Linear Shrinkage Convexification of

Penalized Linear Regression With Missing Data

Seongoh Park[1], Seong Jin Lee[2], Nguyen Thi Hai Yen[3],

Nguyen Phuoc Long[4], Johan Lim[5,*]

[1]*Sungshin Women's University,* [2]*University of North Carolina at Chapel Hill,*

[3]*Inje University College of Medicine,* [4]*Chang Gung University,* [5]*Seoul National University*

Supplementary Material

This supplementary material presents additional simulation results and technical theorems to prove the main results.

S1 Non-asymptotic inequality of the IPW estimator in the spectral norm

In this section, we will derive the concentration inequality of the IPW estimator. More specifically, we are interested in the rate of convergence of $||\widehat{\mathbf{\Sigma}}^{\text{IPW}} - \mathbf{\Sigma}||_2$. Recall the definition of the IPW estimator

$$\widehat{\mathbf{\Sigma}}^{\text{IPW}} = \boldsymbol{S} * \left[\frac{1}{\pi^{xx}_{jk}}, 1 \leq j, k \leq p \right],$$

which is given in (2.5). The random variables $\boldsymbol{x}_i$, $(\delta^x_{i1}, \ldots, \delta^x_{ip})$ used above are assumed to satisfy Assumption 1, 2, and 3. For notational convenience, we write the IPW estimator by $\widehat{\mathbf{\Sigma}}$. Also, we omit the superscript in δ^x_{ij}, $\pi^{xx}_{ij\cdots}$ and K^x.

*To whom all correspondence should be addressed.

Theorem S1. *For $t > 1 \vee \log n$, it holds with probability at least $1 - 3e^{-t}$ that*

$$||\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_2 \leq C\mathrm{tr}(\boldsymbol{\Sigma}) \max\{K^2, 1\} \max\left\{ \sqrt{\frac{\pi^{(4)}_{\max}(t + \log p)}{n}}, (t + \log n)\frac{\pi^{(4)}_{\max}(t + \log p)}{n} \right\},$$

where $C > 0$ is some numerical constant and

$$\pi^{(4)}_{\max} = \max_{k_1,k_2,\ell_1,\ell_2} \frac{\pi_{k_1k_2\ell_1\ell_2}}{\pi_{k_1\ell_1}\pi_{k_2\ell_2}}.$$

Our proof is based on the idea of Lounici (2014), but improve it to address the general missing dependency.

We begin with the following decomposition:

$$||\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_2 \leq ||\mathrm{diag}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})||_2 + ||\mathrm{OD}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})||_2$$

where $\mathrm{diag}(\boldsymbol{A})$ is a diagonal matrix with diagonals inherited from $\boldsymbol{A}$, and $\mathrm{OD}(\boldsymbol{A}) = \boldsymbol{A} - \mathrm{diag}(\boldsymbol{A})$. We deal with each of them separately.

S1.1 Off-diagonal part

To use Bernstein inequality of bounded matrices later, we consider an event $A_i = \{||X_i||_2^2 \leq U\}$ where $U = C \cdot \mathrm{tr}(\boldsymbol{\Sigma})(K^2 + 1)(t + \log n)$ for some numerical constant $C > 0$. We claim the following:

Fact S1. $\mathrm{P}(\cap_{i=1}^n A_i) \geq 1 - e^{-t}$ for any $t > 0$.

Define a matrix Z_i with zero diagonals

$$Z_i = \mathrm{OD}\left(\left[\frac{\tilde{X}_{ik}\tilde{X}_{i\ell}}{\pi_{k\ell}} \right]_{1 \leq k,\ell \leq p} \right),$$

and $\tilde{Z}_i = Z_i \mathrm{I}_{A_i}$. On the event $\cap_{i=1}^n A_i$, we can get $\mathrm{OD}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) = \frac{1}{n}\sum_{i=1}^n (Z_i - \mathbb{E}Z_i) = \frac{1}{n}\sum_{i=1}^n (\tilde{Z}_i - \mathbb{E}\tilde{Z}_i) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}Z_i \mathrm{I}_{A_i^c}$ and thus

$$||\mathrm{OD}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})||_2 \leq ||\frac{1}{n}\sum_{i=1}^n (\tilde{Z}_i - \mathbb{E}\tilde{Z}_i)||_2 + ||\frac{1}{n}\sum_{i=1}^n \mathbb{E}Z_i \mathrm{I}_{A_i^c}||_2. \tag{S1}$$

For the latter term, we get

$$\begin{aligned} ||\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}Z_i\mathrm{I}_{A_i^c}||_2 &= ||\mathbb{E}Z_1\mathrm{I}_{A_1^c}||_2 \\ &= \max_{\theta\in\mathcal{S}_{p-1}}|\mathbb{E}\theta^\top Z_1\theta\mathrm{I}_{A_1^c}| \\ &\le \max_{\theta\in\mathcal{S}_{p-1}}\mathbb{E}|\theta^\top Z_1\theta|\mathrm{I}_{A_1^c} \\ &\le \max_{\theta\in\mathcal{S}_{p-1}}\sqrt{\mathbb{E}(\theta^\top Z_1\theta)^2\mathbb{E}\mathrm{I}_{A_1^c}} \\ &= \sqrt{\max_{\theta\in\mathcal{S}_{p-1}}\mathbb{E}(\theta^\top Z_1\theta)^2\cdot\mathrm{P}(A_1^c)} \equiv t_2 \end{aligned} \tag{S2}$$

Next, note that $\tilde{Z}_1-\mathbb{E}\tilde{Z}_1$ is bounded conditioning on the set A, which is stated and proved more specifically in (F1) of Fact S2. Hence, we can use Bernstein inequality for the former, and get the upper bound of $||\frac{1}{n}\sum_{i=1}^{n}(\tilde{Z}_i-\mathbb{E}\tilde{Z}_i)||_2$. The following result is from Proposition 2 of Lounici (2014). For $t>0$, with probability at least $1-e^{-t}$, we have (conditioning on the set A)

$$||\frac{1}{n}\sum_{i=1}^{n}(\tilde{Z}_i-\mathbb{E}\tilde{Z}_i)||_2 \le 2\max\left\{\sigma_{\tilde{Z}}\sqrt{\frac{t+\log p}{n}}, 2\pi_{\max}^{(2)}U\frac{t+\log p}{n}\right\} \equiv t_1, \tag{S3}$$

where $\sigma_{\tilde{Z}}^2=||\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\tilde{Z}_i-\mathbb{E}\tilde{Z}_i)^2||_2=||\mathbb{E}(\tilde{Z}_1-\mathbb{E}\tilde{Z}_1)^2||_2$.

Combining (S1), (S2), and (S3), we have

$$\begin{aligned} \mathrm{P}(||\mathrm{OD}(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})||_2 > t_1+t_2) &\le \mathrm{P}(||\mathrm{OD}(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})||_2 > t_1+t_2|A)+\mathrm{P}(A^c) \\ &\le \mathrm{P}(||\frac{1}{n}\sum_{i=1}^{n}(\tilde{Z}_i-\mathbb{E}\tilde{Z}_i)||_2 \\ &\qquad +||\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}Z_i\mathrm{I}_{A_i^c}||_2 > t_1+t_2|A)+\mathrm{P}(A^c) \\ &\le \mathrm{P}(||\frac{1}{n}\sum_{i=1}^{n}(\tilde{Z}_i-\mathbb{E}\tilde{Z}_i)||_2 > t_1|A)+\mathrm{P}(A^c) \\ &\le 2e^{-t}. \end{aligned}$$

The remaining part is to prove the boundedness of $\tilde{Z}_i-\mathbb{E}\tilde{Z}_i$ and calculate constants appearing in t_1 and t_2.

Fact S2. The following statements hold in deterministic sense.

(F1) Conditioning on the set $A = \cap_{i=1}^n \{||X_i||_2^2 \le U\}$, we get

$$||\tilde{Z}_1 - \mathbb{E}\tilde{Z}_1||_2 \le 2\pi_{\max}^{(2)} U,$$

where $\pi_{\max}^{(2)} = \max_{k,\ell} 1/\pi_{k\ell}$.

(F2) $\max_{\theta \in \mathcal{S}_{p-1}} \mathbb{E}(\theta^\top Z_1 \theta)^2 \le CK^4 \pi_{\max}^{(4)} (\mathrm{tr}(\mathbf{\Sigma}))^2$ where

$$\pi_{\max}^{(4)} = \max_{k_1,k_2,\ell_1,\ell_2} \frac{\pi_{k_1k_2\ell_1\ell_2}}{\pi_{k_1\ell_1}\pi_{k_2\ell_2}}$$

(F3) $\sigma_{\tilde{Z}}^2 = ||\mathbb{E}(\tilde{Z}_1 - \mathbb{E}\tilde{Z}_1)^2||_2 \le CK^4 \pi_{\max}^{(3)} (\mathrm{tr}(\mathbf{\Sigma}))^2$ where

$$\pi_{\max}^{(3)} = \max_{s,k,\ell} \frac{\pi_{k\ell s}}{\pi_{ks}\pi_{\ell s}}$$

One can easily check that $\pi_{\max}^{(4)} \ge \max\{\pi_{\max}^{(2)}, \pi_{\max}^{(3)}\}$. Thus, some calculations lead to

$$t_1 + t_2 \le C\mathrm{tr}(\mathbf{\Sigma}) \max\{K^2, 1\} \max\left\{ \sqrt{\frac{\pi_{\max}^{(4)}(t + \log p)}{n}}, (t + \log n) \frac{\pi_{\max}^{(4)}(t + \log p)}{n} \right\},$$

for some $C > 0$ if $t > 1 \vee \log n$.

S1.2 Diagonal part

Remark that the Orlicz norm used in Lounici (2014) and ψ_2-norm in this paper are equivalent, up to a constant factor. Moreover, they both satisfies

$$||\tilde{X}_{ik}||_{\psi_2} \le ||X_{ik}||_{\psi_2}, \quad ||\tilde{X}_{ik}^2||_{\psi_1} \le 2||\tilde{X}_{ik}||_{\psi_2}^2.$$

Using these facts, we get

$$||\tilde{X}_{ik}^2||_{\psi_1} \le 2||\tilde{X}_{ik}||_{\psi_2}^2 \le 2||X_{ik}||_{\psi_2}^2 \le 2\sigma_{kk}K^2.$$

By Proposition 1 of Lounici (2014), we get with probability at least $1 - e^{-t}$

$$\left| \frac{\sum_{i=1}^n \tilde{X}_{ik}^2}{n\pi_k} - \Sigma_{kk} \right| \le \frac{C\sigma_{kk}K^2}{\pi_k} \left(\sqrt{\frac{t}{n}} \vee \frac{t}{n} \right).$$

This implies that with probability at most pe^{-t}

$$\max_k \left| \frac{\sum_{i=1}^n \tilde{X}_{ik}^2}{n\pi_k} - \Sigma_{kk} \right| > CK^2 \max_k \frac{\sigma_{kk}}{\pi_k} \left(\sqrt{\frac{t}{n}} \vee \frac{t}{n} \right)$$

Putting $t \leftarrow t + \log p$, we get

$$\mathrm{P}\left[||\mathrm{diag}(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})||_2 > CK^2 \max_k \frac{\sigma_{kk}}{\pi_k} \left\{ \sqrt{\frac{t + \log p}{n}}, \frac{t + \log p}{n} \right\} \right] \leq e^{-t}$$

S1.3 Proof of Fact S1

Proof. $||X_i||_2^2 - \mathbb{E}||X_i||_2^2$ is sub-exponential satisfying its ψ_2-norm bounded by

$$\begin{aligned} \left|\left| ||X_i||_2^2 - \mathbb{E}||X_i||_2^2 \right|\right|_{\psi_2} &\leq \textstyle\sum_{j=1}^p ||X_{ij}^2||_{\psi_2} + \mathrm{tr}(\mathbf{\Sigma}) \\ &\leq \textstyle\sum_{j=1}^p 2\sigma_{jj} K^2 + \mathrm{tr}(\mathbf{\Sigma}) \\ &= \mathrm{tr}(\mathbf{\Sigma})(2K^2 + 1) \end{aligned}$$

By Proposition 1 of Lounici (2014),

$$\mathrm{P}\Big[||X_i||_2^2 > \mathrm{tr}(\mathbf{\Sigma}) \big\{ 1 + C(2K^2 + 1)(\sqrt{t} \vee t) \big\} \Big] \leq e^{-t}, \quad t > 0.$$

Putting $t \leftarrow t + \log n$ for $n > 2$, we get

$$\mathrm{P}\Big[||X_i||_2^2 > \mathrm{tr}(\mathbf{\Sigma}) \big\{ 1 + C(2K^2 + 1)(t + \log n) \big\} \Big] \leq e^{-t}/n, \quad t > 0.$$

Note that we can find another constant $C' > 0$ such that $\mathrm{tr}(\mathbf{\Sigma}) \big\{ 1 + C(2K^2+1)(t+\log n) \big\} \leq C' \cdot \mathrm{tr}(\mathbf{\Sigma})(K^2 + 1)(t + \log n) \equiv U$. By the union argument, we conclude $\mathrm{P}\Big[\cup_{i=1}^n A_i \Big] \leq e^{-t}$, for $t > 0$. □

