Submitted to the Annals of Statistics

SUPPLEMENT TO "TEST OF SIGNIFICANCE FOR HIGH-DIMENSIONAL LONGITUDINAL DATA"

By Ethan X. Fang[*], Yang Ning[‡] and Runze Li[‡]

Pennsylvania State University[*†] *and Cornell University*[‡]

S.1. Technical Lemmas. In this section, we provide some technical lemmas used in our proofs in Sections 2 and 3.

Lemma S.1.1 (Bernstein's inequality for random matrices (Tropp et al., 2015)). *Let $\{\mathbf{V}_k\}$ be a sequence of independent random matrices with dimensions $d_1 \times d_2$. Assume that $\mathbb{E}\mathbf{V}_k = \mathbf{0}$ and $\|\mathbf{V}_k\|_2 \le R_n$ almost surely. Define*

$$\sigma_n^2 = \max\Big\{\Big\|\sum_{k=1}^n \mathbb{E}(\mathbf{V}_k\mathbf{V}_k^\top)\Big\|_2, \Big\|\sum_{k=1}^n \mathbb{E}(\mathbf{V}_k^\top\mathbf{V}_k)\Big\|_2\Big\}.$$

Then, for all $t \ge 0$,

$$\mathbb{P}\Big(\Big\|\sum_{k=1}^n \mathbf{V}_k\Big\|_2 \ge t\Big) \le (d_1+d_2)\exp\Big(-\frac{t^2/2}{\sigma_n^2 + R_n t/3}\Big). \tag{S.1}$$

Moreover, when $\|\mathbf{V}_k\|_2$ is not uniformly bounded, we still have

$$\Big(\mathbb{E}\Big\|\sum_{k=1}^n \mathbf{V}_k\Big\|_2^2\Big)^{1/2} \lesssim \Big(\sigma_n^2\log(d_1+d_2)\Big)^{1/2} + (\mathbb{E}\max_k\|\mathbf{V}_k\|_2^2)^{1/2}\log(d_1+d_2). \tag{S.2}$$

Lemma S.1.2. *Recall that*

$$\bar{\mathbf{S}}_k^*(\boldsymbol{\theta}^*) = \frac{1}{n}\sum_{i=1}^n(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}_k^*)^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{-1/2}(\boldsymbol{\beta}^*)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i^*) \in \mathbb{R}^{d_0}.$$

Suppose that Assumptions 2.1 — 2.5 and $\frac{d_0(\log n)^2\log d_0}{n} = o(1)$ hold. Then

$$\max_{1\le k\le K}\Big\|\nabla\bar{\mathbf{S}}_k^*(\boldsymbol{\theta}^*) - \mathbf{g}_{0k}(\boldsymbol{\theta}^*)\Big\|_2 = \mathcal{O}_{\mathbb{P}}(\sqrt{d_0\log d_0/n}), \tag{S.3}$$

$$\max_{1\le k\le K}\Big\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}_k^*)^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}_k^*) - \mathbf{g}_{0k}(\boldsymbol{\theta}^*)\Big\|_2 = \mathcal{O}_{\mathbb{P}}(\sqrt{d_0\log d_0/n}),$$

and

$$\max_{1\le k\le K}\Big\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\boldsymbol{Z}_i - \mathbb{E}(\boldsymbol{Z}_i\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\boldsymbol{Z}_i)\Big\|_2 = \mathcal{O}_{\mathbb{P}}(\sqrt{d_0\log d_0/n}).$$

[*]Supported by NIH grant P50 DA039838 and NSF grant DMS 1820702.
[†]Supported by NSF grant DMS 1854637.
[‡]Supported by NSF grant DMS 1820702, NIH grants P50 DA039838 and T32 LM012415.
Keywords and phrases: False discovery rate, Generalized estimating equation, Quadratic inference function

Recall that

$$S^*_{ik}(\boldsymbol{\theta}) = (\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}^*_k)^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{-1/2}(\boldsymbol{\beta}^*)\{\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta},\boldsymbol{\gamma}^*)\}, \tag{S.4}$$

$$\mathbf{S}^*_i(\boldsymbol{\theta}) = \big(S^{*T}_{i1}(\boldsymbol{\theta}), ..., S^{*T}_{iK}(\boldsymbol{\theta})\big)^T \in \mathbb{R}^{d_0K}. \tag{S.5}$$

Finally, we have

$$\Big\|\frac{1}{n}\sum_{i=1}^n \mathbf{S}^*_i(\boldsymbol{\theta}^*)\mathbf{S}^{*T}_i(\boldsymbol{\theta}^*) - \mathbf{C}^*\Big\|_2 = \mathcal{O}_{\mathbb{P}}(\sqrt{d_0\log d_0(\log n)^2/n}). \tag{S.6}$$

Proof. We prove this result by applying Lemma S.1.1. In particular, we focus on (S.3) and (S.6) and the rest of the results follow from the same argument. To show (S.3), we take

$$\mathbf{V}_i = n^{-1}[(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}^*_k)^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\boldsymbol{Z}_i - \mathbb{E}\{(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}^*_k)^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\boldsymbol{Z}_i\}].$$

It is easily seen that

$$\begin{aligned}\|\mathbf{V}_i\|_2 &\le n^{-1}\Big[\|(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}^*_k)^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\boldsymbol{Z}_i\|_2 + \|\mathbb{E}\{(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}^*_k)^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\boldsymbol{Z}_i\}\|_2\Big]\\ &\le n^{-1}\Big[\mathrm{tr}((\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}^*_k)^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\boldsymbol{Z}_i) + \|\mathbf{H}_{k\boldsymbol{\theta}\boldsymbol{\theta}} - \mathbf{H}_{k\boldsymbol{\theta}\boldsymbol{\gamma}}\mathbf{H}^{-1}_{k\boldsymbol{\gamma}\boldsymbol{\gamma}}\mathbf{H}_{k\boldsymbol{\gamma}\boldsymbol{\theta}}\|_2\Big]\\ &\lesssim d_0/n,\end{aligned}$$

where in the last step we use boundedness assumption on each entries of $\boldsymbol{Z}_i$ and $\boldsymbol{U}_i\mathbf{W}^*_k$, and

$$\begin{aligned}\|\mathbf{H}_{k\boldsymbol{\theta}\boldsymbol{\theta}} - \mathbf{H}_{k\boldsymbol{\theta}\boldsymbol{\gamma}}\mathbf{H}^{-1}_{k\boldsymbol{\gamma}\boldsymbol{\gamma}}\mathbf{H}_{k\boldsymbol{\gamma}\boldsymbol{\theta}}\|_2 &\le \sup_{\|\boldsymbol{v}\|_2=1}\mathbb{E}[(\boldsymbol{Z}_i\boldsymbol{v})^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k^{-1}\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)(\boldsymbol{Z}_i\boldsymbol{v})]\\ &\le \sup_{\|\boldsymbol{v}\|_2=1}\mathbb{E}[\|\boldsymbol{Z}_i\boldsymbol{v}\|_2^2\|\mathbf{A}_i(\boldsymbol{\beta}^*)\|_2\|\mathbf{T}_k^{-1}\|_2]\\ &\lesssim \lambda_{\max}(\mathbb{E}(\boldsymbol{Z}_i^T\boldsymbol{Z}_i)),\end{aligned}$$

which is bounded by Assumptions 2.5 and 2.4. Denote $\boldsymbol{\Psi}_i = \mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)$. Moreover,

$$\begin{aligned}\Big\|\sum_{i=1}^n\mathbb{E}(\mathbf{V}_i\mathbf{V}_i^\top)\Big\|_2 &= n^{-1}\Big\|\mathbb{E}(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}^*_k)^T\boldsymbol{\Psi}_i\boldsymbol{Z}_i\boldsymbol{Z}_i^T\boldsymbol{\Psi}_i(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}^*_k) - (\mathbf{H}_{k\boldsymbol{\theta}\boldsymbol{\theta}} - \mathbf{H}_{k\boldsymbol{\theta}\boldsymbol{\gamma}}\mathbf{H}^{-1}_{k\boldsymbol{\gamma}\boldsymbol{\gamma}}\mathbf{H}_{k\boldsymbol{\gamma}\boldsymbol{\theta}})^{\otimes 2}\Big\|_2\\ &\le n^{-1}\Big(\|\mathbb{E}(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}^*_k)^T\boldsymbol{\Psi}_i\boldsymbol{Z}_i\boldsymbol{Z}_i^T\boldsymbol{\Psi}_i(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}^*_k)\|_2 + C\Big)\\ &\lesssim d_0/n,\end{aligned}$$

where C is a constant, and the last step follows from

$$\begin{aligned}&\|\mathbb{E}(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}^*_k)^T\boldsymbol{\Psi}_i\boldsymbol{Z}_i\boldsymbol{Z}_i^T\boldsymbol{\Psi}_i(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}^*_k)\|_2\\ &= \sup_{\|\boldsymbol{v}\|_2=1}\mathbb{E}[(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}^*_k)\boldsymbol{v}]^T\boldsymbol{\Psi}_i\boldsymbol{Z}_i\boldsymbol{Z}_i^T\boldsymbol{\Psi}_i[(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}^*_k)\boldsymbol{v}]\\ &\le \sup_{\|\boldsymbol{v}\|_2=1}\mathbb{E}[(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}^*_k)\boldsymbol{v}]^T\boldsymbol{\Psi}_i[(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}^*_k)\boldsymbol{v}]\|\boldsymbol{\Psi}_i^{1/2}\boldsymbol{Z}_i\boldsymbol{Z}_i^T\boldsymbol{\Psi}_i^{1/2}\|_2\\ &\le \sup_{\|\boldsymbol{v}\|_2=1}\mathbb{E}[(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}^*_k)\boldsymbol{v}]^T\boldsymbol{\Psi}_i[(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}^*_k)\boldsymbol{v}]\mathrm{tr}(\boldsymbol{\Psi}_i^{1/2}\boldsymbol{Z}_i\boldsymbol{Z}_i^T\boldsymbol{\Psi}_i^{1/2})\\ &= \sup_{\|\boldsymbol{v}\|_2=1}\mathbb{E}[(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}^*_k)\boldsymbol{v}]^T\boldsymbol{\Psi}_i[(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}^*_k)\boldsymbol{v}]\mathrm{tr}(\boldsymbol{Z}_i^T\boldsymbol{\Psi}_i^{1/2}\boldsymbol{Z}_i)\\ &\lesssim d_0\|\mathbf{H}_{k\boldsymbol{\theta}\boldsymbol{\theta}} - \mathbf{H}_{k\boldsymbol{\theta}\boldsymbol{\gamma}}\mathbf{H}^{-1}_{k\boldsymbol{\gamma}\boldsymbol{\gamma}}\mathbf{H}_{k\boldsymbol{\gamma}\boldsymbol{\theta}}\|_2.\end{aligned}$$

Similarly, we can check

$$\begin{aligned}\Big\|\sum_{i=1}^n \mathbb{E}(\mathbf{V}_i^T\mathbf{V}_i)\Big\|_2 &= n^{-1}\Big\|\mathbb{E}\boldsymbol{Z}_i^T\boldsymbol{\Psi}_i(\boldsymbol{Z}_i-\boldsymbol{U}_i\mathbf{W}_k^*)(\boldsymbol{Z}_i-\boldsymbol{U}_i\mathbf{W}_k^*)^T\boldsymbol{\Psi}_i\boldsymbol{Z}_i-(\mathbf{H}_{k\boldsymbol{\theta}\boldsymbol{\theta}}-\mathbf{H}_{k\boldsymbol{\theta}\boldsymbol{\gamma}}\mathbf{H}_{k\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}\mathbf{H}_{k\boldsymbol{\gamma}\boldsymbol{\theta}})^{\otimes 2}\Big\|_2\\ &\le n^{-1}\Big(\|\mathbb{E}\boldsymbol{Z}_i^T\boldsymbol{\Psi}_i(\boldsymbol{Z}_i-\boldsymbol{U}_i\mathbf{W}_k^*)(\boldsymbol{Z}_i-\boldsymbol{U}_i\mathbf{W}_k^*)^T\boldsymbol{\Psi}_i\boldsymbol{Z}_i\|_2+C\Big)\\ &\lesssim d_0/n,\end{aligned}$$

since after similar calculation we have

$$\|\mathbb{E}\boldsymbol{Z}_i^T\boldsymbol{\Psi}_i(\boldsymbol{Z}_i-\boldsymbol{U}_i\mathbf{W}_k^*)(\boldsymbol{Z}_i-\boldsymbol{U}_i\mathbf{W}_k^*)^T\boldsymbol{\Psi}_i\boldsymbol{Z}_i\|_2 \lesssim d_0\lambda_{\max}(\mathbb{E}(\boldsymbol{Z}_i^T\boldsymbol{Z}_i)).$$

Note that $\sqrt{d_0\log d_0/n}=o(1)$. Now, if we take $t=C\sqrt{d_0\log d_0/n}$ in Lemma S.1.1 (S.1) for some constant C sufficiently large, then we have

$$\mathbb{P}(\|\sum_{k=1}^n \mathbf{V}_k\|_2\ge t)\le 2d_0\exp(-C'\log d_0)$$

for some $C'>1$. Then, the right hand side converges to 0, as $d_0\to\infty$. Finally, we notice that K is finite. This completes the proof of (S.3). To show (S.6), we take

$$\mathbf{V}_i = n^{-1}(\mathbf{S}_i^*(\boldsymbol{\theta}^*)\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*)-\mathbf{C}^*).$$

Thus,

$$\begin{aligned}\Big\|\sum_{i=1}^n \mathbb{E}(\mathbf{V}_i\mathbf{V}_i^\top)\Big\|_2 &= n^{-1}\Big\|\mathbb{E}\mathbf{S}_i^*(\boldsymbol{\theta}^*)\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*)\mathbf{S}_i^*(\boldsymbol{\theta}^*)\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*)-\mathbf{C}^*\mathbf{C}^*\Big\|_2\\ &\le n^{-1}\Big(\|\mathbb{E}\mathbf{S}_i^*(\boldsymbol{\theta}^*)\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*)\mathbf{S}_i^*(\boldsymbol{\theta}^*)\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*)\|_2+C\Big)\\ &\lesssim n^{-1}\Big\|\mathbb{E}\|\mathbf{S}_i^*(\boldsymbol{\theta}^*)\|_2^2\mathbf{S}_i^*(\boldsymbol{\theta}^*)\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*)\Big\|_2\\ &\lesssim d_0n^{-1}\Big\|\mathbb{E}\|\boldsymbol{Y}_i-\boldsymbol{\mu}_i^*\|_2^2\mathbf{S}_i^*(\boldsymbol{\theta}^*)\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*)\Big\|_2,\end{aligned}$$

where the third step holds because $\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*)\mathbf{S}_i^*(\boldsymbol{\theta}^*)=\|\mathbf{S}_i^*(\boldsymbol{\theta}^*)\|_2^2$ and the last step holds by the boundedness assumption implied by the previous proof. From the last line, we will further bound the spectral norm as follows

$$\begin{aligned}&\Big\|\mathbb{E}\|\boldsymbol{Y}_i-\boldsymbol{\mu}_i^*\|_2^2\mathbf{S}_i^*(\boldsymbol{\theta}^*)\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*)\Big\|_2\\ &\le \sup_{\|\boldsymbol{v}\|_2=1}\mathbb{E}\|\boldsymbol{Y}_i-\boldsymbol{\mu}_i^*\|_2^2|\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*)\boldsymbol{v}|^2I(\|\boldsymbol{Y}_i-\boldsymbol{\mu}_i^*\|_2\le\kappa)+\sup_{\|\boldsymbol{v}\|_2=1}\mathbb{E}\|\boldsymbol{Y}_i-\boldsymbol{\mu}_i^*\|_2^2|\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*)\boldsymbol{v}|^2I(\|\boldsymbol{Y}_i-\boldsymbol{\mu}_i^*\|_2>\kappa)\\ &\le \kappa^2\sup_{\|\boldsymbol{v}\|_2=1}\mathbb{E}|\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*)\boldsymbol{v}|^2I(\|\boldsymbol{Y}_i-\boldsymbol{\mu}_i^*\|_2\le\kappa)+\mathbb{E}\|\boldsymbol{Y}_i-\boldsymbol{\mu}_i^*\|_2^2\|\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*)\|_2^2I(\|\boldsymbol{Y}_i-\boldsymbol{\mu}_i^*\|_2>\kappa)\\ &\lesssim \kappa^2\sup_{\|\boldsymbol{v}\|_2=1}\mathbb{E}|\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*)\boldsymbol{v}|^2+d_0\mathbb{E}\|\boldsymbol{Y}_i-\boldsymbol{\mu}_i^*\|_2^4I(\|\boldsymbol{Y}_i-\boldsymbol{\mu}_i^*\|_2>\kappa)\\ &\le \kappa^2\|\mathbf{C}^*\|_2+d_0(\mathbb{E}\|\boldsymbol{Y}_i-\boldsymbol{\mu}_i^*\|_2^8)^{1/2}(\mathbb{P}(\|\boldsymbol{Y}_i-\boldsymbol{\mu}_i^*\|_2>\kappa))^{1/2}\\ &\lesssim (\log n)^2,\end{aligned}$$

where we pick $\kappa = C\log n$ for some sufficiently large constant C. In the last step, we use the sub-exponential property for each entry of $\boldsymbol{Y}_i - \boldsymbol{\mu}_i^* \in \mathbb{R}^m$. First, clearly $\mathbb{E}\|\boldsymbol{Y}_i - \boldsymbol{\mu}_i^*\|_2^8$ is bounded by a constant, and moreover

$$\mathbb{P}(\|\boldsymbol{Y}_i - \boldsymbol{\mu}_i^*\|_2 > \kappa) \le C_1 \exp(-C_2\kappa) \lesssim n^{-2}$$

as we pick a sufficiently large constant C and $d_0/n = o(1)$. Thus, $\sigma_n^2 \lesssim d_0 n^{-1}(\log n)^2$. Then, consider

$$\begin{aligned}
\mathbb{E}\max_i \|\mathbf{V}_i\|_2^2 &\lesssim n^{-2}\mathbb{E}\max_i \|\mathbf{S}_i^*(\boldsymbol{\theta}^*)\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*)\|_2^2 \\
&\le n^{-2}\mathbb{E}\max_i \|\mathbf{S}_i^*(\boldsymbol{\theta}^*)\|_2^4 \\
&\lesssim d_0^2 n^{-2}\mathbb{E}\max_i \|\boldsymbol{Y}_i - \boldsymbol{\mu}_i^*\|_2^4 \\
&\lesssim d_0^2 n^{-2}(\log n)^4,
\end{aligned}$$

where again the last step follows from the sub-exponential property. Thus, we can now apply (S.2), which yields

$$\Big(\mathbb{E}\Big\|\sum_{i=1}^n \mathbf{V}_i\Big\|_2^2\Big)^{1/2} \lesssim \Big(\frac{d_0(\log n)^2\log d_0}{n}\Big)^{1/2} + \frac{d_0(\log n)^2\log d_0}{n}.$$

Thus, (S.6) holds by the Markov inequality and $\frac{d_0(\log n)^2\log d_0}{n} = o(1)$. □

Lemma S.1.3. Under Assumption 2.1, we have

$$\Big\|\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^m \boldsymbol{X}_{ij}\{Y_{ij} - \mu_{ij}(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}^*)\}\Big\|_\infty \lesssim \sqrt{\frac{\log d}{n}}.$$

Proof. We can easily show that for any $1 \le l \le d$,

$$\|\sum_{j=1}^m X_{ijl}\{Y_{ij} - \mu_{ij}(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}^*)\}\|_{\psi_1} \le \sum_{j=1}^m \|X_{ijl}\{Y_{ij} - \mu_{ij}(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}^*)\}\|_{\psi_1} \le Cm\|\epsilon_{ij}\|_{\psi_1} \le C'$$

for some constants $C, C' > 0$. Then, the lemma follows by the Bernstein inequality. □

