Received: 27 December 2019 | Revised: 22 January 2020 | Accepted: 24 January 2020
DOI: 10.1002/sta4.272

WILEY

ORIGINAL ARTICLE

High-dimensional variable screening under multicollinearity

Naifei Zhao[1] | Qingsong Xu[2] | Man-Lai Tang[3] | Binyan Jiang[4] | Ziqi Chen[2] | Hong Wang[2]

[1]Institute of Statistics and Big Data, Renmin University of China, Beijing, 100872, China
[2]School of Mathematics and Statistics, Central South University, Changsha, 410083, China
[3]Department of Mathematics and Statistics, Hang Seng University of Hong Kong, Hong Kong, 999077, China
[4]Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong, 999077, China

Correspondence
Hong Wang, School of Mathematics and Statistics, Central South University, Changsha 410083, China.
Email: wh@csu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 11731011 and 11931014; National Social Science Foundation of China, Grant/Award Number: Grant No.17BTJ019

Variable screening is of fundamental importance in linear regression models when the number of predictors far exceeds the number of observations. Multicollinearity is a common phenomenon in high-dimensional settings, in which two or more predictor variables are highly correlated, leading to the notorious difficulty for high-dimensional variable screening. *Sure independence screening* (SIS) procedure can greatly reduce the dimensionality, but it may break down when the predictors are highly correlated. By combing the factor modelling with SIS, the *profiled independence screening* (PIS) approach was proposed. However, under a spiked population model, the profiled predictors could not be guaranteed to be uncorrelated and PIS may therefore be misleading. Instead of assuming either the predictors are uncorrelated as in SIS or the profiled predictors are uncorrelated as in PIS, a more general and challenging scenario is considered in which the predictors can be highly correlated. A so-called preconditioned PIS (PPIS) method is proposed that produces asymptotically uncorrelated profiled predictors and thus leads to consistent model selection results under a spiked population model. Compared with PIS, the proposed method could handle the complex multicollinearity case, such as a spiked population model with a slow spectrum decay of population covariance matrix, while keeping the calculation simple. The promising performance of the proposed PPIS method will be illustrated via extensive simulation studies and two real examples.

KEYWORDS
high dimensionality, multicollinearity, preconditioning, profiling, screening, spiked population model

1 | INTRODUCTION

Exploring the relationship between the response and some predictors in a linear model is an important topic in statistics. Rapid advances in computing power and other modern technologies drive big data collections across many scientific disciplines (e.g., genomics, functional magnetic resonance imaging, tomography, finance, and chemometrics) in which the predictor dimensions are substantially larger than the sample sizes. In these settings, the classical *ordinary least squares* estimate is no longer applicable, and difficulties are encountered in estimating the regression coefficient vector in linear models. Over the last two decades, various approaches have been proposed to tackle this issue, and they are mainly built on the premise that the number of the variables that actually contribute to the response is relatively small although the predictor dimension is high. Various penalization methods have been proposed to simultaneously perform model selection and parameter estimation; see, for example, the lasso (Tibshirani, 1996; Zhao & Yu, 2006), the smoothly clipped absolute deviation (Fan & Li, 2001; Fan & Peng, 2004), the elastic net (Fu et al., 2011; Zou & Hastie, 2005), the adaptive lasso (Zou, 2006), and the adaptive elastic net (Zou & Zhang, 2009). A drawback of these methods except lasso is that the consistency property for model selection may not be guaranteed if the predictor dimension (vastly) outnumbers the sample size (Zhao & Yu, 2006). Furthermore, they are computationally extremely intensive for high-dimensional settings. Some recent works sought to reduce the high dimensionality rapidly before performing a refined analysis. The *sure independence screening* (SIS; Fan & Lv, 2008), a dimension reduction procedure that screens the marginal correlations to determine which variables should remain in the model, is shown to possess the sure screening property and is computationally very simple (Fan & Lv, 2008; Fan et al., 2009; Fan & Song, 2010). Alternative screening methods using new measures of association between each variable and the response have been proposed and carefully studied, including but not limited to Cho and Fryzlewicz (2012), Huang, Xu, and Liang (2012), Ji and Jin (2012), G. Li, Peng, Zhang, and Zhu (2012), R. Li, Zhong, and Zhu (2012), Wang (2012), Wang and Leng (2016), Witten and Tibshirani (2009), and Zhu, Li, Li, and Zhu (2011).

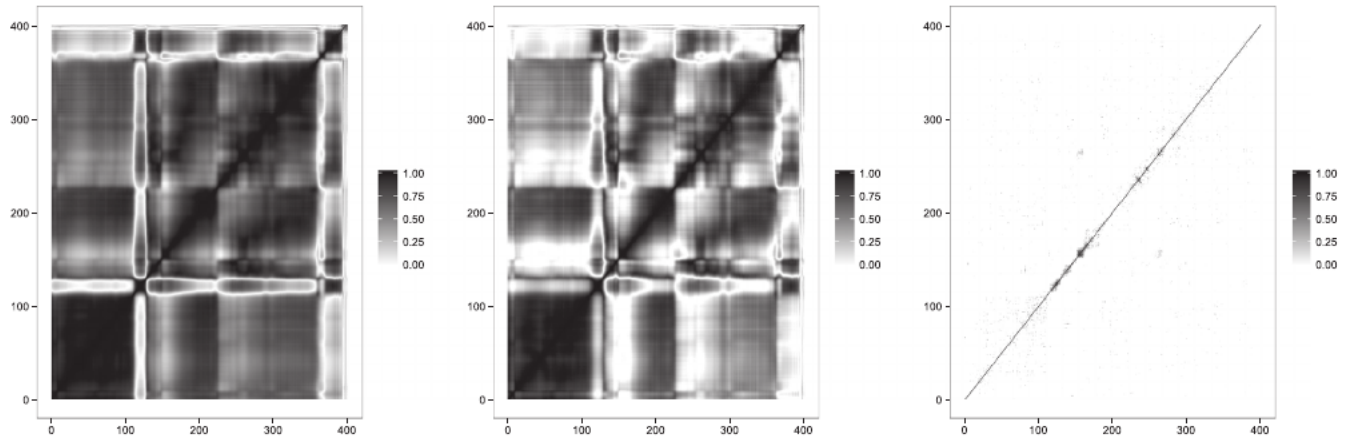

FIGURE 1 Heat maps of the absolute correlation matrices corresponding to the soft-thresholded covariance matrix estimators (Rothman et al., 2009) using the gasoline dataset. The left panel is the absolute correlation matrix of the 401-dimensional predictors. The middle panel is that of the profiled predictors for profiled independence screening (PIS). The right panel is that of the profiled predictors for preconditioned PIS (PPIS)

Multicollinearity is a notorious and frequently encountered phenomenon in high-dimensional data analysis (Fan et al., 2009; Yu, Jiang, & Land, 2015). It usually refers to designs in which two or more predictors are strongly correlated and typically possess a latent factor structure. For example, the gasoline dataset that motivates this study consists of 60 samples with the octane number being the response variable and the wavelength intensities measured at 401 points being the predictors, leading the design matrix of dimension 60×401. We estimate the covariance matrix of predictors by soft-thresholding (Rothman, Levina, & Zhu, 2009), and the heat map of the corresponding absolute correlation matrix is given in the left panel of Figure 1. Each picture in Figure 1 is composed of 401×401 points, and each point represents the absolute value of the element at the corresponding position of correlation matrix. The darker the colour of a point, the closer the absolute value of this element is to one. Similarly, the lighter the colour of a point, the closer the absolute value of this element is to zero. It can be clearly observed that many predictors are highly correlated in the first picture of Figure 1.

