UNIVERSITY OF CALIFORNIA
SANTA CRUZ

BAYESIAN HIERARCHICAL MODELS FOR COUNT DATA

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICAL SCIENCE

by

Kurtis Shuler

June 2020

The Dissertation of Kurtis Shuler
is approved:

Associate Professor Juhee Lee, Chair

Professor Athanasios Kottas

Professor Raquel Prado

Quentin Williams
Acting Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Bayesian Hierarchical Models for Count Data

by

Kurtis Shuler

This dissertation focuses on the development of methodology for the analysis of multivariate count responses. Such contexts present a number of unique modeling challenges that are not well handled by standard models for count data which have restrictive mean-variance and correlation structures. In addition to being high-dimensional, sparse and overdispersed, multivariate count data often exhibits complicated dependencies across categories and samples that must be accounted for in order to obtain accurate inference. Three Bayesian modeling strategies are presented to handle these challenges and produce accurate, interpretable inference with uncertainty quantification. The first model incorporates novel nonlocal priors for variable selection which outperform existing alternatives, and introduces a process convolutions sub-model to handle temporally dependent responses taken over uneven sampling intervals. The second applies Bayesian nonparametric (BNP) methods based on a dependent Dirichlet process mixture to flexibly model how category abundance levels and zero inflation are related to covariates. The BNP approach facilitates community level comparisons across experimental conditions through density estimates that provide additional insights over simple statistical tests or ordination analysis. The third model employs a directed acyclic graph (DAG) to identify related response categories. The graphical model does a better job uncovering network relationships than alternatives based on simple marginal correlations, and, unlike simpler count data models, handles cross-category dependence in a principled manner.

To my family, for their unwavering encouragement

Acknowledgments

People

I would like to thank my collaborators, Professor Sison-Mangus, Professor Chen and her student Dr. Verbanic, both for providing data and for their biological insights. I would also like to thank my committee members, Professor Kottas and Professor Prado, for their thoughtful comments that have strengthened my research. Most of all I would like to thank my advisor, Professor Lee, who I am deeply indebted to, for her support and intellectual contribution to this work.

Publications

Portions of this dissertation are taken from the previously published work Shuler *et al.* (2019a).

Chapter 1

Introduction

The focus of this work is the analysis of multivariate count responses, where each observation is a vector of integers representing the counts of some events of interest. These multivariate response vectors are typically collected into an observation matrix for analysis, with each row or column of the matrix consisting of one of the response vectors, and this matrix is used for downstream analysis. Because tabulating event counts coming from multiple categories is so often the natural sampling method, such data arises in a wide variety of settings. In applied ecology species/event counts are used to assess biodiversity and population dynamics (Johnson *et al.*, 2010; Richards, 2008). Natural language processing applications use counts of words or other tokens for topic modeling, sentiment analysis, and classification (Agarwal *et al.*, 2011; Blei *et al.*, 2003; McCallum *et al.*, 1998; Joachims, 1998). Read counts produced by RNA sequencing have been crucial in producing new insights in transcriptomics and genetics (Lowe *et al.*, 2017; Ozsolak and Milos, 2011; Robinson and Oshlack, 2010). The datasets analyzed in this work are taken from microbiome studies, where taxa counts are used to analyze the community of microbiata in some environment of interest.

Despite their ubiquity, adequately modeling multivariate count responses is

typically not straightforward. Simple count models can be inadequate because facets of the data generating processes and sample collection procedures complicate analysis. Count data often exhibit dependence across both the rows and the columns of the response matrix (Ren *et al.*, 2017b). Across the samples, this dependence can arise due to spatial or temporal structure, or through correlation induced by batch/group-effects (Chen and Li, 2016; Li, 2015; Xu *et al.*, 2017). These effects can manifest via a number of different mechanisms, such as subsets of the samples being produced by different individuals or labs; or through multiple samples being taken from the same subject. Dependence is often observed across the count categories as well (Chen and Li, 2016; Mandal *et al.*, 2015; Ren *et al.*, 2017a; Weiss *et al.*, 2017). As a result, analyzing the counts from each category separately rather than jointly may result in reduced statistical power, inferior uncertainty quantification, and, under the wrong assumptions, biased parameter estimates (Ren *et al.*, 2017b).

In many count data settings the counts do not reflect absolute abundance, but rather are an abundance measure relative to the other counts (Grantham *et al.*, 2017). This is the case for natural language processing applications using "bag-of-words" assumptions which simply count the occurrences of each word or token in a document (Blei *et al.*, 2003; Joachims, 1998; Zhang *et al.*, 2010). In bag-of-words models the features' observed counts are a function of the document length (Li *et al.*, 2017). Similarly, the observed counts for RNA sequencing data is a function of the effort put into the sequencing procedure (Li *et al.*, 2017; Robinson and Oshlack, 2010). To make them interpretable, such counts typically must be normalized before analysis (Li *et al.*, 2012, 2017; McMurdie and Holmes, 2014; Robinson and Oshlack, 2010; Weiss *et al.*, 2017). The choice of normalization procedure is not simple, and can have a significant impact on the results of count

data analysis (McMurdie and Holmes, 2014; Rivera-Pinto *et al.*, 2018; Weiss *et al.*, 2017). Higher count values in these settings may not reflect higher certainty about the abundances across categories or samples, but rather may be an artifact of the effort put into the sampling procedure. Conversely, many simple count models have restrictive mean-variance structures, which, if ignored, may have inappropriate implications for uncertainty quantification and testing (Li *et al.*, 2017; Robinson and Smyth, 2007; Zhang *et al.*, 2017b).

Count data is often sparse, exhibiting far more zero counts than would be expected under standard count distributions (Lee *et al.*, 2018; Jonsson *et al.*, 2018; Xu *et al.*, 2015; Zhang *et al.*, 2017a). Count data is also often overdispersed, with the count categories having higher variance than common Poisson or multinomial models can accommodate (Jonsson *et al.*, 2018; Zhang *et al.*, 2017a). This simultaneous sparsity and overdispersion is in part why naively transforming count data using log-transformations and modeling the data on a continuous scale is not adequate to obtain sensible inference (O'Hara and Kotze, 2010). Zeros must be replaced before the log-transformation, often by adding a small value known as a pseudocount to the entire dataset or to the zero observations. There is no clear consensus on how to pick the pseudocount, and its influence on the results can be non-trivial (Weiss *et al.*, 2017).

This work seeks to address these challenges through the development of Bayesian models for count data with multivariate count responses, which are motivated by microbiome studies. Methods are developed to handle row/column dependence in the response matrix through model components that induce temporal and group dependence across samples, as well as dependence across count categories. Careful consideration is given to proper normalization procedures, as well as to handling counts exhibiting zero inflation and overdispersion. These methods are developed

in regression frameworks so covariate effects can be estimated, and special attention is given to the problem of variable selection in these contexts. The next section briefly describes the motivation for this dissertation, and §1.2 outlines its main contributions and organization.

1.1 Motivation and Literature Review

The advent of widely available and, relative to the past, affordable high-throughput sequencing (HTS) technology has led to a surge in interest in microbiome studies (Clooney *et al.*, 2016; Reuter *et al.*, 2015). Ambitious, high visibility initiatives like the Human Microbiome Project continue to fuel this interest and have underscored that microbiomes study may be the key to answering crucial open questions in ecology, biology, and medicine (Knight *et al.*, 2017; Turnbaugh *et al.*, 2007). In microbiome studies a sample of genetic material is taken from an environment of interest and profiled to characterize the microbiata present in that environment. Often, this profile is produced via 16S ribosomal RNA (rRNA) amplicon sequencing of the genetic material from the sample. The 16S rRNA gene is widespread and contains both highly conserved regions suitable for broad-spectrum polymerase chain reaction (PCR) primer pairs and fast evolving regions which can be used to classify the microbiata present in a sample (Sambo *et al.*, 2018). The foundation of these analyses consists of grouping together similar genetic sequences to form Operational Taxonomic Units (OTUs) (Buza *et al.*, 2019). Counts of the similar sequence reads are used as a proxy for the microbiota present in the sample and their abundance (Kurtz *et al.*, 2015; Wadsworth *et al.*, 2017). Because of the nature of the data generating process, the total number of reads for each OTU cannot be used as an absolute measure of the abundance of that OTU; rather, OTU counts only reflect relative abundance and must be normal-

ized before comparisons can be made across samples (Li *et al.*, 2012, 2017; Weiss *et al.*, 2016, 2017). After pre-processing, the collection of OTU counts for each sample are collected into a multivariate count response vector, and these vectors are organized into a matrix called an OTU table for downstream analysis. Table 1.1 illustrates the layout of such a table. For an introduction to statistical analysis of microbiome samples and a more detailed overview of how they are collected and sequenced see Xia *et al.* (2018).

	OTU 1	OTU 2	...	OTU J	Total
Sample 1	4,928	55	...	0	26,819
Sample 2	2,667	21	...	12	41,167
⋮	⋮	⋮	⋱	⋮	⋮
Sample n	119	0	...	2	1,743

Table 1.1: Example of an OTU table. Each cell represents the count of a particular OTU for that sample. The sum total, shown in the right margin, is commonly used in procedures to normalize the counts across samples.

Broadly speaking, the questions that researchers seek to answer using this data can be organized into three categories. The first concerns questions about global interactions between the microbiome and some environmental factor (e.g. phenotype, experimental condition, etc.). These studies may seek to address how microbial diversity varies with some covariate, or to perform clustering on microbiomes based on community composition (Arumugam *et al.*, 2011; Gilbert *et al.*, 2016; Holmes *et al.*, 2012; Lewis *et al.*, 2015). Second, researchers may be interested in local interactions like which taxa are associated with a particular outcome; such as if certain taxa are present in a disease state (Frank *et al.*, 2007; Gilbert *et al.*, 2016; Kostic *et al.*, 2012; Ley *et al.*, 2006; Mendes *et al.*, 2011; Scher *et al.*, 2013). Third, questions may be posed regarding the interactions of taxa within a microbiome (Levy and Borenstein, 2013; Louca *et al.*, 2016; Zelezniak *et al.*, 2015). Some taxa may have microbe to microbe interactions, and knowing

if and how taxa abundances fluctuate in tandem may help address key ecological questions about the how the community functions.

The approaches taken to answer these questions range from classical statistical tests to sophisticated regression models designed to handle the peculiar challenges of modeling multivariate count data. Testing procedures, often permutation tests, are commonly used to address questions of global interactions in the microbiome (Anderson, 2001; Mann and Whitney, 1947; McArdle and Anderson, 2001; Zhao *et al.*, 2015); or, for a more qualitative view, ordination analysis, like principal coordinate analysis (PCoA, also known as multidimensional scaling), to assess the relative similarity of different microbial communities (Arumugam *et al.*, 2011; Gower, 1966; Oksanen *et al.*, 2007; Ren *et al.*, 2017a). Parametric tests are used as for this purpose as well, such as in the popular DESeq2 software package, which is designed detect changes across experimental conditions using count data modeled with the negative binomial distribution (Love *et al.*, 2014). To answer questions about how covariates are related to taxa abundance many of the more sophisticated statistical approaches involve fitting a generalized linear model to the counts, or some transformation thereof. The popular software package edgeR, for example, estimates the effects of environmental factors on abundance using log-linear models with a negative binomial likelihood for the counts (Robinson *et al.*, 2010). On the Bayesian side, Wadsworth *et al.* (2017) suggest a Dirchlet-Multinomial regression model with a spike-and-slab prior to identify significant relationships between taxa abundance and environmental factors. The use of multinomial likelihoods can simplify the problem of normalization across samples, but it also induces a negative correlation across the count categories, which is often an unrealistic assumption. To address this limitation, Grantham *et al.* (2017) propose a mixed-effects model with a multinomial likelihood that uses a spike-

and-slab prior for variable selection, but also incorporate additional structure to allow for positive cross-taxa correlations. Ren *et al.* (2017b) propose a model with a Dirichlet process prior on the marginal taxa compositions, such that the number of taxa present in a sample are not constrained a-priori, and use the composition vectors to make community level comparisons. Lee *et al.* (2018) model the counts directly using a zero inflated Poisson distribution, with spike-and-slab priors for variable selection and random effects to account for dependence across taxa. Xu *et al.* (2017) propose a zero inflated negative binomial model in a longitudinal context, with random effects to control for dependence structure arising from samples being taken from related individuals (e.g. individuals from the same family). Lee and Sison-Mangus (2018) develop a negative binomial regression model for the counts in a temporal context, using a Laplace shrinkage prior on the regression coefficients to improve their estimation in high-dimensional settings.

1.2 Contribution and Organization

The contribution of this work is the development of Bayesian models for the analysis of multivariate count responses. These models address challenging aspects of count data analysis. Flexible normalization procedures for the responses are proposed, and careful adjustments are made to account for dependence structures within and across samples. Methods to handle excess zero inflation and overdispersed counts are introduced. Because these models were developed in the context of microbiome analysis, special attention is given to features that help address the questions that biologist seek to answer in these settings. Regression models are proposed to answer questions about global interactions between the microbiome and environmental factors. OTU abundances are modeled jointly to facilitate borrowing strength across OTUs, and the relationships between individ-

ual OTU abundances and covariates are made available as well. Variable selection procedures are presented which offer superior performance over existing methods. Bayesian nonparametric approaches which offer a more nuanced way to compare microbial communities than simple statistical tests or ordination methods are introduced. Inference about taxa interactions is given by a graphical model which provides a much clearer view of networked taxa than current state of the art methods which rely on analyzing correlations across taxa abundances.

Chapter 2 describes the development of a Bayesian sparse multivariate regression method to model the relationship between microbe abundance and environmental factors for microbiome data. OTU abundance counts are modeled with a negative binomial distribution that relates covariates to the counts through regression. Relevant covariates and their effect directions are efficiently identified through the construction of asymmetric nonlocal priors for the regression coefficients which extend conventional nonlocal priors. A hierarchical model is built which facilitates pooling of information across OTUs and produces parsimonious results with improved accuracy. Simulation studies 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.

Chapter 3 introduces a Bayesian nonparametric regression model with zero inflation to analyze complex multivariate count data from microbiome studies. The baseline counts of taxa in samples are carefully constructed to obtain improved

estimates of differential abundance. A Bayesian nonparametric approach flexibly models microbial associations with covariates such as environmental factors and clinical characteristics. Importantly, the approach provides straightforward community-level insights into how characteristics of microbial communities such as taxa richness and diversity are related to covariates. Simulation studies show the model outperforms popular alternatives. The model is then applied to a chronic wound microbiome dataset, comparing the microbial communities present in chronic wounds versus in healthy skin

For many microbiome analyses a key research task is to understand the microbiome as a whole, whose structure and function can be heavily affected by microbe-microbe interactions and interactions with its environment. Chapter 4 presents a Bayesian regression model with a graph (BRM-G) for count data to provide a holistic understanding of complex microbial communities. The model employs a directed acyclic graph (DAG) to represent microbe-microbe interactions. Inference is summarized through moralization of the DAG and inferred interactions between microbes can be further validated through additional experiments. A regression component is included to provide insights into how environmental factors and experimental conditions are related to taxa abundance. In addition, the model simultaneously accounts for different sample sequencing depths through model based normalization. A simulation study shows that, compared to BRM-G, alternative methods that do not incorporate the interactions between microbes or are based on simple marginal correlations among microbes perform poorly in uncovering the complex interplay among microbial taxa. The model is also applied to a microbiome dataset to identify groups of related taxa in chronic wounds and healthy skin in human subjects.

Finally, chapter 5 summarizes the main contributions from chapter 2 to 4, and

concludes with some possible future extensions.

Chapter 2

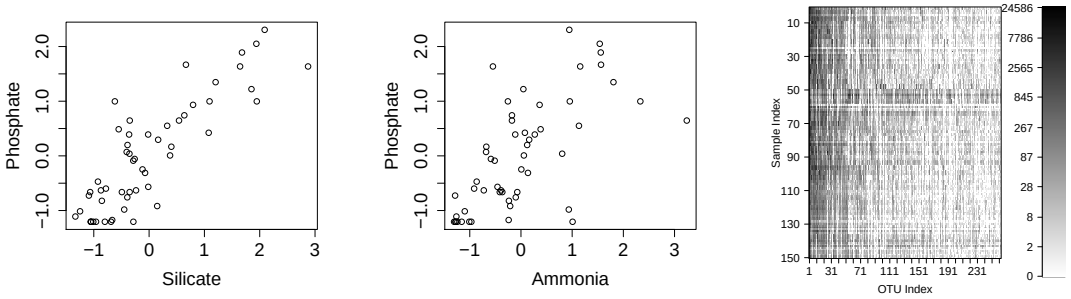
Bayesian Sparse Multivariate Regression with Asymmetric Nonlocal Priors for Microbiome Data Analysis

2.1 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 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 2.1(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 for 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



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

Figure 2.1: [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.

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.*, 2017a). 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.* (2017a)). 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 regres-

sion 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 2.2 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 settings (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 chapter is organized as follows. §2.2 describes the proposed ANLP-SB model. §2.3 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. §2.4 summarizes analyses of the ocean microbiome dataset, and we close with a discussion in §2.5.

2.2 Probability Model

2.2.1 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,

$$Y_{tkj} \mid \boldsymbol{x}_t, \mu_{tkj}, s_j \overset{indep}{\sim} \text{NB}(\mu_{tkj}(\boldsymbol{x}_t), s_j),\ j = 1, \ldots, J. \tag{2.1}$$

The model in (2.1) 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.* (2017a)). In the next section we develop models for $\boldsymbol{\beta}_j$, g_{tkj} and s_j.

2.2.2 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

$$\beta_{jp} \mid \boldsymbol{\pi}_p^\star, \sigma_p^2, \iota_p \overset{indep}{\sim} \pi_{p0}^\star \mathbb{I}(\beta_{jp} = 0) + \pi_{p1}^\star \frac{\phi(\beta_{jp}/\sigma_p)}{\sigma_p\{1 - \Phi(\iota_p)\}} \mathbb{I}\left(\frac{\beta_{jp}}{\sigma_p} > \iota_p\right) + \pi_{p2}^\star \frac{\phi(\beta_{jp}/\sigma_p)}{\sigma_p \Phi(-\iota_p)} \mathbb{I}\left(\frac{\beta_{jp}}{\sigma_p} < -\iota_p\right), \tag{2.2}$$

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 (2.2) 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 (2.2) 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 (2.2). 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 (2.2) 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 2.2.1. *Consider a model with $\boldsymbol{\gamma}_j$ for OTU j. Let $\boldsymbol{\beta}_j^\star$ denote a vector of*

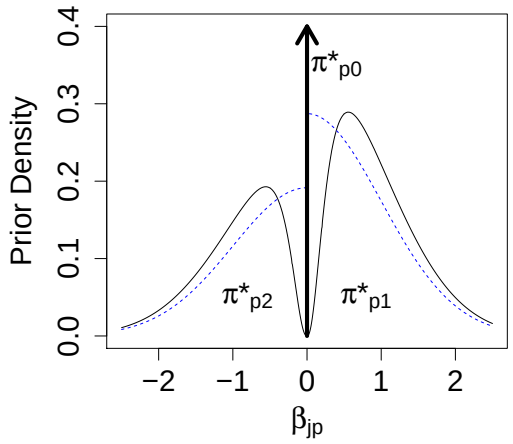


Figure 2.2: Plot of the asymmetric nonlocal prior density function $\mathrm{P}(\beta_{jp}^{\star} \mid \boldsymbol{\pi}^{\star})$ (black, solid) and its corresponding asymmetric local prior density function (blue, dotted). $\boldsymbol{\pi}_p^{\star} = (0.4, 0.36, 0.24)$ and $\iota_p \sim \text{Gamma}(2.5, 10)$ are assumed.

β_{jp} with $\gamma_{jp} \neq 0$ only. Given $\boldsymbol{\gamma}_j$, the joint prior of $\boldsymbol{\beta}_j^{\star}$ can be written as

$$P(\boldsymbol{\beta}_j^{\star} \mid \boldsymbol{\gamma}_j, \boldsymbol{\delta}, \boldsymbol{\iota}) = \prod_{p=1;\gamma_{jp}\neq 0}^{P} \left\{ \pi_{p1} \frac{\phi(\beta_{jp}/\sigma_p)}{\sigma_p\{1-\Phi(\iota_p)\}} \mathbb{I}\left(\frac{\beta_{jp}}{\sigma_p} > \iota_p\right) + \pi_{p2} \frac{\phi(\beta_{jp}/\sigma_p)}{\sigma_p \Phi(-\iota_p)} \mathbb{I}\left(\frac{\beta_{jp}}{\sigma_p} < -\iota_p\right) \right\}, \quad (2.3)$$

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)

$$P^L(\boldsymbol{\beta}_j^{\star} \mid \boldsymbol{\gamma}_j, \boldsymbol{\delta}) = \prod_{p=1;\gamma_{jp}\neq 0}^{P} \left\{ \pi_{p1} \frac{\phi(\beta_{jp}/\sigma_p)}{\sigma_p\{1-\Phi(0)\}} \mathbb{I}\left(\frac{\beta_{jp}}{\sigma_p} > 0\right) + \pi_{p2} \frac{\phi(\beta_{jp}/\sigma_p)}{\sigma_p \Phi(0)} \mathbb{I}\left(\frac{\beta_{jp}}{\sigma_p} < 0\right) \right\}, \quad (2.4)$$

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 (2.3) *and* (2.4) *asymmetric nonlocal priors (ANLPs) and asymmetric local priors (ALPs), respectively.*

Figure 2.2 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 $\beta^{\star}_{jp}$. 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 $d(\boldsymbol{\beta}^{\star}_{j})$ 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 (2.2), $\pi^{\star}_{p0}$ serves as the rate at which the coefficients β_{jp} are exactly zero in the J regression problems. We let $\pi^{\star}_{p0} \overset{iid}{\sim} \text{Be}(a_{\pi 0}, b_{\pi 0})$. 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 $\boldsymbol{\pi}^{\star}_{p}$ 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 $\text{E}((1-\pi^{\star}_{p0})/\pi^{\star}_{p0})$ 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 $\sigma^2_p \overset{iid}{\sim} \text{IG}(a_\sigma, b_\sigma)$ with fixed a_σ and b_σ. Parameters σ^2_p, $\boldsymbol{\pi}^{\star}_{p}$ 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^2_p 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 struc-

ture 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};

$$r_{tk} \overset{iid}{\sim} \sum_{\ell=1}^{L^r} \psi_\ell^r \left\{ w_\ell^r \, \mathrm{N}(\eta_\ell^r, u_r^2) + (1 - w_\ell^r) \, \mathrm{N}\left(\frac{\upsilon_r - w_\ell^r \eta_\ell^r}{1 - w_\ell^r}, u_r^2 \right) \right\}, \tag{2.5}$$

$$\alpha_{0j} \overset{iid}{\sim} \sum_{\ell=1}^{L^\alpha} \psi_\ell^\alpha \left\{ w_\ell^\alpha \, \mathrm{N}(\eta_\ell^\alpha, u_\alpha^2) + (1 - w_\ell^\alpha) \, \mathrm{N}\left(\frac{\upsilon_\alpha - w_\ell^\alpha \eta_\ell^\alpha}{1 - w_\ell^\alpha}, u_\alpha^2 \right) \right\}, \tag{2.6}$$

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 (2.5) and (2.6) 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, 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 §2.3. 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 since the number of bacteria is known to depend on the number of bacteria at previous time points. 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, $\theta_{mj} \overset{iid}{\sim} \mathrm{N}(0, \tau_j^2)$, with $\tau_j^2 \overset{iid}{\sim} \mathrm{IG}(a_\tau, b_\tau)$.

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

2.2.3 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 (2.5) and (2.6), 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

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

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 §2.3 and §2.4. Details of the posterior simulation are in Appendix §A.1. In the appendix, 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`.

2.3 Simulation Studies

2.3.1 Simulation 1

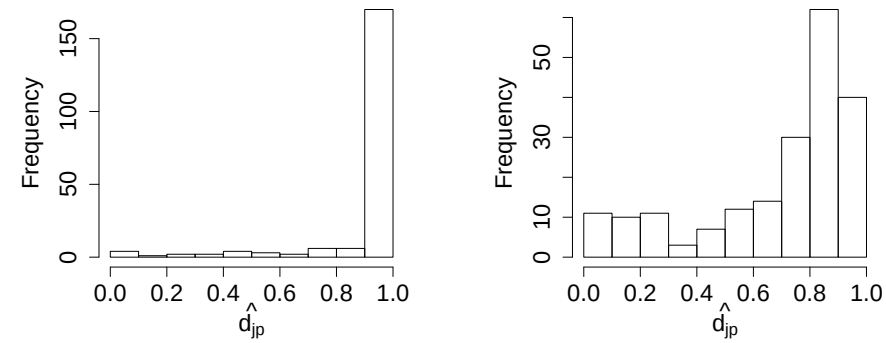
Data Simulation We performed simulation studies to assess the performance of the proposed ANLP-SB model and compared it to alternative models. We assumed $J = 200$ OTUs. We used time points t_i, $i = 1, \dots n$, the number of replicates K_i and some covariates from the ocean microbiome dataset described in §2.4. 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 §2.2. 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, $\alpha_{tj}^{\text{TR}} = A_j \sin\left(\frac{2\pi}{T} h_{ja} t_i - a_j\right) + B_j \sin\left(\frac{2\pi}{T} h_{jb} t_i - b_j\right)$, $0 \leq t \leq T$. 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 $s_j^{\text{TR}} \overset{iid}{\sim} \text{Log-Normal}\,(-1/2, 1/10^2)$. Finally, OTU counts were drawn from $Y_{tkj} \mid \mu_{tkj}^{\text{TR}}, s_j^{\text{TR}} \overset{indep}{\sim} \text{NB}(\mu_{tkj}^{\text{TR}}(\boldsymbol{x}_t),\, s_j^{\text{TR}})$,

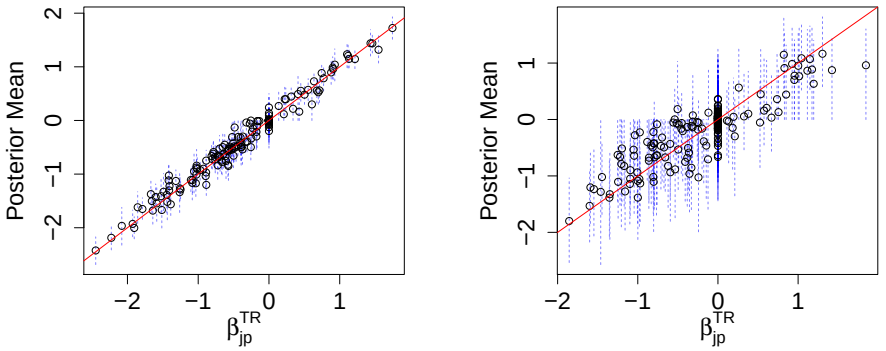
where $\log(\mu^{\mathrm{TR}}_{tkj}(\boldsymbol{x}_t)) = r^{\mathrm{TR}}_{tk} + \alpha^{\mathrm{TR}}_{0j} + \alpha^{\mathrm{TR}}_{tj} + \boldsymbol{x}'_t\boldsymbol{\beta}^{\mathrm{TR}}_j$.

