Weak signals in high-dimensional regression: detection, estimation and prediction

Yanming Li[1], Hyokyoung G. Hong[2,*], S. Ejaz Ahmed[3] and Yi Li[1]

[1] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109 USA
[2] Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824 USA
[3] Department of Mathematics & Statistics, Brock University, St. Catharines, ON L2S 3A1 Canada
[*] *email*: hhong@msu.edu

May 14, 2018

SUMMARY. Regularization methods, including Lasso, group Lasso and SCAD, typically focus on selecting variables with strong effects while ignoring weak signals. This may result in biased prediction, especially when weak signals outnumber strong signals. This paper aims to incorporate weak signals in variable selection, estimation and prediction. We propose a two-stage procedure, consisting of variable selection and post-selection estimation. The variable selection stage involves a covariance-insured screening for detecting weak signals, while the post-selection estimation stage involves a shrinkage estimator for jointly estimating strong and weak signals selected from the first stage. We term the proposed method as the covariance-insured screening based post-selection shrinkage estimator. We establish asymptotic properties for the proposed method and show, via simulations, that incorporating weak signals can improve estimation and prediction performance. We apply the proposed method to predict the annual gross domestic product (GDP) rates based on various socioeconomic indicators for 82 countries.

KEY WORDS: high-dimensional data, Lasso, post-selection shrinkage estimation, variable selection, weak signal detection.

1 Introduction

Given n independent samples, we consider a high-dimensional linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ is an n-vector of responses, $\mathbf{X} = (X_{ij})_{n\times p}$ is an $n \times p$ random design matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a p-vector of regression coefficients and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is an n-vector of independently and identically distributed random errors with mean 0 and variance σ^2. Let $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^{\mathrm{T}}$ denote the true value of $\boldsymbol{\beta}$. We write $\mathbf{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})^{\mathrm{T}} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, where $\mathbf{x}^{(i)} = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ is the i-th row of $\mathbf{X}$ and $\mathbf{x}_j$ is the j-th column of $\mathbf{X}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Without the subject index i, we write y, X_j and ε as the random variables underlying y_i, X_{ij} and ε_i, respectively. We assume that each X_j is independent of ε. We write $\mathbf{x}$ as the random vector underlying $\mathbf{x}^{(i)}$ and assume that $\mathbf{x}$ follows a p-dimensional multivariate sub-Gaussian distribution with mean zeros, variance proxy σ_x^2, and covariance matrix $\boldsymbol{\Sigma}$. Sub-Gaussian distributions contain a wide range of distributions such as Gaussian, binary and all bounded random variables. Therefore, our proposed framework can accommodate more data types, as opposed to the conventional Gaussian distributions.

We assume that model (1) is *sparse*. That is, the number of nonzero $\boldsymbol{\beta}^*$ components is less than n. When $p > n$, the essential problem is to recover the set of predictors with nonzero coefficients. The past two decades have seen many regularization methods developed for variable selection and estimation in high-dimensional settings, including Lasso (Tibshirani, 1996), adaptive Lasso (Zou, 2006), group Lasso (Yuan and Lin, 2006), SCAD (Fan and Li, 2001) and MCP (Zhang, 2010), among many others. Most regularization methods assume the restrictive β-min condition which requires that the strength of nonzero β_j^*'s is larger than a certain noise level (Zhang and Zhang, 2014). Hence, regularization methods may fail to detect weak signals with nonzero but small β_j^*'s, and this will result in biased estimates and inaccurate predictions, especially when weak signals outnumber strong signals.

Detection of weak signals is challenging. However, if weak signals are partially correlated with strong signals which satisfy the β-min condition, they may be more reliably detected. To elaborate on this idea, first notice that the regression coefficient β_j^* can be written as

$$\beta_j^* = \sum_{1 \le j' \le p} \Omega_{jj'} \mathrm{cov}(X_{j'}, y), \quad j = 1, \ldots, p, \tag{2}$$

where $\Omega_{jj'}$ is the jj'-th entry of $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$, the precision matrix of $\mathbf{x}$. Let $\rho_{jj'}$ be the partial correlation of X_j and $X_{j'}$, i.e. the correlation between the residuals of X_j and $X_{j'}$ after regressing them on all the other X variables. It can be shown that $\rho_{jj'} = -\Omega_{jj'}/\sqrt{\Omega_{jj}\Omega_{j'j'}}$. Hence, that X_j and $X_{j'}$ are partially uncorrelated is equivalent to $\Omega_{jj'} = 0$. Assume that $\mathbf{\Omega}$ is a sparse matrix with only a few nonzero entries in $\mathbf{\Omega}$. When the right hand side of (2) can be accurately evaluated, weak signals can be distinguished from those of noises. In high-dimensional settings, it is impossible to accurately evaluate $\sum_{1 \le j' \le p} \Omega_{jj'} \mathrm{cov}(X_{j'}, y)$. However, under the faithfulness condition that will be introduced in Section 3, a variable, say, indexed by j', satisfing the β-min condition will have a nonzero $\mathrm{cov}(X_{j'}, y)$. Once we identify such strong signals, we set to discover variables that are partially correlated with them.

For brevity, we term weak signals which are partially correlated with strong signals as "weak but correlated" (WBC) signals. This paper aims to incorporate WBC signals in variable selection, estimation and prediction. We propose a two-stage procedure which consists of variable selection and post-selection estimation. The variable selection stage involves a covariance-insured screening for detecting weak signals, and the post-selection estimation stage involves a shrinkage estimator for jointly estimating strong and weak signals selected from the first stage. We call the proposed method as the covariance-insured screening based post-selection shrinkage estimator (CIS-PSE). Our simulation studies demonstrate that by incorporating WBC signals, CIS-PSE improves estimation and prediction accuracy. We also establish the asymptotic selection consistency of CIS-PSE.

The paper is organized as follows. We outline the proposed CIS-PSE method in Section 2 and

investigate its asymptotic properties in Section 3. We evaluate the finite-sample performance of CIS-PSE via simulations in Section 4, and apply the proposed method to predict the annual gross domestic product (GDP) rates based on the socioeconomic status for 82 countries in Section 5. We conclude the paper with a brief discussion in Section 6. All technical proofs are provided in Appendix.

2 Methods

2.1 Notation

We use scripted upper-case letters, such as $\mathcal{S}$, to denote the subsets of $\{1, \ldots, p\}$. Denote by $|\mathcal{S}|$ the cardinality of $\mathcal{S}$ and by $\mathcal{S}^c$ the complement of $\mathcal{S}$. For a vector $\mathbf{v}$, we denote a subvector of $\mathbf{v}$ indexed by $\mathcal{S}$ by $\mathbf{v}_{\mathcal{S}}$. Let $\mathbf{X}_{\mathcal{S}} = (\mathbf{x}_j, j \in \mathcal{S})$ be a submatrix of the design matrix $\mathbf{X}$ restricted to the columns indexed by $\mathcal{S}$. For the symmetric covariance matrix $\mathbf{\Sigma}$, denote by $\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}'}$ its submatix with the row and column indices restricted to subsets $\mathcal{S}$ and $\mathcal{S}'$, respectively. When $\mathcal{S} = \mathcal{S}'$, we write $\mathbf{\Sigma}_{\mathcal{S}} = \mathbf{\Sigma}_{\mathcal{S}\mathcal{S}'}$ for short. The notation also applies to its sample version $\hat{\mathbf{\Sigma}}$.

Denote by $\mathcal{G}(\mathcal{V}, \mathcal{E}; \mathbf{\Omega})$ the graph induced by $\mathbf{\Omega}$, where the node set is $\mathcal{V} = \{1, \ldots, p\}$ and the set of edges is denoted by $\mathcal{E}$. An edge is a pair of nodes, say, k and k', with $\Omega_{kk'} \neq 0$. For a subset $\mathcal{V}_l \subset \mathcal{V}$, denote by $\mathbf{\Omega}_l$ the principal submatrix of $\mathbf{\Omega}$ with its row and column indices restricted to $\mathcal{V}_l$ and by $\mathcal{E}_l$ the corresponding edge set. The subgraph $\mathcal{G}(\mathcal{V}_l, \mathcal{E}_l, \mathbf{\Omega}_l)$ is a connected component of $\mathcal{G}(\mathcal{V}, \mathcal{E}; \mathbf{\Omega})$ if any two nodes in $\mathcal{V}_l$ are connected by edges in $\mathcal{E}$, and if $k \in \mathcal{V}_l^c$, then $\Omega_{kk'} = 0$ for any $k' \in \mathcal{V}_l$.

For a symmetric matrix $\mathbf{A}$, denote by $tr(\mathbf{A})$ the trace of $\mathbf{A}$, and denote by $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ the minimum and maximum eigenvalues of $\mathbf{A}$. We define the operator norm and the Frobenius norm as $\|\mathbf{A}\| = \lambda_{\max}^{1/2}(\mathbf{A}^{\mathrm{T}}\mathbf{A})$ and $\|\mathbf{A}\|_F = tr(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{1/2}$, respectively. For a p-vector $\mathbf{v}$, denote its L_q norm by $\|\mathbf{v}\|_q = (\sum_{j=1}^p |v_j|^q)^{1/q}$ with $q \geq 1$. For any real numbers a and b, denote by $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$.

Denote the sample covariance matrix and the marginal sample covariance between X_j and y, $j = 1, \ldots, p$, by

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}^{(i)}(\mathbf{x}^{(i)})^{\mathrm{T}} \quad \text{and} \quad \widehat{\mathrm{cov}}(X_j, y) = \frac{1}{n}\sum_{i=1}^{n} X_{ij} y_i.$$

For a vector $\mathbf{V} = (V_1, \ldots, V_p)^{\mathrm{T}}$, denote $\widehat{\mathrm{cov}}(\mathbf{V}, y) = (\widehat{\mathrm{cov}}(V_1, y), \ldots, \widehat{\mathrm{cov}}(V_p, y))^{\mathrm{T}}$.

2.2 Defining strong and weak signals

Consider a low-dimensional linear regression model where $p < n$. The ordinary least squares (OLS) estimator $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} = \hat{\mathbf{\Sigma}}^{-1}\widehat{\mathrm{cov}}(\mathbf{x}, y) = \hat{\mathbf{\Omega}}\widehat{\mathrm{cov}}(\mathbf{x}, y)$ minimizes the prediction error, where $\hat{\mathbf{\Omega}} = \hat{\mathbf{\Sigma}}^{-1}$ is the empirical precision matrix. It is also known that $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ is an unbiased estimator of $\boldsymbol{\beta}^*$ and yields the best outcome prediction $\hat{\mathbf{y}}^{\mathrm{best}} = \mathbf{X}\hat{\mathbf{\Omega}}\widehat{\mathrm{cov}}(\mathbf{x}, y)$ with the minimal prediction error.

