Submitted to the Bernoulli

Supplement to "Inference in latent factor regression with clusterable features"

XIN BING[1,*] FLORENTINA BUNEA[1,**] and MARTEN WEGKAMP[1,2]

[1]*Department of Statistics and Data Science, Cornell University, Ithaca, New York, USA.*
E-mail: [*]xb43@cornell.edu; [**]fb238@cornell.edu

[2]*Department of Mathematics, Cornell University, Ithaca, New York, USA.*
E-mail: mhw73@cornell.edu

Contents

Appendix A: The identifiability of A when Γ is known

We state and prove the identifiability of A when Γ is known under Assumption 1 but when *(A1)* is replaced by

(A1') For each $k \in [K]$, there exists at least one index $i \in [p]$ such that $A_{i\bullet} = \mathbf{e}_k$.

Corollary 1. *Under (A0), (A2) of Assumption 1 and (A1'), suppose Γ is known. Then the matrix A is identifiable from Σ, up to a $K \times K$ signed permuation matrix.*

Proof. When Γ is known, one can identify $A\Sigma_Z A^\top = \Sigma - \Gamma$. If I can be identified, then A is identifiable by the proof of Theorem 2 in [2]. It remains to show that I is identifiable from $A\Sigma_Z A^\top$. This can be shown by repeating the same arguments of proving [2, Theorem 1] except that we replace the definitions in (2.2) and (2.3) of [2] by

$$M_i := \arg\max_{j\in[p]} |\Sigma_{ij}|, \quad S_i := \{j \in [p] : |\Sigma_{ij}| = M_i\}$$

for each $1 \le i \le p$. □

Appendix B: Proofs of Proposition 1 and Theorem 2

B.1. Proof of Proposition 1: the identifiability of β

From the structure of Σ together with Assumption 1, Theorem 1 in [2] can be directly invoked to show that I and its partition $\mathcal{I}$ are identifiable up to a label permutation. In

addition, A is identifiable up to a signed permutation. Suppose we identify $\widetilde{A} = AP$ for some signed permutation P and, in particular, $\widetilde{A}_{I\bullet} = A_{I\bullet}P$.

First observe that for any $a, b \in [K]$, $[\Sigma_Z]_{ab}$ is recovered by

$$[\Sigma_Z]_{ab} = A_{ia}A_{jb}\Sigma_{ij}, \quad \text{for } i \in I_a,\ j \in I_b,\ i \neq j. \tag{1}$$

We prove this as follows. Since $\Sigma = A\Sigma_Z A^\top + \Gamma$ with Γ diagonal,

$$\Sigma_{ij} = \sum_{c,d=1}^{K} A_{ic}[\Sigma_Z]_{cd}A_{jc} = A_{ia}A_{jb}[\Sigma_Z]_{ab},$$

where in the second step we use that $i \in I_a$, $j \in I_b$. Then (1) follows from $|A_{ia}| = |A_{jb}| = 1$.

From $\widetilde{A}$ and the corresponding partition $\{\widetilde{I}_a\}_{a\in[K]}$, we can define $\widetilde{\Sigma}_Z$ to be the matrix with elements $[\widetilde{\Sigma}_Z]_{ab} = \widetilde{A}_{i_a a}\widetilde{A}_{j_b b}\Sigma_{i_a j_b}$ for some $i_a \in \widetilde{I}_a$, $j_b \in \widetilde{I}_b$, and $i_a \neq j_b$. Let $\pi : [K] \to [K]$ be the permutation mapping corresponding to P. Specifically, $\pi(a)$ equals the unique $b \in [K]$ such that $|P_{ba}| = 1$. Then for any $i \in [p]$ and $a \in [K]$, $\widetilde{A} = AP$ implies $\widetilde{A}_{ia} = P_{\pi(a)a}A_{i\pi(a)}$ and $\widetilde{I}_a = I_{\pi(a)}$, so

$$[\widetilde{\Sigma}_Z]_{ab} = P_{\pi(a)a}P_{\pi(b)b}A_{i_a\pi(a)}A_{j_b\pi(b)}\Sigma_{i_a j_b} = P_{\pi(a)a}P_{\pi(b)b}[\Sigma_Z]_{\pi(a)\pi(b)} = [P^\top\Sigma_Z P]_{ab},$$

where we use (1) in the second equality since $i_a \in I_{\pi(a)}$, $j_b \in I_{\pi(b)}$, $i_a \neq j_b$. This shows $\widetilde{\Sigma}_Z = P^\top\Sigma_Z P$. Finally, we have

$$\widetilde{\beta} = \widetilde{\Sigma}_Z^{-1}(\widetilde{A}_{I\bullet}^\top\widetilde{A}_{I\bullet})^{-1}\widetilde{A}_{I\bullet}^\top\mathrm{Cov}(X_I, Y) = \widetilde{\Sigma}_Z^{-1}(\widetilde{A}_{I\bullet}^\top\widetilde{A}_{I\bullet})^{-1}\widetilde{A}_{I\bullet}^\top A_{I\bullet}\Sigma_Z\beta = P^\top\beta$$

by using $\widetilde{A}_{I\bullet} = A_{I\bullet}P$ and $\widetilde{\Sigma}_Z = P^\top\Sigma_Z P$ in the last step. This concludes the proof. □

B.2. Proof of Theorem 2: the minimax lower bounds for estimators of β

Let $\mathbb{P}_\beta$ and $\mathbb{P}_{\beta'}$ denote the joint distribution of (X_i, Y_i) for $i = 1, \ldots, n$, parametrized by the same (A, Σ_Z) but different β and β', respectively. Denote by $\mathrm{KL}(\mathbb{P}_\beta, \mathbb{P}_{\beta'})$ the Kullback-Leibler divergence between these two distributions. Since

$$\begin{aligned}\sup_{(\beta,A,\Sigma_Z)\in\mathcal{S}(R,m)} \mathbb{P}_{A,\Sigma_Z,\beta}\left\{\|\widehat{\beta} - \beta\| \geq c'\left(1 \vee \frac{R}{\sqrt{m}}\right)\cdot\sqrt{\frac{K}{n}}\right\} \\ \geq \sup_{\|\beta\|\leq R} \mathbb{P}_{A^*,\Sigma_Z^*,\beta}\left\{\|\widehat{\beta} - \beta\| \geq c'\left(1 \vee \frac{R}{\sqrt{m}}\right)\cdot\sqrt{\frac{K}{n}}\right\},\end{aligned}$$

for any fixed A^* and Σ_Z^*, we let $\mathbb{P}_\beta := \mathbb{P}_{A^*,\Sigma_Z^*,\beta}$ for A^* and Σ_Z^* defined below and aim to prove

$$\inf_{\widehat{\beta}} \sup_{\|\beta\|\leq R} \mathbb{P}_\beta\left\{\|\widehat{\beta} - \beta\| \geq c'\left(1 \vee \frac{R}{\sqrt{m}}\right)\cdot\sqrt{\frac{K}{n}}\right\} \geq c''.$$

We choose Σ_Z^* such that $0 < C_{\min} \le \lambda_{\min}(\Sigma_Z^*) \le \lambda_{\max}(\Sigma_Z^*) \le C_{\max} < \infty$ and

$$A^* := \begin{bmatrix} \mathbf{1}_m \otimes \mathbf{I}_K \\ 0 \end{bmatrix} \tag{2}$$

with $\otimes$ denoting the kronecker product and $\mathbf{1}_d$ denoting the vector in $\mathbb{R}^d$ with all ones.

We start by constructing a set of hypothesis S for β. From Lemma 2, stated in Section B.3. with $s = k = K-1$, we can find a subset S_0 of the set of binary sequences $\{0,1\}^{K-1}$ such that

(i) $\log |S_0| \ge c_1(K-1)$,
(ii) $c_2(K-1) \le \|a\|_0 \le (K-1)$, for all $a \in S_0$.
(iii) $\|a-b\|^2 \ge c_3(K-1)$, for all $a, b \in S_0$ and $a \ne b$,

where $c_1, c_2, c_3 > 0$ are absolute constants. Let $v^{(0)} = (1, 0, \ldots, 0) \in \mathbb{R}^K$ and $v^{(j)} = (0, a) \in \mathbb{R}^K$ for all $a \in S_0$ so that $j \in \{1, \ldots, |S_0|\}$. We then define $\beta^{(0)} = (R, 0, \ldots, 0)$ and

$$\beta^{(j)} := \frac{R}{\sqrt{1+\eta^2(K-1)}} \left(v^{(0)} + \eta v^{(j)} \right) \quad \text{for all } j \in \{1, \ldots, |S_0|\}, \tag{3}$$

with η to be chosen later.

It is easy to verify that $\|\beta^{(j)}\| \le R$ so that $(\beta^{(j)}, \Sigma_Z^*, A^*) \in \mathcal{S}(R, m)$ for $j \in \{0, 1, \ldots, |S_0|\}$. Moreover, (iii) above implies that, for any $j, \ell \ge 1$ with $j \ne \ell$,

$$\|\beta^{(j)} - \beta^{(\ell)}\|^2 = \frac{R^2\eta^2}{1+\eta^2(K-1)} \|v^{(j)} - v^{(\ell)}\|^2 \ge c_3 \frac{R^2\eta^2(K-1)}{1+\eta^2(K-1)}, \tag{4}$$

and (ii) above guarantees that, for any $j \ge 1$,

$$\begin{aligned} \|\beta^{(j)} - \beta^{(0)}\|^2 &= \frac{R^2(\sqrt{1+\eta^2(K-1)}-1)^2}{1+\eta^2(K-1)} + \frac{R^2\eta^2}{1+\eta^2(K-1)} \|v^{(j)}\|^2 \qquad (5) \\ &\ge \frac{R^2\eta^2}{1+\eta^2(K-1)} \|v^{(j)}\|^2 \\ &\ge c_2 \frac{R^2\eta^2(K-1)}{1+\eta^2(K-1)}. \qquad (6) \end{aligned}$$

On the other hand, for any $j \in \{1, \ldots, |S_0|\}$, Lemma 3 in Section B.3 implies

$$\begin{aligned} &\frac{1}{n} \mathrm{KL}(\mathbb{P}_{\beta^{(j)}}, \mathbb{P}_{\beta^{(0)}}) \\ &\le \frac{|(\beta^{(j)})^\top G^{-1} \beta^{(j)} - (\beta^{(0)})^\top G^{-1} \beta^{(0)}| + \|\Sigma_Z - G^{-1}\|_{\mathrm{op}} \|\beta^{(j)} - \beta^{(0)}\|^2}{\sigma^2 + \min(\|\beta^{(j)}\|^2, \|\beta^{(0)}\|^2)/\|G\|_{\mathrm{op}}} \end{aligned}$$

with $G = \Sigma_Z^{-1} + \tau^{-2} A^\top A$. By using (3), the definition of $\beta^{(0)}$ and (ii), we further have

$$\begin{aligned}
&|(\beta^{(j)})^\top G^{-1}\beta^{(j)} - (\beta^{(0)})^\top G^{-1}\beta^{(0)}| \\
&= \frac{R^2\eta}{1+\eta^2(K-1)} \left| 2(v^{(0)})^\top G^{-1} v^{(j)} + \eta (v^{(j)})^\top G^{-1} v^{(j)} - \eta(K-1)(v^{(0)})^\top G^{-1} v^{(0)} \right| \\
&= \frac{R^2\eta^2}{1+\eta^2(K-1)} \left| (v^{(j)})^\top G^{-1} v^{(j)} - (K-1)(G^{-1})_{11} \right| \\
&\le \frac{R^2\eta^2(K-1)}{1+\eta^2(K-1)} \|G^{-1}\|_{\mathrm{op}}.
\end{aligned}$$

Also note that $\|\beta^{(0)}\|^2 = R^2$ and

$$\|\beta^{(j)}\|^2 = \frac{R^2}{1+\eta^2(K-1)} \left\| v^{(0)} + \eta v^{(j)} \right\|^2 = \frac{R^2(1+\eta^2\|v^{(j)}\|^2)}{1+\eta^2(K-1)} \overset{(ii)}{\ge} cR^2.$$

Together with

$$\|\beta^{(j)} - \beta^{(0)}\|^2 \le \frac{R^2\eta^2(K-1)}{4[1+\eta^2(K-1)]} + \frac{R^2\eta^2}{1+\eta^2(K-1)} \|v^{(j)}\|^2 \overset{(ii)}{\le} \frac{5R^2\eta^2(K-1)}{4[1+\eta^2(K-1)]},$$

from (5) and the fact that $f(x) = \sqrt{x}$ is concave for $x > 0$, we obtain

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{\beta^{(j)}}, \mathbb{P}_{\beta^{(0)}}) &\le \frac{5}{4} \cdot \frac{nR^2\eta^2(K-1)}{1+\eta^2(K-1)} \cdot \frac{\|G^{-1}\|_{\mathrm{op}} + \|\Sigma_Z - G^{-1}\|_{\mathrm{op}}}{\sigma^2 + \min(\|\beta^{(j)}\|^2, \|\beta^{(0)}\|^2)/\|G\|_{\mathrm{op}}} \\
&\le 3n\eta^2(K-1) \cdot \frac{R^2 C_{\max}}{\sigma^2 + cR^2/\|G\|_{\mathrm{op}}}.
\end{aligned}$$

where in the second line we use

$$\|G^{-1}\|_{\mathrm{op}} \le \|\Sigma_Z\|_{\mathrm{op}} \left\| \left[\mathbf{I}_K + (\Sigma_Z)^{1/2} A^\top \Gamma^{-1} A (\Sigma_Z)^{1/2} \right]^{-1} \right\|_{\mathrm{op}} \le \|\Sigma_Z\|_{\mathrm{op}}.$$

Further note that

$$\|G\|_{\mathrm{op}} \le \|\Sigma_Z^{-1}\|_{\mathrm{op}} + \tau^{-2}\|A^\top A\|_{\mathrm{op}} \le C_{\min}^{-1} + m/\tau^2 \le c'm/\tau^2. \tag{7}$$

Choosing

$$\eta^2 = c \cdot \frac{\sigma^2 + c\tau^2 R^2/(c'm)}{nR^2 C_{\max}} \tag{8}$$

yields

$$\mathrm{KL}(\mathbb{P}_{\beta^{(j)}}, \mathbb{P}_{\beta^{(0)}}) \le c \log |S_0|$$

for any $j \ge 1$, and

$$\|\beta^{(j)} - \beta^{(\ell)}\|^2 \ge c \left(\frac{\sigma^2}{C_{\max}} \vee \frac{\tau^2 R^2}{c' C_{\max} m} \right) \frac{(K-1)}{n} \cdot \frac{1}{1+\eta^2(K-1)}$$

for any $j \neq \ell$, from (4) and (6). Since condition $K \leq \bar{c}(R^2 \vee m)n$ guarantees that $\eta^2(K-1) \leq c = c(\bar{c})$, invoking Theorem 2.5 in [7] concludes

$$\inf_{\widehat{\beta}} \sup_{\|\beta\| \leq R} \mathbb{P}_\beta \left\{ \|\widehat{\beta} - \beta\| \geq c \left(1 \vee \frac{R}{\sqrt{m}}\right) \sqrt{\frac{K-1}{n}} \right\} \geq c',$$

which completes the proof. □

B.3. Lemmas used in the proof of Theorem 2

Lemma 2. *Let $k \geq 2$ and $s \geq 1$ be integers, $s \leq k$. There exists a subset S_0 of the set of binary sequences $\{0,1\}^k$ such that*

(i) $\log |S_0| \geq c_1^* s \log(ek/s)$,
(ii) $c_2^* s \leq \|a\|_0 \leq s$, *for all* $a \in S_0$, *and all* $\|a\|_0 = s$ *for* $a \in S_0$, *if* $s \leq k/2$,
(iii) $\|a - b\|^2 \geq c_3^* s$, *for all* $a, b \in S_0$ *and* $a \neq b$,

where $c_j^* > 0$, $j = 1, 2, 3$ are absolute constants.

Proof. This lemma is proved in [4, Lemma 16]. □

Lemma 3. *Suppose that (X_i, Y_i) are i.i.d. Gaussian from model (1). Then, for any $\beta, \beta' \in \mathbb{R}^K$,*

$$\frac{1}{n} KL(\mathbb{P}_\beta, \mathbb{P}_{\beta'}) \leq \frac{|\beta^\top G^{-1}\beta - \beta'^\top G^{-1}\beta'| + \|\Sigma_Z - G^{-1}\|_{\mathrm{op}} \|\beta - \beta'\|^2}{\sigma^2 + \min(\|\beta\|^2, \|\beta'\|^2)/\|G\|_{\mathrm{op}}}$$

where $G = \Sigma_Z^{-1} + A^\top \Gamma^{-1} A$.

Proof. By the additivity of the Kullback-Leibler divergence, it suffices to consider one data pair (X_i, Y_i). We remove the subscript i to lighten the notation. Note that, for given β_j with $j = 1, 2$, model (1) implies

$$\begin{bmatrix} Y \\ X \end{bmatrix} \sim N_{p+1}\left(0, \begin{bmatrix} \beta_j^\top \Sigma_Z \beta_j + \sigma^2 & \beta_j^\top \Sigma_Z A^\top \\ A\Sigma_Z \beta_j & \Sigma \end{bmatrix}\right)$$

with $\Sigma = A\Sigma_Z A^\top + \Gamma$. This further yields

$$\begin{aligned} Y|X &\sim N(\beta_j^\top \Sigma_Z A^\top \Sigma^{-1} X, \sigma^2 + \beta_j^\top (\Sigma_Z - \Sigma_Z A^\top \Sigma^{-1} A \Sigma_Z)\beta_j) \\ &:= N(\mu_j, \sigma_j^2). \end{aligned} \tag{9}$$

Since the marginal distribution of X does not depend on β, we observe that

$$\begin{aligned}
&\frac{1}{n}\mathrm{KL}(\mathbb{P}_{\beta_1}, \mathbb{P}_{\beta_2}) \\
&\quad = \mathbb{E}_{\beta_1}\left[\log f_{\beta_1}(Y|X)\right] - \mathbb{E}_{\beta_1}\left[\log f_{\beta_2}(Y|X)\right] \\
&\quad = -\frac{1}{2}\log\sigma_1^2 - \frac{1}{2\sigma_1^2}\mathbb{E}_{\beta_1}\left[(Y-\mu_1)^2\right] + \frac{1}{2}\log\sigma_2^2 + \frac{1}{2\sigma_2^2}\mathbb{E}_{\beta_1}\left[(Y-\mu_2)^2\right] \\
&\quad = \frac{1}{2}(\log\sigma_2^2 - \log\sigma_1^2) + \frac{1}{2\sigma_2^2}\mathbb{E}_{\beta_1}\left[(Y-\mu_2)^2 - (Y-\mu_1)^2\right] \\
&\qquad + \frac{\sigma_1^2-\sigma_2^2}{2\sigma_1^2\sigma_2^2}\cdot\mathbb{E}_{\beta_1}\left[(Y-\mu_1)^2\right] \\
&\quad = \frac{1}{2}(\log\sigma_2^2 - \log\sigma_1^2) + \frac{1}{2\sigma_2^2}\mathbb{E}_{\beta_1}\left[(Y-\mu_2)^2 - (Y-\mu_1)^2\right] + \frac{\sigma_1^2-\sigma_2^2}{2\sigma_2^2}
\end{aligned}$$

where the last inequality uses $\mathbb{E}_{\beta_1}[(Y-\mu_1)^2] = \sigma_1^2$ from (9). To calculate the expectation, it follows from $\mathbb{E}[XX^\top] = \Sigma$ and (9) that

$$\begin{aligned}
\mathbb{E}_{\beta_1}\left[(Y-\mu_2)^2 - (Y-\mu_1)^2\right] &= \mathbb{E}_{\beta_1}\left[\mu_2^2 - \mu_1^2 + 2Y(\mu_1-\mu_2)\right] \\
&= \beta_2^\top\Sigma_Z A^\top\Sigma^{-1}A\Sigma_Z\beta_2 - \beta_1^\top\Sigma_Z A^\top\Sigma^{-1}A\Sigma_Z\beta_1 \\
&\quad + 2\mathbb{E}_{\beta_1}\left[Y(\beta_1-\beta_2)^\top\Sigma_Z A^\top\Sigma^{-1}X\right] \\
&= \beta_2^\top\Sigma_Z A^\top\Sigma^{-1}A\Sigma_Z\beta_2 - \beta_1^\top\Sigma_Z A^\top\Sigma^{-1}A\Sigma_Z\beta_1 \\
&\quad + 2\mathbb{E}\left[(\beta_1-\beta_2)^\top\Sigma_Z A^\top\Sigma^{-1}AZZ^\top\beta_1\right] \\
&= (\beta_1-\beta_2)^\top\Sigma_Z A^\top\Sigma^{-1}A\Sigma_Z(\beta_1-\beta_2),
\end{aligned}$$

where in the fourth line we use model (1). Plugging this into the KL-divergence yields

$$\begin{aligned}
\frac{1}{n}\mathrm{KL}(\mathbb{P}_{\beta_1}, \mathbb{P}_{\beta_2}) = \frac{1}{2}(\log\sigma_2^2 - \log\sigma_1^2) + \frac{\sigma_1^2-\sigma_2^2}{2\sigma_2^2} \\
+ \frac{1}{2\sigma_2^2}(\beta_1-\beta_2)^\top\Sigma_Z A^\top\Sigma^{-1}A\Sigma_Z(\beta_1-\beta_2).
\end{aligned}$$

Note that $\Sigma_Z - \Sigma_Z A^\top\Sigma^{-1}A\Sigma_Z = G^{-1}$ from Fact 1. Recalling that $\sigma_j^2 = \sigma^2 + \beta_j^\top G^{-1}\beta_j$ from (9) and Fact 1 and using the inequality

$$\left|\log(\sigma^2+t_2) - \log(\sigma^2+t_1)\right| \le \frac{|t_2-t_1|}{\min(\sigma^2+t_1, \sigma^2+t_2)}$$

for $t_1, t_2 > 0$ gives

$$\frac{1}{n}\mathrm{KL}(\mathbb{P}_{\beta_1}, \mathbb{P}_{\beta_2}) \le \frac{|\beta_1^\top G^{-1}\beta_1 - \beta_2^\top G^{-1}\beta_2|}{\min(\sigma_1^2, \sigma_2^2)} + \frac{(\beta_1-\beta_2)^\top(\Sigma_Z - G^{-1})(\beta_1-\beta_2)}{2\sigma_2^2}.$$

Using $\sigma_j^2 \ge \sigma^2 + \|\beta_j\|^2\lambda_{\min}(G^{-1}) = \sigma^2 + \|\beta_j\|^2/\|G\|_{\mathrm{op}}$ completes the proof. □

The following fact is used in the proof of Lemma 3.

Fact 1. *Let* $\Sigma = A\Sigma_Z A^\top + \Gamma$, $G = \Omega + A^\top \Gamma^{-1} A$ *with* $\Omega = \Sigma_Z^{-1}$.

$$\Sigma^{-1} A \Sigma_Z = \Gamma^{-1} A G^{-1}, \quad \Sigma_Z - \Sigma_Z A^\top \Sigma^{-1} A \Sigma_Z = G^{-1}.$$

Proof. The Sherman-Morrison-Woodbury formula gives

$$\begin{aligned}\Sigma^{-1} A \Sigma_Z &= \left[\Gamma^{-1} - \Gamma^{-1} A (\Omega + A^\top \Gamma^{-1} A)^{-1} A^\top \Gamma^{-1}\right] A \Sigma_Z \\ &= \Gamma^{-1} A \Sigma_Z - \Gamma^{-1} A (\Omega + A^\top \Gamma^{-1} A)^{-1} (\Omega + A^\top \Gamma^{-1} A - \Omega) \Sigma_Z \\ &= \Gamma^{-1} A (\Omega + A^\top \Gamma^{-1} A)^{-1},\end{aligned}$$

which concludes the proof of the first statement. The second part follows immediately by noting that

$$\Sigma_Z - \Sigma_Z A^\top \Sigma^{-1} A \Sigma_Z = \Sigma_Z - \Sigma_Z A^\top \Gamma^{-1} A (\Omega + A^\top \Gamma^{-1} A)^{-1} = (\Omega + A^\top \Gamma^{-1} A)^{-1}.$$

□

Appendix C: Preliminaries

C.1. General principles

Throughout the proofs of the main results, we will work on the event

$$\mathcal{E} := \left\{ \max_{1 \le k \le K} \frac{1}{n} \sum_{t=1}^{n} \mathbf{Z}_{ik}^2 \le B_z \right\} \bigcap \left\{ \max_{1 \le j < \ell \le p} |\widehat{\Sigma}_{j\ell} - \Sigma_{j\ell}| \le \delta \right\} \tag{10}$$

with B_z defined in Assumption 2 and $\delta := c\delta_n$ some constant $c > 0$ and

$$\delta_n = \sqrt{\log(p \vee n)/n}.$$

Provided $\log p \lesssim n$, Lemma 4 and Lemma 6 below guarantee that $\mathbb{P}(\mathcal{E}) \ge 1 - (p \vee n)^{-\alpha}$ for some constant $\alpha > 0$. This event plays an important role. The proof of Theorem 3 in [2] reveals that on this event $\mathcal{E}$, we have

- $\widehat{K} = K$,
- $I_k \subseteq \widehat{I}_{\pi(k)} \subseteq I_k \cup J_1^k$ with $J_1^k = \{j \in J : \ |A_{jk}| \ge 1 - 4\delta/v\}$, for all $k \in [K]$,

for some permutation $\pi : [K] \to [K]$. *To lighten the notation, we assume throughout the remainder of the appendix that* π *is the identity group permutation and* $A_{ik} = 1$ *for any* $i \in I_k$ *and* $k \in [K]$. In the general case, the signed permutation matrix P can be traced throughout the proofs.

This means in particular that the dimensions of the parameter β and its estimate $\widehat{\beta}$ are the same. Unfortunately, $\widehat{I}$ is only guaranteed to satisfy $\widehat{I} \supseteq I$. This is the main source of many technical challenges in the proofs. We emphasize that signal conditions on A could prevent this, and lead to guarantees of $\widehat{I} = I$. However, an assumption that the entries of A are either zero or exceed a certain threshold in absolute value, is rather unnatural and instead we only rely on a mild separation condition on the matrix Σ_Z in Assumptions 1 and 4.

The next subsections contain notation and some auxiliary results, that may be skipped during the first reading of this manuscript.

C.2. Notation

We write

$$\widehat{\Pi} = \widehat{A}_{\widehat{I}\bullet}(\widehat{A}_{\widehat{I}\bullet}^\top \widehat{A}_{\widehat{I}\bullet})^{-1} \tag{11}$$

and

$$\widetilde{\mathbf{X}} := \mathbf{X}_{\bullet\widehat{I}}\widehat{\Pi}, \qquad \widetilde{\mathbf{Z}} := \mathbf{Z}A_{\widehat{I}\bullet}^\top\widehat{\Pi}, \qquad \widetilde{\mathbf{W}} := \mathbf{W}_{\bullet\widehat{I}}\widehat{\Pi}, \tag{12}$$

so that $\widetilde{\mathbf{X}} = \widetilde{\mathbf{Z}} + \widetilde{\mathbf{W}}$. Similarly, we write

$$\Pi = A_{I\bullet}(A_{I\bullet}^\top A_{I\bullet})^{-1} \tag{13}$$

and

$$\overline{\mathbf{X}} := \mathbf{X}_{\bullet I}\Pi, \qquad \overline{\mathbf{W}} := \mathbf{W}_{\bullet I}\Pi, \tag{14}$$

so that $\overline{\mathbf{X}} = \mathbf{Z} + \overline{\mathbf{W}}$. For all $k \in [K]$, set

$$m_k := |I_k| \quad \text{and} \quad m = \min_{k\in[K]} m_k.$$

On the event $\mathcal{E}$, we further define $L_k := \widehat{I}_k \setminus I_k \subseteq J_1^k$ so that

$$\widehat{m}_k := |\widehat{I}_k| = m_k + |L_k|,$$

and

$$D_\rho = \mathrm{diag}(\rho_1, \dots, \rho_K), \quad \text{with} \quad \rho_k = \frac{|L_k|}{\widehat{m}_k}. \tag{15}$$

On the event $\mathcal{E}$, we have $|L_k| \le |J_1^k|$, and we will frequently make use of the inequality

$$\|\rho\|_2^2 = \sum_{k=1}^K \left(\frac{|L_k|}{m_k + |L_k|}\right)^2 \le \sum_{k=1}^K \left(\frac{|J_1^k|}{m_k + |J_1^k|}\right)^2 := \bar{\rho}^2. \tag{16}$$

This inequality uses the fact that the function $x/(m_k + x)$ is increasing for $x \ge 0$. Finally, the inequality

$$\max_{k\in[K],\ j\in J_1^k} \|A_{j\bullet} - \mathbf{e}_k\|_1 \le 8\delta/\nu \lesssim \delta_n \tag{17}$$

is a direct consequence of the definition of J_1^k, and will be repeatedly used in our proofs. Here ν is the constant defined in Assumption 1. Recall $\Theta = A\Sigma_Z$ and define the matrices

$$\Omega := \Sigma_Z^{-1} \quad \text{and} \quad H := A(\Sigma_Z)^{1/2} = \Theta(\Sigma_Z)^{-1/2} = \Theta\Omega^{1/2}$$

and $H^+ = (H^\top H)^{-1}H^\top$. Further, on the event $\mathcal{E}$ to ensure that $\widehat{K} = K$, write

$$\widehat{H} := \widehat{\Theta}\Omega^{1/2} \tag{18}$$

with $\widehat{\Theta}$ defined in (12), and

$$\Delta = A_{\widehat{I}\bullet}^\top \widehat{A}_{\widehat{I}\bullet}(\widehat{A}_{\widehat{I}\bullet}^\top \widehat{A}_{\widehat{I}\bullet})^{-1} - \mathbf{I}_K = A_{\widehat{I}\bullet}^\top \widehat{\Pi} - \mathbf{I}_K. \tag{19}$$

C.3. Preliminary lemmas

We now state three preliminary lemmas that will be used throughout the proofs that follow. Their proofs are relegated to Section C.4.1.

Lemma 4. *Suppose Assumption 2 holds.*

(1) For any fixed $v \in \mathbb{R}^K$ and $t \in [n]$, $\langle \mathbf{Z}_{t\bullet}, v\rangle$ is $\gamma_z\sqrt{v^\top\Sigma_Z v}$–sub-Gaussian. In particular, $\mathbf{Z}_{tk}$ is γ_z'-sub-Gaussian with $\gamma_z' := \gamma_z\sqrt{B_z}$ for any $k \in [K]$.
(2) Provided Assumption 1 holds as well, $\mathbf{X}_{tj}$ is $(\gamma_z' + \gamma_w)$-sub-Gaussian for any $t \in [n]$ and $j \in [p]$.

Lemma 5. *Suppose Assumption 2 holds. Let $v \in \mathbb{R}^K$ and $\alpha \in \mathbb{R}^p$ be any fixed vectors. Then, for any $t \in [n]$,*

(1) $\alpha^\top \mathbf{W}_{t\bullet}$ is $\gamma_w\|\alpha\|_2$-sub-Gaussian;
(2) $v^\top H^+ \mathbf{W}_{t\bullet}$ is $\gamma_w\sqrt{v^\top(H^\top H)^{-1}v}$-sub-Gaussian;
(3) $\overline{\mathbf{W}}_{tk}$ is $\gamma_w/\sqrt{m}$-sub-Gaussian for any $k \in [K]$.

Lemma 6. *Let $\{X_t\}_{t=1}^n$ and $\{Y_t\}_{t=1}^n$ be any two sequences, each with zero mean independent γ_x-sub-Gaussian and γ_y-sub-Gaussian elements. Then, for some constants $c, c' > 0$, we have*

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{t=1}^n (X_tY_t - \mathbb{E}[X_tY_t])\right| \le c\gamma_x\gamma_y t\right\} \ge 1 - 2\exp\left\{-c'\min\left(t^2, t\right)n\right\}.$$

In particular, when $\log p \le c''n$ for some constant $c'' > 0$, one has

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{t=1}^n (X_tY_t - \mathbb{E}[X_tY_t])\right| \le c\gamma_x\gamma_y\sqrt{\frac{\log(p\vee n)}{n}}\right\} \ge 1 - 2(p\vee n)^{-c'}.$$

Finally we collect several results that control different random quantities of interest.

Lemma 7. *Let $u, v \in \mathbb{R}^K$ and $\alpha \in \mathbb{R}^p$ be fixed vectors. Under Assumption 2, each of the following statements holds with probability $1 - (p \vee n)^{-c}$ for some constant $c > 0$,*

$$\frac{1}{n}|u^\top \mathbf{Z}^\top \varepsilon| \lesssim \delta_n \sqrt{u^\top \Sigma_Z u}, \tag{20}$$

$$\frac{1}{n}|\alpha^\top \mathbf{W}^\top \varepsilon| \lesssim \delta_n \|\alpha\|_2, \tag{21}$$

$$\frac{1}{n}|\alpha^\top \mathbf{W}^\top \mathbf{Z} u| \lesssim \delta_n \|\alpha\|_2 \sqrt{u^\top \Sigma_Z u}, \tag{22}$$

$$\frac{1}{n}|u^\top H^+ \mathbf{W}^\top \varepsilon| \lesssim \delta_n \sqrt{u^\top (H^\top H)^{-1} u}, \tag{23}$$

$$\frac{1}{n}|u^\top H^+ \mathbf{W}^\top \mathbf{Z} v| \lesssim \delta_n \sqrt{v^\top \Sigma_Z v} \sqrt{u^\top (H^\top H)^{-1} u}, \tag{24}$$

$$\left| u^\top \left(\frac{1}{n} \mathbf{Z}^\top \mathbf{Z} - \Sigma_Z \right) v \right| \lesssim \delta_n \sqrt{u^\top \Sigma_Z u} \sqrt{v^\top \Sigma_Z v}, \tag{25}$$

$$\frac{1}{n}|u^\top \mathbf{Z}^\top \mathbf{Z} v| \lesssim |u^\top \Sigma_Z v| + \delta_n \sqrt{u^\top \Sigma_Z u} \sqrt{v^\top \Sigma_Z v}. \tag{26}$$

If, in addition, Assumption 1 holds, we have, on an event that holds with probability $1 - C(p \vee n)^{-\alpha}$ for some constants $C, \alpha > 0$, $\widehat{K} = K$ and each of the following inequalities holds,

$$\frac{1}{n}|u^\top \mathbf{Z}^\top \widetilde{\mathbf{W}} v| \lesssim \delta_n \|v\|_2 \sqrt{u^\top \Sigma_Z u} \left(\frac{1}{\sqrt{m}} + \bar{\rho} \right) \tag{27}$$

$$\left| u^\top H^+ \left(\frac{1}{n} \mathbf{W}^\top \widetilde{\mathbf{W}} - \widehat{\Gamma}_{\cdot \widehat{I}} \widehat{\Pi} \right) v \right| \lesssim \delta_n \frac{\|u\|_2 \|v\|_2}{\sigma_K(H)} \left(\frac{1}{\sqrt{m}} + \bar{\rho} \right) \tag{28}$$

$$\left| \alpha^\top \left(\frac{1}{n} \mathbf{W}^\top \widetilde{\mathbf{W}} - \widehat{\Gamma}_{\cdot \widehat{I}} \widehat{\Pi} \right) v \right| \lesssim \delta_n \|\alpha\|_2 \|v\|_2 \left(\frac{1}{\sqrt{m}} + \bar{\rho} \right). \tag{29}$$

Proof. This lemma is proved in Section C.4.2. We emphasize that the expressions on the left hand side of (27), (28) and (29) are well defined on the event $\widehat{K} = K$. □

Lemma 8. *Under Assumption 2, on an event that holds with probability $1 - C(p \vee n)^{-K\alpha}$*

for some constants $C, \alpha > 0$, $\widehat{K} = K$ as well as

$$\frac{1}{n}\|H^+\mathbf{W}^\top\mathbf{Z}\Omega^{1/2}\|_{\text{op}} \lesssim \frac{\delta_n}{\sigma_K(H)}\sqrt{K}, \tag{30}$$

$$\left\|\Omega^{1/2}\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{Z} - \Sigma_Z\right)\Omega^{1/2}\right\|_{\text{op}} \lesssim \delta_n\sqrt{K}, \tag{31}$$

$$\frac{1}{n}\left\|\Omega^{1/2}\mathbf{Z}^\top\widetilde{\mathbf{W}}\right\|_{\text{op}} \lesssim \left(\frac{1}{\sqrt{m}} + \bar{\rho}\right)\delta_n\sqrt{K}, \tag{32}$$

$$\left\|H^+\left(\frac{1}{n}\mathbf{W}^\top\widetilde{\mathbf{W}} - \widehat{\Gamma}_{\cdot\widehat{I}}\widehat{\Pi}\right)\right\|_{\text{op}} \lesssim \frac{1}{\sigma_K(H)}\left(\frac{1}{\sqrt{m}} + \bar{\rho}\right)\delta_n\sqrt{K}. \tag{33}$$

Proof. This lemma is proved in Section C.4.3. □

Lemma 9. *Under Assumptions 1 and 2, we have, for any fixed $v \in \mathbb{R}^K$, on an event that holds with probability at least $1 - C(p \vee n)^{-\alpha}$ for some constants $C, \alpha > 0$, $\widehat{K} = K$ and*

$$\|\Delta v\|_1 \lesssim \|v\|_2\bar{\rho}\delta_n.$$

Furthermore, for any $K \times K$-matrix Q, on an event that holds with probability $1 - C(p \vee n)^{-\alpha}$ for some constants $C, \alpha > 0$, $\widehat{K} = K$ and

$$\|Q\Delta\|_{\text{op}} \le \|Q\|_{2,\infty} \cdot \bar{\rho} \cdot \delta_n.$$

Proof. This lemma is proved in Section C.4.4. □

Lemma 10. *Suppose Assumptions 1 and 2 hold, and that there exists some sufficiently small constant $c_0 > 0$ such that*

$$\delta_n\sqrt{K} \cdot T_n \le c_0 \tag{34}$$

with

$$T_n := 1 + \left(\frac{1}{\sqrt{m}} + \bar{\rho} + \bar{\rho}\delta_n\right) \Big/ \sqrt{C_{\min}} \tag{35}$$

Then, for some constants $0 < c_1 < 1$ and $c_2 > 0$, on an event that holds with probability at least $1 - C(p \vee n)^{-\alpha}$ for some constants $C, \alpha > 0$, $\widehat{K} = K$ and the following inequalities hold

(a) $\|H^+(\widehat{H} - H)\|_{\text{op}} \le c'\delta_n T_n\sqrt{K} \le c_1$;
(b) $\lambda_K(\widehat{H}^\top\widehat{H}) \ge (1 - c_1)^2\lambda_K\left(H^\top H\right)$;
(c) $\lambda_K(\widehat{\Theta}^\top\widehat{\Theta}) \ge (1 - c_1)^2 C_{\min}\lambda_K(H^\top H)$;
(d) $|u^\top(\widehat{H} - H)v| \lesssim \|u\|_2\|v\|_2 T_n\delta_n$ *for any fixed* $u \in \mathbb{R}^p$ *and* $v \in \mathbb{R}^K$;
(e) $\|\widehat{H} - H\|_{\text{op}} \le c_2 T_n\delta_n\sqrt{pK}$.