Lemma S.1.4. Without loss of generality, assume that $d_0 = 1$. Under Assumptions 2.1, 2.2, 2.3, 2.4 and 2.5 and $s\sqrt{\frac{\log d}{n}} = o(1)$, for any set $\mathcal{S} \subset \{1, ..., d-1\}$ where $|\mathcal{S}| \asymp s'$ and any vector $\mathbf{v}$ belonging to the cone $\mathcal{C}(\xi, \mathcal{S}) = \{\mathbf{v} \in \mathbb{R}^{d-1} : \|\mathbf{v}_{\bar{\mathcal{S}}}\|_1 \le \xi\|\mathbf{v}_{\mathcal{S}}\|_1\}$ for some positive constant ξ, it holds that

$$\inf_{\mathbf{0}\neq\mathbf{v}\in\mathcal{C}(\xi,\mathcal{S})} \frac{s^{1/2}\{\mathbf{v}^T\widehat{\mathbf{W}}\mathbf{v}\}^{1/2}}{\|\mathbf{v}_{\mathcal{S}}\|_1} \ge C > 0,$$

with probability tending to one, where

$$\widehat{\mathbf{W}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{U}_i^T\mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}})\mathbf{T}_k\mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}})\boldsymbol{U}_i.$$

Proof. Let

$$\mathbf{W} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{U}_i^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\boldsymbol{U}_i.$$

Denote

$$\boldsymbol{\Psi}_i = \mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*), \quad \widehat{\boldsymbol{\Psi}}_i = \mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}})\mathbf{T}_k\mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}}).$$

By the consistency of $\widehat{\boldsymbol{\beta}}$, we can easily verify that

$$\begin{aligned}
|\mathbf{v}^T(\widehat{\mathbf{W}} - \mathbf{W})\mathbf{v}| &= \Big|\frac{1}{n}\sum_{i=1}^n \mathbf{v}^T\boldsymbol{U}_i^T(\widehat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i)\boldsymbol{U}_i\mathbf{v}\Big| \\
&\lesssim \Big|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{U}_i\mathbf{v})^T\boldsymbol{\Psi}_i(\boldsymbol{U}_i\mathbf{v})\Big| \max_i \|\widehat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i\|_2 \\
&= o_p(\mathbf{v}^T\mathbf{W}\mathbf{v}).
\end{aligned}$$

By $\mathbf{v}^T\widehat{\mathbf{W}}\mathbf{v} \geq \mathbf{v}^T\mathbf{W}\mathbf{v} - |\mathbf{v}^T(\widehat{\mathbf{W}} - \mathbf{W})\mathbf{v}| \geq (1 - o_p(1))\mathbf{v}^T\mathbf{W}\mathbf{v}$, we only need to separately bound $\mathbf{v}^T\mathbf{W}\mathbf{v}$ from the below. Since $\lambda_{\min}(\boldsymbol{\Psi}_i) \geq C > 0$, we have

$$\begin{aligned}
\mathbf{v}^T\mathbf{W}\mathbf{v} &\geq C \cdot \frac{1}{n}\sum_{i=1}^n \mathbf{v}^T\boldsymbol{U}_i^T\boldsymbol{U}_i\mathbf{v} \\
&\geq \lambda_{\min}(\mathbb{E}\boldsymbol{U}_i^T\boldsymbol{U}_i)\|\mathbf{v}\|_2^2 - \|\mathbf{v}\|_1^2\Big\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{U}_i^T\boldsymbol{U}_i - \mathbb{E}\boldsymbol{U}_i^T\boldsymbol{U}_i\Big\|_{\max} \\
&\geq C\|\mathbf{v}_{\mathcal{S}}\|_1^2/s - (1+\xi)^2\|\mathbf{v}_{\mathcal{S}}\|_1^2\sqrt{\frac{\log d}{n}} \\
&\geq C's^{-1}\|\mathbf{v}_{\mathcal{S}}\|_1^2,
\end{aligned}$$

where we apply the concentration result similar to Lemma S.1.5 and the last step holds due to the condition $s\sqrt{\frac{\log d}{n}} = o(1)$. Thus,

$$\frac{s^{1/2}\{\mathbf{v}^T\widehat{\mathbf{W}}\mathbf{v}\}^{1/2}}{\|\mathbf{v}_{\mathcal{S}}\|_1} \geq \frac{(1 - o_p(1))C'^{1/2}\|\mathbf{v}_{\mathcal{S}}\|_1}{\|\mathbf{v}_{\mathcal{S}}\|_1} \geq C'' > 0.$$

This completes the proof. □

LEMMA S.1.5. Suppose that Assumptions 2.1 — 2.5 hold. We have

$$\Big\|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}_k^*)^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\boldsymbol{U}_i\Big\|_{\max} \lesssim \sqrt{\frac{\log d}{n}},$$

$$\Big\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{U}_i^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\theta^*, \boldsymbol{\gamma}^*))\Big\|_{\infty} \lesssim \sqrt{\frac{\log d}{n}}.$$

PROOF. The proof is similar to Lemma S.1.3 by noting that for $j \in [d - d_0]$ and $\ell \in [d_0]$

$$|(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}_k^*)_\ell^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\boldsymbol{U}_{ij}| \leq \|(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}_k^*)_\ell\|_2\|\mathbf{A}_i(\boldsymbol{\beta}^*)\|_2\|\mathbf{T}_k\|_2\|\boldsymbol{U}_{ij}\|_2$$

is bounded, $\mathbb{E}[(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}_k^*)^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\boldsymbol{U}_i] = 0$ and the use of Hoeffding inequality and the union bound. To show the second result, we denote the jth row of $\boldsymbol{U}_i^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)$ as

$\boldsymbol{h}_j \in \mathbb{R}^m$. Following the similar argument, we can verify that $\|\boldsymbol{h}_j\|_\infty \le C$. Then

$$\begin{aligned}\mathbb{P}(|\boldsymbol{h}_j(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\theta^*, \boldsymbol{\gamma}^*))| \ge t) &\le \mathbb{P}(C\|\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\theta^*, \boldsymbol{\gamma}^*)\|_1 \ge t) \\ &\le \sum_{k=1}^m \mathbb{P}(|(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\theta^*, \boldsymbol{\gamma}^*))_k| \ge \frac{t}{Cm}) \le c_1 \exp(-c_2 t),\end{aligned}$$

for some constants $c_1, c_2 > 0$, where the last step follows from the sub-exponential assumption in 2.1. The conclusion holds by applying the Bernstein inequality for the average of sub-exponential random variables and the union bound. □

Define

$$\|\mathbf{A}\|_{L_1} := \max_j \sum_i |A_{ij}|$$

to be the maximum absolute column sum of the matrix $\mathbf{A}$.

LEMMA S.1.6. Under Assumptions 2.1, 2.2, 2.3, 2.4 and 2.5, if $\lambda \asymp \lambda' \asymp \sqrt{n^{-1}\log d}$ and $s\sqrt{\frac{\log d}{n}} = o(1)$, we have that

$$\sup_{1\le k\le K} \|\widehat{\mathbf{W}}_k - \mathbf{W}_k^*\|_{L_1} = \mathcal{O}_{\mathbb{P}}\Big(\max(s, s')\sqrt{\frac{\log d}{n}}\Big), \tag{S.7}$$

(S.8)

$$\sup_{1\le k\le K} \mathrm{tr}\Big(\frac{1}{n}\sum_{i=1}^n [\boldsymbol{U}_i(\widehat{\mathbf{W}}_k - \mathbf{W}_k^*)]^T \mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k \mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)[\boldsymbol{U}_i(\widehat{\mathbf{W}}_k - \mathbf{W}_k^*)]\Big) = \mathcal{O}_{\mathbb{P}}\Big(\frac{\max(s, s')d_0 \log d}{n}\Big).$$

PROOF OF LEMMA S.1.6. Since the minimization problem is decomposable for $\widehat{\mathbf{W}}_k$, we will focus on the property of $(\widehat{\mathbf{W}}_k)_{\cdot j}$ for $1 \le j \le d_0$. For notational simplicity, denote $\widehat{\mathbf{w}} = (\widehat{\mathbf{W}}_k)_{\cdot j}$, $\mathbf{w}^* = (\mathbf{W}_k^*)_{\cdot j}$, $\widehat{\boldsymbol{\Delta}} = \widehat{\mathbf{w}} - \mathbf{w}^*$,

$$\boldsymbol{\Psi}_i = \mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k \mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*), \quad \widehat{\boldsymbol{\Psi}}_i = \mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}})\mathbf{T}_k \mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}}),$$

and $\boldsymbol{z}_{ij}$ to be the jth column of $\boldsymbol{Z}_i$. By definition of $\widehat{\mathbf{W}}_k$, it is easily seen that

$$\frac{1}{n}\sum_{i=1}^n (\boldsymbol{z}_{ij} - \boldsymbol{U}_i\widehat{\mathbf{w}})^T \widehat{\boldsymbol{\Psi}}_i(\boldsymbol{z}_{ij} - \boldsymbol{U}_i\widehat{\mathbf{w}}) + \lambda'\|\widehat{\mathbf{w}}\|_1 \le \frac{1}{n}\sum_{i=1}^n (\boldsymbol{z}_{ij} - \boldsymbol{U}_i\mathbf{w}^*)^T \widehat{\boldsymbol{\Psi}}_i(\boldsymbol{z}_{ij} - \boldsymbol{U}_i\mathbf{w}^*) + \lambda'\|\mathbf{w}^*\|_1.$$

Let

$$V = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{U}_i\widehat{\boldsymbol{\Delta}})^T \widehat{\boldsymbol{\Psi}}_i(\boldsymbol{U}_i\widehat{\boldsymbol{\Delta}}), \quad I = \frac{2}{n}\sum_{i=1}^n (\boldsymbol{z}_{ij} - \boldsymbol{U}_i\mathbf{w}^*)^T \widehat{\boldsymbol{\Psi}}_i(\boldsymbol{U}_i\widehat{\boldsymbol{\Delta}}).$$

The above inequality is equivalent to

$$V \le I + \lambda'\|\mathbf{w}^*\|_1 - \lambda'\|\widehat{\mathbf{w}}\|_1. \tag{S.9}$$

Let S' denote the support set of $\mathbf{w}^*$ and $\bar{S}'$ denote its complement. We have

$$\lambda'\|\mathbf{w}^*\|_1 - \lambda'\|\widehat{\mathbf{w}}\|_1 \le \lambda'\|\mathbf{w}^*_{S'}\|_1 - \lambda'\|\widehat{\mathbf{w}}_{S'}\|_1 - \lambda'\|\widehat{\mathbf{w}}_{\bar{S}'}\|_1 \le \lambda'\|\widehat{\boldsymbol{\Delta}}_{S'}\|_1 - \lambda'\|\widehat{\boldsymbol{\Delta}}_{\bar{S}'}\|_1.$$

Let $\boldsymbol{r}_{ij} = \boldsymbol{z}_{ij} - \boldsymbol{U}_i \mathbf{w}^*$. Note that for I, we have

$$
\begin{aligned}
I &= \frac{2}{n}\sum_{i=1}^{n} \boldsymbol{r}_{ij}^T \boldsymbol{\Psi}_i (\boldsymbol{U}_i \widehat{\boldsymbol{\Delta}}) + \frac{2}{n}\sum_{i=1}^{n} \boldsymbol{r}_{ij}^T (\widehat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i)(\boldsymbol{U}_i \widehat{\boldsymbol{\Delta}}) \\
&\le \Big\| \frac{2}{n}\sum_{i=1}^{n} \boldsymbol{r}_{ij}^T \boldsymbol{\Psi}_i \boldsymbol{U}_i \Big\|_\infty \|\widehat{\boldsymbol{\Delta}}\|_1 + \Big| \frac{2}{n}\sum_{i=1}^{n} (\boldsymbol{U}_i \widehat{\boldsymbol{\Delta}})^T \widehat{\boldsymbol{\Psi}}_i (\boldsymbol{U}_i \widehat{\boldsymbol{\Delta}}) \Big|^{1/2} \Big| \frac{2}{n}\sum_{i=1}^{n} \boldsymbol{r}_{ij}^T (\widehat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i) \widehat{\boldsymbol{\Psi}}_i^{-1} (\widehat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i) \boldsymbol{r}_i \Big|^{1/2} \\
&\le C\sqrt{\frac{\log d}{n}} \|\widetilde{\boldsymbol{\Delta}}\|_1 + (2V)^{1/2} \Big(\frac{2}{n}\sum_{i=1}^{n} \|\boldsymbol{r}_{ij}\|_2^2 \|\widehat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i\|_2^2 \|\widehat{\boldsymbol{\Psi}}_i^{-1}\|_2 \Big)^{1/2},
\end{aligned}
$$

where the first inequality follows from the Hölder inequality and Cauchy-Schwartz inequality. In particular, we apply Lemma S.1.5,

$$
\max_{1\le j\le d_0} \Big\| \frac{2}{n}\sum_{i=1}^{n} \boldsymbol{r}_{ij}^T \boldsymbol{\Psi}_i \boldsymbol{U}_i \Big\|_\infty \le C\sqrt{\frac{\log d}{n}},
$$

which holds uniformly over $1 \le j \le d_0$ with high probability. Let $\mathbf{A}_i = \mathbf{A}_i(\boldsymbol{\beta}^*)$ and $\widehat{\mathbf{A}}_i = \mathbf{A}_i(\widehat{\boldsymbol{\beta}})$. It can be shown that

$$
\begin{aligned}
\max_i \|\widehat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i\|_2 &\le \max_i 2\|\mathbf{A}_i^{1/2}\|_2 \|\mathbf{T}_k\|_2 \|\widehat{\mathbf{A}}_i^{1/2} - \mathbf{A}_i^{1/2}\|_2 + \max_i \|\mathbf{T}_k\|_2 \|\widehat{\mathbf{A}}_i^{1/2} - \mathbf{A}_i^{1/2}\|_2^2 \\
&\lesssim \max_i \max_{1\le j\le m} |\mu_{ij}^{1/2}(\widehat{\boldsymbol{\beta}}) - \mu_{ij}^{1/2}(\boldsymbol{\beta}^*)| + \max_i \max_{1\le j\le m} |\mu_{ij}^{1/2}(\widehat{\boldsymbol{\beta}}) - \mu_{ij}^{1/2}(\boldsymbol{\beta}^*)|^2 \\
&\lesssim \max_i \max_{1\le j\le m} |\boldsymbol{X}_{ij}^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)| \lesssim s\sqrt{\frac{\log d}{n}} = o_p(1).
\end{aligned}
$$

In addition,

$$
\lambda_{\min}(\boldsymbol{\Psi}_i) \ge \lambda_{\min}(\mathbf{A}_i)\lambda_{\min}(\mathbf{T}_k) \ge C > 0,
$$

and similarly $\lambda_{\max}(\boldsymbol{\Psi}_i)$ is upper bounded by a constant. Thus, $\|\widehat{\boldsymbol{\Psi}}_i^{-1}\|_2 = \frac{1}{\lambda_{\min}(\widehat{\boldsymbol{\Psi}}_i)} \le \frac{1}{\lambda_{\min}(\boldsymbol{\Psi}_i) - \|\widehat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i\|_2}$ is upper bounded by a constant uniformly over i. Moreover, $\|\boldsymbol{r}_{ij}\|_2$ is bounded uniformly over i, j, we have

$$
\begin{aligned}
\frac{2}{n}\sum_{i=1}^{n} \|\boldsymbol{r}_{ij}\|_2^2 \|\widehat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i\|_2^2 \|\widehat{\boldsymbol{\Psi}}_i^{-1}\|_2 &\lesssim \frac{2}{n}\sum_{i=1}^{n} \max_{1\le j\le m} |\mu_{ij}^{1/2}(\widehat{\boldsymbol{\beta}}) - \mu_{ij}^{1/2}(\boldsymbol{\beta}^*)|^2 \\
&\le \frac{2}{n}\sum_{i=1}^{n} \max_{1\le j\le m} |\boldsymbol{X}_{ij}^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|^2 \\
&\le \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} |\boldsymbol{X}_{ij}^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|^2 \lesssim \frac{s\log d}{n}.
\end{aligned}
$$

Combining the bound for I and (S.9) and taking $\lambda' = 2C\sqrt{\frac{\log d}{n}}$, we obtain

$$
V \le C'\sqrt{\frac{s\log d}{n}} V^{1/2} + 3C\sqrt{\frac{\log d}{n}} \|\widehat{\boldsymbol{\Delta}}_{S'}\|_1 - C\sqrt{\frac{\log d}{n}} \|\widehat{\boldsymbol{\Delta}}_{\bar{S}'}\|_1, \tag{S.10}
$$

for some constants C and C'. In particular, we note that the above inequality holds uniformly over $1 \le j \le d_0$. Now, we consider two situations. If $V^{1/2} \le C'\sqrt{\frac{s\log d}{n}}$, the inequality (S.8) holds trivially. Otherwise, we have $V^{1/2} > C'\sqrt{\frac{s\log d}{n}}$. In this case, $V - C'\sqrt{\frac{s\log d}{n}}V^{1/2} > 0$, together with (S.10), this implies $3\|\widehat{\mathbf{\Delta}}_{S'}\|_1 \geq \|\widehat{\mathbf{\Delta}}_{\bar{S}'}\|_1$. Due to this cone condition for $\widehat{\mathbf{\Delta}}$, we can apply Lemma S.1.4 with $\xi = 3$ to conclude that $\|\widehat{\mathbf{\Delta}}_{S'}\|_1 \leq C''\sqrt{s'}V^{1/2}$ for some constant C''. By (S.10),

$$V \leq C'\sqrt{\frac{s\log d}{n}}V^{1/2} + 3CC''\sqrt{\frac{s'\log d}{n}}V^{1/2},$$

which implies that $V^{1/2} = \mathcal{O}_{\mathbb{P}}(\sqrt{\frac{\max(s,s')\log d}{n}})$.

