Received: 4 December 2019 | Accepted: 5 April 2021

DOI: 10.1111/rssc.12493

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Royal Statistical Society
Series C
Applied Statistics

ORIGINAL ARTICLE

A Bayesian nonparametric analysis for zero-inflated multivariate count data with application to microbiome study

Kurtis Shuler[1] | **Samuel Verbanic**[2] | **Irene A. Chen**[2] | **Juhee Lee**[3]

[1]Sandia National Laboratories in Albuquerque, Albuquerque, NM, USA

[2]Department of Chemical and Biomolecular Engineering, University of California Los Angeles, Los Angeles, CA, USA

[3]Department of Statistics, University of California Santa Cruz, Santa Cruz, CA, USA

Correspondence
Kurtis Shuler, Sandia National Laboratories, Albuquerque, New Mexico, USA.
Email: kwshule@sandia.gov

Funding information
NIH, Grant/Award Number: DP2 GM123457–01; NSF, Grant/Award Number: DMS-1662427

Abstract

High-throughput sequencing technology has enabled researchers to profile microbial communities from a variety of environments, but analysis of multivariate taxon count data remains challenging. We develop a Bayesian nonparametric (BNP) regression model with zero inflation to analyse multivariate count data from microbiome studies. A BNP approach flexibly models microbial associations with covariates, such as environmental factors and clinical characteristics. The model produces estimates for probability distributions which relate microbial diversity and differential abundance to covariates, and facilitates community comparisons beyond those provided by simple statistical tests. We compare the model to simpler models and popular alternatives in simulation studies, showing, in addition to these additional community-level insights, it yields superior parameter estimates and model fit in various settings. The model's utility is demonstrated by applying it to a chronic wound microbiome data set and a Human Microbiome Project data set, where it is used to compare microbial communities present in different environments.

KEYWORDS

Bayesian nonparametrics, dependent Dirichlet process, high-throughput sequencing, microbiome, multivariate count, normalization, operational taxonomic unit, zero inflation

1 | INTRODUCTION

The statistical community has increasingly focused on developing techniques to model high-throughput sequencing (HTS) data produced by microbiome studies. Although HTS data has been successfully used to profile complex microbial communities, analysis of such data remains challenging. In this work, we focus on the analysis of multivariate count data with excess zeros, in particular, read count data of taxa produced by 16S ribosomal RNA (rRNA) sequencing. As a motivating application, we consider the chronic wound microbiome data in Verbanic et al. (2019), which consists of microbiome samples taken from human subjects' chronic wounds, both pre- and post-debridement, as well as from their healthy skin. Verbanic et al. (2019) studied changes to the chronic wound microbiome by debridement, which is known to be an effective treatment for chronic wounds. We present a Bayesian nonparametric regression model that includes a submodel for zero inflation and flexibly accommodates covariates such as environmental factors and clinical characteristics for differential abundance analysis. The model provides an inferential framework to gain further insights into complex microbial communities.

In microbiome studies, samples are taken from some environment of interest, and the 16S rRNA gene in DNA extracts of the samples is amplified and sequenced using HTS. Counts of the resulting sequence reads are produced by comparing the reads to a database and grouping them into operational taxonomic units (OTUs) that exhibit some degree of similarity. The data from each sample are summarized in a multivariate vector of OTU counts. These counts commonly exhibit zero inflation and overdispersion, making their analysis more complicated. Standard errors will be underestimated if the model does not properly accommodate overdispersion. Failing to account for zero inflation can bias estimation of the relationships between covariates and OTU abundance, and lead to incorrect predictions. Total counts in the samples vary due to experimental artefacts such as the sequencing depth, and the raw counts do not reflect the absolute microbial abundance in the samples. Consequently, the OTU counts need to be normalized for meaningful comparison across samples, and determining whether a zero count is due to an OTU truly being absent from the environment versus a detection failure is not straightforward.

Various statistical models haven been proposed for microbiome data analysis that take these features into account. Zero-inflated count models, including zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB), are common choices to address the problem of excessive zeros. To detect associations or differential abundance, these models generally relate OTU abundance to a set of covariates by modelling the mean counts or some transformation of the counts via a link function. Some of these models, such as Chen and Li (2016) analyse each OTU individually, while many more recent models analyse OTUs jointly through some hierarchical structure. Hierarchical models allow for borrowing strength across taxa for enhanced estimation of covariate effects or increased power to detect differential abundance. In this vein, Jonsson et al. (2018) model the counts directly using a ZIP model with OTU and sample specific random effects to account for overdispersion. Lee et al. (2018) use a ZIP model with spike-and-slab priors for variable selection on regression parameters related to taxa abundances and zero inflation. This model also includes a multivariate random effect to account for interdependence among OTU counts in a sample. Paulson et al. (2013) developed a zero-inflated Gaussian mixture model, called metagenomeSeq, on log-transformed counts after adding the value of 1 to avoid numerical problems. Sohn et al. (2015) proposed a similar approach, called RAIDA, which first selects an OTU that has non-zero counts in all samples as a common divisor and uses a zero-inflated log-normal model on the ratios of OTU counts to the count of the chosen divisor. See Sankaran and Holmes (2018), Tang and Chen (2018) and Kaul et al. (2017) among many others for more examples of using zero inflated models. We also note that there are statistical models that

account for relationships across taxa using a latent factor model or a graph in modelling taxa abundances. For example, see Grantham et al. (2020), Mao et al. (2020) and Ren et al. (2020). But those models do not address a potential problem of zero inflation.

We develop a Bayesian nonparametric multivariate regression model with zero inflation that enables assessment of taxa richness and diversity that potentially varies with covariates. We use a ZINB distribution for OTU counts and assume an OTU count is either equal to zero or follows a NB distribution. The ZINB model properly accounts for the overdispersion and excess zeros that are common in microbiome data. We build nonparametric regression prior models on the probability of an OTU count being zero and the mean count of an OTU to study the effects of covariates $\boldsymbol{x}$ on microbial communities. The probit of the probability of an OTU count being zero, ξ, and the logarithm of the OTU's differential abundance compared to the baseline counts, θ, are assumed to follow unknown distribution functions indexed by $\boldsymbol{x}$, $F^{\xi}_{\boldsymbol{x}}$ and $F^{\theta}_{\boldsymbol{x}}$, respectively. We use a dependent Dirichlet process (DDP) (MacEachern, 1999, 2000), a flexible nonparametric Bayesian model to model $F^{\xi}_{\boldsymbol{x}}$ and $F^{\theta}_{\boldsymbol{x}}$. The DDP is a popular choice to model a set of random functions related through $\boldsymbol{x}$. Our model is highly flexible with regard to the nature of the relationship of the covariates and an OTU's abundance and presence. In addition to inference on the association of individual taxa with covariates through ξ and θ, $F^{\xi}_{\boldsymbol{x}}$ and $F^{\theta}_{\boldsymbol{x}}$ provide community-level insights related to alpha-diversity and species evenness, which distinguishes our method from other commonly used models for differential abundance analysis. To improve the inference on $F^{\xi}_{\boldsymbol{x}}$ and $F^{\theta}_{\boldsymbol{x}}$, we construct an elaborate model for the baseline abundance of OTUs in samples. The baseline count of an OTU in a sample is modelled as a function of a sample-specific size factor and an OTU-specific baseline abundance factor to account for count variation related to sequencing depth and different baseline abundances of OTUs. The baseline abundance factor of an OTU is shared by samples from a group, such as the subject or location where each sample was collected, to reflect the dependent taxa abundance levels shared across these samples. These two factors constitute a basis for the estimation and meaningful interpretation of ξ and θ.

In the remainder of the paper, we describe the model and its applications. Section 2 describes the proposed Bayesian nonparametric multivariate NB regression model with zero-inflation (called 'BNP-ZIMNR') and Section 3 has results from the model applied to some simulation studies. Section 4 has results from the model applied to a chronic wound microbiome data set and an additional human microbiome data set collected from NIH Human Microbiome Project, and Section 5 concludes with some discussion of the results and areas of future research.

2 | PROBABILITY MODEL

2.1 | Sampling model

Assume that non-negative integer counts Y_{ij} are observed for OTU j in sample i, $j = 1, \ldots, J$ and $i = 1, \ldots, n$, and are organized in a $n \times J$ table, $\boldsymbol{Y} = [Y_{ij}]$. Let a sample have a categorical covariate $x_i \in \mathcal{X} = \{1, \ldots, K\}$ and a grouping factor $u_i \in \mathcal{U} = \{1, \ldots, M\}$. In our motivating data set, skin type provides three levels of a covariate, that is, $\mathcal{X} = \{1, 2, 3\}$. The samples were taken from 18 subjects, which we use as a grouping factor, $\mathcal{U} = \{1, \ldots, 18\}$ with $M = 18$. Although we use a setting with one categorical covariate to present the model, it can be easily extended to accommodate more factors and continuous covariates. We use a ZINB regression model. For OTU count Y_{ij} with covariate level x_i and grouping factor u_i,

$$Y_{ij} \,|\, \epsilon_{j,x_i},\, \mu_{ij},\, s_j \overset{\text{indep}}{\sim} \epsilon_{j,x_i}\delta_{\{0\}}(Y_{ij}) + \left(1 - \epsilon_{j,x_i}\right) \text{NB}\left(\mu_{ij}(x_i, u_i),\, s_j\right), \tag{1}$$

where $\delta_A(\cdot)$ is the Dirac measure at A and NB(μ, s) the negative binomial (NB) distribution with mean μ and dispersion parameter s (so the variance is $\mu + s\mu^2$). The zero-inflated model in Equation (1) assumes that abundance is conditional on the presence of an OTU. $(1 - \epsilon_{j,x_i})$ is the probability of presence for OTU j in sample i, and is a function of covariate x_i. With probability $(1 - \epsilon_{j,x_i})$ the NB generates counts, some of which can be zero. The model specification implies that a zero count can be produced in two ways. An OTU may truly be absent in a sample with x_i. Conversely, zero counts may be produced for rare OTUs even when those OTUs are truly present if the sequencing effort is not sufficient to surface their presence. HTS data is commonly modelled using NB models, as in Equation (1), which are more flexible in accommodating overdispersion than their single-parameter Poisson counterparts, distributions for which the mean must be equal to the variance. Overdispersion parameter s_j controls the amount of overdispersion, with larger s_j indicating a greater amount of overdispersion, and the equivalent Poisson model with mean μ_{ij} is recovered as $s_j \to 0$. We let the overdispersion parameters $s_j \overset{\text{iid}}{\sim} \text{Log-Normal}(a_s, b_s^2)$ with a_s and b_s^2 fixed. The mixture model in Equation (1) can be represented with latent indicator variables $\delta_{ij} \in \{0, 1\}$ for presence and absence of OTU j in sample i. We assume $\delta_{ij} \overset{\text{indep}}{\sim} \text{Ber}(1 - \epsilon_{j,x_i})$, and let $Y_{ij} = 0$ for $\delta_{ij} = 0$ and $Y_{ij} \overset{\text{indep}}{\sim} \text{NB}\left(\mu_{ij}(x_i, u_i), s_j\right)$ for $\delta_{ij} = 1$.

