Direct covariance matrix estimation with compositional data

Aaron J. Molstad[⋆,∗], Karl Oskar Ekvall[⋆,†], and Piotr M. Suder[‡]
Department of Statistics[⋆] and Genetics Institute[∗], University of Florida
Division of Biostatistics, Institute of Environmental Medicine,
Karolinska Institutet[†]
Department of Statistical Science, Duke University[‡]

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. Through simulation studies, we demonstrate that our estimator can outperform existing estimators. We show that our method provides more reliable estimates than competitors in an analysis of microbiome data from subjects with chronic fatigue syndrome.

Keywords: Compositional data, covariance matrix estimation, microbiome data analysis, convex optimization, positive definiteness, joint estimation

1 Introduction

High-dimensional compositional data arise in many areas of modern science. To study the human gut microbiome, for example, scientists measure the relative abundances of various microbes using next-generation sequencing followed by alignment and normalization (Shi et al., 2016). For each subject in a study, the resulting measurement is a p-dimensional vector which has nonnegative entries and sums to one (Huson et al., 2007). More generally, compositional data arise when, for example, one observes count-valued data wherein the total counts in a sample is not of interest. Here, we focus on compositional data which belong to the $(p-1)$-dimensional probability simplex, defined as

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

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

In many studies involving compositional microbiome data, scientists are interested in modeling the interactions and dependencies between microbes (Ma et al., 2021). For instance, one may want to estimate whether two microbes occur in higher abundance jointly. Covariance matrices provide one route for addressing such questions, but are not straightforward to estimate in this context. In particular, standard estimators of a covariance matrix perform poorly when the data are compositional (Fang et al., 2015; Cao et al., 2019).

To make matters concrete, let $X = (X_1, \ldots, X_p)^\top \in \mathbb{C}^p$ 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, we assume

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

To quantify the dependence between any two components from the compositional vector, the parameter of interest is the basis covariance matrix $\Omega^* \in \mathbb{S}^p_+$ where

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

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

Because the W_j are latent, we use independent realizations of the compositional X to estimate Ω^*. A common approach relies on the estimation of the variation matrix Θ^* (Aitchison, 2003, Chapter 4), defined elementwise by

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

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

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

To define an estimator of Θ^*, let $x_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{C}^p$, $i \in [n]$, denote independent realizations of X. Let also $z_{ijk} = \log(x_{ij}/x_{ik})$ and $\bar{z}_{jk} = n^{-1} \sum_{i=1}^{n} z_{ijk}$, $(j,k) \in [p] \times [p]$. 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 for 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 $\widehat{\Theta}$ are zero, so there are $p(p-1)/2$ unique equalities but $p(p-1)/2+p$ unknowns in $\widehat{\Omega}$. However, if one assumes that many entries of Ω^* are zero, then it can be estimated from (1). Cao et al. (2019) 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 we assume that only $s < p-1$ off-diagonal entries of Ω^* are nonzero, we may consider the estimator

$$\arg\min_{\Omega=\Omega^\top} \|\widehat{\Theta} - \omega\mathbb{1}_p^\top - \mathbb{1}_p\omega^\top + 2\Omega\|_F^2 \text{ subject to } \|\Omega^-\|_0 \leq s, \tag{2}$$

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 $\|A\|_0 = \sum_{j,k} \mathbf{1}(A_{jk} \neq 0)$ is the norm which counts the number of nonzero entries. Due to the L_0 constraint, (2) is the solution to a nonconvex optimization problem.

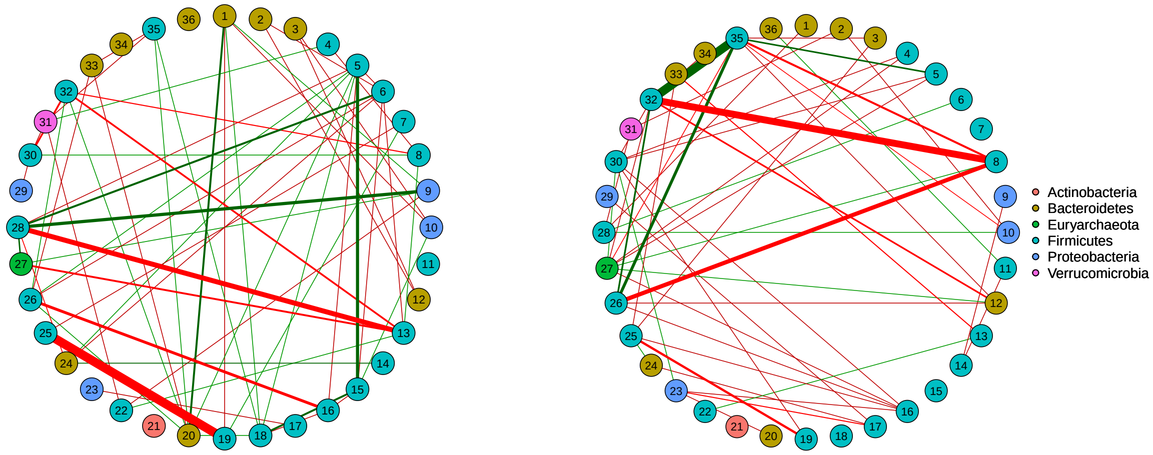
In view of (2), and given that the L_1 norm is a convex relaxation of the L_0 norm, a natural estimator is

$$\arg\min_{\Omega=\Omega^\top} \left\{ \|\widehat{\Theta} - \omega\mathbb{1}_p^\top - \mathbb{1}_p\omega^\top + 2\Omega\|_F^2 + \lambda\|\Omega^-\|_1 \right\}. \tag{3}$$

Cao et al. (2019) mentioned (3) as a direct version of their method, but did not study it. Appealingly, the problem in (3) can be recast as an L_1-penalized least squares problem and computed via existing algorithms. However, neither (2), (3), nor the method of Cao et al. (2019) provide estimates which are guaranteed to be positive definite, or even nonnegative definite (e.g., see Section 2.2). Replacing the feasible set in (2) or (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 W_j are observable), enforcing sparsity and positive definiteness is challenging (Bien and Tibshirani, 2011; Rothman, 2012; Xue et al., 2012).

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 chronic fatigue syndrome versus controls (Giloteaux et al., 2016). To estimate the two basis covariance matrices, one could apply existing estimators to each of the populations (CFS and controls) separately. However, sample sizes are often small relative to the dimension of the basis covariance. For example, there are only 36 and 47 control and chronic fatigue patients, respectively, used to estimate both 36×36 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 structures, exploiting this shared information across populations can substantially improve efficiency. Joint estimation is especially common in the literature on estimating sparse inverse covariance matrices from multivariate normal data (Guo et al., 2011; Danaher et al., 2014; Price et al., 2015; Saegusa and Shojaie, 2016; Price et al., 2021). In the context of



(a) Control

(b) Chronic fatigue syndrome

Figure 1: Estimated correlation networks for controls and patients with chronic fatigue syndrome (Giloteaux et al., 2016) using the method of Cao et al. (2019). Each node corresponds to an OTU as described in Section 6. 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 chronic fatigue syndrome.

estimating covariance matrices from microbiome data, it is natural to assume the covariance matrices have similar sparsity patterns. Biologically, it is often reasonable to assume there are microbes whose abundances are uncorrelated in all the populations in a study. For example, in Section 6, when we estimate the basis covariance matrices for controls and chronic fatigue syndrome patients 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 6.2.

