BAYESIAN MIXED EFFECTS MODELS FOR ZERO-INFLATED COMPOSITIONS IN MICROBIOME DATA ANALYSIS

BY BOYU REN[1,*], SERGIO BACALLADO[2], STEFANO FAVARO[3], TOMMI VATANEN[4],
CURTIS HUTTENHOWER[1,4,†] AND LORENZO TRIPPA[1,‡]

[1]*Department of Biostatistics, Harvard University,* *[*]*bor158@mail.harvard.edu;* [†]*chuttenh@hsph.harvard.edu;*
[‡]*ltrippa@jimmy.harvard.edu*

[2]*Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, sb2116@cam.ac.uk*

[3]*Departimento di Scienze Economico-Sociali e Matematico-Statistiche, Università di Torino and Collegio Carlo Alberto,
stefano.favaro@unito.it*

[4]*Liggins Institute, University of Auckland, t.vatanen@auckland.ac.nz*

Detecting associations between microbial compositions and sample characteristics is one of the most important tasks in microbiome studies. Most of the existing methods apply univariate models to single microbial species separately, with adjustments for multiple hypothesis testing. We propose a Bayesian analysis for a generalized mixed effects linear model tailored to this application. The marginal prior on each microbial composition is a Dirichlet process, and dependence across compositions is induced through a linear combination of individual covariates, such as disease biomarkers or the subject's age, and latent factors. The latent factors capture residual variability and their dimensionality is learned from the data in a fully Bayesian procedure. The proposed model is tested in data analyses and simulation studies with zero-inflated compositions. In these settings and within each sample, a large proportion of counts per microbial species are equal to zero. In our Bayesian model a priori the probability of compositions with absent microbial species is strictly positive. We propose an efficient algorithm to sample from the posterior and visualizations of model parameters which reveal associations between covariates and microbial compositions. We evaluate the proposed method in simulation studies, and then analyze a microbiome dataset for infants with type 1 diabetes which contains a large proportion of zeros in the sample-specific microbial compositions.

1. Introduction. Large-scale studies of the human microbiome have become increasingly common thanks to advances in next generation sequencing (NGS) technologies (Qin et al. (2010), Human Microbiome Project Consortium (2012)). A relevant task in these studies is to measure the association between a sample's microbial composition and individual characteristics, such as biomarkers and aspects of the sample's environment (Kostic et al. (2015), Morgan et al. (2012), Quince et al. (2013)). The abundances of microbial taxa are measured by assigning DNA reads to reference genomes. Some experiments target specific genes, such as the 16S rRNA gene, while others sample the entire bacterial genome. In all cases, the resulting count data for a collection of samples are organized into a contingency table known as the operational taxonomic unit (OTU) table.

Several methods for association studies with microbial data apply ideas from RNA-seq and other high-throughput genomic experiments (Anders and Huber (2010), Robinson, McCarthy and Smyth (2010), Paulson et al. (2013)). These methods use raw or transformed counts of microbial species to test the association of a single species with relevant covariates. Typically, these tests are carried out one species at a time by using generalized linear models (GLMs) combined with families of distributions that are overdispersed and zero inflated (Xu et al.

(2015)) to accommodate well-known characteristics of microbial abundance data (Li (2015)). The major drawback of this approach is that it models species independently. This approach does not take into account correlations across microbial species and does not allow borrowing of information across species.

The outlined limitation has prompted the introduction of joint models of microbial abundance (Chen and Li (2013), Wadsworth et al. (2017), Xia et al. (2013), Grantham et al. (2019)). These methods model the counts of I microbial species ($n_{i,j}; i = 1, \ldots, I$), of a specific sample j, say a saliva sample, with a multinomial distribution. To account for the overdispersion, these methods assume the multinomial parameter $P^j = (P_1^j, \ldots, P_I^j)$ is random and distributed accordingly to a parametric model. For example, in Chen and Li (2013) and Wadsworth et al. (2017), P^j's follow independent Dirichlet distributions and in Xia et al. (2013) and Grantham et al. (2019), P^j's follow multivariate logistic-normal distributions. To associate the covariates to the microbial compositions, all models link the parameters of each distribution of P^j (Dirichlet or logistic normal) to covariates of sample j via a regression function. Inference on the regression coefficients indicates whether a covariate is associated with the abundance of a species or not. Although these joint models overcome limitations of separate modeling of single species, the assumed distributions of the P^j's in these methods are restrictive. For instance, P_i^j is strictly positive for all i and j. This does not reflect the fact that some species can be completely absent in sample j. In addition, the variation of P^j's across samples might be mainly associated to some latent characteristics that are not observed. In this case, the methods which link model parameters exclusively to covariates do not capture dependence across species-specific residuals.

Bayesian nonparametric methods that jointly model the compositions P^j offer flexible alternatives. A widely used class of nonparametric models stems from the hierarchical Dirichlet process (Teh et al. (2006)). In its simplest form, the hierarchical Dirichlet process (HDP) assumes samples are exchangeable and the compositions P^j over $I = \infty$ species are identically distributed. The exchangeability assumption in the HDP does not capture potential association between P^j's and covariates. Nonparametric models with covariates explicitly embedded are ideal candidates for modeling dependence of compositions P^j on covariates. There are only a few such models discussed in literature. A relevant class of nonparametric models embedding covariates utilizes the Chinese restaurant processes representation (Johnson et al. (2013)). A second class of such models utilizes completely random measures (Lijoi, Nipoti and Prünster (2014)). A third class of models follows the idea in MacEachern (2000). Among this class of models, Rodríguez and Dunson (2011), Müller, Quintana and Rosner (2011), Griffin, Kolossiatis and Steel (2013) and Arbel, Mengersen and Rousseau (2016) construct dependent random measures using stick-breaking processes with atoms and weights specified through covariate-indexed stochastic processes.

Recently, a Bayesian nonparametric model for microbiome data specified through sample-specific latent factors has been discussed in Ren et al. (2017). This construction induces a marginal Dirichlet process prior for each composition P^j and introduces dependences across samples by associating microbial compositions P^j to linear combinations of latent factors. In addition, the authors introduced a link function with hard truncation at zero to model zero inflation in microbiome data. This model employs a shrinkage prior on the latent factors to produce parsimonious estimates that concentrate on a low-dimensional space.

This manuscript builds on the model of Ren et al. (2017), linking the microbial composition P^j to covariate effects as well as to the latent factors. The resulting extended model takes into account overdispersion and zero inflation in microbiome data. More importantly, it can also enable association studies for microbiome data with efficient computations. By estimating coefficients for linear combinations of relevant covariates, we visualize and infer whether a given covariate is associated with the microbial compositions or not.

We performed an extensive simulation analysis to compare the performances of the proposed model and a parametric model with latent factors (Grantham et al. (2019)) that is used in microbiome studies. The simulation study suggests that our model can accurately recover population-level trends of microbial abundances over covariates even when the model is misspecified. Our model has better performance than Grantham et al. (2019) in estimating the relationship between covariates and microbial abundances when the level of zero inflation in the data increases. We also discuss the interpretation of model parameters and propose approaches to visualize covariates' effects.

The paper is organized as follows. In Section 2 we specify the Bayesian model and discuss the identifiability of relevant model parameters. Section 3 is dedicated to computational aspects and provides an overview of the sampling algorithm for posterior inference. Section 4 presents simulation studies, and in Section 5 we discuss an application of the model to data from type 1 diabetes studies which collected longitudinal measurements from a cohort of infants. Section 6 concludes and discusses possible extensions of the proposed analyses.

2. Prior model. In this section we first review the construction of the dependent Dirichlet processes in Ren et al. (2017) and then provide a new version of the model which incorporates covariates. We also discuss the identifiability of the model parameters, including the parameters that correspond to the covariates' effects. The model will be used in the next sections to analyze the OTU table $\mathbf{n} = (n_{i,j}; i \leq I, j \leq J)$, where $n_{i,j}$ is the observed count of the microbial species i in sample j. I and J are the total number of species and samples, respectively. Our aim is to extract from the OTU table information on the relationships between microbial composition and observed samples' characteristics.

2.1. *Dependent Dirichlet processes.* In Table 1 we illustrate a subset of the OTU table from the DIABIMMUNE project (Vatanen et al. (2016)). The goal of the DIABIMMUNE project is to compare microbiome communities in infants with type 1 diabetes (T1D) or serum autoantibodies (markers predicting the onset of T1D) and healthy controls in three countries: Finland (FIN), Estonia (EST) and Russia (RUS). The study is prospective and longitudinal, and the microbial abundances are measured with shotgun sequencing. Table 1 records the counts of 10 microbial species in three Russian samples and three Finnish samples based on 16S rRNA sequencing. We denote the ith recorded species by Z_i. For instance, Z_1 is Bifidobacterium longum in Table 1.

For sample j, we assume the vector $(n_{1,j}, \ldots, n_{I,j})$ follows a multinomial distribution with unknown parameters. Our analyses extend easily to the case in which the counts $n_{i,j}$

TABLE 1
An example of OTU table (Vatanen et al. (2016))

Species	RUS1	RUS2	RUS3	FIN1	FIN2	FIN3
Bifidobacterium longum	0	73,222	3,014,074	14,294	7291	9228
Bifidobacterium bifidum	3,594,189	49,223	0	11,177	11,656,816	14,759
Escherichia coli	4,210,380	23,025	635,855	29,700	7508	556,208
Bifidobacterium breve	0	136	245,827	19,312	7,223,273	0
Bacteroides fragilis	0	88,751	0	6,257,732	343	75,506
Bacteroides vulgatus	0	7454	0	4745	0	25,859
Bacteroides dorei	0	0	0	0	0	0
Bifidobacterium adolescentis	0	111,248	1,626,357	735,715	1194	0
Bacteroides uniformis	0	3901	0	5859	1633	28,638
Ruminococcus gnavus	145,485	33,004	92,101	253,830	29	1,186,774

are Poisson random variables with unknown means. The sequencing depth $n_j = \sum_{i=1}^{I} n_{i,j}$ and the sample-specific multinomial probabilities $(P_1^j, \ldots, P_I^j)$ determine the distribution of $(n_{i,j}; i \le I)$. The probabilities $(P_1^j, \ldots, P_I^j)$ represent the microbial composition of sample j. We use $P^j(\{Z_i\}) = P_i^j$ to denote the relative abundance of Z_i in sample j. The vectors P^j vary across samples according to heterogeneity of either measured or unknown characteristics of the J samples. For example, in Table 1, the maximum likelihood estimates (MLE) of abundances of Bifidobacterium longum ($P^j(\{Z_1\})$) tend to be higher in Russian samples than in Finnish samples.

We describe the Bayesian model for the unknown compositions P^j, $j = 1, \ldots, J$ in Ren et al. (2017). Let $\mathcal{Z}$ be the set of all microbial species and $Z_i \in \mathcal{Z}$, $i \ge 1$ be a sequence of distinct species. The model does not constrain a priori the number of species present in the J samples. The relative abundance of OTU Z_i in sample j is defined as

$$(2.1) \qquad P^j(\{Z_i\}) = \frac{\sigma_i \langle \mathbf{X}_i, \mathbf{Y}_j \rangle_+^2}{\sum_{i'} \sigma_{i'} \langle \mathbf{X}_{i'}, \mathbf{Y}_j \rangle_+^2},$$

where $\sigma_i \in (0, 1)$, $\sigma_1 > \sigma_2 > \sigma_3 > \cdots$ and $\mathbf{X}_i, \mathbf{Y}_j \in \mathbb{R}^K$. The kth components of $\mathbf{X}_i$ and $\mathbf{Y}_j$ are denoted as $X_{k,i}$ and $Y_{k,j}$. We will explain the definitions of σ_i, $\mathbf{X}_i$, $\mathbf{Y}_j$ and K in the next paragraph. $\mathbb{I}(\cdot)$ is the indicator function, and $x_+ = x \times \mathbb{I}(x > 0)$. $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $\mathbb{R}^K$. We define $Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}_j \rangle$. In addition, $\boldsymbol{\sigma} = (\sigma_i; i \ge 1)$, $\mathbf{X} = (\mathbf{X}_i; i \ge 1)$, $\mathbf{Y} = (\mathbf{Y}_j; j \le J)$ and $\mathbf{Q} = (Q_{i,j}; i \ge 1, j \le J)$.

We can interpret $\sigma_i > 0$ as a summary of the overall abundance of species i across samples. We call $\mathbf{X}_i$ and $\mathbf{Y}_j$ species vector and sample vector, respectively. $\mathbf{X}_i$ and $\mathbf{Y}_j$ are latent components of the probability model. Differences across compositions P^j are determined by the $\mathbf{Y}_j$ latent vectors. Vectors $\mathbf{Y}_j$ can be interpreted as latent characteristics of the samples that affect their microbial compositions. The model assumes that there are K latent characteristics and $\mathbf{X}_i$ corresponds to the effects of these K latent characteristics on the abundance of the species Z_i.

The construction above implies that the angle $\phi_{j,j'}$ between $\mathbf{Y}_j$ and $\mathbf{Y}_{j'}$ determines the degree of similarity between compositions P^j and $P^{j'}$. Specifically, small $\phi_{j,j'}$ indicates that P^j and $P^{j'}$ are similar. When $\phi_{j,j'} = 0$, compositions P^j and $P^{j'}$ are identical. Symmetrically, the angle $\varphi_{i,i'}$ between $\mathbf{X}_i$ and $\mathbf{X}_{i'}$ can be viewed as a measure of similarity between species Z_i and $Z_{i'}$. When $\varphi_{i,i'}$ decreases toward zero, the correlation between $(P^j(\{Z_i\}); j \le J)$ and $(P^j(\{Z_{i'}\}); j \le J)$ increases to one.

