### Title

A Bayesian Semiparametric Regression Model for Joint Analysis of Microbiome Data.




### Authors

Lee, Juhee

Sison-Mangus, Marilou



### Publication Date

2018

### DOI

10.3389/fmicb.2018.00522

Peer reviewed

frontiers in Microbiology

METHODS
published: 26 March 2018
doi: 10.3389/fmicb.2018.00522

# A Bayesian Semiparametric Regression Model for Joint Analysis of Microbiome Data

*Juhee Lee*[1]* *and Marilou Sison-Mangus*[2]

[1] Department of Applied Mathematics and Statistics, University of California, Santa Cruz, Santa Cruz, CA, United States,
[2] Department of Ocean Sciences, University of California, Santa Cruz, Santa Cruz, CA, United States

The successional dynamics of microbial communities are influenced by the synergistic interactions of physical and biological factors. In our motivating data, ocean microbiome samples were collected from the Santa Cruz Municipal Wharf, Monterey Bay at multiple time points and then 16S ribosomal RNA (rRNA) sequenced. We develop a Bayesian semiparametric regression model to investigate how microbial abundance and succession change with covarying physical and biological factors including algal bloom and domoic acid concentration level using 16S rRNA sequencing data. A generalized linear regression model is built using the Laplace prior, a sparse inducing prior, to improve estimation of covariate effects on mean abundances of microbial species represented by operational taxonomic units (OTUs). A nonparametric prior model is used to facilitate borrowing strength across OTUs, across samples and across time points. It flexibly estimates baseline mean abundances of OTUs and provides the basis for improved quantification of covariate effects. The proposed method does not require prior normalization of OTU counts to adjust differences in sample total counts. Instead, the normalization and estimation of covariate effects on OTU abundance are simultaneously carried out for joint analysis of all OTUs. Using simulation studies and a real data analysis, we demonstrate improved inference compared to an existing method.

**Keywords: count data, Laplace prior, metagenomics, microbiome, regularizing prior, process convolution, negative binomial model, 16S ribosomal RNA sequencing**

OPEN ACCESS

***Edited by:***
*Michele Guindani,*
*University of California, Irvine,*
*United States*

***Reviewed by:***
*Yanxun Xu,*
*Johns Hopkins University,*
*United States*
*Michael Pester,*
*Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ), Germany*

***Correspondence:**
*Juhee Lee*
*juheelee@soe.ucsc.edu*

***Specialty section:***
*This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology*

***Received:*** *15 November 2018*
***Accepted:*** *08 March 2018*
***Published:*** *26 March 2018*

***Citation:***
*Lee J and Sison-Mangus M (2018) A Bayesian Semiparametric Regression Model for Joint Analysis of Microbiome Data. Front. Microbiol. 9:522. doi: 10.3389/fmicb.2018.00522*

## 1. INTRODUCTION

Microbial communities are influenced by several factors whether they live in the host's guts or other occupied niches. Their successional dynamics could further change in response to perturbations of the host or of the surrounding environments (Turnbaugh et al., 2009; Needham and Fuhrman, 2016). Understanding how abiotic and biotic factors influence the dynamics of microbial communities is of great interest in the field of microbiome studies. Recent revolutionary advances in next-generation sequencing (NGS) technologies along with rapidly decreasing costs, have facilitated the accumulation of large datasets of 16S ribosomal RNA (rRNA) amplicon sequences across various disciplines such as medicine, biology, ecology, and environmental sciences (Woo et al., 2008). Sequencing data is usually pre-treated for quality filtering, noise removal and chimera checking through bioinformatics algorithms and the filtered sequences are clustered into Operational Taxonomic Units (OTUs), which represent similar organisms (microbial species) based on sequence homology (called OTU picking). An OTU abundance table is generated,

recording counts for OTUs in samples. Further statistical data analyses are then performed using the OTU table to answer biological and ecological questions.

Analysis of huge NGS data is computationally expensive and challenging. One of the key challenges is the normalization of counts across samples. Total counts (often called library size or sequencing depth) may vastly vary across different samples due to technical reasons. Thus, observed counts are not directly comparable across samples and cannot be used as a measure of the abundance of an OTU. Normalized counts through rarefaction or relative frequencies are commonly used for easy comparison of OTU abundance across samples. However, such *ad hoc* normalization procedures have been criticized from a statistical perspective since using pre-normalized quantities may undermine the performance of downstream analysis (McMurdie and Holmes, 2014). Another challenge is high dimensionality and sparsity in OTU count data. A dataset typically includes hundreds or thousands of OTUs and a majority of them has zero or very low frequencies in most of samples. For example, **Figure 1A** illustrates a heatmap of OTU counts in our motivating dataset described in section 2.3. It shows that a majority of OTUs has very low counts (gray) in a sample, and the set of OTUs having large counts (blue) vary across samples. Due to such sparsity in data, borrowing strength across OTUs through joint analysis of all OTUs is crucial for improved inference. Recently, various statistical methods including Romero et al. (2014), Chen and Li (2016), Gibbons et al. (2017), and Zhang et al. (2017) have been developed for microbiome studies using NGS data. For example, Zhang et al. (2017) used a negative binomial mixed regression model to study interactions between the microbiome and host environmental/clinical factors. Random effects are used to induce correlation among samples from a group. Common to most of recent methods including Zhang et al. (2017) is separately analyzing each OTU at a time.

We develop a Bayesian semiparametric generalized linear regression model to study the effects of physical and biological factors on abundance of microbes. The proposed method performs mode-based normalization through a hierarchical model, which enables direct modeling of OTU counts. Furthermore, the hierarchical model facilitates borrowing strength between OTUs, between samples, and between time points through joint analysis and improves inference on the effects of covariates $\boldsymbol{X}$ on OTU abundance which are the parameters of our primary interest. Specifically, a negative binomial (NB) distribution parameterized by a mean parameter $\mu$ and an overdispersion parameter $s$ is assumed for OTU counts. The NB distribution flexibly accommodates overdispersion often seen in NGS data and is commonly used as a robust alternative to a Poisson distribution (Anders and Huber, 2010). The expected count $\mu$ of an OTU is decomposed as a product of factors, a baseline mean count $g$ and a nonnegative function $\eta(\boldsymbol{X})$ of covariates that describes their effects on the mean count. We use the log link function for $\eta(\boldsymbol{X})$ and assume that change in a covariate has a multiplicative effect on mean count, where the associated coefficient quantifies the size and direction of the effect. We consider a Laplace prior for the coefficients, a shrinkage prior that is essential in a high dimensional regression setting. Shrinkage priors in regression yield sparse point estimates of the coefficients, where many of the coefficients have values close to zero and few have large values. The sparse estimates improve out-of-sample prediction and produce more interpretable models (Park and Casella, 2008). In addition, shrinkage priors such as a Laplace prior in a regression problem mitigate potential problems by multicollinearity and yield improved coefficient estimates when covariates are high-dimensional and potentially highly correlated (Polson and Scott, 2012). For baseline mean counts, we develop a nonparametric model to combine all OTUs for joint analysis. Baseline mean counts may vary across samples and OTUs. Also, as in our motivating data for which samples were taken over time, there may be temporal dependence in baseline mean counts. To tackle the problem, we further decompose the baseline count $g$ into sample size factor ($r$), OTU size factor ($\alpha_0$), and OTU and time factor ($\alpha_t$), that is, $g = r \times \alpha_0 \times \alpha_t$. Due to the overparametrization of the baseline mean abundance, individual factors are not identifiable. To avoid identifiability issues, we place the regularizing priors with mean constraints (Li et al., 2017) for sample size factor $r$ and OTU size factor $\alpha_0$. In addition, we model a temporal dependence structure between the baseline expected counts for an OTU through a convolutional Gaussian process (Higdon, 1998). The process convolution approach is often used as an alternative approach of the Gaussian process to construct a dependent process due to its efficient computation (Lee et al., 2005; Liang and Lee, 2014). Through simulation studies, we show that estimates of individual parameters $r$, $\alpha_0$, and $\alpha_t$ are not fully interpretable under the proposed model, but baseline mean counts $g$ are identifiable. The model also provides a posterior distribution of $g$ for uncertainty quantification.

The rest of the paper is organized as follows. In section 2 we describe the proposed model and discuss the prior formulations and the resulting posterior inference. We perform simulation studies to assess the proposed model and perform comparison with an existing method that analyzes one OTU at a time. We then apply the proposed model to an ocean microbiome dataset. Section 3 presents the performance of the proposed model from the simulation experiment and the ocean microbime data. Section 4 concludes the paper with a discussion on limitations and possible extensions.

## 2. MATERIALS AND METHODS

### 2.1. Bayesian Semiparametric Regression Model

Suppose that samples are taken at $n$ different time points, $0 \leq t_i \leq T$, $i = 1, \ldots, n$, and with $K_i$ replicates at time point $t_i$. We consider count $y_{t_i,k,j}$ of OTU $j$ in replicate $k$ taken at time $t_i$, where $i = 1, \ldots, n$, $k = 1, \ldots, K_i$, and $j = 1, \ldots, J$. A sample is thus indexed by $t_i$ and $k$. We let the total number of samples $N = \sum_{i=1}^{n} K_i$. Let $\boldsymbol{Y} = [y_{t_i,k,j}]$ denote the $N \times J$ matrix of counts, where $y_{t_i,k,j}$ is integer-valued and nonnegative. Also, suppose that covariates $\boldsymbol{X}_{t_i} = (X_{t_i,1}, \ldots, X_{t_i,P})'$ are recorded at

