Joint mean and covariance estimation with unreplicated matrix-variate data *

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

November 13, 2016

Abstract

It has been proposed that complex populations, such as those that arise in genomics studies, may exhibit dependencies among the observations as well as among the variables. This gives rise to the challenging problem of analyzing unreplicated high-dimensional data with unknown mean and dependence structures. Matrix-variate approaches that impose various forms of (inverse) covariance sparsity allow flexible dependence structures to be estimated, but cannot directly be applied when the mean and dependences are estimated jointly. We present a practical method utilizing penalized (inverse) covariance estimation and generalized least squares to address this challenge. The advantages of our approaches are: (i) dependence graphs and covariance structures can be estimated in the presence of unknown mean structure, (ii) the mean structure becomes more efficiently estimated when accounting for the dependence structure(s); and (iii) inferences about the mean parameters become correctly calibrated. We establish consistency and obtain rates of convergence for estimating the mean parameters and covariance matrices. We use simulation studies and analysis of genomic data from a twin study of ulcerative colitis to illustrate the statistical convergence and the performance of the procedures in practical settings. Furthermore, several lines of evidence show that the test statistics for differential gene expression produced by our method are correctly calibrated.

Keywords: two-group comparison, sparsity, genomics, generalized least squares, graphical modeling

*The research is supported in part by NSF Grant DMS-13-16731 and Elizabeth C. Crosby Research Award to SZ from the University of Michigan.

1 Introduction

Understanding how changes in gene expression are related to changes in biological state is one of the fundamental tasks in genomics research, and is a prototypical example of "large scale inference" (Efron, 2010). While some genomics datasets have within-subject replicates or other known clustering factors that could lead to dependence among the observations, many genomics datasets are viewed as population cross-sections or convenience samples, and are usually analyzed by taking the observations (biological samples) to be statistically independent of each other. Countering this conventional view, Efron (2009) proposed that there may be unanticipated correlations between samples even when the study design would not suggest it. These correlations may lead to inflated evidence for mean differences, and could be one explanation for the claimed lack of reproducibility in genomics research (Sugden et al., 2013; Leek et al., 2010).

To identify and adjust for unanticipated sample-wise correlations, Efron (2009) proposed an empirical Bayes approach utilizing the sample moments of the data. Subsequently, Allen and Tibshirani (2012) proposed a two-stage approach which first decorrelates the residuals using estimates of the sample-wise and gene-wise covariance matrices, and then applies two-sample testing to the resulting adjusted data. We propose an alternative approach to jointly estimate the mean and covariance with a single instance of the design matrix, exploiting recent developments in two-way covariance estimation for matrix-variate data (Zhou, 2014a) combined with the classical idea of generalized least squares (GLS) (Aitken, 1936). We motivate the approach using differential expression analysis in a genomics context, but the method is broadly applicable to the matrix-variate data having unknown mean and covariance structures, with or without replications.

The basic idea of our approach is to remove as much of the mean structure as needed to allow the sample-wise covariance matrix to be accurately estimated. We then plug this covariance matrix estimate into a GLS estimator of the mean structure, and use Wald-type statistics to conduct inference. Such an approach has the potential to improve the accuracy of mean structure estimation, due to the Gauss-Markov theorem from classical least squares analysis. Just as importantly, the estimated covariance matrix can be used in uncertainty assessment and formal testing of the mean parameters, thereby improving calibration of the inferences. In addition, estimates of the variable-wise and sample-wise covariance matrices can be used for other purposes such as graphical modeling of dependence structures.

A persistent problem in genomics research is that test statistics for mean parameters (e.g. t-statistics for two-group comparisons) often appear to be incorrectly calibrated (Efron, 2005; Allen

and Tibshirani, 2012). For example, if the test statistics are uniformly more dispersed, as can sometimes be seen in a quantile plot, this is usually taken to be an indication of miscalibration, rather than reflecting a nearly global pattern of differential effects (Efron, 2007). Adjustments such as genomic control (Devlin and Roeder, 1999) are often used to account for this. We illustrate, using simulation studies and a genomic study of ulcerative colitis, that estimating and accounting for the sample-wise dependence using our procedure can dramatically improve the calibration of test statistics. This suggests that at least in some cases, unanticipated sample-wise dependence exhibits a strong effect on the statistical inferences.

One major challenge we face is that the variables (e.g. genes or mRNA transcripts) have a complex correlation structure that exists together with any correlations among the observations. As pointed out by Efron (2009) and others, the presence of correlations among the samples makes it more difficult to estimate correlations among variables, and vice versa. A second major challenge is that due to dependence among both observations and variables, there is no independent replication in the data, that is, we have a single sample of size one to use for covariance estimation. This challenge is addressed in Zhou (2014a) when the mean structure is taken to be zero. A third major challenge that is unique to our framework is that the covariance structures can only be estimated after removing the mean structure, a fact that is generally not considered in most work on high dimensional covariance estimation, where the population mean is taken to be zero. We elaborate on this challenge next.

Two obvious approaches for removing the mean structure in our setting are to globally center each column of the data matrix (containing the data for one variable), or to center each column within each group. Globally centering each column by ignoring the mean structure may result in a sample covariance matrix that may not be consistent. Group centering all genes, by contrast, may lead to a consistent covariance estimate, as shown in Theorem 2 with regard to our Algorithm 1. However, group centering all genes introduces extraneous noise when the true vector of mean differences is sparse. We find that there is a complex interplay between the mean and covariance estimation tasks, such that overly flexible modeling of the mean structure can introduce large systematic errors into the dependence structure estimates. To mitigate this effect, we aim to center the data based on model selection. More specifically, we adopt a model selection centering approach in which only the mean parameters having a sufficiently large effect size are targeted for removal. This refined approach performs well in simulations.

1.1 Our approach and contributions

In this work, we focus on a joint mean and covariance modeling framework, where the data matrix $X \in \mathbb{R}^{n\times m}$ has one row for each individual and one column for each variable. We focus on the two group setting, where n_1 individuals in group one have mean $\beta^{(1)} \in \mathbb{R}^m$, and n_2 individuals in group two have mean $\beta^{(2)} \in \mathbb{R}^m$. Our goal is to estimate the group mean vectors $\beta^{(1)}, \beta^{(2)}$, the vector of mean differences between two groups $\gamma = \beta^{(2)} - \beta^{(1)} \in \mathbb{R}^m$, the row-wise covariance matrix $B \in \mathbb{R}^{n\times n}$, and the column-wise covariance matrix $A \in \mathbb{R}^{m\times m}$. In Section 1.2, we focus on the estimation of $\gamma \in \mathbb{R}^m$, while in Section 2 we focus on joint mean and covariance estimation for γ, A and B, under the separable covariance structure which we elaborate now.

Our approach to covariance modeling builds on the Gemini method (Zhou, 2014a), which is designed to estimate a separable covariance matrix for data with two-way dependencies. Loosely speaking, we say that an $n \times m$ random matrix X follows a matrix variate distribution with mean $M \in \mathbb{R}^{n\times m}$ and a separable covariance matrix $\Sigma = A \otimes B$, which we write

$$X_{n\times m} \sim \mathcal{L}_{n,m}(M, A_{m\times m} \otimes B_{n\times n}), \tag{1}$$

is equivalent to say $\mathrm{vec}\{\,\mathrm{X}\,\}$ has mean $\mathrm{vec}\{\,\mathrm{M}\,\}$ and covariance $\Sigma = A \otimes B$. For matrices $A \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{n\times n}$, the Kronecker product $A \otimes B \in \mathbb{R}^{mn\times nm}$ is the block matrix for which the (i, j)th block is $a_{ij}B$, for $i, j \in \{1, \ldots, m\}$. Here $\mathrm{vec}\{\,\mathrm{X}\,\}$ is formed by stacking the columns of X into a vector in $\mathbb{R}^{mn}$. As mentioned, for the mean matrix M, we focus on the two-group setting to be defined in (5). Intuitively, A describes the covariance between columns of X while B describes the covariance between rows of X. Note that we can only estimate A and B up to a scaled factor, as $A\eta \otimes \frac{1}{\eta}B = A \otimes B$ for any $\eta > 0$, and hence this will be our goal of the paper, and precisely what we mean, when we say we are interested in estimating covariances A and B.

In the separable covariance model, the matrix A represents the shared covariance among variables for each sample, while B represents the covariance among observations which in turn is shared by all genes. In our motivating genomics applications, B is often incorrectly anticipated to have a simple known structure, for example, B is taken to be diagonal if the observations are assumed to be uncorrelated. However, we show by example in Section 5 that departures from the anticipated diagonal structure may occur, corroborating earlier claims of this type by Efron (2009) and others. There, we will show that the test statistics for differential gene expression produced by our method are correctly calibrated (c.f. (22) and Figure 8). We will show that our theory and analysis works with a model much more general than the matrix variate normal model, which we will define in

Section 1.2.

We characterize the following theoretical contributions of our method. In Section 1.2, we elaborate upon the two group mean model and the GLS algorithm, where we do not impose a separable covariance model in the sense of (1). We then present the rate of convergence for mean estimation in the two-group model for subgaussian data in Theorem 1. In particular, we bound the statistical error on estimating each column of the mean matrix using the GLS procedure so long as each column of X shares the same covariance matrix B, for which we have a close approximation. In Section 2.1, we specialize Theorem 1 to the setting where $\mathrm{vec}\{\mathrm{X}\}$ now follows a separable covariance model, for which we proposed two centering algorithms. We illustrate their interactions with the iterative covariance estimation procedures in Section 2 as well as in simulation studies. We then provide a joint rate of convergence for the mean and covariance estimation using Algorithm 1 in Theorem 2. We expect the same analysis to go through for Algorithm 2. Finally, in the analysis of the observation-wise sample covariance, we characterize an interesting bias-variance trade-off in the entry-wise error, which we make explicit in the supplement in Section D (c.f. (85) and (86)). There, we elaborate that the source of the bias comes from the statistical error in estimating the high-dimensional mean vectors.

1.2 The model and the method for mean estimation

In this section we present our model and method for joint mean and covariance estimation. We will state an initial theoretical characterization of the GLS-based approach for mean estimation while defer the theory and more general methodology to Section 2. Our results apply to subgaussian data. Before we present the model, we define subgaussian random vectors and the ψ_2 norm. For a vector $y = (y_1, \ldots, y_p) \in \mathbb{R}^p$, denote by $\|y\|_2 = \sqrt{\sum_{i=1}^p y_i^2}$.

Definition 1.1. Let Y be a random vector in $\mathbb{R}^p$.

1. Y is called isotropic if for every $y \in \mathbb{R}^p$, $E[|\langle Y, y\rangle|^2] = \|y\|_2^2$.
2. Y is ψ_2 with a constant α if for every $y \in \mathbb{R}^p$,

$$\|\langle Y, y\rangle\|_{\psi_2} := \inf\{t : E[\exp(\langle Y, y\rangle^2/t^2)] \leqslant 2\} \leqslant \alpha\|y\|_2. \tag{2}$$

The ψ_2 condition on a scalar random variable V is equivalent to the subgaussian tail decay of V, which means $P(|V| > t) \leqslant 2\exp(-t^2/c^2)$, for all $t > 0$.

The model. Our model for the matrix-variate data X can be expressed as a mean matrix plus a noise term,

$$X = M + \varepsilon, \tag{3}$$

where the columns of ε are subgaussian. Let $u, v, \in \mathbb{R}^n$ be defined as

$$u = (\underbrace{1, \ldots, 1}_{n_1}, \underbrace{0, \ldots, 0}_{n_2}) \in \mathbb{R}^n \quad \text{and} \quad v = (\underbrace{0, \ldots, 0}_{n_1}, \underbrace{1, \ldots, 1}_{n_2}) \in \mathbb{R}^n. \tag{4}$$

Let $\mathbf{1}_n \in \mathbb{R}^n$ denote a vector of ones. For the two-group model, we take the mean matrix to have the form

$$M = D\beta = \begin{bmatrix} \mathbf{1}_{n_1}\beta^{(1)T} \\ \mathbf{1}_{n_2}\beta^{(2)T} \end{bmatrix} \in \mathbb{R}^{n \times m}, \quad \text{where} \quad D = \begin{bmatrix} u & v \end{bmatrix} \in \mathbb{R}^{n \times 2} \tag{5}$$

is the design matrix and $\beta = (\beta^{(1)}, \beta^{(2)})^T \in \mathbb{R}^{2 \times m}$ is a matrix of group means. Let $\gamma = \beta^{(1)} - \beta^{(2)} \in \mathbb{R}^m$ denote the vector of mean differences, and let $d_0 = |\operatorname{supp}(\gamma)| = |\{j : \gamma_j \neq 0\}|$ denote the size of the support of γ. To estimate the group means, we use a GLS estimator,

$$\widehat{\beta}(\widehat{B}^{-1}) := (D^T \widehat{B}^{-1} D)^{-1} D^T \widehat{B}^{-1} X \in \mathbb{R}^{2 \times m}, \tag{6}$$

where $\widehat{B}^{-1}$ is an estimate of the observation-wise inverse covariance matrix. Throughout the paper, we denote by $\widehat{\beta}(B^{-1})$ the oracle GLS estimator, since it depends on the unknown true covariance B. Also, we denote the estimated vector of mean differences as $\widehat{\gamma}(\widehat{B}^{-1}) = \delta^T \widehat{\beta}(\widehat{B}^{-1}) \in \mathbb{R}^m$, where $\delta = (1, -1) \in \mathbb{R}^2$. Recall that the operator norm $\|B\|_2$ is given by $\sqrt{\varphi_{\max}(BB^T)}$, where $\varphi_{\max}(BB^T)$ denotes the largest eigenvalue of BB^T. We now state a theorem on the rate of convergence of the GLS estimator (6) where we use a fixed approximation $B_{n,m}^{-1}$ to B^{-1} to obtain $\widehat{\beta}(B_{n,m}^{-1})$, where the operator norm of $\Delta_{n,m} = B_{n,m}^{-1} - B^{-1}$ is small in the sense of (7). We will specialize Theorem 1 to the case where B^{-1} is estimated using the baseline methods in Zhou (2014a) when X follows matrix-variate distributions.

Theorem 1. *Let Z be an $n \times m$ random matrix with independent entries Z_{ij} satisfying $\mathbb{E}Z_{ij} = 0$, $1 = \mathbb{E}Z_{ij}^2 \leqslant \|Z_{ij}\|_{\psi_2} \leqslant K$. Let $Z_1, \ldots, Z_m \in \mathbb{R}^n$ be the columns of Z. Suppose the jth column of the data matrix satisfies $X_j \sim B^{1/2} Z_j$. Suppose $B_{n,m} \in \mathbb{R}^{n \times n}$ is a positive definite symmetric matrix. Let $\Delta_{n,m} := B_{n,m}^{-1} - B^{-1}$. Suppose*

$$\|\Delta_{n,m}\|_2 < \frac{1}{(n_{\max}/n_{\min})\,\|B\|_2}, \text{ where } n_{\min} = \min(n_1, n_2) \text{ and } n_{\max} = \max(n_1, n_2). \tag{7}$$

Then with probability at least $1 - 4/(m \vee n)^2$*, for some absolute constants* C*,* C'*,*

$$\|\widehat{\beta}_j(B_{n,m}^{-1}) - \beta_j^*\|_2 \leqslant r_{n,m} := s_{n,m} + t_{n,m}, \quad \textit{where} \tag{8}$$

$$s_{n,m} = C\sqrt{\frac{\log(m)\|B\|_2}{n_{\min}}} \quad \textit{and} \quad t_{n,m} = C'\frac{\|\Delta_{n,m}\|_2}{n_{\min}^{1/2}}. \tag{9}$$

Moreover, with probability at least $1 - 4/(n \vee m)^2$*,*

$$\|\widehat{\gamma}(B_{n,m}) - \gamma\|_\infty \leqslant \sqrt{2}\left(C\sqrt{\frac{\log(m)\|B\|_2}{n_{\min}}} + C' n_{\min}^{-1/2}\|\Delta_{n,m}\|_2\right). \tag{10}$$

We prove Theorem 1 in Section 6. Additional technical lemmas are proved in Section B.