In this article, we study (3) under positive definite constraints, and propose a generalization of (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 both simulation studies and our analysis of the chronic fatigue syndrome microbiome data, we demonstrate that our methods can provide more accurate and 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 the introduction, one can estimate any $\Theta^*_{(h)}$ using

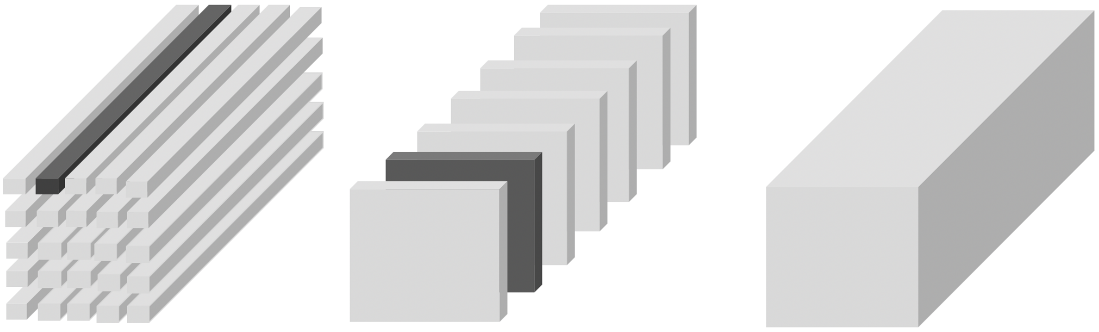
$$\widehat{\Theta}_{(h)jk} = \frac{1}{n_{(h)}} \sum_{i=1}^{n_{(h)}} (z_{(h)ijk} - \bar{z}_{(h)jk})^2, \quad (j,k) \in [p] \times [p],$$

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 $\boldsymbol{\Omega} \in \mathbb{R}^{H \times p \times p}$ to be the three-way tensor where $\boldsymbol{\Omega}_{h\cdot\cdot} = \Omega_{(h)} \in \mathbb{R}^{p \times p}$ for $h \in [H]$ and $\boldsymbol{\Omega}_{\cdot jk} = (\Omega_{(1)jk}, \ldots, \Omega_{(H)jk})^\top \in \mathbb{R}^H$ for $(j,k) \in [p] \times [p]$. We present a visualization of the tensor $\boldsymbol{\Omega}$ in Figure 2. The mode-1 fibers, $\boldsymbol{\Omega}_{\cdot jk}$, 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 $\boldsymbol{\Omega}^*_{\cdot jk} = 0$ for many pairs (j,k).

Generalizing (3) with an additional positive definiteness constraint, we propose to estimate $\boldsymbol{\Omega}^*$ using

$$\operatorname*{arg\,min}_{\boldsymbol{\Omega} \in \mathbb{R}^{H \times p \times p}} \left\{ \sum_{h=1}^{H} \left(\|\widehat{\Theta}_{(h)} - \omega_{(h)} \mathbb{1}_p^\top - \mathbb{1}_p \omega_{(h)}^\top + 2\Omega_{(h)}\|_F^2 + \lambda \|\Omega^-_{(h)}\|_1 \right) + \gamma \sum_{j \neq k} \|\boldsymbol{\Omega}_{\cdot jk}\|_2 \right\} \quad (4)$$



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

Figure 2: Visualization of (a) the fibers of $\boldsymbol{\Omega}$ which are penalized by the final term in (4), (b) the organization of $\boldsymbol{\Omega}$ into the $\Omega_{(h)}$, and (c) the three way tensor $\boldsymbol{\Omega}$.

$$\text{subject to } \ \omega_{(h)} = \operatorname{diag}(\Omega_{(h)}),\ \Omega_{(h)} = \Omega_{(h)}^{\top},\ \Omega_{(h)} \succcurlyeq \epsilon I_p, \quad \text{for all } \ h \in [H],$$

where $\lambda \geq 0$, $\gamma \geq 0$, and $\epsilon \geq 0$ are user-specified tuning parameters, $\|A\|_1 = \sum_{j,k} |A_{jk}|$ for a matrix A, $\|\cdot\|_2$ denotes the Euclidean norm of a vector, and the notation $A \succcurlyeq \epsilon I_p$ means that $A - \epsilon I_p$ is positive semidefinite.

The estimator (4) 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 $\boldsymbol{\Omega}$. Note that if $H = 1$, taking either $\lambda = 0$ or $\gamma = 0$ (with the other nonzero), (4) simplifies to (3) with a positive definiteness constraint. For example, if $\gamma = 0$, then (4) simplifies to the estimator

$$\operatorname*{arg\,min}_{\Omega_{(h)}=\Omega_{(h)}^{\top}} \left(\|\widehat{\Theta}_{(h)} - \omega_{(h)}\mathbb{1}_p^{\top} - \mathbb{1}_p\omega_{(h)}^{\top} + 2\Omega_{(h)}\|_F^2 + \lambda\|\Omega_{(h)}^{-}\|_1 \right), \tag{5}$$

$$\text{subject to } \ \omega_{(h)} = \operatorname{diag}(\Omega_{(h)}),\ \Omega_{(h)} \succcurlyeq \epsilon I_p$$

applied to each of the H populations separately. The estimator (5) can be seen as a convex approximation to (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 et al. (2012).

By taking $\gamma > 0$, however, (4) ties the estimators of $\Omega_{(1)}^{*}, \ldots, \Omega_{(H)}^{*}$ together. For large values of the tuning parameter γ, the second penalty in (4) will require that the solution to (4) has some $\boldsymbol{\Omega}_{\cdot jk} = 0$, i.e., that sparsity is partially shared across all H basis covariance matrix estimates. In the leftmost subfigure 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, (4) 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 (2) admits a closed-form solution.

Proposition 1. *If $p \geq 3$, the solution to (2) with $s = 0$ (or equivalently, (3) with $\lambda = \infty$) is unique and is given by*

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

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, $\hat{\omega}_j$ 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 $x_i \in \mathbb{C}^3$ by drawing the $\log(W_i)$ 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, $\hat{\omega}_3 = (2.45 + 1.24)/2 - 3.83/2 = -0.07$. Intuitively, negative variance estimates are more likely when p is large relative to n.

