Statistica Sinica Preprint No: SS- 2021-0147						
Title	Robust Shape Matrix Estimation for High-Dimensional					
	Compositional Data with Application to Microbial					
	Inter-Taxa Analysis					
Manuscript ID	SS-2021-0147					
URL	http://www.stat.sinica.edu.tw/statistica/					
DOI	10.5705/ss.202021.0147					
Complete List of Authors	Danning Li,					
	Arun Srinivasan,					
	Lingzhou Xue and					
	Xiang Zhan					
<b>Corresponding Authors</b>	Lingzhou Xue					
	Xiang Zhan					
E-mails	lzxue@psu.edu					
	zhanx@bjmu.edu.cn					
Notice: Accepted version subject to English editing.						

\* Danning Li and Arun Srinivasan contributed equally to this work. Lingzhou Xue and Xiang Zhan are co-corresponding authors. Statistica Sinica

# Robust Shape Matrix Estimation for High-Dimensional Compositional Data with Application to Microbial Inter-Taxa Analysis

Danning Li<sup>1</sup>, Arun Srinivasan<sup>2</sup>, Lingzhou Xue<sup>2</sup> and Xiang Zhan<sup>3</sup>

<sup>1</sup>KLAS and School of Mathematics & Statistics, Northeast Normal University
 <sup>2</sup>Department of Statistics, Pennsylvania State University
 <sup>3</sup>Department of Biostatistics, School of Public Health and BICMR, Peking University

Abstract:

Estimating dependence structure is a key task in compositional data analysis. Real-world compositional datasets are often complex due to factors such as high-dimensionality, heavy tails, and existence of possible outliers. We consider a general class of elliptical distributions to model the heavy-tailed distribution of latent log-basis variables, which is characterized by the latent shape matrix. The latent shape matrix is a scalar multiple of the latent covariance matrix when it exists, and it can preserve the directional properties of the dependence in a distribution when the covariance matrix does not exist. We propose a robust composition-adjusted thresholding procedure based on Tyler's M-estimator (Tyler, 1987) for jointly estimating the latent shape matrices of high-dimensional compositional data from different groups. We prove appealing theoretical properties under the high-dimensional setting. Simulation studies and a real application to microbial inter-taxa analysis are used to demonstrate the numerical properties.

*Key words and phrases:* Elliptical distribution, Compositional data; Human microbiome research, Shape matrix, Thresholding, Tyler's M Estimation.

## 1. Introduction

Compositional data naturally arise in a large number of research topics in biology, ecology, finance, geology, and many others. For example, compositional data are used to assess the relative proportions of chemicals within stones across different geographical locations in geology (Thomas and Aitchison, 2005) or analyze relative market share while dynamically accounting for the total market size in economics (Arata and Onozaki, 2017). This paper is motivated by the inter-taxa analysis of microbiome compositional data in the rapidly growing field of human microbiome research (Cho and Blaser, 2012). It is known that the accurate estimation of the dependence structure between bacteria leads to a better understanding of the underlying mechanisms of disease development (Friedman and Alm, 2012; Goodrich et al., 2014). We focus on the estimation of the dependence structure of high-dimensional compositional data, which is a fundamental problem in microbial inter-taxa analysis.

Sparse estimation of large covariance or correlation matrices studies the dependence structure under the assumption that only a small proportion of variables are correlated with one another. See Bickel and Levina (2008), Rothman et al. (2009), Cai and Liu (2011), Bien and Tibshirani (2011), Cai et al. (2012), Rothman (2012), Xue et al. (2012), Xue and Zou (2014a), Fan et al. (2016), Bien (2019), among others, for related papers in this topic. However, it is non-trivial to estimate the sparse covariance matrix of high-dimensional compositional data, whose data matrix induces a sum-to-one constraint for each row and is intrinsically not full rank. The compositional data analysis framework proposed by Aitchison (1986) lays the bedrock for the study of the dependency structure of compositional data. Under this latent framework, the Sparse Correlations for Compositional Data (sparCC) method proposed by Friedman and Alm (2012) employed an iterative algorithm to estimate the correlation matrix. The sparCC method does not guarantee that the estimator is positivedefinite or that correlations are bounded between [-1, 1]. To solve these issues, Fang et al. (2015) developed the Correlation Inference for Compositional Data through Lasso (CCLasso) method which employs an  $\ell_1$ -penalization to estimate a sparse representation of the latent correlation matrix and ensure the positivedefiniteness. Recently, work by Cao et al. (2019) introduced the composition adjusted thresholding (COAT) procedure to estimate the sparse covariance matrix of compositional data under a latent data framework.

Existing covariance or correlation estimation procedures for compositional data, including the aforementioned sparCC, CCLasso, and COAT, are essen-

tially built on the regularized sample covariance matrix. They all assume the sub-Gaussian tails or even normal distributions for the latent log-basis variables (Cao et al., 2019). However, it is known that the sample covariance matrix behaves poorly when the data significantly deviates from normality (Tyler, 1987; Nordhausen and Tyler, 2015). Due to machine error, natural variation, or experimental procedure, real-world compositional datasets are riddled with outliers and heavy tails.

Another key rationale for developing our method is to create a method which naturally fits into a practical data analysis pipeline. While classic estimation of a single dependency structure is common, researchers often wish to compare across multiple cohorts simultaneously (Morgan et al., 2015). This motivates a joint analysis viewpoint, as the estimation of multiple dependence structures provides additional insights when we study the compositions across multiple groups. For example, one may be interested in learning how the dependence of the microbiome changes under the presence or absence of antibiotics regimes, or between disease cohorts. However, across cohorts, there may be key dependencies that remain the same regardless of clinical covariate. Therefore, joint estimation is useful as it shares information across groups in order to improve accuracy.

Our primary contribution is the development of the HeavyCOAT procedure,

which provides an accurate joint estimation of the orientation and spread of the underlying elliptical contours even when the covariance matrix may not exist. We focus on estimating the dependence structure from the shape matrix in wide class of elliptical distributions. We propose a robust composition-adjusted thresholding procedure, called HeavyCOAT, to estimate the sparse latent shape matrix of high-dimensional compositional data under the general elliptical distribution framework. The HeavyCOAT procedure first estimates the shape matrix of the transformed data  $X_c$  across each cohort of interest and then uses the estimated shape matrices to construct a sparse estimation of the latent shape matrices of each row of  $\mathbb{Y}$  for each cohort by solving a positive-definite thresholding problem. By using either a fused or group penalty (Tibshirani et al., 2005; Friedman et al., 2010; Danaher et al., 2014) in the positive-definite thresholding step, the HeavyCOAT method can be used to jointly estimate multiple sparse latent shape matrices when the observed compositions come from different groups and their shape matrices share certain similarity.

Theoretically, we study the asymptotic behaviours of the proposed method. Primarily, we show that the HeavyCOAT procedure is a robust estimator that effectively recovers the sparsity pattern of the dependence structure and achieves the sign consistency with high probability under the broad class of elliptical distributions. Also, we derive the convergence rate under the spectral norm for the estimation of sparse latent shape sub-matrix as well as the explicit expected risk bound under the squared spectral norm. The derived convergence rate and risk bound are comparable to those optimal results under Gaussian or sub-Gaussian tail conditions. The finite-sample properties of the HeavyCOAT procedure and a real application to microbial inter-taxa analysis are also demonstrated.

The rest of this paper is organized as follows. After introducing a general class of elliptical distributions and notation in Section 2, we present the Heavy-COAT procedure in Section 3. Section 4 presents the asymptotic properties including the selection consistency, sign consistency, convergence rate, and risk bound. We evaluate the finite-sample properties through simulation studies in Section 5 and a real application to microbial inter-taxa analysis in Section 6. We give the concluding remarks in Section 7. Technical proofs and additional simulations are presented in a supplemental note.

# 2. Preliminaries

## 2.1 Problem Setting

The challenges of compositional data analysis arise from the transformation of the unconstrained features to the compositional space (Aitchison, 1986). We will first introduce the notation under the single group setting and the latent view of compositional data analysis popularized by Cao et al. (2019). Let  $\mathbf{W}_0 \in \mathbb{R}^{p_0}_+$  be the vector of basis variables and  $\mathbf{Y}_0 = \log(\mathbf{W}_0)$  be the vector of latent log-basis variables. The corresponding compositional random vector  $\mathbf{X}_{0,i} = (X_{0,i,1}, \dots, X_{0,i,p_0})'$  is generated by normalizing  $X_{0,i} = W_{0,i} / \sum_{i=1}^{p_0} W_{0,i}$ . Often, we only observe the compositional data matrix  $\mathbb{X}_0 = (\mathbf{X}_{0,1}, \dots, \mathbf{X}_{0,n})'$  in practice, instead of the basis matrix  $\mathbb{W}_0 = (\mathbf{W}_{0,1}, \dots, \mathbf{W}_{0,n})'$ , where *n* denotes the number of observations. For example, in microbiome research, raw DNA totals vary greatly between samples and thus the relative proportions  $\mathbb{X}_0$  are reported (Li, 2015). To account for the compositional structure of  $\mathbb{X}_0$ , the centered log-ratio (clr) transformation is used as the pre-processing step on  $\mathbb{X}_0$ . Specifically, we have  $\mathbb{X}_c = \operatorname{clr}(\mathbb{X}_0) = (\log(X_{0,i,j}/g(\mathbf{X}_{0,i})))$  for i = 1, ..., n and j = $1, ..., p_0$ , where  $g(\mathbf{X}_{0,i})$  is the geometric mean of  $\mathbf{X}_{0,i} = (X_{0,i,1}, \dots, X_{0,i,p_0})'$ . The goal of this paper is to use observations  $\mathbf{X}_{0,i}$  where i = 1, ..., n to make inference on the dependence structure of  $\mathbf{Y}_0 = (\mathbf{Y}_{0,1}, \ldots, \mathbf{Y}_{0,n})'$  where i =1, ..., n which is the true dependence relationship we wish to capture.