Remarks. If the operator norm of B is bounded, that is $\|B\|_2 < W$, then condition (7) is equivalent to $\|\Delta_{n,m}\|_2 < 1/(W n_{\text{ratio}})$. The term $t_{n,m}$ in (9) reflects the error due to approximating B^{-1} with $B_{n,m}^{-1}$, whereas $s_{n,m}$ reflects the error in estimating the mean matrix (6) as defined using GLS with the true B^{-1} for the random design X. The term $s_{n,m}$ is $O(\sqrt{\log(m)/n})$, whereas $t_{n,m}$ is $O(1/\sqrt{n})$. If $\|B\|_2$ is bounded below, then $s_{n,m}$ dominates $t_{n,m}$ because of the additional factor of $\sqrt{\log(m)}$. The term $s_{n,m}$ in (9) can be replaced by the tighter bound, namely,

$$s'_{n,m} = C' \log^{1/2}(m)\sqrt{\delta^T(D^T B^{-1} D)^{-1}\delta},$$

with $\delta = (1, -1) \in \mathbb{R}^2$. This bound correctly drops the factor of $\|B\|_2$ present in (9) while revealing that variation aligned with the column space of D is especially important in this problem. Note that $\delta^T(D^T B^{-1} D)^{-1}\delta = \text{Var}(\widehat{\gamma}_j(B^{-1}))$ for all $j = 1, \ldots, m$, that is, the factor appearing above in $s'_{n,m}$ is the variance of the "oracle" GLS estimator (6) of γ_j using the true B. The "design effect" $\delta^T(D^T B^{-1} D)^{-1}\delta$ contributes to the rate of the data-driven procedure characterized in Theorem 1. In Section 4, we present simulation results that demonstrate the advantage of the oracle GLS and non-oracle GLS (6) methods over the sample mean-based method (c.f. (18) and (28)) for mean estimation as well as the related variable selection problem with respect to γ. There, we scrutinize this quantity and its estimation procedures to details.

1.3 Related work

Efron (2009) introduced an approach for inference on mean differences in data with with two-way dependence. The approach uses empirical Bayes' ideas and tools from large scale inference, and also explores the challenge of conducting inference on mean-parameters when there is uncharacter-

ized dependence among samples. Allen and Tibshirani (2012) also considered this question and developed a very different approach based on decorrelating residuals.

Another line of relevant research has focused on hypothesis testing of high-dimensional means, exploiting assumed sparsity of effects, and developing theoretical results using techniques from high dimensional estimation theory. Work of this type include Cai and Xia (2014); Chen et al. (2014); Bai and Saranadasa (1996); Chen et al. (2010).

Our inference procedures are based on Z-scores and associated FDR values for mean comparisons of individual variables. This relies on earlier work for false discovery rate estimation using correlated data, including Owen (2005); Benjamini and Yekutieli (2001); Cai et al. (2011); Li and Zhong (2014); Benjamini and Hochberg (1995); Storey (2003). Taking a different approach, Hall et al. (2010) develop the innovated higher criticism test statistic to detect differences in means in the presence of correlations between genes.

Our method builds on the Gemini estimator introduced by (Zhou, 2014a), for covariance matrices when both the rows and columns of the data matrix are dependent. In the setting where the correlations exist along only one axis of the array, researchers have proposed various covariance estimators and studied their theoretical and numerical properties (Banerjee et al., 2008; Fan et al., 2009; Friedman et al., 2008; Lam and Fan, 2009; Meinshausen and Bühlmann, 2006; Peng et al., 2009; Ravikumar et al., 2011; Rothman et al., 2008; Yuan and Lin, 2007; Zhou et al., 2010). Although we focus on the setting of Kronecker products, or separable covariance structures, Cai et al. (2015) proposed a covariance estimator for a model with several populations, each of which may have a different variable-wise covariance matrix.

1.4 Organization

The remainder of the paper is organized as follows. In Section 2, we present our matrix-variate modeling framework and methods on joint mean and covariance estimation. In Section 3, we present our main theorem on the rate of convergence for estimating the mean and covariance matrices when the data follows a matrix-variate model. After stating our theoretical results, we review related work and put our results in context. In Section 4, we demonstrate through simulations that in several realistic settings, our algorithms outperform ordinary least squares estimators in terms of accuracy and variable selection consistency. In Section 5, we analyze a gene expression data set, and show that our method corrects test statistic overdispersion that is clearly present when using sample mean-based methods (c.f. Section 4.3). We provide a proof sketch for Theorem 1 in Section 6. We conclude in Section 7. We place all technical proofs in the supplementary material. In

Sections A and C, we provide proofs for Theorem 2. In Section B, we prove technical lemmas for Theorem 1. In Section D, we prove entry-wise rates of convergence for the sample covariance matrices, which might be of independent interests.

Notation. Before we leave this section, we introduce the notation needed for the technical sections. Let $e_1, \ldots, e_p$ be the canonical basis of $\mathbb{R}^p$. For a matrix $A = (a_{ij})_{1 \leqslant i,j \leqslant m}$, let $|A|$ denote the determinant and $\mathrm{tr}(A)$ be the trace of A. Let $\|A\|_{\max} = \max_{i,j} |a_{ij}|$ denote the entry-wise max norm. Let $\|A\|_1 = \max_j \sum_{i=1}^m |a_{ij}|$ denote the matrix ℓ_1 norm. The Frobenius norm is given by $\|A\|_F^2 = \sum_i \sum_j a_{ij}^2$. Let $\varphi_i(A)$ denote the ith largest eigenvalue of A, with $\varphi_{\max}(A)$ and $\varphi_{\min}(A)$ denoting the largest and smallest eigenvalues, respectively, and let $\kappa(A)$ be the condition number for matrix A. Let $|A|_{1,\mathrm{off}} = \sum_{i \neq j} |a_{ij}|$ denote the sum of the absolute values of the off-diagonal entries and let $|A|_{0,\mathrm{off}}$ denote the number of non-zero off-diagonal entries. Let $a_{\max} = \max_i a_{ii}$. Denote by $r(A)$ the stable rank $\|A\|_F^2 / \|A\|_2^2$. We write $\mathrm{diag}(A)$ for a diagonal matrix with the same diagonal as A. Let I be the identity matrix. We let $C, C_1, c, c_1, \ldots$ be constants which may change from line to line. For two numbers a, b, $a \wedge b := \min(a, b)$ and $a \vee b := \max(a, b)$. Let $(a)_+ := a \vee 0$. We write $a = O(b)$ if $a \leqslant Cb$ for some positive absolute constants C which are independent of n and m or sparsity parameters and write $a \asymp b$ if $ca \leqslant b \leqslant Ca$. We write $a_n = o(b_n)$ if $\lim_{n \to \infty} a_n / b_n = 0$. For random variables X and Y, let $X \sim Y$ denote that X and Y follow the same distribution.

2 Joint mean and covariance modeling and estimation

In the previous section, we have not yet explicitly constructed an estimator of B^{-1}. To address this need, we model the data matrix X with a matrix-variate distribution with a separable covariance matrix, namely, the covariance of $\mathrm{vec}\{\mathrm{X}\}$ follows a Kronecker product covariance model. Let Z be the same as in Theorem 1. In the Kronecker product covariance model, the noise term has the form $\varepsilon = B^{1/2} Z A^{1/2}$ for a mean-zero noise matrix Z with independent entries such that $\mathrm{vec}\{\varepsilon\} = \mathrm{A} \otimes \mathrm{B}$; We focus on joint mean and covariance estimation for this setting in the present work, while we mention in passing that our methods can be generalized to other covariance models, see for example Cai et al. (2015).

Let S_A and S_B denote sample covariance matrices, and let the corresponding sample correlation matrices be calculated as

$$\widehat{\Gamma}_{ij}(A) = \frac{(S_A)_{ij}}{\sqrt{(S_A)_{ii}(S_A)_{jj}}} \quad \text{and} \quad \widehat{\Gamma}_{ij}(B) = \frac{(S_B)_{ij}}{\sqrt{(S_B)_{ii}(S_B)_{jj}}}. \tag{11}$$

The baseline Gemini estimators Zhou (2014a) are defined as follows, using a pair of penalized

estimators for the correlation matrices $\rho(A) = (a_{ij}/\sqrt{a_{ii}a_{jj}})$ and $\rho(B) = (b_{ij}/\sqrt{b_{ii}b_{jj}})$

$$\widehat{A}_\rho = \underset{A_\rho \succ 0}{\arg\min} \left\{ \operatorname{tr}\left(\widehat{\Gamma}(A)A_\rho^{-1}\right) + \log|A_\rho| + \lambda_B |A_\rho^{-1}|_{1,\text{off}} \right\}, \tag{12a}$$

$$\widehat{B}_\rho = \underset{B_\rho \succ 0}{\arg\min} \left\{ \operatorname{tr}\left(\widehat{\Gamma}(B)B_\rho^{-1}\right) + \log|B_\rho| + \lambda_A |B_\rho^{-1}|_{1,\text{off}} \right\}. \tag{12b}$$

where the input are a pair of sample correlation matrices as defined in (11).

Remarks.

1. When $M = 0$ is known, S_A and S_B can be the usual Gram matrices, and the theory in Zhou (2014a) guarantees that $t_{n,m}$ has rate $C_A\sqrt{\log(m)/m}$, with $C_A = \sqrt{m}\|A\|_F/\operatorname{tr}(A)$. However in our setting, M in general is nonzero. In Sections 2.1 and 2.2 we provide two constructions for S_A and S_B, which differ in how the data are centered. These constructions have a different bound $t_{n,m}$, as discussed below in more detail.

2. We note that even with perfect knowledge of M, $E[\varepsilon\varepsilon^T]$, and $E[\varepsilon^T\varepsilon]$, A and B are not identifiable, because for any $\eta > 0$, $A \otimes B = \eta A \otimes B/\eta$. However, this lack of identifiability does not affect the GLS estimate, because the GLS estimate is invariant to rescaling the estimate of B^{-1}. For identifiability, and convenience, we define

$$A^* = \frac{m}{\operatorname{tr}(A)}A \quad \text{and} \quad B^* = \frac{\operatorname{tr}(A)}{m}B, \tag{13}$$

   with the scaling chosen so that A^* has trace m. For the rest of the paper A and B refer to A^* and B^*, respectively.

3. When Z is i.i.d. Gaussian, ε follows a matrix-variate normal distribution $\mathcal{N}_{n,m}(0, A \otimes B)$, as considered in Zhou (2014a). In this case, the support of B^{-1} encodes conditional independence relationships between samples; Likewise, the support of A^{-1} encodes conditional independence relationships between genes. The inverse covariance matrices A^{-1} and B^{-1} have the same supports as their respective correlation matrices, so the edges of the dependence graph are identifiable under the model $\operatorname{Cov}(\operatorname{vec}(\varepsilon)) = A \otimes B$. Our results are more general in the sense that we analyze the subgaussian correspondent of the matrix variate normal model as intensively studied in Zhou (2014a).

We now define quantities related to the sample covariance matrices to be used in our Algorithms 1 and 2. Let $\widehat{M}$ denote the estimator of the mean matrix M in (1). Denote the centered data matrix

as

$$X_{\text{cen}} = X - \widehat{M}, \quad \text{for } \widehat{M} \text{ to be specified in Algorithms 1 and 2 below,} \tag{14}$$

and define the sample covariance matrices,

$$S_B = X_{\text{cen}} X_{\text{cen}}^T / m, \quad \text{and} \quad S_A = X_{\text{cen}}^T X_{\text{cen}} / n, \tag{15}$$

Define the diagonal matrices of sample standard deviations as

$$\widehat{W}_1 = \sqrt{n}\text{diag}(S_A)^{1/2} \in \mathbb{R}^{m\times m}, \quad \text{and} \quad \widehat{W}_2 = \sqrt{m}\text{diag}(S_B)^{1/2} \in \mathbb{R}^{n\times n}. \tag{16}$$

2.1 Group Based Centering Method

In this section we discuss our first method for estimation and inference with respect to the vector of mean differences $\gamma = \beta^{(1)} - \beta^{(2)}$, for $\beta^{(1)}$ and $\beta^{(2)}$ as in (5). Our approach here is to remove all possible mean effects by centering each variable within every group. The centered data matrix used to calculate S_A and S_B for Algorithm 1 is $X_{\text{cen}} = (I - P_2)X$, where P_2 is the projection matrix that performs within-group centering,

$$P_2 = \begin{bmatrix} n_1^{-1} 1_{n_1} 1_{n_1}^T & 0 \\ 0 & n_2^{-1} 1_{n_2} 1_{n_2}^T \end{bmatrix} = uu^T/n_1 + vv^T/n_2, \tag{17}$$

for u and v as defined in (4).

Algorithm 1: GLS-Global group centering

Input: X; and $\mathcal{G}(1), \mathcal{G}(2)$: indices of group one and two, respectively.

Output: $\widehat{A}^{-1}$, $\widehat{B}^{-1}$, $\widehat{A \otimes B}$, $\widehat{\gamma}$, T_j for all j

---

1. Group center the data. Let Y_i denote the ith row of the data matrix.

To estimate the group mean vectors $\beta^{(1)}, \beta^{(2)} \in \mathbb{R}^m$: Compute sample mean vectors

$$\widetilde{\beta}^{(1)} = \frac{1}{n_1} \sum_{i\in\mathcal{G}(1)} Y_i \quad \text{and} \quad \widetilde{\beta}^{(2)} = \frac{1}{n_2} \sum_{i\in\mathcal{G}(2)} Y_i; \quad \text{set} \quad \widehat{\gamma}^{\text{OLS}} = \widetilde{\beta}^{(1)} - \widetilde{\beta}^{(2)}. \tag{18}$$

Center the data by $X_{\text{cen}} = X - \widehat{M}$, with $\widehat{M} = \begin{bmatrix} 1_{n_1} \widetilde{\beta}^{(1)T} \\ 1_{n_2} \widetilde{\beta}^{(2)T} \end{bmatrix}$.

2. Obtain regularized correlation estimates. (2a) Compute sample covariance matrices based on group-centered data:

$$\begin{aligned} S_A &= X_{\text{cen}}^T X_{\text{cen}}/n = X^T(I-P_2)X/n \\ S_B &= X_{\text{cen}} X_{\text{cen}}^T/m = (I-P_2)XX^T(I-P_2)/m. \end{aligned}$$

Obtain estimates of the correlation matrices $\widehat{A}_\rho$ and $\widehat{B}_\rho$ using the Gemini estimators as defined in (12a) and (12b) with the tuning parameters that appear in (24) respectively.