S1.4 Proof of (F1) of Fact S2

Proof. Define $V_1 = \left[\frac{Y_{1k}Y_{1\ell}}{\pi_{k\ell}}\right]_{1\le k,\ell\le p}$ and $W_1 = \mathrm{diag}(V_1)$, and thus $Z_1 = V_1 - W_1$ holds. Since $V_1 - Z_1 = W_1 \succcurlyeq 0$, we begin with

$$\begin{aligned} ||Z_1||_2 &\le ||V_1||_2 \\ &= \max_{\theta\in\mathcal{S}_{p-1}} \left| \sum_{k,\ell} \frac{Y_{1k}Y_{1\ell}\theta_k\theta_\ell}{\pi_{k\ell}} \cdot \mathrm{I}_{A_1} \right| \\ &\le \max_{\theta\in\mathcal{S}_{p-1}} \sqrt{\sum_{k,\ell} \frac{Y_{1k}^2 Y_{1\ell}^2}{\pi_{k\ell}^2} \sum_{k,\ell} \theta_k^2\theta_\ell^2} \\ &\le \max_{\theta\in\mathcal{S}_{p-1}} \pi_{\max}^{(2)} \sqrt{\sum_{k,\ell} Y_{1k}^2 Y_{1\ell}^2 \sum_{k,\ell} \theta_k^2\theta_\ell^2} \\ &= \pi_{\max}^{(2)} ||Y_1||_2^2 \end{aligned} \tag{S4}$$

where we used the Cauchy-Schwartz inequality and $\pi_{\max}^{(2)} = \max_{k,\ell} 1/\pi_{k\ell}$. Moreover, we know that

$$||Y_1||_2^2 \le ||X_1||_2^2 \le U,$$

where the last inequality holds conditional on the event A. Combining these with (S4), we can get $||\tilde{Z}_1||_2 \le \pi_{\max}^{(2)} U$. Then, since $||\mathbb{E}\tilde{Z}_1||_2 \le \mathbb{E}||\tilde{Z}_1||_2 \le \mathbb{E}||Z_1||_2$, we get

$$||\tilde{Z}_1 - \mathbb{E}\tilde{Z}_1||_2 \le ||\tilde{Z}_1||_2 + ||\mathbb{E}\tilde{Z}_1||_2 \le ||Z_1||_2 + \mathbb{E}||Z_1||_2 \le 2\pi_{\max}^{(2)} U$$

□

S1.5 Proof of (F2) of Fact S2

Proof. We can get

$$\begin{aligned}
\mathbb{E}(\theta^\top Z_1\theta)^2 &= \mathbb{E}\left(\sum_{1\le k\neq \ell\le p}\frac{Y_{1k}Y_{1\ell}\theta_k\theta_\ell}{\pi_{k\ell}}\right)^2 \\
&= \mathbb{E}\sum_{(k_1,k_2)\neq(\ell_1,\ell_2)}\frac{Y_{1k_1}Y_{1\ell_1}\theta_{k_1}\theta_{\ell_1}}{\pi_{k_1\ell_1}}\frac{Y_{1k_2}Y_{1\ell_2}\theta_{k_2}\theta_{\ell_2}}{\pi_{k_2\ell_2}} \\
&= \mathbb{E}\sum_{k_1,k_2,\ell_1,\ell_2}\frac{Y_{1k_1}Y_{1\ell_1}\theta_{k_1}\theta_{\ell_1}}{\pi_{k_1\ell_1}}\frac{Y_{1k_2}Y_{1\ell_2}\theta_{k_2}\theta_{\ell_2}}{\pi_{k_2\ell_2}} - \mathbb{E}\sum_{k_1,k_2}\frac{Y_{1k_1}^2Y_{1k_2}^2\theta_{k_1}^2\theta_{k_2}^2}{\pi_{k_1}\pi_{k_2}} \\
&\le \sum_{k_1,k_2,\ell_1,\ell_2}\frac{\pi_{k_1k_2\ell_1\ell_2}}{\pi_{k_1\ell_1}\pi_{k_2\ell_2}}\mathbb{E}(X_{1k_1}X_{1k_2}X_{1\ell_1}X_{1\ell_2})\theta_{k_1}\theta_{k_2}\theta_{\ell_1}\theta_{\ell_2} \\
&\le \sqrt{\sum_{k_1,k_2,\ell_1,\ell_2}\left(\frac{\pi_{k_1k_2\ell_1\ell_2}}{\pi_{k_1\ell_1}\pi_{k_2\ell_2}}\right)^2(\mathbb{E}X_{1k_1}X_{1k_2}X_{1\ell_1}X_{1\ell_2})^2\sum_{k_1,k_2,\ell_1,\ell_2}\theta_{k_1}^2\theta_{k_2}^2\theta_{\ell_1}^2\theta_{\ell_2}^2} \\
&\le \pi^{(4)}_{\max}\sqrt{\sum_{k_1,k_2,\ell_1,\ell_2}(\mathbb{E}X_{1k_1}X_{1k_2}X_{1\ell_1}X_{1\ell_2})^2},
\end{aligned}$$

where we used Cauchy-Schwartz inequality in the second inequality. In the third inequality, we define $\pi^{(4)}_{\max} = \max_{k_1,k_2,\ell_1,\ell_2}\frac{\pi_{k_1k_2\ell_1\ell_2}}{\pi_{k_1\ell_1}\pi_{k_2\ell_2}}$. Applying Cauchy-Schwartz inequality twice, we get

$$\mathbb{E}X_{1k_1}X_{1k_2}X_{1\ell_1}X_{1\ell_2} \le \sqrt{\mathbb{E}X_{1k_1}^2X_{1k_2}^2\mathbb{E}X_{1\ell_1}^2X_{1\ell_2}^2} \le \left(\mathbb{E}X_{1k_1}^4\mathbb{E}X_{1k_2}^4\mathbb{E}X_{1\ell_1}^4\mathbb{E}X_{1\ell_2}^4\right)^{1/4}.$$

Thus, we get for any $\theta\in\mathcal{S}^{p-1}$

$$\mathbb{E}(\theta^\top Z_1\theta)^2 \le \pi^{(4)}_{\max}\left(\sum_k\sqrt{\mathbb{E}X_{1k}^4}\right)^2.$$

Finally, using equation (2.1) in Lounici (2014), we get

$$\mathbb{E}X_{1k}^4 \le C||X_{1k}||_{\psi_2}^4 \le CK^4\sigma_{kk}^2, \tag{S5}$$

which concludes the proof.

□

S1.6 Proof of (F3) of Fact S2

Proof. We observe that

$$||\mathbb{E}(\tilde{Z}_1 - \mathbb{E}\tilde{Z}_1)^2||_2 \leq ||\mathbb{E}(\tilde{Z}_1)^2||_2$$

since $\mathbb{E}(\tilde{Z}_1)^2 - \mathbb{E}(\tilde{Z}_1 - \mathbb{E}\tilde{Z}_1)^2 = (\mathbb{E}\tilde{Z}_1)^2 \succcurlyeq 0$. Moreover, we get $||\mathbb{E}(\tilde{Z}_1)^2||_2 = \max_{\theta \in \mathcal{S}_{p-1}} \theta^\top \mathbb{E}(Z_1)^2 \theta \mathrm{I}_{A_1} = ||\mathbb{E}(Z_1)^2||_2$.

Also, recall the relationship $Z_1 = V_1 - W_1$, which implies with the triangular inequality that $||\mathbb{E}(Z_1)^2||_2 = ||\mathbb{E}V_1^2 + \mathbb{E}W_1^2 - \mathbb{E}V_1W_1 - \mathbb{E}W_1V_1||_2 \leq ||\mathbb{E}V_1^2||_2 + ||\mathbb{E}W_1^2||_2 + 2||\mathbb{E}V_1W_1||_2$. Note that

$$\begin{aligned} ||\mathbb{E}V_1W_1||_2 &= \max_{\theta \in \mathcal{S}_{p-1}} |\mathbb{E}\theta^\top V_1 W_1 \theta| \\ &\leq \max_{\theta \in \mathcal{S}_{p-1}} \sqrt{\mathbb{E}(\theta^\top V_1^2 \theta)\mathbb{E}(\theta^\top W_1^2 \theta)} \\ &\leq \sqrt{||\mathbb{E}V_1^2||_2 ||\mathbb{E}W_1^2||_2}. \end{aligned}$$

Therefore, we get $||\mathbb{E}(Z_1)^2||_2 \leq \left(\sqrt{||\mathbb{E}V_1^2||_2} + \sqrt{||\mathbb{E}W_1^2||_2}\right)^2$. We now calculate the last two terms.

First, we calculate $||\mathbb{E}W_1^2||_2$.

$$||\mathbb{E}(W_1)^2||_2 = \sum_k \mathbb{E}Y_{1k}^4 \theta_k^2/\pi_k^2 = \sum_k \mathbb{E}X_{1k}^4 \theta_k^2/\pi_k = \max_k \mathbb{E}X_{1k}^4/\pi_k.$$

Secondly, we compute $||\mathbb{E}(V_1)^2||_2$.

$$\begin{aligned} ||\mathbb{E}(V_1)^2||_2 &= \max_{\theta \in \mathcal{S}_{p-1}} \sum_{k,\ell,s} \frac{\mathbb{E}Y_{1k}Y_{1\ell}Y_{1s}^2}{\pi_{ks}\pi_{\ell s}} \theta_k \theta_\ell \\ &= \max_{\theta \in \mathcal{S}_{p-1}} \sum_s \sum_{k,\ell} \frac{\pi_{k\ell s}}{\pi_{ks}\pi_{\ell s}} \mathbb{E}X_{1s}^2 X_{1k} X_{1\ell} \theta_k \theta_\ell \\ &\leq \max_{\theta \in \mathcal{S}_{p-1}} \sum_s \sqrt{\sum_{k,\ell} \left(\mathbb{E}\frac{\pi_{k\ell s}}{\pi_{ks}\pi_{\ell s}} X_{1s}^2 X_{1k} X_{1\ell}\right)^2 \sum_{k,\ell} \theta_k^2 \theta_\ell^2} \\ &= \pi_{\max}^{(3)} \sum_s \sqrt{\sum_{k,\ell} \left(\mathbb{E}X_{1s}^2 X_{1k} X_{1\ell}\right)^2} \end{aligned}$$

where we used Cauchy-Schwartz inequality and $\pi^{(3)}_{\max} = \max_{s,k,\ell} \frac{\pi_{k\ell s}}{\pi_{ks}\pi_{\ell s}}$. Due to

$$\mathbb{E}X^2_{1s}X_{1k}X_{1\ell} \leq \sqrt{\mathbb{E}X^4_{1s}\mathbb{E}X^2_{1k}X^2_{1\ell}} \leq \sqrt{\mathbb{E}X^4_{1s}\sqrt{\mathbb{E}X^4_{1k}\mathbb{E}X^4_{1\ell}}},$$

we conclude that

$$||\mathbb{E}(V_1)^2||_2 \leq \pi^{(3)}_{\max}\left(\sum_k \sqrt{\mathbb{E}X^4_{1k}}\right)^2.$$

Finally, combining all of these with equation (S5), we get

$$\begin{aligned}||\mathbb{E}(\tilde{Z}_1 - \mathbb{E}\tilde{Z}_1)^2||_2 &\leq \left(\sqrt{\pi^{(3)}_{\max}}\sum_k \sqrt{\mathbb{E}X^4_{1k}} + \sqrt{\max_k \mathbb{E}X^4_{1k}/\pi_k}\right)^2 \\ &\leq CK^4\left(\sqrt{\pi^{(3)}_{\max}}\mathrm{tr}(\boldsymbol{\Sigma}) + \sqrt{\max_k \sigma^2_{kk}/\pi_k}\right)^2.\end{aligned}$$

which concludes the proof because $\max_k 1/\pi_k \leq \pi^{(3)}_{\max}$ and $\max_k \sigma_{kk} \leq \mathrm{tr}(\boldsymbol{\Sigma})$. □

S2 Miscellaneous results

Without the loss of generality, assume that variables in $\mathcal{A}$ come before those in $\mathcal{A}^c$, or we rearrange them to do so. In all the following proofs, we denote block matrices of $\boldsymbol{A}$ decomposed by the subset $\mathcal{A}$ by $\boldsymbol{A}_{\mathcal{A}\mathcal{A}}, \boldsymbol{A}_{\mathcal{A}\mathcal{A}^c}, \boldsymbol{A}_{\mathcal{A}^c\mathcal{A}}, \boldsymbol{A}_{\mathcal{A}^c\mathcal{A}^c}$, respectively.

S2.1 Proof of Proposition 1

Let us review the three conditions used in Theorem 3.4 of Lee et al. (2015) and apply them to our problem in (2.7).

RSC condiction

The first condition is the restricted strong convexity (RSC).

Assumption S1 (RSC)**.** Let $C \subset \mathbb{R}^p$ be some known convex set containing $\boldsymbol{\theta}^*$. The loss

function ℓ is RSC when $\exists m, L > 0$ such that

$$(1)\ \boldsymbol{t}^T \nabla^2 \ell(\boldsymbol{\theta}) \boldsymbol{t} \geq m \boldsymbol{t}^T \boldsymbol{t}, \quad \forall \boldsymbol{\theta} \in C \cap M, \quad \forall \boldsymbol{t} \in C \cap M - C \cap M$$

$$(2)\ \|\nabla^2 \ell(\boldsymbol{\theta}) - \nabla^2 \ell(\boldsymbol{\theta}^*)\|_2 \leq L \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2, \quad \forall \boldsymbol{\theta} \in C$$

The RSC condition is a relaxed version of strong convexity, which is a commonly used assumption for guaranteeing the properties of given loss functions.

