Supplementary Material for "Statistical Inference on High Dimensional Covariate-Dependent Gaussian Graphical Regressions"

Xuran Meng

Department of Biostatistics, University of Michigan, Ann Arbor, USA
email: xuranm@umich.edu

and

Jingfei Zhang

Department of Information Systems and Operations Management, Emory University, Atlanta, USA
email: emma.zhang@emory.edu

and

Yi Li

Department of Biostatistics, University of Michigan, Ann Arbor, USA
email: yili@umich.edu

S1. Analysis of Debiasing Optimization

We consider the two optimization problems, i.e., (6) and (7), for obtaining $\widehat{\mathbf{M}}_j$. To proceed, recall that singular value decomposition gives $\mathbf{W}_j/\sqrt{n} = \mathbf{U}_j\mathbf{D}_j^{1/2}\mathbf{V}_j^\top$, where $\mathbf{U}_j \in \mathbb{R}^{n\times n}$ and $\mathbf{V}_j \in \mathbb{R}^{p\times n}$ are orthonormal so that $\mathbf{V}_j^\top\mathbf{V}_j = \mathbf{I}_n = \mathbf{U}_j^\top\mathbf{U}_j = \mathbf{U}_j\mathbf{U}_j^\top$. Hence, $\widehat{\mathbf{\Sigma}}_{\mathbf{W}_j} = \mathbf{W}_j^\top\mathbf{W}_j/n = \mathbf{V}_j\mathbf{D}_j\mathbf{V}_j^\top$.

Proposition 1 establishes a connection between these two optimization problems. It becomes clear that the analysis of (6) will be much simpler than that of (7), and proving the existence of $\boldsymbol{\theta}^*$ in (7) is equivalent to proving the existence of a solution in (6). Moreover, once we obtain the solution $\boldsymbol{\theta}^*$ from (7), we can apply $\mathbf{V}_j\boldsymbol{\theta}^*$ to derive a solution for (6). This approach can substantially reduce computational time, especially as the dimension of $\widehat{\mathbf{\Sigma}}_{\mathbf{W}_j}$ increases. The computational cost will primarily depend on the sample size and remain constant as the dimension increases.

Proof. [Proof of Proposition 1] We first prove that if $\mathbf{m}^*$ is a solution of (6), then $\boldsymbol{\theta}^* = \mathbf{V}_j^\top\mathbf{m}^*$ is a solution of (7). It is easy to verify that

$$\|H_\alpha(\mathbf{V}_j\mathbf{D}_j\boldsymbol{\theta}^* - \mathbf{e}_l)\|_{\infty,2} = \|H_\alpha(\mathbf{V}_j\mathbf{D}_j\mathbf{V}_j^\top\mathbf{m}^* - \mathbf{e}_l)\|_{\infty,2} = \|H_\alpha(\widehat{\mathbf{\Sigma}}_{\mathbf{W}_j}\mathbf{m}^* - \mathbf{e}_l)\|_{\infty,2} \leqslant \gamma,$$

where the inequality follows because $\mathbf{m}^*$ is a solution of (6). We conclude that $\boldsymbol{\theta}^*$ satisfies the constraint in (7). For any $\boldsymbol{\theta}$ that belongs to the constraint set in (7), we can always find $\mathbf{m} = \mathbf{V}_j\boldsymbol{\theta}$ such that

$$\|H_\alpha(\widehat{\mathbf{\Sigma}}_{\mathbf{W}_j}\mathbf{m} - \mathbf{e}_l)\|_{\infty,2} = \|H_\alpha(\mathbf{V}_j\mathbf{D}_j\mathbf{V}_j^\top\mathbf{V}_j\boldsymbol{\theta} - \mathbf{e}_l)\|_{\infty,2} = \|H_\alpha(\mathbf{V}_j\mathbf{D}_j\boldsymbol{\theta} - \mathbf{e}_l)\|_{\infty,2} \leqslant \gamma.$$

Hence, for any $\boldsymbol{\theta}$ that satisfies the constraint in (7), there exists an $\mathbf{m}$ satisfying the constraint in (6) such that

$$\begin{aligned}\boldsymbol{\theta}^{*\top}\mathbf{D}_j\boldsymbol{\theta}^* &= \boldsymbol{\theta}^{*\top}\mathbf{V}_j^\top\mathbf{V}_j\mathbf{D}_j\mathbf{V}_j^\top\mathbf{V}_j\boldsymbol{\theta}^* \\ &= \boldsymbol{\theta}^{*\top}\mathbf{V}_j^\top\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j}\mathbf{V}_j\boldsymbol{\theta}^* \\ &= \mathbf{m}^{*\top}\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j}\mathbf{m}^* \\ &\leqslant \mathbf{m}^\top\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j}\mathbf{m} \\ &= \boldsymbol{\theta}^\top\mathbf{V}_j^\top\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j}\mathbf{V}_j\boldsymbol{\theta} = \boldsymbol{\theta}^\top\mathbf{D}_j\boldsymbol{\theta}.\end{aligned}$$

The first equality is by $\mathbf{V}_j^\top\mathbf{V}_j = \mathbf{I}_n$, the second equality is by $\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j} = \mathbf{V}_j\mathbf{D}_j\mathbf{V}_j^\top$, the third equality is by $\mathbf{m}^* = \mathbf{V}_j\boldsymbol{\theta}^*$, the first inequality comes from $\mathbf{m}^*$ is the solution of (6) and the last two equalities are by direct calculations. We conclude that for any $\boldsymbol{\theta}$ satisfying the constraint in (7), $\boldsymbol{\theta}^{*\top}\mathbf{D}_j\boldsymbol{\theta}^* \leqslant \boldsymbol{\theta}^\top\mathbf{D}_j\boldsymbol{\theta}$.

Next, we prove that if $\boldsymbol{\theta}^*$ is a solution of (7), then $\mathbf{m}^* = \mathbf{V}_j\boldsymbol{\theta}^*$ is a solution of (6). Let $\widetilde{\mathbf{m}}$ be a solution of (6), implying

$$\|H_\alpha(\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j}\widetilde{\mathbf{m}} - \mathbf{e}_l)\|_{\infty,2} \leqslant \gamma,$$

and $\widetilde{\mathbf{m}}^\top\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j}\widetilde{\mathbf{m}}$ has the minimum value in this constraint. Further denote by $\boldsymbol{\alpha}_1 = \mathbf{V}_j^\top\widetilde{\mathbf{m}}$, it can be seen that

$$\|H_\alpha(\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j}\widetilde{\mathbf{m}} - \mathbf{e}_l)\|_{\infty,2} = \|H_\alpha(\mathbf{V}_j\mathbf{D}_j\mathbf{V}_j^\top\widetilde{\mathbf{m}} - \mathbf{e}_l)\|_{\infty,2} = \|H_\alpha(\mathbf{V}_j\mathbf{D}_j\boldsymbol{\alpha}_1 - \mathbf{e}_l)\|_{\infty,2} \leqslant \gamma,$$

which means $\boldsymbol{\alpha}_1 = \mathbf{V}_j^\top\widetilde{\mathbf{m}}$ satisfies the constraint in (7). Note that $\boldsymbol{\theta}^*$ is a solution of (7), we conclude that

$$\begin{aligned}\mathbf{m}^{*\top}\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j}\mathbf{m}^* &= \boldsymbol{\theta}^{*\top}\mathbf{V}_j^\top\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j}\mathbf{V}_j\boldsymbol{\theta}^* \\ &= \boldsymbol{\theta}^{*\top}\mathbf{V}_j^\top\mathbf{V}_j\mathbf{D}_j\mathbf{V}_j^\top\mathbf{V}_j\boldsymbol{\theta}^* \\ &= \boldsymbol{\theta}^{*\top}\mathbf{D}_j\boldsymbol{\theta}^* \\ &\leqslant \boldsymbol{\alpha}_1^\top\mathbf{D}_j\boldsymbol{\alpha}_1 \\ &= \widetilde{\mathbf{m}}^\top\mathbf{V}_j\mathbf{D}_j\mathbf{V}_j^\top\widetilde{\mathbf{m}} = \widetilde{\mathbf{m}}^\top\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j}\widetilde{\mathbf{m}}.\end{aligned}$$

The first equality is by $\mathbf{m}^* = \mathbf{V}_j\boldsymbol{\theta}^*$, the second equality is by $\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j} = \mathbf{V}_j\mathbf{D}_j\mathbf{V}_j^\top$, the third equality is by $\mathbf{V}_j^\top \mathbf{V}_j = \mathbf{I}_n$, the first inequality comes from $\boldsymbol{\theta}^*$ is a solution of (7) and $\boldsymbol{\alpha}_1$ satisfies the constraint in (7), the fourth equality is by $\mathbf{V}_j^\top \widetilde{\mathbf{m}}$ and the last equality is by $\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j} = \mathbf{V}_j \mathbf{D}_j \mathbf{V}_j^\top$. We conclude that

$$\mathbf{m}^{*\top} \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j} \mathbf{m}^* \leqslant \widetilde{\mathbf{m}}^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j} \widetilde{\mathbf{m}}.$$

Moreover, it holds that

$$\|H_\alpha(\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j} \mathbf{m}^* - \mathbf{e}_l)\|_{\infty,2} = \|H_\alpha(\mathbf{V}_j \mathbf{D}_j \mathbf{V}_j^\top \mathbf{V}_j \boldsymbol{\theta}^* - \mathbf{e}_l)\|_{\infty,2} = \|H_\alpha(\mathbf{V}_j \mathbf{D}_j \boldsymbol{\theta}^* - \mathbf{e}_l)\|_{\infty,2} \leqslant \gamma,$$

which means $\mathbf{m}^*$ satisfies the constraint of (6). Given that $\widetilde{\mathbf{m}}$ is a solution of (6), $\mathbf{m}^{*\top} \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j} \mathbf{m}^* \leqslant \widetilde{\mathbf{m}}^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j} \widetilde{\mathbf{m}}$ and $\mathbf{m}^*$ satisfies the constraint of (6), we conclude that $\mathbf{m}^*$ is a solution of (6). This completes the proof of Proposition 1.

Remark 1: The construction of $\boldsymbol{\Xi}_j = \frac{\mathbf{W}_j \mathbf{W}_j^\top}{n}$ and $\mathbf{V}_j$ helps reduce computational time. We utilize MATLAB's `fmincon` function for solving constraint optimization problems. When applying the debiased methods outlined in Section 3 to solve the matrix $\widehat{\mathbf{M}}_j$, we need to solve (7) $(p-1)(q+1)$ times. If each optimization process for $l \in [(p-1)(q+1)]$ requires T iterations, the total computational cost, including the decomposition of $\boldsymbol{\Xi}_j$, is $O(n^3 + nTpq)$. Here, n^3 corresponds to the time needed for eigenvalue decomposition, and $nTpq$ reflects the optimization time, where each iteration costs $O(n)$. Solving (7) across all $l \in [(p-1)(q+1)]$ yields a total cost of $O(nTpq)$. In contrast, directly solving (6) incurs a computational cost of at least $O(p^2q^2)$ due to the complexity of the objective function and constraints. When factoring in the optimization iterations and the entire set $l \in [(p-1)(q+1)]$, this cost increases to $O(p^3q^3T)$, much higher than the cost of the proposed debiased method.

The proposed debiased procedure also allows us to focus only on the specific parameters of interest, achieving further computational efficiency. For instance, if we are only interested in s parameters in $(\boldsymbol{\beta}_j)_{\mathcal{S}}$, where $s = |\mathcal{S}|$, we only need to compute $\widehat{\mathbf{M}}_{j\mathcal{S}}$. This reduces the computational complexity to $n^3 + nTs$. In contrast, directly optimizing (6) would still require at least $O(sTp^2q^2)$, which is substantially more intensive.

REMARK 2: We note the work of Banerjee et al. (2025), which also exploits low-rank structures to accelerate debiasing. However, key differences prevent direct applications of their framework to our setting. First, our estimator uses a sparse group Lasso penalty, inducing constraints outside their analysis. Second, in our covariate-dependent Gaussian setting, $\mathbf{W}_j$ can be heavy-tailed with non-diagonal covariance and unbounded eigenvalues, violating key regularity conditions (e.g., Assumptions D.2–D.3 in Banerjee et al. (2025)). Hence, their concentration-based results may not apply, and we instead develop a structured projection approach tailored to this framework.

S2. Proof of Theorem 2

We present the following lemmas before proving Theorem 2. First, Lemma 1 gives the property of $H_\alpha(\cdot)$, which is used in the proof of Theorem 2.

LEMMA 1 (Lemma 8 in Cai et al. (2022)): *1. Suppose $a, b > 0, x, y \in \mathbb{R}, H.(\cdot)$ is the soft-threshold operator satisfying $H_a(x) = \mathrm{sgn}(x) \cdot (|x| - a)_+$. Then the following triangular inequality holds,*

$$|H_{a+b}(x+y)| \leqslant |H_a(x)| + |H_b(y)|.$$

2. Suppose $a, b > 0, \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, if $\|H_a(\mathbf{x})\|_{\infty,2} \leqslant b$, then

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leqslant a\|\mathbf{y}\|_1 + b\|\mathbf{y}\|_{1,2}.$$

Below, Lemma 2 shows the existence of $\widehat{\mathbf{M}}_j$, and Lemma 3 shows that $\mathbf{e}_l^\top \frac{\mathbf{W}_j^\top \mathbf{W}_j}{n} \mathbf{e}_l$ has the lower bound with high probability, which are also applied in the proof of Theorem 2.