However, when $p > n$, $\hat{\mathbf{\Sigma}}$ becomes non-invertible, and thus $\boldsymbol{\beta}$ cannot be estimated using all $\mathbf{X}$ variables. Let $\mathcal{S}_0 = \{j : \beta_j^* \neq 0\}$ be the true signal set and assume that $|\mathcal{S}_0| < n$. If $\mathcal{S}_0$ were known, the predicted outcome, $\hat{\mathbf{y}}^{\mathrm{best}} = \mathbf{X}_{\mathcal{S}_0}\hat{\mathbf{\Sigma}}_{\mathcal{S}_0}^{-1}\widehat{\mathrm{cov}}(\mathbf{x}_{\mathcal{S}_0}, y)$, would have the smallest prediction error. In practice, $\mathcal{S}_0$ is unknown and some variable selection method must be applied first to identify $\mathcal{S}_0$. We define the set of strong signals as

$$\mathcal{S}_1 = \left\{ j : |\beta_j^*| > c\sqrt{\log p/n} \text{ for some } c > 0, 1 \leq j \leq p \right\} \tag{3}$$

and let $\mathcal{S}_2 = \mathcal{S}_0 \setminus \mathcal{S}_1$ be the set of weak signals. Then, the OLS estimator and the best outcome prediction are given by

$$\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_{\mathcal{S}_1}^{\mathrm{OLS}} \\ \hat{\boldsymbol{\beta}}_{\mathcal{S}_2}^{\mathrm{OLS}} \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{\Omega}}_{11}\widehat{\mathrm{cov}}(\mathbf{x}_{\mathcal{S}_1}, y) + \hat{\mathbf{\Omega}}_{12}\widehat{\mathrm{cov}}(\mathbf{x}_{\mathcal{S}_2}, y) \\ \hat{\mathbf{\Omega}}_{21}\widehat{\mathrm{cov}}(\mathbf{x}_{\mathcal{S}_1}, y) + \hat{\mathbf{\Omega}}_{22}\widehat{\mathrm{cov}}(\mathbf{x}_{\mathcal{S}_2}, y) \end{pmatrix} \quad \text{and}$$

$$\hat{\mathbf{y}}^{\mathrm{best}} = \mathbf{X}_{\mathcal{S}_1}\hat{\mathbf{\Omega}}_{11}\widehat{\mathrm{cov}}(\mathbf{x}_{\mathcal{S}_1}, y) + \mathbf{X}_{\mathcal{S}_2}\hat{\mathbf{\Omega}}_{21}\widehat{\mathrm{cov}}(\mathbf{x}_{\mathcal{S}_1}, y) + \mathbf{X}_{\mathcal{S}_1}\hat{\mathbf{\Omega}}_{12}\widehat{\mathrm{cov}}(\mathbf{x}_{\mathcal{S}_2}, y) + \mathbf{X}_{\mathcal{S}_2}\hat{\mathbf{\Omega}}_{22}\widehat{\mathrm{cov}}(\mathbf{x}_{\mathcal{S}_2}, y),$$

where $\begin{pmatrix} \hat{\mathbf{\Omega}}_{11} & \hat{\mathbf{\Omega}}_{12} \\ \hat{\mathbf{\Omega}}_{21} & \hat{\mathbf{\Omega}}_{22} \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{\Sigma}}_{\mathcal{S}_1} & \hat{\mathbf{\Sigma}}_{\mathcal{S}_1\mathcal{S}_2} \\ \hat{\mathbf{\Sigma}}_{\mathcal{S}_2\mathcal{S}_1} & \hat{\mathbf{\Sigma}}_{\mathcal{S}_2} \end{pmatrix}^{-1}$ is the partitioned empirical precision matrix. We observe that the partial correlations between the variables in $\mathcal{S}_1$ and $\mathcal{S}_2$ contribute to the estimation of $\boldsymbol{\beta}_{\mathcal{S}_1}$ and $\boldsymbol{\beta}_{\mathcal{S}_2}$ as well as outcome prediction. Therefore, incorporating WBC signals helps reduce the estimation bias and prediction error.

We further decompose $\mathcal{S}_2 = \mathcal{S}_{\text{WBC}} \cup \mathcal{S}_{2*}$, where $\mathcal{S}_{\text{WBC}}$ and $\mathcal{S}_{\text{WBC}}$ are the sets of weak signals with nonzero and zero partial correlations with the signals in $\mathcal{S}_1$, respectively. Formally, with c given in (3),

$$\mathcal{S}_{\text{WBC}} = \left\{ j : 0 < |\beta_j^*| < c\sqrt{\log p/n} \text{ and } \Omega_{jj'} \neq 0 \text{ for some } j' \in \mathcal{S}_1, 1 \leq j \leq p \right\}$$

and

$$\mathcal{S}_{2*} = \left\{ j : 0 < |\beta_j^*| < c\sqrt{\log p/n} \text{ and } \Omega_{jj'} = 0 \text{ for any } j' \in \mathcal{S}_1, 1 \leq j \leq p \right\}.$$

Thus, $\{1, \ldots, p\} = \mathcal{S}_1 \cup \mathcal{S}_{\text{WBC}} \cup \mathcal{S}_{2*} \cup \mathcal{S}_{\text{null}}$, where $\mathcal{S}_{\text{null}} = \{j : \beta_j^* = 0\}$. We assume that $|\mathcal{S}_1| = p_1$, $|\mathcal{S}_{\text{WBC}}| = p_{\text{WBC}}$ and $|\mathcal{S}_{2*}| = p_{2*}$.

2.3 Covariance-insured screening based post-selection shrinkage estimator (CIS-PSE)

Our proposed CIS-PSE method consists of the variable selection and post-shrinkage estimation steps.

Variable selection: First, we detect strong signals by regularization methods such as Lasso or adaptive Lasso. Denote by $\hat{\mathcal{S}}_1$ the set of detected strong signals. To identify WBC signals, we evaluate (2) for each $j \in \hat{\mathcal{S}}_1^c$. When there is no confusion, we use a j' to denote a strong signal.

Though estimating $\text{cov}(X_{j'}, y)$ for every $1 \leq j' \leq p$ can be easily done, identifying and estimating nonzero entries in $\boldsymbol{\Omega}$ is still challenging in high-dimensional settings. However, for identifying WBC signals, it is unnecessary to estimate the whole $\boldsymbol{\Omega}$ matrix. Leveraging intra-feature correlations among predictors, we introduce a computationally efficient method for detecting nonzero $\Omega_{jj'}$'s.

Variables that are partially correlated with signals in $\hat{\mathcal{S}}_1$ form the connected components of $\mathcal{G}(\mathcal{V}, \mathcal{E}; \boldsymbol{\Omega})$ that contain at least one element of $\hat{\mathcal{S}}_1$. Therefore, for detecting WBC signals, it suffices to focus on such connected components. Under the sparsity assumptions of $\boldsymbol{\beta}^*$ and $\boldsymbol{\Omega}$,

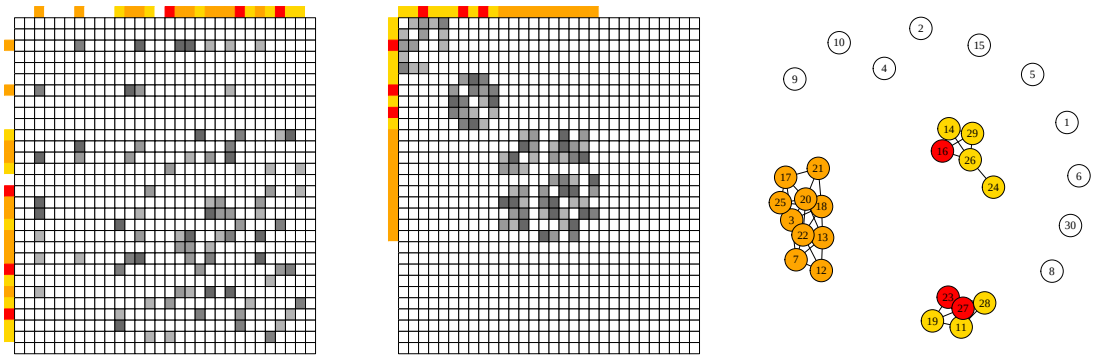

Figure 1: An illustrative example of marginally strong signals and their connected components in $\mathcal{G}(\mathcal{V}, \mathcal{E}; \mathbf{\Omega})$. Left panel: structure of $\mathbf{\Omega}$; Middle panel: structure of $\mathbf{\Omega}$ after properly reordering the row and column indices of $\mathbf{\Omega}$; Right panel: the corresponding graph structure and connected components of the strong signals. Signals in $\mathcal{S}_1$ are colored red. Signals in $\mathcal{S}_{2*}$ are colored orange. WBC signals in $\mathcal{S}_{\text{WBC}}$ are colored yellow.

the size of such connected components is relatively small. For example, as shown in Figure 1, the first two diagonal blocks of a moderate size are relevant for detection of WBC signals.

Under the sparsity assumption of $\mathbf{\Omega}$, the connected components of $\mathbf{\Omega}$ can be inferred from those of the thresholded sample covariance matrix (Mazumder and Hastie, 2012; Bickel and Levina, 2008; Fan et al., 2011; Shao et al., 2011), which is much easier to estimate and can be calculated in a parallel manner. Denote by $\tilde{\mathbf{\Sigma}}^{\alpha}$ the thresholded sample covariance matrix with a thresholding parameter α, where $\tilde{\Sigma}^{\alpha}_{kk'} = \hat{\Sigma}_{kk'} 1\{|\hat{\Sigma}_{kk'}| \geq \alpha\}$, $1 \leq k, k' \leq p$ with $1(\cdot)$ being the indicator function. Denote by $\mathcal{G}(\mathcal{V}, \tilde{\mathcal{E}}; \tilde{\mathbf{\Sigma}}^{\alpha})$ the graph corresponding to $\tilde{\mathbf{\Sigma}}^{\alpha}$. For variable k, $1 \leq k \leq p$, denote by $\mathcal{C}_{[k]}$ the vertex set of the connected component in $\mathcal{G}(\mathcal{V}, \mathcal{E}; \mathbf{\Omega})$ containing k. If variables k and k' belong to the same connected component, $1 \leq k \neq k' \leq p$, then $\mathcal{C}_{[k]} = \mathcal{C}_{[k']}$. For example, $\mathcal{C}_{[14]} = \mathcal{C}_{[16]} = \mathcal{C}_{[24]} = \mathcal{C}_{[26]} = \mathcal{C}_{[29]}$ in the third panel of Figure 1. Clearly, when $k' \notin \mathcal{C}_{[k]}$, $\Omega_{kk'} = 0$, evaluating (2) is equivalent to estimating

$$\beta^*_j = \sum_{j' \in \mathcal{C}_{[j]}} \Omega_{jj'} \text{cov}(X_{j'}, y), \qquad j = 1, \ldots, p. \tag{4}$$

Correspondingly, for a variable k, $1 \leq k \leq p$, denote by $\hat{\mathcal{C}}_{[k]}$ the vertex set of the connected component in $\mathcal{G}(\mathcal{V}, \tilde{\mathcal{E}}; \tilde{\mathbf{\Sigma}}^{\alpha})$ containing k. When $\mathbf{x}$ follows a multivariate Gaussian distribution,

Mazumder and Hastie (2012) showed that $\mathcal{C}_{[k]}$'s can be exactly recovered from $\hat{\mathcal{C}}_{[k]}$'s with a properly chosen α. For a multivariate sub-Gaussian $\mathbf{x}$, the same results follow as shown in the following lemma.

Lemma 2.1. *Suppose that the maximum size of a connected component in* $\mathbf{\Omega}$ *containing a variable in* $\mathcal{S}_0$ *is of order* $O(\exp(n^{\xi}))$*, for some* $0 < \xi < 1$*, then under Assumption (A7) specified in Section 3, with an* $\alpha = O(\sqrt{n^{\xi-1}})$ *and for any variable* k*,* $1 \le k \le p$*, we have*

$$P(\mathcal{C}_{[k]} = \hat{\mathcal{C}}_{[k]}) \ge 1 - C_1 n^{\xi} \exp\left(-C_2 n^{1+\xi}\right) \to 1 \tag{5}$$

for some positive constants C_1 *and* C_2*.*

We summarize the variable selection procedure for $\mathcal{S}_1$ and $\mathcal{S}_{\text{WBC}}$.

