Supplement to "Joint mean and covariance estimation with unreplicated matrix-variate data"

Michael Hornstein, Roger Fan, Kerby Shedden, Shuheng Zhou
Department of Statistics, University of Michigan

Outline

We provide additional simulation and data analysis results in Section A and B. We state some preliminary results and notation in Section C. We prove Theorem 1 in Section D and Corollary 2 in Section D.2. We prove Theorem 3 in Section E, with additional lemmas proved in Section F. We prove entrywise convergence of the sample correlation matrices for Algorithm 1 in Section G. We prove Theorem 4 in Section H, and we prove additional lemmas used in the proof of Theorem 4 in Section I. In Section J we provide additional comparisons between our method and some related methods on both simulated and real data.

A Additional simulation results

Figure S1 demonstrates the effect of mean structure on covariance estimation. As expected, when there is no mean structure Gemini performs competitively. As more mean structure is added, however, its performance quickly decays to be worse than Algorithm 2. This also provides evidence that the plug-in estimator $\widehat{\tau}_{\text{init}}$ used in Algorithm 2 is appropriately selecting genes to group center, as when there are no or very few differentially expressed genes Algorithm 2 is still never worse than Gemini. Algorithm 1 does not perform as well as Algorithm 2 but still tends to eventually outperform Gemini as more mean structure is added. As the sample size increases, the difference between Algorithm 2 and Algorithm 1 decreases as the added noise from group centering becomes less of a factor. We still recommend using Algorithm 2 in most realistic scenarios, but this reinforces our theoretical finding that the two algorithms have the same error rates.

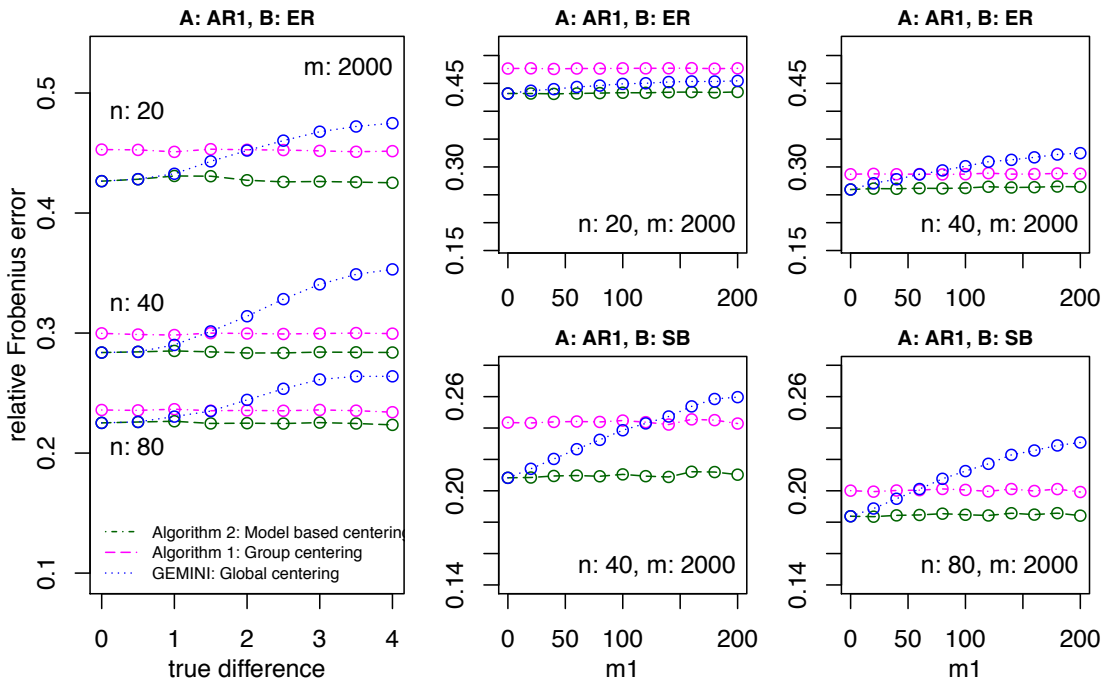
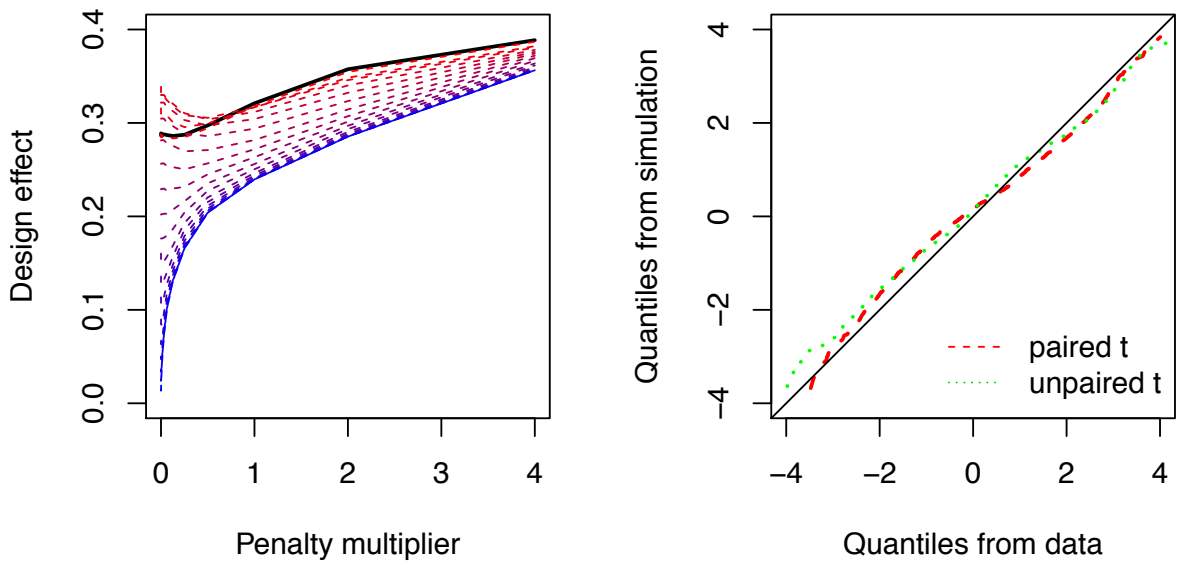


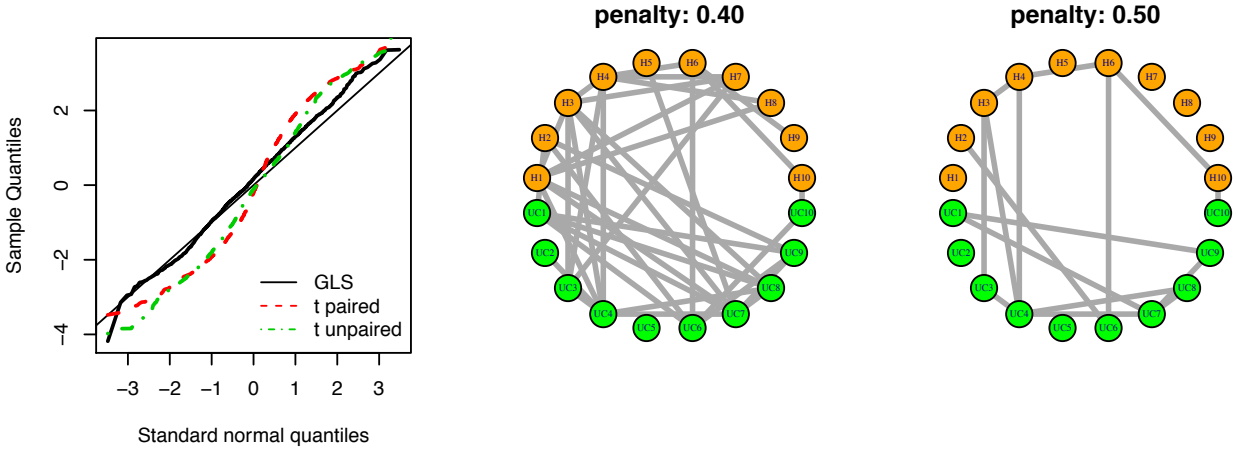
Figure S1: Performance of Gemini, Algorithm 1, and Algorithm 2 for estimating B under different mean and covariance structures. As the sample size increases, we can see that Algorithm 1 improves relative to Gemini and begins to catch up to Algorithm 2. Gemini's performance always degrades as the true differences grow or more differentially expressed genes are added, while Algorithm 1 and 2 are stable. We set B^{-1} as Erdős-Rényi (ER) or star-block with blocks of size 4 (SB). All plots use A from an AR1(0.8) model with $m = 2000$ and are averaged over 200 replications. In the left plot the first 50 genes are differentially expressed at the level specified on the x-axis. As indicated, the three groups of lines correspond to $n = 20$, 40, and 80. In the right two columns there are `m1` number of genes with exponentially decaying true differences between groups, scaled so that the largest difference is 5 (resulting in an average difference of approximately 1).

B Additional data analysis

As discussed in Section 3.1, it is particularly important that the design effect is accurately estimated in order for the test statistics to be properly calibrated. The first plot of Figure S2a displays the sensitivity of the estimated design effect (21) for Algorithm 2 to the GLasso penalty parameter and the number of group centered columns. In the case that all columns are group centered, Algorithm 2 reduces to Algorithm 1. If we group center all genes, the estimated design effect is sensitive to the penalty parameter, but if we group center a small proportion of genes, it is less sensitive to the penalty parameter. This is further evidence that it may be advantageous to avoid over-centering the data when the true mean difference



(a) The first plot displays the estimated design effect vs. the penalty multiplier for Algorithm 2. Each curve corresponds to a different number of columns being group centered. As the curves go from top to bottom, the number of group centered columns increases from 10 to 2000. The second plot shows a quantile plot of test statistics from the data vs. simulated test statistics; in the simulation, the population person-person covariance matrix is $\widehat{B}$, as estimated from the UC data.



(b) Quantile plot and inverse covariance graphs. The first two plots correspond to $\lambda = 0.4$ and 128 group centered genes. The third plot corresponds to $\lambda = 0.5$ and 128 group centered genes. Green circles correspond to twins with UC, orange circles to twins without UC. Twins are aligned vertically.

Figure S2

vector γ may be sparse. The second plot of Figure S2a shows a quantile plot comparing the distribution of test statistics from the UC data to test statistics from a simulation whose population correlation structure is matched to the UC data. The quantile plot reveals that we can reproduce the pattern of overdispersion in the test statistics using simulated data having person-person as well as gene-gene correlations. Such correlations therefore provide a possible explanation for the overdispersion of the test statistics.

Figure S2b displays a quantile plot and inverse covariance graph for $\lambda = 0.4$ and 128 group centered genes. Under these settings the test statistics appear correctly calibrated, coinciding with the central portion of the reference line. Furthermore, the inverse covariance graph is sparse (38 edges). In the inverse covariance graph, there are more edges between subjects with UC than between the healthy subjects, which could be explained by the existence of subtypes of UC inducing correlations between subsets of subjects. The third plot of Figure S2b displays a sparser inverse covariance graph, corresponding to a larger penalty $\lambda = 0.5$. There are three edges between twin pairs, and there are more edges between subjects with UC than between those without UC.

B.1 Stability simulation

Table S1 shows the results from a simulation analogous to Table 2, demonstrating stability across iterations of the procedure. Iteration 1 begins by group centering 1280 genes and this number is halved in each successive iteration. We can see from the table that the gene rankings generated by Algorithm 2 are robust to misspecifying the number of differentially expressed genes. When the number of group centered genes is 160 or below (iterations 4 through 8), the commonly selected genes among the top 20 genes are stable. Furthermore, the true positives remain stable as we decrease the amount of genes centered, while the false positives decrease.

Table S1: Number of genes in common among genes ranked in the top 20 when different numbers of genes are group centered. This simulation is analogous to Table 2. Note that the maximum possible value for any entry of the table is 20; if entry (i, j) is 20, then iterations i and j selected the same top twenty genes. The first 10 genes have a difference of 1.5 and the second 10 have a difference of 1. All remaining genes have a true mean difference of zero. We use B as estimated from the UC data, and A is from an AR1(0.8) model. These simulations have $n = 20$ individuals and 2000 genes and are averaged over 200 replications. The last two rows display the average number of true and false positives among the genes ranked in the top 20 of each iteration.

	1	2	3	4	5	6	7	8
1	20.0	17.6	15.8	14.8	14.3	14.0	14.0	13.9
2	17.6	20.0	17.9	16.8	16.2	15.9	15.8	15.8
3	15.8	17.9	20.0	18.7	18.1	17.8	17.7	17.6
4	14.8	16.8	18.7	20.0	19.3	19.0	18.9	18.8
5	14.3	16.2	18.1	19.3	20.0	19.6	19.5	19.4
6	14.0	15.9	17.8	19.0	19.6	20.0	19.8	19.7
7	14.0	15.8	17.7	18.9	19.5	19.8	20.0	19.8
8	13.9	15.8	17.6	18.8	19.4	19.7	19.8	20.0
TP	12.7	14.3	15.6	16.4	16.7	16.8	16.8	16.8
FP	7.3	5.7	4.4	3.6	3.3	3.2	3.2	3.2

C Preliminary results

In this section, we refresh notation and introduce propositions that are shared in the proofs of the theorems. For convenience, we first restate some notation.

$$D = \begin{bmatrix} 1_{n_1} & 0 \\ 0 & 1_{n_2} \end{bmatrix} \in \mathbb{R}^{n \times 2} \tag{S1}$$

$$\Omega = (D^T B^{-1} D)^{-1} \text{ and } \Omega_{n,m} = (D^T B_{n,m}^{-1} D)^{-1} \tag{S2}$$

$$\Delta = B_{n,m}^{-1} - B^{-1} \tag{S3}$$

$$\widehat{\beta}(\widehat{B}^{-1}) = (D^T \widehat{B}^{-1} D)^{-1} D^T \widehat{B}^{-1} X \in \mathbb{R}^{2 \times m} \tag{S4}$$

When D has the form (S1), the singular values are $\sigma_{\max}(D) = \sqrt{n_{\max}}$ and $\sigma_{\min}(D) = \sqrt{n_{\min}}$. The condition number is $\kappa(D) = \sigma_{\max}(D)/\sigma_{\min}(D) = \sqrt{n_{\text{ratio}}}$ where $n_{\text{ratio}} = \max(n_1, n_2)/\min(n_1, n_2)$.

We first state some convenient notation and bounds.

$$
\begin{aligned}
&r_a := a_{\max}/a_{\min} \text{ and } r_b := b_{\max}/b_{\min};\\
&1/\varphi_{\min}(A) = \|A^{-1}\|_2 \leqslant \|\rho(A)^{-1}\|_2/a_{\min} = \frac{1}{a_{\min}\varphi_{\min}(\rho(A))}, \qquad \text{(S5)}\\
&1/\varphi_{\min}(B) = \|B^{-1}\|_2 \leqslant \|\rho(B)^{-1}\|_2/b_{\min} = \frac{1}{b_{\min}\varphi_{\min}(\rho(B))}, \qquad \text{(S6)}\\
&1/\varphi_{\min}(\rho(A)) = \|\rho(A)^{-1}\|_2 \leqslant a_{\max}\|A^{-1}\|_2, \qquad \text{(S7)}\\
&1/\varphi_{\min}(\rho(B)) = \|\rho(B)^{-1}\|_2 \leqslant b_{\max}\|B^{-1}\|_2 \qquad \text{(S8)}\\
&\|A\|_2 \leqslant a_{\max}\|\rho(A)\|_2, \quad \|B\|_2 \leqslant b_{\max}\|\rho(B)\|_2, \qquad \text{(S9)}\\
&\|\rho(A)\|_2 \leqslant \|A\|_2/a_{\min}, \quad \text{and} \quad \|\rho(B)\|_2 \leqslant \|B\|_2/b_{\min}. \qquad \text{(S10)}
\end{aligned}
$$

The eigenvalues of the correlation matrices satisfy

$$
0 < \varphi_{\min}(\rho(A)) \leqslant 1 \leqslant \varphi_{\max}(\rho(A)) \text{ and } 0 < \varphi_{\min}(\rho(B)) \leqslant 1 \leqslant \varphi_{\max}(\rho(B)). \qquad \text{(S11)}
$$

In the remainder of this section, we state preliminary results and highlight important intermediate steps that are used in the proofs of Theorems 1 and 3. First we state propositions used in mean estimation for Theorems 1 and 3.

C.1 Propositions

We now state propositions used in the proofs of Lemmas S5 and S6. We defer the proof of Proposition S1 to Section D.5.

Proposition S1. *For Ω as defined in (S2) and some design matrix D,*

$$
\|\Omega\|_2 \leqslant \|B\|_2/\sigma^2_{\min}(D)
$$

In the case that D is defined as in (S1), we have $\|\Omega\|_2 \leqslant \|B\|_2/n_{\min}$.

Furthermore,

$$
\lambda_{\min}(\Omega) \geqslant \frac{\lambda_{\min}(B)}{n_{\max}}. \qquad \text{(S12)}
$$

We state the following perturbation bound.

Theorem S2 (Golub & Van Loan, Theorem 2.3.4)**.** *If A is invertible and $\|A^{-1}E\|_p < 1$, then $A + E$ is invertible and*

$$\|(A+E)^{-1} - A^{-1}\|_p \leqslant \frac{\|E\|_p \|A^{-1}\|_p^2}{1 - \|A^{-1}E\|_p} \leqslant \frac{\|E\|_p \|A^{-1}\|_p^2}{1 - \|A^{-1}\|_p \|E\|_p}.$$

In Proposition S3, we provide auxiliary upper bounds that depend on $\|\Delta\|_2$, $\|B\|_2$, $\kappa(D)$, and $\sigma_{\min}(D)$. We defer the proof of Proposition S3 to the end of this section, for clarity of presentation.

Proposition S3. *Let $\Delta = B_{n,m}^{-1} - B^{-1}$.*

$$\delta_0(\Delta) := \|\Omega_{n,m} - \Omega\|_2 \leqslant \frac{1}{\sigma_{\min}^2(D)} \frac{\|B\|_2^2 \|\Delta\|_2}{1/\kappa^2(D) - \|B\|_2 \|\Delta\|_2} \tag{S13}$$

$$\delta_1(\Delta) := \left\|\Omega D^T \Delta\right\|_2 \leqslant \sigma_{\max}(D) \|B\|_2 \|\Delta\|_2 / \sigma_{\min}^2(D) = \frac{\sqrt{n_{\max}}}{n_{\min}} \|B\|_2 \|\Delta\|_2. \tag{S14}$$

If $\|(D^T B^{-1} D)^{-1} D^T \Delta D\|_2 < 1$, then

$$\delta_2(\Delta) := \left\|(\Omega_{n,m} - \Omega) D^T \Delta\right\|_2 \leqslant \frac{\kappa(D)}{\sigma_{\min}(D)} \frac{\|B\|_2^2 \|\Delta\|_2^2}{1/\kappa^2(D) - \|B\|_2 \|\Delta\|_2} \tag{S15}$$

$$\delta_3(\Delta) := \left\|(\Omega_{n,m} - \Omega) D^T B^{-1}\right\|_2 \leqslant \frac{\kappa(D)}{\sigma_{\min}(D)} \frac{\|B\|_2^2 \|B^{-1}\|_2 \|\Delta\|_2}{1/\kappa^2(D) - \|B\|_2 \|_2 \|\Delta\|_2} \tag{S16}$$

The following proposition is a corollary of Proposition S3.

Proposition S4. *When D has the form (S1), and Ω is as defined in* (S2)*,*

$$\delta_0(\Delta) = \|\Omega_{n,m} - \Omega\|_2 \leqslant \frac{1}{n_{\min}} \frac{\|B\|_2^2 \|\Delta\|_2}{1/n_{ratio} - \|B\|_2 \|\Delta\|_2}$$

$$\delta_1(\Delta) = \left\|\Omega D^T \Delta\right\|_2 \leqslant \frac{\sqrt{n_{ratio}}}{\sqrt{n_{\min}}} \|B\|_2 \|\Delta\|_2$$

$$\delta_2(\Delta) = \left\|(\Omega_{n,m} - \Omega) D^T \Delta\right\|_2 \leqslant \frac{\sqrt{n_{ratio}}}{\sqrt{n_{\min}}} \frac{\|B\|_2^2 \|\Delta\|_2^2}{1/n_{ratio} - \|B\|_2 \|\Delta\|_2}$$

Let K be defined as in Theorem 1. We express the entrywise rates of convergence of the sample correlation matrices $\widehat{\Gamma}(B)$ and $\widehat{\Gamma}(A)$, respectively, in terms of the following quantities:

$$\widetilde{\alpha} = C_A K \frac{\log^{1/2}(m)}{\sqrt{m}} \left(1 + \frac{\|B\|_1}{n}\right) + \frac{\|B\|_1}{n_{\min}} \quad \text{and} \quad \widetilde{\eta} = C_B K \frac{\log^{1/2}(m \vee n)}{\sqrt{n}} + \frac{\|B\|_1}{n}. \tag{S17}$$

D Proof of Theorem 1 and Corollary 2

D.1 Proof of Theorem 1

Let $B_{n,m} \in \mathbb{R}^{n\times n}$ denote a fixed positive definite matrix. Let D be as defined as in (4). Define $\Delta_{n,m} = B_{n,m}^{-1} - B^{-1}$ and

$$\Omega = (D^T B^{-1} D)^{-1} \text{ and } \Omega_{n,m} = (D^T B_{n,m}^{-1} D)^{-1}. \tag{S18}$$

Note that we can decompose the error for all j as

$$\|\widehat{\beta}_j(B_{n,m}^{-1}) - \beta_j^*\|_2 \leqslant \|\widehat{\beta}_j(B^{-1}) - \beta_j^*\|_2 + \|\widehat{\beta}_j(B_{n,m}^{-1}) - \widehat{\beta}_j(B^{-1})\|_2 =: \mathrm{I} + \mathrm{II}. \tag{S19}$$

We will use the following lemmas, which are proved in subsections D.4 and D.3, to bound these two terms on the right-hand side, respectively.

Lemma S5. *Let $\mathcal{E}_2$ denote the event*

$$\mathcal{E}_2 = \left\{\|\widehat{\beta}_j(B^{-1}) - \beta_j^*\|_2 \leqslant s_{n,m}\right\}, \quad \textit{with} \quad s_{n,m} = C_3 d^{1/2}\sqrt{\frac{\log(m)\|B\|_2}{n_{\min}}}. \tag{S20}$$

Then $P(\mathcal{E}_2) \geqslant 1 - 2/m^d$.

Lemma S6. *Let $B_{n,m} \in \mathbb{R}^{n\times n}$ denote a fixed matrix such that $B_{n,m} \succ 0$. Let $X_j \in \mathbb{R}^n$ denote the jth column of X, where X is a realization of model (2). Let $\mathcal{E}_3$ denote the event*

$$\mathcal{E}_3 = \left\{\|\widehat{\beta}_j(B_{n,m}^{-1}) - \widehat{\beta}_j(B^{-1})\|_2 \leqslant t_{n,m}\right\}, \quad \textit{with} \quad t_{n,m} = \widetilde{C} n_{\min}^{-1/2}\|\Delta_{n,m}\|_2. \tag{S21}$$

for some absolute constant $\widetilde{C}$. Then $P(\mathcal{E}_3) \geqslant 1 - 2/m^d$.