LEMMA 2: *Under the conditions of Theorem 2 and given $\mathbf{U} \in \mathbb{R}^{n \times (q+1)}$, for $n > q + 1$, there exists a matrix $\mathbf{A}_l$ depend on $\mathbf{U}$, with probability at least $1 - \exp\left\{ -C'' \frac{s_e \log(epq)}{s_g} \right\}$ for some constant $C'' > 0$, the following inequalities holds:*

$$\max_{1 \leqslant l \leqslant (p-1)(q+1)} \left\| H_\alpha \left(\mathbf{e}_l - \frac{\mathbf{W}_j^\top \mathbf{W}_j}{n} \mathbf{A}_l \mathbf{e}_l \right) \right\|_{\infty,2} \leqslant \gamma.$$

The proof of Lemma 2 is given in Section S2.1. We can conclude the solution of (6) exists with Lemma 2 conditional on $\mathbf{u}$. Moreover, since the concentration bound in Lemma 2 holds uniformly when $\mathbf{U}$ is bounded, the existence result extends to the unconditional setting. Then the solution of (7) also exists by Proposition 1.

LEMMA 3: *Under the conditions of Theorem 2, with probability at least*

$$1 - C_1 \exp\Big\{ - C_2 \frac{s_e \log(ep) + s_g \log(eq/s_g)}{s_e} \Big\}$$

for some constants $C_1, C_2 > 0$*, the following inequality holds:*

$$\left| \mathbf{e}_l^\top \frac{\mathbf{W}_j^\top \mathbf{W}_j}{n} \mathbf{e}_l - \mathbb{E}\, \mathbf{e}_l^\top \frac{\mathbf{W}_j^\top \mathbf{W}_j}{n} \mathbf{e}_l \right| \precsim \frac{\log^2 n \cdot \{s_e \log(epq) + s_g \log(eq^2/s_g)\}^2}{s_e^2 n}.$$

With the same probability it holds that

$$\phi_1/(2\phi_0) \leqslant \mathbf{e}_l^\top \frac{\mathbf{W}_j^\top \mathbf{W}_j}{n} \mathbf{e}_l \leqslant M^2 \phi_2 + 1.$$

The proof of Lemma 3 is given in Section S2.2. In the following, Lemmas 4 and 5 give some properties on the vector $\boldsymbol{\delta}_j = \widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j$.

LEMMA 4 (Improved Theorem 3 in Zhang and Li (2025)): *Let* $\boldsymbol{\delta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$*. Under the conditions of Theorem 2, there exist constants* $C_1, C_2 > 0$ *such that, with probability at least* $1 - C_1 \exp[-C_2\{s_e \log(ep) + s_g \log(eq/s_g)\}/s_e]$*, the following holds:*

$$\|\boldsymbol{\delta}\|_2^2 = O\bigg(\frac{s_e \log(ep) + s_g \log(eq/s_g)}{n} \bigg).$$

LEMMA 5: *Under the conditions of Theorem 2, there exist constants* $C_1, C_2 > 0$ *such that, with probability at least* $1 - C_1 \exp[-C_2\{s_e \log(ep) + s_g \log(eq/s_g)\}/s_e]$*, the following holds:*

$$\frac{\|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1}{\sqrt{s_e}} + \frac{\|\boldsymbol{\delta}_{(\mathcal{G}^c)}\|_{1,2}}{\sqrt{s_g}} \leqslant 7\sqrt{\frac{2s_e \log(ep) + s_g \log(eq/s_g)}{n}} + \frac{\|\boldsymbol{\delta}_{\mathcal{S}}\|_1}{\sqrt{s_e}} + \frac{\|(\boldsymbol{\delta})_{(\mathcal{G})}\|_{1,2}}{\sqrt{s_g}}.$$

The proofs of Lemmas 4 and 5 are given in Section S2.3. We next give the following lemma, which provides a consistent estimator of $\frac{1}{\sigma^{jj}}$ by using $\widehat{\boldsymbol{\beta}}$.

LEMMA 6 (Consistent estimator for σ^{jj}): *Under the conditions of Theorem 2, define* $\frac{1}{\widehat{\sigma}^{jj}} = \|\mathbf{z}_j - \mathbf{W}_j\widehat{\boldsymbol{\beta}}_j^{OLS}\|_2^2/(n-\widehat{s}_j)$*, where* $\widehat{s}_j$ *is the number of non-zero values of* $\widehat{\boldsymbol{\beta}}_j^{mul}$*, and* $\widehat{\boldsymbol{\beta}}_j^{OLS}$ *is the OLS estimator constrained on the set* $\widehat{\mathcal{S}}_j$ *satisfying* $\left(\widehat{\boldsymbol{\beta}}_j^{OLS}\right)_{\widehat{\mathcal{S}}_j} = \left\{(\mathbf{W}_j)_{\widehat{\mathcal{S}}_j}^\top(\mathbf{W}_j)_{\widehat{\mathcal{S}}_j}\right\}^{-1}(\mathbf{W}_j)_{\widehat{\mathcal{S}}_j}^\top\mathbf{z}_j$ *and* $\left(\widehat{\boldsymbol{\beta}}_j^{OLS}\right)_{\widehat{\mathcal{S}}_j^c} = \mathbf{0}$*. If* $\mathcal{S}_j \subseteq \widehat{\mathcal{S}}_j$*, it holds that*

$$\frac{\widehat{\sigma}^{jj}}{\sigma^{jj}} \xrightarrow{p} 1.$$

The proof of Lemma 6 is provided in Section S2.4. The condition of Lemma 6 is also mild, and $\widehat{\mathcal{S}}_j$ can even converge to $\mathcal{S}_j$ under a similar proof of Theorem 2 in Zhang and Li (2023). This lemma offers a consistent estimator for σ^{jj}, enabling us to construct confidence intervals for statistical inference. With these lemmas, we prove Theorem 2.

Proof. [Proof of Theorem 2] We first have that

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\beta}}_j^u - \boldsymbol{\beta}_j) &= \sqrt{n}\{\widehat{\boldsymbol{\beta}}_j + \frac{1}{n}\widehat{\mathbf{M}}_j\mathbf{W}_j^\top(\mathbf{z}_j - \mathbf{W}_j\widehat{\boldsymbol{\beta}}_j) - \boldsymbol{\beta}_j\} \\ &= \sqrt{n}\big(\mathbf{I} - \frac{1}{n}\widehat{\mathbf{M}}_j\mathbf{W}_j^\top\mathbf{W}_j\big)(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) + \frac{1}{\sqrt{n}}\widehat{\mathbf{M}}_j\mathbf{W}_j^\top\boldsymbol{\varepsilon}_j.\end{aligned}$$

Then, conditional on $\mathbf{W}_j$, we have

$$\frac{1}{\sqrt{n}}\widehat{\mathbf{M}}_j\mathbf{W}_j^\top\boldsymbol{\varepsilon}_j \mid \mathbf{W}_j \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\sigma^{jj}}\widehat{\mathbf{M}}_j\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j}\widehat{\mathbf{M}}_j^\top\right).$$

We then investigate the term $\boldsymbol{\Delta}_j = \sqrt{n}\big(\mathbf{I} - \frac{1}{n}\widehat{\mathbf{M}}\mathbf{W}_j^\top\mathbf{W}_j\big)(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)$. Recall the definition $\boldsymbol{\delta}_j = \widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j$, we have

$$\|\boldsymbol{\Delta}_j\|_\infty = \sqrt{n} \cdot \max_l \big|\langle \mathbf{e}_l - \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j}\widehat{\mathbf{M}}_j^\top\mathbf{e}_l, \boldsymbol{\delta}_j\rangle\big|. \tag{S2.1}$$

From Lemma 2, it is easy to see that $\boldsymbol{\Sigma}_{\mathbf{W}_j}^{-1}$ is in the feasible set of (6), therefore $\widehat{\mathbf{M}}_j$ exists. By the constraint in (6), we have

$$\|H_\alpha(\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j}\widehat{\mathbf{M}}_j\mathbf{e}_l - \mathbf{e}_l)\|_{\infty,2} = \|H_\alpha(\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j}\widehat{\mathbf{m}}_{jl} - \mathbf{e}_l)\|_{\infty,2} \leqslant \gamma. \tag{S2.2}$$

Combining this equation above with Lemma 1, we have

$$\big|\langle \mathbf{e}_l - \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j}\widehat{\mathbf{M}}^\top\mathbf{e}_l, \boldsymbol{\delta}_j\rangle\big| \leqslant \alpha \cdot \|\boldsymbol{\delta}_j\|_1 + \gamma\|\boldsymbol{\delta}_j\|_{1,2}. \tag{S2.3}$$

From (S2.3), it only remains for us to investigate the bound of $\|\boldsymbol{\delta}_j\|_1$ and $\|\boldsymbol{\delta}_j\|_{1,2}$. From

Cauchy-Schwarz inequality, it is easy to see that

$$\frac{\|(\boldsymbol{\delta}_j)_{\mathcal{S}_j}\|_1}{\sqrt{s_e}} \leqslant \frac{\|\boldsymbol{\delta}_{\mathcal{S}}\|_1}{\sqrt{s_e}} \leqslant \|\boldsymbol{\delta}\|_2, \quad \frac{\|\boldsymbol{\delta}_j\|_{1,2}}{\sqrt{s_g}} \leqslant \frac{\|(\boldsymbol{\delta})_{\mathcal{G}}\|_{1,2}}{\sqrt{s_g}} \leqslant \|\boldsymbol{\delta}\|_2. \tag{S2.4}$$

Combining (S2.4) with Lemma 5, we have

$$\frac{\|\boldsymbol{\delta}_j\|_1}{\sqrt{s_e}} \leqslant \frac{\|\boldsymbol{\delta}\|_1}{\sqrt{s_e}} \leqslant 7\sqrt{\frac{2s_e \log(ep) + s_g \log(eq/s_g)}{n}} + 3\|\boldsymbol{\delta}\|_2,$$
$$\frac{\|\boldsymbol{\delta}_j\|_{1,2}}{\sqrt{s_g}} \leqslant \frac{\|(\boldsymbol{\delta})_{\mathcal{G}}\|_{1,2}}{\sqrt{s_g}} \leqslant 7\sqrt{\frac{2s_e \log(ep) + s_g \log(eq/s_g)}{n}} + 3\|\boldsymbol{\delta}\|_2$$

with probability at least $1 - C_1 \exp[-C_2\{s_e \log(ep) + s_g \log(eq/s_g)\}/s_e]$. Hence, combining the results above with Lemma 4, we write the upper bound in (S2.3) as

$$\left|\langle \mathbf{e}_l - \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j} \widehat{\mathbf{M}}_j^\top \mathbf{e}_l, \boldsymbol{\delta}_j \rangle\right| \precsim \frac{s_e \log(epq)}{n}.$$

Then combining the equation above with (S2.1) completes the upper bound of $\|\boldsymbol{\Delta}_j\|_\infty$.

We next give the lower bound of $\widehat{\mathbf{m}}_{jl} \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j} \widehat{\mathbf{m}}_{jl}$. By (S2.2) and Lemma 1, we have

$$1 - \mathbf{e}_l^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j} \widehat{\mathbf{m}}_{jl} \leqslant \alpha \|\mathbf{e}_l\|_1 + \gamma \|\mathbf{e}_l\|_{1,2} = \alpha + \gamma.$$

Hence for any $c \geqslant 0$ we have

$$\begin{aligned}\widehat{\mathbf{m}}_{jl}^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j} \widehat{\mathbf{m}}_{jl} &\geqslant \widehat{\mathbf{m}}_{jl}^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j} \widehat{\mathbf{m}}_{jl} + c(1 - \alpha - \gamma) - c\mathbf{e}_l^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j} \widehat{\mathbf{m}}_{jl} \\ &\geqslant \min_{\mathbf{m}} \mathbf{m}^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j} \mathbf{m} + c(1 - \alpha - \gamma) - c\mathbf{e}_l^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j} \mathbf{m}.\end{aligned}$$

As $\mathbf{m} = c\mathbf{e}_l/2$ achieves the minimum of the right hand side, we have

$$\widehat{\mathbf{m}}_{jl}^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j} \widehat{\mathbf{m}}_{jl} \geqslant c(1 - \alpha - \gamma) - \frac{c^2}{4}\big(\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j}\big)_{l,l} \geqslant \frac{(1 - \alpha - \gamma)^2}{\big(\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j}\big)_{l,l}}. \tag{S2.5}$$

The last inequality is by setting $c = 2(1 - \alpha - \gamma)/\big(\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j}\big)_{l,l}$. By Lemma 3, we have with probability at least $1 - C_1 \exp\Big(-C_2 \frac{s_e \log(ep) + s_g \log(eq/s_g)}{s_e}\Big)$,

$$\phi_1/(2\phi_0) \leqslant \mathbf{e}_l^\top \frac{\mathbf{W}_j^\top \mathbf{W}_j}{n} \mathbf{e}_l \leqslant M^2\phi_2 + 1, \quad \forall 1 \leqslant l \leqslant (p-1)(q+1).$$

Combining the equation above with (S2.5), we have

$$\widehat{\mathbf{m}}_{jl}^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j} \widehat{\mathbf{m}}_{jl} \geqslant \frac{1}{2(M^2\phi_2 + 1)}.$$

This completes the proof of Theorem 2 with the SAGE estimator $\widehat{\boldsymbol{\beta}}_j^u$ defined by (5) and (6).