To prove (S.7), we still consider two situations. First, if $6\|\widehat{\mathbf{\Delta}}_{S'}\|_1 \geq \|\widehat{\mathbf{\Delta}}_{\bar{S}'}\|_1$, then we have $\|\widehat{\mathbf{\Delta}}\|_1 \leq 7\|\widehat{\mathbf{\Delta}}_{S'}\|_1 \leq C''\sqrt{s'}V^{1/2}$, by Lemma S.1.4 with $\xi = 6$. Therefore, we obtain $\|\widehat{\mathbf{\Delta}}\|_1 = \mathcal{O}_{\mathbb{P}}(\max(s,s')\sqrt{\frac{\log d}{n}})$. Otherwise, we have $6\|\widehat{\mathbf{\Delta}}_{S'}\|_1 < \|\widehat{\mathbf{\Delta}}_{\bar{S}'}\|_1$. Together with (S.10), this implies that

$$V \leq C'\sqrt{\frac{s\log d}{n}}V^{1/2} - 3C\sqrt{\frac{\log d}{n}}\|\widehat{\mathbf{\Delta}}_{S'}\|_1.$$

Hence,

$$\|\widehat{\mathbf{\Delta}}\|_1 \leq \frac{7}{6}\|\widehat{\mathbf{\Delta}}_{\bar{S}'}\|_1 \leq \frac{7}{6}\Big(3C\sqrt{\frac{\log d}{n}}\Big)^{-1}\Big(C'\sqrt{\frac{s\log d}{n}}V^{1/2} - V\Big) \leq \frac{7C'}{18C}\sqrt{s}V^{1/2}.$$

Since $V^{1/2} = \mathcal{O}_{\mathbb{P}}(\sqrt{\frac{\max(s,s')\log d}{n}})$, we obtain (S.7). Finally,

$$\begin{aligned}
&\Big|V - \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{U}_i\widehat{\mathbf{\Delta}})^T\mathbf{\Psi}_i(\boldsymbol{U}_i\widehat{\mathbf{\Delta}})\Big| \leq \Big|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{U}_i\widehat{\mathbf{\Delta}})^T(\widehat{\mathbf{\Psi}}_i - \mathbf{\Psi}_i)(\boldsymbol{U}_i\widehat{\mathbf{\Delta}})\Big| \\
&\leq \Big|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{U}_i\widehat{\mathbf{\Delta}})^T\widehat{\mathbf{\Psi}}_i(\boldsymbol{U}_i\widehat{\mathbf{\Delta}})\Big|\max_i\|\widehat{\mathbf{\Psi}}_i^{-1}\|_2\|\widehat{\mathbf{\Psi}}_i - \mathbf{\Psi}_i\|_2 = o_p\Big(\frac{\max(s,s')\log d}{n}\Big).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathrm{tr}\Big(\frac{1}{n}\sum_{i=1}^{n}[\boldsymbol{U}_i(\widehat{\mathbf{W}}_k - \mathbf{W}_k^*)]^T\mathbf{\Psi}_i[\boldsymbol{U}_i(\widehat{\mathbf{W}}_k - \mathbf{W}_k^*)]\Big) \\
&\leq d_0\max_{1\leq j\leq d_0}\Big|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{U}_i(\widehat{\mathbf{W}}_k - \mathbf{W}_k^*)_{\cdot j})^T\mathbf{\Psi}_i(\boldsymbol{U}_i(\widehat{\mathbf{W}}_k - \mathbf{W}_k^*)_{\cdot j})\Big| \\
&\lesssim \frac{\max(s,s')d_0\log d}{n}.
\end{aligned}$$

This completes the proof of (S.8) $\square$

Lemma S.1.7. Recall that $\widehat{\mathbf{g}}(\boldsymbol{\theta}) = \nabla\overline{\mathbf{S}}_n(\boldsymbol{\theta})$ and $\mathbf{g}_0(\boldsymbol{\theta}) = \mathbb{E}\{\nabla\mathbf{S}_i^*(\boldsymbol{\theta})\}$. Under the conditions in Theorem 2.7, we have

$$\|\widehat{\mathbf{g}}(\widehat{\boldsymbol{\theta}}) - \mathbf{g}_0(\boldsymbol{\theta}^*)\|_2 = \mathcal{O}_{\mathbb{P}}\Big(\sqrt{\frac{(s\vee s')d_0\log d}{n}}\Big).$$

$$\|\widehat{\mathbf{g}}(\boldsymbol{\theta}^*) - \mathbf{g}_0(\boldsymbol{\theta}^*)\|_2 = \mathcal{O}_{\mathbb{P}}\Big(\sqrt{\frac{(s\vee s')d_0\log d}{n}}\Big).$$

PROOF. Note that

$$\|\widehat{\mathbf{g}}(\widehat{\boldsymbol{\theta}}) - \mathbf{g}_0(\boldsymbol{\theta}^*)\|_2 \le \|\nabla \bar{\mathbf{S}}_n^*(\boldsymbol{\theta}^*) - \mathbf{g}_0(\boldsymbol{\theta}^*)\|_2 + \|\nabla \bar{\mathbf{S}}_n^*(\boldsymbol{\theta}^*) - \nabla \bar{\mathbf{S}}_n(\widehat{\boldsymbol{\theta}})\|_2. \tag{S.11}$$

For notational simplicity, we only look at the kth block of $\nabla \bar{\mathbf{S}}_n^*(\boldsymbol{\theta}^*)$, denoted by

$$\nabla \bar{\mathbf{S}}_k^*(\boldsymbol{\theta}^*) = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{Z}_i - \boldsymbol{U}_i \mathbf{W}_k^*)^T \mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k \mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\boldsymbol{Z}_i \in \mathbb{R}^{d_0 \times d_0}.$$

The corresponding block in $\mathbf{g}_0(\boldsymbol{\theta}^*)$ is denoted by $\mathbf{g}_{0k}(\boldsymbol{\theta}^*)$. Lemma S.1.2 implies

$$\|\bar{\mathbf{S}}_n^*(\boldsymbol{\theta}^*) - \mathbf{g}_0(\boldsymbol{\theta}^*)\|_2 \lesssim \max_{1\le k\le K} \|\nabla \bar{\mathbf{S}}_k^*(\boldsymbol{\theta}^*) - \mathbf{g}_{0k}(\boldsymbol{\theta}^*)\|_2 = \mathcal{O}_{\mathbb{P}}(\sqrt{d_0 \log d_0/n}) \tag{S.12}$$

which gives the rate for the first term in (S.11). Now, we consider the second term,

$$\|\nabla \bar{\mathbf{S}}_n^*(\boldsymbol{\theta}^*) - \nabla \bar{\mathbf{S}}_n(\widehat{\boldsymbol{\theta}})\|_2 \lesssim \max_{1\le k\le K} \|\nabla \bar{\mathbf{S}}_k^*(\boldsymbol{\theta}^*) - \nabla \bar{\mathbf{S}}_k(\widehat{\boldsymbol{\theta}})\|_2,$$

where again with a slight abuse of notation we use $\nabla \bar{\mathbf{S}}_k^*(\boldsymbol{\theta}^*)$ to denote the kth block of $\nabla \bar{\mathbf{S}}_n^*(\boldsymbol{\theta}^*)$. Recall that $\boldsymbol{\Psi}_i = \mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k \mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)$, $\widehat{\boldsymbol{\Psi}}_i = \mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}})\mathbf{T}_k \mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}})$. Then

$$\begin{aligned}
&\|\nabla \bar{\mathbf{S}}_k^*(\boldsymbol{\theta}^*) - \nabla \bar{\mathbf{S}}_k(\widehat{\boldsymbol{\theta}})\|_2 \\
&\le \Big\|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{Z}_i - \boldsymbol{U}_i \mathbf{W}_k^*)^T (\widehat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i)\boldsymbol{Z}_i\Big\|_2 + \Big\|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{U}_i \mathbf{W}_k^* - \boldsymbol{U}_i \widehat{\mathbf{W}}_k)^T \boldsymbol{\Psi}_i \boldsymbol{Z}_i\Big\|_2 \\
&\quad + \Big\|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{U}_i \mathbf{W}_k^* - \boldsymbol{U}_i \widehat{\mathbf{W}}_k)^T (\widehat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i)\boldsymbol{Z}_i\Big\|_2 := T_1 + T_2 + T_3.
\end{aligned}$$

For T_1, Lemma S.1.16 implies

$$\begin{aligned}
T_1 &\le \Big\|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{Z}_i - \boldsymbol{U}_i \mathbf{W}_k^*)^T \boldsymbol{\Psi}_i (\boldsymbol{Z}_i - \boldsymbol{U}_i \mathbf{W}_k^*)\Big\|_2^{1/2} \Big\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i^T (\widehat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i)\boldsymbol{\Psi}_i^{-1}(\widehat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i)\boldsymbol{Z}_i\Big\|_2^{1/2} \\
&= \mathcal{O}_{\mathbb{P}}\Big(\sqrt{\frac{s d_0 \log d}{n}}\Big),
\end{aligned}$$

where the last step is from Lemma S.1.2

$$\begin{aligned}
\Big\|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{Z}_i - \boldsymbol{U}_i \mathbf{W}_k^*)^T \boldsymbol{\Psi}_i (\boldsymbol{Z}_i - \boldsymbol{U}_i \mathbf{W}_k^*)\Big\|_2 &= \|\mathbf{H}_{k\boldsymbol{\theta}\boldsymbol{\theta}} - \mathbf{H}_{k\boldsymbol{\theta}\boldsymbol{\gamma}}\mathbf{H}_{k\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}\mathbf{H}_{k\boldsymbol{\gamma}\boldsymbol{\theta}}\|_2 + \mathcal{O}_{\mathbb{P}}(\sqrt{d_0 \log d_0/n}) \\
&= \mathcal{O}_{\mathbb{P}}(1),
\end{aligned}$$

and also Lemma S.1.9. For T_2, by Lemma S.1.16

$$\begin{aligned}
T_2 &\le \Big\|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{U}_i \mathbf{W}_k^* - \boldsymbol{U}_i \widehat{\mathbf{W}}_k)^T \boldsymbol{\Psi}_i (\boldsymbol{U}_i \mathbf{W}_k^* - \boldsymbol{U}_i \widehat{\mathbf{W}}_k)\Big\|_2^{1/2} \Big\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i^T \boldsymbol{\Psi}_i \boldsymbol{Z}_i\Big\|_2^{1/2} \\
&\lesssim \Big[\mathrm{tr}\Big(\frac{1}{n}\sum_{i=1}^n (\boldsymbol{U}_i \mathbf{W}_k^* - \boldsymbol{U}_i \widehat{\mathbf{W}}_k)^T \boldsymbol{\Psi}_i (\boldsymbol{U}_i \mathbf{W}_k^* - \boldsymbol{U}_i \widehat{\mathbf{W}}_k)\Big)\Big]^{1/2} \\
&\lesssim \sqrt{\frac{(s \vee s') d_0 \log d}{n}},
\end{aligned}$$

where the second step follows from Lemma S.1.2 and the last step is from Lemma S.1.6. Finally, for T_3, similar argument implies

$$\begin{aligned}
T_3 &\le \Big\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{U}_i\mathbf{W}_k^* - \boldsymbol{U}_i\widehat{\mathbf{W}}_k)^T\boldsymbol{\Psi}_i(\boldsymbol{U}_i\mathbf{W}_k^* - \boldsymbol{U}_i\widehat{\mathbf{W}}_k)\Big\|_2^{1/2}\Big\|\frac{1}{n}\sum_{i=1}^n\boldsymbol{Z}_i^T(\widehat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i)\boldsymbol{\Psi}_i^{-1}(\widehat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i)\boldsymbol{Z}_i\Big\|_2^{1/2}\\
&\lesssim \Big[\mathrm{tr}\Big(\frac{1}{n}\sum_{i=1}^n(\boldsymbol{U}_i\mathbf{W}_k^* - \boldsymbol{U}_i\widehat{\mathbf{W}}_k)^T\boldsymbol{\Psi}_i(\boldsymbol{U}_i\mathbf{W}_k^* - \boldsymbol{U}_i\widehat{\mathbf{W}}_k)\Big)\Big]^{1/2}\sqrt{\frac{sd_0\log d}{n}}\\
&\lesssim \frac{(s\vee s')d_0\log d}{n}.
\end{aligned}$$

Thus, combining (S.11) and (S.12), we have

$$\|\widehat{\mathbf{g}}(\widehat{\boldsymbol{\theta}}) - \mathbf{g}_0(\boldsymbol{\theta}^*)\|_2 = \mathcal{O}_{\mathbb{P}}\Big(\sqrt{\frac{(s\vee s')d_0\log d}{n}}\Big).$$

It is easily seen that similar argument implies

$$\|\widehat{\mathbf{g}}(\boldsymbol{\theta}^*) - \mathbf{g}_0(\boldsymbol{\theta}^*)\|_2 = \mathcal{O}_{\mathbb{P}}\Big(\sqrt{\frac{(s\vee s')d_0\log d}{n}}\Big).$$

Our claim holds as desired. □

LEMMA S.1.8. Recall that $\widehat{\mathbf{C}} = n^{-1}\sum_{i=1}^n\widehat{\mathbf{S}}_i(\widehat{\theta})\widehat{\mathbf{S}}_i^T(\widehat{\theta})$ in (2.9), where $\widehat{\mathbf{S}}_i(\theta) = \{\widehat{S}_{i1}(\theta),...,\widehat{S}_{iK}(\theta)\}^T$, and $\mathbf{C}^* = \mathbb{E}\{\mathbf{S}_i^*(\theta^*)\mathbf{S}_i^*(\theta^*)^T\}$ in (2.16), where $\mathbf{S}_i^*(\theta^*) = (S_{i1}^*(\theta^*),\ldots,S_{iK}^*(\theta^*))^T$. Under the conditions in Theorem 2.7, we have

$$\|\widehat{\mathbf{C}} - \mathbf{C}^*\|_2 = \mathcal{O}_{\mathbb{P}}\Big(\sqrt{\frac{d_0(s\vee s')\log d(\log n)^2}{n}}\Big).$$

PROOF. We have that

$$\|\widehat{\mathbf{C}} - \mathbf{C}^*\|_2 \le \Big\|\frac{1}{n}\sum_{i=1}^n\mathbf{S}_i^*(\boldsymbol{\theta}^*)\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*) - \frac{1}{n}\sum_{i=1}^n\widehat{\mathbf{S}}_i(\widehat{\boldsymbol{\theta}})\widehat{\mathbf{S}}_i^T(\widehat{\boldsymbol{\theta}})\Big\|_2 + \Big\|\frac{1}{n}\sum_{i=1}^n\mathbf{S}_i^*(\boldsymbol{\theta}^*)\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*) - \mathbf{C}^*\Big\|_2.$$

The second term is of order $\mathcal{O}_{\mathbb{P}}(\sqrt{d_0\log d_0(\log n)^2/n})$ by Lemma S.1.2. For the first term,

$$\begin{aligned}
&\Big\|\frac{1}{n}\sum_{i=1}^n\mathbf{S}_i^*(\boldsymbol{\theta}^*)\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*) - \frac{1}{n}\sum_{i=1}^n\widehat{\mathbf{S}}_i(\widehat{\boldsymbol{\theta}})\widehat{\mathbf{S}}_i^T(\widehat{\boldsymbol{\theta}})\Big\|_2\\
&\le 2\Big\|\frac{1}{n}\sum_{i=1}^n[\mathbf{S}_i^*(\boldsymbol{\theta}^*) - \widehat{\mathbf{S}}_i(\widehat{\boldsymbol{\theta}})]\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*)\Big\|_2 + \Big\|\frac{1}{n}\sum_{i=1}^n[\mathbf{S}_i^*(\boldsymbol{\theta}^*) - \widehat{\mathbf{S}}_i(\widehat{\boldsymbol{\theta}})][\mathbf{S}_i^*(\boldsymbol{\theta}^*) - \widehat{\mathbf{S}}_i(\widehat{\boldsymbol{\theta}})]^T\Big\|_2 := T_{11} + T_{12}.
\end{aligned}$$

We first consider the T_{12} term. For simplicity of notation, we assume $K = 1$. After some tedious calculation, we can show that

$$T_{12} \lesssim A_1 + A_2 + A_3 + A_4 + A_5,$$

where $\widehat{\mathbf{\Delta}}_i = \mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}})\mathbf{T}_k\mathbf{A}_i^{-1/2}(\widehat{\boldsymbol{\beta}})$ and $\mathbf{\Delta}_i = \mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{-1/2}(\boldsymbol{\beta}^*)$ and

$$A_1 = \Big\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}_k^*)^T\mathbf{\Delta}_i(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i^*)(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i^*)^T\mathbf{\Delta}_i^T(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}_k^*)\Big\|_2$$
$$A_2 = \Big\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}_k^*)^T(\widehat{\mathbf{\Delta}}_i - \mathbf{\Delta}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i^*)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i^*)^T(\widehat{\mathbf{\Delta}}_i - \mathbf{\Delta}_i)^T(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}_k^*)\Big\|_2$$
$$A_3 = \Big\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}_k^*)^T(\widehat{\mathbf{\Delta}}_i - \mathbf{\Delta}_i)(\boldsymbol{\mu}_i^* - \widehat{\boldsymbol{\mu}}_i)(\boldsymbol{\mu}_i^* - \widehat{\boldsymbol{\mu}}_i)^T(\widehat{\mathbf{\Delta}}_i - \mathbf{\Delta}_i)^T(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}_k^*)\Big\|_2$$
$$A_4 = \Big\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{U}_i\mathbf{W}_k^* - \boldsymbol{U}_i\widehat{\mathbf{W}}_k)^T\widehat{\mathbf{\Delta}}_i(\boldsymbol{Y}_i - \boldsymbol{\mu}_i^*)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i^*)^T\widehat{\mathbf{\Delta}}_i^T(\boldsymbol{U}_i\mathbf{W}_k^* - \boldsymbol{U}_i\widehat{\mathbf{W}}_k)\Big\|_2$$
$$A_5 = \Big\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{U}_i\mathbf{W}_k^* - \boldsymbol{U}_i\widehat{\mathbf{W}}_k)^T\widehat{\mathbf{\Delta}}_i(\boldsymbol{\mu}_i^* - \widehat{\boldsymbol{\mu}}_i)(\boldsymbol{\mu}_i^* - \widehat{\boldsymbol{\mu}}_i)^T\widehat{\mathbf{\Delta}}_i^T(\boldsymbol{U}_i\mathbf{W}_k^* - \boldsymbol{U}_i\widehat{\mathbf{W}}_k)\Big\|_2.$$

In the following, we give the bound for the above five terms. It holds that

$$\begin{aligned} A_1 &= \frac{1}{n}\sum_{i=1}^n\|\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}_k^*\|_2^2\|\mathbf{\Delta}_i\|_2^2\|\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i^*\|_2^2 \\ &\lesssim d_0\frac{1}{n}\sum_{i=1}^n\|\boldsymbol{X}_i^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 \lesssim \frac{d_0 s\log d}{n}. \end{aligned}$$

The similar arguments and the proof of Lemma S.1.9 yield

$$\begin{aligned} A_2 &\lesssim d_0\frac{1}{n}\sum_{i=1}^n\|\widehat{\mathbf{\Delta}}_i - \mathbf{\Delta}_i\|_2^2\|\boldsymbol{Y}_i - \boldsymbol{\mu}_i^*\|_2^2 \\ &\lesssim d_0(\log n)^2\frac{1}{n}\sum_{i=1}^n\|\boldsymbol{X}_i^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 \lesssim \frac{d_0 s\log d(\log n)^2}{n}, \end{aligned}$$

where the sub-exponential condition implies $\max_i\|\boldsymbol{Y}_i - \boldsymbol{\mu}_i^*\|_2 \lesssim \log n$, and similarly $A_3 \lesssim \frac{d_0 s\log d}{n}\frac{s^2\log d}{n}$.

$$\begin{aligned} A_4 &\lesssim \max_i\|\boldsymbol{Y}_i - \boldsymbol{\mu}_i^*\|_2^2\mathrm{tr}\Big(\frac{1}{n}\sum_{i=1}^n[\boldsymbol{U}_i(\widehat{\mathbf{W}}_k - \mathbf{W}_k^*)]^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)[\boldsymbol{U}_i(\widehat{\mathbf{W}}_k - \mathbf{W}_k^*)]\Big) \\ &\lesssim \frac{d_0(s\vee s')\log d(\log n)^2}{n}, \end{aligned}$$

and $A_5 \lesssim \frac{d_0(s\vee s')\log d}{n}\frac{s^2\log d}{n}$. Since $\frac{s^2\log d}{n} = o(1)$, we have

$$T_{12} \lesssim \frac{d_0(s\vee s')\log d(\log n)^2}{n}.$$

For T_{11},

$$\begin{aligned} T_{11} &\lesssim \Big\|\frac{1}{n}\sum_{i=1}^n[\mathbf{S}_i^*(\boldsymbol{\theta}^*) - \widehat{\mathbf{S}}_i(\widehat{\boldsymbol{\theta}})][\mathbf{S}_i^*(\boldsymbol{\theta}^*) - \widehat{\mathbf{S}}_i(\widehat{\boldsymbol{\theta}})]^T\Big\|_2^{1/2}\Big\|\frac{1}{n}\sum_{i=1}^n\mathbf{S}_i^*(\boldsymbol{\theta}^*)\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*)\Big\|_2^{1/2} \\ &\lesssim \sqrt{\frac{d_0(s\vee s')\log d(\log n)^2}{n}}, \end{aligned}$$

by the convergence rate of T_{12} established above, Lemma S.1.2

$$\Big\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{S}_i^*(\boldsymbol{\theta}^*)\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*)-\mathbf{C}^*\Big\|_2=\mathcal{O}_{\mathbb{P}}(\sqrt{d_0\log d_0(\log n)^2/n}),$$

and $\|\mathbf{C}^*\|_2=\mathcal{O}(1)$. Our claim holds as desired. □

LEMMA S.1.9. Under the same conditions as in Theorem 2.7, for any $\boldsymbol{u}_i,\boldsymbol{v}_i\in\mathbb{R}^m$

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i^T(\widehat{\boldsymbol{\Delta}}_i-\boldsymbol{\Delta}_i)\boldsymbol{v}_i\Big|\lesssim\Big[\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{u}_i\|_2^2\|\boldsymbol{v}_i\|_2^2\Big]^{1/2}\sqrt{\frac{s\log d}{n}},\tag{S.13}$$

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i^T(\widehat{\boldsymbol{\Delta}}_i-\boldsymbol{\Delta}_i)\mathbf{A}_i^{-1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k^{-1}\mathbf{A}_i^{-1/2}(\boldsymbol{\beta}^*)(\widehat{\boldsymbol{\Delta}}_i-\boldsymbol{\Delta}_i)\boldsymbol{u}_i\Big|\lesssim\Big[\max_i\|\boldsymbol{u}_i\|_2^2\Big]\frac{s\log d}{n},\tag{S.14}$$

where $\widehat{\boldsymbol{\Delta}}_i=\mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}})\mathbf{T}_k\mathbf{A}_i^{-1/2}(\widehat{\boldsymbol{\beta}})$ and $\boldsymbol{\Delta}_i=\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{-1/2}(\boldsymbol{\beta}^*)$. In addition, for any matrix $\mathbf{M}_i\in\mathbb{R}^{m\times d_0}$ with $\max_i\|\mathbf{M}_i\|_{\max}=\mathcal{O}(1)$,

$$\Big\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{M}_i^T(\widehat{\boldsymbol{\Psi}}_i-\boldsymbol{\Psi}_i)\boldsymbol{\Psi}_i^{-1}(\widehat{\boldsymbol{\Psi}}_i-\boldsymbol{\Psi}_i)\mathbf{M}_i^T\Big\|_2\lesssim\frac{sd_0\log d}{n},\tag{S.15}$$

where $\boldsymbol{\Psi}_i=\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)$ and $\widehat{\boldsymbol{\Psi}}_i=\mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}})\mathbf{T}_k\mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}})$.