In our specified problem, $\nabla^2 \ell(\boldsymbol{\theta}) = \widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}}$. Thus, the RSC condition (2) is satisfied with L with any positive value. Moreover, for ℓ_1-norm, the model space is $M = \{\boldsymbol{\theta} \in \mathbb{R}^p : \boldsymbol{\theta}_{\mathcal{A}^c} = 0\}$ where $\mathcal{A} \subset [p]$ is the support of the true parameter. We note that

$$\min_{\boldsymbol{t} \in \mathbb{R}^p : \|\boldsymbol{t}\|_2 = 1, \boldsymbol{t}_{\mathcal{A}^c} = 0} \boldsymbol{t}^\top \widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}} \boldsymbol{t} = \alpha \lambda_{\min}(\widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}}) + \mu(1-\alpha) \geq \min\{\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}}), \mu\}.$$

Using Weyl's inequality, $||\widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}} - \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}||_2 \leq 0.5 \lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}})$ implies that $\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}}) \geq 0.5 \lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}})$. Now, we set $m = \min\{0.5 \lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}), \mu\}$.

RE condition

The second condition is the irrepresentibility (IR) condition. Let us define a few notions to introduce IR condition. The *support function* on a convex subset $C \subset \mathbb{R}^p$ is defined as:

$$h_C(\boldsymbol{x}) = \sup\{\boldsymbol{x}^\top \boldsymbol{y} : \boldsymbol{y} \in C\}.$$

We say the penalty function ρ is *geometrically decomposable* in terms of $D, I, E \subset \mathbb{R}^p$ if it is decomposed as a sum of support functions:

$$\rho(\boldsymbol{\theta}) = h_D(\boldsymbol{\theta}) + h_I(\boldsymbol{\theta}) + h_{E^\perp}(\boldsymbol{\theta}),$$

where D is a convex bounded set, I is a convex bounded set which contains a relative neighborhood of the origin (i.e. $0 \in \mathrm{relint}(E)$) and E is a subspace. Now, we can define our second condition, IR condition.

Assumption S2 (IR)**.** $\exists \tau \in (0,1)$ such that

$$\sup_{\boldsymbol{z} \in \partial h_D(M)} V\Big[\boldsymbol{P}_{M^\perp}\{\boldsymbol{Q}\boldsymbol{P}_M(\boldsymbol{P}_M\boldsymbol{Q}\boldsymbol{P}_M)^\dagger \boldsymbol{P}_M \boldsymbol{z} - \boldsymbol{z}\}\Big] \leq 1-\tau$$

where $\boldsymbol{Q} = \nabla^2 \ell(\boldsymbol{\theta}^*) = \widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}}$, $\boldsymbol{P}_B$ is the projection matrix to B,

$$\partial h_D(M) = \bigcup_{\boldsymbol{u} \in M} \partial h_D(\boldsymbol{u})$$

$$\gamma_C(\boldsymbol{x}) = \inf\{\lambda : \boldsymbol{x} \in \lambda C\}$$

$$V(\boldsymbol{u}) = \inf\{\gamma_I(\boldsymbol{y}) + \mathbf{1}_{E^\perp}(\boldsymbol{u}-\boldsymbol{y})\} = \inf_{t \in E^\perp} \gamma_I(\boldsymbol{u}-\boldsymbol{t}),$$

We can easily check that ρ is geometrically decomposed with the terms of

$$E = \mathbb{R}^p$$

$$D = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_\infty \leq 1, \boldsymbol{\theta}_{\mathcal{A}^c} = 0\}, \quad \mathrm{span}(D) = M$$

$$I = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_\infty \leq 1, \boldsymbol{\theta}_{\mathcal{A}} = 0\}, \quad \mathrm{span}(I) = M^\perp$$

$$h_D(\boldsymbol{\theta}) = \|\boldsymbol{\theta}_{\mathcal{A}}\|_1, \quad h_I(\boldsymbol{\theta}) = \|\boldsymbol{\theta}_{\mathcal{A}^c}\|_1.$$

Then, the RE condition becomes equivalent to:

$$\exists \tau \in (0,1) \quad \text{s.t.} \quad \|\widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}}_{\mathcal{A}^c\mathcal{A}}(\widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}}_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty \leq 1-\tau \tag{S6}$$

which is the classical irrepresentability, proposed in Zhao and Yu (2006).

Proof of (S6).

$$\begin{aligned}
\partial h_D(\boldsymbol{\theta}) &= \{\boldsymbol{y} \in D : \boldsymbol{y}^\top \boldsymbol{\theta} = h_D(\boldsymbol{\theta})\} \\
&= \{\boldsymbol{y} \in D : \boldsymbol{y}^\top \boldsymbol{\theta} = \|\boldsymbol{\theta}_{\mathcal{A}}\|_1\} \\
&= \mathrm{sgn}(\boldsymbol{\theta}) \\
\partial h_D(M) &= \{\mathrm{sgn}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in M\}
\end{aligned}$$

$$\boldsymbol{P}_M = \begin{bmatrix} \mathbf{I}_{|\mathcal{A}|} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \qquad \boldsymbol{P}_{M^\perp} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-|\mathcal{A}|} \end{bmatrix}$$

$$\begin{aligned} (\boldsymbol{P}_M \boldsymbol{Q} \boldsymbol{P}_M)^\dagger &= \begin{bmatrix} \boldsymbol{Q}_{\mathcal{A}\mathcal{A}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}^\dagger \\ &= \begin{bmatrix} (\boldsymbol{Q}^*_{\mathcal{A}\mathcal{A}} \boldsymbol{Q}_{\mathcal{A}\mathcal{A}})^\dagger \boldsymbol{Q}^*_{\mathcal{A}\mathcal{A}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \boldsymbol{P}_{M^\perp}\{\boldsymbol{Q}\boldsymbol{P}_M(\boldsymbol{P}_M\boldsymbol{Q}\boldsymbol{P}_M)^\dagger \boldsymbol{P}_M \boldsymbol{z} - \boldsymbol{z}\} &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \boldsymbol{Q}_{\mathcal{A}^c\mathcal{A}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} (\boldsymbol{Q}^*_{\mathcal{A}\mathcal{A}} \boldsymbol{Q}_{\mathcal{A}\mathcal{A}})^\dagger \boldsymbol{Q}^*_{\mathcal{A}\mathcal{A}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_1 \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{0} \\ \boldsymbol{z}_2 \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} \\ \boldsymbol{Q}_{\mathcal{A}^c\mathcal{A}} (\boldsymbol{Q}^*_{\mathcal{A}\mathcal{A}} \boldsymbol{Q}_{\mathcal{A}\mathcal{A}})^\dagger \boldsymbol{Q}^*_{\mathcal{A}\mathcal{A}} \boldsymbol{z}_1 - \boldsymbol{z}_2 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} &\sup_{\boldsymbol{z} \in \partial h_D(M)} V\Big[\boldsymbol{P}_{M^\perp}\{\boldsymbol{Q}\boldsymbol{P}_M(\boldsymbol{P}_M\boldsymbol{Q}\boldsymbol{P}_M)^\dagger \boldsymbol{P}_M \boldsymbol{z} - \boldsymbol{z}\}\Big] \\ &= \sup_{\boldsymbol{z} \in \partial h_D(M)} V\left(\begin{bmatrix} \mathbf{0} \\ \boldsymbol{Q}_{\mathcal{A}^c\mathcal{A}} (\boldsymbol{Q}^*_{\mathcal{A}\mathcal{A}} \boldsymbol{Q}_{\mathcal{A}\mathcal{A}})^\dagger \boldsymbol{Q}^*_{\mathcal{A}\mathcal{A}} \boldsymbol{z}_1 - \boldsymbol{z}_2 \end{bmatrix}\right) \\ &= \sup_{\boldsymbol{\theta}_1 \in \mathbb{R}^{|\mathcal{A}|}} V\left(\begin{bmatrix} \mathbf{0} \\ \boldsymbol{Q}_{\mathcal{A}^c\mathcal{A}} (\boldsymbol{Q}^*_{\mathcal{A}\mathcal{A}} \boldsymbol{Q}_{\mathcal{A}\mathcal{A}})^\dagger \boldsymbol{Q}^*_{\mathcal{A}\mathcal{A}} \mathrm{sgn}(\boldsymbol{\theta}_1) \end{bmatrix}\right) \\ &= \sup_{\boldsymbol{\theta}_1 \in \mathbb{R}^{|\mathcal{A}|}} \|\boldsymbol{Q}_{\mathcal{A}^c\mathcal{A}} (\boldsymbol{Q}^*_{\mathcal{A}\mathcal{A}} \boldsymbol{Q}_{\mathcal{A}\mathcal{A}})^\dagger \boldsymbol{Q}^*_{\mathcal{A}\mathcal{A}} \mathrm{sgn}(\boldsymbol{\theta}_1)\|_\infty \end{aligned}$$

Since $\boldsymbol{Q}_{\mathcal{A}\mathcal{A}}$ is invertible due to Assumption 4, we have

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}_1 \in \mathbb{R}^{|\mathcal{A}|}} \|\boldsymbol{Q}_{\mathcal{A}^c\mathcal{A}}(\boldsymbol{Q}^*_{\mathcal{A}\mathcal{A}}\boldsymbol{Q}_{\mathcal{A}\mathcal{A}})^\dagger \boldsymbol{Q}^*_{\mathcal{A}\mathcal{A}}\mathrm{sgn}(\boldsymbol{\theta}_1)\|_\infty \\
=&\sup_{\boldsymbol{\theta}_1 \in \mathbb{R}^{|\mathcal{A}|}} \|\boldsymbol{Q}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{Q}^{-1}_{\mathcal{A}\mathcal{A}}\mathrm{sgn}(\boldsymbol{\theta}_1)\|_\infty \\
=&\|\boldsymbol{Q}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{Q}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty
\end{aligned}$$

□

BG condition

The last condition is the bounded gradient (BG) condition. Let us first define related constants. The compatibility constant, denoted by κ_ρ, between ρ and ℓ_2-norm on M is defined by

$$\kappa_\rho = \sup_{\boldsymbol{\theta}}\{\rho(\boldsymbol{\theta}) | \boldsymbol{\theta} \in \mathcal{B}_2 \cap M\},$$

where $\mathcal{B}_2$ is the ℓ_2-unit ball. The compatibility constant between the irrepresentable term and ρ^* is given as

$$\kappa_{\mathrm{IC}} = \sup_{\rho^*(\mathbf{z}) \le \mathbf{1}} V\Big[\boldsymbol{P}_{M^\perp}\{\boldsymbol{Q}\boldsymbol{P}_M(\boldsymbol{P}_M\boldsymbol{Q}\boldsymbol{P}_M)^\dagger \boldsymbol{P}_M\mathbf{z} - \mathbf{z}\}\Big].$$

We can state the third condition with the constants κ_ρ and κ_{IC}, which decides a suitable range of a tuning parameter λ.

Assumption S3 (BG)**.**

$$\frac{4\kappa_{\mathrm{IC}}}{\tau}\rho^*(\nabla\ell(\boldsymbol{\theta}^*)) < \lambda < \frac{m^2}{2L}\left(2\kappa_\rho + \frac{\kappa_\rho}{\kappa_{\mathrm{IC}}}\frac{\tau}{2}\right)^{-2}\frac{\tau}{\kappa_{\rho^*}\kappa_{\mathrm{IC}}}.$$

Now, we check the preliminaries for the BG condition. In our case, ρ is the ℓ_1-norm, $\kappa_\rho =$

$\sqrt{|\mathcal{A}|}$ and $\kappa_{\rho^*} = 1$. As for κ_{IC}:

$$\begin{aligned}\kappa_{\text{IC}} &= \sup_{\rho^*(\mathbf{z})\le \mathbf{1}} V\Big[\boldsymbol{P}_{M^\perp}\{\boldsymbol{Q}\boldsymbol{P}_M(\boldsymbol{P}_M\boldsymbol{Q}\boldsymbol{P}_M)^\dagger \boldsymbol{P}_M\mathbf{z}-\mathbf{z}\}\Big] \\ &= \sup_{\|\mathbf{z}\|_\infty\le 1} \|\boldsymbol{Q}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{Q}_{\mathcal{A}\mathcal{A}}^{-1}\mathbf{z}_1 - \mathbf{z}_2\|_\infty \\ &= \|\boldsymbol{Q}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{Q}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty + 1\end{aligned}$$

Recall the BG condition for λ:

$$\frac{4\kappa_{\text{IC}}}{\tau}\rho^*(\nabla\ell(\boldsymbol{\theta}^*)) < \lambda < \frac{m^2}{2L}\left(2\kappa_\rho + \frac{\kappa_\rho}{\kappa_{\text{IC}}}\frac{\tau}{2}\right)^{-2}\frac{\tau}{\kappa_{\rho^*}\kappa_{\text{IC}}}.$$

With the IR condition, we have $\kappa_{\text{IC}} \le 2-\tau$. Also, since L can be of any value, the right side of the BG condition holds. So, the following is sufficient for the BG condition:

$$\frac{4(2-\tau)}{\tau}\|\nabla\ell(\boldsymbol{\theta}^*)\|_\infty < \lambda.$$

Conclusion

Under the three conditions above, Lee et al. (2015) concluded the following results for the solution.

1. The minimizer is unique.
2. ℓ_2 consistency: $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \le \frac{2}{m}\left(\kappa_\rho + \frac{\tau}{4}\frac{\kappa_\rho}{\kappa_{\text{IC}}}\right)\lambda$
3. Model selection consistency : $\hat{\boldsymbol{\theta}} \in M$.