Step 1 (Detection of $\mathcal{S}_1$): Obtain a candidate subset $\hat{\mathcal{S}}_1$ of strong signals using a penalized regression method. We consider the following penalized least squares (PLS) estimator:

$$\hat{\boldsymbol{\beta}}^{\text{PLS}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{p} \text{Pen}_{\lambda}(\beta_j) \right\}, \tag{6}$$

where $\text{Pen}_{\lambda}(\beta_j)$ is a penalty on each individual β_j to shrink the weak effects toward zeros and select the strong signals, with the tuning parameter $\lambda > 0$ controlling the size of the candidate subset $\hat{\mathcal{S}}_1$. Commonly used penalties are $\text{Pen}_{\lambda}(\beta_j) = \lambda|\beta_j|$ and $\text{Pen}_{\lambda}(\beta_j) = \lambda\omega_j|\beta_j|$ for Lasso and adaptive Lasso, where $\omega_j > 0$ is a known weight.

Step 2 (Detection of $\mathcal{S}_{\text{WBC}}$): First, for a given threshold α, construct a sparse estimate of the covariance matrix $\tilde{\mathbf{\Sigma}}^{\alpha}$. Next, for each selected variable $m \in \hat{\mathcal{S}}_1$, detect $\hat{\mathcal{C}}_{[m]}$, its connected component in $\mathcal{G}(\mathcal{V}, \tilde{\mathcal{E}}; \tilde{\mathbf{\Sigma}}^{\alpha})$. Let $\mathcal{U} = \bigcup_{m \in \hat{\mathcal{S}}_1} \hat{\mathcal{C}}_{[m]}$. According to (4), it suffices to identify WBC signals within $\mathcal{U}$. Let $\tilde{\mathbf{\Sigma}}^{\alpha}_{[m]}$ be the submatrix by restricting the row and column indices of $\tilde{\mathbf{\Sigma}}^{\alpha}$ to $\hat{\mathcal{C}}_{[m]}$. Then by properly re-arranging the rows and columns of $\tilde{\mathbf{\Sigma}}^{\alpha}$ according to $\mathcal{U}$, we can tranform $\tilde{\mathbf{\Sigma}}^{\alpha}$ into a block diagonal matrix as illustrated in Figure 1,

and $(\tilde{\boldsymbol{\Sigma}}^{\alpha})^{-1}$ can be easily computed. Denote $(\tilde{\boldsymbol{\Sigma}}^{\alpha})^{-1}_{jj'}$ as the entry of $(\tilde{\boldsymbol{\Sigma}}^{\alpha})^{-1}$ corresponding to variables j and j'. We then evaluate (4) and select WBC variables by

$$\hat{\mathcal{S}}_{\text{WBC}} = \left\{ j \in \hat{\mathcal{S}}_1^c \cap \mathcal{U} : \left| \sum_{j' \in \hat{\mathcal{S}}_1 \cap \mathcal{U}} (\tilde{\boldsymbol{\Sigma}}^{\alpha})^{-1}_{jj'} \widehat{\text{cov}}(X_{j'}, y) \right| \geq \nu_n \right\} \tag{7}$$

for some pre-specified $\nu_n > 0$.

Step 3 (Detection of $\mathcal{S}_{2*}$): To identify $\hat{\mathcal{S}}_{2*}$, we first solve a regression problem with a ridge penalty only on variables in $\hat{\mathcal{S}}^c_{1\text{WBC}}$, where $\hat{\mathcal{S}}_{1\text{WBC}} = \hat{\mathcal{S}}_1 \cup \hat{\mathcal{S}}_{\text{WBC}}$. That is,

$$\hat{\boldsymbol{\beta}}^{\text{r}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \tilde{\lambda}_n \|\boldsymbol{\beta}_{\hat{\mathcal{S}}^c_{1\text{WBC}}}\|_2^2 \right\}, \tag{8}$$

where $\tilde{\lambda}_n > 0$ is a tuning parameter controlling the overall strength of the variables selected in $\hat{\mathcal{S}}^c_{1\text{WBC}}$. Then a post-selection weighted ridge (WR) estimator $\hat{\boldsymbol{\beta}}^{\text{WR}}$ has the form

$$\hat{\beta}_j^{\text{WR}} = \begin{cases} \hat{\beta}_j^{\text{r}}, & j \in \hat{\mathcal{S}}_{1\text{WBC}}, \\ \hat{\beta}_j^{\text{r}} 1(|\hat{\beta}_j^{\text{r}}| > a_n), & j \in \hat{\mathcal{S}}^c_{1\text{WBC}}, \end{cases} \tag{9}$$

where a_n is a thresholding parameter. Then the candidate subset $\hat{\mathcal{S}}_{2*}$ is obtained by

$$\hat{\mathcal{S}}_{2*} = \{ j \in \hat{\mathcal{S}}^c_{1\text{WBC}} : \hat{\beta}_j^{\text{WR}} \neq 0, 1 \leq j \leq p \}. \tag{10}$$

Post-selection shrinkage estimation: We consider the following two cases when performing the post-selection shrinkage estimation.

Case 1: $\hat{p}_1 + \hat{p}_{\text{WBC}} + \hat{p}_{2*} < n$. We obtain the CIS-PSE on $\hat{\mathcal{S}}_0$ by

$$\hat{\boldsymbol{\beta}}^{\text{CIS-PSE}}_{\hat{\mathcal{S}}_0} = \hat{\boldsymbol{\Sigma}}^{-1}_{\hat{\mathcal{S}}_0} \widehat{\text{cov}}(\mathbf{x}_{\hat{\mathcal{S}}_0}, y),$$

where $\hat{\mathcal{S}}_0 = \hat{\mathcal{S}}_1 \cup \hat{\mathcal{S}}_{\text{WBC}} \cup \hat{\mathcal{S}}_{2*}$. Then $\hat{\boldsymbol{\beta}}^{\text{CIS-PSE}}_{\hat{\mathcal{S}}_1}$ and $\hat{\boldsymbol{\beta}}^{\text{CIS-PSE}}_{\hat{\mathcal{S}}_{\text{WBC}}}$ can be obtained by restricting $\hat{\boldsymbol{\beta}}^{\text{CIS-PSE}}_{\hat{\mathcal{S}}_0}$ to $\hat{\mathcal{S}}_1$ and $\hat{\mathcal{S}}_{\text{WBC}}$, respectively.

Case 2: $\hat{p}_1 + \hat{p}_{\text{WBC}} + \hat{p}_{2*} \geq n$. Recall that $\hat{\boldsymbol{\beta}}^{\text{WR}}_{\hat{\mathcal{S}}_{1\text{WBC}}} = (\hat{\beta}^{\text{r}}_j,\ j \in \hat{\mathcal{S}}_{1\text{WBC}})^{\text{T}}$ and $\hat{\boldsymbol{\beta}}^{\text{WR}}_{\hat{\mathcal{S}}_{2*}} = (\hat{\beta}^{\text{r}}_j 1(|\hat{\beta}^{\text{r}}_j| > a_n),\ j \in \hat{\mathcal{S}}_{2*})^{\text{T}}$. We obtain the CIS-PSE of $\boldsymbol{\beta}_{\hat{\mathcal{S}}_{1\text{WBC}}}$ by

$$\hat{\boldsymbol{\beta}}^{\text{CIS-PSE}}_{\hat{\mathcal{S}}_{1\text{WBC}}} = \hat{\boldsymbol{\beta}}^{\text{WR}}_{\hat{\mathcal{S}}_{1\text{WBC}}} - \left(\frac{\hat{s}_2 - 2}{\hat{T}_n} \wedge 1\right)(\hat{\boldsymbol{\beta}}^{\text{WR}}_{\hat{\mathcal{S}}_{1\text{WBC}}} - \hat{\boldsymbol{\beta}}^{\text{RE}}_{\hat{\mathcal{S}}_{1\text{WBC}}}), \tag{11}$$

where $\hat{s}_2 = |\hat{\mathcal{S}}_{2*}|$, the post-selection OLS estimator $\hat{\boldsymbol{\beta}}^{\text{RE}}_{\hat{\mathcal{S}}_{1\text{WBC}}}$ restricted to $\hat{\mathcal{S}}_{1\text{WBC}}$ is constructed by

$$\hat{\boldsymbol{\beta}}^{\text{RE}}_{\hat{\mathcal{S}}_{1\text{WBC}}} = \hat{\boldsymbol{\Sigma}}^{-1}_{\hat{\mathcal{S}}_{1\text{WBC}}} \widehat{\text{cov}}(\mathbf{x}_{\hat{\mathcal{S}}_{1\text{WBC}}}, y),$$

and $\hat{T}_n$ is as defined by

$$\hat{T}_n = (\hat{\boldsymbol{\beta}}^{\text{WR}}_{\hat{\mathcal{S}}_{2*}})^{\text{T}}(\mathbf{X}^{\text{T}}_{\hat{\mathcal{S}}_{2*}} \mathbf{M}_{\hat{\mathcal{S}}_{1\text{WBC}}} \mathbf{X}_{\hat{\mathcal{S}}_{2*}}) \hat{\boldsymbol{\beta}}^{\text{WR}}_{\hat{\mathcal{S}}_{2*}} / \hat{\sigma}^2, \tag{12}$$

with $\mathbf{M}_{\hat{\mathcal{S}}_{1\text{WBC}}} = \mathbf{I}_n - \mathbf{X}_{\hat{\mathcal{S}}_{1\text{WBC}}}(\mathbf{X}^{\text{T}}_{\hat{\mathcal{S}}_{1\text{WBC}}} \mathbf{X}_{\hat{\mathcal{S}}_{1\text{WBC}}})^{-1} \mathbf{X}^{\text{T}}_{\hat{\mathcal{S}}_{1\text{WBC}}}$ and $\hat{\sigma}^2 = \sum_{i=1}^n (y_i - \mathbf{X}^{\text{T}}_{\hat{\mathcal{S}}_{2*}} \hat{\boldsymbol{\beta}}^{\text{WR}}_{\hat{\mathcal{S}}_{2*}})^2/(n - \hat{s}_2)$. If $\mathbf{X}^{\text{T}}_{\hat{\mathcal{S}}_{1\text{WBC}}} \mathbf{X}_{\hat{\mathcal{S}}_{1\text{WBC}}}$ is singular, we replace $(\mathbf{X}^{\text{T}}_{\hat{\mathcal{S}}_{1\text{WBC}}} \mathbf{X}_{\hat{\mathcal{S}}_{1\text{WBC}}})^{-1}$ with its generalized inverse. Then $\hat{\boldsymbol{\beta}}^{\text{CIS-PSE}}_{\hat{\mathcal{S}}_1}$ and $\hat{\boldsymbol{\beta}}^{\text{CIS-PSE}}_{\hat{\mathcal{S}}_{\text{WBC}}}$ can be obtained by restricting $\hat{\boldsymbol{\beta}}^{\text{CIS-PSE}}_{\hat{\mathcal{S}}_{1\text{WBC}}}$ to $\hat{\mathcal{S}}_1$ and $\hat{\mathcal{S}}_{\text{WBC}}$, respectively.

2.4 Selection of tuning parameters

When selecting strong signals, the tuning parameter λ in Lasso or adaptive Lasso can be chosen by BIC (Zou, 2006). To choose ν_n for the selection of WBC signals according to (7), we rank variables $j \in \hat{\mathcal{S}}_1^c \cap \mathcal{U}(\alpha, \hat{\mathcal{S}}_1)$ according to the magnitude of $\left|\sum_{j' \in \hat{\mathcal{C}}_{[j]}} (\tilde{\boldsymbol{\Sigma}}^{\alpha}_{[j]})^{-1}_{jj'} \widehat{\text{cov}}(X_{j'}, y)\right|$, and select the first $r \leq n - |\hat{\mathcal{S}}_1|$ variables to be $\hat{\mathcal{S}}_{\text{WBC}}$. Specifically, r can be chosen such that $\hat{\mathcal{S}}_{1\text{WBC}}$ minimizes the average prediction error in an independent validation dataset. For tuning parameter α, we set $\alpha = c_3 \log(n)$, for some positive constant c_3, as suggested in Shao et al. (2011). Our empirical experiments show that $\alpha = c_3 \log(n)$ tends to give the larger true positives and the smaller false positives in identifying WBC variables. Figure 7 in Appendix reveals that in order to find the optimal α that minimizes the prediction error on a validation dataset, it suffices to conduct a grid search with only a few proposed values of α. In our numerical studies, instead of thresholding the sample covariance matrix, we threshold the sample correlation matrix. As

correlations are ranged between -1 and 1, it is easier to set a target range for α. To detect signals in $\mathcal{S}_{2*}$, we follow Gao et al. (2017) to use cross-validation to choose $\tilde{\lambda}_n$ and a_n in (8) and (9), respectively. In particular, we set $\tilde{\lambda}_n = c_1 a_n^{-2}(\log\log n)^3 \log(n \vee p)$ and $a_n = c_2 n^{-1/8}$ for some positive constants c_1 and c_2. In the training dataset we fix the tuning parameters and fit the model, and in the validation dataset we compute the prediction error of the model. We repeat this procedure for various c_1 and c_2, and choose a pair that gives the smallest prediction error on the validation dataset.

