Sparse Bayesian Group Factor Model for Feature Interactions in Multiple Count Tables Data

Shuangjie Zhang *
Department of Statistics, University of California Santa Cruz

Yuning Shen
Department of Chemical and Biomolecular Engineering, University of California Los Angeles†

Irene A. Chen
Department of Chemical and Biomolecular Engineering, University of California Los Angeles

Juhee Lee
Department of Statistics, University of California Santa Cruz

May 14, 2024

Abstract

Group factor models have been developed to infer relationships between multiple co-occurring multivariate continuous responses. Motivated by complex count data from multi-domain microbiome studies using next-generation sequencing, we develop a sparse Bayesian group factor model (Sp-BGFM) for multiple count table data that captures the interaction between microorganisms in different domains. Sp-BGFM uses a rounded kernel mixture model using a Dirichlet process (DP) prior with log-normal mixture kernels for count vectors. A group factor model is used to model the covariance matrix of the mixing kernel that describes microorganism interaction. We construct a Dirichlet-Horseshoe (Dir-HS) shrinkage prior and use it as a joint prior for factor loading vectors. Joint sparsity induced by a Dir-HS prior greatly improves the performance in high-dimensional applications. We further model the effects of covariates on microbial abundances using regression. The semiparametric model flexibly accommodates large variability in observed counts and excess zero counts and provides a basis for robust estimation of the interaction and covariate effects. We evaluate Sp-BGFM using simulation studies and real data analysis, comparing it to popular alternatives. Our results highlight the necessity of joint sparsity induced by the Dir-HS prior, and the benefits of a flexible DP model for baseline abundances.

Keywords: Dirichlet Horseshoe Distributions; Dirichlet Process Mixtures; High Dimensionality; Joint Sparsity; Rounded Kernel Model.

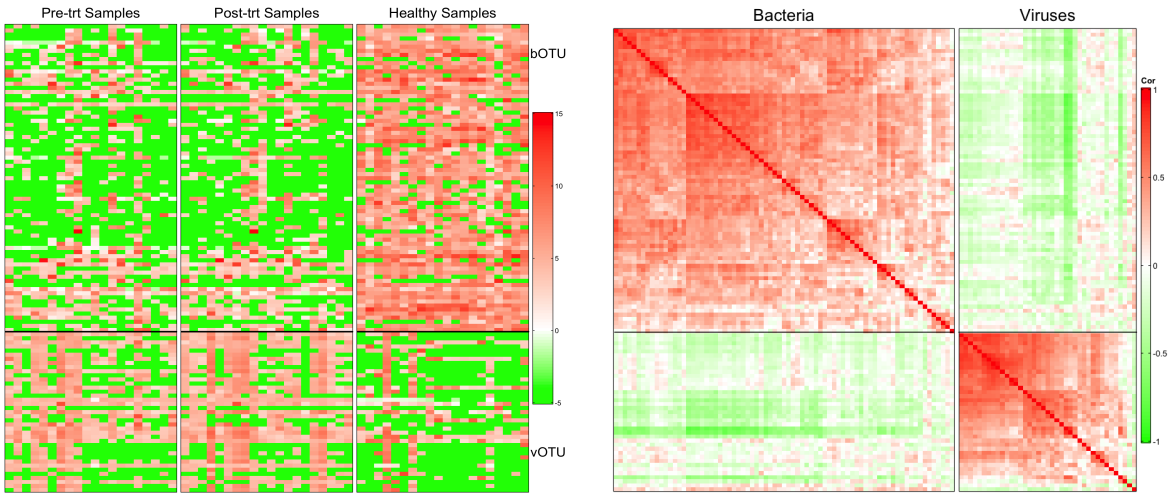
*Address for Correspondence: 1156 High St, Santa Cruz, CA 95064. E-mail: szhan209@ucsc.edu.
†The work is conducted during UCLA and the current affiliation is ByteDance Research.

1 Introduction

1.1 Motivation and Multi-Domain Microbiome Data

Statistical methods that capture correlations in different responses can be helpful in the multiple output case. For example, canonical correlation analysis (CCA) and inter-battery factor analysis (IBFA) are useful tools that combine two multivariate responses and provide inference on cross-covariance between the responses (Browne, 1979; Bach and Jordan, 2005; Klami et al., 2013). Group factor analysis extends traditional factor analysis to infer joint variability between two or more multivariate responses (Virtanen et al., 2012; Klami et al., 2014; Zhao et al., 2016). However, they may not be suitable for the analysis of multiple intercorrelated multivariate count variables because those methods consider continuous responses and assume a multivariate normal distribution.

Motivated by a high-throughput sequencing dataset from the multi-domain chronic wounds microbiome study in Verbanic et al. (2020, 2022); Zhang et al. (2023), we develop a Bayesian group factor model that accounts for the discreteness of data with *multiple count responses*. Microorganisms, including bacteria, viruses, fungi, and archaea, coexist in diverse communities and form polymicrobial communities within the human body (Peters et al., 2012). Polymicrobial infection is one of the leading impediments to chronic wound healing. Appropriately inferring the intricate interactions among microorganisms, both within a specific domain and across different domains, as well as their associations with the environment, is crucial to a better understanding of the healing of chronic wounds. The dataset consists of multiple count tables, with each count table representing a specific microorganism domain. In these count tables, the counts correspond to the abundances of microbial operational taxonomic units (OTUs), which are commonly used as a proxy for microbial species. The motivating study investigated bacteria and bacteriophages (bacterial viruses) in the wound microbiome. Bacteriophages play a role in regulating bacterial



(a) Log-transformed normalized OTU counts

(b) Empirical correlation estimates

Figure 1: [Multi-domain skin microbiome data] Panel (a) has a heatmap of the log-transformed normalized OTU counts. The counts are normalized using cumulative sum scaling. A pseudocount of 0.01 is added for log transformation. Panel (b) illustrates empirical correlation estimates using the log-transformed normalized OTU counts. The OTUs are rearranged within a domain.

abundance and influencing their metabolism and fitness. They are essential components of the wound microbiome. However, the interaction between bacterial and viral communities in wound microbiomes has received relatively limited attention. Verbanic et al. (2020) and Zhang et al. (2023) focused on the bacterial fraction of the microbial community in the dataset and examined its taxonomic associations with debridement - a common treatment for chronic wounds, whereas Verbanic et al. (2022) explored the viral content of wound surfaces in the same dataset but did not analyze it together with bacteria. To gain a comprehensive understanding of wound microbiomes and their association with treatment, it is essential to consider both bacteria and bacteriophages.

More specifically, the study collected wound swabs from 20 patients attending an outpatient wound care clinic. Samples were obtained from chronic wounds before and after a treatment event, as well as from a control site on the skin. This resulted in a dataset of 60 samples from 20 subjects, along with a categorical covariate with three levels: healthy, pre-treatment and post-treatment. The abundance of bacteria in the samples was measured by

high-throughput sequencing of the V1–V3 loops of 16S rRNA genes, and the abundance of viral contents by high-throughput sequencing of DNA from virus-like particles (VLPs) isolated from the samples. Counts of bacterial OTUs (bOTU) were aggregated at the genus level, and counts of viral OTUs (vOTUs) at the host level. To ensure reliable inference, we removed OTUs having extremely low counts on average or having zero counts in a significant number of samples. The preprocessing details are described in § 4. After preprocessing, the dataset comprises counts of 75 bOTUs and 39 vOTUs in the two domains, bacteria and viruses, for the 60 samples. Fig 1(a) shows a heatmap of the log-transformed normalized OTU counts. The counts are normalized using cumulative sum scaling (CSS) in Paulson et al. (2013). CSS normalization involves summing the OTU counts up to a pre-specified quantile of a sample and generating normalized counts by dividing the counts by the sum. The sample medians are used for the illustration. It corrects potential bias introduced by total-sum normalization (TSS) in differential abundance analysis. To avoid problems with the log transformation of zero counts, a pseudocount of 0.01 is added. From the figure, the bOTUs exhibit higher richness in the healthy skin samples than in the wound samples. On the other hand, the vOTUs are more enriched in the wound samples than in the healthy skin samples. Fig 1(b) illustrates empirical correlation estimates using the log-transformed normalized counts from all 60 samples obtained under the three different experimental conditions. Also, empirical correlation estimates are computed separately for each condition and presented in Supp. Fig. 15. The figures indicate potential interactions between OTUs within and across different domains.

1.2 Statistical Challenges

Besides discreteness, microbiome data presents several challenges for statistical modeling, including compositionality, excess zeros, high dimensionality and large inter-sample variability. Typically, microbiome data is represented as a table of counts, where the total

number of reads can vary between samples due to experimental artifacts such as sequencing depth. Raw counts in an OTU table thus represent only relative abundances in a sample (i.e., compositionality), and it requires appropriate normalization of raw counts for modeling. Supp. Fig 16 illustrates histograms of the logarithm of the total counts in the skin microbiome dataset. The total counts greatly vary across samples, with the variability differing according to the domain. In addition, OTU count tables contain excess zeros because of the absence of OTUs and/or limited sequencing depth, with counts of an OTU greatly varying due to a large amount of inter-subject or inter-sample variability. Fig 1(a) reveals a substantial degree of variability in OTU counts among samples even after taking into account the difference in sample total counts through normalization. The figure also illustrates excess zeros in the dataset. Furthermore, in the presence of environmental factors, the underlying data-generating structure becomes even more complicated. These make statistical analysis challenging, and any method that does not address them appropriately may produce erroneous inferences such as spurious estimates of correlations between microorganisms.

1.3 Current Approaches and Limitations

Various statistical methods have been developed to explore the associations among microorganisms, mainly with a focus on a single domain (i.e., a count table of a single group). Typically, a covariance or precision (i.e., inverse covariance) matrix is utilized to infer the associations. Most of these methods use a penalized estimation method after normalizing and/or transforming raw counts. The graphical lasso in Friedman et al. (2008) is one of the popular penalized methods for estimating the precision matrix Σ^{-1} that forms an undirected graph in a high-dimensional setting. In a Gaussian graphical model, the off-diagonal values of zero and non-zero in Σ^{-1} represent conditional independence or dependence between the OTUs. The ℓ_1 penalty encourages sparsity in Σ^{-1}. Examples of

the graphical model based approach include SPIEC-EASI (SParse InversE Covariance Estimation for Ecological Association Inference) (Kurtz et al., 2015), Zi-LN (Zero-inflated Log-Normal model) (Prost et al., 2021), Comp-gLASSO (Compositional graphical LASSO method) (Tian et al., 2023) and PhyloBCG (Phylogenetically-informed Bayesian Copula Graphical model) (Chung et al., 2022) among many others. All these methods are designed for single-domain microbiome data analysis. Specifically, SPIEC-EASI first applies the centered log-ratio (clr) transformation to raw OTU counts to account for the compositionality and discreteness. It then assumes a Gaussian distribution with mean zero and precision matrix Σ^{-1} for the clr transformed data and estimates Σ^{-1} with the ℓ_1 penalty to obtain an interaction graph. This method was later extended to allow for multi-domain analysis by applying the clr transformation separately to an OTU table from each domain and estimating the precision matrix using a concatenated transformed composition vector (Tipton et al., 2018). Other penalized estimation methods of the covariance matrix Σ include REBECCA (Regularized Estimation of the Basis Covariance Based on Compositional Data) (Ban et al., 2015) and COAT (COmposition-Adjusted Thresholding Method) (Cao et al., 2019) that are developed for single group data analysis. Alternatively, low-rank approximations can be used for the estimation of Σ. For example, see MOFA (Multi-Omics Factor Analysis) (Argelaguet et al., 2018) and ZI-MLN (Zero-inflated Multivariate Log-normal Kernel Model) (Zhang et al., 2023). In particular, MOFA builds a Bayesian group factor model for clr-transformed multi-group count table data. The data is recentered by subtracting the sample mean for each OTU, and subsequently it assumes a normal distribution with mean zero and covariance Σ. Σ is estimated by a factor model that assumes two-level sparsity priors for factor loadings to obtain fast computation and robust estimation. While there are several methods available for inferring microorganism interactions across multiple domains, a need remains for more robust approaches to address the aforementioned challenges.

