UNIVERSITY OF CALIFORNIA
SANTA CRUZ

FLEXIBLE BAYESIAN MODELING OF MULTIVARIATE COUNT DATA

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICAL SCIENCE

by

Shuangjie Zhang

September 2024

The Dissertation of Shuangjie Zhang
is approved:

Juhee Lee, Chair

Bruno Sansó

Zehang Richard Li

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Flexible Bayesian Modeling of Multivariate Count Data

by

Shuangjie Zhang

The analysis of multivariate count data presents significant statistical challenges due to its discrete nature, excess zeroes, over-dispersion, and high dimensionality, which are often encountered in practical applications. These challenges are further complicated by the presence of covariates. Traditional methods frequently struggle with these complexities, potentially leading to inferior performance in estimating feature abundance and their dependencies. This thesis develops flexible Bayesian statistical methodologies, particularly for cases where the distribution of a multivariate random vector exhibits non-Gaussianity, heterogeneity, and heteroscedasticity, using count table data from microbiome studies as motivating examples. First, we propose a Bayesian zero-inflated rounded log-normal kernel method that infers feature interdependencies through the covariance between features measured in counts. We employ a factor model that assumes a lower-dimensional structure for the covariance matrix, and impose joint sparsity on its factor loadings using a Dirichlet-Laplace (Dir-Laplace) prior. This sparse spiked covariance structure reduces the number of parameters and robustifies the estimation in high-dimensional settings. A regression model is used to characterize changes in mean feature abundance with covariates, and a Bayesian nonparametric approach is adopted to handle large variability across samples. For problems involving multiple count ta-

bles obtained from different groups, we extend the sparse factor model and develop a Bayesian group factor model that infers within-group and across-group feature interdependencies. We incorporate a flexible infinite mixture of log-normal rounded kernels through the Dirichlet process prior directly for count vectors and construct a Dirichlet-Horseshoe (Dir-HS) shrinkage prior for factor loadings to more efficiently induce joint sparsity for the greater number of features in a multiple group setting. Lastly, we develop a covariate-dependent factor model that flexibly estimates heteroscedasticity in the covariance matrix due to covariates, addressing the problem of the mean and covariance structure of a multivariate count vector varying with covariates. Our approach employs covariance regression through linear regression on the lower-dimensional factor loading matrix. This formulation, combined with joint sparsity imposed by the Dir-HS prior, provides robust estimation of covariate-dependent covariance in high-dimensional settings. For all developed models, we carefully explore their properties and perform extensive simulation studies to examine their performance. In addition, real data examples from microbiome studies are used for illustration.

Acknowledgments

I would like first to thank my advisor, Juhee Lee, for her mentorship. Juhee is a truly helpful and responsible mentor who advocates me during my Ph.D research and future career. She dedicates ample time to discussions when I need it most and patiently explains. I am also grateful to my family, who gave me accompany during this journey. I would also like to thank: Bruno Sansó and Zehang Richard Li for serving on my dissertation reading committee, and offering invaluable suggestions that significantly enhanced the quality of my dissertation; Raquel Prado for serving as my Ph.D advancement committee chair and providing excellent suggestions; Irene A. Chen and Yuning Shen for the collaboration on projects involving microbiome data in Chapters 2 and 3; Michael Patnode for the collaboration of mice gut microbiome data in Chapter 4.

The text of Chapter 2 includes an adapted reprint of the following previously published article:

Zhang, S., Shen, Y., Chen, I. A., & 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.

A manuscript based on the work from Chapter 3 is currently under its second revision for journal publication. The co-authors listed in this publication supervised the research that forms the basis for the chapter. I also acknowledge the editors of the journals and anonymous referees for helpful suggestions.

Chapter 1

Introduction

In the realm of statistical analysis, multivariate count data presents unique challenges and opportunities. This data type, characterized by observations that are discrete and typically represent counts of occurrences, is pervasive in various fields such as genomics (Schloissnig et al., 2013), epidemiology (Papoz et al., 1996), social sciences (Böhning et al., 1997), and marketing (Ravishanker et al., 2016). The complexity increases exponentially when dealing with high-dimensional data, where each observation consists of counts across numerous variables. Traditional statistical methods often falter under these complexities due to the discrete nature and the high dimensionality of the data. Over a long history of tackling the discreteness obstacle, adding a small pseudo-count plus a transformation is a common strategy, such as square root transformation (Bartlett, 1936) and log transformation (Sokal and Rohlf, 1995) among many others. The primary purpose of the transformation is to let transformed data meet the assumptions required for statistical methods for continuous data, such as linear regression,

ANOVA, and t-tests. The generalized linear model(GLM) (Myers and Montgomery, 1997) is another popular approach which utilizes Poisson distribution or negative binomial distribution for the count data. GLM generalizes linear regression by allowing the linear model to be related to counts via a link function. However, multivariate Poisson distribution or multivariate negative binomial distribution still requires more foundation for practical use. Alternatively, copula models can be used as a statistical tool for modeling multivariate count data(e.g. Safari-Katesari et al. (2020)). A copula allows for modeling of dependencies between random variables, regardless of their marginal distributions. This is particularly useful in multivariate settings where traditional models struggle with capturing complex dependencies. Some common choices are modeling each count variable with an appropriate marginal distribution (e.g., Poisson(Cook et al., 2010), Negative Binomial(Shi and Valdez, 2014), Zero-Inflated models(Alqawba and Diawara, 2021)) while using the copula to flexibly capture the complex dependencies between them.

High-dimensional count data often exhibit intricate patterns of covariance that need to be accurately captured to make meaningful inferences. Ignoring these dependencies can lead to biased estimates of mean structure and suboptimal model fit. The dependence structure itself is also of interest for inference. Additionally in the high-dimensional setting, the number of features exceeds the number of samples, further making traditional estimates of covariance biased and unstable. Dimensionality reduction techniques offer a solution by transforming the high-dimensional data into a lower-dimensional space, and regularization approaches have been proposed for large co-

variance estimation. Wu and Pourahmadi (2003); Bickel and Levina (2008b) construct covariance estimators via banding the sample covariance matrix directly and banding the Cholesky factor matrix of the precision matrix. Bickel and Levina (2008a); Rothman et al. (2009) combine thresholding with shrinkage and study generalized thresholding of the sample covariance matrix in high dimensions. Among those, factor analysis is a commonly used method in Bayesian inference to access multivariate dependence (Rummel, 1988). The factor model identifies a smaller number of unobservable latent variables, known as factors, that account for the covariance among the observed variables. The basic premise of factor analysis is that many variables are influenced by common underlying latent factors, which can significantly simplify the data structure without losing essential information. Bernardo et al. (2003) propose the Bayesian factor model to reduce dimensionality in two ways: the number of latent factors smaller than dimension and the factor loadings matrix having a lot of zeros. This structure is well motivated in many biomedical applications. To further induce sparsity on the factor loading matrix, a variable selection-type mixture prior has been designed for loadings (Lucas et al., 2006; Carvalho et al., 2008). Heavy-tailed default prior (Ghosh and Dunson, 2009), multiplicative gamma process shrinkage prior (Bhattacharya and Dunson, 2011) and Dirichlet-Laplace prior (Bhattacharya et al., 2015) are also developed for efficient shrinkage and robust estimate of large covariance matrix. Recently, extended methodologies of factor model such as group factor analysis (Klami et al., 2014; Virtanen et al., 2012a), multi-study factor analysis (De Vito et al., 2019) and perturbed factor analysis (Roy et al., 2021) are built for different analyze goals. Although Schiavon et al.

(2022) proposes generalized factor models for binary data, it still calls for statistical methods modeling high-dimensional multivariate count tables with added complexity.

In addition, factor models have been used to address the problem of heteroscedasticity. The assumption of homoscedasticity, constant variance across all levels of the covariates, is fundamental for the validity of many statistical techniques, including all methods mentioned above. However, in real-world applications, this assumption is often violated when dealing with the phenomenon where the variance-covariance matrix varies across different combinations of covariates. It presents a new significant challenge to traditional statistical methods. The implications of heteroscedasticity in univariate settings have been extensively studied and are well-documented. The development of a linear or generalized linear model with a link function on the variance can be found in Carroll and Ruppert (1982); Rutemiller and Bowers (1968); Smyth (1989). Multivariate heteroscedasticity, especially in high-dimensional count data, further adds up to the complexity of estimation with an exponentially increasing number of parameters. Direct modeling of each element of the covariance matrix, such as in a log scale to ensure non-negativity (Chiu et al., 1996; Pourahmadi, 2011; Battey, 2017), is hard to extend in high-dimensional setting. More recently, researchers incorporate covariates into the dimensionality reduction technique, allowing the lower-dimensional structure varying with covariates. Pourahmadi (1999) and Hoff and Niu (2012) relate covariates to the Cholesky decomposition and factor analysis, respectively. Fox and Dunson (2015) further build a flexible Bayesian nonparametric covariance regression model by putting a Gaussian process prior on the factor loading matrix. There is very few literature

addressing the multivariate heteroscedasticity in high-dimensional count data; careful and thoughtful designs are still needed.

One of the motivating applications of multivariate count data in this thesis is microbiome data which comprises the collective genomes of microorganisms in experimental subjects (Marchesi and Ravel, 2015). Contemporary studies have demonstrated that microbes, such as bacteria, viruses and fungi, play an important role in the process of disease infection and illness recovery (Lloyd-Price et al., 2019; Verbanic et al., 2020). Through high-throughput sequencing (HTS) sequencing or shotgun metagenomic sequencing technologies, it generates multivariate Operational Taxonomic Units (OTUs) for downstream analysis. OTU tables are multivariate count tables, where each variable represents the abundance of an OTU in a sample. Analyzing OTU tables, such as inferring interaction between microbes, helps to understand the mechanism of microbial ecology and interactions between microbes. Besides the aforementioned statistical challenges in modeling multivariate count vectors, there are additional complexities due to compositionality; the raw counts do not reflect absolute abundance but rather are relative abundance compared to the other counts, due to the experimental artifacts such as the sequencing depth. To make counts more comparable across samples, a normalization is required. For example, SparCC (Friedman and Alm, 2012) normalizes raw counts by adding pseudo counts and then dividing by the sample's total counts. It models log-transformed ratios of these normalized counts to infer correlations between OTUs through sparse networks. Similarly, CCLasso in Fang et al. (2015) uses ℓ_1 penalty to estimate the correlation network of log-transformed counts. Kurtz et al. (2015) devel-

ops SPIEC-EASI first applying the centered log-ratio (clr) transformation to raw OTU counts. It then uses graphical lasso (Friedman et al., 2008), a popular penalized method outputting the association of undirected graphs, to obtain a robust precision matrix estimate. See REBECCA (Ban et al., 2015), COAT (Cao et al., 2019), MOFA (Argelaguet et al., 2018) and ZI-MLN (Zhang et al., 2023a) for more. Copula-based methods such as SparseDOSSA(Ma et al., 2021) fit a penalized multivariate Gaussian copula model with a zero-inflated log-normal distribution on the absolute count abundances. Deek and Li (2023) proposes using copula models with a zero-inflated beta marginal to estimate covariance between taxa using normalized microbial relative abundance data. Although several methods exist for inferring microorganism interactions in microbiome studies, there is still a need for comprehensive approaches to address all aforementioned challenges.

The contribution of this work is the development of Bayesian modeling techniques for multivariate count data, focusing on methods that effectively manage discrete high dimensionality and intricate covariance structures. Bayesian methods offer a robust framework for incorporating prior knowledge and uncertainty, making them particularly suitable for complex data structures (Gelman et al., 1995). In the context of multivariate count data, Bayesian models can effectively handle the intricacies of discrete distributions and allow for the explicit modeling of covariance structures through hierarchical models and latent variable approaches. The proposed methods can uncover underlying relationships between variables, providing insights that might be missed by simple models. In addition, they carefully address other complexities commonly arising

from count table data analysis; excess zeros, over-dispersion and large variability across samples, using flexible Bayesian nonparametric methods.

We begin first by constructing a Bayesian zero-inflated log-normal rounded kernel model in Chapter 2. The rounded kernel model (Canale and Dunson, 2011) introduces latent multivariate log-normal variables to model the interaction between counts. We put a sparse factor model with Dirichlet Laplace prior (Bhattacharya et al., 2015) on the factor loading matrix and induce sparsity on the covariance matrix of the multivariate log-normal kernel. The model also performs model-based normalization estimation and estimates the differential abundance of count features associated with covariates through a log-linear regression. The zero-inflation proportion and heavy-tailed log-normal distribution account for zeros and over-dispersion in count data. Simulation studies show the proposed model identifies count abundance differences and yields covariance estimates with favorable accuracy compared with the alternatives. The proposed model is applied to analyze two real datasets: skin microbiome data and human gut microbiome data.

Chapter 3 describes a sparse Bayesian group factor model for the analysis of multiple multivariate count data to obtain desired inferences among multiple sources of count tables. This approach uses Bayesian nonparametric mixtures of rounded multivariate log-normal kernels to obtain a flexible joint distribution of count vectors. Another primary novelty of this method is constructing a new Dirichlet-Horseshoe (Dir-HS) shrinkage prior on the joint sparsity of factor loadings. We carefully study the property of the new prior and compare it to existing priors. The nonparametric approach flexibly

addresses excess zeros and heterogeneity issues. Extensive numerical studies indicate the model's superior recovery of the underlying data-generating process of multiple count tables. We apply the model to analyze multi-domain skin microbiome data. The model outputs valuable abundance estimates for different types of microbes and reveals the associations among domains.

We build a Bayesian covariate-dependent rounded kernel model in Chapter 4 to provide insight into covariate effects on interactions in a count vector. The model follows the covariance regression strategy with a graceful multiplicative effect on the factor loading matrix, allowing the covariance to vary with general covariates. The parametric construction gains computational efficiency in estimating the high-dimensional covariance matrix of each sample, significantly reducing the number of parameters to estimate. A flexible Dirichlet process mixture (DPM) model is used for the count distribution, helping to address the aforementioned challenges of count data. A regression formulation is used on the mean abundance to detect covariate effects on the count abundance. Thus, this method simultaneously explores covariate effects on the mean and the covariance of the count. We use our model to analyze a mice gut microbiome dataset.

Finally, chapter 5 summarizes the main contributions of this thesis, and concludes with some possible future extensions.

Chapter 2

Bayesian Modeling of Interaction between Features in Sparse Multivariate Count Data with Application to Microbiome Study

2.1 Introduction

High-throughput sequencing (HTS) technologies in microbial ecology generate multivariate count data to characterize and analyze microbial communities from a variety of habitats such as human body sites, soil and water. Widely used sequencing methods in microbiome research include 16S ribosomal RNA (rRNA) sequencing and shotgun metagenomic sequencing (Jovel et al., 2016). 16S rRNA gene sequencing utilizes PCR to target and amplify some portions of the bacterial 16S rRNA subunit

gene for sequencing. The sequence reads are then clustered based on their similarity into operational taxonomic units (OTUs), which represent bacteria types. Following some initial preprocessing procedures, 16S rRNA sequencing data is summarized into a large count matrix (referred to as an OTU table) for downstream analyses, where the columns represent samples, and the rows contain multivariate count vectors of sequences corresponding to OTUs in the samples. Different from marker gene-based community profiling, shotgun metagenomic sequencing sequences a sample's entire metagenome and offers finer resolution at a higher cost. After some bioinformatic preprocessing, it also produces multivariate count table data that has structure and properties similar to those of an OTU table for downstream analyses. 16S rRNA sequencing datasets are used for illustrations of the statistical method developed in this work, but it can be considered for analysis of the data generated by either sequencing technique. We note that their analysis units are different, and the resulting statistical inferences may have different biological interpretations. In the human gut microbiome data, one of our real data examples in § 2.4.2, 16S rRNA sequencing data was collected to study how the composition of the gut microbiome is associated with inflammatory bowel disease (IBD) such as Crohn's disease (CD) or ulcerative colitis (UC) (Lloyd-Price et al., 2019). Understanding how the composition of the human gut microbiome is associated with covariates such as disease status and age is important to provide insights on its role in human health and disease. Also, detecting and investigating the structure of microbial interactions is critical to better characterize microbial communities. Accurately accounting for the interactions can further improve the quantification of covariate effects

on microbial abundances.

HTS sequencing data in microbiome study presents various challenges for statistical analysis due to high dimensionality and some added complexity. Total OTU counts vary in samples due to experimental artifacts such as the sequencing depth, and raw counts do not reflect actual microbial abundances (called compositionality). Consequently, normalization of OTU counts is needed for meaningful comparison across samples. In addition, the high-dimensional structure with excess zeros and over-dispersion further complicates the analysis of an OTU table and calls for flexible statistical models. While various statistical models have been proposed for microbiome data analysis, most existing methods focus on either inference on the effects of environmental factors (i.e., covariate) on microbial abundances or their absence/presence or inference on associations between microbes. For studying associations with covariates, generalized regression models are popular. For example, Poisson or negative binomial (NB) regression models are one of the common approaches, where covariates are related to expected counts through a log-linear regression framework. Those models include sample size factors for normalization. Zero-inflated (ZI) Poisson or ZI-NB models are also utilized to address excess zeros. Under a ZI model, a count is distributed as a mixture, a component of which is the distribution with a point mass of one at zero. See Li et al. (2017), Zhang et al. (2017), Jiang et al. (2021), Shuler et al. (2021a) among many others, for examples of using Poisson or NB regression models. Another common regression approach uses multinomial or ZI multinomial models, where a similar log-linear regression framework is used to relate covariates to (unconstrained) occurrence

probability vectors, e.g., Xia et al. (2013), Wadsworth et al. (2017), Ren et al. (2017), Tang and Chen (2019) and Grantham et al. (2020) among many others. In particular, Grantham et al. (2020) proposed a Bayesian multinomial regression model that assumes a mixed effects model for unconstrained occurrence probabilities and uses a latent factor model for the covariance matrix of the prior distribution of the unconstrained probabilities. However, the implication of the covariance among unconstrained probabilities for microbial interactions is not clear due to the fixed total count constraint under the assumed multinomial distribution. Approaches of using a Dirichlet-tree multinomial model were also proposed to exploit the tree structure information via a phylogenetic tree, e.g., Wang and Zhao (2017), Mao et al. (2020) and Wang et al. (2021). They assume potential associations between microbes that have similar sequences but do not attempt to infer microbial interactions. Alternatively, Paulson et al. (2013) assumed a univariate log-normal distribution for individual counts after adding a pseudo count to observed counts and used regression to relate covariates to OTU abundances. For inferences on microbial interactions, correlations between pairs of microbes based on some transformed OTU counts are commonly used as a measure. The task of estimating correlations between microbes is complicated due to the aforementioned challenges. Centered-log-ratio (clr) transformation is usually applied to raw counts prior to analysis for compositionality, and small pseudocounts are added to avoid numerical issues of excess zeros. To address high dimensionality, an additional structure such as sparsity through ℓ_1 penalty is often imposed on the covariance matrix or precision matrix for reliable inference. For example, SparCC in Friedman and Alm (2012) normalizes raw

counts by sample total counts after adding pseudo counts and models log-transformed ratios of the normalized counts to infer correlations between OTUs. CCLasso in Fang et al. (2015) models log-transformed counts and provides a least squares estimate of a correlation matrix with ℓ_1 penalty under some constraint for compositionality of microbiome data. SPIEC-EASI in Kurtz et al. (2015) builds an undirected graphical model for clr transformed data and yields inference on an association network between OTUs through a precision matrix. Sparsity is assumed for the underlying association network. Schwager et al. (2017) uses a Bayesian log-normal graphical model for unconstrained counts. A LASSO prior is used for the precision matrix. Similarly, Prost et al. (2021) developed a likelihood-based zero-inflated log-normal graphical model (Zi-LN) that appropriately accounts for excess zeros in microbiome data. Graphical LASSO (Friedman et al., 2008) is used for estimation of the precision matrix. While existing methods can provide useful insights on microbial communities, methods that jointly infer associations between microbes and their associations with covariates are still lacking. Furthermore, statistical methods that carefully address excess zeros, compositionality and high dimensionality are needed for accurate inference on the associations.

To obtain a better understanding of the underlying biological processes, we develop a Bayesian rounded kernel regression model with zero inflation. The model enables a direct assessment of interrelationships between OTUs and their associations with covariates. The developed method directly models raw counts and simultaneously performs model-based normalization through random sample scale factors for compositionality. Specifically, we use a multivariate log-normal distribution as the kernel and

define multivariate count responses $\mathbf{Y} = (Y_1, \ldots, Y_J)$ of J OTUs in terms of multivariate log-normal latent variables $\mathbf{Y}^\star = (Y_1^\star, \ldots, Y_J^\star)$ using fixed thresholds. We then relate covariates $\mathbf{x}$ to the mean vector $\boldsymbol{\mu}$ of the distribution of $\mathbf{Y}^\star$ through regression and use the covariance matrix Σ to characterize interrelationship among OTUs. $\boldsymbol{\mu}$ also includes sample size factors for normalization. For Σ, we assume joint sparsity to reliably learn a high dimensional covariance structure with a small sample size. Sparsity assumption is commonly used in the covariance matrix estimation when $p \gg n$ (e.g., Cai et al. (2016), Pati et al. (2014), Gao and Zhou (2015), Xie et al. (2018)). Specifically, we develop a joint sparse latent factor model for Σ, where we let the number of factors much smaller than the number of OTUs (features), and a majority of OTUs can have factor loadings close to zero, i.e., feature selection. The model greatly reduces the number of parameters to estimate and provides a simple interpretation of the interrelationship structure. The representation of the model with independent latent factors also allows introducing zero inflation in a convenient manner. The model appropriately accounts for excess zeros due to the absence of an OTU or the undersampling of a rare OTU, and Σ provides inferences on the interrelationship structure among OTUs present in a sample. In addition, overdispersion is accommodated through random effects, resulting in further improvement in the inference.

In the remainder of this chapter, we describe the model and its applications. § 2.2 describes the zero-inflated multivariate log-normal kernel model (called "ZI-MLN"), and § 2.3 has results of simulation studies to evaluate the performance of our method. § 2.4 has results from the model applied to two real datasets, and § 2.5

concludes with some discussion of the results and areas of future research.

2.2 Statistical Model

2.2.1 Sampling Distribution and Prior Specification

Consider multivariate count data obtained for J OTUs in a microbiome study. We let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})$ denote a J-dimensional random count vector of OTU counts of sample $i = 1, \ldots, N$ taken from subject $g_i \in \{1, \ldots, M\}$, where $Y_{ij} \in \mathbb{N}^0$ is the count of OTU $j = 1, \ldots, J$ in sample i. We let n_m be the number of samples taken from subject m and have $N = \sum_{m=1}^{M} n_m$. In addition, data may include a set of P covariates, $\mathbf{x}_i = (x_{i1}, \ldots, x_{iP})$. Our skin microbiome dataset in § 2.4.1 consists of observed counts of 187 OTUs in 20 samples, one sample from each of 20 subjects. The dataset does not have covariates besides the subject factor. Human gut microbiome data in § 2.4.2 includes 67 samples collected from multiple biopsy sites of 37 patients. 107 OTUs are included with covariates such as disease phenotype and age for analysis. The model simultaneously infers the interaction structure of OTUs and the differential abundance of OTUs by covariates. It can also be easily simplified if no covariate is available, as we will show later.

We consider a Bayesian rounded multivariate log-normal kernel model for $\mathbf{Y}_i$ in Canale and Dunson (2011). We first introduce continuous latent variables $\mathbf{Y}_i^\star =$

$(Y_{i1}^\star, \ldots, Y_{iJ}^\star)$ with $Y_{ij}^\star \in \mathbb{R}^+$, $i = 1, \ldots, n$ and $j = 1, \ldots, J$, and assume

$$\mathbf{Y}_i^\star \mid \boldsymbol{\mu}_i, \Sigma \overset{indep}{\sim} \text{log-N}_J(\boldsymbol{\mu}_i, \Sigma), \tag{2.1}$$

where parameters $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iJ})' \in \mathbb{R}^J$ and $\Sigma > 0$. In (2.1), we have the mean $\text{E}(Y_{ij}^\star \mid \boldsymbol{\mu}_i, \Sigma) = \exp(\mu_{ij} + \frac{1}{2}\Sigma_{jj})$, the median $Q_{0.5} = \exp(\mu_{ij})$ and covariance $\text{Cov}(Y_{ij}^\star, Y_{ij'}^\star) = \exp\{\mu_{ij}+\mu_{ij'}+\frac{1}{2}(\Sigma_{jj}+\Sigma_{j'j'})\}\{\exp(\Sigma_{jj'})-1\} = \text{E}(Y_{ij}^\star)\text{E}(Y_{ij'}^\star)\{\exp(\Sigma_{jj'})-1\}$. We next use a threshold function to relate $Y_{ij}^\star$ to Y_{ij} by letting $Y_{ij} = y_j$ if $y_j \leq Y_{ij}^\star < (y_j + 1)$. The multivariate log-normal density is zero for a vector with negative values, and the kernel defines a valid multivariate distribution for $\mathbf{Y}$. We further let $\tilde{\mathbf{Y}}_i^\star = (\tilde{Y}_{i1}^\star, \ldots, \tilde{Y}_{iJ}^\star)$ with $\tilde{Y}_{ij}^\star = \log(Y_{ij}^\star) \in \mathbb{R}$ and have

$$\begin{aligned} \text{P}(\mathbf{Y}_i = \mathbf{y}_i \mid \boldsymbol{\mu}_i, \Sigma) = & \int_{A(\mathbf{y}_i)} f_{\mathbf{y}^\star}(\mathbf{y}^\star \mid \boldsymbol{\mu}_i, \Sigma) d\mathbf{y}^\star \\ = & \int_{\tilde{A}(\mathbf{y}_i)} \phi_J(\tilde{\mathbf{y}}^\star \mid \boldsymbol{\mu}_i, \Sigma) d\tilde{\mathbf{y}}^\star, \end{aligned} \tag{2.2}$$

where $f_{\mathbf{y}^\star}$ represents the density function of the J-dimensional log-normal distribution with parameters $\boldsymbol{\mu}_i$ and Σ, and ϕ_J the density function of a J-dimensional normal distribution. The regions of integration are $A(\mathbf{y}_i) = \{\mathbf{y}^\star \mid y_{i1} \leq y_1^\star < y_{i1} + 1, \ldots, y_{iJ} \leq y_J^\star < y_{iJ}+1\}$ and $\tilde{A}(\mathbf{y}_i) = \{\tilde{\mathbf{y}}^\star \mid \log(y_{i1}) \leq \tilde{y}_1^\star < \log(y_{i1}+1), \ldots, \log(y_{iJ}) \leq \tilde{y}_J^\star < \log(y_{iJ}+1)\}$. The properties of the distribution of Y_{ij}'s such as their means and covariances can be easily computed from (2.2). For example, we find $\text{E}(Y_{ij} \mid \mu_{ij}, \Sigma_{jj}) = \sum_{b=0}^{\infty} b\text{P}(Y_{ij} = b \mid \mu_{ij}, \Sigma_{jj})$ with $\text{P}(Y_{ij} = b \mid \mu_{ij}, \Sigma_{jj}) = \Phi_1(\log(b+1) \mid \mu_{ij}, \Sigma_{jj}) - \Phi_1(\log(b) \mid \mu_{ij}, \Sigma_{jj})$, where $\Phi_d(\cdot \mid a, \boldsymbol{B})$ is the cdf of the $d-$ variate normal distribution with mean a and (co)variance

$\boldsymbol{B}$. A large value of μ_{ij} thus implies high abundance of OTU j in sample i. We express μ_i as a function of covariates, sample-size factor and OTU-size factor. The factors account for differences in sample total counts and variability in baseline OTU abundances. We will give a regression model for $\boldsymbol{\mu}_i$ below. We can also compute variances and covariances of the counts. In particular, $\mathrm{Cov}(Y_{ij}, Y_{ij'} \mid \boldsymbol{\mu}_i, \Sigma) = \sum_{b=0}^{\infty}\sum_{b'=0}^{\infty} bb'\mathrm{P}(Y_{ij}=b, Y_{ij'}=b' \mid \boldsymbol{\mu}_i, \Sigma) - \mathrm{E}(Y_{ij} \mid \mu_{ij}, \Sigma_{jj})\mathrm{E}(Y_{ij'} \mid \mu_{ij'}, \Sigma_{j'j'})$. $\mathrm{P}(Y_{ij}=b, Y_{ij'}=b' \mid \boldsymbol{\mu}_i, \Sigma)$ can be computed with a bivariate normal distribution in a way similar to $\mathrm{P}(Y_{ij}=b \mid \mu_{ij}, \Sigma_{jj})$. Under (2.2), the counts of OTUs j and j' are dependent if $\Sigma_{jj'} \neq 0$. That is, Σ characterizes microbial interactions with a straightforward interpretation. In addition, overdispersion is known to be common in sequencing data and can be properly accommodated through heavy tails of a log-normal distribution.

We next build a prior distribution for Σ. The number of OTUs J is often much greater than the sample size N in microbiome studies, i.e., $J \gg N$. In a high-dimensional setting, the sample covariance matrix is singular and provides an unstable estimate for Σ. To overcome the difficulty, it is common that structural assumptions are imposed on Σ (Cai et al., 2016). For example, Friedman et al. (2008), Bien and Tibshirani (2011) and Cai et al. (2011) consider the sparsity assumption that most of the elements in Σ (or Σ^{-1}) are zero or negligible for marginal independencies between features (or conditional independencies). In particular, ℓ_1 penalty is used to shrink the elements of Σ (or Σ^{-1}) to zero. Alternatively, a low-rank structure is considered, sometimes jointly with the sparsity assumption (called joint sparsity). For example, see Cai et al. (2015); Bhattacharya et al. (2015) and Xie et al. (2018). The joint sparsity

structure allows to achieve good theoretical properties, such as faster minimax rate of convergence and tighter posterior contraction rate for estimating a covariance matrix (Cai et al., 2015; Xie et al., 2018). Taking the latter approach, we first decompose Σ as

$$\Sigma = \Lambda\Lambda' + \sigma^2 \mathrm{I}_J, \tag{2.3}$$

where $\boldsymbol{\lambda}_j = [\lambda_{j1}, \ldots, \lambda_{jk}]'$ and $\Lambda = [\boldsymbol{\lambda}_1', \ldots, \boldsymbol{\lambda}_J']'$ is a $J \times K$ factor loading matrix with $K \ll J$. The model assumes most of the covariance structure between OTUs is explained by a small number of factors to obtain a more accurate and reliable estimate of Σ in the case of $N \ll J$. We assume an isotropic noise and consider a conditionally conjugate prior distribution $\sigma^2 \sim \text{inv-Ga}(a_\sigma, b_\sigma)$ with fixed a_σ and b_σ for easy computation. If needed, independent idiosyncratic noise can be considered by letting $\Sigma = \Lambda\Lambda' + \text{diag}(\sigma_j^2)$ and $\sigma_j^2 \overset{iid}{\sim} \text{inv-Ga}(a_\sigma, b_\sigma)$. We introduce joint sparsity on Σ by considering a Dirichlet-Laplace prior in Bhattacharya et al. (2015),

$$\begin{aligned} \tau_k \mid a_\tau, b_\tau &\overset{iid}{\sim} \text{Ga}(a_\tau, b_\tau), \\ \boldsymbol{\phi} = (\phi_1, \ldots, \phi_J) \mid a_\phi &\sim \text{Dir}(a_\phi, \ldots, a_\phi), \\ \lambda_{jk} \mid \phi_j, \tau_k &\overset{indep}{\sim} \text{DE}(\phi_j \tau_k), \end{aligned} \tag{2.4}$$

where $\text{DE}(a)$ represents the double-exponential (Laplace) distribution with scale parameter a, and $\text{Ga}(a, b)$ is the gamma distribution with shape parameter a and scale parameter b (so mean a/b). Under the model in (2.4), a small value of ϕ_j shrinks λ_{jk}

toward zero for all k, and $\Sigma_{jj'}$ tends to have small values for all j'. That is, ϕ_j induces joint sparsity for Σ together with K. OTUs with a small value of ϕ_j may be those less interacting with other OTUs. The model provides an easy interpretation of the interrelationships between OTUs and reliable inference even for cases with $N \ll J$. The double-exponential distribution for λ_{jk} has heavier tails and a more pointed center than the normal distribution that is a convenient choice, and facilitates sparsity in λ_{jk}, resulting in sparsity in Σ. Theorem 3.1 of Bhattacharya et al. (2015) proves that when a_ϕ is set to be $J^{-(1+b)}$ for any $b > 0$, the posterior contraction rate of λ_{jk} achieves the minimax rate. However, our simulation studies show that the model with $a_\phi = 1/J$ tends to overshrink λ_{jk} even when only a small number of OTUs interact, and we fix $a_\phi = 1/2$ with soften conditions for the contraction rate. We fix the factor dimension K at a reasonably large value for computational convenience. If desired, an exponentially decaying prior such as a Poisson distribution can be placed on K to attain optimal posterior contraction rate (Pati et al., 2014). Pati et al. (2014) used the Dirichlet-Laplace prior for vectorized loadings vec(Λ) in a Bayesian factor model for a multivariate normal outcome vector with mean zero and did not attempt to induce a joint sparsity structure. Xie et al. (2018) used a spike-and-slab prior for ϕ_j and developed a matrix spike-and-slab LASSO prior under the Gaussian sampling distribution assumption. However, placing spike-and-slab priors for individual matrix elements may cause computational difficulties, especially for large J. Similar to Bhattacharya and Dunson (2011) and Xie et al. (2018), we do not place any constraints on Λ such as orthogonality of the columns nor attempt to interpret latent factors since the primary interest of inference is on Σ.

We re-write the model in (2.1) and (2.3) by introducing a latent normal vector $\boldsymbol{\eta}_i \overset{iid}{\sim} \mathrm{N}_K(0, \mathrm{I}_K)$;

$$\tilde{Y}_{ij}^{\star} \mid \mu_{ij}, \boldsymbol{\lambda}_j, \boldsymbol{\eta}_i, \sigma^2 \overset{indep}{\sim} \mathrm{N}_1(\mu_{ij} + \boldsymbol{\lambda}_j' \boldsymbol{\eta}_i, \sigma^2). \tag{2.5}$$

By integrating over $\boldsymbol{\eta}_i$, we obtain the normal distribution with covariance matrix Σ in (2.3) for $\tilde{\mathbf{Y}}_i^{\star}$. The conditional independence between $\tilde{Y}_{ij}^{\star}$ given $\boldsymbol{\eta}_i$ in (2.5) greatly facilitates the posterior computation. Furthermore, it enables easy implementation of a zero-inflated model. Excess zeros in microbiome data are very common. If excess zeros are not compatible with the distribution in (2.2), the resulting inferences can be distorted. For a zero-inflated model, we introduce binary indicators δ_{ij} that represent the absence/presence of OTUs, and assume $\delta_{ij} \mid \epsilon_{ij} \overset{indep}{\sim} \mathrm{Ber}(\epsilon_{ij})$, where ϵ_{ij} is the probability of OTU j being absent in sample i. We let $\delta_{ij} = 1$ indicate the absence of OTU j in sample i, so $Y_{ij} = 0$. Given $\delta_{ij} = 0$, we assume, for $y = 0, 1, 2, \ldots$,

$$\begin{aligned} \mathrm{P}(Y_{ij} = y \mid \mu_{ij}, \boldsymbol{\lambda}_j, \boldsymbol{\eta}_i, \sigma^2, \delta_{ij} = 0) = \; & \Phi_1(\log(y+1) \mid \mu_{ij} + \boldsymbol{\lambda}_j' \boldsymbol{\eta}_i, \sigma^2) \\ & - \Phi_1(\log(y) \mid \mu_{ij} + \boldsymbol{\lambda}_j' \boldsymbol{\eta}_i, \sigma^2). \end{aligned} \tag{2.6}$$

Given the presence of an OTU, the model in (2.6) generates counts, some of which can be zero. Given $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_{iJ})$, a vector of $\tilde{Y}_{ij}^{\star}$ with $\delta_{ij} = 0$ follows a multivariate normal distribution, and its mean vector and covariance matrix are a subvector of $\boldsymbol{\mu}_i$ omitting the elements with $\delta_{ij} = 1$ and a submatrix of Σ omitting the rows and columns with $\delta_{ij} = 1$, respectively. That is, $\boldsymbol{\mu}_i$ and Σ provide inferences on the mean abundance

and interrelationship structure even when the zero inflation component is added to the model. We relate covariates $\mathbf{x}_i$ to the probability of $\delta_{ij} = 1$ by using a probit link function,

$$\epsilon_{ij} = \Phi_1(\kappa_{j0} + \mathbf{x}_i'\boldsymbol{\kappa}_j \mid 0, 1), \tag{2.7}$$

where κ_{j0} and $\boldsymbol{\kappa}_j = (\kappa_{j1}, \ldots, \kappa_{jP})'$ are parameters that quantify the effects of $\mathbf{x}_i$ on ϵ_{ij}. We consider a normal distribution for the prior of κ_{jp}, $\kappa_{jp} \overset{iid}{\sim} \mathrm{N}(\bar{\kappa}_p, u_\kappa^2)$, $p = 0, \ldots, P$. With a high proportion of zero counts, adding subject specific random effects into ϵ_{ij} may produce unstable model fitting (Agarwal et al., 2002). Thus, the model in (2.7) does not include subject specific random effects.

Lastly, we relate covariates $\mathbf{x}_i$ and subject-specific group factors g_i to the mean OTU abundances through μ_{ij};

$$\mu_{ij} = r_i + \alpha_j + \mathbf{x}_i'\boldsymbol{\beta}_j + s_{g_i,j}. \tag{2.8}$$

r_i and α_j are sample size factors and OTU size factors, respectively. The observed OTU counts are a product of both the library size (total number of reads) and the OTU baseline abundance. r_i's normalize OTU counts across samples, and α_j's account for variability in OTU baseline abundances. We let r_i and α_j random. Thus, the model performs model-based normalization and addresses compositionality. We will specify priors of r_i and α_j below. In (2.8), regression coefficients β_{jp} quantify the change in

the abundance of OTU j from the mean abundance by x_{ip} (so-called a factor effects model in an ANOVA setting). Under the formulation, choosing a reference category for a categorical covariate is not required, and an implicit assumption of the presence of an OTU under the arbitrarily chosen reference category is not needed to infer the effects of the other categories. When any covariate is categorical, $\mathbf{x}_i$ in (2.8) is different from that in (2.7) due to a different parameterization of the covariate. An example will be illustrated in § 2.3.2. When no covariate is available as in Simulation 1 in § 2.3.1 and the skin microbiome data in § 2.4.1, we simply drop the regression terms $\mathbf{x}_i'\boldsymbol{\kappa}_j$ and $\mathbf{x}_i'\boldsymbol{\beta}_j$ from (2.7) and (2.8), respectively, and use the simplified model to infer OTU interaction structure. $s_{g_i,j}$'s in (2.8) are random effects to account for between-subject heterogeneity and induce dependence among the samples collected from the same subject. We assume normal priors $\beta_{jp} \overset{iid}{\sim} \mathrm{N}(0, u_\beta^2)$ with fixed u_β^2. In addition, we place a sum-to-zero constraint on the prior of β_{jp}'s corresponding to the categories of a categorical covariate, and the model ensures meaningful inference on β_{jp}. If desired, a joint prior distribution of $\boldsymbol{\kappa}_j$ and $\boldsymbol{\beta}_j$ can be consider. For example, we assume $(\boldsymbol{\kappa}_j', \boldsymbol{\beta}_j')' \overset{iid}{\sim} \mathrm{N}(\mathbf{0}, \boldsymbol{V})$, and $\boldsymbol{V}$ accommodates potential association between covariates' effects on presence/absence of an OTU and their effects on the abundance of the OTU. We let $s_{g_i,j} \mid u_s^2 \overset{iid}{\sim} \mathrm{N}(0, u_s^2)$ and $u_s^2 \sim \mathrm{Ga}(a_s, b_s)$. Due to $s_{g_i,j}$, the marginal covariance matrix of $\widetilde{\mathbf{Y}}_i^\star$ is $\Omega = \Sigma + u_s^2 \mathrm{I}_J$, and the marginal correlations between OTUs j and j' are $\rho_{jj'} = \{\Sigma_{jj'} + u_s^2 1(j = j')\}/\sqrt{(\Sigma_{jj} + u_s^2)(\Sigma_{j'j'} + u_s^2)} \in (-1, 1)$. While any of parameters, Σ, Ω and $\rho_{jj'}$, can be considered as a measure of dependence between OTUs, we use $\rho_{jj'}$ for easy interpretation in the simulation studies and real data analyses illustrated

later.

Recall that the mean and median of $Y_{ij}^{\star}$ are proportional to $\exp(r_i + \alpha_j)$, implying that r_i and α_j are not identifiable. To circumvent potential identifiability issues, we follow Li et al. (2017) and use the mean-constrained prior with a mixture of mixture of normals on r_i and α_j;

$$\begin{aligned} r_i \mid \boldsymbol{\psi}^r, \boldsymbol{\omega}^r, \boldsymbol{\xi}^r &\overset{iid}{\sim} \sum_{l=1}^{L^r} \psi_l^r \left\{ \omega_l^r \mathrm{N}(\xi_l^r, u_r^2) + (1-\omega_l^r)\mathrm{N}\left(\frac{v_r - \omega_l^r \xi_l^r}{1-\omega_l^r}, u_r^2\right)\right\}, \\ \alpha_j \mid \boldsymbol{\psi}^\alpha, \boldsymbol{\omega}^\alpha, \boldsymbol{\xi}^\alpha &\overset{iid}{\sim} \sum_{l=1}^{L^\alpha} \psi_l^\alpha \left\{ \omega_l^\alpha \mathrm{N}(\xi_l^\alpha, u_\alpha^2) + (1-\omega_l^\alpha)\mathrm{N}\left(\frac{v_\alpha - \omega_l^\alpha \xi_l^\alpha}{1-\omega_l^\alpha}, u_\alpha^2\right)\right\}, \end{aligned} \tag{2.9}$$

where v_r and v_α are prespecified mean constraints for the distributions of r_i and α_j, respectively. u_r^2 and u_α^2 are fixed. Different from a multinomial model that conditions on sample total counts, our model assumes $\mathrm{E}(Y_{ij}^{\star} \mid \mu_{ij}, \Sigma) \propto \exp(\mu_{ij}) = \exp(r_i + \alpha_j + \mathbf{x}_i'\boldsymbol{\beta}_j + s_{g_i,j})$ in (2.8), and simultaneously performs model-based normalization through random r_i's. It flexibly accounts for compositionality in microbiome data and improves the inference on parameters of primary interest compared to a model using plug-in empirical estimates for normalizing factors (Shuler et al., 2021a). To specify the value of v_r, we obtain sample scale factor estimates by the cumulative sum scaling (CSS) normalization method in Paulson et al. (2013), and fix v_r at the average of those estimates. Specifically, we let $v_r = \frac{1}{N}\sum_{i=1}^{N} \log(\sum_{j=1|Y_{ij}\le q_i}^{J} Y_{ij})$, where q_i is set as the largest quantile such that the difference in quantiles across samples is small enough. Then we set $v_\alpha = \frac{1}{NJ}\sum_{i=1}^{N}\sum_{j=1}^{J} \log(Y_{ij}+0.01) - v_r$. Lee and Sison-Mangus (2018) and Shuler et al. (2021a) showed that overall means $r_i + \alpha_j$ can be well estimated under the

mean-constrained prior and their posterior inference is not sensitive to the choice of v_r and v_α. To complete the specification of the mean-constrained prior, we place Dirichlet priors for $\boldsymbol{\psi}^\chi = (\psi_1^\chi, \ldots, \psi_{L^\chi}^\chi)$ and beta priors for ω_l^χ, $\chi \in \{r, \alpha\}$, $\boldsymbol{\psi}^\chi \sim \text{Dir}(a_\psi^\chi, \ldots, a_\psi^\chi)$ and $\omega_l^\chi \overset{iid}{\sim} \text{Be}(a_\omega^\chi, b_\omega^\chi)$, where the hyperparameters a_ψ^χ, a_ω^χ and b_ω^χ are fixed. Finally, we set $\xi_l^\chi \overset{iid}{\sim} \text{N}(\bar{\xi}^\chi, v_\chi^2)$ with fixed $\bar{\xi}^\chi$ and v_χ^2. With random mixture weights, ω_l^χ and ψ_l^χ, and random locations ξ_l^χ, the mixture models in (2.9) flexibly capture various shapes of distributions, while keeping their means at v_χ and provide reasonable estimates of $r_i + \alpha_j$.

2.2.2 Posterior Computation

Let $\boldsymbol{\theta} = \{\lambda_{jk}, \phi_j, \tau_k, \kappa_{jp}, \delta_{ij}, \eta_i, \sigma^2, r_i, \alpha_j, \beta_{jp}, s_{g_i,j}, u_s^2, \omega_l^\alpha, \psi_l^\alpha, \xi_l^\alpha, \omega_l^r, \psi_l^r, \xi_l^r\}$ be a vector of all random parameters. We use Markov chain Monte Carlo (MCMC) methods to draw samples from the posterior distribution of $\boldsymbol{\theta}$. We write a Laplace distribution in (2.4) as a normal scale mixture to facilitate the posterior computation, and introduce latent mixture indicators for easy computation in updating ω_l^χ, ψ_l^χ and ξ_l^χ, $\chi \in \{r, \alpha\}$. Given the latent variables, all parameters except for ϕ_j and ω_l^χ are in standard conjugate forms and can be easily updated through a data augmented Gibbs step. Details of the posterior computation are given in Appendix § A.1. We examined the mixing and convergence of the Markov chains using trace plots and autocorrelation plots and did not find evidence of poor mixing or bad convergence for both the upcoming simulation examples and the real data analyses. The open-source code that implements the model is available online at https://github.com/Zsj950708/ZI-MLN. The detailed instructions

of implementation are in Appendix § A.2.

2.3 Simulation Studies

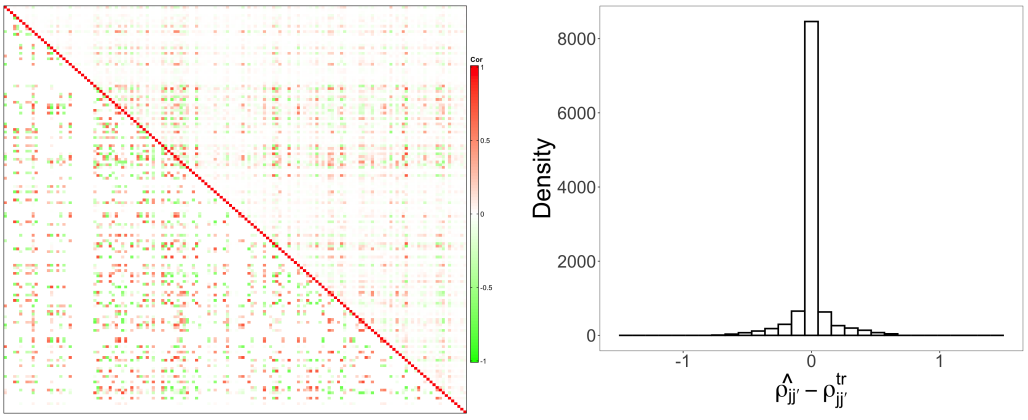
2.3.1 Simulation 1

We performed simulation studies and assessed the performance of the zero-inflated multivariate log-normal kernel model (ZI-MLN). For Simulation 1, we considered a case where no covariate is included, and each subject has one sample. We fitted a simplified model that has $\mu_{ij} = r_i + \alpha_j + s_{g_i,j}$ and $\epsilon_{ij} = \Phi_1(\kappa_{j0} \mid 0, 1)$. The simplified model is useful in estimating the interactions between OTUs for data without covariates. We let $J = 150$ OTUs and $N = 20$ samples, a sample from each of $M = 20$ subjects. For joint sparsity, we set $K^{\text{tr}} = 5$ and generated $e_{jk} \overset{iid}{\sim} \text{Ber}(g)$ with sparsity level $g = 0.8$. We then let $\lambda_{jk}^{\text{tr}} = 0$ if $e_{jk} = 1$ and otherwise, simulated $\lambda_{jk}^{\text{tr}} \overset{iid}{\sim} \text{Unif}(-3, 3)$. We let $\Sigma^{\text{tr}} = \Lambda^{\text{tr}}\Lambda^{\text{tr},\prime} + \sigma^{2,\text{tr}}\text{I}_J$ with $\sigma^{2,\text{tr}} = 1$. We also simulated random effects $s_{g_i,j}^{\text{tr}} \overset{iid}{\sim} \text{N}(0, u_s^{2,\text{tr}})$ with $u_s^{2,\text{tr}} = 1$, sample size factors $r_i^{\text{tr}} \overset{iid}{\sim} \text{Unif}(3, 7)$ and OTU size factors $\alpha_j^{\text{tr}} \overset{iid}{\sim} \text{Unif}(0, 2)$. We then simulated $\mathbf{Y}_i^{\star,\text{tr}} \overset{indep}{\sim} \text{log-N}_J(r_i^{\text{tr}} 1_J + \boldsymbol{\alpha}^{\text{tr}} + \boldsymbol{s}_i^{\text{tr}}, \Sigma^{\text{tr}})$. For excess zeros, we generated $\kappa_{j0}^{\text{tr}} \overset{iid}{\sim} \text{Unif}(-1, 0)$ and simulated $\delta_{ij}^{\text{tr}} \mid \epsilon_j^{\text{tr}} \overset{indep}{\sim} \text{Ber}(\epsilon_j^{\text{tr}})$ with $\epsilon_j^{\text{tr}} = \Phi_1(\kappa_{j0}^{\text{tr}} \mid 0, 1)$. We then let $Y_{ij} = 0$ if $\delta_{ij}^{\text{tr}} = 1$ and otherwise, let $Y_{ij} = \lfloor Y_{ij}^{\star,\text{tr}} \rfloor$. It yielded approximately 40% of Y_{ij} being 0. The lower left triangle of the heatmap in Fig 2.1(a) illustrates the true marginal correlation matrix $\rho_{jj'}^{\text{tr}} = \{\Sigma_{jj'}^{\text{tr}} + u_s^{2,\text{tr}} 1(j = j')\} / \sqrt{(\Sigma_{jj}^{\text{tr}} + u_s^{2,\text{tr}})(\Sigma_{j'j'}^{\text{tr}} + u_s^{2,\text{tr}})}$. Empirical correlation estimates $\rho_{jj'}^{\text{em}}$ are computed using transformed raw counts and illustrated in Appendix Fig A.2(a). It shows that naive correlation estimates are noisy

and do not capture the true interrelationship between OTUs.

To fit the model, we set the fixed hyperparameters as follows; For the mean-constrained priors of r_i and α_j, we let $L^r = 5, L^\alpha = 10$, $a^r_\psi = a^\alpha_\psi = 1$, and $a^r_\omega = b^r_\omega = a^\alpha_\omega = b^\alpha_\omega = 5$. The values of the mean constraints v^r and v^α were specified through the empirical approach described in § 2.2.1. We set the prior mean and variance of κ_{j0}, $\bar{\kappa}_0 = 0$ and $u^2_\kappa = 3$. Also, we set $a_\sigma = b_\sigma = 3$ and $a_s = b_s = 1$. Lastly, we set $K = 10$, $a_\phi = 1/2$, $a_\tau = 1$ and $b_\tau = 1/50$. We simulated posterior samples through MCMC described in § 2.2.2. We discarded the first 15,000 draws for burn-in and kept the next 15,000 draws for posterior inference. It took 25 minutes for every 5,000 iterations on a M1 Mac. Assessment of MCMC simulation convergence is discussed in Appendix § A.3. We also checked the posterior distributions of τ_k to examine if a greater value of K is needed. The posterior distributions of some τ_k's are greatly concentrated close to zero, indicating that $K = 10$ is sufficiently large for the dataset. We also performed sensitivity analyses to the specification of a_ϕ and b_τ to examine the robustness of the model in estimating Σ.

Posterior inference on the marginal correlations $\rho_{jj'}$ is illustrated in Fig 2.1. The heatmap in panel (a) compares posterior mean estimates $\hat{\rho}_{jj'}$ in the upper right triangle to their truth $\rho^{\text{tr}}_{jj'}$ in the lower left triangle. Panel (b) shows a histogram of the differences $\hat{\rho}_{jj'} - \rho^{\text{tr}}_{jj'}$, $j < j'$. In the histogram, the differences are tightly centered around 0, indicating that the method provides good estimates of the correlations. Our method identifies the truly inactive OTUs successfully, and the true OTU interrelationship structure is reasonably well captured even when the sample size is much smaller than the

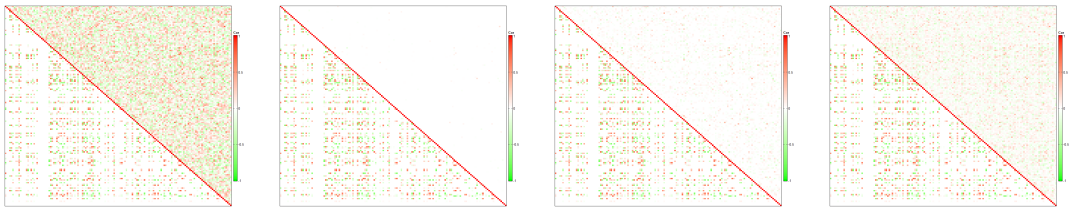


(a) $\rho^{\text{tr}}_{jj'}$ versus $\hat{\rho}_{jj'}$ (b) $\hat{\rho}_{jj'} - \rho^{\text{tr}}_{jj'}$

Figure 2.1: [Simulation 1] The upper right and lower left triangles of the heatmap in panel (a) illustrate posterior estimates of correlations $\hat{\rho}_{jj'}$ and their true values $\rho^{\text{tr}}_{jj'}$, respectively. Panel (b) has a histogram of differences between $\hat{\rho}_{jj'}$ and $\rho^{\text{tr}}_{jj'}$.

number of OTUs ($N = 20$ and $J = 150$), and excess zeros are present. Appendix Fig A.3 compares posterior mean estimates of baseline abundances $r_i + \alpha_j$ and probabilities ϵ_{ij} of an OTU being absent to their truth. In the figure, the absence/presence of OTUs and OTU baseline abundances are well estimated, which provides a crucial basis for the estimation of the parameters of primary interest, such as Σ. We performed posterior predictive checking to examine model fit under ZI-MLN. Fig 2.4(a) compares posterior predictive median estimates $\hat{y}^{\text{pred}}_{ij}$ of OTU counts to the observed counts y_{ij} and shows that our model provides a good model fit to the data.

For comparison, we applied SparCC (Friedman and Alm, 2012), SPIEC-EASI (Kurtz et al., 2015), CCLasso (Fang et al., 2015) and Zi-LN (Prost et al., 2021) that are briefly described in § 2.1. The comparators infer dependence structure between OTUs through the estimation of covariance or precision matrix under some sparsity assumptions and yield correlation estimates $\hat{\rho}_{jj'}$. The tuning parameter for sparsity in SparCC,



(a) SparCC (b) SPIEC-EASI (c) CCLasso (d) Zi-LN

Figure 2.2: [Simulation 1: Comparison] The upper right and lower left triangles of each heatmap illustrate estimates $\hat{\rho}_{jj'}$ of correlations between OTUs and their true values $\rho^{\text{tr}}_{jj'}$, respectively. Panels (a)-(d) are for SparCC, SPIEC-EASI, CCLasso and Zi-LN, respectively.

Table 2.1: [Simulation 1: Comparison] RMSEs are computed for correlations $\rho_{jj'}, j < j'$, binary indicator δ_{ij} of an OTU being absent in a sample and mean abundance μ_{ij} under ZI-MLN and comparators.

Model	$\rho_{jj'}$
ZI-MLN	**0.130**
SparCC	0.258
SPIEC-EASI	0.167
CCLasso	0.166
Zi-LN	0.173

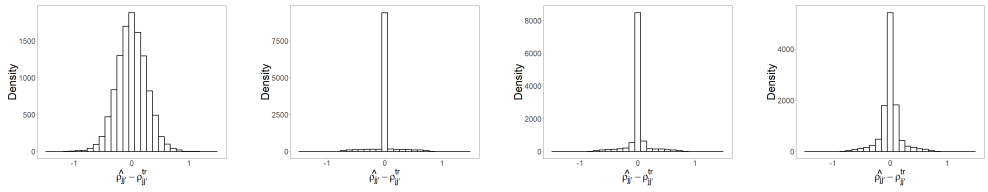
(a) $\rho_{jj'}$

Model	δ_{ij}	μ_{ij}
ZI-MLN	**0.084**	**0.453**
ZI-MLN without Λ	0.088	0.543
MetagenomeSeq	0.095	1.717

(b) δ_{ij} and μ_{ij}

SPIEC-EASI and Zi-LN is chosen by cross-validation. $\hat{\rho}_{jj'}$ under the comparators are compared to the true values $\rho^{\text{tr}}_{jj'}$ in Fig 2.2. Fig 2.3 illustrates histograms of differences $\hat{\rho}_{jj'} - \rho^{\text{tr}}_{jj'}$. Root mean square error (RMSE) for $\rho_{jj'}, j < j'$ for the models including ZI-MLN is shown in Tab 2.1(a). ZI-MLN outperforms in recovering the dependence structure between OTUs. Poor performance of the comparators can be because they do not account for overdispersion and/or excess zeros and/or they lack flexible normalization for compositionality. In addition, we compare our method to ZI-MLN without Λ, a simpler version of our ZI-MLN, and metagenomeSeq in Paulson et al. (2013) for comparison of the estimation of μ_{ij} and δ_{ij}. We simplified our ZI-MLN by letting $\Sigma = \sigma^2 \mathrm{I}_J$ and kept the remaining model components including zero-inflation and subject-specific random effects the same. We call it "ZI-MLN without Λ." MetagenomeSeq is a likelihood-based model that uses transformed counts $\log_2(y_{ij} + 1)$ and assumes a zero-inflated normal mixture model separately for individual OTUs, where the mean has a regression function of covariates, a sample size factor fixed at estimates by CSS normalization method and an OTU size factor similar to ZI-MLN. Under metagenomeSeq, the zero inflation probabilities of y are common for all OTUs in a sample and regressed on the sample total counts through a logit link. An EM algorithm is used to estimate unknown parameters. The additional comparators do not account for the interrelationships between OTUs and do not provide any inference on OTU interaction. We compared parameter estimates of μ_{ij} and δ_{ij} under each of the three models, including ZI-MLN, to the truth and computed RMSE for the parameters, summarized in Tab 2.1(b). The table shows that our model outperforms the comparators in the estimation of OTU mean abun-

dances and absence/presence. Especially, comparison to ZI-MLN without Λ indicates that ignoring the dependence structure among counts when it is present can deteriorate the inference on the other parameters, including μ_{ij}. It is also indicated from posterior predictive checking under ZI-MLN without Λ shown in Fig 2.4(b). Comparison of mean abundance estimates $\hat{\mu}_{ij}$ by metegenomSeq to observed counts in Fig 2.4(c) also shows potential model misfit under metagenomeSeq.

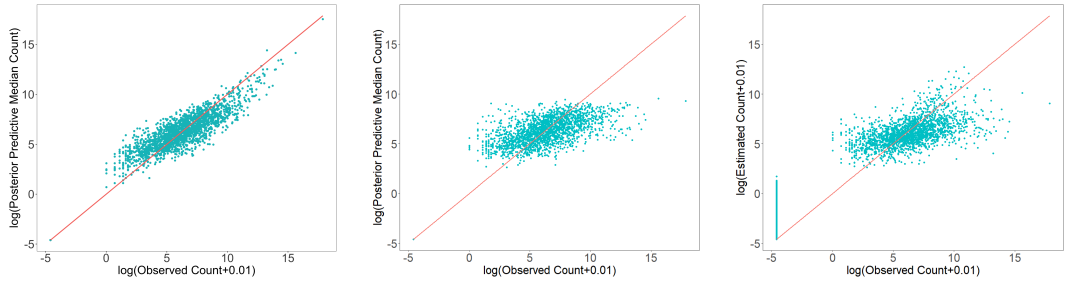


(a) SparCC (b) SPIEC-EASI (c) CCLasso (d) ZiLN

Figure 2.3: [Simulation 1: Comparison] A histogram of differences between $\hat{\rho}_{jj'}$ under SparCC, SPIEC-EASI, CCLasso and Zi-LN and $\rho^{\text{tr}}_{jj'}$, in panels (a)-(d), respectively.

2.3.2 Simulation 2

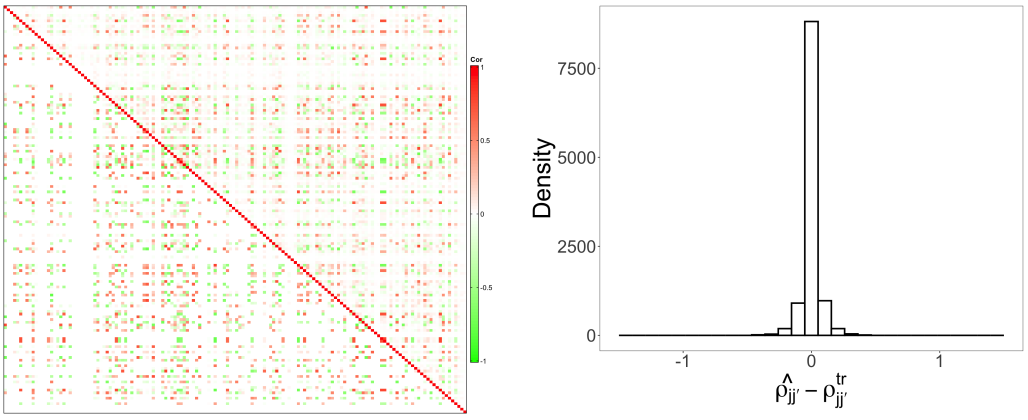
We conducted Simulation 2 for a case having covariates. We examined the estimation of covariate effects on OTU abundances and their presence/absence in addition to the estimation of Σ. We set the number of OTUs $J = 150$ and assumed two samples from each of $M = 35$ subjects under two experimental conditions. We thus have the number of samples $N = 70$ and $g_i \in \{1, \ldots, M\}$ with $n_{g_i} = 2$ for all g_i. The remaining setup is similar to that of Simulation 1. We set $K^{\text{tr}} = 5$, $\sigma^{2,\text{tr}} = 1$ and $u_s^{2,\text{tr}} = 1$, and simulated λ^{tr}_{jk}, r^{tr}_i, α^{tr}_j and $s^{\text{tr}}_{g_i,j}$, as done in Simulation 1. We included a binary covariate that represents the experimental conditions using a pair of dummy variables



(a) ZI-MLN (b) ZI-MLN without Λ (c) MetagenomeSeq

Figure 2.4: [Simulation 1] Scatter plots of observed $\log(y_{ij}+0.01)$ versus $\log(\hat{y}_{ij}^{\text{pred}}+0.01)$ estimated by ZI-MLN with Λ and ZI-MLN without Λ are shown in panels (a) and (b), respectively. $\hat{y}_{ij}^{\text{pred}}$ is the median estimate of the posterior predictive distribution. Panel (c) is the scatter plots of observed $\log(y_{ij}+0.01)$ versus $\log(\hat{\mu}_{ij}+0.01)$, where $\hat{\mu}_{ij}$ are mean abundances of OTUs estimated by metagenomeSeq.

$(x_{i1}, x_{i2}) \in \{(1,0),(0,1)\}$. The corresponding coefficients β_{j1} and β_{j2} thus quantify changes in mean abundance by a condition compared to the overall mean abundance $r_i^{\text{tr}} + \alpha_j^{\text{tr}}$. In addition, we included a continuous covariate, x_{i3} generated from $\text{N}(0,1)$, so we have $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})'$ with $P = 3$. For the coefficients, we set $\beta_{jp}^{\text{tr}} \overset{iid}{\sim} \text{N}(0,1)$ for $p = 1, \ldots, P$. For ϵ_{ij}, we let $\tilde{\mathbf{x}}_i = (x_{i2}, x_{i3})'$ with $P_\kappa = 2$ using x_{i1} as a reference category, and simulated $\kappa_{jp}^{\text{tr}} \overset{iid}{\sim} \text{Unif}(-0.5, 0)$, $p = 0, \ldots, P_\kappa$. We finally generated counts Y_{ij} as follows; we simulated $\mathbf{Y}_i^{\star,\text{tr}} \overset{indep}{\sim} \text{log-N}_J(r_i^{\text{tr}} 1_J + \boldsymbol{\alpha}^{\text{tr}} + \mathbf{x}_i' \boldsymbol{\beta}^{\text{tr}} + \boldsymbol{s}_i^{\text{tr}}, \Sigma^{\text{tr}})$, with $\Sigma^{\text{tr}} = \Lambda^{\text{tr}} \Lambda^{\text{tr},\prime} + \sigma^{2,\text{tr}} \text{I}_J$ and $\boldsymbol{\beta}^{\text{tr}}$ being a $J \times P$ matrix of β_{jp}^{tr}. We also generated binary indicators $\delta_{ij}^{\text{tr}} \mid \epsilon_{ij}^{\text{tr}} \overset{indep}{\sim} \text{Ber}(\epsilon_j^{\text{tr}})$ with $\epsilon_{ij}^{\text{tr}} = \Phi(\kappa_{j0}^{\text{tr}} + \kappa_j^{\text{tr},\prime} \tilde{\mathbf{x}}_i \mid 0, 1)$. We then let $Y_{ij} = 0$ if $\delta_{ij}^{\text{tr}} = 1$, and let $Y_{ij} = \lfloor Y_{ij}^{\star,\text{tr}} \rfloor$, otherwise. The simulated dataset has approximately 40% of counts being zero. Fig 2.5(a) and Appendix Fig A.5(a) illustrate the true marginal correlations $\rho_{jj'}^{\text{tr}}$ and their naive empirical estimates $\rho_{jj'}^{\text{em}}$ using transformed counts after the normalization, respectively.