PROOF OF LEMMA S.1.9. Let $\delta_{ij}=\boldsymbol{X}_{ij}^T(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)$. Recall that by the mean value theorem,

$$\begin{aligned}\widehat{\boldsymbol{\Delta}}_i-\boldsymbol{\Delta}_i=&\begin{bmatrix}\frac{1}{2}\widetilde{\mu}_{i1}^{\prime-0.5}\widetilde{\mu}_{i1}''\delta_{i1}&&&\\&\frac{1}{2}\widetilde{\mu}_{i2}^{\prime-0.5}\widetilde{\mu}_{i1}''\delta_{i2}&&\\&&\cdots&\\&&&\frac{1}{2}\widetilde{\mu}_{im}^{\prime-0.5}\widetilde{\mu}_{im}''\delta_{im}\end{bmatrix}\mathbf{T}_k\mathbf{A}_i^{-1/2}(\boldsymbol{\beta}^*)\\&-\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\begin{bmatrix}\frac{1}{2}\widetilde{\mu}_{i1}^{\prime-1.5}\widetilde{\mu}_{i1}''\delta_{i1}&&&\\&\frac{1}{2}\widetilde{\mu}_{i2}^{\prime-1.5}\widetilde{\mu}_{i1}''\delta_{i2}&&\\&&\cdots&\\&&&\frac{1}{2}\widetilde{\mu}_{im}^{\prime-1.5}\widetilde{\mu}_{im}''\delta_{im}\end{bmatrix}\end{aligned}$$

where $\widetilde{\mu}_{ij}'=d\mu_{ij}(\widetilde{\eta}_{ij})/d\eta_{ij}$ and $\widetilde{\eta}_{ij}$ is an intermediate value between $\boldsymbol{X}_{ij}^T\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{X}_{ij}^T\widehat{\boldsymbol{\beta}}$ and $\widetilde{\mu}_{ij}''$ is defined similarly. To show (S.13), note that

$$\begin{aligned}\Big|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i^T(\widehat{\boldsymbol{\Delta}}_i-\boldsymbol{\Delta}_i)\boldsymbol{v}_i\Big|&\le\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{u}_i\|_2\|\boldsymbol{v}_i\|_2\|\widehat{\boldsymbol{\Delta}}_i-\boldsymbol{\Delta}_i\|_2\\&\le\Big[\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{u}_i\|_2^2\|\boldsymbol{v}_i\|_2^2\Big]^{1/2}\Big[\frac{1}{n}\sum_{i=1}^{n}\|\widehat{\boldsymbol{\Delta}}_i-\boldsymbol{\Delta}_i\|_2^2\Big]^{1/2}\\&\lesssim\Big[\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{u}_i\|_2^2\|\boldsymbol{v}_i\|_2^2\Big]^{1/2}\Big[\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\delta_{ij}^2\Big]^{1/2}\\&\lesssim\Big[\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{u}_i\|_2^2\|\boldsymbol{v}_i\|_2^2\Big]^{1/2}\sqrt{\frac{s\log d}{n}},\end{aligned}$$

where the third step follows from the assumption that $\|\mathbf{T}_k\|_2$ is bounded and Assumption 2.4, and the last step is implied by the convergence rate of the initial estimators. The similar argument can be used to show (S.14). For (S.15),

$$\Big\|\frac{1}{n}\sum_{i=1}^n \mathbf{M}_i^T(\widehat{\mathbf{\Psi}}_i-\mathbf{\Psi}_i)\mathbf{\Psi}_i^{-1}(\widehat{\mathbf{\Psi}}_i-\mathbf{\Psi}_i)\mathbf{M}_i^T\Big\|_2 \le \frac{1}{n}\sum_{i=1}^n \|\widehat{\mathbf{\Psi}}_i-\mathbf{\Psi}_i\|_2^2\|\mathbf{\Psi}_i^{-1}\|_2\|\mathbf{M}_i\|_2^2$$
$$\lesssim \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^m \delta_{ij}^2 \mathrm{tr}(\mathbf{M}_i\mathbf{M}_i^T) \lesssim \frac{sd_0\log d}{n}.$$

The proof is complete. □

Lemma S.1.10. Under the same conditions as in Theorem 2.7, we have,

$$\max_{1\le k\le K}\Big\|\bar{\mathbf{S}}_{nk}(\boldsymbol{\theta}^*)-\frac{1}{n}\sum_{i=1}^n \mathbf{S}_{ik}^*(\boldsymbol{\theta}^*)\Big\|_2 = \mathcal{O}_{\mathbb{P}}\Big(\frac{d_0^{1/2}(s\vee s')\log d\log n}{n}\Big).$$

Proof of Lemma S.1.10. For notational simplicity, denote

$$\widehat{\mathbf{\Delta}}_i = \mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}})\mathbf{T}_k\mathbf{A}_i^{-1/2}(\widehat{\boldsymbol{\beta}}), \mathbf{\Delta}_i = \mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{-1/2}(\boldsymbol{\beta}^*),$$
$$\boldsymbol{T}_i = \boldsymbol{Y}_i - \boldsymbol{\mu}_i(\theta^*,\boldsymbol{\gamma}^*), \quad \text{and} \quad \widehat{\boldsymbol{T}}_i = \boldsymbol{Y}_i - \boldsymbol{\mu}_i(\theta^*,\widehat{\boldsymbol{\gamma}}).$$

We can rewrite it as

$$\Big\|\bar{\mathbf{S}}_{nk}(\theta^*)-\frac{1}{n}\sum_{i=1}^n \mathbf{S}_{ik}^*(\theta^*)\Big\|_2 = \Big\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{Z}_i-\boldsymbol{U}_i\widehat{\mathbf{W}}_k)^T\widehat{\mathbf{\Delta}}_i\widehat{\boldsymbol{T}}_i - (\boldsymbol{Z}_i-\boldsymbol{U}_i\mathbf{W}_k^*)^T\mathbf{\Delta}_i\boldsymbol{T}_i\Big\|_2$$
$$\le \Big\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{Z}_i-\boldsymbol{U}_i\mathbf{W}_k^*)^T[\widehat{\mathbf{\Delta}}_i\widehat{\boldsymbol{T}}_i-\mathbf{\Delta}_i\boldsymbol{T}_i]\Big\|_2 + \Big\|\frac{1}{n}\sum_{i=1}^n[\boldsymbol{U}_i(\widehat{\mathbf{W}}_k-\mathbf{W}_k^*)]^T\widehat{\mathbf{\Delta}}_i\widehat{\boldsymbol{T}}_i\Big\|_2.$$

We call the last two terms as I_1 and I_2. For I_2,

$$I_2 \le \Big\|\frac{1}{n}\sum_{i=1}^n[\boldsymbol{U}_i(\widehat{\mathbf{W}}_k-\mathbf{W}_k^*)]^T\widehat{\mathbf{\Delta}}_i\boldsymbol{T}_i\Big\|_2 + \Big\|\frac{1}{n}\sum_{i=1}^n[\boldsymbol{U}_i(\widehat{\mathbf{W}}_k-\mathbf{W}_k^*)]^T\widehat{\mathbf{\Delta}}_i(\widehat{\boldsymbol{\mu}}_i-\boldsymbol{\mu}_i)\Big\|_2 := I_{21}+I_{22},$$

where $\widehat{\boldsymbol{\mu}}_i=\boldsymbol{\mu}_i(\theta^*,\widehat{\boldsymbol{\gamma}})$ and $\boldsymbol{\mu}_i=\boldsymbol{\mu}_i(\theta^*,\boldsymbol{\gamma}^*)$. For I_{21}, let $\boldsymbol{R}_{ij}\in\mathbb{R}^m$ denote the jth row of the matrix $[\boldsymbol{U}_i(\widehat{\mathbf{W}}_k-\mathbf{W}_k^*)]^T\mathbf{\Psi}_i^{1/2}$. Note that

$$\Big\|\frac{1}{n}\sum_{i=1}^n[\boldsymbol{U}_i(\widehat{\mathbf{W}}_k-\mathbf{W}_k^*)]^T[\widehat{\mathbf{\Delta}}_i-\mathbf{\Delta}_i]\boldsymbol{T}_i\Big\|_2^2$$
$$= \sum_{j=1}^{d_0}\Big|\frac{1}{n}\sum_{i=1}^n\boldsymbol{R}_{ij}\mathbf{\Psi}_i^{-1/2}[\widehat{\mathbf{\Delta}}_i-\mathbf{\Delta}_i]\boldsymbol{T}_i\Big|^2$$
$$\le \sum_{j=1}^{d_0}\Big|\frac{1}{n}\sum_{i=1}^n\boldsymbol{R}_{ij}\boldsymbol{R}_{ij}^T\Big| \times \Big|\frac{1}{n}\sum_{i=1}^n\boldsymbol{T}_i^T(\widehat{\mathbf{\Delta}}_i-\mathbf{\Delta}_i)\mathbf{\Psi}^{-1}(\widehat{\mathbf{\Delta}}_i-\mathbf{\Delta}_i)\boldsymbol{T}_i\Big|$$
$$= \mathrm{tr}\Big(\frac{1}{n}\sum_{i=1}^n[\boldsymbol{U}_i(\widehat{\mathbf{W}}_k-\mathbf{W}_k^*)]^T\mathbf{\Psi}_i[\boldsymbol{U}_i(\widehat{\mathbf{W}}_k-\mathbf{W}_k^*)]\Big) \times \Big|\frac{1}{n}\sum_{i=1}^n\boldsymbol{T}_i^T(\widehat{\mathbf{\Delta}}_i-\mathbf{\Delta}_i)\mathbf{\Psi}^{-1}(\widehat{\mathbf{\Delta}}_i-\mathbf{\Delta}_i)\boldsymbol{T}_i\Big|$$
$$\lesssim \frac{d_0(s\vee s')\log d}{n}\cdot\frac{s\log d(\log n)^2}{n},$$

where the last step follows from Lemma S.1.9 and S.1.6, and the sub-exponential condition implies $\max_i \|\boldsymbol{Y}_i - \boldsymbol{\mu}_i^*\|_2 \lesssim \log n$. In addition,

$$\Big\|\frac{1}{n}\sum_{i=1}^n [\boldsymbol{U}_i(\widehat{\mathbf{W}}_k - \mathbf{W}_k^*)]^T \boldsymbol{\Delta}_i \boldsymbol{T}_i\Big\|_2 \le d_0^{1/2}\|\widehat{\mathbf{W}}_k - \mathbf{W}_k^*\|_{L_1}\Big\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{U}_i^T \boldsymbol{\Delta}_i \boldsymbol{T}_i\Big\|_\infty \lesssim \frac{d_0^{1/2}(s \vee s')\log d}{n}.$$

Thus, combining these results,

$$\begin{aligned} I_{21} &\le \Big\|\frac{1}{n}\sum_{i=1}^n [\boldsymbol{U}_i(\widehat{\mathbf{W}}_k - \mathbf{W}_k^*)]^T \boldsymbol{\Delta}_i \boldsymbol{T}_i\Big\|_2 + \Big\|\frac{1}{n}\sum_{i=1}^n [\boldsymbol{U}_i(\widehat{\mathbf{W}}_k - \mathbf{W}_k^*)]^T [\widehat{\boldsymbol{\Delta}}_i - \boldsymbol{\Delta}_i]\boldsymbol{T}_i\Big\|_2 \\ &\lesssim \frac{d_0^{1/2}(s \vee s')\log d}{n} + \frac{d_0^{1/2}(s \vee s')\log d \log n}{n}. \end{aligned}$$

For I_{22}, we have

$$\begin{aligned} |I_{22}|^2 &= \sum_{j=1}^{d_0}\Big|\frac{1}{n}\sum_{i=1}^n \boldsymbol{R}_{ij}\boldsymbol{\Psi}_i^{-1/2}\widehat{\boldsymbol{\Delta}}_i(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)\Big|^2 \\ &\le \sum_{j=1}^{d_0}\Big|\frac{1}{n}\sum_{i=1}^n \boldsymbol{R}_{ij}\boldsymbol{R}_{ij}^T\Big| \times \Big|\frac{1}{n}\sum_{i=1}^n (\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)^T \widehat{\boldsymbol{\Delta}}_i^T \boldsymbol{\Psi}_i^{-1}\widehat{\boldsymbol{\Delta}}_i(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)\Big| \\ &\lesssim \operatorname{tr}\Big(\frac{1}{n}\sum_{i=1}^n [\boldsymbol{U}_i(\widehat{\mathbf{W}}_k - \mathbf{W}_k^*)]^T \boldsymbol{\Psi}_i [\boldsymbol{U}_i(\widehat{\mathbf{W}}_k - \mathbf{W}_k^*)]\Big) \times \Big|\frac{1}{n}\sum_{i=1}^n \|\boldsymbol{U}_i(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)\|_2^2\Big| \\ &\lesssim \frac{d_0(s \vee s')\log d}{n} \cdot \frac{s\log d}{n}. \end{aligned}$$

Combining above results, we have $I_2 = \mathcal{O}_p(\frac{d_0^{1/2}(s\vee s')\log d \log n}{n})$. Then

$$\begin{aligned} I_1 \le& \Big\|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}_k^*)^T(\widehat{\boldsymbol{\Delta}}_i - \boldsymbol{\Delta}_i)\boldsymbol{T}_i\Big\|_2 + \Big\|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}_k^*)^T \boldsymbol{\Delta}_i(\widehat{\boldsymbol{T}}_i - \boldsymbol{T}_i)\Big\|_2 \\ &+ \Big\|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}_k^*)^T(\widehat{\boldsymbol{\Delta}}_i - \boldsymbol{\Delta}_i)(\widehat{\boldsymbol{T}}_i - \boldsymbol{T}_i)\Big\|_2. \end{aligned}$$

Similarly, we can show that $I_1 = \mathcal{O}_p(\frac{d_0^{1/2}(s\vee s')\log d \log n}{n})$. This completes the proof. □

LEMMA S.1.11. Recall that $Q_n^*(\boldsymbol{\theta}) = \bar{\mathbf{S}}_n^*(\boldsymbol{\theta})^T \mathbf{C}^{*-1}\bar{\mathbf{S}}_n^*(\boldsymbol{\theta})$, where $\bar{\mathbf{S}}^*(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \mathbf{S}_i^*(\boldsymbol{\theta})$. Under the same conditions as in Theorem 2.7, we have

$$\|\nabla Q_n^*(\boldsymbol{\theta}^*)\|_2 = \mathcal{O}_{\mathbb{P}}\Big(\sqrt{\frac{d_0}{n}}\Big), \tag{S.16}$$

$$\|\nabla \widetilde{Q}_n(\boldsymbol{\theta}^*) - \nabla Q_n^*(\boldsymbol{\theta}^*)\|_2 = \mathcal{O}_{\mathbb{P}}\Big(\frac{d_0\{(s \vee s')\log d(\log n)^2\}^{1/2}}{n} + \frac{d_0^{1/2}(s \vee s')\log d \log n}{n}\Big). \tag{S.17}$$

PROOF. To show (S.16), we note that

$$\begin{aligned} \|\nabla Q_n^*(\boldsymbol{\theta}^*)\|_2 &\le 2\|\nabla \bar{\mathbf{S}}_n^*(\boldsymbol{\theta}^*)\|_2\|\mathbf{C}^{*-1}\|_2\|\bar{\mathbf{S}}_n^*(\boldsymbol{\theta}^*)\|_2 \\ &\le 2(\|\mathbf{g}_0(\boldsymbol{\theta}^*)\|_2 + \|\nabla \bar{\mathbf{S}}_n^*(\boldsymbol{\theta}^*) - \mathbf{g}_0(\boldsymbol{\theta}^*)\|_2)\|\mathbf{C}^{*-1}\|_2\|\bar{\mathbf{S}}_n^*(\boldsymbol{\theta})\|_2. \end{aligned} \tag{S.18}$$

We will consider the terms in the last line one by one. We first show that $\|\bar{\mathbf{S}}_n^*(\boldsymbol{\theta}^*)\|_2 = \mathcal{O}_{\mathbb{P}}(\sqrt{d_0/n})$. This follows from the Markov inequality and the fact that

$$\text{(S.19)} \qquad \mathbb{E}\|\bar{\mathbf{S}}_n^*(\boldsymbol{\theta}^*)\|_2^2 = \frac{1}{n}\mathbb{E}[\mathbf{S}_i^*(\boldsymbol{\theta}^*)^T \mathbf{S}_i^*(\boldsymbol{\theta}^*)] = \frac{1}{n}\mathrm{tr}(\mathbf{C}^*) = \mathcal{O}(\frac{d_0}{n}),$$

where the last step follows from $\mathrm{tr}(\mathbf{C}^*) \le d_0 K \lambda_{\max}(\mathbf{C}^*)$ and the condition that $\lambda_{\max}(\mathbf{C}^*)$ is upper bounded. In addition, we note that for any matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ with the same number of rows, the matrix $(\mathbf{A}_1, \mathbf{A}_2)$ satisfies $\|(\mathbf{A}_1, \mathbf{A}_2)\|_2 \le \|\mathbf{A}_1\|_2 + \|\mathbf{A}_2\|_2$. Applying this result repeatedly and the definition of $\mathbf{H}_{k\boldsymbol{\gamma}\boldsymbol{\gamma}}$ and $\mathbf{H}_{k\boldsymbol{\gamma}\boldsymbol{\theta}}$, we have

$$\|\mathbf{g}_0(\boldsymbol{\theta}^*)\|_2 \lesssim \sum_{k=1}^K \|\mathbf{H}_{k\boldsymbol{\theta}\boldsymbol{\theta}} - \mathbf{H}_{k\boldsymbol{\theta}\boldsymbol{\gamma}}\mathbf{H}_{k\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}\mathbf{H}_{k\boldsymbol{\gamma}\boldsymbol{\theta}}\|_2 \le \sum_{k=1}^K \lambda_{\max}(\mathbf{H}_{k\boldsymbol{\theta}\boldsymbol{\theta}}).$$

Furthermore, by the definition of $\mathbf{H}_{k\boldsymbol{\theta}\boldsymbol{\theta}}$, we have

$$\begin{aligned}\lambda_{\max}(\mathbf{H}_{k\boldsymbol{\theta}\boldsymbol{\theta}}) &= \sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E}[(\boldsymbol{Z}_i\boldsymbol{v})^T \mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k^{-1}\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)(\boldsymbol{Z}_i\boldsymbol{v})] \\ &\le \sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E}[\|\boldsymbol{Z}_i\boldsymbol{v}\|_2^2 \|\mathbf{A}_i(\boldsymbol{\beta}^*)\|_2 \|\mathbf{T}_k^{-1}\|_2] \\ &\lesssim \lambda_{\max}(\mathbb{E}(\boldsymbol{Z}_i^T\boldsymbol{Z}_i)),\end{aligned}$$

which is bounded by Assumptions 2.5 and 2.4. Thus, $\|\mathbf{g}_0(\boldsymbol{\theta}^*)\|_2 = \mathcal{O}(1)$. For notational simplicity, we only look at the kth block of $\nabla\bar{\mathbf{S}}_n^*(\boldsymbol{\theta}^*)$, denoted by

$$\nabla\bar{\mathbf{S}}_k^*(\boldsymbol{\theta}^*) = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}_k^*)^T \mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\boldsymbol{Z}_i \in \mathbb{R}^{d_0\times d_0}.$$

