Statistica Sinica **32** (2022), 1881-1909
doi:https://doi.org/10.5705/ss.202020.0495

SEMI-STANDARD PARTIAL COVARIANCE VARIABLE SELECTION WHEN IRREPRESENTABLE CONDITIONS FAIL

Fei Xue and Annie Qu

Purdue Univeristy and University of California Irvine

Abstract: Traditional variable selection methods could fail to be sign consistent when irrepresentable conditions are violated. This is especially critical in high-dimensional settings when the number of predictors exceeds the sample size. In this paper, we propose a new semi-standard partial covariance (SPAC) approach that is capable of reducing the correlation effects from other covariates, while fully capturing the magnitude of the coefficients. The proposed SPAC is effective in choosing covariates that have direct effects on the response variable, while eliminating predictors that are not directly associated with the response, but are highly correlated with the relevant predictors. We show that the proposed SPAC method with the Lasso penalty or the smoothly clipped absolute deviation (SCAD) penalty possesses strong sign consistency in high-dimensional settings. Numerical studies and a post-traumatic stress disorder data application confirm that the proposed method outperforms the existing Lasso, adaptive Lasso, SCAD, Peter–Clark-simple algorithm, and factor-adjusted regularized model selection methods when the irrepresentable conditions fail.

Key words and phrases: Irrepresentable condition, Lasso, model selection consistency, partial correlation, smoothly clipped absolute deviation.

1. Introduction

Variable selection is an important model-building tool for selecting covariates relevant to the response variable, which is fundamental for the construction of a sparse model when the number of relevant covariates is much smaller than the total number of observed covariates. This is especially crucial under high dimensionality, where the number of covariates far exceeds the number of observations. For high-dimensional data, traditional regularization variable selection methods (Tibshirani (1996); Fan and Li (2001); Zou and Hastie (2005); Yuan and Lin (2006); Zou (2006); Candes and Tao (2007); Zhang (2010)) are effective in achieving model selection and parameter estimation simultaneously

Corresponding author: Annie Qu, Department of Statistics, University of California Irvine, Irvine, CA 92697, USA. E-mail: aqu2@uci.edu.

under irrepresentable conditions (Zhao and Yu (2006); Fan and Lv (2011); Kim, Choi and Oh (2008)), which assume that the correlations between relevant and irrelevant covariates are relatively weak compared with those between relevant covariates.

However, the irrepresentable conditions could fail, regardless of whether or not the dimension is high. For example, in a mediation analysis seeking to identify mediators that transmit effects from an exposure factor to an outcome variable, spurious mediators (irrelevant covariates) could be strongly correlated with the exposure factor and the true mediators (relevant covariates) (Jérolon et al. (2021); Chén et al. (2018); Imai and Yamamoto (2013)). Although modified model selection methods have been proposed that incorporate strongly correlated covariates, they either do not possess variable selection consistency (Wang and Wang (2014); Maier and Rodríguez-Salas (2017); Hilafu and Yin (2017); Bühlmann et al. (2013)), or they impose a more restrictive condition, such as knowing the true number of relevant covariates (Javanmard and Montanari (2013)). In particular, several existing methods (Sharma, Bondell and Zhang (2013); Fu et al. (2014); Zeng and Xie (2012); Huang et al. (2016)) tend to group and select highly correlated relevant and irrelevant predictors together. Jia and Rohe (2015) propose transforming the design matrix so that the irrepresentable conditions are satisfied. However, the error terms are no longer independent from each other after the transformation. More importantly, a model-based transformation loses its original interpretation, in practice.

Under high-dimensional settings (Fan, Shao and Zhou (2018)), sure independence screening (Fan and Lv (2008)) screens out variables using the marginal correlations between the response and the covariates. However, the marginal correlations between the irrelevant covariates and the response could increase when the irrelevant covariates are strongly correlated with the relevant covariates, which may reduce the effectiveness of the sure independence screening. The Peter–Clark-simple (PC-simple) algorithm (Bühlmann, Kalisch and Maathuis (2010)) was developed to screen variables using partial correlation to solve the correlation problem. Moreover, Cho and Fryzlewicz (2012) generalize the partial correlation to a tilted correlation, and Li et al. (2016) and Jin, Zhang and Zhang (2014) incorporate inter-feature correlations to improve the detection of marginally weakly associated covariates. In addition, Bradic (2016) proposes a subsample bootstrap aggregation approach to circumvent the irrepresentable conditions, and Fan, Ke and Wang (2020) developed the factor-adjusted regularized model selection (Farm-Select) method to decorrelate highly-correlated covariates.

The partial correlation approach measures each individual covariate effect after removing other covariate effects (Peng et al. (2009); Bühlmann, Kalisch and Maathuis (2010); Li, Liu and Lou (2017); Tang, Wang and Barut (2017)). However, the range of the partial correlation is bounded between minus one and one, and therefore the partial correlation may not fully capture strong signals of some relevant covariates. This motivates us to develop a new semi-standard partial covariance (SPAC) approach to fully use the magnitude of the signal strength. The proposed SPAC is more powerful than the partial correlation in identifying relevant covariates.

Compared with traditional regularization methods, the proposed method encourages selecting covariates that have direct effects on the response variable, while discouraging the selection of irrelevant covariates that are strongly correlated with relevant covariates. We demonstrate the estimation consistency and variable selection consistency for the proposed SPAC method with the Lasso penalty (SPAC-Lasso) and the smoothly clipped absolute deviation (SCAD) penalty (SPAC-SCAD). The proposed method can handle both fixed-dimensional settings and high-dimensional settings when relevant and irrelevant covariates are highly correlated with each other.

Our work has the following contributions. First, the proposed variable selection approach can mitigate the bias of model selection caused by the violation of irrepresentable conditions for the Lasso or the SCAD method. We show that the proposed SPAC-Lasso and SPAC-SCAD are still sign consistent, and are especially effective when the correlations between the relevant and irrelevant covariates are higher than those between the relevant covariates. Second, the proposed SPAC is more effective in acquiring the signal strength, and thus is more powerful in selecting relevant predictors than is the traditional partial correlation. Numerical studies confirm that the proposed method outperforms traditional penalty-based variable selection methods, namely, the PC-simple algorithm and the Farm-Select method, for highly dependent covariates.

The remainder of the paper is organized as follows. Section 2 provides the model framework for the variable selection problem. Section 3 introduces the SPAC and presents the proposed methodology. Section 4 establishes theoretical properties of the SPAC-Lasso and SPAC-SCAD. Section 5 discusses the implementation of the proposed method. Section 6 presents various simulation studies. Section 7 illustrates a real-data application to a study on post-traumatic stress disorder (PTSD) in African Americans. Section 8 concludes the paper.

2. Model Framework and Notation

We formulate the variable selection problem under a linear regression setting,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ consists of samples for the response variable Y, $\boldsymbol{X} = (x_{ij})$ is an $n \times p$ random design matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a coefficient vector, and the noise vector $\boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I}_n)$ is uncorrelated with $\boldsymbol{X}$. Let $\boldsymbol{x}_j$ be the jth column (jth covariate) of $\boldsymbol{X}$, for each $j = 1, \ldots, p$. Without loss of generality, we assume that each column is standardized from independently and identically distributed (i.i.d.) samples; that is, $\boldsymbol{x}_j^T \boldsymbol{x}_j = n$ and mean $\sum_{i=1}^n x_{ij} = 0$, for $j = 1, \ldots. n$. Then, each row of $\boldsymbol{X}$ is identically distributed from a p-dimensional random vector $\boldsymbol{\mathcal{X}} = (X_1, \ldots, X_p)^T$ with mean $\boldsymbol{0}$ and positive-definite covariance matrix $\boldsymbol{C}_{p\times p}$, with diagonal elements all ones. In addition, we assume that the response variable is standardized with $\sum_{i=1}^n y_i = 0$, and thus the intercept can be omitted.

Here, we assume that the linear model in (2.1) is sparse, where most covariates have zero coefficients and are irrelevant to the response Y. That is, only the first q covariates in $\boldsymbol{X}$ have nonzero coefficients and are relevant to the response variable, and let $\beta_i = 0$ if and only if $i > q$. In addition, we let $\boldsymbol{\Sigma} = \mathrm{Cov}(Y, X_1, \ldots, X_p)$ and $\boldsymbol{\Sigma}^{-1} = (\sigma^{ij})$, where $i, j \in \{Y, 1, 2, \ldots, p\}$.

Under the sparsity assumption, the penalized least squares regression methods (Tibshirani (1996); Fu (1998)) select variables by minimizing the penalized least squares function

$$L(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^p p_\lambda(\beta_j), \tag{2.2}$$

where $\|\cdot\|$ represents the Euclidean norm, $p_\lambda(\cdot)$ is a penalty function, and λ is a tuning parameter. Here, $p_\lambda(\beta_j)$ could be the Lasso, adaptive Lasso, or SCAD penalty, which have the forms $p_{Lasso,\lambda}(\beta_j) = \lambda|\beta_j|$, $p_{ALasso,\lambda}(\beta_j) = \lambda|\beta_j|/|\hat{\beta}_{0j}|$, and

$$p_{SCAD,\lambda}(\beta_j) = \begin{cases} \lambda|\beta_j| & \text{if } 0 \le |\beta_j| \le \lambda \\ \dfrac{a\lambda|\beta_j| - 0.5(|\beta_j|^2 + \lambda^2)}{a-1} & \text{if } \lambda < |\beta_j| \le a\lambda \\ \dfrac{\lambda^2(a^2-1)}{2(a-1)} & \text{if } |\beta_j| > a\lambda, \end{cases} \tag{2.3}$$

respectively, where “ALasso” represents the adaptive Lasso penalty, $a > 2$, and $\hat{\beta}_{0j}$ is an initial estimator of β_j.

3. A New Variable Selection Method

In this section, we propose a semi-standard partial covariance (SPAC) variable selection approach to achieve selection consistency when the original irrepresentable conditions (Zhao and Yu (2006); Fan and Lv (2011)) fail; that is, there exist strong correlations between the relevant and the irrelevant covariates. The proposed SPAC is able to capture the relationship between a relevant covariate and the response variable, conditional on other covariate effects, because we derive this SPAC from the notion of partial correlation. For each $j = 1, \ldots, p$, let $\rho_j = \mathrm{Corr}(\varepsilon_Y, \varepsilon_j)$ be the partial correlation between the response Y and the covariate X_j, where ε_Y and ε_j are the residuals of linear regression models with Y and X_j as responses, respectively, and with $X_{-j} = \{X_k : k = 1, \ldots, j-1, j+1, \ldots, p\}$ as predictors.