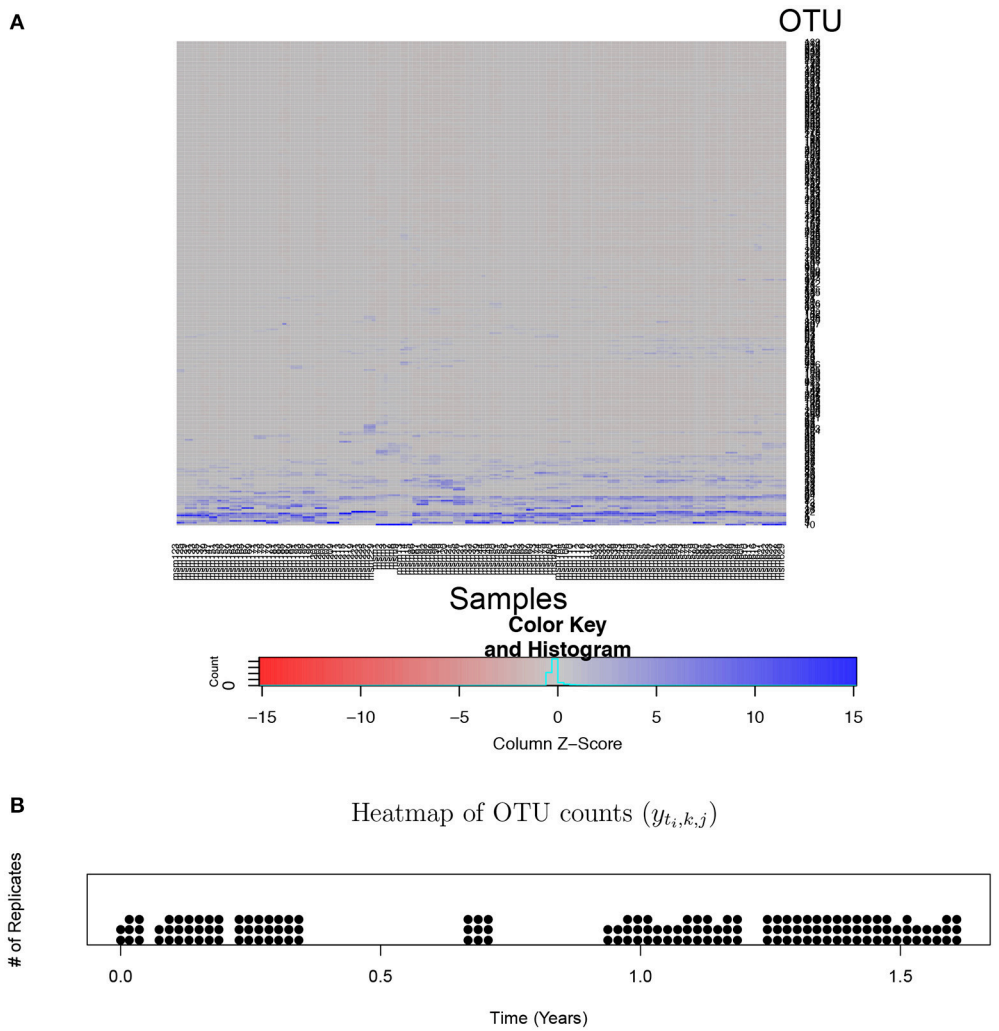


**FIGURE 1 |** Ocean microbiome data. **(A)** Heatmap of OTU counts ($y_{t_i,k,j}$). OTU and samples are in rows and columns, respectively. OTU counts are rescaled within a sample for better illustration. **(B)** 55 time points where ocean microbiome samples were collected are marked on the X-axis and the number of dots at a time point represents the number of replicates ($K_i$) at the time point.

time $t_i$. For example, covariates are physical and biological factors in our motivating data.

### 2.1.1. Sampling Model

Count data by NGS methods is often modeled through a Poisson distribution. The assumption under the Poisson distribution that the variance is equal to the mean is often too restrictive to accommodate overdispersion that variation in data exceeds the mean. The negative binomial (NB) distribution is a popular and convenient alternative to address the overdispersion problem and is widely recognized as a model that provides improved inference to NGS count data (for example, see Robinson and Smyth, 2007; Anders and Huber, 2010). A NB distribution can be characterized by mean and overdispersion parameters. We suppress index $i$ for simpler notation and assume a NB model for count $y_{t,k,j}$ of OTU $j$ in replicate $k$ at time $t$,

$$y_{t,k,j} \overset{indep}{\sim} \text{NB}(\mu_{t,k,j},\ s_j), \tag{1}$$

where mean count $\mu_{t,k,j} > 0$ and overdispersion parameter $s_j > 0$. The model in Equation (1) implies that count of OTU $j$ in replicate $k$ at time $t$ has mean $\text{E}(y_{t,k,j} \mid \mu_{t,k,j}) = \mu_{t,k,j}$ and variance $\text{Var}(y_{t,k,j} \mid \mu_{t,k,j}, s_j) = \mu_{t,k,j} + \mu_{t,k,j}^2 s_j$. The model allows different dispersion levels across OTUs through OTU-specific overdispersion parameters $s_j$. In the limit as $s_j \to 0$, the model in Equation (1) yields the Poisson distribution with mean $\mu_{t,k,j}$. We assume a gamma distribution for a prior distribution of $s_j$, $s_j \overset{iid}{\sim} \text{Ga}(a_s, b_s), j = 1, \ldots, J$, with fixed $a_s$ and $b_s$.

### 2.1.2. Model for Regression

We next model the mean count $\mu_{t,k,j}$ of $y_{t,k,j}$. We decompose the mean count into factors, a baseline mean count and a function of covariates, $\mu_{t,k,j} = g_{t,k,j}\eta_j(\boldsymbol{X}_t)$. Here parameter $g_{t,k,j}$ denotes the baseline mean abundance of OTU $j$ in sample $(t,k)$ and $\eta_j(\boldsymbol{X}_t)$ is a function of covariates $\boldsymbol{X}_t$ for OTU $j$ to model the covariate effects. We construct a generalized regression model by letting $\log(\eta_j(\boldsymbol{X}_t)) = \boldsymbol{X}_t'\beta_j$ , where $\beta_j = (\beta_{j1},\ldots,\beta_{jP})'$ is a $P$-dimensional vector of regression coefficients of OTU $j$ (Lawless, 1987; McCullagh and Nelder, 1989). The coefficient $\beta_{j,p}$ quantifies the effect of covariate $p$ $X_p$ on the mean abundance of OTU $j$. A vector $\beta_j$ close to the zero vector produces a value of $\eta_j(\boldsymbol{X}_t)$ close to 1, and the mean count remains similar to the baseline mean count $g_{t,k,j}$, implying insignificant covariate effects. A negative (positive) of $\beta_{j,p}$ implies a negative (positive) association between mean counts and the $p$-th covariate, and a larger value of $X_{j,p}$ decreases (increases) the mean count, while holding the other covariates constant. We consider a Laplace prior on $\beta_j$. Specifically, we express the Laplace distribution as a scale mixture of normals and assume for $j = 1,\ldots,J$ and $p = 1,\ldots,P$,

$$\begin{aligned} \beta_{j,p} \mid \sigma_j^2, \phi_{j,p} &\overset{indep}{\sim} \mathrm{N}(0, \sigma_j^2\phi_{j,p}), \quad \phi_{j,p} \overset{indep}{\sim} \mathrm{Exp}(\frac{\lambda_j^2}{2}), \\ \lambda_j^2 &\overset{iid}{\sim} \mathrm{Ga}(a_\lambda, b_\lambda), \quad \sigma_j^2 \overset{iid}{\sim} \mathrm{IG}(a_\sigma, b_\sigma), \end{aligned} \tag{2}$$

where $a_\lambda$, $b_\lambda$, $a_\sigma$, and $b_\sigma$ are fixed. $\sigma_j^2$ and $\phi_{j,p}$ denote the global and local shrinkage parameters, respectively, for OTU $j$. After integrating $\phi_{j,p}$ out, the prior distribution of $\beta_{j,p}$ is the Laplace distribution with location parameter 0 and scale parameter $\sqrt{\sigma_j^2}/\lambda_j$, that is, $p(\beta_{j,p} \mid \lambda_j^2, \sigma_j^2) \propto \exp\left(-\lambda_j|\beta_{j,p}|/\sqrt{\sigma_j^2}\right)$. Compared to a normal distribution that is a common choice for the prior of $\beta_{j,p}$, the Laplace distribution has more concentration around zero but allows heavier tails. The regularized regression through the Laplace prior more shrinks the coefficients of insignificantly related covariates into zero and less pulls the coefficients of important covariates toward zero. Shrinkage of $\beta$ estimates through the model in Equation (2) mitigates possible issues due to multicollinearity and efficiently improves estimation of $\beta$ in a high dimensional setting (Polson and Scott, 2012).

### 2.1.3. Model for Baseline Mean Count

We next build a prior probability model for the baseline mean count $g_{t,k,j}$ of OTU $j$ in sample $(t,k)$. We assume $g_{t,k,j} = r_{t,k}\alpha_{0,j}\alpha_{t,j}$ to separate sample $(r_{t,k})$, OTU $(\alpha_{0,j})$, and OTU-time $(\alpha_{t,j})$ factors. Sample total counts $y_{t,k,\cdot} = \sum_{j=1}^{J} y_{t,k,j}$ may greatly differ for different samples possibly due to experimental artifacts. For example, counts of an OTU even in the replicates taken at a time point may vastly differ. Sample specific size factors $r_{t,k}$ account for different total counts in different samples and expected counts normalized by $r_{t,k}$ are comparable across samples. Factor $\alpha_{0,j}$ explains variabilities in baseline mean abundances of OTUs and $\alpha_{t,j}$ models temporal dependence of the mean counts for an OTU, respectively. Factors $\alpha_{0,j}$ and $\alpha_{t,j}$ are not indexed by replicate $k$ and account for stochastic change over time in normalized baseline expected counts of OTU $j$. Collecting all, we write the mean count as

$$\mu_{t,k,j} = g_{t,k,j}\eta_j(\boldsymbol{X}_t) = r_{t,k}\alpha_{0,j}\alpha_{t,j}\eta_j(\boldsymbol{X}_t), \tag{3}$$

The model for $g_{t,k,j}$ in Equation (3) is overparameterized and the individual parameters are not identifiable. To avoid potential identifiability issues, many of NB models rely on some form of approximation for the baseline mean counts. For example, one may find the maximum likelihood estimates (MLEs) of baseline mean abundance under some constraints and plug in those estimates to infer the mean abundance levels $\mu_{t_i,j}$ of OTUs (Witten, 2011). Plugging in MLEs is simple but may not be robust. In particular, the inference is greatly affected by a small change in a few OTUs that have large counts. Moreover, the errors introduced in the baseline mean count estimation will not be reflected in the inference. Several approaches to robustify the estimates are proposed (for example, see Anders and Huber, 2010; Witten, 2011). To circumvent the identifiability issue and provide uncertainty quantification for estimation of $g_{t,k,j}$, we take an alternative in Li et al. (2017) by imposing regularizing priors with mean constraints for $r_{t,k}$ and $\alpha_{0,j}$. We let the logarithm of the factors $\tilde{r}_{t,k} = \log(r_{t,k})$ and $\tilde{\alpha}_{0,j} = \log(\alpha_{0,j})$, and assume the regularizing prior distribution with mean constraints,

$$\begin{aligned} \tilde{r}_{t_i,k} \mid \psi^r, \eta^r, w^r, \nu_r^2, c_r &\overset{iid}{\sim} \sum_{\ell=1}^{L^r} \psi_\ell^r \left\{ w_\ell^r \phi(\eta_\ell^r, \nu_r^2) + (1-w_\ell^r)\phi\left(\frac{c_r - w_\ell^r\eta_\ell^r}{1-w_\ell^r}, \nu_r^2\right)\right\}, \\ \tilde{\alpha}_{0,j} \mid \psi^\alpha, \eta^\alpha, w^\alpha, \nu_\alpha^2, c_\alpha &\overset{iid}{\sim} \sum_{\ell=1}^{L^\alpha} \psi_\ell^\alpha \left\{ w_\ell^\alpha \phi(\eta_\ell^\alpha, \nu_\alpha^2) + (1-w_\ell^\alpha)\phi\left(\frac{c_\alpha - w_\ell^\alpha\eta_\ell^\alpha}{1-w_\ell^\alpha}, \nu_\alpha^2\right)\right\}, \end{aligned} \tag{4}$$