We decompose the mean abundance μ_{ij} for OTU j present in sample i as follows: For sample with $x_i = k$ and $u_i = m$,

$$\log(\mu_{ij}(k, m)) = \alpha_{jm} + r_i + \theta_{jk}. \tag{2}$$

A baseline abundance factor of OTU j for samples from group m, α_{jm} accounts for different baseline abundances of OTUs. It is shared by the samples from group $u_i = m$ and induces dependence among Y_{ij} with $u_i = m$. r_i is a sample specific normalization factor to account for different library sizes across samples. Parameters α_{jm} and r_i together form the baseline count of OTU j in sample i. It is common that r_i is set to the logarithm of the total counts $Y_{i\bullet} = \sum_{j=1}^{J} Y_{ij}$ as an offset variable (e.g. see Lee et al. (2018) and Zhang et al. (2017)). We instead let r_i be random, which enables full model-based inference with appropriate uncertainty quantification. θ_{jk} in Equation (2) represents a multiplicative change in abundance of OTU j for covariate level k compared to its baseline abundance. A value of θ_{jk} close to zero implies that the abundance of an OTU is close to the baseline abundance, that is, non-differentially abundant, and positive or negative values of θ_{jk} imply low or high abundance of OTU j in a sample with $x_i = k$, respectively. Comparison of θ_{jk} across k can be used to infer differential abundance of OTU j. Similarly, comparison of θ_{jk} across j provides insights on relative abundances of OTUs in a sample with level k, such as species diversity compared to the baseline.

Using regression models for ε_{jk} and θ_{jk} is common to quantify covariate effects on the occurrence of excess zeros and differential abundances. Using our motivating data set as a specific example, one may choose one k' of the levels as a reference and let $\theta_{jk'} = 0$. θ_{jk}, $k \neq k'$ is then interpreted as an effect size relative to the abundance of OTU j under the reference. A potential drawback of this approach is that θ_{jk}, $k \neq k'$ cannot be meaningfully estimated if the OTU is absent under the reference level. A common workaround to address this issue is to replace zeros with a small value, known as pseudo count, if an OTU has zeros in all samples of the reference level. However, this arbitrary modification of the data may result in biased inference. On the other hand, the decomposition of μ in Equation (2) can avoid potential biases because θ_{jk} represents differential abundance compared to the baseline abundance $r_i + \alpha_{jm}$. The baseline count of an OTU can be estimated if an OTU exists for at least one k. We let $\theta_{jk} = 0$ if an OTU is present only for one level of k so that θ_{jk} can be fully interpreted. For ε_{jk}, we use a probit link function, $\Phi^{-1}(\varepsilon_{jk}) = \xi_{jk}$, where $\Phi^{-1}(\cdot)$ is a inverse cumulative distribution function of the standard normal distribution. In the presence of a high proportion of zeros, differentiating the event $\delta_{ij} = 0$ from the event $\delta_{ij} = 1$ for the cases of $Y_{ij} = 0$ is challenging. Specifically, more than 65%

of the OTU counts are equal to zero in two conditions for our application in Section 4. As discussed in Agarwal et al. (2002), in such cases including random group effects for ε may result in unstable model fitting and computational intractability. For this reason, we let ε_{jk} be a function of x_i only. The dependence of ε_{jk} on x_i only is in contrast with μ_{ij}, which depends on both u_i and x_i. If non-zero counts are observed for most of Y_{ij} or enough samples are obtained from each group, group-specific random effects could be included in the model for ε similar to the approach in Jonsson et al. (2018) to account for potential heterogeneity between groups. Simulation studies in Section 3 show that the proposed model without group random effects for ε performs reasonably well even when there is mild between-group heterogeneity in ε or the zero inflation levels are not very high. In the following, we consider a flexible BNP approach to model ξ_{jk} and θ_{jk} to improve inference on presence/absence and differential abundance.

2.2 | Prior

We assume $\xi_{jk} \overset{\text{iid}}{\sim} F_k^{\xi}$ and $\theta_{jk} \overset{\text{iid}}{\sim} F_k^{\theta}$, and use a BNP approach to build a model for F_k^{ξ} and F_k^{θ}. In addition to inference on individual OTUs through ξ_{jk} and θ_{jk}, their distributions F_k^{ξ} and F_k^{θ} capture useful information relating microbial communities with different levels of the covariate, and provide biological insights into community changes in k. In particular, F_x^{ξ} describes the distribution of the probabilities of OTUs in a community under condition x, and is closely related to species richness (number of different species in a community). For F_k^{ξ} that assigns more probability mass to small values, OTUs in a sample with $x_i = k$ are more likely to be present and have non-zero counts, potentially implying higher microbial species richness for the sample. Similarly, F_k^{θ} captures the distribution of differential abundance of OTUs present in a sample with $x_i = k$. If F_k^{θ} is greatly concentrated around zero, many OTUs in a sample with $x_i = k$ are not differentially abundant compared to their baseline counts. Comparison of F_k^{ξ} and F_k^{θ} across k tells how community composition changes by covariates. To build flexible prior models for F_k^{ξ} and F_k^{θ} that are possibly related across different k, we consider a dependent Dirichlet process (DDP) model in a Dirichlet process (DP) mixture model. For OTU j in a sample with $x_i = k$, we assume

$$\xi_{jk} \overset{\text{iid}}{\sim} F_k^{\xi} = \sum_{\ell=1}^{\infty} \psi_{\ell}^{\xi} \, \text{N}\left(\xi_{k\ell}^{\star}, \, \sigma_{\xi k}^2\right) \quad \text{and} \quad \theta_{jk} \overset{\text{iid}}{\sim} F_k^{\theta} = \sum_{\ell=1}^{\infty} \psi_{\ell}^{\theta} \, \text{N}\left(\theta_{k\ell}^{\star}, \, \sigma_{\theta k}^2\right). \tag{3}$$

The mixture locations $\xi_{k\ell}^{\star}$ and $\theta_{k\ell}^{\star}$ depend on k and we let $\xi_{k\ell}^{\star} \overset{\text{iid}}{\sim} \text{N}(\bar{\xi}^{\star}, \, \tau_{\xi}^2)$ and $\theta_{k\ell}^{\star} \overset{\text{iid}}{\sim} \text{N}(\bar{\theta}^{\star}, \, \tau_{\theta}^2)$. The covariate-independent weights ψ_{ℓ}^{χ}, $\chi \in \{\theta, \, \xi\}$ take the form $\psi_{\ell}^{\chi} = v_{\ell}^{\chi} \prod_{\ell'=1}^{\ell-1} \left(1 - v_{\ell'}^{\chi}\right)$ with $v_{\ell}^{\chi} \overset{\text{iid}}{\sim} \text{Be}(1, \, \rho^{\chi})$. That is, the 'single-p' DDPs that assume predictor-independent weights are used in Equation (3) as priors over the distributions of the mixture locations. MacEachern (1999, 2000) proposed the DDP to model related random probability distributions. When flexible point mass processes are considered for $\boldsymbol{\theta}_{\ell}^{\star} = \{\theta_{x\ell}^{\star}, \, x \in \mathcal{X}\}$ and $\boldsymbol{\xi}_{\ell}^{\star} = \{\xi_{x\ell}, \, x \in \mathcal{X}\}$, the 'single-p' DDP has full weak support, implying that the prior model is flexible enough to generate sample paths sufficiently close to any probability distribution. DDP and its variations have been successfully used to model related probability distributions in many applications including ANOVA (De Iorio et al., 2004), survival (De Iorio et al., 2009; Jara et al., 2010), time series analysis (Griffin & Steel, 2011; Nieto-Barajas et al., 2012) and spatial modelling (Gelfand et al., 2005) among many others. The DDP mixture formulation in Equation (3) allows us to flexibly specify and, after fitting the model, analyse and compare, F_x^{θ} and F_x^{ξ} without restrictive parametric assumptions about their functional forms. We assume $\sigma_{\chi k}^2 \overset{\text{iid}}{\sim} \text{IG}(a_{\sigma}^{\chi}, \, b_{\sigma}^{\chi})$, $\chi \in \{\xi, \, \theta\}$. The model can be

further extended to accommodate additional categorical/continuous covariates; for example, if the effects of additional covariates can be reasonably assumed to be simple, the additional covariates can be included by adding a conventional regression function in Equation (2), similar to the constructions in edgeR (Robinson et al., 2010) and DESeq2 (Love et al., 2014). A similar extension can be used to accommodate additional covariates in modelling ξ. When a fully nonparametric approach is more desirable, a stochastic process such as Gaussian process prior can be placed on on $\theta_\ell^\star(\boldsymbol{x})$ and $\xi_\ell^\star(\boldsymbol{x})$ in Equation (3) as a function of $\boldsymbol{x}$. Thus, the dependence of $F_{\boldsymbol{x}}^\chi, \chi \in \{\theta, \xi\}$, is induced over a continuum of covariates and the model can capture more general relationships between $F_{\boldsymbol{x}}^\chi$ and $\boldsymbol{x}$. We refer the reader to MacEachern (1999, 2000) for details.

