Supplement to "Statistical Inference for high dimensional regression with proxy data"

A Proofs

A.1 Proofs in Section 3

For a matrix $A \in \mathbb{R}^{n_1 \times n_2}$, let $\|A\|_S$ denote the spectral norm of A.

Proof for Theorem 1. The following oracle inequality holds for the two-sample Lasso:

$$\frac{1}{2}(\hat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})^{\intercal}\widetilde{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}) \leq |(\hat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})(\widehat{\boldsymbol{S}} - \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\beta})| + \lambda^{(ts)}\|\boldsymbol{\beta}\|_1 - \lambda^{(ts)}\|\hat{\boldsymbol{\beta}}^{(ts)}\|_1.$$

Define an event

$$E_0 = \left\{(\boldsymbol{u})^{\intercal}\widetilde{\boldsymbol{\Sigma}}\boldsymbol{u} \geq C\|\boldsymbol{u}\|_2^2 - \frac{\log p}{\tilde{n}}\|\boldsymbol{u}\|_1^2, \quad \|\widehat{\boldsymbol{S}} - \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\beta}\|_\infty \leq \frac{\lambda^{(ts)}}{2}\right\}.$$

In event E_0, we have

$$\frac{C}{2}\|\hat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})\|_2^2 \leq +\frac{3}{2}\lambda^{(ts)}\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_S\|_1 - \frac{\lambda^{(ts)}}{2}\|(\hat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})_{S^c}\|_1 + \frac{4\log p}{\tilde{n}}\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_S\|_1^2.$$

Standard analysis lead to

$$\begin{aligned}
&\|\widetilde{\boldsymbol{\Sigma}}^{1/2}(\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})\|_2^2 \vee \|\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}\|_2^2 \leq Cs(\lambda^{(ts)})^2 \\
&\|\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}\|_1 \leq Cs\lambda^{(ts)}
\end{aligned}$$

under the condition $\tilde{n} \gg s\log p$.

We are left to prove $\mathbb{P}(E_0) \to 1$ for $\lambda^{(ts)}$ given in Theorem 1. The first statement in E_0 holds with probability at least $1 - \exp(-c_1\tilde{n})$ by Theorem 1 in Raskutti et al. (2010). For the second statement in E_0,

$$\|\widehat{\boldsymbol{S}} - \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\beta}\|_\infty \leq \|\widehat{\boldsymbol{S}} - \mathbb{E}[\widehat{\boldsymbol{S}}]\|_\infty + \|(\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}})\boldsymbol{\beta}\|_\infty.$$

By the Bernstein's inequality for sub-exponential random variables,

$$\mathbb{P}(\|\widehat{\boldsymbol{S}} - \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\beta}\|_\infty \geq \lambda^{(ts)}) \leq C\exp(-c\log p)$$

when $n \wedge \tilde{n} \gg \log p$ for large enough c_1 in $\lambda^{(ts)}$.

□

Proof of Theorem 3. Let $\boldsymbol{w}_j = \boldsymbol{\Omega}_{\cdot,j}$.

$$\begin{aligned}
\hat{\beta}_j^{(ts-db)} - \beta_j &= \hat{\beta}_j^{(ts)} - \beta_j + \tilde{\boldsymbol{w}}_j^\intercal \widehat{\boldsymbol{S}} - \tilde{\boldsymbol{w}}_j^\intercal \widetilde{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\beta}}^{(ts)} \\
&= \hat{\beta}_j^{(ts)} - \beta_j + \tilde{\boldsymbol{w}}_j^\intercal (\boldsymbol{X}^\intercal \boldsymbol{\epsilon}/n + \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}) - \tilde{\boldsymbol{w}}_j^\intercal \widetilde{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\beta}}^{(ts)} \\
&= \tilde{\boldsymbol{w}}_j^\intercal \boldsymbol{X}^\intercal \boldsymbol{\epsilon}/n + \boldsymbol{e}_j^\intercal (\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}) + \tilde{\boldsymbol{w}}_j^\intercal \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \tilde{\boldsymbol{w}}_j^\intercal \widetilde{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\beta}}^{(ts)} \\
&= \tilde{\boldsymbol{w}}_j^\intercal \boldsymbol{X}^\intercal \boldsymbol{\epsilon}/n + \boldsymbol{e}_j^\intercal (\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}) + \tilde{\boldsymbol{w}}_j^\intercal \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \tilde{\boldsymbol{w}}_j^\intercal \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\beta} + \tilde{\boldsymbol{w}}_j^\intercal \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \tilde{\boldsymbol{w}}_j^\intercal \widetilde{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\beta}}^{(ts)} \\
&= \tilde{\boldsymbol{w}}_j^\intercal (\boldsymbol{X}^\intercal \boldsymbol{y}/n - \boldsymbol{\Sigma}\boldsymbol{\beta}) + \underbrace{(\boldsymbol{e}_j - \widetilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{w}}_j)^\intercal (\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})}_{R_1} + \tilde{\boldsymbol{w}}_j^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta} - \tilde{\boldsymbol{w}}_j^\intercal \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\beta} \\
&= \underbrace{\boldsymbol{w}_j^\intercal (\boldsymbol{X}^\intercal \boldsymbol{y}/n - \boldsymbol{\Sigma}\boldsymbol{\beta}) + \boldsymbol{w}_j^\intercal (\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}})\boldsymbol{\beta}}_{R_0} + R_1 \\
&\quad + \underbrace{(\tilde{\boldsymbol{w}}_j - \boldsymbol{w}_j)^\intercal (\boldsymbol{X}^\intercal \boldsymbol{y}/n - \boldsymbol{\Sigma}\boldsymbol{\beta}) + (\tilde{\boldsymbol{w}}_j - \boldsymbol{w}_j)^\intercal (\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}})\boldsymbol{\beta}}_{R_2}.
\end{aligned}$$

For R_0, $\boldsymbol{w}_j^\intercal (\boldsymbol{X}^\intercal \boldsymbol{y}/n - \boldsymbol{\Sigma}\boldsymbol{\beta})$ and $\boldsymbol{w}_j^\intercal (\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}})\boldsymbol{\beta}$ are independent. Using the formula for the higher moments of Gaussian distribution, we know that

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{w}_j^\intercal (\boldsymbol{X}^\intercal \boldsymbol{y}/n - \boldsymbol{\Sigma}\boldsymbol{\beta})) &= \frac{1}{n}(\mathbb{E}[(\boldsymbol{x}_i^\intercal \boldsymbol{w}_j y_i)^2] - (\boldsymbol{w}_j^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta})^2) \\
&= \frac{\boldsymbol{w}_j^\intercal \boldsymbol{\Sigma}\boldsymbol{w}_j M + (\boldsymbol{w}_j^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta})^2}{n} \\
\mathrm{Var}(\boldsymbol{w}_j^\intercal (\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}})\boldsymbol{\beta}) &= \frac{\boldsymbol{w}_j^\intercal \boldsymbol{\Sigma}\boldsymbol{w}_j \boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta} + (\boldsymbol{w}_j^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta})^2}{\tilde{n}}
\end{aligned}$$

Notice that $\boldsymbol{w}_j^\intercal \boldsymbol{\Sigma}\boldsymbol{w}_j = \Omega_{j,j}$ and $\boldsymbol{w}_j^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta} = \boldsymbol{\Omega}_{j,\cdot}\boldsymbol{\Sigma}\boldsymbol{\beta} = \beta_j$, it is easy to show that

$$\frac{R_0}{(V_j^{(ts)})^{1/2}} \xrightarrow{D} N(0,1).$$

for

$$V_j^{(ts)} = \Omega_{j,j}(\frac{M}{n} + \frac{1}{\tilde{n}}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^2) + \beta_j^2(\frac{1}{n} + \frac{1}{\tilde{n}}) = \frac{\Omega_{j,j}}{n}\gamma_{n,\tilde{n}} + \beta_j^2(\frac{1}{n} + \frac{1}{\tilde{n}}).$$

We first show that (3.3) is feasible with high probability. Using the sub-exponential property of $(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{w}_j$, we have if $\lambda_j \geq c_j\sqrt{\log p/\tilde{n}}$ for $c_j \geq c_1\sqrt{\boldsymbol{\Sigma}_{j,j}\Omega_{j,j}}$ with some large enough constant c_1, then

$$\mathbb{P}(\|\widetilde{\boldsymbol{\Sigma}}_j - \boldsymbol{e}_j\|_\infty \leq \lambda_j) \geq 1 - \exp\{-c_2 \log p\}.$$

For R_1, we have

$$|R_1| \leq \|\boldsymbol{e}_j - \widetilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{w}}_j\|_\infty \|\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}\|_1 \leq \lambda_j \|\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}\|_1.$$

For R_2, we can first use the results of Dantzig selector to show that

$$\mathbb{P}\left(\|\tilde{\boldsymbol{w}}_j - \boldsymbol{w}_j\|_1 \geq C s_j \lambda_j\right) \leq \exp\{-c_1 \log p\}$$

given that $\tilde{n} \gg s_j \log p$. By the sub-exponential properties of $\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\beta}$ and $\boldsymbol{X}^\intercal \boldsymbol{y}/n$, we have

$$|R_2| \leq \|\tilde{\boldsymbol{w}}_j - \boldsymbol{w}_j\|_1 (\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta}\|_\infty + \|\boldsymbol{X}^\intercal \boldsymbol{y}/n - \mathbb{E}[\boldsymbol{X}^\intercal \boldsymbol{y}/n]\|_\infty) = O_P\left(s_j \lambda^{(ts)} \lambda_j\right).$$

To summarize,

$$\hat{\beta}_j^{(ts-db)} - \beta_j = R_0 + O_P\left((s + s_j)\lambda^{(ts)}\lambda_j\right),$$

where $(V_j^{(ts)})^{-1/2} R_0 \xrightarrow{D} N(0,1)$.

To obtain (3.4), we need

$$(s + s_j)\lambda^{(ts)}\lambda_j = o(\sqrt{V_j^{(ts)}})$$

and a sufficient condition is

$$(s + s_j)\gamma_{n,\tilde{n}}^{1/2} \frac{\log p}{\sqrt{n\tilde{n}}} \ll \sqrt{\frac{\gamma_{n,\tilde{n}}}{n}}.$$

The proof is complete now. □

Proof of Lemma 1. Notice that $\Omega_{j,j} = \boldsymbol{w}_j^\intercal \boldsymbol{\Sigma} \boldsymbol{w}_j$. It is easy to show that

$$\begin{aligned}
|\tilde{\boldsymbol{w}}_j^\intercal \widetilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{w}}_j - \Omega_{j,j}| &\leq 2|(\tilde{\boldsymbol{w}}_j - \boldsymbol{w}_j)^\intercal \widetilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{w}}_j| + |(\tilde{\boldsymbol{w}}_j - \boldsymbol{w}_j)^\intercal \widetilde{\boldsymbol{\Sigma}} (\tilde{\boldsymbol{w}}_j - \boldsymbol{w}_j)| \\
&\leq 2\|\tilde{\boldsymbol{w}}_j - \boldsymbol{w}_j\|_1 \lambda_j^{(ts)} + |(\tilde{\boldsymbol{w}}_j - \boldsymbol{w}_j)^\intercal \widetilde{\boldsymbol{\Sigma}} (\tilde{\boldsymbol{w}}_j - \boldsymbol{w}_j)| = O_P(\frac{s_j \log p}{\tilde{n}}).
\end{aligned}$$

Morevoer, $\|\boldsymbol{y}\|_2^2/n = \mathbb{E}[y_i^2] + O_P(\{M/n\}^{1/2})$. Finally,

$$\begin{aligned}
&2(\widehat{\boldsymbol{\beta}}^{(ts)})^\intercal \widehat{\boldsymbol{S}} - (\widehat{\boldsymbol{\beta}}^{(ts)})^\intercal \widetilde{\boldsymbol{\Sigma}} (\widehat{\boldsymbol{\beta}}^{(ts)}) - \boldsymbol{\beta}^\intercal \boldsymbol{\Sigma} \boldsymbol{\beta} \\
&= 2\boldsymbol{\beta}^\intercal \widehat{\boldsymbol{S}} - (\widehat{\boldsymbol{\beta}}^{(ts)})^\intercal \widetilde{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}^\intercal \boldsymbol{\Sigma} \boldsymbol{\beta} + 2(\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})^\intercal \widehat{\boldsymbol{S}} \\
&= \underbrace{2\boldsymbol{\beta}^\intercal \widehat{\boldsymbol{S}} - \boldsymbol{\beta}^\intercal \boldsymbol{\Sigma} \boldsymbol{\beta} - \boldsymbol{\beta}^\intercal \widetilde{\boldsymbol{\Sigma}} \boldsymbol{\beta}}_{F_1} \\
&\quad + \underbrace{2(\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})^\intercal \widehat{\boldsymbol{S}} - 2(\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})^\intercal \widetilde{\boldsymbol{\Sigma}} \boldsymbol{\beta}}_{F_2} - \underbrace{(\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})^\intercal \widetilde{\boldsymbol{\Sigma}} (\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})}_{F_3},
\end{aligned}$$

where

$$F_2 = O_P(\gamma_{n,\tilde{n}} \frac{s \log p}{n}), \quad F_3 = O_P(\gamma_{n,\tilde{n}} \frac{s \log p}{n}).$$

For F_1, using the Gaussian property of $\boldsymbol{X}$, we have

$$\begin{aligned} F_1 &= O_P(\sqrt{\frac{var(\boldsymbol{x}_i^\intercal \boldsymbol{\beta} + y_i)}{n}}) + O_P(\sqrt{\frac{var(\tilde{\boldsymbol{x}}_i^\intercal \boldsymbol{\beta})}{\tilde{n}}}) \\ &= O_P(\sqrt{\frac{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^4 + M\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^2}{n}}) + O_P(\sqrt{\frac{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^4}{\tilde{n}}}) \\ &= O_P(\frac{\gamma_{n,\tilde{n}}}{\sqrt{n}}). \end{aligned}$$

Hence,

$$n\left|\tilde{\boldsymbol{w}}_j^\intercal \widetilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{w}}_j(\frac{\|\boldsymbol{y}\|_2^2}{n^2} + \frac{2(\widehat{\boldsymbol{\beta}}^{(ts)})^\intercal \widehat{\boldsymbol{S}} - (\widehat{\boldsymbol{\beta}}^{(ts)})^\intercal \widetilde{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}}^{(ts)}}{\tilde{n}}) - \frac{\Omega_{j,j}\gamma_{n,\tilde{n}}}{n}\right| = O_P(\gamma_{n,\tilde{n}}\frac{s_j\log p}{\tilde{n}} + \gamma_{n,\tilde{n}}\frac{s\log p}{n} + \frac{\gamma_{n,\tilde{n}}}{\sqrt{n}}).$$

By Theorem 3,

$$|(\hat{\beta}_j^{(ts-db)})^2 - \beta_j^2|(1+\frac{n}{\tilde{n}}) \le |\hat{\beta}_j^{(ts-db)} - \beta_j|\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2(1+\frac{n}{\tilde{n}})$$

By some simple algebra,

$$\begin{aligned} \frac{|\widehat{V}_j^{(ts)} - V_j^{(ts)}|}{V_j^{(ts)}} &= O_P(\frac{\gamma_{n,\tilde{n}}\frac{s_j\log p}{\tilde{n}} + \gamma_{n,\tilde{n}}\frac{s\log p}{n} + \frac{\gamma_{n,\tilde{n}}}{\sqrt{n}}}{nV_j^{(ts)}}) + O_P(\frac{|\hat{\beta}_j^{(ts-db)} - \beta_j|\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2(1+\frac{n}{\tilde{n}})}{V_j^{(ts)}}) \\ &= O_P(\gamma_{n,\tilde{n}}\frac{s_j\log p}{\tilde{n}} + \gamma_{n,\tilde{n}}\frac{s\log p}{n} + \frac{\gamma_{n,\tilde{n}}}{\sqrt{n}}) + O_P(\frac{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2(1+\frac{n}{\tilde{n}})}{\sqrt{V_j^{(ts)}}}) \\ &= O_P(\gamma_{n,\tilde{n}}\frac{s_j\log p}{\tilde{n}} + \gamma_{n,\tilde{n}}\frac{s\log p}{n} + \frac{\gamma_{n,\tilde{n}}}{\sqrt{n}}) + O_P(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{\tilde{n}}}). \end{aligned}$$