(2b) Rescale the estimated correlation matrices to obtain covariance estimates,

$$\widehat{A \otimes B} = \left(\widehat{W}_A \widehat{A}_\rho \widehat{W}_A\right) \otimes \left(\widehat{W}_B \widehat{B}_\rho \widehat{W}_B\right) / \|X_{\text{cen}}\|_F^2, \quad \text{where} \tag{19}$$

$$\widehat{B}^{-1} = m\widehat{W}_2^{-1} \widehat{B}_\rho \widehat{W}_2^{-1} \tag{20}$$

$$\text{and} \quad \widehat{A}^{-1} = (\|X_{\text{cen}}\|_F^2/m)\widehat{W}_1^{-1} \widehat{A}_\rho \widehat{W}_1^{-1}. \tag{21}$$

4. Estimate the mean matrix. Estimate the group mean matrix using the GLS estimator as defined in (6).

5. Obtain test statistics. The jth test statistic is defined as

$$T_j = \frac{\widehat{\gamma}_j(\widehat{B}^{-1})}{\sqrt{\delta^T (D^T \widehat{B}^{-1} D)^{-1} \delta}}, \qquad \text{with } \delta = (1,-1) \in \mathbb{R}^2, \tag{22}$$

and $\widehat{\gamma}_j(\widehat{B}^{-1}) = \delta^T \widehat{\beta}_j(\widehat{B}^{-1})$, for $j = 1, \ldots, m$. Note that T_j as defined in (22) is essentially a Wald test and the denominator is a plug-in standard error of $\widehat{\gamma}_j(B^{-1})$.

2.2 Model Selection Centering Method

In this section we present an alternative algorithm, which aims to remove only the mean effects that are strong enough to have an impact on covariance estimation. In the case that γ is sparse, we anticipate that this approach could perform (much) better than the approach in Section 2.1. The strategy here is to use a preliminary model selection step to identify the variables with strong mean effects.

Algorithm 2: GLS-Model selection centering

Input: X, and $\mathcal{G}(1), \mathcal{G}(2)$: indices of group one and two, respectively.

Output: $\widehat{A}^{-1}$, $\widehat{B}^{-1}$, $\widehat{A \otimes B}$, $\widehat{\gamma}$, T_j for all j

---

1. **Run Algorithm 1.** Use the group centering method to obtain initial estimates $\widehat{\gamma}_j^{\text{init}} = \widehat{\beta}_j^{(1)} - \widehat{\beta}_j^{(2)}$ for all $j = 1, \ldots, m$.

2. **Select genes with large estimated differences in means.** Let $\widetilde{J}_0 = \{j : |\widehat{\gamma}_j^{\text{init}}| > 2\tau_{\text{init}}\}$ denote the set of genes which we consider as having strong mean effects, where

$$\tau_{\text{init}} \asymp \left(\frac{\log^{1/2}(m)}{\sqrt{m}} + \frac{\|B\|_1}{n_{\min}} \right) \sqrt{\frac{n_{\text{ratio}} \left(|B^{-1}|_{0,\text{off}} \vee 1 \right)}{n_{\min}}} + \sqrt{\log(m)} \|(D^T B^{-1} D)^{-1}\|_2^{1/2}. \quad (23)$$

3. **Calculate Gram matrices based on model selection centering.** Global centering can be expressed in terms of the projection matrix $P_1 = n^{-1} \mathbf{1}_n \mathbf{1}_n^T$. Compute the centered data matrix as

$$X_{\text{cen},j} = \begin{cases} X_j - P_2 X_j & \text{if } j \in \widetilde{J}_0 \\ X_j - P_1 X_j & \text{if } j \in \widetilde{J}_0^c, \end{cases}$$

where $X_{\text{cen},j}$ is the jth column of X_{cen}. Compute the sample covariance matrices with the centered data matrix as in (15).

4. **Estimate covariances and means.** [(a)]

   Obtain estimates of the correlation matrices $\widehat{B}_\rho$ and $\widehat{A}_\rho$ using Gemini estimators as defined in (12a) and (12b) with the tuning parameters of the same order as those in (24).

   2. Obtain inverse covariance estimates $\widehat{B}^{-1}$, $\widehat{A}^{-1}$ as in (20), (21).
   3. Calculate the GLS estimator $\widehat{\beta}(\widehat{B}^{-1})$ as in (6), as well as the vector of mean differences $\widehat{\gamma}(\widehat{B}^{-1}) = \delta^T \widehat{\beta}(\widehat{B}^{-1})$, for $\delta = (1, -1) \in \mathbb{R}^2$.

5. **Obtain test statistics.** Calculate test statistics as in (22), now using $\widehat{B}^{-1}$ as estimated in Step 4.

Although (23) gives the theoretical threshold, in reality we do not know the values of the parameters that depend on B, so in simulations we select a set of genes $\breve{J}_0$ to group center such that $|\breve{J}_0| = d'$, where d' is understood to be an upper bound on $d_0 = |\operatorname{supp}(\gamma)|$, for example, chosen in advance based on prior knowledge. We select $\breve{J}_0$ by ranking the components of the estimated vector of mean differences $\widehat{\gamma}$. In the data analysis we do this in an iterative manner by successively

halving the number of selected genes, choosing at each step the genes with largest estimated mean differences.

3 Theoretical results

We provide a theorem stating the rates of convergence for estimating the group mean matrix $\beta \in \mathbf{R}^{2\times m}$ and the covariance matrices A, B, and their inverses simultaneously. In part I of Theorem 2, we state the rates of convergence of the Gemini estimators of B^{-1} and A^{-1} when the input sample covariance matrix uses group centering as defined in Algorithm 1. In part II, we state the rate of convergence of the corresponding GLS estimator of $\{\beta_j^*\}$. We state the following assumptions.

(A1) The number of nonzero off-diagonal entries of A^{-1} and B^{-1} satisfy

$$\begin{aligned} \left|A^{-1}\right|_{0,\text{off}} &= o(n/\log(m\vee n)) \qquad\qquad (n,m\to\infty) \quad \text{and} \\ \left|B^{-1}\right|_{0,\text{off}} &= o\left([m/\log(m\vee n)]\vee\left[n^2/\|B\|_1^2\right]\right) \qquad (n,m\to\infty). \end{aligned}$$

(A2) (A2) The eigenvalues of A and B are bounded away from 0 and $+\infty$. We assume that the stable ranks satisfy $r(A), r(B) \geqslant 4\log(m\vee n)$. Let $C_A = \sqrt{m}\|A\|_F/\operatorname{tr}(A)$ and $C_B = \sqrt{n}\|B\|_F/\operatorname{tr}(B)$.

Theorem 2. ***(I)*** *Suppose that (A1) and (A2) hold. Consider the data as generated from model* (3) *with* $\varepsilon = B^{1/2}ZA^{1/2}$, *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, satisfying* $\mathbb{E}Z_{ij}=0$, $1=\mathbb{E}Z_{ij}^2\leqslant\|Z_{ij}\|_{\psi_2}\leqslant K$. *Let* λ_A *and* λ_B *denote the penalty parameters for* (12b) *and* (12a) *respectively. Suppose*

$$\lambda_A \geqslant C\left(C_A K\frac{\log^{1/2}(m\vee n)}{\sqrt{m}}+\frac{\|B\|_1}{n_{\min}}\right) \quad \textit{and} \quad \lambda_B \geqslant C'\left(C_B K\frac{\log^{1/2}(m\vee n)}{\sqrt{n}}+\frac{\|B\|_1}{n_{\min}}\right) \tag{24}$$

for some absolute constants C *and* C'. *Then with probability at least* $1-C''/(m\vee n)^2$,

$$\begin{aligned} &\|\widehat{A\otimes B}-A\otimes B\|_2 \leqslant \|A\|_2\|B\|_2\delta, \\ &\|\widehat{A\otimes B}^{-1}-A^{-1}\otimes B^{-1}\|_2 \leqslant \|A^{-1}\|_2\|B^{-1}\|_2\delta', \\ \textit{where}\qquad &\delta,\delta' = O\left(\lambda_A\sqrt{|B^{-1}|_{0,\textit{off}}\vee 1}+\lambda_B\sqrt{|A^{-1}|_{0,\textit{off}}\vee 1}\right). \end{aligned}$$

Furthermore, with probability at least $1 - C'''/(m \vee n)^2$,

$$\begin{aligned} &\|\widehat{A \otimes B} - A \otimes B\|_F \leqslant \|A\|_2 \|B\|_2 \eta, \\ &\|\widehat{A \otimes B}^{-1} - A^{-1} \otimes B^{-1}\|_F \leqslant \|A^{-1}\|_2 \|B^{-1}\|_2 \eta', \\ \textit{where} \qquad &\eta, \eta' = O\left(\lambda_A \sqrt{|B^{-1}|_{0,off} \vee n}/\sqrt{n} + \lambda_B \sqrt{|A^{-1}|_{0,off} \vee m}/\sqrt{m}\right). \end{aligned}$$

(II) *Let* $\widehat{\beta}$ *be defined as in (6) with* $\widehat{B}^{-1}$ *being defined as in* (20) *and* D *as in (5). Then with probability at least* $1 - C/m^d$, *for all* j,

$$\|\widehat{\beta}_j(\widehat{B}^{-1}) - \beta_j^*\|_2 \leqslant C_1 \lambda_A \sqrt{\frac{n_{ratio}\left(|B^{-1}|_{0,off} \vee 1\right)}{n_{\min}}} + C_2 \sqrt{\log(m)} \|(D^T B^{-1} D)^{-1}\|_2^{1/2}. \tag{25}$$

We prove Theorem 2 part I in Section A; this relies on rates of convergence of $\widehat{B}^{-1}$ and $\widehat{A}^{-1}$ in the operator and the Frobenius norm, which are established in Lemma 5. We prove part II in Section A.2.

Remarks. We find that the additional complexity of estimating the mean matrix leads to an additional additive term of order $1/n$ appearing in the convergence rate for covariance estimation for B and A. In part I of the main theorem, λ_A is decomposed into two terms, one term reflecting the variance of S_B, and one term reflecting the bias due to group centering. The variance term goes to zero as m increases, and the bias term goes to zero as n increases. To analyze the error in the GLS estimator based on $\widehat{B}^{-1}$, we decompose $\|\widehat{\beta}_j(\widehat{B}^{-1}) - \beta_j^*\|_2$ as

$$\|\widehat{\beta}_j(\widehat{B}^{-1}) - \beta_j^*\|_2 \leqslant \|\widehat{\beta}_j(\widehat{B}^{-1}) - \widehat{\beta}_j(B^{-1})\|_2 + \|\widehat{\beta}_j(B^{-1}) - \beta_j^*\|_2, \tag{26}$$

where the first term is the error due to not knowing B^{-1}, and the second term is the error due to not knowing β_j^*. The rate of convergence given in (25) of the main theorem reflects this decomposition. For Algorithm 2, we have analogous rates of convergence for both mean and covariance estimation. Simulations suggest that the constants in the rates of convergence of Algorithm 2 are smaller. We aim to provide a sharper analysis for Algorithm 2 in our future work.

4 Simulations

We present simulations to compare Algorithms 1 and 2 to analysis based on the sample means and to oracle algorithms based on knowledge of the true correlation structures A and B. We use several population structures.

4.1 Notation and Models

We construct covariance matrices for A and B from one of:

- AR1(ρ) model. The covariance matrix is of the form $B = \{\rho^{|i-j|}\}_{i,j}$, and the graph corresponding to B^{-1} is a chain.
- Star-Block model. The covariance matrix is block-diagonal with equal-sized blocks whose inverses correspond to star structured graphs, where $B_{ii} = 1$, for all i. In each subgraph, a central hub node connects to all other nodes in the subgraph, with no additional edges. The covariance matrix for each block S in B is generated as in Ravikumar et al. (2011): $S_{ij} = \rho = 0.5$ if $(i, j) \in E$ and $S_{ij} = \rho^2$ otherwise.
- Erdős-Rényi model. We use use the random concentration matrix model in Zhou et al. (2010). The graph is generated according to a type of Erdős-Rényi random graph. Initially we set $B^{-1} = 0.25 I_{n\times n}$. Then, we randomly select d edges and update B^{-1} as follows: for each new edge (i, j), a weight $w > 0$ is chosen uniformly at random from $[w_{\min}, w_{\max}]$ where $w_{\min} = 0.6$ and $w_{\max} = 0.8$; we subtract w from B^{-1}_{ij} and B^{-1}_{ji}, and increase B^{-1}_{ii} and B^{-1}_{jj} by w. This keeps B^{-1} positive definite. We then rescale so that B^{-1} is an inverse correlation matrix.

4.2 Accuracy of $\widehat{\gamma}$ and its implication for variable ranking

Table 1 displays metrics that reflect the difficulty of the different population structures. Column 2 is a measure discussed by Efron (2007). Column 3 appears directly in the theoretical analysis, reflecting the entry-wise error in the sample correlation $\widehat{\Gamma}(B)$. Columns 4 and 5 analogously reflect the entry-wise error for the Flip-Flop procedure in (Zhou, 2014a), and are included here for completeness. Column 6 displays what we call the standard deviation ratio, namely

$$\sqrt{\frac{u^T B u}{\delta^T (D^T B^{-1} D)^{-1} \delta}}, \tag{27}$$

where $u = (\underbrace{1/n_1, \ldots, 1/n_1}_{n_1}, \underbrace{-1/n_2, \ldots, -1/n_2}_{n_2}) \in \mathbb{R}^n$ and $\delta = (1, -1) \in \mathbb{R}^2$, which reflects the potential efficiency gain for GLS over sample-mean based method (18) for estimating γ. The values in Column 6 show that substantial improvement is possible in mean estimation. For an AR1 covariance matrix, the standard deviation ratio increases as the AR1 parameter increases; as the correlations get stronger, the potential improvement in mean estimation due to GLS grows. For the Star Block model with fixed block size, the standard deviation ratio is stable as n increases.

