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

Shuangjie Zhang *

Dept of Statistics, Uni. of California Santa Cruz

Yuning Shen

Chemical and Biomolecular Engineering Dept, Uni. of California Los Angeles

Irene A. Chen

Chemical and Biomolecular Engineering Dept, Uni. of California Los Angeles

Juhee Lee

Dept of Statistics, Uni. of California Santa Cruz

Abstract

Many statistical methods have been developed for the analysis of microbial community profiles, but due to the complexity of typical microbiome measurements, inference of interactions between microbial features remains challenging. We develop a Bayesian zero-inflated rounded log-normal kernel method to model interaction between microbial features in a community using multivariate count data, in the presence of covariates and excess zeros. The model carefully constructs the interaction structure by imposing joint sparsity on the covariance matrix of the kernel, and obtains a reliable estimate of the structure with a small sample size. The model also includes zero inflation to account for excess zeros observed in data and infers differential abundance of microbial features associated with covariates through log-linear regression. We provide simulation studies and real data analysis examples to demonstrate the developed model. Comparison of the model to a simpler model and popular alternatives in simulation studies shows

*Address for Correspondence: Department of Statistics, Baskin School of Engineering, University of California 1156 High Street Santa Cruz, CA 95064 USA. E-mail: szhan209@ucsc.edu.

that in addition to an added and important insight on the feature interaction, it yields superior parameter estimates and model fit in various settings.

Keywords: Covariance Matrix; Differential Abundance; Factor Model; Joint Sparsity; Multivariate Count Data; Rounded Kernel Model; Zero Inflation.

1 Introduction

16S ribosomal RNA (rRNA) sequencing technology in microbial ecology generates multivariate count data to characterize and analyze microbial communities from a variety of habitats such as human body sites, soil and water. 16S rRNA genes from complex communities in biological samples are PCR-amplified and sequenced using high-throughput sequencing (HTS). The sequence reads are then clustered based on their similarity into operational taxonomic units (OTUs), which represent bacteria types. Following some initial preprocessing procedures, 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. In the human gut microbiome data, one of our real data examples in § 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.

16S rRNA sequencing data 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 in samples. 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 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 are to infer

the associations of microbial abundances or their absence/presence with environmental factors, i.e., covariates, based on generalized regression models. For example, Poisson or negative binomial (NB) regression models are one of 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 a distribution with point mass of one at zero. See Li et al. (2017), Zhang et al. (2017), Jiang et al. (2021), Shuler et al. (2021) among many others, for examples of using Poisson or NB regression models. Another common approach uses multinomial or ZI multinomial regression 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, it is not clear the implication of the covariance among unconstrained probabilities on the marginal relationships between microbes due to the fixed total count constraint under the assumed multinomial distribution. Some approaches use a Dirichlet-tree multinomial model that exploits 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 pseudocount to observed counts and used regression to relate covariates to OTU abundances. Modeling microbial interactions is an important task in microbiome studies, but methods that infer the dependency structure between features in multivariate count data are limited. The task is often further complicated in microbiome studies due to a small sample size and the aforementioned challenges.

We develop a Bayesian rounded kernel regression model with zero inflation that enables a direct assessment of interrelationships between OTUs. 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 to the mean vector $\boldsymbol{\mu}$ of the distribution of $\mathbf{Y}^\star$ through regression, and use the covariance matrix Σ to infer interrelationship among OTUs. 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 developed joint sparse factor 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 to introduce 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 the remainder of the paper, we describe the model and its applications. § 2 describes the zero-inflated multivariate log-normal kernel model (called "ZI-MLN") and § 3 has results of simulation studies to evaluate the performance of our method. § 4 has results from the model applied to two real datasets, and § 5 concludes with some discussion of the results and areas of future research.

2 Statistical Model

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 § 4.1 consists of observed counts of 187 OTUs in 20 samples, a sample from each of 20 subjects. The dataset does not have covariates besides the subject factor. Human gut microbiome data in § 4.2 includes samples collected from multiple biopsy sites of 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 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

$$\mathbf{Y}_i^\star \mid \boldsymbol{\mu}_i, \Sigma \overset{indep}{\sim} \text{log-N}_J(\boldsymbol{\mu}_i, \Sigma), \tag{1}$$

where parameters $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iJ})' \in \mathbb{R}^J$ and $\Sigma > 0$. In (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 $\text{Cov}(Y_{ij}^\star, Y_{ij'}^\star) = \exp\{\mu_{ij} + \mu_{ij'} + \frac{1}{2}(\Sigma_{jj} + \Sigma_{j'j'})\}\{\exp(\Sigma_{jj'}) - 1\} = \text{E}(Y_{ij}^\star)\text{E}(Y_{ij'}^\star)\{\exp(\Sigma_{jj'}) - 1\}$. Overdispersion is known to be common in sequencing data and can be properly accommodated through heavy tails of a log-normal distribution. Also, the multivariate log-normal distribution provides a straightforward interpretation on the interaction structure between OTU counts through Σ. 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

$$\begin{aligned} \text{P}(\mathbf{Y}_i = \mathbf{y}_i \mid \boldsymbol{\mu}_i, \Sigma) = & \int_{A(\mathbf{y}_i)} f_{\mathbf{y}^\star}(\mathbf{y}^\star \mid \boldsymbol{\mu}_i, \Sigma) d\mathbf{y}^\star \\ = & \int_{\tilde{A}(\mathbf{y}_i)} \phi_J(\tilde{\mathbf{y}}^\star \mid \boldsymbol{\mu}_i, \Sigma) d\tilde{\mathbf{y}}^\star, \end{aligned} \tag{2}$$

