information about the maximal absolute pairwise sample correlation between covariates, and is applicable to large-scale problems. Simulation experiments and real-data studies demonstrate that the proposed method performs well when important variables are highly correlated, compared with existing approaches. Future research can consider other extensions to our proposed method, such as the Cox model for survival data.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

Supplementary Material

The online Supplementary Material contains proofs of Corollary 1 and Theorem 2, and additional numerical studies.

A. Technical Proofs

We present some regularity conditions and key proofs in the appendix.

Regularity Conditions for Sure Independence Screening Define $\mathbf{z} = \Sigma^{-1/2}\mathbf{x}$, $Z = X\Sigma^{-1/2}$. Let $\mathcal{M}^*$ be the index set of covariates with nonzero 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

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

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

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

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

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

Proof of Theorem 1. To prove Theorem 1, we need to use the following lemma, which is

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

from Arratia et al. [1989].

Lemma 1. *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 $\lambda = \sum_{\alpha\in I} \mathrm{P}(\eta_\alpha > t)$. Then*

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

where $b_1 = \sum_{\alpha\in I}\sum_{\beta\in B_\alpha} \mathrm{P}(\eta_\alpha > t)\,\mathrm{P}(\eta_\beta > t)$, $b_2 = \sum_{\alpha\in I}\sum_{\alpha\neq\beta\in B_\alpha} \mathrm{P}(\eta_\alpha > t, \eta_\beta > t)$ and $b_3 = \sum_{\alpha\in I} E|\mathrm{P}(\eta_\alpha > t|\sigma(\eta_\beta, \beta \notin B_\alpha)) - \mathrm{P}(\eta_\alpha > t)|$, 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 $I = \{(i,j); 1 \le i \le j \le p\}$. 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 $A_\alpha = A_{ij} = \{|\rho_{i,j}|^2 \ge t\}$, where $\rho_{i,j} = |\widehat{\mathrm{Corr}}(X_i, X_j)|$. Let $W_{pn} = \max_{1\le i<j\le p} |\rho_{i,j}|$, by the Chen-Stein method (in particular, Lemma 6.2 in Cai and Jiang (2011)),

$$|P(W_{pn}^2 \le t) - e^{-\lambda_{p,n}}| \le b_1 + b_2, \tag{A.19}$$

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 $b_1 \le 2p^3 P(A_{12})^2$ and $b_2 \le 2p^3 P(A_{12}A_{13})$.

Since $X_1, \cdots, X_p$ are independent, A_{12} and A_{13} are also independent with equal proba-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

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

$$\begin{aligned} P(A_{12}^*) &= \frac{2(1-t^*)^{(n-2)/2}}{B(\frac{1}{2}, \frac{n-2}{2})(n-2)\sqrt{t^*}}(1 + O(\frac{1}{\log(p)})). \\ &= p^{-2}\big(1 - \frac{2}{n-2}x\big)^{\frac{n-2}{2}} \sqrt{\frac{1-p^{-4/(n-2)}}{a_{p,n}}}\big(1 + (\frac{b_{p,n}}{a_{p,n}}x)\big)^{-1/2}\big(1 + O(\log^{-1}(p))\big). \\ &= p^{-2}\big(1 - \frac{2}{n-2}x\big)^{\frac{n-2}{2}} \big(1 + O(\frac{\log\log(p)}{\log(p)})\big)\big(1 + O(\log^{-1}(p))\big)^2 \\ &= p^{-2}\big(1 - \frac{2}{n-2}x\big)^{\frac{n-2}{2}} \big(1 + O(\frac{\log\log(p)}{\log(p)})\big) \end{aligned} \tag{A.20}$$

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 (A.19) that uniformly for any $n \geq 3$ and $x \leq \frac{n-2}{2}$,

$$\lim_{p\to\infty} |P(W_{pn}^2 \leq t^*) - \exp\big\{-\frac{1}{2}\big(1 - \frac{2}{n-2}x\big)^{\frac{n-2}{2}}\big\}| = 0. \tag{A.21}$$