Proof. This lemma is proved in Section C.4.5. □

C.4. Proofs of the preliminary lemmas

C.4.1. Proof of lemmas 4, 5 and 6

Proof of Lemma 4. For any $t \in [n]$, Assumption 2 implies that

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda \left\langle \mathbf{Z}_{t\bullet}, v\right\rangle\right)\right] &= \mathbb{E}\left[\exp\left(\lambda \langle \Sigma_Z^{-1/2}\mathbf{Z}_{t\bullet}, \Sigma_Z^{1/2} v\rangle\right)\right] \\ &\leq \mathbb{E}\left[\exp\left(\lambda^2 \gamma_z^2 v^\top \Sigma_Z v/2\right)\right], \qquad \forall\ \lambda \in \mathbb{R}.\end{aligned}$$

This proves the statement of $\langle \mathbf{Z}_{t\bullet}, v\rangle$. Taking $v = \mathbf{e}_k$ and using $\|\Sigma_Z\|_\infty \leq B_z$ concludes the proof of (1). Part (1) and the independence between Z and W yield

$$\begin{aligned}\mathbb{E}[\exp(\lambda \mathbf{X}_{tj})] &= \mathbb{E}\left[\exp(\lambda \langle A_{j\bullet}, \mathbf{Z}_{t\bullet}\rangle)\right] \mathbb{E}\left[\exp(\lambda \mathbf{W}_{tj})\right] \\ &\leq \exp\left(\lambda^2 \gamma_z^2 A_{j\bullet}^\top \Sigma_Z A_{j\bullet}/2\right) \exp\left(\lambda^2 \gamma_w^2/2\right) \\ &\leq \exp\left\{\lambda^2(\gamma_z^2 \|\Sigma_Z\|_\infty + \gamma_w^2)/2\right\} \qquad (\|A_{j\bullet}\|_1 \leq 1)\end{aligned}$$

for any $\lambda \in \mathbb{R}$ and any $1 \leq j \leq p$. This completes the proof. □

Proof of Lemma 5. By Assumption 2, $\mathbf{W}_{tj}$ is γ_w-sub-Gaussian for all $t \in [n]$ and $j \in [p]$, so

$$\mathbb{E}[\exp(\lambda \mathbf{W}_{ti})] \leq \exp(\lambda^2 \gamma_w^2/2), \quad \forall \lambda \in \mathbb{R}.$$

Again by Assumption 2, the $\mathbf{W}_{ti}$ are independent across index i, whence

$$\begin{aligned}\mathbb{E}[\exp(t\alpha^\top W_{t\bullet})] &= \prod_{j=1}^{p} \mathbb{E}\left[\exp(t\alpha_j \mathbf{W}_{tj})\right] \\ &\leq \prod_{j=1}^{p} \exp\left[t^2 \gamma_w^2 \alpha_j^2/2\right] \\ &\leq \exp\left[t^2 \gamma_w^2 \|\alpha\|_2^2/2\right],\end{aligned}$$

proving that $\sum_j \alpha_j \mathbf{W}_{tj}$ is $\gamma_w \|\alpha\|_2$-sub-Gaussian. The second statement follows by taking $\alpha = (H^+)^\top v$. Similarly,

$$\begin{aligned}\mathbb{E}[\exp(t\overline{\mathbf{W}}_{tk})] &= \prod_{i \in I_k} \mathbb{E}\left[\exp(t\mathbf{W}_{ti}/m_k)\right] \\ &\leq \prod_{i \in I_k} \exp\left[t^2 \gamma_w^2/(2m_k^2)\right] \\ &\leq \exp\left[t^2 \gamma_w^2/(2m)\right],\end{aligned}$$

proving that $\overline{\mathbf{W}}_{tk}$ is $\gamma_w/\sqrt{m}$-sub-Gaussian. This concludes the proof. □

Proof of Lemma 6. Let $\|\cdot\|_{\psi_1}$ and $\|\cdot\|_{\psi_2}$ denote, respectively, the sub-exponential norm and the sub-gaussian norm (Definitions 5.7 and 5.13 in [8]) as

$$\|X\|_{\psi_1} = \sup_{p\geq 1} p^{-1} \left(\mathbb{E}[|X|^p]\right)^{1/p}, \quad \|X\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2} \left(\mathbb{E}[|X|^p]\right)^{1/p}$$

for any random variable X. Then, $\|X_t\|_{\psi_2} \leq c\gamma_x$ and $\|Y_t\|_{\psi_2} \leq c\gamma_y$ and an application of the Hölder's inequality yields $\|X_tY_t\|_{\psi_1} \leq \|X_t\|_{\psi_2}\|Y_t\|_{\psi_2} \leq c\gamma_x\gamma_y$. The proof of the first statement follows by Corollary 5.17 in [8]. The second statement follows by taking $t = \max(\delta_n, \delta_n^2)$ and observing that $\min(t^2, t)n = \log(p \vee n)$. □

C.4.2. Proof of Lemma 7

The proofs of (20) – (26) are all based on applying Lemma 6 in conjunction with Assumption 2, Lemmas 4 and 5 and the fact that $\mathbb{E}[\mathbf{Z}_{t\bullet}\mathbf{Z}_{t\bullet}^\top] = \Sigma_Z$ for all $t \in [n]$.

We work on the event $\mathcal{E}$ defined in (10) that has probability at least $1 - (p \vee n)^{-c}$ for some $c > 0$. Recall that it implies $\widehat{K} = K$ and $I_k \subseteq \widehat{I}_k \subseteq I_k \cup J_1^k$ for all $k \in [K]$.

To prove (27), since

$$\widetilde{\mathbf{W}}_{\bullet k} = \overline{\mathbf{W}}_{\bullet k} + \frac{|L_k|}{\widehat{m}_k}\left(\frac{1}{|L_k|}\sum_{i\in L_k}\mathbf{W}_{\bullet i} - \overline{\mathbf{W}}_{\bullet k}\right)$$

from (12) and (14), we have

$$\widetilde{\mathbf{W}} - \overline{\mathbf{W}} = \mathbf{W}_{\bullet L}\widehat{A}_{L\bullet}D_{\widehat{m}} - \overline{\mathbf{W}}D_\rho \tag{36}$$

where $D_{\widehat{m}} = \mathrm{diag}(1/\widehat{m}_1, \ldots, 1/\widehat{m}_K)$ and D_ρ is defined in (15). It then follows that

$$\begin{aligned}
&\frac{1}{n}|u^\top\mathbf{Z}^\top\widetilde{\mathbf{W}}v| \\
&\leq \frac{1}{n}|u^\top\mathbf{Z}^\top\overline{\mathbf{W}}v| + \frac{1}{n}|u^\top\mathbf{Z}^\top\mathbf{W}_{\bullet L}\widehat{A}_{L\bullet}D_{\widehat{m}}v| + \frac{1}{n}|u^\top\mathbf{Z}^\top\overline{\mathbf{W}}D_\rho v| \\
&\leq \frac{1}{n}|u^\top\mathbf{Z}^\top\overline{\mathbf{W}}v| + \max_{i\in J_1}\frac{1}{n}|u^\top\mathbf{Z}^\top\mathbf{W}_{\bullet i}|\cdot\|\widehat{A}_{L\bullet}D_{\widehat{m}}v\|_1 + \max_k\frac{1}{n}|u^\top\mathbf{Z}^\top\overline{\mathbf{W}}_{\bullet k}|\cdot\|D_\rho v\|_1.
\end{aligned} \tag{37}$$

We find that

$$\|\widehat{A}_{L\bullet}D_{\widehat{m}}v\|_1 \leq \bar{\rho}\|v\|_2 \leq \|v\|_2\|\rho\|_2 \overset{(16)}{\leq} \|v\|_2\bar{\rho}, \tag{38}$$

after we apply Lemmas 4, 5 and 6 and take a union bound. This concludes the proof of (27).

To prove the results of (28) and (29), we first have

$$\frac{1}{n}\mathbf{W}^\top\widetilde{\mathbf{W}} - \widehat{\Gamma}_{\bullet\widehat{I}}\widehat{\Pi} = \frac{1}{n}\mathbf{W}^\top(\widetilde{\mathbf{W}} - \overline{\mathbf{W}}) + \frac{1}{n}\mathbf{W}^\top\overline{\mathbf{W}} - \Gamma_{\bullet I}\Pi + \Gamma_{\bullet I}\Pi - \widehat{\Gamma}_{\bullet\widehat{I}}\widehat{\Pi}. \tag{39}$$

Then by plugging (36) into (39), we obtain

$$\frac{1}{n}\mathbf{W}^\top\widetilde{\mathbf{W}} - \widehat{\Gamma}_{\cdot\widehat{I}}\widehat{\Pi} = \left(\frac{1}{n}\mathbf{W}^\top\mathbf{W}_{\cdot L} - \Gamma_{\cdot L}\right)\widehat{A}_{L\cdot}D_{\widehat{m}} - \left(\frac{1}{n}\mathbf{W}^\top\overline{\mathbf{W}} - \Gamma_{\cdot I}\Pi\right)D_\rho + \frac{1}{n}\mathbf{W}^\top\overline{\mathbf{W}} - \Gamma_{\cdot I}\Pi + \Delta_d \tag{40}$$

where

$$\Delta_d = \Gamma_{\cdot I}\Pi - \widehat{\Gamma}_{\cdot\widehat{I}}\widehat{\Pi} + \Gamma_{\cdot L}\widehat{A}_{L\cdot}D_{\widehat{m}} - \Gamma_{\cdot I}\Pi D_\rho. \tag{41}$$

We first prove (29) and fix any $\alpha \in \mathbb{R}^p$ and $v \in \mathbb{R}^K$. From display (40), we only need to upper bound

$$\left\|\alpha^\top\left(\frac{1}{n}\mathbf{W}^\top\mathbf{W}_{\cdot L} - \Gamma_{\cdot L}\right)\right\|_\infty \|\widehat{A}_{L\cdot}D_{\widehat{m}}v\|_1 + \left\|\alpha^\top\left(\frac{1}{n}\mathbf{W}^\top\overline{\mathbf{W}} - \Gamma_{\cdot I}\Pi\right)\right\|_\infty \|D_\rho v\|_1 + \left|\alpha^\top\left(\frac{1}{n}\mathbf{W}^\top\overline{\mathbf{W}} - \Gamma_{\cdot I}\Pi\right)v\right| + |\alpha^\top\Delta_d v|.$$

Invoke Lemmas 5 and 6 and apply a union bound to derive

$$\max_{i\in J_1}\left|\alpha^\top\left(\frac{1}{n}\mathbf{W}^\top\mathbf{W}_{\cdot i} - \Gamma_{\cdot i}\right)\right| \lesssim \|\alpha\|_2\delta_n,$$
$$\left|\alpha^\top\left(\frac{1}{n}\mathbf{W}^\top\overline{\mathbf{W}} - \Gamma_{\cdot I}\Pi\right)v\right| \lesssim \|\alpha\|_2\|v\|_2\delta_n/\sqrt{m}$$

Each inequality holds with probability $1-(p\vee n)^{-c}$. By further using (38), we conclude that, with probability $1-c'(p\vee n)^{-c}$,

$$\left\|\alpha^\top\left(\frac{1}{n}\mathbf{W}^\top\mathbf{W}_{\cdot L} - \Gamma_{\cdot L}\right)\right\|_\infty \|\widehat{A}_{L\cdot}D_{\widehat{m}}v\|_1 \lesssim \bar{\rho}\|v\|_2\|\alpha\|_2\delta_n, \tag{42}$$
$$\left\|\alpha^\top\left(\frac{1}{n}\mathbf{W}^\top\overline{\mathbf{W}} - \Gamma_{\cdot I}\Pi\right)\right\|_\infty \|D_\rho v\|_1 \lesssim \bar{\rho}\|v\|_2\|\alpha\|_2\delta_n/\sqrt{m}, \tag{43}$$
$$\left|\alpha^\top\left(\frac{1}{n}\mathbf{W}^\top\overline{\mathbf{W}} - \Gamma_{\cdot I}\Pi\right)v\right| \lesssim \|\alpha\|_2\|v\|_2\delta_n/\sqrt{m}. \tag{44}$$

We upper bound the remaining term $|\alpha^\top\Delta_d v|$. First note that

$$[\Delta_d]_{ik} = \frac{\Gamma_{ii} - \widehat{\Gamma}_{ii}}{\widehat{m}_k}, \quad \forall i \in \widehat{I}_k, \qquad [\Delta_d]_{ik} = 0, \text{ otherwise .}$$

Furthermore, since $\widehat{I} \subseteq I \cup J_1$ on the event $\mathcal{E}$, we have

$$\|\Delta_d v\|_2^2 = \sum_{k=1}^K \sum_{i\in\widehat{I}_k} \frac{v_k^2}{\widehat{m}_k^2}\left(\widehat{\Gamma}_{ii} - \Gamma_{ii}\right)^2 \le \frac{\|v\|_2^2}{m}\max_{i\in I\cup J_1}\left(\widehat{\Gamma}_{ii} - \Gamma_{ii}\right)^2.$$

We thus obtain

$$|\alpha^\top \Delta_d v| \le \|\alpha\|_2 \|\Delta_d v\|_2 \lesssim \|\alpha\|_2 \|v\|_2 \delta_n/\sqrt{m} \tag{45}$$

on the event $\mathcal{E}$ intersected with

$$\mathcal{E}_W := \left\{ \max_{i\in I\cup J_1} \left|\widehat{\Gamma}_{ii} - \Gamma_{ii}\right| \lesssim \delta_n \right\}. \tag{46}$$

Since $\mathcal{E}_W$ holds with probability $1-(p\vee n)^{-c}$ by Lemma 21, in conjunction with (42) – (44), we conclude (29). The result of (28) follows immediately by taking $\alpha = u^\top H^+$ and noting that $\|\alpha\|_2 \le \|u\|_2/\sigma_K(H)$. □

C.4.3. Proof of Lemma 8

All results (30) – (33) can be proved based on Lemma 7 together with a classical discretization method to prove the uniformity over the unit sphere. We only prove display (32) since the other results can be shown by similar arguments.

Again, we work on the event $\mathcal{E}$ defined in (10) that has probability at least $1-(p\vee n)^{-c}$ for some $c>0$. Recall that it implies $\widehat{K}=K$ and $I_k \subseteq \widehat{I}_k \subseteq I_k \cup J_1^k$ for all $k\in[K]$.

Let $\mathcal{N}_\varepsilon \subset \mathcal{S}^{K-1}$ be a minimal ε-net of $\mathcal{S}^{K-1}$, i.e. a set with minimum cardinality such that the collection of ε-balls centered at points in $\mathcal{N}_\varepsilon$ covers $\mathcal{S}^{K-1}$.

To prove (32), a standard discretization argument, for instance, see [5, Proof of Proposition 2.4] gives, on the event $\mathcal{E}$,

$$\frac{1}{n}\|\Omega^{1/2}\mathbf{Z}^\top \widetilde{\mathbf{W}}\|_{\mathrm{op}} = \sup_{u,v\in\mathcal{S}^{K-1}} \frac{1}{n}|u^\top \Omega^{1/2}\mathbf{Z}^\top \widetilde{\mathbf{W}} v| \le 4 \max_{u,v\in\mathcal{N}_{1/2}} \frac{1}{n}|u^\top \Omega^{1/2}\mathbf{Z}^\top \widetilde{\mathbf{W}} v|.$$

Recall that (37) and (38) imply that, for any $u,v\in\mathcal{S}^{K-1}$,

$$\begin{aligned}\frac{1}{n}|u^\top \Omega^{1/2}\mathbf{Z}^\top \widetilde{\mathbf{W}} v| \le{}& \frac{1}{n}|u^\top \Omega^{1/2}\mathbf{Z}^\top \overline{\mathbf{W}} v| + \max_{i\in J_1}\frac{1}{n}|u^\top \Omega^{1/2}\mathbf{Z}^\top \mathbf{W}_{\bullet i}|\cdot\bar\rho\|v\|_2 \\ &+ \max_{1\le k\le K}\frac{1}{n}|u^\top \Omega^{1/2}\mathbf{Z}^\top \overline{\mathbf{W}}_{\bullet k}|\cdot\bar\rho\|v\|_2,\end{aligned}$$

on the event $\mathcal{E}$. Use the classical bound $|\mathcal{N}_{1/2}|\le 5^K$ by Lemma 5.2 in [8], apply Lemma 5 and Lemma 6 with $t=\delta_n\sqrt{K}$ for each random term in the above display and take the union bound over $u,v\in\mathcal{N}_{1/2}$, $i\in J_1$ and $k\in[K]$, and use the restriction $K\log(p\vee n)\le cn$ to obtain

$$\max_{u,v\in\mathcal{N}_{1/2}} \frac{1}{n}|u^\top \Omega^{1/2}\mathbf{Z}^\top \widetilde{\mathbf{W}} v| \lesssim \delta_n\sqrt{K}\left(\frac{1}{\sqrt{m}}+\bar\rho\right)$$

with probability $1-4\cdot 5^{2K}(p\vee n)^{-cK}$. This finishes the proof of (32). □

C.4.4. Proof of Lemma 9

We work on the event $\mathcal{E}$. Recall that it $\widehat{K} = K$, but also $\widehat{A}_{ik} = A_{ik}$ for all $i \in I$ and $k \in [K]$ by Theorem 4.2. Recall the definition $\Delta = A_{\widehat{I}\bullet}^\top \widehat{A}_{\widehat{I}\bullet}(\widehat{A}_{\widehat{I}\bullet}^\top \widehat{A}_{\widehat{I}\bullet})^{-1} - \mathbf{I}_K$. For any fixed vector $v \in \mathbb{R}^K$, we have

$$\begin{aligned}
\|\Delta v\|_1 &= \sum_{k=1}^K \left| \sum_{a=1}^K \frac{v_a}{\widehat{m}_a} \sum_{i \in \widehat{I}_a} (\widehat{A}_{ik} - A_{ia}) \right| \\
&\le \sum_{k=1}^K \sum_{a=1}^K \frac{|v_a|}{\widehat{m}_a} \sum_{i \in L_a} \left| \widehat{A}_{ik} - A_{ia} \right| \\
&= \sum_{a=1}^K \frac{|v_a|}{\widehat{m}_a} \sum_{i \in L_a} \left\| \widehat{A}_{i\bullet} - A_{i\bullet} \right\|_1 \\
&\le \max_{i \in J_1} \left\| \widehat{A}_{i\bullet} - A_{i\bullet} \right\|_1 \sum_{a=1}^K \rho_a |v_a| \\
&\lesssim \delta_n \bar{\rho} \|v\|_2.
\end{aligned}$$

We used $\widehat{I}_a \subseteq I_a \cup L_a$ on the event $\mathcal{E}$ in the second line and (15), (16) and (17) in the last line. This proves the first result.

To prove the second claim, we argue that for any matrix $Q \in \mathbb{R}^{d \times K}$, by using the dual norm inequality and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\|Q\Delta\|_{\mathrm{op}} &= \sup_{\alpha \in \mathcal{S}^{d-1}, u \in \mathcal{S}^{K-1}} \alpha^\top Q \Delta u \\
&\le \sup_{\alpha \in \mathcal{S}^{d-1}, u \in \mathcal{S}^{K-1}} \|\alpha^\top Q\|_\infty \|\Delta u\|_1 \\
&\lesssim \sup_{u \in \mathcal{S}^{K-1}} \|Q\|_{2,\infty} \cdot \delta_n \|D_\rho u\|_1.
\end{aligned}$$

The result then follows from $\|D_\rho u\|_1 \le \|\rho\|_2 \le \bar{\rho}$ by using (16). □

C.4.5. Proof of Lemma 10

We work on the event $\mathcal{E}$ defined in (10) that has probability at least $1 - (p \vee n)^{-c}$ for some $c > 0$. Recall that it implies $\widehat{K} = K$ and $I_k \subseteq \widehat{I}_k \subseteq I_k \cup J_1^k$ for all $k \in [K]$.

We first prove (a). The definition of $\widehat{\Theta}$ in (12) gives

$$\begin{aligned}\widehat{\Theta}-\Theta &= \frac{1}{n}\mathbf{X}^\top\widetilde{\mathbf{X}} - \widehat{\Gamma}_{\cdot\widehat{I}}\widehat{\Pi} - A\Sigma_Z \\ &= \frac{1}{n}\mathbf{X}^\top\mathbf{Z} + \frac{1}{n}\mathbf{X}^\top(\widetilde{\mathbf{X}}-\mathbf{Z}) - \widehat{\Gamma}_{\cdot\widehat{I}}\widehat{\Pi} - A\Sigma_Z \\ &= A\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{Z}-\Sigma_Z\right) + A\frac{1}{n}\mathbf{Z}^\top\widetilde{\mathbf{W}} + A\frac{1}{n}\mathbf{Z}^\top\mathbf{Z}\Delta \\ &\quad + \frac{1}{n}\mathbf{W}^\top\widetilde{\mathbf{Z}} + \frac{1}{n}\mathbf{W}^\top\widetilde{\mathbf{W}} - \widehat{\Gamma}_{\cdot\widehat{I}}\widehat{\Pi}\end{aligned} \tag{47}$$

where we use $\mathbf{X} = \mathbf{Z}A^\top + \mathbf{W}$ and the definition (19) in the third equality. By writing

$$\Delta_1 := \Omega^{1/2}\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{Z}-\Sigma_Z + \frac{1}{n}\mathbf{Z}^\top\widetilde{\mathbf{W}} + \frac{1}{n}\mathbf{Z}^\top\mathbf{Z}\Delta\right)\Omega^{1/2}, \tag{48}$$

$$\Delta_2 := \left(\frac{1}{n}\mathbf{W}^\top\widetilde{\mathbf{Z}} + \frac{1}{n}\mathbf{W}^\top\widetilde{\mathbf{W}} - \widehat{\Gamma}_{\cdot\widehat{I}}\widehat{\Pi}\right)\Omega^{1/2}, \tag{49}$$

we have $(\widehat{\Theta}-\Theta)\Omega^{1/2} = \widehat{H} - H = H\Delta_1 + \Delta_2$ such that

$$\|H^+(\widehat{H}-H)\|_{\mathrm{op}} \le \|\Delta_1\|_{\mathrm{op}} + \|H^+\Delta_2\|_{\mathrm{op}}.$$

By triangle inequality and the definition of operator norm, we can upper bound $\|\Delta_1\|_{\mathrm{op}}$ by

$$\left\|\Omega^{1/2}\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{Z}-\Sigma_Z\right)\Omega^{1/2}\right\|_{\mathrm{op}} + \frac{1}{n}\|\Omega^{1/2}\mathbf{Z}^\top\widetilde{\mathbf{W}}\|_{\mathrm{op}}\|\Omega^{1/2}\|_{\mathrm{op}} + \frac{1}{n}\|\Omega^{1/2}\mathbf{Z}^\top\mathbf{Z}\Delta\|_{\mathrm{op}}\|\Omega^{1/2}\|_{\mathrm{op}}.$$

By adding and subtracting Z, $\|H^+\Delta_2\|_{\mathrm{op}}$ is upper bounded by

$$\frac{1}{n}\|H^+\mathbf{W}^\top\mathbf{Z}\Omega^{1/2}\|_{\mathrm{op}} + \frac{1}{n}\|H^+\mathbf{W}^\top\mathbf{Z}\Delta\|_{\mathrm{op}}\|\Omega^{1/2}\|_{\mathrm{op}} + \left\|H^+\left(\frac{1}{n}\mathbf{W}^\top\widetilde{\mathbf{W}} - \widehat{\Gamma}_{\cdot\widehat{I}}\widehat{\Pi}\right)\right\|_{\mathrm{op}}\|\Omega^{1/2}\|_{\mathrm{op}}.$$

Note that Lemma 9 gives

$$\frac{1}{n}\|\Omega^{1/2}\mathbf{Z}^\top\mathbf{Z}\Delta\|_{\mathrm{op}} \lesssim \frac{1}{n}\max_k\|\Omega^{1/2}\mathbf{Z}^\top\mathbf{Z}_k\|_2\cdot\bar{\rho}\delta_n \le \sqrt{K}\frac{1}{n}\max_{k,k'}|\mathbf{e}_{k'}^\top\Omega^{1/2}\mathbf{Z}^\top\mathbf{Z}\mathbf{e}_k|\cdot\bar{\rho}\delta_n$$

with probability $1-(p\vee n)^{-c}$, where in the second inequality we used $\|v\|_2 \le \sqrt{K}\|v\|_\infty$ for any $v\in\mathbb{R}^K$. And the term $H^+\mathbf{W}^\top\mathbf{Z}\Delta$ can be bounded by the same way. By invoking results (25), (27), (26), (22), and (29) in Lemma 8 and Lemma 9 and taking the union bounds over $k,k'\in[K]$, one has

$$\|\Delta_1\|_{\mathrm{op}} \lesssim \delta_n\sqrt{K}\cdot T_n, \tag{50}$$

$$\|H^+\Delta_2\|_{\mathrm{op}} \lesssim \frac{\delta_n\sqrt{K}}{\sigma_K(H)}\cdot T_n \tag{51}$$

with probability $1 - c(p \vee n)^{-c'}$ and T_n defined in (35). These two displays and condition (34) yield the result in (a).

For part (b), to lower bound the Kth eigenvalue of $\widehat{H}^\top \widehat{H}$, observe that

$$\begin{aligned}\lambda_K\left(\widehat{H}^\top \widehat{H}\right) &\geq \lambda_K\left(\widehat{H}^\top H(H^\top H)^{-1} H^\top \widehat{H}\right) \\ &\geq \lambda_K(H^\top H) \cdot \lambda_K\left(\widehat{H}^\top H(H^\top H)^{-2} H^\top \widehat{H}\ \right) \\ &= \lambda_K(H^\top H) \cdot \sigma_K^2\left((H^\top H)^{-1} H^\top \widehat{H}\ \right).\end{aligned}$$

Since Weyl's inequality gives

$$\sigma_K\left(H^+ \widehat{H}\ \right) \geq 1 - \|H^+(\widehat{H} - H)\|_{\mathrm{op}},$$

invoking (a) finishes the proof of (b). Part (c) follows immediately from part (b) and noting that

$$\lambda_K(\widehat{H}^\top \widehat{H}) = \lambda_K(\Omega^{1/2} \widehat{\Theta}^\top \widehat{\Theta} \Omega^{1/2}) \leq \lambda_K(\widehat{\Theta}^\top \widehat{\Theta})/C_{\min}.$$

To prove (d), starting from $\widehat{H} - H = H\Delta_1 + \Delta_2$, it suffices to upper bound $|u^\top H \Delta_1 v|$ and $|u^\top \Delta_2 v|$. For the first term, recalling that $H = A\Sigma_Z^{1/2}$, we have

$$\begin{aligned}|u^\top H \Delta_1 v| &\leq \|u^\top A\|_1 \cdot \max_k |\mathbf{e}_k^\top (\Sigma_Z)^{1/2} \Delta_Z v| \\ &\leq \|u^\top A\|_1 \cdot \max_k \left|\mathbf{e}_k^T \left(\frac{1}{n}\mathbf{Z}^\top \mathbf{Z} - \Sigma_Z + \frac{1}{n}\mathbf{Z}^\top \widetilde{\mathbf{W}} + \frac{1}{n}\mathbf{Z}^\top \mathbf{Z}\Delta\right) \Omega^{1/2} v\right|\end{aligned}$$

where we used the definition of Δ_1 in (48) in the third line. Noting that $|\mathbf{e}_k^T \mathbf{Z}^\top \mathbf{Z} \Delta \Omega^{1/2} v| \leq \max_a |\mathbf{Z}_{\bullet k}^T \mathbf{Z}_{\bullet a}| \|\Delta \Omega^{1/2} v\|_1$, invoking (25), (27), (26) together with Lemma 9 gives

$$|u^\top H \Delta_1 v| \lesssim \|u^\top A\|_1 \cdot \delta_n \|v\|_2 \cdot T_n$$

with probability greater than $1 - c(p \vee n)^{-c'}$. On the other hand, by the definition of Δ_2 in (49) and using $\widetilde{\mathbf{Z}} = \mathbf{Z} + \mathbf{Z}\Delta$, we can upper bound $|u^\top \Delta_2 v|$ by

$$\left\{\frac{1}{n}|u^\top \mathbf{W}^\top \mathbf{Z} \Omega^{1/2} v| + \frac{1}{n}|u^\top \mathbf{W}^\top \mathbf{Z} \Delta \Omega^{1/2} v| + \left|u^T \left(\frac{1}{n}\mathbf{W}^\top \widetilde{\mathbf{W}} - \widehat{\Gamma}_{\bullet \widehat{I}} \widehat{\Pi}\right) \Omega^{1/2} v\right|\right\}.$$

Invoking (22) and (29) in Lemma 7 together with Lemma 9 again, gives

$$\|u^\top \Delta_2 v| \lesssim \delta_n \|u\|_2 \|v\|_2 T_n$$

with probability $1 - c(p \vee n)^{-c'}$. This concludes the result of (d).