To allow for heavy-tails or possible outliers, we assume that each  $\mathbf{Y}_{0,i} \in \mathbb{R}^{p_0}$  follows the elliptical distribution such that  $\mathbf{Y}_{0,i} \sim \mathcal{E}_{p_0}(\mu, \Sigma_0, \phi)$ , where  $\mathcal{E}_{p_0}(\mu, \Sigma_0, \phi)$  is a  $p_0$ -dimensional elliptical distribution with the location parameter  $\mu \in \mathbb{R}^{p_0}$ , positive-definite shape matrix  $\Sigma_0$ , and density generator  $\phi$  (Cambanis et al., 1981; Tyler, 1987; Fang et al., 1990). By Cambanis et al. (1981),  $\mathbf{Y}_{0,i}$  is equivalently represented as

# 2.1 Problem Setting

$$\mathbf{Y}_{0,i} = \mu + u_i (\Sigma_0)^{1/2} \xi_i,$$

where  $\xi_i$  is drawn uniformly from  $\mathbb{S}^{p_0-1}$ , and  $u_i$  is an arbitrary random variable or deterministic nonzero scalar, independent of  $\xi_i$ . Note,  $\mathcal{E}_{p_0}(\mu, \Sigma_0, \phi)$ consists of a large class of distributions with elliptically shaped contours, including the Gaussian distribution and also the heavy-tailed distributions such as the Laplace and Cauchy distributions, which are often used to efficiently generate data from the elliptical distribution (Andrews and Mallows, 1974; Goes et al., 2020; Müller and Richter, 2019). Let  $u_i$  be a scalar random variable, and  $\xi_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{p_0 \times p_0})$ . The Gaussian scale mixture random variable  $\mathbf{Y}_{0,i} =$  $\mu + u_i(\Sigma_0)^{1/2}\xi_i$  is a useful elliptical distribution (Goes et al., 2020). For ease of representation, we refer to these distributions according the form of  $u_i$ , for instance,  $\mathbf{Y}_{0,i}$  follows a Laplace scale mixture distribution when  $u_i$  is generated from a Laplace distribution. The underlying dependence structure is characterized by the shape matrix. If the covariance matrix exists, the shape matrix is proportional to the covariance matrix. When the covariance matrix does not exist (e.g. Cauchy distribution), the shape matrix is still a reliable measure of directional dependence (Tyler, 1987; Nordhausen and Tyler, 2015).

## 2.2 Real Data Motivation

As a practical illustration of this problem setting, we use the mucosal membrane dataset collected by Morgan et al. (2015) to illustrate the non-normality and heavy-tails in the microbiome compositional data.We conduct several normality tests (e.g., the Cramer-von Mises test (Cramér, 1928), Lilliefors test (Conover, 1998), and Shapiro-Francia test (Shapiro and Francia, 1972)) to check if each column of  $Y_0$  is normally distributed with an  $\alpha$ -level threshold 0.05 or the Bonferroni-adjusted  $\alpha$ -level threshold. The results of these tests is summarized in Table 1. Across all normality tests, a vast majority of the taxa fail to satisfy the assumed normal distribution, suggesting the presence of non-normal behavior.

Table 1: Number of rejections of three different tests of normality among the 200 most abundant taxa. Each test is evaluated at a threshold of 0.05 and the Bonferroni-adjusted threshold.

	Test of normality for each taxa							
	Method	Threshold	Number of Rejections					
C	ramer-von Mises	.05	200					
		.05/200	197					
	Lilliefors	.05	199					
		.05/200	198					
·	Shaprio-Francia	.05	199					
		.05/200	198					

#### 2.3 The Shape Matrix

To study the dependence structure of compositional data, we consider a general class of elliptical distributions characterized by the shape matrix (Cambanis et al., 1981; Fang et al., 1990). It is known that the shape matrix is a scalar multiple of the covariance matrix when it exists, and it can preserve the directional properties of the dependence in a distribution without requiring the existence of moments (Tyler, 1987; Nordhausen and Tyler, 2015; Wiesel and Zhang, 2015). This is particularly useful in microbiome analysis as the shape matrix allows us to recover the linear relationship between microbial taxa, even in extremely noisy practical data settings. In these practical settings, the error within the data may not fit the classical Gaussian assumption. For example, if the (i, j) element of the shape matrix is positive, this indicates that taxa *i* is positively, linearly, associated with taxa *j*. Therefore, as the abundance of taxa *i* increases, the abundance of taxa *j* increases. If the (i, j) element of the shape matrix is 0, then there is no association between the abundances of taxa *i* and taxa *j*.

In view of these appealing properties, the shape matrix can be viewed as a promising alternative to the covariance matrix for heavy-tailed distributions. The shape matrix can be accurately estimated even when the second moment is not bounded (Tyler, 1987). We use a simulation study to explore the interpretation of the shape matrix and also the impact of heavy-tails. We generated n = 500 in-

dependent bivariate realizations from a Gaussian scale mixture distribution. We provide more insight into scale mixture distributions in Section 2. For this toy example, we considered three different choices of  $u_i$  corresponding to: the multivariate normal distribution with  $\{u_i\}_{i=1}^n = 1$ , the Laplace scale mixture with  $u_i \sim \text{Laplace}(0, 1)$  and the Cauchy scale mixture with  $u_i \sim \text{Cauchy}(0, 1/100)$ . The diagonal elements of the shape matrix were fixed as 1 and we varied the off-diagonal dependence strength  $C \in \{-.5, 0, .5\}$ . In the first two settings, the covariance matrix existed and C coincided with the correlation.

We compared four different measures of bivariate dependence in this simulation study: Pearson correlation coefficient (denoted by P), Kendall correlation coefficient (denoted by K), Spearman correlation coefficient (denoted by S), and the normalized Tyler's M-estimator (Tyler, 1987) of the shape matrix (denoted by T). As shown in Table 2, all four measures perform well in the Gaussian setting. Tyler's M-estimation and rank correlations can capture the underlying dependence under all three settings, while the Pearson correlation coefficient performs poorly in both Laplace and Cauchy settings due to heavy-tails. Overall, Tyler's M-estimation provides a more accurate estimation than both rank correlations and the Pearson correlation estimates, which motivates us to propose new methods for robust estimation of the shape matrix  $\Sigma_0$  of latent variables  $\mathbb{Y}_0$ . Table 2: Illustration of the shape matrix measuring dependence. The estimated correlation is listed for a given true correlation C under each scale mixture.

			0							
Method	Gaussian			]	Laplace			Cauchy		
C	-0.5	0	0.5	-0.5	0	0.5	-0.5	0	0.5	
T	-0.48	0.04	0.42	-0.44	-0.01	0.51	-0.50	0.08	0.48	
K	-0.34	0.02	0.30	-0.30	0.00	0.37	-0.37	0.03	0.33	
S	-0.49	0.04	0.44	-0.39	0.00	0.50	-0.48	0.02	0.42	
P	-0.50	0.03	0.45	-0.38	0.03	0.35	-0.89	-0.95	0.93	

#### 3. Methodology

This section proposes a robust positive-definite estimation of sparse shape matrices for compositional data. The observed compositions may come from different cohorts, and their shape matrices may share the similarity. Suppose that there are K independent compositional datasets  $\mathbb{X}_{0,k} = (\mathbf{X}_{0,k,1}, \dots, \mathbf{X}_{0,k,n_k})' \in \mathbb{R}^{n_k \times p_0}$ for  $k = 1, 2, \dots, K$ , and  $\mathbb{Y}_{0,k} = (\mathbf{Y}_{0,k,1}, \dots, \mathbf{Y}_{0,k,n_k})' \in \mathbb{R}^{n_k \times p_0}$  are the corresponding latent log-basis variables. For any  $p < p_0$ , the proposed method estimates the  $p \times p$  sub-matrix of the true shape matrix of  $\mathbb{Y}_{0,k}$ . The estimation of a sub-matrix is not overly restrictive as it is typical to screen out taxa with low counts or low accuracy prior to running an analysis in practice. Further, when  $p = p_0 - 1$ , this is analogous to dropping the reference component whose dependency structure is not of interest to a researcher. We employ this in practice in the microbial inter-taxa analysis of Section 6, we drop the column of the OTU matrix referring to the counts of *unclassified* microbes. As this OTU encompasses all taxa that were not matched to known taxa of interest, the dependency captured by this OTU is likely uninformative in understanding the underlying biological mechanism and we still recover the vast majority of dependencies between taxa.