Under the normality assumption

$$(Y, X_1, \ldots, X_p)^T \sim N_{p+1}(\mathbf{0}, \mathbf{\Sigma}), \tag{3.1}$$

it is well known that $\rho_j = \mathrm{Corr}(Y, X_j \mid X_{-j})$ (Baba, Shibata and Sibuya (2004)), indicating that a partial correlation measures the linear relationship between Y and X_j, conditional on other covariates. Moreover, nonzero partial correlations correspond to relevant covariates, whereas zero partial correlations correspond to irrelevant covariates.

However, a partial correlation is unable to fully capture the signal strength, which is the magnitude of β_j, owing to its bounded range. To overcome this limitation, we propose the following SPAC, and provide the association between the SPAC and a partial correlation in Lemma 1.

Definition 1. The semi-standard partial covariance (SPAC) between a response Y and a covariate X_j is $\gamma_j = \beta_j / d_{jj}^{1/2}$, for $j = 1, \ldots, p$, where d_{jj} is the jth diagonal element of the precision matrix $\boldsymbol{D} = \boldsymbol{C}^{-1}$.

The exponent $1/2$ of d_{jj} in Definition 1 ensures that γ_j does not depend on the scale of $X_j \mid X_{-j}$, as stated in the following lemma.

Lemma 1. *Let* $s_j = \{\mathrm{Var}(Y \mid X_{-j})\}^{1/2}$, *for each* $j = 1, \ldots, p$. *Under the normality assumption* (3.1), *we have*

$$\gamma_j = \rho_j s_j = \frac{\mathrm{Cov}(Y, X_j \mid X_{-j})}{\{\mathrm{Var}(X_j \mid X_{-j})\}^{1/2}}, \quad s_j^2 = \frac{1/\sigma^{YY}}{1 - \rho_j^2} = \frac{\beta_j^2}{d_{jj}} + \sigma_\varepsilon^2.$$

By definition, $\gamma_j = 0$ if and only if $\beta_j = 0$, for each $j = 1, \ldots, p$, implying that we can select relevant covariates by identifying nonzero SPACs. Lemma 1 shows that the SPAC is equivalent to multiplying the partial correlation by s_j under the normality assumption. Moreover, the proposed γ_j standardizes the partial covariance $\mathrm{Cov}(Y, X_j \mid X_{-j})$ by $\{\mathrm{Var}(X_j \mid X_{-j})\}^{1/2}$, instead of both s_j and $\{\mathrm{Var}(X_j \mid X_{-j})\}^{1/2}$ as in the partial correlation, which is why we refer to γ_j as the "semi-standard" partial covariance. We involve s_j in the SPAC, because it is an increasing function of the partial correlation ρ_j, and it incorporates the magnitude of the coefficient β_j, as indicated in Lemma 1. Therefore, the proposed SPAC is able to fully capture the signal strength of relevant predictors, while removing the effects of other covariates.

We illustrate the SPAC and compare it with a partial correlation from a geometric perspective using a toy example. Let $\boldsymbol{y} = \beta_1\boldsymbol{x}_1 + \beta_2\boldsymbol{x}_2 + \boldsymbol{\varepsilon}$, with $\beta_1 \neq 0$ and $\beta_2 = 0$; that is, $\boldsymbol{x}_1$ is relevant, but $\boldsymbol{x}_2$ is irrelevant. We also assume that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are correlated. By definition, $\gamma_1 \neq 0$ and $\gamma_2 = 0$.

We plot the relationships of $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{y}$ in Figure 1. As shown in the left graph, $\hat{\omega}_1$ is the angle between the two bold lines, which represent residuals of projections from $\boldsymbol{y}$ and $\boldsymbol{x}_1$ onto $\boldsymbol{x}_2$. Then, $\hat{\rho}_1 = \cos(\hat{\omega}_1)$ is the sample partial correlation based on samples in $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{y}$. The length of the bold line for residuals of $\boldsymbol{y}$ is a sample estimator of s_1, denoted by $\hat{s}_1$. By Lemma 1, $\hat{\gamma}_1 = \hat{s}_1 \cos(\hat{\omega}_1)$ is a sample estimator for γ_1, which is also the projection from residuals of $\boldsymbol{y}$ onto residuals of $\boldsymbol{x}_1$, represented by the dotted line in the left graph.

Similarly, in the right graph of Figure 1, $\hat{\rho}_2 = \cos(\hat{\omega}_2)$ and $\hat{\gamma}_2$ are the sample partial correlation and sample SPAC for $\boldsymbol{x}_2$, respectively. Here, $\hat{\gamma}_2$ is not exactly zero owing to sample variation. The differences between the sample SPACs and the sample partial correlations come from $\hat{s}_1$ and $\hat{s}_2$. As shown in Figure 1, $\hat{s}_2$ is just the sample variance of the error term, while $\hat{s}_1$ contains the error variation and increases with the signal coefficient β_1, implying that $\hat{s}_1$ should be larger than $\hat{s}_2$. Therefore, the SPAC is more effective in distinguishing relevant covariates from irrelevant covariates than is a partial correlation.

Compared with the coefficients $\boldsymbol{\beta}$, the SPAC takes account of correlation effects from other covariates. Specifically, because $1/d_{jj}^{1/2} = \{\mathrm{Var}(X_j \mid X_{-j})\}^{1/2} = (1 - R_j^2)^{1/2}$ (Lauritzen (1996); Raveh (1985)), the SPAC for covariate X_j is

$$\gamma_j = \beta_j \{\mathrm{Var}(X_j \mid X_{-j})\}^{1/2} = \beta_j \left(1 - R_j^2\right)^{1/2},$$

where R_j is the coefficient of the multiple correlation between X_j and all other covariates. When X_j is independent of the other covariates, γ_j is the same as β_j. On the other hand, when X_j is correlated with the other covariates, the SPAC mitigates the correlation effects from other covariates by multiplying β_j

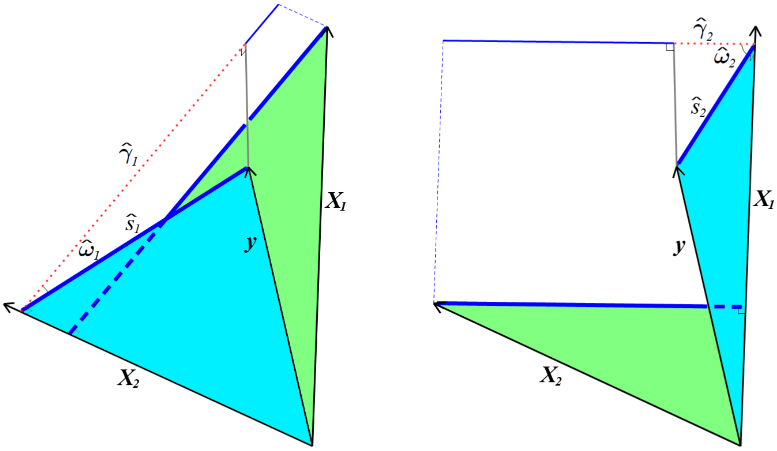

Figure 1. Illustrations of the SPAC and partial correlation when X_1 and X_2 are correlated.

by $(1-R_j^2)^{1/2}$. Thus, we propose estimating SPAC γ_j instead of the coefficient β_j to achieve model selection consistency for data with strong correlations between the irrelevant covariates and the relevant covariates.

Specifically, we replace the coefficient β_j in the penalized least squares function (2.2) with $\hat{d}_{jj}^{1/2}\gamma_j$, for each $j = 1, \ldots, p$, and estimate $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$ by minimizing

$$L(\boldsymbol{\gamma}, \hat{\boldsymbol{d}}) = \frac{1}{2}\|\boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{x}_j \hat{d}_{jj}^{1/2}\gamma_j\|^2 + \sum_{j=1}^{p} p_\lambda(\gamma_j)\hat{d}_{jj}, \tag{3.2}$$

where $\hat{\boldsymbol{d}} = (\hat{d}_{11}, \ldots, \hat{d}_{pp})^T$ is a consistent estimator of the diagonal elements $\boldsymbol{d} = (d_{11}, \ldots, d_{pp})^T$. Substituting $\boldsymbol{\beta}$ by $\boldsymbol{\gamma}$, we obtain a new matrix $\boldsymbol{X}^* = (\boldsymbol{x}_1\hat{d}_{11}^{1/2}, \ldots, \boldsymbol{x}_p\hat{d}_{pp}^{1/2})$, which serves as a design matrix for $\boldsymbol{\gamma}$. The squared Euclidean norm of the jth column in $\boldsymbol{X}^*$ is $\hat{d}_{jj}\boldsymbol{x}_j^T\boldsymbol{x}_j = \hat{d}_{jj}n$, for $j = 1, \ldots, p$, which leads to different weights on the penalizations for different covariates. However, the SPAC of each covariate could be equally important. To avoid unequal weighting, we reweight the penalization term by multiplying the penalty $p_\lambda(\gamma_j)$ in (3.2) by $\hat{d}_{jj}$, for each $j = 1, \ldots, p$. Consequently, the proposed SPAC estimator is

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}}{\text{argmin}}\ L(\boldsymbol{\gamma}, \hat{\boldsymbol{d}}),$$

and the corresponding estimator for the coefficients is $\hat{\boldsymbol{\beta}} = (\hat{d}_{11}^{1/2}\hat{\gamma}_1, \ldots, \hat{d}_{pp}^{1/2}\hat{\gamma}_p)^T$.

We adopt the Lasso, adaptive Lasso, and SCAD penalty functions to shrink the SPACs in Examples 1–3, respectively, and refer to the corresponding estimators as SPAC-Lasso, SPAC-ALasso, and SPAC-SCAD, respectively. We compare these estimators with the original Lasso, adaptive Lasso, and SCAD estimators in Sections 6 and 7.