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 $\boldsymbol{a}^r_\phi = \mathbf{1}$, $a^r_w = 0.5$, $b^r_w = 0.5$, $u^2_r = 0.1$, $b^2_{\eta^r} = 0.3$, $\boldsymbol{a}^\alpha_\psi = \mathbf{1}$, $a^\alpha_w = 0.5$, $b^\alpha_w = 0.5$ and $b^2_{\eta^\alpha} = 1$, 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 $a_h = -10$, $b^2_h = 100$, $a_\kappa = 10^{-5}$ and $b_\kappa = 10^{-5}$. We initialized θ_{mj} and β_{jp} using observed y_{tkj}. We generated initial values for σ^2_p 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 §A.2.

Figure 2.3(a) and (b) show histograms of posterior estimates of $\hat{d}_{jp} = \hat{\mathrm{P}}(\gamma_{jp} = \gamma^{\mathrm{TR}}_{jp} \mid \boldsymbol{Y})$, 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 2.3(c) and (d) compare posterior mean estimates $\hat{\beta}_{jp}$ of β_{jp} to their true values β^{TR}_{jp} 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^{TR}_{jp} with narrower interval estimates for the continuous covariates. Supplementary Figures A.1 and A.2 show histograms of $\hat{d}_{jp}$ and plots of $\hat{\beta}_{jp}$ versus β^{TR}_{jp} for all covariates. We next compare posterior esti-



(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 2.3: [Simulation 1] Panels (a) and (b): Histograms of the posterior estimates of $\hat{d}_{jp} = \hat{\mathrm{P}}(\gamma_{jp} = \gamma_{jp}^{\mathrm{TR}})$ 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.

mates $\hat{g}_{tkj}$ of the baseline mean counts to their true values. Supplementary Figure A.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 A.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 §A.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 (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

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

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 $BS = \frac{1}{J \times P} \sum_{jpq} \left(\hat{z}_{jpq} - \mathbb{I}(\gamma_{jp}^{\mathrm{TR}} = q) \right)^2 \in [0, 1]$, 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 proportion of true positives among "selected" covariates (also called precision) and the proportion of "selected" covariates among

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

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 2.1(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 (2.4) (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 $\beta_{jp} \mid \gamma_{jp} = 1 \overset{indep}{\sim} \text{N}(0, \sigma_p^2)$ 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.* (2017a), 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 2.1(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, $\sqrt{\sum_{jp}(\hat{\beta}_{jp} - \beta^{\mathrm{TR}}_{jp})^2/(100JP)}$. Columns 1-3 of Table 2.1(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 2.1(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.

2.3.2 Simulations 2 and 3

We conducted additional simulations for further examination of the proposed ANLP-SB model. The simulation setup for Simulations 2 and 3 is similar to Simulation 1's setup, including the specification for $\boldsymbol{x}_t$, r^{TR}_{tk} and $\alpha^{\mathrm{TR}}_{0j}$. Simulation 2 was

Model	MCC	ACC	AUC	Brier-Score	F_1
ANLP-SB	**0.613** (0.044)	**0.802** (0.021)	**0.886** (0.018)	**0.286** (0.031)	**0.788** (0.021)
ALP-SB	0.294 (0.042)	0.606 (0.029)	0.781 (0.025)	0.547 (0.047)	0.710 (0.026)
SLP-SB	0.288 (0.039)	0.604 (0.028)	0.775 (0.023)	–	0.708 (0.026)
BayesReg	0.530 (0.035)	0.741 (0.023)	0.799 (0.019)	–	0.676 (0.024)
edgeR-L	-0.003 (0.031)	0.498 (0.016)	0.499 (0.020)	–	0.442 (0.027)
edgeR-Q	0.002 (0.031)	0.501 (0.016)	0.502 (0.019)	–	0.474 (0.025)
BhGLM	0.231 (0.043)	0.602 (0.024)	0.635 (0.026)	–	0.491 (0.032)

(a) Variable Selection

Model	RMSE			DIC	LPML
	β_{jp}	$\pi^{\star}_{p0}$	g_{tkj}		
ANLP-SB	**0.280** (0.023)	**0.088** (0.025)	**0.322** (0.030)	**241,947** (6,214)	**-4.036** (0.104)
ALP-SB	0.297 (0.017)	0.282 (0.031)	0.339 (0.015)	242,030 (6,210)	-4.038 (0.104)
SLP-SB	0.303 (0.016)	0.281 (0.031)	0.350 (0.018)	242,058 (6,209)	-4.039 (0.104)
BayesReg	0.304 (0.016)	–	0.355 (0.029)	242,177 (6,219)	-4.044 (0.104)
edgeR-L	0.870 (0.032)	–	–	–	–
edgeR-Q	0.861 (0.028)	–	–	–	–
BhGLM	0.976 (0.063)	–	–	–	–

(b) Parameter Estimation and Model Fit

Table 2.2: [Simulation 2: Comparison] Performance metric averages over 100 simulated datasets with standard deviations in parenthesis. The best performances are in bold.

setup to investigate the models' performance when the regression coefficients for irrelevant covariates are not exactly zero. In particular, we let $\beta_{jp}^{\text{TR}} \overset{indep}{\sim} \text{N}(0, (\iota_p/6)^2)$ for covariates with $\gamma_{jp}^{\text{TR}} \neq 0$, giving these covariates negligible but non-zero effects on OTU abundance. Recall that the model assumes $\beta_{jp}^{\text{TR}} = 0$ when $\gamma_{jp}^{\text{TR}} = 0$. For Simulation 3, we let $\alpha_{tj}^{\text{TR}} \overset{indep}{\sim} \text{N}(0, (2/3)^2)$; that is, no temporal dependence in the simulation truth was assumed. Tables 2.2 and 2.3 show performance metrics for Simulations 2 and 3, respectively. In both simulations, ANLP-SB outperforms the competing models, especially with regard to variable selection. ANLP-SB performs notably better than the other Bayesian models in terms of the RMSE of $\pi^{\star}_{p0}$, MCC, AUC, and Brier score. The four Bayesian models have similar performance for parameter estimation based on the RMSE and similar model fit based on DIC and LPML.

Model	MCC	ACC	AUC	Brier-Score	F_1
ANLP-SB	**0.461** (0.038)	**0.730** (0.019)	**0.826** (0.016)	**0.393** (0.027)	**0.739** (0.018)
ALP-SB	0.247 (0.037)	0.587 (0.028)	0.743 (0.021)	0.609 (0.043)	0.698 (0.027)
SLP-SB	0.240 (0.037)	0.584 (0.028)	0.736 (0.019)	–	0.696 (0.027)
BayesReg	0.450 (0.035)	0.711 (0.022)	0.759 (0.020)	–	0.652 (0.024)
edgeR-L	-0.003 (0.033)	0.498 (0.022)	0.498 (0.018)	–	0.328 (0.034)
edgeR-Q	0.001 (0.031)	0.499 (0.023)	0.502 (0.017)	–	0.282 (0.029)
BhGLM	0.389 (0.064)	0.665 (0.029)	0.732 (0.038)	–	0.549 (0.039)

(a) Variable Selection

Model	RMSE			DIC	LPML
	β_{jp}	$\pi^{\star}_{p0}$	g_{tkj}		
ANLP-SB	0.393 (0.021)	**0.102** (0.030)	0.716 (0.014)	**247,071** (6,073)	**-4.124** (0.101)
ALP-SB	**0.390** (0.019)	0.285 (0.031)	0.712 (0.012)	247,078 (6,070)	**-4.124** (0.101)
SLP-SB	0.393 (0.016)	0.285 (0.031)	**0.703** (0.011)	247,089 (6,073)	-4.125 (0.101)
BayesReg	**0.390** (0.016)	–	0.707 (0.015)	247,742 (6,089)	-4.140 (0.102)
edgeR-L	0.487 (0.014)	–	–	–	–
edgeR-Q	0.491 (0.014)	–	–	–	–
BhGLM	0.526 (0.062)	–	–	–	–

(b) Parameter Estimation and Model Fit

Table 2.3: [Simulation 3: Comparison] Performance metric averages over 100 simulated datasets with standard deviations in parenthesis. The best performances are in bold.

	J	P	n	N	Run-time
Simulation 1	200	10	54	150	9
Simulation 2	200	10	54	150	8
Simulation 3	200	10	54	150	9
Simulation 4	400	10	50	100	12
Simulation 5	400	20	50	100	16
Simulation 6	400	10	100	200	23
Simulation 7	400	20	100	200	29
Simulation 8	200	50	100	200	22

Table 2.4: Simulation setups (J, P, n, N) for Simulations 1–8. The run-times (in minutes) for 1,000 MCMC iterations are reported in the last column.

2.3.3 Simulations 4–8

We performed Simulations 4–8 to examine the models' performance in higher dimensional settings by using different numbers of OTUs (J), covariates (P), and samples (N). We fixed $K = 2$ and used values of (J, P, N) in Table 2.4. For all

Model	MCC	ACC	AUC	Brier-Score	F_1
ANLP-SB	**0.703** (0.137)	**0.853** (0.073)	**0.925** (0.062)	**0.253** (0.141)	**0.851** (0.076)
ALP-SB	0.298 (0.093)	0.595 (0.128)	0.785 (0.025)	0.647 (0.193)	0.687 (0.127)
SLP-SB	0.298 (0.095)	0.588 (0.135)	0.799 (0.026)	–	0.687 (0.128)
BayesReg	0.589 (0.078)	0.788 (0.059)	0.831 (0.027)	–	0.744 (0.032)
edgeR-L	-0.043 (0.098)	0.501 (0.087)	0.475 (0.062)	–	0.301 (0.106)
edgeR-Q	-0.038 (0.113)	0.501 (0.081)	0.477 (0.071)	–	0.335 (0.109)
BhGLM	0.337 (0.054)	0.621 (0.105)	0.720 (0.042)	–	0.386 (0.058)

(a) Variable Selection

Model	RMSE			DIC	LPML
	β_{jp}	$\pi^{\star}_{p0}$	g_{tkj}		
ANLP-SB	**0.125** (0.028)	**0.246** (0.130)	**0.368** (0.071)	**359,138** (8,220)	**-4.506** (0.103)
ALP-SB	0.157 (0.020)	0.474 (0.113)	0.476 (0.099)	359,381 (8,188)	-4.512 (0.103)
SLP-SB	0.176 (0.035)	0.462 (0.112)	0.633 (0.187)	359,515 (8,187)	-4.514 (0.103)
BayesReg	0.217 (0.037)	–	0.815 (0.214)	361,391 (8,196)	-4.545 (0.102)
edgeR-L	0.341 (0.031)	–	–	–	–
edgeR-Q	0.313 (0.031)	–	–	–	–
BhGLM	0.357 (0.044)	–	–	–	–

(b) Parameter Estimation and Model Fit

Table 2.5: [Simulation 4: Comparison] Results from the simulation study with $N = 100$ samples taken at $n = 50$ time points with $J = 400$ OTUs and $P = 10$ covariates.

simulations, we let $\boldsymbol{x}_t \overset{iid}{\sim} \mathrm{N}(\mathbf{0}, \Sigma)$, where the diagonal of $\boldsymbol{\Sigma}$ was 1, and the off diagonals were 1/4. Similar to Simulations 2-3, we used r'_{tk} and α'_{0j} computed from the ocean microbiome data to set r^{TR}_{tk} and $\alpha^{\mathrm{TR}}_{0j}$; let $r^{\mathrm{TR}}_{tk} = \log(r''_{tk} + \epsilon^{r}_{tk})$ and $\alpha^{\mathrm{TR}}_{0j} = \log(\alpha''_{0j} + \epsilon^{\alpha}_{j})$, where r''_{tk} and α''_{0j} were random draws from $\{r'_{tk}\}$ and $\left\{\alpha'_{0j}\right\}$, respectively, $\epsilon^{r}_{tk} \overset{iid}{\sim} \mathrm{N}(0, 10^{-5})$ and $\epsilon^{\alpha}_{j} \overset{iid}{\sim} \mathrm{N}(0, 1/10)$. We let $\pi^{\star,\mathrm{TR}}_{p0} = 0.95$ or 0.05 with equal probability, and $\pi^{\mathrm{TR}}_{p1} \overset{iid}{\sim} \mathrm{Be}(5, 10)$. We let $\gamma^{\mathrm{TR}}_{jp} = 0$, 1 or 2 with probabilities, $\boldsymbol{\pi}^{\star,\mathrm{TR}}_{p} = (\pi^{\star,\mathrm{TR}}_{p0}, (1 - \pi^{\star,\mathrm{TR}}_{p0})\pi^{\mathrm{TR}}_{p1}, (1 - \pi^{\star,\mathrm{TR}}_{p0})(1 - \pi^{\mathrm{TR}}_{p1}))$. We generated $\sigma^{2^{\mathrm{TR}}}_{p} \overset{iid}{\sim} \mathrm{Unif}(3/10, 4/10)$ and $\iota^{\mathrm{TR}}_{p} \overset{iid}{\sim} \mathrm{Unif}(1/10, 3/10)$. We simulated β^{TR}_{jp} conditional on $\gamma^{\mathrm{TR}}_{jp}$; if $\gamma^{\mathrm{TR}}_{jp} = 0$, then $\beta^{\mathrm{TR}}_{jp} = 0$. For the cases of $\gamma^{\mathrm{TR}}_{jp} \neq 0$, we generated β^{TR}_{jp} from the normal distributions with mean 0 and variance $\sigma^{2,\mathrm{TR}}_{p}$ truncated from below at $\iota^{\mathrm{TR}}_{p}\sigma^{\mathrm{TR}}_{p}$ if $\gamma^{\mathrm{TR}}_{jp} = 1$ and from above at $-\iota^{\mathrm{TR}}_{p}\sigma^{\mathrm{TR}}_{p}$ if $\gamma^{\mathrm{TR}}_{jp} = 2$. We kept the same simulation setup for $\alpha^{\mathrm{TR}}_{tj}$ and s_j. We then drew the OTU counts

Model	MCC	ACC	AUC	Brier-Score	F_1
ANLP-SB	**0.741** (0.089)	**0.872** (0.043)	**0.957** (0.034)	**0.204** (0.075)	**0.860** (0.042)
ALP-SB	0.227 (0.061)	0.553 (0.106)	0.811 (0.020)	0.668 (0.154)	0.672 (0.102)
SLP-SB	0.202 (0.067)	0.538 (0.111)	0.832 (0.022)	–	0.667 (0.103)
BayesReg	0.567 (0.052)	0.768 (0.051)	0.815 (0.018)	–	0.704 (0.024)
edgeR-L	-0.028 (0.060)	0.505 (0.069)	0.483 (0.037)	–	0.272 (0.078)
edgeR-Q	-0.003 (0.082)	0.514 (0.068)	0.496 (0.052)	–	0.307 (0.085)
BhGLM	0.285 (0.042)	0.602 (0.086)	0.683 (0.030)	–	0.331 (0.056)

(a) Variable Selection

Model	RMSE			DIC	LPML
	β_{jp}	$\pi^{\star}_{p0}$	g_{tkj}		
ANLP-SB	**0.140** (0.030)	**0.194** (0.066)	**0.491** (0.103)	**366,347** (13,317)	**-4.611** (0.168)
ALP-SB	0.167 (0.017)	0.492 (0.077)	0.618 (0.110)	366,577 (13,222)	-4.625 (0.167)
SLP-SB	0.185 (0.025)	0.483 (0.077)	0.975 (0.260)	366,873 (13,318)	-4.629 (0.167)
BayesReg	0.222 (0.024)	–	1.198 (0.269)	368,725 (13,276)	-4.669 (0.167)
edgeR-L	0.407 (0.031)	–	–	–	–
edgeR-Q	0.377 (0.033)	–	–	–	–
BhGLM	0.439 (0.066)	–	–	–	–

(b) Parameter Estimation and Model Fit

Table 2.6: [Simulation 5: Comparison] Results from the simulation study with $N = 100$ samples taken at $n = 50$ time points with $J = 400$ OTUs and $P = 20$ covariates.

from $Y_{tkj} \mid \mu^{\text{TR}}_{tkj}, s^{\text{TR}}_j \overset{indep}{\sim} \text{NB}(\mu^{\text{TR}}_{tkj}(\boldsymbol{x}_t),\, s^{\text{TR}}_j)$, where $\mu^{\text{TR}}_{tkj}(\boldsymbol{x}_t) = \exp(r^{\text{TR}}_{tk} + \alpha^{\text{TR}}_{0j} + \alpha^{\text{TR}}_{tj} + \boldsymbol{x}'_t\boldsymbol{\beta}^{\text{TR}}_j)$. A total of 100 datasets were simulated under each scenario. The models, including ANLP-SB and the competing models, are compared under eight criteria. Tables 2.5-2.9 summarize results for Scenarios 4-8, respectively. The averages for the metrics are listed in the table along with standard deviations in parenthesis. Under all scenarios, ANLP-SB outperforms the other models in terms of variable selection. For Scenarios 4-7, ANLP-SB yields better parameter estimates as well. In Simulation 8, ALP-SB and SLP-SB obtain better RMSE for β_{jp}, ALP-SB for RMSE for g_{tkj}, and BayeReg for DIC, while ANLP-SB is still very close to the best performers under those criteria. The results demonstrate that ANLP-SB is well-suited for scaling up to higher dimensional settings.

The last column of Table 2.4 lists the run-times in minutes for 1,000 MCMC

Model	MCC	ACC	AUC	Brier-Score	F_1
ANLP-SB	**0.754** (0.148)	**0.879** (0.076)	**0.930** (0.063)	**0.218** (0.150)	**0.876** (0.095)
ALP-SB	0.378 (0.091)	0.643 (0.121)	0.837 (0.018)	0.581 (0.193)	0.717 (0.132)
SLP-SB	0.388 (0.096)	0.640 (0.129)	0.852 (0.020)	–	0.718 (0.134)
BayesReg	0.671 (0.065)	0.838 (0.040)	0.878 (0.021)	–	0.818 (0.031)
edgeR-L	-0.043 (0.137)	0.505 (0.085)	0.474 (0.084)	–	0.392 (0.145)
edgeR-Q	-0.043 (0.146)	0.503 (0.078)	0.475 (0.090)	–	0.422 (0.150)
BhGLM	0.467 (0.064)	0.697 (0.081)	0.788 (0.033)	–	0.578 (0.041)

(a) Variable Selection

Model	RMSE			DIC	LPML
	β_{jp}	$\pi^{\star}_{p0}$	g_{tkj}		
ANLP-SB	**0.097** (0.025)	**0.231** (0.144)	**0.302** (0.072)	**616,719** (13,581)	**-3.859** (0.085)
ALP-SB	0.125 (0.019)	0.439 (0.115)	0.398 (0.075)	617,165 (13,543)	-3.864 (0.085)
SLP-SB	0.144 (0.031)	0.421 (0.113)	0.520 (0.143)	617,280 (13,526)	-3.865 (0.085)
BayesReg	0.187 (0.032)	–	0.695 (0.169)	617,529 (13,513)	-3.868 (0.085)
edgeR-L	0.222 (0.021)	–	–	–	–
edgeR-Q	0.205 (0.020)	–	–	–	–
BhGLM	0.266 (0.039)	–	–	–	–

(b) Parameter Estimation and Model Fit

Table 2.7: [Simulation 6: Comparison] Results from the simulation study with $N = 200$ samples taken at $n = 100$ time points with $J = 400$ OTUs and $P = 10$ covariates.

iterations using the setups from Simulations 1-8. Run times of ANLP-SB depend on the size of the data (J, P, n, N) as well as some fixed hyperparameters (L^r, L^α, M). The number of time points (n) and replicates at a time point (K_i) determine the total number of samples (N). Comparing the run times of Simulation 4 and 5 to those of Simulations 6 and 7, respectively, shows the impact of an increase in n on the computational cost. The number of sample-specific size factors r_{tk} increases in N, and updating them involves updating two mixture indicators, c^r_{tk} and λ^r_{tk} for each r_{tk}, potentially resulting in a substantial increase in the computational cost. The computational cost for α_{0j} scales in a similar way to that of r_{tk}, but with respect to the number of OTUs J. Another factor that may significantly increase the computational cost of the model is the number of candidate covariates, P, especially with large J, as indicated from comparing run times of Simulations 4

Model	MCC	ACC	AUC	Brier-Score	F_1
ANLP-SB	**0.752** (0.112)	**0.878** (0.056)	**0.937** (0.051)	**0.213** (0.108)	**0.878** (0.056)
ALP-SB	0.334 (0.075)	0.609 (0.099)	0.873 (0.017)	0.618 (0.156)	0.705 (0.095)
SLP-SB	0.314 (0.079)	0.591 (0.107)	0.894 (0.018)	–	0.698 (0.098)
BayesReg	0.681 (0.058)	0.837 (0.037)	0.877 (0.018)	–	0.815 (0.020)
edgeR-L	-0.026 (0.104)	0.502 (0.060)	0.484 (0.061)	–	0.378 (0.105)
edgeR-Q	-0.032 (0.095)	0.498 (0.058)	0.480 (0.058)	–	0.399 (0.092)
BhGLM	0.437 (0.051)	0.678 (0.069)	0.766 (0.028)	–	0.533 (0.031)

(a) Variable Selection

Model	RMSE			DIC	LPML
	β_{jp}	$\pi^{\star}_{p0}$	g_{tkj}		
ANLP-SB	**0.111** (0.024)	**0.224** (0.099)	**0.426** (0.102)	**633,665** (18,953)	**-3.971** (0.119)
ALP-SB	0.126 (0.011)	0.470 (0.083)	0.495 (0.084)	634,361 (18,864)	-3.979 (0.118)
SLP-SB	0.145 (0.025)	0.452 (0.081)	0.800 (0.268)	634,656 (18,839)	-3.981 (0.118)
BayesReg	0.183 (0.026)	–	1.088 (0.302)	635,145 (18,811)	-3.989 (0.118)
edgeR-L	0.245 (0.017)	–	–	–	–
edgeR-Q	0.224 (0.017)	–	–	–	–
BhGLM	0.291 (0.032)	–	–	–	–

(b) Parameter Estimation and Model Fit

Table 2.8: [Simulation 7: Comparison] Results from simulation study with $N = 200$ samples taken at $n = 100$ time points with $J = 400$ OTUs and $P = 20$ covariates.

and 6 to those of Simulations 5 and 7, respectively. The computation time for β_{jp} and γ_{jp} scales with both J and P. A large P also increases the number of other parameters such as σ_p^2, ι_p and $\boldsymbol{\pi}_p$. The amount of computation required may rapidly escalate with increasing P when J is large. We note that compared to N and J, the values of L^r, L^α, and M do not significantly impact run times because they are related to hyperparameters at a high level of the model. Also, parameters $\alpha_{tj} = \sum_{m=1}^{M} K(t - u_m)\theta_{mj}$ are deterministically calculated given θ_{mj} and prespecified kernel K, and the computation time for α_{tj} scales with M and J.

Model	MCC	ACC	AUC	Brier-Score	F_1
ANLP-SB	**0.744** (0.033)	**0.866** (0.023)	**0.959** (0.007)	**0.236** (0.041)	**0.835** (0.015)
ALP-SB	0.081 (0.035)	0.480 (0.051)	0.817 (0.014)	0.699 (0.072)	0.630 (0.050)
SLP-SB	0.038 (0.029)	0.466 (0.054)	0.840 (0.019)	–	0.627 (0.051)
BayesReg	0.645 (0.032)	0.813 (0.026)	0.850 (0.013)	–	0.756 (0.017)
edgeR-L	-0.001 (0.048)	0.522 (0.035)	0.498 (0.029)	–	0.334 (0.051)
edgeR-Q	-0.015 ()0.050	0.515 (0.038)	0.491 (0.030)	–	0.336 (0.046)
BhGLM	0.321 (0.037)	0.646 (0.042)	0.687 (0.022)	–	0.426 (0.042)

(a) Variable Selection

Model	RMSE			DIC	LPML
	β_{jp}	$\pi^{\star}_{p0}$	g_{tkj}		
ANLP-SB	0.171 (0.022)	**0.249** (0.023)	1.087 (0.261)	334,422 (15,563)	**-4.211** (0.196)
ALP-SB	**0.152** (0.011)	0.456 (0.023)	**0.655** (0.078)	334,196 (15,512)	-4.232 (0.195)
SLP-SB	0.153 (0.017)	0.449 (0.023)	1.120 (0.323)	334,599 (15,495)	-4.238 (0.195)
BayesReg	0.185 (0.016)	–	1.742 (0.353)	**334,161** (15,526)	-4.252 (0.195)
edgeR-L	0.344 (0.022)	–	–	–	–
edgeR-Q	0.320 (0.021)	–	–	–	–
BhGLM	0.423 (0.035)	–	–	–	–

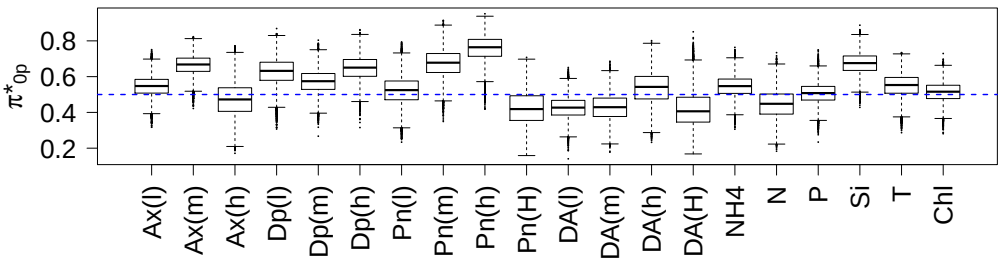
(b) Parameter Estimation and Model Fit

Table 2.9: [Simulation 8: Comparison] Results from simulation study with $N = 200$ samples taken at $n = 100$ time points with $J = 200$ OTUs and $P = 50$ covariates.

