Electronic Journal of Statistics
Vol. 18 (2024) 1702–1748
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2222

Direct covariance matrix estimation with compositional data

Aaron J. Molstad

School of Statistics, University of Minnesota
e-mail: amolstad@umn.edu

Karl Oskar Ekvall

Department of Statistics, University of Florida
Division of Biostatistics, Institute of Environmental Medicine, Karolinska Institutet
e-mail: k.ekvall@ufl.edu

Piotr M. Suder

Department of Statistical Science, Duke University
e-mail: piotr.suder@duke.edu

Abstract: Compositional data arise in many areas of research in the natural and biomedical sciences. One prominent example is in the study of the human gut microbiome, where one can measure the relative abundance of many distinct microorganisms in a subject's gut. Often, practitioners are interested in learning how the dependencies between microbes vary across distinct populations or experimental conditions. In statistical terms, the goal is to estimate a covariance matrix for the (latent) log-abundances of the microbes in each of the populations. However, the compositional nature of the data prevents the use of standard estimators for these covariance matrices. In this article, we propose an estimator of multiple covariance matrices which allows for information sharing across distinct populations of samples. Compared to some existing estimators, which estimate the covariance matrices of interest indirectly, our estimator is direct, ensures positive definiteness, and is the solution to a convex optimization problem. We compute our estimator using a proximal-proximal gradient descent algorithm. Asymptotic properties of our estimator reveal that it can perform well in high-dimensional settings. We show that our method provides more reliable estimates than competitors in an analysis of microbiome data from subjects with myalgic encephalomyelitis/chronic fatigue syndrome and through simulation studies.

Keywords and phrases: Compositional data, covariance matrix estimation, microbiome data analysis, convex optimization, positive definiteness.

Received March 2023.

Contents

1. Introduction

High-dimensional compositional data arise in many areas of modern science. To study the human gut microbiome, for example, practitioners measure the relative abundances of various microbes using next-generation sequencing followed by alignment and normalization [14]. For each subject in a study, the resulting measurement is a p-dimensional vector which has nonnegative entries and sums to one [18, 33]. More generally, compositional data arise when, for example, one observes multivariate count-valued data wherein the total counts in a sample is an experimental artifact. Here, we focus on compositional data which belong to the set

$$\mathbb{C}^{p-1} = \left\{ x \in \mathbb{R}^p : \sum_{j=1}^{p} x_j = 1, x_j > 0 \text{ for each } j \in [p] \right\},$$

where $[p] = \{1, \ldots, p\}$ for a positive integer p.

To make matters concrete, let $X = (X_1, \ldots, X_p)^\top \in \mathbb{C}^{p-1}$ be a random composition whose components correspond to the variables of interest. Letting $W = (W_1, \ldots, W_p)^\top$ denote the corresponding latent abundances, also known as the basis [1], we assume

$$X_j = \frac{W_j}{\sum_{k=1}^p W_k}, \quad j \in [p],$$

where each $W_j \in (0, \infty)$. When characterizing the dependence between any two components from the compositional vector, the parameter of interest is often the basis covariance matrix $\Omega^* \in \mathbb{S}_+^p$, where

$$\Omega^*_{jk} = \mathrm{Cov}\left\{\log(W_j), \log(W_k)\right\}, \quad (j,k) \in [p] \times [p],$$

and $\mathbb{S}_+^p$ denotes the set of $p \times p$ symmetric positive definite matrices.

In many studies involving compositional microbiome data, practitioners are interested in modeling the interactions and dependencies between microbe abundance [11, 22]. For instance, one may want to estimate whether two microbes occur in higher frequencies jointly. The basis covariance matrix Ω^* provides one route for addressing such questions [19, 23, 16], but is not straightforward to estimate from independent realizations of X because W is latent. One common approach relies on the estimation of the variation matrix Θ^* [2, Chapter 4], defined elementwise by

$$\begin{aligned}\Theta^*_{jk} &= \mathrm{Var}\left\{\log(X_j/X_k)\right\}, \\ &= \mathrm{Var}\left\{\log(W_j) - \log(W_k)\right\} \\ &= \mathrm{Var}\left\{\log(W_j)\right\} + \mathrm{Var}\left\{\log(W_k)\right\} - 2\mathrm{Cov}\left\{\log(W_j), \log(W_k)\right\}.\end{aligned}$$

Thus, letting $\omega^* = \mathrm{diag}(\Omega^*) \in \mathbb{R}^p$ and $\mathbb{1}_p = (1, 1, \ldots, 1)^\top \in \mathbb{R}^p$,

$$\Theta^* = \omega^* \mathbb{1}_p^\top + \mathbb{1}_p \omega^{*\top} - 2\Omega^*. \tag{1.1}$$

To define an estimator of Θ^*, let $x_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{C}^{p-1}$, $i \in [n]$, denote independent realizations of X. Let also $z_{ijk} = \log(x_{ij}/x_{ik})$ and $\bar{z}_{jk} = n^{-1}\sum_{i=1}^n z_{ijk}$ for all $(j,k) \in [p] \times [p]$. The sample estimator $\widehat{\Theta}$ is defined elementwise by

$$\widehat{\Theta}_{jk} = \frac{1}{n}\sum_{i=1}^n (z_{ijk} - \bar{z}_{jk})^2.$$

While $\widehat{\Theta}$ is a natural estimator of Θ^*, it is unclear how to use it to estimate Ω^* in general because there are infinitely many $\widehat{\Omega}$ such that $\widehat{\Theta} = \widehat{\omega}\mathbb{1}_p^\top + \mathbb{1}_p\widehat{\omega}^\top - 2\widehat{\Omega}$. Namely, the diagonal entries of $\widehat{\Theta}$ and $\widehat{\omega}\mathbb{1}_p^\top + \mathbb{1}_p\widehat{\omega}^\top - 2\widehat{\Omega}$ are zero, so there are $p(p-1)/2$ unique equalities but $p(p+1)/2$ unknowns in $\widehat{\Omega}$. However, if one assumes that many entries of Ω^* are zero, then it can be estimated based on (1.1). Cao, Lin and Li [7] proved that if $p \geq 5$ and Ω^* has fewer than $(p-1)$ nonzero off-diagonal entries, then no two Ω^* correspond to the same Θ^*. Thus,

if one could assume $s < p-1$ off-diagonal entries of Ω^* are nonzero, one could consider the estimator

$$\underset{\Omega=\Omega^\top}{\arg\min} \|\widehat{\Theta} - \omega \mathbb{1}_p^\top - \mathbb{1}_p \omega^\top + 2\Omega\|_F^2 \quad \text{subject to } \|\Omega^-\|_0 \le s, \tag{1.2}$$

where Ω^- denotes the matrix Ω with its diagonal entries set to zero, $\|A\|_F^2 = \operatorname{tr}(A^\top A) = \sum_{j,k} A_{jk}^2$ is the squared Frobenius norm of a matrix A, and $\|A\|_0 = \sum_{j,k} \mathbf{1}(A_{jk} \neq 0)$ counts the number of nonzero entries in A. In practice, assuming only $s < p-1$ off-diagonal elements are non-zero is often too restrictive. Of course, (1.2) could also be used with $s \ge p-1$, but due to the L_0 constraint, (1.2) is the solution to a nonconvex optimization problem and is computationally challenging for large p.

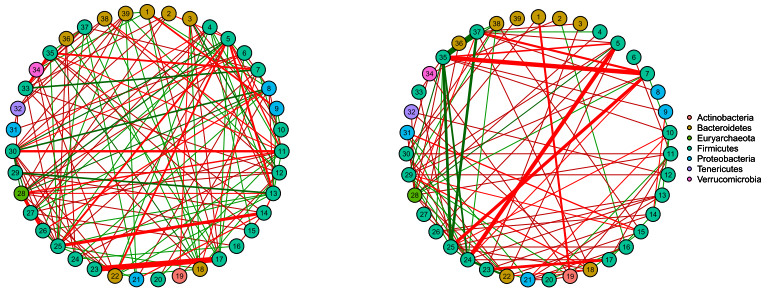
In view of (1.2) and its limitations, and given that the L_1 norm is a convex relaxation of the L_0 norm, a natural alternative is

$$\underset{\Omega=\Omega^\top}{\arg\min} \left\{ \|\widehat{\Theta} - \omega \mathbb{1}_p^\top - \mathbb{1}_p \omega^\top + 2\Omega\|_F^2 + \lambda \|\Omega^-\|_1 \right\}, \tag{1.3}$$

where $\|A\|_1 = \sum_{j,k} |A_{jk}|$ for a matrix A. Cao, Lin and Li [7] described their estimator as a "one-step approximation to (1.3)", but did not study (1.3). Appealingly, the problem in (1.3) can be recast as an L_1-penalized least squares problem and computed via existing algorithms. However, neither (1.2), (1.3), nor the method of Cao, Lin and Li [7] provide estimates which are guaranteed to be positive definite, or even nonnegative definite (see Section 2.2). Replacing the feasible set in (1.2) or (1.3) by $\mathbb{S}_+^p$, or a subset thereof, complicates computation substantially. For example, even in the context of standard covariance matrix estimation (i.e., when the $\log(W_j)$ are observable), enforcing sparsity and positive definiteness simultaneously is challenging [4, 31, 38, 37].

In many applications, one requires a basis covariance matrix estimate from multiple distinct populations. For example, in our motivating data analysis, the goal is to compare how the microbes interact in the gut of patients with myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) versus controls [13]. To estimate the two basis covariance matrices, one could apply existing estimators to each of the populations (ME/CFS patients and controls) separately. However, sample sizes are often small relative to the dimension of the basis covariance. For example, there are only 37 and 47 control and ME/CFS patients, respectively, used to estimate both 39×39 basis covariance matrices.

A more efficient approach would estimate the two covariance matrices jointly in order to borrow information across populations. If, for instance, the basis covariances have similar sparsity patterns, exploiting this shared information across populations can substantially improve efficiency. Joint estimation is especially common in the literature on estimating sparse covariance and inverse covariance matrices from multivariate normal data collected on multiple populations [5, 15, 8, 28, 21, 6, 32, 29]. In the context of estimating basis covariance matrices from microbiome data, it is natural to assume the covariance matrices have similar sparsity patterns. Biologically, it is often reasonable to assume



(a) Controls (b) ME/CFS

FIG 1. *Estimated correlation networks for controls and patients with ME/CFS [13] using the method of Cao, Lin and Li [7]. Each node corresponds to an OTU as described in Section 7. Green edges denote positive estimated correlations, red edges denote negative estimated correlations, and the absence of an edge indicates an estimated correlation of zero. The thickness of an edge indicates the magnitude of the correlation: thicker corresponds to a larger magnitude. Panel (a) is the network estimated from control patients while panel (b) is the network estimated from patients with ME/CFS.*

there are microbes whose abundances are uncorrelated in all the populations in a study. For example, in Section 7, when we estimate the basis covariance matrices for ME/CFS patients and controls using our method, which shares information across populations, we estimate identical sparsity patterns. In contrast, when we estimate these matrices separately using an existing method, few estimated nonzero correlations are shared between populations (Figure 1). We investigate the reliability of these estimates in Section 7.2.

In this article, we study (1.3) under positive definite constraints, and propose a generalization of (1.3) for estimating multiple covariance matrices simultaneously. We establish asymptotic error bounds for both the single and multiple population versions of our estimator, and we propose an efficient algorithm for their computation. In simulation studies and our analysis of the ME/CFS microbiome data, we demonstrate that our methods can provide more reliable estimates of the covariance matrices of interest than existing competitors.

2. Methodology

2.1. Multiple basis covariance matrix estimation

In the remainder of this article, we let the subscript (h) denote data or population parameters from the hth population, $h \in [H]$ for some $H \geq 1$. For example, $x_{(h)i} \in \mathbb{R}^p$ is the vector with compositional data for observation $i \in [n_{(h)}]$ in the hth population. Similarly, $\Omega^*_{(h)}$ is the basis covariance for the hth population.

We focus on estimating $\Omega^*_{(1)}, \ldots, \Omega^*_{(H)}$ using the data $\{x_{(h)i} \in \mathbb{R}^p : h \in$

$[H], i \in n_{(h)}\}$. As argued in Section 1, one can estimate any $\Theta^*_{(h)}$ using

$$\widehat{\Theta}_{(h)jk} = \frac{1}{n_{(h)}} \sum_{i=1}^{n_{(h)}} (z_{(h)ijk} - \bar{z}_{(h)jk})^2, \quad (j,k) \in [p] \times [p],$$

where $z_{(h)ijk} = \log(x_{(h)ij}/x_{(h)ik})$ and $\bar{z}_{(h)jk} = n_{(h)}^{-1} \sum_{i=1}^{n_{(h)}} z_{(h)ijk}$.

To describe our estimator, define $\mathbf{\Omega} \in \mathbb{R}^{H \times p \times p}$ as the three-way tensor where $\mathbf{\Omega}_{h\cdot\cdot} = \Omega_{(h)} \in \mathbb{R}^{p \times p}$ for $h \in [H]$ and $\mathbf{\Omega}_{\cdot jk} = (\Omega_{(1)jk}, \ldots, \Omega_{(H)jk})^\top \in \mathbb{R}^H$ for $(j,k) \in [p] \times [p]$. We present a visualization of the tensor $\mathbf{\Omega}$ in Figure 2. The mode-1 fibers, $\mathbf{\Omega}_{\cdot jk}$, are vectors containing the (j,k)th entry of all the $\Omega_{(h)}$. Assuming sparsity patterns are shared across populations is thus equivalent to assuming $\mathbf{\Omega}^*_{\cdot jk} = 0$ for many pairs (j,k). Finally, let $\omega_{(h)} = \operatorname{diag}(\Omega_{(h)})$ for $h \in [H]$.

Generalizing (1.3) with an additional positive definiteness constraint, we propose to estimate $\mathbf{\Omega}^*$ using

$$\begin{aligned} \underset{\mathbf{\Omega} \in \mathbb{R}^{H \times p \times p}}{\arg\min} \sum_{h=1}^{H} \left(\|\widehat{\Theta}_{(h)} - \omega_{(h)} \mathbb{1}_p^\top - \mathbb{1}_p \omega_{(h)}^\top + 2\Omega_{(h)}\|_F^2 + \lambda \|\Omega^-_{(h)}\|_1 \right) + \gamma \sum_{j \neq k} \|\mathbf{\Omega}_{\cdot jk}\|_2 \\ \text{subject to } \ \Omega_{(h)} = \Omega_{(h)}^\top, \ \Omega_{(h)} \succcurlyeq \epsilon I_p \ \text{ for all } \ h \in [H], \end{aligned} \tag{2.1}$$

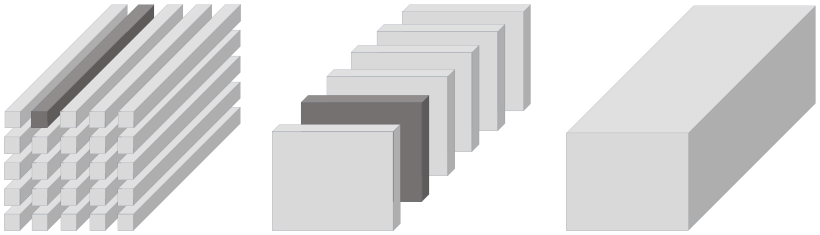
where $\lambda \geq 0$, $\gamma \geq 0$, and $\epsilon \geq 0$ are user-specified tuning parameters, $\|\cdot\|_2$ denotes the Euclidean norm of a vector, and $A \succcurlyeq \epsilon I_p$ means that $A - \epsilon I_p$ is positive semidefinite.

The estimator (2.1) imposes both a lasso-type penalty on the off-diagonal entries of the $\Omega_{(h)}$, as well as a group lasso penalty on the mode-1 fibers of the tensor $\mathbf{\Omega}$. Note that if $H = 1$, taking either $\lambda = 0$ or $\gamma = 0$ with the other nonzero, (2.1) simplifies to (1.3) with a positive definiteness constraint. For example, if $\gamma = 0$, then (2.1) simplifies to the estimator

$$\begin{aligned} \underset{\Omega_{(h)} \in \mathbb{R}^{p \times p}}{\arg\min} \left(\|\widehat{\Theta}_{(h)} - \omega_{(h)} \mathbb{1}_p^\top - \mathbb{1}_p \omega_{(h)}^\top + 2\Omega_{(h)}\|_F^2 + \lambda \|\Omega^-_{(h)}\|_1 \right), \\ \text{subject to } \ \Omega_{(h)} = \Omega_{(h)}^\top, \ \ \Omega_{(h)} \succcurlyeq \epsilon I_p, \end{aligned} \tag{2.2}$$

applied to each of the H populations separately. The estimator (2.2) can be seen as a convex approximation to (1.2) where we have replaced the L_0 constraint with an L_1 constraint, and replaced the feasible set with the closed convex set $\{\Omega \in \mathbb{R}^{p \times p} : \Omega = \Omega^\top, \Omega \succcurlyeq \epsilon I\}$. The tuning parameter ϵ serves as a lower bound on the smallest eigenvalue of the solution. For this reason, we do not recommend tuning ϵ, but rather fixing it at some reasonably small quantity like 10^{-4}, as in Xue, Ma and Zou [38].

By taking $\gamma > 0$, however, (2.1) ties the estimators of $\Omega^*_{(1)}, \ldots, \Omega^*_{(H)}$ together. For large values of the tuning parameter γ, the second penalty in (2.1) will require that the solution to (2.1) has some $\mathbf{\Omega}_{\cdot jk} = 0$, i.e., that sparsity is partially shared across all H basis covariance matrix estimates. In the leftmost subfigure



(a) $\boldsymbol{\Omega}_{\cdot 12} \in \mathbb{R}^H$ (b) $\boldsymbol{\Omega}_{2\cdot\cdot} = \Omega_{(2)} \in \mathbb{R}^{p\times p}$ (c) $\boldsymbol{\Omega} \in \mathbb{R}^{H\times p\times p}$

FIG 2. *Visualization of (a) the fibers of* $\boldsymbol{\Omega}$ *which are penalized by the final term in* (2.1)*, (b) the organization of* $\boldsymbol{\Omega}$ *into the* $\Omega_{(h)}$*, and (c) the three way tensor* $\boldsymbol{\Omega}$*.*

of Figure 2, (a), we provide an example of the group of parameters—the (1,2)th element of each $\Omega_{(h)}$—which are jointly penalized by the group lasso penalty. The tuning parameter γ controls whether this group is entirely zero or not, whereas the tuning parameter λ controls sparsity in the individual entries of the $\Omega_{(h)}$, as displayed in subfigure (b).

Importantly, (2.1) is a convex optimization problem and as we discuss in Section 3, can be solved using first-order methods.

2.2. *Positive definiteness*

To understand why enforcing positive definiteness can be necessary, consider estimating a single Ω^* whose off-diagonal entries are assumed to be zero. Then the problem reduces to estimation of the variances of the log abundances and (1.2) admits a closed-form solution.

Proposition 1. *If* $p \geq 3$*, the solution to* (1.2) *with* $s = 0$ *(or equivalently,* (1.3) *with* $\lambda = \infty$*) is unique and is given by*

$$\widehat{\omega}_j = \frac{1}{p-1}\sum_{k\neq j}\widehat{\Theta}_{jk} - \frac{1}{2(p-1)(p-2)}\sum_{k\neq j}\sum_{l\neq j}\widehat{\Theta}_{lk}, \quad j \in [p].$$

The factor 2 in the denominator is due to the double sum running over both the upper and lower triangular parts of the symmetric $\widehat{\Theta}$. The proposition reveals variance estimates can be negative if positive definiteness is not enforced. Roughly speaking, for large p, $\widehat{\omega}_j$ will be negative if the average of the elements in $\widehat{\Theta}$ not in the jth row or column is larger than the average of the elements in the jth row and column. It is not difficult to produce such examples. As an illustration, the following $\widehat{\Theta}$ resulted from simulating $n = 10$ compositional $x_i \in \mathbb{C}^2$ by drawing the $\log(W)$ independently from a multivariate normal distribution with mean zero and identity covariance matrix:

$$\widehat{\Theta} = \begin{pmatrix} 0 & 3.83 & 2.45 \\ 3.83 & 0 & 1.24 \\ 2.45 & 1.24 & 0 \end{pmatrix}.$$

Thus, $\widehat{\omega}_3 = (2.45 + 1.24)/2 - 3.83/2 = -0.07$. Intuitively, negative variance estimates are more likely when p is large relative to n.