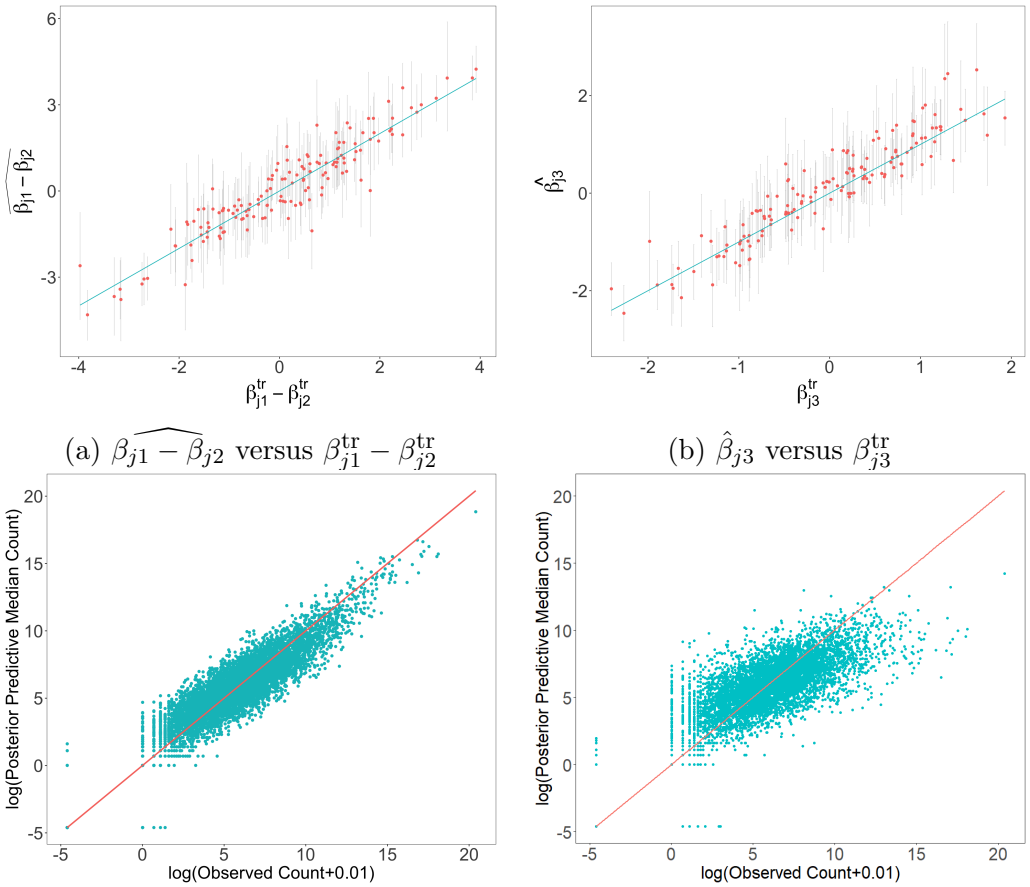


(a) $\rho^{\text{tr}}_{jj'}$ versus $\hat{\rho}_{jj'}$ (b) $\hat{\rho}_{jj'} - \rho^{\text{tr}}_{jj'}$

Figure 2.5: [Simulation 2] The upper right and lower left triangles of the heatmap in panel (a) illustrate posterior estimates of correlations $\hat{\rho}_{jj'}$ and the true values of the correlations $\rho^{\text{tr}}_{jj'}$, respectively. Panel (b) has a histogram of differences between $\hat{\rho}_{jj'}$ and $\rho^{\text{tr}}_{jj'}$.

We specified the fixed hyperparameter values similar to those in Simulation 1. We set $L^r = 8$ due to a larger sample size. We set $u^2_\beta = 25$ for the prior of β_{jp} and placed the sum-to-zero constraint for β_{j1} and β_{j2} for identifiability. We set $\bar{\kappa}_p = 0$ for all p and $u^2_\kappa = 3$. The MCMC simulation was run over 30,000 iterations, with the first 15,000 iterations discarded as burn-in. A discussion on the chain's convergence and mixing is in Appendix § A.3.2. It took 0.7 hours on average for every 5,000 iterations on a M1 Mac.

Fig 2.5 illustrates posterior mean estimates $\hat{\rho}_{jj'}$ of marginal correlations between OTUs j and j', $j \neq j'$. The figure shows that the underlying interrelationships between OTUs are well captured even with small sample size and excess zero counts. The histogram in panel (b) shows the differences $\hat{\rho}_{jj'} - \rho^{\text{tr}}_{jj'}$ are close to zero. Figs 2.6(a)-(b) and Appendix Figs A.6 (a)-(c) compare regression coefficient estimates, $\hat{\beta}_{jp}$ and $\hat{\kappa}_{jp}$



(a) $\widehat{\beta_{j1}-\beta_{j2}}$ versus $\beta^{\text{tr}}_{j1}-\beta^{\text{tr}}_{j2}$

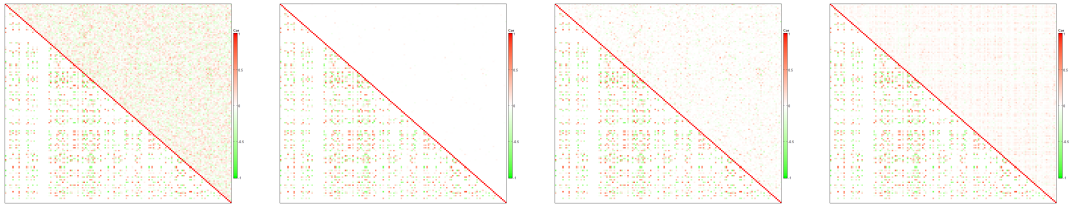
(b) $\hat{\beta}_{j3}$ versus β^{tr}_{j3}

(c) ZI-MLN with Λ

(d) ZI-MLN without Λ

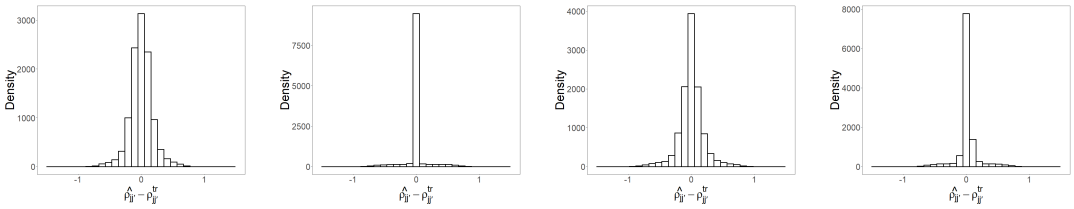
Figure 2.6: [Simulation 2] Panels (a) and (b) compare posterior estimates of regression coefficients $\widehat{\beta_{j1}-\beta_{j2}}$ and $\hat{\beta}_{j3}$ to the truth $\beta^{\text{tr}}_{j1}-\beta^{\text{tr}}_{j2}$ and β^{tr}_{j3}, respectively, where the vertical lines represent 95% credible intervals. Panels (c) and (d) compare posterior predictive median count estimates to their observed counts on the logarithm scale, $\log(y_{ij}+0.01)$ versus $\log(\hat{y}^{\text{pred}}_{ij}+0.01)$. ZI-MLN with Λ and ZI-MLN without Λ are used in panels (c) and (d), respectively.

to their true values. From Figs 2.6(a)-(b), posterior mean estimates of $\beta_{j1}-\beta_{j2}$ and β_{j3} are close to the true values. Here, $\beta_{j1}-\beta_{j2}$ quantifies the difference in the mean abundances between two categories of the binary covariate. Their posterior 95% credible intervals capture the truth well. Appendix Figs A.7 shows that posterior estimates $\widehat{r_i+\alpha_j}$ and $\hat{\epsilon}_{ij}$ are also close to their true values. To check the model fit, we compare



(a) SparCC (b) SPIEC-EASI (c) CCLasso (d) Zi-LN

Figure 2.7: [Simulation 2: Comparison] The upper right and lower left triangles of each heatmap illustrate estimates $\hat{\rho}_{jj'}$ of correlations between OTUs and their true values $\rho^{\text{tr}}_{jj'}$, respectively. Panels (a)-(d) are for SparCC, SPIEC-EASI, CCLasso and Zi-LN, respectively.



(a) SparCC (b) SPIEC-EASI (c) CCLasso (d) Zi-LN

Figure 2.8: [Simulation 2: Comparison] A histogram of differences between $\hat{\rho}_{jj'}$ under SparCC, SPIEC-EASI, CCLasso and ZiLN and $\rho^{\text{tr}}_{jj'}$, in panels (a)-(d), respectively.

median estimates $\hat{y}^{\text{pred}}_{ij}$ of the posterior predictive distributions to the observed counts. Fig 2.6(c) provides evidence for a good model fit under ZI-MLN.

For comparison, we applied the four comparators that provide estimates of associations between OTUs, SparCC, SPIEC-EASI, CCLasso and Zi-LN, to the simulated data. The heatmaps in Fig 2.7 and histograms in Fig 2.8 compare their estimates $\hat{\rho}_{jj'}$ to the truth $\rho^{\text{tr}}_{jj'}$. RMSE for $\rho_{jj'}$ are computed for comparison between the models including ZI-MLN. Tab 2.2(a) shows that ZI-MLN outperforms the comparators in estimating the dependencies between OTUs. Note that the comparators do not account for covariate effects, potentially resulting in poor performance. Also, we applied three other comparators, ZI-MLN without Λ, metagenomeSeq and edgeR (Robinson et al., 2010)

Table 2.2: [Simulation 2: Comparison] RMSEs are computed for $\rho_{jj'}, j < j'$, δ_{ij}, μ_{ij}, $\beta_{j2} - \beta_{j1}$, β_{j3} and κ_{jp} under ZI-MLN and comparators.

Model	$\rho_{jj'}$
ZI-MLN	**0.064**
SparCC	0.176
SPIEC-EASI	0.158
CCLasso	0.155
Zi-LN	0.157

(a) $\rho_{jj'}$

Model	δ_{ij}	μ_{ij}	$\beta_{j2} - \beta_{j1}$	β_{j3}	κ_{j0}	κ_{j1}	κ_{j2}
ZI-MLN	**0.096**	**1.096**	**0.597**	**0.385**	**0.215**	**0.184**	**0.334**
ZI-MLN without Λ	0.123	1.172	0.750	0.426	0.234	0.191	0.361
MetagenomeSeq	0.130	1.962	1.409	0.843	-	-	-
EdgeR	-	2.205	0.902	0.585	-	-	-

(b) δ_{ij}, μ_{ij}, $\beta_{j2} - \beta_{j1}$, β_{j3} and κ_{jp}

and compared the abundance and absence/presence related model parameters. EdgeR is a likelihood-based method that uses a negative binomial generalized linear regression approach for the analysis of HTS data. It uses the normalization factors estimated by an empirical Bayes strategy and does not account for excess zeros. Similar to ZI-MLN without Λ and metagenomeSeq, edgeR does not account for the dependence structure among OTUs and does not provide inferences on the relationship among OTUs. MetagenomeSeq and edgeR require selecting a category of a discrete covariate as a reference category, and their β_{jp}'s estimate changes in the mean abundance relative to that in the reference category. We chose x_{i1} as the reference for those methods. Appendix Figs A.6 (d)-(f) and A.8 compare estimates of β_{jp} and κ_{jp} under the comparators to the truth. RMSE for each of the four models, including ZI-MLN, is computed and summarized in Tab 2.2(b). RMSE of κ_{jp} is not computed for metagenomeSeq since it has a logit regression of ϵ_{ij} on the total sample count, but not on covariates. The results show that our model outperforms the comparators in the estimation of the parameters, δ_{ij}, μ_{ij}, β_{jp} and κ_{jp}. We also performed posterior predictive checking for ZI-MLN without Λ

by comparing $\hat{y}_{ij}^{\text{pred}}$ under the model to the observed counts. As shown in Fig 2.6(d), ZI-MLN without Λ provides a poor fit to the data. Their posterior mean estimates of σ^2 and u_s^2 are greatly inflated compared to their true value. Estimates $\hat{\sigma}^2$ and $\hat{u}_s^2$ are 3.86 and 0.77, respectively, while their true values are $\sigma^{2,\text{tr}} = 1$ and $u_s^{2,\text{tr}} = 1$. The comparison of the inference under ZI-MLN to that under ZI-MLN without Λ shows the necessity of modeling the dependence structure between OTUs to enhance the inference on the other parameters such as covariate effects when the interactions between OTUs are present. Estimates of the mean abundances under metagenomeSeq and edgeR are compared to the observed counts in Appendix Fig A.9.

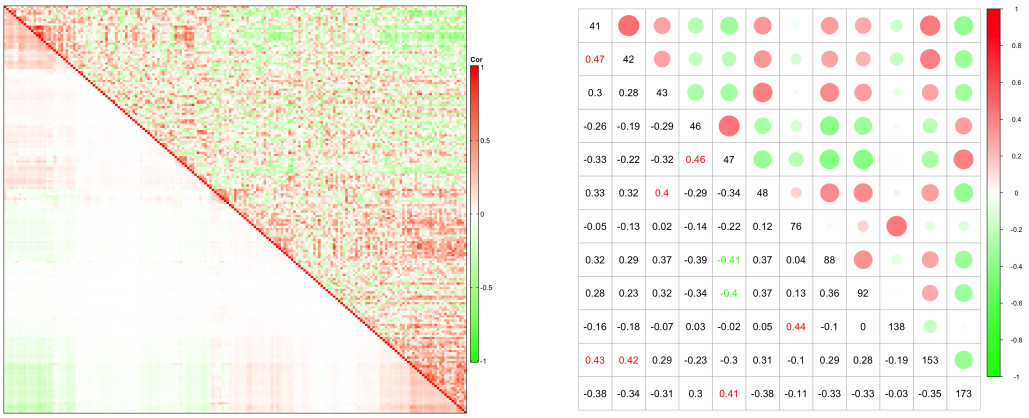
Additional Simulations We conducted additional simulation studies, Simulations 3-5 to examine the performance of our model under various settings. In Simulation 3, we first generated correlated mean vectors $\tilde{\boldsymbol{\mu}}_i^{\text{tr}} = (\tilde{\mu}_{i1}^{\text{tr}}, \ldots, \tilde{\mu}_{iJ}^{\text{tr}})$ from a multivariate normal distribution and simulated OTU counts from zero-inflated Poisson distributions with means $\exp(\tilde{\mu}_{ij}^{\text{tr}})$. The simulation results show that our model provides reasonable estimates of the parameters even when the simulation truth is different from the assumed model, showing the robustness of the model. Importantly, the OTU interaction structure is also reasonably well reconstructed even when the dependency is embedded in the mean abundances, and the sampling distribution is incorrectly specified. In Simulation 4, we kept the simulation setup the same as in Simulation 2, but let $\Sigma^{\text{tr}} = \sigma^{2,\text{tr}} \mathrm{I}_J$, i.e., OTU counts are independent given the mean parameters. Although the simulation truth is closer to the assumption made under ZI-MLN without Λ, the results show that ZI-MLN

performs almost the same as well. For Simulation 5, we used SparseDOSSA in Ma et al. (2021) to simulate a dataset. SparseDOSSA takes a real microbiome dataset as an input, estimates some input parameters of their data-generating model, and generates a realistic microbiome dataset that has a dependence structure between OTUs using the estimates. We used the skin microbiome dataset in § 2.4.1 as an input dataset. An open-source software, *SparseDOSSA2* is provided by the authors. SparseDOSSA estimates a precision matrix, one of the input parameters, with ℓ_1 penalty for sparsity. The sparsity assumption is similar to that under some of the comparators, SPIEC-EASI and CCLasso. It simulates count vectors from a multinomial distribution conditioning random total counts. The dataset in the scenario was thus simulated from a model significantly different from ZI-MLN. The results greatly demonstrate the robustness of ZI-MLN. The model-based normalization appropriately accounts for differences in total counts. More importantly, the model does a good job of capturing the dependence between OTUs in the truth and recovers the truly underlying between-OTU structure reasonably well. In all simulation studies, the results also show that our model compares very favorably relative to the comparators for estimation of covariate effects and of dependence structure between OTUs. More details of Simulations 3- 5 are in Appendix § A.3.3-A.3.5, respectively. In addition, we assumed a different sparsity level for Λ^{tr} by generating $e_{jk} \overset{iid}{\sim} \text{Ber}(g)$ with $g = 0.5$, and reran analyses under the settings of Simulations 1-4. The results show that ZI-MLN recovers the truth well with a lower sparsity level and works better than the comparators under the comparison metrics.

2.4 Real Data Analyses

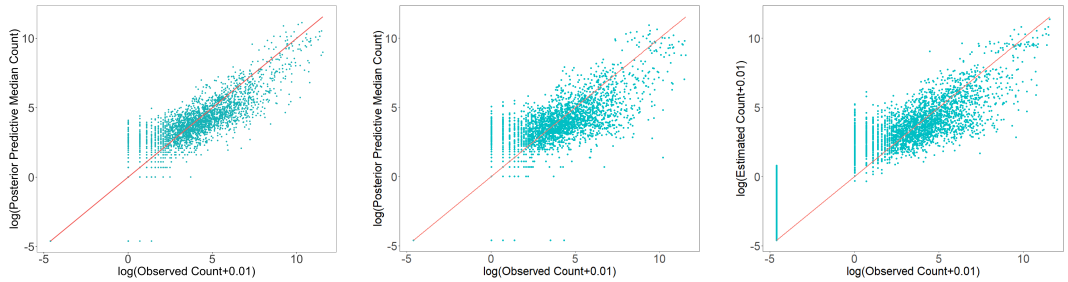
2.4.1 Skin Microbiome Data

We applied our ZI-MLN to a subset of the chronic wound microbiome data in Verbanic et al. (2020). The study was conducted to investigate the effect of debridement on the wound microbial community. Skin swab samples were collected under three conditions, healthy skin, pre-debridement, and post-debridement conditions. The skin microbiome dataset was analyzed Shuler et al. (2021a), which showed changes in the community-level microbial richness and abundance diversity by the experimental conditions. For an illustration of ZI-MLN without covariates, we used a subset of the data that consists of $N = 20$ healthy skin samples collected from $M = 20$ subjects and investigated the interaction structure between OTUs in the healthy skin samples. We removed OTUs that have zero counts in more than 50% of the samples, leaving $J = 187$ OTUs for analysis. The threshold of 50% was chosen so that each OTU has at least 10 nonzero counts, and the model parameters such as α_j can be reliably estimated. Manual inspection of the curated OTU list indicated that the threshold chosen did not eliminate OTUs of major biological importance. In addition, we performed sensitivity analysis to the specification of the threshold. We found that any reasonable choice has little impact on the posterior inference, showing robustness of our model. Details of the sensitivity analysis are summarized in Appendix § A.4.1. Fig 2.9(a) shows empirical correlation estimates $\rho_{jj'}^{\text{em}}$ computed using $\log(y_{ij} + 0.01)$ after normalization with CSS sample size factor estimates. To fit ZI-MLN, the values of the fixed hyperparameter



(a) $\hat{\rho}_{jj'}$ versus $\rho^{\text{em}}_{jj'}$ (b) $\hat{\rho}_{jj'}$ for some selected OTUs

Figure 2.9: [Skin Microbiome Data] Posterior correlation estimates $\hat{\rho}_{jj'}$ (lower left triangle) and empirical correlation estimates $\rho^{\text{em}}_{jj'}$ (upper right triangle) are shown in panel (a). Panel (b) have the OTUs having $|\hat{\rho}_{j,j'}| \geq 0.40$ for any $j' \neq j$.



(a) ZI-MLN (b) ZI-MLN without Λ (c) metagenomeSeq

Figure 2.10: [Skin Microbiome Data: Comparison] Panels (a) and (b) have scatter plots of observed $\log(y_{ij} + 0.01)$ versus $\log(\hat{y}^{\text{pred}}_{ij} + 0.01)$ under ZI-MLN and ZI-MLN without Λ, respectively. Panel (c) is the scatter plots of observed $\log(y_{ij} + 0.01)$ versus mean abundance estimates $\log(\hat{\mu}_{ij} + 0.01)$ by metagenomeSeq.

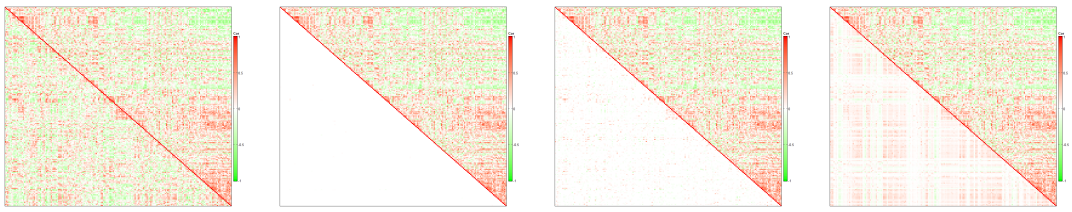
values were set similar to those of Simulation 1 in § 2.3.1. The MCMC simulation was run over 30,000 iterations, with the first 15,000 iterations discarded as burn-in. It took 25 minutes for every 5,000 iterations on a M1 Mac.

Fig 2.9(a) illustrates posterior mean estimates $\hat{\rho}_{jj'}$ of the marginal correlations for all OTUs. From panel (a), correlation estimates are overall small for most of (j, j'),

implying weak interactions between OTUs. Compared to $\rho_{jj'}^{\text{em}}$, $\hat{\rho}_{jj'}$'s are shrunken toward zero for many OTUs. The overall weak correlations among OTUs in the skin samples are consistent with previous analysis. Specifically, Bashan et al. (2016) analyzed data from the Human Microbiome Project and the Student Microbiome Project, and compared samples from the gut and oral microbiome to those from the skin microbiome. They reported that, while the gut and mouth microbiome samples appeared to exhibit universal dynamics of inter-species interactions, the extent of such interactions in the skin microbiome samples was relatively low. Fig 2.9(b) presents $\hat{\rho}_{jj'}$ for the OTUs that have $|\hat{\rho}_{jj'}| \geq 0.40$ for any $j' \neq j$, where the value of 0.4 is arbitrarily chosen to make the estimates readable. Taxonomic information of the OTUs in Fig 2.9(b) is given in Appendix Tab A.4. From panel (b) and the Appendix Tab A.4, OTUs 43 and 88 belonging to genera *Porphyromonas* and *Peptoniphilus*, respectively, are estimated to be positively correlated with $\hat{\rho} = 0.37$. Interestingly, they were found to co-occur in a large sample of genitourinary microbiome samples (Qin et al., 2021) as well as vaginal samples (Xiaoming et al., 2021) and were suggested to be 'keystone' species, i.e., strongly interacting species that help define their ecological system. These species are also found to co-occur in skin samples (Chattopadhyay et al., 2021), where they are more abundant in patients with diabetic foot ulcers (Park et al., 2019). OTUs 43 and 48 having correlation estimate $\hat{\rho} = 0.40$ belong to genera *Porphyromonas* and *Campylobacter*, respectively, that are both potentially pathogenic. *Porphyromonas* is a known pathogenic genus in periodontitis and is a risk factor in inflammatory bowel disease, while *Campylobacteri* is a known gut and oral pathogen with a role in inflammatory

bowel disease. Their positive correlation estimate may reflect a tendency to co-occur, as both are observed in inflammatory bowel disease (Cai et al., 2021). From Appendix Tab A.4, OTUs that have a large positive value of $\hat{\rho}_{j,j'}$ tend to be phylogenetically closely related. For example, OTUs 41 and 42 having $\hat{\rho} = 0.47$ belong to the same order *Micrococcales*. Similarly, OTUs 46 and 47 with $\hat{\rho} = 0.45$ having are in family *Chitinophagaceae*. On the other hand, some OTUs are estimated to have a positive association with phylogenetically distant OTUs. For example, the correlation estimates between OTU 153 and OTUs 41 and 42 are $\hat{\rho} = 0.44$ and 0.41, respectively, but OTU 153 is not phylogenetically closely related to OTUs 41 and 42. Interestingly, OTU 153 has similar interaction patterns with OTUs 41 and 42 in the same genus. Fig 2.10(a) has a scatter plot comparing the posterior predictive median estimates $\hat{y}_{ij}^{\text{pred}}$ to the observed counts. The posterior predictive checking indicates a good model fit by ZI-MLN.

We also applied the comparators, SparCC, SPIEC-EASI, CCLasso and Zi-LN to the skin microbiome data for comparison. Their correlation estimates $\hat{\rho}_{jj'}$ are illustrated in Fig 2.11 with the naive estimates of the correlations. SPIEC-EASI and CCLasso produce $\hat{\rho}_{jj'}$ very close or equal to zero for most OTU pairs, while SparCC has nonzero estimates for a majority of $\rho_{jj'}$. In addition, ZI-MLN without Λ and metagenomeSeq are applied for further comparison. In Fig 2.10(b), the posterior predictive median estimates $\hat{y}_{ij}^{\text{pred}}$ under ZI-MLN without Λ are compared to the observed counts. In panel (c), mean abundance estimates under metagenomeSeq are compared to the observed counts. A comparison of those plots to that in panel (a) indicates that our ZI-MLN provides a better model fit, possibly because our model accounts for microbial



(a) SparCC (b) SPIEC-EASI (c) CCLasso (d) Zi-LN

Figure 2.11: [Skin Microbiome Data: Comparison] Correlation estimates $\hat{\rho}_{jj'}$ (lower left triangle) and empirical correlation estimates $\rho^{\text{em}}_{jj'}$ (upper right triangle) are shown. The estimates in panel (a)-(d) are obtained by SparCC, SPIEC-EASI, CCLasso and Zi-LN, respectively.

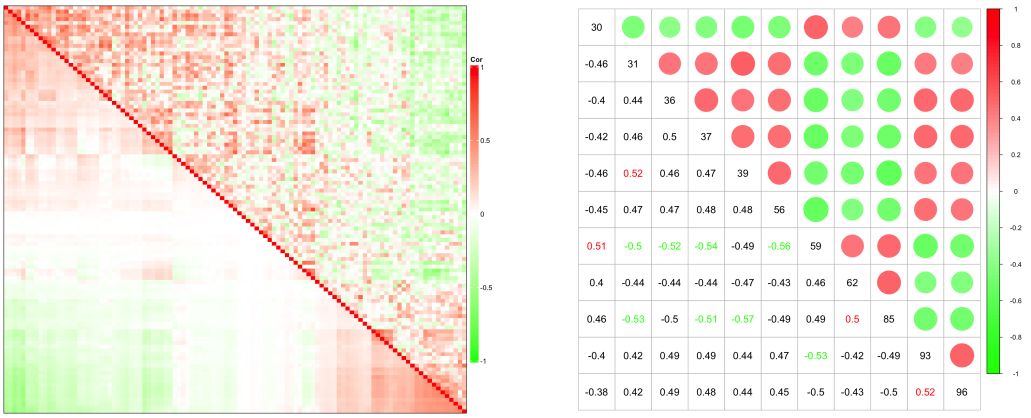
interactions.

2.4.2 Human Gut Microbiome Data

We analyzed the microbiome dataset available from the inflammatory bowel disease (IBD) multi-omics database (https://ibdmdb.org/) with our ZI-MLN. Crohn's disease (CD) and ulcerative colitis (UC) are the most prevalent forms of IBD and are characterized by chronic inflammation of the gastrointestinal tract. As part of the Integrative Human Microbiome Project (iHMP), Lloyd-Price et al. (2019) conducted an integrated study of multiple molecular features of the gut microbiome to investigate host- and microbiome-specific taxonomic and molecular features related to IBD and how they vary over time. In the study, biopsies were taken during the initial screening colonoscopy from the participants who were recruited from multiple medical centers and sequenced using 16S rRNA gene amplicon sequencing. For an illustration of our statistical model, we used part of their 16S rRNA sequencing data. In particular, we included the samples obtained from 37 pediatric participants from two recruitment sites,

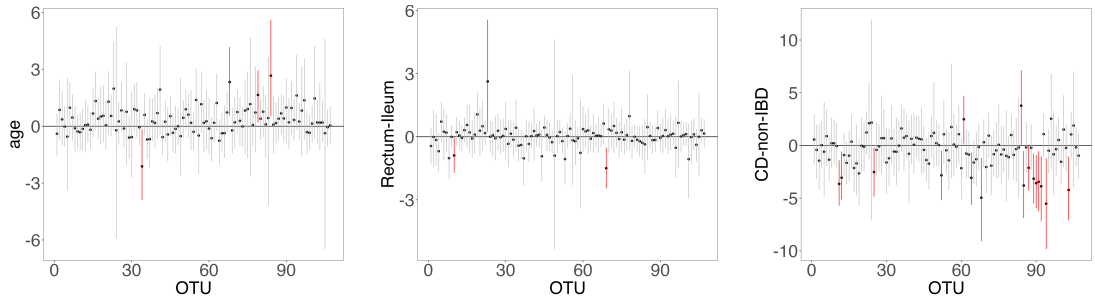
Cincinnati Children's Hospital and Massachusetts General Hospital (MGH) Pediatrics. For some subjects, two samples were collected from different biopsy locations, resulting in a total of 67 samples. In addition to biopsy locations, we included one continuous covariate, age and five categorical covariates such as sex, race, recruitment site and disease phenotype. Disease phenotype is a trinary covariate taking a value of UC, CD or non-IBD, and the others are binary, resulting in $P = 12$ after adding dummy variables to indicate the categories of the discrete covariates. Appendix Tab A.5 lists all covariates with their supports. We removed OTUs having zero count in more than 80% of the samples or average counts smaller than five. $J = 107$ OTUs are left after the preprocessing. With the threshold of 80%, each OTU has approximately 13.4 nonzero counts, similar to that in the skin microbiome data analysis, to ensure reliable estimates of κ_{jp}, $\boldsymbol{\beta}_{jp}$ and Σ. We specified hyperparameters similar to those in § 2.3.2. The MCMC simulation was run over 30,000 iterations, with the first 15,000 iterations discarded as burn-in. It took 0.75 hours for every 5,000 iterations on a M1 Mac.

Posterior mean estimates $\hat{\rho}_{jj'}$ of the marginal correlations (lower left triangle) are illustrated with naive empirical correlation estimates $\rho^{\text{em}}_{jj'}$ (upper right triangle) in Fig 2.12(a). The figure shows relatively rich microbial interactions in the gut microbiome samples as reported in Bashan et al. (2016). Fig 2.12(b) reports $\hat{\rho}_{jj'}$ for the OTUs having $|\hat{\rho}_{jj'}| > 0.5$ for any $j' \neq j$, where the value of 0.5 is chosen to make the estimates in the figure readable. Taxonomic information of the OTUs in panel (b) is in Appendix Tab A.6. In panel (b), a group of OTUs 31, 37, 39, 44, 56, 93 and 96 that are positively correlated with each other, are taxa that are found to indicate dysbiotic



(a) $\hat{\rho}_{jj'}$ versus $\rho^{\text{em}}_{jj'}$ (b) $\hat{\rho}_{jj'}$ for some selected OTUs

Figure 2.12: [Human Gut Microbiome Data]: Posterior marginal correlation estimates $\hat{\rho}_{jj'}$ (lower left triangle) and empirical correlation estimates $\rho^{\text{em}}_{jj'}$ (upper right triangle) are shown in panel (a). Panel (b) illustrates the OTUs having $|\hat{\rho}_{jj'}| > 0.5$ for any $j' \neq j$.



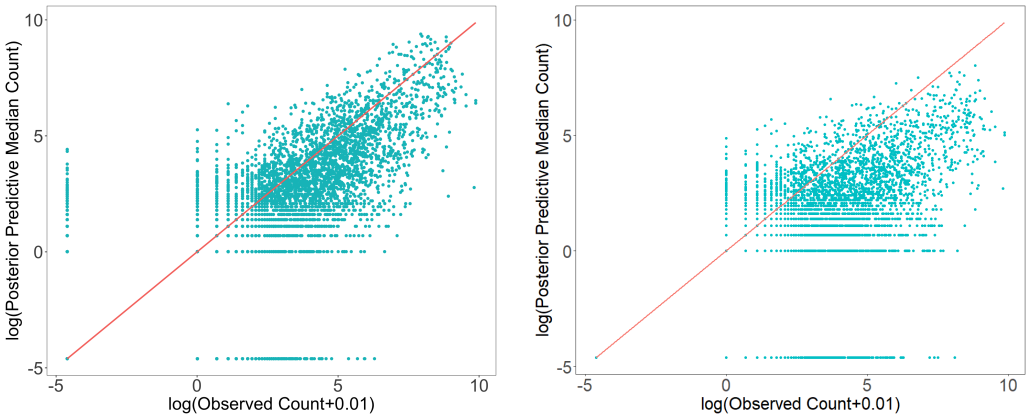
(a) β_{age} (b) $\beta_{Rectum} - \beta_{Ileum}$ (c) $\beta_{CD} - \beta_{non-IBD}$

Figure 2.13: [Human Gut Microbiome Data] Posterior inference of regression coefficients β_{age}, $\beta_{Rectum} - \beta_{Ileum}$, and $\beta_{CD} - \beta_{non-IBD}$, where the posterior mean estimates are denoted by dots, and the 95% credible estimates with vertical lines. The intervals that do not contain zero are marked.

microbiota from gastrointestinal diseases. For example, OTUs 31 and 39 that belong to family *Erysipelotrichaceae* are observed to be related to gastrointestinal inflammatory disorders (Kaakoush, 2015). And some species in *Escherichia* (OTU 93) (e.g., E. Coli (Mirsepasi-Lauridsen et al., 2019)) and *Clostridium* (OTUs 31 and 96) (e.g., C. difficile (Nitzan et al., 2013)) are known to be related to the development of IBD. Another

group of OTUs that are positively associated with each other includes genera, *Bacteroides* (OTU 59), *Faecalibacterium* (OTU 30), and *Ruminococcaceae* (OTU 85). The group of those genera contains species that were found active in metabolic processes and can produce short-chain fatty acids (Parada Venegas et al., 2019). These species might interact though exchanging metabolic products; for example, *Bacteroides thetaiotaomicron* and *Faecalibacterium prausnitzii* were found metabolically complementary, where the former is an acetate producer, and the latter is acetate consumer and butyrate producer (Wrzosek et al., 2013). Furthermore, such metabolic functions might be part of a complex interplay between the microbiota and disease states. For example, butyrate is an anti-inflammation promoter, and the decrease of butyrate producers might also indicate dysbiotic gut microbiota (Andrade et al., 2020). Interestingly, the OTUs in those two groups are negatively associated. The correlation patterns between the groups indicate how gut microbiota may shift from dysbiosis and may suggest further investigation through experiments. From taxonomic information in Appendix Tab A.6, the OTUs in the groups belong to different families and orders, indicating that phylogenetically distant OTUs interact in gut microbiota.

Fig 2.13 and Appendix Figs A.39 (a)-(b) illustrate posterior mean estimates $\hat{\beta}_{jp}$ and $\hat{\kappa}_{jp}$ of the regression coefficients, respectively, with their 95% credible intervals for some selected covariates. Dots represent point estimates and vertical lines interval estimates. In the figures, β_{jp} and κ_{jp} that do not contain zero in their 95% credible interval are marked. In addition, Appendix Tabs A.7 and A.8 provide taxonomic information of the OTUs whose abundance or presence/absence is statistically associated



(a) ZI-MLN (b) ZI-MLN without Λ

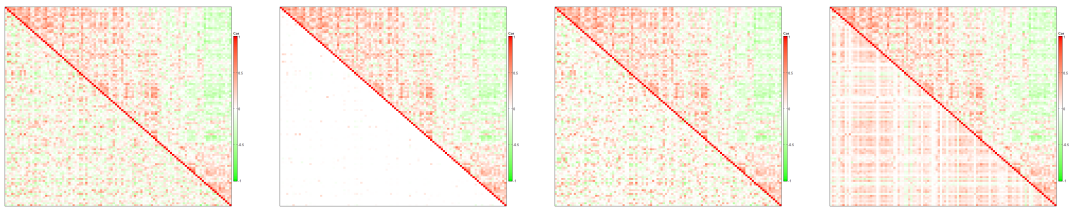
Figure 2.14: [Human Gut Microbiome Data: Comparison]: Panels (a) and (b) have scatter plots of observed $\log(y_{ij} + 0.01)$ versus $\log(\hat{y}_{ij}^{\text{pred}} + 0.01)$ under ZI-MLN and ZI-MLN without Λ, respectively.

with change in covariates. Overall, the covariate effects are statistically significant for a small number of OTUs. From panel (c), the effect of having condition CD compared to non-IBD $\beta_{CD} - \beta_{non-IBD}$ is statistically significant for 14 OTUs. The effect estimates are negative for those except for OTU 84, which implies that their abundance is lower for a subject with CD than for a subject with non-IBD. Also, among those, 13 OTUs belong to phylum *Firmicutes* and order *Clostridiales*. Significant decrease in abundance of phylum *Firmicutes* (*Clostridium leptum* and *Clostridium coccoides* groups) in active IBD subjects compared to that in non-IBD subjects is reported in Sokol et al. (2009), Vester-Andersen et al. (2019) and Alam et al. (2020). Lloyd-Price et al. (2019) also reported a statistically significant decrease in abundance of *Clostridium leptum* in active IBD subjects. We compare posterior predictive median estimates of OTU counts to the observed data in Fig 2.14(a) to access the model fit. The figure shows that the model fits the data well.

For comparison, we applied the comparators, SparCC, SPIEC-EASI, CCLasso and Zi-LN to the gut microbiome data. Fig 2.15 illustrates $\hat{\rho}_{jj'}$ under the comparators. Also, additional comparators, ZI-MLN without Λ, metagenomeSeq and edgeR were applied. The first set of the comparators does not account for covariate effects, and the second set does not infer the dependence structure between OTUs. SPIEC-EASI yields a very sparse estimate, whereas the other comparators produce very dense estimates. Appendix Figs A.39(c)-(d) and A.40 illustrate posterior estimates of regression coefficients β_{jp} and κ_{jp} obtained by the second set of the comparators. While ZI-MLN without Λ yields similar estimates, the estimates under metagenomeSeq and edgeR are greatly different from those under ZI-MLN. Specifically, under metagenomeSeq, the effects of covariate *age* are positive and statistically significant for most OTUs. A similar pattern is also observed from edgeR. For ZI-MLN without Λ, we further examine posterior predictive distributions of OTU counts (shown in Fig 2.14(b)). Compared to the fit under ZI-MLN, ZI-MLN without Λ yields a poor fit, especially for large counts. Appendix Fig A.41 compares mean abundance estimates under edgeR and metagenomeSeq to the observed counts and indicates poor model fit under those models.

2.5 Discussion

We have presented a Bayesian zero-inflated rounded log-normal kernel model to analyze multivariate count data with excess zeros. Different from most existing models, the model directly infers interrelationships between counts and produces reliable



(a) SparCC (b) SPIEC-EASI (c) CCLasso (d) Zi-LN

Figure 2.15: [Human Gut Microbiome Data: Comparison] Correlation estimates $\hat{\rho}_{jj'}$ by SparCC, SPIEC-EASI, CCLasso and Zi-LN (lower left triangle) and empirical correlation estimates $\rho^{\text{em}}_{jj'}$ (upper right triangle) are shown in panel (a)-(d), respectively.

inference on microbial interaction with a small sample size. It offers a straightforward interpretation of microbial dependence structures. Furthermore, the model simultaneously incorporates covariates and accounts for excess zeros. The simulations showed that the developed model compares very favorably in parameter estimation and model fit to a model that ignores between-OTUs' dependence structure and some popular alternatives that do not model covariate effects and/or dependence structure.

ZI-MLN can be further extended to accommodate more complex data structures. Specifically, Lloyd-Price et al. (2019) collected multi-omics data to obtain a comprehensive understanding of the IBD microbial ecosystem. Multi-omic measurements from the same subject may be interrelated, and joint analysis of bacterial sequencing data with other types of sequencing data such as viral sequencing data can be useful. In general, latent factor models provide a convenient way to model complex interrelationship structures in multivariate data and can be extended to accommodate multiple coupled observation matrices, e.g., a group factor model (Zhao et al., 2016). In that vein, our ZI-MLN can be extended to jointly analyze multiple correlated count matrices from a multi-omics study using an approach of a group factor model. Another possible

extension is to incorporate phylogenetic information into the model. Investigating potential interactions between phylogenetically related microbes is biologically interesting, e.g., see Faust et al. (2012); Connor et al. (2017); Kamneva (2017). Similar to Lo and Marculescu (2018), phylogenetic information can be utilized in building a prior model of Σ.

Chapter 3

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

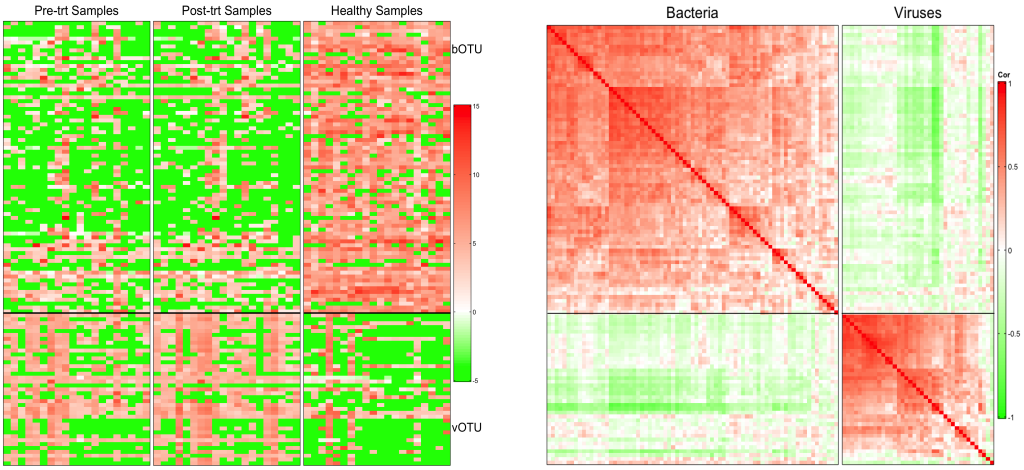
3.1 Introduction

3.1.1 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., 2012b; 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. (2023a), 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 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. (2023a)



(a) Log-transformed normalized OTU counts (b) Empirical correlation estimates

Figure 3.1: [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.

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 § 3.4. After preprocessing, the dataset comprises counts of 75 bOTUs and 39 vOTUs in the two domains, bacteria and viruses, for the 60 samples. Fig 3.1(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 3.1(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 Fig B.15. The figures indicate potential interactions between OTUs within and across different domains.

3.1.2 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. Fig B.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 3.1(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.

3.1.3 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., 2023a). 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 mod-

els 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 chapter is organized as follows. § 3.2 details the development of Sp-BGFM and describes the prior specification and posterior computation. In § 3.3, we evaluate the performance of Sp-BGF under different simulation settings and compare it to several popular alternatives. § 3.4 demonstrates the application of our method to the multi-domain skin microbiome dataset. Finally, § 3.5 provides a brief discussion and

conclusion.

3.2 Model and Posterior Inference

3.2.1 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 $\mathbf{y}_i^\star \in \mathbf{R}_+^J$ and build a flexible model for $\mathbf{y}_i^\star$. For sample i from subject s_i, we assume

$$\mathbf{y}_i^\star \mid \mathbf{r}_i, \boldsymbol{\alpha}_{s_i}, \Sigma \overset{indep}{\sim} \text{log-N}_J(\mathbf{y}^\star \mid \boldsymbol{\alpha}_{s_i} + \mathbf{r}_i, \Sigma), \quad i = 1, \ldots, N, \tag{3.1}$$

$$\boldsymbol{\alpha}_{s_i} \mid G \overset{iid}{\sim} G(\boldsymbol{\alpha}), \quad s_i \in \{1, \ldots, S\}. \tag{3.2}$$

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;

$$\mathrm{P}(\mathbf{y}_i = \mathbf{y} \mid \mathbf{r}_i, \boldsymbol{\alpha}_{s_i}, \Sigma) = \int_{A(\mathbf{y})} f_{\mathbf{y}^\star}(\mathbf{y}^\star \mid \boldsymbol{\alpha}_{s_i} + \mathbf{r}_i, \Sigma)\, d\mathbf{y}^\star, \tag{3.3}$$

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. $\boldsymbol{\alpha}_{s_i} = [\boldsymbol{\alpha}_{s_i 1}, \ldots, \boldsymbol{\alpha}_{s_i M}]'$ is a $J-$dim vector of OTU abundances, where a subvector $\boldsymbol{\alpha}_{s_i m} = (\alpha_{s_i mj})$, $j = 1, \ldots, J_m$ 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 (3.1), $\exp(\alpha_{s_i mj} + r_{im})$ is the median of $y^\star_{imj}$ 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. $\alpha_{s_i mj}$ 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 $\Sigma^{mm'}_{jj'}$ denote the element of Σ corresponding to the covariance between OTU j of group m and OTU j' of group m'. Letting $\mu_{imj} = \alpha_{s_i mj} + r_{im}$, 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., 2018). 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, $k = 1, \ldots, K$,

$$
\begin{aligned}
\tau_k \mid a_\tau, b_\tau &\overset{iid}{\sim} \mathrm{Ga}(a_\tau, b_\tau/J), \\
\boldsymbol{\phi}_k = (\phi_{11k}, \ldots, \phi_{MJ_Mk}) \mid a_\phi &\overset{iid}{\sim} \mathrm{Dir}(a_\phi, \ldots, a_\phi), \\
\zeta_{mjk} &\overset{iid}{\sim} \mathrm{C}^+(0,1), \quad m = 1, \ldots, M, \ j = 1, \ldots, J_m, \\
\lambda_{mjk} \mid \phi_{mjk}, \tau_k, \zeta_{mjk} &\overset{indep}{\sim} \mathrm{N}(0, \zeta_{mjk}^2 \phi_{mjk} \tau_k),
\end{aligned}
\tag{3.4}
$$

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 (3.4), $\boldsymbol{\phi}_k$ 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.

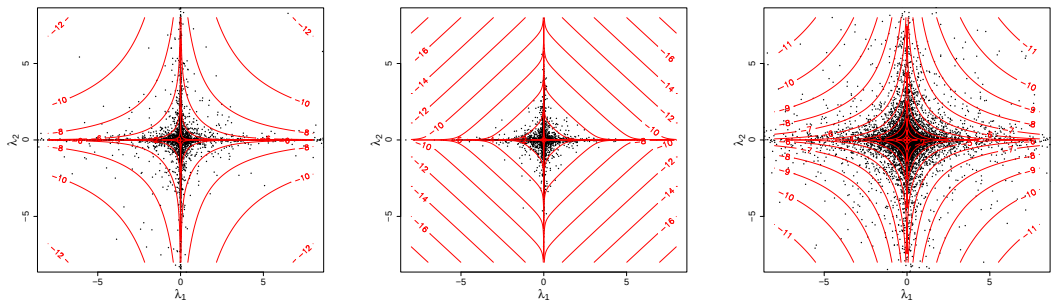
Theorem 3.2.1. *Let $J = 2$. Assume $\phi_1 \sim Be(a_\phi, a_\phi)$ and let $\phi_2 = 1 - \phi_1$. Assume the Dir-HS distribution in* (3.4) *as a joint distribution for $\boldsymbol{\lambda} = (\lambda_1, \lambda_2) \in \mathrm{R}^2$ given τ. Without loss of generality, let $\tau = 1$. The marginal 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 $\lambda_1 \neq 0$,*

$$\begin{aligned} 2^{2a_\phi - \frac{5}{2}} \pi^{-2} \frac{\Gamma^2(a_\phi + 1/2)}{\Gamma(2a_\phi + 1/2)} \frac{4}{\lambda_1^2} \, {}_3F_2\left(1, 1, a_\phi + 1/2; 2, 2a_\phi + 1/2; -\frac{4}{\lambda_1^2}\right) < \quad & \Pi_{Dir\text{-}HS}(\lambda_1) \\ < \quad 2^{2a_\phi - \frac{3}{2}} \pi^{-2} \frac{\Gamma^2(a_\phi + 1/2)}{\Gamma(2a_\phi + 1/2)} \frac{2}{\lambda_1^2} \, {}_3F_2\left(1, 1, a_\phi + 1/2; 2, 2a_\phi + 1/2; -\frac{2}{\lambda_1^2}\right), & \end{aligned} \tag{3.5}$$

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

$$\frac{1}{\sqrt{2\pi^5}} \left\{\sinh^{-1}(2/|\lambda_1|)\right\}^2 \; < \; \Pi_{Dir\text{-}HS}(\lambda_1) \; < \; \sqrt{\frac{2}{\pi^5}} \left\{\sinh^{-1}(\sqrt{2}/|\lambda_1|)\right\}^2, \tag{3.6}$$

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



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

Figure 3.2: 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.

A proof is given in Appendix B.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 3.2(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. Appendix Proposition B.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. Appendix Figs B.1 and B.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 (3.1)-(3.3), 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 (3.2),

$$\begin{aligned} G(\boldsymbol{\alpha}) &= \prod_{m=1}^{M}\prod_{j=1}^{J_m} G_{mj}(\alpha_{mj}) \\ &= \prod_{m=1}^{M}\prod_{j=1}^{J_m}\left[\sum_{l=1}^{\infty}\psi_{ml}^{\alpha}\left\{\omega_{ml}^{\alpha}\delta_{\xi_{mjl}^{\alpha}} + (1-\omega_{ml}^{\alpha})\delta_{\left(\frac{\nu_{mj}^{\alpha}-\omega_{ml}^{\alpha}\xi_{mjl}^{\alpha}}{1-\omega_{ml}^{\alpha}}\right)}\right\}\right], \end{aligned} \tag{3.7}$$

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 (3.7) are constructed using a stick-breaking process (Sethuraman, 1994); let $\psi_{m1}^{\alpha} = V_{m1}^{\alpha}$ and $\psi_{ml}^{\alpha} = V_{ml}^{\alpha}\prod_{l'=1}^{l-1}(1-V_{ml'}^{\alpha})$, $l > 1$ 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. Observe that individual parameters α_{s_imj} and r_{im} in μ_{imj} are not identifiable due to the multiplicative structure, $\mathrm{E}(\log(y_{imj}^{\star}) \mid \alpha_{s_imj}, r_{im}) = \alpha_{s_imj} + r_{im}$. Under (3.7), the prior and posterior means of α_{s_imj} are fixed at ν_{mj}^{α}, and $\mathrm{E}(\log(y_{imj}^{\star}) \mid G_{mj}, r_{im})$ fixed at $\nu_{mj}^{\alpha} + r_{im}$. 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 $y_{imj}^{\star}$ can be written as a Dirichlet process mixture with a log-normal mixture kernel in Antoniak (1974). Also, the model in (3.7) 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_{ml}^{α}'s and ω_{ml}^{α}'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 (3.7), we consider a flexible infinite mixture model for r_{im};

$$r_{im} \mid \psi^r_{ml}, \omega^r_{ml} \overset{indep}{\sim} H_m = \sum_{l=1}^{\infty} \psi^r_{ml} \left\{ \omega^r_{ml} \mathrm{N}(\xi^r_{ml}, u^2_r) + (1-\omega^r_{ml}) \mathrm{N}\left(\frac{\nu^r_m - \omega^r_{ml}\xi^r_{ml}}{1-\omega^r_{ml}}, u^2_r \right) \right\}, \quad (3.8)$$

where ν^r_m and u^2_r are fixed. The prior and posterior expectations of r_{im} are ν^r_m in (3.8), and $\mathrm{E}(\log(y^\star_{imj}) \mid G_{mj}, H_m)$ fixed at $\nu^\alpha_{mj} + \nu^r_m$. Each group has different means, as indicated in our motivating application as illustrated in Appendix Fig B.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 $\xi^r_{ml} \mid \nu^r_m, u^2_{\xi^r} \overset{iid}{\sim} \mathrm{N}(\nu^r_m, u^2_{\xi^r})$, $\omega^r_{ml} \mid a^r_\omega, b^r_\omega \overset{iid}{\sim} \mathrm{Be}(a^r_\omega, b^r_\omega)$, $\psi^r_{m1} = V^r_{m1}$ 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};

$$\mu_{imj} = r_{im} + \alpha_{s_i mj} + \mathbf{x}_i'\boldsymbol{\beta}_{mj}. \quad (3.9)$$

Assume $\beta_{mjp} \overset{iid}{\sim} \mathrm{N}(0, u^2_\beta)$ 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

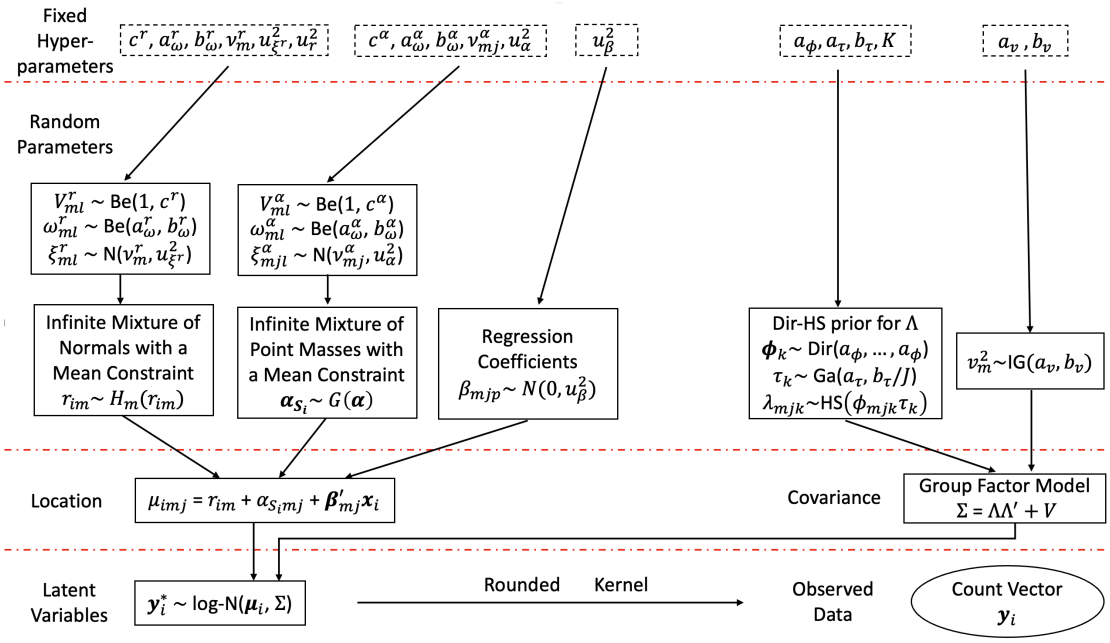


Figure 3.3: A graphical representation of Sp-BGFM. Fixed hyperparameters are in boxes with dashed lines, while random parameters are in boxes with solid lines. Observables are represented within circles.

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 3.3. In Appendix § B.2, we illustrate the distribution of observables under Sp-BGFM to examine the distributions of OTUs' count. Specifically, how the model with (3.1)-(3.3) and (3.7) accommodates the dependence between OTU counts, excess zeros and large between-sample variability is illustrated with various examples. The moments such as expectation and correlation of count vectors are also derived, further illustrating the interpretation of Σ in terms of counts.

3.2.2 Prior Calibration and Posterior Computation

The prior of Σ in (3.4) 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 (3.4), 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 § 3.3 and § 3.4. 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 Appendix § B.6 for sensitivity analyses related to the real data analysis in

§ 3.4.

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

$$p(\boldsymbol{\theta}, \mathbf{y}^{\star}, \boldsymbol{\eta} \mid \mathbf{y}, \mathbf{x}) \propto \prod_{i=1}^{N} \prod_{m=1}^{M} \prod_{j=1}^{J_m} p(y_{imj} \leq y_{imj}^{\star} < y_{imj} + 1 \mid \boldsymbol{\eta}_i, \boldsymbol{\theta}) \prod_{i=1}^{N} p(\boldsymbol{\eta}_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}). \quad (3.10)$$

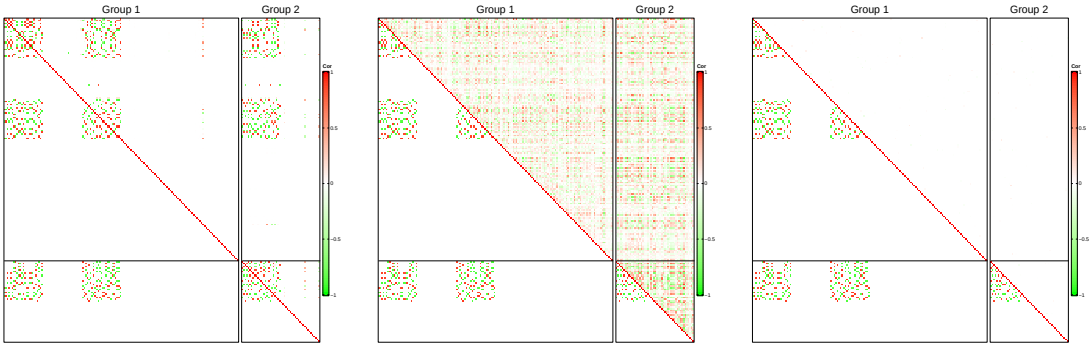
We further augment the model by introducing latent variables to facilitate updates of $\mathbf{r}_i$, $\boldsymbol{\alpha}_{s_i}$, and ζ_{mkj}. We use the blocked Gibbs sampling algorithm (Ishwaran and James, 2001) by considering a finite-dimensional truncation of the stick-breaking processes in (3.7) and (3.8). 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 Appendix § B.3. Details of the reproducing code are in Appendix § B.4.

3.3 Simulation

3.3.1 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 $\text{N}(0, 1)$ 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 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. The correlation matrix corresponding to Σ^{tr} is illustrated in the lower triangle of Fig 3.4(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 § 3.2.2. 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

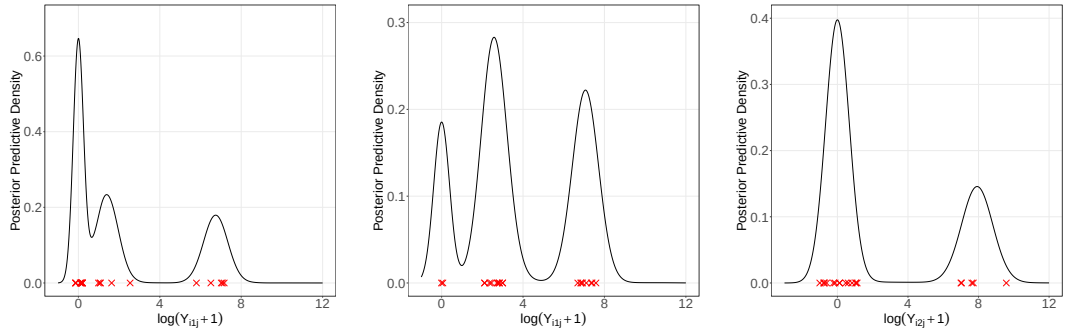


(a) Sp-BGFM (b) MOFA (c) SPIEC-EASI

Figure 3.4: [Simulation 1] 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.

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 $\rho_{jj'}^{mm'} = \Sigma_{jj'}^{mm'}/(\Sigma_{jj}^{mm}\Sigma_{j'j'}^{m'm'})$ instead of Σ. Fig 3.4 (a) compares posterior median estimates $\hat{\rho}_{jj'}^{mm'}$ of correlations to their truth. As shown in the figure, Sp-BGFM capably identifies zeroes in the truth and efficiently shrinks 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 3.5 illustrates marginal predictive distribution estimates



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

Figure 3.5: [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}.

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 methods. Prior to fitting, the OTU counts were clr-transformed and

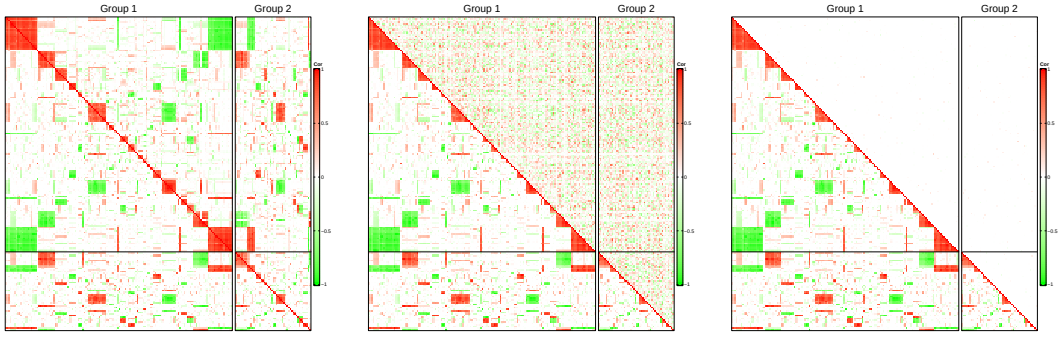
Method	Sim 1	Sim 2	Sim 3	Sim 4	Sim 5
Sp-BGFM	**0.031**	**0.109**	**0.232**	**0.000**	**0.124**
MOFA	0.229	0.364	0.316	0.107	0.235
SPIEC-EASI	0.150	0.306	0.306	0.004	0.205

Table 3.1: 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.

re-centered with default settings in the packages. Their correlation estimates $\hat{\rho}_{jj'}^{mm'}$ are compared to the truth in Fig 3.4 (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 3.1. 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 Appendix § B.5.1. Comparing their estimates to the truth, those alternative methods perform poorly in uncovering the true dependence among the OTUs.

3.3.2 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 3.6(a). The OTUs are rearranged within a group for a better illustration. For abundances, we generated $\alpha_{s_i mj}^{\text{tr}}$ and r_{im}^{tr} similarly

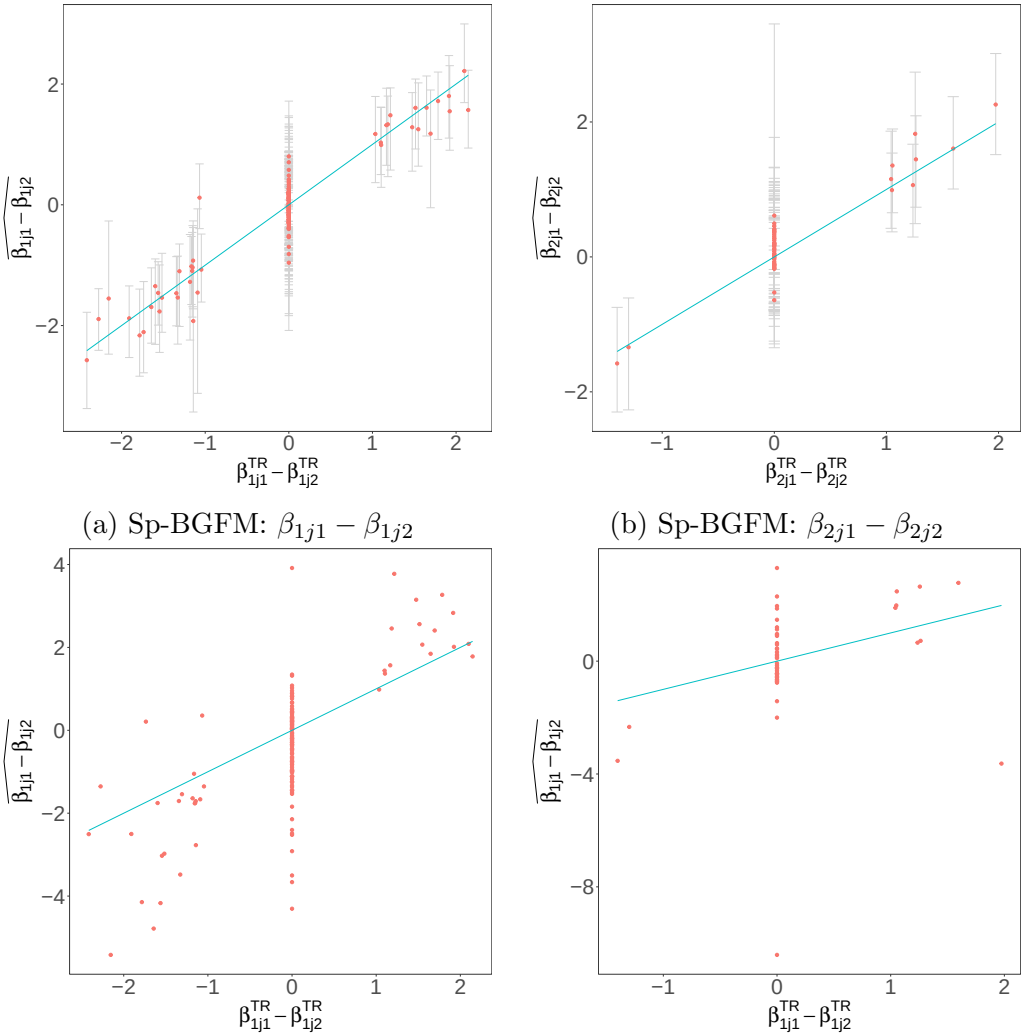


(a) Sp-BGFM (b) MOFA (c) SPIEC-EASI

Figure 3.6: [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.

as in Simulation 1, but we used the empirical proportions of zero counts from the multi-domain skin microbiome dataset in § 3.4 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 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 $\mu_{imj}^{\text{tr}} = r_{im}^{\text{tr}} + \alpha_{s_imj}^{\text{tr}} + \mathbf{x}_i'\boldsymbol{\beta}_{mj}^{\text{tr}}$ 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 the simulation setup are in Appendix § B.5.2.

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



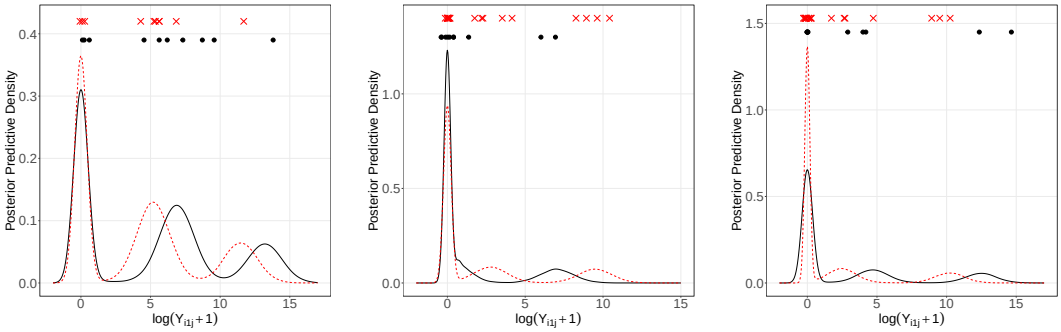
(a) Sp-BGFM: $\beta_{1j1} - \beta_{1j2}$ (b) Sp-BGFM: $\beta_{2j1} - \beta_{2j2}$

(c) metagenomeSeq: $\beta_{1j1} - \beta_{1j2}$ (d) metagenomeSeq: $\beta_{2j1} - \beta_{2j2}$

Figure 3.7: [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.

inferences. The trace plots demonstrated a good mixing of the MCMC chain.

