Large Covariance Estimation for Compositional Data via Composition-Adjusted Thresholding

Yuanpei Cao, Wei Lin, and Hongzhe Li

Abstract

High-dimensional compositional data arise naturally in many applications such as metagenomic data analysis. The observed data lie in a high-dimensional simplex, and conventional statistical methods often fail to produce sensible results due to the unit-sum constraint. In this article, we address the problem of covariance estimation for high-dimensional compositional data, and introduce a composition-adjusted thresholding (COAT) method under the assumption that the basis covariance matrix is sparse. Our method is based on a decomposition relating the compositional covariance to the basis covariance, which is approximately identifiable as the dimensionality tends to infinity. The resulting procedure can be viewed as thresholding the sample centered log-ratio covariance matrix and hence is scalable for large covariance matrices. We rigorously characterize the identifiability of the covariance parameters, derive rates of convergence under the spectral norm, and provide theoretical guarantees on support recovery. Simulation studies demonstrate that the COAT estimator outperforms some existing optimization-based estimators. We apply the proposed method to the analysis of a microbiome dataset in order to understand the dependence structure among bacterial taxa in the human gut.

Key words: Adaptive thresholding; Basis covariance; Centered log-ratio covariance; High dimensionality; Microbiome; Regularization.

Yuanpei Cao is Postdoctoral Researcher (E-mail: *yuanpeic@sas.upenn.edu*) and Hongzhe Li is Professor (E-mail: *hongzhe@upenn.edu*), Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104. Wei Lin is Assistant Professor, School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing 100871, China (E-mail: *weilin@math.pku.edu.cn*).

1 Introduction

Compositional data, which represent the proportions or fractions of a whole, arise naturally in a wide range of applications; examples include geochemical compositions of rocks, household patterns of expenditures, species compositions of biological communities, and topic compositions of documents, among many others. This article is particularly motivated by the metagenomic analysis of microbiome data in order to understand the dependence structure among microbial taxa within communities. The human microbiome is the totality of all microbes at various body sites, whose importance in human health and disease has increasingly been recognized. Recent studies have revealed that microbiome composition varies based on diet, health, and the environment (The Human Microbiome Project Consortium 2012), and may play a key role in complex diseases such as obesity, atherosclerosis, and Crohn's disease (Turnbaugh et al. 2009; Koeth et al. 2013; Lewis et al. 2015).

With the development of next-generation sequencing technologies, it is now possible to survey the microbiome composition using direct DNA sequencing of either marker genes or the whole metagenomes. After aligning these sequence reads to the reference microbial genomes, one can quantify the relative abundances of microbial taxa. These sequencing-based microbiome studies, however, only provide a relative, rather than absolute, measure of the abundances of community components. The counts comprising these data (e.g., 16S rRNA gene reads or shotgun metagenomic reads) are set by the amount of genetic material extracted from the community or the sequencing depth, and analysis typically begins by normalizing the observed data by the total number of counts. The resulting fractions thus fall into a class of high-dimensional compositional data that we focus in this article. The high dimensionality refers to the fact that the number of taxa may be comparable to or much larger than the sample size.

To fix the notation, we consider a microbial community with p taxa. Let $\mathbf{W} = (W_1, \ldots, W_p)^T$ with $W_j > 0$ for all j be a vector of latent variables that represent the absolute abundances of the

p taxa, called the *basis*, which generate the observed data via the normalization

$$X_j = \frac{W_j}{\sum_{i=1}^{p} W_i}, \quad j = 1, \ldots, p. \tag{1}$$

The observed data $\mathbf{X} = (X_1, \ldots, X_p)^T$ are *compositional* in the sense that they satisfy the simplex constraint

$$X_j > 0, \quad j = 1, \ldots, p, \quad \sum_{j=1}^{p} X_j = 1.$$

Define also the *basis covariance matrix* $\mathbf{\Omega}_0 = (\omega_{ij}^0)_{p\times p}$ by

$$\omega_{ij}^0 = \mathrm{Cov}(Y_i, Y_j), \tag{2}$$

where $Y_j = \log W_j$. An important question in metagenomic studies is to understand the co-occurrence and co-exclusion relationships between microbial taxa, which would provide valuable insights into the complex ecology of microbial communities (Faust et al. 2012). Ideally, such relationships are described by the basis covariance matrix $\mathbf{\Omega}_0$ and could be easily estimated if the absolute abundances $\mathbf{W}$ were observable. In practice, however, such absolute abundances reflecting the bacterial loads are rarely available. Instead, much recent effort has focused on estimating $\mathbf{\Omega}_0$ based on the relative abundances $\mathbf{X}$ measured through 16S or metagenomic sequencing.

Owing to the difficulties arising from the simplex constraint, it has been a long-standing question how to appropriately model, estimate, and interpret the covariance structure of compositional data. It is well known that standard correlation analysis from the raw proportions can lead to spurious results due to the unit-sum constraint; the proportions tend to be correlated even if the absolute abundances are independent. Such effects are biologically irrelevant and must be removed in an analysis in order to make valid inferences about the underlying biological processes. The compositional effects are further magnified by the low diversity of microbiome data, that is, a few taxa make up the overwhelming majority of the microbiome (Li 2015).

The pioneering work of Aitchison (1982, 2003) introduced several equivalent matrix specifications of compositional covariance structures via the log-ratio transformations. Statistical methods based on these covariance models respect the unique features of compositional data and prove use-

ful in a variety of applications such as geochemical analysis. A potential disadvantage of these models, however, is that they lack a direct interpretation in the usual sense of covariances and correlations; as a result, it is unclear how to impose certain structures such as sparsity in high dimensions, which is crucial for our applications to microbiome data analysis.

A relationship connecting the basis and compositional covariance structures, which is due to Aitchison (2003, Section 4.11), has recently been exploited to develop algorithms for inferring correlation networks from metagenomic data. Friedman and Alm (2012) introduced an approximation approach, SparCC, to infer the basis correlation matrix under certain sparsity assumptions. Their method, however, consists of a series of approximations whose behavior is difficult to analyze. In addition, the estimated covariance matrix is not guaranteed to be positive definite and the estimated correlation coefficients may fall outside the interval $[-1, 1]$. Fang et al. (2015) proposed a CCLasso method that combines a weighted least squares loss with the ℓ_1 penalty to infer the basis correlation matrix. Ban, An, and Jiang (2015) developed a penalized estimation method, REBACCA, to estimate the basis covariance by finding a sparse solution to an underdetermined linear system. While these methods build on similar ideas and seem to work in some practical scenarios, none of the aforementioned work provide theoretical performance guarantees or make explicit the statistical assumptions required for their methods to work effectively. Moreover, all the existing methods involve computationally expensive iterative procedures and do not scale to large p.

Our contributions in this article are to turn the above idea into a principled approach to sparse covariance matrix estimation and provide statistical insights into the issue of identifiability and the impacts of dimensionality. By exploring a decomposition relating the compositional covariance to the basis covariance, we show that the nonidentifiability of the basis covariance vanishes asymptotically as the dimensionality grows under certain sparsity assumptions. In other words, $\mathbf{\Omega}_0$ is approximately identifiable as long as it belongs to a class of large sparse covariance matrices. This somewhat surprising "blessing of dimensionality" allows us to develop a simple, two-step method by first extracting a rank-2 component from the decomposition and then estimating the sparse component $\mathbf{\Omega}_0$ by thresholding the residual matrix. The resulting procedure can equivalently be viewed

as thresholding the sample centered log-ratio covariance matrix, and hence is optimization-free and scalable for large covariance matrices. We call our method *composition-adjusted thresholding* (COAT), which removes the "coat" of compositional effects from the covariance structure. We derive rates of convergence under the spectral norm and provide theoretical guarantees on support recovery. Simulation studies demonstrate that the COAT estimator outperforms some existing optimization-based estimators. We illustrate our method by analyzing a microbiome dataset in order to understand the dependence structure among bacterial taxa in the human gut.

A fast-expanding literature on large covariance estimation can be found for unconstrained high-dimensional data. Bickel and Levina (2008) and El Karoui (2008) introduced regularized estimators by hard thresholding for large covariance matrices that satisfy certain notions of sparsity. Rothman, Levina, and Zhu (2009) considered a wider class of thresholding functions, and Cai and Liu (2011) proposed adaptive thresholding procedures that take into account the variability of individual entries. Fan, Fan, and Lv (2008) and Fan, Liao, and Mincheva (2013) considered large covariance estimation for factor-based models. The standard assumptions made in the literature, however, do not generally hold for constrained data. Our work fills this important gap by adapting covariance thresholding methods to compositional data, a common type of constrained data in many scientific applications.

The rest of the article is organized as follows. Section 2 reviews the covariance relationship and addresses the issue of identifiability. Section 3 introduces the COAT methodology. Section 4 investigates the theoretical properties of the COAT estimator in terms of convergence rates and support recovery. Simulation studies and an application to human gut microbiome data are presented in Sections 5 and 6, respectively. We conclude the article with some discussion in Section 7 and relegate all proofs to the Appendix.

2 Identifiability of the Covariance Model

We first introduce some notation. Denote by $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_F$, and $\|\cdot\|_{\max}$ the matrix L_1-norm, spectral norm, Frobenius norm, and entrywise L_∞-norm, defined for a matrix $\mathbf{A} = (a_{ij})$ by

$\|\mathbf{A}\|_1 = \max_j \sum_i |a_{ij}|$, $\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}(\mathbf{A}^T\mathbf{A})}$, $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$, and $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$, where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue. For two matrices $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$ of the same dimension, define the Frobenius inner product $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{tr}(\mathbf{A}^T\mathbf{B}) = \sum_i \sum_j a_{ij} b_{ij}$.

In the latent variable covariance model (1) and (2), the basis covariance matrix $\mathbf{\Omega}_0$ is the parameter of interest. One of the matrix specifications of compositional covariance structures introduced by Aitchison (2003) is the *variation matrix* $\mathbf{T}_0 = (\tau_{ij}^0)_{p\times p}$ defined by

$$\tau_{ij}^0 = \mathrm{Var}\{\log(X_i/X_j)\}. \tag{3}$$

In view of the relationship (1), we can decompose τ_{ij}^0 as

$$\begin{aligned} \tau_{ij}^0 &= \mathrm{Var}(\log W_i - \log W_j) \\ &= \mathrm{Var}(Y_i) + \mathrm{Var}(Y_j) - 2\,\mathrm{Cov}(Y_i, Y_j) \\ &= \omega_{ii}^0 + \omega_{jj}^0 - 2\omega_{ij}^0, \end{aligned} \tag{4}$$

or in matrix form,

$$\mathbf{T}_0 = \boldsymbol{\omega}_0 \mathbf{1}^T + \mathbf{1}\boldsymbol{\omega}_0^T - 2\mathbf{\Omega}_0, \tag{5}$$

where $\boldsymbol{\omega}_0 = (\omega_{11}^0, \ldots, \omega_{pp}^0)^T$ and $\mathbf{1} = (1, \ldots, 1)^T$. Corresponding to the many-to-one relationship between bases and compositions, the basis covariance matrix $\mathbf{\Omega}_0$ is unidentifiable from the decomposition (5), since $\boldsymbol{\omega}_0 \mathbf{1}^T + \mathbf{1}\boldsymbol{\omega}_0^T$ and $\mathbf{\Omega}_0$ are in general not orthogonal to each other with respect to the Frobenius inner product. In fact, using the *centered log-ratio covariance matrix* $\mathbf{\Gamma}_0 = (\gamma_{ij}^0)_{p\times p}$ (Aitchison 1982) defined by

$$\gamma_{ij}^0 = \mathrm{Cov}\{\log(X_i/g(\mathbf{X})), \log(X_j/g(\mathbf{X}))\},$$

where $g(\mathbf{x}) = (\prod_{j=1}^p x_j)^{1/p}$ is the geometric mean of a vector $\mathbf{x} = (x_1, \ldots, x_p)^T$, we can similarly write

$$\begin{aligned} \tau_{ij}^0 &= \mathrm{Var}\{\log(X_i/g(\mathbf{X})) - \log(X_j/g(\mathbf{X}))\} \\ &= \mathrm{Var}\{\log(X_i/g(\mathbf{X}))\} + \mathrm{Var}\{\log(X_j/g(\mathbf{X}))\} - 2\,\mathrm{Cov}\{\log(X_i/g(\mathbf{X}), \log(X_j/g(\mathbf{X}))\} \end{aligned}$$

$$= \gamma^0_{ii} + \gamma^0_{jj} - 2\gamma^0_{ij},$$

or in matrix form,

$$\mathbf{T}_0 = \boldsymbol{\gamma}_0 \mathbf{1}^T + \mathbf{1}\boldsymbol{\gamma}_0^T - 2\boldsymbol{\Gamma}_0, \tag{6}$$

where $\boldsymbol{\gamma}_0 = (\gamma^0_{11}, \ldots, \gamma^0_{pp})^T$ and $\mathbf{1} = (1, \ldots, 1)^T$. Unlike (5), the following proposition shows that (6) is an orthogonal decomposition and hence the components $\boldsymbol{\gamma}_0 \mathbf{1}^T + \mathbf{1}\boldsymbol{\gamma}_0^T$ and $\boldsymbol{\Gamma}_0$ are identifiable. In addition, by comparing the decompositions (5) and (6), we can bound the difference between $\boldsymbol{\Omega}_0$ and its identifiable counterpart $\boldsymbol{\Gamma}_0$ as follows.

Proposition 1. *The components $\boldsymbol{\gamma}_0 \mathbf{1}^T + \mathbf{1}\boldsymbol{\gamma}_0^T$ and $\boldsymbol{\Gamma}_0$ in the decomposition* (6) *are orthogonal to each other with respect to the Frobenius inner product. Moreover, for the covariance parameters $\boldsymbol{\Omega}_0$ and $\boldsymbol{\Gamma}_0$ in the decompositions* (5) *and* (6)*,*

$$\|\boldsymbol{\Omega}_0 - \boldsymbol{\Gamma}_0\|_{\max} \leq 3p^{-1}\|\boldsymbol{\Omega}_0\|_1.$$

Proposition 1 entails that the covariance parameter $\boldsymbol{\Omega}_0$ is *approximately* identifiable as long as $\|\boldsymbol{\Omega}_0\|_1 = o(p)$. In particular, suppose that $\boldsymbol{\Omega}_0$ belongs to a class of sparse covariance matrices considered by Bickel and Levina (2008),

$$\mathcal{U}(q, s_0(p), M) \equiv \left\{ \boldsymbol{\Omega} \colon \boldsymbol{\Omega} \succ 0, \max_j \omega_{jj} \leq M, \max_i \sum_{j=1}^{p} |\omega_{ij}|^q \leq s_0(p) \right\}, \tag{7}$$

where $0 \leq q < 1$ and $\boldsymbol{\Omega} \succ 0$ denotes that $\boldsymbol{\Omega}$ is positive definite. Then

$$\|\boldsymbol{\Omega}_0\|_1 = \max_i \sum_{j=1}^{p} |\omega^0_{ij}|^{1-q} |\omega^0_{ij}|^q \leq \max_i \sum_{j=1}^{p} (\omega^0_{ii}\omega^0_{jj})^{(1-q)/2} |\omega^0_{ij}|^q \leq M^{1-q} s_0(p),$$

and hence the parameters $\boldsymbol{\Omega}_0$ and $\boldsymbol{\Gamma}_0$ are asymptotically indistinguishable when $s_0(p) = o(p)$. This allows us to use $\boldsymbol{\Gamma}_0$ as a proxy for $\boldsymbol{\Omega}_0$ and greatly facilitates the development of new methodology and associated theory. The intuition behind the approximate identifiability under the sparsity assumption is that the rank-2 component $\boldsymbol{\omega}_0 \mathbf{1}^T + \mathbf{1}\boldsymbol{\omega}_0^T$ represents a global effect that spreads across all rows and columns, while the sparse component $\boldsymbol{\Omega}_0$ represents a local effect that is confined to individual entries.

Also of interest is the *exact* identifiability of $\mathbf{\Omega}_0$ over L_0-balls, which has been studied by Fang et al. (2015) and Ban, An, and Jiang (2015). The following result provides a sufficient and necessary condition for the exact identifiability of $\mathbf{\Omega}_0$ by confining it to an L_0-ball.