Parameters r_i and α_{jm} form the baseline count of OTU j in a sample with $u_i = m$, and serve as an 'overall mean'. Observe that the parameters in Equation (2) are not identifiable due to the multiplicative structure, $\mathrm{E}(Y_{ij} \mid \delta_{ij} = 0) = e^{r_i + \alpha_{jm} + \theta_{jk}}$. We place constraints on the distributions of both r_i and α_{jm} to circumvent the identifiability issue in estimating the baseline counts, $\exp(r_i + \alpha_{jm})$. More importantly, the constraints allow parameters of primary interest θ_{jk} and F_k^θ to be identified. Specifically, we use mean-constrained priors with a mixture-of-mixtures structure (Li et al., 2017) for r_i and α_{jm},

$$\begin{aligned} r_i &\overset{\text{iid}}{\sim} \sum_{\ell=1}^{L^r} \psi_\ell^r \left\{ w_\ell^r \,\mathrm{N}(\eta_\ell^r,\, u_r^2) + (1 - w_\ell^r)\,\mathrm{N}\left(\frac{\upsilon_r - w_\ell^r \eta_\ell^r}{1 - w_\ell^r},\, u_r^2\right)\right\}, \\ \alpha_{jm} &\overset{\text{iid}}{\sim} \sum_{\ell=1}^{L^\alpha} \psi_\ell^\alpha \left\{ w_\ell^\alpha \,\mathrm{N}(\eta_\ell^\alpha,\, u_\alpha^2) + (1 - w_\ell^\alpha)\,\mathrm{N}\left(\frac{\upsilon_\alpha - w_\ell^\alpha \eta_\ell^\alpha}{1 - w_\ell^\alpha},\, u_\alpha^2\right)\right\}, \end{aligned} \tag{4}$$

where $\upsilon_\chi, \chi \in \{r, \alpha\}$ are the distribution's fixed, prespecified mean constraints, and $\boldsymbol{\psi}_\ell^\chi$ and w_ℓ^χ are mixture weights with $\sum_{\ell=1}^{L^\chi} \psi_\ell^\chi = 1$ and $0 < \psi_\ell^\chi,\, w_\ell^\chi < 1$. Although mean-constrained, the mixture-of-mixture formulation provides significant flexibility, as it can accurately characterize a wide range of distributions, including multi-modal and skewed distributions. Lee and Sison-Mangus (2018) and Shuler et al. (2019) used the distributions in Equation (4) for model based normalization in similar settings, and their results indicate the baseline abundance and covariate effects can be estimated without issues related to identifiability. In contrast to using plug-in empirical estimates for normalizing factors, the flexible model-based approach can further improve estimation of ξ_{jk} and θ_{jk}, and thus enhance estimation of F_k^ξ and F_k^θ. We follow Li et al. (2017) and set $\upsilon_r = 0$, which can be interpreted as on average no scaling adjustment; although other approaches are available, such as using an empirical estimate like in Shuler et al. (2019) or setting the constraint using prior information if it is available. We use an empirical approach to set υ_α. We compute $\tilde{r}_i = \log\left(Y_{i\bullet}/Y_{\bullet\bullet}\right) - \frac{1}{N}\sum_{i'} \log\left(Y_{i'\bullet}/Y_{\bullet\bullet}\right)$ with $Y_{\bullet\bullet} = \sum_{i,j} Y_{ij}$ as mean zero empirical estimates of r_i and set $\upsilon_\alpha = \left[\sum_{i,j|Y_{ij}>0} \left\{\log(Y_{ij}) - \tilde{r}_i\right\}\right] / \left\{\sum_{i,j} 1(Y_{ij} > 0)\right\}$. Inference on θ and ε is not sensitive to specification of υ_r and υ_α (Lee & Sison-Mangus, 2018; Shuler et al., 2019). Our simulation studies and real data analyses also show robustness of inference to different specifications of υ_r and υ_α. We place a Dirichlet prior on the outer mixture weights and a beta prior on the inner mixture weights, letting $\boldsymbol{\psi}_\ell^\chi = (\psi_1^\chi, \ldots, \psi_{L^\chi}^\chi) \sim \mathrm{Dir}(\boldsymbol{a}_\psi^\chi)$ and $w_\ell^\chi \overset{\text{iid}}{\sim} \mathrm{Be}(a_w^\chi, b_w^\chi), \chi \in \{r, \alpha\}$, where $\boldsymbol{a}_\psi^\chi = (a_{\psi 1}^\chi, \ldots, a_{\psi, L^\chi}^\chi)$, a_w^χ and b_w^χ are fixed hyperparameters. We let $\eta_\ell^\chi \overset{\text{iid}}{\sim} \mathrm{N}(\upsilon_\chi, b_{\eta^\chi}^2)$ with $b_{\eta^\chi}^2$ fixed.

2.3 | Posterior computation

Let $\underline{\boldsymbol{\theta}} = [s_j, \delta_{ij}, r_i, \alpha_{jm}, \xi_{jk}, \theta_{jk}, (\chi_{k\ell}^\star, v_\ell^\chi, \sigma_{\chi k}^2, \chi \in \{\theta, \xi\}), (\psi_\ell^\chi, w_\ell^\chi, \eta_\ell^\chi, \chi \in \{r, \alpha\})]$ denote the vector of all unknown parameters. The joint posterior distribution is $\mathrm{P}(\underline{\boldsymbol{\theta}} \mid \boldsymbol{Y}, \boldsymbol{x}, \boldsymbol{u}) \propto \mathrm{P}(\boldsymbol{Y} \mid \underline{\boldsymbol{\theta}}, \boldsymbol{x}, \boldsymbol{u})\,\mathrm{P}(\underline{\boldsymbol{\theta}})$. We use standard Markov chain Monte Carlo (MCMC) methods consisting of Gibbs and Metropolis steps to draw samples from the posterior distribution. As is standard in mixture modelling we

introduce auxiliary variables to indicate the mixture components from which the parameters of interest belong. We add auxiliary variables of this type to aid in the posterior computation for r_i, α_{jm}, θ_{jk}, and ξ_{jk}. For computational convenience, when fitting the model we approximate the DDP in Equation (3) by truncating the number of mixture components of F_k^{χ} to L^{χ}, $\chi \in \{\xi, \theta\}$. The final weight $\psi_{L^{\chi}}^{\chi} = 1 - \sum_{\ell=1}^{L^{\chi}-1} \psi_{\ell}^{\chi}$ is set to ensure F_k^{ξ} is proper. With large enough L^{χ} the truncated process produces inference almost identical to that with the infinite process (Ishwaran & James, 2001; Rodriguez & Dunson, 2011). As discussed in Rodriguez and Dunson (2011) if there is discrepancy between the posterior distributions under the truncated and infinite processes, the model is typically sensitive to the choice of L^{χ}. We examined the posterior distribution of $\psi_{L^{\chi}}^{\chi}$ and the sensitivity of the model to a choice of L^{χ}. We found that the truncated process is robust to a choice of L^{χ} if L^{χ} is sufficiently large. We diagnose convergence and mixing of the described posterior MCMC simulation using trace plots and autocorrelation plots of imputed parameters. For both the upcoming simulation examples and the data analysis, we found no evidence of practical convergence problems. An R package for the model, `bnpzimnr`, is available at https://github.com/kurtis-s/bnpzimnr. Details of posterior computation are given in Supplementary Section 1.

3 | SIMULATION STUDIES

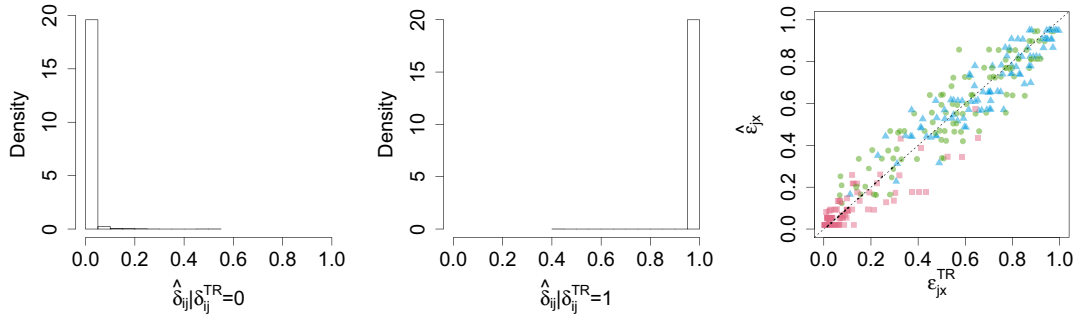
To assess the performance of the proposed model, BNP-ZIMNR, we performed simulation studies and compared its performance to alternative models. We included a factor with three levels and simulated data for 100 OTUs from 20 subjects, that is, $J = 100$, $M = 20$ and $K = 3$, resulting in $n = 60$ samples, a covariate $x_i \in \{1, 2, 3\}$, $i = 1, \ldots, N$ and a grouping factor $u_i \in \{1, \ldots, 20\}$. We used Gaussian mixtures to set the simulation truth for $F_k^{\xi,\mathrm{TR}}$ and $F_k^{\theta,\mathrm{TR}}$, $k = 1, 2, 3$; let $F_1^{\xi,\mathrm{TR}} = 0.6\,\mathrm{N}(-2, 0.25) + 0.4\,\mathrm{N}(-1, 0.5)$, $F_2^{\xi,\mathrm{TR}} = 0.2\,\mathrm{N}(-0.5, 0.25) + 0.8\,\mathrm{N}(0.5, 0.5)$ and $F_3^{\xi,\mathrm{TR}} = 0.5\,\mathrm{N}(0, 0.25) + 0.5\,\mathrm{N}(1, 0.5)$. Similarly, we set to $F_1^{\theta,\mathrm{TR}} = 0.3\,\mathrm{N}(3, 0.25) + 0.6\,\mathrm{N}(2, 0.25) + 0.1\,\mathrm{N}(-1.5, 0.5)$, $F_2^{\theta,\mathrm{TR}} = 0.3\,\mathrm{N}(2, 0.5) + 0.6\,\mathrm{N}(-1, 0.25) + 0.1\,\mathrm{N}(-2, 0.25)$ and $F_3^{\theta,\mathrm{TR}} = 0.3\,\mathrm{N}(2, 0.5) + 0.35\,\mathrm{N}(-1, 0.25) + 0.35\,\mathrm{N}(-2, 0.25)$. $F_k^{\xi,\mathrm{TR}}$ and $F_k^{\theta,\mathrm{TR}}$ are illustrated with the solid black lines in Figure 3. F_1^{ξ} generally favours smaller values of ξ_{jk}, indicating greater species richness in level 1 than in the other levels. When an OTU is present in a sample with $k = 1$, it tends to have a value of θ_{jk} greater than zero, that is, a higher abundance. On the other hand, for levels $k = 2, 3$, OTUs are likely to be absent, and when they are present, their abundances are low with large probability. In a simulated data set, the three levels of x_i approximately have 9%, 59% and 69% of Y_{ij} being equal to 0, respectively. We drew ξ_{jk}^{TR} independently from $F_k^{\xi,\mathrm{TR}}$ and generated $\delta_{ij}^{\mathrm{TR}} \overset{\mathrm{indep}}{\sim} \mathrm{Ber}(1 - \epsilon_{jk}^{\mathrm{TR}})$ for a sample with $x_i = k$, where $\epsilon_{jk}^{\mathrm{TR}} = \Phi(\xi_{jk}^{\mathrm{TR}})$. If an OTU is present for two or more levels of the factor, that is, differential abundance can be meaningfully defined, then we drew $\theta_{jk}^{\mathrm{TR}}$ from $F_k^{\theta,\mathrm{TR}}$. If an OTU is present for only one level $\theta_{jk}^{\mathrm{TR}} = 0$. Otherwise, $\theta_{jk}^{\mathrm{TR}}$ is not defined. We simulated group factors $\alpha_{j,u_i}^{\mathrm{TR}} \overset{\mathrm{iid}}{\sim} \mathrm{N}(10, 1)$, normalization factors $(\exp(r_1^{\mathrm{TR}}), \ldots, \exp(r_N^{\mathrm{TR}})) \sim \mathrm{Dir}(5, \ldots, 5)$ and dispersion parameters $s_j^{\mathrm{TR}} \overset{\mathrm{iid}}{\sim} \text{Log-Normal}\left(-2, (1/10)^2\right)$. For (i,j) with $\delta_{ij}^{\mathrm{TR}} = 1$, we simulated OTU counts Y_{ij} using the NB distribution with mean $\mu_{ij}^{\mathrm{TR}} = \exp(\alpha_{j,u_i}^{\mathrm{TR}} + r_i^{\mathrm{TR}} + \theta_{j,x_i}^{\mathrm{TR}})$ and dispersion s_j^{TR}. When $\delta_{ij}^{\mathrm{TR}} = 0$, we set $\mu_{ij}^{\mathrm{TR}} = 0$ and $Y_{ij} = 0$.