The prior specification in the model is as follows. First, $\sigma_1 > \sigma_2 > \sigma_3 \ldots$ are a priori ordered points from a Poisson process on $(0, 1)$ with intensity $\nu(\sigma) = \alpha \sigma^{-1}(1 - \sigma)^{-1/2}$. Second, the $X_{k,i}$ random variables are independent Gaussian $\mathcal{N}(0, 1)$, $i = 1, 2, \ldots$, $k = 1, 2, \ldots, K$. We can assume for the moment that the $\mathbf{Y}_j$'s are fixed.

The resulting marginal prior distribution on the composition P^j is a Dirichlet process (Ren et al. (2017)). In addition, P^j and $P^{j'}$ are dependent for $j \ne j'$. To provide some intuition on this construction of Dirichlet process, we consider a similar model with $I < \infty$ species. For simplicity, we set $\|\mathbf{Y}_j\| = 1$, where $\|\cdot\|$ is the Euclidean norm of a real vector. The prior on $\mathbf{X}_i$'s induces a standard normal distribution on $(Q_{1,j}, \ldots, Q_{I,j})$. The prior distribution of $(Q_{i,j})_+^2$ is therefore a mixture of a point mass at zero and a Gamma$(1/2, 1/2)$ distribution. Assume $(\sigma_1, \ldots, \sigma_I)$ are independent Beta$(\alpha/I, 1/2 - \alpha/I)$ variables. It can be verified by moment generating function that the joint law of these ordered independent Beta random variables converges to the law of a Poisson process on $(0, 1)$ with intensity $\alpha \sigma^{-1}(1 - \sigma)^{-1/2}$ when $I \to \infty$. The products $(\sigma_i(Q_{i,j})_+^2, i = 1, \ldots, I)$ then follow a mixture distribution of a point mass at zero and a Gamma$(\alpha/I, 1/2)$. The normalized vector $(\sigma_i(Q_{i,j})_+^2 / \sum_{i'} \sigma_{i'}(Q_{i',j})_+^2, i = 1, \ldots, I)$, conditioned on $(\mathbb{I}(Q_{1,j} > 0), \ldots, \mathbb{I}(Q_{I,j} > 0))$,

follows a Dirichlet distribution with weights proportional to $\mathbb{I}(Q_{i,j} > 0)$. If $I \to \infty$, we know that the $(\sigma_1, \ldots, \sigma_I)$ converges in distribution to the Poisson process with intensity ν, and $(\sigma_i(Q_{i,j})_+^2 / \sum_{i'} \sigma_{i'}(Q_{i',j})_+^2, i = 1, \ldots, I)$ becomes a Dirichlet process (Ferguson (1973)). This holds also when $\|\mathbf{Y}_j\| \neq 1$ because the distribution of $(\sigma_i(Q_{i,j})_+^2 / \sum_{i'} \sigma_{i'}(Q_{i',j})_+^2, i = 1, \ldots, I)$ does not depend on $\|\mathbf{Y}_j\|$.

For inferential and visualization purposes it is desirable that the $\mathbf{Y}_j$ latent vectors concentrate approximately on a low-dimensional space. The resulting $\mathbf{Y}_j$ are parsimonious latent factors that capture the variability of observed species abundances across samples. To this end, the model applies the prior studied in Bhattacharya and Dunson (2011),

$$\mathbf{Y}_j \sim \mathcal{N}(\mathbf{0}, \operatorname{diag}\{\gamma_1, \ldots, \gamma_K\}),$$

where γ_k rapidly decrease with k. The prior formalizes the desiderata of having the norm $\|\mathbf{Y}_j\|$ mostly driven by the first few components of $\mathbf{Y}_j$, say the first three components $(Y_{1,j}, Y_{2,j}, Y_{3,j})$ and the rest of the components, $(Y_{4,j}, \ldots, Y_{K,j})$, vanish with negligible values. In different words, only a small set of $\mathbf{Y}_j$ entries—three in the example—are relevant. This approach is preferable to a hyperprior on the dimensionality of $\mathbf{Y}_j$ mainly for computational convenience.

2.2. *Fixed effects.* The goal of this subsection is to model relationships between microbial compositions and samples' characteristics. For example, in studies of inflammatory bowel disease (IBD) (Gevers et al. (2014), Greenblum, Turnbaugh and Borenstein (2012), Morgan et al. (2012)) researchers were interested in identifying microbes that correlate with the onset of IBD to develop therapeutic hypotheses. These analyses typically utilize regression models where the outcomes coincide with OTU abundances. Following a similar strategy, we expand the model in Section 2.1.

Assume there are $L \geq 1$ observed covariates. We use $\mathbf{w}_j = (w_{l,j}; l = 1, \ldots, L)$ to denote the covariates' values for sample j. The effects of this set of covariates on species i are $\mathbf{v}_i = (v_{l,i}; l = 1, \ldots, L)$. The collection of all $\mathbf{w}_j$ and $\mathbf{v}_i$ are $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_J)$ and $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_I)$. Our extended model directly modifies the random variables $Q_{i,j}$'s introduced in the definition of the model (2.1), by adding a linear function of $\mathbf{w}_j$ and an error term

$$(2.2) \qquad Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}_j \rangle + \langle \mathbf{v}_i, \mathbf{w}_j \rangle + \epsilon_{i,j},$$

where $\epsilon_{i,j} \overset{iid}{\sim} \mathcal{N}(0, 1)$ is the error term. Thus,

$$P^j(\{Z_i\}) = \frac{\sigma_i(Q_{i,j})_+^2}{\sum_{i'} \sigma_{i'}(Q_{i',j})_+^2}.$$

The inner product $\langle \mathbf{v}_i, \mathbf{w}_j \rangle$ represents the fixed effects of our model, whereas $\langle \mathbf{X}_i, \mathbf{Y}_j \rangle$ represents the random effects. We fix the variance of the errors to one since the model for P^j is invariant if we rescale all $Q_{i,j}$ variables by a fixed multiplicative term.

In this construction $\mathbf{v}_i$ and $\mathbf{w}_j$ can be viewed as additional dimensions of $\mathbf{X}_i$ and $\mathbf{Y}_j$, respectively. The angle between $(\mathbf{w}_j, \mathbf{Y}_j)$ and $(\mathbf{w}_{j'}, \mathbf{Y}_{j'})$, denoted as $\tilde{\phi}_{j,j'}$, measures the similarity between the microbial compositions P^j and $P^{j'}$. As in model (2.1), one can verify that the correlation $\operatorname{cor}(P^j(A), P^{j'}(A))$ is monotone with respect to $\tilde{\phi}_{j,j'}$. Similarly, the angle between $(\mathbf{v}_i, \mathbf{X}_i)$ and $(\mathbf{v}_{i'}, \mathbf{X}_{i'})$, $\tilde{\varphi}_{i,i'}$, is representative of the correlation between abundances of species i and i' across samples. A small $\tilde{\varphi}_{i,i'}$ value makes the correlation between vectors $(P^j(\{Z_i\}); j \leq J)$ and $(P^j(\{Z_{i'}\}); j \leq J)$ close to one.

The coefficients $\mathbf{v}_i$ are a priori independent normal random variables with mean zero and variance one. When the latent factors $\mathbf{Y}$ are fixed and the prior for $\mathbf{X}_i$ and σ_i remains the same as in Section 2.1, the microbial composition P^j, for each $j = 1, \ldots, J$, retains a marginal
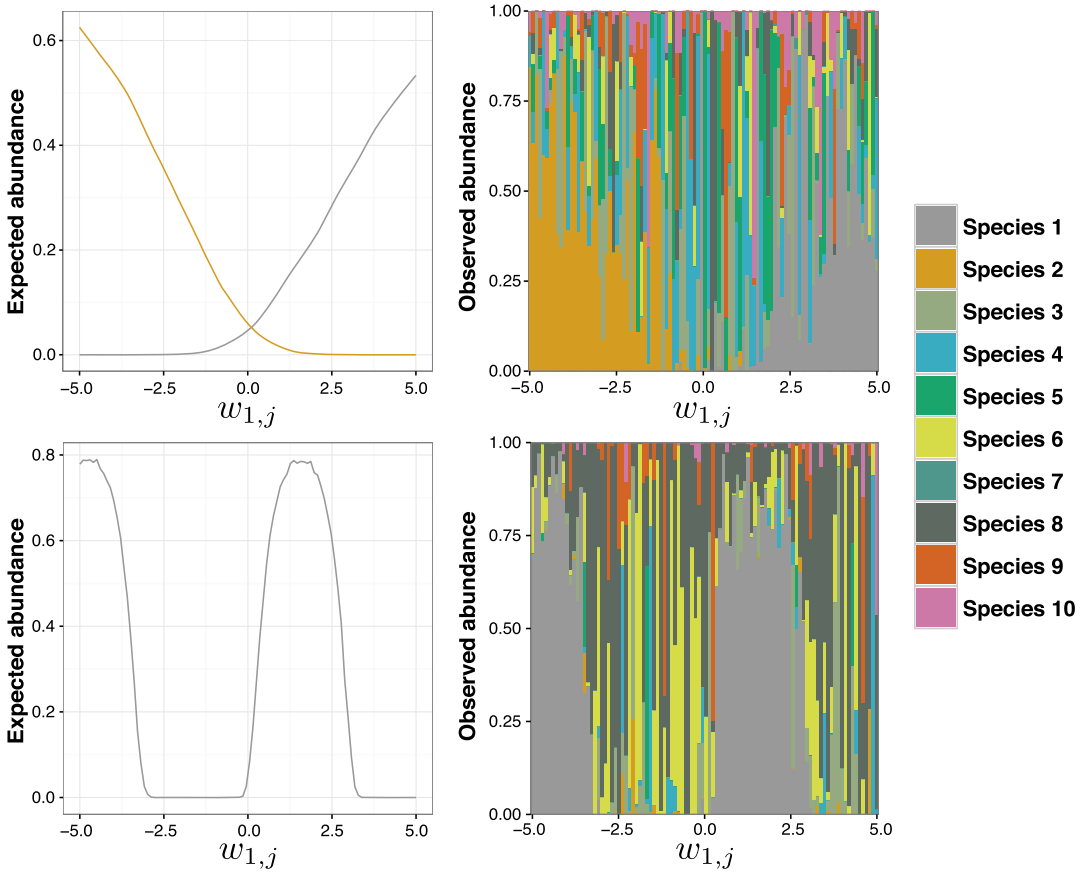
FIG. 1. *Effect of a single covariate $w_{1,j}$ on microbial species abundances. We illustrate the expected abundances of species* 1 *and* 2 *when $w_{1,j}$ varies* (Left) *and the observed abundances of species* 1–10 *in one simulated dataset as $w_{1,j}$ changes* (Right). *We focus on a single sample j and fix the random effects $\langle \mathbf{X}_i, \mathbf{Y}_j \rangle$ in all simulations. Only the value of $w_{1,j}$ and the error terms $\epsilon_{i,j}$'s vary. The expected abundances are calculated by averaging over* 1000 *simulation replicates. We consider the cases where $Q_{i,j} = v_{1,i} w_{1,j} + \langle \mathbf{X}_i, \mathbf{Y}_j \rangle + \epsilon_{i,j}$ with $v_{1,1} = 5$, $v_{1,2} = -5$ and $v_{1,i} = 0$ for $i > 2$* (Top) *and $Q_{i,j} = v_{1,i} \sin(w_{1,j}) + \langle \mathbf{X}_i, \mathbf{Y}_j \rangle + \epsilon_{i,j}$ with $v_{1,1} = 5$ and $v_{1,i} = 0$ for $i > 1$* (Bottom). *The covariate $w_{1,j}$ varies from* −5 *to* 5 *with* 0.1 *increments.*

Dirichlet process distribution. More precisely, P^j is a Dirichlet process with concentration parameter α. This can be shown using the same argument as in Section 2.1.

It is important not to misinterpret the coefficients $\mathbf{v}_i$. The species abundances are not linear functions of the covariates (see Figure 1). In certain cases the relationship between the covariates and the species abundances is not monotone. Consider a single covariate $w_{1,j}$, and assume $Q_{i,j} = v_{1,i} w_{1,j} + \langle \mathbf{X}_i, \mathbf{Y}_j \rangle + \epsilon_{i,j}$, where $v_{1,1} = 5$, $v_{1,2} = 1$ and $v_{1,i} = 0$ when $i > 2$. For simplicity, assume in addition σ_i variables all equal to 0.5. When $w_{1,j}$ is small, say $w_{1,j} \in (0, 0.5)$, the abundances of species 1 and 2 increases with $w_{1,j}$. However, as $w_{1,j}$ gets larger, say $w_{1,j} > 5$, species 1 will dominate all other species, and the abundance of species 2 decreases to nearly zero.

2.2.1. *Models for data analysis.* In our analyses we considered longitudinal data with repeated measurements over time for each individual. Assume samples $j = 1, \ldots, J$ are partitioned into U groups, that is, U distinct individuals. We use u_j to identify the individual associated to sample j. We enforce the samples j and j' from the same individual u ($u_j = u_{j'} = u$) to share common latent factors $\mathbf{Y}_u$. The *longitudinal* version of model (2.2)

utilizes

$$Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}_{u_j} \rangle + \langle \mathbf{v}_i, \mathbf{w}_j \rangle + \epsilon_{i,j}. \tag{2.3}$$

The rationale for this model is that samples derived from the same individual tend to be similar. The covariates $\mathbf{w}_j$ will include time information (e.g., individual's age) for each sample j. This version of the model is tailored toward longitudinal data and studies with repeated measurements, and it allows one to visualize time trends of microbial compositions.