The proof of (18) follows from the union bound $P(\mathcal{E}_2 \cap \mathcal{E}_3) \geqslant 1 - P(\mathcal{E}_2) - P(\mathcal{E}_3) \geqslant 1 - 4/m^d$. Next we prove (20). Let $r_{n,m} = s_{n,m} + t_{n,m}$, as defined in (18). Let $\delta = (1, -1) \in \mathbb{R}^2$. Then

$$|\widehat{\gamma}_j(B_{n,m}^{-1}) - \gamma_j| = \left|\delta^T\left(\widehat{\beta}_j(B_{n,m}^{-1}) - \beta_j^*\right)\right| \leqslant \|\delta\|_2\|\widehat{\beta}_j(B_{n,m}^{-1}) - \beta_j^*\|_2 = \sqrt{2}\|\widehat{\beta}_j(B_{n,m}^{-1}) - \beta_j^*\|_2,$$

where we used the Cauchy-Schwarz inequality. Hence if $\|\widehat{\beta}_j(B_{n,m}^{-1}) - \beta_j\|_2 \leqslant r_{n,m}$, it follows

that $|\widehat{\gamma}_j(B_{n,m}^{-1}) - \gamma_j| \leqslant \sqrt{2} r_{n,m}$. The result holds by applying a union bound over the variables $j = 1, \ldots, m$. $\square$

This completes the proof of Theorem 1.

D.2 Proof of Corollary 2 and Corollary 5

First note that by Proposition S4,

$$\begin{aligned}
\left|\delta^T (D^T \widehat{B}^{-1} D)^{-1} \delta - \delta^T (D^T B^{-1} D)^{-1} \delta\right| &= \left|\delta^T \left((D^T \widehat{B}^{-1} D)^{-1} - (D^T B^{-1} D)^{-1}\right) \delta\right| \\
&\leqslant \|\delta\|_2^2 \left\|(D^T \widehat{B}^{-1} D)^{-1} - (D^T B^{-1} D)^{-1}\right\|_2 \\
&= 2 \left\|(D^T \widehat{B}^{-1} D)^{-1} - (D^T B^{-1} D)^{-1}\right\|_2 \\
&\leqslant 2 \frac{\|B\|_2^2 \, \|\Delta\|_2}{n_{\min}}. \qquad \text{(S22)}
\end{aligned}$$

Note that by Proposition S1,

$$|\delta^T \Omega \delta| \geqslant \frac{\lambda_{\min}(B)}{n_{\max}}. \qquad \text{(S23)}$$

Corollary 2 follows from (S22) and (S23), which provide an upper bound on the numerator and lower bound on the denominator, respectively.

Corollary 5 holds because by (28) of Theorem 4,

$$\left|\delta^T \left(\widehat{\Omega} - \Omega\right) \delta\right| \leqslant 2 \frac{\|B\|_2^2}{n_{\min}} \left(\frac{C' \lambda_A \sqrt{|B^{-1}|_{0,\text{off}} \vee 1}}{b_{\min} \varphi_{\min}^2(\rho(B))} \right) \leqslant 2C' \frac{\kappa(B)}{n_{\min}} \lambda_A \sqrt{|B^{-1}|_{0,\text{off}} \vee 1} \qquad \text{(S24)}$$

D.3 Proof of Lemma S5

First, we show that

$$\|\Omega^{1/2}\|_F + d^{1/2} K^2 \sqrt{\log(m)} \|\Omega\|_2^{1/2} / \sqrt{c} \leqslant s_{n,m}, \qquad \text{(S25)}$$

with $s_{n,m}$ as defined in (19). Because $\|\Omega^{1/2}\|_F \leqslant \sqrt{2}\|\Omega^{1/2}\|_2$, it follows that

$$\begin{aligned}\|\Omega^{1/2}\|_F + d^{1/2}K^2\sqrt{\log(m)}\|\Omega\|_2^{1/2}/\sqrt{c} &\leqslant \left(\sqrt{2} + d^{1/2}K^2\sqrt{\log(m)}/\sqrt{c}\right)\|\Omega\|_2^{1/2} \\ &\leqslant C_3 d^{1/2}\sqrt{\log(m)}\|\Omega\|_2^{1/2} \leqslant C_3 d^{1/2}\sqrt{\frac{\log(m)\|B\|_2}{n_{\min}}},\end{aligned}$$

where the last step follows from Proposition S1. Next, we express $\widehat{\beta}_j(B^{-1}) - \beta_j^*$ as

$$\widehat{\beta}_j(B^{-1}) - \beta_j^* = \Omega^{1/2}\eta_j, \quad \text{where} \quad \eta_j = \Omega^{-1/2}\left(\widehat{\beta}_j(B^{-1}) - \beta_j^*\right).$$

By the bound (S25), event $\mathcal{E}_2^c$ implies $\{\|\Omega^{1/2}\eta_j\|_2 > \|\Omega^{1/2}\|_F + d^{1/2}K^2\sqrt{\log(m)}\|\Omega\|_2^{1/2}/\sqrt{c}\}$. Therefore,

$$\begin{aligned}P\left(\|\Omega\eta_j\|_2 \geqslant s_{n,m}\right) &\leqslant P\left(\|\Omega\eta_j\|_2 > \|\Omega^{1/2}\|_F + d^{1/2}K^2\sqrt{\log(m)}\|\Omega\|_2^{1/2}/\sqrt{c}\right) \\ &\leqslant P\left(\left|\|\Omega^{1/2}\eta_j\|_2 - \|\Omega^{1/2}\|_F\right| > d^{1/2}K^2\sqrt{\log(m)}\|\Omega\|_2^{1/2}/\sqrt{c}\right) \\ &\leqslant 2\exp\left(\frac{-c\left(d^{1/2}K^2\sqrt{\log(m)}\|\Omega\|_2^{1/2}/\sqrt{c}\right)^2}{K^4\|\Omega^{1/2}\|_2^2}\right) \\ &= 2\exp\left(\frac{-d\log(m)\|\Omega\|_2}{\|\Omega^{1/2}\|_2^2}\right) = 2\exp\left(-d\log(m)\right) = 2/m^d.\end{aligned}$$

□

D.4 Proof of Lemma S6

The proof will proceed in the following steps. First, we show that $\widehat{\beta}_j(B_{n,m}^{-1}) - \widehat{\beta}_j(B^{-1})$ can be expressed as VZ_j, where

$$V = \left(\Omega_{n,m}D^T B_{n,m}^{-1} - \Omega D^T B^{-1}\right)B^{1/2} \in \mathbb{R}^{2\times m}$$

is a fixed matrix, and $Z_j = B^{-1/2}X_j$. Second, we show that

$$\|V\|_F + d^{1/2}K^2\log^{1/2}(m)\|V\|_2/\sqrt{c} \leqslant \widetilde{C}n_{\min}^{-1/2}\|\Delta\|_2.$$

Third, we use the first and second steps combined with the Hanson-Wright inequality to show that with high probability, $\|VZ_j\|_2$ is at most $\widetilde{C}n_{\min}^{-1/2}\|\Delta\|_2$.

For the first step of the proof, let $Z_j = B^{-1/2}X_j$, and note that $\widehat{\beta}_j(B_{n,m}^{-1}) - \widehat{\beta}_j(B^{-1}) = VZ_j$, where $V \in \mathbb{R}^{2\times m}$ is a fixed matrix, because

$$\begin{aligned}\widehat{\beta}_j(B_{n,m}^{-1}) - \widehat{\beta}_j(B^{-1}) &= \left[(D^TB_{n,m}^{-1}D)^{-1}D^TB_{n,m}^{-1} - \Omega D^TB^{-1}\right]B^{1/2}(B^{-1/2}X_j)\\ &= \left[(D^TB_{n,m}^{-1}D)^{-1}D^TB_{n,m}^{-1} - \Omega D^TB^{-1}\right]B^{1/2}Z_j.\end{aligned}$$

For the second step of the proof, we show that $\|V\|_F + d^{1/2}K^2\log^{1/2}(m)\|V\|_2/\sqrt{c} \leqslant \widetilde{C}n_{\min}^{-1/2}\|\Delta\|_2$. First we obtain an upper bound on V. By the triangle inequality,

$$\begin{aligned}\|\Omega_{n,m}D^TB_{n,m}^{-1} - \Omega D^TB^{-1}\|_2 &= \left\|\Omega_{n,m}D^TB_{n,m}^{-1} - \Omega D^TB^{-1}\right\|_2\\ &\leqslant \left\|(\Omega_{n,m}-\Omega)\,D^T(B_{n,m}^{-1}-B^{-1})\right\|_2 + \left\|(\Omega_{n,m}-\Omega)\,D^TB^{-1}\right\|_2 + \left\|\Omega D^T\Delta\right\|_2\\ &= \delta_2(\Delta)+\delta_3(\Delta)+\delta_1(\Delta).\end{aligned}$$

We bound each of the three terms using Proposition S3,

$$\begin{aligned}\delta_2(\Delta) &= \left\|(\Omega_{n,m}-\Omega)\,D^T\Delta\right\|_2 \leqslant \frac{\sqrt{n_{\text{ratio}}}}{\sqrt{n_{\min}}}\frac{\|B\|_2^2\|\Delta\|_2^2}{1/n_{\text{ratio}} - \|B\|_2\|\Delta\|_2}\\ \delta_3(\Delta) &= \left\|(\Omega_{n,m}-\Omega)\,D^TB^{-1}\right\|_2 \leqslant \frac{\sqrt{n_{\text{ratio}}}}{\sqrt{n_{\min}}}\frac{\|B\|_2^2\|B^{-1}\|_2\|\Delta\|_2}{1/n_{\text{ratio}} - \|B\|_2\|_2\|\Delta\|_2}\\ \delta_1(\Delta) &= \left\|\Omega D^T\Delta\right\|_2 \leqslant \frac{\sqrt{n_{\text{ratio}}}}{\sqrt{n_{\min}}}\|B\|_2\|\Delta\|_2.\end{aligned}$$

Applying the above bounds yields

$$\begin{aligned}\|V\|_2 &\leqslant \frac{\sqrt{n_{\text{ratio}}}}{\sqrt{n_{\min}}}\|\Delta\|_2\|B\|_2^{1/2}\left(\frac{\|B\|_2^2\|\Delta\|_2}{1/\kappa^2(D) - \|B\|_2\|\Delta\|_2} + \frac{\|B\|_2^2\|B^{-1}\|_2}{1/\kappa^2(D) - \|B\|_2\|_2\|\Delta\|_2} + \|B\|_2\right)\\ &\leqslant \widetilde{C}n_{\min}^{-1/2}\|\Delta\|_2.\end{aligned}$$

For the third step of the proof, we use the Hanson-Wright inequality to bound $\|VZ_j\|_2$:

$$\begin{aligned}
P\left(\|VZ_j\|_2 > \widetilde{C} n_{\min}^{-1/2}\|\Delta\|_2\right) &\leqslant P\left(\|VZ_j\|_2 > \|V\|_F + d^{1/2}K^2\log^{1/2}(m)\|V\|_2/\sqrt{c}\right) \\
&= P\left(\|VZ_j\|_2 - \|V\|_F > d^{1/2}K^2\log^{1/2}(m)\|V\|_2/\sqrt{c}\right) \\
&\leqslant P\left(|\|VZ_j\|_2 - \|V\|_F| > d^{1/2}K^2\log^{1/2}(m)\|V\|_2/\sqrt{c}\right) \\
&\leqslant 2\exp\left(-\frac{c\left(d^{1/2}K^2\log^{1/2}(m)\|V\|_2/\sqrt{c}\right)^2}{K^4\|V\|_2^2}\right) \qquad \text{(Hanson-Wright inequality)} \\
&= 2\exp\left(-d\log(m)\right) = 2/m^d.
\end{aligned}$$

□

D.5 Proof of Proposition S1

Let $D = U\Psi V^T$ be the singular value decomposition of D, with $U \in \mathbb{R}^{n\times 2}$, $\Psi \in \mathbb{R}^{2\times 2}$, and $V \in \mathbb{R}^{2\times 2}$. Then $(D^TB^{-1}D)^{-1} = (V\Psi U^TB^{-1}U\Psi V^T)^{-1} = V\Psi^{-1}(U^TB^{-1}U)^{-1}\Psi^{-1}V^T$. Thus

$$\begin{aligned}
\|(D^TB^{-1}D)^{-1}\|_2 &= \|\Psi^{-1}(U^TB^{-1}U)^{-1}\Psi^{-1}\|_2 \qquad \text{(because } V \text{ is square, orthonormal)} \\
&\leqslant \|\Psi^{-1}\|_2^2\|(U^TB^{-1}U)^{-1}\|_2 \qquad \text{(sub-multiplicative property)} \\
&= \sigma_{\max}^2(\Psi^{-1})\|(U^TB^{-1}U)^{-1}\|_2 \\
&= \|(U^TB^{-1}U)^{-1}\|_2/\sigma_{\min}^2(\Psi) = \|(U^TB^{-1}U)^{-1}\|_2/\sigma_{\min}^2(D),
\end{aligned}$$

where $\sigma_{\min}(D) = \sigma_{\min}(\Psi)$, because Ψ is the diagonal matrix of singular values of D. Next, note that $\|(U^TB^{-1}U)^{-1}\|_2 = 1/\varphi_{\min}(U^TB^{-1}U)$ and

$$\varphi_{\min}(U^TB^{-1}U) = \min_{\eta\in\mathbb{R}^2}\eta^TU^TB^{-1}U\eta/\eta^T\eta.$$

We perform the change of variables $\gamma = U\eta$, under which $\eta^T\eta = \gamma^T U^T U\gamma = \gamma^T\gamma$ (that is, U preserves the length of η because the columns of U are orthonormal). Hence

$$\begin{aligned}\varphi_{\min}(U^T B^{-1} U) &= \min_{\gamma\in\mathrm{col}(U),\gamma\neq 0} \gamma^T B^{-1}\gamma/\gamma^T\gamma \\ &\geqslant \min_{\gamma\neq 0} \gamma^T B^{-1}\gamma/\gamma^T\gamma \\ &= \varphi_{\min}(B^{-1}) = 1/\|B\|_2.\end{aligned}$$

We have shown that $1/\varphi_{\min}(U^T B^{-1} U) \leqslant \|B\|_2$, which implies that

$$\|(U^T B^{-1} U)^{-1}\|_2 \leqslant \|B\|_2.$$

Therefore

$$\|(D^T B^{-1} D)^{-1}\|_2 \leqslant \|B\|_2/\sigma^2_{\min}(D).$$

In the special case of the two-group design matrix, $\sigma^2_{\min}(D) = n_{\min}$, so $\|(D^T B^{-1} D)^{-1}\|_2 \leqslant \|B\|_2/n_{\min}$.

The proof of (S12) is as follows:

$$\lambda_{\min}(\Omega) = \frac{1}{\lambda_{\max}(\Omega^{-1})} = \frac{1}{\lambda_{\max}(D^T B^{-1} D)} \geqslant \frac{1}{\|D\|_2^2\,\lambda_{\max}(B^{-1})} = \frac{\lambda_{\min}(B)}{\|D\|_2^2} = \frac{\lambda_{\min}(B)}{n_{\max}}.$$

□

D.6 Proof of Proposition S3

By the definitions of $\Omega_{n,m}$ in (S2) and $\Delta = B_{n,m}^{-1} - B^{-1}$, we have by Theorem S2

$$\begin{aligned}\|\Omega_{n,m} - \Omega\|_2 &= \|(D^T B_{n,m} D)^{-1} - \Omega\|_2 \\ &= \left\|\left(D^T B_{n,m}^{-1} D - D^T B^{-1} D + D^T B^{-1} D\right)^{-1} - \Omega\right\|_2 \\ &= \left\|\left(D^T B^{-1} D + D^T \Delta D\right)^{-1} - \Omega\right\|_2 \\ &\leqslant \frac{\|D^T\Delta D\|_2 \|\Omega\|_2^2}{1 - \|\Omega\|_2\|D^T\Delta D\|_2} \qquad \text{(by Theorem S2)} \\ &\leqslant \frac{(\sigma^2_{\max}(D)/\sigma^4_{\min}(D))\,\|B\|_2^2\|\Delta\|_2)}{1 - \kappa^2(D)\|B\|_2\|\Delta\|_2}.\end{aligned}$$

In the last step we apply Proposition S1. Thus

$$\begin{aligned}\|\Omega_{n,m} - \Omega\|_2 &\leqslant \frac{1}{\sigma^2_{\min}(D)} \frac{\kappa^2(D)\|B\|_2^2\|\Delta\|_2}{1 - \kappa^2(D)\|B\|_2\|\Delta\|_2} \\ &= \frac{1}{\sigma^2_{\min}(D)} \frac{\|B\|_2^2\|\Delta\|_2}{(1/\kappa^2(D)) - \|B\|_2\|\Delta\|_2}.\end{aligned}$$

We prove (S14) using the submultiplicative property of the operator norm and Proposition S1:

$$\left\|\Omega D^T \Delta\right\|_2 \leqslant \frac{\|B\|_2}{\sigma^2_{\min}(D)} \sigma_{\max}(D)\|\Delta\|_2 = \frac{\kappa(D)}{\sigma_{\min}(D)}\|B\|_2\|\Delta\|_2.$$

We prove (S15), as follows:

$$\begin{aligned}\left\|\left(\Omega_{n,m} - \Omega\right) D^T \Delta\right\|_2 &\leqslant \|\Omega_{n,m} - \Omega\|_2 \left\|D^T\right\|_2 \|\Delta\|_2 \\ &\leqslant \left[\frac{1}{\sigma^2_{\min}(D)} \frac{\|B\|_2^2\|\Delta\|_2}{(1/\kappa^2(D)) - \|B\|_2\|\Delta\|_2}\right] \sigma_{\max}(D)\|\Delta\|_2 \qquad \text{(by Proposition S3)} \\ &= \frac{\kappa(D)}{\sigma_{\min}(D)} \frac{\|B\|_2^2\|\Delta\|_2^2}{(1/\kappa^2(D)) - \|B\|_2\|\Delta\|_2}.\end{aligned}$$

The proof of (S16) is analogous. □

E Proof of Theorem 3

Note that the proof in the current Section follows exactly the same steps as the proof of Theorems 3.1 and 3.2 in Zhou (2014a). Theorem 3 **Part II** is proved in Section E.2. To prove Theorem 3 **Part I**, we first state Lemma S7, which establishes rates of convergence for estimating A^{-1} and B^{-1} in the operator and the Frobenius norm. We then state the auxiliary Lemma S8, which is identical to that for Theorems 11.1 and 11.2 of Zhou (2014a), except that we plug in $\widetilde{\alpha}$ and $\widetilde{\eta}$ as defined in (S17). Putting these results together proves Theorem 3, **Part I.** We prove these auxiliary results in Section F.

Let $\mathcal{X}_0$ denote the event

$$\forall i,j \qquad \left| \frac{(e_i - p_i)^T X X^T (e_j - p_j)}{\operatorname{tr}(A^*)\sqrt{b^*_{ii} b^*_{jj}}} - \rho_{ij}(B) \right| \leqslant \widetilde{\alpha} \tag{S26}$$

$$\forall i,j \qquad \left| \frac{X_i^T (I - P_2) X_j}{\operatorname{tr}(B^*)\sqrt{a^*_{ii} a^*_{jj}}} - \rho_{ij}(A) \right| \leqslant \widetilde{\eta}, \tag{S27}$$

with $\mathcal{X}_0(B)$ and $\mathcal{X}_0(A)$ denoting the events defined by equations (S26) and (S27), respectively.

Let $\widetilde{\alpha}$ and $\widetilde{\eta}$ be as defined in (S17). On event $\mathcal{X}_0(A)$, for all j, $\widehat{\Gamma}_{jj}(A) = \rho_{jj}(A) = 1$ and

$$\max_{j,k,j\neq k} |\widehat{\Gamma}_{jk}(A) - \rho_{jk}(A)| \leqslant \frac{2\widetilde{\eta}}{1 - \widetilde{\eta}} \tag{S28}$$

On event $\mathcal{X}_0(B)$, for all j, $\widehat{\Gamma}_{jj}(B) = \rho_{jj}(B) = 1$ and

$$\max_{j,k,j\neq k} |\widehat{\Gamma}_{jk}(B) - \rho_{jk}(B)| \leqslant \frac{2\widetilde{\alpha}}{1 - \widetilde{\alpha}}. \tag{S29}$$

Lemma S7. *Suppose (A1) and (A2) hold. Let $\widehat{W}_1$ and $\widehat{W}_2$ be as defined in (10). Let $\widehat{A}_\rho$ and $\widehat{B}_\rho$ be as defined in (8a) and (8b). For some absolute constants $18 < C, C' < 36$, the following events hold with probability at least $1 - 2/(n \vee m)^2$,*

$$\delta_{A,2} := \|\widehat{W}_1 \widehat{A}_\rho \widehat{W}_1 / \operatorname{tr}(B) - A\|_2 \leqslant C a_{\max} \kappa(\rho(A))^2 \lambda_B \sqrt{|A^{-1}|_{0,off} \vee 1} \tag{S30}$$

$$\delta_{B,2} := \|\widehat{W}_2 \widehat{B}_\rho \widehat{W}_2 / \operatorname{tr}(A) - B\|_2 \leqslant C' b_{\max} \kappa(\rho(B))^2 \lambda_A \sqrt{|B^{-1}|_{0,off} \vee 1} \tag{S31}$$

$$\delta_{A,F} := \|\widehat{W}_1 \widehat{A}_\rho \widehat{W}_1 / \operatorname{tr}(B) - A\|_F \leqslant C a_{\max} \kappa(\rho(A))^2 \lambda_B \sqrt{|A^{-1}|_{0,off} \vee m} \tag{S32}$$

$$\delta_{B,F} := \|\widehat{W}_2 \widehat{B}_\rho \widehat{W}_2 / \operatorname{tr}(A) - B\|_F \leqslant C' b_{\max} \kappa(\rho(B))^2 \lambda_A \sqrt{|B^{-1}|_{0,off} \vee n}; \tag{S33}$$

and for some $10 < C, C' < 19$,

$$\begin{aligned}
\delta^-_{A,2} &:= \left\| \mathrm{tr}(B) \left(\widehat{W}_1 \widehat{A}_\rho \widehat{W}_1 \right)^{-1} - A^{-1} \right\|_2 \leqslant \frac{C\lambda_B \sqrt{|A^{-1}|_{0,off} \vee 1}}{a_{\min} \varphi^2_{\min}(\rho(A))} \\
\delta^-_{B,2} &:= \left\| \mathrm{tr}(A) \left(\widehat{W}_2 \widehat{B}_\rho \widehat{W}_2 \right)^{-1} - B^{-1} \right\|_2 \leqslant \frac{C'\lambda_A \sqrt{|B^{-1}|_{0,off} \vee 1}}{b_{\min} \varphi^2_{\min}(\rho(B))} \\
\delta^-_{A,F} &:= \left\| \mathrm{tr}(B) \left(\widehat{W}_1 \widehat{A}_\rho \widehat{W}_1 \right)^{-1} - A^{-1} \right\|_F \leqslant \frac{C\lambda_B \sqrt{|A^{-1}|_{0,off} \vee m}}{a_{\min} \varphi^2_{\min}(\rho(A))} \\
\delta^-_{B,F} &:= \left\| \mathrm{tr}(A) \left(\widehat{W}_2 \widehat{B}_\rho \widehat{W}_2 \right)^{-1} - B^{-1} \right\|_F \leqslant \frac{C'\lambda_A \sqrt{|B^{-1}|_{0,off} \vee n}}{b_{\min} \varphi^2_{\min}(\rho(B))}.
\end{aligned}$$

Lemma S8 follows from Theorems 11.1 and 11.2 of Zhou (2014a,b), where we now plug in $\widetilde{\alpha}$ and $\widetilde{\eta}$ as defined in (S17). For completeness, we provide a sketch in Section F.2.

