Bayesian Analysis (2020) **15**, Number 2, pp. 559–578

Bayesian Sparse Multivariate Regression with Asymmetric Nonlocal Priors for Microbiome Data Analysis

Kurtis Shuler*, Marilou Sison-Mangus† and Juhee Lee‡

Abstract. We propose a Bayesian sparse multivariate regression method to model the relationship between microbe abundance and environmental factors for microbiome data. We model abundance counts of operational taxonomic units (OTUs) with a negative binomial distribution and relate covariates to the counts through regression. Extending conventional nonlocal priors, we construct asymmetric nonlocal priors for regression coefficients to efficiently identify relevant covariates and their effect directions. We build a hierarchical model to facilitate pooling of information across OTUs that produces parsimonious results with improved accuracy. We present simulation studies that compare variable selection performance under the proposed model to those under Bayesian sparse regression models with asymmetric and symmetric local priors and two frequentist models. The simulations show the proposed model identifies important covariates and yields coefficient estimates with favorable accuracy compared with the alternatives. The proposed model is applied to analyze an ocean microbiome dataset collected over time to study the association of harmful algal bloom conditions with microbial communities.

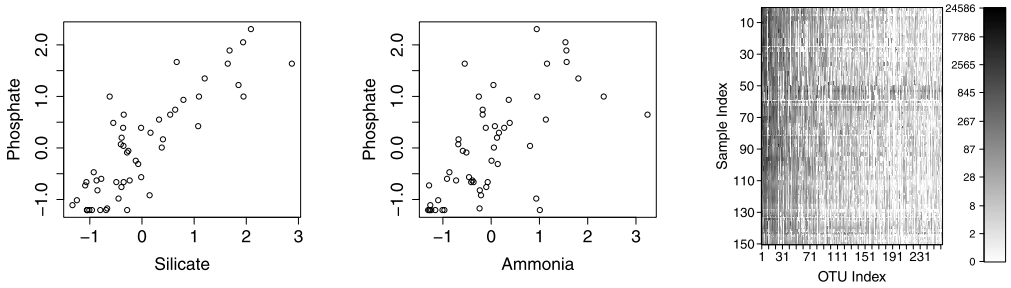
Keywords: count data, harmful algal bloom, microbiome, negative binomial, next-generation sequencing, nonlocal prior, stochastic search variable selection.

1 Introduction

Microbiome data are widely used in exploring microbial communities across many disciplines including medicine, toxicology, immunology, ecology and environmental sciences (Clooney et al., 2016; Knight et al., 2017; Aguiar-Pulido et al., 2016). High-throughput sequencing of 16S ribosomal RNA (rRNA) gene amplicons has enabled thorough profiling of the genetic contents of microbial communities, and provided opportunities to understand the interactions of microbes with their environment and their hosts. Estimating changes in microbe abundance in the community with respect to changes in candidate predictors can be formulated as a multivariate regression problem. When there are many candidate variables, some variables may be redundant or irrelevant. Variable selection procedures are commonly used to identify biologically interpretable and predictive covariates, and subsequently to quantify their associations with microbial

*Department of Statistics, University of California Santa Cruz, Santa Cruz, CA, kshuler@ucsc.edu
†Department of Ocean Sciences, University of California Santa Cruz, Santa Cruz, CA, msisonma@ucsc.edu
‡Department of Statistics, University of California Santa Cruz, Santa Cruz, CA, juheelee@soe.ucsc.edu

 https://doi.org/10.1214/19-BA1164



(a) Silicate vs Phosphate (b) Ammonia vs Phosphate (c) OTU counts

Figure 1: [Ocean Microbiome Data] Panels (a) and (b): Scatterplots of selected environmental factors from the ocean microbiome dataset. Panel (c): Heatmap of the ocean microbiome OTU counts. Darker shades indicate larger counts.

communities. As a specific example, we consider the ocean microbiome dataset in Lee and Sison-Mangus (2018) that consists of 263 operational taxonomic units (OTUs) in 150 samples collected at 54 time points. Ten candidate predictor variables, including abundance levels of harmful algal bloom species (HAB species) as well as nutrient and physical variables, were recorded to investigate their potential associations with microbial communities. Nutrients such as ammonia, phosphate, and silicate in seawater are closely related to each other, as shown in Figure 1(a) and (b), because they are controlled by biological cycling in the ocean. In such contexts, parsimonious models that include only a subset of the covariates truly associated with microbial abundances are preferable. Microbiome data is typically high-dimensional, sparse, and over-dispersed; and sampling procedures can introduce complex dependencies in the resulting data. Constructing a sparse model that allows a flexible dependence structure across samples is crucial to obtain a better understanding of the underlying biological processes.

An OTU represents a microbial taxa based on DNA sequence similarity of taxonomic marker genes, such as the 16S rRNA gene, and microbiome data is typically summarized with an OTU abundance table in a $J \times N$ matrix, where J and N are the numbers of OTUs and samples, respectively. Such data presents a number of analytical challenges. The elements of the table are OTU counts which can be used as a proxy for taxa abundances in a sample. However, the raw OTU counts depend on the amount of effort put into the sequencing procedure for each sample (the "sequencing depth") and do not reflect absolute OTU abundances in the environment of interest, making abundance comparisons more difficult. For statistical analysis OTU counts are commonly converted to normalized counts (relative abundances) by dividing the raw counts by the total sample count or by normalizing factors estimated through some other method (Witten, 2011; Zhang et al., 2017). While appealing for their simplicity, these normalization procedures may introduce bias in parameter estimation, and their inflexibility can make inference less robust (Li et al., 2017). Moreover, microbiome data typically has a large J, and building models that can adequately limit false positive rates but still can identify significant relationships between OTU abundance and environmental factors is challenging. In addition, the variance of OTU counts tends to

be greater than the variance of multinomial or Poisson data, and a large proportion of OTUs have negligible counts in most of the samples.

Many statistical methods have been proposed for microbiome data analysis, including models to characterize community structure and to identify relationships between OTUs and covariates. For association studies, Poisson, multinomial, and negative binomial models are popular for modeling OTU counts, oftentimes with the distribution means related to covariates through a link function (Paulson et al., 2013). Some of those works consider each OTU individually, ignoring community structure (e.g., edgeR in Robinson et al. (2010) and negative binomial mixed model (BhGLM) in Zhang et al. (2017)). More recently, approaches of jointly modeling all OTUs, mostly through a multinomial distribution, have been developed to improve inference by borrowing strength across OTUs. See Chen and Li (2013); Xia et al. (2013); Grantham et al. (2017); Wadsworth et al. (2017); Ren et al. (2017a,b); Mao et al. (2017); Lee and Sison-Mangus (2018) among many others. Wadsworth et al. (2017) and Mao et al. (2017) used a multinomial-Dirichlet (MD) regression model to relate a set of covariates to abundance counts. Wadsworth et al. (2017) used spike-and-slab mixture priors to identify significantly associated covariates. Mao et al. (2017) exploited a graph with the MD regression model to efficiently detect difference in microbiome composition across different groups. Ren et al. (2017a,b) proposed a Bayesian nonparametric approach for microbiome data analysis using a multinomial likelihood and a Dirichlet process prior. Xia et al. (2013) assumed a logistic normal multinomial model and used a group ℓ_1 penalized likelihood to estimate coefficients with variable selection. Chen and Li (2013) also used a sparse group ℓ_1 penalty with a MD regression model. Lee and Sison-Mangus (2018) proposed a Bayesian regression model using a negative binomial likelihood with a Laplace prior for regression coefficients.

To enhance the search for an optimal subset of variables, we build on the model in Lee and Sison-Mangus (2018) and develop a Bayesian sparse multivariate regression model equipped with a variable selection method using asymmetric nonlocal priors (ANLPs), called ANLP-SB. We model counts Y_{ij} of OTU j in sample i with a negative binomial distribution and utilize a log link function to relate the mean counts μ_{ij} to covariates. We let $\log(\mu_{ij}) = g_{ij} + \boldsymbol{x}_i'\boldsymbol{\beta}_j$, where g_{ij} represents the baseline mean count (intercept) of OTU j in sample i and $\boldsymbol{\beta}_j$ is a vector of regression parameters of size P for OTU j. The inferential goal is the estimation of a $J \times P$ regression coefficient matrix, where the β_{jp}s are sparse and possibly interrelated across OTUs. Motivated in part by the particular interest that biologists often place on identifying the directions of covariate effects on OTU abundance in microbiome studies, we construct ANLPs using a truncation mixture with three components for β_{jp}, each for exactly zero, positive and negative effects, where the mixture weights are $\boldsymbol{\pi}_p^\star = (\pi_{p0}^\star, \pi_{p1}^\star, \pi_{p2}^\star)$. While assuming a point mass at zero for $\beta_{jp} = 0$, we assume normal distributions truncated below and above at latent truncation parameter ι_p for positive and negative values of β_{jp}. The marginal prior for nonzero β_{jp} after integrating out ι_p defines a valid NLP (Rossell and Telesca, 2017) and, due to $\pi_{p1}^\star \neq \pi_{p2}^\star$, our NLP is asymmetric. NLPs place zero probability density on $\{0\}$ (see Figure 2 for an illustration) and are competitive against a suite of other variable selection techniques (Johnson and Rossell, 2012; Wu, 2016; Shin et al., 2018). Furthermore, NLPs improve both shrinkage and variable selection in high-dimensional estimation set-

tings (Rossell and Telesca, 2017). In our ocean microbiome data, the abundance levels of many OTUs may have similar relationships with environmental factors including nutrient concentration and phytoplankton abundances inherently, because these variables are trophically-linked. Statistical inference can thus be improved by combining the regression problems of individual OTUs through a hierarchical model. The hierarchical structure enables borrowing of information across OTUs, increasing power for detecting important covariates and estimating their effects. We compare the proposed ANLPs to the corresponding asymmetric local priors (ALPs) that assume normal distributions truncated below and above at zero for $\beta_{jp} > 0$ and $\beta_{jp} < 0$, and conventional symmetric local priors (SLPs) that assume $\mathrm{N}(0, \sigma_p^2)$ for $\beta_{jp} \neq 0$. Simulation studies and real data analysis show favorable performance of ANLPs in identifying relevant covariates and coefficient estimation. For the baseline mean count, we decompose g_{ij} into terms, each of which accounts for differences in sequencing depth, variability in baseline OTU abundances, and dependence across samples within an OTU. The model based normalization through g_{ij} alleviates some pitfalls of using plug-in normalizing factors, and can further improve identification of important covariates and estimation of their effects.