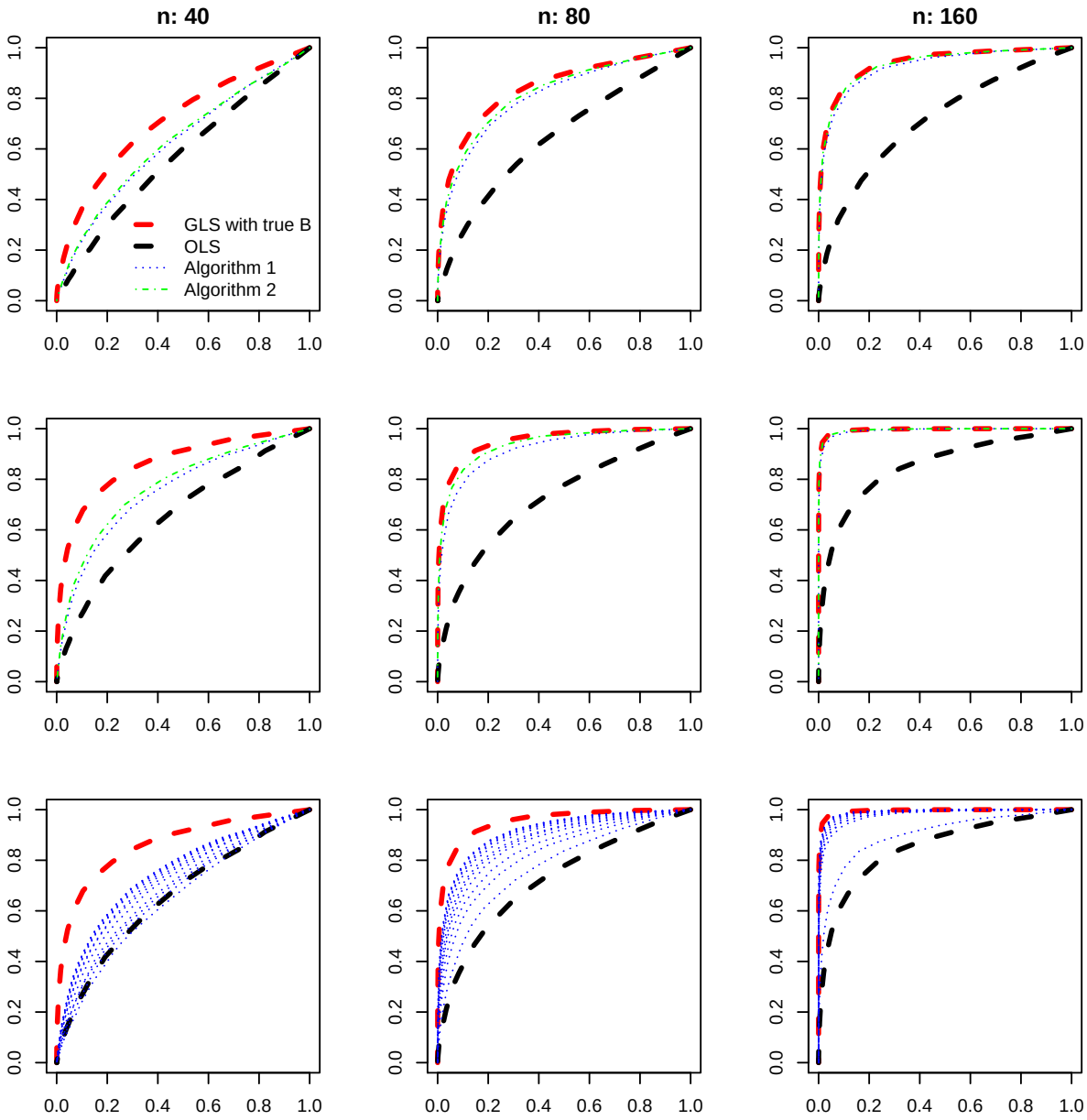

Figure 1: ROC curves. For each plot, the horizontal axis is FPR and the vertical axis is TPR, as we vary a threshold for classifying variables as null or non-null. The covariance matrices A and B are both AR1 with parameter 0.8, with $m = 2000$ and $n = 40$, 80, and 160 in column one, two, and three, respectively. Ten variables in γ have nonzero entries. On each trial, the group labels are randomly assigned, with equal sample sizes. The marginal variance of each entry of the data matrix is equal to one. For the first row of plots, the magnitude of each nonzero entry of γ is 0.2, and for the second and third rows of plots, the magnitude of each nonzero entry of γ is 0.3. In the first two rows we display ROC curves for Algorithms 1 and 2 with penalty parameters chosen to maximize area under the curve. The third row displays an ROC curves for Algorithm 1, sweeping out penalty parameters.

	B	ρ_B^2	$\|B\|_F/\mathrm{tr}(B)$	$\left\|\rho(B)^{-1}\right\|_{1,\mathrm{off}}$	$\left\|\rho(B)^{-1}\right\|_1$	sd ratio
$n = 80$						
1	AR1(0.2)	0.00	0.12	32.92	119.50	1.00
2	AR1(0.4)	0.00	0.13	75.24	185.33	1.02
3	AR1(0.6)	0.01	0.16	148.12	317.00	1.07
4	AR1(0.8)	0.04	0.24	351.11	712.00	1.32
5	StarBlock(4, 20)	0.02	0.18	101.33	232.00	1.51
6	ER(0.6, 0.8)	0.01	0.14	92.75	217.57	1.21
$n = 40$						
1	AR1(0.2)	0.00	0.16	16.25	59.50	1.01
2	AR1(0.4)	0.01	0.19	37.14	92.00	1.03
3	AR1(0.6)	0.03	0.23	73.12	157.00	1.12
4	AR1(0.8)	0.08	0.33	173.33	352.00	1.47
5	StarBlock(2, 20)	0.04	0.25	50.67	116.00	1.51
6	ER(0.6, 0.8)	0.02	0.21	47.24	110.32	1.23

Table 1: Assessment of the difficulty of estimating B^{-1} and the potential gain from GLS. The total correlation ρ_B is the average squared off-diagonal value of the correlation matrix $\rho(B)$. The last column (sd ratio) presents the ratio of the standard deviation of the difference in sample means in the sample mean based method (18) to the standard deviation of the GLS estimator of the difference in means. The first four columns of the table reflect the difficulty of estimating B, whereas the last column reflects the potential improvement of GLS over the sample mean based method (18).

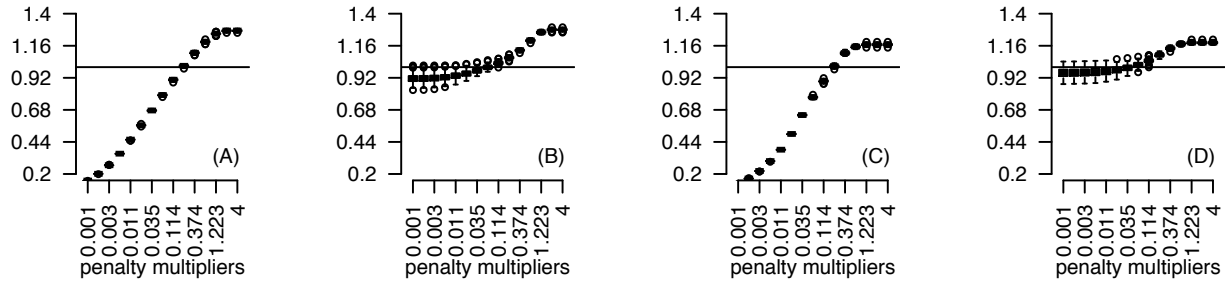

Figure 2: Ratio of estimated design effect to true design effect when B^{-1} is Erdős-Rényi, and A is AR1(0.8). Figures (A) and (B) correspond to sample size $n = 80$; (C) and (D) correspond to $n = 40$. Figures (A) and (C) correspond to Algorithm 1; Figures (B) and (D) correspond to Algorithm 2, with ten columns group centered. These results are based on dimension parameter $m = 2000$ and 250 simulation replications.

In our first experiment, we set the population covariance matrices A and B as AR1(0.8), with $m = 2000$, and $n = 40$, 80, and 160. In Figure 1, we use ROC curves to illustrate the sensitivity and specificity for variable selection in the sense of how well we can identify the support for $\{i : \gamma_i \neq 0\}$ when we threshold $\widehat{\gamma}_i$ at various values: $\widehat{\gamma}$ is the output of Algorithm 1, Algorithm 2, the oracle GLS, and the the sample mean based method (18), respectively. This corresponds to the four curves on each plot of the top two rows of plots. We compare the results for Algorithms 1 and 2

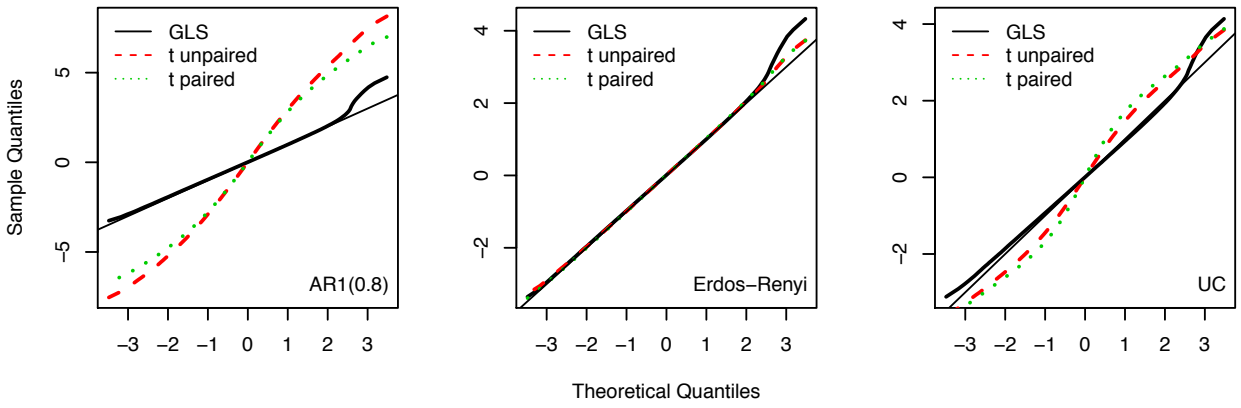

Figure 3: Quantile plots of test statistics. Ten genes have nonzero mean differences equal to 2, 0.8, and 1 in the three plots, respectively. In each plot A is AR1(0.8). Covariance structures are as indicated. In the third plot, the true B is set to $\widehat{B}$ for the ulcerative colitis data, described in Section 5. For the first two plots there are $n = 40$ samples and $m = 2000$ variables. For the third plot there are $n = 20$ samples and $m = 2000$ variables. Each plot has 250 simulation replications.

to the results of the oracle GLS and the sample mean based method (18). We find that Algorithm 1 and Algorithm 2 perform better than the sample mean based method (18), and in some cases perform comparably to the oracle GLS. Plots in the third row of Figure 1 illustrate the sensitivity of Algorithm 1 to the choice of GLasso penalty parameter; the performance can degenerate to that of the sample mean based method (18), if the penalty is too high.

4.3 Inference for the mean difference $\widehat{\gamma}$

Two basic approaches to conducting inferences for mean differences are paired and unpaired t statistics. The unpaired t statistic is defined as follows. Let $X = (X_{ij})$. Then the jth unpaired t statistic is

$$T_j = \left(\widetilde{\beta}_j^{(1)} - \widetilde{\beta}_j^{(2)}\right) \widehat{\sigma}_j^{-1} (n_1^{-1} + n_2^{-1})^{-1/2}, \tag{28}$$

where

$$\widehat{\sigma}_j^2 = (n_1 + n_2 - 2)^{-1} \sum_{k=1}^{2} \sum_{i \in \mathcal{G}_k} \left(X_{ij} - \widetilde{\beta}_j^{(k)}\right)^2,$$

where $\widetilde{\beta}_j^{(k)}$, $k = 1, 2$, and $j = 1, \ldots, m$, denotes the sample mean of group k and variable j as defined in (18), and $\mathcal{G}_k$ is the set of indices corresponding to group k. When there is a natural basis for pairing the observations, and paired units are anticipated to be positively correlated, we can calculate paired t statistics. For the paired t statistic, suppose observations i and $i' = i + n/2$ are paired, for $i \in \{1, \ldots, n/2\}$. Note that samples can always be permuted so as to be paired in this

way. Define the paired differences $d_{ij} = X_{ij} - X_{i'j}$, for $i \in \{1, \ldots, n/2\}$. Then the paired t statistic is $\overline{d}_j (n/2-1)^{1/2} / \left(\sum_{i=1}^{n/2} (d_{ij} - \overline{d}_j)^2 \right)^{1/2}$, where $\overline{d}_j = (n/2)^{-1} \sum_{i=1}^{n/2} d_{ij}$.

Figure 2 considers estimation of the "design effect," defined as

$$\delta^T (D^T B^{-1} D)^{-1} \delta, \tag{29}$$

with $\delta = (1, -1)^T$. The estimated design effect appears in the GLS estimate of β (6), and also is a scale factor in the test statistics for $\widehat{\gamma}$ (22). The design effect is estimated via $\delta^T (D^T \widehat{B}^{-1} D)^{-1} \delta$, with $\widehat{B}^{-1}$ from Algorithm 1 or 2. The GLasso penalty parameters are chosen as

$$\lambda_A = f_A \left(C_A K \frac{\log^{1/2}(m \vee n)}{\sqrt{m}} + \frac{\|B\|_1}{n_{\min}} \right) \tag{30}$$

where we sweep over the factor f_A, referred to as the penalty multiplier. Figure 2 displays boxplots of the ratio

$$\delta^T (D^T \widehat{B}^{-1} D)^{-1} \delta / \delta^T (D^T B^{-1} D)^{-1} \delta,$$

over 250 replications for each setting of the penalty multiplier f_B. In Figure 2, B^{-1} follows the Erdős-Rényi model, and A is AR1(0.8), with $m = 2000$, and $n = 40$ and 80. Figure 2 shows that Algorithm 2 (plots B and D) estimates the design effect to high accuracy and is quite insensitive to the penalty multiplier as long as it is less than 1, as predicted by the theoretical analysis. Algorithm 1 also estimates the design effect with high accuracy, but with somewhat greater sensitivity to the tuning parameter. The best penalty parameter for Algorithm 1 is around 0.1, whereas reasonable penalty parameters for Algorithm 2 are in the range 0.01 to 0.1. This is consistent with smaller entrywise error in the sample covariance for model selection centering than for group centering.

We next compare the results from Algorithm 2 to results obtained using paired and unpaired t statistics. Figure 3 illustrates the calibration and power of plug-in Z-scores, $\widehat{\gamma}_j / \widehat{\mathrm{SE}}(\widehat{\gamma}_j)$ derived from Algorithm 2 for three population settings. The standard error is calculated as $\sqrt{\delta^T (D^T \widehat{B}^{-1} D)^{-1} \delta}$, with $\delta = (1, -1)$, so depends directly on our ability to estimate B. In the first and second plots, the data was simulated from AR1(0.8) and Erdős-Rényi, respectively. In the third plot, the data was simulated from $\widehat{B}$ for ulcerative colitis data described in Section 5. To obtain $\widehat{B}$, we apply Algorithm 2 to the ulcerative colitis data, using a Glasso penalty of $\lambda \approx 0.5[(\log(m)/m) + 3/n]$ in step 1, followed by group centering the top ten genes in step 2, and using a Glasso penalty of $\lambda \approx 0.1[(\log(m)/m) + 3/n]$ in step 4. In all cases A is AR1(0.8). In each case, we introduced 10 variables with different population means in the two groups, by setting $\gamma = 0.8$ for those variables,

with the remaining γ values equal to zero. The ideal Q-Q plot would follow the diagonal except at the upper end of the range, as does our plug-in GLS test statistic. The t statistics (ignoring dependence) are seen to be overly dispersed throughout the range, and are less sensitive to the real effects.

4.4 Covariance Estimation

Figure 4 displays the RMSE for estimating γ as well as the relative Frobenius error for estimating B^{-1} as a function of the sample size parameter n. The population structures for B are Erdős-Rényi and Star Block. The figure illustrates that Algorithm 2 can outperform both Algorithm 1 and the sample mean based method (18). Figure 5 shows the relative Frobenius error in estimating A^{-1} as n grows, for fixed m. The horizontal axis is $n/\log(m)$, scaled so that the curves align. Because $\|A^{-1}\|_F$ is of order $\sqrt{m}$, the vertical axis essentially displays $\|\widehat{A}^{-1} - A^{-1}\|_F/\sqrt{m}$. For estimating A^{-1}, the rate of convergence is of order $\sqrt{\log(m)/n}$. For each of the three population structures, accuracy increases with respect to n.