2.3. Existing estimators

An alternative estimator of a single basis covariance matrix Ω^* is based on the centered log-ratio covariance matrix [2], Γ^*, whose (j,k)th entry is

$$\Gamma^*_{jk} = \mathrm{Cov}\left[\log\{X_j/g(X)\}, \log\{X_k/g(X)\}\right]$$

where $g(X) = (\prod_{i=1}^p X_i)^{1/p}$ is the geometric mean of X. Specifically,

$$\begin{aligned}\Theta^*_{jk} &= \mathrm{Var}\{\log(X_j/X_k)\} \\ &= \mathrm{Var}[\log\{X_j/g(X)\} - \log\{X_k/g(X)\}] \\ &= \mathrm{Var}[\log\{X_j/g(X)\}] + \mathrm{Var}[\log\{X_k/g(X)\}] \\ &\quad - 2\mathrm{Cov}[\log\{X_j/g(X)\}, \log\{X_k/g(X)\}]\end{aligned}$$

so that $\Theta^* = \gamma^* \mathbb{1}_p^\top + \mathbb{1}_p \gamma^{*\top} - 2\Gamma^*$, where $\gamma^* = \mathrm{diag}(\Gamma^*)$. Cao, Lin and Li [7] show there exists a unique Γ^* such that $\Theta^* = \gamma^* \mathbb{1}_p^\top + \mathbb{1}_p \gamma^{*\top} - 2\Gamma^*$ and that $\max_{j,k} |\Omega^*_{jk} - \Gamma^*_{jk}| \leq (3/p)(\max_{j\in[p]} \sum_{k=1}^p |\Omega^*_{jk}|)$. Thus, by proposing a two-step procedure to get an estimate of Γ^* from $\widehat{\Theta}$, they also get an indirect estimate of Ω^* that can perform well when p is large. However, their estimator is not guaranteed to be positive definite, nor is it the solution to an optimization problem amenable to analysis.

Fang et al. [10] proposed a different estimator, using that with $F = I_p - \mathbb{1}_p \mathbb{1}_p^\top / p$, $F\Omega^* F = F\mathrm{Cov}(\log X)F$. Thus, replacing $\mathrm{Cov}(\log X)$ with its sample version, say, $\widehat{\Omega}_X$, a natural estimating equation is $F(\Omega - \widehat{\Omega}_X)F = 0$. To account for differing variances in each element of $F(\Omega^* - \widehat{\Omega}_X)F$, they propose the weighted least squares estimator

$$\underset{\Omega=\Omega^\top}{\arg\min} \left\{ \frac{1}{2} \|F(\Omega - \widehat{\Omega}_X)F\|_V^2 + \lambda \|\Omega^-\|_1 \right\}, \tag{2.3}$$

where V is a diagonal matrix with $\mathrm{diag}(V) = \{\mathrm{diag}(F\widehat{\Omega}_X F)\}^{-1}$ and $\|A\|_V^2 = \mathrm{tr}(A^\top V A)$. While Fang et al. [10] suggest including a positive definiteness constraint on the optimization variable Ω in (2.3), their computational algorithm does not enforce this constraint. Instead, if the solution to (2.3) is not positive definite, they estimate Ω^* using its nearest positive definite matrix. This can be appropriate, but often leads to a non-sparse estimate [35].

In contrast to the methods of Cao, Lin and Li [7] and Fang et al. [10], our estimator is "direct" in the sense that we do not rely on estimation of intermediate quantities like Γ^*, nor do we rely on post-hoc adjustments to achieve positive definiteness.

Many other estimators of Ω^* exist, though we do not cover them in detail here. In general, these estimators do not enforce both positive definiteness and

sparsity, and are not specifically designed to estimate multiple covariance matrices simultaneously: see Friedman and Alm [12], Ban, An and Jiang [3], He et al. [16], Li et al. [20], for example, and see Ma, Yue and Shojaie [22] for a comprehensive review.

As mentioned in Section 1, our work is related to the literature on jointly estimating sparse precision (inverse covariance) matrices for multiple populations [e.g., 15, 8, 28, 32, 29]. However, our work is distinct in at least two important ways. First, these existing methods are largely focused on estimating precision matrices, rather than covariance matrices. Sparse precision matrix estimation and sparse covariance matrix estimation are fundamentally different tasks. Second, these methods, broadly speaking, assume the observed data are normally distributed or utilize a normal negative log-likelihood as a loss function. These methods are thus not directly applicable to either covariance matrix estimation, nor the analysis of compositional data.

Differences in motivation aside, there are some similarities between our work and that of Guo et al. [15] and Danaher, Wang and Witten [8], for example. Danaher, Wang and Witten [8] use the same penalty as (2.1), but applied to precision matrices for normally distributed data. Guo et al. [15] use a group lasso-type penalty to encourage shared sparsity patterns across populations in the same context as Danaher, Wang and Witten [8]. The existing work most closely related to that of our own is Bigot et al. [5], who use a group lasso penalty in the context of estimating multiple sparse covariance matrices from data observed with additive noise. One could not straightforwardly use their method for estimating basis covariance matrices from compositional data, and moreover, neither their theory nor algorithms apply to our estimator.

3. Computation

3.1. Proximal-proximal gradient descent algorithm

In order to solve the optimization problem to compute (2.1), we must address both the nondifferentiability of the objective function and the positive definiteness constraint. To do so, we use the proximal-proximal gradient descent algorithm [9], which allows us to handle the nondifferentiable penalty and positive definiteness constraint separately. The algorithm generalizes the well-known proximal gradient descent algorithm [25, Section 4.2] to handle problems where the objective function to be minimized is the sum of three convex functions. Specifically, supposing f and g are closed, proper, and convex functions; and ℓ is convex and differentiable with β^{-1}-Lipschitz continuous gradient for some $\beta > 0$; consider a problem of the form

$$\underset{u \in \mathbb{R}^d}{\text{minimize}} \left\{ \ell(u) + f(u) + g(u) \right\}. \tag{3.1}$$

Further suppose there exists $u^\star \in \mathbb{R}^d$ such that $0 \in \partial f(u^\star) + \partial g(u^\star) + \nabla \ell(u^\star)$ where $\partial f(u)$ denotes the subdifferential of f at u. The proximal operator of a

function f evaluated at u is

$$\mathbf{prox}_f(u) = \underset{y \in \mathrm{dom} f}{\arg\min} \left\{ \frac{1}{2}\|u - y\|_2^2 + f(y) \right\}.$$

Davis and Yin [9] show that (3.1) can be solved by an algorithm whose (t)th iterates are computed using the updating equations

$$\begin{aligned} u_g^{(t)} &= \mathbf{prox}_{\alpha g}(v^{(t)}) \\ u_f^{(t)} &= \mathbf{prox}_{\alpha f}\{2u_g^{(t)} - v^{(t)} - \alpha \nabla \ell(u_g^{(t)})\} \\ v^{(t+1)} &= v^{(t)} + u_f^{(t)} - u_g^{(t)}, \end{aligned}$$

where $v^{(0)}$ is an arbitrary point in $\mathbb{R}^d$ and $\alpha \in (0, 2\beta)$ is fixed. Here, the superscript (t) denotes the (t)th iterate. As $t \to \infty$, $u_g^{(t)} \to u^\star$ and $u_f^{(t)} \to u^\star$ [9]. In practice, however, this algorithm can be slow to converge: fixing the step size $\alpha \in (0, 2\beta)$ can sometimes lead to incremental progress. Therefore, we use a modified version of the proximal-proximal gradient descent algorithm proposed by Pedregosa and Gidel [26], which allows us to start with a step size α larger than 2β and reduce its value as needed, thus potentially accelerating the descent. The $(t+1)$th iterates of the algorithm use the updating equations

$$u_f^{(t+1)} = \mathbf{prox}_{\alpha f}\{u_g^{(t)} - \alpha v^{(t)} - \alpha \nabla \ell(u_g^{(t)})\} \tag{3.2}$$

$$u_g^{(t+1)} = \mathbf{prox}_{\alpha g}(u_f^{(t+1)} + \alpha v^{(t)}) \tag{3.3}$$

$$v^{(t+1)} = v^{(t)} + \alpha^{-1}(u_f^{(t+1)} - u_g^{(t+1)}). \tag{3.4}$$

At each step, after (3.2) is carried out, the value

$$Q(u_f^{(t+1)}, \alpha) = \ell(u_g^{(t)}) + \langle \nabla \ell(u_g^{(t)}), u_f^{(t+1)} - u_g^{(t)} \rangle + \frac{1}{2\alpha}\|u_f^{(t+1)} - u_g^{(t)}\|_2^2$$

is compared to $\ell(u_f^{(t+1)})$. If $\ell(u_f^{(t+1)}) \leq Q(u_f^{(t+1)}, \alpha)$, then the algorithm proceeds to (3.3). If $\ell(u_f^{(t+1)}) > Q(u_f^{(t+1)}, \alpha)$, then α is replaced with $\tau\alpha$, where $\tau \in (0, 1)$ is a constant, and (3.2) is carried out again. This process is repeated until $\ell(u_f^{(t+1)}) \leq Q(u_f^{(t+1)}, \alpha)$.

The efficiency of this algorithm hinges on the ability to compute the proximal operators of the functions g and f efficiently. As we will show momentarily, we can write the optimization problem from (2.1) as (3.1) and the corresponding g and f have proximal operators which can be solved in closed form.

3.2. Application to proposed estimator

In order to express the problem in (2.1) in a form analogous to (3.1), we must define the corresponding ℓ, f, and g. First, let $\chi_\epsilon : \mathbb{R}^{p \times p} \to \{0, \infty\}$ be the

function $\chi_\epsilon(\Omega) = \infty \cdot \mathbf{1}(\{\epsilon I_p \succ \Omega\} \cup \{\Omega \neq \Omega^\top\})$, with the convention $\infty \cdot 0 = 0$. Then, the unconstrained objective function from (2.1) is

$$\sum_{h=1}^{H} \{ \|\widehat{\Theta}_{(h)} - \omega_{(h)}\mathbb{1}_p^\top - \mathbb{1}_p\omega_{(h)}^\top + 2\Omega_{(h)}\|_F^2 + \lambda\|\Omega_{(h)}^-\|_1 + \chi_\epsilon(\Omega_{(h)})\} + \gamma \sum_{j\neq k} \|\boldsymbol{\Omega}_{\cdot jk}\|_2. \tag{3.5}$$

If we minimize (3.5) over all $\boldsymbol{\Omega} \in \mathbb{R}^{H\times p\times p}$, the minimizer with respect to each $\Omega_{(h)}$ must belong to the set $\{\Omega \in \mathbb{R}^{p\times p} : \Omega = \Omega^\top, \Omega \succcurlyeq \epsilon I_p\}$. Thus, defining $\ell(\boldsymbol{\Omega}) = \sum_{h=1}^H \|\widehat{\Theta}_{(h)} - \omega_{(h)}\mathbb{1}^\top - \mathbb{1}\omega_{(h)}^\top + 2\Omega_{(h)}\|_F^2$, $f(\boldsymbol{\Omega}) = \lambda\sum_{h=1}^H \|\Omega_{(h)}^-\|_1 + \gamma\sum_{j\neq k}\|\boldsymbol{\Omega}_{\cdot jk}\|_2$, and $g(\boldsymbol{\Omega}) = \sum_{h=1}^H \chi_\epsilon(\Omega_{(h)})$, (3.5) has the form of (3.1). Moreover, f and g are closed, proper, and convex functions; and the function ℓ is convex and differentiable with Lipschitz continuous gradient.

Specifically, letting $\widehat{\boldsymbol{\Theta}} \in \mathbb{R}^{H\times p\times p}$ be the three-way tensor made up of $\widehat{\Theta}_{(1)}, \ldots, \widehat{\Theta}_{(H)}$ so that $\widehat{\boldsymbol{\Theta}}_{hjk} = \widehat{\Theta}_{(h)jk}$, the function ℓ is differentiable with respect to $\boldsymbol{\Omega}$ with gradient

$$[\nabla\ell(\boldsymbol{\Omega})]_{hjk} = \begin{cases} \sum_{l\in[p]\setminus\{j\}} (4\boldsymbol{\Omega}_{hjj} - 4\widehat{\boldsymbol{\Theta}}_{hjl} - 8\boldsymbol{\Omega}_{hjl} + 4\boldsymbol{\Omega}_{hll}) & : j = k \\ 8\boldsymbol{\Omega}_{hjk} - 4\boldsymbol{\Omega}_{hjj} - 4\boldsymbol{\Omega}_{hkk} + 4\widehat{\boldsymbol{\Theta}}_{hjk} & : j \neq k \end{cases},$$

for all $(h, j, k) \in [H]\times[p]\times[p]$. The updating equations corresponding to (3.2)–(3.4) are

$$\boldsymbol{\Omega}^{(t+1)} = \underset{\boldsymbol{\Omega}\in\mathbb{R}^{H\times p\times p}}{\arg\min} \left\{ \frac{1}{2}|||\boldsymbol{\Omega} - \boldsymbol{\Lambda}^{(t)}|||_F^2 + \alpha\lambda\sum_{h=1}^H \|\boldsymbol{\Omega}_{(h)}^-\|_1 + \alpha\gamma\sum_{j\neq k}\|\boldsymbol{\Omega}_{\cdot jk}\|_2 \right\} \tag{3.6}$$

$$\tilde{\boldsymbol{\Omega}}^{(t+1)} = \underset{\boldsymbol{\Omega}\in\mathbb{R}^{H\times p\times p}}{\arg\min} \left\{ \frac{1}{2}|||\boldsymbol{\Omega} - \boldsymbol{\Omega}^{(t+1)} - \alpha\boldsymbol{\Psi}^{(t)}|||_F^2 + \alpha\sum_{h=1}^H \chi_\epsilon(\Omega_{(h)}) \right\} \tag{3.7}$$

$$\boldsymbol{\Psi}^{(t+1)} = \boldsymbol{\Psi}^{(t)} + \alpha^{-1}(\boldsymbol{\Omega}^{(t+1)} - \tilde{\boldsymbol{\Omega}}^{(t+1)}), \tag{3.8}$$

where $\boldsymbol{\Lambda}^{(t)} = \tilde{\boldsymbol{\Omega}}^{(t)} - \alpha\boldsymbol{\Psi}^{(t)} - \alpha\nabla\ell(\tilde{\boldsymbol{\Omega}}^{(t)})$ and $|||\boldsymbol{A}|||_F^2 = \sum_{h,j,k}\boldsymbol{A}_{hjk}^2$ for a three-way tensor $\boldsymbol{A}$. Because (3.8) is immediate, we focus on (3.6) and (3.7).

First, (3.6) can be separated across the second and third mode of $\boldsymbol{\Omega}$ since for all $(j, k) \in [p]\times[p]$ such that $j \neq k$,

$$\boldsymbol{\Omega}_{\cdot jk}^{(t+1)} = \underset{\nu\in\mathbb{R}^H}{\arg\min} \left\{ \frac{1}{2}\|\nu - \boldsymbol{\Lambda}_{\cdot jk}^{(t)}\|_2^2 + \alpha\lambda\|\nu\|_1 + \alpha\gamma\|\nu\|_2 \right\}, \tag{3.9}$$

and $\boldsymbol{\Omega}_{\cdot jj}^{(t+1)} = \boldsymbol{\Lambda}_{\cdot jj}^{(t)}$ for $j \in [p]$. The solution to (3.9) is

$$\boldsymbol{\Omega}_{\cdot jk}^{(t+1)} = \left(1 - \frac{\alpha\gamma}{\|\mathbf{soft}(\boldsymbol{\Lambda}_{\cdot jk}^{(t)}, \alpha\lambda)\|_2}\right)_+ \mathbf{soft}\left(\boldsymbol{\Lambda}_{\cdot jk}^{(t)}, \alpha\lambda\right),$$

where $(a)_+ = \max(a, 0)$ and $\mathbf{soft}(y, \tau) = \max(|y| - \tau, 0)\text{sign}(y)$ is applied elementwise [34]. The second step, (3.7), also has a closed form solution. In particular, (3.7) can be solved with respect to each $\Omega_{(h)}$ separately, in parallel, using that

$$\begin{aligned}\Omega_{(h)}^{(t+1)} &= \underset{\Omega_{(h)} \in \mathbb{R}^{p\times p}}{\arg\min} \left\{ \frac{1}{2} \|\Omega_{(h)} - \Omega_{(h)}^{(t+1)} - \alpha \mathbf{\Psi}_{h\cdot\cdot}^{(t)}\|_F^2 + \chi_\epsilon \left(\Omega_{(h)}\right) \right\} \\ &= \sum_{j=1}^{p} u_{(h)j} u_{(h)j}^\top \max(\xi_{(h)j}, \epsilon),\end{aligned}$$

where $u_{(h)j}$ and $\xi_{(h)j}$ are the jth eigenvector and eigenvalue of $\Omega_{(h)}^{(t+1)} + \alpha \mathbf{\Psi}_{h\cdot\cdot}^{(t)}$, respectively, for $h \in [H]$ [17]. This is the projection of $\Omega_{(h)}^{(t+1)} + \alpha \mathbf{\Psi}_{h\cdot\cdot}^{(t)}$ onto the convex set $\{\Omega \in \mathbb{R}^{p\times p} : \Omega \succcurlyeq \epsilon I_p \text{ and } \Omega = \Omega^\top\}$.

The convergence of the algorithm follows immediately from results in Pedregosa and Gidel [26]. The specific algorithm we implement is given in Algorithm 1. Without the positive definiteness constraint (e.g., by taking $\epsilon = -\infty$), a version of this algorithm simplifies to the standard proximal gradient descent algorithm [25, Chapter 4.2].

Enforcing the positive definiteness constraint on each of the $\Omega_{(h)}$ requires the eigendecomposition of a $p \times p$ matrix, a computation costing $O(p^3)$ floating point operations. To reduce the computational burden imposed by this additional constraint, our software implementation first solves (2.1) without the positive definiteness constraint. To do so, we use accelerated proximal gradient descent [25, Section 4.3]. If the solution to the unconstrained problem satisfies the constraint, then we know this is also the solution to the constrained problem. If the solution does not satisfy the constraint, we then apply the proximal-proximal gradient descent algorithm, Algorithm 1, initializing at the solution to the unconstrained problem. This scheme can significantly improve the computing time, especially when n is large relative to p.

An R package implementing our estimators, along with code for reproducing the results in Section 6, can be downloaded from `https://github.com/ajmolstad/SpPDCC`.

4. Practical considerations

To select tuning parameters (λ, γ), we use V-fold cross-validation. Given candidate tuning parameter sets $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$, for (2.1) we select tuning parameters according to

$$\underset{(\lambda,\gamma)\in\boldsymbol{\lambda}\times\boldsymbol{\gamma}}{\arg\min} \sum_{v=1}^{V}\sum_{h=1}^{H} \|\widehat{\Theta}_{(h),v} - \tilde{\omega}_{(h),-v}^{\lambda,\gamma} \mathbb{1}_p^\top - \mathbb{1}_p [\tilde{\omega}_{(h),-v}^{\lambda,\gamma}]^\top + 2\,\widetilde{\Omega}_{(h),-v}^{\lambda,\gamma}\|_F^2,$$

where $\widetilde{\mathbf{\Omega}}_{-v}^{\lambda,\gamma}$ is the solution to (2.1) with input sample variation matrices $\widehat{\mathbf{\Theta}}_{-v}$, which are computed using all the data from outside the vth fold.

Algorithm 1 Adaptive proximal-proximal gradient descent algorithm for computing multiple covariance matrices for compositional data.