The corresponding block in $\mathbf{g}_0(\boldsymbol{\theta}^*)$ is denoted by $\mathbf{g}_{0k}(\boldsymbol{\theta}^*)$. We have

$$\|\nabla\bar{\mathbf{S}}_n^*(\boldsymbol{\theta}^*) - \mathbf{g}_0(\boldsymbol{\theta}^*)\|_2 \lesssim \max_{1\le k\le K} \|\nabla\bar{\mathbf{S}}_k^*(\boldsymbol{\theta}^*) - \mathbf{g}_{0k}(\boldsymbol{\theta}^*)\|_2 = \mathcal{O}_{\mathbb{P}}(\sqrt{d_0\log d_0/n}),$$

by Lemma S.1.2. Thus, (S.16) holds by applying (S.18), (S.19) and the above bound. We will consider (S.17) as follows. For notational simplicity, we ignore the $\boldsymbol{\theta}^*$ in the rest of the proof, unless stated otherwise. Thus,

$$\begin{aligned}\|\nabla\widetilde{Q}_n - \nabla Q_n^*\|_2 &= \|(\nabla\bar{\mathbf{S}}_n)\widehat{\mathbf{C}}^{-1}\bar{\mathbf{S}}_n - (\nabla\bar{\mathbf{S}}_n^*)\mathbf{C}^{*-1}\bar{\mathbf{S}}_n^*\|_2 \\ &\le \|(\nabla\bar{\mathbf{S}}_n - \nabla\bar{\mathbf{S}}_n^*)\widehat{\mathbf{C}}^{-1}\bar{\mathbf{S}}_n\|_2 + \|(\nabla\bar{\mathbf{S}}_n^*)(\widehat{\mathbf{C}}^{-1} - \mathbf{C}^{*-1})\bar{\mathbf{S}}_n\|_2 \\ &\quad + \|(\nabla\bar{\mathbf{S}}_n^*)\mathbf{C}^{*-1}(\bar{\mathbf{S}}_n - \bar{\mathbf{S}}_n^*)\|_2.\end{aligned}$$

If we can show

$$\text{(S.20)} \qquad \|\widehat{\mathbf{C}}^{-1} - \mathbf{C}^{*-1}\|_2 = \mathcal{O}_{\mathbb{P}}\Big(\sqrt{\frac{d_0(s\vee s')\log d(\log n)^2}{n}}\Big),$$

$$\text{(S.21)} \qquad \|\nabla\bar{\mathbf{S}}_n - \nabla\bar{\mathbf{S}}_n^*\|_2 = \mathcal{O}_{\mathbb{P}}\Big(\sqrt{\frac{(s\vee s')d_0\log d}{n}}\Big),$$

$$\text{(S.22)} \qquad \|\bar{\mathbf{S}}_n - \bar{\mathbf{S}}_n^*\|_2 = \mathcal{O}_{\mathbb{P}}\Big(\frac{d_0^{1/2}(s\vee s')\log d\log n}{n}\Big),$$

then combined with $\|\nabla\bar{\mathbf{S}}_n^*\|_2 = \mathcal{O}_{\mathbb{P}}(1)$ and $\|\bar{\mathbf{S}}_n^*(\boldsymbol{\theta}^*)\|_2 = \mathcal{O}_{\mathbb{P}}(\sqrt{d_0/n})$, we have (S.17) holds. In the following, we will show (S.20)–(S.22). For (S.20),

$$\|\widehat{\mathbf{C}}^{-1} - \mathbf{C}^{*-1}\|_2 \le \|\widehat{\mathbf{C}}^{-1}\|_2\|\mathbf{C}^{*-1}\|_2\|\widehat{\mathbf{C}} - \mathbf{C}^*\|_2 = \mathcal{O}_{\mathbb{P}}\Big(\sqrt{\frac{d_0(s \vee s')\log d(\log n)^2}{n}}\Big),$$

implied by the Weyl's inequality

$$\|\widehat{\mathbf{C}}^{-1}\|_2 = \frac{1}{\lambda_{\min}(\widehat{\mathbf{C}} - \mathbf{C}^* + \mathbf{C}^*)} \le \frac{1}{\lambda_{\min}(\mathbf{C}^*) - \|\widehat{\mathbf{C}} - \mathbf{C}^*\|_2} = \mathcal{O}_{\mathbb{P}}(1)$$

and Lemma S.1.8. In addition, (S.21) is implied by the proof of Lemma S.1.7. Finally, (S.22) is implied by Lemma S.1.10. This completes the proof. □

LEMMA S.1.12. Let c denote a small constant. Under the conditions in Theorem 2.7, uniformly over $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \le c d_0^{-1/2}$ it holds that with probability tending to one

$$\widetilde{Q}_n(\boldsymbol{\theta}) - \widetilde{Q}_n(\boldsymbol{\theta}^*) - \nabla\widetilde{Q}_n(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \ge C\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2$$

for some positive constant C.

PROOF. Applying the mean value theorem, there exists $\bar{\boldsymbol{\theta}} = t\boldsymbol{\theta}^* + (1-t)\boldsymbol{\theta}$ for some $t \in (0,1)$ such that

$$\begin{aligned}
&\widetilde{Q}_n(\boldsymbol{\theta}) - \widetilde{Q}_n(\boldsymbol{\theta}^*) - \nabla\widetilde{Q}_n(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\\
&= (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T\nabla^2\widetilde{Q}_n(\bar{\boldsymbol{\theta}})(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\\
&= (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T\widehat{\mathbf{g}}(\bar{\boldsymbol{\theta}})^T\widehat{\mathbf{C}}^{-1}\widehat{\mathbf{g}}(\bar{\boldsymbol{\theta}})(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - (\mathbf{P}_{i1}, ..., \mathbf{P}_{iK})\widehat{\mathbf{C}}^{-1}\bar{\mathbf{S}}_n(\bar{\boldsymbol{\theta}}),
\end{aligned} \tag{S.23}$$

where

$$\mathbf{P}_{ik} = \frac{1}{n}\sum_{i=1}^n(\delta_{i1}^2, ..., \delta_{im}^2)\mathbf{M}_i\widehat{\boldsymbol{\Delta}}_{ik}(\boldsymbol{Z}_i - \boldsymbol{U}_i\widehat{\mathbf{W}}_k),$$

with $\widehat{\boldsymbol{\Delta}}_{ik} = \mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}})\mathbf{T}_k\mathbf{A}_i^{-1/2}(\widehat{\boldsymbol{\beta}})$, $\delta_{ik}^2 = (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T\boldsymbol{Z}_{ik}$, and

$$\mathbf{M}_i = \begin{bmatrix} \mu_{i1}''(\bar{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}}) & & & \\ & \mu_{i2}''(\bar{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}}) & & \\ & & \cdots & \\ & & & \mu_{im}''(\bar{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}}) \end{bmatrix}.$$

We first note that

$$\begin{aligned}
&|(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T[\widehat{\mathbf{g}}(\bar{\boldsymbol{\theta}})^T\widehat{\mathbf{C}}^{-1}\widehat{\mathbf{g}}(\bar{\boldsymbol{\theta}}) - \mathbf{g}_0(\boldsymbol{\theta}^*)^T\mathbf{C}^{*-1}\mathbf{g}_0(\boldsymbol{\theta}^*)](\boldsymbol{\theta} - \boldsymbol{\theta}^*)|\\
&\le |(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T[\{\widehat{\mathbf{g}}(\bar{\boldsymbol{\theta}}) - \mathbf{g}_0(\boldsymbol{\theta}^*)\}^T\mathbf{C}^{*-1}\mathbf{g}_0(\boldsymbol{\theta}^*)](\boldsymbol{\theta} - \boldsymbol{\theta}^*)| + |(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T[\widehat{\mathbf{g}}(\bar{\boldsymbol{\theta}})^T(\widehat{\mathbf{C}}^{-1} - \mathbf{C}^{*-1})\mathbf{g}_0(\boldsymbol{\theta}^*)](\boldsymbol{\theta} - \boldsymbol{\theta}^*)|\\
&\quad + |(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T[\widehat{\mathbf{g}}(\bar{\boldsymbol{\theta}})^T\mathbf{C}^{*-1}\{\widehat{\mathbf{g}}(\bar{\boldsymbol{\theta}}) - \mathbf{g}_0(\boldsymbol{\theta}^*)\}](\boldsymbol{\theta} - \boldsymbol{\theta}^*)|\\
&\lesssim \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2\{\|\widehat{\mathbf{g}}(\bar{\boldsymbol{\theta}}) - \mathbf{g}_0(\boldsymbol{\theta}^*)\|_2 + \|\widehat{\mathbf{C}}^{-1} - \mathbf{C}^{*-1}\|_2\}\\
&\lesssim \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2\Big(d_0^{1/2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 + \sqrt{\frac{d_0(s \vee s')\log d(\log n)^2}{n}}\Big),
\end{aligned}$$

where the last step follows from the same argument in the proof of Lemma S.1.7 and (S.20). Since $d_0^{1/2}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2 \le c$, with probability tending to one,

$$(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T\widehat{\mathbf{g}}(\bar{\boldsymbol{\theta}})^T\widehat{\mathbf{C}}^{-1}\widehat{\mathbf{g}}(\bar{\boldsymbol{\theta}})(\boldsymbol{\theta}-\boldsymbol{\theta}^*) \ge (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T\mathbf{g}_0(\boldsymbol{\theta}^*)^T\mathbf{C}^{*-1}\mathbf{g}_0(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*) - c'\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2 \tag{S.24}$$

for some small constant c'. Furthermore, after some tedious calculation, we can show that

$$\begin{aligned}
|(\mathbf{P}_{i1},...,\mathbf{P}_{iK})\widehat{\mathbf{C}}^{-1}\bar{\mathbf{S}}_n(\bar{\boldsymbol{\theta}})| &\lesssim \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^m \delta_{ij}^2 d_0^{1/2}\|\widehat{\mathbf{C}}^{-1}\|_2\|\bar{\mathbf{S}}_n(\bar{\boldsymbol{\theta}})\|_2\\
&\lesssim \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2\sum_{j=1}^m\Big\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_{ij}\boldsymbol{Z}_{ij}^T\Big\|_2 d_0^{1/2}\Big(\sqrt{\frac{d_0}{n}}+\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2\Big)\\
&\lesssim \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2 d_0^{1/2}\Big(\sqrt{\frac{d_0}{n}}+\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2\Big),
\end{aligned}$$

where we bound each term in $\mathbf{P}_{ik}$ by using Lemma S.1.2 and $\|\bar{\mathbf{S}}_n(\bar{\boldsymbol{\theta}})-\bar{\mathbf{S}}_n(\boldsymbol{\theta}^*)\|_2 \lesssim \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2$ by Lemma S.1.7 and the proof of Lemma S.1.11. Since $d_0/n^{1/2}=o(1)$, we can combine (S.23) and (S.24),

$$\begin{aligned}
&\widetilde{Q}_n(\boldsymbol{\theta})-\widetilde{Q}_n(\boldsymbol{\theta}^*)-\nabla\widetilde{Q}_n(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\\
&\ge (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T\mathbf{g}_0(\boldsymbol{\theta}^*)^T\mathbf{C}^{*-1}\mathbf{g}_0(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*) - c'\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2\\
&\quad - C\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2 d_0^{1/2}\Big(\sqrt{\frac{d_0}{n}}+\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2\Big) && \text{(S.25)}\\
&\ge (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T\mathbf{g}_0(\boldsymbol{\theta}^*)^T\mathbf{C}^{*-1}\mathbf{g}_0(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*) - c''\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2, && \text{(S.26)}
\end{aligned}$$

where c'' is a small constant. Finally, we shall bound the minimum eigenvalue of $\mathbf{g}_0(\boldsymbol{\theta}^*)^T\mathbf{C}^{*-1}\mathbf{g}_0(\boldsymbol{\theta}^*)$. Assume that the corresponding eigenvector is denoted by $\boldsymbol{v}^*$. Then

$$\begin{aligned}
\lambda_{\min}\Big[\mathbf{g}_0(\boldsymbol{\theta}^*)^T\mathbf{C}^{*-1}\mathbf{g}_0(\boldsymbol{\theta}^*)\Big] &= \frac{\boldsymbol{v}^{*T}\mathbf{g}_0(\boldsymbol{\theta}^*)^T\mathbf{C}^{*-1}\mathbf{g}_0(\boldsymbol{\theta}^*)\boldsymbol{v}^*}{\|\boldsymbol{v}^*\|_2^2}\\
&= \frac{\eta^{*T}\mathbf{C}^{*-1}\eta^*}{\|(\mathbf{g}_0(\boldsymbol{\theta}^*)^T\mathbf{g}_0(\boldsymbol{\theta}^*))^{-1}\mathbf{g}_0(\boldsymbol{\theta}^*)^T\eta^*\|_2^2}\\
&\ge \frac{\eta^{*T}\mathbf{C}^{*-1}\eta^*}{\|(\mathbf{g}_0(\boldsymbol{\theta}^*)^T\mathbf{g}_0(\boldsymbol{\theta}^*))^{-1}\mathbf{g}_0(\boldsymbol{\theta}^*)^T\|_2^2\|\eta^*\|_2^2},
\end{aligned}$$

where $\eta^*=\mathbf{g}_0(\boldsymbol{\theta}^*)\boldsymbol{v}^*\in\mathbb{R}^{d_0K}$ and therefore $\boldsymbol{v}^*=(\mathbf{g}_0(\boldsymbol{\theta}^*)^T\mathbf{g}_0(\boldsymbol{\theta}^*))^{-1}\mathbf{g}_0(\boldsymbol{\theta}^*)^T\eta^*$. Recall that $\mathbf{g}_0(\boldsymbol{\theta}^*)^T=[\mathbf{H}_1,...,\mathbf{H}_K]\in\mathbb{R}^{d_0\times d_0K}$, where $\mathbf{H}_k=\mathbf{H}_{k\boldsymbol{\theta}\boldsymbol{\theta}}-\mathbf{H}_{k\boldsymbol{\theta}\boldsymbol{\gamma}}\mathbf{H}_{k\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}\mathbf{H}_{k\boldsymbol{\gamma}\boldsymbol{\theta}}$

$$\begin{aligned}
\|(\mathbf{g}_0(\boldsymbol{\theta}^*)^T\mathbf{g}_0(\boldsymbol{\theta}^*))^{-1}\mathbf{g}_0(\boldsymbol{\theta}^*)^T\|_2^2 &= \|(\mathbf{g}_0(\boldsymbol{\theta}^*)^T\mathbf{g}_0(\boldsymbol{\theta}^*))^{-1}\mathbf{g}_0(\boldsymbol{\theta}^*)^T\mathbf{g}_0(\boldsymbol{\theta}^*)(\mathbf{g}_0(\boldsymbol{\theta}^*)^T\mathbf{g}_0(\boldsymbol{\theta}^*))^{-1}\|_2\\
&= \|(\sum_{k=1}^K\mathbf{H}_k\mathbf{H}_k)^{-1}\|_2 = \frac{1}{\lambda_{\min}(\sum_{k=1}^K\mathbf{H}_k\mathbf{H}_k)}\\
&\le \frac{1}{\lambda_{\min}(\mathbf{H}_k\mathbf{H}_k)} \le \frac{1}{\lambda_{\min}^2(\mathbf{H}_k)}
\end{aligned}$$

and

$$\begin{aligned}\lambda_{\min}(\mathbf{H}_k) &\geq \inf_{\|\boldsymbol{v}\|_2=1} \mathbb{E}[(\boldsymbol{X}_i\boldsymbol{v})^T \mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k^{-1}\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)(\boldsymbol{X}_i\boldsymbol{v})] \\ &\geq \inf_{\|\boldsymbol{v}\|_2=1} \mathbb{E}[\|\boldsymbol{X}_i\boldsymbol{v}\|_2^2 \lambda_{\min}(\mathbf{A}_i(\boldsymbol{\beta}^*))\lambda_{\min}(\mathbf{T}_k^{-1})] \\ &\geq C\lambda_{\min}(\mathbb{E}(\boldsymbol{X}_i^T\boldsymbol{X}_i)) \geq C',\end{aligned}$$

for some constants $C, C' > 0$. Thus

$$\lambda_{\min}\Big[\mathbf{g}_0(\boldsymbol{\theta}^*)^T\mathbf{C}^{*-1}\mathbf{g}_0(\boldsymbol{\theta}^*)\Big] \geq C\frac{\eta^{*T}\mathbf{C}^{*-1}\eta^*}{\|\eta^*\|_2^2} \geq C\lambda_{\min}(\mathbf{C}^{*-1}) \geq C'.$$

Plugging above bound into (S.26), it leads to

$$\widetilde{Q}_n(\boldsymbol{\theta}) - \widetilde{Q}_n(\boldsymbol{\theta}^*) - \nabla\widetilde{Q}_n(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*) \geq (C'-c')\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2,$$

uniformly over $\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2 \leq cd_0^{-1/2}$, as long as the constant c is small enough. □

LEMMA S.1.13. Under the same assumptions as in Theorem 2.8, we have

$$\frac{n(\boldsymbol{\theta}^*-\widetilde{\boldsymbol{\theta}})^T\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}^*-\widetilde{\boldsymbol{\theta}}) - d_0}{\sqrt{2d_0}} \rightsquigarrow N(0,1), \quad \text{as} \quad n, d_0 \to \infty,$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ are defined in (2.21) and (2.23), respectively. If d_0 is fixed, under the same condition,

$$n(\boldsymbol{\theta}^*-\widetilde{\boldsymbol{\theta}})^T\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}^*-\widetilde{\boldsymbol{\theta}}) \rightsquigarrow \chi^2_{d_0},$$

where $\chi^2_{d_0}$ denotes the chi-square distribution with d_0 degrees of freedom.

PROOF. The proof follows from Theorem 2.8, Slutsky's Theorem and the fact that

$$\begin{aligned}&d_0^{-1/2}n\big|(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)^T(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{-1}-\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1})(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\big| \\ &\leq d_0^{-1/2}n\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2^2\|\widehat{\mathbf{g}}(\widehat{\boldsymbol{\theta}})^T\widehat{\mathbf{C}}^{-1}\widehat{\mathbf{g}}(\widehat{\boldsymbol{\theta}}) - \mathbf{g}_0(\boldsymbol{\theta}^*)^T\mathbf{C}^{*-1}\mathbf{g}_0(\boldsymbol{\theta}^*)\|_2 \\ &\lesssim \frac{n}{d_0^{1/2}}\frac{d_0}{n}\sqrt{\frac{d_0(s\vee s')\log d(\log n)^2}{n}} \\ &\lesssim \sqrt{\frac{d_0^2(s\vee s')\log d(\log n)^2}{n}} = o_{\mathbb{P}}(1).\end{aligned}$$

This completes the proof. □

LEMMA S.1.14. Suppose Assumptions 2.1, 2.2, 2.3, 2.4 and 2.5 hold for all β_j^* for $j \in [d]$, and suppose that $\lambda \asymp \lambda' \asymp \sqrt{n^{-1}\log d}$ and $n^{-1/2}s(\log d)(\log n) = o(1/\log d)$. Let the "ideal" version of $\widetilde{\beta}_j$ be

$$\text{(S.27)} \qquad \frac{1}{n}\sum_{i=1}^n A_{ij} = \frac{1}{n}\sum_{i=1}^n \big\{\mathbf{g}_{0j}(\beta_j^*)\mathbf{C}_j^{*-1}\mathbf{g}_{0j}(\beta_j^*)\big\}^{-1}\mathbf{g}_{0j}(\beta_j^*)\mathbf{C}_j^{*-1}\cdot\mathbf{S}_{ij}^*(\beta_j^*),$$

where $\mathbf{g}_{0j}^*$, $\mathbf{C}_j^*$ and $\mathbf{S}_{ij}^*(\beta_j^*)$ denote corresponding $\mathbf{g}_0$, $\mathbf{C}^*$ and $\mathbf{S}_i^*(\beta_j^*)$ in the previous section for β_j^*, and let the ideal version of $\widehat{T}_j$ be

$$T_j = \sqrt{n}\sigma_j^{-1}\Big(\frac{1}{n}\sum_{i=1}^n A_{ij}\Big),$$

where σ_j is defined in (2.21) . We have that $\sqrt{n}\widetilde{\beta}_j$ converges to $\frac{1}{\sqrt{n}}\sum_{i=1}^n A_{ij}$ such that as $n\to\infty$,

$$\max_{j\in\mathcal{H}_0}\Big|\sqrt{n}\widetilde{\beta}_j - \frac{1}{\sqrt{n}}\sum_{i=1}^n A_{ij}\Big| = \mathcal{O}_{\mathbb{P}}(n^{-1/2}s(\log d)(\log n)),$$

and

$$\max_j \big|\widehat{T}_j - T_j\big| = \mathcal{O}_{\mathbb{P}}(n^{-1/2}s(\log d)(\log n)).$$

PROOF. The proof follows from Theorem 2.8 and Lemma S.1.13. We omit the details. □

LEMMA S.1.15. Suppose the conditions in Theorem 3.1 holds. Let r_d be any sequence such that $r_d\to\infty$ as $d\to\infty$, and $r_d = o(|\mathcal{S}_1|)$. Then we have

$$\sup_{r_d/d\le u\le 1}\bigg|\frac{\sum_{j\in\mathcal{S}_0}\mathbb{1}\{P_j\le u\}}{u\cdot|\mathcal{S}_0|} - 1\bigg| \to 0,$$

in probability.

