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

July 18, 2023

Abstract

Group factor models have been developed to infer relationships between multiple co-occurring multivariate continuous responses. Motivated by complex count data from multi-domian 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.

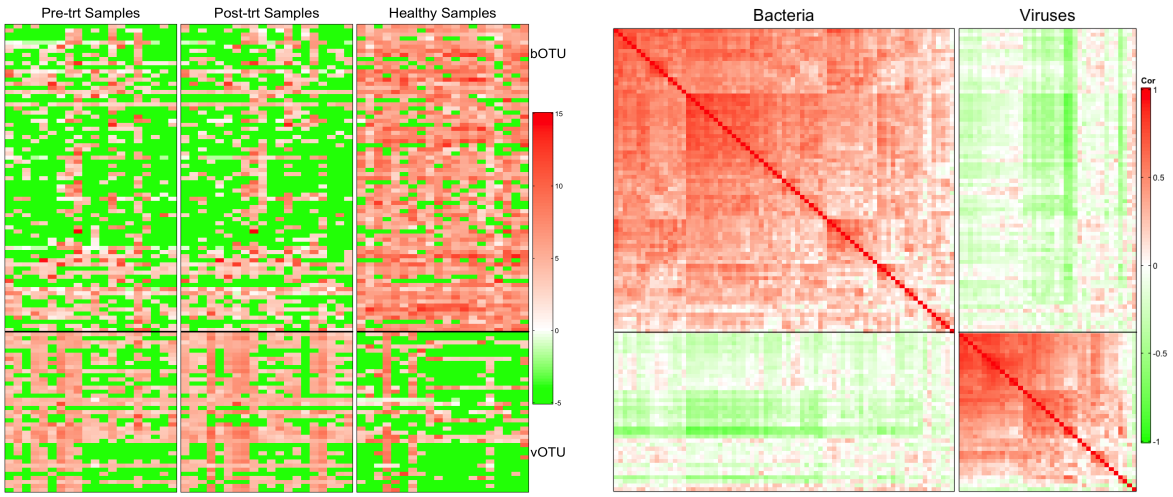
*Address for Correspondence: 1156 High St, Santa Cruz, CA 95064. E-mail: szhan209@ucsc.edu.
†The work is conducted during UCLA and the current affiliation is ByteDance Research.

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.

The proposed method is motivated by a dataset from the multi-domain skin microbiome study in Verbanic et al. (2020); Zhang et al. (2023); Verbanic et al. (2022). The dataset consists of *multiple count tables*, with each count table representing a specific 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. Microorganisms, including bacteria, viruses, fungi, and archaea, coexist in diverse communities within the human body (Peters et al., 2012). However, most previous studies on the human microbiome have primarily focused on a single domain alone (usually bacteria) or have examined different microbial domains separately. In the context of chronic wounds, the microbiome comprises a mixture of different microbial species, forming polymicrobial communities. The motivating study investigated bacteria and bacteriophages (bacterial viruses) in the wound microbiome. Bacteriophages play a role in regulating bacterial 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



(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.

has received relatively limited attention. While 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, 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. Statistical methods that account for the discreteness of data with multiple count responses can be crucial to appropriately infer the intricate interactions among microorganisms, both within a specific domain and across different domains, as well as their associations with the environment, potentially leading to understanding of healing of chronic wounds.

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 debridement 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 from the patients. 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. The figure indicates 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 sequenc-

ing 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 6 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. 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. 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 avoid identifiability problems. The covariance matrix of the kernel, which is the parameter of main interest for understanding microorganism interactions, is estimated through latent factors. For a joint prior of factor loading vectors, we construct the Dirichlet-Horseshoe distribution to efficiently induce joint sparsity, and the model provides reliable inferences on a high-dimensional interaction structure both within and across the 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

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

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

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 the details 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 $A(\mathbf{y}) = \{\mathbf{y}^\star \mid y_{11} \leq y^\star_{11} < y_{11} + 1, \ldots, y_{MJ_M} \leq y^\star_{MJ_M} < y_{MJ_M} + 1\}$ 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. In (1), $\boldsymbol{\alpha}_{s_i} = [\boldsymbol{\alpha}_{s_i1}, \ldots, \boldsymbol{\alpha}_{s_iM}]'$ 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. $\mathbf{r}_i$ is a vector of sample scale factors, $\mathbf{r}_i = [r_{i1}\mathbf{1}_{J_1}, \ldots, r_{iM}\mathbf{1}_{J_M}]'$, where r_{im} is a scalar. 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. It is shared by all samples from subject s_i, and dependence among those samples is induced. The entire joint distribution of $\mathbf{y}$ in (3) can be used to infer the marginals and dependence structure of the counts. Let $\mu_{imj} = \alpha_{s_imj} + r_{im}$. $\exp(\mu_{imj})$ represents the median of $y^\star_{imj}$ and explains the location of the distribution of y_{imj} (i.e, OTU abundance). $\Sigma > 0$ is a $J \times J$ covariance matrix, and let $\Sigma^{mm'}_{jj'}$ denote the element of Σ corresponding to the covariance between OTU j of group m and OTU j' of group m'. We have $\mathrm{E}(y^\star_{imj}) = \exp(\mu_{imj} + \Sigma^{mm}_{jj}/2)$ and $\mathrm{Cov}(y^\star_{imj}, y^\star_{im'j'}) = \mathrm{E}(y^\star_{imj})\mathrm{E}(y^\star_{im'j'})\left\{\exp(\Sigma^{mm'}_{jj'}) - 1\right\}$, $m, m' \in \{1, \ldots, M\}$, $j \in \{1, \ldots J_m\}$ 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 $J \times K$ 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 $V^{mm'} = \mathbf{0}_{J_m \times J_{m'}}$, $m \neq m'$. The within-group and cross-group covariances are $\Sigma^{mm} = \Lambda_m\Lambda_m' + V^{mm}$ and $\Sigma^{mm'} = \Lambda_m\Lambda_{m'}'$, $m \neq m'$. We construct a Dirichlet-Horseshoe (Dir-HS) prior for columns $\boldsymbol{\lambda}_k$ of Λ to efficiently induce