Posterior Inference. When fitting the model, we set the hyperparameters as follows: For the mean-constrained distribution of normalization factors r_i, let $\upsilon_r = 0$, $L^r = 20$, $\boldsymbol{a}_{\psi}^r = \mathbf{1}$, $a_w^r = 5$, $b_w^r = 5$, $u_r^2 = 0.05$ and $b_{\eta^r}^2 = 0.25$. Similarly, for the group-specific baseline abundance of OTU j α_{jm}, let υ_{α} be specified using the empirical approach described in Section 2.2, $L^{\alpha} = 150$, $\boldsymbol{a}_{\psi}^{\alpha} = \mathbf{1}$, $a_w^{\alpha} = 1$, $b_w^{\alpha} = 1$, $u_{\alpha}^2 = 2$ and $b_{\eta^{\alpha}}^2 = 1$. For the DDP priors, we let $\rho^{\theta} = 1$, $\overline{\theta}^{\star} = 0$ and $\tau_{\theta}^2 = 10$. For the DDP prior of ξ_{jk}, we used $\rho^{\xi} = 1$, $\overline{\xi}^{\star} = 2$ and $\tau_{\xi}^2 = 1$, which encourages a preference for a

higher probability for zero inflation, but is still flexible enough to accommodate OTUs with little sparsity. For the mixtures' kernel dispersions, we let $a_\sigma^\chi = b_\sigma^\chi = 1, \chi \in \{\xi, \theta\}$. We set the DDP truncation levels to $L^\theta = L^\xi = 50$. Finally, we used $a_s = 0.3$, $b_s^2 = 0.1$ for the prior of OTU-specific dispersion parameters s_j. To run the MCMC simulation, we used data to initialize the parameters. For example, we initialized r_i with the empirical sample size factors $\tilde{r}_i$ used to set υ_r. Empirical proportions of zero counts, $p_{jk} = \frac{1}{M}\sum_{i=1\,|\,x_i=k}^{n} 1(y_{ij} = 0)$ were used to set initial values of ε_{jk} and $\xi^\star_{k\ell}$. We ran the MCMC for 70,000 iterations, discarding the first 20,000 iterations, and thinned to use every fifth sample, resulting in 10,000 samples from the posterior distribution. On a 3.2 GHz Intel i5-6500 CPU running Ubuntu Linux the MCMC took approximately 12 mins for every 5000 iterations of the MCMC.

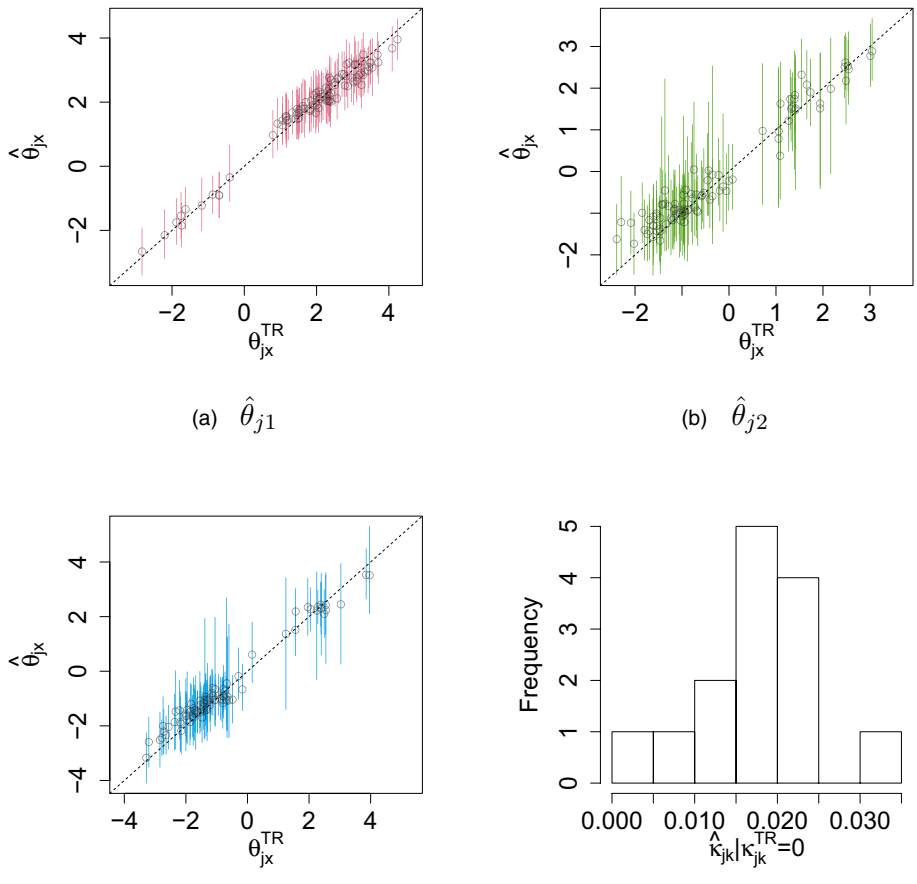
We first examine the inference on species richness in samples with k. Recall that $\delta_{ij} = 1$ implies the presence of OTU j in sample i. We used posterior means of δ_{ij} as their point estimates $\widehat{\delta}_{ij} = \widehat{\mathrm{P}}(\delta_{ij} = 1 \mid \boldsymbol{y})$. The model recovers the indicators for zero inflation well, as shown by the histograms of $\widehat{\delta}_{ij}$ when $\delta_{ij}^{\mathrm{TR}} = 0$ and 1 in Figure 1a and b, respectively. The model yields good estimates of $\epsilon_{jk}^{\mathrm{TR}}$, as seen in Figure 1c, which shows posterior estimates of ε_{jk} plotted against the simulation truth. Figure 2 shows the resulting posterior inference on θ_{jk} for individual OTUs. To account for zero inflation, we define $\kappa_{jk} = 1\{\sum_{i=1;x_i=k}^{N} 1(\delta_{ij} = 1) > 0\}$, a binary indicator taking 0 if OTU j is absent in all samples from level k, or 1 otherwise. Note that θ_{jk} is defined only when $\kappa_{jk} = 1$. We incorporate κ_{jk} and compute point posterior estimates of θ_{jk}; $\hat{\theta}_{jk} = \sum_{b=1}^{B}\kappa_{jk}^{(b)} \times \theta_{jk}^{(b)} / \sum_{b=1}^{B}\kappa_{jk}^{(b)}$, where $b = 1, \ldots, B$ indexes the posterior samples and $\kappa_{jk}^{(b)} = 1\{\sum_{i=1\,|\,x_i=k}^{N} 1(\delta_{jk}^{(b)} = 0) > 0\}$. $\widehat{\theta}_{jk}$ along with 95% credible intervals (CIs) are shown. The plots show that the model provides good estimates for differential abundance in different levels of the factor. The differences between the estimates and truth and CI lengths are greater for levels $k = 2$ and 3 because fewer non-zero counts are observed due to the high prevalence of absence. Panel (d) shows posterior estimates of $\widehat{\kappa}_{jk} = \frac{1}{B}\sum_{b=1}^{B}\kappa_{jk}^{(b)}$ when $\kappa_{jk}^{\mathrm{TR}} = 0$ in the simulation truth. The plot illustrates the model does a good job of identifying the absence in factor levels and further enhances the estimation of θ_{jk}. Figure 3 shows posterior inference for communities through $\widehat{f}_k^{\xi}$ and $\widehat{f}_k^{\theta}$. In each panel, the posterior estimates are shown by dashed coloured lines with shaded 95% pointwise CIs, and the simulation truth is shown in solid black. From the plot, the BNP regression approach flexibly captures non-Gaussian patterns such as bimodality and skewness in the distributions. Even for levels $k = 2$, 3, where many OTUs are not present, the model produces good estimates of f_k^{θ}, potentially because it borrows information across different levels through the DDP as well as across different OTUs. We also examined estimates of baseline counts of OTU j in sample i, $r_i + \alpha_{jm}$. These estimates are shown in Supplementary Figure 1. The posterior estimates recover the true baseline counts well. There is no indication that the model suffers identifiability problems.