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} such as their means and covariances can be easily computed from (2). For example, we find $\text{E}(Y_{ij} \mid \mu_{ij}, \Sigma_{jj}) = \sum_{b=0}^{\infty} b\{\Phi(\log(b+1) \mid \mu_{ij}, \Sigma_{jj}) - \Phi(\log(b) \mid \mu_{ij}, \Sigma_{jj})\}$, where $\Phi(\cdot \mid a, b^2)$ is the cdf of the normal distribution with mean a and variance b^2. A large value of μ_{ij} thus implies high abundance of OTU j in sample i. We relate $\mathbf{x}_i$ and g_i to OTU abundance through μ_{ij} and will give a regression model for $\boldsymbol{\mu}_i$ below. Similarly, we can compute covariance

$\text{Cov}(Y_{ij}, Y_{ij'} \mid \boldsymbol{\mu}_i, \Sigma)$. Under (2), the counts of OTUs j and j' are dependent if $\Sigma_{jj'} \neq 0$. In other words, microbial interactions are characterized through Σ.

We next build a prior distribution for Σ. The number J of OTUs is oftentimes much greater than sample size N in microbiome studies. To achieve reliable inference on Σ, we develop a joint sparse latent factor model. We decompose Σ as

$$\Sigma = \Lambda\Lambda' + \sigma^2 \mathrm{I}_J, \tag{3}$$

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. We let $\sigma^2 \sim \text{inv-Ga}(a_\sigma, b_\sigma)$ with fixed a_σ and b_σ. 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),

$$\begin{aligned} \tau_k \mid a_\tau, b_\tau &\overset{iid}{\sim} \text{Ga}(a_\tau, b_\tau), \\ \boldsymbol{\phi} = (\phi_1, \ldots, \phi_J) \mid a_\phi &\sim \text{Dir}(a_\phi, \ldots, a_\phi), \\ \lambda_{jk} \mid \phi_j, \tau_k &\overset{indep}{\sim} \text{DE}(\phi_j \tau_k), \end{aligned} \tag{4}$$

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 (4), a small value of ϕ_j more shrinks λ_{jk} toward zero for all k, and $\Sigma_{j,j'}$ 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 $N \ll J$. 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 a factor dimension K at a reasonably large value for computational convenience. If more desirable, an exponentially decaying prior such as a Poisson distribution can be placed for 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 does 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 since primary interest of inference is on Σ.

We re-write the model in (1) and (3) by introducing a latent normal vector $\boldsymbol{\eta}_i \overset{iid}{\sim} \mathrm{N}_K(0, \mathrm{I}_K)$;

$$\tilde{Y}_{ij}^{\star} \mid \mu_{ij}, \boldsymbol{\lambda}_j, \boldsymbol{\eta}_i, \sigma^2 \overset{indep}{\sim} \mathrm{N}_1(\mu_{ij} + \boldsymbol{\lambda}_j' \boldsymbol{\eta}_i, \sigma^2). \tag{5}$$

By integrating over $\boldsymbol{\eta}_i$, we obtain the normal distribution with covariance matrix Σ in (3) for $\tilde{\mathbf{Y}}_i^{\star}$. The conditional independence between $\tilde{Y}_{ij}^{\star}$ given $\boldsymbol{\eta}_i$ in (5) greatly facilitates the posterior computation. Furthermore, it enables to easily implement a zero-inflated model. Excess zeros in microbiome data are very common. If excess zeros are not compatible with the distribution in (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 $y = 0, 1, 2, \ldots$,

$$\begin{aligned} \mathrm{P}(Y_{ij} = y \mid \mu_{ij}, \boldsymbol{\lambda}_j, \boldsymbol{\eta}_i, \sigma^2, \delta_{ij} = 0) = \Phi(\log(y+1) \mid \mu_{ij} + \boldsymbol{\lambda}_j' \boldsymbol{\eta}_i, \sigma^2) \\ - \Phi(\log(y) \mid \mu_{ij} + \boldsymbol{\lambda}_j' \boldsymbol{\eta}_i, \sigma^2). \end{aligned} \tag{6}$$

Given the presence of an OTU, the model in (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,

$$\epsilon_{ij} = \Phi(\kappa_{j0} + \mathbf{x}_i'\boldsymbol{\kappa}_j, 1), \tag{7}$$

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}, $\kappa_{jp} \overset{iid}{\sim} \mathrm{N}(\bar{\kappa}_p, u_\kappa^2)$, $p = 0, \ldots, P$. The model in (7) does not include subject specific random effects. With a high proportion of zero counts, adding subject specific random effects into ϵ_{ij} may produce unstable model fitting (Agarwal et al., 2002). We thus do not include random effects in (7).

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

$$\mu_{ij} = r_i + \alpha_j + \mathbf{x}_i'\boldsymbol{\beta}_j + s_{g_i,j}. \tag{8}$$

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 abundance. r_i's normalize OTU counts across samples. Similarly, α_j's account for variability in OTU baseline abundances. Similar to the construction in Shuler et al. (2021), $r_i + \alpha_j$ constitutes the overall mean abundance of OTU j in sample i, and β_{jp} quantifies 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 (8) is different from that in (7) due to a different parameterization of the covariate. An example will be illustrated in § 3.2. When no covariate is available as in Simulation 1 in § 3.1 and the skin microbiome data in § 4.1, we simply drop the regression terms $\mathbf{x}_i'\boldsymbol{\beta}_j$ and $\mathbf{x}_i'\boldsymbol{\kappa}_j$ from (7) and (8), respectively, and use the simplified model to infer OTU interaction structure. $s_{g_i,j}$'s in (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}. 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 the random effects, 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 $\rho_{jj'} = \{\Sigma_{jj'} + u_s^2 1(j = j')\}/\sqrt{(\Sigma_{jj} + u_s^2)(\Sigma_{j'j'} + u_s^2)}$.

