A negative binomial latent factor model for paired microbiome sequencing data

Hyotae Kim, Nazema Siddiqui, Lisa Karstens, and Li Ma [*]

Abstract

Motivation: Microbiome compositional data are often collected from several body sites and exhibit dependency among them. Analyzing microbial compositions from different sites jointly allows for effective borrowing of information by exploiting the underlying cross-site correlation, which can lead to more effective statistical analysis, especially when the sample size at one or both sites is limited. To this end, we introduce a joint model for microbiome compositions at two (or more) sites within the same subjects. Our model incorporates (i) latent factors shared across two body sites to explain the common subject effects and to serve as the source of correlation between the two sites; and (ii) mixtures of latent factors to allow heterogeneity among the subjects in their level of cross-site association. The model is illustrated with synthetic data and we apply it in a case study involving samples of the urinary and vaginal microbiome collected from women. **Results:** Simulation studies show

[*] Hyotae Kim (hyotae.kim@duke.edu) is Postdoctoral Associate, Department of Biostatistics & Bioinformatics, Duke University, Nazema Siddiqui (nazema.siddiqui@duke.edu) is Associate Professor, Department of Obstetrics and Gynecology, Duke University, Lisa Karstens(karstens@ohsu.edu) is Associate Professor, Department of Medical Informatics and Clinical Epidemiology and Department of Obstetrics and Gynecology, Oregon Health & Science University, and Li Ma (li.ma@duke.edu) is Professor, Department of Statistical Science and Department of Biostatistics and Bioinformatics, Duke University.

how common subject effects influence regression analysis results; a stronger association between two sites in the data causes a greater degree of bias in the analysis. The model with latent factors mitigates the bias present in the model without latent factors, whereas the two models perform comparably for the data set without paired associations. In a case study involving samples collected from a study on the female urogenital microbiome with aging, our model leads to the detection of covariate associations of the vaginal and urinary microbiome composition that are otherwise not statistically significant under a similar regression model applied to the two sites separately. Our model also enables prediction of the microbial abundance at one site based on observations from another, improving predictive power. We consider a model extension that allows the clustering of subjects (samples) and cluster-specific levels of paired association. Under the extended modeling framework, the clusters can be classified according to their association strengths.

KEY WORDS: Bayesian modeling, latent factor models; paired microbiome data; Pólya-Gamma augmentation; compositional data.

1 Introduction

The advent of next-generation sequencing (NGS) technology enables the identification of a wide range of microbes from environmental samples without the need for cultivation, thus facilitating the exploration of microbial communities. The two most widely used methods for sequencing microbial communities are universal marker gene amplicon sequencing, such as the 16S rRNA gene, and whole metagenome shotgun sequencing (WMS) of all microbial genomes. These methods produce sequencing reads, which are subsequently mapped to taxa at various taxonomic levels using bioinformatic preprocessing pipelines, such as DADA2 [Callahan et al., 2016] and MetaPhlAn [Blanco-Míguez et al., 2023]. As a result of this taxonomic profiling, a large, highly sparse count table of taxa per sample is produced, with typically a finer taxonomic resolution in WMS than in 16S amplicon sequencing.

To gain a comprehensive understanding of the human microbiome, researchers often collect data from multiple body sites for each individual and compare the composition and functions of the microbial communities in different parts of the body (e.g., [HMPC, 2012a, HMPC, 2012b]). The UMICRO data set, which we will use as a case study, includes vaginal-urine paired samples obtained by vaginal swabbing of the distal vagina and urine collection by transurethral catheterization, with the aim of identifying the microbial composition of the two communities and analyzing their variations with respect to the menopausal status of the participants. As the sample pairs were collected from two different body sites of the same subject, common effects associated with the subject are likely to be present in both vaginal and urine samples. This data set motivates the development of a model to capture the potential associations between the vaginal and urine niches.

For the analysis of microbial compositional data in the form of a count table with rows for samples and columns for taxa, modeling methods originally developed for RNA sequencing (RNA-seq) or single-cell RNA sequencing (scRNA-seq) data can be adopted. Similarly to microbial compositional data, RNA sequencing data provide sparse count tables that report the number of sequence fragments assigned to each gene per sample or cell, for which the following models have been devised: negative binomial models [Robinson et al., 2010, Love et al., 2014], zero-inflated negative binomial models [Risso et al., 2018], Poisson zero-inflated log-normal models [Wang et al., 2018], and truncated Gaussian hurdle model [Finak et al., 2015]. In addition, [Martin et al., 2020, Morton et al., 2019, Paulson et al., 201 Sohn et al., 2015] introduced beta-binomial, multinomial, and zero-inflated Gaussian models, which were specifically developed for microbiome data. Although zero-inflated models were created to accommodate the pronounced sparsity of microbiome data, their validity in count-based models for sequencing reads remains controversial. As discussed in [Silverman et al., 2020, Sarkar and Stephens, 2021], in many common settings involving sparse sequencing count data, the abundance of zeros can often be adequately accommodated by simply incorporating overdispersion into count-based sampling models, such as in negative binomial models, without needing an additional zero-inflation component. We share this viewpoint, and because we are considering overdispersed count-based models in this paper, we do not by default incorporate an additional zero-inflation component, but in cases where such a component is indeed justified, incorporating it into the model is straightforward.

We propose a model-based approach to jointly account for microbiome compositions at two (or more) sites, adopting a negative binomial regression model for each site while incorporating a shared latent factor to parsimoniously capture potential correlation in paired samples from different body sites in some taxa. Specifically, two negative binomial distributions, one for each site, share a set of latent factors, which are interpreted as unobserved common effects that contribute to their correlation. Under suitable choices of priors on the model parameters, posterior samples can be drawn efficiently using a Gibbs sampler that employs a data-augmentation technique called Pólya-Gamma augmentation. We extend our model on the latent factors to accommodate the common assumption that the samples are not homogeneous but rather form subgroups or clusters, each with their own cross-site correlation patterns. We present versions of the model that work both when the subgrouping is observed and when it is not. In the latter case the subgrouping is inferred from the data. We carry out a case study on the UMICRO data in which the study subjects come from three subgroups by study design.

The rest of the paper is organized as follows. Section 2 introduces our negative binomial latent variable model with Section 2.1 for the base modeling framework and Section 2.2

for some model extensions. In Section 2.3, we present a hierarchical representation of our model with prior specifications for the model parameters, followed by a posterior prediction approach. The proposed model is illustrated through synthetic data in Section 3.1 and the UMICRO study in Section 3.2. Finally, Section 4 concludes.

2 Methods

2.1 Joint Negative Binomial Model

For a given taxon (e.g., genus), we use y_{si} to denote the observed count in sample i at body site s, with $i = 1, \dots, n$ and $s = 1, 2$. Note that the data are actually also indexed by taxa, but for simplicity, throughout our description of the model and the computational recipes, we suppress the index for taxa, as our model is taxon-specific and is applied to each of the taxa separately. Let $\mathrm{NB}(\mu, \alpha)$ be a negative binomial distribution with mean μ and variance $(1 + \mu\alpha)\mu$. We consider the following joint negative binomial model (JNBM) that relates the counts from two sampling sites through a taxon-specific latent factor, γ_i, in the form of a multiplicative factor on the mean count for the taxon,

$$\begin{aligned} y_{1i} &\overset{ind.}{\sim} \mathrm{NB}(\mu_{1i}, \alpha_1), \quad \mu_{1i} = \exp(\gamma_i) N_{1i} \exp(X_i' \beta_1) \\ y_{2i} &\overset{ind.}{\sim} \mathrm{NB}(\mu_{2i}, \alpha_2), \quad \mu_{2i} = \exp(\gamma_i) N_{2i} \exp(X_i' \beta_2) \\ \gamma_i &\overset{ind.}{\sim} \mathrm{N}(-\phi^2/2, \phi^2), \end{aligned} \tag{1}$$

where X_i denotes the vector of covariates for sample i, and β_s is the regression coefficient vector for body site s. α_s is the overdispersion parameter of the negative binomial distribution. N_{si} indicates the total number of read counts, that is, the sum of y_{si} over all taxa of interest. The latent factor γ_i $\big($ or its exponential $\exp(\gamma_i)\big)$ is an unobserved variable that represents a site-invariant sample-specific effect. The multiplicative effect of $\exp(\gamma_i)$

on the mean, μ_{si}, has $\mathrm{E}(\exp(\gamma_i)) = 1$ and $\mathrm{Var}(\exp(\gamma_i)) = \exp(\phi^2) - 1$. This distributional assumption with the mean of 1 enables the random effects of $\exp(\gamma_i)$ to have no preference for a positive impact (> 1) or negative impact (< 1). The parameter ϕ^2 measures the strength of the association between the two body sites; when ϕ^2 becomes zero, the model is reduced to two independent negative binomial distributions for each body site, that is, $\mathrm{NB}(N_{si} \exp(X_i' \beta_s), \alpha_s)$ for $s = 1, 2$. We will refer to it as the separate negative binomial model (SNBM). Section 3.2 will compare our joint two-site model with SNBM in a case study on the UMICRO data.

To perform JNBM-based Bayesian inference, we need priors for $(\alpha_s, \beta_s, \phi^2)$; a description of the complete Bayesian hierarchical model can be found in Section 2.3.1. A fully conjugate sampling recipe for posterior inference based on the Pólya-Gamma data augmentation technique is detailed in Supplementary Material B.