□

A.2 Proof of minimax lower bound

Proof of Theorem 2. Recall that $\mathcal{Z} = \{\mathcal{Z}_1, \mathcal{Z}_2\} = \{\widehat{\boldsymbol{S}}, \widetilde{\boldsymbol{\Sigma}}\}$. Let $KL(f_1(\mathcal{Z}), f_2(\mathcal{Z})) = \int \log(f_1(\mathcal{Z})/f_2(\mathcal{Z})f_1(\mathcal{Z}$ denote the KL-divergence between $f_1(\mathcal{Z})$ and $f_2(\mathcal{Z})$. As $\mathcal{Z}_1$ is independent of $\mathcal{Z}_2$,

$$KL(f_1(\mathcal{Z}), f_2(\mathcal{Z})) \le KL(f_1(\mathcal{Z}_1), f_2(\mathcal{Z}_1)) + KL(f_1(\mathcal{Z}_2), f_2(\mathcal{Z}_2)).$$

Next, we consider a reduction of the parameter space. Let $\mathcal{B}_0(s) = \{\boldsymbol{u} \in \mathbb{R}^p : \|u\|_0 \le s\}$. Define a η-packing of $\mathcal{B}_0(s)$ in metric ρ as $\{\boldsymbol{b}^{(1)}, \ldots, \boldsymbol{b}^{(M)}\} \subseteq \mathcal{B}_0(s)$ such that $\rho(\boldsymbol{b}^{(i)}, \boldsymbol{b}^{(j)}) > \eta$ for all $i \neq j$. Denote the η-packing of a set $\mathcal{S}$ in metric ρ as $M(\eta, s, \rho)$. Let M denote the cardinality of $M(\eta, s, \rho)$. This is the same definition as in Definition 1 of Raskutti et al. (2011). Then

$$\inf_{\hat{\boldsymbol{\beta}} \in \mathcal{F}(\mathcal{Z})} \sup_{\boldsymbol{\beta} \in \Xi(s, M_0, \sigma_0^2)} \mathbb{P}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \geq \eta) \geq \inf_{\hat{\boldsymbol{\beta}} \in \mathcal{F}(\mathcal{Z})} \sup_{\boldsymbol{\beta} \in \kappa M(\eta, s, \|\cdot\|_2)} \mathbb{P}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \geq \eta), \tag{A.1}$$

where κ is a positive constant based on our choice such that $s\kappa^2 \leq M_0$. Let $\hat{\psi}$ be the test function depending on $\mathcal{Z}$ such that $\hat{\psi} = j$ if $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(j)}\|_2 < \min_{m \neq j} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(m)}\|_2$ for $\boldsymbol{\beta}^{(j)}$ and $\boldsymbol{\beta}^{(m)}$ in $\kappa M(\eta, s, \|\cdot\|_2)$. Notice that $\hat{\psi}$ is unique for a given $\hat{\boldsymbol{\beta}}$. Then

$$\begin{aligned} \inf_{\hat{\boldsymbol{\beta}} \in \mathcal{F}(\mathcal{Z})} \sup_{\boldsymbol{\beta} \in \kappa M(\eta, s, \|\cdot\|_2)} \mathbb{P}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \geq \eta) &\geq \max_{1 \leq m \leq M} \mathbb{P}_m(\hat{\psi} \neq m) \\ &\geq 1 - \frac{\xi + \log 2}{\log M}, \end{aligned} \tag{A.2}$$

where $\xi = \max_{j,m \leq M} KL(f_j(\mathcal{Z}), f_m(\mathcal{Z}))$. The last step is due to Fano's inequality.

We first construct a packing of $\mathcal{B}_0(s)$. Define

$$\mathcal{H}(s) = \{z \in \{-1, 0, 1\}^p : \|z\|_0 = s\}.$$

By Lemma 4 in Raskutti et al. (2011), there exists a subset $\tilde{\mathcal{H}}(k) \subset \mathcal{H}(k)$ such that $\rho_H(z, z') = \sum_{j=1}^p \mathbb{1}(z_j \neq z'_j) \geq k/2$ for any $z \neq z' \in \tilde{\mathcal{H}}(k)$ and the cardinality of $\tilde{\mathcal{H}}(k) \geq \exp\{\frac{k}{2} \log \frac{p-k}{k/2}\}$. Hence, $\|z_j - z'_j\|_2 \geq \sqrt{k/2}$ for any $z \neq z' \in \tilde{\mathcal{H}}(k)$. Hence, $\eta\sqrt{\frac{2}{k}}\tilde{\mathcal{H}}(k)$ is a η-packing of $\mathcal{B}_0(k)$ in ℓ_2-norm. Let $\boldsymbol{\delta}^{(1)}, \ldots, \boldsymbol{\delta}^{(M)} \in \eta\sqrt{\frac{2}{s-1}}\tilde{\mathcal{H}}(s-1)$ be a η-packing of $\mathcal{B}_0(s-1)$ for $\eta^2 = c_\eta s \log p/(2\tilde{n})$ for some constant $c_\eta > 0$. That is, $\boldsymbol{\delta}^{(1)}, \ldots, \boldsymbol{\delta}^{(M)} \in c_\eta\sqrt{\frac{\log p}{\tilde{n}}}\tilde{\mathcal{H}}(s-1)$. The cardinality $M \geq \exp\{\frac{s-1}{2} \log \frac{p-s}{(s-1)/2}\}$.

For some constant $0 < c_1 \leq 1/3$, let

$$\bar{M} = \frac{M_0}{2} + \sigma_0^2 \quad \text{and} \quad \rho = \begin{cases} (1 + c_1)\sqrt{\frac{M_0}{2}} + c_1\sqrt{\frac{\bar{M}\tilde{n}}{n}} & \text{if } \sqrt{\frac{\bar{M}\tilde{n}}{n}} \geq \sqrt{\frac{M_0}{2}} \\ c_1\sqrt{\frac{\bar{M}\tilde{n}}{n}} & \text{if } \sqrt{\frac{\bar{M}\tilde{n}}{n}} < \sqrt{\frac{M_0}{2}}. \end{cases}$$

Notice that

$$\rho \leq 2\sqrt{\frac{\bar{M}\tilde{n}}{n}} \quad \text{and} \quad |\rho - \sqrt{\frac{M_0}{2}}| \geq c_1(\sqrt{\frac{M_0}{2}} + \sqrt{\frac{\bar{M}\tilde{n}}{n}}). \tag{A.3}$$

We specify f_i, $i = 1, \ldots, M$ with parameters

$$\boldsymbol{\Sigma}^{(i)} = \begin{pmatrix} 1 & (\boldsymbol{\delta}^{(i)})^\intercal \\ \boldsymbol{\delta}^{(i)} & I_{p-1} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta}^{(i)} = (\boldsymbol{\Sigma}^{(i)})^{-1}(\sqrt{\frac{M_0}{2}}, \rho(\boldsymbol{\delta}^{(i)})^\intercal)^\intercal = \begin{pmatrix} \frac{\sqrt{\frac{M_0}{2}} - \rho\|\boldsymbol{\delta}^{(i)}\|_2^2}{1 - \|\boldsymbol{\delta}^{(i)}\|_2^2} \\ \frac{\rho - \sqrt{\frac{M_0}{2}}}{1 - \|\boldsymbol{\delta}^{(i)}\|_2^2}\boldsymbol{\delta}^{(i)} \end{pmatrix}$$

$$(\sigma^{(i)})^2 = \bar{M} - (\boldsymbol{\beta}^{(i)})^\intercal \boldsymbol{\Sigma}^{(i)} \boldsymbol{\beta}^{(i)} = \sigma_0^2 - \frac{(\rho - \sqrt{\frac{M_0}{2}})^2 \|\boldsymbol{\delta}^{(i)}\|_2^2}{1 - \|\boldsymbol{\delta}^{(i)}\|_2^2}. \tag{A.4}$$

We first verify that $\{\boldsymbol{\beta}^{(i)}, \boldsymbol{\Sigma}^{(i)}, (\sigma^{(i)})^2\}_{i=1}^M \subseteq \Xi(s, M_0, \sigma_0^2)$ with some small enough constants c_η and c_ρ. In fact, we can verify that

$$\max_{1\le i\le M} \|\boldsymbol{\beta}^{(i)}\|_0 \le s,\ \ \max_{1\le i\le M} (\boldsymbol{\beta}^{(i)})^\intercal \boldsymbol{\Sigma}^{(i)} \boldsymbol{\beta}^{(i)} = M_0/2 + \max_{1\le i\le M} \frac{(\rho - \sqrt{\frac{M_0}{2}})^2 \|\boldsymbol{\delta}^{(i)}\|_2^2}{1 - \|\boldsymbol{\delta}^{(i)}\|_2^2},$$
$$\max_{1\le i\le M} \Lambda_{\max}(\boldsymbol{\Sigma}^{(i)}) = 1,\ \ \min_{1\le i\le M} \Lambda_{\min}(\boldsymbol{\Sigma}^{(i)}) \ge 1 - \max_{1\le i\le M} \|\boldsymbol{\delta}^{(i)}\|_2^2,$$
$$0 < (\sigma^{(i)})^2 = \sigma_0^2 - \frac{(\rho - \sqrt{\frac{M_0}{2}})^2 \|\boldsymbol{\delta}^{(i)}\|_2^2}{1 - \|\boldsymbol{\delta}^{(i)}\|_2^2} \le \sigma_0^2.$$

Hence, for small enough c_0 and c_1 such that $(\rho - \sqrt{\frac{M_0}{2}})^2 \|\boldsymbol{\delta}^{(i)}\|_2^2 \le c_1 \min\{M_0, \sigma_0^2\}$ and $\|\boldsymbol{\delta}^{(i)}\|_2^2 \le c_1$, $\{\boldsymbol{\beta}^{(i)}, \boldsymbol{\Sigma}^{(i)}, \sigma_i^2\}_{i=0}^K \subseteq \Xi(s, M_0, \sigma_0^2)$. It suffices to require

$$\frac{\bar{M} s \log p}{n} + \frac{M_0 s \log p}{\tilde{n}} \le c_1 \min\{\sigma_0^2, M_0\} \text{ and } \frac{s \log p}{\tilde{n} \wedge n} \le c_2 \tag{A.5}$$

for small enough constants c_1 and c_2. Moreover, $\boldsymbol{\beta}_{-1}^{(i)} \in 2c_\eta \sqrt{\frac{\log p}{\tilde{n}}} (\rho - \sqrt{\frac{M_0}{2}}) \tilde{\mathcal{H}}(s-1)$. For

$$\eta = 4c_\eta^2 \frac{s \log p}{\tilde{n}} (\rho - \sqrt{\frac{M_0}{2}})^2 \ge c_1 (\frac{\bar{M} s \log p}{n} + \frac{M_0 s \log p}{\tilde{n}}),$$

by (A.2),

$$\inf_{\hat{\boldsymbol{\beta}} \in \mathcal{F}(\mathcal{Z})} \sup_{\Xi(s, M_0, \sigma^2)} \mathbb{P}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \ge \eta) \ge \inf_{\hat{\boldsymbol{\beta}} \in \mathcal{F}(\mathcal{Z})} \sup_{\boldsymbol{\beta}_{-1} \in 2c_\eta \sqrt{\frac{\log p}{\tilde{n}}} (\rho - \sqrt{\frac{M_0}{2}}) \tilde{\mathcal{H}}(s-1)} \mathbb{P}(\|\hat{\boldsymbol{\beta}}_{-1} - \boldsymbol{\beta}_{-1}\|_2 \ge \eta)$$
$$\ge 1 - \frac{\xi + \log 2}{\log M}.$$

We are left to find

$$\max_{i,j\le M} KL(f_i(\mathcal{Z}), f_j(\mathcal{Z})) \le \max_{i,j\le M} KL(f_i(\mathcal{Z}_1), f_j(\mathcal{Z}_1)) + \max_{i,j\le M} KL(f_i(\mathcal{Z}_2), f_j(\mathcal{Z}_2)).$$

First,

$$KL(f_i(\mathcal{Z}_2), f_j(\mathcal{Z}_2)) = \frac{\tilde{n}}{2} \{\mathrm{Tr}(\{\boldsymbol{\Sigma}^{(j)}\}^{-1} \boldsymbol{\Sigma}^{(i)}) - \log det(\{\boldsymbol{\Sigma}^{(j)}\}^{-1} \boldsymbol{\Sigma}^{(i)}) - p\},$$

where

$$\{\boldsymbol{\Sigma}^{(j)}\}^{-1} = \begin{pmatrix} \frac{1}{1-\|\boldsymbol{\delta}^{(j)}\|_2^2} & -\frac{(\boldsymbol{\delta}^{(j)})^\intercal}{1-\|\boldsymbol{\delta}^{(j)}\|_2^2} \\ -\frac{\boldsymbol{\delta}^{(j)}}{1-\|\boldsymbol{\delta}^{(j)}\|_2^2} & I_{p-1} + \frac{\boldsymbol{\delta}^{(j)}\boldsymbol{\delta}^{(j)})^\intercal}{1-\|\boldsymbol{\delta}^{(j)}\|_2^2} \end{pmatrix}$$

$$\{\boldsymbol{\Sigma}^{(j)}\}^{-1}\boldsymbol{\Sigma}^{(i)} = \begin{pmatrix} \frac{1-(\boldsymbol{\delta}^{(j)})^\intercal\boldsymbol{\delta}^{(i)}}{1-\|\boldsymbol{\delta}^{(j)}\|_2^2} & \frac{(\boldsymbol{\delta}^{(i)}-\boldsymbol{\delta}^{(j)})}{1-\|\boldsymbol{\delta}^{(j)}\|_2^2} \\ \frac{(\boldsymbol{\delta}^{(i)}-\boldsymbol{\delta}^{(j)})-\boldsymbol{\delta}^{(i)}\|\boldsymbol{\delta}^{(j)}\|_2^2+(\boldsymbol{\delta}^{(j)})^\intercal\boldsymbol{\delta}^{(i)}\boldsymbol{\delta}^{(j)}}{1-\|\boldsymbol{\delta}^{(j)}\|_2^2} & I_{p-1} + \frac{\boldsymbol{\delta}^{(j)}(\boldsymbol{\delta}^{(i)}-\boldsymbol{\delta}^{(j)})^\intercal}{1-\|\boldsymbol{\delta}^{(j)}\|_2^2} \end{pmatrix}$$

$$= \begin{pmatrix} 1 - \frac{(\boldsymbol{\delta}^{(j)})^\intercal(\boldsymbol{\delta}^{(i)}-\boldsymbol{\delta}^{(j)})}{1-\|\boldsymbol{\delta}^{(j)}\|_2^2} & \frac{(\boldsymbol{\delta}^{(i)}-\boldsymbol{\delta}^{(j)})}{1-\|\boldsymbol{\delta}^{(j)}\|_2^2} \\ (\boldsymbol{\delta}^{(i)} - \boldsymbol{\delta}^{(j)}) + \frac{(\boldsymbol{\delta}^{(j)})^\intercal(\boldsymbol{\delta}^{(i)}-\boldsymbol{\delta}^{(j)})\boldsymbol{\delta}^{(j)}}{1-\|\boldsymbol{\delta}^{(j)}\|_2^2} & I_{p-1} - \frac{\boldsymbol{\delta}^{(j)}(\boldsymbol{\delta}^{(i)}-\boldsymbol{\delta}^{(j)})^\intercal}{1-\|\boldsymbol{\delta}^{(j)}\|_2^2} \end{pmatrix}$$

Hence,

$$\mathrm{Tr}(\{\boldsymbol{\Sigma}^{(j)}\}^{-1}\boldsymbol{\Sigma}^{(i)}) = p - 1 + \frac{1 - (\boldsymbol{\delta}^{(j)})^\intercal\boldsymbol{\delta}^{(i)} - (\boldsymbol{\delta}^{(j)})^\intercal(\boldsymbol{\delta}^{(i)} - \boldsymbol{\delta}^{(j)})}{1 - \|\boldsymbol{\delta}^{(j)}\|_2^2} = p + \frac{2(\boldsymbol{\delta}^{(j)} - \boldsymbol{\delta}^{(i)})^\intercal\boldsymbol{\delta}^{(j)}}{1 - \|\boldsymbol{\delta}^{(j)}\|_2^2}$$

$$det(\{\boldsymbol{\Sigma}^{(j)}\}^{-1}\boldsymbol{\Sigma}^{(i)}) = det(\boldsymbol{\Sigma}^{(i)})/det(\boldsymbol{\Sigma}^{(j)}) = 1 - \frac{\|\boldsymbol{\delta}^{(i)}\|_2^2 - \|\boldsymbol{\delta}^{(j)}\|_2^2}{1 - \|\boldsymbol{\delta}^{(j)}\|_2^2}.$$

Recall the inequality $1/(1-x) \leq \exp\{2x\}$ for $x \in [0, \log 2/2]$. Given that $|\|\boldsymbol{\delta}^{(i)}\|_2^2-\|\boldsymbol{\delta}^{(j)}\|_2^2| < \log 2/4$, we have

$$\max_{i,j\leq M} KL(f_i(\mathcal{Z}_2), f_j(\mathcal{Z}_2)) \leq 4\tilde{n} \max_{i,j\leq M} |(\boldsymbol{\delta}^{(j)} - \boldsymbol{\delta}^{(i)})^\intercal\boldsymbol{\delta}^{(j)}| + 4\tilde{n} \max_{i,j\leq M} \|\boldsymbol{\delta}^{(j)}\|_2^2 \leq Cs\log p \quad \text{(A.6)}$$

for some constant C depending on c_η.