The result of (e) follows by using $\|\widehat{H} - H\|_{\mathrm{op}} \leq \sqrt{pK} \max_{j,k} |\mathbf{e}_j^\top (\widehat{H} - H)\mathbf{e}_k|$ together with choosing $u = \mathbf{e}_j$ and $v = \mathbf{e}_K$ in part (d) and taking a union bound over $1 \leq j \leq p$ and $1 \leq k \leq K$. □

Appendix D: Proof of Theorem 3: convergence rate of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$

D.1. Main proof of Theorem 3

Throughout this proof, we work on the event that $\widehat{\Theta}^\top \widehat{\Theta}$ can be inverted intersected with the event $\mathcal{E}$ defined in (10). This event holds with probability $1 - c(p \vee n)^{-c'}$ by recalling that $\mathbb{P}(\mathcal{E}) \geq 1 - (p \vee n)^{-c}$ and using Lemma 10 in Section C. Further recall that $\mathcal{E}$ implies $\widehat{K} = K$ and $I_k \subseteq \widehat{I}_k \subseteq I_k \cup J_1^k$ with $J_1^k = \{j \in J : \ |A_{jk}| \geq 1 - 4\delta/v\}$, for all $k \in [K]$. Define

$$\widehat{\Theta}^+ := [\widehat{\Theta}^\top \widehat{\Theta}]^{-1} \widehat{\Theta}^\top$$

so that $\widehat{\Theta}^+ \widehat{\Theta} = \mathbf{I}_K$. In the same way, using $\text{rank}(\Theta) = K$, define $\Theta^+ := (\Theta^\top \Theta)^{-1} \Theta^\top$. Recall that

$$\widehat{\beta} = \widehat{\Theta}^+ \frac{1}{n} \mathbf{X}^\top \mathbf{y},$$

hence

$$\begin{aligned} \widehat{\beta} - \beta &= \widehat{\Theta}^+ \left(\frac{1}{n} \mathbf{X}^\top \mathbf{y} - \widehat{\Theta} \beta \right) \\ &= \Theta^+ \left(\frac{1}{n} \mathbf{X}^\top \mathbf{y} - \widehat{\Theta} \beta \right) + \left(\widehat{\Theta}^+ - \Theta^+ \right) \left(\frac{1}{n} \mathbf{X}^\top \mathbf{y} - \widehat{\Theta} \beta \right). \end{aligned} \tag{52}$$

Let $\delta_n^2 := \log(p \vee n)/n$ and let $\bar{\rho}$ be as defined in (18). Lemma 13 in Section D.2 implies that, with probability $1 - (p \vee n)^{-c}$,

$$\begin{aligned} &\left\| \left(\widehat{\Theta}^+ - \Theta^+ \right) \left(\frac{1}{n} \mathbf{X}^\top \mathbf{y} - \widehat{\Theta} \beta \right) \right\|_2 \\ &\quad \lesssim C_{\min}^{-1/2} \cdot \frac{\delta_n \sqrt{p}}{\sigma_K(A \Sigma_Z^{1/2})} \cdot \left\{ 1 + \frac{\sqrt{p}}{\sigma_K(A \Sigma_Z^{1/2})} \right\} \delta_n \sqrt{K} \left\{ 1 + \frac{\|\beta\|_2}{\sqrt{m}} + \bar{\rho} \|\beta\|_2 \right\}. \end{aligned} \tag{53}$$

It remains to study the first term on the right of (52). Recall the definition of $\widehat{\Theta}$ in (12),

$$\widehat{\Theta} = \left(\widehat{\Sigma}_{\cdot \widehat{I}} - \widehat{\Gamma}_{\cdot \widehat{I}} \right) \widehat{\Pi}$$

with

$$\widehat{\Pi} := \widehat{A}_{\widehat{I} \cdot} \left[\widehat{A}_{\widehat{I} \cdot}^\top \widehat{A}_{\widehat{I} \cdot} \right]^{-1}.$$

Use the fact that $\mathbf{X} = \mathbf{Z} A^T + \mathbf{W}$ to obtain

$$\begin{aligned} \widehat{\Theta} &= \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X}_{\cdot \widehat{I}} - \widehat{\Gamma}_{\cdot \widehat{I}} \right) \widehat{\Pi} \\ &= \left[A \frac{1}{n} \mathbf{Z}^\top \mathbf{Z} A_{\widehat{I} \cdot}^\top + A \frac{1}{n} \mathbf{Z}^\top \mathbf{W}_{\cdot \widehat{\mathrm{I}}} + \frac{1}{n} \mathbf{W}^\top \mathbf{Z} A_{\widehat{I} \cdot}^\top + \left(\frac{1}{n} \mathbf{W}^\top \mathbf{W}_{\cdot \widehat{I}} - \widehat{\Gamma}_{\cdot \widehat{I}} \right) \right] \widehat{\Pi}. \end{aligned} \tag{54}$$

Combine this expression with the expansion

$$\frac{1}{n}\mathbf{X}^\top \mathbf{y} = A\frac{1}{n}\mathbf{Z}^\top \mathbf{Z}\beta + A\frac{1}{n}\mathbf{Z}^\top \varepsilon + \frac{1}{n}\mathbf{W}^\top \mathbf{Z}\beta + \frac{1}{n}\mathbf{W}^\top \varepsilon, \tag{55}$$

to arrive at

$$\frac{1}{n}\mathbf{X}^\top \mathbf{y} - \widehat{\Theta}\beta = A\Delta_Z + \Delta_W \tag{56}$$

with

$$\Delta_Z := \frac{1}{n}\mathbf{Z}^\top \varepsilon - \frac{1}{n}\mathbf{Z}^\top \mathbf{W}_{\bullet \widehat{I}}\widehat{\Pi}\beta + \frac{1}{n}\mathbf{Z}^\top \mathbf{Z}\left(\mathbf{I}_K - A^\top_{\widehat{I}\bullet}\widehat{\Pi}\right)\beta, \tag{57}$$

$$\Delta_W := \frac{1}{n}\mathbf{W}^\top \varepsilon + \frac{1}{n}\mathbf{W}^\top \mathbf{Z}\left(\mathbf{I}_K - A^\top_{\widehat{I}\bullet}\widehat{\Pi}\right)\beta + \left(\widehat{\Gamma}_{\bullet \widehat{I}} - \frac{1}{n}\mathbf{W}^\top \mathbf{W}_{\bullet \widehat{I}}\right)\widehat{\Pi}\beta. \tag{58}$$

Recall that $\Theta = A\Sigma_Z$ and $\Theta^+\Theta = \mathbf{I}_K$, to arrive at the identity

$$\begin{aligned}\Theta^+\left(\frac{1}{n}\mathbf{X}^\top \mathbf{y} - \widehat{\Theta}\beta\right) &= \Theta^+ A\Sigma_Z\Sigma_Z^{-1}\Delta_Z + \Theta^+\Delta_W \\ &= \Sigma_Z^{-1}\Delta_Z + \Theta^+\Delta_W. \end{aligned}\tag{59}$$

Use the inequality $\|\Sigma_Z^{-1}\Delta_Z\|_2 \le \|\Sigma_Z^{-1/2}\|_{\mathrm{op}}\|\Sigma_Z^{-1/2}\Delta_Z\|_2 \le C_{\min}^{-1/2}\|\Sigma_Z^{-1/2}\Delta_Z\|_2$ and invoke Lemma 11 and 12 in Section D.2 to obtain the bounds

$$\left\|\Sigma_Z^{-1}\Delta_Z\right\|_2 \lesssim C_{\min}^{-1/2}\delta_n\sqrt{K}\left(1 + \frac{\|\beta\|_2}{\sqrt{m}} + \bar{\rho}\|\beta\|_2\right), \tag{60}$$

$$\|\Theta^+\Delta_W\|_2 \lesssim C_{\min}^{-1/2}\frac{\delta_n\sqrt{K}}{\sigma_K(A\Sigma_Z^{1/2})}\left(1 + \frac{\|\beta\|_2}{\sqrt{m}} + \bar{\rho}\|\beta\|_2\right). \tag{61}$$

These inequalities hold with probability $1 - C(p \vee n)^{-c}$. Use the inequality

$$\sigma_K(A\Sigma_Z^{1/2}) \ge C_{\min}^{1/2}\sigma_K(A) \ge C_{\min}^{1/2}\sigma_K(A_{I\bullet}) \ge C_{\min}^{1/2}\sqrt{m} \ge \sqrt{2C_{\min}},$$

and collect (53), (60) and (61) to conclude that, with probability $1 - C(p \vee n)^{-c}$,

$$\begin{aligned}\|\widehat{\beta} - \beta\|_2 &\le \left\|\Theta^+\left(\frac{1}{n}\mathbf{X}^\top \mathbf{y} - \widehat{\Theta}\beta\right)\right\|_2 + \left\|\left(\widehat{\Theta}^+ - \Theta^+\right)\left(\frac{1}{n}\mathbf{X}^\top \mathbf{y} - \widehat{\Theta}\beta\right)\right\|_2 \\ &\le C_{\min}^{-1/2}\left\|\Sigma_Z^{-1}\Delta_Z\right\|_2 + \|\Theta^+\Delta_W\|_2 + \left\|\left(\widehat{\Theta}^+ - \Theta^+\right)\left(\frac{1}{n}\mathbf{X}^\top \mathbf{y} - \widehat{\Theta}\beta\right)\right\|_2 \\ &\lesssim C_{\min}^{-1/2}\delta_n\sqrt{K}\left\{1 + \frac{\|\beta\|_2}{\sqrt{m}} + \bar{\rho}\|\beta\|_2\right\}\left\{1 + \frac{\delta_n\sqrt{p}}{\sigma_K(A\Sigma_Z^{1/2})}\cdot\left(1 + \frac{\sqrt{p}}{\sigma_K(A\Sigma_Z^{1/2})}\right)\right\}.\end{aligned}$$

Recall that $\lambda_K := \lambda_K(A\Sigma_Z A^\top) = \sigma_K^2(A\Sigma_Z^{1/2})$ and invoke Lemma 14 in Section D.2 to deduce $p/\lambda_K \geq K/B_z$. This implies

$$\frac{\delta_n\sqrt{p}}{\sigma_K(A\Sigma_Z^{1/2})}\cdot\left\{1+\frac{\sqrt{p}}{\sigma_K(A\Sigma_Z^{1/2})}\right\} = \delta_n\sqrt{\frac{p}{\lambda_K}} + \delta_n\frac{p}{\lambda_K} \lesssim \delta_n\frac{p}{\lambda_K}, \tag{62}$$

whence

$$\|\widehat{\beta}-\beta\|_2 \quad \lesssim \quad C_{\min}^{-1/2}\delta_n\sqrt{K}\left(1+\frac{\|\beta\|_2}{\sqrt{m}}+\bar{\rho}\|\beta\|_2\right)\left(1+\delta_n\frac{p}{\lambda_K}\right).$$

Finally, invoke Assumptions 3 and 5 to complete the proof. □

D.2. Main lemmas used in the proof of Theorem 3

The first two lemmas provide upper bounds for the Euclidean norm of $(\Sigma_Z)^{-1}\Delta_Z$ and Δ_W defined in (57) and (58). The third lemma controls the sup-norm and ℓ_2 norm of $(\widehat{\Theta}^+ - \Theta^+)(\frac{1}{n}\mathbf{X}^\top\mathbf{y} - \widehat{\Theta}\beta)$ in (52). The final lemma gives a lower bound on the "signal strength" $\lambda_K(A\Sigma_Z A^\top)$. All lemmas are proved in Section D.3. We use throughout the notation

$$H := A[\Sigma_Z]^{1/2} = \Theta[\Sigma_Z^{-1}]^{1/2}.$$

All statements are valid on some events that are subsets of $\mathcal{E}$ and the probabilities of these events are greater than $1 - C(p\vee n)^{-\alpha}$ for some positive constants C, α. This is an important observation since on the event $\mathcal{E}$, the dimensions $\widehat{K}$ and K are equal, which ensures that the various quantities in the statements are well-defined. For instance, Δ_Z, Δ_W and $\widehat{\Theta}-\Theta$ are only defined if $\widehat{K} = K$.

Lemma 11. *Under the conditions of Theorem 3, with probability* $1 - c(p\vee n)^{-c'}$,

$$\left\|(\Sigma_Z)^{-1/2}\Delta_Z\right\|_2 \lesssim \delta_n\sqrt{K}\left(1+\frac{\|\beta\|_2}{\sqrt{m}}+\bar{\rho}\|\beta\|_2\right)$$

and

$$\|\Delta_Z\|_\infty \lesssim \delta_n\left(1+\frac{\|\beta\|_2}{\sqrt{m}}+\bar{\rho}\|\beta\|_2\right). \tag{63}$$

Lemma 12. *Under the conditions of Theorem 3, with probability* $1 - c(p\vee n)^{-c'}$,

$$\|\Theta^+\Delta_W\|_2 \lesssim C_{\min}^{-1/2}\frac{\delta_n\sqrt{K}}{\sigma_K(H)}\left(1+\frac{\|\beta\|_2}{\sqrt{m}}+\bar{\rho}\|\beta\|_2\right)$$

and

$$\|\Delta_W\|_\infty \lesssim \delta_n\left(1+\frac{\|\beta\|_2}{\sqrt{m}}+\bar{\rho}\|\beta\|_2\right). \tag{64}$$

Lemma 13. *Under the conditions of Theorem 3, with probability* $1 - c(p \vee n)^{-c'}$*, we have*

$$\left|\mathbf{e}_k^\top \left(\widehat{\Theta}^+ - \Theta^+\right)\left(\frac{1}{n}\mathbf{X}^\top \mathbf{y} - \widehat{\Theta}\beta\right)\right| \lesssim [\Sigma_Z^{-1}]_{kk}^{1/2} \cdot \frac{\delta_n\sqrt{p}}{\sigma_K(H)} \cdot \left(\sqrt{K} + \frac{\sqrt{p}}{\sigma_K(H)}\right)\delta_n\left(1 + \frac{\|\beta\|_2}{\sqrt{m}} + \bar{\rho}\|\beta\|_2\right)$$

for any $1 \le k \le K$*. Moreover, with the same probability,*

$$\left\|\left(\widehat{\Theta}^+ - \Theta^+\right)\left(\frac{1}{n}\mathbf{X}^\top \mathbf{y} - \widehat{\Theta}\beta\right)\right\|_2 \lesssim C_{\min}^{-1/2} \cdot \frac{\delta_n\sqrt{p}}{\sigma_K(H)} \cdot \left(1 + \frac{\sqrt{p}}{\sigma_K(H)}\right)\delta_n\sqrt{K}\left(1 + \frac{\|\beta\|_2}{\sqrt{m}} + \bar{\rho}\|\beta\|_2\right)$$

Lemma 14. *Under Assumptions 1 and 2, we have*

$$\lambda_K(A\Sigma_Z A^\top) \le B_z \frac{p}{K}.$$

D.3. Proof of lemmas in Section D.2

D.3.1. Proof of Lemma 11

We work on the event $\mathcal{E}$ defined in (10) that has probability at least $1 - (p \vee n)^{-c}$ for some $c > 0$. Recall that it implies $\widehat{K} = K$ and $I_k \subseteq \widehat{I}_k \subseteq I_k \cup J_1^k$ for all $k \in [K]$.

By writing $\Omega = \Sigma_Z^{-1}$, note that

$$\|\Omega\Delta_Z\|_2 \le \|\Omega^{1/2}\|_{\text{op}} \cdot \|\Omega^{1/2}\Delta_Z\|_2 \le C_{\min}^{-1/2} \cdot \sqrt{K} \cdot \|\Omega^{1/2}\Delta_Z\|_\infty.$$

From (57), it suffices to bound

$$\max_k \left[\frac{1}{n}|\Omega_{k\bullet}^{1/2}\mathbf{Z}^\top\varepsilon| + \frac{1}{n}|\Omega_{k\bullet}^{1/2}\mathbf{Z}^\top\mathbf{W}_{\bullet\widehat{I}}\widehat{\Pi}\beta| + \frac{1}{n}\left|\Omega_{k\bullet}^{1/2}\mathbf{Z}^\top\mathbf{Z}\left(\mathbf{I}_K - A_{\widehat{I}\bullet}^\top\widehat{\Pi}\right)\beta\right|\right].$$

Note that, by definition (19),

$$\left|\Omega_{k\bullet}^{1/2}\mathbf{Z}^\top\mathbf{Z}\left(\mathbf{I}_K - A_{\widehat{I}\bullet}^\top\widehat{\Pi}\right)\beta\right| \le \max_a \left|\Omega_{k\bullet}^{1/2}\mathbf{Z}^\top\mathbf{Z}_{\bullet a}\right| \|\Delta\beta\|_1.$$

Invoking (20), (27), (26) in Lemma 7 together with Lemma 9 concludes the proof of the first result. The second result follows immediately by the same arguments applied to $\max_k |\mathbf{e}_k^\top\Delta_Z|$. □

D.3.2. Proof of Lemma 12

We work on the event $\mathcal{E}$ defined in (10) that has probability at least $1-(p\vee n)^{-c}$ for some $c>0$. Recall that it implies $\widehat{K}=K$ and $I_k\subseteq\widehat{I}_k\subseteq I_k\cup J_1^k$ for all $k\in[K]$.

First recall that $H\Sigma_Z^{1/2}=\Theta$ implies $\Theta^+=\Omega^{1/2}H^+$. Note that

$$\|\Theta^+\Delta_W\|_2=\|\Omega^{1/2}H^+\Delta_W\|_2\le C_{\min}^{-1/2}\sqrt{K}\|H^+\Delta_W\|_\infty.$$

By definitions (58) and (19), we can bound

$$\max_k\left\{\frac{1}{n}\left|\mathbf{e}_k^\top H^+\mathbf{W}^\top\varepsilon\right|+\frac{1}{n}\left\|\mathbf{e}_k^\top H^+\mathbf{W}^\top\mathbf{Z}\right\|_\infty\|\Delta\beta\|_1+\left|\mathbf{e}_k^\top H^+\left(\frac{1}{n}\mathbf{W}^\top\mathbf{W}_{\cdot\widehat{I}}-\widehat{\Gamma}_{\cdot\widehat{I}}\right)\widehat{\Pi}\beta\right|\right\}.$$

Apply (23), (24) and (28) in Lemma 7, apply Lemma 9 and take an union bound to complete the proof of the first result. To prove (64), since

$$\|\Delta_W\|_\infty\le\max_j\left\{\frac{1}{n}\left|\mathbf{W}_{\cdot j}^\top\varepsilon\right|+\frac{1}{n}\left\|\mathbf{W}_{\cdot j}^\top\mathbf{Z}\right\|_\infty\|\Delta\beta\|_1+\left|\mathbf{e}_j^\top\left(\frac{1}{n}\mathbf{W}^\top\mathbf{W}_{\cdot\widehat{I}}-\widehat{\Gamma}_{\cdot\widehat{I}}\right)\widehat{\Pi}\beta\right|\right\},$$

invoking (21), (22) and (29) in Lemma 7 and applying Lemma 9 completes the proof. □

D.3.3. Proof of Lemma 13

We work on the event $\mathcal{E}$ defined in (10) that has probability at least $1-(p\vee n)^{-c}$ for some $c>0$. Recall that it implies $\widehat{K}=K$ and $I_k\subseteq\widehat{I}_k\subseteq I_k\cup J_1^k$ for all $k\in[K]$.

Using the identity (56) and $\Theta=A\Sigma_Z$ gives

$$\begin{aligned}\left(\widehat{\Theta}^+-\Theta^+\right)\left(\frac{1}{n}\mathbf{X}^\top\mathbf{y}-\widehat{\Theta}\beta\right)&=\left(\widehat{\Theta}^+-\Theta^+\right)(A\Delta_Z+\Delta_W)\\&=\left([\Theta^\top\Theta]^{-1}[\widehat{\Theta}-\Theta]^\top P_{\widehat{\Theta}}^\perp+\Theta^+[\Theta-\widehat{\Theta}]\widehat{\Theta}^+\right)(A\Delta_Z+\Delta_W).\end{aligned}\tag{65}$$

In the last line, we used $P_{\widehat{\Theta}}^\perp=\mathbf{I}_p-\widehat{\Theta}\widehat{\Theta}^+$ and the identity

$$\widehat{\Theta}^+-\Theta^+=(\Theta^\top\Theta)^{-1}(\widehat{\Theta}-\Theta)^\top P_{\widehat{\Theta}}^\perp+\Theta^+(\Theta-\widehat{\Theta})\widehat{\Theta}^+.\tag{66}$$

To prove the first statement of the lemma, recall that $\Theta\Omega^{1/2}=H$ and $\widehat{\Theta}\Omega^{1/2}=\widehat{H}$, so that we can write

$$\begin{aligned}&\left|\mathbf{e}_k^\top\left(\widehat{\Theta}^+-\Theta^+\right)\left(\frac{1}{n}\mathbf{X}^\top\mathbf{y}-\widehat{\Theta}\beta\right)\right|\\&=\left|\mathbf{e}_k^\top\left([\Theta^\top\Theta]^{-1}[\widehat{\Theta}-\Theta]^\top P_{\widehat{\Theta}}^\perp+\Theta^+[\Theta-\widehat{\Theta}]\widehat{\Theta}^+\right)(A\Delta_Z+\Delta_W)\right|\\&=\left|\mathbf{e}_k^\top\Omega^{1/2}\left([H^\top H]^{-1}[\widehat{H}-H]^\top P_{\widehat{\Theta}}^\perp+H^+[H-\widehat{H}]\widehat{H}^+\right)(A\Delta_Z+\Delta_W)\right|\\&\le\left\|\mathbf{e}_k^\top\Omega^{1/2}\left([H^\top H]^{-1}[\widehat{H}-H]^\top P_{\widehat{\Theta}}^\perp+H^+[H-\widehat{H}]\widehat{H}^+\right)\right\|_2(\|A\Delta_Z\|_2+\|\Delta_W\|_2)\end{aligned}$$

by using the Cauchy-Schwarz inequality in the last step. Note that

$$\|A\Delta_Z\|_2 + \|\Delta_W\|_2 \le \sqrt{p}\max_j \left(|A_{j\bullet}^\top \Delta_Z| + \|\Delta_W\|_\infty\right) \le \sqrt{p}(\|\Delta_Z\|_\infty + \|\Delta_W\|_\infty)$$

from the fact that $\|A_{j\bullet}\|_1 \le 1$. The first term can be bounded above via

$$\begin{aligned}
&\left\|\mathbf{e}_k^\top \Omega^{1/2}(H^\top H)^{-1}(\widehat{H}-H)^\top P_{\widehat{\Theta}}^\perp\right\|_2 + \left\|\mathbf{e}_k^\top \Omega^{1/2} H^+(\widehat{H}-H)\widehat{H}^+\right\|_2 \\
&\le \left\|\mathbf{e}_k^\top \Omega^{1/2}(H^\top H)^{-1}(\widehat{H}-H)^\top\right\|_2 + \Omega_{kk}^{1/2}\frac{\|H^+(\widehat{H}-H)\|_{\mathrm{op}}}{\sigma_K(\widehat{H})}.
\end{aligned}$$

Apply Lemma 10 to obtain

$$\sigma_K(\widehat{H}) \ge (1-c_1)\sigma_K(H), \qquad \|H^+(\widehat{H}-H)\|_{\mathrm{op}} \lesssim \delta_n\sqrt{K} \tag{67}$$

and

$$\begin{aligned}
\left\|\mathbf{e}_k^\top \Omega^{1/2}(H^\top H)^{-1}(\widehat{H}-H)^\top\right\|_2 &\le \sqrt{p}\max_{j\in[K]}\left|\mathbf{e}_k^\top \Omega^{1/2}(H^\top H)^{-1}(\widehat{H}-H)^\top \mathbf{e}_j\right| \\
&\lesssim \sqrt{p}\cdot\delta_n\left\|\mathbf{e}_k^\top \Omega^{1/2}(H^\top H)^{-1}\right\|_2 \\
&\le \frac{\delta_n\sqrt{p}}{\lambda_K(H^\top H)}\Omega_{kk}^{1/2}.
\end{aligned} \tag{68}$$

Hence,

$$\begin{aligned}
&\left\|\mathbf{e}_k^\top \Omega^{1/2}(H^\top H)^{-1}(\widehat{H}-H)^\top P_{\widehat{\Theta}}^\perp\right\|_2 + \left\|\mathbf{e}_k^\top \Omega^{1/2} H^+(\widehat{H}-H)\widehat{H}^+\right\|_2 \\
&\lesssim \left(\frac{\delta_n\sqrt{p}}{\lambda_K(H^\top H)} + \frac{\delta_n\sqrt{K}}{\sigma_K(H)}\right)\Omega_{kk}^{1/2},
\end{aligned} \tag{69}$$

and, using (63) and (64), we conclude

$$\begin{aligned}
&\left\|\mathbf{e}_k^\top \Omega^{1/2}\left([H^\top H]^{-1}[\widehat{H}-H]^\top P_{\widehat{\Theta}}^\perp + H^+(H-\widehat{H})\widehat{H}^+\right)\right\|_2 (\|A\Delta_Z\|_2 + \|\Delta_W\|_2) \\
&\lesssim \Omega_{kk}^{1/2}\left(\frac{\delta_n p}{\lambda_K(H^\top H)} + \frac{\delta_n\sqrt{pK}}{\sigma_K(H)}\right)\cdot\delta_n\left(1 + \frac{\|\beta\|_2}{\sqrt{m}} + \bar{\rho}\|\beta\|_2\right)
\end{aligned}$$

with probability $1-(p\vee n)^{-c}$. This proves the first statement.

We now prove the second statement. From (65), observe that

$$\begin{aligned}
&\left\|\left([\Theta^\top\Theta]^{-1}[\widehat{\Theta}-\Theta]^\top P_{\widehat{\Theta}}^\perp + \Theta^+[\Theta-\widehat{\Theta}]\widehat{\Theta}^+\right)(A\Delta_Z + \Delta_W)\right\|_2 \\
&\le \|\Omega^{1/2}\|_{\mathrm{op}}\left\|\left([H^\top H]^{-1}[\widehat{H}-H]^\top P_{\widehat{\Theta}}^\perp + H^+[H-\widehat{H}]\widehat{H}^+\right)(A\Delta_Z + \Delta_W)\right\|_2 \\
&\le \|\Omega^{1/2}\|_{\mathrm{op}}\left[\frac{\|\widehat{H}-H\|_{\mathrm{op}}}{\lambda_K(H^\top H)} + \frac{\|H^+(\widehat{H}-H)\|_{\mathrm{op}}}{\sigma_K(\widehat{H})}\right]\|A\Delta_Z + \Delta_W\|_2.
\end{aligned}$$

Invoking (67), (69), (63) and Lemma 10 concludes the proof. □

D.3.4. Proof of Lemma 14

We argue that

$$\begin{aligned}\lambda_K(A\Sigma_Z A^\top) &= \lambda_{\min}(\Sigma_Z^{1/2} A^\top A \Sigma_Z^{1/2}) \\ &\le \min_{k\in[K]} \mathbf{e}_k^\top \Sigma_Z^{1/2} A^\top A \Sigma_Z^{1/2} \mathbf{e}_k \\ &\le \frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{p} A_{j\bullet}^\top \Sigma_Z^{1/2} \mathbf{e}_k \mathbf{e}_k^\top \Sigma_Z^{1/2} A_{j\bullet} \\ &= \frac{1}{K}\sum_{j=1}^{p} A_{j\bullet}^\top \Sigma_Z A_{j\bullet}.\end{aligned}$$

By $\|A_{j\bullet}\|_1 \le 1$, the result follows from the inequality

$$A_{j\bullet}^\top \Sigma_Z A_{j\bullet} \le \|A_{j\bullet}\|_1^2 \|\Sigma_Z\|_\infty \le B_z.$$

□

Appendix E: Proof of Theorem 4: Asymptotic normality of $\widehat{\beta}$

E.1. Main proof of Theorem 4

We work throughout this proof on the event $\mathcal{E}$ so that $\widehat{K} = K$ and $I_k \subseteq \widehat{I}_k \subseteq I_k \cup J_1^k$ with $J_1^k = \{j \in J : |A_{jk}| \ge 1 - 4\delta/v\}$, for all $k \in [K]$, and the event that the inverse of $\widehat{\Theta}^\top \widehat{\Theta}$ exists. This event holds with probability $1 - c(p \vee n)^{-c'}$ by recalling that $\mathbb{P}(\mathcal{E}) \ge 1 - (p \vee n)^{-c}$ and Lemma 10. Recall from (52) the identity

$$\widehat{\beta} - \beta = \Theta^+ \left(\frac{1}{n}\mathbf{X}^\top \mathbf{y} - \widehat{\Theta}\beta\right) + \left(\widehat{\Theta}^+ - \Theta^+\right)\left(\frac{1}{n}\mathbf{X}^\top \mathbf{y} - \widehat{\Theta}\beta\right). \tag{70}$$

The proof consists of four main steps:

(1) We derive the decomposition of $\widehat{\beta}_k - \beta_k$ as an average of independent, mean zero random variables $n^{-1}\sum_{i=1}^n \xi_{ik}$ and two remainder terms;
(2) We verify that $\mathbb{E}[\xi_{ik}] = 0$ and calculate the variance $\mathbb{E}[\xi_{ik}^2]$;
(3) We apply Lyapunov's central limit theorem for triangular arrays to $\sum_{t=1}^n \xi_{tk}$;
(4) We show that the two remainder terms are asymptotically negligible.

Step 1: Set $\Pi := A_{I\bullet}[A_{I\bullet}^\top A_{I\bullet}]^{-1}$. Recall (56), (57) and (58), to obtain the identity

$$\frac{1}{n}\mathbf{X}^\top\mathbf{y} - \widehat{\Theta}\beta = A\Delta_Z + \Delta_W$$

and rewrite Δ_Z and Δ_W by adding and subtracting appropriate terms as

$$\Delta_Z = \underbrace{\frac{1}{n}\mathbf{Z}^\top\varepsilon - \frac{1}{n}\mathbf{Z}^\top\mathbf{W}_{\bullet I}\Pi\beta}_{\Delta_{Z,I}} + \underbrace{\frac{1}{n}\mathbf{Z}^\top\left(\mathbf{W}_{\bullet I}\Pi - \mathbf{W}_{\bullet\widehat{I}}\widehat{\Pi}\right)\beta + \frac{1}{n}\mathbf{Z}^\top\mathbf{Z}\left(\mathbf{I}_K - A_{\widehat{I}\bullet}^\top\widehat{\Pi}\right)\beta}_{\Delta_{Z,J_1}},$$

$$\Delta_W = \underbrace{\frac{1}{n}\mathbf{W}^\top\varepsilon + \left(\widetilde{\Gamma}_{\bullet I} - \frac{1}{n}\mathbf{W}^\top\mathbf{W}_{\bullet I}\right)\Pi\beta}_{\Delta_{W,I}}$$
$$+ \underbrace{\frac{1}{n}\mathbf{W}^\top\mathbf{Z}\left(\mathbf{I}_K - A_{\widehat{I}\bullet}^\top\widehat{\Pi}\right)\beta + \left[\left(\widehat{\Gamma}_{\bullet\widehat{I}} - \frac{1}{n}\mathbf{W}^\top\mathbf{W}_{\bullet\widehat{I}}\right)\widehat{\Pi} - \left(\widetilde{\Gamma}_{\bullet I} - \frac{1}{n}\mathbf{W}^\top\mathbf{W}_{\bullet I}\right)\Pi\right]\beta}_{\Delta_{W,J_1}}.$$

Here, in view of (11), the matrix $\widetilde{\Gamma}$ is a diagonal $p\times p$ matrix with

$$\widetilde{\Gamma}_{ii} := \widehat{\Sigma}_{ii} - \frac{1}{|I_a|(|I_a|-1)}\sum_{j,\ell\in I_a,\ j\neq\ell}\widehat{\Sigma}_{j\ell} \quad \forall i\in I_a,\ a\in[K]. \tag{71}$$

Plug (59) into (70), use the above expressions of Δ_Z and Δ_W, and find, for each $k\in[K]$,

$$\begin{aligned}
\widehat{\beta}_k - \beta_k &= \mathbf{e}_k^\top\Theta^+\left(\frac{1}{n}\mathbf{X}^\top\mathbf{y} - \widehat{\Theta}\beta\right) + \mathbf{e}_k^\top\left(\widehat{\Theta}^+ - \Theta^+\right)\left(\frac{1}{n}\mathbf{X}^\top\mathbf{y} - \widehat{\Theta}\beta\right)\\
&= \mathbf{e}_k^\top\left(\Sigma_Z^{-1}\Delta_Z + \Theta^+\Delta_W\right) + \mathbf{e}_k^\top\left(\widehat{\Theta}^+ - \Theta^+\right)\left(\frac{1}{n}\mathbf{X}^\top\mathbf{y} - \widehat{\Theta}\beta\right)\\
&= \mathbf{e}_k^\top\left(\Sigma_Z^{-1}\Delta_{Z,I} + \Theta^+\Delta_{W,I}\right) + [\mathrm{Rem}_1]_k + [\mathrm{Rem}_2]_k
\end{aligned}$$

with

$$\mathrm{Rem}_1 = \left(\widehat{\Theta}^+ - \Theta^+\right)\left(\frac{1}{n}\mathbf{X}^\top\mathbf{y} - \widehat{\Theta}\beta\right), \qquad \mathrm{Rem}_2 = \Sigma_Z^{-1}\Delta_{Z,J_1} + \Theta^+\Delta_{W,J_1}. \tag{72}$$

We now write $\mathbf{e}_k^\top\left(\Sigma_Z^{-1}\Delta_{Z,I} + \Theta^+\Delta_{W,I}\right)$ as a sum of independent variables. First observe that

$$\begin{aligned}
\mathbf{e}_k^\top\Sigma_Z^{-1}\Delta_{Z,I} &= \mathbf{e}_k^\top\Sigma_Z^{-1}\left(\frac{1}{n}\mathbf{Z}^\top\varepsilon - \frac{1}{n}\mathbf{Z}^\top\mathbf{W}_{\bullet I}\Pi\beta\right)\\
&= \mathbf{e}_k^\top\Sigma_Z^{-1}\frac{1}{n}\sum_{t=1}^n\mathbf{Z}_{t\bullet}\left(\varepsilon_t - \mathbf{W}_{tI}\Pi\beta\right)
\end{aligned}$$

and

$$\begin{aligned}\mathbf{e}_k^\top \Theta^+ \Delta_{W,I} &= \mathbf{e}_k^\top \Theta^+ \left\{ \frac{1}{n}\mathbf{W}^\top \varepsilon + \left(\widetilde{\Gamma}_{\bullet I} - \frac{1}{n}\mathbf{W}^\top \mathbf{W}_{\bullet I} \right) \Pi\beta \right\} \\ &= \mathbf{e}_k^\top \Theta^+ \left\{ \frac{1}{n}\sum_{t=1}^n \mathbf{W}_{t\bullet}\varepsilon_t + \frac{1}{n}\sum_{t=1}^n \widetilde{\Gamma}_{\bullet I}^{(t)} \Pi\beta - \frac{1}{n}\sum_{t=1}^n \mathbf{W}_{t\bullet}\mathbf{W}_{tI}\Pi\beta \right\}\end{aligned}$$

with, for any $t \in [n]$, $\widetilde{\Gamma}_{ij}^{(t)} := 0$ for all $i \neq j$ and

$$\widetilde{\Gamma}_{ii}^{(t)} := \mathbf{X}_{ti}^2 - \frac{1}{|I_a|(|I_a|-1)} \sum_{j,\ell \in I_a,\ j\neq \ell} \mathbf{X}_{tj}\mathbf{X}_{t\ell}, \text{ for all } i \in I_a,\ a \in [K], \tag{73}$$

so that $\widetilde{\Gamma}_{\bullet I} = n^{-1}\sum_{t=1}^n \widetilde{\Gamma}_{\bullet I}^{(t)}$. Finally, define

$$\xi_{tk} := \mathbf{e}_k^\top \Sigma_Z^{-1} \mathbf{Z}_{t\bullet} \left[\varepsilon_t - \mathbf{W}_{t\bullet}\Pi\beta \right] + \mathbf{e}_k^\top \Theta^+ \left[\mathbf{W}_{t\bullet}\varepsilon_t + \left(\widetilde{\Gamma}_{\bullet I}^{(t)} - \mathbf{W}_{t\bullet}\mathbf{W}_{tI} \right) \Pi\beta \right] \tag{74}$$

and conclude that

$$\widehat{\beta}_k - \beta_k = \frac{1}{n}\sum_{t=1}^n \xi_{tk} + [\mathrm{Rem}_1]_k + [\mathrm{Rem}_2]_k. \tag{75}$$

It is easily verified that, for each k, the random variables $\xi_{1k}, \ldots, \xi_{nk}$ are independent.

Step 2: Next we calculate the first two moments of ξ_{tk}. Lemma 15 in Section E.2 shows that $\mathbb{E}[\xi_{tk}] = 0$ and

$$\begin{aligned}\mathbb{E}[\xi_{tk}^2] = &\left\{ \sigma^2 + \sum_{\ell=1}^K \beta_\ell^2 \left(\frac{1}{|I_\ell|^2} \sum_{i\in I_\ell} \tau_i^2 \right) \right\} \left\{ [\Sigma_Z^{-1}]_{kk} + \mathbf{e}_k^\top (\Theta^\top\Theta)^{-1}\Theta^\top \Gamma \Theta (\Theta^\top\Theta)^{-1}\mathbf{e}_k \right\} \\ &+ \sum_{\ell=1}^K \sum_{a\in I_\ell} \left(\mathbf{e}_k^\top (\Theta^\top\Theta)^{-1}\Theta^\top \mathbf{e}_a \right)^2 \frac{\beta_\ell^2}{m_\ell} \sum_{i\in I_\ell} \tau_i^2 \left(\frac{1}{(|I_\ell|-1)^2} \sum_{j\in I_\ell \setminus \{i\}} \tau_j^2 - \frac{1}{|I_\ell|^2}\sum_{j\in I_\ell} \tau_j^2 \right),\end{aligned} \tag{76}$$

for all $1 \le t \le n$. In case $|I_1| = \cdots = |I_K| = m$ and $\tau_1^2 = \cdots = \tau_p^2 = \tau^2$, the above expression simplifies to

$$\left(\sigma^2 + \tau^2 \frac{\|\beta\|_2^2}{m} \right) \left([\Sigma_Z^{-1}]_{kk} + \tau^2 \mathbf{e}_k^\top [\Theta^\top\Theta]^{-1}\mathbf{e}_k \right) + \frac{\tau^4}{m(m-1)} \sum_{\ell=1}^K \beta_\ell^2 \sum_{i\in I_\ell} \left(\mathbf{e}_k^\top [\Theta^\top\Theta]^{-1}\Theta^\top \mathbf{e}_i \right)^2. \tag{77}$$

This corresponds to the expression for V_k in (28). Furthermore, since

$$\tau^2 \mathbf{e}_k^\top [\Theta^\top\Theta]^{-1}\mathbf{e}_k \le \frac{\tau^2 [\Sigma_Z^{-1}]_{kk}}{\lambda_K(A\Sigma_Z A^\top)} \tag{78}$$

and the second term in (77) is always smaller than the first term (as shown below), it follows that V_k reduces to (29), provided $\lambda_K(A\Sigma_Z A^\top)/\tau^2 \to \infty$. To appreciate why the first term dominates the second term in (77), observe that

$$\begin{aligned}
\sum_{\ell=1}^{K} \frac{\beta_\ell^2}{m} \sum_{i\in I_\ell} \left(\mathbf{e}_k^\top [\Theta^\top\Theta]^{-1}\Theta^\top \mathbf{e}_i\right)^2 &= \sum_{\ell=1}^{K} \left(\mathbf{e}_k^\top [\Theta^\top\Theta]^{-1}\Sigma_Z \mathbf{e}_\ell\right)^2 \beta_\ell^2 \\
&\le \|\beta\|_2^2 \max_\ell \left|\mathbf{e}_k^\top (\Theta^\top\Theta)^{-1}\Sigma_Z \mathbf{e}_\ell\right|^2 \\
&\le \|\beta\|_2^2 \cdot \mathbf{e}_k^\top (\Theta^\top\Theta)^{-1}\mathbf{e}_k \cdot \max_\ell \left|\mathbf{e}_\ell^\top \Sigma_Z (\Theta^\top\Theta)^{-1}\Sigma_Z \mathbf{e}_\ell\right| \\
&= \|\beta\|_2^2 \cdot \mathbf{e}_k^\top (\Theta^\top\Theta)^{-1}\mathbf{e}_k \cdot \max_\ell \left|\mathbf{e}_\ell^\top (A^\top A)^{-1}\mathbf{e}_\ell\right| \\
&\le \|\beta\|_2^2 \cdot \mathbf{e}_k^\top (\Theta^\top\Theta)^{-1}\mathbf{e}_k \cdot \lambda_{\min}^{-1}(A^\top A) \\
&\le \|\beta\|_2^2 \mathbf{e}_k^\top (\Theta^\top\Theta)^{-1}\mathbf{e}_k/m. \qquad (79)
\end{aligned}$$

We used the identity $\Theta = A\Sigma_Z$, $A_{I_\ell \bullet} = |\mathbf{e}_\ell|$ for all $i \in I_\ell$ and $|I_\ell| = m$ in the first line, the Cauchy-Schwarz inequality in the second line, $\Theta = A\Sigma_Z$ in the third line and the inequality $\lambda_{\min}(A^\top A) \ge \lambda_{\min}(A_{I_\bullet}^\top A_{I_\bullet}) \ge m$ in the last line.