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;

$$\begin{aligned} r_i \mid \boldsymbol{\psi}^r, \boldsymbol{\omega}^r, \boldsymbol{\xi}^r &\overset{iid}{\sim} \sum_{l=1}^{L^r} \psi_l^r \left\{ \omega_l^r \mathrm{N}(\xi_l^r, u_r^2) + (1-\omega_l^r)\mathrm{N}\left(\frac{v_r - \omega_l^r \xi_l^r}{1-\omega_l^r}, u_r^2\right)\right\}, \\ \alpha_j \mid \boldsymbol{\psi}^\alpha, \boldsymbol{\omega}^\alpha, \boldsymbol{\xi}^\alpha &\overset{iid}{\sim} \sum_{l=1}^{L^\alpha} \psi_l^\alpha \left\{ \omega_l^\alpha \mathrm{N}(\xi_l^\alpha, u_\alpha^2) + (1-\omega_l^\alpha)\mathrm{N}\left(\frac{v_\alpha - \omega_l^\alpha \xi_l^\alpha}{1-\omega_l^\alpha}, u_\alpha^2\right)\right\}, \end{aligned} \tag{9}$$

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. 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. (2021) 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, $\chi \in \{r, \alpha\}$, $\boldsymbol{\psi}^\chi \sim \mathrm{Dir}(a_\psi^\chi, \ldots, a_\psi^\chi)$ and $\omega_l^\chi \overset{iid}{\sim} \mathrm{Be}(a_\omega^\chi, b_\omega^\chi)$, where the hyperparameters a_ψ^χ, a_ω^χ and b_ω^χ are fixed. Finally, we set $\xi_l^\chi \overset{iid}{\sim} \mathrm{N}(\bar{\xi}^\chi, v_\chi^2)$ 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 (9) flexibly capture various shapes of distributions, while keeping their means at v_χ. They thus provide reasonable estimates of $r_i + \alpha_j$ and may further improve estimates of the parameters of main interest including Σ, $\boldsymbol{\beta}_j$ and $\boldsymbol{\kappa}_j$.

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 (4) as a normal scale mixture to facilitate the posterior computation, and intro-

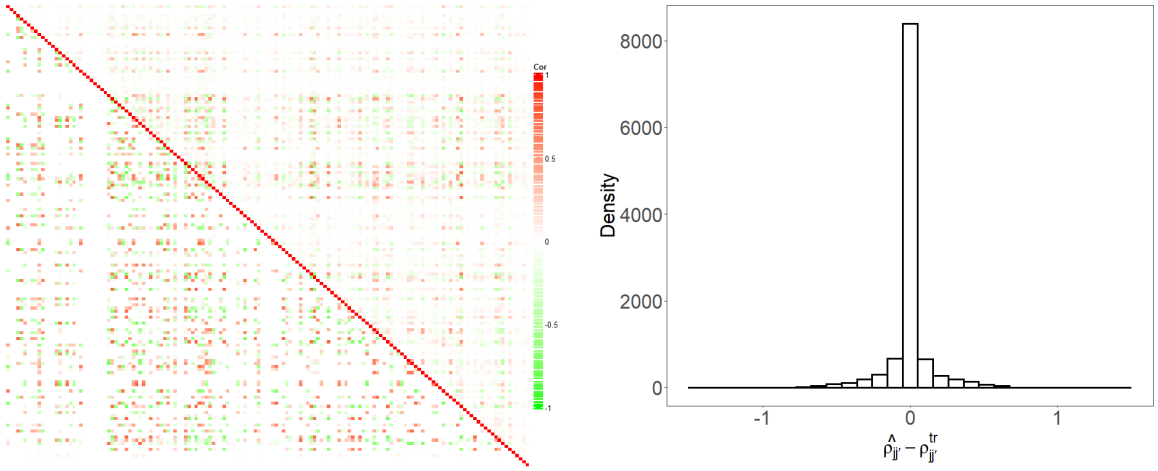
duce latent mixture indicators for easy computation in updating ω_l^χ, ψ_l^χ and ξ_l^χ. Given the latent variables, all parameters except for ϕ_j are in standard conjugate forms, and can be easily updated through a data augmented Gibbs step. Details of posterior computation are given in Supp. §1. We examined 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/Bayesian-Factor-Model.

3 Simulation Studies

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 have $\mu_{ij} = r_i + \alpha_j + s_{g_i,j}$ and $\epsilon_{ij} = \Phi(\kappa_{j0})$ to a simulated dataset. The simplified model is useful in estimating the interactions between OTUs for data without covariates. We let $J = 150$ OTUs and $N = 20$ samples, one 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 $\epsilon_j^{\text{tr}} = \Phi(\kappa_{j0}^{\text{tr}} \mid 0, 1)$. 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 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 Supp. Fig 1. It shows that naive correlation estimates are noisy and do not capture the true interrelationship between OTUs.

To fit the model, we set the hyperparameters as follows; For the mean-constrained priors of r_i and α_j, we let $L^r = 5, L^\alpha = 10$, $a_\psi^r = a_\psi^\alpha = 1$, and $a_\omega^r = b_\omega^r = a_\omega^\alpha = b_\omega^\alpha = 5$. The values of the mean constraints v^r and v^α were specified through the empirical approach described in § 2.1.