where $\phi(\eta,\nu^2)$ is the probability density function of the normal distribution with mean $\eta$ and variance $\nu^2$, constraints for the mixture weights $\sum_{\ell=1}^{L^r}\psi_\ell^r = \sum_{\ell=1}^{L^\alpha}\psi_\ell^\alpha = 1$ with $0 < \psi_\ell^r < 1$ and $0 < \psi_\ell^\alpha < 1$ , $0 < w_\ell^r < 1$, and $0 < w_\ell^\alpha < 1$ for all $\ell$. Mixture models as in Equation (4) are often used as a basis to approximate any distribution. Each component in Equation (4) is further composed of a mixture of two normals, $\mathrm{N}(\eta_\ell,\nu^2)$ and $\mathrm{N}\left(\frac{(c-w_\ell\eta_\ell)}{(1-w_\ell)},\nu^2\right)$ with weights $w_\ell$ and $1-w_\ell$, respectively, and the mean of the component is $c$. In consequence, the prior and posterior of $\tilde{r}$ and $\tilde{\alpha}$ under the model in Equation (4) satisfy their prespecified mean constraints $c_r$ and $c_\alpha$, respectively. Li et al. (2017) showed that the model in Equation (4) flexibly accommodates various features in a distribution such as skeweness or multi-modality while satisfying the constraints. Furthermore, the model based normalization through Equation (4) enables joint analysis of all OTUs and can further improve

estimation of the covariate effects. With the regularizing priors, baseline mean counts $g_{t,k,j}$ are identifiable, while $r_{t,k}$, $\alpha_{0,j}$, and $\alpha_{t,j}$ are not directly interpretable. More importantly, the parameters of primary interest $\eta_j(\boldsymbol{X}_t)$ can be uniquely estimated and $\beta_{j,p}$'s keep their interpretation as parameters that quantify the effects of covariates on mean counts. We used an empirical approach to fix the mean constraints $c_r$ and $c_\alpha$. Sensitivity analyses were conducted to assess the robustness to the specification of $c_r$ and $c_\alpha$ and show that the model provides reasonable estimates of $g_{t,k,j}$ and moderate changes in the values of $c_r$ and $c_\alpha$ minimally change the estimates. More details of the specification of $c_r$ and $c_\alpha$ are discussed in section 3.1. We fix the numbers of mixture components, $L^r$ and $L^\alpha$ and variances $v_r^2$ and $v_\alpha^2$. We let $\eta_\ell^r \overset{iid}{\sim} \mathrm{N}(c_r, \omega_r^2)$ and $\eta_\ell^\alpha \overset{iid}{\sim} \mathrm{N}(c_\alpha, \omega_\alpha^2)$, where $\omega_r^2$ and $\omega_\alpha^2$ are fixed. We assume $\psi^r = (\psi_1^r, \ldots, \psi_{L^r}^r) \sim \mathrm{Dir}(d_r, \ldots, d_r)$ and $\psi^\alpha = (\psi_1^\alpha, \ldots, \psi_{L^\alpha}^\alpha) \sim \mathrm{Dir}(d_\alpha, \ldots, d_\alpha)$, with fixed $d_r$ and $d_\alpha$. We let $w_\ell^r \overset{iid}{\sim} \mathrm{Be}(a_r, b_r)$, $\ell = 1, \ldots, L^r$ and $w_\ell^\alpha \overset{iid}{\sim} \mathrm{Be}(a_\alpha, b_\alpha)$, $\ell = 1, \ldots, L^\alpha$ with fixed $a_r$, $b_r$, $a_\alpha$, and $b_\alpha$.

Recall that samples are collected over time points $t_1, \ldots, t_n$ in $[0, T]$ and $\alpha_{t,j}$ accounts for temporal dependence in the baseline mean count for an OTU. We let $\tilde{\alpha}_{t,j} = \log(\alpha_{t,j})$ a function in time $t$ and use a stochastic process to model temporal dependence among $\mu_{t,k,j}$. The Gaussian process (GP) is one of the most popular stochastic models for the underlying process in spatial and spatio-temporal data (for example, see Cressie, 1992; Banerjee et al., 2014 among many others). The GP effectively represents the underlying phenomenon in a variety of applications, but it has some drawbacks such as a complex computation that requires a matrix decomposition and problematic estimation of the parameters in its covariance function, potentially leading to difficulties in exploring the posterior distribution (Lee et al., 2005; Liang and Lee, 2014). To alleviate such difficulties of GP models while still maintaining their flexibility and adaptiveness, we use a convolution approach with a kernel function developed in Higdon (1998, 2002). For each OTU, we specify the latent process $\boldsymbol{\theta}_j(t)$ to be nonzero only at the time points $u_1, \ldots, u_M$ in $[0, T]$. Specifically, we consider the GP convolution model,

$$\tilde{\alpha}_{t,j} = \sum_{m=1}^{M} Z(t - u_m)\theta_{m,j},$$

where $\{u_1, \ldots, u_M\}$ a set of basis points in $[-t_1', T + t_2']$ with $t_1', t_2' > 0$, and $Z(t - u_m)$ a Gaussian kernel centered at $u_m$, $Z(t - u_m) = \frac{1}{\sqrt{2\pi\gamma^2}} \exp\{\frac{-(t-u_m)^2}{2\gamma^2}\}$. The number of basis points $M$, their locations $u_m$ and the range parameter $\gamma$ can be treated as random variables by placing prior distributions, e.g., consider a gamma prior for $\gamma$. For simplicity, we fix them as follows. We first choose a value for $M$ and let $u_m$ evenly spaced over time $[-t_1', T + t_2']$. Following Xiao (2015), we let the range parameter $\gamma^2 = ((2T + t_1' + t_2')/M)^2$, that is, the range parameter depends on the value of $M$. Through simulations, we studied the impact of different values of $M$ on the posterior inference of $g_{t,k,j}$. A discussion is included in section 3.1. Given the number of basis points $M$, we assume $\theta_{m,j} \mid \tau_j^2 \overset{indep}{\sim} \mathrm{N}(0, \tau_j^2)$ and $\tau_j^2 \overset{iid}{\sim} \mathrm{IG}(a_\tau, b_\tau)$, $m = 1, \ldots, M$ and $j = 1, \ldots, J$.

We implement posterior inference on the parameters $\tilde{\boldsymbol{\theta}} = (\beta_j, \sigma_j^2, \lambda_j^2, \phi_{j,p}, \tilde{r}_{t,k}, \psi^r, w_\ell^r, \eta_\ell^r, \tilde{\alpha}_{0,j}, \ \tilde{\alpha}_{t,j}, \psi^\alpha, w_\ell^\alpha, \eta_\ell^\alpha, \boldsymbol{\theta}_j, \tau_j^2, s_j)$ via a Markov chain Monte Carlo (MCMC) method based on Metropolis-Hastings algorithm and Gibbs sampling. Each of the parameters is iteratively updated conditional on the currently computed values of all other parameters to simulate a sample from the posterior distribution. The parameters $\tilde{r}$ and $\tilde{\alpha}_0$ jointly determine baseline mean counts and joint updating of $\tilde{r}$ and $\tilde{\alpha}_0$ may greatly improve the mixing. In our ocean microbiome data, some discretized covariates are missing. We treat them as random variables by assuming a uniform distribution over possible categories, and impute their values in MCMC simulation. Full details of our MCMC algorithm are given in Supplementary section 1. We diagnose convergence and mixing of the described posterior MCMC simulation using trace plots and autocorrelation plots of imputed parameters. For the upcoming simulation examples and the data analysis, we found no evidence of practical convergence problems. An R package of the code used for simulations and the analysis of the ocean microbiome dataset in the following sections is available from the authors website https://users.soe.ucsc.edu/~juheelee/.

## 2.2. Simulation Experiment: Data Generation and Comparative Study

We conducted simulation studies to assess the performance of our model. We compared the model to an alternative model, the negative binomial mixed model (NBMM) in Zhang et al. (2017). We assumed a sample of $J = 200$ OTUs. We used the same time points ($t_i$) and numbers of replicates ($K_i$) of our ocean microbiome data as shown in **Figure 1B**. We let $\beta_{j,p}^{\mathrm{TR}} = 0$ with probability 0.85. For $\beta_{j,p}^{\mathrm{TR}} \neq 0$ we simulated $\beta_{j,p}^{\mathrm{TR}}$ from either of $\mathrm{N}(-1.5, 0.05^2)$ or $\mathrm{N}(1.5, 0.05^2)$ with equal probability, where $\mathrm{N}(a, b^2)$ denotes the normal distribution with mean $a$ and variance $b^2$. It implies that a covariate has no effect on OTU abundance with probability 0.85 or may significantly affect mean abundance with the remaining probability 0.15. To specify $r_{t,k}^{\mathrm{TR}}$ and $\alpha_{0,j}^{\mathrm{TR}}$, we did not assume any distribution and used their classical estimates from our ocean microbiome data; following Witten (2011), we first computed estimates of sample size factors $r'_{t_i,k}$ and OTU size factors $\alpha'_{0,j}$ using the ocean microbiome data, $r'_{t_i,k} = y_{t_i,k,\cdot}/y_{\cdots}$ and $\alpha'_{0,j} = \frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{K_i} y_{t_i,k,j}/r'_{t_i,k}$ where $y_{t_i,k,\cdot} = \sum_{j=1}^{J} y_{t_i,k,j}$ and $y_{\cdots} = \sum_{j=1}^{J} y_{\cdot\cdot,j}$. We then randomly sampled from the pool of $r'_{t,k}$ and $\alpha'_{0,j}$ to specify the true values. To simulate temporal dependence in OTU abundance, we let $\tilde{\alpha}_{t_i,j}^{\mathrm{TR}} = a_{t_i,j}\cos(2\pi(\tilde{t}_i - b_{t_i,j})) + c_{t_i,j}(\tilde{t}_i - \tilde{t}^\star)^2$. Here $\tilde{t}_i$ denotes time $t_i$ in year and $\tilde{t}^\star$ the median of $\tilde{t}_i$. We let $a_{t,j} \overset{iid}{\sim} \mathrm{N}(0.15, 0.1^2)$, $b_{t,j} \overset{iid}{\sim} \mathrm{N}(0, 0.5^2)$, and $c_{t,j} \overset{iid}{\sim} \mathrm{N}(0.1, 0.1^2)$ to have different patterns for OTUs. For some OTUs, $\tilde{\alpha}_{t_i,j}^{\mathrm{TR}}$ are illustrated in red squares in **Figures 4E–G**. We generated $s_j^{\mathrm{TR}} \overset{iid}{\sim} \mathrm{Ga}(1, 10)$. We used the covariate matrix of the ocean microbiome data illustrated in **Figure 2** for the simulation study. For the missing covariates in

the data, we generated a value of possible categories with equal probability. We finally simulated OTU counts $y_{t_i,k,j}$ from the negative binomial distribution $y_{t_i,k,j} \overset{indep}{\sim} \text{NB}(\mu^{\text{TR}}_{t_i,k,j}, s^{\text{TR}}_j)$, where $\mu^{\text{TR}}_{t_i,k,j} = r^{\text{TR}}_{t_i,k}\alpha^{\text{TR}}_{0,j}\exp(\tilde{\alpha}^{\text{TR}}_{t_i,j} + \boldsymbol{X}^{\text{TR}}_t\beta^{\text{TR}}_j)$.