Initialize $\mathbf{\Psi}^{(0)} \in \mathbb{R}^{H\times p\times p}, \tilde{\mathbf{\Omega}}^{(0)} \in \mathbb{R}^{H\times p\times p}$, $\alpha > 0$, and $\tau \in (0,1)$. Set $t = 0$ and proceed to **1.**

1. For $(j,k) \in [p] \times [p]$

1.1. If $j = k$

1.1.1. Set $\mathbf{\Omega}^{(t+1)}_{\cdot jj} = \tilde{\mathbf{\Omega}}^{(t)}_{\cdot jj} - \alpha\mathbf{\Psi}^{(t)}_{\cdot jj} - \alpha[\nabla\ell(\tilde{\mathbf{\Omega}}^{(t)})]_{\cdot jj}$

1.2. If $j \neq k$

1.2.1. Set $y = \tilde{\mathbf{\Omega}}^{(t)}_{\cdot jk} - \alpha\mathbf{\Psi}^{(t)}_{\cdot jk} - \alpha[\nabla\ell(\tilde{\mathbf{\Omega}}^{(t)})]_{\cdot jk}$

1.2.2. Set $w_h = (|y_h| - \alpha\lambda)_+ \text{sign}(y_h)$ for $h \in [H]$

1.2.3. Set $\mathbf{\Omega}^{(t+1)}_{\cdot jk} = \left(1 - \frac{\alpha\gamma}{\|w\|_2}\right)_+ w$

2. If $\ell(\mathbf{\Omega}^{(t+1)}) \leq Q(\mathbf{\Omega}^{(t+1)}, \alpha)$, proceed to **3.** Else, set $\alpha = \tau\alpha$, and return to **1.**

3. For $h \in [H]$

3.1. Decompose $\mathbf{\Omega}^{(t+1)}_{h\cdot\cdot} + \alpha\mathbf{\Psi}^{(t)}_{h\cdot\cdot} = \sum_{j=1}^p \xi_j \boldsymbol{u}_j\boldsymbol{u}_j^\top$, where $\boldsymbol{u}_j^\top\boldsymbol{u}_k = 0$ for $j \neq k$ and $\|\boldsymbol{u}_j\|_2 = 1$ for all $j \in [p]$

3.2. Set $\tilde{\mathbf{\Omega}}^{(t+1)}_{h\cdot\cdot} = \sum_{j=1}^p \max(\xi_j, \epsilon)\boldsymbol{u}_j\boldsymbol{u}_j^\top$

4. Set $\mathbf{\Psi}^{(t+1)} = \mathbf{\Psi}^{(t)} + \alpha^{-1}(\mathbf{\Omega}^{(t+1)} - \tilde{\mathbf{\Omega}}^{(t+1)})$

5. If the objective function value converged, terminate. Else, set $t = t+1$ and go to **1.**

In some applications, it may be preferable to let populations with larger samples sizes have a greater effect on the objective function. To do so, we propose an alternative variation of (2.1), defined as the argument minimizing

$$\sum_{h=1}^H \left(\frac{n_{(h)}}{N}\|\widehat{\Theta}_{(h)} - \omega_{(h)}\mathbb{1}_p^\top - \mathbb{1}_p\omega_{(h)}^\top + 2\Omega_{(h)}\|_F^2 + \lambda\|\Omega^-_{(h)}\|_1\right) + \gamma\sum_{j\neq k}\|\mathbf{\Omega}_{\cdot jk}\|_2 \tag{4.1}$$

subject to $\Omega_{(h)} = \Omega_{(h)}^\top$, $\Omega_{(h)} \succcurlyeq \epsilon I_p$ for all $h \in [H]$, where $N = \sum_{h=1}^H n_{(h)}$. The objective function (4.1) accounts for the distinct sample sizes by weighting each populations' contribution to the likelihood according to its relative contribution to the total sample size N. When using this estimator, we also recommend modifying the cross-validation criterion so that the tuning parameter pair selected is

$$\underset{(\lambda,\gamma)\in\boldsymbol{\lambda}\times\boldsymbol{\gamma}}{\arg\min}\sum_{v=1}^V\sum_{h=1}^H \frac{n_{(h),v}}{N_v}\|\widehat{\Theta}_{(h),v} - \tilde{\omega}^{\lambda,\gamma}_{(h),-v}\mathbb{1}_p^\top - \mathbb{1}_p[\tilde{\omega}^{\lambda,\gamma}_{(h),-v}]^\top + 2\,\widetilde{\Omega}^{\lambda,\gamma}_{(h),-v}\|_F^2,$$

where $n_{(h),v}$ is the number of samples in the vth fold from the hth population, and $N_v = \sum_{h=1}^H n_{(h),v}$. We compare the performance of (4.1) to (2.1), among other competitors, in Section A.3.

5. Statistical properties

5.1. Asymptotics for single population estimator

Though our primary focus is the multipopulation estimator (2.1), the estimator (2.2) is itself a novel and useful estimator of a single basis covariance matrix. In this section, we study its asymptotic properties. Specifically, we study $\widehat{\Omega}$ defined as

$$\arg\min_{\Omega=\Omega^\top} \left\{ \|\widehat{\Theta} - \omega \mathbb{1}_p^\top - \mathbb{1}_p \omega^\top + 2\Omega\|_F^2 + \lambda \|\Omega^-\|_1 \right\} \quad \text{subject to} \quad \Omega \succcurlyeq \epsilon I_p. \tag{5.1}$$

We will require the following assumptions.

A1. (Sub-Gaussian log abundances). The sample variation matrix, $\widehat{\Theta}$, is computed from n independent and identically distributed samples $W = (W_1, \ldots, W_p)^\top$ such that each $\log(W_j)$ is sub-Gaussian and $\mathrm{Cov}\{\log(W)\} = \Omega^*$.

A2. (Row-wise sparsity). As $n \to \infty$, $\max_j s_j/p \to 0$ where s_j is the number of nonzero off-diagonal entries in the jth row of Ω^*.

A3. (Alignment of n and p). As $n \to \infty$, $p \to \infty$ and $\log(p)/n \to 0$.

The assumptions **A1**—**A3** are natural in the context of high-dimensional compositional data. Assumption **A2** is needed to establish the restricted strong convexity of the loss function $\|\widehat{\Theta} - \omega \mathbb{1}_p^\top - \mathbb{1}_p \omega^\top + 2\Omega\|_F^2$ in a neighborhood of Ω^*. Cao, Lin and Li [7] assume similar sparsity, which in their setting ensures approximate identifiability (see their Proposition 1 and the comments following it). An analogous interpretation is possible here: unidentifiable parameters often lead to loss functions that are constant in some directions, but Assumption **A2** ensures the loss function is strictly convex around Ω^* in the directions that matter [see 24, Section 2.4, for details]. For this assumption to hold, we need p to grow with n. This is congruous with the assumptions in Cao, Lin and Li [7], who require $p \to \infty$ as $n \to \infty$ for consistency and characterize this as a "blessing of dimensionality". Assumption **A3** is standard in high-dimensional covariance matrix estimation.

We now state our first result concerning the asymptotic error of our estimator. The proof of this and all subsequent results can be found in the Appendix. Recall $s = \sum_{j=1}^p s_j$ and let φ_p be the pth largest eigenvalue of its matrix-valued argument.

Theorem 5.1. *Suppose **A1**–**A3** hold. If $\epsilon < \varphi_p(\Omega^*)$ and $\lambda = \sqrt{c_1 \log(p)/n}$ for fixed constant $c_1 > 0$ sufficiently large, then there exists a constant $b_1 \in (0, \infty)$ such that*

$$\frac{\|[\widehat{\Omega} - \Omega^*]^-\|_F}{\sqrt{p}} + \|\widehat{\omega} - \omega^*\|_2 \leq b_1 \left(\sqrt{\frac{s \log(p)}{pn}} + \sqrt{\frac{p \log(p)}{n}} \right) \tag{5.2}$$

and $\|\widehat{\omega} - \omega^*\|_2 \leq b_1 \sqrt{p \log(p)/n}$ with probability tending to one as $n \to \infty$.

The error bound in (5.2) consists of two parts: the error for estimating off-diagonals and the diagonals. The asymptotic Euclidean norm error for the diagonals is $b_1\sqrt{p\log(p)/n}$. The Frobenius norm error for the off-diagonals, however, cannot be disentangled from the diagonal error. Though our results would seem to suggest that $\|[\widehat{\Omega}-\Omega^*]^-\|_F \leq b_1\sqrt{s\log(p)/n}$ with probability tending to one, we are only able to establish a bound for $\|[\widehat{\Omega}-\Omega^*]^-\|_F/\sqrt{p}+\|\widehat{\omega}-\omega^*\|_2$. We cannot isolate the asymptotic error for the off-diagonals because of the intrinsic connection between the diagonals and off-diagonals in the objective function (5.1). This is in contrast to some traditional covariance matrix estimators, where off-diagonals can be estimated in a way which is not dependent on the diagonals.

Note that although (5.1) is the solution to a penalized least squares problem, we do not assume any type of restricted eigenvalue condition [30]. Instead, in our proof we first show that $\widehat{\Omega}-\Omega^*$ belongs to a restricted set, then establish a quadratic lower bound on $\ell(\widehat{\Omega})-\ell(\Omega^*)-\text{tr}\{\nabla\ell(\Omega^*)^\top(\widehat{\Omega}-\Omega^*)\}$ over this set where here, ℓ is the objective function from (1.3) with $\lambda=0$. Our technique for establishing this bound may be applicable in other penalized least squares problems.

Direct comparison of our estimation error bound to those established in Cao, Lin and Li [7] is not possible as their results are given in terms of the spectral norm, and under a different set of assumptions.

5.2. Asymptotics for multiple population estimator

Next, we consider the multiple population estimator (2.1) with $\lambda=0$. By doing so, we are able to illustrate how our method exploits shared sparsity across the populations. Our results will apply with $N=\sum_{h=1}^{H}n_{(h)}$ tending to infinity. To establish error bounds for this estimator, we will need a slightly different set of assumptions than in the single population case.

A4. (Bounded log abundances). The sample variation matrix $\widehat{\Theta}_{(h)}$ is computed from $n_{(h)}$ independent and identically distributed samples $W_{(h)}=(W_{(h)1},\ldots,W_{(h)p})^\top$ such that $\log(W_{(h)k})\in[-L,L]$ for all $k\in[p]$ and $\text{Cov}\{\log(W_{(h)})\}=\Omega^*_{(h)}$ for all $h\in[H]$.

A5. (Fiber-wise sparsity). As $N\to\infty$, $\max_j \tilde{s}_j/p\to 0$ where $\tilde{s}_j=|\{k:\boldsymbol{\Omega}^*_{\cdot jk}\neq 0, k\neq j\}|$ for $j\in[p]$.

A6. (Nonvanishing $n_{(h)}/N$). There exists $\pi>0$ such that for N sufficiently large, $\min_{h\in[H]}n_{(h)}/N\geq\pi$.

A7. (Alignment of $n_{(h)}$, p, H, and L). As $N\to\infty$, $p\to\infty$, $\log(p)/n_{(h)}\to 0$, and $L^4H/n_{(h)}\to 0$ for all $h\in[H]$.

Assumption **A4** requires that the log abundances take values over the interval $[-L,L]$. When $W_{(h)k}$ is a normalized count—as is standard in microbiome data—this assumption requires that all counts are bounded away from zero and

infinity. A positive lower bound on $W_{(h)k}$ is often assumed implicitly in the analysis of compositional data. Of course, **A4** is stronger than **A1**, but allows us to establish a concentration inequality on the Euclidean norm of the fibers of $\nabla\ell(\boldsymbol{\Omega}^*)$. The quantity L will appear in our asymptotic error bound, so this assumption is not so restrictive since L can be arbitrarily large.

Assumption **A5** requires that the number of nonzero off-diagonal entries in any of the $\Omega^*_{(h)}$ does not grow too quickly with p. Like **A2**, **A5** implicitly requires that p grows as $N \to \infty$. Assumption **A6** is a requirement on how frequently, as $N \to \infty$, we sample from each of the H populations. This assumption requires that we do not systemically undersample from any of the H populations. Our error bounds will depend on π, so we can quantify how sampling affects estimation accuracy.

We are ready to state our asymptotic error bound for (2.1) with $\lambda = 0$. Our bound will depend on $\tilde{s} = \sum_{j=1}^{p} \tilde{s}_j$.

Theorem 5.2. *Suppose **A4**–**A7** hold. Define $\widehat{\boldsymbol{\Omega}}$ as the solution to (2.1) with $\lambda = 0$. Let $\boldsymbol{\omega}^* = (\omega^*_{(1)}, \ldots, \omega^*_{(H)})$ and $\widehat{\boldsymbol{\omega}} = (\widehat{\omega}_{(1)}, \ldots, \widehat{\omega}_{(H)})$. If $\epsilon < \min_{h\in[H]} \varphi_p(\Omega^*_{(h)})$ and $\gamma = \sqrt{c_2 L^4 H/\pi N} + \sqrt{c_2 \log(p)/\pi N}$ for fixed constant $c_2 > 0$ sufficiently large, then there exists a constant $b_2 \in (0, \infty)$ such that*

$$\frac{|||[\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*]^-|||_F}{\sqrt{p}} + \|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}^*\|_F \leq b_2 \left\{ \left(\frac{\sqrt{\tilde{s}} + p}{\sqrt{p}} \right) \left(\sqrt{\frac{L^4 H}{\pi N}} + \sqrt{\frac{\log(p)}{\pi N}} \right) \right\}$$

and

$$\|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}^*\|_F \leq b_2 \left(\sqrt{\frac{pL^4 H}{\pi N}} + \sqrt{\frac{p \log(p)}{\pi N}} \right)$$

with probability tending to one as $N \to \infty$.

The bound in Theorem 5.2 can be interpreted in a similar way as the bound in Theorem 5.1. Specifically, we cannot separate the error for estimating the off-diagonals of the $\Omega^*_{(h)}$ from the error for estimating the diagonals. In particular, where the diagonals and off-diagonals affect the error bound are through their contribution to numerator in the leftmost term of the error bound: the $\sqrt{\tilde{s}}$ comes from having to estimate nonzero entries in $\tilde{s}$ off-diagonals of the $\Omega^*_{(h)}$, whereas the $\sqrt{p}$ comes from having to estimate p diagonal entries in each $\Omega^*_{(h)}$.

Just as in Theorem 5.1, we can establish a bound specifically for the diagonals. If there exists a constant $b_3 \in (0, \infty)$ such that $L^4 H \leq b_3 \log(p)$ for N sufficiently large, which is natural since one would not expect H nor L to grow with N, we then achieve essentially the same result as in Theorem 5.1: there exists a constant $b_4 \in (0, \infty)$ such that $\|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}^*\|_F \leq b_4 \sqrt{p \log(p)/(\pi N)}$ with probability tending to one.

If each $\Omega^*_{(h)}$ had a substantial number of zeros which were not shared across all H populations, (2.1) should perform better with $\lambda > 0$. Specifically, we conjecture that in this case, one could replace **A5** with a combination of **A2** (applied to each population separately) and a relaxed version of **A5**, and obtain

an improved error bound relative to that in Theorem 5.2 by taking $\lambda > 0$. However, proving this type of result is technically challenging, and it is unclear whether some of our intermediate results can be generalized (e.g., Lemma B.1).

6. Numerical experiments

6.1. Data generating models and competing methods

In this section, we compare the proposed estimator, (2.1), to existing estimators under three data generating models, Models 1–3, with $H = 4$ and $(n, p) \in \{50, 100, 150\} \times \{40, 80, 120, 160, 200\}$ where $n_{(1)} = \cdots = n_{(H)} = n$. In each replication, we generate $\log(W_{(h)1}), \ldots, \log(W_{(h)n_{(h)}})$ for $h \in [H]$ independently with each $\log(W_{(h)i}) \in \mathbb{R}^p$ drawn from $\mathrm{N}_p(0, \Omega^*_{(h)})$.

The three models allow us to examine the methods' performance under different types of shared sparsity. In Model 1, all covariance matrices are tridiagonal with $\Omega^*_{(1)} = \Omega^*_{(2)}$ and $\Omega^*_{(3)} = \Omega^*_{(4)}$. This is the ideal scenario for our method since the sparsity patterns are identical across populations. In Model 2, only one $p/4 \times p/4$ diagonal block is nonzero in each covariance matrix, though the block is in a different position for each $\Omega^*_{(h)}$. Finally, in Model 3, $\Omega^*_{(1)}$ and $\Omega^*_{(4)}$ do not share any nonzero off-diagonal elements, though $\Omega^*_{(2)}$ and $\Omega^*_{(3)}$ share some nonzero off-diagonals with each other, and with both $\Omega^*_{(1)}$ and $\Omega^*_{(4)}$.

The specific models we consider are as follows.

Model 1. The $\Omega^*_{(h)}$ are tridiagonal with either all positive or all negative correlations:

$$\mathbf{\Omega}^*_{(h)jk} = \begin{cases} 0.3 \cdot \mathbf{1}(1 \leq |j-k| \leq 2) + \mathbf{1}(j=k) & : h \in \{1, 2\} \\ -0.2 \cdot \mathbf{1}(1 \leq |j-k| \leq 2) + \mathbf{1}(j=k) & : h \in \{3, 4\} \end{cases},$$

for $(h, j, k) \in [4] \times [p] \times [p]$.

Model 2. The $\Omega^*_{(h)}$ are block diagonal with each block having an AR(1) structure:

$$\mathbf{\Omega}^*_{(h)jk} = \begin{cases} 0.8^{|j-k|} & : |j-k| < p/4,\ (j,k) \in \mathcal{A}_h \\ 1 & : (j,k) \notin \mathcal{A}_h,\ j = k \end{cases}, \ (h, j, k) \in [4] \times [p] \times [p],$$

and $\mathcal{A}_1 = [p/4] \times [p/4]$, $\mathcal{A}_2 = \{p/4+1, \ldots, p/2\} \times \{p/4+1, \ldots, p/2\}$, $\mathcal{A}_3 = \{p/2+1, \ldots, 3p/4\} \times \{p/2+1, \ldots, 3p/4\}$, and $\mathcal{A}_4 = \{3p/4+1, \ldots, p\} \times \{3p/4+1, \ldots, p\}$.

Model 3. The $\Omega^*_{(h)}$ have heterogeneous variances and are block diagonal with diagonal blocks having an AR(1) structure, i.e., $\Omega^*_{(h)} = D\Xi^*_{(h)}D$ where

$$\Xi^*_{(h)jk} = \begin{cases} 0.9^{|j-k|} & : (j,k) \in \mathcal{B}_h \\ 1 & : (j,k) \notin \mathcal{B}_h,\ j = k \end{cases}, \ (h, j, k) \in [4] \times [p] \times [p],$$

$D \in \mathbb{R}^{p \times p}$ is a diagonal matrix with diagonal entries equally spaced from 3 to 1, and $\mathcal{B}_1 = [p/2] \times [p/2]$, $\mathcal{B}_2 = \{p/6+1, \ldots, 2p/3\} \times \{p/6+1, \ldots, 2p/3\}$,

$\mathcal{B}_3 = \{p/3+1, \ldots, 5p/6\} \times \{p/3+1, \ldots, 5p/6\}$, and $\mathcal{B}_4 = \{p/2+1, \ldots, p\} \times \{p/2+1, \ldots, p\}$

In order to select tuning parameters for each of the methods, we also generate independent validation sets of the same size as the training set.

To estimate $\mathbf{\Omega}^*$, we consider multiple methods. All methods but (2.1) estimate $\Omega^*_{(1)}, \ldots, \Omega^*_{(4)}$ separately. Specifically, we use the method of Cao, Lin and Li [7] with adaptive soft-thresholding, `COAT`, and use the method of Fang et al. [10], `cclasso`. We also use an oracle estimator, `Oracle`, which is the adaptively soft-thresholded sample covariance matrix of each $\log(W_{(h)1}), \ldots, \log(W_{(h)n_{(h)}})$. This is an oracle method in the sense that we do not have access to the underlying abundances in practice. Finally, we consider two versions of our method, `SCC` and `SCC-H`, where `SCC` is short for sparse compositional covariance matrices. The estimator `SCC` is defined in (2.1), whereas `SCC-H` is (2.2) with a separate tuning parameter λ chosen for each $h \in [H]$. The method `SCC-H` estimates the covariances separately, but using a version of our criterion. Including both `SCC` and `SCC-H` serves to illustrate to what degree the improvement in performance is driven by the loss function versus the sharing of sparsity patterns across the fibers of $\mathbf{\Omega}^*$.

To assess the performance of each method, we measure the average (over H populations and 50 independent replications) Frobenius norm error and L_1 matrix norm error of the estimated covariance matrices on the correlation scale. We use the correlation scale because `cclasso` was designed specifically for correlation matrix estimation. Relative performances under both Frobenius norm and L_1 matrix norm error on the covariance scale are similar and thus relegated to the Appendix.

We also measure true positive (TPR) and true negative rates (TNR) for each method so that we may assess the recovery of nonzero correlations. Given an estimate of $\mathbf{\Omega}^*$, $\widehat{\mathbf{\Omega}}$, TPR and TNR are defined as, respectively,

$$\frac{1}{H}\sum_{h=1}^{H}\frac{|\{(j,k) : \widehat{\mathbf{\Omega}}_{hjk} \neq 0 \cap \mathbf{\Omega}^*_{hjk} \neq 0\}|}{|\{(j,k) : \mathbf{\Omega}^*_{hjk} \neq 0\}|}, \quad \frac{1}{H}\sum_{h=1}^{H}\frac{|\{(j,k) : \widehat{\mathbf{\Omega}}_{hjk} = 0 \cap \mathbf{\Omega}^*_{hjk} = 0\}|}{|\{(j,k) : \mathbf{\Omega}^*_{hjk} = 0\}|}.$$

6.2. Results

In Figure 3, we display average Frobenius norm errors (divided by p). With $p = 40$, the `cclasso` software would sometimes return undefined estimates (`NA` in R), so we omit comparisons in these settings. Unsurprisingly, under Model 1, `SCC` substantially outperforms all of the competitors, including `Oracle`. This illustrates the utility of exploiting shared sparsity patterns when estimating multiple covariance matrices. Notably, `Oracle`, `COAT`, and `SCC-H` all perform similarly in each setting under Model 1. Under Model 2, `SCC` outperforms all competitors, including `Oracle`, once $p \geq 120$. Comparing the competitors which could be used in practice, `SCC-H` performs better than `COAT` and `cclasso` in all situations. The estimator `COAT` performs worse than `cclasso` for small p, but

TABLE 1
True positive and true negative rates for each of the methods averaged over 50 independent replications under Model 1–3.