2.3 Existing estimators

An alternative estimator of an individual basis covariance matrix Ω^* is based on the centered log-ratio covariance matrix (Aitchison, 1982) 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,

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

so that $\Theta^* = \gamma^* \mathbb{1}_p^\top + \mathbb{1}_p \gamma^{*\top} - 2\Gamma^*$, where $\gamma^* = \mathrm{diag}(\Gamma^*)$. Cao et al. (2019) 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. (2015) 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

$$\underset{\Omega = \Omega^\top}{\arg\min} \left\{ \frac{1}{2} \| F(\Omega - \widehat{\Omega}_X)F \|_V^2 + \lambda \|\Omega^-\|_1 \right\}, \tag{6}$$

where $V = \{\mathrm{diag}(F\widehat{\Omega}_X F)\}^{-1}$ and $\|A\|_V^2 = \mathrm{tr}(A^\top V A)$. While Fang et al. (2015) suggests a positive definiteness constraint on (6), their computational algorithm does not enforce this constraint. Instead, if the solution to (6) 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 (Sun and Vandenberghe, 2015).

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 (2012); Ban et al. (2015); He et al. (2021); Li et al. (2022), for example, and see Ma et al. (2021) for a comprehensive review.

3 Computation

3.1 Proximal-proximal gradient descent

In order to solve the optimization problem to compute (4), 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 (Davis and Yin, 2017), which allows us to handle the nondifferentiable penalty and positive definiteness constraint separately. The algorithm generalizes the well-known proximal gradient descent algorithm (Parikh and Boyd, 2014, 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

$$\underset{u \in \mathcal{U}}{\mathrm{minimize}} \left\{ \ell(u) + f(u) + g(u) \right\}. \tag{7}$$

Further suppose there exists $u^\star \in \mathcal{U} \subseteq \mathbb{R}^d$ 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 (2017) show that (7) 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 $\mathcal{U}$ and $\alpha \in (0, 2\beta)$. Here, the superscript (t) denotes the (t)th iterate. As $t \to \infty$, $u_g^{(t)} \to u^\star$ and $u_f^{(t)} \to u^\star$ (Davis and Yin, 2017). 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 (2018), which allows us to start with a value of the step size α which is larger than 2β and reduce its value as needed, thus accelerating the descent. The $(t+1)$th iterates of the algorithm use the updating equations

$$u_f^{(t+1)} = \mathbf{prox}_{\alpha f}\{u_g^{(t)} - \alpha v^{(t)} - \alpha \nabla \ell(u_g^{(t)})\} \tag{8}$$

$$u_g^{(t+1)} = \mathbf{prox}_{\alpha g}(u_f^{(t+1)} + \alpha v^{(t)}) \tag{9}$$

$$v^{(t+1)} = v^{(t)} + \alpha^{-1}(u_f^{(t+1)} - u_g^{(t+1)}) \tag{10}$$

At each step, after (8) is carried out, the value

$$Q(u_f^{(t+1)}, \alpha) := \ell(u_g^{(t)}) + \langle \nabla \ell(u_g^{(t)}), u_f^{(t+1)} - u_g^{(t)} \rangle + \frac{1}{2\alpha} \langle u_f^{(t+1)} - u_g^{(t)}, u_f^{(t+1)} - u_g^{(t)} \rangle \tag{11}$$

is compared to $\ell(u_f^{(t+1)})$, where $\langle \cdot, \cdot \rangle$ is the inner product. If $\ell(u_f^{(t+1)}) \leq Q(u_f^{(t+1)}, \alpha)$, then the algorithm proceedes to (9). 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 (8) 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 (4) as (7) 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 (4) in a form analogous to (7), 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 objective function from (4) is

$$\left[\sum_{h=1}^{H} \left\{ \|\widehat{\Theta}_{(h)} - \omega_{(h)} \mathbb{1}_p^\top - \mathbb{1}_p \omega_{(h)}^\top + 2\Omega_{(h)}\|_F^2 + \lambda \|\Omega_{(h)}^-\|_1 + \chi_\epsilon(\Omega_{(h)}) \right\} + \gamma \sum_{j \neq k} \|\mathbf{\Omega}_{\cdot jk}\|_2 \right]. \tag{12}$$

If we minimize (12) 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 $\ell(\boldsymbol{\Omega}) = \sum_{h=1}^H \|\widehat{\Theta}_{(h)} - \omega_{(h)}\mathbb{1}^\top - \mathbb{1}\omega_{(h)}^\top + 2\Omega_{(h)}\|_F^2$, $f(\boldsymbol{\Omega}) = \lambda\sum_{h=1}^H \|\Omega^-_{(h)}\|_1 + \gamma \sum\sum_{j\neq k}\|\boldsymbol{\Omega}_{\cdot jk}\|_2$, and $g(\boldsymbol{\Omega}) = \sum_{h=1}^H \chi_\epsilon(\Omega_{(h)})$, (12) has the form of (7). 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 $\widehat{\Theta}_{(1)}, \dots, \widehat{\Theta}_{(H)}$ so that $\widehat{\boldsymbol{\Theta}}_{hjk} = \widehat{\Theta}_{(h)jk}$, the function ℓ is differentiable with respect to $\boldsymbol{\Omega}$ with gradient

$$[\nabla\ell(\boldsymbol{\Omega})]_{hjk} = \begin{cases} \sum\limits_{l\in[p]\setminus\{j\}} (4\boldsymbol{\Omega}_{hjj} - 4\widehat{\boldsymbol{\Theta}}_{hjl} - 8\boldsymbol{\Omega}_{hjl} + 4\boldsymbol{\Omega}_{hll}) & : j = k \\ 8\boldsymbol{\Omega}_{hjk} - 4\boldsymbol{\Omega}_{hjj} - 4\boldsymbol{\Omega}_{hkk} + 4\widehat{\boldsymbol{\Theta}}_{hjk} & : j \neq k \end{cases}, \tag{13}$$

for all $(h, j, k) \in [H] \times [p] \times [p]$. The updating equations corresponding to (8)–(10) are

$$\boldsymbol{\Omega}^{(t+1)} = \underset{\boldsymbol{\Omega}\in\mathbb{R}^{H\times p\times p}}{\arg\min} \left\{ \frac{1}{2}|||\boldsymbol{\Omega} - \boldsymbol{\Gamma}^{(t)}|||_F^2 + \alpha\lambda\sum_{h=1}^H \|\boldsymbol{\Omega}^-_{(h)}\|_1 + \alpha\gamma\sum_{j\neq k}\|\boldsymbol{\Omega}_{\cdot jk}\|_2 \right\} \tag{14}$$

$$\tilde{\boldsymbol{\Omega}}^{(t+1)} = \underset{\boldsymbol{\Omega}\in\mathbb{R}^{H\times p\times p}}{\arg\min} \left\{ \frac{1}{2}|||\boldsymbol{\Omega} - \boldsymbol{\Omega}^{(t+1)} - \alpha\boldsymbol{\Psi}^{(t)}|||_F^2 + \alpha\sum_{h=1}^H \chi_\epsilon(\Omega_{(h)}) \right\} \tag{15}$$

$$\boldsymbol{\Psi}^{(t+1)} = \boldsymbol{\Psi}^{(t)} + \alpha^{-1}(\boldsymbol{\Omega}^{(t+1)} - \tilde{\boldsymbol{\Omega}}^{(t+1)}) \tag{16}$$

where $\boldsymbol{\Gamma}^{(t)} = \tilde{\boldsymbol{\Omega}}^{(t)} - \alpha\boldsymbol{\Psi}^{(t)} - \alpha\nabla\ell(\tilde{\boldsymbol{\Omega}}^{(t)})$ and $|||\boldsymbol{A}|||_F^2 = \sum_{h,j,k}\boldsymbol{A}_{hjk}^2$ for a three-way tensor $\boldsymbol{A}$. Because (16) is trivial, we focus on (14) and (15).