Next, we bound $\max_{i,j\leq M} KL(f_i(\mathcal{Z}_1), f_j(\mathcal{Z}_1))$. One can easily calculate that

$$\mathbb{P}_j(\mathcal{Z}_1|\boldsymbol{y}) = N\left(\underbrace{\frac{\|\boldsymbol{y}\|^2}{n\mathbb{E}_j[y_i^2]}\boldsymbol{\Sigma}^{(j)}\boldsymbol{\beta}^{(j)}}_{\boldsymbol{\mu}_j}, \underbrace{\frac{\|\boldsymbol{y}\|^2}{n^2}(\boldsymbol{\Sigma}^{(j)} - \frac{\boldsymbol{\Sigma}^{(j)}\boldsymbol{\beta}^{(j)}(\boldsymbol{\beta}^{(j)})^\intercal\boldsymbol{\Sigma}^{(j)}}{\mathbb{E}_j[y_i^2]})}_{\boldsymbol{W}_j}\right), \quad \text{(A.7)}$$

where $\boldsymbol{W}_j$ is positive definite. Hence,

$$\begin{aligned} &KL(\mathbb{P}_i(\mathcal{Z}_1|\boldsymbol{y}), \mathbb{P}_j(\mathcal{Z}_1|\boldsymbol{y})) \\ &= \frac{1}{2}\{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\intercal\{\boldsymbol{W}_j\}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \mathrm{Tr}(\{\boldsymbol{W}_j\}^{-1}(\boldsymbol{W}_i - \boldsymbol{W}_j)) - \log det(\{\boldsymbol{W}_j\}^{-1}\boldsymbol{W}_i)\}. \end{aligned} \quad \text{(A.8)}$$

By (A.4), we have $\mathbb{E}_j[y_i^2] = M_0/2 + \sigma_0^2 = \bar{M}$ for any $1 \leq i \leq M$. We can calculate that

$$\boldsymbol{\mu}_i - \boldsymbol{\mu}_j = \frac{\|\boldsymbol{y}\|^2}{n\bar{M}}(0, \rho(\boldsymbol{\delta}^{(i)} - \boldsymbol{\delta}^{(j)})^\intercal)^\intercal.$$

$$\boldsymbol{W}_j = \frac{\|\boldsymbol{y}\|_2^2}{n^2}\begin{pmatrix} 1 - \frac{M_0}{2\bar{M}} & (1 - \frac{\rho\sqrt{M_0/2}}{\bar{M}})(\boldsymbol{\delta}^{(j)})^\intercal \\ (1 - \frac{\rho\sqrt{M_0/2}}{\bar{M}})\boldsymbol{\delta}^{(j)} & I_{p-1} - \frac{\rho^2}{\bar{M}}\boldsymbol{\delta}^{(j)}(\boldsymbol{\delta}^{(i)})^\intercal \end{pmatrix}$$

Hence,

$$\boldsymbol{W}_j^{-1} = \frac{n^2}{\|\boldsymbol{y}\|_2^2} \begin{pmatrix} a_{1,1} & -a_{1,1}\frac{1-\rho\sqrt{M_0/2}/\bar{M}}{1-\rho^2\|\boldsymbol{\delta}^{(j)}\|_2^2/M}(\boldsymbol{\delta}^{(j)})^\intercal \\ -a_{1,1}\frac{1-\rho\sqrt{M_0/2}/\bar{M}}{1-\rho^2\|\boldsymbol{\delta}^{(j)}\|_2^2/M}\boldsymbol{\delta}^{(j)} & I_{p-1} + a_{2,2}\boldsymbol{\delta}^{(j)}(\boldsymbol{\delta}^{(j)})^\intercal \end{pmatrix},$$

where

$$a_{1,1} = \{\frac{\sigma_0^2}{\bar{M}} - \frac{(1-\frac{\rho\sqrt{M_0/2}}{\bar{M}})^2\|\boldsymbol{\delta}^{(j)}\|_2^2}{1-\rho^2/\bar{M}\|\boldsymbol{\delta}^{(j)}\|_2^2}\}^{-1}$$

$$a_{2,2} = \frac{\frac{\rho^2}{\bar{M}} + \frac{\bar{M}}{\sigma_0^2}(1-\frac{\rho\sqrt{M_0/2}}{\bar{M}})^2}{1-\{\frac{\rho^2}{\bar{M}} + \frac{\bar{M}}{\sigma_0^2}(1-\frac{\rho\sqrt{M_0/2}}{\bar{M}})^2\}\|\boldsymbol{\delta}^{(j)}\|_2^2}.$$

Notice that (A.5) guarantees that

$$\{\frac{\rho^2}{\bar{M}} + \frac{\bar{M}}{\sigma_0^2}(1-\frac{\rho\sqrt{M_0/2}}{\bar{M}})^2\}\|\boldsymbol{\delta}^{(j)}\|_2^2 \le 1/2 \quad \text{and} \quad \rho^2\|\boldsymbol{\delta}^{(j)}\|_2^2 < \bar{M}/2.$$

Hence, we have

$$a_{2,2} \le \frac{2\rho^2}{\bar{M}} + \frac{2\bar{M}}{\sigma_0^2}(1-\frac{\rho\sqrt{M_0/2}}{\bar{M}})^2.$$

We can calculate that

$$(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\intercal\{\boldsymbol{W}_j\}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \le \frac{n^2}{\|\boldsymbol{y}\|_2^2}\cdot(\frac{\rho\|\boldsymbol{y}\|_2^2}{n\bar{M}})^2\{\|\boldsymbol{\delta}^{(i)} - \boldsymbol{\delta}^{(j)}\|_2^2 + a_{2,2}((\boldsymbol{\delta}^{(i)} - \boldsymbol{\delta}^{(j)})^\intercal\boldsymbol{\delta}^{(j)})^2\}$$

$$\le \frac{\rho^2\|\boldsymbol{y}\|_2^2}{\bar{M}^2}[\|\boldsymbol{\delta}^{(i)} - \boldsymbol{\delta}^{(j)}\|_2^2 + 2\{\frac{\rho^2}{\bar{M}} + \frac{\bar{M}}{\sigma_0^2}(1-\frac{\rho\sqrt{M_0/2}}{\bar{M}})^2\}((\boldsymbol{\delta}^{(i)} - \boldsymbol{\delta}^{(j)})^\intercal\boldsymbol{\delta}^{(j)})^2],$$

assuming (A.5).

Moreover,

$$det(\{\boldsymbol{W}_j\}^{-1}\boldsymbol{W}_i) = \frac{det(\boldsymbol{W}_i)}{det(\boldsymbol{W}_j)} = \frac{1-(\frac{\rho^2}{\bar{M}} + \frac{\bar{M}}{\sigma_0^2}(1-\frac{\rho\sqrt{M_0/2}}{\bar{M}})^2)\|\boldsymbol{\delta}^{(i)}\|_2^2}{1-(\frac{\rho^2}{\bar{M}} + \frac{\bar{M}}{\sigma_0^2}(1-\frac{\rho\sqrt{M_0/2}}{\bar{M}})^2)\|\boldsymbol{\delta}^{(j)}\|_2^2}$$

$$= 1 - \frac{(\frac{\rho^2}{\bar{M}} + \frac{\bar{M}}{\sigma_0^2}(1-\frac{\rho\sqrt{M_0/2}}{\bar{M}})^2)(\|\boldsymbol{\delta}^{(i)}\|_2^2 - \|\boldsymbol{\delta}^{(j)}\|_2^2)}{1-(\frac{\rho^2}{\bar{M}} + \frac{\bar{M}}{\sigma_0^2}(1-\frac{\rho\sqrt{M_0/2}}{\bar{M}})^2)\|\boldsymbol{\delta}^{(j)}\|_2^2}.$$

Recall that $1/(1-x) \le \exp\{2x\}$ for $x \in [0, \log 2/2]$. Hence, under (A.5),

$$\log det(\{\boldsymbol{W}_j\}^{-1}\boldsymbol{W}_i) \le 4(\frac{\rho^2}{\bar{M}} + (\frac{\rho^2}{\bar{M}} + \frac{\bar{M}}{\sigma_0^2}(1-\frac{\rho\sqrt{M_0/2}}{\bar{M}})^2)|\|\boldsymbol{\delta}^{(i)}\|_2^2 - \|\boldsymbol{\delta}^{(j)}\|_2^2|.$$

We can also calculate that

$$\begin{aligned}\mathrm{Tr}(\boldsymbol{W}_j^{-1}(\boldsymbol{W}_i-\boldsymbol{W}_j)) &\le 4a_{1,1}(1-\frac{\rho\sqrt{M_0/2}}{\bar{M}})^2|(\boldsymbol{\delta}^{(j)})^\intercal(\boldsymbol{\delta}^{(i)}-\boldsymbol{\delta}^{(j)})|+\frac{\rho^2}{\bar{M}}(\|\boldsymbol{\delta}^{(j)}\|_2^2+a_{2,2}\|\boldsymbol{\delta}^{(j)}\|_2^4)\\ &\le 4a_{1,1}(1-\frac{\rho\sqrt{M_0/2}}{\bar{M}})^2|(\boldsymbol{\delta}^{(j)})^\intercal(\boldsymbol{\delta}^{(i)}-\boldsymbol{\delta}^{(j)})|+\frac{\rho^2}{\bar{M}}\|\boldsymbol{\delta}^{(j)}\|_2^2(1+a_{2,2}\|\boldsymbol{\delta}^{(j)}\|_2^2).\end{aligned}$$

Under (A.5),

$$a_{1,1}\le\{\frac{\sigma_0^2/2}{\bar{M}-\rho^2\|\boldsymbol{\delta}^{(j)}\|_2^2}\}^{-1}\le\frac{4\bar{M}}{\sigma_0^2}.$$

We have

$$\begin{aligned}&\mathrm{Tr}(\boldsymbol{W}_j^{-1}(\boldsymbol{W}_i-\boldsymbol{W}_j))\\ &\le\frac{C\bar{M}}{\sigma_0^2}(1-\frac{\rho\sqrt{M_0/2}}{\bar{M}})^2\frac{s\log p}{\tilde{n}}+\frac{\rho^2}{\bar{M}}\frac{s\log p}{\tilde{n}}\{1+\frac{2\rho^2}{\bar{M}}\frac{s\log p}{\tilde{n}}+\frac{2\bar{M}}{\sigma_0^2}(1-\frac{\rho\sqrt{M_0/2}}{\bar{M}})^2\frac{s\log p}{\tilde{n}}\}\\ &\le\frac{C\bar{M}}{\sigma_0^2}(1+\frac{\tilde{n}M_0/2}{n\bar{M}})\frac{s\log p}{\tilde{n}}+\frac{Cs\log p}{n}(1+\frac{s\log p}{n}+\frac{\bar{M}s\log p}{\sigma_0^2\tilde{n}}+\frac{\bar{M}_0s\log p}{\sigma_0^2 n})\\ &\le\frac{C\bar{M}s\log p}{\sigma_0^2\tilde{n}}(1+\frac{s\log p}{n})+\frac{CM_0s\log p}{\sigma_0^2 n}\frac{s\log p}{n}+\frac{s\log p}{n}\le C_3 s\log p,\end{aligned}$$

where the last step is due to (A.5).

Tedious calculations lead to

$$\begin{aligned}&KL(\mathbb{P}_i(\mathcal{Z}_1|\boldsymbol{y}),\mathbb{P}_j(\mathcal{Z}_1|\boldsymbol{y}))\le\frac{\rho^2\|\boldsymbol{y}\|^2}{\bar{M}^2}[\|\boldsymbol{\delta}^{(i)}-\boldsymbol{\delta}^{(j)}\|_2^2+2\{\frac{\rho^2}{\bar{M}}+\frac{\bar{M}}{\sigma_0^2}(1-\frac{\rho\sqrt{M_0/2}}{\bar{M}})^2\}((\boldsymbol{\delta}^{(i)}-\boldsymbol{\delta}^{(j)})^\intercal\boldsymbol{\delta}^{(j)})^2]\\ &\quad+c_2\{\frac{\rho^2}{\bar{M}}+\frac{\bar{M}}{\sigma_0^2}(1-\frac{\rho\sqrt{M_0/2}}{\bar{M}})^2\}|\|\boldsymbol{\delta}^{(i)}\|_2^2-\|\boldsymbol{\delta}^{(j)}\|_2^2|+C_3s\log p\\ &\le c_3\frac{\rho^2\|\boldsymbol{y}\|^2}{\bar{M}^2}\frac{s\log p}{\tilde{n}}+(1+c_3\frac{\rho^2\|\boldsymbol{y}\|^2}{\bar{M}^2})\{\frac{\rho^2}{\bar{M}}+\frac{\bar{M}}{\sigma_0^2}(1-\frac{\rho\sqrt{M_0/2}}{\bar{M}})^2\}(\frac{s\log p}{\tilde{n}})^2+C_3s\log p\\ &\le c_3\frac{\|\boldsymbol{y}\|_2^2s\log p}{n\bar{M}}+(1+\frac{\tilde{n}\|\boldsymbol{y}\|_2^2}{n\bar{M}})(\frac{\tilde{n}}{n}+\frac{\bar{M}}{\sigma_0^2}+\frac{\tilde{n}M_0/2}{n\sigma_0^2})(\frac{s\log p}{\tilde{n}})^2+C_3s\log p.\end{aligned}$$

Notice that (A.5) guarantees that

$$(\frac{\tilde{n}}{n}+\frac{\bar{M}}{\sigma_0^2}+\frac{\tilde{n}M_0/2}{n\sigma_0^2})\frac{s\log p}{\tilde{n}}\le c_5$$

for some small enough c_5. Then we arrive at for some large enough constant C_1,

$$KL(\mathbb{P}_i(\mathcal{Z}_1|\boldsymbol{y}),\mathbb{P}_j(\mathcal{Z}_1|\boldsymbol{y}))\le C_1\frac{\|\boldsymbol{y}\|_2^2s\log p}{n\bar{M}}+C_3s\log p$$

under (A.5). Using the properties of Chi-Square distribution for $\|\boldsymbol{y}\|_2^2/\bar{M}$, we have

$$\max_{i,j\le M} KL(\mathbb{P}_i(\mathcal{Z}|\boldsymbol{y}),\mathbb{P}_j(\mathcal{Z}|\boldsymbol{y}))\le Cs\log p.$$

Together with (A.2), the proof is complete.

□

Proof of Theorem 4. Let $\mathrm{TV}(f_1,f_0)=\int|f_1(z)-f_0(z)|dz$ and $\chi^2(f_1,f_0)=\int f_1^2(z)/f_0(z)dz-1$.