We will use a truncated version of model (2.2) or (2.3) in data analyses, which we call the *finite-species* model. Truncating the stick-breaking representation of Dirichlet process has been studied extensively in literature (Ishwaran and Zarepour (2002)). The truncated process can be arbitrarily close to the Dirichlet process in total variation distance if the number of retained atoms is large. In our case we truncate the model (2.3) at the number of observed species, I. This is sufficient for data analysis as the sequencing depth in microbiome studies is generally large enough to capture most of the microbial species of interest. With $I < \infty$ species the finite-species model is defined by

$$Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}_j \rangle + \langle \mathbf{v}_i, \mathbf{w}_j \rangle + \epsilon_{i,j},$$

$$P^j(\{Z_i\}) = \frac{\sigma_i (Q_{i,j})_+^2}{\sum_{i'=1}^{I} \sigma_{i'} (Q_{i',j})_+^2}, \quad i = 1, \dots, I. \tag{2.4}$$

The prior for $\mathbf{X}_i$ and $\mathbf{Y}_j$ remain identical. The prior for σ_i's becomes $\sigma_i \overset{iid}{\sim} \text{Beta}(\alpha/I, 1/2 - \alpha/I)$.

2.3. *Identifiability.* In this subsection we consider the identifiability of the proposed model. Since the model is invariant under simultaneous rotations of the vectors $\mathbf{Y}_j$ and $\mathbf{X}_i$, we cannot learn $\mathbf{Y}$ from the data. We discuss the identifiability of the correlation matrix $\mathbf{S}$ associated to $\boldsymbol{\Sigma} = \mathbf{Y}^\mathsf{T}\mathbf{Y} + \mathbf{I}$, where $\mathbf{I}$ is the $J \times J$ identity matrix. Similarly, since the composition P^j is invariant to scale transformation of $\boldsymbol{\sigma}$, we will discuss identifiability of the ratios $\sigma_i/\sigma_{i'}$ for $i \neq i'$. We assume that the number of samples is finite and that covariates $\mathbf{w}_j$'s are independent with $\mathbb{E}(\mathbf{w}_j \mathbf{w}_j^\mathsf{T})$ of full rank.

We proceed assuming initially that $(P^j(\{Z_i\}); i \geq 1, j \leq J)$ are observable random variables. Recall that

$$(Q_{i,j}; j \leq J)|\mathbf{v}_i, \mathbf{Y}, \mathbf{w} \sim \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{v}_i, \boldsymbol{\Sigma}). \tag{2.5}$$

Since we assume that $P^j(\{Z_i\})$ is observable, we have that $P^j(\{Z_i\}) = 0$ implies $Q_{i,j} \leq 0$. Consider a set of new random variables, denoted as $(\widetilde{P}^j(\{Z_i\}); i \leq I, j \leq J)$, where $\widetilde{P}^j(\{Z_i\}) = \mathbb{I}(P^j(\{Z_i\}) > 0)$. From (2.5) the conditional distribution of $(\widetilde{P}^j(\{Z_i\}); i \geq 1, j \leq J)$ given $(\boldsymbol{\sigma}, \mathbf{Y}, \mathbf{v}, \mathbf{w})$ is

$$p(\widetilde{P}^j(\{Z_i\}), i \geq 1, j \leq J | \boldsymbol{\sigma}, \mathbf{Y}, \mathbf{v}, \mathbf{w})$$
$$\propto \prod_i \left[\int_{\mathcal{A}_i} (2\pi)^{-J/2} |\boldsymbol{\Sigma}|^{-1/2} \times \exp\left(-\frac{1}{2}(\mathbf{Q}_i - \boldsymbol{\mu}_i)^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{Q}_i - \boldsymbol{\mu}_i)\right) d\mathbf{Q}_i \right]. \tag{2.6}$$

Here, $\mathbf{Q}_i = (Q_{i,1}, \dots, Q_{i,J})$, $\boldsymbol{\mu}_i = \mathbf{w}^\mathsf{T}\mathbf{v}_i$, $\mathcal{A}_i = \times_{j=1}^{J} \mathcal{A}_{i,j}$ and $\mathcal{A}_{i,j} = (-\infty, 0]$ if $\widetilde{P}^j(\{Z_i\}) = 0$, while $\mathcal{A}_{i,j} = (0, \infty)$ when $\widetilde{P}^j(\{Z_i\}) = 1$. To illustrate the identifiability of the parameters $(\sigma_i/\sigma_{i'}; i \neq i')$, $\mathbf{S}$ and $\mathbf{v}$, we start with two simplified cases and then give a proposition:

1. *Without random effects* ($\mathbf{Y} = \mathbf{0}$). We first note that conditioning on $\mathbf{w}$, for a fixed i, $(\widetilde{P}^j(\{Z_i\}); j \leq J)$ are samples from a standard probit model (Albert and Chib (1993)), where

$\mathbf{v}_i$ serves as regression coefficients and the sample covariates are $\mathbf{w}_j$. Based on the theory of generalized linear models, $\mathbf{v}_i$ is identifiable when $\mathbb{E}(\mathbf{w}_j \mathbf{w}_j^{\mathsf{T}})$ is of full rank.

We then consider $(\sigma_i/\sigma_{i'}; i \neq i')$. By construction,

$$\frac{P^j(\{Z_i\})}{P^j(\{Z_{i'}\})} = \frac{\sigma_i}{\sigma_{i'}} \frac{(Q_{i,j})_+^2}{(Q_{i',j})_+^2}.$$

Here, we use the convention that the ratio is zero whenever the denominator is zero. To ensure the identifiability of $(\sigma_i/\sigma_{i'}; i \geq 1, j \leq J)$, we want to show that if

$$\left(P^j(\{Z_i\}); i \geq 1, j \leq J\right), \mathbf{w}|\mathbf{v}, \boldsymbol{\sigma} \stackrel{d}{=} \left(P^j(\{Z_i\}); i \geq 1, j \leq J\right), \mathbf{w}|\mathbf{v}', \boldsymbol{\sigma}',$$

then $\sigma_i/\sigma_{i'} = \sigma_i'/\sigma_{i'}'$ for all $i \neq i'$. Using the identifiability of $\mathbf{v}_i$, the above equality in distribution implies $\mathbf{v}_i = \mathbf{v}_i'$, and in turn the equality of the conditional distributions $p(((Q_{i,j})_+^2, (Q_{i',j})_+^2), \mathbf{w}_j|\mathbf{v}, \boldsymbol{\sigma})$ and $p(((Q_{i,j})_+^2, (Q_{i',j})_+^2), \mathbf{w}_j|\mathbf{v}', \boldsymbol{\sigma}')$. This directly implies $\sigma_i/\sigma_{i'} = \sigma_i'/\sigma_{i'}'$ for all $i \neq i'$.

2. *Without fixed effects* ($\mathbf{v}_i = \mathbf{0}$). We consider $\boldsymbol{\sigma}$ and $\mathbf{S}$. The distribution of $(\widetilde{P}^j(\{Z_i\}), \widetilde{P}^{j'}(\{Z_i\}))$ is

$$p(\widetilde{P}^j(\{Z_i\}), \widetilde{P}^{j'}(\{Z_i\})|\boldsymbol{\sigma}, \mathbf{Y}) = \frac{1}{2\pi} \int_{\mathcal{A}_{i,j} \times \mathcal{A}_{i,j'}} (1 - S_{j,j'}^2)^{-1/2} \exp\left(-\frac{1}{2}\mathbf{q}^{\mathsf{T}} \mathbf{S}_{j:j'}^{-1} \mathbf{q}\right) d\mathbf{q},$$

where $S_{j,j'}$ is the correlation between $Q_{i,j}$ and $Q_{i,j'}$ and $\mathbf{S}_{j:j'}$ is the correlation matrix of $(Q_{i,j}, Q_{i,j'})$. $\mathcal{A}_{i,j} = (-\infty, 0]$ if $\widetilde{P}^j(\{Z_i\}) = 0$, while $\mathcal{A}_{i,j} = (0, \infty)$ if $\widetilde{P}^j(\{Z_i\}) = 1$. Corollary 3.12 in Ledoux and Talagrand (2011) shows that $p(\widetilde{P}^j(\{Z_i\}), \widetilde{P}^{j'}(\{Z_i\})|\mathbf{v}_i, \mathbf{Y})$, when $\mathbf{v}_i = \mathbf{0}$, is monotone with respect to $S_{j,j'}$. This implies that $S_{j,j'}$ is identifiable.

Using the same arguments as in the case where no random effect is present, one can show that the ratios $(\sigma_i/\sigma_{i'}; i \neq i')$ remain identifiable.

In the general case the identifiability of the model parameters, with both fixed and random effects, is described through Proposition 1 in Section S1 of the Supplementary Material (Ren et al. (2020)).

3. Posterior simulations and visualization of covariates' effects. In this section we focus on posterior inference and computational aspects. In Section 3.1 we introduce an algorithm for posterior simulations with the model described in Section 2.2. Then, in Section 3.2 we propose graphical visualizations to illustrate associations of microbial compositions and covariates. These representations are relevant for the analysis of microbial abundances because, as we mentioned in Section 2.2, a positive (or negative) element of the vector $\mathbf{v}_i$, say the lth element, does not imply a monotone relation between the lth covariate and the abundances of species i. To illustrate the relation between the lth covariate and species i, we estimate how the abundance of species i would vary at hypothetical values of the lth covariate.

3.1. *Posterior simulations.* We proceed with the finite-species model (2.4). The likelihood function is

$$p(\mathbf{n}|\mathbf{Q}, \boldsymbol{\sigma}) \propto \left(\prod_{j=1}^J \prod_{i=1}^I (\sigma_i (Q_{i,j})_+^2)^{n_{i,j}}\right) \times \prod_{j=1}^J \left(\sum_{i=1}^I \sigma_i (Q_{i,j})_+^2\right)^{-n_j}$$

and

(3.1)
$$p(\boldsymbol{\sigma}, \mathbf{Q}, \mathbf{X}, \mathbf{Y}, \mathbf{v}|\mathbf{n}, \mathbf{w}) \propto \left(\prod_{j=1}^J \prod_{i=1}^I (\sigma_i (Q_{i,j})_+^2)^{n_{i,j}}\right) \times \prod_{j=1}^J \left(\sum_{i=1}^I \sigma_i (Q_{i,j})_+^2\right)^{-n_j}$$

$$\times \pi(\boldsymbol{\sigma}, \mathbf{Q}|\mathbf{X}, \mathbf{Y}, \mathbf{v}, \mathbf{w})\pi(\mathbf{X}, \mathbf{Y}, \mathbf{v}),$$

where π indicates the prior. By introducing positive latent random variables $\mathbf{T} = (T_1, \ldots, T_J)$, as in James, Lijoi and Prünster (2009), we rewrite the conditional distribution

$$
\begin{aligned}
&p(\boldsymbol{\sigma}, \mathbf{Q}, \mathbf{X}, \mathbf{Y}, \mathbf{v} | \mathbf{n}, \mathbf{w}) \\
&\propto \int \pi(\boldsymbol{\sigma}, \mathbf{Q}, \mathbf{X}, \mathbf{Y}, \mathbf{v} | \mathbf{w}) \\
&\quad \times \prod_{j=1}^{J} \left\{ \left(\prod_{i=1}^{I} (\sigma_i (Q_{i,j})_+^2)^{n_{i,j}} \right) T_j^{n_j - 1} \exp\left(-T_j \sum_i \sigma_i (Q_{i,j})_+^2 \right) \right\} d\mathbf{T}.
\end{aligned}
$$
(3.2)

We use a Gibbs sampler to perform posterior simulations. The algorithm iteratively samples $\boldsymbol{\sigma}, \mathbf{T}, \mathbf{Q}, \mathbf{X}, \mathbf{Y}$ and $\mathbf{v}$ from the full conditional distributions. We describe the two components of the algorithm:

1. The first component samples $\boldsymbol{\sigma}, \mathbf{T}$ and $\mathbf{Q}$ from the full-conditional distributions. We note that $\sigma_1, \ldots, \sigma_I$, given the remaining variables, are conditionally independent. The sampling of $(\sigma_1, \ldots, \sigma_I)$ from the full conditional distribution is identical as in Ren et al. (2017). The random variables $T_1, \ldots, T_J$, given $(\mathbf{Q}, \mathbf{n}, \boldsymbol{\sigma})$, are conditionally independent with Gamma distributions. These random variables can be straightforwardly generated from the full conditional distribution. To complete this part of the algorithm we can write

$$
\begin{aligned}
&p(Q_{i,j} | \mathbf{n}, \mathbf{Q}_{-i,-j}, \mathbf{T}, \boldsymbol{\sigma}, \mathbf{X}, \mathbf{Y}, \mathbf{w}, \mathbf{v}) \\
&\propto (Q_{i,j})_+^{2n_{i,j}} \times \exp(-T_j \sigma_i (Q_{i,j})_+^2) \times \exp\left(-\frac{(Q_{i,j} - \langle \mathbf{X}_i, \mathbf{Y}_j \rangle - \langle \mathbf{v}_i, \mathbf{w}_j \rangle)^2}{2} \right),
\end{aligned}
$$
(3.3)

where $\mathbf{Q}_{-i,-j}$ is identical to $\mathbf{Q}$ with the only exception that it does not include $Q_{i,j}$. The density (3.3) indicates that the $Q_{i,j}$'s random variables are conditionally independent. We also note that the density in (3.3) is log concave. We use these arguments to sample $\mathbf{Q}$ from the full conditional distribution.