For brevity in the model description, we assume balanced paired data, i.e., both sites have the same number of samples. However, the model can also be used for unbalanced data where some samples are available only for one site.

2.2 Model Extension

Although JNBM assumes a constant level of cross-site association throughout all paired samples, in real-world examples, this assumption is often unrealistic. However, allowing each sample to have its own level of association may lead to an overly flexible model. We thus compromise and assume that there are subgroups (either observed or unobserved) among the samples for which the extent of cross-site association is comparable. Consequently, we extend the base joint model by allowing different sample groups to exhibit different levels of paired association, which can easily be achieved by replacing the normal distribution for the latent factors with a mixture distribution. The following is a two-component normal mixture-based model, which randomly assigns paired samples to two groups:

$$\gamma_i \overset{ind.}{\sim} \nu \mathrm{N}(-\phi_1^2/2, \phi_1^2) + (1-\nu)\mathrm{N}(-\phi_2^2/2, \phi_2^2), \qquad \phi_1^2 < \phi_2^2. \tag{2}$$

It should be noted that we enforce an inequality constraint on the mixing parameter, ϕ_1^2 and ϕ_2^2, to ensure model identifiability. The general form of the mixture prior is given by $\sum_{l=1}^{L} \nu_l \mathrm{N}(-\phi_l^2/2, \phi_l^2)$ with $0 < \phi_1^2 < \ldots < \phi_L^2$ and $\sum_{l=1}^{L} \nu_l = 1$ for $L > 1$. In the paper, we will call this joint negative binomial model with the two-component mixture distribution JNBM_Mix. Supplementary Material A.4 presents JNBM_Mix's superior predictive performance over JNBM, along with random clustering results for Lactobacillus samples as an example.

The clustering of paired samples can also be accomplished by using a given variable $g(i) = \{1, \ldots, L\}$ rather than random assignment with $\{\nu_l\}$, where differences in association levels across L groups may serve as an indicator of the proximity among them. The mixture distribution for γ_i is adjusted as follows:

$$\gamma_i \overset{ind.}{\sim} \sum_{l=1}^{L} \delta_l\Big(g(i)\Big)\mathrm{N}(-\phi_l^2/2, \phi_l^2) = \mathrm{N}(-\phi_{g(i)}^2/2, \phi_{g(i)}^2), \tag{3}$$

where $\delta_l(x)$ is a delta function with $\delta_l(x) = 1$ for $x = l$ and 0 otherwise. For the UMICRO case study, we use Study Group variable for clustering, such that $g(i) = \{$Postmenopausal with no estrogen, Postmenopausal on estrogen, Premenopausal$\}$, which model is named JNBM_SG. Figure S11 shows the posterior distributions of $\phi_{g(i)}^2$ for each study group under the model, revealing obvious similarities between the two postmenopausal groups and differences in location or/and dispersion between postmenopausal and premenopausal for *Lactobacillus*, *Aerococcus*, *Prevotella* and *Corynebacterium*. JNBM_SG also outperforms SNBM and JNBM in prediction, which will be discussed in Section 3.2 and

Supplementary Material A.2.

2.3 Posterior inference

2.3.1 Hierarchical model representation

Below is a full description of the joint negative binomial model (JNBM) presented in (1) of Section 2.1. Let $\boldsymbol{y} = \{y_{si} : s = 1, 2 \text{ and } i = 1, \ldots, n\}$, $\boldsymbol{\gamma} = \{\gamma_i : i = 1, \ldots, n\}$, $\beta_s = \{\beta_{sp} : p = 1, \ldots, P\}$, and $\boldsymbol{\tau^2} = \{\tau_p^2 : p = 1, \ldots, P\}$. The model is defined as

$$\begin{aligned}
p(\boldsymbol{y}|\alpha_s, \beta_s, \boldsymbol{\gamma}) &= \prod_{s=1}^{2}\prod_{i=1}^{n}\Big[\frac{\Gamma(y_{si}+\alpha_s^{-1})}{\Gamma(y_{si}+1)\Gamma(\alpha_s^{-1})}\Big(\frac{\alpha_s^{-1}}{\alpha_s^{-1}+\mu_{si}}\Big)^{\alpha_s^{-1}} \\
&\quad \times \Big(\frac{\mu_{si}}{\alpha_s^{-1}+\mu_{si}}\Big)^{y_{si}}\Big] \\
\gamma_i|\phi^2 &\overset{ind.}{\sim} \mathrm{N}(-\phi^2/2, \phi^2) \\
p(\alpha_s, \phi^2) &= \mathrm{Exp}(\alpha_s|a_{\alpha_s})\mathrm{Exp}(\phi^2|a_{\phi^2}) \\
p(\beta_s|\boldsymbol{\tau^2}) &= \prod_{p=1}^{P}\mathrm{N}(\beta_{sp}|-\tau_p^2/2, \tau_p^2) \\
\tau_p^2 &\overset{ind.}{\sim} \mathrm{Exp}(a_{\tau_p^2}),
\end{aligned}$$

where $\mu_{si} = \exp\big(\gamma_i + \log(N_{si}) + X_i'\beta_s\big)$. For simplicity, we place exponential priors on $(\alpha_s, \phi^2, \tau_p^2)$; their rate parameters were chosen empirically, which are conservative choices and lead to substantial prior-to-posterior learning. For example, in Section 3.1, we set $(a_{\tau_1^2}, a_{\tau_2^2}, a_{\tau_3^2}) = (1, 10, 1)$ for a simulation study with true regression coefficients of $(\beta_{11}, \beta_{12}, \beta_{13}) = (0.05, 0.001, 0.03)$ and $(\beta_{21}, \beta_{22}, \beta_{23}) = (0.03, 0.002, 0.01)$. With the prior means of $(\mathrm{E}(\tau_1^2), \mathrm{E}(\tau_2^2), \mathrm{E}(\tau_3^2)) = (1, 0.1, 1)$, the 95% prior uncertainty bands for β_{s1}, β_{s2}, and β_{s3} are given by $(-2.5, 1.5)$, $(-0.7, 0.6)$, and $(-2.5, 1.5)$, respectively. The conservative choices result in much tighter 95% posterior uncertainty bands with substantial shifts of the posterior means (from the prior means) toward the true values.

Likewise, we chose conservative hyperparameters of $(a_{\alpha_s}, a_{\phi^2}) = (0.1, 0.1)$. For the case study in Section 3.2, we set the hyperparameters to $(a_{\alpha_s}, a_{\phi^2}) = (0.5, 0.1)$, as well as $(a_{\tau_1^2}, \ldots, a_{\tau_8^2}) = (0.1, 10, 10, 1, 1, 1, 1, 1)$ for coefficients of the intercept, two continuous covariates, and five binary covariates. We observed that this prior specification gives rise to such prior-to-posterior learning, although the true underlying parameter values are unknown in the case study. As in γ_i, the regression coefficients β_{sp} are assigned normal distribution priors with mean $-\tau_p^2/2$ and variance τ_p^2, allowing the multiplicative effects of $\exp(\beta_{sp})$ on μ_{si} to have their means equal to 1 with their variances of $\exp(\tau_p^2) - 1$. The following section discusses the augmented likelihood under the negative binomial modeling framework, which enables the regression coefficients and the latent factors to have full conditionals in closed form.

Similarly, we can represent the extended models – JNBM_Mix and JNBM_SG – with variations in distributional assumptions on γ_i as follows,

[**JNBM_Mix**] : with auxiliary variables $\{\xi_i\}$ for a hierarchical representation of the mixture distribution on γ_i,

$$\begin{aligned} \gamma_i|\phi_1^2, \ldots, \phi_L^2, \xi_i &\overset{ind.}{\sim} \mathrm{N}(-\phi_{\xi_i}^2/2, \phi_{\xi_i}^2), \quad i = 1, \ldots, n \\ p(\xi_i = l|\nu_l) &= \nu_l, \quad l = 1 \ldots, L \\ (\nu_1, \ldots, \nu_L) &\overset{ind.}{\sim} \mathrm{Dir}(p_{\nu_1}, \ldots, p_{\nu_L}) \\ \phi_l^2 &\overset{ind.}{\sim} \mathrm{Exp}(\phi_l^2|a_{\phi_l^2}), \end{aligned} \tag{4}$$

where $\mathrm{Dir}(p_1, \ldots, p_L)$ denotes a Dirichlet distribution with $p_l = 1/L$.

[**JNBM_SG**] :

$$\begin{aligned} \gamma_i|\phi_1^2, \ldots, \phi_L^2 &\overset{ind.}{\sim} \mathrm{N}(-\phi_{g(i)}^2/2, \phi_{g(i)}^2), \quad i = 1, \ldots, n \\ \phi_l^2 &\overset{ind.}{\sim} \mathrm{Exp}(\phi_l^2|a_{\phi_l^2}), \end{aligned}$$

where $g(i)$ is an observed variable, e.g., the study group clinical variable of the UMICRO data in Section 3.2. The hyperparameter $a_{\phi_l^2}$ of the exponential prior for ϕ_l^2 is chosen in the same manner as a_{ϕ^2} of ϕ^2 in JNBM.