The upper triangle of Fig 3.6(a) illustrates the posterior estimates $\hat{\rho}^{mm'}_{jj'}$ under Sp-BGFM. Figs 3.7(a) and (b) show the posterior median estimates of $\beta_{mj1} - \beta_{mj2}$



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

Figure 3.8: [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.

(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.

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

Figs 3.6(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 3.1. The estimates from the additional comparators, REBACCA, COAT and Zi-LN, are shown in Appendix Fig B.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 3.7(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.

3.3.3 Additional Simulations

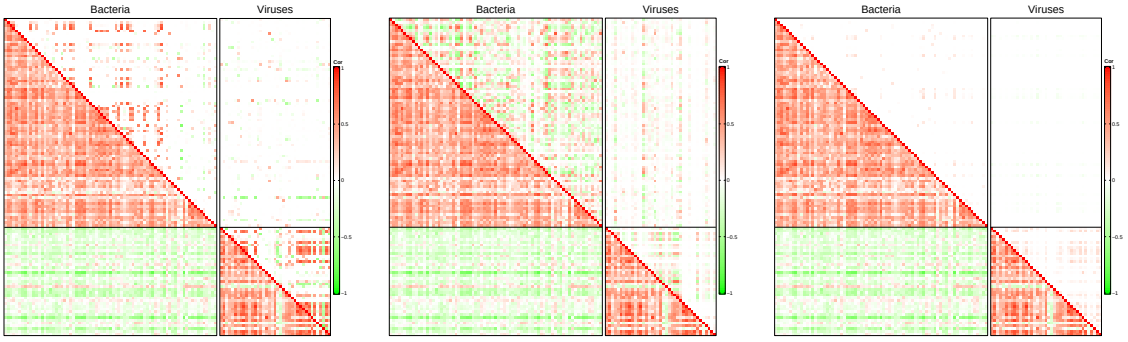
We conducted additional simulation studies, Simulations 3, 4, and 5, to further examine the robustness of Sp-BGFM. In Simulation 3, we kept the setup of Simulation 2 and used Σ^{tr} arbitrarily specified by the vine method in Lewandowski et al. (2009) to generate data. However, no covariate was considered. Sp-BGFM recovers the true microbial interaction structure well, as shown in Appendix Fig B.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 § 3.4. The true OTU dependence structure is well recovered under Sp-BGFM, as shown in Appendix Fig B.10. Especially, Appendix Fig B.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 § 3.4 and correlation matrices randomly generated by the vine method. Appendix Fig B.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 3.1. Details of Simulations 3-5 are reported in §4.3-§4.5 of Appendix B, respectively.

3.4 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 3.9(a). We used $K = 15$, and all other hyperparameters were specified at the same values as in the simulation studies of § 3.3. 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



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

Figure 3.9: [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.

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 Appendix § B.6.

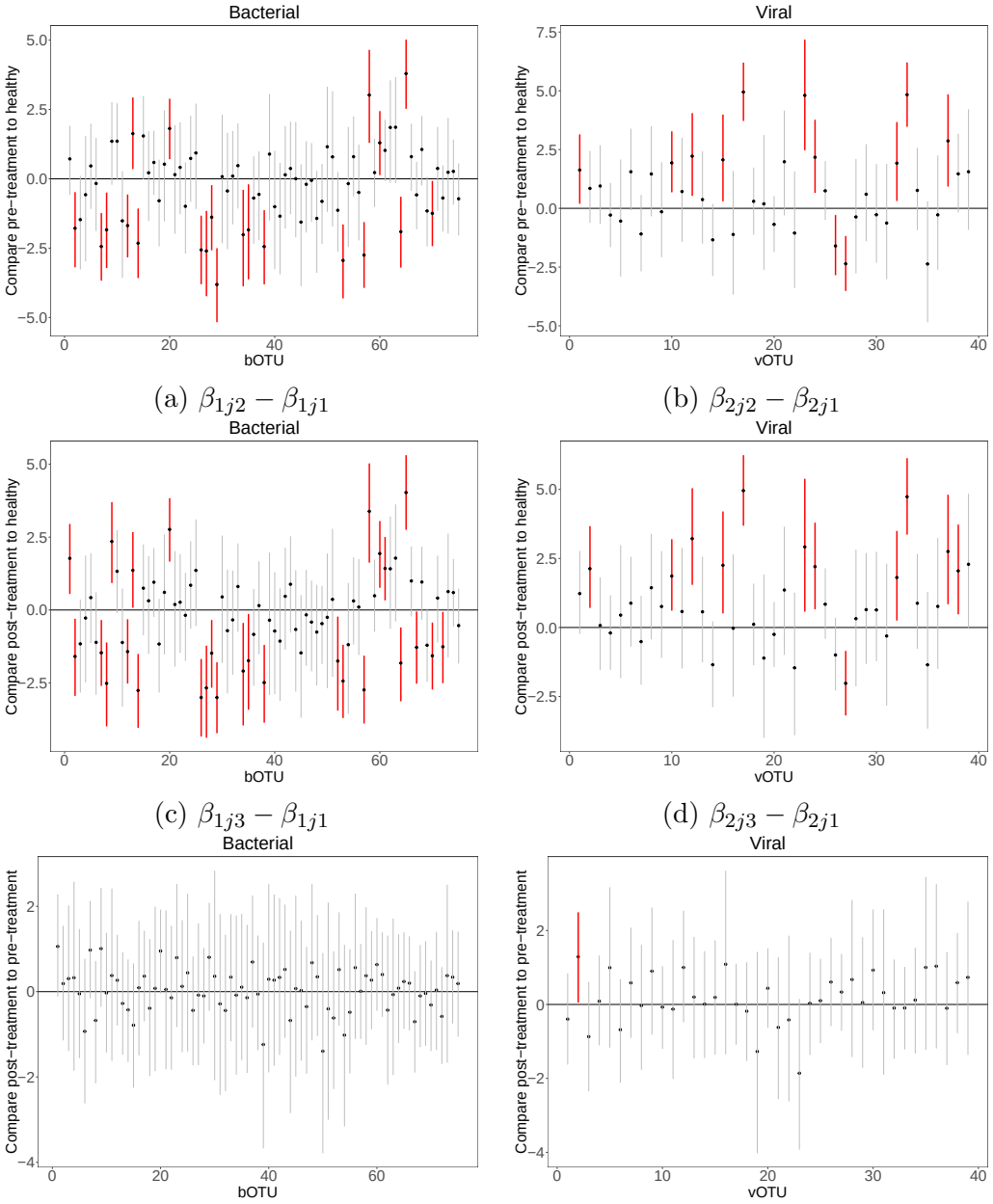
The upper right triangle of Fig 3.9(a) illustrates posterior median estimates $\hat{\rho}_{jj'}^{mm'}$ of correlations. The OTUs are rearranged within a group for a better illustration. Appendix Fig B.17 illustrates $\hat{\rho}_{jj'}^{mm'}$ for the OTUs that have $|\hat{\rho}_{jj'}^{mm'}| > 0.5$ with any other OTU j', $j' \neq j$. Appendix Tabs B.2 and B.3 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 $\hat{\rho}_{jj'}^{mm} \geq 0.71$, $m = 1$. 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 $\hat{\rho}_{jj'}^{mm} \geq 0.61$, $m = 2$ 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 3.9(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 3.10 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



(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 3.10: [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.

zero are in red. Appendix Tabs B.2 and B.3 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.

Appendix Fig B.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 3.9(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 Appendix Fig B.9. Appendix Fig B.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.

3.5 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 (3.2) 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^{α}_{ml} and/or $\xi^{\star}_{mjl}$ of G_{mj} in (3.2) 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.

Chapter 4

Bayesian Covariate-Assisted Interaction Analysis for Multivariate Count Data in Microbiome Study

4.1 Introduction

Covariance estimation is a fundamental task in multivariate statistical analysis, critical for understanding the relationships between variables. Covariance matrices are pivotal in various applications, including principal component analysis (Pearson, 1901), factor analysis (Rummel, 1988), and canonical-correlation analysis (Hotelling, 1992). Traditional methods for covariance estimation, such as the sample covariance matrix, typically assume that the data is identically and independently distributed (i.i.d.). However, this assumption is often violated in real-world scenarios, where data often exhibits heteroscedasticity and covariance changes with external factors, such as

covariates. Ignoring these covariate dependencies can result in inaccurate models and misleading conclusions, necessitating the development of methods that account for these dependencies to provide more reliable estimates of the interrelationships between variables.

Covariance regression has gained significant attention over a long history due to its ability to incorporate covariate information, thereby enhancing the accuracy and interpretability of covariance estimates. Carroll and Ruppert (1982) first considered a heteroscedastic model in which the variances were given by a parametric function of the mean. A linear model for the standard deviation(Rutemiller and Bowers, 1968) and a generalized model with a link function to allow non-negativity of variance (Smyth, 1989) were developed for uni-variate cases. When it comes to multivariate heteroscedasticity, Chiu et al. (1996), Pourahmadi (2011) and Battey (2017) modeled the logarithm of elements of the covariance matrix as a linear function of known matrices to guarantee the positive definiteness of the covariance matrix. However, it is difficult to interpret parameters of covariate effects in the log scale and the number of parameters can be quite large in high-dimensional data. More recently, sparse and low-rank methods for covariate-dependent covariance estimation or its inversion (precision matrix) have been considered to manage high-dimensional data where traditional methods are often inadequate. These approaches leverage structural assumptions, sparsity and low rankness, to enhance estimation accuracy and interpretability. Pourahmadi (1999) modeled the unconstrained elements of the Cholesky decomposition on the precision matrix and linked covariates to elements. Hoff and Niu (2012) expressed the covariance as a baseline

covariance matrix plus a rank-1 positive definite matrix which depends on covariates. They further extended to allow the deviation of each covariate-dependent covariance from the baseline to be any rank. Fox and Dunson (2015) put a Gaussian process prior on the latent factor model and induced a flexible Bayesian nonparametric covariance regression model. The predictor-dependent framework was characterized as a combination of Gaussian process random functions of covariates. Ni et al. (2019) proposed a graphical regression method that estimates directed acyclic graphs for the precision matrix in heterogenous data with additional subject-level covariates. Niu et al. (2023) further modeled continuously varying undirected graphs with additional assistance from any general covariates for underlying heterogeneous multivariate observations.

Besides modeling the covariance matrix with covariates, joint modeling for means and covariances allows for the simultaneous exploration of covariate effects on the mean and the covariance of the data. Pourahmadi (1999) provided a joint mean-covariance model with applications to longitudinal data. In the context of temporal heteroscedasticity, Fong et al. (2006) studied multivariate autoregressive conditionally heteroscedastic (ARCH) models in the financial data. Niu and Hoff (2019) extended their model in Hoff and Niu (2012) to a joint mean and covariance model, studying the covariate effects on both mean and covariance in the application of multiple health outcome measures. Moran et al. (2021) used a parametric covariance regression model to analyze verbal autopsy data. It was designed specifically for cause of death denoted covariance. However, the above approaches are built for continuous data, and they can be inappropriate for analyzing multivariate count data. With the advent of

high-throughput sequencing (HTS) sequencing technologies, multivariate count tables arise in various biological applications. Especially in microbiome studies, 16S ribosomal RNA (16S rRNA) sequencing uses similarity-based clustering algorithms to group 16S rRNA sequences into Operational Taxonomic Units (OTUs), producing multivariate count tables for downstream analysis. Analyzing OTU tables and detecting the structure of microbial interactions is essential for more accurately characterizing microbial communities. Popular methods in microbiome studies such as SparCC (Friedman and Alm, 2012), CCLasso (Fang et al., 2015) and SPIEC-EASI (Kurtz et al., 2015) adopt log-transformed counts or log-transformed ratio for analysis of interactions. Specifically, SparCC (Friedman and Alm, 2012) adds pseudo counts and then divides the raw counts by the sample's total counts for normalization. It models log-transformed ratios of these normalized counts to infer correlations between OTUs through sparse networks. Similarly, CCLasso in Fang et al. (2015) uses ℓ_1 penalty to estimate the correlation network of log-transformed counts. SPIEC-EASI (Kurtz et al., 2015) uses graphical lasso (Friedman et al., 2008), a popular penalized method outputting the association of undirected graphs, to obtain a robust precision matrix estimate. The raw OTU counts are also first centered by log-ratio (clr) transformation. See REBECCA (Ban et al., 2015), COAT (Cao et al., 2019) and MOFA (Argelaguet et al., 2018) for more. However, most methodologies above assume the mean centered at 0 or simply subtract the sample mean. In addition, those covariance estimates remain the same across any covariates.

To circumvent the challenges described above and address the effect of covari-

ates on microbial interactions, we propose a Bayesian covariate-dependent sparse factor model with a rounded kernel. The model assesses interrelationships between OTUs varying as a function of covariates. Furthermore, it simultaneously performs model-based normalization and flexibly accommodates large variability in count data. Specifically, we use nonparametric mixtures of rounded multivariate log-normal kernels to introduce latent continuous random variables. Then, the covariance matrix characterizes interrelationships among OTUs and varies with covariates. We adopt the low-rank structure factor model and induce a Dir-HS prior (Zhang et al., 2024) on the factor loading matrix to effectively learn a high-dimensional covariance structure despite a limited sample size. We further link the covariate as a multiplicative effect to the factor loading matrix, letting the covariance vary with general covariates. The parametric formulation in the covariate-covariance structure significantly reduces the number of parameters to estimate and offers a straightforward interpretation of the interrelationship structure. Moreover, we use a Dirichlet process (DP) prior on relative abundances to obtain a flexible joint distribution of count vectors. It induces an infinite Dirichlet process mixture (DPM) on the count distribution, a flexible mean formulation handling excess zeros and overdispersion in microbiome data. We also relate covariates to the mean to detect different OTU abundances under covariates.

In the rest of the section, we describe the model and its applications. § 4.2 and § 4.3 describe the covariate-dependent rounded multivariate log-normal kernel model, its prior specification and posterior computation. § 4.4 shows the results of simulation studies to evaluate the performance of our method. § 4.5 has results from the model

applied to the real dataset, and § 4.6 concludes with some discussion of the results and areas of future research.

4.2 Model and Prior Specification

In this section, we first introduce a Bayesian sparse factor model and let factors vary with covariates and obtain estimates of covariate-dependent covariance matrices. Then, we build a rounded kernel model that exploits a Bayesian nonparametric approach to induce a flexible joint distribution for multivariate count responses.

4.2.1 Sparse Covariate-dependent Factor Model

Let $\tilde{\mathbf{Y}}_i^\star = (\tilde{Y}_{i1}^\star, \ldots, \tilde{Y}_{iJ}^\star) \in \mathbf{R}^J$ be a J-dimensional normal vector taken from sample i,

$$\tilde{\mathbf{Y}}_i^\star \mid \boldsymbol{\mu}(\mathbf{x}_i), \Sigma(\mathbf{x}_i) \overset{indep}{\sim} \mathrm{N}_J(\boldsymbol{\mu}(\mathbf{x}_i), \Sigma(\mathbf{x}_i)), \quad i = 1, \ldots, N, \tag{4.1}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{iP})$ is a P-dimensional covariates of a sample. We first build a prior probability model for the covariate-dependent covariance $\Sigma(\mathbf{x}_i)$, the main parameter of interest. We will discuss a model for $\boldsymbol{\mu}(\mathbf{x}_i)$ later in § 4.2.2. In the high-dimensional setting, the number of features J is large compared to the sample size N, and sample covariance estimates of a large covariance matrix are usually unstable. Introducing a sparse structure of covariance matrix mitigates the curse of dimensionality, allows for more efficient and interpretable models and leads to more robust statistical inferences

(Cai et al., 2015; Xie et al., 2022). Following this vein, we first decompose $\Sigma(\mathbf{x}_i)$ as

$$\Sigma(\mathbf{x}_i) = \Lambda(\mathbf{x}_i)\Lambda'(\mathbf{x}_i) + \sigma^2 \mathrm{I}_J, \tag{4.2}$$

where $\boldsymbol{\lambda}_j(\mathbf{x}_i) = [\lambda_{j1}(\mathbf{x}_i), \ldots, \lambda_{jK}(\mathbf{x}_i)]'$ and $\Lambda(\mathbf{x}_i) = [\boldsymbol{\lambda}_1(\mathbf{x}_i)', \ldots, \boldsymbol{\lambda}_J(\mathbf{x}_i)']'$ is a $J \times K$ covariate-dependent factor loading matrix. Here, K is the dimension of the subspace that is assumed to capture statistical variability, and typically $K \ll J$. Similar to Bhattacharya and Dunson (2011) and Xie et al. (2018), we do not impose any constraints on $\Lambda(\mathbf{x}_i)$, such as column orthogonality, nor do we seek to interpret latent factors, as our primary focus is on the inference of $\Sigma(\mathbf{x}_i)$. For each factor loading element $\lambda_{jk}(\mathbf{x}_i)$, we further express as

$$\lambda_{jk}(\mathbf{x}_i) = q_{jk}\boldsymbol{f}_k'\mathbf{x}_i, \tag{4.3}$$

where q_{jk} is a local multiplicative effect and P-dimensional random vector $\boldsymbol{f}_k$ is used to have a column-wise covariate-dependent multiplicative effect. Note that q_{jk}'s are OTU and factor specific, and $\boldsymbol{f}_k$ is common for all OTUs. When the local effect q_{jk} is close to 0, the corresponding λ_{jk} becomes small. When f_{kp} is small for all k, $\Sigma(\mathbf{x}_i)$ does not vary much with x_{ip}. Under (4.2) and (4.3), the covariance between OTUs j and j' is a quadratic function of covariates;

$$\Sigma_{jj'}(\mathbf{x}_i) = \begin{cases} \sum_{k=1}^{K} q_{jk}q_{j'k}(\boldsymbol{f}_k'\mathbf{x}_i)^2, & \text{if } j \neq j', \\ \sum_{k=1}^{K} q_{jk}^2(\boldsymbol{f}_k'\mathbf{x}_i)^2 + \sigma^2, & \text{if } j = j'. \end{cases} \tag{4.4}$$

Similar to Hoff and Niu (2012) and Niu and Hoff (2019), including the intercept 1 in $\mathbf{x}_i$ alleviates the constraint that the minimum of covariance is always obtained at $x_{ip} = 0, p > 1$. Instead of directly estimating $J(J+1)/2$ parameters of the covariance matrix for each $\mathbf{x}_i$, we have $JK + KP$ parameters in (4.3). It yields a significant reduction of unknown parameters to estimate, which is crucial for a high-dimensional setting. We introduce joint sparsity through columns $\boldsymbol{q}_k = (q_{1k}, \ldots, q_{Jk})$ by considering the Dirichlet-Horseshoe (Dir-HS) prior in Zhang et al. (2024), for $k = 1, \cdots, K$,

$$\begin{aligned} \tau_k \mid a_\tau, b_\tau &\overset{iid}{\sim} \mathrm{Ga}(a_\tau, b_\tau/J), \\ \boldsymbol{\phi}_k = (\phi_{1k}, \ldots, \phi_{Jk}) \mid a_\phi &\overset{iid}{\sim} \mathrm{Dir}(a_\phi, \ldots, a_\phi), \\ \zeta_{jk} &\overset{iid}{\sim} \mathrm{C}^+(0,1), \\ q_{jk} \mid \phi_{jk}, \tau_k, \zeta_{jk} &\overset{indep}{\sim} \mathrm{N}(0, \zeta_{jk}^2 \phi_{jk} \tau_k). \end{aligned} \tag{4.5}$$

where $\mathrm{C}^+(0,1)$ is the half-Cauchy distribution for $\mathbb{R}^+$ with location and scale parameters 0 and 1, and $\mathrm{Ga}(a,b)$ represents the gamma distribution with mean a/b. Under the model in (4.5), ϕ_{jk} locally shrinks q_{jk} toward zero and in return shrinks $\lambda_{jk}(x_i)$. On the other hand, τ_k controls the global shrinkage for each factor and performs an effective truncation of the number of latent factors. The joint sparsity assumption leads to obtaining a reliable estimate of the structure with a small sample size and achieving good theoretical properties (Cai et al., 2015; Xie et al., 2018). To complete the prior specification, we assume a conditionally conjugate prior on σ^2, $\sigma^2 \sim \text{inv-Ga}(a_\sigma, b_\sigma)$ with fixed a_σ and b_σ and a standard normal prior on $f_{kp} \overset{iid}{\sim} \mathrm{N}(0,1)$.

A nonparametric model can also be considered by expressing $\lambda_{jk}(\mathbf{x}_i)$ in (4.3) as a weighted combination of a set of basis functions from a Gaussian process (GP) prior (Fox and Dunson, 2015). While a nonparametric model using the GP can be more flexible, the computation with many GPs becomes complicated when the number of OTUs increases. Moran et al. (2021) further adopts a similar expression with parametric basis functions, and the covariance regression model in Hoff and Niu (2012) can also be written into a covariate-dependent factor model. Different from these methods, our proposed model includes OTU and factor-specific multiplicative effect with joint sparsity, efficiently reducing the number of parameters and having a straightforward interpretation of coefficients. We present later in Simulation 2 that our parametric design can flexibly capture arbitrary random covariance matrices while parsimonious.

4.2.2 A Nonparametric Model for Mean

Next, we build a Bayesian nonparametric mixture model with a rounded kernel to obtain a flexible multivariate count distribution. Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})$ be the observed J-dimensional random count vector of counts of sample $i = 1, \ldots, N$, where $Y_{ij} \in \mathbb{N}_0$ is the count of feature $j = 1, \ldots, J$ in sample i. Noting the relationship between normal distribution and log-normal distribution, we introduce a latent multivariate log-normal vector $\mathbf{Y}_i^\star = \exp(\tilde{\mathbf{Y}}_i^\star) \in \mathbf{R}_+^J$ by considering the rounded kernel approach for count data in Canale and Dunson (2011) and assume

$$\mathbf{Y}_i^\star \mid \boldsymbol{\mu}(\mathbf{x}_i), \Sigma(\mathbf{x}_i) \overset{indep}{\sim} \text{log-N}_J(\mathbf{y}^\star \mid \boldsymbol{\mu}(\mathbf{x}_i), \Sigma(\mathbf{x}_i)), \quad i = 1, \ldots, N. \tag{4.6}$$

where parameters $\boldsymbol{\mu}(\mathbf{x}_i) = (\mu_{i1}, \ldots, \mu_{iJ})' \in \mathbb{R}^J$ and $\Sigma(\mathbf{x}_i) > 0$. The multivariate log-normal density is zero for a vector with negative values, and the kernel defines a valid multivariate count distribution for $\mathbf{Y}_i$ as follows;

$$\mathrm{P}(\mathbf{Y}_i = \mathbf{y} \mid \boldsymbol{\mu}(\mathbf{x}_i), \Sigma(\mathbf{x}_i)) = \int_{A(\mathbf{y})} f_{\mathbf{y}^\star}(\mathbf{y}^\star \mid \boldsymbol{\mu}(\mathbf{x}_i), \Sigma(\mathbf{x}_i)) d\mathbf{y}^\star, \tag{4.7}$$

where the region of integration $A(\mathbf{y}) = \{\mathbf{y}^\star \mid y_1 \leq y_1^\star < y_1+1, \ldots, y_J \leq y_J^\star < y_J+1\}$ and $f_{\mathbf{y}^\star}(\cdot)$ is a pdf of a $J-$dimensional log-normal distribution with parameters $\boldsymbol{\mu}(\mathbf{x}_i)$ and $\Sigma(\mathbf{x}_i)$. In (4.6), $\exp(\mu_{ij})$ is the median of $Y_{ij}^\star$ representing the location of the distribution. Larger μ_{ij} thus implies a large abundance of feature j in sample i. We consider the model in § 4.2.1 for $\Sigma_{jj'}(\mathbf{x}_i)$. We have the mean $\mathrm{E}(Y_{ij}^\star) = \exp(\mu_{ij} + \frac{1}{2}\Sigma_{jj}(\mathbf{x}_i))$ and $\mathrm{Cov}(Y_{ij}^\star, Y_{ij'}^\star) = \mathrm{E}(Y_{ij}^\star)\mathrm{E}(Y_{ij'}^\star)\left\{\exp(\Sigma_{jj'}(\mathbf{x}_i)) - 1\right\}$. When $\Sigma_{jj'}(\mathbf{x}_i) = 0$, it implies there are no microbial interactions between features. In terms of the count distributions, the mean and covariance of $\mathbf{Y}_i$ can be easily verified finite and computed through probability mass function defined in (4.7).

We relate μ_{ij} to covariates $\mathbf{x}_i$ through regression;

$$\mu_{ij}(\mathbf{x}_i) = r_i + \alpha_j + \tilde{\mathbf{x}}_i'\boldsymbol{\beta}_j. \tag{4.8}$$

r_i is the sample (library) size factor, normalizing counts across samples. α_j represents the normalized baseline abundance of feature $j = 1, \cdots, J$ for all samples, and $\tilde{\mathbf{x}}_i$ omits the intercept in $\mathbf{x}_i$ due to identifiability. Regression coefficients β_{jp} quantify the change

in the abundance of feature j from its baseline abundance α_j by covariates x_{ip}. For example, we have different experimental conditions for $\tilde{\mathbf{x}}_i$ in the mice gut microbiome data in § 4.5. β_{jp} measures the change in the abundance of OTU j by experimental condition with x_{ip}. We first impose a Bayesian nonparametric prior model on α_j to accommodate large variability. An adequately flexible mean model further improves the estimation of $\Sigma(\mathbf{x}_i)$. While $\boldsymbol{\beta}_j$ are identifiable, r_i and α_j in μ_{ij} are not identifiable due to the multiplicative structure, $\mathrm{E}(\log(y^\star_{ij}) \mid r_i, \alpha_j) = r_i + \alpha_j$. To address the issues, we assume a mean-constrained Dirichlet process for the prior of α_j as follows,

$$\alpha_j \mid G \overset{iid}{\sim} G = \sum_{l=1}^{\infty} \psi_l^\alpha \left\{ \omega_l^\alpha \delta_{\xi_l^\alpha} + (1-\omega_l^\alpha) \delta_{\left(\frac{\nu^\alpha - \omega_l^\alpha \xi_l^\alpha}{1-\omega_l^\alpha}\right)} \right\}, \tag{4.9}$$

where δ_ξ is a point mass centered at ξ. We let $\xi_l^\alpha \mid \nu^\alpha, u_\alpha^2 \overset{iid}{\sim} \mathrm{N}(\nu^\alpha, u_\alpha^2)$ with fixed ν^α and u_α^2. The mixture weights ψ_l^α in (4.9) are constructed using a stick-breaking process (Sethuraman, 1994); let $\psi_1^\alpha = V_1^\alpha$ and $\psi_l^\alpha = V_l^\alpha \prod_{l'=1}^{l-1}(1 - V_{l'}^\alpha)$, $l > 1$ with $V_l^\alpha \mid c^\alpha \overset{iid}{\sim} \mathrm{Be}(1, c^\alpha)$, where the concentration parameter c^α is fixed. Assume inner mixture weights $\omega_l^\alpha \mid a_\omega^\alpha, b_\omega^\alpha \overset{iid}{\sim} \mathrm{Be}(a_\omega^\alpha, b_\omega^\alpha)$, where a_ω^α and b_ω^α are fixed. Under (4.9), the prior and posterior means of α_j are fixed at ν^α, and $\mathrm{E}(\log(y^\star_{ij}) \mid G, r_i)$ fixed at $\nu^\alpha + r_i$. We will impose a similar constraint on the prior of r_i below to achieve soft identifiability. Shuler et al. (2021a) and Zhang et al. (2024) showed that overall means can be well estimated under the mean-constrained prior and their posterior inference is not sensitive to the choice of ν^α and u_α^2. (4.6) and (4.9) lead to a Dirichlet process

mixture model for $Y_{ij}^{\star}$,

$$\mathbf{Y}_i^{\star} \mid \boldsymbol{\mu}(\tilde{\mathbf{x}}_i), \Sigma(\mathbf{x}_i) \overset{indep}{\sim} \int \text{log-N}_J(\mathbf{y}^{\star} \mid r_i \mathbf{1}_J + \boldsymbol{\alpha} + \boldsymbol{\beta}\tilde{\mathbf{x}}_i', \Sigma(\mathbf{x}_i)) dG(\alpha), \quad (4.10)$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_J]'$ and $\boldsymbol{\beta}$ is a $J \times P$ coefficient matrix with $\boldsymbol{\beta}_j$ in the rows. With random mixture weights, ω_l^{α} and ψ_l^{α}, and random locations ξ_l^{α} in $G(\alpha)$, the mixture model in (4.10) can flexibly capture various shapes of a distribution and accommodate variability in the count distribution. We also consider an extension of the model in (4.8)-(4.10) to accommodate inter-subject heterogeneity. We illustrate it in Simulations 2 and 3 in detail.

Similar to (4.9), we consider a flexible infinite mixture model for r_i;

$$r_i \mid \psi_l^r, \omega_l^r \overset{iid}{\sim} \sum_{l=1}^{\infty} \psi_l^r \left\{ \omega_l^r \text{N}(\xi_l^r, u_r^2) + (1 - \omega_l^r) \text{N}\left(\frac{\nu^r - \omega_l^r \xi_l^r}{1 - \omega_l^r}, u_r^2 \right) \right\}, \quad (4.11)$$

where ν^r and u_r^2 are fixed. The prior expectation of r_i is ν^r in (4.11), and $\text{E}(\log(y_{ij}^{\star}) \mid G, r_i)$ fixed at $\nu^{\alpha} + \nu^r$ from (4.9) and (4.11). We jointly specify values of ν^{α} and ν^r using observed counts. For example, we first fix ν^r at the average of the logarithm of the total count, $\nu^r = \frac{1}{N} \sum_{i=1}^{N} \log\left(\sum_{j=1}^{J} y_{ij} \right)$, and set $\nu^{\alpha} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{J} \{\log(y_{ij} + 0.01) - \nu^r\}$. We consider similar following priors for ψ_l^r, ω_l^r and ξ_l^r; assume $\xi_l^r \mid \nu^r, u_{\xi^r}^2 \overset{iid}{\sim} \text{N}(\nu^r, u_{\xi^r}^2)$, $\omega_l^r \mid a_{\omega}^r, b_{\omega}^r \overset{iid}{\sim} \text{Be}(a_{\omega}^r, b_{\omega}^r)$, $\psi_1^r = V_1^r$ and $\psi_l^r = V_l^r \prod_{l'=1}^{l-1}(1 - V_{l'}^r)$, $l > 1$, where $V_l^r \mid c^r \overset{iid}{\sim} \text{Be}(1, c^r)$. Here, $u_{\xi^r}^2$, a_{ω}^r, b_{ω}^r, and c^r are fixed. To complete the prior specification of $\boldsymbol{\beta}$, we consider a conjugate standard normal distribution for $\beta_{jp} \overset{iid}{\sim} \text{N}(0, 1)$.

4.3 Prior Calibration and Posterior Computation

The prior of $\Sigma(\mathbf{x}_i)$ in (4.5) requires the specification of the hyperparameters K, a_ϕ, a_τ and b_τ. Selecting the dimension of the latent space K can be challenging. K determines the number of parameters, and a model with a random K requires complicated algorithms such as reversible jump MCMC (Green and Hastie, 2009) for posterior simulation. We set K to a reasonably large value for computational convenience. Empirically, we determine K by performing principle component analysis (PCA) on the sample covariance matrix of log-transformed normalized counts and fixing K such that the K largest eigenvalues explain 95% of the total variance. With a sufficiently large K, the model can let some τ_k small for redundant latent factors. If desired, a geometric or truncated Poisson distribution can be placed on K to achieve an optimal posterior contraction rate (Pati et al., 2014). In terms of the hyperparameters a_ϕ, a_τ and b_τ, we follow the setup in Zhang et al. (2024) and let $a_\phi = 1/(0.2 \times J)$, $a_\tau = 0.1$ and $b_\tau = 1/J$. From simulation studies, we observed that a too small value of a_ϕ tends to overly shrink q_{jk} toward zero, resulting in a poor estimate of $\Sigma(\mathbf{x}_i)$. We also examined sensitivity to the specifications of those hyper-parameters and found that the model's performance remains robust within a reasonable range of those values.

Let $\boldsymbol{\theta} = \{q_{jk}, \phi_{jk}, \tau_k, \zeta_{jk}, f_{kp}, \sigma^2, \alpha_j, \omega_l^\alpha, V_l^\alpha, \xi_{jl}^\alpha, r_i, \omega_l^r, V_l^r, \xi_l^r, \beta_{jp}\}$ a vector of all random parameters. We use Markov Chain Monte Carlo (MCMC) to sample $\boldsymbol{\theta}$ from their posterior distributions. We introduce a latent normal vector $\boldsymbol{\eta}_i \overset{iid}{\sim} \mathrm{N}_K(0, \mathrm{I}_K)$. We then have $Y_{ij}^\star \mid \mu_{ij}(\mathbf{x}_i), \boldsymbol{\lambda}_j(\mathbf{x}_i), \boldsymbol{\eta}_i, \sigma^2 \overset{indep}{\sim} \text{log-N}(\mu_{ij}(\mathbf{x}_i) + \boldsymbol{\lambda}_j'(\mathbf{x}_i)\boldsymbol{\eta}_i, \sigma^2)$ as independent

log-normal variables, which results in significant computational efficiency. The joint posterior distribution of the augmented model is

$$p(\boldsymbol{\theta}, \mathbf{Y}^{\star}, \boldsymbol{\eta} \mid \mathbf{y}, \mathbf{x}) \propto \prod_{i=1}^{N} \prod_{j=1}^{J} p(y_{ij} \leq Y_{ij}^{\star} < y_{ij} + 1 \mid \boldsymbol{\eta}_i, \boldsymbol{\theta}) \prod_{i=1}^{N} p(\boldsymbol{\eta}_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}). \tag{4.12}$$

We use the blocked Gibbs sampling algorithm (Ishwaran and James, 2001) by considering a finite-dimensional truncation of the stick-breaking processes in (4.9) and (4.11). We set the truncation levels L^r and L^{α} to sufficiently large values. Given the latent variables, all parameters except $\boldsymbol{\phi}_k$ can be updated through Gibbs steps. Although f_{kp} has a conjugate full conditional distribution, we found the mixing over f_{kp} could be poor, and used an adaptive MH algorithm (Haario et al., 2001) for an efficient update of $\boldsymbol{\phi}_k$ and f_{kp}. Details of the MCMC algorithm are in Supp. § C.1. The reproducing codes of the proposed model are on https://github.com/shuang-jie/BCAIA. The instructions are on the ReadMe page.

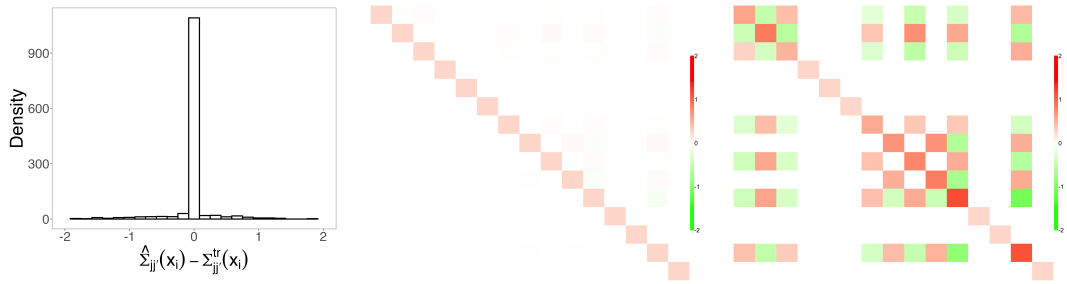
4.4 Simulation Studies

4.4.1 Simulation 1

For Simulation 1, we considered a case mimicking the mice gut microbiome dataset in § 4.5. We include two categorical variables, one with two levels $x_{i1} \in \{0, 1\}$ and the other with three levels $x_{i2} \in \{(0, 0), (0, 1), (1, 0)\}$, and we have $\mathbf{x}_i = (1, x_{i1}, x_{i2})$. We assumed five samples with $J = 15$ OTUs under each of the six conditions, and had

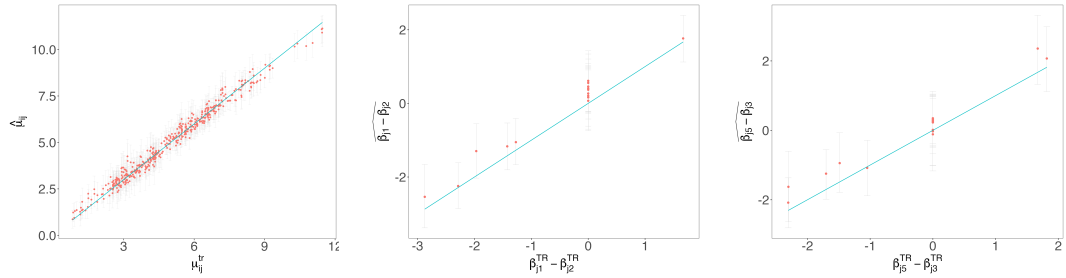
$N = 30$ samples in total. To specify $\Sigma^{\text{tr}}(\mathbf{x}_i)$, we let $K^{\text{tr}} = 2$. We then simulated q_{jk}^{tr} from $\text{N}(0, 1)$ with probability 0.5 and to be zero with probability 0.5. For non-zero entries, we further shifted them away from zero by 1 to have large non-zero covariance. We let $f_{kp}^{\text{tr}} \overset{iid}{\sim} \text{Unif}(-1, 1)$ and $f_{11} = -f_{13}, f_{21} = -f_{22}$, resulting in covariance matrix varying over different levels. The truth covariance matrix $\Sigma^{\text{tr}}(\mathbf{x}_i)$ is illustrated in Fig 4.1(b)-(c). For the mean count abundance, we set $r_i^{\text{tr}} \overset{iid}{\sim} \text{Unif}(0, 2)$ and $\alpha_j^{\text{tr}} \overset{iid}{\sim} 0.3\text{N}(2.5, 0.5^2) + 0.7\text{N}(5, 0.5^2)$. And we sample β_{jp}^{tr} similar to q_{jk} by having probability 0.5 to be 0 and probability 0.5 to be sampled from $\text{N}(0, 1)$ then shifting by 1. Finally, we generated $\mathbf{Y}_i^{\star,\text{tr}}$ from log-$\text{N}_J(\boldsymbol{\mu}^{\text{tr}}(\mathbf{x}_i), \Sigma^{\text{tr}}(\mathbf{x}_i))$ and obtain count vectors $\mathbf{Y}_i = \lfloor \mathbf{Y}_i^{\star,\text{tr}} \rfloor$. There are no zero counts which is the same as the mice dataset in § 4.5. We fit the model by setting the hyperparameters as discussed in § 4.3, $K = 8$, $c^r = c^\alpha = 3$, $L^r = 30, L^\alpha = 35$, $a_\sigma = b_\sigma = 3$, $a_\omega^r = b_\omega^r = a_\omega^\alpha = b_\omega^\alpha = 5$. We ran MCMC for $160,000$ iterations and discarded the first half for burn-in. It took 13 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.

Fig 4.1(a) shows a histogram of the differences $\hat{\Sigma}_{jj'}(\mathbf{x}_i) - \Sigma_{jj'}^{\text{tr}}(\mathbf{x}_i)$ of all six conditions, $j < j'$. The differences tightly centered around 0, indicates that the method provides good estimates of the covariance. We compare posterior median covariance estimates $\hat{\Sigma}_{jj'}(\mathbf{x}_i)$ for two randomly samples to their true values $\Sigma_{jj'}^{\text{tr}}(\mathbf{x}_i)$ in Fig 4.1(b)-(c). For sample 16, all OTU pairs have a small covariance in the truth, but in sample 26, some OTU pairs have a strong covariance. Our method produces reasonable estimates for the samples. Covariates of sample 16 and 26 are $\mathbf{x}_{16} = [1, 1, (0, 0)]$ and $\mathbf{x}_{26} =$



(a) $\hat{\Sigma}_{jj'} - \Sigma^{\text{tr}}_{jj'}$ (b) Sample 16 (c) Sample 26

Figure 4.1: [Simulation 1] Panel (a) has a histogram of differences between $\hat{\Sigma}_{jj'}(\mathbf{x}_i)$ and $\Sigma^{\text{tr}}_{jj'}(\mathbf{x}_i)$ of six conditions. In (b), the lower left and upper right triangles of the heatmap illustrate true values $\Sigma^{\text{tr}}(\mathbf{x}_i)$ and their posterior estimates of covariance $\hat{\Sigma}(\mathbf{x}_i)$, respectively. Samples 16 ($\mathbf{x}_{16} = [1, 1, (0, 0)]$) and 26 ($\mathbf{x}_{26} = [1, 1, (0, 1)]$) from two different levels are used for illustration.



(a) $\hat{\mu}_{ij}$ vs μ^{tr}_{ij} (b) $\widehat{\beta_{j1} - \beta_{j2}}$ vs $\beta^{\text{tr}}_{j1} - \beta^{\text{tr}}_{j2}$ (c) $\widehat{\beta_{j5} - \beta_{j3}}$ vs $\beta^{\text{tr}}_{j5} - \beta^{\text{tr}}_{j3}$

Figure 4.2: [Simulation 1] The posterior median estimate of mean abundance μ_{ij} is plotted against the truth in panels (a). Panel (b)-(c) plot the effect of the first covariate $\beta_{j1} - \beta_{j2}$ and the difference in mean between the first and the third levels of the second covariate $\beta_{j5} - \beta_{j3}$, respectively.

$[1, 1, (0, 1)]$, respectively. Fig 4.2(a) shows the posterior median estimates of mean μ_{ij} with its 95% credible interval estimates. We also check the estimates of categorical covariate effects in Figs 4.2(b)-(c). Our method estimates the mean abundance well, which provides a profound basis for good estimates of covariance.

4.4.2 Simulation 2

In Simulation 2, we extend the model in (4.9) to accommodate inter-subject variability, which is commonly present in microbiome data. Suppose we have multiple samples from a set of subjects, $\{1, \ldots, S\}$. We let s_i denote the subject from which sample i is taken, and $\alpha_{s_i j}$ the normalized abundance of OTU j in the samples form subject s_i. We replace the prior of α_j in (4.9) to with the following for $\alpha_{s_i} = [\alpha_{s_i 1}, \ldots, \alpha_{s_i J}]$,

$$\begin{gathered} \boldsymbol{\alpha}_{s_i} \mid G \overset{iid}{\sim} G(\boldsymbol{\alpha}),\, s_i \in \{1, \ldots, S\}, \\ G(\boldsymbol{\alpha}) = \prod_{j=1}^{J} G_j(\alpha_j) = \prod_{j=1}^{J} \left[\sum_{l=1}^{\infty} \psi_l^{\alpha} \left\{ \omega_l^{\alpha} \delta_{\xi_{jl}^{\alpha}} + (1 - \omega_l^{\alpha}) \delta_{\left(\frac{\nu_j^{\alpha} - \omega_l^{\alpha} \xi_{jl}^{\alpha}}{1 - \omega_l^{\alpha}} \right)} \right\} \right]. \end{gathered} \tag{4.13}$$

where the individual $G_j(\alpha_j)$ has different mixing locations ξ_{jl}^{α} for each feature.

To generate a dataset, we let $S = 25$ subjects and $J = 100$ OTUs. We introduce a continuous $x_{s_i}^c$ and generate $x_{s_i}^c \overset{iid}{\sim} \mathrm{N}(0, 1)$. In addition, we include a binary covariate $x_i^d \in \{0, 1\}$, that represents two experimental conditions. Assuming that a sample is obtained from each condition for a subject, we have $\mathbf{x}_i = (x_{s_i}^c, 0)$ or $(x_{s_i}^c, 1)$ for sample i. Thus, we have $N = 50$ samples in total. In Simulation 2, we also intend to assess our model's ability to recover the common factor model (De Vito et al., 2019, 2021), where some columns of common factor loadings are the same for all $\mathbf{x}_i$. That is, in the simulation truth, we have

$$\Sigma^{\mathrm{tr}}(\mathbf{x}_i) = \Lambda_0^{\mathrm{tr}} \Lambda_0^{\mathrm{tr},\prime} + \Lambda^{\mathrm{tr}}(\mathbf{x}_i) \Lambda^{\mathrm{tr},\prime}(\mathbf{x}_i) + \sigma^{2,\mathrm{tr}} \mathrm{I}_J, \tag{4.14}$$

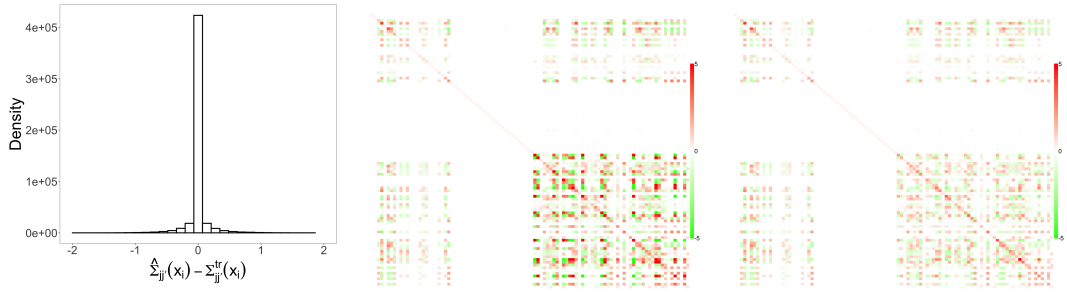
where Λ_0^{tr} is the common factors with dimension $J \times K_0$ resulting in a commonly shared baseline covariance. We first specify $\Sigma^{\text{tr}}(\mathbf{x}_i)$ in (4.14) by having $K_0 = 2$ common factors and $K^{\text{tr}} = 3$ covariate dependent factors. For $K_0 = 2$ common factors, we simulated $\lambda_{0,jk}^{\text{tr}}$ from $\text{N}(0, 1)$ and shifted away from zero by $1/2$ for OTUs 1-25 and 51-100 to ensure that those OTUs have baseline interactions. We let $\lambda_{0,jk}^{\text{tr}} = 0$ for the remaining OTUs. Similarly, for $K^{\text{tr}} = 3$ covariate dependent factors, we set q_{jk}^{tr} also from $\text{N}(0, 1)$ and shifted away from zero by $1/2$ but only for OTUs 51-100. And we have $f_{kp}^{\text{tr}} \overset{iid}{\sim} \text{Unif}(-1, 1)$ and $\sigma^{2,\text{tr}} = 0.5^2$. Under this design, interactions among OTUs 1-25 do not change with covariates, while among OTUs 51-100 change with the binary experimental covariate and the continuous covariate. The covariance matrix corresponding to $\Sigma^{\text{tr}}(\mathbf{x}_i)$ is illustrated in the lower triangle of Fig 4.3(b). Finally, we would like to include a considerable amount of zero counts and large variability over samples. For the normalized abundance level, we first set $\xi_{j1}^{\alpha,\text{tr}} = -5$, $\xi_{j2}^{\alpha,\text{tr}} \sim \text{N}(2.5, 0.5)$ and $\xi_{j3}^{\alpha,\text{tr}} \sim \text{N}(5, 0.5)$ and simulated $\boldsymbol{\psi}_j^{\text{tr}} = (\psi_{j1}^{\text{tr}}, \psi_{j2}^{\text{tr}}, \psi_{j3}^{\text{tr}}) \sim \text{Dir}(30, 40, 30)$. The three values, $\xi_{jl}^{\alpha,\text{tr}}$, $l = 1$, 2 and 3, represent zero, small and large counts, respectively. We then let $\alpha_{s_ij}^{\text{tr}} = \xi_{jl}^{\alpha,\text{tr}}$ with probability ψ_{jl}^{tr} for $s_i \in \{1, \ldots, S\}$. We next simulated size factors $r_i^{\text{tr}} \overset{iid}{\sim} \text{Unif}(0, 2)$ and regression coefficients $\beta_{jp}^{\text{tr}} \overset{iid}{\sim} \text{N}(0, 1)$. We included one binary covariate and one continuous covariate sampled from $\text{N}(0, 1)$. We had $\boldsymbol{\mu}^{\text{tr}}(\mathbf{x}_i) = r_i^{\text{tr}}\mathbf{1}_J + \boldsymbol{\alpha}_{s_i}^{\text{tr}} + \boldsymbol{\beta}'\mathbf{x}_i$. Under this setup, approximately 30.76% of Y_{ij}'s are 0. We specified the hyper-parameters values similar to Simulation 1 with $K = 7$. We ran MCMC for $160,000$ iterations and discarded the first half for burn-in. It took 23 hours on an Apple M1 chip laptop.

Fig 4.3(a) plotting the differences $\hat{\Sigma}_{jj'}(\mathbf{x}_i) - \Sigma^{\text{tr}}_{jj'}(\mathbf{x}_i)$ of all samples shows good estimates of interactions. We check the binary experimental effect and continuous covariate effect on the covariance in Fig 4.3(b)-(c) and Fig 4.4. Fig 4.4 presents the posterior estimation of covariance-covariates dependence among selected OTUs. Our method identifies truly inactive OTUs and estimates the baseline covariance successfully, and the OTU interrelationship with covariate structure is reasonably well captured even when the sample size is smaller than the number of OTUs. Figs 4.5(a) and (b) show the posterior median estimates of β_{jp} with their 95% credible interval estimates for binary covariate and continuous covariate, respectively. Our method effectively captures the covariate effects as well.

We also examine estimates of the factor loading matrix in Supp Fig C.1. Our posterior estimates of q_{jk} resemble two designed common factors, and for redundant factors we estimate extremely small posterior estimates of τ_k to shrink towards 0. Supp. Fig C.2 compares posterior median estimates of sample size factor r_i and the mean abundance μ_{ij} to their truth. In the figure, the library size factor and mean abundances are well estimated, serving as a reliable foundation for estimating the parameters of primary interest, such as $\Sigma(\mathbf{x}_i)$.

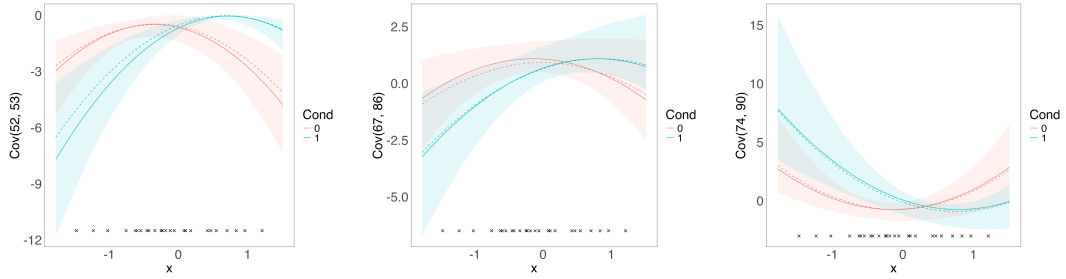
4.4.3 Simulation 3

We conducted Simulation 3 for a case with one binary covariate, where the covariance under each condition is arbitrarily generated using the vine method in Lewandowski et al. (2009). In particular, we simulated partial correlations from linearly



(a) $\hat{\Sigma}_{jj'}(\mathbf{x}) - \Sigma^{\text{tr}}_{jj'}(\mathbf{x})$ (b) $\hat{\Sigma}(\mathbf{x}_2)$ vs $\Sigma^{\text{tr}}(\mathbf{x}_2)$ (c) $\hat{\Sigma}(\mathbf{x}_{27})$ vs $\Sigma^{\text{tr}}(\mathbf{x}_{27})$

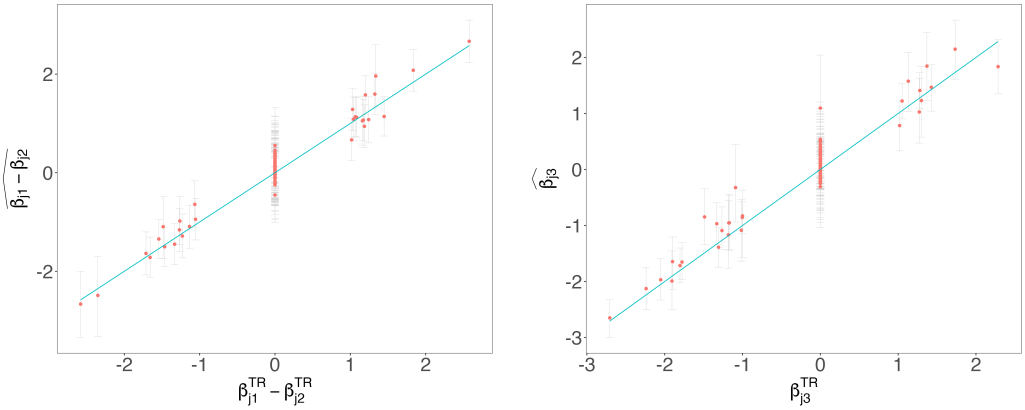
Figure 4.3: [Simulation 2] Panel (a) has a histogram of differences between $\hat{\Sigma}_{jj'}(\mathbf{x}_i)$ and $\Sigma^{\text{tr}}_{jj'}(\mathbf{x}_i)$ of all samples. In (b), the lower left and upper right triangles of the heatmap illustrate true values $\Sigma^{\text{tr}}_{jj'}$ and their posterior estimates of correlations $\hat{\Sigma}_{jj'}$, respectively. Two samples, samples 2 and 27, from subject 2, are arbitrarily chosen for illustration. Their covariates are $\mathbf{x}_2 = (1, -1.23), \mathbf{x}_{27} = (0, -1.23)$.



(a) OTU 52 & 53 (b) OTU 67 & 86 (c) OTU 74 & 90

Figure 4.4: [Simulation 2] Scatter plots of $\Sigma_{jj'}(\mathbf{x})$ (dashed) and $\Sigma^{\text{tr}}_{jj'}(\mathbf{x})$ (solid) are plotted for three arbitrarily chosen OTU pairs, OTUs 52 and 53, OTUs 67 and 86, and OTUs 74 and 90 for model checking. Crosses are observed covariates in the simulated data.

transformed Be$(1, 1)$ distribution over the interval of $(-1, 1)$. To encourage sparsity in $\Sigma^{\text{tr}}(x_i)$, we set the partial correlations below 0.8 to 0 and generated a correlation matrix, $\rho^{\text{tr}}(x_i)$ using their recursive formula. We then sampled $\sigma^{2,\text{tr}}$ independently from Unif$(1, 1.5)$ and let $\Sigma^{\text{tr}}_{jj'}(x_i) = \sigma^{2,\text{tr}}_j \sigma^{2,\text{tr}}_{j'} \rho^{\text{tr}}_{jj'}(x_i)$. Σ^{tr}_0 and Σ^{tr}_1 is shown in the lower triangle of Fig 4.6(b)-(c). For abundances, we kept the same as in Simulation 1 to simulate a count dataset. We used the same fixed hyperparameter values as in Simulation 2

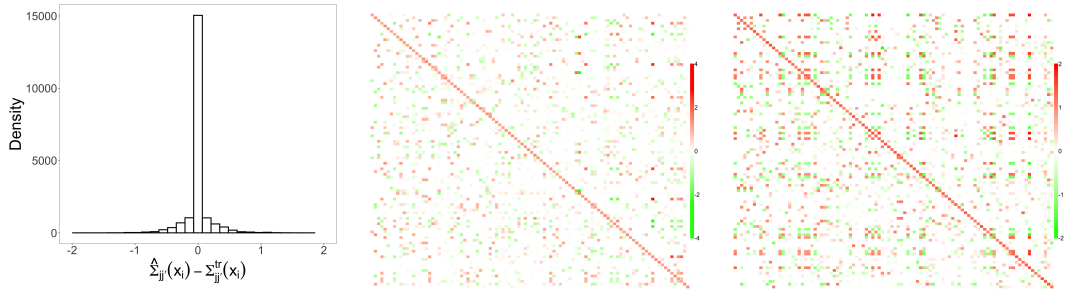


(a) $\widehat{\beta_{j1}-\beta_{j2}}$ versus $\beta_{j1}^{\text{tr}}-\beta_{j2}^{\text{tr}}$ (b) $\hat{\beta}_{j3}$ versus β_{j3}^{tr}

Figure 4.5: [Simulation 2] The posterior median estimates (dots) of β_{jp} for binary and continuous covariates are plotted in (a)-(b), respectively. Vertical lines represent their corresponding 95% credible interval estimates.

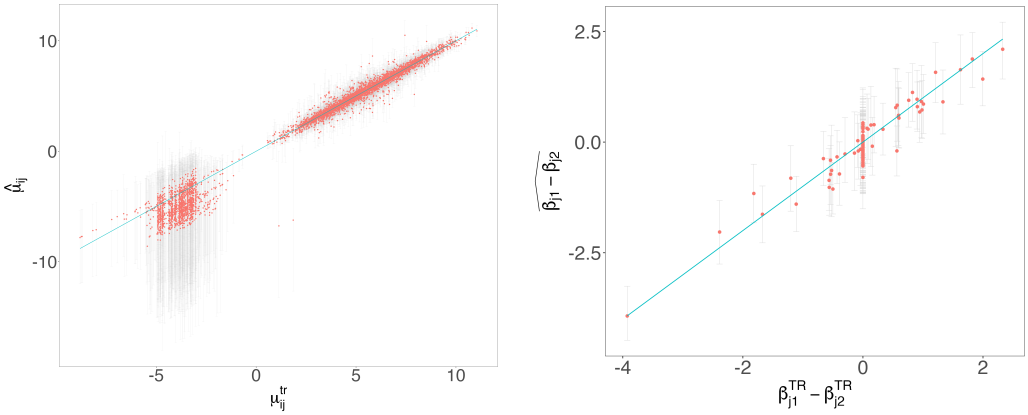
except $K = 25$. We test the sensitivity analysis in the supplementary that large enough K outputs similar estimates of the covariance matrix. We approximated the posterior distribution using MCMC. The examination of the MCMC simulation using traceplots indicated no evidence of convergence or mixing problems.

The upper triangle of Fig 4.6(a)-(b) illustrates the posterior estimates $\hat{\Sigma}_0$ and $\hat{\Sigma}_1$ under our model. Figs 4.7(a) and (b) show the posterior median estimates of mean abundance μ_{ij} of each OTU j in sample i and experimental coefficient $\beta_{j1} - \beta_{j2}$ (dots) with their 95% credible interval estimates (vertical lines), respectively. Our method effectively captures the true dependence structure among the OTUs, even with the arbitrary specification of Σ^{tr} and the added complexity introduced by the covariate in the true data-generating process.



(a) $\hat{\Sigma}_{jj'} - \Sigma^{\text{tr}}_{jj'}$ (b) $\hat{\Sigma}_0$ vs Σ^{tr}_0 (c) $\hat{\Sigma}_1$ vs Σ^{tr}_1

Figure 4.6: [Simulation 3] Panel (a) has a histogram of differences between $\hat{\Sigma}_{jj'}(x_i)$ and $\Sigma^{\text{tr}}_{jj'}(x_i)$ of all samples. In (b), the lower left and upper right triangles of the heatmap illustrate true values $\Sigma^{\text{tr}}_{x_i}$ and their posterior estimates of correlations $\hat{\Sigma}_{x_i}$, respectively.



(a) $\hat{\mu}_{ij}$ vs μ^{tr}_{ij} (b) $\widehat{\beta_{j1} - \beta_{j2}}$

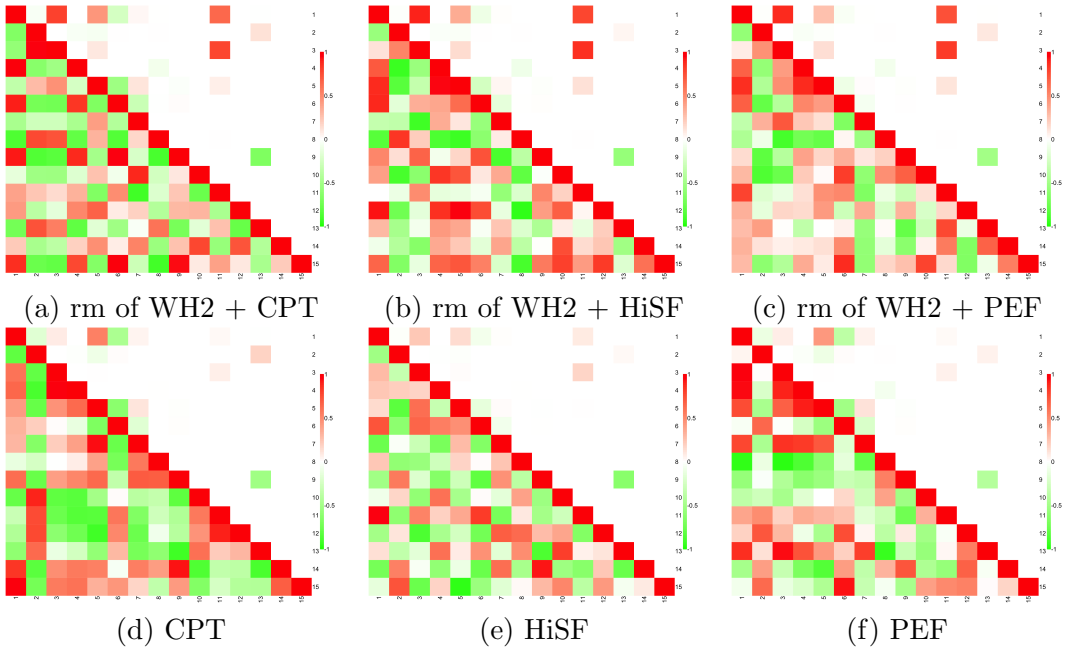
Figure 4.7: [Simulation 3] The posterior median estimates and 95% credible intervals of mean abundance μ_{ij} and experimental regression coefficient β_{jp} are plotted against the truth in panels (a) and (b), respectively.

4.5 Mice Gut Microbiome Data Analysis

We applied our method to the mice gut microbiome data in Patnode et al. (2019). $N = 30$ gnotobiotic mice were fed a human diet supplemented with different combinations of three different fiber preparations, e.g upper tertile of saturated fat (HiSF), 10% Citrus Pectin (CPT) and 10% Pea Fiber (PEF). Besides, one specific

OTU, Bacteroides cellulosilyticus WH2 (WH2), was removed from half of the mice at the beginning of the study. Thus, it resulted in two categorical variables: one with two levels (removal of WH2 or not) and the other with three levels (diets). 16S rRNA gene sequencing was performed, and $J = 15$ OTUs including OTU WH2 were included for analysis. The names of OTUs are included in the supplementary files for further studies. We denote covariates $\mathbf{x}_i = (x_{\text{WH2}}, x_{\text{diet}})$, where $x_{\text{WH2}} \in \{0, 1\}$ and $x_{\text{diet}} \in \{(0, 0), (1, 0), (0, 1)\}$. Since the removal of WH2 at the beginning highly affects the mean abundance of WH2 throughout the study, we have $\tilde{\mathbf{x}}_{ij} = \mathbf{x}_i$ for OTU WH2. For all other OTUs, the study focuses on their abundance changes though interactions with WH2. We set $\tilde{\mathbf{x}}_{ij} = x_{\text{diet}}$ for all other OTUs. Fig 4.8(a)-(f) shows empirical correlation estimates $\rho^{\text{em}}_{jj'}(\mathbf{x}_i)$ computed using $\log(y_{ij} + 0.01)$ after normalization with log of total count sample size factor estimates. To fit our model, the values of the fixed hyperparameter values were set similar to those of Simulation 1 with $K = 8$. The MCMC simulation was run over 160,000 iterations, with the first half iterations discarded as burn-in. It took 22 minutes on a M1 Mac.

Fig 4.8(a)-(f) illustrates posterior mean estimates of correlation $\hat{\rho}_{jj'}(\mathbf{x}_i)$ of all OTUs under six different conditions. We turn to correlation ranging from -1 to 1 for easier illustration in the context of microbiome study. Some common interaction structures are found under all conditions, such as negative correlation between OTUs 9 (Cat Collinsella aerofaciens TSDC17) and 13 (Cat Ruminococcaceae TSDC17). We identify three significantly different correlation pairs in Fig 4.9. Fig 4.9(a)-(b) shows that the correlation between OTU 1 (B ovatus ATCC.8483) and OTU 3 (B thetaiotaomicron

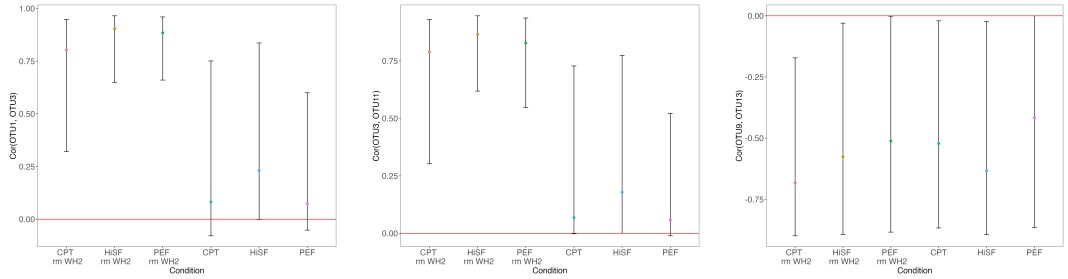


(a) rm of WH2 + CPT (b) rm of WH2 + HiSF (c) rm of WH2 + PEF

(d) CPT (e) HiSF (f) PEF

Figure 4.8: [Mice Data] The lower left and upper right triangles of the heatmap illustrate empirical estimates $\Sigma_{x_i}^{\text{em}}$ and their posterior estimates of correlations $\hat{\Sigma}_{x_i}$ under six different experimental conditions, respectively.