Let  $\mathbb{X}_{k,c} = (\mathbf{X}_{k,c,1}, \dots, \mathbf{X}_{k,c,n_k})' \in \mathbb{R}^{n_k \times p}$  be the sub-matrix of transformed data  $\operatorname{clr}(\mathbb{X}_{0,k})$ , and the columns of  $\operatorname{clr}(\mathbb{X}_{0,k})$  associated with this sub-matrix be enumerated in the index set  $\mathcal{C} \subset \{1, \dots, p_0\}$ . Without loss of generality, we assume that  $\mathcal{C} = \{1, \dots, p\}$  (i.e. we drop the  $p_0^{th}$  component). Let  $A_0 = \mathbf{I}_{p_0 \times p_0} - (1/p_0)\mathbf{1}_{p_0}\mathbf{1}_{p_0}^T$  and  $A = (\mathbf{I}_{p \times p}, \mathbf{0}_{p \times (p_0 - p)}) - (1/p_0)\mathbf{1}_p\mathbf{1}_{p_0}^T$ . Lemma 1 is the key bedrock of our analysis showing how elliptical distributions are retained through transformation.

**Lemma 1.** Suppose each  $\mathbf{Y}_{0,k,i} \in \mathbb{R}^{p_0}$  in  $\mathbb{Y}_{0,k}$  follows the elliptical distribution  $\mathcal{E}_{p_0}(\mu_k, \Sigma_{0,k}, \phi_k)$ , where rank $(\Sigma_{0,k}) = p_0$ . Then, each  $clr(\mathbf{X}_{0,k,i})$  in  $clr(\mathbb{X}_{0,k})$  follows the elliptical distribution  $\mathcal{E}_{p_0-1}(A_0\mu_k, A_0\Sigma_{0,k}A_0^T, \phi_k)$ , and each  $\mathbf{X}_{k,c,i}$  in  $\mathbb{X}_{k,c}$  follows the elliptical distribution  $\mathcal{E}_p(A\mu_k, A\Sigma_{0,k}A^T, \phi_k)$ . Moreover, we have  $rank(A_0\Sigma_{0,k}A_0^T) = p_0 - 1$ .

As shown in Lemma 1, the distributions of both  $\mathbf{X}_{0,k,i}$  and  $\operatorname{clr}(\mathbf{X}_{0,k,i})$  fall in the class of elliptical distributions. Further,  $\operatorname{rank}(A_0\Sigma_{0,k}A_0^T) = p_0 - 1$  and  $\operatorname{rank}(A\Sigma_{0,k}A^T) = p$ . While the shape matrix of  $\operatorname{clr}(\mathbb{X}_{0,k})$ , i.e.,  $A_0\Sigma_{0,k}A_0^T \in \mathbb{R}^{p_0 \times p_0}$ , is degenerate, after removing columns relating to unimportant taxa, the shape sub-matrix of  $\mathbf{X}_{k,c,i}$ , i.e.,  $A\Sigma_{0,k}A^T \in \mathbb{R}^{p \times p}$ , is non-degenerate. For ease of terminology, we will refer to  $A\Sigma_{0,k}A^T$  as a "shape sub-matrix", which is a sub-matrix of the true shape matrix  $\Sigma_{0,k}$ . By Proposition 1 of Cao et al. (2019), the shape sub-matrix is asymptotically indistinguishable from  $\Sigma_{0,k}$  when  $\Sigma_{0,k}$ belongs to the class of sparse shape matrices explored in Section 4. Under this condition, the sparsity  $A\Sigma_{0,k}A^T$  is asymptotically equivalent to the sparsity pattern of  $\Sigma_{0,k}$ , thus allowing the shape sub-matrix to function as a proxy for latent log-basis  $\Sigma_{0,k}$ . Thus, this asymptotic indistinguishability forms the bedrock to the support recovery and convergence properties explored in Section 4.

Recall that the shape matrix characterizes the linear relationships between variables under the elliptical distribution framework in Section 2. Let  $\Sigma_{0,k}$  be the sparse shape sub-matrix for those latent log-basis variables  $\mathbb{Y}_{0,k}$  in the *k*-th group. Let  $\Gamma_k$  be the corresponding shape sub-matrix for transformed variables in  $\mathbb{X}_{k,c}$ . While  $\Sigma_{0,k}$  and  $\Gamma_k$  appear to be different estimation targets, as show by Proposition 1 of Cao et al. (2019), the shape sub-matrix is asymptotically indistinguishable from the true log-basis shape sub-matrix under a weak set of conditions. This relationship is a key fact that motivates the theoretical recovery properties we explore in Section 4. Under the elliptical distribution framework, we propose the HeavyCOAT procedure to first compute a robust estimation of the shape sub-matrix  $\Gamma_k$  of transformed variables in  $\mathbb{X}_{k,c}$  (see Subsection 3.1) and then obtain the sparse estimation of the latent shape sub-matrix of log-basis variables in  $\mathbb{Y}_{0,k}$  (see Subsection 3.2). The HeavyCOAT procedure first uses the Tyler's M-estimation method (Tyler, 1987) to construct a robust shape sub-matrix estimator  $\hat{\Gamma}_k$  based on the transformed data matrix  $\mathbb{X}_{k,c}$  and then obtains the final positive-definite and sparse shape sub-matrix estimator  $\hat{\Sigma}_k$  through a subsequent joint thresholding step. Since the shape matrix is scale invariant, we may assume that  $tr(\Gamma_k) = p$  without loss of generality (Tyler, 1987; Goes et al., 2020). We should point out that, under the K = 1 setting we drop the k-subscript and the sparse estimator  $\hat{\Sigma}$ , constructed using  $\mathbb{X}_c$  as a proxy of the latent  $\mathbb{Y}_0$ , enjoys desirable theoretical properties including the selection consistency and sign consistency, which will be presented in Section 4.

## 3.1 Estimation Step

For the k-th group, to construct the robust shape sub-matrix estimator  $\hat{\Gamma}_k$  of the observed data matrix  $\mathbb{X}_{k,c}$ , we follow Tyler (1987) to solve the constrained optimization problem over the space of all positive-definite matrices satisfying  $\operatorname{tr}(\Gamma_k) = p$ , that is,

$$\min_{\Gamma_k: \text{ tr}(\Gamma_k)=p} \frac{p}{n_k} \sum_{i=1}^{n_k} \log((\mathbf{X}_{k,c,i})^T \Gamma_k^{-1} \mathbf{X}_{k,c,i}) + \log(\det(\Gamma_k)).$$
(3.1)

Optimization problem (3.1) does not require the existence of the covariance matrix and it is also agnostic to the exact functional form of elliptical distributions rather than a specific member of the class (Tyler, 1987). Hence, it can be applied to estimate the shape sub-matrix in the wide class of elliptical distributions. Although (3.1) is not a convex problem, the objective function enjoys geodesic convexity (Duembgen and Tyler, 2016).

Sun et al. (2015) and Goes et al. (2020) introduced an iterative algorithm to solve the high-dimensional Tyler's M-estimation problem. Specifically, starting from  $\tilde{\Gamma}_k^{(1)} = \alpha_k/(1 + \alpha_k) \cdot \mathbf{I}_{p \times p}$  with  $\alpha_k > \max(0, p/n_k - 1)$ , for  $t = 1, 2, \cdots$ , we solve

$$\tilde{\Gamma}_{k}^{(t+1)} = \frac{1}{1+\alpha_{k}} \frac{p}{n_{k}} \sum_{i=1}^{n_{k}} \frac{\mathbf{X}_{k,c,i} (\mathbf{X}_{k,c,i})^{T}}{(\mathbf{X}_{k,c,i})^{T} (\tilde{\Gamma}_{k}^{(t)})^{-1} \mathbf{X}_{k,c,i}} + \frac{\alpha_{k}}{1+\alpha_{k}} \mathbf{I}_{p \times p}.$$
(3.2)

At each iteration, the diagonal element of the current estimate is increased by  $\alpha_k$ to ensure the positive definiteness, where the accuracy of the estimator is robust to the choice of  $\alpha_k$ . It is known that the iterative algorithm attains a unique solution when  $\alpha_k > \max(0, p/n_k - 1)$  (Pascal et al., 2013; Goes et al., 2020) and  $\alpha_k$  primarily controls the speed at which the algorithm converges. Thus, we can define the iterated solution for the *k*-th group as  $\tilde{\Gamma}_k$  and also its normalization,  $\hat{\Gamma}_k$ , according to the trace constraint as

$$\hat{\Gamma}_{k} = \frac{p(\tilde{\Gamma}_{k} - \frac{\alpha_{k}}{1 + \alpha_{k}} \mathbf{I}_{p \times p})}{\operatorname{tr}(\tilde{\Gamma}_{k} - \frac{\alpha_{k}}{1 + \alpha_{k}} \mathbf{I}_{p \times p})}$$
(3.3)

Proposition 18 of Sun et al. (2014) shows that the trace normalized solution converges to the desired global minimum.