2.3.2 Cross-site prediction

In making predictive inferences about unknown observables y^*, we consider the posterior predictive distribution below, from which we draw predictive samples and compute their average. The posterior predictive distribution is

$$p(y^*|y) = \int p(y^*|\boldsymbol{\theta})p(\boldsymbol{\theta}|y)d\boldsymbol{\theta}, \tag{5}$$

where y denotes the observed data and $\boldsymbol{\theta}$ is a vector of model parameters: (α, β) for SNBM and (α, β, γ) for JNBM. Posterior predictive samples for y^* are drawn from the negative binomial sampling distribution $p(y^*|\boldsymbol{\theta})$ with parameters that are substituted with the posterior samples of $\boldsymbol{\theta}$ taken from $p(\boldsymbol{\theta}|y)$. Then, the mean of the posterior predictive samples becomes the predictive value for y^*.

One useful feature of our joint model is the ability to predict counts at one site using observations from the other site. Suppose, for example, that $y_{1i'}$ for $i' \in I \subset A \equiv \{1, \ldots, n\}$ are unobserved and prediction targets. In SNBM, posterior sampling of model parameters (α_1, β_1) is based only on observations $\{y_{1i} : i \in A - I\}$, while, in JNBM, not only $\{y_{1i} : i \in A - I\}$ but also $y_{2i'}$, $i' \in I$, are used for $(\alpha_1, \beta_1, \gamma_{i'})$ posterior sampling, specifically for latent factors $\gamma_{i'}$. By plugging the posterior samples of the model parameters into the sampling distribution $\text{NB}\Big(\exp(\gamma_{i'} + \log(N_{1i'}) + X_{i'}'\beta_1), \alpha_1\Big)$, with $\gamma_{i'} = 0$ for SNBM, we can draw posterior predictive samples for $y_{1i'}$. Let $\tilde{y}_{1i'}^{(b)}$ be a posterior predictive sample from $\text{NB}\Big(\exp(\gamma_{i'}^{(b)} + \log(N_{1i'}) + X_{i'}'\beta_1^{(b)}), \alpha_1^{(b)}\Big)$ with $(\alpha_1^{(b)}, \beta_1^{(b)}, \gamma_i^{(b)})$ acquired at b-th MCMC iteration for $b = 1, \ldots, B$. Then, the predictive value $\tilde{y}_{1i'}$ for $y_{1i'}$ is defined as the posterior mean, that is, $\tilde{y}_{1i'} = \sum_{b=1}^{B} \tilde{y}_{1i'}^{(b)}/B$. Section 3.2 discusses the predictive performance of the models, which is based on the leave-one-out cross-validation (LOOCV) approach; we select one observation from the $2n$ observations across n pairs of samples as a test set (i.e., an unknown observable for prediction), fit the models to the remaining data, and predict the test set value. We repeated the LOOCV $2n$ times to obtain the predictive values for all sample pairs of a given taxon.

3 Results

3.1 Simulation

We simulate paired samples with varying levels of association, comparing separate and joint negative binomial models to empirically grasp the role of latent factors in the joint model. Synthetic data were generated by drawing $n = 300$ pairs of samples from negative binomial distributions with latent factors defined as,

$$\begin{aligned} y_{1i} &\overset{ind.}{\sim} \mathrm{NB}(\exp(\gamma_i + \log(N_{1i}) + X_i'\beta_1), \alpha_1); \\ y_{2i} &\overset{ind.}{\sim} \mathrm{NB}(\exp(\gamma_i + \log(N_{2i}) + X_i'\beta_2), \alpha_2); \\ \gamma_i &\overset{i.i.d.}{\sim} \mathrm{N}(-\phi^2/2, \phi^2). \end{aligned}$$

The dispersion parameters and regression coefficients are arbitrarily chosen to be $\alpha_1 = 2$, $\alpha_2 = 1$, $\beta_1 = (\beta_{11}, \beta_{12}, \beta_{13}) = (0.05, 0.001, 0.03)$, and $\beta_2 = (\beta_{21}, \beta_{22}, \beta_{23}) = (0.03, 0.002, 0.01)$, where $X_i = (x_{i1}, x_{i2}, x_{i3})'$ is a vector consisting of an intercept x_{i1}, a continuous covariate $x_{i2} \sim \mathrm{Unif}(20, 80)$ and a binary covariate $x_{i3} \sim \mathrm{Bern}(0.3)$. $\mathrm{Unif}(a, b)$ and $\mathrm{Bern}(p)$ are (continuous) uniform and Bernoulli distributions with means $(a+b)/2$ and p, respectively. Again, the dispersion parameter ϕ^2 of the latent factors in the underlying distribution determines how strong the association between pairs of samples is, and it is set to 0 for

Scenario	Model	$\overline{e^2}_{12}(\sigma_{\overline{e^2}_{12}}) \times 10^5$	$\overline{e^2}_{13}(\sigma_{\overline{e^2}_{13}}) \times 10^2$	$\overline{e^2}_{22}(\sigma_{\overline{e^2}_{22}}) \times 10^5$	$\overline{e^2}_{23}(\sigma_{\overline{e^2}_{23}}) \times 10^2$
$\phi^2 = 0$	SNBM	**1.53** (0.14)	2.10 (0.18)	0.80 (0.07)	**1.10** (0.09)
	JNBM	1.54 (0.14)	**2.06** (0.18)	**0.77** (0.07)	1.11 (0.10)
$\phi^2 = 2$	SNBM	21.84 (1.48)	7.45 (0.72)	17.90 (0.90)	4.49 (0.43)
	JNBM	**4.37** (0.33)	**2.15** (0.19)	**3.50** (0.20)	**1.41** (0.11)
$\phi^2 = 10$	SNBM	176.89 (9.41)	44.03 (4.33)	185.50 (6.97)	26.21 (2.34)
	JNBM	**14.79** (0.69)	**2.20** (0.27)	**16.25** (0.67)	**1.44** (0.13)

Table 1: MSE, denoted as $\overline{e^2}_{sp}$ with site $s = 1, 2$ and covariate $p = 2, 3$ for regression coefficients $\{\beta_{sp}\}$, with its standard error $\sigma_{\overline{e^2}_{sp}}$ in parentheses.

independent pairs and to 2 or 10 for dependent pairs, with 10 indicating a stronger association. N_{1i} and N_{2i} are offsets, which in real-world applications are used to normalize for sequencing depth (or the sum of abundances across selected taxa of interest). We set $N_{1i}, N_{2i} \overset{i.i.d.}{\sim} \text{NB}(1000, 1) + 1000$; sampling of N_{si} from the negative binomial distribution results in count data with a right-skewness as with the sequencing depths in the UMICRO data set (but on a smaller scale). By repeatedly sampling with the parameters, covariates, and offsets defined above, we produced 300 sets of 300 sample pairs for the following analysis.

We evaluate the separate and joint models with regard to the accuracy of their estimated regression coefficients. Table 1 shows the mean square error (MSE) between the estimated and true coefficients for the two covariates under each model, demonstrating that JNBM is superior to SNBM for the data with non-zero associations ($\phi^2 \neq 0$). This indicates that ignoring the latent sample effect (γ_i) as in SNBM can severely reduce the efficiency in estimating the underlying relationships between the response variables and predictors. As expected, the difference in estimation accuracy between the two models becomes greater as the paired association in the underlying distribution becomes stronger, and therefore the common effects become larger.

3.2 Case Study: the UMICRO Data

We apply JNBM to the UMICRO data set and compare the resulting inferences to those from SNBM in the estimation of regression coefficients and in the prediction of taxa abundance at one of the sites. Again, data were collected from two different body sites by catheterization for urine samples and vaginal swabbing for vaginal samples. Full-length contigs received from Loop Genomics (Element Biosciences, San Diego, CA), which we refer to as LoopSeq, were processed with DADA2 (v 1.24.0) to generate amplicon sequence variants (ASV), using parameters as recommended for synthetic full-length 16S LoopSeq data in [Callahan et al., 2021]. Taxonomic classifications were assigned using BLCA (v 2.2) and the 16S NCBI database (downloaded on 11/16/2021). The processed data contain 68 sample pairs with 151 common taxa (genera) found in both vaginal and urinary samples, while the following analysis focuses on nine taxa of interest: *Gardnerella, Streptococcus, Lactobacillus, Aerococcus, Anaerococcus, Bifidobacterium, Corynebacterium, Fannyhessea, Prevotella.* We incorporate several clinical and demographic metadata for the subjects as covariates in the models: age, body mass index (BMI), diabetes (yes/no), daily yogurt or probiotic consumption (yes/no), race_ethnicity (0: White, 1: Black, 2: Others), the presence or absence of overactive bladder (OAB) (yes/no). In addition, the study group variable on menopausal status (Postmenopausal with no estrogen; Postmenopausal on estrogen; Premenopausal) is used to cluster the samples for JNBM_SG, introduced in Section 2.2.