Additionally, Proposition 1 implies that the solution $\widehat{\mathbf{M}}_j$ obtained based on (7) also satisfies

(6). Hence, the results of the theorem apply to the SAGE estimator $\widehat{\boldsymbol{\beta}}_j^u$ defined by (5) and (7).

S2.1 *Proof of Lemma 2*

We begin by proving the following lemma, which will be used for proving Lemma 2.

LEMMA 7: *Recall that $(\mathbf{W}_j)_{i.}, i = 1, \dots, n$, are the rows of $\mathbf{W}_j$. Conditional on $\mathbf{U}$, let $\mathbf{w}_i$ have the same distribution as $(\mathbf{W}_j)_{i.}$, and $\kappa = \max_i \max_{l\in[(p-1)(q+1)]} \|\langle \mathbf{w}_i, \mathbf{e}_l\rangle\|_{\psi_2}$. Then under the conditions of Theorem 2, for any fixed $\mathbf{U}_0 \in \mathbb{R}^{(p-1)(q+1)\times k}$ and $\mathbf{v} \in \mathbb{R}^{(p-1)(q+1)}$, the following inequality holds:*

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{U}_0^\top(\mathbf{W}_j)_{i.}(\mathbf{W}_j)_{i.}^\top\mathbf{v} - \mathbf{U}_0^\top\boldsymbol{\Sigma}_{\mathbf{W}_j|\mathbf{U}}\mathbf{v}\right\|_2 \geqslant t\|\mathbf{U}_0\|\|\mathbf{v}\|_2\right) \leqslant 2\exp\left(Ck - cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right).$$

Proof. [Proof of Lemma 7] All the proof is conditional on $\mathbf{U}$. By Assumption 1, $\mathbb{E}\mathbf{w} = \mathbf{0}$, $\|\langle \mathbf{w}, \mathbf{v}\rangle\|_{\psi_2} \leqslant C'\kappa$ for any fixed $\|\mathbf{v}\|_2 = 1$, where $C' > 0$ is a constant. Therefore, for any fixed vectors $\mathbf{u}_0$ and $\mathbf{v}$ with $\|\mathbf{u}_0\|_2 = \|\mathbf{v}\|_2 = 1$, there exists constant $C > 0$ such that

$$\|\mathbf{u}_0^\top\mathbf{w}\mathbf{w}^\top\mathbf{v}\|_{\psi_1} \leqslant C\kappa^2.$$

Note that $\mathbb{E}\mathbf{u}_0^\top(\mathbf{W}_j)_{i.}(\mathbf{W}_j)_{i.}^\top\mathbf{v} = \mathbf{u}_0^\top\boldsymbol{\Sigma}_{\mathbf{W}_j|\mathbf{U}}\mathbf{v}$. By Bernstein inequality (Proposition 5.16 in Vershynin (2010)), it holds that

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{u}_0^\top(\mathbf{W}_j)_{i.}(\mathbf{W}_j)_{i.}^\top\mathbf{v} - \mathbf{u}_0^\top\boldsymbol{\Sigma}_{\mathbf{W}_j|\mathbf{U}}\mathbf{v}\right\|_2 \geqslant t\|\mathbf{u}\|_2\|\mathbf{v}\|_2\right) \leqslant 2\exp\left(-cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right) \tag{S2.6}$$

for some constant $c > 0$. Then we can prove Lemma 7 as follows.

For any $\mathbf{r} \in \mathbb{R}^r$, $\|\mathbf{r}\|_2 = 1$, set $\mathbf{u}_r = \mathbf{U}_0\mathbf{r}$. Hence from (S2.6) we have

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{r}^\top\mathbf{U}_0^\top(\mathbf{W}_j)_{i.}(\mathbf{W}_j)_{i.}^\top\mathbf{v} - \mathbf{u}_r^\top\boldsymbol{\Sigma}_{\mathbf{W}_j|\mathbf{U}}\mathbf{v}\right\|_2 \geqslant t\|\mathbf{U}_0\mathbf{r}\|_2\|\mathbf{v}\|_2\right) \leqslant 2\exp\left(-cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right).$$

Define a $\frac{1}{2}$-Net $\mathcal{N}_{\frac{1}{2}}$ of $\mathbb{S}^{r-1} = \{\mathbf{x} \in \mathbb{R}^r : \|\mathbf{x}\|_2 = 1\}$. Easy to see that $|\mathcal{N}_{\frac{1}{2}}| \leqslant 5^r$. Moreover, for any vector $\mathbf{r}_0 \in \mathbb{R}^r$, there exists $\mathbf{x}_{r_0} \in \mathcal{N}_{\frac{1}{2}}$ such that $\|\mathbf{r}_0/\|\mathbf{r}_0\|_2 - \mathbf{x}_{r_0}\|_2 \leqslant 1/2$. This indicates that for any $\mathbf{r}_0 \in \mathbb{R}^r$, there exists $\mathbf{x}_{r_0} \in \mathcal{N}_{\frac{1}{2}}$ such that

$$\|\mathbf{r}_0\|_2 - |\mathbf{x}_{r_0}^\top \mathbf{r}_0| \leqslant \left| \left(\frac{\mathbf{r}_0}{\|\mathbf{r}_0\|_2} - \mathbf{x}_{r_0} \right)^\top \mathbf{r}_0 \right| \leqslant \left\| \frac{\mathbf{r}_0}{\|\mathbf{r}_0\|_2} - \mathbf{x}_{r_0} \right\|_2 \|\mathbf{r}_0\|_2 \leqslant \|\mathbf{r}_0\|_2/2.$$

Hence, we have that

$$\sup_{\|\mathbf{r}\|_2=1} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{r}^\top \mathbf{U}_0^\top (\mathbf{W}_j)_{i.} (\mathbf{W}_j)_{i.}^\top \mathbf{v} - \mathbf{u}_r^\top \boldsymbol{\Sigma}_{\mathbf{W}_j|\mathbf{U}} \mathbf{v} \right| \leqslant 2 \sup_{\mathbf{r} \in \mathcal{N}_{\frac{1}{2}}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{r}^\top \mathbf{U}_0^\top (\mathbf{W}_j)_{i.} (\mathbf{W}_j)_{i.}^\top \mathbf{v} - \mathbf{u}_r^\top \boldsymbol{\Sigma}_{\mathbf{W}_j|\mathbf{U}} \mathbf{v} \right|. \tag{S2.7}$$

The following inequalities hold:

$$\begin{aligned}
&P\left(\left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{U}_0^\top (\mathbf{W}_j)_{i.} (\mathbf{W}_j)_{i.}^\top \mathbf{v} - \mathbf{U}_0^\top \boldsymbol{\Sigma}_{\mathbf{W}_j|\mathbf{U}} \mathbf{v} \right\|_2 \geqslant t \|\mathbf{U}_0\| \|\mathbf{v}\|_2 \right) \\
&\quad = P\left(\sup_{\|\mathbf{r}\|_2=1} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{r}^\top \mathbf{U}_0^\top (\mathbf{W}_j)_{i.} (\mathbf{W}_j)_{i.}^\top \mathbf{v} - \mathbf{u}_r^\top \boldsymbol{\Sigma}_{\mathbf{W}_j|\mathbf{U}} \mathbf{v} \right| \geqslant t \|\mathbf{U}_0\| \|\mathbf{v}\|_2 \right) \\
&\quad \leqslant P\left(\sup_{\mathbf{r} \in \mathcal{N}_{\frac{1}{2}}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{r}^\top \mathbf{U}_0^\top (\mathbf{W}_j)_{i.} (\mathbf{W}_j)_{i.}^\top \mathbf{v} - \mathbf{u}_r^\top \boldsymbol{\Sigma}_{\mathbf{W}_j|\mathbf{U}} \mathbf{v} \right| \geqslant \frac{t}{2} \|\mathbf{U}_0\| \|\mathbf{v}\|_2 \right) \\
&\quad \leqslant 5^r \cdot 2 \exp\left\{ -cn \min\left(\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2} \right) \right\},
\end{aligned}$$

where the first equality is by the definition, the first inequality is by (S2.7), and the second inequality is by the union bound and $|\mathcal{N}_{\frac{1}{2}}| \leqslant 5^r$. This completes the proof of Lemma 7.

Proof. [Proof of Lemma 2] From the Kronecker structure of $\boldsymbol{\Sigma}_{\mathbf{W}_j|\mathbf{U}}$, when $n > q+1$ we can construct a matrix $\mathbf{A}_l$ such that $\|\boldsymbol{\Sigma}_{\mathbf{W}_j|\mathbf{U}} \mathbf{A}_l \mathbf{e}_l - \mathbf{e}_l\|_2 \leqslant \gamma/2$. Applying the union bound, we have that

$$\begin{aligned}
&P\left(\max_{1 \leqslant l \leqslant (p-1)(q+1)} \left\| H_\alpha\left(\mathbf{e}_l - \frac{\mathbf{W}_j^\top \mathbf{W}_j}{n} \mathbf{A}_l \mathbf{e}_l \right) \right\|_{\infty,2} \geqslant \gamma \right) \\
&\quad \leqslant \sum_{l=1}^{(p-1)(q+1)} \sum_{g=1}^{q+1} P\left(\left\| H_\alpha\left((\mathbf{e}_l)_{(g)} - \frac{(\mathbf{W}_j)_{(g)}^\top \mathbf{W}_j}{n} \mathbf{A}_l \mathbf{e}_l \right) \right\|_2 \geqslant \gamma \right).
\end{aligned} \tag{S2.8}$$

It remains to give the upper bound of $P\Big(\Big\| H_\alpha\Big((\mathbf{e}_l)_{(g)} - \frac{\mathbf{W}_j^\top \mathbf{W}_j}{n} \mathbf{A}_l (\mathbf{e}_l)_{(g)} \Big) \Big\|_2 \geqslant \gamma \Big)$. By the definition of $H_\alpha(\cdot)$, it is easy to see that

$$\begin{aligned}
&P\bigg(\bigg\|H_\alpha\bigg((\mathbf{e}_l)_{(g)}-\frac{(\mathbf{W}_j)_{(g)}^\top\mathbf{W}_j}{n}\mathbf{A}_l\mathbf{e}_l\bigg)\bigg\|_2\geqslant\gamma\bigg)\\
&\quad\leqslant P\bigg(\exists\Lambda\subseteq(g),\text{all entries of }\bigg|(\mathbf{e}_l)_\Lambda-\frac{(\mathbf{W}_j)_\Lambda^\top\mathbf{W}_j}{n}\mathbf{A}_l\mathbf{e}_l\bigg|\geqslant\alpha\text{ and }\bigg\|(\mathbf{e}_l)_\Lambda-\frac{(\mathbf{W}_j)_\Lambda^\top\mathbf{W}_j}{n}\mathbf{A}_l\mathbf{e}_l\bigg\|\geqslant\gamma\bigg)\\
&\quad\leqslant P\bigg(\exists\Lambda\subseteq(g),|\Lambda|>\gamma^2/\alpha^2,\text{all entries of }\bigg|(\mathbf{e}_l)_\Lambda-\frac{(\mathbf{W}_j)_\Lambda^\top\mathbf{W}_j}{n}\mathbf{A}_l\mathbf{e}_l\bigg|\geqslant\alpha\bigg)\\
&\qquad+P\bigg(\exists\Lambda\subseteq(g),|\Lambda|\leqslant\gamma^2/\alpha^2,\bigg\|(\mathbf{e}_l)_\Lambda-\frac{(\mathbf{W}_j)_\Lambda^\top\mathbf{W}_j}{n}\mathbf{A}_l\mathbf{e}_l\bigg\|_2\geqslant\gamma\bigg).
\end{aligned}$$