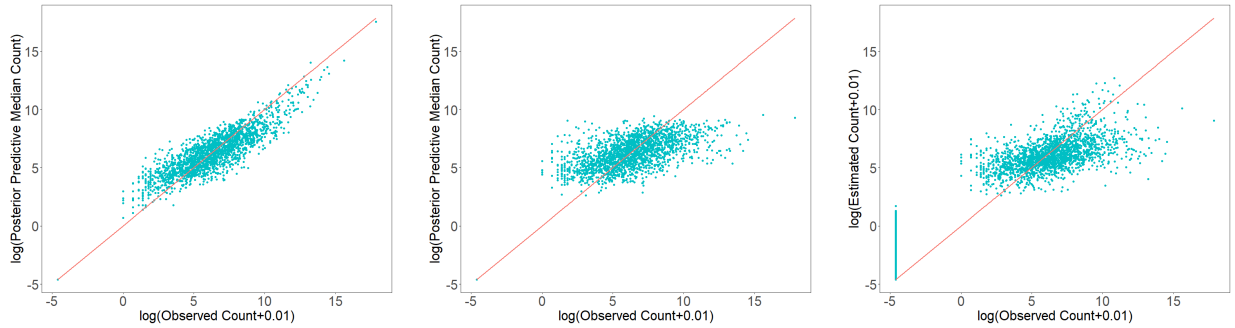


(a) $\rho^{\text{tr}}_{jj'}$ versus $\hat{\rho}_{jj'}$ (b) $\hat{\rho}_{jj'} - \rho^{\text{tr}}_{jj'}$

Figure 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 $\rho^{\text{tr}}_{jj'}$.

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. We discarded the first 15,000 draws for burn-in, and kept the next 15,000 draws for posterior inference. It took 0.7 hours for every 5,000 iterations on a 2.60 GHz Intel i7 laptop. We 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 data. We also performed sensitivity analyses to the specification of a_ϕ and b_τ to examine robustness of the model in estimating Σ.

Posterior inference on the marginal correlations $\rho_{jj'} = \{\Sigma_{jj'} + u^2_s 1(j = j')\}/\sqrt{(\Sigma_{jj} + u^2_s)(\Sigma_{j'j'} + u^2_s)}$ is illustrated in Fig 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 number of OTUs ($N = 20$ and $J = 150$) and excess zeros are present. Supp. Fig 2 compares posterior mean estimates of baseline abundances $r_i + \alpha_j$ and of the probabilities ϵ_{ij} of an OTU being absent to their



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

Figure 2: [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.

Table 1: [Simulation 1: Comparison] RMSEs are computed for binary indicator δ_{ij} of an OTU being absent in a sample and mean abundance μ_{ij} under our ZI-MLN, ZI-MLN without Λ and metagenomeSeq.

Model	δ_{ij}	μ_{ij}
ZI-MLN	**0.084**	**0.449**
ZI-MLN without Λ	0.088	0.543
MetagenomeSeq	0.095	1.717

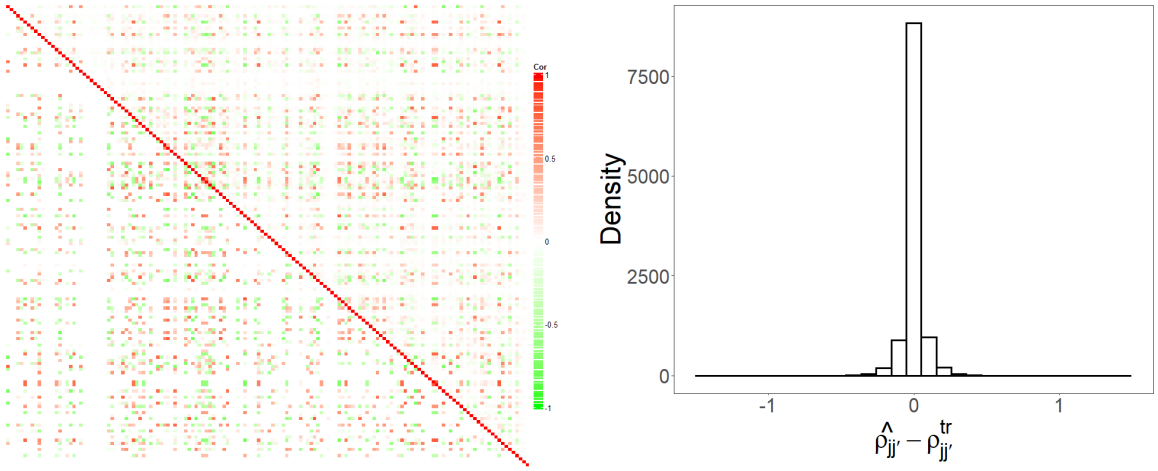
truth. In the figure, absence/presence of OTUs and OTU baseline abundances are well estimated, which provides a crucial basis for the estimation of the parameters of primary interest such as Σ. We performed posterior predictive checking to examine model fit under ZI-MLN. Fig 2(a) compares posterior predictive median estimates $\hat{y}_{ij}^{\text{pred}}$ of OTU counts to the observed counts y_{ij} and shows that our model provides good model fit to the data.

For comparison, we applied ZI-MLN without Λ, a simpler version of our ZI-MLN, and metagenomeSeq in Paulson et al. (2013). 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

probabilities of y coming from the component of the point mass at zero are common for all OTUs in a sample and regressed on the sample total counts through the logit link. An EM algorithm is used to estimate unknown parameters. The comparators do not account for the interrelationships between OTUs and do not provide any inference on OTU interaction. We compared parameter estimates under each of the three models including ZI-MLN to the truth and computed root mean square errors (RMSE) for parameters, δ_{ij} and μ_{ij}, summarized in Tab 1. The table shows that our model outperforms the comparators in the parameter estimation. 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(b). Comparison of mean abundance estimates $\hat{\mu}_{ij}$ by metegenomSeq to observed counts in Fig 2(c) also shows potential model misfit under metagenomeSeq.