Step 3: Next we apply Lyapunov's central limit theorem (see, for instance, [6]) to $\sum_{t=1}^n \xi_{tk}$. The independent $\xi_{1k}, \ldots, \xi_{nk}$ form a triangular array, with variances possibly changing in n, due to the dependence on p and K. We verify Lyapunov's condition

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n \mathbb{E}[|\xi_{ik}|^3]}{(\sum_{j=1}^n \mathbb{E}[\xi_{jk}^2])^{3/2}} = \lim_{n\to\infty} \frac{\sum_{i=1}^n \mathbb{E}[|\xi_{ik}|^3]}{(n\mathbb{E}[\xi_{1k}^2])^{3/2}} = 0 \qquad (80)$$

for the simplified case with $\mathbb{E}[\xi_{1k}^2] := V_k$ in (77) only. The general case using the variance in (76) can be verified in a similar way. Invoke Lemma 16 in Section E.2 to obtain

$$\frac{1}{n\sqrt{n}} \sum_{i=1}^n \mathbb{E}\left[|\xi_{ik}|^3\right] \lesssim \frac{1}{\sqrt{n}} \left\{ \left(\gamma_\varepsilon + \frac{\tau\|\beta\|_2}{\sqrt{m}}\right)^3 [\Sigma_Z^{-1}]_{kk}^{3/2} + \left(\gamma_w \frac{\|\beta\|_2}{\sqrt{m}} + \gamma_\varepsilon\right)^3 \gamma_w^3 \left(\mathbf{e}_k^\top (\Theta^\top\Theta)^{-1}\mathbf{e}_k\right)^{3/2} \right\}.$$

Here γ_ε and γ_w are the sub-Gaussian constants of ε and W. Compare this expression with $V_k = \mathbb{E}[\xi_{1k}^2]$ in (77),

$$V_k^{3/2} \ge \left(\sigma^2 + \tau^2\|\beta\|_2^2/m\right)^{3/2} \left([\Sigma_Z^{-1}]_{kk} + \tau^2 \mathbf{e}_k^\top (\Theta^\top\Theta)^{-1}\mathbf{e}_k\right)^{3/2}$$

and conclude

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n \mathbb{E}[|\xi_{ik}|^3]}{(nV_k)^{3/2}} \lesssim \lim_{n\to\infty} \frac{1}{\sqrt{n}} = 0.$$

The last step used $\gamma_\varepsilon \lesssim \Sigma$ and $\gamma_w \lesssim \tau$. By Lyapunov's central limit theorem, the standardized sum $\sum_{i=1}^n \xi_{ik}/\sqrt{nV_k}$ converges weakly to $N(0,1)$, as $n \to \infty$.

Step 4: Finally, we show that both remainder terms in (72) are $o_p(\sqrt{V_k/n})$. For Rem_1, apply Lemma 13 in Section D.2 and invoke Assumption 3′ to find

$$\begin{aligned}
&\sqrt{n}\,|[\mathrm{Rem}_1]_k| \\
&= O_p\left([\Sigma_Z^{-1}]_{kk}^{1/2}\left\{1+\frac{\|\beta\|_2}{\sqrt{m}}+\bar{\rho}\|\beta\|_2\right\}\frac{\delta_n\sqrt{p\log(p\vee n)}}{\sigma_K(A\Sigma_Z^{1/2})}\left\{\sqrt{K}+\frac{\sqrt{p}}{\sigma_K(A\Sigma_Z^{1/2})}\right\}\right) \\
&= O_p\left([\Sigma_Z^{-1}]_{kk}^{1/2}\left\{1+\frac{\|\beta\|_2}{\sqrt{m}}\right\}\frac{\delta_n\sqrt{p\log(p\vee n)}}{\sigma_K(A\Sigma_Z^{1/2})}\left\{\sqrt{K}+\frac{\sqrt{p}}{\sigma_K(A\Sigma_Z^{1/2})}\right\}\right).
\end{aligned}$$

Invoke Assumption 5′ and write $\lambda_K := \sigma_K^2(A\Sigma_Z^{1/2})$, to get

$$\frac{\delta_n\sqrt{p\log(p\vee n)}}{\sigma_K(A\Sigma_Z^{1/2})}\left\{\sqrt{K}+\frac{\sqrt{p}}{\sigma_K(A\Sigma_Z^{1/2})}\right\} = \sqrt{\frac{K\log^2(p\vee n)}{n}\cdot\frac{p}{\lambda_K}}+\frac{p\log(p\vee n)}{\lambda_K\sqrt{n}} = o(1),$$

as $n\to\infty$, where we also use $\lambda_K \lesssim (p/K)$ from Lemma 14 in conjunction with Assumption 5′ to deduce $K = o(\sqrt{n}/\log(p\vee n))$. This concludes

$$\sqrt{n}\,|[\mathrm{Rem}_1]_k| = o_p\left([\Sigma_Z^{-1}]_{kk}^{1/2}\left(1+\frac{\|\beta\|_2}{\sqrt{m}}\right)\right) = o_p\left(\sqrt{V_k}\right).$$

Finally, invoke Lemma 17 in Section E.2 and Assumption 3′, to obtain

$$\sqrt{n}\,|[\mathrm{Rem}_2]_k| = O_p\left(\bar{\rho}\|\beta\|_2\sqrt{\log(p\vee n)}\sqrt{[\Sigma_Z^{-1}]_{kk}}\right) = o_p\left(\sqrt{V_k}\right).$$

This completes the proof. □

E.2. Lemmas used in the proof of Theorem 4

Let ξ_{tk} be defined in (74) for all $t\in[n]$ and $k\in[K]$. The first lemma states its first and second moments while the second lemma provides upper bounds for its third absolute moment. The third lemma studies the rate of Rem_2 defined in (72). These lemmas are proved in Section E.3.

Lemma 15. *Under Assumption 2, we have* $\mathbb{E}[\xi_{tk}]=0$ *and*

$$\begin{aligned}
\mathbb{E}[\xi_{tk}^2] = &\left(\sigma^2+\sum_{a=1}^K\beta_a^2\bar{\tau}_a^2\right)\left([\Sigma_Z^{-1}]_{kk}+\mathbf{e}_k^\top[\Theta^\top\Theta]^{-1}\Theta^\top\Gamma\Theta[\Theta^\top\Theta]^{-1}\mathbf{e}_k\right) \\
&+\sum_{a=1}^K\frac{\Delta_\tau^{(a)}}{m_a}\sum_{i\in I_a}\left(\mathbf{e}_k^\top[\Theta^\top\Theta]^{-1}\Theta^\top\mathbf{e}_i\right)^2
\end{aligned}$$

for all $t \in [n]$ *and* $k \in [K]$, *with* $\overline{\tau}_a^2 := \sum_{i \in I_a} \tau_i^2/m_a^2$ *and*

$$\Delta_\tau^{(a)} := \beta_a^2 \sum_{i \in I_a} \tau_i^2 \left(\frac{1}{(m_a-1)^2} \sum_{j \in I_a \setminus \{i\}} \tau_j^2 - \frac{1}{m_a^2} \sum_{j \in I_a} \tau_j^2 \right).$$

Lemma 16. *Under the conditions of Theorem 4, we have*

$$\mathbb{E}[|\xi_{tk}|^3] \lesssim \left(\gamma_\varepsilon + \frac{\tau \|\beta\|}{\sqrt{m}} \right)^3 [\Sigma_Z^{-1}]_{kk}^{3/2} + \left(\gamma_w \frac{\|\beta\|}{\sqrt{m}} + \gamma_\varepsilon \right)^3 \gamma_w^3 \left(\mathbf{e}_k^\top [\Theta^\top \Theta]^{-1} \mathbf{e}_k \right)^{3/2}.$$

Lemma 17. *Let* Rem_2 *be defined in (72) on the event* $\mathcal{E}$ *defined in (10). Under the conditions of Theorem 4, we have*

$$\sqrt{n}\, |[\mathrm{Rem}_2]_k| \cdot \mathbf{1}_{\mathcal{E}} = O_p\left(\bar{\rho} \|\beta\|_2 \sqrt{\log(p \vee n)} \sqrt{[\Sigma_Z^{-1}]_{kk}} \right).$$

E.3. Proof of lemmas in Section E.2

E.3.1. Proof of Lemma 15

Fix any $t \in [n]$ and $k \in [K]$. Recall that

$$\xi_{tk} \;:=\; \mathbf{e}_k^\top \Sigma_Z^{-1} \mathbf{Z}_{t\bullet} \left[\varepsilon_t - \mathbf{W}_{t\bullet} \Pi \beta \right] + \mathbf{e}_k^\top \Theta^+ \left[\mathbf{W}_{t\bullet} \varepsilon_t + \left(\widetilde{\Gamma}_{\bullet I}^{(t)} - \mathbf{W}_{t\bullet} \mathbf{W}_{tI} \right) \Pi \beta \right]$$

Since $\Theta = A\Sigma_Z$, we have

$$\Theta^+ \left(\widetilde{\Gamma}_{\bullet I}^{(t)} - \mathbf{W}_{t\bullet} \mathbf{W}_{tI} \right) \Pi \beta = (\Theta^\top \Theta)^{-1} \Sigma_Z A^\top \left(\widetilde{\Gamma}_{\bullet I}^{(t)} - \mathbf{W}_{t\bullet} \mathbf{W}_{tI} \right) \Pi \beta.$$

Write $U^{(t)} := (\widetilde{\Gamma}_{\bullet I}^{(t)} - \mathbf{W}_{t\bullet} \mathbf{W}_{tI}) \Pi \in \mathbb{R}^{p \times K}$, and observe that definition (73) yields

$$\begin{aligned}
U_{ia}^{(t)} &= \mathbf{W}_{ti}^\top \overline{\mathbf{W}}_{ta}, && \forall i \in J,\ a \in [K]; \\
U_{ib}^{(t)} &= \mathbf{W}_{ti}^\top \overline{\mathbf{W}}_{tb}, && \forall i \in I_a,\ a, b \in [K],\ b \neq a; \\
U_{ia}^{(t)} &= \mathbf{W}_{ti}^\top \overline{\mathbf{W}}_{ta} - \frac{1}{m_a} \widetilde{\Gamma}_{ii}^{(t)}, && \forall i \in I_a,\ a \in [K].
\end{aligned}$$

Lemma 18, stated immediately after this proof, gives

$$\frac{1}{m_a} \sum_{i \in I_a} U_{ia}^{(t)} = \overline{\mathbf{W}}_{ta}^\top \overline{\mathbf{W}}_{ta} - \frac{1}{m_a^2} \sum_{i \in I_a} \widetilde{\Gamma}_{ii}^{(t)} = \frac{1}{m_a(m_a-1)} \sum_{j \neq \ell \in I_a} \mathbf{W}_{tj} \mathbf{W}_{t\ell}.$$

Hence, we have

$$A_{J\bullet}^\top U_{J\bullet}^{(t)} = A_{J\bullet}^\top \mathbf{W}_{tJ} \overline{\mathbf{W}}_t \in \mathbb{R}^{K \times K}$$

and

$$\left[A_{I\bullet}^{\top}U_{I\bullet}^{(t)}\right]_{ab} = \frac{1}{m_b}\sum_{j\in I_a,\ell\in I_b}\mathbf{W}_{tj}\mathbf{W}_{t\ell}, \qquad \forall a\neq b,\ a,b\in[K], \tag{81}$$

$$\left[A_{I\bullet}^{\top}U_{I\bullet}^{(t)}\right]_{aa} = \frac{1}{m_a-1}\sum_{j\neq\ell\in I_a}\mathbf{W}_{tj}\mathbf{W}_{t\ell}, \quad \forall a\in[K]. \tag{82}$$

It follows that

$$\begin{aligned}\xi_{tk} = {}& \Omega_{k\bullet}^{\top}\mathbf{Z}_{t\bullet}\varepsilon_t - \Omega_{k\bullet}^{\top}\mathbf{Z}_{t\bullet}\overline{\mathbf{W}}_{t\bullet}^{\top}\beta + \mathbf{e}_k^{\top}\Theta^{+}\mathbf{W}_{t\bullet}\varepsilon_t \\ & - \mathbf{e}_k^{\top}(\Theta^{\top}\Theta)^{-1}\Sigma_Z A_{J\bullet}^{\top}\mathbf{W}_{tJ}\overline{\mathbf{W}}_{t\bullet}^{\top}\beta - \mathbf{e}_k^{\top}(\Theta^{\top}\Theta)^{-1}\Sigma_Z A_{I\bullet}^{\top}U_{I\bullet}^{(t)}\beta.\end{aligned}$$

Since all terms in the above display have zero mean, $\mathbb{E}[\xi_{tk}]=0$. We use Assumption 2 to verify that any two terms are uncorrelated, and we find $\mathbb{E}[\xi_{tk}^2]$ after summing up the second moments for each individual term. We have

$$\begin{aligned}&\mathbb{E}\left[\Omega_{k\bullet}^{\top}\mathbf{Z}_{t\bullet}\varepsilon_t\right]^2 = \sigma^2\Omega_{kk},\\ &\mathbb{E}\left[\Omega_{k\bullet}^{\top}\mathbf{Z}_{t\bullet}\overline{\mathbf{W}}_{t\bullet}^{\top}\beta\right]^2 = \Omega_{kk}\sum_{a=1}^{K}\frac{\beta_a^2}{m_a^2}\sum_{i\in I_a}\tau_i^2 = \Omega_{kk}\sum_{a=1}^{K}\beta_a^2\overline{\tau}_a^2\\ &\mathbb{E}\left[\mathbf{e}_k^{\top}\Theta^{+}W_{t\bullet}\varepsilon_t\right]^2 = \sigma^2\sum_{i=1}^{p}\tau_i^2\left(\mathbf{e}_k^{\top}(\Theta^{\top}\Theta)^{-1}\Theta_{I\bullet}\right)^2\\ &\mathbb{E}\left[\mathbf{e}_k^{\top}(\Theta^{\top}\Theta)^{-1}\Sigma_Z A_{J\bullet}^{\top}\mathbf{W}_{tJ}\overline{\mathbf{W}}_{t\bullet}^{\top}\beta\right]^2 = \sum_{j\in J}\tau_i^2\left(\mathbf{e}_k^{\top}(\Theta^{\top}\Theta)^{-1}\Theta_{i\bullet}\right)^2\sum_{a=1}^{K}\beta_a^2\overline{\tau}_a^2,\end{aligned}$$

by writing $\overline{\tau}_a^2=\sum_{i\in I_a}\tau_i^2/m_a^2$. Finally, we have

$$\begin{aligned}&\mathbb{E}\left[\mathbf{e}_k^{\top}(\Theta^{\top}\Theta)^{-1}\Sigma_Z A_{I\bullet}^{\top}U_{I\bullet}^{(t)}\beta\right]^2\\ &= \sum_{a=1}^{K}\left(\mathbf{e}_k^{\top}(\Theta^{\top}\Theta)^{-1}[\Sigma_Z]_{a\bullet}\right)^2\sum_{b=1}^{K}\beta_b^2\mathbb{E}\left[(A_{I\bullet}^{\top}U_{I\bullet})_{ab}\right]^2\\ &= \sum_{a=1}^{K}\left(\mathbf{e}_k^{\top}(\Theta^{\top}\Theta)^{-1}[\Sigma_Z]_{a\bullet}\right)^2\left[\sum_{b=1}^{K}\frac{\beta_b^2}{m_b^2}\sum_{i\in I_a,}\sum_{j\in I_b}\tau_i^2\tau_j^2+\Delta_\tau^{(a)}\right]\\ &= \sum_{a=1}^{K}\sum_{i\in I_a}\tau_i^2\left(\mathbf{e}_k^{\top}(\Theta^{\top}\Theta)^{-1}[\Sigma_Z]_{a\bullet}\right)^2\sum_{b=1}^{K}\beta_b^2\overline{\tau}_b^2+\sum_{a=1}^{K}\left(\mathbf{e}_k^{\top}(\Theta^{\top}\Theta)^{-1}[\Sigma_Z]_{a\bullet}\right)^2\Delta_\tau^{(a)}\end{aligned}$$

with

$$\begin{aligned}\Delta_\tau^{(a)} &= \frac{\beta_a^2}{(m_a-1)^2}\sum_{i,j\in I_a, i\neq j}\sum \tau_i^2\tau_j^2 - \frac{\beta_a^2}{m_a^2}\sum_{i,j\in I_a}\tau_i^2\tau_j^2 \\ &= \beta_a^2\sum_{i\in I_a}\tau_i^2\left(\frac{1}{(m_a-1)^2}\sum_{j\in I_a\setminus\{i\}}\tau_j^2 - \frac{1}{m_a^2}\sum_{j\in I_a}\tau_j^2\right).\end{aligned}$$

Since $A_{i\bullet} = \mathbf{e}_a$ for any $i \in I_a$ and $\Theta = A\Sigma_Z$, we further find

$$\begin{aligned}&\mathbb{E}\left[\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Sigma_Z A_{I\bullet}^\top U_{I\bullet}^{(t)}\beta\right]^2 \\ &= \sum_{i\in I}\tau_i^2\left(\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Theta_{I\bullet}\right)^2\sum_{b=1}^K\beta_b^2\overline{\tau}_b^2 + \sum_{a=1}^K\frac{\Delta_\tau^{(a)}}{m_a}\sum_{i\in I_a}\left(\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Theta^\top\mathbf{e}_i\right)^2.\end{aligned}$$

Collecting all second moments yields (76) and completes the proof. □

Lemma 18. *Let $\widetilde{\Gamma}_{\bullet I}$ be as defined in (71). For any $a \in [K]$, we have*

$$\frac{1}{n}\overline{\mathbf{W}}_{\bullet a}^\top\overline{\mathbf{W}}_{\bullet a} - \frac{1}{m_a^2}\sum_{i\in I_a}\widetilde{\Gamma}_{ii} = \frac{1}{m_a(m_a-1)}\sum_{j\neq\ell\in I_a}\sum\frac{1}{n}\mathbf{W}_{\bullet j}^\top\mathbf{W}_{\bullet\ell}.$$

Proof of Lemma 18. Fix any $a \in [K]$. Definition (71) yields

$$\begin{aligned}&\frac{1}{m_a^2}\sum_{i\in I_a}\widetilde{\Gamma}_{ii} - \frac{1}{n}\overline{\mathbf{W}}_{\bullet a}^\top\overline{\mathbf{W}}_{\bullet a} \\ &= \frac{1}{m_a^2}\sum_{i\in I_a}\frac{1}{n}\mathbf{X}_{\bullet i}^\top\mathbf{X}_{\bullet i} - \frac{1}{m_a^2(m_a-1)}\sum_{i,j\in I_a, i\neq j}\sum\frac{1}{n}\mathbf{X}_{\bullet i}^\top\mathbf{X}_{\bullet j} - \frac{1}{m_a^2}\sum_{i,j\in I_a}\frac{1}{n}\mathbf{W}_{\bullet i}^\top\mathbf{W}_{\bullet j} \\ &= \frac{1}{m_a^2}\sum_{i\in I_a}\frac{1}{n}\left(\mathbf{Z}_{\bullet a}^\top\mathbf{Z}_{\bullet a} + 2\mathbf{Z}_{\bullet a}^\top\mathbf{W}_{\bullet i} + \mathbf{W}_{\bullet i}^\top\mathbf{W}_{\bullet j}\right) - \frac{1}{m_a^2}\sum_{i\in I_a}\frac{1}{n}\mathbf{W}_{\bullet i}^\top\mathbf{W}_{\bullet i} - \frac{1}{m_a^2}\sum_{i,j\in I_a, i\neq j}\sum\frac{1}{n}\mathbf{W}_{\bullet i}^\top\mathbf{W}_{\bullet j} \\ &\quad - \frac{1}{m_a^2(m_a-1)}\sum_{i,j\in I_a, i\neq j}\sum\frac{1}{n}\left(\mathbf{Z}_{\bullet a}^\top\mathbf{Z}_{\bullet a} + \mathbf{Z}_{\bullet a}^\top\mathbf{W}_{\bullet j} + \mathbf{Z}_{\bullet a}^\top\mathbf{W}_{\bullet i} + \mathbf{W}_{\bullet i}^\top\mathbf{W}_{\bullet j}\right) \\ &= -\frac{1}{m_a(m_a-1)}\sum_{i,j\in I_a, i\neq j}\sum\frac{1}{n}\mathbf{W}_{\bullet i}^\top\mathbf{W}_{\bullet j},\end{aligned}$$

as desired. □

E.3.2. Proof of Lemma 16

By using the inequality $|x+y|^3 \le 4(|x|^3+|y|^3)$ and (74), we obtain

$$\mathbb{E}\left[|\xi_{tk}|^3\right] \le 4\mathbb{E}\left[\left|\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Sigma_Z A_{I\bullet}^\top U_{I\bullet}^{(t)}\beta\right|^3\right] + 4\mathbb{E}\left[\left|\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Sigma_Z A_{J\bullet}^\top \mathbf{W}_{tJ}\overline{\mathbf{W}}_{t\bullet}^\top\beta\right|^3\right] + 4\mathbb{E}\left[\left|\mathbf{e}_k^\top\Theta^+\mathbf{W}_{t\bullet}\varepsilon_t\right|^3\right] + 4\mathbb{E}\left[\left|\Omega_{k\bullet}^\top\mathbf{Z}_{t\cdot}(\varepsilon_t - \langle\overline{\mathbf{W}}_{t\bullet},\beta\rangle\right|^3\right].$$

The upper bound of the first term is established in Lemma 19, stated immediately after this proof. For the second term, by using the inequality $\|XY\|_{\psi_1} \le \|X\|_{\psi_2}\|Y\|_{\psi_2}$[1] for any two sub-Gaussian random variables X and Y, we find

$$\left\|\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Sigma_Z A_{J\bullet}^\top\mathbf{W}_{tJ}\overline{\mathbf{W}}_{t\bullet}^\top\beta\right\|_{\psi_1} \le \left\|\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Sigma_Z A_{J\bullet}^\top\mathbf{W}_{tJ}\right\|_{\psi_2}\left\|\overline{\mathbf{W}}_{t\bullet}^\top\beta\right\|_{\psi_2}.$$

Lemma 5 implies that $\overline{\mathbf{W}}_{t\bullet}^\top\beta$ is $\gamma_w\|\beta\|/\sqrt{m}$-sub-Gaussian, hence $\|\overline{\mathbf{W}}_{t\bullet}^\top\beta\|_{\psi_2} \le c\gamma_w\|\beta\|/\sqrt{m}$ for some constant $c>0$. Similarly,

$$\left\|\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Sigma_Z A_{J\bullet}^\top\mathbf{W}_{tJ}\right\|_{\psi_2} \le c\gamma_w\sqrt{\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Theta_{J\bullet}^\top\Theta_{J\bullet}(\Theta^\top\Theta)^{-1}\mathbf{e}_k} \le c\gamma_w\sqrt{\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k}$$

and

$$\left\|\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Sigma_Z A_{J\bullet}^\top\mathbf{W}_{tJ}\overline{\mathbf{W}}_{t\bullet}^\top\beta\right\|_{\psi_1} \le c\gamma_w^2\frac{\|\beta\|}{\sqrt{m}}\sqrt{\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k}.$$

This inequality and the definition of $\|\cdot\|_{\psi_1}$ imply that

$$\mathbb{E}\left[\left|\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Sigma_Z A_{J\bullet}^\top W_{tJ}\overline{\mathbf{W}}_{t\bullet}^\top\beta\right|^3\right] \le c\gamma_w^6\left(\frac{\|\beta\|^2}{m}\right)^{3/2}\left(\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k\right)^{3/2}. \tag{83}$$

Similarly, observe that $\mathbf{e}_k^\top\Theta^+W_{t\bullet}$ is $\gamma_w\sqrt{e_k^\top(\Theta^\top\Theta)^{-1}e_k}$-sub-Gaussian by Lemma 5, and deduce

$$\left\|\mathbf{e}_k^\top\Theta^+\mathbf{W}_{t\bullet}\varepsilon_t\right\|_{\psi_1} \le \left\|\mathbf{e}_k^\top\Theta^+\mathbf{W}_{t\bullet}\right\|_{\psi_2}\|\varepsilon_t\|_{\psi_2} \le c\gamma_\varepsilon\gamma_w\left(\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k\right)^{3/2}$$

so that

$$\mathbb{E}\left[\left|\mathbf{e}_k^\top\Theta^+\mathbf{W}_{t\bullet}\varepsilon_t\right|^3\right] \le c\gamma_\varepsilon^3\gamma_w^3\left(\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k\right)^{3/2}. \tag{84}$$

Finally, we bound the third term. The independence of Z, W and ε guarantees

$$\mathbb{E}\left[\left|\Omega_{k\bullet}^\top\mathbf{Z}_{t\bullet}(\varepsilon_t - \langle\overline{\mathbf{W}}_{t\bullet},\beta\rangle\right|^3\right] \le \mathbb{E}\left[\left|\Omega_{k\bullet}^\top\mathbf{Z}_{t\bullet}\right|^3\right]\mathbb{E}\left[\left|\varepsilon_t - \langle\overline{\mathbf{W}}_{t\bullet},\beta\rangle\right|^3\right].$$

Observe that $\|\varepsilon_t - \langle\overline{\mathbf{W}}_{t\bullet},\beta\rangle\|_{\psi_2} \le c(\gamma_\varepsilon + \|\beta\|\gamma_w/\sqrt{m})$ and part (1) of Lemma 4 imply that $\langle\Omega_{k\bullet},\mathbf{Z}_{t\bullet}\rangle$ is $(\gamma_z\sqrt{\Omega_{kk}})$-sub-Gaussian. The definition of the Orlicz ψ_2 norm implies

$$\mathbb{E}\left[\left|\Omega_{k\bullet}^\top\mathbf{Z}_{t\bullet}\right|^3\right]\mathbb{E}\left[\left|\varepsilon_t - \langle\overline{\mathbf{W}}_{t\bullet},\beta\rangle\right|^3\right] \le c\left(\gamma_\varepsilon + \frac{\tau\|\beta\|}{\sqrt{m}}\right)^3\Omega_{kk}^{3/2}. \tag{85}$$

[1]For any random variable X, we write $\|X\|_{\psi_1} = \sup_{p\ge1}p^{-1}(\mathbb{E}[|X|^p])^{1/p}$ and $\|X\|_{\psi_2} = \sup_{p\ge1}p^{-1/2}(\mathbb{E}[|X|^p])^{1/p}$.

Collecting (83) – (85) and invoking Lemma 19 concludes the proof. □

Lemma 19. *Under conditions of Theorem 4, we have*

$$\mathbb{E}\left[\left|\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Sigma_Z A_{I\bullet}^\top U_{I\bullet}^{(t)}\beta\right|^3\right] \lesssim \gamma_w^6\left(\frac{\|\beta\|_2^2}{m}\right)^{3/2}\left(\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k\right)^{3/2}$$

for any $1 \le t \le n$.

Proof. From (81), we have

$$\begin{aligned}
&\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Sigma_Z A_{I\bullet}^\top U_{I\bullet}^{(t)}\beta \\
&= \sum_{a,b}\sum \mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}[\Sigma_Z]_{a\bullet}\beta_b[A_{I\bullet}^\top U_{I\bullet}^{(t)}]_{ab} \\
&= \sum_{a=1}^K \mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}[\Sigma_Z]_{a\bullet}\left(\frac{\beta_a}{m_a-1}\sum\sum_{j\neq\ell\in I_a}\mathbf{W}_{tj}\mathbf{W}_{t\ell} + \sum\sum_{b\neq a}\frac{\beta_b}{m_b}\sum\sum_{j\in I_a,\ell\in I_b}\mathbf{W}_{tj}\mathbf{W}_{t\ell}\right) \\
&= \sum_{a=1}^K \mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}[\Sigma_Z]_{a\bullet}\left(\frac{\beta_a}{m_a-1}\sum\sum_{j\neq\ell\in I_a}\mathbf{W}_{tj}\mathbf{W}_{t\ell} - \frac{\beta_a}{m_a}\sum\sum_{j,\ell\in I_a}\mathbf{W}_{tj}\mathbf{W}_{t\ell} + \sum_{b=1}^K\beta_b\sum_{j\in I_a}\mathbf{W}_{tj}\overline{\mathbf{W}}_{tb}\right) \\
&= \sum_{a=1}^K \mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}[\Sigma_Z]_{a\bullet}\beta_a\left(\frac{1}{m_a(m_a-1)}\sum\sum_{j\neq\ell\in I_a}\mathbf{W}_{tj}\mathbf{W}_{t\ell} - \frac{1}{m_a}\sum_{j\in I_a}\mathbf{W}_{tj}^2\right) \\
&\quad + \sum_{a=1}^K \mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}[\Sigma_Z]_{a\bullet}\sum_{j\in I_a}\mathbf{W}_{tj}\beta^\top\overline{\mathbf{W}}_{t\cdot}.
\end{aligned}$$

By using

$$\sum_{a=1}^K \mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}[\Sigma_Z]_{a\bullet}\sum_{j\in I_a}\mathbf{W}_{tj} = \mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Sigma_Z A_{I\bullet}^\top\mathbf{W}_{tI} = \mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Theta_{I\bullet}^\top\mathbf{W}_{tI},$$

after a bit algebra, we obtain

$$\begin{aligned}
&\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Sigma_Z A_{I\bullet}^\top U_{I\bullet}^{(t)}\beta \\
&= \sum_{a=1}^K \mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}[\Sigma_Z]_{a\bullet}\beta_a\left(\frac{1}{m_a(m_a-1)}\sum_{j,\ell\in I_a}\mathbf{W}_{tj}\mathbf{W}_{t\ell} - \frac{m_a-2}{m_a(m_a-1)}\sum_{j\in I_a}\mathbf{W}_{tj}^2\right) \\
&\quad + \mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Theta_{I\bullet}^\top\mathbf{W}_{tI}\beta^\top\overline{\mathbf{W}}_{t\cdot}
\end{aligned}$$

such that

$$\begin{aligned}&\left|\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Sigma_Z A_{I\bullet}^\top U_{I\bullet}^{(t)}\beta\right| \qquad (86)\\ &\le \left|\sum_{a=1}^K \mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}[\Sigma_Z]_{a\bullet}\beta_a\frac{m_a}{m_a-1}\overline{\mathbf{W}}_{ta}^2\right| + \left|\sum_{a=1}^K \mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}[\Sigma_Z]_{a\bullet}\beta_a\frac{m_a-2}{m_a(m_a-1)}\sum_{j\in I_a}\mathbf{W}_{tj}^2\right|\\ &\quad + |\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Theta_{I\bullet}^\top\mathbf{W}_{tI}\beta^\top\overline{\mathbf{W}}_{t\bullet}|.\end{aligned}$$

We now bound the third moment of each term on the right. For the last term, using the inequality $\|XY\|_{\psi_1}\le\|X\|_{\psi_2}\|Y\|_{\psi_2}$ yields

$$\left\|\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Theta_{I\bullet}^\top\mathbf{W}_{tI}\beta^\top\overline{\mathbf{W}}_{t\cdot}\right\|_{\psi_1}\le\left\|\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Theta_{I\bullet}^\top\mathbf{W}_{tI}\right\|_{\psi_2}\left\|\beta^\top\overline{\mathbf{W}}_{t\cdot}\right\|_{\psi_2}.$$

Recall that, by Lemma 5,

$$\left\|\beta^\top\overline{\mathbf{W}}_{t\cdot}\right\|_{\psi_2}\le c\frac{\gamma_w\|\beta\|_2}{\sqrt{m}},\quad \left\|\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Theta_{I\bullet}^\top\mathbf{W}_{tI}\right\|_{\psi_2}\le c\gamma_w\sqrt{\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k}$$

after using $\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Theta_{I\bullet}^\top\Theta_{I\bullet}(\Theta^\top\Theta)^{-1}\mathbf{e}_k\le\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k$. The definition of the Orlicz ψ_1-norm gives

$$\mathbb{E}\left[\left|\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Theta_{I\bullet}^\top\mathbf{W}_{tI}\beta^\top\overline{\mathbf{W}}_{t\cdot}\right|^3\right]\le c\gamma_w^6\left(\frac{\|\beta\|_2^2}{m}\right)^{3/2}\left(\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k\right)^{3/2} \qquad (87)$$

It remains to study the first two terms in (86). By using the independence between $\overline{\mathbf{W}}_{ta}$, $a\in[K]$ and $\mathbf{W}_{tj}$, $j\in I$, and by further writing, for each $a\in[K]$,

$$\begin{aligned}\alpha_a&=\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}[\Sigma_Z]_{a\bullet}\beta_a\frac{m_a}{m_a-1},\\ \widetilde{\alpha}_a&=\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}[\Sigma_Z]_{a\bullet}\beta_a\frac{m_a-2}{m_a(m_a-1)},\end{aligned}$$

two applications of Rosenthal's inequality [3] give

$$\begin{aligned}\mathbb{E}\left[\left|\sum_{a=1}^K\alpha_a\overline{\mathbf{W}}_{ta}^2\right|^3\right]&\le c'\left\{\sum_{a=1}^K|\alpha_a|^3\mathbb{E}[\overline{\mathbf{W}}_{ta}^6]+\left(\sum_{a=1}^K\alpha_a^2\mathbb{E}[\overline{\mathbf{W}}_{ta}^4]\right)^{3/2}\right\},\\ \mathbb{E}\left[\left|\sum_{a=1}^K\sum_{j\in I_a}\widetilde{\alpha}_a\mathbf{W}_{tj}^2\right|^3\right]&\le c'\left\{\sum_{a=1}^K\sum_{j\in I_a}|\widetilde{\alpha}_a|^3\mathbb{E}[\mathbf{W}_{tj}^6]+\left(\sum_{a=1}^K\sum_{j\in I_a}\widetilde{\alpha}_a^2\mathbb{E}[\mathbf{W}_{tj}^4]\right)^{3/2}\right\}\end{aligned}$$

for some absolute constant $c'>0$. The definition of the Orlicz ψ_2-norm and the inequalities $\|\overline{\mathbf{W}}_{ta}\|_{\psi_2}\le c\gamma_w/\sqrt{m}$ and $\|\mathbf{W}_{tj}\|_{\psi_2}\le c\gamma_w$ from Lemma 5 imply

$$\mathbb{E}[\overline{\mathbf{W}}_{ta}^6]\le c\gamma_w^6/m^3,\quad \mathbb{E}[\overline{\mathbf{W}}_{ta}^4]\le c\gamma_w^4/m^2,\quad \mathbb{E}[\mathbf{W}_{tj}^6]\le c\gamma_w^6,\quad \mathbb{E}[\mathbf{W}_{tj}^4]\le c\gamma_w^4.$$

Furthermore, we have

$$\begin{aligned}\sum_{a=1}^{K}\left(\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}[\Sigma_Z]_{a\bullet}\right)^2 &\leq \sum_{a=1}^{K}\left(\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}[\Sigma_Z]_{a\bullet}m_a[\Sigma_Z]_{a\bullet}^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k\right)^2\\ &= \mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Sigma_Z A_{I\bullet}^\top A_{I\bullet}\Sigma_Z(\Theta^\top\Theta)^{-1}\mathbf{e}_k\\ &= \mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\Theta_{I\bullet}^\top\Theta_{I\bullet}(\Theta^\top\Theta)^{-1}\mathbf{e}_k\\ &\leq \mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k, \end{aligned} \tag{88}$$

and we obtain

$$\begin{aligned}\sum_{a=1}^{K}|\alpha_a|^3\mathbb{E}[\overline{\mathbf{W}}_{ta}^6] &\leq c\frac{\gamma_w^6}{m^3}\left(\sum_{a=1}^{K}|\alpha_a|\right)^3\\ &\leq c\frac{\gamma_w^6}{m^3}\left(\sum_{a=1}^{K}|\beta_a|\cdot\left|\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}[\Sigma_Z]_{a\bullet}\right|\right)^3\\ &\leq c\gamma_w^6\left(\frac{\|\beta\|_2^2}{m}\right)^{3/2}\left(\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k\right)^{3/2}\end{aligned}$$

where we use $m \geq 2$, the Cauchy-Schwarz inequality and (88) in the last line. Moreover, by the same reasoning, we find

$$\sum_{a=1}^{K}\alpha_a^2\mathbb{E}[\overline{\mathbf{W}}_{ta}^4] \leq c\frac{\gamma_w^4}{m^2}\sum_{a=1}^{K}\beta_a^2\left(\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}[\Sigma_Z]_{a\bullet}\right)^2 \leq c\gamma_w^4\frac{\|\beta\|_2^2}{m}\mathbf{e}_k^\top[\Theta^\top\Theta]^{-1}\mathbf{e}_k.$$

Combine the previous displays to conclude that

$$\mathbb{E}\left[\left|\sum_{a=1}^{K}\alpha_a\overline{\mathbf{W}}_{ta}^2\right|^3\right] \leq c'c\gamma_w^6\left(\frac{\|\beta\|_2^2}{m}\right)^{3/2}\left(\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k\right)^{3/2}. \tag{89}$$

By similar arguments, it is easy to show that

$$\begin{aligned}\sum_{a=1}^{K}\sum_{j\in I_a}|\widetilde{\alpha}_a|^3\mathbb{E}[\mathbf{W}_{tj}^6] &\leq c\gamma_w^6\left(\sum_{a=1}^{K}|\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}[\Sigma_Z]_{a\bullet}\beta_a|\right)^3\\ &\leq c\gamma_w^6\left(\sum_{a=1}^{K}\frac{|\beta_a|}{\sqrt{m}_a}\cdot|\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}[\Sigma_Z]_{a\bullet}\sqrt{m_a}|\right)^3\\ &\leq c\gamma_w^6\left(\frac{\|\beta\|_2^2}{m}\right)^{3/2}\left(\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k\right)^{3/2}\end{aligned}$$

and

$$\sum_{a=1}^{K}\sum_{j\in I_a}\widetilde{\alpha}_a^2\mathbb{E}[\mathbf{W}_{tj}^4] \leq c\gamma_w^4\sum_{a=1}^{K}\beta_a^2\left(\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}[\Sigma_Z]_{a\bullet}\right)^2 \leq c\gamma_w^4\frac{\|\beta\|_2^2}{m}\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k.$$

Consequently,

$$\mathbb{E}\left[\left|\sum_{a=1}^{K}\sum_{j\in I_a}\widetilde{\alpha}_a \mathbf{W}_{tj}^2\right|^3\right] \le c\gamma_w^6\left(\frac{\|\beta\|_2^2}{m}\right)^{3/2}\left(\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k\right)^{3/2}. \tag{90}$$

Combination of the bounds (87), (89) and (90) concludes the proof. □

E.3.3. Proof of Lemma 17

We work on the event $\mathcal{E}$ defined in (10) that has probability at least $1-(p\vee n)^{-c}$ for some $c>0$. Recall that it implies $\widehat{K}=K$ and $I_k\subseteq\widehat{I}_k\subseteq I_k\cup J_1^k$ for all $k\in[K]$.