Proposition 2. *Suppose that* $\mathbf{\Omega}_0$ *belongs to the* L_0*-ball*

$$\mathcal{B}_0(s_e(p)) \equiv \left\{ \mathbf{\Omega} \colon \sum_{(i,j)\colon i<j} I(\omega_{ij} \neq 0) \leq s_e(p) \right\},$$

where $p \geq 5$*. Then there exist no two values of* $\mathbf{\Omega}_0$ *that correspond to the same* $\mathbf{T}_0$ *in* (5) *if and only if* $s_e(p) < (p-1)/2$*.*

A counterexample is provided in the proof of Proposition 2 to show that the sparsity condition in Fang et al. (2015), which is of the order $O(p^2)$, does not suffice. The identifiability condition in Proposition 2 essentially requires the average degree of the correlation network to be less than 1, which is too restrictive to be useful in practice. This illustrates the importance and necessity of introducing the notion of approximate identifiability.

3 A Sparse Covariance Estimator for Compositional Data

Suppose that $(\mathbf{W}_k, \mathbf{X}_k)$, $k = 1, \ldots, n$, are independent copies of $(\mathbf{W}, \mathbf{X})$, where the compositions $\mathbf{X}_k = (X_{k1}, \ldots, X_{kp})^T$ are observed and the bases $\mathbf{W}_k = (W_{k1}, \ldots, W_{kp})^T$ are latent. In Section 3.1, we rely on the decompositions (5) and (6) and Proposition 1 to develop an estimator of $\mathbf{\Omega}_0$, and in Section 3.2 discuss the selection of the tuning parameter.

3.1 Composition-Adjusted Thresholding

In view of Proposition 1, we wish to estimate the covariance parameter $\mathbf{\Omega}_0$ via the proxy $\mathbf{\Gamma}_0$. To this end, we first construct an empirical estimate of $\mathbf{\Gamma}_0$ and then apply adaptive thresholding to the estimate.

There are two equivalent ways to form the estimate of $\mathbf{\Gamma}_0$. Motivated by the decomposition (6), one can start with the sample counterpart $\widehat{\mathbf{T}} = (\hat{\tau}_{ij})_{p\times p}$ of $\mathbf{T}_0$ defined by

$$\hat{\tau}_{ij} = \frac{1}{n}\sum_{k=1}^{n}(\tau_{kij} - \bar{\tau}_{ij})^2,$$

where $\tau_{kij} = \log(X_{ki}/X_{kj})$ and $\bar{\tau}_{ij} = n^{-1}\sum_{k=1}^{n}\tau_{kij}$. A rank-2 component $\widehat{\boldsymbol{\alpha}}\mathbf{1}^T + \mathbf{1}\widehat{\boldsymbol{\alpha}}^T$ with $\widehat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_p)^T$ can be extracted from the decomposition (6) by projecting $\widehat{\mathbf{T}}$ onto the subspace $\mathcal{A} \equiv \{\boldsymbol{\alpha}\mathbf{1}^T + \mathbf{1}\boldsymbol{\alpha}^T : \boldsymbol{\alpha} \in \mathbb{R}^p\}$, which is given by

$$\hat{\alpha}_i = \hat{\tau}_{i\cdot} - \frac{1}{2}\hat{\tau}_{\cdot\cdot},$$

where $\hat{\tau}_{i\cdot} = p^{-1}\sum_{j=1}^{p}\hat{\tau}_{ij}$ and $\hat{\tau}_{\cdot\cdot} = p^{-2}\sum_{i,j=1}^{p}\hat{\tau}_{ij}$. The residual matrix $\widehat{\mathbf{\Gamma}} = -(\widehat{\mathbf{T}} - \widehat{\boldsymbol{\alpha}}\mathbf{1}^T - \mathbf{1}\widehat{\boldsymbol{\alpha}}^T)/2$, with entries

$$\hat{\gamma}_{ij} = -\frac{1}{2}(\hat{\tau}_{ij} - \hat{\alpha}_i - \hat{\alpha}_j) = -\frac{1}{2}(\hat{\tau}_{ij} - \hat{\tau}_{i\cdot} - \hat{\tau}_{j\cdot} + \hat{\tau}_{\cdot\cdot}),$$

is then an estimate of $\mathbf{\Gamma}_0$. Alternatively, $\widehat{\mathbf{\Gamma}}$ can be obtained directly as the sample counterpart of $\mathbf{\Gamma}_0$ through the expression

$$\hat{\gamma}_{ij} = \frac{1}{n}\sum_{k=1}^{n}(\gamma_{ki} - \bar{\gamma}_i)(\gamma_{kj} - \bar{\gamma}_j), \tag{8}$$

where $\gamma_{kj} = \log(X_{kj}/g(\mathbf{X}_k))$ and $\bar{\gamma}_j = n^{-1}\sum_{k=1}^{n}\gamma_{kj}$.

Now applying adaptive thresholding to $\widehat{\mathbf{\Gamma}}$, we define the *composition-adjusted thresholding* (COAT) estimator

$$\widehat{\mathbf{\Omega}} = (\hat{\omega}_{ij})_{p\times p} \quad \text{with } \hat{\omega}_{ij} = S_{\lambda_{ij}}(\hat{\gamma}_{ij}), \tag{9}$$

where $S_\lambda(\cdot)$ is a general thresholding function and $\lambda_{ij} > 0$ are entry-dependent thresholds.

In this article, we consider a class of general thresholding functions $S_\lambda(\cdot)$ that satisfy the following conditions:

(i) $S_\lambda(z) = 0$ for $|z| \le \lambda$;

(ii) $|S_\lambda(z) - z| \le \lambda$ for all $z \in \mathbb{R}$.

These two conditions were assumed by Rothman, Levina, and Zhu (2009) and Cai and Liu (2011) along with another condition that is not required in our analysis. Examples of thresholding func-

tions belonging to this class include the hard thresholding rule $S_\lambda(z) = zI(|z| \geq \lambda)$, the soft thresholding rule $S_\lambda(z) = \mathrm{sgn}(z)(|z| - \lambda)_+$, and the adaptive lasso rule $S_\lambda(z) = z(1 - |\lambda/z|^\eta)_+$ for $\eta \geq 1$.

The performance of the COAT estimator depends critically on the choice of thresholds. Using entry-adaptive thresholds may in general improve the performance over applying a universal threshold. To derive a data-driven choice of λ_{ij}, define

$$\theta_{ij} = \mathrm{Var}\{(Y_i - \mu_i)(Y_j - \mu_j)\},$$

where $\mu_j = EY_j$. We take λ_{ij} to be of the form

$$\lambda_{ij} = \delta\sqrt{\hat{\theta}_{ij}}, \tag{10}$$

where $\hat{\theta}_{ij}$ are estimates of θ_{ij}, and $\delta > 0$ is a tuning parameter to be chosen, for example, by cross-validation. We rewrite (8) as $\hat{\gamma}_{ij} = n^{-1}\sum_{k=1}^n \gamma_{kij}$, where $\gamma_{kij} = (\gamma_{ki} - \bar{\gamma}_i)(\gamma_{kj} - \bar{\gamma}_j)$. Then θ_{ij} can be estimated by

$$\hat{\theta}_{ij} = \frac{1}{n}\sum_{k=1}^n (\gamma_{kij} - \hat{\gamma}_{ij})^2. \tag{11}$$

The resulting algorithm for computing the COAT estimator with a fixed δ is summarized as follows:

1. Compute the sample centered log-ratio covariance matrix $\widehat{\mathbf{\Gamma}} = (\hat{\gamma}_{ij})_{p\times p}$ using (8).
2. Compute the variance estimates $\hat{\theta}_{ij}$, $i, j = 1, \ldots, p$, using (11).
3. Obtain the COAT estimator $\widehat{\mathbf{\Omega}}$ using (9) with λ_{ij} defined by (10).

We close this subsection with a remark on the related work of Fang et al. (2015). Their method also computes the sample centered log-ratio covariance matrix $\widehat{\mathbf{\Gamma}}$ and relies on its closeness to its population counterpart $\mathbf{\Gamma}_0 = \mathbf{G}\mathbf{\Omega}_0\mathbf{G}$, where $\mathbf{G} = \mathbf{I}_p - p^{-1}\mathbf{1}_p\mathbf{1}_p^T$. This then leads to a weighted least squares loss combined with an ℓ_1 penalization on $\mathbf{\Omega}_0$. By contrast, our method exploits directly the similarity of $\mathbf{\Gamma}_0$ and $\mathbf{\Omega}_0$ provided by Proposition 1, thereby avoiding any optimization procedure.

3.2 Tuning Parameter Selection

The thresholds defined by (10) depend on the tuning parameter δ, which can be chosen through V-fold cross-validation. Denote by $\widehat{\mathbf{\Omega}}^{(-v)}(\delta)$ the COAT estimate based on the training data excluding the vth fold, and $\widehat{\mathbf{\Gamma}}^{(v)}$ the residual matrix (or the sample centered log-ratio covariance matrix) based on the test data including only the vth fold. We choose the optimal value $\hat{\delta}$ of δ that minimizes the cross-validation error

$$\mathrm{CV}(\delta) = \frac{1}{V}\sum_{v=1}^{V} \|\widehat{\mathbf{\Omega}}^{(-v)}(\delta) - \widehat{\mathbf{\Gamma}}^{(v)}\|_F^2.$$

With the optimal $\hat{\delta}$, we then compute the COAT estimate based on the whole dataset as our final estimate. When the positive definiteness of the covariance estimate in finite samples is required for interpretation, we follow the approach of Fan, Liao, and Mincheva (2013) and choose δ in the range where the minimum eigenvalue of the COAT estimate is positive.

4 Theoretical Properties

In this section, we investigate the asymptotic properties of the COAT estimator. As a distinguishing feature of our theoretical analysis, we assume neither the exact identifiability of the parameters nor that the degree of (approximate) identifiability is dominated by the statistical error. Instead, the degree of identifiability enters our analysis and shows up in the resulting rate of convergence. Such theoretical analysis is rare in the literature, but is extremely relevant for latent variable models in the presence of nonidentifiability and is of theoretical interest in its own right. We introduce our assumptions in Section 4.1, and present our main results on rates of convergence and support recovery in Section 4.2.

4.1 Assumptions

Recall that $Y_j = \log W_j$, $\mu_j = EY_j$, and $\theta_{ij} = \mathrm{Var}\{(Y_i - \mu_i)(Y_j - \mu_j)\}$, and define $Y_{kj} = \log W_{kj}$. Without loss of generality, assume $\mu_j = 0$ for all j throughout this section. We need to impose the following moment conditions on the log-basis $\mathbf{Y} = (Y_1, \ldots, Y_p)^T$.

Condition 1. There exists a constant $\alpha > 0$ such that $\max_j E \exp(\alpha Y_j^2) \leq 2$.

Condition 2. The basis covariance matrix $\mathbf{\Omega}_0$ belongs to the class $\mathcal{U}(q, s_0(p), M)$ defined by (7), where $0 \leq q < 1$, $s_0(p) = o(p)$, and $\log p = o(n^{1/5})$.

Condition 3. There exists a constant $\tau > 0$ such that $\min_{i,j} \theta_{ij} \geq \tau$.

Condition 4. There exists a constant $s_1(p)$ depending on p such that

$$\max_{i,j,\ell} \left| \sum_{m=1}^{p} E Y_i Y_j Y_\ell Y_m \right| \leq s_1(p)$$

and $s_1(p) = o(p)$.

Conditions 1–3 are similar to those commonly assumed in the covariance estimation literature; see, for example, Cai and Liu (2011). Condition 1 requires that the log-basis variables be uniformly sub-Gaussian, which is satisfied if $\mathbf{Y}$ is multivariate normal or bounded, and can be relaxed to a sub-exponential or polynomial tail condition at the price of a technically more involved argument. Condition 2 imposes some restrictions on the dimensionality and sparsity of the basis covariance matrix $\mathbf{\Omega}_0$. It is worth mentioning that the sparsity level condition $s_0(p) = o(p)$ is so weak that it suffices to guarantee only approximate identifiability but allows the degree of nonidentifiability to be large relative to the statistical error. Condition 3 is a technical assumption for adaptive thresholding procedures. Condition 4 arises from identifiability considerations in estimating the variances θ_{ij} but is very mild. In fact, if $\mathbf{Y}$ is multivariate normal, then Condition 4 is implied by the assumptions $\mathbf{\Omega}_0 \in \mathcal{U}(q, s_0(p), M)$ and $s_0(p) = o(p)$ in Condition 2, since from Isserlis' theorem (Isserlis 1918) we have

$$\max_{i,j,\ell} \left| \sum_{m=1}^{p} E Y_i Y_j Y_\ell Y_m \right| \leq \max_{i,j,\ell} \sum_{m=1}^{p} \left(|\omega_{ij}^0||\omega_{\ell m}^0| + |\omega_{i\ell}^0||\omega_{jm}^0| + |\omega_{im}^0||\omega_{j\ell}^0| \right) \leq 3M^{2-q} s_0(p).$$

More generally, under Condition 2, Condition 4 also holds for elliptically contoured distributions, since

$$E Y_i Y_j Y_\ell Y_m = \kappa(\omega_{ij}^0 \omega_{\ell m}^0 + \omega_{i\ell}^0 \omega_{jm}^0 + \omega_{im}^0 \omega_{j\ell}^0),$$

where $\kappa = \frac{1}{3}EY_i^4/(\omega_{ii}^0)^2$ (Anderson 2003). Although the abundances of bacterial populations in many applications have been found to follow log-normal distributions (Limpert, Stahel, and Abbt 2001), we will show in simulation studies that the performance of our method is not sensitive to the distributions of the absolute abundances.

4.2 Main Results

We are now in a position to state our main results. The following theorem gives the rate of convergence under the spectral norm for the COAT estimator.

Theorem 1 (Rate of convergence)**.** *Under Conditions 1–4, if the tuning parameter δ in* (10) *is chosen to be*

$$\delta = C_1\sqrt{\frac{\log p}{n}} + C_2\frac{s_0(p)}{p} \tag{12}$$

for sufficiently large $C_1, C_2 > 0$, then the COAT estimator $\widehat{\mathbf{\Omega}}$ in (9) *satisfies*

$$\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}_0\|_2 = O_p\left\{s_0(p)\left(\sqrt{\frac{\log p}{n}} + \frac{s_0(p)}{p}\right)^{1-q}\right\}$$

uniformly on $\mathcal{U}(q, s_0(p), M)$.

The rate of convergence provided by Theorem 1 exhibits an interesting decomposition: the term $s_0(p)\{(\log p)/n\}^{(1-q)/2}$ represents the estimation error due to estimating $\mathbf{\Gamma}_0$, while the term $s_0(p)(s_0(p)/p)^{1-q}$ accounts for the approximation error due to using $\mathbf{\Gamma}_0$ as a proxy for $\mathbf{\Omega}_0$. In particular, if the approximation error is dominated by the estimation error, then the COAT estimator attains the minimax optimal rate under the spectral norm over $\mathcal{U}(q, s_0(p), M)$ (Cai and Zhou 2012). It is important to note that the dimensionality p appears in both terms where it plays opposite roles. We observe a "curse of dimensionality" in the first term, where the growth of dimensionality contributes a logarithmic factor to the estimation error. By contrast, a "blessing of dimensionality" is reflected by the second term in that a diverging dimensionality shrinks the approximation error toward zero at a power rate.

The insights gained from Theorem 1 have important implications for compositional data analysis. In the analysis of many compositional datasets, the dimensionality often depends on the

taxonomic level to be examined. For instance, in metagenomic studies, the dimensionality may range from only a few taxa at the phylum level to thousands of taxa at the operational taxonomic unit (OTU) level. Suppose, for simplicity, that the magnitudes of correlation signals are of about the same order across different taxonomic levels. Then Theorem 1 indicates a tradeoff between an accurate estimation of the covariance structure with low dimensionality and a sensible interpretation in terms of the basis components with high dimensionality. This tradeoff thus suggests the need to analyze compositional data at relatively finer taxonomic levels when a latent variable interpretation is desired.

The proof of Theorem 1 relies on a series of concentration inequalities that take the approximation error term into account, which can be found in the Appendix. As a consequence of these inequalities, we obtain the following result regarding the support recovery property of the COAT estimator. Here the support of $\mathbf{\Omega}_0$ refers to the set of all indices (i, j) with $\omega_{ij}^0 \neq 0$.