In our problem (2.7), the ℓ_2 consistency is

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}^{\text{LPD}} - \boldsymbol{\beta}^*\|_2 &\le \frac{2}{\min\{0.5\lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}), \mu\}}\left(\sqrt{|\mathcal{A}|} + \frac{\tau}{4}\frac{\sqrt{|\mathcal{A}|}}{\|\boldsymbol{Q}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{Q}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty + 1}\right)\lambda \\ &\le \frac{2}{\min\{0.5\lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}), \mu\}}\left(1 + \frac{\tau}{4}\right)\sqrt{|\mathcal{A}|}\lambda,\end{aligned}$$

and the model selection consistency is $\widehat{\boldsymbol{\beta}}^{\text{LPD}}_{\mathcal{A}^c} = 0$.

S2.2 Proof of Proposition 2

Proof. Let $a_{d,\max} = \max_j a_{jj}$, $a_{d,\min} = \min_j a_{jj}$.

$$\begin{aligned}
\left|\left|\Phi_{\mu,\alpha^*} - \boldsymbol{A}\right|\right|_{\max} &= \frac{(\epsilon - \lambda_{\min}(\boldsymbol{A}))\left|\left|\boldsymbol{A} - \mu\mathbf{I}\right|\right|_{\max}}{\mu - \lambda_{\min}(\boldsymbol{A})} \\
&= (\epsilon - \lambda_{\min}(\boldsymbol{A}))\frac{\max_{i\neq j}|a_{ij}| \vee \max_i |a_{ii} - \mu|}{\mu - \lambda_{\min}(\boldsymbol{A})} \\
&= (\epsilon - \lambda_{\min}(\boldsymbol{A}))\frac{\max_{i\neq j}|a_{ij}| \vee |a_{d,\max} - \mu| \vee |a_{d,\min} - \mu|}{\mu - \lambda_{\min}(\boldsymbol{A})}
\end{aligned}$$

We now denote $a_{\text{off},\max} = \max_{i\neq j}|a_{ij}|$, $\Psi(\mu) = \dfrac{a_{\text{off},\max} \vee |a_{d,\max} - \mu| \vee |a_{d,\min} - \mu|}{\mu - \lambda_{\min}(\boldsymbol{A})}$, and consider two disjoint cases: Case (i) $(a_{d,\max} - a_{d,\min})/2 > a_{\text{off},\max}$ and Case (ii) $(a_{d,\max} - a_{d,\min})/2 \leq a_{\text{off},\max}$. For each case, we divide up the value of μ into multiple cases, which is summarized in Figure S1.

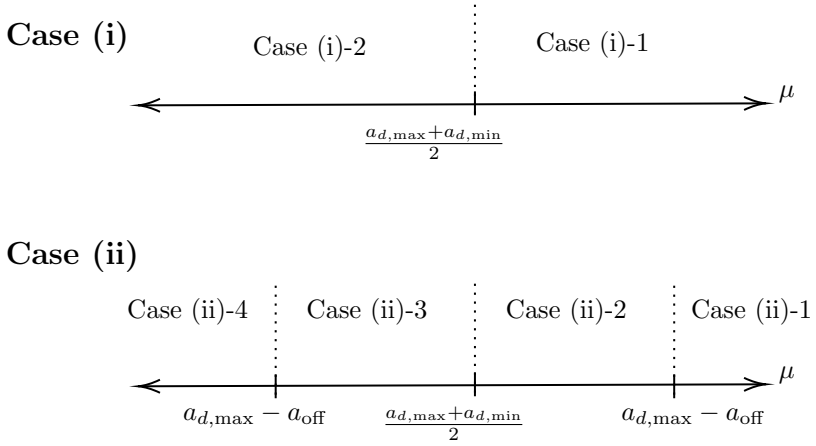

Figure S1: Summary of cases used in the proof. Case (i) (top) and Case (ii) (bottom).

Case (i): $(a_{d,\max} - a_{d,\min})/2 > a_{\mathbf{off},\max}$

For this case, we consider two sub-cases based on the value of μ.

Case (i)-1: $\mu > (a_{d,\max} + a_{d,\min})/2$

Under Case (i)-1, we have $|a_{d,\max} - \mu| < |a_{d,\min} - \mu|$. Moreover, note that by Case (i)

$$\frac{a_{d,\max} + a_{d,\min}}{2} = \frac{a_{d,\max} - a_{d,\min}}{2} + a_{d,\min} > a_{d,\min} + a_{\text{off},\max}$$

and thus $\mu - a_{d,\min} > a_{\text{off},\max}$. Combining these two, we can simplify Ψ by

$$\Psi(\mu) = \frac{|a_{d,\min} - \mu|}{\mu - \lambda_{\min}(\boldsymbol{A})} = \frac{\mu - a_{d,\min}}{\mu - \lambda_{\min}(\boldsymbol{A})} = \frac{\lambda_{\min}(\boldsymbol{A}) - a_{d,\min}}{\mu - \lambda_{\min}(\boldsymbol{A})} + 1. \tag{S7}$$

From the last expression, we can see that Ψ is increasing in μ because $a_{d,\min} > \lambda_{\min}(\boldsymbol{A})$. Thus, the minimum value under the case considered is

$$\min\left\{\Psi(\mu) : \mu > (a_{d,\max} + a_{d,\min})/2\right\} \geq \frac{(a_{d,\max} - a_{d,\min})/2}{(a_{d,\max} + a_{d,\min})/2 - \lambda_{\min}(\boldsymbol{A})},$$

where the right-hand side is achieved by plugging-in $\mu = \dfrac{a_{d,\max} + a_{d,\min}}{2}$ into (S7).

Case (i)-2: $\mu \leq (a_{d,\max} + a_{d,\min})/2$

Under Case (i)-2, we have $|a_{d,\max} - \mu| \geq |a_{d,\min} - \mu|$. Moreover, note that by Case (i)

$$a_{\text{off},\max} < \frac{a_{d,\max} - a_{d,\min}}{2} = a_{d,\max} - \frac{a_{d,\max} + a_{d,\min}}{2}$$

and thus $a_{d,\max} - \mu > a_{\text{off},\max}$. Combining these two, we can simplify Ψ by

$$\Psi(\mu) = \frac{|a_{d,\max} - \mu|}{\mu - \lambda_{\min}(\boldsymbol{A})} = \frac{a_{d,\max} - \mu}{\mu - \lambda_{\min}(\boldsymbol{A})} = \frac{a_{d,\max} - \lambda_{\min}(\boldsymbol{A})}{\mu - \lambda_{\min}(\boldsymbol{A})} - 1. \tag{S8}$$

The last expression tells us that Ψ is decreasing in μ because $a_{d,\max} > \lambda_{\min}(\boldsymbol{A})$. Then, we get

$$\min\left\{\Psi(\mu) : \mu \leq (a_{d,\max} + a_{d,\min})/2\right\} = \frac{(a_{d,\max} - a_{d,\min})/2}{(a_{d,\max} + a_{d,\min})/2 - \lambda_{\min}(\boldsymbol{A})}.$$

Combining the two results from Case (i)-1,2, we conclude that if $(a_{d,\max} - a_{d,\min})/2 > a_{\text{off},\max}$, then the minimum value of Ψ is

$$\min_{\mu:\mu\geq\epsilon} \Psi(\mu) = \frac{(a_{d,\max} - a_{d,\min})/2}{(a_{d,\max} + a_{d,\min})/2 - \lambda_{\min}(\boldsymbol{A})}$$

at $\mu = (a_{d,\max} + a_{d,\min})/2$.

Case (ii): $(a_{d,\max} - a_{d,\min})/2 \leq a_{\mathbf{off},\max}$

Similarly to before, we consider sub-cases based on the value of μ.

Case (ii)-1: $\mu > a_{d,\min} + a_{\mathbf{off},\max}$

Note that $a_{d,\min} + a_{\text{off},\max} \geq (a_{d,\max} + a_{d,\min})/2$ under Case (ii). Then, we have $|a_{d,\max} - \mu| < |a_{d,\min} - \mu| = \mu - a_{d,\min}$. Moreover, by Case (ii)-1, $|a_{d,\min} - \mu| = \mu - a_{d,\min} > a_{\text{off},\max}$.

Thus, we can simplify Ψ by

$$\Psi(\mu) = \frac{|a_{d,\min} - \mu|}{\mu - \lambda_{\min}(\boldsymbol{A})} = \frac{\mu - a_{d,\min}}{\mu - \lambda_{\min}(\boldsymbol{A})} = \frac{\lambda_{\min}(\boldsymbol{A}) - a_{d,\min}}{\mu - \lambda_{\min}(\boldsymbol{A})} + 1. \tag{S9}$$

Case (ii)-2: $(a_{d,\max} + a_{d,\min})/2 < \mu \leq a_{d,\min} + a_{\mathbf{off},\max}$

In Case (ii)-2, we still have $|a_{d,\max} - \mu| < |a_{d,\min} - \mu| = \mu - a_{d,\min}$ as in Case (ii)-1, but $|a_{d,\min} - \mu| = \mu - a_{d,\min} \geq a_{\text{off},\max}$ holds.

Case (ii)-3: $a_{d,\max} - a_{\mathbf{off},\max} < \mu \leq (a_{d,\max} + a_{d,\min})/2$

From $\mu \leq (a_{d,\max} + a_{d,\min})/2$, we have $|a_{d,\max} - \mu| \geq |a_{d,\min} - \mu|$. Moreover, since $a_{d,\max} - a_{\text{off},\max} < \mu$, $|a_{d,\max} - \mu| = a_{d,\max} - \mu < a_{\text{off},\max}$.

Case (ii)-4: $\mu \leq a_{d,\max} - a_{\mathbf{off},\max}$

Note that $a_{d,\max} - a_{\text{off},\max} \leq (a_{d,\max} + a_{d,\min})/2$ under Case (ii). Thus, we have $|a_{d,\max} - \mu| \geq |a_{d,\min} - \mu|$. Since $\mu \leq a_{d,\max} - a_{\text{off},\max}$, $|a_{d,\max} - \mu| = a_{d,\max} - \mu \geq a_{\text{off},\max}$.

Combining the four cases, we can summarize that

$$\Psi(\mu) = \begin{cases} \dfrac{\mu - a_{d,\min}}{\mu - \lambda_{\min}(\boldsymbol{A})}, & \text{for Case (ii)-1} \\ \dfrac{a_{\text{off},\max}}{\mu - \lambda_{\min}(\boldsymbol{A})}, & \text{for Case (ii)-2,3} \\ \dfrac{a_{d,\max} - \mu}{\mu - \lambda_{\min}(\boldsymbol{A})}, & \text{for Case (ii)-4} \end{cases}$$

We note that this function decreases until $\mu < a_{d,\min} + a_{\text{off},\max}$ and increases after that point, which implies $\mu = a_{d,\min} + a_{\text{off},\max}$ give the minimum value

$$\min_{\mu:\mu\geq\epsilon} \Psi(\mu) = \frac{a_{\text{off},\max}}{a_{d,\min} + a_{\text{off},\max} - \lambda_{\min}(\boldsymbol{A})}.$$

□

S3 Proof of the main theorems

S3.1 Proof of Theorem 1

The proof of Theorem 1 is based on Theorem S2, S3, which are stated below.

Theorem S2. *Let Assumption 1, 2, 3, 4 hold. Let us focus on the case of the estimator* $\widehat{\boldsymbol{\Sigma}}^{\text{IPW}}$ *such that* $\widehat{\boldsymbol{\Sigma}}^{\text{IPW}}_{\mathcal{A}\mathcal{A}}$ *is non-singular and the smallest eigenvalue satisfies* $\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}^{\text{IPW}}) \leq 0$. *For any* $\mu > \epsilon$, *we construct the LPD estimator* $\Phi_{\mu,\alpha^*}(\widehat{\boldsymbol{\Sigma}}^{\text{IPW}})$ *with* $\alpha^* = (\mu - \epsilon)/(\mu - \lambda_{\min}(\widehat{\boldsymbol{\Sigma}}^{\text{IPW}}))$. *Then, the LPD estimator satisfies the irrepresentability condition for some constant* $\tilde{\tau} \in (0, 1)$, *if the events hold true*

$$\left\|\widehat{\boldsymbol{\Sigma}}^{\text{IPW}}_{\mathcal{A}\mathcal{A}} - \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}\right\|_\infty + \left\|\widehat{\boldsymbol{\Sigma}}^{\text{IPW}}_{\mathcal{A}^c\mathcal{A}} - \boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\right\|_\infty + \frac{\mu}{\mu - \epsilon}\left\|\widehat{\boldsymbol{\Sigma}}^{\text{IPW}} - \boldsymbol{\Sigma}\right\|_2 \leq \frac{\tau}{\left\|\boldsymbol{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\right\|_\infty}, \tag{S10}$$

The proof is pended until Supplementary Materials S3.2. The other case when the smallest eigenvalue is positive is addressed by the following theorem.

Theorem S3. *Let Assumption 1, 2, 3, 4(b) hold where $\tau \in (0,1)$ is the constant from Assumption 4(b). Let us focus on the case of the estimator $\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}$ such that $\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}}$ is non-singular and the smallest eigenvalue satisfies $\lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}) > 0$. Then, the LPD estimator $\Phi_{\mu,\alpha^*}(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}})$, which is reduced to $\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}$ with $\alpha^* = 1$, satisfies the irrepresentability condition for some constant $\tilde{\tau} \in (0,1)$, if the event holds true*

$$\left\|\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\right\|_\infty + \left\|\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}^c\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\right\|_\infty \leq \tau / \left\|\mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\right\|_\infty . \tag{S11}$$

The proof is pended until Supplementary Materials S3.4.