joint sparsity; for each k, $k = 1, \ldots, K$,

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

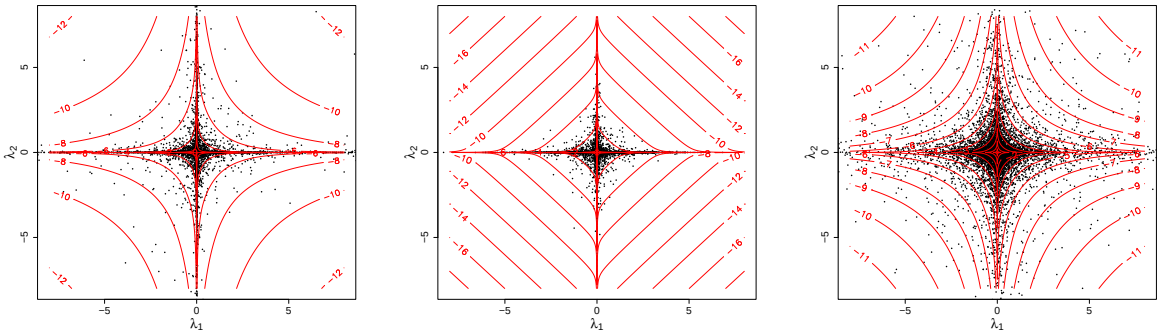
where $\mathrm{C}^+(0,1)$ represents the half-Cauchy distribution for $\mathbb{R}_+$ with location and scale parameters 0 and 1, and $\mathrm{Ga}(a,b)$ is the gamma distribution with mean a/b. For V, we assume $v_m^2 \mid a_v, b_v \overset{iid}{\sim} \text{inv-Ga}(a_v, b_v)$ with fixed a_v and b_v. In (4), while $\boldsymbol{\phi}_k$ chooses active features (OTUs) for factor k, τ_k's globally control individual factors. A small value of τ_k indicates that factors 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 $\lambda_1 \neq 0$,*

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

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 $a_\phi = \frac{1}{2}$,

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



(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 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 any 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 $\zeta_j \overset{iid}{\sim} \text{C}^+(0,1)$, $j = 1, 2$ 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 for continuous responses. They constructed a 'global-factor-local shrinkage' prior for the elements in a factor loading matrix for structured sparsity. The 'global-factor-local shrinkage' 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. Bhattacharya et al. (2015) built a factor model for a continuous response in a single group and considered the Dir-Laplace distribution on the vector constructed by concatenating factor loading vectors. Under factor models, Λ are only identifiable up to orthogonal transformations. Our interest is primarily on the estimation of Σ, and this issue is not of great practical importance.

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),

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

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 $\psi_{m1}^{\alpha} = V_{m1}^{\alpha}$ and $\psi_{ml}^{\alpha} = V_l^{\alpha}\prod_{l'=1}^{l-1}(1-V_{ml'}^{\alpha})$, $\ell > 1$ with $V_{ml}^{\alpha} \mid c^\alpha \overset{iid}{\sim} \mathrm{Be}(1, c^\alpha)$, where the total mass parameter c^α is fixed. Assume inner mixture weights $\omega_{ml}^{\alpha} \mid a_\omega^\alpha, b_\omega^\alpha \overset{iid}{\sim} \mathrm{Be}(a_\omega^\alpha, b_\omega^\alpha)$, where a_ω^α and b_ω^α are fixed. Under the model in (2) and

(7), given r_{im} the prior and posterior medians of $y^{\star}_{imj}$ are fixed at $\exp(r_{im}+\nu^{\alpha}_{mj})$. We will impose a similar constraint on the prior of r_{im} to avoid potential identifiability problems. Individual parameters α_{s_imj} and r_{im} are not identifiable, but μ_{imj}'s are identifiable. 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 $\mathbf{y}^{\star}_{im}$ is a Dirichlet process mixture with a log-normal mixture kernel (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};

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

where ν^{r}_{m} and u^{2}_{r} are fixed. The prior and posterior expectations of r_{im} are ν^{r}_{m} in (8). Each group has different means, as indicated in our motivating application as illustrated in Supp. Fig 6(a) and (b). 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 $\xi^{r}_{ml} \mid \nu^{r}_{m}, u^{2}_{\xi^r} \overset{iid}{\sim} \mathrm{N}(\nu^{r}_{m}, u^{2}_{\xi^r})$. Also, let $\omega^{r}_{ml} \mid a^{r}_{\omega}, b^{r}_{\omega} \overset{iid}{\sim} \mathrm{Be}(a^{r}_{\omega}, b^{r}_{\omega})$, $\psi^{r}_{m1} = V^{r}_{m1}$ and $\psi^{r}_{ml} = V^{r}_{ml}\prod_{l'=1}^{l-1}(1-V^{r}_{ml'})$, $l>1$, where $V^{r}_{ml} \mid c^{r} \overset{iid}{\sim} \mathrm{Be}(1, c^{r})$. 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};

$$\mu_{imj} = r_{im} + \alpha_{s_imj} + \mathbf{x}_i'\boldsymbol{\beta}_{mj}. \quad (9)$$

Assume $\beta_{mjp} \overset{iid}{\sim} \mathrm{N}(0, u_\beta^2)$ with fixed u_β^2. 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 $x_{p'}$.

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.