First, (14) can be separated across the second and third mode of $\boldsymbol{\Omega}$ since for all $(j, k) \in [p] \times [p]$ such that $j \neq k$,

$$\boldsymbol{\Omega}^{(t+1)}_{\cdot jk} = \underset{x\in\mathbb{R}^H}{\arg\min} \left\{ \frac{1}{2}\|x - \boldsymbol{\Gamma}^{(t)}_{\cdot jk}\|_2^2 + \alpha\lambda\|x\|_1 + \alpha\gamma\|x\|_2 \right\}, \tag{17}$$

and $\boldsymbol{\Omega}^{(t+1)}_{\cdot jj} = \boldsymbol{\Gamma}^{(t)}_{\cdot jj}$ for $j \in [p]$. The solution to (17) is

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

where $(a)_+ = \max(a, 0)$ and $\mathbf{soft}(y, \tau) = \max(|y| - \tau, 0)\text{sign}(y)$ is applied elementwise (Simon et al., 2013). The second step, (15), also has a closed form solution. In particular, (15) can be solved with respect to each $\Omega_{(h)}$ separately, in parallel, using that

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

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

```
Initialize $\boldsymbol{\Psi}^{(0)} \in \mathbb{R}^{H\times p\times p}, \tilde{\boldsymbol{\Omega}}^{(0)} \in \mathbb{R}^{H\times p\times p}$, $\alpha > 0$, 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 $\boldsymbol{\Omega}^{(t+1)}_{\cdot jj} = \tilde{\boldsymbol{\Omega}}^{(t)}_{\cdot jj} - \alpha\boldsymbol{\Psi}^{(t)}_{\cdot jj} - \alpha[\nabla\ell(\tilde{\boldsymbol{\Omega}}^{(t)})]_{\cdot jj}$
        1.2. If $j \neq k$
            1.2.1. Set $y = \tilde{\boldsymbol{\Omega}}^{(t)}_{\cdot jk} - \alpha\boldsymbol{\Psi}^{(t)}_{\cdot jk} - \alpha[\nabla\ell(\tilde{\boldsymbol{\Omega}}^{(t)})]_{\cdot jk}$
            1.2.2. Set $w_h = (|y_h| - \alpha\lambda)_+ \mathrm{sign}(y_h)$ for $h \in [H]$
            1.2.3. Set $\boldsymbol{\Omega}^{(t+1)}_{\cdot jk} = \left(1 - \frac{\alpha\gamma}{\|w\|_2}\right)_+ w$
    2. If $\ell(\boldsymbol{\Omega}^{(t+1)}) \leq Q(\boldsymbol{\Omega}^{(t+1)}, \alpha)$, proceed to 3. Else, set $\alpha = \tau\alpha$, and return to 1.
    3. For $h \in [H]$
        3.1. Decompose $\boldsymbol{\Omega}^{(t+1)}_{h\cdot\cdot} + \alpha\boldsymbol{\Psi}^{(t)}_{h\cdot\cdot} = \sum_{j=1}^{p} \xi_j \boldsymbol{u}_j \boldsymbol{u}_j^\top$, 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 $\tilde{\boldsymbol{\Omega}}^{(t+1)}_{h\cdot\cdot} = \sum_{j=1}^{p} \max(\xi_j, \epsilon)\boldsymbol{u}_j\boldsymbol{u}_j^\top$
    4. Set $\boldsymbol{\Psi}^{(t+1)} = \boldsymbol{\Psi}^{(t)} + \alpha^{-1}(\boldsymbol{\Omega}^{(t+1)} - \tilde{\boldsymbol{\Omega}}^{(t+1)})$
    5. If objective function value converged, terminate. Else, set $t = t + 1$ and go to 1.
```

where $u_{(h)j}$ and $\xi_{(h)j}$ are the jth eigenvector and eigenvalue of $\Omega^{(t+1)}_{(h)} + \alpha\boldsymbol{\Psi}^{(t)}_{h\cdot\cdot}$, respectively, for $h \in [H]$ (Henrion and Malick, 2012). This is the projection of $\Omega^{(t+1)}_{(h)} + \alpha\boldsymbol{\Psi}^{(t)}_{h\cdot\cdot}$ 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 (2018). 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.

3.3 Practical considerations

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

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

where $\widetilde{\boldsymbol{\Omega}}^{\lambda,\gamma}_{-v}$ is the solution to (4) with input sample variation matrices $\widehat{\boldsymbol{\Theta}}_{-v}$, which are computed using all the data from outside the vth fold.

4 Statistical properties

4.1 Asymptotics for single population estimator

Though our primary focus is the multipopulation estimator (4), the estimator (5) 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

$$\underset{\Omega=\Omega^\top}{\arg\min} \left\{ \|\widehat{\Theta} - \omega \mathbb{1}_p^\top - \mathbb{1}_p \omega^\top + 2\Omega\|_F^2 + \lambda \|\Omega^-\|_1 \right\} \text{ subject to } \Omega H \epsilon I_p. \tag{18}$$

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 $W = (W_1, \dots, W_p)^\top$ 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 $n \to \infty$, $\log(p)/n \to 0$.

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^*. For this assumption to hold, we need p to grow with n. This is consistent with the assumptions in Cao et al. (2019), who also require $p \to \infty$ as $n \to \infty$ and characterize this as a "blessing of dimensionality". Assumption **A2**, as discussed in an earlier section, also serves to address the identifiability of Ω^*. Assumption **A3** is standard in high-dimensional covariance matrix estimation.

We now state our first result concerning the asymptotic error of our estimator. Let φ_p be the pth largest eigenvalue of its matrix-valued argument.

Theorem 1. *Suppose **A1**–**A3** hold. If $\epsilon < \varphi_p(\Omega^*)$ and $\lambda = \sqrt{c_1 \log(p)/n}$ for constant $c_1 > 0$ sufficiently large, then*

$$\frac{\|[\widehat{\Omega} - \Omega^*]^-\|_F}{\sqrt{p}} + \|\widehat{\omega} - \omega^*\|_2 = O_{\mathrm{P}}\left(\sqrt{\frac{s\log(p)}{np}} + \sqrt{\frac{p\log(p)}{n}} \right), \tag{19}$$

where $s = \sum_{j=1}^p s_j$. Under the same conditions, $\|\widehat{\omega} - \omega^*\|_2 = O_{\mathrm{P}}(\sqrt{p\log(p)/n})$.