2. The second component considers the sampling of $\mathbf{Y}, \mathbf{X}$ and $\mathbf{v}$ from the full conditional distributions. Using expression (3.2) we write

$$
p(\mathbf{X} | \mathbf{n}, \boldsymbol{\sigma}, \mathbf{T}, \mathbf{Q}, \mathbf{Y}, \mathbf{v}, \mathbf{w}) \propto \exp\left(-\sum_{i,j} \frac{(Q_{i,j} - \langle \mathbf{X}_i, \mathbf{Y}_j \rangle - \langle \mathbf{v}_i, \mathbf{w}_j \rangle)^2}{2} \right) \times \pi(\mathbf{X}).
$$

Recall that the $\mathbf{X}_i$'s are a priori independent normal random variables. Therefore, the full conditional distribution of $\mathbf{X}_i$ coincides with the conjugate posterior distribution in a standard linear model (Lindley and Smith (1972)). Sampling of $\mathbf{Y}$ and $\mathbf{v}$ from the full conditional distributions follows identical arguments. Indeed, the prior model studied in Bhattacharya and Dunson (2011), which we use for $\mathbf{Y}$, is conditionally conjugate.

3.2. *Visualization of covariate effects.* We consider the partial derivatives

$$
\frac{\partial P^j(\{Z_i\})}{\partial w_{l,j}} := \partial \left[\frac{\sigma_i (\langle \mathbf{X}_i, \mathbf{Y}_j \rangle + \langle \mathbf{v}_i, \mathbf{w}_j \rangle + \epsilon_{i,j})_+^2}{\sum_{i'} \sigma_{i'} (\langle \mathbf{X}_{i'}, \mathbf{Y}_j \rangle + \langle \mathbf{v}_{i'}, \mathbf{w}_j \rangle + \epsilon_{i',j})_+^2} \right] \Big/ \partial w_{l,j}.
$$

The derivative $\partial P^j(\{Z_i\})/\partial w_{l,j}$ quantifies the abundance variation of species i in sample j in response to an infinitesimal increment of the lth component of $\mathbf{w}_j$. We can estimate these derivatives from the data using the posterior approximation obtained by the algorithm in Section 3.1. We use the estimates $\mathbb{E}(\partial P^j(\{Z_i\})/\partial w_{l,j} | \mathbf{n}, \mathbf{w})$. For example, the top row of Figure 3 in Section 4.2 summarizes the posterior distributions of $\partial P^j(\{Z_i\})/\partial w_{1,j}$, $j = 1, \ldots, 300$, for three species. Details on the figure, including a description of the simulated data that generated the panels, are provided in Section 4.2. In species 1 the estimates of the

derivatives are positive for the majority of the samples and tend to be large when $w_{1,j} > 0$. We also note that the estimates of $\partial P^j(\{Z_i\})/\partial w_{l,j}$ are larger for samples in the subgroup $w_{2,j} = 1$ than in the subgroup $w_{2,j} = 0$. These results indicate that, for any $j = 1, \ldots, 300$, if we could increase (decrease) the value of $w_{1,j}$ while holding $w_{2,j}$ fixed, then one would expect an increase (decrease) of the relative abundance of species 1, and this trend appears more pronounced in those samples with $w_{1,j} > 0$ and $w_{2,j} = 1$.

We also define

$$P^j(\{Z_i\}; \mathbf{w}_0) := \frac{\sigma_i(\langle \mathbf{X}_i, \mathbf{Y}_j \rangle + \langle \mathbf{v}_i, \mathbf{w}_0 \rangle + \epsilon_{i,j})_+^2}{\sum_{i'} \sigma_{i'}(\langle \mathbf{X}_{i'}, \mathbf{Y}_j \rangle + \langle \mathbf{v}_{i'}, \mathbf{w}_0 \rangle + \epsilon_{i',j})_+^2};$$

it is the abundance of species i if the covariates values of sample j could be enforced to be equal to $\mathbf{w}_0$. When estimating the effect of a binary covariate $w_{l,j} \in \{0, 1\}$ on microbial compositions, we replace derivatives by differences

(3.4)
$$\frac{\Delta P^j(\{Z_i\})}{\Delta w_{l,j}} := P^j(\{Z_i\}; \mathbf{w}_{l,j}^1) - P^j(\{Z_i\}; \mathbf{w}_{l,j}^0).$$

Here, $\mathbf{w}_{l,j}^1$ is identical to $\mathbf{w}_j$ with the exception that the lth component $w_{l,j}$ is set to be one and, symmetrically, $\mathbf{w}_{l,j}^0$ is specified with $w_{l,j}$ equal to zero. Therefore, $\Delta P^j(\{Z_i\})/\Delta w_{l,j}$ is the variation of $P^j(\{Z_i\})$ that one would observe by changing the value of a binary covariate.

We also consider the population-level associations between microbial compositions and a specific covariate, say the lth covariate, when adjusting for all other covariates. To this end, we first define $\bar{P}(\{Z_i\}; \mathbf{w}_0)$, the *population average abundance* of species i at a covariate value $\mathbf{w}_0$, by

$$\bar{P}(\{Z_i\}; \mathbf{w}_0) := \frac{1}{J} \left(\sum_{j=1}^J P^j(\{Z_i\}; \mathbf{w}_0) \right),$$

which quantifies the average abundance of species i when all J samples in the study have the same hypothetical covariates values $\mathbf{w}_0$. We estimate $\bar{P}(\{Z_i\}; \mathbf{w}_0)$ from the data with $\mathbb{E}(\bar{P}(\{Z_i\}; \mathbf{w}_0) | \mathbf{n}, \mathbf{w})$.

To illustrate the association between the abundance of species i and the lth covariate, we visualize the variation of $\bar{P}(\{Z_i\}; \mathbf{w}_0)$ as $w_{l,0}$ (the lth entry of $\mathbf{w}_0$) varies and all other covariates remain fixed at $\mathbf{w}_{-l,0}$. This visualization is obtained by plotting the estimated $\bar{P}(\{Z_i\}; \mathbf{w}_0)$ against $w_{l,0}$. We call the resulting curve the *population trend* of species i with respect to the lth covariate at $\mathbf{w}_{-l,0}$. In Figure 3's bottom row, we illustrate population trends of three species with respect to the first covariate at $w_{2,0} = 0$ and at $w_{2,0} = 1$.

Interactions terms for pairs of covariates and, more generally, functions of the covariates can be included in the proposed model. We specify a function $\mathbf{f}: \mathbb{R}^L \to \mathbb{R}^{L'}$ for interaction terms. One example is $\mathbf{f}(\mathbf{w}_j) = w_{1,j} w_{2,j}$. The definition of $Q_{i,j}$ in (2.2), when interactions are incorporated, becomes

$$Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}_j \rangle + \langle \mathbf{v}_i, (\mathbf{w}_j, \mathbf{f}(\mathbf{w}_j)) \rangle + \epsilon_{i,j},$$

where $\mathbf{v}_i \in \mathbb{R}^{L+L'}$. In these cases variations of the lth coordinate of $\mathbf{w}_j$ affect $\mathbf{f}(\mathbf{w}_j)$ and translate into compositional variations equal to $\partial P^j(\{Z_i\})/\partial w_{l,j}$ or $\Delta P^j(\{Z_i\})/\Delta w_{l,j}$.

4. Simulation study. In this section we focus on the model introduced in Section 2.2, and we illustrate that we can transform the model parameters into interpretable results on the relationship between covariates and microbial compositions. We also provide in this section a comparison between our model and a recent published latent factor model MIMIX

TABLE 2
Specification of **v** *in the simulation study*

Species (i)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	...	100
$v_{1,i}$ (for $w_{1,j}$)	5	5	5	5	5	5	5	5	−5	−5	−5	−5	−5	−5	−5	−5	0	...	0
$v_{2,i}$ (for $w_{2,j}$)	5	5	5	5	−5	−5	−5	−5	5	5	5	5	−5	−5	−5	−5	0	...	0
$v_{3,i}$ (for $w_{1,j} \cdot w_{2,j}$)	10	−5	−5	−10	10	−5	−5	−10	−10	5	5	10	−10	5	5	10	0	...	0

(Grantham et al. (2019)) which uses the logistic-normal distribution to link covariates to the relative abundances of species. We illustrate in simulation scenarios that the proposed model has similar performance to the logistic-normal model even when the data is generated from MIMIX. When the degree of zero inflation is large, our model tends to outperform MIMIX regression regardless of the underlying data generating model. The code for replicating the simulation studies is available in the Supplementary Materials (Ren et al. (2020)) and online at https://github.com/boyuren158/DirFactor-fix.

In our simulation study we included $I = 100$ species and $J = 300$ samples. The 300 samples are taken from $U = 50$ individuals (see Section 2.2.1). Each individual is measured six times. The read depth of each sample is $n_j = 10^5$. We simulate $\boldsymbol{\sigma}$ using independent Beta densities with mean 0.2 and variance 0.1. As we discussed in Section 2.1, σ_i represents the average abundance of species i across all samples. We included in the simulation a continuous covariate $w_{1,j}$, generated from independent $\mathcal{N}(0, 1)$ distributions, and a binary covariate $w_{2,j}$, generated from independent Bernoulli(0.5). We also use the interaction term $w_{1,j} \times w_{2,j}$ to specify scenarios where effects of $w_{1,j}$ differ in the groups $w_{2,j} = 0$ and $w_{2,j} = 1$. We will later discuss in Section 5 this type of interaction in a microbiome study for type 1 diabetes.

For the latent factors **Y** we assumed $\mathbf{Y}_u \in \mathbb{R}^4$. For the first half of the individuals, $u = 1, \ldots, 25$, we set $Y_{3,u} = Y_{4,u} = 0$ while for the other half, $u = 26, \ldots, 50$, we set symmetrically $Y_{1,u} = Y_{2,u} = 0$. The nonzero components in $\mathbf{Y}_u$ were simulated independently from a $\mathcal{N}(0, 1)$ density. This specification of **Y** makes the correlation matrix **S** block diagonal (see Figure 2(b)).

We simulate the first eight species with positive $v_{1,i}$'s and the following eight species ($i = 9, \ldots, 16$) with negative $v_{1,i}$'s. As detailed in Table 2, the first 16 species abundances correlate with $w_{2,j}$. Moreover, we make the assumption that some of the trends with respect to $w_{1,j}$ are either amplified or reversed when we contrast the two groups $w_{2,j} = 1$ and $w_{2,j} = 0$. All other species ($i > 16$) have the corresponding $\mathbf{v}_i$ coefficients equal to **0** (Table 2).

We further examine the robustness of our method by checking its performances when the link function between P^j and $Q_{i,j}$ is misspecified. In particular, we apply our method to data simulated using the following specification of $(P^j(\{Z_i\}); i \leq I, j \leq J)$,

$$(4.1) \qquad P^j(\{Z_i\}) = \frac{\sigma_i Q_{i,j}^+}{\sum_{i'} \sigma_{i'} Q_{i',j}^+}.$$

The specification of $\boldsymbol{\sigma}$, **v**, **Y** and **w** remains the same as described in the previous paragraphs.

4.1. *Estimating species and samples parameters* **v** *and* **S**. We first consider estimation of **v** and **S** between individuals when the model is correctly specified. Recall, from Proposition 1 in the Supplementary Material (Ren et al. (2020)), **v** is identifiable when trace($\boldsymbol{\Sigma}$) is assumed fixed at a constant value. We assume trace($\boldsymbol{\Sigma}$) = 1 and compute the posterior distribution of $\mathbf{v}/\sqrt{\text{trace}(\boldsymbol{\Sigma})}$. The performance of the estimate of **S** is measured by the RV-coefficient (Robert and Escoufier (1976)), which is bounded between zero and one, between the estimated **S** and

the actual value of **S**. An RV-coefficient close to one indicates that the estimate is close to the parameter **S** used in simulations.

In Figure 2(a) we illustrate the estimates of $\mathbf{v}_i$, $i = 1, \ldots, 16$ in one simulation. The posterior means of $\mathbf{v}_i$ for the first 16 species are, in general, close to the corresponding values of the simulation scenario. One exception is species 16, whose average relative abundance is the lowest (8.1×10^{-5}) among the first 16 species. In the left panel of Figure 2(b), we illustrate the posterior mean of **S** between individuals in one simulation. The estimate is close to the actual value of **S** with an RV coefficient between them equal to 0.98, although the estimate indicates weak correlation between two independent subgroups (subject 1–25 and subject 26–50).