We provide the upper bounds for the two probabilities mentioned above. It is evident that when $|\Lambda|=\lceil s_e/s_g\rceil$, the first probability on the right-hand side reaches its maximum value. Similarly, when $|\Lambda|=\lfloor s_e/s_g\rfloor$, the second probability on the right-hand side achieves its maximum value. Moreover, when $|\Lambda|=\lceil s_e/s_g\rceil$, that all entries of $\big|(\mathbf{e}_l)_\Lambda-\frac{(\mathbf{W}_j)_\Lambda^\top\mathbf{W}_j}{n}\mathbf{A}_l\mathbf{e}_l\big|\geqslant\alpha$ indicates that $\big\|(\mathbf{e}_l)_\Lambda-\frac{(\mathbf{W}_j)_\Lambda^\top\mathbf{W}_j}{n}\mathbf{A}_l\mathbf{e}_l\big\|_2\geqslant\gamma$, further implying that

$$\begin{aligned}
&P\bigg(\bigg\|H_\alpha\bigg((\mathbf{e}_l)_{(g)}-\frac{(\mathbf{W}_j)_{(g)}^\top\mathbf{W}_j}{n}\mathbf{A}_l\mathbf{e}_l\bigg)\bigg\|_2\geqslant\gamma\bigg)\\
&\quad\leqslant\sum_{\substack{\Lambda\subseteq(g)\\|\Lambda|=\lfloor s_e/s_g\rfloor}}\mathbb{P}\bigg(\bigg\|(\mathbf{e}_l)_\Lambda-\frac{1}{n}(\mathbf{W}_j)_\Lambda^\top\mathbf{W}_j\mathbf{A}_l\mathbf{e}_l\bigg\|_2\geqslant\gamma\bigg)+\sum_{\substack{\Lambda\subseteq(g)\\|\Lambda|=\lceil s_e/s_g\rceil}}\mathbb{P}\bigg(\bigg\|(\mathbf{e}_l)_\Lambda-\frac{1}{n}(\mathbf{W}_j)_\Lambda^\top\mathbf{W}_j\mathbf{A}_l\mathbf{e}_l\bigg\|_2\geqslant\gamma\bigg)\\
&\quad\leqslant\sum_{\substack{\Lambda\subseteq(g)\\|\Lambda|=\lfloor s_e/s_g\rfloor}}\mathbb{P}\bigg(\bigg\|(\boldsymbol{\Sigma}_{\mathbf{W}_j|\mathbf{U}}\mathbf{A}_l\mathbf{e}_l)_\Lambda-\frac{1}{n}(\mathbf{W}_j)_\Lambda^\top\mathbf{W}_j\mathbf{A}_l\mathbf{e}_l\bigg\|_2\geqslant\gamma/2\bigg)+\sum_{\substack{\Lambda\subseteq(g)\\|\Lambda|=\lceil s_e/s_g\rceil}}\mathbb{P}\bigg(\bigg\|(\boldsymbol{\Sigma}_{\mathbf{W}_j|\mathbf{U}}\mathbf{A}_l\mathbf{e}_l)_\Lambda-\frac{1}{n}(\mathbf{W}_j)_\Lambda^\top\mathbf{W}_j\mathbf{A}_l\mathbf{e}_l\bigg\|_2\geqslant\gamma/2\bigg)\\
&\quad\leqslant\bigg\{\binom{p-1}{\lfloor s_e/s_g\rfloor}\cdot2\exp(C\lfloor s_e/s_g\rfloor)+\binom{p-1}{\lceil s_e/s_g\rceil}\cdot2\exp(C\lceil s_e/s_g\rceil)\bigg\}\\
&\qquad\cdot\exp\bigg(-cn\cdot C\frac{s_e\log(epq)}{s_gn}\bigg)\\
&\quad\leqslant4\exp\bigg(-C'\frac{s_e\log(epq)}{s_g}\bigg),
\end{aligned}$$

where $C'>0$ is a large enough constant, the second inequality is by the construction $\|\boldsymbol{\Sigma}_{\mathbf{W}_j|\mathbf{U}}\mathbf{A}_l\mathbf{e}_l-\mathbf{e}_l\|_2\leqslant\gamma/2$, the third inequality is by Lemma 7 and the last inequality is by the direct calculation, the Stirling formula and the definition of γ with a sufficiently large constant $C>0$. Combining the results with (S2.8), we have

$$\begin{aligned}
&P\bigg(\max_{1\leqslant l\leqslant(p-1)(q+1)}\bigg\|H_\alpha\bigg(\mathbf{e}_l-\frac{\mathbf{W}_j^\top\mathbf{W}_j}{n}\mathbf{A}_l\mathbf{e}_l\bigg)\bigg\|_{\infty,2}\geqslant\gamma\bigg)\\
&\quad\leqslant4(p-1)(q+1)^2\cdot\exp\bigg(-C'\frac{s_e\log(epq)}{s_g}\bigg)\leqslant\exp\bigg(-C''\frac{s_e\log(epq)}{s_g}\bigg)
\end{aligned}$$

for some $C''>0$. This completes the proof of Lemma 2.

S2.2 *Proof of Lemma 3*

Recall the definition of $\mathbf{W} = [\mathbf{z}_1, ..., \mathbf{z}_{j-1}, \mathbf{z}_{j+1}, ..., \mathbf{z}_p, \mathbf{z}_1 \odot \mathbf{u}_1, ..., \mathbf{z}_{j-1} \odot \mathbf{u}_1, \mathbf{z}_{j+1} \odot \mathbf{u}_1, ..., \mathbf{z}_p \odot \mathbf{u}_q] \in \mathbb{R}^{n\times(p-1)(q+1)}$. The independent row of $\mathbf{W}$ can be written as

$$(\mathbf{W}_j)_{i.} = (z_1^{(i)}, ..., z_{j-1}^{(i)}, z_{j+1}^{(i)}, ..., z_p^{(i)}, u_1^{(i)} \cdot z_1^{(i)}, ..., z_{j-1}^{(i)} \cdot u_1^{(i)}, z_{j+1}^{(i)} \cdot u_1^{(i)}, ..., z_p^{(i)} \cdot u_q^{(i)}).$$

It can be seen that

$$\left(\widehat{\mathbf{\Sigma}}_{\mathbf{W}_j}\right)_{l,l} = \mathbf{e}_l^\top \frac{\mathbf{W}_j^\top \mathbf{W}_j}{n} \mathbf{e}_l.$$

We first consider the expectation of $\left(\widehat{\mathbf{\Sigma}}_{\mathbf{W}_j}\right)_{l,l}$ and find that

$$\begin{aligned}\mathbb{E}\ \mathbf{e}_l^\top \frac{\mathbf{W}_j^\top \mathbf{W}_j}{n} \mathbf{e}_l &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\ \{\mathbf{e}_l^\top (\mathbf{W}_j)_{i.}\}^2 \\ &= \mathbb{E}\ (\mathbf{e}_l^\top \mathbf{W}_{1.})^2.\end{aligned}$$

By Assumptions 1 and 2, we can easily conclude that conditional on $\mathbf{u}^{(1)}$,

$$\phi_1 \leqslant \lambda_{\min} \operatorname{Cov}(\mathbf{z}^{(i)}) \leqslant \mathbb{E}\big(z_l^{(1)}\big)^2 \leqslant \lambda_{\max} \operatorname{Cov}(\mathbf{z}^{(i)}) \leqslant \phi_2,$$

and

$$1/\phi_0 \leqslant \max_{l\in[q]} \mathbb{E}\ \left(u_l^{(1)}\right)^2 \leqslant M^2.$$

Therefore,

$$\phi_1/\phi_0 \leqslant \mathbb{E}\ \mathbf{e}_l^\top \frac{\mathbf{W}_j^\top \mathbf{W}_j}{n} \mathbf{e}_l \leqslant M^2\phi_2. \tag{S2.9}$$

We next give the distance between $\left(\widehat{\mathbf{\Sigma}}_{\mathbf{W}_j}\right)_{l,l}$ and $\mathbb{E}\ \mathbf{e}_l^\top \frac{\mathbf{W}_j^\top \mathbf{W}_j}{n} \mathbf{e}_l$. We first have the following lemma.

Lemma 8 (Theorem 4.1 in Kuchibhotla and Chakrabortty (2022)): *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent random vectors in $\mathbb{R}^p$. Assume each element of $\mathbf{X}_i$ is sub-exponential with $\|X_{i,j}\|_{\psi_1} < K_2, i \in [n], j \in [p]$. where $\|X_{i,j}\|_{\psi_1} = \sup_{d\geqslant 1} d^{-1}(E|X_{i,j}|^d)^{1/d}$ denotes the subexponential norm. Let $\widehat{\mathbf{\Sigma}}_{\mathbf{X}} = \mathbf{X}^\top\mathbf{X}/n$ and $\mathbf{\Sigma}_{\mathbf{X}} = \mathbb{E}(\mathbf{X}^\top\mathbf{X}/n)$. Define*

$$A_n = \max_{j,k} \frac{1}{n}\sum_{i=1}^n \operatorname{Var}\left(X_{i,j}X_{i,k}\right)$$

Then for any $t > 0$, with probability at least $1 - 3e^{-t}$,

$$\sup_{\|\mathbf{v}\|_0 \leqslant k, \|\mathbf{v}\|_2 \leqslant 1} |\mathbf{v}^\top(\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}} - \boldsymbol{\Sigma}_{\mathbf{X}})\mathbf{v}| \precsim k\sqrt{\frac{A_n(t+\log p)}{n}} + kK_2^2 \frac{\log^2 n \cdot (t+\log p)^2}{n}.$$

By the Cauchy-Schwarz inequality, we have

$$\max_{j,k} \frac{1}{n}\sum_{i=1}^n \mathbb{E}\{(\mathbf{W}_{ij}\mathbf{W}_{ik})^2\} = \max_{l_1,l_2,l_3,l_4} \mathbb{E}(z_{l_1}^{(1)^2} z_{l_2}^{(1)^2} u_{l_3}^{(1)^2} u_{l_4}^{(1)^2}) = O(1).$$

Applying Lemma 8 and letting $t = C_2 \frac{s_e \log(ep) + s_g \log(eq/s_g)}{s_e}$, we see that with probability at least $1 - C_1 \exp\Big\{ - C_2 \frac{s_e \log(ep)+s_g\log(eq/s_e)}{s_e}\Big\}$,

$$\left|\mathbf{e}_l^\top \frac{\mathbf{W}_j^\top \mathbf{W}_j}{n}\mathbf{e}_l - \mathbb{E}\ \mathbf{e}_l^\top \frac{\mathbf{W}_j^\top \mathbf{W}_j}{n}\mathbf{e}_l\right| \precsim \frac{\log^2 n \cdot (s_e \log(epq) + s_g\log(eq/s_g))^2}{s_e^2 n}.$$

By Assumption 3, it holds that $\sqrt{n}/\log(n) \geqslant C\{\log(p) + \log(q)\}$ for a sufficiently large $C > 0$, and the right side of the above (approximate) equality is sufficiently small. Hence combining this with (S2.9) we have

$$\phi_1/(2\phi_0) \leqslant \mathbf{e}_l^\top \frac{\mathbf{W}_j^\top \mathbf{W}_j}{n}\mathbf{e}_l \leqslant M^2\phi_2 + 1.$$

This completes the proof of Lemma 3.

S2.3 *Proofs of Lemmas 4 and 5*

The results in Lemma 4 improve Theorem 3 in Zhang and Li (2025), and we follow a similar stepwise structure used in Zhang and Li (2025) to prove Lemma 4. Specifically, we will provide a concise version of the proof of Theorem 3 from Zhang and Li (2025), while incorporating our improved results.

To proceed, as a preamble on the convex penalty property, we have

$$\frac{1}{2n}\sum_{j=1}^p \|\mathbf{z}_j - \mathbf{W}_j\widehat{\boldsymbol{\beta}}_j\|_2^2 + \lambda_g \sum_{h=1}^q \|\widehat{\mathbf{b}}_h\|_2 + \lambda_e \sum_{h=0}^q \|\widehat{\mathbf{b}}_h\|_1 + \frac{1}{2n}\|\mathbf{W}_j(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\|_2^2$$
$$\leqslant \frac{1}{2n}\sum_{j=1}^p \|\mathbf{z}_j - \mathbf{W}_j\boldsymbol{\beta}_j\|_2^2 + \lambda_g \sum_{h=1}^q \|\mathbf{b}_h\|_2 + \lambda_e \sum_{h=0}^q \|\mathbf{b}_h\|_1,$$

where $\widehat{\mathbf{b}}_h = \big((\widehat{\boldsymbol{\beta}}_1)_{(h)}, \ldots, (\widehat{\boldsymbol{\beta}}_p)_{(h)}\big)$, $h = \{0\} \cup [q]$. Writing $\boldsymbol{\delta}_j = \widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j$, $\quad \boldsymbol{\varepsilon}_j = \mathbf{z}_j - \mathbb{E}(\mathbf{z}_j)$, and reorganizing terms in the above inequality gives

$$\frac{1}{n}\sum_{j=1}^{p}\|\mathbf{W}_j\boldsymbol{\delta}_j\|_2^2+\lambda_g\sum_{h=1}^{q}\|\widehat{\mathbf{b}}_h\|_2+\lambda_e\sum_{h=0}^{q}\|\widehat{\mathbf{b}}_h\|_1\leqslant\frac{1}{n}\sum_{j=1}^{p}\langle\boldsymbol{\varepsilon}_j,\mathbf{W}_j\boldsymbol{\delta}_j\rangle+\lambda_g\sum_{h=1}^{q}\|\widehat{\mathbf{b}}_h\|_2+\lambda_e\sum_{h=0}^{q}\|\widehat{\mathbf{b}}_h\|_1.$$

Recall $\mathcal{W}=\begin{pmatrix}\mathbf{W}_1 & 0 & \dots & 0\\ 0 & \mathbf{W}_2 & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & \mathbf{W}_p\end{pmatrix}$, $\boldsymbol{\varepsilon}=(\boldsymbol{\varepsilon}_1,\dots,\boldsymbol{\varepsilon}_p)$, and $\boldsymbol{\delta}=(\boldsymbol{\delta}_1,\dots,\boldsymbol{\delta}_p)$. Applying the triangle inequality and after some direct algebra, we arrive at

$$\frac{1}{n}\|\mathcal{W}\boldsymbol{\delta}\|_2^2+\lambda_g\|\boldsymbol{\delta}_{(\mathcal{G}^c)}\|_{1,2}+\lambda_e\|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1\leqslant\frac{1}{n}\langle\boldsymbol{\varepsilon},\mathcal{W}\boldsymbol{\delta}\rangle+\lambda_g\|\boldsymbol{\delta}_{(\mathcal{G})}\|_{1,2}+\lambda_e\|\boldsymbol{\delta}_{\mathcal{S}}\|_1. \tag{S2.10}$$

where $\|\boldsymbol{\delta}_{\mathcal{G}}\|_{1,2}=\sum_{j=1}^{p}\sum_{h\in\mathcal{G}}\|(\boldsymbol{\delta}_j)_{(h)}\|_2$, and $\|\boldsymbol{\delta}_{\mathcal{G}^c}\|_1=\sum_{j=1}^{p}\sum_{h\in\mathcal{G}^c}\|(\boldsymbol{\delta}_j)_{(h)}\|_2$. From (S2.10) and by the Cauchy-Schwarz inequality we have that

$$\frac{1}{n}\|\mathcal{W}\boldsymbol{\delta}\|_2^2+\lambda_g\|\boldsymbol{\delta}_{(\mathcal{G}^c)}\|_{1,2}+\lambda_e\|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1\leqslant\frac{1}{2a_1n}\|\mathcal{W}\boldsymbol{\delta}\|_2^2+\frac{a_1}{2n}\|\mathcal{P}_{\widetilde{\mathcal{S}}}\boldsymbol{\varepsilon}\|_2^2+\lambda_g\|\boldsymbol{\delta}_{(\mathcal{G})}\|_{1,2}+\lambda_e\|\boldsymbol{\delta}_{\mathcal{S}}\|_1, \tag{S2.11}$$

where $a_1\geqslant 0$, $\widehat{\mathcal{S}}=\{l:(\widehat{\boldsymbol{\beta}})_l\neq 0, l\in[p(p-1)(q+1)]\}$, $\widetilde{\mathcal{S}}=\widehat{\mathcal{S}}\cup\mathcal{S}$ and $\mathcal{P}_{\widetilde{\mathcal{S}}}$ is the orthogonal projection matrix onto the column space of $\mathcal{W}_{\widetilde{\mathcal{S}}}$.