3 Asymptotic properties

To investigate the asymptotic properties of CIS-PSE, we assume the following.

(A1) The random error ϵ has a finite kurtosis.

(A2) $\log(p) = O(n^\nu)$ for some $0 < \nu < 1$.

(A3) There are positive constants κ_1 and κ_2 such that $0 < \kappa_1 < \lambda_{\min}(\mathbf{\Sigma}) \leq \lambda_{\max}(\mathbf{\Sigma}) < \kappa_2 < \infty$.

(A4) Sparse Riesz condition (SRC): For the random design matrix $\mathbf{X}$, any $\mathcal{S} \subset \{1, \ldots, p\}$ with $|\mathcal{S}| = q$, $q \leq p$, and any vector $\mathbf{v} \in \mathbb{R}^q$, there exist $0 < c_* < c^* < \infty$ such that $c_* \leq \|\mathbf{X}_S^{\mathrm{T}}\mathbf{v}\|_2^2 / \|\mathbf{v}\|_2^2 \leq c^*$ holds with probability tending to 1.

(A5) Faithfulness Assumption: Suppose that

$$\max |\mathbf{\Sigma}_{\mathcal{S}_1^c \mathcal{S}_1} \boldsymbol{\beta}^*_{\mathcal{S}_1}| + \max \left|\mathbf{\Sigma}_{\mathcal{S}_1^c \mathcal{S}_2} \boldsymbol{\beta}^*_{\mathcal{S}_2}\right| + \min \left|\mathbf{\Sigma}_{\mathcal{S}_1 \mathcal{S}_2} \boldsymbol{\beta}^*_{\mathcal{S}_2}\right| < \min |\mathbf{\Sigma}_{\mathcal{S}_1 \mathcal{S}_1} \boldsymbol{\beta}^*_{\mathcal{S}_1}|,$$

where the absolute value function $|\cdot|$ is applied component-wise to its argument vector. The max and min operators are with respect to all individual components in the argument vectors.

(A6) Denote by $C_{\max} = \max_{1 \leq l \leq B} |\mathcal{V}_l|$ the maximum size of the connected components in graph $\mathcal{G}(\mathcal{V}, \mathcal{E}; \mathbf{\Omega})$ that contains at least one signal in $\mathcal{S}_1$, where B is the number of such connected components. Assume $C_{\max} = O(n^\xi)$ for some $\xi \in (0, 1)$.

(A7) Assume $\min_{(k,k') \in \mathcal{E}} |\Sigma_{kk'}| \geq C\sqrt{n^{\xi - 1}}$ for some constant $C > 0$ and $\max_{(k,k') \notin \mathcal{E}} |\Sigma_{kk'}| = o(\sqrt{n^{\xi-1}})$ for the ξ in (A6).

(A8) For any subset $\mathcal{V}_l \subset \{1, \ldots, p\}$ with $|\mathcal{V}_l| = O(n)$, $\sup_j E|X_{ij}1(j \in \mathcal{V}_l)|^{2\zeta} < \infty$ for some positive constant ζ.

(A9) Assume that $\|\beta^*_{\mathcal{S}_{2*}}\|_2 = o(n^\tau)$ for some $0 < \tau < 1$, where $\|\cdot\|_2$ is the Euclidean norm.

(A1), a technical assumption for the asymptotic proofs, is satisfied by many parametric distributions such as Gaussian. The assumption is mild as we do not assume any parametric distributions for ε except that it has finite moments. (A2) and (A3) are commonly assumed in the high-dimensional literature. (A4) guarantees that $\mathcal{S}_1$ can be recovered with probability tending to 1 as $n \to \infty$ (Zhang and Huang, 2008). (A5) ensures that for all $j \in \mathcal{S}_1$, $\min_{j\in\mathcal{S}_1} |\widehat{\text{cov}}(X_j, y)| > \max_{j\in\mathcal{S}_1^c} |\widehat{\text{cov}}(X_j, y)|$ holds with probability tending to 1 (Lemma 4 in Genovese et al., 2012). (A6) implies that the size of each connected component of a strong signal, i.e., $\mathcal{C}_{[j']}$, $j' \in \mathcal{S}_1$, cannot exceed the order of $\exp(n^\xi)$ for some $\xi \in (0, 1)$. This assumption is required for estimating sparse covariance matrices. (A7) guarantees that with a properly chosen thresholding parameter α, X_k and $X_{k'}$ have non-zero thresholded sample covariances for $(k, k') \in \mathcal{E}$, and have zero thresholded sample covariances for $(k, k') \notin \mathcal{E}$. As a result, the connected components of the thresholded sample covariance matrix and those of the precision matrix can be detected with adequate accuracy. (A8) ensures that the precision matrix can be accurately estimated by inverting the thresholded sample covariance matrix; see Shao et al. (2011) and Bickel and Levina (2008) for details. (A9), which bounds the total size of weak signals on $\mathcal{S}_{2*}$, is required for selection consistency on $\mathcal{S}_{2*}$ (Gao et al., 2017).

We show that given a consistently selected $\mathcal{S}_1$, we have selection consistency for $\mathcal{S}_{\text{WBC}}$.

Theorem 3.1. *With (A1)-(A3) and (A6)-(A8),*

$$\lim_{n\to\infty} P\left(\hat{\mathcal{S}}_{WBC} = \mathcal{S}_{WBC} | \hat{\mathcal{S}}_1 = \mathcal{S}_1\right) = 1.$$

The following corollary shows that Theorem 3.1, together with Theorem 2 in Zhang and Huang (2008) and Corollary 2 in Gao et al. (2017), further implies selection consistency for $\mathcal{S}_1 \cup \mathcal{S}_{\text{WBC}} \cup \mathcal{S}_{2*}$.

Corollary 3.2. *Under Assumptions (A1)-(A9), we have*

$$\lim_{n\to\infty} P\left(\{\hat{\mathcal{S}}_1 = \mathcal{S}_1\} \cap \{\hat{\mathcal{S}}_{WBC} = \mathcal{S}_{WBC}\} \cap \{\hat{\mathcal{S}}_{2*} = \mathcal{S}_{2*}\}\right) = 1.$$

Corollary 3.2 implies that CIS-PSE can recover the true set asymptotically. Thus, when $|\mathcal{S}_0| < n$, CIS-PSE gives an OLS estimator with probability going to 1 and has the minimum prediction error asymptotically, among all the unbiased estimators.

4 Simulation studies

We conduct simulations to compare the performance of the proposed CIS-PSE and the post-shrinkage estimator (PSE) by Gao et al. (2017). The key difference between CIS-PSE and PSE lies in that PSE focuses only on $\mathcal{S}_1$ whereas CIS-PSE considers $\mathcal{S}_1 \cup \mathcal{S}_{\text{WBC}}$.

Data are generated according to (1) with

$$\boldsymbol{\beta}^* = (\overbrace{20, 20, 20}^{\mathcal{S}_1}, \overbrace{\underbrace{0.5, \ldots, 0.5}_{30}}^{\mathcal{S}_{\text{WBC}}}, \overbrace{\underbrace{0.5, \ldots, 0.5}_{30}}^{\mathcal{S}_{2*}}, \overbrace{\underbrace{0, \ldots, 0}_{p-63}}^{\mathcal{S}_{\text{null}}})^{\text{T}}. \tag{13}$$

The random errors ϵ_i are independently generated from $N(0, 1)$. We consider the following examples.

Example 1: The first three variables, which belong to $\mathcal{S}_1$, are independently generated from $N(0, 1)$. The first ten, next ten and the last ten signals in $\mathcal{S}_{\text{WBC}}$ belong to the connected component of X_1, X_2 and X_3, respectively. These three connected components are independent of each other. $\mathcal{S}_{2*}$ is independent of $\mathcal{S}_1$ and $\mathcal{S}_{\text{WBC}}$. Each connected component within $\mathcal{S}_1 \cup \mathcal{S}_{\text{WBC}}$ and $\mathcal{S}_{2*}$ are generated from a multivariate normal distribution with mean zeros, variance 1, and a compound symmetric (CS) correlation matrix with correlation coefficient of 0.7. Variables in $\mathcal{S}_{\text{null}}$ are independently generated from $N(0, 1)$.

Example 2: This example is the same as Example 1 except that the three connected components within $\mathcal{S}_1 \cup \mathcal{S}_{\text{WBC}}$ and $\mathcal{S}_{2*}$ follow the first order autocorrelation (AR(1)) structure with correlation coefficient of 0.7.

Example 3: This example is the same as Example 1 except that there are 30 variables in $\mathcal{S}_{\text{null}}$ (i.e., variables X_{64}-X_{93}) are set to be correlated with signals in $\mathcal{S}_1$. That is, X_{64}-X_{73} are correlated with X_1, X_{74}-X_{83} are correlated with X_2, and X_{84}-X_{93} are correlated with X_3. These three connected components within $\mathcal{S}_1 \cup \mathcal{S}_{\text{null}}$ have a CS correlation structure with correlation coefficient of 0.7.

For each example, we conduct 500 independent experiments with p=200, 300, 400 and 500. We generate a training dataset of size $n = 200$, a test dataset of size $n = 100$ to assess the prediction performance, and an independent validation dataset of size $n = 100$ for tuning parameter selection.

First, we compare CIS-PSE and PSE in selecting $\mathcal{S}_0$ under Examples 1–2. We use Lasso and adaptive Lasso to select $\hat{\mathcal{S}}_1$. Since both Lasso and adaptive Lasso give similar results, we report only the Lasso results in this section and present the results of adaptive Lasso in Appendix. We report the number of correctly identified variables (TP) in $\mathcal{S}_0$ and the number of incorrectly selected variables (FP) in $\mathcal{S}_0^c$. Table 1 shows that CIS-PSE outperforms PSE in identifying signals in $\mathcal{S}_0$. We observe that the performance of PSE deteriorates as p increases, whereas CIS-PSE selects $\mathcal{S}_0$ signals consistently even when p increases.

Table 1: The performance of variable selection on $\mathcal{S}_0$

			$p = 200$	$p = 300$	$p = 400$	$p = 500$
Example 1	TP	CIS-PSE	59.6 (1.9)	58.7 (2.1)	57.9 (2.3)	57.7 (2.4)
		PSE	41.2 (4.9)	34.4 (5.1)	25.7 (6.0)	22.6 (5.8)
	FP	CIS-PSE	3.7 (2.4)	5.1 (2.7)	6.9 (3.1)	8.8 (3.3)
		PSE	13.3 (4.6)	18.9 (5.2)	21.7 (5.9)	26.1 (6.0)
Example 2	TP	CIS-PSE	63.0 (0)	62.9 (0.1)	62.9 (0.1)	62.9 (0.1)
		PSE	43.9 (3.9)	37.0 (4.2)	32.8 (5.0)	31.5 (4.3)
	FP	CIS-PSE	3.5 (2.4)	5.0 (2.7)	6.3 (3.3)	8.1 (3.1)
		PSE	12.7 (4.2)	19.5 (5.1)	22.1 (6.3)	27.4 (6.4)

NOTE. TP: true positive; FP: false positive.