Fix any $k\in K$. Definition (72) yields

$$\begin{aligned}|[\mathrm{Rem}_2]_k| \le{}& \frac{1}{n}\left|\mathbf{e}_k^\top\Omega\mathbf{Z}^\top\mathbf{Z}\Delta\beta\right| + \frac{1}{n}\left|\mathbf{e}_k^\top\Theta^+\mathbf{W}^\top\mathbf{Z}\Delta\beta\right| + \frac{1}{n}\left|\mathbf{e}_k^\top\Omega\mathbf{Z}^\top(\widetilde{\mathbf{W}}-\overline{\mathbf{W}})\beta\right| \\ &+ \left|\mathbf{e}_k^\top\Theta^+\left(\widehat{\Gamma}_{\cdot\widehat{I}} - \widetilde{\Gamma}_{\cdot\widehat{I}}\Pi - \frac{1}{n}\mathbf{W}^\top(\widetilde{\mathbf{W}}-\bar{\mathbf{W}})\right)\beta\right|.\end{aligned}$$

First, recall $\Theta^+=\Omega^{1/2}H^+$, and observe that Lemma 7 with $u=\Omega^{1/2}\mathbf{e}_k$ and Lemma 9 imply

$$\frac{1}{n}\left|\mathbf{e}_k^\top\Omega\mathbf{Z}^\top\mathbf{Z}\Delta\beta\right| + \frac{1}{n}\left|\mathbf{e}_k^\top\Theta^+\mathbf{W}^\top\mathbf{Z}\Delta\beta\right| \lesssim \left(\sqrt{\Omega_{kk}} + \sqrt{\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k}\right)\|D_\rho\beta\|_1\delta_n$$

with probability tending to one. Next, we bound $n^{-1}|\mathbf{e}_k^\top\Omega\mathbf{Z}^\top(\widetilde{\mathbf{W}}-\overline{\mathbf{W}})\beta|$. From the identity (36) for $\widetilde{\mathbf{W}}-\overline{\mathbf{W}}$, we need to bound

$$\begin{aligned}&\frac{1}{n}\left|\mathbf{e}_k^\top\Omega\mathbf{Z}^\top\mathbf{W}_{\cdot L}\widehat{A}_{L\cdot}D_{\widehat{m}}\beta\right| + \frac{1}{n}\left|\mathbf{e}_k^\top\Omega\mathbf{Z}^\top\overline{\mathbf{W}}D_\rho\beta\right| \\ &\le \max_{i\in J_1}\frac{1}{n}\left|\mathbf{e}_k^\top\Omega\mathbf{Z}^\top\mathbf{W}_{\cdot i}\right|\cdot\|\widehat{A}_{L\cdot}D_{\widehat{m}}\beta\|_1 + \frac{1}{n}\left\|\mathbf{e}_k^\top\Omega\mathbf{Z}^\top\overline{\mathbf{W}}\right\|_\infty\|D_\rho\beta\|_1 \\ &\le \bar{\rho}\|\beta\|_2\left(\max_{i\in J_1}\frac{1}{n}\left|\mathbf{e}_k^\top\Omega\mathbf{Z}^\top\mathbf{W}_{\cdot i}\right| + \frac{1}{n}\left\|\mathbf{e}_k^\top\Omega\mathbf{Z}^\top\overline{\mathbf{W}}\right\|_\infty\right)\end{aligned}$$

The last inequality uses (38) with $v=\beta$. Observe that $\mathbf{e}_k^\top\Omega\mathbf{Z}_{t\cdot}$ is $\gamma_z\sqrt{\Omega_{kk}}$-sub-Gaussian by Lemma 4, and apply Lemmas 5 and 6 together with a union bound to conclude

$$\max_{i\in J_1}\frac{1}{n}\left|\mathbf{e}_k^\top\Omega\mathbf{Z}^\top\mathbf{W}_{\cdot i}\right| + \frac{1}{n}\left\|\mathbf{e}_k^\top\Omega\mathbf{Z}^\top\overline{\mathbf{W}}\right\|_\infty \lesssim \sqrt{\Omega_{kk}}\delta_n.$$

Finally, from using identity (36) again, we find that

$$\begin{aligned}
&\left|\mathbf{e}_k^\top \Theta^+ \left(\widehat{\Gamma}_{\bullet \widehat{I}} - \widetilde{\Gamma}_{\bullet \widehat{I}}\Pi - n^{-1}\mathbf{W}^\top(\widetilde{\mathbf{W}} - \bar{\mathbf{W}})\right)\beta\right| \\
&= \left|\mathbf{e}_k^\top \Theta^+ \left(\widehat{\Gamma}_{\bullet \widehat{I}} - \widetilde{\Gamma}_{\bullet \widehat{I}}\Pi - \frac{1}{n}\mathbf{W}^\top \mathbf{W}_{\bullet L}\widehat{A}_{L\bullet} D_{\widehat{m}} + \frac{1}{n}\mathbf{W}^\top \overline{\mathbf{W}} D_\rho\right)\beta\right| \\
&\le \left|\mathbf{e}_k^\top \Theta^+ \left(\widehat{\Gamma}_{\bullet \widehat{I}} - \widetilde{\Gamma}_{\bullet \widehat{I}}\Pi - \Gamma_{\bullet L}\widehat{A}_{L\bullet} D_{\widehat{m}} + \Gamma_{\bullet I}\Pi D_\rho\right)\beta\right| \\
&\quad + \left|\mathbf{e}_k^\top \Theta^+ \left(\frac{1}{n}\mathbf{W}^\top \overline{\mathbf{W}} - \Gamma_{\bullet I}\Pi\right) D_\rho \beta\right| + \left|\mathbf{e}_k^\top \Theta^+ \left(\frac{1}{n}\mathbf{W}^\top \mathbf{W}_{\bullet L} - \Gamma_{\bullet L}\right)\widehat{A}_{L\bullet} D_{\widehat{m}}\beta\right|.
\end{aligned}$$

Close inspection of the proof of (29), by taking $\alpha = \mathbf{e}_k^\top \Theta^+$ and $v = \beta$, we can bound the last two terms above by

$$\bar{\rho}\|\beta\|_2 \delta_n \sqrt{\mathbf{e}_k^\top (\Theta^\top \Theta)^{-1} \mathbf{e}_k}$$

with probability $1 - (p \vee n)^{-c}$. It remains to bound $|\mathbf{e}_k^\top \Theta^+ \Delta' \beta|$ with

$$\Delta' = \widehat{\Gamma}_{\bullet \widehat{I}} - \widetilde{\Gamma}_{\bullet \widehat{I}}\Pi - \Gamma_{\bullet L}\widehat{A}_{L\bullet} D_{\widehat{m}} + \Gamma_{\bullet I}\Pi D_\rho.$$

Observe that

$$\begin{aligned}
\Delta'_{ik} &= \frac{\widehat{\Gamma}_{ii} - \Gamma_{ii}}{\widehat{m}_k}, \quad i \in L_k \\
\Delta'_{ik} &= \frac{|L_k|}{m_k \widehat{m}_k}\left(\Gamma_{ii} - \widehat{\Gamma}_{ii}\right) + \frac{\widehat{\Gamma}_{ii} - \widetilde{\Gamma}_{ii}}{m_k}, \quad i \in I_k \\
\Delta'_{ik} &= 0, \quad \text{otherwise.}
\end{aligned}$$

On the event $\mathcal{E}_W$ defined in (46), we find

$$\begin{aligned}
|\mathbf{e}_k^\top \Theta^+ \Delta' \beta| &\le \left|\sum_{a=1}^K \sum_{i \in I_a \cup L_a} \mathbf{e}_k^\top (\Theta^\top \Theta)^{-1} \Theta_{i\bullet} \Delta'_{ia} \beta_a\right| \\
&\le \max_{i \in I \cup J_1} \left|\mathbf{e}_k^\top (\Theta^\top \Theta)^{-1} \Theta_{i\bullet}\right| \cdot \sum_{a=1}^K \left(\sum_{i \in L_a} |\Delta'_{ia}\beta_a| + \sum_{i \in I_a} |\Delta'_{ia}\beta_a|\right) \\
&\lesssim \sqrt{\mathbf{e}_k^\top (\Theta^\top \Theta)^{-1} \mathbf{e}_k} \left(\|D_\rho \beta\|_1 \delta_n + \sum_{a=1}^K |\beta_a| \max_{i \in I_a} |\widehat{\Gamma}_{ii} - \widetilde{\Gamma}_{ii}|\right).
\end{aligned}$$

Since Lemma 20 gives

$$\max_{i \in I_a} |\widehat{\Gamma}_{ii} - \widetilde{\Gamma}_{ii}| = O_p\left(\frac{|J_1^a|}{|J_1^a| + m_a}\delta_n\right),$$

for all $a \in [K]$, we conclude

$$|\mathbf{e}_k^\top \Theta^+ \Delta' \beta| = O_p\left(\bar{\rho}\|\beta\|_2 \delta_n \sqrt{\mathbf{e}_k^\top (\Theta^\top \Theta)^{-1} \mathbf{e}_k}\right).$$

Finally, we complete the proof by collecting all bounds, using the bound (78) for $\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k$ and the inequality $\lambda_K(H^\top H) \geq \lambda_K(A_{I\bullet}^\top A_{I\bullet})C_{\min} \gtrsim 1$. □

Lemma 20. *Let $\widehat{\Gamma}_{ii}$ and $\widetilde{\Gamma}_{ii}$ be defined in (11) and (71), respectively. Under the conditions of Theorem 4, we have*

$$\max_{i\in I_a} |\widehat{\Gamma}_{ii} - \widetilde{\Gamma}_{ii}| = O_p\left(\frac{|J_1^a|}{|J_1^a| + m_a}\delta_n\right), \quad \text{for all } a \in [K].$$

Proof. We work on the event $\mathcal{E}$ defined in (10) that has probability at least $1-(p\vee n)^{-c}$ for some $c > 0$. Recall that it implies $\widehat{K} = K$ and $I_k \subseteq \widehat{I}_k \subseteq I_k \cup J_1^k$ for all $k \in [K]$.

First recall the definitions of $\widehat{\Sigma}_Z$ and $\bar{\Sigma}^z$. For any $a, b \in [K]$,

$$\left[\widehat{\Sigma}_Z\right]_{aa} = \frac{1}{\widehat{m}_a(\widehat{m}_a-1)} \sum_{i,j\in\widehat{I}_a,\ i\neq j} \frac{1}{n}\mathbf{X}_{\bullet i}^\top \mathbf{X}_{\bullet j}, \quad \left[\widehat{\Sigma}_Z\right]_{ab} = \frac{1}{\widehat{m}_a\widehat{m}_b} \sum_{i\in\widehat{I}_a}\sum_{j\in\widehat{I}_b} \frac{1}{n}\mathbf{X}_{\bullet i}^\top \mathbf{X}_{\bullet j}, \quad (91)$$

$$\left[\bar{\Sigma}_Z\right]_{aa} = \frac{1}{m_a(m_a-1)} \sum_{i,j\in I_a,\ i\neq j} \frac{1}{n}\mathbf{X}_{\bullet i}^\top \mathbf{X}_{\bullet j}, \quad \left[\bar{\Sigma}_Z\right]_{ab} = \frac{1}{m_a m_b} \sum_{i\in I_a}\sum_{j\in I_b} \frac{1}{n}\mathbf{X}_{\bullet i}^\top \mathbf{X}_{\bullet j} \quad (92)$$

Choose any $a \in [K]$ and $i \in I_a$. Since $\widehat{\Gamma}_{ii} - \widetilde{\Gamma}_{ii} = [\widehat{\Sigma}_Z]_{aa} - [\bar{\Sigma}_Z]_{aa}$, we have

$$\begin{aligned}
&\left[\widehat{\Sigma}_Z\right]_{aa} - \left[\bar{\Sigma}_Z\right]_{aa} = \\
&\frac{1}{\widehat{m}_a(\widehat{m}_a-1)} \sum_{i,j\in\widehat{I}_a,\ i\neq j} \frac{1}{n}\left(A_{i\bullet}^\top \mathbf{Z}^\top \mathbf{Z} A_{j\bullet} + A_{i\bullet}^\top \mathbf{Z}^\top \mathbf{W}_{\bullet j} + \mathbf{W}_{\bullet i}^\top \mathbf{Z} A_{j\bullet} + \mathbf{W}_{\bullet i}^\top \mathbf{W}_{\bullet j}\right) \\
&\quad - \frac{1}{m_a(m_a-1)} \sum_{i,j\in I_a,\ i\neq j} \frac{1}{n}\left(A_{i\bullet}^\top \mathbf{Z}^\top \mathbf{Z} A_{j\bullet} + A_{i\bullet}^\top \mathbf{Z}^\top \mathbf{W}_{\bullet j} + \mathbf{W}_{\bullet i}^\top \mathbf{Z} A_{j\bullet} + \mathbf{W}_{\bullet i}^\top \mathbf{W}_{\bullet j}\right).
\end{aligned}$$

Use $\mathbf{Z}A_{i\bullet} = \mathbf{Z}_{\bullet a}$ for all $i \in I_a$ and $\widehat{m}_a = m_a + |L_a|$, to argue that

$$\begin{aligned}
&\sum_{i,j\in\widehat{I}_a,\ i\neq j} \frac{1}{n} A_{i\bullet}^\top \mathbf{Z}^\top \mathbf{Z} A_{j\bullet} - \sum_{i,j\in I_a,\ i\neq j} \frac{1}{n} A_{i\bullet}^\top \mathbf{Z}^\top \mathbf{Z} A_{j\bullet} \\
&= \left(2m_a \sum_{j\in L_a} + \sum_{i,j\in L_a,\ i\neq j}\right)\left\{\frac{1}{n} A_{i\bullet}^\top \mathbf{Z}^\top \mathbf{Z} A_{j\bullet} - [\widehat{m}_a(\widehat{m}_a-1) - m_a(m_a-1)]\frac{1}{n}\mathbf{Z}_{\bullet a}^\top \mathbf{Z}_{\bullet a}\right\} \\
&= \left(2m_a \sum_{j\in L_a} + \sum_{i,j\in L_a,\ i\neq j}\right)\left\{\frac{1}{n}(A_{i\bullet}^\top \mathbf{Z}^\top \mathbf{Z} A_{j\bullet} - \mathbf{Z}_{\bullet a}^\top \mathbf{Z}_{\bullet a})\right\} \\
&\leq 2\left\{2m_a|L_a| + |L_a|(|L_a|-1)\right\} \sum_{i\in I_a\cup J_1^a} \left|\frac{1}{n}\mathbf{Z}_{\bullet a}^\top \mathbf{Z}(A_{i\bullet} - \mathbf{e}_a)\right|.
\end{aligned}$$

This implies

$$\begin{aligned}
&\left|\frac{1}{\widehat{m}_a(\widehat{m}_a-1)}\sum_{i,j\in\widehat{I}_a,\ i\neq j}\frac{1}{n}A_{i\bullet}^\top\mathbf{Z}^\top\mathbf{Z}A_{j\bullet}-\frac{1}{m_a(m_a-1)}\sum_{i,j\in I_a,i\neq j}\frac{1}{n}A_{i\bullet}^\top\mathbf{Z}^\top\mathbf{Z}A_{j\bullet}\right| \\
&\leq \frac{2|J_1^a|}{|J_1^a|+m_a}\sum_{i\in I_a\cup J_1^a}\left|\frac{1}{n}\mathbf{Z}_{\bullet a}^\top\mathbf{Z}(A_{i\bullet}-\mathbf{e}_a)\right| \\
&\leq \frac{2|J_1^a|}{|J_1^a|+m_a}\frac{8B_z}{\nu}\delta_n
\end{aligned}$$

The last line follows from inequality (115), and holds with probability $1-C((p\vee n))^{-c}$. For the other two terms, after expanding $\widehat{I}_a = I_a \cup L_a$ to $I_a \cup J_1^a$, we find that

$$\begin{aligned}
&\left|\frac{1}{\widehat{m}_a(\widehat{m}_a-1)}\sum_{i,j\in\widehat{I}_a,\ i\neq j}\frac{1}{n}A_{i\bullet}^\top\mathbf{Z}^\top\mathbf{W}_{\bullet j}-\frac{1}{m_a(m_a-1)}\sum_{i,j\in I_a,\ i\neq j}\frac{1}{n}A_{i\bullet}^\top\mathbf{Z}^\top\mathbf{W}_{\bullet j}\right| \\
&\lesssim \frac{|J_1^a|}{|J_1^a|+m_a}\max_{i,j\in I_a\cup J_1^a,i\neq j}\left|\frac{1}{n}A_{i\bullet}^\top\mathbf{Z}_{\bullet a}^\top\mathbf{W}_{\bullet j}\right|
\end{aligned}$$

and

$$\begin{aligned}
&\left|\frac{1}{\widehat{m}_a(\widehat{m}_a-1)}\sum_{i,j\in\widehat{I}_a,\ i\neq j}\frac{1}{n}\mathbf{W}_{\bullet i}^\top\mathbf{W}_{\bullet j}-\frac{1}{m_a(m_a-1)}\sum_{i,j\in I_a,\ i\neq j}\frac{1}{n}\mathbf{W}_{\bullet i}^\top\mathbf{W}_{\bullet j}\right| \\
&\lesssim \frac{|J_1^a|}{|J_1^a|+m_a}\max_{i,j\in I_a\cup J_1^a,i\neq j}\left|\frac{1}{n}\mathbf{W}_{\bullet i}^\top\mathbf{W}_{\bullet j}\right|.
\end{aligned}$$

Apply the exponential inequality Lemma 6 and use the fact that $\|A_{i\bullet}\|_1 \leq 1$ to arrive at the desired result. □

Appendix F: Proof of Proposition 5: consistent estimation of the asymptotic variance V_k

F.1. Main proof of Proposition 5

We only need to show $\widehat{V}_k^{1/2}/V_k^{1/2} = 1 + o_p(1)$ as $n \to \infty$ since the rest of the proof is a direct consequence of Theorem 4 and Slutsky's lemma. For ease of presentation, we only prove the result for simplified V_k in (77) when $|I_1| = \cdots = |I_K| = m$ and $\tau_1^2 = \cdots = \tau_p^2 = \tau^2$. The general case with the complicated expression for V_k in (76 can be proved by similar arguments.

We work on the event $\mathcal{E}$ defined in (10) intersected with $\widehat{\Theta}^\top\widehat{\Theta}$ and $\widehat{\Sigma}_Z$ are invertible. This event holds with probability $1-c(p\vee n)^{-c'}$ by Lemma 10 and Lemma 23. Also recall that $\mathcal{E}$ implies $\widehat{K} = K$ and $I_k \subseteq \widehat{I}_k \subseteq I_k \cup J_1^k$ for all $k \in [K]$.

Write $\Omega := \Sigma_Z^{-1}$ and $\widehat{\Omega} = \widehat{\Sigma}_Z^{-1}$. By a Taylor expansion, we have

$$\left|\widehat{V}_k^{1/2} V_k^{-1/2} - 1\right| = V_k^{-1} \left|\widehat{V}_k - V_k\right| (1 + o_p(1)),$$

and it suffices to show $V_k^{-1}|\widehat{V}_k - V_k| = o_p(1)$. To this end, we first recall

$$V_k = \underbrace{\left(\sigma^2 + \frac{\tau^2 \|\beta\|_2^2}{m}\right)}_{\Delta_a} \underbrace{\left(\Omega_{kk} + \tau^2 \mathbf{e}_k^\top \left(\Theta^\top \Theta\right)^{-1} \mathbf{e}_k\right)}_{\Delta_b} + \underbrace{\frac{\tau^4}{m-1} \sum_{a=1}^{K} \frac{\beta_a^2}{m} \sum_{i \in I_a} \left(\mathbf{e}_k^\top \Theta^+ \mathbf{e}_i\right)^2}_{\Delta_c}. \tag{93}$$

The corresponding estimator is $\widehat{V}_k = \widehat{\Delta}_a \widehat{\Delta}_b + \widehat{\Delta}_c$ with

$$\widehat{\Delta}_a := \widehat{\sigma}^2 + \widehat{\tau}^2 \frac{\|\widehat{\beta}\|_2^2}{\widehat{m}}, \qquad \widehat{\Delta}_b := \widehat{\Omega}_{kk} + \widehat{\tau}^2 \mathbf{e}_k^\top (\widehat{\Theta}^\top \widehat{\Theta})^{-1} \mathbf{e}_k$$

and

$$\widehat{\Delta}_c := \frac{\widehat{\tau}^4}{\widehat{m} - 1} \sum_{a=1}^{K} \frac{\widehat{\beta}_a^2}{\widehat{m}} \sum_{i \in I_a} \left(\mathbf{e}_k^\top \widehat{\Theta}^+ \mathbf{e}_i\right)^2 \tag{94}$$

writing $\widehat{m} = |\widehat{I}_k|$ and $\widehat{\tau}^2 = \widehat{\tau}_i^2$ for any $k \in [K]$ and $i \in [p]$. Observe that

$$\frac{|\widehat{V}_k - V_k|}{V_k} \le \frac{|\widehat{\Delta}_a - \Delta_a|}{\Delta_a} + \frac{|\widehat{\Delta}_b - \Delta_b|}{\Delta_b} + \frac{|\widehat{\Delta}_c - \Delta_c|}{\Delta_a \Delta_b}, \tag{95}$$

and we will bound each term on the right separately. Lemma 24 in Section F.2 guarantees that $|\widehat{\Delta}_c - \Delta_c| = o_p(\Delta_a \Delta_b)$. From the definition of $\bar{\rho}$ in (18) and the inequality $|\widehat{m} - m|/\widehat{m} \le \bar{\rho}$, we have

$$\begin{aligned}
&|\widehat{\Delta}_a - \Delta_a| \\
&\le |\widehat{\sigma}^2 - \sigma^2| + \widehat{\tau}^2 \|\widehat{\beta}\|_2^2 \frac{|m - \widehat{m}|}{m\widehat{m}} + \frac{\|\beta\|_2^2}{m} |\widehat{\tau}^2 - \tau^2| + \frac{\widehat{\tau}^2}{m} (\|\widehat{\beta}\|_2 + \|\beta\|_2) \|\widehat{\beta} - \beta\|_2 \\
&\le |\widehat{\sigma}^2 - \sigma^2| + \widehat{\tau}^2 \frac{\|\widehat{\beta}\|_2^2}{m} \bar{\rho} + \frac{\|\beta\|_2^2}{m} |\widehat{\tau}^2 - \tau^2| + \frac{\widehat{\tau}^2}{m} (\|\widehat{\beta}\|_2 + \|\beta\|_2) \|\widehat{\beta} - \beta\|_2.
\end{aligned}$$

Since Lemma 14 and Assumption 5′ imply

$$K \log(p \vee n) = o(\sqrt{n}),$$

the rate of $\|\widehat{\beta} - \beta\|_2$ in (22) and the rate of $\max_{1 \le i \le p} |\widehat{\tau}_i^2 - \tau_i^2|$ in Lemma 21 in Section F.2 guarantee that, with probability tending to one,

$$\|\widehat{\beta}\|_2 = O\left(1 \vee \|\beta\|_2\right), \qquad \widehat{\tau}^2 = \tau^2 + o(1) = O(\tau^2). \tag{96}$$

Combine Lemmas 21 and 22 in Section F.2 with the rate of $\|\widehat{\beta} - \beta\|_2$ from (22) to find

$$|\widehat{\Delta}_a - \Delta_a| = O\left(\left(1 \vee \frac{\|\beta\|_2^2}{m}\right)\left(\bar{\rho} + C_{\min}^{-1/2}\delta_n\sqrt{K}\right)\right)$$

with probability tending to one. This bound with $K\log(p \vee n) = o(n)$ and Assumption 3 and the definition $\Delta_a = \sigma^2 + \tau^2\|\beta\|_2^2/m$ yields

$$\Delta_a^{-1}|\widehat{\Delta}_a - \Delta_a| = O\left(\bar{\rho} + C_{\min}^{-1/2}\delta_n\sqrt{K}\right) = o(1) \tag{97}$$

with probability tending to one. We proceed to bound $\Delta_b^{-1}|\widehat{\Delta}_b - \Delta_b|$ by

$$|\widehat{\Delta}_b - \Delta_b| \leq |\widehat{\Omega}_{kk} - \Omega_{kk}| + |\widehat{\tau}^2 - \tau^2| \cdot \mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k + \widehat{\tau}^2\left|\mathbf{e}_k^\top\left[(\widehat{\Theta}^\top\widehat{\Theta})^{-1} - (\Theta^\top\Theta)^{-1}\right]\mathbf{e}_k\right|,$$

and study each term on the right separately. Write $H := A\Sigma_Z^{1/2} = \Theta\Omega^{1/2}$ and $\widehat{H} = \widehat{\Theta}\Omega^{1/2}$. For the first term on the right, observe that the identity

$$\widehat{\Omega} - \Omega = \Omega^{1/2}\left[\Omega^{1/2}(\widehat{\Sigma}_Z - \Sigma_Z)\Omega^{1/2}\right](\Omega^{1/2}\widehat{\Sigma}_Z\Omega^{1/2})^{-1}\Omega^{1/2}, \tag{98}$$

implies

$$|\widehat{\Omega}_{kk} - \Omega_{kk}| \leq \Omega_{kk}\frac{\|\Omega^{1/2}(\Sigma_Z - \widehat{\Sigma}_Z)\Omega^{1/2}\|_{\mathrm{op}}}{\lambda_K(\Omega^{1/2}\widehat{\Sigma}_Z\Omega^{1/2})}.$$

By Weyl's inequality

$$\lambda_K\left(\Omega^{1/2}\widehat{\Sigma}_Z\Omega^{1/2}\right) = \lambda_K\left(\mathbf{I}_K - \Omega^{1/2}(\Sigma_Z - \widehat{\Sigma}_Z)\Omega^{1/2}\right) \geq 1 - \|\Omega^{1/2}(\widehat{\Sigma}_Z - \Sigma_Z)\Omega^{1/2}\|_{\mathrm{op}}, \tag{99}$$

and Lemma 23 in Section F.2 yields

$$|\widehat{\Omega}_{kk} - \Omega_{kk}| = O\left(\Omega_{kk}\delta_n\sqrt{K}\right) = o(\Delta_b)$$

with probability tending to one. Next, invoke Lemma 21 in Section F.2 and the definition of Δ_b to obtain for the second term on the right

$$|\widehat{\tau}^2 - \tau^2| \cdot \mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k = o(\Delta_b)$$

with probability tending to one. Finally, obtain the identity

$$\begin{aligned}
&(\widehat{\Theta}^\top\widehat{\Theta})^{-1} - (\Theta^\top\Theta)^{-1} \\
&= (\Theta^\top\Theta)^{-1}(\Theta^\top\Theta - \widehat{\Theta}^\top\widehat{\Theta})(\widehat{\Theta}^\top\widehat{\Theta})^{-1} \\
&= \Theta^+(\Theta - \widehat{\Theta})(\widehat{\Theta}^\top\widehat{\Theta})^{-1} + (\Theta^\top\Theta)^{-1}(\Theta - \widehat{\Theta})^\top\widehat{\Theta}(\widehat{\Theta}^\top\widehat{\Theta})^{-1} \\
&= \Omega^{1/2}\left[H^+(H - \widehat{H})(\widehat{H}^\top\widehat{H})^{-1} + (H^\top H)^{-1}(H - \widehat{H})^\top\widehat{H}(\widehat{H}^\top\widehat{H})^{-1}\right]\Omega^{1/2},
\end{aligned}$$

and invoke Lemma 10 in Section C to conclude

$$\left|\mathbf{e}_k^\top \left[(\widehat{\Theta}^\top\widehat{\Theta})^{-1} - (\Theta^\top\Theta)^{-1}\right]\mathbf{e}_k\right| = O\left(\frac{\Omega_{kk}}{\lambda_K(H^\top H)}\left(\delta_n\sqrt{K} + \frac{\delta_n\sqrt{pK}}{\sigma_K(H)}\right)\right) = o(\Delta_b)$$

with probability tending to one. The last step uses Assumption 5′. From (96), we get

$$\widehat{\tau}^2\left|\mathbf{e}_k^\top \left[(\widehat{\Theta}^\top\widehat{\Theta})^{-1} - (\Theta^\top\Theta)^{-1}\right]\mathbf{e}_k\right| = o(\Delta_b)$$

with probability tending to one. Collecting all three bounds, we find

$$\Delta_b^{-1}|\widehat{\Delta}_b - \Delta_b| = o(1)$$

with probability tending to one. This concludes completes the proof. □

F.2. Lemmas used in the proof of Proposition 5

We state the lemmas used for proving Proposition 5. Their proofs are deferred to Section F.3. All statements are valid on some events that are subsets of $\mathcal{E}$ and the probabilities of these events are greater than $1 - C(p \vee n)^{-\alpha}$ for some positive constants C, α. This is an important observation since on the event $\mathcal{E}$, the dimensions $\widehat{K}$ and K are equal, which ensures that the various quantities in the statements, for instance, the difference $\Sigma_Z - \Sigma_Z$, are well-defined.

Lemma 21. *Let $\widehat{\tau}_i^2$ be defined in (30). Under the conditions of Theorem 4.2, with probability greater than $1 - (p \vee n)^{-c}$ for some constant $c > 0$, we have*

$$\max_{1\le i\le p} |\widehat{\tau}_i^2 - \tau_i^2| \lesssim \delta_n.$$

Lemma 22. *Let $\widehat{\sigma}^2$ be defined as in (31). Under the conditions of Theorem 3, with probability $1 - c(p \vee n)^{-c'}$,*

$$|\widehat{\sigma}^2 - \sigma^2| \lesssim C_{\min}^{-1}\left(1 \vee \frac{\|\beta\|_2^2}{m}\right)\delta_n\sqrt{K}.$$

Lemma 23. *Under the conditions of Theorem 3, with probability $1 - c(p \vee n)^{-c'}$,*

$$\left\|\Sigma_Z^{-1/2}(\widehat{\Sigma}_Z - \Sigma_Z)\Sigma_Z^{-1/2}\right\|_{\text{op}} \lesssim \delta_n\sqrt{K}.$$

Lemma 24. *Let Δ_a, Δ_b and Δ_c be defined in (93). Let $\widehat{\Delta}_c$ be defined in (94). Under the conditions of Proposition 5, we have*

$$|\widehat{\Delta}_c - \Delta_c| = o_p(\Delta_a\Delta_b).$$

F.3. Proof of lemmas in Section F.2

F.3.1. Proof of Lemma 21

We work on the event $\mathcal{E}$ defined in (10) that has probability at least $1-(p\vee n)^{-c}$ for some $c>0$. Recall that it implies $\widehat{K}=K$ and $I_k\subseteq\widehat{I}_k\subseteq I_k\cup J_1^k$ for all $k\in[K]$.