We take the low-rank approximation approach and develop a sparse Bayesian group factor model (Sp-BGFM) for the analysis of multiple multivariate count data to obtain desired inferences on within-domain and across-domain OTU interactions. Sp-BGFM extends the applicability of a conventional group factor model that handles continuous responses by assuming a Gaussian model with a fixed mean at zero. It directly constructs a discrete distribution for count vectors and simultaneously models mean and variance of a count vector. Specifically, using the approach in Canale and Dunson (2011), Sp-BGFM builds nonparametric mixtures of rounded multivariate continuous kernels using a Dirichlet process (DP) prior to obtain a flexible joint distribution of count vectors. A mean-constrained mixture of log-normals is used as the kernel to capture the location of the count distribution without identifiability problems. A novel prior distribution, the Dirichlet-Horseshoe (Dir-HS) distribution, is constructed as a joint prior on factor loading vectors to efficiently induce joint sparsity and provide reliable inferences on a high-dimensional interaction structure within and across domains, even with a small sample size. The semiparametric formulation flexibly accommodates excess zeros and inter-subject or inter-sample variability in OTU counts and further improves the estimation of OTU interaction. Moreover, the mean function of the kernel is extended through regression to accommodate covariates. Also, our model simultaneously performs model-based normalization for proper uncertainty quantification. Extensive numerical studies show that Sp-BGFM recovers the underlying data-generating process including within- and cross-domain interaction reasonably well and performs very competitively compared to various comparators. The method is then applied to analyze real multi-domain skin microbiome data.

The rest of this article is organized as follows. § 2 details the development of Sp-BGFM and describes the prior specification and posterior computation. In § 3, we evaluate the performance of Sp-BGF under different simulation settings and compare it to several popular alternatives. § 4 demonstrates the application of our method to the multi-domain

skin microbiome dataset. Finally, § 5 provides a brief discussion and conclusion.

2 Model and Posterior Inference

2.1 Sampling Distribution and Prior Specification

Consider random count vectors of M different groups (or domains). Let $\mathbf{y}_{im} = (y_{im1}, \ldots, y_{imJ_m})'$ denote a J_m-dimensional vector of group m of sample i, $i = 1, \ldots, N$ and $m = 1, \ldots, M$. Each $y_{imj} \in \mathbb{N}_0$, $j = 1, \ldots, J_m$, is a non-negative integer that represents an unnormalized abundance of OTU j of group m in sample i. We stack $\mathbf{y}_{im}$ and construct a table $\mathbf{Y}_m$ of size $N \times J_m$, a subset of data corresponding to group m. We assume that $\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iM}$ in sample i are obtained from subject s_i, where $s_i \in \{1, \ldots, S\}$. Also, data may have a vector of P covariates, $\mathbf{x}_i = (x_{i1}, \ldots, x_{iP})$ that may be associated with $\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iM}$.

We concatenate the vectors $\mathbf{y}_{im}$ of sample i and construct $\mathbf{y}_i = (\mathbf{y}'_{i1}, \ldots, \mathbf{y}'_{iM})'$ a J-dim count vector of OTUs in M different groups for sample i, where $J = \sum_{m=1}^{M} J_m$ is the total number of OTUs. Taking the rounded kernel approach for count data in Canale and Dunson (2011), we introduce a continuous random vector $\mathbf{y}^\star_i \in \mathbf{R}^J_+$ and build a flexible model for $\mathbf{y}^\star_i$. For sample i from subject s_i, we assume

$$\mathbf{y}^\star_i \mid \mathbf{r}_i, \boldsymbol{\alpha}_{s_i}, \Sigma \overset{indep}{\sim} \text{log-N}_J(\mathbf{y}^\star \mid \boldsymbol{\alpha}_{s_i} + \mathbf{r}_i, \Sigma), \quad i = 1, \ldots, N, \tag{1}$$

$$\boldsymbol{\alpha}_{s_i} \mid G \overset{iid}{\sim} G(\boldsymbol{\alpha}), \quad s_i \in \{1, \ldots, S\}. \tag{2}$$

We will let G a random probability measure with a DP prior to flexibly accommodate variability in counts across m, s, and j. We will discuss a prior distribution for G later. We use a rounding function and obtain the distribution of $\mathbf{y}_i$ as follows;

$$\text{P}(\mathbf{y}_i = \mathbf{y} \mid \mathbf{r}_i, \boldsymbol{\alpha}_{s_i}, \Sigma) = \int_{A(\mathbf{y})} f_{\mathbf{y}^\star}(\mathbf{y}^\star \mid \boldsymbol{\alpha}_{s_i} + \mathbf{r}_i, \Sigma)\, d\mathbf{y}^\star, \tag{3}$$

where the region of integration $A(\mathbf{y}) = \{\mathbf{y}^\star \mid y_{11} \leq y_{11}^\star < y_{11}+1, \ldots, y_{MJ_M} \leq y_{MJ_M}^\star < y_{MJ_M}+1\}$ and $f_{\mathbf{y}^\star}(\cdot)$ is a pdf of a $J-$dim log-normal distribution with parameters $\boldsymbol{\alpha}_{s_i}+\mathbf{r}_i$ and Σ. $\boldsymbol{\alpha}_{s_i} = [\boldsymbol{\alpha}_{s_i1}, \ldots, \boldsymbol{\alpha}_{s_iM}]'$ is a $J-$dim vector of OTU abundances, where a subvector $\boldsymbol{\alpha}_{s_im} = (\alpha_{s_imj})$, $j = 1, \ldots, J_m$ is for group m. It is shared by all samples from subject s_i, and dependence among those samples is induced. $\mathbf{r}_i$ is a vector of sample scale factors, $\mathbf{r}_i = [r_{i1}\mathbf{1}_{J_1}, \ldots, r_{iM}\mathbf{1}_{J_M}]'$. From (1), $\exp(\alpha_{s_imj}+r_{im})$ is the median of $y_{imj}^\star$ and explains the location of the distribution of y_{imj} (i.e, raw OTU abundance). $\exp(r_{im})$ scales the location for all OTUs in group m of sample i, and r_{im}'s account for difference in total counts across (i,m) due to experimental artifacts. α_{s_imj} thus represents a normalized baseline abundance of OTU j of group m in a sample taken from subject s_i. The dependence structure of the counts can be inferred through a $J \times J$ covariance matrix, $\Sigma > 0$. Let $\Sigma_{jj'}^{mm'}$ denote the element of Σ corresponding to the covariance between OTU j of group m and OTU j' of group m'. Letting $\mu_{imj} = \alpha_{s_imj} + r_{im}$, we have $\mathrm{E}(y_{imj}^\star) = \exp(\mu_{imj} + \Sigma_{jj}^{mm}/2)$ and $\mathrm{Cov}(y_{imj}^\star, y_{im'j'}^\star) = \mathrm{E}(y_{imj}^\star)\mathrm{E}(y_{im'j'}^\star)\left\{\exp(\Sigma_{jj'}^{mm'})-1\right\}$, $m, m' \in \{1, \ldots, M\}$, $j \in \{1, \ldots J_m\}$ and $j' \in \{1, \ldots J_{m'}\}$. That is, Σ^{mm} and $\Sigma^{mm'}$ with $m \neq m'$ describe the within-group and across-group interaction structures, respectively. We will later extend the model to accommodate $\mathbf{x}_i$ through regression in μ_{imj}.

We next build a prior probability model for Σ, the parameter of primary interest. To overcome difficulties due to the high dimensionality, we assume that most pairs do not interact and consider joint sparsity, a structural assumption on Σ (also known as sparse spiked covariance structure) (Cai et al., 2016; Xie et al., 2022). The joint sparsity assumption allows to obtain a faster minimax rate of convergence for a frequentist estimator and improve posterior convergence for a Bayesian estimator. We first decompose a $J \times J$ covariance matrix Σ into $\Sigma = \Lambda\Lambda' + V$. Here, $\Lambda = [\Lambda_1', \ldots, \Lambda_m']'$ is a $J \times K$ factor loading matrix with $J \gg K$, where $\Lambda_m = [\lambda_{mjk}]$ is a $J_m \times K$ matrix. V is a J-dim diagonal matrix, where diagonal submatrices $V^{mm} = v_m^2\mathbf{I}_{J_m}$ and off-diagonal submatrices $V^{mm'} = \mathbf{0}_{J_m \times J_{m'}}$,

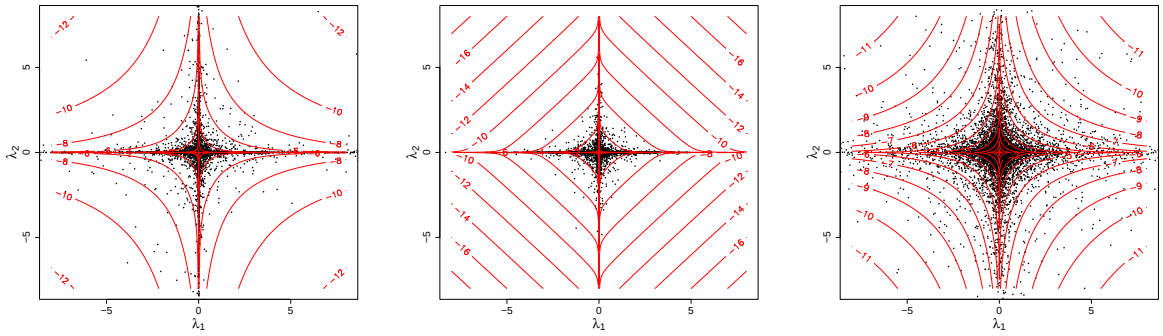
$m \neq m'$. The within-group and cross-group covariances are then $\Sigma^{mm} = \Lambda_m \Lambda_m' + V^{mm}$ and $\Sigma^{mm'} = \Lambda_m \Lambda_{m'}'$, $m \neq m'$. Under factor models, Λ are only identifiable up to orthogonal transformations. Our interest is primarily in the estimation of Σ, and this issue is not of great practical importance. We construct a Dirichlet-Horseshoe (Dir-HS) prior for columns $\boldsymbol{\lambda}_k$ of Λ to efficiently induce joint sparsity; for each k, $k = 1, \ldots, K$,

$$
\begin{aligned}
\tau_k \mid a_\tau, b_\tau &\overset{iid}{\sim} \mathrm{Ga}(a_\tau, b_\tau/J), \\
\boldsymbol{\phi}_k = (\phi_{11k}, \ldots, \phi_{MJ_Mk}) \mid a_\phi &\overset{iid}{\sim} \mathrm{Dir}(a_\phi, \ldots, a_\phi), \\
\zeta_{mjk} &\overset{iid}{\sim} \mathrm{C}^+(0,1), \quad m = 1, \ldots, M, \ j = 1, \ldots, J_m, \\
\lambda_{mjk} \mid \phi_{mjk}, \tau_k, \zeta_{mjk} &\overset{indep}{\sim} \mathrm{N}(0, \zeta_{mjk}^2 \phi_{mjk} \tau_k),
\end{aligned}
\tag{4}
$$

where $\mathrm{C}^+(0,1)$ represents the half-Cauchy distribution for $\mathbb{R}_+$ with location and scale parameters 0 and 1, and $\mathrm{Ga}(a,b)$ is the gamma distribution with mean a/b. For V, we assume $v_m^2 \mid a_v, b_v \overset{iid}{\sim} \text{inv-Ga}(a_v, b_v)$ with fixed a_v and b_v. In (4), $\boldsymbol{\phi}_k$ chooses active features (OTUs) for factor k. On the other hand, τ_k's globally control individual factors, and a small value of τ_k indicates that factor k is negligible in explaining dependence among the OTUs. The Dir-HS distribution can be derived by integrating $\boldsymbol{\phi}_k$ and ζ_{mjk} out. The Dir-HS density function lacks an analytic form, and the following theorem finds tight bounds for the marginal density of λ_{mjk} under the Dir-HS.