Theorem 2 (Support recovery)**.** *Under Conditions 1–4, if the tuning parameter* δ *in* (10) *is chosen as in* (12)*, then the COAT estimator* $\widehat{\mathbf{\Omega}}$ *in* (9) *satisfies*

$$P\left(\hat{\omega}_{ij} = 0 \textit{ for all } (i,j) \textit{ with } \omega_{ij}^0 = 0\right) \to 1. \tag{13}$$

Moreover, if in addition

$$\min_{(i,j):\, \omega_{ij}^0 \neq 0} |\omega_{ij}^0| / \sqrt{\theta_{ij}} \geq C\delta \tag{14}$$

for some constant $C > 3/2$*, then*

$$P\left(\operatorname{sgn}(\hat{\omega}_{ij}) = \operatorname{sgn}(\omega_{ij}^0) \textit{ for all } (i,j)\right) \to 1. \tag{15}$$

Theorem 2 parallels the support recovery results in Rothman, Levina, and Zhu (2009) and Cai and Liu (2011). However, owing to the extra term $s_0(p)/p$ in the expression of δ, the assumption (14) requires in addition that no correlation signals fall below the approximation error. In other words, exact support recovery will break down if any correlation signal is confounded by the compositional effect.

We next turn to the estimator with a data-driven choice of δ as described in Section 3.2. Denote by $\widehat{\mathbf{\Omega}}(\hat{\delta})$ the COAT estimator based on the optimal $\hat{\delta}$ chosen by V-fold cross-validation. For simplicity, assume that $\hat{\delta}$ is chosen from a grid of J points $\delta_j = \delta_0 j/J$, $j = 1, \ldots, J$, where J is fixed and δ_0 is of the form (12) for sufficiently large $C_1, C_2 > 0$. The following result provides the rate of convergence under the Frobenius norm for $\widehat{\mathbf{\Omega}}(\hat{\delta})$, which coincides with that of its theoretical counterpart $\widehat{\mathbf{\Omega}}$ and is minimax optimal (Cai and Zhou 2012).

Theorem 3 (Data-driven choice of δ)**.** *Under Conditions 2 and 3, if* $\mathbf{Y}$ *is multivariate normal and*

$$s_0(p) = O\left\{ p \left(\frac{\log p}{n} \right)^{1-q/2} \right\}, \tag{16}$$

then the COAT estimator $\widehat{\mathbf{\Omega}}(\hat{\delta})$ *with* $\hat{\delta}$ *chosen by* V*-fold cross-validation satisfies*

$$\|\widehat{\mathbf{\Omega}}(\hat{\delta}) - \mathbf{\Omega}_0\|_F^2 = O_p\left\{ p s_0(p) \left(\frac{\log p}{n} \right)^{1-q/2} \right\}$$

uniformly on $\mathcal{U}(q, s_0(p), M)$.

By using a more involved argument, it would be possible to extend the above result to the case of diverging J and rate of convergence under the spectral norm, which we do not pursue further.

5 Simulation Studies

We conducted simulation studies to compare the numerical performance of the COAT estimator with that of the oracle estimator, which assumes that the basis components are observed and the thresholding procedure is applied to the sample covariance matrix of the log-basis $\mathbf{Y}$. For simplicity, soft thresholding is used with both estimators. We also include in our comparison the CCLasso (Fang et al. 2015) and REBACCA (Ban, An, and Jiang 2015) estimators. Since our goal is to estimate the basis covariance, the performance of various estimators should be compared with that of the oracle estimator which estimates $\mathbf{\Omega}_0$ directly using the basis components.

5.1 Simulation Settings

The data $(\mathbf{W}_k, \mathbf{X}_k)$, $k = 1, \ldots, n$, were generated as follows. We first generated the log-basis vectors $\mathbf{Y}_k$ in two different ways:

(i) $\mathbf{Y}_k$ are independent from the multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Omega}_0)$;

(ii) $\mathbf{Y}_k = \boldsymbol{\mu} + \mathbf{F}\mathbf{U}_k/\sqrt{10}$, where $\mathbf{F}\mathbf{F}^T = \boldsymbol{\Omega}_0$ and the components of $\mathbf{U}_k$ are independent gamma variables with shape parameter 10 and scale parameter 1, so that $\mathrm{Var}(\mathbf{Y}_k) = \boldsymbol{\Omega}_0$. Here the matrix $\mathbf{F}$ is obtained by computing the singular value decomposition $\boldsymbol{\Omega}_0 = \mathbf{Q}\mathbf{S}\mathbf{Q}^T$ and letting $\mathbf{F} = \mathbf{Q}\mathbf{S}^{1/2}$.

Then $\mathbf{W}_k = (W_{k1}, \ldots, W_{kp})^T$ and $\mathbf{X}_k = (X_{k1}, \ldots, X_{kp})^T$ were obtained through the transformations $W_{kj} = e^{Y_{kj}}$ and $X_{kj} = W_{kj}/\sum_{i=1}^{p} W_{ki}$, $j = 1, \ldots, p$. Hence, in Case (i), $\mathbf{W}_k$ and $\mathbf{X}_k$ follow multivariate log-normal and logistic-normal distributions (Aitchison and Shen 1980), respectively; the distributions of $\mathbf{W}_k$ and $\mathbf{X}_k$ in Case (ii) can similarly be viewed as a type of multivariate log-gamma and logistic-gamma distributions. In both cases, we took the the components of $\boldsymbol{\mu}$ randomly from the uniform distribution on $[0, 10]$, in order to reflect the fact that compositional data arising from metagenomic studies are often heterogeneous.

The following four models for the basis covariance matrix $\boldsymbol{\Omega}_0$ were considered:

- Model 1 (Identity covariance): $\boldsymbol{\Omega}_0 = \mathbf{I}_p$.
- Model 2 (Hub covariance): The p points were randomly divided into 3 hubs and $p - 3$ non-hub points. Each hub was connected to the other points with probability 0.7 while each pair of non-hub points were connected with probability 0.2. The strength of each edge was set to 0.3 with probability 0.5 and -0.3 with probability 0.5. The diagonal entries were set large enough so that $\boldsymbol{\Omega}_0$ is positive definite.
- Model 3 (Block covariance): The p points were equally divided into 10 blocks. Each pair of points in the same block were connected with probability 0.5 while each pair between different blocks were connected with probability 0.2. The strengths of the off-diagonal and diagonal entries were set as in Model 2.
- Model 4 (Sparse covariance): $\boldsymbol{\Omega}_0 = \mathrm{diag}(\mathbf{A}_1, \mathbf{A}_2)$, where $\mathbf{A}_1 = \mathbf{B} + \varepsilon\mathbf{I}_{p_1}$, $\mathbf{A}_2 = 4\mathbf{I}_{p_2}$, $p_1 = \lfloor 3\sqrt{p} \rfloor$, $p_2 = p - p_1$, and $\mathbf{B}$ is a symmetric matrix whose lower triangular entries are independent from the uniform distribution on $[-1, -0.5] \cup [0.5, 1]$ with probability 0.3 and

equal to 0 with probability 0.7. We set $\varepsilon = \max(-\lambda_{\min}(\mathbf{B}), 0) + 0.01$ to ensure that $\mathbf{A}_1$ is positive definite, where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue.

To facilitate comparisons with the CCLasso and REBACCA procedures which output only the correlation estimates, we further normalize $\mathbf{\Omega}_0$ in Models 2–4 to correlation matrices. Model 1 is an extreme but illustrative case intended for demonstrating spurious correlations when different transformations of the compositional data are applied. The settings of Model 2 and 3 are similar to those in Fang et al. (2015). The setting of Model 4 is typical in the covariance estimation literature (e.g., Cai and Liu 2011). In Sections 5.2 and 5.3, we set the sample size $n = 200$ and the dimension $p = 50$, 100, and 200, while in Section 5.4, we set $p = 500$ and $n = 250$, 500, and 1000. We repeated 100 simulations for each setting.

5.2 Spurious Correlations

The boxplots of sample correlations with simulated data under different transformations in Model 1 are shown in Figure 1. Clearly, the sample centered log-ratio (clr) correlations are centered around zero and have a similar distribution to that of the sample correlations of $\mathbf{Y}$. Such a similarity tends to increase as the dimension p grows, which is consistent with Proposition 1 and provides numerical evidence for the validity of the centered log-ratio covariance matrix $\mathbf{\Gamma}_0$ as a proxy for $\mathbf{\Omega}_0$. In fact, from the proof of Proposition 1 we have, when $\mathbf{\Omega}_0 = \mathbf{I}_p$,

$$\|\mathbf{\Omega}_0 - \mathbf{\Gamma}_0\|_{\max} = \max_{i,j} |\omega^0_{i\cdot} + \omega^0_{j\cdot} - \omega^0_{\cdot\cdot}| = p^{-1}.$$

By contrast, spurious correlations are generally observed when $\log \mathbf{X}$ or $\mathbf{X}$ are used to calculate the correlations between taxa. The sample correlations of $\log \mathbf{X}$ exhibit a severe upward bias, while the sample correlations of $\mathbf{X}$ contain many outliers that would be detected as signals by a thresholding procedure with a threshold level set close to 1. Moreover, the spurious correlations seem to become worse with a gamma log-basis distribution, where the means of the compositional components tend to be more heterogeneous.

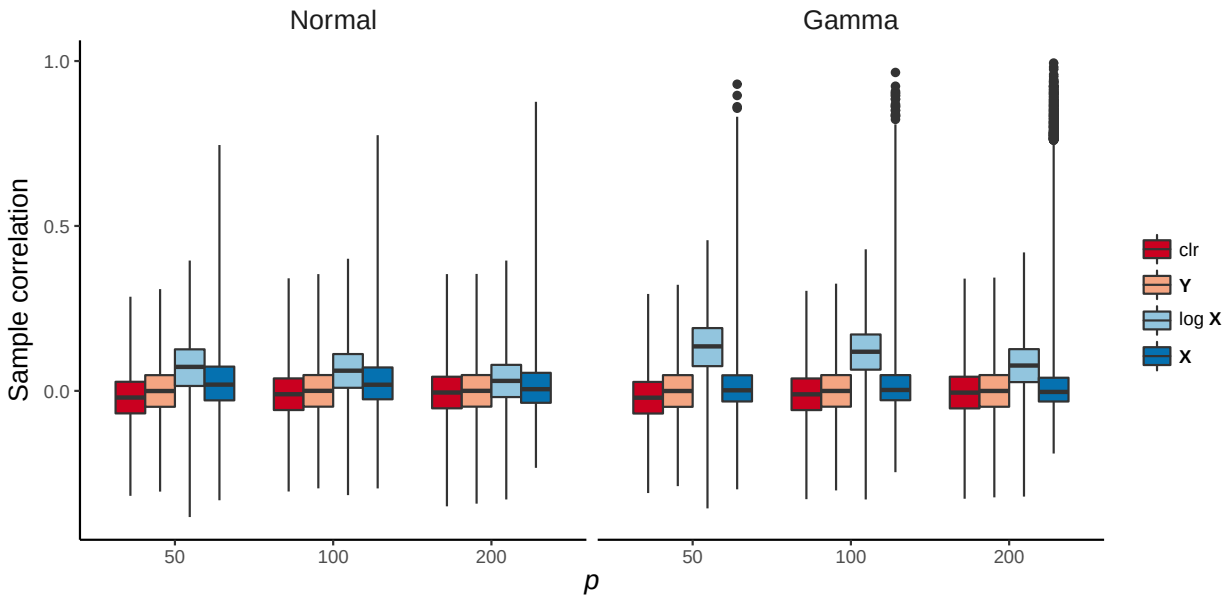


Figure 1: Boxplots of sample correlations based on simulated data under different transformations in Model 1.

5.3 Performance Comparisons

We applied the COAT method with soft thresholding to the simulated data in Models 2–4, where the tuning parameter δ was chosen by fivefold cross-validation. For comparison, we applied the CCLasso and REBACCA methods to estimate the basis correlation matrix. Losses under the matrix L_1-norm, spectral norm, and Frobenius norm were used to evaluate the estimation performance, the true positive rate and false positive rate were used to assess the performance of support recovery, and the run time was taken as a measure of computational efficiency.

The simulation results for Models 2–4 with normal and gamma log-basis distributions are summarized in Tables 1–3, respectively. For all three models, we observe that COAT performs almost equally well as the oracle estimator and uniformly outperforms the other two competitors. Moreover, the performance of COAT gets closer to that of the oracle estimator as p grows. In particular, CCLasso estimates are sensitive to the log-basis distributions in Models 2 and 3, and its estimation losses are much larger when the log-basis follows a gamma distribution. Both CCLasso and REBACCA show inferior performance in terms of support recovery in some scenarios, indicat-

ing that they are not generally model selection consistent. Comparisons on run time suggest that COAT is scalable and computationally the most efficient, whereas the run time of both CCLasso and REBACCA increases dramatically as p grows.

When $p \geq n$, the computational complexity of different methods is as follows: CCLasso is in fact an alternating direction method of multipliers (ADMM) and has a per-iteration complexity of $O(p^3)$. REBACCA is essentially a Lasso problem with $p(p-1)/2$ variables and, if implemented with an efficient ADMM algorithm, requires a complexity of $O(np^4)$ for the initial singular value decomposition and a per-iteration complexity of $O(np^2)$. By contrast, COAT is a direct method and the dominating cost is computing the sample covariance matrix and $\hat{\theta}_{ij}$, resulting in a complexity of only $O(np^2)$. This complexity analysis is consistent with the run time reported in Tables 1–3.

To further compare the support recovery performance without selecting a threshold level, we plot the receiver operating characteristic (ROC) curves for all methods in Models 2–4 with normal and gamma log-basis distributions in Figure 2 and 3, respectively. We observe that the ROC curves for the COAT and oracle methods are almost indistinguishable and uniformly dominate those for CCLasso and REBACCA, demonstrating the superiority of the COAT method. In addition, CCLasso tends to outperform REBACCA when the log-basis follows a gamma distribution, but its performance deteriorates as p grows with a normal log-basis distribution.

Two plausible explanations for the outstanding performance of COAT are that (1) it comes with theoretical guarantees, and (2) its computation is straightforward and introduces less computational error. Moreover, its superiority does not seem to depend on specific model settings.

5.4 Performance in High-Dimensional Settings

To investigate the effects of high dimensionality and varying sample sizes, we applied the COAT and oracle methods with soft thresholding to the simulated data in Models 2–4 with $p = 500$ and $n = 250$, 500, and 1000. CCLasso and REBACCA were not included for comparison owing to their prohibitive computational costs. Tables 4 and 5 summarize the performance of the COAT and oracle methods with normal and gamma log-basis distributions, respectively. When the dimension-

Table 1: Comparisons of means (standard errors) of performance measures for four different methods in Model 2 (hub covariance) based on 100 replications.