The error bound in (19) consists of two parts: the error for estimating off-diagonals and the diagonals. The Euclidean norm error for the diagonals is $O_P(\sqrt{p\log(p)/n})$. The Frobenius norm error for the off-diagonals, however, cannot be disentangled from the diagonal error. Though our bound would seem to suggest that $\|[\widehat{\Omega} - \Omega^*]^-\|_F = O_P(\sqrt{s\log(p)/n})$, 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 (18). 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 (18) is the solution to a penalized least squares problem, we do not assume any type of restricted eigenvalue condition (Raskutti et al., 2010). Instead, in our proof we first show that $\widehat{\Omega} - \Omega^*$ belongs to a restricted set, then establish a quadratic lower bound on $\ell(\widehat{\Omega}) - \ell(\Omega^*) - \operatorname{tr}\{\nabla \ell(\Omega^*)^\top (\widehat{\Omega} - \Omega^*)\}$ over this set where here, ℓ is the objective function from (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 et al. (2019) is not possible as their results are given in terms of the spectral norm, and under a different set of assumptions.

4.2 Asymptotics for multiple population estimator

Next, we consider the multiple population estimator (4) with $\lambda = 0$. This version of (4) can exploit shared sparsity patterns across the $\Omega^*_{(h)}$. 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 $W_{(h)} = (W_{(h)1}, \ldots, W_{(h)p})^\top \in \mathbb{R}^p$ such that $\log(W_{(h)k}) \in [-L, L]$ for $k \in [p]$ and $\operatorname{Cov}\{\log(W_{(h)})\} = \Omega^*_{(h)}$ for all $h \in [H]$.

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

A6. (Nonvanishing $n_{(h)}/N$). As $N \to \infty$, there exists $\pi > 0$ such that for sufficiently large N, $\min_h n_{(h)}/N \geq \pi$.

A7. (Alignment of $n_{(h)}$, p, H, and L). As $N \to \infty$, $\log(p)/n_{(h)} \to 0$ and $L^4 H / n_{(h)} \to 0$ 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 $\nabla \ell(\mathbf{\Omega}^*)$. The quantity L will appear in our asymptotic error bound, so this assumption is not so restrictive since L can be arbitairly 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** also requires that p grows as $N \to \infty$. Assumption **A6** is a requirement on how frequently, as $N \to \infty$, we draw a sample from each of the H populations. If the population to sample from was selected randomly, this would require that each population is sampled from with probability bounded away from zero. If the population to sample from was chosen deterministically, this would require that we do not systemically undersample from any of the H populations. Our error bounds will depend on π, so we can quantify how π affects estimation accuracy.

We are ready to state our asymptotic error bound for (4) with $\lambda = 0$. Here, $a_N \lesssim b_N$ means that there exists a constant K such that $a_N \leq K b_N$ for all N.

Theorem 2. *Suppose **A4**—**A7** hold. Define $\widehat{\boldsymbol{\Omega}}$ as the solution to* (4) *with $\lambda = 0$. Let $\boldsymbol{\omega}^* = (\omega^*_{(1)}, \dots, \omega^*_{(H)})$ and $\widehat{\boldsymbol{\omega}} = (\widehat{\omega}_{(1)}, \dots, \widehat{\omega}_{(H)})$. If $\epsilon < \min_h \varphi_p(\Omega^*_{(h)})$ and $\gamma = \sqrt{c_2 L^4 H/\pi N} + \sqrt{c_2 \log(p)/\pi N}$ for constant $c_2 > 0$ sufficiently large, then*

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

with probability tending to one as $N \to \infty$, where $\tilde{s} = \sum_{j=1}^p \tilde{s}_j$. Under the same conditions,

$$\|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}^*\|_F \lesssim \sqrt{\frac{L^4 H}{\pi N}} + \sqrt{\frac{p \log(p)}{\pi N}}$$

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

The bound in Theorem 2 can be intepreted in a similar way as the bound in Theorem 1. Specifically, we cannot separate the error for estimating the off-diagonals of $\Omega^*_{(h)}$ 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 $\sqrt{\tilde{s}}$ 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 1, we can establish a bound specifically for the diagonals. If $HL^4 = O\{p \log(p)\}$, which is natural since one would not expect H nor L to grow with N, we then achieve essentially the same error bound for estimating the diagonals as in Theorem 1: $\|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}^*\|_F \lesssim \sqrt{p \log(p)/\pi N}$ with probability tending to one.

5 Numerical experiments

5.1 Data generating models and competing methods

In this section, we compare the proposed estimator, (4), to existing estimators under various data generating models. Specifically, we consider three distinct data generating models, Models 1–3, with $H = 4$ and $(n, p) \in \{50, 100, 150\} \times \{40, 80, 120, 160, 200\}$. In each replication, we generate $\log(W_{(h)1}), \dots, \log(W_{(h)n_{(h)}})$ 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:

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

  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:

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

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

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

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

  $D \in \mathbb{S}_p^+$ is diagonal matrix with diagonal entries equally spaced from 3 to 1, and $\mathcal{B}_1 = [p/2] \times [p/2]$, $\mathcal{B}_2 = \{p/6+1, \dots, 2p/3\} \times \{p/6+1, \dots, 2p/3\}$, $\mathcal{B}_3 = \{p/3+1, \dots, 5p/6\} \times \{p/3+1, \dots, 5p/6\}$, and $\mathcal{B}_4 = \{p/2+1, \dots, p\} \times \{p/2+1, \dots, p\}$

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 $\boldsymbol{\Omega}_*$, we consider multiple methods. All methods but (4) estimate $\Omega^*_{(1)}, \dots, \Omega^*_{(4)}$ separately. Specifically, we use the method of Cao et al. (2019) with adapative soft-thresholding, `COAT`, and use the method of Fang et al. (2015), `cclasso`. We also use an oracle estimator, `Oracle`, which is the adapatively soft-thresholded sample covariance matrix of each $\log(W_{(h)1}), \dots, \log(W_{(h)n_{(h)}})$. 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, `MCC` and `MCC-H`, where `MCC` is short for multiple compositional covariance. The estimator `MCC` is defined in (4), whereas `MCC-H` is (5) with a separate tuning parameter λ chosen for each $h \in [H]$. The method `MCC-H` estimates the covariances separately, but using a version of our criterion. Including both `MCC` and `MCC-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 $\boldsymbol{\Omega}_*$.