Theorem 2.1. *Let $J = 2$. Assume $\phi_1 \sim Be(a_\phi, a_\phi)$ and let $\phi_2 = 1 - \phi_1$. Assume the Dir-HS distribution in (4) as a joint distribution for $\boldsymbol{\lambda} = (\lambda_1, \lambda_2) \in \mathrm{R}^2$ given τ. Without loss of generality, let $\tau = 1$. The marginal density $\Pi_{Dir\text{-}HS}(\lambda_1)$ of λ_1 satisfies the following: (a) $\lim_{\lambda_1 \to 0} \Pi_{Dir\text{-}HS}(\lambda_1) = \infty$. (b) For $\lambda_1 \neq 0$,*

$$
\begin{aligned}
&2^{2a_\phi - \frac{5}{2}} \pi^{-2} \frac{\Gamma^2(a_\phi + 1/2)}{\Gamma(2a_\phi + 1/2)} \frac{4}{\lambda_1^2} \, {}_3F_2\left(1, 1, a_\phi + 1/2; 2, 2a_\phi + 1/2; -\frac{4}{\lambda_1^2}\right) \\
< \quad &\Pi_{Dir\text{-}HS}(\lambda_1) \quad < \quad 2^{2a_\phi - \frac{3}{2}} \pi^{-2} \frac{\Gamma^2(a_\phi + 1/2)}{\Gamma(2a_\phi + 1/2)} \frac{2}{\lambda_1^2} \, {}_3F_2\left(1, 1, a_\phi + 1/2; 2, 2a_\phi + 1/2; -\frac{2}{\lambda_1^2}\right),
\end{aligned}
\tag{5}
$$



(a) Dir-HS (b) Dir-Laplace (c) Independent HS

Figure 2: Scatter plots of (λ_1, λ_2) simulated from Dir-HS, Dir-Laplace and independent HS are illustrated in panels (a), (b) and (c), respectively. The contours represent their empirical density on the logarithmic scale.

where ${}_pF_q$ *is the generalized hypergeometric function,* ${}_pF_q(\alpha_1, \ldots, \alpha_p; \beta_1, \ldots, \beta_q; x) = \sum_{t=0}^{\infty} \frac{(\alpha_1)_t \ldots (\alpha_p)_t}{(\beta_1)_t \ldots (\beta_q)_t} \frac{x^t}{t!}$. *Especially when* $a_\phi = \frac{1}{2}$,

$$\frac{1}{\sqrt{2\pi^5}} \left\{\sinh^{-1}(2/|\lambda_1|)\right\}^2 \;<\; \Pi_{Dir\text{-}HS}(\lambda_1) \;<\; \sqrt{\frac{2}{\pi^5}} \left\{\sinh^{-1}(\sqrt{2}/|\lambda_1|)\right\}^2, \tag{6}$$

where the inverse hyperbolic sine function $\sinh^{-1}(x) = \log(x + \sqrt{x^2+1})$.

A proof is given in Supp. §1. From the theorem, the marginal density of λ_{mjk} has an unbounded spike at zero for any value of a_ϕ similar to a HS prior (Carvalho et al., 2009). It thus obtains severe shrinkage for λ_{mjk} when needed, while having tail robustness, and can achieve improved performance at handling unknown sparsity with a small number of large signals compared to other joint shrinkage priors such as the Dirichlet-Laplace (Dir-Laplace) prior (Bhattacharya et al., 2015). Fig 2(a) has a scatterplot of (λ_1, λ_2) simulated from the Dir-HS with $a_\phi = 1/20$ and $\tau = 1$. For comparison, panels (b) and (c) have scatterplots from the Dir-Laplace distribution and an independent HS distribution, respectively. Specifically, for the Dir-Laplace, we assume $\phi_1 \sim \text{Be}(a_\phi, a_\phi)$, let $\phi_2 = 1 - \phi_1$ and $\lambda_j \mid \phi_j \overset{indep}{\sim} \text{DE}(\tau\phi_j)$, $j = 1, 2$, where $\text{DE}(b)$ is the Laplace distribution with mean 0 and variance $2b^2$. For independent HS distributions, we assume $\lambda_j \mid \zeta_j \overset{indep}{\sim} \text{N}(0, \zeta_j^2/2)$

and $\zeta_j \overset{iid}{\sim} \mathrm{C}^+(0,1)$, $j = 1, 2$, to match the scale parameter with that under the Dir-HS. Comparing panel (a) to panel (b), the Dir-HS has heavier tails, leading to greater robustness to large signals. Supp. Proposition 1.1 examines the tails of the marginal densities $\Pi_{\text{Dir-HS}}(\lambda_1)$ and $\Pi_{\text{Dir-Laplace}}(\lambda_1)$ of λ_1 under the Dir-HS and Dir-Laplace and shows that $\lim_{\lambda_1 \to \pm\infty} \Pi_{\text{Dir-Laplace}}(\lambda_1)/\Pi_{\text{Dir-HS}}(\lambda_1) = 0$. Also, note that $\Pi_{\text{Dir-Laplace}}(\lambda_1)$ is bounded at 0 given τ when $a_\phi > 1$. The Dir-HS has a higher density along the axes than the independent HS in panel (c) and enables joint sparsity. Supp. Figs 1 and 2 plot joint and marginal densities of the distributions in the central origin and tail regions with various values of a_ϕ.

Previously, Zhao et al. (2016) built a group factor model with mean fixed at zero for continuous responses. They constructed a 'global-factor-local shrinkage' prior for the elements in a factor loading matrix for structured sparsity. Their prior was built with a hierarchical structure that includes global, factor-specific and element-specific hyperparameters. Note that their prior does not induce joint sparsity. Pati et al. (2014) built a factor model with a fixed mean at zero for a continuous response in a single group and considered the Dir-Laplace distribution on the vector constructed by concatenating factor loading vectors.

From (1)-(3), the marginal distribution of $\mathbf{y}_i$ can be obtained by integrating $\boldsymbol{\alpha}$ with respect to mixing distribution G. It is critical to improving the estimation of Σ that the model adequately accommodates large inter-subject variability in counts, which is a common issue in microbiome data analysis. We consider the following infinite mixture model for G in (2),

$$\begin{aligned} G(\boldsymbol{\alpha}) &= \prod_{m=1}^{M} \prod_{j=1}^{J_m} G_{mj}(\alpha_{mj}) \\ &= \prod_{m=1}^{M} \prod_{j=1}^{J_m} \left[\sum_{l=1}^{\infty} \psi_{ml}^{\alpha} \left\{ \omega_{ml}^{\alpha} \delta_{\xi_{mjl}^{\alpha}} + (1-\omega_{ml}^{\alpha}) \delta_{\left(\frac{\nu_{mj}^{\alpha} - \omega_{ml}^{\alpha} \xi_{mjl}^{\alpha}}{1-\omega_{ml}^{\alpha}} \right)} \right\} \right], \end{aligned} \tag{7}$$

where δ_ξ is a point mass centered at ξ. We assume $\xi_{mjl}^{\alpha} \mid \nu_{mj}^{\alpha}, u_\alpha^2 \overset{iid}{\sim} \mathrm{N}(\nu_{mj}^{\alpha}, u_\alpha^2)$ with fixed ν_{mj}^{α} and u_α^2. The mixture weights ψ_{ml}^{α} in (7) are constructed using a stick-breaking

process (Sethuraman, 1994); let $\psi^{\alpha}_{m1} = V^{\alpha}_{m1}$ and $\psi^{\alpha}_{ml} = V^{\alpha}_{ml}\prod_{l'=1}^{l-1}(1-V^{\alpha}_{ml'})$, $l > 1$ with $V^{\alpha}_{ml} \mid c^{\alpha} \overset{iid}{\sim} \text{Be}(1, c^{\alpha})$, where the total mass parameter c^{α} is fixed. Assume inner mixture weights $\omega^{\alpha}_{ml} \mid a^{\alpha}_{\omega}, b^{\alpha}_{\omega} \overset{iid}{\sim} \text{Be}(a^{\alpha}_{\omega}, b^{\alpha}_{\omega})$, where a^{α}_{ω} and b^{α}_{ω} are fixed. Observe that individual parameters α_{s_imj} and r_{im} in μ_{imj} are not identifiable due to the multiplicative structure, $\text{E}(\log(y^{\star}_{imj}) \mid \alpha_{s_imj}, r_{im}) = \alpha_{s_imj} + r_{im}$. Under (7), the prior and posterior means of α_{s_imj} are fixed at ν^{α}_{mj}, and $\text{E}(\log(y^{\star}_{imj}) \mid G_{mj}, r_{im})$ fixed at $\nu^{\alpha}_{mj} + r_{im}$. We will impose a similar constraint on the prior of r_{im} below. The constraints are placed to address potential issues with the identifiability. Note that μ_{imj}'s are identifiable, and Σ, a parameter of primary interest, can be identified. Despite the constraint, G can capture various patterns in the distribution of $\boldsymbol{\alpha}$ due to its inherent flexibility (Müller et al., 2015). Specifically, the distribution of $y^{\star}_{imj}$ can be written as a Dirichlet process mixture with a log-normal mixture kernel in Antoniak (1974). Also, the model in (7) allows to efficiently borrow information across subjects and across OTUs through its hierarchical structure and yield improved estimates of α_{s_imj}. In particular, ψ^{α}_{ml}'s and ω^{α}_{ml}'s are common weights for all OTUs in group m, while the mixture locations vary by j for each m.

Recall that r_{im} is a normalizing factor of group m of sample i. Similar to (7), we consider a flexible infinite mixture model for r_{im};

$$r_{im} \mid \psi^{r}_{ml}, \omega^{r}_{ml} \overset{indep}{\sim} H_m = \sum_{l=1}^{\infty} \psi^{r}_{ml} \left\{ \omega^{r}_{ml} \text{N}(\xi^{r}_{ml}, u^2_r) + (1-\omega^{r}_{ml}) \text{N}\left(\frac{\nu^{r}_{m} - \omega^{r}_{ml}\xi^{r}_{ml}}{1-\omega^{r}_{ml}}, u^2_r \right) \right\}, \quad (8)$$

where ν^{r}_{m} and u^2_r are fixed. The prior and posterior expectations of r_{im} are ν^{r}_{m} in (8), and $\text{E}(\log(y^{\star}_{imj}) \mid G_{mj}, H_m)$ fixed at $\nu^{\alpha}_{mj} + \nu^{r}_{m}$. Each group has different means, as indicated in our motivating application as illustrated in Supp. Fig 16. We jointly specify values of ν^{α}_{mj} and ν^{r}_{m} using observed counts. For example, we first fix ν^{r}_{m} at the average of the logarithm of the total count, $\nu^{r}_{m} = \frac{1}{N}\sum_{i=1}^{N} \log\left(\sum_{j=1}^{J_m} y_{imj}\right)$, and set $\nu^{\alpha}_{mj} = \frac{1}{N}\sum_{i=1}^{N}\{\log(y_{imj}+0.01) - \nu^{r}_{m}\}$. We consider the following priors for ψ^{r}_{ml}, ω^{r}_{ml} and ξ^{r}_{ml}; assume $\xi^{r}_{ml} \mid \nu^{r}_{m}, u^2_{\xi^r} \overset{iid}{\sim} \text{N}(\nu^{r}_{m}, u^2_{\xi^r})$, $\omega^{r}_{ml} \mid a^{r}_{\omega}, b^{r}_{\omega} \overset{iid}{\sim} \text{Be}(a^{r}_{\omega}, b^{r}_{\omega})$, $\psi^{r}_{m1} = V^{r}_{m1}$ and