3.2 Simulation 2

We conducted an additional simulation study, 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_{jk}^{tr}, r_i^{tr}, α_j^{tr} and $s_{g_i,j}^{\text{tr}}$, as done in Simulation 1. We included a binary covariate that represents the experimental conditions using a pair of dummy variables $(x_{i1}, x_{i2}) \in \{(1,0),(0,1)\}$. 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



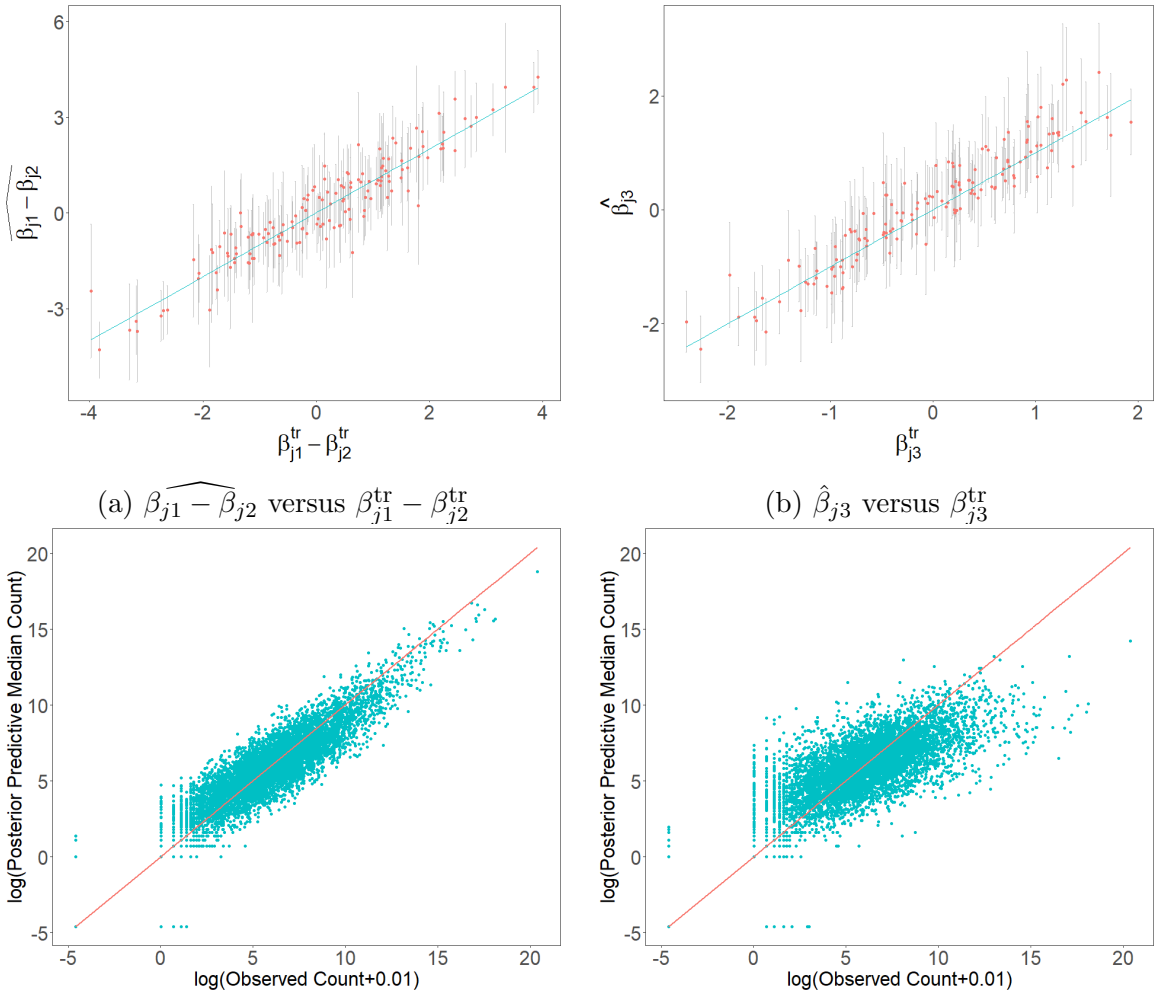
(a) $\rho^{\text{tr}}_{jj'}$ versus $\hat{\rho}_{jj'}$ (b) $\hat{\rho}_{jj'} - \rho^{\text{tr}}_{jj'}$

Figure 3: [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 $\rho^{\text{tr}}_{jj'}$.

simulated data has approximately 40% of counts being zero. Fig 3(a) and Supp. Fig 3 illustrate the true marginal correlations $\rho^{\text{tr}}_{jj'}$ and their naive empirical estimates $\rho^{\text{em}}_{jj'}$ using transformed counts after the normalization, respectively.

We specified hyperparameters 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. It took 1.3 hours on average for every 5,000 iterations on a 2.60 GHz Intel i7 laptop.