Example 1. If we use Lasso penalty, the penalized loss function in (3.2) becomes

$$L_{Lasso}(\boldsymbol{\gamma}, \hat{\boldsymbol{d}}) = \frac{1}{2}\|\boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{x}_j \hat{d}_{jj}^{1/2} \gamma_j\|^2 + \lambda \sum_{j=1}^{p} \hat{d}_{jj} |\gamma_j|. \tag{3.3}$$

Accordingly, the proposed estimator with the Lasso penalty (SPAC-Lasso) is

$$\hat{\boldsymbol{\gamma}}_{Lasso} = \underset{\boldsymbol{\gamma}}{\text{argmin}}\ L_{Lasso}(\boldsymbol{\gamma}, \hat{\boldsymbol{d}}).$$

Example 2. Suppose that $\hat{\boldsymbol{\gamma}}_0 = (\hat{\gamma}_{01}, \ldots, \hat{\gamma}_{0p})^T$ is a consistent initial estimator for $\boldsymbol{\gamma}$. The objective function for the SPAC method with the adaptive Lasso penalty (SPAC-ALasso) is

$$L_{ALasso}(\boldsymbol{\gamma}, \hat{\boldsymbol{d}}) = \frac{1}{2}\|\boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{x}_j \hat{d}_{jj}^{1/2} \gamma_j\|^2 + \lambda \sum_{j=1}^{p} \hat{d}_{jj} \frac{|\gamma_j|}{|\hat{\gamma}_{0j}|^{\mu}}, \tag{3.4}$$

where $\mu > 0$ is a tuning parameter. The corresponding SPAC-ALasso estimator is

$$\hat{\boldsymbol{\gamma}}_{ALasso} = \underset{\boldsymbol{\gamma}}{\text{argmin}}\ L_{ALasso}(\boldsymbol{\gamma}, \hat{\boldsymbol{d}}).$$

Example 3. Similarly, the objective function for the proposed SPAC method with the SCAD penalty (SPAC-SCAD) is

$$L_{SCAD}(\boldsymbol{\gamma}, \hat{\boldsymbol{d}}) = \frac{1}{2}\|\boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{x}_j \hat{d}_{jj}^{1/2} \gamma_j\|^2 + n \sum_{j=1}^{p} p_{SCAD,\lambda}(\gamma_j) \hat{d}_{jj}, \tag{3.5}$$

where $p_{SCAD,\lambda}(\cdot)$ is defined in (2.3), and the corresponding SPAC-SCAD estimator is

$$\hat{\gamma}_{SCAD} = \underset{\boldsymbol{\gamma}}{\text{argmin}}\ L_{SCAD}(\boldsymbol{\gamma}, \hat{\boldsymbol{d}}).$$

4. Consistency Theory

In this section, we demonstrate the asymptotic properties of the proposed SPAC-Lasso and SPAC-SCAD estimators, and provide examples satisfying the conditions for the consistency of the proposed method. Although Lemma 1 is under the normality assumption, we do not require this assumption in the following subsection.

4.1. Consistency under high dimensionality

In this subsection, we establish the variable selection consistency and estimation consistency of the SPAC-Lasso and SPAC-SCAD under high dimensionality, where $p = p_n$, $q = q_n$, and $\boldsymbol{C} = \boldsymbol{C}_n$ increases with n. Similar results for fixed dimensions of p and q are provided in Section S2 in the Supplementary Material. For high-dimensional settings, the Lasso, adaptive Lasso, and SCAD methods require the correlations between relevant and irrelevant covariates to be relatively small compared with those between relevant covariates in order to achieve variable selection consistency (Zhao and Yu (2006); Huang, Ma and Zhang (2008); Kim, Choi and Oh (2008); Fan and Lv (2011)). The proposed SPAC approach mitigates the correlation effects from other covariates to achieve model selection consistency when relevant and irrelevant covariates are strongly correlated and the original irrepresentable conditions fail.

Following similar notation to that in Zhao and Yu (2006), let $\hat{\boldsymbol{\gamma}} =_s \boldsymbol{\gamma}$ if and only if $\text{sign}(\hat{\boldsymbol{\gamma}}) = \text{sign}(\boldsymbol{\gamma})$, and an estimator $\hat{\boldsymbol{\gamma}}$ is **strongly sign consistent** if there exists a tuning parameter λ_n, a function of n, such that

$$\lim_{n\to\infty} P\left\{\hat{\boldsymbol{\gamma}}(\lambda_n) =_s \boldsymbol{\gamma}\right\} = 1,$$

where λ_n is independent of the data.

To show the sign consistency of the proposed method, we define the following notation. Let $\boldsymbol{X}(1)$ and $\boldsymbol{X}(2)$ be the first q_n and the remaining $p_n - q_n$ columns in $\boldsymbol{X}$, respectively, such that $\boldsymbol{X}(1)$ contains relevant covariates, and $\boldsymbol{X}(2)$ consists of irrelevant covariates. Let $\hat{\boldsymbol{C}}_n = \boldsymbol{X}^T\boldsymbol{X}/n$ be the sample covariance matrix of $\boldsymbol{X}$, with diagonal elements all ones, because the covariates are standardized, as mentioned in Section 2. Thus, $\hat{\boldsymbol{C}}_n$ and the true covariance matrix $\boldsymbol{C}_n$ are both correlation matrices, and can be partitioned into blocks

$$\hat{\boldsymbol{C}}_n = \begin{pmatrix} \hat{\boldsymbol{C}}_n^{11} & \hat{\boldsymbol{C}}_n^{12} \\ \hat{\boldsymbol{C}}_n^{21} & \hat{\boldsymbol{C}}_n^{22} \end{pmatrix}, \quad \boldsymbol{C}_n = \begin{pmatrix} \boldsymbol{C}_n^{11} & \boldsymbol{C}_n^{12} \\ \boldsymbol{C}_n^{21} & \boldsymbol{C}_n^{22} \end{pmatrix},$$

according to $\boldsymbol{X} = (\boldsymbol{X}(1), \boldsymbol{X}(2))$. Similarly, we partition $\boldsymbol{\gamma}$ into $\boldsymbol{\gamma}(1) = (\gamma_1, \ldots, \gamma_q)^T$ and $\boldsymbol{\gamma}(2) = (\gamma_{q+1}, \ldots, \gamma_p)^T$, representing the relevant and irrelevant coefficients of the SPACs, respectively.

In addition, we define the following conditions for the proposed SPAC-Lasso and SPAC-SCAD.

Condition 1 (**Irrepresentable condition for SPAC-Lasso**). *There exists a* *positive constant η such that*

$$\left\| \boldsymbol{V}(2)\hat{\boldsymbol{C}}_n^{21}(\hat{\boldsymbol{C}}_n^{11})^{-1}\boldsymbol{V}(1)^{-1}\operatorname{sign}\{\boldsymbol{\beta}(1)\} \right\|_\infty \le 1-\eta,$$

where $\|\cdot\|_\infty$ represents the infinity norm of a matrix, and $\boldsymbol{V}(1)$ and $\boldsymbol{V}(2)$ are diagonal matrices diag$\{1/d_{11}^{1/2},\dots,1/d_{qq}^{1/2}\}$ and diag$\{1/d_{q+1q+1}^{1/2},\dots,1/d_{pp}^{1/2}\}$, respectively.

Condition 2 (**Irrepresentable condition for SPAC-SCAD**)**.** *There exists a positive constant η such that*

$$\mathcal{P}_{\lambda_n^*}(h_{\min}) \left\| \boldsymbol{V}(2)\hat{\boldsymbol{C}}_n^{21}(\hat{\boldsymbol{C}}_n^{11})^{-1}\boldsymbol{V}(1)^{-1} \right\|_\infty \le 1-\eta,$$

where $\mathcal{P}_{\lambda_n^*}(\cdot) = p'_{SCAD,\lambda_n^*}(\cdot)/\lambda_n^*$, $h_{\min} = \min_{1\le j\le q_n}|\beta_j|/2$, and $\lambda_n^* = \lambda_n \max_{1\le j\le q_n}$ $d_{jj}^{1/2}$.

Condition 1 is required for the sign consistency of the SPAC-Lasso, while Condition 2 is required for the SPAC-SCAD under high-dimensional settings. Condition 2 is weaker than Condition 1 when the signals are strong, because the SCAD penalty gradually levels off. The above two conditions are modified from the original irrepresentable conditions proposed in Zhao and Yu (2006) and Fan and Lv (2011) for the Lasso and SCAD, respectively. However, the proposed Conditions 1 and 2 could still hold for cases where the original irrepresentable conditions fail. We illustrate this with examples in Section 4.2.

Condition 3. *For some positive constants $0<\kappa_0,\kappa_2<1/2$, and $\kappa_1>0$, $\log p_n = O(n^{1-2\kappa_0})$, $q_n = O(n^{\kappa_2})$, $h_{\min} \ge (\log p_n/n)^{1/2}$, and $p_n \ge n^{\kappa_1}$.*

Condition 3 allows the number of covariates to grow exponentially, but requires a lower bound of signal strength, similarly to Fan and Lv (2013) and Zheng, Fan and Lv (2014)). The requirement $p_n \ge n^{\kappa_1}$ comes from Cai, Liu and Luo (2011) to ensure the consistency of the constrained L_1-minimization estimator (CLIME) (Cai, Liu and Luo (2011)), which is adopted in the following theorems. We let $\hat{\boldsymbol{d}}$ be the diagonal elements of the CLIME. Then, $\hat{\boldsymbol{d}}$ is consistent under some regularity conditions and a sparsity assumption of the precision matrix (Cai, Liu and Luo (2011)).

Following the notation in Cai, Liu and Luo (2011), we model the sparsity of the precision matrix $\boldsymbol{D}$ by defining

$$\mathcal{G}_u(K_{p_n}, M_{p_n}) = \left\{ \boldsymbol{D} : \max_{1\le j\le p_n} \sum_{i=1}^{p_n} |d_{ij}|^u \le K_{p_n}, \|\boldsymbol{D}\|_1 \le M_{p_n} \right\}, \tag{4.1}$$

where $0 \leq u < 1$, and K_{p_n} and M_{p_n} are positive and allowed to increase with n. We consider data with precision matrices $\boldsymbol{D} \in \mathcal{G}_u(K_{p_n}, M_{p_n})$ throughout this subsection. Details of other regularity Conditions 4, 5, 6, are provided in Section S1 in the Supplementary Material. The proofs for the following theorems are provided in Section S4 in the Supplementary Material.

Theorem 1. *Let $\hat{\boldsymbol{d}}$ be diagonal elements of the CLIME of $\boldsymbol{D}$. If Conditions 3, 4, and 5 are satisfied, and Condition 1 holds with probability at least $1 - O(n^{-\delta})$, then we have the following properties for the minimization of $L_{Lasso}(\boldsymbol{\gamma}, \hat{\boldsymbol{d}})$ in (3.3) with probability at least $1 - O(n^{-\delta})$.*

(1) *Strong sign consistency: There exists a strict local minimizer $\hat{\gamma}_{Lasso}$ such that $\hat{\gamma}_{Lasso} =_s \boldsymbol{\gamma}$.*

(2) *Estimation consistency: The corresponding estimator of the coefficients $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{V}}^{-1}\hat{\gamma}_{Lasso}$ satisfies*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty = O\left\{ \left(\frac{\log p_n}{n} \right)^{1/2} \right\}.$$

Theorem 2. *Let $\hat{\boldsymbol{d}}$ be diagonal elements of the CLIME of $\boldsymbol{D}$. If Conditions 3, 5, and 6 are satisfied, and Condition 2 holds with probability at least $1 - O(n^{-\delta})$, then we have the following properties for the minimization of $L_{SCAD}(\boldsymbol{\gamma}, \hat{\boldsymbol{d}})$ in (3.5) with probability at least $1 - O(n^{-\delta})$.*

(1) *Strong sign consistency: There is a strict local minimizer $\hat{\gamma}_{SCAD}$ such that $\hat{\gamma}_{SCAD} =_s \boldsymbol{\gamma}$.*

(2) *Estimation consistency: The corresponding estimator of the coefficients $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{V}}^{-1}\hat{\gamma}_{SCAD}$ satisfies*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty = O\left\{ \left(\frac{\log p_n}{n} \right)^{1/2} \right\}.$$

Theorems 1 and 2 state that, even though the number of covariates increases exponentially, the proposed SPAC-Lasso and SPAC-SCAD are able to select the true model with probability tending to one under Conditions 1 and 2, respectively. Moreover, the estimators for the coefficients based on the SPAC-Lasso and SPAC-SCAD both converge to the true $\boldsymbol{\beta}$.