Figure 1 shows examples of differences between SNBM and JNBM when it comes to the estimation of regression coefficients. The results indicate that JNBM has relatively pronounced positive effects of BMI and age on *Gardnerella* and *Streptococcus* counts in both vaginal and urinary samples, while these effects are negligible under SNBM. In the previous section, we saw that the joint model provides more accurate estimation results

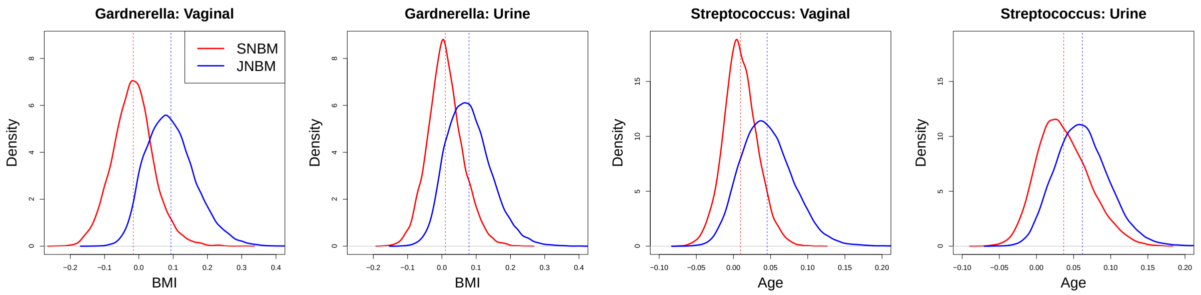

Figure 1: Posterior distributions of regression coefficients for BMI in the vaginal and urine data sets for *Gardnerella* (first two panels); and regression coefficients for Age in the data sets for *Streptococcus* (last two panels). The dashed lines indicate the posterior means.

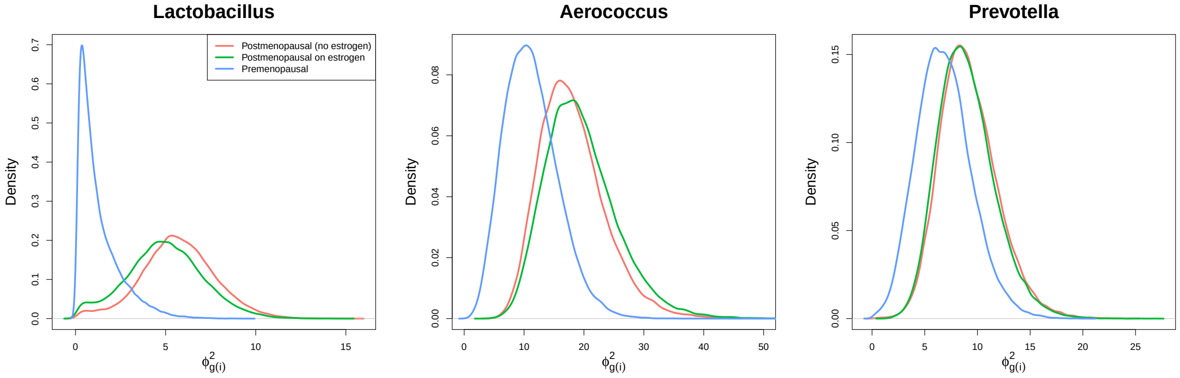

Figure 2: Posterior distributions of $\phi^2_{g(i)}$, the dispersion hyperparameters of latent factors, from JNBM_SG for the three different genera.

when paired associations are present. Here, the posterior means (and standard deviations) of ϕ^2 for *Gardnerella* and *Streptococcus* are given by 17 (4) and 14 (4), respectively. It favors the results of JNBM for the regression coefficients. Furthermore, the positive effects of BMI and age on the taxa under JNBM are consistent with the biological findings reported in [Brookheart et al., 2019, Xu et al., 2020]. More results on regression coefficients can be found in Supplementary Material A.1.

JNBM_SG clusters the samples according to the study group variable, allowing each group to have a different level of paired association. Figure 2 illustrates that, for the three genera, the two postmenopausal groups have similar association strengths but dif-

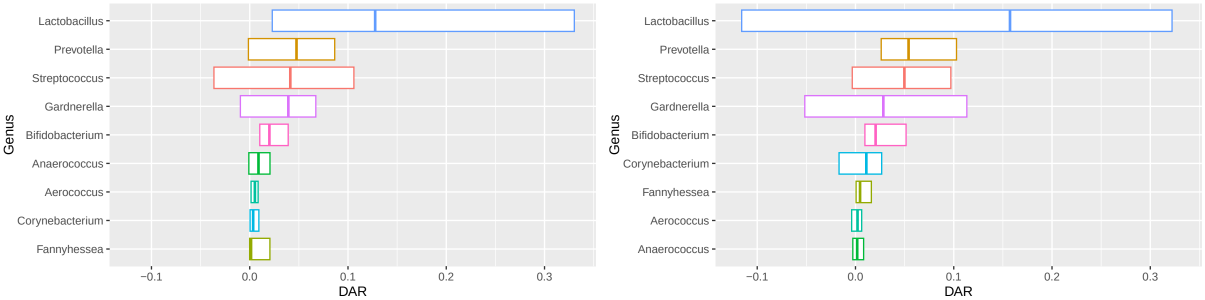

Figure 3: Boxplots (from the 1st to 3rd quartiles) of $\mathrm{DAR}_{sik}^{\text{SNBM-JNBM\_SG}}$, the differences in the absolute residuals between SNBM and JNBM_SG, for each taxon. The genera on the y-axis are sorted by their median DARs in each data set.

fer from the premenopausal group, which has a weaker association. Figure S11 displays the posterior distributions of the parameters for the nine taxa; interestingly, *Gardnerella* and *Fannyhessea* exhibit remarkable differences in strength among the three groups, with postmenopausal with no estrogen showing the strongest association, followed by postmenopausal with estrogen.

To assess the predictive performance of the models for the microbial composition, we calculated the residuals between the observed and predicted relative abundances of each microbial genus, denoted as $r_{sik}^{M} = \frac{y_{sik}}{\sum_{k=1}^{9} y_{sik}} - \frac{\tilde{y}_{sik}^{M}}{\sum_{k=1}^{9} \tilde{y}_{sik}^{M}}$, where y_{sik} is an observed count for the k-th taxon in the i-th sample of the s-th body site and $\tilde{y}_{sik}^{M}$ the corresponding predicted count under model M. Figure 3 presents the difference in the absolute residuals between SNBM and JNBM_SG, that is, $\mathrm{DAR}_{sik}^{\text{SNBM-JNBM\_SG}} = |r_{sik}^{\text{SNBM}}| - |r_{sik}^{\text{JNBM\_SG}}|$. The greater the difference, the better the predictive performance of JNBM_SG. The boxplots represent the distributions of DARs per taxon for each body site, with the positive medians (in most taxa) indicating that JNBM_SG enhances the predictive efficiency of SNBM. The improvement in prediction under the joint model is attributed to the borrowing of information between sites. As such, the base joint model, JNBM, also outperforms SNBM, which has been further improved by accommodating different levels of association for clusters in JNBM_SG (Figure S10) and in JNBM_Mix (Figure S12).

4 Conclusions

We have proposed a joint negative binomial model (JNBM), a latent variable model based on negative binomial distributions, for paired microbiome data. The UMICRO data, gathered from two different body sites for each of the subjects, is a motivating real-world example, and the proposed model incorporates a set of latent factors to capture associations between the two body sites. JNBM has been extended by modifying the distribution for the latent factors, which permits varying levels of paired association across groups of samples.

Our joint negative binomial model provides more accurate regression estimates than the separate negative binomial model without latent factors when there exists a non-zero paired association in the data. In the UMICRO case study, the joint model reveals conspicuous positive effects of BMI and age on *Gardnerella* and *Streptococcus* abundances, respectively, which are in agreement with some previous research findings ([Brookheart et al., 2019, Xu et al., 2020]). Moreover, the joint model outperforms the separate model in prediction, which can be attributed to the use of observations from the opposite body site for prediction via latent factors. With extended models, prediction performance can be further enhanced due to their ability to assign different association strengths to each group. In addition, estimates of the association strengths can also be used to characterize sample groups. For example, we found that the two postmenopausal cohorts (with and without estrogen) had similar associations but were different from the premenopausal cohort, which had a relatively weaker association, for *Lactobacillus*, *Aerococcus*, and *Prevotella* in the UMICRO case study.

Although the case study focuses on nine taxa of interest, we have applied these models to other genera. Unsurprisingly, the models struggled to predict taxa that are prevalent in one site but rare in the other. For example, *Escherichia*, *Klebsiella*, and *Pseudomonas* are pathogens that are common in the urine but scarce in the vagina, showing a high sparsity (high proportion of samples with zero count) in the vaginal data set. It is particularly challenging to predict non-zero counts (urine) using zero counts (vaginal) compared to the reverse scenario. The shared latent factor assumption under our model also becomes questionable when the taxon is rarely present in one of the sites but common in the other.

Acknowledgement

The research was supported in part by NIGMS grant R01-GM135440, NSF grant EEC-2133504, and NIA grant R03-AG060082.

References

[Blanco-Míguez et al., 2023] Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., Manghi, P., Dubois, L., Huang, K. D., Thomas, A. M., et al. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using metaphlan 4. *Nature Biotechnology*, pages 1–12.

[Brookheart et al., 2019] Brookheart, R. T., Lewis, W. G., Peipert, J. F., Lewis, A. L., and Allsworth, J. E. (2019). Association between obesity and bacterial vaginosis as assessed by nugent score. *American journal of obstetrics and gynecology*, 220(5):476–e1.

[Callahan et al., 2021] Callahan, B. J., Grinevich, D., Thakur, S., Balamotis, M. A., and Yehezkel, T. B. (2021). Ultra-accurate microbial amplicon sequencing with synthetic long reads. *Microbiome*, 9(1):130.

[Callahan et al., 2016] Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). Dada2: High-resolution sample inference from illumina amplicon data. *Nature methods*, 13(7):581–583.

[Finak et al., 2015] Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., et al. (2015). Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology*, 16(1):1–13.

[HMPC, 2012a] HMPC (2012a). A framework for human microbiome research. *nature*, 486(7402):215–221.