Collecting terms, let $\boldsymbol{\theta} = \{\lambda_{mjk}, \phi_{mjk}, \tau_k, \zeta_{mkj}, v_m^2, \alpha_{s_imj}, \omega_{ml}^{\alpha}, V_{ml}^{\alpha}, \xi_{mjl}^{\alpha}, r_{im}, \omega_{ml}^{r}, V_{ml}^{r}, \xi_{ml}^{r}, \beta_{mjp}\}$ 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 $y_{imj}^{\star} \mid \mu_{imj}, \boldsymbol{\lambda}_{mj}, \boldsymbol{\eta}_i, v_m^2 \overset{indep}{\sim} \text{log-N}(\mu_{imj} + \boldsymbol{\lambda}_{mj}'\boldsymbol{\eta}_i, v_m^2)$ as independent log-normal variables, which results in significant computational efficiency. The joint posterior distribution of the augmented model is

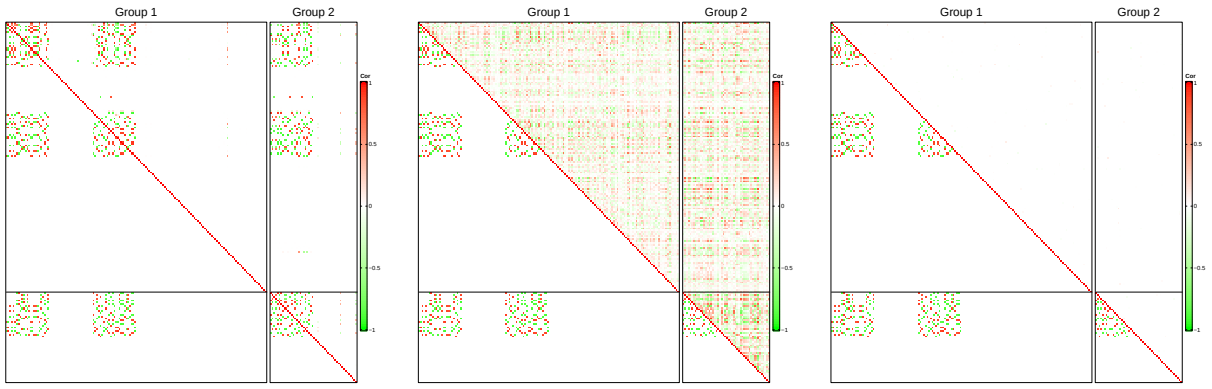
$$p(\boldsymbol{\theta}, \mathbf{y}^{\star}, \boldsymbol{\eta} \mid \mathbf{y}, \mathbf{x}) \propto \prod_{i=1}^{N} \prod_{m=1}^{M} \prod_{j=1}^{J_m} p(y_{imj} \leq y_{imj}^{\star} < y_{imj} + 1 \mid \boldsymbol{\eta}_i, \boldsymbol{\theta}) \prod_{i=1}^{N} p(\boldsymbol{\eta}_i \mid \boldsymbol{\theta})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

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_m^r and L_m^{α} 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. §2. An R package is available at https://github.com/Zsj950708/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 $\lambda_{mjk}^{\text{tr}}$ 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



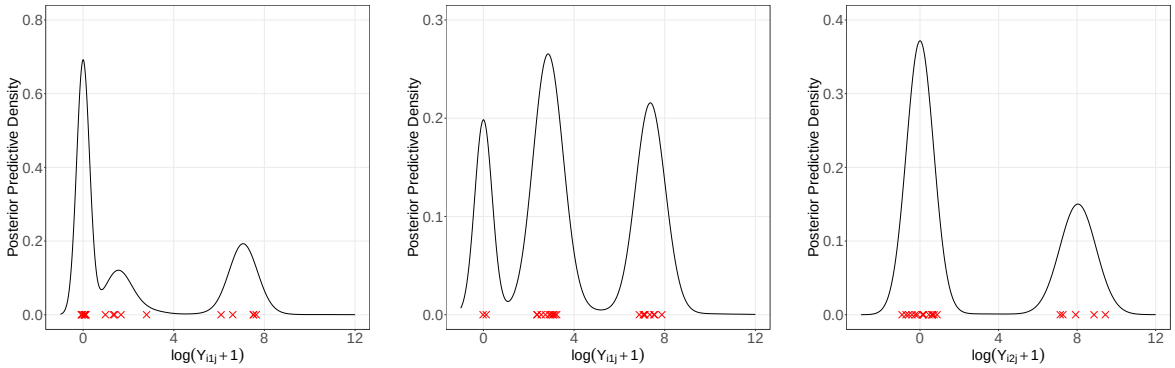
(a) Sp-BGFM (b) MOFA (c) SPIEC-EASI

Figure 3: [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.

OTUs have large covariances. For the remaining OTUs, we let $\lambda^{\text{tr}}_{mjk} = 0$ for all k. Thus, 80% of OTUs do not interact with the other OTUs. We then let $\Sigma^{\text{tr}} = \Lambda^{\text{tr}}\Lambda^{\text{tr},\prime} + V^{\text{tr}}$ with $v^{2,\text{tr}}_m = 0.5^2$ for all m. Σ^{tr} is illustrated in the lower triangle of Fig 3(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-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 3 (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

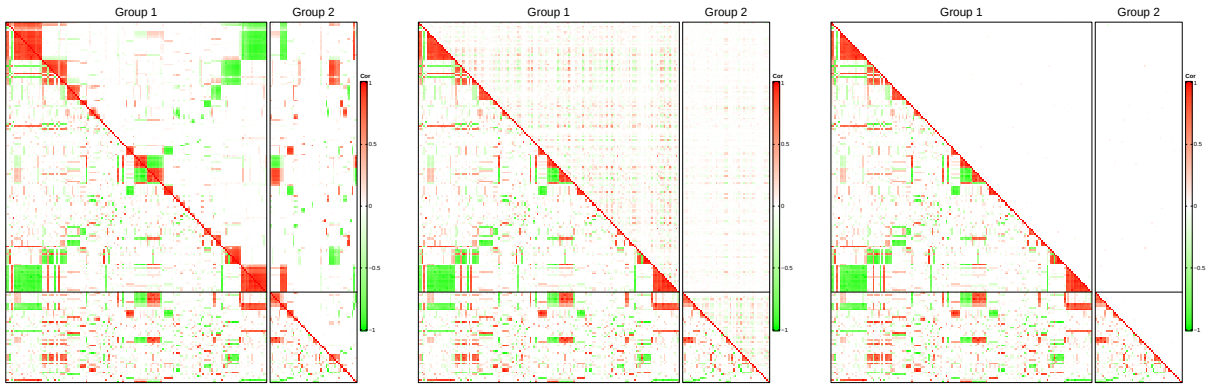


(a) Group 1, OTU 30 (b) Group 1, OTU 133 (c) Group 2, OTU 31

Figure 4: [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 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 4 illustrates marginal predictive distribution estimates of log-transformed counts for three arbitrarily chosen OTUs with $r_m^{\text{pred}} = 0$, $m = 1, 2$. To avoid numerical issues, we added 1 to the posterior predictive samples of $\mathbf{y}$. Log-transformed observed counts are shown with crosses after normalization by a posterior estimate of their scale factor, i.e., $\log(\lfloor y_{imj} / \exp(\hat{r}_{im} - r_m^{\text{pred}}) \rfloor + 1)$, 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 methods. Prior to fitting, the OTU counts were clr-transformed and re-centered with