Fig 3 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. Fig 4(a)-(b) and Supp. Fig 4(a)-(c) compare regression coefficient estimates, $\hat{\beta}_{jp}$ and $\hat{\kappa}_{jp}$ to their true values. From Fig 4(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 the two categories of the binary covariate. Their posterior 95% credible intervals capture the truth well. Supp. Figs 5 show that posterior estimates $\widehat{r_i + \alpha_j}$ and $\hat{\epsilon}_{ij}$ are also close to their true values. To check model fit, we compare median estimates $\hat{y}^{\text{pred}}_{ij}$ of the posterior predictive



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

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

(c) ZI-MLN with Λ

(d) ZI-MLN without Λ

Figure 4: [Simulation 2] Panels (a) and (b) compare posterior estimates of regression coefficients $\widehat{\beta_{j1} - \beta_{j2}}$ and $\hat{\beta}_{j3}$ to the truth $\beta_{j1}^{\text{tr}} - \beta_{j2}^{\text{tr}}$ and β_{j3}^{tr}, 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 $\log(\hat{y}_{ij}^{\text{pred}} + 0.01)$. ZI-MLN with Λ and ZI-MLN without Λ are used in panels (c) and (d), respectively.

distributions to the observed counts. Fig 4(c) provides evidence for good model fit under ZI-MLN.

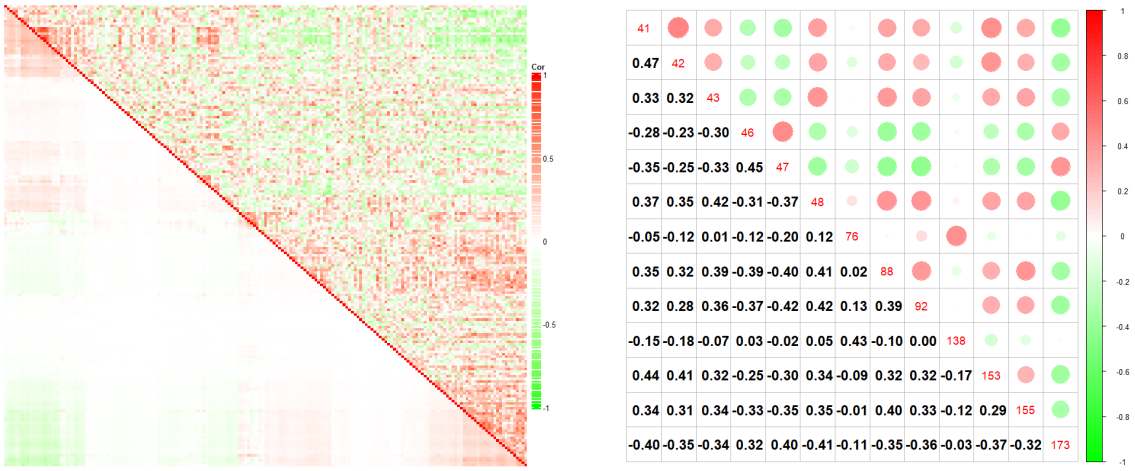
For comparison, we applied three comaprators, ZI-MLN without Λ, metagenomeSeq and edgeR (Robinson et al., 2010) to the simulated data. EdgeR is a likelihood-based method that uses a negative binomial generalized linear regression approach for analysis of HTS data. It uses the normalization factors estimated 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.

Table 2: [Simulation 2: Comparison] RMSEs are computed for δ_{ij}, μ_{ij}, $\beta_{j2}-\beta_{j1}$, β_{j3} and κ_{jp} under ZI-MLN, ZI-MLN without Λ, metagenomeSeq and edgeR.

Model	δ_{ij}	μ_{ij}	$\beta_{j2}-\beta_{j1}$	β_{j3}	κ_{j0}	κ_{j1}	κ_{j2}
ZI-MLN	**0.096**	**1.084**	**0.570**	**0.359**	**0.214**	**0.183**	**0.335**
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	-	-	-

MetagenomeSeq and edgeR require to select 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. Supp. Figs 4(d)-(f) and 6 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. 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-MNL without Λ by comparing $\hat{y}_{ij}^{\text{pred}}$ under the model to the observed counts. As shown in Fig 4(d), ZI-MLN without Λ provides poor fit to the data. Their posterior mean estimates of σ^2 and u_s^2 are greatly inflated compared to their true values. 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 Supp. Fig 7.

Additional Simulations We conducted additional simulation studies, Simulations 3 and 4, 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



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

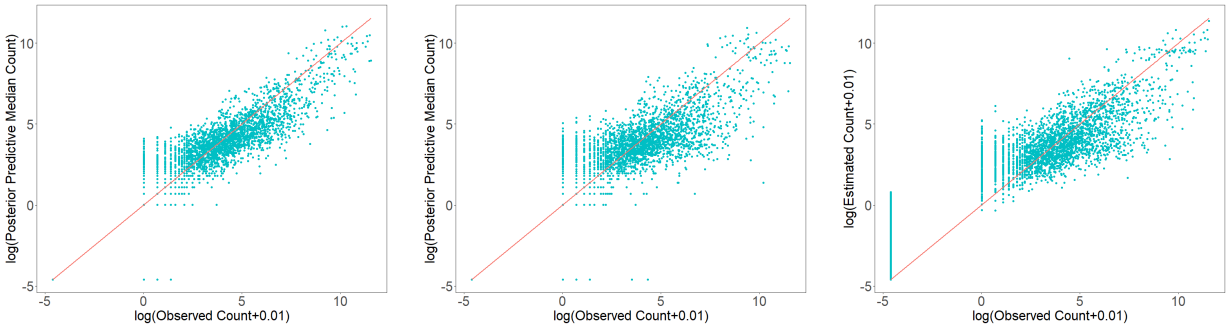
Figure 5: [Skin Microbiome Data] Posterior correlation estimates $\hat{\rho}_{jj'}$ (lower left triangle) and empirical correlation estimates $\rho^{\text{em}}_{jj'}$ (upper right triangle) are shown in panel (a). Panel (b) have the OTUs having $|\hat{\rho}_{j,j'}| \geq 0.40$ for any $j' \neq j$.

