Supplementary Materials

for

Covariance-insured Screening

Proof of Lemma 1

Consider the semi-partial correlation

$$\rho(Y_i, X_{i,j}|\mathbf{X}_{i,-j}) = \frac{\text{Cov}\left[Y_i, X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,-j})\right]}{\left\{\text{Var}(Y_i)\text{Var}(X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,-j}))\right\}^{1/2}} \tag{.1}$$

$$= \frac{\text{Cov}\left[Y_i - E(Y_i|\mathbf{X}_{i,-j}), X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,-j})\right]}{\left\{\text{Var}(Y_i)\text{Var}(X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,-j}))\right\}^{1/2}}, \tag{.2}$$

where the last equality is due to the fact

$$\begin{aligned}
&\text{Cov}\left[E(Y_i|\mathbf{X}_{i,-j})), X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,-j})\right] \\
=&E\left[(\beta_j E(X_{i,j}|\mathbf{X}_{i,-j}) + \mathbf{X}_{i,-j}\boldsymbol{\beta}_{-j})(X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,-j}))\right] \\
=&E\left[\beta_j E(X_{i,j}|\mathbf{X}_{i,-j})(X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,-j}))\right] = 0.
\end{aligned}$$

The numerator of (.2) is equal to

$$\text{Cov}\left[\beta_j(X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,-j})) + \epsilon_i, X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,-j})\right] = \beta_j \text{Var}(X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,-j})).$$

As $\boldsymbol{\Sigma}$ is positive definite,

$$\text{Var}(X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,-j})) = E[(X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,-j}))^2] \neq 0.$$

Therefore,

$$\beta_j = 0 \text{ if and only if } \rho(Y_i, X_{i,j}|\mathbf{X}_{i,-j}) = 0.$$

Proof of Lemma 2

The semi-partial correlation, as defined in (.1), is equal to

$$\begin{aligned}
\rho(Y_i, X_{i,j}|\mathbf{X}_{i,-j}) &= \frac{\mathrm{Cov}\left[Y_i - E(Y_i|\mathbf{X}_{i,-j}), X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,-j})\right]}{\left\{\mathrm{Var}(Y_i)\mathrm{Var}(X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,-j}))\right\}^{1/2}} \\
&= \frac{\mathrm{Cov}\left[\beta_j(X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,\mathcal{S}_g\setminus\{j\}}, \mathbf{X}_{i,\mathcal{S}_g^c})) + \epsilon_i, X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,\mathcal{S}_g\setminus\{j\}}, \mathbf{X}_{i,\mathcal{S}_g^c})\right]}{\left\{\mathrm{Var}(Y_i)\mathrm{Var}(X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,\mathcal{S}_g\setminus\{j\}}, \mathbf{X}_{i,\mathcal{S}_g^c}))\right\}^{1/2}} \\
&= \beta_j\{\mathrm{Var}(X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,\mathcal{S}_g\setminus\{j\}}))/\mathrm{Var}(Y_i)\}^{1/2}. \qquad (.3)
\end{aligned}$$

Similarly, for any $j = 1, \ldots, p$ and some g such that $j \in \mathcal{S}_g$, the block-wise semi-partial correlation

$$\begin{aligned}
\rho(Y_i, X_{i,j}|\mathbf{X}_{i,\mathcal{S}_g\setminus\{j\}}) &= \frac{\mathrm{Cov}\left[Y_i - E(Y_i|\mathbf{X}_{i,\mathcal{S}_g\setminus\{j\}}), X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,\mathcal{S}_g\setminus\{j\}})\right]}{\left\{\mathrm{Var}(Y_i)\mathrm{Var}(X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,\mathcal{S}_g\setminus\{j\}}))\right\}^{1/2}} \\
&= \frac{\mathrm{Cov}\left[\beta_j(X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,\mathcal{S}_g\setminus\{j\}})) + \mathbf{X}_{i,\mathcal{S}_g^c}\boldsymbol{\beta}_{\mathcal{S}_g^c} - E(\mathbf{X}_{i,\mathcal{S}_g^c}\boldsymbol{\beta}_{\mathcal{S}_g^c}) + \epsilon_i, X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,\mathcal{S}_g\setminus\{j\}})\right]}{\left\{\mathrm{Var}(Y_i)\mathrm{Var}(X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,\mathcal{S}_g\setminus\{j\}}))\right\}^{1/2}} \\
&= \beta_j\{\mathrm{Var}(X_{i,j} - E(X_{i,j}|\mathbf{X}_{i,\mathcal{S}_g\setminus\{j\}}))/\mathrm{Var}(Y_i)\}^{1/2}, \qquad (.4)
\end{aligned}$$