For comparison, we used the negative binomial mixed model (NBMM) in Zhang et al. (2017). Similar to the proposed model, the NBMM uses a negative binomial distribution with mean $\mu^{\text{NBMM}}$ and shape parameter $\theta^{\text{NBMM}}$ to model OTU counts and assumes $\log(\mu^{\text{NBMM}}_{t,k,j}) = \log(y_{t,k,\cdot}) + \beta^{\text{NBMM}}_{0,j} + \boldsymbol{X}_t\beta^{\text{NBMM}}_j + \boldsymbol{Z}_{t,k}\boldsymbol{b}^{\text{NBMM}}_j$ where $\boldsymbol{X}_t$ and $\boldsymbol{Z}_{t,k}$ are the covariate matrices for fixed effects and random effects, respectively. It assumes random effects $\boldsymbol{b}^{\text{NBMM}}_j \sim \text{N}(\boldsymbol{0}, \Psi)$. By letting the replicates at a time point share the same random effect, OTU abundances in the replicates at a time point are correlated. The NBMM normalizes OTU counts by sample total counts. That is, sample total counts $y_{t,k,\cdot}$ are used as an offset to adjust for the variability in total counts across samples. Similar to other existing methods, the NBMM performs separate analyses of OTUs. An iterative weighted least squares algorithm is developed to produce the MLEs under the NBMM and implemented in a R function *glmm* in R package *BhGLM*.

## 2.3. Ocean Microbiome Data: Data Description and Preprocessing

We applied the proposed statistical method to ocean microbiome data. Seawater samples were collected weekly at the end of Santa Cruz Municipal Wharf (SCW), Monterey Bay (36.958 °N, −122.017 °W), with an approximate depth of 10 m. SCW is one of the ocean observing sites in Central and Northern California (CenCOOS), where harmful algal bloom species [HAB species: *Alexandrium* (Ax), *Dinophysis* (Dp), *Pseudo-nitzschia* (Pn) etc.] are monitored weekly along with nutrient measurements [ammonia ($NH_4$), silicate (Si), nitrate (N), phosphate (P)], temperature (T), domoic acid (DA) concentration, and chlorophyll (Chl). Details of phytoplankton net tow sampling of measuring phytoplankton abundance, measurement of physical (nutrients and temperature) and biological parameters (chlorophyll $\alpha$ and DA concentration) are described in Sison-Mangus et al. (2016). *Pseudo-nitzschia*, *Dinophysis*, and *Alexandrium* cells were counted with a Sedgewick rafter counter under the microscope. Data for physical and biological factors are available from the website link http://www.sccoos.org/query/?project=Harmful%20Algal%20Blooms&study[]=Santa%20Cruz%20Wharf. Among the 10 variables, the concentration levels of *Alexandrium*, *Dinophysis*, *Pseudo-nitzschia*, and domoic acid have highly right-skewed distributions and are discretized into categories based on their biological properties for our analysis. The ranges of the concentration levels for the discretization are in Supplementary Table 1 and **Figures 2A–J** illustrates all covariates included for analysis. The values of −1, 0, 1, 2, 3, and 4 represent missing values and the categories of None, Low, Medium, High, and Very High for the discretized variables, respectively. Due to high right skeweness, categories corresponding to high concentration levels have low frequencies. Values of the *Dinophysis* concentration level are missing at 20 time points among 55 points used for analysis. Sample correlations between the factors are relatively strong. **Figures 2K,L** shows scatterplots for some selected pairs of the factors.

For bacterial RNA samples, three depth-integrated (0, 5, and 10 ft) water samples were collected at a total of 55 time points between April 2014 and November 2015. Two or three samples are sequenced at each time point. The numbers of replicates at the time points are illustrated in **Figure 1B**. Microbial RNA in the samples was extracted for 16S rRNA sequencing. Post-processing of sequences was performed using the Quantitative Insights Into Microbial Ecology (QIIME 1.9.1) pipeline. A total of nearly 39,823 OTUs were obtained in data after removing singletons. We restricted our attention to OTUs that have greater than or equal to five counts on average. The rule leaves in the end $J = 263$ OTUs for the 150 samples for the analysis. A heatmap of the counts in the filtered data is shown in **Figure 1**. The primary goal of the study is to investigate the effects of physical and biological factors on abundance levels of OTUs, while accounting for baseline abundance levels of OTUs in samples.

# 3. RESULTS

## 3.1. Simulation Experiment: Model Fitting and Comparison

To fit the proposed model for the simulated data designed in section 2.2, we specified hyperparameter values of the model as follows; for the Laplace prior of $\beta_{j,p}$, we let $a_\lambda = b_\lambda = 0.5$ for a gamma prior of $\lambda_j^2$ (with mean $a_\lambda/b_\lambda$ and variance $a_\lambda/b_\lambda^2$) and $a_\sigma = b_\sigma = 0.3$ for an inverse gamma prior for common variances $\sigma_j^2$. For the regularizing priors of $\tilde{r}_{t_i,k}$ and $\tilde{\alpha}_{0,j}$, we fixed $d_\alpha = d_r = 10$, $a_r = b_r = a_\alpha = b_\alpha = 1$, $\omega_r^2 = \omega_\alpha^2 = 1.0$, $v_r^2 = 1$, and $v_\alpha^2 = 2.0$. We also fixed the number of mixture components for the regularizing priors $L_r = 30$ and $L_\alpha = 50$. To specify values of the mean constraints $c_r$ and $c_\alpha$, we took an empirical approach. We used the simulated $y_{t_i,k,j}$, computed estimates of $r_{t_i,k,j}$ and $\alpha_{j,0}$ as described in section 2.2 and fixed the mean constraints at the means of the logarithm of the estimates, respectively. Note that the specified values of $c_r$ and $c_\alpha$ were very different from the means of their true values. For the process convolution prior of OTU-time factor $\tilde{\alpha}_{t_i,j}$, we chose a value of $M$ such that the kernel function at a basis point is not entirely located in a place where no sample is obtained. We let the number of basis $M = 13$ and basis $u_m$, $m = 1, \ldots, M$ evenly spaced between $-10$ and $T_i + 10$. For overdispersion parameter $s_j$ we let $a_s = 1$ and $b_s = 2$. To run MCMC simulation, we initialized the parameters by simulating with their prior distributions. We then implemented posterior inference using MCMC simulation over 25,000 iterations, discarding the first 10,000 iterations as burn-in and choosing every other sample as thinning.

**Figure 3** illustrates the comparison of posterior estimates of $\beta_{j,p}$ to their true values $\beta^{\text{TR}}_{j,p}$ for some selected covariates. In the figure, dots and blue dashed lines represent posterior means $\hat{\beta}_{j,p}$ of $\beta_{j,p}$ and their 95% credible intervals, respectively. $\hat{\beta}_{j,p}$s are around the 45 degree line (red dotted line) for most of $(j, p)$ and most of the interval estimates captures the true values. It implies that the proposed model reasonably recovers $\beta^{\text{TR}}_{j,p}$. For categories 3 and 4 of $X_4$ in **Figures 3I,J**, the credible intervals tend to be wider due to their low frequencies in the data as shown