5 Genomic Study of Ulcerative Colitis

Ulcerative colitis (UC) is a chronic form of inflammatory bowel disease (IBD), resulting from inappropriate immune cell infiltration of the colon. As part of an effort to better understand the molecular pathology of UC, Lepage et al. (2011) reported on a study of mRNA expression in biopsy samples of the colon mucosal epithelium, with the aim being to identify gene transcripts that are differentially expressed between people with UC and healthy controls. The study subjects were discordant identical twins, that is, monozygotic twins such that one twin has UC and the other does not. This allows us to explore dependences among samples (both within and between twins), along with dependences among genes and mean differences between the UC and non-UC subjects.

The data consist of 10 discordant twin pairs, for a total of 20 subjects. Each subject's biopsy sample was assayed for mRNA expression, using the Affymetrix UG 133 Plus 2.0 array, which has 54,675 distinct transcripts. Previous analyses of this data did not consider twin correlations or unanticipated non-twin correlations, and used very different methodology (e.g. Wilcoxon testing). Roughly 70 genes were found to be differentially expressed (Lepage, et al. 2011).

We applied our Algorithm 2 to the UC genomics data as follows. First we selected the 2000 most variable genes based on marginal variance and then rescaled each gene to have unit marginal

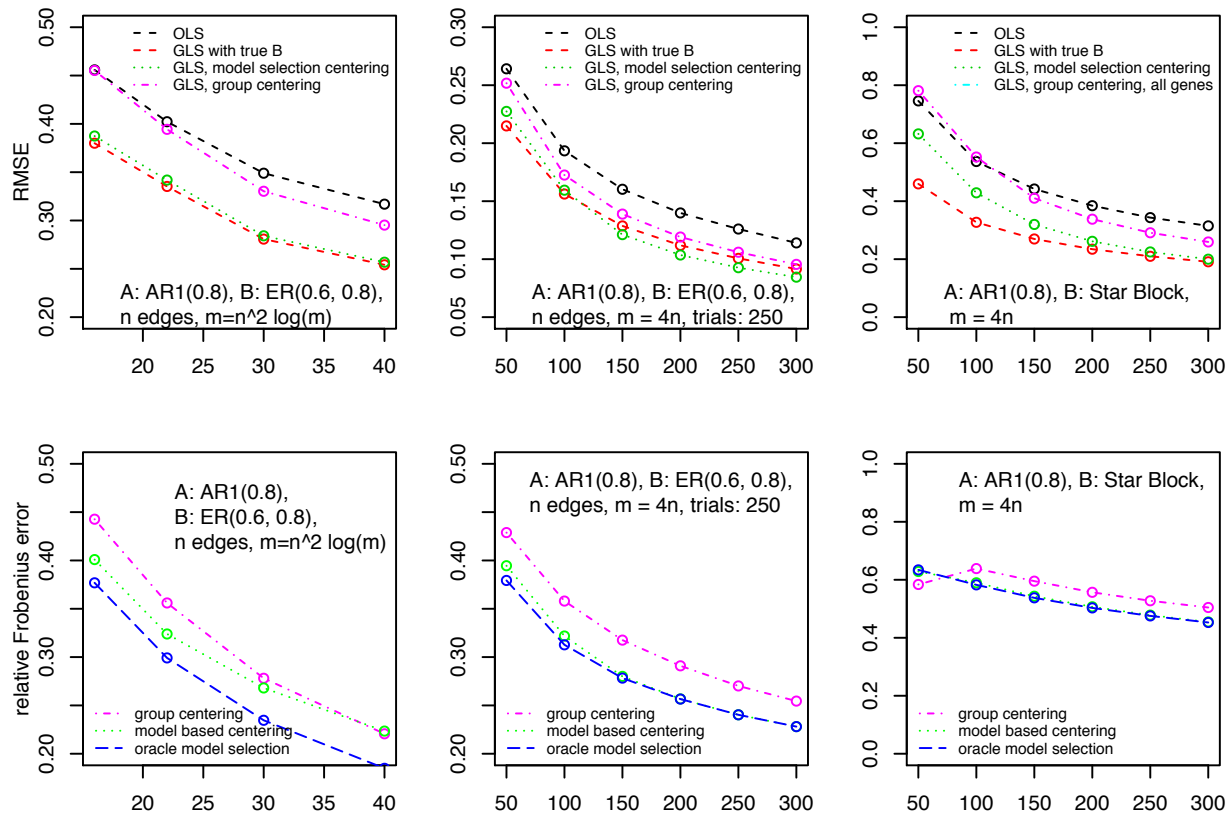

Figure 4: In this figure, we compare centering methods as we vary n and m, with n shown on the horizontal axis. In the first column of plots, the number of edges is proportional to $\sqrt{m/\log(m)}$. In the second and third columns of plots, the number of edges is proportional to m. In the first two columns of plots, B^{-1} is an Erdős-Rényi inverse covariance matrix. In the third column, B^{-1} is star block. The first row of plots shows RMSE for estimating γ, whereas the second row shows average relative Frobenius error in estimating B^{-1}.

variance. We then applied step 1 of Algorithm 2, setting

$$\lambda = 0.1 \approx 0.5 \left(\sqrt{\frac{\log(m)}{m}} + \frac{3}{n} \right)$$

with $m = 2000$ and $n = 20$. For step 2 of the algorithm, we ranked the genes by estimated difference in means, group centered the top ten, and globally centered the remaining genes. We then re-calculated the Gram matrix S_B using the centered data. In step 3, following the Gemini approach, we applied the GLasso to S_B using a regularization parameter $\lambda \approx 0.25(\sqrt{\log(m)/m} + 3/n)$. We obtain estimated differences in means and test statistics via steps 4 through 6.

A natural analysis of these data using more standard methods would be a paired t-test for each mRNA transcript (paired by twin pair). Such an approach is optimized for the situation where

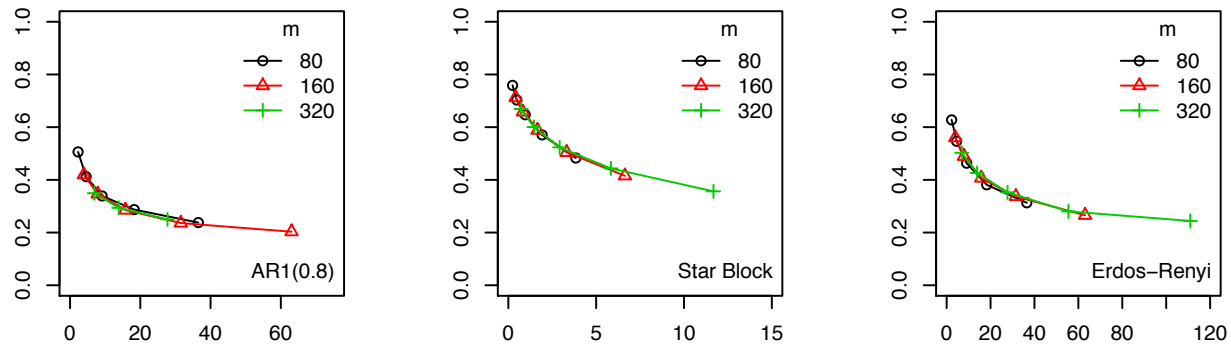

Figure 5: Relative Frobenius error in estimating A^{-1}, as n varies. In each plot the matrix B is AR1(0.8). The vertical axis is relative Frobenius error, and the horizontal axis $n/(d\log(m))$, where d is the maximum node degree. The GLasso penalty is chosen to minimize the relative Frobenius error. Each point is based on 250 Monte Carlo replications.

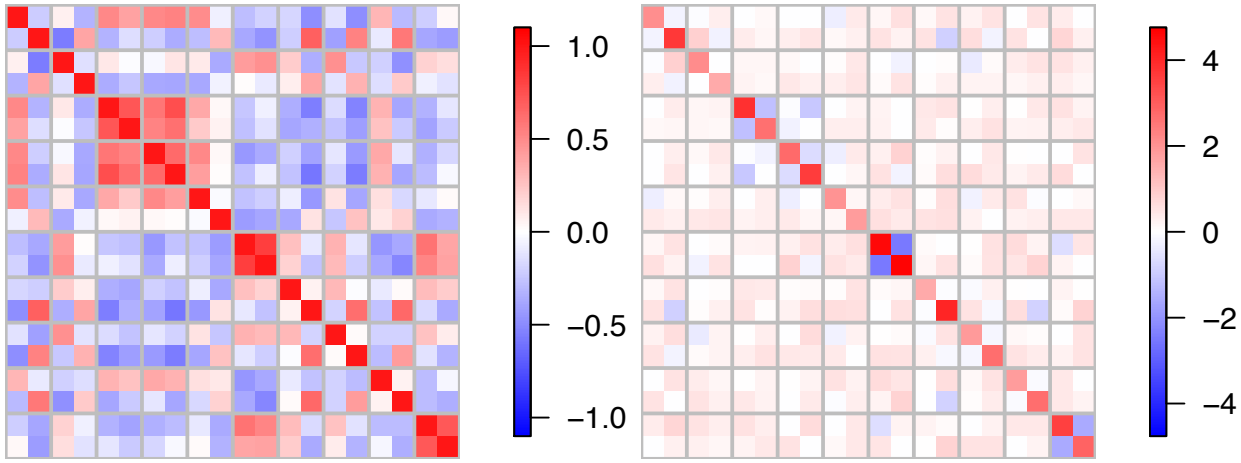

Figure 6: This figure displays estimated person-person correlation matrix and its inverse, estimated using the 2000 genes with largest marginal variance.

there is a constant level of correlation within all of the twin pairs, with no non-twin correlations. However as in Efron (2008), we wish to accommodate unexpected correlations, which in this case would be correlations between non-twin subjects or a lack of correlation between twin subjects. Our approach, developed in section 1.2, does not require pre-specification or parameterization of the dependence structure, thus we were able to consider twin and non-twin correlations simultaneously. Lepage et al. note that UC has lower heritability than other forms of IBD. If UC has a relatively stronger environmental component, this could explain the pattern of correlations that we uncovered.

We also found only a small amount of evidence for differential gene expression between the UC and non-UC subjects. For Algorithm 2, four of the adjusted p-values fell below a threshold of 0.1,

using the Benjamini-Hochberg adjustment; that is, four genes satisfied $2000\widehat{p}_{(i)}/i < 0.1$, where $\widehat{p}_{(i)}$ is the ith order statistic of the p-values calculated using Algorithm 2, for $i = 1, \ldots, 2000$. Based on our theoretical and simulation work showing that our procedure can successfully recover and accommodate dependence among samples, we argue that this is a more meaningful representation of the evidence in the data for differential expression compared to methods that do not adapt to dependence among samples. In particular, below we argue that the sample-wise correlations detected by our approach would be expected to artificially inflate the evidence for differential expression.

5.1 Calibration of test statistics

As noted above, based on the test statistics produced by Algorithm 2, we find evidence for only a small number of genes being differentially expressed. This conclusion, however, depends on the test statistics conforming to the claimed null distribution whenever the group-wise means are equal. In this section, we consider this issue in more detail.

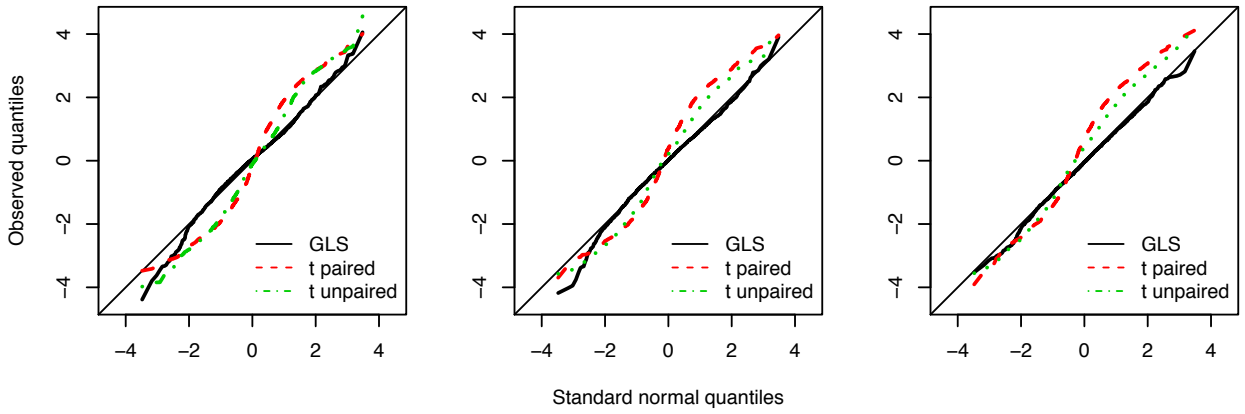

Figure 7: Quantile plots of test statistics for three disjoint gene sets, each consisting of 2000 genes. The genes are partitioned based on marginal variance. GLS statistics are taken from step 5 of Algorithm 2; in step 2, the ten genes with greatest mean differences are selected for group centering.

The first plot of Figure 7 compares the empirical quantiles of $\Phi^{-1}(T_j)$ to the corresponding quantiles of a standard normal distribution, where Φ is the standard normal cdf and the T_j are as defined in (28). Plots 2 and 3 show the same information for successive non-overlapping blocks of two thousand genes sorted by marginal variance. Since this is a discordant twins study, we also show results for the standard paired t statistics, pairing by twin. In all cases, the paired and unpaired statistics are more dispersed relative to the reference distribution. By contrast, the central

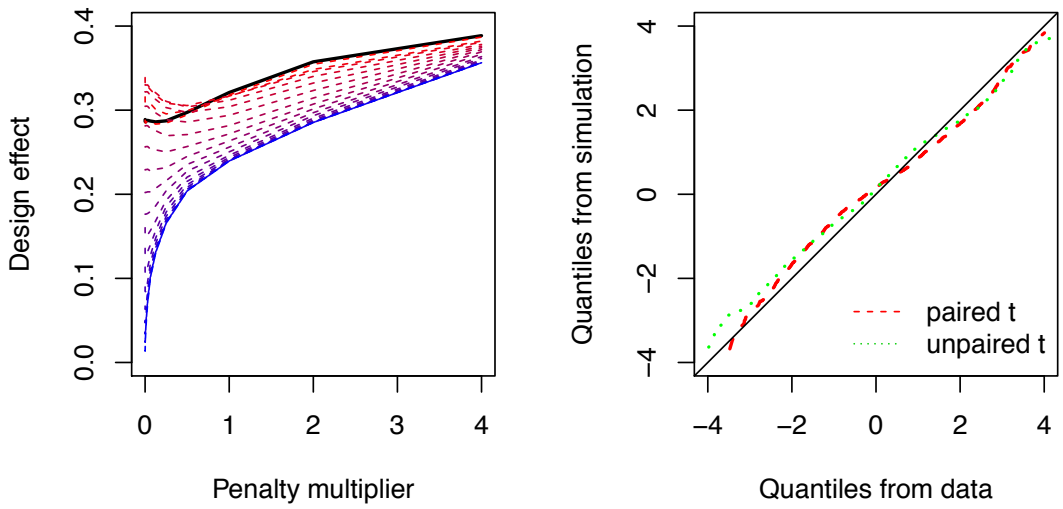

Figure 8: The first plot displays the estimated design effect vs. the penalty multiplier for Algorithm 2. Each curve corresponds to a different number of columns being group centered. As the curves go from top to bottom, the number of group centered columns increases from 10 to 2000. The second plot 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.

portion of the GLS test statistics coincide with the reference line. Overdispersion of test statistics throughout their range is often taken to be evidence of miscalibration (Devlin and Roeder, 1999). In this setting the GLS statistics are calibrated correctly under the null hypothesis, but the paired and unpaired t statistics are not.