		$n = 50$					$n = 100$					$n = 150$				
		40	80	120	160	200	40	80	120	160	200	40	80	120	160	200
Model 1																
TPR	SCC	0.953	0.928	0.904	0.890	0.875	0.999	0.998	0.996	0.996	0.996	1.000	1.000	1.000	1.000	0.999
	SCC-H	0.570	0.499	0.448	0.422	0.397	0.832	0.778	0.747	0.719	0.695	0.924	0.905	0.884	0.867	0.851
	COAT	0.636	0.548	0.484	0.449	0.427	0.888	0.820	0.784	0.757	0.730	0.957	0.930	0.907	0.890	0.876
	Oracle	0.653	0.567	0.498	0.464	0.438	0.888	0.821	0.788	0.760	0.733	0.951	0.927	0.905	0.889	0.875
	cclasso	0.851	0.843	0.811	0.807	0.777	0.990	0.996	0.987	0.984	0.985	1.000	0.999	0.997	0.997	0.998
TNR	SCC	0.718	0.825	0.880	0.901	0.919	0.657	0.798	0.847	0.876	0.889	0.649	0.770	0.827	0.858	0.879
	SCC-H	0.892	0.949	0.968	0.977	0.983	0.800	0.890	0.923	0.941	0.952	0.771	0.860	0.900	0.920	0.933
	COAT	0.828	0.921	0.953	0.966	0.974	0.702	0.857	0.902	0.927	0.940	0.636	0.823	0.880	0.906	0.923
	Oracle	0.858	0.924	0.953	0.966	0.973	0.777	0.874	0.909	0.929	0.942	0.762	0.850	0.890	0.912	0.926
	cclasso	0.231	0.281	0.374	0.395	0.465	0.023	0.028	0.085	0.104	0.111	0.001	0.007	0.024	0.035	0.054
Model 2																
TPR	SCC	0.904	0.709	0.569	0.489	0.414	0.954	0.798	0.653	0.554	0.489	0.970	0.821	0.697	0.589	0.522
	SCC-H	0.766	0.529	0.404	0.333	0.277	0.856	0.631	0.486	0.401	0.340	0.883	0.672	0.529	0.434	0.374
	COAT	0.805	0.607	0.468	0.388	0.320	0.886	0.747	0.590	0.481	0.400	0.921	0.810	0.678	0.548	0.461
	Oracle	0.897	0.632	0.478	0.391	0.326	0.953	0.707	0.544	0.443	0.379	0.975	0.743	0.580	0.476	0.407
	cclasso		0.999	0.997	0.995	0.989	1.000	1.000	1.000	0.999	0.999	1.000	1.000	1.000	1.000	1.000
TNR	SCC	0.238	0.481	0.605	0.663	0.722	0.136	0.374	0.532	0.618	0.661	0.098	0.350	0.490	0.583	0.630
	SCC-H	0.536	0.718	0.792	0.827	0.858	0.462	0.644	0.749	0.797	0.824	0.436	0.625	0.725	0.776	0.807
	COAT	0.326	0.586	0.715	0.774	0.819	0.196	0.412	0.601	0.706	0.764	0.137	0.329	0.511	0.638	0.714
	Oracle	0.576	0.704	0.765	0.799	0.831	0.544	0.667	0.737	0.782	0.808	0.521	0.645	0.726	0.768	0.795
	cclasso		0.002	0.004	0.006	0.013	0.000	0.000	0.000	0.001	0.002	0.000	0.000	0.000	0.000	0.000
Model 3																
TPR	SCC	0.933	0.732	0.623	0.550	0.488	0.984	0.804	0.696	0.616	0.545	0.992	0.880	0.761	0.655	0.597
	SCC-H	0.696	0.556	0.463	0.404	0.342	0.794	0.654	0.567	0.492	0.430	0.831	0.696	0.618	0.534	0.469
	COAT	0.838	0.740	0.638	0.556	0.470	0.920	0.869	0.803	0.715	0.628	0.942	0.920	0.873	0.810	0.731
	Oracle	0.960	0.772	0.620	0.524	0.445	0.988	0.830	0.700	0.596	0.504	0.994	0.872	0.726	0.619	0.544
	cclasso		0.999	0.998	0.997	0.994		1.000	1.000	0.999	0.999	1.000	1.000	1.000	1.000	1.000
TNR	SCC	0.106	0.395	0.521	0.598	0.640	0.025	0.324	0.467	0.555	0.616	0.011	0.210	0.396	0.519	0.571
	SCC-H	0.433	0.652	0.744	0.791	0.825	0.322	0.561	0.668	0.738	0.784	0.292	0.518	0.639	0.713	0.759
	COAT	0.288	0.514	0.630	0.701	0.750	0.137	0.333	0.462	0.575	0.653	0.089	0.225	0.369	0.477	0.565
	Oracle	0.560	0.647	0.697	0.745	0.773	0.515	0.617	0.662	0.708	0.757	0.543	0.587	0.657	0.707	0.745
	cclasso		0.001	0.002	0.003	0.006		0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000

significantly outperforms `cclasso` once $p \geq 120$. Finally, under Model 3, `SCC` and `SCC-H` perform similarly in every setting. The only method to outperform `SCC` and `SCC-H` is `Oracle`, which cannot be used in practice. The fact that `Oracle` outperforms the other methods so substantially speaks to the difficulty of estimating the covariances under this data generating model relative to Model 1 and 2.

One should be careful drawing conclusions based on Frobenius norm error results alone, however, because our methods, `SCC` and `SCC-H`, both minimize a Frobenius norm criterion, whereas `COAT` does not. Thus, these results may be biased in favor of `SCC` and `SCC-H`. For this reason, we also included L_1 matrix norm results in Figure 4. Here, the L_1 matrix norm is the maximum of the L_1 vector-norm of the columns of a matrix. Under Model 1 with $n = 50$, there appears to be little difference between the methods—other than `cclasso`—in terms of L_1 matrix norm. However, when $n \geq 100$, `SCC` significantly outperforms

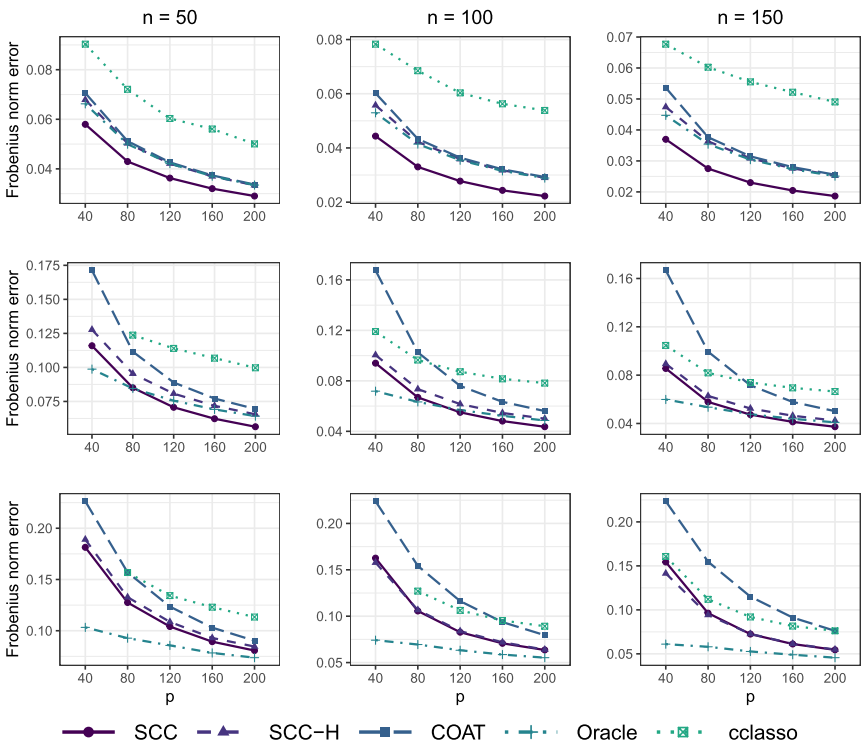


FIG 3. *Average Frobenius norm error divided by p (on the correlation scale) over 50 independent replications under (top row) Model 1, (middle row) Model 2, and (bottom row) Model 3 with* $(n, p) \in \{50, 100, 150\} \times \{40, 80, 120, 160, 200\}$.

all competitors. Under Model 2, the results more closely mirror those in Figure 3: `SCC` outperforms all competitors, including `Oracle`, when $p \geq 120$. The results under Model 3, relatively speaking, are similar to those observed under Model 3 using Frobenius norm error. The method `Oracle` performs best, but among the methods which could be used in practice, `SCC` and `SCC-H` clearly outperform `cclasso` and `COAT`.

The performance of `SCC` and `SCC-H` can be partially explained by their performance in recovering the true set of nonzero off-diagonals. In Model 1, `SCC` has nearly perfect TPR, and TNR only slightly lower than the best performing competitor. `SCC-H` tends to have similar TPR as `COAT`, but also tends to have higher TNR. A similar conclusion can be drawn under Model 2. Under Model 3, however, `COAT` tends to have higher TPR and similar TNR to `SCC`, whereas `SCC-H` has lower TPR and higher TNR than `COAT`. Note that under Model 3, `oracle` does well in part due to the fact that the covariances have varying diagonals.

In Section A of the Appendix, we provide additional simulation study results. First, we present the results from Figures 3 and 4, but on the covariance scale. Relative performances closely mirror those in Figures 3 and 4. Second, we present results comparing `SCC-H` to `COAT` and `cclasso` in terms of estimating

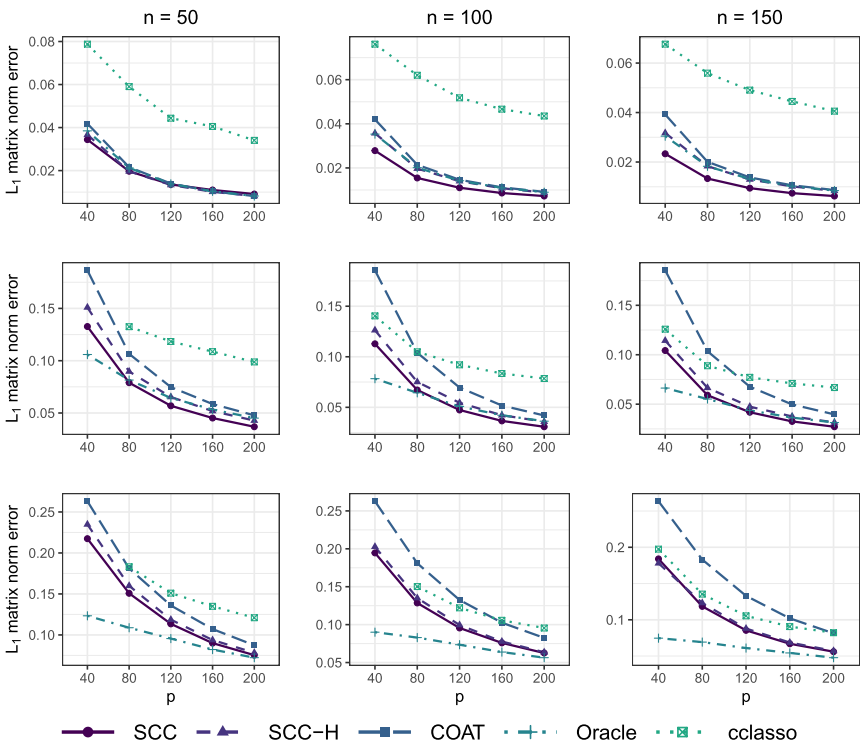


FIG 4. *Average L_1 matrix norm error divided by p (on the correlation scale) over 50 independent replications under (top row) Model 1, (middle row) Model 2, and (bottom row) Model 3 with $(n, p) \in \{50, 100, 150\} \times \{40, 80, 120, 160, 200\}$.*

$\Omega^*_{(1)}$ under Models 1–3. Our results clearly demonstrate that `SCC-H` can significantly outperform both `COAT` and `cclasso` for single population basis covariance matrix estimation. Finally, we also perform additional simulation studies wherein the sample sizes $(n_{(1)}, n_{(2)}, n_{(3)}, n_{(4)}) = (100, 75, 50, 25)$. We compare the same competitors as in Section 6.2, but also include (4.1). Under Model 1, (4.1) outperforms the competitors, though under Models 2 and 3, there is little difference between (2.1) and (4.1).

7. Analysis of microbiome in myalgic encephalomyelitis/chronic fatigue syndrome

7.1. Basis covariance matrix estimation

We illustrate our method by analyzing data on the gut microbiome of patients diagnosed with myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) versus controls from Giloteaux et al. [13]. In order to obtain the microbial profiles, Giloteaux et al. [13] sequenced 16S rRNA genes from stool samples using

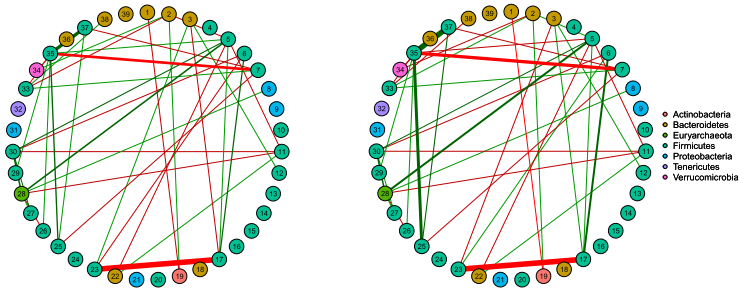
Illumina MiSeq. After first filtering patients based on total reads ($\geq$ 5000), we filter operational taxonomic units (OTUs) to only those that comprise at least 10% of total reads in one or more patients. This reduced the original 138 OTUs to $p = 39$ OTUs, though led to us filtering out only 8% and 10% of each subjects' total reads (on average) in controls and ME/CFS, respectively. Following Cao, Lin and Li [7], we add 0.5 to all counts to avoid zeros before converting counts to compositions. To be clear, these counts $Y_{(h)i}$ are distinct from the latent abundances $W_{(h)i}$, which are assumed to be independent and identically distributed for all $k \in [n_{(h)}]$. For example, we may assume that $Y_{(h)ij} = M_{(h)i}W_{(h)ij}$ where $M_{(h)i}$ is a positive random variable [22], so that $Y_{(h)ij}/\sum_{k=1}^{p} Y_{(h)ik} = X_{(h)ij} = W_{(h)ij}/\sum_{k=1}^{p} W_{(h)ik}$. In Section D of the Appendix, we demonstrate that our estimates do not change much when we use a pseudocount of 0.01 instead of 0.5 to handle zeros.

Our estimates of the covariance matrices, with tuning parameters chosen using ten-fold cross-validation, are in Figure 5. Each node in these graphs represents a unique OTU. The nodes are colored according to OTU's phylum and each node's family, genus, and species is provided in the Table 4. The thickness of the edge corresponds to the strength of the association: stronger associations are represented by thicker edges. Positive and negative correlations are colored, respectively, with green and red, while a zero correlation is represented by the absence of an edge.

Examining the estimated covariance matrices, the majority of associations occur within two OTUs belonging to the same phylum. We also see that our method estimates the two groups' covariance matrices to have identical sparsity patterns, in sharp contrast with the estimates based on `COAT`, the method of Cao, Lin and Li [7] (Figure 1). Most strong positive and negative associations are shared across the two groups. Notably, one of the eight associations whose direction differ across controls and ME/CFS is (23–3; *Ruminococcus bromii–Bacteroides ovatus*). *Ruminococcus bromii* is known to degrade resistant starch particles inaccessible to other bacteria, whereas *Bacteroides ovatus* digests inulin [27]. That these two OTUs are estimated to be associated is interesting since both resistant starch and inulin are fermentable carbohydrates whose joint behavior has been of interest in past studies [39].

In addition, an insight gleaned from our estimates is that more negative associations are observed in chronic fatigue syndrome patients than in controls. This coheres with the reduced diversity in the microbiome communities for ME/CFS patients observed by Giloteaux et al. [13]. Moreover, the positive associations between (25–35) and (6–17) are much stronger in ME/CFS than in controls. This too suggests reduced diversity in ME/CFS as OTUs labeled 6, 35, 25, and 17 all belong to the same phylum, *Firmicutes* (see Table 4 for details).

Estimates using `COAT` are more difficult to interpret. First, there is a larger number of nonzero entries in both estimates, and their sparsity patterns differ substantially. In total, `COAT` identifies 185 associations in one population not present in the other. Moreover, the estimates from `COAT` disagree in terms of their strongest associations. For example, one of the strongest positive associations estimated in controls is between (29–13), whereas in patients with ME/CFS,



(a) Controls (b) ME/CFS

FIG 5. *Estimated correlation networks using* (2.1) *for (a) control patients and (b) ME/CFS patients. The thickness of the edge corresponds to the strength of the association: stronger associations are represented by thicker edges. Positive and negative correlations are colored, respectively, with green and red, while a zero correlation is represented by the lack of an edge.*

their method estimates these two OTUs to be uncorrelated.

Finally, we emphasize that our estimator (2.1) does not require that sparsity patterns are identical across CFS and controls. Instead, the similarity of sparsity patterns is determined by the combination of tuning parameters (γ, λ), which are selected by cross-validation. Thus, in this application, it is the data which suggest that the sparsity patterns are identical.

7.2. Stability assessment

We perform a stability assessment to determine to what degree our respective estimates, displayed in Figures 1 and 5, are reliable. Following Cao, Lin and Li [7], we generate 100 independent bootstrap samples and refit both estimators to the bootstrapped samples. We say an estimated nonzero correlation is stable if it is nonzero in at least 80 of the 100 bootstrap samples. In Table 2, we report the stability of each correlation estimate.

In the first four columns, we assess the stability of all correlations: in rows labeled positive and negative, we report the number of correlations estimated to be positive and negative, respectively, in the estimates displayed in Figures 1 and 5. In the row labeled stability, we report the percentage of these correlations which were estimated to be nonzero at least 80 of the 100 bootstrap samples. For example, `SCC` estimated 22 positive and 16 negative correlations (Figure 5a); of these 38 correlations, 89.5% of them were estimated to be nonzero in at least 80 bootstrap samples. For our method, in both controls and ME/CFS basis covariance matrix estimates, almost all of the edges we estimated to be nonzero are stable. `COAT`, on the other hand, has lower stability in both controls and ME/CFS.

In the first row of the "shared correlations" columns of Table 2, we report the number of estimated correlations where a correlation was positive in both estimates (controls and ME/CFS), or negative in both estimates. Our method

TABLE 2

Stability for all correlations, shared correlations, and distinct correlations over 100 bootstrap samples. For the distinct correlation columns, D1 refers to a correlation which was nonzero in controls, but zero in ME/CFS, whereas D2 refers to a correlation which was zero in controls but nonzero in ME/CFS.

All correlations					Shared correlations			Distinct correlations		
	SCC		COAT							
	Control	ME/CFS	Control	ME/CFS		SCC	COAT		SCC	COAT
Positive	22	21	53	30	Same sign	29	19	D1	0	111
Negative	16	17	82	68	Diff. sign	9	5	D2	0	74
Stability	89.5%	86.8%	83.0%	84.7%	Stability	86.8%	58.3%	Stability	—	00.0%

estimated that all correlations are shared, and has reasonably high stability. In particular, this column suggests that of the 40 shared correlations, 90% were estimated in at least 80 bootstrap samples. `COAT` has slightly lower stability for its shared correlations despite estimating fewer than half as many as our method.

Finally, the most telling result comes in the "distinct correlations" columns of Table 2. Here, we report the number of correlations which were nonzero in controls and zero in ME/CFS (D1) and the number of correlations which were zero in controls and nonzero in ME/CFS (D2). We see that `SCC` estimates no correlations to be distinct, whereas `COAT` estimates 185 correlations to be distinct. However, the stability of these correlations is zero: none of these distinct correlations appeared in 80 or more of the bootstrap samples. These results suggest that the estimates provided by our method may be more reliable than `COAT`.

8. Discussion

In this article, we proposed a new method for estimating basis covariance matrices from compositional data. An important question about our method is whether it could provide reasonable estimates of the basis precision (inverse covariance) matrix. Though our method can provide estimates of $\Omega_{*(h)}^{-1}$ (since our estimates are always positive definite), these estimates will not, in general, be sparse. If a practitioner is interested in sparse precision matrix estimation, we recommend using methods specifically designed for this task, e.g., Zhang, Wang and Lin [40]. To the best of knowledge, there exist no methods for jointly estimating multiple sparse precision matrices from compositional data. This could be a fruitful direction for future research.

There are two aspects of our data analysis which could be improved. First, the original data were counts (reads per OTU), which we converted to compositions. It has been argued that total reads per patient is an experimental artifact, and thus, microbiome sequencing data should be converted to compositions [e.g., see 14, and references therein]. However, as pointed out by a referee, there is nonetheless some loss of information when we ignore total reads per patient. Ideally, an estimator could somehow make use of this additional information.

Second, our method assumes that components of the observed composition are positive with probability one. In 16S rRNA sequencing (microbiome) data, however, it is common to observe many zeros. Thus, as future work, we hope to extend our method to address these two issues.

Appendix A: Supplemental numerical experiments

A.1. Additional results from Section 6

In Figures 6 and 7, we display both the average Frobenius norm and average L_1 matrix norm errors on the covariance scale (i.e., for the $\Omega^*_{(h)}$ directly). In terms of Frobenius norm error, under Models 1 and 2, `SCC` outperforms all competitors. Under Model 3, `SCC-H` can sometimes outperform `SCC`: the same result as the correlation scale. Errors under Model 3 are much larger overall because the elements of the covariance matrices tend to be larger by construction. Result using the L_1 matrix norm as performance metric are similar to those on the correlation scale (Figure 4).