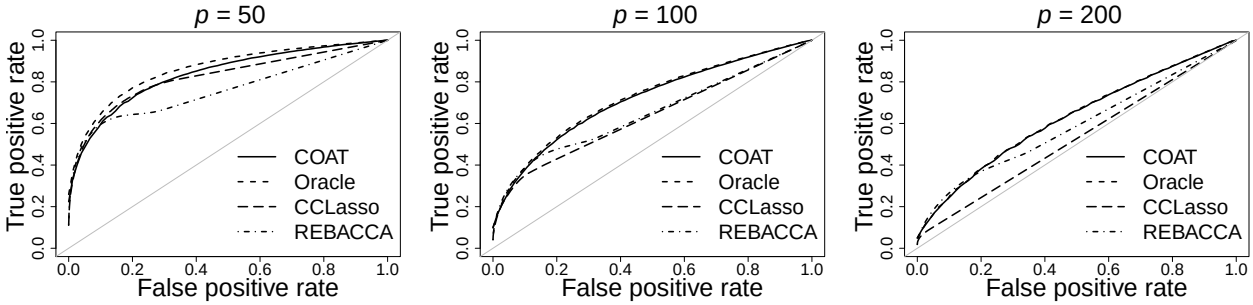
	Normal				Gamma			
p	COAT	Oracle	CCLasso	REBACCA	COAT	Oracle	CCLasso	REBACCA
				Matrix L_1-norm loss				
50	3.14 (0.26)	3.11 (0.26)	3.19 (0.27)	3.64 (0.35)	3.17 (0.21)	3.13 (0.20)	3.29 (0.28)	3.59 (0.38)
100	5.83 (0.22)	5.83 (0.21)	6.43 (0.32)	6.45 (0.36)	5.83 (0.24)	5.82 (0.23)	5.49 (0.48)	6.53 (0.36)
200	9.28 (0.14)	9.30 (0.14)	9.69 (0.21)	11.30 (0.43)	9.27 (0.15)	9.30 (0.14)	10.13 (0.41)	11.32 (0.47)
				Spectral norm loss				
50	0.83 (0.04)	0.79 (0.05)	1.04 (0.09)	1.00 (0.06)	0.82 (0.04)	0.79 (0.05)	1.07 (0.13)	0.99 (0.06)
100	0.92 (0.02)	0.91 (0.02)	1.15 (0.08)	1.36 (0.07)	0.92 (0.03)	0.91 (0.03)	1.22 (0.19)	1.35 (0.06)
200	0.99 (0.01)	1.00 (0.01)	1.08 (0.06)	2.02 (0.06)	1.00 (0.01)	1.00 (0.01)	2.01 (0.11)	2.04 (0.08)
				Frobenius norm loss				
50	2.55 (0.06)	2.46 (0.06)	3.23 (0.12)	3.30 (0.10)	2.55 (0.07)	2.47 (0.08)	3.26 (0.22)	3.30 (0.10)
100	4.14 (0.04)	4.13 (0.06)	5.35 (0.24)	6.14 (0.08)	4.14 (0.04)	4.13 (0.05)	5.27 (0.47)	6.16 (0.09)
200	6.32 (0.03)	6.33 (0.03)	6.97 (0.45)	11.67 (0.13)	6.32 (0.03)	6.33 (0.03)	10.31 (0.32)	11.71 (0.11)
				True positive rate				
50	0.81 (0.03)	0.82 (0.03)	0.80 (0.04)	0.65 (0.03)	0.80 (0.04)	0.81 (0.04)	0.94 (0.03)	0.65 (0.03)
100	0.42 (0.04)	0.41 (0.04)	0.35 (0.08)	0.52 (0.02)	0.42 (0.04)	0.41 (0.04)	0.71 (0.13)	0.52 (0.02)
200	0.10 (0.03)	0.10 (0.03)	0.07 (0.03)	0.46 (0.01)	0.10 (0.03)	0.10 (0.03)	0.87 (0.02)	0.46 (0.01)
				False positive rate				
50	0.30 (0.04)	0.25 (0.03)	0.29 (0.05)	0.27 (0.02)	0.29 (0.04)	0.24 (0.03)	0.70 (0.09)	0.27 (0.02)
100	0.11 (0.02)	0.10 (0.02)	0.10 (0.04)	0.31 (0.02)	0.10 (0.02)	0.10 (0.02)	0.46 (0.20)	0.31 (0.01)
200	0.02 (0.01)	0.02 (0.01)	0.01 (0.01)	0.34 (0.01)	0.02 (0.01)	0.02 (0.01)	0.78 (0.03)	0.34 (0.01)
				Run time (seconds)				
50	0.16 (0.02)	0.16 (0.01)	1.11 (0.07)	1.41 (0.06)	0.18 (0.03)	0.17 (0.03)	1.42 (0.19)	1.45 (0.16)
100	0.48 (0.02)	0.50 (0.04)	5.45 (0.37)	25.38 (1.86)	0.48 (0.02)	0.48 (0.02)	8.90 (1.54)	25.55 (1.55)
200	2.87 (0.12)	3.19 (0.26)	9.29 (0.85)	770.76 (38.34)	2.00 (0.09)	2.00 (0.09)	41.96 (6.53)	1055.51 (64.37)

Table 2: Comparisons of means (standard errors) of performance measures for four different methods in Model 3 (block covariance) based on 100 replications.

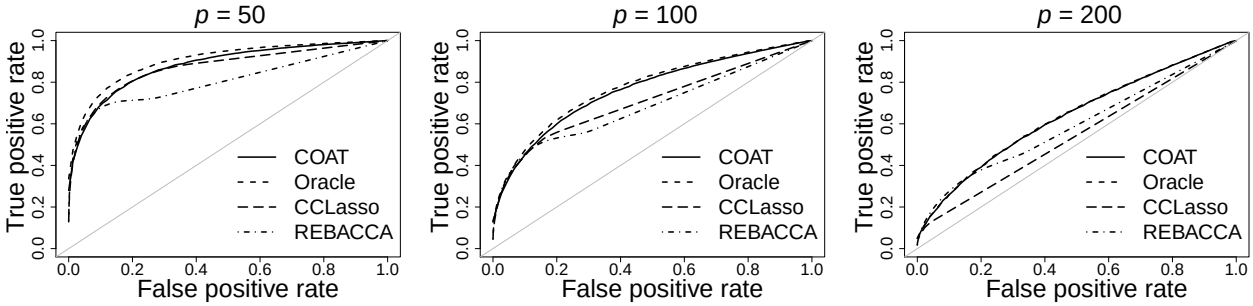
	Normal				Gamma			
p	COAT	Oracle	CCLasso	REBACCA	COAT	Oracle	CCLasso	REBACCA
				Matrix L_1-norm loss				
50	2.13 (0.16)	1.97 (0.16)	2.88 (0.24)	2.70 (0.23)	2.15 (0.17)	1.97 (0.16)	3.43 (0.28)	2.74 (0.26)
100	3.21 (0.13)	3.16 (0.14)	4.60 (0.36)	5.02 (0.26)	3.22 (0.15)	3.17 (0.15)	5.65 (0.74)	4.99 (0.27)
200	4.56 (0.13)	4.56 (0.13)	5.37 (0.40)	9.15 (0.36)	4.58 (0.11)	4.58 (0.11)	10.05 (0.37)	9.17 (0.37)
				Spectral norm loss				
50	0.81 (0.04)	0.74 (0.05)	1.07 (0.09)	0.98 (0.07)	0.81 (0.04)	0.75 (0.05)	1.15 (0.13)	0.99 (0.06)
100	0.96 (0.03)	0.94 (0.02)	1.32 (0.12)	1.38 (0.06)	0.95 (0.03)	0.94 (0.03)	1.45 (0.21)	1.38 (0.07)
200	0.97 (0.01)	0.97 (0.01)	1.17 (0.08)	2.05 (0.07)	0.98 (0.01)	0.98 (0.01)	1.98 (0.11)	2.06 (0.08)
				Frobenius norm loss				
50	2.61 (0.06)	2.50 (0.07)	3.27 (0.12)	3.30 (0.10)	2.61 (0.07)	2.49 (0.08)	3.41 (0.21)	3.32 (0.09)
100	4.54 (0.04)	4.50 (0.04)	5.93 (0.15)	6.31 (0.07)	4.54 (0.05)	4.50 (0.06)	6.00 (0.47)	6.34 (0.09)
200	6.97 (0.03)	6.97 (0.03)	8.18 (0.45)	11.88 (0.10)	6.97 (0.03)	6.98 (0.03)	10.34 (0.31)	11.91 (0.10)
				True positive rate				
50	0.89 (0.02)	0.90 (0.02)	0.88 (0.03)	0.72 (0.03)	0.88 (0.03)	0.90 (0.02)	0.97 (0.02)	0.71 (0.03)
100	0.58 (0.03)	0.58 (0.03)	0.54 (0.04)	0.56 (0.02)	0.58 (0.03)	0.58 (0.03)	0.70 (0.05)	0.55 (0.02)
200	0.15 (0.03)	0.15 (0.02)	0.11 (0.04)	0.46 (0.01)	0.15 (0.02)	0.15 (0.02)	0.87 (0.02)	0.46 (0.01)
				False positive rate				
50	0.32 (0.04)	0.28 (0.03)	0.34 (0.05)	0.26 (0.02)	0.32 (0.04)	0.28 (0.03)	0.73 (0.07)	0.26 (0.02)
100	0.16 (0.02)	0.15 (0.02)	0.16 (0.03)	0.29 (0.01)	0.16 (0.02)	0.15 (0.02)	0.35 (0.07)	0.30 (0.01)
200	0.03 (0.01)	0.03 (0.01)	0.03 (0.01)	0.33 (0.01)	0.03 (0.01)	0.03 (0.01)	0.78 (0.03)	0.33 (0.01)
				Run time (seconds)				
50	0.10 (0.01)	0.10 (0.01)	0.52 (0.03)	0.94 (0.03)	0.10 (0.02)	0.10 (0.01)	0.62 (0.03)	0.93 (0.03)
100	0.33 (0.03)	0.33 (0.01)	3.51 (0.18)	19.25 (0.74)	0.42 (0.04)	0.41 (0.02)	6.52 (0.33)	19.53 (0.93)
200	1.26 (0.03)	1.19 (0.02)	20.64 (1.21)	777.96 (40.39)	1.34 (0.09)	1.31 (0.15)	32.64 (5.37)	954.52 (57.16)

Table 3: Comparisons of means (standard errors) of performance measures for four different methods in Model 4 (sparse covariance) based on 100 replications.

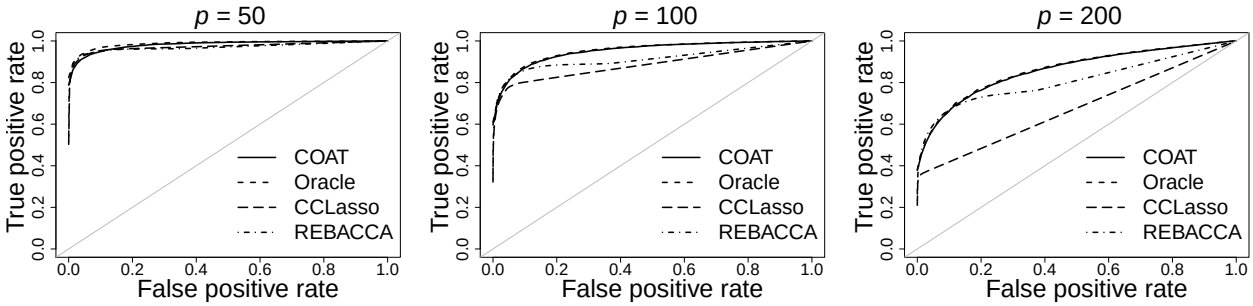
	Normal				Gamma			
p	COAT	Oracle	CCLasso	REBACCA	COAT	Oracle	CCLasso	REBACCA
				Matrix L_1-norm loss				
50	1.17 (0.13)	1.11 (0.12)	1.54 (0.27)	2.24 (0.23)	1.19 (0.14)	1.13 (0.14)	3.04 (0.37)	2.25 (0.22)
100	2.80 (0.15)	2.03 (0.13)	2.23 (0.38)	4.14 (0.28)	2.08 (0.14)	2.04 (0.14)	5.37 (0.98)	4.09 (0.28)
200	2.38 (0.07)	2.38 (0.07)	2.56 (0.22)	7.70 (0.38)	2.37 (0.07)	2.38 (0.07)	10.37 (1.37)	7.71 (0.32)
				Spectral norm loss				
50	0.56 (0.05)	0.56 (0.05)	0.68 (0.10)	0.87 (0.07)	0.57 (0.06)	0.57 (0.06)	0.90 (0.13)	0.86 (0.07)
100	0.79 (0.05)	0.79 (0.05)	0.84 (0.10)	1.29 (0.07)	0.79 (0.05)	0.80 (0.05)	1.22 (0.24)	1.30 (0.08)
200	0.95 (0.02)	0.95 (0.02)	0.94 (0.04)	1.97 (0.07)	0.95 (0.02)	0.95 (0.02)	1.90 (0.29)	1.98 (0.07)
				Frobenius norm loss				
50	1.49 (0.06)	1.44 (0.07)	2.11 (0.22)	2.77 (0.11)	1.49 (0.08)	1.44 (0.09)	2.88 (0.34)	2.76 (0.09)
100	2.39 (0.06)	2.39 (0.06)	3.40 (0.42)	5.53 (0.10)	2.40 (0.07)	2.40 (0.07)	4.91 (0.82)	5.54 (0.10)
200	3.03 (0.04)	3.03 (0.04)	3.57 (0.41)	11.01 (0.13)	3.03 (0.04)	3.03 (0.04)	9.49 (1.41)	11.02 (0.14)
				True positive rate				
50	0.96 (0.02)	0.97 (0.02)	0.96 (0.02)	0.96 (0.02)	0.96 (0.02)	0.96 (0.02)	1.00 (0.00)	0.96 (0.02)
100	0.82 (0.03)	0.82 (0.03)	0.79 (0.07)	0.89 (0.02)	0.82 (0.03)	0.81 (0.04)	0.97 (0.02)	0.89 (0.03)
200	0.39 (0.04)	0.40 (0.04)	0.36 (0.05)	0.76 (0.02)	0.40 (0.04)	0.40 (0.05)	0.93 (0.04)	0.76 (0.02)
				False positive rate				
50	0.12 (0.03)	0.09 (0.02)	0.12 (0.03)	0.35 (0.02)	0.12 (0.03)	0.09 (0.02)	0.78 (0.09)	0.35 (0.02)
100	0.05 (0.01)	0.04 (0.01)	0.06 (0.04)	0.36 (0.01)	0.05 (0.01)	0.04 (0.01)	0.50 (0.13)	0.36 (0.02)
200	0.01 (0.00)	0.01 (0.00)	0.01 (0.01)	0.37 (0.01)	0.01 (0.00)	0.01 (0.00)	0.67 (0.14)	0.37 (0.01)
				Run time (seconds)				
50	0.16 (0.02)	0.16 (0.01)	0.97 (0.06)	1.46 (0.03)	0.10 (0.01)	0.10 (0.01)	0.62 (0.05)	1.00 (0.05)
100	0.32 (0.01)	0.32 (0.02)	3.30 (0.22)	21.00 (0.81)	0.35 (0.03)	0.34 (0.01)	5.19 (0.19)	22.50 (1.04)
200	1.29 (0.07)	1.20 (0.05)	16.71 (1.64)	897.75 (54.93)	1.35 (0.11)	1.30 (0.10)	36.63 (8.05)	1001.78 (83.12)



(a) Model 2: hub covariance



(b) Model 3: block covariance



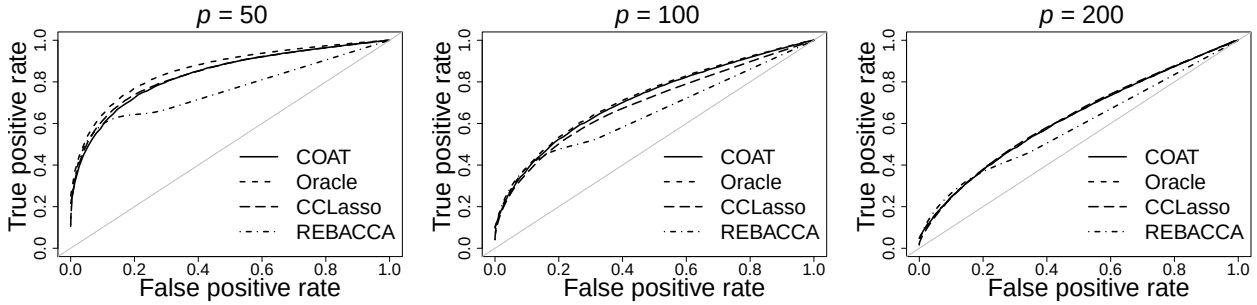
(c) Model 4: sparse covariance

Figure 2: ROC curves for four different methods in Models 2–4 with a normal log-basis distribution.

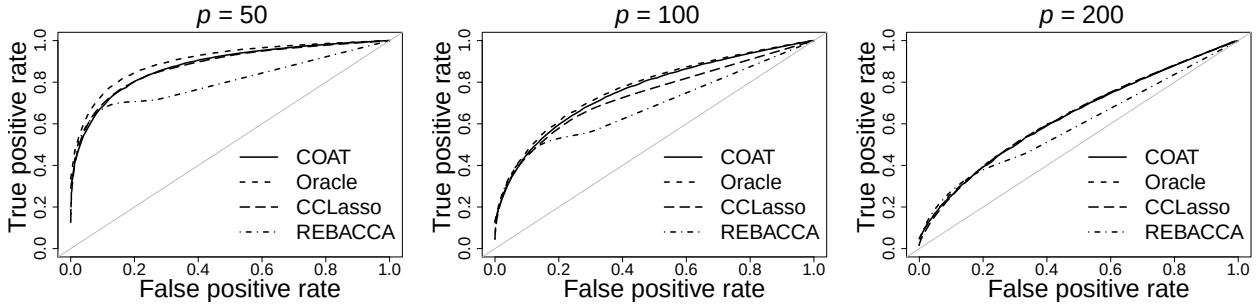
ality is very high, the COAT and oracle methods perform almost identically for all three models and sample sizes, and neither of them is sensitive to the log-basis distributions. In addition, as the sample size increases, all of the matrix losses tend to decrease.