Since the design effect (29) appears as a scaling factor in the test statistics (22), it is particularly important that the design effect is accurately estimated in order for the test statistics to be property calibrated. The first plot of Figure 8 displays the sensitivity of the estimated design effect (29) 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 vector γ may be sparse. The second plot of Figure 8 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.

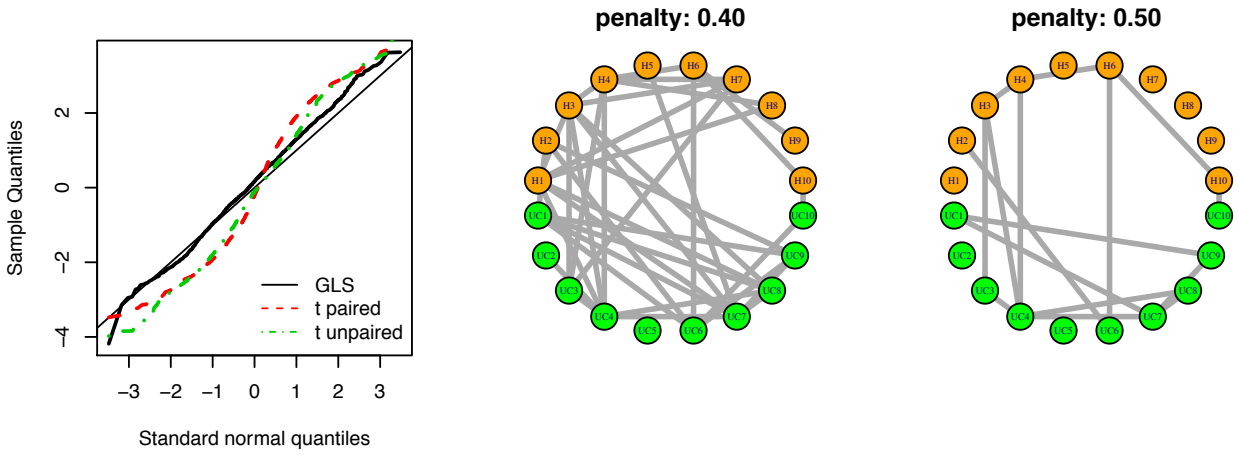

Figure 9: Quantile plot and inverse covariance graphs. The first two plots correspond to $\lambda = 0.4$ and 128 group centered genes. The third plot corresponds to $\lambda = 0.5$ and 128 group centered genes. Green circles correspond to twins with UC, orange circles to twins without UC. Twins are aligned vertically.

5.2 Stability of gene sets

The motivation of our Algorithm 2 is that in many practical settings a relatively small fraction of variables may have differential means, and therefore it is advantageous to avoid centering variables presenting no evidence of a strong mean difference. Here we assess the stability of the estimated mean differences as we vary the number of group centered genes in Algorithm 2. To do so, we successively group center fewer genes, globally centering the remaining genes.

The iterative process is as follows. Let $\widehat{B}^{-1}_{(i)} \in \mathbb{R}^{n\times n}$ denote the estimate of B^{-1} at iteration i, let $\widehat{\beta}_{(i)} \in \mathbb{R}^{2\times m}$ denote the estimates of the group means β on the ith iteration, let $\widehat{\gamma}_{(i)} \in \mathbb{R}^m$ denote the vector of differences in group means between the two groups, and let $\widehat{\mu}_{(i)} \in \mathbb{R}^m$ denote vector of global mean estimates. Let $\widehat{\mu}(B^{-1}) \in \mathbb{R}^m$ denote the result of applying GLS with design matrix $D = 1_n$ to estimate the global means.

Initialize $\widehat{\beta}_{(1)}$, $\widehat{\mu}_{(1)}$ and $\widehat{\gamma}_{(1)}$ using the sample means. On the ith iteration,

1. Rank the genes according to $|\widehat{\gamma}_{(i-1)}|$. Center the highest ranked n_i' genes around $\widehat{\beta}_{(i-1)}$. Center the remaining genes around $\widehat{\mu}_{(i-1)}$.

2. Obtain $\widehat{B}^{-1}_{(i)}$ by applying GLasso to the centered data matrix from step 1.

3. Set $\widehat{\beta}_{(i)} = \widehat{\beta}(\widehat{B}^{-1}_{(i)})$, $\widehat{\mu}_{(i)} = \widehat{\mu}(\widehat{B}^{-1}_{(i)})$, and $\widehat{\gamma}_{(i)} = (1,-1)\widehat{\beta}_{(i)}$.

We assess the stability of the mean estimates by comparing the rankings of the genes across

Table 2: Each iteration k of the algorithm produces a ranking of all 2000 genes. For the top ten genes on each iteration, entry (i, j) of the table shows the number of genes in common in iterations i and j of the algorithm. Note that the maximum possible value for any entry of the table is 10; if entry (i, j) is 10, then iterations i and j selected the same top ten genes.

	1	2	3	4	5	6	7	8	9
1	10	10	7	5	5	3	3	3	3
2	10	10	7	5	5	3	3	3	3
3	7	7	10	6	5	3	3	3	3
4	5	5	6	10	8	5	5	5	5
5	5	5	5	8	10	7	7	7	7
6	3	3	3	5	7	10	10	10	10
7	3	3	3	5	7	10	10	10	10
8	3	3	3	5	7	10	10	10	10
9	3	3	3	5	7	10	10	10	10

Table 3: For the above algorithm, this table shows the number of genes that are significant at an FDR level of 0.1 on each iteration of the algorithm, for different values of the GLasso penalty λ. The top row shows the number of genes group centered on each iteration.

n.group	2000	1024	512	256	128	64	32	16	8
$\lambda = 0.1$	1006	913	327	14	3	1	1	1	1
$\lambda = 0.2$	865	806	262	2	1	1	1	1	0
$\lambda = 0.3$	778	789	303	3	1	1	0	0	0
$\lambda = 0.4$	706	774	452	3	1	0	0	0	0
$\lambda = 0.6$	657	751	587	19	1	1	0	0	0
$\lambda = 0.8$	628	699	493	30	1	1	1	1	1

iterations of the algorithm. Table 2 displays the number of genes in common out of the top ten genes on each pair of iterations of the algorithm. For example, three genes ranked in the top ten on the first iteration of the algorithm are also ranked in the top ten on the last iteration. Iterations six through nine produce the same ranking of the top ten genes. Three genes are ranked among the top ten on every iteration of the algorithm: DPP10-AS1, OLFM4, and PTN.

Table 3 shows the number of genes that fall below an FDR threshold of 0.1 on each iteration, for several values of the GLasso penalty λ. The number of genes below the threshold is more sensitive to the number of group-centered genes than to the GLasso penalty parameter. This is consistent with the first plot Figure 8 where the design effect (in the denominator of the test statistics) is likewise more sensitive to the number of group centered genes than to the GLasso penalty. When fewer than 128 genes are group centered, the number of genes below an FDR threshold of 0.1 is stable across the penalty parameters from $\lambda = 0.1$ to $\lambda = 0.8$.

Figure 9 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 9 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.

6 Proof sketch 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 (5). Define $\Delta_{n,m} = B_{n,m}^{-1} - B^{-1}$ and

$$\Omega = (D^T B^{-1} D)^{-1} \text{ and } \Omega_{n,m} = (D^T B_{n,m}^{-1} D)^{-1}. \tag{31}$$

Note that we can decompose the error for all j as

$$\|\widehat{\beta}_j(B_{n,m}^{-1}) - \beta_j^*\|_2 \leqslant \|\widehat{\beta}_j(B^{-1}) - \beta_j^*\|_2 + \|\widehat{\beta}_j(B_{n,m}^{-1}) - \widehat{\beta}_j(B^{-1})\|_2 =: \mathrm{I} + \mathrm{II}. \tag{32}$$

We will use the following lemmas, which are proved in subsections B.2 and B.1, to bound these two terms on the right-hand side, respectively.

Lemma 3. *Let $\mathcal{E}_2$ denote the event*

$$\mathcal{E}_2 = \left\{ \|\widehat{\beta}_j(B^{-1}) - \beta_j^*\|_2 \leqslant s_{n,m} \right\}, \quad \textit{with} \quad s_{n,m} = C_3 d^{1/2} \sqrt{\frac{\log(m)\|B\|_2}{n_{\min}}}. \tag{33}$$

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

Lemma 4. *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 (3). Let $\mathcal{E}_3$ denote the event*

$$\mathcal{E}_3 = \left\{ \|\widehat{\beta}_j(B_{n,m}^{-1}) - \widehat{\beta}_j(B^{-1})\|_2 \leqslant t_{n,m} \right\}, \quad \textit{with} \quad t_{n,m} = \widetilde{C} n_{\min}^{-1/2} \|\Delta_{n,m}\|_2. \tag{34}$$

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

The proof of (8) 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 (10). Let $r_{n,m} = s_{n,m} + t_{n,m}$, as defined in (8), and 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 $j = 1, \ldots, m$. $\square$

7 Conclusion

It has long been known that heteroscedasticity and dependence between observations impacts the precision and degree of uncertainty for estimates of mean values and regression coefficients. Further, data that are modeled for convenience as containing independent observations may in fact show unanticipated dependence (Kruskal, 1988). This has motivated the development of numerous statistical methods, including generalized/weighted least squares (GLS/WLS), mixed models, and generalized estimating equations (GEE). Our approach utilizes recent developments in high dimensional estimation to permit estimation of an inter-observation dependence structure (reflected in the matrix B in our model). Like GLS/GEE, we use an alternating fitting approach, but provide convergence guarantees and rates for a non-iterative two-step version of the algorithm.

Estimation of dependence or covariance structures usually requires some form of replication, and/or strong models. We require a relatively weak form of replication and a relatively weak model. In our framework, the dependence among observations must be common (up to proportionality) across a set of "quasi-replicates" (the columns of X, or the genes in our UC example). These quasi-replicates may be statistically dependent, and may have different means. We also require the precision matrices for the dependence structures to be sparse, which is a commonly used condition in recent high-dimensional analyses.

In addition to providing theoretical guarantees, we also show through simulations and a genomic data analysis that the approach improves estimation accuracy for the mean structure, and appears to eliminate test statistic overdispersion. The latter observation suggests that undetected dependence among observations may be one reason that genomic analyses are sometimes more inconsistent than traditional statistical methods would suggest, an observation made previously by Efron (2009) and others.

Although our theoretical analysis guarantees the convergence of our procedure even with a single observation of the random matrix X, there are reasons to expect this estimation problem

to be fundamentally challenging. One reason for this as pointed out by Efron (2009) and subsequently explored by Zhou (2014a), is that the row-wise and column-wise dependence structures are somewhat non-orthogonal, in that row-dependence can "leak" into the estimates of column-wise dependence, and vice-versa. Our results suggest that while row-wise correlations make it more difficult to estimate column-wise correlations (and vice-versa), when the emphasis is on mean structure estimation, even a somewhat rough estimate of the dependence structure (B) can improve estimation and inference.

References

Aitken, A. C. (1936). IV.–on least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh* **55** 42–48.

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.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica* 311–329.

Banerjee, O., Ghaoui, L. E. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research* **9** 485–516.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* 289–300.

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

Cai, T., Jessie Jeng, X. and Jin, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73** 629–662.

Cai, T. T., Li, H., Liu, W. and Xie, J. (2015). Joint estimation of multiple high-dimensional precision matrices. *The Annals of Statistics* **38** 2118–2144.

Cai, T. T. and Xia, Y. (2014). High-dimensional sparse manova. *Journal of Multivariate Analysis* **131** 174–196.

Chen, S. X., Li, J. and Zhong, P. S. (2014). Two-sample tests for high dimensional means with thresholding and data transformation. *arXiv preprint arXiv:1410.2848* .

Chen, S. X., Qin, Y.-L. et al. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics* **38** 808–835.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics* **55** 997–1004.

Efron, B. (2005). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association* **99** 96–104.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association* **102**.

Efron, B. (2009). Are a set of microarrays independ of each other? *Ann. App. Statist.* **3** 922–942.

Efron, B. (2010). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, vol. 1. Cambridge University Press.

Fan, J., Feng, Y. and Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics* **3** 521–541.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* **9** 432–441.

Hall, P., Jin, J. et al. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics* **38** 1686–1732.

Kruskal, W. (1988). Miracles and statistics: The casual assumption of independence. *Journal of the American Statistical Association* **83** 929–940.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrices estimation. *Annals of Statistics* **37** 4254–4278.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K. and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* **11** 733–739.

Li, J. and Zhong, P.-S. (2014). A rate optimal procedure for sparse signal recovery under dependence. *arXiv preprint arXiv:1410.2839* .

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Annals of Statistics* **34** 1436–1462.

Owen, A. B. (2005). Variance of the number of false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 411–426.

Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association* **104** 735–746.

Ravikumar, P., Wainwright, M., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics* **4** 935–980.

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

Storey, J. D. (2003). The positive false discovery rate: a bayesian interpretation and the q-value. *Annals of statistics* 2013–2035.

Sugden, L. A., Tackett, M. R., Savva, Y. A., Thompson, W. A. and Lawrence, C. E. (2013). Assessing the validity and reproducibility of genome-scale predictions. *Bioinformatics* **29** 2844–2851.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika* **94** 19–35.

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., Lafferty, J. and Wasserman, L. (2010). Time varying undirected graphs. *Machine Learning* **80** 295–319.

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.