[HMPC, 2012b] HMPC (2012b). Structure, function and diversity of the healthy human microbiome. *nature*, 486(7402):207–214.

[Love et al., 2014] 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(12):1–21.

[Martin et al., 2020] Martin, B. D., Witten, D., and Willis, A. D. (2020). Modeling microbial abundances and dysbiosis with beta-binomial regression. *The annals of applied statistics*, 14(1):94.

[Morton et al., 2019] Morton, J. T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L. S., Edlund, A., Zengler, K., and Knight, R. (2019). Establishing microbial composition measurement standards with reference frames. *Nature communications*, 10(1):2719.

[Paulson et al., 2013] 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.

[Polson et al., 2013] Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349.

[Risso et al., 2018] Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell rna-seq data. *Nature communications*, 9(1):284.

[Robinson et al., 2010] 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.

[Sarkar and Stephens, 2021] Sarkar, A. and Stephens, M. (2021). Separating measurement and expression models clarifies confusion in single-cell rna sequencing analysis. *Nature genetics*, 53(6):770–777.

[Silverman et al., 2020] Silverman, J. D., Roche, K., Mukherjee, S., and David, L. A. (2020). Naught all zeros in sequence count data are the same. *Computational and structural biotechnology journal*, 18:2789–2798.

[Sohn et al., 2015] Sohn, M. B., Du, R., and An, L. (2015). A robust approach for identifying differentially abundant features in metagenomic samples. *Bioinformatics*, 31(14):2269–2275.

[Wang et al., 2018] Wang, J., Huang, M., Torre, E., Dueck, H., Shaffer, S., Murray, J., Raj, A., Li, M., and Zhang, N. R. (2018). Gene expression distribution deconvolution in single-cell rna sequencing. *Proceedings of the National Academy of Sciences*, 115(28):E6437–E6446.

[Xu et al., 2020] Xu, J., Bian, G., Zheng, M., Lu, G., Chan, W.-Y., Li, W., Yang, K., Chen, Z.-J., and Du, Y. (2020). Fertility factors affect the vaginal microbiome in women of reproductive age. *American Journal of Reproductive Immunology*, 83(4):e13220.

Supplementary Material for “Negative binomial latent variable model for paired data in microbiome analysis”

A. Additional figures for real data analysis

A.1 Posterior distributions of β_s

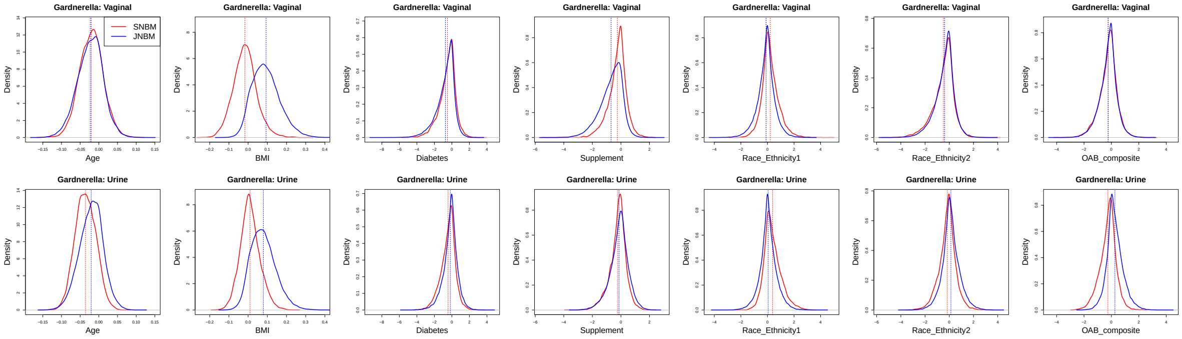

Figure S1: Posterior distributions of regression coefficients in the vaginal (first row) and urine (second row) data sets for *Gardnerella*.

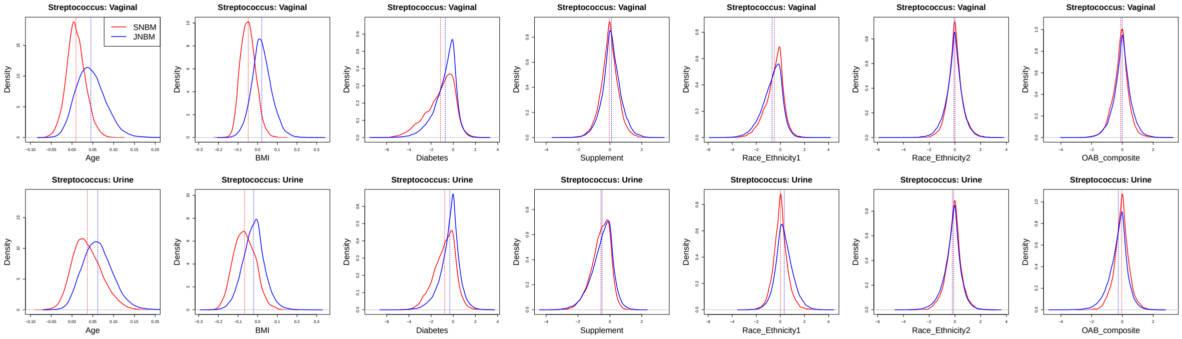

Figure S2: Posterior distributions of regression coefficients in the vaginal (first row) and urine (second row) data sets for *Streptococcus*.

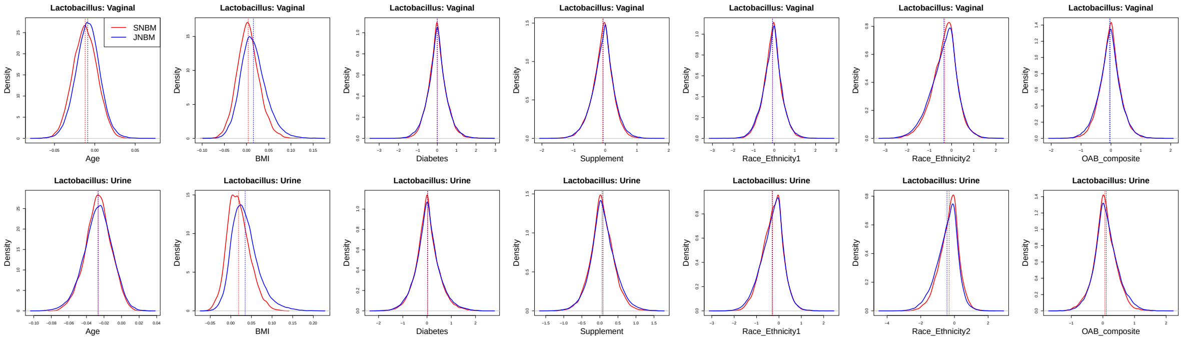

Figure S3: Posterior distributions of regression coefficients in the vaginal (first row) and urine (second row) data sets for *Lactobacillus*.

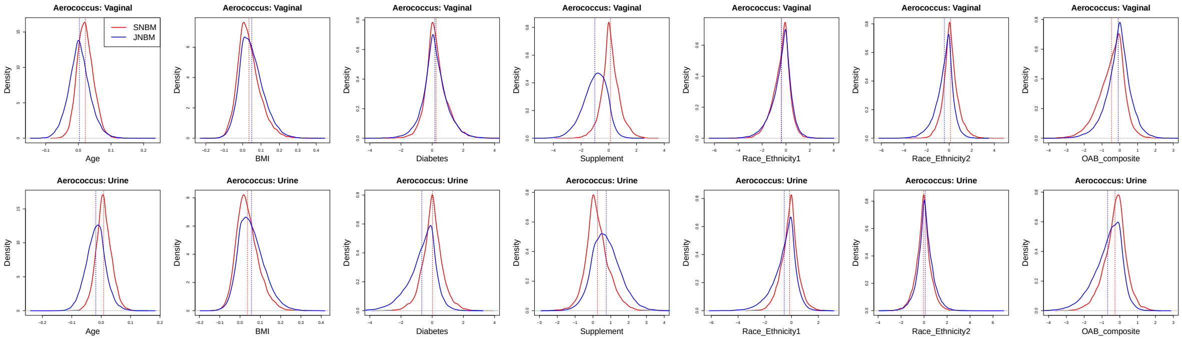

Figure S4: Posterior distributions of regression coefficients in the vaginal (first row) and urine (second row) data sets for *Aerococcus*.

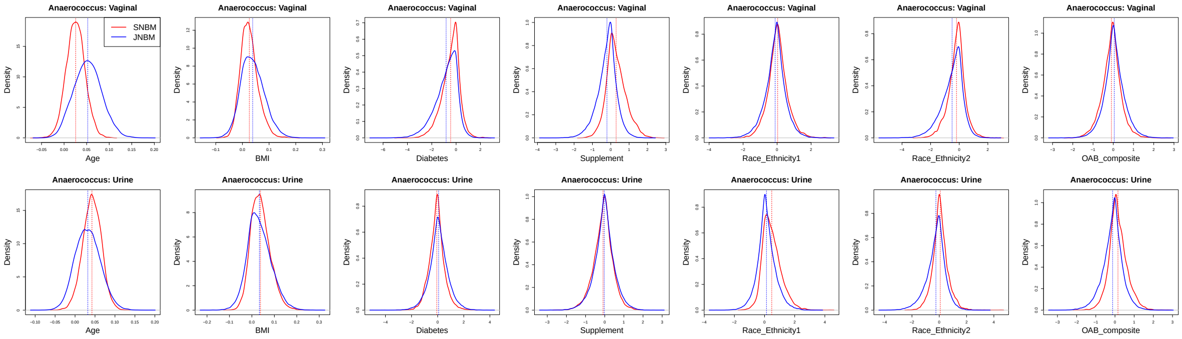

Figure S5: Posterior distributions of regression coefficients in the vaginal (first row) and urine (second row) data sets for *Anaerococcus*.