6 Gut Microbiome Data Analysis

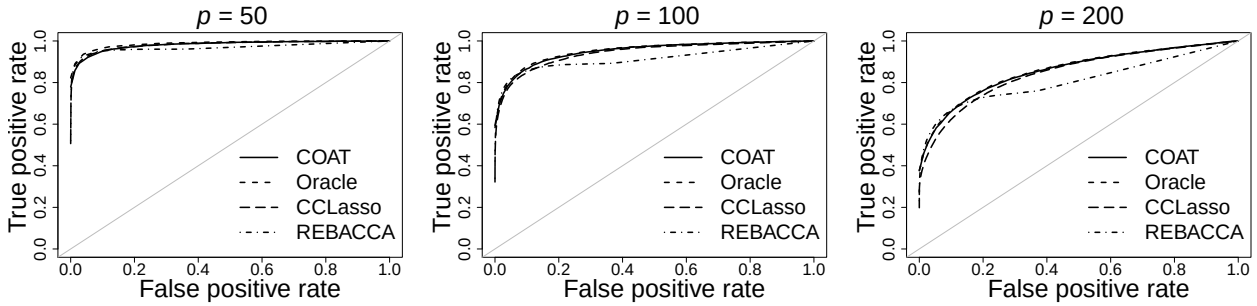
The gut microbiome plays a critical role in energy extraction from the diet and interacts with the immune system to exert a profound influence on human health and disease. Despite an emerging interest in characterizing the ecology of human-associated microbial communities, the complex



(a) Model 2: hub covariance



(b) Model 3: block covariance



(c) Model 4: sparse covariance

Figure 3: ROC curves for four different methods in Models 2–4 with a gamma log-basis distribution.

interactions among microbial taxa remain poorly understood (Coyte, Schluter, and Foster 2015). We now illustrate the proposed method by applying it to a human gut microbiome dataset described by Wu et al. (2011), which was collected from a cross-sectional study of 98 healthy individuals at the University of Pennsylvania. DNA from stool samples of these subjects were analyzed by 454/Roche pyrosequencing of 16S rRNA gene segments, resulting in an average of 9265 reads per sample, with a standard deviation of 3864. Taxonomic assignment yielded 3068 operational taxonomic units, which were further combined into 87 genera that appeared in at least one sample. Demographic information, including body mass index (BMI), was also collected from the subjects.

Table 4: Comparisons of means (standard errors) of performance measures for the COAT and oracle methods in Models 2–4 with $p = 500$ and a normal log-basis distribution based on 100 replications.

	Model 2		Model 3		Model 4	
n	COAT	Oracle	COAT	Oracle	COAT	Oracle
			Matrix L_1-norm loss			
250	15.42 (0.02)	15.42 (0.02)	6.82 (0.04)	6.81 (0.04)	3.37 (0.05)	3.37 (0.05)
500	15.05 (0.11)	15.06 (0.11)	6.66 (0.08)	6.66 (0.08)	3.00 (0.08)	2.99 (0.08)
1000	13.25 (0.19)	13.26 (0.18)	6.14 (0.11)	6.13 (0.11)	2.36 (0.10)	2.35 (0.10)
			Spectral norm loss			
250	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.99 (0.01)	0.99 (0.01)
500	0.98 (0.01)	0.98 (0.01)	0.99 (0.01)	0.99 (0.01)	0.92 (0.02)	0.92 (0.02)
1000	0.92 (0.01)	0.92 (0.01)	0.93 (0.01)	0.93 (0.01)	0.72 (0.02)	0.73 (0.02)
			Frobenius norm loss			
250	9.87 (0.00)	9.87 (0.00)	11.37 (0.01)	11.37 (0.01)	4.13 (0.02)	4.14 (0.02)
500	9.74 (0.02)	9.74 (0.02)	11.04 (0.02)	11.04 (0.02)	3.89 (0.03)	3.88 (0.03)
1000	9.03 (0.02)	9.03 (0.02)	9.93 (0.02)	9.92 (0.03)	3.05 (0.03)	3.04 (0.03)
			True positive rate			
250	0.01 (0.00)	0.01 (0.00)	0.02 (0.01)	0.02 (0.01)	0.30 (0.02)	0.29 (0.02)
500	0.08 (0.01)	0.08 (0.01)	0.14 (0.01)	0.14 (0.01)	0.60 (0.02)	0.60 (0.02)
1000	0.37 (0.01)	0.37 (0.01)	0.49 (0.01)	0.49 (0.01)	0.88 (0.01)	0.88 (0.01)
			False positive rate			
250	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
500	0.01 (0.00)	0.01 (0.00)	0.03 (0.00)	0.03 (0.00)	0.01 (0.00)	0.01 (0.00)
1000	0.08 (0.00)	0.08 (0.00)	0.13 (0.01)	0.13 (0.01)	0.01 (0.00)	0.01 (0.00)
			Run time (seconds)			
250	8.55 (0.07)	8.55 (0.08)	8.44 (0.09)	8.43 (0.08)	8.46 (0.05)	8.46 (0.06)
500	8.69 (0.06)	8.69 (0.06)	10.46 (0.10)	10.45 (0.12)	10.46 (0.11)	10.45 (0.10)
1000	9.37 (0.12)	9.36 (0.14)	11.41 (0.21)	11.40 (0.16)	11.37 (0.13)	11.28 (0.10)

We are interested in identifying and comparing the correlation structures among bacterial genera between lean and obese subjects. We therefore divided the dataset into a lean group ($\mathrm{BMI} < 25$, $n = 63$) and an obese group ($\mathrm{BMI} \geq 25$, $n = 35$), and focused on the $p = 40$ bacterial genera that appeared in at least four samples in each group. The count data were transformed into compositions after zero counts were replaced by 0.5.

Table 5: Comparisons of means (standard errors) of performance measures for the COAT and oracle methods in Models 2–4 with $p = 500$ and a gamma log-basis distribution based on 100 replications.

	Model 2		Model 3		Model 4	
n	COAT	Oracle	COAT	Oracle	COAT	Oracle
			Matrix L_1-norm loss			
250	15.42 (0.02)	15.42 (0.02)	6.82 (0.03)	6.82 (0.03)	3.37 (0.04)	3.37 (0.04)
500	15.08 (0.11)	15.07 (0.11)	6.67 (0.09)	6.66 (0.09)	3.02 (0.09)	3.01 (0.10)
1000	13.26 (0.21)	13.26 (0.20)	6.14 (0.11)	6.13 (0.11)	2.37 (0.09)	2.36 (0.09)
			Spectral norm loss			
250	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.99 (0.01)	0.99 (0.01)
500	0.98 (0.01)	0.98 (0.01)	0.99 (0.01)	0.99 (0.01)	0.92 (0.03)	0.92 (0.03)
1000	0.92 (0.02)	0.92 (0.02)	0.93 (0.01)	0.93 (0.01)	0.73 (0.03)	0.73 (0.03)
			Frobenius norm loss			
250	9.87 (0.00)	9.87 (0.00)	11.37 (0.01)	11.37 (0.01)	4.14 (0.02)	4.14 (0.02)
500	9.75 (0.02)	9.75 (0.02)	11.04 (0.02)	11.04 (0.02)	3.89 (0.04)	3.89 (0.04)
1000	9.03 (0.02)	9.03 (0.02)	9.93 (0.02)	9.92 (0.03)	3.07 (0.04)	3.06 (0.04)
			True positive rate			
250	0.01 (0.00)	0.01 (0.00)	0.02 (0.01)	0.02 (0.01)	0.29 (0.01)	0.30 (0.02)
500	0.08 (0.01)	0.08 (0.01)	0.14 (0.01)	0.14 (0.01)	0.60 (0.02)	0.60 (0.02)
1000	0.37 (0.01)	0.37 (0.01)	0.49 (0.01)	0.49 (0.01)	0.88 (0.01)	0.88 (0.01)
			False positive rate			
250	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
500	0.01 (0.00)	0.01 (0.00)	0.03 (0.00)	0.03 (0.00)	0.01 (0.00)	0.01 (0.00)
1000	0.08 (0.00)	0.08 (0.00)	0.13 (0.01)	0.13 (0.01)	0.01 (0.00)	0.01 (0.00)
			Run time (seconds)			
250	8.39 (0.11)	8.37 (0.08)	8.34 (0.05)	8.33 (0.05)	8.29 (0.05)	8.28 (0.05)
500	8.69 (0.06)	8.68 (0.10)	10.36 (0.10)	10.35 (0.11)	10.34 (0.11)	10.33 (0.10)
1000	9.41 (0.07)	9.39 (0.12)	11.41 (0.21)	11.40 (0.16)	10.18 (0.08)	10.15 (0.08)

We applied the COAT method with soft thresholding to each group, and used tenfold cross-validation to select the tuning parameter. The resulting estimate was represented by a correlation network among the bacterial genera with each edge representing a nonzero correlation. To assess the stability of support recovery, we further generated 100 bootstrap samples for each group and repeated the thresholding procedure on each sample. The stability of the correlation network was measured by the average proportion of edges reproduced by each bootstrap replicate. Finally,

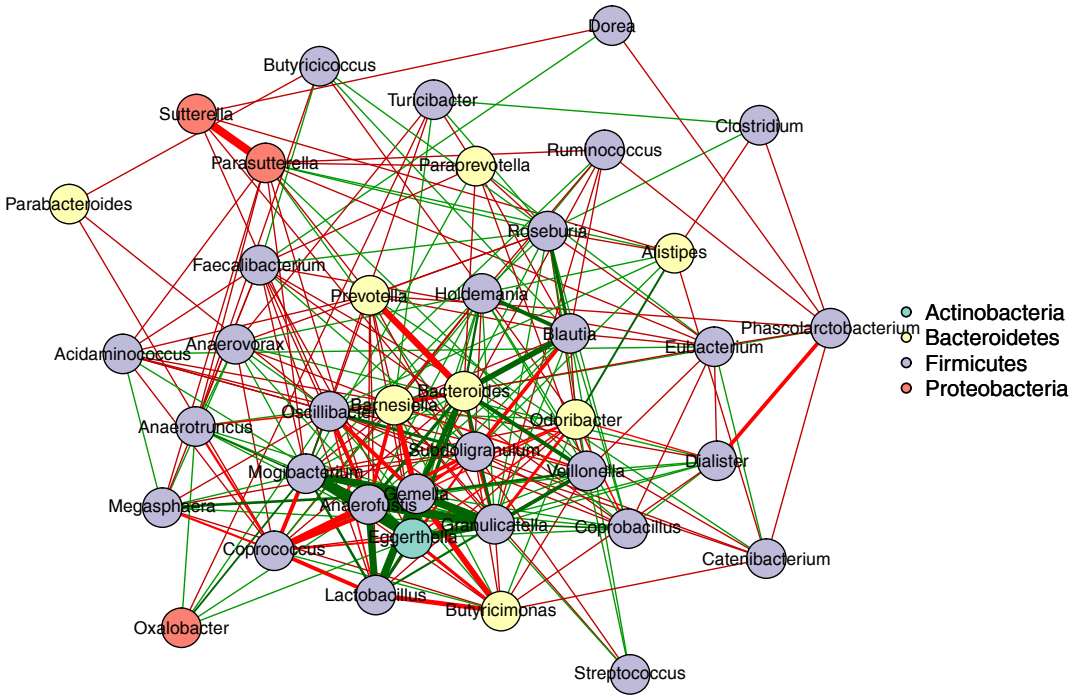
Table 6: Numbers of positive and negative correlations and stability of correlation networks for three different methods applied to the gut microbiome data.

	Lean			Obese		
	COAT	CCLasso	REBACCA	COAT	CCLasso	REBACCA
Positive correlations	114	98	18	61	43	12
Negative correlations	140	95	11	80	43	9
Network stability	0.90	0.86	0.57	0.87	0.80	0.53

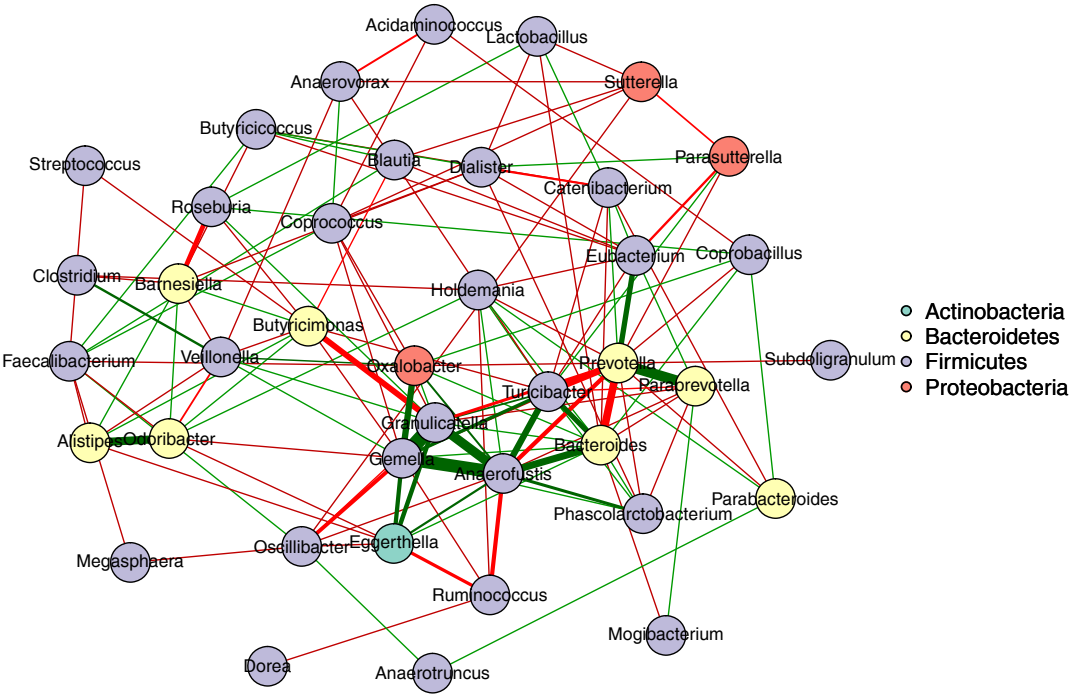
we retained only the edges in the correlation network that were reproduced in at least 85 bootstrap replicates. The numbers of positive and negative correlations and the stability of correlation networks are reported in Table 6; the results for the CCLasso and REBACCA methods are also included for comparison. We see that COAT achieves the highest stability among the three methods and identifies more edges based on the stability criterion. The correlation network identified by REBACCA seems the least stable.

The correlation networks identified by the COAT method for the two groups are displayed in Figure 4. Clearly, the networks for the lean and obese groups show markedly different architecture, indicating that the obese microbiome is less modular with less complex interactions between the modules. This phenomenon has been demonstrated by previous studies and is possibly due to adaptation of the microbiome to low-diversity environments (Greenblum, Turnbaugh, and Borenstein 2012). Table 6 and Figure 4 also suggest that the gut microbial network tends to contain more competitive (negative) interactions than cooperative (positive) ones, which seems consistent with the recent finding that the ecological stability of the gut microbiome can be attributed to the benefits from limiting positive feedbacks and dampening cooperative networks (Coyte, Schluter, and Foster 2015).

A closer inspection of the correlation networks identifies *Bacteroides* and *Prevotella* as two key genera of the gut microbiome. The abundances of these two genera are well known to distinguish two gut microbial enterotypes, which are strongly associated with long-term dietary patterns (Arumugam et al. 2011; Wu et al. 2011). The negative correlations between *Bacteroides* and *Prevotella* (-0.372 in the lean group and -0.377 in the obese group) are well explained by the diet-dependent



(a) Lean



(b) Obese

Figure 4: Correlation networks identified by the COAT method for the lean and obese groups in the gut microbiome data. Positive and negative correlations are displayed in green and red, respectively. The thickness of edges indicates the magnitude of correlations.

enterotypes and the within-body separation of the two genera (Jordán et al. 2015). Moreover, recent studies have suggested several keystone species belonging to the genus *Bacteroides*, through which the structure of gut microbial communities may be influenced by small perturbations (Fisher and Mehta 2014). Also, the Firmicutes-enriched microbiome has been found to hold greater metabolic potential than the Bacteroidetes-enriched microbiome for more efficient energy harvest from the diet (Turnbaugh et al. 2006). Figure 4 seems to support these findings, in view of the central position of *Bacteroides* in the networks and its strong correlations with a few genera belonging to the Firmicutes. Such patterns, however, are less clearly seen in the correlation networks identified by the other two methods.