Part 1. We first show the parametric rate $\sqrt{(M_0+\sigma_0^2)/n}+\sqrt{M_0/\tilde{n}}$. Consider the following two hypotheses.

$$H_0:\ \boldsymbol{\Sigma}^{(0)}=I_p,\ \boldsymbol{\beta}^{(0)}=(0,\sqrt{\frac{M_0}{2}},\mathbf{0}_{p-2}^\intercal)^\intercal,(\sigma^{(0)})^2=\sigma_0^2.$$

$$H_1:\ \boldsymbol{\Sigma}^{(1)}=\begin{pmatrix}1 & \frac{1}{\sqrt{\tilde{n}}} & \mathbf{0}_{p-2}^\intercal\\ \frac{1}{\sqrt{\tilde{n}}} & 1 & \mathbf{0}_{p-2}^\intercal\\ 0 & 0 & I_{p-1}\end{pmatrix},$$

$$\boldsymbol{\beta}^{(1)}=(\boldsymbol{\Sigma}^{(1)})^{-1}(\frac{\rho}{\sqrt{\tilde{n}}},\sqrt{\frac{M_0}{2}},\mathbf{0}_{p-2}^\intercal)=\frac{1}{1-1/\tilde{n}}\begin{pmatrix}\frac{\rho-\sqrt{\frac{M_0}{2}}}{\sqrt{\tilde{n}}}\\ \sqrt{\frac{M_0}{2}}-\frac{\rho}{\tilde{n}}\end{pmatrix},$$

$$(\sigma^{(1)})^2=\sigma_0^2+M_0/2-(\boldsymbol{\beta}^{(1)})^\intercal\boldsymbol{\Sigma}^{(1)}\boldsymbol{\beta}^{(1)}=\sigma_0^2-\frac{(\rho-\sqrt{\frac{M_0}{2}})^2}{\tilde{n}-1}.$$

It is easy to check that for

$$(\boldsymbol{\beta}^{(1)})^\intercal\boldsymbol{\Sigma}^{(1)}\boldsymbol{\beta}^{(1)}=\sigma_0^2-\frac{(\rho-\sqrt{\frac{M_0}{2}})^2}{\tilde{n}-1}\le\min\{\frac{M_0}{2},\sigma_0^2\},\tag{A.9}$$

$\{\boldsymbol{\beta}^{(i)},\boldsymbol{\Sigma}^{(i)},(\sigma^{(i)})^2\}_{i=0}^1\subseteq\Xi(s,M_0,\sigma_0^2)$. Hence, by (A.2),

$$\begin{aligned}&\inf_{\hat{\boldsymbol{\beta}}\in\mathcal{F}(\mathcal{Z})}\sup_{\boldsymbol{\beta}\in\Xi(s,M_0,\sigma_0^2)}\mathbb{P}\left(|\hat{\beta}_1-\beta_1|\ge|(\boldsymbol{\beta}^{(1)}-\boldsymbol{\beta}^{(0)})_1|\right)\ge\inf_{\hat{\boldsymbol{\beta}}\in\mathcal{F}(\mathcal{Z})}\sup_{\{\boldsymbol{\beta}^{(i)},\boldsymbol{\Sigma}^{(i)},(\sigma^{(i)})^2\}_{i=0}^1}\mathbb{E}[|\hat{\beta}_1-\beta_1|]\\&\ge\max_{m\in\{0,1\}}\mathbb{P}_m(\hat{\psi}\neq m)\ge\frac{1-TV(f_0(\mathcal{Z}),f_1(\mathcal{Z}))}{2}\end{aligned}\tag{A.10}$$

as long as $TV(f_0(\mathcal{Z}),f_1(\mathcal{Z}))\le c_0<1$ where the last step is due to Theorem 2.2 in Tsybakov (2008). Notice that

$$|(\boldsymbol{\beta}^{(1)}-\boldsymbol{\beta}^{(0)})_1|=\frac{1}{1-\tilde{n}}\frac{|\rho-\sqrt{M_0/2}|}{\sqrt{\tilde{n}}}\ge c_1\frac{\gamma_{n,\tilde{n}}^{1/2}}{\sqrt{n}},$$

where the last step is due to (A.3). We are left to show $TV(f_0(\mathcal{Z}), f_1(\mathcal{Z})) \le c_0 < 1$. As $TV(f_0(\mathcal{Z}), f_1(\mathcal{Z})) \le \sqrt{KL(f_1, f_0)/2}$, we consider

$$KL(f_1(\mathcal{Z}), f_0(\mathcal{Z})) = KL(f_1(\mathcal{Z}_1), f_0(\mathcal{Z}_1)) + KL(f_1(\mathcal{Z}_2), f_0(\mathcal{Z}_2))$$

We can calculate that

$$KL(f_1(\mathcal{Z}_2), f_0(\mathcal{Z}_2)) = \log(1 - \frac{1}{\tilde{n}}).$$

According to (A.7), we have

$$\begin{aligned}
(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\intercal \boldsymbol{W}_0^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) &= \frac{\|\boldsymbol{y}\|_2^2 \rho^2}{\bar{M}^2 \tilde{n}} \\
det(\boldsymbol{W}_0^{-1}\boldsymbol{W}_1) &= 1 - \frac{\frac{\sigma_0^2\rho^2}{\tilde{n}} + (\frac{\bar{M} - \rho\sqrt{\frac{M_0}{2}}}{\bar{M}\sqrt{\tilde{n}}})^2}{\sigma_0^2/\bar{M}} \\
\mathrm{Tr}(\boldsymbol{W}_0^{-1}(\boldsymbol{W}_1 - \boldsymbol{W}_0)) &= -\frac{\rho^2}{\tilde{n}\bar{M}}
\end{aligned}$$

By (A.8) and (A.7), we can calculate that

$$\begin{aligned}
KL(f_1(\mathcal{Z}_1), f_0(\mathcal{Z}_1)) &= \mathbb{E}\boldsymbol{y}[KL(f_1(\mathcal{Z}_1|\boldsymbol{y}), f_0(\mathcal{Z}_1|\boldsymbol{y}))] \\
&\le \mathbb{E}\boldsymbol{y}[\frac{\|\boldsymbol{y}\|_2^2\rho^2}{\bar{M}^2\tilde{n}} - \frac{\rho^2}{\tilde{n}\bar{M}} + \frac{\frac{\sigma_0^2\rho^2}{\tilde{n}} + (\frac{\bar{M} - \rho\sqrt{\frac{M_0}{2}}}{\bar{M}\sqrt{\tilde{n}}})^2}{\sigma_0^2/\bar{M}}],
\end{aligned}$$

where the last step holds when $\frac{\frac{\sigma_0^2\rho^2}{\tilde{n}} + (\frac{\bar{M} - \rho\sqrt{\frac{M_0}{2}}}{\bar{M}\sqrt{\tilde{n}}})^2}{\sigma_0^2/\bar{M}} \le \log 2/2$.

We can calculate that if $\min\{n, \tilde{n}\} \ge C \max\{1, \bar{M}^2\}$ for large enough C, then

$$KL(f_1(\mathcal{Z}), f_0(\mathcal{Z})) \le 1/2 \text{ and } TV(f_1(\mathcal{Z}), f_0(\mathcal{Z})) \le 1/2.$$

The proof is complete now. $\square$

Theorem A.1 (Lower bound for estimating β_j with known $\boldsymbol{\Sigma}$)**.** *Consider the parameter space* $\Xi(s, M_0, \sigma_0^2) \cap \{\boldsymbol{\Sigma} = I_p\}$. *Suppose that* $\max\{1, (M_0 + \sigma_0^2)\} \le c_1 n$ *for some constant* $c_1 > 0$. *Then for any fixed* $1 \le j \le p$, *there exists some constant* c_2 *that*

$$\inf_{\hat{\beta}_j \in \mathcal{F}(\mathcal{Z})} \sup_{\boldsymbol{\beta} \in \Xi(s, M_0, \sigma_0^2) \cap \{\boldsymbol{\Sigma} = I_p\}} \mathbb{P}\left(|\hat{\beta}_j - \beta_j| \ge c_2 \sqrt{\frac{M_0 + \sigma_0^2}{n}}\right) \ge \frac{1}{2}.$$

Proof of Theorem A.1.

$$H_0: \ \boldsymbol{\Sigma}^{(0)} = I_p, \ \boldsymbol{\beta}^{(0)} = \mathbf{0}_p, (\sigma^{(0)})^2 = \sigma_0^2.$$

$$H_1: \ \boldsymbol{\Sigma}^{(1)} = I_p, \ \boldsymbol{\beta}^{(1)} = (c\sqrt{\frac{\bar{M}}{n}}, \mathbf{0}_{p-1}^\intercal),$$

$$(\sigma^{(1)})^2 = \sigma_0^2 - (\boldsymbol{\beta}^{(1)})^\intercal \boldsymbol{\Sigma}^{(1)} \boldsymbol{\beta}^{(1)} = \sigma_0^2 - \frac{c^2 \bar{M}}{n}.$$

As long as $c^2\bar{M}/n \leq \min\{\sigma_0^2, M_0\}/2$, $\{\boldsymbol{\beta}^{(i)}, \boldsymbol{\Sigma}^{(i)}, (\sigma^{(i)})^2\}_{i=0}^1 \in \Xi(s, M_0, \sigma_0^2)$. We can calculate that $|(\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(0)})_1| = c\sqrt{\bar{M}/n}$. It is left to bound $KL(f_1(\mathcal{Z}), f_0(\mathcal{Z}))$. As the distribution of $\mathcal{Z}_2$ is unchanged, we can easily calculate that for a small enough constant c,

$$KL(f_1(\mathcal{Z}), f_0(\mathcal{Z})) = KL(f_1(\mathcal{Z}_1), f_0(\mathcal{Z}_1)) \leq 1/2.$$

By (A.10), the proof is complete now. □

A.3 Proofs in Section 4

Proof of Theorem 5.

$$\begin{aligned}
\hat{\mu}_*^{(ts-db)} - \mu_* &= \boldsymbol{x}_*^\intercal(\hat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}) + \tilde{\boldsymbol{w}}_*^\intercal(\widehat{\boldsymbol{S}} - \widetilde{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}}^{(ts)}) \\
&= \boldsymbol{x}_*^\intercal(\hat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}) + \tilde{\boldsymbol{w}}_*^\intercal(\boldsymbol{X}^\intercal\boldsymbol{y}/n - \boldsymbol{\Sigma}\boldsymbol{\beta} + \boldsymbol{\Sigma}\boldsymbol{\beta} - \widetilde{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}}^{(ts)}) \\
&= \boldsymbol{x}_*^\intercal(\hat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}) + \tilde{\boldsymbol{w}}_*^\intercal(\boldsymbol{X}^\intercal\boldsymbol{y}/n - \boldsymbol{\Sigma}\boldsymbol{\beta}) + \tilde{\boldsymbol{w}}_*^\intercal(\boldsymbol{\Sigma}\boldsymbol{\beta} - \widetilde{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}}^{(ts)}) \\
&= \boldsymbol{x}_*^\intercal(\hat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}) + \tilde{\boldsymbol{w}}_*^\intercal(\boldsymbol{X}^\intercal\boldsymbol{y}/n - \boldsymbol{\Sigma}\boldsymbol{\beta}) + \tilde{\boldsymbol{w}}_*^\intercal(\boldsymbol{\Sigma}\boldsymbol{\beta} - \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\beta}) \\
&\quad + \tilde{\boldsymbol{w}}_*^\intercal\widetilde{\boldsymbol{\Sigma}}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(ts)}) \\
&= \underbrace{(\boldsymbol{x}_* - \tilde{\boldsymbol{w}}_*^\intercal\widetilde{\boldsymbol{\Sigma}})^\intercal(\hat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})}_{R_1} + \underbrace{\boldsymbol{x}_*^\intercal\boldsymbol{\Omega}^\intercal(\boldsymbol{X}^\intercal\boldsymbol{y}/n - \boldsymbol{\Sigma}\boldsymbol{\beta}) + \boldsymbol{x}_*^\intercal\boldsymbol{\Omega}(\boldsymbol{\Sigma}\boldsymbol{\beta} - \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\beta})}_{R_0} \\
&+ \underbrace{(\tilde{\boldsymbol{w}}_* - \boldsymbol{x}_*^\intercal\boldsymbol{\Omega})^\intercal(\boldsymbol{X}^\intercal\boldsymbol{y}/n - \boldsymbol{\Sigma}\boldsymbol{\beta} + \boldsymbol{\Sigma}\boldsymbol{\beta} - \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\beta})}_{R_2}
\end{aligned}$$

Let $\boldsymbol{w}_* = \boldsymbol{x}_*^\intercal\boldsymbol{\Omega}$. Define an event

$$\begin{aligned}
E_* = \Big\{ &\max_j \|\widetilde{\boldsymbol{\Omega}}_{\cdot,j} - \boldsymbol{\Omega}_{\cdot,j}\|_0 \leq Cs_\Omega, \|\widetilde{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_S \leq Cs_\Omega\tilde{\lambda}, \\
&\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta}\|_{\infty,\infty} \leq C\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2\tilde{\lambda}, \boldsymbol{w}_* \text{ is feasible to (4.2)} \Big\}.
\end{aligned}$$

For R_2, we have

$$\begin{aligned}
|R_2| &\leq |(\tilde{\boldsymbol{w}}_* - \boldsymbol{w}_*)^\intercal(\boldsymbol{X}^\intercal\boldsymbol{y}/n - \boldsymbol{\Sigma}\boldsymbol{\beta} + \boldsymbol{\Sigma}\boldsymbol{\beta} - \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\beta})| \\
&\leq \|\tilde{\boldsymbol{w}}_* - \boldsymbol{w}_*\|_1\|\boldsymbol{X}^\intercal\boldsymbol{y}/n - \boldsymbol{\Sigma}\boldsymbol{\beta} + \boldsymbol{\Sigma}\boldsymbol{\beta} - \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\beta}\|_\infty \\
&\leq (\|\tilde{\boldsymbol{w}}_* - \boldsymbol{x}_*^\intercal\widetilde{\boldsymbol{\Omega}}\|_1 + \|\boldsymbol{x}_*(\widetilde{\boldsymbol{\Omega}} - \boldsymbol{\Omega})\|_1)\lambda^{(ts)}.
\end{aligned}$$

In event E_*, $\boldsymbol{w}_*$ is feasible to (4.2) and hence $\|\tilde{\boldsymbol{w}}_* - \boldsymbol{x}_*^\intercal \widetilde{\boldsymbol{\Omega}}\|_1 \leq \|\boldsymbol{w}_* - \boldsymbol{x}_*^\intercal \widetilde{\boldsymbol{\Omega}}\|_1$. Hence, we have

$$\begin{aligned}
R_2 &\leq 2\|\boldsymbol{x}_*(\widetilde{\boldsymbol{\Omega}} - \boldsymbol{\Omega})\|_1 \lambda^{(ts)} \\
&\leq 2\|\boldsymbol{x}_*\|_2 \|\widetilde{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_{2,1} \lambda^{(ts)} \\
&\leq 2\|\boldsymbol{x}_*\|_2 s_\Omega^{1/2} \|\widetilde{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_S \lambda^{(ts)} \\
&\leq 2\|\boldsymbol{x}_*\|_2 s_\Omega^{3/2} \tilde{\lambda} \lambda^{(ts)}.
\end{aligned}$$

For R_1, using the constraint in (4.2),

$$\begin{aligned}
|R_1| &\leq \|\boldsymbol{x}_* - \tilde{\boldsymbol{w}}_*^\intercal \widetilde{\boldsymbol{\Sigma}}\|_\infty \|\hat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}\|_1 \\
&\leq C\|\boldsymbol{x}_*\|_2 s \lambda^{(ts)} \tilde{\lambda}.
\end{aligned}$$

We can show that

$$\sqrt{\frac{(\boldsymbol{x}_*^\intercal \boldsymbol{\Omega} \boldsymbol{x}_*)\gamma_{n,\tilde{n}}}{n} + \frac{(\boldsymbol{x}_*^\intercal \boldsymbol{\beta})^2}{n} + \frac{(\boldsymbol{x}_*^\intercal \boldsymbol{\beta})^2}{\tilde{n}}} R_0 | \boldsymbol{x}_* \xrightarrow{D} N(0,1).$$

We are left to show $\mathbb{P}(E_*) \to 1$, which follows from standard arguments from CLIME. □

B Power analysis based one-sample and two-sample debiased Lasso

In this section, we evaluate the power of hypothesis testing with proxy data. We have seen in Section 2 that it is necessary to focus on the regime that $s \log p \ll \sqrt{\tilde{n}}$ and $s_j \log p \ll \tilde{n}/\sqrt{n} + \tilde{n}^{1/2}$ for valid inference. For the simplicity of the power analysis, we ignore the asymptotic bias in the debiased estimator. To avoid confusion, we introduce some new notations. Let $\sqrt{n}\hat{z}_j^{(os)}$ be the probabilistic limit of conventional debiased estimator $\sqrt{n}\hat{\beta}_j^{(db)}/\sqrt{V_j}$ where $V_j = \Omega_{j,j}\sigma^2$ (van de Geer et al., 2014). Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. The distribution of one-sample z-score is

$$\sqrt{n}\hat{z}_j^{(os)} - \frac{\sqrt{n}\beta_j}{\sqrt{\Omega_{j,j}\sigma^2}} \sim N(0,1), \ 1 \leq j \leq p. \tag{B.1}$$

For proxy-data based inference, let $\sqrt{n}\hat{z}_j^{(ts)}$ be the probabilistic limit of two-sample debiased estimator $\sqrt{n}\hat{\beta}_j^{(ts-db)}/\sqrt{V_j^{(ts)}}$. The marginal distribution of each two-sample z-score is

$$\sqrt{n}\hat{z}_j^{(ts)} - \frac{\sqrt{n}\beta_j}{\sqrt{\Omega_{j,j}\gamma_{n,\tilde{n}} + \beta_j^2(1 + n/\tilde{n})}} \sim N(0,1), \ 1 \leq j \leq p. \tag{B.2}$$

We see that the distribution of $\hat{z}_j^{(ts)}$ depends not only on β_j but all other coefficients β_{-j}.