Next, while following the steps outlined by Zhang and Li (2025), we demonstrate key distinctions at each major step of the proof. This comparison highlights how our analysis achieves an improvement over their results.

Step 1: In this step, we aim to provide the bound for the term $\sup_{|\mathcal{J}|=s_e,|\mathcal{G}(\mathcal{J})|=s_g}\|\mathcal{P}_{\mathcal{J}}\boldsymbol{\varepsilon}\|_2^2$ for any $\mathcal{J}\subseteq[p(p-1)(q+1)]$. Here, given any $\boldsymbol{\gamma}\in\{0,1\}^{p(p-1)(q+1)}$ satisfying $\boldsymbol{\gamma}_{\mathcal{J}}=\mathbf{1}$ and $\boldsymbol{\gamma}_{\mathcal{J}^c}=\mathbf{0}$, $\mathcal{G}(\mathcal{J})$ is defined by $\mathcal{G}(\mathcal{J})=\{h:(\boldsymbol{\gamma})_{(h)}\neq\mathbf{0}, h\in[q]\}$. We have the following lemma from Zhang and Li (2025).

Lemma 9 (Step 1 in Zhang and Li (2025)): *Suppose that Assumption 3 holds, it holds that*

$$P\left(\sup_{|\mathcal{J}|=s_e,|\mathcal{G}(\mathcal{J})|=s_g}\|\mathcal{P}_{\mathcal{J}}\boldsymbol{\varepsilon}\|_2^2\geqslant 6\{2s_e\log(ep)+s_g\log(eq/s_g)\}+t\right)\leqslant c_1\exp(-c_2t),$$

for some constant $c_1, c_2 > 0$.

Step 2: We modify this step to establish the setup necessary for demonstrating our improved results. From Lemma 9, if we define

$$r_{s_e,s_g} = \sup_{|\mathcal{J}|=s_e,|\mathcal{G}(\mathcal{J})|=s_g} \|\mathcal{P}_{\mathcal{J}}\boldsymbol{\varepsilon}\|_2^2 - 6\{2s_e \log(ep) + s_g \log(eq/s_g)\},$$

it is clear that

$$P(r_{s_e,s_g} \geqslant t) \leqslant c_1 \exp(-c_2 t).$$

Applying the union bound, we have

$$P\left(\sup_{s_e \in [p(p-1)(q+1)], s_g \in [0:q]} r_{s,s_g} \geqslant t\right) \leqslant c_1 p(p-1)(q+1)^2 \exp(-c_2 t).$$

Set $t = 2s_e \log(ep) + s_g \log(eq/s_g)$, with probability at least

$$1 - C_1 \exp\big[-C_2\{2s_e \log(ep) + s_g \log(eq/s_g)\}\big],$$

it holds that

$$\|\mathcal{P}_{\widehat{\mathcal{S}}}\boldsymbol{\varepsilon}\|_2^2 \leqslant 14(s_e + \widehat{s}_e)\log(ep) + 7(s_g + \widehat{s}_g)\log(eq/s_g), \tag{S2.12}$$

where we use the fact that $1 \leqslant s_g \leqslant q$, implying $s_g \log(eq/s_g) = \Omega(\log(q))$.

Step 3: The main improvements compared to Zhang and Li (2025) are presented in this step. It is important to note that Assumption 3 is more relaxed compared to Zhang and Li (2025), which necessitates different treatments of the inequalities in the proof. We provide the proof as follows. Note that $\widehat{\boldsymbol{\beta}}$ is a stationary and minimum point of

$$\frac{1}{2n}\sum_{j=1}^{p} \|\mathbf{z}_j - \mathbf{W}_j\boldsymbol{\beta}_j\|_2^2 + \lambda_e \sum_{h=0}^{q} \|\mathbf{b}_h\|_1 + \lambda_g \sum_{h=1}^{q} \|\boldsymbol{b}_h\|_2.$$

The KKT condition gives us that for any $l \in \widehat{\mathcal{S}}_j \cap \{(0)\}$, $(\widehat{\boldsymbol{\beta}})_l$ satisfy

$$\lambda_e \text{sign}\{(\widehat{\boldsymbol{\beta}})_l\} = \frac{1}{n}\langle \mathcal{W}_{\cdot l}, \mathbf{z}_j - \mathbf{W}_j\widehat{\boldsymbol{\beta}}_j\rangle, \quad \text{for any } l \in \widehat{\mathcal{S}}_j \cap \{(0)\},$$
$$\lambda_e \text{sign}\{(\widehat{\boldsymbol{\beta}})_l\} + \lambda_g \frac{(\widehat{\boldsymbol{\beta}})_l}{\|\boldsymbol{b}_h\|_2} = \frac{1}{n}\langle \mathcal{W}_{\cdot l}, \mathbf{z}_j - \mathbf{W}_j\widehat{\boldsymbol{\beta}}_j\rangle, \quad \text{for any } l \in \widehat{\mathcal{S}}_j \cap (h).$$

Squaring both sides and summing over all $l \in \widehat{\mathcal{S}}$ gives

$$\lambda_g^2 \cdot \widehat{s}_g + \lambda_e^2 \cdot \widehat{s}_e \leqslant \frac{1}{n^2}\sum_{j=1}^{p} \big\|(\mathbf{W}_j)_{\widetilde{\mathcal{S}}_j}^\top(\mathbf{z}_j - \mathbf{W}_j\widehat{\boldsymbol{\beta}}_j)\big\|_2^2 = \frac{1}{n^2}\sum_{j=1}^{p} \big\|(\mathbf{W}_j)_{\widetilde{\mathcal{S}}_j}^\top \mathcal{P}_{\widetilde{\mathcal{S}}_j}(\mathbf{z}_j - \mathbf{W}_j\widehat{\boldsymbol{\beta}}_j)\big\|_2^2.$$

We now bound the term $\|(\mathbf{W}_j)_{\widetilde{\mathcal{S}}_j}\|_{\mathrm{op}}/n$. From Lemma 8 and the proof of Lemma 3, it can be seen that with probability at least $1 - C_1\exp\Big\{-C_2\frac{s_e\log(ep)+s_g\log(eq/s_g)}{s_e}\Big\}$,

$$\sup_{\|\mathbf{v}_j\|_0\leqslant s_\lambda, \|\mathbf{v}_j\|_2\leqslant 1} |\mathbf{v}_j^\top(\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j} - \boldsymbol{\Sigma}_{\mathbf{W}_j})\mathbf{v}_j| \precsim s_\lambda\frac{\log^2 n\cdot\{s_e\log(epq)+s_g\log(eq/s_g)\}^2}{s_e^2 n},$$

where we use that

$$s_\lambda \cdot \sqrt{\frac{s_e\log(epq)+s_g\log(eq/s_g)}{s_e n}} = o(1)$$

by the condition of $s_\lambda(\log p + \log q) = O(\sqrt{n}/\log n)$. As $\sup_{\|\mathbf{v}_j\|_2\leqslant 1}\mathbf{v}_j^\top\boldsymbol{\Sigma}_{\mathbf{W}_j}\mathbf{v}_j$ is also bounded, we conclude that

$$\|(\mathbf{W}_j)_{\widetilde{\mathcal{S}}_j}\|^2/n \leqslant M_1$$

for some constant $M_1 > 0$. From the analysis above, we obtain that with probability at least $1 - C_1\exp\Big\{-C_2\frac{s_e\log(ep)+s_g\log(eq/s_g)}{s_e}\Big\}$,

$$\lambda_e^2\widehat{s}_e + \lambda_g^2\widehat{s}_g \leqslant \frac{2M_1}{n}\|\mathcal{W}\boldsymbol{\delta}\|_2^2 + \frac{2M_1}{n}\|\mathcal{P}_{\widetilde{\mathcal{S}}}\boldsymbol{\varepsilon}\|_2^2. \tag{S2.13}$$

With (S2.12) and (S2.13) and $\lambda_e = C\sqrt{(2s_e\log(ep)+s_g\log(eq/s_g))/(ns_e)}$ for any $C > 0$, direct algebra gives

$$\|\mathcal{P}_{\widetilde{\mathcal{S}}}\boldsymbol{\varepsilon}\|_2^2 \leqslant 7\big\{2s_e\log(ep)+s_g\log(eq/s_g)\big\} + \frac{14M_1}{C^2}\big(\|\mathcal{P}_{\widetilde{\mathcal{S}}}\boldsymbol{\varepsilon}\|_2^2 + \|\mathcal{W}\boldsymbol{\delta}\|_2^2\big). \tag{S2.14}$$

Combining (S2.14) with (S2.11), with probability at least $1-C_1\exp\Big\{-C_2\frac{s_e\log(ep)+s_g\log(eq/s_g)}{s_e}\Big\}$, it holds that

$$\begin{aligned}
&\frac{1}{n}\|\mathcal{W}\boldsymbol{\delta}\|_2^2 + \lambda_e\|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 + \lambda_g\|\boldsymbol{\delta}_{(\mathcal{G}^c)}\|_{1,2} \\
&\leqslant \frac{1}{2a_1 n}\|\mathcal{W}\boldsymbol{\delta}\|_2^2 + \frac{a_1}{2n}\left[\frac{a_2}{1-a_2}\|\mathcal{W}\boldsymbol{\delta}\|_2^2 + \frac{7\big\{2s_e\log(ep)+s_g\log(eq/s_g)\big\}}{1-a_2}\right] + \lambda_e\|\boldsymbol{\delta}_{\mathcal{S}}\|_1 + \lambda_g\|\boldsymbol{\delta}_{(\mathcal{G})}\|_{1,2}
\end{aligned} \tag{S2.15}$$

for any constant $a_1 > 0$ and $0 < a_2 < 1$. Here, $a_2 = 14M_1/C^2$. Let $a_1 = 1, a_2 = 0.5$, then

(S2.15) can be written as

$$\frac{\|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1}{\sqrt{s_e}} + \frac{\|\boldsymbol{\delta}_{(\mathcal{G}^c)}\|_{1,2}}{\sqrt{s_g}} \leqslant 7\sqrt{\frac{2s_e \log(ep) + s_g \log(eq/s_g)}{n}} + \frac{\|\boldsymbol{\delta}_{\mathcal{S}}\|_1}{\sqrt{s_e}} + \frac{\|(\boldsymbol{\delta})_{(\mathcal{G})}\|_{1,2}}{\sqrt{s_g}}.$$

This completes the proof of Lemma 5.