### 3.2 Thresholding Step

After obtaining the robust estimator  $\hat{\Gamma}_k$  for each class of interest, we use a positive-definite thresholding step (Rothman, 2012; Xue et al., 2012; Bien, 2019) to derive the sparse shape sub-matrix estimator  $\hat{\Sigma}_k$  of the latent log-basis variable  $\mathbb{Y}_{0,k}$  for each class by solving the following objective,

$$\hat{\Sigma}_{k} = \operatorname*{argmin}_{\operatorname{each} \Sigma_{k} \succeq \varepsilon \mathbf{I}_{p \times p}} \frac{1}{2} \sum_{k=1}^{K} ||\Sigma_{k} - \hat{\Gamma}_{k}||_{F}^{2} + P(\{\Sigma_{k}\}), \qquad (3.4)$$

where  $\|\cdot\|_F$  denotes the Forbenius norm,  $\{\Sigma_k\}$  is the set of estimated shape matrices for each group k = 1, 2, ..., K, and  $P(\cdot)$  denotes a convex and nonsmooth penalty function of interest. Here,  $\varepsilon > 0$  ensures the positive-definite constraint and in practice can be taken to be a sufficiently small number and its choice does not affect the accuracy. The constrained optimization problem (3.4) is convex and can be efficiently solved by an alternating direction method of multipliers (ADMM) that is presented in Section A1 of the Supplement.

When K = 1 and there is no cohort information to be shared and we may use the  $\ell_1$ -penalty to penalize the off-diagonal elements of  $\hat{\Sigma}$  to yield a sparse estimate. For ease of notation, we drop the subscript k in this special case when estimating  $\hat{\Sigma}$  and  $\hat{\Gamma}$ . Specifically, we use  $P(\hat{\Sigma}) = \lambda \|\hat{\Sigma}\|_{1,\text{off}}$ , where  $\lambda > 0$  is a tuning parameter and  $\|\cdot\|_{1,\text{off}}$  denotes the entry-wise  $\ell_1$  norm of the off-diagonal elements of  $\hat{\Sigma}$ . We present theoretical results including the selection consistency, sign consistency, convergence rate and risk bound for this setting in Section 4.

When K > 1, we may use the specific choice of  $P(\cdot)$  to borrow strength across different groups and encourage similarity across their estimates. Two commonly-used penalty functions are the *fused lasso* (Tibshirani et al., 2005) and the *group lasso* (Yuan and Lin, 2006; Friedman et al., 2010). They are useful when we are interested in learning the differential dependence structures across different groups. The fused penalty

$$P(\{\Sigma_k\}) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\sigma_{k,ij}| + \lambda_2 \sum_{k < l} \sum_{i \neq j} |\sigma_{k,ij} - \sigma_{l,ij}|$$

encourages sparsity of the resulting covariance estimates  $\hat{\Sigma}_k$  and sparsity of their differences, where  $\lambda_1 > 0$  and  $\lambda_2 > 0$  are both tuning parameters and  $\sigma_{k,ij}$ denotes the (i, j)-element of  $\Sigma_k$ . The fused penalty encourages shared entrywise values across different covariance estimates. On the other hand, the group penalty

$$P(\{\Sigma_k\}) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\sigma_{k,ij}| + \lambda_2 \sum_{i \neq j} \sqrt{\sum_{k=1}^{K} \sigma_{k,ij}^2}$$

uses the similarity between groups and encourages sparsity on both individual and grouped levels, where  $\lambda_1 > 0$  and  $\lambda_2 > 0$  are both tuning parameters. While the sparsity pattern of each  $\hat{\Sigma}_k$  are likely to be similar to one another, the group penalty may be desirable when there is a shared sparsity pattern across different covariance estimates. We will explore the effectiveness of these penalty functions through simulation studies in Section 5 and a real application in Section 6. We summarize the Heavy COAT procedure as Algorithm 1. We refer to Section A1 of the Supplement for details on the algorithm and implementation of the thresholding step.

Algorithm 1 The Proposed HeavyCOAT Procedure.

**Step 1.** For k = 1, ..., K, obtain  $clr(\mathbb{X}_{0,k})$  through the centered log-ratio transformation.

Step 2. For k = 1, ..., K, construct sub-matrix  $\mathbb{X}_{k,c}$  by selecting the columns of  $clr(\mathbb{X}_{0,k})$ enumerated in C

**Step 3.** Estimation step. For k = 1, ..., K, solve (3.1) via (3.2) iteratively to obtain  $\hat{\Gamma}_k = \frac{p(\tilde{\Gamma}_k - \frac{\alpha}{1+\alpha} \mathbf{I}_{p \times p})}{\operatorname{tr}(\tilde{\Gamma}_k - \frac{\alpha}{1+\alpha} \mathbf{I}_{p \times p})}.$ 

Step 4. Thresholding step. Obtain the sparse estimator  $\hat{\Sigma}_k$  for k = 1, ..., K by solving (3.4) with the  $\ell_1$  penalty, fused penalty or group penalty via the ADMM that is presented in Section A1 of the Supplement.

# 4. Theoretical Properties

This section studies the theoretical properties of the HeavyCOAT procedure when K = 1. As we focus on K = 1, we drop the k subscript for ease of notation. Assume each latent log-basis random vector  $\mathbf{Y}_{0,i} \in \mathbb{R}^{p_0}$  in  $\mathbb{Y}_0$  follows an elliptical distribution  $\mathcal{E}_{p_0}(\mu, \Sigma_0, \phi)$ . We consider the following parameter space for the sparse latent shape matrix  $\Sigma_0$ :

$$U(q, M, s_0) = \{ \Sigma_0 : \Sigma_0 \succ 0, \ ||\Sigma_0||_2 \le M, \ \max_i \sum_{j=1}^{p_0} |\sigma_{ij}|^q \le s_0 \}$$

where  $M > 0, q \in [0, 1)$ , and  $s_0 > 0$ . Given the above definition of  $U(q, M, s_0)$ , we assume

(A1): 
$$\Sigma_0 \in U(q, M, s_0)$$
 with  $p_0/n \to r \in (0, +\infty)$  and  $s_0/p_0 \to 0$  as  $n \to \infty$ .

Recall that  $\hat{\Gamma}$  is the robust estimator for the shape sub-matrix  $\Gamma$  of the center log-transformed sub-matrix  $\mathbb{X}_c$  and  $\hat{\Sigma} = (\hat{\sigma}_{ij})_{p \times p}$  is the robust estimator for the sparse latent shape sub-matrix  $\Sigma$  of the reduced set of latent log-basis variables in  $\mathbb{Y}$ . In what follows, we first study the support recovery and sign consistency in Theorems 1–2 and then explore the convergence rate under the spectral norm and risk bound under the squared spectral norm in Theorems 3–4.

Theorems 1–2 show that the underlying shape matrix of  $\mathbb{Y}$  such as the sparsity and sign of  $\Sigma = (\sigma_{ij})_{p \times p}$  can be accurately estimated by  $\hat{\Sigma}$  based on  $\mathbb{X}_c$ even when the covariance matrix of  $\mathbb{Y}$  may not exist. Thus, even in the extreme settings, the HeavyCOAT procedure is still able to recover the directionality of the effect.

**Theorem 1.** Under assumption (A1), if  $\alpha > \max\{0, r - 1 + M(1 + \sqrt{r})^2\}$  and  $\lambda = C_1 \sqrt{\frac{\log(p)}{n}} + C_2 \frac{s_0}{p}$ , the sparse shape sub-matrix estimator  $\hat{\Sigma}$  satisfies that

$$P(\hat{\sigma}_{i,j} = 0 \text{ for all } (i,j) \text{ with } \sigma_{ij} = 0) \to 1, \text{ as } n \to \infty.$$
(4.5)

**Theorem 2.** Under assumption (A1), if  $\alpha > \max\{0, r - 1 + M(1 + \sqrt{r})^2\}$ ,  $\lambda = C_1 \sqrt{\frac{\log(p)}{n}} + C_2 \frac{s_0}{p}$  and  $\lambda \leq \frac{2}{3} \min_{(i,j):\sigma_{ij}\neq 0} |\sigma_{ij}|$ , the sparse shape sub-matrix estimator  $\hat{\Sigma}$  satisfies that

$$P(sgn(\hat{\sigma}_{i,j}) = sgn(\sigma_{ij}) \text{ for all } (i,j) \text{ with } \sigma_{ij} \neq 0) \to 1, \text{ as } n \to \infty.$$
(4.6)

Theorem 3 gives the convergence rate for estimating the sparse latent shape sub-matrix  $\Sigma$  under the spectral norm.