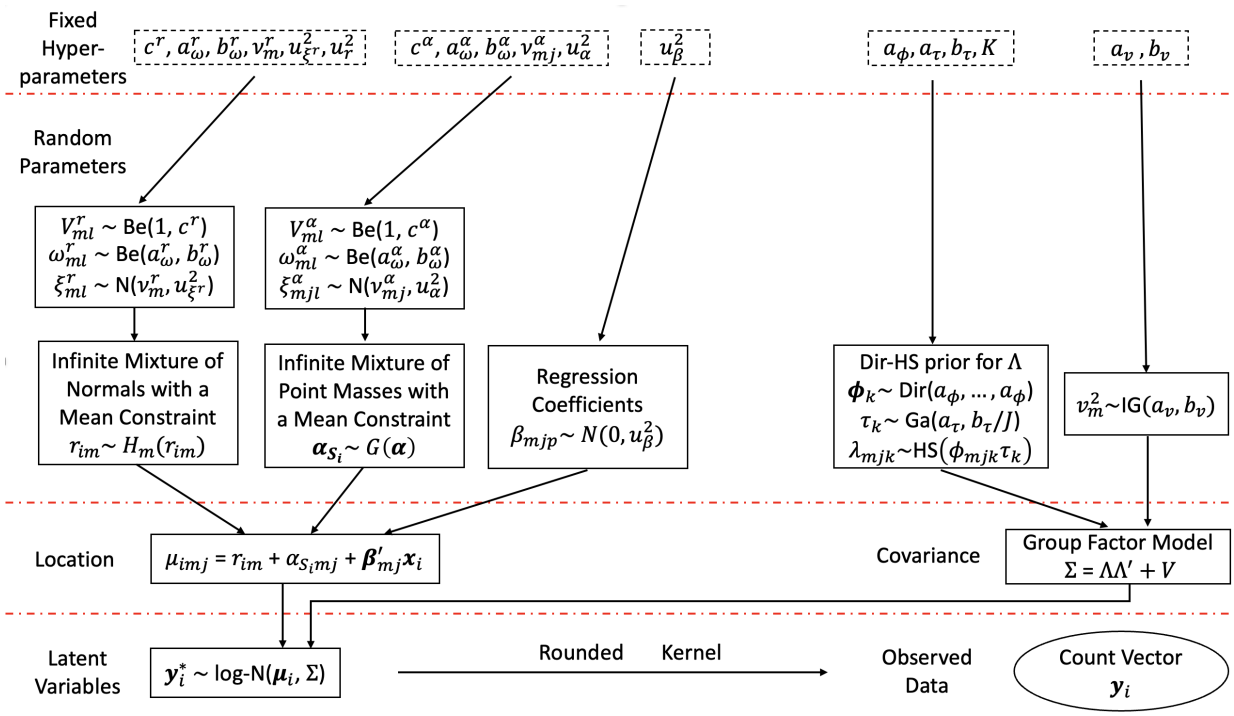


Figure 3: A graphical representation of Sp-BGFM. Fixed hyperparameters are in boxes with dashed lines, while random parameters are in boxes with solid lines. Observables are represented within circles.

$\psi^r_{ml} = V^r_{ml} \prod_{l'=1}^{l-1}(1 - V^r_{ml'})$, $l > 1$, where $V^r_{ml} \mid c^r \overset{iid}{\sim} \text{Be}(1, c^r)$. Here, $u^2_{\xi^r}$, a^r_ω, b^r_ω, and c^r are fixed.

In addition, the model is extended to accommodate covariates $\mathbf{x}_i$ using regression in μ_{imj};

$$\mu_{imj} = r_{im} + \alpha_{s_i mj} + \mathbf{x}'_i \boldsymbol{\beta}_{mj}. \tag{9}$$

Assume $\beta_{mjp} \overset{iid}{\sim} \text{N}(0, u^2_\beta)$ with fixed u^2_β. Regression coefficients β_{mjp} quantify the change in the abundance of OTU j of group m from its baseline abundance by x_{ip}. Especially, in a case of a categorical covariate, β_{mjp} shows an effect on the baseline abundance of the OTU for the level represented by x_p, and $\beta_{mjp} - \beta_{mjp'}$ can be used to infer the effect by the difference in levels between x_p and $x_{p'}$.

A graphical representation of Sp-BGFM is shown in Fig 3. In Supp. §2, we illustrate the

distribution of observables under Sp-BGFM to examine the distributions of OTUs' count. Specifically, how the model with (1)-(3) and (7) accommodates the dependence between OTU counts, excess zeros and large between-sample variability is illustrated with various examples.

2.2 Prior Calibration and Posterior Computation

The prior of Σ in (4) requires specification of fixed hyperparameters K, a_ϕ, a_τ and b_τ. The number K of latent factors is assumed to be fixed. For cases with $N \ll J$, a relatively small value of K is more desirable to obtain reliable estimation of Σ. For our simulation studies and real data analyses, we empirically set a value for K; we perform principle component analysis (PCA) for the sample covariance matrix of log-transformed normalized counts and fix K at a value such that the K largest eigenvalues explain 95% of the total variance. Given a sufficiently large value of K, the model may let τ_k close to 0 for unneeded latent factors. If desired, a prior can be considered for K, e.g., a geometric or truncated Poisson distribution. In addition, specifications of a_ϕ, a_τ and b_τ may need careful attention. Similar to Bhattacharya et al. (2015), we observed that estimates of λ_{mjk} tend to be overly shrunken toward zero with $a_\phi = 1/J$. We also observed that $a_\phi = 1/2$ recommended in Bhattacharya et al. (2015) for the Dir-Laplace distribution does not efficiently produce joint sparsity under the Dir-HS distribution. After careful exploration, we used $a_\phi = 1/(0.2 \times J)$, which gives approximately $1/20$ for a dataset with $J \approx 100$ as in our motivating example. By setting the scale parameter of τ_k to b_τ/J in (4), the prior for λ_{mjk} is appropriately scaled under the constraint $\sum_{m,j} \phi_{mjk} = 1$. We fixed $a_\tau = 0.1$ and $b_\tau = 1/J$ for the analyses in § 3 and § 4. We performed a thorough sensitivity analysis by varying the values of K, a_ϕ, a_τ, and b_τ and found that the model's performance remains robust within a reasonable range of these values. See Supp. §5 for sensitivity analyses related to the real data analysis in § 4.

Collecting terms, let $\boldsymbol{\theta} = \{\lambda_{mjk}, \phi_{mjk}, \tau_k, \zeta_{mkj}, v_m^2, \alpha_{s_i mj}, \omega_{ml}^\alpha, V_{ml}^\alpha, \xi_{mjl}^\alpha, r_{im}, \omega_{ml}^r, V_{ml}^r, \xi_{ml}^r,$

$\beta_{mjp}\}$ a vector of all random parameters. We utilize Markov Chain Monte Carlo (MCMC) simulations to generate samples of $\boldsymbol{\theta}$ from their posterior distribution. To facilitate the posterior computation, we introduce sample-specific latent vectors $\boldsymbol{\eta}_i \overset{iid}{\sim} \mathrm{N}_K(0, \mathrm{I}_K)$. We then have $y^\star_{imj} \mid \mu_{imj}, \boldsymbol{\lambda}_{mj}, \boldsymbol{\eta}_i, v^2_m \overset{indep}{\sim} \text{log-N}(\mu_{imj} + \boldsymbol{\lambda}'_{mj}\boldsymbol{\eta}_i, v^2_m)$ as independent log-normal variables, which results in significant computational efficiency. The joint posterior distribution of the augmented model is

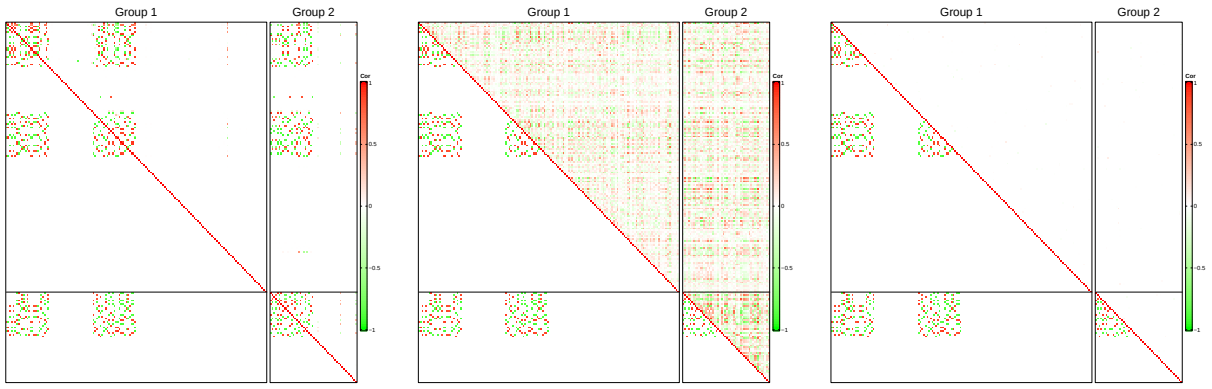
$$p(\boldsymbol{\theta}, \mathbf{y}^\star, \boldsymbol{\eta} \mid \mathbf{y}, \mathbf{x}) \propto \prod_{i=1}^{N}\prod_{m=1}^{M}\prod_{j=1}^{J_m} p(y_{imj} \le y^\star_{imj} < y_{imj} + 1 \mid \boldsymbol{\eta}_i, \boldsymbol{\theta}) \prod_{i=1}^{N} p(\boldsymbol{\eta}_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}). \tag{10}$$

We further augment the model by introducing latent variables to facilitate updates of $\mathbf{r}_i$, $\boldsymbol{\alpha}_{s_i}$, and ζ_{mkj}. We use the blocked Gibbs sampling algorithm (Ishwaran and James, 2001) by considering a finite-dimensional truncation of the stick-breaking processes in (7) and (8). We set the truncation levels L^r_m and L^α_m to sufficiently large values. Under the augmented model, all model parameters except $\boldsymbol{\phi}_k$ can be updated through Gibbs steps. We use adaptive MH algorithm (Haario et al., 2001) for an efficient update of $\boldsymbol{\phi}_k$. Details of the MCMC algorithm are in Supp. §3. R codes are available at https://github.com/shuang-jie/SP-BGFM.

3 Simulation

3.1 Simulation 1

For Simulation 1, we considered a case without covariates and evaluated the estimation of interaction between OTUs in two groups. We let $M = 2$ with $J_1 = 150$ and $J_2 = 50$ OTUs. We assumed one sample from each of $S = 20$ subjects, and we had $N = 20$. To specify Σ^{tr}, we let $K^{\text{tr}} = 5$. We then simulated $\lambda^{\text{tr}}_{mjk}$ from $\mathrm{N}(0, 1)$ and shifted away from zero by 1 for OTUs 1-25 and 51-75 in group 1 and OTUs 1-25 in group 2 to ensure that those OTUs have large covariances. For the remaining OTUs, we let $\lambda^{\text{tr}}_{mjk} = 0$ for all k. Thus, 80% of OTUs



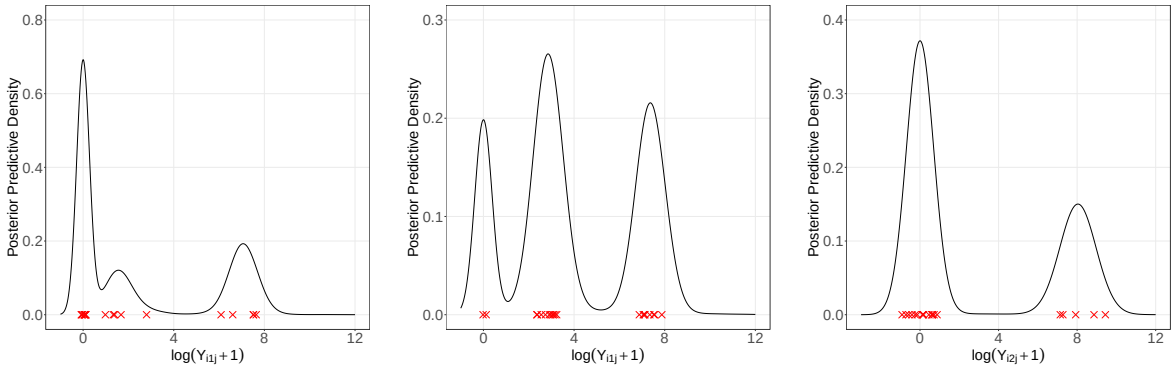
(a) Sp-BGFM (b) MOFA (c) SPIEC-EASI

Figure 4: [Simulation 1] The upper right and lower left triangles of a heatmap illustrate the estimates $\hat{\rho}^{mm'}_{jj'}$ of correlations and their truth, respectively. The horizontal and vertical lines are to divide the groups. The estimates in panels (a)-(c) are from Sp-BGFM, MOFA and SPIEC-EASI.