A Proof of Theorem 2

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 2 **Part II** is proved in Section A.2. To prove Theorem 2 **Part I**, we first state Lemma 5, 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 6, 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 (42). Putting these results together proves Theorem 2, **Part I.** We prove these auxiliary results in Section C. We first state some convenient notation and bounds.

$$
\begin{aligned}
&r_a := a_{\max}/a_{\min} \text{ and } r_b := b_{\max}/b_{\min};\\
&1/\varphi_{\min}(A) = \|A^{-1}\|_2 \leqslant \|\rho(A)^{-1}\|_2/a_{\min} = \frac{1}{a_{\min}\varphi_{\min}(\rho(A))}, &(35)\\
&1/\varphi_{\min}(B) = \|B^{-1}\|_2 \leqslant \|\rho(B)^{-1}\|_2/b_{\min} = \frac{1}{b_{\min}\varphi_{\min}(\rho(B))}, &(36)\\
&1/\varphi_{\min}(\rho(A)) = \|\rho(A)^{-1}\|_2 \leqslant a_{\max}\|A^{-1}\|_2, &(37)\\
&1/\varphi_{\min}(\rho(B)) = \|\rho(B)^{-1}\|_2 \leqslant b_{\max}\|B^{-1}\|_2 &(38)\\
&\|A\|_2 \leqslant a_{\max}\|\rho(A)\|_2, \quad \|B\|_2 \leqslant b_{\max}\|\rho(B)\|_2, &(39)\\
&\|\rho(A)\|_2 \leqslant \|A\|_2/a_{\min}, \quad \text{and} \quad \|\rho(B)\|_2 \leqslant \|B\|_2/b_{\min}. &(40)
\end{aligned}
$$

The eigenvalues of the correlation matrices satisfy

$$
0 < \varphi_{\min}(\rho(A)) \leqslant 1 \leqslant \varphi_{\max}(\rho(A)) \text{ and } 0 < \varphi_{\min}(\rho(B)) \leqslant 1 \leqslant \varphi_{\max}(\rho(B)). \tag{41}
$$

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

$$
\widetilde{\alpha} = C_A K \frac{\log^{1/2}(m)}{\sqrt{m}}\left(1 + \frac{\|B\|_1}{n}\right) + \frac{\|B\|_1}{n_{\min}} \text{ and } \widetilde{\eta} = C_B K \frac{\log^{1/2}(m \vee n)}{\sqrt{n}} + \frac{\|B\|_1}{n}. \tag{42}
$$

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

$$
\forall i,j \qquad \left|\frac{(e_i - p_i)^T X X^T (e_j - p_j)}{\operatorname{tr}(A^*)\sqrt{b^*_{ii}b^*_{jj}}} - \rho_{ij}(B)\right| \leqslant \widetilde{\alpha} \tag{43}
$$

$$
\forall i,j \qquad \left|\frac{X_i^T(I - P_2)X_j}{\operatorname{tr}(B^*)\sqrt{a^*_{ii}a^*_{jj}}} - \rho_{ij}(A)\right| \leqslant \widetilde{\eta}, \tag{44}
$$

with $\mathcal{X}_0(B)$ and $\mathcal{X}_0(A)$ denoting the events defined by equations (43) and (44), respectively.

Let $\widetilde{\alpha}$ and $\widetilde{\eta}$ be as defined in (42). On event $\mathcal{X}_0(A)$, for all j, $\widehat{\Gamma}_{jj}(A) = \rho_{jj}(A) = 1$ and

$$\max_{j,k,j\neq k} |\widehat{\Gamma}_{jk}(A) - \rho_{jk}(A)| \leqslant \frac{2\widetilde{\eta}}{1-\widetilde{\eta}} \tag{45}$$

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

$$\max_{j,k,j\neq k} |\widehat{\Gamma}_{jk}(B) - \rho_{jk}(B)| \leqslant \frac{2\widetilde{\alpha}}{1-\widetilde{\alpha}}. \tag{46}$$

Lemma 5. *Suppose (A1) and (A2) hold. Let $\widehat{W}_1$ and $\widehat{W}_2$ be as defined in* (16)*. Let $\widehat{A}_\rho$ and $\widehat{B}_\rho$ be as defined in (12a) and (12b). For some absolute constants $18 < C, C' < 36$, the following events hold with probability at least $1 - 2/(n \vee m)^2$,*

$$\delta_{A,2} := \|\widehat{W}_1\widehat{A}_\rho\widehat{W}_1/\operatorname{tr}(B) - A\|_2 \leqslant Ca_{\max}\kappa(\rho(A))^2\lambda_B\sqrt{|A^{-1}|_{0,off}\vee 1} \tag{47}$$

$$\delta_{B,2} := \|\widehat{W}_2\widehat{B}_\rho\widehat{W}_2/\operatorname{tr}(A) - B\|_2 \leqslant C'b_{\max}\kappa(\rho(B))^2\lambda_A\sqrt{|B^{-1}|_{0,off}\vee 1} \tag{48}$$

$$\delta_{A,F} := \|\widehat{W}_1\widehat{A}_\rho\widehat{W}_1/\operatorname{tr}(B) - A\|_F \leqslant Ca_{\max}\kappa(\rho(A))^2\lambda_B\sqrt{|A^{-1}|_{0,off}\vee m} \tag{49}$$

$$\delta_{B,F} := \|\widehat{W}_2\widehat{B}_\rho\widehat{W}_2/\operatorname{tr}(A) - B\|_F \leqslant C'b_{\max}\kappa(\rho(B))^2\lambda_A\sqrt{|B^{-1}|_{0,off}\vee n}; \tag{50}$$

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

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

Lemma 6 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 (42). For completeness, we provide a sketch in Section C.2.

Lemma 6. *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 (42), 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 $18 < C, C' < 36$,

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

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

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

A.1 Proof of Theorem 2, Part I

We state additional helpful bounds:

$$(a_{\min} \vee \varphi_{\min}(A))\sqrt{m} \leqslant \|A\|_F = \left(\sum_{i=1}^m \varphi_i^2(A)\right)^{1/2} \leqslant \sqrt{m}\|A\|_2, \tag{51}$$

$$(b_{\min} \vee \varphi_{\min}(B))\sqrt{n} \leqslant \|B\|_F = \left(\sum_{i=1}^m \varphi_i^2(B)\right)^{1/2} \leqslant \sqrt{n}\|B\|_2, \tag{52}$$

$$\sqrt{m}/a_{\max} = \left(\frac{1}{a_{\max}} \vee \frac{1}{\varphi_{\max}(A)}\right)\sqrt{m} \leqslant \|A^{-1}\|_F \leqslant \sqrt{m}\|A^{-1}\|_2, \tag{53}$$

and

$$\sqrt{n}/b_{\max} = \left(\frac{1}{b_{\max}} \vee \frac{1}{\varphi_{\max}(B)}\right)\sqrt{n} \leqslant \|B^{-1}\|_F \leqslant \sqrt{n}\|B^{-1}\|_2. \tag{54}$$

Proof of Theorem 2, **Part I**. We plug in bounds as in (39) and (40) into Lemma 6 to obtain under (A1) and (A2), $\left\|\widehat{A\otimes B} - A\otimes B\right\|_2 \leqslant \|A\|_2\|B\|_2\delta$, where

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

For the inverse, we plug in bounds as in (37) and (38) into Lemma 6 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

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

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

A.2 Proof of Theorem 2, Part II

Let $\widehat{B}^{-1} = \widehat{W}_2\widehat{B}_\rho\widehat{W}_2$, and let $\widehat{\Delta} = \widehat{B}^{-1} - B^{-1}$. Let $\mathcal{E}_0(B)$ denote the event given by equations (51) and (51), which we know has probability at least $1 - 2/(n\vee m)^2$ from Lemma 5, and define the event

$$\mathcal{E}_4 = \left\{\|\widehat{\beta}_j(\widehat{B}^{-1}) - \beta_j^*\|_2 \leqslant s_{n,m} + t'_{n,m}\right\}, \tag{55}$$

where $s_{n,m}$ is as defined in (9) and

$$t'_{n,m} := C\lambda_A\sqrt{\frac{n_{\text{ratio}}\left(|B_0^{-1}|_{0,\text{off}} \vee 1\right)}{n_{\min}}}. \tag{56}$$

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

$$\|\widehat{\Delta}\|_2 \leqslant \frac{C'\lambda_A\sqrt{|B^{-1}|_{0,\text{off}} \vee 1}}{b_{\min}\varphi^2_{\min}(\rho(B))} = o(1). \tag{57}$$

Using Proposition 7 and the fact that $\|D\|_2 = \sqrt{n_{\max}}$, we get that

$$\|\Omega D^T \widehat{\Delta} D\|_2 \leqslant n_{\text{ratio}}\|B\|_2\|\widehat{\Delta}\|_2, \tag{58}$$

From (57) we know that $\|\widehat{\Delta}\|_2 \leqslant 1/(n_{\text{ratio}}\|B\|_2)$, which we can plug into (58) 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,

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

□

B Proofs for Theorem 1

For convenience, we first restate some notation.

$$D = \begin{bmatrix} 1_{n_1} & 0 \\ 0 & 1_{n_2} \end{bmatrix} \in \mathbb{R}^{n \times 2} \tag{59}$$

$$\Omega = (D^T B^{-1} D)^{-1} \text{ and } \Omega_{n,m} = (D^T B_{n,m}^{-1} D)^{-1} \tag{60}$$

$$\Delta = B_{n,m}^{-1} - B^{-1} \tag{61}$$

$$\widehat{\beta}(\widehat{B}^{-1}) = (D^T \widehat{B}^{-1} D)^{-1} D^T \widehat{B}^{-1} X \in \mathbb{R}^{2 \times m} \tag{62}$$

When D has the form (59), 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 defer the proof of Proposition 7 to Section B.3

Proposition 7. *For Ω as defined in (60) and some design matrix D,*

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

In the case that D is defined as in (59), we have $\|\Omega\|_2 \leqslant \|B\|_2/n_{\min}$.

We state the following perturbation bound.

Theorem 8 (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 9, we provide auxiliary upper bounds that depend on $\|\Delta\|_2$, $\|B\|_2$, $\kappa(D)$, and $\sigma_{\min}(D)$. We defer the proof of Proposition 9 to the end of this section, for clarity of presentation.

Proposition 9. *Let $\Delta = B^{-1}_{n,m} - B^{-1}$.*

$$\delta_0(\Delta) := \|\Omega_{n,m} - \Omega\|_2 \leqslant \frac{1}{\sigma^2_{\min}(D)} \frac{\|B\|_2^2\|\Delta\|_2}{1/\kappa^2(D) - \|B\|_2\|\Delta\|_2} \tag{63}$$

$$\delta_1(\Delta) := \left\|\Omega D^T\Delta\right\|_2 \leqslant \sigma_{\max}(D)\|B\|_2\|\Delta\|_2/\sigma^2_{\min}(D) = \frac{\sqrt{n_{\max}}}{n_{\min}}\|B\|_2\|\Delta\|_2. \tag{64}$$

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

$$\delta_2(\Delta) := \left\|(\Omega_{n,m} - \Omega)\, D^T\Delta\right\|_2 \leqslant \frac{\kappa(D)}{\sigma_{\min}(D)} \frac{\|B\|_2^2\|\Delta\|_2^2}{1/\kappa^2(D) - \|B\|_2\|\Delta\|_2} \tag{65}$$

$$\delta_3(\Delta) := \left\|(\Omega_{n,m} - \Omega)\, D^TB^{-1}\right\|_2 \leqslant \frac{\kappa(D)}{\sigma_{\min}(D)} \frac{\|B\|_2^2\|B^{-1}\|_2\|\Delta\|_2}{1/\kappa^2(D) - \|B\|_2\|_2\|\Delta\|_2} \tag{66}$$

When D has the form (59), and Ω is as defined in (60)*,*

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

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

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

B.1 Proof of Lemma 3

First, we show that

$$\|\Omega^{1/2}\|_F + d^{1/2}K^2\sqrt{\log(m)}\|\Omega\|_2^{1/2}/\sqrt{c} \leqslant s_{n,m}, \tag{67}$$

with $s_{n,m}$ as defined in (9). 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 7. 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 (67), 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,

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

□

B.2 Proof of Lemma 4

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

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

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,

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

We bound each of the three terms using Proposition 9,

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

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 $\|VZ_j\|_2$:

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

□

B.3 Proof of Proposition 7

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 $(D^T B^{-1} D)^{-1} = (V\Psi U^T B^{-1} U \Psi V^T)^{-1} = V\Psi^{-1}(U^T B^{-1} U)^{-1}\Psi^{-1} V^T$. Thus

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

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

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

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
$\|(D^T B^{-1} D)^{-1}\|_2 \leqslant \|B\|_2 / n_{\min}$. □

B.4 Proof of Proposition 9

By the definitions of $\Omega_{n,m}$ in (60) and $\Delta = B_{n,m}^{-1} - B^{-1}$, we have by Theorem 8

$$\begin{aligned}
\|\Omega_{n,m} - \Omega\|_2 &= \|(D^T B_{n,m} D)^{-1} - \Omega\|_2 \\
&= \left\| \left(D^T B_{n,m}^{-1} D - D^T B^{-1} D + D^T B^{-1} D\right)^{-1} - \Omega \right\|_2 \\
&= \left\| \left(D^T B^{-1} D + D^T \Delta D\right)^{-1} - \Omega \right\|_2 \\
&\leqslant \frac{\|D^T \Delta D\|_2 \|\Omega\|_2^2}{1 - \|\Omega\|_2 \|D^T \Delta D\|_2} \qquad \text{(by Theorem 8)} \\
&\leqslant \frac{\left(\sigma_{\max}^2(D)/\sigma_{\min}^4(D)\right) \|B\|_2^2 \|\Delta\|_2)}{1 - \kappa^2(D)\|B\|_2\|\Delta\|_2}.
\end{aligned}$$

In the last step we apply Proposition 7. Thus

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

We prove (65), as follows:

$$\begin{aligned}
\left\|(\Omega_{n,m} - \Omega)\, D^T \Delta\right\|_2 &\leqslant \|\Omega_{n,m} - \Omega\|_2 \left\|D^T\right\|_2 \|\Delta\|_2 \\
&\leqslant \left[\frac{1}{\sigma_{\min}^2(D)} \frac{\|B\|_2^2\|\Delta\|_2}{(1/\kappa^2(D)) - \|B\|_2\|\Delta\|_2}\right] \sigma_{\max}(D)\|\Delta\|_2 \qquad \text{(by Proposition 9)} \\
&= \frac{\kappa(D)}{\sigma_{\min}(D)} \frac{\|B\|_2^2\|\Delta\|_2^2}{(1/\kappa^2(D)) - \|B\|_2\|\Delta\|_2}.
\end{aligned}$$

The proof of (66) is analogous. We prove (64) as follows:

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

The inequality follows from the submultiplicative property and Proposition 7. □

C More proofs for Theorem 2

The proof of Lemma 5 appears in Section C.1. The proofs of auxiliary lemmas appear in Section C.2.

C.1 Proof of Lemma 5

In order to prove Lemma 5, we need Theorem 10, which shows explicit non-asymptotic convergence rates in the Frobenius norm for estimating $\rho(A)$, $\rho(B)$, and their inverses. Theorem 10 follows from the standard proof; see Rothman et al. (2008); Zhou et al. (2011) We also need Proposition 12 and Lemma 11, which are stated below and proved in Section C.2.