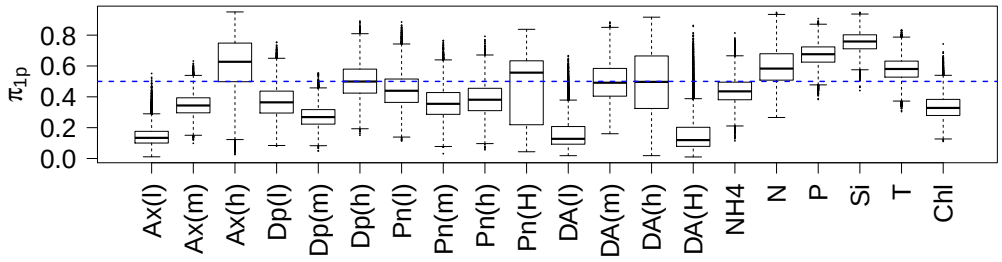
2.4 Ocean Microbiome Data Analysis

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 2.1(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



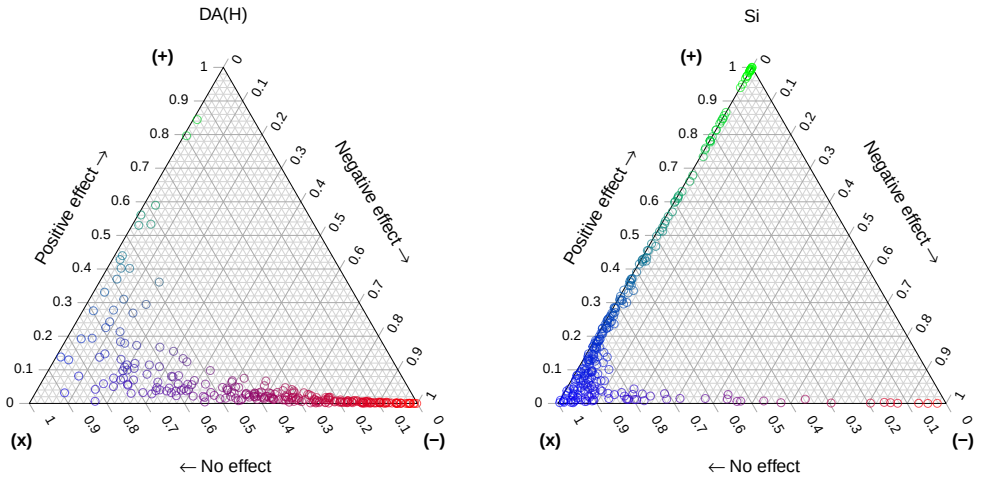
(a) Posterior distributions of $\pi^{\star}_{p0}$



(b) Posterior distributions of π_{p1}

Figure 2.4: [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.

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 A.2 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.



(a) Very high level of domoic acid (b) Silicate

Figure 2.5: [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.

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 2.4 summarizes posterior inferences on overall sparsity parameter $\pi^{\star}_{p0}$, 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 $\pi^{\star}_{p0}$ 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 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 $\hat{\pi}_0^\star = 0.67$, 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 2.5 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 2.4(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 A.7 has simplex plots for all covariates. Supplementary Figure A.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* 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 A.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 A.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 A.3 shows DIC and LPML for the Bayesian models. Both criteria indicate that ANLP-SB gives a better fit to the data.

2.5 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.

Chapter 3

A Bayesian Nonparametric Analysis for Zero Inflated Multivariate Count Data with Application to Microbiome Study

3.1 Introduction

The statistical community has increasingly focused on developing techniques to model high-throughput sequencing (HTS) data produced by microbiome studies. Although HTS data has been successfully used to profile complex microbial communities, analysis of such data remains challenging. In this work, we focus on the analysis of multivariate count data with excess zeros, in particular, read count data of taxa produced by 16S ribosomal RNA (rRNA) sequencing. As a motivating example, we consider the chronic wound microbiome data in Verbanic *et al.* (2019), which consists of microbiome samples taken from human subjects'

chronic wounds, both pre- and post-debridement, as well as from their healthy skin. Verbanic *et al.* (2019) studied changes to the chronic wound microbiome by debridement, which is known to be an effective treatment for chronic wounds. We present a Bayesian nonparametric regression model that includes a submodel for zero inflation and flexibly accommodates covariates such as environmental factors and clinical characteristics for differential abundance analysis. The model provides an inferential framework to gain further insights into complex microbial communities.

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

Various statistical models haven been proposed for microbiome data analysis that take these features into account. Zero-inflated count models including zero

inflated Poisson (ZIP) and zero inflated negative binomial (ZINB) are common choices to address the problem of excessive zeros. To detect associations or differential abundance, these models generally relate OTU abundance to a set of covariates by modeling the mean counts or some transformation of the counts via a link function. Some of these models, such as Chen and Li (2016) analyze each OTU individually, while many more recent models analyze OTUs jointly through some hierarchical structure. Hierarchical models allow for borrowing strength across taxa for enhanced estimation of covariate effects or increased power to detect differential abundance. In this vein, Jonsson *et al.* (2018) model the counts directly using a ZIP model with OTU and sample specific random effects to account for overdispersion. Lee *et al.* (2018) use a ZIP model with spike-and-slab priors for variable selection on regression parameters related to taxa abundances and zero inflation. This model also includes a multivariate random effect to account for interdependence among OTU counts in a sample. See Sankaran and Holmes (2018), Tang and Chen (2018) and Kaul *et al.* (2017) among many others for more examples of using zero inflated models.

We develop a Bayesian nonparametric multivariate regression model with zero inflation that enables assessment of taxa richness and diversity that potentially varies with covariates. We use a ZINB distribution for OTU counts and assume an OTU count is either equal to zero or follows a NB distribution. The ZINB model properly accounts for the overdispersion and excess zeros that are common in HTS data. We build nonparametric regression prior models on the probability of an OTU count being zero and the mean count of an OTU to study the effects of covariates $\boldsymbol{x}$ on microbial communities. The probit of the probability of an OTU count being zero, ξ, and the logarithm of the OTU's differential abundance compared to the baseline counts, θ, are assumed to follow unknown distribution

functions indexed by $\boldsymbol{x}$, $F^{\xi}_{\boldsymbol{x}}$ and $F^{\theta}_{\boldsymbol{x}}$, respectively. We use a dependent Dirichlet process (DDP) (MacEachern, 1999, 2000), a flexible nonparametric Bayesian model to model $F^{\xi}_{\boldsymbol{x}}$ and $F^{\theta}_{\boldsymbol{x}}$. The DDP is a popular choice to model a set of random functions related through $\boldsymbol{x}$. Our model is highly flexible with regard to the nature of the relationship of the covariates and an OTU's abundance and presence. In addition to inference on the association of individual taxa with covariates through ξ and θ, $F^{\xi}_{\boldsymbol{x}}$ and $F^{\theta}_{\boldsymbol{x}}$ provide community-level insights related to alpha-diversity and species evenness, which distinguishes our method from other commonly used models for differential abundance analysis. To improve the inference on $F^{\xi}_{\boldsymbol{x}}$ and $F^{\theta}_{\boldsymbol{x}}$, we construct an elaborate model for the baseline abundance of OTUs in samples. The baseline count of an OTU in a sample is modeled as a function of a sample size factor and an OTU baseline abundance factor to account for count variation related to sequencing depth and different baseline abundances of OTUs. The baseline abundance factors are shared by samples from a group, such as the subject or location where each sample was collected, to reflect the dependent taxa abundance levels shared across these samples. These two factors constitute a basis for the estimation and meaningful interpretation of ξ and θ.

In the remainder of the chapter we describe the model and its applications. §3.2 describes the proposed Bayesian nonparametric multivariate NB regression model with zero inflation (called "BNP-ZIMNR"), §3.3 has results from the model applied to some simulation studies, §3.4 has results from the model applied to a chronic wound microbiome dataset, and §3.5 concludes with some discussion of the results and areas of future research.

3.2 Probability Model

3.2.1 Sampling Model

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

$$Y_{ij} \mid \epsilon_{j,x_i}, \mu_{ij}, s_j \overset{indep}{\sim} \epsilon_{j,x_i}\delta_{\{0\}}(Y_{ij}) + (1 - \epsilon_{j,x_i})\,\mathrm{NB}\left(\mu_{ij}(x_i, u_i),\ s_j\right), \tag{3.1}$$

where $\delta_A(\cdot)$ is the Dirac measure at A and $\mathrm{NB}(\mu, s)$ the negative binomial (NB) distribution with mean μ and dispersion parameter s (so the variance is $\mu + s\mu^2$). The zero inflated model in (3.1) assumes that abundance is conditional on the presence of an OTU. $(1 - \epsilon_{j,x_i})$ is the probability of presence for OTU j in sample i, and is a function of covariate x_i. With probability $(1 - \epsilon_{j,x_i})$ the NB generates counts, some of which can be zero. The model specification implies that a zero count can be produced in two ways. An OTU may truly be absent in a sample with x_i. Conversely, zero counts may be produced for rare OTUs even when those OTUs are truly present if the sequencing effort is not sufficient to surface their presence. HTS data is commonly modeled using NB models, as in (3.1), which are more flexible than their Poisson counterparts in accommodating overdispersion.

Overdispersion parameter s_j controls the amount of overdispersion, with larger s_j indicating a greater amount of overdispersion, and the equivalent Poisson model with mean μ_{ij} is recovered as $s_j \to 0$. We let the overdispersion parameters $s_j \overset{iid}{\sim}$ Log-Normal(a_s, b_s^2) with a_s and b_s^2 fixed. The mixture model in (3.1) can be represented with latent indicator variables $\delta_{ij} \in \{0, 1\}$ for presence and absence of OTU j in sample i. We assume $\delta_{ij} \overset{indep}{\sim} \text{Ber}(1-\epsilon_{j,x_i})$, and let $Y_{ij} = 0$ for $\delta_{ij} = 0$ and $Y_{ij} \overset{indep}{\sim} \text{NB}\left(\mu_{ij}(x_i, u_i),\, s_j\right)$ for $\delta_{ij} = 1$.

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

$$\log(\mu_{ij}(k, m)) = \alpha_{jm} + r_i + \theta_{jk}. \tag{3.2}$$

A baseline abundance factor of OTU j for samples from group m, α_{jm} accounts for different baseline abundances of OTUs. It is shared by the samples from group $u_i = m$ and induces dependence among Y_{ij} with $u_i = m$. r_i is a sample specific normalization factor to account for different library sizes across samples. Parameters α_{jm} and r_i together form the baseline count of OTU j in sample i. It is common that r_i is set to the logarithm of the total counts $Y_{i\bullet} = \sum_{j=1}^{J} Y_{ij}$ as an offset variable (e.g., see Lee *et al.* (2018) and Zhang *et al.* (2017a)). We instead let r_i be random, which enables full model-based inference with appropriate uncertainty quantification. θ_{jk} in (3.2) represents a multiplicative change in abundance of OTU j for covariate level k compared to its baseline abundance. A value of θ_{jk} close to zero implies that the abundance of an OTU is close to the baseline abundance, i.e., non-differentially abundant, and positive or negative values of θ_{jk} imply low or high abundance of OTU j in a sample with $x_i = k$, respectively. Comparison of θ_{jk} across k can be used to infer differential abundance of OTU j. Similarly, comparison of θ_{jk} across j provides insights on relative abundances of

OTUs in a sample with level k, such as species diversity compared to the baseline.

Using regression models for ϵ_{jk} and θ_{jk} is common to quantify covariate effects on the occurrence of excess zeros and differential abundances. Using our motivating dataset as a specific example, one may choose one k' of the levels as a reference and let $\theta_{jk'} = 0$. θ_{jk}, $k \neq k'$ is then interpreted as an effect size relative to the abundance of OTU j under the reference. A potential drawback of this approach is that θ_{jk}, $k \neq k'$ cannot be meaningfully estimated if an OTU is absent under the reference level. A common workaround to address this issue is to replace zeros with a small value, known as pseudo count, if an OTU has zeros in all samples of the reference level. However, this arbitrary modification of the data may result in biased inference. On the other hand, the decomposition of μ in (3.2) can avoid potential biases because θ_{jk} represents differential abundance compared to the baseline abundance $r_i + \alpha_{jm}$. The baseline count of an OTU can be estimated if an OTU exists for at least one k. We let $\theta_{jk} = 0$ if an OTU is present only for one level of k so that θ_{jk} can be fully interpreted. For ϵ_{jk}, we use a probit link function, $\Phi^{-1}(\epsilon_{jk}) = \xi_{jk}$, where $\Phi^{-1}(\cdot)$ is a inverse cumulative distribution function of the standard normal distribution. In the presence of a high proportion of zeros, including random group effects for ϵ may produce highly unstable model fitting and computational intractability (Agarwal *et al.*, 2002), and for this reason we let ϵ_{jk} be a function of x_i only. The dependence of ϵ_{jk} on x_i only is in contrast with μ_{ij}, which depends on both u_i and x_i. In the following, we consider a flexible BNP approach to model ξ_{jk} and θ_{jk} to improve inference on presence/absence and differential abundance.

3.2.2 Prior

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

$$\xi_{jk} \overset{iid}{\sim} F_k^{\xi} = \sum_{\ell=1}^{\infty} \psi_{\ell}^{\xi} \, \mathrm{N}\left(\xi_{k\ell}^{\star}, \, \sigma_{\xi k}^2\right) \quad \text{and} \quad \theta_{jk} \overset{iid}{\sim} F_k^{\theta} = \sum_{\ell=1}^{\infty} \psi_{\ell}^{\theta} \, \mathrm{N}\left(\theta_{k\ell}^{\star}, \, \sigma_{\theta k}^2\right). \tag{3.3}$$

The mixture locations $\xi_{k\ell}^{\star}$ and $\theta_{k\ell}^{\star}$ depend on k and we let $\xi_{k\ell}^{\star} \overset{iid}{\sim} \mathrm{N}(\bar{\xi}^{\star}, \tau_{\xi}^2)$ and $\theta_{k\ell}^{\star} \overset{iid}{\sim} \mathrm{N}(\bar{\theta}^{\star}, \tau_{\theta}^2)$. The covariate independent weights ψ_{ℓ}^{χ}, $\chi \in \{\theta, \xi\}$ take the form $\psi_{\ell}^{\chi} = v_{\ell}^{\chi} \prod_{\ell'=1}^{\ell-1} (1 - v_{\ell'}^{\chi})$ with $v_{\ell}^{\chi} \overset{iid}{\sim} \mathrm{Be}(1, \rho^{\chi})$. That is, the "single-p" DDPs that assume predictor independent weights are used in (3.3) as priors over the distributions of the mixture locations. MacEachern (1999, 2000) proposed the DDP to

model related random probability distributions. When flexible point mass processes are considered for $\boldsymbol{\theta}^{\star}_{\ell} = \{\theta^{\star}_{x\ell}, x \in \mathcal{X}\}$ and $\boldsymbol{\xi}^{\star}_{\ell} = \{\xi_{x\ell}, x \in \mathcal{X}\}$, the "single-p" DDP has full weak support, implying that the prior model is flexible enough to generate sample paths sufficiently close to any probability distribution. DDP and its variations have been successfully used to model related probability distributions in many applications including ANOVA (De Iorio *et al.*, 2004), survival (De Iorio *et al.*, 2009; Jara *et al.*, 2010), time series analysis (Griffin and Steel, 2011; Nieto-Barajas *et al.*, 2012) and spatial modeling (Gelfand *et al.*, 2005) among many others. The DDP mixture formulation in (3.3) allows us to flexibly specify and, after fitting the model, analyze and compare, F^{θ}_{x} and F^{ξ}_{x} without restrictive parametric assumptions about their functional forms. We assume $\sigma^2_{\chi k} \overset{iid}{\sim} \text{IG}(a^{\chi}_{\sigma}, b^{\chi}_{\sigma})$, $\chi \in \{\xi, \theta\}$. With minimal changes, (3.3) can accommodate multiple covariates and continuous covariates; as a simple example, we may consider $\theta^{\star}_{\ell}(\boldsymbol{x}_i) = \boldsymbol{a}'_{\theta,\ell}\boldsymbol{x}_i$ and $\xi^{\star}_{\ell}(\boldsymbol{x}_i) = \boldsymbol{a}'_{\xi,\ell}\boldsymbol{x}_i$ with $\boldsymbol{a}_{\chi,\ell} \overset{iid}{\sim} \text{N}(\bar{\boldsymbol{a}}_{\chi}, \boldsymbol{B}_{\chi})$ with $\boldsymbol{B}_{\chi} > 0$, $\chi \in \{\theta, \xi\}$.

Parameters r_i and α_{jm} construct the baseline count of OTU j in a sample with $u_i = m$, and serve as an "overall mean." Observe that the parameters in (3.2) are not identifiable due to the multiplicative structure, $\text{E}(Y_{ij} \mid \delta_{ij} = 0) = e^{r_i+\alpha_{jm}+\theta_{jk}}$. We place constraints on the distributions of both r_i and α_{jm} to circumvent the identifiability issue in estimating the baseline counts, $\exp(r_i+\alpha_{jm})$. More importantly, the constraints allow parameters of primary interest θ_{jk} and F^{θ}_{k} to be identified. Specifically, we use mean-constrained priors with a mixture-of-mixtures structure (Li *et al.*, 2017) for r_i and α_{jm},

$$\begin{aligned} r_i &\overset{iid}{\sim} \sum_{\ell=1}^{L^r} \psi^r_{\ell} \left\{ w^r_{\ell}\, \text{N}(\eta^r_{\ell}, u^2_r) + (1 - w^r_{\ell})\, \text{N}\left(\frac{\upsilon_r - w^r_{\ell}\eta^r_{\ell}}{1 - w^r_{\ell}}, u^2_r\right)\right\}, \\ \alpha_{jm} &\overset{iid}{\sim} \sum_{\ell=1}^{L^{\alpha}} \psi^{\alpha}_{\ell} \left\{ w^{\alpha}_{\ell}\, \text{N}(\eta^{\alpha}_{\ell}, u^2_{\alpha}) + (1 - w^{\alpha}_{\ell})\, \text{N}\left(\frac{\upsilon_{\alpha} - w^{\alpha}_{\ell}\eta^{\alpha}_{\ell}}{1 - w^{\alpha}_{\ell}}, u^2_{\alpha}\right)\right\}, \end{aligned} \tag{3.4}$$

where υ_χ, $\chi \in \{r, \alpha\}$ are the distribution's fixed, prespecified mean constraints, and ψ_ℓ^χ and w_ℓ^χ are mixture weights with $\sum_{\ell=1}^{L^\chi} \psi_\ell^\chi = 1$ and $0 < \psi_\ell^\chi, w_\ell^\chi < 1$. Although mean-constrained, the mixture-of-mixture formulation provides significant flexibility, as it can accurately characterize a wide range of distributions, including multi-modal and skewed distributions. Lee and Sison-Mangus (2018) and Shuler *et al.* (2019a) used the distributions in (3.4) for model based normalization in similar settings, and their results indicate the baseline abundance and covariate effects can be estimated without issues related to identifiability. In contrast to using plug-in empirical estimates for normalizing factors, the flexible model based approach can further improve estimation of ξ_{jk} and θ_{jk}, and thus enhance estimation of F_k^ξ and F_k^θ. We follow Li *et al.* (2017) and set $\upsilon_r = 0$, which can be interpreted as on average no scaling adjustment; although other approaches are available, such as using an empirical estimate like in Shuler *et al.* (2019a) or setting the constraint using prior information if it is available. We use an empirical approach to set υ_α. We compute $\tilde{r}_i = \log(Y_{i\bullet}/Y_{\bullet\bullet}) - \frac{1}{N}\sum_{i'} \log(Y_{i'\bullet}/Y_{\bullet\bullet})$ with $Y_{\bullet\bullet} = \sum_{i,j} Y_{ij}$ as mean zero empirical estimates of r_i and set $\upsilon_\alpha = \left[\sum_{i,j|Y_{ij}>0} \{\log(Y_{ij}) - \tilde{r}_i\}\right] / \left\{\sum_{i,j} 1(Y_{ij} > 0)\right\}$. Inference on θ and ϵ is not sensitive to specification of υ_r and υ_α (Lee and Sison-Mangus, 2018; Shuler *et al.*, 2019a). Our simulation studies and real data analyses also show robustness of inference to different specifications of υ_r and υ_α. We place a Dirichlet prior on the outer mixture weights and a beta prior on the inner mixture weights, letting $\boldsymbol{\psi}_\ell^\chi = (\psi_1^\chi, \ldots, \psi_{L^\chi}^\chi) \sim \text{Dir}(\boldsymbol{a}_\psi^\chi)$ and $w_\ell^\chi \overset{iid}{\sim} \text{Be}(a_w^\chi, b_w^\chi)$, $\chi \in \{r, \alpha\}$, where $\boldsymbol{a}_\psi^\chi = (a_{\psi 1}^\chi, \ldots, a_{\psi, L^\chi}^\chi)$, a_w^χ and b_w^χ are fixed hyperparameters. We let $\eta_\ell^\chi \overset{iid}{\sim} \text{N}(\upsilon_\chi, b_{\eta^\chi}^2)$ with $b_{\eta^\chi}^2$ fixed.

3.2.3 Posterior Computation

Let $\underline{\boldsymbol{\theta}} = [s_j, \delta_{ij}, r_i, \alpha_{jm}, \xi_{jk}, \theta_{jk}, (\chi^{\star}_{k\ell}, v^{\chi}_{\ell}, \sigma^2_{\chi k}, \chi \in \{\theta, \xi\}), (\psi^{\chi}_{\ell}, w^{\chi}_{\ell}, \eta^{\chi}_{\ell}, \chi \in \{r, \alpha\})]$ denote the vector of all unknown parameters. The joint posterior distribution is $\mathrm{P}\left(\underline{\boldsymbol{\theta}} \mid \boldsymbol{Y}, \boldsymbol{x}, \boldsymbol{u}\right) \propto \mathrm{P}\left(\boldsymbol{Y} \mid \underline{\boldsymbol{\theta}}, \boldsymbol{x}, \boldsymbol{u}\right) \mathrm{P}\left(\underline{\boldsymbol{\theta}}\right)$. We use standard Markov chain Monte Carlo (MCMC) methods consisting of Gibbs and Metropolis steps to draw samples from the posterior distribution. As is standard in mixture modeling we introduce auxiliary variables to indicate the mixture components from which the parameters of interest belong. We add auxiliary variables of this type to aid in the posterior computation for r_i, α_{jm}, θ_{jk}, and ξ_{jk}. For computational convenience, when fitting the model we approximate the DDP in (3.3) by truncating the number of mixture components of F^{χ}_k to L^{χ}, $\chi \in \{\xi, \theta\}$. The final weight $\psi^{\chi}_{L^{\chi}} = 1 - \sum_{\ell=1}^{L^{\chi}-1} \psi^{\chi}_{\ell}$ is set to ensure F^{ξ}_k is proper. With large enough L^{χ} the truncated process produces inference almost identical to that with the infinite process (Ishwaran and James, 2001; Rodriguez and Dunson, 2011). As discussed in Rodriguez and Dunson (2011) if there is discrepancy between the posterior distributions under the truncated and infinite processes, the model is typically sensitive to the choice of L^{χ}. We examined the posterior distribution of $\psi^{\chi}_{L^{\chi}}$ and the sensitivity of the model to a choice of L^{χ}. We found that the truncated process is robust to a choice of L^{χ} if L^{χ} is sufficiently large. We diagnose convergence and mixing of the described posterior MCMC simulation using trace plots and autocorrelation plots of imputed parameters. For both the upcoming simulation examples and the data analysis, we found no evidence of practical convergence problems. An R package for the model, `bnpzimnr`, is available at https://github.com/kurtis-s/bnpzimnr. Details of posterior computation are given in Supplementary §B.1.

3.3 Simulation Studies

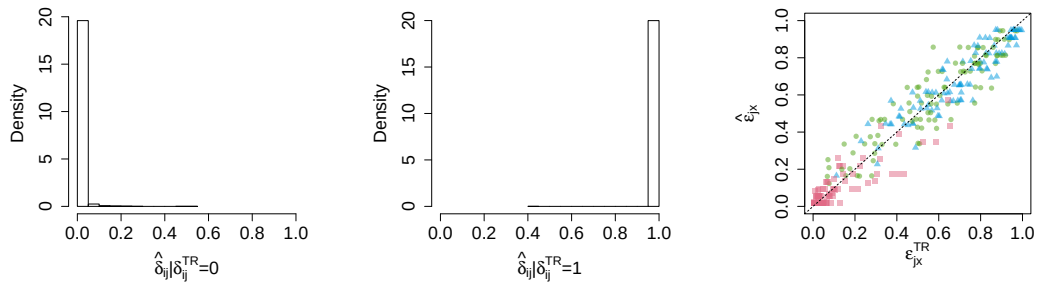
3.3.1 Simulation 1

To assess the performance of the proposed model, BNP-ZIMNR, we performed simulation studies and compared its performance to that of alternative models. We included a factor with three levels and simulated data for 100 OTUs from 20 subjects, i.e., $J = 100$, $M = 20$, and $K = 3$, resulting in $n = 60$ samples, a covariate $x_i \in \{1, 2, 3\}$, $i = 1, \ldots, N$ and a grouping factor $u_i \in \{1, \ldots, 20\}$. We used Gaussian mixtures to set the simulation truth for $F_k^{\xi,\mathrm{TR}}$ and $F_k^{\theta,\mathrm{TR}}$, $k = 1, 2, 3$; let $F_1^{\xi,\mathrm{TR}} = 0.6\,\mathrm{N}(-2,\, 0.25) + 0.4\,\mathrm{N}(-1,\, 0.5)$, $F_2^{\xi,\mathrm{TR}} = 0.2\,\mathrm{N}(-0.5,\, 0.25) + 0.8\,\mathrm{N}(0.5,\, 0.5)$, and $F_3^{\xi,\mathrm{TR}} = 0.5\,\mathrm{N}(0,\, 0.25) + 0.5\,\mathrm{N}(1,\, 0.5)$. Similarly, we set to $F_1^{\theta,\mathrm{TR}} = 0.3\,\mathrm{N}(3,\, 0.25) + 0.6\,\mathrm{N}(2,\, 0.25) + 0.1\,\mathrm{N}(-1.5,\, 0.5)$, $F_2^{\theta,\mathrm{TR}} = 0.3\,\mathrm{N}(2,\, 0.5) + 0.6\,\mathrm{N}(-1,\, 0.25) + 0.1\,\mathrm{N}(-2,\, 0.25)$, and $F_3^{\theta,\mathrm{TR}} = 0.3\,\mathrm{N}(2,\, 0.5) + 0.35\,\mathrm{N}(-1,\, 0.25) + 0.35\,\mathrm{N}(-2,\, 0.25)$. $F_k^{\xi,\mathrm{TR}}$ and $F_k^{\theta,\mathrm{TR}}$ are illustrated with the solid black lines in Figure 3.3. F_1^{ξ} generally favors smaller values of ξ_{jk}, indicating greater species richness in level 1 than in the other levels. When an OTU is present in a sample with $k = 1$, it tends to have a value of θ_{jk} greater than zero, i.e., a higher abundance. On the other hand, for levels $k = 2, 3$, OTUs are likely to be absent, and when they are present, their abundances are low with large probability. In a simulated dataset, the three levels of x_i approximately have 9%, 59% and 69% of Y_{ij} being equal to 0, respectively. We drew ξ_{jk}^{TR} independently from $F_k^{\xi,\mathrm{TR}}$ and generated $\delta_{ij}^{\mathrm{TR}} \overset{indep}{\sim} \mathrm{Ber}(1-\epsilon_{jk}^{\mathrm{TR}})$ for a sample with $x_i = k$, where $\epsilon_{jk}^{\mathrm{TR}} = \Phi(\xi_{jk}^{\mathrm{TR}})$. If an OTU is present for two or more levels of the factor, i.e., differential abundance can be meaningfully defined, then we drew $\theta_{jk}^{\mathrm{TR}}$ from $F_k^{\theta,\mathrm{TR}}$. If an OTU is present for only one level $\theta_{jk}^{\mathrm{TR}} = 0$. Otherwise, $\theta_{jk}^{\mathrm{TR}}$ is not defined. We simulated group factors $\alpha_{j,u_i}^{\mathrm{TR}} \overset{iid}{\sim} \mathrm{N}(10,\, 1)$, normalization factors $(\exp(r_1^{\mathrm{TR}}), \ldots, \exp(r_N^{\mathrm{TR}})) \sim \mathrm{Dir}(5, \ldots, 5)$,

and dispersion parameters $s_j^{\text{TR}} \overset{iid}{\sim} \text{Log-Normal}\,(-2,\,(1/10)^2)$. For (i, j) with $\delta_{ij}^{\text{TR}} = 1$, we simulated OTU counts Y_{ij} using the NB distribution with mean $\mu_{ij}^{\text{TR}} = \exp(\alpha_{j,u_i}^{\text{TR}} + r_i^{\text{TR}} + \theta_{j,x_i}^{\text{TR}})$ and dispersion s_j^{TR}. When $\delta_{ij}^{\text{TR}} = 0$, we set $\mu_{ij}^{\text{TR}} = 0$ and $Y_{ij} = 0$.