7330) and the correlation between 3 and 11 (Cat Odoribacter splanchnicus TSDC17) are highly positively correlated when OTU WH2 is removed from mice at the beginning. But this interaction is much mitigated in the mice having OTU WH2 from the beginning to the end. Fig 4.10 illustrates posterior median estimates $\hat{\beta}_{jp}$ of the two categorical covariates, respectively, with their 95% credible intervals. Fig 4.10(a) verifies the correctness of our results that mice which were removed WH2 at the beginning have a significantly smaller count abundance of OTU 2 (WH2). In Fig 4.10(b)-(d), the effect of having diet PET compared to CPT $\beta_{PET} - \beta_{CPT}$ is statistically significant for 4 OTUs. The effect estimates are positive for OTU 1 and 4 (B thetaiotaomicron VPI.5482) and negative for OTU 7 (Cat Bacteroides finegoldii TSDC17) and 13. Similarly, we also



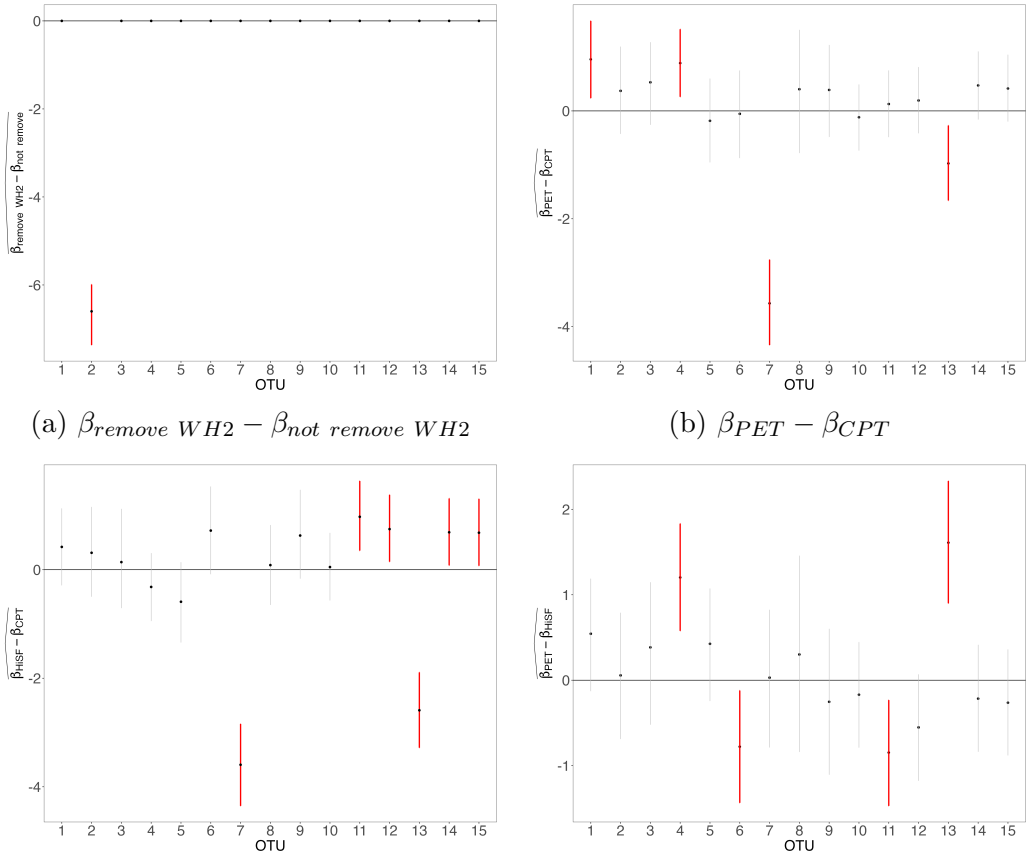
(a) OTU 1 and OTU 3 (b) OTU 3 and OTU 11 (c) OTU 9 and OTU 13

Figure 4.9: [Mice Data] Three representative pairs of OTUs having significantly different correlations under six conditions are plotted in (a)-(c). The points are the posterior estimate of correlation, and 95% credible intervals are in black intervals.

have significantly positive effects for OTU 4 comparing PET to HiSF, and significantly negative for OTU 7 and 13 comparing HiSF to CPT. It implies that OTU 4 is more abundant under diet PET, and OTUs 7,13 are very rare under diet CPT.

4.6 Conclusion

In this paper, we developed a Bayesian joint model of mean and covariance varying with covariates for high-dimensional multivariate count data. This method utilizes a covariate-dependent factor model for the covariance matrix and models the mean abundance using a flexible DP mixture. The model enables the assessment of covariate effects on mean and covariance in tandem. We place a Dir-Horseshoe prior on the covariate-dependent loading matrix to induce sparse feature interactions. The flexible mean mixture kernels handle the excess zeros and over-dispersion problems in the count data. The model is demonstrated through simulations and a real data example with categorical covariates.



(a) $\beta_{remove\ WH2} - \beta_{not\ remove\ WH2}$ (b) $\beta_{PET} - \beta_{CPT}$

(c) $\beta_{HiSF} - \beta_{CPT}$ (d) $\beta_{PET} - \beta_{HiSF}$

Figure 4.10: [Mice Data] Posterior inference of regression coefficients of two categorical covariates, where the posterior mean estimates are denoted by dots, and the 95% credible estimates with vertical lines. The intervals that do not contain zero are marked.

Our methods can be further extended by relaxing the linear covariance regression to a more complicated regression formula, such as introducing the transformation of covariates $\sqrt{x}$ and $x^{\frac{3}{2}}$. Introducing different orders of covariates induces a higher order of covariance-covariate relationship. It is important to note that while this extension offers greater flexibility, the same higher order of covariates entering the mean regression would need more exploration (variables selection). With added parameters, more

samples are needed to obtain reasonable inferences. A further research field is to study covariate-assisted interactions among temporal and/or longitudinal multivariate count data. In spatial and temporal heteroscedasticity, it's natural to consider the conditional dependence through time or locations. Fieuws and Verbeke (2006) discussed a pairwise approach jointly modeling of multivariate longitudinal data using mixed models, providing a foundation for understanding covariance structures in such contexts. Adapting the factor loading matrix to a time or location-dependent one has the potential to enhance the inference of interaction structures in other domains.

Chapter 5

Conclusion

In this thesis, we have developed flexible and efficient Bayesian methodologies for high-dimensional multivariate count data analysis, addressing the unique challenges posed by the complexities of such data including high dimensionality. By leveraging Bayesian frameworks, we introduced robust approaches that effectively capture the underlying dependence structure in high-dimensional count data, while accommodating the inherent uncertainty and variability.

Our exploration began with a zero-inflated Bayesian rounded kernel model tailored for count data, emphasizing its interpretability and adaptability in handling complex datasets. The covariance matrix of the kernel was estimated through a sparse factor model with a Dir-Laplace shrinkage prior on the factor loading matrix. We demonstrated how the prior was particularly beneficial in high-dimensional settings where data sparsity and overfitting are common concerns. The model also yielded reasonable estimates of relative count abundance by simultaneously performing a model-based normalization.

Simulation and real data examples both provide superior performance of the proposed model compared to the alternatives.

We next developed a Bayesian nonparametric method that integrates multiple sources of count tables. A Dirichlet process mixture of rounded kernel provides flexible multivariate distribution for count tables. We further constructed a novel shrinkage prior Dir-HS distribution to effectively induce a sparse factor loading matrix, leading to robust estimates of interactions in high-dimensional data. The theoretical properties of Dir-HS were examined and compared to existing priors. Simulations indicated the proposed model captured various shapes of distributions and recovered arbitrary random covariance matrices. We used the model to analyze a multi-domain skin microbiome dataset to infer interactions among the microbes from different domains.

Finally, we considered the problem of heteroscedasticity in a count vector and developed a covariate-dependent factor model for multivariate count data. We utilized a linear formulation in the lower-dimensional structure that induces a quadratic covariance-covariate relationship. Simulations showed the model with this parsimonious structure can sufficiently approximate arbitrary covariance varying over different experimental conditions. The parametric relationship brought computational efficiency and straightforward interpretability. We further extended this model to accommodate inter-subject heterogeneity. The model was demonstrated with analysis of mice gut microbiome dataset, where competition among microbes may be affected by different diets.

In conclusion, Bayesian high-dimensional multivariate count analysis offers a

comprehensive and powerful framework for analyzing complex count data. By introducing latent variables and combining rigorous statistical modeling with advanced computational techniques, these proposed approaches provide significant advantages in terms of flexibility, interpretability, and accuracy. Future research can build on these foundations by exploring more sophisticated count structures, such as spatial-temporal count data and tree-evolving count tables. The ongoing advancements in this field promise to enhance our ability to extract meaningful insights from high-dimensional count data, driving further innovations in various scientific and applied disciplines.

Bibliography

Agarwal, D. K., Gelfand, A. E., and Citron-Pousty, S. (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological statistics*, 9(4):341–355.

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

Alam, M. T., Amos, G. C., Murphy, A. R., Murch, S., Wellington, E. M., and Arasaradnam, R. P. (2020). Microbial imbalance in inflammatory bowel disease patients at different taxonomic levels. *Gut pathogens*, 12(1):1–8.

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

Alqawba, M. and Diawara, N. (2021). Copula-based markov zero-inflated count time series models with application. *Journal of Applied Statistics*, 48(5):786–803.

Andrade, J. C., Almeida, D., Domingos, M., Seabra, C. L., Machado, D., Freitas, A. C., and Gomes, A. M. (2020). Commensal obligate anaerobic bacteria and health: produc-

tion, storage, and delivery strategies. *Frontiers in Bioengineering and Biotechnology*, 8:550.

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

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., and Stegle, O. (2020). Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biology*, 21:1–17.

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.

Baker, J. L., Bor, B., Agnello, M., Shi, W., and He, X. (2017). Ecology of the oral microbiome: beyond bacteria. *Trends in microbiology*, 25:362–374.

Balakrishnama, S. and Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, 18(1998):1–8.

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

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

Bartlett, M. S. (1936). The square root transformation in analysis of variance. *Supplement to the Journal of the Royal Statistical Society*, 3(1):68–78.

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(7606):259–262.

Battey, H. (2017). Eigen structure of a new class of covariance and inverse covariance matrices. *Bernoulli*, 23(4B):3166 – 3177.

Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A. M., et al. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3. *Elife*, 10:e65088.

Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M.-C. C., Charles, T., Chen, X., Cocolin, L., Eversole, K., Corral, G. H., et al. (2020). Microbiome definition re-visited: old concepts and new challenges. *Microbiome*, 8(1):1–22.

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian statistics*, 7:733–742.

Bhattacharya, A. and Dunson, D. B. (2011). Sparse bayesian infinite factor models. *Biometrika*, pages 291–306.

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(512):1479–1490.

Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577 – 2604.

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199 – 227.

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

Billheimer, D., Guttorp, P., and Fagan, W. F. (2001). Statistical interpretation of species composition. *Journal of the American statistical Association*, 96(456):1205–1214.

Böhning, D., Dietz, E., and Schlattmann, P. (1997). Zero-inflated count models and their applications in public health and social science. *Applications of latent trait and latent class models in the social sciences*, pages 333–344.

Bostanci, N., Allaker, R., Belibasakis, G., Rangarajan, M., Curtis, M., Hughes, F., and McKay, I. (2007). Porphyromonas gingivalis antagonises campylobacter rectus induced cytokine production by human monocytes. *Cytokine*, 39:147–156.

Brogden, K. A., Guthmiller, J. M., and Taylor, C. E. (2005). Human Polymicrobial Infections. *The Lancet*, 365:253–255.

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

Byrd, A. L., Belkaid, Y., and Segre, J. A. (2018). The human skin microbiome. *Nature Reviews Microbiology*, 16(3):143–155.

Cai, T., Liu, W., and Luo, X. (2011). A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Cai, T., Ma, Z., and Wu, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability theory and related fields*, 161(3):781–815.

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.

Cai, Z., Zhu, T., Liu, F., Zhuang, Z., and Zhao, L. (2021). Co-pathogens in periodontitis and inflammatory bowel disease. *Frontiers in Medicine*, 8.

Canale, A. and Dunson, D. B. (2011). Bayesian kernel mixtures for counts. *Journal of the American Statistical Association*, 106(496):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.

Carroll, R. J. and Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. *The annals of statistics*, pages 429–441.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456.

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.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The Horseshoe Estimator for Sparse Signals. *Biometrika*, 97:465–480.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276.

Chandra, N. K., Dunson, D. B., and Xu, J. (2023). Inferring covariance structure from multiple data sources via subspace factor analysis. *arXiv preprint arXiv:2305.04113*.

Chattopadhyay, S., Arnold, J. D., Malayil, L., Hittle, L., Mongodin, E. F., Marathe, K. S., Gomez-Lobo, V., and Sapkota, A. R. (2021). Potential role of the skin and gut microbiota in premenarchal vulvar lichen sclerosus: A pilot case-control study. *PloS one*, 16(1):e0245243.

Chen, E. Z. and Li, H. (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, 32(17):2611–2617.

Chen, L., Reeve, J., Zhang, L., Huang, S., Wang, X., and Chen, J. (2018). Gmpr: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ*, 6:e4600.

Chiu, T. Y., Leonard, T., and Tsui, K.-W. (1996). The matrix-logarithmic covariance model. *Journal of the American Statistical Association*, 91(433):198–210.

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.

Connor, N., Barberán, A., and Clauset, A. (2017). Using null models to infer microbial co-occurrence networks. *PloS one*, 12(5):e0176751.

Cook, R. J., Lawless, J. F., and Lee, K.-A. (2010). A copula-based mixed poisson model for bivariate recurrent events under event-dependent censoring. *Statistics in Medicine*, 29(6):694–707.

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). An anova model

for dependent random measures. *Journal of the American Statistical Association*, 99:205–215.

De Vito, R., Bellio, R., Trippa, L., and Parmigiani, G. (2019). Multi-study factor analysis. *Biometrics*, 75:337–346.

De Vito, R., Bellio, R., Trippa, L., and Parmigiani, G. (2021). Bayesian multistudy factor analysis for high-throughput biological data. *The annals of applied statistics*, 15(4):1723–1741.

Deek, R. A. and Li, H. (2023). Inference of microbial covariation networks using copula models with mixture margins. *Bioinformatics*, 39(7):btad413.

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.

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

Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., and Huttenhower, C. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS computational biology*, 8(7):e1002606.

Ferrari, F. and Dunson, D. B. (2021). Bayesian factor analysis for inference on interactions. *Journal of the American Statistical Association*, 116:1521–1532.

Fieuws, S. and Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, 62(2):424–431.

Fong, P. W., Li, W. K., and An, H.-Z. (2006). A simple multivariate arch model specified by random coefficients. *Computational statistics & data analysis*, 51(3):1779–1802.

Fox, E. B. and Dunson, D. B. (2015). Bayesian nonparametric covariance regression. *The Journal of Machine Learning Research*, 16(1):2501–2542.

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

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

Fryer, M. (1989). Lognormal distributions: Theory and applications.

Gao, C. and Zhou, H. H. (2015). Rate-optimal posterior contraction for sparse pca. *The Annals of Statistics*, 43(2):785–818.

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.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Ghosh, J. and Dunson, D. B. (2009). Default prior distributions and efficient posterior computation in bayesian factor analysis. *Journal of Computational and Graphical Statistics*, 18(2):306–320.

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., and Knight, R. (2018). Current understanding of the human microbiome. *Nature medicine*, 24:392–400.

Gradshteyn, I. S. and Ryzhik, I. M. (2014). *Table of Integrals, Series, and Products*. Academic press.

Grantham, N. S., Guan, Y., Reich, B. J., Borer, E. T., and Gross, K. (2020). Mimix: A bayesian mixed-effects model for microbiome data from designed experiments. *Journal of the American Statistical Association*, 115(530):599–609.

Green, P. J. and Hastie, D. I. (2009). Reversible jump mcmc. *Genetics*, 155(3):1391–1403.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive metropolis algorithm. *Bernoulli*, pages 223–242.

Hayton, J. C., Allen, D. G., and Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational research methods*, 7(2):191–205.

Hinton, G. E. and Roweis, S. (2002). Stochastic neighbor embedding. *Advances in neural information processing systems*, 15.

Hoff, P. D. and Niu, X. (2012). A Covariance Regression Model. *Statistica Sinica*, pages 729–753.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Hotelling, H. (1992). Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*, pages 162–190. Springer.

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J., and Tappu, R. (2016). Megan community edition-interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS computational biology*, 12(6):e1004957.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., FitzGerald, M. G., Fulton, R. S., et al. (2012). Structure, function and diversity of the healthy human microbiome. *nature*, 486(7402):207.

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

Jiang, S., Xiao, G., Koh, A. Y., Kim, J., Li, Q., and Zhan, X. (2021). A bayesian zero-inflated negative binomial regression model for the integrative analysis of microbiome data. *Biostatistics*, 22(3):522–540.

Jonsson, V., Österlund, T., Nerman, O., and Kristiansson, E. (2017). Variability in

metagenomic count data and its influence on the identification of differentially abundant genes. *Journal of Computational Biology*, 24(4):311–326.

Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., Perry, T., Kao, D., Mason, A. L., Madsen, K. L., et al. (2016). Characterization of the gut microbiome using 16s or shotgun metagenomics. *Frontiers in microbiology*, 7:459.

Jung, S. and Takane, Y. (2008). Regularized common factor analysis. *New trends in psychometrics*, pages 141–149.

Kaakoush, N. O. (2015). Insights into the role of erysipelotrichaceae in the human host. *Frontiers in Cellular and Infection Microbiology*, 5:84.

Kamneva, O. K. (2017). Genome composition and phylogeny of microbes predict their co-occurrence in the environment. *PLoS computational biology*, 13(2):e1005366.

Kitamoto, S., Nagao-Kitamoto, H., Hein, R., Schmidt, T., and Kamada, N. (2020). The bacterial connection between the oral cavity and the gut diseases. *Journal of Dental Research*, 99(9):1021–1029.

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(5):e1004226.

La Rosa, P. S., Brooks, J. P., Deych, E., Boone, E. L., Edwards, D. J., Wang, Q., Sodergren, E., Weinstock, G., and Shannon, W. D. (2012). Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PloS one*, 7(12):e52078.

Lee, J. and Sison-Mangus, M. (2018). A bayesian semiparametric regression model for joint analysis of microbiome data. *Frontiers in microbiology*, 9:522.

Lee, T.-W. and Lee, T.-W. (1998). *Independent component analysis*. Springer.

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.

Li, J., Ellen, R., Hoover, C., and Felton, J. (1991). Association of proteases of porphyromonas (bacteroides) gingivalis with its adhesion to actinomyces viscosus. *Journal of dental research*, 70:82–86.

Li, Q., Guindani, M., Reich, B. J., Bondell, H. D., and Vannucci, M. (2017). A bayesian mixture model for clustering and selection of feature occurrence rates under mean constraints. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6):393–409.

Li, Z., Tian, L., O'Malley, A. J., Karagas, M. R., Hoen, A. G., Christensen, B. C.,

Madan, J. C., Wu, Q., Gharaibeh, R. Z., Jobin, C., et al. (2021). Ifaa: robust association identification and inference for absolute abundance in microbiome analyses. *Journal of the American Statistical Association*, 116:1595–1608.

Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., Andrews, E., Ajami, N. J., Bonham, K. S., Brislawn, C. J., et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655–662.

Lo, C. and Marculescu, R. (2018). Pglasso: Microbial community detection through phylogenetic graphical lasso. https://arxiv.org/abs/1807.08039v1.

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, pages 41–67.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15:1–21.

Lu, J., Breitwieser, F. P., Thielen, P., and Salzberg, S. L. (2017). Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science*, 3:e104.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J. R., and West, M. (2006). Sparse statistical modelling in gene expression genomics. *Bayesian inference for gene expression and proteomics*, 1(1):1644.

Ma, S., Ren, B., Mallick, H., Moon, Y. S., Schwager, E., Maharjan, S., Tickle, T. L., Lu, Y., Carmody, R. N., Franzosa, E. A., et al. (2021). A statistical model for

describing and simulating microbial community profiles. *PLoS computational biology*, 17(9):e1008913.

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.

Makalic, E. and Schmidt, D. F. (2015). A Simple Sampler for the Horseshoe Estimator. *IEEE Signal Processing Letters*, 23(1):179–182.

Mao, J., Chen, Y., and Ma, L. (2020). Bayesian graphical compositional regression for microbiome data. *Journal of the American Statistical Association*, 115(530):610–624.

Marchesi, J. R. and Ravel, J. (2015). The vocabulary of microbiome research: a proposal.

Mirsepasi-Lauridsen, H. C., Vallance, B. A., Krogfelt, K. A., and Petersen, A. M. (2019). Escherichia coli pathobionts associated with inflammatory bowel disease. *Clinical microbiology reviews*, 32(2):e00060–18.

Mirzaei, M. K. and Maurice, C. F. (2017). Ménage à trois in the human gut: interactions between host, bacteria and phages. *Nature Reviews Microbiology*, 15:397–408.

Mitra, S., Klar, B., and Huson, D. H. (2009). Visual and statistical comparison of metagenomes. *Bioinformatics*, 25(15):1849–1855.

Moran, K. R., Turner, E. L., Dunson, D., and Herring, A. H. (2021). Bayesian hier-

archical factor regression models to infer cause of death from verbal autopsy data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 70(3):532–557.

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

Myers, R. H. and Montgomery, D. C. (1997). A tutorial on generalized linear models. *Journal of Quality Technology*, 29(3):274–291.

Ni, Y., Stingo, F. C., and Baladandayuthapani, V. (2019). Bayesian graphical regression. *Journal of the American Statistical Association*, 114(525):184–197.

Nitzan, O., Elias, M., Chazan, B., Raz, R., and Saliba, W. (2013). Clostridium difficile and inflammatory bowel disease: role in pathogenesis and implications in treatment. *World journal of gastroenterology: WJG*, 19(43):7577.

Niu, X. and Hoff, P. D. (2019). Joint mean and covariance modeling of multiple health outcome measures. *The annals of applied statistics*, 13(1):321.

Niu, Y., Ni, Y., Pati, D., and Mallick, B. K. (2023). Covariate-assisted bayesian graph learning for heterogeneous data. *Journal of the American Statistical Association*, pages 1–15.

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.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016.

Pallast, N., Wieters, F., Fink, G. R., and Aswendt, M. (2019). Atlas-based imaging data analysis tool for quantitative mouse brain histology (aidahisto). *Journal of neuroscience methods*, 326:108394.

Papoz, L., Balkau, B., and Lellouch, J. (1996). Case counting in epidemiology: limitations of methods based on multiple data sources. *International journal of epidemiology*, 25(3):474–478.

Parada Venegas, D., De la Fuente, M. K., Landskron, G., González, M. J., Quera, R., Dijkstra, G., Harmsen, H. J., Faber, K. N., and Hermoso, M. A. (2019). Short chain fatty acids (scfas)-mediated gut epithelial and immune regulation and its relevance for inflammatory bowel diseases. *Frontiers in immunology*, page 277.

Park, J.-U., Oh, B., Lee, J. P., Choi, M.-H., Lee, M.-J., and Kim, B.-S. (2019). Influence of microbiota on diabetic foot wound in comparison with adjacent normal skin based on the clinical features. *BioMed research international*, 2019.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Pati, D., Bhattacharya, A., Pillai, N. S., Dunson, D., et al. (2014). Posterior contraction in sparse bayesian factor models for massive covariance matrices. *Annals of Statistics*, 42(3):1102–1130.

Patnode, M. L., Beller, Z. W., Han, N. D., Cheng, J., Peters, S. L., Terrapon, N., Henrissat, B., Le Gall, S., Saulnier, L., Hayashi, D. K., et al. (2019). Interspecies competition impacts targeted manipulation of human gut bacteria by fiber-derived glycans. *Cell*, 179(1):59–73.

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

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572.

Pereira, M. B., Wallroth, M., Jonsson, V., and Kristiansson, E. (2018). Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC genomics*, 19(1):1–17.

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.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690.

Pourahmadi, M. (2011). Covariance Estimation: The GLM and Regularization Perspectives. *Statistical Science*, 26(3):369 – 387.

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(6):e1009089.

Qin, J., Shi, X., Xu, J., Yuan, S., Zheng, B., Zhang, E., Huang, G., Li, G., Jiang, G., Gao, S., et al. (2021). Characterization of the genitourinary microbiome of 1,165 middle-aged and elderly healthy individuals. *Frontiers in Microbiology*, 12.

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.

Ravishanker, N., Venkatesan, R., and Hu, S. (2016). Dynamic models for time series of counts with a marketing application. *Handbook of discrete-valued time series*, pages 425–446.

Ren, B., Bacallado, S., Favaro, S., Vatanen, T., Huttenhower, C., and Trippa, L. (2017). Bayesian nonparametric mixed effects models in microbiome data analysis. *arXiv preprint arXiv:1711.01241*.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Ročková, V. and George, E. I. (2016). Fast bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association*, 111(516):1608–1622.

Rohwer, F., Prangishvili, D., and Lindell, D. (2009). Roles of viruses in the environment.

Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186.

Roy, A., Lavine, I., Herring, A. H., and Dunson, D. B. (2021). Perturbed factor analysis: Accounting for group differences in exposure profiles. *The annals of applied statistics*, 15(3):1386.

Rummel, R. J. (1988). *Applied factor analysis*. Northwestern University Press.

Rutemiller, H. C. and Bowers, D. A. (1968). Estimation in a heteroscedastic regression model. *Journal of the American Statistical Association*, 63(322):552–557.

Safari-Katesari, H., Samadi, S. Y., and Zaroudi, S. (2020). Modelling count data via copulas. *Statistics*, 54(6):1329–1355.

Schiavon, L., Canale, A., and Dunson, D. B. (2022). Generalized infinite factorization models. *Biometrika*, 109:817–835.

Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A., Waller, A., Mende, D. R., Kultima, J. R., Martin, J., et al. (2013). Genomic variation landscape of the human gut microbiome. *Nature*, 493(7430):45–50.

Schommer, N. N. and Gallo, R. L. (2013). Structure and function of the human skin microbiome. *Trends in microbiology*, 21(12):660–668.

Schwager, E., Mallick, H., Ventz, S., and Huttenhower, C. (2017). A bayesian method

for detecting pairwise associations in compositional data. *PLoS computational biology*, 13(11):e1005852.

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

Shi, P. and Valdez, E. A. (2014). Multivariate negative binomial models for insurance claim counts. *Insurance: Mathematics and Economics*, 55:18–29.

Shuler, K., Sison-Mangus, M., Lee, J., et al. (2020). Bayesian sparse multivariate regression with asymmetric nonlocal priors for microbiome data analysis. *Bayesian Analysis*, 15(2):559–578.

Shuler, K., Verbanic, S., Chen, I. A., and Lee, J. (2021a). A bayesian nonparametric analysis for zero-inflated multivariate count data with application to microbiome study. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

Shuler, K., Verbanic, S., Chen, I. A., and Lee, J. (2021b). A bayesian nonparametric analysis for zero-inflated multivariate count data with application to microbiome study. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 70.

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 51(1):47–60.

Sokal, R. R. and Rohlf, F. J. (1995). *biometry*. Macmillan.

Sokol, H., Seksik, P., Furet, J., Firmesse, O., Nion-Larmurier, I., Beaugerie, L., Cosnes,

J., Corthier, G., Marteau, P., and Doré, J. (2009). Low counts of faecalibacterium prausnitzii in colitis microbiota. *Inflammatory bowel diseases*, 15(8):1183–1189.

Tang, Z.-Z. and Chen, G. (2019). Zero-inflated generalized dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*, 20(4):698–713.

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.

Van Der Maaten, L., Postma, E., Van den Herik, J., et al. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res*, 10(66-71).

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):1–11.

Vester-Andersen, M., Mirsepasi-Lauridsen, H., Prosberg, M., Mortensen, C., Träger, C., Skovsen, K., Thorkilgaard, T., Nøjgaard, C., Vind, I., Krogfelt, K. A., et al. (2019). Increased abundance of proteobacteria in aggressive crohn's disease seven years after diagnosis. *Scientific reports*, 9(1):1–10.

Virtanen, S., Klami, A., Khan, S., and Kaski, S. (2012a). Bayesian group factor analysis. In *Artificial Intelligence and Statistics*, pages 1269–1277. PMLR.

Virtanen, S., Klami, A., Khan, S., and Kaski, S. (2012b). 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.

Wadsworth, W. D., Argiento, R., Guindani, M., Galloway-Pena, J., Shelburne, S. A., and Vannucci, M. (2017). An integrative bayesian dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC bioinformatics*, 18(1):1–12.

Wang, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 7(4):867–886.

Wang, T. and Zhao, H. (2017). A dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. *Biometrics*, 73(3):792–801.

Wang, Z., Mao, J., and Ma, L. (2021). Logistic-tree normal model for microbiome compositions. *arXiv preprint arXiv:2106.15051*.

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.

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1):1–18.

Winkelmann, R. (2008). *Econometric analysis of count data*. Springer Science & Business Media.

Wrzosek, L., Miquel, S., Noordine, M.-L., Bouet, S., Chevalier-Curt, M. J., Robert, V., Philippe, C., Bridonneau, C., Cherbuy, C., Robbe-Masselot, C., et al. (2013). Bacteroides thetaiotaomicron and faecalibacterium prausnitzii influence the production of mucus glycans and the development of goblet cells in the colonic epithelium of a gnotobiotic model rodent. *BMC biology*, 11(1):1–13.

Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844.

Xia, F., Chen, J., Fung, W. K., and Li, H. (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, 69(4):1053–1063.

Xiaoming, W., Jing, L., Yuchen, P., Huili, L., Miao, Z., and Jing, S. (2021). Characteristics of the vaginal microbiomes in prepubertal girls with and without vulvovaginitis. *European Journal of Clinical Microbiology & Infectious Diseases*, 40(6):1253–1261.

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.

Xie, F., Xu, Y., Priebe, C. E., and Cape, J. (2018). Bayesian estimation of sparse spiked covariance matrices in high dimensions. *arXiv preprint arXiv:1808.07433*.

Xu, L., Paterson, A. D., Turpin, W., and Xu, W. (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PloS one*, 10(7):e0129606.

Xu, T., Demmer, R. T., and Li, G. (2021). Zero-inflated poisson factor model with application to microbiome read counts. *Biometrics*, 77(1):91–101.

Young, V. B. (2017). The role of the microbiome in human health and disease: an introduction for clinicians. *Bmj*, 356.

Zhang, S., Shen, Y., Chen, I. A., and Lee, J. (2023a). 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.

Zhang, S., Shen, Y., Chen, I. A., and Lee, J. (2023b). 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.

Zhang, S., Shen, Y., Chen, I. A., and Lee, J. (2023c). Supplement to "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.

Zhang, S., Shen, Y., Chen, I. A., and Lee, J. (2024). Sparse Bayesian Group Factor Model for Feature Interactions in Multiple Count Tables Data. Technical report, University of California, Santa Cruz.

Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A. K., and Yi, N. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC bioinformatics*, 18(1):1–10.

Zhang, X. and Yi, N. (2020). Fast zero-inflated negative binomial mixed modeling approach for analyzing longitudinal metagenomics data. *Bioinformatics*, 36:2345–2351.

Zhao, S., Gao, C., Mukherjee, S., and Engelhardt, B. E. (2016). Bayesian group factor analysis with structured sparsity. *The Journal of Machine Learning Research*, 17(1):6868–6914.

Appendix A

SUPPLEMENTARY FOR Bayesian Modeling of Interaction between Features in Sparse Multivariate Count Data with Application to Microbiome Study

A.1 Details of Posterior Computation

We use Markov chain Monte Carlo (MCMC) to draw samples of the random parameters from their posterior distribution. Recall that $Y_{ij} \in \mathbb{N}^0$, $i = 1, \ldots, N$ and $j = 1, \ldots, J$ denotes the count of OTU j in sample i, and the model assumes $Y_{ij} = \lfloor Y_{ij}^{\star} \rfloor$ with $Y_{ij}^{\star} \in \mathbb{R}^{+}$. The distribution of $\mathbf{Y}_i^{\star} = (Y_{i1}^{\star}, \ldots, Y_{iJ}^{\star})$ is specified in (1) of the main

text. We also let $\mathbf{X}$ represent a $N \times P$ covariate matrix whose rows have a $P-$dim covariate vector $\mathbf{x}_i$. The probit regression for the probabilities of an OTU being absent may have a different covariate vector $\tilde{\mathbf{x}}_i$ with P_κ even for the same set of covariates due to different parameterizations. We let $\tilde{\mathbf{X}}$ be a $N \times P_\kappa$ covariate matrix having $\tilde{\mathbf{x}}_i$ in rows and $\tilde{\mathbf{X}}_\star = [\mathbf{1}, \tilde{\mathbf{X}}]$ a $N \times (P_\kappa + 1)$ covariate matrix with the first column being a column of 1.

To facilitate updating the parameters related to zero inflation such as δ_{ij} and $\boldsymbol{\kappa}_j$, we introduce a continuous real valued latent variable $z_{ij} \sim \mathrm{N}(\kappa_{j0} + \tilde{\mathbf{x}}_i'\boldsymbol{\kappa_j}, 1)$ and let $\delta_{ij} = 1$ if $z_{ij} < 0$ and otherwise $\delta_{ij} = 0$. We then have $\epsilon_{ij} = \Pr(\delta_{ij} = 1) = \Pr(z_{ij} < 0 \mid \boldsymbol{\kappa_j}) = \Phi(\kappa_{j0} + \tilde{\mathbf{x}}_i'\boldsymbol{\kappa_j})$. The MCMC steps of updating parameters $\boldsymbol{\kappa}_j$ and latent variables δ_{ij} and z_{ij} can be summarized as below;

- $\boldsymbol{\kappa}_j$: The full conditional distribution of $\boldsymbol{\kappa}_j$ is

$$\boldsymbol{\kappa}_j \mid \boldsymbol{\mu}_\kappa, \Sigma_\kappa \sim \mathrm{N}_{(P_\kappa+1)}((\tilde{\mathbf{X}}_\star'\tilde{\mathbf{X}}_\star + u_\kappa \mathrm{I}_{P+1})^{-1}((u_\kappa \mathrm{I}_{P+1})^{-1}\boldsymbol{\mu}_\kappa + \tilde{\mathbf{X}}'\mathbf{z}_j), \\ (\tilde{\mathbf{X}}_\star'\tilde{\mathbf{X}}_\star + u_\kappa \mathrm{I}_{P+1})^{-1}).$$

  Draw $\boldsymbol{\kappa}_j$ from its full conditional.

- z_{ij}: The full conditional distribution is

$$z_{ij} \sim \begin{cases} \mathrm{N}_-(\kappa_{j0} + \tilde{\mathbf{x}}_i'\boldsymbol{\kappa}_j, 1) & \text{if } \delta_{ij} = 1, \\ \mathrm{N}_+(\kappa_{j0} + \tilde{\mathbf{x}}_i'\boldsymbol{\kappa}_j, 1) & \text{if } \delta_{ij} = 0, \end{cases}$$

where N_+ and N_- represent normal distributions truncated below and above at zero, respectively.

- δ_{ij}: For (i,j) with $Y_{ij}=0$, update δ_{ij} using the full conditional;

$$\begin{aligned}\Pr(\delta_{ij}=1 \mid -) &\propto \epsilon_{ij},\\ \Pr(\delta_{ij}=0 \mid -) &\propto (1-\epsilon_{ij})\int_{-\infty}^{0}\phi(\tilde{y}^{\star}_{ij} \mid r_i+\alpha_j+s_{g_i,j}+\lambda_j'\eta_i+\mathbf{x}_i'\beta_j,\sigma^2)d\tilde{y}^{*}_{ij},\end{aligned}$$

where $\epsilon_{ij}=\Phi(\kappa_{j0}+\tilde{\mathbf{x}}_i'\boldsymbol{\kappa}_j,1)$ and $\phi(\cdot \mid a,b^2)$ represents the probability density function of the normal distribution with mean a and variance b^2. If $Y_{ij}>0$, $\delta_{ij}=0$ with probability 1.

Updating of the random parameters related to μ_{ij} and Σ can be more convenient with latent continuous variables $\tilde{Y}^{\star}_{ij}=\log(Y^{\star}_{ij})$ for (i,j) having $\delta_{ij}=0$ imputed as follows;

$$\tilde{Y}^{\star}_{ij}\sim \mathrm{N}(r_i+\alpha_j+s_{g_i,j}+\boldsymbol{\lambda}_j'\boldsymbol{\eta}_i+\mathbf{x}_i'\boldsymbol{\beta}_j,\sigma^2)1(\log(y_{ij})\le \tilde{Y}^{\star}_{ij}<\log(y_{ij})),$$

that is, a truncated normal distribution, where the support is determined by the observed count y_{ij}. Given $\tilde{Y}^{\star}_{ij}$, the full conditionals of the parameters in Σ except for ϕ_j have a standard form, and the Gibbs sampler can be used to update τ_k, λ_{jk} and σ^2. Specifically, we re-write the Laplace distribution as a normal scale mixture to facilitate the step of updating λ_{jk} from its full conditional; $\lambda_{jk} \mid \zeta_{jk},\phi_j,\tau_k \overset{indep}{\sim} \mathrm{N}(0,\zeta_{jk}\phi_j^2\tau_k^2)$ and $\zeta_{jk} \overset{iid}{\sim} \mathrm{Exp}(1/2)$. Then λ_{jk} can be easily obtained through a data augmented Gibbs

step. The full conditional distribution of ζ_{jk}^{-1} and τ_k can be sampled from the inverse Gaussian and generalized inverse Gaussian sampling distribution (Park and Casella, 2008). We update ϕ_j using a Metropolis-Hastings step. We let $\phi_j^\star \overset{iid}{\sim} \text{Ga}(a_\phi, 1)$ and have $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_J) \sim \text{Dir}(a_\phi, \ldots, a_\phi)$ with $\phi_j = \phi_j^\star / \sum_{j'} \phi_{j'}^\star$. The full conditional of $\boldsymbol{\phi}$ is given by

$$\begin{aligned} p(\boldsymbol{\phi} \mid -) &\propto p(\boldsymbol{\lambda} \mid \boldsymbol{\tau}, \boldsymbol{\phi}, \boldsymbol{\zeta}) p(\boldsymbol{\phi}) \\ &\propto \prod_{j=1}^{J} \prod_{k=1}^{K} \text{N}(\lambda_{jk} \mid 0, \zeta_{jk} \phi_j^2 \tau_k^2) \prod_{j=1}^{J} \text{Ga}(\phi_j^\star \mid a_\phi, 1). \end{aligned}$$

In order to explore the posterior distribution of ϕ_j efficiently, the adaptive MH algorithm (Haario et al., 2001) is used. We adjust the MH step size according to the acceptance ratio, and the convergence rate is accelerated.

Recall that we have parameters, r_i, α_j, $\boldsymbol{\beta}_j$, $s_{g_i,j}$ and u_s^2, for μ_{ij} in (8) of the main text, and parameters, $\boldsymbol{\psi}^\chi$, ω_l^χ and ξ_l^χ, $\chi \in \{\alpha, r\}$, in (9) of the main text. The full conditional distributions of the parameters $\boldsymbol{\beta}_j$, $s_{g_i,j}$ and u_s^2 have a standard distribution, and their samples are easily drawn through a usual Bayesian Normal-Gamma model update. Size factors r_i and α_j have a mixture of mixtures as their prior. To facilitate computation, we introduce a pair of auxiliary variables for each $\chi \in \{\alpha, r\}$ that specifies the mixture component from which each particular χ is, i.e., (S_{i1}^r, S_{i2}^r) for r_i, where $S_{i1}^r \in \{1, \ldots, L^r\}$ and $S_{i2}^r \in \{0, 1\}$, and $(S_{j1}^\alpha, S_{j2}^\alpha)$ for α_j, where $S_{i1}^\alpha \in \{1, \ldots, L^\alpha\}$ and $S_{i2}^\alpha \in \{0, 1\}$. We then assume $\text{P}(S_{i1}^r = l) = \psi_l^r$ and $\text{P}(S_{i2}^r = 0 \mid S_{i1}^r = l) = \omega_l^r$, and similarly, assume $\text{P}(S_{i1}^\alpha = l) = \psi_l^\alpha$ and $\text{P}(S_{i2}^\alpha = 0 \mid S_{j1}^r = l) = \omega_l^\alpha$. The conditional prior

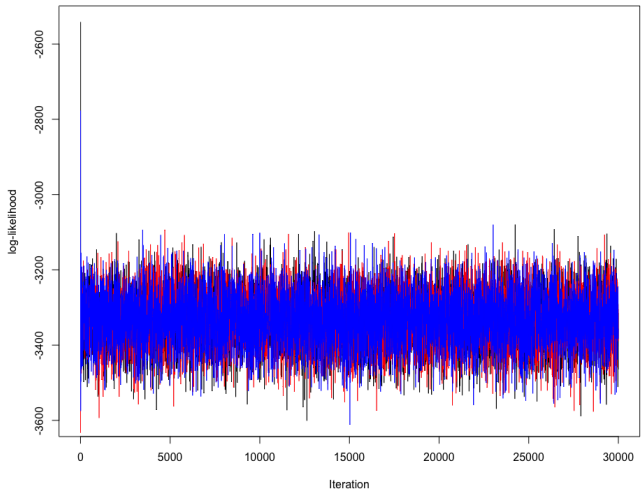


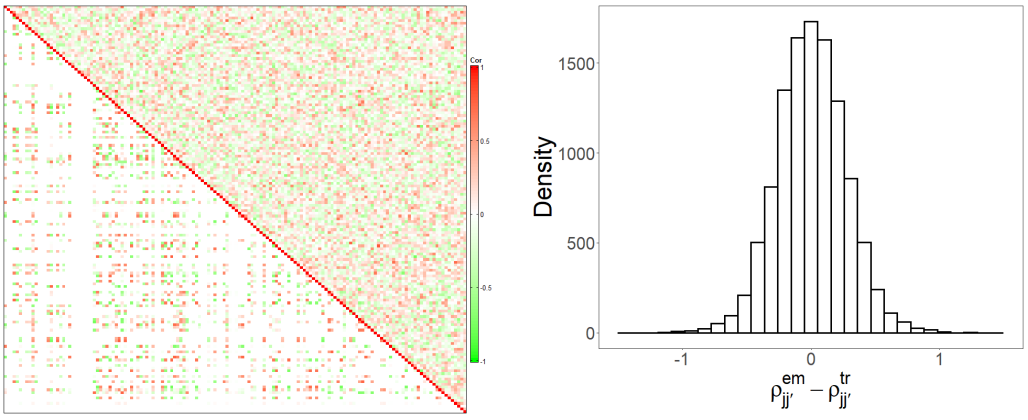
Figure A.1: [Simulation 1] Traceplots of log likelihood under three different initializations are presented in black, red and blue, respectively.

distributions of r_i and α_j are

$$r_i \mid \boldsymbol{\psi}^r, \boldsymbol{\omega}^r, \boldsymbol{\xi}^r, S^r_{i1} = l, S^r_{i2} \sim \begin{cases} \mathrm{N}(\xi^r_l, u^2_r) & \text{if } S^r_{i2} = 0, \\ \mathrm{N}\left(\frac{v_r - \omega^r_l \xi^r_l}{1-\omega^r_l}, u^2_r\right) & \text{if } S^r_{i2} = 1, \end{cases}$$

$$\alpha_j \mid \boldsymbol{\psi}^\alpha, \boldsymbol{\omega}^\alpha, \boldsymbol{\xi}^\alpha, S^\alpha_{i1} = l, S^\alpha_{i2} \sim \begin{cases} \mathrm{N}(\xi^\alpha_l, u^2_\alpha) & \text{if } S^\alpha_{i2} = 0, \\ \mathrm{N}\left(\frac{v_\alpha - \omega^\alpha_l \xi^\alpha_l}{1-\omega^\alpha_l}, u^2_\alpha\right) & \text{if } S^\alpha_{i2} = 1, \end{cases}$$

Conditional on those indicators, $\boldsymbol{\psi}^\chi$ can be drawn through a traditional Multinomial-Dirichlet model update and ω^χ_l through a Metropolis-hasting update. Also, given the indicators, the full conditional distributions of r_i, α_j and ξ^χ_l, $\chi \in \{r, \alpha\}$ have a Gaussian distribution.



(a) $\rho_{jj'}^{\text{tr}}$ versus $\rho_{jj'}^{\text{em}}$ (b) $\rho_{jj'}^{\text{em}} - \rho_{jj'}^{\text{tr}}$

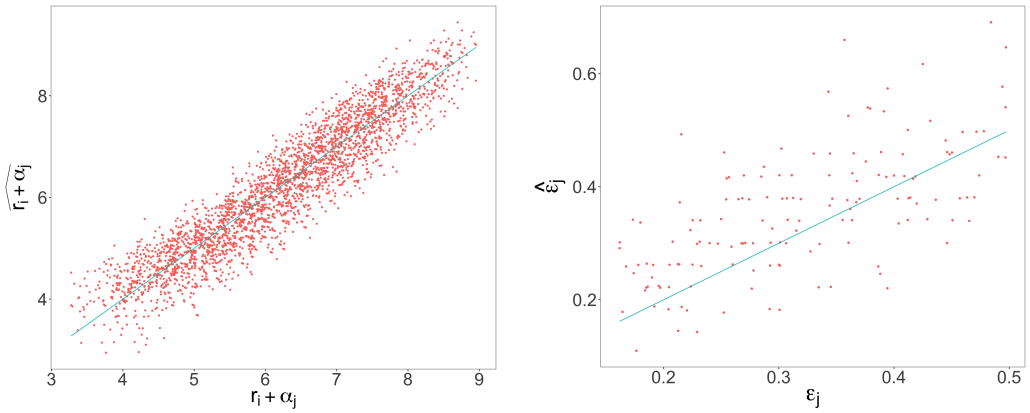
Figure A.2: [Simulation 1] The upper right and lower left triangles of the heatmap in panel (a) illustrates empirical correlation estimates $\rho_{jj'}^{\text{em}}$ of $\log(Y_{ij} + 0.01)$ scaled with CSS and the true values of the correlations $\rho_{jj'}^{\text{tr}}$, respectively. Panel (b) has a histogram of differences between $\rho_{jj'}^{\text{em}}$ and $\rho_{jj'}^{\text{tr}}$.

A.2 Instruction for the R package, ZI-MLN

ZI-MLN is an R package that reproduces the tables and figures in Chapter 2 and evaluates the performance of ZI-MLN. Download and install R from https://www.r-project.org/. It requires R 3.6 or greater. Once installed, open R from the terminal and run the following command to install packages.

```
install.packages(c("statmod", "GIGrvg", "extraDistr", "mvtnorm"))
```

One can also import the GitHub repository https://github.com/shuang-jie/ZI-MLN directly to load all functions. Two main functions in the repository are 'ZI MLN without' for microbiome count tables without covariates and 'ZI MLN with' for cases with covariates. The input count table Y is a raw count table and does not need normalization. For the input count table, samples are in rows and features (OTUs) are in columns. Note we also have an index $m = 1, 2...M$ for subjects. For example,



(a) $r_i + \alpha_j$ (b) ϵ_{ij}

Figure A.3: [Simulation 1] In panel (a), posterior mean estimates of the mean abundance $r_i + \alpha_j$ are plotted against the simulation truth $r_i^{\text{tr}} + \alpha_j^{\text{tr}}$. Panel (b) has a histogram of the differences in posterior mean estimates $\hat{\epsilon}_{ij}$ of probabilities of an OTU being absent and their true values $\epsilon_{ij}^{\text{tr}}$.

$m = 1, 2, 2, 3$ means that the four samples are from subject 1, 2, 2 and 3, respectively. A special case is $m = 1, 2, 3, \ldots, n$, which implies one sample is obtained from each subject. Hyper-parameter specifications are discussed in the simulation part of Chapter 2. We also provide simulation code in 'without covariate.R' to analyze simulated data when there are no covariates on artificially generated data. More details of the code are on the github README page.

A.3 Additional Simulation Studies

A.3.1 Additional Results of Simulation 1

In this subsection, we present additional results from Simulation 1 in § 2.3.1 of the main text. We examined the convergence of the MCMC simulation using trace plots of the log-likelihood. The model was run under different initializations and random

seeds. Traceplots of the log-likelihood in Fig A.1 suggest that the model converged to a similar state under these different initializations. The figure provides practical evidence of the model's convergence. Fig A.2 illustrates empirical estimates of the marginal correlations $\rho^{\text{em}}_{jj'}$ of the logarithm transformed counts, $\log(y_{ij} + 0.01)$ after normalization with sample size factors estimated by CSS. The true values $\rho^{\text{tr}}_{jj'}$ of the correlations are shown in the lower left triangle of the heatmap in panel (a). In panel (b), a histogram of the differences $\rho^{\text{em}}_{jj'} - \rho^{\text{tr}}_{jj'}$ is shown. Fig A.3(a) shows posterior mean estimates of the baseline abundance $r_i + \alpha_j$ of OTU j in sample i compared to the simulation truth $r^{\text{tr}}_i + \alpha^{\text{tr}}_j$. The figure indicates that the mean abundances $r_i + \alpha_j$ are identifiable although r_i and α_j are not individually identifiable, and our model provides good estimates of the mean abundance. Furthermore, it provides a good basis for the estimation of parameters of our main interest such as Σ. We also examined the performance of estimating probabilities ϵ_{ij} of OTUs being absent in samples. Under the setup without covariates, $\epsilon_{ij} = \Phi_1(\kappa_{j0} \mid 0, 1)$ is identical for all i, i.e., $\epsilon_{ij} = \epsilon_j$ for all i. Fig A.3(b) shows that ϵ_j's are well estimated even with a small sample size $N = 20$.

A.3.2 Additional Results of Simulation 2

In this subsection, we include additional results from Simulation 2, described in § 2.3.2 of the main text. We ran the model on the dataset with different initializations and random seeds for the MCMC chain. Traceplots of the log-likelihood under the different random seeds and initializations are shown in Fig A.4. The figure shows the MCMC converges to similar log-likelihood ranges under these different specifications,

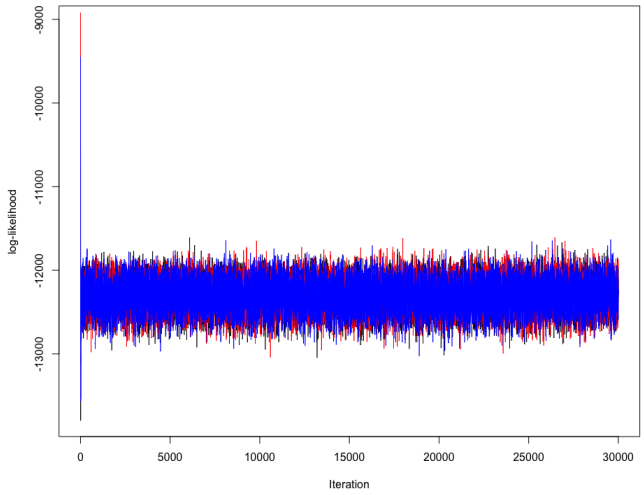
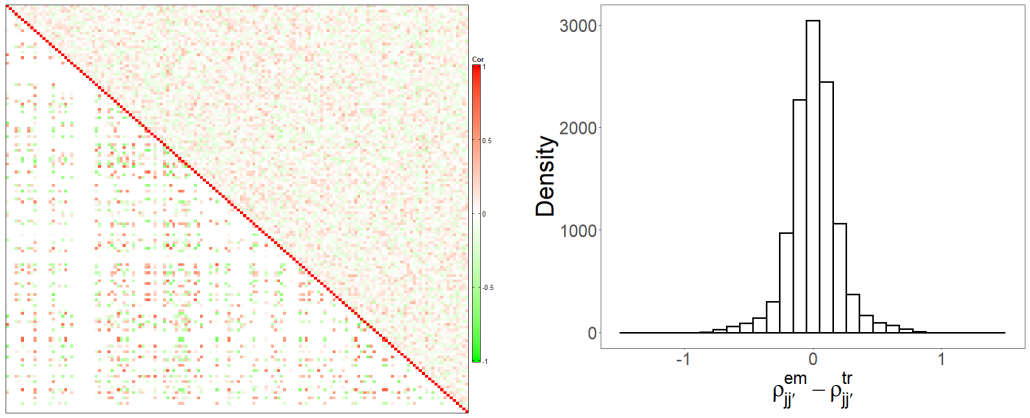


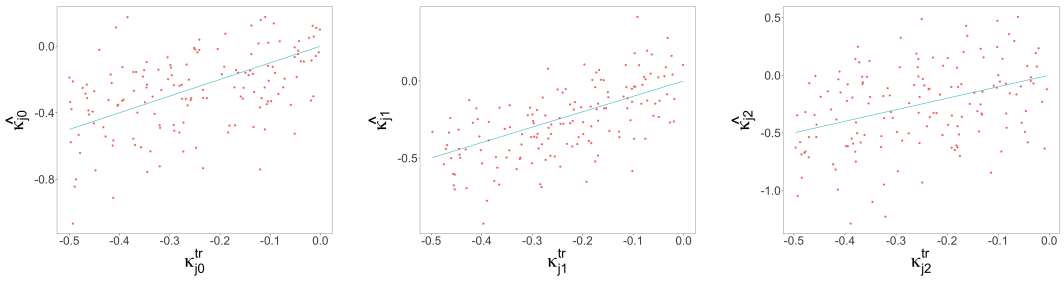
Figure A.4: [Simulation 2] Traceplots of log likelihood under three different initializations are presented in black, red and blue, respectively.



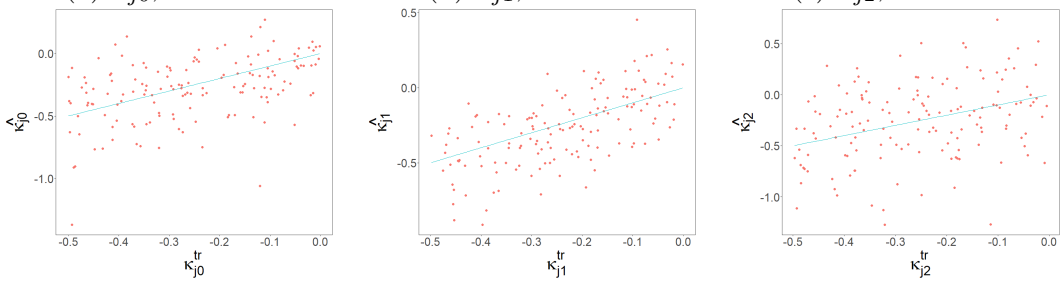
(a) $\rho^{\text{tr}}_{jj'}$ versus $\rho^{\text{em}}_{jj'}$ (b) $\rho^{\text{em}}_{jj'} - \rho^{\text{tr}}_{jj'}$

Figure A.5: [Simulation 2] The upper right and lower left triangles of the heatmap in panel (a) illustrate empirical correlation estimates $\rho^{\text{em}}_{jj'}$ of $\log(Y_{ij} + 0.01)$ scaled with CSS and the true values of the correlations $\rho^{\text{tr}}_{jj'}$, respectively. Panel (b) has a histogram of differences between $\rho^{\text{em}}_{jj'}$ and $\rho^{\text{tr}}_{jj'}$.

and we did not find evidence suggesting the Markov chain failed to converge. Fig A.5 compares empirical estimates $\rho^{\text{em}}_{jj'}$ of the marginal correlations to the true values $\rho^{\text{tr}}_{jj'}$ of

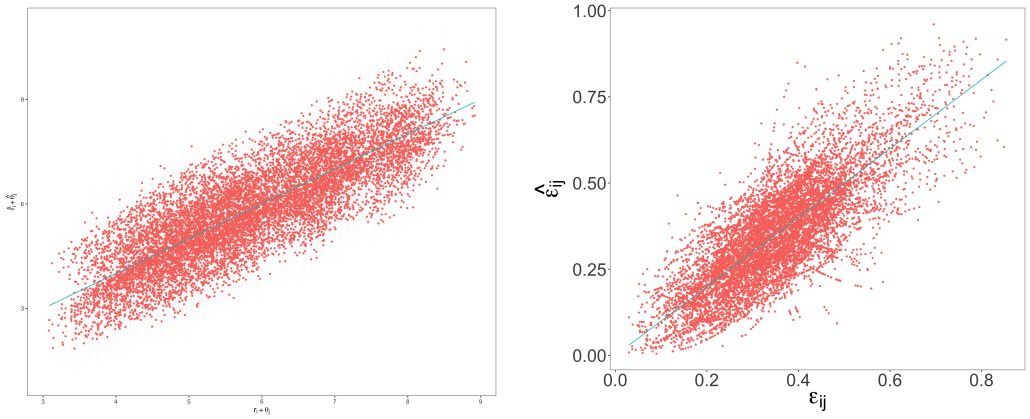


(a) κ_{j0}, ZI-MLN (b) κ_{j1}, ZI-MLN (c) κ_{j2}, ZI-MLN



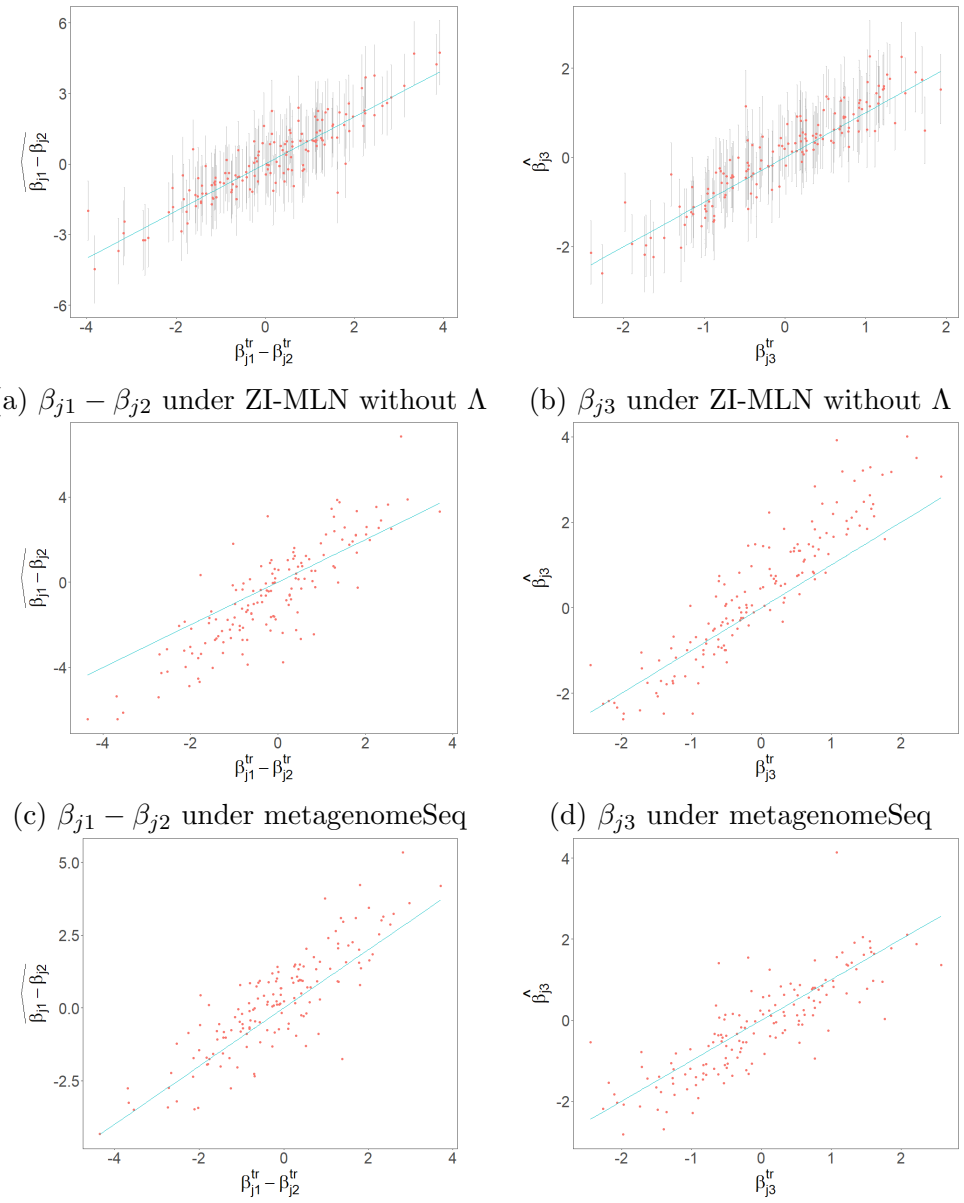
(d) κ_{j0}, ZI-MLN w.o. Λ (e) κ_{j1}, ZI-MLN w.o. Λ (f) κ_{j2}, ZI-MLN w.o. Λ

Figure A.6: [Simulation 2] Posterior mean estimates $\hat{\kappa}_{jp}$ of coefficients on ϵ_{ij} are plotted against the truth for $p = 0, 1, 2$ in columns 1-3, respectively. The top and bottom rows are for ZI-MLN and ZI-MLN without Λ, respectively.



(a) $r_i + \alpha_j$ (b) ϵ_{ij}

Figure A.7: [Simulation 2] In panel (a), posterior mean estimates of the mean abundance $r_i + \alpha_j$ are plotted against the simulation truth $r_i^{\text{tr}} + \alpha_j^{\text{tr}}$. Panel (b) has a histogram of the differences in posterior mean estimates $\hat{\epsilon}_{ij}$ of probabilities of an OTU being absent and their true values $\epsilon_{ij}^{\text{tr}}$.

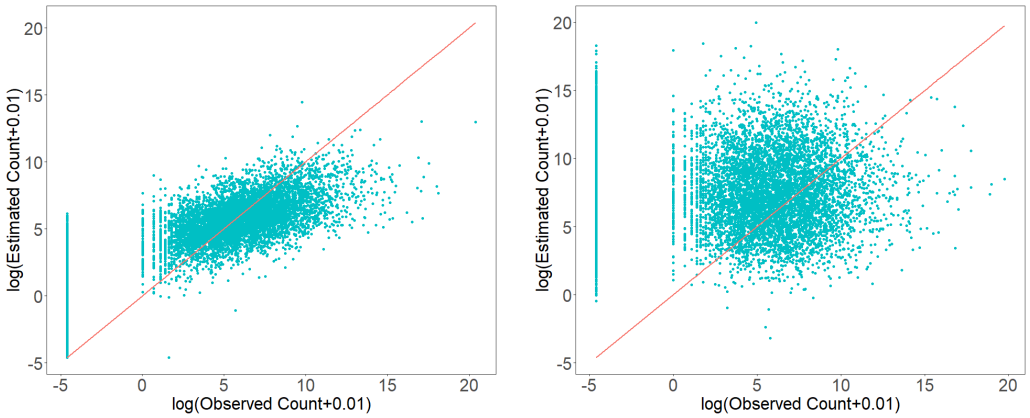


(a) $\beta_{j1} - \beta_{j2}$ under ZI-MLN without Λ (b) β_{j3} under ZI-MLN without Λ

(c) $\beta_{j1} - \beta_{j2}$ under metagenomeSeq (d) β_{j3} under metagenomeSeq

(e) $\beta_{j1} - \beta_{j2}$ under edgeR (f) β_{j3} under edgeR

Figure A.8: [Simulation 2: Comparison] Estimates $\widehat{\beta_{j1} - \beta_{j2}}$ and $\hat{\beta}_{j3}$ of regression coefficients are compared to the truth, $\beta_{j1}^{\text{tr}} - \beta_{j2}^{\text{tr}}$ and β_{j3}^{tr}. The estimates in rows 1-3 are produced by ZI-MLN without Λ, metagenomeSeq and edgeR, respectively.



(a) MetagenomeSeq (b) EdgeR

Figure A.9: [Simulation 2: Comparison] Panels (a) and (b) present scatter plots of observed $\log(y_{ij} + 0.01)$ versus $\log(\hat{\mu}_{ij} + 0.01)$ by metagenomeSeq and edgeR. $\hat{\mu}_{ij}$ are estimated mean abundances of OTUs in samples.

the correlations. $\rho^{\text{em}}_{jj'}$ and $\rho^{\text{tr}}_{jj'}$ are shown in the upper right and lower left triangles of the heatmap in panel (a), respectively. A histogram of differences $\rho^{\text{em}}_{jj'} - \rho^{\text{tr}}_{jj'}$ is in panel (b). Figs A.6(a)-(c) compare posterior mean estimates $\hat{\kappa}_{jp}$ of probit regression coefficients on ϵ_{ij} to their true values under ZI-MLN. In Fig A.7, we examine the estimation of mean abundances $r_i + \alpha_j$ and the probabilities ϵ_{ij} of an OTU being absent in a sample. The figure shows that posterior mean estimates of $r_i + \alpha_j$ and of ϵ_{ij} are close to their simulation truth, and the model reasonably well recovers the simulation truth.

Fig A.8 illustrates posterior mean estimates of regression coefficients β_{jp} under comparators, ZI-MLN without Λ, metagenomeSeq and edgeR. Figs A.6(d)-(f) compare posterior mean estimates of κ_{jp} under ZI-MLN without Λ to their truth. Estimates of OTU mean abundances under metagenomeSeq and edgeR are compared to the observed counts in Fig A.9.