do not interact with the other OTUs. We then let $\Sigma^{\text{tr}} = \Lambda^{\text{tr}}\Lambda^{\text{tr},\prime} + V^{\text{tr}}$ with $v_m^{2,\text{tr}} = 0.5^2$ for all m. The correlation matrix corresponding to Σ^{tr} is illustrated in the lower triangle of Fig 4(a). For the normalized abundance level, we first set $\xi^{\alpha,\text{tr}}_{mj1} = -5$, $\xi^{\alpha,\text{tr}}_{mj2} \sim \text{N}(4,1)$ and $\xi^{\alpha,\text{tr}}_{mj3} \sim \text{N}(10,1)$ and simulated $\boldsymbol{\psi}^{\text{tr}}_{mj} = (\psi^{\text{tr}}_{mj1}, \psi^{\text{tr}}_{mj2}, \psi^{\text{tr}}_{mj3}) \sim \text{Dir}(30,40,30)$ independently for each (m,j). The three values, $\xi^{\alpha,\text{tr}}_{mjl}$, $l = 1$, 2 and 3, represent zero, small and large counts, respectively. We then let $\alpha^{\text{tr}}_{s_imj} = \xi^{\alpha,\text{tr}}_{mjl}$ with probability ψ^{tr}_{mjl} for $s_i \in \{1,\ldots,S\}$. We next simulated size factors $r^{\text{tr}}_{im} \overset{iid}{\sim} \text{Unif}(0,2)$. Finally, we generated $\mathbf{y}^{\star,\text{tr}}_i$ from log-$\text{N}_J(\boldsymbol{\mu}^{\text{tr}}_i, \Sigma^{\text{tr}})$ with $\boldsymbol{\mu}^{\text{tr}}_i = \mathbf{r}^{\text{tr}}_i + \boldsymbol{\alpha}^{\text{tr}}_{s_i}$ and obtain count vectors $\mathbf{y}_i = \lfloor \mathbf{y}^{\star,\text{tr}}_i \rfloor$. Under this setup, approximately 30% of y_{imj}'s are 0.

We specified the hyper-parameters values as discussed in § 2.2. In addition, we let $K = 10$, $c^r = c^\alpha = 1$, $L^r_m = L^\alpha_m = 50$, $a_v = b_v = 3$, $a^r_\omega = b^r_\omega = a^\alpha_\omega = b^\alpha_\omega = 5$. We ran MCMC for 10^5 iterations and discarded the first half for burn-in. It took 67 minutes on an Apple M1 chip laptop. We examined trace plots to assess the convergence and mixing of the MCMC chain and did not observe any evidence of slow mixing and convergence issues.

For easy interpretation, we consider correlations $\rho^{mm'}_{jj'} = \Sigma^{mm'}_{jj'}/(\Sigma^{mm}_{jj}\Sigma^{m'm'}_{j'j'})$ instead of Σ. Fig 4 (a) compares posterior median estimates $\hat{\rho}^{mm'}_{jj'}$ of correlations to their truth. As shown in the figure, Sp-BGFM capably identifies zeroes in the truth and efficiently shrinks



(a) Group 1, OTU 30 (b) Group 1, OTU 133 (c) Group 2, OTU 31

Figure 5: [Simulation 1] Posterior predictive estimates of the marginal distribution of log-transformed counts are plotted for three arbitrarily chosen OTUs, OTUs 30 and 133 of group 1 and OTU 31 of group 2 for model checking. Crosses are log-transformed observed counts after normalization based on a posterior estimate of the scale factors r_{im}.

the corresponding λ_{mjk} to zero, leading to an accurate reconstruction of the truth. We performed posterior predictive checking to assess model fit as follows; we first set the sample size factors $\mathbf{r}^{\text{pred}} = (r_1^{\text{pred}}, r_2^{\text{pred}})$ for an unobserved sample and estimated the posterior predictive distribution of a count vector, $\Pr(\mathbf{y}^{\text{pred}} = \mathbf{y} \mid \mathbf{r}^{\text{pred}}, \mathcal{D}) = \int_{A(\mathbf{y})} \int f(\tilde{\mathbf{y}}^{\star} \mid \mathbf{r}^{\text{pred}}, \boldsymbol{\theta}) f(\boldsymbol{\theta} \mid \mathcal{D}) d\boldsymbol{\theta} d\mathbf{y}$, where $\mathcal{D} = \{\mathbf{Y}_1, \mathbf{Y}_2\}$ denotes observed data. We approximated it with posterior samples of $\boldsymbol{\theta}$ drawn from the posterior simulation. Fig 5 illustrates marginal predictive distribution estimates of log-transformed counts for three arbitrarily chosen OTUs with $r_m^{\text{pred}} = 0$, $m = 1, 2$. If the model fits well, the observed data should look plausible under the posterior predictive distribution (Gelman et al., 2013). To avoid numerical issues, we added 1 to the posterior predictive samples of $\mathbf{y}$. The observed counts, marked with crosses in the figure, are also scaled according to $\mathbf{r}^{\text{pred}}$ after normalization by a posterior estimate of their scale factor for compatibility, $\log(\lfloor y_{imj} / \exp(\hat{r}_{im} - r_m^{\text{pred}}) \rfloor + 1)$, where $\hat{r}_{im}$ is a posterior estimate of r_{im}. The comparison of the predictive density estimates to the empirical distribution of the normalized observed counts suggests that the model offers a good fit to the data, accounting for excess zeros and multimodality, even with $N = 20$ for $J = 200$.

For comparison, we fit MOFA(Argelaguet et al., 2018) and SPIEC-EASI (Tipton et al., 2018) to the simulated data. We used R packages, *MOFA2* and *SpiecEasi* to apply their

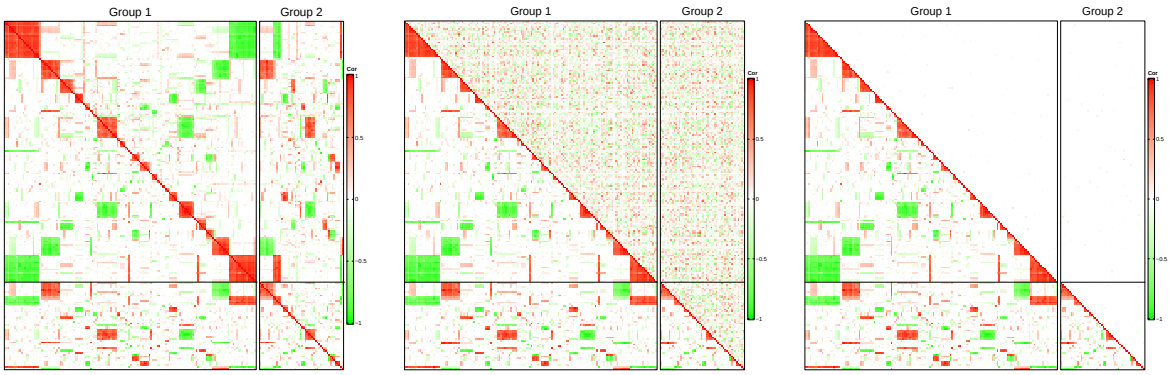
Method	Sim 1	Sim 2	Sim 3	Sim 4	Sim 5
Sp-BGFM	**0.065**	**0.123**	**0.234**	**0.000**	**0.109**
MOFA	0.229	0.364	0.316	0.107	0.235
SPIEC-EASI	0.150	0.306	0.306	0.004	0.205

Table 1: Root mean square error (RMSE) of the correlations $\rho_{jj'}^{mm'}$ is computed for Simulations 1-5. Estimates $\hat{\rho}_{jj'}^{mm'}$ are obtained from three methods, Sp-BGFM, MOFA and SPIEC-EASI. The smallest RMSE is in bold.

methods. Prior to fitting, the OTU counts were clr-transformed and re-centered with default settings in the packages. Their correlation estimates $\hat{\rho}_{jj'}^{mm'}$ are compared to the truth in Fig 4 (b)-(c). They yield poor estimates and fail to recover the true interaction structure, potentially due to their assumption of mean zero and/or the normalization of the observed counts prior to analysis. The root mean square error (RMSE) of $\rho_{jj'}^{mm'}$ is used to quantify the differences between the estimates from Sp-BGFM, MOFA, and SPIEC-EASI and the truth. The results are presented in Tab 1. Additional comparison of Sp-BGFM to REBACCA(Ban et al., 2015), COAT(Cao et al., 2019) and Zi-LN (Prost et al., 2021) that analyze a single count table, is provided in Supp. §4.1. Comparing their estimates to the truth, those alternative methods perform poorly in uncovering the true dependence among the OTUs.

3.2 Simulation 2

For Simulation 2, we set $M = 2$, $J_1 = 150$, $J_2 = 50$, $S = 20$ and $N = 40$ with a binary covariate. We used the vine method in Lewandowski et al. (2009) and generated an arbitrary covariance matrix to specify Σ^{tr}. The correlation matrix corresponding to Σ^{tr} is shown in the lower triangle of Fig 6(a). The OTUs are rearranged within a group for a better illustration. For abundances, we generated $\alpha_{s_imj}^{\text{tr}}$ and r_{im}^{tr} similarly as in Simulation 1, but we used the empirical proportions of zero counts from the multi-domain skin microbiome dataset in § 4 for $\alpha_{s_imj}^{\text{tr}}$ to simulate a dataset closely resembling the skin microbiome dataset. In addition, we incorporated a categorical covariate with two levels to investigate



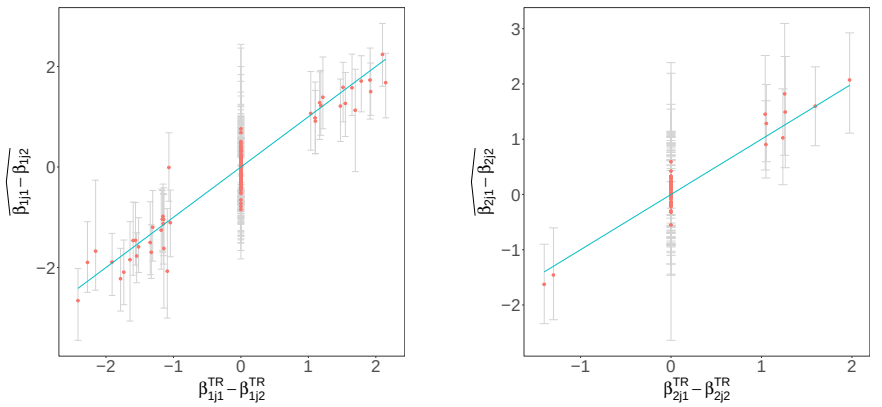
(a) Sp-BGFM (b) MOFA (c) SPIEC-EASI

Figure 6: [Simulation 2] The upper right and lower left triangles of a heatmap illustrate estimates $\hat{\rho}_{jj'}^{mm'}$ of correlations and their truth, respectively. The horizontal and vertical lines are to divide the groups. The estimates in panels (a)-(c) are from Sp-BGFM, MOFA and SPIEC-EASI, respectively.

the estimation of β_{mjp} and Σ in a complex setting. A sample was generated under each level for a subject, resulting in $N = 40$. We imposed sparsity on β^{tr} by letting them zero with a large probability. We then let $\mu_{imj}^{\text{tr}} = r_{im}^{\text{tr}} + \alpha_{s_i mj}^{\text{tr}} + \mathbf{x}_i' \boldsymbol{\beta}_{mj}^{\text{tr}}$ and generated $\mathbf{y}_i^{\star,\text{tr}}$ from log-N$_J(\boldsymbol{\mu}_i^{\text{tr}}, \Sigma^{\text{tr}})$. We finally let count vectors $\mathbf{y}_i = \lfloor \mathbf{y}_i^{\star,\text{tr}} \rfloor$, and the overall zero count rate is 45%. Details of simulation set-up are in Supp. §4.2.