Note that Condition 1 could still be valid even when the original irrepresentable conditions (Zhao and Yu (2006)) for the Lasso are violated. Similarly, Condition 2 is able to accommodate highly correlated covariates when the irrepresentable condition for the SCAD method (Fan and Lv (2011)) fails. We illustrate this point with examples in the following subsection.

4.2. Examples satisfying the proposed conditions

In this subsection, we give some examples where the proposed irrepresentable Conditions 1 and 2 still hold, even when the original irrepresentable conditions for the Lasso and SCAD fail, respectively. We suppose that $\boldsymbol{C}_n$ is a submatrix of $\boldsymbol{C}_{n+1}$ as the dimension increases.

We first consider using an extended block-exchangeable covariance matrix structure, which is defined as a block diagonal matrix consisting of identity matrices and $\boldsymbol{R}$:

$$\boldsymbol{C}_n = diag\{\boldsymbol{I}_{q_n-q_0}, \boldsymbol{R}, \boldsymbol{I}_{p_n-q_n-(p_0-q_0)}\}, \tag{4.2}$$

where

$$\boldsymbol{R}_{p_0\times p_0} = \begin{pmatrix} \boldsymbol{R}^{11} & \boldsymbol{R}^{12} \\ (\boldsymbol{R}^{12})^T & \boldsymbol{R}^{22} \end{pmatrix} \tag{4.3}$$

is block-exchangeable with

$$(\boldsymbol{R}^{11})_{i,j} = \begin{cases} 1, & i=j \\ \alpha_1, & i\neq j \end{cases}, \quad (\boldsymbol{R}^{22})_{i,j} = \begin{cases} 1, & i=j \\ \alpha_3, & i\neq j \end{cases}, \quad (\boldsymbol{R}^{12})_{i,j} = \alpha_2.$$

Here, α_1, α_2, and α_3 are unknown constants, $\boldsymbol{R}^{11}$ is a $q_0 \times q_0$ matrix, and p_0 and q_0 are constants independent of n.

The $\boldsymbol{R}$ is a fixed-dimensional and dense sub-matrix in $\boldsymbol{C}_n$. The number q_0 represents the number of relevant covariates with a nonsparse covariance matrix $\boldsymbol{R}^{11}$, and p_0 represents the total number of covariates with a nonsparse covariance matrix $\boldsymbol{R}$. There are $p_0 - q_0$ irrelevant covariates with a dense covariance matrix $\boldsymbol{R}^{22}$. In addition, $\boldsymbol{R}^{12}$ represents the covariance matrix between the correlated relevant and irrelevant covariates. We use $\boldsymbol{C}_n = diag\{\boldsymbol{I}_{q_n-q_0}, \boldsymbol{R}, \boldsymbol{I}_{p_n-q_n-(p_0-q_0)}\}$ instead of $\boldsymbol{R}$ as the covariance matrix to include a diverging and sparse covariance matrix for high-dimensional settings. Even under the sparse covariance matrix setting, the original irrepresentable conditions could still fail.

Similarly, we define $\boldsymbol{C}_n = diag\{\boldsymbol{I}_{q_n-q_0}, \boldsymbol{R}_{p_0\times p_0}, \boldsymbol{I}_{p_n-q_n-(p_0-q_0)}\}$ as an extended block-autoregressive (block-AR) covariance matrix, where

$$(\boldsymbol{R}^{11})_{i,j} = \alpha_1^{|i-j|}, \quad (\boldsymbol{R}^{22})_{i,j} = \alpha_3^{|i-j|}, \quad (\boldsymbol{R}^{12})_{i,j} = \alpha_2^{|i-(q_0+j)|}. \tag{4.4}$$

When the covariance matrix $\boldsymbol{C}_n$ is extended block-exchangeable, as in (4.2), the sparsity assumption in (4.1) holds with $K = p_0\{(p_0-1)!/\Delta_1\}^u$ and $M = p_0!/\Delta_1$, where $(p_0-1)!$ and $p_0!$ denote the factorials of p_0-1 and p_0, respectively, and $\Delta_1 = (1-\alpha_1)^{q_0}(1-\alpha_3)^{p_0-q_0}q_0(p_0-q_0)(\alpha_1\alpha_3-\alpha_2^2)$. When the covariance matrix is extended block-AR, the sparsity assumption in (4.1) holds with $K = p_0\{(p_0-1)!/\Delta_2\}^u$ and $M = p_0!/\Delta_2$, where $\Delta_2 = (1-\alpha_1^2)^{q_0}(1-\alpha_3^2)^{p_0-q_0}[1-\alpha_2^2(1-\alpha_1\alpha_2)^2(1-\alpha_3\alpha_2)^2/\{(1-\alpha_1^2)(1-\alpha_3^2)(1-\alpha_2^2)^2\}]$. Note that in both cases, the K and M do not depend on p_n. To simplify the following statements, we let $L_0 = q_0/m_0$, where $m_0 = |\sum_{i=q_n-q_0+1}^{q_n}\operatorname{sign}(\beta_i)| = |\sum_{i=q_n-q_0+1}^{q_n}\operatorname{sign}(\gamma_i)| > 0$.

Proposition 1. *Let $p_n = \exp(n^{1-2\kappa_0})$ and $q_n = n^{1/3}$, with $1/3+\tau < \kappa_0 < 1/2$ and $0<\tau<1/6$. Under the normality assumption* (3.1), *suppose that $\boldsymbol{C}_n$ is an extended block-exchangeable covariance matrix of the form in* (4.2), *with $\alpha_1, \alpha_2, \alpha_3 \in (-1,1)$, such that $\alpha_1\alpha_3 \neq \alpha_2^2$ and $\boldsymbol{C}_n$ is positive definite for any large constants q_0 and p_0-q_0, where $q_0 < p_0-q_0$. Then, there exists a constant $0<\delta<1/2$ such that*

$$\|\hat{\boldsymbol{C}}_n^{21}(\hat{\boldsymbol{C}}_n^{11})^{-1}\operatorname{sign}\{\boldsymbol{\beta}(1)\}\|_\infty \geq 1 \qquad \textit{with probability at least } 1-O(n^{-\delta}) \quad (4.5)$$

if $|\alpha_2| > \alpha_1 L_0$. Conversely, (4.5) *implies $|\alpha_2| > \alpha_1 L_0 \geq \alpha_1$, $\alpha_3 \geq |\alpha_2|$, and*

$$|\boldsymbol{V}(2)\hat{\boldsymbol{C}}_n^{21}(\hat{\boldsymbol{C}}_n^{11})^{-1}\boldsymbol{V}(1)^{-1}\operatorname{sign}\{\boldsymbol{\beta}(1)\}| < |\hat{\boldsymbol{C}}_n^{21}(\hat{\boldsymbol{C}}_n^{11})^{-1}\operatorname{sign}\{\boldsymbol{\beta}(1)\}|, \qquad (4.6)$$

for sufficiently large constants m_0 and p_0-q_0 with probability at least $1-O(n^{-\delta})$, where the inequality holds element-wise.

Proposition 2. *Under the conditions of Proposition* 1, *if for some constant $0<\delta<1/2$,*

$$\|\hat{\boldsymbol{C}}_n^{21}(\hat{\boldsymbol{C}}_n^{11})^{-1}\|_\infty \mathcal{P}_\lambda(h_{\min}) \geq 1 \qquad \textit{with probability at least } 1-O(n^{-\delta}), \quad (4.7)$$

then $\alpha_3 \geq |\alpha_2| > \alpha_1$, and

$$\|\boldsymbol{V}(2)\hat{\boldsymbol{C}}_n^{21}(\hat{\boldsymbol{C}}_n^{11})^{-1}\boldsymbol{V}(1)^{-1}\|_\infty \mathcal{P}_\lambda(h_{\min}) < \|\hat{\boldsymbol{C}}_n^{21}(\hat{\boldsymbol{C}}_n^{11})^{-1}\|_\infty \mathcal{P}_\lambda(h_{\min}), \qquad (4.8)$$

for sufficiently large constants m_0 and p_0-q_0 with probability at least $1-O(n^{-\delta})$.

The failure of the original irrepresentable conditions (Zhao and Yu (2006); Fan and Lv (2011)) of the Lasso and SCAD methods implies inequalities (4.5) and (4.7), respectively. By Propositions 1 and 2, if the original irrepresentable conditions fail, then the correlations between the relevant covariates are the smallest among the correlations of all covariates, followed by those between the relevant and irrelevant covariates. More importantly, the inequalities in (4.6) and (4.8) hold even when the original irrepresentable conditions are violated, indicating that the new irrepresentable Conditions 1 and 2 for the SPAC-Lasso and SPAC-SCAD, respectively, can still be valid.

The following corollaries provide sufficient conditions for the SPAC-Lasso to be strongly sign consistent when the true covariance matrix is extended block-exchangeable, as in (4.2), or extended block-AR, with $\boldsymbol{R}$ defined as in (4.4). We also provide a similar corollary in Section S4 in the Supplementary Material for the strong sign consistency of the SPAC-SCAD under the extended block-exchangeable covariance matrix structure.

Corollary 1. *Let $\hat{\boldsymbol{d}}$ be diagonal elements of the CLIME of $\boldsymbol{D}$. Suppose that the conditions of Proposition* 1 *and Condition* 4 *are satisfied, and that $h_{\min} \geq n^{-\kappa_0}$. If there exists a positive constant η such that*

$$|\alpha_2| \leq (1-\eta)\left(\frac{1-\alpha_1}{1-\alpha_3}\right)^{1/2} \alpha_1 L_0, \tag{4.9}$$

then the SPAC-Lasso possesses strong sign consistency, and the estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{V}}^{-1}\hat{\gamma}_{Lasso}$ is consistent for sufficiently large q_0 and $p_0 - q_0$.