We next prove Lemma 4. Define $\boldsymbol{\Sigma}_{\mathcal{W}} = \mathbb{E}\, \mathcal{W}^\top \mathcal{W}/n$, it is clear that $\lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{W}}) \geqslant \phi_1/\phi_0$. We first bound $\lambda_e \|\boldsymbol{\delta}_{\mathcal{S}}\|_1 + \lambda_e \|\boldsymbol{\delta}_{\mathcal{G}}\|_{1,2}$ as

$$\begin{aligned}\lambda_e \|\boldsymbol{\delta}_{\mathcal{S}}\|_1 + \lambda_e \|\boldsymbol{\delta}_{\mathcal{G}}\|_{1,2} &\leqslant (\lambda_e \sqrt{s_e} + \lambda_g \sqrt{s_g}) \|\boldsymbol{\delta}\|_2 \\ &\leqslant a_3 C \frac{\phi_1}{\phi_0} \frac{2s_e \log(ep) + s_g \log(eq/s_g)}{n} + \frac{\|\boldsymbol{\Sigma}_{\mathcal{W}} \boldsymbol{\delta}\|_2^2}{a_3},\end{aligned}$$

where the inequalities come from Cauchy-Schwarz inequality. Combining the inequality above with (S2.15), and letting $a_1 = 2$, $a_2 = 0.2$ and $a_3 = 3$, we have

$$\frac{1}{2n}\|\mathcal{W}\boldsymbol{\delta}\|_2^2 + \lambda_e \|\boldsymbol{\delta}_{\mathcal{S}_j^c}\|_1 + \lambda_g \|\boldsymbol{\delta}_{(\mathcal{G}_j^c)}\|_{1,2} \leqslant \left(\frac{35}{4} + 3C\frac{\phi_1}{\phi_0}\right) \frac{2s_e \log(ep) + s_g \log(eq/s_g)}{n} + \frac{\|\boldsymbol{\Sigma}_{\mathcal{W}}^{1/2} \boldsymbol{\delta}\|_2^2}{3}$$

with probability at least $1 - C_1 \exp\Big\{ - C_2 \frac{s_e \log(ep) + s_g \log(eq/s_g)}{s_e} \Big\}$. Therefore it holds that

$$\frac{\|\boldsymbol{\Sigma}_{\mathcal{W}}^{1/2} \boldsymbol{\delta}\|_2^2}{6} \precsim \frac{2s_e \log(ep) + s_g \log(eq/s_g)}{n} + \frac{1}{2}\left|\boldsymbol{\delta}\left(\frac{\mathcal{W}^\top \mathcal{W}}{n} - \boldsymbol{\Sigma}_{\mathcal{W}}\right)\boldsymbol{\delta}\right|$$

with probability at least $1 - C_1 \exp\Big\{ - C_2 \frac{s_e \log(ep) + s_g \log(eq/s_g)}{s_e} \Big\}$. From Lemma 8 and the Assumption 3, we have

$$\begin{aligned}\sup_{\|\mathbf{v}\|_0 \leqslant s_e, \|\mathbf{v}\|_2 \leqslant 1} |\mathbf{v}^\top (\widehat{\boldsymbol{\Sigma}}_{\mathcal{W}} - \boldsymbol{\Sigma}_{\mathcal{W}}) \mathbf{v}| &\leqslant \sup_{\|\mathbf{v}\|_0 \leqslant s_e, \|\mathbf{v}\|_2 \leqslant 1} \sum_j |\mathbf{v}_j^\top (\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_j} - \boldsymbol{\Sigma}_{\mathbf{W}_j}) \mathbf{v}_j| \\ &\precsim s_e \sqrt{\frac{\log p + \log q}{n}} + s_e \frac{\log^2 n \cdot \{s_e \log(epq) + s_g \log(eq/s_g)\}^2}{s_e^2 n},\end{aligned}$$

where the last inequality utilizes the sparsity on vector $\mathbf{v}$. By Assumption 3, the right side of the inequality above is a sufficiently small constant. Combining this with Lemma 12 in Loh and Wainwright (2011), it holds that with the same probability,

$$\frac{1}{2}\left|\boldsymbol{\delta}\left(\frac{\mathcal{W}^\top \mathcal{W}}{n} - \boldsymbol{\Sigma}_{\mathcal{W}}\right)\boldsymbol{\delta}\right| = o\bigg(\|\boldsymbol{\delta}\|_2^2 + \frac{1}{s_e}\|\boldsymbol{\delta}\|_1^2\bigg).$$

Hence we have that

$$\|\boldsymbol{\delta}\|_2^2 \precsim \frac{\|\boldsymbol{\Sigma}_{\mathcal{W}}^{1/2} \boldsymbol{\delta}\|_2^2}{6} \precsim \frac{2s_e \log(ep) + s_g \log(eq/s_g)}{n} + o\bigg(\|\boldsymbol{\delta}\|_2^2 + \frac{1}{s_e}\|\boldsymbol{\delta}_e\|_1^2\bigg).$$

Note also that from (S2.15), $\|\boldsymbol{\delta}\|_1/\sqrt{s_e} \precsim \sqrt{(2s_e \log(ep) + s_g \log(eq/s_g))/n} + \|\boldsymbol{\delta}\|_2$, we conclude that

$$\|\boldsymbol{\delta}\|_2^2 \precsim \frac{2s_e \log(ep) + s_g \log(eq/s_g)}{n} + o\big(\|\boldsymbol{\delta}\|_2^2\big)$$

with probability at least $1 - C_1 \exp\Big\{ - C_2 \frac{s_e \log(ep) + s_g \log(eq/s_g)}{s_e} \Big\}$. This indicates that

$$\|\boldsymbol{\delta}\|_2^2 \precsim \frac{2s_e \log(ep) + s_g \log(eq/s_g)}{n},$$

which completes the proof of Lemma 4.

S2.4 *Proof of Lemma 6*

It follows that

$$\begin{aligned}\frac{1}{\widehat{\sigma}^{jj}} &= \frac{\|\mathbf{z}_j - \mathbf{W}_j \widehat{\boldsymbol{\beta}}_j^{OLS}\|_2^2}{n - \widehat{s}_j} = \frac{1}{n - \widehat{s}_j} \cdot \big\|(\mathbf{W}_j)_{\mathcal{S}_j}(\boldsymbol{\beta}_j)_{\mathcal{S}_j} + \boldsymbol{\varepsilon}_j - \mathcal{P}_{\widehat{\mathcal{S}}_j}(\mathbf{W}_j)_{\mathcal{S}_j}(\boldsymbol{\beta}_j)_{\mathcal{S}_j} - \mathcal{P}_{\widehat{\mathcal{S}}_j}\boldsymbol{\varepsilon}_j\big\| \\ &= \frac{1}{n - \widehat{s}_j}\boldsymbol{\varepsilon}_j^\top(\mathbf{I}_n - \mathcal{P}_{\widehat{\mathcal{S}}_j})\boldsymbol{\varepsilon}_j,\end{aligned}$$

where the second equality comes from the definition of $\mathcal{P}_{\widehat{\mathcal{S}}_j}$ and $\mathcal{S}_j \subseteq \widehat{\mathcal{S}}_j$. From the proof in Section S2.3, (S2.14) indicates that $\boldsymbol{\varepsilon}_j^\top \mathcal{P}_{\widehat{\mathcal{S}}_j}\boldsymbol{\varepsilon}_j \precsim s_e \log(ep) + s_g \log(eq/s_g) = o(n)$. Combining the results that $\widehat{s}_j \leqslant s_\lambda = o(n)$, we conclude that

$$\frac{\widehat{\sigma}^{jj}}{\sigma^{jj}} \xrightarrow{p} 1$$

as $n \to \infty$, which completes the proof.

S3. Additional Simulation Results

S3.1 *Varying Sample Sizes*

We provide additional results considering sample sizes $n = 200$ and $n = 800$.

As shown in Tables S.1 and S.2, the SAGE estimator demonstrates a strong performance in finite sample cases, even performing comparably to the oracle estimator $\widehat{\boldsymbol{\beta}}^{oracle}$, highlighting the effectiveness of our approach. However, as the sample size n decreases, the coverage probability of confidence intervals also declines, as expected. With smaller n, the asymptotic results are less likely to hold, leading to a greater discrepancy between the empirical variance and the variance based on asymptotic theory. This difference results in reduced coverage probability for confidence intervals constructed using large-sample theory.

[Table S.1 about here.]

[Table S.2 about here.]

S3.2 *Comparison with the decorelation method*

A natural application of our debiasing method is point-wise hypothesis testing, which allows for a direct comparison with the inferential framework of Ning and Liu (2017). We still implement their methods in node wise regression by treating Z_j as the response and $\mathbf{W}_j$ as the design matrix. For a fair evaluation, we report the true positive rate (TPR) and the false positive rate (FPR) across multiple replications. Both procedures are benchmarked under a practical selection then inference protocol: SAGE debiases and tests the nonzero coordinates selected by the sparse group lasso, while the decorrelation method of Ning and Liu (2017) is applied to the coordinates selected by lasso. TPR and FPR are then computed against the ground truth, and the results are summarized in Table S.3.

Table S.3 reports the empirical performance of our proposed SAGE estimator against the decorrelation method of Ning and Liu (2017). Both procedures achieve high true positive rates (TPR), and the power of SAGE rapidly approaches one as the sample size increases. Moreover, SAGE delivers substantially smaller false positive rates (FPR) across all settings. For instance, when $n = 400$, both methods achieve nearly perfect TPR, but the FPR of SAGE is close to zero, whereas the decorrelation method yields an FPR an order of magnitude larger. Even in the small-sample regime ($n = 200$), SAGE maintains competitive power while significantly reducing false discoveries. With a smaller significance level α, the TPR of both methods remains high and close to one, but the FPR of SAGE is much lower than that of the decorrelation method, demonstrating that SAGE provides more accurate inference. Overall, these results highlight the practical advantage of SAGE in achieving reliable inference with strong power and superior control of spurious rejections. In our covariate-dependent framework, the distribution of the design matrix is more intricate than the standard sub-Gaussian case typically assumed in literature. While decorrelation-based procedures are powerful in many conventional settings, our results suggest that methods like SAGE are particularly well-suited to handle the added structural complexities present here, making them a natural choice for this type of covariate-dependent setting.

[Table S.3 about here.]

S3.3 *Additional Results with Unknown Means*

We follow the two-step procedure proposed by Zhang and Li (2023) to estimate $\boldsymbol{\beta}$ when $\boldsymbol{\Gamma}$ is unknown. In Step 1, we assume that $\boldsymbol{\Gamma}$ is sparse and obtain its estimate via ℓ_1-penalized regression. In Step 2, we construct the approximation

$$\widehat{\mathbf{z}}^{(i)} = \mathbf{x}^{(i)} - \widehat{\boldsymbol{\Gamma}}\mathbf{u}^{(i)},$$

where $\widehat{\boldsymbol{\Gamma}}$ is the estimator from Step 1. With the resulting pairs $\left(\widehat{\mathbf{z}}^{(i)}, \mathbf{u}^{(i)}\right)$, we then apply the debiasing procedure described in Section 3. Specifically,

Step 1: Denote the covariate matrix $\mathbf{U}$ by $\mathbf{U} = (\mathbf{u}^{(1)}, ..., \mathbf{u}^{(n)})^\top \in \mathbb{R}^{n\times q}$ and the sample of j-the variable $\mathbf{x}_j$ by $(x_j^{(1)}, ..., x_j^{(n)})^\top$. We first estimate $\boldsymbol{\Gamma}$ with

$$\widehat{\boldsymbol{\gamma}}_j = \mathrm{argmin}_{\boldsymbol{\gamma}\in\mathbb{R}^q} \frac{1}{2n}\|\mathbf{x}_j - \mathbf{U}\boldsymbol{\gamma}\|_2^2 + \lambda_u\|\boldsymbol{\gamma}\|_1, \tag{S3.1}$$

and $\widehat{\boldsymbol{\Gamma}} = (\widehat{\boldsymbol{\gamma}}_1, ..., \widehat{\boldsymbol{\gamma}}_p)^\top$.

Step 2: With $\widehat{\boldsymbol{\Gamma}}$ estimated in Step 1, we obtain $\widehat{\mathbf{z}}^{(i)} = \mathbf{x}^{(i)} - \widehat{\boldsymbol{\Gamma}}\mathbf{u}^{(i)}$, then we can get the estimated $\widehat{\mathbf{W}}$ with expression

$$\widehat{\mathbf{W}}_j = [\widehat{\mathbf{z}}_1, ..., \widehat{\mathbf{z}}_{j-1}, \widehat{\mathbf{z}}_{j+1}, ..., \widehat{\mathbf{z}}_p, \widehat{\mathbf{z}}_1 \odot \mathbf{u}_1, ..., \widehat{\mathbf{z}}_{j-1} \odot \mathbf{u}_1, \widehat{\mathbf{z}}_{j+1} \odot \mathbf{u}_1, ..., \widehat{\mathbf{z}}_p \odot \mathbf{u}_q] \in \mathbb{R}^{n\times(p-1)(q+1)},$$

where $\widehat{\mathbf{z}}_j = (\widehat{z}_j^{(1)}, ..., \widehat{z}_j^{(n)})$. The estimated vector $\widehat{\boldsymbol{\beta}}$ is given by $\widehat{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}} \frac{1}{2n}\left\|\widehat{\mathbf{y}} - \widehat{\mathcal{W}}\boldsymbol{\beta}\right\|_2^2 + \lambda_e\|\boldsymbol{\beta}\|_1 + \lambda_g \sum_{h=1}^q \|\boldsymbol{b}_h\|_2$, where $\widehat{\mathcal{W}} = \mathrm{diag}(\widehat{\mathbf{W}}_1, ..., \widehat{\mathbf{W}}_p)$, $\widehat{\mathbf{y}} = (\widehat{z}_1^{(1)}, ..., \widehat{z}_1^{(n)}, \widehat{z}_2^{(1)}, ..., \widehat{z}_2^{(n)}, ..., z_p^{(n)})^\top \in \mathbb{R}^{np}$ and $\boldsymbol{b}_h$ is defined in Section 2.2.

The two steps involve three tuning parameters λ_u, λ_e and λ_g. In Step 1, we select λ_u via 5-fold cross validation and obtain the corresponding estimator $\widehat{\mathbf{\Gamma}}$. With this $\widehat{\mathbf{\Gamma}}$, we calculate $\widehat{\mathbf{z}}^{(i)}$ and then proceed to Step 2. The remaining tuning parameters λ_e and λ_g are chosen in the same way as in Section 4, and the subsequent estimation follows the same procedure. In the simulation design, we set $(\boldsymbol{\gamma}_1)_2 = 2$ and $(\boldsymbol{\gamma}_2)_3 = -2$ to induce sparsity in $\mathbf{\Gamma}$, while all other settings remain the same as in the previous simulation studies.