The remainder of the paper is organized as follows. Section 2 describes the proposed ANLP-SB model. Section 3 reports simulation studies to evaluate ANLP-SB and compare it to alternative models including Bayesian regression models with the ALP, SLP, and likelihood based methods. Section 4 summarizes analyses of the ocean microbiome dataset, and we close with a discussion in Section 5.

2 Probability Model

2.1 Sampling Model

Samples are collected at n different time points, $0 < t_1 < t_2 < \ldots < t_n < T$ with K_i replicates at time point t_i, $i = 1, \ldots, n$; and a sample is indexed by t_i and k. $N = \sum_{i=1}^{n} K_i$ is the total number of samples. We let $\boldsymbol{Y}_j = [Y_{t_1 1 j}, \ldots, Y_{t_n K_n j}]'$ represent a N-dimensional response vector of OTU j, where $Y_{t_i k j}$ denotes the count of OTU j in sample (t_i, k). Let $\boldsymbol{x}_{t_i} = [x_{t_i 1}, \ldots, x_{t_i P}]'$ be a P-dimensional vector of covariates, where $x_{t_i p}$ is the value of covariate p at time point t_i. In the remainder of the model description we suppress index i for simpler notation. For OTU j, we consider a negative binomial (NB) regression model,

$$Y_{tkj} \mid \boldsymbol{x}_t, \mu_{tkj}, s_j \overset{indep}{\sim} \mathrm{NB}(\mu_{tkj}(\boldsymbol{x}_t), s_j), \ j = 1, \ldots, J. \tag{1}$$

The model in (1) is parameterized such that the mean and variance of Y_{tkj} are μ_{tkj} and $\mu_{tkj} + \mu_{tkj}^2 s_j$, respectively. We consider a log-linear model $\log(\mu_{tkj}) = g_{tkj} + \boldsymbol{\beta}_j' \boldsymbol{x}_t$, where g_{tkj} represents the baseline mean count of OTU j in sample (t, k) and $\boldsymbol{\beta}_j = [\beta_{j1}, \ldots, \beta_{jP}]'$ is a P-dimensional regression coefficient vector for OTU j. The second term $\boldsymbol{\beta}_j' \boldsymbol{x}_t$ explains the dependence of μ_{tkj} on $\boldsymbol{x}_t$, where each effect acts multiplicatively on μ_{tkj}. Our principal inferential interest lies in the estimation of the $J \times P$ matrix of coefficients β_{jp}. The baseline mean count g_{tkj} accounts for different sample total counts and different baseline abundances across OTUs. g_{tkj} may have additional dependence across samples in an OTU, such as temporal dependence in data collected over time.

$s_j > 0$ is an unknown over-dispersion parameter for OTU j. Unlike a Poisson model for which the variance is equal to the mean, the NB model has an extra component $\mu_{tkj}^2 s_j$ in the variance. For count data such as next generation sequencing (NGS) data, it is common that the observed variance exceeds the assumed variance of the multinomial or Poisson distributions, and the negative binomial distribution is used as a popular alternative to accommodate overdispersion of counts (e.g. Robinson et al. (2010); Zhang et al. (2017)). In the next section we develop models for $\boldsymbol{\beta}_j$, g_{tkj} and s_j.

2.2 Prior

Covariate Effects To achieve a model with parsimony and good predictive power, we build a prior model for $\boldsymbol{\beta}_j$, $j = 1, \ldots, J$ by employing a variable selection approach. To effectively combine J related regression problems, we extend NLPs for $\boldsymbol{\beta}_j$ and construct ANLPs using truncation mixtures. For $j = 1, \ldots, J$ and $p = 1, \ldots, P$, let

$$\begin{aligned}\beta_{jp} \mid \boldsymbol{\pi}_p^\star, \sigma_p^2, \iota_p \quad \overset{indep}{\sim} \quad & \pi_{p0}^\star \mathbb{I}(\beta_{jp} = 0) \; + \; \pi_{p1}^\star \frac{\phi(\beta_{jp}/\sigma_p)}{\sigma_p\{1 - \Phi(\iota_p)\}} \, \mathbb{I}\left(\frac{\beta_{jp}}{\sigma_p} > \iota_p\right) \\ & + \pi_{p2}^\star \frac{\phi(\beta_{jp}/\sigma_p)}{\sigma_p \Phi(-\iota_p)} \, \mathbb{I}\left(\frac{\beta_{jp}}{\sigma_p} < -\iota_p\right), \end{aligned} \tag{2}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ represent the pdf and cdf of the standard normal distribution, respectively, $\mathbb{I}(\beta \in A)$ is a binary indicator function taking the value 1 if $\beta \in A$ or 0 otherwise, and $\iota_p > 0$ is a truncation parameter. As opposed to a conventional approach that has two mixture components for variable selection, the model in (2) has three components, each of which represents the cases of no, positive, and negative effects. We let $\boldsymbol{\pi}_p^\star = (\pi_{p0}^\star, \pi_{p1}^\star, \pi_{p2}^\star)$ be a mixture weight vector with $\sum_{q=0}^2 \pi_{pq}^\star = 1$ and $0 < \pi_{pq}^\star < 1$, $q = 0, 1, 2$. The truncation parameter ι_p can be viewed as a practical significance threshold for the p^{th} covariate. For any $\beta_{jp} \neq 0$ the signal-to-noise ratio $|\beta_{jp}| / \sigma_p$ is greater than ι_p. The mixture model in (2) can be represented with latent indicator variables, $\gamma_{jp} \in \{0, 1, 2\}$, where the values of $\{0, 1, 2\}$ indicate the events of $\{\beta_{jp} = 0\}$, $\{\beta_{jp}/\sigma_p > \iota_p\}$ and $\{\beta_{jp}/\sigma_p < -\iota_p\}$, respectively. We let $\mathrm{P}(\gamma_{jp} = q) = \pi_{pq}^\star$, $q = 0, 1, 2$. If $\gamma_{jp} = 0$, β_{jp} is exactly equal to 0, meaning that covariate p is irrelevant or redundant to modeling counts of OTU j. Covariates with $\gamma_{jp} \neq 0$ are important variables selected for modeling and have large effects following truncated normal distributions. After integrating out γ_{jp}, we recover the prior for β_{jp} in (2). We will specify priors for ι_p and $\boldsymbol{\pi}_p$. The indicator vector $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jP})$ defines a model for OTU j that contains only β_{jp} with $\gamma_{jp} \neq 0$. The estimation of $\boldsymbol{\gamma}_j$ can be viewed as a model selection problem and (2) assigns a priori probability $\prod_{p=1}^P \prod_{q=0}^2 (\pi_{pq}^\star)^{\mathbb{I}(\gamma_{jp}=q)}$ to a model defined by $\boldsymbol{\gamma}_j$.

Remark 2.1. *Consider a model with $\boldsymbol{\gamma}_j$ for OTU j. Let $\boldsymbol{\beta}_j^\star$ denote a vector of β_{jp} with $\gamma_{jp} \neq 0$ only. Given $\boldsymbol{\gamma}_j$, the joint prior of $\boldsymbol{\beta}_j^\star$ can be written as*

$$\begin{aligned} P(\boldsymbol{\beta}_j^\star \mid \boldsymbol{\gamma}_j, \boldsymbol{\delta}, \boldsymbol{\iota}) \quad = \quad & \prod_{p=1;\gamma_{jp}\neq 0}^{P} \left\{ \pi_{p1} \frac{\phi(\beta_{jp}/\sigma_p)}{\sigma_p\{1 - \Phi(\iota_p)\}} \, \mathbb{I}\left(\frac{\beta_{jp}}{\sigma_p} > \iota_p\right) \right. \\ & \left. + \pi_{p2} \frac{\phi(\beta_{jp}/\sigma_p)}{\sigma_p \Phi(-\iota_p)} \, \mathbb{I}\left(\frac{\beta_{jp}}{\sigma_p} < -\iota_p\right) \right\}, \end{aligned} \tag{3}$$