Variable screening under multicollinearity for high-dimensional dataset is challenging and has not been well addressed (Ke, Jin, & Fan, 2014). Screening methods such as SIS based on marginal correlation between each predictor and the response would have nonzero probabilities of including irrelevant variables. The *profiled independence screening* (PIS) approach proposed by Wang (2012) provides a computationally efficient way for consistent variable screening. It uses profiled factor operation to eliminate the correlation between the predictors. However, the success of PIS hinges on the condition that the profiled predictors are also uncorrelated, which may still be impossible under a spiked population model (Baik & Silverstein, 2006; Johnstone, 2001) in the case of slow decay for eigenvalues of the population covariance matrix. With the gasoline dataset, we again apply the soft-thresholding method to estimate the covariance matrix of the profiled predictors by PIS. The heat map of the corresponding absolute correlation matrix is shown in the middle panel of Figure 1. Clearly, there are still many nonzero elements outside the main diagonal of the correlation matrix, indicating that many profiled predictors are still highly correlated.

In this article, we propose a novel method called preconditioned PIS (PPIS) for high-dimensional variable screening. The major advantage of PPIS is that it is as simple as PIS and produces asymptotically uncorrelated profiled predictors under a spiked population model. The key of our method is the twice decorrelation of predictors: factor profiling and preconditioning. We show that the preconditioning procedure can guarantee that the profiled predictors are asymptotically uncorrelated to each other. Preconditioning is a commonly used technique, and several preconditioners have been proposed to deal with high-dimensional linear regressions (Jia & Rohe, 2012; Wang, Dunson, & Leng, 2016). As an empirical evidence, the right panel of Figure 1 provides the heat map of the absolute correlation matrix corresponding to the soft-thresholded covariance matrix estimator of the profiled predictors for PPIS using the gasoline dataset. It can be seen that the correlation matrix of the profiled predictors for PPIS is very close to an identity matrix. Although the uncorrelated assumption for the profiled predictors is crucial for the success of PIS in variable screening, our PPIS approach is more promising for analysing datasets with highly correlated predictors. Theoretical justifications regarding the consistency of variable screening of PPIS are provided.

The rest of this paper is organized as follows. In the next section, we first give an introduction to factor profiling. Then, the spiked population model is introduced. We then present our PPIS approach with theoretical justifications. Simulation studies and real data analyses are reported in Section 3. We conclude the article in Section 4 and put the technical details to the Supporting Information.

2 | THE METHODOLOGY AND THEORY

2.1 | Linear regression with factor profiling

Let $\{y_i, x_i\}$ be the collected observations for the ith subject ($1 \le i \le n$), where $y_i \in \mathbf{R}$ is the response and $x_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbf{R}^p$ is the p-dimensional predictor vector with $p > n$. The relationship between y_i and x_i can be depicted as a simple linear regression

$$y_i = x_i^T \beta + \varepsilon_i, \tag{1}$$

where $\varepsilon_i \sim N(0, \sigma^2)$ is the random noise and $\beta = (\beta_1, \ldots, \beta_p)^T \in \mathbf{R}^p$ is the regression coefficient vector with the true value $\beta_0 = (\beta_{01}, \ldots, \beta_{0p})^T \in \mathbf{R}^p$. In this paper, we assume that x_i and ε_i are independent and β_0 is sparse in the sense that most of its elements are zeros. Let $y = (y_1, \ldots, y_n)^T \in \mathbf{R}^n$, $X = (x_1, \ldots, x_n)^T \in \mathbf{R}^{n\times p}$, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbf{R}^n$ be the response vector, the design matrix, and the noise vector, respectively. The relationship between y and X is given as

$$y = X\beta + \varepsilon. \tag{2}$$

If predictors define the notation X^*_S are uncorrelated, then SIS is expected to perform well. However, this condition is easily violated and may be inappropriate for high-dimensional data. By employing a factor model, H. Wang (2012) proposed a factor profiling operator $Q(Z_I) = I_n - Z_I(Z_I^T Z_I)^{-1} Z_I^T$ to eliminate the correlation of predictors and apply the SIS on the profiled data. $Z_I \in \mathbf{R}^{n\times d}$ is latent factor matrix of X, and d is the number of latent factors. Then factor profiling is as follows:

$$Q(Z_I)y = Q(Z_I)X\beta + Q(Z_I)\varepsilon. \tag{3}$$

In Equation (3), $Q(Z_I)y$ is the profiled response and the columns of $Q(Z_I)X$ are the profiled predictors.

2.2 | Spiked population model

We assume that the joint distribution of the design matrix $X \in \mathbf{R}^{n\times p}$ is Gaussian with zero mean and covariance matrix $\Sigma_p = \mathrm{cov}(x)$. $x \in \mathbf{R}^p$ is the random predictor vector. Define $Z^* \in \mathbf{R}^{n\times p}$ and $z^* \in \mathbf{R}^p$ respectively as

$$Z^* = X\Sigma_p^{-1/2},\ z^* = \Sigma_p^{-1/2}x. \tag{4}$$

The covariance matrix of z^* is an identity matrix obviously. Suppose that the spectral decomposition of Σ_p is given by $\Sigma_p = \sum_{j=1}^p l_j u_j^* u_j^{*T}$, where $l_1 \geq \ldots \geq l_p \geq 0$ and $u_1^*, \ldots, u_p^*$ form an orthonormal basis of $\mathbf{R}^p$. Consider a spiked population model as follows:

$$\begin{aligned} l_j &= \lambda_j + \sigma_0^2, && j = 1, \ldots, d, \\ l_j &= \omega_{j-d} + \sigma_0^2, && j = d+1, \ldots, d+m, \\ l_j &= \sigma_0^2, && j = d+m+1, \ldots, p, \end{aligned} \tag{5}$$

where $\lambda_1 \geq \ldots \geq \lambda_d > \omega_1 \geq \ldots \geq \omega_m > 0$ and σ_0^2 is a positive constant. We define a subscript j of l_j be a change point when $l_j / l_{j+1} \to \infty$ as $p \to \infty$. Then we define d be the biggest change point that

$$\frac{l_d / l_{d+1}}{\max_{1\leq j\leq p, j\neq d} l_j / l_{j+1}} \to \infty \ as\ p \to \infty. \tag{6}$$

Furthermore, we assume that d and m are fixed and they have true values with d_0 and m_0, respectively. Equation (6) means that the change point d is the number of large eigenvalues of Σ_p in the case where $l_{d+1}, \ldots, l_p$ are decreasing and sufficiently well separated from $l_1, \ldots, l_d$. Consequently, the maximum value of l_j / l_{j+1} would be expected to happen at $j = d$. Under a Gaussian assumption, the design matrix X can be expressed as

$$X = \sum_{j=1}^{d} \sqrt{\lambda_j} z_j u_j^{*T} + \sum_{k=1}^{m} \sqrt{\omega_k} z_{d+k} u_{d+k}^{*T} + \sigma_0^2 \Lambda, \tag{7}$$

where $z_1, \ldots, z_{d+m}$ are i.i.d. $N(0, I_n)$ vectors. Λ is viewed as a noise matrix that is an $n \times p$ matrix with i.i.d. $N(0, 1)$ entries and is independent of $z_1, \ldots, z_{d+m}$. In the analysis presented in this paper throughout, we use Equation (7) as the model for X. Define $Z_I = (z_I, \ldots, z_d) \in \mathbf{R}^{n\times d}$, $X_I^* = \sum_{j=1}^d \sqrt{\lambda_j} z_j u_j^{*T}$, $X_{II}^* = \sum_{k=1}^m \sqrt{\omega_k} z_{d+k} u_{d+k}^{*T}$, and $X_{III}^* = \sigma_0^2 \Lambda$. PIS can select d consistently by the maximum eigenvalue ratio criterion (MERC; Luo et al., 2009; Wang, 2012). Then factor profiling operator $Q(Z_I)$ is used to remove the effect of $z_1, \ldots, z_d$, as follows:

$$\begin{aligned} Q(Z_I)X &= Q(Z_I)(X_I^* + X_{II}^* + X_{III}^*) \\ &\approx X_{II}^* + X_{III}^*. \end{aligned} \tag{8}$$

From Equation (8), X_{II}^* is still in the profiled data after factor profiling. And it may lead to correlated profiled predictors in the case that population covariance matrix has eigenvalues with slow decay. Our simulation results of Examples 5 and 6 in Section 3 illustrate this phenomenon. To overcome this issue, we propose and apply a novel preconditioned profiling operator to the data, which guarantees the resulting predictors to be asymptotically uncorrelated. Consequently, by combining the preconditioned profiling operator and SIS, we propose a novel screening procedure named PPIS.