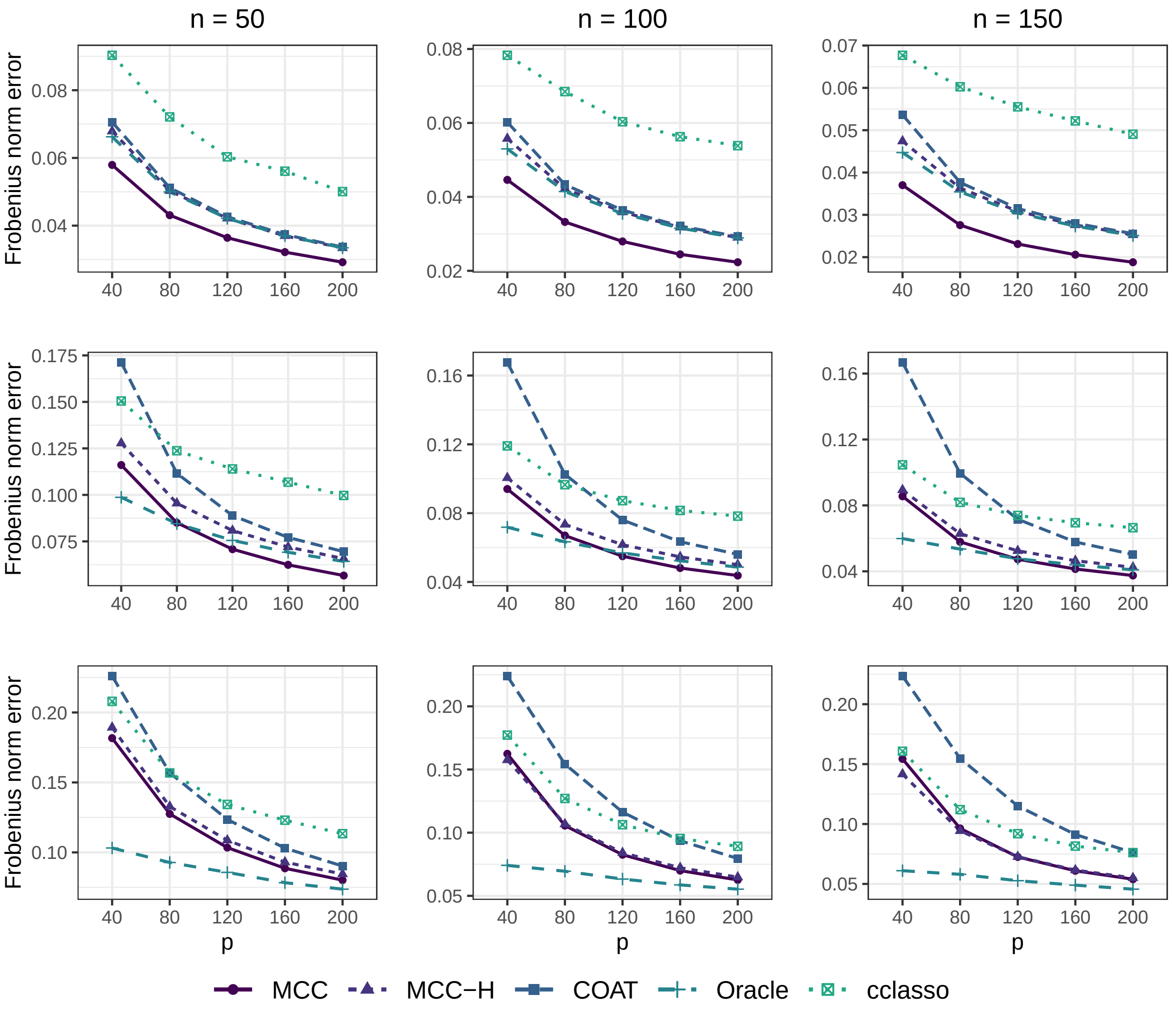


Figure 3: Average Frobenius norm error (divided by p) over 50 independent replications under (top row) Model 1, (middle row) Model 2, and (bottom row) Model 3 with $(n, p) \in \{50, 100, 150\} \times \{40, 80, 120, 160, 200\}$.

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. We provide both Frobenius norm and L_1 matrix-norm error on the covariance scale in the Supplementary Material.