PROOF. We denote $W_{nj} = n^{-1/2}\sum_{i=1}^n A_{ij}$ and $z_u := \Phi^{-1}(u)$. We first show that

$$\sup_{r_d/d\le u\le 1}\bigg|\frac{\sum_{i\in\mathcal{S}_0}\mathbb{1}\{|W_{nj}|\ge z_{1-u/2}\}}{|\mathcal{S}_0|\cdot u} - 1\bigg| \xrightarrow{L_2} 0. \tag{S.28}$$

Let $G(u) := \sum_{i\in\mathcal{S}_0}(\mathbb{1}\{|W_{nj}|\ge z_{1-u/2}\} - u)/(|\mathcal{S}_0|\cdot u)$. For any $u\in[r_d/d,1]$, we have

$$\begin{aligned}\mathbb{E}[G(u)^2] = \frac{1}{|\mathcal{S}_0|^2\cdot u^2}\sum_{j,k\in\mathcal{S}_0}\Big\{&\mathbb{P}(|W_{nj}|\ge z_{1-u/2}, |W_{nk}|\ge z_{1-u/2}) + u^2\\ &- u\cdot\mathbb{P}(|W_{nj}|\ge z_{1-u/2}) - u\cdot\mathbb{P}(|W_{nk}|\ge z_{1-u/2})\Big\}\end{aligned}$$

By Lemma 6.1 in Liu (2013),

$$\frac{\big|\sum_{j,k\in\mathcal{S}_0}\{u\cdot\mathbb{P}(|W_{nj}|\ge z_{1-u/2}) - u^2\}\big|}{u^2|\mathcal{S}_0|^2} \le \frac{1}{|\mathcal{S}_0|}\sum_{j\in\mathcal{S}_0}\bigg|\frac{\mathbb{P}(|W_{nj}|\ge z_{1-u/2})}{u} - 1\bigg| \le C(\log d)^{-3/2},$$

and similar inequality holds for the last term. Hence

$$\mathbb{E}[G(u)^2] \le \frac{1}{|\mathcal{S}_0|^2\cdot u^2}\sum_{j,k\in\mathcal{S}_0}\big\{\mathbb{P}(|W_{nj}|\ge z_{1-u/2}, |W_{nk}|\ge z_{1-u/2}) - u^2\big\} + C(\log d)^{-3/2}. \tag{S.29}$$

Now consider the cases $j = k$. We have

(S.30)
$$\frac{1}{|\mathcal{S}_0|^2 \cdot u^2} \sum_{j \in \mathcal{S}_0} \left\{ \mathbb{P}(|W_{nj}| \geq z_{1-u/2}) - u^2 \right\} \leq \frac{1}{|\mathcal{S}_0| \cdot u} \frac{\sum_{j \in \mathcal{S}_0} \mathbb{P}(|W_{nj}| \geq z_{1-u/2})}{|\mathcal{S}_0| \cdot u} \leq \frac{1}{|\mathcal{S}_0| \cdot u} \big(1 + o(1)\big).$$

For the case that $j \neq k$, let $\mathcal{A} := \{(j,k) : j \neq k, \Omega_{jk} \geq (\log d)^{-5/2}\}$. By Lemma 6.1 of Liu (2013), it holds

$$\text{(S.31)} \qquad \frac{1}{|\mathcal{S}_0|^2 \cdot u^2} \sum_{(j,k) \in \mathcal{A}^c} \left| \mathbb{P}(|W_{nj}| \geq z_{1-u/2}, |W_{nk}| \geq z_{1-u/2}) - u^2 \right| \leq C(\log d)^{-3/2}.$$

If $(j,k) \in \mathcal{A}$, then by Lemma 6.2 of Liu (2013), we have

$$\begin{aligned} &\frac{1}{|\mathcal{S}_0|^2 \cdot u^2} \sum_{(j,k) \in \mathcal{A}} \left\{ \mathbb{P}(|W_{nj}| \geq z_{1-u/2}, |W_{nk}| \geq z_{1-u/2}) - u^2 \right\} \\ &\leq \frac{1}{|\mathcal{S}_0|^2 \cdot u^2} \sum_{(j,k) \in \mathcal{A}} \mathbb{P}(|W_{nj}| \geq z_{1-u/2}, |W_{nk}| \geq z_{1-u/2}) \\ \text{(S.32)} \qquad &\leq \frac{1}{|\mathcal{S}_0|^2} \sum_{(j,k) \in \mathcal{A}} \frac{C \cdot (z_{1-u/2} + 1)^{-2} \exp(-z_{1-u/2}^2 / (1 + |\Omega_{jk}|))}{u^2}. \end{aligned}$$

When $0 \leq z_{1-u/2} \leq 1$, we have $u \geq 2(1 - \Phi(1))$, hence each summand in (S.32) is uniformly upper bounded by a constant. If $1 < z_{1-u/2} \leq z_{1-r_d/(2d)}$, we have $u = 2(1 - \Phi(z_{1-u/2})) \geq (1/\sqrt{2\pi}) \cdot z_{1-u/2}/(z_{1-u/2}^2 + 1) \exp(-z_{1-u/2}^2/2)$, and hence

$$\begin{aligned} \sum_{(j,k) \in \mathcal{A}} \frac{(z_{1-u/2} + 1)^{-2} \exp(-z_{1-u/2}^2 / (1 + |\Omega_{jk}|))}{|\mathcal{S}_0|^2 \cdot u^2} &\leq \frac{1}{|\mathcal{S}_0|^2} \sum_{j,k \in \mathcal{A}} \frac{2\pi \cdot (z_{1-u/2}^2 + 1)^2}{(z_{1-u/2} + 1)^2 z_{1-u/2}^2} \exp\left\{ \frac{z_{1-u/2}^2 |\Omega_{jk}|}{(1 + |\Omega_{jk}|)} \right\} \\ &\leq \frac{1}{|\mathcal{S}_0|^2} \sum_{j,k \in \mathcal{A}} 4\pi \cdot d^{\frac{2|\Omega_{jk}|}{1 + |\Omega_{jk}|}} \lesssim (\log d)^{-3/2}, \end{aligned}$$

where in the first inequality we used the fact that $z_{1-r_d/(2d)} \leq \sqrt{2 \log d}$, and the last inequality is by the condition (3.2) in the theorem and the fact that $|\mathcal{S}_0|/d \to 1$. Therefore, by the above arguments, we conclude that

$$\text{(S.33)} \qquad \frac{1}{|\mathcal{S}_0|^2 \cdot u^2} \sum_{(j,k) \in \mathcal{A}} \left\{ \mathbb{P}(|W_{nj}| \geq z_{1-u/2}, |W_{nk}| \geq z_{1-u/2}) - u^2 \right\} \lesssim (\log d)^{-3/2}.$$

Combining (S.29), (S.30), (S.31) and (S.33), we obtain

$$\text{(S.34)} \qquad \mathbb{E}[G(u)^2] \leq C \cdot \frac{1}{|\mathcal{S}_0| \cdot u} + C' \cdot (\log d)^{-3/2},$$

for any $r_d/d \leq u \leq 1$. As $r_d \to \infty$ and $|\mathcal{S}_0|/d \to 1$, we get $1/(|\mathcal{S}_0| \cdot u) \leq d/(r_d \cdot |\mathcal{S}_0|) \to 0$, and so $G(u) \xrightarrow{L_2} 0$. We use the following argument to show that the convergence is uniform in $r_d/d \leq u \leq 1$. Define $u_\ell := r_d/d + \sqrt{r_d} \exp(\ell^{4/5})/d$, for $\ell = 1, \ldots, L-1$, and $u_0 = r_d/d$, $u_L = 1$, where $L = \lceil \log^{5/4}(d/\sqrt{r_d} - \sqrt{r_d}) \rceil$. For any $u \in [r_d/d, 1]$, there exists $\ell \in \{1, \ldots, L-1\}$ such that $u_\ell \le u \le u_{\ell+1}$. Hence,

(S.35)
$$G(u) \le \max\left\{ \left| \frac{\sum_{i\in\mathcal{S}_0}(\mathbb{1}\{|W_{nj}| \ge z_{1-u_{\ell+1}/2}\} - u_\ell)}{d\cdot u} \right|, \left| \frac{\sum_{i\in\mathcal{S}_0}(\mathbb{1}\{|W_{nj}| \ge z_{1-u_\ell/2}\} - u_{\ell+1})}{d\cdot u} \right| \right\},$$

and we have

$$\left| \frac{\sum_{i\in\mathcal{S}_0}(\mathbb{1}\{|W_{nj}| \ge z_{1-u_{\ell+1}/2}\} - u_\ell)}{d\cdot u} \right| \le \left(|I(u_{\ell+1})| + \left| \frac{u_\ell}{u_{\ell+1}} - 1 \right| \right) \cdot \frac{u_{\ell+1}}{u_\ell},$$

where

$$I(u) = \frac{\sum_{i\in\mathcal{S}_0}(\mathbb{1}\{|W_{nj}| \ge z_{1-u/2}\} - u)}{d\cdot u}.$$

Similar argument also applies to the second term in the R.H.S. of (S.35). Note that $1 \ge u_\ell/u_{\ell+1} = (\sqrt{r_d} + \exp(\ell^{4/5}))/(\sqrt{r_d} + \exp((\ell+1)^{4/5})) \ge \sqrt{r_d}/(\sqrt{r_d} + e) \to 1$ for all ℓ, which implies that $u_\ell/u_{\ell+1} \to 1$ uniformly over $\ell \in \{0, \ldots, L\}$ as $d \to \infty$. Therefore, for large enough d, we have

$$\mathbb{E}\Big[\sup_{r_d/d \le u \le 1} G(u)^2 \Big] \le 4 \cdot \mathbb{E}\Big[\max_{0\le\ell\le L} G(u_\ell)^2 + o(1) \Big] \le 4 \sum_{\ell=0}^{L} \mathbb{E}[G(u_\ell)^2] + o(1).$$

Note that $L \le (\log d)^{5/4}$ and $|\mathcal{S}_0|/d \to 1$. Hence by (S.34), we get

$$\sum_{\ell=0}^{L} \mathbb{E}[G(u_\ell)^2] \le \sum_{\ell=0}^{L} C \cdot \frac{1}{|\mathcal{S}_0|(r_d/d + \sqrt{r_d}\exp(\ell^{4/5})/d)} + L \cdot C' \cdot (\log d)^{-3/2} = o(1).$$

Therefore, we proved (S.28). Lemma S.1.14 further implies

$$\Lambda_{nj} = (W_{nj} + o_{\mathbb{P}}(1/\log d))^2,$$

uniformly in $j \in \mathcal{S}_0$. Note that $2(1 - \Phi(z_{1-u/2} + 1/\log d))/u \to 0$ uniformly over $r_d/d \le u \le 1$. Hence by (S.28) and similar argument as (S.35), we have

$$\sup_{r_d/d\le u\le 1} \left| \frac{\sum_{j\in\mathcal{S}_0} \mathbb{1}\{P_j \le u\}}{u\cdot|\mathcal{S}_0|} - 1 \right| = \sup_{r_d/d\le u\le 1} \left| \frac{\sum_{j\in\mathcal{S}_0} \mathbb{1}\{|W_{nj} + o_{\mathbb{P}}(1/\log d)| \ge z_{1-u/2}\}}{u\cdot|\mathcal{S}_0|} - 1 \right| \to 0,$$

in probability. This concludes the proof. □

LEMMA S.1.16. For any matrices $\mathbf{A}_i \in \mathbb{R}^{a\times c}$ and $\mathbf{B}_i \in \mathbb{R}^{c\times b}$,

$$\Big\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{A}_i\mathbf{B}_i \Big\|_2 \le \Big\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{A}_i\mathbf{A}_i^T \Big\|_2^{1/2} \Big\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{B}_i^T\mathbf{B}_i \Big\|_2^{1/2}$$

The proof of this simple lemma follows from the definition of the spectral norm and Cauchy-Schwartz inequality. We omit the proof.

S.1.1. *Proof of Corollary 2.10.* The first statement $\text{Avar}(\widetilde{\theta}) \le \text{Avar}(\widetilde{\theta}_I)$ is a standard result in the generalized method of moment. We omit the proof. To show the second statement, the efficient influence function for $\boldsymbol{\beta}$ (e.g., Tsiatis (2007) Eq. (4.54)) is

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i^T \mathbf{A}_i^{1/2}(\boldsymbol{\beta})\mathbf{R}^{-1}\mathbf{A}_i^{-1/2}(\boldsymbol{\beta})\big\{\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\big\},$$

which yields the semiparametric information bound $\mathbb{E}(\boldsymbol{X}^T\mathbf{A}^{1/2}\mathbf{R}^{-1}\mathbf{A}^{1/2}\boldsymbol{X})^{-1}$. The corresponding efficient influence function for the parameter of interest θ is

$$\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{v}^*)^T \mathbf{A}_i^{1/2}(\boldsymbol{\beta})\mathbf{R}^{-1}\mathbf{A}_i^{-1/2}(\boldsymbol{\beta})\big\{\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\big\},$$

where $\mathbf{v}^* = \mathbb{E}(\boldsymbol{U}_i^T\mathbf{A}^{1/2}\mathbf{R}^{-1}\mathbf{A}^{1/2}\boldsymbol{U}_i)^{-1}\mathbb{E}(\boldsymbol{U}_i^T\mathbf{A}^{1/2}\mathbf{R}^{-1}\mathbf{A}^{1/2}\boldsymbol{Z}_i)$. Recall that $\widetilde{\boldsymbol{Z}}_i = \mathbf{A}^{1/2}\boldsymbol{Z}_i$ and $\widetilde{\boldsymbol{U}}_i = \mathbf{A}^{1/2}\boldsymbol{U}_i$ and (2.17) holds. We have that for all i,

$$\mathbb{E}(\boldsymbol{U}_i^T\mathbf{A}^{1/2}\mathbf{R}^{-1}\mathbf{A}^{1/2}\boldsymbol{Z}_i) = \mathbb{E}(\widetilde{\boldsymbol{U}}_i^T\mathbf{R}^{-1}\widetilde{\boldsymbol{Z}}_i) = \mathbb{E}(\widetilde{\boldsymbol{U}}_i^T\mathbf{R}^{-1}\widetilde{\boldsymbol{U}}_i\mathbf{w}^* + \widetilde{\boldsymbol{U}}_i^T\mathbf{R}^{-1}\boldsymbol{\delta}) = \mathbb{E}(\widetilde{\boldsymbol{U}}_i^T\mathbf{R}^{-1}\widetilde{\boldsymbol{U}}_i)\mathbf{w}^*,$$

and therefore $\mathbf{w}^* = \mathbf{v}^*$. Finally, the estimating equation theory implies that the proposed method is optimal among the class of estimating equations

$$\sum_{k=1}^{K} C_k \sum_{i=1}^{n}(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{w}^*)^T \mathbf{A}_i^{1/2}(\boldsymbol{\beta})\mathbf{T}_k\mathbf{A}_i^{-1/2}(\boldsymbol{\beta})\big\{\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\big\},$$

where $C_1, ..., C_K \in \mathbb{R}$. Since $\mathbf{R}^{-1} = \sum_{k=1}^{K} a_k\mathbf{T}_k$ for some constants $a_1, ..., a_K$, by taking $C_k = a_k$ the above estimating equation reduces to the efficient influence function for the parameter of interest θ. This implies that our method is efficient in the sense that it attains the semiparametric information bound.

S.2. A Brief Description of Generalized Estimating Equation. In this section, we briefly review the GEE method for the analysis of longitudinal data.

Recall notation in Section 2, Liang and Zeger (1986) propose to estimate $\boldsymbol{\beta}^*$ by solving the following generalized estimating equation (GEE),

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\boldsymbol{\mu}_i(\eta_i)}{\partial\boldsymbol{\beta}}\mathbf{V}_{i(w)}^{-1}\big\{\mathbf{Y}_i - \boldsymbol{\mu}_i(\eta_i)\big\} = 0,$$

where $\eta_i = (\eta_{i1}, ..., \eta_{im})^T = (\boldsymbol{X}_{i1}^T\boldsymbol{\beta}, ..., \boldsymbol{X}_{im}^T\boldsymbol{\beta})^T \in \mathbb{R}^m$, $\boldsymbol{\mu}_i(\eta_i) = \big\{\mu_{i1}(\eta_{i1}), ..., \mu_{im}(\eta_{im})\big\}^T \in \mathbb{R}^m$, and $\mathbf{V}_{i(w)}$ is a working $m\times m$ working covariance matrix of $\mathbf{Y}_i$. Let $\mathbf{A}_i(\boldsymbol{\beta}) = \text{diag}\big\{\mu'_{i1}(\eta_{i1}), ..., \mu'_{im}(\eta_{im})\big\}$. Then the covariance matrix $\mathbf{V}_{i(w)}$ can be decomposed as $\mathbf{V}_{i(w)} = \mathbf{A}_i^{1/2}(\boldsymbol{\beta})\mathbf{R}(\boldsymbol{\phi})\mathbf{A}_i^{1/2}(\boldsymbol{\beta})$, where $\mathbf{R}(\boldsymbol{\phi})$ is a working $m \times m$ correlation matrix of $\mathbf{Y}_i$ indexed by a finite dimensional parameter $\boldsymbol{\phi}$. For the canonical exponential family, we have $\partial\boldsymbol{\mu}_i(\eta_i)/\partial\boldsymbol{\beta} = \boldsymbol{X}_i^T\mathbf{A}_i(\boldsymbol{\beta})$. For notational simplicity, we rewrite $\boldsymbol{\mu}_i(\eta_i)$ as $\boldsymbol{\mu}_i(\boldsymbol{\beta})$, since η_i is parameterized by $\boldsymbol{\beta}$. Then the GEE can be reformulated as

$$\text{(S.1)} \qquad \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i^T \mathbf{A}_i^{1/2}(\boldsymbol{\beta})\widehat{\mathbf{R}}^{-1}\mathbf{A}_i^{-1/2}(\boldsymbol{\beta})\big\{\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\big\} = 0,$$

where $\widehat{\mathbf{R}}$ is an estimator of $\mathbf{R}(\boldsymbol{\phi})$. We denote the root of this equation by $\widehat{\boldsymbol{\beta}}$. When d is fixed and n goes to infinity, Liang and Zeger (1986) show that

$$\sqrt{n}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}) \rightsquigarrow N(\mathbf{0}, \mathbf{V}), \tag{S.2}$$

where the covariance matrix $\mathbf{V}$ is

$$\mathbf{V} = \Big(\sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_{i(w)}^{-1} \mathbf{D}_i\Big)^{-1} \Big\{\sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_{i(w)}^{-1} \operatorname{Cov}(\boldsymbol{Y}_i) \mathbf{V}_{i(w)}^{-1} \mathbf{D}_i\Big\} \Big(\sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_{i(w)}^{-1} \mathbf{D}_i\Big)^{-1},$$

and $\mathbf{D}_i = \partial \boldsymbol{\mu}_i(\boldsymbol{\beta}^*)/\partial \boldsymbol{\beta}$. Xie and Yang (2003); Ma (2012) further establish the large sample theory for $\widehat{\boldsymbol{\beta}}$, when the cluster size m is large and potentially increases with n. Wang (2011) extends the asymptotics to the "large n, diverging d" regime and establishes $\sqrt{d/n}$-consistency of the GEE estimator (S.1) when $n^{-1}d^2 = o(1)$. In addition, assuming the sparsity of $\boldsymbol{\beta}^*$, Wang et al. (2012) propose a penalized generalized estimating equation framework, and show that the penalized estimator achieves the oracle property when $d = O(n)$. The result is further extended to the generalized additive partial linear models by Lian et al. (2014).