Next, we evaluate estimation accuracy on the targeted sub-model $\mathcal{S}_1 \cup \mathcal{S}_{\text{WBC}}$ using the mean squared error (MSE) as the criterion under Examples 1–2. Figure 2 indicates that the proposed

CIS-PSE detects WBC signals and provides more accurate and precise estimates. Figure 3 shows that CIS-PSE also improves the estimation of $\beta_{\mathcal{S}_1}$ compared to PSE.

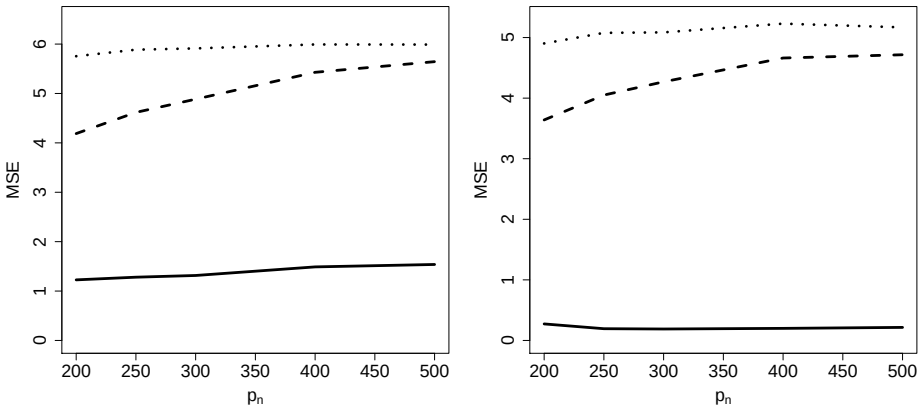

Figure 2: The mean squared error (MSE) of $\hat{\boldsymbol{\beta}}_{\mathcal{S}_{\mathrm{WBC}}}$ for different p's under Example 1 (Left panel) and Example 2 (Right panel). Solid lines represent CIS-PSE, dashed lines are for PSE, and dotted lines indicate Lasso.

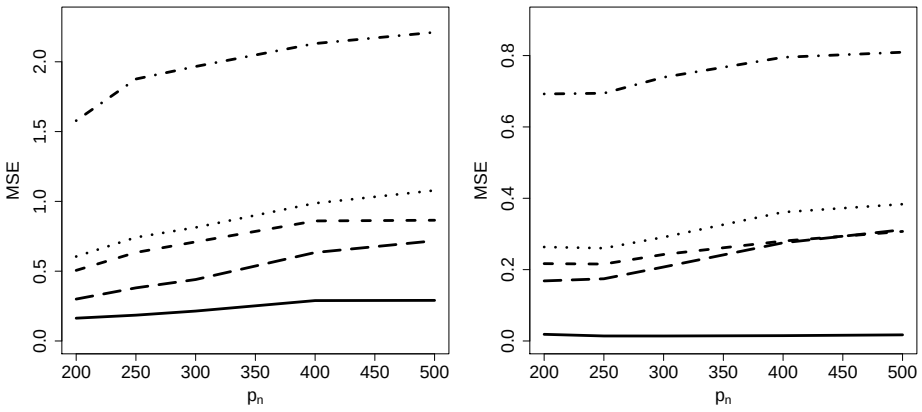

Figure 3: The mean squared error (MSE) of $\hat{\boldsymbol{\beta}}_{\mathcal{S}_1}$ for different p's under Example 1 (Left panel) and Example 2 (Right panel). Solid lines represent CIS-PSE, dashed lines are for PSE, dotted lines indicate Lasso RE defined as $\hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{\hat{\mathcal{S}}_1} = \hat{\boldsymbol{\Sigma}}^{-1}_{\hat{\mathcal{S}}_1} \widehat{\mathrm{cov}}(\mathbf{x}_{\hat{\mathcal{S}}_1}, y)$, dot-dashed lines represent Lasso, and long-dashed lines are for WR in (9).

We explore the prediction performance under Examples 1–2 using the mean squared prediction error (MSPE), defined as $\|\hat{\mathbf{y}} - \mathbf{y}^{\mathrm{test}}\|_2^2/n_{\mathrm{test}} = \|\mathbf{X}\hat{\boldsymbol{\beta}}^{\diamond} - \mathbf{y}^{\mathrm{test}}\|_2^2/n_{\mathrm{test}}$, where $\hat{\boldsymbol{\beta}}^{\diamond}$ is obtained from the training data, $\mathbf{y}^{\mathrm{test}}$ is the response variable for the test dataset, n_{test} is the size of test

dataset, and $\diamond$ represents either the proposed CIS-PSE or PSE. Table 2, which summarizes the results, shows that CIS-PSE outperforms PSE, suggesting incorporating WBC signals helps to improve the prediction accuracy.

Table 2: Mean squared prediction error (MSPE) of the predicted outcomes

	p	200	300	400	500
Example 1	CIS-PSE	3.17 (0.80)	3.19 (0.78)	3.25 (0.77)	3.32 (0.77)
	PSE	4.19 (0.83)	4.93 (1.02)	5.28 (1.07)	5.50 (1.16)
	Lasso	10.28 (5.68)	10.02 (5.58)	9.77 (4.96)	9.78 (4.54)
Example 2	CIS-PSE	0.65 (0.14)	0.92 (0.19)	1.30 (0.16)	2.43 (0.64)
	PSE	2.89 (0.61)	3.55 (0.75)	4.09 (0.68)	4.34 (0.97)
	Lasso	4.20 (0.79)	4.50 (0.88)	4.68 (0.90)	4.73 (0.97)

Lastly, we consider the setting where a subset of $\mathcal{S}_{\text{null}}$ is correlated with a subset of $\mathcal{S}_1$; see Example 3. Compared to Example 1, the results that are summarized in Table 3 show that the number of false positives only slightly increases, when some variables in $\mathcal{S}_{\text{null}}$ are correlated with variables in $\mathcal{S}_1$.

5 A real data example

We apply the proposed CIS-PSE method to analyze the gross domestic product (GDP) growth data studied in Gao et al. (2017) and Barro and Lee (1994). Our goal is to identify factors that are associated with the long-run GDP growth rate. The dataset includes the GDP growth rates and 45 socioeconomic variables for 82 countries from 1960 to 1985. We consider the following

Table 3: Comparison of false positives (standard deviations in parentheses) between Examples 1 and 3

	$p = 200$	$p = 300$	$p = 400$	$p = 500$
Example 1	3.7 (2.4)	5.1 (2.7)	6.9 (3.1)	8.8 (3.3)
Example 3	4.6 (3.5)	6.1 (3.4)	7.8 (3.8)	10.9 (4.2)

model:

$$\mathrm{GR}_i = \beta_0 + \beta_1 \log(\mathrm{GDP60}_i) + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}_2 + 1(\mathrm{GDP60}_i < 2898)(\delta_0 + \delta_1 \log(\mathrm{GDP60}_i) + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\delta}_2) + \varepsilon_i, \quad (14)$$

where i is the country indicator, $i = 1, \ldots, 82$, GR_i is the annualized GDP growth rate of country i from 1960 to 1985, $\mathrm{GDP60}_i$ is the GDP per capita in 1960, and $\mathbf{z}_i$ are 45 socioeconomic covariates, the details of which can be found in Gao et al. (2017). The β_1 and $\boldsymbol{\beta}_2$ represent the coefficients of log(GDP60) and socioeconomic predictors, respectively. The δ_0 represents the coefficient of whether the GDP per capita in 1960 is below a threshold (=2898) or not. The δ_1 represents the coefficient of log(GDP60) when GDP per capita in 1960 is below 2898. The $\boldsymbol{\delta}_2$ represent the coefficients of the interactions between the $\mathrm{GDP60}_i < 2898$ and the socioeconomic predictors when GDP per capita in 1960 is below 2898.

We apply the proposed CIS-PSE and PSE by Gao et al. (2017) to detect $\mathcal{S}_1$. Additionally, CIS-PSE is used to further identify $\mathcal{S}_{\mathrm{WBC}}$. Effects of covariates in $\hat{\mathcal{S}}_1$ are estimated by Lasso, adaptive Lasso, PSE and CIS-PSE. Effects of covariates in $\hat{\mathcal{S}}_{\mathrm{WBC}}$ are estimated by CIS-PSE. The sample correlations between variables in $\hat{\mathcal{S}}_1$ and $\hat{\mathcal{S}}_{\mathrm{WBC}}$ are also provided. Table 4 reports the selected variables and their estimated coefficients.

Next, we evaluate the accuracy of predicted GR using a leave-one-out cross-validation. For each country, we treat it itself as the test set while using all other countries as the training set. We apply Lasso, adaptive Lasso, PSE and CIS-PSE. All tuning parameters are selected as described in Section 4. The prediction results in Figure 4 show that CIS-PSE has the smallest prediction errors compared to PSE, Lasso and adaptive Lasso, with $\hat{\mathcal{S}}_1$ detected by either Lasso or adaptive Lasso.

6 Discussion

To improve the estimation and prediction accuracy in high-dimensional linear regressions, we introduce the concept of weak but correlated (WBC) signals, which are commonly missed by

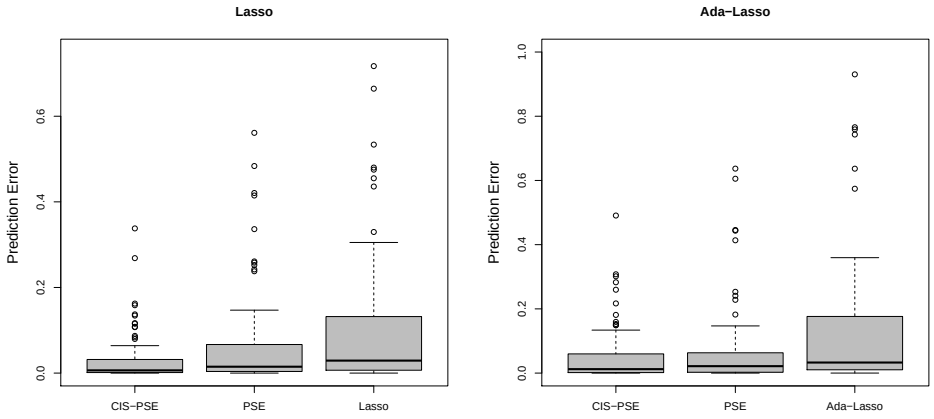

Figure 4: Prediction errors from post-selection shrinkage estimators: CIS-PSE, PSE and two penalized estimators (Lasso and adaptive Lasso). $\hat{\mathcal{S}}_1$ is detected by Lasso in the left panel and by adaptive Lasso in the right panel.