As shown in Table S.4, when $\mathbf{\Gamma}$ is unknown, the two-step procedure still achieves a relatively high rejection rate at zero. Moreover, the coverage probability after debiasing remains around 95%, indicating that the proposed procedure performs well in practice with unknown $\mathbf{\Gamma}$, providing both accurate coverage and competitive power.

However, several theoretical difficulties arise. The first challenge comes from the proof of existence of the solution of equation (6) in our manuscript, which gives us the matrix $\widehat{\mathbf{M}}_j$. Our previous proof relies on the independence of each row in $(\mathbf{W}_j)_{i\cdot}$. However, once $\widehat{\mathbf{\Gamma}}$ and $\widehat{\mathbf{z}}_j$ are introduced, this independence is violated, leading to complicated dependence structures across both rows and columns. The second challenges come from the debiasing step itself. When $\mathbf{\Gamma}$ is replaced by its estimate $\widehat{\mathbf{\Gamma}}$, the resulting $\widehat{\mathbf{W}}_j$ becomes intricately dependent on $\boldsymbol{\varepsilon}_j$, and the debiasing methods are no longer directly applicable when conditioning on $\mathbf{W}_j$. In addition, one must carefully account for the extra error terms arising from the discrepancies between $\widehat{\mathbf{\Gamma}}$ and $\mathbf{\Gamma}$, as well as between $\widehat{\mathbf{W}}_j$ and $\mathbf{W}_j$. One possible way to address the dependence is through sample splitting, $\widehat{\mathbf{\Gamma}}$ can be treated as fixed when conditional on auxiliary sample set. By following the technique of Zhang and Li (2023), we can establish the convergence rate, and potentially extend the same approach to develop valid inference methods. This strategy, however, typically loses some efficiency, and a full investigation of it lies beyond the scope of the present paper. We think it is an interesting and promising future direction.

[Table S.4 about here.]

S3.4 *Different pairs of* (n, p, q)

We conduct more simulations by varying (n, p, q) to show the consistency of our results. Table S.5, S.6 and S.7 give the results with sample size $n = 200$, 400 and 800 for various (p, q), respectively.

[Table S.5 about here.]

[Table S.6 about here.]

[Table S.7 about here.]

S3.5 *Comparison with Computational Time*

We provide the comparison of computation time with simulations conducted on a MAC Pro with M3 Pro chips.

[Table S.8 about here.]

S4. Additional Real Data Results

We show the effects of "rs6701524", "rs723210", and "rs503314" in Figure S.1, and "rs9303511", "rs728655", and "rs306098" in Figure S.2.

[Figure S.1 about here.]

[Figure S.2 about here.]

References

Banerjee, S., Saunderson, J., Srivastava, R., and Rajwade, A. (2025). Fast debiasing of the lasso estimator. *arXiv preprint arXiv:2502.19825* .

Cai, T. T., Zhang, A. R., and Zhou, Y. (2022). Sparse group lasso: Optimal sample complexity, convergence rate, and statistical inference. *IEEE Transactions on Information Theory* **68,** 5975–6002.

Kuchibhotla, A. K. and Chakrabortty, A. (2022). Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA* **11,** 1389–1456.

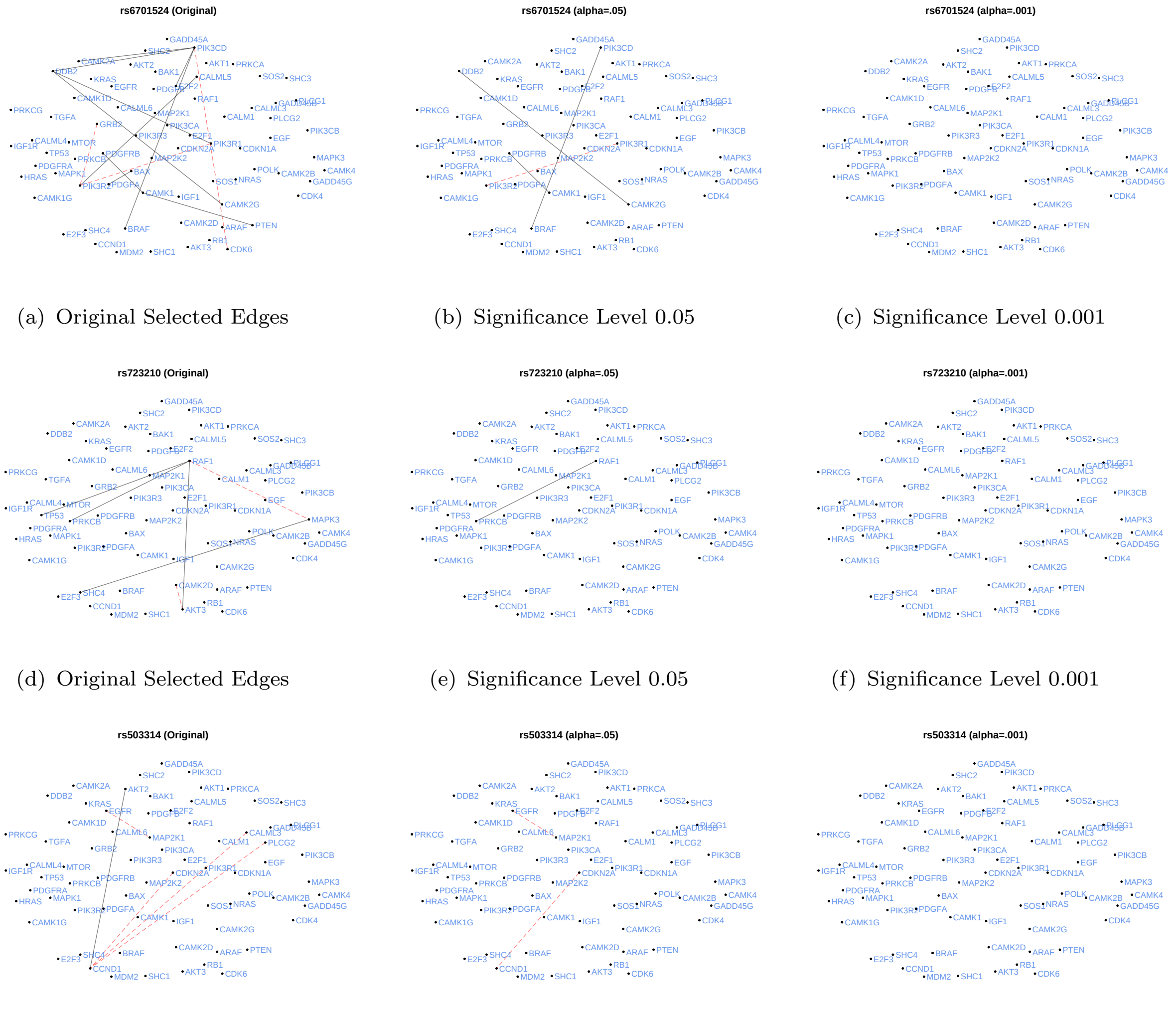
Loh, P.-L. and Wainwright, M. J. (2011). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Advances in Neural Information Processing Systems* **24,**.

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *THE ANNALS* **45,** 158–195.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* .

Zhang, J. and Li, Y. (2023). High-dimensional Gaussian graphical regression models with covariates. *Journal of the American Statistical Association* **118,** 2088–2100.

Zhang, J. and Li, Y. (2025). Multi-task learning for Gaussian graphical regressions with high dimensional covariates. *Journal of Computational and Graphical Statistics* **34,** 961–970.

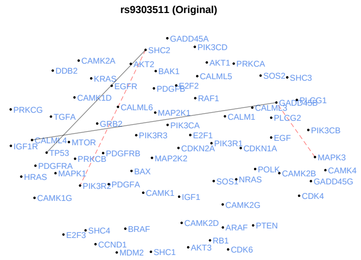


(a) Original Selected Edges (b) Significance Level 0.05 (c) Significance Level 0.001

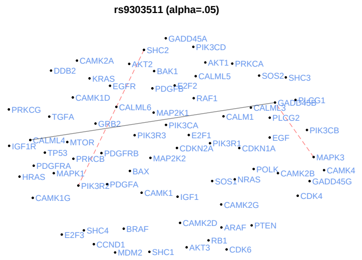
(d) Original Selected Edges (e) Significance Level 0.05 (f) Significance Level 0.001

(g) Original Selected Edges (h) Significance Level 0.05 (i) Significance Level 0.001

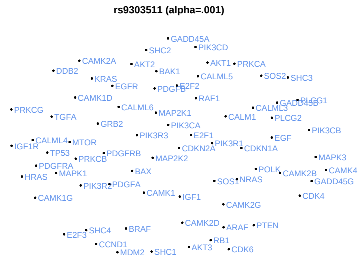
Figure S.1: Additional results of the effects of different SNPs on the gene co-expression. Positive partial correlations are shown with red dashed lines, while negative correlations are indicated by black solid lines.



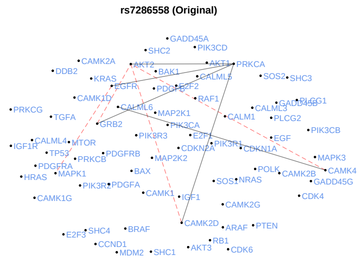
(a) Original Selected Edges



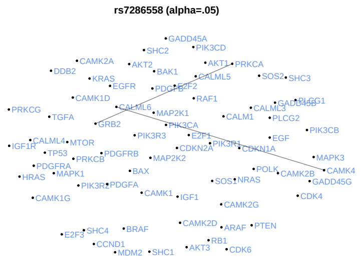
(b) Significance Level 0.05



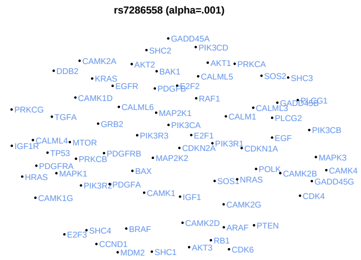
(c) Significance Level 0.001



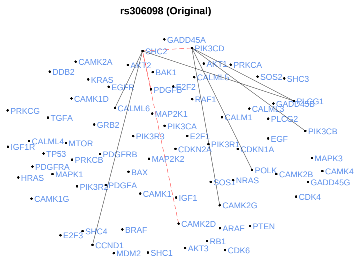
(d) Original Selected Edges



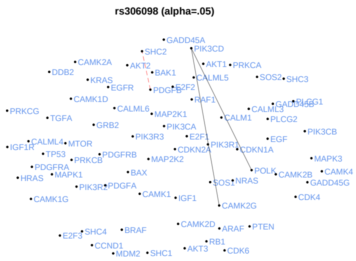
(e) Significance Level 0.05



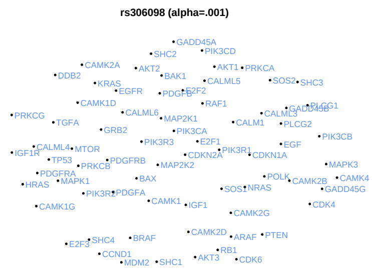
(f) Significance Level 0.001



(g) Original Selected Edges



(h) Significance Level 0.05



(i) Significance Level 0.001

Figure S.2: Additional results of the effects of different SNPs on the gene co-expression. Positive partial correlations are shown with red dashed lines, while negative correlations are indicated by black solid lines.

Table S.1: Results for sample size $n = 200$.

(λ_e, λ_g)		$\boldsymbol{\beta}_{\text{ind}_1}$	$\boldsymbol{\beta}_{\text{ind}_2}$	$\boldsymbol{\beta}_{\text{ind}_3}$	$\boldsymbol{\beta}_{\text{ind}_4}$
(.3,.212)	Pre-Bias	.177(.052)	.175(.054)	.178(.049)	.175(.053)
	Post-Bias	.051(.073)	.050(.075)	.055(.068)	.053(.073)
	EmpSD	1.23	1.30	1.22	1.30
	Cov-Prob	81.0%	81.0%	79.0%	79.0%
	Rej-0	95.5%	96.5%	98.0%	96.0%
(.6,.414)	Pre-Bias	.281(.018)	.280(.018)	.281(.018)	.280(.018)
	Post-Bias	.024(.057)	.025(.060)	.031(.056)	.029(.055)
	Emp-SD	1.02	1.07	1.04	1.04
	Cov-Prob	92.5%	91.5%	91.5%	92.5%
	Rej-0	99.5%	99.5%	99.5%	100%
Cross Validation	Pre-Bias	.227(.059)	.225(.062)	.229(.056)	.225(.062)
	Post-Bias	.023(.064)	.022(.066)	.030(.062)	.026(.065)
	Emp-SD	1.28	1.34	1.28	1.30
	Cov-Prob	88.0%	86.5%	85.5%	84.5%
	Rej-0	99.5%	100%	99.5%	100%
	AIL	.197(.009)	.197(.009)	.195(.009)	.195(.008)
$\widehat{\boldsymbol{\beta}}^{oracle}$	Bias	.001(.065)	-.002(.066)	.006(.060)	.003(.064)
	Emp-SD	.917	.902	.849	.902
	Cov-Prob	96.5%	98.0%	97.0%	97.0%
	Rej-0	99%	99%	99.5%	99.5%
	AIL	.277			
Post Selection	Bias	.079(.073)	.079(.077)	.080(.070)	.082(.077)
	Cov-Prob	96.5%	92.5%	91.5%	93.0%
	AIL	2.26(1.73)	2.42(1.83)	2.23(1.63)	2.38(1.88)
	Rej-0	32.0%	27.5%	29.5%	26.0%

Table S.2: Simulation results with $n = 800$.