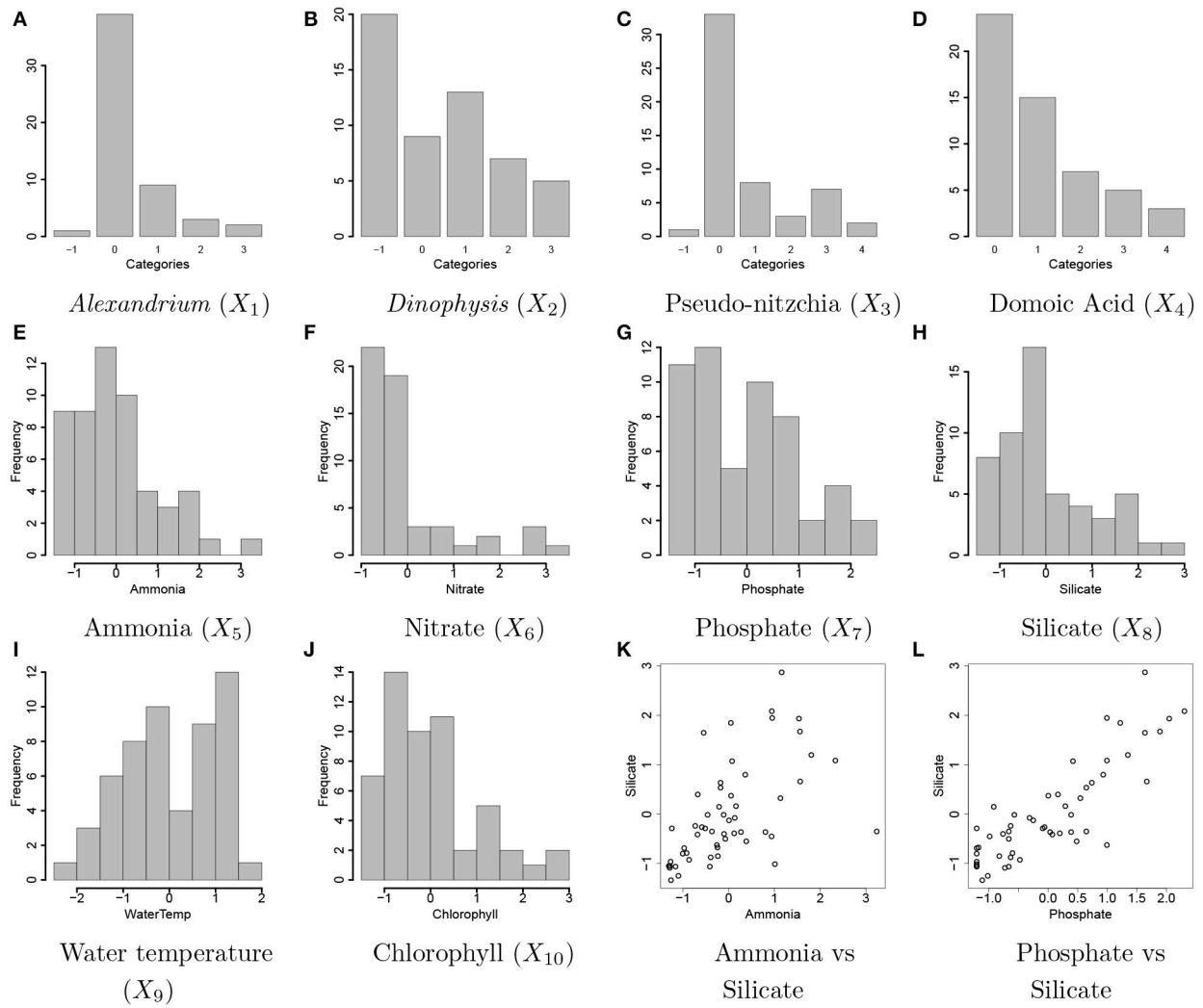


**FIGURE 2 |** Ocean microbiome data. Bar plots of discretized covariates, concentration levels of *Alexandrium* (Ax, $X_1$) and *Dinophysis* (Dp, $X_2$), Pseudo-nitzchia (Pn, $X_3$), domoic acid (DA, $X_4$) in **(A–D)**. The values of $-1$, 0, 1, 2, 3, and 4 represent a missing value, none and low, medium, high, and highest concentration levels, respectively. Histograms of continuous covariates, concentration levels of ammonia ($NH_4$, $X_5$), nitrate (N, $X_6$), phosphate (P, $X_7$), silicate (Si, $X_8$), water temperature (T, $X_9$), concentration level of chlorophyll (Chl, $X_{10}$) in **(E–J)**. The variables are standardized to have mean 0 and variance 1 prior to analysis. A scatterplot of the concentrations of ammonia and silicate and a scatterplot of the concentrations of phosphate and silicate are shown in **(K,L)**, respectively.

in **Figure 2D**. The insert plot in each panel illustrates a scatter plot of $\hat{\beta}_{j,p}$ for $(j, p)$ with $\beta^{\mathrm{TR}}_{j,p} = 0$ . It shows that the proposed regression model with the Laplace prior effectively shrinks $\beta_{j,p}$ with $\beta^{\mathrm{TR}}_{j,p} = 0$ to zero, as is desired in our simulation setup. Supplementary Figure 1 has plots for all covariates.

**Figures 4A–C** illustrate plots of $g^{\mathrm{TR}}_{t,k,j}$ vs estimates of $g_{t,k,j}$ with their means (black dots) and 95% credible intervals (blue vertical lines) for some selected OTUs, $j = 8, 34$, and 48. Recall that we do not attempt to recover the true values of individual $r_{t_i,k}$, $\alpha_{0,j}$, and $\alpha_{t,j}$, but we rather aim to reasonably recover the true baseline mean counts, $g^{\mathrm{TR}}_{t_i,k,j} = r^{\mathrm{TR}}_{t_i,k}\alpha^{\mathrm{TR}}_{0,j}\exp(\tilde{\alpha}^{\mathrm{TR}}_{t_i,j})$. In the figure the estimates are tightly around the 45 degree line, providing evidence that reasonable estimates of baseline mean counts are obtained under the proposed model. **Figure 4D** has a histogram of averaged differences between baseline mean count estimates and their true values, $D_j = \sum_{i=1}^{n}\sum_{k=1}^{K_i}(\hat{g}_{t_i,k,j} - g^{\mathrm{TR}}_{t_i,k,j})/N$. The averaged differences are around zero, implying that the proposed model provides reasonable estimates of baseline mean counts for most of OTUs. We further examined individual parameters. **Figures 4E–G** shows the comparison of estimates of $\tilde{\alpha}_{0,j} + \tilde{\alpha}_{t,j}$ to their true values over time for the same OTUs in **Figures 4A–C**. Black dots and blue vertical lines represent estimates of posterior means of $\tilde{\alpha}_{0,j} + \tilde{\alpha}_{t,j}$ and their 95% credible intervals, respectively. Red squares represent their true values. From the figure, the estimates of $\tilde{\alpha}_{0,j} + \tilde{\alpha}_{t,j}$ are consistently greater than their true values at all time points, but capture their overall temporal trend. **Figure 4H** illustrates a scatterplot of $\tilde{r}^{\mathrm{TR}}_{t,k}$ and their posterior estimates of $\tilde{r}_{t,k}$, where dots and blue vertical

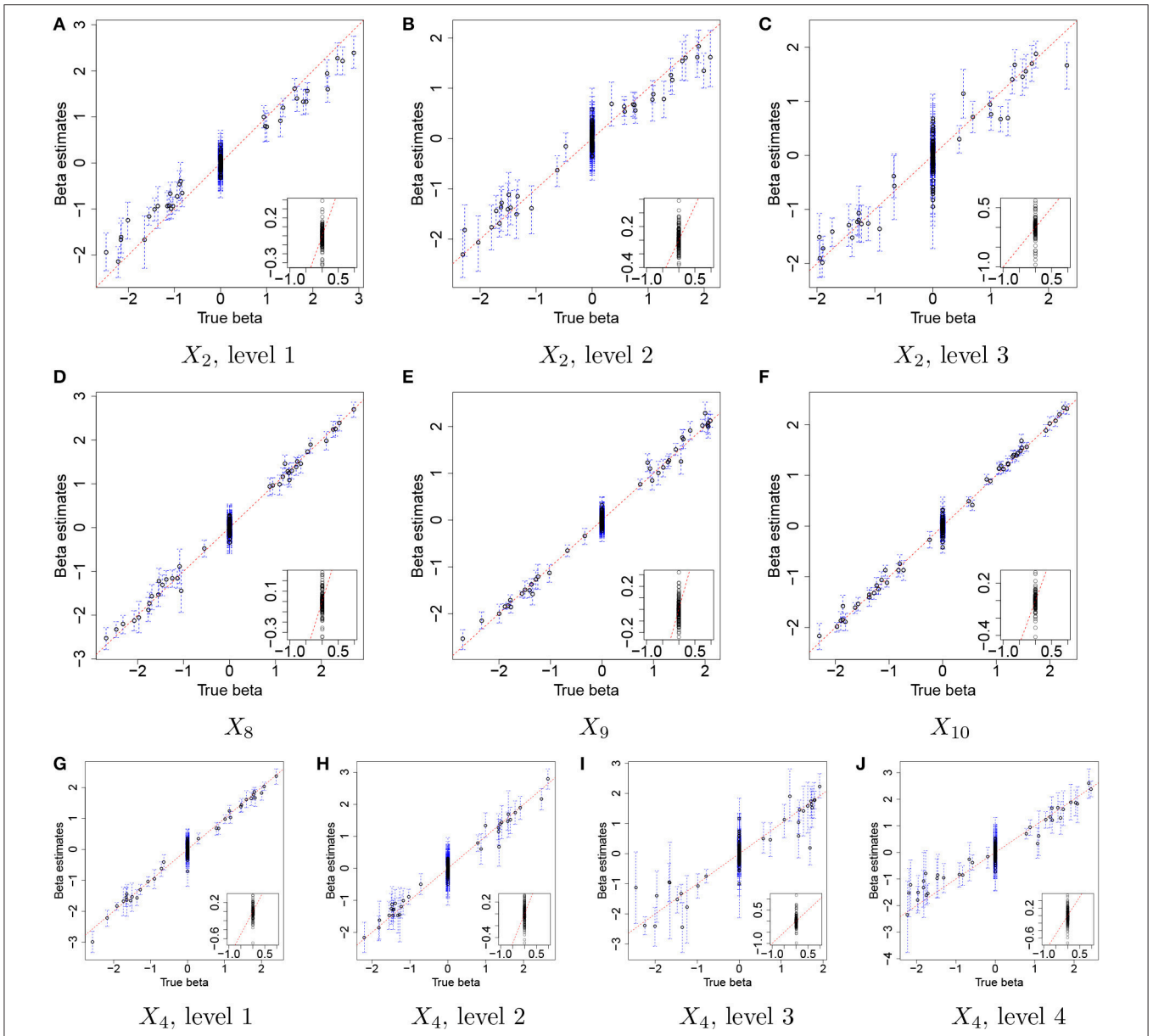


**FIGURE 3 |** Simulation 1—proposed model. Comparison of the true values $\beta_{j,p}^{\mathrm{TR}}$ and the posterior estimates of $\beta_{j,p}$ under the proposed model for some selected covariates. Dots and blue dashed lines represent estimates of posterior means $\hat{\beta}_{j,p}$ of $\beta_{j,p}$ and 95% credible intervals (CIs) of $\beta_{j,p}$, respectively. The insert plot in each panel is a scatter plot of $\hat{\beta}_{j,p}$ and $\beta_{j,p}^{\mathrm{TR}}$ for $(j,p)$ with $\beta_{j,p}^{\mathrm{TR}} = 0$.

intervals denote estimates of posterior means and 95% credible intervals, respectively, and the gray horizontal line is at $c_r$ used for analysis. Different from the estimates of $\tilde{\alpha}_{0,j} + \tilde{\alpha}_{t,j}$, the estimates of $\tilde{r}_{t,k}$ fall below the 45 degree line approximately by the same distance for all OTUs. It shows that estimates of $\tilde{\alpha}_{0,j} + \tilde{\alpha}_{t,j}$ and $\tilde{r}_{t,k}$ have discrepancies from their true values but in the opposite direction and the model can produce reasonable estimates of $g_{t,k,j}$ as seen in **Figures 4A–D**. The true overdispersion parameters $s_j^{\mathrm{TR}}$ are reasonably well estimated as shown in **Figure 4I**. We check the posterior predictive distribution of $Y_{t,k,j}$. The posterior predicted values of $Y_{t_i,k,j}$ with their 95% predictive intervals for OTUs $j = 8, 34$, and 48 are compared to their observed values in Supplementary Figure 2. The figure indicates a reasonable model fit.