Table 4: Estimation results of $\mathcal{S}_1$ and $\mathcal{S}_{\text{WBC}}$ from the growth rate data

$\hat{\mathcal{S}}_1$ is selected by Lasso					
$\hat{\mathcal{S}}_1$	$\hat{\beta}^{\text{Lasso}}_{\hat{\mathcal{S}}_1}$	$\hat{\beta}^{\text{PSE}}_{\hat{\mathcal{S}}_1}$	$\hat{\beta}^{\text{CIS-PSE}}_{\hat{\mathcal{S}}_1}$	$\hat{\mathcal{S}}_{\text{WBC}}$	$\hat{\beta}^{\text{CIS-PSE}}_{\hat{\mathcal{S}}_{\text{WBC}}}$
TOT	0.06	2.85	3.73	–	–
LFERT	1.66	1.75	2.55	LLIFE	1.88
				NOM60	0.12
				NOF60	-0.10
				LGDP60×NOM60	-0.02
PRIF60	-0.001	-0.002	-0.12	LGDP60×PRIF60	0.02
				LGDP60×PRIM60	-0.02
				LGDP60×NOF60	0.02
$\hat{\mathcal{S}}_1$ selected by adaptive Lasso					
$\hat{\mathcal{S}}_1$	$\hat{\beta}^{\text{Ada-Lasso}}_{\hat{\mathcal{S}}_1}$	$\hat{\beta}^{\text{PSE}}_{\hat{\mathcal{S}}_1}$	$\hat{\beta}^{\text{CIS-PSE}}_{\hat{\mathcal{S}}_1}$	$\hat{\mathcal{S}}_{\text{WBC}}$	$\hat{\beta}^{\text{CIS-PSE}}_{\hat{\mathcal{S}}_{\text{WBC}}}$
LFERT	1.98	2.04	2.54	LLIFE	1.77
				NOM60	0.08
				NOF60	-0.07
				LGDP60×NOM60	-0.01

NOTE. TOT: the term of trade shock; LFERT: log of fertility rate (children per woman) averaged over 1960-1985; LLIFE: log of life expectancy at age 0 averaged over 1960-1985; NOM60: percentage of no schooling in the male population in 1960; NOF60: percentage of no schooling in the female population in 1960; LGDP60: log GDP per capita in 1960 (1985 price); PRIF60: percentage of primary schooling attained in female population in 1960; PRIM60: percentage of primary schooling attained in male population in 1960.

the Lasso-type variable selection methods. We show that these variables can be easily detected with the help of their partial correlations with strong signals. We propose a CIS-PSE procedure for high-dimensional variable selection and estimation, particularly for WBC signal detection and estimation. We show that, by incorporating WBC signals, it significantly improves the estimation and prediction accuracy.

An alternative approach to weak signal detection would be to group them according to a known group structure and then select by their grouped effects (Bodmer and Bonilla, 2008; Li and Leal, 2008; Wu et al., 2011; Yuan and Lin, 2006). However, grouping strategies require prior knowledge on the group structure, and, in some situations, may not amplify the grouped effects of weak signals. For example, as pointed out in Bühlmann et al. (2013) and Shah and Samworth (2013), when a pair of highly negatively correlated variables are grouped together, they cancel out each other's effect. On the other hand, our CIS-PSE method is based on detecting partial correlations and can accommodate the "canceling out" scenarios. Hence, when the grouping structure is known, it is worth combining the grouping strategy and CIS-PSE for weak signal detection. We will pursue this in the future.

7 Appendix

We provide technical proofs for Theorem 3.1, Corollary 3.2 and lemmas in this section. We first list some definitions and auxiliary lemmas.

Definition 7.1. *A random vector* $\mathbf{Z} = (Z_1, \ldots, Z_p)$ *is a sub-Gaussian with mean vector* $\boldsymbol{\mu}$ *and variance proxy* σ_z^2*, if for any* $\mathbf{a} \in \mathbb{R}^p$*,* $E[\exp\{\mathbf{a}^{\mathrm{T}}(\mathbf{Z} - \boldsymbol{\mu})\}] \leq \exp(\sigma_z^2 \|\mathbf{a}\|_2^2/2)$.

Let Z be a sub-Gaussian random variable with variance proxy σ_z^2. The sub-Gaussian tail inequality is given as, for any $t > 0$,

$$P(Z > t) \leq e^{-\frac{t^2}{2\sigma_z^2}} \text{ and } P(Z < -t) \leq e^{-\frac{t^2}{2\sigma_z^2}}.$$

The following Lemma 7.2 ensures that the set of signals with non-vanishing marginal sample correlations with y coincides with $\mathcal{S}_1$ with probability tending to 1. Therefore, evaluating condition (4) for a covariate j is equivalent to estimating nonzero $\Omega_{jj'}$'s for every $j' \in \mathcal{S}_1$. Let $r_{j'}$ be the rank of variable j' according to the magnitude of $|\widehat{\mathrm{cov}}(X_{j'}, y)|$, $j' = 1, \ldots, p$. Denote by $\tilde{\mathcal{S}}_1(k) = \{j' : r_{j'} \leq k\}$ the first k covariates with the largest absolute marginal correlations with y, for $k = 1, \ldots, p$. Recall that $s_1 = |\mathcal{S}_1|$.

Lemma 7.2. *Under Assumption (A5), we have*

$$\lim_{n\to\infty} P\left(\tilde{\mathcal{S}}_1(s_1) = \mathcal{S}_1\right) = 1.$$

Proof of Lemma 7.2. By the definition of $\tilde{\mathcal{S}}_1(s_1)$, it is suffice to show that with probability tending to 1, as $n \to \infty$,

$$\max \left| \frac{1}{n} \mathbf{X}_{\mathcal{S}_1^c}^{\mathrm{T}} \mathbf{y} \right| < \min \left| \frac{1}{n} \mathbf{X}_{\mathcal{S}_1}^{\mathrm{T}} \mathbf{y} \right|.$$

Since $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon} = \mathbf{X}_{\mathcal{S}_0}\boldsymbol{\beta}^*_{\mathcal{S}_0} + \boldsymbol{\varepsilon} = \mathbf{X}_{\mathcal{S}_1}\boldsymbol{\beta}^*_{\mathcal{S}_1} + \mathbf{X}_{\mathcal{S}_2}\boldsymbol{\beta}^*_{\mathcal{S}_2} + \boldsymbol{\varepsilon}$, we have

$$\frac{1}{n}\mathbf{X}_{\mathcal{S}_1^c}^{\mathrm{T}}\mathbf{y} = \frac{1}{n}\mathbf{X}_{\mathcal{S}_1^c}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}_1}\boldsymbol{\beta}^*_{\mathcal{S}_1} + \frac{1}{n}\mathbf{X}_{\mathcal{S}_1^c}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}_2}\boldsymbol{\beta}^*_{\mathcal{S}_2} + \frac{1}{n}\mathbf{X}_{\mathcal{S}_1^c}^{\mathrm{T}}\boldsymbol{\varepsilon}.$$

Notice that for each $j' \in \mathcal{S}_1^c$, $\frac{1}{n}\mathbf{x}_{j'}^{\mathrm{T}}\boldsymbol{\varepsilon} \to \mathrm{cov}(X_{j'}, \varepsilon) = 0$ in probability, then $\max|\frac{1}{n}\mathbf{X}_{\mathcal{S}_1^c}^{\mathrm{T}}\boldsymbol{\varepsilon}| = o_P(1)$ as $n \to \infty$.

It follows that when $n \to \infty$,

$$\max \left| \frac{1}{n} \mathbf{X}_{\mathcal{S}_1^c}^{\mathrm{T}} \mathbf{y} \right| \leq \max \left| \boldsymbol{\Sigma}_{\mathcal{S}_1^c\mathcal{S}_1} \boldsymbol{\beta}^*_{\mathcal{S}_1} \right| + \max \left| \boldsymbol{\Sigma}_{\mathcal{S}_1^c\mathcal{S}_2} \boldsymbol{\beta}^*_{\mathcal{S}_2} \right| + o_P(1).$$

Similarly, when $n \to \infty$,

$$\min \left| \frac{1}{n} \mathbf{X}_{\mathcal{S}_1}^{\mathrm{T}} \mathbf{y} \right| \geq \min \left| \boldsymbol{\Sigma}_{\mathcal{S}_1\mathcal{S}_1} \boldsymbol{\beta}^*_{\mathcal{S}_1} \right| - \min \left| \boldsymbol{\Sigma}_{\mathcal{S}_1\mathcal{S}_2} \boldsymbol{\beta}^*_{\mathcal{S}_2} \right| + o_P(1).$$

Lemma 7.2 is concluded by combining the above two inequalities with the faithfulness condition.

□

Bickel and Levina (2008) showed that for $\alpha = O(\sqrt{\log|\mathcal{S}_{[j]}|/n})$, $\|\tilde{\mathbf{\Sigma}}^{\alpha}_{[j]} - \mathbf{\Sigma}_{[j]}\| = O_P(\rho_n)$, where $\rho_n = O(n^{-1}C^{1/4}_{\max})$ with $C^{1/4}_{\max}$ given in (A6). Furthermore, Bickel and Levina (2008) and Fan et al. (2011) showed that the estimation error for each connected component of the precision matrix is bounded by

$$\|\hat{\mathbf{\Omega}}_{[j]} - \mathbf{\Omega}_{[j]}\| = \|(\tilde{\mathbf{\Sigma}}^{\alpha}_{[j]})^{-1} - \mathbf{\Sigma}^{-1}_{[j]}\| = O_P(\rho_n). \tag{15}$$

To detect the connected components of the thresholded sample covariance matrix, we adopt the recursive labeling Algorithm as in Shapiro and Stockman (2002).

Without loss of generality, suppose that the strong signals in $\hat{\mathcal{S}}_1$ belong to distinct connected components of $\tilde{\mathbf{\Sigma}}^{\alpha}$. We rearrange the indices in $\hat{\mathcal{S}}_1$ as $\{1,\ldots,|\hat{\mathcal{S}}_1|\}$ and write the submatrix of $\tilde{\mathbf{\Sigma}}^{\alpha}$ corresponding to j', $1 \le j' \le |\hat{\mathcal{S}}_1|$, as $\tilde{\mathbf{\Sigma}}^{\alpha}_{j'}$. For notational convenience, we rewrite

$$\hat{\mathbf{\Omega}} = \mathrm{diag}((\tilde{\mathbf{\Sigma}}^{\alpha}_1)^{-1},\ldots,(\tilde{\mathbf{\Sigma}}^{\alpha}_{|\hat{\mathcal{S}}^{(\alpha)}_1|})^{-1},\mathbf{0})_{p\times p}.$$

The following Lemma 7.3 is useful for controlling the size of $\hat{\mathcal{C}}_{[j]}$.

Lemma 7.3. *Under (A6)-(A7), when* $\mathbf{x}$ *is from a multivariate sub-Gaussian distribution, we have* $P\left(|\hat{\mathcal{C}}_{[j]}| \le O(n^{\xi})\right) \ge 1 - n^{\xi}C_1\exp(-C_2n^{1+\xi})$ *for some positive constants* C_1 *and* C_2.

Lemma 7.3 is a direct conclusion of Lemma 2.1 and Assumption (A6). Next we prove Theorem 3.1.

Proof of Theorem 3.1. Notice that $\beta^*_j = \sum_{j'\in\mathcal{C}_{[j]}}\Omega_{jj'}\mathrm{cov}(X_{j'},y)$ and $\hat{\beta}_j = \sum_{j'\in\hat{\mathcal{C}}_{[j]}}\hat{\Omega}_{jj'}\widehat{\mathrm{cov}}(X_{j'},y)$.