Fix any $i\in[p]$. The decomposition $\mathbf{X}_{\bullet i}=\mathbf{Z}A_{i\bullet}+\mathbf{W}_{\bullet i}$ readily gives

$$|\widehat{\tau}_i^2-\tau_i^2|\le\left|\frac{1}{n}A_{i\bullet}^\top\mathbf{Z}^\top\mathbf{Z}A_{i\bullet}-\widehat{A}_{i\bullet}^\top\widehat{\Sigma}_Z\widehat{A}_{i\bullet}\right|+\left|\frac{1}{n}\mathbf{W}_{\bullet i}^\top\mathbf{W}_{\bullet i}-\tau_i^2\right|+\left|\frac{2}{n}A_{i\bullet}^\top\mathbf{Z}^\top\mathbf{W}_{\bullet i}\right|$$

Since $\mathbb{E}[n^{-1}\mathbf{W}_{\bullet i}^\top\mathbf{W}_{\bullet i}]=\tau_i^2$ and $|A_{i\bullet}^\top\mathbf{Z}^\top\mathbf{W}_{\bullet i}|\le\|\mathbf{Z}^\top\mathbf{W}_{\bullet i}\|_\infty$ by $\|A_{i\bullet}\|_1\le1$, an application of the exponential inequality in Lemma 6 followed by the union bound yields

$$\mathbb{P}\left\{\left|\frac{1}{n}\mathbf{W}_{\bullet i}^\top\mathbf{W}_{\bullet i}-\tau_i^2\right|+\left|\frac{2}{n}A_{i\bullet}^\top\mathbf{Z}^\top\mathbf{W}_{\bullet i}\right|\lesssim\delta_n\right\}\ge1-(p\vee n)^{-c}. \tag{100}$$

For the first term, we argue that, for any $i\in I_k$ and $k\in[K]$, $\widehat{A}_{i\bullet}=A_{i\bullet}$ (on the event $\mathcal{E}$) so that

$$\left|\frac{1}{n}A_{i\bullet}^\top\mathbf{Z}^\top\mathbf{Z}A_{i\bullet}-\widehat{A}_{i\bullet}^\top\widehat{\Sigma}_Z\widehat{A}_{i\bullet}\right|=\left|\frac{1}{n}\mathbf{Z}_{\bullet k}^\top\mathbf{Z}_{\bullet k}-[\Sigma_Z]_{kk}\right|.$$

This can be bounded by applying (25) in Lemma 7 with $u=v=\mathbf{e}_k$. Taking the union bound concludes the proof of $\widehat{\tau}_i^2-\tau_i^2$ for $i\in I$.

For any $i\in J$,

$$\begin{aligned}|\tau_i^2-\widehat{\tau}_i^2|=\left|\frac{1}{n}A_{i\bullet}^\top\mathbf{Z}^\top\mathbf{Z}A_{i\bullet}-\widehat{A}_{i\bullet}^\top\widehat{\Sigma}_Z\widehat{A}_{i\bullet}\right|&\le\left|(\widehat{A}_{i\bullet}-A_{i\bullet})^\top\widehat{\Sigma}_Z\widehat{A}_{i\bullet}\right|+\left|(\widehat{A}_{i\bullet}-A_{i\bullet})\widehat{\Sigma}_ZA_{i\bullet}\right|\\&\quad+\left|A_{i\bullet}^\top\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{Z}-\widehat{\Sigma}_Z\right)A_{i\bullet}\right|.\end{aligned}$$

By the Cauchy-Schwarz inequality, $\|\widehat{A}_{I\bullet}\|_1\le1$, $\|A_{i\bullet}\|_1\le1$ and Lemma 25, we obtain

$$\left|(\widehat{A}_{i\bullet}-A_{i\bullet})\widehat{\Sigma}_ZA_{i\bullet}\right|+\left|(\widehat{A}_{i\bullet}-A_{i\bullet})^\top\widehat{\Sigma}_Z\widehat{A}_{i\bullet}\right|\le2\|\widehat{\Sigma}_Z(\widehat{A}_{i\bullet}-A_{i\bullet})\|_\infty\lesssim\delta_n$$

with probability $1-(p\vee n)^{-c}$. Regarding the third term, apply Lemma 29 to find that

$$\begin{aligned}\left|A_{i\bullet}^\top\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{Z}-\widehat{\Sigma}_Z\right)A_{i\bullet}\right|&\le\|n^{-1}\mathbf{Z}^\top\mathbf{Z}-\widehat{\Sigma}_Z\|_\infty\\&\le\left\|n^{-1}\mathbf{Z}^\top\mathbf{Z}-\Sigma_Z\right\|_\infty+\|\widehat{\Sigma}_Z-\Sigma_Z\|_\infty\lesssim\delta_n\end{aligned}$$

with probability $1-(p\vee n)^{-c}$. Combining these two displays completes our proof. □

Lemma 25. *Under the same conditions of Theorem 4.2, let $\widehat{\Sigma}_Z$ be constructed as (8) and $\widehat{A}$ as in display (33) together with (9) – (10). Let $s_j = \|A_{j\bullet}\|_0$ for any $j \in [p]$. With probability greater than $1 - (p \vee n)^{-c}$ for some constant $c > 0$,*

$$\|\widehat{A}_{j\bullet} - A_{j\bullet}\|_2 \lesssim C_{\min}^{-1}\delta_n\sqrt{s_j}, \quad \|\widehat{\Sigma}_Z(\widehat{A}_{j\bullet} - A_{j\bullet})\|_\infty \lesssim \delta_n$$

hold for all $1 \le j \le p$.

Proof. Write $s := s_j$. The rate of $\widehat{A}_{j\bullet} - A_{j\bullet}$ follows immediately from [2, display (36) in Theorem 5] by observing that

$$\begin{aligned}
\kappa_2(\Sigma_Z, s) := \inf_{|S|\le s}\inf_{\substack{\|v\|=1\\ v\in\mathcal{C}_S}} \|\Sigma_Z v\|_\infty &\ge \inf_{|S|\le s}\inf_{\substack{\|v\|=1\\ \mathrm{supp}(v)=S}} \|\Sigma_Z v\|_\infty \\
&\ge \inf_{|S|\le s}\inf_{\substack{\|v\|=1\\ \mathrm{supp}(v)=S}} \|[\Sigma_Z]_{SS}\cdot v_S\|_\infty \\
&\ge \inf_{|S|\le s}\inf_{\substack{\|v\|=1\\ \mathrm{supp}(v)=S}} \|[\Sigma_Z]_{SS}\cdot v_S\|/\sqrt{s} \\
&\ge C_{\min}/\sqrt{s}
\end{aligned}$$

with $\mathcal{C}_S := \{v \in \mathbb{R}^K : \|v_{S^c}\|_1 \le \|v_S\|_1\}$ and $S \subseteq [K]$ with $|S| \le s$. The proof of the bound for $\|\widehat{\Sigma}_Z(\widehat{A}_{j\bullet} - A_{j\bullet})\|_\infty$ follows from, writing $\widehat{\Pi} := \widehat{A}_{\widehat{I}\bullet}[\widehat{A}_{\widehat{I}\bullet}^\top\widehat{A}_{\widehat{I}\bullet}]^{-1}$,

$$\|\widehat{\Sigma}_Z(\widehat{A}_{j\bullet} - A_{j\bullet})\|_\infty \le \|\widehat{\Sigma}_Z\widehat{A}_{j\bullet} - \widehat{\Pi}^\top\widehat{\Sigma}_{\widehat{I}j}\|_\infty + \|\widehat{\Sigma}_Z A_{j\bullet} - \widehat{\Pi}^\top\widehat{\Sigma}_{\widehat{I}j}\|_\infty = O(\delta_n)$$

by using the feasibility of both $\widehat{A}_{j\bullet}$ and $A_{j\bullet}$. □

F.3.2. Proof of Lemma 22

We work on the event $\mathcal{E}$ defined in (10) that has probability at least $1 - (p \vee n)^{-c}$ for some $c > 0$. Recall that it implies $\widehat{K} = K$ and $I_k \subseteq \widehat{I}_k \subseteq I_k \cup J_1^k$ for all $k \in [K]$.

Recall that, from the definition of $\widehat{h}$ in (32),

$$\widehat{h} = \frac{1}{n}(\widehat{A}_{\widehat{I}\bullet}^\top\widehat{A}_{\widehat{I}\bullet})^{-1}\widehat{A}_{\widehat{I}\bullet}^\top X_{\bullet\widehat{I}}^\top\mathbf{y} \overset{(12)}{=} \frac{1}{n}\widetilde{\mathbf{X}}^\top\mathbf{y}.$$

We observe that the definition of $\widehat{\Sigma}_Z$ in (8) yields

$$\left[\widehat{\Sigma}_Z\right]_{aa} = \frac{1}{n}\widetilde{\mathbf{X}}_{\bullet a}^\top\widetilde{\mathbf{X}}_{\bullet a} - d_a,\ \forall a \in [K]; \qquad \left[\widehat{\Sigma}_Z\right]_{ab} = \frac{1}{n}\widetilde{\mathbf{X}}_{\bullet a}^\top\widetilde{\mathbf{X}}_{\bullet b},\ \forall a \ne b \in [K] \tag{101}$$

with

$$d_a = \frac{1}{\widehat{m}_a^2}\sum_{i\in\widehat{I}_a}\frac{1}{n}\mathbf{X}_{\bullet i}^\top\mathbf{X}_{\bullet i} - \frac{1}{\widehat{m}_a^2(\widehat{m}_a - 1)}\sum_{i,j\in\widehat{I}_a, i\ne j}\frac{1}{n}\mathbf{X}_{\bullet i}^\top\mathbf{X}_{\bullet i}. \tag{102}$$

Let $D = \mathrm{diag}(d_1, \ldots, d_K)$ so that $\widehat{\Sigma}_Z = n^{-1}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} - D$, and define $\Delta_\beta := \mathbf{Z}\beta - \widetilde{\mathbf{Z}}\widehat{\beta}$. Since $\mathbf{y} = \mathbf{Z}\beta + \varepsilon$ and $\widetilde{\mathbf{X}} = \widetilde{\mathbf{Z}} + \widetilde{\mathbf{W}}$, we find

$$\begin{aligned}
\widehat{\sigma}^2 &= \frac{1}{n}\mathbf{y}^\top\mathbf{y} - \frac{2}{n}\widehat{\beta}^\top\widetilde{\mathbf{X}}^\top\mathbf{y} + \widehat{\beta}^\top\widehat{\Sigma}_Z\widehat{\beta} \\
&= \frac{1}{n}\|\mathbf{y} - \widetilde{\mathbf{X}}\widehat{\beta}\|_2^2 - \widehat{\beta}^\top\left(\frac{1}{n}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} - \widehat{\Sigma}_Z\right)\widehat{\beta} \\
&= \frac{1}{n}\|\mathbf{Z}\beta + \varepsilon - (\widetilde{\mathbf{Z}} + \widetilde{\mathbf{W}})\widehat{\beta}\|_2^2 - \widehat{\beta}^\top D\widehat{\beta} \\
&= \frac{1}{n}\|\Delta_\beta\|_2^2 + \frac{1}{n}\varepsilon^\top\varepsilon + \frac{2}{n}\varepsilon^\top\Delta_\beta - \frac{2}{n}\varepsilon^\top\widetilde{\mathbf{W}}\widehat{\beta} - \frac{2}{n}\widehat{\beta}^\top\widetilde{\mathbf{W}}^\top\Delta_\beta + \widehat{\beta}^\top\left(\frac{1}{n}\widetilde{\mathbf{W}}^\top\widetilde{\mathbf{W}} - D\right)\widehat{\beta},
\end{aligned}$$

and consequently we have

$$\begin{aligned}
|\widehat{\sigma}^2 - \sigma^2| \le \left|\frac{1}{n}\varepsilon^\top\varepsilon - \sigma^2\right| + \left\|\frac{1}{n}\widetilde{\mathbf{W}}^\top\widetilde{\mathbf{W}} - D\right\|_{\mathrm{op}}\|\widehat{\beta}\|^2 + \frac{2}{n}|\varepsilon^\top\widetilde{\mathbf{W}}^\top\widehat{\beta}| \\
+ \frac{1}{n}\|\Delta_\beta\|_2^2 + \frac{2}{n}|\varepsilon^\top\Delta_\beta| + \frac{2}{n}|\widehat{\beta}^\top\widetilde{\mathbf{W}}^\top\Delta_\beta|.
\end{aligned}$$

The rate of $\|\widehat{\beta} - \beta\|_2$ in Theorem 3 together with $K\log(p \vee n) = o(n)$ implies

$$\|\widehat{\beta}\| \lesssim 1 \vee \|\beta\|_2 \tag{103}$$

with probability $1 - (p \vee n)^{-c}$. From Lemma 6 and Lemma 28 with the same discretization arguments in the proof of Lemma 23, we obtain

$$\left|\frac{1}{n}\varepsilon^\top\varepsilon - \sigma^2\right| \lesssim \delta_n, \qquad \left\|\frac{1}{n}\widetilde{\mathbf{W}}^\top\widetilde{\mathbf{W}} - D\right\|_{\mathrm{op}} \lesssim \frac{\delta_n\sqrt{K}}{m}, \tag{104}$$

with probability greater than $1 - (p \vee n)^{-c}$. Using similar arguments as in the proof of Lemma 27, by substituting Z_k by ε, we find, with probability $1 - (p \vee n)^{-c}$,

$$\frac{1}{n}|\varepsilon^\top\widetilde{\mathbf{W}}v| \lesssim \left(\frac{\|v\|_2}{\sqrt{m}} + \rho\|v\|_1\right)\delta_n\sqrt{K}$$

for any fixed $v \in \mathbb{R}^K$. By choosing $v = \beta$ and $v = \mathbf{e}_k$ for $k \in [K]$ and using the bound

$$\frac{1}{n}\left|\varepsilon^\top\widetilde{\mathbf{W}}\widehat{\beta}\right| \le \frac{1}{n}\left|\varepsilon^\top\widetilde{\mathbf{W}}\beta\right| + \frac{1}{n}\|\varepsilon^\top\widetilde{\mathbf{W}}\|_\infty\sqrt{K}\|\widehat{\beta} - \beta\|_2,$$

we have, with probability $1 - c(p \vee n)^{-c'}$,

$$\frac{1}{n}\left|\varepsilon^\top\widetilde{\mathbf{W}}\widehat{\beta}\right| \lesssim C_{\min}^{-1}\left(1 \vee \frac{\|\beta\|_2}{\sqrt{m}}\right)\delta_n\sqrt{K}. \tag{105}$$

We proceed to bound the remaining three terms involving Δ_β. Recall that

$$\Delta_\beta = \mathbf{Z}(\beta - \widehat{\beta}) + (\mathbf{Z} - \widetilde{\mathbf{Z}})\widehat{\beta} = \mathbf{Z}(\beta - \widehat{\beta}) + \mathbf{Z}\Delta\widehat{\beta}$$

with Δ defined in (19). Observe that

$$\frac{1}{n}\|\Delta_\beta\|_2^2 \le \frac{1}{n}\|\mathbf{Z}(\widehat{\beta}-\beta)\|_2^2 + \frac{1}{n}\|\mathbf{Z}^\top\mathbf{Z}\|_\infty\|\Delta\|_{1,\infty}^2\|\widehat{\beta}\|_1^2,$$
$$\frac{1}{n}|\varepsilon^\top\Delta_\beta| \le \frac{1}{n}\|\varepsilon^\top\mathbf{Z}\|_\infty\sqrt{K}\|\widehat{\beta}-\beta\|_2 + \frac{1}{n}\|\varepsilon^\top\mathbf{Z}\|_\infty\|\Delta\widehat{\beta}\|_1,$$

and $\|\widehat{\beta}\|_1 \le \sqrt{K}\|\widehat{\beta}\|_2$. Hence, display (103), Lemmas 6, 9 and 26 and Assumptions 3 and 5 yield

$$\frac{1}{n}\|\Delta_\beta\|_2^2 + \frac{2}{n}|\varepsilon^\top\Delta_\beta| \lesssim C_{\min}^{-1}\left(1 \vee \frac{\|\beta\|_2^2}{m}\right)\delta_n\sqrt{K} \tag{106}$$

with probability at least $1 - c(p \vee n)^{-c'}$. Finally, we obtain the bound

$$\begin{aligned}\frac{1}{n}|\widehat{\beta}^\top\widetilde{\mathbf{W}}^\top\Delta| &\le \frac{1}{n}|\beta^\top\widetilde{\mathbf{W}}^\top\Delta| + \frac{1}{n}|(\widehat{\beta}-\beta)^\top\widetilde{\mathbf{W}}^\top\Delta| \\ &\le \frac{1}{n}\|\beta^\top\widetilde{\mathbf{W}}^\top\mathbf{Z}\|_\infty\sqrt{K}\|\widehat{\beta}-\beta\|_2 + \frac{1}{n}\|\beta^\top\widetilde{\mathbf{W}}^\top\mathbf{Z}\|_\infty\|\Delta\widehat{\beta}\|_1 \\ &\quad + \frac{1}{n}\|\widetilde{\mathbf{W}}^\top\mathbf{Z}\|_\infty\sqrt{K}\|\widehat{\beta}-\beta\|_2\|\Delta\widehat{\beta}\|_1 + \frac{1}{n}|(\widehat{\beta}-\beta)^\top\widetilde{\mathbf{W}}^\top\mathbf{Z}(\widehat{\beta}-\beta)| \\ &= O_p\left(C_{\min}^{-1/2}\left(1 \vee \frac{\|\beta\|_2^2}{m}\right)\delta_n\sqrt{K}\right) + \frac{1}{n}|(\widehat{\beta}-\beta)^\top\widetilde{\mathbf{W}}^\top\mathbf{Z}(\widehat{\beta}-\beta)|,\end{aligned}$$

using the same arguments as above, in combination with Lemma 27. We control the final term by

$$\begin{aligned}\frac{1}{n}|(\widehat{\beta}-\beta)^\top\mathbf{Z}^\top\widetilde{\mathbf{W}}(\widehat{\beta}-\beta)| &\le \left\|\Sigma_Z^{1/2}(\widehat{\beta}-\beta)\right\|_2^2 \sup_{v\in\mathcal{S}^{K-1}}\frac{1}{n}\left|v^\top\Omega^{1/2}\mathbf{Z}^\top\widetilde{\mathbf{W}}\Omega^{1/2}v\right| \\ &= O_p\left(C_{\min}^{-1/2}\left(1 \vee \frac{\|\beta\|_2^2}{m}\right)\delta_n\sqrt{K}\right)\end{aligned} \tag{107}$$

using inequality (131), Lemma 26 and $K\log(p\vee n) = o(n)$. Collecting (104) – (107) gives the desired result. □

The following lemmas are used in the proof of Lemma 22.

Lemma 26. *Under conditions of Theorem 3, with probability* $1 - (p \vee n)^{-c}$ *for some constant* $c > 0$,

$$\frac{1}{n}\|\mathbf{Z}\widehat{\beta} - \mathbf{Z}\beta\|_2^2 \lesssim \left(1 \vee \frac{\|\beta\|_2^2}{m}\right)\delta_n^2 K, \qquad \left\|\Sigma_Z^{1/2}(\widehat{\beta}-\beta)\right\|_2^2 \lesssim \left(1 \vee \frac{\|\beta\|_2^2}{m}\right)\delta_n^2 K.$$

Proof. From (57), (58) and (52), $\widehat{\Theta}\Omega^{1/2} = \widehat{H}$ and $\Theta\Omega^{1/2} = H$, it follows that, on the event $\mathcal{E}$,

$$\frac{1}{n}\|\mathbf{Z}(\widehat{\beta}-\beta)\|_2^2 \le \|\Delta_1\|_2 + \|H^+\Delta_2\|_2 + \|H^+(\widehat{H}-H)\Delta_1\|_2 + \|(\Sigma_Z)^{1/2}(\widehat{\Theta}^+ - \Theta^+)\Delta_2\|_{\text{op}}.$$

The first result follows immediately by invoking the upper bounds of each term in the proof of Theorem 3, except that there is no $C_{\min}^{-1/2}$. The second result follows from Lemma 23 and the inequality, on the event $\mathcal{E}$,

$$\|\mathbf{Z}(\widehat{\beta}-\beta)\|_2 \le \|\mathbf{Z}\Omega^{1/2}\|_{\mathrm{op}}\|\Sigma_Z^{1/2}(\widehat{\beta}-\beta)\|_2.$$

□

Lemma 27. *Let $u, v \in \mathbb{R}^K$ be any fixed vectors. Under the conditions of Theorem 4.2, with probability $1-(p\vee n)^{-c}$, one has*

$$\frac{1}{n}\left|u^\top \widetilde{\mathbf{Z}}^\top \widetilde{\mathbf{W}} v\right| \lesssim \left(\sqrt{u^\top \Sigma_Z u}+\bar{\rho}\|u\|_2\right)\left(\frac{\|v\|_2}{\sqrt{m}}+\bar{\rho}\|v\|_2\right)\delta_n. \tag{108}$$

Proof. We work on the event $\mathcal{E}$ defined in (10) that has probability at least $1-(p\vee n)^{-c}$ for some $c>0$. Recall that it implies $\widehat{K}=K$ and $I_k \subseteq \widehat{I}_k \subseteq I_k \cup J_1^k$ for all $k\in[K]$.

Since $\widetilde{\mathbf{Z}} = \mathbf{Z}+\mathbf{Z}\Delta$. from (19), we have

$$\begin{aligned}\frac{1}{n}|u^\top \widetilde{\mathbf{Z}}^\top \widetilde{\mathbf{W}} v| &= \frac{1}{n}|u^\top \mathbf{Z}^\top \widetilde{\mathbf{W}} v| + \frac{1}{n}|u^\top \Delta^\top \mathbf{Z}^\top \widetilde{\mathbf{W}} v| \\ &\le \frac{1}{n}|u^\top \mathbf{Z}^\top \widetilde{\mathbf{W}} v| + \max_{1\le k\le K}\frac{1}{n}|\mathbf{Z}_k^\top \widetilde{\mathbf{W}} v|\cdot\|\Delta u\|_1. \end{aligned}\tag{109}$$

Repeated application of Lemma 7 in conjunction with Lemma 9 gives

$$\begin{aligned}\frac{1}{n}|u^\top \widetilde{\mathbf{Z}}^\top \widetilde{\mathbf{W}} v| &\lesssim \delta_n\sqrt{u^\top\Sigma_Z u}\left(\frac{\|v\|_2}{\sqrt{m}}+\bar{\rho}\|v\|_2\right)+\delta_n\left(\frac{\|v\|_2}{\sqrt{m}}+\bar{\rho}\|v\|_2\right)\rho\|u\|_1\delta_n \\ &\lesssim \delta_n\left(\sqrt{u^\top\Sigma_Z u}+\bar{\rho}\|u\|_2\right)\left(\frac{\|v\|_2}{\sqrt{m}}+\bar{\rho}\|v\|_2\right).\end{aligned}$$

The last inequality uses $\delta_n \lesssim 1$. □

Lemma 28. *Let $\widetilde{\mathbf{W}}$ and $D=\mathrm{diag}(d_1,\dots d_K)$ be defined in (12) and (102), respectively. Under the conditions of Theorem 4.2, for any fixed vectors $u, v\in\mathbb{R}^K$, with probability $1-(p\vee n)^{-c}$,*

$$\left|u^\top\left(\frac{1}{n}\widetilde{\mathbf{W}}^\top\widetilde{\mathbf{W}}-D\right)v\right| \lesssim \left(\frac{1}{m}+\bar{\rho}^2\right)\|u\|_2\|v\|_2\delta_n.$$

Proof. We work on the event $\mathcal{E}$ defined in (10) that has probability at least $1-(p\vee n)^{-c}$ for some $c>0$. Recall that it implies $\widehat{K}=K$ and $I_k\subseteq\widehat{I}_k\subseteq I_k\cup J_1^k$ for all $k\in[K]$.

Recall that $\overline{\mathbf{W}} := \mathbf{W}A_{I\bullet}[A_{I\bullet}^\top A_{I\bullet}]^{-1} = \mathbf{W}\Pi$. For any $a \in [K]$, we have

$$\widetilde{\mathbf{W}}_{\bullet a} = \frac{1}{\widehat{m}_a}\sum_{i\in I_a}\mathbf{W}_{\bullet i} + \frac{1}{\widehat{m}_a}\sum_{i\in L_a}\mathbf{W}_{\bullet i} = \overline{\mathbf{W}}_{\bullet a} + \underbrace{\frac{|L_a|}{\widehat{m}_a}\left(\frac{1}{|L_a|}\sum_{i\in L_a}\mathbf{W}_{\bullet i} - \overline{\mathbf{W}}_{\bullet a}\right)}_{R_a}. \tag{110}$$

Expand the quadratic term as

$$\begin{aligned} u^\top\left(\frac{1}{n}\widetilde{\mathbf{W}}^\top\widetilde{\mathbf{W}} - D\right)v = \sum_a u_a v_a\left(\frac{1}{n}\widetilde{\mathbf{W}}_{\bullet a}^\top\widetilde{\mathbf{W}}_{\bullet a} - d_a\right) \\ + \frac{1}{n}\sum_{a,b\in[K],a\neq b}\sum u_a v_b\left(\overline{\mathbf{W}}_{\bullet a}^\top\overline{\mathbf{W}}_{\bullet b} + \overline{\mathbf{W}}_{\bullet a}^\top R_b + \overline{\mathbf{W}}_{\bullet b}^\top R_a + R_a^\top R_b\right). \end{aligned}$$

Since $\mathbf{X}_{\bullet i} = \mathbf{Z}A_{i\bullet} + \mathbf{W}_{\bullet i}$, by the definition of d_a in (102), we find

$$\begin{aligned} d_a &= \frac{1}{\widehat{m}_a^2}\sum_{i\in\widehat{I}_a}\frac{1}{n}\mathbf{X}_{\bullet i}^\top\mathbf{X}_{\bullet i} - \frac{1}{\widehat{m}_a^2(\widehat{m}_a-1)}\sum_{i,j\in\widehat{I}_a,i\neq j}\sum\frac{1}{n}\mathbf{X}_{\bullet i}^\top\mathbf{X}_{\bullet i} \\ &= \frac{1}{\widehat{m}_a^2}\sum_{i\in\widehat{I}_a}\left(\frac{1}{n}A_{i\bullet}^\top\mathbf{Z}^\top\mathbf{Z}A_{i\bullet} + \frac{1}{n}\mathbf{W}_{\bullet i}^\top\mathbf{W}_{\bullet i} + \frac{2}{n}A_{i\bullet}^\top\mathbf{Z}^\top\mathbf{W}_{\bullet i}\right) \\ &\quad - \frac{1}{\widehat{m}_a^2(\widehat{m}_a-1)}\sum_{i,j\in\widehat{I}_a,i\neq j}\sum\left[\frac{1}{n}A_{i\bullet}^\top\mathbf{Z}^\top\mathbf{Z}A_{j\bullet} + \frac{1}{n}\mathbf{W}_{\bullet i}^\top\mathbf{W}_{\bullet j} + \frac{2}{n}A_{i\bullet}^\top\mathbf{Z}^\top\mathbf{W}_{\bullet j}\right]. \end{aligned}$$

Rearranging terms yields

$$\begin{aligned} d_a = \frac{1}{\widehat{m}_a^2}\sum_{i,j\in\widehat{I}_a}\sum\frac{1}{n}\mathbf{W}_{\bullet i}^\top\mathbf{W}_{\bullet j} - \underbrace{\frac{1}{\widehat{m}_a(\widehat{m}_a-1)}\sum_{i,j\in\widehat{I}_a,i\neq j}\sum\frac{1}{n}\mathbf{W}_{\bullet i}^\top\mathbf{W}_{\bullet j}}_{T_1^a} \\ + \underbrace{\frac{1}{\widehat{m}_a^2}\sum_{i\in\widehat{I}_a}\frac{1}{n}A_{i\bullet}^\top\mathbf{Z}^\top\mathbf{Z}A_{i\bullet} - \frac{1}{\widehat{m}_a^2(\widehat{m}_a-1)}\sum_{i,j\in\widehat{I}_a,i\neq j}\sum\frac{1}{n}A_{i\bullet}^\top\mathbf{Z}^\top\mathbf{Z}A_{j\bullet}}_{T_2^a} \\ + \underbrace{\frac{2}{\widehat{m}_a^2}\sum_{i\in\widehat{I}_a}\frac{1}{n}A_{i\bullet}^\top\mathbf{Z}^\top\mathbf{W}_{\bullet i} - \frac{2}{\widehat{m}_a^2(\widehat{m}_a-1)}\sum_{i,j\in\widehat{I}_a,i\neq j}\sum\frac{1}{n}A_{i\bullet}^\top\mathbf{Z}^\top\mathbf{W}_{\bullet j}}_{T_3^a}. \end{aligned} \tag{111}$$

The first term on the right equals $n^{-1}\widetilde{\mathbf{W}}_{\bullet a}^{\top}\widetilde{\mathbf{W}}_{\bullet a}$. We further have

$$\begin{aligned}\left|u^{\top}\left(n^{-1}\widetilde{\mathbf{W}}^{\top}\widetilde{\mathbf{W}}-D\right)v\right| &\leq \|u\|_2\|v\|_2 \max_a(|T_1^a|+|T_2^a|+|T_3^a|) \\ &\quad + \frac{1}{n}\left|\sum_{a,b\in[K],a\neq b}\sum u_a v_b\left(\overline{\mathbf{W}}_{\bullet a}^{\top}\overline{\mathbf{W}}_{\bullet b}+\overline{\mathbf{W}}_{\bullet a}^{\top}R_b+\overline{\mathbf{W}}_{\bullet b}^{\top}R_a+R_a^{\top}R_b\right)\right| \qquad (112)\\ &:= \Delta_1+\Delta_2\end{aligned}$$

To bound Δ_1: We first study T_1^a, T_2^a and T_3^a, separately. For T_1^a, expand $\widehat{I}_a = I_a \cup L_a$ to get the bound

$$\frac{1}{\widehat{m}_a(\widehat{m}_a-1)}\left\{\left|\sum_{i,j\in I_a,i\neq j}\sum\frac{1}{n}\mathbf{W}_{\bullet i}^{\top}\mathbf{W}_{\bullet j}\right|+2\left|\sum_{i\in I_a,j\in L_a}\sum\frac{1}{n}\mathbf{W}_{\bullet i}^{\top}\mathbf{W}_{\bullet j}\right|+\left|\sum_{i\neq j\in L_a}\sum\frac{1}{n}\mathbf{W}_{\bullet i}^{\top}\mathbf{W}_{\bullet j}\right|\right\} \qquad (113)$$

$$\begin{aligned}&\leq \frac{1}{\widehat{m}_a(\widehat{m}_a-1)}\left|\sum_{i,j\in I_a,i\neq j}\sum\frac{1}{n}\mathbf{W}_{\bullet i}^{\top}\mathbf{W}_{\bullet j}\right|+\frac{2|L_a|}{\widehat{m}_a(\widehat{m}_a-1)}\max_{j\in L_a}\left|\sum_{i\in I_a}\frac{1}{n}\mathbf{W}_{\bullet i}^{\top}\mathbf{W}_{\bullet j}\right| \\ &\quad + \frac{|L_a|(|L_a|-1)}{\widehat{m}_a(\widehat{m}_a-1)}\max_{i,j\in\widehat{I}_a,i\neq j}\left|\frac{1}{n}\mathbf{W}_{\bullet i}^{\top}\mathbf{W}_{\bullet j}\right|.\end{aligned}$$

The first term is no greater than

$$\frac{1}{n}\left|\sum_{t=1}^{n}\frac{1}{m_a}\sum_{i\in I_a}\mathbf{W}_{ti}\frac{1}{m_a-1}\sum_{j\in I_a\setminus\{i\}}\mathbf{W}_{tj}\right| = \frac{1}{n}\left|\sum_{t=1}^{n}\overline{\mathbf{W}}_{ta}\frac{1}{m_a-1}\sum_{j\in I_a\setminus\{i\}}\mathbf{W}_{tj}\right|.$$

Since $(m_a-1)^{-1}\sum_{j\in I_a\setminus\{i\}}\mathbf{W}_{tj}$ is $(\gamma_w/\sqrt{m_a-1})$-sub-Gaussian by the arguments of Lemma 5 and

$$\mathbb{E}\left[\sum_{i\in I_a}\mathbf{W}_{ti}\sum_{j\neq i}\mathbf{W}_{tj}\right]=0,$$

invoking Lemma 6 gives

$$\frac{1}{\widehat{m}_a(\widehat{m}_a-1)}\left|\sum_{i,j\in I_a,i\neq j}\sum\frac{1}{n}\mathbf{W}_{\bullet i}^{\top}\mathbf{W}_{\bullet j}\right|\leq c\gamma_w^2\sqrt{\frac{\log(p\vee n)}{nm_a(m_a-1)}}$$

with probability $1-(p\vee n)^{-c'}$. Note that $|L_a|/\widehat{m}_a\leq\|\rho\|_2\leq\bar{\rho}$ and $\widehat{m}_a-1\geq m_a$ when $L_a\neq\varnothing$. We further obtain

$$|T_1^a|\leq\frac{1}{\widehat{m}_a(\widehat{m}_a-1)}\left|\sum_{i,j\in I_a,i\neq j}\sum\frac{1}{n}\mathbf{W}_{\bullet i}^{\top}\mathbf{W}_{\bullet j}\right|+2\bar{\rho}\max_{j\in L_a}\frac{1}{n}\left|\sum_{t=1}^{n}\mathbf{W}_{tj}\overline{\mathbf{W}}_{ta}\right|+\bar{\rho}^2\max_{i,j\in\widehat{I}_a,i\neq j}\left|\frac{1}{n}\mathbf{W}_{\bullet i}^{\top}\mathbf{W}_{\bullet j}\right|.$$

Invoking Lemma 6 and taking an union bound yields

$$\max_a |T_1^a| \le c\gamma_w^2 \left(\frac{1}{\sqrt{m(m-1)}} \vee \frac{\bar\rho}{\sqrt{m}} \vee \bar\rho^2 \right) \delta_n, \tag{114}$$

with probability greater than $1 - c(p \vee n)^{-c'}$. We bound T_2^a by writing

$$\begin{aligned} T_2^a &= \frac{m_a}{\widehat m_a^2} \frac{1}{n} \mathbf{Z}_{\bullet a}^\top \mathbf{Z}_{\bullet a} + \frac{1}{\widehat m_a^2} \sum_{i \in L_a} \frac{1}{n} A_{i\bullet}^\top \mathbf{Z}^\top \mathbf{Z} A_{i\bullet} \\ &\quad - \frac{1}{\widehat m_a^2(\widehat m_a - 1)} \Big\{ m_a(m_a-1) \frac{1}{n} \mathbf{Z}_{\bullet a}^\top \mathbf{Z}_{\bullet a} + 2m_a \sum_{i \in L_a} \frac{1}{n} \mathbf{Z}_{\bullet a}^\top \mathbf{Z} A_{i\bullet} \\ &\qquad + \sum_{i \ne j \in L_a} \frac{1}{n} A_{i\bullet}^\top \mathbf{Z}^\top \mathbf{Z} A_{j\bullet} \Big\} \\ &= \frac{1}{\widehat m_a^2} \sum_{i \in L_a} \left(\frac{1}{n} A_{i\bullet}^\top \mathbf{Z}^\top \mathbf{Z} A_{i\bullet} - \frac{1}{n} \mathbf{Z}_{\bullet a}^\top \mathbf{Z}_{\bullet a} \right) - \frac{2m_a}{\widehat m_a^2(\widehat m_a - 1)} \sum_{i \in L_a} \frac{1}{n} \mathbf{Z}_{\bullet a}^\top \mathbf{Z} (A_{i\bullet} - \mathbf{e}_a) \\ &\quad - \frac{1}{\widehat m_a^2(\widehat m_a - 1)} \sum_{i \ne j \in L_a} \left(\frac{1}{n} A_{i\bullet}^\top \mathbf{Z}^\top \mathbf{Z} A_{j\bullet} - \frac{1}{n} \mathbf{Z}_{\bullet a}^\top \mathbf{Z}_{\bullet a} \right). \end{aligned}$$

From (17), on the event $\mathcal{E}$ defined in (10), we find

$$\left| \frac{1}{n} \mathbf{Z}_{\bullet a}^\top \mathbf{Z} (A_{I\bullet} - \mathbf{e}_b) \right| \le \frac{8\delta_n}{\nu} \max_\ell \left| \frac{1}{n} \mathbf{Z}_{\bullet a}^\top \mathbf{Z}_\ell \right| \overset{\mathcal{E}}{\le} \frac{8B_z}{\nu} \delta_n, \tag{115}$$

for any $i \in L_b$ and $a \in [K]$, which in conjunction with $\|A_{j\bullet}\|_1 \le 1$ further gives

$$\begin{aligned} \left| \frac{1}{n} A_{i\bullet}^\top \mathbf{Z}^\top \mathbf{Z} A_{j\bullet} - \frac{1}{n} \mathbf{Z}_{\bullet a}^\top \mathbf{Z}_{\bullet a} \right| &\le \left| \frac{1}{n} (A_{i\bullet} - \mathbf{e}_a)^\top \mathbf{Z}^\top \mathbf{Z} A_{j\bullet} \right| + \left| \frac{1}{n} \mathbf{Z}_{\bullet a}^\top \mathbf{Z} (A_{j\bullet}^\top - \mathbf{e}_a) \right| \\ &\le \frac{16 B_z}{\nu} \delta_n. \end{aligned}$$

We thus obtain, on the event $\mathcal{E}$,

$$\begin{aligned} \max_a |T_2^a| &\le \max_a \left\{ \frac{|L_a|}{\widehat m_a^2} + \frac{m_a |L_a|}{\widehat m_a^2(\widehat m_a - 1)} + \frac{|L_a|(|L_a|-1)}{\widehat m_a^2(\widehat m_a - 1)} \right\} \frac{16 B_z}{\nu} \delta_n \\ &\le \frac{32 B_z}{\nu} \frac{\bar\rho \delta_n}{m} \lesssim \bar\rho \sqrt{\frac{\log(p \vee n)}{n m^2}}. \end{aligned} \tag{116}$$

Regarding T_3^a, notice that

$$\sum_{i \in I_a} \frac{1}{n} A_{i\bullet}^\top \mathbf{Z}^\top \mathbf{W}_{\bullet i} = m_a \frac{1}{n} \mathbf{Z}_{\bullet a}^\top \overline{\mathbf{W}}_{\bullet a} + \sum_{i \in L_a} \frac{1}{n} A_{i\bullet}^\top \mathbf{Z}^\top \mathbf{W}_{\bullet i}$$

and

$$\sum_{j,\ell\in\widehat{I}_a,j\neq\ell}\sum \frac{1}{n}A_{i\bullet}^{\top}\mathbf{Z}^{\top}\mathbf{W}_{\bullet i}$$
$$= m_a(m_a-1)\frac{1}{n}\mathbf{Z}_{\bullet a}^{\top}\overline{\mathbf{W}}_{\bullet a}+m_a\sum_{\ell\in L_a}\frac{1}{n}\mathbf{Z}_{\bullet a}^{\top}\mathbf{W}_{\bullet\ell}+\sum_{j\in L_a,\ell\in\widehat{I}_a\setminus\{j\}}\sum\frac{1}{n}A_{j\bullet}^{\top}\mathbf{Z}^{\top}\mathbf{W}_{\bullet\ell}.$$

After a bit algebra, by also using $\|A_{j\bullet}\|_1\leq 1$ for $1\leq j\leq p$, we can establish that

$$\begin{aligned}|T_3^a| &\lesssim \frac{|L_a|}{\widehat{m}_a^2}\max_{i\in\widehat{I}_a}\frac{1}{n}\|\mathbf{Z}^{\top}\mathbf{W}_{\bullet i}\|_{\infty}+\frac{m_a|L_a|}{\widehat{m}_a^2(\widehat{m}_a-1)}\frac{1}{n}|\mathbf{Z}_{\bullet a}^{\top}\overline{\mathbf{W}}_{\bullet a}|\\ &\leq \frac{|L_a|}{\widehat{m}_a^2}\max_{i\in\widehat{I}_a}\frac{1}{n}\|\mathbf{Z}^{\top}\mathbf{W}_{\bullet i}\|_{\infty}+\frac{|L_a|}{\widehat{m}_a^2}\frac{1}{n}|\mathbf{Z}_{\bullet a}^{\top}\overline{\mathbf{W}}_{\bullet a}|.\end{aligned}$$

By (15) and (16), invoking Lemmas 5 – 6 and taking the union bounds over $a\in[K]$, $i,j\in I_k\cup J_1^k$ yield

$$\max_a|T_3^a|\lesssim\bar{\rho}\sqrt{\frac{\log(p\vee n)}{nm^2}} \tag{117}$$

with probability $1-(p\vee n)^{-c}$. Collecting displays (114) - (117) concludes

$$\Delta_1\lesssim\|u\|_2\|v\|_2\left(\frac{1}{\sqrt{m(m-1)}}\vee\frac{\bar{\rho}}{\sqrt{m}}\vee\bar{\rho}^2\right)\sqrt{\frac{\log(p\vee n)}{n}}, \tag{118}$$

with probability greater than $1-c(p\vee n)^{-c'}$.