S.3. Motivating Examples in High-dimensional Longitudinal Data. In this section, we present two concrete motivating examples under $d = O(n)$ and $\log d = o(n)$, respectively.

Example 1: BMI genomic dataset. The Framingham Heart Study (FHS) is a long-term, ongoing cardiovascular study beginning in 1948 under the direction of the National Heart, Lung and Blood institute (NHLBI) on residents of the town of Framingham, Massachusetts. The objective is to identify the important characteristics that contribute to cardiovascular disease. We refer to Jaquish (2007) for more details of the study. Recently, 913,854 SNPs from 24 chromosomes have been genotyped from the Offspring Cohort study. The goal is to investigate the issue of obesity as the body mass index (BMI), which is measured at multiple time points. Our dataset contains BMI of 977 samples, where each samples BMI value is collected at 26 times. This example falls into the ultra-high dimensional setting, i.e., $\log d = o(n)$. We will evaluate our proposed method using this example in Section 4.2.

Example 2: Yeast cell-cycle gene expression data. The dataset analyzed by Wang et al. (2012) contains 297 samples of cell-cycle-regularized genes. The goal of the analysis is to understand how the gene expression level is associated with 96 transcription factors. Both the gene expression and the score of each transcription factor are measured longitudinally at 5 common time points. In this example, the number of samples is comparable to the number of features, and the asymptotic analysis under the $d = O(n)$ setting seems appropriate. Finally, we note that the proposed method developed under the ultra-high dimensional setting is directly applicable to the $d = O(n)$ setting.

S.4. More Simulations. In this section, we conduct more simulation studies to empirically test the proposed method under different scenarios.

S.4.1. *Settings where $d_0 > 1$.* We investigate the empirical coverage rates for the confidence region when $d_0 = 3$ and 5, where d_0 is the number of parameters of interest. We follow settings corresponding to the settings in Table 1 in the main text. We consider confidence regions for $\beta_1, \beta_2, ..., \beta_{d_0}$, where we set $\beta_1^* = \beta_2^* = 0$, and the rest are generated from either Dirac or Unif$[0, 2]$. We report the empirical coverage rate where the nominal level is set as 95% in Table S1. It is seen that when $d_0 = 3$, our method performs reasonably well. When $d_0 = 5$, as the sparsity level s increases, especially when $s = 20$, the empirical coverage probabilities deviate from the desired nominal level as expected.

Table S1

Empirical coverage probability (%) under equal-correlation with correlation parameter being 0.5 and $m = 3$. The nominal level is set to be 95%.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
(n, d)	s	Dirac	U[0, 2]	Dirac	U[0, 2]	Dirac	U[0, 2]	Dirac	U[0, 2]
					$d_0 = 3$				
(50,200)	5	94.3	95.5	95.7	93.8	94.5	96.3	93.0	94.2
	10	95.2	96.3	93.4	96.2	94.3	96.3	95.5	93.9
	20	93.7	94.8	94.6	95.3	94.1	96.1	95.9	93.6
(100,500)	5	94.6	95.6	93.8	96.0	94.9	94.2	97.0	94.3
	10	96.3	93.7	94.0	95.9	96.4	93.4	93.9	96.8
	20	95.5	94.4	93.2	93.5	92.8	92.5	93.3	92.6
(100,1000)	5	93.9	92.6	94.5	92.8	93.2	94.1	92.5	92.9
	10	93.1	93.5	92.4	94.0	92.9	93.3	93.0	93.4
	20	93.4	92.7	94.2	93.8	92.9	91.8	93.5	92.0
					$d_0 = 5$				
(50,200)	5	93.8	94.3	92.6	93.9	93.0	92.9	92.7	93.1
	10	90.3	88.5	91.5	89.9	89.3	86.8	90.8	87.0
	20	87.6	85.3	86.8	86.2	87.4	85.1	86.1	84.3
(100,500)	5	92.0	93.1	92.8	93.4	93.5	93.0	92.4	93.8
	10	89.5	87.4	88.8	85.6	88.0	84.6	86.2	83.5
	20	84.6	83.8	84.0	84.9	83.7	83.4	82.2	82.6
(100,1000)	5	93.4	93.7	92.6	92.5	93.0	92.4	91.8	93.1
	10	88.3	86.2	87.9	83.8	84.3	86.6	85.0	83.1
	20	82.0	85.7	84.5	79.9	82.4	78.3	81.7	76.3

S.4.2. *Different within-sample correlation structure.* We consider the case where the within-sample correlation matrix is a convex combination of equal and AR-correlation matrices with equal weight. We report the empirical Type I error and FDR in Tables S2 and S3, respectively. We find that the proposed method works well in this setting.

S.4.3. *Different within-sample observations m_i.* We consider settings where numbers, m_i's, of within-sample observations is different. While we assume $m_i = m$ in the main text, the proposed method can be generalized to unequal m_i if the within sample correlation structures are the same, like AR/equal correlation models with the same parameter. To demonstrate the applicability of our method, consider the following numerical study. Specifically, we consider both equal and AR correlation models where half of the subjects have $m_1 = 3$ samples, and the other half have $m_2 = 5$ samples. We follow the same data generating schemes in Tables 1 and 2 in the main text. We investigate the empirical Type I error rate and the false discovery rate. We present the results in Tables S4, S5, S6 and S7. We see that our proposed methods still work as desired.

Finally, we note that in a more complicated setting, where Y_{ij} is the response measured at a random time T_{ij} and the number of observations m_i on each subject is considered random, reflecting sparse and irregular design, assuming AR/equal correlation is very strong and thus the QIF approach may not be directly applicable. A more flexible approach as shown by Yao et al. (2005) is to treat the data as sparse functional (or longitudinal) data and apply functional component analysis approach. The extension to sparse longitudinal data is beyond the scope of this paper, and is left for future research.

Table S2

Empirical Type I error rate (%) under the convex combination of equal-correlation and AR-correlation with correlation parameter being 0.5 and 0.6, respectively. The nominal level is set to be 5%.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
(n, d)	s	Dirac	U[0, 2]	Dirac	U[0, 2]	Dirac	U[0, 2]	Dirac	U[0, 2]
					$m = 3$				
(50,200)	5	5.3	5.6	5.2	5.5	5.1	4.8	5.3	5.7
	10	5.5	4.7	5.0	5.6	5.8	5.2	5.7	5.5
	20	5.4	6.0	5.8	6.1	5.9	5.5	6.0	5.3
(100,500)	5	5.7	5.9	5.2	5.5	4.7	5.9	5.4	5.6
	10	5.9	5.2	5.9	5.5	5.4	5.7	5.3	5.6
	20	5.4	5.8	5.5	5.3	6.0	5.7	5.8	5.9
(100,1000)	5	5.1	5.0	5.6	5.3	5.5	5.6	4.9	5.3
	10	5.4	5.4	5.7	5.2	4.7	5.3	5.8	5.5
	20	5.8	5.6	6.0	5.5	5.7	6.2	5.9	6.0
					$m = 5$				
(50,200)	5	5.5	5.3	5.0	5.2	4.9	5.3	5.7	5.4
	10	5.3	5.4	5.0	5.5	4.6	6.0	5.9	5.3
	20	5.7	5.6	5.8	5.2	5.9	5.0	6.0	5.7
(100,500)	5	5.4	5.1	5.3	5.5	5.0	4.9	4.7	5.4
	10	5.4	5.2	5.5	5.8	6.0	5.7	5.9	5.5
	20	5.7	5.5	5.2	5.6	5.9	5.4	4.9	6.2
(100,1000)	5	5.2	5.5	5.4	5.1	5.0	5.7	5.4	5.8
	10	5.6	5.5	5.8	5.2	5.7	5.9	5.6	5.4
	20	5.5	6.1	5.3	5.4	5.9	5.5	6.1	6.0

S.4.4. *Logistic Model.* Under the logistic model, we follow the simulation setup as in Cho and Qu (2015) with equal correlation structures. We report the resulting empirical Type I error rate and FDR in Table S8 and S9. It is seen that our method works well under the logistic model.

S.4.5. *Low-dimensional settings.* We consider the performance of the proposed method QDIF in comparison with the classical QIF method. In particular, we first consider a low-dimensional setting where $d \ll n$ with $m = 3$, and we set $\beta_1^* = 0$, and consider the Dirac and Unif$[0, 2]$ settings to generate other components. In Tables S10 and S11, we consider equal-correlation and AR-correlation structure with correlation parameter being 0.5. It is seen that both of the two methods well control the empirical Type I error rate as desired. In addition, we consider some moderate dimensional settings in Tables S12 and S13, where we set $d = 100, 150$ and 200. We find that the traditional QIF method does not successfully control the Type I error rate, while the proposed QDIF method still works.

S.4.6. *Comparison with GEE.* To compare the performance of QIF and GEE in high dimension, we need to consider a decorrelated GEE approach, as the traditional GEE estimator (and the inference result) is not applicable in high dimension. The decorrelated GEE estimator is defined as follows. We use the same initial estimator $\widehat{\boldsymbol{\beta}}$ defined in section 2.3 in main text. Assuming the working correlation $\boldsymbol{R}(\eta)$ is parameterized by a finite dimensional parameter η, we plug in the initial estimator $\widehat{\boldsymbol{\beta}}$ to construct a moment estimator for η, say $\widehat{\eta}$. Then, we may simply apply our method with $K = 1$ and $T_1 = \left\{\boldsymbol{R}(\widehat{\eta})\right\}^{-1}$. The root of the quasi-score function $\bar{\mathbf{S}}_n(\boldsymbol{\theta})$ in (2.7) in

TABLE S3
Empirical false discovery rate (%) at level $\alpha = 0.1$ and 0.2 under the convex combination of equal-correlation and AR-correlation with correlation parameter being 0.5 and 0.6, respectively.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
	α	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2
(n, d)	s	$m = 3$							
(50,200)	5	10.3	19.6	10.2	20.4	9.5	19.4	9.8	20.6
	10	9.8	19.4	9.5	19.5	10.3	20.3	10.7	19.1
	20	10.8	20.5	10.6	19.7	9.5	19.0	9.6	20.8
(100,500)	5	10.0	19.2	9.7	19.6	10.5	20.3	9.4	20.9
	10	9.5	19.7	9.1	20.4	9.4	20.8	10.0	20.5
	20	10.8	18.9	10.4	19.6	9.1	19.2	10.3	19.0
(100,1000)	5	9.8	20.3	9.5	20.8	9.3	19.6	10.4	19.2
	10	10.0	19.4	9.7	20.3	9.5	20.8	10.4	19.2
	20	9.3	19.0	9.6	20.6	9.5	20.9	9.1	19.1
		$m = 5$							
(50,200)	5	9.9	20.5	9.3	20.3	10.4	19.7	10.1	20.8
	10	9.5	20.3	10.4	19.5	10.0	20.4	10.6	19.2
	20	10.4	20.8	9.1	20.0	9.5	21.0	8.9	20.1
(100,500)	5	9.5	20.6	9.2	19.4	10.3	20.2	10.6	20.9
	10	9.1	20.5	9.8	20.1	9.0	19.3	10.3	20.5
	20	10.6	18.8	10.1	19.5	10.8	20.3	11.0	21.0
(100,1000)	5	9.6	20.4	10.5	19.6	10.2	19.7	9.7	20.1
	10	10.7	19.4	10.0	19.8	9.4	20.4	10.5	20.6
	20	9.4	20.5	9.7	19.4	9.0	18.9	9.1	19.0

TABLE S4
Empirical Type I error rate (%) under equal-correlation with correlation parameter being 0.5, where half of the subjects have $m_1 = 3$ samples, and the other half have $m_2 = 5$ samples. The nominal level is set to be 5%.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
(n, d)	s	Dirac	U[0, 2]	Dirac	U[0, 2]	Dirac	U[0, 2]	Dirac	U[0, 2]
(50,200)	5	5.4	5.6	5.3	5.6	5.0	5.2	5.4	5.6
	10	5.6	5.5	5.3	5.2	4.7	4.9	5.4	5.1
	20	5.5	5.2	5.6	5.7	5.1	6.2	4.5	5.9
(100,500)	5	5.3	5.6	5.0	4.8	5.4	5.8	5.2	5.5
	10	5.5	5.2	5.8	5.4	5.0	5.4	6.0	5.9
	20	5.9	6.1	5.5	4.7	5.4	5.8	5.9	5.6
(100,1000)	5	5.2	5.4	5.5	5.3	5.8	5.4	4.8	5.3
	10	5.5	5.8	5.4	5.6	5.8	5.5	5.7	5.0
	20	5.9	5.5	5.8	6.2	5.3	5.6	4.7	6.3

main text (which is now a d_0 dimensional vector as $K = 1$) is our decorrelated GEE estimator.

We then conduct simulation to empirically test the decorrelated GEE method, where we consider the cases corresponding to Tables 1 and 2 in the main text where the true within sample correlation matrices are equal-correlation and AR-correlation matrices, respectively. The working correlation $\boldsymbol{R}(\eta)$ in decorrelated GEE approach is the equal-correlation matrix. We report the empirical Type I error rates in Tables S14 and S15. It is seen that when the within sample correlation matrix is

Table S5

Empirical Type I error rate (%) under AR-correlation with correlation parameter being 0.5, where half of the subjects have $m_1 = 3$ samples, and the other half have $m_2 = 5$ samples. The nominal level is set to be 5%.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
(n, d)	s	Dirac	U[0, 2]	Dirac	U[0, 2]	Dirac	U[0, 2]	Dirac	U[0, 2]
(50,200)	5	5.1	5.4	5.0	4.7	5.2	5.3	5.1	5.4
	10	5.3	4.8	5.1	5.4	5.0	5.3	5.2	5.5
	20	5.4	5.6	5.2	5.5	5.6	5.9	6.1	4.7
(100,500)	5	5.5	5.1	5.2	4.7	5.3	5.2	5.0	5.4
	10	5.7	5.4	5.2	5.1	5.3	5.8	4.9	5.6
	20	5.4	5.8	5.7	6.0	6.1	5.5	5.2	5.9
(100,1000)	5	5.4	5.1	5.3	5.6	5.2	5.7	4.7	5.2
	10	5.2	5.3	5.6	5.2	5.7	4.6	5.0	5.9
	20	5.5	5.8	5.4	6.1	5.5	5.9	6.4	5.7

Table S6

Empirical false discovery rate (%) at level $\alpha = 0.1$ and 0.2 under equal-correlation structure with correlation parameter being 0.5 where half of the subjects have 3 samples, and the other half subjects have 5 samples.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
	α	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2
(n, d)	s								
(50,200)	5	10.4	20.7	9.8	19.5	9.5	20.3	10.1	20.1
	10	9.3	20.1	9.0	20.9	9.8	20.4	10.7	19.3
	20	9.0	19.3	9.3	19.1	9.5	20.3	9.4	18.8
(100,500)	5	10.3	19.5	10.9	19.2	10.6	20.4	9.6	21.0
	10	9.5	20.4	9.9	20.9	9.2	20.7	10.3	19.4
	20	9.2	19.3	8.8	19.5	10.0	20.9	10.7	20.6
(100,1000)	5	10.3	19.5	9.7	20.4	9.7	20.9	10.2	20.3
	10	9.9	20.4	9.6	19.4	9.8	19.1	21.1	19.6
	20	9.1	20.9	10.9	20.6	10.6	21.3	10.7	20.8

Table S7

Empirical false discovery rate (%) at level $\alpha = 0.1$ and 0.2 under AR-correlation structure with correlation parameter being 0.5 where half of the subjects have 3 samples, and the other half subjects have 5 samples.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
	α	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2
(n, d)	s								
(50,200)	5	10.2	19.8	9.6	20.5	10.4	20.0	9.3	19.4
	10	10.4	19.1	9.6	19.8	9.3	20.5	11.0	20.8
	20	9.5	19.5	9.2	19.4	9.5	20.6	9.0	21.2
(100,500)	5	10.6	19.2	9.7	19.5	10.4	20.6	9.5	20.1
	10	9.8	20.5	9.4	19.6	9.5	20.1	10.3	20.9
	20	8.9	19.3	9.5	20.2	9.3	21.1	10.8	20.7
(100,1000)	5	9.7	20.3	10.0	20.8	9.6	19.4	10.5	19.7
	10	10.2	20.5	9.4	20.1	10.5	19.3	10.9	19.7
	20	9.2	20.6	10.5	19.5	10.9	18.8	11.0	21.2

Table S8

Empirical Type I error rate (%) under equal-correlation with correlation parameter being 0.5 under the logistic model. The nominal level is set to be 5%.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
(n, d)	s	Dirac	U[0, 2]	Dirac	U[0, 2]	Dirac	U[0, 2]	Dirac	U[0, 2]
					$m = 3$				
(50,200)	5	5.4	5.7	4.8	4.7	5.3	4.9	5.6	5.4
	10	6.0	5.5	5.3	5.7	5.4	4.8	4.1	5.0
	20	5.7	5.0	4.5	4.3	6.3	3.9	5.5	5.9
(100,500)	5	5.2	5.9	5.3	4.3	5.0	5.9	4.1	4.0
	10	6.1	6.2	5.8	5.7	4.6	4.8	5.5	6.2
	20	5.5	5.9	5.2	5.9	6.2	6.0	5.8	6.1
(100,1000)	5	5.0	5.4	6.1	6.2	5.7	6.0	5.4	4.8
	10	6.0	5.5	5.2	5.8	5.5	6.1	6.5	5.9
	20	6.3	5.9	5.6	6.0	5.3	4.9	6.3	6.1
					$m = 5$				
(50,200)	5	5.3	5.5	5.4	5.0	4.8	4.9	5.6	5.7
	10	5.7	5.6	5.9	5.5	5.6	6.0	6.1	5.9
	20	5.8	5.9	5.5	5.1	6.0	6.2	5.8	6.1
(100,500)	5	5.4	5.2	5.7	6.0	4.8	5.9	5.8	5.6
	10	5.8	5.7	6.1	5.8	5.7	6.0	5.9	6.2
	20	6.0	5.5	5.7	5.9	5.5	4.9	6.0	6.3
(100,1000)	5	5.4	5.1	5.8	5.1	5.7	5.9	5.5	5.9
	10	5.9	6.1	5.5	5.8	6.0	5.7	6.3	6.2
	20	5.7	5.9	6.0	5.4	6.2	5.9	6.4	5.7

correctly specified, the performance of the decorrelated GEE is similar to QDIF as shown in Table S14. However, when the correlation matrix is misspecified, the performance of GEE is much worse than QDIF, by comparing the type I errors in Table S15 for the decorrelated GEE estimator with Table 2 in main text for QDIF. This result shows that in practice misspecification of the correlation structure has a large effect on the performance of the GEE estimators in high dimension. We think this is mainly due to the instability of the estimated correlation $\boldsymbol{R}(\widehat{\eta})$. This confirms that the proposed QDIF approach is preferred.

S.4.7. *Different error distribution.* We consider the case where the errors follow the T distribution with 5 degrees of freedom. In particular, we consider the same data generating schemes as in Tables 1 and 2 in the main text, except that we generate the noises by the T distribution and we standardize the error's marginal variance to be 1. We report empirical Type I error rates in Tables S16 and S17, and the empirical FDR in Tables S18 and S19. It is seen that our method still controls the empirical Type I error and FDR.