Consider a sequence of thresholding parameters $\nu_n = O(n^{3\xi/2})$ with a decreasing series of positive numbers $u_n = 1 + n^{-\xi/4}$ such that $\lim_{n\to\infty}u_n = 1$,

$$\begin{aligned} & P\left(\{j\notin\mathcal{S}_1 : |\beta^*_j| > \nu_n u_n\} \subseteq \{j\notin\mathcal{S}_1 : |\hat{\beta}_j| > \nu_n\}\right) \\ \ge\ & 1 - P\left(\bigcup_{j\notin\mathcal{S}_1, |\beta^*_j|>\nu_n u_n}\{|\hat{\beta}_j| \le \nu_n\}\right). \end{aligned} \tag{16}$$

Moreover, since $\{|\hat{\beta}_j| \le \nu_n\}$ and $\{|\beta_j^*| > \nu_n u_n\}$, we have $|\hat{\beta}_j - \beta_j^*| \ge \nu_n(u_n - 1)$. As a result,

$$\begin{aligned} & 1 - P\left(\bigcup_{j\notin\mathcal{S}_1, |\beta_j^*|>\nu_n u_n} \{|\hat{\beta}_j| \le \nu_n\}\right) \\ \ge\ & 1 - \sum_{j\notin\mathcal{S}_1, |\beta_j^*|>\nu_n u_n} P\left(|\hat{\beta}_j - \beta_j^*| \ge \nu_n(u_n-1)\right). \end{aligned} \tag{17}$$

Notice that from Lemma 2.1, $P(\mathcal{C}_{[j]} \subseteq \hat{\mathcal{C}}_{[j]}) \ge 1 - C_1 n^{\xi}\exp(-C_2 n^{\xi})$ for some positive constants C_1 and C_2 and $0 < \xi < 1$. Therefore, we further have

$$\begin{aligned} & P\left(|\hat{\beta}_j - \beta_j^*| \ge \nu_n(u_n-1)\right) \\ =\ & P\left(\left|\sum_{j'\in\hat{\mathcal{C}}_{[j]}} \hat{\Omega}_{jj'}\widehat{\mathrm{cov}}(X_{j'}, y) - \sum_{j'\in\mathcal{E}_j}\Omega_{jj'}\mathrm{cov}(X_{j'}, y)\right| \ge \nu_n(u_n-1)\right) \\ =\ & P\left(\left|\sum_{j'\in\hat{\mathcal{C}}_{[j]}} \hat{\Omega}_{jj'}\widehat{\mathrm{cov}}(X_{j'}, y) - \sum_{j'\in\hat{\mathcal{C}}_{[j]}}\Omega_{jj'}\mathrm{cov}(X_{j'}, y)\right| \ge \nu_n(u_n-1)\right) + C_1 n^{\xi}\exp(-C_2 n^{\xi}) \\ \le\ & P\left(\sum_{j'\in\hat{\mathcal{C}}_{[j]}} \left|(\hat{\Omega}_{jj'} - \Omega_{jj'})\widehat{\mathrm{cov}}(X_{j'}, y) + \Omega_{jj'}(\widehat{\mathrm{cov}}(X_{j'}, y) - \mathrm{cov}(X_{j'}, y))\right| \ge \nu_n(u_n-1)\right) + \\ & C_1 n^{\xi}\exp(-C_2 n^{\xi}) \\ \le\ & P\left(\sum_{j'\in\hat{\mathcal{C}}_{[j]}} \left|(\hat{\Omega}_{jj'} - \Omega_{jj'})\widehat{\mathrm{cov}}(X_{j'}, y)\right| \ge \nu_n(u_n-1)/2\right) + \\ & P\left(\sum_{j'\in\hat{\mathcal{C}}_{[j]}} |\Omega_{jj'}(\widehat{\mathrm{cov}}(X_{j'}, y) - \mathrm{cov}(X_{j'}, y))| \ge \nu_n(u_n-1)/2\right) + C_1 n^{\xi}\exp(-C_2 n^{\xi}). \end{aligned} \tag{18}$$

By (15) and Assumption (A6), $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\| \le M n^{-(1-\xi/4)}$, for some $0 < M < \infty$. As $n \to \infty$, the

first term in (18) can be shown as:

$$
\begin{aligned}
& P\left(\sum_{j'\in\hat{\mathcal{C}}_{[j]}} |(\hat{\Omega}_{jj'}-\Omega_{jj'})\widehat{\mathrm{cov}}(X_{j'},y)| > \nu_n(u_n-1)/2\right) \\
\leq\ & P\left(\frac{1}{n^2}\mathbf{y}^{\mathrm{T}}\mathbf{x}_{\hat{\mathcal{C}}_{[j]}}(\hat{\mathbf{\Omega}}-\mathbf{\Omega})^{\mathrm{T}}_{j\hat{\mathcal{C}}_{[j]}}(\hat{\mathbf{\Omega}}-\mathbf{\Omega})_{j\hat{\mathcal{C}}_{[j]}}\mathbf{x}^{\mathrm{T}}_{\hat{\mathcal{C}}_{[j]}}\mathbf{y} > \frac{\nu_n^2(u_n-1)^2}{4}\right) \\
\leq\ & P\left(\lambda_{\max}\left(\frac{1}{n}\mathbf{x}_{\hat{\mathcal{C}}_{[j]}}(\hat{\mathbf{\Omega}}-\mathbf{\Omega})^{\mathrm{T}}_{j\hat{\mathcal{C}}_{[j]}}(\hat{\mathbf{\Omega}}-\mathbf{\Omega})_{j\hat{\mathcal{C}}_{[j]}}\mathbf{x}^{\mathrm{T}}_{\hat{\mathcal{C}}_{[j]}}\right)\frac{1}{n}\mathbf{y}^{\mathrm{T}}\mathbf{y} > \frac{\nu_n^2(u_n-1)^2}{4}\right). \\
\leq\ & P\left(\lambda_{\max}\left(\frac{1}{n}\mathbf{x}_{\hat{\mathcal{C}}_{[j]}}\mathbf{x}^{\mathrm{T}}_{\hat{\mathcal{C}}_{[j]}}\right)\lambda_{\max}\left((\hat{\mathbf{\Omega}}-\mathbf{\Omega})^{\mathrm{T}}_{j\hat{\mathcal{C}}_{[j]}}(\hat{\mathbf{\Omega}}-\mathbf{\Omega})_{j\hat{\mathcal{C}}_{[j]}}\right)\frac{1}{n}\mathbf{y}^{\mathrm{T}}\mathbf{y} > \frac{\nu_n^2(u_n-1)^2}{4}\right). \qquad (19)
\end{aligned}
$$

Notice that $E[\mathbf{y}^{\mathrm{T}}\mathbf{y}/n] = E(y^2) = \sigma^2+(\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{\Sigma}\boldsymbol{\beta}^* \leq \sigma^2+\lambda_{\max}(\mathbf{\Sigma})\|\boldsymbol{\beta}^*\|^2 < \tilde{C}_1$ for some positive constant $\tilde{C}_1$. For sufficiently large n, $\lambda_{\max}\left(\mathbf{x}_{\hat{\mathcal{C}}_{[j]}}\mathbf{x}^{\mathrm{T}}_{\hat{\mathcal{C}}_{[j]}}/n\right) \leq \lambda_{\max}\left(\hat{\mathbf{\Sigma}}\right) \leq \kappa_2+\lambda_{\max}\left(\hat{\mathbf{\Sigma}}-\mathbf{\Sigma}\right) = \kappa_2+\|\hat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|^{1/2} = \kappa_2+O(\rho_n^{1/2}) < \tilde{C}_2$ for some positive constant $\tilde{C}_2$. And $\lambda_{\max}\left((\hat{\mathbf{\Omega}}-\mathbf{\Omega})^{\mathrm{T}}_{j\hat{\mathcal{C}}_{[j]}}(\hat{\mathbf{\Omega}}-\mathbf{\Omega})_{j\hat{\mathcal{C}}_{[j]}}\right) \leq \|\hat{\mathbf{\Omega}}-\mathbf{\Omega}\|^2 \leq M^2 n^{-(2-\xi/2)}$. Therefore, from (19), for sufficiently large n,

$$
\begin{aligned}
& P\left(\sum_{j'\in\hat{\mathcal{C}}_{[j]}} |(\hat{\Omega}_{jj'}-\Omega_{jj'})\widehat{\mathrm{cov}}(X_{j'},y)| > \nu_n(u_n-1)/2\right) \\
& \leq P\left(\tilde{C}_2M^2\frac{1}{n}\mathbf{y}^{\mathrm{T}}\mathbf{y} > n^{(2-\xi/2)}\frac{\nu_n^2(u_n-1)^2}{4}\right) \\
& \leq P\left(\frac{1}{n}\mathbf{y}^{\mathrm{T}}\mathbf{y} > \frac{1}{\tilde{C}_2M^2}\frac{n^2}{4}\right) \leq \frac{4\tilde{C}_2M^2E[\mathbf{y}^{\mathrm{T}}\mathbf{y}/n]}{n^2} \leq \frac{4\tilde{C}_1\tilde{C}_2M^2}{n^2} \to 0, \qquad (20)
\end{aligned}
$$

where the second last step is from applying Markov inequality to the positively valued random variable $\mathbf{y}^{\mathrm{T}}\mathbf{y}/n$.

For the second term in (18), let $\tilde{\mathbf{z}} = (\tilde{Z}_1,\ldots,\tilde{Z}_p)^{\mathrm{T}} = (\mathbf{x}_1^{\mathrm{T}}\mathbf{y}/n-\mathrm{cov}(X_1,y),\ldots,\mathbf{x}_p^{\mathrm{T}}\mathbf{y}/n-\mathrm{cov}(X_p,y))^{\mathrm{T}}$, then we have

$$
\begin{aligned}
& P\left(\sum_{j'\in\hat{\mathcal{C}}_{[j]}} \left|\Omega_{jj'}(\mathbf{x}_{j'}^{\mathrm{T}}\mathbf{y}/n-\mathrm{cov}(X_{j'},y))\right| \geq \nu_n(u_n-1)/2\right) \\
\leq\ & P\left(\lambda_{\max}\left(\mathbf{\Omega}^{\mathrm{T}}\mathbf{\Omega}\right)\|\tilde{\mathbf{z}}_{\hat{\mathcal{C}}_{[j]}}\|_2^2 \geq \frac{\nu_n^2(u_n-1)^2}{4}\right) \\
\leq\ & P\left(|\tilde{Z}_{j'}| > \frac{\nu_n\kappa_1(u_n-1)}{2|\hat{\mathcal{C}}_{[j]}|}\right)
\end{aligned}
$$

for some $j' \in \hat{\mathcal{C}}_{[j]}$. Notice that $E[\tilde{Z}_{j'}] = E[X_{j'}y - \text{cov}(X_{j'}, y)] = 0$. Also $E[\tilde{Z}_{j'}^2] = Var[X_{j'}y] \leq E[X_{j'}^2 y^2] \leq (E(X_{j'}^4) + E(y^4))/2 < \tilde{C}_3$ for some positive constant $\tilde{C}_3$ as $E(X_{j'}^4) \leq \lambda_{\max}^2(\mathbf{\Sigma})$ and $E(y^4) \leq (\lambda_{\max}(\mathbf{\Sigma})\|\boldsymbol{\beta}\|^2)^2 + E(\varepsilon^4) < \infty$. Therefore, from Jensen's inequality, $E[|\tilde{Z}_{j'}|] = E[(\tilde{Z}_{j'}^2)^{1/2}] \leq (E[\tilde{Z}_{j'}^2])^{1/2} \leq \tilde{C}_3^{1/2}$. Then using Lemma 7.3 to control the size of $\hat{\mathcal{C}}_{[j]}$ and applying Markov inequality on $|\tilde{Z}_{j'}|$, we have

$$P\left(|\tilde{Z}_{j'}| > \frac{\nu_n \kappa_1 (u_n - 1)}{2|\hat{\mathcal{C}}_{[j]}|}\right) \leq P\left(|\tilde{Z}_{j'}| > \frac{\kappa_1 n^{\xi/4}}{2}\right) \leq \frac{2E[|\tilde{Z}_{j'}|]}{\kappa_1 n^{\xi/4}}$$
$$\leq \frac{2\tilde{C}_3^{1/2}}{\kappa_1 n^{\xi/4}} \to 0. \tag{21}$$

Plugging (20) and (21) into (18) and then plugging (18) into (17) gives

$$\lim_{n\to\infty} P\left(\{j \notin \mathcal{S}_1 : |\beta_j^*| > \nu_n u_n\} \subseteq \{j \notin \mathcal{S}_1 : |\hat{\beta}_j| > \nu_n\}\right) = 1. \tag{22}$$

By a similar argument, we also have

$$\lim_{n\to\infty} P\left(\{j \notin \mathcal{S}_1 : |\beta_j^*| > \nu_n / u_n\} \subseteq \{j \notin \mathcal{S}_1 : |\hat{\beta}_j| > \nu_n\}\right) = 1. \tag{23}$$