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

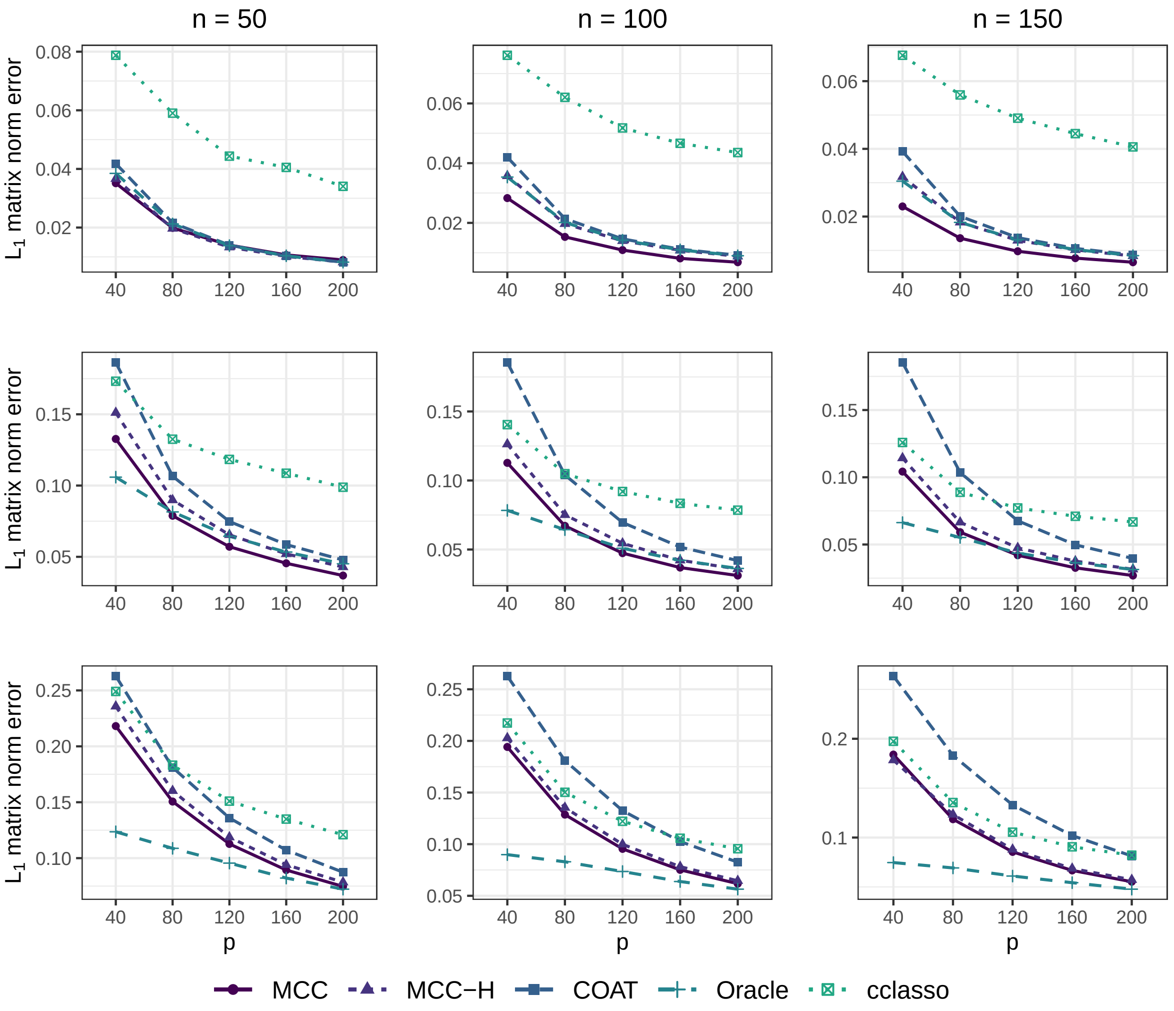


Figure 4: Average L_1 matrix norm error (divided by p) over 50 independent replications under (top row) Model 1, (middle row) Model 2, and (bottom row) Model 3 with $(n, p) \in \{50, 100, 150\} \times \{40, 80, 120, 160, 200\}$.

and TNR are defined as, respectively,

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

5.2 Results

In Figure 3, we display average Frobenius norm errors (divided by p). Unsurprisingly, under Model 1, `MCC` substantially outperforms all of the competitors, including `Oracle`. This illustrates the utility of exploiting shared sparsity patterns when estimating multiple

		$n = 50$					$n = 100$					$n = 150$				
		40	80	120	160	200	40	80	120	160	200	40	80	120	160	200
Model 1																
TPR	MCC	0.954	0.925	0.899	0.877	0.859	0.998	0.996	0.994	0.995	0.995	1.000	0.999	0.999	0.999	0.998
	MCC-H	0.570	0.499	0.447	0.422	0.398	0.832	0.778	0.747	0.719	0.696	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	MCC	0.701	0.824	0.874	0.910	0.925	0.647	0.809	0.855	0.894	0.904	0.667	0.761	0.816	0.848	0.871
	MCC-H	0.892	0.949	0.968	0.977	0.983	0.800	0.890	0.923	0.941	0.952	0.770	0.861	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	MCC	0.902	0.706	0.567	0.489	0.403	0.953	0.791	0.655	0.559	0.489	0.969	0.821	0.695	0.583	0.504
	MCC-H	0.765	0.529	0.404	0.331	0.274	0.856	0.631	0.486	0.402	0.342	0.883	0.672	0.529	0.435	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	1.000	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	MCC	0.250	0.486	0.608	0.664	0.734	0.137	0.386	0.528	0.612	0.661	0.098	0.350	0.491	0.590	0.653
	MCC-H	0.540	0.718	0.792	0.829	0.861	0.462	0.644	0.749	0.797	0.823	0.435	0.624	0.725	0.776	0.806
	COAT	0.328	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	0.577	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	0.000	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	MCC	0.933	0.721	0.596	0.525	0.455	0.981	0.807	0.689	0.595	0.520	0.992	0.880	0.761	0.646	0.564
	MCC-H	0.691	0.557	0.463	0.404	0.343	0.794	0.655	0.567	0.491	0.426	0.831	0.696	0.618	0.535	0.469
	COAT	0.840	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	0.959	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	0.998	0.999	0.998	0.997	0.994	1.000	1.000	1.000	0.999	0.999	1.000	1.000	1.000	1.000	1.000
TNR	MCC	0.104	0.412	0.563	0.633	0.684	0.028	0.320	0.480	0.589	0.654	0.012	0.210	0.395	0.535	0.619
	MCC-H	0.437	0.651	0.742	0.790	0.823	0.321	0.560	0.668	0.738	0.783	0.291	0.518	0.639	0.713	0.760
	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	0.555	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	0.001	0.002	0.003	0.006	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000

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

covariance matrices. Notably, `Oracle`, `COAT`, and `MCC-H` all perform similarly in each setting under Model 1. Under Model 2, `MCC` outperforms all competitors, including `Oracle`, once $p \geq 120$. Comparing the competitors which could be used in practice, `MCC-H` performs better than `COAT` and `cclasso` in all situations. The estimator `COAT` performs worse than `cclasso` for small p, but significantly outperforms `cclasso` once $p \geq 120$. Finally, under Model 3, `MCC` and `MCC-H` perform similarly in every setting. The only method to outperform `MCC` and `MCC-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 drawning conclusions based on Frobenius norm error results alone, however, because our methods, `MCC` and `MCC-H`, both minimize a Frobenius norm criterion, whereas `COAT` does not. Thus, these results may be biased in favor of `MCC` and `MCC-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 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 $n \geq 100$, MCC significantly outperforms all competitors. Under Model 2, the results more closely mirror those in Figure 3: MCC 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, MCC and MCC-H clearly outperform cclasso and COAT.

The performance of MCC and MCC-H can be partially explained by their performance in recovering the true set of nonzero off-diagonals. In Model 1, MCC has nearly perfect TPR, and TNR only slightly lower than the best performing competitor. MCC-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 MCC, whereas MCC-H has lower TPR and higher TNR than COAT.

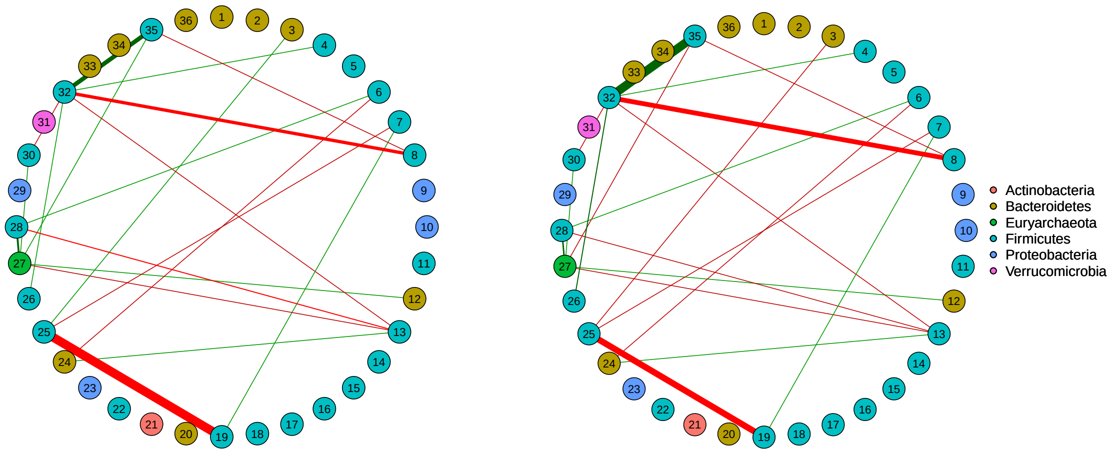
6 Analysis of microbiome in chronic fatigue syndrome

6.1 Basis covariance matrix estimation

We compare our method to that of Cao et al. (2019) on a dataset comparing the gut microbiome of patients diagnosed with myalgic encephalomyelitis/chronic fatigue syndrome (CFS) versus controls from Giloteaux et al. (2016). In order to obtain the microbial profiles, Giloteaux et al. (2016) sequenced 16S rRNA genes from stool samples using Illumina MiSeq. After filtering patients based on total reads (≥ 5000), and filtering operational taxonomic units (OTUs) based on total reads (≥ 5000), we were left with count data for 37 patients in the control group and 47 patients with CFS. Following Cao et al. (2019), we add 0.5 to all counts to avoid zeros before converting counts to compositions.