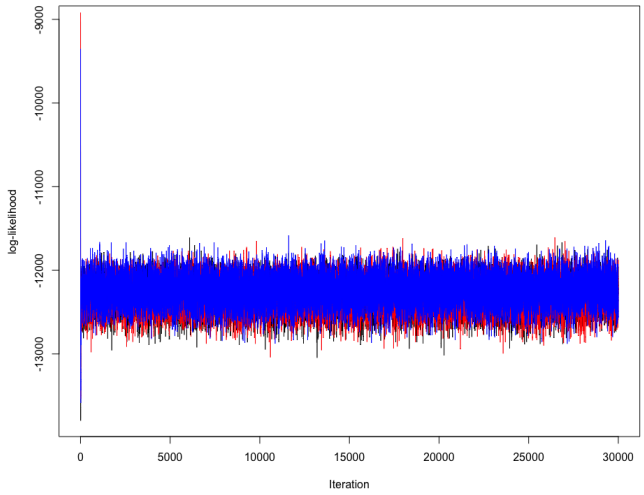
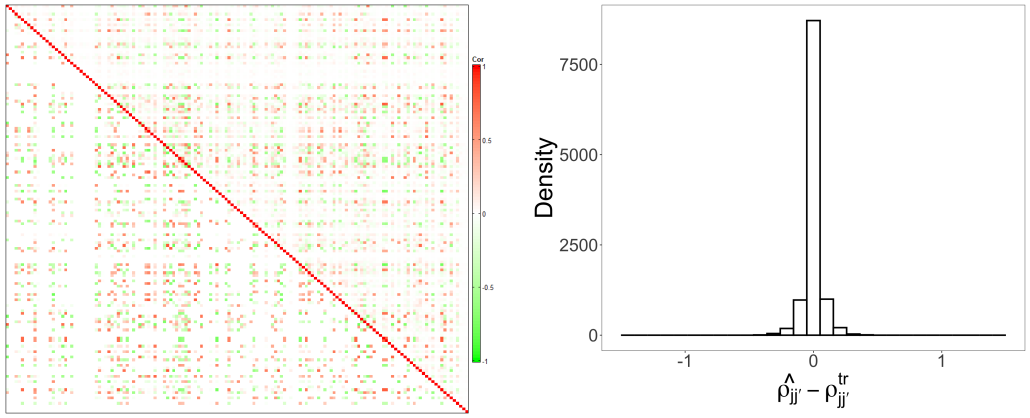


Figure A.10: [Simulation 3] Traceplots of log likelihood under three different initializations are presented in black, red and blue, respectively.

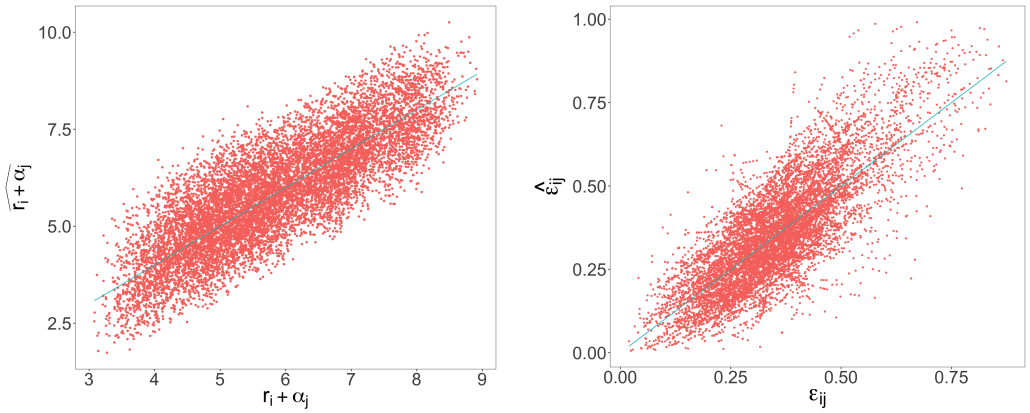


(a) $\rho^{\text{tr}}_{jj'}$ versus $\hat{\rho}_{jj'}$ (b) $\hat{\rho}_{jj'} - \rho^{\text{tr}}_{jj'}$

Figure A.11: [Simulation 3] The upper right and lower left triangles of the heatmap in panel (a) illustrate posterior estimates of correlations $\hat{\rho}_{jj'}$ and their true values $\rho^{\text{tr}}_{jj'}$, respectively. Panel (b) has a histogram of differences between $\hat{\rho}_{jj'}$ and $\rho^{\text{tr}}_{jj'}$.

A.3.3 Simulation 3

We performed an additional simulation study, Simulation 3, where OTU counts are generated from Poisson distributions with correlated means and examined the ro-



(a) $r_i + \alpha_j$ (b) ϵ_{ij}

Figure A.12: [Simulation 3] In panel (a), posterior mean estimates of the mean abundance $r_i+\alpha_j$ are plotted against the simulation truth $r_i^{\text{tr}}+\alpha_j^{\text{tr}}$. Panel (b) has a histogram of the differences in posterior mean estimates $\hat{\epsilon}_{ij}$ of probabilities of an OTU being absent and their true values $\epsilon_{ij}^{\text{tr}}$.

Table A.1: [Simulation 3: Comparison] RMSEs are computed for $\rho_{jj'}, j < j'$, δ_{ij}, μ_{ij}, $\beta_{j2} - \beta_{j1}$, β_{j3} and κ_{jp} under ZI-MLN and comparators.

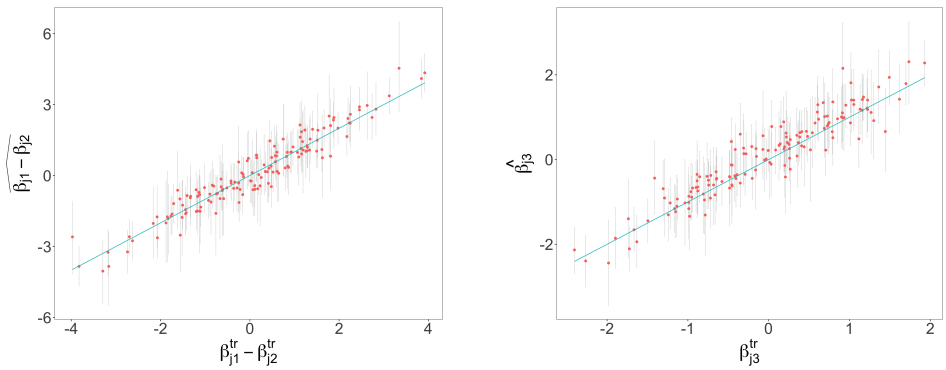
Model	$\rho_{jj'}$
ZI-MLN	**0.064**
SparCC	0.178
SPIEC-EASI	0.158
CCLasso	0.154
Zi-LN	0.160

(a) $\rho_{jj'}$

Model	δ_{ij}	μ_{ij}	$\beta_{j2} - \beta_{j1}$	β_{j3}	κ_{j0}	κ_{j1}	κ_{j2}
ZI-MLN	**0.096**	**1.697**	**0.465**	**0.353**	**0.222**	**0.201**	**0.339**
ZI-MLN without Λ	0.115	1.731	0.601	0.397	0.242	0.222	0.350
MetagenomeSeq	0.113	1.913	1.245	0.729	-	-	-
EdgeR	-	3.400	0.952	0.595	-	-	-

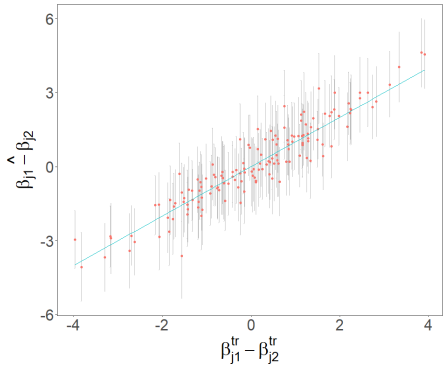
(b) δ_{ij}, μ_{ij}, $\beta_{j2} - \beta_{j1}$, β_{j3} and κ_{jp}

bustness of our ZI-MLN. We kept most of the setup of Simulation 2 the same for Simulation 3; we assumed the number of OTUs $J = 150$ and the number of subjects $M = 35$ assuming two samples from each subject (so the number of samples $N = 70$). We simulated λ_{jk}^{tr} with $K^{\text{tr}} = 5$ assuming sparsity level $g = 0.8$ for joint sparsity, and set $\sigma^{2,\text{tr}} = 1$ and $v_s^{2,\text{tr}} = 1$. We included a pair of dummy variables $(x_{i1}, x_{i2}) \in \{(1,0),(0,1)\}$ to represent a binary covariate, and a continuous covariate, x_{i3} generated from $\text{N}(0,1)$.

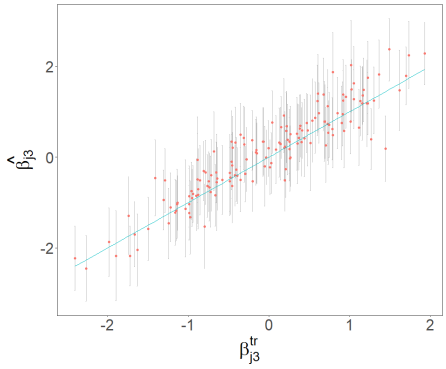


(a) $\beta_{j1} - \beta_{j2}$ under ZI-MLN

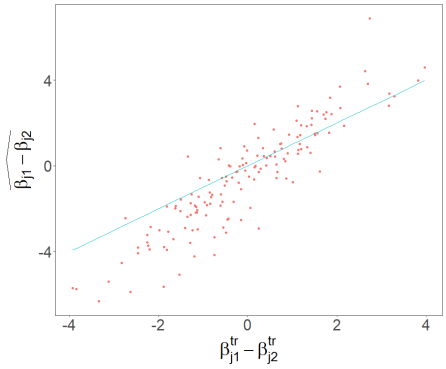
(b) β_{j3} under ZI-MLN



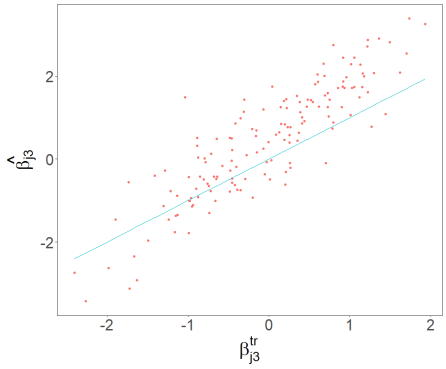
(c) $\beta_{j1} - \beta_{j2}$ under ZI-MLN without Λ



(d) β_{j3} under ZI-MLN without Λ

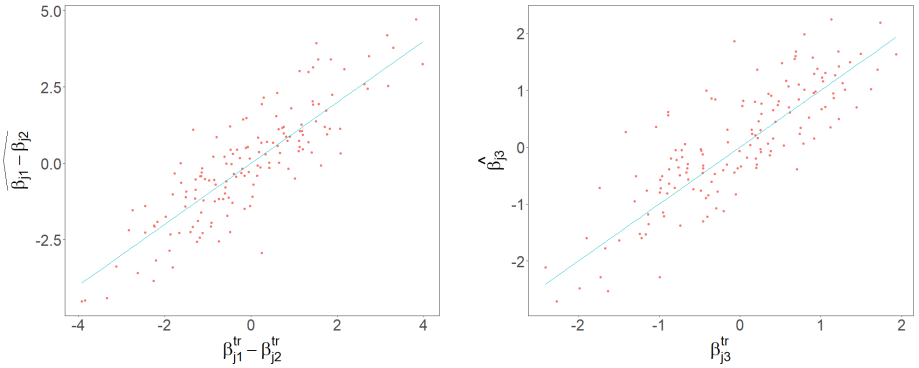


(e) $\beta_{j1} - \beta_{j2}$ under metagemoneSeq



(f) β_{j3} under metagemoneSeq

Figure A.13: [Simulation 3] Estimates $\widehat{\beta_{j1} - \beta_{j2}}$ and $\hat{\beta}_{j3}$ of regression coefficients are compared to the truth $\beta^{\text{tr}}_{j1} - \beta^{\text{tr}}_{j2}$ and β^{tr}_{j3}. Estimates in the four rows are obtained by ZI-MLN, ZI-MLN without Λ, metagenomeSeq and edgeR, respectively. The left and right columns are for $\widehat{\beta_{j1} - \beta_{j2}}$ and $\hat{\beta}_{j3}$, respectively.



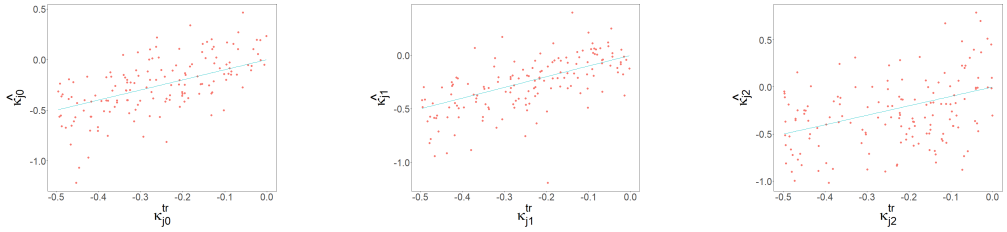
(g) $\widehat{\beta_{j1} - \beta_{j2}}$ under edgeR (h) $\hat{\beta}_{j3}$ under edgeR

Figure A.14: Fig A.13 continued [Simulation 3] Estimates of regression coefficients $\widehat{\beta_{j1} - \beta_{j2}}$ and $\hat{\beta}_{j3}$ are compared to the truth $\beta^{\text{tr}}_{j1} - \beta^{\text{tr}}_{j2}$ and β^{tr}_{j3}. Estimates in the four rows are obtained by ZI-MLN, ZI-MLN without Λ, metagenomeSeq and edgeR, respectively. The left and right columns are for $\widehat{\beta_{j1} - \beta_{j2}}$ and $\hat{\beta}_{j3}$, respectively.

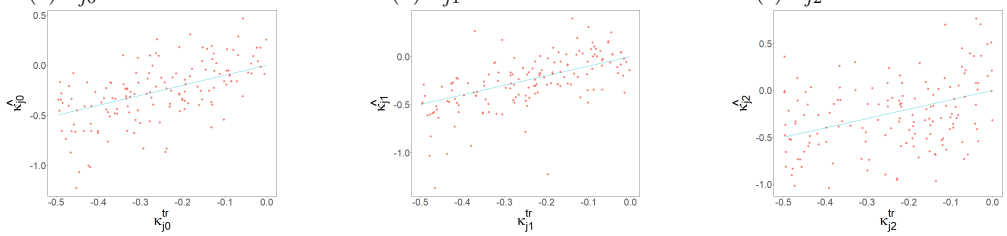
We have $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})$ for the mean abundance μ_{ij} and $\tilde{\mathbf{x}}_i = (x_{i2}, x_{i3})$ for the probability ϵ_{ij} of an OTU being absent. We then simulated sample size factors r^{tr}_i, OTU size factors α^{tr}_j, subject-specific random effects $s^{\text{tr}}_{g_i,j}$, regression coefficients for mean abundances β^{tr}_{jp} and regression coefficients for zero inflation κ^{tr}_{jp}, the same as done in Simulation 2. We finally generated counts Y_{ij} for Poisson distributions as follows;

$$\begin{aligned} \delta^{\text{tr}}_{ij} \mid \epsilon^{\text{tr}}_{ij} &\overset{indep}{\sim} \text{Ber}(\epsilon^{\text{tr}}_{ij}), \quad \text{where } \epsilon^{\text{tr}}_{ij} = \Phi_1(\kappa^{\text{tr}}_{j0} + \tilde{\mathbf{x}}'_i \boldsymbol{\kappa}^{\text{tr}}_j \mid 0, 1), \\ \tilde{\boldsymbol{\mu}}^{\text{tr}}_i &\overset{iid}{\sim} \text{N}_J(r^{\text{tr}}_i \mathbf{1}_J + \boldsymbol{\alpha}^{\text{tr}} + \mathbf{s}^{\text{tr}}_{g_i} + \boldsymbol{\beta}^{\text{tr}} \mathbf{x}_i, \Sigma^{\text{tr}}), \end{aligned}$$

$$\begin{cases} y_{ij} \mid \tilde{\mu}^{\text{tr}}_{ij} \overset{indep}{\sim} \text{Poi}(\exp(\tilde{\mu}^{\text{tr}}_{ij})) & \text{if } \delta^{\text{tr}}_{ij} = 0, \\ y_{ij} = 0 & \text{if } \delta^{\text{tr}}_{ij} = 1, \end{cases}$$



(a) κ_{j0} under ZI-MLN (b) κ_{j1} under ZI-MLN (c) κ_{j2} under ZI-MLN

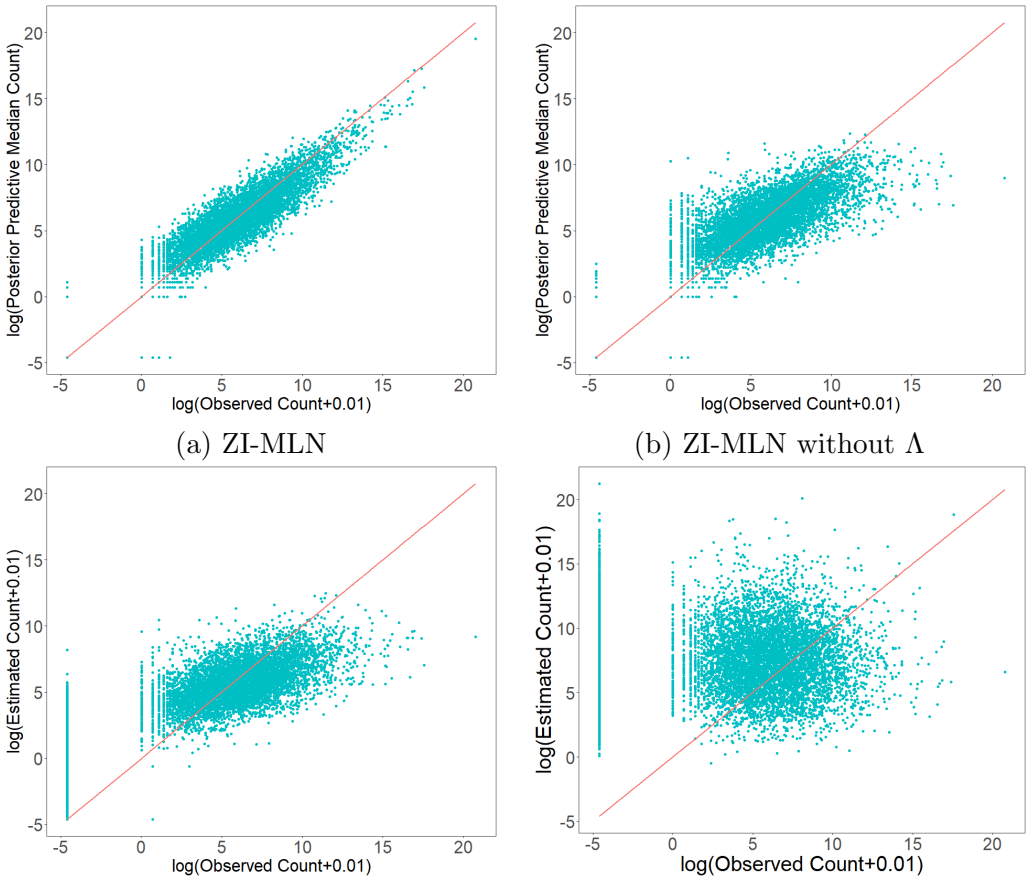


(a) κ_{j0} under ZI-MLN without Λ (b) κ_{j1} under ZI-MLN without Λ (c) κ_{j2} under ZI-MLN without Λ

Figure A.15: [Simulation 3] Posterior mean estimates $\hat{\kappa}_{jp}$ of coefficients on ϵ_{ij} are plotted against the truth for $p = 0, 1, 2$. Estimates in the first and second rows are obtained from ZI-MLN and ZI-MLN without Λ, respectively.

where $\Sigma^{\text{tr}} = \Lambda^{\text{tr}}(\Lambda^{\text{tr}})' + \sigma^{2,\text{tr}}\text{I}_J$. Here, $\boldsymbol{\beta}^{\text{tr}}$ is the $J \times P$ matrix of the true β_{jp}^{tr}. To fit the model, we specified the fixed hyperparameter values similar to those in Simulation 2 and approximated the posterior distribution using MCMC. The MCMC simulation was run for 30,000 iterations, discarding the first 15,000 iterations. Reasonable convergence was achieved and the chain mixed well from checking traceplots and auto-correlation plots. Fig A.10 shows traceplots of the log-likelihood from MCMC chains under different random seeds and initializations. The MCMC converges to similar log-likelihood ranges, showing no evidence of poor mixing or convergence problem.

The results are shown in Figs A.11-A.16. Fig A.11 compares posterior mean estimates of the marginal correlation $\rho_{jj'}$ to the truth. The figure shows that the dependency structure is well recovered. Fig A.12(a) shows a scatter plot of posterior mean estimates $\widehat{r_i + \alpha_j}$ of mean abundances compared to the truth $r_i^{\text{tr}} + \alpha_j^{\text{tr}}$. In panel (b) of the

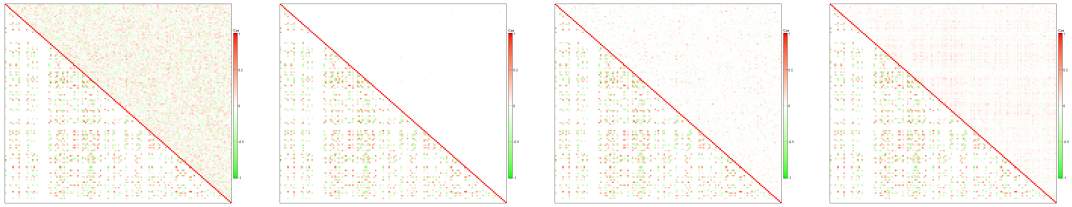


(a) ZI-MLN (b) ZI-MLN without Λ

(c) MetagenomeSeq (d) EdgeR

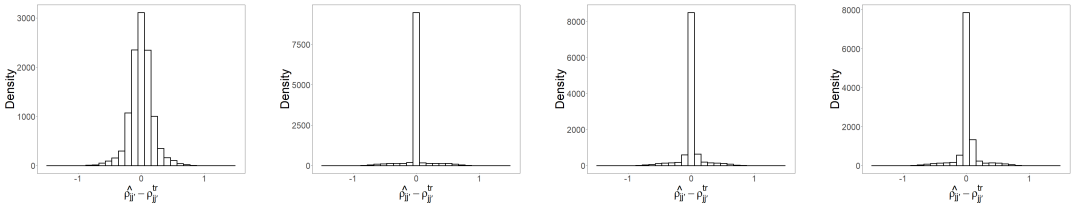
Figure A.16: [Simulation 3] Panels (a) and (b) compare posterior predictive median counts to their observed counts on the logarithm scale, $\log(y_{ij}+0.01)$ versus $\log(\hat{y}_{ij}^{\text{pred}}+0.01)$. $\hat{y}_{ij}^{\text{pred}}$ are estimated with ZI-MLN with Λ in (a) and without Λ in (b). Panels (c) and (d) present scatter plots of observed $\log(y_{ij}+0.01)$ versus $\log(\hat{\mu}_{ij}+0.01)$ by metagenomeSeq and edgeR, respectively, where $\hat{\mu}_{ij}$ are estimated mean abundances of OTUs in samples.

figure, posterior estimates of ϵ_{ij} are compared to the truth $\epsilon_{ij}^{\text{tr}}$. Posterior mean estimates of regression coefficients $\beta_{j1}-\beta_{j2}$ and β_{j3} are compared to their truth in Fig A.13(a) and (b). Posterior mean estimates of probit regression coefficients κ_{jp} are compared to their truth in Fig A.15(a)-(c). In Fig A.16(a), posterior predictive median estimates of OTU counts are plotted against the observed counts for model checking. Overall,



(a) SparCC (b) SPIEC-EASI (c) CCLasso (d) Zi-LN

Figure A.17: [Simulation 3: Comparison] The upper right and lower left triangles of each heatmap illustrate estimates $\hat{\rho}_{jj'}$ of correlations between OTUs and their true values $\rho^{\text{tr}}_{jj'}$. Panels (a)-(d) are for SparCC, SPIEC-EASI, CCLasso and Zi-LN, respectively.



(a) SparCC (b) SPIEC-EASI (c) CCLasso (d) Zi-LN

Figure A.18: [Simulation 3: Comparison] A histogram of differences between $\hat{\rho}_{jj'}$ under SparCC, SPIEC-EASI, CCLasso and ZiLN and $\rho^{\text{tr}}_{jj'}$, in panels (a)-(d), respectively.

the underlying data generation structure is reasonably well approximated although the simulation truth is greatly different from the assumption that ZI-NNL makes, and the model provides a reasonable fit to the data.

We also applied our comparators, SparCC, SPIEC-EASI, CCLasso and Zi-LN, to the simulated dataset. Estimates $\hat{\rho}_{jj'}$ of the correlations under the comparators are compared to the truth $\rho^{\text{tr}}_{jj'}$ in Figs A.17 and A.18. The comparators fail to recover the true dependence structure between OTUs. The RMSE computed for $\rho_{jj'}$ is shown in Tab A.1(a) for all methods in comparison including ZI-MLN. The RMSE under ZI-MLN is much smaller than those under the comparators. It is possibly because the comparators do not attempt to estimate covariate effects on OTU abundances. Addi-

Table A.2: [Simulation 4: Comparison] RMSEs are computed for $\rho_{jj'}, j < j'$, δ_{ij}, μ_{ij}, $\beta_{j2} - \beta_{j1}$, β_{j3} and κ_{jp} under ZI-MLN and comparators.

Model	$\rho_{jj'}$
ZI-MLN	**0.0011**
SparCC	0.128
SPIEC-EASI	0.0042
CCLasso	0.053
Zi-LN	0.051

(a) $\rho_{jj'}$

Model	δ_{ij}	μ_{ij}	$\beta_{j2} - \beta_{j1}$	β_{j3}	κ_{j0}	κ_{j1}	κ_{j2}
ZI-MLN	**0.052**	**0.783**	**0.325**	**0.234**	**0.223**	**0.170**	**0.170**
ZI-MLN without Λ	**0.052**	**0.783**	0.370	0.241	**0.223**	**0.170**	**0.170**
metagenomeSeq	0.078	2.161	1.092	0.711	-	-	-
edgeR	-	1.491	0.741	0.452	-	-	-

(b) δ_{ij}, μ_{ij}, $\beta_{j2} - \beta_{j1}$, β_{j3} and κ_{jp}

tional comparators, ZI-MLN without Λ, metagenomeSeq and edgeR were also applied to compare estimates of the covariate effects. The RMSEs are computed for the parameters δ_{ij}, μ_{ij}, $\beta_{j2} - \beta_{j1}$, β_{j3} and κ_{jp} and summarized in Tab A.1(b). Our ZI-MLN outperforms the comparators even when the counts are generated from Poisson distributions. Estimates of regression coefficients $\beta_{j1} - \beta_{j2}$ and β_{j3} obtained from the comparators are compared to their truth in Figs A.13(c)-(h). Figs A.15(d)-(f) illustrate a comparison of posterior mean estimates of probit regression coefficients κ_{jp} under ZI-MLN without Λ to the truth. Predictive model checking for ZI-MLN without Λ is reported in Fig A.16(b) by comparing their posterior predictive median estimates $\hat{y}_{ij}^{\text{pred}}$ of OTU counts to the observed counts. Estimates of mean abundance levels under metagenomeSeq and edgeR are plotted against the observed counts in Figs A.16(c) and (d), respectively. The comparison indicates that ignoring the interrelationship between OTUs may distort inferences on mean abundances and the absence/presence of OTUs. Also, comparison of ZI-MLN to edgeR shows that ignoring excess zeros may lead to poor estimation of mean abundance levels.

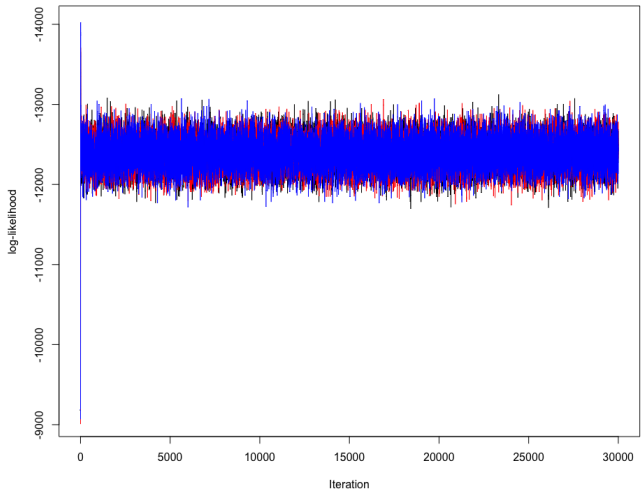
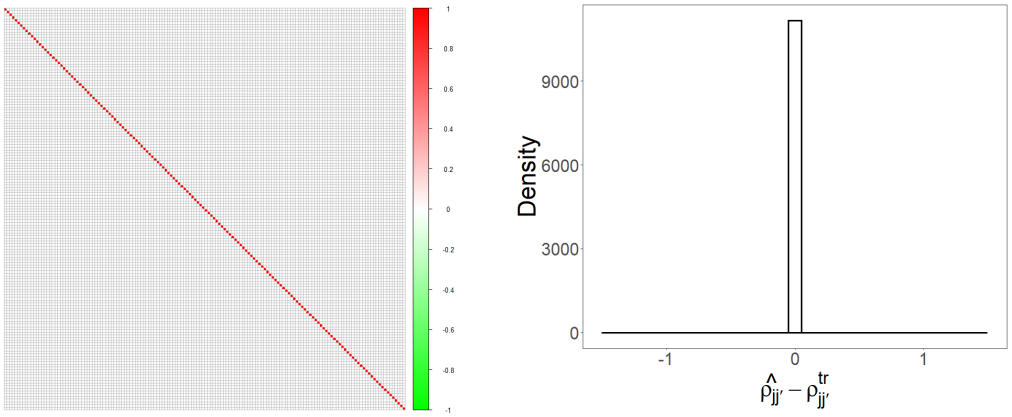


Figure A.19: [Simulation 4] Traceplots of log likelihood under three different initializations are presented in black, red and blue, respectively.

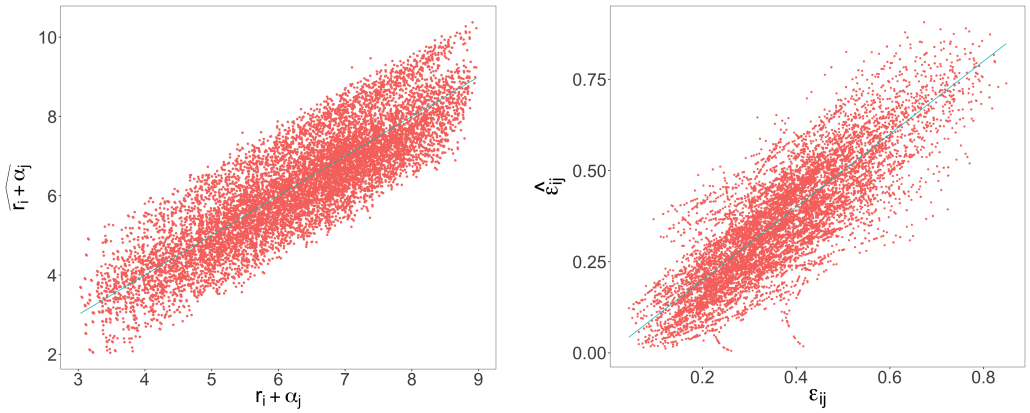


(a) $\rho^{\text{tr}}_{jj'}$ versus $\hat{\rho}_{jj'}$ (b) $\hat{\rho}_{jj'} - \rho^{\text{tr}}_{jj'}$

Figure A.20: [Simulation 4] The upper right and lower left triangles of the heatmap in panel (a) illustrate posterior estimates of correlations $\hat{\rho}_{jj'}$ and their true values $\rho^{\text{tr}}_{jj'}$, respectively. Panel (b) has a histogram of differences between $\hat{\rho}_{jj'}$ and $\rho^{\text{tr}}_{jj'}$.

A.3.4 Simulation 4

In Simulation 4, we considered a dataset without any dependency structure between OTUs. We let $\lambda^{\text{tr}}_{jk} = 0$ for all (j, k) and had $\Sigma^{\text{tr}} = \sigma^{2,\text{tr}} \text{I}_J$ with $\sigma^{2,\text{tr}} =$

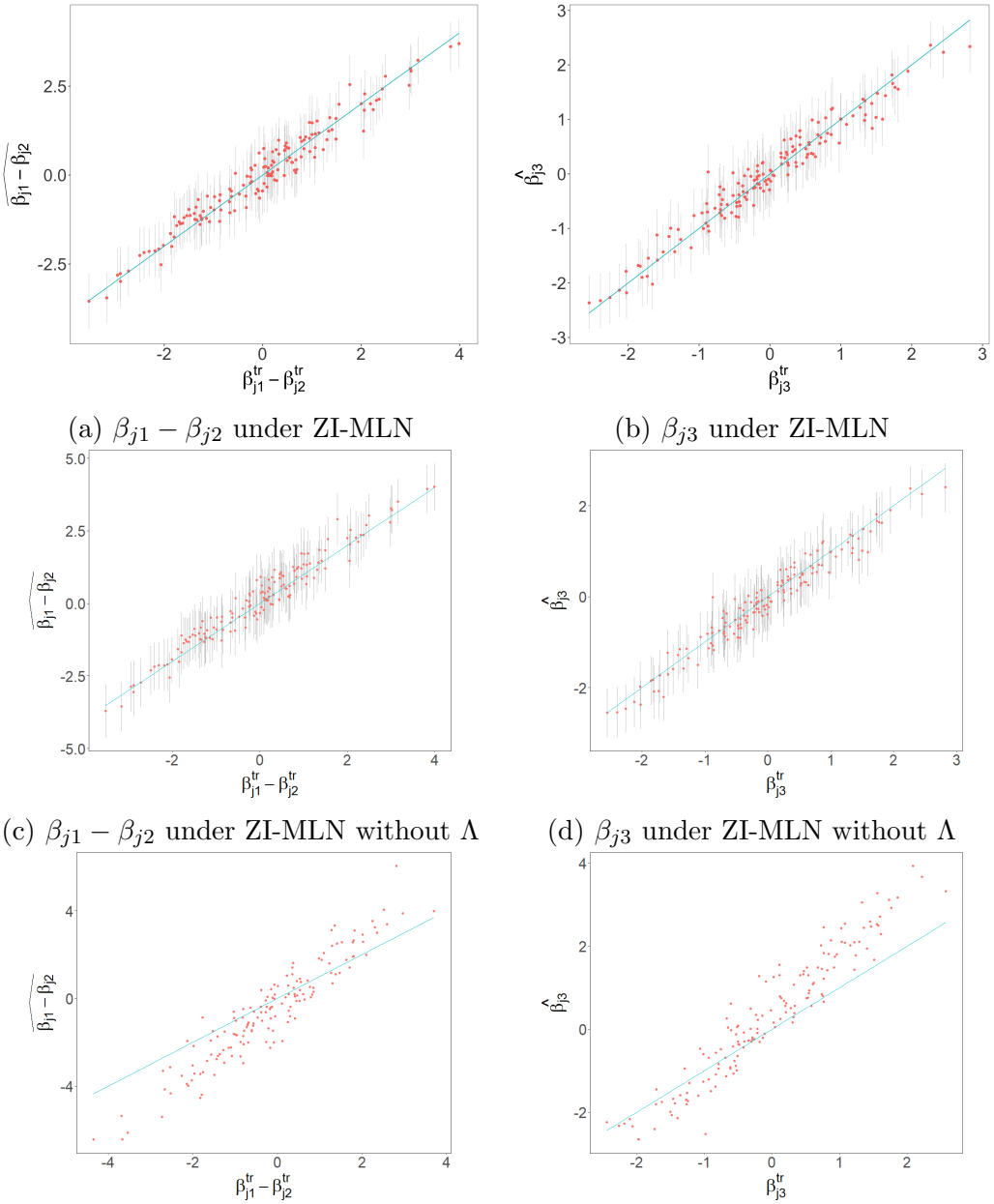


(a) $r_i + \alpha_j$ (b) ϵ_{ij}

Figure A.21: [Simulation 4] In panel (a), posterior mean estimates of the mean abundance $r_i+\alpha_j$ are plotted against the simulation truth $r_i^{\text{tr}}+\alpha_j^{\text{tr}}$. Panel (b) has a histogram of the differences in posterior mean estimates $\hat{\epsilon}_{ij}$ of probabilities of an OTU being absent and their true values $\epsilon_{ij}^{\text{tr}}$.

1. We kept the remaining simulation setup the same as in Simulation 2. The fixed hyperparameters are set the same as in Simulation 2 to fit the model, and the posterior samples were drawn from the posterior distribution via MCMC. We discarded the first 15,000 iterates for burn-in and kept the next 15,000 iterates for posterior inference. We examined the mixing and convergence of the Markov chains using trace plots and did not find evidence of poor mixing or bad convergence. For example, Fig A.19 shows traceplots of the log-likelihood from MCMC chains under different random seeds and initializations and does not indicate evidence of poor mixing or convergence problem.

Posterior inference is summarized in Figs A.20-A.25. Fig A.20 shows that our posterior mean estimates of marginal correlations are close to the truth, $\rho_{jj}^{\text{tr}} = 1$ and $\rho_{jj'}^{\text{tr}} = 0$, $j \neq j'$. Posterior mean estimates of σ^2 and u_s^2 are $(1.002, 0.922)$, which are close to their true values of 1. Also, we check the estimation of the mean abundances.



(a) $\beta_{j1} - \beta_{j2}$ under ZI-MLN

(b) β_{j3} under ZI-MLN

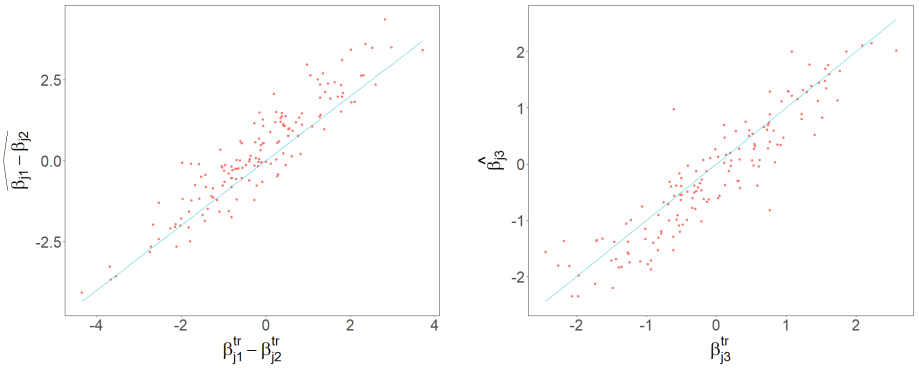
(c) $\beta_{j1} - \beta_{j2}$ under ZI-MLN without Λ

(d) β_{j3} under ZI-MLN without Λ

(e) $\beta_{j1} - \beta_{j2}$ under metagemoneSeq

(f) β_{j3} under metagemoneSeq

Figure A.22: [Simulation 4] Estimates $\widehat{\beta_{j1} - \beta_{j2}}$ and $\hat{\beta}_{j3}$ of regression coefficients are compared to the truth $\beta^{\text{tr}}_{j1} - \beta^{\text{tr}}_{j2}$ and β^{tr}_{j3}. Estimates in the four rows are obtained by ZI-MLN, ZI-MLN without Λ, metagenomeSeq and edgeR, respectively. The left and right columns are for $\widehat{\beta_{j1} - \beta_{j2}}$ and $\hat{\beta}_{j3}$, respectively.

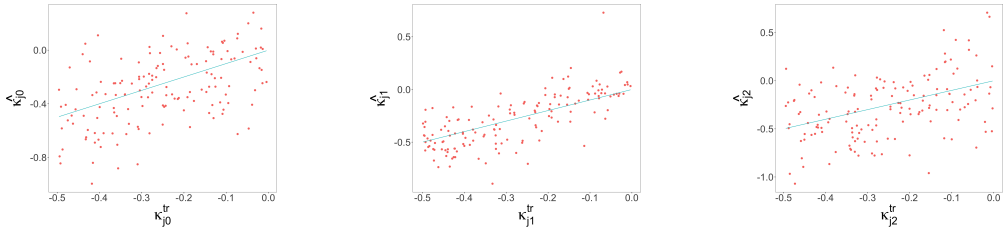


(g) $\beta_{j1} - \beta_{j2}$ under edgeR (h) β_{j3} under edgeR

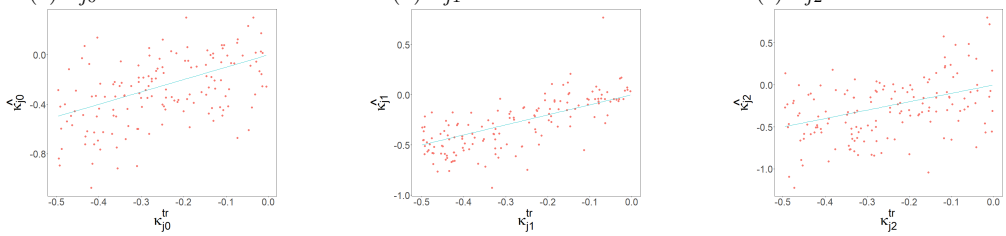
Figure A.23: Fig A.22 continued [Simulation 4] Estimates of regression coefficients $\widehat{\beta_{j1} - \beta_{j2}}$ and $\hat{\beta}_{j3}$ are compared to the truth $\beta^{\text{tr}}_{j1} - \beta^{\text{tr}}_{j2}$ and β^{tr}_{j3}. Estimates in the four rows are obtained by ZI-MLN, ZI-MLN without Λ, metagenomeSeq and edgeR, respectively. The left and right columns are for $\widehat{\beta_{j1} - \beta_{j2}}$ and $\hat{\beta}_{j3}$, respectively.

Specifically, from Fig A.21(a) posterior mean estimates $\widehat{r_i + \alpha_j}$ are tightly around $r^{\text{tr}}_i + \alpha^{\text{tr}}_j$. Fig A.21(b) compares posterior estimates of the probabilities of OTUs being absent, ϵ_{ij} to the truth, indicating reasonable inferences on the absence/presence of OTUs. Figs A.22(a)-(b) and A.24(a)-(c) show that estimates of the regression coefficients are reasonably well estimated. Posterior predictive checking is illustrated in Fig A.25(a). The plot shows that our model provides a good fit even when there is no dependence structure assumed in the truth.

We applied the comparators, SparCC, SPIEC-EASI, CCLasso and Zi-LN, to the simulated data and compared their performance of estimating $\rho_{jj'}$ to that of our ZI-MLN. Figs A.26 and A.27 compare the estimates of $\rho_{jj'}$ under the comparators to the truth. The RMSE of $\rho_{jj'}$ is computed and given in Tab A.2(a). ZI-MLN outperforms the other methods in comparison for estimating $\rho_{jj'}$. SparCC yields estimates not close to zero for many $\rho_{jj'}$ and yields the largest RMSE for this simulated dataset. The addi-



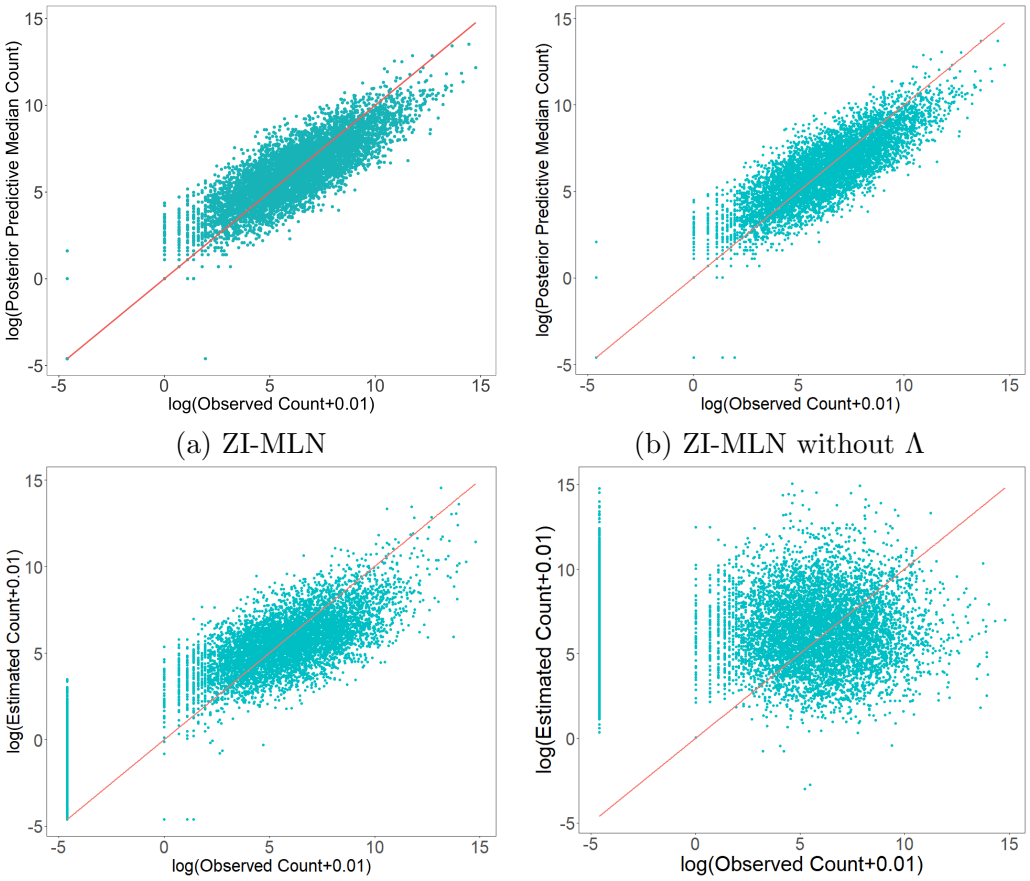
(a) κ_{j0} under ZI-MLN (b) κ_{j1} under ZI-MLN (c) κ_{j2} under ZI-MLN



(b) κ_{j0} under ZI-MLN without Λ (b) κ_{j1} under ZI-MLN without Λ (c) κ_{j2} under ZI-MLN without Λ

Figure A.24: [Simulation 4] Posterior mean estimates $\hat{\kappa}_{jp}$ of coefficients on ϵ_{ij} are plotted against the truth for $p = 0, 1, 2$. Estimates in the first and second rows are obtained from ZI-MLN and ZI-MLN without Λ, respectively.

tional comparators, ZI-MLN without Λ, metagenomeSeq and edgeR were applied to the dataset for further comparison. RMSEs for the parameters, δ_{ij}, μ_{ij}, $\beta_{j2} - \beta_{j1}$, β_{j3} and κ_{jp}, are computed and listed in Tab A.2(b). ZI-MLN performs the best, very closely followed by ZI-MLN without Λ, or the same as ZI-MLN without Λ although the simulation truth is closer to the assumption made under ZI-MLN without Λ. Also, the truth is close to the assumption that metagenomeSeq, but RMSE of mean abundances μ_{ij} is large compared to those under the other methods. Figs A.22(c)-(h) compare estimates of the regression coefficients on OTU abundances under the comparators. Figs A.24(d)-(f) show posterior mean estimates of κ_{jp} under ZI-MLN without Λ. Posterior predictive checking is shown in Fig A.25(b) for ZI-MLN without Λ. Estimates of mean abundances under metagenomeSeq and edgeR are shown in Figs A.25(c) and (d), respectively. The results indicate that there is no degradation in the performance of ZI-MLN in a case

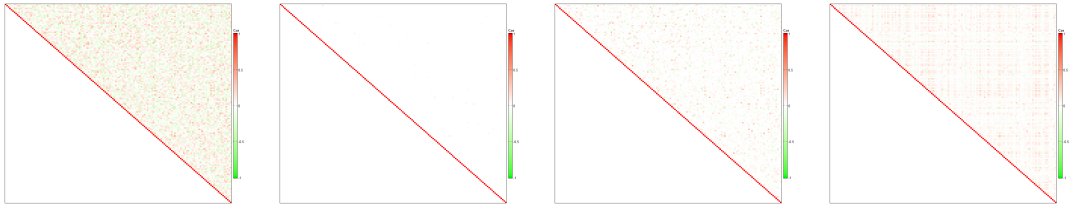


(a) ZI-MLN (b) ZI-MLN without Λ

(c) MetagenomeSeq (d) EdgeR

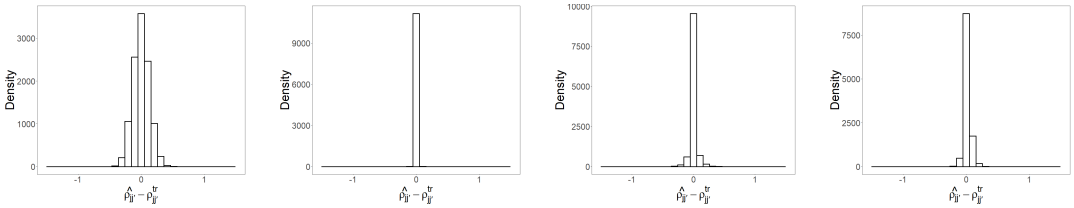
Figure A.25: [Simulation 4] Panels (a) and (b) compare posterior predictive median counts to their observed counts on the logarithm scale, $\log(y_{ij}+0.01)$ versus $\log(\hat{y}_{ij}^{\text{pred}}+0.01)$. $\hat{y}_{ij}^{\text{pred}}$ are estimated with ZI-MLN with Λ in (a) and without Λ in (b). Panels (c) and (d) present scatter plots of observed $\log(y_{ij}+0.01)$ versus $\log(\hat{\mu}_{ij}+0.01)$ by metagenomeSeq and edgeR, respectively, where $\hat{\mu}_{ij}$ are estimated mean abundances of OTUs in samples.

where there is no dependence in the data generating process. Note that Simulations 1-3 indicate the additional flexibility of ZI-MLN allows to outperform the comparators when data has a dependence structure.



(a) SparCC (b) SPIEC-EASI (c) CCLasso (d) Zi-LN

Figure A.26: [Simulation 4: Comparison] The upper right and lower left triangles of each heatmap illustrate estimates $\hat{\rho}_{jj'}$ of correlations between OTUs and their true values $\rho^{\text{tr}}_{jj'}$. Panels (a)-(d) are for SparCC, SPIEC-EASI, CCLasso and Zi-LN, respectively.



(a) SparCC (b) SPIEC-EASI (c) CCLasso (d) Zi-LN

Figure A.27: [Simulation 4: Comparison] A histogram of differences between $\hat{\rho}_{jj'}$ under SparCC, SPIEC-EASI, CCLasso and ZiLN and $\rho^{\text{tr}}_{jj'}$, in panels (a)-(d), respectively.

Table A.3: [Simulation 5: Comparison] RMSEs are computed for $\rho_{jj'}, j < j'$, δ_{ij} and $\tilde{\mu}_{ij}$ under ZI-MLN and comparators. $\tilde{\mu}_{ij}$ is the mean abundance adjusted by a sample total count.

Model	$\rho_{jj'}$
ZI-MLN	**0.034**
SparCC	0.253
SPIEC-EASI	**0.034**
CCLasso	0.142
Zi-LN	0.048

(a) $\rho_{jj'}$

Model	δ_{ij}	$\tilde{\mu}_{ij}$
ZI-MLN	**0.070**	**0.917**
ZI-MLN without Λ	0.071	0.917
metagenomeSeq	0.091	1.922

(b) δ_{ij} and $\tilde{\mu}_{ij}$

A.3.5 Simulation 5

In this simulation study, we used SparseDOSSA (Ma et al., 2021) to generate a dataset. An open-source software *SparseDOSSA2* is available from the authors'

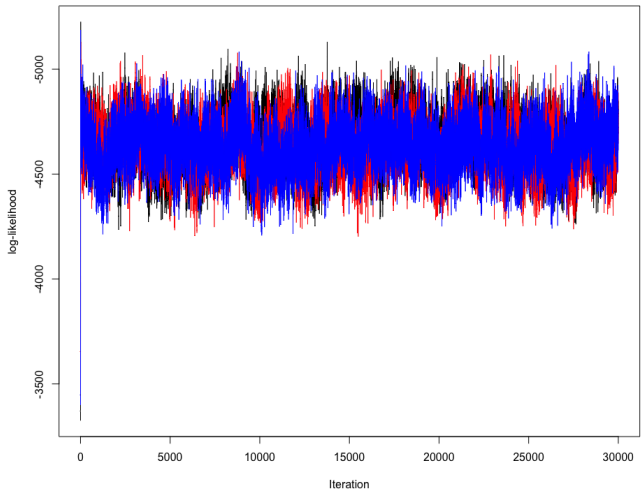
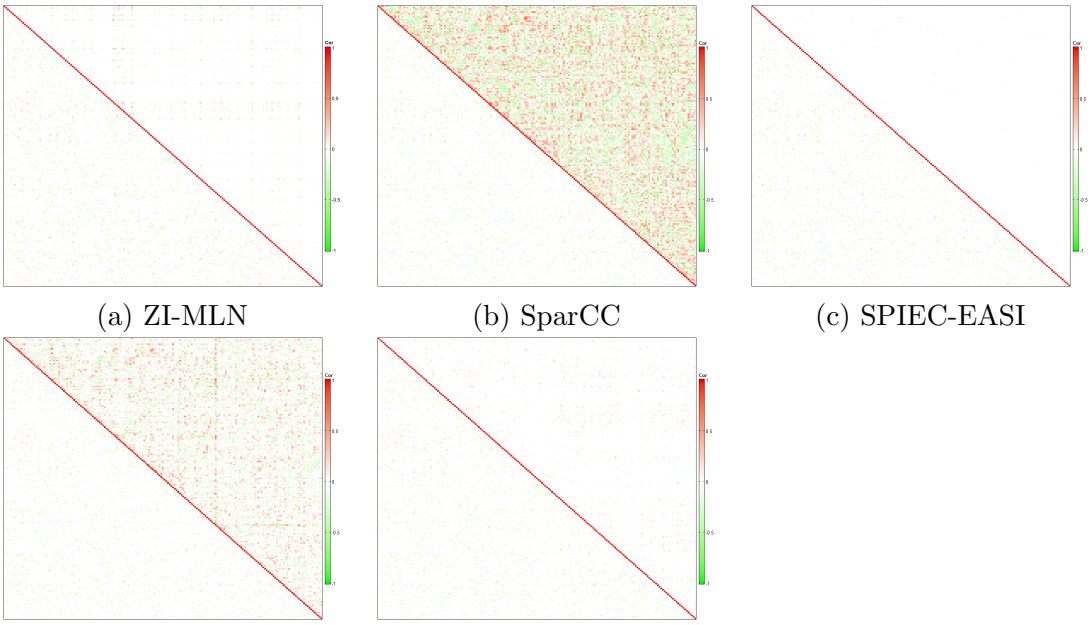


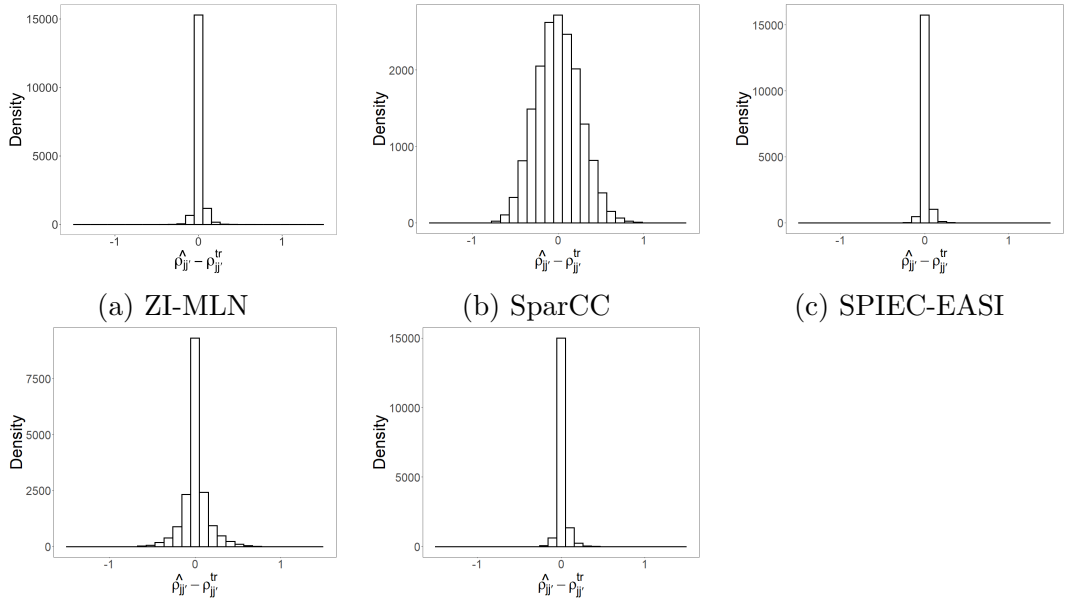
Figure A.28: [Simulation 5] Traceplots of log likelihood under three different initializations are presented in black, red and blue, respectively.



(a) ZI-MLN (b) SparCC (c) SPIEC-EASI

(d) CCLasso (e) Zi-LN

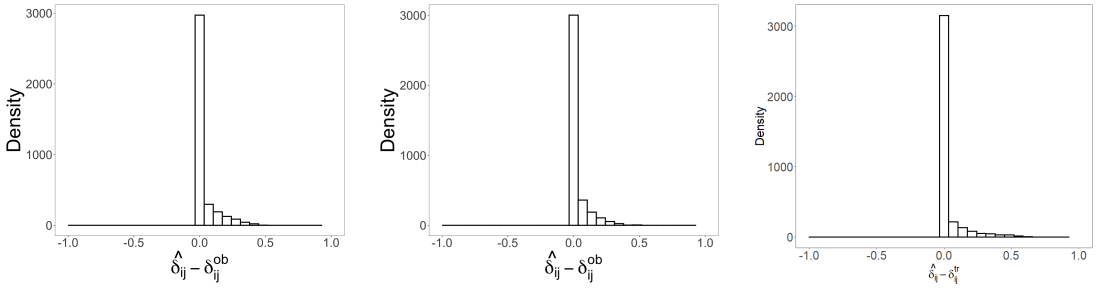
Figure A.29: [Simulation 5] The upper right and lower left triangles of each heatmap illustrate estimates $\hat{\rho}_{jj'}$ of correlations between OTUs and their true values $\rho^{\text{tr}}_{jj'}$. Panels (a)-(e) are for ZI-MLN, SparCC, SPIEC-EASI, CCLasso and Zi-LN, respectively.



(a) ZI-MLN (b) SparCC (c) SPIEC-EASI

(d) CCLasso (e) Zi-LN

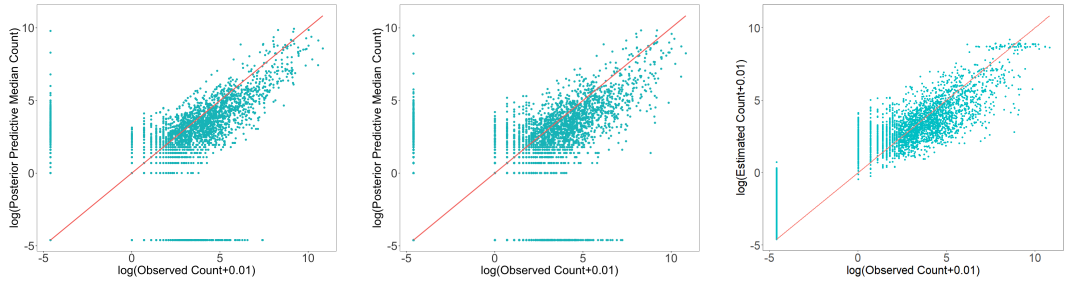
Figure A.30: [Simulation 5] A histogram of differences between $\hat{\rho}_{jj'}$ under ZI-MLN, SparCC, SPIEC-EASI, CCLasso and ZiLN and $\rho^{\text{tr}}_{jj'}$, in panels (a)-(e), respectively.



(a) ZI-MLN (b) ZI-MLN without Λ (c) metagenomeSeq

Figure A.31: [Simulation 5] Histograms of the differences between $\hat{\delta}_{ij}$ and the observed zero indicator $1(Y_{ij} = 0)$ under ZI-MLN, ZI-MLN without Λ and metagenomeSeq, respectively. Posterior mean estimates are used for ZI-MLN and ZI-MLN without Λ.

webpage. SparseDOSSA takes a real microbiome dataset as an input and generates a realistic microbiome dataset. Generated OTU counts in a sample are constrained to sum up to a constant (compositionality), are enriched for zero counts (zero-inflated), and are dependent due to microbial interactions. SparseDOSSA assumes a Gaussian copula



(a) ZI-MLN (b) ZI-MLN without Λ (c) metagenomeSeq

Figure A.32: [Simulation 5] Panels (a) and (b) have scatter plots of observed $\log(y_{ij} + 0.01)$ versus $\log(\hat{y}_{ij}^{\text{pred}} + 0.01)$ under ZI-MLN and ZI-MLN without Λ, respectively. Panel (c) is the scatter plots of observed $\log(y_{ij} + 0.01)$ versus mean abundance estimates $\log(\hat{\mu}_{ij} + 0.01)$ by metagenomeSeq.

model with a zero-inflated log-normal distribution for latent absolute (unnormalized) OTU abundances and generates OTU count vectors from a multinomial distribution with the relative abundances normalized from the absolute abundances. For the multinomial distribution, sample total counts are independently simulated from a log-normal distribution. SparseDOSSA does not include random effects, and the same relative abundance vector is assumed for the samples. Thus, a simulated dataset may not exhibit overdispersion that ZI-MLN accounts for through subject group factor effects $\boldsymbol{s}_{g_i}$. SparseDOSSA estimates the model parameters using an input dataset, e.g., the mean vector and precision matrix of the absolute abundance vector, and then sets the true input parameter values at its estimates. The parameters are estimated by an EM algorithm. In particular, its precision matrix is estimated with a ℓ_1 penalty function for sparsity. The correlation matrix estimated from SparseDOSSA is shown in the lower triangles of the heatmaps in Fig A.29. The data-generating process under SparseDOSSA is greatly different from that assumed for ZI-MLN. We used the skin microbiome dataset

in §2.4.1 of the main text as an input dataset. The dataset has $N = 20$ samples and $J = 187$ OTUs. Similar to that of the skin microbiome data, the simulated dataset has zeros for 30% of the counts.

To fit ZI-MLN, we specified the fixed hyperparameter values similar to those in the previous simulation studies. For this simulation study, we set $a_\phi = 1/10$ for greater sparsity. We ran MCMC for 30,000 iterations with the first 15,000 iterations to burn-in. To examine mixing and convergence, we ran multiple chains under different initial values and random seeds. Fig A.28 shows the chains with different initial values and random seeds converge to similar log-likelihood ranges, indicating no empirical evidence of bad mixing or convergence.

Posterior inferences under ZI-MLN are summarized in Figs A.29(a), A.30(a) and A.31(a). Figs A.29(a) and A.30(a) show that the true underlying between-OTU dependence structure is well recovered although the dataset was generated from a very different model. Especially, the true data-generating process assumes a multimonimal distribution that conditions on sample total counts. The model-based normalization through sample size factors r_i under ZI-MLN accounts for compositionality reasonably well, and the model provides reasonable estimates of $\rho_{jj'}$. From Fig A.31(a), the absence/presence of OTUs is also well estimated. Posterior predicted mean counts are compared to observed counts in Fig A.32(a). The plot indicates that ZI-MLN fits the data well.

For comparison, we applied SparCC, SPIEC-EASI, CCLasso and Zi-LN to the dataset simulated by SparseDOSSA. Note that SPIEC-EASI and CCLasso use ℓ_1

penalty to estimate dependence structure, similar to SparseDOSSA. Estimates $\hat{\rho}_{jj'}$ of the correlations obtained under the comparators are compared to their true values $\rho^{\text{tr}}_{jj'}$ in Figs Figs A.29(b)-(e) and A.30(b)-(e). We also computed the RMSE of $\rho_{jj'}$ under all methods in comparison including ZI-MLN. From Tab A.3(a), ZI-MLN and SPIEC-EASI produce the smallest value of RMSE. It is noticeable that correlation matrix estimates under SparCC and CCLasso are very dense, resulting in very large values of RMSE. Also, we applied the additional comparators, ZI-MLN without Λ and metagenomeSeq, to the dataset. RMSE of δ_{ij} and μ_{ij} are computed and summarized in Tab A.3(b). Since the counts were generated from a multinomial distribution, we adjusted estimates of μ_{ij} by the total sample counts, $\tilde{\mu}_{ij} = \hat{\mu}_{ij} - \log(\sum_j Y_{ij})$, and compared $\tilde{\mu}_{ij}$ to the true normalized abundance of SparseDOSSA. ZI-MLN outperforms ZI-MLN without Λ and metagenomeSeq in estimating the presence/absence of OTUs and their mean abundances. $\hat{\delta}_{ij}$ under ZI-MLN without Λ and metagenomeSeq are compared to the observed zero indicators $1(Y_{ij} = 0)$ in Fig A.31(b) and (c), respectively. Posterior predictive mean counts under ZI-MLN without Λ are plotted against the observed counts in Fig A.32(b). ZI-MLN without Λ yielded a poorer fit to the data than ZI-MLN. The mean abundance estimates under metagenomeSeq are compared to transformed observed counts in Fig A.32(c)

Table A.4: [Skin Microbiome Data] Taxonomic information for the OTUs illustrated in in Fig 9(b) of the main text.