To bound Δ_2: We study the first term

$$\frac{1}{n}\sum_{a,b\in[K],a\neq b}\sum u_a v_b\overline{\mathbf{W}}_{\bullet a}^{\top}\overline{\mathbf{W}}_{\bullet b}=\frac{1}{n}\sum_{t=1}^{n}\sum_a u_a\overline{\mathbf{W}}_{ta}\sum_{b\neq a}v_b\overline{\mathbf{W}}_{tb}.$$

Since Lemma 5 guarantees $\sum_a u_a\overline{\mathbf{W}}_{ta}$ is $(\|u\|_2\gamma_w/\sqrt{m})$-sub-Gaussian and, similarly, $\sum_{b\neq a}v_b\overline{\mathbf{W}}_{tb}$ is $(\|v\|_2\gamma_w/\sqrt{m})$-sub-Gaussian, invoking Lemma 6 and noting that

$$\mathbb{E}\left[\sum_a u_a\overline{\mathbf{W}}_{ta}\sum_{b\neq a}v_b\overline{\mathbf{W}}_{tb}\right]=\sum_{a,b\in[K],a\neq b}\sum u_a v_b\mathbb{E}\left[\overline{\mathbf{W}}_{ta}\overline{\mathbf{W}}_{tb}\right]=0$$

gives

$$\mathbb{P}\left\{\frac{1}{n}\sum_{a,b\in[K],a\neq b}\sum u_a v_b\overline{\mathbf{W}}_{\bullet a}^{\top}\overline{\mathbf{W}}_{\bullet b}\lesssim\|u\|_2\|v\|_2\gamma_w^2\sqrt{\frac{\log(p\vee n)}{nm^2}}\right\}\geq 1-(p\vee n)^{-c}. \tag{119}$$

We then bound the rest term by term. By recalling (110), we obtain

$$\begin{aligned}
&\frac{1}{n}\left|\sum_{a,b\in[K],a\neq b} u_a v_b \overline{\mathbf{W}}_{\bullet a}^\top R_b\right| \\
&= \frac{2}{n}\left|\sum_{b=1}^K \frac{v_b}{\widehat{m}_b}\sum_{t=1}^n \left(|L_b|\overline{\mathbf{W}}_{tb} - \sum_{i\in L_b}\mathbf{W}_{ti}\right)\sum_{a,b\in[K],a\neq b} u_a \overline{\mathbf{W}}_{ta}\right| \\
&\le 2\|D_\rho v\|_1 \max_{b\in[K]}\left|\frac{1}{n}\sum_{t=1}^n \left(\overline{\mathbf{W}}_{tb} - \frac{1}{|L_b|}\sum_{i\in L_b}\mathbf{W}_{ti}\right)\sum_{a,b\in[K],a\neq b} u_a \overline{\mathbf{W}}_{ta}\right| \\
&\le 2\bar\rho\|v\|_2 \max_{b\in[K]}\left|\frac{1}{n}\sum_{t=1}^n \overline{\mathbf{W}}_{tb}\sum_{a,b\in[K],a\neq b} u_a \overline{\mathbf{W}}_{ta}\right| + 2\bar\rho\|v\|_2 \max_{\substack{b\in[K]\\ i\in L_b}}\left|\frac{1}{n}\sum_{t=1}^n \mathbf{W}_{ti}\sum_{a,b\in[K],a\neq b} u_a \overline{\mathbf{W}}_{ta}\right|
\end{aligned}$$

where we used (15) and (16) in the last two lines. Invoke Lemmas 5 and 6 and take union bounds to conclude

$$\frac{1}{n}\left|\sum_{a,b\in[K],a\neq b} u_a v_b \overline{\mathbf{W}}_{\bullet a}^\top R_b\right| \le c\bar\rho\|u\|_2\|v\|_2\sqrt{\frac{\log(p\vee n)}{nm}} \tag{120}$$

with probability $1-(p\vee n)^{-c'}$. A similar bound can be obtained for $|\sum\sum_{a,b\in[K],a\neq b} u_a v_b \overline{\mathbf{W}}_{\bullet b}^\top R_a|$. Finally, to bound the last term, we write

$$\begin{aligned}
&\frac{1}{n}\left|\sum_{a,b\in[K],a\neq b} u_a v_b R_b^\top R_a\right| \\
&\quad= \frac{1}{n}\left|\sum_{a\neq b}\frac{u_a v_b}{\widehat{m}_b\widehat{m}_a}\sum_{t=1}^n \left(|L_b|\overline{\mathbf{W}}_{tb} - \sum_{i\in L_b}\mathbf{W}_{ti}\right)\left(|L_a|\overline{\mathbf{W}}_{ta} - \sum_{i\in L_a}\mathbf{W}_{ti}\right)\right|,
\end{aligned}$$

which can be bounded above by

$$\begin{aligned}
&\bar\rho^2\|u\|_2\|v\|_2 \max_{a\neq b}\left|\frac{1}{n}\sum_{t=1}^n \left(\overline{\mathbf{W}}_{tb} - \frac{1}{|L_b|}\sum_{i\in L_b}\mathbf{W}_{ti}\right)\left(\overline{\mathbf{W}}_{ta} - \frac{1}{|L_a|}\sum_{i\in L_a}\mathbf{W}_{ti}\right)\right| \\
&\le \bar\rho^2\|u\|_2\|v\|_2 \max_{a\neq b}\left|\frac{1}{n}\sum_{t=1}^n \overline{\mathbf{W}}_{tb}\overline{\mathbf{W}}_{ta}\right| + 2\bar\rho^2\|u\|_2\|v\|_2 \max_{\substack{a\neq b\in[K]\\ i\in L_b}}\left|\frac{1}{n}\sum_{t=1}^n \mathbf{W}_{ti}\overline{\mathbf{W}}_{ta}\right| \\
&\quad + \bar\rho^2\|u\|_2\|v\|_2 \max_{\substack{a\neq b\in[K]\\ i\in L_a, j\in L_b}}\left|\frac{1}{n}\sum_{t=1}^n \mathbf{W}_{ti}\mathbf{W}_{tj}\right|.
\end{aligned} \tag{121}$$

Invoke Lemmas 5 and 6 and take the union bounds to obtain

$$\frac{1}{n}\left|\sum_{a,b\in[K],a\neq b}\sum u_a v_b R_a^\top R_b\right| \lesssim \bar{\rho}^2\|u\|_2\|v\|_2\sqrt{\frac{\log(p\vee n)}{n}} \tag{122}$$

with probability $1-(p\vee n)^{-c}$. Combining (119), (120) and (122) concludes

$$\Delta_2 \lesssim \left(\frac{\|u\|_2}{\sqrt{m}}+\bar{\rho}\|u\|_2\right)\left(\frac{\|v\|_2}{\sqrt{m}}+\bar{\rho}\|v\|_2\right)\sqrt{\frac{\log(p\vee n)}{n}} \tag{123}$$

with probability $1-c(p\vee n)^{-c'}$. Finally, combine (118) and (123) to complete the proof. □

F.3.3. Proof of Lemma 23

We first state and prove the following lemma which is used for proving Lemma 23.

Lemma 29. *Under the conditions of Theorem 4.2, for any fixed vector $v,u\in\mathbb{R}^K$, we have*

$$\left|u^\top(\widehat{\Sigma}_Z-\Sigma_Z)v\right| \lesssim \left(\sqrt{u^\top\Sigma_Z u}+\bar{\rho}\|u\|_2\right)\left(\sqrt{v^\top\Sigma_Z v}+\bar{\rho}\|v\|_2\right)\delta_n$$

with probability greater than $1-c(p\vee n)^{-c'}$ for some constant $c>0$.

Proof. We work on the event $\mathcal{E}$ defined in (10) that has probability at least $1-(p\vee n)^{-c}$ for some $c>0$. Recall that it implies $\widehat{K}=K$ and $I_k\subseteq\widehat{I}_k\subseteq I_k\cup J_1^k$ for all $k\in[K]$.

Recall that, from the proof of Theorem 3, $\widehat{\Sigma}_Z=n^{-1}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}-D$ with $D=\mathrm{diag}(d_1,\ldots,d_K)$ and d_k defined in (102). By $\widetilde{\mathbf{X}}=\widetilde{\mathbf{Z}}+\widetilde{\mathbf{W}}$, we obtain

$$\begin{aligned}
u^\top(\widehat{\Sigma}_Z-\Sigma_Z)v &= u^\top\left(\frac{1}{n}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}-\Sigma_Z\right)v-u^\top Dv\\
&= u^\top\left(\frac{1}{n}\widetilde{\mathbf{Z}}^\top\widetilde{\mathbf{Z}}-\Sigma_Z\right)v+u^\top\left(\frac{1}{n}\widetilde{\mathbf{W}}^\top\widetilde{\mathbf{W}}-D\right)v\\
&\quad+\frac{1}{n}u^\top\widetilde{\mathbf{Z}}^\top\widetilde{\mathbf{W}}v+\frac{1}{n}v^\top\widetilde{\mathbf{Z}}^\top\widetilde{\mathbf{W}}u\\
&:= T_1+T_2+T_3+T_4.
\end{aligned}$$

Plugging $\widetilde{\mathbf{Z}}=\mathbf{Z}+\mathbf{Z}\Delta$ into T_1 yields

$$\begin{aligned}
|T_1| &\le \left|u^\top\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{Z}-\Sigma_Z\right)v\right|+\frac{1}{n}\left|u^\top\mathbf{Z}^\top\mathbf{Z}\Delta v\right|+\frac{1}{n}\left|u^\top\Delta^\top\mathbf{Z}^\top\mathbf{Z}v\right|+\frac{1}{n}\left|u^\top\Delta^\top\mathbf{Z}^\top\mathbf{Z}\Delta v\right|\\
&\le \left|u^\top\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{Z}-\Sigma_Z\right)v\right|+\frac{1}{n}\left\|u^\top\mathbf{Z}^\top\mathbf{Z}\right\|_\infty\|\Delta v\|_1+\frac{1}{n}\left\|\mathbf{Z}^\top\mathbf{Z}v\right\|_\infty\|\Delta u\|_1\\
&\quad+\frac{1}{n}\max_a|\mathbf{Z}_{\bullet a}^\top\mathbf{Z}_{\bullet a}|\cdot\|\Delta u\|_1\|\Delta v\|_1
\end{aligned}\tag{124}$$

By using Lemmas 6 and 9, with probability $1-(p\vee n)^{-c}$, one has

$$|T_1| \lesssim \delta_n\sqrt{u^\top\Sigma_Z u}\sqrt{v^\top\Sigma_Z v}+\delta_n\bar\rho\left(\|u\|_2\|v\|_2\bar\rho\delta_n+\|u\|_2\sqrt{v^\top\Sigma_Z v}+\|v\|_2\sqrt{u^\top\Sigma_Z u}\right)$$
$$\lesssim \delta_n\left(\sqrt{u^\top\Sigma_Z u}+\bar\rho\|u\|_2\right)\left(\sqrt{v^\top\Sigma_Z v}+\bar\rho\|v\|_2\right) \tag{125}$$

where we have used $\delta_n\lesssim 1$ in the second line. The proof is completed by invoking Lemmas 27 – 28 for T_2 – T_4 and using $u^\top\Sigma_Z u\ge C_{\min}\|u\|_2^2$ to simplify expressions. □

Proof of Lemma 23. We again work on the event $\mathcal{E}$ such that $\widehat K=K$ and $I_k\subseteq\widehat I_k\subseteq I_k\cup J_1^k$ for all $k\in[K]$.

To prove the upper bound for $\|\Omega^{1/2}(\widehat\Sigma_Z-\Sigma_Z)\Omega^{1/2}\|_{\rm op}$, as in the proof of Lemma 29, we consider the terms T_1 – T_4 separately (except here $T_3=T_4$ since $u=v$). Specifically, we will upper bound

$$T_1'+T_2'+T_3' := \sup_{v\in\mathcal{S}^{K-1}}\left|v^\top\Omega^{1/2}\left(\frac1n\widetilde{\mathbf{Z}}^\top\widetilde{\mathbf{Z}}-\Sigma_Z\right)\Omega^{1/2}v\right|+\|\Omega\|_{\rm op}\sup_{v\in\mathcal{S}^{K-1}}\left|v^\top\left(\frac1n\widetilde{\mathbf{W}}^\top\widetilde{\mathbf{W}}-D\right)v\right|$$
$$+\sup_{v\in\mathcal{S}^{K-1}}\frac2n\left|v^\top\Omega^{1/2}\widetilde{\mathbf{Z}}^\top\widetilde{\mathbf{W}}\Omega^{1/2}v\right|.$$

For T_1', (124) implies

$$T_1' \le \left\|\Omega^{1/2}\left(\frac1n\mathbf{Z}^\top\mathbf{Z}-\Sigma_Z\right)\Omega^{1/2}\right\|_{\rm op}+2\sup_{v\in\mathcal{S}^{K-1}}\frac1n\left\|v^\top\Omega^{1/2}\mathbf{Z}^\top\mathbf{Z}\right\|_\infty\|\Delta\Omega^{1/2}v\|_1$$
$$+\sup_{v\in\mathcal{S}^{K-1}}\frac1n\max_{1\le a\le K}|\mathbf{Z}_{\bullet a}^\top\mathbf{Z}_{\bullet a}|\cdot\|\Delta\Omega^{1/2}v\|_1^2$$
$$\lesssim \delta_n\sqrt K+\bar\rho\delta_n\sqrt{K/C_{\min}}+\bar\rho^2\delta_n^2\sqrt K/C_{\min}$$

with probability $1-(p\vee n)^{-c}-p^{-cK}$. The last inequality uses Lemma 9 with $\|\Omega^{1/2}v\|_2\le\|v\|_2/\sqrt{C_{\min}}$, (31) in Lemma 8 and (26) in Lemma 7. We thus find

$$T_1'\lesssim\delta_n\sqrt K(1\vee\bar\rho^2/C_{\min}) \tag{126}$$

with probability $1-(p\vee n)^{-c'}$. To bound T_2', from (112), it suffices to bound $\sup_{v\in\mathcal{S}^{K-1}}\Delta_2$ only, since the bound of Δ_1 is uniformly in v. Display (112) gives

$$\sup_{v\in\mathcal{S}^{K-1}}\Delta_2=\sup_{v\in\mathcal{S}^{K-1}}\frac1n\left|\sum_{a,b\in[K],a\ne b}v_av_b\overline{\mathbf{W}}_{\bullet a}^\top\overline{\mathbf{W}}_{\bullet b}\right|$$
$$+\sup_{v\in\mathcal{S}^{K-1}}\frac1n\left|\sum_{a,b\in[K],a\ne b}v_av_b\left(2\overline{\mathbf{W}}_{\bullet a}^\top R_b+R_a^\top R_b\right)\right|.$$

By repeating a discretization argument similar to the one above, we can show from (119) that

$$\mathbb{P}\left\{\sup_{v\in\mathcal{S}^{K-1}}\frac{1}{n}\left|\sum_{a,b\in[K],a\neq b}v_a v_b\overline{\mathbf{W}}_{\bullet a}^{\top}\overline{\mathbf{W}}_{\bullet b}\right|\leq c\sqrt{\frac{K\log(p\vee n)}{nm^2}}\right\}\geq 1-(p\vee n)^{-c'K} \tag{127}$$

and that, from (120),

$$\sup_{v\in\mathcal{S}^{K-1}}\frac{2}{n}\left|\sum_{a,b\in[K],a\neq b}v_a v_b\overline{\mathbf{W}}_{\bullet a}^{\top}R_b\right|\leq c\bar{\rho}\sqrt{\frac{K\log(p\vee n)}{nm}} \tag{128}$$

with probability $1-(p\vee n)^{-c'K}$. From (121), the last term $n^{-1}\sum v_a v_b R_a^{\top}R_b$ can be upper bounded by

$$\begin{aligned}
&\sup_{v\in\mathcal{S}^{K-1}}\frac{1}{n}\left|\sum_{a,b\in[K],a\neq b}v_a v_b R_a^{\top}R_b\right| \\
&\leq\bar{\rho}^2\max_{a\neq b}\left|\frac{1}{n}\sum_{t=1}^{n}\overline{\mathbf{W}}_{tb}\overline{\mathbf{W}}_{ta}\right|+\bar{\rho}^2\max_{\substack{a\neq b\in[K]\\ i\in L_b}}\left|\frac{1}{n}\sum_{t=1}^{n}\mathbf{W}_{ti}\overline{\mathbf{W}}_{ta}\right| \\
&\quad+\bar{\rho}^2\max_{\substack{a\neq b\in[K]\\ i\in L_a}}\left|\frac{1}{n}\sum_{t=1}^{n}\mathbf{W}_{ti}\overline{\mathbf{W}}_{tb}\right|+\bar{\rho}^2\max_{\substack{a\neq b\in[K]\\ i\in L_a,j\in L_b}}\left|\frac{1}{n}\sum_{t=1}^{n}\mathbf{W}_{ti}\mathbf{W}_{tj}\right|.
\end{aligned}$$

Invoking Lemmas 5 - 6 and taking the union bound conclude that

$$\sup_{v\in\mathcal{S}^{K-1}}\frac{1}{n}\left|\sum_{a,b\in[K],a\neq b}v_a v_b R_a^{\top}R_b\right|\leq c\bar{\rho}^2\delta_n \tag{129}$$

with probability $1-(p\vee n)^{-c'}$. Finally, from $\|\Omega\|_{\mathrm{op}}\leq C_{\min}^{-1}$, collecting (127) – (129) and invoking the bound of Δ_1 via (114), (116) and (117) yield, with probability $1-c(p\vee n)^{-c'}$,

$$T_2'\leq c\left(\frac{1}{m}\vee\bar{\rho}^2\right)C_{\min}^{-1}\delta_n\sqrt{K}. \tag{130}$$

We then proceed to bound T_3'. From (109), by using Lemma 9 and $\|\Omega^{1/2}v\|_2\leq 1/\sqrt{C_{\min}}$ for any $v\in\mathcal{S}^{K-1}$, we know

$$\frac{1}{n}|v^{\top}\Omega^{1/2}\widetilde{\mathbf{Z}}^{\top}\widetilde{\mathbf{W}}v|\leq\frac{1}{n}|v^{\top}\Omega^{1/2}\mathbf{Z}^{\top}\widetilde{\mathbf{W}}\Omega^{1/2}v|+\max_k\frac{1}{n}|\mathbf{Z}_k^{\top}\widetilde{\mathbf{W}}\Omega^{1/2}v|\cdot\bar{\rho}\delta_n/\sqrt{C_{\min}}$$

with probability $1-(p\vee n)^{-c}$. By (32) in Lemma 8 and $K\log(p\vee n)=O(n)$, we have

$$T_3'\leq c\left(\frac{1}{\sqrt{m}}\vee\bar{\rho}\right)\delta_n\sqrt{K/C_{\min}} \tag{131}$$

with probability $1-c(p\vee n)^{-c'}$. Collecting the bounds for T_1', T_2' and T_3' completes the proof. □

F.3.4. Proof of Lemma 24

We work on the event $\mathcal{E}$ defined in (10) that has probability at least $1-(p\vee n)^{-c}$ for some $c>0$. Recall that it implies $\widehat{K}=K$ and $I_k\subseteq\widehat{I}_k\subseteq I_k\cup J_1^k$ for all $k\in[K]$.

From definition, by adding and subtracting terms and using the fact $\widehat{m}\geq 2$, we have

$$
\begin{aligned}
&|\widehat{\Delta}_c-\Delta_c|\\
&\leq\left|\frac{\widehat{\tau}^4}{\widehat{m}-1}\sum_{a=1}^K\frac{\widehat{\beta}_a^2}{\widehat{m}}\sum_{i\in I_a}\left[\left(\mathbf{e}_k^\top\widehat{\Theta}^+\mathbf{e}_i\right)^2-\left(\mathbf{e}_k^\top\Theta^+\mathbf{e}_i\right)^2\right]\right|\\
&\quad+\left|\frac{\widehat{\tau}^4}{\widehat{m}-1}\sum_{a=1}^K\left[\frac{\widehat{\beta}_a^2}{\widehat{m}}-\frac{\beta_a^2}{m}\right]\sum_{i\in I_a}\left(\mathbf{e}_k^\top\Theta^+\mathbf{e}_i\right)^2\right|+\left|\frac{\widehat{\tau}^4}{\widehat{m}-1}-\frac{\tau^4}{m-1}\right|\sum_{a=1}^K\frac{\beta_a^2}{m}\sum_{i\in I_a}\left(\mathbf{e}_k^\top\Theta^+\mathbf{e}_i\right)^2\\
&\leq\widehat{\tau}^4\frac{\|\widehat{\beta}\|_2^2}{\widehat{m}}\max_{a\in[K]}\sum_{i\in I_a}\left|\left(\mathbf{e}_k^\top\widehat{\Theta}^+\mathbf{e}_i\right)^2-\left(\mathbf{e}_k^\top\Theta^+\mathbf{e}_i\right)^2\right|\\
&\quad+\widehat{\tau}^4\max_{a\in[K]}\left|\frac{\widehat{\beta}_a^2}{\widehat{m}}-\frac{\beta_a^2}{m}\right|\sum_{a=1}^K\sum_{i\in I_a}\left(\mathbf{e}_k^\top\Theta^+\mathbf{e}_i\right)^2+\left|\frac{\widehat{\tau}^4}{\widehat{m}-1}-\frac{\tau^4}{m-1}\right|\sum_{a=1}^K\frac{\beta_a^2}{m}\sum_{i\in I_a}\left(\mathbf{e}_k^\top\Theta^+\mathbf{e}_i\right)^2.
\end{aligned}
$$

We bound each term separately. First, note that Theorem 4.2 guarantees $\widehat{m}\geq m$ and (96) yields

$$\widehat{\tau}^4\leq\tau^4+(\tau^2+\widehat{\tau}^2)|\widehat{\tau}^2-\tau^2|=O_p(\tau^4).$$

Recall from (79) that

$$|\mathbf{e}_k^\top\Theta^+\mathbf{e}_i|\leq\sqrt{\frac{\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k}{m}}.\tag{132}$$

Provided that

$$\max_{i\in I}\left|\mathbf{e}_k^\top\widehat{\Theta}^+\mathbf{e}_i-\mathbf{e}_k^\top\Theta^+\mathbf{e}_i\right|=o_p\left(\sqrt{\frac{\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k}{m}}\right),\tag{133}$$

we can conclude

$$
\begin{aligned}
&\widehat{\tau}^4\frac{\|\widehat{\beta}\|_2^2}{\widehat{m}}\max_{1\leq a\leq K}\sum_{i\in I_a}\left|\left(\mathbf{e}_k^\top\widehat{\Theta}^+\mathbf{e}_i\right)^2-\left(\mathbf{e}_k^\top\Theta^+\mathbf{e}_i\right)^2\right|\\
&\leq\widehat{\tau}^4\frac{\|\widehat{\beta}\|_2^2}{\widehat{m}}\cdot m\cdot\max_{i\in I}\left(\left|\mathbf{e}_k^\top\widehat{\Theta}^+e_i\right|+\left|\mathbf{e}_k^\top\Theta^+\mathbf{e}_i\right|\right)\left|\mathbf{e}_k^\top\widehat{\Theta}^+\mathbf{e}_i-\mathbf{e}_k^\top\Theta^+\mathbf{e}_i\right|\\
&=o_p\left(\tau^4\frac{\|\beta\|_2^2}{m}\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k\right)=o_p(\Delta_a\Delta_b).
\end{aligned}
$$

For the other two terms, note that

$$\left|\frac{\widehat{\beta}_a^2}{\widehat{m}} - \frac{\beta_a^2}{m}\right| \le (\|\widehat{\beta}\|_2 + \|\beta\|_2)\frac{\|\widehat{\beta} - \beta\|_2}{\widehat{m}} + \|\beta\|_2^2\frac{|m - \widehat{m}|}{\widehat{m}m}$$
$$= O_p\left(\left(1 \vee \frac{\|\beta\|_2^2}{m}\right)\left(C_{\min}^{-1/2}\delta_n\sqrt{K} + \bar{\rho}\right)\right) = o_p(\Delta_a)$$

by using (96), $|m - \widehat{m}|/\widehat{m} \le \bar{\rho}$, (22) and $\delta_n\sqrt{K} = o(1)$. Moreover, $\widehat{\tau}^2 = O_p(\tau^2)$, $\widehat{m} \ge 2$, $m \ge 2$, $|m - \widehat{m}|/\widehat{m} \le \bar{\rho}$ and Lemma 21 yield

$$\left|\frac{\widehat{\tau}^4}{\widehat{m} - 1} - \frac{\tau^4}{m - 1}\right| \le \frac{(\widehat{\tau}^2 + \tau^2)|\widehat{\tau}^2 - \tau^2|}{\widehat{m} - 1} + \frac{\tau^4|m - \widehat{m}|}{(\widehat{m} - 1)(m - 1)} = o_p(1).$$

By observing

$$\sum_{a=1}^K \frac{\beta_a^2}{m}\sum_{i\in I_a}\left(\mathbf{e}_k^\top\Theta^+\mathbf{e}_i\right)^2 \le \frac{\|\beta\|_2^2}{m}\left(\mathbf{e}_k^\top\Theta^+\mathbf{e}_i\right)^2 \overset{(132)}{\le} \frac{\|\beta\|_2^2}{m}\frac{\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k}{m},$$

and also using (132), we conclude

$$\widehat{\tau}^4 \max_a \left|\frac{\widehat{\beta}_a^2}{\widehat{m}} - \frac{\beta_a^2}{m}\right| \sum_{a=1}^K \sum_{i\in I_a}\left(\mathbf{e}_k^\top\Theta^+\mathbf{e}_i\right)^2 = o_p(\Delta_a\Delta_b),$$
$$\left|\frac{\widehat{\tau}^4}{\widehat{m} - 1} - \frac{\tau^4}{m - 1}\right| \sum_{a=1}^K \frac{\beta_a^2}{m}\sum_{i\in I_a}\left(\mathbf{e}_k^\top\Theta^+\mathbf{e}_i\right)^2 = o_p(\Delta_a\Delta_b).$$

It then suffices to verify (133). By (66), for any $i \in I$, we have

$$\left|\mathbf{e}_k^\top(\widehat{\Theta}^+ - \Theta^+)\mathbf{e}_i\right| \le \left|\mathbf{e}_k^\top(\Theta^T\Theta)^{-1}(\widehat{\Theta} - \Theta)^T P_{\widehat{\Theta}}^\perp\mathbf{e}_i\right| + \left|\mathbf{e}_k^\top\Theta^+(\Theta - \widehat{\Theta})\widehat{\Theta}^+\mathbf{e}_i\right|$$
$$\le \left\|\mathbf{e}_k^\top\Omega^{1/2}(H^TH)^{-1}(\widehat{H} - H)^T\right\|_2 + \left|\mathbf{e}_k^\top\Omega^{1/2}H^+(H - \widehat{H})\widehat{H}^+\mathbf{e}_i\right|$$
$$\le \sqrt{p}\max_j\left|\mathbf{e}_k^\top\Omega^{1/2}(H^TH)^{-1}(\widehat{H} - H)^T\mathbf{e}_j\right| + \Omega_{kk}^{1/2}\frac{\|H^+(\widehat{H} - H)\|_{op}}{\sigma_K(\widehat{H})}.$$

Invoking (67) and part (d) of Lemma 10, we gives

$$\left|\mathbf{e}_k^\top(\widehat{\Theta}^+ - \Theta^+)\mathbf{e}_i\right| \lesssim \sqrt{p}\delta_n\sqrt{\mathbf{e}_k^\top\Omega^{1/2}(H^\top H)^{-2}\Omega^{1/2}\mathbf{e}_k} + \Omega_{kk}^{1/2}\frac{\delta_n\sqrt{K}}{\sigma_K(H)}$$
$$\lesssim \delta_n\sqrt{\frac{p}{\lambda_K(H^\top H)}}\sqrt{\mathbf{e}_k^\top(\Theta^\top\Theta)^{-1}\mathbf{e}_k} + \Omega_{kk}^{1/2}\frac{\delta_n\sqrt{K}}{\sigma_K(H)}$$

with probability $1 - c(p\vee n)^{-c'}$. By $\sigma_K^2(H) = \lambda_K(H^\top H) \ge mC_{\min}$, invoking Assumption 5′ concludes the proof. □

Appendix G: Theoretical guarantees of $\widehat{\boldsymbol{\beta}}^{(I)}$: convergence rate and asymptotic normality

We provide statistical guarantees for the estimator defined in (27) with $\widehat{\Sigma}_Z$ defined in (8) and $\widehat{A}_{\widehat{I}\bullet}$ obtained from (9) – (10). Their proofs can be found in [1]. The following theorem states the convergence rate of $\min_{P\in\mathcal{H}_K}\|\widehat{\beta}^{(I)}-P\beta\|_2$.

Theorem 30. *Assume Assumptions 1 – 4 hold. Let $K\log(p\vee n)\le cn$ for some sufficiently small constant $c>0$. Then, with probability greater than $1-(p\vee n)^{-c'}$ for some constant $c'>0$, $\widehat{K}=K$, the matrix $\widehat{\Sigma}_Z$ is non-singular and the estimator $\widehat{\beta}_{(I)}$ given by (27) satisfies:*

$$\min_{P\in\mathcal{H}_K}\left\|\widehat{\beta}^{(I)}-P\beta\right\|_2\lesssim\left(1\vee\frac{\|\beta\|_2}{\sqrt{m}}\right)\sqrt{\frac{K\log(p\vee n)}{n}}. \tag{134}$$

The following theorem establishes the asymptotic normality of each coordinate of $\widehat{\beta}^{(I)}$. For ease of the presentation, we assume $\Gamma=\tau^2\mathbf{I}_p$ and $|I_1|=\ldots=|I_K|=m$ while the proof holds for the general case.

Theorem 31. *Under Assumptions 1, 2, 3′, 4 and $K\log(p\vee n)=o(\sqrt{n})$, assume $\gamma_\varepsilon/\sigma=O(1)$ and $\gamma_w/\tau=O(1)$. Then $\widehat{\Sigma}_Z$ is non-singular with probability tending to one, and for any $1\le k\le K$,*

$$\sqrt{n/U_k}\left(\widehat{\beta}_k^{(I)}-\beta_k\right)\xrightarrow{d}N(0,1),\qquad as\quad n\to\infty,$$

where

$$U_k=\left(\sigma^2+\frac{\tau^2}{m}\|\beta\|_2^2\right)\left(\Omega_{kk}+\frac{\tau^2}{m}\|\Omega_{k\bullet}\|_2^2\right)+\frac{\tau^4}{m^2(m-1)}\sum_{a=1}^{K}\beta_a^2\Omega_{ka}^2.$$

To estimate the asymptotic variance U_k, we can also use a plug-in estimator by using $|\widehat{I}_k|$ for each $1\le k\le K$, $\widehat{\Sigma}_Z^{-1}$, $\widehat{\beta}^{(I)}$, $\widehat{\tau}_i^2$ for $1\le i\le p$ obtained as (30) and $\widehat{\sigma}^2$ obtained as (31) by using $\widehat{\beta}^{(I)}$. The following proposition shows that the plug-in estimator $\widehat{U}_k$ consistently estimates the asymptotic variance U_k of $\widehat{\beta}_k^{(I)}$.