7 Discussion

Understanding the dependence structure among microbial taxa within communities, including the co-occurrence and co-exclusion relationships between microbial taxa, is an important problem in microbiome research. Such structures provide biological insights into the community dynamics and factors that change the community structures. To overcome the difficulties arising from the unit-sum constraint of the observed compositional data, we have developed a COAT method to estimate the sparse covariance matrix of the latent log-basis components. Our method is based on a decomposition of the variation matrix into a rank-2 component and a sparse component. The resulting procedure is equivalent to thresholding the sample centered log-ratio covariance matrix, and thus is optimization-free and scalable for large covariance matrices. Our method also bears some resemblance to the POET method proposed by Fan, Liao, and Mincheva (2013) in that underlying both methods is a low-rank plus sparse matrix decomposition. The rank-2 component in our method, however, arises from the covariance structure of compositional data rather than a factor model assumption. As a result, it can be obtained by simple algebraic operations without computing the principal components.

Our simulation results demonstrate that COAT performs almost as well as the oracle estimator that assumes the basis components are observed, and outperforms CCLasso and REBACCA in

terms of both estimation and support recovery. COAT performs consistently better even when the log-basis has a skewed distribution such as the gamma, as is often observed in microbiome studies. Besides, COAT can be tens to hundreds of times faster than existing optimization-based estimators. In the application to gut microbiome data, COAT leads to more stable and biologically more interpretable results for comparing the dependence structures of lean and obese microbiomes.

We have provided conditions for the approximate and exact identifiability of the covariance parameters, and have established rates of convergence and support recovery guarantees for the COAT estimator. The rate of convergence under the spectral norm includes an extra term of $s_0(p)(s_0(p)/p)^{1-q}$ in addition to the minimax optimal rate for sparse covariance estimation. This term represents an approximation error due to using $\mathbf{\Gamma}_0$ as a proxy for $\mathbf{\Omega}_0$, which vanishes asymptotically under mild assumptions as the dimensionality increases. Although it reflects the level of parameter identifiability and cannot be removed without more stringent assumptions on the parameter space, it remains an open question whether it is rate-optimal.

The sparsity assumption $s_0(p) = o(p)$ is very mild and seems to be supported by the empirical literature on ecological networks. For instance, Rejmánek and Starý (1979) and Yodzis (1980) showed that connectance, which is defined as the fraction of interactions in a network, may decline considerably as the number of species increases. More recently, Dunne, Williams, and Martinez (2002) analyzed 16 high-quality food webs and found that connectance ranges from 0.026 to 0.315, indicating that most ecological networks are indeed sparse.

The proposed methodology may be extended in several ways. First, it would be possible to develop an iterative optimization procedure based on the decomposition (5). For example, one may consider the regularized estimator

$$\widehat{\mathbf{\Omega}}_{\text{opt}} = \arg\min_{\mathbf{\Omega}} \{\|\widehat{\mathbf{T}} - \boldsymbol{\omega}\mathbf{1}^T - \mathbf{1}\boldsymbol{\omega}^T + 2\mathbf{\Omega}\|_F^2 + P_\lambda(\mathbf{\Omega})\},$$

where $\boldsymbol{\omega}$ consists of the diagonal entries of $\mathbf{\Omega}$ and $P_\lambda(\cdot)$ is a sparsity-inducing penalty function. Note that COAT can be viewed as a one-step approximation to $\widehat{\mathbf{\Omega}}_{\text{opt}}$ with an appropriately chosen penalty function. Solving the full optimization problem is computationally more expensive but

could improve on the performance of COAT. Another worthwhile extension would be to develop a method that deals with zero counts directly. One may, in principle, combine the ideas presented here with models that account for sampling and structural zeros. The issues of identifiability and computational feasibility are the major concerns with such extensions.

Appendix: Proofs

A.1 Proof of Proposition 1

Using the fact that the centered log-ratio covariance matrix $\mathbf{\Gamma}_0$ is symmetric and has all zero row sums (Aitchison 2003, Property 4.6), we have

$$\begin{aligned}\langle \boldsymbol{\gamma}_0 \mathbf{1}^T + \mathbf{1}\boldsymbol{\gamma}_0^T, \mathbf{\Gamma}_0 \rangle &= \mathrm{tr}\{(\boldsymbol{\gamma}_0 \mathbf{1}^T + \mathbf{1}\boldsymbol{\gamma}_0^T)^T \mathbf{\Gamma}_0\} = \mathrm{tr}(\mathbf{1}\boldsymbol{\gamma}_0^T \mathbf{\Gamma}_0) + \mathrm{tr}(\boldsymbol{\gamma}_0 \mathbf{1}^T \mathbf{\Gamma}_0) \\ &= \mathrm{tr}(\boldsymbol{\gamma}_0^T \mathbf{\Gamma}_0 \mathbf{1}) + \mathrm{tr}(\boldsymbol{\gamma}_0 \mathbf{1}^T \mathbf{\Gamma}_0) = 0,\end{aligned}$$

that is, the components $\boldsymbol{\gamma}_0 \mathbf{1}^T + \mathbf{1}\boldsymbol{\gamma}_0^T$ and $\mathbf{\Gamma}_0$ are orthogonal to each other.

To show the desired inequality, by the identity (4.35) of Aitchison (2003), we have

$$\omega_{ij}^0 - \gamma_{ij}^0 = \omega_{ij}^0 - (\omega_{ij}^0 - \omega_{i\cdot}^0 - \omega_{j\cdot}^0 + \omega_{\cdot\cdot}^0) = \omega_{i\cdot}^0 + \omega_{j\cdot}^0 - \omega_{\cdot\cdot}^0.$$

Therefore,

$$\|\mathbf{\Omega}_0 - \mathbf{\Gamma}_0\|_{\max} \le \max_{i,j}(|\omega_{i\cdot}^0| + |\omega_{j\cdot}^0| + |\omega_{\cdot\cdot}^0|) \le 3p^{-1}\|\mathbf{\Omega}_0\|_1.$$

A.2 Proof of Proposition 2

We first claim that if $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T \neq \mathbf{0}$, then the matrix $\mathbf{A} \equiv \boldsymbol{\alpha}\mathbf{1}^T + \mathbf{1}\boldsymbol{\alpha}^T$ has at least $p-1$ nonzero upper-triangular entries. To prove this, without loss of generality, assume $\alpha_1 \neq 0$ and that the last q entries of the first row of $\mathbf{A}$ are zero, where $0 \le q \le p-1$; that is, $\alpha_1 + \alpha_j \neq 0$ for $1 \le j \le p-q$, and $\alpha_1 + \alpha_{p-q+1} = \cdots = \alpha_1 + \alpha_p = 0$. The latter implies $\alpha_{p-q+1} = \cdots = \alpha_p = -\alpha_1 \neq 0$, which gives rise to $\binom{q}{2} = q(q-1)/2$ nonzero entries at positions (i,j) with $p-q+1 \le i < j \le p$. Putting these pieces together, we obtain that the number of nonzero

upper-triangular entries in $\mathbf{A}$ is at least

$$f(q) \equiv p - q - 1 + \frac{q(q-1)}{2} \geq f(1) = f(2) = p - 2.$$

To show that the lower bound $p-2$ is not attainable, note that if there are only $p-2$ nonzero upper-triangular entries, then $q = 1$ or 2, and we have $\alpha_2 + \alpha_p = \cdots = \alpha_{p-2} + \alpha_p = 0$, which implies $\alpha_2 = \cdots = \alpha_{p-2} = -\alpha_p = \alpha_1 \neq 0$. Since $p \geq 5$, this gives rise to at least one nonzero entry at positions (i, j) with $2 \leq i < j \leq p-2$, which is a contradiction.

Now suppose $s_e(p) < (p-1)/2$ and that $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ in $\mathcal{B}_0(s_e(p))$ lead to $\mathbf{T}_1 = \mathbf{T}_2$, that is,

$$(\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2)\mathbf{1}^T + \mathbf{1}(\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2)^T = 2(\mathbf{\Omega}_1 - \mathbf{\Omega}_2).$$

Note that the right-hand side has fewer than $p-1$ nonzero upper-triangular entries. Then it follows from the above claim that $\mathbf{\Omega}_1 = \mathbf{\Omega}_2$.

We prove the other direction by showing that, if $s_e(p) \geq (p-1)/2$, then there exist $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ in $\mathcal{B}_0(s_e(p))$ with $\mathbf{\Omega}_1 \neq \mathbf{\Omega}_2$ that lead to $\mathbf{T}_1 = \mathbf{T}_2$. Indeed, let

$$\mathbf{\Omega}_1 = \begin{pmatrix} 1+c & c\mathbf{1}_{p_1}^T & \mathbf{0}_{p_2}^T \\ c\mathbf{1}_{p_1} & \mathbf{I}_{p_1} & \mathbf{0} \\ \mathbf{0}_{p_2} & \mathbf{0} & \mathbf{I}_{p_2} \end{pmatrix}, \quad \mathbf{\Omega}_2 = \begin{pmatrix} 1-c & \mathbf{0}_{p_1}^T & -c\mathbf{1}_{p_2}^T \\ \mathbf{0}_{p_1} & \mathbf{I}_{p_1} & \mathbf{0} \\ -c\mathbf{1}_{p_2} & \mathbf{0} & \mathbf{I}_{p_2} \end{pmatrix},$$

where $p_1 = \lfloor (p-1)/2 \rfloor$, $p_2 = p - 1 - p_1$, and $0 < |c| < 1$. Then it is easy to verify that

$$\mathbf{T}_1 = \mathbf{T}_2 = \begin{pmatrix} 0 & (2-c)\mathbf{1}_{p_1}^T & (2+c)\mathbf{1}_{p_2}^T \\ (2-c)\mathbf{1}_{p_1} & 2(\mathbf{1}_{p_1}\mathbf{1}_{p_1}^T - \mathbf{I}_{p_1}) & 2\mathbf{1}_{p_1}\mathbf{1}_{p_2}^T \\ (2+c)\mathbf{1}_{p_2} & 2\mathbf{1}_{p_2}\mathbf{1}_{p_1}^T & 2(\mathbf{1}_{p_2}\mathbf{1}_{p_2}^T - \mathbf{I}_{p_2}) \end{pmatrix}.$$

This completes the proof.

A.3 Concentration Inequalities

To prepare for the proofs of Theorems 1 and 2, we first establish some useful concentration inequalities. For notational simplicity, the constants $C_1, C_2, \ldots$ below may vary from line to line.

Lemma 1. *Under Condition 1, there exist constants $C_1, C_2 > 0$ such that*

$$P\left(\max_j \left|\frac{1}{n}\sum_{k=1}^n Y_{kj}\right| \geq t\right) \leq C_1 p e^{-C_2 n t^2} \tag{A.1}$$

and

$$P\left(\max_{i,j} \left|\frac{1}{n}\sum_{k=1}^n Y_{ki}Y_{kj} - EY_iY_j\right| \geq t\right) \leq C_1 p^2 e^{-C_2 n t^2} \tag{A.2}$$

for sufficiently small $t > 0$. Moreover, if $\log p = o(n^{1/5})$, then there exists a constant $C_3 > 0$ such that

$$P\left(\max_{i,j,\ell,m} \left|\frac{1}{n}\sum_{k=1}^n Y_{ki}Y_{kj}Y_{k\ell}Y_{km} - EY_iY_jY_\ell Y_m\right| \geq \varepsilon\right) = O(p^{-C_3}) \tag{A.3}$$

for every constant $\varepsilon > 0$.

Proof. Inequalities (A.1) and (A.2) follow, for example, from Exercise 2.27 of Boucheron, Lugosi, and Massart (2013); see also Bickel and Levina (2008).

To prove (A.3), let $Z_{kijlm} = Y_{ki}Y_{kj}Y_{k\ell}Y_{km}$ and $Z_{ijlm} = Y_iY_jY_\ell Y_m$. Note first that, by Condition 1 and the sub-Gaussian tail bound, for any $K > 0$ and i, j, ℓ, m,

$$P(|Z_{ijlm}| > K) \leq 4P(|Y_j| > K^{1/4}) \leq 8e^{-\alpha\sqrt{K}/8}.$$

Hence,

$$\begin{aligned}
E|Z_{ijlm}|I(|Z_{ijlm}| > K) &= \int_0^\infty P(|Z_{ijlm}|I(|Z_{ijlm}| > K) > z)\,dz \\
&= KP(|Z_{ijlm}| > K) + \int_K^\infty P(|Z_{ijlm}| > z)\,dz \\
&\leq 8Ke^{-\alpha\sqrt{K}/8} + \int_K^\infty 8e^{-\alpha\sqrt{z}/8}\,dz \\
&= \frac{8}{\alpha^2}(\alpha^2 K + 16\alpha\sqrt{K} + 128)e^{-\alpha\sqrt{K}/8},
\end{aligned}$$

which is less than $\varepsilon/4$ if we choose K sufficiently large. Then we have

$$\begin{aligned}
&P\left(\max_{i,j,\ell,m} \left|\frac{1}{n}\sum_{k=1}^n Z_{kijlm} - EZ_{ijlm}\right| \geq \varepsilon\right) \\
&\leq P\left(\max_{i,j,\ell,m} \left|\frac{1}{n}\sum_{k=1}^n Z_{kijlm}I(|Z_{kijlm}| \leq K) - EZ_{ijlm}I(|Z_{ijlm}| \leq K)\right| \geq \frac{\varepsilon}{2}\right)
\end{aligned}$$

$$+ P\left(\max_{i,j,\ell,m}\left|\frac{1}{n}\sum_{k=1}^{n} Z_{kijlm} I(|Z_{kijlm}| > K)\right| \geq \frac{\varepsilon}{4}\right)$$
$$\equiv T_1 + T_2.$$

By Hoeffding's inequality and the union bound,

$$T_1 \leq 2p^4 \exp\left(-\frac{n\varepsilon^2}{8K^2}\right).$$

Also, by Condition 1 and the sub-Gaussian tail bound,

$$T_2 \leq P\left(\max_{k,i,j,\ell,m} |Z_{kijlm}| > K\right) \leq P\left(\max_{k,j} |Y_{kj}| > K^{1/4}\right) \leq 2npe^{-\alpha\sqrt{K}/8}.$$

Combining both terms, choosing $K = C^2(\log p + \log n)^2$ with $C > 8/\alpha$, and noting $\log p = o(n^{1/5})$, we arrive at

$$\begin{aligned} &P\left(\max_{i,j,\ell,m}\left|\frac{1}{n}\sum_{k=1}^{n} Z_{kijlm} - EZ_{ijlm}\right| \geq \varepsilon\right) \\ &\leq 2p^4 \exp\left(-\frac{n\varepsilon^2}{8C^4(\log p + \log n)^4}\right) + 2(np)^{1-C\alpha/8} \\ &= O(p^{-C_3}) \end{aligned}$$

for some $C_3 > 0$. This proves (A.3) and completes the proof.

Lemma 2. *Under Conditions 1–4, there exist constants $C_1, C_2, C_3 > 0$ such that*

$$P\left(\max_{i,j} |\hat{\theta}_{ij} - \theta_{ij}| \geq \varepsilon\right) = O(p^{-C_3}) \tag{A.4}$$

and

$$P\left(\max_{i,j} |\hat{\gamma}_{ij} - \omega_{ij}^0| / \sqrt{\hat{\theta}_{ij}} \geq C_1\sqrt{\frac{\log p}{n}} + C_2\frac{s_0(p)}{p}\right) = O(p^{-C_3}) \tag{A.5}$$

for every constant $\varepsilon > 0$.