We evaluate the power of single hypothesis testing, $\mathcal{H}_{0,j}: \ \beta_j = 0$ vs $\mathcal{H}_{1,j}: \ \beta_j \neq 0$ based on $\hat{z}_j^{(os)}$ and $\hat{z}_j^{(ts)}$, respectively. By definition, the power of these two statistics can be expressed as

$$\text{Power}_j^{(os)}(\alpha; \boldsymbol{b}) = \mathbb{P}(|\hat{z}_j^{(os)}| \geq \tau_\alpha | \beta_j = b_j, \boldsymbol{\beta}_{-j} = \boldsymbol{b}_{-j}).$$
$$\text{Power}_j^{(ts)}(\alpha, \boldsymbol{b}) = \mathbb{P}(|\hat{z}_j^{(ts)}| \geq \tau_\alpha | \beta_j = b_j, \boldsymbol{\beta}_{-j} = \boldsymbol{b}_{-j}).$$

Based on (B.1), we know that $\text{Power}_j^{(os)}(\alpha; \boldsymbol{b})$ is independent of $\boldsymbol{b}_{-j}$.

Theorem B.1 (Power of two-sided single hypothesis testing)**.** *For any $\boldsymbol{b} \in \mathbb{R}^p$, it holds that*

$$\text{Power}_j^{(os)}(\alpha, \boldsymbol{b}) = \Phi(|\eta_j^{(os)}| - \tau_\alpha) + \Phi(-|\eta_j^{(os)}| - \tau_\alpha), \text{ where } \eta_j^{(os)} = \frac{\sqrt{n} b_j}{\Omega_{j,j}^{1/2} \sigma}.$$
$$\text{Power}_j^{(ts)}(\alpha, \boldsymbol{b}) = \Phi\left(|\eta_j^{(ts)}| - \tau_\alpha\right) + \Phi\left(-|\eta_j^{(ts)}| - \tau_\alpha\right), \text{ where}$$
$$\eta_j^{(ts)} = \frac{\sqrt{n} b_j}{\sqrt{(\Omega_{j,j} \boldsymbol{b}^\intercal \boldsymbol{\Sigma} \boldsymbol{b} + b_j^2)(1 + n/\tilde{n}) + \Omega_{j,j} \sigma^2}}.$$

The one-sample and two-sample power functions are increasing functions of $|\eta_j^{(os)}|$ and $|\eta_j^{(ts)}|$, respectively. The power based on proxy data is smaller than that based on one-sample data for $\beta_j \neq 0$. Furthermore, the power based on $\hat{z}_j^{(os)}$ is independent of effect size distribution but the power based on $\hat{z}_j^{(ts)}$ depends on the effect size distribution. Specifically, if the effect size vector $\boldsymbol{b}$ has sparsity s and approximately equal signal strength, then $\eta_j^{(ts)}$ is approximately $\sqrt{n} b_j / \sqrt{\Omega_{j,j}\{s b_j^2 (1 + n/\tilde{n}) + \sigma^2\}}$. We see that $|\eta_j^{(ts)}|$ is bounded away from infinity for finite $n \wedge \tilde{n}$ even if $b_j \to \infty$.

To further demonstrate this phenomenon, we consider the equal signal strength model

$$b_j \in \{-b_0, 0, b_0\} \text{ and } \|\boldsymbol{b}\|_0 = s. \tag{B.3}$$

Corollary B.1 (Power in the equal signal strength model)**.** *Consider the equal strength model (B.3) with $\boldsymbol{\Sigma} = I_p$. The results of Theorem B.1 hold with*

$$|\eta_j^{(os)}| = \frac{\sqrt{n}|b_0|}{\sigma} \text{ and } |\eta_j^{(ts)}| \leq \min\left\{\sqrt{\frac{n \wedge \tilde{n}}{s}}, \frac{\sqrt{n}|b_0|}{\sigma}\right\}. \tag{B.4}$$

In the two-sample setting, the signal strength $|\eta_j^{(ts)}|$ and hence the power is upper-truncated. As long as $|b_0| > \sigma\sqrt{(n \wedge \tilde{n})/n/s}$, the power based on proxy-data is strictly lower than the power based on one-sample data. Second, for any finite $n \wedge \tilde{n}$ and s, the right-hand side of (B.4) is strictly below one no matter how large $|b_0|/\sigma$ is. This analysis demonstrates a significant loss of power in proxy-data based inference in comparison to the one-sample based inference.

Proof of Theorem B.1 and Corollary B.1. We note that

$$\gamma_{n,\tilde{n}} = \frac{\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma} \boldsymbol{\beta} + \sigma^2}{n} + \boldsymbol{\beta}^\intercal \boldsymbol{\Sigma} \boldsymbol{\beta}/\tilde{n} = \frac{1}{n}\{\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma} \boldsymbol{\beta}(1 + \frac{n}{\tilde{n}}) + \sigma^2\}.$$

By (B.1), (B.2), the proof of B.1 is obvious and is omitted.

For (B.4), notice that in (B.3),

$$\text{Power}^{(ts)}(\alpha, \boldsymbol{b}) = \Phi\left(\frac{\sqrt{n} b_j}{\Omega_{j,j}^{1/2}\sqrt{\boldsymbol{b}^\intercal \boldsymbol{\Sigma} \boldsymbol{b}(1 + n/\tilde{n}) + \sigma^2}} - \tau_\alpha\right).$$

The proof is complete by noting that if $b_j = b_0$,

$$\frac{\sqrt{n} b_0}{\Omega_{j,j}^{1/2}\sqrt{s b_0^2(1 + n/\tilde{n}) + \sigma^2}} \leq \min\{\frac{b_0}{\sqrt{s b_0^2/n + s b_0^2/\tilde{n}}}, \frac{\sqrt{n} b_0}{\sigma}\} \leq \min\{\sqrt{\frac{n}{s}}, \sqrt{\frac{\tilde{n}}{s}}, \frac{\sqrt{n} b_0}{\sigma}\}.$$

□

C Further results on Section 5

C.1 Data generation process

For the "ind" case, the true effect $\boldsymbol{\beta}$ is obtained from GWAS "ieu-b-40" for BMI from the IEU open GWAS project (Elsworth et al., 2020). We truncate $\boldsymbol{\beta}$ to keep its top-s largest values and set other coordinates of $\boldsymbol{\beta}$ to be zero. For the "cfw" case, let $\beta_{4(k-1)+1:4k} = (0.2, 0.1, -0.1, -0.2)^\intercal$ for $k = 1, \ldots, s/4$ and $\beta_k = 0$ otherwise. As in Section 6, we arrive at 27911 SNPs after prescreening. In each Monte Carlo experiment, we randomly sample 2000 SNPs from 27911 SNPs as the covariates. Therefore, the snps are correlated.

C.2 Implementations in Section 5

We describe a convenient implementation of (4.2) based on the established R packages. We first realize $\widetilde{\boldsymbol{\Omega}}$ based on the "fastclime" package in R or based on Lasso estimate of $\boldsymbol{w}_j$,

$j = 1, \dots, p$. For (4.2), it can be parameterized as

$$\tilde{\boldsymbol{w}}_* = \rho \widetilde{\boldsymbol{\Omega}} \boldsymbol{x}_* + \arg\min_{\boldsymbol{\delta} \in \mathbb{R}^p} \|\boldsymbol{\delta}\|_1$$
$$\text{subject to } \|\widetilde{\boldsymbol{\Sigma}}(\boldsymbol{\delta} + \rho \widetilde{\boldsymbol{\Omega}} \boldsymbol{x}_*) - \boldsymbol{x}_*\|_\infty \le \|\boldsymbol{x}_*\|_2 \tilde{\lambda},$$

where $\rho \in \mathbb{R}$ can be estimated by regression $\boldsymbol{x}_*$ on $\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\Omega}}\boldsymbol{x}_*$ and δ can be estimated via a Dantzig selector. We know that the Dantzig selector and the Lasso are asymptotically equivalent (Bickel et al., 2009). Hence, we solve the Lasso counterpart to obtain an estimated δ:

$$\hat{\delta} = \arg\min_{\boldsymbol{\delta} \in \mathbb{R}^p} \left\{ \frac{1}{2} \boldsymbol{\delta}^\intercal \widetilde{\boldsymbol{\Sigma}} \boldsymbol{\delta} - (\widehat{\boldsymbol{S}} - \hat{\rho} \widetilde{\boldsymbol{\Sigma}} \widetilde{\boldsymbol{\Omega}} \boldsymbol{x}_*) + \|\boldsymbol{x}_*\|_2 \tilde{\lambda} \|\boldsymbol{\delta}\|_1 \right\}.$$

C.3 More results for prediction and estimation

We consider $(n, \tilde{n}) \in \{(150, 450), (300, 300), (450, 150)\}$, which corresponds to $n/\tilde{n} \in \{1/3, 1, 3\}$. For prediction and estimation tasks, we consider $p = 2000$ and $s \in \{4, 8, 12\}$. Let $\beta_{4(k-1)+1:4k} = (0.5, -0.5, 0.2, -0.2)^\intercal$ for $k = 1, \dots, s/4$ and $\beta_k = 0$ otherwise. We consider the identity covariance matrix, the Toeplitz covariance matrix with $\boldsymbol{\Sigma}_{j,k} = 0.7^{|j-k|}$, and the equi-correlated $\boldsymbol{\Sigma}$ with $\boldsymbol{\Sigma}_{j,k} = 1$ if $j = k$ and $\boldsymbol{\Sigma}_{j,k} = 0.3$ if $j \neq k$. The identity $\boldsymbol{\Sigma}$ is a simple sparse matrix with $s_\Omega = 1$, the Toeplitz $\boldsymbol{\Sigma}$ is approximately sparse, and the equi-correlated $\boldsymbol{\Sigma}$ is non-sparse.

In Table C.1, we report the estimation and prediction results with one-sample and two-sample Lasso. As one-sample method only uses n individual samples, its estimation errors and test errors decrease as n increases for any given s. For the two-sample Lasso, the test and estimation errors are always larger than those with one-sample Lasso. When $n/\tilde{n}$ is fixed, both methods have errors increasing as the sparsity s increases. The errors of two-sample Lasso increase more significantly. This is because, according to Theorem 1, as s increases, SNR increases which leads to larger estimation errors.

C.4 More results on inference

In Table C.2, we evaluate the performance of statistical inference for the regression coefficients based on one-sample and two-sample debiased Lasso. We see that the one-sample and two-sample debiased Lasso have coverage probabilities close to the nominal level in various configurations. For strong signals, both methods have slightly lower coverage when s is larger. The standard deviations based on proxy data are larger than those based on one-sample data especially when $n/\tilde{n}$ is large. The gap increases as SNR increases. This agrees

Table C.1: Sum of squared errors for estimating $\boldsymbol{\beta}$ and test errors based on one-sample Lasso (OS) and two-sample Lasso (TS) for $n/\tilde{n} \in \{1/3, 1, 3\}$ and for identity (Ident), Toeplitz (Toep), and equi-correlated (Equi) covariance matrices.

			Sum of Squared Errors			Test Errors		
$\boldsymbol{\Sigma}$	s	Method	1/3	1	3	1/3	1	3
Ident	4	OS	0.357	0.241	0.161	0.359	0.226	0.156
		TS	0.557	0.607	0.604	0.562	0.571	0.575
	8	OS	0.739	0.334	0.234	0.706	0.351	0.249
		TS	1.209	0.928	1.010	1.130	0.971	1.074
	12	OS	1.095	0.520	0.333	1.089	0.506	0.336
		TS	1.724	1.485	1.628	1.693	1.469	1.614
Toep	4	OS	0.667	0.603	0.572	0.580	0.579	0.569
		TS	0.675	0.616	0.585	0.586	0.581	0.579
	8	OS	1.189	1.098	1.151	1.159	1.157	1.146
		TS	1.200	1.108	1.163	1.168	1.159	1.160
	12	OS	1.811	1.642	1.587	1.738	1.735	1.722
		TS	1.824	1.647	1.590	1.750	1.740	1.738
Equi	4	OS	0.460	0.269	0.226	0.455	0.287	0.222
		TS	0.585	0.537	0.583	0.575	0.570	0.570
	8	OS	0.958	0.465	0.306	0.883	0.464	0.335
		TS	1.207	1.111	0.959	1.144	1.063	1.091
	12	OS	1.344	0.689	0.445	1.350	0.662	0.463
		TS	1.723	1.596	1.644	1.715	1.516	1.703

Table C.2: Coverage probabilities and standard deviations based on one-sample (OS) and two-sample (TS) debiased Lasso for $n/\tilde{n} \in \{1/3, 1, 3\}$ and identity covariance matrix.

			Average Coverage			Average SD		
β_j	s	Method	1/3	1	3	1/3	1	3
0	4	OS	0.965	0.930	0.950	0.09	0.06	0.05
		TS	0.980	0.935	0.980	0.10	0.07	0.06
	8	OS	0.975	0.965	0.960	0.10	0.06	0.05
		TS	0.965	0.990	0.975	0.12	0.09	0.08
	12	OS	0.945	0.960	0.930	0.11	0.07	0.05
		TS	0.945	0.980	0.945	0.14	0.10	0.09
0.2	4	OS	0.950	0.980	0.940	0.09	0.06	0.05
		TS	0.940	0.965	0.940	0.11	0.07	0.06
	8	OS	0.970	0.955	0.950	0.10	0.06	0.05
		TS	0.960	0.950	0.940	0.12	0.09	0.08
	12	OS	0.940	0.925	0.935	0.11	0.07	0.05
		TS	0.945	0.975	0.935	0.14	0.11	0.09
0.5	4	OS	0.930	0.935	0.970	0.09	0.06	0.05
		TS	0.925	0.940	0.945	0.12	0.08	0.08
	8	OS	0.930	0.940	0.970	0.10	0.06	0.05
		TS	0.925	0.955	0.925	0.13	0.10	0.09
	12	OS	0.915	0.880	0.905	0.11	0.07	0.05
		TS	0.925	0.915	0.895	0.14	0.11	0.10

with our analysis in Theorem 3. From Table C.3 and Table C.4, we see similar patterns with non-identity $\mathbf{\Sigma}$. As the precision matrix $\mathbf{\Omega}$ gets less sparse, the coverage probability gets lower, especially for strong signals. The effects of s_Ω on asymptotic normality have also been revealed in our theoretical results.

			Average Coverage			Average SD		
β_j	s	Method	1/3	1	3	1/3	1	3
0	4	OS	0.950	0.970	0.950	0.10	0.08	0.07
		TS	0.960	0.965	0.940	0.11	0.07	0.06
	8	OS	0.930	0.975	0.935	0.11	0.08	0.07
		TS	0.920	0.980	0.940	0.11	0.08	0.06
	12	OS	0.955	0.940	0.950	0.11	0.09	0.07
		TS	0.955	0.940	0.960	0.12	0.08	0.06
-0.2	4	OS	0.940	0.940	0.945	0.10	0.08	0.07
		TS	0.940	0.950	0.895	0.11	0.07	0.06
	8	OS	0.815	0.870	0.825	0.11	0.08	0.07
		TS	0.880	0.775	0.320	0.11	0.08	0.06
	12	OS	0.825	0.915	0.760	0.11	0.09	0.07
		TS	0.840	0.850	0.255	0.12	0.08	0.06
0.5	4	OS	0.580	0.570	0.625	0.10	0.09	0.06
		TS	0.720	0.450	0.100	0.11	0.09	0.06
	8	OS	0.660	0.710	0.665	0.10	0.09	0.06
		TS	0.775	0.660	0.250	0.11	0.10	0.06
	12	OS	0.760	0.745	0.760	0.11	0.09	0.07
		TS	0.830	0.710	0.360	0.12	0.10	0.07

Table C.3: Coverage probabilities and standard deviations based on one-sample (OS) and two-sample (TS) debiased Lasso with $n/\tilde{n} \in \{1/3, 1, 3\}$ and Toeplitz covariance matrix.