When the model is misspecified (see equation (4.1)), the estimates of **v** are not comparable to the corresponding values of the simulation scenarios. However, this result does not discourage the application of model (2.2) when estimating effects of covariates on microbial compositions. The model can still capture the derivatives and population trends (see Section 4.2) which directly describe the covariates' effects. The estimate of **S**, on the other hand,
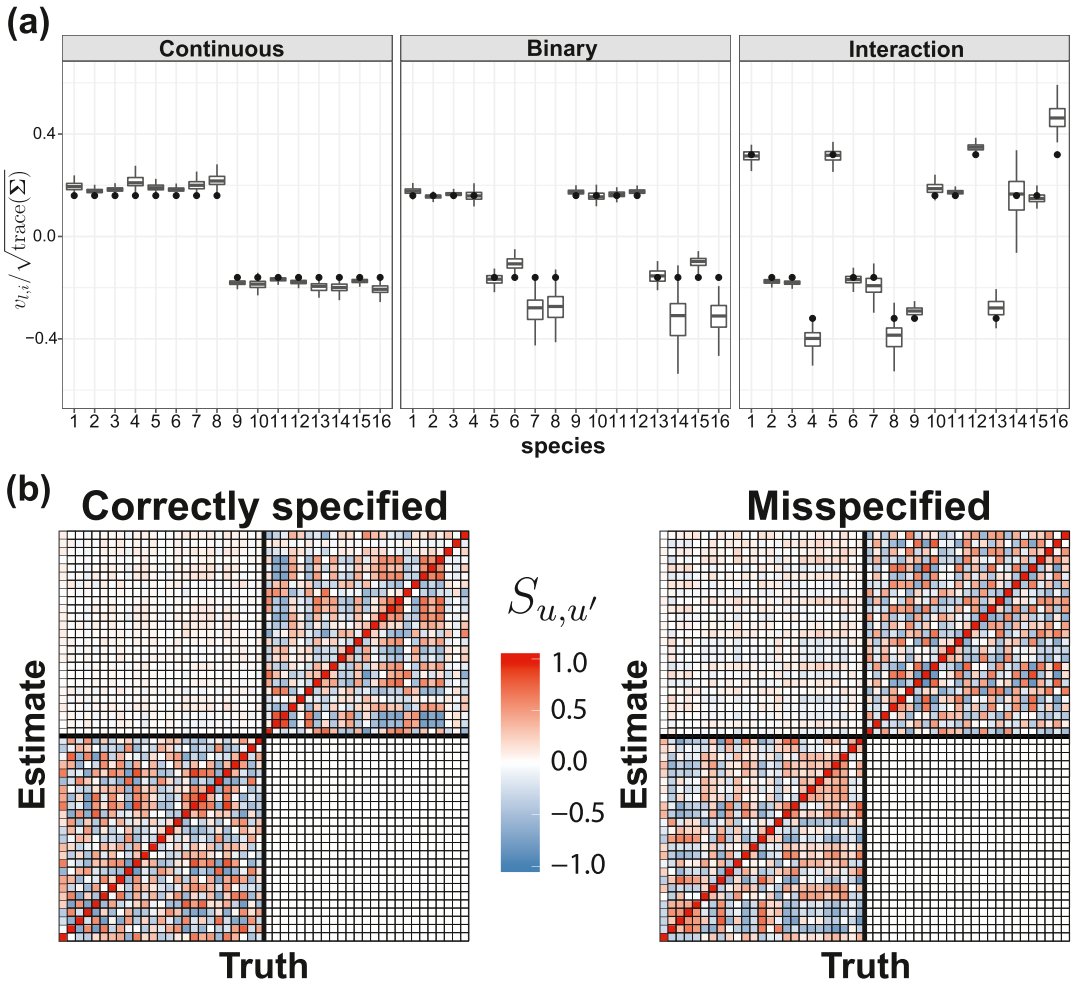

FIG. 2. *Estimates of* $\mathbf{v}_i$, $i = 1, \ldots, 16$, *and* **S** *between individuals.* (a) *Posterior distributions of* $\mathbf{v}_i / \sqrt{\text{trace}(\mathbf{\Sigma})}$, $i = 1, \ldots, 16$. *The posterior distributions are visualized by boxplots. The corresponding values of* $\mathbf{v}_i / \sqrt{\text{trace}(\mathbf{\Sigma})}$ *used for data simulation are indicated by dots.* (b) *Posterior mean of the correlation matrix* **S** *between individuals (values above the main diagonal) compared to the truth (values below the main diagonal) in one simulation when the model is correctly specified* (Left) *and misspecified* (Right).

is only minimally affected by model misspecification and preserves its closeness to the actual value of **S** (Figure 2(b), right panel). The RV coefficient between the estimate and the actual value of **S** is 0.96 in this case.

We then repeat the simulation for 50 times under the correctly specified model as well as the misspecified model to verify the observed accuracy levels. We fix across all simulation replicates the values of $\boldsymbol{\sigma}$. When the model is correctly specified, the mean squared errors (MSEs) between the posterior means of rescaled $\mathbf{v}_i$ coefficients and the values of the simulation scenario across the 50 simulation replicates are comparable across species. The smallest average MSE across 50 replicates is 4.3×10^{-6} for species 18 (the standard deviation of the estimate is 5.8×10^{-6}) and the largest average MSE is 5.2×10^{-3} for species 14 (the standard deviation of the estimate is 2.3×10^{-2}). The RV coefficients between the posterior means of **S** and the actual value of **S** across 50 replicates are close to one, whether the model is correctly specified or not. When the model is correctly specified, the mean and the standard deviation of the RV-coefficients are 0.964 and 0.009. When the model is misspecified, the mean and the standard deviation are 0.960 and 0.012. We diagnosed the mixing of the MCMC sampler for our model with $\hat{R}$ statistics (Brooks and Gelman (1998)). The $\hat{R}$ statistics indicate that, when the model is correctly specified, mixing is reached for rescaled parameter $v_{l,j}$ and eigenvalues of **S** after 60,000 iteration. See Section S2 of the Supplementary Material (Ren et al. (2020)) for details.

The Bayesian model can be embedded into a permutation procedure to detect whether a covariate $w_{l,j}$ is associated with the microbial composition or not. The null and alternative hypotheses that we consider are $H_0 : \mathbf{v}_l = \mathbf{0}_I$ vs. $H_A : \mathbf{v}_l \neq \mathbf{0}_I$, where $\mathbf{0}_I$ is a vector of zeros. We permute covariate values $w_{l,j}$ across samples and estimate, under H_0, the distribution of $\|\hat{\mathbf{v}}_l\|$, where $\hat{\mathbf{v}}_l$ is the posterior mean of $\mathbf{v}_l$. Permutation is one possible approach to estimate the $\|\hat{\mathbf{v}}_l\|$ distribution under H_0, which is applicable if covariates are independent or nearly independent. We finally compare the actual $\|\hat{\mathbf{v}}_l\|$ value with the estimated distribution (see Section S3 of the Supplementary Material (Ren et al. (2020))) for an example). One could apply other approaches to generate in silico datasets under H_0. For example, the parametric bootstrap can replace the observed $w_{l,j}$ values with samples from estimates of the conditional distributions $p(w_{l,j}|w_{-l,j})$, where $w_{-l,j}$ indicates the values of all covariates except $w_{l,j}$.

4.2. *Visualizing the relationship between covariates and microbial compositions.* As we mentioned in Section 2.2, the values of **v** do not directly express the sign and the magnitude of the covariates' effects on microbial compositions. Recall, for example, that a positive $v_{l,i}$ might correspond to a decreasing trend with respect to the covariate $w_{l,j}$. This can happen when the $v_{l,i'}$ of another species i' is larger than $v_{l,i}$. The goal of this subsection is to evaluate if we can estimate responses of species abundances to variations of covariates of interest.

We consider the visualization approaches described in Section 3.2. We first focus on the estimates of the derivatives $\partial P^j(\{Z_i\})/\partial w_{l,j}$. These provide, for each individual sample, estimates of the variation in microbial abundance resulting from an infinitesimal increment of a specific covariate $w_{l,j}$, while the other covariates remain fixed. The results for three representative species are summarized in the top panels of Figure 3. The X-axes indicate the value of $w_{l,j}$ and the Y-axes the value of $\partial P^j(\{Z_i\})/\partial w_{l,j}$. Each solid curve in these figures is generated by computing the posterior means of $\partial P^j(\{Z_i\})/\partial w_{l,j}$, for each sample j, which then become the input of a LOWESS algorithm. We also calculate the actual values of $\partial P^j(\{Z_i\})/\partial w_{l,j}$ using the $\boldsymbol{\sigma}$, **X**, **Y** and **v** parameters that generated the data. The actual values of the partial derivatives are visualized with dash lines. In Section S4 of the Supplementary Material (Ren et al. (2020)), we also plot the posterior mean of $\partial P^j(\{Z_i\})/\partial w_{l,j}$ vs. the actual value of $\partial P^j(\{Z_i\})/\partial w_{l,j}$ for each sample j along with the 95% credible intervals for $\partial P^j(\{Z_i\})/\partial w_{l,j}$.
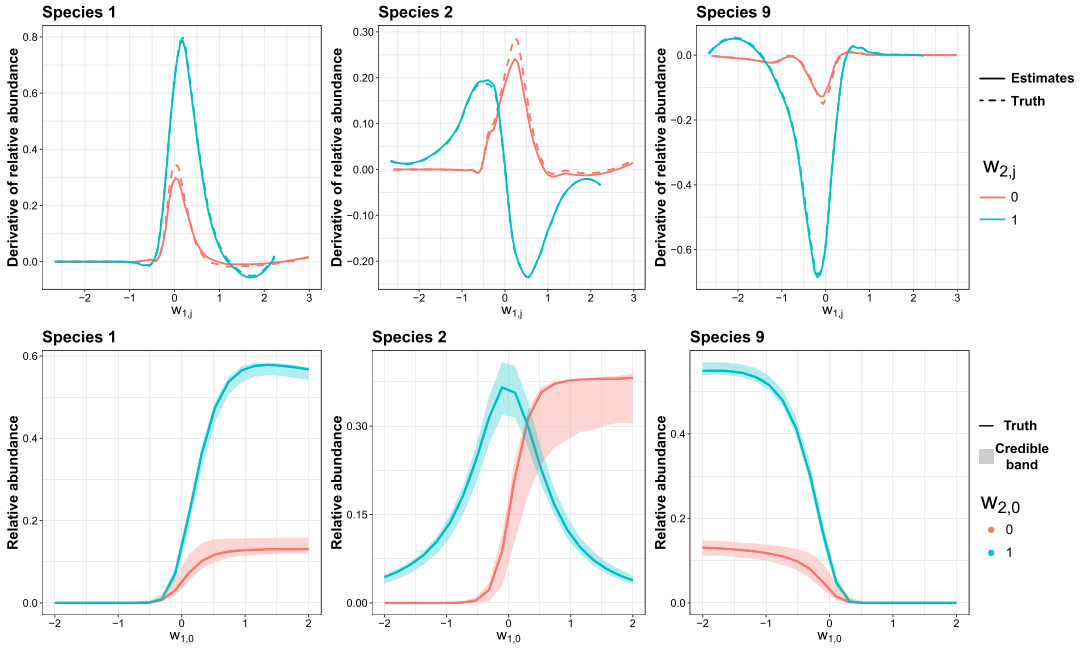
FIG. 3. *Posterior estimates of individual-level and population-level relationship between covariate $l = 1$ and relative abundances when the model is correctly specified.* (Left) *Increasing trend for the group $w_{2,j} = 0$ and the group $w_{2,j} = 1$.* (Middle) *Increasing trend for the group $w_{2,j} = 0$ and nonmonotone trend for the group $w_{2,j} = 1$.* (Right) *Decreasing trend for the group $w_{2,j} = 0$ and the group $w_{2,j} = 1$. Each curve in the* top *panels is generated by computing the individual posterior estimates of $\partial P^j(\{Z_i\})/\partial w_{l,j}$, for each sample j, which then become the input of a LOWESS procedure. The* bottom *panels illustrate the posterior distribution of the population trends.*

We then focus on the population level estimates of covariates' effects by visualizing the population trend of species i with respect to a given covariate (see Section 3.2). Population trends of three representative species with respect to $w_{1,0}$ at different values of $w_{2,0}$ are summarized in Figure 3's bottom panels. The X-axes indicate the value of $w_{1,0}$ and the Y-axes the population average abundance $\bar{P}(\{Z_i\}; \mathbf{w}_0)$. The shaded areas indicate the pointwise 95% credible bands of population trends.

When the model is misspecified, the comparisons of estimated derivatives and population trends to the truth are shown in Figure 4. To compute the actual derivatives and population trends, we use the specification of $P^j(\{Z_i\})$ in (4.1). From the top panels of Figure 4, we observe that the estimates of the derivatives capture the sign of the actual values. However, the estimates are not as close to the actual values of the derivatives, as in the case where the model is correctly specified. This result is expected as we erroneously assume that $P^j(\{Z_i\})$ depends on $(Q_{i,j})_+^2$ instead of $(Q_{i,j})_+$. Bottom panels of Figure 4 illustrate that the estimated population trends follow the actual trends, but the posterior credible bands do not cover the truth, as in the previous example where the model is correctly specified.

We repeat the simulation for 50 times under the correctly specified model as well as under the misspecified model. For each species i, we use MSE between the posterior mean of $(P^j(\{Z_i\}); j \leq J)$ derivatives and the corresponding values of our simulation model. In the Supplementary Material, Figure S5.1's top panel (Ren et al. (2020)), we plot the distributions of MSEs across simulation replicates. This figure confirms the results in Figure 3 and Figure 4. The estimates of derivatives in the correctly specified model are closer to the truth compared to the estimates with the misspecified model. For both correctly specified and misspecified models, the mean MSE across replicates reaches its maximum for species 9, with
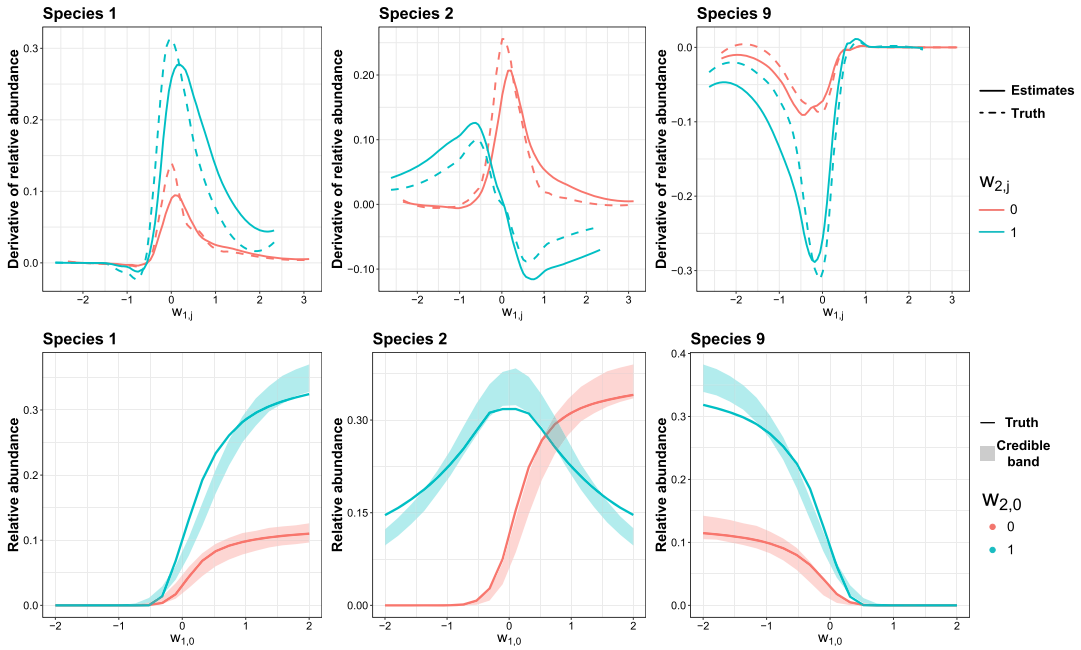
FIG. 4.   *Posterior estimates of individual-level and population-level relationship between covariate $l = 1$ and relative abundances when the model is misspecified.*

value 8.7×10^{-4} under the correctly specified model and 3.8×10^{-2} under the misspecified model.