A.2. Performance of `SCC-H` with $H = 1$

Though our simulation study focused on the performance of `SCC` with $H = 4$, our simulation study can also provide some insight as to how `SCC-H` would perform in a setting where $H = 1$. In an application where $H = 1$, `COAT` and `SCC-H` are direct competitors in the sense that both are designed for estimation of a single sparse basis covariance matrix. To assess how the two estimators compare with $H = 1$, we consider the same simulation settings as in Section 6, but focus our attention on the estimation of $\Omega^*_{(1)}$ under Models 1–3. Average Frobenius norm errors $p^{-1}\|\Omega^*_{(1)} - \widehat{\Omega}_{(1)}\|_F$ and L_1 matrix norm errors $p^{-1}\max_{j\in[p]}(\sum_{k=1}^{p}|\Omega^*_{(1)jk} - \widehat{\Omega}_{(1)jk}|)$ are displayed in Figures 8 and 9. In both figures, `SCC-H` outperforms `COAT` and `cclasso` under all three models. The difference between `SCC-H` and `COAT` is quite small under Model 1 when p is large, but in all other scenarios, `SCC-H` clearly outperforms competitors. This suggests that our method could also be quite useful for estimating a single covariance basis covariance matrix.

A.3. Performance with imbalanced sample sizes

We consider the performance of our estimators when the sample sizes differ across the four populations. When the sample sizes are imbalanced, it may be useful to weight each populations' contribution to the loss function. In this section, we compare the same competitors as in Section 6 to `wSCC`, the estimator defined in (4.1). In the simulation settings considered in Section 6, (4.1) and (2.1) are equivalent as we set $n_{(h)} = n$ for each $h \in [H]$. Here, we consider the same data generating models, Model 1–3, but set $(n_{(1)}, n_{(2)}, n_{(3)}, n_{(4)}) = (100, 75, 50, 25)$.

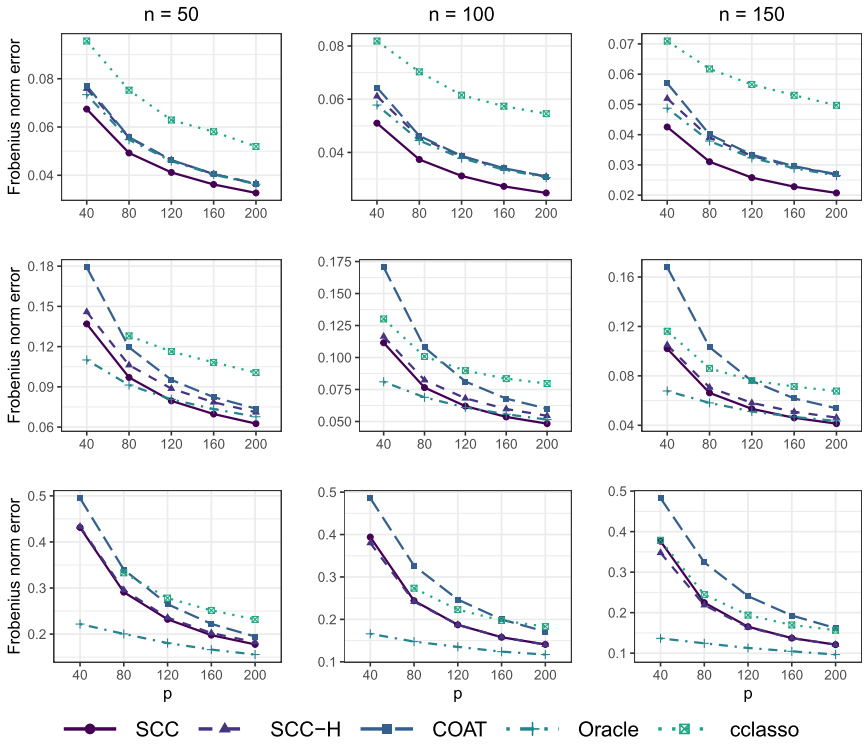


FIG 6. *Average Frobenius norm error divided by p (on the covariance scale) over 50 independent replications under (top row) Model 1, (middle row) Model 2, and (bottom row) Model 3 with* $(n, p) \in \{50, 100, 150\} \times \{40, 80, 120, 160, 200\}$.

Results are presented in Figure 10. In many replications with $p = 40$ or $p = 80$, the software for `cclasso` returned undefined estimates of some of the covariance matrices (`NA`s in R), so we omit `cclasso` in this case. Overall, under Model 1, `wSCC` appears to provide a substantial performance gain over `SCC`, `SCC-H`, `COAT`, and `cclasso`. Recall that it was Model 1 under which we saw the most substantial performance improvement in `SCC` relative to `SCC-H` in the simulation settings from Section 6. Under Models 2 and 3, however, the improvement `wSCC` provides relative to `SCC` is less substantial than under Model 1. We recommend that practicioners consider other variables when determining which version of our loss function to use: for example, is it reasonable to expect $\Omega^*_{(1)}$ and $\Omega^*_{(2)}$ to have elements on similar scales? If not, then perhaps another reweighting scheme (i.e., replacing $n_{(h)}/N$ in (4.1) with properly calibrated weights $w_h > 0$) would be preferable.

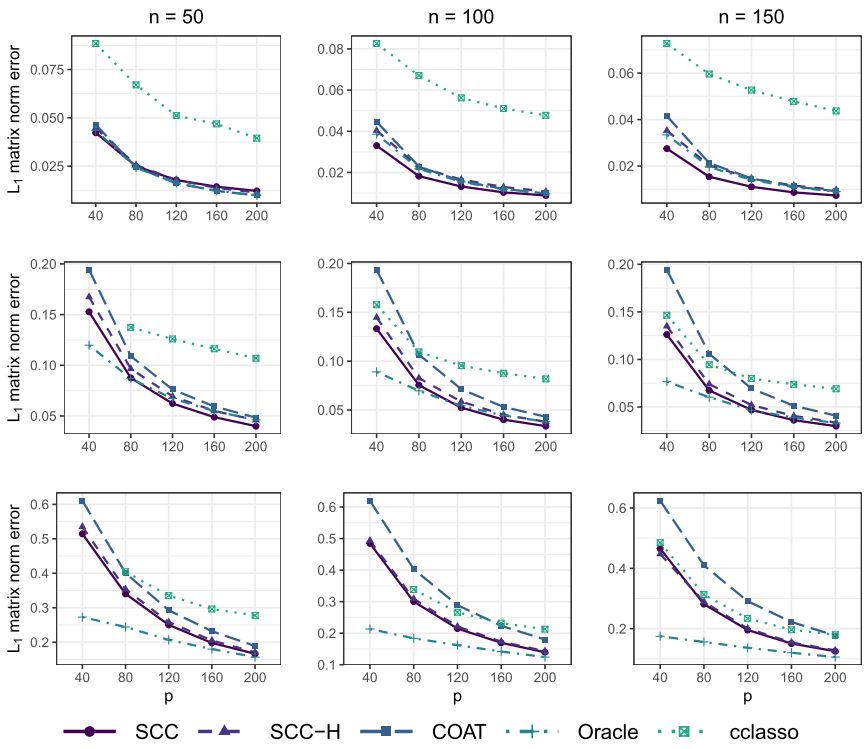


FIG 7. *Average L_1 matrix norm error divided by p (on the covariance scale) over 50 independent replications under (top row) Model 1, (middle row) Model 2, and (bottom row) Model 3 with $(n, p) \in \{50, 100, 150\} \times \{40, 80, 120, 160, 200\}$.*

Appendix B: Theorem proofs

B.1. Notation and key lemmas

For a tensor $\boldsymbol{\Delta} \in \mathbb{R}^{H\times p\times p}$, define the norm $|||\boldsymbol{\Delta}|||_{1,2} = \sum_{j,k} \|\boldsymbol{\Delta}_{\cdot jk}\|_2$ and for a set $\mathcal{M} \subseteq [p] \times [p]$ define $\boldsymbol{\Delta}_{\mathcal{M}}$ as the tensor whose (h, j, k)th entry equals $\boldsymbol{\Delta}_{hjk}$ if $(j, k) \in \mathcal{M}$ and zero otherwise. Let $\boldsymbol{\Delta}^-$ be the tensor which has (h, j, j)th entry equal to zero for all $j \in [p]$ and $h \in [H]$, but is otherwise equal to $\boldsymbol{\Delta}$, and let $\boldsymbol{\Delta}^+ = \boldsymbol{\Delta} - \boldsymbol{\Delta}^-$. Define $\boldsymbol{\xi} = \max_{j\neq k} \|\widehat{\boldsymbol{\Theta}}_{\cdot jk} - \boldsymbol{\tau}^*_{\cdot jk}\|_2$, $\boldsymbol{\eta} = \{\sum_{j=1}^p \|\sum_{k\neq j}(\widehat{\boldsymbol{\Theta}}_{\cdot jk} - \boldsymbol{\tau}^*_{\cdot jk})\|_2^2\}^{1/2}$, and $\boldsymbol{\tau}^*_{(h)jk} = \omega^*_{(h)j} + \omega^*_{(h)k} - 2\Omega^*_{(h)jk}$. Let $\check{s} = \max_j \tilde{s}_j$ and note that, when $H = 1$, $\tilde{s}_j = s_j$. When $H > 1$, define $n_{\min} = \min_{h\in[H]} n_{(h)}$. Lastly, define $\mathcal{S} = \{(j, k) : \Omega^*_{(h)jk} \neq 0 \text{ for any } h \in [H],\ (j, k) \in [p] \times [p]\}$ and $\mathcal{S}^c = [p] \times [p] \setminus \mathcal{S}$.

Let $\ell(\boldsymbol{\Omega}) = \sum_{h=1}^H \ell_{(h)}(\Omega_{(h)})$ where

$$\ell_{(h)}(\Omega) = \|\widehat{\Theta}_{(h)} - \omega \mathbb{1}_p^\top - \mathbb{1}_p \omega^\top + 2\Omega\|_F^2.$$

We denote the Hessian of $\mathrm{vec}(\Omega) \mapsto \ell_{(h)}(\Omega)$ by $\nabla^2\ell \in \mathbb{R}^{p^2\times p^2}$; it does not depend on Ω or h since $\ell_{(h)}$ is quadratic.

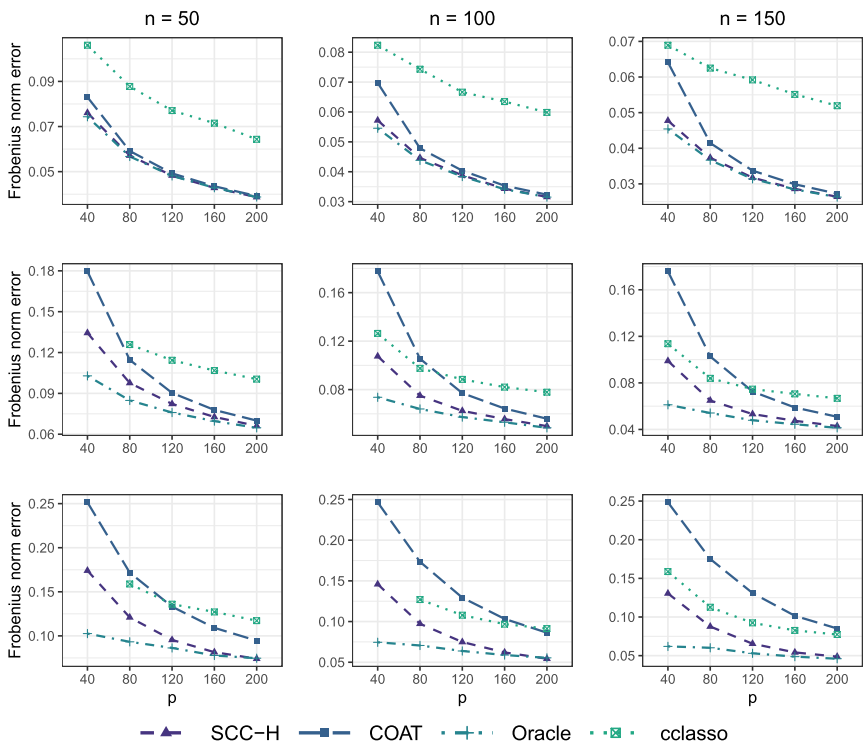


FIG 8. *Average Frobenius norm error divided by p for estimating* $\Omega^*_{(1)}$ *(on the correlation scale) over 50 independent replications under (top row) Model 1, (middle row) Model 2, and (bottom row) Model 3 with* $(n, p) \in \{50, 100, 150\} \times \{40, 80, 120, 160, 200\}$.

We will use the following lemmas. Proofs of the lemmas are in the subsequent section.

Lemma B.1 (Quadratic lower bound)**.** *For any* $a_1 > 0$ *and* $a_2 > 0$*, it holds for every* $\mathbf{\Delta} \in \mathbb{C}_a := \{\mathbf{\Delta} \in \mathbb{R}^{H\times p\times p} : |||\mathbf{\Delta}^-_{\mathcal{S}^c}|||_{1,2} \le a_1|||\mathbf{\Delta}^-_{\mathcal{S}}|||_{1,2} + a_2, \mathbf{\Delta}_{(h)\cdot\cdot} = \mathbf{\Delta}^\top_{(h)\cdot\cdot}, h \in [H]\}$ *that, when* $p \ge 5$*,*

$$\sum_{h\in[H]} \mathrm{vec}(\Delta_{(h)})^\top \nabla^2 \ell\, \mathrm{vec}(\Delta_{(h)}) \ge \left(4 - \frac{32\check{s}(1+a_1)^2}{p-4}\right)|||\mathbf{\Delta}^-|||_F^2 + p|||\mathbf{\Delta}^+|||_F^2 - \frac{32a_2^2}{p(p-4)}.$$

Lemma B.2. *Letting* $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Omega}} - \mathbf{\Omega}^*$*, it holds that*

$$\gamma(|||\widehat{\mathbf{\Delta}}^-_{\mathcal{S}}|||_{1,2} - |||\widehat{\mathbf{\Delta}}^-_{\mathcal{S}^c}|||_{1,2}) \ge \frac{1}{2}\sum_{h\in[H]} \mathrm{vec}(\widehat{\Delta}_{(h)})^\top \nabla^2 \ell\, \mathrm{vec}(\widehat{\Delta}_{(h)}) - 4\boldsymbol{\xi}|||\widehat{\mathbf{\Delta}}^-|||_{1,2} - 4\boldsymbol{\eta}|||\widehat{\mathbf{\Delta}}^+|||_F.$$

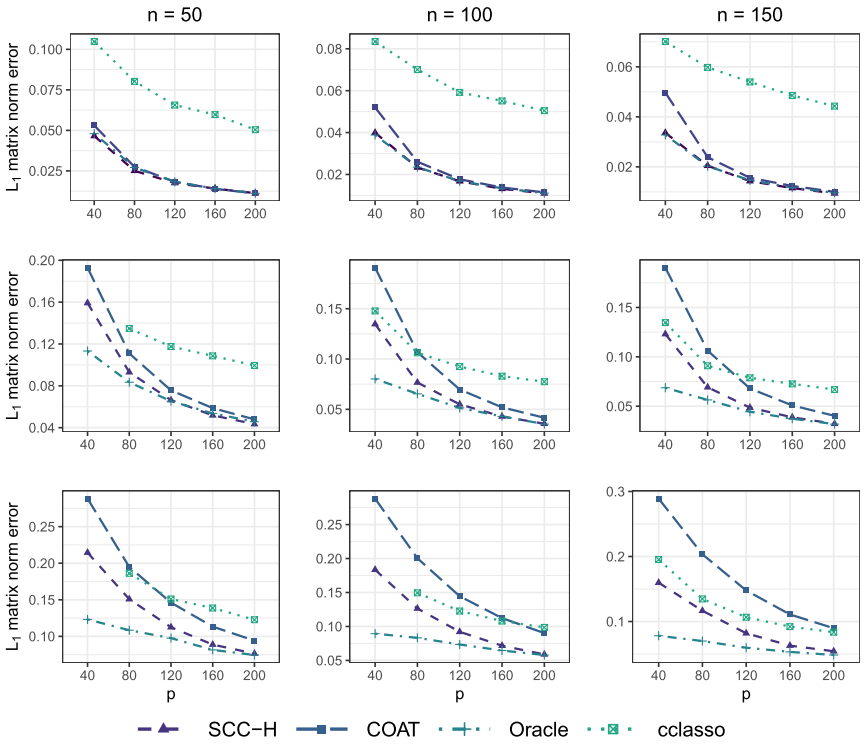


FIG 9. *L_1 matrix norm error divided by p for estimating $\Omega^*_{(1)}$ (on the correlation scale) over 50 independent replications under (top row) Model 1, (middle row) Model 2, and (bottom row) Model 3 with $(n_{(1)}, p) \in \{50, 100, 150\} \times \{40, 80, 120, 160, 200\}$.*

For the remainder, when $H = 1$, let W_{ij} be the jth component of ith observation's random basis vector. When $H > 1$, let $W_{(h)ij}$ be the jth component of the random basis vector for observation $i \in [n_{(h)}]$ from the hth population. The following lemma is used only when analyzing the proposed estimator for one population, that is, when $H = 1$.

Lemma B.3. *Suppose $H = 1$ and that the $\log(W_{ij})$ are independent over $i \in [n]$ with sub-Gaussian norms bounded by $K < \infty$. Let $c_1 > 0$ be a fixed constant. If $\lambda = \sqrt{c_1 \log(p)/n} \to 0$, then for n sufficiently large*

$$P\left(\max_{j \neq k} |\widehat{\Theta}_{jk} - \tau^*_{jk}| \geq \lambda\right) \leq 6p^{2-\nu c_1/K^4},$$

where $\nu > 0$ is a universal constant.

When $H > 1$, the following lemma is used in place of the previous.

Lemma B.4. *Suppose that the $\log(W_{(h)ij})$ are independent over $i \in [n_{(h)}], h \in [H]$, and that each $\log(W_{(h)ij})$ is supported on $[-L, L]$ for $L \in (0, \infty)$. If $\gamma =$*

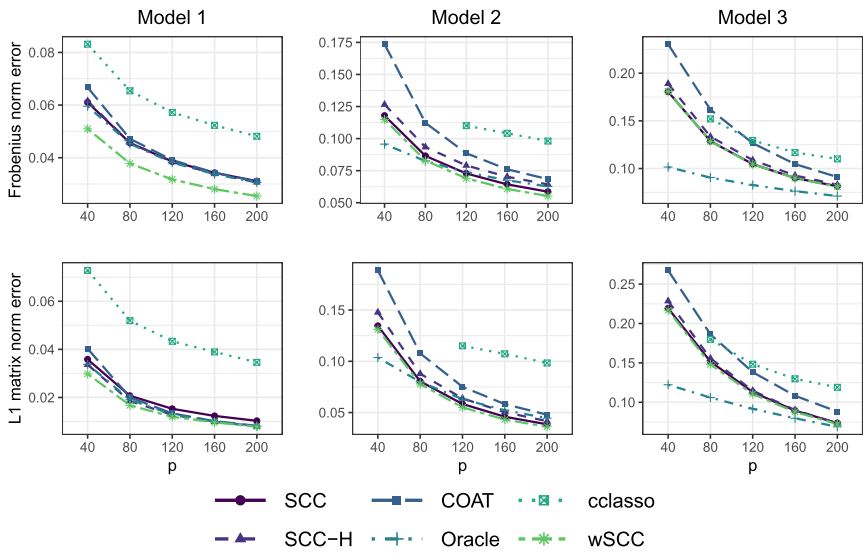


FIG 10. *Average Frobenius and L_1 matrix norm errors divided by p (on the correlation scale) over 50 independent replications under (left column) Model 1, (middle column) Model 2, and (right column) Model 3 with* $p \in \{40, 80, 120, 160, 200\}$ *and* $(n_{(1)}, n_{(2)}, n_{(3)}, n_{(4)}) = (100, 75, 50, 25)$.

$\{\sqrt{c_2 H L^4 / n_{\min}} + \sqrt{c_2 \log(p)/n_{\min}}\}$ *for fixed constant* $c_2 > 0$ *sufficiently large, then there exists a constant* $d_1 \in (0, \infty)$ *such that*

$$P\left(\max_{j \neq k} \|\widehat{\boldsymbol{\Theta}}_{\cdot jk} - \boldsymbol{\tau}^*_{\cdot jk}\|_2 \geq \gamma\right) \leq p^{2 - \frac{2c_2 n_{\min}}{d_1^2 L^4 N}}.$$

Note that the conditions of Lemma B.3 are satisfied under **A1** and **A3**, whereas the conditions of Lemma B.4 are satisfied under **A4**.