We report the inference results of the true PRS μ_* in Table C.5. The coverage probabilities are close to the nominal level in different settings. The standard deviations based on two-sample debiased Lasso are still larger than those based on one-sample debiased Lasso.

C.5 Tuning parameter selection

We evaluated three methods for selecting the tuning parameters as follows.

1. A modified BIC-based selection criteria. Recall that $\widehat{M} = \|y\|_2^2/n$ and $\widehat{S} = X^T y/n$.

			Average Coverage			Average SD		
β_j	s	Method	1/3	1	3	1/3	1	3
0	4	OS	0.935	0.975	0.980	0.10	0.07	0.05
		TS	0.940	0.970	0.945	0.10	0.07	0.06
	8	OS	0.960	0.995	0.980	0.10	0.07	0.06
		TS	0.945	0.955	0.955	0.12	0.09	0.07
	12	OS	0.970	0.950	0.965	0.11	0.07	0.06
		TS	0.915	0.930	0.925	0.13	0.10	0.07
0.2	4	OS	0.885	0.935	0.940	0.10	0.07	0.05
		TS	0.890	0.935	0.920	0.10	0.07	0.06
	8	OS	0.905	0.885	0.940	0.10	0.07	0.06
		TS	0.930	0.865	0.905	0.12	0.09	0.07
	12	OS	0.945	0.930	0.925	0.11	0.07	0.06
		TS	0.925	0.945	0.935	0.13	0.10	0.08
0.5	4	OS	0.820	0.890	0.905	0.10	0.07	0.05
		TS	0.885	0.855	0.790	0.11	0.08	0.07
	8	OS	0.745	0.905	0.930	0.10	0.07	0.06
		TS	0.865	0.890	0.885	0.12	0.09	0.08
	12	OS	0.810	0.855	0.855	0.11	0.07	0.06
		TS	0.925	0.840	0.780	0.14	0.11	0.08

Table C.4: Coverage probabilities and standard deviations based on one-sample (OS) and two-sample (TS) debiased Lasso with $n/\tilde{n} \in \{1/3, 1, 3\}$ and Equi-correlated covariance matrix.

Table C.5: Coverage probabilities and standard deviations based on one-sample (OS) and two-sample (TS) debiased Lasso for μ_* with $n/\tilde{n} \in \{1/3, 1, 3\}$ and the identity (Ident), Toeplitz (Toep), and Equi-correlated (Equi) covariance matrices.

			Average Coverage			Average SD		
$\mathbf{\Sigma}$	s	Method	1/3	1	3	1/3	1	3
Ident	4	OS	0.965	0.935	0.965	2.05	1.39	1.11
		TS	0.955	0.930	0.955	2.30	1.61	1.32
	8	OS	0.935	0.935	0.950	2.23	1.44	1.15
		TS	0.920	0.950	0.945	2.66	1.99	1.63
	12	OS	0.955	0.960	0.920	2.41	1.50	1.17
		TS	0.965	0.950	0.930	2.99	2.25	1.85
Toep	4	OS	0.945	0.960	0.965	2.35	1.53	1.20
		TS	0.945	0.900	0.860	2.08	1.54	1.37
	8	OS	0.965	0.930	0.950	2.45	1.59	1.25
		TS	0.965	0.935	0.860	2.19	1.61	1.43
	12	OS	0.945	0.945	0.955	2.59	1.67	1.31
		TS	0.910	0.930	0.840	2.31	1.68	1.50
Equi	4	OS	0.970	0.945	0.965	1.34	1.03	0.91
		TS	0.845	0.760	0.630	1.75	1.14	0.86
	8	OS	0.950	0.950	0.950	1.48	1.06	0.92
		TS	0.865	0.815	0.575	1.98	1.32	1.00
	12	OS	0.935	0.935	0.960	1.62	1.11	0.94
		TS	0.835	0.750	0.650	2.21	1.49	1.08

For each candidate estimate $b_\lambda \in \mathbb{R}^p$ based on tuning parameter λ, the BIC based on two-sample summary data is defined as

$$\text{BIC}_0(b_\lambda) = \log(\widehat{M} - 2b_\lambda^T\widehat{S} + b_\lambda^T\widetilde{\mathbf{\Sigma}}b_\lambda) + \frac{\sqrt{\log\min(n,\tilde{n})}}{\min(n,\tilde{n})}\|b_\lambda\|_0.$$

However, the term $\widehat{M} - 2b_\lambda^T\widehat{S} + b_\lambda^T\widetilde{\mathbf{\Sigma}}b_\lambda$ can be negative because $\widehat{M}$ and $\widetilde{\mathbf{\Sigma}}$ are based on independent samples. Rewriting this term, we have

$$\widehat{M} - 2b_\lambda^T\widehat{S} + b_\lambda^T\widetilde{\mathbf{\Sigma}}b_\lambda = \frac{1}{n}\|y - Xb_\lambda\|_2^2 + b_\lambda^T(\widetilde{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}})b_\lambda.$$

The last term is mean zero but can be negative in finite samples. Hence, we take the uncertainty of last term into account and propose a modified BIC as

$$\text{BIC}(b_\lambda) = \log\left(\widehat{M} - 2b_\lambda^T\widehat{S} + b_\lambda^T\widetilde{\mathbf{\Sigma}}b_\lambda + \sqrt{2b_\lambda^T\widetilde{\mathbf{\Sigma}}b_\lambda(\frac{1}{n} + \frac{1}{\tilde{n}})}\right) + \frac{\sqrt{\log\min(n,\tilde{n})}}{\min(n,\tilde{n})}\|b_\lambda\|_0.$$

2. Resampling-based selection criteria. Resampling GWAS summary statistics is a common practice for evaluating its uncertainty in many applications such as Mendelian Randomization (Bowden et al., 2016, 2019). Specifically, let $\hat{\eta}_j$ denote the estimated standard deviation of $\widehat{S}_j$. We generate

   $$\widehat{S}_j^* \sim_{ind} N(\widehat{S}_j, \hat{\eta}_j),\ j = 1, \ldots, p.$$

   For each candidate estimate $b_\lambda \in \mathbb{R}^p$, we evaluate its "out-of-sample" loss

   $$L(b_\lambda) = \widehat{M} - 2b_\lambda^T\widehat{S}^* + b_\lambda^T\widetilde{\mathbf{\Sigma}}b_\lambda + \lambda\|b_\lambda\|_1.$$

   Note that b_λ is estimated based on $\widehat{S}$ and $\widetilde{\mathbf{\Sigma}}$ but its loss is evaluated based on $\widehat{S}^*$ and $\widetilde{\mathbf{\Sigma}}$.

3. Pseudo-validation (Mak et al., 2017).

   The pseudo-validation method is introduced in Section 2.2 of Mak et al. (2017). This methods chooses λ such that its corresponding b_λ maximizes

   $$\frac{b_\lambda^T\hat{r}}{\sqrt{b_\lambda^T\widetilde{\mathbf{\Sigma}}b_\lambda}},$$

   where $\hat{r}_j = \widehat{S}_j(1 - \text{fdr}_j)$ and fdr_j is the local false discovery rate of SNP j.

We evaluate three tuning parameter selection methods in the two settings considered in the simulation section.

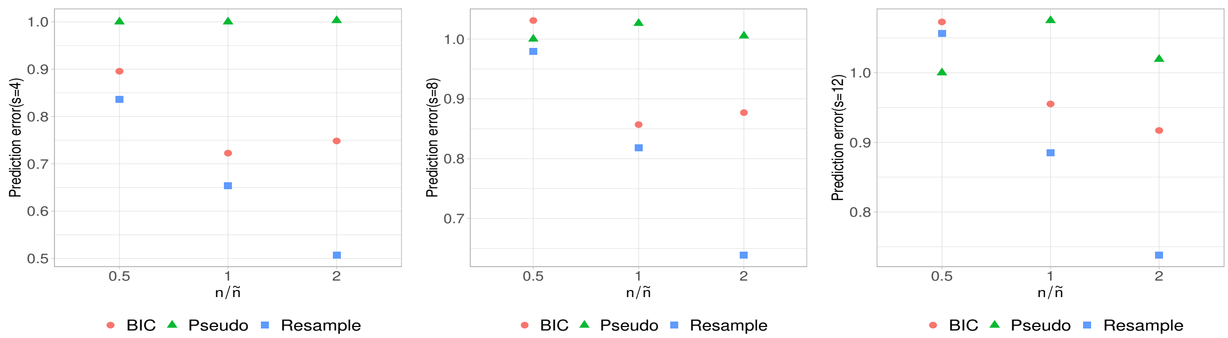

Figure C.1: Prediction errors based on three tuning parameter selection methods in setting 1 with independent SNPs. Each point is based on 200 Monte Carlo experiments.

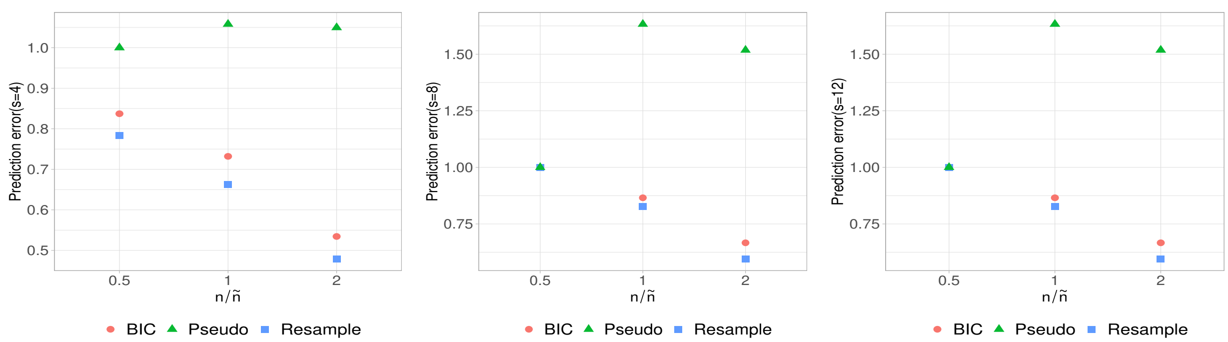

Figure C.2: Prediction errors based on three tuning parameter selection methods in setting 2 with correlated SNPs. Each point is based on 200 Monte Carlo experiments.

We can see from Figure C.1 and Figure C.2 that the resampling-based tuning parameter selection gives to smallest prediction errors in almost all the scenarios. The pseudo-validation performs well when s is large and n is small. When the covariates have correlations, the BIC-based method is comparable to the resampling-based method at different levels of $n/\tilde{n}$ and s. Therefore, we apply resampling-based method for tuning parameter selection in the simulation studies and in real data analysis.

D Testing heritability based on two-sample summary data

The inference of heritability has been studied in the semi-supervised setting based on individual data (Cai and Guo, 2020). With two-sample summary data, we can similarly propose a test statistic

$$\widehat{H} = 2\widehat{M}^T\widehat{\boldsymbol{\beta}}^{(ts)} - (\widehat{\boldsymbol{\beta}}^{(ts)})^T\widetilde{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}}^{(ts)},$$

where $\hat{\boldsymbol{\beta}}^{(ts)}$ is the two-sample Lasso estimate proposed in (2.1). Let $H = \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}$.

Lemma D.1 (Limiting distribution of $\widehat{H}$)**.** *Under the conditions of Theorem 1, for* $\sqrt{n}z/\sigma_h \xrightarrow{D} N(0,1)$ *for* $\sigma_h^2 = 2\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}(\gamma_{n,\tilde{n}} + \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta})$*, we have*

$$\frac{\sqrt{n}(\widehat{H} - H)}{\sigma_h} \xrightarrow{D} N(0,1)$$

given that $s\log p \ll \sqrt{\min(n,\tilde{n})}$ *and* $\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} \gg \frac{\sigma^2(s\log p)^2}{n}$.

Based on Lemma D.1, we estimate the variance via $\hat{\sigma}_h^2 = 2\widehat{H}(2\hat{\sigma}^2 + \widehat{H}(n/\tilde{n}+1))$. The test statistic we use is $\sqrt{n}\widehat{H}/\hat{\sigma}_h$.

Proof of Lemma D.1. We can calculate that

$$\begin{aligned}
\widehat{H} - H &= 2\widehat{M}^T\widehat{\boldsymbol{\beta}}^{(ts)} - (\widehat{\boldsymbol{\beta}}^{(ts)})^T\widetilde{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} \\
&= \frac{2}{n}\langle\epsilon, X\widehat{\boldsymbol{\beta}}^{(ts)}\rangle + 2\boldsymbol{\beta}^T\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}}^{(ts)} - (\widehat{\boldsymbol{\beta}}^{(ts)})^T\widetilde{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} \\
&= \frac{2}{n}\langle\epsilon, X\widehat{\boldsymbol{\beta}}^{(ts)}\rangle + \boldsymbol{\beta}^T(\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\Sigma}\boldsymbol{\beta}) + (\widehat{\boldsymbol{\beta}}^{(ts)})^T(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \widetilde{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}}^{(ts)}) \\
&= \frac{2}{n}\langle\epsilon, X\widehat{\boldsymbol{\beta}}^{(ts)}\rangle + \boldsymbol{\beta}^T\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}) + \boldsymbol{\beta}^T(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta} + (\widehat{\boldsymbol{\beta}}^{(ts)})^T\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(ts)}) \\
&+ (\widehat{\boldsymbol{\beta}}^{(ts)})^T(\widehat{\boldsymbol{\Sigma}} - \widetilde{\boldsymbol{\Sigma}})\widehat{\boldsymbol{\beta}}^{(ts)} \\
&= \frac{2}{n}\langle\epsilon, X\widehat{\boldsymbol{\beta}}^{(ts)}\rangle + (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(ts)})^T\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}) + \boldsymbol{\beta}^T(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta} + (\widehat{\boldsymbol{\beta}}^{(ts)})^T(\widehat{\boldsymbol{\Sigma}} - \widetilde{\boldsymbol{\Sigma}})\widehat{\boldsymbol{\beta}}^{(ts)} \\
&= \frac{2}{n}\langle\epsilon, X\widehat{\boldsymbol{\beta}}^{(ts)}\rangle - (\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})^T\widetilde{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}) + \boldsymbol{\beta}^T(2\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}})\boldsymbol{\beta} + 2(\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})^T(\widehat{\boldsymbol{\Sigma}} - \widetilde{\boldsymbol{\Sigma}})\boldsymbol{\beta} \\
&= \underbrace{\frac{2}{n}\langle\epsilon, X\boldsymbol{\beta}\rangle + \boldsymbol{\beta}^T(2\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}})\boldsymbol{\beta}}_{T_1} \\
&\quad + \underbrace{\frac{2}{n}\langle\epsilon, X(\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})\rangle - (\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})^T\widetilde{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}) + 2(\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})^T(\widehat{\boldsymbol{\Sigma}} - \widetilde{\boldsymbol{\Sigma}})\boldsymbol{\beta}}_{T_2}
\end{aligned}$$

We have proved that $\hat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta} \in \mathcal{C}(S,3)$ where $\mathcal{C}(S,v) = \{u \in \mathbb{R}^p : \|u_S\| \geq v\|u_{S^c}\|_1\}$. For T_2, in the event that

$$\{\|\frac{X^T\epsilon}{n}\|_\infty \leq C\sqrt{\frac{\log p}{n}}, \max_{u\in\mathcal{C}(S,3):\|u\|_2=1} u^T\widetilde{\boldsymbol{\Sigma}}u \leq C\|u\|_2^2, \|(\widehat{\boldsymbol{\Sigma}}-\widetilde{\boldsymbol{\Sigma}})\boldsymbol{\beta}\|_\infty \leq C, \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2\sqrt{\frac{\log p}{n\wedge\tilde{n}}}\},$$

it is easy to prove that

$$\begin{aligned}
|T_2| &\leq 2\|\frac{X^T\epsilon}{n}\|_\infty\|\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}\|_1 + \|\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}\|_2^2 \max_{u\in\mathcal{C}(S,3):\|u\|_2=1} u^T\widetilde{\boldsymbol{\Sigma}}u + 2\|\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}\|_1\|(\widehat{\boldsymbol{\Sigma}} - \widetilde{\boldsymbol{\Sigma}})\boldsymbol{\beta}\|_\infty \\
&\lesssim \sqrt{\frac{\log p}{n}}s\lambda^{(ts)} + s(\lambda^{(ts)})^2 + s\lambda^{(ts)}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2\sqrt{\frac{\log p}{n\wedge\tilde{n}}} \\
&\lesssim \frac{\gamma_{n,\tilde{n}}s\log p}{n}.
\end{aligned}$$

It is easy to show that the above event holds true with probability at least $1-\exp\{-c_1\log p\}$ given that $s\log p = o(\tilde{n})$.