where the last two equalities follow by the assumption of independent blocks. Therefore, (.3) and (.4) together imply

$$\rho(Y_i, X_{i,j}|\mathbf{X}_{i,-j}) = \rho(Y_i, X_{i,j}|\mathbf{X}_{i,\mathcal{S}_g\setminus\{j\}}).$$

We remark that the purpose of Lemma 2 is to provide the intuition behind the proposed method. The assumption of independent blocks is not required for the proposed method. The proposed method is valid for more general settings. Indeed, as long as the correlation between blocks is small, the semi-partial correlation based on the relevant blocks is able to adequately assess the true contributions of each covariate to the response; see the proofs for Theorem 1.

The illustrating example referred in Section 4.2

Consider a simple example,

$$Y_i = -1.5X_{i,1} + X_{i,2} - 0.5X_{i,3} + X_{i,4} + \epsilon_i,$$

where $\epsilon_i \sim N(0,1)$ is independent of $(X_{i,1}, X_{i,2}, X_{i,3}, X_{i,4})$ that is normally distributed with mean zero and a covariance matrix

$$\begin{bmatrix} 1 & 0.5 & 0 & 0 \\ 0.5 & 1 & 0.5 & 0 \\ 0 & 0.5 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

In this case, both the marginal correlation between Y_i and $X_{i,3}$ and the partial correlation between Y_i and $X_{i,3}$ conditional on $X_{i,2}$ (the only variable that is correlated with $X_{i,3}$) are zero and hence $X_{i,3}$ will likely be missed by the Tilting procedure.

An FDR-driven threshold for the selection frequency

To determine data-driven thresholds for the selection frequency ψ and directly control the false discovery rate, we further propose a random permutation based approach.

We randomly permute the outcomes S times to decouple the relation between the covariates and the outcomes. On each permuted dataset, say $s = 1, \ldots, S$, we perform CIS-Stable and compute the empirical probability that each variable is selected, denoted by $\widetilde{\Psi}_j^{(s)}$ for $j = 1, \ldots, p$. Order these empirical probabilities such that $\widetilde{\Psi}_{(1)}^{(s)} > \widetilde{\Psi}_{(2)}^{(s)} > \ldots > \widetilde{\Psi}_{(p)}^{(s)}$. Define $\overline{\Psi}_{(j)} = \frac{1}{S}\sum_{s=1}^{S} \widetilde{\Psi}_{(j)}^{(s)}$. Likewise, let $\widehat{\Psi}_{(j)}$ be the j-th largest value of the empirical probabilities based on the original data. For a positive constant $\Delta > 0$, define $C(\Delta) = \min\{\widehat{\Psi}_{(j)} : \widehat{\Psi}_{(j)} - \overline{\Psi}_{(j)} \geq \Delta\}$, $R(\Delta) = \sum_{j=1}^{p} I(\widehat{\Psi}_j \geq C(\Delta))$, and $\tilde{R}(\Delta) = \sum_{j=1}^{p}\sum_{s=1}^{S} I(\widetilde{\Psi}_j^{(s)} \geq C(\Delta))$. The empirical Bayes false discovery rate corresponding to the Δ can be estimated by