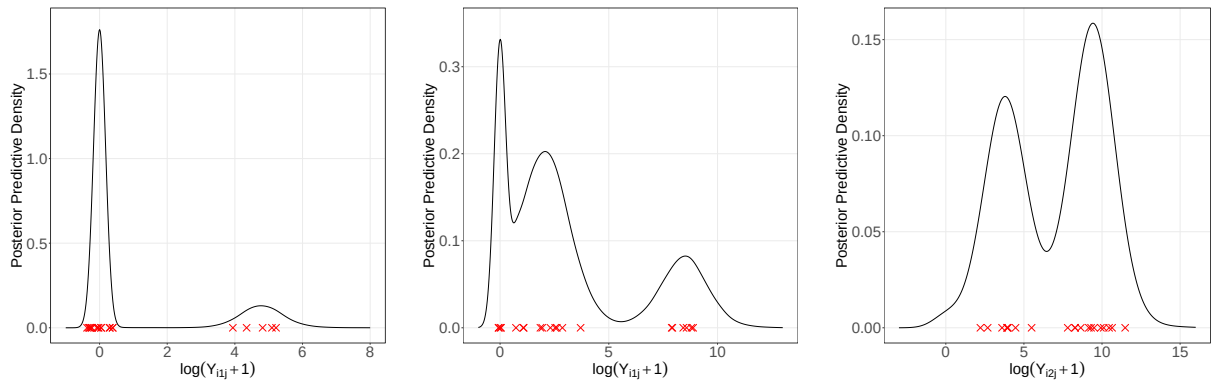
(a) Sp-BGFM (b) MOFA (c) SPIEC-EASI

Figure 5: [Simulation 2] The upper right and lower left triangles of a heatmap illustrate the 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.

default settings in the packages. Their correlation estimates $\hat{\rho}_{jj'}^{mm'}$ are compared to the truth in Fig 3 (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. 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. §3. 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 kept $M = 2$, $J_1 = 150$, $J_2 = 50$, $S = 20$ and $N = 20$ the same as in Simulation 1. We generated an arbitrary covariance matrix to specify Σ^{tr}; we used the vine method in Lewandowski et al. (2009) and generated a random $J \times J$ correlation matrix based on partial correlations. In particular, we simulated partial correlations from linearly transformed Be$(1, 1)$ distribution over the interval of $(-1, 1)$. To encourage sparsity in Σ^{tr}, we set the partial correlations below 0.8 to 0 and generated a correlation matrix, $\rho_{jj'}^{mm',\text{tr}}$ using their recursive formula. We then sampled v_{mj}^{tr} independently from Unif$(1, 1.5)$ and

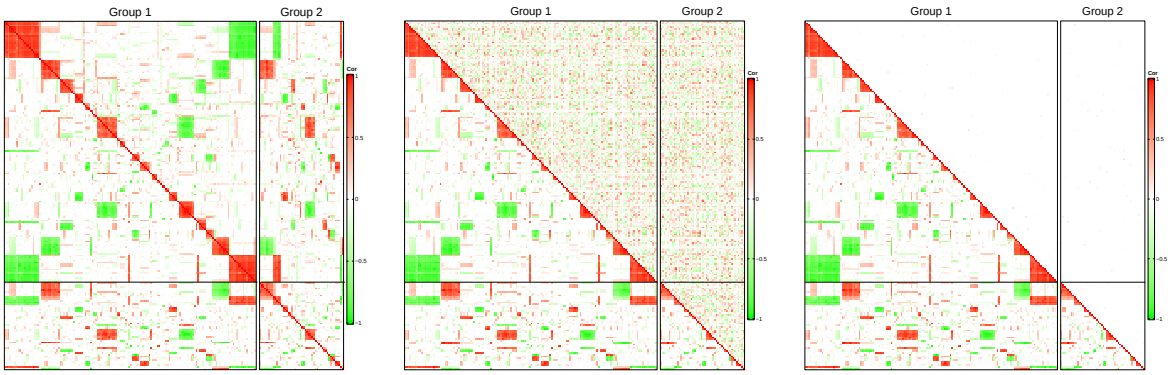


(a) Group 1 OTU 67 (b) Group 1 OTU 118 (c) Group 2 OTU 47

Figure 6: [Simulation 2] Posterior predictive estimates of the marginal distribution of log-transformed counts for three arbitrarily chosen OTUs, OTUs 67 and 118 of group 1 and OTU 47 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}.

let $\Sigma_{jj'}^{mm',\mathrm{tr}} = v_{mj}^{\mathrm{tr}} v_{m'j'}^{\mathrm{tr}} \rho_{jj'}^{mm',\mathrm{tr}}$. Σ^{tr} is shown in the lower triangle of Fig 5(a). The OTUs are rearranged within a group for a better illustration. For abundances, we computed the empirical proportions $\tilde{\psi}_{mj}$ of zero counts in the multi-domain skin microbiome dataset in § 4. To set the values of ψ_{mj1}^{tr} for a group, we sampled with replacement from the corresponding set of $\tilde{\psi}_{mj}$. We let $\psi_{mj2}^{\mathrm{tr}} = 0.6 \times (1 - \psi_{mj1}^{\mathrm{tr}})$ and $\psi_{mj2}^{\mathrm{tr}} = 0.4 \times (1 - \psi_{mj1}^{\mathrm{tr}})$. We then specified $\xi_{mjl}^{\alpha,\mathrm{tr}}$ and simulated $\boldsymbol{\alpha}_{s_i}^{\mathrm{tr}}$, $\mathbf{r}_i^{\mathrm{tr}}$, $\mathbf{y}_i^{\star}$ and $\mathbf{y}_i$ the same as done in Simulation 1. Approximately 40% of the counts in the dataset were zero, which is comparable to the proportion of zeros in the skin microbiome dataset. We used the same fixed hyperparameter values as in Simulation 1, and we approximated the posterior distribution using MCMC. The examination of the MCMC simulation using traceplots indicated no evidence of convergence or mixing problems.