Under the extended block-exchangeable structure with large n, the weak irrepresentable condition (Zhao and Yu (2006)) of the Lasso holds for large α_1. However, when $\alpha_1 < |\alpha_2|/L_0 \leq |\alpha_2|$, the Lasso is not sign consistent, by Proposition 1. In contrast, Corollary 1 shows that the SPAC-Lasso is strongly sign consistent, given that α_3 is sufficiently large, even when α_1 is small.

Corollary 2. *Suppose that the conditions of Corollary* 1 *are satisfied, except that $\boldsymbol{C}_n$ is an extended block-AR covariance matrix with $\boldsymbol{R}$ defined in* (4.4) *and $\alpha_1, \alpha_2, \alpha_3 \in (0,1)$, such that $2|\alpha_2 - z|\{\alpha_2 + 1/(1+z)\} < 1$, where $z = \alpha_1$ or α_3. Further, suppose that the true coefficients of the relevant covariates have the same sign. If* (4.5) *is satisfied, then $\alpha_2 > \alpha_1$, and the SPAC-Lasso is strongly sign consistent when there is a constant $\eta > 0$ such that*

$$\max\left\{\frac{\alpha_2}{|\alpha_2-\alpha_3|}, 1\right\}\left(\frac{1-\alpha_3^2}{1-\alpha_1^2}\right)^{1/2}\frac{\alpha_2(1-\alpha_1\alpha_2)}{(1+\alpha_1)(1-\alpha_2)} \leq 1-\eta. \tag{4.10}$$

Corollary 2 states that, under the extended block-AR structure, the failure of the weak irrepresentable condition of the Lasso also implies that the correlations between the relevant and irrelevant covariates are stronger than those between the relevant covariates, that is, $\alpha_2 > \alpha_1$. More importantly, even when the weak irrepresentable condition fails, the SPAC-Lasso is still strongly sign consistent, given that α_3 is sufficiently large. This is consistent with the results of the extended block-exchangeable example.

In the following proposition, we present another sufficient condition for Conditions 1 and 2 of the proposed method when the correlation structure does not have a specific form. We first introduce some notation. Let $\boldsymbol{C}_n = (c_{ij})_{p\times p}$ with $c_{ij} \geq 0$, and let $\boldsymbol{C}_{n,i}$ be a submatrix of $\boldsymbol{C}_n$ with the ith row and ith column removed for each $i = 1, \ldots, p$. Denote the ith column of $\boldsymbol{C}_n$ with the ith entry removed as $\boldsymbol{v}_i$; that is, $\boldsymbol{v}_i = (c_{1i}, \ldots, c_{i-1i}, c_{i+1i}, \ldots, c_{pi})^T$. In addition, let φ_i be the largest angle between $\boldsymbol{v}_i$ and any column vector in $\boldsymbol{C}_{n,i}$, and let $\lambda_{\min,i}$ and $\lambda_{\max,i}$ be the smallest and the largest eigenvalues of $\boldsymbol{C}_{n,i}$, respectively.

Proposition 3. *Suppose that the normality assumption* (3.1) *and Conditions* 3 *and* 4 *are satisfied with* $\kappa_0 > \max\{\kappa_2 + \kappa_3, (\kappa_2 + \kappa_4)/2\}$. *If*

$$0 \leq \frac{1 - \|\boldsymbol{v}_j\|_2^2/\lambda_{\max,j}}{1 - \|\boldsymbol{v}_i\|_2^2/\lambda_{\max,i} - \|\boldsymbol{v}_i\|_2^2 \sin^2 \varphi_i/\lambda_{\min,i}} < g_n^2$$

holds for all $i \in \{1, \ldots, q_n\}$ *and* $j \in \{q_n + 1, \ldots, p_n\}$, *with* $g_n = (1-\eta)/\|\boldsymbol{C}_n^{21} (\boldsymbol{C}_n^{11})^{-1}\|_\infty$ *for some* $\eta > 0$, *then* $\|\boldsymbol{V}(2)\hat{\boldsymbol{C}}_n^{21}(\hat{\boldsymbol{C}}_n^{11})^{-1}\boldsymbol{V}(1)^{-1}\|_\infty \leq 1 - \eta$ *with probability at least* $1 - O(n^{-\delta})$.

In general, when the correlations between relevant and irrelevant covariates are larger than those between the relevant covariates, the original irrepresentable conditions are likely to fail. In this case, correlations between irrelevant covariates could be high, owing to the positive-definiteness constraint on the correlation matrices. This indicates that, for each pair of relevant and irrelevant covariates, variables other than such a pair are more correlated with the irrelevant covariate than they are with the relevant one. Then, the irrepresentable conditions of the proposed SPAC method are likely to hold, by Proposition 3. Consequently, the irrepresentable conditions for the proposed SPAC method can still be satisfied when the original irrepresentable conditions are violated.

5. Implementation

In this section, we discuss the implementation of the proposed method with the Lasso, adaptive Lasso, or SCAD penalty. To estimate the diagonal elements $\boldsymbol{d}$, we apply the CLIME in our algorithms under high-dimensional settings, which estimates the jth column of the precision matrix $\boldsymbol{D}$ using the following minimization problem:

$$\min_{\boldsymbol{b}\in\mathbb{R}^p} |\boldsymbol{b}|_1 \quad \text{subject to } |\hat{\boldsymbol{C}}_n\boldsymbol{b} - \boldsymbol{e}_j|_\infty \le \lambda_d, \tag{5.1}$$

where $1 \le j \le p$, $\boldsymbol{e}_j \in \mathbb{R}^p$ is a vector with one in the jth coordinate and zero in the others, λ_d is a tuning parameter, and $|\cdot|_1$ and $|\cdot|_\infty$ represent the 1-norm and infinity norm, respectively, of a vector. We solve the problem (5.1) using the "fastclime" R package (`https://cran.r-project.org/web/packages/fastclime/index.html`), and then let $\hat{d}_{jj}$ be the jth element of the solution. In the fixed-dimensional settings, we use the sample precision matrix to estimate the diagonal elements

$$\hat{d}_{jj} = \{(n^{-1}\boldsymbol{X}^T\boldsymbol{X})^{-1}\}_{jj},\ j = 1, \dots, p. \tag{5.2}$$

For the SPAC-ALasso, we estimate the initial estimator $\hat{\boldsymbol{\gamma}}_0 = (\hat{\gamma}_{01}, \dots, \hat{\gamma}_{0p})^T$ in (3.4) using $\hat{\gamma}_{0j} = \hat{\beta}_{0j}/\hat{d}_{jj}^{1/2}$ $(1 \le j \le p)$, which implies that an initial estimator $\hat{\boldsymbol{\beta}}_0$ for $\boldsymbol{\beta}$ is required. We use the ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$ as the initial estimator $\hat{\boldsymbol{\beta}}_0$ under fixed-dimensional situations. For high-dimensional settings, we first select the variables using the SPAC-Lasso, and then compute the OLS estimators of the coefficients for the selected variables. We let $\hat{\boldsymbol{\beta}}_0$ be the vector consisting of the OLS estimators for the selected covariates, and zeros for the nonselected covariates. For the tuning parameter related to the adaptive Lasso penalty in (3.4), we let $\mu = 1$, and compare the proposed SPAC-ALasso with the traditional adaptive Lasso method with $\mu = 1$.

We use the coordinate descent algorithm (Fu (1998); Breheny and Huang (2011)) to solve the minimization problems with objective functions in (3.3), (3.4), and (3.5) for the SPAC-Lasso, SPAC-ALasso, and SPAC-SCAD, respectively. We illustrate this with $p = 1$ first. The unpenalized least squares solution of the univariate setting is $z = \boldsymbol{X}^T\boldsymbol{y}/(n\hat{d}^{1/2})$. Accordingly, the proposed SPAC-Lasso, SPAC-ALasso, and SPAC-SCAD estimators have closed forms

$$\hat{\gamma}_{Lasso}(z,\lambda) = \text{sign}(z)(|z| - \lambda)_+, \quad \hat{\gamma}_{ALasso}(z,\lambda,\hat{\gamma}_0) = \text{sign}(z)\left(|z| - \frac{\lambda}{|\hat{\gamma}_0|}\right)_+, \tag{5.3}$$

$$\hat{\gamma}_{SCAD}(z,\lambda,a) = \begin{cases} \text{sign}(z)(|z| - \lambda)_+ & \text{if } |z| \le 2\lambda \\ \dfrac{(a-1)z - \text{sign}(z)a\lambda}{a-2} & \text{if } 2\lambda < |z| \le a\lambda \\ z & \text{if } |z| > a\lambda, \end{cases} \tag{5.4}$$

respectively.

For a multivariate case, we use these univariate solutions to obtain coordinate-wise minimizers, except that we replace z in (5.3) and (5.4) with the unpenalized

Algorithm 1: SPAC-SCAD.

```
1. Set $l = 1$. Set tolerance $\epsilon$, initial values $\boldsymbol{\gamma}^{(0)}$, and tuning parameters $\lambda$ and $a$.
2. Calculate $\hat{\boldsymbol{d}}$ using (5.2) or by solving (5.1), for $j = 1, \dots, p$.
3. Calculate $\boldsymbol{r}^{(0)} = \boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{x}_j \hat{d}_{jj}^{1/2} \gamma_j^{(0)}$.
4. For $j = 1, \dots, p$, estimate $\gamma_j^{(l)}$ as follows:
       Calculate $z_j$ using (5.5);
       Calculate $\gamma_j^{(l)} = \hat{\gamma}_{SCAD}(z_j, \lambda, a)$ using (5.4);
       Update $\boldsymbol{r}^{(l)} = \boldsymbol{r}^{(l-1)} - \boldsymbol{x}_j \hat{d}_{jj}^{1/2} (\gamma_j^{(l)} - \gamma_j^{(l-1)})$
5. Iterate Step 4 until a convergence criterion is satisfied, for example,
   $\min_j \{ |(\gamma_j^{(l)} - \gamma_j^{(l-1)}) / \gamma_j^{(l-1)}| \} < \epsilon$.
```

solution of the regression with the partial residual of $\boldsymbol{x}_j$ $(1 \leq j \leq p)$ as the response

$$z_j = \frac{\boldsymbol{x}_j^T \boldsymbol{r}_{-j}}{n \hat{d}_{jj}^{1/2}} = \frac{\boldsymbol{x}_j^T \boldsymbol{r}}{n \hat{d}_{jj}^{1/2}} + \gamma_j^{(l-1)}, \tag{5.5}$$

where $\boldsymbol{r}_{-j} = \boldsymbol{y} - \sum_{i \neq j} \boldsymbol{x}_i \hat{d}_{ii}^{1/2} \gamma_i^*$ is $\boldsymbol{x}_j$'s partial residual, $\boldsymbol{r} = \boldsymbol{y} - \sum_{i=1}^{p} \boldsymbol{x}_i \hat{d}_{ii}^{1/2} \gamma_i^*$, and $\boldsymbol{\gamma}^* = (\gamma_1^*, \dots, \gamma_p^*)^T$ is the most recent updated estimator for $\boldsymbol{\gamma}$. A complete algorithm for the SPAC-SCAD is provided in Algorithm 1, including the estimation of $\boldsymbol{d}$ in Step 2 and the coordinate descent method in Step 4. Algorithms of the SPAC-Lasso and SPAC-ALasso are similar to Algorithm 1, except that we replace $\hat{\gamma}_{SCAD}$ in Step 4 with $\hat{\gamma}_{Lasso}$ or $\hat{\gamma}_{ALasso}$ in (5.3), respectively.