OTU	Taxonomic information (Kingdom/ Phylum/ Class/ Order/ Family / Genus)
41	Bacteria - Actinobacteria - Actinobacteria - Micrococcales - NA - NA
42	Bacteria - Actinobacteria - Actinobacteria - Micrococcales - Micrococcaceae - Glutamicibacter
43	Bacteria - Proteobacteria - Epsilonproteobacteria - Campylobacterales - Campylobacteraceae - Campylobacter
46	Bacteria - Bacteroidetes - Sphingobacteriia - Sphingobacteriales - Chitinophagaceae - uncultured
47	Bacteria - Bacteroidetes - Sphingobacteriia - Sphingobacteriales - Chitinophagaceae - Segetibacter
48	Bacteria - Bacteroidetes - Bacteroidia - Bacteroidales - Porphyromonadaceae - Porphyromonas
76	Bacteria - Proteobacteria - Gammaproteobacteria - Pseudomonadales - Moraxellaceae - Enhydrobacter
88	Bacteria - Firmicutes - Clostridia - Clostridiales - Family XI - Peptoniphilus
92	Bacteria - Firmicutes - Clostridia - Clostridiales - Family XIII - uncultured
138	Bacteria - Proteobacteria - Alphaproteobacteria - Caulobacterales - Caulobacteraceae - Brevundimonas
153	Bacteria - Firmicutes - Bacilli - Lactobacillales - Aerococcaceae - uncultured
173	Bacteria - Actinobacteria - Actinobacteria - Streptomycetales - Streptomycetaceae - Streptomyces

A.4 Additional Results for Real Data Analyses

A.4.1 Additional Results for Skin Microbiome Data Analysis

Multiple MCMC chains were run with different initial values and random seeds to examine the mixing and convergence of the MCMC. Fig A.33 illustrates traceplot of the log-likehood of the MCMC runs and shows no evidence of poor mixing or convergence issues. Tab A.4 has taxonomic information for the OTUs illustrated in Fig 9(b) of the main text. We include the comparison of the observed zero inflation rate against the observed zero indicators $1(Y_{ij} = 0)$ under ZI-MLN, ZI-MLN without Λ and metagenomeSeq in Fig A.34.

To examine robustness to the specification of the threshold used for data pre-

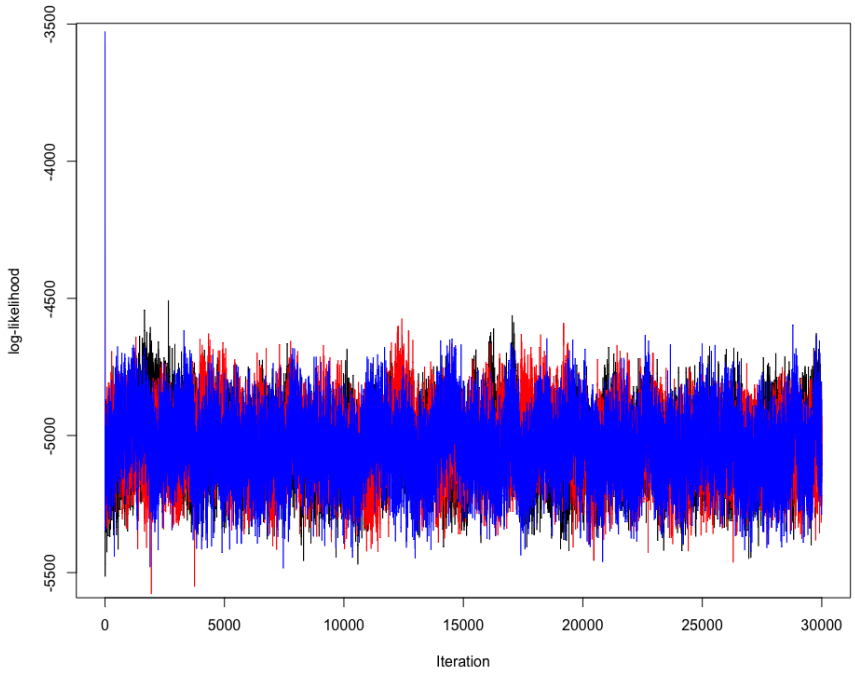
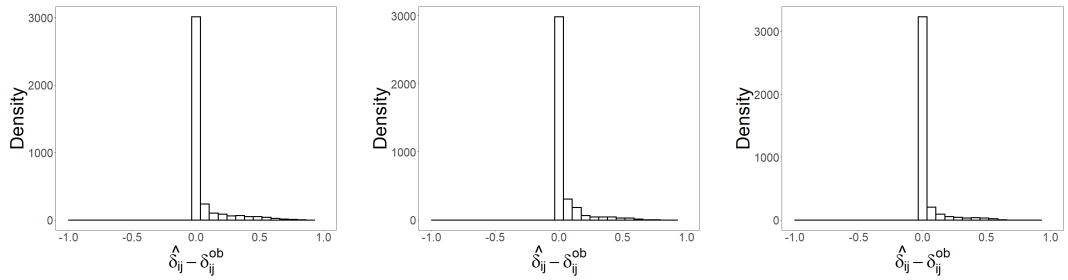


Figure A.33: [Skin Microbiome Data] Traceplots of log likelihood under three different initializations are presented in black, red and blue, respectively.



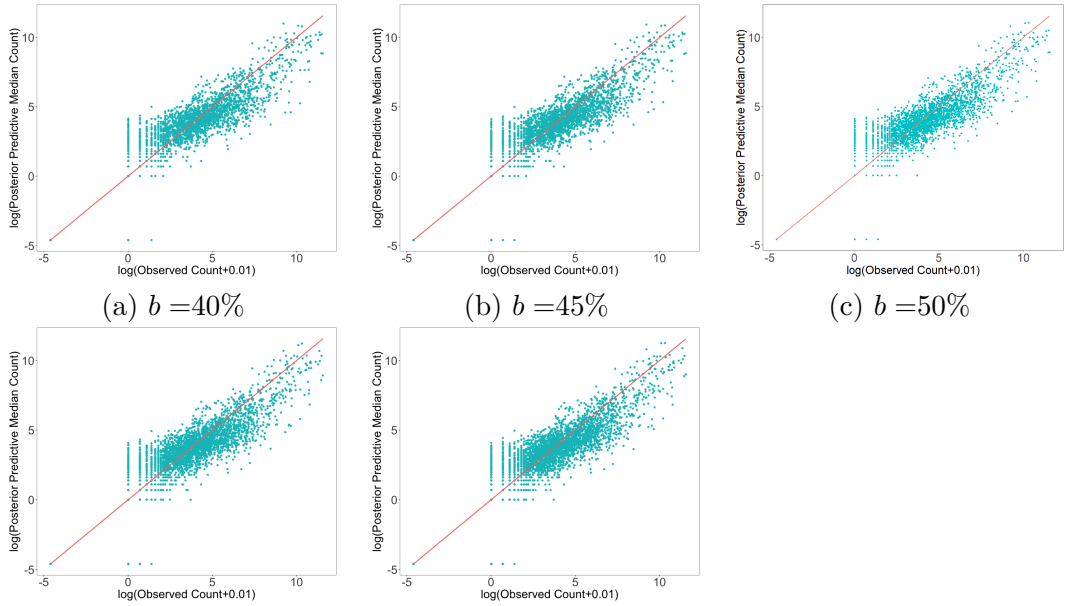
(a) ZI-MLN (b) ZI-MLN without Λ (c) metagenomeSeq

Figure A.34: [Skin Microbiome Data] Histograms of the differences between $\hat{\delta}_{ij}$ and the observed zero indicators $1(Y_{ij} = 0)$ under ZI-MLN, ZI-MLN without Λ and metagenomeSeq, respectively. Posterior mean estimates are used for ZI-MLN and ZI-MLN without Λ.

processing, we performed a sensitivity analysis. We used five different values of the threshold to remove OTUs that have zeros in too many samples from analysis. In particular, OTUs that have zero counts in more than b% of the samples were removed, where 40%, 45%, 50%, 55% and 60% are used for b. Note that b =50% is used for the analysis in §2.4.1 of the main text. The skin microbiome data has a total of 20 samples, and those cutoff values remove OTUs who have zero counts in more than 8, 9, 10, 11 and 12 samples, resulting that 147, 163, 187, 213 and 238 OTUs are included for analysis. We fitted the model to each of the preprocessed datasets and compared posterior inferences and model fit. We used the same hyperparameter values. From the posterior predictive checking illustrated in Fig A.35, we observe that the model provides a good fit to all datasets the mode fit does not change much by the value of b. We also examined correlation estimates for the OTUs that are included in all five preprocessed datasets and compared. Fig A.36 shows the posterior mean estimates $\hat{\rho}_{jj'}$ for seven OTUs that are arbitrarily chosen among the OTUs included in all datasets for illustration. The figure shows that the correlation estimates remain almost unchanged by the value of b, indicating the robustness of the model to the specification of b for preprocessing.

A.4.2 Additional Results from Human Gut Microbiome Data Analysis

Tab A.5 presents the names of the covariates included for human gut microbiome data analysis and their support. The dataset has 37 children subjects collected from two different recruitment sites. The biopsy samples were taken from either of two biopsy locations, ileum or rectum or both locations. The model was run for the



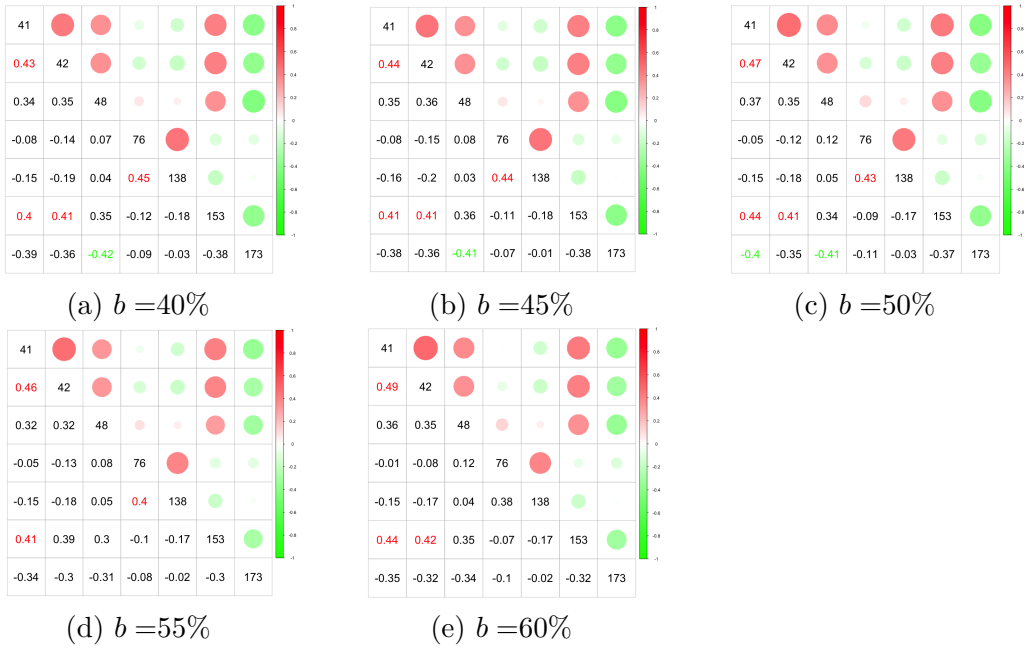
(a) b =40% (b) b =45% (c) b =50%

(d) b =55% (e) b =60%

Figure A.35: [Sensitivity Analysis for the Skin Microbiome Data] Scatter plots of observed $\log(y_{ij} + 0.01)$ versus posterior predictive log count $\log(\hat{y}_{ij}^{\text{pred}} + 0.01)$ estimated by ZI-MLN. Different threshold values are used for data preprocessing. b=40%, 45%, 50%, 55% and 60% are used for panels (a)-(e), respectively.

dataset three times under different initializations and random seeds. Traceplots of the log-likelihood shown in Fig A.37. The plot suggests that the model converged to a similar state under these alternative specifications, and provides practical evidence of the chain's convergence. Fig A.39 illustrates posterior mean estimates of κ_{jp} for two selected covariates, age and binary indicator of a subject being white, where posterior mean estimates are represented with black dots, and 95% credible intervals with vertical lines. κ_{jp} whose credible interval does not contain zero are in red.

Tab A.6 has taxonomic information of the OTUs in Fig 12(b) of the main text. Fig A.38 compares posterior mean estimates $\hat{\delta}_{ij}$ to the observed indicator $1(Y_{ij} = 0)$. Fig A.39 presents posterior estimates of coefficients κ_{jp} of the probit regression for two



(a) b =40% (b) b =45% (c) b =50%

(d) b =55% (e) b =60%

Figure A.36: [Sensitivity Analysis for the Skin Microbiome Data] The posterior mean estimates $\hat{\rho}_{jj'}$ of correlations for seven OTUs. The OTUs are arbitrarily chosen for illustration among the OTUs that are included in datasets preprocessed with different threshold values. The value of a preprecessing threshold, b=40%, 45%, 50%, 55% and 60% are used for panels (a)-(e), respectively.

selected covariates, age and race. $\hat{\kappa}_{jp}$ for age is larger than 0 and $\hat{\kappa}_{jp}$ for race smaller than 0 for many OTUs, although they are not statistically significant. Tabs A.7 and A.8 provide taxonomic information of the OTUs whose abundance and absence/presence are statistically significantly associated with the covariates, respectively.

Fig A.40 provides point estimates for β_{jp} for some selected covariates under the comparators. In panel (d), age under metagenomeSeq has statistically significant positive effects for most OTUs. EdgeR does not provide interval estimates, and Fig A.40 (g)-(i) illustrate point estimates only. Fig A.41 compares posterior predictive median estimates under ZI-MLN without Λ and mean abundance estimates under metagenome-

Table A.5: [Human Gut Microbiome Data] Covairates names with their support

Covariate Name	Support
Age	6 to 17
Gender	Male or Female
Race	White or non-white
Site Name	Cincinnati Children's Hospital or Massachusetts General Hospital (MGH) Pediatrics
Biopsy location	Ileum or Rectum
Disease phenotype	UC, CD or non-IBD

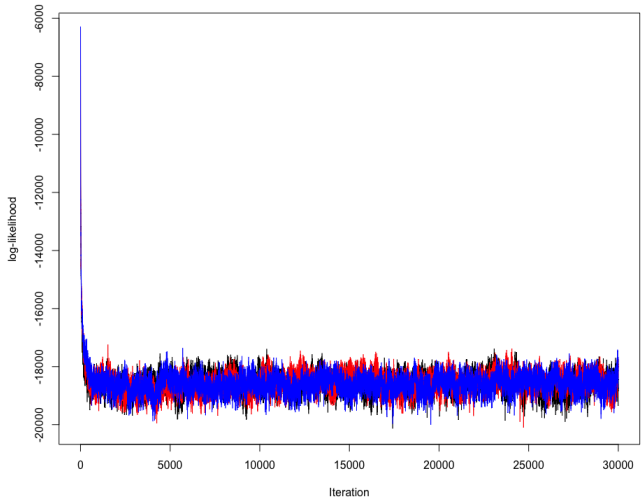


Figure A.37: [Human Gut Microbiome Data] Traceplots of log likelihood under three different initializations are presented in black, red and blue, respectively.

Seq and edgeR to the observed counts.

Table A.6: [Human Gut Microbiome Data] Taxonomic information for the OTUs illustrated in Fig 12(b) of the main text.

OTU	Taxonomic information (Kingdom/ Phylum/ Class/ Order/ Family / Genus)
30	Bacteria - Firmicutes - Clostridia - Clostridiales - Ruminococcaceae - Faecalibacterium
31	Bacteria - Firmicutes - Erysipelotrichia - Erysipelotrichales - Erysipelotrichaceae - Clostridium innocuum group
36	Bacteria - Fusobacteria - Fusobacteriia - Fusobacteriales - Fusobacteriaceae - Fusobacterium
37	Bacteria - Proteobacteria - Betaproteobacteria - Neisseriales - Neisseriaceae - Eikenella
39	Bacteria - Firmicutes - Erysipelotrichia - Erysipelotrichales - Erysipelotrichaceae - Erysipelatoclostridium
56	Bacteria - Firmicutes - Clostridia - Clostridiales - FamilyXI - Anaerococcus
59	Bacteria - Bacteroidetes - Bacteroidia - Bacteroidales - Bacteroidaceae - Bacteroides
62	Bacteria - Proteobacteria - Gammaproteobacteria - Xanthomonadales - Xanthomonadaceae - Stenotrophomonas - LachnospiraceaeUCG010
85	Bacteria - Firmicutes - Clostridia - Clostridiales - Ruminococcaceae - RuminococcaceaeNK4A214group
93	Bacteria - Proteobacteria - Gammaproteobacteria - Enterobacteriales - Enterobacteriaceae - Escherichia Shigella
96	Bacteria - Firmicutes - Clostridia - Clostridiales - Clostridiaceae1 - Clostridium Sensu Stricto 1

Table A.7: [Human Gut Microbiome Data] Taxonomic information for the OTUs, for which a 95% posterior credible interval estimate of κ_{jp} does not contain zero for covariates.

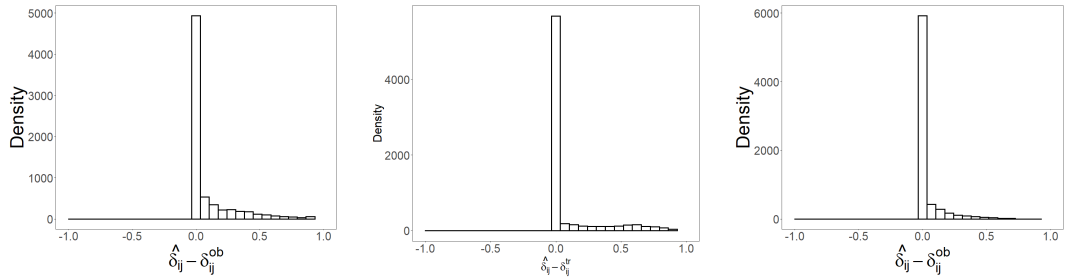
OTU	Covariate	Pos mean	95% credible interval	Taxonomic information (Kingdom/ Phylum/ Class/ Order / Family / Genus)
10	white−non-white	-2.72	[-5.44, -0.22]	Bacteria - Firmicutes - Bacilli - Lactobacillales - Streptococcaceae - Streptococcus
23	MGH −Cincinnati	-2.19	[-4.68, -0.05]	Bacteria - Proteobacteria - Alphaproteobacteria - Sphingomonadales
30	age	-2.83	[-4.91, -0.96]	Bacteria - Firmicutes - Clostridia - Clostridiales - Ruminococcaceae - Faecalibacterium
36	age	-0.89	[-1.85, -0.02]	Bacteria - Fusobacteria - Fusobacteriia - Fusobacteriales - Fusobacteriaceae - Fusobacterium
	white−non-white	-2.03	[-4.01, -0.24]	
49	Rectum−ileum	-2.32	[-4.25, -0.74]	Bacteria - Bacteroidetes - Flavobacteriia - Flavobacteriales - Flavobacteriaceae - Cloacibacterium
55	age	2.26	[0.66, 4.29]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - uncultured
68	MGH −Cincinnati	-2.32	[-4.79, -0.26]	Bacteria - Firmicutes - Erysipelotrichia - Erysipelotrichales - Erysipelotrichaceae - Erysipelatoclostridium
84	age	1.93	[0.38, 3.63]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - Lachnospiraceae UCG010
	male−female	2.21	[0.52, 3.98]	
102	male−female	-4.62	[-2.19, -0.10]	Bacteria - Firmicutes - Clostridia - Clostridiales - Ruminococcaceae - Faecalibacterium
104	male−female	-4.82	[-4.91, -0.53]	Bacteria - Firmicutes - Erysipelotrichia - Erysipelotrichales - Erysipelotrichaceae - Erysipelotrichaceae UCG003

Table A.8: [Human Gut Microbiome Data] Taxonomic information for the OTUs, for which a 95% posterior credible interval estimate of β_{jp} does not contain zero for covariates

OTU	Covariate	Pos mean	95% credible interval	Taxonomic information (Kingdom/ Phylum/ Class/ Order / Family / Genus)
9	white−non-white	3.44	[0.22 , 6.48]	Bacteria - Bacteroidetes - Bacteroidia - Bacteroidales - Bacteroidaceae - Bacteroides
10	Rectum-Ileum	-0.89	[-1.72, -0.03]	Bacteria - Firmicutes - Clostridia - Clostridiales - Family XI -Helcococcus
11	CD−non-IBD	-3.65	[-5.71, -1.40]	Bacteria - Firmicutes - Clostridia - Clostridiales - Peptostreptococcaceae - Intestinibacter
12	CD−non-IBD	-3.06	[-5.12, -0.58]	Bacteria - Bacteroidetes - Bacteroidia - Bacteroidales - Bacteroidaceae - Bacteroides
23	age	2.63	[0.23, 5.57]	Bacteria - Actinobacteria - Actinobacteria - Corynebacteriales - Corynebacteriaceae- Corynebacterium
25	CD−non-IBD	-2.53	[-4.85, -0.21]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - Coprococcus1
26	male−female	-2.49	[-4.71, -0.07]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - uncultured
32	MGH −Cincinnati	-3.11	[-5.57, -0.64]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - Tyzzerella
34	age	-2.12	[-3.91, -0.15]	Bacteria - Bacteroidetes - Bacteroidia - Bacteroidales - Prevotellaceae - Prevotella9
	UC−non-IBD	5.49	[0.68, 10.64]	
43	white−non-white	-3.41	[-6.65 , -0.11]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - NK4A136group
48	MGH −Cincinnati	3.19	[1.10 , 5.31]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - Lachnoclostridium
52	CD−non-IBD	-2.83	[-5.18, -0.51]	Bacteria - Firmicutes - Clostridia - Clostridiales - Ruminococcaceae - uncultured
55	male−female	-3.95	[-6.78, -1.51]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - uncultured
59	male−female	3.48	[0.11, 7.01]	Bacteria - Bacteroidetes - Bacteroidia - Bacteroidales - Bacteroidaceae - Bacteroides
	MGH −Cincinnati	-4.33	[-10.50 , -0.65]	
61	UC−non-IBD	3.09	[0.40, 5.86]	Bacteria - Proteobacteria - Gammaproteobacteria - Enterobacteriales - Enterobacteriaceae - Citrobacter
63	MGH −Cincinnati	-2.81	[-5.11, -0.43]	Bacteria - Bacteroidetes - Bacteroidia - Bacteroidales - Porphyromonadaceae - Parabacteroides
64	CD−non-IBD	-3.07	[-5.64, -0.68]	Bacteria - Firmicutes - Clostridia - Clostridiales - Ruminococcaceae - Ruminococcaceae UCG 013
68	age	2.33	[0.57, 4.17]	Bacteria - Firmicutes - Erysipelotrichia - Erysipelotrichales - Erysipelotrichaceae - Erysipelatoclostridium
69	Rectum−ileum	-1.50	[-2.46, -0.53]	Bacteria - Firmicutes - Bacilli - Lactobacillales - Carnobacteriaceae - Granulicatella

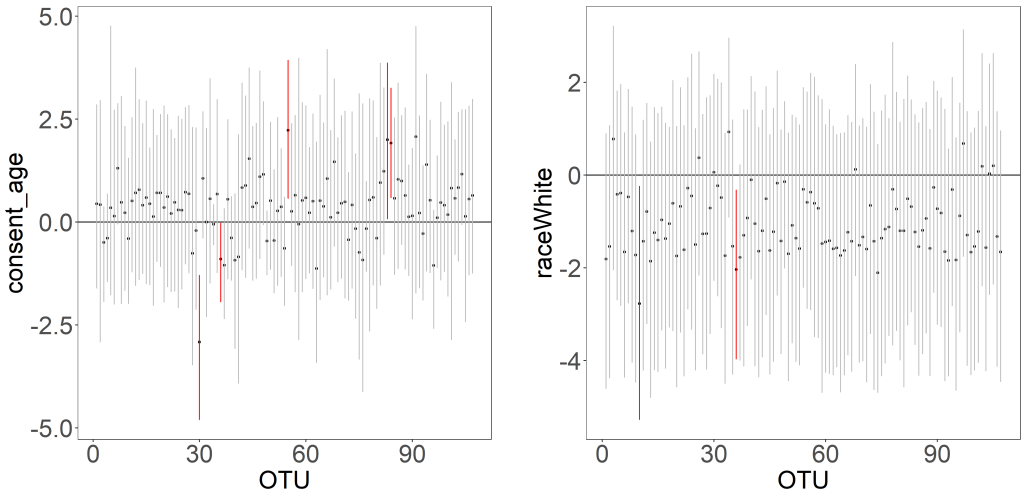
Table A.9: Tab A.8 continued [Human Gut Microbiome Data] Taxonomic information for the OTUs, for which a 95% posterior credible interval estimate of β_{jp} does not contain zero for covariates

OTU	Covariate	Pos mean	95% credible interval	Taxonomic information (Kingdom/ Phylum/ Class/ Order / Family / Genus)
79	age	1.65	[0.20, 2.94]	Bacteria - Firmicutes - Clostridia - Clostridiales
	male−female	-1.87	[-3.68, -0.14]	- Ruminococcaceae - Subdoligranulum
84	age	2.66	[0.53, 5.62]	Bacteria - Firmicutes - Clostridia - Clostridiales
	CD−non-IBD	3.77	[0.07, 7.11]	- Lachnospiraceae - Lachnospiraceae UCG 010
85	CD−non-IBD	-3.80	[-6.89,-0.52]	Bacteria - Firmicutes - Clostridia - Clostridiales
				- Ruminococcaceae - NK4A214group
86	white−non-white	-3.66	[-7.63 , -0.33]	Bacteria - Firmicutes - Clostridia - Clostridiales
				- Ruminococcaceae - Ruminiclostridium5
87	CD−non-IBD	-2.12	[-4.27, -0.01]	Bacteria - Firmicutes - Clostridia - Clostridiales
				- Ruminococcaceae - Subdoligranulum
89	CD−non-IBD	-3.16	[-5.47, -0.78]	Bacteria - Firmicutes - Clostridia - Clostridiales
				- Lachnospiraceae - Eubacterium Ventriosum Group
90	CD−non-IBD	-3.58	[-5.96, -0.91]	Bacteria - Firmicutes - Clostridia - Clostridiales
				- Lachnospiraceae - Lachnospira
91	white−non-white	3.99	[0.07, 8.68]	Bacteria - Firmicutes - Clostridia - Clostridiales
	CD−non-IBD	-3.44	[-6.23, -0.56]	- Clostridiaceae1 - Clostridium Sensu Stricto 1
92	CD−non-IBD	-3.88	[-7.20, -1.05]	Bacteria - Firmicutes - Clostridia - Clostridiales
				- Ruminococcaceae - Ruminiclostridium 6
93	MGH −Cincinnati	-2.81	[-4.96, -0.34]	Bacteria - Proteobacteria - Gammaproteobacteria
				- Enterobacteriales - Enterobacteriaceae
				- Escherichia Shigella
94	age	1.68	[0.17, 3.44]	Bacteria - Firmicutes - Clostridia - Clostridiales
	white−non-white	-5.52	[-9.84, -1.20]	- Ruminococcaceae - Ruminococcus 1
96	MGH −Cincinnati	-5.51	[-9.88, -0.35]	Bacteria - Firmicutes - Clostridia - Clostridiales
				- Clostridiaceae1 - Clostridium Sensu Stricto 1
103	CD−non-IBD	-4.22	[-7.10, -1.08]	Bacteria - Firmicutes - Clostridia - Clostridiales
				- Lachnospiraceae - Eubacterium eligens group
104	male−female	-2.21	[-4.43 , -0.13]	Bacteria - Firmicutes - Erysipelotrichia - Erysipelotrichales
	MGH −Cincinnati	1.92	[0.07, 3.99]	- Erysipelotrichaceae - Erysipelotrichaceae UCG 003

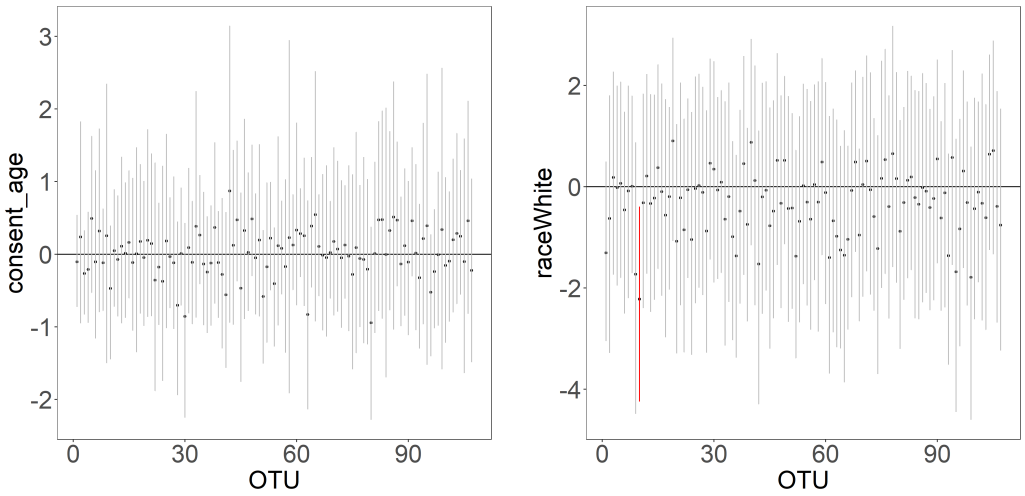


(a) ZI-MLN (b) ZI-MLN without Λ (c) metagenomeSeq

Figure A.38: [Human Gut Microbiome Data] Histograms of the differences between $\hat{\delta}_{ij}$ and the observed zero indicators $1(Y_{ij} = 0)$ under ZI-MLN, ZI-MLN without Λ and metagenomeSeq, respectively. Posterior mean estimates are used for ZI-MLN and ZI-MLN without Λ.

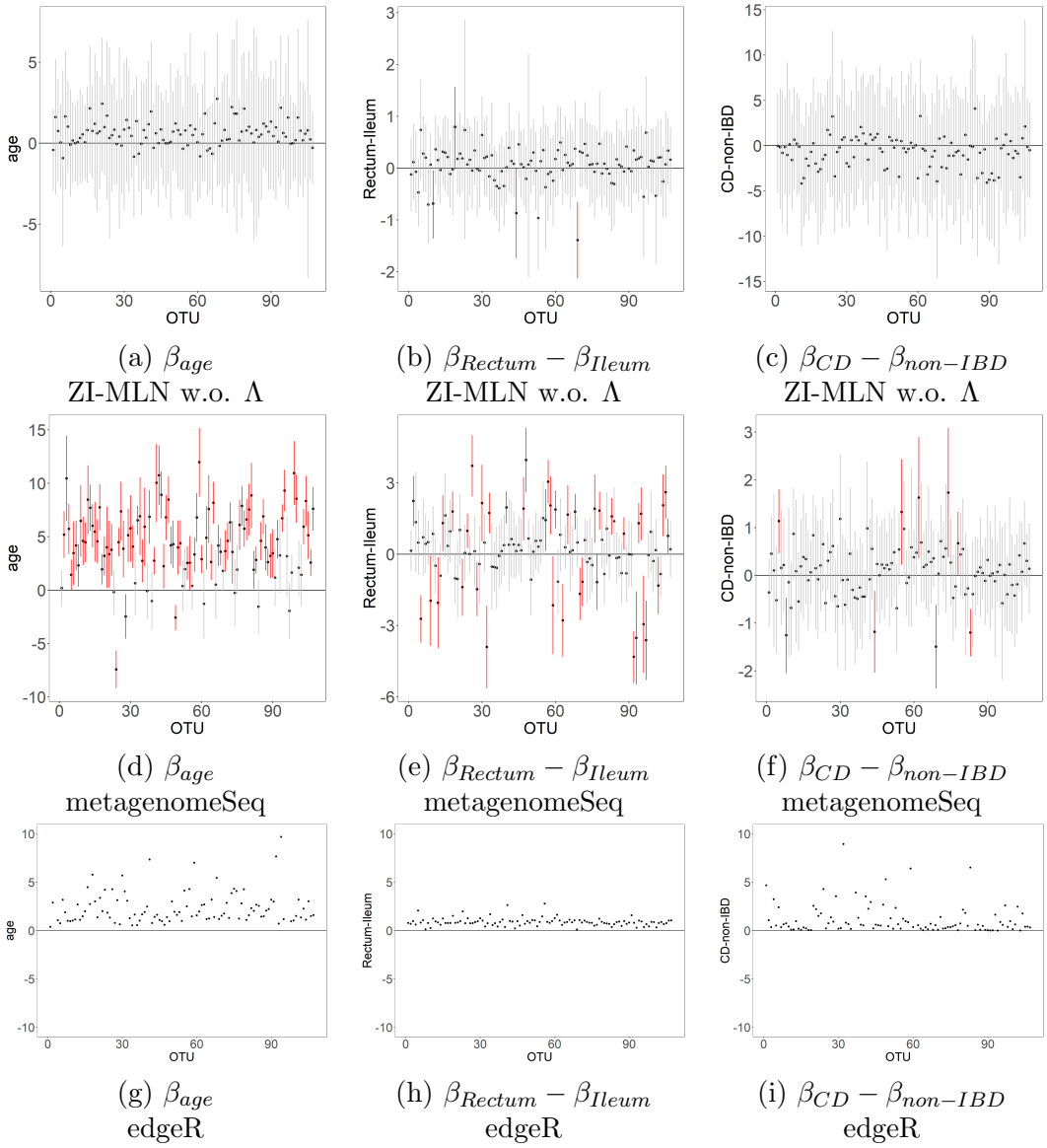


(a) κ_{age}, ZI-MLN (b) $\kappa_{\text{race=white}}$, ZI-MLN



(c) κ_{age}, ZI-MLN without Λ (d) $\kappa_{\text{race=white}}$, ZI-MLN without Λ

Figure A.39: [Human Gut Microbiome Data] Posterior estimates of regression coefficients κ_{age} and $\kappa_{\text{race=white}}$ under ZI-MLN and ZI-MLN without Λ for two selected covariates, where black dots are posterior mean estimates with vertical lines for 95% credible intervals. The intervals that do not contain zero are marked in red.

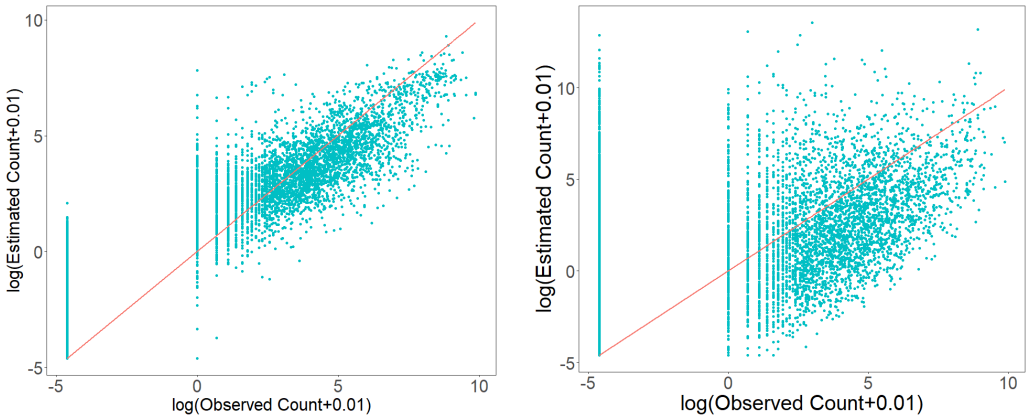


(a) β_{age} ZI-MLN w.o. Λ (b) $\beta_{Rectum} - \beta_{Ileum}$ ZI-MLN w.o. Λ (c) $\beta_{CD} - \beta_{non-IBD}$ ZI-MLN w.o. Λ

(d) β_{age} metagenomeSeq (e) $\beta_{Rectum} - \beta_{Ileum}$ metagenomeSeq (f) $\beta_{CD} - \beta_{non-IBD}$ metagenomeSeq

(g) β_{age} edgeR (h) $\beta_{Rectum} - \beta_{Ileum}$ edgeR (i) $\beta_{CD} - \beta_{non-IBD}$ edgeR

Figure A.40: [Human Gut Microbiome Data: Comparison] Posterior mean estimates of β_{jp} under the comparators for some selected covariates. Rows 1-3 are for ZI-MLN without Λ, metagenomeSeq and edgeR, respectively. Black dots and vertical lines represent point estimates and 95% confidence intervals. The intervals that do not contain zero are marked in red.



(a) MetagenomeSeq (b) EdgeR

Figure A.41: [Human Gut Microbiome Data: Comparison] Panels (a) and (b) present scatter plots of observed $\log(y_{ij} + 0.01)$ versus mean estimated $\log(\hat{y}_{ij} + 0.01)$ by metagenomeSeq and edgeR, respectively.

Appendix B

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

B.1 Properties of the Dirichlet-Horseshoe Distribution

We assume a Dirichlet-Horseshoe (Dir-HS) distribution for $\boldsymbol{\lambda}$ and examine the marginal distribution of λ_j. For a simple illustration, we consider a bivariate case with

$J = 2$. The Dir-HS distribution of $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ can be expressed as follows; first assume

$$\begin{aligned} \phi_1 &\sim \text{Be}(a_\phi, a_\phi), \text{ and let } \phi_2 = 1 - \phi_1, \\ \zeta_j &\overset{iid}{\sim} \text{C}^+(0,1), \quad j = 1, 2, \\ \lambda_j \mid \tau, \phi_j, \zeta_j &\overset{indep}{\sim} \text{N}(0, \zeta_j^2 \phi_j \tau), \quad j = 1, 2. \end{aligned} \tag{B.1}$$

The Dir-HS distribution of $\boldsymbol{\lambda}$ can be obtained by integrating out ζ_j and ϕ_1. Note that a gamma prior is placed for τ in (4) of the main text, while τ is assumed to be fixed in (B.1). Theorem 2.1 of the main text provides the bounds of the marginal density of λ_1 under a Dir-HS distribution in (B.1), and a proof is given below.

Proof. From the construction, we have

$$\Pi_{\text{Dir-HS}}(\lambda_1) = \int_0^1 \Pi(\lambda_1 \mid \phi_1) p(\phi_1) d\phi_1, \quad \text{where} \quad \Pi(\lambda_1 \mid \phi_1) = \int_0^\infty \Pi(\lambda_1 \mid \zeta_1, \phi_1) p(\zeta_1) d\zeta_1.$$

We recognize that $\Pi(\lambda_1 \mid \phi_1)$ is the HS distribution given ϕ_1, and we find the bounds of $\Pi(\lambda_1 \mid \phi_1)$ using Theorem 1 in Carvalho et al. (2010);

$$2^{-\frac{3}{2}} \pi^{-\frac{3}{2}} \phi_1^{-\frac{1}{2}} \log\left(1 + \frac{4\phi_1}{\lambda_1^2}\right) \;<\; \Pi(\lambda_1 \mid \phi_1) \;<\; 2^{-\frac{1}{2}} \pi^{-\frac{3}{2}} \phi_1^{-\frac{1}{2}} \log\left(1 + \frac{2\phi_1}{\lambda_1^2}\right).$$

Under the beta prior $\text{Be}(a_\phi, a_\phi)$ for ϕ_1, the bounds for $\Pi(\lambda_1)$ are

$$2^{-\frac{3}{2}} \pi^{-\frac{3}{2}} \frac{\Gamma(2a_\phi)}{\Gamma(a_\phi)\Gamma(a_\phi)} \int_0^1 \phi_1^{a_\phi - \frac{3}{2}} (1 - \phi_1)^{a_\phi - 1} \log\left(1 + \frac{4\phi_1}{\lambda_1^2}\right) d\phi_1, \tag{B.2}$$

and

$$2^{-\frac{1}{2}}\pi^{-\frac{3}{2}}\frac{\Gamma(2a_\phi)}{\Gamma(a_\phi)\Gamma(a_\phi)}\int_0^1 \phi_1^{a_\phi-\frac{3}{2}}(1-\phi_1)^{a_\phi-1}\log\left(1+\frac{2\phi_1}{\lambda_1^2}\right)d\phi_1. \tag{B.3}$$

We use the Taylor expansion of $\log\left(1+\frac{4\phi_1}{\lambda_1^2}\right)$, $\log\left(1+\frac{4\phi_1}{\lambda_1^2}\right)=\sum_{k=1}^{\infty}\frac{(-1)^{k+1}(4\phi_1/\lambda_1^2)^k}{k}$ and complete the integrals. Using the results in Gradshteyn and Ryzhik (2014), we then obtain the lower bound,

$$\begin{aligned}
&2^{-\frac{3}{2}}\pi^{-\frac{3}{2}}\frac{\Gamma(2a_\phi)}{\Gamma(a_\phi)\Gamma(a_\phi)}\int_0^1 \phi_1^{a_\phi-\frac{3}{2}}(1-\phi_1)^{a_\phi-1}\log\left(1+\frac{4\phi_1}{\lambda_1^2}\right)d\phi_1\\
&\quad=2^{-\frac{3}{2}}\pi^{-\frac{3}{2}}\frac{\Gamma(2a_\phi)}{\Gamma(a_\phi)\Gamma(a_\phi)}\sum_{k=1}^{\infty}\frac{\Gamma(a_\phi)\Gamma(a_\phi+k-1/2)}{\Gamma(2a_\phi+k-1/2)}\frac{(-1)^{k+1}(4/\lambda_1^2)^k}{k}\\
&\quad=2^{2a_\phi-\frac{5}{2}}\pi^{-2}\frac{4}{\lambda_1^2}\sum_{k=0}^{\infty}\frac{\Gamma(a_\phi+1/2)\Gamma(a_\phi+k+1/2)}{\Gamma(2a_\phi+k+1/2)}\frac{(-4/\lambda_1^2)^k}{k+1}\\
&\quad=2^{2a_\phi-\frac{5}{2}}\pi^{-2}\frac{\Gamma^2(a_\phi+1/2)}{\Gamma(2a_\phi+1/2)}\frac{4}{\lambda_1^2}\sum_{k=0}^{\infty}\frac{\frac{\Gamma(k+1)}{\Gamma(1)}\frac{\Gamma(k+1)}{\Gamma(1)}\frac{\Gamma(a_\phi+k+1/2)}{\Gamma(a_\phi+1/2)}}{\frac{\Gamma(k+2)}{\Gamma(2)}\frac{\Gamma(2a_\phi+k+1/2)}{\Gamma(2a_\phi+1/2)}}\frac{(-4/\lambda_1^2)^k}{k!}\\
&\quad=2^{2a_\phi-\frac{5}{2}}\pi^{-2}\frac{\Gamma^2(a_\phi+1/2)}{\Gamma(2a_\phi+1/2)}\frac{4}{\lambda_1^2}\,{}_3F_2\left(1,1,a_\phi+1/2;2,2a_\phi+1/2;-\frac{4}{\lambda_1^2}\right),
\end{aligned} \tag{B.4}$$

where generalized hypergeometric series ${}_pF_q(\alpha_1,\ldots,\alpha_p;\beta_1,\ldots,\beta_q;x)=\sum_{t=0}^{\infty}\frac{(\alpha_1)_t\ldots(\alpha_p)_t}{(\beta_1)_t\ldots(\beta_q)_t}\frac{x^t}{t!}$. We obtain the upper bound in a similar fashion,

$$\begin{aligned}
&2^{-\frac{1}{2}}\pi^{-\frac{3}{2}}\frac{\Gamma(2a_\phi)}{\Gamma(a_\phi)\Gamma(a_\phi)}\int_0^1 \phi_1^{a_\phi-\frac{3}{2}}(1-\phi_1)^{a_\phi-1}\log\left(1+\frac{2\phi_1}{\lambda_1^2}\right)d\phi_1\\
&\quad=2^{2a_\phi-\frac{3}{2}}\pi^{-2}\frac{\Gamma^2(a_\phi+1/2)}{\Gamma(2a_\phi+1/2)}\frac{2}{\lambda_1^2}\,{}_3F_2\left(1,1,a_\phi+1/2;2,2a_\phi+1/2;-\frac{2}{\lambda_1^2}\right).
\end{aligned} \tag{B.5}$$

When $a_\phi = 1/2$, the integrals in (B.2) and (B.3) are in a simpler form;

$$\left(\frac{1}{\sqrt{2\pi^5}} \log^2 \left(\frac{2}{|\lambda_1|} + \sqrt{\frac{4}{\lambda_1^2} + 1} \right), \sqrt{\frac{2}{\pi^5}} \log^2 \left(\frac{\sqrt{2}}{|\lambda_1|} + \sqrt{\frac{2}{\lambda_1^2} + 1} \right) \right).$$

□

We next compare the marginal density of a Dir-HS to that of a Dir-Laplace. Recall that we set $J = 2$. The Dir-Laplace is defined as follows; given τ,

$$\begin{aligned} \phi_1 &\sim \text{Be}(a_\phi, a_\phi), \text{ and let } \phi_2 = 1 - \phi_1, \\ \lambda_j \mid \phi_j &\overset{indep}{\sim} \text{DE}(\tau\phi_j), \quad j = 1, 2, \end{aligned} \tag{B.6}$$

where $\text{DE}(b)$ is the Laplace distribution with mean 0 and variance $2b^2$. The model in Bhattacharya et al. (2015) places a gamma prior on τ.

Proposition B.1.1. *Let $\Pi_{Dir\text{-}HS}(\lambda_1)$ denote the marginal distribution of λ_1 obtained from the Dir-HS distribution in* (B.1) *for $\boldsymbol{\lambda} = (\lambda_1, \lambda_2) \in \text{R}^2$ with $\tau \in \text{R}^+$. Similarly, Let $\Pi_{Dir\text{-}Laplace}(\lambda_1)$ denote the marginal distribution of ϕ_1 obtained from the Dir-Laplace distribution in* (B.6) *for $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ with fixed τ. The limits of the ratio of $\Pi_{Dir\text{-}Laplace}(\lambda_1)$ to $\Pi_{Dir\text{-}HS}(\lambda_1)$ are*

$$\lim_{\lambda_1 \to \pm\infty} \frac{\Pi_{Dir\text{-}Laplace}(\lambda_1)}{\Pi_{Dir\text{-}HS}(\lambda_1)} = 0. \tag{B.7}$$

Proof. Without loss of generality, we fix $\tau = 1$. From the construction of Dir-Laplace

distributions, we have the marginal distribution

$$\begin{aligned}\Pi_{\text{Dir-Laplace}}(\lambda_1) &= \int_0^1 \frac{1}{2\phi_1} e^{-\frac{|\lambda_1|}{\phi_1}} \frac{\Gamma(2a_\phi)}{\Gamma(a_\phi)\Gamma(a_\phi)} \phi_1^{a_\phi-1}(1-\phi_1)^{a_\phi-1} d\phi_1 \\ &= \frac{\Gamma(2a_\phi)}{2\Gamma(a_\phi)\Gamma(a_\phi)} \int_0^1 \phi_1^{a_\phi-2}(1-\phi_1)^{a_\phi-1} e^{-\frac{|\lambda_1|}{\phi_1}} d\phi_1.\end{aligned}$$

From (B.4), we have

$$\frac{\Pi_{\text{Dir-Laplace}}(\lambda_1)}{\Pi_{\text{Dir-HS}}(\lambda_1)} \leq \frac{\frac{\Gamma(2a_\phi)}{2\Gamma(a_\phi)\Gamma(a_\phi)} \int_0^1 \phi_1^{a_\phi-2}(1-\phi_1)^{a_\phi-1} e^{-\frac{|\lambda_1|}{\phi_1}} d\phi_1}{2^{2a_\phi-\frac{5}{2}}\pi^{-2}\Gamma(a_\phi+1/2)\frac{4}{\lambda_1^2}\sum_{k=0}^{\infty}\frac{\Gamma(a_\phi+k+1/2)}{\Gamma(2a_\phi+k+1/2)}\frac{(-4/\lambda_1^2)^k}{k+1}}. \tag{B.8}$$

We first observe $e^{-\frac{|\lambda_1|}{\phi_1}}/\phi_1 \leq e^{-|\lambda_1|}$ for any $0 < \phi_1 < 1$ if $|\lambda_1| > 1$. Given $|\lambda_1| > 1$, we have

$$\frac{\Gamma(2a_\phi)}{2\Gamma(a_\phi)\Gamma(a_\phi)} \int_0^1 \phi_1^{a_\phi-2}(1-\phi_1)^{a_\phi-1} e^{-\frac{|\lambda_1|}{\phi_1}} d\phi_1 \leq \frac{e^{-|\lambda_1|}}{2}.$$

Then from (B.8) we have

$$\frac{\frac{\Gamma(2a_\phi)}{2\Gamma(a_\phi)\Gamma(a_\phi)} \int_0^1 \phi_1^{a_\phi-2}(1-\phi_1)^{a_\phi-1} e^{-\frac{|\lambda_1|}{\phi_1}} d\phi_1}{2^{2a_\phi-\frac{3}{2}}\pi^{-2}\Gamma(a_\phi+1/2)\frac{2}{\lambda_1^2}\sum_{k=0}^{\infty}\frac{\Gamma(a_\phi+k+1/2)}{\Gamma(2a_\phi+k+1/2)}\frac{(-2/\lambda_1^2)^k}{k+1}} \leq \frac{\frac{e^{-|\lambda_1|}}{2}}{2^{2a_\phi-\frac{3}{2}}\pi^{-2}\Gamma(a_\phi+1/2)\frac{2}{\lambda_1^2}\sum_{k=0}^{\infty}\frac{\Gamma(a_\phi+k+1/2)}{\Gamma(2a_\phi+k+1/2)}\frac{(-2/\lambda_1^2)^k}{k+1}}$$

and observe

$$\lim_{\lambda_1 \to \pm\infty} \frac{\frac{e^{-|\lambda_1|}}{2}}{2^{2a_\phi-\frac{3}{2}}\pi^{-2}\Gamma(a_\phi+1/2)\frac{2}{\lambda_1^2}\sum_{k=0}^{\infty}\frac{\Gamma(a_\phi+k+1/2)}{\Gamma(2a_\phi+k+1/2)}\frac{(-2/\lambda_1^2)^k}{k+1}} = 0.$$

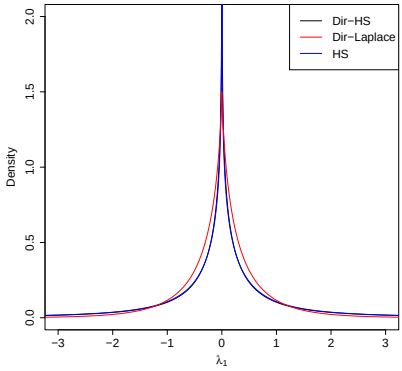
Therefore, we obtain (B.7). □

Proposition B.1.1 compares tails of the Dir-HS and Dir-Laplace distributions and states a Dir-HS distribution has heavier tails than a Dir-Laplace distribution.

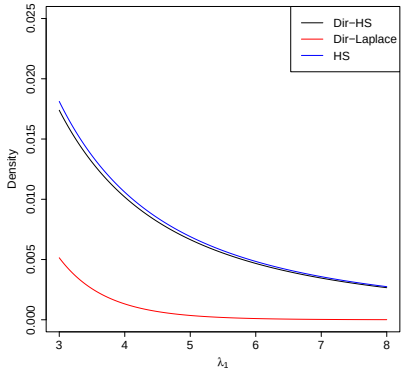
We next use numerical simulations and examine joint distributions of $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ assuming Dir-HS, Dir-Laplace and independent HS distributions. We fix $\tau = 1$ for the Dir-HS and Dir-Laplace distributions. For the independent HS distributions, we generate $\lambda_j \mid \zeta_j \overset{indep}{\sim} \mathrm{N}(0, \zeta_j^2/2)$ and $\zeta_j \overset{iid}{\sim} \mathrm{C}^+(0,1)$, $j = 1, 2$ to match the scale parameter with that under the Dir-HS. We vary the value of a_ϕ to examine how it affects the joint distributions. Figs B.1 and B.2 illustrate the joint densities with $a_\phi = 2, 1/2$ and $1/20$. As explained in the main text with $a_\phi = 1/20$, Dir-HS distributions have higher densities along the axes than the independent HS distributions. It illustrates joint sparsity under the Dir-HS by shrinking one component toward zero more than the other component. Compared to the Dir-Laplace, the Dir-HS has thicker tails. The Dir-HS has unbounded density around zero for any value of a_ϕ, but the Dir-Laplace has bounded density around zero if $a_\phi > 1$.

B.2 Exploration of the Distributions of OTU Counts Under Sp-BGFM

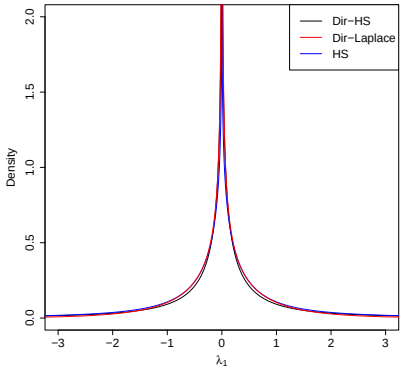
In this section, we explore the marginal distribution of an OTU's count and the joint distribution of the counts of a pair of OTUs under Sp-BGFM, using examples. Specifically, we illustrate how Sp-BGFM addresses statistical challenges outlined in



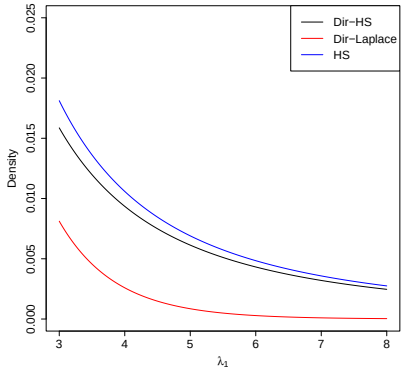
(a) Central density: $a_\phi = 2$



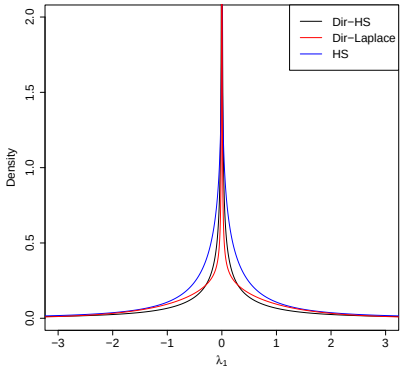
(b) Tail density: $a_\phi = 2$



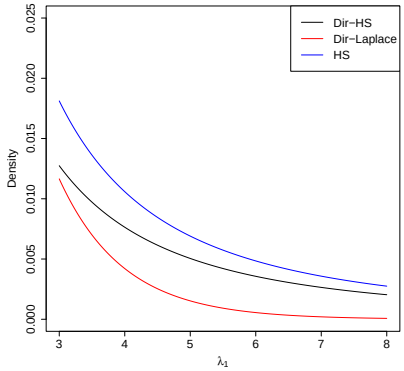
(c) Central density: $a_\phi = 1/2$



(d) Tail density: $a_\phi = 1/2$

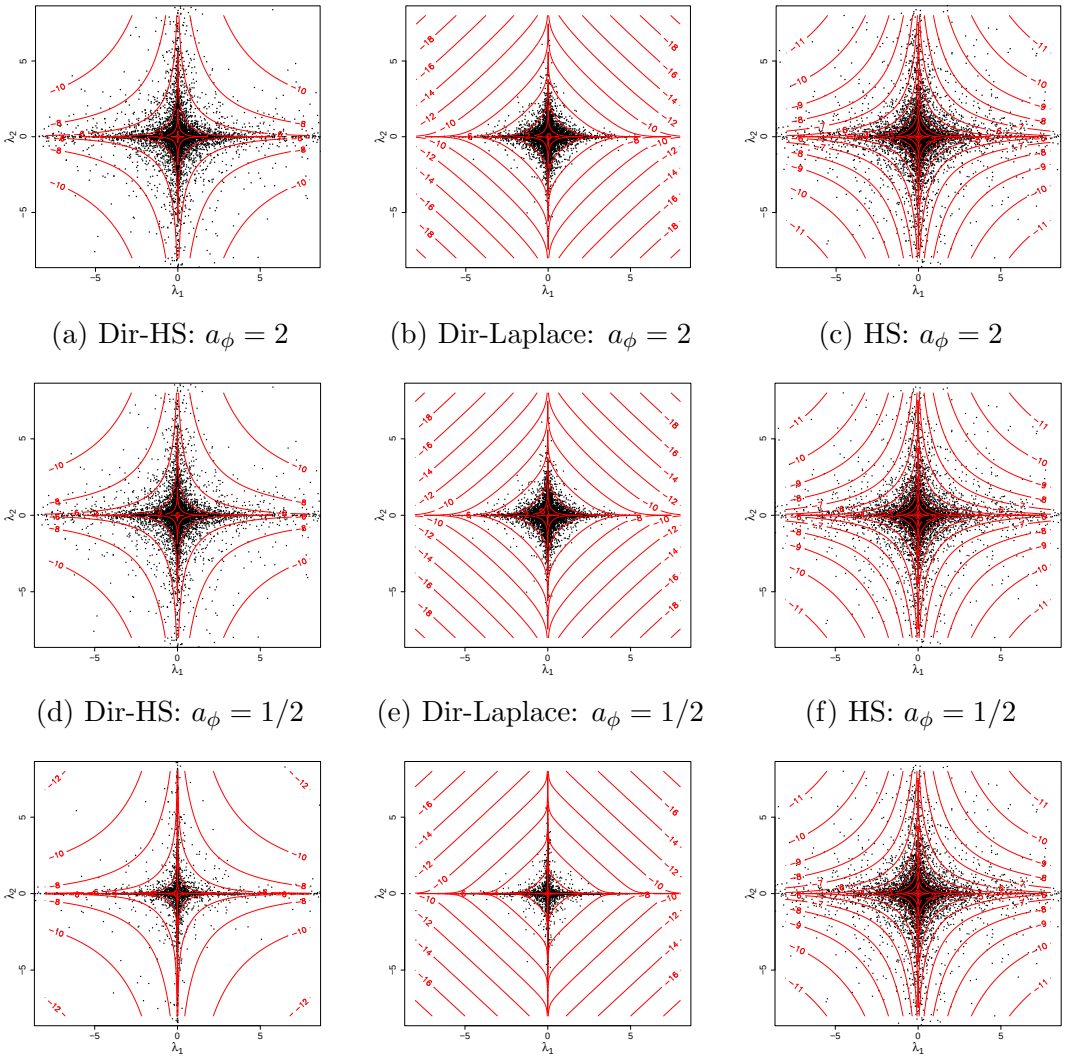


(e) Central density: $a_\phi = 1/20$



(f) Tail density: $a_\phi = 1/20$

Figure B.1: Marginal densities of λ_1 are numerically evaluated at the central and tail areas for the Dir-HS prior, Dir-Laplace, and HS with different values of a_ϕ, $a_\phi = 2, 1/2, 1/20$. The Dir-HS, Dir-Laplace and independent HS distributions are in black, red and blue, respectively.



(a) Dir-HS: $a_\phi = 2$ (b) Dir-Laplace: $a_\phi = 2$ (c) HS: $a_\phi = 2$

(d) Dir-HS: $a_\phi = 1/2$ (e) Dir-Laplace: $a_\phi = 1/2$ (f) HS: $a_\phi = 1/2$

(g) Dir-HS: $a_\phi = 1/20$ (h) Dir-Laplace: $a_\phi = 1/20$ (i) HS: $a_\phi = 1/20$

Figure B.2: Scatter plots of $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ are shown. $\boldsymbol{\lambda}$ are generated from three different prior distributions: Dir-HS in the leftmost column, Dir-Laplace in the middle column, and independent HS priors in the rightmost column. The values of a_ϕ used for the plots are 2, 1/2, and 1/20 for the top, middle, and bottom plots, respectively. The contour plots of the empirical joint densities are shown in red on a logarithmic scale.

§3.1.2 of the main text, including sparsity, between-sample variability, and dependence between OTUs. For simplicity, we consider a model without regression, i.e., $\boldsymbol{\mu}_i =$

$\mathbf{r}_i + \boldsymbol{\alpha}_{s_i}$.

From (1)-(3) of the main text, the marginal distribution of an OTU's count is

$$\mathrm{P}(y_{imj} = y \mid \boldsymbol{\mu}_i, \Sigma) \quad = \quad \int_{y}^{y+1} f_{y^\star}(y^\star_{imj} \mid \mu_{imj}, \Sigma^{mm}_{jj})\, dy^\star_{imj}, \quad y = 0, 1, 2, \ldots. \quad \text{(B.9)}$$

where $\mu_{imj} = r_{im} + \alpha_{s_i mj}$. Here, $f_{y^\star}$ is the density of a univariate log-normal distribution with parameters μ_{imj} and Σ^{mm}_{jj}. After integrating $\alpha_{s_i mj}$ out with respect to G_{mj} in (7) of the main text, we obtain

$$\begin{aligned}\mathrm{P}(y_{imj} = y \mid r_{im}, G_{mj}, \Sigma) = \int_{y}^{y+1} \sum_{l=1}^{\infty} \psi^\alpha_{ml} \Big\{ & \omega^\alpha_{ml} f_{y^\star}(y^\star_{imj} \mid r_{im} + \xi^\alpha_{mjl}, \Sigma^{mm}_{jj}) \\ & + (1 - \omega^\alpha_{ml}) f_{y^\star}\left(y^\star_{imj} \mid r_{im} + \frac{\nu^\alpha_{mj} - \omega^\alpha_{mj}\xi^\alpha_{mjl}}{1 - \omega^\alpha_{ml}}, \Sigma^{mm}_{jj}\right) \Big\}\, dy^\star_{imj},\end{aligned} \quad \text{(B.10)}$$

for $y = 0, 1, 2, \ldots$. Fig B.3 illustrates the marginal distribution of an OTU's count using the rounded kernel method with a log-normal distribution. A single log-normal distribution, log-N$(\alpha_{s_i mj} + r_{im}, \Sigma^{mm}_{jj})$ is used to generate the distribution of y_{imj} for panels (a)-(c). We varied $\alpha_{s_i mj}$ and Σ^{mm}_{jj}, while $r_{im} = 0$ is fixed. In panels (d)-(f), we used a mixture of two log-normals with a constraint ν^α_{mj};

$$\omega^\alpha_m \text{log-N}(r_{im} + \xi^\alpha_m, \Sigma^{mm}_{jj}) + (1 - \omega^\alpha_m)\text{log-N}\left(r_{im} + \frac{\nu^\alpha_{mj} - \omega^\alpha_{mj}\xi^\alpha_{mj}}{1 - \omega^\alpha_m}, \Sigma^{mm}_{jj}\right). \quad \text{(B.11)}$$

We varied the mixture weights ω^α as well as the location ξ^α and variance Σ^{mm}_{jj}, while fixing ν^α_{mj} at 3. In particular, we generate $\mathrm{P}(y_{imj} = y)$ with the following specifications;

- Mixture Case I in panel (d): $\omega^\alpha = 0.5$ and $\xi^\alpha = -1$

- Mixture Case II in panel (e): $\omega^{\alpha} = 0.25$ and $\xi^{\alpha} = 0$
- Mixture Case III in panel (f): $\omega^{\alpha} = 0.1$ and $\xi^{\alpha} = 1$

where Σ_{jj}^{mm} are also varied. The figure shows that (B.11) accommodates excess zeros, multimodality and variability in counts. The model in (B.10) has an infinite mixture for $\alpha_{s_i mj}$ and allows greater flexibility to accommodate various patterns in the distribution of an OTU count.