Proof of Theorem 1. We calculate the probability of the event E that the LPD estimator satisfies the irrepresentability condition as follows. Let the event $A = \{\lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}) > 0\}$.

$$\begin{aligned}
\mathrm{P}(E) &= \mathrm{P}\left(E\middle|A\right)\mathrm{P}(A) + \mathrm{P}\left(E\middle|A^c\right)\mathrm{P}(A^c) \\
&\geq \mathrm{P}\left((S11)\text{ holds}\middle|A\right)\mathrm{P}(A) + \mathrm{P}\left((S10)\text{ holds}\middle|A^c\right)\mathrm{P}(A^c)\,(\because \text{ Theorem S2, S3}) \\
&\geq \mathrm{P}\left((S10)\text{ holds}\middle|A\right)\mathrm{P}(A) + \mathrm{P}\left((S10)\text{ holds}\middle|A^c\right)\mathrm{P}(A^c)\,(\because (S10) \Rightarrow (S11)) \\
&= \mathrm{P}\left((S10)\text{ holds}\right).
\end{aligned}$$

Note that for $\tilde{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}^{\mathrm{IPW}} - \mathbf{\Sigma}$, we have

$$\left\|\tilde{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\right\|_\infty + \left\|\tilde{\mathbf{\Sigma}}_{\mathcal{A}^c\mathcal{A}}\right\|_\infty \leq 2\left\|\tilde{\mathbf{\Sigma}}\right\|_{\infty,\mathcal{A}} = 2\left\|\tilde{\mathbf{\Sigma}}\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\right\|_\infty \leq 2\left\|\tilde{\mathbf{\Sigma}}\right\|_\infty \leq 2\left\|\tilde{\mathbf{\Sigma}}\right\|_2 .$$

Then, using $\mu/(\mu-\epsilon) \leq 2$ for $\mu \geq 2\epsilon$, a sufficient condition for (S10) is

$$\left\|\tilde{\mathbf{\Sigma}}\right\|_2 \leq \frac{\tau}{4\left\|\mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\right\|_\infty}.$$

Theorem S1 states that for any $u > 0$, if $n > \pi^{(4)}_{\max}(u+1)^3\log^3(p \vee n)$, then it holds with probability at least $1 - 3/p^u$

$$||\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}} - \mathbf{\Sigma}||_2 \leq C\mathrm{tr}(\mathbf{\Sigma})\max\{(K^x)^2, 1\}\sqrt{u+1}\sqrt{\frac{\pi^{(4)}_{\max}\log p}{n}}.$$

Hence, if the following condition is satisfied

$$C\mathrm{tr}(\mathbf{\Sigma})\max\{(K^x)^2,1\}\sqrt{u+1}\sqrt{\frac{\pi^{(4)}_{\max}\log p}{n}} \le \frac{\tau}{4\left\|\mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\right\|_\infty},$$

then we can guarantee $\mathrm{P}\left((S10)\text{ holds}\right) \ge 1-3/p^u$, where the above gives another sample size condition:

$$n/(\pi^{(4)}_{\max}\log p) \ge 4C\left\{\frac{\mathrm{tr}(\mathbf{\Sigma})\max\{(K^x)^2,1\}\sqrt{u+1}}{\tau/\left\|\mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\right\|_\infty}\right\}^2.$$

Finally, we deal with (C3) of Proposition 1. By Weyl's inequality, the condition is satisfied if $||\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}||_2 \le 0.5\lambda_{\min}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})$ holds. Following the proof of Theorem 1, we can have a similar probabilistic argument for the event $\{||\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}||_2 \le 0.5\lambda_{\min}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})\}$. That is, $||\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}||_2 \le 0.5\lambda_{\min}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})$ with probability greater than $1-3/p^u$ for $u>0$ if the sample size satisfies

$$\frac{n}{\pi^{(4)}_{\max,\mathcal{A}}\log|\mathcal{A}|} \ge c\left\{\frac{\mathrm{tr}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})\max\{(K^x)^2,1\}\sqrt{u+1}}{1/\lambda_{\min}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})}\right\}^2, \quad n > c\,\pi^{(4)}_{\max,\mathcal{A}}(u+1)^3\log^3(|\mathcal{A}|\vee n),$$

for some $c>0$. Here, $\pi^{(4)}_{\max,\mathcal{A}} = \max_{k_1,k_2,\ell_1,\ell_2\in\mathcal{A}}\pi^{xx}_{k_1k_2\ell_1\ell_2}/(\pi^{xx}_{k_1\ell_1}\pi^{xx}_{k_2\ell_2})$.

□

S3.2 Proof of Theorem S2

It should be noted that the proof of the theorem only depends on the distances between $\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}$ and $\mathbf{\Sigma}$ (or their block matrices), but not any other characteristic of the IPW estimate or the population covariance matrix.

We define the matrix norms that appear in the following proof.

$$\eta_1 = \left\|\mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\right\|_\infty, \quad \eta_2 = \left\|\mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\right\|_\infty$$

$$\delta_1 = \left\|\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\right\|_\infty, \quad \delta_2 = \left\|\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}^c\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\right\|_\infty, \quad \delta_3 = \left\|\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}} - \mathbf{\Sigma}\right\|_2.$$

We first introduce the lemma to ease calculation.

Lemma S1. *Let* $\widehat{\mathbf{\Sigma}}^{\mathrm{LPD}} = \Phi_{\mu,\alpha}(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}})$. *Assume*

$$\eta_1\delta_1 < 1 \text{ and } \quad \frac{(1-\alpha)\mu}{\alpha}\|\big(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}}\big)^{-1}\|_\infty < 1. \tag{S12}$$

Then, we have

$$\Big\|\widehat{\mathbf{\Sigma}}^{\mathrm{LPD}}_{\mathcal{A}^c\mathcal{A}}\big(\widehat{\mathbf{\Sigma}}^{\mathrm{LPD}}_{\mathcal{A}\mathcal{A}}\big)^{-1}\Big\|_\infty \leq \frac{\eta_1\delta_2+\eta_2}{1-\eta_1\delta_1-\alpha^{-1}(1-\alpha)\mu\eta_1}.$$

The proof is given in Supplementary Materials S3.3. Using Lemma S1 and the irrpresentability condition for $\mathbf{\Sigma}$ (i.e. $\eta_2 < 1-\tau$) together, we get

$$\Big\|\widehat{\mathbf{\Sigma}}^{\mathrm{LPD}}_{\mathcal{A}^c\mathcal{A}}\big(\widehat{\mathbf{\Sigma}}^{\mathrm{LPD}}_{\mathcal{A}\mathcal{A}}\big)^{-1}\Big\|_\infty < \frac{\eta_1\delta_2+1-\tau}{1-\eta_1\delta_1-\alpha^{-1}(1-\alpha)\mu\eta_1}. \tag{S13}$$

It remains to claim the right-hand side of the above is strictly less than 1, which is equivalent to show

$$\delta_1+\delta_2 < \tau/\eta_1 - \alpha^{-1}(1-\alpha)\mu.$$

Plugging-in $\alpha^* = (\mu-\epsilon)/(\mu-\lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}))$ and using $\lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}) \geq -\delta_3 + \lambda_{\min}(\mathbf{\Sigma})$ derived by Weyl's inequality, we get a sufficient condition for (S13)

$$\delta_1+\delta_2+\frac{\mu\delta_3}{\mu-\epsilon} < \frac{\tau}{\eta_1} + \frac{\mu(\lambda_{\min}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})-\epsilon)}{\mu-\epsilon}. \tag{S14}$$

Remark that the right-hand side term is greater than 0 if $\min\{\mu, \lambda_{\min}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})\} > \epsilon$.

We remain to show (S12) holds with high probability when plugging-in $\alpha^* = (\mu-\epsilon)/(\mu-\lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}))$, but instead, we will calculate the probability of another sufficient condition (S15) described in the following lemma. One can easily check that (S15) is implied by (S14) because $\mu/(\mu-\epsilon) > 1$ and $\tau < 1$, which concludes the proof.

Lemma S2. *Consider the class of covariance matrices such that* $1/\eta_1 - \epsilon + \lambda_{\min}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}) > 0$. *Let us focus on the case of the estimator* $\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}$ *with* $\lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}) < 0$. *If we choose* $\mu > \epsilon$, *then*

$$\delta_1 + \frac{\mu\delta_3}{\mu-\epsilon} \leq 1/\eta_1 + \frac{\mu(\lambda_{\min}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})-\epsilon)}{\mu-\epsilon}, \tag{S15}$$

implies (S12).

The proof of the lemma is given in Supplementary Materials S3.3.

S3.3 Proof of lemmas used in Theorem S2

Proof of Lemma S1. We introduce three inequalities and suspend their proofs.

$$\|\widehat{\mathbf{\Sigma}}^{\mathrm{LPD}}_{\mathcal{A}^c\mathcal{A}}\big(\widehat{\mathbf{\Sigma}}^{\mathrm{LPD}}_{\mathcal{A}\mathcal{A}}\big)^{-1}\|_\infty \leq \frac{\|\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}^c\mathcal{A}}\big(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}}\big)^{-1}\|_\infty}{1-\alpha^{-1}(1-\alpha)\mu\|\big(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}}\big)^{-1}\|_\infty}, \quad \text{if } \frac{(1-\alpha)\mu}{\alpha}\|\big(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}}\big)^{-1}\|_\infty < 1, \tag{S16}$$

$$\|\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}^c\mathcal{A}}\big(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}}\big)^{-1} - \mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty \leq \frac{\eta_1\left(\eta_2\delta_1+\delta_2\right)}{1-\eta_1\delta_1}, \quad \text{if } \eta_1\delta_1<1, \tag{S17}$$

$$\|\big(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}}\big)^{-1}\|_\infty \leq \frac{\eta_1}{1-\eta_1\delta_1}, \quad \text{if } \eta_1\delta_1<1, \tag{S18}$$

Combining the triangular inequality with (S16), we get

$$\|\widehat{\mathbf{\Sigma}}^{\mathrm{LPD}}_{\mathcal{A}^c\mathcal{A}}\big(\widehat{\mathbf{\Sigma}}^{\mathrm{LPD}}_{\mathcal{A}\mathcal{A}}\big)^{-1}\|_\infty \leq \frac{\|\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}^c\mathcal{A}}\big(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}}\big)^{-1} - \mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty + \|\mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty}{1-\alpha^{-1}(1-\alpha)\mu\|\big(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}}\big)^{-1}\|_\infty}.$$

This completes the proof if (S17), (S18) are combined with the upper bound.

We now prove the above inequalities. The proofs of (S17) and (S18) are from that of Lemma A2 by Mai et al. (2012), but we show them here for completeness. Using the basic property of operator norms,

$$\begin{aligned}
\|\big(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}}\big)^{-1} - \mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty &= \|\mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})\big(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}}\big)^{-1}\|_\infty \\
&\leq \|\mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty \cdot \|\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\|_\infty \cdot \|\big(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}}\big)^{-1}\|_\infty \\
&\leq \|\mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty \times \|\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\|_\infty \\
&\quad \times \big(\|\big(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}}\big)^{-1} - \mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty + \|\mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty\big).
\end{aligned}$$

Arranging the inequality, we get

$$\|\big(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}}\big)^{-1} - \mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty \leq \frac{\|\mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\|^2_\infty\|\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\|_\infty}{1-\|\mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty\|\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\|_\infty},$$

since $\|\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}} - \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}\|_\infty < 1$ by the assumption. Then, by the triangular inequality,

$$\begin{aligned}\|\big(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}}\big)^{-1}\|_\infty &\leq \|\big(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}}\big)^{-1} - \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty + \|\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \\ &\leq \frac{\|\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty^2\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}} - \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}\|_\infty}{1 - \|\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}} - \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}\|_\infty} + \|\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty,\end{aligned} \tag{S19}$$

which achieves (S18). Next, we also exploit the basic properties of norms to get

$$\begin{aligned}&\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}^c\mathcal{A}}^{\mathrm{IPW}}\big(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}}\big)^{-1} - \boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \\ &\quad= \|(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}^c\mathcal{A}} - \boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}})\big(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}}\big)^{-1}\|_\infty \\ &\quad= \|(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}^c\mathcal{A}} - \boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}} + \boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}} - \boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}})\big(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}}\big)^{-1}\|_\infty \\ &\quad\leq \|\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}^c\mathcal{A}} - \boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}} + \boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}(\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}} - \widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}})\|_\infty\|\big(\big(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}}\big)^{-1}\|_\infty \\ &\quad\leq \big(\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}^c\mathcal{A}}^{\mathrm{IPW}} - \boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\|_\infty + \|\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}} - \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}\|_\infty\big)\|\big(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}}\big)^{-1}\|_\infty.\end{aligned}$$