2.3 | Preconditioned profiled independence screening

According to the singular value decomposition, the $n \times p$ matrix X almost surely has n positive singular values (Fan & Lv, 2008; Klema & Laub, 1980). Let $\mu_1, \ldots, \mu_n$ be the n positive singular values such that $\mu_1 \geq \ldots \geq \mu_n > 0$. Therefore, there exist matrices $U = (u_1, \ldots, u_n) \in \mathbf{R}^{n\times n}$ and $V = (v_1, \ldots, v_n) \in \mathbf{R}^{p\times n}$ with $U^T U = V^T V = I_n$ and a diagonal matrix $D = \mathrm{diag}(\mu_1, \ldots, \mu_n) \in \mathbf{R}^{n\times n}$ such that

$$X = UDV^T, \tag{9}$$

where $u_i = (u_{1i}, \dots, u_{ni})^T \in \mathbf{R}^n$, $v_i = (v_{1i}, \dots, v_{pi})^T \in \mathbf{R}^p$ ($i = 1, \dots, n$), and I_n is the identity matrix of size n. We partition U and V respectively as $U = (U_{\mathrm{I}} \vdots U_{\mathrm{II}})$ and $V = (V_{\mathrm{I}} \vdots V_{\mathrm{II}})$, where $U_{\mathrm{I}} = (u_1, \dots, u_d) \in \mathbf{R}^{n\times d}$, $U_{\mathrm{II}} = (u_{d+1}, \dots, u_n) \in \mathbf{R}^{n\times(n-d)}$, $V_{\mathrm{I}} = (v_1, \dots, v_d) \in \mathbf{R}^{p\times d}$, and $V_{\mathrm{II}} = (v_{d+1}, \dots, v_n) \in \mathbf{R}^{p\times(n-d)}$. Similarly, we partition the diagonal matrix D into

$$D = \begin{bmatrix} D_{\mathrm{I}} & 0 \\ 0 & D_{\mathrm{II}} \end{bmatrix},$$

where $D_{\mathrm{I}} = \mathrm{diag}(\mu_1, \dots, \mu_d)$ and $D_{\mathrm{II}} = \mathrm{diag}(\mu_{d+1}, \dots, \mu_n)$. Consequently, Equation (9) reduces to

$$X = X_{\mathrm{I}} + X_{\mathrm{II}}, \tag{10}$$

where $X_{\mathrm{I}} = U_{\mathrm{I}} D_{\mathrm{I}} V_{\mathrm{I}}^T$ and $X_{\mathrm{II}} = U_{\mathrm{II}} D_{\mathrm{II}} V_{\mathrm{II}}^T$.

We define the preconditioned profiling operator as

$$F = U_{\mathrm{II}} D_{\mathrm{II}}^{-1} U_{\mathrm{II}}^T (I_n - U_{\mathrm{I}} U_{\mathrm{I}}^T) = U_{\mathrm{II}} D_{\mathrm{II}}^{-1} U_{\mathrm{II}}^T, \tag{11}$$

By applying the preconditioned profiling operator F to Equation (2), we have $Fy = FX\beta + F\varepsilon$, which can be further reduced to

$$\tilde{y} = \tilde{X}\beta + \tilde{\varepsilon}, \tag{12}$$

where $\tilde{y} = U_{\mathrm{II}} D_{\mathrm{II}}^{-1} U_{\mathrm{II}}^T y$, $\tilde{X} = U_{\mathrm{II}} V_{\mathrm{II}}^T$, and $\tilde{\varepsilon} = U_{\mathrm{II}} D_{\mathrm{II}}^{-1} U_{\mathrm{II}}^T \varepsilon$. As X and ε are assumed to be independent, $\tilde{X}$ and $\tilde{\varepsilon}$ are uncorrelated. Theorem 1 below indicates that our proposed transformation from X to $\tilde{X}$ leads to uncorrelated profiled predictors asymptotically, and its proof is in the Supporting Information.

Theorem 1. *Under Conditions 1–4 in the next section, for fixed i and j, if $\tilde{\Sigma} = \tilde{X}^T\tilde{X}$, then*

$$\frac{|\tilde{\sigma}_{ij}|}{|\tilde{\sigma}_{ii}|} = O_p\left(\sqrt{\frac{\log p}{n}}\right), \ 1 \le i < j \le p,$$

where $\tilde{\sigma}_{ij}$ is the (i, j) element of $\tilde{\Sigma}$.

Obviously, $F = FQ(\hat{Z}_{\mathrm{I}})$, where $Q(\hat{Z}_{\mathrm{I}}) = I_n - U_{\mathrm{I}} U_{\mathrm{I}}^T$ and $\hat{Z}_{\mathrm{I}} (\hat{Z}_{\mathrm{I}}^T \hat{Z}_{\mathrm{I}})^{-1} \hat{Z}_{\mathrm{I}}^T = U_{\mathrm{I}} U_{\mathrm{I}}^T$. It is noted that $U_{\mathrm{I}} U_{\mathrm{I}}^T$ is used to estimate $Z_{\mathrm{I}} (Z_{\mathrm{I}}^T Z_{\mathrm{I}})^{-1} Z_{\mathrm{I}}^T$. This estimator is discussed in Xia (2007), Wang and Xia (2008), and Wang (2012). Lemma 4 in the Supporting Information of this paper promises the consistency of the estimation and quantifies the accuracy under a spiked population model. On one hand, $Q(\hat{Z}_I)$ is a factor profiling operator that filters out the effects of the first d factors in model (7). On the other hand, from Theorem 1, we can see that F plays a role of a preconditioner that decorrelates the profiled predictors X_{II}. Similar to the Puffer transformation in Jia and Rohe (2012), the proposed operator F does not change the linear relationship in the model. It finds a proper way to "standardize" the high-dimensional profiled regressors without having to estimate its high-dimensional inverse covariance matrix. This makes a nice contribution to the literature on variable screening under correlated designs. Our PPIS procedure can then be obtained by applying SIS to the preconditioned and profiled data:

$$\hat{y} = Fy = U_{\mathrm{II}} D_{\mathrm{II}}^{-1} U_{\mathrm{II}}^T (I_n - U_{\mathrm{I}} U_{\mathrm{I}}^T) y, \tag{13}$$

$$\hat{X} = FX = U_{\mathrm{II}} D_{\mathrm{II}}^{-1} U_{\mathrm{II}}^T (I_n - U_{\mathrm{I}} U_{\mathrm{I}}^T) X. \tag{14}$$