$$\overline{FDR}(\Delta) = \frac{1}{S}\frac{\tilde{R}(\Delta)}{R(\Delta)}.$$

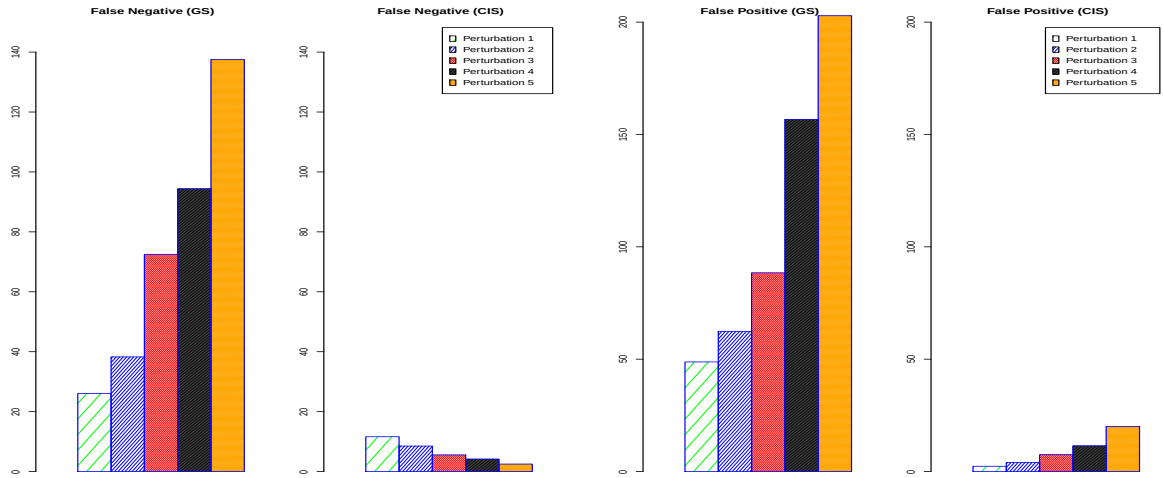
For a pre-specified value $q \in (0, 1)$, finding a Δ that produces $\overline{FDR}(\Delta) \leq q$ will ensure that at most q proportion of the selected variables would be false positives. Finally, the threshold for the selection frequency equals $C(\Delta)$.

More Simulation Results

Comparisons of Iterative CIS and GS

We compare the performance of iterative CIS (with $\delta = 0.3$) and iterative GS with various choices of tuning parameters, using simulation Model E described on page 15 of the main manuscript. Specifically, we consider five perturbation settings for each of the method. Let ϵ_p^* and τ^* be the optimal tuning parameters for GS (defined as in Jin et al., 2014). We define ρ as a binary random variable with equal probability of being 1 or -1. Perturbation 1 for GS: $\epsilon_p = \epsilon_p^*$, $\tau = \tau^*$; Perturbation 2 for GS: $\epsilon_p = (1+0.1\rho)\epsilon_p^*$, $\tau = (1+0.1\rho)\tau^*$; Perturbation 3 for GS: $\epsilon_p = (1+0.2\rho)\epsilon_p^*$, $\tau = (1+0.2\rho)\tau^*$; Perturbation 4 for GS: $\epsilon_p = (1+0.3\rho)\epsilon_p^*$, $\tau = (1+0.3\rho)\tau^*$; Perturbation 5 for GS: $\epsilon_p = (1+0.4\rho)\epsilon_p^*$, $\tau = (1+0.4\rho)\tau^*$. Perturbation 1 for CIS: ν is chosen such that the top $0.1n/\log(n)$ variables are selected; Perturbation 2 for CIS: ν is chosen such that the top $0.15n/\log(n)$ variables are selected; Perturbation 3 for CIS: ν is chosen such that the top $0.2n/\log(n)$ variables are selected; Perturbation 4 for CIS: ν is chosen such that the top $0.25n/\log(n)$ variables are selected; Perturbation 5 for CIS: ν is chosen such that the top $0.3n/\log(n)$ variables are selected. The results suggest that the perturbation of tuning parameters ν has relatively small effects on the proposed iterative CIS, which always outperforms GS with the optimal tuning parameters. In contrast, the number of false negatives and false positives increases for iterative GS as their tuning parameters depart from the optimal values.

Moreover, we assess the proposed CIS with respect to various choices of δ_n (e.g. $\delta_n = \min(1, K\sqrt{\log(p)/n})$, $K = 1, 2, \ldots$). We used Models A, B and C as described on pages



(a) Number of False Negative (b) Number of False Positive

Figure A1: Comparison of iterative CIS and GS with various tuning parameters.

14-15 of the revision. In addition, we varied n and p to assess the performance of CIS under different scenarios. Table A1 compares the minimum model size (number of selected variables to include the true model). When $\delta_n > \rho$, where ρ is the within-block auto-correlation parameter, each identified block tends to include only one predictor. In the extreme case, CIS reduces to SIS, which does not work well for Model A with high correlations (because the marginal correlation condition is not satisfied). On the other hand, if K is too small, the corresponding $q_n > n$ and the proposed method is not applicable. Based on the results with different n, p and ρ, it seems that $K = 5$ or 6 works the best in most cases. This also justifies the use of $K = 5$ in the original version.