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 the phyla and each node's genera is provided in the Supplementary Material. 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.

Examining the estimated covariance matrices, the majority of associations occur within two OTUs belonging to the same phyla. We also see that our method estimates the two groups' covariance matrices to have identical sparsity patterns, in contrast to the estimates based on the method of Cao et al. (2019). Strong positive and negative associations are shared across the two groups. Notably, the only difference between the two is the direction of association between (3–25, *Ruminoscoccus bromii-Bateroides ovatus*) and (27–35, *Methanobrevibacter-Blautia producta*): both are positive in controls, but negative in patients with CFS. *Ruminoscoccus bromii* is known to digest resistant starch particles inaccessible to other bacteria, whereas



(a) Controls

(b) Chronic fatigue syndrome

Figure 5: Estimated covariance matrices using (4) for (a) control patients and (b) patients with chronic fatigue syndrome. 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.

Bateroides ovatus digests inulin (Porter and Martens, 2016). That these two OTUs are estimated to be associated is not suprising since both resistant starch and inulin are fermentable carbohydrates whose joint behavior has been of interest in past studies (Younes et al., 2001).

In addition, an insight gleaned from our estimates is that more negative associations are observed in chronic fatigue patients than in controls. This coheres with the reduced diversity in the microbiome communities for these patients observed by Giloteaux et al. (2016).

Estimates using the method of Cao et al. (2019), displayed in Figure 1, 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, the method of Cao et al. (2019) identifies 104 assocations in one population not present in the other. Moreover, the estimates from Cao et al. (2019) disagree in terms of their strongest associations. For example, the strongest positive association estimated in controls is between (9–28), whereas in patients with CFS, their method estimates these two OTUs to be uncorrelated.

6.2 Stability assessment

We perform a stability assessment to determine to what degree our respective estimates are reliable. Following Cao et al. (2019), we generate 100 independent bootstrap datasets and refit both estimators to the bootstraped 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 row labeled stability, report the percentage of these

	All correlations				Shared correlations			Distinct correlations		
	MCC		COAT							
	Control	CFS	Control	CFS		MCC	COAT		MCC	COAT
Positive	28	18	18	40	Same sign	38	6	D1	0	26
Negative	20	30	14	78	Diff sign	10	0	D2	0	112
Stability	100	97.6	94.1	80.5	Stability	95.8	100.00	Stability	—	0.00

Table 2: Stability for all correlations, shared correlations, and distinct correlations over 100 bootstrap samles. For the distinct correlation columns, D1 refers to a correlation which was nonzero in controls, but zero in CFS, whereas D2 refers to a correlation which was zero in controls but nonzero in CFS.

correlations which were estimated to be nonzero at least 95 of the 100 bootstrap samples. For our method, in both controls and CFS basis covariance matrix estimates, almost all of the edges we estimated to be nonzero are stable. COAT, on the other hand, has substantially lower stability in both controls and CFS.

In the "shared correlations" columns of Table 2, we report the number of estimated correlations where a correlation was positive in both estimates (controls and CFS), or negative in both estimates. We see that COAT did not estimate any correlation to be negative in both controls and CFS. Our method, despite estimating a much larger number of shared correlations, has relatively high stability. COAT too has high stability, but based on only six shared correlations.

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 CFS (D1) and the number of correlations which were zero in controls and nonzero in CFS (D2). We see that MCC estimates no edges to be distinct, whereas COAT estimates 138 correlations to be distinct. However, the stability of these edges is zero: none of these correlations appeared in 95 or more of the bootstrap samples. This suggests that the estimates provided by our method are more reliable.

Acknowledgements

Molstad's research was supported in part by a grant from the National Science Foundation (DMS-2113589). Suder's research was supported in part by the University Scholars Program at the University of Florida.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160.

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

Ban, Y., An, L., and Jiang, H. (2015). Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics*, 31(20):3322–3329.

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

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(526):759–772.

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(2):373–397.

Davis, D. and Yin, W. (2017). A three-operator splitting scheme and its optimization applications. *Set-Valued and Variational Analysis*, 25(4):829–858.

Fang, H., Huang, C., Zhao, H., and Deng, M. (2015). Cclasso: correlation inference for compositional data through lasso. *Bioinformatics*, 31(19):3172–3180.

Friedman, J. and Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLOS Computational Biology*, 8(9):1–11.

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(1):30.

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

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(15):3499–3515.

Henrion, D. and Malick, J. (2012). Projection methods in conic optimization. *Handbook on Semidefinite, Conic and Polynomial Optimization*, pages 565–600.

Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, 17(3):377–386.

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*, pages 1–11.

Ma, J., Yue, K., and Shojaie, A. (2021). Networks for compositional data. *Statistical Analysis of Microbiome Data*, pages 311–336.

Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239.

Pedregosa, F. and Gidel, G. (2018). Adaptive three operator splitting. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4085–4094. PMLR.

Porter, N. T. and Martens, E. C. (2016). Love thy neighbor: Sharing and cooperativity in the gut microbiota. *Cell Host and Microbe*, 19(6):745–746.

Price, B. S., Geyer, C. J., and Rothman, A. J. (2015). Ridge fusion in statistical learning. *Journal of Computational and Graphical Statistics*, 24(2):439–454.

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(4):823–834.

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.

Rothman, A. J. (2012). Positive definite estimators of large covariance matrices. *Biometrika*, 99(3):733–740.

Saegusa, T. and Shojaie, A. (2016). Joint estimation of precision matrices in heterogeneous populations. *Electronic Journal of Statistics*, 10(1):1341.

Shi, P., Zhang, A., and Li, H. (2016). Regression analysis for microbiome compositional data. *The Annals of Applied Statistics*, 10(2):1019–1040.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Sun, Y. and Vandenberghe, L. (2015). Decomposition methods for sparse matrix nearness problems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1691–1717.

Xue, L., Ma, S., and Zou, H. (2012). Positive-definite l1-penalized estimation of large covariance matrices. *Journal of the American Statistical Association*, 107(500):1480–1491.

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(4):479–485.