Combining (22) and (23), we have

$$\lim_{n\to\infty} P(\hat{\mathcal{S}}_{\text{WBC}} = \mathcal{S}_{\text{WBC}} | \hat{\mathcal{S}}_1 = \mathcal{S}_1).$$

□

Proof of Corollary 3.2. Notice that

$$P\left(\{\hat{\mathcal{S}}_{\text{WBC}} = \mathcal{S}_{\text{WBC}}\} \cap \{\hat{\mathcal{S}}_{2*} = \mathcal{S}_{2*}\} \cap \{\hat{\mathcal{S}}_1 = \mathcal{S}_1\}\right)$$
$$= P\left(\hat{\mathcal{S}}_{2*} = \mathcal{S}_{2*} | \{\hat{\mathcal{S}}_{\text{WBC}} = \mathcal{S}_{\text{WBC}}\} \cap \{\hat{\mathcal{S}}_1 = \mathcal{S}_1\}\right) P\left(\{\hat{\mathcal{S}}_{\text{WBC}} = \mathcal{S}_{\text{WBC}}\} \cap \{\hat{\mathcal{S}}_1 = \mathcal{S}_1\}\right)$$
$$= P\left(\hat{\mathcal{S}}_{2*} = \mathcal{S}_{2*} | \hat{\mathcal{S}}_{1\text{WBC}} = \mathcal{S}_{1\text{WBC}}\right) P\left(\hat{\mathcal{S}}_{\text{WBC}} = \mathcal{S}_{\text{WBC}} | \hat{\mathcal{S}}_1 = \mathcal{S}_1\right) P\left(\hat{\mathcal{S}}_1 = \mathcal{S}_1\right). \tag{24}$$

Under the SRC in (A4), by Lemma 1 in Gao et al. (2017) or Theorem 2 in Zhang and Huang (2008),

$$\lim_{n\to\infty} P\left(\hat{\mathcal{S}}_1 = \mathcal{S}_1\right) = 1. \tag{25}$$

From Theorem 3.1,

$$\lim_{n\to\infty} P\left(\hat{\mathcal{S}}_{\text{WBC}} = \mathcal{S}_{\text{WBC}} | \hat{\mathcal{S}}_1 = \mathcal{S}_1\right) = 1. \tag{26}$$

Equations (25) and (26) together give $\lim_{n\to\infty} P(\hat{\mathcal{S}}_{\text{WBC}} = \mathcal{S}_{\text{WBC}}) = 1$. This further gives that $P(\mathcal{S}_1 \subset \hat{\mathcal{S}}_{1\text{WBC}} \subset \{\mathcal{S}_{1\text{WBC}} \cup \mathcal{S}_2\}) \to 1$. Then by Corollary 2 in Gao et al. (2017), we also have

$$\lim_{n\to\infty} P(\{\hat{\mathcal{S}}_{2*} = \mathcal{S}_{2*}\} | \{\hat{\mathcal{S}}_{1\text{WBC}} = \mathcal{S}_{1\text{WBC}}\}) = 1. \tag{27}$$

Combining (25), (26), (27) and (24) completes the proof. □

The following Tables 5 - 6 and Figures 5 - 6 give the selection, estimation and prediction results under Examples 1 and 2 when $\hat{\mathcal{S}}_1$ is selected by adaptive Lasso.

Table 5: The performance of variable selection on $\mathcal{S}_0$ when $\hat{\mathcal{S}}_1$ is selected by adaptive Lasso

			$p = 200$	$p = 300$	$p = 400$	$p = 500$
Example 1	TP	CIS-PSE	61.4 (2.4)	61.1 (2.5)	61.0 (2.6)	61.1 (2.6)
		PSE	41.2 (5.0)	34.4 (5.1)	25.7 (6.1)	22.6 (5.8)
	FP	CIS-PSE	2.5 (2.1)	4.6 (2.5)	6.7 (3.2)	8.6 (3.0)
		PSE	15.1 (4.9)	19.7 (5.0)	22.3 (5.4)	28.0 (6.2)
Example 2	TP	CIS-PSE	62.5 (0.8)	58.0 (2.2)	54.6 (2.9)	52.0 (3.5)
		PSE	43.9 (3.8)	37 (4.2)	32.8 (4.3)	31.5 (4.2)
	FP	CIS-PSE	3.4 (2.6)	5.2 (3.0)	6.0 (3.5)	7.7 (4.1)
		PSE	13.1 (3.7)	18.6 (4.7)	21.9 (5.5)	28.0 (6.1)

Table 6: Mean squared prediction error (MSPE) of the predicted outcomes when $\hat{\mathcal{S}}_1$ is selected by adaptive Lasso

	p	200	300	400	500
Example 1	CIS-PSE	2.16 (0.57)	3.06 (0.63)	3.30 (0.72)	3.60 (0.81)
	PSE	3.91 (0.71)	4.85 (0.86)	5.71 (1.02)	6.27 (1.17)
	adaptive Lasso	15.02 (4.37)	14.67 (3.90)	14.49 (4.00)	14.74 (4.29)
Example 2	CIS-PSE	2.69 (0.58)	2.96 (0.72)	3.23 (0.81)	3.31 (0.83)
	PSE	3.77 (0.77)	4.76 (1.08)	5.50 (1.17)	5.82 (0.18)
	adaptive Lasso	4.48 (0.79)	4.81 (0.91)	4.94 (0.97)	4.97 (1.01)

Figure 7 shows the averaged sum of squared prediction error (SSPE) on the validation datasets across 500 independent experiments for different α's.

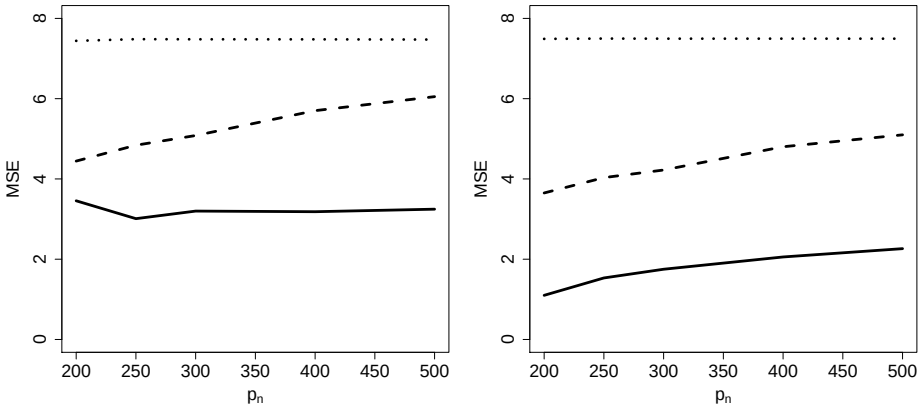

Figure 5: The mean squared error (MSE) of $\hat{\boldsymbol{\beta}}_{\mathcal{S}_{\text{WBC}}}$ for different p's when $\hat{\mathcal{S}}_1$ is selected by adaptive Lasso under Example 1 (Left panel) and Example 2 (Right panel). Solid lines represent CIS-PSE, dashed lines are for PSE, and dotted lines indicate adaptive Lasso.

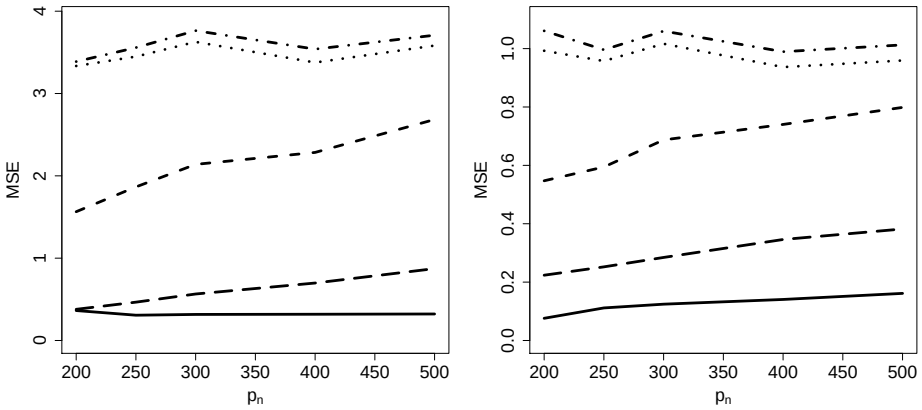

Figure 6: The mean squared error (MSE) of $\hat{\boldsymbol{\beta}}_{\mathcal{S}_1}$ for different p's when $\hat{\mathcal{S}}_1$ is selected by adaptive Lasso under Example 1 (Left panel) and Example 2 (Right panel). Solid lines represent CIS-PSE, dashed lines are for PSE, dotted lines indicate adaptive Lasso RE defined as $\hat{\boldsymbol{\beta}}^{\text{RE}}_{\hat{\mathcal{S}}_1} = \hat{\boldsymbol{\Sigma}}^{-1}_{\hat{\mathcal{S}}_1} \widehat{\text{cov}}(\mathbf{x}_{\hat{\mathcal{S}}_1}, y)$, dot-dashed lines represent adaptive Lasso, and long-dashed lines are for WR in (9).

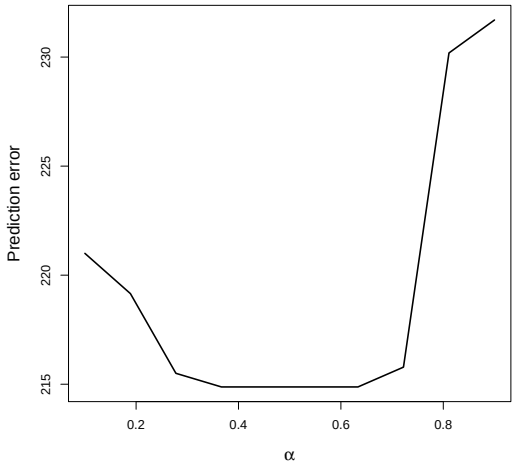

Figure 7: The sum of squared prediction error (SSPE) corresponding to different α's.

References

Barro, R. and Lee, J. (1994). *Data set for a panel of 139 countries, NBER.* available at http://admin.nber.org/pub/barro.lee/.

Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Stat.*, 36(6):2577–2604.

Bodmer, W. and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.*, 40(6):695–701.

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.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.*, 96:1348–1360.

Fan, J., Liao, Y., and Min, M. (2011). High-dimensional covariance matrix estimation in approxiamte factor models. *Ann. Stat.*, 39(6):3320–3356.

Gao, X., Ahmed, S. E., and Feng, Y. (2017). Post selection shrinkage estimation for high-dimensional data analysis. *Applied Stochastic Models in Business and Industry*, 33(2):97–120.

Li, B. and Leal, S. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, 83(3):311–321.

Mazumder, R. and Hastie, T. (2012). Exact covariance thresholding into connected components for large-scale graphical lasso. *Journal of Machine Learning Research*, 12:723–736.

Shah, R. D. and Samworth, R. J. (2013). Discussion of 'correlated variables in regression: clustering and sparse estimation'. *Journal of Statistical Planning and Inference*, 143:1866–1868.

Shao, J., Wang, Y., Deng, X., and Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high-dimensional data. *Ann. Stat.*, 39(2):1241–1265.

Shapiro, L. and Stockman, G. (2002). *Computer Vision*. Prentice Hall.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc. B.*, 58:267–288.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet.*, 89(1):82–93.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B.*, 68:49–67.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.*, 38(2):1567–1594.

Zhang, C. H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high dimensional linear regression. *Ann. Stat.*, 36:1567–1594.

Zhang, C. H. and Zhang, S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc. B*, 76(1):217–242.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Am. Statist. Assoc.*, 101:1418–1429.