Posterior Inference When fitting the model, we set the hyperparameters as follows: For the mean-constrained distribution of normalization factors r_i, let $\upsilon_r = 0$, $L^r = 20$, $\boldsymbol{a}_\psi^r = \mathbf{1}$, $a_w^r = 5$, $b_w^r = 5$, $u_r^2 = 0.05$, and $b_{\eta^r}^2 = 0.25$. Similarly, for the group specific baseline abundance of OTU j α_{jm}, let υ_α be specified using the empirical approach described in § 3.2.2, $L^\alpha = 150$, $\boldsymbol{a}_\psi^\alpha = \mathbf{1}$, $a_w^\alpha = 1$, $b_w^\alpha = 1$, $u_\alpha^2 = 2$ and $b_{\eta^\alpha}^2 = 1$. For the DDP priors, we let $\rho^\theta = 1$, $\bar{\theta}^\star = 0$ and $\tau_\theta^2 = 10$. For the DDP prior of ξ_{jk}, we used $\rho^\xi = 1$, $\bar{\xi}^\star = 2$ and $\tau_\xi^2 = 1$, which encourages a preference for a higher probability for zero inflation, but is still flexible enough to accommodate OTUs with little sparsity. For the mixtures' kernel dispersions let $a_\sigma^\chi = b_\sigma^\chi = 1$, $\chi \in \{\xi, \theta\}$. We set the DDP truncation levels to $L^\theta = L^\xi = 50$. Finally, we used $a_s = 0.25$, $b_s^2 = 0.25$ for the prior of OTU-specific dispersion parameters s_j. To run the MCMC simulation, we used data to initialize the parameters. For example, we initialized r_i with the empirical sample size factors $\tilde{r}_i$ used to set υ_r. Empirical proportions of zero counts, $p_{jk} = \frac{1}{M}\sum_{i=1|x_i=k}^{n} 1(y_{ij} = 0)$ were used to set initial values of ϵ_{jk} and $\xi_{k\ell}^\star$. We ran the MCMC for 70,000 iterations, discarding the first 20,000 iterations, and thinned to use every fifth sample, resulting in 10,000 samples from the posterior distribution. On a 3.2GHz Intel i5-6500 CPU running Ubuntu Linux the MCMC took approximately 12 minutes for every 5,000 iterations of the MCMC.

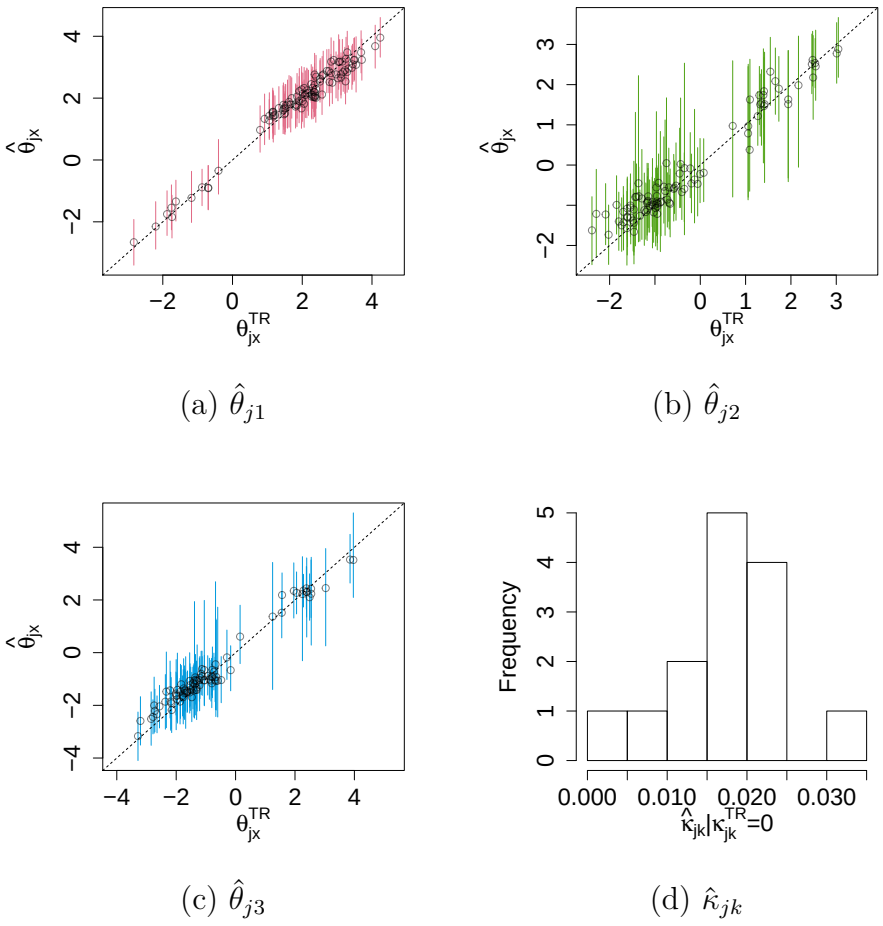
We first examine the inference on species richness in samples with k. Recall that $\delta_{ij} = 1$ implies the presence of OTU j in sample i. We used posterior means of δ_{ij} as their point estimates $\hat{\delta}_{ij} = \hat{\text{P}}(\delta_{ij} = 1 \mid \boldsymbol{y})$. The model recovers



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

Figure 3.1: [Simulation 1] Panels (a) and (b): Histograms of $\hat{\delta}_{ij} = \hat{\mathrm{P}}(\delta_{ij} = 1)$ when $\delta_{ij}^{\mathrm{TR}} = 0$ and $\delta_{ij}^{\mathrm{TR}} = 1$. Panel (c): Posterior means of ϵ_{jk} plotted against the simulation truth. Colors/shapes indicate the factor levels: $k = 1$, red squares; $k = 2$, green circles; $k = 3$, blue triangles.

the indicators for zero inflation well, as shown by the histograms of $\hat{\delta}_{ij}$ when $\delta_{ij}^{\mathrm{TR}} = 0$ and 1 in Figure 3.1(a) and (b), respectively. The model yields good estimates of $\epsilon_{jk}^{\mathrm{TR}}$, as seen in Figure 3.1(c), which shows posterior estimates of ϵ_{jk} plotted against the simulation truth. Figure 3.2 shows the resulting posterior inference on θ_{jk} for individual OTUs. To account for zero inflation, we define $\kappa_{jk} = 1\{\sum_{i=1;x_i=k}^{N} 1(\delta_{ij} = 1) > 0\}$, a binary indicator taking 0 if OTU j is absent in all samples from level k, or 1 otherwise. Note that θ_{jk} is defined only when $\kappa_{jk} = 1$. We incorporate κ_{jk} and compute point posterior estimates of θ_{jk}; $\hat{\theta}_{jk} = \sum_{b=1}^{B} \kappa_{jk}^{(b)} \times \theta_{jk}^{(b)} / \sum_{b=1}^{B} \kappa_{jk}^{(b)}$, where $b = 1, \ldots, B$ indexes the posterior samples and $\kappa_{jk}^{(b)} = 1\{\sum_{i=1|x_i=k}^{N} 1(\delta_{jk}^{(b)} = 0) > 0\}$. $\hat{\theta}_{jk}$ along with 95% credible intervals (CIs) are shown. The plots show that the model provides good estimates for differential abundance in different levels of the factor. The differences between the estimates and truth and CI lengths are greater for levels $k = 2$ and 3 because fewer non-zero counts are observed due to high prevalence of absence. Panel (d) shows posterior estimates of $\hat{\kappa}_{jk} = \frac{1}{B} \sum_{b=1}^{B} \kappa_{jk}^{(b)}$ when $\kappa_{jk}^{\mathrm{TR}} = 0$ in the simulation truth. The plot illustrates the model does a good job of identifying absence in factor levels and

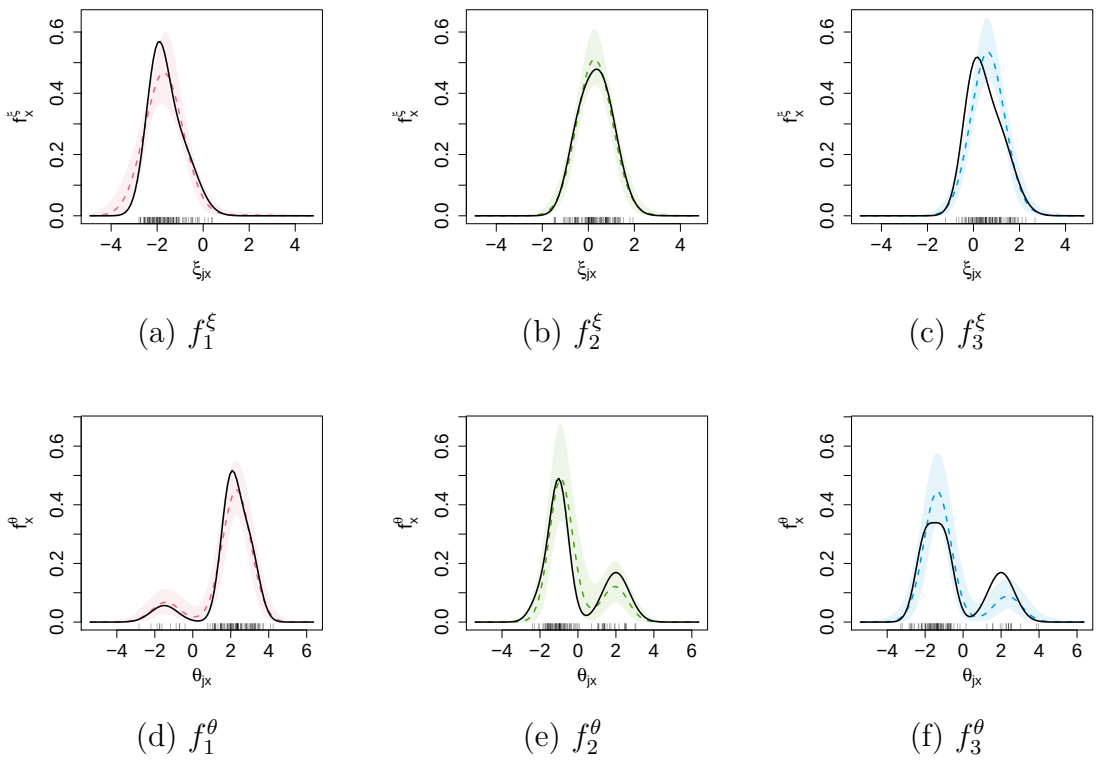


(a) $\hat{\theta}_{j1}$ (b) $\hat{\theta}_{j2}$

(c) $\hat{\theta}_{j3}$ (d) $\hat{\kappa}_{jk}$

Figure 3.2: [Simulation 1] Panels (a)-(c): Posterior means of differential abundances θ_{jk} for $k = 1, 2, 3$, respectively, along with 95% credible intervals and reference lines. Panel (d): Posterior estimates of κ_{jk} for cases of (j, k) with $\kappa_{jk}^{\text{TR}} = 0$, i.e., when OTU j is absent in all samples with level k.

further enhances the estimation of θ_{jk}. Figure 3.3 shows posterior inference for communities through $\hat{f}_k^{\xi}$ and $\hat{f}_k^{\theta}$. In each panel, the posterior estimates are shown by dashed colored lines with shaded 95% pointwise CIs, and the simulation truth is shown in solid black. From the plot, the BNP regression approach flexibly captures non-Gaussian patterns such as bimodality and skewness in the distributions. Even for levels $k = 2, 3$, where many OTUs are not present, the model produces good



(a) f_1^{ξ} (b) f_2^{ξ} (c) f_3^{ξ}

(d) f_1^{θ} (e) f_2^{θ} (f) f_3^{θ}

Figure 3.3: [Simulation 1] Panels (a)-(c) shows posterior estimates of f_k^{ξ} for each k, $k = 1, 2, 3$, and panels (d)-(f) of f_k^{θ}. Dashed colored lines are estimates with shaded 95% pointwise credible intervals. Black solid lines represent the simulation truth. Rugs show ξ_{jk}^{TR} and θ_{jk}^{TR}.

estimates of f_k^{θ}, potentially because it borrows information across different levels through the DDP as well as across different OTUs. We also examined estimates of baseline counts of OTU j in sample i, $r_i + \alpha_{jm}$. These estimates are shown in supplementary Figure B.1. The posterior estimates recover the true baseline counts well. There is no indication that the model suffers identifiability problems. Posterior predictive performance indicates reasonable model fit. Replications Y_{ij}^{rep} from the posterior predictive distribution were compared to the observed counts Y_{ij}. Figure 3.4 compares the average value of these replicates $\hat{Y}_{ij}^{\text{rep}}$ to Y_{ij} on the log scale. An offset of 0.1 was added so that the zero counts could be visualized.

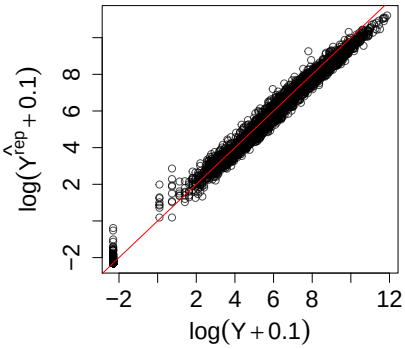


Figure 3.4: [Simulation 1] Average value of replicated counts Y_{ij}^{rep} drawn from the posterior predictive distribution compared to Y_{ij}.

The figure provides no evidence indicating serious failings of the model. We also computed 95% posterior predictive intervals and compared them to the observed values. We observed that the posterior predictive intervals are conservative.

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

Comparison We used 100 simulated datasets to compare results of our BNP-ZIMNR to those of alternative models: A Bayesian nonparametric multivariate regression model with NB (BNP-MNR), a Bayesian nonparametric multivariate Poisson regression model with zero inflation (BNP-ZIMPR), a Bayesian nonparametric multivariate NB regression model with zero inflation but with fixed normalization factors (FN-BNP-ZIMNR), the zero inflated overdispersed Poisson (ZoP) model (Jonsson *et al.*, 2018) and edgeR (Robinson *et al.*, 2010). BNP-MNR is similar to our BNP-ZIMNR, but does not include the submodel in (3.1) for zero inflation. BNP-ZIMPR is likewise similar to BNP-ZIMNR, but uses a Poisson likelihood instead of a negative binomial likelihood. FN-BNP-ZIMNR incorporates the same elements as BNP-ZIMNR, but uses fixed normalization (FN) factors $r_i = \log(Y_{i\bullet})$ rather than using the mean-constrained prior specification of (3.4). ZoP is a Bayesian generalized linear model that uses a zero inflated Poisson distribution for OTU counts, and beta and normal priors for the probability of being zero and the regression coefficients, respectively. Under ZoP, each Y_{ij} has a random effect, i.e., sample and OTU specific random effects to handle overdispersion. EdgeR, one of popular likelihood based methods, uses a NB generalized linear regression approach. It uses OTU specific plugin estimates for the normalization factors produced by an empirical Bayes strategy and analyzes individual OTUs separately. EdgeR does not include random effects for the group factor and does not account for the dependence among samples taken from the same subject. ZoP and edgeR set one level of a factor as a reference level to formulate the regression, and their regression coefficients represent differential abundance compared to the abundance in the reference level. ZoP uses the pseudo count approach when all samples of the reference level have zeros. Both methods include library sizes $Y_{i\bullet}$ as plugin offsets for normalization. EdgeR has an option to use empirically pre-

Model	δ_{ij}	$\theta_{j2} - \theta_{j1}$	$\theta_{j3} - \theta_{j1}$	μ_{ij}
BNP-ZIMNR	**0.019 (0.005)**	**0.308 (0.060)**	**0.325 (0.057)**	**3,154 (818)**
BNP-MNR	–	3.909 (0.504)	4.762 (0.504)	6,5190,628 (89,816,163)
FN-BNP-ZIMNR	0.021 (0.005)	2.234 (0.279)	2.386 (0.263)	4,680 (2,032)
BNP-ZIMPR	0.022 (0.006)	1.650 (0.282)	1.706 (0.258)	3,686 (1,289)
ZoP	0.200 (0.033)	2.759 (0.278)	3.156 (0.249)	3,769 (1,281)
edgeR	–	2.218 (0.303)	2.693 (0.303)	7,924 (1,860)

(a) Parameter Estimation

Model	$\theta_{j1} - \theta_{j3}$	$\theta_{j2} - \theta_{j3}$
BNP-ZIMNR	**0.325 (0.057)**	**0.393 (0.054)**
BNP-MNR	4.762 (0.504)	4.468 (0.446)
FN-BNP-ZIMNR	2.386 (0.263)	0.610 (0.182)
BNP-ZIMPR	1.706 (0.258)	0.617 (0.126)
ZoP	4.348 (0.356)	3.636 (0.388)
edgeR	2.693 (0.303)	3.302 (0.380)

(b) Estimation of Difference in θ with $k = 3$ as a Reference

Table 3.1: [Simulation 1: Comparison] RMSEs of δ_{ij}, $\theta_{jk} - \theta_{j1}$, $k = 2, 3$, and μ_{ij} are shown in (a). Performance metric averages over 100 simulated datasets with standard deviations in parenthesis. $k = 1$ is used as the reference group for the difference in θ. For (b), $k = 3$ is used as the reference group and RMSE of $\theta_{jk} - \theta_{j3}$, $k = 1, 2$ is given.

estimated sample size factors instead of $Y_{i\bullet}$, but we used their default option using $Y_{i\bullet}$.

For comparison, we fit each of the models and compared parameter estimates to their truth using root mean square error (RMSE). The different formulation for the regression model under ZoP and edgeR precludes a direct comparison of their differential abundance estimates to θ_{jk}^{TR}. As an alternative, we arbitrarily set the reference to the first level k = 1 and compare the model performances on the estimation of differences $\theta_{jk} - \theta_{j1}$, $k = 2, 3$. The RMSE computed for δ_{jk}, $\theta_{jk} - \theta_{j1}$ and μ_{jk} is shown in Table 3.1(a). For BNP-MNR, we used the posterior mean estimates of μ_{ij} as a point estimate $\hat{\mu}_{ij}$. For the zero inflated models, similar to $\hat{\theta}$ we computed $\hat{\mu}_{ij} = \sum_{b=1}^{B} \delta_{ij}^{(b)} \times \mu_{ij}^{(b)}/B$. BNP-ZIMNR outperforms the other methods in comparison for estimating δ_{ij} and $(\theta_{jk} - \theta_{j1})$. BNP-ZIMNR is the best

performer in terms of estimating μ_{ij}, closely followed by ZoP and BNP-ZIMPR. Due to OTU and sample specific random effects under ZoP, it obtains good estimates of μ_{ij}, but may tend to overfit the data, leading to worse estimates for $(\theta_{jk} - \theta_{j1})$, as is indicated by model comparison described later. The detrimental impact of excluding zero inflation can be seen by the much larger RMSE of μ_{ij} for the BNP-MNR. Failing to account for overdispersion biases the estimates of θ_{jk}, as can be seen from the performance of BNP-ZIMPR. Fixed normalization factors, like those used in FN-BNP-ZIMNR, lead to poorer estimates of both θ_{jk} and μ_{ij}. Since selecting a level for the reference is arbitrary, we re-fit the data using a different level of the factor as the reference for ZoP and edgeR and computed the RMSE of the differences in θ_{jk}. Table 3.1(b) illustrates the RMSE of $(\theta_{jk} - \theta_{j3})$ with $k = 3$ instead of k=1 as the reference level. Recall that level k=3 has a higher degree of zero inflation than level k=1 in the truth. The performances of ZoP and edgeR degrade when using this sparser factor level as the reference, indicating bias in the estimation of θ due to using arbitrary pseudo counts. In contrast, the inference on θ_{jk} under BNP-ZIMNR and BNP-MNR is invariant to the choice of reference level.

For further comparison of model fit among the Bayesian models, the log pseudo marginal likelihood (LPML) and the deviance information criterion (DIC) were calculated for BNP-ZIMNR, BNP-MNR, FN-BNP-ZIMNR, BNP-ZIMPR and ZoP. These metrics are summarized in Table 3.2(a). Similar to other information criterion, DIC assesses model performance based on the model's predictive accuracy with a penalty for model complexity (Spiegelhalter *et al.*, 2002). Lower values of DIC are preferred. LPML is the sum of the logarithms of conditional predictive ordinates (Gelfand *et al.*, 1992; Gelfand and Dey, 1994). It gives a measure of the leave-one-out cross validated posterior predictive prob-

Model	DIC	LPML
BNP-ZIMNR	**50,994 (1,107)**	**-26,391 (528)**
BNP-MNR	62,909 (1,317)	-32,328 (647)
FN-BNP-ZIMNR	51,780 (1,098)	-26,964 (529)
BNP-ZIMPR	128,182 (11,411)	-74,781 (6,319)
ZoP	2,598,963 (90,051)	-486,810 (30,653)

(a) DIC and LPML

Model	F_1^θ	F_2^θ	F_3^θ
BNP-ZIMNR	**0.158 (0.063)**	**0.195 (0.073)**	**0.163 (0.060)**
BNP-MNR	0.209 (0.069)	0.489 (0.033)	0.510 (0.039)
FN-BNP-ZIMNR	0.775 (0.029)	0.269 (0.108)	0.304 (0.116)
BNP-ZIMPR	0.795 (0.020)	0.317 (0.045)	0.278 (0.050)

(b) Total Variation Distance between $F_k^{\theta,\mathrm{TR}}$ and $\hat{F}_k^\theta$

Table 3.2: [Simulation 1: Comparison] (a) Average model comparison metrics over 100 simulated datasets with standard deviations in parenthesis. (b) Average total variation distance of F_k^θ as compared to the simulation truth both with and without zero inflation. Standard deviations in parenthesis.

ability, with higher values preferred. For more reliable comparison, we evaluated DIC and LPML based on the partially marginalized likelihood that integrates out random effects at the observation level for the ZoP (Millar, 2009). The table shows BNP-ZIMNR has greatly improved model fit compared to BNP-MNR, BNP-ZIMPR and ZoP. DIC and LPML based on the partially marginalized likelihood indicate that BNP-ZIMNR fits the data better, potentially implying overfit under ZoP. Different from ZoP and edgeR, the BNP models also provide community-level inferences. To assess the impact of omitting zero inflation, using fixed normalization factors, or using a Poisson likelihood in the estimation of F_k^θ, we considered the total variation distance between $F_k^{\theta,\mathrm{TR}}$ and $\hat{F}_k^\theta$ estimated from BNP-ZIMNR and its variants. Letting $\mathcal{B}$ denote the class of all Borel sets in $\mathbb{R}$, the total variation distance measures the closeness between two densities as $\sup_{B\in\mathcal{B}} \left| \int_B f_k^{\theta,\mathrm{TR}} d\theta - \int_B \hat{f}_k^\theta d\theta \right| = \frac{1}{2} \int \left| f_k^{\theta,\mathrm{TR}} - \hat{f}_k^\theta \right| d\theta$, where $f_k^{\theta,\mathrm{TR}}$ and $\hat{f}_k^\theta$ denote the densities of $F_k^{\theta,\mathrm{TR}}$ and $\hat{F}_k^\theta$ (Devroye and Lugosi, 2001). Table 3.2(b) shows the

computed total variation distances. We use median estimates of f_k^θ as our point estimate $\hat{f}_k^\theta$. The benefits of incorporating zero inflation into the model are clearly observed for estimating a distribution of differential abundances. The total variation distance under BNP-ZIMNR is notably reduced, especially for $k = 2$ and 3, the levels with higher probability of OTU absence. The use of fixed normalization factors or failing to accommodate overdispersion hinders estimation of F_k^θ as well, as seen by the inferior performance of FNP-BNP-ZIMNR and BNP-ZIMPR.