B.2. Proof of Theorem 5.1

Proof of Theorem 5.1. Specializing notation to the case with $H = 1$ and dropping the nonnegative quadratic term, Lemma B.2 implies

$$(\lambda - 4\xi)\|\widehat{\Delta}^-_{\mathcal{S}^c}\|_1 \leq (\lambda + 4\xi)\|\widehat{\Delta}^-_{\mathcal{S}}\|_1 + 4\eta\|\widehat{\Delta}^+\|_F \leq (\lambda + 4\xi)\|\widehat{\Delta}^-_{\mathcal{S}}\|_1 + 4\xi p^{3/2}\|\widehat{\Delta}^+\|_F,$$

where the last inequality follows from $\eta \leq \xi p^{3/2}$. Thus, by Lemma B.3, we can, for any $v_1 > 0$, pick $c_1 > 0$ sufficiently large so that with probability tending to one,

$$\|\widehat{\Delta}^-_{\mathcal{S}^c}\|_1 \leq 2\|\widehat{\Delta}^-_{\mathcal{S}}\|_1 + (p^{3/2}/v_1)\|\widehat{\Delta}^+\|_F. \tag{B.1}$$

Next, by Lemma B.1, specializing notation to the case $H = 1$ with $a_1 = 2$ and $a_2 = (p^{3/2}/v_1)\|\widehat{\Delta}^+\|_F$, (B.1) implies

$$\operatorname{vec}(\widehat{\Delta})^\top \nabla^2 \ell \operatorname{vec}(\widehat{\Delta}) \geq \left(4 - \frac{288\check{s}}{p-4}\right)\|\widehat{\Delta}^-\|_F^2 + p\|\widehat{\Delta}^+\|_F^2 - \frac{32p^3}{v_1^2 p(p-4)}\|\widehat{\Delta}^+\|_F^2.$$

Thus, for v_1 large enough, it holds with probability tending to one that

$$\operatorname{vec}(\widehat{\Delta})^\top \nabla^2 \ell \operatorname{vec}(\widehat{\Delta}) \geq 2\|\widehat{\Delta}^-\|_F^2 + \frac{p}{2}\|\widehat{\Delta}^+\|_F^2, \tag{B.2}$$

which follows from assumption **A2**. Next, by Lemma B.2 and (B.2)

$$\begin{aligned}
0 &\geq \frac{1}{2}\operatorname{vec}(\widehat{\Delta})\nabla^2\ell \operatorname{vec}(\widehat{\Delta}) - 4\xi\|\widehat{\Delta}^-\|_1 - 4p^{3/2}\xi\|\widehat{\Delta}^+\|_F + \lambda\|\widehat{\Delta}^-_{\mathcal{S}^c}\|_1 - \lambda\|\widehat{\Delta}^-_{\mathcal{S}}\|_1 \\
&\geq \|\widehat{\Delta}^-\|_F^2 + \frac{p}{4}\|\widehat{\Delta}^+\|_F^2 + (\lambda - 4\xi)\|\widehat{\Delta}^-_{\mathcal{S}^c}\|_1 - (\lambda + 4\xi)\|\widehat{\Delta}^-_{\mathcal{S}}\|_1 - 4p^{3/2}\xi\|\widehat{\Delta}^+\|_F \\
&\geq \|\widehat{\Delta}^-\|_F^2 + \frac{p}{4}\|\widehat{\Delta}^+\|_F^2 - (\lambda + 4\xi)\|\widehat{\Delta}^-_{\mathcal{S}}\|_1 - 4p^{3/2}\xi\|\widehat{\Delta}^+\|_F \\
&\geq \|\widehat{\Delta}^-\|_F^2 + \frac{p}{4}\|\widehat{\Delta}^+\|_F^2 - (\lambda + 4\xi)\sqrt{s}\|\widehat{\Delta}^-\|_F - 4p^{3/2}\xi\|\widehat{\Delta}^+\|_F.
\end{aligned}$$

where $s = \sum_{j=1}^p s_j$. Recall that $\lambda = \sqrt{c_1 \log(p)/n}$. Thus, by Lemma B.3, for finite $v_2 > 0$ and $v_3 > 0$ sufficiently large, $\lambda + 4\xi \leq v_2 \log(p)/n$, $4\xi \leq v_3 \log(p)/n$ and

$$0 \geq \|\widehat{\Delta}^-\|_F^2 - v_2\sqrt{\frac{s\log(p)}{n}}\|\widehat{\Delta}^-\|_F + \frac{p}{4}\|\widehat{\Delta}^+\|_F^2 - v_3\sqrt{\frac{p^3\log(p)}{n}}\|\widehat{\Delta}^+\|_F, \tag{B.3}$$

with probability tending to one for $c_1 > 0$ sufficiently large.

To establish the error bound for $\|\widehat{\Delta}^+\|_F$, suppose for the sake of contradiction that $\|\widehat{\Delta}^+\|_F \geq 8v_3\sqrt{p\log(p)/n}$. Then the sum of the last two terms on the right-hand side of (B.3) is no smaller than $16v_3^2p^2\log(p)/n - 8v_3^2p^2\log(p)/n = 8v_3^2p^2\log(p)/n$. Additionally, by minimizing the quadratic in $\|\widehat{\Delta}^-\|_F$, one gets that the sum of the first two terms is no smaller than $-v_2^2 s\log(p)/(4n)$. Thus, since $\max_j s_j = o(p)$ and $s \leq p\max_j s_j$ the last right-hand side in (B.3) is positive for large enough p, so $\widehat{\Omega}$ is not a minimizer. This is the desired contradiction, so we conclude $\|\widehat{\Delta}^+\|_F < 8v_3\sqrt{p\log(p)/n}$ with probability tending to one.

To establish the main result, observe (B.3) implies

$$\begin{aligned}
\|\widehat{\Delta}^-\|_F^2 - v_2\sqrt{\frac{s\log(p)}{n}}\left\|\widehat{\Delta}^- + \sqrt{\frac{p}{4}}\widehat{\Delta}^+\right\|_F + \frac{p}{4}\|\widehat{\Delta}^+\|_F^2 \\
- v_3\sqrt{\frac{4p^2\log(p)}{n}}\left\|\widehat{\Delta}^- + \sqrt{\frac{p}{4}}\widehat{\Delta}^+\right\|_F \leq 0.
\end{aligned}$$

Rearranging terms,

$$\|\widehat{\Delta}^-\|_F^2 + \frac{p}{4}\|\widehat{\Delta}^+\|_F^2 \leq \left(v_2\sqrt{\frac{s\log(p)}{n}} + v_3\sqrt{\frac{4p^2\log(p)}{n}}\right)\left\|\widehat{\Delta}^- + \sqrt{\frac{p}{4}}\widehat{\Delta}^+\right\|_F,$$

so that by dividing both sides by $\|\widehat{\Delta}^- + \sqrt{\frac{p}{4}}\widehat{\Delta}^+\|_F$, we have

$$\left\|\widehat{\Delta}^- + \sqrt{\frac{p}{4}}\widehat{\Delta}^+\right\|_F \leq \left(v_2\sqrt{\frac{s\log(p)}{n}} + v_3\sqrt{\frac{4p^2\log(p)}{n}}\right). \tag{B.4}$$

Finally, because $\|\widehat{\Delta}^-\|_F + \sqrt{\frac{p}{4}}\|\widehat{\Delta}^+\|_F \leq \sqrt{2}\|\widehat{\Delta}^- + \sqrt{\frac{p}{4}}\widehat{\Delta}^+\|_F$, we can conclude

$$\frac{\|\widehat{\Delta}^-\|_F}{\sqrt{p}} + \|\widehat{\Delta}^+\|_F \leq \sqrt{\frac{4}{p}}\|\widehat{\Delta}^-\|_F + \|\widehat{\Delta}^+\|_F \leq \sqrt{\frac{8}{p}}\left(v_2\sqrt{\frac{s\log(p)}{n}} + v_3\sqrt{\frac{4p^2\log(p)}{n}}\right)$$

with probability tending to one.

□

B.3. Proof of Theorem 5.2

Proof of Theorem 5.2. In order to prove Theorem 5.2, we use a similar series of arguments as in the proof of Theorem 5.1. First, notice that Lemma B.2 implies

$$\begin{aligned}(\gamma - 4\boldsymbol{\xi})|||\widehat{\boldsymbol{\Delta}}^-_{\mathcal{S}^c}|||_{1,2} &\leq (\gamma + 4\boldsymbol{\xi})|||\widehat{\boldsymbol{\Delta}}^-_{\mathcal{S}}|||_{1,2} + 4\boldsymbol{\eta}|||\widehat{\boldsymbol{\Delta}}^+|||_F \\ &\leq (\gamma + 4\boldsymbol{\xi})|||\widehat{\boldsymbol{\Delta}}^-_{\mathcal{S}}|||_{1,2} + 4\boldsymbol{\xi}p^{3/2}|||\widehat{\boldsymbol{\Delta}}^+|||_F,\end{aligned}$$

where the last inequality follows from $\boldsymbol{\eta} \leq \boldsymbol{\xi}p^{3/2}$. By Lemma B.4, we can, for any $v_1 > 0$, pick $c_2 > 0$ sufficiently large so that with probability tending to one,

$$|||\widehat{\boldsymbol{\Delta}}^-_{\mathcal{S}^c}|||_{1,2} \leq 2|||\widehat{\boldsymbol{\Delta}}^-_{\mathcal{S}}|||_{1,2} + (p^{3/2}/v_1)|||\widehat{\boldsymbol{\Delta}}^+|||_F. \tag{B.5}$$

Next, by Lemma B.1 with $a_1 = 2$ and $a_2 = (p^{3/2}/v_1)|||\widehat{\boldsymbol{\Delta}}^+|||_F$, (B.5) implies

$$\begin{aligned}\sum_{h=1}^{H}\operatorname{vec}(\widehat{\Delta}_{(h)})^\top\nabla^2\ell\operatorname{vec}(\widehat{\Delta}_{(h)}) \geq \\ \left(4 - \frac{288\check{s}}{p-4}\right)|||\widehat{\boldsymbol{\Delta}}^-|||_F^2 + p|||\widehat{\boldsymbol{\Delta}}^+|||_F^2 - \frac{32p^3}{v_1^2p(p-4)}|||\widehat{\boldsymbol{\Delta}}^+|||_F^2.\end{aligned}$$

Thus, for v_1 large enough, it holds with probability tending to one that

$$\sum_{h=1}^{H}\operatorname{vec}(\widehat{\Delta}_{(h)})^\top\nabla^2\ell\operatorname{vec}(\widehat{\Delta}_{(h)}) \geq 2|||\widehat{\boldsymbol{\Delta}}^-|||_F^2 + \frac{p}{2}|||\widehat{\boldsymbol{\Delta}}^+|||_F^2,$$

which follows from assumption **A5**. Next, by Lemma B.2,

$$\begin{aligned}0 &\geq \frac{1}{2}\sum_{h=1}^{H}\operatorname{vec}(\widehat{\Delta}_{(h)})^\top\nabla^2\ell\operatorname{vec}(\widehat{\Delta}_{(h)}) - 4\boldsymbol{\xi}|||\widehat{\boldsymbol{\Delta}}^-|||_{1,2} - 4p^{3/2}\boldsymbol{\xi}|||\widehat{\boldsymbol{\Delta}}^+|||_F \\ &\qquad + \gamma|||\widehat{\boldsymbol{\Delta}}^-_{\mathcal{S}^c}|||_{1,2} - \gamma|||\widehat{\boldsymbol{\Delta}}^-_{\mathcal{S}}|||_{1,2} \\ &\geq |||\widehat{\boldsymbol{\Delta}}^-|||_F^2 + \frac{p}{4}|||\widehat{\boldsymbol{\Delta}}^+|||_F^2 + (\gamma - 4\boldsymbol{\xi})|||\widehat{\boldsymbol{\Delta}}^-_{\mathcal{S}^c}|||_{1,2} \\ &\qquad - (\gamma + 4\boldsymbol{\xi})|||\widehat{\boldsymbol{\Delta}}^-_{\mathcal{S}}|||_{1,2} - 4p^{3/2}\boldsymbol{\xi}|||\widehat{\boldsymbol{\Delta}}^+|||_F \\ &\geq |||\widehat{\boldsymbol{\Delta}}^-|||_F^2 + \frac{p}{4}|||\widehat{\boldsymbol{\Delta}}^+|||_F^2 - (\gamma + 4\boldsymbol{\xi})|||\widehat{\boldsymbol{\Delta}}^-_{\mathcal{S}}|||_{1,2} - 4p^{3/2}\boldsymbol{\xi}|||\widehat{\boldsymbol{\Delta}}^+|||_F\end{aligned}$$

$$\geq |||\widehat{\boldsymbol{\Delta}}^{-}|||_F^2 + \frac{p}{4}|||\widehat{\boldsymbol{\Delta}}^{+}|||_F^2 - (\gamma + 4\boldsymbol{\xi})\sqrt{\tilde{s}}|||\widehat{\boldsymbol{\Delta}}^{-}|||_F - 4p^{3/2}\boldsymbol{\xi}|||\widehat{\boldsymbol{\Delta}}^{+}|||_F$$

where $\tilde{s} = \sum_{j=1}^{p} \tilde{s}_j$. Then, based on our choice of γ with $c_2 > 0$ sufficiently large, there exists finite $v_2 > 0$ and $v_3 > 0$ such that $(\gamma + 4\boldsymbol{\xi}) \leq v_2[(HL^4/n_{\min})^{1/2} + \{\log(p)/n_{\min}\}^{1/2}]$ and $4\boldsymbol{\xi} \leq v_3[(HL^4/n_{\min})^{1/2} + \{\log(p)/n_{\min}\}^{1/2}]$ with probability tending to one by Lemma B.4. Thus, with probability tending to one

$$0 \geq |||\widehat{\boldsymbol{\Delta}}^{-}|||_F^2 - v_2\sqrt{\tilde{s}}\left(\sqrt{\frac{HL^4}{n_{\min}}} + \sqrt{\frac{\log(p)}{n_{\min}}}\right)|||\widehat{\boldsymbol{\Delta}}^{-}|||_F \tag{B.6}$$
$$+ \frac{p}{4}|||\widehat{\boldsymbol{\Delta}}^{+}|||_F^2 - v_3 p^{3/2}\left(\sqrt{\frac{HL^4}{n_{\min}}} + \sqrt{\frac{\log(p)}{n_{\min}}}\right)|||\widehat{\boldsymbol{\Delta}}^{+}|||_F.$$

By the same arguments as in the proof of Theorem 5.1, one can show that (B.6) implies

$$|||\widehat{\boldsymbol{\Delta}}^{+}|||_F < 8v_3\left(\sqrt{\frac{pHL^4}{n_{\min}}} + \sqrt{\frac{p\log(p)}{n_{\min}}}\right)$$

with probability tending to one, and also that there exists $v_4 \in (0, \infty)$ such that

$$|||\widehat{\boldsymbol{\Delta}}^{-} + \sqrt{p/4}\widehat{\boldsymbol{\Delta}}^{+}|||_F \leq v_4(\sqrt{\tilde{s}} + p)\left(\sqrt{\frac{HL^4}{n_{\min}}} + \sqrt{\frac{\log(p)}{n_{\min}}}\right),$$

with probability tending to one. The conclusion as stated follows from the arguments after equation (B.4), along with the fact that under **A6**, $n_{\min} \geq \pi N$ for N sufficiently large. □

Appendix C: Proofs of Lemmas

Proof of Lemma B.1. Inspecting the definition of $\ell_{(h)}$ shows

$$\sum_{h=1}^{H} \operatorname{vec}(\Delta_{(h)})^\top \nabla^2 \ell \operatorname{vec}(\Delta_{(h)}) = \sum_{i=1}^{p}\sum_{k\neq j}\sum_{h=1}^{H}\{\Delta_{(h)jj} + \Delta_{(h)kk} - 2\Delta_{(h)jk}\}^2.$$

Expanding the squares gives

$$\sum_{h=1}^{H} \operatorname{vec}(\Delta_{(h)})^\top \nabla^2 \ell \operatorname{vec}(\Delta_{(h)}) = \sum_{h=1}^{H}\sum_{j=1}^{p}\sum_{k\neq j}(\Delta_{(h)jj} + \Delta_{(h)kk})^2$$
$$- 4\sum_{h=1}^{H}\sum_{j=1}^{p}\sum_{k\neq j}(\Delta_{(h)jj} + \Delta_{(h)kk})\Delta_{(h)jk} + 4\sum_{h=1}^{H}\sum_{j=1}^{p}\sum_{k\neq j}\Delta_{(h)jk}^2 = \mathrm{I} + \mathrm{II} + \mathrm{III}.$$

Note $\mathrm{III} = 4|||\boldsymbol{\Delta}^-|||_F^2$. Next, using that

$$\sum_{j=1}^{p}\sum_{k\neq j}\Delta_{(h)kk}^2 = \sum_{j=1}^{p}\sum_{k=1}^{p}\Delta_{(h)kk}^2 - \sum_{j=1}^{p}\Delta_{(h)jj}^2 = (p-1)\|\Delta_{(h)}^+\|_F^2$$

and

$$\begin{aligned}\sum_{j=1}^{p}\sum_{k\neq j}\Delta_{(h)jj}\Delta_{(h)kk} &= \sum_{j=1}^{p}\left\{\Delta_{(h)jj}\left(\sum_{k=1}^{p}\Delta_{(h)kk} - \Delta_{(h)jj}\right)\right\}\\ &= \left(\sum_{j=1}^{p}\Delta_{(h)jj}\right)^2 - \|\Delta_{(h)}^+\|_F^2,\end{aligned}$$

we get

$$\begin{aligned}\mathrm{I} &= \sum_{h=1}^{H}\sum_{j=1}^{p}\sum_{k\neq j}\Delta_{(h)jj}^2 + \sum_{h=1}^{H}\sum_{j=1}^{p}\sum_{k\neq j}\Delta_{(h)kk}^2 + 2\sum_{h=1}^{H}\sum_{j=1}^{p}\sum_{k\neq j}\Delta_{(h)jj}\Delta_{(h)kk}\\ &= 2(p-2)|||\boldsymbol{\Delta}^+|||_F^2 + 2\sum_{h=1}^{H}\left(\sum_{j=1}^{p}\Delta_{(h)jj}\right)^2\\ &\geq 2(p-2)|||\boldsymbol{\Delta}^+|||_F^2.\end{aligned}$$

Using this inequality for I, along with the Cauchy–Schwarz inequality, we have

$$\begin{aligned}\mathrm{I}+\mathrm{II}+\mathrm{III} &\geq 4|||\boldsymbol{\Delta}^-|||_F^2 + 2(p-2)|||\boldsymbol{\Delta}^+|||_F^2 - 4\sum_{h=1}^{H}\sum_{j=1}^{p}\sum_{k\neq j}\Delta_{(h)jk}\Delta_{(h)jj}\\ &\quad - 4\sum_{h=1}^{H}\sum_{j=1}^{p}\sum_{k\neq j}\Delta_{(h)jk}\Delta_{(h)kk}\\ &= 4|||\boldsymbol{\Delta}^-|||_F^2 + 2(p-2)|||\boldsymbol{\Delta}^+|||_F^2 - 8\sum_{j=1}^{p}\sum_{k\neq j}\sum_{h=1}^{H}\Delta_{(h)jk}\Delta_{(h)jj} \quad \text{(C.1)}\\ &\geq 4|||\boldsymbol{\Delta}^-|||_F^2 + 2(p-2)|||\boldsymbol{\Delta}^+|||_F^2 - 8\sum_{j=1}^{p}\sum_{k\neq j}\|\boldsymbol{\Delta}_{\cdot jk}\|_2\|\boldsymbol{\Delta}_{\cdot jj}\|_2\\ &= 4|||\boldsymbol{\Delta}^-|||_F^2 + p|||\boldsymbol{\Delta}^+|||_F^2 \quad \text{(C.2)}\\ &\quad + \sum_{j=1}^{p}\left\{(p-4)\|\boldsymbol{\Delta}_{\cdot jj}\|_2^2 - 8\|\boldsymbol{\Delta}_{\cdot jj}\|_2\sum_{k\neq j}\|\boldsymbol{\Delta}_{\cdot jk}\|_2\right\}.\end{aligned}$$

Here, the equality (C.1) follows from

$$\sum_{j=1}^{p}\sum_{k\neq j}\Delta_{(h)jk}\Delta_{(h)jj} = \sum_{j=1}^{p}\sum_{k=1}^{p}1(j\neq k)\Delta_{(h)jj}\Delta_{(h)jk} = \sum_{k=1}^{p}\sum_{j\neq k}\Delta_{(h)jj}\Delta_{(h)jk}.$$

To simplify notation, let $D_j = \sum_{k \neq j} \|\boldsymbol{\Delta}_{\cdot jk}\|_2$ be the sum of the Euclidean norms of the off-diagonal fibers in the jth row. Completing the square in the jth summand of the last lower bound of $\mathrm{I} + \mathrm{II} + \mathrm{III}$ gives

$$\left(\sqrt{(p-4)}\|\boldsymbol{\Delta}_{\cdot jj}\| - \sqrt{\frac{16}{p-4}}D_j\right)^2 - \frac{16}{p-4}D_j^2,$$

and hence, we have shown

$$\sum_{h=1}^{H} \mathrm{vec}(\Delta_{(h)})^\top \nabla^2 \ell \, \mathrm{vec}(\Delta_{(h)}) = \mathrm{I} + \mathrm{II} + \mathrm{III}$$
$$\geq 4|||\boldsymbol{\Delta}^-|||_F^2 + p|||\boldsymbol{\Delta}^+|||_F^2 - \frac{16}{p-4}\sum_{j=1}^{p} D_j^2.$$

To bound $\sum_{j=1}^p D_j^2$, fix an arbitrary $a_3 > 0$ and consider the optimization problem

$$\max_{\boldsymbol{\Delta} \in \mathbb{R}^{H \times p \times p}} \sum_{j=1}^{p} D_j^2 \text{ s.t. } |||\boldsymbol{\Delta}^-_{\mathcal{S}^c}|||_{1,2} \leq a_1|||\boldsymbol{\Delta}^-_{\mathcal{S}}|||_{1,2} + a_2, \ \Delta_{(h)} = \Delta_{(h)}^\top, \ |||\boldsymbol{\Delta}|||_F^2 \leq a_3.$$

The maximum becomes no smaller if we drop the symmetry constraint and add a constraint that all elements be positive. Moreover, we may assume $H = 1$ since only the Euclidean norms of fibers affect the objective function and constraints. Thus, $\boldsymbol{\Delta} = \Delta$ and $D_j = \|\Delta^j\|_1$, where Δ^j is Δ with all elements but the off-diagonal ones in the jth row set to zero.