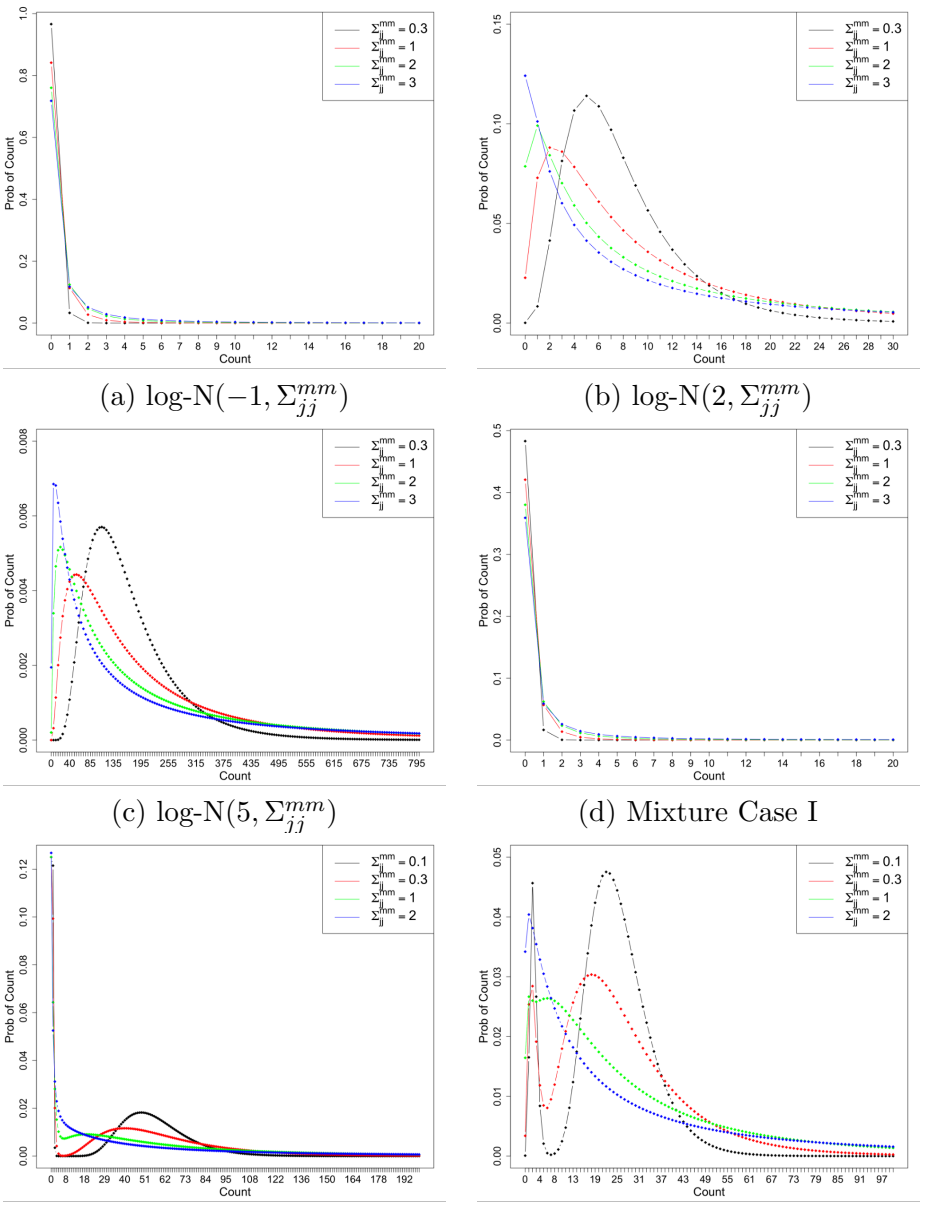
We next explore the joint distribution of counts of a pair of OTUs, y_{imj} and $y_{im'j'}$. Similar to (B.9), we have

$$\mathrm{P}(y_{imj} = y, y_{im'j'} = y' \mid \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}) \quad = \quad \int_{y}^{y+1} \int_{y'}^{y'+1} f_{\mathbf{y}^\star}(y^\star_{imj}, y^\star_{im'j'} \mid \tilde{\boldsymbol{\mu}}, \tilde{\Sigma})\, dy^\star_{im'j'} dy^\star_{imj}, \tag{B.12}$$

where $y, y' = 0, 1, 2, \ldots$. In (B.12), $f_{\mathbf{y}^\star}$ is the density of the bivariate log-normal distribution with parameters

$$\tilde{\boldsymbol{\mu}} = \begin{bmatrix} \mu_{imj} \\ \\ \mu_{im'j'} \end{bmatrix} = \begin{bmatrix} r_{im} + \alpha_{s_i mj} \\ \\ r_{im'} + \alpha_{s_i m'j'} \end{bmatrix} \quad \text{and} \quad \tilde{\Sigma} = \begin{bmatrix} \Sigma_{jj}^{mm} & \Sigma_{jj'}^{mm'} \\ \\ \Sigma_{jj'}^{mm'} & \Sigma_{j'j'}^{m'm'} \end{bmatrix}.$$

From (2) and (7) of the main text, we have $\alpha_{s_i mj} \sim G_{mj}$ and $\alpha_{s_i m'j'} \sim G_{m'j'}$, where G_{mj} and $G_{m'j'}$ are an infinite mixture of point masses with a mean constraint. Fig B.4 illustrates how the joint distribution of y_{imj} and $y_{im'j'}$ varies with $\tilde{\boldsymbol{\mu}}$ and $\tilde{\Sigma}$ using a single bivariate log-normal distribution. From the figures, the dependence between y_{imj} and $y_{im'j'}$ varies with changes in $\tilde{\Sigma}$. For Fig B.5, a mixture of two point masses with a



(a) log-N$(-1, \Sigma_{jj}^{mm})$ (b) log-N$(2, \Sigma_{jj}^{mm})$

(c) log-N$(5, \Sigma_{jj}^{mm})$ (d) Mixture Case I

(e) Mixture Case II (f) Mixture Case III

Figure B.3: [Distribution of an OTU's Count] The probability distribution of an OTU's count is computed from a rounded kernel method with log-normal distributions. For panels (a)-(c), a single log-normal distribution is used, and for panels (d)-(f), a mixture of two log-normals with a constraint in (B.11) is used. The detailed specifications are in § B.2.

mean constraint is used for each of G_{mj} and $G_{m'j'}$;

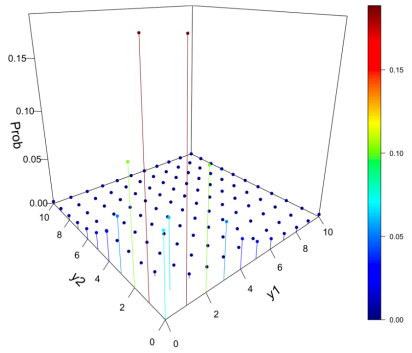
$$G_{\tilde{m}\tilde{j}} = \omega^{\alpha}_{\tilde{m}}\delta_{\xi^{\alpha}_{\tilde{m}\tilde{j}}} + (1-\omega^{\alpha}_{\tilde{m}})\delta_{(\nu^{\alpha}_{\tilde{m}\tilde{j}}-\omega^{\alpha}_{\tilde{m}\tilde{j}}\xi^{\alpha}_{\tilde{m}\tilde{j}})/(1-\omega^{\alpha}_{\tilde{m}})}, \quad (\tilde{m},\tilde{j}) \in \{(m,j),(m',j')\}, \quad \text{(B.13)}$$

where δ_{ξ} is a point mass at ξ. In particular, ν^{α} is fixed at 1.5 and 0.5 for (m,j) and (m',j'), respectively, for both panels, and r_{im} and $r_{im'}$ are fixed at 0. We then generate $\mathrm{P}(y_{imj} = y, y_{im'j'} = y')$ with the following specifications;

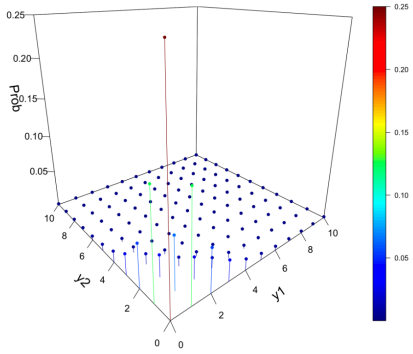
- Case I in panel (a): $\omega^{\alpha}_{m} = 0.5, \omega^{\alpha}_{m'} = 0.4, \xi^{\alpha}_{mj} = 1, \xi^{\alpha}_{m'j'} = 0, \Sigma^{mm}_{jj} = \Sigma^{m'm'}_{j'j'} = 0.5^2, \Sigma^{mm'}_{jj'} = 0.5^2 \times 0.9$
- Case II in panel (b): $\omega^{\alpha}_{m} = 0.1, \omega^{\alpha}_{m'} = 0.6, \xi^{\alpha}_{mj} = 1, \xi^{\alpha}_{m'j'} = 2, \Sigma^{mm}_{jj} = \Sigma^{m'm'}_{j'j'} = 0.5^2, \Sigma^{mm'}_{jj'} = 0.5^2 \times -0.9$

Fig B.5 demonstrates the flexibility of the model, even with fixed ν^{α}. The infinite mixture for $G(\boldsymbol{\alpha})$ in Eq. (7) of the main text can provide more flexibility to accommodate the potential complexity of real data.

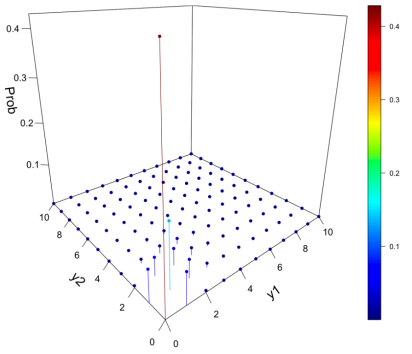
We also calculate the expectation of count variables and their correlation. The moment of order k can be computed through $\mathrm{E}(Y^{k}_{ij} \mid \mu_{ij}, \Sigma_{jj}) = \sum_{b=0}^{\infty} b^{k}\mathrm{P}(Y_{ij} = b \mid \mu_{ij}, \Sigma_{jj})$ with $\mathrm{P}(Y_{ij} = b \mid \mu_{ij}, \Sigma_{jj}) = \Phi_1(\log(b+1) \mid \mu_{ij}, \Sigma_{jj}) - \Phi_1(\log(b) \mid \mu_{ij}, \Sigma_{jj})$, where $\Phi_d(\cdot \mid a, \boldsymbol{B})$ is the cdf of the $d-$ variate normal distribution with mean a and (co)variance $\boldsymbol{B}$. The covariance and correlation of any two count random variables are calculated by $\mathrm{Cov}(Y_{ij}, Y_{ij'} \mid \boldsymbol{\mu}_i, \Sigma) = \sum_{b=0}^{\infty}\sum_{b'=0}^{\infty} bb'\mathrm{P}(Y_{ij} = b, Y_{ij'} = b' \mid \boldsymbol{\mu}_i, \Sigma) - \mathrm{E}(Y_{ij} \mid \mu_{ij}, \Sigma_{jj})\mathrm{E}(Y_{ij'} \mid \mu_{ij'}, \Sigma_{j'j'})$ and $\mathrm{Cor}(Y_{ij}, Y_{ij'} \mid \boldsymbol{\mu}_i, \Sigma) = \mathrm{Cov}(Y_{ij}, Y_{ij'})/\sqrt{\mathrm{Var}(Y_{ij})\mathrm{Var}(Y_{ij'})}$. $\mathrm{P}(Y_{ij} = b, Y_{ij'} = b' \mid \boldsymbol{\mu}_i, \Sigma)$ can be computed with a bivariate normal distribution in



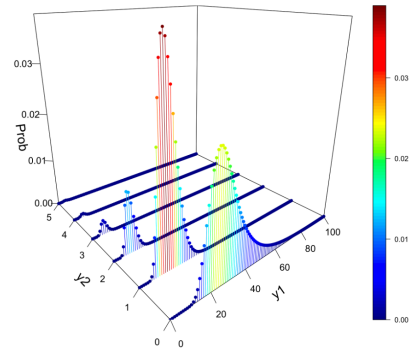
(a) $\tilde{\boldsymbol{\mu}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\tilde{\Sigma} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$



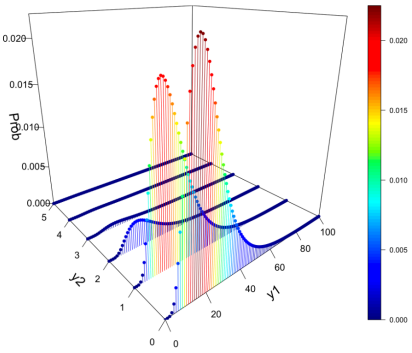
(b) $\tilde{\boldsymbol{\mu}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\tilde{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$



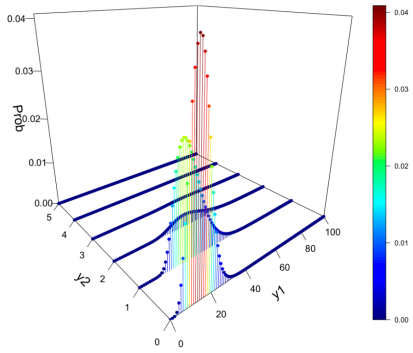
(c) $\tilde{\boldsymbol{\mu}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\tilde{\Sigma} = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$



(d) $\tilde{\boldsymbol{\mu}} = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$, $\tilde{\Sigma} = \begin{bmatrix} 0.5^2 & -0.9 \times 0.5^2 \\ -0.9 \times 0.5^2 & 0.5^2 \end{bmatrix}$

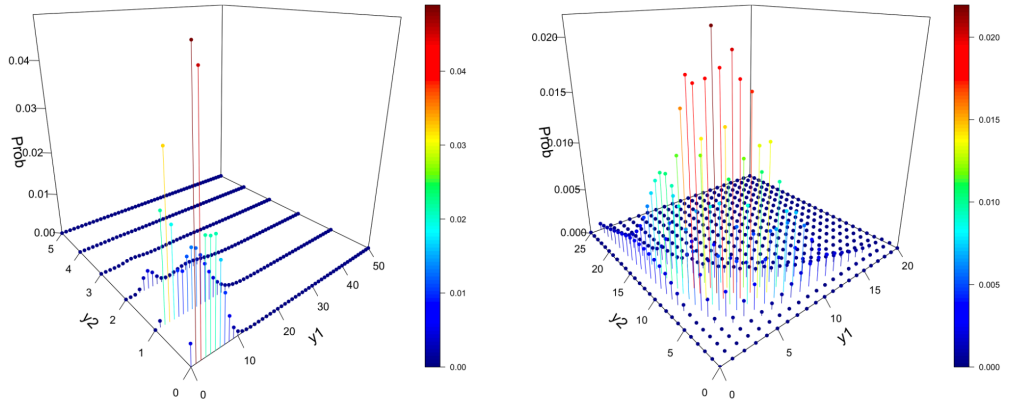


(e) $\tilde{\boldsymbol{\mu}} = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$, $\tilde{\Sigma} = \begin{bmatrix} 0.5^2 & 0 \\ 0 & 0.5^2 \end{bmatrix}$



(f) $\tilde{\boldsymbol{\mu}} = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$, $\tilde{\Sigma} = \begin{bmatrix} 0.5^2 & 0.9 \times 0.5^2 \\ 0.9 \times 0.5^2 & 0.5^2 \end{bmatrix}$

Figure B.4: [Distribution of Counts of a Pair of OTUs I] The joint distribution of counts of a pair of OTUs is computed for a rounded kernel method with bivariate log-normals, log-$\mathrm{N}_2(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma})$. Different combinations of $\tilde{\boldsymbol{\mu}}$ and $\tilde{\Sigma}$ are used.



(a) Case I (b) Case II

Figure B.5: [Distribution of Counts of a Pair of OTUs II] The joint distribution of counts of a pair of OTUs is computed for a rounded kernel method with a mixture of bivariate log-normals in (B.13). ν^{α} is fixed at 1.5 and 0.5 for two OTUs, while the mixture weights and locations vary. The detailed specifications are in § B.2.

a way similar to $\mathrm{P}(Y_{ij} = b \mid \mu_{ij}, \Sigma_{jj})$. Tab B.1 presents the moments of the count variables illustrated in Fig B.4 and Fig B.5. The cases of $\rho = 0$ represent independence between count random variables. The moments of the count distribution vary with μ and Σ, the parameters of the distribution of their latent continuous variables. Thus, posterior inferences on μ and Σ provide inference on the distribution of count vectors.

B.3 Details of Posterior Computation

We use Markov chain Monte Carlo (MCMC) techniques to obtain samples of the random parameters $\boldsymbol{\theta}$ from their posterior distributions, where $\boldsymbol{\theta} = \{\lambda_{mjk}, \phi_{mjk}, \tau_k, \zeta_{mkj}, v_m^2, \alpha_{s_i mj}, \omega_{ml}^{\alpha}, V_{ml}^{\alpha}, \xi_{mjl}^{\alpha}, r_{im}, \omega_{ml}^{r}, V_{ml}^{r}, \xi_{ml}^{r}, \beta_{mjp}\}$. Recall that $Y_{imj} \in \mathbb{N}^0$, $i = 1, \ldots, N$, $m = 1, \ldots, M$ and $j = 1, \ldots, J_m$ denotes the count of OTU j of group m in

Table B.1: [Moments of bivariate count vectors] Moments of bivariate count vectors in Fig B.4 and Fig B.5 are presented. Moments are referred to the marginal expectation, variance and correlation of bivariate count vectors.

Figure	$\mathrm{E}(Y_1)$	$\mathrm{E}(Y_2)$	$\mathrm{Cov}(Y_1)$	$\mathrm{Cov}(Y_2)$	$\mathrm{Cor}(Y_1, Y_2)$
Fig B.4(a)	1.170	1.170	4.636	4.636	-0.277
Fig B.4(b)	1.170	1.170	4.636	4.636	0
Fig B.4(c)	1.170	1.170	4.636	4.636	0.835
Fig B.4(d)	22.188	0.600	141.608	0.486	-0.643
Fig B.4(e)	22.188	0.600	141.608	0.486	0
Fig B.4(f)	22.188	0.600	141.608	0.486	0.801
Fig B.5(a)	5.225	231.841	18.308	60728.64	0.312
Fig B.5(b)	4.639	146.836	8.188	43223.17	-0.256

sample i, r_{im} the sample size factor of group m of sample i, and $\alpha_{s_i mj}$ the normalized baseline abundance level of OTU j of group m in sample i obtained from subject $s_i = 1, \ldots, S$. We also have covariate $\mathbf{X}$, a $N \times P$ covariate matrix whose rows have a $P-$dim covariate vector $\mathbf{x}_i$.

To facilitate the posterior simulation, we introduce the latent continuous variable $y^\star_{imj} \in \mathbb{R}^+$ and have $y_{imj} = \lfloor y^\star_{imj} \rfloor$. We then impute $y^\star_{imj} = \exp(\tilde{y}^\star_{imj})$ from a truncated log-normal distribution

$$\tilde{y}^\star_{imj} \mid \boldsymbol{\theta}, \boldsymbol{\eta}_i, y_{imj} \sim \mathrm{N}(r_{im} + \alpha_{s_i mj} + \boldsymbol{\lambda}'_{mj}\boldsymbol{\eta}_i + \mathbf{x}'_i\boldsymbol{\beta}_{mj}, v_m^2) 1(\log(y_{imj}) \leq \tilde{y}^\star_{imj} < \log(y_{imj}+1)).$$

Given $\tilde{y}^\star_{imj}$, parameters $\boldsymbol{\beta}_{mj}$, $\boldsymbol{\eta}_i$, and v_m^2 can be conveniently updated through normal/inv-gamma Gibbs steps. For ζ_{mjk}, we utilize the following to achieve conjugacy (Makalic

and Schmidt, 2015);

$$\zeta_{mjk} \overset{iid}{\sim} \mathrm{C}^{+}(0,1) \;\Leftrightarrow\; \zeta^2_{mjk} \mid Z_{mjk} \overset{iid}{\sim} \text{inv-Ga}\left(\frac{1}{2},\ \frac{1}{Z_{mjk}}\right),\ Z_{mjk} \overset{iid}{\sim} \text{inv-Ga}\left(\frac{1}{2}, 1\right).$$

ζ_{mjk} can be easily updated via Gibbs steps. Also, recall that parameters $\boldsymbol{r}_i$ and $\boldsymbol{\alpha}_{s_i}$ are from infinite mixtures of mixtures. For computational convenience, when fitting the model, we approximate the infinite mixtures in (5) and (6) of the main text by truncating the number of mixture components to L^α and L^r. The final weights $\psi^\alpha_{mL^\alpha} = 1 - \sum_{l=1}^{L^\alpha - 1} \psi^\alpha_{ml}$ and $\psi^r_{mL^r} = 1 - \sum_{l=1}^{L^r-1} \psi^r_{ml}$ is set to ensure the distributions are proper. With sufficiently large L^α and L^r, the truncated process produces inference almost identical to that with the infinite process (Ishwaran and James, 2001). We further introduce a pair of membership variables (I^r_{im1}, I^r_{im2}) with $I^r_{im1} \in \{1, \ldots, L^r\}$ and $I^r_{im2} \in \{0, 1\}$ for each r_{im} and $(I^\alpha_{s_imj1}, I^\alpha_{s_imj2})$ with $I^\alpha_{s_imj1} \in \{1, \ldots, L^\alpha\}$ and $I^\alpha_{s_imj2} \in \{0, 1\}$ for each α_{s_imj}. We then assume $\mathrm{P}(I^r_{im1} = l) = \psi^r_{ml}$ and $\mathrm{P}(I^r_{im2} = 0 \mid I^r_{im1} = l) = \omega^r_{ml}$, and similarly, assume $\mathrm{P}(I^\alpha_{s_imj1} = l) = \psi^\alpha_{ml}$ and $\mathrm{P}(I^\alpha_{s_imj2} = 0 \mid I^\alpha_{s_imj1} = l) = \omega^\alpha_{ml}$. Given the membership indicator vectors, the conditional distributions of r_{im} and α_{s_imj} are

$$r_{im} \mid \boldsymbol{\psi}^r, \boldsymbol{\omega}^r, \boldsymbol{\xi}^r, I^r_{im1} = l, I^r_{im2} \sim \begin{cases} \mathrm{N}(\xi^r_{ml}, u^2_r) & \text{if } I^r_{im2} = 1, \\ \mathrm{N}(\frac{v^r_m - \omega^r_{ml}\xi^r_{ml}}{1-\omega^r_{ml}}, u^2_r) & \text{if } I^r_{im2} = 0, \end{cases}$$

$$\alpha_{s_imj} \mid \boldsymbol{\psi}^\alpha, \boldsymbol{\omega}^\alpha, \boldsymbol{\xi}^\alpha, I^\alpha_{S_imj1} = l, I^\alpha_{S_imj2} = \begin{cases} \xi^\alpha_{mjl} & \text{if } I^\alpha_{s_imj2} = 1, \\ \frac{v^\alpha_{mj} - \omega^\alpha_l \xi^\alpha_{ml}}{1-\omega^\alpha_l} & \text{if } I^\alpha_{s_imj2} = 0. \end{cases}$$

Given the latent variables, all parameters except $\boldsymbol{\phi}_k$ are updated through Gibbs steps. We update $\boldsymbol{\phi}_k$ using a Metropolis-Hastings step. We let $\phi^\star_{mjk} \overset{iid}{\sim} \text{Ga}(a_\phi, 1)$ and have $\phi_{mjk} = \phi^\star_{mjk} / \sum_{m',\, j'} \phi^\star_{m'j'k}$. The full conditional of $\boldsymbol{\phi}_k$ is given by

$$p(\boldsymbol{\phi}_k \mid -) \propto p(\boldsymbol{\lambda}_k \mid \tau_k, \boldsymbol{\phi}_k, \boldsymbol{\zeta}_k) p(\boldsymbol{\phi}_k) \propto \prod_{m=1}^{M} \prod_{j=1}^{J_m} \text{N}(\lambda_{mjk} \mid 0, \zeta^2_{mjk} \phi_{mjk} \tau_k) \prod_{m=1}^{M} \prod_{j=1}^{J_m} \text{Ga}(\phi^\star_{mjk} \mid a_\phi, 1).$$

To efficiently update $\boldsymbol{\phi}_k$, the adaptive MH algorithm (Haario et al., 2001) is applied to adjust the MH step size according to the acceptance ratio, and the convergence rate is accelerated.

We sample sequentially by alternating conditional sampling. The full conditionals are given below;

- Update $\tilde{y}^\star_{imj}$ given $y_{imj}, r_{im}, \alpha_{S_i mj}, \boldsymbol{\lambda}_{mj}, \boldsymbol{\eta}_i, v^2_m, \boldsymbol{\beta}_{mj}, \mathbf{x}_i$

$$\tilde{y}^\star_{imj} \sim \text{N}(r_{im} + \alpha_{S_i mj} + \boldsymbol{\lambda}'_{mj} \boldsymbol{\eta}_i + \mathbf{x}'_i \boldsymbol{\beta}_{mj}, v^2_m) \mathbb{1}(\log(y_{imj}) \leq \tilde{y}^\star_{imj} < \log(y_{imj} + 1)).$$

- parameters related to r_{im}

  - Update $\boldsymbol{\psi}^r_m$ given I^r_{im1}

$$\psi^r_{m1} = V^r_{m1}, \psi^r_{ml} = V^r_{ml} \prod_{h=1}^{l-1} (1 - V^r_{mh}), \text{ for } l = 2, \ldots, L^r - 1, \ \psi^r_{mL^r} = 1 -$$

$\sum_{l=1}^{L^r-1} \psi^r_{ml}$,

$$V^r_{ml} \sim \text{Be}(1 + \sum_{i=1}^{N} \mathbb{1}(I^r_{im1} = l), \ldots, c^r + \sum_{i=1}^{N} \sum_{h>l} \mathbb{1}(I^r_{im1} = h)).$$

– Update ω^r_{ml} given I^r_{im1}, I^r_{im2}

$$p(\omega^r_{ml} \mid -) \propto {\omega^r_{ml}}^{a^r_\omega + \sum_{i=1}^N \mathbb{1}(I^r_{im1}=l, I^r_{im2}=1)} (1-\omega^r_{ml})^{b^r_\omega + \sum_{i=1}^N \mathbb{1}(I^r_{im1}=l, I^r_{im2}=0)} \prod_{i=1}^{N} \prod_{j=1}^{J} \text{N}(\tilde{y}^\star_{imj} \mid \mu_{imj}, \sigma^2).$$

We use logistic transformation and adaptive Metropolis-Hasting algorithm (Haario et al., 2001) to update ω^r_{ml}.

– Update (I^r_{im1}, I^r_{im2}) given $\psi^r_{ml}, \omega^r_{ml}, r_{im}, v^r_m, \xi^r_{ml}, u^2_r$

$$\begin{aligned} \Pr(I^r_{im1} = l, I^r_{im2} = 1) &\propto \psi^r_{ml}\omega^r_{ml}\text{N}(r_{im} \mid \xi^r_{ml}, u^2_r), \\ \Pr(I^r_{im1} = l, I^r_{im2} = 0) &\propto \psi^r_{ml}(1-\omega^r_{ml})\text{N}(r_{im} \mid \frac{\nu^r_m - \omega^r_{ml}\xi^r_{ml}}{1-\omega^r_{ml}}, u^2_r), \end{aligned}$$

– Update ξ^r_{ml} given $I^r_{im1}, I^r_{im2}, r_{im}, \omega^r_{ml}$

$$\xi^r_{ml} \sim \text{N}(\tilde{u}^2_{\xi^r}(\tfrac{\nu^r_m}{u^2_{\xi^r}} + \sum_{i:I^r_{im1}=l, I^r_{im2}=1} \tfrac{r_{im}}{u^2_r} - \sum_{i:I^r_{im1}=l, I^r_{im2}=0} \tfrac{\frac{\omega^r_{ml}}{1-\omega^r_{ml}} r_{im} - \frac{\omega^r_{ml}}{(1-\omega^r_{ml})^2}\nu^r_m}{u^2_r}), \tilde{u}^2_{\xi^r}),$$

where $\tilde{u}^2_{\xi^r} = (1/u^2_{\xi^r} + \sum_{i=1}^N \mathbb{1}(I^r_{im1} = l, I^r_{im2} = 1)/u^2_r + \omega^{r,2}_{ml} \sum_{i=1}^N \omega^r_{ml}\mathbb{1}(I^r_{im1} = l, I^r_{im2} = 0)/u^2_r(1-\omega^r_{ml})^2)^{-1}$.

– Update r_{im} given $\alpha_{S_imj}, \boldsymbol{\lambda}_{mj}, \boldsymbol{\eta}_i, \boldsymbol{\beta}_{mj}$

$$r_{im} \sim \mathrm{N}\left((\tfrac{c}{u_r^2} + \tfrac{\sum_{j=1}^{J_m}(\tilde{y}^{\star}_{imj} - \alpha_{S_i mj} - \boldsymbol{\lambda}'_{mj}\boldsymbol{\eta}_i - \mathbf{x}'_i\boldsymbol{\beta}_{mj})}{\sigma_m^2})(\tfrac{1}{u_r^2} + \tfrac{J_m}{\sigma_m^2})^{-1}, (\tfrac{1}{u_r^2} + \tfrac{J_m}{\sigma_m^2})^{-1} \right),$$

where prior mean

$$c = \textstyle\sum_{j=1}^{J_m}(\mathbb{1}(I^r_{im2} = 1)\xi^r_{m,I^r_{im1}} + \mathbb{1}(I^r_{im2} = 0)\frac{\nu^r_m - \omega^r_{m,I^r_{im1}}\xi^r_{m,I^r_{im1}}}{1-\omega^r_{m,I^r_{im1}}}).$$

- parameters related to $\alpha_{s_i mj}$

  - Update $\boldsymbol{\psi}^{\alpha}_m$ given $I^{\alpha}_{S_i mj1}$

    $\psi^{\alpha}_{m1} = V^{\alpha}_{m1}, \psi^{\alpha}_{ml} = V^{\alpha}_{ml}\prod_{h=1}^{l-1}(1 - V^{\alpha}_{mh})$, for $l = 2, \ldots, L^{\alpha} - 1$, $\psi^{\alpha}_{L^{\alpha}} = 1 - \sum_{l=1}^{mL^{\alpha}-1}\psi^{\alpha}_{ml}$

$$V^{\alpha}_{ml} \sim \mathrm{Be}(1 + \sum_{i=1}^{N}\sum_{j=1}^{J}\mathbb{1}(I^{\alpha}_{S_i mj1} = l), c^{\alpha} + \sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{h>l}\mathbb{1}(I^r_{S_i mj1} = h)).$$

  - Update ω^{α}_{ml} given $I^{\alpha}_{S_i mj1}, I^{\alpha}_{S_i mj2}$

$$p(\omega^{\alpha}_{ml} \mid -) \propto {\omega^{\alpha}_{ml}}^{a^{\alpha}_{\omega} + \sum_{i=1}^{N}\sum_{j=1}^{J}\mathbb{1}(I^{\alpha}_{S_i mj1} = l, I^{\alpha}_{S_i mj2} = 1)}$$
$$(1 - \omega^{\alpha}_{ml})^{b^{\alpha}_{\omega} + \sum_{i=1}^{N}\sum_{j=1}^{J}\mathbb{1}(I^{\alpha}_{S_i mj1} = l, I^{\alpha}_{S_i mj2} = 0)}\prod_{i=1}^{N}\prod_{j=1}^{J_m}\mathrm{N}(\tilde{y}^{\star}_{imj} \mid \mu_{imj}, \sigma_m^2).$$

    We use logistic transformation and adaptive Metropolis-Hasting algorithm (Haario et al., 2001) to update ω^{α}_{ml}.

– Update ω_{ml}^{α} given $I_{S_imj1}^{\alpha}, I_{S_imj2}^{\alpha}$

$$\omega_{ml}^{\alpha} \sim \text{Be}(a_{\omega}^{\alpha} + \sum_{i=1}^{N}\sum_{j=1}^{J_m} \mathbb{1}(I_{S_imj1}^{\alpha} = l, I_{S_imj2}^{\alpha} = 1),$$
$$b_{\omega}^{\alpha} + \sum_{i=1}^{N}\sum_{j=1}^{J_m} \mathbb{1}(I_{S_imj1}^{\alpha} = l, I_{S_imj2}^{\alpha} = 0)).$$

– Update $(I_{smj1}^{\alpha}, I_{smj2}^{\alpha})$ given $\psi_{ml}^{\alpha}, \omega_{ml}^{\alpha}$

$$\Pr(I_{smj1}^{\alpha} = l, I_{smj2}^{\alpha} = 1) \propto \psi_{ml}^{\alpha}\omega_{ml}^{\alpha}$$
$$\prod_{i:S_i=s} \text{N}(\tilde{y}_{imj}^{\star} \mid r_{im} + \xi_{mjl}^{\alpha} + \boldsymbol{\lambda}_{mj}'\boldsymbol{\eta}_i + \mathbf{x}_i'\boldsymbol{\beta}_{mj}, v_m^2),$$
$$\Pr(I_{smj1}^{\alpha} = l, I_{smj2}^{\alpha} = 0) \propto \psi_{ml}^{\alpha}(1 - \omega_{ml}^{\alpha})$$
$$\prod_{i:S_i=s} \text{N}(\tilde{y}_{imj}^{\star} \mid r_{im} + \frac{\nu_{mj}^{\alpha} - \omega_{ml}^{\alpha}\xi_{mjl}^{\alpha}}{1 - \omega_{ml}^{r}} + \boldsymbol{\lambda}_{mj}'\boldsymbol{\eta}_i + \mathbf{x}_i'\boldsymbol{\beta}_{mj}, v_m^2).$$

– Update ξ_{mjl}^{α} given $\tilde{y}_{imj}^{\star}, r_{im}, \boldsymbol{\lambda}_{mj}', \boldsymbol{\eta}_i, \mathbf{x}_i, \boldsymbol{\beta}_{mj}$

$$\xi_{mjl}^{\alpha} \sim \text{N}(\tilde{u}_{\alpha}^2(\nu_{mj}^{\alpha}/u_{\alpha}^2 + \sum_{i:I_{S_imj1}^{\alpha}=l, I_{S_imj2}^{\alpha}=1} (\tilde{y}_{imj}^{\star} - r_{im} - \boldsymbol{\lambda}_{mj}'\boldsymbol{\eta}_i -$$
$$\mathbf{x}_i'\boldsymbol{\beta}_{mj})/v_m^2 - \sum_{i:I_{S_imj1}^{\alpha}=l, I_{S_imj2}^{\alpha}=0} (\tfrac{\omega_{ml}^{\alpha}}{1-\omega_{ml}^{\alpha}}(\tilde{y}_{imj}^{\star} - r_{im} - \boldsymbol{\lambda}_{mj}'\boldsymbol{\eta}_i - \mathbf{x}_i'\boldsymbol{\beta}_{mj})$$
$$- \tfrac{\omega_{ml}^{\alpha}}{(1-\omega_{ml}^{\alpha})^2}\nu_{mj}^{\alpha})/v_m^2), \tilde{u}_{\alpha}^2),$$

where $\tilde{u}_{\alpha}^2 = (1/u_{\alpha}^2 + \sum_{i=1}^{N} \mathbb{1}(I_{S_imj1}^{\alpha} = l, I_{S_imj2}^{\alpha} = 1)/v_m^2 + \omega_{ml}^{\alpha,2}\sum_{i=1}^{N}\omega_{ml}^{\alpha}$
$\mathbb{1}(I_{S_imj1}^{\alpha} = l, I_{S_imj2}^{\alpha} = 0)/v_m^2(1 - \omega_{ml}^{\alpha})^2)^{-1}$.

- Update $\boldsymbol{\lambda}_{mj}$ given $\boldsymbol{Y}^{\star}_{mj}, \mathbf{r}_m, \boldsymbol{\alpha}_{mj}, \mathbf{X}, \boldsymbol{\beta}_{mj}, \boldsymbol{\eta}, \boldsymbol{\zeta}_{mj}, \boldsymbol{\phi}_{mj}, \boldsymbol{\tau}$

$$\boldsymbol{\lambda}_{mj} \sim \mathrm{N}((v_m^{-2}\boldsymbol{\eta}'\boldsymbol{\eta} + V_\lambda^{-2})^{-1} v_m^{-2}\boldsymbol{\eta}'(\tilde{\boldsymbol{Y}}^{\star}_{mj} - \mathbf{r}_m - \boldsymbol{\alpha}_{mj} - \mathbf{X}\boldsymbol{\beta}_{mj}), (v_m^{-2}\boldsymbol{\eta}'\boldsymbol{\eta} + V_\lambda^{-2})^{-1}),$$

where $V_\lambda = \mathrm{diag}(\zeta^2_{mj1}\phi_{mj1}\tau_1, \ldots, \zeta^2_{mjK}\phi_{mjK}\tau_K)$.

- Update $\boldsymbol{\phi}_k$ using adaptive M-H by proposing from a normalized $\mathrm{Ga}(a_\phi, 1)$.

We let $\phi^{\star}_{mjk} \overset{iid}{\sim} \mathrm{Ga}(a_\phi, 1)$ and have $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_J) \sim \mathrm{Dir}(a_\phi, \ldots, a_\phi)$ with $\phi_{mjk} = \phi^{\star}_{mjk} / \sum_{j'} \phi^{\star}_{mj'k}$. The full conditional of $\boldsymbol{\phi}_k$ is given by

$$p(\boldsymbol{\phi}_k \mid -) \propto \prod_{m=1}^{M}\prod_{j=1}^{J_m} \mathrm{N}(\lambda_{mjk} \mid 0, \zeta^2_{mjk}\phi_{mjk}\tau_k) \prod_{m=1}^{M}\prod_{j=1}^{J_m} \mathrm{Ga}(\phi^{\star}_{mjk} \mid a_\phi, 1).$$

We reject or accept the proposal by utilizing the adaptive MH algorithm (Haario et al., 2001).

- Update ζ^2_{mjk} given $Z_{mjk}, \lambda_{mjk}, \phi_{mjk}, \tau_k$

$$\zeta^2_{mjk} \sim \text{inv-Ga}(1, 1/Z_{mjk} + \lambda^2_{mjk}/(2\phi_{mjk}\tau_k), \sum_{m=1}^{M}\sum_{j=1}^{J_m}(\lambda^2_{mjk}/\phi_{mjk}\tau_k)).$$

- Update Z_{mjk} given $\zeta_{mjk} \overset{indep}{\sim} \text{inv-Ga}(1, 1 + 1/\zeta^2_{mjk})$.

- Update $\tau_k \mid \lambda_{mjk}, \zeta_{mjk}, \phi_{mjk}$

$$\tau_k \sim \text{Generalized inverse Gaussian } (a_\tau - J/2, 2b_\tau, \sum_{m=1}^{M}\sum_{j=1}^{J_m} \lambda^2_{mjk}/\zeta^2_{mjk}\phi_{mjk}).$$

- Update $\boldsymbol{\eta}_i$ given $\Lambda, V, \boldsymbol{Y}_i^\star, \mathbf{r}_i, \boldsymbol{\alpha}_{s_i}, \boldsymbol{\beta}, \mathbf{X}_i$

$$\boldsymbol{\eta}_i \sim \mathrm{N}((I_K + \Lambda' V^{-1}\Lambda)^{-1}\Lambda' V^{-1}(\boldsymbol{Y}_i^\star - \mathbf{r}_i - \boldsymbol{\alpha}_{s_i} - \boldsymbol{\beta}\mathbf{X}_i), (I_K + \Lambda' V^{-1}\Lambda)^{-1}).$$

- Update $\boldsymbol{\beta}_{mj}$ given $\tilde{\boldsymbol{Y}}_{mj}^\star, \mathbf{r}_m, \boldsymbol{\alpha}_{mj}, \boldsymbol{\lambda}_{mj}, \boldsymbol{\eta}, \mathbf{X}$

$$\boldsymbol{\beta}_{mj} \sim \mathrm{N}((\frac{\mathbf{X}'\mathbf{X}}{v_m^2} + v_\beta^{-2} I_p)^{-1} v_m^{-2}\mathbf{X}'(\tilde{\boldsymbol{Y}}_{mj}^\star - \mathbf{r}_m - \boldsymbol{\alpha}_{mj} - \boldsymbol{\eta}\boldsymbol{\lambda}_{mj}), (v_m^{-2}\mathbf{X}'\mathbf{X} + v_\beta^{-2} I_p)^{-1}).$$

- Update v_m^2 given $\tilde{y}_{imj}^\star, r_{im}, \alpha_{S_i,j}, \boldsymbol{\lambda}'_{mj}, \boldsymbol{\eta}_i, \mathbf{x}_i, \boldsymbol{\beta}_{mj}$

$$v_m^2 \sim \text{inv-Ga}(a_v + n \times J_m/2, b_v + \sum_{i=1}^{n}\sum_{j=1}^{J_m} \frac{(\tilde{y}_{imj}^\star - r_{im} - \alpha_{S_i,j} - \boldsymbol{\lambda}'_{mj}\boldsymbol{\eta}_i - \mathbf{x}_i'\boldsymbol{\beta}_{mj})^2}{2}).$$

B.4 Instruction of reproducing codes

SP-BGFM requires R 3.6 or greater to reproduce the tables and graphics in Chapter 3. Download and install R from https://www.r-project.org/. Once installed, open R from the terminal and run the following command to install packages especially Rcpp and RcppArmadillo for Rc++ functions:

```
install.packages(c("Rcpp", "RcppArmadillo", "statmod", "GIGrvg",
"extraDistr", "abind", "mvnfast", "mvnfast", "statmod", "extraDistr")).
```

For a comparison of the SPIEC-EASI method, we need to install `install.packages("SpiecEasi")`.

Make sure that the C++ compiler is correctly installed. Mac users need to install Xcode from command line tools. Execute the command 'xcode-select –install' on Terminal. One can also import the GitHub repository https://github.com/shuang-jie/ZI-MLN directly to load all functions. Under the simulation code folder, Sim 1-5.R produces the results displayed in Sim 1-5 in Chapter 3, and please save the result as RData with the respective sim ID. To reproduce the real data, Filtered7539OTUs.RData has the multi-domain skin microbiome data in Chapter 3.4. In the real data, it contains:

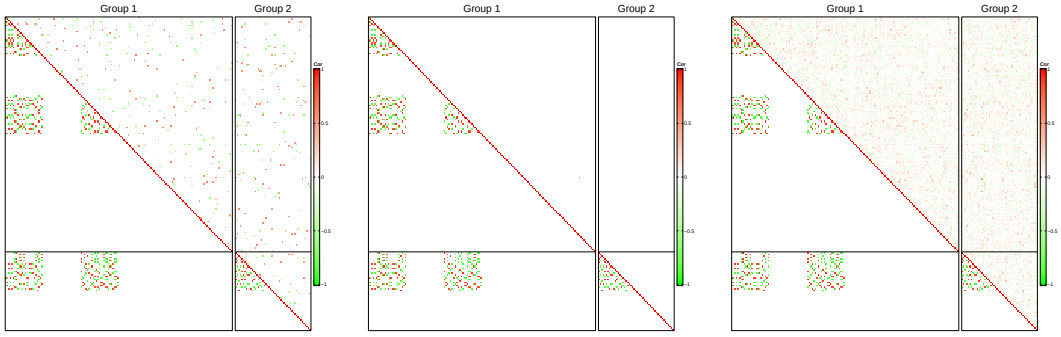
- Y1 : bacterial microbiome count table. 60 samples × 75 OTUs. Each row is a sample, and each column is a bacterial OTU.
- Y2 : viral microbiome count table. 60 samples × 39 OTUs. Each row is a sample, and each column is a viral OTU.
- Y : combined multi-domain skin microbiome data. 60 samples × 114(75+39) OTUs.
- X : a categorical covariate representing experimental conditions. (1,0,0) pre-treatment & (0,1,0) post-treatment & (0,0,1) healthy condition.
- J : number of OTUs in each domain. (75, 39)
- Jsum : number of total OTUs. 114
- n : number of samples. 60
- S : number of subjects. 20

Real Data.R reproduces the results illustrated in Chapter 3.4. Save the result as RData with Real Data.RData. One can further access the performance of SP-BGFM and reproduce Figures 3.1-3.10 using Folder figures-codes.

B.5 Additional Simulation Studies

B.5.1 Additional Results of Simulation 1

We present results from additional comparators REBACCA(Ban et al., 2015), COAT(Cao et al., 2019) and Zi-LN (Prost et al., 2021) for Simulation 1 presented in §3.3.1 of the main text. Those comparators are for a single group count table analysis. To apply those methods for count table data of two groups, we first combined $\mathbf{Y}_1$ of size $N \times J_1$ and $\mathbf{Y}_2$ of size $N \times J_2$ and had a single count matrix of $N \times J$ with $J = J_1 + J_2$. We then applied their normalization or transformation procedures for those methods. REBECCA uses sample proportions by normalizing the observed data by the total number of counts and estimates the covariance matrix of the log-transformed latent basis abundances with the ℓ_1 penalty. COAT further develops REBECCA using a procedure of thresholding the sample centered log-ratio covariance matrix. It can avoid optimization and is scalable for large covariance matrices. Zi-LN is a likelihood-based zero-inflation model on a single transformed count table. Zi-LN normalizes the observed counts using a modified centered log-ratio (clr) transformation prior to analysis to account for zeros, and uses graphical lasso for estimating the precision matrix. The three methods do not include covariates in their model.



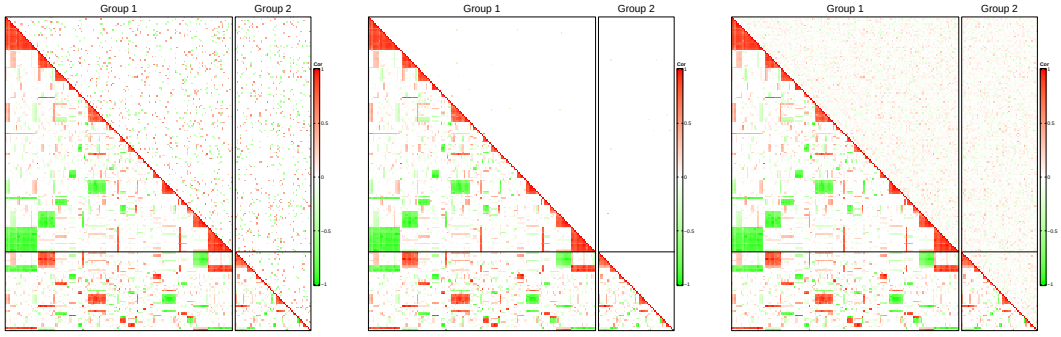
(a) REBACCA (b) COAT (c) Zi-LN

Figure B.6: [Simulation 1] 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 REBACCA, COAT and Zi-LN.

The upper triangles of the heatmaps in Fig B.6 illustrate the estimate of the correlations $\hat{\rho}_{jj'}$ obtained from the additional comparators for Simulation 1. The true values $\rho_{jj'}^{\text{tr}}$ of the correlations are shown in the lower triangles of the heatmaps. Compared to the estimate under Sp-BGFM in panel (a) of Fig 4 of the main text, the comparators perform poorly and do not capture the true interaction patterns among OTUs both within and across groups. This could be due to limitations such as a single-domain analysis and/or failure to account for inter-subject heterogeneity.

B.5.2 Additional Details of Simulation 2

For Simulation 2, we set $M = 2$, $J_1 = 150$, $J_2 = 50$, $S = 20$ the same as in Simulation 1. We incorporated a binary covariate to represent two experimental conditions. To denote 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,



(a) REBACCA (b) COAT (c) Zi-LN

Figure B.7: [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 REBACCA, COAT and Zi-LN.

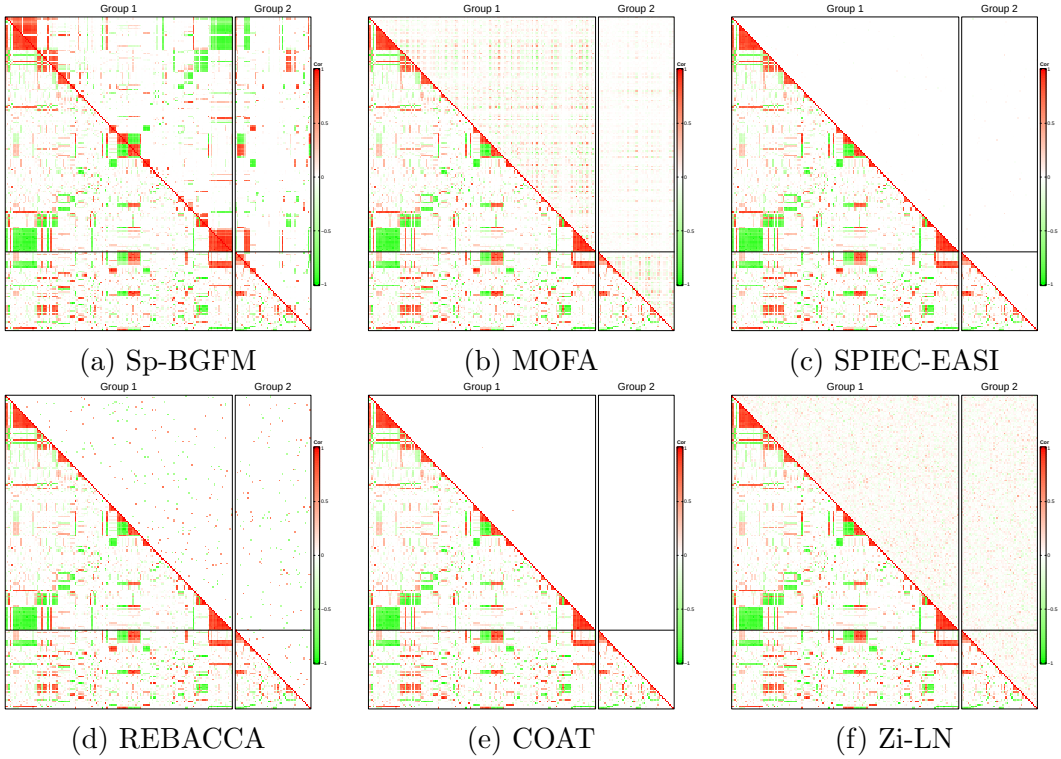
one from each of the levels, resulting in a total of $N = 40$ samples. We used the vine method in Lewandowski et al. (2009) to generate an arbitrary Σ^{tr}. In particular, we simulated partial correlations from linearly transformed $\text{Be}(1, 1)$ distribution over the interval of $(-1, 1)$. To encourage 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 $\text{Unif}(1, 1.5)$ and let $\Sigma_{jj'}^{mm',\text{tr}} = v_{mj}^{\text{tr}} v_{m'j'}^{\text{tr}} \rho_{jj'}^{mm',\text{tr}}$. For abundances, we computed the empirical proportions $\tilde{\psi}_{mj}$ of zero counts in the multi-domain skin microbiome dataset in §3.4 of the main text. 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 $\psi_{mj2}^{\text{tr}} = 0.6 \times (1 - \psi_{mj1}^{\text{tr}})$ and $\psi_{mj2}^{\text{tr}} = 0.4 \times (1 - \psi_{mj1}^{\text{tr}})$. In addition, we introduced a categorical covariate with two levels. For covariate effects, 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.

The upper triangles of the heatmaps in Fig B.7 illustrate the estimate of the correlations $\hat{\rho}_{jj'}$ obtained from REBACCA, COAT and Zi-LN, which are developed for single-domain analysis. The true values $\rho^{\text{tr}}_{jj'}$ of the correlations are shown in the lower triangles of the heatmaps. Compared to the estimate under Sp-BGFM in panel (a) of Fig 6 of the main text, their performance is poor. Note that they do not account for covariates in addition to not properly considering data from a multi-domain study.

B.5.3 Simulation 3

For Simulation 3, we kept $M = 2$, $J_1 = 150$, $J_2 = 50$, $S = 20$ and $N = 20$ the same as in Simulation 2, but removed the covariate to closely examine the estimation of Σ. Specifically, we used the vine method in Lewandowski et al. (2009) to have an arbitrarily specified Σ^{tr}. We used the empirical proportions $\tilde{\psi}_{mj}$ of zero counts from the multi-domain skin microbiome dataset to have a dataset close to the real dataset. 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 B.8(a) compares posterior estimates $\hat{\rho}^{mm'}_{jj'}$ of correlations (upper triangle) to the truth (lower triangle). Recall that $\rho^{mm',\text{tr}}_{jj'}$ 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 pre-



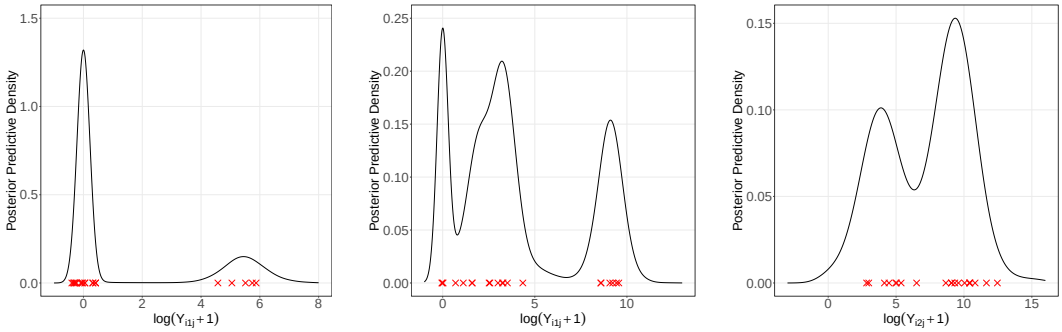
(a) Sp-BGFM (b) MOFA (c) SPIEC-EASI

(d) REBACCA (e) COAT (f) Zi-LN

Figure B.8: [Simulation 3] 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)-(f) are from Sp-BGFM, MOFA, SPIEC-EASI, REBACCA, COAT and Zi-LN, respectively.

dictive 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 some selected OTUs are illustrated with the normalized observed counts in crosses in Fig B.9. 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 B.8(b) and (c). The RMSE of $\rho_{jj'}^{mm'}$ is computed for Sp-BGFM, MOFA, and SPIEC-EASI, and is included in Tab 1 of the main text. Fig B.8



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

Figure B.9: [Simulation 3] 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}.

(d)-(f) compare 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.

B.5.4 Simulation 4

In Simulation 4, we further assessed the robustness of Sp-BGFM by simulating count vectors from a distribution different from the assumed model. Specifically, we simulated count vectors $\mathbf{y}_{im}$ from multinomial distributions, $\mathbf{y}_{im} \overset{indep}{\sim} \text{Multinomial}(\tilde{\boldsymbol{c}}_{im}, \tilde{N}_{im})$, $i = 1, \ldots, N$ and $m = 1, \ldots, M$, where $\tilde{\boldsymbol{c}}_m$ is a J_m-dim probability vector and $\tilde{N}_{im}$ the fixed total count. Let $N = 20$, $M = 2$, $J_1 = 150$, and $J_2 = 50$. To specify $\tilde{\boldsymbol{c}}_{im}$, we first generated counts c_{imj} independently from negative binomal distributions, $c_{imj} \overset{indep}{\sim} \text{NB}(\exp(\alpha_{imj}), s_{imj})$ with mean $\exp(\alpha_{imj}) > 0$ and dispersion parameter

$s_{imj} > 0$. We specified the values of α_{imj} by generating them from the same mixture distribution used in Simulation 1, and s_{mj}'s were sampled independently from log-N$(-2, 0.1^2)$. We then set $\tilde{c}_{imj} = c_{imj} / \sum_{j'=1}^{J_m} c_{imj'}$. The total counts $\tilde{N}_{im}$ were i.i.d sampled from N$(10.54, 1.68)$ and N$(9.26, 1.49)$ respectively for two groups. The values are the sample mean and variance of total counts of the multi-domain skin microbiome dataset in §3.4 of the main text. Note that the true dependence structure does not have any OTU interaction, and the total counts are fixed at $\tilde{N}_{im}$. We specified the fixed hyperparameter values similar to those in the previous simulation studies and applied Sp-BGFM to the dataset. MCMC was run for 100,000 iterations, with the initial half of iterations discarded as burn-in and the remaining half used for posterior inference.

The posterior estimate $\hat{\rho}_{jj'}^{mm'}$ of the correlation matrix is compared to the truth in Fig B.10(a). The model captures the pattern that the OTUs have no dependence structure well. To understand how the model-based normalization through r_{im} works, posterior mean estimates $\hat{r}_{im}$ of r_{im} are compared to the logarithm of total counts, $\log(\sum_{j=1}^{J_m} y_{ijm})$, $i = 1, \ldots, N$ and $m = 1, \ldots, M$ in Fig B.11. Note that due to the simulation setup, we have $\log(\sum_{j=1}^{J_m} y_{imj}) = \log(\tilde{N}_{im})$, where $\tilde{N}_{im}$ that are randomly generated. The figure illustrates that as the total count increases, $\hat{r}_{im}$ tends to increase, providing evidence that the model performs reasonable normalization to account for differences in total counts. Fig B.12 compares the posterior predictive distribution to the empirical distribution of observed counts for some selected OTUs. The observed counts are transformed for better illustration as described in §3.3.1 of the main text. The plot demonstrates that our model offers a reasonable fit, even in cases where the

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

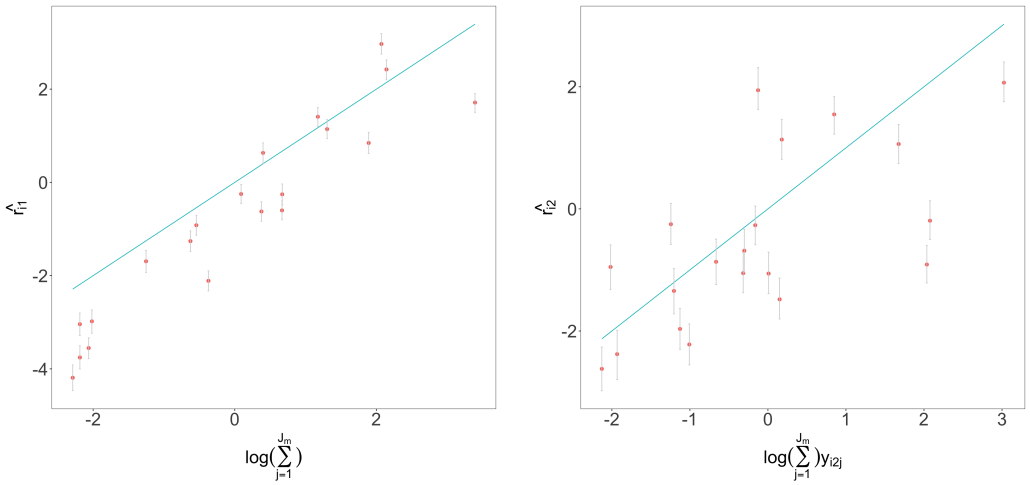
Figure B.10: [Simulation 4] The upper right and lower left triangles of a heatmap illustrate the estimates $\hat{\rho}_{jj'}^{mm'}$ of correlations and their truth, respectively. The horizontal and vertical lines are to divide the groups. The estimates in panels (a)-(c) are from Sp-BGFM, MOFA and SPIEC-EASI.

data were generated from a model significantly different from the assumed model.

For comparison, MOFA and SPIEC-EASI were fitted to the simulated data, and their estimates $\hat{\rho}_{jj'}^{mm'}$ of correlations are compared to the truth in Figs B.10(b)-(c). SPIEC-EASI produces the estimates close to zero, indicating no dependence structure. However, MOFA produces estimates that are not close to zero, especially for the OTUs in group 1, even though the true values are zero, and the inference under MOFA does not accurately capture the true dependence structure. The RMSE of $\rho_{jj'}$ is computed and presented in Tab 1 of the main text.

B.5.5 Simulation 5

In this simulation study, we utilized functions from the R package *SpiecEasi* (Kurtz et al., 2015) to generate a synthetic dataset. The package, available on the authors' GitHub page, provides a function that takes a real microbiome dataset and a correlation matrix as input to generate realistic synthetic OTU count data. It simulates



(a) Group 1 (b) Group 2

Figure B.11: [Simulation 4] The posterior mean estimates of r_{im} are plotted against the logarithm of the total counts, $\tilde{N}_{im} = \log(\sum_{j=1}^{J_m} y_{imj})$, $i = 1, \ldots, N$ and $m = 1$ or 2. Panels (a) and (b) correspond to the two groups, $m = 1$ and $m = 2$, respectively.

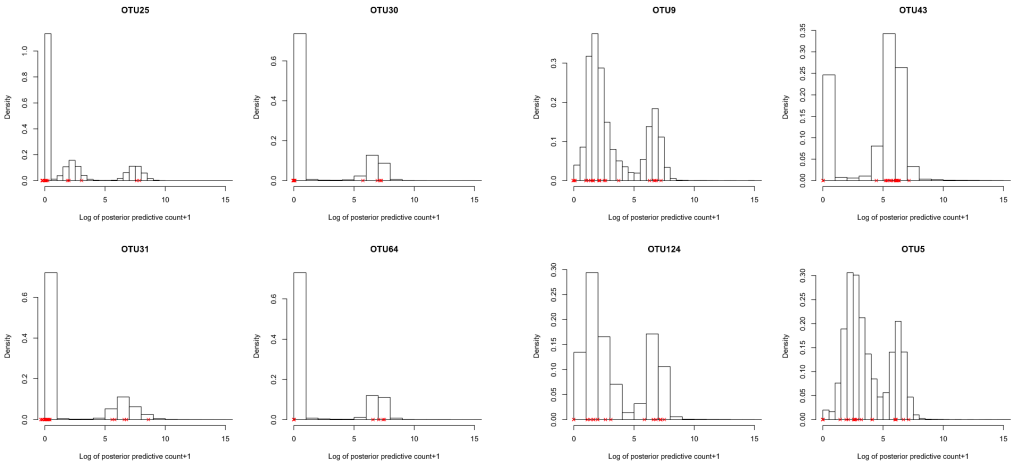
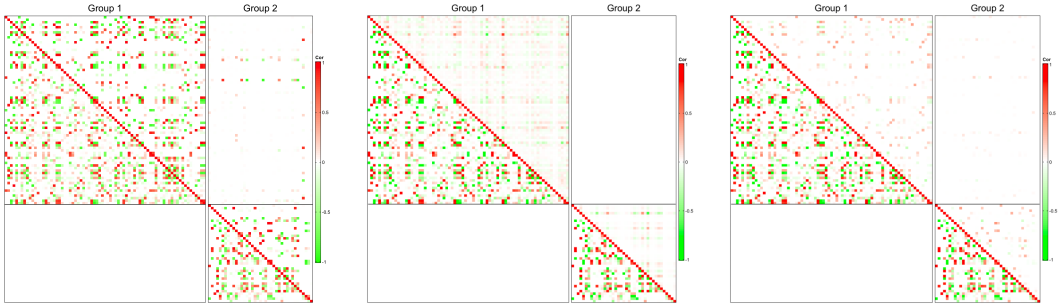


Figure B.12: [Simulation 4] Posterior predictive estimates of the marginal distribution of log-transformed counts are plotted for three arbitrarily chosen OTUs, OTUs 30 and 133 of group 1 and OTU 31 of group 2 for model checking. Crosses are log-transformed observed counts after normalization based on a posterior estimate of the scale factors r_{im}.

counts from zero-inflated negative binomial distributions using normal-copula functions. The simulated counts have a dependence structure as in the provided correlation matrix and marginally have a distribution similar to the empirical distribution of counts in the provided real data. The R functions do not consider complex data structures such as multiple domains, covariates and repeated samples from a subject. We thus generated a table for each group separately, and then concatenated the simulated count tables to have a multi-domain microbiome dataset. We used the bacterial and viral tables of the multi-domain skin microbiome dataset as a real data input, and we had $N = S = 60$, $M = 2$, $J_1 = 75$ and $J_2 = 39$. The covariates and potential dependence in repeated samples that the skin microbiome dataset has were not taken account of for both data simulation and analysis. We used the vine method in Lewandowski et al. (2009) to randomly generate within-domain dependence structure, $\Sigma^{\text{tr},mm}$. Since a dataset is generated separately for each domain, there is no cross-domain dependence among OTUs in the ground truth, i.e., $\Sigma^{\text{tr},mm'} = \mathbf{0}$, $m \neq m'$. Σ^{tr} is plotted in the lower triangles of Figs B.13. 72% of the counts are zero, similar to the zero rate of the multi-domain skin microbiome data similar to that of the skin microbiome dataset. We specified the values of the fixed hyperparameter similar to those in the previous simulation studies and ran MCMC for 100,000 iterations. The initial half of iterations was discarded as burn-in, and the second half was used for inference.

Posterior estimates $\hat{\rho}_{jj'}^{mm'}$ are in Fig B.13(a). The figure indicates that the true dependence structure among OTUs is well captured although the dataset was simulated from a model very different from the assumed model. Also, the posterior predictive



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

Figure B.13: [Simulation 5] The upper right and lower left triangles of each heatmap illustrate estimates $\hat{\rho}_{jj'}$ of correlations between OTUs and their true values $\rho^{\text{tr}}_{jj'} = 0$. Panels (a)-(c) are for Sp-BGFM, MOFA, and SPIEC-EASI, respectively.

distributions in Fig B.14 indicate that Sp-BGFM yields reasonable model fit.

For comparison, we applied SPIEC-EASI and MOFA to the simulated data and plotted their estimates $\hat{\rho}^{mm'}_{jj'}$ in Fig B.13 (b)-(c). Comparing their estimates to the truth indicates that they fail to recover the true dependence structure. A similar conclusion is obtained from comparing the RMSE of $\rho^{mm'}_{jj'}$ in Tab 1 of the main text across the three methods.

B.6 Additional Results from Multi-domain Skin Microbiome Data Analysis

Data exploration: In this section, we present additional results from the multi-domain skin microbiome data analysis. Fig B.15 illustrates empirical correlation estimates computed using the samples from each of the experimental conditions: healthy, pre-treatment and post-treatment. The log-transformed normalized counts were used. Fig B.16 illustrates histograms of the logarithm of the sample total counts for each

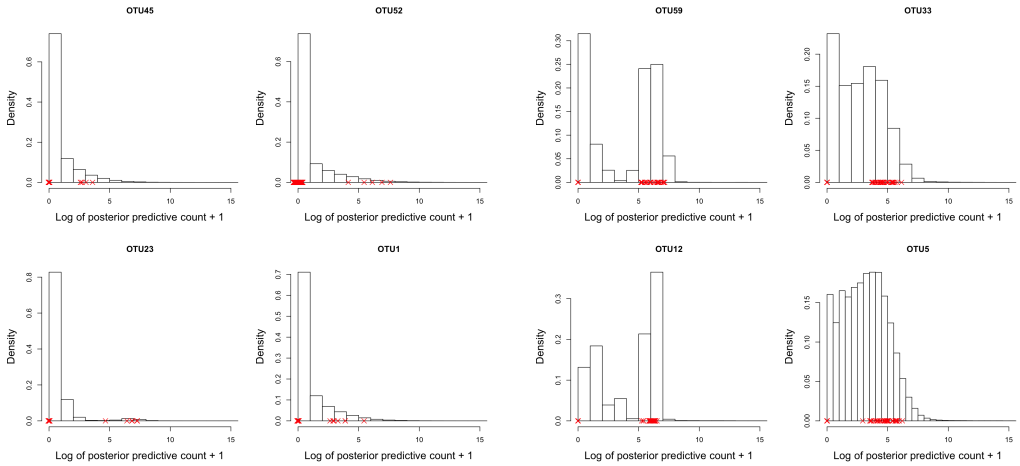
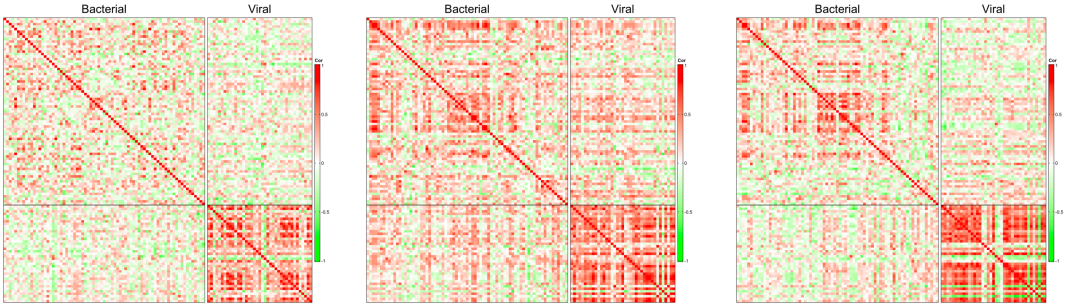


Figure B.14: [Simulation 5: Checking] Posterior predictive estimates of the marginal distribution of log-transformed counts are plotted for arbitrarily chosen OTUs for model checking. Crosses are log-transformed observed counts after normalization based on a posterior estimate of the scale factors r_{im}.

group (domain). The distributions are different by group, which indicates the need for separately modeling r_{im} for each group.

Additional results on $\hat{\rho}_{jj'}^{mm'}$: Fig B.17 illustrates $\hat{\rho}_{jj'}^{mm'}$ for the OTUs that have $|\hat{\rho}_{jj'}^{mm'}| > 0.5$ with any other OTU j', $j' \neq j$. Here, 0.5 is an arbitrary choice to illustrate a smaller set of OTUs that have large estimates. Tabs B.2 and B.3 have taxonomic information of the OTUs whose abundance changes statistically significantly by any of the experimental conditions or the OTUs that have $|\hat{\rho}_{jj'}^{mm'}| > 0.5$ with any other OTUs.