3.3.2 Simulation 2

In this section we present results from Simulation 2, which we performed as an additional assessment of the model's performance and scalability. The setup for Simulation 2 was similar to Simulation 1, but includes $K = 6$ different factor levels instead of $K = 3$ as was done in Simulation 1. We simulated data for 100 OTUs for 20 subjects, i.e., $J = 100$, $M = 20$, resulting in $n = 120$ samples, a covariate $x_i \in \{1, \ldots, 6\}$, $i = 1, \ldots, n$ and a grouping factor $u_i \in \{1, \ldots, 20\}$. We used Gaussian mixtures to set the simulation truth for $F_k^{\xi,\mathrm{TR}}$ and $F_k^{\theta,\mathrm{TR}}$, $k = 1, \ldots, 6$. We let $F_k^{\xi,\mathrm{TR}} = 0.6\,\mathrm{N}(-2, 0.25) + 0.4\,\mathrm{N}(-1, 0.5)$ for $k = 1$ and 6, $F_k^{\xi,\mathrm{TR}} = 0.2\,\mathrm{N}(-0.5, 0.25) + 0.8\,\mathrm{N}(0.5, 0.5)$ for $k = 2$ and 4, and $F_k^{\xi,\mathrm{TR}} = 0.5\,\mathrm{N}(0, 0.25) + 0.5\,\mathrm{N}(1, 0.5)$, $k = 3$ and 5. We let $F_k^{\theta,\mathrm{TR}} = 0.3\,\mathrm{N}(3, 0.25) + 0.6\,\mathrm{N}(2, 0.25) + 0.1\,\mathrm{N}(-1.5, 0.5)$, $k = 1$ and 4, $F_k^{\theta,\mathrm{TR}} = 0.3\,\mathrm{N}(2, 0.5) + 0.6\,\mathrm{N}(-1, 0.25) + 0.1\,\mathrm{N}(-2, 0.25)$, $k = 2$ and 5, and $F_k^{\theta,\mathrm{TR}} = 0.3\,\mathrm{N}(2, 0.5) + 0.35\,\mathrm{N}(-1, 0.25) + 0.35\,\mathrm{N}(-2, 0.25)$, $k = 3$ and 6. In the dataset used for the second simulation 10%, 58%, 62%, 57%, 66% and 14% of Y_{ij} were equal to 0, respectively, for the 6 factor levels. The hyperparameter and truncation levels were set in a manner similar to Simulation 1: $L^r = 20$, $\boldsymbol{a}_\psi^r = \mathbf{1}$, $a_w^r = 5$, $b_w^r = 5$, $u_r^2 = 0.05$, $b_{\eta^r}^2 = 0.25$, $L^\alpha = 150$, $\boldsymbol{a}_\psi^\alpha = \mathbf{1}$, $a_w^\alpha = 1$, $b_w^\alpha = 1$,

$u_\alpha^2 = 2$, $b_{\eta^\alpha}^2 = 1$, $L^\theta = 50$, $\rho^\theta = 1$, $\bar{\theta}^\star = 0$, $\tau_\theta^2 = 10$, $L^\xi = 50$, $\rho^\xi = 1$, $\bar{\xi}^\star = 2$, $\tau_\xi^2 = 1$, $a_\sigma^\xi = b_\sigma^\xi = a_\sigma^\theta = b_\sigma^\theta = 1$, $a_s = 0.3$, and $b_s^2 = 0.1$. $\upsilon_r = 0$ is set and υ_α was specified using the empirical approach described in §3.2.2. The MCMC was run for 70,000 iterations, discarding the first 20,000 iterations, and thinned to use every fifth sample, resulting in 10,000 samples from the posterior distribution. For this larger simulation every $5,000$ iterations of the MCMC took approximately 21 minutes on a 3.2GHz Intel i5-6500 CPU running Ubuntu Linux.

Figure 3.5 shows the resulting posterior inference on differential abundance parameters θ_{jk} for individual OTUs. The figure shows the model is able to provide good estimates for differential abundance under the larger simulation study. Figure 3.6(a) considers estimates of $\kappa_{jk} = 1\{\sum_{i=1;x_i=k}^{N} 1(\delta_{ij} = 1) > 0\}$, letting $\hat{\kappa}_{jk} = \frac{1}{B}\sum_{b=1}^{B} \kappa_{jk}^{(b)}$ when $\kappa_{jk}^{\text{TR}} = 0$ in the simulation truth (and b indices the MCMC iteration). Panel(b) shows ϵ_{jk} plotted against the simulation truth. The results indicate the model continues to do a good job handling zero inflation and OTU absence when fit on the larger dataset. Figures 3.7 and 3.8 show posterior inference on F_k^θ and F_k^ξ. As in simulation 1, the BNP approach is able to provide accurate community level inference, capturing the bimodality and skewness in the distributions. Figure 3.9 shows the average $\hat{Y}_{ij}^{\text{rep}}$ value of replicates drawn from the posterior predictive distribution plotted against the observed counts on the log scale. Like in simulation 1 we find the posterior predictive inference is reasonable.

Comparison We used 100 simulated datasets with $K = 6$ levels to compare results from BNP-ZIMNR to those from the alternative models. For ZoP and edgeR we set the reference to the first level $k = 1$. The method produces estimates of difference between θ_{jk} and θ_{j1}, i.e., $\theta_{jk} - \theta_{j1}$, $k = 2, \ldots, 6$. As in Simulation 1, we compared parameter estimates to their truth using the root mean square error (RMSE) of $\theta_{jk} - \theta_{j1}$, $k \neq 1$, δ_{ij}, and μ_{ij}. Table 3.3(a) and (b) summarize the results.

Model	δ_{ij}	μ_{ij}
BNP-ZIMNR	**0.033 (0.004)**	**1,025 (210)**
BNP-MNR	–	3,891,833 (3,865,212)
FN-BNP-ZIMNR	0.035 (0.005)	2,212 (1,220)
BNP-ZIMPR	0.035 (0.004)	1,547 (426)
ZoP	0.180 (0.020)	1,697 (423)
edgeR	–	3,652 (800)

(a) RMSE of δ and μ

Model	$\theta_{j2} - \theta_{j1}$	$\theta_{j3} - \theta_{j1}$	$\theta_{j4} - \theta_{j1}$	$\theta_{j5} - \theta_{j1}$	$\theta_{j6} - \theta_{j1}$
BNP-ZIMNR	**0.278 (0.056)**	**0.283 (0.051)**	**0.263 (0.047)**	**0.286 (0.058)**	**0.194 (0.057)**
BNP-MNR	3.165 (0.434)	4.054 (0.528)	3.627 (0.537)	4.069 (0.505)	0.570 (0.120)
FN-BNP-ZIMNR	2.205 (0.268)	2.373 (0.281)	0.934 (0.157)	2.342 (0.235)	1.393 (0.225)
BNP-ZIMPR	1.661 (0.253)	1.760 (0.263)	0.654 (0.146)	1.749 (0.229)	1.345 (0.245)
ZoP	2.590 (0.223)	2.942 (0.216)	2.099 (0.419)	2.922 (0.198)	1.424 (0.232)
edgeR	2.109 (0.238)	2.505 (0.259)	1.801 (0.296)	2.473 (0.254)	1.486 (0.223)

(b) RMSE of $\theta_{jk'} - \theta_{j1}$, $k' \neq 1$ with $k = 1$ as a Reference

Model	$\theta_{j1} - \theta_{j3}$	$\theta_{j2} - \theta_{j3}$	$\theta_{j4} - \theta_{j3}$	$\theta_{j5} - \theta_{j3}$	$\theta_{j6} - \theta_{j3}$
BNP-ZIMNR	**0.283 (0.051)**	**0.336 (0.050)**	**0.325 (0.045)**	**0.345 (0.043)**	**0.282 (0.047)**
BNP-MNR	4.054 (0.528)	3.903 (0.469)	4.131 (0.450)	4.274 (0.471)	4.059 (0.538)
FN-BNP-ZIMNR	2.373 (0.281)	0.597 (0.167)	1.542 (0.299)	0.637 (0.297)	1.039 (0.300)
BNP-ZIMPR	1.760 (0.263)	0.572 (0.120)	1.472 (0.243)	0.591 (0.129)	0.569 (0.186)
ZoP	4.179 (0.336)	3.417 (0.268)	3.346 (0.356)	3.475 (0.358)	2.551 (0.287)
edgeR	2.505 (0.259)	3.083 (0.325)	3.166 (0.303)	3.400 (0.319)	2.568 (0.320)

(c) RMSE of $\theta_{jk'} - \theta_{j3}$, $k' \neq 3$ with $k = 3$ as a Reference

Table 3.3: [Simulation 2: Comparison] Performance metric averages over 100 simulated datasets with standard deviations in parenthesis. $k = 1$ is used the reference group and RMSE of δ, μ and $\theta_{jk} - \theta_{j1}$, $k \neq 1$ are shown in (a) and (b). For (c), $k = 3$ is used as the reference and RMSE of $\theta_{jk} - \theta_{j3}$, $k \neq 3$ is computed.

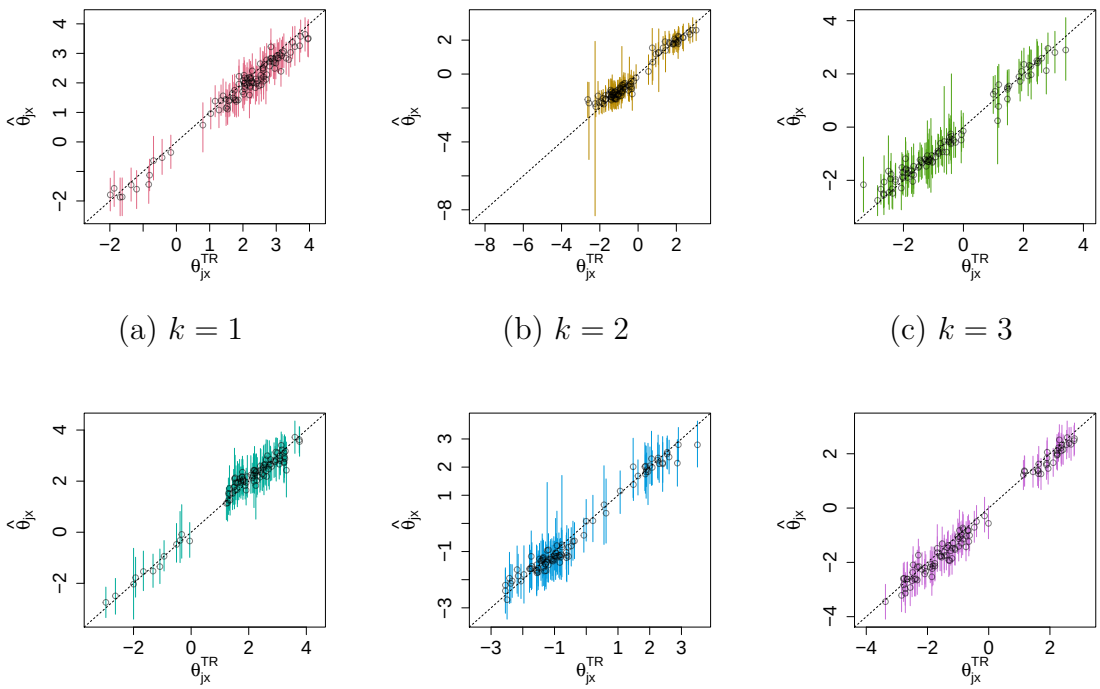
Model	DIC	LPML
BNP-ZIMNR	**84,548 (1,685)**	**-43,038 (832)**
BNP-MNR	112,868 (2,063)	-57,383 (1,004)
FN-BNP-ZIMNR	87,089 (1,738)	-44,667 (873)
BNP-ZIMPR	319,745 (19,183)	-186,249 (10298)
ZoP	3,892,519 (128,904)	-615,820 (29,828)

(a) DIC and LPML

Model	F_1^θ	F_2^θ	F_3^θ	F_4^θ	F_5^θ	F_6^θ
BNP-ZIMNR	**0.126 (0.052)**	**0.157 (0.058)**	**0.132 (0.046)**	**0.133 (0.049)**	**0.157 (0.059)**	**0.126 (0.047)**
BNP-MNR	0.203 (0.062)	0.453 (0.034)	0.472 (0.039)	0.432 (0.061)	0.493 (0.030)	0.182 (0.059)
FN-BNP-ZIMNR	0.749 (0.026)	0.296 (0.110)	0.337 (0.107)	0.450 (0.056)	0.369 (0.110)	0.168 (0.064)
BNP-ZIMPR	0.801 (0.018)	0.374 (0.073)	0.308 (0.081)	0.781 (0.025)	0.346 (0.071)	0.414 (0.065)

(b) Total Variation Distance between $F_k^{\theta,\mathrm{TR}}$ and $\hat{F}_k^\theta$

Table 3.4: [Simulation 2: Comparison] (a) Average model comparison metrics over 100 simulated datasets with standard deviations in parenthesis. (b) Average total variation distance of F_k^θ as compared to the simulation truth both with and without zero inflation. Standard deviations in parenthesis.

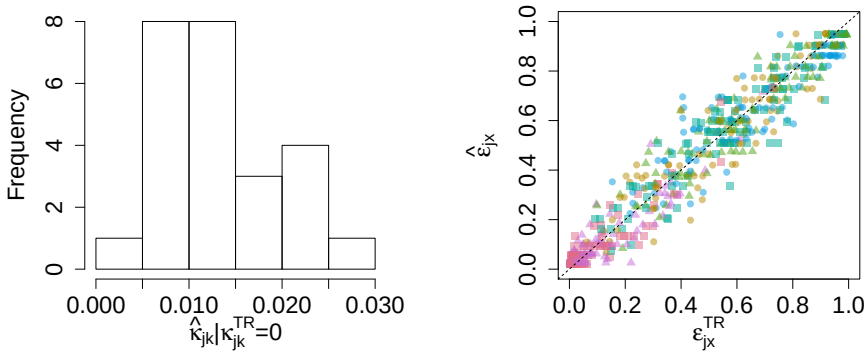


(a) $k = 1$ (b) $k = 2$ (c) $k = 3$

(d) $k = 4$ (e) $k = 5$ (f) $k = 6$

Figure 3.5: [Simulation 2] Posterior means of θ_{jk} for the six levels of the factor along with 95% credible intervals.

For each RMSE metric BNP-ZIMNR outperforms the alternative models. We examined the sensitivity of selecting k as the reference for ZoP and edgeR. For these two models we consider using $k = 3$ an alternative reference group and provide the RMSE for $\theta_{jk} - \theta_{j3}$, $k \neq 3$. The results are shown in Table 3.4(c). BNP-ZIMNR, BNP-MNR, FN-BNP-ZIMNR and BNP-ZIMPR do not require a reference group. BNP-ZIMNR's superior performance in terms of estimating $\theta_{jk} - \theta_{j3}$, $k \neq 3$ is even greater when the reference group is set to $k = 3$, which has greater zero inflation than the $k = 1$ level. DIC and LPML for BNP-ZIMNR, BNP-MNR, FN-BNP-ZIMNR, BNP-ZIMPR and ZoP are shown in the first panel of Table 3.4. For the DIC and LPML calculations the random effects from ZoP were marginalized



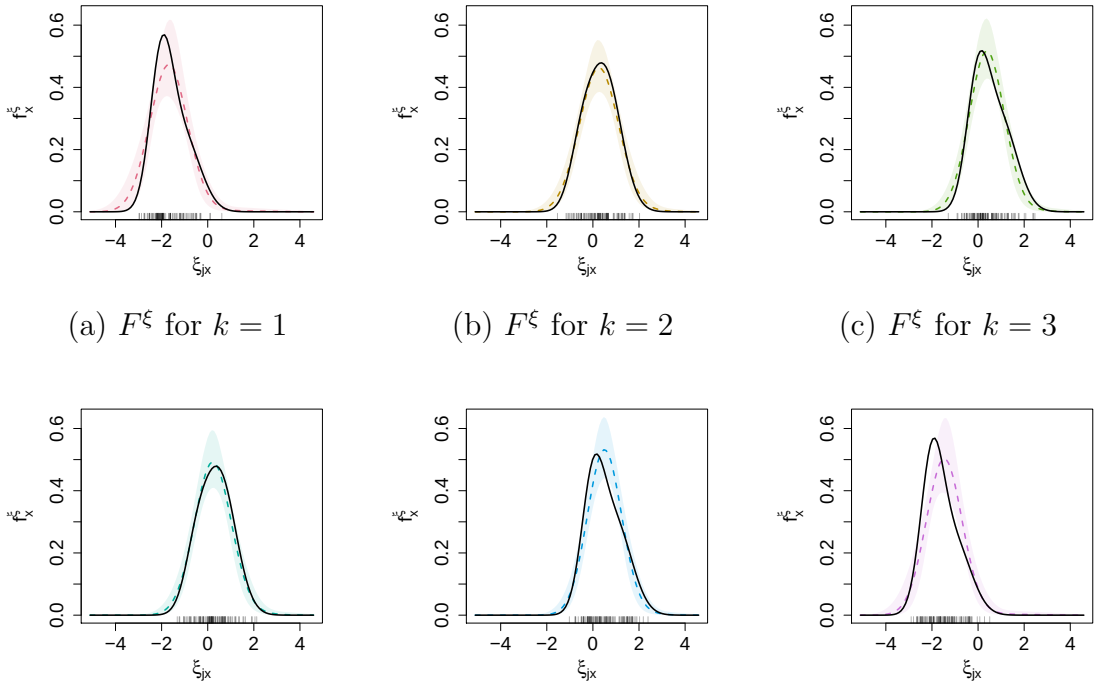
(a) (b)

Figure 3.6: [Simulation 2] Panel (a): Posterior estimates of κ_{jk} for cases of (j, k) with $\kappa_{jk}^{\text{TR}} = 0$, i.e., when OTU j is absent in all samples with level k. Panel (b): Posterior means of ϵ_{jk} plotted against the simulation truth. Shapes/colors indicate factor levels.

out as in Simulation 1, like is described in §3.3. BNP-ZIMNR outperformed the other models in terms of these predictive metrics. The advantage of BNP-ZIMNR over the alternative models is further illustrated by the second panel of Table 3.4 which lists the total variation distance of F_k^{θ} for the BNP-ZIMNR, BNP-MNR, FN-BNP-ZIMNR and BNP-ZIMPR. BNP-ZIMNR outperforms the model without zero inflation, the model with fixed normalization factors, and the model with a Poisson likelihood for all six of the regression coefficient distributions.

3.4 Chronic Wound Microbiome Data Analysis

In this section we apply BNP-ZIMNR to study chronic wound microbiomes using the dataset in Verbanic *et al.* (2019). The dataset consists of microbiome samples collected from $M = 18$ subjects with chronic wounds. Swab samples were collected from chronic wounds pre- and post-debridement, along with a healthy

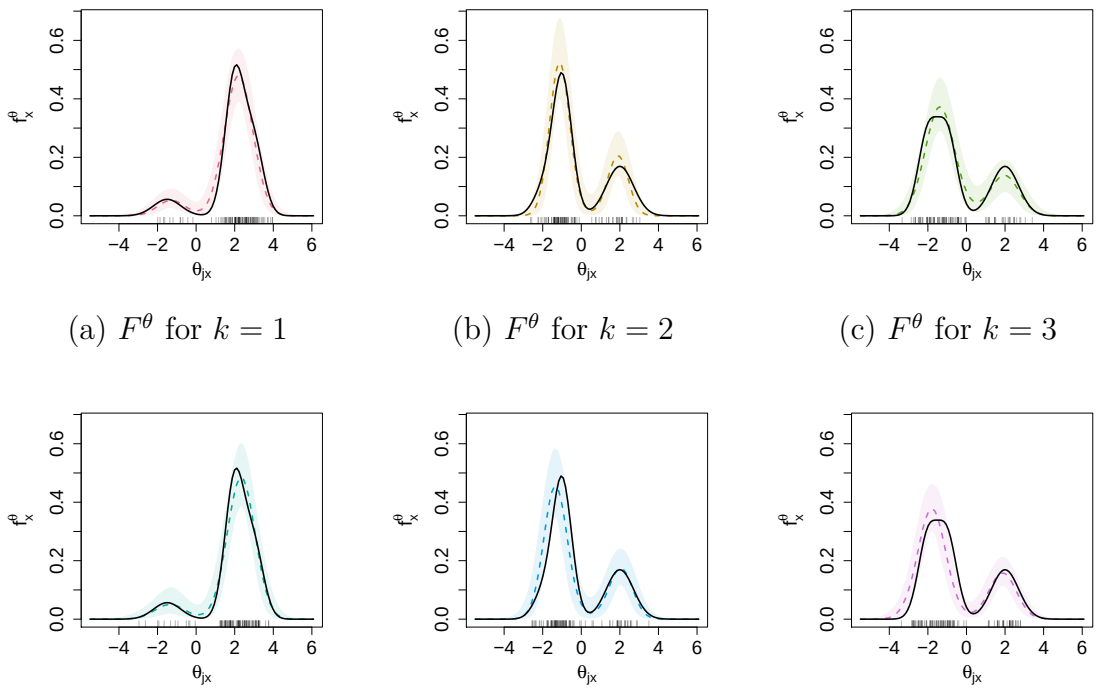


(a) F^{ξ} for $k = 1$ (b) F^{ξ} for $k = 2$ (c) F^{ξ} for $k = 3$

(d) F^{ξ} for $k = 4$ (e) F^{ξ} for $k = 5$ (f) F^{ξ} for $k = 6$

Figure 3.7: [Simulation 2] Posterior estimates of F_k^{ξ} for each k, $k = 1, \ldots, 6$. Solid black lines are the simulation truth. Shaded regions represent 95% pointwise credible intervals. Rugs show ξ_{jk}^{TR}.

skin swab sample from a control site, for each of the subjects. The $K = 3$ experimental conditions result in $n = 54$ samples in total. We let $k = 1, 2$, and 3 represent healthy skin, pre-debridement wound swabs, and post-debridement wound swabs, respectively. The study aims to investigate how debridement influences the composition of the microbial community of the wound, and also to compare the microbial composition of the wound surface to that of healthy skin. We analyzed the data to infer changes in the community-level microbial richness and diversity as well as differential abundances of individual OTUs. Better understanding of the wound microbiome and the effects of debridement on the wound microbiota can further elucidate the role of the microbiome on wound healing. From the



(a) F^θ for $k = 1$ (b) F^θ for $k = 2$ (c) F^θ for $k = 3$

(d) F^θ for $k = 4$ (e) F^θ for $k = 5$ (f) F^θ for $k = 6$

Figure 3.8: [Simulation 2] Posterior estimates of F_k^θ for each k, $k = 1, \ldots, 6$. Solid black lines are the simulation truth. Shaded regions represent 95% pointwise credible intervals. Rugs show θ_{jk}^{TR}.

swab samples, the 16S rRNA gene was amplified by PCR and sequenced using high throughput sequencing, and the sequence reads were organized into an OTU table for analysis. A total of 22,753 OTUs were observed after removing singletons. We restricted our attention to OTUs with nonzero counts in more than 20% of the samples for at least one experimental condition. After pre-processing, $J = 92$ OTUs were included in the analysis. The degree of zero inflation varies widely by experimental condition, with 8% of the OTU counts equal to zero from the healthy skin samples, versus 65% and 67% of the OTU counts equal to zero in the pre-debridement and post-debridement conditions, respectively. Figure 3.10 (a)-(c) illustrates histograms of the empirical proportions p_{jk} of zero counts in

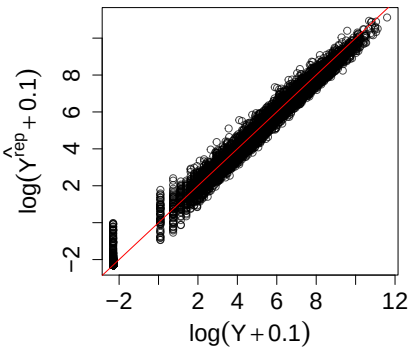
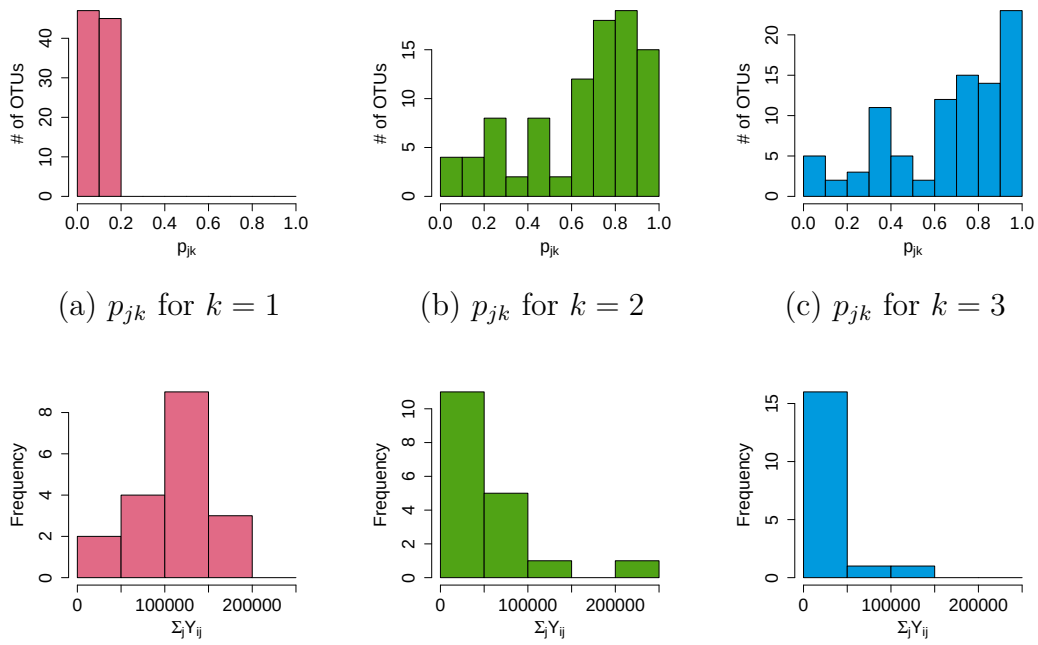


Figure 3.9: [Simulation 2] Average value of replicated counts Y_{ij}^{rep} drawn from the posterior predictive distribution compared to simulated counts Y_{ij}.

the samples for the conditions. Panels (d)-(f) show histograms of total counts $Y_{i\bullet}$ in samples for each k. From the figures, the samples from conditions $k = 2$ and 3 have more zeros and have lower total counts. The observed zeros in the pre/post-debridement conditions may be due to the absence of the OTUs under those conditions. Figure 3.12(a) compares the posterior mean estimates $\hat{\epsilon}_{jk}$ of ϵ_{jk} with the empirical proportions p_{jk}. For many OTUs in conditions $k = 2$ or 3 (green or blue), differences between $\hat{\epsilon}_{jk}$ and p_{jk} are relatively large for some (j, k). That is, the model infers that some zeros were observed even when OTUs were present, possibly because of the small total counts under those conditions as seen from Figure 3.10(e) and (f).

We specified hyperparameters similar to those in the simulations. The MCMC simulation was run over 140,000 iterations, with the first 40,000 iterations discarded as burn-in and every fifth sample kept as thinning and used for inference. The MCMC took approximately 11 minutes for every 5,000 iterations of the MCMC on a 3.2GHz Intel i5-6500 CPU running Ubuntu Linux.

Community level inference provided by f_k^{ξ} and f_k^{θ} is shown in Figure 3.11.