For T_1, We first derive the limiting distribution of $\frac{2}{n}\langle\epsilon, X\boldsymbol{\beta}\rangle + 2\boldsymbol{\beta}^T(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta}$. Note that

$$\frac{2}{n}\langle\epsilon, X\boldsymbol{\beta}\rangle + 2\boldsymbol{\beta}^T(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta} = \frac{2}{n^2}\sum_{i=1}^n\{(x_i^T\boldsymbol{\beta})y_i - \mathbb{E}[(x_i^T\boldsymbol{\beta})y_i]\}.$$

Define

$$\begin{aligned}
(\sigma_h^{(1)})^2 &= \frac{2}{n^2}\sum_{i=1}^n var((x_i^T\boldsymbol{\beta})y_i - \mathbb{E}[(x_i^T\boldsymbol{\beta})y_i]) \\
&= \frac{2}{n}(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}M + (\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta})^2).
\end{aligned}$$

By Lyapunov's central limit theorem, we have for $\sigma_h^{(1)} \neq 0$,

$$\frac{\frac{2}{n}\langle\epsilon, X\boldsymbol{\beta}\rangle + 2\boldsymbol{\beta}^T(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta}}{\sigma_h^{(1)}/\sqrt{n}} \xrightarrow{D} N(0,1).$$

Similarly, we can establish the asymptotic normality of $\boldsymbol{\beta}^T(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta}$. It is easy to show that for $\boldsymbol{\beta} \neq 0$,

$$\frac{\boldsymbol{\beta}^T(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta}}{\sqrt{2(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta})/\tilde{n}}} \xrightarrow{D} N(0,1).$$

Using the independence of (X,y) and $\widetilde{X}$, we have

$$\frac{T_1}{\sigma_h/\sqrt{n}} \xrightarrow{D} N(0,1),$$

where

$$\sigma_h^2 = 2\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}[M + \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}\frac{n}{\tilde{n}}] = 2(\gamma_{n,\tilde{n}} + \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}.$$

Hence, $\widehat{H}$ is asymptotic normal given that $\sigma_h/\sqrt{n} \gg \gamma_{n,\tilde{n}} s \log p/n$. We now derive sufficient conditions for asymptotic normality.

(i) If $\sigma^2 \gtrsim \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}(n/\tilde{n}+2)$, then it suffices to require

$$\sqrt{\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}}\sigma \gg \sigma^2 s \log p/\sqrt{n},$$

which is equivalent to $\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} \gg \frac{\sigma^2(s\log p)^2}{n}$. That is, we need

$$\frac{\sigma^2(s\log p)^2}{n} \ll \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} \lesssim \sigma^2\frac{\min(n,\tilde{n})}{n}.$$

(ii)If $\sigma^2 \lesssim \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}(n/\tilde{n}+2)$, then it suffices to require

$$\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}\sqrt{\frac{n}{\tilde{n}}+2} \gg \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}(n/\tilde{n}+1)s\log p/\sqrt{n},$$

which is equivalent to $s\log p \ll \sqrt{\min(n,\tilde{n})}$ and $\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} \gtrsim \sigma^2\frac{\min(n,\tilde{n})}{n}$.

Hence,

$$\frac{\sqrt{n}(\widehat{H}-H)}{\sigma_h} \xrightarrow{D} N(0,1)$$

given that

$$s\log p \ll \sqrt{\min(n,\tilde{n})} \text{ and } \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} \gg \frac{\sigma^2(s\log p)^2}{n}.$$

□

E Inference with truncated LD matrix

Zou et al. (2022) considers a Ridge-type shrinkage

$$\widetilde{\boldsymbol{\Sigma}}^{(\nu)} = \nu I_p + (1-\nu)\widetilde{\boldsymbol{\Sigma}},$$

where ν is a positive constant. On the other hand, Wen and Stephens (2010); Zhu and Stephens (2017) consider

$$\check{\boldsymbol{\Sigma}}^{(\nu)} = \nu I_p + (1-\nu)\check{\boldsymbol{\Sigma}},$$

where

$$\check{\boldsymbol{\Sigma}}_{i,j} = \begin{cases} \widetilde{\boldsymbol{\Sigma}}_{i,j} & \text{if } i=j \\ \widetilde{\boldsymbol{\Sigma}}_{i,j}\exp\{-\delta_{i,j}\} & \text{if } i\neq j. \end{cases}$$

$\delta_{i,j}$ is related to the recombination rate between SNPs i and j. For SNPs i and j that are distant, $\exp\{-\delta_{i,j}\} \approx 0$.

To demonstrate the effect of truncation, as you suggested, we consider a truncated estimator

$$\check{\boldsymbol{\Sigma}}^{(t)}_{i,j} = \widetilde{\boldsymbol{\Sigma}}_{i,j} \mathbb{1}(|i-j| < t).$$

That is, if SNPs i and j have a distance larger than t, then their correlation is thresholded to 0. We assume $\boldsymbol{\Sigma}$ to have an auto-correlation structure such that $\boldsymbol{\Sigma}_{j,k} = \rho^{|j-k|}$ for some $0 < \rho < 1$. This assumption agrees with common observations in empirical studies. It is easy to show $\check{\boldsymbol{\Sigma}}^{(t)}$ has smaller estimation errors than $\widetilde{\boldsymbol{\Sigma}}$ with proper choice of t. We now study the estimation and inference of $\boldsymbol{\beta}$ with $\check{\boldsymbol{\Sigma}}^{(t)}$ and $\widehat{\boldsymbol{S}}$.

Let

$$\check{\boldsymbol{\beta}}^{(ts)} = \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min} \left\{ \boldsymbol{b}^T \check{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{b} - 2\boldsymbol{b}^T \widehat{\boldsymbol{S}} + \check{\lambda}^{(ts)} \|\boldsymbol{b}\|_1 \right\}.$$

Let $T_j = \{k : |j-k| < t\}$ for $j = 1, \ldots, p$. Let

$$\check{\gamma}_{n,\tilde{n}} = \max_{j\le p} \{ \sqrt{\boldsymbol{\beta}^T_{T_j} \boldsymbol{\Sigma}_{T_j,T_j} \boldsymbol{\beta}_{T_j}} \sqrt{\frac{\log p}{\tilde{n}}} + \|\boldsymbol{\Sigma}_{j,T_j^c}\|_2 \|\boldsymbol{\beta}_{T_j^c}\|_2 \}.$$

Lemma E.1 (Convergence rate with truncated LD matrix)**.** *Assume Conditions 2.1 and 2.2. For* $\check{\lambda}^{(ts)} \ge c_1 \check{\gamma}_{n,\tilde{n}} \sqrt{\log p / n}$ *with large enough* c_1*, if* $s \log p = o(\tilde{n})$*, it holds that*

$$\|\check{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}\|_2^2 \le C \frac{\check{\gamma}_{n,\tilde{n}} s \log p}{n}$$
$$\|\check{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}\|_1 \le C s \sqrt{\frac{\check{\gamma}_{n,\tilde{n}} \log p}{n}}$$

with probability at least $1 - \exp(-c_2 \log p) - \exp(-c_3 \tilde{n})$ *for some positive constants* c_2 *and* c_3*.*

Proof of Lemma E.1. The following oracle inequality holds for the two-sample Lasso:

$$\frac{1}{2}(\hat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})^\intercal \check{\boldsymbol{\Sigma}}^{(t)} (\hat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}) \le |(\hat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})(\widehat{\boldsymbol{S}} - \check{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{\beta})| + \lambda^{(ts)} \|\boldsymbol{\beta}\|_1 - \lambda^{(ts)} \|\hat{\boldsymbol{\beta}}^{(ts)}\|_1.$$

Define an event

$$E_0 = \left\{ \inf_{\|\boldsymbol{u}_{S^c}\|_1 \le 3\|\boldsymbol{u}_S\|_1} \boldsymbol{u}^\intercal \check{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{u} \ge C\|\boldsymbol{u}\|_2^2 - \frac{\log p}{\tilde{n}} \|\boldsymbol{u}\|_1^2, \quad \|\widehat{\boldsymbol{S}} - \check{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{\beta}\|_\infty \le \frac{\lambda^{(ts)}}{2} \right\}.$$

In event E_0, we have

$$\frac{C}{2}\|\hat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})\|_2^2 \le +\frac{3}{2}\lambda^{(ts)} \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_S\|_1 - \frac{\lambda^{(ts)}}{2} \|(\hat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})_{S^c}\|_1 + \frac{4 \log p}{\tilde{n}} \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_S\|_1^2.$$

Standard analysis leads to

$$\|\check{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}\|_2^2 \le Cs(\lambda^{(ts)})^2$$
$$\|\check{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}\|_1 \le Cs\lambda^{(ts)}$$

under the condition $\tilde{n} \gg s \log p$.

We are left to prove $\mathbb{P}(E_0) \to 1$ for $\lambda^{(ts)}$ given in Lemma E.1.

For the second statement in E_0,

$$\|\widehat{\boldsymbol{S}} - \check{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{\beta}\|_\infty \le \|\widehat{\boldsymbol{S}} - \mathbb{E}[\widehat{\boldsymbol{S}}]\|_\infty + \|(\boldsymbol{\Sigma} - \check{\boldsymbol{\Sigma}}^{(t)})\boldsymbol{\beta}\|_\infty. \tag{E.1}$$

By the Bernstein's inequality for sub-exponential random variables,

$$\mathbb{P}\left(\|\widehat{\boldsymbol{S}} - \mathbb{E}[\widehat{\boldsymbol{S}}]\|_\infty \ge \sqrt{\frac{M \log p}{n}}\right) \le C\exp(-c\log p)$$

when $n \gg \log p$.

For the second term of (E.1), we can prove that

$$\begin{aligned}\max_{j\le p} |(\check{\boldsymbol{\Sigma}}^{(t)}_{j,\cdot} - \boldsymbol{\Sigma}_{j,\cdot})\boldsymbol{\beta}| &\le \max_{j\le p}\{|(\check{\boldsymbol{\Sigma}}^{(t)}_{j,T_j} - \boldsymbol{\Sigma}_{j,T_j})\boldsymbol{\beta}_{T_j}| + |(\check{\boldsymbol{\Sigma}}^{(t)}_{j,T_j^c} - \boldsymbol{\Sigma}_{j,T_j^c})\boldsymbol{\beta}_{T_j^c}|\} \\ &= \max_{j\le p}\{|(\check{\boldsymbol{\Sigma}}^{(t)}_{j,T_j} - \boldsymbol{\Sigma}_{j,T_j})\boldsymbol{\beta}_{T_j}| + |\boldsymbol{\Sigma}_{j,T_j^c}\boldsymbol{\beta}_{T_j^c}|\}.\end{aligned}$$

Since ρ is bounded away from 1,

$$|\boldsymbol{\Sigma}_{j,T_j^c}\boldsymbol{\beta}_{T_j^c}| \le \|\boldsymbol{\Sigma}_{j,T_j^c}\|_2\|\boldsymbol{\beta}_{T_j^c}\|_2.$$

Therefore,

$$\mathbb{P}\left(\max_{j\le p} |(\check{\boldsymbol{\Sigma}}^{(t)}_{j,\cdot} - \boldsymbol{\Sigma}_{j,\cdot})\boldsymbol{\beta}| \ge C\max_{j\le p}\{\sqrt{\boldsymbol{\beta}_{T_j}^T\boldsymbol{\Sigma}_{T_j,T_j}\boldsymbol{\beta}_{T_j}}\sqrt{\frac{\log p}{\tilde{n}}} + \|\boldsymbol{\Sigma}_{j,T_j^c}\|_2\|\boldsymbol{\beta}_{T_j^c}\|_2\}\right) \le \exp\{-c_1\log p\}.$$

Therefore, it suffices to take $\check{\lambda}^{(ts)} \ge C\max_{j\le p}\{\sqrt{\frac{M\log p}{n}} + \sqrt{\boldsymbol{\beta}_{T_j}^T\boldsymbol{\Sigma}_{T_j,T_j}\boldsymbol{\beta}_{T_j}}\sqrt{\frac{\log p}{\tilde{n}}} + \|\boldsymbol{\Sigma}_{j,T_j^c}\|_2\|\boldsymbol{\beta}_{T_j^c}\|_2\}$ for some large enough constant C. For such $\lambda^{(ts)}$ the oracle inequality implies that

$$3\|(\check{\boldsymbol{\beta}} - \boldsymbol{\beta})_S\|_1 \ge \|(\check{\boldsymbol{\beta}} - \boldsymbol{\beta})_{S^c}\|_1.$$

Now we prove the first statement in E_0. Let $\boldsymbol{\Sigma}^{(t)}_{j,k} = \boldsymbol{\Sigma}_{j,k}\mathbb{1}(|j-k| \le t)$. Note that

$$\begin{aligned}\boldsymbol{u}^\intercal\check{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{u} &\ge \boldsymbol{u}^\intercal\boldsymbol{\Sigma}^{(t)}\boldsymbol{u} - |\boldsymbol{u}^\intercal(\check{\boldsymbol{\Sigma}}^{(t)} - \boldsymbol{\Sigma})\boldsymbol{u}| \\ &\ge \boldsymbol{u}^\intercal\boldsymbol{\Sigma}\boldsymbol{u} - |\boldsymbol{u}^\intercal(\check{\boldsymbol{\Sigma}}^{(t)} - \boldsymbol{\Sigma})\boldsymbol{u}|.\end{aligned}$$

It is left to bound

$$\sup_{\boldsymbol{u}:\|\boldsymbol{u}_{S^c}\|_1\leq 3\|\boldsymbol{u}_S\|_1} |\boldsymbol{u}^\intercal(\check{\boldsymbol{\Sigma}}^{(t)} - \boldsymbol{\Sigma})\boldsymbol{u}|,$$

where $\mathcal{C}(S,3) = \{u \in \mathbb{R}^p : \|\boldsymbol{u}_S\|_1 \geq 3\|\boldsymbol{u}_{S^c}\|_1\}$. By Lemma 14 of Rudelson and Zhou (2012), we have for some constant $0 < \delta < 1$,

$$C(S,3) \cap \mathbb{S}^{p-1} \subseteq (1-\delta)^{-1}\text{conv}\left(\cup_{|J|\leq d} V_J \cap \mathbb{S}^{p-1}\right)$$

for $\mathbb{S} = \{u \in \mathbb{R}^p : \|u\|_2 = 1\}$, $d = Cs$ where C is a constant depending on δ, V_J is the space spanned by $\{e_j\}_{j\in J}$, and e_j is the canonical basis. Therefore,

$$\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}\cap C(S,3)} |\boldsymbol{u}^\intercal(\check{\boldsymbol{\Sigma}}^{(t)} - \boldsymbol{\Sigma})\boldsymbol{u}| \leq \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}\cap C(S,3),\boldsymbol{v}\in\mathbb{S}^{p-1}\cap C(S,3)} |\boldsymbol{u}^\intercal(\check{\boldsymbol{\Sigma}}^{(t)} - \boldsymbol{\Sigma})\boldsymbol{v}|.$$