Theorem 10. *Suppose that (A2) holds. Let $\widehat{A}_\rho$ and $\widehat{B}_\rho$ be the unique minimizers defined by* (12a) *and* (12b) *with sample correlation matrices $\widehat{\Gamma}(A)$ and $\widehat{\Gamma}(B)$ as their input. Suppose that event $\mathcal{X}_0$ holds, with*

$$\widetilde{\eta}\sqrt{|A^{-1}|_{0,off} \vee 1} = o(1) \quad \text{and} \quad \widetilde{\alpha}\sqrt{|B^{-1}|_{0,off} \vee 1} = o(1).$$
$$\text{Set for some } \ 0 < \epsilon, \varepsilon < 1, \ \ \lambda_B = \widetilde{\alpha}/\varepsilon \ \text{ and } \ \lambda_A = \widetilde{\eta}/\epsilon. \tag{68}$$

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

$$\left\|\widehat{A}_\rho^{-1} - \rho(A)^{-1}\right\|_2 \leqslant \left\|\widehat{A}_\rho^{-1} - \rho(A)^{-1}\right\|_F < \frac{C\lambda_B\sqrt{|A^{-1}|_{0,off} \vee 1}}{2\varphi^2_{\min}(\rho(A))}, \tag{69}$$

$$\left\|\widehat{B}_\rho^{-1} - \rho(B)^{-1}\right\|_2 \leqslant \left\|\widehat{B}_\rho^{-1} - \rho(B)^{-1}\right\|_F \leqslant \frac{C\lambda_A\sqrt{|B^{-1}|_{0,off} \vee 1}}{2\varphi^2_{\min}(\rho(B))}. \tag{70}$$

We now state an auxiliary result, Lemma 11, 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 11. *Let $\widehat{W}_1$ and $\widehat{W}_2$ be as defined in* (16). *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* (43), (44). *For $\eta' := \frac{\widetilde{\eta}}{\sqrt{1-\widetilde{\eta}}} \leqslant \frac{\lambda_B}{6}$ and $\alpha' := \frac{\widetilde{\alpha}}{\sqrt{1-\widetilde{\alpha}}} \leqslant \frac{\lambda_A}{6}$,*

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

Proposition 12. *(Zhou, 2014a). Let $\widehat{W}$ and W be diagonal positive definite matrices. Let $\hat{\Psi}$ and Ψ be symmetric positive definite matrices. Then*

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

Proof of Lemma 5. Assume that event $\mathcal{X}_0$ holds. The proof follows exactly that of Lemma 15.3 in Zhou (2014a,b), in view of Theorem 10, Lemma 11 and Proposition 15.2 from Zhou (2014a,b), which is restated immediately above in Proposition 12. □

It remains to prove Lemma 11.

Proof of Lemma 11. 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

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

□

C.2 Proof of Lemma 6

In order to prove Lemma 6, we state Lemma 13, Lemma 14, and Proposition 15. Let $\|\cdot\|$ denote a matrix norm such that $\|A \otimes B\| = \|A\|\|B\|$. Let

$$\Delta := \widehat{W}_1\widehat{A}_\rho\widehat{W}_1 \otimes \widehat{W}_2\widehat{B}_\rho\widehat{W}_2/\operatorname{tr}(A)\operatorname{tr}(B) - A \otimes B, \tag{71}$$

$$\Delta' := \operatorname{tr}(A)\operatorname{tr}(B)\left(\widehat{W}_1\widehat{A}_\rho\widehat{W}_1\right)^{-1} \otimes \left(\widehat{W}_2\widehat{B}_\rho\widehat{W}_2\right)^{-1} - A^{-1} \otimes B^{-1}. \tag{72}$$

Lemma 13 is identical to Lemma 15.5 of Zhou (2014a), except that we now plug in quantities $\widetilde{\alpha}$ and $\widetilde{\eta}$ as defined in (42). Likewise, Proposition 15 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 13. *Let $\widehat{A \otimes B}$ be as in (19). Then for $\Sigma = A \otimes B$,*

$$\left\| \widehat{A \otimes B}^{-1} - \Sigma^{-1} \right\| \leqslant (\widetilde{\alpha} \wedge \widetilde{\eta}) \|A^{-1}\| \|B^{-1}\| + (1 + \widetilde{\alpha} \wedge \widetilde{\eta}) \|\Delta'\| \tag{73}$$

$$\left\| \widehat{A \otimes B} - \Sigma \right\| \leqslant \frac{\lambda_A \wedge \lambda_B}{2} \|A\| \|B\| + (1 + \frac{\lambda_A \wedge \lambda_B}{2}) \|\Delta\|. \tag{74}$$

Lemma 14 is a helpful bound on the difference of Kronecker products.

Lemma 14. *(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 15. *Under the event $\mathcal{X}_0$, as defined in as defined in* (43), (44),

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

Proof of Lemma 6. Assume that event $\mathcal{X}_0$ as defined in (43), (44) holds. The proof follows exactly the steps in Theorems 11.1 and 11.2 in Supplementary Material of Zhou (2014a,b). □

Proof of Lemma 13. By the triangle inequality and the sub-multiplicativity of the norm $\|\cdot\|$, with Δ and Δ' as defined in (71) and (72),

$$\mathrm{tr}(A)\,\mathrm{tr}(B) \left\| \left(\widehat{W}_1^{-1} \widehat{A}_\rho^{-1} \widehat{W}_1^{-1} \right) \otimes \left(\widehat{W}_2^{-1} \widehat{B}_\rho^{-1} \widehat{W}_2^{-1} \right) \right\| \leqslant \|A^{-1}\| \|B^{-1}\| + \|\Delta'\| \tag{75}$$

$$\left\| \left(\widehat{W}_1 \widehat{A}_\rho \widehat{W}_1 \right) \otimes \left(\widehat{W}_2 \widehat{B}_\rho \widehat{W}_2 \right) / \mathrm{tr}(A)\,\mathrm{tr}(B) \right\| \leqslant \|A\| \|B\| + \|\Delta\|. \tag{76}$$

Following proof of Lemma 15.5 Zhou (2014a,b), we have by definition of Δ', and Proposition 15, and (75),

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

By Proposition 15, we have for $\lambda_A \geqslant 3\widetilde{\alpha}$, $\lambda_B \geqslant 3\widetilde{\eta}$, where $\widetilde{\alpha} \wedge \widetilde{\eta} \leqslant \frac{\lambda_A \wedge \lambda_B}{3}$,

$$\begin{aligned}
&\left|\frac{1}{\|(I-P_2)X\|_F^2} - \frac{1}{\mathrm{tr}(A)\,\mathrm{tr}(B)}\right| = \left|\frac{\|(I-P_2)X\|_F^2 - \mathrm{tr}(A)\,\mathrm{tr}(B)}{\|(I-P_2)X\|_F^2\,\mathrm{tr}(A)\,\mathrm{tr}(B)}\right| \\
&\leqslant \left|\frac{\widetilde{\alpha}\wedge\widetilde{\eta}}{\|(I-P_2)X\|_F^2}\right| \leqslant \frac{\widetilde{\alpha}\wedge\widetilde{\eta}}{\mathrm{tr}(A)\,\mathrm{tr}(B)(1-\widetilde{\alpha}\wedge\widetilde{\eta})} \\
&\text{thus } \left|\frac{\mathrm{tr}(A)\,\mathrm{tr}(B)}{\|(I-P_2)X\|_F^2} - 1\right| \leqslant \frac{\widetilde{\alpha}\wedge\widetilde{\eta}}{1-\widetilde{\alpha}\wedge\widetilde{\eta}} \leqslant \frac{\lambda_A\wedge\lambda_B}{2}. \qquad (77)
\end{aligned}$$

By the triangle inequality, the definition of Δ in (71), and (76) and (77),

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

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

Proof of Proposition 15. 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 $\left|\|(I-P_2)X\|_F^2 - \mathrm{tr}(A)\mathrm{tr}(B)\right| \leqslant (\widetilde{\alpha}\wedge\widetilde{\eta})\mathrm{tr}(A)\mathrm{tr}(B)$. □

D Entrywise convergence of sample correlations

In this section we prove entrywise rates of convergence for the sample correlation matrices in Theorem 16. 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 (5). Let P_2 be as defined in (17). 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

$$S(B^*) = \frac{1}{m} \sum_{j=1}^{m} (I - P_2) X_j X_j^T (I - P_2), \tag{78}$$

$$S(A^*) = X^T (I - P_2) X / n, \tag{79}$$

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

Theorem 16. *Consider a data generating random matrix as in (3). Let C be some absolute constant. Let $\widetilde{\alpha}$ and $\widetilde{\eta}$ be as defined in* (42)*. 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 (11),*

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

We state three results used in the proof of Theorem 16: Proposition 17 provides an entrywise rate of convergence of $S(B^*)$, Proposition 18 provides an entrywise rate of convergence of $S(A^*)$, and Lemma 19 states that these entrywise rates imply $\mathcal{X}_0$. Let

$$\widetilde{B} := (I - P_2) B^* (I - P_2) = \mathrm{Cov}((I - P_2) X_j), \tag{80}$$

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

Proposition 17. *Let $d > 2$. Then with probability at least $1 - 2/m^{d-2}$,*

$$\forall i, j \ \left| S_{ij}(B^*) - b^*_{ij} \right| \leqslant \phi_{B,ij}, \tag{81}$$

with

$$\phi_{B,ij} = C\frac{\log^{1/2}(m)}{\sqrt{m}}\frac{\|A^*\|_F}{\sqrt{m}}\sqrt{\widetilde{b}_{ii}\widetilde{b}_{jj}} + \frac{3\|B^*\|_1}{n_{\min}}. \tag{82}$$

Proposition 18. *Let $d > 2$. Then with probability at least $1 - 2/n^{d-2}$,*

$$\forall i, j \ \left|S_{ij}(A^*) - a^*_{ij}\,\mathrm{tr}\,(B^*)\,/n\right| > \phi_{A,ij}, \tag{83}$$

with

$$\phi_{A,ij} = (a^*_{ij}/n)\left|\mathrm{tr}\left(\widetilde{B}\right) - \mathrm{tr}\,(B^*)\right| + d^{1/2}K\log^{1/2}(n\vee m)(1/n)\sqrt{a^{*2}_{ij} + a^*_{ii}a^*_{jj}}\|\widetilde{B}\|_F. \tag{84}$$

Lemma 19. *Suppose that (A2) holds and that $m \vee n \geqslant 2$. The event* (83) *defined in Proposition 18 implies that $\mathcal{X}_0(A)$ holds. Similarly, the event* (81) *defined in Proposition 17 implies $\mathcal{X}_0(B)$. Hence $\mathbb{P}\left(\mathcal{X}_0\right) \geqslant 1 - \frac{3}{(m\vee n)^2}$.*

Proposition 17 is proved in section D.1. Proposition 18 is proved in section D.2. Lemma 19 is proved in section D.3. Note that Lemma 19 follows from Propositions 17 and 18. We now prove Theorem 16, which follows from Lemma 19.

Proof of Theorem 16. Let q_i denote the ith column of $I - P_2$, so that $q_i^T XX^T q_j$ is the 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 19, 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. □

D.1 Proof of Proposition 17

We first present Lemma 20 and Lemma 21, which decompose the rate of convergence into a bias term and a variance term, respectively. We then combines the rates for the bias and variance terms to prove the entrywise rate of convergence for the sample covariance. Define

$$\mathcal{B}(B^*) := E[S(B^*)] - B^* \quad \text{and} \tag{85}$$

$$\sigma(B^*) := S(B^*) - E[S(B^*)] \tag{86}$$

We state maximum entrywise bounds on $\mathcal{B}(B^*)$ and $\sigma(B^*)$ in Lemma 20 and Lemma 21, respectively. Proofs for these lemmas are provided in Section D.4 and D.5 respectively.

Lemma 20. *For $\mathcal{B}(B^*)$ as defined in* (85)*,*

$$\|\mathcal{B}(B^*)\|_{\max} \leqslant \frac{3\|B^*\|_1}{n_{\min}}. \tag{87}$$

Lemma 21. *Let $\sigma(B^*)$ be as defined in* (86)*. With probability at least* $1 - 2/m^d$*,*

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

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

Proof of Proposition 17. 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 20 and 21. □

Remark. Note that the first term of (82) is of order $\log^{1/2}(m)/\sqrt{m}$, and the second term is of order $\|B^*\|_1/n_{\min}$.

D.2 Proof of Proposition 18

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

$$\Phi = (1/2n)\left((A^*_{\{1,2\}})^{1/2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} (A^*_{\{1,2\}})^{1/2} \right) \otimes B^{1/2}(I - P_2)B^{1/2}. \tag{88}$$

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

$$\mathrm{tr}\left((A^*_{\{1,2\}})^{1/2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} (A^*_{\{1,2\}})^{1/2} \right) = \mathrm{tr}\left(\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} A^*_{\{1,2\}} \right) = 2a^*_{12}. \tag{89}$$

For the Frobenius norm, note that

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

Therefore the trace of Φ is

$$\mathrm{tr}\,(\Phi) = a^*_{12}\,\mathrm{tr}\left(\widetilde{B}\right)/n, \tag{90}$$

and the Frobenius norm of Φ is

$$\|\Phi\|_F = (1/n)\sqrt{a^{*2}_{12} + a^*_{11}a^*_{22}}\|\widetilde{B}\|_F. \tag{91}$$

Applying the Hanson-Wright inequality yields

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

By the union bound,

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

□

D.3 Proof of Lemma 19

For the event (81) from Proposition 17,

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

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

$$\left|\frac{q_i X X^T q_j}{\operatorname{tr}(A^*)\sqrt{b_{ii}^* b_{jj}^*}} - \rho_{ij}(B)\right| < K^2 d C_A \frac{\log^{1/2}(m)}{\sqrt{m}} \sqrt{\frac{\widetilde{b}_{ii}\widetilde{b}_{jj}}{b_{ii}^* b_{jj}^*}} + \frac{\left|b_{ij} - \widetilde{b}_{ij}\right|}{\sqrt{b_{ii}^* b_{jj}^*}}. \tag{92}$$

By Lemma 20,

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

so the right-hand side of (92) is less than or equal to $\widetilde{\alpha}$. Hence event (81) implies $\mathcal{X}_0(B)$. Therefore, we know that $P(\mathcal{X}_0(B)) \geqslant 1 - 2/m^{d-2}$.

Similarly, event (83) in Proposition 18:

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

implies that

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

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. □

D.4 Proofs of Lemma 20

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 $p_i^T B^* p_j$ as an inner product $(B^{1/2}p_i)^T(B^{1/2}p_j)$, so

$$|(P_2 B^* P_2)_{ij}| = |(B^{1/2}p_i)^T(B^{1/2}p_j)| \leqslant \left(p_i^T B^* p_i\right)^{1/2}\left(p_j^T B^* p_j\right)^{1/2} \tag{93}$$

$$\leqslant \|p_i\|_2\|p_j\|_2\|B\|_2 \leqslant \|B^*\|_2/n_{\min}, \tag{94}$$

where (93) follows from the Cauchy Schwarz inequality, and (94) holds because

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

□

D.5 Proof of Lemma 21

Let $B^{1/2}$ denote the symmetric square root of B^*, and 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 $S_{ij}(B^*) = \operatorname{vec}(Z)^T \Phi^{ij} \operatorname{vec}(Z)$, with

$$\Phi^{ij} \;=\; (1/m)A^* \otimes B^{1/2}(e_j - p_j)(e_i - p_i)^T B^{1/2}. \tag{95}$$

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

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

where

$$\operatorname{tr}(\Phi^{ij}) = \operatorname{tr}(B^{1/2}(e_j - p_j)(e_i - p_i)^T B^{1/2}) = (e_i - p_i)^T B^*(e_j - p_j) = \widetilde{b}_{ij}, \tag{96}$$

$$\begin{aligned} \|\Phi^{ij}\|_F &= \frac{1}{m}\|A^*\|_F \|B^{1/2}(e_j - p_j)(e_i - p_i)^T B^{1/2}\|_F \qquad (97) \\ &= \frac{1}{m}\|A^*\|_F \left((e_i - p_i)^T B^*(e_i - p_i)\right)^{1/2}\left((e_j - p_j)^T B^*(e_j - p_j)\right)^{1/2} = \frac{1}{m}\|A^*\|_F \sqrt{\widetilde{b}_{ii}\widetilde{b}_{jj}}. \end{aligned}$$

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. $\square$