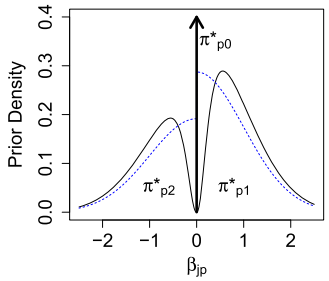


Figure 2: Plot of the asymmetric nonlocal prior density function $\mathrm{P}(\beta_{jp}^{\star} \mid \boldsymbol{\pi}^{\star})$ (black, solid) and its corresponding asymmetric local prior density function (blue, dotted). $\boldsymbol{\pi}_p^{\star} = (0.4, 0.36, 0.24)$ and $\iota_p \sim \text{Gamma}(2.5, 10)$ are assumed.

where $\pi_{pq} = \pi_{pq}^{\star}/(1-\pi_{p0}^{\star})$, $q = 1, 2$, $\boldsymbol{\delta} = \{\sigma_p^2, \boldsymbol{\pi}_p, p = 1, \ldots, P\}$, *and* $\boldsymbol{\iota} = \{\iota_p, p = 1, \ldots, P\}$. *We observe* $P(\boldsymbol{\beta}_j^{\star} \mid \boldsymbol{\gamma}_j, \boldsymbol{\delta}, \boldsymbol{\iota}) \propto d(\boldsymbol{\beta}_j^{\star}) P^L(\boldsymbol{\beta}_j^{\star} \mid \boldsymbol{\gamma}_j, \boldsymbol{\delta})$, *where a local prior (LP)*

$$P^L(\boldsymbol{\beta}_j^{\star} \mid \boldsymbol{\gamma}_j, \boldsymbol{\delta}) = \prod_{p=1;\gamma_{jp}\neq 0}^{P} \left\{ \pi_{p1} \frac{\phi(\beta_{jp}/\sigma_p)}{\sigma_p\{1-\Phi(0)\}} \mathbb{I}\left(\frac{\beta_{jp}}{\sigma_p} > 0\right) + \pi_{p2} \frac{\phi(\beta_{jp}/\sigma_p)}{\sigma_p \Phi(0)} \mathbb{I}\left(\frac{\beta_{jp}}{\sigma_p} < 0\right) \right\}, \tag{4}$$

and a penalty term $d(\boldsymbol{\beta}_j^{\star}) = \prod_{p=1;\gamma_{jp}\neq 0}^{P} \mathbb{I}(|\beta_{jp}|/\sigma_p > \iota_p)$. *Following Corollary 1 of Rossell and Telesca (2017), the prior* $P(\boldsymbol{\beta}_j^{\star} \mid \boldsymbol{\gamma}_j, \boldsymbol{\delta}) = \int P(\boldsymbol{\beta}_j^{\star} \mid \boldsymbol{\gamma}_j, \boldsymbol{\delta}, \boldsymbol{\iota}) P(\boldsymbol{\iota}) d\boldsymbol{\iota}$ *defines a valid nonlocal prior (NLP) if* $P(\boldsymbol{\iota})$ *is absolutely continuous. We call the priors in* (3) *and* (4) *asymmetric nonlocal priors (ANLPs) and asymmetric local priors (ALPs), respectively.*

Figure 2 illustrates an example of the ANLP with a gamma prior for ι_p (black solid line). In contrast with the corresponding ALP (blue dotted line), the ANLP separates the hypotheses $\beta_{jp} = 0$ vs $\beta_{jp} \neq 0$ by assigning small probability to values of β_{jp} close to zero. Furthermore, ANLPs assign different weights to positive and negative values of $\beta_{jp}^{\star}$. Under the NLP, the probability assigned to a model that contains spurious β_{jp} converges to 0 as the sample size grows (Johnson and Rossell, 2012; Wu, 2016; Rossell and Telesca, 2017). The penalty term $d(\boldsymbol{\beta}_j^{\star})$ facilitates model selection (i.e., estimation of $\boldsymbol{\gamma}_j$), and NLPs improve the accuracy of $\boldsymbol{\beta}_j$ estimates compared to LPs. We assume $\iota_p \overset{iid}{\sim} \text{Gamma}(a_\iota, b_\iota)$ with fixed a_ι and b_ι. In (2), $\pi_{p0}^{\star}$ serves as the rate at which the coefficients β_{jp} are exactly zero in the J regression problems. We let $\pi_{p0}^{\star} \overset{iid}{\sim} \text{Be}(a_{\pi 0}, b_{\pi 0})$. We assume the conditional probability of having a positive effect given a covariate is identified as important, $\pi_{p1} \overset{iid}{\sim} \text{Be}(a_{\pi 1}, b_{\pi 1})$ with $\pi_{p2} = 1 - \pi_{p1}$. Priors on $\boldsymbol{\pi}_p^{\star}$ provide an automatic multiplicity correction in variable selection (Scott and Berger, 2010). Following Rossell and Telesca (2017), we let $a_{\pi 0} = P$ and $b_{\pi 0} = 1$, implying the prior inclusion odds $\mathrm{E}((1-\pi_{p0}^{\star})/\pi_{p0}^{\star})$ are $1/(P-1)$. From simulation studies,

we found that with larger P, an informative prior on $\pi^{\star}_{p0}$ favoring very large values (i.e., $a_{\pi 0} \ll b_{\pi 0}$) yields better performance. We let $\sigma_p^2 \overset{iid}{\sim} \text{IG}(a_\sigma, b_\sigma)$ with fixed a_σ and b_σ. Parameters σ_p^2, $\boldsymbol{\pi}^{\star}_p$ and ι_p allow variable specific selection processes. The model can easily be modified to use common σ^2, $\boldsymbol{\pi}^{\star}$ and ι for all covariates if the problem domain does not demand this additional complexity. The hierarchical model construction for β_{jp} through priors on ι_p, $\boldsymbol{\pi}^{\star}_p$ and σ_p^2 facilitates pooling information across OTUs, and improves accuracy of the inference in detecting a parsimonious association between OTUs and covariates, especially for OTUs having small counts in many samples. For example, a large value of $\pi^{\star}_{p1}$ implies positive effect on the abundance (i.e., $\gamma_{jp} = 1$) of most OTUs and the posterior inference on $\pi^{\star}_{p1}$ is informed from all OTUs through the hierarchical structure. In this fashion, the model structure incorporates biological knowledge that environmental factors may have, on average, similar effect directions on OTU abundances.

Baseline Mean Counts We next construct a model for the baseline mean counts g_{tkj} similar to Lee and Sison-Mangus (2018). We first decompose $g_{tkj} = r_{tk} + \alpha_{0j} + \alpha_{tj}$, where terms r_{tk}, α_{0j} and α_{tj} account for different library sizes, different baseline abundances between OTUs, and additional dependence in abundances of an OTU across samples, respectively. Due to its multiplicative structure, the individual terms in g_{tkj} are non-identifiable, whereas g_{tkj} and $\boldsymbol{\beta}_j$ are identifiable. Instead of fixing some terms, we let all the terms be random, and we use distributions with some moment constraints as priors for r_{tk} and α_{0j} to circumvent poor convergence in posterior Markov Chain Monte Carlo (MCMC) simulation. Specifically, we consider the mean-constrained distribution in Li et al. (2017) for r_{tk} and α_{0j};

$$r_{tk} \overset{iid}{\sim} \sum_{\ell=1}^{L^r} \psi_\ell^r \left\{ w_\ell^r \text{N}(\eta_\ell^r, u_r^2) + (1 - w_\ell^r) \text{N}\left(\frac{\upsilon_r - w_\ell^r \eta_\ell^r}{1 - w_\ell^r}, u_r^2 \right) \right\}, \tag{5}$$

$$\alpha_{0j} \overset{iid}{\sim} \sum_{\ell=1}^{L^\alpha} \psi_\ell^\alpha \left\{ w_\ell^\alpha \text{N}(\eta_\ell^\alpha, u_\alpha^2) + (1 - w_\ell^\alpha) \text{N}\left(\frac{\upsilon_\alpha - w_\ell^\alpha \eta_\ell^\alpha}{1 - w_\ell^\alpha}, u_\alpha^2 \right) \right\}, \tag{6}$$

where υ_χ, $\chi = r$ and α, are the prespecified values for the mean constraints and mixture weights ψ_ℓ^χ and w_ℓ^χ with constraints $\sum_{\ell=1}^{L^\chi} \psi_\ell^\chi = 1$ and $0 < \psi_\ell^\chi, w_\ell^\chi < 1$. We fix the number of components L^χ and variances u_χ^2 for $\chi = r, \alpha$. The mixture components in (5) and (6) are convex combinations weighted by w_ℓ^r and w_ℓ^α, respectively. The mixture-of-mixtures formulation encompasses a wide class of distributions, such as multi-modal and skewed distributions. The substantial flexibility of the prior is in contrast with inflexible plug-in estimates of normalizing constants, and this flexibility improves estimation of g_{tkj} and $(\boldsymbol{\gamma}_j, \boldsymbol{\beta}_j)$. Following Lee and Sison-Mangus (2018), we take an empirical approach and use observed counts to specify the values of the mean constraints υ_r and υ_α. We set υ_r to the mean $r'_{tk} = \log(\tilde{r}_{tk})$, where $\tilde{r}_{tk} = \sum_j Y_{tkj} / \sum_{tkj} Y_{tkj}$, and υ_α to the mean of α'_{0j}, where $\alpha'_{0j} = \log(\frac{1}{N} \sum_{tk} Y_{tkj} / \tilde{r}_{tk})$. The particular specification of υ_r and υ_α does not preclude the use of other estimates for the scaling factors. Alternative methods can be used to empirically estimate the mean constraints of scaling factors, for example, maximum likelihood estimates (MLEs) or quantiles in Witten (2011). In the absence of

prior information an empirical approach can yield sensible parameter estimates (Casella, 1985). Alternatively, the mean constraint can be set to 0 as in Li et al. (2017), which can be interpreted as no scaling adjustment on average, or if some prior information is available, priors can be placed on υ_r and υ_α to avoid potential problems with empirical Bayesian approaches (e.g., Scott and Berger (2010)). Our sensitivity analysis to the specification of υ_r and υ_α shows robustness of the model in estimating parameters of interest β_{jp} as well as g_{tkj}; details are in Section 3. We finally let $w_\ell^\chi \overset{iid}{\sim} \text{Be}\,(a_{w^\chi}, b_{w^\chi})$ with fixed a_{w^χ} and b_{w^χ}, $\eta_\ell^\chi \overset{iid}{\sim} \text{N}\left(\upsilon_\chi, b^2_{\eta^\chi}\right)$ with fixed $b^2_{\eta^\chi}$, and $\psi_\ell^\chi \sim \text{Dir}(\boldsymbol{a}_{\psi^\chi})$ with fixed $\boldsymbol{a}_{\psi^\chi}$ for $\chi = r$ and α.