In addition, we conducted a sensitivity analysis to the specification of mean constraints $c_r$ and $c_\alpha$ for the priors of $\tilde{r}$ and $\tilde{\alpha}_0$. We used different values for $c_r$ and $c_\alpha$ and compared the estimates of $g_{t,k,j}$ to their truth. Supplementary Figures 3a–c has histograms of averaged differences $D_j$ between $\hat{g}_{t_i,k,j}$ and $g_{t_i,k,j}^{\mathrm{TR}}$ for different specification of $c_r$ and $c_\alpha$. The histograms show

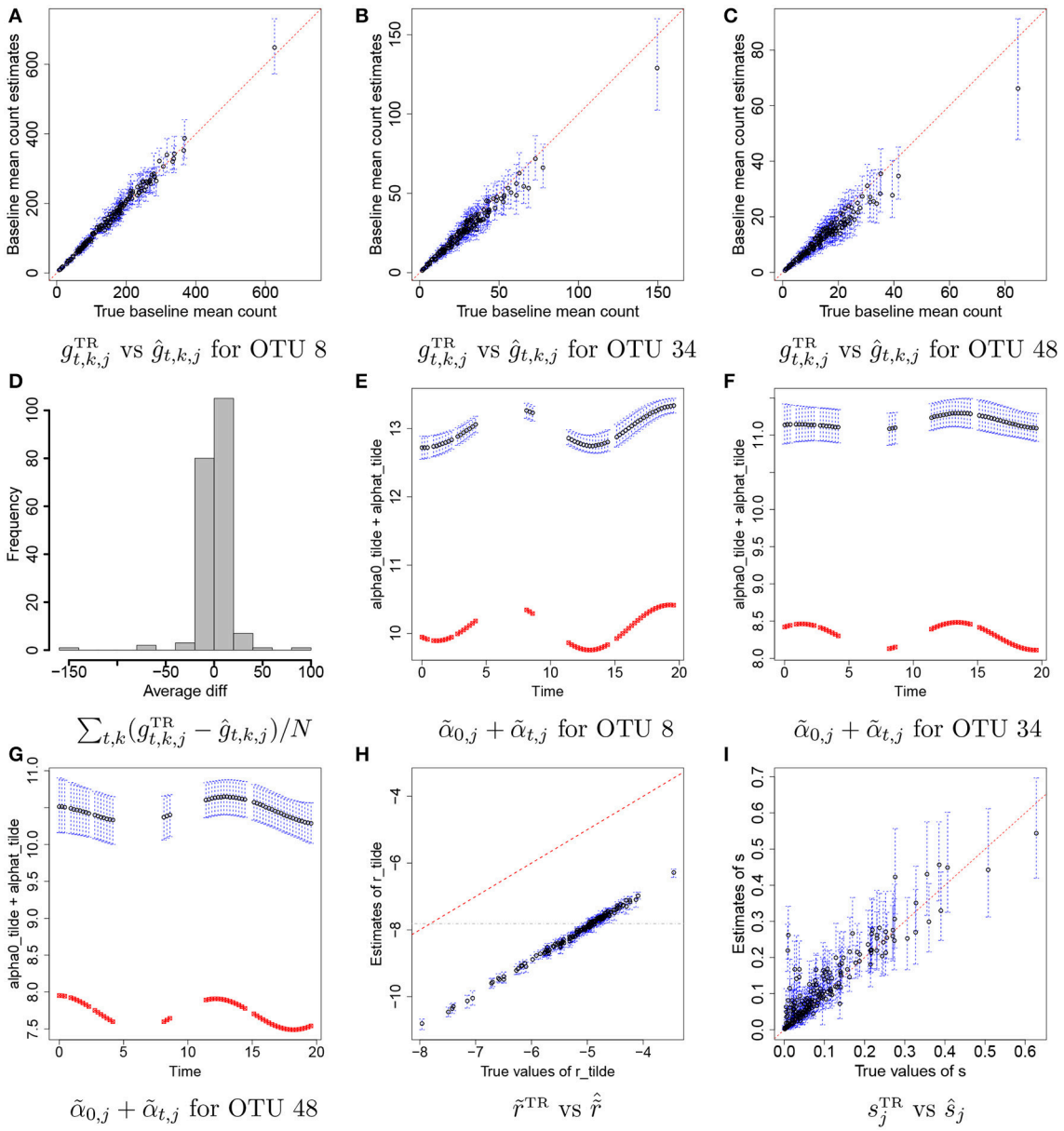


**FIGURE 4 |** Simulation 1—proposed model. Panels **(A–C)** illustrate plots of the true baseline mean counts $g^{TR}_{t,k,j}$ vs their estimates $\hat{g}_{t,k,j}$ for some selected OTUs $j = 8, 34, 48$. Panel **(D)** shows a histogram of averaged differences between $g^{TR}_{t,k,j}$ and $\hat{g}_{t,k,j}$ for each OTU. Panels **(E–G)** show plots of estimates of $\tilde{\alpha}_{0,j} + \tilde{\alpha}_{t,j}$ over time for OTUs $j = 8, 34, 48$. Panel **(H)** has a scatterplot of $\hat{\tilde{r}}_{t,k}$ vs. $\tilde{r}^{TR}$. Panel **(I)** has a scatterplot of ŝ vs. $s^{TR}$. Dots represent posterior mean estimates and blue vertical dotted lines 95% credible intervals. Red squares represent the true values.

minor change in estimates of $g_{t_i,k,j}$ under different specifications of $c_r$ and $c_\alpha$. An sensitivity analysis to the specification of the number $M$ of basis points in the GP convolution model for $\tilde{\alpha}_{t,j}$ was also performed. We used $M = 8$, 13, and 18 and examined estimates of the baseline mean counts, $g_{t,k,j}$. Supplementary Figures 3a,d,e has histograms of averaged differences $D_j$ for each of $M$. The results indicate that the baseline mean counts are reasonably estimated for a range of values of $M$ in the simulation study.

For comparison, we used the NBMM to the simulated data. Since the NBMM does not accommodate missing covariates, we used $\boldsymbol{X}^{TR}$ to fit the NBMM. **Figure 5** compares the MLEs $\hat{\beta}^{NBMM}_{j,p}$ of $\beta_{j,p}$ to the true values for the same covariates used in **Figure 3**. Dots and blue vertical lines represent the MLEs under the NBMM and their 95% confidence intervals, respectively. Comparing **Figure 5** to **Figure 3**, the NBMM produces poor estimates. The MLEs are biased for some covariates (e.g., **Figure 5A**). Also, confidence intervals under the NBMM often

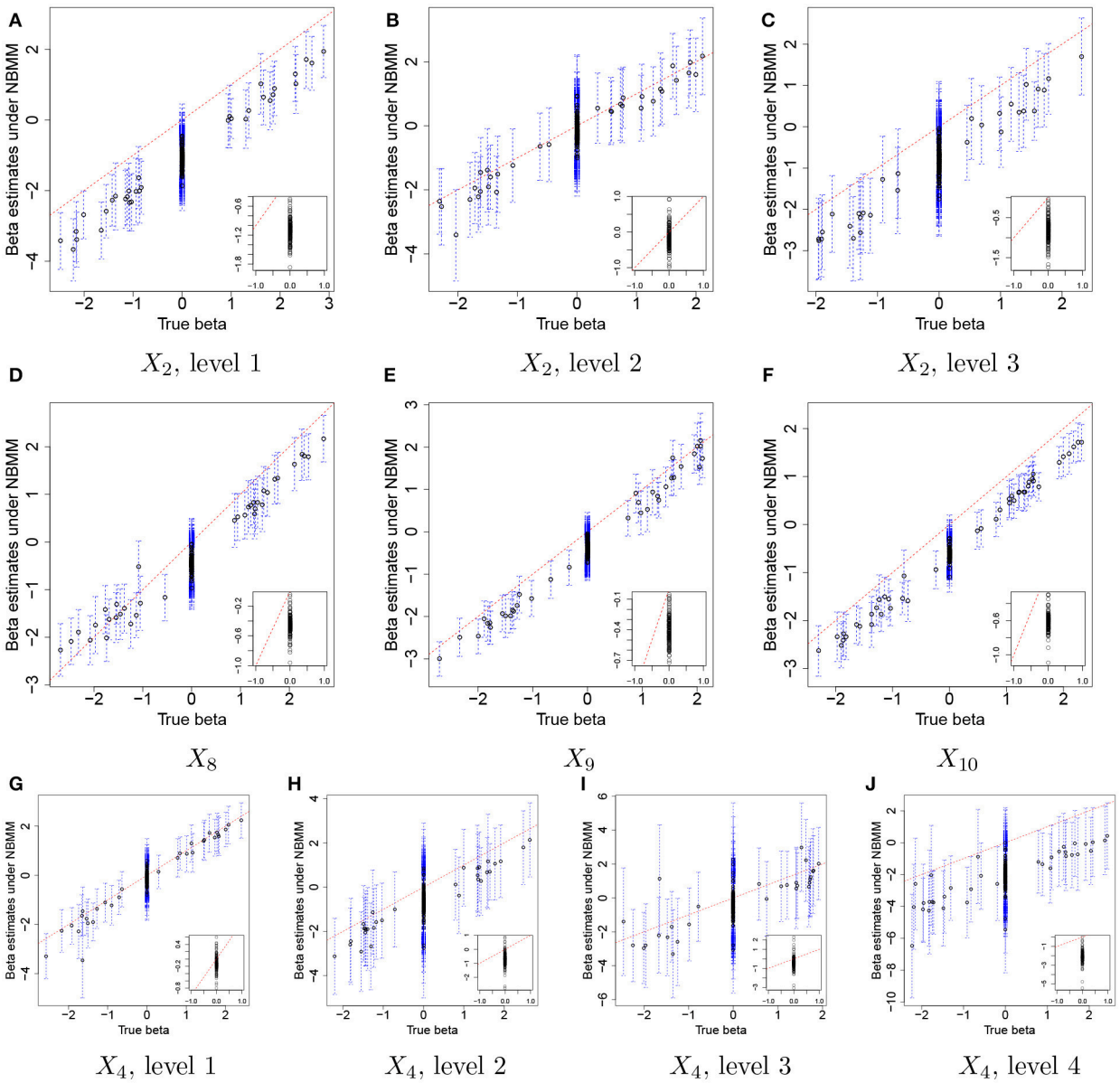


**FIGURE 5 |** Simulation 1—NBMM. Comparison of the true values $\beta_{j,p}^{\mathrm{TR}}$ and maximum likelihood estimates $\hat{\beta}_{j,p}^{\mathrm{NBMM}}$ of $\beta_{j,p}$ under the negative binomial mixed model (NBMM) for some selected covariates. Dots and blue dashed lines represent $\hat{\beta}^{\mathrm{NBMM}}$ and their 95% confidence intervals, respectively. The insert plot in each panel is a scatter plot of $\hat{\beta}_{j,p}^{\mathrm{NBMM}}$ and $\beta_{j,p}^{\mathrm{TR}}$ for $(j,p)$ with $\beta_{j,p}^{\mathrm{TR}} = 0$.

fail to capture the true values and their interval estimates under the NBMM tend to be much wider than those under the proposed model. Normalization through observed sample total counts and inducing correlation in replicates through iid (independent and identically distributed) random effects under the NBMM may lead to poor estimation of the baseline mean abundance for the simulated data, resulting in deterioration of coefficient estimation. In addition, separate analyses of OTUs under the NBMM do not allow to strengthen estimates through combining information across OTUs. Comparing the insert plots in **Figure 5** to those in **Figure 3**, $\hat{\beta}_{j,p}^{\mathrm{NBMM}}$ with $\beta_{j,p}^{\mathrm{TR}} = 0$ tends to more widely spread out from zero and often their confidence intervals fail to capture zero. Supplementary Figure 4 has plots of $\beta_{j,p}$ for all covariates. Supplementary Figures 4, 5 shows the comparison of the estimates $\hat{\theta}^{\mathrm{NBMM}}$ of overdispersion parameters under the NBMM to their true values. Note that $\theta^{\mathrm{NBMM}}$ is the inverse of $s$ in our model. The NBMM underestimates $s_j$ for many OTUs, and yields poor predicted values, implying the lack of a fit.