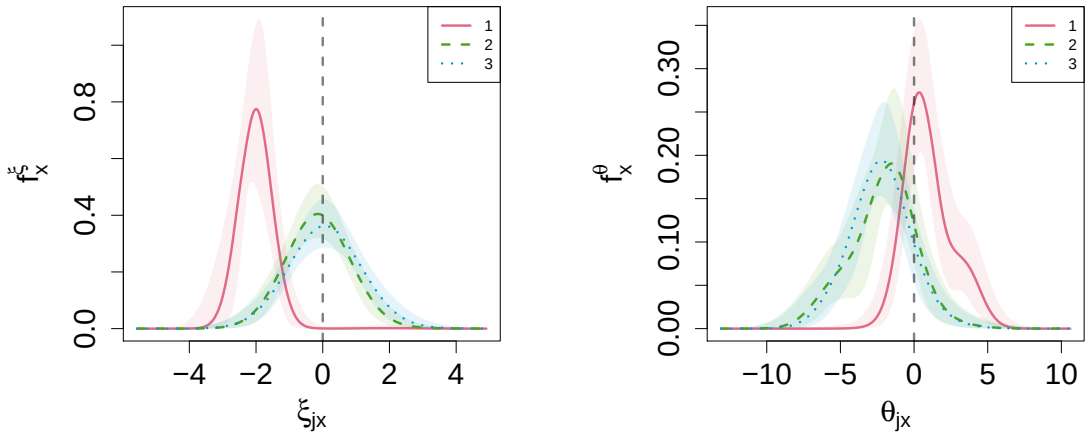


(a) p_{jk} for $k = 1$ (b) p_{jk} for $k = 2$ (c) p_{jk} for $k = 3$

(d) $Y_{i\bullet}$ for $k = 1$ (e) $Y_{i\bullet}$ for $k = 2$ (f) $Y_{i\bullet}$ for $k = 3$

Figure 3.10: [Chronic Wound Data] Panels (a)-(c): Histograms of empirical proportions of zero counts for each condition, $p_{jk} = \frac{1}{M}\sum_{i=1|x_i=k}^{n} 1(Y_{ij} = 0)$, $k = 1, 2, 3$, where $k = 1, 2, 3$ represents the healthy skin, pre- and post-debridement, respectively. Panels (d)-(f): Histograms of total OTU counts of samples for each experimental condition, $Y_{i\bullet}$ for $x_i = k$, $k = 1, 2, 3$.

Posterior estimates of f_k^{ξ} and f_k^{θ} are shown by the colored lines, with pointwise 95% CIs shown by the shaded regions, where the colors, red, blue and green, represents the healthy skin ($k = 1$), pre-debridement wound ($k = 2$), and post-debridement wound ($k = 3$), respectively. The differences between the estimates under the healthy skin condition and those under the wound conditions are substantial, but the wound microbial community does not change immediately after debridement, similar to the previous findings in Gardiner *et al.* (2017); Verbanic *et al.* (2019). In panel (a), $\hat{f}_k^{\xi}$ is stochastically lower for the healthy skin condition, suggesting greater species richness in a healthy skin sample than in a wound sample. For the wound conditions, $\hat{f}_k^{\xi}$ assigns more density to larger values and also has higher



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

Figure 3.11: [Chronic Wound Data] Estimates of f_k^{ξ} and f_k^{θ} are shown in panels (a) and (b). The three experimental conditions, healthy skin ($k = 1$), pre-debridement ($k = 2$) and post-debridement ($k = 3$), are indicated by the colors red, green and blue, respectively. 95% pointwise credible intervals for each condition are shown by the shaded areas.

dispersion. Panel (b) shows that $\hat{f}_k^{\theta}$ assigns more density to higher values in the healthy skin condition than in the pre-/post-debridement conditions. The bulk of the density for the wound conditions is given to values less than zero and the density estimates have long left tails. The distributions imply that on average OTUs in the wound conditions tend to have low abundance compared to their baseline.

The model also provides inference for individual OTUs. Figure 3.13 illustrates the posterior distributions of ϵ_{jk} and θ_{jk} for some selected OTUs, $j = 28$, 34 and 75. From panels (b), (c), (e) and (f), OTUs 34 and 75 that belong to genus *Micrococcus* and *Corynebacterium*, respectively, are highly abundant in skin, but not in wounds. The OTUs are absent in wounds with high probability. The increased likelihood of absence from wound samples and the depleted abundance in

wound samples when present are consistent with the previous findings in Verbanic *et al.* (2019) and Grice *et al.* (2009), indicating these OTUs are associated with a healthy skin microbiome. OTU $j = 28$ belonging to genus *Pseudomonas* is noted to be significantly associated with wounds (Verbanic *et al.*, 2019), and is also known to be a pathogen in chronic wounds (Wolcott *et al.*, 2016; Loesche *et al.*, 2017; Kalan *et al.*, 2019). However, panels (a) and (d) do not show a significant association with wounds. The lack of significant differences may be due to the high variability of wound composition among patients and small sample size. Posterior predictive checks indicate the model produces sensible inference. Figure 3.14 shows the average value of replicated counts Y_{ij}^{rep} drawn from the posterior predictive distribution plotted against the true counts on the log scale. An offset of 0.1 was added to Y_{ij}^{rep} and Y_{ij} so that zero values could be visualized. 97.5% of the true OTU counts were covered by their respective 95% credible intervals from the posterior predictive distribution, suggesting reasonable model fit. We also conducted sensitivity analyses to the specification of some fixed hyperparameters, L^{θ}, L^{ξ}, L^{r}, L^{α}, υ_r and υ_{α}. Changes in the posterior inference was minimal under these alternative specifications. More details are discussed in Supplementary §B.3.

The comparators are applied to the chronic wound data and their inferences are compared to the posterior inference under our BNP-ZIMNR. The healthy skin condition is used as the reference group for ZoP and edgeR to infer differential abundance for individual OTUs. Figure 3.12(b)-(f) compare estimates of $\theta_{jk} - \theta_{j1}$, $k = 2$ and 3, from the comparators to those from BNP-ZIMNR. In the figure, $\theta_{jk} - \theta_{j1}$'s are denoted by symbols + and x in green and blue for $k = 2$ and 3, respectively. From panels (b) and (e), OTUs that are less abundant in the wound conditions under BNP-ZIMNR tend to be less abundant to a greater degree under BNP-MNR and ZoP. FN-BNP-ZIMNR in panel (c), on the other hand, tends to

Model	DIC	LPML
BNP-ZIMNR	**28,271**	**-15,984**
BNP-MNR	30,904	-17,318
FN-BNP-ZIMNR	28,689	-16,262
BNP-ZIMPR	133,899	-73,816
ZoP	1,415,594	-249,758

Table 3.5: [Chronic Wound Data] Model comparison metrics for the chronic wound microbiome dataset.

predict higher OTU abundance under the wound conditions than BNP-ZIMNR. Panel (f) shows that edgeR also indicates greater abundance in the wound conditions for more OTUs, though less consistently than FN-BNP-ZIMNR. Table 3.5 shows DIC and LPML under the Bayesian models, BNP-ZIMNR, BNP-MNR, FN-BNP-ZIMNR, BNP-ZIMPR and ZoP for the chronic wound microbiome dataset. The metrics reported for ZoP were calculated using the partially marginalized likelihoods as was done in the simulation studies. BNP-ZIMNR outperforms the other models using both metrics. A possible explanation for the notably worse performance metrics for ZoP is most of the variability in the data is explained by sample and OTU specific random effects under that model. Because the model fit evaluation is based on marginalization over the random effects ZoP's performance metrics suffer.

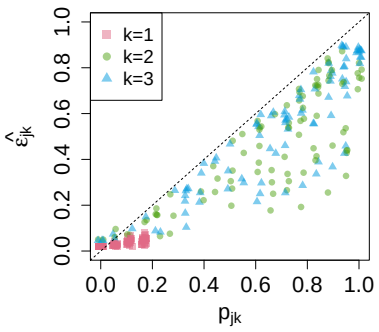
3.5 Discussion

We have presented a Bayesian nonparametric regression approach to model count data in the presence of high zero inflation, with application to microbiome studies. The model incorporates a DDP which avoids restrictive distributional assumptions, and flexibly estimates the degree of zero inflation and differential abundance across covariates and OTUs. Through this development we introduce

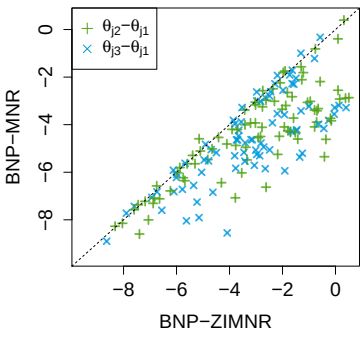
convenient methods for community level inference through examination of distributions related to taxa richness and differential abundance. Furthermore, by carefully considering the parameters' identifiability we remove the need to set a reference condition, allowing such community level inference to be made even when it is unclear which experimental condition should serve as the baseline, or when it is inconvenient to set an experimental condition as the baseline.

Our simulation studies showed that BNP-ZIMNR provides better estimation of differential abundance across different environmental factors or experimental conditions as compared to popular alternative models. The results indicate incorporating zero inflation into the model provided better estimation of OTU abundance and community level inference. The application of BNP-ZIMNR to analyze chronic wound microbiomes illustrates that the model successfully facilitates community-level inference.

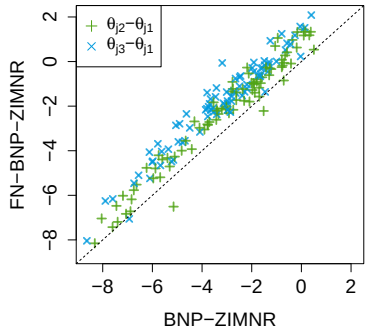
BNP-ZIMNR may be extended to accommodate more complex data structures, such as spatial and temporal dependence. Spatial and temporal changes in human microbiota were studied in Parfrey and Knight (2012) and Galloway-Peña *et al.* (2017). An extended variation of our BNP-ZIMNR can be used to characterize variability in microbiome over time and/or space at the community level as well as at the individual taxa level. The DDP has been successfully applied as a prior for a time series of random probability distributions (e.g., Griffin and Steel (2011) and Nieto-Barajas *et al.* (2012)). Also, Gelfand *et al.* (2005) and Duan *et al.* (2007) developed a variation of the DDP to flexibly model spatial dependence for point-referenced data. These are potential areas for future research.



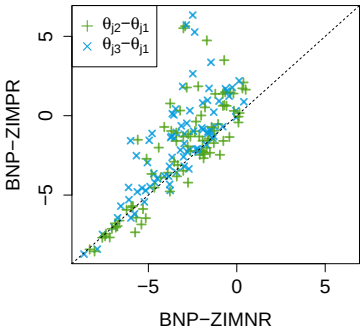
(a) $\hat{\epsilon}_{jk}$ vs p_{jk}



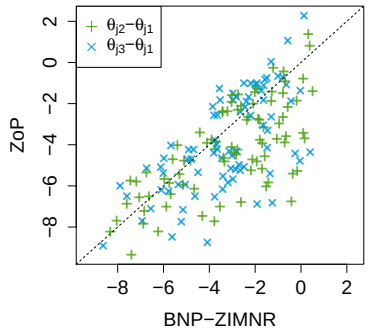
(b) BNP-ZIMNR vs BNP-MNR



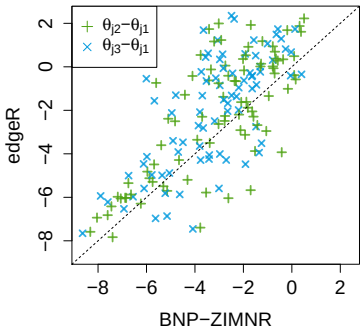
(c) BNP-ZIMNR vs FN-BNP-ZIMNR



(d) BNP-ZIMNR vs BNP-ZIMPR

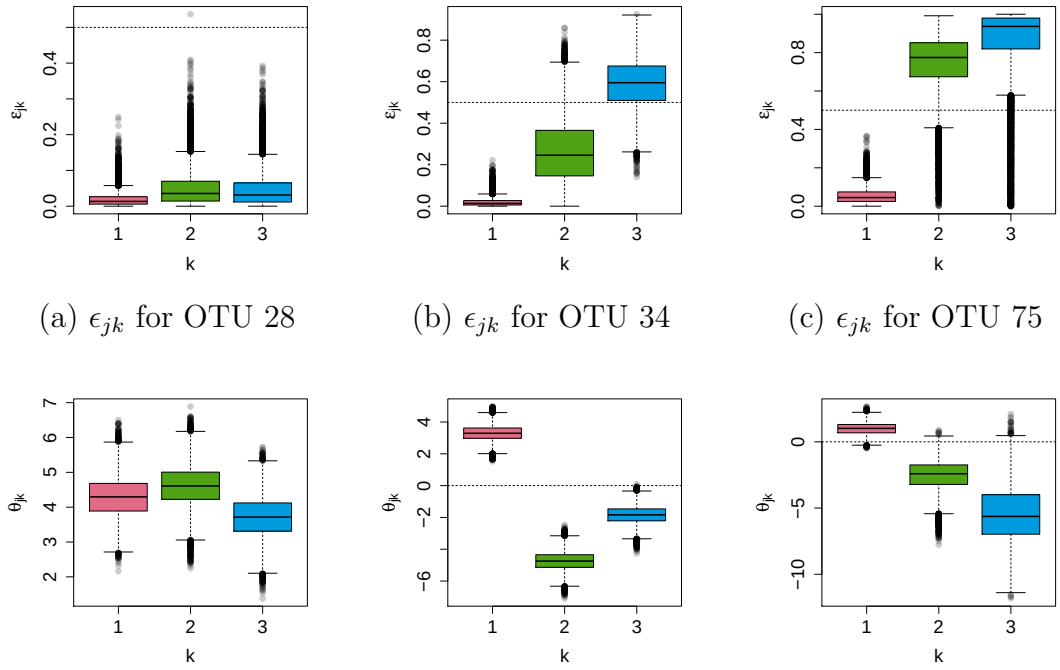


(e) BNP-ZIMNR vs ZoP



(f) BNP-ZIMNR vs edgeR

Figure 3.12: [Chronic Wound Data] Panel (a) shows a plot of empirical proportions p_{jk} of zero counts for each condition versus posterior mean estimates of ϵ_{kj}. Colors, red, green and blue represent different conditions (red for healthy, green for pre-debridement and blue for post-debridement). In panels (b)-(f) estimates of $\theta_{jk} - \theta_{j1}$, $k = 2$ and 3, under the comparators vs BNP-ZIMNR. Differences of the conditions, pre-debridement ($k = 2$) and post-debridement ($k = 3$), are indicated by the colors green and blue, respectively.



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

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

Figure 3.13: [Chronic Wound Data] Panels (a)-(c) illustrate the posterior distributions of ϵ_{jk} for each of the conditions for three selected OTUs $j = 28$, 34, 75. Panels (d)-(f) have the posterior distributions of θ_{jk}. $k = 1$, 2, and 3 denote healthy skin, pre-debridement, and post-debridement, respectively.

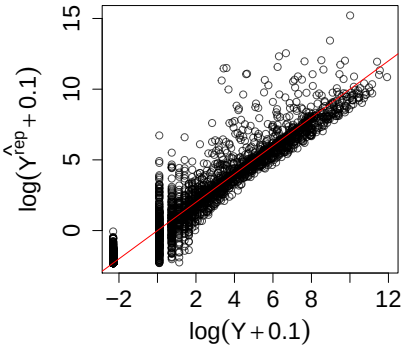


Figure 3.14: [Chronic Wound Data] Average value of replicated counts Y_{ij}^{rep} drawn from the posterior predictive distribution compared to the real OTU counts Y_{ij}.

Chapter 4

Bayesian Graphical Modeling of Microbial Community Composition

4.1 Introduction

Next generation sequencing technology has provided an advanced way to profile and analyze microbial communities and their environments. A typical analysis pipeline involves taking samples from the environment of interest and applying high-throughput sequencing (HTS) to produce read count data on the 16S rRNA gene of the taxa present in the environment. After sequencing, similar reads are grouped together into operational taxonomic units (OTUs), and the read counts of these OTUs are used for further downstream analysis. Such data can potentially answer key open research questions in biology, including the relationships among microbiota present in microbiomes and the effects of environmental factors on their abundances (Banerjee *et al.*, 2018; Gilbert *et al.*, 2018, 2012; Huttenhower *et al.*,

2012; Consortium, 2012). In many cases, the abundance levels of taxa present in the microbiome are not believed to be independent, but identifying symbiotic/antagonistic relationships among OTUs is challenging (Cirri and Pohnert, 2019; Faust and Raes, 2012; Kurtz *et al.*, 2015; Weiss *et al.*, 2016; Zhou *et al.*, 2010). Learning the latent dependence structure between OTUs from noisy data is oftentimes of primary interest to biologists, and accounting for the structure may further enhance inference on other parameters.

Graphical models are powerful tools in genomics and other studies of learning biological systems where network relationships are expected (for example, see Rodríguez *et al.* (2011); Ni *et al.* (2015); Peterson *et al.* (2015) among many others), but less so in microbiome studies. Graphical models allow for the mathematical expression of conditional independence structure between a set of random variables $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)'$, and provide estimates on the interaction patterns with associated uncertainties from noisy data. In a graphical model, variables are represented by a set of nodes corresponding to the variables $V = \{1, \ldots, J\}$ and their associated interactions are represented by edges E. A graph G is defined with a set of nodes V and a set of edges E. A graph characterizes conditional independence structure through presence or absence of edges between pairs of variables. The absence of an edge usually represents the conditional independence between the corresponding pair of random variables given a certain set of other variables. The certain set of other variables are defined by a chosen graph. Among various kinds of graphs, undirected graphs (UG) and directed acyclic graphs (DAG) are most commonly used because they can successfully learn the dependence structure in biological processes even with small sample sizes (Altomare *et al.*, 2013).

Although well suited to genomic studies, Gaussian graphical models typically cannot be directly applied to microbiome studies as the responses in microbiome

analysis are often multivariate count data coming from HTS. Furthermore, HTS samples must be normalized before analysis, otherwise spurious graphical relationships may be inferred (Faust and Raes, 2012; Faust *et al.*, 2012). For graphical analysis normalization must be applied both across samples, to account for varying levels of effort in the sequencing procedures; and across OTUs, to account for different abundance rates across taxa. Common approaches to normalization are to include total OTU counts as an offset in the model (e.g., Zhang *et al.* (2017a); Lee *et al.* (2018)) or to use a multinomial likelihood for the OTU counts conditioning on the total counts (e.g., Wadsworth *et al.* (2017); Tang and Chen (2018)). However, these may lead to undesirable inferences such as underestimated uncertainty with the resulting inference or an unappealing assumption on the relationship between OTUs. In particular, a multinomial model inherently induces a negative correlations between taxa OTU counts, which is at odds with biological knowledge suggesting some taxa will have mutualistic relationships. A simple and common approach to identify graphical structure is using marginal correlations between OTU counts or normalized counts (e.g., see Faust and Raes (2012); Berry and Widder (2014); Layeghifard *et al.* (2017) among many others). Deng *et al.* (2012) empirically standardized OTU counts to have mean zero and variance one in a sample and built a network between OTUs based on Pearson correlations between the standardized counts. They used random matrix theory to automatically select a threshold for similarity scores based on the correlations. Another example is Lee *et al.* (2018), which models the OTU counts through a zero inflated Poisson regression model. They incorporated correlation matrices to account for the taxa dependence structure for zero inflation and abundances. Likely due to their simplicity, marginal correlation based methods are the most common method, but often fail to infer complex interplay between OTUs. Our

simulation studies indicate that the use of marginal correlations to identify graphical structure may lead to noisier estimates of OTU relationships than encoding those through a graph directly, which may be related to the paucity of samples as compared to taxa. Oftentimes analyses of microbiome data is further complicated due to dependence structure arising from the sampling procedure, such dependence from samples taken across time/space, or dependence coming from multiple samples taken from the same subject, must be accounted for before the graphical structure can be accurately inferred.

In this work, we develop a Bayesian regression model with a graph (called BRM-G) to to simultaneously identify interactions between OTUs and estimate effects of covariates on OTUs' abundances. Our BRM-G assumes a negative binomial model to properly model OTU counts with potential overdispersion. The model normalizes mean counts across samples to reflect the data's compositional nature, as well as across taxa to account for different taxa abundance rates, using the mean-constrained mixture-of-mixtures model in Li *et al.* (2017). The model-based normalization allows accurate quantification of inferential uncertainty for the underlying microbioal structure. BRM-G utilizes a DAG (Pearl, 1988) through normalized OTU abundances for theoretical and computational convenience. A DAG directly encodes conditional dependence/independence relationships among OTUs, and assumes that the prevalence of one OTU may influence the expected prevalence of another OTU in a conditionally independent manner. Careful normalization prevents spurious edge detection in the graph and helps answer key biological questions with higher accuracy. Conditional independence relationships do not uniquely specify a DAG, and DAGs with the same set of conditional independence relationships are said to be *Markov equivalent* (Pearl, 1988). Without outside knowledge such as the prior ordering of OTUs or experimental intervention

the underlying DAG structure can only be recovered up to Markov equivalence (Altomare *et al.*, 2013; Ni *et al.*, 2015, 2018; Radhakrishnan *et al.*, 2018). For this reason we focus on recovery of the DAG's 'moralized' counterpart. The model also simultaneously infers the association of the normalized abundances with covariates through regression. BRM-G builds a hierarchical model to integrate information across OTUs and samples for improved inference. We show the advantages of the graphical approach we propose by comparing it to similar models that either omit this graphical component or use marginal correlations between OTUs through simulation studies. We illustrate the method with the chronic wound microbiome dataset in Verbanic *et al.* (2019). They studied the effects of debridement, a treatment for chronic wounds, on the microbiome of human subjects with stalled skin healing.

The remainder of the chapter describes BRM-G and its application to the chronic wound microbiome dataset. We describe the model, its assumptions, and the procedures we used for fitting the model in §4.2. In §4.3 we illustrate the performance of the model on synthetic experiments and compare it to other models for microbiome analysis. §4.4 describes the results from fitting the model to the chronic wound microbiome dataset, and we conclude with §4.5 with a brief discussion and potential areas for further research.

4.2 Probability Model

4.2.1 Sampling Model

Assume that non-negative counts Y_{ij} for each of the OTUs under consideration are observed in n samples, where $i = 1, \ldots, n$ and $j = 1, \ldots, J$. Each sample has an associated covariate vector $\boldsymbol{x}_i' = [x_{i1}, \ldots, x_{ip}]$, such as experimental conditions

associated with sample i, and an associated grouping factor $u_i \in \mathcal{U} = \{1, \ldots, M\}$, such as the subject from which the sample was obtained. The OTU counts are organized into an $n \times J$ table of counts, $\boldsymbol{Y}$. The pattern over OTUs may be similar in the samples taken from a subject or in the samples collected under the same experimental conditions. Also, the counts of an OTU may depend on those of the other OTUs due to their interactions in a sample. We model the counts using a negative binomial (NB) model as

$$Y_{ij} \mid \mu_{ij}, s_j \overset{indep}{\sim} \text{NB}(\mu_{ij}(\boldsymbol{x}_i, u_i, \boldsymbol{\mu}_{i,-j}),\, s_j), \tag{4.1}$$

where $\mu_{ij}(\boldsymbol{x}_i, u_i, \boldsymbol{\mu}_{i,-j})$ is the mean abundance of OTU j in sample i, and $\boldsymbol{\mu}_{i,-j}$ the vector of $\mu_{ij'}$ with μ_{ij} removed, i.e., $\boldsymbol{\mu}_{i,-j} = (\mu_{i1}, \ldots, \mu_{i,j-1}, \mu_{i,j+1}, \ldots, \mu_{iJ})$. $\mu_{ij}(\boldsymbol{x}_i, u_i, \boldsymbol{\mu}_{i,-j})$ is sample and OTU specific, and is a function of covariates ($\boldsymbol{x}_i$), subjects (u_i) and the abundances ($\boldsymbol{\mu}_{i,-j}$) of the other OTUs in the sample. For notational simplicity in the following we denote the mean $\mu_{ij} = \mu_{ij}(\boldsymbol{x}_i, u_i, \boldsymbol{\mu}_{i,-j})$ if it is self-contained. We parameterize the model as

$$\text{P}(Y_{ij} \mid \mu_{ij}, s_j) = \frac{\Gamma\left(Y_{ij} + 1/s_j\right)}{Y_{ij}!\,\Gamma\left(1/s_j\right)} \left(\frac{\mu_{ij} s_j}{1 + \mu_{ij} s_j}\right)^{Y_{ij}} \left(\frac{1}{1 + \mu_{ij} s_j}\right)^{1/s_j},$$

enabling the overdispersion of each OTU to be modeled separately from the mean, with $\text{Var}(Y_{ij}) = \mu_{ij} + s_j \mu_{ij}^2$, and the equivalent Poisson model recovered as $s_j \to 0$. HTS data is known to exhibit a high degree of overdispersion, with the OTU counts across samples sometimes varying by several orders of magnitude. The Poisson distribution's mean is necessarily equal to its variance, potentially leading to biased parameter estimates or underestimated estimation uncertainty when the counts are overdispersed. As a result, models that do not account for potential overdispersion, such as the Poisson distribution, may not be well suited to model

OTU counts. Models based on the negative binomial distribution are more flexible in this regard, as the additional parameter s_j models overdispersion separately.

We create a log-linear model for the average OTU abundance by decomposing μ_{ij} to

$$\log\left(\mu_{ij}(\boldsymbol{x}_i, u_i, \boldsymbol{\mu}_{i,-j})\right) = r_i + \alpha_j + \theta_{mj}(\boldsymbol{\mu}_{i,-j}) + \boldsymbol{x}_i'\boldsymbol{\beta}_j. \tag{4.2}$$

The factor r_i represents a sample size factor to account for different library sizes across the samples, and allows for meaningful comparison across samples. The intercept term α_j accounts for different baseline abundances across OTUs. Together r_i and α_j represent the baseline rate of occurrence for the j^{th} OTU in the i^{th} sample. Correction for a sample specific factor r_i and an OTU-specific factor α_j facilitates meaningful comparison of OTU abundances across samples and OTUs. $\{\log\left(\mu_{ij}(\boldsymbol{x}_i, u_i, \boldsymbol{\mu}_{i,-j})\right) - r_i - \alpha_j\}$ can be viewed as a normalized factor of relative abundance of OTU j in a sample from subject u_i on the logarithmic scale. $\theta_{mj}(\boldsymbol{\mu}_{i,-j})$ is a random effect of OTU j in a sample taken from subject u_i, and allows for variability in OTU abundances across subjects. We assume dependence between OTU abundances through the normalized factors, i.e., $\theta_{mj}(\boldsymbol{\mu}_{i,-j}) = \theta_{mj}(\boldsymbol{\theta}_{m,-j})$, and consider a graphical model through θ_{mj} to encode dependencies among OTUs. Oftentimes, the dependence structure and effects of covariates are of primary inferential interest in microbiome studies.