More specifically, denote $\hat{X} = (\hat{x}_{(1)}, \dots, \hat{x}_{(p)}) \in \mathbf{R}^{n\times p}$. The PPIS procedure uses the following estimator of β_j as a measure for screening:

$$\hat{\beta_j} = (\hat{x}_{(j)}^T \hat{x}_{(j)})^{-1} (\hat{y}^T \hat{x}_{(j)}), j = 1, \dots, p. \tag{15}$$

As the same assumption in Wang (2012), to embody the sorting idea of screening, it is assumed further that the predictor indices have been appropriately relabelled so that $|\hat{\beta}_1| > \dots > |\hat{\beta}_p|$ without loss of generality. A candidate model can be represented as $\mathcal{S} = \{j_1, \dots, j_m\}$, which includes the j_ith column of X for every $j_i \in \mathcal{S}$, and $|\mathcal{S}|$ denotes the corresponding model size. The full model is $\mathcal{S}_{\mathrm{F}} = \{j : j = 1, \dots, p\}$, and the true model is $\mathcal{S}_T = \{j : \beta_{0j} \neq 0\}$. Here, a solution path is $\mathbf{S} = \{\mathcal{S}_k : k = 1, \dots, p\}$ with $\mathcal{S}_0 = \varnothing$ and $\mathcal{S}_k = \{1, \dots, k\}$ for $k = 1, \dots, p$. In addition, we denote the design matrix that corresponds to model $\mathcal{S}_k$ as $X(\mathcal{S}_k) \in \mathbf{R}^{n\times k}$. Hence, the solution path can be used to screen the predictors directly.

The following theorem indicates that our proposed PPIS is path consistent; that is, $\Pr\{\mathcal{S}_T \in \mathbf{S}\} \to 1$ as $n \to \infty$. The definitions of t, s, and h in the theorem can be found in Conditions 1 and 2 in the next section. A proof is given in the Supporting Information.

Theorem 2. *Under Conditions 1–4 in the next section, if $d = d_0$*

$$\max_{1\le j\le p} |\hat{\beta}_j - \beta_{0j}| = O_p\left(n^{-\frac{s-t}{2}} \vee \sqrt{\frac{\log p}{n}}\right) \text{ as } n \to \infty, \tag{16}$$

where $n^{-\frac{s-t}{2}} \vee \sqrt{\frac{\log p}{n}} = \max\{n^{-\frac{s-t}{2}}, \sqrt{\frac{\log p}{n}}\}$.

In practice, we use the following Bayesian information criterion (BIC)-type criterion to estimate the value of $|\mathcal{S}_T|$, which is used in Wang (2012). More specifically, we choose the model such that the following score is minimized:

$$\mathrm{BIC}^*(\mathcal{S}) = \log \mathrm{RSS}\,(\mathcal{S}) + (n^{-1}$$

where RSS $(\mathcal{S})$ is the residual sum of squares. In this paper, we consider the situation that β is highly sparse, and we assume that the value of $|\mathcal{S}_T|$ is less than the number of samples n. For every candidate model $\mathcal{S}_k$, we use the simple least square estimate

$$\hat{\beta}(\mathcal{S}_k)^{ls} = (X(\mathcal{S}_k)^T X(\mathcal{S}_k))^{-1} X(\mathcal{S}_k)^T y, \tag{18}$$

to compute the RSS $(\mathcal{S})$ in Equation (17) as

$$\text{RSS}\ (\mathcal{S}_k) = \|y - X(\mathcal{S}_k)\ \hat{\beta}(\mathcal{S}_k)^{ls}\|^2, \tag{19}$$

and

$$\text{BIC}^*(\mathcal{S}_k) = \log \text{RSS}\ (\mathcal{S}_k) + (n^{-1} \log p)|\mathcal{S}_k| \log n. \tag{20}$$

Therefore, the estimate of $|\mathcal{S}_T|$ is

$$|\hat{\mathcal{S}}_T| = \underset{1 \le k \le n}{\arg\min}\ \text{BIC}^*(\mathcal{S}_k), \tag{21}$$

and the estimate of β in Equation (2) is

$$\hat{\beta}(\hat{\mathcal{S}}_T)^{ls} = (X(\hat{\mathcal{S}}_T)^T X(\hat{\mathcal{S}}_T))^{-1} X(\hat{\mathcal{S}}_T)^T y. \tag{22}$$

2.4 | Conditions and assumptions

We use c_i and C_i to denote positive constants independent of the sample size n and the dimensionality p in this paper throughout. For an arbitrary matrix $A \in \mathbf{R}^{a_1 \times a_2}$, denote the Frobenius norm matrix A as $\|A\|^2 = \text{tr}(A^T A) = \text{tr}(AA^T)$. Let $e_j = (0, \ldots, 1, \ldots, 0)^T$ be a unit vector in $\mathbf{R}^p$ with the jth element being 1 and 0 elsewhere, $j = 1, \ldots, p$.

Due to a significant impact of ε's tail behaviour on the screening performance, Wang and Leng (2016) used a q-exponential tail condition as a characterization of different distribution families. The tail condition is useful for the theoretical proofs in this paper and presented in the following definition.

Definition 1. A zero mean distribution $\mathcal{F}$ is said to have a q-exponential tail, if any $N \ge 1$ independent random variables $\varepsilon_i \sim \mathcal{F}$ satisfy that for any $a \in \mathbf{R}^N$ with $\|a\|_2 = 1$, the following inequality holds:

$$\Pr\left(|\sum_{i=1}^{N} a_i \varepsilon_i| > \tau\right) \le \exp\ (1 - q(\tau)) \tag{23}$$

for any $\tau > 0$ and some function $q(\cdot)$.

In this paper, we assume that $\varepsilon_i \sim N(0, \sigma^2)$. As shown in Wang and Leng (2016), with the classical bound on the Gaussian tail, the Gaussian distribution admits a square-exponential tail in that $q(\tau) = \tau^2/2$. The following conditions are necessary in the theoretical proofs for the theorems in this paper.

Condition 1: There exists a specification for model (7), such that the vectors z_i $(i = 1, \ldots, d+m)$ are i.i.d. $N(0, I_n)$ vectors and Λ is an $n \times p$ matrix with i.i.d. $N(0,1)$ entries and is independent of $z_1, \ldots, z_{d+m}$. Furthermore, we assume that the eigenvalues of the covariance matrix Σ_p are satisfied for Equation (5). Especially, σ_0^2 in Equation (5) is some positive constant, and there are constants $0 \le t < s \le 1$, $c_1 > 0, c_2 > 0$, and $c_3 > 0$ such that

$$\lambda_1 \le c_1 n,\ \lambda_d \ge c_2 n^s,\ \text{and}\ \omega_1 \le c_3 n^t. \tag{24}$$

Condition 2: The true model size $|\mathcal{S}_T|$ is fixed, whereas the sample size $n \to \infty$. Moreover, we assume that $\log p = c_4 n^h$ for some $c_4 > 0$ and $0 < h < 1$.

Condition 3: The random error ε in model (2) is normally distributed with mean zero and standard deviation σ and is independent of X.

Condition 4: The transformed z^* in Equation (4) has a spherically symmetric distribution, and there exist some $c_5 > 1$ and $C_1 > 0$ such that $\Pr\{\lambda_{\max}(p^{-1}Z^*Z^{*T}) > c_5$ *or* $\lambda_{\min}(p^{-1}Z^*Z^{*T}) < 1/c_5\} \le e^{-C_1 n}$, where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ are the largest and smallest nonzero eigenvalues of a matrix, respectively.