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 $n \geq 3$,

$$\lim_{p\to\infty} P(W_{pn} \leq t^*) = 1 \tag{A.22}$$

Combining (A.21) and (A.22) we have uniformly for any $n \geq 3$,

$$\lim_{p\to\infty} |P(W_{pn} \leq t^*) - I(x \leq \frac{n-2}{2})\exp\big\{-\frac{1}{2}\big(1 - \frac{2}{n-2}x\big)^{\frac{n-2}{2}}\big\} - I(x > \frac{n-2}{2})| = 0. \tag{A.23}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

Or equivalently,

$$\lim_{p\to\infty} |P(\frac{W_{pn}^2 - a_{p,n}}{b_{p,n}} \leq x) - I(x \leq \frac{n-2}{2}) \exp\big\{ -\frac{1}{2}\big(1 - \frac{2}{n-2}x\big)^{\frac{n-2}{2}}\big\} - I(x > \frac{n-2}{2})| = 0. \tag{A.24}$$

□

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 $\hat{\rho}_{ij} = |\widehat{\text{Corr}}(X_i, X_j)|$, $f(n,p,\alpha)$ 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

$$P(\mathcal{C} \cap \mathcal{M} \subset \mathcal{M}^*) \geq P(A \cap B) \geq P(A) + P(B) - 1. \tag{A.25}$$

For the event A, we have

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

Under the assumption $(B1)$, $(n/\log(n))^2 p^{-(4+\delta)/2} \to 0$ 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

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

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

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

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

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

$$
\begin{aligned}
\mathrm{P}\left(\hat{\rho}_{ij} \geq \delta_{p,n}\right) =& \mathrm{P}\left(n^{1/2}(z(n)-\xi) \geq n^{1/2}(\frac{1}{2}\log\frac{1+\delta_{p,n}}{1-\delta_{p,n}} - \xi)\right) \\
=& \mathrm{P}\left(Z \geq n^{1/2}(\frac{1}{2}\log\frac{1+\delta_{p,n}}{1-\delta_{p,n}} - \xi) + o_n(1)\right) \leq \frac{e^{-C_{p,n}n}}{\sqrt{2\pi n}C_{p,n}},
\end{aligned} \tag{A.26}
$$

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

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

yields $C_{p,n} \to \infty$. Then the tail probability in (A.26) 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 (A.26) 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 $P(\mathcal{C} \cap \mathcal{M} \subset \mathcal{M}^*) \to 1$ as $n \to \infty$. □

Proof of Theorem 4. It follows from (4.17) directly that

$$\|(C_{21}^{(2)} - C_{21}^{(1)}(C_{11}^{(11)})^{-1}C_{11}^{(12)})\big(C_{11}^{(22)} - C_{11}^{(21)}(C_{11}^{(11)})^{-1}C_{11}^{(12)}\big)^{-1}\text{sign}(\boldsymbol{\beta}_1^{(2)})\|_{\max} \leq 1 - \xi, \quad \text{(A.27)}$$

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

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

$\|C_{21}^{(1)}(C_{11}^{(11)})^{-1}C_{11}^{(12)}\|_{\max} = o_p(1)$ as $n \to \infty$. Therefore

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

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

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

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

References

Richard Arratia, Larry Goldstein, and Louis Gordon. Two moments suffice for poisson approximations: the chen-stein method. *The Annals of Probability*, 17(1):9–25, 1989.

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.

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.

T 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.

T Tony Cai, Tiefeng Jiang, et al. 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.

Emmanuel Candès and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

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.

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

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

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2003.

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

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.

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

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

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

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)*, 58(1):267–288, 1996.

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.

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.

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0170

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.

Siliang Gong

University of Pennsylvania

E-mail: (siliang@pennmedicine.upenn.edu)

Kai Zhang

The University of North Carolina at Chapel Hill

E-mail: (zhangk@email.unc.edu)

Yufeng Liu

The University of North Carolina at Chapel Hill

E-mail: (yfliu@email.unc.edu)