Lemma S8. *Suppose (A1) and (A2) hold. For* $\varepsilon_1, \varepsilon_2 \in (0,1)$, *let*

$$\lambda_A = \widetilde{\eta}/\varepsilon_1, \quad \lambda_B = \widetilde{\alpha}/\varepsilon_2,$$

for $\widetilde{\alpha}$, $\widetilde{\eta}$ *as defined in (S17), and suppose* $\lambda_A, \lambda_B < 1$. *Then on event* $\mathcal{X}_0$, *for* $18 < C, C' < 36$,

$$\begin{aligned}
&\|\widehat{A \otimes B} - A \otimes B\|_2 \leqslant \frac{\lambda_A \wedge \lambda_B}{2} \|A\|_2 \|B\|_2 + C\lambda_B a_{\max} \|B\|_2 \kappa(\rho(A))^2 \sqrt{|A^{-1}|_{0,off} \vee 1} \\
&+ C'\lambda_A b_{\max} \|A\|_2 \kappa(\rho(B))^2 \sqrt{|B^{-1}|_{0,off} \vee 1} \\
&+ 2 \left[C'\lambda_A b_{\max} \kappa(\rho(B))^2 \sqrt{|B^{-1}|_{0,off} \vee 1} \right] \left[C\lambda_B a_{\max} \kappa(\rho(A))^2 \sqrt{|A^{-1}|_{0,off} \vee 1} \right],
\end{aligned}$$

and for $10 < C, C' < 19$,

$$\begin{aligned}
\|\widehat{A \otimes B}^{-1} - A^{-1} \otimes B^{-1}\|_2 &\leqslant \frac{\lambda_A \wedge \lambda_B}{3} \|A^{-1}\|_2 \|B^{-1}\|_2 + C\lambda_B \|B^{-1}\|_2 \frac{\sqrt{|A^{-1}|_{0,off} \vee 1}}{a_{\min} \varphi^2_{\min}(\rho(A))} \\
&+ C'\lambda_A \|A^{-1}\|_2 \frac{\sqrt{|B^{-1}|_{0,off} \vee 1}}{b_{\min} \varphi^2_{\min}(\rho(B))} + \frac{3}{2} \left[C\lambda_B \frac{\sqrt{|A^{-1}|_{0,off} \vee 1}}{a_{\min} \varphi^2_{\min}(\rho(A))} \right] \left[C'\lambda_A \frac{\sqrt{|B^{-1}|_{0,off} \vee 1}}{b_{\min} \varphi^2_{\min}(\rho(B))} \right];
\end{aligned}$$

For $18 < C, C' < 36$*,*

$$\|\widehat{A \otimes B} - A \otimes B\|_F \leqslant \frac{\lambda_A \wedge \lambda_B}{2}\|A\|_F\|B\|_F + C\lambda_B a_{\max}\|B\|_F\kappa(\rho(A))^2\sqrt{|A^{-1}|_{0,off} \vee m}$$
$$+C'\lambda_A b_{\max}\|A\|_F\kappa(\rho(B))^2\sqrt{|B^{-1}|_{0,off} \vee n}$$
$$+2\left[C'\lambda_A b_{\max}\kappa(\rho(B))^2\sqrt{|B^{-1}|_{0,off} \vee n}\right]\left[C\lambda_B a_{\max}\kappa(\rho(A))^2\sqrt{|A^{-1}|_{0,off} \vee m}\right],$$

and for $10 < C, C' < 19$*,*

$$\|\widehat{A \otimes B}^{-1} - A^{-1} \otimes B^{-1}\|_F \leqslant \frac{\lambda_A \wedge \lambda_B}{3}\|A^{-1}\|_2\|B^{-1}\|_F + C\lambda_B\|B^{-1}\|_F\frac{\sqrt{|A^{-1}|_{0,off} \vee m}}{a_{\min}\varphi^2_{\min}(\rho(A))}$$
$$+ C'\lambda_A\|A^{-1}\|_F\frac{\sqrt{|B^{-1}|_{0,off} \vee n}}{b_{\min}\varphi^2_{\min}(\rho(B))} + \frac{7}{5}\left[C\lambda_B\frac{\sqrt{|A^{-1}|_{0,off} \vee m}}{a_{\min}\varphi^2_{\min}(\rho(A))}\right]\left[C'\lambda_A\frac{\sqrt{|B^{-1}|_{0,off} \vee n}}{b_{\min}\varphi^2_{\min}(\rho(B))}\right].$$

E.1 Proof of Theorem 3, Part I

We state additional helpful bounds:

$$(a_{\min} \vee \varphi_{\min}(A))\sqrt{m} \leqslant \|A\|_F = \left(\sum_{i=1}^m \varphi_i^2(A)\right)^{1/2} \leqslant \sqrt{m}\|A\|_2, \tag{S34}$$

$$(b_{\min} \vee \varphi_{\min}(B))\sqrt{n} \leqslant \|B\|_F = \left(\sum_{i=1}^m \varphi_i^2(B)\right)^{1/2} \leqslant \sqrt{n}\|B\|_2, \tag{S35}$$

$$\sqrt{m}/a_{\max} = \left(\frac{1}{a_{\max}} \vee \frac{1}{\varphi_{\max}(A)}\right)\sqrt{m} \leqslant \|A^{-1}\|_F \leqslant \sqrt{m}\|A^{-1}\|_2, \tag{S36}$$

and

$$\sqrt{n}/b_{\max} = \left(\frac{1}{b_{\max}} \vee \frac{1}{\varphi_{\max}(B)}\right)\sqrt{n} \leqslant \|B^{-1}\|_F \leqslant \sqrt{n}\|B^{-1}\|_2. \tag{S37}$$

Proof of Theorem 3, **Part I**. We plug in bounds as in (S9) and (S10) into Lemma

S8 to obtain under (A1) and (A2), $\left\|\widehat{A \otimes B} - A \otimes B\right\|_2 \leqslant \|A\|_2 \|B\|_2 \delta$, where

$$\begin{aligned}
\delta &= \frac{\lambda_A \wedge \lambda_B}{2} + \frac{C r_a \kappa(\rho(A))}{\varphi_{\min}(\rho(A))} \lambda_B \sqrt{|A^{-1}|_{0,\text{off}} \vee 1} + \frac{C' r_b \kappa(\rho(B))}{\varphi_{\min}(\rho(B))} \lambda_A \sqrt{|B^{-1}|_{0,\text{off}} \vee 1} \\
&+ 2 \left[\frac{C r_a \kappa(\rho(A))}{\varphi_{\min}(\rho(A))} \lambda_B \sqrt{|A^{-1}|_{0,\text{off}} \vee 1} \right] \left[\frac{C' r_b \kappa(\rho(B))}{\varphi_{\min}(\rho(B))} \lambda_A \sqrt{|B^{-1}|_{0,\text{off}} \vee 1} \right] \\
&= \frac{\lambda_A \wedge \lambda_B}{2} + \log^{1/2}(m \vee n) \left(\sqrt{\frac{|A^{-1}|_{0,\text{off}} \vee 1}{m}} + \sqrt{\frac{|B^{-1}|_{0,\text{off}} \vee 1}{n}} \right) + o(1).
\end{aligned}$$

For the inverse, we plug in bounds as in (S7) and (S8) into Lemma S8 to obtain under (A1) and (A2), $\left\|\widehat{A \otimes B}^{-1} - A^{-1} \otimes B^{-1}\right\|_2 \leqslant \|A^{-1}\|_2 \|B^{-1}\|_2 \delta'$, where

$$\begin{aligned}
\delta' &= \frac{\lambda_A \wedge \lambda_B}{3} + \frac{C r_a \lambda_B \sqrt{|A^{-1}|_{0,\text{off}} \vee 1}}{\varphi_{\min}(\rho(A))} + \frac{C' r_b \lambda_A \sqrt{|B^{-1}|_{0,\text{off}} \vee 1}}{\varphi_{\min}(\rho(B))} \\
&+ \frac{3}{2} \left[\frac{C r_a \lambda_B \sqrt{|A^{-1}|_{0,\text{off}} \vee 1}}{\varphi_{\min}(\rho(A))} \right] \left[\frac{C' r_b \lambda_A \sqrt{|B^{-1}|_{0,\text{off}} \vee 1}}{\varphi_{\min}(\rho(B))} \right] \\
&\asymp \frac{\lambda_A \wedge \lambda_B}{3} + \log^{1/2}(m \vee n) \left(\sqrt{\frac{|A^{-1}|_{0,\text{off}} \vee 1}{m}} + \sqrt{\frac{|B^{-1}|_{0,\text{off}} \vee 1}{n}} \right) + o(1).
\end{aligned}$$

The bounds in the Frobenius norm are proved in a similar manner; see Zhou (2014a) to finish. □

E.2 Proof of Theorem 3, Part II

Let $\widehat{B}^{-1} = \widehat{W}_2 \widehat{B}_\rho \widehat{W}_2$. Let $\widehat{\Delta} = \widehat{B}^{-1} - B^{-1}$. Let $\mathcal{E}_0(B)$ denote the event given by equations (S34) and (S34), which we know has probability at least $1 - 2/(n \vee m)^2$ from Lemma S7, and define the event

$$\mathcal{E}_4 = \left\{ \|\widehat{\beta}_j(\widehat{B}^{-1}) - \beta_j^*\|_2 \leqslant s_{n,m} + t'_{n,m} \right\}, \tag{S38}$$

where $s_{n,m}$ is as defined in (19) and

$$t'_{n,m} := C \lambda_A \sqrt{\frac{n_{\text{ratio}} \left(|B_0^{-1}|_{0,\text{off}} \vee 1 \right)}{n_{\min}}}. \tag{S39}$$

Under $\mathcal{E}_0(B)$, we see that

$$\|\widehat{\Delta}\|_2 \leqslant \frac{C'\lambda_A\sqrt{|B^{-1}|_{0,\text{off}} \vee 1}}{b_{\min}\varphi^2_{\min}(\rho(B))} = o(1). \tag{S40}$$

Using Proposition S1 and the fact that $\|D\|_2 = \sqrt{n_{\max}}$, we get that

$$\|\Omega D^T \widehat{\Delta} D\|_2 \leqslant n_{\text{ratio}}\|B\|_2\|\widehat{\Delta}\|_2, \tag{S41}$$

From (S40) we know that $\|\widehat{\Delta}\|_2 \leqslant 1/(n_{\text{ratio}}\|B\|_2)$, which we can plug into (S41) to show that $\|\Omega D^T \widehat{\Delta} D\|_2 < 1$. This implies that $\widetilde{C} n_{\min}^{-1/2}\|\widehat{\Delta}\|_2 \leqslant t'_{n,m}$. Therefore, we can apply Theorem 1 to get that the conditional probability of $\mathcal{E}_4$ given $\mathcal{E}_0(B)$ is at least $1 - 4/(n \vee m)^2$.

We can then bound the unconditional probability,

$$\begin{aligned} P(\mathcal{E}_4^c) &\leqslant P\left(\mathcal{E}_4^c \mid \mathcal{E}_0(B)\right) P\left(\mathcal{E}_0(B)\right) + P\left(\mathcal{E}_0(B)^c\right) \\ &\leqslant P\left(\mathcal{E}_4^c \mid \mathcal{E}_0(B)\right) + P\left(\mathcal{E}_0(B)^c\right) \\ &\leqslant \frac{4}{(n \vee m)^2} + \frac{2}{(n \vee m)^2}. \end{aligned}$$

□

F More proofs for Theorem 3

The proof of Lemma S7 appears in Section F.1. The proofs of auxiliary lemmas appear in Section F.2.

F.1 Proof of Lemma S7

In order to prove Lemma S7, we need Theorem S9, which shows explicit non-asymptotic convergence rates in the Frobenius norm for estimating $\rho(A)$, $\rho(B)$, and their inverses. Theorem S9 follows from the standard proof; see Rothman et al. (2008); Zhou et al. (2011) We also need Proposition S11 and Lemma S10, which are stated below and proved in Section F.2.

Theorem S9. *Suppose that (A2) holds. Let $\widehat{A}_\rho$ and $\widehat{B}_\rho$ be the unique minimizers defined by* (8a) *and* (8b) *with sample correlation matrices $\widehat{\Gamma}(A)$ and $\widehat{\Gamma}(B)$ as their input.*

Suppose that event $\mathcal{X}_0$ *holds, with*

$$\widetilde{\eta}\sqrt{|A^{-1}|_{0,off} \vee 1} = o(1) \quad \textit{and} \quad \widetilde{\alpha}\sqrt{|B^{-1}|_{0,off} \vee 1} = o(1).$$
$$\textit{Set for some } \ 0 < \epsilon, \varepsilon < 1, \ \ \lambda_B = \widetilde{\alpha}/\varepsilon \ \ \textit{and} \ \ \lambda_A = \widetilde{\eta}/\epsilon. \tag{S42}$$

Then on event $\mathcal{X}_0$*, we have for* $9 < C < 18$

$$\left\|\widehat{A}_\rho - \rho(A)\right\|_2 \leqslant \left\|\widehat{A}_\rho - \rho(A)\right\|_F \leqslant C\kappa(\rho(A))^2\lambda_B\sqrt{|A^{-1}|_{0,off} \vee 1},$$
$$\left\|\widehat{B}_\rho - \rho(B)\right\|_2 \leqslant \left\|\widehat{B}_\rho - \rho(B)\right\|_F \leqslant C\kappa(\rho(B))^2\lambda_A\sqrt{|B^{-1}|_{0,off} \vee 1},$$

and

$$\left\|\widehat{A}_\rho^{-1} - \rho(A)^{-1}\right\|_2 \leqslant \left\|\widehat{A}_\rho^{-1} - \rho(A)^{-1}\right\|_F < \frac{C\lambda_B\sqrt{|A^{-1}|_{0,off} \vee 1}}{2\varphi^2_{\min}(\rho(A))}, \tag{S43}$$

$$\left\|\widehat{B}_\rho^{-1} - \rho(B)^{-1}\right\|_2 \leqslant \left\|\widehat{B}_\rho^{-1} - \rho(B)^{-1}\right\|_F \leqslant \frac{C\lambda_A\sqrt{|B^{-1}|_{0,off} \vee 1}}{2\varphi^2_{\min}(\rho(B))}. \tag{S44}$$

We now state an auxiliary result, Lemma S10, where we prove a bound on the error in the diagonal entries of the covariance matrices, and on their reciprocals. The following Lemma provides bounds analogous to those in Claim 15.1 Zhou (2014a,b).

Lemma S10. *Let* $\widehat{W}_1$ *and* $\widehat{W}_2$ *be as defined in* (10)*. Let* $W_1 = \sqrt{\mathrm{tr}(B)}\,\mathrm{diag}(A)^{1/2}$ *and* $W_2 = \sqrt{\mathrm{tr}(A)}\,\mathrm{diag}(B)^{1/2}$*. Suppose event* $\mathcal{X}_0$ *holds, as defined in* (S26), (S27)*. For* $\eta' := \frac{\widetilde{\eta}}{\sqrt{1-\widetilde{\eta}}} \leqslant \frac{\lambda_B}{6}$ *and* $\alpha' := \frac{\widetilde{\alpha}}{\sqrt{1-\widetilde{\alpha}}} \leqslant \frac{\lambda_A}{6}$*,*

$$\left\|\widehat{W}_1 - W_1\right\|_2 \leqslant \widetilde{\eta}\sqrt{\mathrm{tr}\,(B)}\sqrt{a_{\max}}, \qquad \left\|\widehat{W}_1^{-1} - W_1^{-1}\right\|_2 \leqslant \frac{\widetilde{\eta}}{1-\widetilde{\eta}}/\sqrt{\mathrm{tr}\,(B)}\sqrt{a_{\min}},$$
$$\left\|\widehat{W}_2 - W_2\right\|_2 \leqslant \widetilde{\alpha}\sqrt{\mathrm{tr}\,(A)}\sqrt{b_{\max}}, \ \textit{and} \ \left\|\widehat{W}_2^{-1} - W_2^{-1}\right\|_2 \leqslant \frac{\widetilde{\alpha}}{1-\widetilde{\alpha}}/\sqrt{\mathrm{tr}\,(A)}\sqrt{b_{\min}}.$$

Proposition S11. *(Zhou, 2014a). Let* $\widehat{W}$ *and* W *be diagonal positive definite matrices. Let*

$\widehat{\Psi}$ *and* Ψ *be symmetric positive definite matrices. Then*

$$\begin{aligned}\left\|\widehat{W}\widehat{\Psi}\widehat{W} - W\Psi W\right\|_2 &\leqslant \left(\left\|\widehat{W} - W\right\|_2 + \|W\|_2\right)^2 \left\|\widehat{\Psi} - \Psi\right\|_2 \\ &\quad + \left\|\widehat{W} - W\right\|_2 \left(\left\|\widehat{W} - W\right\|_2 + 2\right) \|\Psi\|_2 \\ \left\|\widehat{W}\widehat{\Psi}\widehat{W} - W\Psi W\right\|_F &\leqslant \left(\left\|\widehat{W} - W\right\|_2 + \|W\|_2\right)^2 \left\|\widehat{\Psi} - \Psi\right\|_F \\ &\quad + \left\|\widehat{W} - W\right\|_2 \left(\left\|\widehat{W} - W\right\|_2 + 2\right) \|\Psi\|_F.\end{aligned}$$

Proof of Lemma S7. Assume that event $\mathcal{X}_0$ holds. The proof follows exactly that of Lemma 15.3 in Zhou (2014a,b), in view of Theorem S9, Lemma S10 and Proposition 15.2 from Zhou (2014a,b), which is restated immediately above in Proposition S11. □

It remains to prove Lemma S10.

Proof of Lemma S10. Suppose that event $\mathcal{X}_0$ holds. Then

$$\max_{i=1,\ldots,m} \left| \frac{\sqrt{X_i^T(I - P_2)X_i}}{\sqrt{a_{ii}\operatorname{tr}(B)}} - 1 \right| \leqslant \left(1 - \sqrt{1 - \widetilde{\eta}}\right) \bigvee \left(\sqrt{1 + \widetilde{\eta}} - 1\right) \leqslant \widetilde{\eta}.$$

Thus for all i,

$$\frac{1}{\sqrt{1 + \widetilde{\eta}}} \leqslant \frac{\sqrt{a_{ii}\operatorname{tr}(B)}}{\sqrt{X_i^T(I - P_2)X_i}} \leqslant \frac{1}{\sqrt{1 - \widetilde{\eta}}},$$

so

$$\left| \frac{\sqrt{a_{ii}\operatorname{tr}(B)}}{\sqrt{X_i^T(I - P_2)X_i}} - 1 \right| \leqslant \left(\frac{1 - \sqrt{1 - \widetilde{\eta}}}{\sqrt{1 - \widetilde{\eta}}} \right) \bigvee \left(\frac{\sqrt{1 + \widetilde{\eta}} - 1}{\sqrt{1 + \widetilde{\eta}}} \right) \leqslant \frac{\widetilde{\eta}}{\sqrt{1 - \widetilde{\eta}}}.$$

□

F.2 Proof of Lemma S8

In order to prove Lemma S8, we state Lemma S12, Lemma S13, and Proposition S14. Let $\|\cdot\|$ denote a matrix norm such that $\|A \otimes B\| = \|A\|\|B\|$. Let

$$\Delta := \widehat{W}_1\widehat{A}_\rho\widehat{W}_1 \otimes \widehat{W}_2\widehat{B}_\rho\widehat{W}_2/\operatorname{tr}(A)\operatorname{tr}(B) - A \otimes B, \tag{S45}$$

$$\Delta' := \operatorname{tr}(A)\operatorname{tr}(B)\left(\widehat{W}_1\widehat{A}_\rho\widehat{W}_1\right)^{-1} \otimes \left(\widehat{W}_2\widehat{B}_\rho\widehat{W}_2\right)^{-1} - A^{-1} \otimes B^{-1}. \tag{S46}$$

Lemma S12 is identical to Lemma 15.5 of Zhou (2014a), except that we now plug in quantities $\widetilde{\alpha}$ and $\widetilde{\eta}$ as defined in (S17). Likewise, Proposition S14 is analogous to (20) in Theorem 4.1 of Zhou (2014a), except that we now use the centered data matrix $(I - P_2)X$, together with the rates $\widetilde{\alpha}$, $\widetilde{\eta}$.

Lemma S12. *Let $\widehat{A \otimes B}$ be as in (11). Then for $\Sigma = A \otimes B$,*

$$\left\| \widehat{A \otimes B}^{-1} - \Sigma^{-1} \right\| \leqslant (\widetilde{\alpha} \wedge \widetilde{\eta}) \|A^{-1}\| \|B^{-1}\| + (1 + \widetilde{\alpha} \wedge \widetilde{\eta}) \|\Delta'\| \tag{S47}$$

$$\left\| \widehat{A \otimes B} - \Sigma \right\| \leqslant \frac{\lambda_A \wedge \lambda_B}{2} \|A\| \|B\| + (1 + \frac{\lambda_A \wedge \lambda_B}{2}) \|\Delta\|. \tag{S48}$$

Lemma S13 is a helpful bound on the difference of Kronecker products.