It is noteworthy that (a) in Condition 1, s and t control the maximum value of eigenvalue ratio. When $s = 1$ and $t = 0$, the setting of eigenvalues of the population covariance matrix in Condition 1 degenerates to a setting of a factor model. Condition 1 allows slow spectrum decay of the population covariance matrix so that it is weaker than assumption A2 in Wang (2012). Condition 2 is similar to assumption A3 in Wang (2012) in that it allows the predictor dimension to be much larger than the sample size. (b) Condition 3 gives the tail behaviour of ε, and its Gaussian tail can be bounded by Equation (23). (c) Condition 4 is the same as assumption A1 in Wang and Leng (2016). It is similar to but weaker than the concentration property in Fan and Lv (2008). The proof in Fan and Lv (2008) can be directly applied to show that Condition 4 is true for the Gaussian distribution.

2.5 | Detecting the biggest change point d

In essence, the true value d_0 of the biggest change point d is unknown in real practice, and it has to be estimated based on data. Under a spiked population model, we propose a novel method named as maximum modified eigenvalue ratio criterion (MMERC), which is more stable than the commonly used MERC. And we also provide a theoretical justification by showing that the consistent estimator of d_0 can be obtained by using MMERC. Recall that $\mu_1 \ge \mu_2 \ge \ldots \ge \mu_n > 0$ be the nonzero singular values of X from Equation (9). The MERC finds the estimator of d_0 by

$$\hat{d} = \arg\max_{1 \le i \le n-1} \frac{\mu_i^2}{\mu_{i+1}^2}. \tag{25}$$

During our finite sample simulation study, we found interestingly that for a given dataset, the MERC estimator $\hat{d}$ could substantially vary when we repeat the experiment with different subsampling data from the same dataset. To illustrate this phenomenon, we use the gasoline dataset introduced in Section 1. Each time, we use Monte Carlo sampling (Xu, Liang, & Du, 2004) without replacement to select $s = 48$ observations randomly from the 60 samples of gasoline dataset to form a design matrix $X^{(s)} \in \mathbb{R}^{s \times p}$, and we let $\mu_1^{(s)} \ge \mu_2^{(s)} \ge \ldots \ge \mu_s^{(s)} > 0$ be the nonzero singular values of $X^{(s)}$. We then use MERC to get $\hat{d}$ of $X^{(s)}$. We repeat this process 100 times, and the possible values for $\hat{d}$ are solely 1 and 4 with frequencies being 26 and 74, respectively. Here, the chosen s is 80% of the total number of samples, and it is large enough to report different $\hat{d}$s in this illustration.

To obtain a stable estimator of d_0, we propose the following MMERC estimator of d_0:

$$\hat{d}_M = \arg\max_{1 \le i \le n-1} \frac{i\mu_i^2}{(i+1)\mu_{i+1}^2}. \tag{26}$$

The following theorem provides a theoretical justification for our proposed MMERC, and the proof is provided in the Supporting Information. To compare the performance with MERC, we conduct the same finite sample simulation study on the basis of the gasoline dataset using our proposed MMERC. Interestingly, all the 100 repetitions yield the same estimate (i.e., $\hat{d}_M$) being 4. The theorem below proves that MMERC is a consistent ratio, and its proof can be found in the Supporting Information.

Theorem 3. *Under Conditions 1–4,* $\Pr\{\hat{d}_M = d_0\} \to 1$ *as* $n \to \infty$.

According to Theorem 3, d_0 can be estimated consistently by MMERC. Therefore, we use MMERC to detect the change point d and focus on the decorrelated performance of PPIS via all the simulation studies and real examples in the next section.

3 | NUMERICAL STUDIES

To evaluate the finite sample performance of the proposed method, we compare the proposed PPIS with SIS, PIS, and high-dimensional ordinary least squares projection (HOLP) via six simulation experiments and two real examples. HOLP is a screening method that uses a preconditioner to guarantee the sure screening property and give a consistent variable screening without strong correlation assumptions (Wang & Leng, 2016). For each method, we (a) compute the solution path, (b) use the simple least squares estimator in Equation (18) to evaluate the estimation accuracy of variable screening, and (c) use the BIC-type criterion in Section 2.5 to select the model size.

3.1 | Simulation study

In Examples 1–3, we follow the settings in Fan and Lv (2008), Wang (2012), and Tibshirani (1996) and set $n = 100$ or 300, $p = 1,000$. Example 4 discusses the situation of collinear among predictors. Examples 5 and 6 discuss spiked population models. The results are evaluated over 100 replications in all of the six examples.

Example 1

$$y = \beta x_{(1)} + \beta x_{(2)} + \beta x_{(3)} - 3\beta\sqrt{\varphi} x_{(4)} + \varepsilon,$$

where $\varepsilon \sim N(0, I_n)$, and $(x_1, \ldots, x_{ip})^T$ are generated from a multivariate normal distribution $N(0, \Sigma)$ independently for $i = 1, \ldots, n$. The population covariance matrix $\Sigma = (\Sigma_{ij})_{j,k=1}^{p}$ satisfies $\Sigma_{jj} = 1$ $(j = 1, \ldots, p)$ and $\Sigma_{jk} = \varphi$ $(j \ne k)$, except $\Sigma_{4,k} = \Sigma_{j,4} = \sqrt{\varphi}$ $(j, k \ne 4)$, and consequently, $x_{(4)}$ is marginally uncorrelated with y at the population level. Here, $\beta = 5$ and $\varphi = 0.5, 0.95$ are used to investigate the performance of the four variable screening methods.

Example 2

$$y = \beta x_{(1)} + \beta x_{(2)} + \beta x_{(3)} - 3\beta\sqrt{\varphi} x_{(4)} + \beta x_{(5)} + \varepsilon,$$

with the population covariance matrix of X being described in Example 1 except that $\Sigma_{5j} = \Sigma_{j5} = 0$ for any $j \ne 5$, that is, $x_{(5)}$ is relevant in model whereas it is uncorrelated with any other predictors.

Example 3 We set $d = 3$, and the latent factor $U_{\alpha i} \in \mathbb{R}^3 (i = 1, \ldots, n)$ is generated from a three-dimensional normal random vector. A sample of predictors $x_i \in \mathbb{R}^p$ is then simulated as $x_i = BU_{\alpha i} + \check{x}_i$, where $B = (b_{jk}) \in \mathbb{R}^{p \times 3}$, $b_{jk} \sim N(0, 1)$, and $\check{x}_i$ follows a p-dimensional normal distribution with $E(\check{x}_{ij}) = 0$ and $\mathrm{cov}(\check{x}_{ij_1}, \check{x}_{ij_2}) = 0.95^{|j_1 - j_2|}$. Here, y_i is simulated according to $y_i = 3x_{i1} + 1.5x_{i2} + 2x_{i3} + \varepsilon_i$. Finally, ε_i is generated according to $\varepsilon_i = U_{\alpha i}^T \alpha_0 + \check{\varepsilon}_i$, where $\alpha_0 = 0.8\sigma_\varepsilon(3^{-1/2}, 3^{-1/2}, 3^{-1/2})^T \in \mathbb{R}^3$, and the variance of $\check{\varepsilon}_i$ is $\check{\sigma}_\varepsilon^2 = 0.36\sigma^2$. Given $X = (x_1, \ldots, x_n)^T$ and $\beta_0 = (3, 1.5, 2)$, σ^2 is particularly selected so that the signal-to-noise ratio, $\mathrm{var}(X\beta_0)/\sigma^2$, is 1, 2, or 5. For this example, the PIS fails to select the true model, whereas the assumption that the columns of $\check{X}$ is uncorrelated is not satisfied.