Fig 5(a) compares posterior estimates $\hat{\rho}_{jj'}^{mm'}$ of correlations (upper triangle) to the truth (lower triangle). Recall that $\rho_{jj'}^{mm',\mathrm{tr}}$ is specified arbitrarily. Sp-BGFM effectively recovers the underlying interaction structure with a high degree of accuracy even in a case of $N = 20$ and $J = 200$. To assess the fit of the model, we compared predictive distribution estimates to the empirical distribution of the normalized observed counts, using a procedure the same as that employed in Simulation 1. Marginal posterior predictive distribution estimates of



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

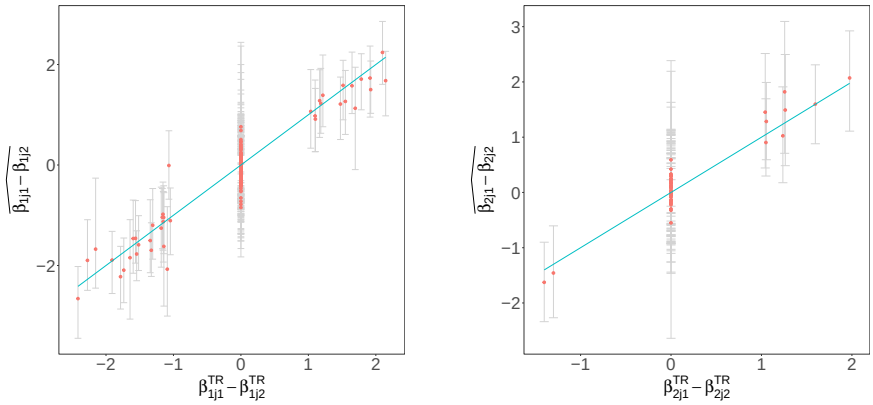
Figure 7: [Simulation 3] The upper right and lower left triangles of a heatmap illustrate estimates $\hat{\rho}_{jj'}^{mm'}$ of correlations and their truth, respectively. The horizontal and vertical lines are to divide the groups. The estimates in panels (a)-(c) are from Sp-BGFM, MOFA and SPIEC-EASI, respectively.

some selected OTUs are illustrated with the normalized observed counts in crosses in Fig 6. The plots do not show any systematic discrepancy and indicate a reasonable model fit.

In addition, correlation estimates are obtained from MOFA and SPIEC-EASI and compared to the truth in Fig 5(b) and (c). Supp. Fig. 4 compares correlation estimates under the additional comparators, REBACCA, COAT and Zi-LN, to the truth. The comparators fail to capture the true dependence structure. Our Sp-BGFM yields superior estimates of $\rho_{jj'}^{mm'}$ and outperforms the other methods in comparison.

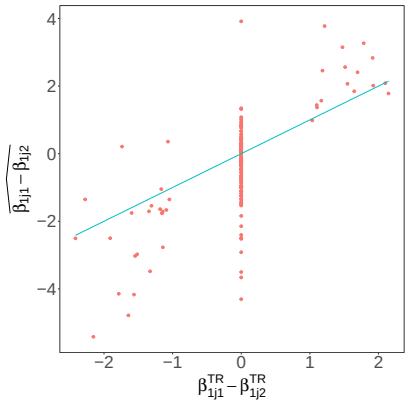
3.3 Simulation 3

For Simulation 3, we built upon the set-up of Simulation 2 and incorporated a categorical covariate with two levels to investigate the estimation of β_{mjp} and Σ in a complex setting. To represent the two levels, we introduced a pair of binary indicators $\mathbf{x}_i = (x_{i1}, x_{i2}) \in \{(1,0),(0,1)\}$. We generated two samples for each of the $S = 20$ subjects, one from each of the levels, resulting in a total of $N = 40$ samples. We set $\beta_{mj1}^{\text{tr}} = 0$ for all (m, j). We let $\beta_{mj2}^{\text{tr}} = 0$ with probability 0.8. For non-zero β_{mj2}^{tr}, we simulated $\beta_{mj2}^{\text{tr}} \sim \text{N}(0, 1/3)$ and shifted away from zero by 1. We simulated r_{im}^{tr}, $\alpha_{s_i mj}^{\text{tr}}$ and Σ^{tr} the same as in Simulation 2. 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