Proof. We first prove (A.4). Define

$$\tilde{\theta}_{ij} = \frac{1}{n}\sum_{k=1}^{n} (\gamma_{ki}\gamma_{kj} - \tilde{\gamma}_{ij})^2,$$

where $\tilde{\gamma}_{ij} = n^{-1}\sum_{k=1}^{n}\gamma_{ki}\gamma_{kj}$. We then write

$$\begin{aligned}\hat{\theta}_{ij} - \tilde{\theta}_{ij} &= \frac{1}{n}\sum_{k=1}^{n}\{(\gamma_{ki}\gamma_{kj} - \tilde{\gamma}_{ij}) - \gamma_{ki}\bar{\gamma}_j - \gamma_{kj}\bar{\gamma}_i + 2\bar{\gamma}_i\bar{\gamma}_j\}^2 - \frac{1}{n}\sum_{k=1}^{n}(\gamma_{ki}\gamma_{kj} - \tilde{\gamma}_{ij})^2 \\ &= \frac{2}{n}\sum_{k=1}^{n}(\gamma_{ki}\gamma_{kj} - \tilde{\gamma}_{ij})(-\gamma_{ki}\bar{\gamma}_j - \gamma_{kj}\bar{\gamma}_i + 2\bar{\gamma}_i\bar{\gamma}_j) + \frac{1}{n}\sum_{k=1}^{n}(-\gamma_{ki}\bar{\gamma}_j - \gamma_{kj}\bar{\gamma}_i + 2\bar{\gamma}_i\bar{\gamma}_j)^2.\end{aligned} \tag{A.6}$$

Note that, by definition, $\gamma_{kj} = Y_{kj} - \bar{Y}_k$, where $\bar{Y}_k = p^{-1}\sum_{j=1}^{p}Y_{kj}$. Define $\gamma_j = Y_j - \bar{Y}$, where $\bar{Y} = p^{-1}\sum_{j=1}^{p}Y_j$. Since Y_j are uniformly sub-Gaussian by Condition 1, γ_j are also uniformly sub-Gaussian. Using a truncation argument similar to that for proving (A.3), we can show that

$$P\left(\max_{i,j}\left|\frac{1}{n}\sum_{k=1}^{n}\gamma_{ki}^2\gamma_{kj} - E\gamma_i^2\gamma_j\right| \geq C_1\right) = O(p^{-C_3})$$

for some $C_1, C_3 > 0$. The sub-Gaussian tails imply also that $E\gamma_i^2|\gamma_j| \leq \frac{1}{2}(E\gamma_i^4 + E\gamma_j^2) = O(1)$. Combining these two pieces yields

$$P\left(\max_{i,j}\left|\frac{1}{n}\sum_{k=1}^{n}\gamma_{ki}^2\gamma_{kj}\right| \geq C_1\right) = O(p^{-C_3}). \tag{A.7}$$

Note that

$$\max_j|\bar{\gamma}_j| = \max_j\left|\frac{1}{n}\sum_{k=1}^{n}\gamma_{kj}\right| = \left|\frac{1}{n}\sum_{k=1}^{n}Y_{kj} - \frac{1}{np}\sum_{k=1}^{n}\sum_{i=1}^{p}Y_{ki}\right| \leq 2\max_j\left|\frac{1}{n}\sum_{k=1}^{n}Y_{kj}\right|.$$

Also, it follows from (A.1) in Lemma 1 that

$$P\left(\max_j\left|\frac{1}{n}\sum_{k=1}^{n}Y_{kj}\right| \geq C_1\sqrt{\frac{\log p}{n}}\right) = O(p^{-C_3})$$

for some $C_1, C_3 > 0$. Hence,

$$P\left(\max_j|\bar{\gamma}_j| \geq C_1\sqrt{\frac{\log p}{n}}\right) = O(p^{-C_3}). \tag{A.8}$$

Inequalities (A.7) and (A.8) together imply

$$P\left(\max_{i,j}\left|\frac{1}{n}\sum_{k=1}^{n}\gamma_{ki}^2\gamma_{kj}\bar{\gamma}_j\right| \geq C_1\sqrt{\frac{\log p}{n}}\right) = O(p^{-C_3}). \tag{A.9}$$

We can similarly bound the other terms in (A.6) and obtain

$$P\left(\max_{i,j}|\hat{\theta}_{ij} - \tilde{\theta}_{ij}| \geq C_1\sqrt{\frac{\log p}{n}}\right) = O(p^{-C_3}). \tag{A.10}$$

Next, write

$$\begin{aligned}\tilde{\theta}_{ij} - \theta_{ij} &= \frac{1}{n}\sum_{k=1}^{n}(\gamma_{ki}\gamma_{kj} - \tilde{\gamma}_{ij})^2 - \mathrm{Var}(Y_iY_j) \\ &= \frac{1}{n}\sum_{k=1}^{n}\gamma_{ki}^2\gamma_{kj}^2 - EY_i^2Y_j^2 - \{\tilde{\gamma}_{ij}^2 - (\omega_{ij}^0)^2\} \\ &\equiv T_1 + T_2.\end{aligned}$$

To bound the term T_1, we further write

$$\begin{aligned}T_1 &= \frac{1}{n}\sum_{k=1}^{n}\{(Y_{ki} - \bar{Y}_k)(Y_{kj} - \bar{Y}_k)\}^2 - EY_i^2Y_j^2 \\ &= \frac{1}{n}\sum_{k=1}^{n}\left(Y_{ki}Y_{kj} - Y_{ki}\bar{Y}_k - Y_{kj}\bar{Y}_k + \bar{Y}_k^2\right)^2 - EY_i^2Y_j^2 \\ &= \frac{1}{n}\sum_{k=1}^{n}Y_{ki}^2Y_{kj}^2 - EY_i^2Y_j^2 + \frac{2}{n}\sum_{k=1}^{n}Y_{ki}Y_{kj}(-Y_{ki}\bar{Y}_k - Y_{kj}\bar{Y}_k + \bar{Y}_k^2) \\ &\quad + \frac{1}{n}(-Y_{ki}\bar{Y}_k - Y_{kj}\bar{Y}_k + \bar{Y}_k^2)^2.\end{aligned}$$

Consider the event A_1 on which

$$\max_{i,j,\ell,m}\left|\frac{1}{n}\sum_{k=1}^{n}Y_{ki}Y_{kj}Y_{k\ell}Y_{km} - EY_iY_jY_\ell Y_m\right| \le \varepsilon_1.$$

Then, on A_1, we have

$$\left|\frac{1}{n}\sum_{k=1}^{n}Y_{ki}^2Y_{kj}^2 - EY_i^2Y_j^2\right| \le \varepsilon_1.$$

To bound the next term in T_1, we write

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}Y_{ki}^2Y_{kj}\bar{Y}_k &= \frac{1}{n}\sum_{k=1}^{n}Y_{ki}^2Y_{kj}\bar{Y}_k - EY_i^2Y_j\bar{Y} + EY_i^2Y_j\bar{Y} \\ &= \frac{1}{p}\sum_{\ell=1}^{p}\left(\frac{1}{n}\sum_{k=1}^{n}Y_{ki}^2Y_{kj}Y_{k\ell} - EY_i^2Y_jY_\ell\right) + \frac{1}{p}\sum_{\ell=1}^{p}EY_i^2Y_jY_\ell,\end{aligned}$$

which, on A_1 and by Condition 4, is bounded by $\varepsilon_1 + s_1(p)/p$. We can similarly bound the other terms in T_1 and obtain, on A_1,

$$|T_1| \le 16\varepsilon_1 + 15s_1(p)/p. \tag{A.11}$$

To bound the term T_2, note that

$$\begin{aligned}\tilde{\gamma}_{ij} - \omega_{ij}^0 &= \frac{1}{n}\sum_{k=1}^{n}(Y_{ki} - \bar{Y}_k)(Y_{kj} - \bar{Y}_k) - EY_iY_j \\ &= \frac{1}{n}\sum_{k=1}^{n} Y_{ki}Y_{kj} - EY_iY_j + \frac{1}{n}\sum_{k=1}^{n}(-Y_{ki}\bar{Y}_k - Y_{kj}\bar{Y}_k + \bar{Y}_k^2). \end{aligned} \tag{A.12}$$

Consider the event A_2 on which

$$\max_{i,j}\left|\frac{1}{n}\sum_{k=1}^{n} Y_{ki}Y_{kj} - EY_iY_j\right| \le \varepsilon_2.$$

To bound the next term in (A.12), we write

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n} Y_{ki}\bar{Y}_k &= \frac{1}{n}\sum_{k=1}^{n} Y_{ki}\bar{Y}_k - EY_i\bar{Y} + EY_i\bar{Y} \\ &= \frac{1}{p}\sum_{j=1}^{p}\left(\frac{1}{n}\sum_{k=1}^{n} Y_{ki}Y_{kj} - EY_iY_j\right) + \frac{1}{p}\sum_{j=1}^{p}\omega_{ij}^0,\end{aligned}$$

which, on A_2 and by Condition 2, is bounded by $\varepsilon_2 + M^{1-q}s_0(p)/p$. We can similarly bound the other terms in (A.12) and obtain, on A_2,

$$|\tilde{\gamma}_{ij} - \omega_{ij}^0| \le 4\varepsilon_2 + 3M^{1-q}s_0(p)/p. \tag{A.13}$$

Note also that, on A_2,

$$|\tilde{\gamma}_{ij} + \omega_{ij}^0| \le |\tilde{\gamma}_{ij} - \omega_{ij}^0| + 2|\omega_{ij}^0| \le 4\varepsilon_2 + 3M^{1-q}s_0(p)/p + 2M.$$

Hence, on A_2, we have

$$|T_2| = |\tilde{\gamma}_{ij} - \omega_{ij}^0||\tilde{\gamma}_{ij} + \omega_{ij}^0| \le (4\varepsilon_2 + 3M^{1-q}s_0(p)/p)(4\varepsilon_2 + 3M^{1-q}s_0(p)/p + 2M). \tag{A.14}$$

Finally, it follows from Lemma 1 that the event $A_1 \cap A_2$ occurs with probability at least $1 - O(p^{-C_3})$ for all constants $\varepsilon_1, \varepsilon_2 > 0$ and some constant $C_3 > 0$. Combining (A.10), (A.11), and (A.14) and noting $\log p = o(n)$, $s_0(p) = o(p)$, and $s_1(p) = o(p)$, we arrive at (A.4).

It remains to prove (A.5). We first write

$$\hat{\gamma}_{ij} - \tilde{\gamma}_{ij} = \frac{1}{n}\sum_{k=1}^{n}(\gamma_{ki} - \bar{\gamma}_i)(\gamma_{kj} - \bar{\gamma}_j) - \frac{1}{n}\sum_{k=1}^{n}\gamma_{ki}\gamma_{kj}$$

$$= \frac{1}{n}\sum_{k=1}^{n}(-\gamma_{ki}\bar{\gamma}_i - \gamma_{kj}\bar{\gamma}_j + \bar{\gamma}_i\bar{\gamma}_j).$$

Using arguments similar to those for proving (A.9), we can show that

$$P\left(\max_{i,j}\left|\frac{1}{n}\sum_{k=1}^{n}\gamma_{ki}\bar{\gamma}_j\right| \geq C_1\sqrt{\frac{\log p}{n}}\right) = O(p^{-C_3}).$$

We can similarly bound the other two terms and obtain

$$P\left(\max_{i,j}|\hat{\gamma}_{ij} - \tilde{\gamma}_{ij}| \geq C_1\sqrt{\frac{\log p}{n}}\right) = O(p^{-C_3}).$$

Taking $\varepsilon_2 = C_1\sqrt{(\log p)/n}$ in (A.13), we have

$$P\left(\max_{i,j}|\tilde{\gamma}_{ij} - \omega_{ij}^0| \geq C_1\sqrt{\frac{\log p}{n}} + C_2\frac{s_0(p)}{p}\right) = O(p^{-C_3}).$$

The above two inequalities together imply

$$P\left(\max_{i,j}|\hat{\gamma}_{ij} - \omega_{ij}^0| \geq C_1\sqrt{\frac{\log p}{n}} + C_2\frac{s_0(p)}{p}\right) = O(p^{-C_3}). \tag{A.15}$$

From Condition 3 and (A.4) with $\varepsilon_2 = \tau/2$, it follows that $|\hat{\theta}_{ij}| \geq \tau/2$ with probability at least $1 - O(p^{-C_3})$. This, together with (A.15), implies (A.5) and completes the proof.

A.4 Proof of Theorem 1

By the triangle inequality, we have

$$\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}_0\|_1 \leq \sum_{j=1}^{p}|S_{\lambda_{ij}}(\omega_{ij}^0) - \omega_{ij}^0| + \sum_{j=1}^{p}|S_{\lambda_{ij}}(\hat{\gamma}_{ij}) - S_{\lambda_{ij}}(\omega_{ij}^0)|. \tag{A.16}$$

Using Conditions (i) and (ii) that define a general thresholding function, the first term above is bounded by

$$\begin{aligned}
&\sum_{j=1}^{p}|\omega_{ij}^0|I(|\omega_{ij}^0| \leq \lambda_{ij}) + \sum_{j=1}^{p}\lambda_{ij}I(|\omega_{ij}^0| > \lambda_{ij}) \\
&\quad= \sum_{j=1}^{p}|\omega_{ij}^0|^q|\omega_{ij}^0|^{1-q}I(|\omega_{ij}^0| \leq \lambda_{ij}) + \sum_{j=1}^{p}\lambda_{ij}^q\lambda_{ij}^{1-q}I(|\omega_{ij}^0| > \lambda_{ij}) \\
&\quad\leq \sum_{j=1}^{p}|\omega_{ij}^0|^q\lambda_{ij}^{1-q}.
\end{aligned}$$

On the other hand, the second term in (A.16) is bounded by

$$\begin{aligned}
&2\sum_{j=1}^{p}|\hat{\gamma}_{ij}|I(|\hat{\gamma}_{ij}|>\lambda_{ij},|\omega_{ij}^0|\leq\lambda_{ij})+2\sum_{j=1}^{p}|\omega_{ij}^0|I(|\hat{\gamma}_{ij}|\leq\lambda_{ij},|\omega_{ij}^0|>\lambda_{ij})\\
&\quad+\sum_{j=1}^{p}|S_{\lambda_{ij}}(\hat{\gamma}_{ij})-S_{\lambda_{ij}}(\omega_{ij}^0)|I(|\hat{\gamma}_{ij}|>\lambda_{ij},|\omega_{ij}^0|>\lambda_{ij})\\
&\quad\equiv T_1+T_2+T_3.
\end{aligned}$$

To bound the term T_1, we write

$$\begin{aligned}
\frac{T_1}{2}&\leq\sum_{j=1}^{p}|\hat{\gamma}_{ij}-\omega_{ij}^0|I(|\hat{\gamma}_{ij}|>\lambda_{ij},|\omega_{ij}^0|\leq\lambda_{ij}/2)\\
&\quad+\sum_{j=1}^{p}|\hat{\gamma}_{ij}-\omega_{ij}^0|I(|\hat{\gamma}_{ij}|>\lambda_{ij},\lambda_{ij}/2<|\omega_{ij}^0|\leq\lambda_{ij})+\sum_{j=1}^{p}|\omega_{ij}^0|I(|\hat{\gamma}_{ij}|>\lambda_{ij},|\omega_{ij}^0|\leq\lambda_{ij})\\
&\equiv T_4+T_5+T_6.
\end{aligned}$$

Consider the event B_1 on which $|\hat{\gamma}_{ij}-\omega_{ij}^0|\leq\lambda_{ij}/2$ for all i,j. On B_1, we have

$$\begin{aligned}
T_4&\leq\sum_{j=1}^{p}|\hat{\gamma}_{ij}-\omega_{ij}^0|I(|\hat{\gamma}_{ij}-\omega_{ij}^0|>\lambda_{ij}/2)=0,\\
T_5&\leq\sum_{j=1}^{p}\left(\frac{\lambda_{ij}}{2}\right)^q\left(\frac{\lambda_{ij}}{2}\right)^{1-q}I(|\hat{\gamma}_{ij}|>\lambda_{ij},\lambda_{ij}/2<|\omega_{ij}^0|\leq\lambda_{ij})\leq\frac{1}{2^{1-q}}\sum_{j=1}^{p}|\omega_{ij}^0|^q\lambda_{ij}^{1-q},
\end{aligned}$$

and

$$T_6\leq\sum_{j=1}^{p}|\omega_{ij}^0|^q\lambda_{ij}^{1-q}.$$

Combining these pieces yields

$$T_1\leq 2\left(1+\frac{1}{2^{1-q}}\right)\sum_{j=1}^{p}|\omega_{ij}^0|^q\lambda_{ij}^{1-q}\leq 4\sum_{j=1}^{p}|\omega_{ij}^0|^q\lambda_{ij}^{1-q}.$$