The model is complex and we performed prior robustness diagnostics. From the diagnostics, specification of the prior for $\xi_k^\star$ may need careful attention. For a particular condition, the observed proportion of zero counts is commonly either 0 or 1. That is, $p_{jk} = \frac{1}{M}\sum_{i=1\,|\,x_i=k}^{n} 1(Y_{ij} = 0) = 0$ or $p_{jk} = 1$, meaning for every subject in that condition an OTU count $= 0$ was observed, or alternatively, for every subject an OTU count > 0 was observed. For such cases, a wide range of small/large values of ξ_{jk} can almost equally well explain the observed p_{jk}, and a large value of τ_ξ^2 may result in undesirable inference on f_k^{ξ}. We also re-fit the model with different values of the fixed parameters including L^r, L^α, L^θ and L^ξ, and examined the robustness of the model. Changes in the posterior inference by specification of other parameters such as L^r, L^α, L^θ and L^ξ are minimal. We did not observe evidence of convergence or mixing problems. In addition, the model shows robustness to the estimation of the baseline counts $r_i + \alpha_{jm}$ with different specifications of the fixed hyperparameter values. A discussion including more details of sensitivity analyses, the chain's convergence and run-time is in Supplementary Section 2.



(a) $\hat{\delta}_{ij}$ for $\delta_{ij}^{\mathrm{TR}} = 0$ (b) $\hat{\delta}_{ij}$ for $\delta_{ij}^{\mathrm{TR}} = 1$ (c) $\epsilon_{jk}^{\mathrm{TR}}$ vs $\hat{\epsilon}_{jk}$

FIGURE 1 [Simulation 1] Panels (a) and (b): Histograms of $\hat{\delta}_{ij} = \hat{P}(\delta_{ij} = 1)$ when $\delta_{ij}^{\mathrm{TR}} = 0$ and $\delta_{ij}^{\mathrm{TR}} = 1$. Panel (c): Posterior means of ϵ_{jk} plotted against the simulation truth. Colours/shapes indicate the factor levels: $k = 1$, red squares; $k = 2$, green circles; $k = 3$, blue triangles [Colour figure can be viewed at wileyonlinelibrary.com]



(a) $\hat{\theta}_{j1}$ (b) $\hat{\theta}_{j2}$ (c) $\hat{\theta}_{j3}$ (d) $\hat{\kappa}_{jk}$

FIGURE 2 [Simulation 1] Panels (a)–(c): Posterior means of differential abundances θ_{jk} for $k = 1, 2, 3$, respectively, along with 95% credible intervals and reference lines. Panel (d): Posterior estimates of κ_{jk} for cases of (j, k) with $\kappa_{jk}^{\mathrm{TR}} = 0$, that is, when OTU j is absent in all samples with level k [Colour figure can be viewed at wileyonlinelibrary.com]

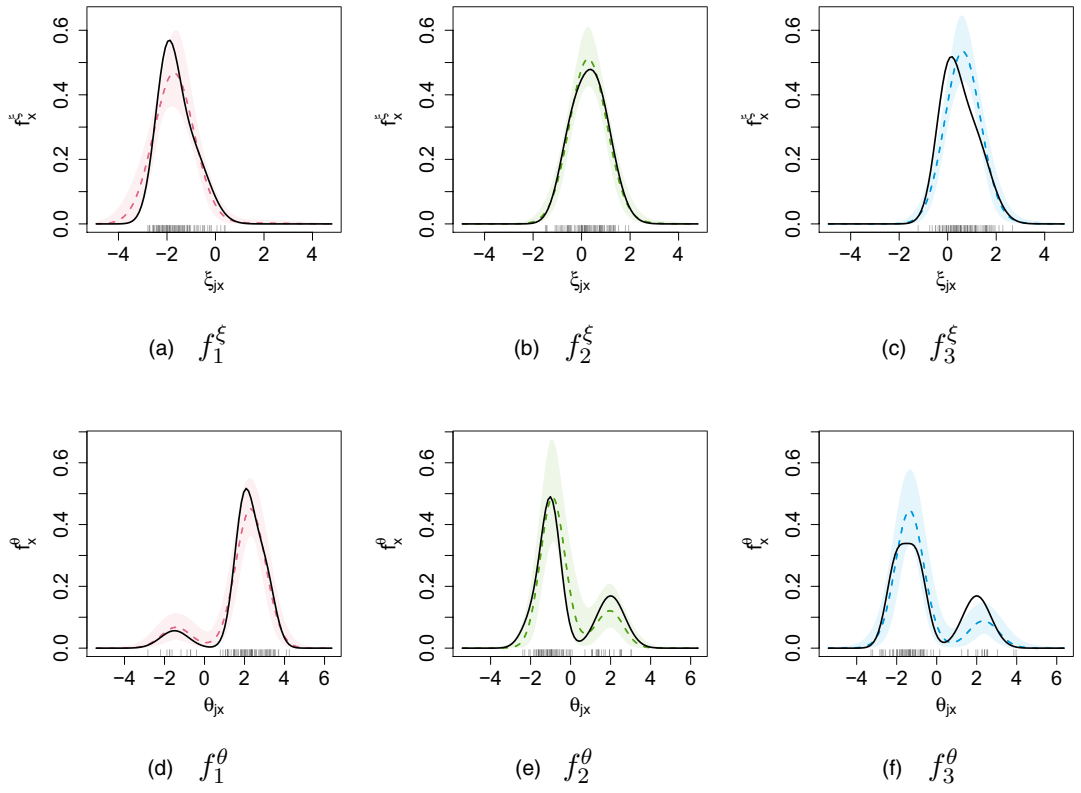


FIGURE 3 [Simulation 1] Panels (a)–(c) shows posterior estimates of f_k^ξ for each k, $k = 1, 2, 3$, and panels (d)–(f) of f_k^θ. Dashed coloured lines are estimates with shaded 95% pointwise credible intervals. Black solid lines represent the simulation truth. Rugs show ξ_{jk}^{TR} and θ_{jk}^{TR} [Colour figure can be viewed at wileyonlinelibrary.com]

Comparison. We used 100 simulated data sets to compare results of BNP-ZIMNR to those of alternative models: A Bayesian nonparametric multivariate regression model with NB (BNP-MNR), a Bayesian nonparametric multivariate regression model with fixed normalization factors (BNP-ZIMNR-FN), a Bayesian multivariate regression model (B-ZIMNR), the zero inflated overdispersed Poisson (ZoP) model (Jonsson et al., 2018), edgeR (Robinson et al., 2010), and DESeq2 (Love et al., 2014). BNP-MNR is similar to BNP-ZIMNR, but does not include the submodel for zero inflation in Equation (1). BNP-ZIMNR-FN likewise is similar to BNP-ZIMNR, but does not use the mean constrained priors for r_i as in Equation (4). Rather, BNP-ZIMNR-FN uses fixed, plug-in estimates for r_i, set to the logarithm of the total OTU counts for each sample. Unlike BNP-ZIMNR, B-ZIMNR does not utilize a Dirchlet Process Mixture (DPM) to model F_k^ξ and F_k^θ. Instead, B-ZIMNR assumes F_k^ξ and F_k^θ are single Gaussian distributions. ZoP is a Bayesian generalized linear model that uses a zero-inflated Poisson distribution for OTU counts, and beta and normal priors for the probability of being zero and the regression coefficients, respectively. Under ZoP, each Y_{ij} has a random effect, that is, sample and OTU-specific random effects to handle overdispersion. EdgeR, one of popular likelihood based methods, uses a NB generalized linear regression approach. It uses OTU-specific plugin estimates for the normalization factors produced by an empirical Bayes strategy and analyses individual OTUs separately. DESeq2 is another popular likelihood based method which models counts using a NB log-linear model. EdgeR and DESeq2 do not include random effects for the group factor and do

not account for the dependence among samples taken from the same subject. A primary difference between edgeR and DESeq2 is in the estimation of OTU-specific dispersion parameters s_j. For more details, we refer to the papers. Although edgeR and DESeq2 were originally designed for gene count data, they have been successfully adapted for amplicon data and are frequently used for microbiome analyses (McMurdie & Holmes, 2014). For this reason, we include them in our comparison. ZoP, edgeR and DESeq2 set one level of a factor as a reference level to formulate the regression, and their regression coefficients represent differential abundance compared to the abundance in the reference level. ZoP uses the pseudo count approach when all samples of the reference level have zeros. Both methods include library sizes $Y_{i\bullet}$ as plugin offsets for normalization. EdgeR has an option to use empirically pre-estimated sample size factors instead of $Y_{i\bullet}$, but we used their default option using $Y_{i\bullet}$.

For comparison, we fit each of the models and compared parameter estimates to their truth using root mean square error (RMSE). The different formulation for the regression model under ZoP, edgeR and DESeq2 precludes a direct comparison of their differential abundance estimates to $\theta_{jk}^{\mathrm{TR}}$. As an alternative, we arbitrarily set the reference to the first level $k = 1$ and compare the model performances on the estimation of differences $\theta_{jk} - \theta_{j1}$, $k = 2, 3$. By default, DESeq2 produces regression coefficient estimates for log base 2 changes in the taxa abundance; for purposes of comparison to the other models, we adjust these estimates to be on the scale of the natural logarithm. The RMSE computed for δ_{jk}, $\theta_{jk} - \theta_{j1}$ and μ_{jk} is shown in Table 1a. For BNP-MNR, we used the posterior mean estimates of μ_{ij} as a point estimate $\widehat{\mu}_{ij}$. For the zero-inflated models, similar to $\widehat{\theta}$ we computed $\widehat{\mu}_{ij} = \sum_{b=1}^{B} \delta_{ij}^{(b)} \times \mu_{ij}^{(b)}/B$. BNP-ZIMNR outperforms the other methods in comparison for estimating δ_{ij} and $(\theta_{jk} - \theta_{j1})$. BNP-ZIMNR is the best performer in terms of estimating μ_{ij}, closely followed by B-ZIMNR and ZoP. Due

TABLE 1 [Simulation 1: Comparison] RMSEs of δ_{ij}, $\theta_{jk} - \theta_{j1}$, $k = 2,3$, and μ_{ij} are shown in (a). Performance metric averages over 100 simulated data sets with standard deviations in parenthesis. $k = 1$ is used as the reference group for the difference in θ. For (b), $k = 3$ is used as the reference group and RMSE of $\theta_{jk} - \theta_{j3}$, $k = 1, 2$ is given