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 $\Sigma^{\text{tr}} = \sigma^{2,\text{tr}} \text{I}_J$, 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 as well. In both simulation studies, the results also show that our model compares very favorably to the comparators. More details of Simulations 3 and 4 are in Supp. §3 and §4, 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.

4 Real Data Analyses

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 commu-



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

Figure 6: [Skin Microbiome Data] 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-MNL 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.

nity. Skin swab samples were collected under three conditions, healthy skin, pre-debridement, and post-debridement conditions. The skin microbiome dataset was analyzed by Shuler et al. (2021), which showed changes in the community-level microbial richness and abundance diversity by the experiment 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. For reliable inferences, we removed OTUs that have zero counts in more than 50% of the samples, leaving $J = 187$ OTUs for analysis. Fig 5(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 values were set similar to those of Simulation 1 in § 3.1. 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 2.60 GHz Intel i7 laptop.

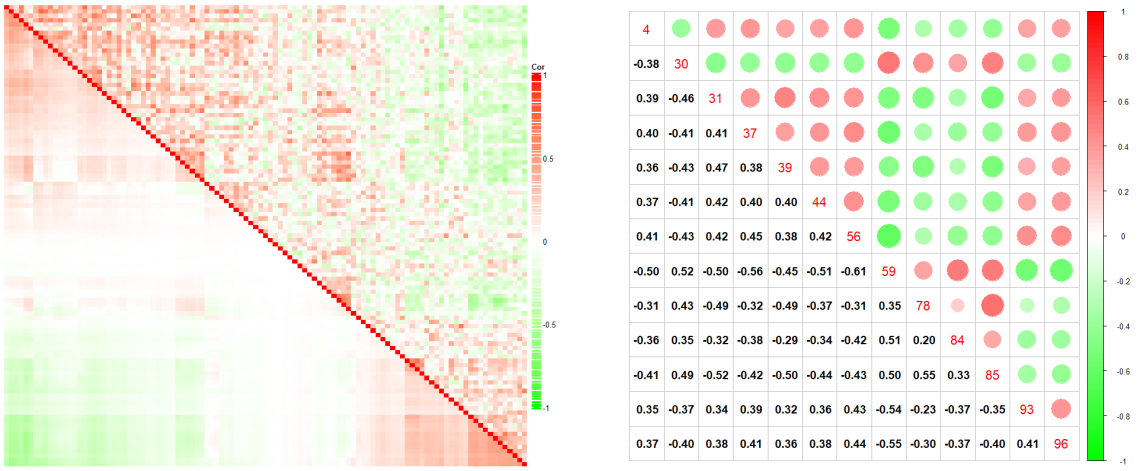
Fig 5(a) illustrates posterior estimates $\hat{\rho}_{jj'}$ of the marginal correlations for all OTUs. Fig 5(b) presents $\hat{\rho}_{jj'}$ for the OTUs that have $|\hat{\rho}_{jj'}| \geq 0.40$ for any $j' \neq j$. Taxonomic information of those selected OTUs is in Supp. Tab 3. From panel (a), correlation 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 a 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. From panel (b), OTUs 43 and 88 belonging to genera *Porphyomonas* and *Peptoniphilus*, respectively, are estimated to be positively correlated with $\hat{\rho} = 0.39$. 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. 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 $\hat{\rho} = 0.42$ belong to genera *Porphyromonas* and *Campylobacter*, respectively, that are both potentially pathogenic. Their positive correlation estimate may reflect a tendency to co-occur, as both are observed in inflammatory bowel disease (Cai et al., 2021). Fig 6(a) has a scatter plot of comparing the posterior predictive median estimates $\hat{y}_{ij}^{\text{pred}}$ to the observed counts. The posterior predictive checking indicates good model fit by ZI-MLN.

We also applied the comparators, ZI-MLN without Λ and metagenomeSeq to the skin microbiome data for comparison. In Fig 6(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. Comparison of those plots to that in panel (a) indicates that our ZI-MLN provides better model fit, possibly because our model accounts for microbial interactions.

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 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 illustration of our statistical model, we used part of their

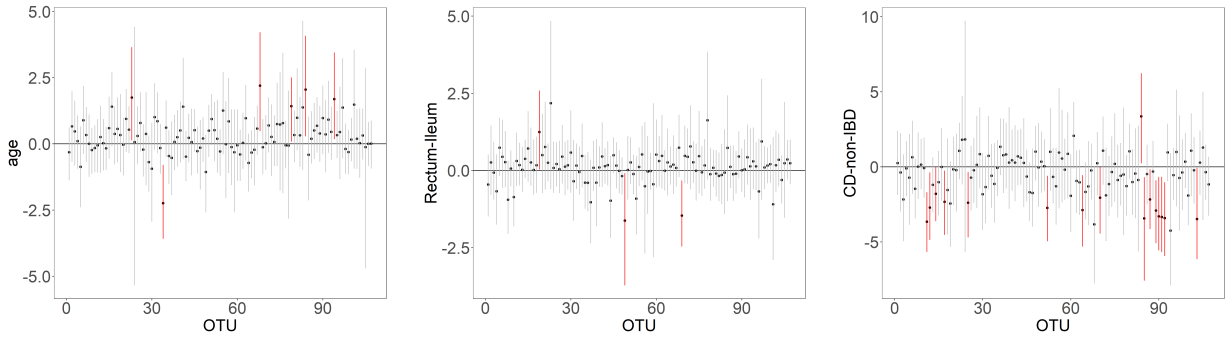


(a) $\hat{\rho}_{jj'}$ versus $\rho^{\text{em}}_{jj'}$ (b) $\hat{\rho}_{jj'}$ for some selected OTUs

Figure 7: [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 $j' \neq j$.