We can similarly bound the terms T_2 and T_3 on B_1:

$$\begin{aligned}
T_2&\leq 2\sum_{j=1}^{p}\left(|\hat{\gamma}_{ij}-\omega_{ij}^0|+|\hat{\gamma}_{ij}|\right)I(|\hat{\gamma}_{ij}|\leq\lambda_{ij},|\omega_{ij}^0|>\lambda_{ij})\\
&\leq 2\sum_{j=1}^{p}\left(\frac{\lambda_{ij}}{2}+\lambda_{ij}\right)I(|\hat{\gamma}_{ij}|\leq\lambda_{ij},|\omega_{ij}^0|>\lambda_{ij})\leq 3\sum_{j=1}^{p}|\omega_{ij}^0|^q\lambda_{ij}^{1-q},
\end{aligned}$$

$$T_3 \leq \sum_{j=1}^{p} \left(|\hat{\gamma}_{ij} - \omega_{ij}^0| + |S_{\lambda_{ij}}(\hat{\gamma}_{ij}) - \hat{\gamma}_{ij}| + |S_{\lambda_{ij}}(\omega_{ij}^0) - \omega_{ij}^0| \right) I(|\hat{\gamma}_{ij}| > \lambda_{ij}, |\omega_{ij}^0| > \lambda_{ij})$$
$$\leq \sum_{j=1}^{p} \left(\frac{\lambda_{ij}}{2} + \lambda_{ij} + \lambda_{ij} \right) I(|\hat{\gamma}_{ij}| > \lambda_{ij}, |\omega_{ij}^0| > \lambda_{ij}) \leq \frac{5}{2} \sum_{j=1}^{p} |\omega_{ij}^0|^q \lambda_{ij}^{1-q}.$$

Collecting all terms, we obtain, on B_1,

$$\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}_0\|_1 \leq \frac{21}{2} \sum_{j=1}^{p} |\omega_{ij}^0|^q \lambda_{ij}^{1-q}. \tag{A.17}$$

Next, we consider the event B_2 on which $|\hat{\theta}_{ij} - \theta_{ij}| \leq \tau$ for all i, j. From Condition 3 we have, on B_2,

$$\hat{\theta}_{ij} \leq |\hat{\theta}_{ij} - \theta_{ij}| + \theta_{ij} \leq \tau + \theta_{ij} \leq 2\theta_{ij}. \tag{A.18}$$

Note that, by Condition 1,

$$\theta_{ij} \leq EY_i^2 Y_j^2 \leq \frac{1}{2}(EY_i^4 + EY_j^4) \leq \frac{2}{\alpha^2}. \tag{A.19}$$

Taking $\lambda_{ij} = \delta\sqrt{\hat{\theta}_{ij}}$ with $\delta = C_1\sqrt{(\log p)/n} + C_2 s_0(p)/p$ in (A.17) and applying (A.18) and (A.19), we obtain, on $B_1 \cap B_2$,

$$\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}_0\|_1 \leq \frac{21}{2} \sum_{j=1}^{p} |\omega_{ij}^0|^q \delta^{1-q} \left(\frac{2}{\alpha}\right)^{1-q} \leq \frac{21}{\alpha} s_0(p) \left(C_1 \sqrt{\frac{\log p}{n}} + C_2 \frac{s_0(p)}{p} \right)^{1-q}.$$

We conclude the proof by noting that the event $B_1 \cap B_2$ occurs with probability $1 - O(p^{-C_3})$ by Lemma 2 and that the spectral norm is bounded by the matrix L_1-norm.

A.5 Proof of Theorem 2

It follows from Condition (i) and (A.5) that

$$P\left(\hat{\omega}_{ij} \neq 0, \omega_{ij}^0 = 0 \text{ for some } i, j\right) \leq P\left(\max_{i,j} |\hat{\gamma}_{ij} - \omega_{ij}^0| \geq \lambda_{ij} \right)$$
$$= P\left(\max_{i,j} |\hat{\gamma}_{ij} - \omega_{ij}^0| / \sqrt{\hat{\theta}_{ij}} \geq C_1 \sqrt{\frac{\log p}{n}} + C_2 \frac{s_0(p)}{p} \right) = O(p^{-C_3}),$$

which proves (13).

To prove (15), note that, by Condition (ii),

$$P\left(\mathrm{sgn}(\hat{\omega}_{ij}) \neq \mathrm{sgn}(\omega_{ij}^0), \omega_{ij}^0 \neq 0 \text{ for some } i,j\right) \leq P\left(|\hat{\gamma}_{ij} - \omega_{ij}^0| \geq |\omega_{ij}^0| - \lambda_{ij} \text{ for some } i,j\right).$$

Also, by taking $\varepsilon = 3\tau/4$ in (A.4), we have, with probability $1 - O(p^{-C_3})$,

$$\left|\sqrt{\hat{\theta}_{ij}} - \sqrt{\theta_{ij}}\right| = \frac{|\hat{\theta}_{ij} - \theta_{ij}|}{\sqrt{\hat{\theta}_{ij}} + \sqrt{\theta_{ij}}} \leq \frac{3\tau/4}{\sqrt{\tau/4} + \sqrt{\tau}} = \frac{\sqrt{\tau}}{2},$$

and hence

$$\begin{aligned} |\omega_{ij}^0| - \lambda_{ij} &\geq C\delta\sqrt{\theta_{ij}} - \delta\left(\sqrt{\hat{\theta}_{ij}} - \sqrt{\theta_{ij}} + \sqrt{\theta_{ij}}\right) \\ &\geq (C-1)\delta\sqrt{\tau} - \delta\frac{\sqrt{\tau}}{2} = \left(C - \frac{3}{2}\right)\delta\sqrt{\tau} \end{aligned}$$

for all i, j. Now applying (A.15) yields

$$P\left(\mathrm{sgn}(\hat{\omega}_{ij}) \neq \mathrm{sgn}(\omega_{ij}^0), \omega_{ij}^0 \neq 0 \text{ for some } i,j\right) = O(p^{-C_3}),$$

which, together with (13), proves the result.

A.6 Proof of Theorem 3

By the argument of Bickel and Levina (2008, Section 3.3), it suffices to consider the procedure where the sample is randomly split into a training set of size n_1 and a test set of size n_2 with $n_1 \asymp n_2 \asymp n$. Denote by $\widehat{\mathbf{\Omega}}_1(\delta)$ the COAT estimator based on the training set and $\widehat{\mathbf{\Gamma}}_2$ the sample centered log-ratio covariance matrix based on the test set. Define the oracle tuning parameter $\hat{\delta}_* = \delta_{\hat{j}_*}$ with

$$\hat{j}_* = \arg\min_j \|\widehat{\mathbf{\Omega}}_1(\delta_j) - \mathbf{\Omega}_0\|_F^2.$$

It follows from Lemma 2 and the proof of Theorem 1 that

$$\begin{aligned} \|\widehat{\mathbf{\Omega}}_1(\hat{\delta}_*) - \mathbf{\Omega}_0\|_F^2 &\leq p\|\widehat{\mathbf{\Omega}}_1(\hat{\delta}_*) - \mathbf{\Omega}_0\|_{\max}\|\widehat{\mathbf{\Omega}}_1(\hat{\delta}_*) - \mathbf{\Omega}_0\|_1 \\ &= pO_p\left(\sqrt{\frac{\log p}{n}} + \frac{s_0(p)}{p}\right) O_p\left\{s_0(p)\left(\sqrt{\frac{\log p}{n}} + \frac{s_0(p)}{p}\right)^{1-q}\right\} \end{aligned}$$

$$= O_p \left\{ p s_0(p) \left(\sqrt{\frac{\log p}{n}} + \frac{s_0(p)}{p} \right)^{2-q} \right\}. \tag{A.20}$$

By the definition of $\hat{\delta}$, we have

$$\|\widehat{\mathbf{\Omega}}_1(\hat{\delta}) - \widehat{\mathbf{\Gamma}}_2\|_F^2 \leq \|\widehat{\mathbf{\Omega}}_1(\hat{\delta}_*) - \widehat{\mathbf{\Gamma}}_2\|_F^2,$$

or, after some algebra,

$$\|\widehat{\mathbf{\Omega}}_1(\hat{\delta}) - \mathbf{\Omega}_0\|_F^2 - \|\widehat{\mathbf{\Omega}}_1(\hat{\delta}_*) - \mathbf{\Omega}_0\|_F^2 \leq 2\langle \widehat{\mathbf{\Omega}}_1(\hat{\delta}) - \widehat{\mathbf{\Omega}}_1(\hat{\delta}_*), \widehat{\mathbf{\Gamma}}_2 - \mathbf{\Omega}_0 \rangle.$$

To bound the right-hand side, note that, by Lemma A.3 of Bickel and Levina (2008),

$$E \max_j |\langle \mathbf{V}, \widehat{\mathbf{\Gamma}}_2 - \mathbf{\Gamma}_0 \rangle| = O\big(\sqrt{p/n}\big)$$

for any $p \times p$ matrix $\mathbf{V}$ with $\|\mathbf{V}\|_F = 1$. Also, by Proposition 1,

$$\|\mathbf{\Gamma}_0 - \mathbf{\Omega}_0\|_F \leq p \|\mathbf{\Gamma}_0 - \mathbf{\Omega}_0\|_{\max} = O(s_0(p)).$$

Combining these two pieces yields

$$E \max_j |\langle \mathbf{V}, \widehat{\mathbf{\Gamma}}_2 - \mathbf{\Omega}_0 \rangle| = O\big(\sqrt{p/n} + s_0(p)\big).$$

Therefore, letting $\|\widehat{\mathbf{\Omega}}_1(\hat{\delta}) - \mathbf{\Omega}_0\|_F = a_n$ and $\|\widehat{\mathbf{\Omega}}_1(\hat{\delta}_*) - \mathbf{\Omega}_0\|_F = r_n$, we obtain

$$a_n^2 - r_n^2 \leq O_p\big(\sqrt{p/n} + s_0(p)\big)(a_n + r_n),$$

or

$$a_n \leq r_n + O_p\big(\sqrt{p/n} + s_0(p)\big).$$

This, together with (A.20) and the assumption (16), implies

$$a_n^2 = O_p \left\{ p s_0(p) \left(\frac{\log p}{n} \right)^{1-q/2} \right\},$$

which concludes the proof.

Acknowledgments

The authors thank Professor Hongmei Jiang and Dr. Huaying Fang for sharing R code and the Associate Editor and two reviewers for helpful comments.

Funding

Cao and Li's research was supported in part by NIH grants CA127334 and GM097505. Lin's research was supported in part by NSFC grants 11671018 and 71532001.

References

Aitchison, J. (1982), "The Statistical Analysis of Compositional Data" (with discussion), *Journal of the Royal Statistical Society*, Series B, 44, 139–177.

——— (2003), *The Statistical Analysis of Compositional Data*, Caldwell, NJ: Blackburn Press.

Aitchison, J., and Shen, S. M. (1980), "Logistic-Normal Distributions: Some Properties and Uses," *Biometrika*, 67, 261–272.

Anderson, T. W. (2003), *An Introduction to Multivariate Statistical Analysis* (3rd ed.), New York: Wiley.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H. B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E. G., Wang, J., Guarner, F., Pedersen, O., de Vos, W. M., Brunak, S., Doré, J., MetaHIT Consortium, Weissenbach, J., Ehrlich, S. D., and Bork, P. (2011), "Enterotypes of the Human Gut Microbiome," *Nature*, 473, 174–180.

Ban, Y., An, L., and Jiang, H. (2015), "Investigating Microbial Co-Ocurrence Patterns Based on Metagenomic Compositional Data," *Bioinformatics*, 31, 3322–3329.

Bickel, P. J., and Levina, E. (2008), "Covariance Regularization by Thresholding," *The Annals of Statistics*, 36, 2577–2604.

Boucheron, S., Lugosi, G., and Massart, P. (2013), *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford: Oxford University Press.

Cai, T., and Liu, W. (2011), "Adaptive Thresholding for Sparse Covariance Matrix Estimation," *Journal of the American Statistical Association*, 106, 672–684.

Cai, T. T., and Zhou, H. H. (2012), "Optimal Rates of Convergence for Sparse Covariance Matrix Estimation," *The Annals of Statistics*, 40, 2389–2420.

Coyte, K. Z., Schluter, J., and Foster, K. R. (2015), "The Ecology of the Microbiome: Networks, Competition, and Stability," *Science*, 350, 663–666.

Dunne, J. A., Williams, R. J., and Martinez, N. D. (2002), "Food-Web Structure and Network Theory: The Role of Connectance and Size," *Proceedings of the National Academy of Sciences*, 99, 12917–12922.

El Karoui, N. (2008), "Operator Norm Consistent Estimation of Large-Dimensional Sparse Covariance Matrices," *The Annals of Statistics*, 36, 2717–2756.

Fan, J., Fan, Y., and Lv, J. (2008), "High Dimensional Covariance Matrix Estimation Using a Factor Model," *Journal of Econometrics*, 147, 186–197.

Fan, J., Liao, Y., and Mincheva, M. (2013), "Large Covariance Estimation by Thresholding Principal Orthogonal Complements" (with discussion), *Journal of the Royal Statistical Society*, Series B, 75, 603–680.

Fang, H., Huang, C., Zhao, H., and Deng, M. (2015), "CCLasso: Correlation Inference for Compositional Data Through Lasso," *Bioinformatics*, 31, 3172–3180.

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.

Fisher, C. K., and Mehta, P. (2014), "Identifying Keystone Species in the Human Gut Microbiome From Metagenomic Timeseries Using Sparse Linear Regression," *PLoS ONE*, 9, e102451.

Friedman, J., and Alm, E. J. (2012), "Inferring Correlation Networks From Genomic Survey Data," *PLoS Computational Biology*, 8, e1002687.

Greenblum, S., Turnbaugh, P. J., and Borenstein, E. (2012), "Metagenomic Systems Biology of the Human Gut Microbiome Reveals Topological Shifts Associated With Obesity and Inflammatory Bowel Disease," *Proceedings of the National Academy of Sciences*, 109, 594–599.

Isserlis, L. (1918), "On a Formula for the Product-Moment Coefficient of Any Order of a Normal Frequency Distribution in Any Number of Variables," *Biometrika*, 12, 134–139.

Jordán, F., Lauria, M., Scotti, M., Nguyen, T.-P., Praveen, P., Morine, M., and Priami, C. (2015), "Diversity of Key Players in the Microbial Ecosystems of the Human Body," *Scientific Reports*, 5, 15920.

Koeth, R. A., Wang, Z., Levison, B. S., Buffa, J. A., Org, E., Sheehy, B. T., Britt, E. B., Fu, X., Wu, Y., Li, L., Smith, J. D., DiDonato, J. A., Chen, J., Li, H., Wu, G. D., Lewis, J. D., Warrier, M., Brown, J. M., Krauss, R. M., Tang, W. H. W., Bushman, F. D., Lusis, A. J., and Hazen, S. L. (2013), "Intestinal Microbiota Metabolism of L-Carnitine, a Nutrient in Red Meat, Promotes Atherosclerosis," *Nature Medicine*, 19, 576–585.

Lewis, J. D., Chen, E. Z., Baldassano, R. N., Otley, A. R., Griffiths, A. M., Lee, D., Bittinger, K., Bailey, A., Friedman, E. S., Hoffmann, C., Albenberg, L., Sinha, R., Compher, C., Gilroy, E.,

Nessel, L., Grant, A., Chehoud, C., Li, H., Wu, G. D., and Bushman, F. D. (2015), "Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease," *Cell Host & Microbe*, 18, 489–500.

Li, H. (2015), "Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis," *Annual Review of Statistics and Its Application*, 2, 73–94.

Limpert, E., Stahel, W. A., and Abbt, M. (2001), "Log-Normal Distributions Across the Sciences: Keys and Clues," *BioScience*, 51, 341–352.

Rejmánek, M., and Starý, P. (1979), "Connectance in Real Biotic Communities and Critical Values for Stability of Model Ecosystems," *Nature*, 280, 311–313.

Rothman, A. J., Levina, E., and Zhu, J. (2009), "Generalized Thresholding of Large Covariance Matrices," *Journal of the American Statistical Association*, 104, 177–186.

The Human Microbiome Project Consortium (2012), "A Framework for Human Microbiome Research," *Nature*, 486, 215–221.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., Affourtit, J. P., Egholm, M., Henrissat, B., Heath, A. C., Knight, R., and Gordon, J. I. (2009), "A Core Gut Microbiome in Obese and Lean Twins," *Nature*, 457, 480–484.

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006), "An Obesity-Associated Gut Microbiome With Increased Capacity for Energy Harvest," *Nature*, 444, 1027–1031.

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F. D., and Lewis, J. D. (2011), "Linking Long-Term Dietary Patterns With Gut Microbial Enterotypes," *Science*, 334, 105–108.

Yodzis, P. (1980), "The Connectance of Real Ecosystems," *Nature*, 284, 544–545.