Model	$\boldsymbol{\delta_{ij}}$	$\boldsymbol{\theta_{j2} - \theta_{j1}}$	$\boldsymbol{\theta_{j3} - \theta_{j1}}$	$\boldsymbol{\mu_{ij}}$
(a) Parameter estimation				
BNP-ZIMNR	**0.019 (0.005)**	**0.308 (0.060)**	**0.325 (0.057)**	**3154 (818)**
BNP-MNR	–	3.909 (0.504)	4.762 (0.504)	65,190,628 (89,816,163)
BNP-ZIMNR-FN	0.021 (0.005)	2.234 (0.279)	2.386 (0.263)	4680 (2032)
B-ZIMNR	**0.019 (0.005)**	0.325 (0.066)	0.340 (0.056)	3277 (839)
ZoP	0.200 (0.033)	2.759 (0.278)	3.156 (0.249)	3769 (1282)
edgeR	–	2.218 (0.303)	2.693 (0.303)	7924 (1860)
DESeq2	–	3.157 (0.640)	4.085 (0.712)	8200 (1954)

Model	$\boldsymbol{\theta_{j1} - \theta_{j3}}$	$\boldsymbol{\theta_{j2} - \theta_{j3}}$
(b) Estimation of difference in θ with $k = 3$ as a reference		
BNP-ZIMNR	**0.325 (0.057)**	**0.393 (0.054)**
BNP-MNR	4.762 (0.504)	4.468 (0.446)
BNP-ZIMNR-FN	2.386 (0.263)	0.610 (0.182)
B-ZIMNR	0.340 (0.056)	0.418 (0.053)
ZoP	4.348 (0.356)	3.636 (0.388)
edgeR	2.693 (0.303)	3.302 (0.380)
DESeq2	4.102 (0.663)	5.184 (0.801)

The bold is used to indicate the best performing model.

to OTU and sample specific random effects under ZoP, it obtains good estimates of μ_{ij}, but may tend to overfit the data, leading to worse estimates for $(\theta_{jk} - \theta_{j1})$, as is indicated by model comparison described later. The detrimental impact of excluding zero inflation can be seen by the much larger RMSE of μ_{ij} for BNP-MNR. The comparison of BNP-ZIMNR to BNP-ZIMNR-FN and B-ZIMNR indicates the model-based normalization and the use of a nonparametric approach to modelling of F^{ξ} and F^{θ} improve inference under this simulation setting. Since selecting a level for the reference is arbitrary, we re-fit the data using a different level of the factor as the reference for ZoP, edgeR and DESeq2, and computed the RMSE of the differences in θ_{jk}. Table 1b illustrates the RMSE of $(\theta_{jk} - \theta_{j3})$ with $k = 3$ instead of $k = 1$ as the reference level. Recall that level $k = 3$ has a higher degree of zero inflation than level $k = 1$ in the truth. The performances of ZoP, edgeR and DESeq2 degrade when using this sparser factor level as the reference, indicating bias in the estimation of θ due to using arbitrary pseudo counts. In contrast, the inference on θ_{jk} under BNP-ZIMNR and its variants do not depend on the choice of reference level.

For further comparison of model fit among the Bayesian models, the log pseudo marginal likelihood (LPML) and the deviance information criterion (DIC) were calculated for the Bayesian models. These metrics are summarized in Table 2a. Similar to other information criterion, DIC assesses model performance based on the model's predictive accuracy with a penalty for model complexity (Spiegelhalter et al., 2002). Lower values of DIC are preferred. LPML is the sum of the logarithms of conditional predictive ordinates (Gelfand & Dey, 1994; Gelfand et al., 1992). It gives a measure of the leave-one-out cross validated posterior predictive probability, with higher values preferred. For more reliable comparison, we evaluated DIC and LPML based on the partially marginalized likelihood that integrates out random effects at the observation level for the ZoP (Millar, 2009). The table shows BNP-ZIMNR has improved model fit compared to the Bayesian competitors. DIC and LPML based on the partially marginalized likelihood indicate that BNP-ZIMNR fits the data better, potentially implying overfit under ZoP. Different from ZoP, edgeR and DESeq2, BNP-ZIMNR and its variants also provide community-level inferences. To assess the impact of

TABLE 2 [Simulation 1: Comparison] (a) Average model comparison metrics over 100 simulated data sets with standard deviations in parenthesis. (b) Average total variation distance of F_k^{θ} as compared to the simulation truth both with and without zero inflation. Standard deviations in parenthesis

Model	DIC	LPML
(a) DIC and LPML		
BNP-ZIMNR	**50,994 (1107)**	**−26,391 (528)**
BNP-MNR	62,909 (1317)	−32,328 (647)
BNP-ZIMNR-FN	51,780 (1098)	−26,964 (529)
B-ZIMNR	51,017 (1109)	−26,413 (535)
ZoP	2,600,574 (91,077)	−486,874 (31,409)

Model	F_1^{θ}	F_2^{θ}	F_3^{θ}
(b) Total variation distance between $F_k^{\theta,\mathrm{TR}}$ and $\widehat{F}_k^{\theta}$			
BNP-ZIMNR	**0.158 (0.063)**	**0.195 (0.073)**	**0.163 (0.060)**
BNP-MNR	0.209 (0.069)	0.489 (0.033)	0.510 (0.039)
BNP-ZIMNR-FN	0.775 (0.029)	0.269 (0.108)	0.304 (0.116)
B-ZIMNR	0.330 (0.037)	0.341 (0.005)	0.330 (0.006)

The bold is used to indicate the best performing model.

omitting zero inflation in the estimation of F_k^θ, we considered the total variation distance between $F_k^{\theta,\mathrm{TR}}$ and $\widehat{F}_k^\theta$ estimated from BNP-ZIMNR and its variants. Letting $\mathscr{B}$ denote the class of all Borel sets in $\mathbb{R}$, the total variation distance measures the closeness between two densities as $\sup_{B\in\mathscr{B}}\left|\int_B f_k^{\theta,\mathrm{TR}}\,d\theta - \int_B \widehat{f}_k^\theta\,d\theta\right| = \frac{1}{2}\int\left|f_k^{\theta,\mathrm{TR}} - \widehat{f}_k^\theta\right|\,d\theta$, where $f_k^{\theta,\mathrm{TR}}$ and $\widehat{f}_k^\theta$ denote the densities of $F_k^{\theta,\mathrm{TR}}$ and $\widehat{F}_k^\theta$ (Devroye & Lugosi, 2001). Table 2b shows the computed total variation distances. We use median estimates of f_k^θ as our point estimate $\widehat{f}_k^\theta$. The benefits of incorporating zero inflation into the model, not using fixed normalization factors, and the flexibility of the DPM over simple Gaussians are clearly observed for estimating a distribution of differential abundances. The total variation distance under BNP-ZIMNR is notably reduced, especially for $k = 2$ and 3, the levels with higher probability of OTU absence.

Additional simulations. We performed additional simulations, Simulations 2–7, to further assess the model's performance and scalability. For Simulation 2, we used a data simulation setup similar to the one for Simulation 1, but assumed a more complex structure with $K = 6$ different levels of a factor. We fit BNP-ZIMNR and the comparators to 100 simulated data sets using a specification similar to Simulation 1. In the simulation, BNP-ZIMNR outperformed the comparator models for all of the metrics that we considered. We found that the model scaled well with additional factor levels, providing accurate OTU level inference via θ_{jk}, as well as community level inference via F_k^θ and F_k^ξ. For Simulation 3 we assumed that $F_k^{\chi,\mathrm{TR}}$, $\chi \in \{\theta, \xi\}$ is a single Gaussian distribution as assumed under one of the comparators, B-ZIMNR, and kept the remaining simulation setup similar to that of Simulation 1. Although the simulation truth is closer to the assumptions made under B-ZIMNR, the results show that BNP-ZIMNR performs almost as well, and it exceeds that of B-ZIMNR for some criteria. For Simulation 4 we assumed a greater number of OTUs, $J = 500$, with $K = 3$. We assumed $M = 30$ subjects without replicates, resulting in fewer samples, $n = 30$. We further introduced between-subject heterogeneity for the zero inflation levels, which is different from the assumption under BNP-ZIMR, and assumed fewer excess zeros for conditions $k = 2$ and 3. BNP-ZIMNR performs better under most of the comparison criteria. It yields better estimates of δ_{ij} and θ_{jk}, and better predictive metrics than the other models. We find, however, that BNP-ZIMNR and its variants suffer in the estimation of μ_{ij}, possibly due to the smaller sample size with no replicates, as we show in Simulation 5, which has a similar setup to Simulation 4 but with replicates across the conditions. The results of Simulation 5 show that the estimation of μ_{ij} under BNP-ZIMNR and its variants is improved by replicates. For Simulation 6, we considered a case where two continuous covariates, $\boldsymbol{z}_n = (z_{n1}, z_{n2})$ are present in addition to the experimental conditions, $x_i \in \{1, \ldots, K\}$. Although BNP-ZIMNR can accommodate $\boldsymbol{z}_i$ through $F_{x,\boldsymbol{z}}^\chi$ in Equation (3) fully nonparametrically, we considered a linear regression similar to edgeR and DESeq2, for a simple and more comparable exercise. In particular, we let $\log(\mu_{ij}(x_i, \boldsymbol{z}_i, m)) = \alpha_{jm} + r_i + \theta_{jk} + \boldsymbol{z}_i'\boldsymbol{\beta}_j$, and placed normal priors on the regression coefficients, $\boldsymbol{\beta}_{jp}$, $\beta_{jp} \mid \tau^2_{\beta,p} \overset{\text{indep}}{\sim} \mathrm{N}(0, \tau^2_{\beta,p})$, and $\tau^2_{\beta,p} \overset{\text{iid}}{\sim} \mathrm{IG}(a^\beta_\tau, b^\beta_\tau)$, $a^\beta_\tau = b^\beta_\tau = 1$. We also considered the same extension for the other Bayesian models in comparison. Because ZoP does not allow continuous covariates, we did not include ZoP for comparison. BNP-ZIMNR performs very well for the estimation of $\boldsymbol{\beta}_{jp}$ as well as of δ_{ij} and θ_{jk}. The estimation of $\boldsymbol{\beta}_{jp}$ is significantly better under BNP-ZIMNR than under edgeR and DESeq2, potentially due to better estimation of θ_{jk} under BNP-ZIMNR. For simulation 7, we fit the model to a data set with fewer samples, using $M = 5$ subjects and $n = 10$ total samples, with $J = 50$ OTUs over $K = 2$ conditions. For most of the metrics considered, BZNP-ZIMNR outperforms the comparators under a setting with a smaller sample size. ZoP produces better estimates of μ_{ij}'s due to its sample and OTU specific random effects, but their estimates of $\theta_{j,k}$'s are very poor. The detailed results of the additional simulation studies are shown in Supplementary Section 3.