By using (S18) in the last inequality, we obtain (S17). To prove (S16), we observe

$$\begin{aligned}\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}^c\mathcal{A}}^{\mathrm{LPD}}\big(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{LPD}}\big)^{-1}\|_\infty &= \|\alpha\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}^c\mathcal{A}}^{\mathrm{IPW}}(\alpha\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}} + (1-\alpha)\mu\mathbf{I})^{-1}\|_\infty \\ &= \|\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}^c\mathcal{A}}^{\mathrm{IPW}}\big(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}}\big)^{-1}(\mathbf{I} + \alpha^{-1}(1-\alpha)\mu\big(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}}\big)^{-1})^{-1}\|_\infty \\ &\leq \|\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}^c\mathcal{A}}^{\mathrm{IPW}}\big(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}}\big)^{-1}\|_\infty\|(\mathbf{I} + \alpha^{-1}(1-\alpha)\mu\big(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}}\big)^{-1})^{-1}\|_\infty \\ &\leq \|\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}^c\mathcal{A}}^{\mathrm{IPW}}\big(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}}\big)^{-1}\|_\infty\big(1 - \alpha^{-1}(1-\alpha)\mu\|\big(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}}\big)^{-1}\|_\infty\big)^{-1}\end{aligned}$$

where the last inequality depends on that for any operator norm $\|\cdot\|$ and a matrix $\boldsymbol{U}$,

$$\|(\mathbf{I} + \boldsymbol{U})^{-1}\| \leq \frac{1}{1 - \|\boldsymbol{U}\|}, \quad \text{if } \|\boldsymbol{U}\| < 1.$$

To use it, we need the following condition

$$\alpha^{-1}(1-\alpha)\mu\|\big(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{\mathrm{IPW}}\big)^{-1}\|_\infty < 1.$$

□

Proof of Lemma S2. Putting $\alpha^* = (\mu - \epsilon)/(\mu - \lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\text{IPW}}))$, we want to show

$$\frac{(1-\alpha^*)\mu}{\alpha^*}\|\big(\widehat{\mathbf{\Sigma}}^{\text{IPW}}_{\mathcal{A}\mathcal{A}}\big)^{-1}\|_\infty = \frac{\mu}{\mu-\epsilon}(\epsilon - \lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\text{IPW}}))\|\big(\widehat{\mathbf{\Sigma}}^{\text{IPW}}_{\mathcal{A}\mathcal{A}}\big)^{-1}\|_\infty < 1. \tag{S20}$$

Remark that by Weyl's inequality

$$\lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\text{IPW}}) \geq -\left\|\widehat{\mathbf{\Sigma}}^{\text{IPW}} - \mathbf{\Sigma}\right\|_2 + \lambda_{\min}(\mathbf{\Sigma}),$$

and recall (S19)

$$\left\|(\widehat{\mathbf{\Sigma}}^{\text{IPW}}_{\mathcal{A}\mathcal{A}})^{-1}\right\|_\infty \leq \frac{\eta_1}{1-\eta_1\delta_1}.$$

Some basic algebra with these two leads to a sufficient condition of (S20):

$$\left\|\widehat{\mathbf{\Sigma}}^{\text{IPW}}_{\mathcal{A}\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\right\|_\infty + \frac{\mu\left\|\widehat{\mathbf{\Sigma}}^{\text{IPW}} - \mathbf{\Sigma}\right\|_2}{\mu-\epsilon} \leq 1/\left\|\mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\right\|_\infty + \frac{\mu(\lambda_{\min}(\mathbf{\Sigma})-\epsilon)}{\mu-\epsilon}.$$

□

S3.4 Proof of Theorem S3

Proof. If the smallest eigenvalue of the IPW estimator is positive, the LPD estimator of it is the IPW estimator, i.e. $\alpha^* = 1$. By following the same proof of Lemma S1, we have

$$\left\|\widehat{\mathbf{\Sigma}}^{\text{IPW}}_{\mathcal{A}^c\mathcal{A}}\big(\widehat{\mathbf{\Sigma}}^{\text{IPW}}_{\mathcal{A}\mathcal{A}}\big)^{-1}\right\|_\infty \leq \frac{\eta_1\delta_2+\eta_2}{1-\eta_1\delta_1}, \quad \text{if } \eta_1\delta_1 < 1.$$

where we use the same definitions of the matrix norms:

$$\eta_1 = \left\|\mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\right\|_\infty, \quad \eta_2 = \left\|\mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\right\|_\infty$$
$$\delta_1 = \left\|\widehat{\mathbf{\Sigma}}^{\text{IPW}}_{\mathcal{A}\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\right\|_\infty, \quad \delta_2 = \left\|\widehat{\mathbf{\Sigma}}^{\text{IPW}}_{\mathcal{A}^c\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\right\|_\infty.$$

Using $\eta_2 < 1-\tau$, it is sufficient for the irrepresentability condition of $\widehat{\mathbf{\Sigma}}^{\text{IPW}}$ to show

$$\frac{\eta_1\delta_2 + 1 - \tau}{1-\eta_1\delta_1} < 1.$$

The above is equivalent to $\delta_1 + \delta_2 < \tau/\eta_1$.

□

S3.5 Proof of Theorem 2

Proof. Using $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}^* + \epsilon_i$ in calculating $\hat{\boldsymbol{\rho}}^{\mathrm{IPW}}$, we can obtain

$$\begin{aligned}\nabla \ell(\boldsymbol{\beta}^*; \widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}}, \hat{\boldsymbol{\rho}}^{\mathrm{IPW}}) &= \widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}} \boldsymbol{\beta}^* - \hat{\boldsymbol{\rho}}^{\mathrm{IPW}} \\ &= \left(\widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}} - \boldsymbol{V}\right) \boldsymbol{\beta}^* - \boldsymbol{w}\end{aligned}$$

where $\boldsymbol{V} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{w} \in \mathbb{R}^p$ have its element respectively by

$$\begin{aligned} v_{jk} &= n^{-1} \sum_{i=1}^n x_{ij} x_{ik} \delta_{ij}^x \delta_i^y / \pi_j^{xy}, \quad 1 \le j, k \le p, \\ w_j &= n^{-1} \sum_{i=1}^n x_{ij} \epsilon_i \delta_{ij}^x \delta_i^y / \pi_j^{xy}, \quad 1 \le j \le p \end{aligned}$$

where $\pi_j^{xy} = \mathrm{P}(\delta_1^y = \delta_{1j}^x = 1)$. Hence, the norm of the gradient is

$$\begin{aligned}\|\nabla \ell(\boldsymbol{\beta}^*; \widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}}, \hat{\boldsymbol{\rho}}^{\mathrm{IPW}})\|_\infty &\le \left\| \left(\widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}} - \boldsymbol{V}\right) \boldsymbol{\beta}^* \right\|_\infty + \|\boldsymbol{w}\|_\infty \\ &= \max_{1\le j\le p} \sum_{k\in\mathcal{A}} \left| \left(\widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}} - \boldsymbol{V}\right)_{jk} \right| |\beta_k^*| + \|\boldsymbol{w}\|_\infty \\ &\le \|\widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}} - \boldsymbol{V}\|_{\infty,\mathcal{A}} \, \beta_{\max}^* + \|\boldsymbol{w}\|_\infty\end{aligned}$$

where the first inequality is from the triangular inequality, the next equality holds because $\beta_k^* = 0$ for $k \in \mathcal{A}^{\mathsf{c}}$, and the last inequality is obvious from definitions $\beta_{\max}^* = \max_{1\le j\le p} |\beta_j^*|$ and $\|\boldsymbol{B}\|_{\infty,\mathcal{A}} = \max_{1\le j\le p} \sum_{k\in\mathcal{A}} |b_{jk}|$ for any matrix $\boldsymbol{B} = (b_{jk})_{p\times p}$. Note that $\|\boldsymbol{B}\|_{\infty,\mathcal{A}}$ is a semi-norm on $\mathbb{R}^{p\times p}$ given a non-empty set $\mathcal{A}$ (i.e. $\|\boldsymbol{B}\|_{\infty,\mathcal{A}} = 0$ does not imply $\boldsymbol{B} = 0$). Finally, using $\widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}} - \boldsymbol{V} = \alpha^*(\widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}} - \boldsymbol{\Sigma}) + (1-\alpha^*)(\mu\mathbf{I} - \boldsymbol{\Sigma}) - (\boldsymbol{V} - \boldsymbol{\Sigma})$ and the triangular inequality, we get

$$\begin{aligned}\|\nabla \ell(\boldsymbol{\beta}^*; \widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}}, \hat{\boldsymbol{\rho}}^{\mathrm{IPW}})\|_\infty \le \Big(&\|\widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}} - \boldsymbol{\Sigma}\|_{\infty,\mathcal{A}} + (1-\alpha^*)\|\mu\mathbf{I} - \boldsymbol{\Sigma}\|_{\infty,\mathcal{A}} \\ &+ \|\boldsymbol{\Sigma} - \boldsymbol{V}\|_{\infty,\mathcal{A}} \Big) \beta_{\max}^* + \|\boldsymbol{w}\|_\infty.\end{aligned} \tag{S21}$$

We use Lemma 1 of Park et al. (2023) to the terms above except the second. Let us define a function f by

$$f(n, p, \mathcal{B}) = |\mathcal{B}| \sqrt{\frac{2\log p + \log|\mathcal{B}|}{2n}}, \quad \mathcal{B} \subset [p],$$

$\sigma_{\max} = \max_{jj} \sigma_{jj}$, and probabilities $\pi^{xx}_{\min,\mathcal{A}} = \min_{1\le j\le p, k\in\mathcal{A}} \pi^{xx}_{jk}$, $\pi^{xx}_{\min} = \min_{1\le j,k\le p} \pi^{xx}_{jk}$, $\pi^{xy}_{\min} = \min_{1\le j\le p} \pi^{xy}_{j}$. Then, we can easily get the followings: for some numerical constants $c_1, c_2, c_3, C_1, C_2, C_3 > 0$ such that

$$\mathrm{P}_{\delta,x}\left(\|\widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}} - \boldsymbol{\Sigma}\|_{\infty,\mathcal{A}} \ge \frac{C_1 (K^x)^2 \sigma_{\max}}{\sqrt{\pi^{xx}_{\min,\mathcal{A}}}} f(n,p,\mathcal{A}) \right) \le 2/p, \tag{S22}$$

if $\dfrac{n}{2\log p + \log|\mathcal{A}|} > \dfrac{1}{c_1 \pi^{xx}_{\min,\mathcal{A}}}$,

$$\mathrm{P}_{\delta,x}\left(\|\boldsymbol{V} - \boldsymbol{\Sigma}\|_{\infty,\mathcal{A}} \ge \frac{C_2 (K^x)^2 \sigma_{\max}}{\sqrt{\pi^{xy}_{\min}}} f(n,p,\mathcal{A}) \right) \le 2/p, \tag{S23}$$

if $\dfrac{n}{2\log p + \log|\mathcal{A}|} > \dfrac{1}{c_2 \pi^{xy}_{\min}}$, and

$$\mathrm{P}_{\delta,x}\left(\|\boldsymbol{w}\|_{\infty} \ge \frac{C_3 \sqrt{\sigma_{\max}\sigma_{\epsilon\epsilon}} K^x K^{\epsilon}}{\sqrt{\pi^{xy}_{\min}}} f(n,p,[1]) \right) \le 2/p, \tag{S24}$$

if $\dfrac{n}{3\log p} > \dfrac{1}{c_3 \pi^{xy}_{\min}}$. Moreover, we get the concentration of the second term: for some $c_4, C_4 > 0$

$$\begin{aligned} \mathrm{P}_{\delta,x}\Bigg((1-\alpha^*)\|\mu\mathbf{I} - \boldsymbol{\Sigma}\|_{\infty,\mathcal{A}} \ge C_4 \mathrm{tr}(\boldsymbol{\Sigma}) \max\{(K^x)^2, 1\} \\ \times \left(1 + \frac{\|\boldsymbol{\Sigma}\|_{\infty,\mathcal{A}}}{\mu} \right) \sqrt{\pi^{(4)}_{\max}} f(n,p,[1]) \Bigg) \le 3/p, \end{aligned} \tag{S25}$$

if $n > c_4 \pi^{(4)}_{\max} \log^3(p \vee n)$. The proof of (S25) is pended until the end of the proof.

Combining these results, it holds with probability greater than $1 - 9/p$

$$\|\nabla \ell(\boldsymbol{\beta}^*; \widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}}, \hat{\boldsymbol{\rho}}^{\mathrm{IPW}})\|_{\infty} \le L \cdot f(n,p,\mathcal{A}),$$

if $n > c \max\left\{ \log p/\pi^{xy}_{\min}, \pi^{(4)}_{\max} \log^3(p \vee n) \right\}$ for some numerical constant $c > 0$. The factor $L > 0$ is a function of parameters given by

$$\begin{aligned} L \propto \beta^*_{\max} \max\{(K^x)^2, 1\} \sqrt{\pi^{(4)}_{\max}} \mathrm{tr}(\boldsymbol{\Sigma}) \left(1 + \frac{\|\boldsymbol{\Sigma}\|_{\infty,\mathcal{A}}}{\mu} \right) \\ + \frac{\max\left\{ \sqrt{\sigma_{\max}\sigma_{\epsilon\epsilon}} K^x K^{\epsilon}, \sigma_{\max}(K^x)^2 \right\}}{\sqrt{\pi^{xy}_{\min}}}. \end{aligned}$$

To derive the constant L, we used $\pi^{(4)}_{\max} \geq 1/\pi^{xx}_{\min,\mathcal{A}}$. Note that if $\lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\text{IPW}}) > 0$, the second term in (S21) no longer exists since $\alpha^* = 0$. Then, we only need to combine (S22), (S23), (S24), which leads to another $L' > 0$ smaller than L. The constant given in the statement of the theorem is deriven considering it.

Now, we prove (S25), which depends on the following lemma.