**Theorem 3.** Under assumption (A1), if  $\alpha > \max\{0, r - 1 + M(1 + \sqrt{r})^2\}$ and  $\lambda = C_1 \sqrt{\frac{\log(p)}{n}} + C_2 \frac{s_0}{p}$ , the sparse shape sub-matrix estimator  $\hat{\Sigma}$  from (3.4) satisfies that

$$||\hat{\Sigma} - \Sigma||_2 = O_p \left( s_0 \left( \sqrt{\frac{\log(p)}{n}} + \frac{s_0}{p} \right)^{1-q} \right)$$
(4.7)

The convergence rate in (4.7) is comparable to that of the COAT method (Cao et al., 2019). We consider the robust estimation under a larger class of elliptical distributions, while COAT imposes the sub-Gaussian tail condition. Similar to COAT, (4.7) can be decomposed into the estimation error  $(\sqrt{\log(p)/n})$  of  $\Sigma$  and the approximation error  $(s_0/p)$  from using the transformed data  $X_c$  as a proxy for the latent log-basis variables in  $\mathbb{Y}$ . Thus, our method also shows the appealing "blessing of dimensionality" as the estimation error dominates the approximation error as p increases. Given the convergence rate derived in Theorem 3, we further derive the risk bound of HeavyCOAT under the squared spectral norm in Theorem 4.

**Theorem 4.** Under the assumption (A1), if  $\alpha > \max\{0, r - 1 + M(1 + \sqrt{r})^2\}$ and  $\lambda = C_1 \sqrt{\frac{\log(p)}{n}} + C_2 \frac{s_0}{p}$ , the sparse shape sub-matrix estimator  $\hat{\Sigma}$  from (3.4) satisfies that

$$\sup_{\Sigma \in U(q,s_0,M)} E||\hat{\Sigma} - \Sigma||_2^2 \le C_3 s_0^2 \left( C_1 \sqrt{\frac{\log(p)}{n}} + C_2 \frac{s_0}{p} \right)^{2-2q} . \tag{4.8}$$

Theorem 4 is a new theoretical result for high-dimensional compositional data analysis. Although  $\hat{\Sigma}$  is a robust estimator under the class of elliptical distributions, the obtained risk bound matches the risk bound derived under the polynomial tails in Lemma 4 of Cai and Liu (2011).

## 5. Numerical Properties

In this section, we study the numerical effectiveness of our proposed Heavy-COAT procedure under a variety of settings. We analyze the K = 2 setting to demonstrate the effectiveness of joint estimation. Analysis of the K = 1setting can be found in the Supplement. We compare HeavyCOAT to three contemporary methods. As an alternative method to Tyler's M-Estimator, we consider a rank-based covariance estimator based on Spearman's  $\rho$  and Kendall's  $\tau$  (Xue and Zou, 2012, 2014b,a; Avella-Medina et al., 2018). Thus, by substituting Tyler's M-Estimator from Section 3.1 with an associated rank-based estimator, we denote the methods as kCOAT for a Kendall's  $\tau$ -based estimator and sCOAT for the Spearman's  $\rho$ -based estimator. Finally, we compare HeavyCOAT to the sample covariance-based COAT procedure (Cao et al., 2019). In all cases, we employ the universal thresholding procedure employed in Section 3.2 to ensure a positive-definite final estimate. If the group penalty is used, we append a -G suffix, and if the fused penalty is used, we append an -F suffix.

# 5.1 Performance Analysis

We study each method across a variety of scale mixture settings as outlined in Section 2. Namely we study the Gaussian setting where  $u_i = 1$  for all *i*, the Laplace setting where  $u_i \sim \text{Laplace}(0, 1)$ , the  $T_5$  setting where  $u_i \sim t_5$ , and the extreme Cauchy setting where  $u_i \sim \text{Cauchy}(0, 1)$ . For brevity, we present results for the heavy-tailed Laplace and the  $T_5$  setting and the Gaussian and Cauchy results can be found in the supplement.

We study the following two covariance models. Let the operation bdiag(A, B) denote creating a block-diagonal matrix with diagonal blocks A and B.

1. (Sparse Covariance)  $\Sigma_{0,1} = \text{bdiag}(B_1, 4\mathbf{I}_{p_2 \times p_2})$  is a block diagonal matrix with  $B_1 = B + \varepsilon \mathbf{I}_{p_1}$  and  $\varepsilon = \max(-\lambda_{\min}(B), 0) + .01$  for  $p_1 = \lfloor 2\sqrt{p_0} \rfloor$  and  $p_2 = p_0 - p_1$ . The sub-matrix *B* is constructed to be a random symmetric matrix whose lower triangular elements are drawn uniformly from  $U(-2, -4) \cup U(2, 4)$  with probability 0.15 and 0 otherwise. Given  $\Sigma_{0,1}$ , we generate by  $\Sigma_{0,2}$  by replacing each non-zero, non-diagonal element of  $\Sigma_{0,2}$  with 0 with probability 0.7.

2. (Block Covariance) Let  $A_r$  for r = 1, ..., 10 be  $\frac{p_0}{10} \times \frac{p_0}{10}$  matrices where  $A_{r,i,j} = 4(.7^{|i-j|})$ , and  $B = 3I_{\frac{p_0}{10} \times \frac{p_0}{10}}$ . Construct  $\Sigma_{0,1} = \text{bdiag}(A_1, ..., A_{10})$ and  $\Sigma_{0,2} = \text{bdiag}(A_1, ..., A_7, B, B, B)$  as two block diagonal matrices.

The Sparse Covariance structure is similar to the sparse setting studied by Cai et al. (2007). The Block Covariance structure consists of 10 blocks following an AR model similar to one studied in Bickel and Levina (2008). In this case, while each block matrix may not be sparse, the magnitudes of each  $A_r$ block decay rapidly. In both cases, we construct  $\Sigma_{0,2}$  by removing non-zero elements of  $\Sigma_{0,1}$ . Thus, for the non-removed elements,  $\Sigma_{0,2}$  shares structural and magnitude information with  $\Sigma_{0,1}$ . To match the trace constraint that is assumed for the shape matrix, we normalize  $\Sigma_{0,1}$  and  $\Sigma_{0,2}$  such that  $tr(\Sigma_{0,1}) = p$  and  $tr(\Sigma_{0,2}) = p$  as in (3.3).

We analyze the setting where n = 100 and  $p_0 \in \{150, 200, 250\}$ . Given this, our target for estimation is the  $p \times p$  sub-matrix where  $p = p_0 - 1$  of  $\Sigma_{0,1}$  and  $\Sigma_{0,2}$  denoted as  $\Sigma_1$  and  $\Sigma_2$  formed by dropping the  $p_0^{\text{th}}$  column of  $\mathbb{X}_c$ . We repeat

#### 5.1 Performance Analysis

the estimation process over s = 100 replications and assess the performance of our method by comparing the average Forbenius norm  $\frac{1}{2}\sum_{k=1}^{2} ||\Sigma_k - \hat{\Sigma}_k||_F$ , average spectral norm  $\frac{1}{2}\sum_{k=1}^{2} ||\Sigma_k - \hat{\Sigma}_k||_2$ , average true positive rate (TPR)  $\frac{1}{2}\sum_{k=1}^{2} \text{TPR}_k$ , and average false positive rate (FPR)  $\frac{1}{2}\sum_{k=1}^{2} \text{FPR}_k$  where,

$$\text{TPR}_{k} = \frac{\#\{(i, j) : \hat{\sigma}_{k, i, j} \neq 0 \text{ and } \sigma_{k, i, j} \neq 0\}}{\#\{(i, j) : \sigma_{k, i, j} \neq 0\}}$$

$$FPR_{k} = \frac{\#\{(i,j) : \hat{\sigma}_{k,i,j} \neq 0 \text{ and } \sigma_{k,i,j} = 0\}}{\#\{(i,j) : \hat{\sigma}_{k,i,j} \neq 0\}}$$

Note FPR<sub>k</sub> is defined to be 0 if  $\#\{(i, j) : \hat{\sigma}_{k,i,j} \neq 0\} = 0$ . We select  $\lambda_1$  and  $\lambda_2$  via the cross-validation procedure outlined in Danaher et al. (2014). Following Xue et al. (2012) and Goes et al. (2020), we set  $\psi = 2$  and  $\alpha = \max(p/n - 1, 0) + 1$ . The results for the Sparse Covariance setting can be seen in Table 4 and the results for the Block Covariance setting are found in Table 5.