6. Simulations

In this section, we compare the performance of the proposed method with that of existing model selection approaches in simulation studies. We generate data 100 times based on a linear regression model, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + N_n(\boldsymbol{0}, \boldsymbol{I}_n)$, where $\boldsymbol{X}$ is an $n \times p$ matrix and $\boldsymbol{\beta}$ is a $p \times 1$ vector. Each row of the design matrix $\boldsymbol{X}$ is i.i.d. from a multivariate normal distribution with mean $\boldsymbol{0}_{p \times 1}$ and a block-exchangeable covariance matrix $\boldsymbol{C}_{p \times p}$ of the form in (4.3) with the parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T$. The first q elements in the coefficient vector $\boldsymbol{\beta}$ are nonzero and take the value β_s; the remaining elements are zero.

We implement the Lasso, adaptive Lasso, and SCAD methods using the coordinate descent algorithm (Fu (1998); Breheny and Huang (2011)). Because the purpose of the proposed method is to provide model selection consistency when the traditional methods fail, we first check whether the original weak irrepresentable condition is satisfied for the covariates selected by the Lasso. If the condition is violated, we adopt the proposed method; otherwise, the standard Lasso, adaptive Lasso, and SCAD methods can still be applied. We use the "pcalg" R package (`https://cran.r-project.org/web/packages/pcalg/index.html`) to implement the PC-simple algorithm with a significance level of 0.05, which is a method based on partial correlations. The Farm-Select method is implemented using the "FarmSelect" R package (`https://cran.r-project.org/web/packages/FarmSelect/index.html`). In each penalty-based method, the tuning parameter λ is selected using the extended BIC (EBIC), which is effective for small n, but large p (Chen and Chen (2008)). For the SCAD method and the proposed SPAC-SCAD method, we choose $a = 3.7$ (Fan and Li (2001)). For the adaptive Lasso, we apply the Lasso estimator as the initial estimator for the weighting.

To evaluate the performance of each method, we compute the false negative rate (FNR) and false positive rate (FPR), as follows:

$$\frac{\sum_{j=1}^{p} I(\hat{\beta}_j = 0, \beta_j \neq 0)}{\sum_{j=1}^{p} I(\beta_j \neq 0)}, \quad \frac{\sum_{j=1}^{p} I(\hat{\beta}_j \neq 0, \beta_j = 0)}{\sum_{j=1}^{p} I(\beta_j = 0)},$$

respectively, where $I(\cdot)$ is an indicator function. The FNR represents the proportion of relevant covariates that are not selected, while the FPR represents the proportion of selected irrelevant covariates. We define the overall false rate of a method as the summation of the FNR and the FPR. A method with smaller overall false rate exhibits better performance in terms of model selection. We calculate the mean FNR and FPR for each method using 100 replications.

Setting 1: Let $p = 250$, $q = 5$, $n = 80$, $\beta_s = 0.4$, and $\boldsymbol{\alpha} = (0.1, 0.3, 0.8)^T$, $(0.2, 0.4, 0.8)^T$, $(0.3, 0.5, 0.8)^T$, or $(0.5, 0.7, 0.9)^T$.

Table 1 shows that the proposed method outperforms existing model selection approaches under Setting 1. Specifically, the ratio of overall false rate of each penalty-based method to that of the proposed method with the same penalty function is greater than one across all covariance matrices. Furthermore, the overall false rates of the Farm-Select method and the PC-simple algorithm are both larger than that of the proposed SPAC-SCAD. In particular, the ratio of overall false rates is the largest when $\boldsymbol{\alpha} = (0.5, 0.7, 0.9)^T$, where the covariates are most correlated. For example, the ratio between the traditional Lasso and the proposed SPAC-Lasso is 6.637 when $\boldsymbol{\alpha} = (0.5, 0.7, 0.9)^T$, which is much larger than the corresponding ratios under other $\boldsymbol{\alpha}$.

Moreover, the FNRs of the SPAC-Lasso, SPAC-ALasso, and SPAC-SCAD are smaller than those of the traditional Lasso, adaptive Lasso, and SCAD methods, respectively, given each $\boldsymbol{\alpha}$. This also holds for the FPR. In addition, we present the violation rates in the last row of Table 1, which is the percentage of the original weak irrepresentable condition being violated based on 100 simulated data. The violation rates are all close to one, because the original weak irrepresentable condition does not hold for the true covariance matrices in this setting.

Setting 2: Let $p = 1000$, $q = 20$, $n = 150$, $\boldsymbol{\alpha} = (0.3, 0.5, 0.8)^T$, and $\beta_s = 0.2, 0.3, 0.4, 0.5$, or 0.6.

We consider high-dimensional situations with 1,000 covariates in Setting 2. The results in Table 2 show that the proposed method still outperforms other competing methods in terms of overall false rate. In addition, the ratios between the overall false rates are larger for scenarios with larger β_s, indicating that the proposed method shows a greater improvement over existing methods when the signals are stronger. The FNR and FPR of the PC-simple algorithm for relatively larger β_s are not provided in Table 2 because the PC-simple algorithm is quite time consuming under settings with strong signals and thousands of correlated potential predictors. It takes more than a few hours to run the algorithm for only one replication. However, we can still observe that the proposed SPAC-SCAD outperforms the PC-simple algorithm based on the results under $\beta_s = 0.2$ and $\beta_s = 0.3$.

We incorporate binary covariates in Setting 3. We first simulate data from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{C}$ of the form in (4.3), and then transform two relevant and 60 irrelevant covariates X_j to $\text{sign}(X_j)$.

Setting 3: Let $p = 250$, $q = 5$, $n = 80$, $\boldsymbol{\alpha} = (0.5, 0.7, 0.9)^T$, and $\beta_s = 0.2, 0.3, 0.4, 0.5, 0.6$, or 0.7.

The proposed method also outperforms the other methods when we have binary potential predictors, according to the results in Table 2. For instance, when $\beta_s = 0.5$, the FNR of the SPAC-ALasso is 0.320, only 43.4% of the FNR of the adaptive Lasso method, indicating that the proposed SPAC-ALasso selects a greater number of relevant covariates. Similarly, the FPR of the SPAC-ALasso is smaller than that of the adaptive Lasso, implying that the proposed SPAC-ALasso selects fewer irrelevant covariates. In addition, the overall false rate of

Table 1. Results for Setting 1. The “Ratio” for each penalty-based approach is the ratio of FPR+FNR calculated using the traditional method to the FPR+FNR from the proposed method with the same penalty. The “Ratio” for Farm-Select (or PC-simple) is the ratio of FPR+FNR for Farm-Select (or PC-simple) to that of SPAC-SCAD. “Violate” represents the percentage of the original weak irrepresentable condition being violated based on Lasso selection results for 100 simulated data.

$\boldsymbol{\alpha}$		(0.1, 0.3, 0.8)	(0.2, 0.4, 0.8)	(0.3, 0.5, 0.8)	(0.5, 0.7, 0.9)
Lasso	FNR	0.804	0.718	0.744	0.744
	FPR	0.002	0.005	0.007	0.018
SPAC-Lasso	FNR	0.510	0.382	0.460	0.112
	FPR	0.002	0.003	0.006	0.003
	Ratio	1.576	1.876	1.614	**6.637**
ALasso	FNR	0.794	0.778	0.794	0.890
	FPR	0.001	0.001	0.003	0.006
SPAC-ALasso	FNR	0.500	0.430	0.528	0.384
	FPR	0.000	0.001	0.002	0.002
	Ratio	1.589	1.808	1.505	**2.321**
SCAD	FNR	0.148	0.196	0.380	0.859
	FPR	0.126	0.093	0.057	0.004
SPAC-SCAD	FNR	0.126	0.120	0.214	0.303
	FPR	0.052	0.043	0.038	0.003
	Ratio	1.542	1.775	1.734	**2.821**
Farm-Select	FNR	0.200	0.600	0.200	1.000
	FPR	0.065	0.029	0.065	0.004
	Ratio	1.489	**3.859**	1.052	**3.281**
PC-simple	FNR	0.496	0.530	0.696	0.892
	FPR	0.003	0.004	0.005	0.007
	Ratio	**2.809**	**3.273**	**2.782**	**2.937**
Violate		0.900	0.940	0.860	0.970

the SPAC-SCAD decreases much faster than that of the PC-simple algorithm as β_s increases, which is consistent with the fact that a partial correlation is unable to fully use the signal strength, owing to its bounded range.

Because the estimation of the diagonal elements $\boldsymbol{d}$ could be inaccurate, we investigate the robustness of the proposed method with respect to the estimation of $\boldsymbol{d}$ in Setting 4. In this setting, we replace $\hat{d}_{jj}$ in the implementation of SPAC-Lasso with $\hat{d}_{jj} + u_j$, for each $j = 1, \ldots, p$, where u_j are i.i.d. from a truncated normal distribution with minimum value $\max_{1 \le j \le p}\{-\hat{d}_{jj}\}$, mean zero, and variance σ_u^2. Here, we require the random noise $u_j \ge \max_{1 \le j \le p}\{-\hat{d}_{jj}\}$ to ensure that $\hat{d}_{jj} + u_j$ is positive for each $j = 1, \ldots, p$.

Setting 4: Let $p = 500$, $q = 6$, $n = 100$, $\beta_s = 0.3$, and $\boldsymbol{\alpha} = (0.1, 0.3, 0.8)^T$ or

Table 2. Results for Settings 2 and 3. The "Ratio" for each penalty-based approach is the ratio of FPR+FNR calculated using the traditional method to the FPR+FNR from the proposed method with the same penalty. The "Ratio" for Farm-Select (or PC-simple) is the ratio of FPR+FNR for Farm-Select (or PC-simple) to that of SPAC-SCAD. "Violate" represents the percentage of the original weak irrepresentable condition being violated based on Lasso selection results for 100 simulated data.