Lemma S3. *Assume ϵ is smaller than the smallest eigenvalue of $\mathbf{\Sigma}$. For $\alpha^* = \mathrm{I}(\lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\text{IPW}}) > 0) + (\mu - \epsilon)/(\mu - \lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\text{IPW}}))\mathrm{I}(\lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\text{IPW}}) \leq 0)$, we have*

$$1 - \alpha^* \leq \|\widehat{\mathbf{\Sigma}}^{\text{IPW}} - \mathbf{\Sigma}\|_2/\mu$$

Proof. By definition of α^*, we have

$$1 - \alpha^* = (\epsilon - \lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\text{IPW}}))/(\mu - \lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\text{IPW}}))\mathrm{I}(\lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\text{IPW}}) \leq 0).$$

Now, we observe

$$\begin{aligned}
\frac{\epsilon - \lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\text{IPW}})}{\mu - \lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\text{IPW}})}\mathrm{I}(\lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\text{IPW}}) \leq 0) &\leq \frac{(\epsilon - \lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\text{IPW}}))_+}{\mu} \\
&\leq \frac{(\lambda_{\min}(\mathbf{\Sigma}) - \lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\text{IPW}}))_+}{\mu} \\
&\leq \frac{\|\widehat{\mathbf{\Sigma}}^{\text{IPW}} - \mathbf{\Sigma}\|_2}{\mu}
\end{aligned}$$

where we use Weyl's inequality in the last inequality. □

By applying Lemma S3, we get

$$(1-\alpha^*)\|\mu\mathbf{I} - \mathbf{\Sigma}\|_{\infty,\mathcal{A}} \leq \|\widehat{\mathbf{\Sigma}}^{\text{IPW}} - \mathbf{\Sigma}\|_2 \frac{\|\mu\mathbf{I} - \mathbf{\Sigma}\|_{\infty,\mathcal{A}}}{\mu} \leq \|\widehat{\mathbf{\Sigma}}^{\text{IPW}} - \mathbf{\Sigma}\|_2 \left(1 + \frac{\|\mathbf{\Sigma}\|_{\infty,\mathcal{A}}}{\mu}\right), \quad \text{(S26)}$$

From Theorem S1, if the sample size condition $n > \pi^{(4)}_{\max}(\alpha+1)^3 \log^3(p \vee n)$ is satisfied, it holds with probability at least $1 - 3/p^\alpha$ that

$$||\widehat{\mathbf{\Sigma}}^{\text{IPW}} - \mathbf{\Sigma}||_2 \leq C\mathrm{tr}(\mathbf{\Sigma}) \max\{(K^x)^2, 1\} \sqrt{\frac{\pi^{(4)}_{\max}(\alpha+1)\log p}{n}}, \quad \text{(S27)}$$

where $C > 0$ is some numerical constant. This concludes that if $n > 16\pi_{\max}^{(4)} \log^3(p \vee n)$

$$\mathrm{P}_{\delta,x}\Bigg((1-\alpha^*)\|\mu\mathbf{I} - \mathbf{\Sigma}\|_{\infty,\mathcal{A}} \geq C\mathrm{tr}(\mathbf{\Sigma}) \max\{(K^x)^2, 1\} \\ \times \left(1 + \frac{\|\mathbf{\Sigma}\|_{\infty,\mathcal{A}}}{\mu}\right) \sqrt{\frac{2\pi_{\max}^{(4)} \log p}{n}} \Bigg) \leq 3/p.$$

□

S4 Additional details/results of simulation study

S4.1 The corrected cross-validation

For the cross-validation, we split data into K folds. Let $\widehat{\boldsymbol{\beta}}_k(\lambda)$ be the solution of any penalized regression estimated with tuning parameter at λ and with all samples but in the k-th fold. Given a set Λ of candidates, we aim to find the best one that minimizes the prediction error on the k-th fold:

$$\hat{\lambda}_{opt} = \arg\min_{\lambda \in \Lambda} \sum_{k=1}^{K} (\widehat{\boldsymbol{\beta}}_k(\lambda))^\top (\widehat{\mathbf{\Sigma}}_k^{\mathrm{IPW}})_+ \widehat{\boldsymbol{\beta}}_k(\lambda) - 2\hat{\boldsymbol{\rho}}_k \widehat{\boldsymbol{\beta}}_k(\lambda).$$

Here, we define

$$(\widehat{\mathbf{\Sigma}}_k^{\mathrm{IPW}})_+ = \begin{cases} \mu\alpha\widehat{\mathbf{\Sigma}}_k^{\mathrm{IPW}} + (1-\alpha)\mathbf{I}, & \text{for cases of LPD, NCL} \\ \min_{\mathbf{\Sigma} \succeq 0} \left\| \widehat{\mathbf{\Sigma}}_k^{\mathrm{IPW}} - \mathbf{\Sigma} \right\|_{\max}, & \text{for cases of CoCo,} \end{cases}$$

and $\widehat{\mathbf{\Sigma}}_k^{\mathrm{IPW}}$ is the IPW estimate calculated over samples in the k-th fold, and $\hat{\boldsymbol{\rho}}_k$ is similarly defined.

S4.2 Method comparison

We focus on comparing a list of variants of LPD. For spectral norm and ℓ_∞-norm, any value over some lower bound, say μ_{lb}, will do, so we suggest trying $k \cdot \mu_{lwr}$, $k = 1, 3, 5$, to see how

much their performances are different. Considering these variants, we name our proposals by LPD-*norm*-k where $norm \in \{\mathrm{S}, \mathrm{F}, \mathrm{I}, \mathrm{E}\}$ and $k \in \{1, 3, 5\}$, resulting 8 estimators (LPD-S-1, LPD-S-3, LPD-S-5, LPD-F-1, LPD-I-1, LPD-I-3, LPD-I-5, LPD-E-1).

	$p = 200, s = 0.05$					
	PE	MSE	pAUC	F_1	TP	FP
TL	1.915 (0.609)	3.656 (1.145)	0.953 (0.031)	0.439 (0.071)	9.680 (0.513)	25.560 (7.484)
NL	3.694 (1.034)	6.160 (1.638)	0.879 (0.063)	0.396 (0.069)	8.620 (1.086)	25.720 (7.420)
CoCo	3.385 (0.927)	6.441 (1.772)	0.830 (0.065)	0.400 (0.076)	8.440 (1.163)	24.460 (6.102)
NCL	5.158 (1.222)	6.292 (1.601)	0.508 (0.075)	0.453 (0.093)	8.140 (1.309)	19.060 (10.442)
LPD-E-1	3.290 (0.840)	6.308 (1.659)	0.879 (0.054)	0.369 (0.070)	8.780 (0.996)	29.840 (7.313)
LPD-F-1	3.608 (0.927)	6.534 (1.708)	0.881 (0.053)	0.350 (0.063)	8.880 (0.982)	32.920 (7.948)
LPD-L-1	3.311 (0.867)	6.262 (1.640)	0.879 (0.053)	0.370 (0.066)	8.800 (1.050)	29.640 (7.551)
LPD-L-3	3.242 (0.844)	6.131 (1.548)	0.878 (0.056)	0.377 (0.062)	8.780 (1.036)	28.320 (5.223)
LPD-L-5	3.260 (0.806)	6.182 (1.515)	0.880 (0.054)	0.376 (0.066)	8.820 (1.004)	28.780 (6.075)
LPD-S-1	3.256 (0.828)	6.181 (1.572)	0.879 (0.055)	0.376 (0.067)	8.780 (0.996)	28.680 (6.149)
LPD-S-3	3.251 (0.817)	6.165 (1.530)	0.878 (0.054)	0.376 (0.064)	8.800 (1.050)	28.680 (5.527)
LPD-S-5	3.300 (0.839)	6.282 (1.578)	0.878 (0.055)	0.363 (0.067)	8.780 (0.996)	30.560 (7.654)
	$p = 500, s = 0.05$					
	PE	MSE	pAUC	F_1	TP	FP
TL	6.039 (1.193)	11.825 (2.347)	0.809 (0.048)	0.420 (0.050)	22.980 (1.286)	62.980 (16.109)
NL	17.374 (4.272)	27.698 (3.981)	0.535 (0.081)	0.278 (0.055)	12.240 (2.966)	50.440 (9.311)
CoCo	16.370 (2.833)	31.179 (4.848)	0.596 (0.046)	0.276 (0.051)	11.880 (2.847)	49.060 (9.421)
NCL	28.492 (7.734)	27.538 (3.863)	0.504 (0.061)	0.212 (0.055)	14.560 (5.035)	106.460 (55.869)
LPD-E-1	18.634 (3.463)	29.315 (4.630)	0.703 (0.057)	0.247 (0.044)	14.760 (2.959)	80.900 (19.125)
LPD-F-1	26.511 (6.173)	31.870 (5.696)	0.702 (0.054)	0.238 (0.045)	14.920 (2.687)	88.020 (25.206)
LPD-L-1	14.017 (2.209)	26.636 (3.549)	0.703 (0.056)	0.250 (0.045)	14.580 (2.829)	78.020 (17.977)
LPD-L-3	14.030 (2.391)	26.661 (4.044)	0.704 (0.054)	0.251 (0.044)	14.560 (2.865)	77.400 (17.331)
LPD-L-5	13.869 (2.186)	26.393 (3.570)	0.704 (0.055)	0.252 (0.043)	14.540 (2.887)	76.380 (14.380)
LPD-S-1	13.923 (2.078)	26.499 (3.362)	0.704 (0.055)	0.251 (0.042)	14.440 (2.786)	76.700 (17.765)
LPD-S-3	13.853 (2.097)	26.377 (3.434)	0.703 (0.053)	0.253 (0.043)	14.520 (2.880)	75.660 (15.904)
LPD-S-5	14.129 (2.182)	26.761 (3.763)	0.703 (0.055)	0.251 (0.047)	14.600 (2.871)	78.200 (21.832)

Table S1: Method comparison for $p = 200, 500$ and $s = 0.05, 0.1$. Each performance measures are averaged over $R = 100$ repetitions (standard deviation in parenthesis).

	$p = 200, s = 0.1$					
	PE	MSE	pAUC	F_1	TP	FP
TL	3.220 (0.763)	6.251 (1.483)	0.916 (0.034)	0.532 (0.066)	19.600 (0.606)	35.220 (9.790)
NL	11.020 (3.241)	15.799 (3.181)	0.755 (0.061)	0.434 (0.059)	14.240 (2.273)	31.440 (5.444)
CoCo	9.878 (2.507)	17.890 (4.268)	0.715 (0.053)	0.431 (0.068)	13.640 (2.145)	29.980 (7.150)
NCL	17.212 (3.866)	17.602 (2.613)	0.614 (0.045)	0.386 (0.100)	14.280 (2.241)	46.520 (27.309)
LPD-E-1	9.085 (1.956)	17.196 (3.661)	0.765 (0.054)	0.406 (0.056)	14.880 (2.086)	38.960 (9.167)
LPD-F-1	10.020 (2.320)	17.907 (3.941)	0.765 (0.054)	0.394 (0.054)	14.900 (2.082)	41.260 (8.689)
LPD-L-1	8.914 (2.040)	16.123 (3.352)	0.764 (0.054)	0.414 (0.056)	14.700 (2.053)	36.660 (7.176)
LPD-L-3	8.868 (1.969)	16.161 (3.436)	0.768 (0.054)	0.415 (0.055)	14.780 (2.122)	36.660 (6.394)
LPD-L-5	8.916 (2.131)	16.137 (3.395)	0.765 (0.055)	0.414 (0.056)	14.780 (2.141)	36.800 (6.958)
LPD-S-1	8.819 (2.044)	16.157 (3.432)	0.765 (0.055)	0.413 (0.052)	14.740 (2.058)	36.780 (6.538)
LPD-S-3	8.840 (2.057)	16.113 (3.424)	0.764 (0.053)	0.414 (0.056)	14.700 (2.112)	36.500 (6.519)
LPD-S-5	9.045 (2.218)	16.381 (3.655)	0.764 (0.056)	0.411 (0.059)	14.760 (2.036)	37.660 (8.277)
	$p = 500, s = 0.1$					
	PE	MSE	pAUC	F_1	TP	FP
TL	14.102 (2.010)	27.752 (4.021)	0.684 (0.045)	0.474 (0.048)	43.740 (2.284)	92.480 (21.073)
NL	48.511 (11.754)	75.830 (9.527)	0.392 (0.062)	0.272 (0.056)	16.840 (3.966)	56.320 (7.377)
CoCo	47.069 (8.296)	90.279 (15.734)	0.547 (0.032)	0.254 (0.048)	15.180 (3.336)	53.820 (8.075)
NCL	76.743 (26.682)	64.362 (9.807)	0.492 (0.038)	0.245 (0.038)	25.380 (7.545)	130.100 (42.421)
LPD-E-1	59.310 (12.606)	81.429 (11.177)	0.606 (0.045)	0.260 (0.047)	20.820 (4.341)	89.180 (17.235)
LPD-F-1	93.961 (23.197)	91.393 (14.167)	0.606 (0.044)	0.252 (0.044)	21.160 (4.560)	96.360 (18.729)
LPD-L-1	37.572 (5.268)	72.016 (9.589)	0.601 (0.044)	0.261 (0.044)	20.900 (4.273)	89.580 (15.831)
LPD-L-3	37.343 (5.633)	71.308 (10.009)	0.606 (0.043)	0.263 (0.047)	20.620 (4.125)	86.680 (17.115)
LPD-L-5	37.214 (5.183)	71.073 (9.155)	0.606 (0.044)	0.263 (0.047)	20.800 (4.536)	87.240 (14.981)
LPD-S-1	37.091 (4.728)	70.722 (8.250)	0.603 (0.042)	0.264 (0.046)	20.600 (4.267)	85.180 (16.184)
LPD-S-3	36.894 (4.797)	70.567 (8.786)	0.604 (0.045)	0.264 (0.049)	20.600 (4.290)	85.440 (14.098)
LPD-S-5	36.937 (5.200)	70.630 (9.674)	0.605 (0.046)	0.264 (0.048)	20.420 (4.121)	84.700 (15.538)

Table S2: Method comparison for $p = 200, 500$ and $s = 0.05, 0.1$. Each performance measures are averaged over $R = 100$ repetitions (standard deviation in parenthesis).