We then consider the estimates of population trends in the 50 replicates. In the Supplementary Material, Figure S5.1's middle and bottom panels (Ren et al. (2020)), we illustrate the estimated population trends in three species when the model is correctly specified and misspecified. When the model is misspecified, the overall shape of each band still mirrors the actual trend, but the confidence band does not cover the actual trend in a few intervals of $w_{1,0}$.

4.3. *The logistic-normal model.*   We conclude the simulation study with a comparison of our model (referred to as DirFactor) to MIMIX (Grantham et al. (2019)), a logistic-normal model with latent factors. MIMIX employs a low-dimensional latent structure that is shared by both the fixed effects and the random effects to highlight the relationships between microbial species. The major difference between MIMIX and our model lies in the distribution assumption for P^j. In MIMIX the distribution of P^j follows a logistic-normal distribution

$$(4.2) \qquad P^j(\{Z_i\}) = \frac{\exp(Q_{i,j})}{\sum_{i'} \exp(Q_{i',j})}.$$

A characteristic of this specification is that the relative abundances of species are strictly positive and not tailored to zero-inflated microbiome data. By contrast, our specification of P^j assigns nonzero mass to zero which means that our model allows for explicit zero-inflation. In this subsection we are interested in comparing the estimation performance of our model to that of MIMIX.

We focus on the accuracy of the estimated population trends for the continuous covariates $w_{1,j}$. The accuracy is evaluated in two aspects: root mean-squared errors (RMSE) of the estimated population trends and coverages of the estimated credible bands of the population trends. The first metric is a universal summary of the bias and variance of the estimates, while

the second metric is used to evaluate the reported uncertainty on the estimates. We generate two sets of simulation datasets. The first set of data is generated using the link function (2.1) in our model whereas the second set uses the link function (4.2) in MIMIX. The specifications of $\mathbf{v}$, $\mathbf{w}$, $\mathbf{X}$, $\mathbf{Y}$ are the same for both sets and are described at the beginning of Section 4.

For each simulated dataset we impose additional zero inflation via hard truncation of P^j at 10^{-3} and 10^{-2}. The larger the threshold the higher the degree of zero inflation introduced in the simulated dataset. We also examine the effect of overdispersion. Specifically, for fixed $\mathbf{v}$, $\mathbf{w}$, $\mathbf{X}$, $\mathbf{Y}$, we generate three datasets based on them with $\mathrm{var}(\epsilon_{i,j}) = 1$, $\mathrm{var}(\epsilon_{i,j}) = 5$ and $\mathrm{var}(\epsilon_{i,j}) = 10$ to represent low overdispersion, medium overdispersion and high overdispersion. We finally consider the effect of overdispersion in the distribution of read depths n_j. Once relative abundances $(P^j(\{Z_i\}); j \leq J, i \leq I)$ are simulated, we generate the OTU counts with three different distributions of n_j: a Poisson distribution with mean 10^5, a negative binomial distribution with mean 10^5 and variance 10^9 (moderate overdispersion) and a negative binomial distribution with mean 10^5 and variance 4×10^{10} (large overdispersion).

We use a B-spline basis of $w_{1,j}$ both when we produce inference based on our model or MIMIX in the simulation study (i.e., we don't directly incorporate the $w_{1,j}$ values within the models). This adds flexibility in the relation between covariates and microbial compositions. We recommend the use of splines or other transformations when the number of covariates is considerably lower than the number of samples, as in our simulation study. The B-spline basis we used is of degree three with internal knots at -1, 0 and 1 and two boundary knots at -2 and 2. We simulate 50 instances of $\mathbf{v}$, $\mathbf{w}$, $\mathbf{X}$ and $\mathbf{Y}$. For each simulation replicate of $\mathbf{v}$, $\mathbf{w}$, $\mathbf{X}$ and $\mathbf{Y}$, we generate datasets based on combinations of different link function (2.1) and (4.2), three different truncation levels, three overdispersion levels and three distributions of n_j.

In each simulation replicate we estimate the population average abundance (see Section 3.2 for its definition) of each species at 20 different values of $w_{1,0}$ equally spaced between -2 and 2. We report the average RMSE between the resulting vector of estimates and simulation scenario parameters across all species and 50 simulation replicates as well as two values of $w_{2,0}$. We also report the coverage of the 95% credible intervals of the population trends for $w_{1,0} \in (-2, 2)$ in the 50 replicates averaging across all species and two values of $w_{2,0} = 0, 1$. For n_j generated from the Poisson distribution, the RMSEs are shown in Table 3, and the coverage probabilities are included in Table 4. For the other two distributions of the n_j counts we illustrate the results in Section S6 of the Supplementary Material (Ren et al. (2020)).

From the results we find that the proposed DirFactor model shows little sensitivity to the degree of zero inflation. Setting $P^j(\{Z_i\})$ to be zero when its value is below a given threshold

TABLE 3

Average RMSE of estimated population mean abundances at 20 *different values of* $w_{1,0}$ *equally spaced between* -2 *and* 2 *across simulation replicates for our model (DirFactor) and MIMIX. The threshold parameter indicates at which value we truncate the simulated* $P^j(\{Z_i\})$*'s to zero. We consider two scenarios where the data is generated from DirFactor and MIMIX, respectively. The read depths are generated from a Poisson distribution with mean* 10^5*. All RMSEs in the table are multiplied by* 10^3

| | Simulated from DirFactor | | | | | | Simulated from MIMIX | | | | | |
| | DirFactor | | | MIMIX | | | DirFactor | | | MIMIX | | |
Threshold	0	10^{-3}	10^{-2}	0	10^{-3}	10^{-2}	0	10^{-3}	10^{-2}	0	10^{-3}	10^{-2}
$\mathrm{var}(\epsilon_{i,j}) = 1$	1.0	1.2	1.4	15.3	24.3	44.8	1.3	2.5	3.8	1.3	1.5	1.8
$\mathrm{var}(\epsilon_{i,j}) = 5$	3.5	3.5	3.4	35.4	65.9	74.1	5.4	7.9	8.8	1.4	2.9	4.1
$\mathrm{var}(\epsilon_{i,j}) = 10$	4.5	4.6	4.7	66.0	96.6	154.3	10.6	11.3	11.7	1.7	3	4.7

TABLE 4

Coverage of the posterior distribution of the population trend (defined in Section 3.2). We average across species and across $w_{1,0}$ values between -2 and 2 and $w_{2,0} = 0$ and $w_{2,0} = 1$. The threshold parameter indicates at which value we truncate the simulated $P^j(\{Z_i\})$'s to zero. The coverage is calculated using simulation replicates. The read depths are generated from a Poisson distribution with mean 10^5. We consider two scenarios where the data is generated from DirFactor and MIMIX, respectively

| | Simulated from DirFactor | | | | | | Simulated from MIMIX | | | | | |
| | DirFactor | | | MIMIX | | | DirFactor | | | MIMIX | | |
Threshold	0	10^{-3}	10^{-2}	0	10^{-3}	10^{-2}	0	10^{-3}	10^{-2}	0	10^{-3}	10^{-2}
$\text{var}(\epsilon_{i,j}) = 1$	0.99	0.97	0.95	0.92	0.86	0.80	0.95	0.89	0.89	1.00	0.93	0.89
$\text{var}(\epsilon_{i,j}) = 5$	0.97	0.97	0.95	0.94	0.87	0.80	0.96	0.91	0.90	1.00	0.95	0.90
$\text{var}(\epsilon_{i,j}) = 10$	0.94	0.91	0.90	0.94	0.86	0.77	0.90	0.88	0.83	0.94	0.90	0.84

does not affect accuracy. On the other hand, when the threshold for truncating $P^j(\{Z_i\})$ increases, the accuracy of MIMIX tends to decrease. The RMSE of MIMIX increases with this threshold parameter, regardless of the data generating models (2.1) and (4.2) and the level of overdispersion $\text{var}(\epsilon_{i,j})$. The performances in terms of coverage of the two models appear comparable even when $\text{var}(\epsilon_{i,j})$ is large. These findings are confirmed when the distribution of n_j counts is a negative binomial distribution. But prediction accuracy and coverage of the two models decrease significantly when the overdispersion of the negative binomial distribution is large (mean $= 10^5$ and variance $= 4 \times 10^{10}$). See the Supplementary Material, Tables S6.3 and S6.4, for details (Ren et al. (2020)).

We conclude this subsection with a posterior predictive procedure to evaluate and compare Bayesian models. For distinct Bayesian models we compute leave-one-out 95% posterior predictive intervals of the relative abundance ($P^j(\{Z_i\})$) of a species i in sample j using the available data, with sample j excluded. The predictive intervals are generated using Pareto smoothed importance sampling (Vehtari, Gelman and Gabry (2015), Vehtari, Gelman and Gabry (2017)). We calculate the predictive intervals for all samples and all species in the data. We then derive the proportion of samples whose observed relative abundances $n_{i,j}/n_j$ of species i are covered by the corresponding leave-one-out predictive intervals. We define the mean coverage probability of the model by averaging these proportions across species. In Section S7 of the Supplementary Material (Ren et al. (2020)), we illustrate the approach in the comparison of our Bayesian model and MIMIX (Grantham et al. (2019)). Limitations of leave-one-out cross-validation in terms of stability have been previously discussed (Kohavi (1995)); the use of the procedure in our work serves the main purpose of producing interpretable summaries that integrate our evaluations and comparisons of regression methods.

5. Microbiome analyses for type 1 diabetes in early infancy. We use the longitudinal model in Section 2.2.1 to evaluate associations between gut microbiome compositions, clinical variables and demographic characteristics of infants in the DIABIMMUNE project (Vatanen et al. (2016)). The DIABIMMUNE project collected longitudinal microbiome data in 157 infants over a period up to 1600 days after birth. Infants were enrolled from Finland, Estonia and Russia. Dietary information has been collected from each participant. The main goal of this project is to examine the relationship between type 1 diabetes (T1D) associated autoantibody seropositivity (seroconverted), which is an indicator of T1D onset, and the infants' gut microbiome. In this project, seven out of 157 infants are seroconverted.

The dataset contains a total of 55 microbial genera and 762 samples from 157 infants. A large collection of potential associations between relative abundances of microbial taxa and covariates has been previously discussed in Vatanen et al. (2016). Among these associations,

the most significant ones link nationality and age to 44 microbial genera. Due to moderate sample size, only limited evidence of variations of the microbiome profile associated with seroconversion has been reported.

We present analyses based on the proposed Bayesian model. The set of covariates is composed by nationality, age, seroconversion and the interaction between age and nationality. We want to verify consistency of our posterior inference with the results discussed in the literature. Additionally, we want to quantify the uncertainty of the estimated relationship between seroconversion and microbial compositions in human gut.