We get the concave optimization problem

$$\max_{\Delta \in \mathbb{R}^{p \times p}} \sum_{j=1}^{p} \|\Delta^j\|_1^2 \text{ s.t. } \sum_{j=1}^{p}\sum_{k=1}^{p}(\Delta^-_{\mathcal{S}^c})_{jk} \leq a_1 \sum_{j=1}^{p}\sum_{k=1}^{p}(\Delta^-_{\mathcal{S}})_{jk} + a_2, \qquad \text{(C.3)}$$
$$\|\Delta\|_F^2 \leq a_3, \quad \Delta_{jk} \geq 0.$$

Note that, except for the definition of $\mathcal{S}$, the order of elements within rows does not matter in (C.3). Thus, we may, without affecting the maximum, assume $\mathcal{S} = \{(j,k) \in [p] \times [p] : k \leq s_j\}$ if we also redefine Δ^j to be Δ with all elements but the first $p-1$ in the jth row set to zero, Δ^- to be Δ with the last column set to zero, and $\Delta^+ = \Delta - \Delta^-$. Effectively, this interchanges the diagonal element in each row with the last element in that row and puts the support indicated by $\mathcal{S}$ on the first s_j elements in the jth row. Now, the feasible set becomes no smaller if $s_j = \check{s}$ for each j, so we can assume this without decreasing the maximum.

With these definitions, pick a feasible Δ and note any Δ' obtained by permuting the rows of Δ is feasible and attains the same value. Thus, by concavity, the convex combination which puts equal weight $1/p!$ on all $p!$ row-permutations of Δ, say $\tilde{\Delta}$, is feasible and attains at least the same value as Δ. By construction, every row of $\tilde{\Delta}$ is the same. Thus, we have shown that for any feasible Δ there is a feasible $\tilde{\Delta}$ that attains at least the same value and has all rows equal.

Pick an arbitrary feasible Δ and a corresponding $\tilde{\Delta}$. Since all rows of $\tilde{\Delta}$ are the same, the constraint $\|\tilde{\Delta}^-_{\mathcal{S}^c}\|_1 \le a_1\|\tilde{\Delta}^-_{\mathcal{S}}\|_1 + a_2$ implies $\|\tilde{\Delta}^j_{\mathcal{S}^c}\|_1 \le a_1\|\tilde{\Delta}^j_{\mathcal{S}}\|_1 + a_2/p$ for every $j \in [p]$. Thus,

$$\sum_{j=1}^p \|\tilde{\Delta}^j\|_1^2 \le \sum_{j=1}^p \{(1+a_1)\|\tilde{\Delta}^j_{\mathcal{S}}\|_1 + a_2/p\}^2 \le \sum_{j=1}^p \{(1+a_1)\sqrt{\check{s}}\|\tilde{\Delta}^j_{\mathcal{S}}\|_F + a_2/p\}^2.$$

Moreover,

$$\{(1+a_1)\sqrt{\check{s}}\|\tilde{\Delta}^j_{\mathcal{S}}\|_F + a_2/p\}^2 \le 2(1+a_1)^2\check{s}\|\tilde{\Delta}^j\|_F^2 + 2a_2^2/p^2$$

and hence

$$\begin{aligned}\sum_{j=1}^p \|\Delta^j\|_1^2 \le \sum_{j=1}^p \|\tilde{\Delta}^j\|_1^2 &\le \sum_{j=1}^p \{2(1+a_1)^2\check{s}\|\tilde{\Delta}^j\|_F^2 + 2a_2^2/p^2\} \\ &= 2(1+a_1)^2\check{s}\|\Delta^-\|_F^2 + 2a_2^2/p.\end{aligned}$$

Thus,

$$\begin{aligned}\mathrm{I}+\mathrm{II}+\mathrm{III} &\ge 4\|\Delta^-\|_F^2 + p\|\Delta^+\|_F^2 - \frac{16}{p-4}\{2(1+a_1)^2\check{s}\|\Delta^-\|_F^2 + 2a_2^2/p\} \\ &= \left(4 - \frac{32\check{s}(1+a_1)^2}{p-4}\right)\|\Delta^-\|_F^2 + p\|\Delta^+\|_F^2 - \frac{32a_2^2}{p(p-4)},\end{aligned}$$

which completes the proof. □

We will use the following to prove Lemma B.3.

Lemma C.1. *Supposing $H = 1$, if the $\log(W_{ij})$ are independent over $i \in [n]$ and have sub-Gaussian norms bounded by $K < \infty$, then for $j \ne k$ and any $\epsilon > 0$,*

$$P\left(|\widehat{\Theta}_{jk} - \tau^*_{jk}| \ge \epsilon\right) \le 6\exp\{-\nu n \min(\epsilon/K^2, \epsilon^2/K^4)\}$$

where $\nu > 0$ is a universal constant.

Proof of Lemma C.1. Suppose first $\widehat{\Theta}_{jk} = n^{-1}\sum_{i=1}^n \log(X_{ij}/X_{ik})^2$ and note this makes sense since $\mathbb{E}\{\log(X_{ij}/X_{ik})\} = \mathbb{E}\{\log(W_{ij}) - \log(W_{ik})\} = 0$. Now

$$n^{-1}\sum_{i=1}^n \log(X_{ij}/X_{ik})^2 = n^{-1}\sum_{i=1}^n \{\log(W_{ij})^2 + \log(W_{ik})^2 - 2\log(W_{ij})\log(W_{ik})\}.$$

We thus have, since $\mathbb{E}\{\log(W_{ij})\} = 0$,

$$\begin{aligned}P\left(|\widehat{\Theta}_{jk} - \omega^*_j - \omega^*_k + 2\Omega^*_{jk}| \ge \epsilon\right) &\le P\left(|\sum_{i=1}^n \{\log(W_{ij})^2 - \omega^*_j\}| \ge n\epsilon/3\right) \\ &\quad + P\left(|\sum_{i=1}^n \{\log(W_{ik})^2 - \omega^*_k\}| \ge n\epsilon/3\right)\end{aligned}$$

$$+ P\left(|\sum_{i=1}^{n}\{\log(W_{ij})\log(W_{ik}) - \Omega^*_{jk}\}| \geq n\epsilon/6\right).$$

Each of these terms enjoys sub-exponential concentration, meaning they are upper bounded by $2\exp\{-c\min(\epsilon/K^2, \epsilon^2/K^4)n\}$, where K^2 is the sub-Exponential norm bound of the $\log(W_{ij})^2$ and $\nu > 0$ a universal constant [36, Lemma 2.7.5 and Corollary 2.8.4].

Now, with $\widehat{\Theta}_{jk} = n^{-1}\sum_{i=1}^{n}\left\{\log(X_{ij}/X_{ik}) - n^{-1}\sum_{l=1}^{n}\log(X_{lj}/X_{lk})\right\}^2$, which is equal to $n^{-1}\sum_{i=1}\log(X_{ij}/X_{ik})^2 - \left\{n^{-1}\sum_{i=1}\log(X_{ij}/X_{ik})\right\}^2$, routine arguments for the concentration of the sample mean of sub-Gaussian random variables show $\left\{n^{-1}\sum_{i=1}\log(X_{ij}/X_{ik})\right\}^2$ is of smaller order than

$$n^{-1}\sum_{i=1}\log(X_{ij}/X_{ik})^2 - \tau^*_{jk}$$

and so essentially the same proof applies. We omit the details for brevity. □

Proof of Lemma B.3. Let $\epsilon = \sqrt{c_1\log(p)/n} \to 0$ for fixed constant $c_1 > 0$. By Lemma C.1 and the union bound, for n sufficiently large

$$P\left(\max_{j,k}|\widehat{\Theta}_{jk} - \tau^*_{jk}| \geq \epsilon\right) \leq 6p^2\exp(-\nu n\epsilon^2/K^4) = 6\exp\{(2 - \nu c_1/K^4)\log(p)\},$$

which completes the proof. □

Proof of Lemma B.2. By definition, $\ell(\widehat{\boldsymbol{\Omega}}) + \gamma|||\widehat{\boldsymbol{\Omega}}^-|||_{1,2} \leq \ell(\boldsymbol{\Omega}^*) + \gamma|||[\boldsymbol{\Omega}^*]^-|||_{1,2}$, and hence

$$\ell(\widehat{\boldsymbol{\Omega}}) - \ell(\boldsymbol{\Omega}^*) \leq \gamma\left(|||[\boldsymbol{\Omega}^*]^-|||_{1,2} - |||\widehat{\boldsymbol{\Omega}}^-|||_{1,2}\right) \leq \gamma\left(|||\widehat{\boldsymbol{\Delta}}^-_{\mathcal{S}}|||_{1,2} - |||\widehat{\boldsymbol{\Delta}}^-_{\mathcal{S}^c}|||_{1,2}\right),$$

where the last inequality follows from $\boldsymbol{\Omega}^*_{\mathcal{S}} = \boldsymbol{\Omega}^*$, $\boldsymbol{\Omega}^*_{\mathcal{S}^c} = 0$, and the triangle inequality. On the other hand, for any $\boldsymbol{\Omega}$ and $\boldsymbol{\Delta} = \boldsymbol{\Omega} - \boldsymbol{\Omega}^*$,

$$\begin{aligned}
\ell(\boldsymbol{\Omega}) - \ell(\boldsymbol{\Omega}^*) &= \sum_{h=1}^{H}\Bigg\{\frac{1}{2}\operatorname{vec}(\Delta_{(h)})^\top\nabla^2\ell\operatorname{vec}(\Delta_{(h)}) \\
&\qquad + 4\sum_{j=1}^{p}\sum_{k\neq j}(\widehat{\Theta}_{(h)jk} - \tau^*_{(h)jk})(\boldsymbol{\Delta}_{(h)jk} - \boldsymbol{\Delta}_{(h)jj})\Bigg\} \\
&= \frac{1}{2}\sum_{h=1}^{H}\operatorname{vec}(\Delta_{(h)})^\top\nabla^2\ell\operatorname{vec}(\Delta_{(h)}) + 4\sum_{h=1}^{H}\sum_{j=1}^{p}\sum_{k\neq j}(\widehat{\Theta}_{(h)jk} - \tau^*_{(h)jk})\boldsymbol{\Delta}_{(h)jk} \\
&\quad - 4\sum_{h=1}^{H}\sum_{j=1}^{p}\boldsymbol{\Delta}_{(h)jj}\sum_{k\neq j}(\widehat{\Theta}_{(h)jk} - \tau^*_{(h)jk}) \\
&= \frac{1}{2}\sum_{h=1}^{H}\operatorname{vec}(\Delta_{(h)})^\top\nabla^2\ell\operatorname{vec}(\Delta_{(h)}) + 4\sum_{j=1}^{p}\sum_{k\neq j}(\widehat{\boldsymbol{\Theta}}_{\cdot jk} - \tau^*_{\cdot jk})^\top\boldsymbol{\Delta}_{\cdot jk}
\end{aligned}$$

$$- 4\sum_{j=1}^{p} \boldsymbol{\Delta}_{\cdot jj}^{\top} \left\{ \sum_{k \neq j} (\widehat{\boldsymbol{\Theta}}_{\cdot jk} - \tau_{\cdot jk}^{*}) \right\}$$

$$\geq \frac{1}{2} \sum_{h=1}^{H} \operatorname{vec}(\Delta_{(h)})^{\top} \nabla^2 \ell \operatorname{vec}(\Delta_{(h)}) - 4\boldsymbol{\xi} |||\boldsymbol{\Delta}^{-}|||_{1,2} - 4\boldsymbol{\eta} \|\boldsymbol{\Delta}^{+}\|_F,$$

which completes the proof. □

Proof of Lemma B.4. First, recall that we assume $\log(W_{(h)ij})$ is supported on $[-L, L]$. Thus, defining $v_{(h)ilm} = \log(W_{(h)il}/W_{(h)im})$, it can be easily verified that $v_{(h)ilm}^2$ is supported on $[0, 4L^2]$.

To establish the desired concentration inequality, we will use McDiarmid's inequality [36, Theorem 2.9.1] to get a high-probability bound for $\|\widehat{\boldsymbol{\Theta}}_{\cdot lm} - \boldsymbol{\tau}_{\cdot lm}^{*}\|_2$ for an arbitrary pair (l, m) with $l \neq m$, and then apply the union bound to control $\max_{l \neq m} \|\widehat{\boldsymbol{\Theta}}_{\cdot lm} - \boldsymbol{\tau}_{\cdot lm}^{*}\|_2$.

Let $\widehat{\boldsymbol{\Theta}}$ be the tensor of sample variation matrices based on $\{W_{(h)1}, \ldots, W_{(h)n_{(h)}}\}_{h=1}^{H}$ and let $\widetilde{\boldsymbol{\Theta}}$ be the tensor of sample variation matrices based on $\{\widetilde{W}_{(h)1}, \ldots, \widetilde{W}_{(h)n_{(h)}}\}_{h=1}^{H}$ where $\widetilde{W}_{(h)i} = W_{(h)i}$ for all but a single pair $(h^\star, i^\star)$, i.e., $\widetilde{W}_{(h^\star)i^\star} \neq W_{(h^\star)i^\star}$. To apply McDiarmid's inequality, we need to find a $c_{h^\star, i^\star}$ such that

$$|\|\widehat{\boldsymbol{\Theta}}_{\cdot lm} - \boldsymbol{\tau}_{\cdot lm}^{*}\|_2 - \|\widetilde{\boldsymbol{\Theta}}_{\cdot lm} - \boldsymbol{\tau}_{\cdot lm}^{*}\|_2| \leq \|\widehat{\boldsymbol{\Theta}}_{\cdot lm} - \widetilde{\boldsymbol{\Theta}}_{\cdot lm}\|_2 \leq c_{h^\star, i^\star}$$

for each pair $(h^\star, i^\star)$. Notice that bounding $\|\widehat{\boldsymbol{\Theta}}_{\cdot lm} - \widetilde{\boldsymbol{\Theta}}_{\cdot lm}\|_2^2$ is a matter of bounding

$$\|\widehat{\boldsymbol{\Theta}}_{\cdot lm} - \widetilde{\boldsymbol{\Theta}}_{\cdot lm}\|_2^2 = \sum_{h=1}^{H} \left(\frac{1}{n_{(h)}} \sum_{i=1}^{n_{(h)}} \left[\left\{ \log\left(\frac{W_{(h)il}}{W_{(h)im}}\right) - n_{(h)}^{-1} \sum_{j=1}^{n_{(h)}} \log\left(\frac{W_{(h)jl}}{W_{(h)jm}}\right) \right\}^2 \right. \right.$$

$$\left. \left. - \left\{ \log\left(\frac{\widetilde{W}_{(h)il}}{\widetilde{W}_{(h)im}}\right) - n_{(h)}^{-1} \sum_{j=1}^{n_{(h)}} \log\left(\frac{\widetilde{W}_{(h)jl}}{\widetilde{W}_{(h)jm}}\right) \right\}^2 \right] \right)^2.$$

Now, using that by definition $v_{(h)ilm} = \log\left(\frac{W_{(h)il}}{W_{(h)im}}\right)$, and since $v_{(h)ilm} = \tilde{v}_{(h)ilm}$ for all $i \in [n_{(h)}]$ when $h \neq h^\star$, we have that

$$\|\widehat{\boldsymbol{\Theta}}_{\cdot lm} - \widetilde{\boldsymbol{\Theta}}_{\cdot lm}\|_2^2$$

$$= \left(\frac{1}{n_{(h^\star)}} \sum_{i=1}^{n_{(h^\star)}} \left[\left\{ v_{(h^\star)ilm} - n_{(h^\star)}^{-1} \sum_{j=1}^{n_{(h^\star)}} v_{(h^\star)jlm} \right\}^2 \right. \right.$$

$$\left. \left. - \left\{ \tilde{v}_{(h^\star)ilm} - n_{(h^\star)}^{-1} \sum_{j=1}^{n_{(h^\star)}} v_{(h^\star)jlm} - n_{(h^\star)}^{-1} \tilde{v}_{(h^\star)i^\star lm} + n_{(h^\star)}^{-1} v_{(h^\star)i^\star lm} \right\}^2 \right] \right)^2$$

$$= \left[\frac{1}{n_{(h^\star)}} \sum_{i=1}^{n_{(h^\star)}} \left\{ a_i^2 - (\tilde{a}_i + b)^2 \right\} \right]^2 = \left\{ \frac{1}{n_{(h^\star)}} \sum_{i=1}^{n_{(h^\star)}} \left(a_i^2 - \tilde{a}_i^2 - 2\tilde{a}_i b - b^2 \right) \right\}^2$$

where $a_i = v_{(h^\star)ilm} - n_{(h^\star)}^{-1} \sum_{j=1}^{n_{(h^\star)}} v_{(h^\star)jlm}$, $\tilde{a}_i = \tilde{v}_{(h^\star)ilm} - n_{(h^\star)}^{-1} \sum_{j=1}^{n_{(h^\star)}} v_{(h^\star)jlm}$, and $b = n_{(h^\star)}^{-1} v_{(h^\star)i^\star lm} - n_{(h^\star)}^{-1} \tilde{v}_{(h^\star)i^\star lm}$. Because $a_i = \tilde{a}_i$ for all $i \neq i^\star$, there exists a constant $d_1 \in (0, \infty)$ such that

$$\begin{aligned}
&= \left\{ \frac{(a_{i^\star}^2 - \tilde{a}_{i^\star}^2)}{n_{(h^\star)}} - \frac{1}{n_{(h^\star)}} \sum_{i=1}^{n_{(h^\star)}} \left(2\tilde{a}_i b + b^2 \right) \right\}^2 \\
&= \left\{ \frac{(a_{i^\star}^2 - \tilde{a}_{i^\star}^2)}{n_{(h^\star)}} + \frac{2b}{n_{(h^\star)}} \underbrace{(v_{(h^\star)i^\star lm} - \tilde{v}_{(h^\star)i^\star lm})}_{= -\sum_{i=1}^{n_{(h^\star)}} \tilde{a}_i = b\, n_{(h^\star)}} - b^2 \right\}^2 \\
&= \left\{ \frac{(a_{i^\star}^2 - \tilde{a}_{i^\star}^2)}{n_{(h^\star)}} + b^2 \right\}^2 \leq \frac{4a_{i^\star}^4 + 4\tilde{a}_{i^\star}^4}{n_{(h^\star)}^2} + 2b^4 \leq \frac{d_1^2 L^4}{n_{(h^\star)}^2} \leq \frac{d_1^2 L^4}{n_{\min}^2}.
\end{aligned}$$

The second to last inequality above follows from the fact that

$$\begin{aligned}
a_i^2 &= \left(v_{(h^\star)ilm} - n_{(h^\star)}^{-1} \sum_{j=1}^{n_{(h^\star)}} v_{(h^\star)jlm)} \right)^2 \\
&\leq 2v_{(h^\star)ilm}^2 + 2 \left(\frac{1}{n_{(h^\star)}} \sum_{j=1}^{n_{(h^\star)}} v_{(h^\star)jlm} \right)^2 \leq 16L^2,
\end{aligned}$$

and similarly for $\tilde{a}_i^2$, and also that

$$b^2 = \frac{1}{n_{(h^\star)}^2} (v_{(h^\star)i^\star lm} - \tilde{v}_{(h^\star)i^\star lm})^2 \leq \frac{2}{n_{(h^\star)}^2} (v_{(h^\star)i^\star lm}^2 + \tilde{v}_{(h^\star)i^\star lm}^2) \leq \frac{16L^2}{n_{(h^\star)}^2}.$$