Among four matrix norms considered here, ℓ_∞-norm (LPD-L) and spectral norm (LPD-S) perform best, while different μ values do not result in any significant changes in practice. The other two norms do not achieve comparative results when the dimension increases to $p = 500$.

S4.3 Missing rate and missing mechanism

We try different missing rates and mechanisms to investigate the robustness of each method under other scenarios of missing data generation. This is similar to the idea of sensitivity

analysis in missing data literature (Kolar and Xing, 2012; van Buuren, 2018). We generate missing values by the three mechanisms known as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Following Kolar and Xing (2012), every third variable ($j = 1, \ldots, \lfloor p/3 \rfloor$) is subject to missing; for MAR case, $\delta^x_{i,3j} = 0$ if $X_{i,3j-2} < \Phi^{-1}(1-\theta)$ and for MNAR case, $\delta^x_{i,3j} = 0$ if $X_{i,3j} < \Phi^{-1}(1-\theta)$. Here, we fix $s = 0.05$ and $p = 200$.

Table S3 confirms that a higher rate of missing in data can lead to worse performance. Also, the performance gets poorer as the missing mechanism changes from MCAR to MAR, MNAR, but interestingly, the results on relative performance are not much different.

	$\theta = 0.9$, MAR					
	PE	MSE	pAUC	F_1	TP	FP
TL	1.860 (0.536)	3.558 (1.059)	0.948 (0.039)	0.455 (0.063)	9.700 (0.544)	23.640 (5.784)
NL	3.654 (1.052)	5.989 (1.528)	0.866 (0.067)	0.389 (0.076)	8.500 (1.074)	26.220 (7.731)
CoCo	3.229 (0.861)	6.179 (1.627)	0.832 (0.064)	0.387 (0.084)	8.340 (1.171)	25.980 (8.482)
NCL	4.823 (1.126)	6.149 (1.613)	0.548 (0.091)	0.428 (0.113)	8.080 (1.275)	23.260 (17.444)
LPD-E-1	3.316 (0.907)	6.227 (1.672)	0.879 (0.058)	0.346 (0.071)	8.680 (0.935)	32.940 (9.182)
LPD-F-1	3.451 (0.937)	6.240 (1.652)	0.877 (0.059)	0.343 (0.065)	8.740 (0.944)	33.660 (9.164)
LPD-L-1	3.147 (0.836)	5.934 (1.482)	0.876 (0.060)	0.371 (0.065)	8.520 (1.054)	28.240 (6.962)
LPD-S-1	3.094 (0.815)	5.893 (1.484)	0.877 (0.060)	0.366 (0.065)	8.500 (1.015)	28.760 (6.133)
	$\theta = 0.7$, MAR					
	PE	MSE	pAUC	F_1	TP	FP
TL	1.828 (0.490)	3.512 (0.991)	0.956 (0.037)	0.438 (0.076)	9.740 (0.600)	26.040 (7.982)
NL	9.796 (2.676)	8.887 (1.463)	0.718 (0.100)	0.290 (0.073)	5.600 (1.400)	24.060 (9.646)
CoCo	6.027 (1.422)	10.851 (2.433)	0.666 (0.096)	0.303 (0.075)	5.480 (1.344)	21.080 (5.606)
NCL	6.813 (1.513)	10.039 (1.974)	0.466 (0.081)	0.312 (0.091)	4.980 (1.363)	17.500 (5.694)
LPD-E-1	7.048 (3.141)	11.014 (3.025)	0.743 (0.093)	0.253 (0.060)	6.400 (1.539)	34.400 (7.910)
LPD-F-1	21.120 (34.859)	14.843 (8.075)	0.746 (0.096)	0.235 (0.078)	6.140 (2.204)	36.020 (9.079)
LPD-L-1	5.344 (1.177)	9.132 (1.592)	0.744 (0.096)	0.285 (0.061)	6.540 (1.216)	29.960 (5.577)
LPD-S-1	5.238 (1.050)	9.163 (1.526)	0.742 (0.093)	0.283 (0.060)	6.520 (1.233)	30.180 (6.521)
	$\theta = 0.9$, MNAR					
	PE	MSE	pAUC	F_1	TP	FP
TL	1.937 (0.558)	3.697 (1.087)	0.951 (0.033)	0.430 (0.073)	9.700 (0.463)	26.700 (8.122)
NL	3.952 (1.097)	6.682 (1.552)	0.857 (0.063)	0.369 (0.077)	8.080 (1.412)	26.500 (7.492)
CoCo	3.698 (1.010)	7.055 (1.988)	0.817 (0.066)	0.361 (0.075)	8.060 (1.219)	27.820 (8.578)
NCL	5.062 (1.149)	6.917 (1.581)	0.584 (0.070)	0.372 (0.109)	7.720 (1.325)	28.600 (19.799)
LPD-E-1	3.624 (0.817)	6.807 (1.588)	0.852 (0.063)	0.341 (0.065)	8.200 (1.229)	30.840 (7.980)
LPD-F-1	3.679 (0.758)	6.784 (1.474)	0.851 (0.064)	0.336 (0.050)	8.320 (1.186)	31.680 (6.485)
LPD-L-1	3.470 (0.893)	6.602 (1.685)	0.850 (0.064)	0.351 (0.064)	8.220 (1.217)	29.360 (7.331)
LPD-S-1	3.478 (0.786)	6.586 (1.509)	0.851 (0.061)	0.353 (0.066)	8.220 (1.282)	29.300 (8.117)
	$\theta = 0.7$, MNAR					
	PE	MSE	pAUC	F_1	TP	FP
TL	1.927 (0.536)	3.708 (1.036)	0.945 (0.039)	0.426 (0.064)	9.700 (0.505)	27.000 (8.732)
NL	10.107 (3.407)	9.440 (1.697)	0.688 (0.080)	0.286 (0.089)	5.280 (1.371)	22.620 (6.648)
CoCo	6.750 (2.215)	12.217 (4.246)	0.660 (0.072)	0.286 (0.082)	5.080 (1.226)	21.100 (5.486)
NCL	7.116 (1.667)	10.195 (2.007)	0.472 (0.073)	0.306 (0.093)	4.820 (1.466)	17.400 (7.741)
LPD-E-1	6.930 (2.367)	10.865 (2.421)	0.759 (0.082)	0.251 (0.064)	6.320 (1.362)	35.020 (7.878)
LPD-F-1	10.617 (5.046)	13.477 (4.554)	0.759 (0.084)	0.234 (0.067)	6.500 (1.821)	39.740 (11.940)
LPD-L-1	5.384 (1.176)	9.481 (1.686)	0.756 (0.083)	0.255 (0.063)	6.320 (1.504)	33.760 (7.224)
LPD-S-1	5.351 (1.223)	9.491 (1.843)	0.760 (0.082)	0.260 (0.066)	6.300 (1.432)	32.740 (6.452)

Table S3: Sensitivity analysis for $\theta = 0.7, 0.9$ and different missing mechanisms. Each performance measures are averaged over $R = 100$ repetitions (standard deviation in parenthesis).

S4.4 Timing

For both LPD and CoCo, the first step is to modify the estimate of covariance matrix to be PD, and the second step is to solve the penalized regression (e.g. (2.7) for LPD) with the

modified estimate. We separately measure the time elapsed for the steps, positive definite modification (PD) and lasso regression (Lasso), which is shown in Table S4. We use ℓ_∞-norm for LPD since the other norms take roughly the same amount of time. In this experiment, we fix the tuning parameter λ at the middle of endpoints of search grids.

In step "Lasso", both methods solve a strictly convex quadratic programming problem, which is very fast. It took less than a second for both methods and does not have much difference between the two methods. However, in step "PD", CoCo takes much longer than LPD, for example, around 50 seconds when $p = 1000$ compared to 0.128 seconds for LPD. Thus, "PD" step is dominant in the whole process of CoCo, while it does not scale up the total time of LPD.

Method	Step	p =200	p =500	p =1000
CoCo	Lasso	0.146	0.507	0.538
CoCo	PD	0.174	3.849	49.587
LPD	Lasso	0.103	0.382	0.515
LPD	PD	0.004	0.033	0.128

Table S4: The elapsed times (unit: second) for (1) lasso estimation at a fixed tuning parameter (Lasso) and (2) positive definite modification (PD). We average over 100 independent datasets generated under $n = 200$, $s = 0.05$, and p varying over $200, 500, 1000$.

S4.5 Empirical analysis of the rate of convergence

We investigate the empirical scaling of the estimation error, as suggested by one of the reviewers. We calculate the mean squared error (MSE) $||\widehat{\beta} - \beta^*||_2^2$ while varying the sample size and the dimension: $n, p = 100, 200, \ldots, 500$. We also consider different covariance structures: the compound symmetry $\mathbf{\Sigma} = \rho 1 1^\top + (1 - \rho)\mathbf{I}$, the autoregressive structure $\mathbf{\Sigma} = (\rho^{|i-j|})_{1 \le i,j \le p}$, and the independent structure $\mathbf{\Sigma} = \mathbf{I}$. Here, we fix $\rho = 0.5$. We generate

10% missing values under the MCAR mechanism. The regression coefficient β^* is set to have ones in the first 10 entries and zeros elsewhere.

Figure S2 shows the results where the logarithm of MSE (y-axis) is plotted against $\log(\sqrt{\log p/n})$ (x-axis). Regardless of the models ("CoCo", "LPD", and "TL"), the error curves align together for different dimensions, meaning that the logarithm of MSE scales with $\log(\sqrt{\log p/n})$ up to an additive constant independent of n and p. This implies the convergence rate of LPD does not depend on the trace term, and thus is close to $O_p(\sqrt{\log p/n})$. In other words, the proposed method does not require as many samples as claimed in our theory for accurate estimation.

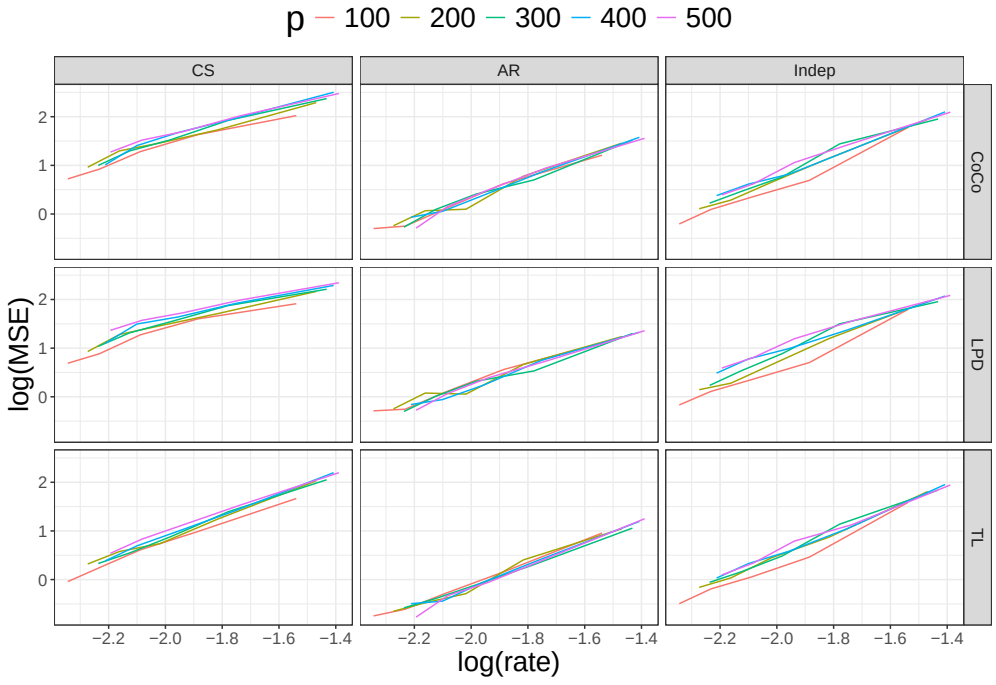

Figure S2: Estimation error against the aimed rate based on 30 replications. "CoCo" is CoColasso by Datta and Zou (2017), "LPD" is the proposed method, and "TL" is the lasso using the complete data.

Bibliography

Datta, A. and H. Zou (2017). CoCoLasso for high-dimensional error-in-variables regression. *The Annals of Statistics 45*(6), 2400 – 2426.

Kolar, M. and E. P. Xing (2012). Estimating sparse precision matrices from data with missing values. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, USA, pp. 635–642. Omnipress.

Lee, J. D., Y. Sun, and J. E. Taylor (2015). On model selection consistency of regularized M-estimators. *Electronic Journal of Statistics 9*(1), 608 – 642.

Lounici, K. (2014, 08). High-dimensional covariance matrix estimation with missing observations. *Bernoulli 20*(3), 1029–1058.

Mai, Q., H. Zou, and M. Yuan (2012, 12). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika 99*(1), 29–42.

Park, S., X. Wang, and J. Lim (2023). Sparse HansonWright inequality for a bilinear form of sub-gaussian variables. *Stat 12*(1), e539.

van Buuren, S. (2018). *Flexible imputation of missing data*. CRC Press.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *Journal of Machine Learning Research 7*(90), 2541–2563.