Setting		Setting 2					Setting 3					
β_s		0.2	0.3	0.4	0.5	0.6	0.2	0.3	0.4	0.5	0.6	0.7
Lasso	FNR	0.986	0.972	0.852	0.586	0.415	0.976	0.912	0.782	0.642	0.388	0.180
	FPR	0.007	0.008	0.010	0.016	0.019	0.003	0.009	0.016	0.018	0.025	0.028
SPAC-Lasso	FNR	0.921	0.639	0.357	0.094	0.020	0.904	0.692	0.334	0.182	0.044	0.006
	FPR	0.003	0.007	0.012	0.019	0.021	0.002	0.003	0.009	0.011	0.013	0.014
	Ratio	1.075	1.516	**2.334**	**5.328**	**10.625**	1.082	1.325	**2.329**	**3.425**	**7.214**	**10.557**
ALasso	FNR	0.998	0.990	0.936	0.659	0.466	0.982	0.954	0.854	0.738	0.496	0.300
	FPR	0.002	0.002	0.002	0.003	0.002	0.002	0.004	0.005	0.006	0.005	0.005
SPAC-ALasso	FNR	0.969	0.806	0.565	0.340	0.249	0.914	0.744	0.478	0.320	0.108	0.042
	FPR	0.001	0.001	0.001	0.001	0.000	0.001	0.002	0.002	0.002	0.002	0.001
	Ratio	1.031	1.229	1.657	1.939	1.881	1.075	1.284	1.788	**2.309**	**4.569**	**7.080**
SCAD	FNR	0.827	0.702	0.580	0.323	0.073	0.824	0.830	0.780	0.612	0.476	0.242
	FPR	0.005	0.008	0.008	0.007	0.004	0.024	0.008	0.008	0.009	0.006	0.005
SPAC-SCAD	FNR	0.513	0.296	0.108	0.013	0.001	0.774	0.499	0.298	0.158	0.078	0.032
	FPR	0.007	0.007	0.004	0.002	0.001	0.005	0.006	0.007	0.004	0.004	0.002
	Ratio	1.599	**2.339**	**5.235**	**22.118**	**44.648**	1.090	1.659	**2.585**	**3.825**	**5.908**	**7.275**
Farm-Select	FNR	1.000	1.000	1.000	0.850	0.800	1.000	1.000	1.000	1.000	0.400	0.200
	FPR	0.000	0.001	0.005	0.006	0.006	0.016	0.016	0.024	0.000	0.029	0.029
	Ratio	1.923	**3.296**	**8.944**	**>50**	**>50**	1.306	**2.011**	**3.359**	**6.154**	**5.250**	**6.712**
PC-simple	FNR	0.992	0.997	—	—	—	0.918	0.876	0.844	0.786	0.726	0.682
	FPR	0.005	0.007	—	—	—	0.005	0.006	0.007	0.007	0.007	0.007
	Ratio	1.917	**3.305**	—	—	—	1.186	1.746	**2.789**	**4.880**	**8.981**	**20.242**
Violate		0.980	0.970	1.000	1.000	1.000	0.909	0.899	0.980	0.960	0.970	1.000

$(0.2, 0.4, 0.8)^T$. The variance parameter $\sigma_u = 0, 1, 3$, or 5.

The results for Setting 4 in Table 3 show that the overall false rate of the SPAC-Lasso with noise increases as the variance σ_u^2 increases, but that this overall false rate is still smaller than that of the Lasso method. For example, when $\boldsymbol{\alpha} = (0.1, 0.3, 0.8)^T$, the overall false rate of the SPAC-Lasso with $\sigma_u = 5$ is 0.919 larger than that with $\sigma_u = 0$, but smaller than 0.989, which is the overall false rate of the Lasso method. Thus, the proposed method is robust to certain errors in the estimation of $\boldsymbol{d}$. Although we use the CLIME for the estimation of d in this study, other consistent estimators can also be used.

Setting 5: Let $p = 150$, $q = 3$, $n = 80$, and $\beta_s = 0.5$. The parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T$ are $(0.2, 0.4, 0.8)$, $(0.8, 0.4, 0.2)$, $(0.1, 0.3, 0.8)$, $(0.8, 0.3, 0.1)$, $(0.2, 0.4, 0.7)$, $(0.7, 0.4, 0.2)$, $(0.4, 0.5, 0.7)$, or $(0.7, 0.5, 0.4)$.

Table 3. Results for Setting 4.

$\boldsymbol{\alpha}$	(0.1, 0.3, 0.8)					(0.2, 0.4, 0.8)				
Method	Lasso	SPAC-Lasso				Lasso	SPAC-Lasso			
σ_u	—	0	1	3	5	—	0	1	3	5
FNR	0.988	0.843	0.903	0.895	0.918	0.958	0.848	0.867	0.915	0.908
FPR	0.001	0.000	0.000	0.001	0.001	0.001	0.001	0.001	0.002	0.002
FNR+FPR	0.989	0.843	0.903	0.896	0.919	0.960	0.849	0.868	0.917	0.910

In Setting 5, we examine the robustness of the proposed method when the original weak irrepresentable condition holds. As shown in Tables 4–5, the proposed method still outperforms the existing methods in terms of FNR+FNP when $\alpha_3 > \alpha_1$. In addition, the proposed method performs comparably to the existing methods when $\alpha_1 > \alpha_3$, where the original weak irrepresentable condition holds. For example, the ratios of the overall false rates of the adaptive Lasso method to those of the proposed SPAC-ALasso are greater than 1.5 when $\alpha_3 > \alpha_1$, and are equal to or quite close to one when $\alpha_1 > \alpha_3$. In summary, the proposed method performs similarly to the regular penalization method when the weak irrepresentable condition holds, but performs much better than the existing method when the condition fails.

7. Real-Data Application

In this section, we apply the proposed method to high-dimensional genetic data collected in the Detroit neighborhood health study (`https://dnhs.unc.edu/`), a representative study focusing on post-traumatic stress disorder (PTSD) of African American adults in Detroit, Michigan. This study collects gene expression data and post-traumatic checklists based on incident trauma exposures, which is a 17-item set of self-reported measures of PTSD symptoms. We treat the average of the 17 post-traumatic checklist scores as the response Y. Studies (Logue et al. (2015); Kuan et al. (2017b)) show that gene expression is associated with PTSD. To identify gene probes that are relevant to PTSD, we consider using all gene probes as potential predictors.

Because there are more than 15,000 gene probes and the sample size is only 93, we first screen the gene probes based on the correlations between probes and the marginal correlations between probes and Y. For each probe X_j, we let $\boldsymbol{c}_j$ denote the vector consisting of correlations between this probe and other probes. Because the proposed method targets correlated data, we consider X_j to be correlated with others and select it if the average absolute value of the elements in

Table 4. Results for Setting 5. The "Ratio" for each penalty-based approach is the ratio of FPR+FNR calculated using the traditional method to the FPR+FNR from the proposed method with the same penalty. "Violate" represents the percentage of the original weak irrepresentable condition being violated based on Lasso selection results for 100 simulated data.

$\boldsymbol{\alpha}$		(0.2, 0.4, 0.8)	(0.8, 0.4, 0.2)	(0.1, 0.3, 0.8)	(0.8, 0.3, 0.1)
Lasso	FNR	0.240	0.110	0.357	0.103
	FPR	0.009	0.002	0.006	0.002
SPAC-Lasso	FNR	0.143	0.123	0.123	0.110
	FPR	0.003	0.002	0.003	0.002
	Ratio	1.699	0.895	2.871	0.942
ALasso	FNR	0.297	0.393	0.257	0.373
	FPR	0.002	0.001	0.001	0.001
SPAC-ALasso	FNR	0.160	0.413	0.147	0.373
	FPR	0.001	0.001	0.001	0.001
	Ratio	1.852	0.952	1.748	1.000
SCAD	FNR	0.073	0.647	0.057	0.660
	FPR	0.012	0.002	0.009	0.002
SPAC-SCAD	FNR	0.050	0.657	0.030	0.663
	FPR	0.007	0.002	0.006	0.002
	Ratio	1.493	0.985	1.830	0.995
Violate		0.797	0.037	0.743	0.007

$\boldsymbol{c}_j$ is greater than 0.1. Moreover, we calculate the marginal correlations between selected probes and the response variable, and filter out probes with absolute values of the marginal correlations less than 0.15, which are unlikely to be important probes. After the screening, we retain $3,591$ gene probes for further analysis.

To evaluate the performance of the different methods, we randomly partition all observations into 95% for training and 5% for testing, 100 times. For each method, we estimate the parameters using the training sets, calculate the mean number of selected probes, and compute the average prediction mean squared errors (PMSEs) from testing sets based on 100 replications. However, the PMSEs of the PC-simple algorithm and the Farm-Select method are unavailable; the former only provides variable selection results, without a coefficient estimation, and the R package of the Farm-Select method does not have an intercept in the model. To calculate prediction errors for the two methods and compare them to those of other methods, we adopt the OLS to estimate the coefficients of the probes selected by each method, and calculate the PMSE based on the OLS estimation, denoted by OLS-PMSE. The original weak irrepresentable condition fails in each training set based on the selection results of Lasso, indicating that

Table 5. Results for Setting 5. The "Ratio" for each penalty-based approach is the ratio of FPR+FNR calculated using the traditional method to the FPR+FNR from the proposed method with the same penalty. "Violate" represents the percentage of the original weak irrepresentable condition being violated based on Lasso selection results for 100 simulated data.

$\boldsymbol{\alpha}$		(0.2, 0.4, 0.7)	(0.7, 0.4, 0.2)	(0.4, 0.5, 0.7)	(0.7, 0.5, 0.4)
Lasso	FNR	0.520	0.103	0.353	0.170
	FPR	0.010	0.004	0.013	0.004
SPAC-Lasso	FNR	0.303	0.110	0.163	0.193
	FPR	0.007	0.004	0.006	0.004
	Ratio	1.708	0.943	2.158	0.884
ALasso	FNR	0.593	0.310	0.377	0.327
	FPR	0.002	0.002	0.003	0.002
SPAC-ALasso	FNR	0.380	0.310	0.200	0.350
	FPR	0.001	0.002	0.002	0.003
	Ratio	1.562	1.000	1.882	0.931
SCAD	FNR	0.280	0.463	0.187	0.493
	FPR	0.012	0.008	0.015	0.009
SPAC-SCAD	FNR	0.140	0.473	0.133	0.497
	FPR	0.006	0.008	0.010	0.010
	Ratio	1.991	0.980	1.409	0.991
Violate		0.883	0.007	0.890	0.133

the proposed method is more suitable for the data than traditional methods are.