Across the Laplace and  $T_5$  error settings, regardless of covariance structure, the HeavyCOAT procedure performs well in terms of both norms (specifically the spectral norm) and selection consistency. Both kCOAT and sCOAT have theoretical convergence and accuracy guarantees in cases where the fourth moment is bounded; though, their effectiveness in more extreme settings is unclear (Avella-Medina et al., 2018). This may be a key reason as to the good performance of the rank-based methods in the  $T_5$  and Laplace settings. In comparison,

#### 5.1 Performance Analysis

HeavyCOAT does not require the same moment assumptions; thus, can be applied more flexibly to situations where the underlying distribution may not be known apriori. We further note that HeavyCOAT appears to be less conservative than the rank-based methods, allowing for a higher TPR across settings. While the FPR appears to be slightly inflated in comparison to kCOAT and sCOAT, these differences may be explained through differences in the  $\lambda_1$  and  $\lambda_2$  selected by the cross-validation procedure. To better understand the selection power of HeavyCOAT compared to kCOAT and sCOAT, we compare these methods under the sparse covariance setting via an ROC analysis in Section 5.2.

## Table 3 and Table 4 are about here.

In both the block and sparse settings, we see the improvement of robust analysis over the sample covariance under the presence of heavy tails. COAT performs poorly in terms of error compared to all robust methods as the sample covariance suffers as the tails become heavier. Further, the FPR of COAT methods may be more than twice as large as HeavyCOAT depending on the choice of penalty. This is likely due to poor estimation in the Estimation Step causing errors to propagate down to the Thresholding Step, yielding inappropriate results. Thus, in settings where one may suspect the presence of heavy tails, a robust viewpoint is a necessity.

### 5.2 ROC Analysis

In this section, we perform compare the performance of HeavyCOAT, kCOAT, and sCOAT through the receiver operating characteristic (ROC) curves. Through this analysis, we demonstrate the improved performance of our proposed Heavy-COAT procedure across various different choices of the tuning parameters  $\lambda_1$ and  $\lambda_2$ . We focus our analysis on comparing the sparse covariance across the Gaussian, Laplace, and  $T_5$  settings. Further, we analyze the ROC curves under both the group-penalty (Figure 1) and fused-penalty (Figure 2).

Through these curves, we note that our proposed HeavyCOAT procedure dominates the competitive procedures regardless of the error type or penalty function. The kCOAT and sCOAT procedures are similar to one another, which matches our previous numerical results. In the Gaussian setting, all methods are relatively comparable which is to be expected as this setting exhibits no heavytailed behavior.

#### 6. Application to Microbial Inter-Taxa Analysis

In this section, we illustrate the potential usefulness of our method using microbiome data collected from n = 255 patients which exhibit forms of gastrointestinal disease (Morgan et al., 2015). For these individuals, various experimental factors were recorded, namely the use of antibiotics and the specific type of



Figure 1: ROC curves for the choice of  $\lambda_1$  and  $\lambda_2$  for the sparse covariance setting under the group lasso penalty. We compare HeavyCOAT, kCOAT, and sCOAT across the Gaussian, Laplace, and  $T_5$  scale mixture models.



Figure 2: ROC curves for the choice of  $\lambda_1$  and  $\lambda_2$  for the sparse covariance setting under the fused lasso penalty. We compare HeavyCOAT, kCOAT, and sCOAT across the Gaussian, Laplace, and  $T_5$  scale mixture models.

disease. The use of antibiotics has been well-studied and greatly reduces microbiome diversity in patients (Morgan et al., 2015; Dudek-Wicher et al., 2018). In regards to disease type, pouchitis refers to the inflammation of the ileal pouch which may become a chronic condition that often requires surgical intervention. The composition of the microbiome has been linked to the development of pouchitis. However, individuals with Familal Adenomatous Polyposis (FAP) undergo similar surgical intervention as individuals with inflammatory bowel disease though often do not develop pouchitis. Thus, as pouchitis is influenced by microbial composition, it is of interest to study the microbial dependence relationships of FAP patients to elucidate any differences. For the disease type analysis, the patients are split into two groups: FAP and non-FAP. In both the antibiotics and the disease type analysis, it is of particular interest to study the relationships between these bacterial taxa, as the inter-taxa dependence is linked to various risk factors (Morgan et al., 2015; Becker et al., 2015).

In this data, the microbial community was measured by 16s rRNA sequencing and sequence counts were clustered as Operational Taxonomic Units (OTU), representing biological taxa (Morgan et al., 2015). In particular, the target microbiome dataset consists of 7000 species-level OTUs, which have been further classified into  $p_0 = 303$  genera to reduce sequencing errors (Li, 2015). We focus on the sub-matrix of p = 302 genera by omitting the category of *unclassified*  taxa. Of the 255 patients,  $n_1 = 66$  had used antibiotics in the previous month before sampling and  $n_2 = 189$  did not take antibiotics.

The patients were also classified by disease type, where  $n_1 = 39$  patients were classified with FAP and  $n_2 = 216$  were classified as non-FAP. In these two-class cases, the choice of penalty function should be motivated by the underlying biological framework. It is known that antibiotics are extremely influential and can greatly alter the microbiome composition (Morgan et al., 2015; Dudek-Wicher et al., 2018). It is likely that the microbes that survive after an antibiotics regimen retain similar dependence relationships, but it is unlikely that the antibiotics group and the non-antibiotics group share similar values between their underlying shape matrices. Thus, we use the group lasso penalty to encourage the shared sparsity pattern, and borrows information less aggressively. On the other hand, as noted by Morgan et al. (2015), the difference between microbiome composition between disease-type is not a major factor. It is reasonable to assume that the dependence structures and strengths between both disease-type groups are similar. Employing the fused penalty in this setting is desirable as it will improve estimation by strongly leveraging information on both sparsity structure and covariance magnitude.

The correlation matrices were constructed from applying HeavyCOAT with the group penalty for the antibiotics group and the non-antibiotics group and apply the fused penalty for the disease type analysis. To visualize these interactions, we represent the correlation matrices via network graphs. To ensure network stability, we implement a bootstrapping procedure to capture the relevant edges. For each setting, we construct networks from applying our HeavyCOAT procedure across 50 bootstrapped samples. We retain the edges that appear at least 95% of the bootstrapped replicates, thereby presenting the most stable interactions within the microbial network. The edge thickness represents the strength of the correlation between nodes. The results for the antibiotics setting can be found in Figure 3 and disease setting can be found in Figure 4. To better understand the structure of these networks, we further implement the Louvain method for community detection (Blondel et al., 2008) to identify sub-communities of taxa within the correlation networks. These communities are denoted by the circular or square node shape in Figure 3 and Figure 4. In all settings, we identify two unique clustering of taxa. Finally, we compute key network statistics (average correlation and average node degree) for each setting in Table 3.

For ease of presentation, we present the relationships between the top 40 genera of these phyla as they are likely to be the most influential. We color each node by the corresponding phyla. When the effect of antibiotics in Figure 3 we observed a marked change in the sparsity pattern as expected. Firstly, after antibiotic use, the overall diversity of the microbial network is drastically reduced

Table 3: Average network statistics across group settings. The Degree column denotes the average degree across all nodes within the network, and the Correlation column lists the average correlation coefficient between taxa within each network.

Setting	Degree	Correlation
Antibiotics	7.10	0.162
No Antibiotics	16.27	0.158
FAP	11.36	0.193
non-FAP	11.36	0.200

with far fewer active edges, which is can be seen by the large reduction in average degree in the antibiotic setting. This reduction in microbial diversity is well studied (Hildebrand et al., 2019), as there may be fundamental disruptions in these microbial systems after antibiotic use (Schwartz et al., 2020; Xu et al., 2020; Seelbinder et al., 2020). For example, we observe that genera primarily of the *Firmicutes* phylum remain active. Recent literature has shown that members of the *Firmicutes* phylum may opportunistically dominate other phyla in the post-antibiotics ecosystem (Ng et al., 2019).

When analyzing the key community memberships, we focus on the four taxa identified by Morgan et al. (2015) as playing a pivotal role: *Escherichia, Roseburia, Bifidobacterium, Sutterella*. When no antibiotics were used, we see that *Escherichia* is in a unique community while *Sutterella, Roseburia, and Bi-fidobacterium* are in a unique community. However, after antibiotics use, *Bifidobacterium* and *Escherichia* are in a separate cluster from *Roseburia*, and *Sut-*

*terella* disappears all together. This shift in community membership influenced by the use of antibiotics may point to further evidence that the use of antibiotics can fundamentally affect the relationships between surviving microbial taxa.

When comparing individuals of the non-FAP disease type to those of the FAP disease type, the difference between the covariance structures are less extreme. The two settings have identical correlation structure which is to be expected as the fused penalty aggressively ensures similar sparsity structures between groups. However, the edge weights between groups may vary. The correlation strength in the FAP group is slightly less than that of the non-FAP group, but the difference between disease types is less extreme than the effect of antibiotics which is expected via the analysis conducted by Morgan et al. (2015). Community membership also remains consistent across groups, further suggesting that influence of disease type is not as impactful in differentiating microbial communities. In both groups, the *Firmicutes* and *Bacteroidetes* phyla were the most active which is to be expected as these two phyla are dominant within the gut microbiome (Hildebrand et al., 2019; Morgan et al., 2015).