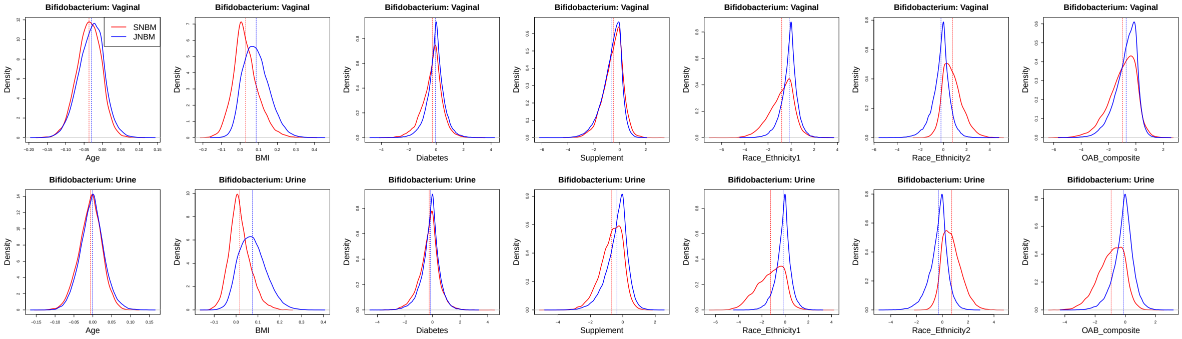

Figure S6: Posterior distributions of regression coefficients in the vaginal (first row) and urine (second row) data sets for *Bifidobacterium*.

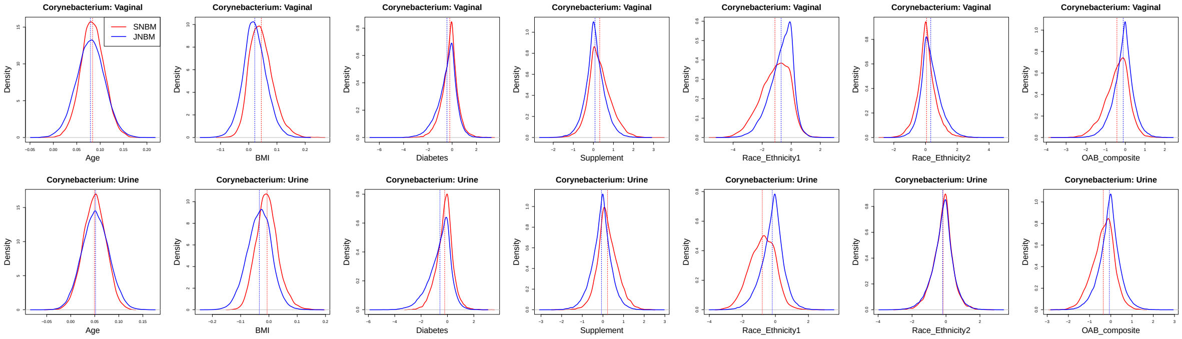

Figure S7: Posterior distributions of regression coefficients in the vaginal (first row) and urine (second row) data sets for *Corynebacterium*.

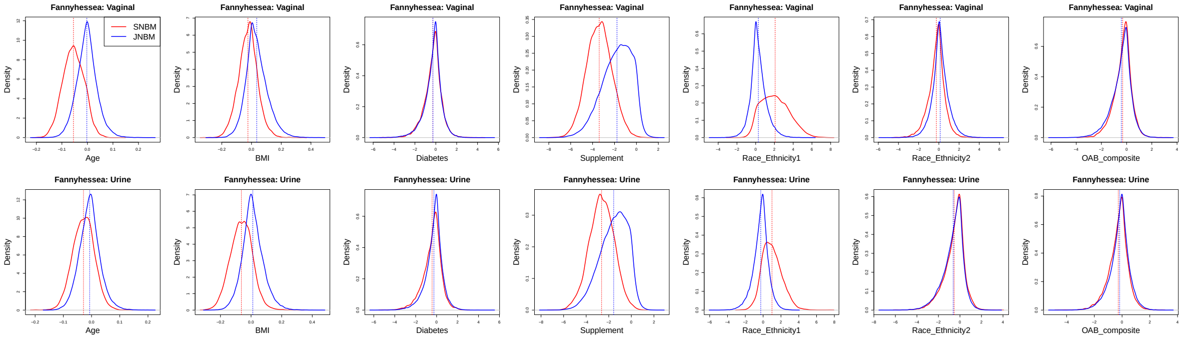

Figure S8: Posterior distributions of regression coefficients in the vaginal (first row) and urine (second row) data sets for *Fannyhessea*.

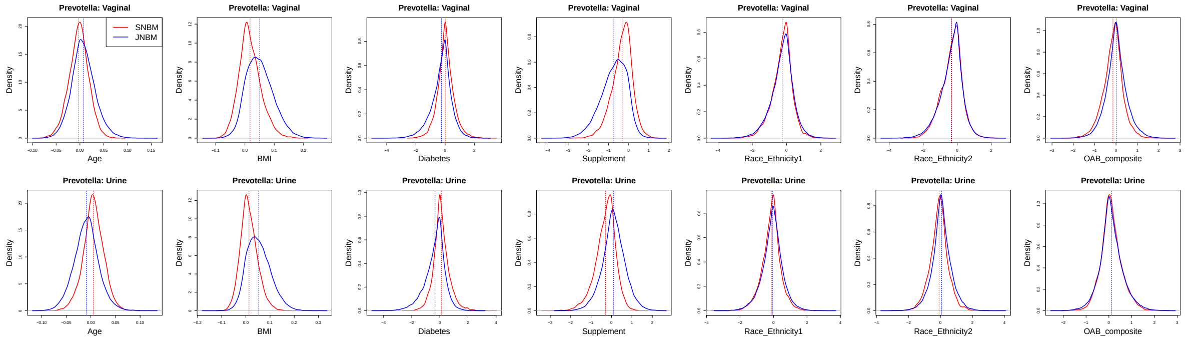

Figure S9: Posterior distributions of regression coefficients in the vaginal (first row) and urine (second row) data sets for *Prevotella*.

A.2 Boxplots of DARs

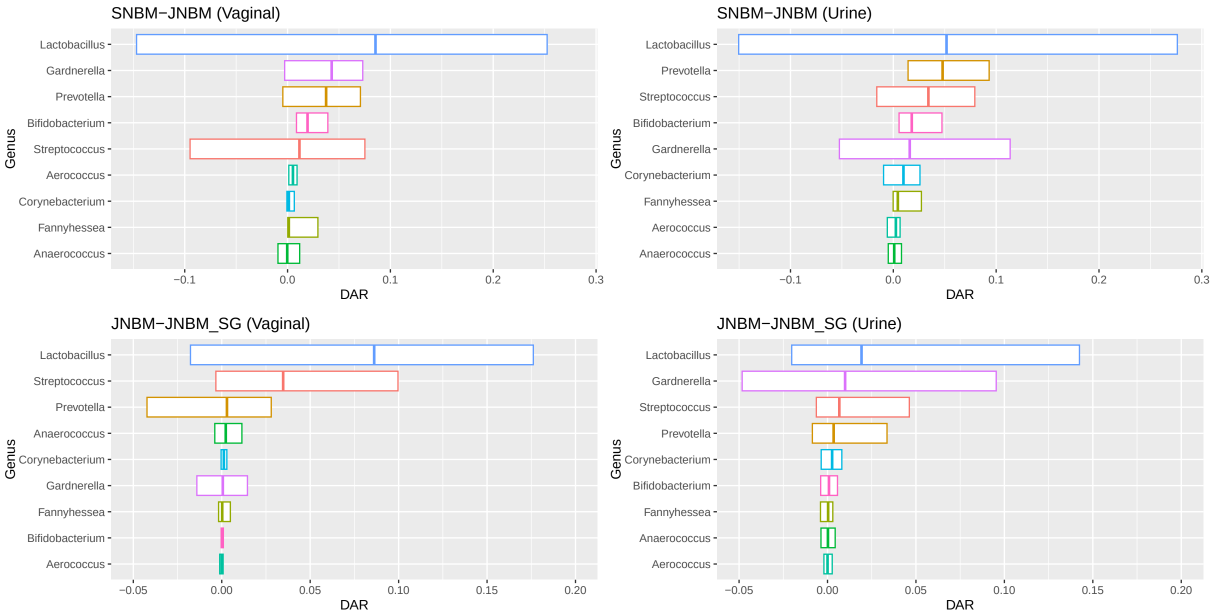

Figure S10: Boxplots of $\text{DAR}_{sik}^{\text{SNBM-JNBM}}$ (top row) and $\text{DAR}_{sik}^{\text{JNBM-JNBM\_SG}}$ (bottom row).