Lemma S13. *(Zhou, 2014a). For matrices A_1 and B_1, let $\Delta_A := A_1 - A$ and $\Delta_B := B_1 - B$. Then*

$$\|A_1 \otimes B_1 - A \otimes B\| \leqslant \|\Delta_A\| \|B\| + \|\Delta_B\| \|A\| + \|\Delta_A\| \|\Delta_B\|.$$

Proposition S14. *Under the event $\mathcal{X}_0$, as defined in as defined in* (S26), (S27),

$$\left| \|(I - P_2)X\|_F^2 - \mathrm{tr}(A)\mathrm{tr}(B) \right| \leqslant (\widetilde{\alpha} \wedge \widetilde{\eta}) \mathrm{tr}(A)\mathrm{tr}(B).$$

Proof of Lemma S8. Assume that event $\mathcal{X}_0$ as defined in (S26), (S27) holds. The proof follows exactly the steps in Theorems 11.1 and 11.2 in Supplementary Material of Zhou (2014a,b). □

Proof of Lemma S12. By the triangle inequality and the sub-multiplicativity of the norm $\|\cdot\|$, with Δ and Δ' as defined in (S45) and (S46),

$$\mathrm{tr}(A)\,\mathrm{tr}(B) \left\| \left(\widehat{W}_1^{-1} \widehat{A}_\rho^{-1} \widehat{W}_1^{-1} \right) \otimes \left(\widehat{W}_2^{-1} \widehat{B}_\rho^{-1} \widehat{W}_2^{-1} \right) \right\| \leqslant \|A^{-1}\| \|B^{-1}\| + \|\Delta'\| \tag{S49}$$

$$\left\| \left(\widehat{W}_1 \widehat{A}_\rho \widehat{W}_1 \right) \otimes \left(\widehat{W}_2 \widehat{B}_\rho \widehat{W}_2 \right) / \mathrm{tr}(A)\,\mathrm{tr}(B) \right\| \leqslant \|A\| \|B\| + \|\Delta\|. \tag{S50}$$

Following proof of Lemma 15.5 Zhou (2014a,b), we have by definition of Δ', and Proposition S14, and (S49),

$$\left\| \widehat{A \otimes B}^{-1} - A^{-1} \otimes B^{-1} \right\| \;\leqslant\; (\widetilde{\alpha} \wedge \widetilde{\eta}) \left(\|A^{-1}\| \|B^{-1}\| + \|\Delta'\| \right) + \|\Delta'\|.$$

By Proposition S14, we have for $\lambda_A \geqslant 3\widetilde{\alpha}$, $\lambda_B \geqslant 3\widetilde{\eta}$, where $\widetilde{\alpha} \wedge \widetilde{\eta} \leqslant \frac{\lambda_A \wedge \lambda_B}{3}$,

$$\begin{aligned}
&\left|\frac{1}{\|(I-P_2)X\|_F^2} - \frac{1}{\mathrm{tr}(A)\,\mathrm{tr}(B)}\right| = \left|\frac{\|(I-P_2)X\|_F^2 - \mathrm{tr}(A)\,\mathrm{tr}(B)}{\|(I-P_2)X\|_F^2\,\mathrm{tr}(A)\,\mathrm{tr}(B)}\right| \\
&\leqslant \left|\frac{\widetilde{\alpha}\wedge\widetilde{\eta}}{\|(I-P_2)X\|_F^2}\right| \leqslant \frac{\widetilde{\alpha}\wedge\widetilde{\eta}}{\mathrm{tr}(A)\,\mathrm{tr}(B)(1-\widetilde{\alpha}\wedge\widetilde{\eta})} \\
&\text{thus } \left|\frac{\mathrm{tr}(A)\,\mathrm{tr}(B)}{\|(I-P_2)X\|_F^2} - 1\right| \leqslant \frac{\widetilde{\alpha}\wedge\widetilde{\eta}}{1-\widetilde{\alpha}\wedge\widetilde{\eta}} \leqslant \frac{\lambda_A\wedge\lambda_B}{2}. \qquad \text{(S51)}
\end{aligned}$$

By the triangle inequality, the definition of Δ in (S45), and (S50) and (S51),

$$\left\|\widehat{A\otimes B} - A\otimes B\right\| \;\leqslant\; \frac{\lambda_A+\lambda_B}{2}\|A\|\|B\| + (1+\frac{\lambda_A+\lambda_B}{2})\|\Delta\|;$$

See the proof of Lemma 15.5 Zhou (2014a,b). □

Proof of Proposition S14. Suppose event $\mathcal{X}_0$ holds. Note that

$$E[\|(I-P_2)X\|_F^2] = \mathrm{tr}\left((I-P_2)E[XX^T](I-P_2)\right) = \mathrm{tr}(A)\mathrm{tr}(\widetilde{B})$$

Decomposing by columns, we obtain the inequality,

$$\begin{aligned}
&\left|\|(I-P_2)X\|_F^2 - \mathrm{tr}(A)\mathrm{tr}(B)\right| = \left|\sum_{j=1}^m \|(I-P_2)X_j\|_2^2 - a_{jj}\mathrm{tr}(B)\right| \\
&\leqslant \sum_{j=1}^m \left|X_j^T(I-P_2)X_j - a_{jj}\mathrm{tr}(B)\right| \leqslant \sum_{j=1}^m \widetilde{\eta}_{jj}a_{jj}\mathrm{tr}(B) \leqslant \widetilde{\eta}\mathrm{tr}(A)\mathrm{tr}(B).
\end{aligned}$$

Decomposing by rows, we obtain the inequality,

$$\begin{aligned}
&\left|\|(I-P_2)X\|_F^2 - \mathrm{tr}(A)\mathrm{tr}(B)\right| = \left|\sum_{i=1}^n \|(e_i-p_i)^TX\|_2^2 - b_{ii}\mathrm{tr}(A)\right| \\
&\leqslant \sum_{i=1}^n \left|(e_i-p_i)^TXX^T(e_i-p_i) - b_{ii}\mathrm{tr}(A)\right| \leqslant \sum_{i=1}^n \widetilde{\alpha}_{ii}b_{ii}\mathrm{tr}(A) \leqslant \widetilde{\alpha}\mathrm{tr}(A)\mathrm{tr}(B).
\end{aligned}$$

Therefore $|\|(I-P_2)X\|_F^2 - \mathrm{tr}(A)\mathrm{tr}(B)| \leqslant (\widetilde{\alpha}\wedge\widetilde{\eta})\mathrm{tr}(A)\mathrm{tr}(B)$. □

G Entrywise convergence of sample correlations

In this section we prove entrywise rates of convergence for the sample correlation matrices in Theorem S15. The theorem applies to the Kronecker product model, $\mathrm{Cov}(\mathrm{vec}(X)) = A^* \otimes B^*$, where for identifiability we define the sample covariance matrices as

$$A^* = \frac{m}{\mathrm{tr}(A)} A \quad \text{and} \quad B^* = \frac{\mathrm{tr}(A)}{m} B,$$

with the scaling chosen so that A^* has trace m. Let $\rho(A) \in \mathbb{R}^{m \times m}$ and $\rho(B) \in \mathbb{R}^{n \times n}$ denote the correlation matrices corresponding to covariance matrices A^* and B^*, respectively. Assume that that the mean of X satisfies the two-group model (4). Let P_2 be as defined in (13). The matrix $I - P_2$ is a projection matrix of rank $n-2$ that performs within-group centering. The sample covariance matrices are defined as

$$S(B^*) = \frac{1}{m} \sum_{j=1}^{m} (I - P_2) X_j X_j^T (I - P_2), \tag{S52}$$

$$S(A^*) = X^T (I - P_2) X / n, \tag{S53}$$

where $S(B^*)$ has null space of dimension two.

Theorem S15. *Consider a data generating random matrix as in (2). Let C be some absolute constant. Let $\widetilde{\alpha}$ and $\widetilde{\eta}$ be as defined in* (S17)*. Let $m \vee n \geqslant 2$. Then with probability at least $1 - \frac{3}{(m \vee n)^2}$, for $\widetilde{\alpha}, \widetilde{\eta} < 1/3$, and $\widehat{\Gamma}(A)$ and $\widehat{\Gamma}(B)$ as in (7),*

$$\forall i \neq j, \ \left| \widehat{\Gamma}_{ij}(B) - \rho_{ij}(B) \right| \leqslant \frac{\widetilde{\alpha}}{1 - \widetilde{\alpha}} + |\rho_{ij}(B)| \frac{\widetilde{\alpha}}{1 - \widetilde{\alpha}} \leqslant 3\widetilde{\alpha},$$
$$\forall i \neq j, \ \left| \widehat{\Gamma}_{ij}(A) - \rho_{ij}(A) \right| \leqslant \frac{\widetilde{\eta}}{1 - \widetilde{\eta}} + |\rho_{ij}(A)| \frac{\widetilde{\eta}}{1 - \widetilde{\eta}} \leqslant 3\widetilde{\eta}.$$

We state three results used in the proof of Theorem S15: Proposition S16 provides an entrywise rate of convergence of $S(B^*)$, Proposition S17 provides an entrywise rate of convergence of $S(A^*)$, and Lemma S18 states that these entrywise rates imply $\mathcal{X}_0$. Let

$$\widetilde{B} := (I - P_2) B^* (I - P_2) = \mathrm{Cov}((I - P_2) X_j), \tag{S54}$$

where X_j is the jth column of X. Let $\widetilde{b}_{ij}$ denote the (i,j)th entry of $\widetilde{B}$.

Proposition S16. *Let $d > 2$. Then with probability at least $1 - 2/m^{d-2}$,*

$$\forall i, j \ \left|S_{ij}(B^*) - b^*_{ij}\right| \leqslant \phi_{B,ij}, \tag{S55}$$

with

$$\phi_{B,ij} = C\frac{\log^{1/2}(m)}{\sqrt{m}}\frac{\|A^*\|_F}{\sqrt{m}}\sqrt{\widetilde{b}_{ii}\widetilde{b}_{jj}} + \frac{3\|B^*\|_1}{n_{\min}}. \tag{S56}$$

Proposition S17. *Let $d > 2$. Then with probability at least $1 - 2/n^{d-2}$,*

$$\forall i, j \ \left|S_{ij}(A^*) - a^*_{ij}\operatorname{tr}(B^*)/n\right| > \phi_{A,ij}, \tag{S57}$$

with

$$\phi_{A,ij} = (a^*_{ij}/n)\left|\operatorname{tr}\left(\widetilde{B}\right) - \operatorname{tr}(B^*)\right| + d^{1/2}K\log^{1/2}(n \vee m)(1/n)\sqrt{a^{*2}_{ij} + a^*_{ii}a^*_{jj}}\|\widetilde{B}\|_F. \tag{S58}$$

Lemma S18. *Suppose that (A2) holds and that $m \vee n \geqslant 2$. The event* (S57) *defined in Proposition S17 implies that $\mathcal{X}_0(A)$ holds. Similarly, the event* (S55) *defined in Proposition S16 implies $\mathcal{X}_0(B)$. Hence $\mathbb{P}(\mathcal{X}_0) \geqslant 1 - \frac{3}{(m\vee n)^2}$.*

Proposition S16 is proved in section G.1. Proposition S17 is proved in section G.2. Lemma S18 is proved in section G.3. Note that Lemma S18 follows from Propositions S16 and S17. We now prove Theorem S15, which follows from Lemma S18.

Proof of Theorem S15. Let q_i denote the ith column of $I - P_2$, so that $q_i^T XX^T q_j$ is the (i,j)th entry of $(I - P_2)XX^T(I - P_2)$. Under $\mathcal{X}_0(B)$, the sample correlation $\widehat{\Gamma}(B)$

satisfies the following bound:

$$\begin{aligned}
\left|\widehat{\Gamma}_{ij}(B)-\rho_{ij}(B)\right| &= \left|\frac{q_i^T XX^T q_j}{\sqrt{q_i^T XX^T q_i}\sqrt{q_j^T XX^T q_j}}-\rho_{ij}(B)\right| \\
&= \left|\frac{q_i^T XX^T q_j/\left(\mathrm{tr}(A^*)\sqrt{b_{ii}^* b_{jj}^*}\right)}{\sqrt{q_i^T XX^T q_i/\left(b_{ii}^*\mathrm{tr}(A^*)\right)}\sqrt{q_j^T XX^T q_j/\left(b_{jj}^*\mathrm{tr}(A^*)\right)}}-\rho_{ij}(B)\right| \\
&\leqslant \left|\frac{q_i^T XX^T q_j/\left(\mathrm{tr}(A^*)\sqrt{b_{ii}^* b_{jj}^*}\right)-\rho_{ij}(B)}{\sqrt{q_i^T XX^T q_i/\left(b_{ii}^*\mathrm{tr}(A^*)\right)}\sqrt{q_j^T XX^T q_j/\left(b_{jj}^*\mathrm{tr}(A^*)\right)}}\right| \\
&+ \left|\frac{\rho_{ij}(B)}{\sqrt{q_i^T XX^T q_i/\left(b_{ii}^*\mathrm{tr}(A^*)\right)}\sqrt{q_j^T XX^T q_j/\left(b_{jj}^*\mathrm{tr}(A^*)\right)}}-\rho_{ij}(B)\right| \\
&\leqslant \frac{\widetilde{\alpha}}{1-\widetilde{\alpha}}+|\rho_{ij}(B)|\left|\frac{1}{1-\widetilde{\alpha}}-1\right| \\
&\leqslant 3\widetilde{\alpha},
\end{aligned}$$

where the first inequality holds by $\mathcal{X}_0(B)$ and the second inequality holds for $\widetilde{\alpha} \leqslant 1/3$. Similarly, under $\mathcal{X}_0(A)$ we obtain an entrywise bound on the sample correlation $\widehat{\Gamma}(A)$:

$$\begin{aligned}
\left|\widehat{\Gamma}_{ij}(A)-\rho_{ij}(A)\right| &= \left|\frac{X_i^T(I-P_2)X_j}{\sqrt{X_i^T(I-P_2)X_i}\sqrt{X_j^T(I-P_2)X_j}}-\rho_{ij}(A)\right| \\
&= \left|\frac{X_i^T(I-P_2)X_j/\left(\mathrm{tr}(B^*)\sqrt{a_{ii}^* a_{jj}^*}\right)}{\sqrt{X_i^T(I-P_2)X_i/\left(a_{ii}^*\mathrm{tr}(B^*)\right)}\sqrt{X_j^T(I-P_2)X_j/\left(a_{jj}^*\mathrm{tr}(B^*)\right)}}-\rho_{ij}(A)\right| \\
&\leqslant \left|\frac{X_i^T(I-P_2)X_j/\left(\mathrm{tr}(B^*)\sqrt{a_{ii}^* a_{jj}^*}\right)-\rho_{ij}(A)}{\sqrt{X_i^T(I-P_2)X_i/\left(a_{ii}^*\mathrm{tr}(B^*)\right)}\sqrt{X_j^T(I-P_2)X_j/\left(a_{jj}^*\mathrm{tr}(B^*)\right)}}\right| \\
&+ \left|\frac{\rho_{ij}(A)}{\sqrt{X_i^T(I-P_2)X_i/\left(a_{ii}^*\mathrm{tr}(B^*)\right)}\sqrt{X_j^T(I-P_2)X_j/\left(a_{jj}^*\mathrm{tr}(B^*)\right)}}-\rho_{ij}(A)\right| \\
&\leqslant \frac{\widetilde{\eta}}{1-\widetilde{\eta}}+|\rho_{ij}(A)|\left|\frac{1}{1-\widetilde{\eta}}-1\right| \leqslant 3\widetilde{\eta},
\end{aligned}$$

where the first inequality holds by $\mathcal{X}_0(A)$, and the second inequality holds for $\widetilde{\eta} < 1/3$.

By Lemma S18, the event $\mathcal{X}_0 = \mathcal{X}_0(B) \cap \mathcal{X}_0(A)$ holds with probability at least $1 - 3/(n \vee m)^2$, which completes the proof. $\square$

G.1 Proof of Proposition S16

We first present Lemma S19 and Lemma S20, which decompose the rate of convergence into a bias term and a variance term, respectively. We then combine the rates for the bias and variance terms to prove the entrywise rate of convergence for the sample covariance. Define

$$\mathcal{B}(B^*) := E[S(B^*)] - B^* \quad \text{and} \tag{S59}$$

$$\sigma(B^*) := S(B^*) - E[S(B^*)]. \tag{S60}$$

We state maximum entrywise bounds on $\mathcal{B}(B^*)$ and $\sigma(B^*)$ in Lemma S19 and Lemma S20, respectively. Proofs for these lemmas are provided in Section G.4 and G.5 respectively.

Lemma S19. *For $\mathcal{B}(B^*)$ as defined in* (S59)*,*

$$\|\mathcal{B}(B^*)\|_{\max} \leqslant \frac{3\|B^*\|_1}{n_{\min}}. \tag{S61}$$

Lemma S20. *Let $\sigma(B^*)$ be as defined in* (S60)*. With probability at least $1 - 2/m^d$,*

$$|\sigma_{ij}(B^*)| = \left|S_{ij}(B^*) - b^*_{ij}\right| < C \log^{1/2}(m) \frac{\|A^*\|_F}{\operatorname{tr}(A^*)} \sqrt{\widetilde{b}_{ii}\widetilde{b}_{jj}}.$$

We now prove the entrywise rate of convergence for the sample covariance $S(B^*)$.

Proof of Proposition S16. By the triangle inequality,

$$\begin{aligned}
\left|S_{ij}(B^*) - b^*_{ij}\right| &\leqslant |S_{ij}(B^*) - E[S_{ij}(B^*)]| + \left|E[S_{ij}(B^*)] - b^*_{ij}\right| \\
&= |\mathcal{B}_{ij}(B^*)| + |\sigma_{ij}(B^*)| \\
&\leqslant \phi_{B,ij},
\end{aligned}$$

where the last step follows from Lemmas S19 and S20. □

Remark. Note that the first term of (S56) is of order $\log^{1/2}(m)/\sqrt{m}$, and the second term is of order $\|B^*\|_1/n_{\min}$.

G.2 Proof of Proposition S17

We express the (i,j)th entry of $S(A^*)$ as a quadratic form in order to apply the Hanson-Wright inequality to obtain an entrywise large deviation bound. Without loss of generality, let $i=1$, $j=2$. The $(1,2)$ entry of $S(A^*)$ can be expressed as a quadratic form, as follows,

$$\begin{aligned} S_{12}(A^*) &= X_1^T(I-P_2)X_2/n \\ &= (1/2)\begin{bmatrix} X_1^T & X_2^T \end{bmatrix}\begin{bmatrix} 0 & (I-P_2) \\ (I-P_2) & 0 \end{bmatrix}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}/n \\ &= (1/2)\begin{bmatrix} X_1^T & X_2^T \end{bmatrix}\left(\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\otimes(I-P_2)\right)\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}/n. \end{aligned}$$

We decorrelate the random vector $(X_1, X_2) \in \mathbb{R}^{2n}$ so that we can apply the Hanson-Wright inequality. The covariance matrix used for decorrelation is

$$\mathrm{Cov}\left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\right) = \begin{bmatrix} a_{11}^* & a_{12}^* \\ a_{21}^* & a_{22}^* \end{bmatrix}\otimes B^* =: A_{\{1,2\}}^*\otimes B^*,$$

with

$$A_{\{1,2\}}^* = \begin{bmatrix} a_{11}^* & a_{12}^* \\ a_{21}^* & a_{22}^* \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

Decorrelating the quadratic form yields

$$S_{12}(A^*) = Z^T\Phi Z,$$

where $Z \in \mathbb{R}^{2n}$, with $E[Z]=0$ and $\mathrm{Cov}(Z) = I_{2n\times 2n}$, and

$$\Phi = (1/2n)\left((A_{\{1,2\}}^*)^{1/2}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}(A_{\{1,2\}}^*)^{1/2}\right)\otimes B^{1/2}(I-P_2)B^{1/2}. \tag{S62}$$

To apply the Hanson-Wright inequality, we first find the trace and Frobenius norm of Φ.

For the trace, note that

$$\operatorname{tr}\left((A^*_{\{1,2\}})^{1/2}\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}(A^*_{\{1,2\}})^{1/2}\right) = \operatorname{tr}\left(\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}A^*_{\{1,2\}}\right) = 2a^*_{12}. \tag{S63}$$

For the Frobenius norm, note that

$$\begin{aligned}\left\|(A^*_{\{1,2\}})^{1/2}\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}(A^*_{\{1,2\}})^{1/2}\right\|_F^2 &= \operatorname{tr}\left(\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}A^*_{\{1,2\}}\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}A^*_{\{1,2\}}\right)\\ &= \operatorname{tr}\left(\begin{bmatrix}a^{*2}_{12}+a^*_{11}a^*_{22} & 2a^*_{12}a^*_{22}\\ 2a^*_{12}a^*_{22} & a^{*2}_{12}+a^*_{11}a^*_{22}\end{bmatrix}\right)\\ &= 2a^{*2}_{12}+2a^*_{11}a^*_{22},\end{aligned}$$

Therefore the trace of Φ is

$$\operatorname{tr}(\Phi) = a^*_{12}\operatorname{tr}\left(\widetilde{B}\right)/n, \tag{S64}$$

and the Frobenius norm of Φ is

$$\|\Phi\|_F = (1/n)\sqrt{a^{*2}_{12}+a^*_{11}a^*_{22}}\|\widetilde{B}\|_F. \tag{S65}$$

Applying the Hanson-Wright inequality yields

$$\begin{aligned}&P\left(|S_{12}(A^*) - a^*_{12}\operatorname{tr}(B^*)/n| > \phi_{A,12}\right)\\ &\leqslant P\left(\left|S_{12}(A^*) - a^*_{12}\operatorname{tr}\left(\widetilde{B}\right)/n\right| + (a^*_{12}/n)\left|\operatorname{tr}\left(\widetilde{B}\right) - \operatorname{tr}(B^*)\right| > \phi_{A,12}\right)\\ &= P\left(\left|S_{12}(A) - a^*_{12}\operatorname{tr}\left(\widetilde{B}\right)/n\right| > d^{1/2}K\log^{1/2}(n\vee m)\|\Phi\|_F\right)\\ &\leqslant 2/(n\vee m)^d.\end{aligned}$$

By the union bound,

$$\begin{aligned} & P\left(\forall i, j\left|S_{ij}(A^*)-a_{ij}\operatorname{tr}\left(B^*\right)/n\right|<\phi_{A,ij}\right) \\ \geqslant\ & 1-\sum_{i=1}^{m}\sum_{j=1}^{m}P\left(\left|S_{ij}(A^*)-a_{ij}\operatorname{tr}\left(B^*\right)/n\right|>\phi_{A,ij}\right) \\ \geqslant\ & 1-2m^2/(n\vee m)^d \geqslant 2/(n\vee m)^{d-2}. \end{aligned}$$

□

G.3 Proof of Lemma S18

For the event (S55) from Proposition S16,

$$\left|S_{ij}(B^*)-b^*_{ij}\right|<\phi_{B,ij}=K^2d\frac{\log^{1/2}(m)}{\sqrt{m}}C_A\sqrt{\widetilde{b}_{ii}\widetilde{b}_{jj}}+\left|b^*_{ij}-\widetilde{b}_{ij}\right|,$$

dividing by $\sqrt{b^*_{ii}b^*_{jj}}$ yields

$$\left|\frac{q_iXX^Tq_j}{\operatorname{tr}(A^*)\sqrt{b^*_{ii}b^*_{jj}}}-\rho_{ij}(B)\right|<K^2dC_A\frac{\log^{1/2}(m)}{\sqrt{m}}\sqrt{\frac{\widetilde{b}_{ii}\widetilde{b}_{jj}}{b^*_{ii}b^*_{jj}}}+\frac{\left|b_{ij}-\widetilde{b}_{ij}\right|}{\sqrt{b^*_{ii}b^*_{jj}}}. \tag{S66}$$

By Lemma S19,

$$\widetilde{b}_{ij}=b_{ij}\left[1+O\left(\frac{\|B\|_1}{n}\right)\right],$$

so the right-hand side of (S66) is less than or equal to $\widetilde{\alpha}$. Hence event (S55) implies $\mathcal{X}_0(B)$. Therefore, we know that $P(\mathcal{X}_0(B))\geqslant 1-2/m^{d-2}$.

Similarly, event (S57) in Proposition S17:

$$\begin{aligned} & \left|S_{ij}(A^*)-a^*_{ij}\operatorname{tr}\left(B^*\right)/n\right|<\phi_{A,ij} \\ & =(a^*_{ij}/n)\left|\operatorname{tr}\left(\widetilde{B}\right)-\operatorname{tr}\left(B\right)\right|+d^{1/2}K\log^{1/2}(n\vee m)(1/n)\sqrt{a^{*2}_{ij}+a^*_{ii}a^*_{jj}}\|\widetilde{B}\|_F, \end{aligned}$$

implies that

$$\begin{aligned}
&\left|\frac{X_j^T(I-P_2)X_t}{\mathrm{tr}(B^*)\sqrt{a^*_{jj}a^*_{tt}}} - \rho_{jt}(A)\right| \\
&< |\rho_{jt}(A)|\frac{\left|\mathrm{tr}\left(\widetilde{B}\right) - \mathrm{tr}\left(B^*\right)\right|}{\mathrm{tr}(B^*)} + d^{1/2}K\log^{1/2}(n\vee m)\sqrt{\rho_{jt}(A)^2+1}\frac{\|\widetilde{B}\|_F}{\mathrm{tr}(B^*)} \\
&= |\rho_{jt}(A)|\frac{\left|\mathrm{tr}\left(\widetilde{B}\right) - \mathrm{tr}\left(B^*\right)\right|}{\mathrm{tr}(B^*)} + d^{1/2}KC_B\frac{\|\widetilde{B}\|_F}{\|B^*\|_F}\sqrt{\rho_{jt}(A)^2+1}\frac{\log^{1/2}(n\vee m)}{\sqrt{n}} \\
&\leqslant \widetilde{\eta},
\end{aligned}$$

which is the event $\mathcal{X}_0(A)$. Therefore, we get that $P(\mathcal{X}_0(A)) \geqslant 1 - 2/(n\vee m)^d$.