Figure 3 and Figure 4 are about here.

## 7. Conclusion

We proposed the HeavyCOAT procedure to estimate the latent shape sub-matrix of high-dimensional compositional data across multiple groups, which is a scalar multiple of the covariance matrix when it exists. Unlike existing methods, Heavy-COAT has improved estimation accuracy by modeling the heavy-tails using a large class of elliptical distributions. We showed that when K = 1, HeavyCOAT has competitive theoretical properties. When the number of groups increases, as may be common in practical microbiome data analysis, the use of the fused and group penalties during the thresholding step allow us to leverage power across the multiple groups and improve estimation. Finally, we applied HeavyCOAT to gut microbiome data and identified biologically significant dependence patterns.

Currently, our method is agnostic to known structure within the data. For example, in the microbiome setting, bacterial taxa can be naturally organized by a phylogentic tree structure and we expect similar taxa to have a similar dependence relationships. Further avenues of exploration include embedding known structure into the thresholding scheme. For example, one could apply group thresholding to vary the thresholding intensity across groups to account for this structure.

#### **Supplementary Materials**

In the supplementary note, we describe the ADMM for solving multiple shape matrices in the thresholding step, present simulation results under additional settings, and provide detailed technical derivations for the Lemmas and Theorems presented in Section 2.

### Acknowledgements

We would like to thank the Editor, Associate Editor and referees for their helpful comments and suggestions. Danning Li was partially supported by the National Natural Science Foundation of China grant 12101116. Arun Srinivasan and Lingzhou Xue were partially supported by the National Institutes of Health grants R21AI144765 and T32GM102057, and the National Science Foundation grants DMS-1811552 and DMS-1953189.

## References

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. London, UK: Chapman and Hall.

- Andrews, D. F. and C. L. Mallows (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Methodological)* 36(1), 99–102.
- Arata, Y. and T. Onozaki (2017). A compositional data analysis of market share dynamics. Discussion papers, Research Institute of Economy, Trade and Industry (RIETI).

Avella-Medina, M., H. S. Battey, J. Fan, and Q. Li (2018). Robust estimation of high-dimensional covariance and precision matrices. *Biometrika* 105(2), 271–284.

Becker, C., M. F. Neurath, and S. Wirtz (2015). The intestinal microbiota in inflammatory bowel disease.

ILAR Journal 56(2), 192-204.

Bickel, P. J. and E. Levina (2008). Covariance regularization by thresholding. The Annals of Statistics 36(6),

2577 - 2604.

Bien, J. (2019). Graph-guided banding of the covariance matrix. *Journal of the American Statistical* Association 114(526), 782–792.

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

- Blondel, V., J.-L. Guillaume, R. Lambiotte, and E. Lefebvre (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics Theory and Experiment 2008*, P10008.
- Cai, D., X. He, and J. Han (2007). Semi-supervised discriminant analysis. In 2007 IEEE 11th International Conference on Computer Vision, pp. 1–7.
- Cai, T. and W. Liu (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association 106*(494), 672–684.
- Cai, T. T., H. H. Zhou, et al. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics* 40(5), 2389–2420.
- Cambanis, S., S. Huang, and G. Simons (1981). On the theory of elliptically contoured distributions. Journal of Multivariate Analysis 11(3), 368–385.

- Cao, Y., W. Lin, and H. Li (2019). Large covariance estimation for compositional data via compositionadjusted thresholding. *Journal of the American Statistical Association 114*(526), 759–772.
- Cho, I. and M. J. Blaser (2012). The human microbiome: at the interface of health and disease. *Nature Reviews Genetics* 13(4), 260.

Conover, W. J. (1998). Practical Nonparametric Statistics, 3rd Edition. New York; Wiley.

Cramér, H. (1928). On the composition of elementary errors. Scandinavian Actuarial Journal 1928(1),

13-74.

- Danaher, P., P. Wang, and D. M. Witten (2014, Mar). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 76(2), 373–397.
- Dudek-Wicher, R., A. Junka, and M. Bartoszewicz (2018). The influence of antibiotics and dietary components on gut microbiota. *Gastroenterology Review 13*(2), 85–92.
- Duembgen, L. and D. E. Tyler (2016). Geodesic convexity and regularized scatter estimators. *arXiv preprint arXiv:1607.05455*.
- Fan, J., L. Xue, and H. Zou (2016). Multitask quantile regression under the transnormal model. *Journal of the American Statistical Association 111*(516), 1726–1735.
- Fang, H., C. Huang, H. Zhao, and M. Deng (2015). Cclasso: Correlation inference for copositional data through lasso. *Bioinformatics* 31(19), 3172–3180.

Fang, K. T., S. Kotz, and K. W. Ng (1990). Symmetric Multivariate and Related Distributions. Chapman

#### and Hall/CRC.

Friedman, J. and E. Alm (2012). Inferring correlation networks from genomic survey data. PLoS Compu-

tational Biology 8(9), e1002687.

- Friedman, J., T. Hastie, and R. Tibshirani (2010). A note on the group lasso and a sparse group lasso. *arXiv* preprint arXiv:1001.0736.
- Goes, J., G. Lerman, and B. Nadler (2020). Robust sparse covariance estimation by thresholding Tyler's M-estimator. *The Annals of Statistics* 48(1), 86 110.
- Goodrich, J. K., J. L. Waters, A. C. Poole, J. L. Sutter, O. Koren, R. Blekhman, M. Beaumont, W. Van Treuren, R. Knight, J. T. Bell, T. D. Spector, A. G. Clark, and R. E. Ley (2014, Nov). Human genetics shape the gut microbiome. *Cell 159*(4), 789–799.
- Hildebrand, F., L. Moitinho-Silva, S. Blasche, M. T. Jahn, T. I. Gossmann, J. Huerta-Cepas, R. Hercog, M. Luetge, M. Bahram, A. Pryszlak, R. J. Alves, S. M. Waszak, A. Zhu, L. Ye, P. I. Costea, S. Aalvink, C. Belzer, S. K. Forslund, S. Sunagawa, U. Hentschel, C. Merten, K. R. Patil, V. Benes, and P. Bork (2019). Antibiotics-induced monodominance of a novel gut bacterial order. *Gut 68*(10), 1781–1790.
- Li, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application* 2, 73–94.
- Morgan, X. C., B. Kabakchiev, L. Waldron, A. D. Tyler, T. L. Tickle, R. Milgrom, J. M. Stempak, D. Gevers, R. J. Xavier, M. S. Silverberg, et al. (2015). Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome Biology 16*(1), 67.

Müller, K. and W.-D. Richter (2019). On p-generalized elliptical random processes. *Journal of Statistical Distributions and Applications 6*(1), 2195–5832.

- Ng, K. M., A. Aranda-Díaz, C. Tropini, M. R. Frankel, W. Van Treuren, C. T. O'Loughlin, B. D. Merrill,
  F. B. Yu, K. M. Pruss, R. A. Oliveira, et al. (2019). Recovery of the gut microbiota after antibiotics depends on host diet, community context, and environmental reservoirs. *Cell host & microbe* 26(5), 650–665.
- Nordhausen, K. and D. E. Tyler (2015). A cautionary note on robust covariance plug-in methods. Biometrika 102(3), 573–588.
- Pascal, F., Y. Chitour, and Y. Quek (2013). Generalized robust shrinkage estimator and its application to stap detection problem. *IEEE Transactions on Signal Processing* 62(21), 5640–5651.
- Rothman, A. J. (2012). Positive definite estimators of large covariance matrices. *Biometrika* 99(3), 733–740.
- Rothman, A. J., E. Levina, and J. Zhu (2009). Generalized thresholding of large covariance matrices. Journal of the American Statistical Association 104(485), 177–186.
- Schwartz, D., A. Langdon, and G. Dantas (2020). Understanding the impact of antibiotic perturbation on the human microbiome. *Genome Medicine 12*, 82.
- Seelbinder, B., J. Chen, S. Brunke, R. Vazquez-Uribe, R. Santhaman, A.-C. Meyer, F. S. de Oliveira Lino,
  K.-F. Chan, D. Loos, L. Imamovic, C.-C. Tsang, R. Lam, S. Sridhar, K. Kang, B. Hube, P. Woo,
  M. Sommer, and G. Panagiotou (2020). Antibiotics create a shift from mutualism to competition in human gut communities with a longer-lasting impact on fungi than bacteria. *Microbiome* 8, 133.

Shapiro, S. S. and R. S. Francia (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association* 67(337), 215–216.