5.1. *Estimating the effects of age.* We estimated the effects of age on microbial compositions using the visualization approaches in Section 3.2. In the top panels of Figure 5, we illustrate the estimated derivatives of microbial abundances with respect to age, $\partial P^j(\{Z_i\})/\partial w_{3,j}$, for two genera, Bifidobacterium and Bacteroides. We only plot the $\partial P^j(\{Z_i\})/\partial w_{3,j}$'s for 150 randomly selected samples for visual clarity. We show the 95% credible intervals for derivatives with bars, and the sizes of points are proportional to the observed abundances.
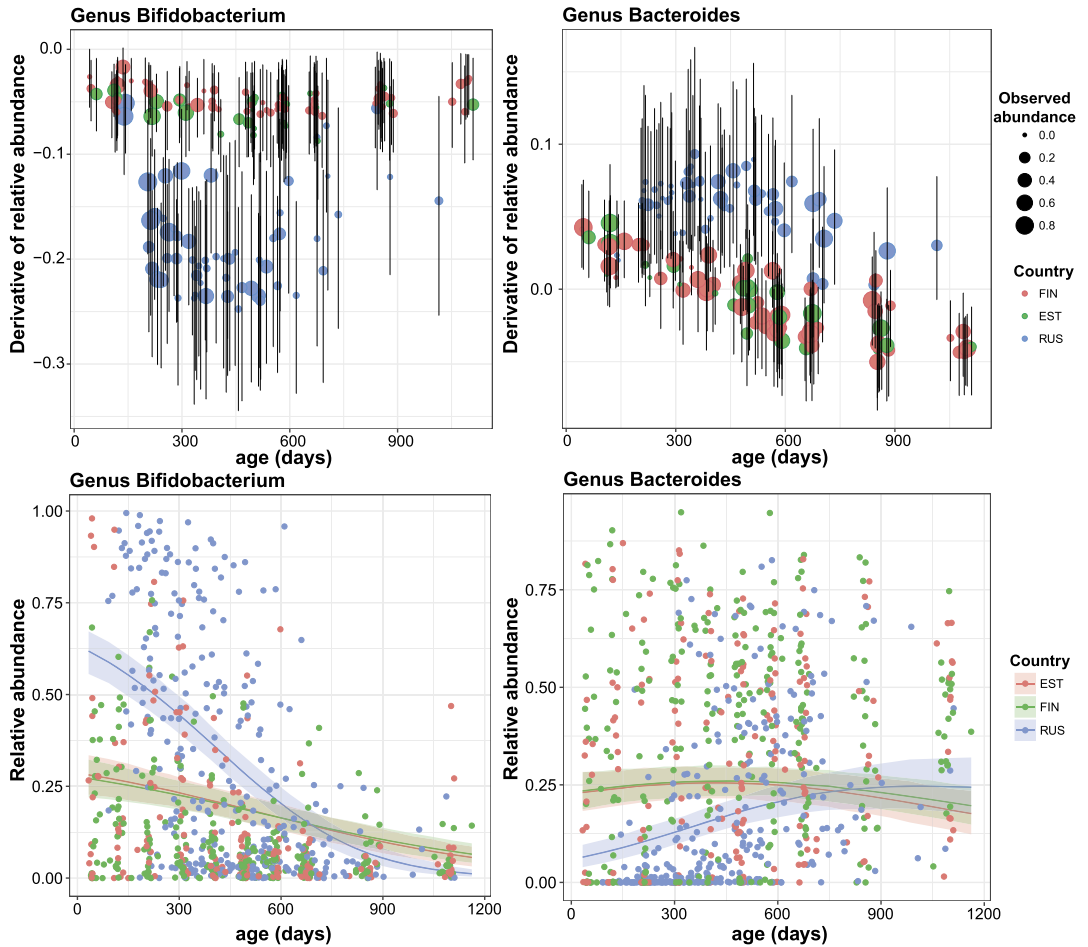

FIG. 5. (Top) *Estimated $\partial P^j(\{Z_i\})/\partial w_{3,j}$ for two genera. Each point represents a sample. Colors indicate nationalities, and the sizes of points are proportional to the observed abundances. The error bars indicate* 95% *credible intervals. We only plot* 150 *randomly selected samples.* (Bottom) *Estimated population trends of Bifidobacterium and Bacteroides for Estonian, Finnish and Russian infants. The infants are assumed to be nonseroconverted. Curves represent the estimated population trends and the shaded areas illustrate pointwise* 95% *credible bands. Points indicate the observed abundances of Bifidobacterium or Bacteroides in all samples. We use colors to indicate nationalities.*

In the bottom panels of Figure 5, we plot the estimated population trends of the same genera with respect to age. We consider the population trends for Estonian, Finnish and Russian infants and assume that the infants are not seroconverted. Posterior credible bands for the population trends are visualized by shaded areas. The observed abundances of Bifidobacterium and Bacteroides in all samples are illustrated by scatter plots together with the estimated population trends.

The estimated derivatives with respect to age for Bifidobacterium are significantly smaller than zero in most of the samples, indicating that the abundances of Bifidobacterium in infants' gut microbiome tend to decrease with age. Since bacteria from this genus are associated with breastfeeding (Fanaro et al. (2003)), this is to some extent expected. The results on derivatives are consistent with the estimated population trends. In all three populations (Estonian, Finnish and Russian), Bifidobacterium is estimated to have a decreasing population trend with respect to age. The trends for Finnish and Estonian infants are similar, while for Russian infants the decrease is faster for infants that are less than 600 days old.

The association between genus Bacteroides and age is less pronounced. The derivatives of Bacteroides tend to be positive in samples taken before 300 days. When the infants get older, the derivatives become slightly negative in Estonian and Finnish infants but remain positive in the Russian group. The population trends in this case are also consistent with the estimated derivatives. For nonseroconverted Estonian and Finnish infants the estimated population abundances of Bifidobactrium increase with age when the infants are less than 450 days old and start to decrease slowly afterward. In Russian infants the initial increasing trend is more pronounced with a narrower credible band than the other two populations until 900 days. After 900 days, the population average abundance reaches a plateau and the credible band widens.

5.2. *Estimating effects of nationalities and seroconversion.*   We make inference about the associations between the gut microbial compositions and nationalities using the differences $\Delta P^j(\{Z_i\})/\Delta w_{l,j}$ defined in (3.4). For each sample, we estimate $\Delta P^j(\{Z_i\})/\Delta w_{1,j}$, which is the difference associated to the change of nationality from Finland (FIN) to Estonia (EST), as well as $\Delta P^j(\{Z_i\})/\Delta w_{2,j}$, the difference associated to the change from Finland (FIN) to Russia (RUS). We consider the averages of $\Delta P^j(\{Z_i\})/\Delta w_{1,j}$ and $\Delta P^j(\{Z_i\})/\Delta w_{2,j}$ in each of five consecutive age groups. The posterior distributions of these population averages (Figure 6) illustrate the effect of nationality. In both panels of Figure 6, the X-axis identifies age groups, and the Y-axis indicates the value of $\Delta P^j(\{Z_i\})/\Delta w_{1,j}$ and $\Delta P^j(\{Z_i\})/\Delta w_{2,j}$. Each box plot approximates, using posterior simulations, the posterior distribution of the average $\Delta P^j(\{Z_i\})/\Delta w_{l,j}, l = 1, 2$. These averages are defined by integration within a specific age group.

There is an increase of Bifidobacterium abundance when we compare FIN to RUS nationalities. This increase diminishes with age. In the last age group (670–1160 days), the posterior distribution of $\Delta P^j(\{Z_i\})/\Delta w_{2,j}$ indicates that the abundances of Bifidobacterium in samples collected from infants older than 670 days remain comparable across nationalities. In the second comparison of nationalities, FIN to EST, only minor changes of Bifidobacterium abundance levels are observed. The abundances of Bacteroides are smaller in RUS than in FIN nationalities. This difference again diminishes with age. The difference of Bacteroides abundances between EST and FIN are also minor.

We also explored associations between microbial compositions and seroconversion status. In this case we again examine the posterior distributions of average $\Delta P^j(\{Z_i\})/\Delta w_{4,j}$ in five consecutive age groups. We do not find evidence in our analyses of any genus associated to seroconversion, due to high uncertainty of the estimated average $\Delta P^j(\{Z_i\})/\Delta w_{4,j}$ in all age groups.
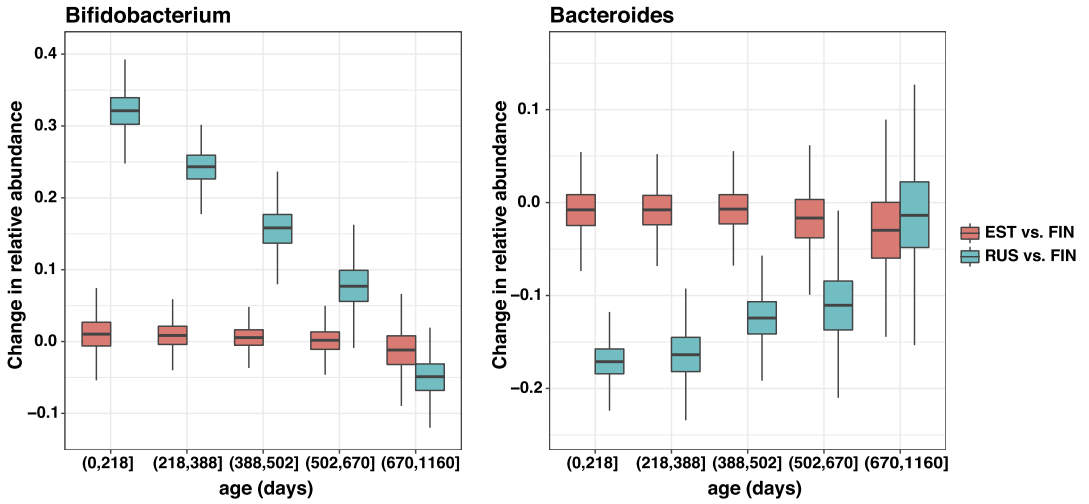
FIG. 6.   *Posterior distributions of the average difference* $\Delta P^j(\{Z_i\})/\Delta w_{1,j}$ *(red) and* $\Delta P^j(\{Z_i\})/\Delta w_{2,j}$ *(green) in five consecutive age groups. We plot the results for Bifidobacterium* (Left) *and Bacteroides* (Right).

5.3. *Similarities of microbial genera.*   In this subsection we focus on similarities between microbial genera. We first consider the simple approach where the similarity of two genera is measured by the correlation between their observed relative abundances across all samples. The result of this approach is a correlation matrix, denoted by $\mathbf{S}_{\text{raw}} = (S_{\text{raw}}(i, i'); i, i' \leq I)$, where $S_{\text{raw}}(i, i') = \text{cor}[(n_{i,j}/n_j; j \leq J), (n_{i',j}/n_j; j \leq J)]$. We then consider two approaches which utilize the proposed model. The first approach uses the cosine of the angle between $\mathbf{v}_i$ and $\mathbf{v}_{i'}$ to quantify the similarity of genera i and i', whereas the second approach uses the cosine of the angle between $\mathbf{X}_i$ and $\mathbf{X}_{i'}$. The results of these two approaches are normalized Gram matrices, denoted as $\mathbf{S}_\mathbf{v}$ and $\mathbf{S}_\mathbf{X}$, respectively. In the top panels of Figure 7, we illustrate the estimates of $\mathbf{S}_{\text{raw}}$, $\mathbf{S}_\mathbf{v}$ and $\mathbf{S}_\mathbf{X}$ by heat-maps. Each row or column of the heat map represents a specific genus, and the color of each tile represents the estimated similarity of two genera.

We then focus on examining the concordance of $\mathbf{S}_{\text{raw}}$, $\mathbf{S}_\mathbf{v}$ and $\mathbf{S}_\mathbf{X}$ to the phylogenetic relations of the observed genera. To this end, we compare the phylogenetic tree of the observed genera published in Segata et al. (2013) to the heat maps. If an estimated correlation matrix indicates clusters of genera that share similarities with the phylogenetic tree, then we conclude that the estimate is consistent with phylogenetic relations.

From the figures we can find that $\mathbf{S}_{\text{raw}}$ indicates little between-genera similarity and does not recover phylogenetic relations of the observed genera. On the other hand, both $\mathbf{S}_\mathbf{X}$ and $\mathbf{S}_\mathbf{v}$ indicate clusters of genera that are consistent with the phylogenetic tree. For instance, the cluster in the middle of the heat maps of $\mathbf{S}_\mathbf{X}$ and $\mathbf{S}_\mathbf{v}$ corresponds to 13 genera from phylum Firmicutes (Clostridium, Ruminococcus, etc). These results suggest that both $\mathbf{S}_\mathbf{X}$ and $\mathbf{S}_\mathbf{v}$ capture the phylogenetic relations of the observed genera. The ordination plot of genera based on $\mathbf{S}_\mathbf{X}$ in the bottom panel of Figure 7 further confirms this conclusion. We generate the ordination plot using the method in Ren et al. (2017) which represents each genus by a region instead of a single point. In the ordination plot we find that genera from the same cluster in $\mathbf{S}_\mathbf{X}$ or $\mathbf{S}_\mathbf{v}$ are close to each other.