Example 4

$$y = \beta x_{(1)} + \beta x_{(2)} + \beta x_{(3)} - 3\beta\sqrt{\varphi} x_{(4)} + \beta x_{(5)} + \varepsilon,$$

with the population covariance matrix of X being described in Example 2 except that $x_{(6)} = 0.8x_{(5)} + \varepsilon_x$, where $\varepsilon_x \sim N(0, \sigma_x^2)$. Here, $x_{(5)}$ and $x_{(6)}$ are collinear, but $x_{(6)}$ is a noise variable in this example. We set $\sigma_x = 0.05, 0.1, 0.2$, $n = 100$, $p = 1,000$, and $\varphi = 0.95$.

Example 5 A spiked population model in Equation (7) is considered in this simulation study. We set $n = 200$, $p = 1{,}000$, $d = 3$, $m = 40$, and $\sigma_0^2 = 1$; and $z_j \in \mathbf{R}^n$ $(j = 1, \ldots, d+m)$ is generated from an n-dimensional standard normal random vector. The design matrix $X \in \mathbf{R}^{n\times p}$ is then simulated as

$$X = \sum_{j=1}^{d} z_j b_j^T + \sum_{j=d+1}^{d+m} n^{\frac{-(j-d+9)}{m+10}} z_j b_j^T + \tilde{x}, \tag{27}$$

where $b_j = (b_{jk}) \in \mathbf{R}^{p\times 1}$, $b_{jk} \sim N(0,1)$, $\tilde{x} = (\tilde{x}_1, \ldots \tilde{x}_n)^T$, and $\tilde{x}_i$ follows a p-dimensional normal distribution in that $E(\tilde{x}_{ij}) = 0$ and $\text{cov}(\tilde{x}_{ij_1}, \tilde{x}_{ij_2}) = I_p$. Here, y_i is simulated according to $y_i = 3x_{i1} + 3x_{i2} + 3x_{i3} + \varepsilon_i$. Finally, $\varepsilon_i \sim N(0, \sigma^2)$. Given $X = (x_1, \ldots, x_n)^T$ and $\beta_0 = (3,3,3)$, σ^2 is particularly selected so that the signal-to-noise ratio $\text{var}(X\beta_0)/\sigma^2 = 5$.

Example 6 With a spiked population model, we consider the endogeneity problem in this simulation study, which means that the residual might be correlated with the predictor. We use the spiked population model, which is described in Example 5 except that ε_i is generated according to $\varepsilon_i = \sum_{j=1}^{d+m} z_j + \check{\varepsilon}_i$, where $\check{\varepsilon}_i \sim N(0, 0.36\sigma^2)$. Given $X = (x_1, \ldots, x_n)^T$ and $\beta_0 = (3,3,3)$, σ^2 is particularly selected so that the signal-to-noise ratio $\text{var}(X\beta_0)/\sigma^2 = 5$.

For each method and simulation setting, the following measures are adopted to evaluate the performance of variable screening (Cho & Fryzlewicz, 2012): the number of false negatives (FNs; i.e., the number of relevant variables incorrectly identified as irrelevant), the number of false positives (FPs; i.e., the number of irrelevant variables incorrectly identified as relevant), and the L2 distance $\|\beta_0 - \hat{\beta}\|_2^2$. Tables 1--4 summarize the averaged FN, averaged FP, averaged L2, and total number of times that a specific relevant variable is being correctly selected over 100 repetitions.

Under the multicollinearity settings in Example 1 (see Table 1), both SIS and HOLP perform poorly when the sample size is small. In particular, we find that SIS always and HOLP often miss the relevant variable $x_{(4)}$. These are not surprising observations for SIS because $x_{(4)}$ has no marginal correlation with y, although SIS is a method based on marginal correlation estimation. Therefore, increment of the sample size does not improve the performance of SIS. Performance of HOLP improves when the sample size increases. Among the four methods under consideration, PIS and PPIS have satisfactory performance. In particularly, PPIS has the best performance under all the settings being considered in the sense that it

TABLE 1 Averaged FN, FP, FN + FP, and L2 and the total number of times for each relevant variable being selected for Examples 1 and 2

	n	φ	Methods	FN	FP	FN + FP	L2	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$
Example 1	$n = 100$	0.5	SIS	3.04	0.10	3.14	168.93	43	22	31	0	
			HOLP	2.04	0.31	2.35	114.49	52	58	54	32	
			PIS	0.27	0.03	0.30	15.28	94	94	94	91	
			PPIS	0.21	0.09	0.30	11.91	94	98	94	93	
		0.95	SIS	3.43	0.48	3.91	286.10	18	22	17	0	
			HOLP	2.25	0.04	2.29	215.10	49	51	50	25	
			PIS	0.45	0.07	0.52	43.25	91	91	87	86	
			PPIS	0.34	0.08	0.42	33.43	92	93	92	89	
	$n = 300$	0.5	SIS	1.92	2.49	4.41	152.18	73	61	74	0	
			HOLP	0.26	1.43	1.69	26.42	99	96	97	82	
			PIS	0.06	0.32	0.38	6.09	100	99	99	96	
			PPIS	0.05	0.27	0.32	4.53	100	99	99	97	
		0.95	SIS	2.75	1.18	3.93	276.75	47	41	37	0	
			HOLP	0.31	0.27	0.58	40.99	94	95	95	85	
			PIS	0.13	0.20	0.33	23.70	98	100	98	91	
			PPIS	0.13	0.12	0.25	19.00	99	98	97	93	
Example 2	$n = 100$	0.5	SIS	3.85	0	3.85	184.53	5	5	5	0	100
			HOLP	2.30	0.76	3.06	125.09	54	54	54	30	78
			PIS	0.32	0.62	0.94	16.98	91	92	94	94	97
			PPIS	0.15	0.69	0.84	9.65	97	96	98	96	98
		0.95	SIS	4.00	0	4.00	288.80	0	0	0	0	100
			HOLP	1.99	0.49	2.48	205.82	56	55	58	32	100
			PIS	4.09	0.09	4.18	313.11	1	1	0	89	0
			PPIS	0.81	0.42	1.23	62.82	76	74	76	93	100
	$n = 300$	0.5	SIS	2.16	1.83	3.99	156.01	66	57	61	0	100
			HOLP	0.24	1.17	1.41	29.91	98	100	99	79	100
			PIS	0.08	0.52	0.60	9.81	100	100	99	93	100
			PPIS	0.05	0.32	0.37	7.12	100	100	100	95	100
		0.95	SIS	3.98	0.02	4.00	288.56	1	0	1	0	100
			HOLP	0.36	0.35	0.71	43.88	93	93	94	84	100
			PIS	0.71	1.05	1.76	46.67	80	78	73	98	100
			PPIS	0.15	0.12	0.27	19.34	97	98	97	93	100

Abbreviations: FN, false negative; FP, false positive; HOLP, high-dimensional ordinary least squares projection; PIS, profiled independence screening; PPIS, preconditioned PIS; SIS, sure independence screening.