We further examined the performance of the proposed model through additional simulation studies, Simulations 2 and 3 in

Supplementary section 2. In these simulations, we studied the robustness of the model when different simulation setups are used to simulate data. In Simulation 2, we assumed no temporal dependence among OTU abundance and generated independent samples from a normal distribution for $\tilde{\alpha}_{t,j}$. We assumed that all $\beta_{j,p}^{\mathrm{TR}}$ has nonzero effects for all OTUs and simulated $\beta_{j,p}$ from a mixture of normals. The performance of our model is almost the same as in Simulation 1 (see Supplementary Figures 6–8). Interestingly, the NBMM that assumes iid random effects performs poorly for $\beta$ estimation. In Simulation 3, we simulated $\tilde{\alpha}_{t,j}^{\mathrm{TR}}$ from a discontinuous function. The results show that when the temporal dependence pattern is not smooth as assumed for the GP, estimates of the baseline mean counts under the proposed model are slightly deteriorated but the model produces reasonable inference on $\beta_{j,p}$ (see Supplementary Figures 9–11). A more detailed summary of the additional simulations is given in Supplementary section 2.

## 3.2. Ocean Microbiome Data: Model Fitting and Comparison

We specified hyperparameters similar to those in the simulations and analyzed the microbiome data in section 2.3. The MCMC simulation was run over 25,000 iterations. The first 15,000 iterations were discarded as burn-in and every other sample was kept as thinning and used for inference. **Figure 6** illustrates inference on covariate effects for some selected OTUs, $j = 16$, 34, and 49, taxonomically belonging to *Alteromonadales*, *Halomonas* sp., and *Alteromonadales* in the Gamma-proteobacteria phyla, respectively. Dots and vertical solid lines represent the posterior mean $\hat{\beta}_{j,p}$ and 95% credible interval estimates, respectively. Similar to the results of the simulation study, the credible intervals for high and highest levels of the discretized covariates tend to be wider due to their low frequencies in the data. From **Figure 6A**, on average the medium concentration level of domoic acid (DA, $X_4$) and the concentration level of nitrate (N, $X_6$) significantly decrease the mean abundance of OTU 16 by a multiplicative factor of $\exp(-0.572) = 0.564$ and $\exp(-0.260) = 0.771$, respectively. One may infer that the medium concentration level of domoic acid is significantly associated with lower expected counts for the OTU compared to those with category none of the domoic acid concentration level. A similar argument can be applied to the inference on the nitrate concentration level. Interestingly, we observed statistically significant reduction in abundance from many OTUs belonging to Gamma-proteobacteria including those OTUs for increasing domoic acid concentration (not shown). The resulting inference was further validated through a lab experiment. Most notably, four bacterial cultured isolates belonging to Gamma-proteobacteria (three among them are *Alteromonadales*) were observed to be severely retarded in growth after 2 days of exposure to increasing domoic acid of 0 to 150 $\mu$g/ml in the experiment (Sison-Mangus, unpublished data). This demonstrates that the proposed model successfully identifies important OTUs in ocean bacterial community dynamics for further investigation. More results are presented in Supplementary section 3. Supplementary Figures 12a–c illustrates the posterior estimates of baseline expected counts $\tilde{\alpha}_{0,j} + \tilde{\alpha}_{t,j}$ normalized by sample size factors for the OTUs. From the figure, the baseline expected counts vary over time for those OTUs and the temporal pattern of OTU $j = 34$ is different from those of OTUs $j = 16$ and 49. Histograms of the posterior mean estimates $\hat{\beta}_{j,p}$ of $\beta_{j,p}$, are illustrated in Supplementary Figure 13. The figure does not show clear overall tendency in the direction of association between covariates and OTU counts. Posterior inference on sample size factors $r_{t_i,k}$ and OTU specific overdispersion parameters $s_j$ is illustrated in Supplementary Figures 12d,e.

For comparison, we fitted the NBMM to the data. Since the NBMM does not account for missing values, we use the maximum a posteriori estimates of the missing values under the proposed for the NBMM. We used the R function *glmm* and the algorithm produced warning messages on convergence for 32 OTUs. **Figure 7** illustrates $\hat{\beta}_{j,p}^{\mathrm{NBMM}}$ (dots) with their 95% confidence intervals (solid vertical lines) for OTUs $j = 16, 34$, and 49. Inference on the covariate effects is different from that under the proposed. For example, domoic acid (DA) levels do not have significant effects on the mean counts for OTU $j = 16$ and 49 from **Figures 7A,C**. Comparing **Figures 7A,C** to **Figure 7**, the NBMM produces wider interval estimates for $\beta_{j,p}$. Histograms of the MLEs of $\beta_{j,p}$, $\hat{\beta}_{j,p}^{\mathrm{NBMM}}$ under the NBMM are shown in Supplementary Figure 14. The histograms are much dispersed than those under the proposed model shown in Supplementary Figure 13. Estimates $\hat{\beta}_{j,p}$ and $\hat{\beta}_{j,p}^{\mathrm{NBMM}}$ for all covariates are also compared in Supplementary Figure 15. From the figure, the NBMM yields extremely large or small values for $\hat{\beta}_{j,p}$ for some OTUs, possibly due to the convergence problem. The insert plots show that for regions of small values of $\hat{\beta}_{j,p}$, the estimates under the proposed are more shrunken toward zero than those under the NBMM, similar to the results in section 3.1. The overdispersion parameter estimates under the NBMM tend to be smaller than those under the proposed (shown in Supplementary Figure 12f), which may lead to different predictive distributions of OTU counts.

# 4. DISCUSSION AND CONCLUSIONS

In this paper, we developed a Bayesian semiparametric regression model for joint analysis of microbiome data. We formulated the mean counts of OTUs as a product of factors and built models for the factors. We utilized the regularizing priors with mean constraints to avoid possible idenfiability issues, and the process convolution model to capture the temporal dependence structure in the baseline mean abundance of an OTU. The flexible model developed for baseline abundance enables joint analysis of all OTUs in the data and allows borrowing information across OTUs, across samples, and across time points. The model produces accurate estimates of the baseline mean counts and yields improved estimates of the effects of the covariates. We incorporated the Laplace distribution, a sparsity inducing shrinkage prior for the coefficients and the proposed model produces sparse estimates that is more desirable when the problem is high-dimensional and covariates are highly

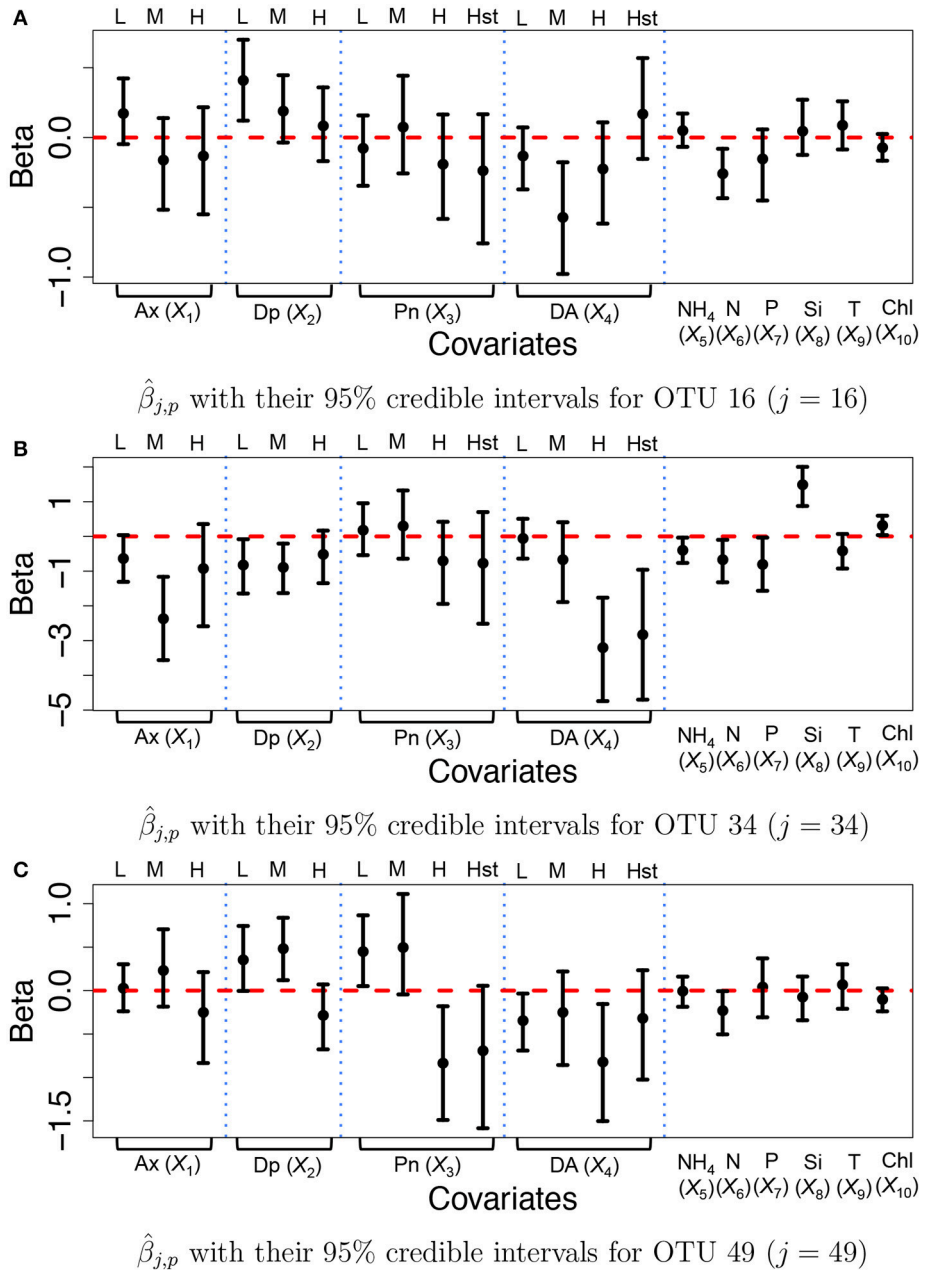


**FIGURE 6 |** Ocean microbiome data—proposed model. Posterior Inference on $\beta_j$ for some selected OTUs ($j = 16, 34, 49$). Dots represent the posterior means $\hat{\beta}_{j,p}$ of $\beta_{j,p}$. Each vertical line connects the lower bounds and the upper bounds of 95% credible intervals.

correlated. We compared the proposed model to a comparable frequentist model that does separate analyses for individual OTUs. The comparisons through the simulation study and real data analysis show better performance of the proposed model.