We also verify quantitatively the consistency of $\mathbf{S}_\mathbf{X}$ and $\mathbf{S}_\mathbf{v}$ to the phylogenetic relations. We first calculate the pair-wise phylogenetic distance matrix of the observed genera using unweighted-Unifrac dissimilarity (Lozupone et al. (2011)). We then convert this distance matrix into a normalized Gram matrix $\mathbf{S}_{\text{unifrac}}$ by Torgerson Classical Scaling (Borg and Groenen (2005)) and compare $\mathbf{S}_{\text{unifrac}}$ to $\mathbf{S}_{\text{raw}}$, $\mathbf{S}_\mathbf{X}$ and $\mathbf{S}_\mathbf{v}$. The estimated $\mathbf{S}_\mathbf{X}$ and $\mathbf{S}_\mathbf{v}$ are both
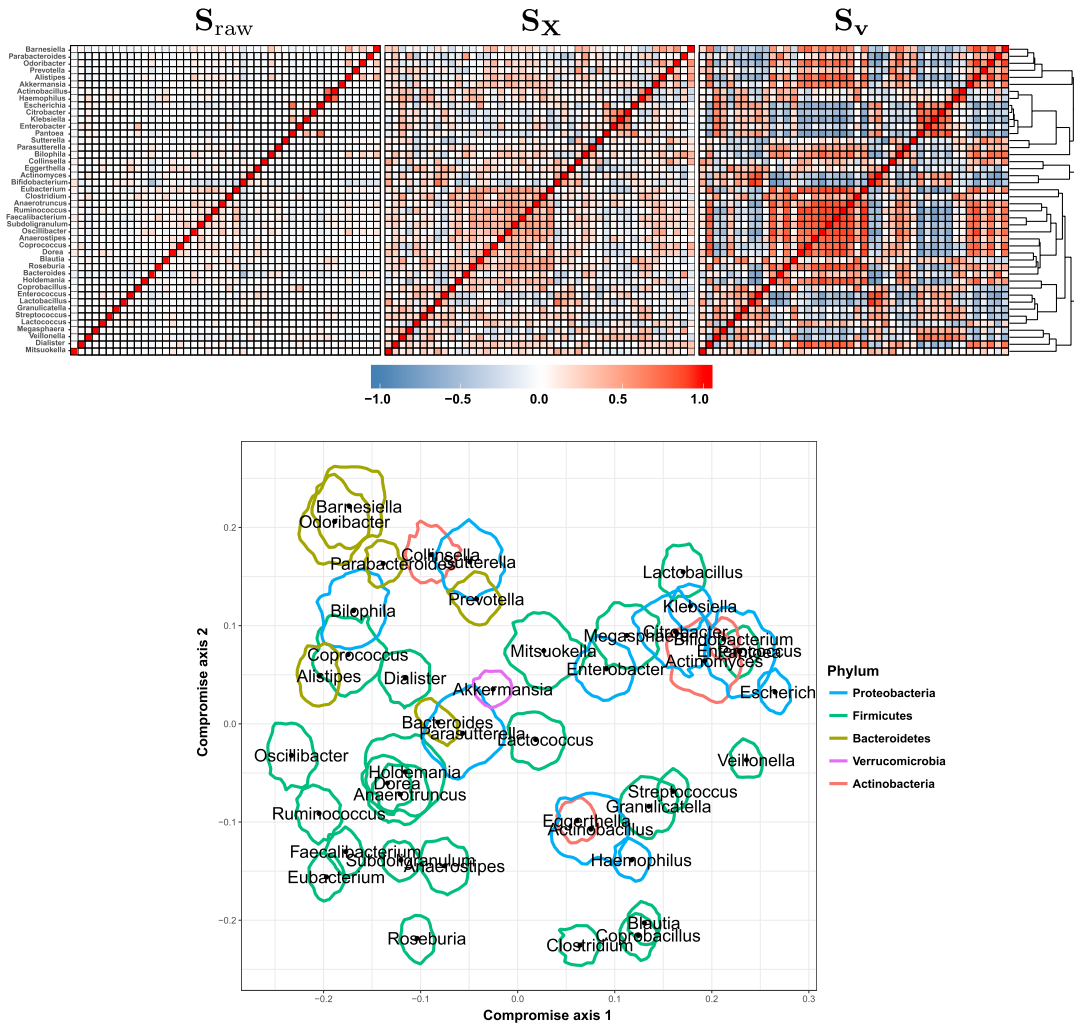
FIG. 7.   *Estimated similarities of genera.* (Top) *Estimates of* $\mathbf{S_X}$, $\mathbf{S_v}$ *and* $\mathbf{S_{raw}}$. *Each row or column in the heat maps correspond to a specific genus. The color of each entry is determined by the estimated pair-wise similarity. The rows and columns in heat maps are reordered so that adjacent rows or columns correspond to genera that are close phylogenetically. The phylogenetic tree for these genera are plotted at the right side of the figure.* (Bottom) *Ordination of genera based on* $\mathbf{S_X}$. *The contour lines indicate uncertainty regions in the ordination configuration. The contour line of a genus is colored accordingly to the phylum of the genus.*

similar to $\mathbf{S_{unifrac}}$ with RV-coefficients 0.66 and 0.76, respectively, while the RV-coefficient between $\mathbf{S_{raw}}$ and $\mathbf{S_{unifrac}}$ is 0.32.

5.4. *Goodness-of-fit of the model.*   We conducted goodness-of-fit analyses for our model based on the model evaluation approach proposed in Section 4.3; see, for example, the results shown in Section S8 of the Supplementary Material (Ren et al. (2020)). We use posterior predictive evaluations to examine whether the observed distributions of reads for the two species discussed in this section, Bacteroides and Bifidobacterium, are close to the corresponding posterior predictive distributions. We construct the leave-one-out 95% posterior predictive intervals of the relative abundances of Bacteroides and Bifidobacterium in biological sample j based on data with biological sample j excluded. We then check if the leave-one-out posterior predictive intervals cover the observed abundances of Bacteroides and Bifidobacterium. In our case the predictive intervals for 93.2% of all biological samples cover the observed

relative abundances of Bacteroides and 96.3% of all biological samples for Bifidobacterium. The high proportions of coverage for both genera indicate that there is no systematic discrepancy between the observed data and the fitted model.

6. Discussion.   We proposed a Bayesian mixed effects regression model to perform multivariate analyses for microbiome data. This regression analysis estimates the effects of covariates on microbial composition while allowing for correlations of the residuals. We illustrate that the model parameters are identifiable. This result is consistent with our simulation study. The model allows us to infer the relationship between covariates and microbial compositions with two visualization approaches. In simulations we showed that both the individual-level and the population-level relationships between covariates and microbial compositions can be accurately estimated. Moreover, our model is more robust against zero inflation than a latent factor model based on logistic-normal distribution. We finally applied the model to a longitudinal microbiome dataset and compared our results with those previously reported in the literature.

The current posterior computation is implemented with a Gibbs sampler. This can be inefficient when the number of parameters is large. The computation time increases approximately linearly with the number of samples and, similarly, with the number of microbial species. For the longitudinal microbial dataset that we analyzed, the computation time of one chain with 10^5 iterations is around 90 minutes. A possible substantial improvement in computation time can probably be obtained with Hamiltonian Monte Carlo or variational Bayes methods.

In the future we would also like to investigate appropriate variable selection techniques for the fixed effects. This is particularly helpful in settings with large collections of covariates. A more flexible model for the fixed effects is also desirable. Currently, the relationship between microbial abundances and covariates are depicted by linear functions of the samples characteristics, possibly augmented by prespecified transformations of the covariates. Finally, the current prior specification ignores potential relationship across regression vectors $\mathbf{v}_i$ associated to similar microbial species. A systematic way to incorporate such information would involve the specification of a prior distribution on $\mathbf{v}$ that mirrors the phylogeny of microbial species.

SUPPLEMENTARY MATERIAL

Source code for "Bayesian mixed effects models for zero-inflated compositions in microbiome data analysis" (DOI: 10.1214/19-AOAS1295SUPPA; .zip). R source code for replicating results in this paper and data files for the microbiome dataset.

Supplement to "Bayesian mixed effects models for zero-inflated compositions in microbiome data analysis" (DOI: 10.1214/19-AOAS1295SUPPB; .pdf). We provide the proof of the proposition for model identifiability in the general setting. We also include additional supporting plots and tables for the simulation studies and data application.

REFERENCES

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394

ANDERS, S. and HUBER, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* **11** R106. https://doi.org/10.1186/gb-2010-11-10-r106

ARBEL, J., MENGERSEN, K. and ROUSSEAU, J. (2016). Bayesian nonparametric dependent model for partially replicated data: The influence of fuel spills on species diversity. *Ann. Appl. Stat.* **10** 1496–1516. MR3553233 https://doi.org/10.1214/16-AOAS944

BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. MR2806429 https://doi.org/10.1093/biomet/asr013

BORG, I. and GROENEN, P. J. F. (2005). *Modern Multidimensional Scaling*: *Theory and Applications*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2158691

BROOKS, S. P. and GELMAN, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.* **7** 434–455. MR1665662 https://doi.org/10.2307/1390675

CHEN, J. and LI, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Stat.* **7** 418–442. MR3086425 https://doi.org/10.1214/12-AOAS592

FANARO, S., CHIERICI, R., GUERRINI, P. and VIGI, V. (2003). Intestinal microflora in early infancy: Composition and development. *Acta Pdæiatr.* **92** 48–55.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

GEVERS, D., KUGATHASAN, S., DENSON, L. A., VÁZQUEZ-BAEZA, Y., VAN TREUREN, W., REN, B., SCHWAGER, E., KNIGHTS, D., SONG, S. J. et al. (2014). The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host & Microbe* **15** 382–392. https://doi.org/10.1016/j.chom.2014.02.005

GRANTHAM, N. S., GUAN, Y., REICH, B. J., BORER, E. T. and GROSS, K. (2019). MIMIX: A Bayesian mixed-effects model for microbiome data from designed experiments. *J. Amer. Statist. Assoc.* **0** 1–16. https://doi.org/10.1080/01621459.2019.1626242

GREENBLUM, S., TURNBAUGH, P. J. and BORENSTEIN, E. (2012). Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc. Natl. Acad. Sci. USA* **109** 594–599.

GRIFFIN, J. E., KOLOSSIATIS, M. and STEEL, M. F. J. (2013). Comparing distributions by using dependent normalized random-measure mixtures. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 499–529. MR3065477 https://doi.org/10.1111/rssb.12002

HUMAN MICROBIOME PROJECT CONSORTIUM (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486** 207–214.

ISHWARAN, H. and ZAREPOUR, M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canad. J. Statist.* **30** 269–283. MR1926065 https://doi.org/10.2307/3315951

JAMES, L. F., LIJOI, A. and PRÜNSTER, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scand. J. Stat.* **36** 76–97. MR2508332 https://doi.org/10.1111/j.1467-9469.2008.00609.x

JOHNSON, D. S., REAM, R. R., TOWELL, R. G., WILLIAMS, M. T. and LEON GUERRERO, J. D. (2013). Bayesian clustering of animal abundance trends for inference and dimension reduction. *J. Agric. Biol. Environ. Stat.* **18** 299–313. MR3110895 https://doi.org/10.1007/s13253-013-0143-0

KOHAVI, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the* 14*th International Joint Conference on Artificial Intelligence*, *Vol.* 2. *IJCAI'*95 1137–1143. Morgan Kaufmann Publishers Inc., San Francisco, CA.

KOSTIC, A. D., GEVERS, D., SILJANDER, H., VATANEN, T., HYÖTYLÄINEN, T., HÄMÄLÄINEN, A.-M., PEET, A., TILLMANN, V., PÖHÖ, P. et al. (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host & Microbe* **17** 260–273. https://doi.org/10.1016/j.chom.2015.01.001

LEDOUX, M. and TALAGRAND, M. (2011). *Probability in Banach Spaces*: *Isoperimetry and Processes*. *Classics in Mathematics*. Springer, Berlin. Reprint of the 1991 edition. MR2814399

LI, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annu. Rev. Stat. Appl.* **2** 73–94.

LIJOI, A., NIPOTI, B. and PRÜNSTER, I. (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli* **20** 1260–1291. MR3217444 https://doi.org/10.3150/13-BEJ521

LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. Ser. B* **34** 1–41. MR0415861

LOZUPONE, C., LLADSER, M. E., KNIGHTS, D., STOMBAUGH, J. and KNIGHT, R. (2011). UniFrac: An effective distance metric for microbial community comparison. *ISME J.* **5** 169–172.

MacEachern, S. N. (2000). Dependent Dirichlet processes. Technical Report, Dept. Statistics, The Ohio State Univ.

Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., Reyes, J. A., Shah, S. A., LeLeiko, N. et al. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol*. **13** R79. https://doi.org/10.1186/gb-2012-13-9-r79

Müller, P., Quintana, F. and Rosner, G. L. (2011). A product partition model with regression on covariates. *J. Comput. Graph. Statist*. **20** 260–278. Supplementary material available online. MR2816548 https://doi.org/10.1198/jcgs.2011.09066

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

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F. et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464** 59–65.

Quince, C., Lundin, E. E., Andreasson, A. N., Greco, D., Rafter, J., Talley, N. J., Agreus, L., Andersson, A. F., Engstrand, L. et al. (2013). The impact of Crohn's disease genes on healthy human gut microbiota: A pilot study. *Gut* **62** 952–954.

Ren, B., Bacallado, S., Favaro, S., Holmes, S. and Trippa, L. (2017). Bayesian nonparametric ordination for the analysis of microbial communities. *J. Amer. Statist. Assoc*. **112** 1430–1442. MR3750866 https://doi.org/10.1080/01621459.2017.1288631

Ren, B., Bacallado, S., Favaro, S., Vatanen, T., Huttenhower, C. and Trippa, L. (2020). Supplement to "Bayesian mixed effects models for zero-inflated compositions in microbiome data analysis." https://doi.org/10.1214/19-AOAS1295SUPPA, https://doi.org/10.1214/19-AOAS1295SUPPB.

Robert, P. and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: The RV-coefficient. *J. R. Stat. Soc. Ser. C. Appl. Stat*. **25** 257–265. MR0440801 https://doi.org/10.2307/2347233

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** 139–140.

Rodríguez, A. and Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Anal*. **6** 145–177. MR2781811 https://doi.org/10.1214/11-BA605

Segata, N., Börnigen, D., Morgan, X. C. and Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun*. **4** 2304. https://doi.org/10.1038/ncomms3304

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101** 1566–1581. https://doi.org/10.1198/016214506000000302

Vatanen, T., Kostic, A. D., d'Hennezel, E., Siljander, H., Franzosa, E. A., Yassour, M., Kolde, R., Vlamakis, H., Arthur, T. D. et al. (2016). Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell* **165** 842–853. https://doi.org/10.1016/j.cell.2016.04.007

Vehtari, A., Gelman, A. and Gabry, J. (2015). Pareto smoothed importance sampling. arXiv preprint arXiv:1507.02646.

Vehtari, A., Gelman, A. and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput*. **27** 1413–1432. MR3647105 https://doi.org/10.1007/s11222-016-9696-4

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 Bioinform*. **18** 94. https://doi.org/10.1186/s12859-017-1516-0

Xia, F., Chen, J., Fung, W. K. and Li, H. (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics* **69** 1053–1063. MR3146800 https://doi.org/10.1111/biom.12079

Xu, L., Paterson, A. D., Turpin, W. and Xu, W. (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PLoS ONE* **10** 1–30. https://doi.org/10.1371/journal.pone.0129606