In the ocean microbiome data the samples were collected over time and the baseline mean count g_{tkj} of OTU j may be dependent over time. We model temporal dependence in the baseline mean counts by letting α_{tj} change over time. We use a process convolution model (Higdon, 2002) and let $\alpha_{tj} = \sum_{m=1}^{M} K(t - u_m)\theta_{mj}$. The process convolution model provides a good approximation to a continuous underlying process without a large burden in computation (Lee et al., 2005). Accounting for the dependence structure in temporally adjacent samples can further enhance the estimation of $\boldsymbol{\gamma}_j$ and $\boldsymbol{\beta}_j$. We place the knots u_m, $m = 1, \ldots, M$ on a uniform grid spanning the times when the samples were collected, $[-T', t_n + T']$ with $T' > 0$. We use a Gaussian kernel $\text{N}(0, \tau_j^2)$ for $K(\cdot)$, and following Xiao (2015), fix the variance/range parameter at $2n/M$. Finally, we place independent normal priors centered at zero on the convolution component coefficients, $\theta_{mj} \overset{iid}{\sim} \text{N}(0, \tau_j^2)$, with $\tau_j^2 \overset{iid}{\sim} \text{IG}(a_\tau, b_\tau)$.

We assume OTU specific overdispersion parameters $s_j \overset{iid}{\sim} \text{Log-Normal}(h, \kappa^2)$, with $h \sim \text{N}(a_h, b_h^2)$ and $\kappa^2 \sim \text{IG}(a_\kappa, b_\kappa)$, where a_h, b_h^2, a_κ and b_κ are fixed hyperparameters. NGS data does not have enough information for precise estimation of individual s_j and the hierarchical model can yield improved estimates.

2.3 Posterior Computation

To aid in the posterior computation, as is common in finite mixture models, we introduce auxiliary variables $(c_{tk}^r, \lambda_{tk}^r)$ and $(c_j^\alpha, \lambda_{tk}^\alpha)$, which indicate a mixture component for r_{tk} and α_{0j} in (5) and (6), where $c_{tk}^\chi \in \{1, \ldots, L^\chi\}$ and $\lambda_{tk}^\chi \in \{0, 1\}$, $\chi = r, \alpha$. Similar to γ_{jp}, we define the distribution of r_{tk} and α_{0j} conditional on the auxiliary variables. Let $\underline{\boldsymbol{\theta}} = \{\boldsymbol{s}, \boldsymbol{\alpha}_0, \boldsymbol{\theta}_m, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\pi}_0, \boldsymbol{\pi}_1, h, \kappa^2, \tilde{\boldsymbol{r}}, \boldsymbol{\psi}^r, \boldsymbol{\eta}^r, \boldsymbol{w}^r, \boldsymbol{c}^r, \boldsymbol{\lambda}^\alpha, \boldsymbol{\psi}^\alpha, \boldsymbol{\eta}^\alpha, \boldsymbol{w}^\alpha, \boldsymbol{c}^\alpha, \boldsymbol{\lambda}^\alpha, \boldsymbol{\iota}\}$ denote the vector of all unknown parameters. In the ocean microbiome data, some of the categorical covariates were missing at random for some samples. For missing values we assume that the categories are a priori equally likely and impute their values during posterior simulation. Let $\boldsymbol{X}_{\text{miss}}$ and $\boldsymbol{X}_{\text{obs}}$ denote the missing categorical covariates and observed covariates, respectively, so that $\boldsymbol{X} = \{\boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{miss}}\}$ a $n \times P$ matrix of covariates. The joint posterior probability model of parameters under the proposed model is

$$\text{P}(\underline{\boldsymbol{\theta}}, \boldsymbol{X}_{\text{miss}} \mid \boldsymbol{Y}, \boldsymbol{X}_{\text{obs}}) \propto \text{P}(\boldsymbol{Y} \mid \boldsymbol{X}, \underline{\boldsymbol{\theta}})\text{P}(\underline{\boldsymbol{\theta}}, \boldsymbol{X}_{\text{miss}}),$$

where $\boldsymbol{Y}$ denotes a $N \times J$ matrix of OTU counts. We use standard MCMC methods to implement posterior inference on the parameters. Usual MCMC posterior simulation

proceeds by iteratively updating each of the parameters conditional on the currently computed values of all other parameters. In addition, we do a joint update of β_{jp} and γ_{jp} through the Metropolis-Hastings algorithm for better mixing.

We assessed convergence and mixing of posterior MCMC simulation and found no evidence of practical convergence problems for the simulation examples and the data analysis in Section 3 and Section 4. Details of the posterior simulation are in Supplementary Section 1 (Shuler et al., 2019). In the supplementary, we also include full conditional derivations and some suggestions to improve mixing and convergence. An R package, `anlpsb`, is also available from https://github.com/kurtis-s/anlpsb.

3 Simulation Studies

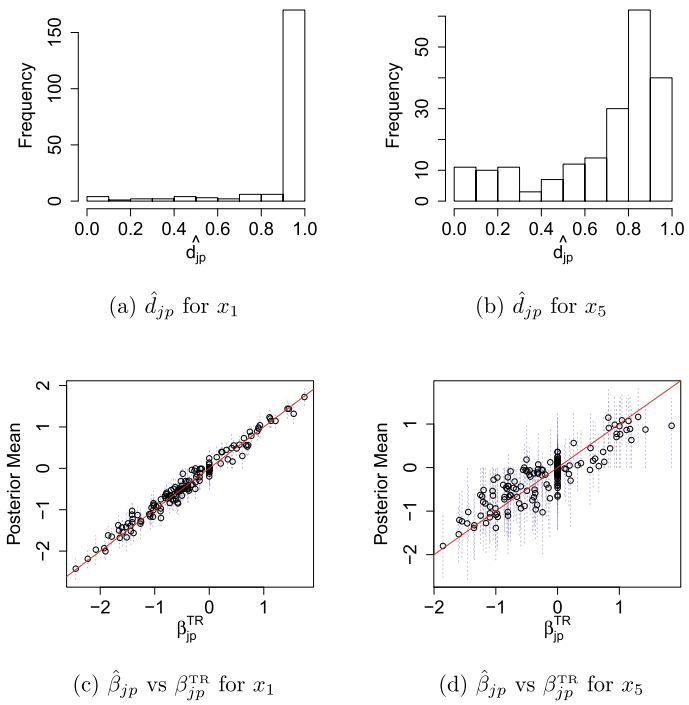
Data Simulation We performed simulation studies to assess the performance of the proposed ANLP-SB model and compared it to alternative models. We assumed $J = 200$ OTUs. We used time points t_i, $i = 1, \ldots n$, the number of replicates K_i and some covariates from the ocean microbiome dataset described in Section 4. Like the ocean microbiome dataset, the simulated data has $n = 54$ time points and total number of samples $N = \sum_i K_i = 150$. We included three continuous covariates, x_1 (silicate), x_2 (water temperature) and x_3 (chlorophyll), and created binary indicator variables for two categorical covariates, the *Alexandrium* (Ax) abundance level and the domoic acid (DA) concentration level. Using the "none" category as the reference category, $x_4 - x_6$ are binary indicators for low, medium, and high abundance levels of Ax, respectively; and $x_7 - x_{10}$ for low, medium, high, and very high concentration levels of DA, respectively. Using these covariates results in $P = 10$. For missing values of Ax, we randomly generated a category for the simulation truth. For the simulation studies and the ocean microbiome data analysis in the following section, the continuous covariates were standardized to have mean 0 and variance 1 before applying the model, as is common in other variable selection techniques. In the ocean microbiome data, covariates were measured in different units (e.g., silicate in μg and water temperature in degree Celsius), and the means and standard deviations of the raw values greatly vary across covariates. The standardization can prevent covariates from being included or discarded purely as a consequence of scale. In our model, common hyperpriors for ι_p and σ_p are used for all p, and use of unstandardized covariates may require more complicated hyperpriors. We used the ocean microbiome data to set r_{tk}^{TR} and α_{0j}^{TR}. We used the OTU counts from the ocean microbiome dataset and computed r'_{tk}, and α'_{0j} as defined in Section 2. r_{tk}^{TR} were then set by randomly permuting $\{r'_{tk}; i = 1, \ldots, n, k = 1, \ldots, K_i\}$, and α_{0j}^{TR} was specified by drawing a random sample of size $J = 200$ from $\{\alpha'_{0j}\}$. We simulated $\pi_{p0}^{\star,\text{TR}} \overset{iid}{\sim} \text{Be}(10, 10)$ and $\pi_{p1}^{\text{TR}} \overset{iid}{\sim} \text{Be}(5, 10)$. We then let $\gamma_{jp}^{\text{TR}} = 0$, 1 or 2 with probabilities, $\boldsymbol{\pi}_p^{\star,\text{TR}} = (\pi_{p0}^{\star,\text{TR}}, (1 - \pi_{p0}^{\star,\text{TR}})\pi_{p1}^{\text{TR}}, (1 - \pi_{p0}^{\star,\text{TR}})(1 - \pi_{p1}^{\text{TR}}))$. We generated ${\sigma_p^2}^{\text{TR}} \overset{iid}{\sim} \text{Unif}(1/2, 1)$ and $\iota_p^{\text{TR}} \overset{iid}{\sim} \text{Unif}(1/10, 3/10)$. We then simulated β_{jp}^{TR} conditional on γ_{jp}^{TR}; if $\gamma_{jp}^{\text{TR}} = 0$, let then $\beta_{jp}^{\text{TR}} = 0$. For the cases of $\gamma_{jp}^{\text{TR}} \neq 0$, we generated β_{jp}^{TR} from the normal distributions with mean 0 and variance $\sigma_p^{2,\text{TR}}$ truncated from below at $\iota_p^{\text{TR}}\sigma_p^{\text{TR}}$ if $\gamma_{jp}^{\text{TR}} = 1$ and from above at $-\iota_p^{\text{TR}}\sigma_p^{\text{TR}}$ if $\gamma_{jp}^{\text{TR}} = 2$. We induced dependence across samples in an OTU using a linear combination of trigonometric functions,