We can obtain the $P(\mathcal{X}_0)$ by using a union bound put together $P(\mathcal{X}_0(B))$ and $P(\mathcal{X}_0(A))$, completing the proof. □

G.4 Proof of Lemma S19

Recall that $\widetilde{B} = (I-P_2)B^*(I-P_2)$. The matrix $\widetilde{B} - B^*$ can be expressed as

$$\widetilde{B} - B^* = (I-P_2)B^*(I-P_2) - B^* = -P_2B^* - B^*P_2 + P_2B^*P_2.$$

By the triangle inequality, $\|\widetilde{B} - B^*\|_{\max} \leqslant \|P_2B^*\|_{\max} + \|B^*P_2\|_{\max} + \|P_2B^*P_2\|_{\max}$. We bound each term on the right-hand side.

First we bound $\|P_2B^*\|_{\max}$ and $\|B^*P_2\|_{\max}$. Let p_i denote the ith column of P_2. The (i,j)th entry satisfies

$$|p_i^Tb_j^*| \leqslant \|B^*p_i\|_\infty \leqslant \|B^*\|_\infty\|p_i\|_\infty = \|B^*\|_1\|p_i\|_\infty = \|B^*\|_1/n_{\min},$$

so $\|P_2B^*\|_{\max} \leqslant \|B^*\|_1/n_{\min}$. Because P_2 and B^* are symmetric, $\|P_2B^*\|_{\max} = \|B^*P_2\|_{\max}$.

We now bound $\|P_2B^*P_2\|_{\max}$. Let $B^{1/2}$ denote the symmetric square root of B^*. We can express $p_i^TB^*p_j$ as an inner product $(B^{1/2}p_i)^T(B^{1/2}p_j)$, so

$$\begin{aligned}
|(P_2B^*P_2)_{ij}| = |(B^{1/2}p_i)^T(B^{1/2}p_j)| &\leqslant \left(p_i^TB^*p_i\right)^{1/2}\left(p_j^TB^*p_j\right)^{1/2} &\text{(S67)}\\
&\leqslant \|p_i\|_2\|p_j\|_2\|B\|_2 \leqslant \|B^*\|_2/n_{\min}, &\text{(S68)}
\end{aligned}$$

where (S67) follows from the Cauchy Schwarz inequality, and (S68) holds because

$$\|p_i\|_2 = \begin{cases} 1/\sqrt{n_1} & \text{if } i \in \{1, \ldots, n_1\} \\ 1/\sqrt{n_2} & \text{if } i \in \{n_1+1, \ldots, n\}. \end{cases}$$

□

G.5 Proof of Lemma S20

Let $B^{1/2}$ denote the symmetric square root of B^*. Let $Z_j = (a^*_{jj}B^*)^{-1/2}X_j$. We express $S_{ij}(B^*)$ as a quadratic form in order to use the Hanson-Wright inequality to prove a large deviation bound. That is, we show that $S_{ij}(B^*) = \text{vec}(Z)^T\Phi^{ij}\,\text{vec}(Z)$, with

$$\Phi^{ij} \;=\; (1/m)A^* \otimes B^{1/2}(e_j - p_j)(e_i - p_i)^T B^{1/2}. \tag{S69}$$

We express $S_{ij}(B^*)$ as a quadratic form, as follows:

$$\begin{aligned} S_{ij}(B^*) &= \frac{1}{m}\sum_{k=1}^m (e_i - p_i)^T X_k X_k^T (e_j - p_j) = \frac{1}{m}\sum_{k=1}^m \text{tr}\left[(e_i - p_i)^T X_k X_k^T (e_j - p_j)\right] \\ &= \frac{1}{m}\sum_{k=1}^m X_k^T (e_j - p_j)(e_i - p_i)^T X_k \\ &= \frac{1}{m}\,\text{vec}(X)^T\left(I_{m\times m} \otimes (e_j - p_j)(e_i - p_i)^T\right)\text{vec}(X) \\ &= \text{vec}(Z)^T\Phi^{ij}\,\text{vec}(Z) \end{aligned}$$

where

$$\text{tr}(\Phi^{ij}) \;=\; \text{tr}(B^{1/2}(e_j - p_j)(e_i - p_i)^T B^{1/2}) = (e_i - p_i)^T B^*(e_j - p_j) = \widetilde{b}_{ij}, \tag{S70}$$

$$\begin{aligned} \|\Phi^{ij}\|_F &= \frac{1}{m}\|A^*\|_F \|B^{1/2}(e_j - p_j)(e_i - p_i)^T B^{1/2}\|_F \qquad \text{(S71)} \\ &= \frac{1}{m}\|A^*\|_F \left((e_i - p_i)^T B^*(e_i - p_i)\right)^{1/2}\left((e_j - p_j)^T B^*(e_j - p_j)\right)^{1/2} = \frac{1}{m}\|A^*\|_F\sqrt{\widetilde{b}_{ii}\widetilde{b}_{jj}}. \end{aligned}$$

Therefore, we get that

$$\begin{aligned}
& P\left(\forall i,j \ \left|S_{ij}(B^*) - \widetilde{b}_{ij}\right| \leqslant K^2 d \log^{1/2}(m) \|\Phi^{ij}\|_F / c'\right) \\
&= \ P\left(\forall i,j \ \left|\mathrm{vec}(Z)^T \Phi^{ij} \,\mathrm{vec}(Z) - \mathrm{tr}\left(\Phi^{ij}\right)\right| \leqslant K^2 d \log^{1/2}(m) \|\Phi^{ij}\|_F / c'\right) \\
&\geqslant \ 1 - 2m^2 \exp\left(-c \min\left(d^2 \log(m)/c'^2, \frac{d \log^{1/2}(m) \|\Phi^{ij}\|_F / c'}{\|\Phi^{ij}\|_2}\right)\right) \\
&\geqslant \ 1 - 2/m^{d-2}.
\end{aligned}$$

If the event $\left\{\forall i,j \ \left|S_{ij}(B^*) - \widetilde{b}_{ij}\right| \leqslant K^2 d \log^{1/2}(m) \|\Phi^{ij}\|_F / c'\right\}$ holds, it follows that

$$\left|S_{ij}(B^*) - b^*_{ij}\right| \leqslant \left|S_{ij}(B^*) - \widetilde{b}_{ij}\right| + |b^*_{ij} - \widetilde{b}_{ij}| \leqslant K^2 d \log^{1/2}(m) \|\Phi^{ij}\|_F / c' + |b_{ij} - \widetilde{b}_{ij}|.$$

The Lemma is thus proved. $\square$

H Proof of Theorem 4

H.1 Notation

Notation	Meaning
Mean structure	
$\mu \in \mathbb{R}^m$	Vector of grand means of each gene
$\gamma \in \mathbb{R}^m$	Vector of mean differences for each gene
$\nu = \frac{1}{2}\left[\frac{1}{n_1}1_{n_1}^T \quad \frac{1}{n_2}1_{n_2}^T\right]^T \in \mathbb{R}^n$	Inner product with ν computes global mean
Outcome of model selection step	
$J_0 \subset \{1, 2, \ldots, m\}$	Indices selected for group centering
$J_1 \subset \{1, 2, \ldots, m\}$	Indices selected for global centering
Sizes of gene subsets	
$m_0 = \lvert J_0 \rvert$	Number of group centered genes
$m_1 = \lvert J_1 \rvert$	Number of globally centered genes
Projection matrices	
$P_1 = 1_n \nu^T$	Projection matrix that performs global centering
P_2 (as in (S81))	Projection matrix that performs group centering
Sample covariance matrices	
$S(B, J_0, J_1) = \frac{m_1}{m} S_1(B) + \frac{m_0}{m} S_2(B)$	Model selection sample covariance matrix
$S_1(B, J_1) = \frac{1}{m_1} \sum_{j \in J_1} (I - P_1) X_j X_j^T (I - P_1)$	Globally centered sample covariance matrix
$S_2(B, J_0) = \frac{1}{m_0} \sum_{j \in J_0} (I - P_2) X_j X_j^T (I - P_2)$	Group centered sample covariance matrix
Decomposition of $S(B, J_0, J_1)$	
$S_{\mathrm{I}} = S(B, J_0, J_1) - \mathbb{E}\left[S(B, J_0, J_1)\right]$	Bias
$S_{\mathrm{II}} = \frac{1}{m}(I - P_1) M_{J_1} M_{J_1}^T (I - P_1)$	False negatives (deterministic)
$S_{\mathrm{III}} = \frac{1}{m}(I - P_1) M_{J_1} \varepsilon^T (I - P_1)$	False negatives (random)
$S_{\mathrm{IV}} = m^{-1}(I - P_2)\varepsilon_{J_0}\varepsilon_{J_0}^T(I - P_2) +$ $m^{-1}(I - P_1)\varepsilon_{J_1}\varepsilon_{J_1}^T(I - P_1)$	True negatives

H.2 Two-Group Model and Centering

We begin by introducing some relevant notation for the two-group model and centering. Define the group membership vector $\delta_n \in \mathbb{R}^n$ as

$$\delta_n := \begin{bmatrix} 1_{n_1}^T & -1_{n_2}^T \end{bmatrix}^T \in \mathbb{R}^n. \tag{S72}$$

In the two-group model, the mean matrix M can be expressed as

$$M = 1_n \mu^T + (1/2)\delta_n \gamma^T, \tag{S73}$$

where $\mu \in \mathbb{R}^m$ is a vector of grand means, and $\gamma \in \mathbb{R}^m$ is the vector of mean differences. According to (S73), the (i,j)th entry of M can be expressed as

$$m_{ij} = \begin{cases} \mu_j + \gamma_j/2 & \text{if sample } i \text{ is in group one} \\ \mu_j - \gamma_j/2 & \text{if sample } i \text{ is in group two.} \end{cases} \tag{S74}$$

Define the vector $\nu \in \mathbb{R}^n$ as

$$\nu = \frac{1}{2}\begin{bmatrix} \frac{1}{n_1} 1_{n_1}^T & \frac{1}{n_2} 1_{n_2}^T \end{bmatrix}^T \in \mathbb{R}^n, \tag{S75}$$

so that for the jth column of the data matrix $X_j \in \mathbb{R}^n$,

$$\mathbb{E}\left(\nu^T X_j\right) = \frac{1}{2}\mathbb{E}\left(\frac{1}{n_1}\sum_{k=1}^{n_1} X_{jk} + \frac{1}{n_2}\sum_{k=n_1+1}^{n} X_{jk}\right) = \mu_j. \tag{S76}$$

Note that

$$\nu^T 1_n = (1/2)(1+1) = 1, \quad \text{and} \quad \nu^T \delta_n = (1/2)(1-1) = 0. \tag{S77}$$

Next we define a projection matrix that performs global centering. Define the non-orthogonal projection matrix

$$P_1 := 1_n \nu^T \in \mathbb{R}^{n \times n}. \tag{S78}$$

Applying the projection matrix to the mean matrix yields

$$P_1 M = 1_n \nu^T \left(1_n \mu^T + (1/2)\delta_n \gamma^T\right) = 1_n \mu^T + (1/2)(\nu^T \delta_n) 1_n \gamma^T = 1_n \mu^T, \tag{S79}$$

with residuals

$$(I - P_1)M = M - P_1 M = M - 1_n \mu^T = (1/2)\delta_n \gamma^T. \tag{S80}$$

Define

$$P_2 = \begin{bmatrix} n_1^{-1} 1_{n_1} 1_{n_1}^T & \\ & n_2^{-1} 1_{n_2} 1_{n_2}^T \end{bmatrix}. \tag{S81}$$

Note that $P_2 1_n = 1_n$ and $P_2 \delta_n = \delta_n$, so

$$P_2 M = P_2 1_n \mu^T + (1/2) P_2 \delta_n \gamma^T = 1_n \mu^T + (1/2)\delta_n \gamma^T = M, \tag{S82}$$

and therefore $(I - P_2)M = 0$.

Define

$$\breve{B} = (I - P_1)B(I - P_1) = \left(\breve{b}_{ij}\right) \tag{S83}$$

$$\widetilde{B} = (I - P_2)B(I - P_2) = \left(\widetilde{b}_{ij}\right) \tag{S84}$$

$$\breve{\breve{B}} = (I - P_1)B(I - P_2) = \left(\breve{\breve{b}}_{ij}\right). \tag{S85}$$

Let $\breve{b}_{\max}$, $\widetilde{b}_{\max}$, and $\breve{\breve{b}}_{\max}$ denote the maximum diagonal entries of $\breve{B}$, $\widetilde{B}$, and $\breve{\breve{B}}$, respectively.

H.3 Model Selection Centering

For a subset $J \subset \{1, \ldots, m\}$, let X_J denote the submatrix of X consisting of columns indexed by J. For the fixed sets of genes J_0 and J_1, define the sample covariance

$$S(B, J_0, J_1) = m^{-1} \sum_{k \in J_0} (I - P_2) X_k X_k^T (I - P_2)^T + m^{-1} \sum_{k \in J_1} (I - P_1) X_k X_k^T (I - P_1)^T =: \mathrm{I} + \mathrm{II}\,. \tag{S86}$$

Note that $\mathbb{E}\left[S(B, J_0, J_1)\right] = B^\sharp$, with

$$B^\sharp = \frac{\mathrm{tr}\,(A_{J_0})}{m}(I - P_2)B(I - P_2) + \frac{\mathrm{tr}\,(A_{J_1})}{m}(I - P_1)B(I - P_1). \tag{S87}$$

Define the sample correlation matrix,

$$\widehat{\Gamma}_{ij}(B) = \frac{(S(B, J_0, J_1))_{ij}}{\sqrt{(S(B, J_0, J_1))_{ii}(S(B, J_0, J_1))_{jj}}}. \tag{S88}$$

The baseline Gemini estimators Zhou (2014a) are then defined as follows, using a pair of penalized estimators for the correlation matrices $\rho(A) = (a_{ij}/\sqrt{a_{ii}a_{jj}})$ and $\rho(B) = (b_{ij}/\sqrt{b_{ii}b_{jj}})$:

$$\widehat{A}_\rho = \underset{A_\rho \succ 0}{\arg\min} \left\{ \mathrm{tr}\left(\widehat{\Gamma}(A)A_\rho^{-1}\right) + \log|A_\rho| + \lambda_B |A_\rho^{-1}|_{1,\mathrm{off}} \right\}, \tag{S89a}$$

$$\widehat{B}_\rho = \underset{B_\rho \succ 0}{\arg\min} \left\{ \mathrm{tr}\left(\widehat{\Gamma}(B)B_\rho^{-1}\right) + \log|B_\rho| + \lambda_A |B_\rho^{-1}|_{1,\mathrm{off}} \right\}. \tag{S89b}$$

We will focus on $\widehat{B}_\rho$ using the input as defined in (S88).

The proof proceeds as follows. Lemma S22, the equivalent of Proposition S16 for Algorithm 1, establishes entry-wise convergence rates of the sample covariance matrix for fixed sets of group and globally centered genes. We use this to prove Theorem S21 below in Section H.4 and to prove Theorem 4 in Section H.5.

H.4 Convergence for fixed gene sets

We first state a standalone result, Theorem S21, which provides rates of convergence when $S(B, J_0, J_1)$ as in (S86) is calculated using fixed sets of group centered and globally centered genes, J_0 and J_1, respectively. This result shows how the algorithm used in the preliminary step to choose which genes to group center can be decoupled from the rest of the estimation procedure. The proof is presented below in Section H.4.2.

Theorem S21. *Suppose that (A1), (A2'), and (A3) hold. Let J_0 and J_1 denote sets such that $J_0 \cap J_1 = \varnothing$ and $J_0 \cup J_1 = \{1, \ldots, m\}$. Let $m_0 = |J_0|$ and $m_1 = |J_1|$ denote the sizes of the sets. Let $\tau_{global} > 0$ satisfy*

$$\max_{j \in J_1} |\gamma_j| \leqslant \tau_{global}, \tag{S90}$$

for $\tau_{global} = C\sqrt{\log(m)}\|(D^T B^{-1} D)^{-1}\|_2^{1/2} \asymp \sqrt{\frac{\log(m)}{n}}$.

Consider the data as generated from model (S73) *with $\varepsilon = B^{1/2} Z A^{1/2}$, where $A \in \mathbb{R}^{m \times m}$ and $B \in \mathbb{R}^{n \times n}$ are positive definite matrices, and Z is an $n \times m$ random matrix as defined*

in Theorem 1. Let λ_A denote the penalty parameter for estimating B. Suppose the penalty parameter λ_A in (S89b) *satisfies*

$$\lambda_A \geqslant C'' \left[C_A K \frac{\log^{1/2}(m \vee n)}{\sqrt{m}} + \frac{\|B\|_1}{n_{\min}} \right]. \tag{S91}$$

where C'' is an absolute constant.

(I) *Let $\mathcal{E}_4(J_0, J_1)$ be the event such that*

$$\left\| \mathrm{tr}\,(A) \left(\widehat{W}_2 \widehat{B}_\rho \widehat{W}_2 \right)^{-1} - B^{-1} \right\|_2 \leqslant \frac{C' \lambda_A \sqrt{|B^{-1}|_{0,\mathrm{off}} \vee 1}}{b_{\min} \varphi^2_{\min}(\rho(B))}. \tag{S92}$$

Then $P(\mathcal{E}_4(J_0, J_1)) \geqslant 1 - C/m^d$.

(II) *With probability at least $1 - C'/m^d$, for all j,*

$$\|\widehat{\beta}_j(\widehat{B}^{-1}) - \beta_j^*\|_2 \leqslant C_1 \lambda_A \sqrt{\frac{n_{ratio}\,(|B^{-1}|_{0,off} \vee 1)}{n_{\min}}} + C_2 \sqrt{\log(m)} \|(D^T B^{-1} D)^{-1}\|_2^{1/2}. \tag{S93}$$

H.4.1 Decomposition of sample covariance matrix

The error in the sample covariance $S(B, J_0, J_1)$ can be decomposed as

$$S(B, J_0, J_1) - B = \left[B^\sharp - B \right] + \left[S(B, J_0, J_1) - B^\sharp \right], \tag{S94}$$

where the first term corresponds to bias and the second term to variance. We now further decompose the variance term. The first term of $S(B, J_0, J_1)$ in (S86) can be decomposed as,

$$\begin{aligned}
\mathrm{I} &= m^{-1}(I - P_2) X_{J_0} X_{J_0}^T (I - P_2) \\
&= m^{-1}(I - P_2)(M_{J_0} + \varepsilon_{J_0})(M_{J_0} + \varepsilon_{J_0})^T (I - P_2) \\
&= m^{-1}(I - P_2)\varepsilon_{J_0}\varepsilon_{J_0}^T (I - P_2) + m^{-1}(I - P_2) M_{J_0} \varepsilon_{J_0}^T (I - P_2) \\
&\quad + m^{-1}(I - P_2)\varepsilon_{J_0} M_{J_0}^T (I - P_2) + m^{-1}(I - P_2) M_{J_0} M_{J_0}^T (I - P_2),
\end{aligned} \tag{S95}$$

and the second term can be decomposed analogously, as

$$\begin{aligned}\mathrm{II} = {} & m^{-1}(I-P_1)\varepsilon_{J_1}\varepsilon_{J_1}^T(I-P_1) + m^{-1}(I-P_1)M_{J_1}\varepsilon_{J_1}^T(I-P_1) \\ & + m^{-1}(I-P_1)\varepsilon_{J_1}M_{J_1}^T(I-P_1) + m^{-1}(I-P_1)M_{J_1}M_{J_1}^T(I-P_1). \end{aligned} \tag{S96}$$

By the above decompositions, it follows that $S(B, J_0, J_1)$ can be expressed as

$$S(B, J_0, J_1) = S_{\mathrm{II}} + S_{\mathrm{III}} + S_{\mathrm{III}}^T + S_{\mathrm{IV}}, \tag{S97}$$

with

$$S_{\mathrm{II}} = m^{-1}(I-P_2)M_{J_0}M_{J_0}^T(I-P_2) + m^{-1}(I-P_1)M_{J_1}M_{J_1}^T(I-P_1). \tag{S98}$$

$$S_{\mathrm{III}} = m^{-1}(I-P_2)M_{J_0}\varepsilon_{J_0}^T(I-P_2) + m^{-1}(I-P_1)M_{J_1}\varepsilon_{J_1}^T(I-P_1) \tag{S99}$$

$$S_{\mathrm{IV}} = m^{-1}(I-P_2)\varepsilon_{J_0}\varepsilon_{J_0}^T(I-P_2) + m^{-1}(I-P_1)\varepsilon_{J_1}\varepsilon_{J_1}^T(I-P_1) \tag{S100}$$

For each of S_{II}, S_{III}, and S_{IV}, the first term comes from (S96) and the second term comes from (S97).

The terms S_{II} and S_{III} can be simplified, as follows. Because $(I-P_2)M_{J_0} = 0$, it follows that the first term of S_{II} is zero:

$$m^{-1}(I-P_2)M_{J_0}M_{J_0}^T(I-P_2) = 0.$$

and the first term of S_{III} is also zero,

$$m^{-1}(I-P_2)M_{J_0}\varepsilon_{J_0}^T(I-P_2) = 0,$$

Therefore the terms S_{II} and S_{III} are equal to

$$S_{\mathrm{II}} = m^{-1}(I-P_1)M_{J_1}M_{J_1}^T(I-P_1), \tag{S101}$$

$$S_{\mathrm{III}} = m^{-1}(I-P_1)M_{J_1}\varepsilon_{J_1}^T(I-P_1). \tag{S102}$$

Let $S_{\mathrm{I}} = B^{\sharp} - B$. We have thus decomposed the error in the sample covariance as

$$S(B, J_0, J_1) - B = \underbrace{S_{\mathrm{I}}}_{\text{bias}} + \underbrace{\left[\left(S_{\mathrm{IV}} - B^{\sharp}\right) + S_{\mathrm{III}} + S_{\mathrm{II}}\right]}_{\text{variance}}. \tag{S103}$$

In Lemma S23, we provide an error bound for each term in the decomposition (S104).

We next state Lemma S22, which establishes the maximum of entry-wise errors for estimating B using the sample covariance for fixed gene sets as defined in (S104). Lemma S22 is used in the proof of Theorem S21. Following, we state Lemma S23, which is used in the proof of Lemma S22.

Lemma S22. *Suppose the conditions of Theorem S21 hold. Let $\mathcal{E}_6(J_0, J_1)$ denote the event*

$$\mathcal{E}_6(J_0, J_1) = \left\{ \left\| S(B, J_0, J_1) - B \right\|_{\infty} \leqslant C_A K \frac{\log^{1/2}(m \vee n)}{\sqrt{m}} + \frac{\|B\|_1}{n_{\min}} \right\}. \tag{S104}$$

Then $\mathcal{E}_6(J_0, J_1)$ holds with probability at least $1 - \frac{8}{(m \vee n)^2}$.

Lemma S23. *Let the model selection-based sample covariance $S(B, J_0, J_1)$ be as defined in* (S86)*, where J_1 and J_0 are fixed sets of variables that are globally centered, and group centered, respectively. Let $m_0 = |J_0|$ and $m_1 = |J_1|$. Define the rates*

$$r_1 = \frac{3 \|B\|_1}{n_{\min}}, \tag{S105}$$

$$r_2 = (4m)^{-1} \left\|\gamma_{J_1}\right\|_2^2, \tag{S106}$$

$$r_3 = C_3 d^{1/2} K^2 \log^{1/2}(m) m^{-1} \left(\gamma_{J_1}^T A_{J_1} \gamma_{J_1}\right)^{1/2} \breve{b}_{\max}^{1/2}, \tag{S107}$$

$$r_4 = C_4 d^{1/2} K \log^{1/2}(m) m^{-1} \left\|A\right\|_F \left\|B\right\|_2. \tag{S108}$$

(I) *Deterministically,*

$$\left\|B^{\sharp} - B\right\|_{\infty} \leqslant r_1 \quad \text{and} \quad \left\|S_{\mathrm{II}}\right\|_{\infty} \leqslant r_2. \tag{S109}$$

(II) *Define the events*

$$\mathcal{E}_{\mathrm{I}} = \left\{ \left\|S_{\mathrm{IV}} - B^{\sharp}\right\|_{\infty} \leqslant r_4 \right\} \quad \text{and} \quad \mathcal{E}_{\mathrm{II}} = \left\{ \left\|S_{\mathrm{III}}\right\|_{\infty} \leqslant r_3 \right\}. \tag{S110}$$

Then $\mathcal{E}_{\mathrm{I}}$ and $\mathcal{E}_{\mathrm{II}}$ occur with probability at least $1 - 2/m^d$.