4.2.2 Prior

A key feature of the model is a graphical component to characterize dependence between J OTUs. θ_{mj} in (4.2) forms the basis for the graphical component of the model. For large problems, such as large J, inferring the interaction pattern between OTUs is challenging. We utilize a DAG since it is flexible yet compu-

tationally tractable. Under a DAG, all the edges of the graph are directed and the graph has no cycles. We encode a DAG $G = \{V, E\}$, where $V = \{1, \ldots, J\}$ represents the set of J OTUs and and $E \subseteq \{(\ell \to j), \ell, j \in V, \ell \neq j\}$ the set of edges characterizing G. Under DAGs, edge $(\ell \to j) \in E$ indicates that OTU ℓ is a parent of OTU j, and let $\text{Pa}(j)$ be the set $\{\ell : (\ell \to j) \in E\}$ of OTUs which are parents of OTU j. DAGs assume that OTU j is conditionally independent of all of its nondescendants in G given its parents $\text{Pa}(j)$. In turn, it implies that an OTU is independent of all other OTUs given the set of OTUs consisting of its parents, its children, and the other parents of its children (Friedman and Koller, 2003). Given a DAG, the joint distribution of $\boldsymbol{\theta}_m$ can be written as the product of conditional densities of each of θ_{mj} conditioned on their parents. We thus have, for subject m,

$$\theta_{mj} \mid G, \boldsymbol{\gamma}, \boldsymbol{\theta}_{m,-j}, \sigma^2 \overset{indep}{\sim} \text{N}\left(\sum_{\ell \in \text{Pa}(j)} \gamma_{\ell j}\theta_{m\ell},\, \sigma^2\right), j = 1, \ldots, J, \tag{4.3}$$

where $\boldsymbol{\theta}_{m,-j}$ is the set of $\theta_{m,j'}$ values with θ_{mj} removed. Through the structure in (4.3), the normalized relative abundance of OTU j in a sample from subject m depends on those of the other OTUs having a directed edge towards OTU j. The strength and direction of these associations are controlled by coefficients $\gamma_{\ell j}$. We consider the normal-inverse-gamma prior for $\boldsymbol{\gamma}$ and σ^2; let $\gamma_{\ell j} \mid G, \sigma^2, \kappa \overset{iid}{\sim}$ $\text{N}(0,\, \kappa\sigma^2)$, and $\sigma^2 \sim \text{IG}(a_\sigma,\, b_\sigma)$, where κ, a_σ and b_σ are fixed hyperparameters. Estimating G can be viewed as a model selection problem for the linear regression model in (4.3). Specifying the fixed hyperparameter values such as κ, a_σ and b_σ is closely tied with the regularization problem (Chipman *et al.*, 2001; Raftery *et al.*, 1997). When expert information is available, subjective elicitation of those parameters is desirable. If not, we choose their values such that the prior is relatively flat over the region of plausible values of γ. In particular, we choose a

large enough value for κ to reduce prior influence on the estimation of G. Such a value of κ can also yield parsimonious and interpretable estimates of G by ignoring edges with negligible $\gamma_{\ell j}$. For example, Raftery *et al.* (1997) suggests $\kappa = 2.85^2$. We let $p(G) \propto P_G^{|E|}$, where $P_G \in (0, 1)$ is a fixed probability for including an edge and $|E|$ is the number of edges in the graph. Our prior on G is similar to the prior in Telesca *et al.* (2012) *a priori* assuming no edge and penalizing for adding edges. With small P_G, it induces parsimony in G. When expert knowledge on interactions in microbiome is available, one can construct G_0 using the information and define an informative prior $p(G)$ centered around G_0. For details, see Telesca *et al.* (2012). Another simple and common choice of $p(G)$ is a uniform prior over G. Together κ and p_G determine the complexity of G, and we conduct sensitivity analyses to examine robustness to changes in those.

Recall we use a log-linear regression model to accommodate effects of covariates on the OTU abundances. We let $\beta_{jp} \mid \tau_p^2 \overset{indep}{\sim} \mathrm{N}(0, \tau_p^2)$, and $\tau_p^2 \overset{iid}{\sim} \mathrm{IG}(a_\tau, b_\tau)$, where a_τ and b_τ are fixed hyperparameters. τ_p^2 is indexed by p but shared by all j. This prior allows for borrowing strength across OTUs and enhance inferences on β_{jp}. We next build priors for the sample scale factor r_i and the baseline abundance factor of an OTU, α_j. The raw OTU counts from HTS data do not reflect absolute OTU abundance in a sample, as the magnitude of the counts depends on the effort put into the sequencing procedure. In order to account for sequencing depth, the OTU counts must be normalized. Many normalization methods have been proposed, including rarefying the OTU counts to induce similar library sizes across samples before analysis, and using fixed plug-in estimates for r_i. A common choice is to let r_i be equal to the logarithm of the library size (e.g. Lee *et al.* (2018); Robinson *et al.* (2010); Zhang *et al.* (2017a), among others), although many other plug-in estimates have been proposed (Anders and

Huber, 2010; Bullard *et al.*, 2010; Weiss *et al.*, 2017; Witten, 2011). In order to avoid bias in posterior uncertainties that can accompany such approaches, we instead employ a model-based method for estimating r_i, imposing a moment constraint on its distribution to avoid identifiability issues. Due to the multiplicative structure of $\mathrm{E}(Y_{ij}) = \exp(r_i + \alpha_j + \boldsymbol{x}_i'\boldsymbol{\beta}_j + \theta_{mj})$ the individual factors r_i and α_j are not identifiable. To avoid issues in their estimation we use mean-constrained mixture-of-mixtures priors of Li *et al.* (2017) for r_i and α_j

$$\begin{aligned} r_i &\overset{iid}{\sim} \sum_{\ell=1}^{L^r} \psi_\ell^r \left\{ w_\ell^r \,\mathrm{N}(\eta_\ell^r, u_r^2) + (1 - w_\ell^r)\,\mathrm{N}\left(\frac{\upsilon_r - w_\ell^r \eta_\ell^r}{1 - w_\ell^r}, u_r^2\right)\right\}, \\ \alpha_j &\overset{iid}{\sim} \sum_{\ell=1}^{L^\alpha} \psi_\ell^\alpha \left\{ w_\ell^\alpha \,\mathrm{N}(\eta_\ell^\alpha, u_\alpha^2) + (1 - w_\ell^\alpha)\,\mathrm{N}\left(\frac{\upsilon_\alpha - w_\ell^\alpha \eta_\ell^\alpha}{1 - w_\ell^\alpha}, u_\alpha^2\right)\right\}. \end{aligned} \tag{4.4}$$

The inner mixture component of these distributions is a convex combination of Gaussian distributions, with $w_\ell^\chi \in (0,1)$, $\chi \in \{r, \alpha\}$, and outer mixture weights $\sum_\ell^{L^\chi} \psi_\ell^\chi = 1$, with a fixed, marginal mean of υ_χ. Following Li *et al.* (2017) we set the mean constraint for r_i to have no scaling adjustment and let $\upsilon_r = 0$. We use an empirical approach to set υ_α. We use $\tilde{r}_i = \log(Y_{i\bullet}/Y_{\bullet\bullet}) - \frac{1}{N}\sum_{i'} \log(Y_{i'\bullet}/Y_{\bullet\bullet})$ with $Y_{\bullet\bullet} = \sum_{i,j} Y_{ij}$ as mean zero empirical estimates of r_i and let $\upsilon_\alpha = \left[\sum_{i,j} \{\log(Y_{ij}+1) - \tilde{r}_i\}\right]/(n \times J)$. We have found inference is robust to misspecification of the mean constraints (Lee and Sison-Mangus, 2018; Shuler *et al.*, 2019a,b); a more detailed sensitivity analysis of the effects of the mean constraints on inference on G and β_{jp} is shown in Appendix §C.2.1. We fix the kernel variances u_χ^2, noting their specification is not critical, though the number of mixture components required to accurately describe the distribution may be larger if u_χ^2 does not match the scale of χ well. We let $\boldsymbol{\psi}_\ell^\chi = (\psi_1^\chi, \ldots, \psi_{L^\chi}^\chi) \sim \mathrm{Dir}(\boldsymbol{a}_\psi^\chi)$ and $w_\ell^\chi \overset{iid}{\sim} \mathrm{Be}(a_w^\chi, b_w^\chi)$, where $\boldsymbol{a}_\psi^\chi = (a_{\psi 1}^\chi, \ldots, a_{\psi, L^\chi}^\chi)$, a_w^χ and b_w^χ are fixed hyperparameters, and let $\eta_\ell^\chi \overset{iid}{\sim} \mathrm{N}(\upsilon_\chi, b_{\eta^\chi}^2)$ with $b_{\eta^\chi}^2$ fixed. The mixture-of-mixtures formulation

is highly flexible, and enables the estimation of r_i and α_j despite their lack of identifiability with minimal assumptions about their distributions. We complete the model by letting overdispersion parameters $s_j \overset{iid}{\sim}$ Log-Normal(a_s, b_s^2), where a_s and b_s^2 are fixed hyperparameters.

4.2.3 Posterior Computation

Let $\underline{\boldsymbol{\theta}} = [s_j, r_i, \alpha_j, (\psi_\ell^\chi, w_\ell^\chi, \eta_\ell^\chi, \chi \in \{r, \alpha\}), \beta_{jp}, \tau_p^2, \theta_{mj}, \sigma^2, \gamma_{\ell j}, G]$ be the vector of all unknown parameters. By Bayes' rule the joint posterior distribution of $\underline{\boldsymbol{\theta}}$ is given by $\mathrm{P}(\underline{\boldsymbol{\theta}} \mid \boldsymbol{Y}, \boldsymbol{X}) \propto \mathrm{P}(\underline{\boldsymbol{\theta}})\mathrm{P}(\boldsymbol{Y} \mid \boldsymbol{X}, \underline{\boldsymbol{\theta}})$. We sample from the joint posterior using MCMC methods, the majority of which are straightforward Gibbs and Metropolis-within Gibbs parameter updates. The space of graphs requires efficient algorithms, especially for high dimensional problems. To this end, we exploit MC^3 in Madigan *et al.* (1995); Giudici and Castelo (2003). Here we briefly describe the steps to update the graph. We split the graph update into three cases and update via a Metropolis step on a selected edge resulting in a proposed: (1) birth, (2) death, or (3) switch of the edge. At each MCMC iteration we choose an edge $(\ell \to j)$ at random. If $(\ell \to j) \notin E$ & $(j \to \ell) \notin E$ we propose birth through the addition of $(\ell \to j)$ to E. If $(\ell \to j) \in E$ we propose death by removing $(\ell \to j)$ from E. Finally, if $(\ell \to j) \notin E$ & $(\ell \leftarrow j) \in E$, we propose switching the edge direction by removing $(\ell \leftarrow j)$ and adding $(\ell \to j)$ to E. When necessary, we propose new values for $\gamma_{\ell j}$ by drawing them from its prior, and the acceptance probability can be easily evaluated. For better mixing we repeat the edge selection and graph update procedure several times at each iteration. At each iteration we also consider updating the graph by proposing a move of death of an existing edge for an OTU jointly with a birth of an edge for the OTU, i.e., proposing a switch of a parent. Our empirical examination of the performance of

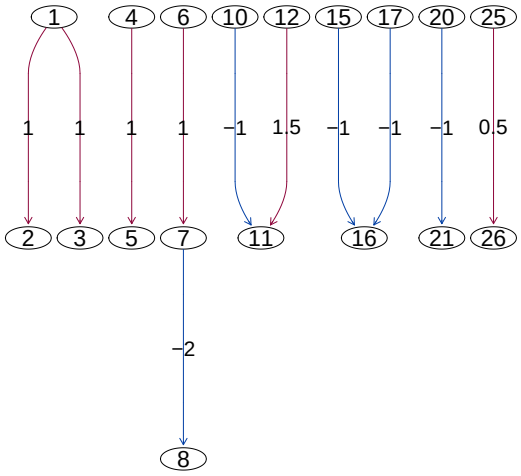
the algorithm through simulated data and real data does not indicate bad mixing or poor convergence. To improve convergence and mixing further, other methods can be considered. For example, see Grzegorczyk and Husmeier (2008), Barker *et al.* (2010) and Goudie and Mukherjee (2016). More details about the graph update and the other parameter updates are described in Appendix §C.1.

Although a DAG is an efficient tool for structure learning, our model cannot distinguish the DAGs in a Markov equivalence class, where all DAGs induce the same set of conditional independence structure relationships, from observation data (Chickering, 2002; Castelletti *et al.*, 2019). For more meaningful posterior inference, we learn the dependence structure encoded in DAGs through moral graphs G^m. A moral graph can be formed by 'marrying' parents nodes having a common child and then removing the graph's edge directions. Using the posterior Monte Carlo sample we approximately evaluate the marginal posterior $p(G^m \mid \boldsymbol{y})$ and determine a point estimate $\hat{G}^m$ for G^m. Specifically, we construct the moral graph $G^{m,(b)}$ from each MCMC sample of G, $G^{(b)}$ indexed by $b = 1, \ldots, B$. We let $m_{\ell j}^{(b)} \in \{0, 1\}$ indicate whether $G^{m,(b)}$ has an edge between OTUs ℓ and j. We let $\hat{G}^m$ be the point estimate obtained by including an edge if its posterior probability of inclusion $\bar{m}_{\ell j} = \frac{1}{B} \sum_{b=1}^{B} m_{\ell j}^{(b)}$ is > 0.5.

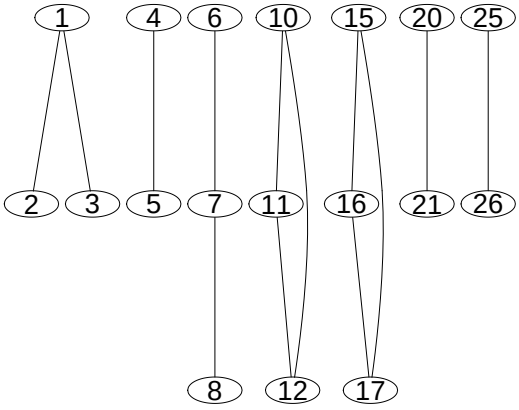
4.3 Simulation Studies

4.3.1 Simulation 1

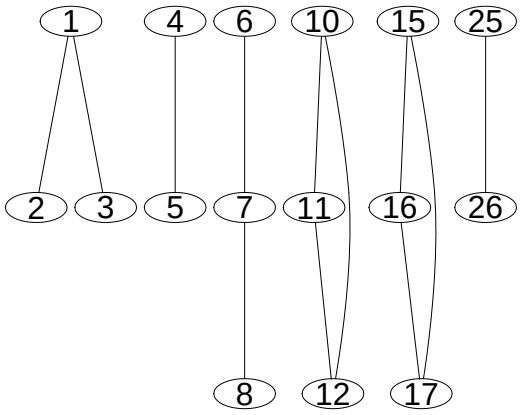
We evaluated the model's performance through simulation studies and compared the proposed model to alternative models. We consider a dataset comprised of simulated OTU counts for $J = 50$ OTUs from $M = 20$ subjects. As in the Chronic Wound Microbiome dataset we assume three experimental conditions,



(a) G^{TR} with $\gamma_{\ell j}^{\text{TR}}$



(b) True moral graph



(c) $\hat{G}^m$

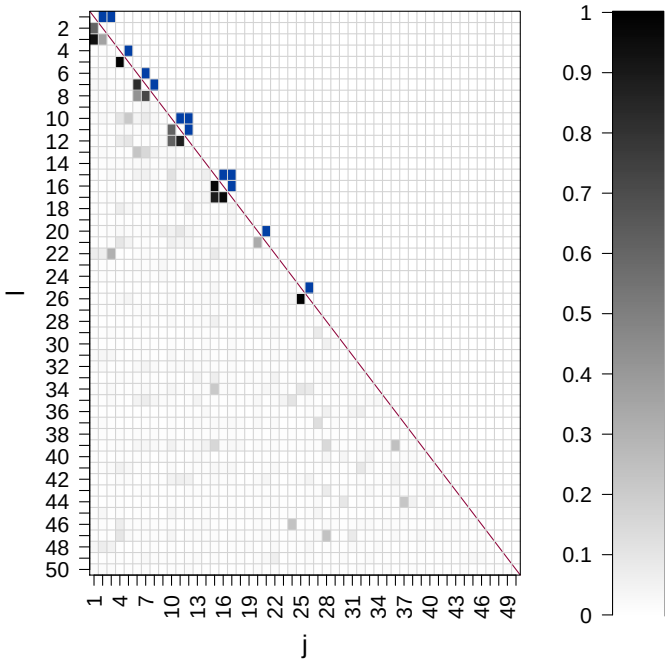
Figure 4.1: [Simulation 1 Truth] True DAG G^{TR} with its associated coefficients $\gamma_{\ell j}^{\text{TR}}$ is shown in (a), where positive effects are in red and negative effects in blue. Panels (b) and (c) show the true moral graph $G^{m,\text{TR}}$ and its posterior point estimate $\hat{G}^m$, respectively.

with one sample from each subject in each condition, resulting in $n = M \times 3 = 60$ samples in total and $P = 2$ dichotomous covariates indicating the experimental condition corresponding to each sample. We note that the model accommodates any covariate type. The true DAG G^{TR} assumed to generate data is shown in Figure 4.1(a). The true values $\gamma_{\ell j}^{\text{TR}}$ are shown along the edges of G^{TR}. We evaluate the model's ability to uncover relationships among OTUs by recasting the DAG into its corresponding moral graph, $G^{m,\text{TR}}$ and comparing it to a moral graph estimate $\hat{G}^m$ recovered by the model. $G^{m,\text{TR}}$ is illustrated in Figure 4.1(b). For OTUs with $\text{Pa}(j) = \emptyset$ and $j \in \text{Pa}(\ell)$ for any $\ell \neq j$, we set θ_{mj}^{TR} by sampling uniformly from $\{-3, 0, 3\}$; for OTUs with $\text{Pa}(j) \neq \emptyset$ we set θ_{mj}^{TR} as in equation (4.3), with $\gamma_{\ell,j}^{\text{TR}}$ specified above and $\sigma^{2,\text{TR}} = 1/2$; for the remaining OTUs with $\text{Pa}(j) = \emptyset$ and $j \notin \text{Pa}(\ell)$ for any $\ell \neq j$ we let $\theta_{mj}^{\text{TR}} \sim \text{N}(u_{mj}, 0.25)$ with u_{mj} sampled uniformly from $\{-1.5, 1.5\}$. We generated the true values for the regression coefficients by letting $\beta_{jp}^{\text{TR}} \overset{iid}{\sim} \text{N}(0, \tau_p^2)$, with $\tau_p^{\text{TR}} \overset{iid}{\sim} \text{Gamma}(5, 5)$ parameterized such that $\text{E}(\tau_p^2) = 1$. We let $\alpha_j^{\text{TR}} \overset{iid}{\sim} \text{N}(7, 2^2)$, $r_i^{\text{TR}} \overset{iid}{\sim} \text{N}(-5, 1)$, $s_j^{\text{TR}} \overset{iid}{\sim} \text{Log-Normal}(-2, 0.01)$. OTU counts were generated by setting μ_{ij}^{TR} using (4.2) and drawing Y_{ij} from the NB distribution with μ_{ij}^{TR} and s_j^{TR}.

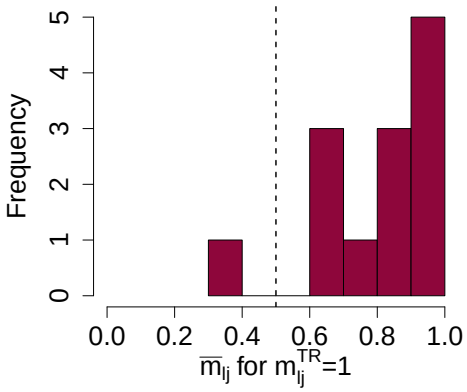
For the model fit we set the hyperparameters to $a_s = \log(0.01)$, $b_s^2 = 0.02$, $\boldsymbol{a}_\psi^r = \mathbf{1}$, $a_w^r = b_w^r = 5$, $u_r^2 = 0.05$, $b_{\eta^r}^2 = 0.25$, $\boldsymbol{a}_\psi^\alpha = \mathbf{1}$, $a_w^\alpha = b_w^\alpha = 5$, $u_\alpha^2 = 2$, $b_{\eta^\alpha}^2 = 2$, $a_\tau = 1$, $b_\tau = 1$, $a_\sigma = b_\sigma = 1$, $\kappa = 10$, and $P_g = 0.05$; and we set the number of mixture components for the mean-constrained priors for r_i and α_j to $L^r = 15$ and $L^\alpha = 5$. We used empirical partial correlations of $\log(Y_{ij} + 1)$ to initialize G; letting $(\ell \to j) \in E$, $\ell < j$ if the absolute value of a partial correlation is > 0.5. We only consider edges with OTUs $\ell < j$ to ensure the resulting G is a DAG. We then initialized $\gamma_{\ell j} \overset{iid}{\sim} \text{N}(0, 0.1^2)$ for the edges in the initial G. We also used the empirical estimates $\tilde{\boldsymbol{r}} = [\tilde{r}_1, \ldots, \tilde{r}_n]'$ to initialize the normal-

ization factors. To initialize α_j, β_{jp}, and θ_{mj} we used estimates obtained from a likelihood-based fit of a linear mixed-effects model of $\boldsymbol{x}_i$ onto $(\log(Y_{ij}+1)-\tilde{r}_i)$ individually for each OTU, with a random-intercept term for each subject. Because of the large number of potential OTU interactions we run the graph update step 625 times each MCMC iteration. We ran the MCMC chain for 140,000 iterations, discarding the first 40,000 iterations as burn-in and thinning every 10 iterations, resulting in 10,000 samples from the joint posterior distribution. On our 3.2GHz Intel i5-6500 machine it took approximately 7.5 minutes for every 10,000 draws. We analyzed the chain's convergence by examining parameter traceplots and comparing multiple chains with different initial values for the random parameters. We did not find evidence in this analysis suggesting the chain did not converge. More details about the chain's convergence are described in Appendix §C.2.1.

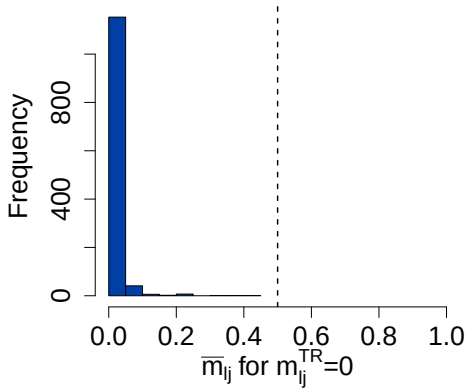
A point estimate $\hat{G}^m$ for the moral graph G^m and posterior inclusion probabilities of the edges $[\bar{m}_{\ell j}]$ are illustrated in Fig 4.1(c) and Fig 4.2(a), respectively. $\hat{G}^m$ recovers conditional independence structure reasonably well, but misses the edge between OTUs 20 and 21. $\bar{m}_{\ell j}$'s are larger for the edges with $m^{\text{TR}}_{\ell j}=1$, while small for the edges with $m^{\text{TR}}_{\ell j}=0$ as shown Fig 4.2(b) and (c). Fig 4.3(b) illustrates posterior estimates of $\gamma_{\ell j}$ given the directed edge $(\ell \to j)$ is included for the pairs of OTUs with $m^{\text{TR}}_{\ell j}=1$. In particular, we compute $\hat{\gamma}_{\ell j}=(\sum_{b=1}^{B} a^{(b)}_{\ell j}\gamma^{(b)}_{\ell j})/(\sum_{b=1}^{B} a^{(b)}_{\ell j})$, where $a^{(b)}_{\ell j}$ is a binary indicator taking 1 if edge $(\ell \to j)$ is included in $G^{(b)}$, or 0 otherwise. The posterior probability estimates of including directed edges $(\ell \to j)$ $\bar{a}_{\ell j}=\sum_{b=1}^{B} a^{(b)}_{\ell j}/B$, are shown in (a) of the figure. Although the directions of the edges are not recovered with high accuracy, $\hat{\gamma}_{\ell j}$ are very close to their truth conditional on the inclusion of the directed edges. Posterior inference on β_{jp} is shown in Figure 4.4 (a) and (d). We use the posterior mean $\hat{\beta}_{jp}$ as a point estimate and use vertical lines to illustrate the associated 95% credible intervals.



(a) $[\bar{m}_{\ell j}]$

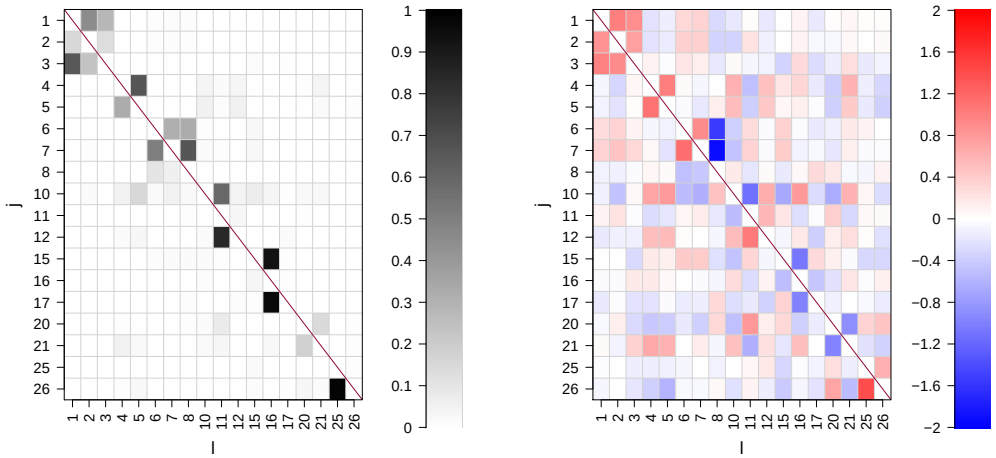


(b) $\bar{m}_{\ell j}$ for $m^{\text{TR}}_{\ell j} = 1$



(c) $\bar{m}_{\ell j}$ for $m^{\text{TR}}_{\ell j} = 0$

Figure 4.2: [Simulation 1] (a) $[\bar{m}_{\ell j}]$ under BRM-G is shown in the the lower triangle and $m^{\text{TR}}_{\ell j}$ in the upper triangle, where blue and white represent $m^{\text{TR}}_{\ell j} = 1$ and 0, respectively. Histograms of $\bar{m}_{\ell j}$ are in panels (b) and (c), separately for $m^{\text{TR}}_{\ell j} = 1$ and $m^{\text{TR}}_{\ell j} = 0$, respectively.