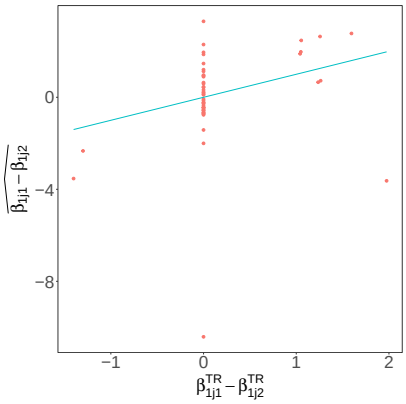


(a) Sp-BGFM: $\beta_{1j1} - \beta_{1j2}$

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



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



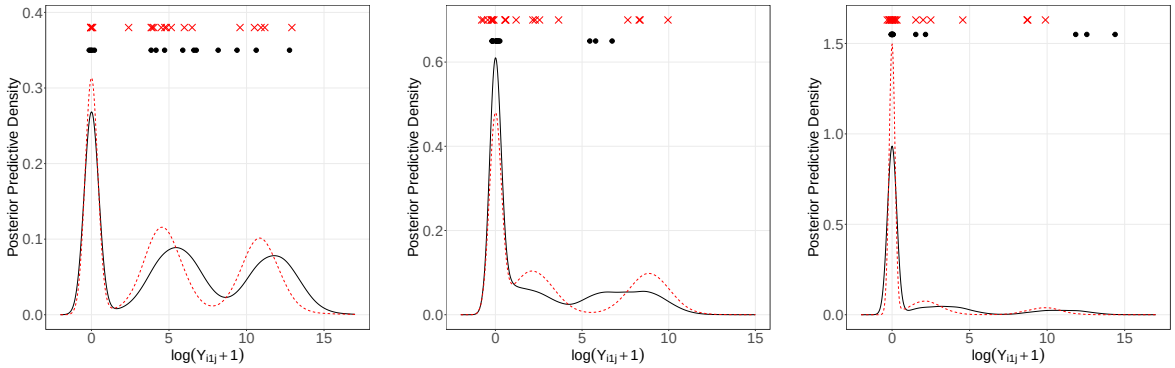
(d) metagenomeSeq: $\beta_{2j1} - \beta_{2j2}$

Figure 8: [Simulation 3] 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.

finally let count vectors $\mathbf{y}_i = \lfloor \mathbf{y}_i^{\star,\text{tr}} \rfloor$, and the overall zero count rate is 45%.

The fixed hyperparameters are specified the same as those in Simulations 1 and 2. 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 7(a) illustrates the posterior estimates $\hat{\rho}_{jj'}^{mm'}$ under Sp-BGFM. Fig 8(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



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

Figure 9: [Simulation 3] 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.

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 are well estimated.

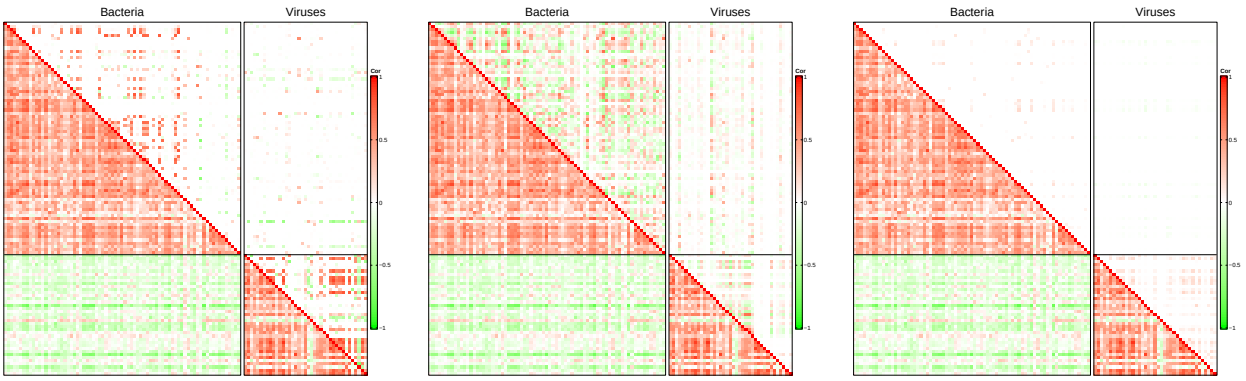
We also check the model fit using posterior predictive checking. We set $r_m^{\text{pred}} = 0$ for $m = 1, 2$ and estimate the distribution of $\mathbf{y}^{\text{pred}}$ for the two conditions, $\mathbf{x} = (1, 0)$ and $(0, 1)$, similar to the procedure used in Simulation 1. The predictive distribution estimates are illustrated in Fig 9 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 the covariate's impact on the prediction of counts for those OTUs.

Fig 7(b) and (c) compare the correlation estimates obtained from MOFA and SPIEC-EASI to the truth. The estimates from the additional comparators, REBACCA, COAT and

Zi-LN, are shown in Supp. Fig. 5. 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. Fig 8(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.

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 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 10(a). The fixed 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.