Lemmas S22 and S23 are proved in Section I. We analyze term S_{I} in Section I.2, term S_{II} in Section I.3, term S_{III} in Section I.4, and term S_{IV} in Section I.5.

H.4.2 Proof of Theorem S21

Let us first define the event $\mathcal{E}_{\mathrm{global}}$, that is, the GLS error based on the true B^{-1} is small:

$$\mathcal{E}_{\mathrm{global}} = \left\{ \left\| \widehat{\gamma}(B^{-1}) - \gamma \right\|_\infty < \sqrt{\log(m)} \|(D^T B^{-1} D)^{-1}\|_2^{1/2} \right\}. \tag{S111}$$

Let $\mathcal{E}_4(J_0, J_1)$ be defined as in (S93), denoting small operator norm error in estimating B^{-1}:

$$\mathcal{E}_4(J_0, J_1) = \left\{ \left\| \mathrm{tr}\,(A) \left(\widehat{W}_2 \widehat{B}_\rho \widehat{W}_2 \right)^{-1} - B^{-1} \right\|_2 \leqslant \frac{C' \lambda_A \sqrt{|B^{-1}|_{0,\mathrm{off}} \vee 1}}{b_{\min} \varphi^2_{\min}(\rho(B))} \right\}. \tag{S112}$$

Note that $\mathcal{E}_4(J_0, J_1)$ holds deterministically under event $\mathcal{E}_6(J_0, J_1)$ as defined in (S105) of Lemma S22.

Define the event bounding the perturbation in mean estimation due to error in estimating B^{-1}:

$$\mathcal{E}_5(J_0, J_1) = \left\{ \left\| \widehat{\gamma}(\widehat{B}^{-1}) - \widehat{\gamma}(B^{-1}) \right\|_\infty < C n_{\min}^{-1/2} \left\| \widehat{B}^{-1} - B^{-1} \right\|_2 \right\}. \tag{S113}$$

Conditional on a fixed matrix $\widehat{B}^{-1}$ that satisfies $\mathcal{E}_4(J_0, J_1)$, event $\mathcal{E}_5(J_0, J_1)$ holds with probability at least $1 - C/m^d$, by Lemma S6 (used in the proof of Theorem 1).

The overall rate of convergence follows by applying the union bound to the events $\mathcal{E}_{\mathrm{global}} \cap \mathcal{E}_4(J_0, J_1) \cap \mathcal{E}_5(J_0, J_1)$, as follows:

$$\begin{aligned}
&P(\mathcal{E}^c_{\mathrm{global}} \cup \mathcal{E}_4(J_0, J_1)^c \cup \mathcal{E}_5(J_0, J_1)^c) \\
&\leqslant P(\mathcal{E}^c_{\mathrm{global}}) + P(\mathcal{E}_4(J_0, J_1)^c) + P(\mathcal{E}_5(J_0, J_1)^c \mid \mathcal{E}_4(J_0, J_1)) P(\mathcal{E}_4(J_0, J_1)) \\
&\quad + P(\mathcal{E}_5(J_0, J_1)^c \mid \mathcal{E}_4(J_0, J_1)^c) P(\mathcal{E}_4(J_0, J_1)^c) \\
&\leqslant P(\mathcal{E}^c_{\mathrm{global}}) + P(\mathcal{E}_4(J_0, J_1)^c) + P(\mathcal{E}_4(J_0, J_1)^c) + P(\mathcal{E}_5(J_0, J_1)^c \mid \mathcal{E}_4(J_0, J_1)) \\
&= P(\mathcal{E}^c_{\mathrm{global}}) + 2P(\mathcal{E}_4(J_0, J_1)^c) + P(\mathcal{E}_5(J_0, J_1)^c \mid \mathcal{E}_4(J_0, J_1)),
\end{aligned}$$

where $P(\mathcal{E}^c_{\text{global}})$ and $P(\mathcal{E}_5(J_0, J_1)^c \mid \mathcal{E}_4(J_0, J_1))$ are bounded in Theorem 1, and $P(\mathcal{E}_4(J_0, J_1)^c)$ has high probability under Lemma S22.

H.5 Proof of Theorem 4

Let $\widehat{\gamma}^{\text{init}}$ denote the output from Algorithm 1. By our choice of the threshold parameter τ_{init} as in (16), that is,

$$\tau_{\text{init}} = C\left(\frac{\log^{1/2}(m)}{\sqrt{m}} + \frac{\|B\|_1}{n_{\min}}\right)\sqrt{\frac{n_{\text{ratio}}\left(|B^{-1}|_{0,\text{off}} \vee 1\right)}{n_{\min}}} + C\sqrt{\log(m)}\|(D^T B^{-1} D)^{-1}\|_2^{1/2},$$

we have a partition $(\widetilde{J}_0, \widetilde{J}_1)$ such that $\widetilde{J}_0$ is the set of variables selected for group centering and $\widetilde{J}_1$ is the set of variables selected for global centering. The partition results in a sample covariance matrix $S(B, \widetilde{J}_0, \widetilde{J}_1)$ as defined in (S86). Define the event that the Algorithm 1 estimate $\widehat{\gamma}^{\text{init}}$ is close to γ in the sense that

$$\mathcal{E}_{A1} = \left\{\left\|\widehat{\gamma}^{\text{init}} - \gamma\right\|_\infty < \tau_{\text{init}}\right\}. \tag{S114}$$

Note that the event $\mathcal{E}_{A1}$ implies that the false negatives have small true mean differences. That is, on event $\mathcal{E}_{A1}$, by the triangle inequality,

$$\left\|\gamma_{\widetilde{J}_1}\right\|_\infty \leqslant \left\|\gamma_{\widetilde{J}_1} - \widehat{\gamma}^{\text{init}}_{\widetilde{J}_1}\right\|_\infty + \left\|\widehat{\gamma}^{\text{init}}_{\widetilde{J}_1}\right\|_\infty \leqslant \tau_{\text{init}} + \tau_{\text{init}} = 2\tau_{\text{init}}, \tag{S115}$$

where $\left\|\widehat{\gamma}^{\text{init}}_{\widetilde{J}_1}\right\|_\infty < \tau_{\text{init}}$ by definition of $\mathcal{E}_{A1}$, and $\left\|\gamma_{\widetilde{J}_1} - \widehat{\gamma}^{\text{init}}_{\widetilde{J}_1}\right\|_\infty < \tau_{\text{init}}$ by definition of the thresholding set $\widetilde{J}_1$.

Under the assumptions of Theorem S21, $\tau_{\text{init}} \leqslant \tau_{\text{global}}$ with τ_{global} as defined in (S90), so condition (S90) of Theorem S21 is satisfied. Under the conditions of Theorem S21, event $\mathcal{E}_6(J_0, J_1)$ as defined in Lemma S22 holds with high probability; that is, the entrywise error in the sample covariance matrix is small.

Let $\mathcal{E}_B$ denote event (28) in Theorem 4. In view of Theorem S9 and Lemma S10, event

$\mathcal{E}_B$ holds on $\mathcal{E}_6(J_0, J_1)$. Hence

$$\begin{aligned} P\left(\mathcal{E}_B^c\right) &= P\left(\mathcal{E}_6(J_0,J_1)^c \mid \mathcal{E}_{A1}\right) P\left(\mathcal{E}_{A1}\right) + P\left(\mathcal{E}_6(J_0,J_1)^c \mid \mathcal{E}_{A1}^c\right) P\left(\mathcal{E}_{A1}^c\right) \\ &\leqslant P\left(\mathcal{E}_6(J_0,J_1)^c \mid \mathcal{E}_{A1}\right) + P\left(\mathcal{E}_{A1}^c\right) \\ &\leqslant 2/m^d + 2/m^d, \end{aligned}$$

where the first term is bounded in Lemma S22 and the second in Theorem 3.

Recall the event $\mathcal{E}_{\text{global}}$ as defined in (S112). Event (29) in Theorem 4 holds under the intersection of events $\mathcal{E}_{\text{global}} \cap \mathcal{E}_5(\widetilde{J}_0, \widetilde{J}_1) \cap \mathcal{E}_B \cap \mathcal{E}_{A1}$. Hence the probability of (29) can be bounded as follows:

$$\begin{aligned} &P(\mathcal{E}_{\text{global}}^c \cup \mathcal{E}_5(\widetilde{J}_0, \widetilde{J}_1)^c \cup \mathcal{E}_B^c \cup \mathcal{E}_{A1}^c) \\ &\leqslant P(\mathcal{E}_{\text{global}}^c) + P(\mathcal{E}_B^c) + P(\mathcal{E}_5(\widetilde{J}_0, \widetilde{J}_1)^c \mid \mathcal{E}_B) P(\mathcal{E}_B) \\ &\quad + P(\mathcal{E}_5(\widetilde{J}_0, \widetilde{J}_1)^c \mid \mathcal{E}_B^c) P(\mathcal{E}_B^c) + P\left(\mathcal{E}_{A1}^c\right) \\ &\leqslant P(\mathcal{E}_{\text{global}}^c) + P(\mathcal{E}_B^c) + P(\mathcal{E}_B^c) + P(\mathcal{E}_5(\widetilde{J}_0, \widetilde{J}_1)^c \mid \mathcal{E}_B) + P\left(\mathcal{E}_{A1}^c\right) \\ &= P(\mathcal{E}_{\text{global}}^c) + 2P(\mathcal{E}_B^c) + P(\mathcal{E}_5(\widetilde{J}_0, \widetilde{J}_1)^c \mid \mathcal{E}_B) + P\left(\mathcal{E}_{A1}^c\right), \end{aligned}$$

where $P(\mathcal{E}_{\text{global}}^c)$ and $P(\mathcal{E}_5(\widetilde{J}_0, \widetilde{J}_1)^c \mid \mathcal{E}_B)$ are bounded in Theorem 1, $P(\mathcal{E}_B^c)$ is bounded above, and $P\left(\mathcal{E}_{A1}^c\right)$ is bounded in Theorem 3.

I Proof of Lemmas S22 and S23

We first prove Lemma S22 in Section I.1. The rest of the section contains the proof of Lemma S23, where part I is proved in Sections I.2 and I.3 and part II in Sections I.4 and I.5.

I.1 Proof of Lemma S22

The entrywise error in the sample covariance matrix (S86) can be decomposed as

$$\begin{aligned} \|S(B, J_0, J_1) - B\|_\infty &\leqslant \left\|S(B, J_0, J_1) - B^\sharp\right\|_\infty + \left\|B^\sharp - B\right\|_\infty && \text{(S116)} \\ &\leqslant \left\|S_{\text{IV}} - B^\sharp\right\|_\infty + 2\left\|S_{\text{III}}\right\|_\infty + \left\|S_{\text{II}}\right\|_\infty + \left\|B^\sharp - B\right\|_\infty. && \text{(S117)} \end{aligned}$$

Let $r_{n,m} = r_1 + r_2 + 2r_3 + r_4$. By parts I and II of Lemma S23,

$$
\begin{aligned}
&P\left(\left\|S(B, J_0, J_1) - B\right\|_\infty \geqslant r_{n,m}\right) \\
&\leqslant P\left(\left\|S_{\mathrm{IV}} - B^\sharp\right\|_\infty + 2\left\|S_{\mathrm{III}}\right\|_\infty + \left\|S_{\mathrm{II}}\right\|_\infty + \left\|B^\sharp - B\right\|_\infty \geqslant r_{n,m}\right) \quad \text{(by (S118))} \\
&\leqslant P\left(\left\|S_{\mathrm{IV}} - B^\sharp\right\|_\infty + 2\left\|S_{\mathrm{III}}\right\|_\infty + r_2 + r_1 \geqslant r_{n,m}\right) \quad \text{(by (S110))} \\
&= P\left(\left\|S_{\mathrm{IV}} - B^\sharp\right\|_\infty + 2\left\|S_{\mathrm{III}}\right\|_\infty \geqslant r_4 + 2r_3\right) \\
&\leqslant P\left(\left\|S_{\mathrm{IV}} - B^\sharp\right\|_\infty \geqslant r_4\right) + P\left(2\left\|S_{\mathrm{III}}\right\|_\infty \geqslant 2r_3\right) \quad \text{(by (S111))} \\
&\leqslant \frac{2}{m^d} + \frac{2}{m^d} = \frac{4}{m^d}.
\end{aligned}
$$

We show that under the assumptions of Theorem S21, the entrywise error in terms S_{II} and S_{III} is $O\left(C_A\sqrt{\frac{\log(m)}{m}}\right)$. Recall that the entrywise rates of convergence of S_{II} and S_{III} are stated in equations (S107) and (S108), respectively. Let $s = |\mathrm{supp}(\gamma)|$ denote the sparsity of γ. Let $m_{01} = |\mathrm{supp}\left(\gamma_{J_1}\right)|$ denote the number of false negatives.

First, we express the entrywise rate of convergence of S_{II} in terms of τ_{global}. By (S90), $\left\|\gamma_{J_1}\right\|_\infty \leqslant \tau_{\mathrm{global}}$, which implies that $\left\|\gamma_{J_1}\right\|_2^2 \leqslant m_{01}\tau_{\mathrm{global}}^2 \leqslant s\tau_{\mathrm{global}}^2$, where the last inequality holds because $m_{01} \leqslant s$ by definition. Therefore,

$$
r_2 = (4m)^{-1}\left\|\gamma_{J_1}\right\|_2^2 \leqslant \frac{s\tau_{\mathrm{global}}^2}{4m} \leqslant C\frac{s\log(m)}{4nm}\left\|B\right\|_2, \tag{S118}
$$

where the last step holds because $\tau_{\mathrm{global}} = C\sqrt{\log(m)}\|(D^TB^{-1}D)^{-1}\|_2^{1/2} \asymp \sqrt{\frac{\log(m)}{n}}\left\|B\right\|_2^{1/2}$ by assumption. Applying (A3) to the right-hand side of (S119) implies that $r_2 = O\left(C_A\sqrt{\frac{\log(m)}{m}}\right)$.

Next, consider term S_{III}. First note that

$$
\gamma_{J_1}^T A_{J_1}\gamma_{J_1} \leqslant \left\|\gamma_{J_1}\right\|_2^2\left\|A_{J_1}\right\|_2 \leqslant m_{01}\tau_{\mathrm{global}}^2\left\|A_{J_1}\right\|_2, \tag{S119}
$$

where the last inequality holds by (S90). This implies that r_3 is on the order

$$
\begin{aligned}
\frac{\log^{1/2}(m)}{m}\left(\breve{b}_{\max}\gamma_{J_1}^T A_{J_1}\gamma_{J_1}\right)^{1/2} &\leqslant \breve{b}_{\max}^{1/2}\left\|A_{J_1}\right\|_2^{1/2}\left(\frac{\log^{1/2}(m)m_{01}^{1/2}}{m}\right)\tau_{\mathrm{global}} \\
&\leqslant C\frac{\log(m)}{\sqrt{n}}\frac{\sqrt{s}}{m}\left\|A_{J_1}\right\|_2^{1/2}\left\|B\right\|_2^{1/2}\breve{b}_{\max}^{1/2},
\end{aligned} \tag{S120}
$$

where the last inequality holds because $m_{01} \leqslant s \leqslant m$ and $\tau_{\text{global}} \asymp \sqrt{\frac{\log(m)}{n}} \|B\|_2^{1/2}$. Under (A2'), the right-hand side of (S121) satisfies

$$\frac{\log(m)}{\sqrt{n}} \frac{\sqrt{s}}{m} \|A_{J_1}\|_2^{1/2} \|B\|_2^{1/2} \breve{b}_{\max}^{1/2} \leqslant \sqrt{\log(m)} \frac{\sqrt{s}}{m} C_A \frac{\|A_{J_1}\|_2^{1/2}}{\|A\|_2^{1/2}} \leqslant C_A \sqrt{\frac{\log(m)}{m}}, \tag{S121}$$

where the last inequality holds because $s \leqslant m$.

I.2 Proof of part I of Lemma S23, term I

We bound the entrywise bias,

$$\begin{aligned} \left\|B^\sharp - B\right\|_{\max} &= \left\| \frac{\operatorname{tr}(A_{J_0})}{m} \widetilde{B} + \frac{\operatorname{tr}(A_{J_1})}{m} \breve{B} - B \right\|_{\max} \\ &\leqslant \frac{\operatorname{tr}(A_{J_0})}{m} \left\| \widetilde{B} - B \right\|_{\max} + \frac{\operatorname{tr}(A_{J_1})}{m} \left\| \breve{B} - B \right\|_{\max}. \end{aligned} \tag{S122}$$

Note that

$$\begin{aligned} \left\| \breve{B} - B \right\|_{\max} &= \|(I - P_1)B(I - P_1) - B\|_{\max} = \|P_1 B P_1 - P_1 B - B P_1\|_{\max} \\ &\leqslant \|P_1 B P_1\|_{\max} + \|P_1 B\|_{\max} + \|B P_1\|_{\max}. \end{aligned} \tag{S123}$$

We bound the first term of (S124) as follows:

$$\left| (P_1 B P_1)_{ij} \right| \leqslant \left\| p_i^{(1)} \right\|_2 \left\| p_j^{(1)} \right\|_2 \|B\|_2 \leqslant \frac{\|B\|_2}{n_{\min}}.$$

For the second term of (S124),

$$(P_1 B)_{ij} = \left| b_i^T p_j^{(1)} \right| \leqslant \|b_i\|_1 \left\| p_j^{(1)} \right\|_\infty \leqslant \|B\|_1 \left\| p_j^{(1)} \right\|_\infty \leqslant \frac{\|B\|_1}{n_{\min}},$$

where $\left\| p_j^{(1)} \right\|_\infty \leqslant \frac{1}{n_{\min}}$ by the definition of P_1 in (S78). We have shown $\|B P_1\|_{\max} \leqslant \frac{\|B\|_1}{n_{\min}}$. Likewise, $\|B P_1\|_{\max} \leqslant \frac{\|B\|_1}{n_{\min}}$. Therefore,

$$\left\| \breve{B} - B \right\|_{\max} \leqslant 3 \frac{\|B\|_1}{n_{\min}}. \tag{S124}$$

Because the projection matrix P_2 satisfies $\left\|p_j^{(2)}\right\|_\infty \leqslant \frac{1}{n_{\min}}$, an analogous proof shows that

$$\left\|\widetilde{B} - B\right\|_{\max} \leqslant \frac{3\left\|B\right\|_1}{n_{\min}}. \tag{S125}$$

Substituting (S125) and (S126) into (S123) yields

$$\begin{aligned}
\left\|B^\sharp - B\right\|_{\max} &\leqslant \frac{\operatorname{tr}(A_{J_0})}{m}\left\|\check{B} - B\right\|_{\max} + \frac{\operatorname{tr}(A_{J_1})}{m}\left\|\widetilde{B} - B\right\|_{\max} \\
&\leqslant \left(\frac{\operatorname{tr}(A_{J_0})}{m} + \frac{\operatorname{tr}(A_{J_1})}{m}\right)\frac{3\left\|B\right\|_1}{n_{\min}} \\
&= \frac{\operatorname{tr}(A)}{m}\frac{3\left\|B\right\|_1}{n_{\min}} \\
&= \frac{3\left\|B\right\|_1}{n_{\min}}.
\end{aligned} \tag{S126}$$

I.3 Proof of part I of Lemma S23, term II

In this section we prove a deterministic entrywise bound on S_{II}. By (S80), it follows that

$$(I - P_1)M_{J_1}M_{J_1}^T(I - P_1) = (1/4)\left\|\gamma_{J_1}\right\|_2^2 \delta_n\delta_n^T,$$

which implies

$$\left\|(I - P_1)M_{J_1}M_{J_1}^T(I - P_1)\right\|_\infty = \left\|(1/4)\left\|\gamma_{J_1}\right\|_2^2 \delta_n\delta_n^T\right\|_\infty = (1/4)\left\|\gamma_{J_1}\right\|_2^2.$$

Therefore S_{II} satisfies the maximum entrywise bound

$$\left\|S_{\mathrm{II}}\right\|_\infty = \left\|m^{-1}(I - P_1)M_{J_1}M_{J_1}^T(I - P_1)\right\|_\infty = \left\|(4m)^{-1}\left\|\gamma_{J_1}\right\|_2^2 \delta_n\delta_n^T\right\|_\infty = (4m)^{-1}\left\|\gamma_{J_1}\right\|_2^2,$$

so

$$\left\|S_{\mathrm{II}}\right\|_\infty = r_2.$$

Note that if J_1 is chosen so that $\left\|\gamma_{J_1}\right\|_\infty \leqslant \tau$, then $\left\|\gamma_{J_1}\right\|_2^2 \leqslant m_{01}\tau^2$, where m_{01} is the number of false negatives, so

$$\frac{\left\|\gamma_1\right\|_2^2}{4m} \leqslant \frac{m_{01}}{4m}\tau^2 \leqslant \frac{\tau^2}{4}. \tag{S127}$$

which implies that the entrywise rate of convergence of S_{II} is $O(\tau^2)$.