Although we focused on the analysis of NGS count data, the proposed model is quite general and can be applied for analyses of any count data. Future work will explore alternative approaches to model the effects of covariates on the mean counts. For example, one may consider a nonparametric model using linear combinations of basis functions (Kohn et al., 2001) to flexibly capture shape in the response function. In such a case, an elaborate development of the prior model may be needed to produce a robust inference since both the baseline mean counts and the covariate effects are nonparametrically modeled. Other possible extensions are to include a variable selection method such as a stochastic search variable selection (George and McCulloch, 1993) if it is reasonable to assume that some covariate effects are exactly zero, and to let coefficients vary over time if covariate effects evolve with time. For time varying coefficients, we may use the random walk process in Leybourne (1993) to

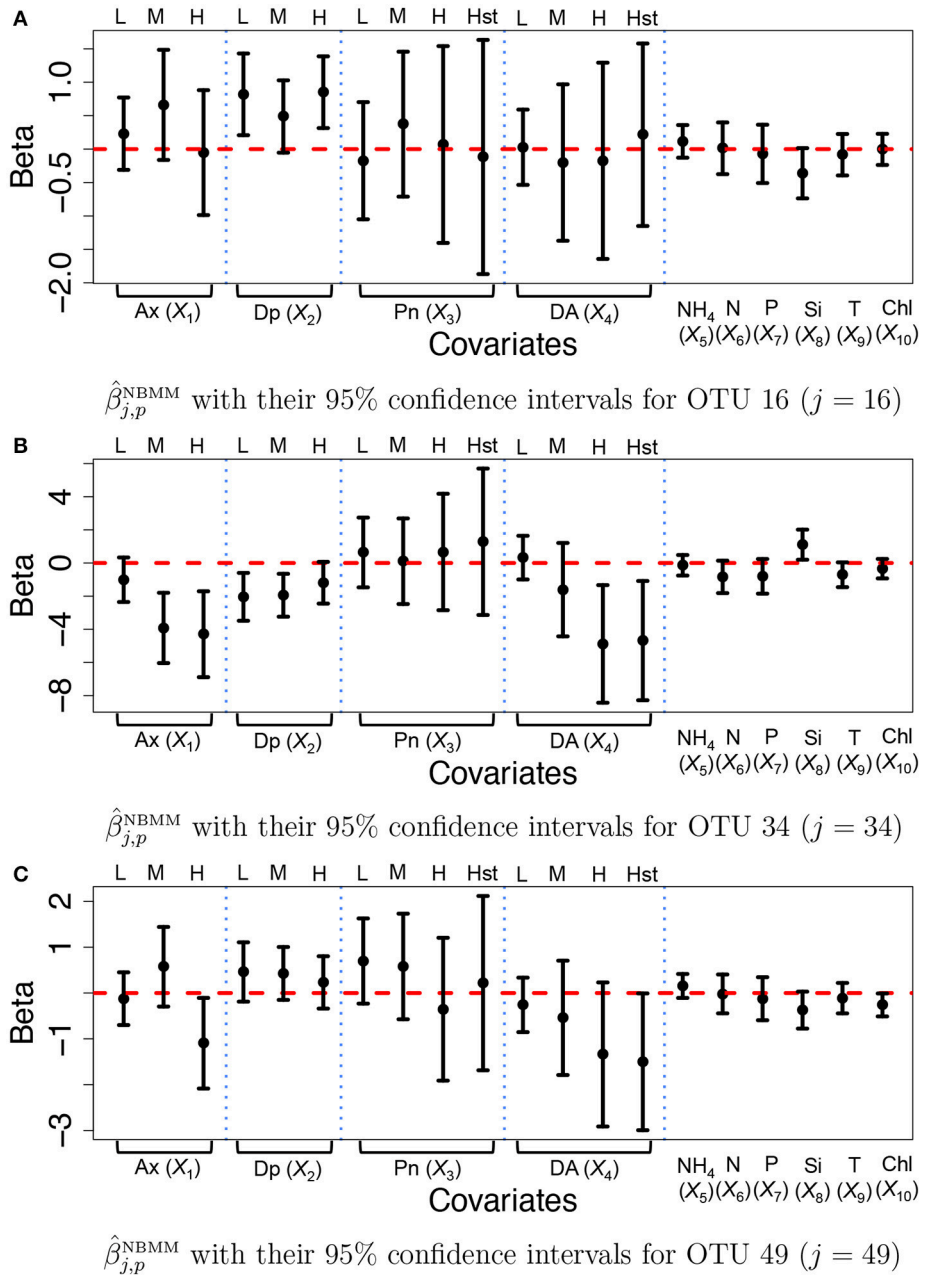


**FIGURE 7 |** Ocean microbiome data—NBMM. Inference on $\beta_j$ for some selected OTUs ($j = 16, 34, 49$) under the negative binomial mixed model (NBMM). Dots represent the maximum likelihood estimates $\hat{\beta}_{j,p}^{\mathrm{NBMM}}$ of $\beta_{j,p}$. Each vertical line connects the lower bounds and the upper bounds of 95% confidence intervals.

induce relationship between $\beta_{j,p,t-1}$ and $\beta_{j,p,t}$. Considering the high dimensionality in OTU data, posterior computation may need to be carefully handled. Also, prior information may be needed to produce sensible inference due to sparsity in data.

## AUTHOR CONTRIBUTIONS

JL developed the statistical model and conducted simulation studies and data analysis. She also prepared the first draft and led the collaboration with MS-M for statistical analysis. MS-M provided the ocean microbiome data, participated the statistical model development, provided biological interpretation of the resulting inference and edited the manuscript.

## FUNDING

This work was supported by NSF grant DMS-1662427 (JL) and NOAA-ECOHAB PROGRAM (Grant No. NA17NOS4780183,

ECOHAB #905) (MS-M and JL). 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.

## ACKNOWLEDGMENTS

We gratefully acknowledge Raphael Kudela who provided the environmental data in this study, Michael Kempnich and Sanjin Mehic for doing the water sampling and processing the water samples for DNA extraction.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.00522/full#supplementary-material

## REFERENCES

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

Banerjee, S. Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data. 2nd Edn.* Boca Raton, FL: CRC Press; Chapman & Hall.

Chen, E. Z., and Li, H. (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics* 32, 2611–2617. doi: 10.1093/bioinformatics/btw308

Cressie, N. (1992). Statistics for spatial data. *Terra Nova* 4, 613–617.

George, E. I., and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *J. Am. Stat. Assoc.* 88, 881–889.

Gibbons, S. M., Kearney, S. M., Smillie, C. S., and Alm, E. J. (2017). Two dynamic regimes in the human gut microbiome. *PLoS Comput. Biol.* 13:e1005364. doi: 10.1371/journal.pcbi.1005364

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the north atlantic ocean. *Environ. Ecol. Stat.* 5, 173–190.

Higdon, D. (2002). "Space and space-time modeling using process convolutions," in *Quantitative Methods for Current Environmental Issues*, eds C. W. Anderson, V. Barnett, P. C. Chatwin, and A. H. El-Shaarawi (London: Springer), 37–56.

Kohn, R., Smith, M., and Chan, D. (2001). Nonparametric regression using linear combinations of basis functions. *Stat. Comput.* 11, 313–322. doi: 10.1023/A:1011916902934

Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *Can. J. Stat.* 15, 209–225.

Leybourne, S. J. (1993). Estimation and testing of time-varying coefficient regression models in the presence of linear restrictions. *J. Forecast.* 12, 49–62.

Lee, H. K., Higdon, D. M., Calder, C. A., and Holloman, C. H. (2005). Efficient models for correlated data via convolutions of intrinsic processes. *Stat. Model.* 5, 53–74. doi: 10.1191/1471082X05st085oa

Li, Q., Guindani, M., Reich, B., Bondell, H., and Vannucci, M. (2017). A bayesian mixture model for clustering and selection of feature occurrence rates under mean constraints. *Stat. Anal. Data Mining* 10, 393–409. doi: 10.1002/sam.11350

Liang, W. W., and Lee, H. K. (2014). Sequential process convolution gaussian process models via particle learning. *Stat. Interface* 7, 465–475. doi: 10.4310/SII.2014.v7.n4.a4

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models, No. 37 in Monograph on Statistics and Applied Probability.* Boca Raton, FL: Chapman & Hall/CRC.

McMurdie, P. J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* 10:e1003531. doi: 10.1371/journal.pcbi.1003531

Needham, D. M., and Fuhrman, J. A. (2016). Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nat. Microbiol.* 1:16005. doi: 10.1038/nmicrobiol.2016.5

Park, T., and Casella, G. (2008). The bayesian lasso. *J. Am. Stat. Assoc.* 103, 681–686. doi: 10.1198/016214508000000337

Polson, N. G., and Scott, J. G. (2012). Local shrinkage rules, lévy processes and regularized regression. *J. R. Stat. Soc. Ser. B* 74, 287–311. doi: 10.1111/j.1467-9868.2011.01015.x

Robinson, M. D., and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* 23, 2881–2887. doi: 10.1093/bioinformatics/btm453

Romero, R., Hassan, S. S., Gajer, P., Tarca, A. L., Fadrosh, D. W., Nikita, L., et al. (2014). The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome* 2:4. doi: 10.1186/2049-2618-2-4

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. *Front. Microbiol.* 7:1433. doi: 10.3389/fmicb.2016.01433

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., et al. (2009). A core gut microbiome in obese and lean twins. *Nature* 457, 480–484. doi: 10.1038/nature07540

Witten, D. M. (2011). Classification and clustering of sequencing data using a poisson model. *Ann. Appl. Stat.* 5, 2493–2518. doi: 10.1214/11-AOAS493

Woo, P., Lau, S., Teng, J., Tse, H., and Yuen, K.-Y. (2008). Then and now: use of 16s rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. *Clin. Microbiol. Infect.* 14, 908–934. doi: 10.1111/j.1469-0691.2008.02070.x

Xiao, S. (2015). *Bayesian Nonparametric Modeling for Some Classes of Temporal Point Processes*. Ph.D. thesis, University of California, Santa Cruz.

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* 18:4. doi: 10.1186/s12859-016-1441-7

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.