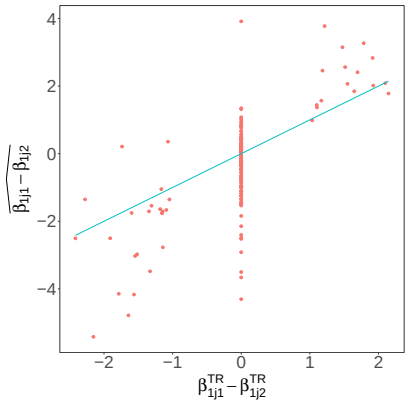
The fixed hyperparameters are specified the same as those in Simulation 1. For the prior of β_{mjp}, we set $u_\beta^2 = 3$. The MCMC simulation, consisting of 10^5 iterations, took approximately 98 minutes to complete on an Apple M1 chip laptop. We discarded the first half of the iterations as burn-in, and the remaining half was used for making inferences. The trace plots demonstrated a good mixing of the MCMC chain.

The upper triangle of Fig 6(a) illustrates the posterior estimates $\hat{\rho}_{jj'}^{mm'}$ under Sp-BGFM. Figs 7(a) and (b) show the posterior median estimates of $\beta_{mj1} - \beta_{mj2}$ (dots) with their 95% credible interval estimates (vertical lines) for groups 1 and 2, respectively. Sp-BGFM performs well in capturing the true within-domain and across-domain dependence structure among the OTUs, despite the arbitrary specification of Σ^{tr} and the added complexity due to the covariate in the true data generating process. In addition, the covariate effects β_{mjp} are well estimated.

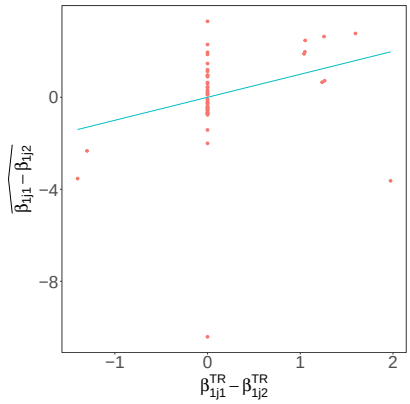


(a) Sp-BGFM: $\beta_{1j1} - \beta_{1j2}$

(b) Sp-BGFM: $\beta_{2j1} - \beta_{2j2}$



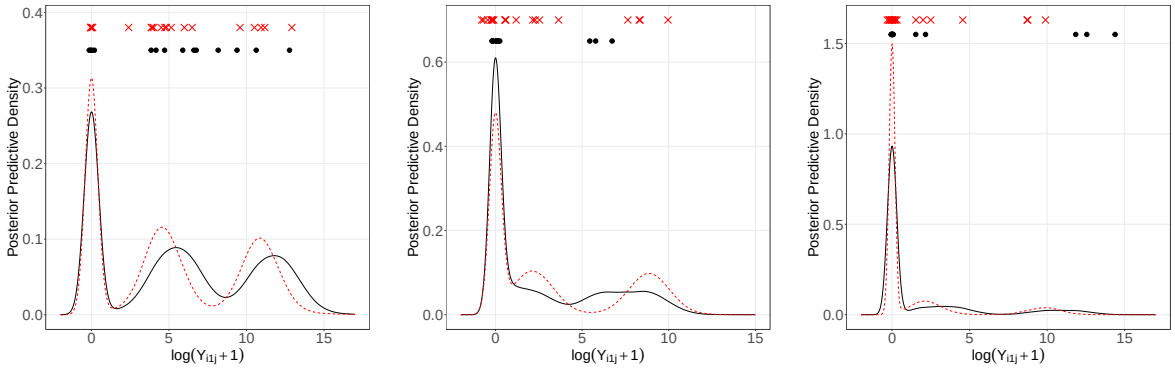
(c) metagenomeSeq: $\beta_{1j1} - \beta_{1j2}$



(d) metagenomeSeq: $\beta_{2j1} - \beta_{2j2}$

Figure 7: [Simulation 2] Posterior estimates of covariate effect $\beta_{mj1} - \beta_{mj2}$ under Sp-BGFM are plotted against the truth in panels (a) and (b) for two groups, $m = 1$ and 2. The posterior median estimates are denoted by dots, and the 95% credible estimates with vertical lines. In panels (c) and (d), the estimates of β_{mjp} under metagenomeSeq are plotted for two groups.

We also check the model fit using posterior predictive checking. We set $r_m^{\text{pred}} = 0$ for $m = 1, 2$ and estimate the distribution of $\mathbf{y}^{\text{pred}}$ for the two conditions, $\mathbf{x} = (1, 0)$ and $(0, 1)$, similar to the procedure used in Simulation 1. The predictive distribution estimates are illustrated in Fig 8 for some selected OTUs. The solid and dashed lines are for conditions, $\mathbf{x} = (1, 0)$ and $(0, 1)$, respectively. The observed normalized counts are shown with dots and crosses on the top of the figures after log transformation. For the OTUs in the figure, posterior estimates of $\beta_{mj1} - \beta_{jm2}$, are 1.68, -2.65 and 2.07 with 95% credible intervals (0.98, 2.26), (-3.44, -2.02), and (1.11, 2.92), respectively. Their true values are 2.15, -2.42, and 1.97, respectively. The figures show an adequate model fit under Sp-BGFM and depict



(a) Group 1 OTU 12 (e) Group 1 OTU 32 (f) Group 2 OTU 11

Figure 8: [Simulation 2] Posterior predictive estimates of the marginal distribution of log-transformed counts for three arbitrarily chosen OTUs, OTUs 1 and 32 of group 1 and OTU 161 of group 2 for model checking. Dots and crosses are log-transformed observed counts after normalization based on a posterior estimate of the scale factors r_{im} for $\mathbf{x}$ =(1, 0) and (0, 1), respectively. The solid and dashed lines represent the conditions with $\mathbf{x}$ =(1, 0) and (0, 1), respectively.

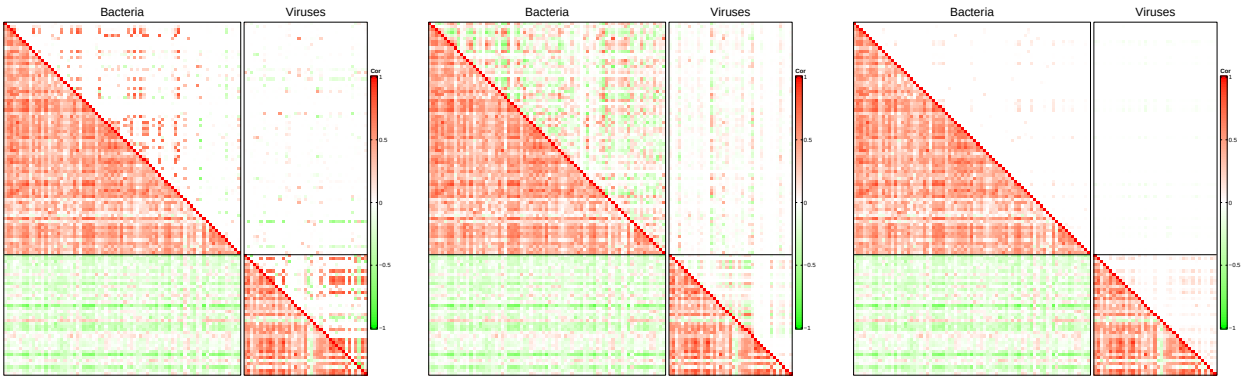
the covariate's impact on the prediction of counts for those OTUs.

Figs 6(b) and (c) compare the correlation estimates obtained from MOFA and SPIEC-EASI to the truth. For Sp-BGFM, MOFA, and SPIEC-EASI, RMSEs of $\rho_{jj'}^{mm'}$ are computed and shown in Tab 1. The estimates from the additional comparators, REBACCA, COAT and Zi-LN, are shown in Supp. Fig. 7. The estimates of the comparators are very poor and fail to recover Σ^{tr}, potentially due to a lack of consideration for covariates and/or assumption of mean zero. In addition, we compare our Sp-BGFM to metagenomeSeq (Paulson et al., 2013) in the estimation of β_{mjp}. MetagenomeSeq transforms counts $\log_2(y_{imj}+1)$ and builds a zero-inflated normal mixture model. For the non-zero part, the mean function is modeled through regression. It uses the CSS normalization method to estimate sample size factors and includes as an offset to account for differences between samples in sequencing depth. Figs 7(c) and (d) illustrate point estimates of $\beta_{mj1} - \beta_{mj2}$ under metagenomeSeq. MetagenomeSeq does not provide interval estimates. Comparison of the plots in panels (a) and (b) to those in panels (c) and (d) suggests that Sp-BGFM offers more accurate estimates of covariate effects with uncertainty quantification.

3.3 Additional Simulations

We conducted additional simulation studies, Simulations 3, 4, and 5, to further examine the robustness of Sp-BGFM. In Simulation 3, we kept the setup of Simulation 2 and used Σ^{tr} arbitrarily specified by the vine method in Lewandowski et al. (2009) to generate data. However, no covariate was considered. Sp-BGFM recovers the true microbial interaction structure well, as shown in Supp. Fig. 8. In Simulation 4, we simulated count vectors from multinomial distributions, where the total count, i.e., the number of trials, was simulated from a normal distribution whose parameters were empirically specified using the real dataset in § 4. The true OTU dependence structure is well recovered under Sp-BGFM, as shown in Supp. Fig 10. Especially, Supp. Fig 11 illustrates that the model-based normalization through r_{im} provides a reasonable basis for estimating $\boldsymbol{\alpha}$ and Σ. For Simulation 5, we generated a multi-domain count dataset using the functions in R package *SpiecEasi* (Kurtz et al., 2015). The functions take a real microbiome count dataset and a correlation matrix as input and generate a count table from a zero-inflated negative binomial distribution through normal-copula functions. OTU counts have a dependence structure as in the provided correlation matrix, and their marginal distributions are similar to those in the provided dataset. We used the multi-domain skin microbiome dataset in § 4 and correlation matrices randomly generated by the vine method. Supp. Fig 13 demonstrates that Sp-BGFM does an excellent job of capturing the true within-domain and cross-domain dependence structure and provides a reasonable fit to the simulated data, although the dataset was generated from a model significantly different from the assumed model.

For comparisons, we fit the comparators, MOFA and SPEIC-SASI, to the datasets of Simulations 3-5 and compared their results to the truth and those of Sp-BGFM, indicating favorable performance of Sp-BGFM. The RMSEs of $\rho_{jj'}^{mm'}$ are computed for Sp-BGFM and the comparators, and they are presented in Tab 1. Details of Simulations 3-5 are reported in §4.3-§4.5 of the Supplementary Materials, respectively.



(a) Sp-BGFM (b) MOFA (c) SPIEC-EASI

Figure 9: [Multi-domain skin microbiome] The upper right triangle of the heatmaps in (a)-(c) has correlation estimates $\hat{\rho}_{jj'}^{mm'}$ under Sp-BGFM, MOFA and SPIEC-EASI, respectively. Empirical correlation estimates $\widetilde{\rho}_{jj'}^{mm'}$ are shown in the lower triangles.