TABLE 2 Averaged FN, FP, FN + FP, and L2 and the total number of times for each relevant variable being selected for Example 3

n	$\text{var}(X\beta_0)/\sigma_\varepsilon^2$	Methods	FN	FP	FN + FP	L2	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$
$n = 100$	1	SIS	2.45	0.60	3.05	30.27	23	15	17
		HOLP	1.68	0.80	2.48	25.66	44	46	42
		PIS	1.76	1.14	2.90	30.49	44	45	35
		PPIS	1.73	1.04	2.77	27.37	41	44	42
	2	SIS	2.14	0.52	2.66	21.82	38	20	28
		HOLP	1.25	1.01	2.26	24.73	58	57	60
		PIS	1.63	1.60	3.23	28.03	48	44	45
		PPIS	1.37	1.16	2.53	24.28	51	55	57
	5	SIS	1.92	0.57	2.49	17.09	54	22	32
		HOLP	0.75	0.93	1.68	11.58	77	77	71
		PIS	1.29	1.63	2.92	16.41	61	62	48
		PPIS	0.83	1.05	1.88	12.77	74	75	68
$n = 300$	1	SIS	2.23	1.12	3.35	22.55	40	18	19
		HOLP	1.52	2.53	4.05	23.99	59	52	37
		PIS	1.38	2.21	3.59	43.54	56	51	55
		PPIS	1.45	2.24	3.69	21.88	61	52	42
	2	SIS	1.78	1.44	3.22	19.11	59	33	30
		HOLP	0.69	1.34	2.03	14.26	82	79	70
		PIS	1.12	2.45	3.57	21.69	72	58	58
		PPIS	0.66	1.34	2.00	15.47	83	79	72
	5	SIS	1.37	1.97	3.34	15.55	71	39	53
		HOLP	0.17	0.85	1.02	6.77	99	92	92
		PIS	0.91	2.86	3.77	12.89	77	66	66
		PPIS	0.18	0.93	1.11	6.57	96	95	91

Abbreviations: FN, false negative; FP, false positive; HOLP, high-dimensional ordinary least squares projection; PIS, profiled independence screening; PPIS, preconditioned PIS; SIS, sure independence screening.

TABLE 3 Averaged FN, FP, FN + FP, and L2 and the total number of times for each relevant variable and the noise variable $x_{(6)}$ being selected for Example 4

σ_x	Methods	FN	FP	FN + FP	L2	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$
0.05	SIS	4.27	0.27	4.54	305.67	0	0	0	0	73	27
	HOLP	2.17	0.60	2.77	214.82	57	57	54	33	82	43
	PIS	4.05	0.05	4.10	313.07	0	0	1	94	0	0
	PPIS	0.96	0.73	1.69	79.45	77	77	74	93	83	62
0.1	SIS	4.05	0.05	4.10	291.96	0	0	0	0	95	5
	HOLP	2.07	0.42	2.49	218.49	54	58	59	27	95	28
	PIS	4.03	0.03	4.06	312.95	0	1	0	96	0	0
	PPIS	0.75	0.60	1.35	65.66	80	81	78	89	97	52
0.2	SIS	4.01	0.01	4.02	289.39	0	0	0	0	99	1
	HOLP	2.12	0.27	2.39	217.93	52	54	54	28	100	8
	PIS	4.06	0.06	4.12	316.28	1	1	1	91	0	0
	PPIS	0.80	0.26	1.06	66.50	76	74	77	93	100	15

Abbreviations: FN, false negative; FP, false positive; HOLP, high-dimensional ordinary least squares projection; PIS, profiled independence screening; PPIS, preconditioned PIS; SIS, sure independence screening.

generally has the smallest averaged FNs, FPs, and L2s and the largest hit rates for individual relevant variables. This supports our theory that PPIS is more suitable for variable screening for high-dimensional models under multicollinearity.

For Example 2 (see Table 1), it is noted that $x_{(5)}$ is uncorrelated with other predictors and can therefore be successfully selected by SIS. However, SIS still poorly misses those relevant variables under the multicollinearity issue. Similar to those observations in Example 1, HOLP still does not perform satisfactorily especially for small-sample settings. PIS performs satisfactorily when $\varphi = 0.5$. However, it has the worst performance when $\varphi = 0.95$ and the sample sizes are small. Again, our proposed PPIS yields the best performance under all settings being considered in Example 2. It is noteworthy that PPIS is also more suitable to cope with situations in which some of the true variables are marginally uncorrelated with y in the linear model.

For Example 3 (see Table 2), it should be noticed that the profiled predictors produced by PIS are correlated. Therefore, both SIS and PIS are inferior to PPIS and HOLP. As PPIS can produce uncorrelated profiled predictors by the alliance between preconditioning and factor profiling, it performs satisfactorily. For all the methods being considered, it is noteworthy that the larger the signal-to-noise ratio, the better the performance.

The results of Example 4 are listed in Table 3. It can be seen that PPIS has the least FN + FP and the best performance for important variable screening in every situation. It is noteworthy that PPIS can select $x_{(5)}$ and reject $x_{(6)}$ more successfully when σ_x is larger.

	Methods	FN	FP	FN + FP	L2	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$
Example 5	SIS	2.17	0.58	2.75	23.63	26	28	29
	HOLP	0.21	0.32	0.53	2.25	89	95	95
	PIS	1.03	1.29	2.32	10.82	65	69	63
	PPIS	0.26	0.36	0.62	2.83	86	94	94
Example 6	SIS	2.28	0.48	2.76	25.44	21	26	25
	HOLP	0.63	0.22	0.85	6.82	82	76	79
	PIS	1.99	0.95	2.94	21.25	30	31	40
	PPIS	0.66	0.34	1.00	7.06	83	76	75

TABLE 4 Averaged FN, FP, FN + FP, and L2 and the total number of times for each relevant variable being selected for Examples 5 and 6

Abbreviations: FN, false negative; FP, false positive; HOLP, high-dimensional ordinary least squares projection; PIS, profiled independence screening; PPIS, preconditioned PIS; SIS, sure independence screening.

Dataset	Method	RMSEP	Selected wavelengths (selected times)
Corn dataset	SIS	0.2916	2,478 (64).
	HOLP	0.0003	1,906 (37), 1,908 (100), 2,108 (100)
	PIS	0.2562	1,416 (33), 1,418 (69), 1,420 (52)
	PPIS	0.0003	1,906 (38), 1,908 (100), 2,108 (100)
Gasoline dataset	SIS	0.6477	1208 (97)
	HOLP	0.4059	1,218 (84), 1,224 (58)
	PIS	1.0796	1,218 (36)
	PPIS	0.3836	1,218 (86), 1,224 (31), 1,414 (35), 1,416 (56)

TABLE 5 Average RMSEP, selected wavelengths, and their total selected times for two real datasets

Note. Predictors that are selected more than 30 times are listed. Abbreviations: HOLP, high-dimensional ordinary least squares projection; PIS, profiled independence screening; PPIS, preconditioned PIS; RMSEP, root-mean-square prediction error; SIS, sure independence screening.

For Examples 5 and 6 (see Table 4), both SIS and PIS perform poorly in the two examples. Under the spiked population model with slow spectrum decay of population covariance matrix, PIS could not always guarantee to obtain uncorrelated profiled predictors, and therefore, some relevant variables are missing in the model. Even worse, as described in Example 6, if there is an endogeneity problem in the model, and SIS and PIS will be misleading. Both PPIS and HOLP perform better than PIS and SIS in the two examples.

3.2 | Application to real data

In this section, we apply SIS, HOLP, PIS, and PPIS to analyse two near-infrared (NIR) spectral datasets. NIR spectra are an important type of data in chemometrics. It is of great challenge to select important predictors for NIR spectrum research. It is well-known that the predictors of NIR spectral datasets are always high dimensional and high correlated. This character coincides with our proposed approach and other compared methods to distinguish their different performance as big as possible. Therefore, to obtain the comprehensive performance of our proposed method, we use two NIR spectral datasets to report the study results. Below are brief descriptions of the two datasets:

Corn dataset The corn dataset consists of 80 samples, downloaded from http://software.eigenvector.com/data/index.html. The response variable is the corn moisture values, and the predictors are the wavelength intensities at 700 points ranging from 1,100 to 2,498 nm at 2-nm intervals. The design matrix is of dimension 80×700.

Gasoline dataset The gasoline dataset (Kalivas, 1997) is a NIR spectral dataset with NIR spectra and octane numbers of 60 gasoline samples. The NIR spectra were measured from 900 to 1,700 nm in 2-nm intervals, giving 401 wavelengths (variables).