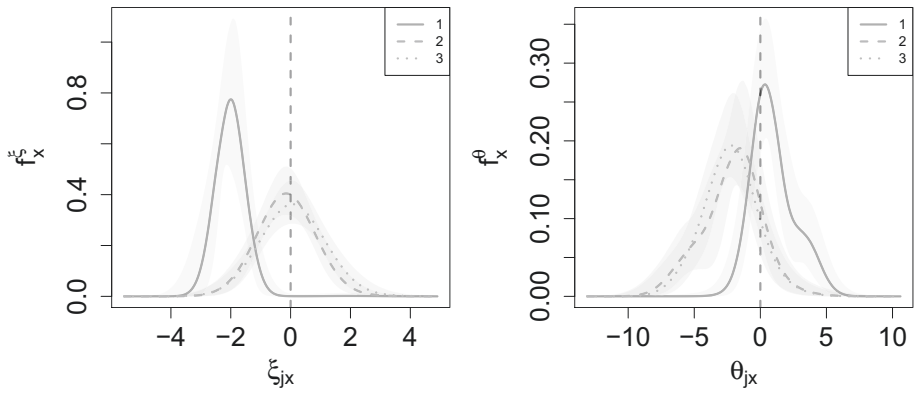
4 | CHRONIC WOUND MICROBIOME DATA ANALYSIS

In this section, we apply BNP-ZIMNR to study chronic wound microbiomes using the data set in Verbanic et al. (2019). The data set consists of microbiome samples collected from $M = 18$ subjects with chronic wounds. Swab samples were collected from chronic wounds pre- and post-debridement, along with a healthy skin swab sample from a control site, for each of the subjects. The $K = 3$ experimental conditions result in $n = 54$ samples in total. We let $k = 1$, 2, and 3 represent healthy skin, pre-debridement wound swabs, and post-debridement wound swabs, respectively. The study aims to investigate how debridement influences the composition of the microbial community of the wound, and also to compare the microbial composition of the wound surface to that of healthy skin. We analysed the data to infer changes in the community-level microbial richness and diversity as well as differential abundances of individual OTUs. Better understanding of the wound microbiome and the effects of debridement on the wound microbiota can further elucidate the role of the microbiome on wound healing. From the swab samples, the 16S rRNA gene was amplified by PCR and sequenced using high throughput sequencing, and the sequence reads were organized into an OTU table for statistical analysis. A total of 22,753 OTUs were observed after removing singletons. We restricted our attention to OTUs with nonzero counts in more than 20% of the samples for at least one experimental condition to obtain reliable inference. After pre-processing, $J = 92$ OTUs were included in the analysis. It was checked by our biological collaborators that biologically interesting OTUs were not removed from analysis. The degree of zero inflation varies widely by experimental condition, with 8% of the OTU counts equal to zero from the healthy skin samples, versus 65% and 67% of the OTU counts equal to zero in the pre-debridement and post-debridement conditions, respectively. Supplementary Figure 8a–c illustrate histograms of the empirical proportions p_{jk} of zero counts in the samples for the conditions. Panels (d)–(f) show histograms of total counts $Y_{i\bullet}$ in samples for each k. From the figures, the samples from conditions $k = 2$ and 3 have more zeros and have lower total counts. The observed zeros in the pre/post-debridement conditions may be due to the absence of the OTUs under those conditions.

We specified hyperparameters similar to those in the simulations. The MCMC simulation was run over 140,000 iterations, with the first 40,000 iterations discarded as burn-in and every fifth sample kept as thinning and used for inference. The MCMC took approximately 11 mins for every 5000 iterations of the MCMC on a 3.2 GHz Intel i5-6500 CPU running Ubuntu Linux.

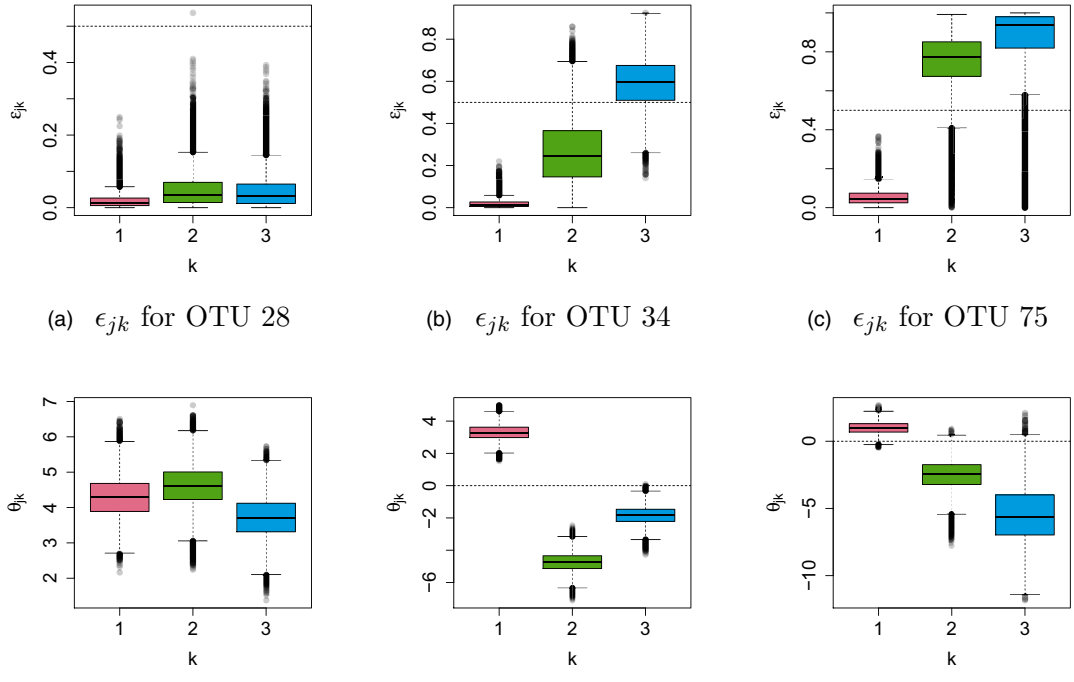
Community level inference provided by f_k^{ξ} and f_k^{θ} is shown in Figure 4. Posterior estimates of f_k^{ξ} and f_k^{θ} are shown by the coloured lines, with pointwise 95% CIs shown by the shaded regions, where the colours, red, blue and green, represents the healthy skin ($k = 1$), pre-debridement wound ($k = 2$), and post-debridement wound ($k = 3$), respectively. The differences between the estimates under the healthy skin condition and those under the wound conditions are substantial, but the wound microbial community does not change immediately after debridement, similar to the previous findings in Gardiner et al. (2017) and Verbanic et al. (2019). In panel (a), $\widehat{f}_k^{\xi}$ is stochastically lower for the healthy skin condition, suggesting greater species richness in a healthy skin sample than in a wound sample. For the wound conditions, $\widehat{f}_k^{\xi}$ assigns more density to larger values and also has higher dispersion. Panel (b) shows that $\widehat{f}_k^{\theta}$ assigns more density to higher values in the healthy skin condition than in the pre-/post-debridement conditions. The bulk of the density for the wound conditions is given to values less than zero and the density estimates have long left tails. The distributions imply that on average OTUs in the wound conditions tend to have low abundance compared to their baseline. $\widehat{f}_k^{\theta}$ are slightly skewed, but overall $\widehat{f}_k^{\xi}$ and $\widehat{f}_k^{\theta}$ do not show a substantial departure from unimodal symmetric distributions.

The model also provides inference for individual OTUs. Figure 5 illustrates the posterior distributions of ϵ_{jk} and θ_{jk} for some selected OTUs, $j = 28$, 34 and 75. From panels (b), (c), (e) and (f), OTUs 34 and 75 that belong to genus *Micrococcus* and *Corynebacterium*, respectively, are highly abundant



(a) $\hat{f}_k^{\xi}$ (b) $\hat{f}_k^{\theta}$

FIGURE 4 [Chronic Wound Data] Estimates of f_k^{ξ} and f_k^{θ} are shown in panels (a) and (b). The three experimental conditions, healthy skin ($k = 1$), pre-debridement ($k = 2$) and post-debridement ($k = 3$), are indicated by the colours red, green and blue, respectively. 95% pointwise credible intervals for each condition are shown by the shaded areas [Colour figure can be viewed at wileyonlinelibrary.com]



(a) ϵ_{jk} for OTU 28 (b) ϵ_{jk} for OTU 34 (c) ϵ_{jk} for OTU 75

(d) θ_{jk} for OTU 28 (e) θ_{jk} for OTU 34 (f) θ_{jk} for OTU 75

FIGURE 5 [Chronic Wound Data] Panels (a)–(c) illustrate the posterior distributions of ε_{jk} for each of the conditions for three selected OTUs $j = 28$, 34, 75. Panels (d)–(f) have the posterior distributions of θ_{jk}. $k = 1$, 2, and 3 denote healthy skin, pre-debridement, and post-debridement, respectively [Colour figure can be viewed at wileyonlinelibrary.com]

in skin, but not in wounds. The OTUs are absent in wounds with high probability. The increased likelihood of absence from wound samples and the depleted abundance in wound samples when present are consistent with the previous findings in Verbanic et al. (2019) and Grice et al. (2009), indicating these OTUs are associated with a healthy skin microbiome. OTU $j = 28$ belonging to genus *Pseudomonas* is noted to be significantly associated with wounds (Verbanic et al., 2019), and is also known to be a pathogen in chronic wounds (Kalan et al., 2019; Loesche et al., 2017; Wolcott et al., 2016). However, panels (a) and (d) do not show a significant association with wounds. The lack of significant differences may be due to the high variability of wound composition among patients and small sample size. We also conducted sensitivity analyses to the specification of some fixed hyperparameters, L^{θ}, L^{ξ}, L^{r}, L^{α}, υ_r and υ_α. Changes in the posterior inference was minimal under these alternative specifications. More details are discussed in Supplementary Section 4.