Applying the arguments in Lemma 16 of Bradic et al. (2019), we have

$$\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}\cap C(S,3)} |\boldsymbol{u}^\intercal(\check{\boldsymbol{\Sigma}}^{(t)} - \boldsymbol{\Sigma}^{(t)})\boldsymbol{u}| \leq c_1 \max_{\boldsymbol{u}\in\cup_{|J|\leq d}V_J\cap\mathbb{S}^{p-1}} |\boldsymbol{u}^\intercal(\check{\boldsymbol{\Sigma}}^{(t)} - \boldsymbol{\Sigma}^{(t)})\boldsymbol{u}|,$$

where $d \asymp s$ depending on δ,V_J is the space spanned by $\{e_j\}_{j\in J}$ and e_j denotes the j-th canonical basis. It is easy to show that for fixed $\boldsymbol{u} \in V_J \cap \mathbb{S}^{p-1}$ and $|J| = d$,

$$\mathbb{P}(|\boldsymbol{u}^\intercal(\check{\boldsymbol{\Sigma}}^{(t)} - \boldsymbol{\Sigma}^{(t)})\boldsymbol{u}| \geq v) \leq \exp\{-c_1 \min(\frac{\tilde{n}v^2}{c\Lambda_{\max}(\boldsymbol{\Sigma})}, \frac{\tilde{n}v}{\Lambda_{\max}(\boldsymbol{\Sigma})})\}.$$

Therefore,

$$\begin{aligned}
&\mathbb{P}(\sup_{u\in\cup_{|J|\leq d}V_J\cap\mathbb{S}^{p-1}} |\boldsymbol{u}^\intercal(\check{\boldsymbol{\Sigma}}^{(t)} - \boldsymbol{\Sigma}^{(t)})\boldsymbol{u}| \geq v) \leq \exp\{-c_1 \min(\frac{\tilde{n}v^2}{c\Lambda_{\max}(\boldsymbol{\Sigma})}, \frac{\tilde{n}v}{\Lambda_{\max}(\boldsymbol{\Sigma})})\} \\
&\leq \sum_{k=1}^{d} \binom{p}{k} \exp\{k\} \sup_{u\in\cup_{|J|\leq d}V_J\cap\mathbb{S}^{p-1}} \mathbb{P}(\boldsymbol{u}^\intercal(\check{\boldsymbol{\Sigma}}^{(t)} - \boldsymbol{\Sigma}^{(t)})\boldsymbol{u}| \geq v) \\
&\leq \exp\{c_2 d\log p\} \exp\{-c_1 \min(\frac{\tilde{n}v^2}{c\Lambda_{\max}(\boldsymbol{\Sigma})}, \frac{\tilde{n}v}{\Lambda_{\max}(\boldsymbol{\Sigma})})\}.
\end{aligned}$$

For $v \geq c_1\sqrt{s\log p/\tilde{n}}$, we have shown that

$$\mathbb{P}(\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}\cap C(S,3)} |\boldsymbol{u}^\intercal(\check{\boldsymbol{\Sigma}}^{(t)} - \boldsymbol{\Sigma})\boldsymbol{u}| \geq c_1\sqrt{\frac{s\log p}{n}}) \leq \exp\{-c_1 s\log p\}$$

given that $\tilde{n} \gg s\log p$. □

We see that if $\max_{j\leq p}\|\boldsymbol{\Sigma}_{j,T_j^c}\|_2$ is sufficiently small, then $\check{\gamma}_{n,\tilde{n}}$ can be smaller than $\gamma_{n,\tilde{n}}$ and hence, $\check{\boldsymbol{\beta}}^{(ts)}$ can have a faster convergence rate than $\hat{\boldsymbol{\beta}}^{(ts)}$. If $\|\boldsymbol{\Sigma}_{j,T_j^c}\|_2$ is large, then the truncation can lead to significant bias and $\check{\boldsymbol{\beta}}^{(ts)}$ can be worse than $\hat{\boldsymbol{\beta}}^{(ts)}$.

Now consider the debiased Lasso estimator based on $\check{\boldsymbol{\beta}}^{(ts)}$.

$$\check{\beta}_j^{(ts-db)} = \check{\beta}_j^{(ts)} + \check{\boldsymbol{w}}_j^{\intercal}(\widehat{\boldsymbol{S}} - \check{\boldsymbol{\Sigma}}^{(t)}\check{\boldsymbol{\beta}}^{(ts)}), \tag{E.2}$$

where

$$\check{\boldsymbol{w}}_j = \mathop{\arg\min}_{\boldsymbol{w}\in\mathbb{R}^p}\|\boldsymbol{w}\|_1 \quad \text{subject to } \|\check{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{w} - e_j\|_\infty \leq \check{\lambda}_j. \tag{E.3}$$

Theorem E.1 (Limiting distribution of two-sample debiased Lasso)**.** *Assume that Conditions 2.1 and 2.2 hold and $\tilde{n} \gg (s\vee s_j)\log p$. Then it holds that*

$$\check{\beta}_j^{(ts-db)} - \beta_j = b_j + (\check{V}_j^{(ts)})^{1/2}z_j + O_P\left((s+s_j)\check{\lambda}^{(ts)}\sqrt{\frac{\log p}{\tilde{n}}}\right), \tag{E.4}$$

where $b_j = \boldsymbol{w}_j^T(\boldsymbol{\Sigma}^{(t)}-\boldsymbol{\Sigma})\boldsymbol{\beta}$, $z_j \xrightarrow{D} N(0,1)$, and

$$\check{V}_j^{(ts)} = \frac{\Omega_{j,j}M+\beta_j^2}{n} + \frac{\sum_{k,k'\leq p}\{(\boldsymbol{w}_j)_k\boldsymbol{\Sigma}_{k,k'}w_{k',j}\boldsymbol{\beta}_{T_k}^T\boldsymbol{\Sigma}_{T_k,T_k'}\boldsymbol{\beta}_{T_{k'}} + (\boldsymbol{w}_j)_k\boldsymbol{\Sigma}_{k,T_{k'}}\boldsymbol{\beta}_{T_{k'}}(\boldsymbol{w}_j)_{k'}\boldsymbol{\Sigma}_{k',T_k}\boldsymbol{\beta}_{T_k}\}}{\tilde{n}}.$$

Proof of Theorem E.1. We first bound $\|\check{\boldsymbol{w}}_j - \boldsymbol{w}_j\|_1$ for $\boldsymbol{w}_j = \{\boldsymbol{\Sigma}^{(t)}\}^{-1}_{\cdot,j}$. We first show that $\boldsymbol{w}_j$ is feasible to (E.3) for $\check{\lambda}_j \geq C\sqrt{\log p/\tilde{n}}$. Note that

$$\begin{aligned}\|\check{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{w} - e_j\|_\infty &\leq \|(\check{\boldsymbol{\Sigma}}^{(t)}-\boldsymbol{\Sigma})\boldsymbol{w}_j\|_\infty \\ &\leq C\|\boldsymbol{w}_j\|_2\sqrt{\frac{\log p}{\tilde{n}}} \leq C'\sqrt{\frac{\log p}{\tilde{n}}},\end{aligned}$$

with probability at least $1-\exp\{-c_1\log p\}$. Hence, standard analysis leads to

$$\|\hat{\boldsymbol{w}}_j - \boldsymbol{w}_j\|_1 \leq C\|\boldsymbol{w}_j\|_0\sqrt{\frac{\log p}{\tilde{n}}} \text{ and } \|\hat{\boldsymbol{w}}_j - \boldsymbol{w}_j\|_2^2 \leq C\|\boldsymbol{w}_j\|_0\frac{\log p}{\tilde{n}}$$

with probability at least $1-\exp\{-c_1\log p\}$.

$$
\begin{aligned}
\check{\beta}_j^{(ts-db)} - \beta_j &= (e_j - \check{\boldsymbol{w}}_j^{\intercal}\check{\boldsymbol{\Sigma}}^{(t)})(\check{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}) + \check{\boldsymbol{w}}_j^{\intercal}(\widehat{\boldsymbol{\Sigma}} - \check{\boldsymbol{\Sigma}}^{(t)})\boldsymbol{\beta} + \frac{\check{\boldsymbol{w}}_j^T X^T \epsilon}{n} \\
&= \underbrace{(e_j - \check{\boldsymbol{w}}_j^{\intercal}\check{\boldsymbol{\Sigma}}^{(t)})(\check{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})}_{R_1} \\
&\quad + \underbrace{(\check{\boldsymbol{w}}_j - \boldsymbol{w}_j)^{\intercal}(\widehat{\boldsymbol{\Sigma}} - \check{\boldsymbol{\Sigma}}^{(t)})\boldsymbol{\beta} + \frac{(\check{\boldsymbol{w}}_j - \boldsymbol{w}_j)^T X^T \epsilon}{n}}_{R_2} \\
&\quad + \underbrace{\boldsymbol{w}_j^{\intercal}(\widehat{\boldsymbol{\Sigma}} - \check{\boldsymbol{\Sigma}}^{(t)})\boldsymbol{\beta} + \frac{\boldsymbol{w}_j^T X^T \epsilon}{n}}_{R_3}
\end{aligned}
$$

It is easy to show that

$$|R_1| \le \check{\lambda}_j \|\check{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}\|_1 \le s\check{\lambda}_j\check{\lambda}^{(ts)}.$$

For R_2, we have

$$
\begin{aligned}
|R_2| &\le \|\check{\boldsymbol{w}}_j - \boldsymbol{w}_j\|_1 \|(\widehat{\boldsymbol{\Sigma}} - \check{\boldsymbol{\Sigma}}^{(t)})\boldsymbol{\beta}\|_\infty + \|\check{\boldsymbol{w}}_j - \boldsymbol{w}_j\|_1 \|\frac{X^T\epsilon}{n}\|_\infty \\
&\le \|\boldsymbol{w}_j\|_0 \check{\lambda}_j \check{\lambda}^{(ts)}.
\end{aligned}
$$

For R_3,

$$R_3 = \boldsymbol{w}_j^T \widehat{M} - \mathbb{E}[\boldsymbol{w}_j^T \widehat{M}] - \boldsymbol{w}_j^T(\check{\boldsymbol{\Sigma}}^{(t)} - \boldsymbol{\Sigma})\boldsymbol{\beta},$$

where the first and second terms are independent. The first term is asymptotic normal with mean zero and variance

$$\frac{\boldsymbol{w}_j^T \boldsymbol{\Sigma} \boldsymbol{w}_j M + \beta_j^2}{n}.$$

For the second term,

$$
\begin{aligned}
\boldsymbol{w}_j^T(\check{\boldsymbol{\Sigma}}^{(t)} - \boldsymbol{\Sigma})\boldsymbol{\beta} &= \boldsymbol{w}_j^T(\check{\boldsymbol{\Sigma}}^{(t)} - \boldsymbol{\Sigma}^{(t)})\boldsymbol{\beta} + \boldsymbol{w}_j^T(\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma})\boldsymbol{\beta} \\
&= \sum_{k=1}^{p} (\boldsymbol{w}_j)_k (\widetilde{\boldsymbol{\Sigma}}_{k,T_k} - \boldsymbol{\Sigma}_{k,T_k})\boldsymbol{\beta}_{T_k} + \boldsymbol{w}_j^T(\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma})\boldsymbol{\beta}.
\end{aligned}
$$

It is easy to see that it is asymptotic normal with mean $\boldsymbol{w}_j^T(\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma})\boldsymbol{\beta}$ and variance

$$
\begin{aligned}
var(\sum_{k=1}^{p} (\boldsymbol{w}_j)_k \widetilde{\boldsymbol{\Sigma}}_{k,T_k}\boldsymbol{\beta}_{T_k}) &= var(\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \sum_{k=1}^{p} \boldsymbol{w}_{k,j} \boldsymbol{x}_{i,k} \boldsymbol{x}_{i,T_k}^T \boldsymbol{\beta}_{T_k}) \\
&= \frac{1}{\tilde{n}} var(\sum_{k=1}^{p} \boldsymbol{w}_{k,j} \boldsymbol{x}_{i,k} \boldsymbol{x}_{i,T_k}^T \boldsymbol{\beta}_{T_k}) \\
&= \frac{\sum_{k,k' \le p} \{(\boldsymbol{w}_j)_k \boldsymbol{\Sigma}_{k,k'} (\boldsymbol{w}_j)_{k'} \boldsymbol{\beta}_{T_k}^T \boldsymbol{\Sigma}_{T_k,T_{k'}} \boldsymbol{\beta}_{T_{k'}} + (\boldsymbol{w}_j)_k \boldsymbol{\Sigma}_{k,T_{k'}} \boldsymbol{\beta}_{T_{k'}} (\boldsymbol{w}_j)_{k'} \boldsymbol{\Sigma}_{k',T_k} \boldsymbol{\beta}_{T_k}\}}{\tilde{n}}.
\end{aligned}
$$

To summarize,

$$\check{\beta}_j - \beta_j = \check{\sigma}_j z + b_j + O_P((s + s_j)\check{\lambda}_j \check{\lambda}^{(ts)}),$$

where $z \xrightarrow{D} N(0,1)$, $b_j = \boldsymbol{w}_j^T(\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma})\boldsymbol{\beta}$, and

$$\check{V}_j^{(ts)} = \frac{\Omega_{j,j} M + \beta_j^2}{n} + \frac{\sum_{k,k' \le p}\{(\boldsymbol{w}_j)_k \boldsymbol{\Sigma}_{k,k'}(\boldsymbol{w}_j)_{k'} \boldsymbol{\beta}_{T_k}^T \boldsymbol{\Sigma}_{T_k,T_{k'}} \boldsymbol{\beta}_{T_{k'}} + (\boldsymbol{w}_j)_k \boldsymbol{\Sigma}_{k,T_{k'}} \boldsymbol{\beta}_{T_{k'}} (\boldsymbol{w}_j)_{k'} \boldsymbol{\Sigma}_{k',T_k} \boldsymbol{\beta}_{T_k}\}}{\tilde{n}}.$$

□

We see that $\check{\beta}_j^{(ts-db)}$ has an asymptotic normal distribution. Compared to $\hat{\beta}_j^{(ts-db)}$, $\check{\beta}_j^{(ts-db)}$ has an extra bias term b_j due to the truncation step. Again, if $\max_{j \le p} \|\boldsymbol{\Sigma}_{j,T_j^c}\|_2$ is sufficiently small, then b_j is dominated by other terms. If this is true, then $\check{\beta}_j^{(ts-db)}$ has smaller bias and variance in comparison to $\hat{\beta}_j^{(ts-db)}$. Otherwise, b_j can be dominant and $\check{\beta}_j^{(ts-db)}$ cannot be used for inference purposes.

References

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics 37*(4), 1705–1732.

Bowden, J., G. Davey Smith, P. C. Haycock, and S. Burgess (2016). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic epidemiology 40*(4), 304–314.

Bowden, J., F. Del Greco M, C. Minelli, Q. Zhao, D. A. Lawlor, N. A. Sheehan, J. Thompson, and G. Davey Smith (2019). Improving the accuracy of two-sample summary-data mendelian randomization: moving beyond the nome assumption. *International Journal of Epidemiology 48*(3), 728–742.

Bradic, J., S. Wager, and Y. Zhu (2019). Sparsity double robust inference of average treatment effects. *arXiv preprint arXiv:1905.00744*.

Cai, T. T. and Z. Guo (2020). Semisupervised inference for explained variance in high dimensional linear regression and its applications. *Journal of the Royal Statistical Society Series B: Statistical Methodology 82*(2), 391–419.

Elsworth, B., M. Lyon, T. Alexander, Y. Liu, P. Matthews, J. Hallett, P. Bates, T. Palmer, V. Haberland, G. D. Smith, et al. (2020). The mrc ieu opengwas data infrastructure. *BioRxiv*, 2020–08.

Mak, T. S. H., R. M. Porsch, S. W. Choi, X. Zhou, and P. C. Sham (2017). Polygenic scores via penalized regression on summary statistics. *Genetic epidemiology 41*(6), 469–480.

Raskutti, G., M. J. Wainwright, and B. Yu (2010). Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research 11*, 2241–2259.

Raskutti, G., M. J. Wainwright, and B. Yu (2011). Minimax rates of estimation for high-dimensional linear regression over lq-balls. *IEEE transactions on information theory 57*(10), 6976–6994.

Rudelson, M. and S. Zhou (2012). Reconstruction from anisotropic random measurements. In *Conference on Learning Theory*, pp. 10–1. JMLR Workshop and Conference Proceedings.

Tsybakov, A. B. (2008). *Introduction to nonparametric estimation.* Springer Science & Business Media.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*(3), 1166–1202.

Wen, X. and M. Stephens (2010). Using linear predictors to impute allele frequencies from summary or pooled genotype data. *The Annals of Applied Statistics 4*(3), 1158.

Zhu, X. and M. Stephens (2017). Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *The Annals of Applied Statistics 11*(3), 1561.

Zou, Y., P. Carbonetto, G. Wang, and M. Stephens (2022). Fine-mapping from summary data with the "sum of single effects" model. *PLoS Genetics 18*(7), e1010299.