A.3 Posterior distributions of $\phi^2_{g(i)}$ under JNBM_SG

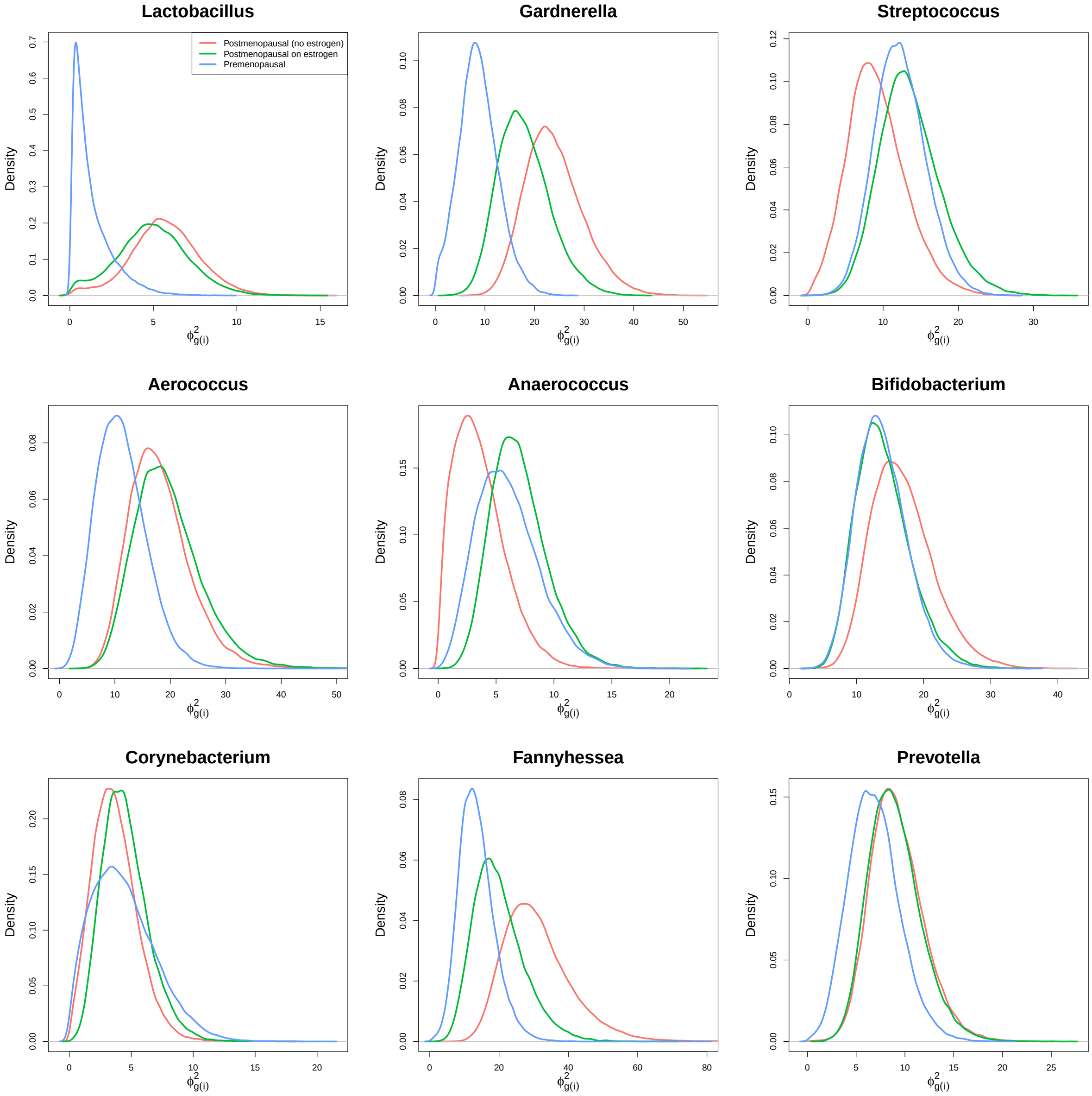

Figure S11: Posterior distributions of $\phi^2_{g(i)}$ colored by Study Group $g(i)$.

A.4 Prediction results of JNBM_Mix

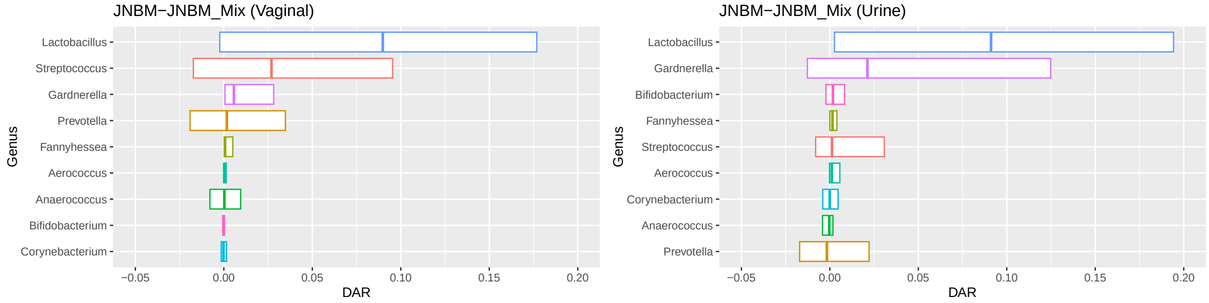

Figure S12: Boxplots of $\text{DAR}_{sik}^{\text{JNBM-JNBM\_Mix}}$.

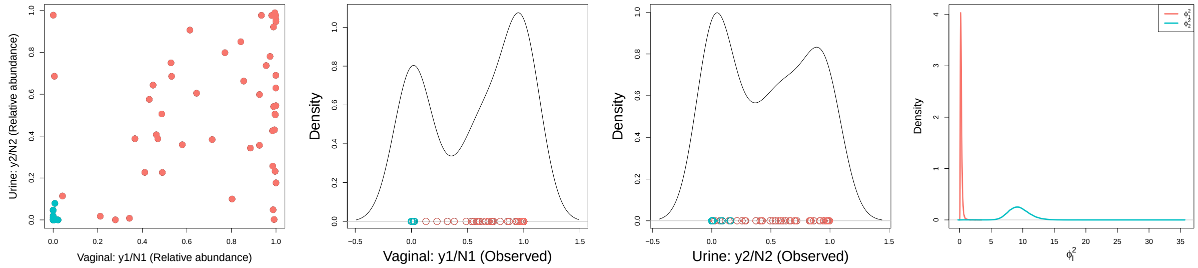

Figure S13: Lactobacillus: On the left is a scatter plot of observed relative abundances colored by posterior estimates of the auxiliary variable ξ_i for the mixture prior; red for $\hat{\xi}_i = 1$ and green for $\hat{\xi}_i = 2$, where $\hat{\xi}_i$ is the posterior median estimate. The next two panels show marginal density estimates of the relative abundances, with observed values at the bottom. The last panel presents the posterior distributions of two dispersion hyperparameters ϕ_1^2 and ϕ_2^2.

B. Computational details on posterior inference

B.1 Pólya-Gamma augmentation for negative binomial models

This section describes the Pólya Gamma augmentation scheme for the proposed negative binomial model. For sample i at body site s, the negative binomial probability mass function is given by

$$\begin{aligned}
p(y_{si}|\alpha_s,\beta_s,\gamma_i) &= \frac{\Gamma(y_{si}+\alpha_s^{-1})}{\Gamma(y_{si}+1)\Gamma(\alpha_s^{-1})}\Big(\frac{\alpha_s^{-1}}{\alpha_s^{-1}+\mu_{si}}\Big)^{\alpha_s^{-1}}\Big(\frac{\mu_{si}}{\alpha_s^{-1}+\mu_{si}}\Big)^{y_{si}} \\
&= \frac{\Gamma(y_{si}+\alpha_s^{-1})}{\Gamma(y_{si}+1)\Gamma(\alpha_s^{-1})}(\alpha_s^{-1})^{\alpha_s^{-1}}\Big[\frac{\mu_{si}^{y_{si}}}{(\alpha_s^{-1}+\mu_{si})^{\alpha_s^{-1}+y_{si}}}\Big] \\
&= \frac{\Gamma(y_{si}+\alpha_s^{-1})}{\Gamma(y_{si}+1)\Gamma(\alpha_s^{-1})}\Big[\frac{(\mu_{si}\alpha_s)^{y_{si}}}{(1+\mu_{si}\alpha_s)^{\alpha_s^{-1}+y_{si}}}\Big],
\end{aligned}$$

where $\mu_{si} = \exp(\gamma_i + \log(N_{si}) + X_i'\beta_s)$ under JNBM. Following [Polson et al., 2013], the last term in the square brackets can be expressed as,

$$\begin{aligned}
\frac{(\mu_{si}\alpha_s)^{y_{si}}}{(1+\mu_{si}\alpha_s)^{\alpha_s^{-1}+y_{si}}} &= \frac{(\exp(Z_{si}))^{y_{si}}}{\Big(1+\exp(Z_{si})\Big)^{\alpha_s^{-1}+y_{si}}} \\
&= 2^{-(\alpha_s^{-1}+y_{si})}\exp\Big(Z_{si}(y_{si}-(\alpha_s^{-1}+y_{si})/2)\Big) \\
&\times \int_0^\infty \exp\Big(-\omega_{si}Z_{si}^2/2\Big)\mathrm{PG}(\omega_{si}|\alpha_s^{-1}+y_{si},0)d\omega_{si},
\end{aligned} \tag{6}$$

where $Z_{si} \equiv \log(\mu_{si}) + \log(\alpha_s)$. $\mathrm{PG}(\omega|a,b)$ denotes the probability density function of the Pólya-Gamma distribution $\mathrm{PG}(a,b)$. Using the (Pólya-Gamma) auxiliary variable ω_{si}, the equation (6) can be represented hierarchically without the integration. Then, the joint probability function for y_{si} and ω_{si} is derived as,

$$\begin{aligned}
p(y_{si},\omega_{si}|\alpha_s,\beta_s,\gamma_i) = &\frac{\Gamma(y_{si}+\alpha_s^{-1})}{\Gamma(y_{si}+1)\Gamma(\alpha_s^{-1})}2^{-(\alpha_s^{-1}+y_{si})}\exp\Big(Z_{si}(y_{si}-(\alpha_s^{-1}+y_{si})/2)\Big) \\
&\times \exp\Big(-\omega_{si}Z_{si}^2/2\Big),
\end{aligned}$$

with $\omega_{si} \overset{ind.}{\sim} \mathrm{PG}(\alpha_s^{-1}+y_{si},0)$ for $i = 1,\ldots,n$. The augmented likelihood with the Pólya-Gamma auxiliary variables ensures (normal) prior conjugacy for (β_s,γ_i), facilitating posterior inference, which will be discussed in the following section.