Table 6 provides the average PMSE and OLS-PMSE and the number of selected probes for all the methods. According to the table, the proposed method produces a smaller PMSE and a smaller OLS-PMSE than those of existing methods. In particular, the average OLS-PMSE of the Lasso is 18.7% more than that of the SPAC-Lasso. Similarly, the average PMSE of the traditional adaptive Lasso and SCAD methods are 16.2% and 17.3% more than those of the proposed SPAC-ALasso and SPAC-SCAD, respectively. Moreover, in terms of the OLS-PMSE, the Farm-Select method and PC-simple algorithm perform worse than the proposed method. Among all the methods, the SPAC-ALasso produces the smallest PMSE with relatively fewer selected probes. In addition, the prediction errors of the methods with the SCAD penalty are larger than those of methods with other penalties.

In addition, we apply these methods to all of the samples, and summarize the selected probes in tables in Section S3 of the Supplementary Material. On the one hand, *ILMN_1716728*, *ILMN_1682259*, *ILMN_3307729*, *ILMN_1670134*, *ILMN_1793201*, *ILMN_1811507*, *ILMN_1656111*, and *ILMN_3248844* are common probes selected by the

Table 6. Average results for the real data.

	PMSE	OLS-PMSE	NS
Lasso	0.9306	0.9868	73
SPAC-Lasso	**0.8283**	**0.8310**	74
ALasso	0.9568	1.0406	20
SPAC-ALasso	**0.8232**	**0.9101**	22
SCAD	1.3353	1.3164	38
SPAC-SCAD	**1.1387**	**1.1298**	39
Farm-Select	—	1.2429	40
PC-simple	—	1.3278	5

Lasso, SPAC-Lasso, ALasso, SPAC-ALasso, SCAD, SPAC-SCAD, and Farm-Select. Thus, these probes are very likely to be associated with the response. Of these, *ILMN_1716728*, *ILMN_3307729*, and *ILMN_3248844* are also selected by the PC-simple algorithm, indicating that these three probes are extremely likely to be relevant to PTSD. On the other hand, *ILMN_1663035* from the *SREBF1* gene is only selected by the proposed SPAC-Lasso and SPAC-ALasso. According to the existing literature (Kuan et al. (2017a)), the *SREBF1* gene is indeed associated with PTSD.

In conclusion, the proposed method leads to a smaller PMSE and OLS-PMSE than existing variable selection methods with similar numbers of selected probes, showing that the proposed SPAC strategy improves the accuracy of variable selection.

8. Conclusion

We have proposed a new variable selection approach to address the problem in which the original irrepresentable conditions fail due to a strong dependency between relevant and irrelevant covariates. The violation of the irrepresentable conditions leads to inconsistency of model selection based on traditional methods. In this paper, we introduce a semi-standard partial covariance (SPAC), which has a clear geometric interpretation based on projections, and takes advantage of both coefficients $\boldsymbol{\beta}$ and partial correlations. Moreover, we develop a SPAC method that penalizes SPACs instead of coefficients $\boldsymbol{\beta}$ or partial correlations alone to mitigate the selection of irrelevant covariates that are strongly correlated with relevant covariates.

We establish the strong sign consistency of the proposed SPAC-Lasso and SPAC-SCAD under high dimensionality. Specifically, we transform irrepresentable conditions to achieve variable selection consistency, thus solving the problem of when the Lasso or SCAD method is not sign consistent. Because we focus on situations in which traditional methods fail, we first check whether the original weak irrepresentable condition holds. If it is violated, numerical studies show that the proposed approach is more effective and outperforms the traditional variable selection methods.

In contrast to partial correlation approaches, such as the PC-simple algorithm, the proposed method takes full advantage of the signal strength, because SPACs incorporate the magnitudes of the coefficients. This is also reflected in the numerical studies, where the SPAC-ALasso and the PC-simple algorithm both produce relatively small FNRs but large FPRs, because they tend to select fewer covariates compared with other methods. However, as the signal strength increases, the false positive rate of the SPAC-ALasso decreases significantly compared with that of the PC-simple algorithm. Additionally, the proposed method can still achieve sign consistency for nonGaussian distributed covariates, such as categorical covariates, where a partial correlation is unable to capture the conditional independence. In simulation settings with binary covariates, the proposed method performs much better than the PC-simple algorithm in terms of overall false rate.

Although we do not provide the theoretical properties on the consistency of the SPAC-ALasso, the proof should be similar to that of the SPAC-Lasso. Moreover, the SPAC idea is flexible and can be readily applied to other penalty-based methods and the generalized linear model framework.

Supplementary Material

We provide additional conditions, theorems, tables, and corollaries, as well as proofs for Lemma 1, all theorems, propositions, and corollaries, in the online Supplementary Material.

Acknowledgments

We would like to acknowledge the support for this project from the National Science Foundation (NSF grant DMS 1821198).

References

Baba, K., Shibata, R. and Sibuya, M. (2004). Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics* **46**, 657–664.

Bradic, J. (2016). Randomized maximum-contrast selection: Subagging for large-scale regression. *Electronic Journal of Statistics* **10**, 121–170.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics* **5**, 232–252.

Bühlmann, P., Kalisch, M. and Maathuis, M. H. (2010). Variable selection in high-dimensional linear models: Partially faithful distributions and the PC-simple algorithm. *Biometrika* **97**, 261–278.

Bühlmann, P., Rütimann, P., van de Geer, S. and Zhang, C.-H. (2013). Correlated variables in regression: Clustering and sparse estimation. *Journal of Statistical Planning and Inference* **143**, 1835–1858.

Cai, T., Liu, W. and Luo, X. (2011). A constrained l_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106**, 594–607.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics* **35**, 2313–2351.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759–771.

Chén, O. Y., Crainiceanu, C., Ogburn, E. L., Caffo, B. S., Wager, T. D. and Lindquist, M. A. (2018). High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics* **19**, 121–136.

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

Fan, J., Ke, Y. and Wang, K. (2020). Factor-adjusted regularized model selection. *Journal of Econometrics* **216**, 71–85.

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

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

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory* **57**, 5467–5484.

Fan, J., Shao, Q.-M. and Zhou, W.-X. (2018). Are discoveries spurious? Distributions of maximum spurious correlations and their applications. *The Annals of Statistics* **46**, 989–1017.

Fan, Y. and Lv, J. (2013). Asymptotic equivalence of regularization methods in thresholded parameter space. *Journal of the American Statistical Association* **108**, 1044–1061.

Fu, G.-H., Zhang, W.-M., Dai, L. and Fu, Y.-Z. (2014). Group variable selection with oracle property by weight-fused adaptive elastic net model for strongly correlated data. *Communications in Statistics-Simulation and Computation* **43**, 2468–2481.

Fu, W. J. (1998). Penalized regressions: The bridge versus the Lasso. *Journal of Computational and Graphical Statistics* **7**, 397–416.

Hilafu, H. and Yin, X. (2017). Sufficient dimension reduction and variable selection for large-p-small-n data with highly correlated predictors. *Journal of Computational and Graphical Statistics* **26**, 26–34.

Huang, J., Breheny, P., Lee, S., Ma, S. and Zhang, C.-H. (2016). The Mnet method for variable selection. *Statistica Sinica* **26**, 903–923.

Huang, J., Ma, S. and Zhang, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica* **18**, 1603–1618.

Imai, K. and Yamamoto, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis* **21**, 141–171.

Javanmard, A. and Montanari, A. (2013). Model selection for high-dimensional regression under the generalized irrepresentability condition. In *Proceedings of the 26th International Conference Neural Information Processing Systems* (Edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger), 3012–3020. Curran Associates Inc., Red Hook.

Jérolon, A., Baglietto, L., Birmelé, E., Alarcon, F. and Perduca, V. (2021). Causal mediation analysis in presence of multiple mediators uncausally related. *The International Journal of Biostatistics* **17**, 191–221.

Jia, J. and Rohe, K. (2015). Preconditioning the Lasso for sign consistency. *Electronic Journal of Statistics* **9**, 1150–1172.

Jin, J., Zhang, C.-H. and Zhang, Q. (2014). Optimality of graphlet screening in high-dimensional variable selection. *Journal of Machine Learning Research* **15**, 2723–2772.

Kim, Y., Choi, H. and Oh, H.-S. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association* **103**, 1665–1673.

Kuan, P., Waszczuk, M., Kotov, R., Marsit, C., Guffanti, G., Gonzalez, A. et al. (2017a). An epigenome-wide DNA methylation study of PTSD and depression in World Trade Center responders. *Translational Psychiatry* **7**, e1158.

Kuan, P.-F., Waszczuk, M. A., Kotov, R., Clouston, S., Yang, X., Singh, P. K. et al. (2017b). Gene expression associated with PTSD in World Trade Center responders: An RNA sequencing study. *Translational Psychiatry* **7**, 1297.

Lauritzen, S. L. (1996). *Graphical Models.* Clarendon Press, Oxford.

Li, R., Liu, J. and Lou, L. (2017). Variable selection via partial correlation. *Statistica Sinica* **27**, 983–996.

Li, Y., Hong, H., Kang, J., He, K., Zhu, J. and Li, Y. (2016). Classification with ultrahigh-dimensional features. *arXiv preprint arXiv:1611.01541.*

Logue, M. W., Smith, A. K., Baldwin, C., Wolf, E. J., Guffanti, G., Ratanatharathorn, A. et al. (2015). An analysis of gene expression in PTSD implicates genes involved in the glucocorticoid receptor pathway and neural responses to stress. *Psychoneuroendocrinology* **57**, 1–13.

Maier, A. and Rodríguez-Salas, D. (2017). Fast and robust selection of highly-correlated features in regression problems. In *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA),* 482–485. IEEE, New York.

Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association* **104**, 735–746.

Raveh, A. (1985). On the use of the inverse of the correlation matrix in multivariate data analysis. *The American Statistician* **39**, 39–42.

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

Tang, Y., Wang, H. J. and Barut, E. (2017). Testing for the presence of significant covariates through conditional marginal regression. *Biometrika* **105**, 57–71.

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

Wang, X. and Wang, M. (2014). Combination of nonconvex penalties and ridge regression for high-dimensional linear models. *Journal of Mathematical Research with Applications* **34**, 743–753.

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

Zeng, L. and Xie, J. (2012). Group variable selection for data with dependent structures. *Journal of Statistical Computation and Simulation* **82**, 95–106.

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

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

Zheng, Z., Fan, Y. and Lv, J. (2014). High dimensional thresholded regression and shrinkage effect. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 627–649.

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

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

Fei Xue
Department of Statistics, Purdue University, West Lafayette, IN 47906, USA.
E-mail: feixue@purdue.edu
Annie Qu
Department of Statistics, University of California Irvine, Irvine, CA 92697, USA.
E-mail: aqu2@uci.edu

(Received March 2020; accepted February 2021)