Table A2 assesses the proposed ICIS with respect to various choices of δ_n. The trend is similar to that in Table A1. When $K = 2$ to 7 the results are almost the same.

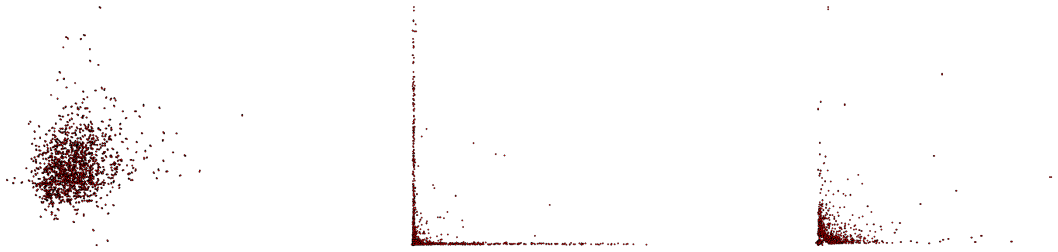
Finally, we vary the number of permutations (S) or the number of bootstraps (B) to assess the performance of the proposed method. As illustrated in Figure A3, the performance is relatively robust to the number of permutations, while 50-100 bootstraps are sufficient for

Table A1: The minimum model size to include the true model for CIS with various δ_n (e.g. $\delta_n = \min(1, K\sqrt{\log(p)/n})$, $K = 1, 2, \ldots$); Model A, B and C; assess various combinations of sample size (n) and number of predictors (p); NA: no results are reported because the corresponding $q_n > n$; q_n: maximal number of variables in the blocks.

Model	(n, p)	ρ	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$	$K=9$	$K=10$
A	(1,000, 10,000)	0.9	NA	72.9	73.7	73.7	73.7	73.7	73.7	73.7	73.7	6861.22
		0.7	NA	10.0	10.0	10.0	10.0	10.3	35.5	782.8	782.8	782.8
		0.5	NA	10.0	10.0	10.0	10.0	12.7	13.6	19.4	19.4	19.4
		0.3	NA	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
		0.1	NA	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
A	(500, 1,000)	0.9	NA	161.5	161.5	161.5	161.5	161.5	161.5	988.75	988.75	988.75
		0.7	NA	18.2	18.0	18.0	18.0	18.0	286.45	379.56	379.56	379.56
		0.5	NA	10.8	10.8	14.8	50.1	50.2	50.2	50.2	50.2	50.2
		0.3	NA	10.2	10.4	11.9	11.9	11.9	11.9	11.9	11.9	11.9
		0.1	NA	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
B	(1,000, 10,000)	0.9	NA	68.1	68.6	68.6	68.6	68.6	68.6	68.6	68.6	65.1
		0.7	NA	10.0	10.0	10.0	10.0	10.0	10.0	10.7	10.7	10.7
		0.5	NA	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
		0.3	NA	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
		0.1	NA	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
B	(500, 1,000)	0.9	NA	116.9	116.9	116.9	116.9	116.9	116.9	116.9	31.0	31.0
		0.7	NA	10.5	10.5	10.5	10.5	10.5	15.9	13.1	13.1	13.1
		0.5	NA	10.0	10.0	10.1	10.5	10.4	10.4	10.4	10.4	10.4
		0.3	NA	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
		0.1	NA	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
C	(1,000, 5,000)	0.9	NA	NA	NA	NA	NA	NA	NA	NA	NA	33.3
		0.7	NA	NA	NA	NA	NA	NA	NA	NA	11.3	11.3
		0.5	NA	NA	NA	NA	NA	10.1	10.1	10.1	10.1	10.1
		0.3	NA	NA	NA	10.0	10.0	10.0	10.0	10.0	10.0	10.0
		0.1	NA	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
C	(500, 1,000)	0.9	NA	NA	NA	NA	NA	NA	NA	48.0	48.0	48.0
		0.7	NA	NA	NA	NA	NA	NA	NA	16.0	16.0	16.0
		0.5	NA	NA	NA	NA	NA	10.8	10.8	10.8	10.8	10.8
		0.3	NA	NA	NA	10.0	10.0	10.0	10.0	10.0	10.0	10.0
		0.1	NA	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0

Table A2: Number of false positives (FP) and number of false negatives (FN) for ICIS with various δ_n (e.g. $\delta_n = \min(1, K\sqrt{\log(p)/n})$, $K = 1, 2, \ldots$); Model D (with $\rho = 0.9$ and $p = 1,000$) and E (with $p = 5,000$); NA: no results are reported because the corresponding $q_n > n$; q_n: maximal number of variables in the blocks.