(λ_e, λ_g)		$\boldsymbol{\beta}_{\text{ind}_1}$	$\boldsymbol{\beta}_{\text{ind}_2}$	$\boldsymbol{\beta}_{\text{ind}_3}$	$\boldsymbol{\beta}_{\text{ind}_4}$
(.3,.212)	Pre-Bias	.142(.028)	.140(.027)	.142(.028)	.141(.027)
	Post-Bias	.003(.033)	.002(.034)	.004(.032)	.003(.033)
	EmpSD	1.16	1.16	1.11	1.14
	Cov-Prob	92.5%	93.0%	93.0%	93.5%
	Rej-0	100%	100%	100%	100%
(.6,.414)	Pre-Bias	.278(.015)	.278(.015)	.278(.015)	.278(.015)
	Post-Bias	.005(.031)	.005(.031)	.007(.030)	.006(.031)
	Emp-SD	1.01	1.01	0.96	1.00
	Cov-Prob	94.5%	95.0%	95.0%	95.0%
	Rej-0	100%	100%	100%	100%
	AIL	.109(.003)			
$\widehat{\boldsymbol{\beta}}^{oracle}$	Bias	.000(.033)	-.001(.033)	.001(.032)	.000(.032)
	Emp-SD	.933	.931	.903	.895
	Cov-Prob	97.0%	96.0%	97.5%	98.0%
	Rej-0	100%	100%	100%	100%
	AIL	.139			
Post Selection	Bias	.019(.037)	.018(.041)	.022(.036)	.018(.037)
	Cov-Prob	94.5%	96.5%	95.0%	96.0%
	AIL	0.880(.652)	.848(.679)	.834(.680)	.873(.751)
	Rej-0	62.0%	56.0%	66.5%	63.5%

Table S.3: Compared Results for SAGE and Decorelation.

Sample size	Methods	Significanve level $\alpha = 0.05$			
		Mean of TPR	Std of TPR	Mean of FPR ($\times 10^{-3}$)	Std of FPR ($\times 10^{-3}$)
$n = 200$	SAGE	.840	.268	.079	.255
	Ning and Liu (2017)	.913	.209	1.35	1.55
$n = 400$	SAGE	.998	.035	.036	.107
	Ning and Liu (2017)	1	0	1.55	1.93
$n = 800$	SAGE	1	0	.008	.030
	Ning and Liu (2017)	1	0	1.40	1.90
Sample size	Methods	Significanve level $\alpha = 0.001$			
		Mean of TPR	Std of TPR	Mean of FPR ($\times 10^{-3}$)	Std of FPR ($\times 10^{-3}$)
$n = 200$	SAGE	.840	.268	.050	.123
	Ning and Liu (2017)	.818	.284	.483	.551
$n = 400$	SAGE	.998	.035	.031	.087
	Ning and Liu (2017)	.995	.050	.575	.514
$n = 800$	SAGE	1	0	.006	.262
	Ning and Liu (2017)	1	0	.566	.479

Table S.4: Simulation results with $\boldsymbol{\Gamma}$ unknown.

Sample size	(λ_e, λ_g)		$\boldsymbol{\beta}_{\text{ind}_1}$	$\boldsymbol{\beta}_{\text{ind}_2}$	$\boldsymbol{\beta}_{\text{ind}_3}$	$\boldsymbol{\beta}_{\text{ind}_4}$
$n = 400$	(.3,.212)	Pre-Bias	.151(.042)	.145(.039)	.149(.043)	.143(.038)
		Post-Bias	.012(.055)	.006(.052)	.011(.057)	.004(.048)
		EmpSD	1.24	1.18	1.28	1.09
		Cov-Prob	87.0%	91.0%	88.0%	91.5%
		Rej-0	100%	100%	100%	100%
	(.6,.414)	Pre-Bias	.281(.017)	.279(.016)	.281(.017)	.279(.016)
		Post-Bias	.009(.050)	.003(.048)	.008(.052)	.003(.044)
		EmpSD	1.07	1.02	1.09	0.94
		Cov-Prob	94.5%	94.5%	93.0%	95.0%
		Rej-0	100%	100%	100%	100%
		AIL	.165(.006)	.165(.006)	.166(.007)	.166(.007)
$n = 800$	(.3,.212)	Pre-Bias	.146(.027)	.140(.026)	.145(.027)	.139(.027)
		Post-Bias	.001(.032)	-.004(.033)	.002(.032)	-.005(.035)
		EmpSD	1.05	1.07	1.06	1.15
		Cov-Prob	94.0%	94.5%	93.0%	92.0%
		Rej-0	100%	100%	100%	100%
	(.6,.414)	Pre-Bias	.281(.015)	.277(.015)	.281(.015)	.276(.015)
		Post-Bias	-.004(.031)	-.008(.032)	-.002(.031)	-.008(.034)
		EmpSD	0.93	0.96	0.95	1.03
		Cov-Prob	95.5%	97.0%	96.0%	96.0%
		Rej-0	100%	100%	100%	100%
		AIL	.116(.003)	.116(.003)	.116(.003)	.116(.003)

Table S.5: Results for sample size $n = 200$.

(p, q)	(λ_e, λ_g)		$\boldsymbol{\beta}_{\text{ind}_1}$	$\boldsymbol{\beta}_{\text{ind}_2}$	$\boldsymbol{\beta}_{\text{ind}_3}$	$\boldsymbol{\beta}_{\text{ind}_4}$
$(120, 20)$	$(.3, .212)$	Pre-Bias	.156(.049)	.150(.049)	.154(.049)	.148(.049)
		Post-Bias	.044(.062)	.038(.068)	.041(.062)	.034(.067)
		Emp-SD	1.21	1.35	1.20	1.31
		Cov-Prob	81.5%	80.5%	83.5%	79.5%
		Rej-0	100%	99.5%	100%	99.5%
	$(.6, .424)$	Pre-Bias	.281(.018)	.277(.018)	.281(.018)	.277(.018)
		Post-Bias	.032(.044)	.028(.058)	.029(.043)	.024(.056)
		Emp-SD	0.87	1.14	0.85	1.09
		Cov-Prob	92%	88.5%	94.5%	89%
		Rej-0	100%	100%	100%	100%
	Cross Validation	Pre-Bias	.182(.066)	.175(.064)	.182(.067)	.173(.066)
		Post-Bias	.030(.059)	.023(.066)	.067(.028)	.019(.064)
		Emp-SD	1.18	1.30	1.18	1.26
		Cov-Prob	86%	85.5%	85.5%	84.5%
		Rej-0	100%	100%	100%	100%
$(120, 20)$	$\widehat{\boldsymbol{\beta}}^{oracle}$	Bias	.010(.064)	.002(.063)	.007(.065)	-.004(.064)
		Emp-SD	0.90	0.90	0.92	0.91
		Cov-Prob	98%	98.5%	97%	98%
		Rej-0	99.5%	99.5%	100%	100%
$(20, 120)$	$(.3, .212)$	Pre-Bias	.165(.057)	.160(.054)	.163(.059)	.160(.055)
		Post-Bias	.030(.069)	.026(.074)	.026(.072)	.023(.073)
		Emp-SD	1.30	1.36	1.38	1.36
		Cov-Prob	81.5%	85.5%	82.5%	81.5%
		Rej-0	99.5%	99.5%	99.5%	99.5%
	$(.6, .424)$	Pre-Bias	.281(.018)	.280(.018)	.281(.018)	.280(.018)
		Post-Bias	.021(.049)	.020(.060)	.017(.049)	.016(.060)
		Emp-SD	0.92	1.13	0.92	1.12
		Cov-Prob	95%	90.5%	96.5%	92%
		Rej-0	99.5%	99.5%	100%	100%
	Cross Validation	Pre-Bias	.157(.067)	.154(.065)	.155(.069)	.153(.066)
		Post-Bias	.023(.064)	.020(.067)	.020(.069)	.016(.067)
		Emp-SD	1.20	1.26	1.28	1.26
		Cov-Prob	87%	88.5%	85%	86%
		Rej-0	99.5%	99.5%	100%	100%
$(20, 120)$	$\widehat{\boldsymbol{\beta}}^{oracle}$	Bias	.005(.064)	.001(.064)	.000(.069)	-.004(.065)
		Emp-SD	0.91	0.91	0.97 133	0.91
		Cov-Prob	96.5%	95.5%	95%	97%
		Rej-0	99%	99.5%	99%	100%

Table S.6: Simulation results with $n = 400$.

(p, q)	(λ_e, λ_g)		$\boldsymbol{\beta}_{\text{ind}_1}$	$\boldsymbol{\beta}_{\text{ind}_2}$	$\boldsymbol{\beta}_{\text{ind}_3}$	$\boldsymbol{\beta}_{\text{ind}_4}$
$(120, 20)$	$(.3, .212)$	Pre-Bias	.137(.036)	.143(.037)	.138(.036)	.143(.038)
		Post-Bias	.005(.047)	.011(.046)	.006(.045)	.011(.046)
		Emp-SD	1.22	1.20	1.18	1.21
		Cov-Prob	87.5%	89.5%	89%	85.5%
	$(.6, .424)$	Pre-Bias	.278(.016)	.280(.017)	.278(.016)	.280(.017)
		Post-Bias	.013(.033)	.014(041)	.013(.032)	.015(.042)
		Emp-SD	0.87	1.08	0.83	1.11
		Cov-Prob	97%	91.5%	97.5%	90%
	Cross Validation	Pre-Bias	.116(.048)	.122(.047)	.117(.046)	.122(.047)
		Post-Bias	.004(.047)	.010(.046)	.005(.045)	.010(.046)
		Emp-SD	1.23	1.20	1.17	1.20
		Cov-Prob	89%	89.5%	89.5%	86.5%
$(120, 20)$	$\widehat{\boldsymbol{\beta}}^{oracle}$	Bias	-.007(.046)	.001(.045)	-.003(.044)	.002(.045)
		Emp-SD	0.91	0.89	0.88	0.90
		Cov-Prob	98%	98.5%	97.5%	99%
$(20, 120)$	$(.3, .212)$	Pre-Bias	.142(.035)	.146(.040)	.143(.035)	.148(.040)
		Post-Bias	.002(.044)	.006(.049)	.006(.043)	.010(.047)
		Emp-SD	1.11	1.23	1.09	1.20
		Cov-Prob	94%	90%	93.5%	88.5%
	$(.6, .424)$	Pre-Bias	.279(.017)	.279(.017)	.279(.017)	.279(.016)
		Post-Bias	.009(.035)	.009(.045)	.013(.034)	.012(.044)
		Emp-SD	0.89	1.14	0.85	1.12
		Cov-Prob	96.5%	89.5%	96.5%	92%
	Cross Validation	Pre-Bias	.104(.043)	.108(.048)	.107(.043)	.111(.048)
		Post-Bias	.004(.044)	.008(.048)	.008(.043)	.012(.048)
		Emp-SD	1.11	1.22	1.10	1.21
		Cov-Prob	93.5%	90.5%	93%	87.5%
$(20, 120)$	$\widehat{\boldsymbol{\beta}}^{oracle}$	Bias	-.003(.043)	.001(.047)	.002(.043)	.006(.046)
		Emp-SD	0.86	0.94	0.86	0.91
		Cov-Prob	98.5%	95.5%	97.5%	96%

Table S.7: Simulation results with $n = 800$.

(p, q)	(λ_e, λ_g)		$\boldsymbol{\beta}_{\text{ind}_1}$	$\boldsymbol{\beta}_{\text{ind}_2}$	$\boldsymbol{\beta}_{\text{ind}_3}$	$\boldsymbol{\beta}_{\text{ind}_4}$
$(120, 20)$	$(.3, .212)$	Pre-Bias	.144(.029)	.139(.026)	.144(.028)	.139(.027)
		Post-Bias	005(.034)	-.001(.031)	.006(.033)	.001(.032)
		Emp-SD	1.21	1.10	1.17	1.16
		Cov-Prob	91%	91.5%	90%	91%
$(20, 120)$	$(.3, .212)$	Pre-Bias	.141(.025)	.140(.028)	.143(.026)	.141(.027)
		Post-Bias	.001(.030)	-.000(.034)	.005(.031)	.003(.032)
		Emp-SD	1.07	1.18	1.11	1.13
		Cov-Prob	93%	87.5%	92%	90.5%

Table S.8: Computation time in seconds with varying n, p while $q = 20$.

Time (s)	$n = 50$			$n = 100$			$n = 200$		
	$p = 20$	$p = 50$	$p = 100$	$p = 20$	$p = 50$	$p = 100$	$p = 20$	$p = 50$	$p = 100$
(6)	1.36	15.76	99.14	1.55	17.52	117.92	1.25	12.56	98.38
(7)	0.11	0.10	0.13	0.16	0.17	0.29	0.26	0.57	0.67