$\alpha_{tj}^{\text{TR}} = A_j \sin\left(\frac{2\pi}{T} h_{ja} t_i - a_j\right) + B_j \sin\left(\frac{2\pi}{T} h_{jb} t_i - b_j\right)$, $0 \le t \le T$. The amplitudes, A_j and B_j, and the frequencies, h_{ja} and h_{jb}, were iid draws from Unif(1, 2) and the phase offsets, a_j and b_j iid draws from Unif(0, T). We generated OTU specific over-dispersion parameters from $s_j^{\text{TR}} \overset{iid}{\sim}$ Log-Normal $\left(-1/2, 1/10^2\right)$. Finally, OTU counts were drawn from $Y_{tkj} \mid \mu_{tkj}^{\text{TR}}, s_j^{\text{TR}} \overset{indep}{\sim} \text{NB}(\mu_{tkj}^{\text{TR}}(\boldsymbol{x}_t),\, s_j^{\text{TR}})$, where $\log(\mu_{tkj}^{\text{TR}}(\boldsymbol{x}_t)) = r_{tk}^{\text{TR}} + \alpha_{0j}^{\text{TR}} + \alpha_{tj}^{\text{TR}} + \boldsymbol{x}_t' \boldsymbol{\beta}_j^{\text{TR}}$.

Posterior Inference To fit the proposed model, we fix the hyperparameters as follows; let $a_\sigma = 1$, $b_\sigma = 1$, $a_\iota = 2.5$, $b_\iota = 10$, $a_{\pi 0} = 1$, $b_{\pi 0} = P$, $a_{\pi 1} = 5$, and $b_{\pi 1} = 5$. For the prior on r_{tk}, α_{0j} and α_{tj}, we let $\boldsymbol{a}_\phi^r = \mathbf{1}$, $a_w^r = 0.5$, $b_w^r = 0.5$, $u_r^2 = 0.1$, $b_{\eta^r}^2 = 0.3$, $\boldsymbol{a}_\psi^\alpha = \mathbf{1}$, $a_w^\alpha = 0.5$, $b_w^\alpha = 0.5$ and $b_{\eta^\alpha}^2 = 1$, hyperparameters for α_{tj}, $a_\tau = 1$ and $b_\tau = 1$. We set the number of knot points to $M = 70$, and the mixture truncation levels to $L^r = L^\alpha = 50$. For the prior on over-dispersion parameter s_j, we set $a_h = -10$, $b_h^2 = 100$, $a_\kappa = 10^{-5}$ and $b_\kappa = 10^{-5}$. We initialized θ_{mj} and β_{jp} using observed y_{tkj}. We generated initial values for σ_p^2 by taking the variance of the initial values for β_{jp}. We ran the MCMC simulation over 50,000 iterations, discarding the first 10,000 iterations as initial burn-in and choosing every fifth sample as thinning. Assessment of MCMC simulation convergence is discussed in Supplementary Section 2.

Figure 3(a) and (b) show histograms of posterior estimates of $\hat{d}_{jp} = \hat{\text{P}}(\gamma_{jp} = \gamma_{jp}^{\text{TR}} \mid \boldsymbol{Y})$, the probabilities that β_{jp} is correctly selected and its effect direction identified for selected covariates x_1 (continuous) and x_5 (binary). Recall that γ_{jp} takes a value of $\{0, 1, 2\}$ representing no, positive, and negative effects. The histograms have a high spike near 1 indicating that ANLP-SB identifies important covariates with their true effect direction with high accuracy. $\hat{d}_{jp}$ tends to be closer to 1 for continuous covariates, while less concentrated around 1 for binary covariates due to small counts for each level. Figure 3(c) and (d) compare posterior mean estimates $\hat{\beta}_{jp}$ of β_{jp} to their true values β_{jp}^{TR} with posterior 95% credible interval estimates. The plots show that the model also provides good estimates of β_{jp}. Similar to $\hat{d}_{jp}$, $\hat{\beta}_{jp}$ is closer to β_{jp}^{TR} with narrower interval estimates for the continuous covariates. Supplementary Figures 1 and 2 show histograms of $\hat{d}_{jp}$ and plots of $\hat{\beta}_{jp}$ versus β_{jp}^{TR} for all covariates. We next compare posterior estimates $\hat{g}_{tkj}$ of the baseline mean counts to their true values. Supplementary Figure 3(a) shows that g_{tkj} are well estimated, which enables the model to produce good estimates of γ_{jp} and β_{jp}. Recall that terms r_{tk}, α_{0j} and α_{tj} in g_{tkj} are not identifiable. Supplementary Figures 3(b)–(f) compare the estimates of r_{tk}, α_{0j} and α_{tj} to the true values. From the figures, the model recovers the parameters only up to a scaling factor and does a good job of capturing the dependence across samples in the truth. In addition, we performed sensitivity analysis to the specification of values of some parameters including (a_ι, b_ι), (a_σ, b_σ), υ_r, υ_α and M. We found that any reasonable choice of those fixed parameters has little impact on the posterior inference, showing robustness of our model. Details of the sensitivity analysis are summarized in Supplementary Section 2.

We further assessed the performance of our model by considering variable selection results from applying the model to 100 replicated datasets. For each dataset, we used the posterior distribution of γ_{jp} and computed the Matthews correlation coefficient (MCC), accuracy (ACC), area under the receiver operating curve (AUC), Brier score



(a) $\hat{d}_{jp}$ for x_1 (b) $\hat{d}_{jp}$ for x_5

(c) $\hat{\beta}_{jp}$ vs β_{jp}^{TR} for x_1 (d) $\hat{\beta}_{jp}$ vs β_{jp}^{TR} for x_5

Figure 3: [Simulation 1] Panels (a) and (b): Histograms of the posterior estimates of $\hat{d}_{jp} = \hat{\text{P}}(\gamma_{jp} = \gamma_{jp}^{\text{TR}})$ for x_1 (Silicate) and x_5 (low concentration of Alexandrium). Panels (c) and (d): Posterior means of the regression coefficients $\hat{\beta}_{jp}$ versus their true values β_{jp}^{TR} for x_1 (Silicate) and x_5 (low concentration of Alexandrium). The dashed blue lines show 95% posterior credible intervals, and the solid red lines are 45 degree reference lines.