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. Supp. Tab 4 lists all covariates with their supports. For our analysis, 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. We specified hyperparameters similar to those in § 3.2. The MCMC simulation was run over 30,000 iterations, with the first 15,000 iterations discarded as burn-in. It took 0.6 hours for every 5,000 iterations on a 2.60 GHz Intel i7 laptop.

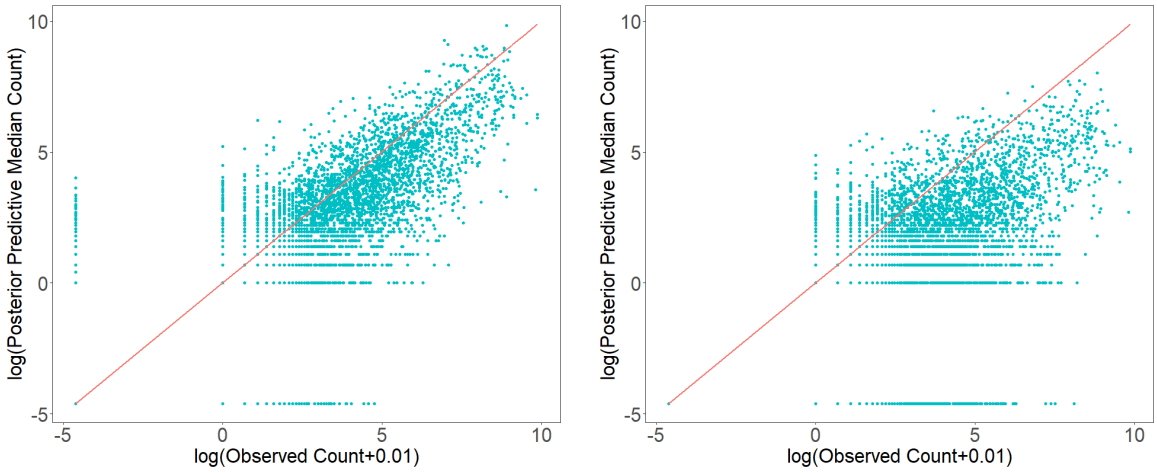
Posterior mean estimates $\hat{\rho}_{jj'}$ of the marginal correlations (lower left triangle) are illustrated with naive empirical correlation estimates $\rho^{\text{em}}_{jj'}$ (upper right triangle) in Fig 7(a). Fig 7(b) reports $\hat{\rho}_{jj'}$ for the OTUs having $|\hat{\rho}_{jj'}| > 0.5$ for any $j' \neq j$. The taxonomic information of the OTU in panel (b) is in Supp. Tab 5. Fig 7(a) shows relatively rich microbial interactions in the gut microbiome



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

Figure 8: [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 black dots, and the 95% credible estimates with vertical lines. The intervals that do not contain zero are in red.

samples as reported in Bashan et al. (2016). In panel (b), OTUs 31, 36, 37, 39, 44, 56, 93 and 96 that are positively correlated with each other, are taxa that are found to indicate dysbiotic microbiota from gastrointestinal diseases. For example, genera *Fusobacterium* and *Parvimonas* that OTUs 36 and 44 belong to, are oral microbes that were found enriched in the gut for patients with colorectal cancer (Kitamoto et al., 2020). OTUs 31 and 39 that belong to family *Erysipelotrichaceae* are observed related to gastrointestinal inflammatory disorders (Kaakoush, 2015). Also, our inference suggests that OTUs 30, 59, 84 and 85 are commensal bacteria. Fig 8 and Supp. Fig 20(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, where back 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 red. 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 16 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, 14 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). We compare posterior predictive median estimates of OTU counts to the observed data in Fig 9(a) to access the



(a) ZI-MLN (b) ZI-MLN without Λ

Figure 9: [Human Gut Microbiome Data]: 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-MNL without Λ, respectively.

model fit. The figure shows that the model fits the data well.

We also applied the comparators, ZI-MLN without Λ, metagenomeSeq and edgeR to the gut microbiome data. Supp. Figs 20(c)-(d) and 21 illustrate posterior estimates of regression coefficients β_{jp} and κ_{jp} from 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 examined posterior predictive distributions of OTU counts (shown in Fig 9(b)). Compared to the fit under ZI-MLN, ZI-MLN without Λ yields poor fit especially for large counts. Supp. Fig 22 compares mean abundance estimates under edgeR and metagenomeSeq to the observed counts, and indicates poor model fit under those models.

5 Conclusion

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 inference on microbial interaction with a small sample size. It offers a straightforward interpretation of microbial dependence structure. The

model also 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' dependency structure and some popular alternatives.

ZI-MLN can be further extended to accommodate more complex data structures. Specifically, Lloyd-Price et al. (2019) collected multi-omics data to obtain comprehensive understanding of the IBD microbial ecosystem. Multi-omic measurements from the same subject may be interrelated, and a 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 structure in multivariate data and can be extended to accommodate multiple coupled observation matrices, e.g., a group factor model in 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.

Acknowledgements

This work was supported by NIH: DP2 GM123457-01 to IAC (Irene Chen) and NSF grant DMS-1662427 (Juhee Lee).

Supplementary File

There are three sections in the Supplementary Materials. In Section § 1 we establish details of posterior computation. Then second Section § 2 contains additional results from some extra simulation studies. Section § 3 contains more supporting plots of the skin data and human gut data.

References

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.

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.

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.

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.

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.

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.

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

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.

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.

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

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.

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

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.

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.

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.

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.

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.

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.

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.

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.

Shuler, K., Verbanic, S., Chen, I. A., and Lee, J. (2021). 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)*.

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.

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.

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, 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*.

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

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.

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.