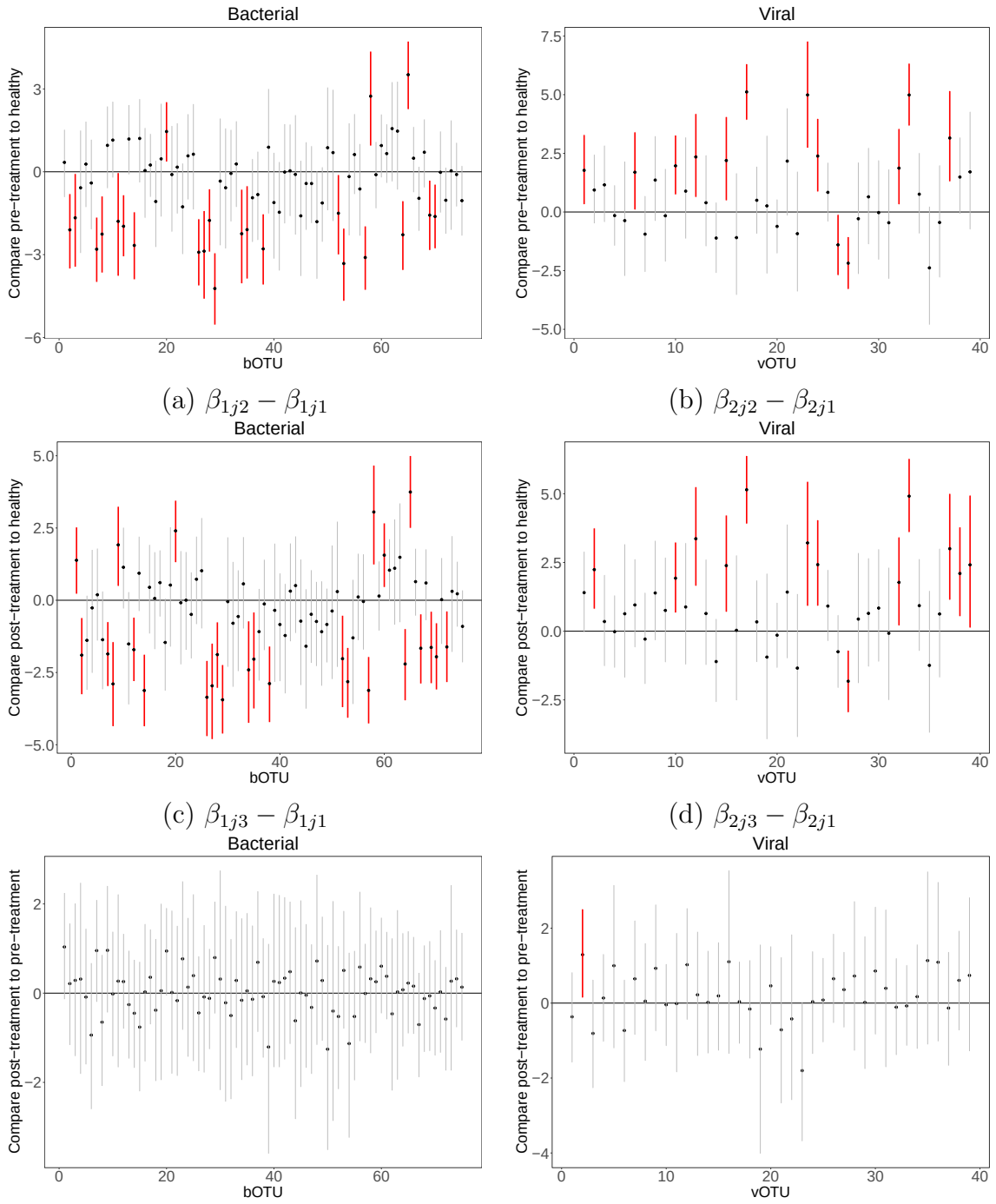
(a) Sp-BGFM (b) MOFA (c) SPIEC-EASI

Figure 10: [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.

The upper right triangle of Fig 10(a) illustrates posterior median estimates $\hat{\rho}_{jj'}^{mm'}$ of correlations. The OTUs are rearranged within a group for a better illustration. Supp. Fig 7 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. 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* are facultative aerobics found in patients with diabetic foot osteomyelitis (Van Asten et al., 2016). Additionally, *Actinomyces* infections are known to be polymicrobial where major concomitant or coinfecting microbes include species of *Campylobacter* and *Phorphylomonas* (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. 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. 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 dysbiosis 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 relationship label (*Pseudomonas* - *Pseudomonas* phage).

Fig 11 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 bold. Supp. Tabs 1 and 2 have taxonomic information of the OTUs



(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 11: [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.

whose interval estimates do not contain zero. Overall, the bOTUs tend to be enriched in the healthy condition compared to the pre- and post-debridement conditions. In contrast,

vOTUs tend to be enriched in the pre- and post-debridement conditions. Changes in abundance between pre- and post-debridement conditions are relatively minimal for bOTUs and vOTUs. Within the wound samples, vOTUs 1, 18 and 23, corresponding to *Acinetobacter* phage, *Proteus* phage and *Staphylococcus* phage, are founded 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.

Supp. Fig. 8 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. The figure 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 10(b) and (c) illustrate $\hat{\rho}_{jj'}^{mm'}$ under the comparators. 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. 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 the model assumptions even further. For instance, one possible extension is to incorporate a hierarchical Dirichlet process (HDP) in Teh et al. (2004) or a common atom model as proposed in Denti et al. (2023). These approaches allow for the construction of domain and OTU specific distributions using a hierarchical structure. In particular, an HDP enables G_{mj} in (2) to share mixture components, while the mixture weights differ across OTUs. Another extension involves using a fully nonparametric regression model to accommodate covariates $\mathbf{x}$ in a more flexible manner. This can be achieved through a dependent Dirichlet process (DDP) model (MacEachern, 1999, 2000) by letting ψ^{α}_{ml} and/or $\xi^{\star}_{mjl}$ 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 $\mathbf{x}$. The DDP model possesses desirable theoretical properties, such as full support (Barrientos et al., 2012). It is important to note that while these extended models offer greater flexibility, they may require a sufficiently large sample size to obtain inference with reasonable uncertainty bounds.

An interesting avenue for future research would be to integrate taxonomy rank information into the analysis. In microbiome studies, the widely used 16S rRNA gene sequencing produces a phylogenetic tree estimate that contains crucial phylogenetic information on the evolutionary relationships among OTUs. Because closely related organisms often share similar characteristics, we may improve the estimation of interactions between OTUs by incorporating a phylogenetic tree into the analysis (Washburne et al., 2018). For example, Chung et al. (2022) incorporated the branch split information by using a latent position model and built a truncated Gaussian copula model. Adapting a similar idea, Sp-BGFM can be extended by introducing taxonomy level specific factor loadings, e.g. Λ^{T}_{m}, where T

is a taxonomy level in the phylogenetic tree. By letting OTUs have latent factor loadings according to their phylogeny, the interaction structure that integrates the phylogenetic relatedness among OTUs can be obtained. This approach has the potential to enhance the inference of interaction structures in other domains, such as the virome, as well as improve the overall understanding of the interaction patterns across multiple domains.

SUPPLEMENTARY MATERIAL

SUPPLEMENTARY: The Supplement contains an examination of the properties of the Dir-HS distribution and a detailed description of the MCMC sampling algorithm. Additionally, it presents additional 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 (Irene A. Chen); and NSF under Grant DMS-2015428 (Juhee Lee).

References

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. The annals of statistics 2(6), 1152–1174.

Argelaguet, R., B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, and O. Stegle (2018). Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. Molecular systems biology 14(6), e8124.

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

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

Barrientos, A. F., A. Jara, and F. A. Quintana (2012). On the Support of MacEachern's Dependent Dirichlet Processes and Extensions. Bayesian Analysis 7(2), 277 – 310.

Bhattacharya, A., D. Pati, N. S. Pillai, and D. B. Dunson (2015). Dirichlet–laplace priors for optimal shrinkage. Journal of the American Statistical Association 110(512), 1479–1490.