(Brier, 1950), and F_1 score. MCC is a combined measure of overall variable selection performance that accounts for an unbalanced number of true positive and false positive cases. MCC ranges between -1 and 1, with MCC $= 1$ indicating perfect selection performance. MCC $= 0$ is expected under random selection, and MCC $= -1$ indicates perfect disagreement between the model's selections and the truth. MCC is defined as

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}},$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. The Brier score is a probability score metric for categorical prediction, defined as $BS = \frac{1}{J \times P} \sum_{jpq} \left(\hat{z}_{jpq} - \mathbb{I}(\gamma_{jp}^{\text{TR}} = q)\right)^2 \in [0, 1]$, where $\hat{z}_{jpq}$ is the posterior probability that $\gamma_{jp} = q$, $q \in \{0, 1, 2\}$. The Brier score is a proper scoring rule (Gneiting and Raftery, 2007), and a lower Brier score indicates better performance. The F_1 score is a metric for binary classification defined as the harmonic mean of the

Model	MCC	ACC	AUC	Brier Score	F_1
ANLP-SB	**0.615** (0.049)	**0.802** (0.023)	**0.885** (0.024)	**0.287** (0.038)	**0.786** (0.026)
ALP-SB	0.302 (0.038)	0.609 (0.030)	0.781 (0.023)	0.546 (0.049)	0.712 (0.027)
SLP-SB	0.295 (0.038)	0.606 (0.029)	0.774 (0.021)	–	0.710 (0.027)
BayesReg	0.539 (0.040)	0.744 (0.026)	0.800 (0.020)	–	0.678 (0.028)
edgeR-L	-0.001 (0.028)	0.499 (0.015)	0.498 (0.017)	–	0.443 (0.028)
edgeR-Q	0.000 (0.029)	0.500 (0.015)	0.498 (0.018)	–	0.472 (0.026)
BhGLM	0.227 (0.049)	0.601 (0.026)	0.632 (0.028)	–	0.488 (0.034)

(a) Variable Selection

Model	RMSE			DIC	LPML
	β_{jp}	$\pi^{\star}_{p0}$	g_{tkj}		
ANLP-SB	**0.279** (0.023)	**0.092** (0.030)	**0.328** (0.043)	**240,430** (6331)	**-4.011** (0.105)
ALP-SB	0.298 (0.018)	0.282 (0.033)	0.341 (0.017)	240,525 (6335)	-4.013 (0.106)
SLP-SB	0.303 (0.015)	0.281 (0.032)	0.353 (0.021)	240,554 (6333)	-4.013 (0.106)
BayesReg	0.302 (0.016)	–	0.356 (0.031)	240,688 (6356)	-4.020 (0.107)
edgeR-L	0.873 (0.030)	–	–	–	–
edgeR-Q	0.864 (0.028)	–	–	–	–
BhGLM	0.979 (0.071)	–	–	–	–

(b) Parameter Estimation and Model Fit

Table 1: [Simulation 1: Comparison] Performance metric averages over 100 simulated datasets with standard deviations in parenthesis. The best performances are in bold.

proportion of true positives among "selected" covariates (also called precision) and the proportion of "selected" covariates among true positive covariates (also called recall). The F_1 score ranges between 0 and 1, with a higher score indicating better performance. For MCC, AUC and F_1, we identified covariates as selected if their posterior probability of $(\gamma_{jp} = 0)$ was less than 0.5. Results from ANLP-SB are summarized in the first row of Table 1(a), where the numbers are averages over the 100 datasets with standard deviations in parenthesis. The scores show ANLP-SB performs well in terms of variable selection and in terms of identifying effect directions.

Comparison We compared the performance of ANLP-SB based on the 100 simulated dataset to alternative models. We include three Bayesian models, sparse regression models with the ALP in (4) (called ALP-SB) and with the symmetric LP for β_{jp} (called SLP-SB) and BayesReg in Lee and Sison-Mangus (2018). For SLP-SB, we assumed equal probability for effect directions, $\gamma_{jp} \overset{indep}{\sim} \text{Ber}(\pi^{\star}_{p0})$ and $\beta_{jp} \mid \gamma_{jp} = 1 \overset{indep}{\sim} \text{N}(0, \sigma^2_p)$ while letting $\beta_{jp} = 0$ for $\gamma_{jp} = 0$. BayesReg assumes Laplace priors for β_{jp} for more shrinkage of the coefficients of insignificant covariates towards zero. We also include the likelihood-based methods edgeR in Robinson et al. (2010) (one of the popular models in practice for NGS data analysis) and the generalized linear regression model with mixed effects (called BhGLM) in Zhang et al. (2017), for comparison. Both methods assume a negative binomial likelihood and use a generalized linear model to accommodate covariate effects similar to the ANLP-SB model. edgeR normalizes raw counts using the trimmed mean of M-values normalization method (Robinson and Oshlack, 2010) to adjust library sizes. It estimates OTU specific overdispersion parameters prior to

analysis through an empirical Bayes approach and uses these estimates to fit the model. edgeR does not explicitly handle dependence structure among samples such as temporal dependence, and we included a term linear in time (edgeR-L) and terms linear and quadratic in time (edgeR-Q) as additional covariates. BhGLM uses the total counts for library size adjustment and induces dependence in samples with shared random effects. The Bayesian comparators hierarchically combine J regression problems similar to ANLP-SB, but edgeR and BhGLM separately analyze each of the OTUs. R package `BhGLM` and Bioconductor package `edgeR` are available for those models. Because edgeR and BhGLM do not handle missing covariates, the true covariate values were used in their simulations.

Under each of the comparators, we computed MCC, ACC, AUC, Brier scores and F_1. The results are summarized in Table 1(a). BayesReg, edgeR, and BhGLM do not explicitly perform variable selection. For BayesReg, we used posterior 95% credible intervals for selection. We considered a variable "selected" if its posterior 95% credible interval did not include zero. For edgeR and BhGLM, selection was performed using p-values with the multiple testing correction of Benjamini and Hochberg (1995) at an α level of 0.05. Brier scores are applicable only for ANLP-SB and ALP-SB, which have a ternary indicator γ_{jp}. The results show that ANLP-SB outperforms the comparators under all metrics. In particular, comparison of ANLP-SB to ALP-SB shows that the performance in variable selection can be greatly improved by the NLP. We also computed estimates of β_{jp}, g_{tkj}, and $\pi^\star_{p0}$, and used them to evaluate root-mean-square error (RMSE) based on the 100 datasets, e.g., $\sqrt{\sum_{jp}(\hat{\beta}_{jp} - \beta^{\text{TR}}_{jp})^2/(100JP)}$. Columns 1–3 of Table 1(b) show that the model with the ANLP also provides better estimates of the parameters, especially for the overall sparsity parameter $\pi^\star_{p0}$. For more comparison among the Bayesian models, the deviance information criterion (DIC) (Spiegelhalter et al., 2002) and log pseudo marginal likelihood (LPML) (Gelfand et al., 1992; Gelfand and Dey, 1994) are computed. DIC measures posterior prediction error based on deviance penalized by model complexity, similar to the Akaike information criterion, where lower values are preferable. LPML is a metric based on cross validated posterior predictive probability with higher values indicating a better model fit. It is defined as the sum of the logarithms of conditional predictive ordinates (CPOs) (Geisser and Eddy, 1979; Geisser, 1993). Columns 4–5 of Table 1(b) show DIC and LPML averaged over the replicated datasets with the standard deviation in parenthesis. DIC and LPML indicate that ANLP-SB provides a better fit to the data than the competing Bayesian models.

Additional Simulations We further examined the performance of our model through additional simulation studies, Simulations 2–8. In Simulations 2–3, we kept most of the simulation set-up used in Simulation 1, including the specification of $\boldsymbol{x}$, r^{TR}_{tk} and α^{TR}_{0j}. In Simulation 2, we assumed that truly irrelevant covariates have negligible effect sizes rather than no effect, that is, $\beta^{\text{TR}}_{jp} \overset{indep}{\sim} \text{N}(0, (\iota_p/6)^2)$ for β_{jp} with $\gamma^{\text{TR}}_{jp} = 0$. The results are summarized in Supplementary Table 2. ANLP-SB obtains good parameter estimates, especially for γ_{jp} and $\pi^\star_{p0}$. It outperforms the competing models, particularly in terms of variable selection, and provides better model fit. For Simulation 3, we simulated the baseline counts from a model different from the assumed model. For this simulation we

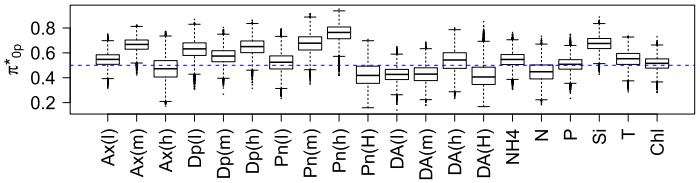
assumed no temporal dependence in the truth and generated $\alpha_{tj}^{\text{TR}} \overset{iid}{\sim} \text{N}(0, (2/3)^2)$. Supplementary Table 3 shows that ANLP-SB recovers good estimates of the associations between the covariates and OTU abundances even when the assumed model for the baseline counts is violated. The comparison shows that our model outperforms the competing models. Simulations 4–8 investigate the performance of ANLP-SB in higher dimensional settings. We increased the values of J, P, n and N and compared its performance to that of the competing models. The results in Supplementary Tables 5–9 show that the ANLP-SB is well-suited for scaling up to higher dimensional settings. ANLP-SB performs favorably relative to the competing models especially for variable selection. More results of the additional simulations, including run-times, are in Supplementary Section 3.

4 Ocean Microbiome Data Analysis

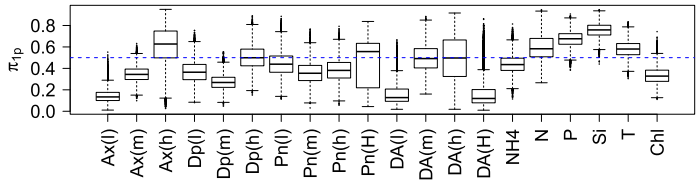
In this section, we summarize our analyses of the ocean microbiome dataset in Lee and Sison-Mangus (2018). Bacterial RNA samples were collected at a total of 54 time points between April 2014 and November 2015 with two or three replicates at a time point, resulting in $N = 150$ samples. Microbial 16s rRNA in the samples was sequenced and a 39,823×150 OTU table was obtained after post-processing of the sequences. We removed OTUs having smaller than 5 counts on average and included $J = 263$ OTUs for our analysis. Figure 1(c) shows a heatmap of the OTU counts in our ocean microbiome data.