4 Multi-domain Skin Microbiome Data Analysis

To fit Sp-BGFM for the multi-domain skin microbiome data, we removed OTUs having extremely low counts on average or having zero counts in too many samples. In particular, we included only the OTUs that have a non-zero count in at least two samples under each condition and an average count larger than ten under each condition for analysis. After pre-processing, 75 bOTUs and 39 vOTUs were left for analysis, so $J_1 = 75$ and $J_2 = 39$. The proportions of zeros are 42.97% and 44.10% for bOTUs and vOTUs, respectively. Empirical correlation estimates $\widetilde{\rho}_{jj'}^{mm'}$ among the OTUs are computed using the OTU counts normalized using CSS, and illustrated in the lower triangle of Fig 9(a). We used $K = 15$, and all other hyperparameters were specified at the same values as in the simulation studies of § 3. We implemented posterior inference using MCMC posterior simulation. The Markov chain ran for 10^5 iterations, and the initial half was discarded as burn-in. The posterior simulation took approximately 4.82 minutes for every 10,000 iterations on an Apple M1 chip laptop. The trace plots indicated that the MCMC chain mixed well. We also performed sensitivity analysis on the specification of the fixed hyperparameters. Details of MCMC simulation diagnostics and prior sensitivity analyses are included in Supp. §5.

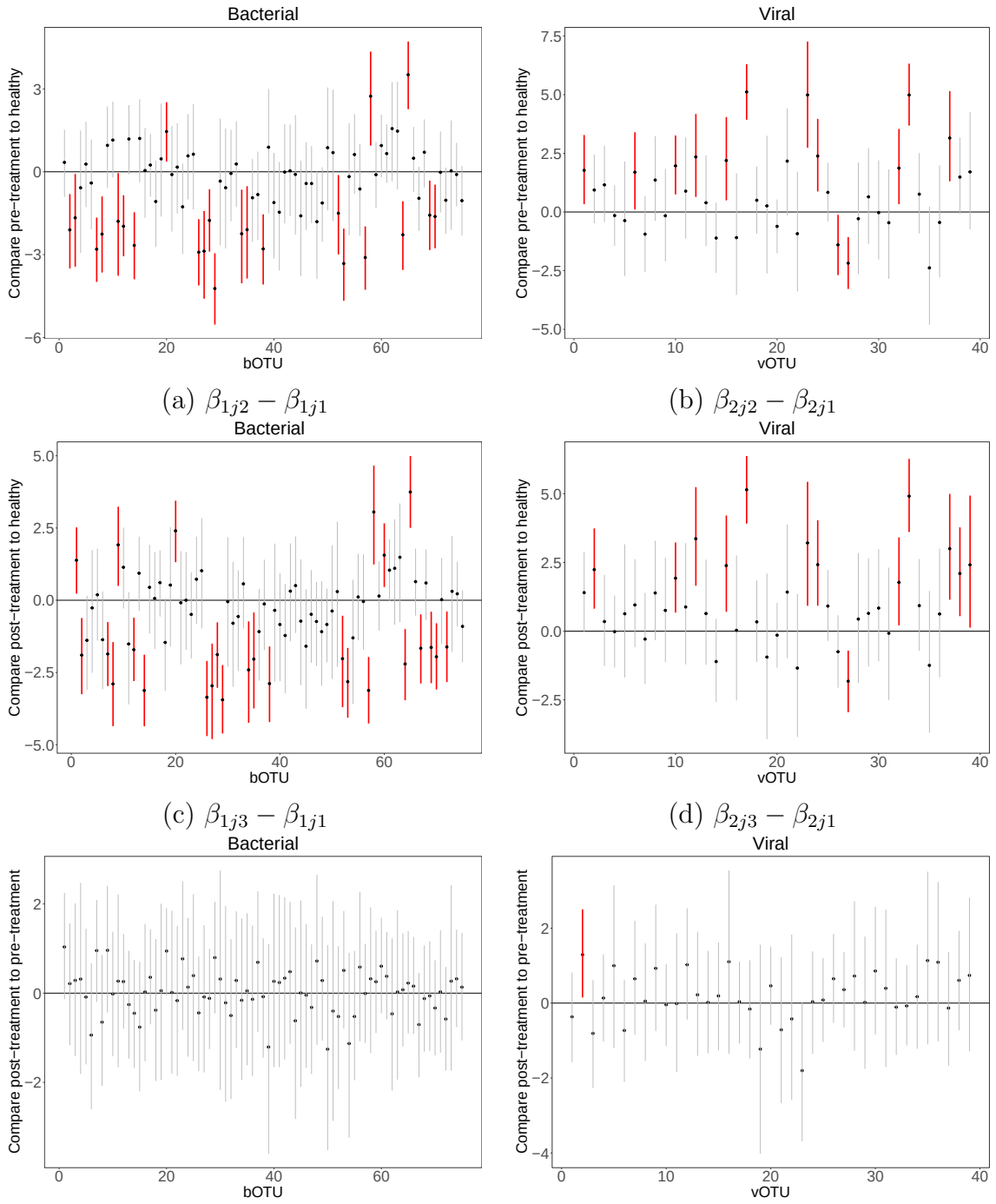
The upper right triangle of Fig 9(a) illustrates posterior median estimates $\hat{\rho}_{jj'}^{mm'}$ of cor-

relations. The OTUs are rearranged within a group for a better illustration. Supp. Fig 17 illustrates $\hat{\rho}_{jj'}^{mm'}$ for the OTUs that have $|\hat{\rho}_{jj'}^{mm'}| > 0.5$ with any other OTU j', $j' \neq j$. Supp. Tabs 1 and 2 have taxonomic information of those OTUs. Here, 0.5 is an arbitrary choice to illustrate a smaller set of OTUs that have large estimates. While the overall estimated interaction structure is sparse, some OTU subsets within a group have large positive values of $\hat{\rho}_{jj'}^{mm}$. Interestingly, many of these OTUs have zero counts across samples concurrently, potentially suggesting potential microbial co-existence patterns. Positive correlations among bacteria are expected because some bacterial infections are known to be polymicrobial. That is, infections occur with microorganisms from different genera. Specifically, the genera, *Actinomyces*, *Actinotignum*, *Campylobacter*, *Helcococcus* and *Porphyromonas*, which are bOTUs 3, 4, 10, 24 and 56, respectively, have large positive correlation estimates with $\hat{\rho}_{jj'}^{mm} \geq 0.72$, $m = 1$. Previous research has indicated potential relations between some of the species of those OTUs. *Actinomyces* and *Helcococcus*, which are bacteria that can adapt and survive in environments with or without oxygen, were found in diabetic patients with osteomyelitis, a serious bone infection typically in the foot (Van Asten et al., 2016). Additionally, *Actinomyces*-associated infections are frequently found to occur with other bacteria including *Campylobacter* and *Porphyromonas* that might synergistically enhance the infection process (Könönen and Wade, 2015). In the oral microbiome, species of *Actinomyces*, *Campylobacter*, and *Porphyronomas* are also known to be related to periodontal diseases (Noiri et al., 1997). Synergistic interactions between the microbes of these OTUs have not been found in chronic wounds. However, the identified positive correlations align with previous findings under other biological contexts and support further investigations into the relationship between these bacterial species in the context of chronic wound healing. In addition, vOTUs 2, 9, 10, 13, 29, 32, 34 and 38 are estimated to have $\hat{\rho}_{jj'}^{mm} \geq 0.65$, $m = 2$ with each other, implying that they coexist and their abundance is related with that of the others. Especially, vOTUs 2, 9, 10 and 13, corresponding to *Aquisalimonas* phage,

Grimontella phage, *Klebsiella* phage, and *Methylomonas* phage, are annotated. With the exception of *Klebsiella* which is a pathogen in the human microbiome, little is known about those phage hosts. The positive correlation estimates among those vOTUs may reflect the richness or scarcity of the common environment, as virion production is influenced by environmental factors such as nutrient availability. Correlations among the phages reflect potential interactions among the hosts, the phages, or the phages and hosts, and the results may suggest the need for further studies to gain additional biological context.

Different from the previous analyses that focused on a single domain, Sp-BGFM provides inference on interactions among microorganisms in both within and different domains. From Fig 9(a), the overall cross-domain interaction is scarce, except for *Staphylococcus aureus* (bOTU 65), a prominent skin pathogen. Interestingly, it has a negative correlation estimate with a subset of phages, vOTU 2, 6, 8, 9, 10, 13, 28, 29, 31, 32, 34, 36 and 38, that are positively correlated with each other. The colonization of *S. aureus* is found associated with disruption in the healthy composition of skin microbiota (Di Domenico et al., 2019). The negative correlations may suggest potential adversarial relationships between *S. aureus* and these phages (or their host) and call for further investigation to enhance our understanding of the underlying biological process. Additionally, the pair, *Pseudomonas* (bOTU 59) and *Pseudomonas* phage (vOTU 18), is estimated to have a positive correlation 0.38, aligning with their inherent ecological relations (i.e., *Pseudomonas* phage occurs with *Pseudomonas* bacteria).

In contrast to MOFA and SPEICE-EASI, Sp-BGFM also produces inferences on mean microbial abundances and their association with covariates. Fig 10 illustrates inference on covariate effects $\beta_{mjp} - \beta_{mjp'}$, $p \neq p'$. Recall that β_{mjp}, $p = 1$, 2 and 3, quantify changes in abundance compared to the baseline abundance. In the figure, dots represent the posterior median estimates of $\beta_{mjp} - \beta_{mjp'}$, while vertical lines illustrate their 95% credible interval estimates. The interval estimates that do not contain zero are in red. Supp. Tabs 1



(a) $\beta_{1j2} - \beta_{1j1}$ (b) $\beta_{2j2} - \beta_{2j1}$

(c) $\beta_{1j3} - \beta_{1j1}$ (d) $\beta_{2j3} - \beta_{2j1}$

(e) $\beta_{1j3} - \beta_{1j2}$ (f) $\beta_{2j3} - \beta_{2j2}$

Figure 10: [Multi-domain skin microbiome] The left and right columns display the posterior median estimates of $\beta_{mjp} - \beta_{mjp'}$ for bacterial and viral OTUs, respectively. Vertical lines represent their corresponding 95% credible interval estimates. The interval estimates that do not include 0 are marked in red bold.

and 2 have taxonomic information of the OTUs whose interval estimates do not contain zero. Overall, the bOTUs tend to be enriched in the healthy condition compared to the

pre- and post-treatment conditions. In contrast, vOTUs tend to be enriched in the pre- and post-treatment conditions. Changes in abundance between pre- and post-treatment conditions are relatively minimal for bOTUs and vOTUs. This could be due to the fact that the post-treatment samples were taken quite quickly after the treatment, while any significant changes might take longer to occur. Within the wound samples, vOTUs 1, 18 and 23, corresponding to *Acinetobacter* phage, *Proteus* phage and *Staphylococcus* phage, are found enriched as also reported in Verbanic et al. (2022). Similar to the findings in Fig 2 of Verbanic et al. (2020), bOTUs 27, 29 and 53, corresponding to the genera, *Kocuria*, *Micrococcus* and *Paracoccus*, are significantly more abundant in the healthy skin samples. Interestingly, the abundance of vOTU 2 (*Aquisalimonas* phage) is found to be statistically significantly different between the pre- and post-treatment conditions. Little is known about this phage, and the result suggests follow-up experiments for further examination.

Supp. Fig. 18 illustrates posterior predictive density estimates of an OTU's count under the different conditions for some selected OTUs, bOTUs 1, bOTU 69 and vOTU 17, for model assessment. The figure also demonstrates the effects of the experimental conditions on the prediction. Overall, the comparison of the posterior predictive density estimates to empirical distributions of the observed counts indicates a reasonable model fit to the data.