Proposition 32. *Under conditions of Theorem 31, we have*

$$\left|\widehat{U}_k^{1/2}/U_k^{1/2}-1\right|=o_p(1).$$

Consequently, we have

$$\sqrt{n/\widehat{U}_k}\left(\widehat{\beta}_k^{(I)}-\beta_k\right)\xrightarrow{d}N(0,1),\qquad as\ n\to\infty,\quad k\in[K].$$

Appendix H: Data-driven choice of the tuning parameter δ in Algorithm 1

A selection procedure of choosing δ is proposed in [2, Section 5.1]. For the reader's convenience, we restate it here as well as an illustrative example in [2].

Display (17) specifies the theoretical rate of δ, but only up to constants that depend on the underlying data generating mechanism. We propose below a data-dependent way to select δ, based on data splitting. Specifically, we split the data set into two independent parts, of equal sizes. On the first set, we calculate the sample covariance matrix $\widehat{\Sigma}^{(1)}$. On the second set, we choose a fine grid of values $\delta_\ell = c_\ell\sqrt{\log p/n}$, with $1 \le \ell \le M$, for δ, by varying the proportionality constants c_ℓ. For each δ_ℓ, we obtain the estimated number of clusters $\widehat{K}(\ell)$ and the pure variable set $\widehat{I}(\ell)$ with its partition $\widehat{\mathcal{I}}(\ell)$. Then we construct the $|\widehat{I}(\ell)| \times \widehat{K}(\ell)$ submatrix $\widehat{A}_{\widehat{I}(\ell)}$ of $\widehat{A}$, and estimate $\widehat{\Sigma}_Z(\ell)$ via formula (8). Finally, we calculate the $|\widehat{I}(\ell)| \times |\widehat{I}(\ell)|$ matrix $W_\ell = \widehat{A}_{\widehat{I}(\ell)}\widehat{\Sigma}_Z(\ell)\widehat{A}^T_{\widehat{I}(\ell)}$. In the end, we have constructed a family $\mathcal{F} = \{W_1, \ldots, W_M\}$ of the fitted matrices W_ℓ, each corresponding to different $\widehat{\mathcal{I}}(\ell)$ that depend in turn on δ_ℓ, for $\ell \in \{1, \ldots, M\}$. Define

$$CV(\widehat{\mathcal{I}}(\ell)) := \frac{1}{\sqrt{|\widehat{I}(\ell)|(|\widehat{I}(\ell)|-1)}} \left\| \widehat{\Sigma}^{(1)}_{\widehat{I}(\ell)\widehat{I}(\ell)} - W_\ell \right\|_{\text{F-off}}, \tag{135}$$

where $\|B\|_{\text{F-off}} := \|B - \text{diag}(B)\|_F$ denotes the Frobenius norm over the off-diagonal elements of a square matrix B. We choose δ^{cv} as the value δ_ℓ that minimizes $CV(\widehat{\mathcal{I}}(\ell))$ over the grid $\ell \in [M]$. To illustrate how the selection procedure works, we provide an example below.

We consider a simple case, when Σ_Z is diagonal and the signed permutation matrix P is the identity, to illustrate our cross-validation method.

Example 1. Let $\Sigma_Z = \text{diag}(\tau, \tau, \tau)$, $\mathcal{I} = \big\{\{1,2\},\{3,4\},\{5,6\}\big\}$ and

$$A = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & -1 \\ 0.4 & 0.6 & 0 \\ -0.5 & 0 & 0.4 \end{bmatrix}, \quad A_I \Sigma_Z A_I^T = \begin{bmatrix} * & \tau & 0 & 0 & 0 & 0 \\ \tau & * & 0 & 0 & 0 & 0 \\ 0 & 0 & * & \tau & 0 & 0 \\ 0 & 0 & \tau & * & 0 & 0 \\ 0 & 0 & 0 & 0 & * & \tau \\ 0 & 0 & 0 & 0 & \tau & * \end{bmatrix},$$

where we use $*$ to reflect the fact that our algorithm ignores the diagonal elements. For

the true I and $\mathcal{I}$, we have $\widehat{A}_I = A_I$,

$$\begin{aligned}\left\|\widehat{\Sigma}^{(1)}_{II} - A_I\widehat{\Sigma}_Z A_I^T\right\|_{\text{F-off}} &\le \left\|\widehat{\Sigma}^{(1)}_{II} - \Sigma_{II}\right\|_{\text{F-off}} + \left\|A_I\widehat{\Sigma}_Z A_I^T - \Sigma_{II}\right\|_{\text{F-off}} \\ &\le \left\|\widehat{\Sigma}^{(1)}_{II} - \Sigma_{II}\right\|_{\text{F-off}} + \sqrt{|I|(|I|-1)}\cdot\|\widehat{\Sigma}_Z - \Sigma_Z\|_\infty.\end{aligned}$$

For

$$\epsilon = \left(\max_{i\neq j}\left|\widehat{\Sigma}^{(1)}_{ij} - \Sigma_{ij}\right|\right) \vee \left(\max_{i\neq j}\left|\widehat{\Sigma}^{(2)}_{ij} - \Sigma_{ij}\right|\right),$$

we obtain

$$CV(\mathcal{I}) = \frac{1}{\sqrt{|I|\big(|I|-1\big)}}\left\|\widehat{\Sigma}^{(1)}_{II} - A_I\widehat{\Sigma}_Z A_I^T\right\|_{\text{F-off}} \le 2\epsilon.$$

Suppose that $\widehat{\mathcal{I}} = \big\{\{1,2\},\{3,5\},\{4,6\}\big\}$, so $\widehat{I} = I$, yet $\widehat{\mathcal{I}} \neq \mathcal{I}$, we would have

$$\widehat{A}_{\widehat{I}}\widehat{\Sigma}_Z\widehat{A}_{\widehat{I}}^T = \begin{bmatrix} * & \widehat{\tau}_1 & 0 & 0 & 0 & 0 \\ \widehat{\tau}_1 & * & 0 & 0 & 0 & 0 \\ 0 & 0 & * & 0 & \widehat{\tau}_2 & 0 \\ 0 & 0 & 0 & * & 0 & \widehat{\tau}_3 \\ 0 & 0 & \widehat{\tau}_2 & 0 & * & 0 \\ 0 & 0 & 0 & \widehat{\tau}_3 & 0 & * \end{bmatrix}$$

and

$$\widehat{A}_{\widehat{I}}\widehat{C}\widehat{A}_{\widehat{I}}^T - \Sigma_{\widehat{I}\widehat{I}} = \begin{bmatrix} * & \Delta\tau_1 & 0 & 0 & 0 & 0 \\ \Delta\tau_1 & * & 0 & 0 & 0 & 0 \\ 0 & 0 & * & \boldsymbol{-\tau} & \boldsymbol{\widehat{\tau}_2} & 0 \\ 0 & 0 & \boldsymbol{-\tau} & * & 0 & \boldsymbol{\widehat{\tau}_3} \\ 0 & 0 & \boldsymbol{\widehat{\tau}_2} & 0 & * & \boldsymbol{-\tau} \\ 0 & 0 & 0 & \boldsymbol{\widehat{\tau}_3} & \boldsymbol{-\tau} & * \end{bmatrix}.$$

Here $\Delta\tau_a = \widehat{\tau}_a - \tau_a$, using estimates $\widehat{\tau}_a$ defined in lieu of $[\widehat{\Sigma}_Z]_{aa}$ from (8) for each $a \in [\widehat{K}]$. Thus, the cross-validation criterion in (135) would satisfy

$$\begin{aligned}CV(\widehat{\mathcal{I}}) &\ge \frac{1}{\sqrt{|\widehat{I}|\big(|\widehat{I}|-1\big)}}\left\|\widehat{A}_{\widehat{I}}\widehat{\Sigma}_Z\widehat{A}_{\widehat{I}}^T - \Sigma_{\widehat{I}\widehat{I}}\right\|_{\text{F}-\text{off}} - \frac{1}{\sqrt{|\widehat{I}|\big(|\widehat{I}|-1\big)}}\left\|\widehat{\Sigma}^{(1)}_{\widehat{I}\widehat{I}} - \Sigma_{\widehat{I}\widehat{I}}\right\|_{\text{F-off}} \\ &\ge \sqrt{\frac{4\tau^2 + 2\widehat{\tau}_2^2 + 2\widehat{\tau}_3^2}{|\widehat{I}|\big(|\widehat{I}|-1\big)}} - 2\epsilon.\end{aligned}$$

From noting that $|\widehat{\tau}_a - \tau| \le \epsilon$, for $a = 2, 3$, it gives

$$CV(\widehat{\mathcal{I}}) \ge \sqrt{\frac{4\tau^2 - 4\tau\epsilon + 2\epsilon^2}{15}} - 2\epsilon > 2\epsilon \ge CV(\mathcal{I}),$$

for $\tau \geq 9\epsilon$. We conclude in this example, with $\widehat{I} = I$, incorrectly specifying $\mathcal{I}$ will induce a large loss. It is easily verified that this is also the case when $\widehat{I} = I$ but $\widehat{K} \neq K$ and $\widehat{\mathcal{I}} \neq \mathcal{I}$.

On the other hand, suppose we mistakenly included some non-pure variable in $\widehat{I}$. For instance, suppose we found $\widehat{\mathcal{I}} = \big\{\{1,2\},\{3,4\},\{5,6,7\}\big\}$. Then we would have

$$\Sigma_{\widehat{I}'\widehat{I}'} = \begin{bmatrix} * & \tau & 0 & 0 & 0 & 0 & 0.4\tau \\ \tau & * & 0 & 0 & 0 & 0 & -0.4\tau \\ 0 & 0 & * & \tau & 0 & 0 & 0.6\tau \\ 0 & 0 & \tau & * & 0 & 0 & 0.6\tau \\ 0 & 0 & 0 & 0 & * & \tau & 0 \\ 0 & 0 & 0 & 0 & \tau & * & 0 \\ 0.4\tau & -0.4\tau & 0.6\tau & 0.6\tau & 0 & 0 & * \end{bmatrix}$$

and

$$\widehat{A}_{\widehat{I}'}\widehat{\Sigma}_Z\widehat{A}_{\widehat{I}'}^T = \begin{bmatrix} * & \widehat{\tau}_1 & 0 & 0 & 0 & 0 & 0 \\ \widehat{\tau}_1 & * & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & * & \widehat{\tau}_2 & 0 & 0 & 0 \\ 0 & 0 & \widehat{\tau}_2 & * & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & * & \widehat{\tau}_3 & \widehat{\tau}_3 \\ 0 & 0 & 0 & 0 & \widehat{\tau}_3 & * & \widehat{\tau}_3 \\ 0 & 0 & 0 & 0 & \widehat{\tau}_3 & \widehat{\tau}_3 & * \end{bmatrix}.$$

We thus have

$$\widehat{A}_{\widehat{I}'}\widehat{\Sigma}_Z\widehat{A}_{\widehat{I}'}^T - \Sigma_{\widehat{I}'\widehat{I}'} = \begin{bmatrix} * & \Delta\tau_1 & 0 & 0 & 0 & 0 & \boldsymbol{-0.4\tau} \\ \Delta\tau_1 & * & 0 & 0 & 0 & 0 & \boldsymbol{0.4\tau} \\ 0 & 0 & * & \Delta\tau_2 & 0 & 0 & \boldsymbol{-0.6\tau} \\ 0 & 0 & \Delta\tau_2 & * & 0 & 0 & \boldsymbol{-0.6\tau} \\ 0 & 0 & 0 & 0 & * & \Delta\tau_3 & \boldsymbol{\widehat{\tau}_3} \\ 0 & 0 & 0 & 0 & \Delta\tau_3 & * & \boldsymbol{\widehat{\tau}_3} \\ \boldsymbol{-0.4\tau} & \boldsymbol{0.4\tau} & \boldsymbol{-0.6\tau} & \boldsymbol{-0.6\tau} & \boldsymbol{\widehat{\tau}_3} & \boldsymbol{\widehat{\tau}_3} & * \end{bmatrix}$$

and, by similar arguments, for $\tau \geq 12\epsilon$, we find

$$CV(\widehat{\mathcal{I}}') \geq \sqrt{\frac{4\widehat{\tau}_3^2 + 4 \times 0.36\tau^2 + 4 \times 0.16\tau^2}{42}} - 2\epsilon > 2\epsilon.$$

Thus, the cross-validation loss in this example will be large even if only one non-pure variable is mistakenly classified as pure variable. In rare cases, the cross-validation criterion might miss a very small subset of I but this can be rectified in our later estimation of A_J.

Appendix I: Simulations

In this section, we complement and support our theoretical findings with simulations, focusing on the ℓ_2 convergence rate of $\widehat{\beta}$ and on the attained coverage of the corresponding

95% confidence interval (CI) for β. Additional simulation results are provided in Appendix J to investigate the impact of a potential inconsistent estimation of the number of factors on subsequent estimation steps, and on inference for β.

Data generating mechanism: We first describe how we generate A, Σ_Z, Γ, and β. Recall that A can be partitioned into $A_{I\bullet}$ and $A_{J\bullet}$. To generate $A_{I\bullet}$, we set $|I_k| = m$ for each $k \in [K]$ and choose $A_{I\bullet} = \mathbf{I}_K \otimes \mathbf{1}_m$, where $\otimes$ denotes the Kronecker product. Each row of $A_{J\bullet}$ is generated by first randomly selecting its support with cardinality drawn from $\{2, 3 \ldots, K\}$ and then by sampling its non-zero entries from Uniform$(0, 1)$. In the end, we rescale $A_{J\bullet}$ such that the ℓ_1-norm of each row is no greater than 1. We multiply all entries in $A_{J\bullet}$ by independent random signs. To generate Σ_Z, we range its K diagonal entries from 2.5 to 3 with equal increments. The off-diagonal elements of Σ_Z are then chosen as $[\Sigma_Z]_{ij} = (-1)^{(i+j)}([\Sigma_Z]_{ii} \wedge [\Sigma_Z]_{jj})(0.3)^{|i-j|}$ for any $i \neq j \in [K]$. Finally, Γ is chosen by randomly sampling its diagonal elements from Unif$(1, 3)$.

Next, we generate the $n \times K$ matrix $\mathbf{Z}$ and the $n \times p$ noise matrix $\mathbf{W}$ by generating i.i.d. rows from $N_K(0, \Sigma_Z)$ and $N_p(0, \Gamma)$, respectively. Finally, we set $\mathbf{X} = \mathbf{Z}A^\top + \mathbf{W}$ and $\mathbf{y} = \mathbf{Z}\beta + \varepsilon$ where the n components of ε are i.i.d. $N(0, 1)$. For each setting, we repeat generating pairs $(\mathbf{X}, \mathbf{y})$ 200 times and record the corresponding results.

Convergence rate and the coverage of the 95% confidence interval

We consider the following four settings:

(1) Fix $p = 400$, $K = 10$, $m = 5$, and vary $n \in \{200, 400, 600, 800\}$;
(2) Fix $n = 300$, $K = 10$, $m = 5$, and vary $p \in \{100, 300, 500, 700\}$;
(3) Fix $n = 300$, $p = 400$, $m = 5$ and vary $K \in \{5, 10, 15, 20\}$;
(4) Fix $n = 300$, $p = 400$, $K = 10$ and vary $m \in \{2, 5, 10, 15, 20\}$.

The entries of β are independently sampled from Unif$(1, 3)$. For each setting, we calculate the averaged ℓ_2 errors $\|\widehat{\beta} - \beta\|_2$ of the following four estimators, the proposed estimator and three other estimates, and report them in Table 1:

- Our proposed estimator: $\widehat{\beta}$ constructed in (13);
- $\widehat{\beta}^{(A)}$ defined in (25);
- $\widehat{\beta}^{(I)}$ defined in (27);
- $\widehat{\beta}_{\text{naive}}$ obtained by naively regressing $\mathbf{y}$ on $\bar{\mathbf{X}} = \mathbf{X}\widehat{A}[\widehat{A}^\top \widehat{A}]^{-1}$;
- $\widehat{\beta}_{\text{oracle}} = (\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{y}$, the oracle least squares estimator based on the true matrix $\mathbf{Z}$.

We focus on the best feasible estimators, with respect to their respective ℓ_2 error, $\widehat{\beta}$ and $\widehat{\beta}^{(I)}$, and check the coverage of their corresponding 95% confidence intervals (CI). Table 1 shows the coverage and average length of the CI's for β_1, respectively constructed from $\widehat{\beta}$, based on Theorem 4 and Proposition 5 and $\widehat{\beta}^{(I)}$, based on Theorem 31 and Proposition 32 in the Appendix.

Summary: As expected, the oracle estimator $\widehat{\beta}_{\text{oracle}}$ is the best performer, since it uses the true $\mathbf{Z}$, not available to the other estimates. Among the remaining estimators, our proposed estimator $\widehat{\beta}$ outperforms the other three estimators in all cases. The gap between $\widehat{\beta}$ and $\widehat{\beta}_{\text{oracle}}$ decreases as either n or m increases or K decreases. We find that $\widehat{\beta}^{(I)}$ has better performance than $\widehat{\beta}_{\text{naive}}$ and $\widehat{\beta}^{(A)}$. While dominated by $\widehat{\beta}_{\text{oracle}}$ and $\widehat{\beta}$, we find that for large m, $\widehat{\beta}^{(I)}$ performs similarly to $\widehat{\beta}$. Overall, the naive estimator $\widehat{\beta}_{\text{naive}}$ has the worst performance, which supports our findings in Section 5.

The estimation errors of $\widehat{\beta}$ and $\widehat{\beta}^{(I)}$ decrease as n and/or m increase. The estimation errors increase in K. It is worth mentioning that increasing p barely affects the estimation errors. These findings support Theorem 3.

Regarding the CIs of β based on $\widehat{\beta}$, the average coverage, over 200 repetitions, are close to the nominal 95% in most settings, especially for moderately large sample size n. This further supports the results of Section 4.5. The coverage level of the intervals based on $\widehat{\beta}$ are closer to the 95% level than those based on $\widehat{\beta}^{(I)}$. The averaged CI lengths corresponding to $\widehat{\beta}$ are also smaller than those relative to $\widehat{\beta}^{(I)}$, in most of the settings we considered. This suggests that $\widehat{\beta}$ is more efficient than $\widehat{\beta}^{(I)}$. We further corroborate the validity of Theorem 4 and Proposition 5 in Figure 1. This figure depicts histograms based on 200 values of $\sqrt{n/\widehat{V}_k}(\widehat{\beta}_1 - \beta_1)$.

Appendix J: Additional simulation results: estimation and inference of $\boldsymbol{\beta}$ when the number of latent factors is not consistently estimated

We discuss the impact of selecting $\widehat{K}$ with $\widehat{K} \neq K$ on the estimation and inference of β, particularizing to the case when some columns of A have very weak signals such that our estimate $\widehat{K}$ is smaller than the true K. We start by offering some intuition here. Intuitively, when the submatrix $A_{\cdot S^c}$ contains many zero entries for some index set $S \subseteq [K]$, our procedure of estimating K is likely to miss the latent factors in S^c, but may still recover Z_S. As a result, only $A_{\cdot S}$ can be well estimated. Consider the simple case $\Sigma_Z = \mathbf{I_K}$ and recall that $\beta = (A^T A)^{-1} A^T \Sigma_{XY}$. Replacing A by $A_{\cdot S}$ yields

$$(A_{\cdot S}^T A_{\cdot S})^{-1} A_{\cdot S}^T \Sigma_{XY} = (A_{\cdot S}^T A_{\cdot S})^{-1} A_{\cdot S}^T A \beta = \beta_S + \underbrace{(A_{\cdot S}^T A_{\cdot S})^{-1} A_{\cdot S}^T A_{\cdot S^c} \beta_{S^c}}_{\Delta}.$$

Therefore, when $A_{\cdot S^c}$ is small such that Δ is small, we still have

$$(A_{\cdot S}^T A_{\cdot S})^{-1} A_{\cdot S}^T \Sigma_{XY} \;\approx\; \beta_S.$$

When Σ_Z is not diagonal, the dependence among the latent factors also affects this approximation.

	$\widehat{\beta}$	$\widehat{\beta}^{(I)}$	$\widehat{\beta}_{\text{naive}}$	$\widehat{\beta}^{(A)}$	$\widehat{\beta}_{\text{oracle}}$	CIs of $\widehat{\beta}$		CIs of $\widehat{\beta}^{(I)}$	
Vary n **with** $p = 400$, $K = 10$, $m = 5$						coverage	length	coverage	length
$n = 200$	0.045	0.052	0.204	0.106	0.002	91.0	0.76	92.5	0.85
$n = 400$	0.019	0.023	0.135	0.052	0.001	95.0	0.53	93.5	0.58
$n = 600$	0.013	0.016	0.112	0.036	0.001	93.5	0.44	94.0	0.47
$n = 800$	0.009	0.011	0.101	0.029	0.001	94.0	0.38	93.5	0.40
Vary p **with** $n = 300$, $K = 10$, $m = 5$									
$p = 100$	0.029	0.035	0.183	0.029	0.002	94.5	0.69	92.5	0.72
$p = 300$	0.029	0.035	0.181	0.067	0.002	94.0	0.64	92.0	0.69
$p = 500$	0.031	0.039	0.179	0.088	0.002	95.0	0.65	95.0	0.72
$p = 700$	0.030	0.037	0.176	0.103	0.001	94.5	0.63	94.0	0.70
Vary K **with** $n = 300$, $p = 400$, $m = 5$									
$K = 5$	0.016	0.020	0.064	0.041	0.001	91.0	0.44	91.0	0.45
$K = 10$	0.026	0.032	0.177	0.080	0.001	94.5	0.63	92.0	0.68
$K = 15$	0.052	0.065	0.282	0.099	0.002	93.0	0.88	94.0	0.96
$K = 20$	0.046	0.057	0.192	0.064	0.002	97.5	0.81	96.5	0.89
Vary m **with** $n = 300$, $p = 400$, $K = 10$									
$m = 2$	0.116	0.191	0.301	0.171	0.002	90.1	1.02	90.1	1.34
$m = 5$	0.030	0.035	0.173	0.080	0.001	93.5	0.65	91.0	0.70
$m = 10$	0.015	0.016	0.097	0.036	0.002	94.0	0.48	93.0	0.48
$m = 15$	0.011	0.012	0.060	0.022	0.002	92.0	0.40	90.5	0.39
$m = 20$	0.008	0.008	0.035	0.012	0.001	96.5	0.35	96.0	0.34

Table 1. ℓ_2 error of different estimators and the coverages and the averaged lengths of the 95% CIs of β_1.

To empirically investigate the impact of the estimating error of $\widehat{K}$ on the subsequent inferential result, we conduct the following two simulation studies.

(1) We take the same generating mechanism as described in Section 6 and consider $p = 400$, $n = 300$, $K = 10$ and $m = 5$. The matrix Σ_Z is set to be $\sigma_Z^2 \mathbf{I_K}$ with $\sigma_Z^2 = 3$. After generating the matrix A, we draw an index set from p i.i.d. Bernoulli$(1-\theta)$ with $\theta \in (0,1)$, and manually change the entries A_{jK} into zero, for all j in this index set. The parameter θ controls the overall sparsity (or signal) of the Kth column of A. Small values of θ correspond to small signal $A_{\cdot K}$. Note that A no longer necessarily meets our model requirement (A1), since the entries in the Kth column of A corresponding to pure variables are not necessarily 1, and allowed to be set to zero. We allowed for this misspecification since we empirically found that as long as there exists at least two entries in the Kth column of A corresponding to pure variables (even though the rest

entries of $A_{\cdot K}$ are zero), our algorithm continues to consistently estimate K.

Our goal is to verify if the resulting estimator $\widehat{\beta}$ estimates β well expect for the Kth coordinate and to examine the empirical coverage of the 95% confidence intervals of our proposed estimator. The same estimators mentioned in Section 6 of the main paper are considered. For the oracle estimator $\widehat{\beta}_{\text{oracle}}$, to illustrate the effect of using a subset of Z, we change it to

$$\widehat{\beta}_{\text{oracle}} = \left(\mathbf{Z}_{\cdot\widehat{K}}^T \mathbf{Z}_{\cdot\widehat{K}}\right)^{-1} \mathbf{Z}_{\cdot\widehat{K}}^T \mathbf{Y}$$

where $\widehat{K}$ is estimated from Algorithm 1. We vary $\theta \in \{0.1, 0.3, 0.5, 0.8\}$ and, for each setting, we calculate the averaged mean squared error of the subvector $[\widehat{\beta} - \beta]_{S(1)}$, with $S(1) := \{1, 2, \ldots, K-1\}$, for different estimators, as well as the averaged coverage of the 95% confidence intervals for β_1, over 200 repetitions. These results together with the averaged estimated K are reported in Table 2.

Result: When θ is small, that is, $A_{\cdot K}$ has very weak signal, our procedure of estimating K is likely to miss the Kth latent factor, leading to $\widehat{K} = 9$, as expected. The estimation of the first $K-1$ entries of β is only slightly affected. Our proposed estimator $\widehat{\beta}$ has the second best performance, after the oracle estimator. The coverage of the 95% confidence intervals constructed from $\widehat{\beta}$ is slightly under the nominal 95% level. As θ increases ($\theta \geq 0.5$), that is, the signal of $A_{\cdot K}$ gets larger, we consistently estimate K, and consequently, our proposed estimator performs increasingly better in terms of both the estimation error and the coverage of 95% confidence intervals.

(2) We now consider the case when there are multiple, sparse columns of A. We take the same setting $p = 400$, $n = 300$, $K = 10$, $m = 5$ and $\Sigma_Z = \sigma_Z^2 \mathbf{I_K}$ with $\sigma_Z^2 = 3$. For each $n_A \in \{1, 2, \ldots, 5\}$, we threshold the last n_A columns of A according to the procedure described in the previous paragraph with $\theta = 0.1$. For each n_A, let $S(n_A) = \{1, 2, \ldots, K - n_A\}$. The averaged mean squared error of $[\widehat{\beta} - \beta]_{S(n_A)}$, the averaged coverage of the 95% confidence intervals for β_1 and the averaged estimated K are reported in Table 2.

Result: As expected, the estimated number of factors is smaller than the true value, 10, and is close to $K - n_A$, the number of columns of A that have strong signals. As long as $n_A \leq 4$, the estimation of $\beta_{S(n_A)}$ seems fairly good, and $\widehat{\beta}$ still has the best performance (after $\widehat{\beta}_{\text{oracle}}$). The coverage of the 95% confidence intervals constructed from $\widehat{\beta}$ is still close to, though slightly lower than, 95%. As n_A increases, more latent factors have weak signals. This makes estimation of K more difficult and clearly affects the estimation of β and the coverage of confidence intervals.

(3) We further investigate the effect of the correlation of Z on the impact of inconsistently estimating K. We choose Σ_Z as $[\Sigma_Z]_{ij} = \sigma_Z^2 (-1)^{i+j} \rho^{|i-j|}$ for each $i, j \in [K]$ with $\sigma_Z^2 = 3$. We vary $\rho \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and fix $\theta = 0.1$, $n_A = 1$, $p = 400$, $n = 300$, $K = 10$ and $m = 5$. The averaged mean squared error of the subvector $[\widehat{\beta} - \beta]_{S(1)}$, the averaged coverage of the 95% confidence intervals for β_1 and the averaged estimated K are reported in Table 2.

Result: The estimation errors of all estimators (including $\widehat{\beta}_{\text{oracle}}$) increase as ρ gets larger. $\widehat{\beta}^{(I)}$ turns out to be more robust than $\widehat{\beta}$ in the presence of correlated factors. One possible explanation is that $\widehat{\beta}^{(I)}$ targets $\Sigma_Z^{-1}(A_{I\cdot}^T A_{I\cdot})^{-1}A_{I\cdot}^T\text{Cov}(X_I, Y)$ which, due to the diagonal structure of $A_{I\cdot}^T A_{I\cdot}$, is less affected by a non-diagonal Σ_Z compared to $\widehat{\beta}$. The coverage of the 95% confidence intervals constructed from both $\widehat{\beta}$ and $\widehat{\beta}^{(I)}$ is close to 95%, and slightly decreases as ρ gets larger.

	$\widehat{\beta}$	$\widehat{\beta}^{(I)}$	$\widehat{\beta}_{\text{naive}}$	$\widehat{\beta}^{(A)}$	$\widehat{\beta}_{\text{oracle}}$	CIs of $\widehat{\beta}$		CIs of $\widehat{\beta}^{(I)}$		$\widehat{K}$
Vary θ with $n = 300$, $p = 400$, $K = 10$, $m = 5$						coverage	length	coverage	length	
$\theta = 0.1$	0.07	0.08	0.12	0.08	0.02	0.90	0.84	0.91	0.88	9.1
$\theta = 0.3$	0.06	0.06	0.11	0.07	0.03	0.90	0.84	0.92	0.87	9.0
$\theta = 0.5$	0.03	0.04	0.10	0.06	0.00	0.93	0.68	0.93	0.72	10.0
$\theta = 0.8$	0.03	0.03	0.11	0.06	0.00	0.95	0.64	0.94	0.67	10.0
Vary n_A with $n = 300$, $p = 400$, $K = 10$, $m = 5$										
$n_A = 1$	0.07	0.07	0.12	0.08	0.02	0.92	0.83	0.92	0.87	9.1
$n_A = 2$	0.10	0.11	0.13	0.11	0.05	0.91	1.05	0.93	1.09	8.1
$n_A = 3$	0.12	0.13	0.14	0.13	0.07	0.91	1.14	0.93	1.19	7.1
$n_A = 4$	0.13	0.15	0.13	0.13	0.07	0.93	1.16	0.93	1.21	6.1
$n_A = 5$	0.18	0.20	0.18	0.18	0.10	0.89	1.32	0.89	1.37	5.2
Vary ρ with $n = 300$, $p = 400$, $K = 10$, $m = 5$										
$\rho = 0.1$	0.07	0.07	0.16	0.10	0.03	0.94	0.83	0.94	0.87	9.1
$\rho = 0.2$	0.11	0.09	0.24	0.16	0.06	0.95	0.83	0.96	0.88	9.0
$\rho = 0.3$	0.16	0.13	0.33	0.22	0.09	0.92	0.85	0.94	0.89	9.0
$\rho = 0.4$	0.24	0.19	0.46	0.31	0.15	0.92	0.85	0.93	0.91	9.0
$\rho = 0.5$	0.34	0.28	0.61	0.42	0.21	0.91	0.87	0.93	0.96	9.0

Table 2. ℓ_2 error of various estimators, the coverage and the averaged length of the 95% CIs of β_1 and the estimated number of latent factors.

References

[1] Bing, X., Bunea, F., Wegkamp, M. and Strimas-Mackey, S. (2019). Essential regression.

[2] Bing, X., Bunea, F., Yang, N. and Wegkamp, M. (2019). Adaptive Estimation in Structured Factor Models with Applications to Overlapping Clustering. *To appear in the Annals of Statistics*.

[3] Ibragimov, R. and Sharakhmetov, S. (1998). Short Communications: On an Exact Constant for the Rosenthal Inequality. *Theory of Probability & Its Applications* **42** 294-302.

[4] Klopp, O., Lu, Y., Tsybakov, A. B. and Zhou, H. H. (2017). Structured Matrix Estimation and Completion. *ArXiv e-prints*.

[5] Rudelson, M. and Vershynin, R. (2010). *Non-asymptotic Theory of Random Matrices: Extreme Singular Values* 1576-1602.

[6] Shorack, G. R. (2017). *Probability for Statisticians* 23–38. Springer International Publishing, Cham.

[7] Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York.

[8] Vershynin, R. (2012). *Introduction to the non-asymptotic analysis of random matrices* 210 – 268. Cambridge University Press.

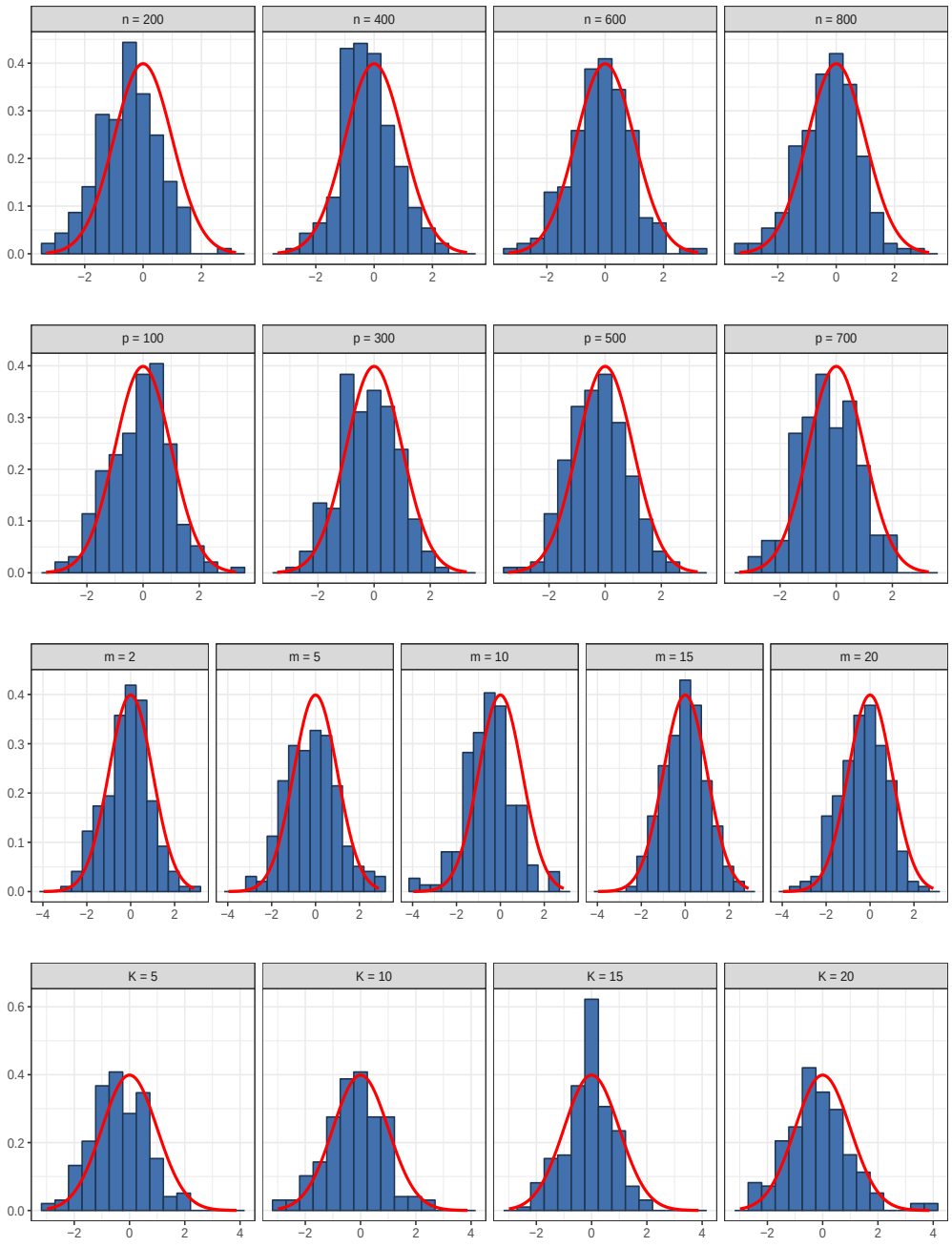

Figure 1: Histograms of the standardized $\widehat{\beta}_1$. The red curves are the density of $N(0,1)$.