The dataset also has continuous and categorical covariates recorded at the same time points. Continuous variables include ammonia (NH_4), silicate (Si), nitrate (N), phosphate (P), temperature (T) and chlorophyll (Chl); and categorical variables include abundance levels of *Alexandrium* (Ax), *Dinophysis* (Dp) and *Pseudo-nitzschia* (Pn), and the domoic acid (DA) concentration level. Binary indicators were created to represent low (ℓ), medium (m), high (h) and very high (H) levels of the categorical variables with the 'none' category used as the reference group. In total, we have $P = 20$ covariates. Supplementary Table 10 lists all covariates. For more details of the dataset, see Lee and Sison-Mangus (2018) and Sison-Mangus et al. (2016). The primary goal of this study is to identify important covariates related to changes in OTU abundance levels and to quantify the effects of those identified covariates.

We specified hyperparameters similar to those in the simulations for the Bayesian models. The MCMC simulation was run over 125,000 iterations, with the first 25,000 iterations discarded as burn-in and every fifth sample kept as thinning and used for inference. It took about 21 minutes for 1,000 iterations on a 3.20GHz Intel i5-6500 processor. Figure 4 summarizes posterior inferences on overall sparsity parameter $\pi_{p0}^{\star}$, and on conditional probability π_{p1} that a covariate has a positive effect given that it has a significant effect. Panel (a) shows that low, medium, and very high DA concentration levels have estimates of $\pi_{p0}^{\star}$ smaller than 0.5, implying that they are significantly related to OTU abundance with probability greater than 0.5. From panel (b), the low and very high concentration levels of DA are associated with depressed OTU abundance with larger probability when they are identified as significant. DA is a chemical secreted by toxic *Pseudo-nitzschia* species whose ecological role is currently unknown. However, previous reports suggest that it could have antibacterial activities (Bates et al., 1995). Both



(a) Posterior distributions of $\pi^\star_{p0}$

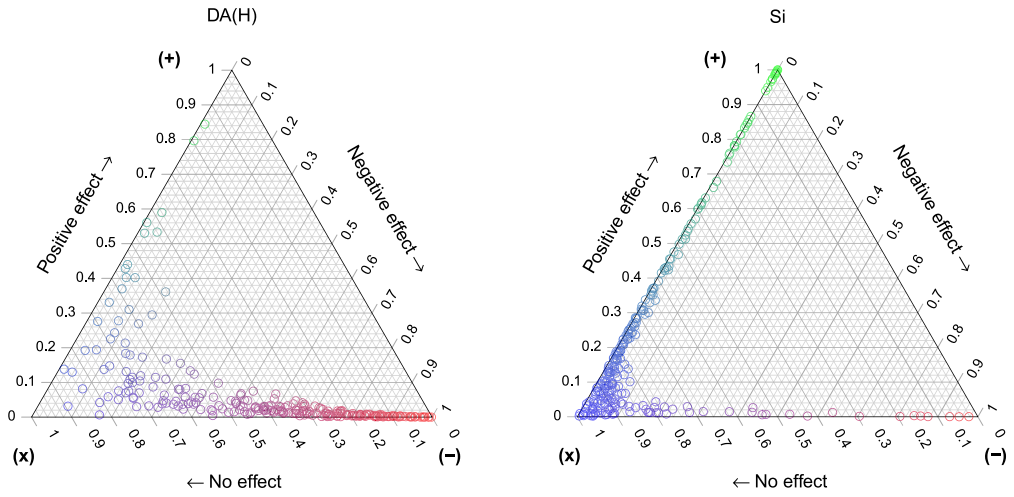


(b) Posterior distributions of π_{p1}

Figure 4: [Ocean Microbiome Data] Panel (a): Boxplots of the posterior distributions of $\pi^\star_{p0}$, the probability of a non-zero effect on OTU abundance. Panel (b): Boxplots the posterior distributions of π_{p1}, the conditional probability of a positive effect direction given the covariate has a non-zero effect.

our preliminary laboratory and ocean studies suggest that it can depress the abundance and growth of some bacterial taxa, while promoting others (Sison-Mangus et al. unpublished). Panel (a) also indicates that silicate is identified as irrelevant with probability $\hat{\pi}^\star_0 = 0.67$, and when it is significant, its effect is positive with probability $\hat{\pi}_1 = 0.75$. Silicate concentration is normally associated with diatom growth as this nutrient is required for silica frustule formation. The breakdown of diatom organic carbon and silicate matter is enhanced by particular groups of bacteria from Flavobacteriales (Bacteroidetes) and Alteromonadales family (Gamma-proteobacteria) (Bidle and Azam, 2001). Moreover, bacterial production is intimately tied to diatom primary production, which biologically explains positive effects of silicate to abundance of some bacterial OTUs.

Figure 5 has simplex plots of a probability vector $\hat{\boldsymbol{z}}_{jp} = (\hat{z}_{jp0}, \hat{z}_{jp1}, \hat{z}_{jp2})$ with $\hat{z}_{jpq}$ being a posterior probability estimate that $\gamma_{jp} = q$, $q \in \{0, 1, 2\}$ for silicate and for the very high concentration level of DA. Circles represent individual OTUs. OTUs having no association with a covariate lie in the bottom-left corner of the plot, those with negative relationships in the bottom-right corner, and those with positive relationships at the apex. Similar to Figure 4(b), the figure indicates silicate tends to not be associated with abundance for many OTUs, while very high DA concentration tends to be negatively associated with abundance for many OTUs. Supplementary Figure 7 has simplex plots for all covariates. Supplementary Figure 8 illustrates posterior inference of β_{jp} and $\mathrm{P}(\gamma_{jp} = 2)$ for the OTUs belonging to class *Gamma-proteobacteria*. The figure shows that many of those OTUs have negative associations with DA, especially with the very high concentration level of DA, compared to the reference level, 'none'. The findings were further validated through a lab experiment using a cultured *Gamma-proteobacteria*



(a) Very high level of domoic acid

(b) Silicate

Figure 5: [Ocean Microbiome Data] Simplex plots of the posterior means $\hat{\boldsymbol{z}}_{jp} = (\hat{z}_{jp0}, \hat{z}_{jp1}, \hat{z}_{jp2})$ of $\gamma_{jp} = 0$, (no effect), $\gamma_{jp} = 1$, (positive effect) and $\gamma_{jp} = 2$, (negative effect). The colors, blue, red and green, indicate no relationship, a negative relationship, and a positive relationship with OTU abundance, respectively.

strain. This bacterial isolate was exposed to different concentrations of DA for 24 to 48 hours followed by growth measurement (Optical Density at 600 nm). We found that the bacteria was significantly affected by DA at concentrations ranging from 25 to 50 $\mu g/ml$, suggesting that DA can indeed inhibit the growth of bacteria (Supplementary Figure 9).

For comparison, we fit the alternative Bayesian models to the dataset. Posterior inferences on $\pi^{\star}_{p0}$ and π_{p1} under ALP-SB and SLP-SB are summarized in Supplementary Figure 11. Under those models, the posterior distributions of $\pi^{\star}_{p0}$ are mostly concentrated in the region between 0.2 and 0.4 for all covariates. ANLP-SB encourages a more parsimonious fit, which is desirable as a sparser fit may better elucidate the biological mechanisms at play. Supplementary Table 11 shows DIC and LPML for the Bayesian models. Both criteria indicate that ANLP-SB gives a better fit to the data.

5 Discussion

We have presented a Bayesian sparse multivariate regression model for microbiome data analysis. We extended NLPs to allow asymmetric probabilities for a coefficient being negative/positive and used the extended ANLPs as a prior for regression coefficients to yield good performance in identification of important covariates related to changes in OTU abundances. By assuming common threshold parameters and overall sparsity parameters, the proposed method makes use of information from all OTUs and yields improved statistical inferences on all OTUs. Taking a probabilistic modeling approach,

our model propagates uncertainties at all levels and provides an assessment of the uncertainty of the selection process. In addition, ANLP-SB simultaneously adjusts for differences in library sizes and accounts for dependence structure in samples via process convolutions.

Our simulation studies and analysis of the ocean microbiome data show that utilizing the ANLPs greatly improves posterior inferences in terms of variable selection and in terms of identifying the direction of relationships between covariates and OTU abundance. In the simulations, ANLP-SB showed robustness to mild violations of the modeling assumptions on effect sizes of irrelevant variables and on dependence structure in samples. ANLP-SB compared favorably to two Bayesian models that used an ALP and an SLP, and to the likelihood-based methods, edgeR and BhGLM. ANLP-SB also appears to yield improved parameter estimates, both at the community and individual OTU levels.

Our ANLP-SB model can be used for analyses of any count data in various fields such as biomedical sciences and economics and can be further extended to accommodate more complex data structures. For example, interaction effects between OTUs can be modeled through graphical models. In particular, Gaussian graphical models use a covariance matrix to represent conditional interdependencies between OTUs and can provide a convenient framework for analyzing and interpreting relationships between OTUs (Dempster, 1972). These are potential areas for future research.

Supplementary Material

Supplementary Materials: Bayesian Sparse Multivariate Regression with Asymmetric Nonlocal Priors for Microbiome Data Analysis (DOI: 10.1214/19-BA1164SUPP; .pdf).