Model	Measure	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$	$K=9$	$K=10$
D	FP	NA	0.72	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73
	FN	NA	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
E	FP	NA	20.13	20.13	20.13	20.13	20.13	20.13	30.91	38.26	38.26
	FN	NA	2.55	2.55	2.55	2.55	2.55	2.55	4.94	6.68	6.68

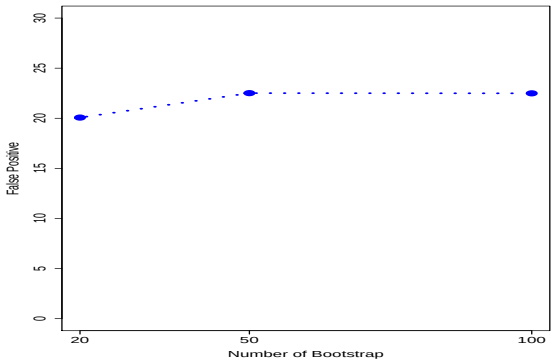


(a) $\delta = 0.3$ (b) $\delta = 0.4$ (c) $\delta = 0.5$

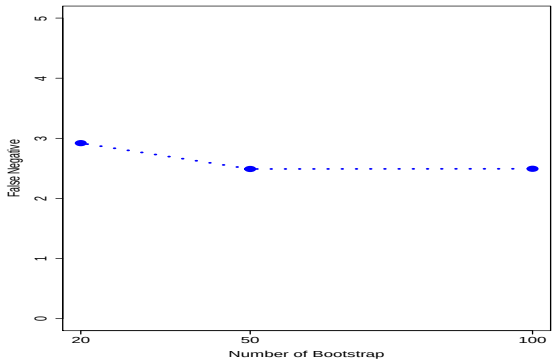
Figure A2: Block structures; (a): $\delta = 0.3$, number of blocks=1,250; maximum number of predictors per block=4; (b): $\delta = 0.4$, number of blocks=1,250; maximum number of predictors per block=4; (c): $\delta = 0.5$, number of blocks=5,000; maximum number of predictors per block=1.

reliable estimations. For practical implementation, we recommend S=10 and B=50.

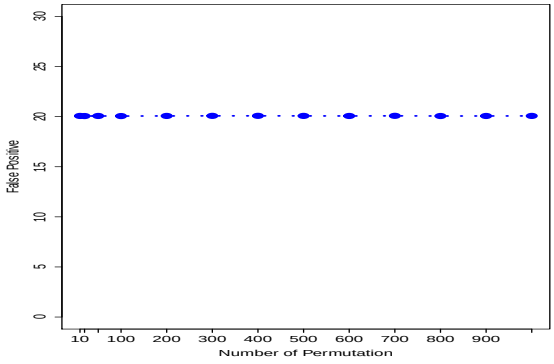
References James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) *An Introduction to Statistical Learning with Applications in R*. Springer-Verlag, New York.



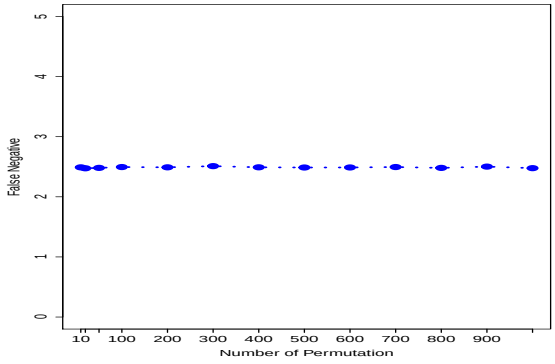
(a) Number of False Positive for various B



(b) Number of False Negative for various B



(c) Number of False Positive for various S



(d) Number of False Negative for various S

Figure A3: Performance with various numbers of bootstrapping (B) and permutation (S).