Hence, we have shown that there exists a constant $d_1 \in (0, \infty)$ such that

$$\|\widehat{\boldsymbol{\Theta}}_{\cdot lm} - \widetilde{\boldsymbol{\Theta}}_{\cdot lm}\|_2 \leq \frac{d_1 L^2}{n_{\min}}$$

for all pairs $(h^\star, i^\star)$. Thus applying McDiarmid's inequality, for any $\epsilon > 0$

$$\begin{aligned}
P\left(\|\widehat{\boldsymbol{\Theta}}_{\cdot lm} - \boldsymbol{\tau}_{\cdot lm}^*\|_2 \geq \mathbb{E}\|\widehat{\boldsymbol{\Theta}}_{\cdot lm} - \boldsymbol{\tau}_{\cdot lm}^*\|_2 + \epsilon \right) &\leq \exp\left(- \frac{2\epsilon^2}{d_1^2 \sum_{h=1}^{H} \sum_{i=1}^{n_{(h)}} \frac{L^4}{n_{\min}^2}} \right) \\
&= \exp\left(- \frac{2n_{\min}^2 \epsilon^2}{d_1^2 L^4 N} \right).
\end{aligned}$$

All that remains is to bound $\mathbb{E}\|\widehat{\boldsymbol{\Theta}}_{\cdot lm} - \boldsymbol{\tau}^*_{\cdot lm}\|_2$. First applying Jensen's inequality,

$$\begin{aligned}
&\mathbb{E}\|\widehat{\boldsymbol{\Theta}}_{\cdot lm} - \boldsymbol{\tau}^*_{\cdot lm}\|_2 \\
&\quad \le \sqrt{\sum_{h=1}^{H} \mathbb{E}\left[n_{(h)}^{-1} \sum_{i=1}^{n_{(h)}} \left\{ \left(v_{(h)ilm} - n_{(h)}^{-1} \sum_{j=1}^{n_{(h)}} v_{(h)jlm} \right)^2 - \boldsymbol{\tau}^*_{\cdot lm} \right\} \right]}^2
\end{aligned}$$

and letting $\hat{v}_{(h)ilm} = v_{(h)ilm} - \mathbb{E}(v_{(h)ilm})$,

$$\begin{aligned}
&= \sqrt{\sum_{h=1}^{H} \mathbb{E}\left[n_{(h)}^{-1} \sum_{i=1}^{n_{(h)}} \left\{ \left(\hat{v}_{(h)ilm} - n_{(h)}^{-1} \sum_{j=1}^{n_{(h)}} \hat{v}_{(h)jlm} \right)^2 - \boldsymbol{\tau}^*_{\cdot lm} \right\} \right]}^2 \\
&= \sqrt{\sum_{h=1}^{H} \mathbb{E}\left[n_{(h)}^{-1} \sum_{i=1}^{n_{(h)}} (\hat{v}^2_{(h)ilm} - \boldsymbol{\tau}^*_{\cdot lm}) - \left(n_{(h)}^{-1} \sum_{j=1}^{n_{(h)}} \hat{v}_{(h)jlm} \right)^2 \right]}^2 \\
&\le \sqrt{2 \sum_{h=1}^{H} \underbrace{\mathbb{E}\left\{ n_{(h)}^{-1} \sum_{i=1}^{n_{(h)}} (\hat{v}^2_{(h)ilm} - \boldsymbol{\tau}^*_{\cdot lm}) \right\}^2}_{=:\mathrm{IV}} + 2 \sum_{h=1}^{H} \underbrace{\mathbb{E}\left(n_{(h)}^{-1} \sum_{i=1}^{n_{(h)}} \hat{v}_{(h)ilm} \right)^4}_{=:\mathrm{V}}}.
\end{aligned}$$

Next, we bound IV and V separately. Notice that because $v_{(h)ilm} \in [-2L, 2L]$, with $\hat{v}_{(h)ilm} = v_{(h)ilm} - \mathbb{E}(v_{(h)ilm})$ it follows that $\hat{v}_{(h)ilm} \in [-4L, 4L]$. Therefore

$$\begin{aligned}
\mathrm{IV} = \mathbb{E}\left\{ n_{(h)}^{-1} \sum_{i=1}^{n_{(h)}} (\hat{v}^2_{(h)ilm} - \boldsymbol{\tau}^*_{\cdot lm}) \right\}^2 &= \mathrm{Var}\left\{ n_{(h)}^{-1} \sum_{i=1}^{n_{(h)}} (\hat{v}^2_{(h)ilm} - \boldsymbol{\tau}^*_{\cdot lm}) \right\} \\
&= \sum_{i=1}^{n_{(h)}} \frac{\mathrm{Var}\left(\hat{v}^2_{(h)ilm} \right)}{n_{(h)}^2} \le \frac{16^2 L^4}{4 n_{(h)}}
\end{aligned}$$

because for bounded random variable $Y \in [a, b]$, $\mathrm{Var}(Y) \le \frac{1}{4}(b-a)^2$. For V, notice that

$$\begin{aligned}
\mathrm{V} = \mathbb{E}\left(n_{(h)}^{-1} \sum_{i=1}^{n_{(h)}} \hat{v}_{(h)ilm} \right)^4 &= 16^2 L^4 \mathbb{E}\left(n_{(h)}^{-1} \sum_{i=1}^{n_{(h)}} \underbrace{\left(\hat{v}_{(h)ilm} / 4L \right)}_{=:u_{(h)ilm}} \right)^4 \\
&\le 16^2 L^4 \mathbb{E}\left(n_{(h)}^{-1} \sum_{i=1}^{n_{(h)}} u_{(h)ilm} \right)^2 = \frac{16^2 L^4}{n_{(h)}^2} \sum_{i=1}^{n_{(h)}} \mathrm{Var}(u_{(h)ilm}) \le \frac{16^2 L^4}{n_{(h)}}, \quad \text{(C.4)}
\end{aligned}$$

where the first inequality in (C.4) follows from the fact that $u_{(h)il} \in [-1, 1]$ and the equality follows from $\mathbb{E}(u_{(h)ilm}) = 0$ and independence. Putting the pieces

together, we have shown that there exists a constant $d_2 \in (0, \infty)$ such that

$$\mathbb{E}\|\widehat{\boldsymbol{\Theta}}_{\cdot lm} - \boldsymbol{\tau}^*_{\cdot lm}\|_2 \leq \sqrt{\frac{d_2 H L^4}{n_{\min}}}.$$

Thus, we conclude that

$$P\left(\|\widehat{\boldsymbol{\Theta}}_{\cdot lm} - \boldsymbol{\tau}^*_{\cdot lm}\|_2 \geq \sqrt{\frac{d_2 H L^4}{n_{\min}}} + \epsilon\right) \leq \exp\left(-\frac{2n_{\min}^2 \epsilon^2}{d_1^2 L^4 N}\right).$$

Finally, applying the union bound with $\epsilon = \sqrt{c_2 \log(p)/n_{\min}}$ and $c_2 \geq d_2$,

$$\begin{aligned}
P\left(\max_{l \neq m} \|\widehat{\boldsymbol{\Theta}}_{\cdot lm} - \boldsymbol{\tau}^*_{\cdot lm}\|_2 \geq \sqrt{\frac{c_2 H L^4}{n_{\min}}} + \sqrt{\frac{c_2 \log(p)}{n_{\min}}}\right) \\
\leq P\left(\max_{l \neq m} \|\widehat{\boldsymbol{\Theta}}_{\cdot lm} - \boldsymbol{\tau}^*_{\cdot lm}\|_2 \geq \sqrt{\frac{d_2 H L^4}{n_{\min}}} + \sqrt{\frac{c_2 \log(p)}{n_{\min}}}\right) \\
\leq p^2 \; P\left(\|\widehat{\boldsymbol{\Theta}}_{\cdot lm} - \boldsymbol{\tau}^*_{\cdot lm}\|_2 \geq \sqrt{\frac{d_2 H L^4}{n_{\min}}} + \sqrt{\frac{c_2 \log(p)}{n_{\min}}}\right) \\
\leq p^2 \exp\left(-\frac{2n_{\min}^2}{d_1^2 L^4 N} \frac{c_2 \log(p)}{n_{\min}}\right) = p^{2 - \frac{2c_2 n_{\min}}{d_1^2 L^4 N}}.
\end{aligned}$$

□

Appendix D: Additional details for microbiome data analysis

First, we consider how our estimate of the log-abundance covariance matrix is affected by the amount added to zero counts, i.e., the pseudocount. Namely, we tried adding 0.01, instead of 0.5, to each of the counts before conversion to compositions. We used the same training/testing splits, though the set of candidate tuning parameters is slightly different since these are determined from the data in our software. The set of covariance matrix estimates are presented in Figure 11. Differences between these and those from Section 7 are relatively minor: the largest correlations in magnitude are consistent across both estimates, and the sparsity patterns are largely identical. Of the 76 off diagonals which were nonzero using the 0.5 pseudocount, 64 of those also appeared with the 0.01 pseudocount version of the data. Compared to the 0.05 pseudocount data, the covariance matrix estimates based on the 0.01 pseudocount do have more edges, though the correlations for many of these edges are estimated to be closer to zero. Note that we do not intend these results to imply that the pseudocount will not affect estimates–it will, and does. We only provide these to verify that estimates are reasonably similar across two distinct versions of the dataset.

In Table 4, we provide the phyla, family, genus, and species for each of the 39 OTUs from our data analyis. Finally, In Table 3, we include a version of

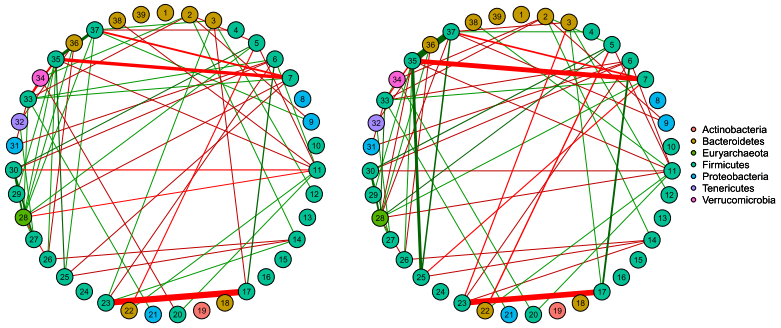


FIG 11. *Covariance matrix estimates with 0.01, instead of 0.5, added to all counts before conversion to compositions.*

TABLE 3

Stability for all correlations, shared correlations, and distinct correlations over 100 bootstrap samples. For the distinct correlation columns, D1 refers to a correlation which was nonzero in controls, but zero in ME/CFS, whereas D2 refers to a correlation which was zero in controls but nonzero in ME/CFS.

	All correlations						Shared correlations				Distinct correlations			
	SCC		COAT		SCC-H									
	Control	ME/CFS	Control	ME/CFS	Control	ME/CFS		SCC	COAT	SCC-H		SCC	COAT	SCC-H
Positive corr.	22	21	53	30	28	11	Same sign	29	19	6	D1	0	111	55
Negiatve corr.	16	17	82	68	33	8	Diff sign	9	5	0	D2	0	74	13
Stability	89.5%	86.8%	83.0%	84.7%	26.2%	73.7%	Stability	86.8%	58.3%	33.3%	Stability	—	00.0%	01.5%

Table 2 which includes SCC-H. Here, like COAT, SCC-H has lower stability than SCC. However, in general, SCC-H provides far sparser estimates than COAT, which may partially explain its lower stability. Either way, these results further suggest that the joint estimation procedure SCC provides more reliable estimates than separate estimators of the covariance matrices.

TABLE 4
Detailed information about the 39 OTUs from the data analyzed in Section 7.

Node	Phyla	Family	Genus	Species
1	Bacteroidetes	Porphyromonadaceae	Parabacteroides	
2	Bacteroidetes	Bacteroidaceae	Bacteroides	caccae
3	Bacteroidetes	Bacteroidaceae	Bacteroides	ovatus
4	Firmicutes	Lachnospiraceae	Blautia	
5	Firmicutes	Ruminococcaceae	Ruminococcus	
6	Firmicutes	Lachnospiraceae	Roseburia	faecis
7	Firmicutes	Lachnospiraceae	Ruminococcus	
8	Proteobacteria	Enterobacteriaceae	Escherichia	coli
9	Proteobacteria	Alcaligenaceae	Sutterella	
10	Firmicutes	Lachnospiraceae	Coprococcus	
11	Firmicutes	Lachnospiraceae	Blautia	producta
12	Firmicutes	Veillonellaceae	Phascolarctobacterium	
13	Firmicutes	Lachnospiraceae		
14	Firmicutes	Ruminococcaceae	Ruminococcus	
15	Firmicutes	Lachnospiraceae		
16	Firmicutes	Lachnospiraceae	Coprococcus	
17	Firmicutes	Ruminococcaceae	Ruminococcus	bromii
18	Bacteroidetes	Porphyromonadaceae	Parabacteroides	distasonis
19	Actinobacteria	Bifidobacteriaceae	Bifidobacterium	adolescentis
20	Firmicutes	Lachnospiraceae	Coprococcus	
21	Proteobacteria	Pseudomonadaceae	Pseudomonas	fragi
22	Bacteroidetes	Barnesiellaceae		
23	Firmicutes	Ruminococcaceae	Ruminococcus	bromii
24	Firmicutes			
25	Firmicutes	Ruminococcaceae	Oscillospira	
26	Firmicutes	Erysipelotrichaceae	Clostridium	saccharogumia
27	Firmicutes	Ruminococcaceae	Oscillospira	
28	Euryarchaeota	Methanobacteriaceae	Methanobrevibacter	
29	Firmicutes	Streptococcaceae	Streptococcus	
30	Firmicutes	Ruminococcaceae	Ruminococcus	
31	Proteobacteria	Enterobacteriaceae	Klebsiella	
32	Tenericutes			
33	Firmicutes	Lachnospiraceae	Lachnobacterium	
34	Verrucomicrobia	Verrucomicrobiaceae	Akkermansia	muciniphila
35	Firmicutes	Lachnospiraceae	Blautia	
36	Bacteroidetes	Barnesiellaceae		
37	Firmicutes	Lachnospiraceae	Blautia	producta
38	Bacteroidetes	S24-7		
39	Bacteroidetes	Bacteroidaceae	Bacteroides	uniformis

Acknowledgments

The authors thank the associate editor and two referees for their insightful comments and suggestions.

Funding

Aaron J. Molstad was supported in part by NSF DMS-2113589. Piotr M. Suder was supported in part by University Scholars Program at the University of Florida.

References

[1] Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)* **44** 139–160. MR0676206

[2] Aitchison, J. (2003). *The Statistical Analysis of Compositional Data*. Blackburn Press. MR0865647

[3] Ban, Y., An, L. and Jiang, H. (2015). Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics* **31** 3322–3329.

[4] Bien, J. and Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika* **98** 807–820. MR2860325

[5] Bigot, J., Biscay, R. J., Loubes, J.-M. and Muñiz-Alvarez, L. (2011). Group lasso estimation of high-dimensional covariance matrices. *The Journal of Machine Learning Research* **12** 3187–3225. MR2877598

[6] Cai, T. T., Li, H., Liu, W. and Xie, J. (2016). Joint estimation of multiple high-dimensional precision matrices. *Statistica Sinica* **26** 445-464. MR3497754

[7] Cao, Y., Lin, W. and Li, H. (2019). Large Covariance Estimation for Compositional Data Via Composition-Adjusted Thresholding. *Journal of the American Statistical Association* **114** 759-772. https://doi.org/10.1080/01621459.2018.1442340 MR3963178

[8] Danaher, P., Wang, P. and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 373–397. MR3164871

[9] Davis, D. and Yin, W. (2017). A Three-Operator Splitting Scheme and its Optimization Applications. *Set-Valued and Variational Analysis* **25** 829-858. https://doi.org/10.1007/s11228-017-0421-z MR3740519

[10] Fang, H., Huang, C., Zhao, H. and Deng, M. (2015). CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics* **31** 3172–3180.

[11] Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J. and Huttenhower, C. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS computational biology* **8** e1002606. MR3251360

[12] Friedman, J. and Alm, E. J. (2012). Inferring Correlation Networks from Genomic Survey Data. *PLOS Computational Biology* **8** 1-11. https://doi.org/10.1371/journal.pcbi.1002687

[13] Giloteaux, L., Goodrich, J. K., Walters, W. A., Levine, S. M., Ley, R. E. and Hanson, M. R. (2016). Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome. *Microbiome* **4** 30. https://doi.org/10.1186/s40168-016-0171-4

[14] Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology* **8** 2224. MR3992128

[15] Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98** 1–15. MR2804206

[16] He, Y., Liu, P., Zhang, X. and Zhou, W. (2021). Robust covariance estimation for high-dimensional compositional data with application to microbial communities analysis. *Statistics in Medicine* **40** 3499–3515. MR4269066

[17] Henrion, D. and Malick, J. (2012). Projection Methods in Conic Optimization. *Handbook on Semidefinite, Conic and Polynomial Optimization* 565-600. https://doi.org/10.1007/978-1-4614-0769-0_20 MR2894664

[18] Huson, D. H., Auch, A. F., Qi, J. and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research* **17** 377–386.

[19] Jiang, D., Armour, C. R., Hu, C., Mei, M., Tian, C., Sharpton, T. J. and Jiang, Y. (2019). Microbiome multi-omics network analysis: statistical considerations, limitations, and opportunities. *Frontiers in genetics* **10** 995.

[20] Li, D., Srinivasan, A., Chen, Q. and Xue, L. (2022). Robust Covariance Matrix Estimation for High-Dimensional Compositional Data with Application to Sales Data Analysis. *Journal of Business and Economic Statistics* 1–11. MR4650447

[21] Ma, J. and Michailidis, G. (2016). Joint structural estimation of multiple graphical models. *The Journal of Machine Learning Research* **17** 5777–5824. MR3555057

[22] Ma, J., Yue, K. and Shojaie, A. (2021). Networks for Compositional Data. *Statistical Analysis of Microbiome Data* 311–336.

[23] Matchado, M. S., Lauber, M., Reitmeier, S., Kacprowski, T., Baumbach, J., Haller, D. and List, M. (2021). Network analysis methods for studying microbial communities: A mini review. *Computational and*

structural biotechnology journal **19** 2687–2698.

[24] Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A Unified Framework for High-Dimensional Analysis of M-estimators with Decomposable Regularizers. *Statistical Science* **27**. https://doi.org/10.1214/12-STS400 MR3025133

[25] Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in Optimization* **1** 127–239.

[26] Pedregosa, F. and Gidel, G. (2018). Adaptive Three Operator Splitting. In *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.). *Proceedings of Machine Learning Research* **80** 4085–4094. PMLR.

[27] Porter, N. T. and Martens, E. C. (2016). Love thy neighbor: Sharing and cooperativity in the gut microbiota. *Cell Host and Microbe* **19** 745–746.

[28] Price, B. S., Geyer, C. J. and Rothman, A. J. (2015). Ridge fusion in statistical learning. *Journal of Computational and Graphical Statistics* **24** 439–454. MR3357389

[29] Price, B. S., Molstad, A. J. and Sherwood, B. (2021). Estimating multiple precision matrices with cluster fusion regularization. *Journal of Computational and Graphical Statistics* **30** 823–834. MR4356588

[30] Raskutti, G., Wainwright, M. J. and Yu, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *The Journal of Machine Learning Research* **11** 2241–2259. MR2719855

[31] Rothman, A. J. (2012). Positive definite estimators of large covariance matrices. *Biometrika* **99** 733–740. MR2966781

[32] Saegusa, T. and Shojaie, A. (2016). Joint estimation of precision matrices in heterogeneous populations. *Electronic Journal of Statistics* **10** 1341. MR3507368

[33] Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O. and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods* **9** 811–814.

[34] Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2013). A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics* **22** 231-245. https://doi.org/10.1080/10618600.2012.681250 MR3173712

[35] Sun, Y. and Vandenberghe, L. (2015). Decomposition methods for sparse matrix nearness problems. *SIAM Journal on Matrix Analysis and Applications* **36** 1691–1717. MR3432149

[36] Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science* **47**. Cambridge University Press. MR3837109

[37] Xu, J. and Lange, K. (2022). A proximal distance algorithm for

likelihood-based sparse covariance estimation. *Biometrika* **109** 1047-1066. https://doi.org/10.1093/biomet/asac011 MR4519115

[38] Xue, L., Ma, S. and Zou, H. (2012). Positive-definite L1-penalized estimation of large covariance matrices. *Journal of the American Statistical Association* **107** 1480–1491. MR3036409

[39] Younes, H., Coudray, C., Bellanger, J., Demigné, C., Rayssiguier, Y. and Rémésy, C. (2001). Effects of two fermentable carbohydrates (inulin and resistant starch) and their combination on calcium and magnesium balance in rats. *British Journal of Nutrition* **86** 479–485.

[40] Zhang, S., Wang, H. and Lin, W. (2023). CARE: Large Precision Matrix Estimation for Compositional Data. *arXiv preprint arXiv:2309.06985*.