For each dataset, we used Monte Carlo sampling without replacement to select 80% of the original samples as the training dataset and the rest 20% samples as the testing dataset. SIS, HOLP, PIS, and PPIS are used to select the best model on the basis of the training dataset. This process is repeated independently 100 times. The prediction accuracies of these methods are measured by the root-mean-square prediction error (RMSEP) on the basis of the testing data with size, say N. Here, the RMSEP is computed as $\sqrt{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2/N}$, where y_is are the observations of the response variable in the testing dataset, whereas $\hat{y}_i$s are the predicted values of y_is for any selected model. In essence, we did not know the real model of real data. And we could not say which predictor is the true variable. In this paper, we use RMSEP to represent the performance of prediction. And we think the best model must have the smallest RMSEP.

The results of the two real examples are summarized in Table 5. Here, we report in Table 5 those predictors that are selected more than 30 times and their selected times. It is observed that PPIS produces the smallest RMSEPs in both datasets. It is interesting to note that the selected predictors and RMSEPs are similar for PPIS and HOLP in both two datasets. In particular, the selected predictors of PPIS and HOLP have important chemical meaning in the corn dataset. Briefly, the 1,906-, 1,908-, and 2,108-nm wavelengths are in the region of water absorption and the combination of the O–H areas, which can be regarded as an important predictor for the response variable (i.e., corn moisture; Huang et al., 2012). For the gasoline dataset, 1,218- and 1,224-nm wavelengths are in the spectral region, which are the most useful for the determination of

paraffin and isoparaffin concentrations, and they could be correlated to the response variable (i.e., octane number; Maggard, 1994). In conclusion, we observe that PPIS is a reliable variable screening procedure for high-dimensional datasets with a multicollinearity issue. Most importantly, it gives the smallest prediction errors and produces more reasonable results for the two NIR spectral datasets.

4 | CONCLUSION

In this article, we propose a so-called PPIS approach for selecting variables for high-dimensional linear regression models with highly correlated predictors. Our proposed simple preconditioning and factor profiling procedures are shown to successfully remove the multicollinearity among the (profiled) predictors. According to our simulation studies, although the famous PIS may perform satisfactorily for situations in which the correlations are low among the predictors, it may not be useful for heavy multicollinearity situations, as its profiled predictors may still be correlated. Compared with PIS and other existing approaches, our proposed PPIS performs very well when the predictors are highly correlated in high-dimensional settings. The good performance of the PPIS in the two real data analyses also indicates that our proposed method could be a very good alternative for variable screening task for datasets with high-dimensional and highly correlated predictors.

One may be sceptical if the profiled predictors from PIS are still correlated and preconditioning is good enough to decorrelate the predictors: Is it necessary for PPIS to use both the factor profiling and preconditioning to achieve the uncorrelated predictors asymptotically? To indicate why factor profiling is still necessary, we consider in the simulation studies of Examples 1 and 2, which were adopted in Fan and Lv (2008) and Cho and Fryzlewicz (2012). It is noteworthy to point out that in both examples, a relevant variable in the model has no marginal correlation with the response variable. As a result, SIS and HOLP are unable to select the true variables correctly, whereas the sorting of SIS or HOLP depends on the marginal correlation between each predictor and the response variable. This interesting phenomenon is due to the latent factor structure of the predictors, and the factor profiling can successfully eliminate the effect from the latent factor structure. This is indeed the main difference between our method and other preconditioning methods.

DATA ACCESSIBILITY

The corn dataset is available at http://software.eigenvector.com/data/index.html, and the gasoline dataset is available in "pls" R package at CRAN.

ACKNOWLEDGEMENTS

This research is financially supported by the National Natural Science Foundation of China (Grants 11731011 and 11931014), a FDS research grant under the University Grants Committee (UGC) of the Hong Kong Special Administrative Region (Project No. UGC/FDS14/P02/18), National Social Science Foundation of China (Grant 17BTJ019), and Hunan Provincial Social Science Foundation of China (Grant 16YBA367). The authors are grateful to Dr. Guo Shaojun for his beneficial discussion that helped us improve the article substantially.

ORCID

Hong Wang https://orcid.org/0000-0002-6938-9507

REFERENCES

Baik, J., & Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, *97*(6), 1382–1408.

Cho, H., & Fryzlewicz, P. (2012). High dimensional variable selection via tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*(3), 593–622.

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

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

Fan, J., & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, *32*(3), 928–961.

Fan, J., Samworth, R., & Wu, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *Journal of Machine Learning Research*, *10*(Sep), 2013–2038.

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

Fu, G.-H., Xu, Q.-S., Li, H.-D., Cao, D.-S., & Liang, Y.-Z. (2011). Elastic net grouping variable selection combined with partial least squares regression (EN-PLSR) for the analysis of strongly multi-collinear spectroscopic data. *Applied Spectroscopy*, *65*(4), 402–408.

Huang, X., Xu, Q.-S., & Liang, Y.-Z. (2012). PLS regression based on sure independence screening for multivariate calibration. *Analytical Methods*, *4*(9), 2815–2821.

Ji, P., & Jin, J. (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *The Annals of Statistics*, *40*(1), 73–103.

Jia, J., & Rohe, K. (2012). Preconditioning to comply with the irrepresentable condition. ArXiv Preprint ArXiv:1208.5584.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, *29*(2), 295–327.

Kalivas, J. H. (1997). Two data sets of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, *37*(2), 255–259.

Ke, T., Jin, J., & Fan, J. (2014). Covariance assisted screening and estimation. *The Annals of Statistics*, *42*(6), 2202–2242.

Klema, V., & Laub, A. (1980). The singular value decomposition: Its computation and some applications. *IEEE Transactions on Automatic Control*, *25*(2), 164–176.

Li, G., Peng, H., Zhang, J., & Zhu, L. (2012). Robust rank correlation based screening. *The Annals of Statistics*, *40*(3), 1846–1877.

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

Luo, R., Wang, H., & Tsai, C.-L. (2009). Contour projected dimension reduction. *The Annals of Statistics*, *37*(6B), 3743–3778.

Maggard, S. M. (1994). Near infrared analysis of piano constituents and octane number of hydrocarbons. Google Patents. US Patent 5,349,188.

Rothman, A. J., Levina, E., & Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, *104*(485), 177–186.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *58*(1), 267–288.

Wang, H. (2012). Factor profiled sure independence screening. *Biometrika*, *99*(1), 15–28.

Wang, X., Dunson, D. B., & Leng, C. (2016). Decorrelated feature space partitioning for distributed sparse regression. In *Advances in Neural Information Processing Systems*, *29*, 802–810.

Wang, X., & Leng, C. (2016). High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*(3), 589–611.

Wang, H., & Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association*, *103*(482), 811–821.

Witten, D. M., & Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, *8*(1), 1–27.

Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics*, *35*(6), 2654–2690.

Xu, Q.-S., Liang, Y.-Z., & Du, Y.-P. (2004). Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *18*(2), 112–120.

Yu, H., Jiang, S., & Land, K. C. (2015). Multicollinearity in hierarchical linear models. *Social Science Research*, *53*, 118–136.

Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, *7*, 2541–2563.

Zhu, L.-P., Li, L., Li, R., & Zhu, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, *106*(496), 1464–1475.

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

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

Zou, H., & Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, *37*(4), 1733–1751.

How to cite this article: Zhao N, Xu Q, Tang M-L, Jiang B, Chen Z, Wang H. High-dimensional variable screening under multicollinearity. *Stat*. 2020;9:e272. https://doi.org/10.1002/sta4.272