I.4 Proof of part II of Lemma S23, term III

Let p_i denote the ith column of P_1^T, for $i = 1, \ldots, n$. Let m_k denote the kth column of M. Let ε_k denote the kth column of ε. The term S_{III} can be expressed as

$$\begin{aligned}
(S_{\mathrm{III}})_{ij} &= m^{-1}(e_i - p_i)^T M_{J_1} \varepsilon_{J_1}^T (e_j - p_j) \\
&= m^{-1} \operatorname{tr}\left(\varepsilon_{J_1}^T (e_j - p_j)(e_i - p_i)^T M_{J_1}\right) \\
&= m^{-1} \sum_{k \in J_1} \varepsilon_k^T (e_j - p_j)(e_i - p_i)^T m_k \\
&= m^{-1} \operatorname{vec}\{\varepsilon_{J_1}\}^T \left(I_{m_1} \otimes (e_j - p_j)(e_i - p_i)^T\right) \operatorname{vec}\{M_{J_1}\} \\
&= m^{-1} \operatorname{vec}\{Z\}^T \left(A_{J_1}^{1/2} \otimes B^{1/2}(e_j - p_j)(e_i - p_i)^T\right) \operatorname{vec}\{M_{J_1}\} \\
&= \operatorname{vec}\{Z\}^T \psi_{ij},
\end{aligned}$$

where

$$\psi_{ij} := m^{-1} \left(A_{J_1}^{1/2} \otimes B^{1/2}(e_j - p_j)(e_i - p_i)^T\right) \operatorname{vec}\{M_{J_1}\}. \tag{S128}$$

The squared Euclidean norm of ψ_{ij} is

$$\begin{aligned}
\|\psi_{ij}\|_2^2 &= \operatorname{vec}\{M_{J_1}\}^T \left(A_{J_1} \otimes (e_i - p_i)(e_j - p_j)^T B (e_j - p_j)(e_i - p_i)^T\right) \operatorname{vec}\{M_{J_1}\} / m^2 \\
&= \operatorname{vec}\{M_{J_1}\}^T \left(A_{J_1} \otimes \breve{b}_{jj}(e_i - p_i)(e_i - p_i)^T\right) \operatorname{vec}\{M_{J_1}\} / m^2 \\
&= \breve{b}_{jj} \sum_{k \in J_1} \sum_{\ell \in J_1} a_{k\ell} m_k^T (e_i - p_i)(e_i - p_i)^T m_\ell / m^2 \\
&= \breve{b}_{jj} \sum_{k \in J_1} \sum_{\ell \in J_1} a_{k\ell} (\delta_n)_i \gamma_k (\delta_n)_i \gamma_\ell / \left(4m^2\right) \\
&= \breve{b}_{jj} \sum_{k \in J_1} \sum_{\ell \in J_1} a_{k\ell} \gamma_k \gamma_\ell / \left(4m^2\right) \\
&= \breve{b}_{jj} \gamma_{J_1}^T A_{J_1} \gamma_{J_1} / \left(4m^2\right).
\end{aligned} \tag{S129}$$

By the Hanson-Wright inequality (Theorem 2.1),

$$\mathbb{P}\left(\left|\operatorname{vec}\{Z\}^T \psi_{ij} - \|\psi_{ij}\|_2\right| > d^{1/2} K^2 \sqrt{\log(m)}\, \|\psi_{ij}\|_2\right) \leqslant 2\exp\{-d\log(m)\} = 2/m^d. \tag{S130}$$

Therefore

$$\begin{aligned}\mathbb{P}\left(|(S_{\mathrm{III}})_{ij}| > \left(1 + d^{1/2}K^2\sqrt{\log(m)}\right)\|\psi_{ij}\|_2\right) &= \mathbb{P}\left(\left|\mathrm{vec}\,\{Z\}^T\psi_{ij}\right| > \|\psi_{ij}\|_2 + d^{1/2}K^2\sqrt{\log(m)}\,\|\psi_{ij}\|_2\right)\\ &\leqslant \mathbb{P}\left(\left|\mathrm{vec}\,\{Z\}^T\psi_{ij} - \|\psi_{ij}\|_2\right| > d^{1/2}K^2\sqrt{\log(m)}\,\|\psi_{ij}\|_2\right)\\ &\leqslant 2/m^d,\end{aligned}$$

where the last step follows from (S131). By (S130), it follows that

$$\left(1 + d^{1/2}K^2\sqrt{\log(m)}\right)\|\psi_{ij}\|_2 \leqslant r_3, \tag{S131}$$

so

$$\mathbb{P}\left(|(S_{\mathrm{III}})_{ij}| > r_3\right) \leqslant \mathbb{P}\left(|(S_{\mathrm{III}})_{ij}| > \left(1 + d^{1/2}K^2\sqrt{\log(m)}\right)\|\psi_{ij}\|_2\right) \leqslant 2/m^d, \tag{S132}$$

by (S132). By the union bound,

$$\mathbb{P}\left(\|S_{\mathrm{III}}\|_\infty > r_3\right) \leqslant \sum_{i=1}^m\sum_{j=1}^m \mathbb{P}\left(|(S_{\mathrm{III}})_{ij}| > r_3\right) \leqslant 2/m^{d-2}.$$

I.5 Proof of part II of Lemma S23, term IV

We now analyze term S_{IV}. To do so, we express S_{IV} as a quadratic form in order to apply the Hanson-Wright inequality.

Let $p_i^{(1)}$ denote the ith column of P_1^T. Let $p_i^{(2)}$ denote the ith column of P_2^T. Define

$$H^{ij}_{\mathrm{group}} = I_{m_0} \otimes \left(e_j - p_j^{(2)}\right)\left(e_j - p_j^{(2)}\right)^T \qquad \text{and} \qquad H^{ij}_{\mathrm{global}} = I_{m_1} \otimes \left(e_j - p_j^{(1)}\right)\left(e_j - p_j^{(1)}\right)^T, \tag{S133}$$

and let

$$H^{ij}(J_0, J_1) = \begin{bmatrix} H^{ij}_{\mathrm{group}} & \\ & H^{ij}_{\mathrm{global}} \end{bmatrix}, \tag{S134}$$

where $H^{ij}_{\mathrm{group}} \in \mathbb{R}^{m_0 n \times m_0 n}$, $H^{ij}_{\mathrm{global}} \in \mathbb{R}^{m_1 n \times m_1 n}$, and $H^{ij}(J_0, J_1) \in \mathbb{R}^{mn \times mn}$. Recall that

$$S_{\mathrm{IV}} = m^{-1}(I - P_2)\varepsilon_{J_0}\varepsilon_{J_0}^T(I - P_2) + m^{-1}(I - P_1)\varepsilon_{J_1}\varepsilon_{J_1}^T(I - P_1).$$

The second term of S_{IV} can be expressed as a quadratic form, as follows (where ε_k denotes the kth column of $\varepsilon \in \mathbb{R}^{n\times m}$):

$$\begin{aligned} m^{-1}(I-P_1)\varepsilon_{J_1}\varepsilon_{J_1}^T(I-P_1) &= m^{-1}\sum_{k\in J_1}\left(e_i-p_i^{(1)}\right)^T\varepsilon_k\varepsilon_k^T\left(e_j-p_j^{(1)}\right) \\ &= m^{-1}\sum_{k\in J_1}\operatorname{tr}\left(\left(e_i-p_i^{(1)}\right)^T\varepsilon_k\varepsilon_k^T\left(e_j-p_j^{(1)}\right)\right) \\ &= m^{-1}\sum_{k\in J_1}\varepsilon_k^T\left(e_j-p_j^{(1)}\right)\left(e_i-p_i^{(1)}\right)^T\varepsilon_k \\ &= m^{-1}\operatorname{vec}\left\{\varepsilon_{J_1}\right\}^T\left(I_{m_1}\otimes\left(e_j-p_j^{(1)}\right)\left(e_i-p_i^{(1)}\right)^T\right)\operatorname{vec}\left\{\varepsilon_{J_1}\right\}^T \\ &= m^{-1}\operatorname{vec}\left\{\varepsilon_{J_1}\right\}^T H^{ij}_{\text{global}}\operatorname{vec}\left\{\varepsilon_{J_1}\right\}^T. \end{aligned} \tag{S135}$$

Analogously, the first term of S_{IV} can be expressed as a quadratic form:

$$\begin{aligned} m^{-1}(I-P_2)\varepsilon_{J_0}\varepsilon_{J_0}^T(I-P_2) &= m^{-1}\sum_{k\in J_0}\left(e_i-p_i^{(2)}\right)^T\varepsilon_k\varepsilon_k^T\left(e_j-p_j^{(2)}\right) \\ &= m^{-1}\operatorname{vec}\left\{\varepsilon_{J_0}\right\}^T H^{ij}_{\text{group}}\operatorname{vec}\left\{\varepsilon_{J_0}\right\}^T. \end{aligned} \tag{S136}$$

We now express S_{IV} as a quadratic form. Let $\pi(X)$ denote the matrix X with reordered columns:

$$\pi(X) = \begin{bmatrix} X_{J_0} & X_{J_1} \end{bmatrix} \quad \text{and} \quad \pi(A) = \operatorname{Cov}\left(\operatorname{vec}\left\{\pi(X)\right\}\right). \tag{S137}$$

Then by (S136) and (S137),

$$\begin{aligned} (S_{\text{IV}})_{ij} &= m^{-1}\operatorname{vec}\left\{\varepsilon_{J_0}\right\}^T H^{ij}_{\text{group}}\operatorname{vec}\left\{\varepsilon_{J_0}\right\}^T + m^{-1}\operatorname{vec}\left\{\varepsilon_{J_1}\right\}^T H^{ij}_{\text{global}}\operatorname{vec}\left\{\varepsilon_{J_1}\right\}^T \\ &= m^{-1}\operatorname{vec}\left\{\pi\left(\varepsilon\right)\right\}^T H^{ij}(J_0,J_1)\operatorname{vec}\left\{\pi\left(\varepsilon\right)\right\} \\ &= m^{-1}\operatorname{vec}\left\{Z\right\}^T\left(\left(\pi(A)^{1/2}\otimes B^{1/2}\right)H^{ij}(J_0,J_1)\left(\pi(A)^{1/2}\otimes B^{1/2}\right)\right)\operatorname{vec}\left\{Z\right\}, \end{aligned}$$

where the last step holds by decorrelation, with $Z \in \mathbb{R}^{n\times m}$ as a random matrix with independent subgaussian entries.

Note that the (i,j)th entry of S_{IV} can be expressed as

$$(S_{\text{IV}})_{ij} = \operatorname{vec}\left\{Z\right\}^T\Phi_{i,j}\operatorname{vec}\left\{Z\right\}, \tag{S138}$$

with

$$\Phi_{i,j} = m^{-1} \left(\pi(A)^{1/2} \otimes B^{1/2}\right) H^{ij}(J_0, J_1) \left(\pi(A)^{1/2} \otimes B^{1/2}\right). \tag{S139}$$

Having expressed $(S_{\mathrm{IV}})_{ij}$ as a quadratic form in (S139), we find the trace and Frobenius norm of $\Phi_{i,j}$, then apply the Hanson-Wright inequality. First we find the trace of $\Phi_{i,j}$. Let

$$\mathcal{I}_0 = \begin{bmatrix} I_{m_0 \times m_0} & 0_{m_0 \times m_1} \\ 0_{m_1 \times m_0} & 0_{m_1 \times m_1} \end{bmatrix} \quad \text{and} \quad \mathcal{I}_1 = \begin{bmatrix} 0_{m_0 \times m_0} & 0_{m_0 \times m_1} \\ 0_{m_1 \times m_0} & I_{m_1 \times m_1} \end{bmatrix}. \tag{S140}$$

Note that $H^{ij}(J_0, J_1)$ can be written as a sum of Kronecker products,

$$H^{ij}(J_0, J_1) = \mathcal{I}_0 \otimes \left(e_j - p_j^{(2)}\right) \left(e_i - p_i^{(2)}\right)^T + \mathcal{I}_1 \otimes \left(e_j - p_j^{(1)}\right) \left(e_i - p_i^{(1)}\right)^T, \tag{S141}$$

hence (S140) can be expressed as

$$m^{-1} \left(\pi(A)^{1/2} \otimes B^{1/2}\right) \left(\mathcal{I}_0 \otimes \left(e_j - p_j^{(2)}\right) \left(e_i - p_i^{(2)}\right)^T\right) \left(\pi(A)^{1/2} \otimes B^{1/2}\right) \tag{S142}$$

$$+ m^{-1} \left(\pi(A)^{1/2} \otimes B^{1/2}\right) \left(\mathcal{I}_1 \otimes \left(e_j - p_j^{(1)}\right) \left(e_i - p_i^{(1)}\right)^T\right) \left(\pi(A)^{1/2} \otimes B^{1/2}\right). \tag{S143}$$

The trace of the term (S143) is

$$\begin{aligned}
&m^{-1} \operatorname{tr}\left(\left(\pi(A)^{1/2} \otimes B^{1/2}\right) \left(\mathcal{I}_0 \otimes \left(e_j - p_j^{(2)}\right) \left(e_i - p_i^{(2)}\right)^T\right) \left(\pi(A)^{1/2} \otimes B^{1/2}\right)\right) \\
&= m^{-1} \operatorname{tr}\left(\pi(A)^{1/2} \mathcal{I}_0 \pi(A)^{1/2} \otimes B^{1/2} \left(e_j - p_j^{(2)}\right) \left(e_i - p_i^{(2)}\right)^T B^{1/2}\right) \\
&= m^{-1} \operatorname{tr}\left(\pi(A)^{1/2} \mathcal{I}_0 \pi(A)^{1/2}\right) \operatorname{tr}\left(B^{1/2} \left(e_j - p_j^{(2)}\right) \left(e_i - p_i^{(2)}\right)^T B^{1/2}\right) \\
&= m^{-1} \operatorname{tr}\left(\mathcal{I}_0 \pi(A)\right) \left(\left(e_i - p_i^{(2)}\right)^T B \left(e_j - p_j^{(2)}\right)\right) \\
&= m^{-1} \operatorname{tr}\left(A_{J_0}\right) \left[(I - P_2) B (I - P_2))\right]_{ij} \\
&= m^{-1} \operatorname{tr}\left(A_{J_0}\right) \widetilde{b}_{ij}.
\end{aligned}$$

Analogously, the trace of the term (S144) is

$$\begin{aligned}
&m^{-1}\operatorname{tr}\left(\left(\pi(A)^{1/2}\otimes B^{1/2}\right)\left(\mathcal{I}_1\otimes\left(e_j-p_j^{(1)}\right)\left(e_i-p_i^{(1)}\right)^T\right)\left(\pi(A)^{1/2}\otimes B^{1/2}\right)\right)\\
&=m^{-1}\operatorname{tr}\left(A_{J_1}\right)\left[(I-P_1)B(I-P_1)\right]_{ij}\\
&=m^{-1}\operatorname{tr}\left(A_{J_1}\right)\breve{b}_{ij}.
\end{aligned}$$

Let $b^\sharp_{ij}$ denote the (i,j)th entry of $B^\sharp$ defined in (S87). We have shown that the trace of $\Phi_{i,j}$ (as defined in (S140)) is

$$\operatorname{tr}\left(\Phi_{i,j}\right)=m^{-1}\operatorname{tr}\left(A_{J_0}\right)\widetilde{b}_{ij}+m^{-1}\operatorname{tr}\left(A_{J_1}\right)\breve{b}_{ij}=b^\sharp_{ij}. \tag{S144}$$

Next, we find the Frobenius norm of $\Phi_{i,j}$. For convenience, define

$$\mathcal{A}_0=\pi(A)^{1/2}\mathcal{I}_0\pi(A)^{1/2}\quad\text{and}\quad\mathcal{A}_1=\pi(A)^{1/2}\mathcal{I}_1\pi(A)^{1/2} \tag{S145}$$

$$\mathcal{B}_{2,ij}=B^{1/2}\left(e_j-p_j^{(2)}\right)\left(e_i-p_i^{(2)}\right)^T B^{1/2}\quad\text{and}\quad\mathcal{B}_{1,ij}=B^{1/2}\left(e_j-p_j^{(1)}\right)\left(e_i-p_i^{(1)}\right)^T B^{1/2}. \tag{S146}$$

Then

$$\begin{aligned}
\left\|\Phi_{i,j}\right\|_F^2&=\left\|m^{-1}\left(\pi(A)^{1/2}\otimes B^{1/2}\right)H^{ij}(J_0,J_1)\left(\pi(A)^{1/2}\otimes B^{1/2}\right)\right\|_F^2\\
&=m^{-2}\left\|\mathcal{A}_0\otimes\mathcal{B}_{2,ij}+\mathcal{A}_1\otimes\mathcal{B}_{1,ij}\right\|_F^2\\
&=m^{-2}\operatorname{tr}\left(\left(\mathcal{A}_0\otimes\mathcal{B}_{2,ij}+\mathcal{A}_1\mathcal{B}_{1,ij}\right)^T\left(\mathcal{A}_0\otimes\mathcal{B}_{2,ij}+\mathcal{A}_1\otimes\mathcal{B}_{1,ij}\right)\right)\\
&=m^{-2}\operatorname{tr}\left(\mathcal{A}_0^T\mathcal{A}_0\otimes\mathcal{B}_{2,ij}^T\mathcal{B}_{2,ij}\right)+m^{-2}\operatorname{tr}\left(\mathcal{A}_1^T\mathcal{A}_1\otimes\mathcal{B}_{1,ij}^T\mathcal{B}_{1,ij}\right)\\
&\quad+m^{-2}\operatorname{tr}\left(\mathcal{A}_0^T\mathcal{A}_1\otimes\mathcal{B}_{2,ij}^T\mathcal{B}_{1,ij}\right)+m^{-2}\operatorname{tr}\left(\mathcal{A}_1^T\mathcal{A}_0\otimes\mathcal{B}_{1,ij}^T\mathcal{B}_{2,ij}\right).
\end{aligned} \tag{S147}$$

We now find the traces of each of the terms in (S148). First, note that

$$\operatorname{tr}\left(\mathcal{A}_0^T\mathcal{A}_0\right)=\operatorname{tr}\left(\mathcal{I}_0\pi(A)\mathcal{I}_0\pi(A)\right)=\operatorname{tr}\left(A_{J_0}^2\right)=\left\|A_{J_0}\right\|_F^2. \tag{S148}$$

Analogously,

$$\operatorname{tr}\left(\mathcal{A}_1^T\mathcal{A}_1\right)=\left\|A_{J_1}\right\|_F^2. \tag{S149}$$

For the cross-term, let $A_{J_0J_1}$ denote the $m_0 \times m_1$ submatrix of $\pi(A)$ given by columns of A in J_0 and rows of A in J_1. Then

$$\begin{aligned}
\operatorname{tr}\left(\mathcal{A}_0^T \mathcal{A}_1\right) &= \operatorname{tr}\left(\mathcal{I}_0 \pi(A) \mathcal{I}_1 \pi(A)\right) \\
&= \operatorname{tr}\left(\begin{bmatrix} 0_{m_0 \times m_0} & A_{J_0J_1} \\ 0_{m_1 \times m_0} & 0_{m_1 \times m_1} \end{bmatrix} \pi(A)\right) \\
&= \operatorname{tr}\left(A_{J_0J_1}^T A_{J_0J_1}\right) \\
&= \left\|A_{J_0J_1}\right\|_F^2 . \qquad \text{(S150)}
\end{aligned}$$

Next,

$$\begin{aligned}
\operatorname{tr}\left(\mathcal{B}_{1,ij}^T \mathcal{B}_{1,ij}\right) &= \operatorname{tr}\left(B^{1/2}\left(e_i - p_i^{(1)}\right)\left(e_j - p_j^{(1)}\right)^T B \left(e_j - p_j^{(1)}\right)\left(e_i - p_i^{(1)}\right)^T B^{1/2}\right) \\
&= \left(\left(e_j - p_j^{(1)}\right)^T B \left(e_j - p_j^{(1)}\right)\right)\left(\left(e_i - p_i^{(1)}\right)^T B \left(e_i - p_i^{(1)}\right)\right) \\
&= \breve{b}_{jj}\breve{b}_{ii}. \qquad \text{(S151)}
\end{aligned}$$

Analogously,

$$\begin{aligned}
\operatorname{tr}\left(\mathcal{B}_{2,ij}^T \mathcal{B}_{2,ij}\right) &= \left(\left(e_j - p_j^{(2)}\right)^T B \left(e_j - p_j^{(2)}\right)\right)\left(\left(e_i - p_i^{(2)}\right)^T B \left(e_i - p_i^{(2)}\right)\right) \\
&= \widetilde{b}_{jj}\widetilde{b}_{ii}. \qquad \text{(S152)}
\end{aligned}$$

The cross-terms yield

$$\operatorname{tr}\left(\mathcal{B}_{1,ij}^T \mathcal{B}_{2,ij}\right) = \left(\left(e_j - p_j^{(1)}\right)^T B \left(e_j - p_j^{(2)}\right)\right)\left(\left(e_i - p_i^{(2)}\right)^T B \left(e_i - p_i^{(1)}\right)\right) = \breve{b}_{ii}\breve{b}_{jj}. \qquad \text{(S153)}$$

The squared Frobenius norm of $\Phi_{i,j}$ is

$$\begin{aligned}
\left\|\Phi_{i,j}\right\|_F^2 &= \frac{1}{m^2}\left(\left\|A_{J_0}\right\|_F^2 \breve{b}_{ii}\breve{b}_{jj} + \left\|A_{J_1}\right\|_F^2 \widetilde{b}_{ii}\widetilde{b}_{jj} + 2\left\|A_{J_0,J_1}\right\|_F^2 \breve{b}_{ii}\breve{b}_{jj}\right) \\
&\leqslant \frac{1}{m^2} C\left(\left\|A_{J_0}\right\|_F^2 + \left\|A_{J_1}\right\|_F^2 + 2\left\|A_{J_0J_1}\right\|_F^2\right)\|B\|_2^2 \\
&= C\frac{1}{m^2}\|A\|_F^2 \|B\|_2^2 .
\end{aligned}$$

We now apply the Hanson-Wright inequality,

$$\begin{aligned}\mathbb{P}\left(\left|(S_{\mathrm{I}})_{ij} - b^{\sharp}_{ij}\right| > r_4\right) &= \mathbb{P}\left(\left|\mathrm{vec}\left\{Z\right\}^T \Phi_{i,j}\,\mathrm{vec}\left\{Z\right\} - \mathrm{tr}\left(\Phi_{i,j}\right)\right| > r_4\right) \\ &\leqslant 2\exp\left(-c\min\left\{d\log(m), d^{1/2}\sqrt{\log(m)}\frac{\|\Phi_{i,j}\|_F}{\|\Phi_{i,j}\|_2}\right\}\right) \\ &\leqslant 2\max\left(m^{-d}, \exp\left(d^{1/2}\sqrt{\log(m)}r^{1/2}(\Phi_{i,j})\right)\right).\end{aligned}$$

The first step holds by (S139) and (S145).

J Comparisons to related methods

The most similar existing method to ours is the sphering approach from Allen and Tibshirani (2012). Both methods use a preliminary demeaned version of the data to generate covariance estimates, then use these estimates to adjust the gene-wise t-tests. The largest difference between the procedures lies in this last step. The sphering approach produces an adjusted data set based on decorrelating residuals from a preliminary mean estimate and performs testing and mean estimation on this adjusted data using traditional OLS techniques. Though their approach is well-motivated at the population level, they do not provide theoretical support for their plug-in procedure, and in particular do not explore how noise in the initial mean estimate may complicate their decorrelation procedure. In contrast, our approach uses a generalized least squares approach motivated by classical statistical results including the Gauss Markov theorem.

The sphering approach also involves decorrelating a data matrix along both axes. Our work, including the theoretical analysis in Zhou (2014a), suggests that when the data matrix is non-square, attempting to decorrelate along the longer axis generally degrades performance. For genetics applications, where there are usually many more genes than samples, this suggests that decorrelating along the genes may hurt the performance of the sphering method. Fortunately, for gene-level analyses it is not necessary to decorrelate along the gene axis, since inference methods like false discovery rate are robust to a certain level of dependence among the variables (genes) (Benjamini and Yekutieli, 2001). Therefore, we also consider a modification of the sphering algorithm that only decorrelates along the samples.

Confounder adjustment is another related line of work that deals with similar issues when attempting to discover mean differences. In particular, a part of that literature posits models where row-wise connections arise from the additive effects of potential latent variables. Sun et al. (2012) and Wang et al. (2015) use models of the form

$$X_{n\times m} = D_{n\times 1}\beta^T_{m\times 1} + Z_{n\times r}\Gamma^T_{m\times r} + E_{n\times m}$$
$$Z_{n\times r} = D_{n\times 1}\alpha^T_{r\times 1} + W_{n\times r}$$

where Z is an unobserved matrix of r latent factors. Rewriting these equations into the following form lets us better contrast the confounder model to our matrix-variate setup:

$$X = D(\beta + \Gamma\alpha)^T + W\Gamma^T + E. \tag{S154}$$

These models are generally estimated by using some form of factor analysis to estimate Γ and then using regression methods with additive outlier detection to identify β, methodology that is quite different from our GLS-based methods.