References

Aguiar-Pulido, V., Huang, W., Suarez-Ulloa, V., Cickovski, T., Mathee, K., and Narasimhan, G. (2016). "Metagenomics, Metatranscriptomics, and Metabolomics Approaches for Microbiome Analysis: Supplementary Issue: Bioinformatics Methods and Applications for Big Metagenomics Data." *Evolutionary Bioinformatics*, 12s1: EBO.S36436. URL https://doi.org/10.4137/EBO.S36436. 559

Bates, S. S., Douglas, D. J., Doucette, G. J., and Leger, C. (1995). "Enhancement of domoic acid production by reintroducing bacteria to axenic cultures of the diatom Pseudo-nitzschia multiseries." *Natural Toxins*, 3(6): 428–435. 572

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1): 289–300. URL http://www.jstor.org/stable/2346101. MR1325392. 571

Bidle, K. D. and Azam, F. (2001). "Bacterial control of silicon regeneration from diatom detritus: significance of bacterial ectohydrolases and species identity." *Limnology and Oceanography*, 46(7): 1606–1623. 573

Brier, G. (1950). "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review*, 78: 1. 569

Casella, G. (1985). "An Introduction to Empirical Bayes Data Analysis." *The American Statistician*, 39(2): 83–87. URL https://amstat.tandfonline.com/doi/abs/10.1080/00031305.1985.10479400. MR0789118. doi: https://doi.org/10.2307/2682801. 566

Chen, J. and Li, H. (2013). "Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis." *The annals of applied statistics*, 7(1). MR3086425. doi: https://doi.org/10.1214/12-AOAS592. 561

Clooney, A. G., Fouhy, F., Sleator, R. D., O' Driscoll, A., Stanton, C., Cotter, P. D., and Claesson, M. J. (2016). "Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis." *PLOS ONE*, 11(2): e0148028. URL https://doi.org/10.1371/journal.pone.0148028. 559

Dempster, A. P. (1972). "Covariance selection." *Biometrics*, 157–175. 575

Geisser, S. (1993). *Predictive Inference*, volume 55. CRC Press. MR1252174. doi: https://doi.org/10.1007/978-1-4899-4467-2. 571

Geisser, S. and Eddy, W. F. (1979). "A Predictive Approach to Model Selection." *Journal of the American Statistical Association*, 74(365): 153–160. URL http://www.jstor.org/stable/2286745. MR0529531. 571

Gelfand, A. E. and Dey, D. K. (1994). "Bayesian model choice: asymptotics and exact calculations." *Journal of the Royal Statistical Society. Series B (Methodological)*, 501–514. MR1278223. 571

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). "Model determination using predictive distributions with implementation via sampling-based methods." Technical report, Stanford. MR1380275. 571

Gneiting, T. and Raftery, A. E. (2007). "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association*, 102(477): 359–378. MR2345548. doi: https://doi.org/10.1198/016214506000001437. 569

Grantham, N. S., Reich, B. J., Borer, E. T., and Gross, K. (2017). "MIMIX: a Bayesian Mixed-Effects Model for Microbiome Data from Designed Experiments." *arXiv preprint* arXiv:1703.07747. 561

Higdon, D. (2002). "Space and Space-Time Modeling using Process Convolutions." In *Quantitative Methods for Current Environmental Issues*, 37–56. Springer. URL http://link.springer.com/10.1007/978-1-4471-0657-9_2. MR2059819. 566

Johnson, V. E. and Rossell, D. (2012). "Bayesian Model Selection in High-Dimensional Settings." *Journal of the American Statistical Association*, 107(498): 10.1080/01621459.2012.682536. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3867525/. MR3036423. doi: https://doi.org/10.1080/01621459.2012.742822. 561, 564

Knight, R., Callewaert, C., Marotz, C., Hyde, E. R., Debelius, J. W., McDonald, D., and Sogin, M. L. (2017). "The Microbiome and Human Biology." *Annual Re-*

view of Genomics and Human Genetics, 18(1): 65–86. URL https://doi.org/10.1146/annurev-genom-083115-022438. 559

Lee, H. K. H., Higdon, D. M., Calder, C. A., and Holloman, C. H. (2005). "Efficient models for correlated data via convolutions of intrinsic processes." *Statistical Modelling*, 5(1): 53–74. MR2133528. doi: https://doi.org/10.1191/1471082X05st085oa. 566

Lee, J. and Sison-Mangus, M. (2018). "A Bayesian Semiparametric Regression Model for Joint Analysis of Microbiome Data." *Frontiers in Microbiology*, 9: 522. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5879107/. 560, 561, 565, 570, 572

Li, Q., Guindani, M., Reich, B. J., Bondell, H. D., and Vannucci, M. (2017). "A Bayesian mixture model for clustering and selection of feature occurrence rates under mean constraints." *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6): 393–409. MR3733613. doi: https://doi.org/10.1002/sam.11350. 560, 565, 566

Mao, J., Chen, Y., and Ma, L. (2017). "Bayesian graphical compositional regression for microbiome data." *arXiv preprint* arXiv:1712.04723. MR3844772. 561

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). "Differential abundance analysis for microbial marker-gene surveys." *Nature Methods*, 10: 1200. URL http://dx.doi.org/10.1038/nmeth.2658, https://www.nature.com/articles/nmeth.2658#supplementary-information. 561

Ren, B., Bacallado, S., Favaro, S., Holmes, S., and Trippa, L. (2017a). "Bayesian nonparametric ordination for the analysis of microbial communities." *Journal of the American Statistical Association*, 112(520): 1430–1442. MR3750866. doi: https://doi.org/10.1080/01621459.2017.1288631. 561

Ren, B., Bacallado, S., Favaro, S., Vatanen, T., Huttenhower, C., and Trippa, L. (2017b). "Bayesian Nonparametric Mixed Effects Models in Microbiome Data Analysis." *arXiv preprint* arXiv:1711.01241. 561

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics*, 26(1): 139–140. 561, 563, 570

Robinson, M. D. and Oshlack, A. (2010). "A scaling normalization method for differential expression analysis of RNA-seq data." *Genome biology*, 11(3): R25. 570

Rossell, D. and Telesca, D. (2017). "Nonlocal Priors for High-Dimensional Estimation." *Journal of the American Statistical Association*, 112(517): 254–265. MR3646569. doi: https://doi.org/10.1080/01621459.2015.1130634. 561, 562, 564

Scott, J. G. and Berger, J. O. (2010). "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *The Annals of Statistics*, 2587–2619. MR2722450. doi: https://doi.org/10.1214/10-AOS792. 564, 566

Shin, M., Bhattacharya, A., and Johnson, V. E. (2018). "Scalable Bayesian Variable Selection Using Nonlocal Prior Densities in Ultrahigh-dimensional Settings." *Statistica Sinica*, 28(2): 1053–1078. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5891168/. MR3791100. 561

Shuler, K., Sison-Mangusy, M., and Lee, J. (2019). "Supplementary Materials: Bayesian Sparse Multivariate Regression with Asymmetric Nonlocal Priors for Microbiome Data Analysis." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1164SUPP. 567

Sison-Mangus, M. P., Jiang, S., Kudela, R. M., and Mehic, S. (2016). "Phytoplankton-Associated Bacterial Community Composition and Succession during Toxic Diatom Bloom and Non-Bloom Events." *Frontiers in Microbiology*, 7: 1433. URL https://www.frontiersin.org/article/10.3389/fmicb.2016.01433. 572

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and 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. MR1979380. doi: https://doi.org/10.1111/1467-9868.00353. 571

Wadsworth, W. D., Argiento, R., Guindani, M., Galloway-Pena, J., Shelburne, S. A., and Vannucci, M. (2017). "An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data." *BMC Bioinformatics*, 18(1): 94. URL https://doi.org/10.1186/s12859-017-1516-0. 561

Witten, D. M. (2011). "Classification and clustering of sequencing data using a poisson model." *Annals of Applied Statistics*, 5(4): 2493–2518. MR2907124. doi: https://doi.org/10.1214/11-AOAS493. 560, 565

Wu, H.-H. (2016). "Nonlocal Priors for Bayesian Variable Selection in Generalized Linear Models and Generalized Linear Mixed Models and Their Applications in Biology Data." Ph.d. thesis, The University of Missouri. MR3698950. 561, 564

Xia, F., Chen, J., Fung, W. K., and Li, H. (2013). "A logistic normal multinomial regression model for microbiome compositional data analysis." *Biometrics*, 69(4): 1053–1063. MR3146800. doi: https://doi.org/10.1111/biom.12079. 561

Xiao, S. (2015). "Bayesian nonparametric modeling for some classes of temporal point processes." Ph.D. thesis, University of California Santa Cruz, Santa Cruz. URL https://search.proquest.com/docview/1674523183?accountid=14523, http://ucelinks.cdlib.org:8888/sfx_local?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&genre=dissertations+%26+theses&sid=ProQ:Dissertations+%26+Theses+%40+University+of+Ca. 566

Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A. K., and Yi, N. (2017). "Negative binomial mixed models for analyzing microbiome count data." *BMC Bioinformatics*, 18(1): 4. URL https://doi.org/10.1186/s12859-016-1441-7. 560, 561, 563, 570

Acknowledgments

This work was supported by NSF grant DMS-1662427 (Juhee Lee) and NOAA-ECOHAB PROGRAM (Grant No. NA17NOS4780183, ECOHAB #940) (Marilou Sison-Mangus and Juhee Lee). Collection of environmental data from the Santa Cruz Municipal Wharf was supported by Cal-PReEMPT with funding from the NOAA-MERHAB (#206), ECOHAB and IOOS programs.