The comparators are applied to the chronic wound data and their inferences are compared to the posterior inference under our BNP-ZIMNR. The healthy skin condition is used as the reference group for ZoP, edgeR, and DESeq2 to infer differential abundance for individual OTUs. Supplementary Figure 11b–g compare estimates of $\theta_{jk} - \theta_{j1}$, $k = 2$ and 3, from the comparators to those from BNP-ZIMNR. For the Bayesian models, we computed DIC and LPML. Both DIC and LPML indicate BNP-ZIMNR provides a better fit to the data than BNP-MNR, BNP-ZIMNR-FN and ZoP as shown in Supplementary Table 3. It is observed that the inferences under BNP-ZIMNR and B-ZIMNR are similar as implied by $\widehat{f}_k^{\xi}$ and $\widehat{f}_k^{\theta}$ in Figure 4. DIC and LPML indicate the two models yield close model fit.

Additional Real Microbiome Analysis. For more illustration, we analysed a second real microbiome data from the Human Microbiome Project (HMB), where microbiome samples were collected from two different skin subsites for 90 subjects. We constructed four experimental conditions, one for each combination of the subsites and sex of the subjects. The number of samples from a subject varies from 1 to 4, and the data set has a total of 146 samples. Zeros are less prevalent in this data set than in the wound microbiome data. In this analysis, we find that BNP-ZIMNR sensibly characterizes differential abundance across subsites and sex, indicating potential for BNP-ZIMNR's broad applicability to microbiome studies. More details are included in Supplementary Section 5.

5 | DISCUSSION

We have presented BNP-ZIMNR, a Bayesian nonparametric regression approach to model count data in the presence of high zero inflation, with application to microbiome studies. Estimates of $F_k^{\chi}, \chi \in \{\theta, \xi\}$, which are produced from the BNP modelling approach, give a different, more nuanced look at how diversity and differential abundance are related to covariates than statistical tests alone. Our model construction for baseline abundances through mean-constrained regularizing priors removes the need to arbitrarily set a reference condition which would affect posterior inference. The simulation studies indicate BNP-ZIMNR provides better parameter estimates than popular alternatives across a range of different settings, but more importantly show the model can recover F_k^{χ} after accounting for sequencing depth, zero inflation, and baseline taxa abundance levels. Direct, community-level comparison of differential abundance and diversity is thus possible by examining F_k^{χ} across different values of k. This use was illustrated by applying BNP-ZIMNR to two real data sets, where F_k^{χ} confirmed findings from previous literature, and also provided a richer view of differential abundance and diversity in these communities.

BNP-ZIMNR may be extended to accommodate more complex data structures. In microbiome studies, where samples are taken from different geographic locations or from the same environment over time, the composition of the microbial communities is expected to change by spatial locations/time points $\boldsymbol{x}$. Parfrey and Knight (2012) and Galloway-Peña et al. (2017) studied spatial and temporal

changes in the human microbiota, the latter understanding longitudinal variability in the microbiome as 'critical' to the development and use of microbiome-based therapeutics in clinical practice. In general, DDPs provide a convenient way to model a collection of distributions which may be related to each other across $\boldsymbol{x}$. Griffin and Steel (2011) and Nieto-Barajas et al. (2012), for example, use a DDP prior for a time series of random probability distributions, and Gelfand et al. (2005) and Duan et al. (2007) developed a variation of the DDP to flexibly model spatial dependence for point-referenced spatial data. In this vein, BNP-ZIMNR can be extended to accommodate spatial/temporal dependence in random distributions $F_{\boldsymbol{x}}^{\chi}$, and may offer a different way of exploring temporal/spatial changes in microbial abundance and diversity.

ACKNOWLEDGEMENTS

This work was supported by NIH: DP2 GM123457–01 to IAC (Irene Chen) and NSF grant DMS-1662427 (Juhee Lee). This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. Sandia National Laboratories, a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

REFERENCES

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

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

De Iorio, M., Müller, P., Rosner, G.L. & MacEachern, S.N. (2004) An ANOVA model for dependent random measures. *Journal of the American Statistical Association* 99(465), 205–215.

De Iorio, M., Johnson, W.O., Müller, P. & Rosner, G.L. (2009) Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics*, 65(3), 762–771.

Devroye, L. & Lugosi, G. (2001) *Total variation*. New York, NY: Springer, pp. 38–46.

Duan, J.A., Guindani, M. & Gelfand, A.E. (2007) Generalized spatial Dirichlet process models. *Biometrika*, 94(4), 809–825.

Galloway-Peña, J.R., Smith, D.P., Sahasrabhojane, P., Wadsworth, W.D., Fellman, B.M., Ajami, N.J. et al. (2017) Characterization of oral and gut microbiome temporal variability in hospitalized cancer patients. *Genome Medicine*, 9(1), 21.

Gardiner, M., Vicaretti, M., Sparks, J., Bansal, S., Bush, S., Liu, M. et al. (2017) A longitudinal study of the diabetic skin and wound microbiome. *PeerJ*, 5, e3543.

Gelfand, A.E. & Dey, D.K. (1994) Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3), 501–514.

Gelfand, A.E., Dey, D.K. & Chang, H. (1992) Model determination using predictive distributions with implementation via sampling-based methods. Technical report, Stanford.

Gelfand, A.E., Kottas, A. & MacEachern, S.N. (2005) Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100(471), 1021–1035.

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

Grice, E.A., Kong, H.H., Conlan, S., Deming, C.B., Davis, J., Young, A.C. et al. (2009) Topographical and temporal diversity of the human skin microbiome. *Science* 324(5931), 1190–1192.

Griffin, J.E. & Steel, M.F. (2011) Stick-breaking autoregressive processes. *Journal of Econometrics*, 162(2), 383–396.

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

Jara, A., Lesaffre, E., De Iorio, M. & Quintana, F. (2010) Bayesian semiparametric inference for multivariate doubly-interval-censored data. *The Annals of Applied Statistics*, 4(4), 2126–2149.

Jonsson, V., Österlund, T., Nerman, O. & Kristiansson, E. (2018) Modelling of zero-inflation improves inference of metagenomic gene count data. *Statistical Methods in Medical Research*, 28, 3712–3728.

Kalan, L.R., Meisel, J.S., Loesche, M.A., Horwinski, J., Soaita, I., Chen, X. et al. (2019) Strain-and species-level variation in the microbiome of diabetic wounds is associated with clinical outcomes and therapeutic efficacy. *Cell Host & Microbe*, 25(5), 641–655.

Kaul, A., Davidov, O. & Peddada, S.D. (2017) Structural zeros in high-dimensional data with applications to microbiome studies. *Biostatistics*, 18(3), 422–433.

Lee, J. & Sison-Mangus, M. (2018) A Bayesian semiparametric regression model for joint analysis of microbiome data. *Frontiers in Microbiology*, 9, 522.

Lee, K.H., Coull, B.A., Moscicki, A.-B., Paster, B.J. & Starr, J.R. (2018) Bayesian variable selection for multivariate zero-inflated models: application to microbiome count data. *Biostatistics*, 21, 499–517.

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

Loesche, M., Gardner, S.E., Kalan, L., Horwinski, J., Zheng, Q., Hodkinson, B.P. et al. (2017) Temporal stability in chronic wound microbiota is associated with poor healing. *Journal of Investigative Dermatology*, 137(1), 237–244.

Love, M.I., Huber, W. & Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550.

MacEachern, S.N. (1999) Dependent nonparametric processes. In: *ASA proceedings of the section on Bayesian statistical science*. American Statistical Association.

MacEachern, S.N. (2000) Dependent Dirichlet processes. Unpublished manuscript, Department of Statistics, The Ohio State University, 1–40.

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

McMurdie, P.J. & Holmes, S. (2014) Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Computational Biology*, 10(4), e1003531.

Millar, R.B. (2009) Comparison of hierarchical Bayesian models for overdispersed count data using DIC and Bayes' factors. *Biometrics*, 65(3), 962–969.

Nieto-Barajas, L.E., Müller, P., Ji, Y., Lu, Y. & Mills, G.B. (2012) A time-series DDP for functional proteomics profiles. *Biometrics*, 68(3), 859–868.

Parfrey, L. & Knight, R. (2012) Spatial and temporal variability of the human microbiota. *Clinical Microbiology and Infection*, 18, 5–7.

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

Ren, B., Bacallado, S., Favaro, S., Vatanen, T., Huttenhower, C. & Trippa, L. (2020) Bayesian mixed effects models for zero-in ated compositions in microbiome data analysis. *Annals of Applied Statistics*, 14(1), 494–517.

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

Rodriguez, A. & Dunson, D.B. (2011) Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis (Online)*, 6(1), 145–177.

Sankaran, K. & Holmes, S.P. (2018) Latent variable modeling for the microbiome. *Biostatistics*, 20(4), 599–614.

Shuler, K., Sison-Mangus, M. & Lee, J. (2019) Bayesian sparse multivariate regression with asymmetric nonlocal priors for microbiome data analysis. *Bayesian Analysis*, 15, 559–578.

Sohn, M.B., Du, R. & An, L. (2015) A robust approach for identifying differentially abundant features in metagenomic samples. *Bioinformatics (Oxford, England)*, 31(14), 2269–2275.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(4), 583–639.

Tang, Z.-Z. & Chen, G. (2018) Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*, 20(4), 698–713.

Verbanic, S., Shen, Y., Lee, J., Deacon, J.M. & Chen, I.A. (2019) Microbial predictors of healing and short-term effect of debridement on the microbiome of chronic wounds: the role of facultative anaerobes. Technical report, University of California Santa Barbara.

Wolcott, R.D., Hanson, J.D., Rees, E.J., Koenig, L.D., Phillips, C.D., Wolcott, R.A. et al. (2016) Analysis of the chronic wound microbiota of 2,963 patients by 16S rDNA pyrosequencing. *Wound Repair and Regeneration*, 24(1), 163–174.

Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A.K. et al. (2017) Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics*, 18(1), 4.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Shuler K, Verbanic S, Chen IA, Lee J. A Bayesian nonparametric analysis for zero-inflated multivariate count data with application to microbiome study. *J R Stat Soc Series C*. 2021;70:961–979. https://doi.org/10.1111/rssc.12493