- Sun, Y., P. Babu, and D. P. Palomar (2014). Regularized tyler's scatter estimator: Existence, uniqueness, and algorithms. *IEEE Transactions on Signal Processing* 62(19), 5143–5156.
- Sun, Y., P. Babu, and D. P. Palomar (2015). Robust estimation of structured covariance matrix for heavytailed elliptical distributions. *IEEE Transactions on Signal Processing* 64(14), 3576–3590.
- Thomas, C. and J. Aitchison (2005). Compositional data analysis of geological variability and process: A case study. *Mathematical Geology* 37, 753–772.
- Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(1), 91–108.
- Tyler, D. E. (1987). A distribution-free *m*-estimator of multivariate scatter. *The Annals of Statistics 15*(1), 234–251.
- Wiesel, A. and T. Zhang (2015). Structured robust covariance estimation. *Foundations and Trends*® *in Signal Processing* 8(3), 127–216.

Xu, L., A. Surathu, I. Raplee, A. Chockalingam, S. Stewart, L. Walker, L. Sacks, V. Patel, Z. Li, and

- R. Rouse (2020). The effect of antibiotics on the gut microbiome: a metagenomics analysis of microbial shift and gut antibiotic resistance in antibiotic treated mice. *BMC Genomics* 21, 263.
- Xue, L., S. Ma, and H. Zou (2012). Positive-definite 11-penalized estimation of large covariance matrices. Journal of the American Statistical Association 107(500), 1480–1491.

Xue, L. and H. Zou (2012). Regularized rank-based estimation of high-dimensional nonparanomal graphi-

cal models. The Annals of Statistics 40(5), 2541-2571.

Xue, L. and H. Zou (2014a). Optimal estimation of sparse correlation matrices of semiparametric gaussian

copulas. Statistics and its Interface 7(2), 201-209.

Xue, L. and H. Zou (2014b). Rank-based tapering estimation of bandable correlation matrices. Statistica

Sinica 24(1), 83-100.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. Journal

of the Royal Statistical Society: Series B (Statistical Methodology) 68(1), 49–67.

Table 4: Comparison of the estimation and selection performance under the sparse covariance setting with n = 100 and  $p_0 = 150$ , 200, and 250 over 100 independent repetitions.

	Laplace Mixture Setting			$T_5$ Mixture Setting					
	$p_0 = 150$	$p_0 = 200$	$p_0 = 250$	$p_0 = 150$	$p_0 = 200$	$p_0 = 250$			
		For	benius Norm						
HeavyCOAT-G	10.75	13.33	15.90	10.71	13.31	15.90			
kCOAT-G	8.78	10.71	12.43	8.82	10.73	12.55			
sCOAT-G	8.28	10.12	11.73	8.44	10.25	12.02			
COAT-G	97.43	127.18	153.96	78.03	106.70	122.63			
HeavyCOAT-F	9.99	12.14	14.13	9.96	12.14	14.13			
kCOAT-F	8.44	10.13	11.47	8.40	10.05	11.42			
sCOAT-F	8.44	10.13	11.47	8.40	10.05	11.42			
COAT-F	89.36	112.69	142.33	70.47	92.45	103.77			
		Sp	ectral Norm						
HeavyCOAT-G	2.57	2.80	3.12	2.54	2.79	3.10			
kCOAT-G	2.56	2.79	3.07	2.56	2.81	3.08			
sCOAT-G	2.58	2.83	3.14	2.61	2.90	3.19			
COAT-G	53.84	71.43	83.09	42.72	64.42	70.08			
HeavyCOAT-F	9.99	12.14	14.13	9.96	12.14	14.13			
kCOAT-F	8.44	10.13	11.47	8.40	10.05	11.42			
sCOAT-F	8.44	10.13	11.47	8.40	10.05	11.42			
COAT-F	89.36	112.69	142.33	70.47	92.45	103.77			
		True	e Positive Rate						
HeavyCOAT-G	0.80	0.80	0.82	0.80	0.81	0.82			
kCOAT-G	0.76	0.76	0.78	0.76	0.76	0.78			
sCOAT-G	0.75	0.74	0.77	0.75	0.75	0.77			
COAT-G	0.80	0.80	0.82	0.80	0.80	0.82			
HeavyCOAT-F	0.78	0.77	0.80	0.77	0.77	0.80			
kCOAT-F	0.74	0.74	0.77	0.74	0.74	0.76			
sCOAT-F	0.74	0.74	0.77	0.74	0.74	0.76			
COAT-F	0.80	0.80	0.82	0.79	0.79	0.81			
	False Positive Rate								
HeavyCOAT-G	0.06	0.07	0.12	0.06	0.07	0.12			
kCOAT-G	0.03	0.03	0.06	0.03	0.04	0.06			
sCOAT-G	0.03	0.03	0.05	0.03	0.03	0.06			
COAT-G	0.27	0.27	0.29	0.27	0.27	0.29			
HeavyCOAT-F	0.05	0.05	0.10	0.05	0.05	0.10			
kCOAT-F	0.02	0.02	0.05	0.02	0.03	0.05			
sCOAT-F	0.02	0.02	0.05	0.02	0.03	0.05			
COAT-F	0.34	0.34	0.36	0.33	0.33	0.35			

Table 5: Comparison of the estimation and selection performance under the block covariance setting with n = 100 and  $p_0 = 150$ , 200, and 250 over 100 independent repetitions.

	Lapla	ce Mixture S	Setting	$T_5$ .	$T_5$ Mixture Setting				
	$p_0 = 150$	$p_0 = 200$	$p_0 = 250$	$p_0 = 150$	$p_0 = 200$	$p_0 = 250$			
		For	rbenius Norm						
HeavyCOAT-G	15.70	19.00	22.01	15.70	19.02	22.00			
kCOAT-G	15.00	17.83	20.46	14.88	17.75	20.29			
sCOAT-G	14.73	17.39	19.88	14.57	17.29	19.68			
COAT-G	61.31	81.85	94.76	49.86	64.11	72.14			
HeavyCOAT-F	15.86	18.96	21.69	15.85	18.96	21.66			
kCOAT-F	15.16	17.93	20.39	15.03	17.81	20.21			
sCOAT-F	14.89	17.56	19.91	14.72	17.39	19.68			
COAT-F	42.73	54.68	63.72	35.10	43.51	51.44			
		Sp	pectral Norm						
HeavyCOAT-G	3.12	3.38	3.48	3.12	3.36	3.49			
kCOAT-G	3.91	4.18	4.33	3.88	4.14	4.29			
sCOAT-G	4.15	4.42	4.58	4.11	4.38	4.53			
COAT-G	31.85	41.24	49.06	25.76	32.56	33.02			
HeavyCOAT-F	3.38	3.65	3.79	3.37	3.64	3.79			
kCOAT-F	4.06	4.34	4.51	4.03	4.32	4.48			
sCOAT-F	4.28	4.57	4.74	4.24	4.53	4.69			
COAT-F	19.05	24.61	28.46	15.35	18.45	21.81			
		True	e Positive Rate	e					
HeavyCOAT-G	0.26	0.20	0.16	0.26	0.20	0.16			
kCOAT-G	0.19	0.15	0.12	0.19	0.15	0.12			
sCOAT-G	0.17	0.13	0.11	0.17	0.13	0.11			
COAT-G	0.30	0.27	0.24	0.28	0.25	0.22			
HeavyCOAT-F	0.19	0.14	0.12	0.19	0.15	0.12			
kCOAT-F	0.16	0.12	0.10	0.16	0.12	0.10			
sCOAT-F	0.15	0.11	0.09	0.15	0.11	0.09			
COAT-F	0.19	0.16	0.13	0.17	0.14	0.12			
False Positive Rate									
HeavyCOAT-G	0.00	0.00	0.00	0.00	0.00	0.00			
kCOAT-G	0.00	0.00	0.00	0.00	0.00	0.00			
sCOAT-G	0.00	0.00	0.00	0.00	0.00	0.00			
COAT-G	0.23	0.23	0.23	0.23	0.23	0.23			
HeavyCOAT-F	0.00	0.00	0.00	0.00	0.00	0.00			
kCOAT-F	0.00	0.00	0.00	0.00	0.00	0.00			
sCOAT-F	0.00	0.00	0.00	0.00	0.00	0.00			
COAT-F	0.21	0.21	0.21	0.18	0.19	0.19			



Figure 3: Networks for the estimated correlation networks between major phyla of the antibiotic group (Left) and the non-antibiotic group (Right). Taxa are sorted into communities via the Louvain method of community detection and are identified by a circle shape and square shape respectively.



Figure 4: Networks for the estimated correlation networks between major phyla of the FAP group (Left) and the non-FAP group (Right). Taxa are sorted into communities via the Louvain method of community detection and are identified by a circle shape and square shape respectively.

KLAS and School of Mathematics & Statistics, Northeast Normal University

E-mail: lidn040@nenu.edu.cn

Department of Statistics, Pennsylvania State University

E-mail: uus91@psu.edu

Department of Statistics, Pennsylvania State University

E-mail: lzxue@psu.edu

Department of Biostatistics, School of Public Health and Beijing International Center for Mathematical

Research, Peking University

E-mail: zhanx@bjmu.edu.cn