Table S9

Empirical false discovery rate (%) at level $\alpha = 0.1$ and 0.2 under equal-correlation structure with correlation parameter being 0.5 for the logistic model.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
	α	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2
(n, d)	s				$m = 3$				
(50,200)	5	9.6	18.7	9.1	19.3	8.8	19.7	9.6	19.5
	10	8.5	18.5	8.8	19.3	8.7	19.8	9.5	20.4
	20	9.3	18.8	9.0	19.4	8.5	18.7	8.8	19.0
(100,500)	5	9.3	18.5	9.0	18.9	10.3	18.7	9.2	20.2
	10	9.1	19.3	8.5	20.3	8.8	18.6	9.7	20.9
	20	8.9	18.7	9.0	18.4	8.5	19.3	9.0	18.9
(100,1000)	5	10.5	18.7	9.4	19.8	9.1	19.3	9.9	19.5
	10	9.6	18.8	9.3	18.6	9.6	20.4	8.7	19.0
	20	9.2	19.3	9.0	20.5	8.9	21.3	8.6	20.8
					$m = 5$				
(50,200)	5	9.3	20.4	8.9	19.6	9.7	20.8	10.6	19.3
	10	9.6	20.0	9.2	19.7	9.1	18.9	11.2	18.5
	20	8.7	18.7	9.2	18.5	8.8	19.3	9.3	19.0
(100,500)	5	9.0	21.5	8.8	20.9	9.1	21.2	9.6	20.3
	10	8.9	19.3	9.4	19.3	8.7	20.6	8.5	21.0
	20	8.7	19.0	8.6	18.7	9.4	18.5	9.1	19.3
(100,1000)	5	9.2	19.8	10.0	18.9	9.3	19.6	8.6	21.3
	10	8.8	19.5	8.9	19.0	8.5	21.0	8.5	18.6
	20	8.7	18.7	8.3	18.3	8.7	19.0	8.2	19.1

Table S10

Empirical Type I error rate (%) under equal-correlation with correlation parameter being 0.5 under low-dimensional settings. The nominal level is set to be 5%.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
(n, d)	Method	Dirac	$\mathrm{U}[0, 2]$	Dirac	$\mathrm{U}[0, 2]$	Dirac	$\mathrm{U}[0, 2]$	Dirac	$\mathrm{U}[0, 2]$
					$m = 3$				
(200,5)	QIF	5.2	4.9	5.3	4.8	5.2	5.0	4.9	5.2
	QDIF	5.1	5.3	4.8	4.7	4.9	5.1	5.0	5.0
(200,8)	QIF	5.3	5.1	4.8	5.0	4.9	5.2	4.7	5.3
	QDIF	5.3	4.9	4.9	5.4	5.1	5.1	5.2	5.0
(200,10)	QIF	4.8	4.9	5.3	4.9	5.0	5.1	5.2	5.2
	QDIF	5.1	4.8	4.7	4.6	5.3	5.2	5.4	4.9
					$m = 5$				
(200,5)	QIF	5.0	5.1	4.9	5.2	4.8	4.9	5.3	5.1
	QDIF	5.1	5.3	4.8	4.9	5.0	5.3	5.2	4.9
(200,8)	QIF	4.7	5.1	5.2	4.9	4.8	5.2	5.2	4.7
	QDIF	5.0	5.2	4.8	4.9	5.3	4.8	5.3	5.1
(200,10)	QIF	4.9	4.9	5.3	4.8	5.2	4.7	5.0	5.3
	QDIF	4.7	4.9	5.2	5.1	4.8	4.8	4.9	5.0

Table S11

Empirical Type I error rate (%) under AR-correlation with correlation parameter being 0.5 under low-dimensional settings. The nominal level is set to be 5%.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
(n, d)	Method	Dirac	U[0, 2]	Dirac	U[0, 2]	Dirac	U[0, 2]	Dirac	U[0, 2]
					$m = 3$				
(200,5)	QIF	4.9	5.2	5.0	5.0	4.8	5.3	5.1	5.1
	QDIF	5.0	5.1	5.3	5.3	5.1	4.8	4.9	4.7
(200,8)	QIF	4.8	5.3	5.1	4.7	4.8	5.4	5.0	4.9
	DQIF	4.7	4.6	5.0	5.1	4.9	4.8	5.1	5.3
(200,10)	QIF	5.3	5.1	5.4	4.8	4.8	5.2	4.9	5.0
	DQIF	5.0	4.9	5.3	4.9	5.1	5.3	5.0	5.3
					$m = 5$				
(200,5)	QIF	4.7	4.9	4.8	5.3	5.3	4.9	5.1	4.9
	DQIF	5.1	4.8	4.7	5.2	5.1	4.9	5.0	4.8
(200,8)	QIF	4.8	5.2	5.1	4.7	4.9	5.3	5.0	5.1
	DQIF	5.1	5.2	4.7	5.1	5.0	4.9	4.8	5.2
(200,10)	QIF	5.3	5.0	5.2	4.9	4.7	5.1	4.7	5.2
	DQIF	4.9	4.8	5.1	4.9	5.2	5.3	4.8	5.1

Table S12

Empirical Type I error rate (%) under equal-correlation with correlation parameter being 0.5 under modestly high-dimensional settings, where we set the sparsity level $s = 10$. The nominal level is set to be 5%.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
(n, d)	Method	Dirac	U[0, 2]	Dirac	U[0, 2]	Dirac	U[0, 2]	Dirac	U[0, 2]
					$m = 3$				
(200,100)	QIF	15.4	25.5	18.7	29.4	17.1	30.5	19.3	27.8
	DQIF	5.0	5.4	4.9	5.1	5.2	4.8	4.9	5.2
(200,150)	QIF	18.2	28.0	20.5	29.7	19.3	27.3	18.4	28.6
	DQIF	4.7	4.8	5.3	5.1	5.0	4.8	4.9	5.3
(200,200)	QIF	20.3	31.4	22.9	33.1	21.7	34.8	23.9	33.5
	DQIF	5.3	5.5	4.7	4.5	5.3	5.0	4.9	5.4
					$m = 5$				
(200,100)	QIF	13.8	20.4	15.1	21.7	16.3	22.3	18.0	21.8
	DQIF	4.8	4.9	5.2	4.7	5.3	5.1	4.7	5.2
(200,150)	QIF	14.2	21.8	14.9	23.0	16.0	24.5	15.6	23.4
	QDIF	5.4	5.1	5.2	4.7	5.2	4.6	4.8	5.3
(200,200)	QIF	15.0	23.1	16.1	23.8	16.5	24.1	17.2	24.9
	QDIF	5.3	5.4	4.7	5.2	4.6	4.9	5.2	5.5

Table S13

Empirical Type I error rate (%) under AR-correlation with correlation parameter being 0.5 under modestly high-dimensional settings, where we set the sparsity level $s = 10$. The nominal level is set to be 5%.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
(n, d)	Method	Dirac	U$[0, 2]$	Dirac	U$[0, 2]$	Dirac	U$[0, 2]$	Dirac	U$[0, 2]$
				$m = 3$					
(200,100)	QIF	14.5	23.4	15.1	24.9	14.8	26.4	16.2	25.9
	QDIF	5.3	5.1	4.7	5.2	4.8	5.3	4.7	5.4
(200,150)	QIF	15.0	24.2	16.3	25.8	15.8	26.2	16.7	26.5
	QDIF	4.8	4.9	5.3	5.0	4.7	4.6	4.9	5.4
(200,200)	QIF	16.8	24.8	17.3	25.7	18.4	26.5	18.0	25.9
	QDIF	5.2	5.1	4.8	4.6	5.4	5.2	5.1	4.7
				$m = 5$					
(200,100)	QIF	12.7	20.2	13.9	19.7	13.5	20.9	15.3	21.5
	QDIF	5.0	5.3	5.1	4.7	4.8	5.3	4.8	5.3
(200,150)	QIF	13.3	23.9	13.8	25.1	14.1	24.8	15.0	25.6
	QDIF	4.8	5.3	4.9	4.8	4.9	5.4	5.2	5.3
(200,200)	QIF	13.5	24.1	14.2	24.4	14.5	25.1	15.3	26.3
	QDIF	4.9	5.3	5.2	4.6	4.8	5.0	5.1	4.7

Table S14

Empirical Type I error rate (%) under equal-correlation with correlation parameter being 0.5 using the decorrelated GEE method. The nominal level is set to be 5%.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
(n, d)	s	Dirac	U$[0, 2]$	Dirac	U$[0, 2]$	Dirac	U$[0, 2]$	Dirac	U$[0, 2]$
				$m = 3$					
(50,200)	5	5.4	5.7	5.4	5.5	5.6	5.0	4.8	5.3
	10	5.5	5.8	5.4	5.6	4.9	5.3	5.6	5.3
	20	5.8	6.0	5.7	5.3	5.9	6.1	5.0	5.9
(100,500)	5	5.5	5.8	4.9	5.5	5.9	5.6	5.5	5.3
	10	5.5	5.1	5.0	5.7	5.8	5.1	4.2	5.7
	20	5.6	5.9	5.6	5.4	4.3	5.7	5.5	5.8
(100,1000)	5	5.5	5.7	5.2	4.5	5.5	5.9	6.1	5.6
	10	5.2	5.7	5.1	5.6	5.5	4.3	5.2	5.6
	20	5.8	6.0	5.7	5.2	5.1	5.7	5.1	4.5
				$m = 5$					
(50,200)	5	4.5	5.5	5.4	5.1	5.3	5.5	5.2	5.0
	10	5.1	5.5	5.2	5.3	5.4	5.2	5.0	4.7
	20	5.5	5.2	5.4	5.0	4.5	5.2	4.9	5.4
(100,500)	5	4.8	5.0	5.2	5.5	5.3	5.5	5.3	5.4
	10	5.3	5.1	5.3	5.6	5.7	5.2	5.3	5.5
	20	5.4	5.5	5.1	4.6	5.5	6.0	5.6	6.2
(100,1000)	5	5.3	5.1	5.6	4.8	4.7	5.9	5.8	5.3
	10	5.5	5.8	5.7	5.2	5.9	4.5	5.1	4.6
	20	6.1	5.2	5.7	5.3	6.2	5.3	5.8	5.4

Table S15

Empirical Type I error rate (%) under AR-correlation structure with correlation parameter being 0.6 using the decorrelated GEE. The nominal level is set to be 5%.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
(n, d)	s	Dirac	U$[0, 2]$	Dirac	U$[0, 2]$	Dirac	U$[0, 2]$	Dirac	U$[0, 2]$
					$m = 3$				
(50,200)	5	6.9	7.3	8.1	7.5	7.4	7.7	8.1	7.4
	10	8.5	9.2	8.7	8.3	8.8	8.5	8.9	7.8
	20	8.9	8.7	8.5	9.3	9.0	8.8	8.7	9.2
(100,500)	5	7.4	8.1	7.8	7.9	8.3	8.0	8.2	8.9
	10	9.1	9.0	9.5	8.7	8.9	9.5	9.3	9.7
	20	9.4	9.1	9.6	9.3	9.0	9.7	10.1	9.5
(100,1000)	5	8.2	7.7	8.0	8.4	8.6	8.7	8.2	8.8
	10	9.4	9.7	8.9	9.1	9.2	9.5	9.9	9.3
	20	9.9	9.8	10.3	9.9	10.2	10.1	9.7	10.4
					$m = 5$				
(50,200)	5	7.0	7.4	7.1	6.9	7.3	7.2	7.4	7.0
	10	8.9	9.2	8.7	9.3	8.8	9.0	9.5	9.2
	20	9.5	10.2	9.8	9.5	9.7	10.0	9.4	9.3
(100,500)	5	7.5	7.9	7.8	8.2	7.1	7.6	8.5	8.0
	10	9.1	9.3	9.4	9.0	9.7	10.2	10.5	9.8
	20	10.5	10.7	9.9	10.4	11.2	10.3	10.9	10.5
(100,1000)	5	8.3	7.9	8.5	8.3	8.0	7.9	8.2	8.0
	10	10.6	10.1	10.5	10.3	10.4	10.3	9.9	10.2
	20	10.9	11.2	10.4	11.4	10.7	11.6	10.8	10.7

Table S16

Empirical Type I error rate (%) under equal-correlation with correlation parameter being 0.5 and the noises are generated from T distribution with 5 degrees of freedom. The nominal level is set to be 5%.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
(n, d)	s	Dirac	U$[0,2]$	Dirac	U$[0,2]$	Dirac	U$[0,2]$	Dirac	U$[0,2]$
					$m = 3$				
(50,200)	5	5.7	5.4	5.9	5.2	5.0	5.5	4.9	5.8
	10	5.8	5.3	5.9	5.5	5.0	5.9	6.1	5.4
	20	5.5	6.3	6.1	6.2	5.7	5.3	6.5	5.8
(100,500)	5	6.0	6.2	5.3	5.9	5.4	6.1	4.6	5.9
	10	5.3	6.3	5.7	5.4	6.0	6.2	5.9	6.3
	20	6.2	6.4	5.4	6.0	5.8	5.5	6.1	6.0
(100,1000)	5	5.2	5.6	6.1	5.0	5.7	6.3	6.0	5.8
	10	5.8	5.5	5.6	5.4	6.2	6.1	4.9	6.2
	20	6.0	6.1	6.3	5.5	6.7	6.1	5.8	6.5
					$m = 5$				
(50,200)	5	5.2	5.9	5.4	5.0	5.2	4.8	5.5	5.7
	10	5.7	5.5	5.0	5.5	4.6	6.0	5.9	5.3
	20	5.8	5.5	5.4	5.8	6.2	5.9	5.7	6.2
(100,500)	5	5.3	5.2	5.7	5.4	5.7	5.7	5.3	4.7
	10	5.7	5.3	5.1	5.4	5.6	6.2	5.8	6.0
	20	5.6	5.4	6.0	5.8	6.1	6.4	5.9	6.3
(100,1000)	5	5.5	5.9	5.8	5.3	5.4	6.0	5.2	5.3
	10	5.9	5.2	5.5	6.1	5.3	5.7	6.2	5.8
	20	5.9	5.6	5.8	5.5	6.0	6.1	5.9	6.3

Table S17

Empirical Type I error rate (%) under AR-correlation structure with correlation parameter being 0.6 and the noises are generated from T distribution with 5 degrees of freedom. The nominal level is set to be 5%.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
(n, d)	s	Dirac	U[0, 2]	Dirac	U[0, 2]	Dirac	U[0, 2]	Dirac	U[0, 2]
					$m = 3$				
(50,200)	5	4.8	5.4	5.3	4.9	5.7	4.9	5.2	5.5
	10	5.3	5.2	4.9	5.5	5.7	5.4	5.6	4.5
	20	5.5	5.5	6.0	5.4	5.6	5.3	5.2	5.9
(100,500)	5	5.2	5.8	5.6	5.9	5.4	5.7	5.4	5.0
	10	5.7	5.3	5.4	5.9	6.2	6.1	5.8	6.3
	20	5.9	6.1	5.9	5.5	6.3	6.5	6.2	6.5
(100,1000)	5	5.3	5.6	5.4	5.1	5.3	5.5	6.1	5.8
	10	5.8	5.7	5.3	6.2	6.0	5.9	5.5	6.4
	20	5.4	6.2	5.9	6.1	5.8	5.8	6.3	6.6
					$m = 5$				
(50,200)	5	5.0	5.3	5.2	4.8	5.5	5.7	5.2	5.4
	10	5.8	5.4	5.7	6.1	5.9	5.6	5.8	4.9
	20	5.5	6.2	5.2	5.7	5.8	5.5	6.0	5.9
(100,500)	5	5.3	5.8	5.6	4.9	5.0	5.5	5.3	5.4
	10	5.2	5.5	5.4	5.8	5.3	6.0	6.1	5.9
	20	6.3	5.6	5.8	6.2	5.9	6.5	6.4	6.9
(100,1000)	5	5.8	5.9	5.4	5.6	6.0	5.3	5.4	5.9
	10	5.4	5.3	5.9	4.8	5.5	5.0	5.8	6.0
	20	5.8	5.5	6.0	5.4	6.0	5.8	6.2	6.5

Table S18

Empirical false discovery rate (%) at level $\alpha = 0.1$ and 0.2 under equal-correlation structure with correlation parameter being 0.5 and the noises are generated from T distribution with 5 degrees of freedom.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
	α	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2
(n, d)	s				$m = 3$				
(50,200)	5	9.8	19.3	8.9	20.4	9.5	19.5	10.3	19.3
	10	9.4	19.2	9.3	20.6	9.7	20.7	9.1	19.3
	20	9.0	18.7	9.2	19.0	8.8	20.4	8.6	21.0
(100,500)	5	9.5	19.4	9.2	21.1	10.7	19.2	9.3	19.7
	10	9.8	20.9	9.1	20.2	9.0	19.3	10.5	20.6
	20	8.7	19.0	10.9	21.2	8.9	20.3	9.3	18.6
(100,1000)	5	10.3	20.4	10.6	19.5	9.3	18.9	9.4	20.3
	10	9.4	19.3	10.6	19.1	9.9	21.4	9.5	19.5
	20	9.0	20.8	8.7	21.3	11.0	20.9	9.1	18.5
					$m = 5$				
(50,200)	5	9.6	21.3	9.1	19.9	9.0	20.5	10.4	20.5
	10	9.3	21.1	10.6	20.3	9.0	19.3	10.3	19.8
	20	9.7	19.5	11.1	20.5	10.7	18.7	10.8	20.9
(100,500)	5	9.3	20.9	9.4	21.4	10.0	20.5	9.7	19.3
	10	8.9	20.5	9.5	19.4	9.8	20.3	9.8	19.7
	20	9.4	20.2	9.3	19.4	9.0	21.3	8.7	21.0
(100,1000)	5	9.5	19.3	10.4	19.7	9.8	20.3	9.1	20.8
	10	9.1	20.0	9.3	20.5	8.9	19.4	8.9	21.2
	20	9.4	18.9	9.7	19.5	9.2	21.3	9.1	20.8

Table S19

Empirical false discovery rate (%) at level $\alpha = 0.1$ and 0.2 under AR-correlation structure with correlation parameter being 0.6 and the noises are generated from T distribution with 5 degrees of freedom.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
	α	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2
(n, d)	s				$m = 3$				
(50,200)	5	10.3	20.5	9.9	19.5	10.8	20.9	10.7	20.4
	10	9.5	19.4	9.2	19.6	9.4	19.2	10.7	20.1
	20	9.2	19.6	9.6	20.7	10.6	20.2	9.1	21.3
(100,500)	5	9.7	19.2	10.5	20.4	10.3	19.3	10.6	20.6
	10	9.4	18.9	10.7	20.8	9.6	20.4	10.7	19.7
	20	10.3	20.9	9.6	20.4	10.3	21.1	9.8	19.0
(100,1000)	5	9.8	19.5	9.7	20.4	10.6	19.8	10.5	19.4
	10	10.7	20.5	9.7	19.3	10.7	20.8	9.3	20.4
	20	10.1	19.4	9.4	21.0	9.1	19.5	10.9	21.3
					$m = 5$				
(50,200)	5	10.5	20.8	10.6	19.4	10.2	19.6	9.3	20.0
	10	10.2	20.5	9.4	19.7	10.8	20.4	9.2	20.8
	20	10.4	20.8	9.2	20.5	9.0	19.2	9.1	21.1
(100,500)	5	9.1	20.4	9.6	19.5	10.7	19.3	10.3	20.7
	10	10.3	20.1	9.6	20.4	9.5	19.7	9.0	20.4
	20	9.6	19.6	9.0	20.7	10.9	20.9	11.0	20.6
(100,1000)	5	10.3	20.5	10.2	20.2	9.4	19.6	9.6	19.3
	10	9.3	20.8	9.5	20.3	10.4	19.5	9.2	21.0
	20	9.9	20.2	10.4	20.9	8.8	18.7	10.9	21.3

References.

Cho, H. and Qu, A. (2015). Efficient estimation for longitudinal data by combining large-dimensional moment conditions. *Electronic Journal of Statistics*, **9** 1315–1334.

Jaquish, C. E. (2007). The framingham heart study, on its way to becoming the gold standard for cardiovascular genetic epidemiology? *BMC Med. Genet.*, **8** 63.

Lian, H., Liang, H. and Wang, L. (2014). Generalized additive partial linear models for clustered data with diverging number of covariates using gee. *Stat. Sinica*, **24** 173–196.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73** 13–22.

Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *Ann. Statist.*, **41** 2948–2978.

Ma, S. (2012). Two-step spline estimating equations for generalized additive partially linear models with large cluster sizes. *Ann. Statist.*, **40** 2943–2972.

Tropp, J. A. et al. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, **8** 1–230.

Tsiatis, A. (2007). *Semiparametric Theory and Missing Data*. Springer Science & Business Media.

Wang, L. (2011). GEE analysis of clustered binary data with diverging number of covariates. *Ann. Statist.*, **39** 389–417.

Wang, L., Zhou, J. and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, **68** 353–360.

Xie, M. and Yang, Y. (2003). Asymptotics for generalized estimating equations with large cluster sizes. *Ann. Statist.*, **31** 310–347.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.*, **100** 577–590.

Ethan X. Fang and Runze Li
Department of Statistics
the Pennsylvania State University
University Park, PA 16802-2111
USA
E-mail: xxf13@psu.edu
E-mail: rzli@psu.edu

Yang Ning
Department of Statistics
Cornell University
Ithaca, NY 14850
USA
E-mail: yn265@cornell.edu