B.2 Posterior simulation

The Gibbs sampler is used to draw posterior samples of model parameters for inference. The Pólya-Gamma augmentation enables us to obtain the full conditionals in closed form for most of JNBM parameters except $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2\}$, ϕ^2, and $\boldsymbol{\tau^2} = \{\tau_1^2, \dots, \tau_p^2\}$, for which we employ the Metropolis–Hastings algorithm. The following are the full conditionals for the parameters of our joint models.

Let $\boldsymbol{y_{s\cdot}} = \{y_{si} : i = 1, \dots, n\}$, $\boldsymbol{\gamma} = \{\gamma_i : i = 1, \dots, n\}$, and $\boldsymbol{\omega_{s\cdot}} = \{\omega_{si} : i = 1, \dots, n\}$ for $s = 1, 2$. With the normal prior $\mathrm{N}(-\tau_p^2/2, \tau_p^2)$, the full conditional for regression coefficients, $\beta_s = (\beta_{s1}, \beta_{s2}, \dots, \beta_{sP})'$, is derived as,

$$\beta_s | \alpha_s, \boldsymbol{\gamma}, \boldsymbol{\omega_{s\cdot}}, \boldsymbol{y_{s\cdot}} \overset{ind.}{\sim} \mathrm{N}(\xi_s, U_s),$$

where $\xi_s = U_s\Big\{X'\Big[\Big(\omega_{s1}(\log(\alpha_s^{-1}) - \log(N_{s1}) - \gamma_1) + (y_{s1} - \alpha_s^{-1})/2\Big), \dots, \Big(\omega_{sn}(\log(\alpha_s^{-1}) - \log(N_{sn}) - \gamma_n) + (y_{sn} - \alpha_s^{-1})/2\Big)\Big]' + T^{-1}(-\tau_1^2/2, \dots, -\tau_P^2/2)'\Big\}$ and $U_s = (X'\Omega_s X + T^{-1})^{-1}$. X is a design matrix consisting of the intercept and the covariates for all samples, that is, $X = (X_1, \dots, X_n)'$ with $X_i = (1, x_{i2}, \dots, x_{ip})'$. Ω_s indicates a diagonal matrix of ω_{si}, such that $\Omega_s = \mathrm{diag}(\omega_{s1}, \dots, \omega_{sn})$. T is also a diagonal matrix of $\mathrm{diag}(\tau_1^2, \dots, \tau_p^2)$.

Similarly, the full conditional for the latent factors $\boldsymbol{\gamma}$, with the normal distribution assumption of $\mathrm{N}(-\phi_i^2/2, \phi_i^2)$ with $\phi_i^2 = \phi^2$, is given by

$$(\gamma_1, \dots, \gamma_n)' | \alpha_1, \alpha_2, \beta_1, \beta_2, \phi^2, \boldsymbol{\omega_{1\cdot}}, \boldsymbol{\omega_{2\cdot}}, \boldsymbol{y_{1\cdot}}, \boldsymbol{y_{2\cdot}} \overset{ind.}{\sim} \mathrm{N}(\zeta, V),$$

where $V = (\Omega_1 + \Omega_2 + \Phi^{-1})^{-1}$ and $\zeta = V\Big\{\Big[\Big(\Big(\omega_{11}(\log(\alpha_1^{-1}) - \log(N_{11}) - X_1'\beta_1) + (y_{11} - \alpha_1^{-1})/2\Big), \dots, \Big(\omega_{1n}(\log(\alpha_1^{-1}) - \log(N_{1n}) - X_n'\beta_1) + (y_{1n} - \alpha_1^{-1})/2\Big)\Big)' + \Big(\Big(\omega_{21}(\log(\alpha_2^{-1}) - \log(N_{21}) - X_1'\beta_2) + (y_{21} - \alpha_2^{-1})/2\Big), \dots, \Big(\omega_{2n}(\log(\alpha_2^{-1}) - \log(N_{2n}) - X_n'\beta_2) + (y_{2n} - \alpha_2^{-1})/2\Big)\Big)'\Big] + \Phi^{-1}(-\phi_1^2/2, \dots, -\phi_n^2/2)'\Big\}$. Φ is a diagonal matrix of $\mathrm{diag}(\phi_1^2, \dots, \phi_n^2)$.

The full conditional for the auxiliary variables of ω_{si}, with the Pólya-Gamma distribution $\mathrm{PG}(\alpha_s^{-1} + y_{si}, 0)$ prior, takes the form of

$$p(\omega_{si}|\alpha_s, \beta_s, \gamma_i, y_{si}) \propto \exp\Big(-\omega_{si}\big(\log(N_{si}) + X_i'\beta_s + \gamma_i + \log(\alpha_s)\big)^2/2\Big)\mathrm{PG}(\omega_{si}|\alpha_s^{-1} + y_{si}, 0).$$

According to Theorem 1 of [Polson et al., 2013], this is proportional to a density function of a Pólya-Gamma distribution $\mathrm{PG}\Big(\alpha_s^{-1} + y_{si}, \big(\log(N_{si}) + X_i'\beta_s + \gamma_i + \log(\alpha_s)\big)\Big)$; using the prior conjugacy, the posterior samples of ω_{si} can be taken from the updated Pólya-Gamma distribution.

Other parameters, $(\boldsymbol{\alpha}, \phi^2, \boldsymbol{\tau^2})$, have no closed-form full conditionals with the joint full conditional density

$$\begin{aligned}
p(\boldsymbol{\alpha}, \phi^2, \boldsymbol{\tau^2}|\boldsymbol{y}, \beta_s, \boldsymbol{\gamma}) \propto & \prod_{s=1}^{2}\prod_{i=1}^{n}\Big[\frac{\Gamma(y_{si} + \alpha_s^{-1})}{\Gamma(y_{si} + 1)\Gamma(\alpha_s^{-1})}\Big(\frac{\alpha_s^{-1}}{\alpha_s^{-1} + \mu_{si}}\Big)^{\alpha_s^{-1}}\Big(\frac{\mu_{si}}{\alpha_s^{-1} + \mu_{si}}\Big)^{y_{si}}\Big] \\
& \times \mathrm{Exp}(\alpha_1|a_{\alpha_1})\mathrm{Exp}(\alpha_2|a_{\alpha_2}) \\
& \times \Big[\prod_{i=1}^{n}\mathrm{N}(\gamma_i| - \phi^2/2, \phi^2)\Big]\mathrm{Exp}(\phi^2|a_{\phi^2}) \\
& \times \Big[\prod_{p=1}^{P}\mathrm{N}(\beta_{sp}| - \tau_p^2/2, \tau_p^2)\mathrm{Exp}(\tau_p^2|a_{\tau_p^2})\Big].
\end{aligned}$$

Hence, the parameters are updated with Metropolis-Hastings steps in MCMC, using log-normal proposal distributions.

For the extended models, the dispersion parameters ϕ_l^2, $l = 1, \ldots, L$, of latent factors are updated using the Metropolis-Hastings algorithm, too. Unlike JNBM_SG, JNBM_Mix has additional parameters $\{\nu_l\}$ and $\{\xi_i\}$, given by (4). As $\mathrm{Dir}(p_{\nu_1}, \ldots, p_{\nu_L})$ is a conjugate prior for $(\nu_1, \ldots, \nu_L)$, the mixture weight parameters are updated using the Dirichlet distribution with parameters $\big(p_{\nu_1} + |\{i : \xi_i = 1,\ i = 1, \ldots, n\}|, \ldots, p_{\nu_L} + |\{i : \xi_i = L,\ i = 1, \ldots, n\}|\big)$, where $|A|$ indicates the cardinality of set A. Finally, posterior samples of auxiliary variables $\{\xi_i\}$ can be drawn from the updated discrete probability function: $\Pr(\xi_i = l) \propto \mathrm{N}(\gamma_i| - \phi_l^2/2, \phi_l^2)\nu_l$, $i = 1, \ldots, n$ and $l = 1, \ldots, L$.

As our modeling strategy is taxon-by-taxon, posterior estimates of model parameters for the microbial community (of multiple taxa) can be obtained by repeating the above Gibbs samplers multiple times, but can be done in parallel.