Browne, M. W. (1979). The maximum-likelihood solution in inter-battery factor analysis. British Journal of Mathematical and Statistical Psychology 32(1), 75–86.

Cai, T. T., Z. Ren, and H. H. Zhou (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 D. B. Dunson (2011). Bayesian kernel mixtures for counts. Journal of the American Statistical Association 106(496), 1528–1539.

Cao, Y., W. Lin, and H. Li (2019). Large covariance estimation for compositional data via composition-adjusted thresholding. Journal of the American Statistical Association 114(526), 759–772.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2009). Handling sparsity via the horseshoe. In Artificial intelligence and statistics, pp. 73–80. PMLR.

Chung, H. C., I. Gaynanova, and Y. Ni (2022). Phylogenetically informed bayesian truncated copula graphical models for microbial association networks. The Annals of Applied Statistics 16(4), 2437–2457.

Denti, F., F. Camerlenghi, M. Guindani, and A. Mira (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., I. Cavallo, B. Capitanio, F. Ascenzioni, F. Pimpinelli, A. Morrone, and F. Ensoli (2019). Staphylococcus aureus and the cutaneous microbiota biofilms in the pathogenesis of atopic dermatitis. Microorganisms 7(9), 301.

Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3), 432–441.

Haario, H., E. Saksman, and J. Tamminen (2001). An adaptive Metropolis algorithm. Bernoulli 7(2), 223 – 242.

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association 96(453), 161–173.

Klami, A., S. Virtanen, and S. Kaski (2013). Bayesian canonical correlation analysis. Journal of Machine Learning Research 14(30), 965–1003.

Klami, A., S. Virtanen, E. Leppäaho, and S. Kaski (2014). Group factor analysis. IEEE transactions on neural networks and learning systems 26(9), 2136–2147.

Könönen, E. and W. G. Wade (2015). Actinomyces and related organisms in human infections. Clinical microbiology reviews 28(2), 419–442.

Kurtz, Z. D., C. L. Müller, E. R. Miraldi, D. R. Littman, M. J. Blaser, and R. A. Bonneau (2015). Sparse and compositionally robust inference of microbial ecological networks. PLoS computational biology 11(5), e1004226.

Lewandowski, D., D. Kurowicka, and H. Joe (2009). Generating random correlation matrices based on vines and extended onion method. Journal of Multivariate Analysis 100(9), 1989–2001.

MacEachern, S. (2000). Dependent dirichlet processes (tech. rep.). Columbus, OH: Ohio State University.

MacEachern, S. N. (1999). Dependent nonparametric processes. In ASA proceedings of the section on Bayesian statistical science, Volume 1, pp. 50–55. Alexandria, Virginia. Virginia: American Statistical Association; 1999.

Müller, P., F. A. Quintana, A. Jara, and T. Hanson (2015). Bayesian nonparametric data analysis, Volume 1. Springer.

Noiri, Y., K. Ozaki, H. Nakae, T. Matsuo, and S. Ebisu (1997). An immunohistochemical study on the localization of porphyromonas gingivalis, campylobacter rectus and actinomyces viscosus in human periodontal pockets. Journal of periodontal research 32(7), 598–607.

Paulson, J. N., O. C. Stine, H. C. Bravo, and M. Pop (2013). Differential abundance analysis for microbial marker-gene surveys. Nature methods 10(12), 1200–1202.

Peters, B. M., M. A. Jabra-Rizk, G. A. O'May, J. W. Costerton, and M. E. Shirtliff (2012). Polymicrobial interactions: impact on pathogenesis and human disease. Clinical microbiology reviews 25(1), 193–213.

Prost, V., S. Gazut, and T. Brüls (2021). A zero inflated log-normal model for inference of sparse microbial association networks. PLoS Computational Biology 17(6), e1009089.

Sethuraman, J. (1994, 01). A constructive definition of the dirichlet prior. Statistica Sinica 4, 639–650.

Teh, Y., M. Jordan, M. Beal, and D. Blei (2004). Sharing clusters among related groups: Hierarchical dirichlet processes. Advances in neural information processing systems 17.

Tian, C., D. Jiang, A. Hammer, T. Sharpton, and Y. Jiang (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 0(0), 1–15.

Tipton, L., C. L. Müller, Z. D. Kurtz, L. Huang, E. Kleerup, A. Morris, R. Bonneau, and E. Ghedin (2018). Fungi stabilize connectivity in the lung and skin microbial ecosystems. Microbiome 6, 1–14.

Van Asten, S., J. La Fontaine, E. Peters, K. Bhavan, P. Kim, and L. Lavery (2016). The microbiome of diabetic foot osteomyelitis. European Journal of Clinical Microbiology & Infectious Diseases 35, 293–298.

Verbanic, S., J. M. Deacon, and I. A. Chen (2022). The chronic wound phageome: Phage diversity and associations with wounds and healing outcomes. Microbiology Spectrum 10(3), e02777–21.

Verbanic, S., Y. Shen, J. Lee, J. M. Deacon, and I. A. Chen (2020). Microbial predictors of healing and short-term effect of debridement on the microbiome of chronic wounds. NPJ biofilms and microbiomes 6(1), 1–11.

Virtanen, S., A. Klami, S. Khan, and S. Kaski (2012, 21–23 Apr). Bayesian group factor analysis. In N. D. Lawrence and M. Girolami (Eds.), Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, Volume 22 of Proceedings of Machine Learning Research, La Palma, Canary Islands, pp. 1269–1277. PMLR.

Washburne, A. D., J. T. Morton, J. Sanders, D. McDonald, Q. Zhu, A. M. Oliverio, and R. Knight (2018). Methods for phylogenetic analysis of microbiome data. Nature microbiology 3(6), 652–661.

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

Zhang, S., Y. Shen, I. A. Chen, and J. Lee (2023+). Bayesian modeling of interaction between features in sparse multivariate count data with application to microbiome study. The Annals of Applied Statistics.

Zhao, S., C. Gao, S. Mukherjee, and B. E. Engelhardt (2016). Bayesian group factor analysis with structured sparsity. Journal of Machine Learning Research 17(196), 1–47.