For comparison, we applied MOFA and SPIEC-EASI to the skin microbiome data. Fig 9(b) and (c) illustrate $\hat{\rho}_{jj'}^{mm'}$ under the comparators. The inference under MOFA suggests a large number of interactions compared to that under Sp-BGFM. While some interactions have been identified, such as the interaction between *Staphylococcus* and other species (Alonzo III, 2022; Christensen et al., 2016), it is unclear whether the high number of interactions aligns with the relative scarcity of known interspecies interactions in the skin and the lack of universal dynamics compared to the gut microbiome (Bashan et al., 2016). On the other hand, in contrast, SPIEC-EASI does not suggest any significant interactions and fails to capture interactions related to known mechanisms for chemical communication

among species (e.g., secreted by *Staphylococcus* species). The estimates from the additional comparators, REBACCA, COAT and Zi-LN, are in Supp. Fig. 9. Supp. Fig. 10 illustrates estimates of covariate effects under metagenomeSeq. The point estimates of coefficients under metagenomeSeq suggest that abundance of the bOTUs tends to be higher in the healthy condition compared to the post-treatment condition, which is similar to the inference under Sp-BGFM. However, it does not provide any uncertainty associated with the point estimates, and their statistical significance cannot be determined. Note that the comparators for estimating OTU interactions do not take into account covariates, and metagenomeSeq that estimates covariate effects does not consider potential interactions among OTUs.

5 Conclusions

We developed Sp-BGFM, a sparse Bayesian group factor model for analyzing multiple count tables data from multi-domain microbiome studies. The Dir-HS distribution was developed to efficiently induce joint sparsity and used as a prior for factor loadings. The model produces a reliable estimate of covariance matrices even with small sample sizes. Additionally, Sp-BGFM incorporates nonparametric mixtures of multivariate rounded kernels to capture inter-subject variability and improves inference on the dependence structure. The model also accommodates covariates through regression. Simulation studies and real data analysis confirm the robust performance of Sp-BGFM compared to other alternatives. The model is applicable to the analysis of multiple count tables data in any application.

Sp-BGFM can be extended by relaxing model assumptions further. One possible extension is to incorporate a hierarchical Dirichlet process (HDP) in Teh et al. (2004) or to adopt a common atom model in Denti et al. (2023). These approaches facilitate the construction of domain and OTU-specific distributions through a hierarchical structure. Specifically, an HDP allows G_{mj} in (2) to share mixture components, with mixture weights differing

across OTUs. Another extension incorporates a fully nonparametric regression model to accommodate covariates $\mathbf{x}$ more flexibly. This can be achieved using a dependent Dirichlet process (DDP) model (MacEachern, 1999; Quintana et al., 2022) by letting ψ_{ml}^{α} and/or $\xi_{mjl}^{\star}$ of G_{mj} in (2) depend on $\mathbf{x}$. The distribution of $\mathbf{y}$ is marginally a DP-distributed random probability distribution that varies flexibly with $\mathbf{x}$. It is important to note that while these extended models offer greater flexibility, obtaining inference with reasonable uncertainty bounds may require a sufficiently large sample size.

A potentially interesting avenue for further research is to integrate taxonomy rank information into analysis. In microbiome studies, utilizing a phylogenetic tree from 16S rRNA gene sequencing can enhance OTU interaction estimation (Washburne et al., 2018). For example, Chung et al. (2022) incorporated branch split information using a latent position model and a truncated Gaussian copula model. Adapting a similar idea, Sp-BGFM can include taxonomy level-specific factor loadings, denoted as Λ_m^T. Assigning OTUs latent factor loadings based on their phylogeny may allow to capture interaction structures integrating phylogenetic relatedness. This approach has the potential to enhance the inference of interaction structures in other domains.

SUPPLEMENTARY MATERIAL

SUPPLEMENTARY: The Supplement includes an examination of the properties of the Dir-HS distribution and the distributions of OTUs' count under Sp-BGFM. It also provides a detailed description of the MCMC sampling algorithm. Additionally, the supplement presents further results from simulation studies and the analysis of multi-domain skin microbiome data. (pdf file)

FUNDING DETAILS

This work was supported by the NIH under Grant DP2GM123457 and R35GM148249 (Irene A. Chen); and NSF under Grant DMS-2015428 (Juhee Lee).

References

Alonzo III, F. (2022). Toward Uncovering the Complexities of Bacterial Interspecies Communication and Competition on the Skin. *Mbio*, 13:e01320–22.

Antoniak, C. E. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2:1152–1174.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14:e8124.

Bach, F. R. and Jordan, M. I. (2005). A probabilistic interpretation of canonical correlation analysis. Technical report, Department of Statistics, University of California, Berkeley.

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

Bashan, A., Gibson, T. E., Friedman, J., Carey, V. J., Weiss, S. T., Hohmann, E. L., and Liu, Y.-Y. (2016). Universality of Human Microbial Dynamics. *Nature*, 534:259–262.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet–Laplace Priors for Optimal Shrinkage. *Journal of the American Statistical Association*, 110:1479–1490.

Browne, M. W. (1979). The Maximum-Likelihood Solution in Inter-Battery Factor Analysis. *British Journal of Mathematical and Statistical Psychology*, 32:75–86.

Cai, T. T., Ren, Z., and Zhou, H. H. (2016). Estimating Structured High-Dimensional Covariance and Precision Matrices: Optimal Rates and Adaptive Estimation. *Electronic Journal of Statistics*, 10(1):1 – 59.

Canale, A. and Dunson, D. B. (2011). Bayesian Kernel Mixtures for Counts. *Journal of the American Statistical Association*, 106:1528–1539.

Cao, Y., Lin, W., and Li, H. (2019). Large Covariance Estimation for Compositional Data via Composition-Adjusted Thresholding. *Journal of the American Statistical Association*, 114:759–772.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling Sparsity via the Horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80. PMLR.

Christensen, G. J., Scholz, C. F., Enghild, J., Rohde, H., Kilian, M., Thürmer, A., Brzuszkiewicz, E., Lomholt, H. B., and Brüggemann, H. (2016). Antagonism between Staphylococcus Epidermidis and Propionibacterium Acnes and Its Genomic Basis. *BMC genomics*, 17:1–14.

Chung, H. C., Gaynanova, I., and Ni, Y. (2022). Phylogenetically Informed Bayesian Truncated Copula Graphical Models for Microbial Association Networks. *The Annals of Applied Statistics*, 16:2437–2457.

Denti, F., Camerlenghi, F., Guindani, M., and Mira, A. (2023). A Common Atoms Model for the Bayesian Nonparametric Analysis of Nested Data. *Journal of the American Statistical Association*, 118(541):405–416. PMID: 37089274.

Di Domenico, E. G., Cavallo, I., Capitanio, B., Ascenzioni, F., Pimpinelli, F., Morrone, A., and Ensoli, F. (2019). Staphylococcus Aureus and the Cutaneous Microbiota Biofilms in the Pathogenesis of Atopic Dermatitis. *Microorganisms*, 7:301.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, 9:432–441.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Haario, H., Saksman, E., and Tamminen, J. (2001). An Adaptive Metropolis Algorithm. *Bernoulli*, 7(2):223 – 242.

Ishwaran, H. and James, L. F. (2001). Gibbs Sampling Methods for Stick-breaking Priors. *Journal of the American Statistical Association*, 96:161–173.

Klami, A., Virtanen, S., and Kaski, S. (2013). Bayesian Canonical Correlation Analysis. *Journal of Machine Learning Research*, 14:965–1003.

Klami, A., Virtanen, S., Leppäaho, E., and Kaski, S. (2014). Group Factor Analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 26:2136–2147.

Könönen, E. and Wade, W. G. (2015). Actinomyces and Related Organisms in Human Infections. *Clinical Microbiology Reviews*, 28:419–442.

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLoS Computational Biology*, 11:e1004226.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating Random Correlation Matrices Based on Vines and Extended Onion Method. *Journal of Multivariate Analysis*, 100(9):1989–2001.

MacEachern, S. N. (1999). Dependent Nonparametric Processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, volume 1, pages 50–55. Alexandria, Virginia. Virginia: American Statistical Association; 1999.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*, volume 1. Springer.

Noiri, Y., Ozaki, K., Nakae, H., Matsuo, T., and Ebisu, S. (1997). An Immunohistochemical Study on the Localization of Porphyromonas Gingivalis, Campylobacter Rectus and Actinomyces Viscosus in Human Periodontal Pockets. *Journal of Periodontal Research*, 32:598–607.

Pati, D., Bhattacharya, A., Pillai, N. S., and Dunson, D. (2014). Posterior contraction in sparse bayesian factor models for massive covariance matrices. *The Annals of Statistics*, 42:1102–1130.

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential Abundance Analysis for Microbial Marker-gene Surveys. *Nature Methods*, 10:1200–1202.

Peters, B. M., Jabra-Rizk, M. A., O'May, G. A., Costerton, J. W., and Shirtliff, M. E. (2012). Polymicrobial Interactions: Impact on Pathogenesis and Human Disease. *Clinical Microbiology Reviews*, 25:193–213.

Prost, V., Gazut, S., and Brüls, T. (2021). A Zero Inflated Log-Normal Model for Inference of Sparse Microbial Association Networks. *PLoS Computational Biology*, 17:e1009089.

Quintana, F. A., Müller, P., Jara, A., and MacEachern, S. N. (2022). The Dependent Dirichlet Process and Related Models. *Statistical Science*, 37(1):24 – 41.

Sethuraman, J. (1994). A Constructive Definition of the Dirichlet Prior. *Statistica Sinica*, 4:639–650.

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2004). Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes. *Advances in Neural Information Processing Systems*, 17.

Tian, C., Jiang, D., Hammer, A., Sharpton, T., and Jiang, Y. (2023). Compositional Graphical Lasso Resolves the Impact of Parasitic Infection on Gut Microbial Interaction Networks in a Zebrafish Model. *Journal of the American Statistical Association*, 118(543):1–15.

Tipton, L., Müller, C. L., Kurtz, Z. D., Huang, L., Kleerup, E., Morris, A., Bonneau, R., and Ghedin, E. (2018). Fungi Stabilize Connectivity in the Lung and Skin Microbial Ecosystems. *Microbiome*, 6:1–14.

Van Asten, S., La Fontaine, J., Peters, E., Bhavan, K., Kim, P., and Lavery, L. (2016). The Microbiome of Diabetic Foot Osteomyelitis. *European Journal of Clinical Microbiology & Infectious Diseases*, 35:293–298.

Verbanic, S., Deacon, J. M., and Chen, I. A. (2022). The Chronic Wound Phageome: Phage Diversity and Associations with Wounds and Healing Outcomes. *Microbiology Spectrum*, 10:e02777–21.

Verbanic, S., Shen, Y., Lee, J., Deacon, J. M., and Chen, I. A. (2020). Microbial Predictors of Healing and Short-term Effect of Debridement on the Microbiome of Chronic Wounds. *NPJ Biofilms and Microbiomes*, 6:1–11.

Virtanen, S., Klami, A., Khan, S., and Kaski, S. (2012). Bayesian Group Factor Analysis. In Lawrence, N. D. and Girolami, M., editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1269–1277, La Palma, Canary Islands. PMLR.

Washburne, A. D., Morton, J. T., Sanders, J., McDonald, D., Zhu, Q., Oliverio, A. M., and Knight, R. (2018). Methods for Phylogenetic Analysis of Microbiome Data. *Nature Microbiology*, 3:652–661.

Xie, F., Cape, J., Priebe, C. E., and Xu, Y. (2022). Bayesian Sparse Spiked Covariance Model with a Continuous Matrix Shrinkage Prior. *Bayesian Analysis*, 17(4):1193 – 1217.

Zhang, S., Shen, Y., Chen, I. A., and Lee, J. (2023). Bayesian Modeling of Interaction between Features in Sparse Multivariate Count Data with Application to Microbiome Study. *The Annals of Applied Statistics*, 17(3):1861 – 1883.

Zhao, S., Gao, C., Mukherjee, S., and Engelhardt, B. E. (2016). Bayesian Group Factor Analysis with Structured Sparsity. *Journal of Machine Learning Research*, 17(196):1–47.