For the two-group model, in the case of a globally centered data matrix X, the design matrix D in (S155) takes the form

$$D^T_{n\times 1} = \begin{bmatrix} -1 & \cdots & -1 & 1 \cdots & 1 \end{bmatrix} = \begin{bmatrix} -1^T_{n_1} & 1^T_{n_2} \end{bmatrix}, \tag{S155}$$

and 2β represents the vector of true mean differences between the groups. The vector β is estimated via OLS, yielding $\widehat{\beta}_{\text{OLS}}$, and CATE considers whether the residual $X - D_{n\times 1}\widehat{\beta}_{\text{OLS}}$ has a low-rank covariance structure plus noise. If so, $\widehat{\Gamma}\widehat{\alpha}$ aims to take out the residual low-rank structure through $D(\widehat{\Gamma\alpha})^T$. As illustrated in simulation and data analysis, this improves upon inference based only on $\widehat{\beta}_{\text{OLS}}$. When applying the CATE and related methods to data originated from the generative model as described in the present paper, CATE (and in particular, the related LEAPP) method essentially seeks a sparse approximation of $\widehat{\beta}_{\text{OLS}}$; Moreover in LEAPP, this is essentially achieved via hard thresholding of coefficients of $\widehat{\beta}_{\text{OLS}}$, leading to improvements in performance in variable selection and its subsequence inference when the vector of true mean differences is presumed to be sparse. In our setting, we improve upon OLS using GLS.

J.1 Simulation results

Figure S3 compares the performance of Algorithm 2 to the sphering method of Allen and Tibshirani (2012) and the robust regression confounder adjustment method of Wang et al. (2015) on simulated matrix variate data motivated by the ulcerative colitis dataset described in Section 5. Note that this robust regression confounder adjustment is a minor modification of the LEAPP algorithm introduced in Sun et al. (2012). As discussed above, we also consider a modification of Allen and Tibshirani (2012) that only decorrelates along the rows.

We can see that across a range of dataset sizes our method consistently outperforms sphering in terms of sensitivity and specificity for identifying mean differences. In some settings, CATE improves on Tsphere and t-statistics despite being applied on misspecified models, because CATE takes out the additional rank two structure from the mean after OLS regression and does some approximate thresholding on the coefficients. Our method using GLS performs significantly better than CATE in the setting of non-identity B, with edges present both within and between groups.

Figure S5 fixes the sample size and repeats these comparisons on different sample correlation structures (which are described in Section 4). Figure S6 is analogous to Figure S5, but with A as the identity matrix. Algorithm 2 is competitive or superior to the competing methods across a range of topologies.

J.2 Comparison on UC data

We apply both Algorithm 2 and CATE on the ulcerative colitis data to compare their respective findings on real data. Figure S7 presents the test statistics from these algorithms. The test statistics have a correlation of 0.75. As expected, both methods find that the bulk of genes have small test statistics. Note that the regression line of the CATE test statistics on Algorithm 2's test statistics has a slope less than 1. This implies that Algorithm 2 generates more dispersed test statistics than CATE, and, given that we have shown in Figures 5 and 8 that Algorithm 2 provides well-calibrated test statistics, that it also has more power in this situation.

Using a threshold of FDR adjusted p-values smaller than 0.1, both methods find four genes with significant mean differences. However, there is only one gene (DPP10-AS1) that

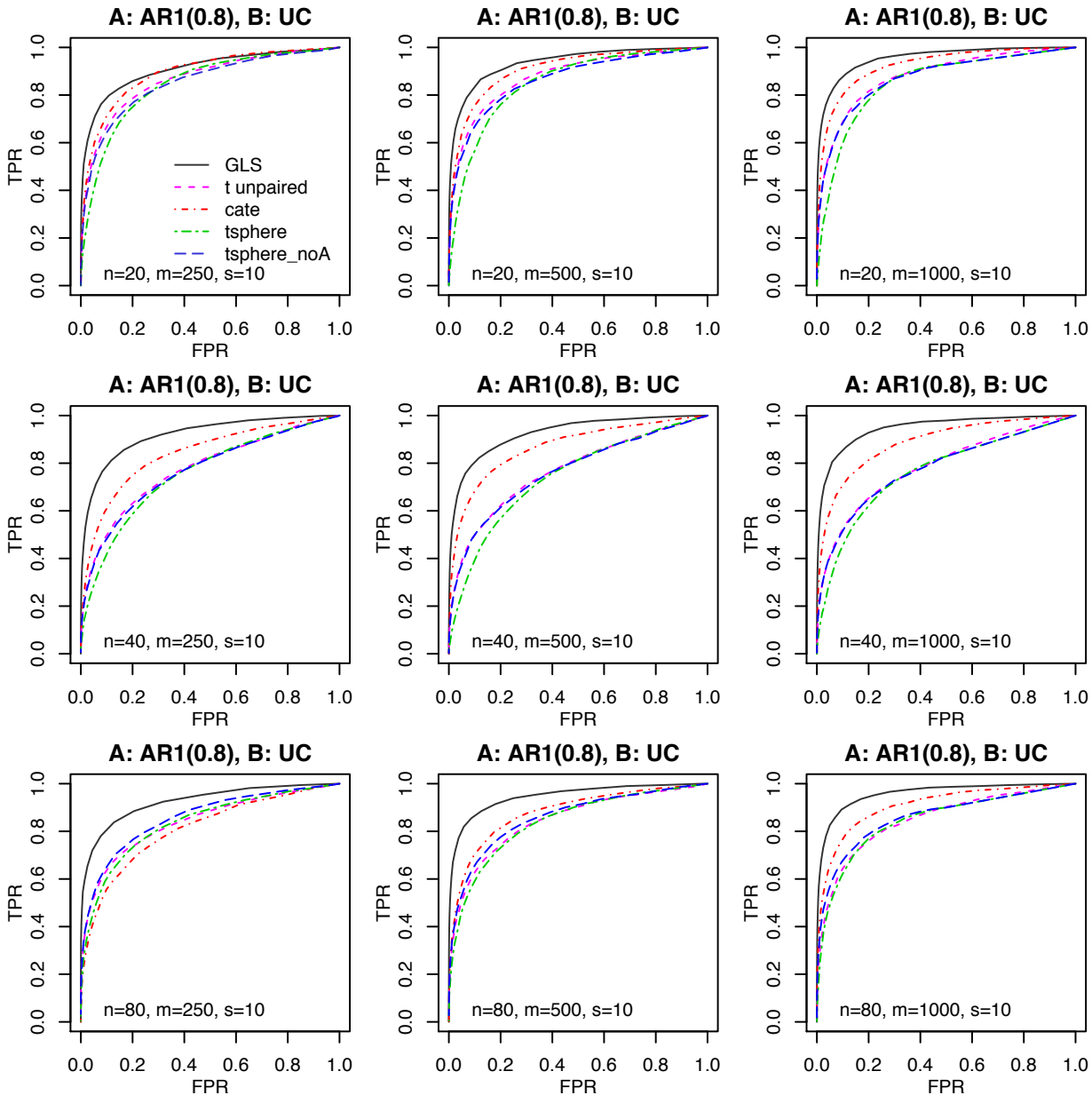


Figure S3: Performance of Algorithm 2 (GLS) relative to sphering and confounder adjustment methods, labeled as `tsphere` and `cate`, respectively. These are ROC curves for identifying true mean differences. An implementation of the sphering algorithm that does not adjust for A is also included, labeled as `tsphere_noA`. Each panel shows the average ROC curves over 200 simulations. We simulate matrix variate data with gene correlations from an AR1(0.8) model and let $s = 10$ genes have true mean differences of 0.8, 0.6, and 0.4 for the first, second and third rows, respectively. For all of these the true B is set to $\widehat{B}$ from the ulcerative colitis data (using a repeated block structure for larger n values), described in Section 5 and evenly-sized groups are assigned randomly.

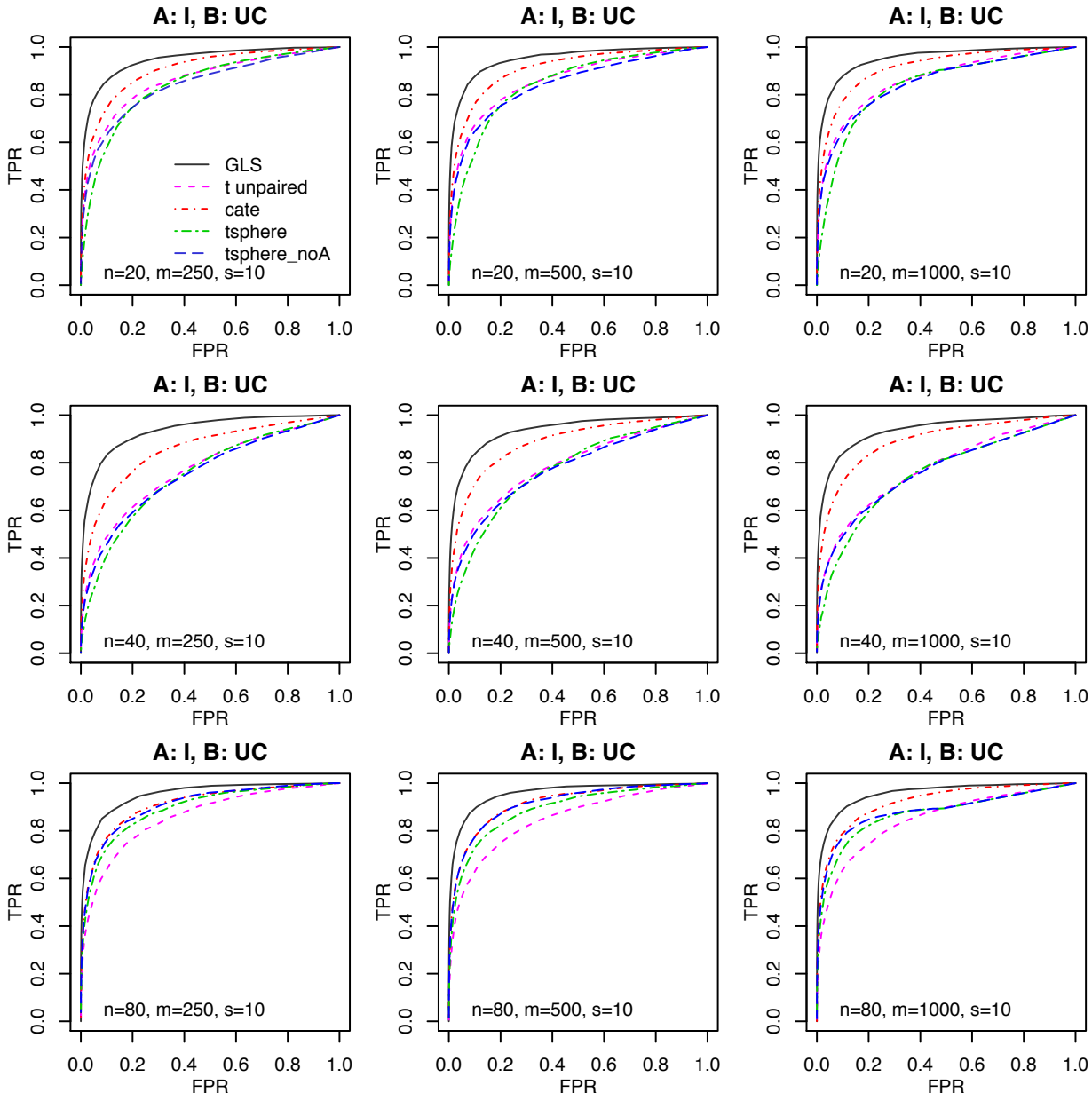


Figure S4: Performance of Algorithm 2 (GLS) relative to sphering and confounder adjustment, labeled as `tsphere` and `cate`, respectively. These are ROC curves for identifying true mean differences. An implementation of the sphering algorithm that does not adjust for A is also included, labeled as `tsphere_noA`. Each panel shows the average ROC curves over 200 simulations. We simulate matrix variate data with no gene-wise correlations ($A = I$) and let $s = 10$ genes have true mean differences of 0.8, 0.6, and 0.4 for the first, second and third rows, respectively. For all of these the true B is set to $\widehat{B}$ from the ulcerative colitis data (using a repeated block structure for larger n values), described in Section 5 and evenly-sized groups are assigned randomly.

both methods identify. So, although there is significant correlation between the test statistics, the methods do not necessarily identify the same genes.

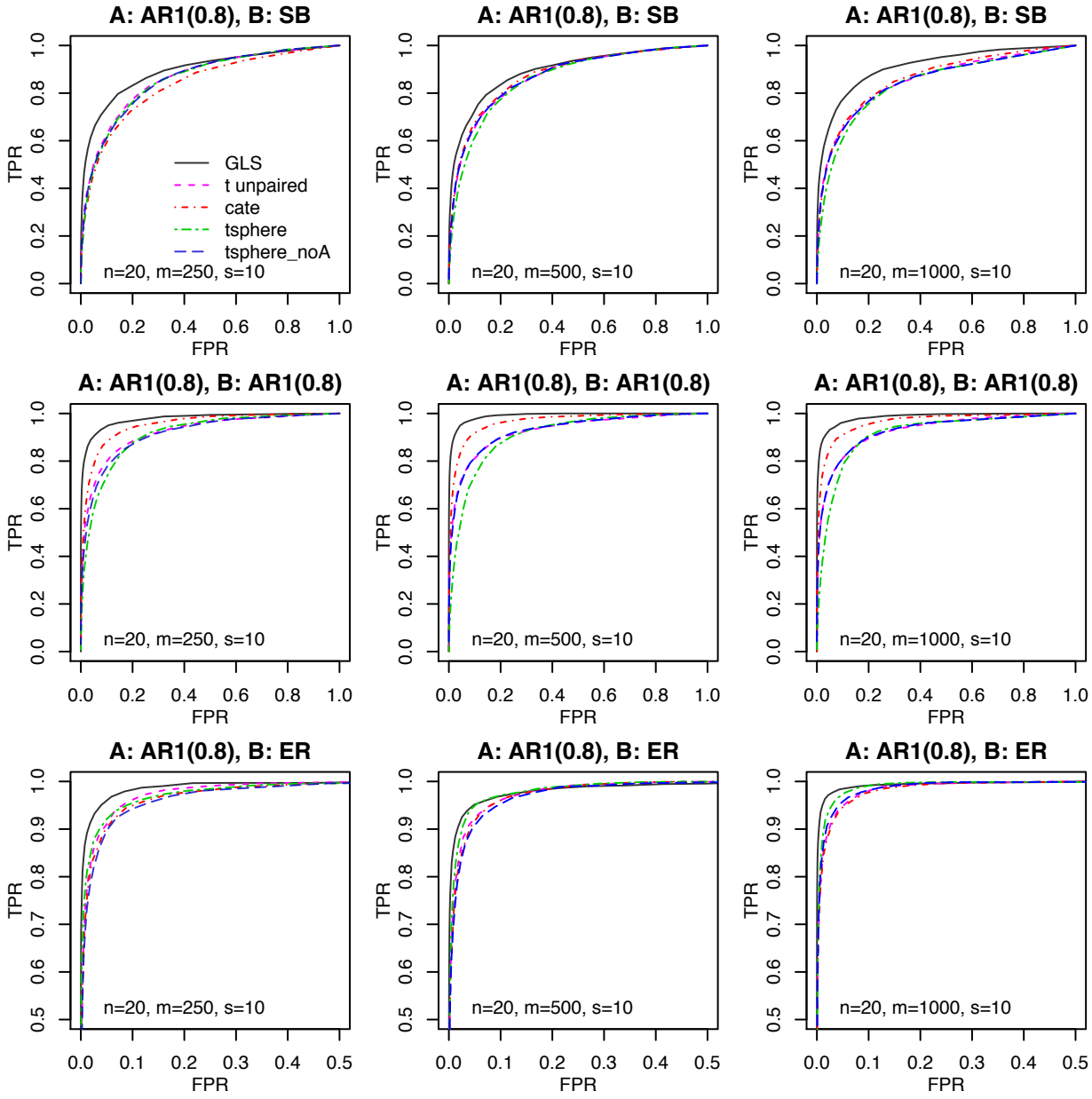


Figure S5: Performance of Algorithm 2 (GLS) relative to sphering and confounder adjustment, labeled as `tsphere` and `cate`, respectively. These are ROC curves for identifying true mean differences. An implementation of the sphering algorithm that does not adjust for A is also included, labeled as `tsphere_noA`. Each panel shows the average ROC curves over 200 simulations. We simulate matrix variate data with an AR1(0.8) model for A and let $s = 10$ genes have true mean differences of 0.8. B is constructed according to a Star-Block model with blocks of size 4, an AR1(0.8), and an Erdős-Rényi random graph with $d = n \log n$ edges. All of these use $n = 20$ and randomly assign 10 observations to each group.

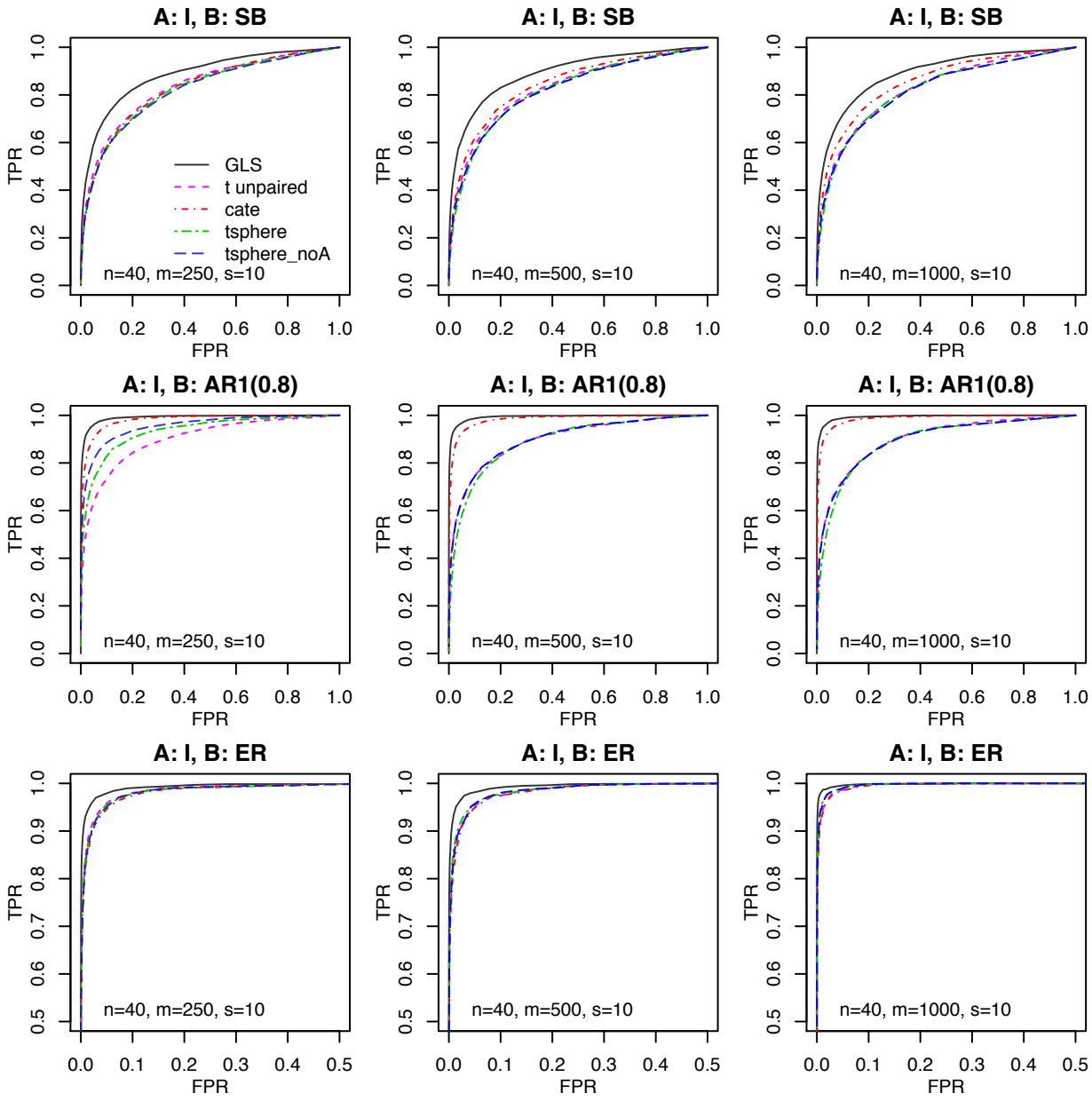


Figure S6: Performance of Algorithm 2 (GLS) relative to sphering and confounder adjustment, labeled as `tsphere` and `cate`, respectively. These are ROC curves for identifying true mean differences. An implementation of the sphering algorithm that does not adjust for A is also included, labeled as `tsphere_noA`. Each panel shows the average ROC curves over 200 simulations. We simulate matrix variate data with no gene-wise correlations ($A = I$) and let $s = 10$ genes have true mean differences of 0.6. B is constructed according to a Star-Block model with blocks of size 4, an AR1(0.8), and an Erdős-Rényi random graph with $d = n \log n$ edges. All of these use $n = 40$ and randomly assign 20 observations to each group.

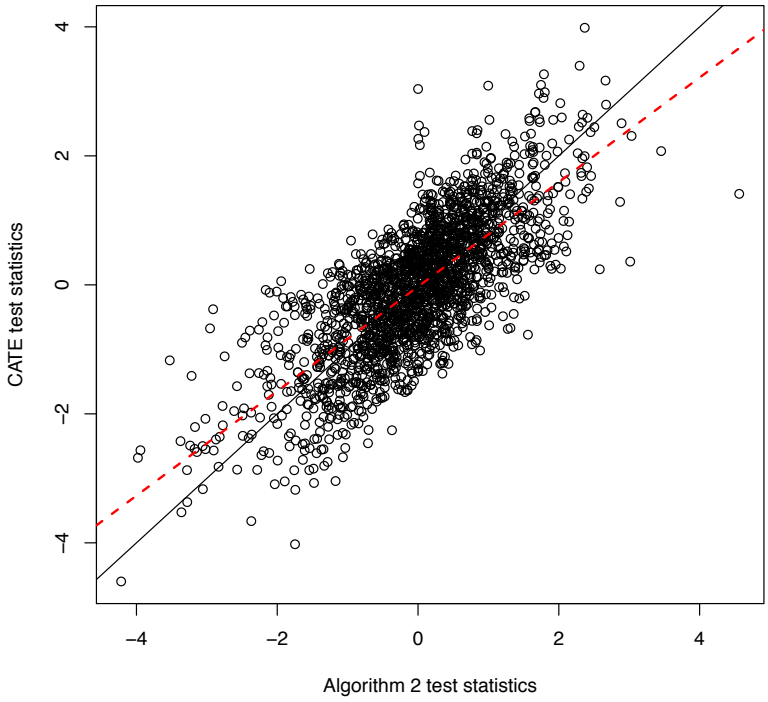


Figure S7: Scatterplot of t-statistics for CATE and Algorithm 2 applied on the ulcerative colitis data. The 45-degree line is included in black while red dashed line is the linear fit.

References

Allen, G. I. and Tibshirani, R. (2012). Inference with transposable data: modelling the effects of row and column correlations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 721–743.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics* 1165–1188.

Rothman, A., Bickel, P., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2** 494–515.

Sun, Y., Zhang, N. R. and Owen, A. B. (2012). Multiple hypothesis testing adjusted for latent variables, with an application to the agemap gene expression data. *The Annals of Applied Statistics* 1664–1688.

Wang, J., Zhao, Q., Hastie, T. and Owen, A. B. (2015). Confounder adjustment in multiple hypothesis testing. *arXiv preprint arXiv:1508.04178* .

Zhou, S. (2014a). Gemini: Graph estimation with matrix variate normal instances. *Annals of Statistics* **42** 532–562.

Zhou, S. (2014b). Supplement to "Gemini: Graph estimation with matrix variate normal instances". *Annals of Statistics* DOI:10.1214/13-AOS1187SUPP.

Zhou, S., Rütimann, P., Xu, M. and Bühlmann, P. (2011). High-dimensional covariance estimation based on Gaussian graphical models. *Journal of Machine Learning Research* **12** 2975–3026.