Predictive checking: Fig B.18 has posterior predictive density estimates of log-transformed counts for some selected OTUs, bOTU 1, bOTU 69 and vOTU 17, where black solid, red and blue dashed represent healthy, pre-treatment and post-treatment conditions, respectively. We set $r_m^{\text{pred}} = 0$ for $m = 1$ and 2. Red crosses represent

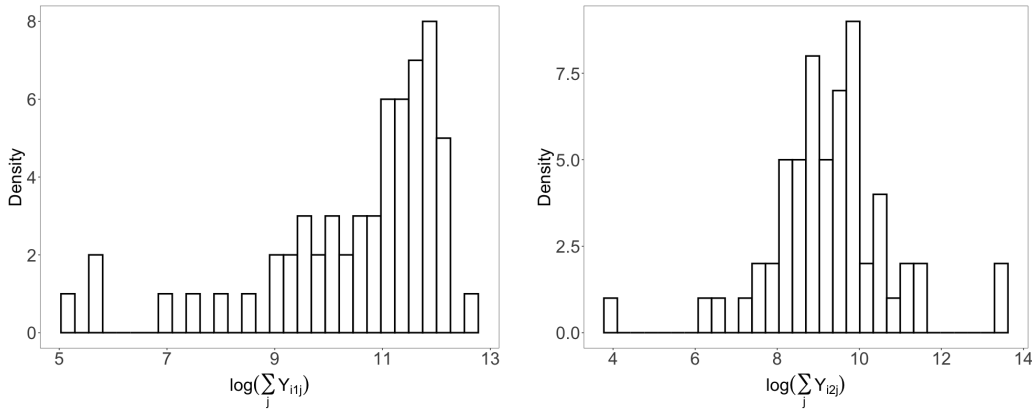


(a) Healthy (b) Pre-treatment (c) Post-treatment samples

Figure B.15: [Multi-domain skin microbiome data] Empirical correlation estimates $\tilde{\rho}_{jj'}^{mm'}$ are plotted in panels (a)-(c) for each experimental condition, healthy, pre-treatment, and post-treatment. The counts are normalized using cumulative sum scaling and log-transformed, with the addition of a pseudocount of 0.01 for the log-transformation.

log-transformed observed counts after normalization, $\log(\lfloor y_{imj}/\exp(\hat{r}_{im} - r_m^{\text{pred}})\rfloor + 1)$, where $\hat{r}_{im}$ is a posterior estimate of r_{im}. Posterior estimates of $\beta_{mj2} - \beta_{mj1}$, $\beta_{mj3} - \beta_{mj1}$ and $\beta_{mj3} - \beta_{mj2}$ are 0.340, 1.384 and 1.036 with 95% credible intervals (-0.914, 1.522), (0.227, 2.521) and (-0.134, 2.244), respectively for bOTU 1, -1.571, -1.633 and -0.062 with 95% credible intervals (-2.832, -0.321), (-2.869, -0.397) and (-1.294, 1.164), respectively for bOTU 69, and 5.118, 5.146 and 0.037 with 95% credible intervals (3.935, 6.299), (3.916, 6.372) and (1.078, 1.104), respectively for vOTU 17.

Additional comparison: Fig B.19 plots correlation estimates from the additional comparators, REBACCA, COAT and Zi-LN. Recall that the methods are developed for a single-domain microbiome data analysis and do not include covariates. Compared to the estimates under Sp-BGFM presented in the main text, the comparators produce very dense correlation estimates. Fig B.20 shows the estimates of coefficient effects $\beta_{mjp} - \beta_{mjp'}$ under metagenomeSeq.



(a) Bacterial group (b) Viral group

Figure B.16: [Multi-domain skin microbiome data] Histograms of the logarithm of the sample total counts $\log(\sum_j Y_{imj})$ are shown for the bacterial and viral groups in the left and right panels, respectively.

Assessing convergence of MCMC simulation: To assess the convergence of the MCMC simulation, we conducted Markov chain simulations with various initial values. Fig B.21 displays traceplots of the log-likelihood and some selected random parameters, v_1^2, $\beta_{1,2,2} - \beta_{1,2,1}$ and $\beta_{2,2,3} - \beta_{2,2,2}$. The traceplots indicate that the Markov chains converge to a similar state, providing practical evidence of the MCMC simulation's convergence.

Prior sensitivity analyses: We conducted sensitivity analyses to assess the robustness of Sp-BGFM to the fixed hyperparameter specifications. Specifically, we investigated the sensitivity of the posterior inference on $\rho_{jj'}^{mm'}$ to the values of K, a_ϕ, and a_τ, the hyperparameters of the priors on Σ. Recall that the results presented in §3.4 of the main text are obtained with $K = 15$, $a_\phi = 1/20$ and $a_\tau = 1/10$.

We first varied the value of K, setting $K = 13$, 17, and 20, while keeping a_ϕ

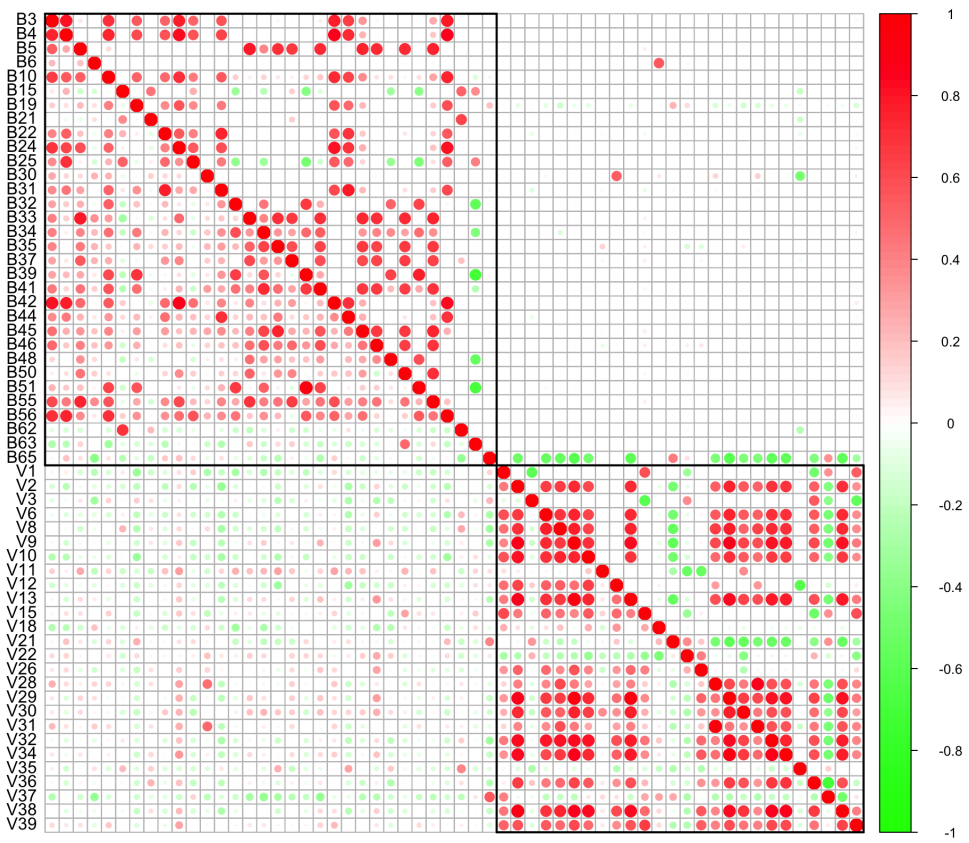
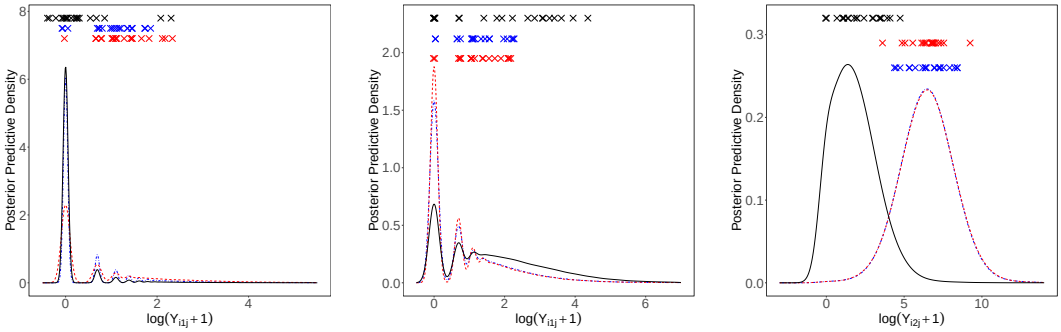


Figure B.17: [Multi-domain skin microbiome data] Posterior correlation estimates $\hat{\rho}_{jj'}^{mm'}$ (upper right triangle) and empirical correlation estimates $\widetilde{\rho}_{jj'}^{mm'}$ (lower left triangle) are plotted for the OTUs having $|\hat{\rho}_{jj'}^{mm'}| > 0.5$

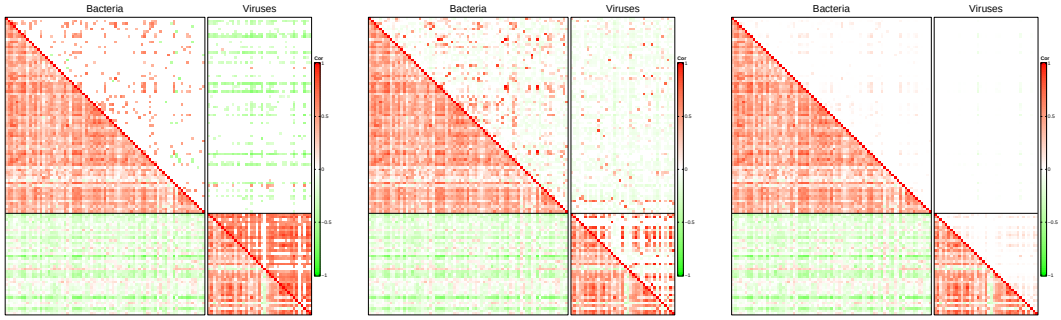
and a_τ the same. The traceplots of the log-likelihood in Fig B.22(a) show that log-likelihoods under $K = 15$, 17 and 20 converge to a similar state, but it has a much smaller value for $K = 13$. Posterior estimates $\hat{\rho}_{jj'}^{mm'}$ are presented in Fig B.22(b)-(e). From comparison with the estimates in Fig 9(a) of the main text, it is observed that the posterior estimates of $\rho_{jj'}^{mm'}$ show minimal changes across the different values of K.

We then examined how the posterior estimates of $\rho_{jj'}^{mm'}$ change by the spec-



(a) Bacterial OTU 1 (b) Bacterial OTU 69 (c) Viral OTU 17

Figure B.18: [Multi-domain skin microbiome data] In panels (a)-(c), posterior predictive density estimates of log-transformed counts $\log(y^{\text{pred}} + 1)$ are plotted for some OTUs. Solid, blue and red dashed lines denote healthy, pre-debridement and post-debridement conditions, respectively. Log-transformed observed counts are plotted with crosses after normalization.



(a) REBACCA (b) COAT (c) Zi-LN

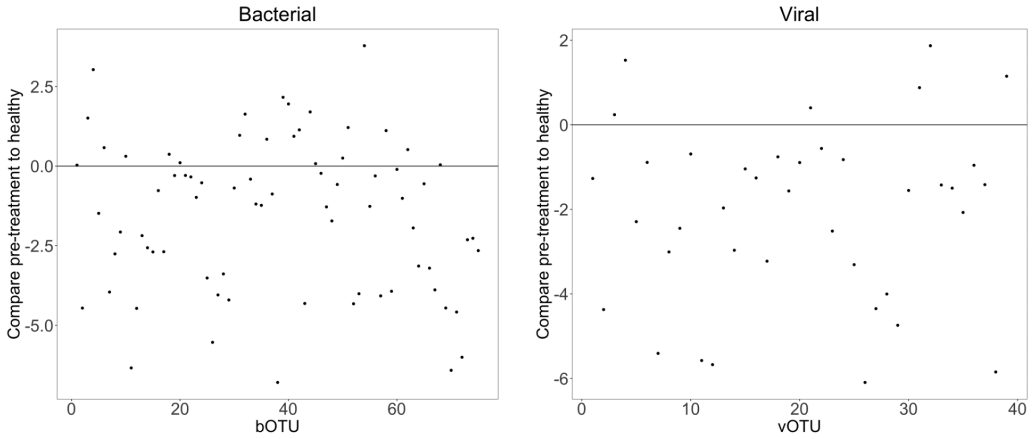
Figure B.19: [Multi-domain skin microbiome data] The upper right triangle of the heatmaps in panels (a)-(c) illustrates the correlations estimates $\hat{\rho}_{jj'}^{mm'}$ under REBACCA, COAT and Zi-LN, respectively. The lower left triangles have the empirical correlation estimate $\widetilde{\rho}_{jj'}^{mm'}$.

ification of a_ϕ. Specifically, we used three different values: $a_\phi = 1/2, 1/10$, and $1/50$. The traceplots of the log-likelihood in Fig B.23(a) show that the Markov chains with $a_\phi = 1/10, 1/20$, and $1/50$ reached a similar state, while the chain with $a_\phi = 1/2$ converged to a much smaller value. The posterior estimates $\hat{\rho}_{jj'}^{mm'}$ in Fig B.22(b)-(e) show that a larger value of a_ϕ results in less sparsity, and $\hat{\rho}_{jj'}^{mm'}$ with the values of

$a_\phi = 1/10, 1/20$, and $1/50$ do not change much.

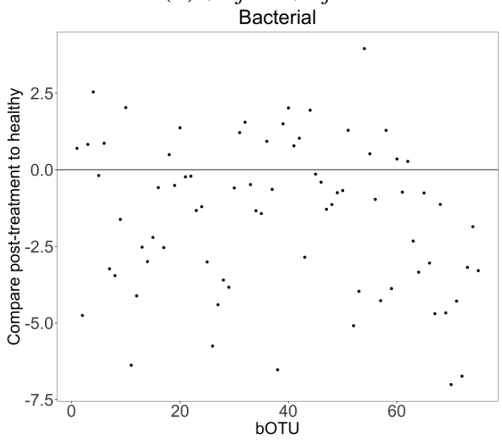
Lastly, Fig B.24 presents the results obtained by varying the value of a_τ. Specifically, we used $a_\tau = 1/100, 1/2, 1, 2$. The traceplots in Fig B.24(a) and the posterior estimates $\hat{\rho}_{jj'}^{mm'}$ in Fig B.24(b)-(d) demonstrate the robustness of the model under different specifications of a_τ.

In summary, while a small value of K leads to a significant reduction in computation, a complex dependence structure in Σ requires a large value of K. As suggested in §3.2.2 of the main text, one way to specify the value of K is by choosing a sufficiently large value based on principal component analysis using the sample covariance matrix. A large value of a_ϕ may result in negligibly small values of $\hat{\rho}_{jj'}^{mm'}$ for many pairs of j and j' (i.e., less sparsity). The model's performance does not change significantly within a reasonable range of values for a_τ.

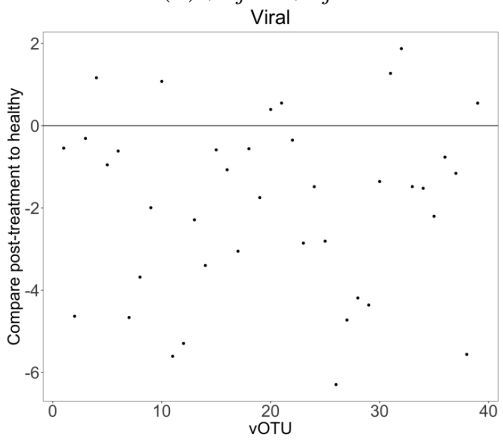


(a) $\beta_{1j2} - \beta_{1j1}$

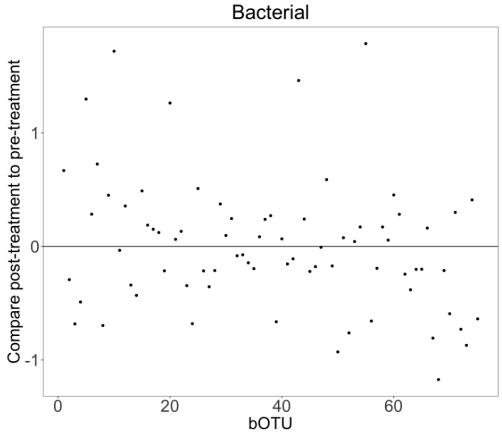
(b) $\beta_{2j2} - \beta_{2j1}$



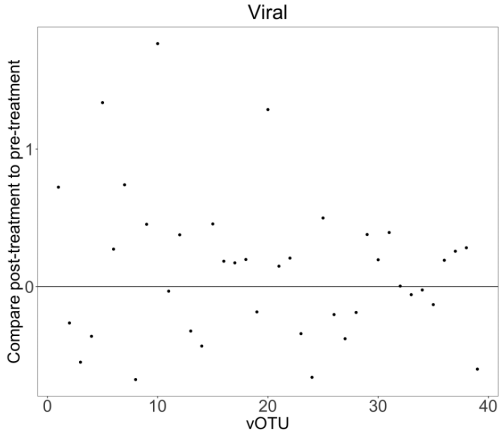
(c) $\beta_{1j3} - \beta_{1j1}$



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



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



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

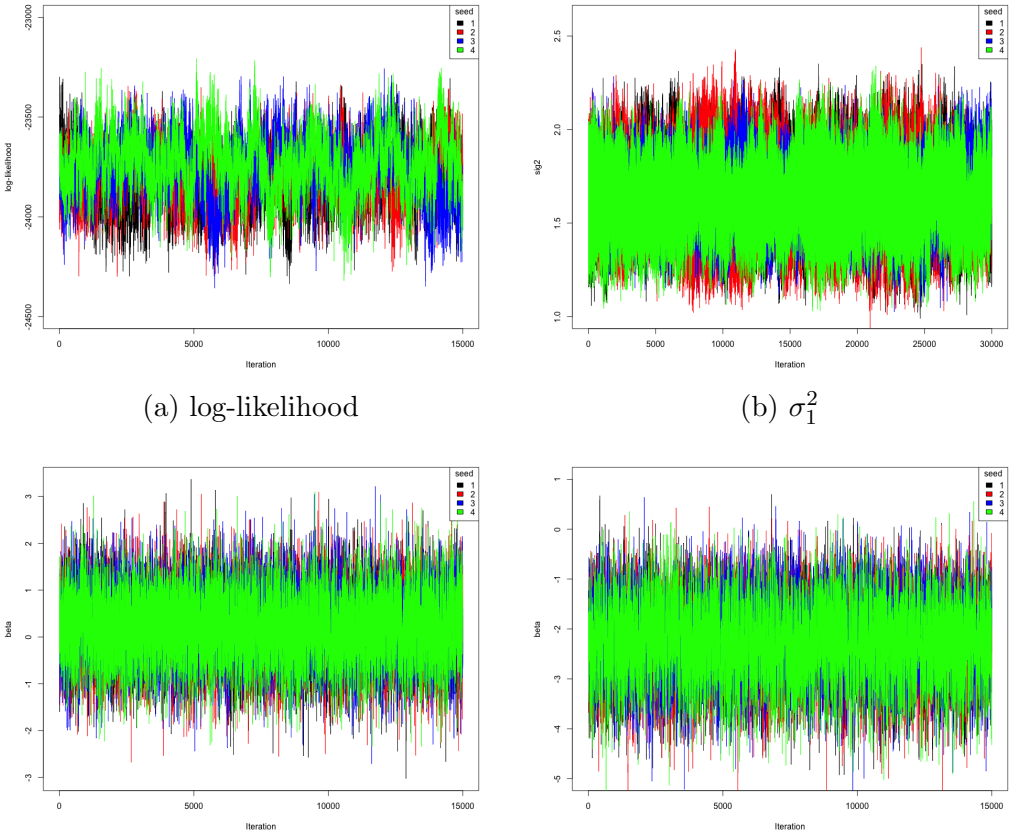
Figure B.20: [Multi-domain skin microbiome] The point estiamte of regression coefficient effect $\beta_{mjp} - \beta_{mjp'}$ under metagenomeSeq is plotted in panels (a) - (f).

Table B.2: [Multi-domain skin microbiome data] Taxonomic information of the bacterial OTUs whose abundance changes statistically significantly by any of the experimental conditions or the OTUs that have $|\hat{\rho}_{jj'}^{mm'}| > 0.5$ with any other OTUs. The OTUs that have $|\hat{\rho}_{jj'}^{mm'}| > 0.5$ and abundances significantly changing by an experimental condition are in blue. The OTUs that have $|\hat{\rho}_{jj'}^{mm'}| > 0.5$ and abundances significantly changing by an experimental condition are in blue *italic*.

OTU	Phylum	Class	Order	Family	Genus/Genus species
B1	Proteobacteria	Alphaproteobacteria	Rhizobiales	Brucellaceae	Unassigned
B2	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter
B3	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinomyces
B4	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinotignum
B5	Firmicutes	Clostridia	Clostridiales	Family XI	Anaerococcus
B6	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
B7	Actinobacteria	Actinobacteria	Micrococcales	Brevibacteriaceae	Brevibacterium
B8	Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	Brevundimonas
B9	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Burkholderia-Paraburkholderia
B10	Proteobacteria	Epsilonproteobacteria	Campylobacterales	Campylobacteraceae	Campylobacter
B11	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	Chryseobacterium
B12	Actinobacteria	Actinobacteria	Corynebacteriales	Corynebacteriaceae	Corynebacterium 1
B14	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Enhydrobacter
B15	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Enterobacter
B19	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Fastidiosipila
B20	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Unassigned
B21	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Unassigned
B22	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Glutamicibacter
B24	Firmicutes	Clostridia	Clostridiales	Family XI	Helcococcus
B25	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Unassigned
B26	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Kocuria
B27	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae	Massilia
B28	Proteobacteria	Alphaproteobacteria	Rhizobiales	Phyllobacteriaceae	Mesorhizobium
B29	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Micrococcus
B30	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Moraxella
B31	Actinobacteria	Actinobacteria	Micrococcales	Unassigned	Unassigned
B32	Unassigned	Unassigned	Unassigned	Unassigned	Unassigned
B33	Unassigned	Unassigned	Unassigned	Unassigned	Unassigned
B34	Unassigned	Unassigned	Unassigned	Unassigned	Unassigned
B35	Unassigned	Unassigned	Unassigned	Unassigned	Unassigned
B37	Actinobacteria	Actinobacteria	Micrococcales	Dermabacteraceae	Unassigned
B38	Unassigned	Unassigned	Unassigned	Unassigned	Unassigned
B39	Unassigned	Unassigned	Unassigned	Unassigned	Unassigned
B41	Unassigned	Unassigned	Unassigned	Unassigned	Unassigned
B42	Unassigned	Unassigned	Unassigned	Unassigned	Unassigned
B44	Firmicutes	Bacilli	Lactobacillales	Aerococcaceae	uncultured
B45	Unassigned	Unassigned	Unassigned	Unassigned	Unassigned
B46	Unassigned	Unassigned	Unassigned	Unassigned	Unassigned
B48	Unassigned	Unassigned	Unassigned	Unassigned	Unassigned
B50	Unassigned	Unassigned	Unassigned	Unassigned	Unassigned
B51	Unassigned	Unassigned	Unassigned	Unassigned	Unassigned
B52	Actinobacteria	Actinobacteria	Propionibacteriales	Nocardioidaceae	Nocardioides
B53	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Paracoccus
B55	Firmicutes	Clostridia	Clostridiales	Family XI	Peptoniphilus
B56	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	Porphyromonas
B57	Actinobacteria	Actinobacteria	Propionibacteriales	Propionibacteriaceae	Propionibacterium
B58	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Proteus
B60	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Ralstonia
B62	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Salmonella
B63	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Serratia
B64	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingomonas
B65	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus aureus
B67	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus cohnii
B70	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus hominis
B72	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus pettenkoferi

Table B.3: [Multi-domain skin microbiome data] Taxonomic information of the viral OTUs whose abundance changes statistically significantly by any of the experimental conditions or the OTUs that have $|\hat{\rho}_{jj'}^{mm'}| > 0.5$ with any other OTUs. The OTUs that have $|\hat{\rho}_{jj'}^{mm'}| > 0.5$ and abundances significantly changing by an experimental condition are in blue. The OTUs that have $|\hat{\rho}_{jj'}^{mm'}| > 0.5$ and abundances significantly changing by an experimental condition are in blue *italic*.

OTU	Type	Resolution
V1	Acinetobacter phage	Defined
V2	Aquisalimonas phage	Defined
V3	Bacillus phage	Defined
V6	Citrobacter phage	Defined
V8	Enterobacter phage	Defined
V9	Grimontella phage	Defined
V10	Klebsiella phage	Defined
V11	Leptotrichia phage	Defined
V12	Mannheimia phage	Defined
V13	Methylomonas phage	Defined
V15	Prevotella phage	Defined
V17	Proteus phage	Defined
V18	Pseudomonas phage	Defined
V21	Staphylococcus aureus phage	Defined
V22	Staphylococcus phage	Defined
V23	Staphylococcus phage	Defined
V24	Streptococcus phage	Defined
V26	Vibrio phage	Defined
V28	Unknown host type	Ambiguous
V29	Other phage	Ambiguous
V30	Other phage	Ambiguous
V31	Other phage	Ambiguous
V32	Other phage	Ambiguous
V33	Unknown host type	Ambiguous
V34	Other phage	Ambiguous
V35	Other phage	Ambiguous
V36	Other phage	Ambiguous
V37	Other phage	Ambiguous
V38	Other phage	Ambiguous
V39	Other phage	Ambiguous



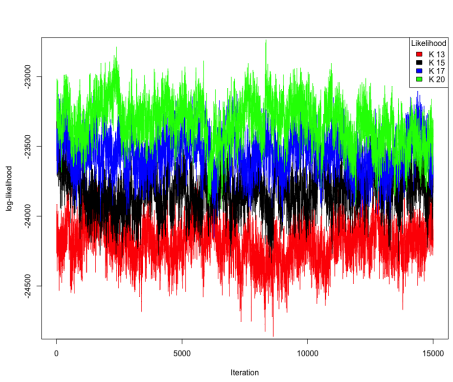
(a) log-likelihood

(b) σ_1^2

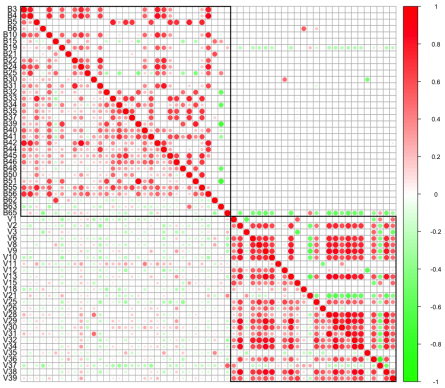
(c) $\beta_{1,2,2} - \beta_{1,2,1}$

(d) $\beta_{2,2,3} - \beta_{2,2,2}$

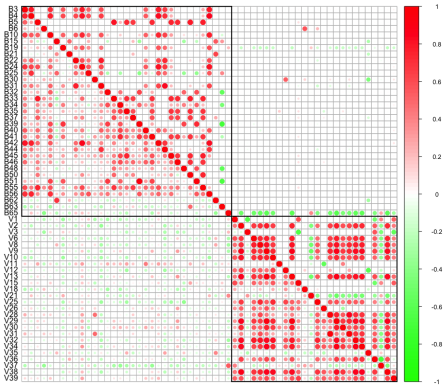
Figure B.21: [Convergence checking] Traceplots of log-likelihood and some selected parameters, v_1^2, $\beta_{1,2,2} - \beta_{1,2,1}$ and $\beta_{2,2,3} - \beta_{2,2,2}$. MCMC simulations were ran with four different initial values.



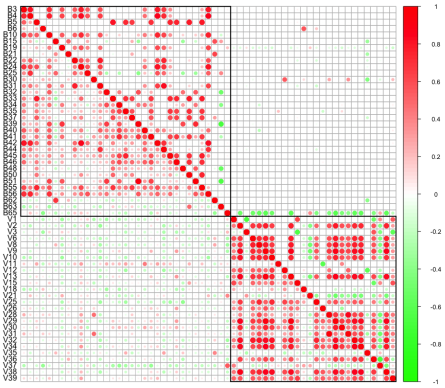
(a) Traceplot of log-likelihood



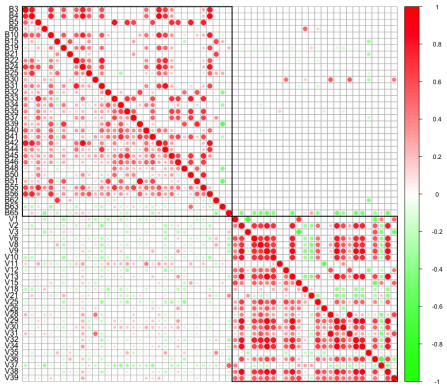
(b) $K = 13$



(c) $K = 15$

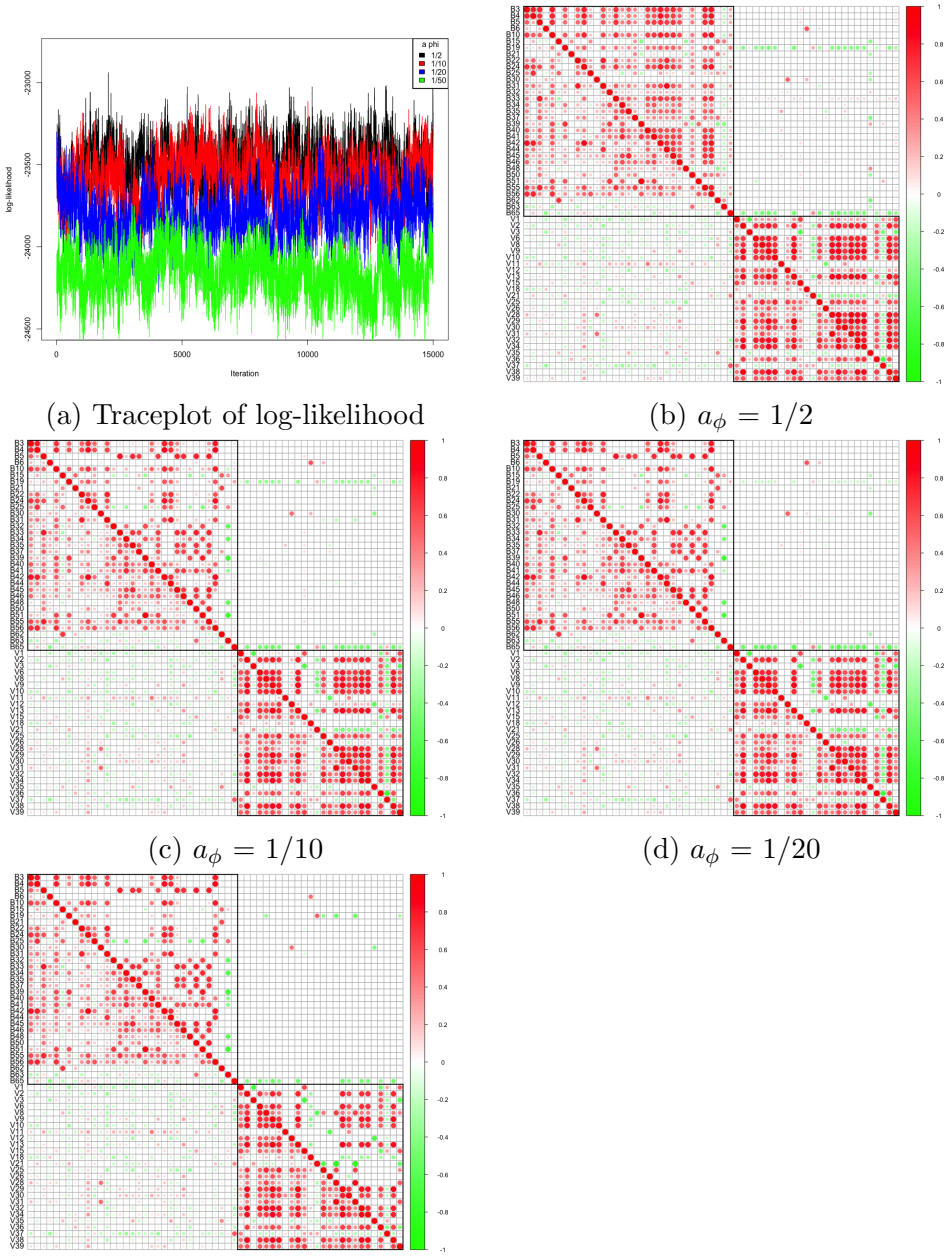


(d) $K = 17$



(e) $K = 20$

Figure B.22: [Sensitivity to the specification of K] Traceplots of log-likelihood under different values of K ($K = 13, 15, 17, 20$) are presented in distinct colors. In panels (b)-(e), posterior estimates $\hat{\rho}_{jj'}^{mm'}$ of the correlation matrix are displayed in the upper triangles for each value of K. Additionally, empirical correlation estimates are shown in the lower triangles. The estimates with $K = 15$ in § 3.4 of the main text are included for easy comparison.



(a) Traceplot of log-likelihood

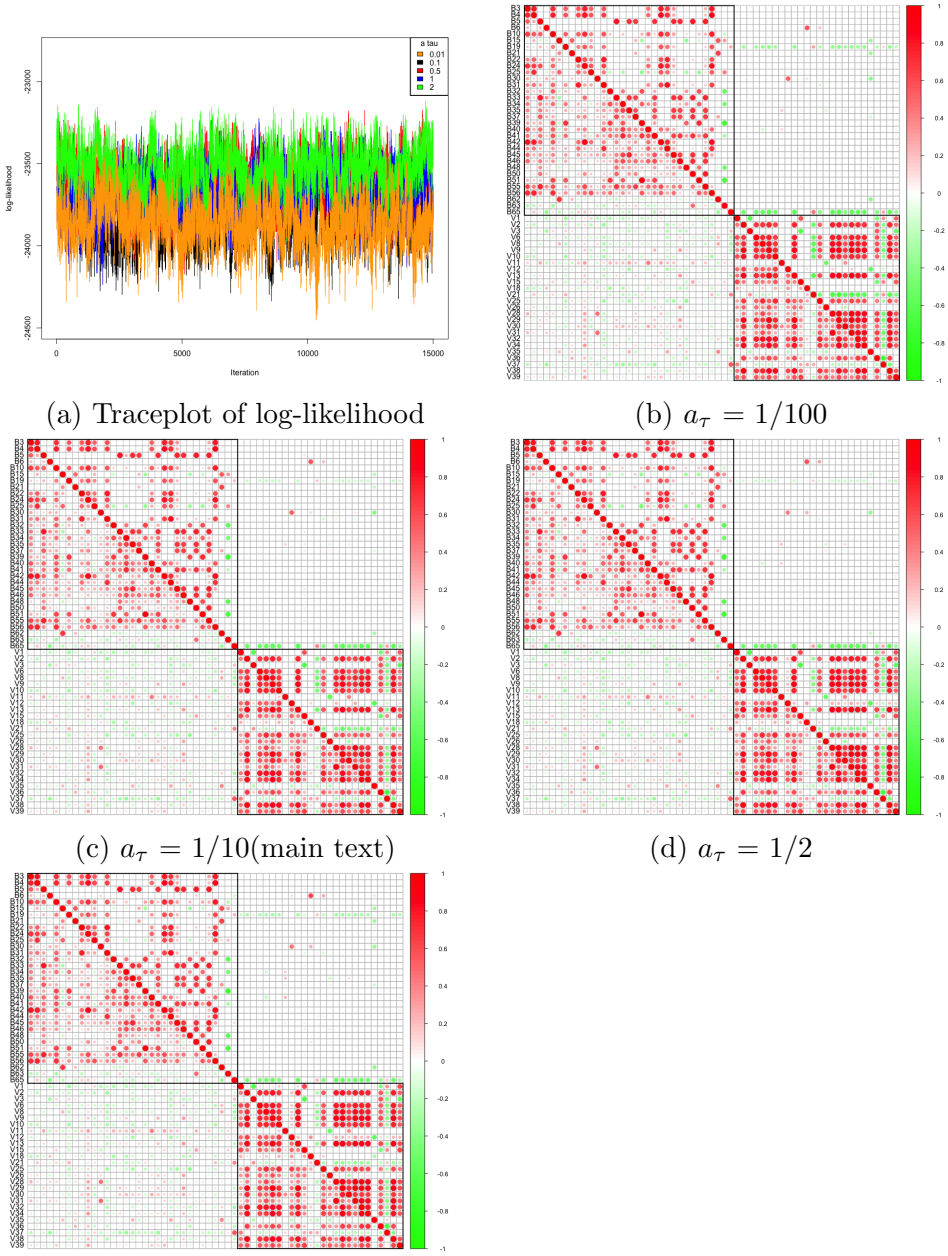
(b) $a_\phi = 1/2$

(c) $a_\phi = 1/10$

(d) $a_\phi = 1/20$

(e) $a_\phi = 1/50$

Figure B.23: [Sensitivity to the specification of a_ϕ] Traceplots of log-likelihood under different values of a_ϕ ($a_\phi = 1/2$, $1/10$, $1/20$, $1/50$) are presented in distinct colors. In panels (b)-(e), posterior estimates $\hat{\rho}_{jj'}^{mm'}$ of the correlation matrix are displayed in the upper triangles for each value of a_ϕ. Additionally, empirical correlation estimates are shown in the lower triangles. The estimates with $a_\phi = 1/20$ in § 3.4 of the main text are included for easy comparison.



(a) Traceplot of log-likelihood

(b) $a_\tau = 1/100$

(c) $a_\tau = 1/10$(main text)

(d) $a_\tau = 1/2$

(e) $a_\tau = 2$

Figure B.24: [Sensitivity to the specification of a_τ] Traceplots of log-likelihood under different values of a_τ ($a_\tau = 1/100,\ 1/10,\ 1/2,\ 2$) are presented in distinct colors. In panels (b)-(f), posterior estimates $\hat{\rho}_{jj'}^{mm'}$ of the correlation matrix are displayed in the upper triangles for each value of a_τ. Additionally, empirical correlation estimates are shown in the lower triangles. The estimates with $a_\tau = 1/10$ in § 3.4 of the main text are included for easy comparison.

Appendix C

SUPPLEMENTARY FOR Bayesian Covariate-Assisted Interaction Analysis for Multivariate Count Data in Microbiome Study

C.1 Details of Posterior Computation

We use Markov chain Monte Carlo (MCMC) techniques to obtain samples of the random parameters $\boldsymbol{\theta}$ from their posterior distributions, where $\boldsymbol{\theta} = \{q_{jk}, f_{kp}, \phi_{jk}, \tau_k, \zeta_{jk}, \sigma^2, \alpha_j, \omega_l^\alpha, V_l^\alpha, \xi_l^\alpha, r_i, \omega_l^r, V_l^r, \xi_l^r, \beta_{jp}\}$. To facilitate the posterior simulation, we introduce the latent continuous variable $y_{ij}^\star \in \mathbb{R}^+$ and have $y_{ij} = \lfloor y_{ij}^\star \rfloor$. We then impute

$y^{\star}_{ij} = \exp(\tilde{y}^{\star}_{ij})$ from a truncated log-normal distribution

$$\tilde{y}^{\star}_{ij} \mid \boldsymbol{\theta}, \boldsymbol{\eta}_i, y_{ij} \sim \mathrm{N}(r_i + \alpha_j + \boldsymbol{\lambda}'_j(\mathbf{x}_i)\boldsymbol{\eta}_i + \mathbf{x}'_i\boldsymbol{\beta}_j, \sigma^2)1(\log(y_{ij}) \leq \tilde{y}^{\star}_{ij} < \log(y_{ij}+1)).$$

Given $\tilde{y}^{\star}_{ij}$, parameters $\boldsymbol{\beta}_j$, $\boldsymbol{\eta}_i$, and σ^2 can be conveniently updated through normal/inv-gamma Gibbs steps. For ζ_{jk}, we utilize the following to achieve conjugacy (Makalic and Schmidt, 2015);

$$\zeta_{jk} \overset{iid}{\sim} \mathrm{C}^{+}(0,1) \;\Leftrightarrow\; \zeta^2_{jk} \mid Z_{jk} \overset{iid}{\sim} \text{inv-Ga}\left(\frac{1}{2}, \frac{1}{Z_{jk}}\right),\; Z_{jk} \overset{iid}{\sim} \text{inv-Ga}\left(\frac{1}{2}, 1\right). \qquad \text{(C.1)}$$

ζ_{jk} can be easily updated via Gibbs steps. Also, recall that parameters r_i and α_j are from infinite mixtures of mixtures. For computational convenience, when fitting the model, we approximate the infinite mixtures by truncating the number of mixture components to L^{α} and L^{r}. The final weights $\psi^{\alpha}_{L^{\alpha}} = 1 - \sum_{l=1}^{L^{\alpha}-1} \psi^{\alpha}_{l}$ and $\psi^{r}_{L^{r}} = 1 - \sum_{l=1}^{L^{r}-1} \psi^{r}_{l}$ is set to ensure the distributions are proper. With sufficiently large L^{α} and L^{r}, the truncated process produces inference almost identical to that with the infinite process (Ishwaran and James, 2001). We further introduce a pair of membership variables (I^{r}_{i1}, I^{r}_{i2}) with $I^{r}_{i1} \in \{1, \ldots, L^{r}\}$ and $I^{r}_{i2} \in \{0, 1\}$ for each r_i and $(I^{\alpha}_{j1}, I^{\alpha}_{j2})$ with $I^{\alpha}_{j1} \in \{1, \ldots, L^{\alpha}\}$ and $I^{\alpha}_{j2} \in \{0, 1\}$ for each α_j. We then assume $\mathrm{P}(I^{r}_{i1} = l) = \psi^{r}_{l}$ and $\mathrm{P}(I^{r}_{i2} = 0 \mid I^{r}_{i1} = l) = \omega^{r}_{l}$, and similarly, assume $\mathrm{P}(I^{\alpha}_{j1} = l) = \psi^{\alpha}_{l}$ and $\mathrm{P}(I^{\alpha}_{j2} = 0 \mid I^{\alpha}_{j1} = l) = \omega^{\alpha}_{l}$. Given the

membership indicator vectors, the conditional distributions of r_i and α_j are

$$r_i \mid \boldsymbol{\psi}^r, \boldsymbol{\omega}^r, \boldsymbol{\xi}^r, I^r_{i1} = l, I^r_{i2} \sim \begin{cases} \mathrm{N}(\xi^r_l, u^2_r) & \text{if } I^r_{i2} = 1, \\ \mathrm{N}(\frac{v^r - \omega^r_l \xi^r_l}{1-\omega^r_l}, u^2_r) & \text{if } I^r_{i2} = 0, \end{cases}$$

$$\alpha_j \mid \boldsymbol{\psi}^\alpha, \boldsymbol{\omega}^\alpha, \boldsymbol{\xi}^\alpha, I^\alpha_{j1} = l, I^\alpha_{j2} = \begin{cases} \xi^\alpha_l & \text{if } I^\alpha_{j2} = 1, \\ \frac{v^\alpha - \omega^\alpha_l \xi^\alpha_l}{1-\omega^\alpha_l} & \text{if } I^\alpha_{j2} = 0. \end{cases}$$

Given the latent variables, parameters $\boldsymbol{\phi}_k, f_{kp}$ are not updated through Gibbs steps. We update $\boldsymbol{\phi}_k, f_{kp}$ using a Metropolis-Hastings step. We let $\phi^\star_{jk} \overset{iid}{\sim} \mathrm{Ga}(a_\phi, 1)$ and have $\phi_{jk} = \phi^\star_{jk} / \sum_{m',\ j'} \phi^\star_{m'j'k}$. The full conditional of $\boldsymbol{\phi}_k$ is given by

$$p(\boldsymbol{\phi}_k \mid -) \propto p(\boldsymbol{q}_k \mid \tau_k, \boldsymbol{\phi}_k, \boldsymbol{\zeta}_k) p(\boldsymbol{\phi}_k) \propto \prod_{j=1}^{J} \mathrm{N}(q_{jk} \mid 0, \zeta^2_{jk} \phi_{jk} \tau_k) \prod_{j=1}^{J} \mathrm{Ga}(\phi^\star_{jk} \mid a_\phi, 1).$$

To efficiently update $\boldsymbol{\phi}_k$, the adaptive MH algorithm (Haario et al., 2001) is applied to adjust the MH step size according to the acceptance ratio, and the convergence rate is accelerated.

We sample sequentially by alternating conditional sampling. The full conditionals are given below;

- Update $\tilde{y}^\star_{ij}$ given $y_{ij}, r_i, \alpha_j, \boldsymbol{\lambda}_j(\mathbf{x}_i), \boldsymbol{\eta}_i, \sigma^2, \boldsymbol{\beta}_j, \mathbf{x}_i$

$$\tilde{y}^\star_{ij} \sim \mathrm{N}(r_i + \alpha_j + \boldsymbol{\lambda}'_j(\mathbf{x}_i)\boldsymbol{\eta}_i + \mathbf{x}'_i\boldsymbol{\beta}_j, \sigma^2) \mathbb{1}(\log(y_{ij}) \leq \tilde{y}^\star_{ij} < \log(y_{ij}+1)).$$

- parameters related to r_i

  - Update $\boldsymbol{\psi}^r$ given I^r_{i1}

    $\psi^r_1 = V^r_1, \psi^r_l = V^r_l \prod_{h=1}^{l-1}(1-V^r_h)$, for $l = 2, \ldots, L^r - 1$, $\psi^r_{L^r} = 1 - \sum_{l=1}^{L^r-1} \psi^r_l$

    $$V^r_l \sim \text{Be}(1 + \sum_{i=1}^{N} \mathbb{1}(I^r_{i1} = l), \ldots, c^r + \sum_{i=1}^{N}\sum_{h>l} \mathbb{1}(I^r_{i1} = h)).$$

  - Update ω^r_l given I^r_{i1}, I^r_{i2}

    $$p(\omega^r_l \mid -) \propto {\omega^r_l}^{a^r_\omega + \sum_{i=1}^N \mathbb{1}(I^r_{i1}=l, I^r_{i2}=1)} (1-\omega^r_l)^{b^r_\omega + \sum_{i=1}^N \mathbb{1}(I^r_{i1}=l, I^r_{i2}=0)} \prod_{i=1}^{N}\prod_{j=1}^{J} \text{N}(\tilde{y}^\star_{ij} \mid \mu_{ij}, \sigma^2). \tag{C.2}$$

    We use logistic transformation and adaptive Metropolis-Hasting algorithm (Haario et al., 2001) to update ω^r_l.

  - Update (I^r_{i1}, I^r_{i2}) given $\psi^r_l, \omega^r_l, r_i, v^r_m, \xi^r_l, u^2_r$

    $$\begin{aligned} \Pr(I^r_{i1} = l, I^r_{i2} = 1) &\propto \psi^r_l \omega^r_l \text{N}(r_i \mid \xi^r_l, u^2_r), \\ \Pr(I^r_{i1} = l, I^r_{i2} = 0) &\propto \psi^r_l (1-\omega^r_l) \text{N}(r_i \mid \frac{\nu^r - \omega^r_l \xi^r_l}{1-\omega^r_l}, u^2_r), \end{aligned}$$

  - Update ξ^r_l given $I^r_{i1}, I^r_{i2}, r_i, \omega^r_l$

    $$\xi^r_l \sim \text{N}\left(\tilde{u}^2_{\xi^r} (\frac{\nu^r}{u^2_{\xi^r}} + \sum_{i: I^r_{i1}=l, I^r_{i2}=1} \frac{r_i}{u^2_r} - \sum_{i: I^r_{i1}=l, I^r_{i2}=0} \frac{\frac{\omega^r_l}{1-\omega^r_l} r_i - \frac{\omega^r_l}{(1-\omega^r_l)^2}\nu^r}{u^2_r}), \tilde{u}^2_{\xi^r} \right),$$

where $\tilde{u}^2_{\xi^r} = (1/u^2_{\xi^r} + \sum_{i=1}^N \mathbb{1}(I^r_{i1} = l, I^r_{i2} = 1)/u^2_r + \omega^{r,2}_l \sum_{i=1}^N \omega^r_l \mathbb{1}(I^r_{i1} = l, I^r_{i2} = 0)/u^2_r(1-\omega^r_l)^2)^{-1}$.

– Update r_i given $\alpha_j, \boldsymbol{\lambda}_j(\mathbf{x}_i), \boldsymbol{\eta}_i, \boldsymbol{\beta}_j$

$$r_i \sim \mathrm{N}((\frac{c}{u^2_r} + \frac{\sum_{j=1}^J (\tilde{y}^\star_{ij} - \alpha_j - \boldsymbol{\lambda}'_j(\mathbf{x}_i)\boldsymbol{\eta}_i - \mathbf{x}'_i\boldsymbol{\beta}_j)}{\sigma^2})(\frac{1}{u^2_r} + \frac{J}{\sigma^2})^{-1}, \\ (\frac{1}{u^2_r} + \frac{J}{\sigma^2})^{-1}),$$

where prior mean $c = \sum_{j=1}^J (\mathbb{1}(I^r_{i2} = 1)\xi^r_{I^r_{i1}} + \mathbb{1}(I^r_{i2} = 0)\frac{\nu^r - \omega^r_{I^r_{i1}}\xi^r_{I^r_{i1}}}{1-\omega^r_{I^r_{i1}}})$.

- parameters related to α_j

– Update $\boldsymbol{\psi}^\alpha$ given I^α_{j1}

$\psi^\alpha_1 = V^\alpha_1, \psi^\alpha_l = V^\alpha_l \prod_{h=1}^{l-1}(1-V^\alpha_h)$, for $l = 2, \ldots, L^\alpha - 1$, $\psi^\alpha_{L^\alpha} = 1 - \sum_{l=1}^{L^\alpha - 1} \psi^\alpha_l$

$$V^\alpha_l \sim \mathrm{Be}(1 + \sum_{j=1}^J \mathbb{1}(I^\alpha_{j1} = l), \ldots, c^\alpha + \sum_{j=1}^J \sum_{h>l} \mathbb{1}(I^r_{j1} = h)).$$

– Update ω^α_l given $I^\alpha_{j1}, I^\alpha_{j2}$

$$p(\omega^\alpha_l \mid -) \propto \omega^{\alpha a^\alpha_\omega + \sum_{j=1}^J \mathbb{1}(I^\alpha_{j1}=l, I^\alpha_{j2}=1)}_l (1-\omega^\alpha_l)^{b^\alpha_\omega + \sum_{j=1}^J \mathbb{1}(I^\alpha_{j1}=l, I^\alpha_{j2}=0)} \\ \prod_{i=1}^N \prod_{j=1}^J \mathrm{N}(\tilde{y}^\star_{ij} \mid \mu_{ij}, \sigma^2).$$

We use logistic transformation and adaptive Metropolis-Hasting algorithm (Haario et al., 2001) to update ω^α_l.

– Update $(I_{j1}^{\alpha}, I_{j2}^{\alpha})$ given $\psi_l^{\alpha}, \omega_l^{\alpha}$

$$\Pr(I_{j1}^{\alpha} = l, I_{j2}^{\alpha} = 1) \propto \psi_l^{\alpha}\omega_l^{\alpha} \prod_{i=1}^{N} \mathrm{N}(\tilde{y}_{ij}^{\star} \mid r_i + \xi_l^{\alpha} + \boldsymbol{\lambda}_j'(\mathbf{x}_i)\boldsymbol{\eta}_i + \mathbf{x}_i'\boldsymbol{\beta}_j, \sigma^2),$$

$$\Pr(I_{j1}^{\alpha} = l, I_{j2}^{\alpha} = 0) \propto \psi_l^{\alpha}(1-\omega_l^{\alpha}) \prod_{i=1}^{N} \mathrm{N}(\tilde{y}_{ij}^{\star} \mid r_i + \tfrac{\nu_j^{\alpha} - \omega_l^{\alpha}\xi_l^{\alpha}}{1-\omega_l^{\alpha}} + \boldsymbol{\lambda}_j'(\mathbf{x}_i)\boldsymbol{\eta}_i + \mathbf{x}_i'\boldsymbol{\beta}_j, \sigma^2).$$

– Update ξ_l^{α} given $\tilde{y}_{ij}^{\star}, r_i, \boldsymbol{\lambda}_j(\mathbf{x}_i), \boldsymbol{\eta}_i, \mathbf{x}_i, \boldsymbol{\beta}_j$

$$\xi_l^{\alpha} \sim \mathrm{N}(\tilde{u}_{\alpha}^2(\nu_j^{\alpha}/u_{\alpha}^2 + \sum_{i:I_{j1}^{\alpha}=l, I_{j2}^{\alpha}=1} (\tilde{y}_{ij}^{\star} - r_i - \boldsymbol{\lambda}_j'\boldsymbol{\eta}_i - \mathbf{x}_i'\boldsymbol{\beta}_j)/\sigma^2 - \sum_{i:I_{j1}^{\alpha}=l, I_{j2}^{\alpha}=0} (\tfrac{\omega_l^{\alpha}}{1-\omega_l^{\alpha}}(\tilde{y}_{ij}^{\star} - r_i - \boldsymbol{\lambda}_j'\boldsymbol{\eta}_i - \mathbf{x}_i'\boldsymbol{\beta}_j) - \tfrac{\omega_l^{\alpha}}{(1-\omega_l^{\alpha})^2}\nu_j^{\alpha})/\sigma^2), \tilde{u}_{\alpha}^2),$$

where $\tilde{u}_{\alpha}^2 = (1/u_{\alpha}^2 + N\sum_{j=1}^{J} \mathbb{1}(I_{j1}^{\alpha} = l, I_{j2}^{\alpha} = 1)/\sigma^2 + N\omega_l^{\alpha,2}\sum_{j=1}^{J} \omega_l^{\alpha}\mathbb{1}(I_{j1}^{\alpha} = l, I_{j2}^{\alpha} = 0)/\sigma^2(1-\omega_l^{\alpha})^2)^{-1}$.

- Update f_{kp} given $\tilde{\boldsymbol{Y}}_{ij}^{\star}, r_i, \alpha_j, \mathbf{x}_i, \boldsymbol{\beta}_j, \boldsymbol{\eta}$

Although f_{kp} has the posterior distribution in closed form, due to non-identifiability between f_{kp} and $\boldsymbol{q}_j$, we adaptive MH algorithm by proposing from a normal distribution. The full conditional of f_{kp} is given by

$$p(f_{kp} \mid -) \propto \prod_{i=1}^{N}\prod_{j=1}^{J} \mathrm{N}(\tilde{Y}_{ij}^{\star} \mid r_i + \alpha_j + \boldsymbol{\lambda}_j'(\mathbf{x}_i)\boldsymbol{\eta}_i + \mathbf{x}_i'\boldsymbol{\beta}_j, \sigma^2)\mathrm{N}(f_{kp} \mid 0, 1),$$

where $\lambda_{jk}(x_i) = q_{jk}\boldsymbol{f}_k'\mathbf{x}_i$, we reject or accept the proposal by utilizing the adaptive MH algorithm.

- Update $\boldsymbol{q}_j$ given $\tilde{\boldsymbol{Y}}_j^\star, \boldsymbol{r}, \alpha_j, \mathbf{X}, \boldsymbol{\beta}_j, \boldsymbol{\eta}, \boldsymbol{\zeta}_j, \boldsymbol{\phi}_j, \boldsymbol{F}$

$$\boldsymbol{q}_j \sim \mathrm{N}(V_q \sigma^{-2} (\mathbf{X}\boldsymbol{F}' \circ \boldsymbol{\eta})' (\tilde{\boldsymbol{Y}}_j^\star - \mathbf{r} - \boldsymbol{\alpha}_j \mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}_j), V_q),$$

where $V = \mathrm{diag}(\zeta_{j1}^2 \phi_{j1} \tau_1, \ldots, \zeta_{jk}^2 \phi_{jk} \tau_K)$, $V_q = (\sigma^{-2} (\mathbf{X}\boldsymbol{F}' \circ \boldsymbol{\eta})' (\mathbf{X}\boldsymbol{F}' \circ \boldsymbol{\eta}) + V^{-1})^{-1}$.

$$p(\boldsymbol{\phi}_k \mid -) \propto \prod_{j=1}^{J} \mathrm{N}(q_{jk} \mid 0, \zeta_{jk}^2 \phi_{jk} \tau_k) \prod_{j=1}^{J} \mathrm{Ga}(\phi_{jk}^\star \mid a_\phi, 1).$$

- Update $\boldsymbol{\phi}_k$ using adaptive MH algorithm by proposing from a normalized $\mathrm{Ga}(a_\phi, 1)$.

  We let $\phi_{jk}^\star \overset{iid}{\sim} \mathrm{Ga}(a_\phi, 1)$ and have $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_J) \sim \mathrm{Dir}(a_\phi, \ldots, a_\phi)$ with $\phi_{jk} = \phi_{jk}^\star / \sum_{j'} \phi_{j'k}^\star$. The full conditional of $\boldsymbol{\phi}_k$ is given by

$$p(\boldsymbol{\phi}_k \mid -) \propto \prod_{j=1}^{J} \mathrm{N}(q_{jk} \mid 0, \zeta_{jk}^2 \phi_{jk} \tau_k) \prod_{j=1}^{J} \mathrm{Ga}(\phi_{jk}^\star \mid a_\phi, 1).$$

  We reject or accept the proposal by utilizing the adaptive MH algorithm.

- Update ζ_{jk}^2 given $Z_{jk}, q_{jk}, \phi_{jk}, \tau_k$

$$\zeta_{jk}^2 \sim \text{inv-Ga}(1, 1/Z_{jk} + q_{jk}^2/(2\phi_{jk}\tau_k)).$$

- Update Z_{jk} given $\zeta_{jk} \overset{indep}{\sim} \text{inv-Ga}(1, 1 + 1/\zeta_{jk}^2)$.

- Update $\tau_k \mid q_{jk}, \zeta_{jk}, \phi_{jk}$

$$\tau_k \sim \text{Generalized inverse Gaussian } (a_\tau - J/2, 2b_\tau, \sum_{m=1}^{M}\sum_{j=1}^{J} q_{jk}^2/\zeta_{jk}^2\phi_{jk}).$$

- Update $\boldsymbol{\eta}_i$ given $\Lambda(\mathbf{x}_i), \tilde{\boldsymbol{Y}}_i^\star, \mathbf{r}_i, \boldsymbol{\alpha}_{s_i}, \boldsymbol{\beta}, \mathbf{X}_i$

$$\boldsymbol{\eta}_i \sim \text{N}((I_K + \frac{\Lambda'(\mathbf{x}_i)\Lambda(\mathbf{x}_i)}{\sigma^2})^{-1}\frac{\Lambda'}{\sigma^2}(\boldsymbol{Y}_i^\star - r_i\mathbf{1}_J - \boldsymbol{\alpha} - \boldsymbol{\beta}\mathbf{X}_i), (I_K + \frac{\Lambda'(\mathbf{x}_i)\Lambda(\mathbf{x}_i)}{\sigma^2})^{-1}).$$

- Update $\boldsymbol{\beta}_j$ given $\tilde{\boldsymbol{Y}}_j^\star, \mathbf{r}, \boldsymbol{\alpha}_j, \boldsymbol{\lambda}_j, \boldsymbol{\eta}, \tilde{\mathbf{X}}$

$$\boldsymbol{\beta}_j \sim \text{N}((\frac{\tilde{\mathbf{X}}'\tilde{\mathbf{X}}}{\sigma^2} + v_\beta^{-2}I_p)^{-1}\frac{\tilde{\mathbf{X}}'}{\sigma^2}(\tilde{\boldsymbol{Y}}_j^\star - \mathbf{r} - \alpha_j\mathbf{1}_n - \boldsymbol{\eta}\boldsymbol{\lambda}_j(\mathbf{X})), (\frac{\tilde{\mathbf{X}}'\tilde{\mathbf{X}}}{\sigma^2} + \frac{\beta I_p}{v^{-2}})^{-1}).$$

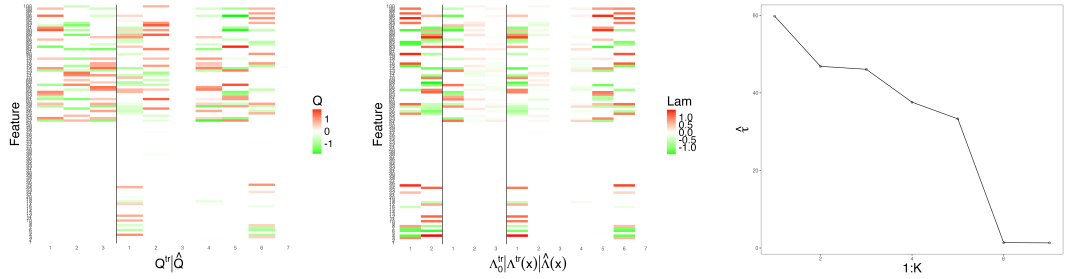
- Update σ^2 given $\tilde{y}_{ij}^\star, r_i, \alpha_j, \boldsymbol{\lambda}_j(\mathbf{x}_i), \boldsymbol{\eta}_i, \mathbf{x}_i, \boldsymbol{\beta}_j$

$$\sigma^2 \sim \text{inv-Ga}\left(a_\sigma + \frac{nJ}{2}, b_\sigma + \sum_{i=1}^{n}\sum_{j=1}^{J}(\tilde{y}_{ij}^\star - r_i - \alpha_j - \boldsymbol{\lambda}_j'(\mathbf{x}_i)\boldsymbol{\eta}_i - \mathbf{x}_i'\boldsymbol{\beta}_j)^2/2\right).$$

C.2 Additional Simulation Studies

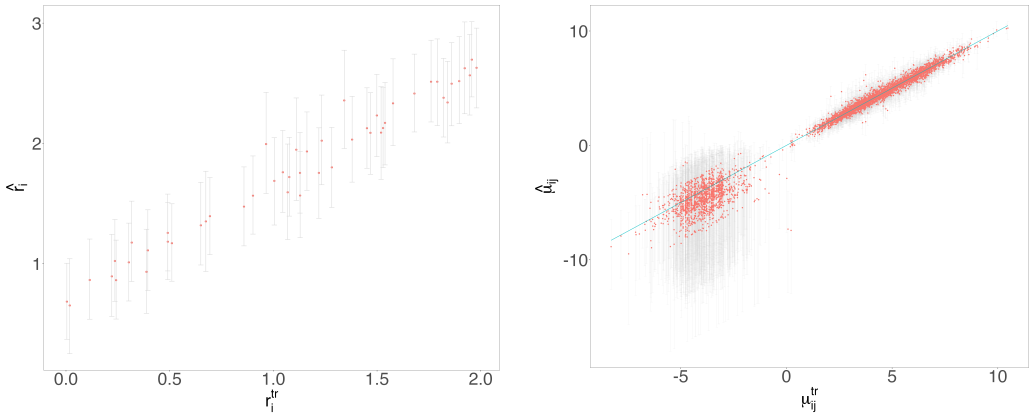
C.2.1 Additional results of Sim 1

The posterior median estimates of q_{jk} and posterior median estimates of τ_k are plotted in Fig C.1 (a) and (b). The estimate of q_{jk} and $\lambda_{jk}(x)$ both show a good estimate of baseline covariance. And Fig C.1(c) plots the posterior median estimates of τ_k in a decreasing order. As designed in the prior of q_{jk}, a small value of τ_k shrinks



(a) $\hat{q}_{jk}$ vs q_{jk}^{tr} (b) $\Lambda(x_1)$ vs $\Lambda^{\text{tr}}(x_1)$ (c) $\hat{\tau}_k$

Figure C.1: [Simulation 1] Heatmap of q_{jk}^{tr} and posterior median estimates $\hat{q}_{jk}$ are plotted in panel (a). In (b), we have a heatmap of $\Lambda_0^{\text{tr}}, \Lambda^{\text{tr}}(x), \hat{\Lambda}(x)$. We use sample 1 as an example. The scree plot of posterior estimates of τ_k is plotted in panel (c).



(a) $\hat{r}_i$ vs r_i^{tr} (b) $\hat{\mu}_{ij}$ vs μ_{ij}^{tr}

Figure C.2: [Simulation 1] The posterior median estimates of sample size factor r_i and mean abundance μ_{ij} are plotted against the truth in panels (a) and (b), respectively.

column-wise elements of q_{jk} toward 0. It leaves the redundant factor to 0, contributing little to the covariance. We also check the posterior estimates of sample size factor r_i versus the truth and the mean estimates μ_{ij} in Fig C.2. Our model provides accurate estimates of the mean abundance, serving as a reliable foundation for estimating the parameters of primary interest $\Sigma(x_i)$.

Table C.1: [Mice Gut Microbiome Data] OTUs information in the mice gut microbiome data.

OTU1	OTU2	OTU3
B_ovatus_ATCC.8483	B_sp_WH2	B_thetaiotaomicron_7330
OTU4	OTU5	OTU6
B_thetaiotaomicron_VPI.5482	B_vulgatus_ATCC.8482	Cat_Bacteroides_caccae_TSDC17
OTU7	OTU8	OTU9
Cat_Bacteroides_finegoldii_TSDC17	Cat_Bacteroides_massiliensis_TSDC17	Cat_Collinsella_aerofaciens_TSDC17
OTU10	OTU11	OTU12
Cat_Escherichia_coli_TSDC17	Cat_Odoribacter_splanchnicus_TSDC17	Cat_Parabacteroides_distasonis_TSDC17
OTU13	OTU14	OTU15
Cat_Ruminococcaceae_TSDC17	Cat_Ruminococcus_albus_TSDC17	Cat_Subdoligranulum_variabile_TSDC17

C.3 Additional Results of Mice Gut Microbiome Data

Tab C.1 provides information of the OTUs in the mice gut microbiome data.