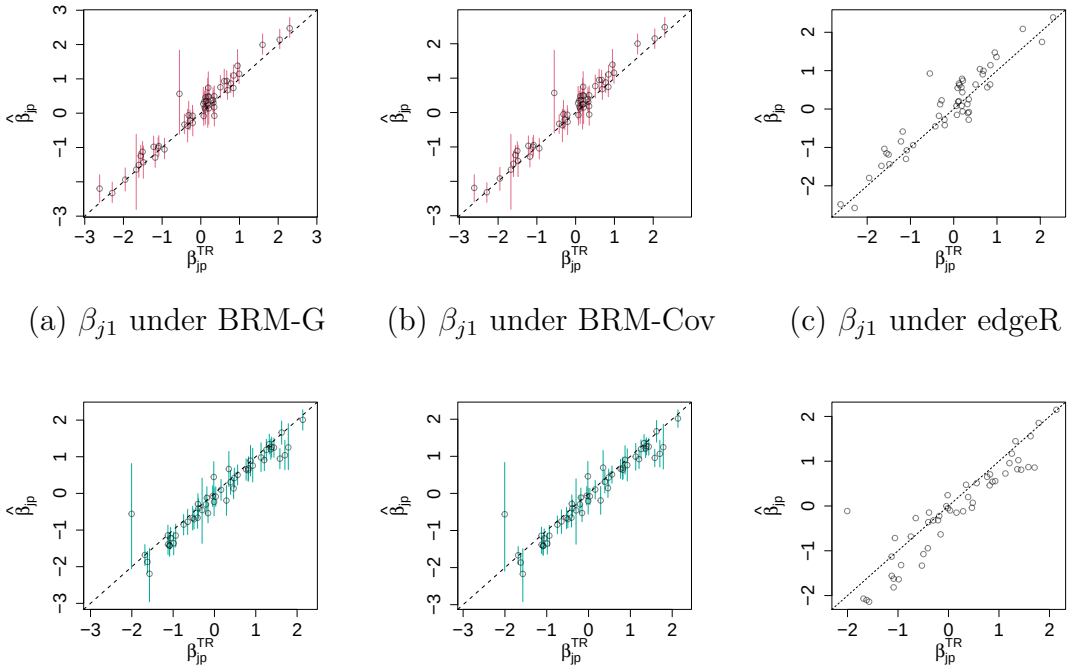


(a) $\bar{a}_{\ell j}$ (b) $\hat{\gamma}_{\ell j}$

Figure 4.3: [Simulation 1] Posterior estimates $\bar{a}_{\ell j}$ of the probabilities of including the directed edges $(\ell \to j)$ for the pairs with $m_{\ell j}^{\text{TR}} = 1$ are in panel (a). Panel (b) has the posterior mean estimates of $\gamma_{\ell j}$ given that $(\ell \to j)$ is included. The OTUs with $j \leq 26$ only are shown for better illustration.

The model yields reasonable estimates for the regression coefficients, accurately characterizing the direction and effect size of the different experimental conditions on OTU abundance. Individually r_i and α_j are not identifiable, but the baseline abundance $r_i + \alpha_j$ can be recovered. The estimation of r_i and α_j is shown in Figure 4.5, which shows the posterior means of those parameters plotted against the simulation truth. The estimates for α_j are smaller than the simulation truth, but this underestimation is compensated by estimates for r_i which are higher than the simulation truth. The central tendencies implied by υ_r and υ_α can be seen in the figure, with the estimates of r_i and α_j clustering around their respective mean constraints. On average the estimates for the baseline abundance $r_i + \alpha_j$ are unbiased, as is seen in the last panel of the figure. Unlike with plug-in normalizing factors uncertainty is propagated, resulting in more honest uncertainty quantification for the parameters of interest, β_{jp} and θ_{mj}.

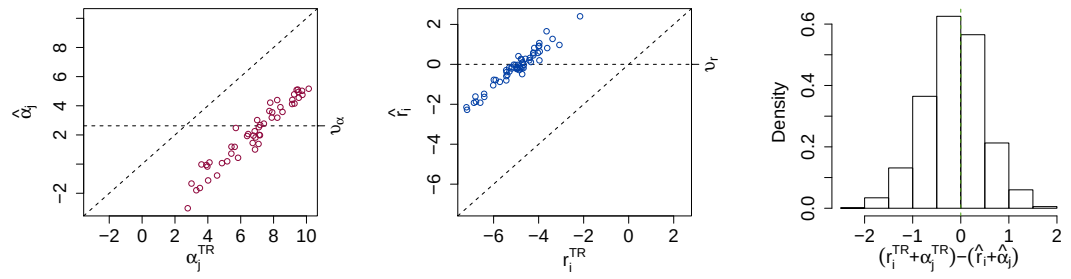
We assess the sensitivity of our inference with respect to the specifications



(a) β_{j1} under BRM-G (b) β_{j1} under BRM-Cov (c) β_{j1} under edgeR

(d) β_{j2} under BRM-G (e) β_{j2} under BRM-Cov (f) β_{j2} under edgeR

Figure 4.4: [Simulation 1] Posterior means $\hat{\beta}_{jp}$ and 95% credible intervals under BRM-G are plotted against the simulation truth β_{jp}^{TR} in (a) and (b) for $p = 1$ and 2, respectively.

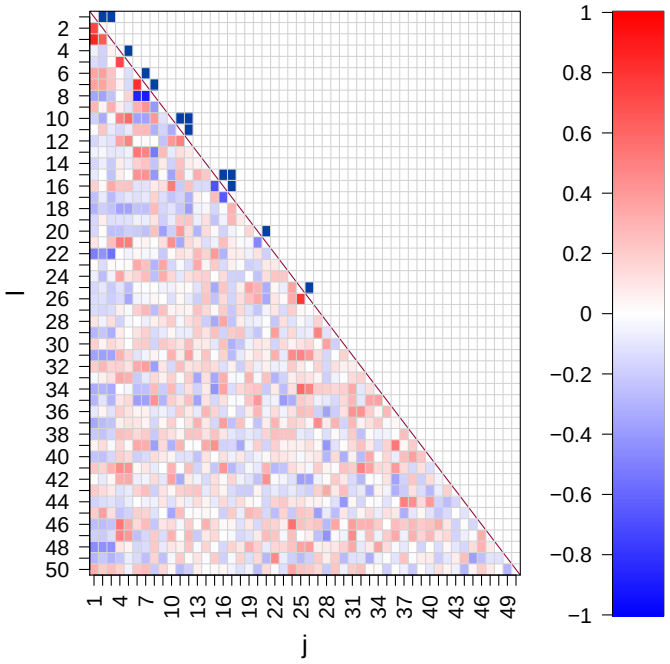


(a) α_j (b) r_i (c) Baseline differences

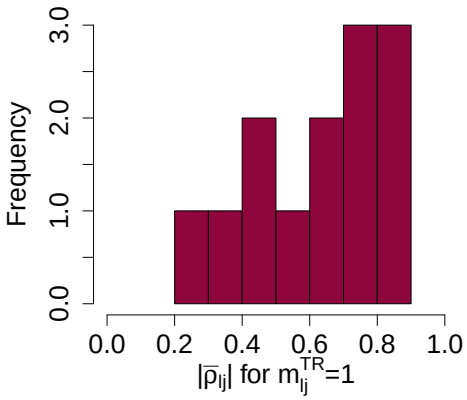
Figure 4.5: [Simulation 1] Posterior means $\hat{\alpha}_j$ and $\hat{r}_i$ are plotted against the simulation truth in panels (a) and (b). Horizontal reference lines show the parameters' respective mean constraints. In (c) differences of the baseline abundance from the simulation truth are compared to the baseline abundance estimated by $\hat{\alpha}_j + \hat{r}_i$.

of υ_r, υ_α, κ and P_G by repeating our analysis for a range of those values. We found that our inference on G^m is robust to varying the values of κ and P_G over reasonable ranges. Also, we observed that the specification of the values of υ_r and υ_α only minimally affects the baseline abundance estimation. More details are discussed in Supplementary §C.2.1.

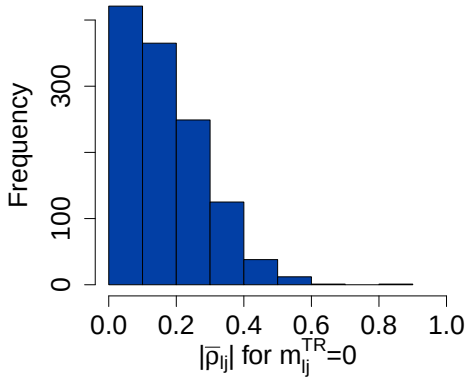
We compare BRM-G to two alternative models. We consider a model that replaces the graph in BRM-G with a covaraince matrix $\boldsymbol{S}$ for a J-dim vector $\boldsymbol{\theta}_m = [\theta_{m1}, \dots \theta_{mJ}]'$, while keeping the remaining parts of BRM-G the same. We call it a Bayesian regression model with a covariance matrix for interactions between OTUs (BRM-Cov). We let a J-dim vector of random effects $\boldsymbol{\theta}_m \mid \boldsymbol{S} \overset{iid}{\sim} \mathrm{N}_J(\mathbf{0}, \boldsymbol{S})$ and a $J \times J$ positive definite matrix $\boldsymbol{S} \sim \mathrm{IW}\,(J, \mathrm{diag}(10, \dots, 10))$. BRM-Cov is similar to our BRM-G, but omits the graph component in favor of subject-specific random effects for the OTU abundances. We simulated posterior samples from BRM-Cov using MCMC, similar to that used for BRM-G. For the second comparator, we include edgeR, a popular likelihood based method for microbiome analysis that does not include features for recovering graphical relationships among OTUs (Robinson *et al.*, 2010). edgeR is a negative binomial generalized log-linear model which uses the sample library sizes to generate plug-in estimates for the normalization factors and an empirical Bayes procedure to produce estimates for the OTUs' degrees of overdispersion. Posterior inference under BRM-Cov is summarized in Figure 4.6. We consider the pairwise correlations between OTUs produced by BRM-Cov as point estimates to infer OTUs' interactions. Specifically, we produce elementwise averages of posterior correlation samples, $\bar{\rho}_{\ell j} = \frac{1}{B}\sum_{b=1}^{B} \rho_{\ell j}^{(b)}$, where $\rho_{\ell j}^{(b)}$ is the pairwise correlation of OTUs ℓ and j computed from the b-th sample of $\boldsymbol{S}$, $\boldsymbol{S}^{(b)}$. $[\bar{\rho}_{\ell j}]$ is shown in panel (a) of the figure. For an easy comparison to the truth, $m_{\ell j}^{\mathrm{TR}}$ is in the upper triangle. $|\bar{\rho}_{\ell j}|$'s are large for the pairs of OTUs ℓ and j with $m_{\ell j}^{\mathrm{TR}} = 1$.



(a) $[\bar{\rho}_{\ell j}]$



(b) $|\bar{\rho}_{\ell j}|$ for $m^{\text{TR}}_{\ell j} = 1$



(c) $|\bar{\rho}_{\ell j}|$ for $m^{\text{TR}}_{\ell j} = 0$

Figure 4.6: [Simulation 1- BRM-Cov] (a) Elementwise posterior mean of pairwise correlations $[\bar{\rho}_{\ell j}]$ under BM-Cov is shown in the the lower triangle and $m^{\text{TR}}_{\ell j}$ in the upper triangle, where blue and white represent $m^{\text{TR}}_{\ell j} = 1$ and 0, respectively. Histograms of $|\bar{\rho}_{\ell j}|$ are in panels (b) and (c), for $m^{\text{TR}}_{\ell j} = 1$ and $m^{\text{TR}}_{\ell j} = 0$, respectively.

Model	β_{j1}	β_{j2}	μ_{ij}
BRM-G	**0.261 (0.065)**	**0.260 (0.054)**	10,533 (28,451)
BRM-Cov	0.262 (0.064)	0.260 (0.055)	**10,478 (27,981)**
edgeR	0.555 (0.313)	0.556 (0.236)	33,449 (83,049)

(a) Parameter Estimation

Model	DIC	LPML
BRM-G	**-1394 (168)**	**-11341 (869)**
BRM-Cov	-1382 (167)	-11,365 (865)

(b) Model Fit

Table 4.1: [Simulation 1] Performance metrics on 100 simulated datasets. RMSE's for β_{j1}, β_{j2}, and μ_{ij} are shown in (a). DIC and LPML for the Bayesian models are in (b). Standard deviations in parenthesis.

However, inference on the dependence structure among OTUs using $\rho_{\ell j}$ fails to recover important features of the true structure. In particular, $\bar{\rho}_{2,3} = 0.65$ although OTUs 2 and 3 are conditional independent given OTU 1 in the truth. Similarly, $\bar{\rho}_{6,8} = -0.86$ is far from satisfactory to infer their conditional independence given OTU 7. Furthermore, the v-structure among OTUs 10, 11 and 12 in the truth is not noticeable, whereas our $\hat{G}^m$ successfully detects such relationships. Panels (b) and (c) have histograms of $\bar{\rho}_{\ell j}$ for the pairs with $m_{\ell j}^{\text{TR}} = 1$ and 0, respectively. Compared to $\bar{m}_{\ell j}$ under BRM-G, $|\bar{\rho}_{\ell j}|$'s are more dispersed and a thresholding approach based on $|\bar{\rho}_{\ell j}|$ may lead to more incorrect conclusions. On the other hand, the inference on individual β_{jp} under BRM-Cov is almost the same as that under BRM-G, as shown in Figure 4.4(b) and (e). EdgeR produces estimates of β_{jp} but dose not attempt to infer any dependence structure between OTUs. Figure 4.4(c) and (f) compares the estimates of β_{jp} by edgeR to the truth.

To further compare the performance of BRM-G to that of the other models we fit each model to 100 simulated datasets. We investigated BRM-G's ability to recover the true graph on average as follows; we find $\hat{G}_k^m$, $k = 1, \ldots, 100$ for the simulated dataset as described earlier, and compute proportions of inclusion

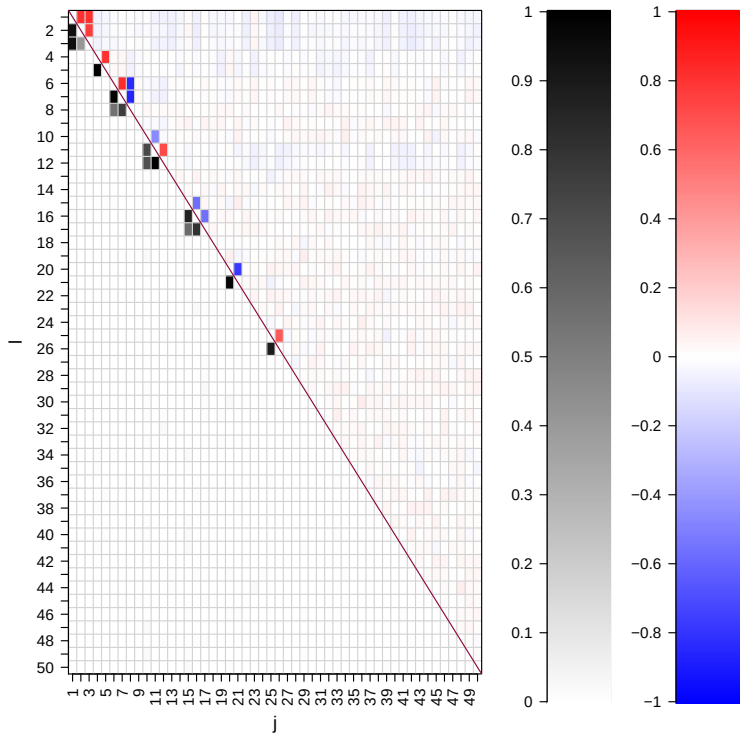


Figure 4.7: [Simulation 1] Results based on 100 simulated datasets. Proportions of edge inclusions over $\hat{G}_k^m$, $k = 1, \ldots, 100$ computed under BRM-G are in the lower triangle. Averages of $\bar{\rho}_{\ell,j,k}$, $k = 1, \ldots, 100$ computed under BRM-Cov are in the upper triangle.

of individual edges in $\hat{G}_k^m$'s. We let $m^\star_{\ell j}$ denote the proportion of inclusion of the edge between ℓ and j in $\hat{G}_k^m$. Similarly, we compute $\rho^\star_{\ell,j}$ by taking averages of $\bar{\rho}_{\ell,j,k}$ computed for the k-th simulated dataset over all datasets. Figure 4.7 illustrates $[m^\star_{\ell j}]$ and $[\rho^\star_{\ell j}]$ in the lower and upper triangles of a $J \times J$ matrix, respectively. The advantages of utilizing a graph over correlations are evident from its performance of recovering subgraphs with more than 2 OTUs. The nature of the relationships among OTUs in these groups is clearer under BRM-G than under BRM-Cov. In particular, BRM-G recovers the conditional independence structures well, e.g., dependence structures among OTUs 1, 2 and 3. Also, it detects the interrelationship between three OTUs in a v-shape form most of time,

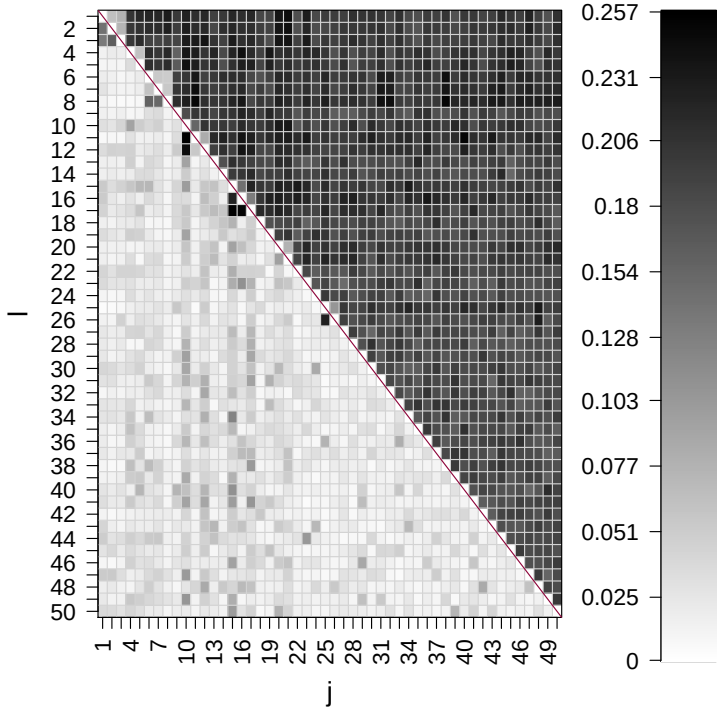


Figure 4.8: [Simulation 1] Variability of the OTU dependence structure over 100 simulated datasets. Standard deviation of posterior probabilities of edge inclusion under BRM-G and of the correlation estimates under BRM-Cov are shown in the lower and upper triangles, respectively.

e.g., dependence structures among OTUs 10- 12, and among 15-17. Figure 4.8 shows the standard deviations of the posterior correlation estimates $\bar{\rho}_{\ell,j,k}$ for BRM-Cov and of the posterior edge probabilities $\bar{m}_{\ell,j,k}$ for BRM-G. Note that their scales are different. Variability in the OTU relationships indicated by the models is higher for BRM-Cov than for BRM-G. The correlation estimates are notably noisier than the posterior edge probability estimates, especially when there is not an edge in the simulation truth. For some, but not all, of the OTU groups BRM-G's estimates are noisier when there is an edge in the simulation truth, suggesting BRM-G tends to be conservative with regard to identifying related OTUs than BRM-Cov.

As additional criteria to evaluate their performance we considered the root mean square error (RMSE) on β_{j1}, β_{j2} and μ_{ij}. For the Bayesian models we used the posterior means as point estimates for the parameter values. For BRM-G and BRM-Cov in addition to RMSE we also include model the Deviance Information Criterion (DIC) and log pseudo marginal likelihood (LPML). DIC is an information criterion similar to AIC for hierarchical models which simultaneously considers model fit and model complexity, with lower values indicating super model performance (Spiegelhalter *et al.*, 2002). LPML is a measure of the model's leave-one-out cross validation performance, using the likelihood as the evaluation criterion (Gelfand *et al.*, 1992; Gelfand and Dey, 1994). For LPML higher values indicate superior performance. The results of the model fits on the 100 simulated datasets are shown in Table 4.1. Both BRM-G and BRM-Cov produce better estimates for the regression coefficients and OTU abundance than edgeR. The performance of the two Bayesian models is very similar across all of the evaluation criteria, demonstrating it is possible for BRM-G to recover graphical relationships among the OTUs without diminishing its ability to estimate covariate effects on OTU abundance.

4.3.2 Simulation 2

In this section we present results from Simulation 2, which incorporates a larger graph with more complicated structure than the graph from Simulation 1. As in Simulation 1, in Simulation 2 we produce simulated OTU counts for $J = 50$ OTUs from $M = 20$ patients. We assume three experimental conditions and $P = 2$ dichotomous covariates for a total of $n = 60$ samples. For Simulation 2 we use a more complicated graph than Simulation 1 that has a greater number of relationships among OTUs and different sizes for the network effects. Simulation

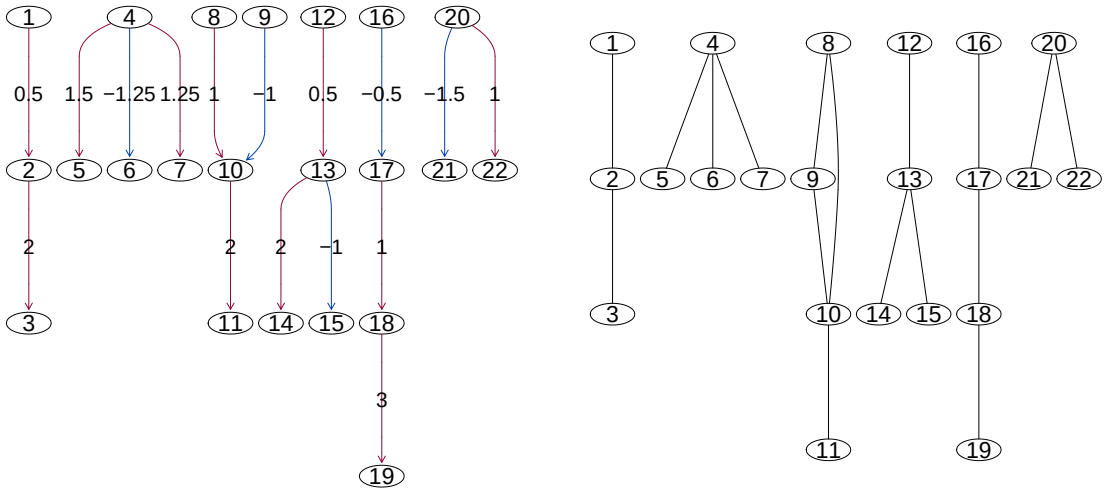
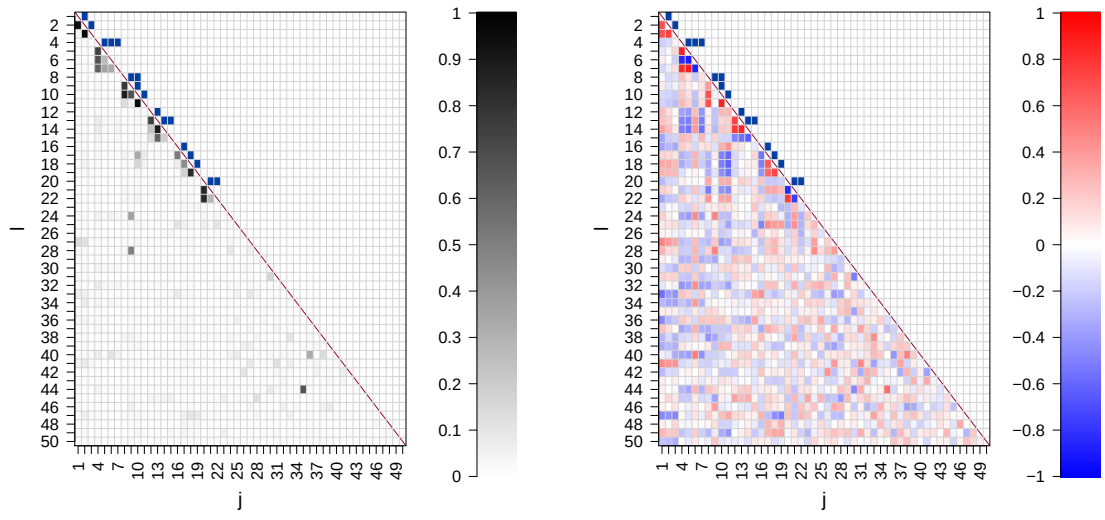


Figure 4.9: [Simulation 2 Truth] (a) True DAG and associated coefficients γ_{lj}^{TR}. Positive effects in red, negative effects in blue. (b) True moral graph.

2's true DAG and its corresponding moral graph are shown in Figure 4.9. The true moral graph and the corresponding posterior edge probabilities are illustrated in Figure 4.10. For comparison the point estimate for the correlation matrix $[\bar{\rho}_{\ell j}]$ produced by BRM-Cov using the methods described is also shown. The moral graph point estimate produced by BRM-G by including edges with posterior probability > 0.5 is shown in Figure 4.11. BRM-G generally does a good job recovering the graphical structure, although there are spurious edges between OTUs 9 and 28, and between OTUs 35 and 44. An edge between OTUs 17 and 18 is missing. Nonetheless, the OTU relationships are better recovered and more well defined under BRM-G than under BRM-Cov. Histograms of the posterior probabilities of edge inclusion under BRM-G and the pairwise correlations $\bar{\rho}_{lj}$ conditional on the true graph having/lacking an edge are shown in Figure 4.12. BRM-G more clearly discriminates when OTUs have/lack abundance relationships with other OTUs as compared to BRM-Cov. Estimates of the regression coefficients β_{jp} and



(a) $[\bar{m}_{\ell j}]$ (b) $[\bar{\rho}_{\ell j}]$

Figure 4.10: [Simulation 2] (a) Upper-diagonal: Edges of the true moral graph M^{TR}. Lower-diagonal: Posterior probabilities of edge inclusion $\bar{m}_{lj}$ under BRM-G. (b) Elementwise posterior mean of pairwise correlations $[\bar{\rho}_{\ell j}]$ under BM-Cov is shown in the the lower triangle and $m^{\text{TR}}_{\ell j}$ in the upper triangle, where blue and white represent $m^{\text{TR}}_{\ell j} = 1$ and 0, respectively.

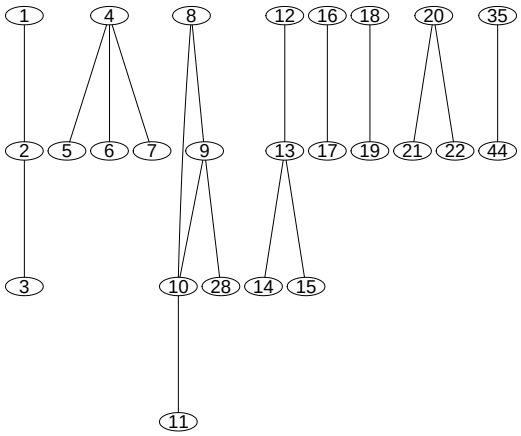


Figure 4.11: [Simulation 2] Estimated moral graph $\hat{G}^m$

corresponding 95% credible intervals are shown in Figure 4.13. Even with the

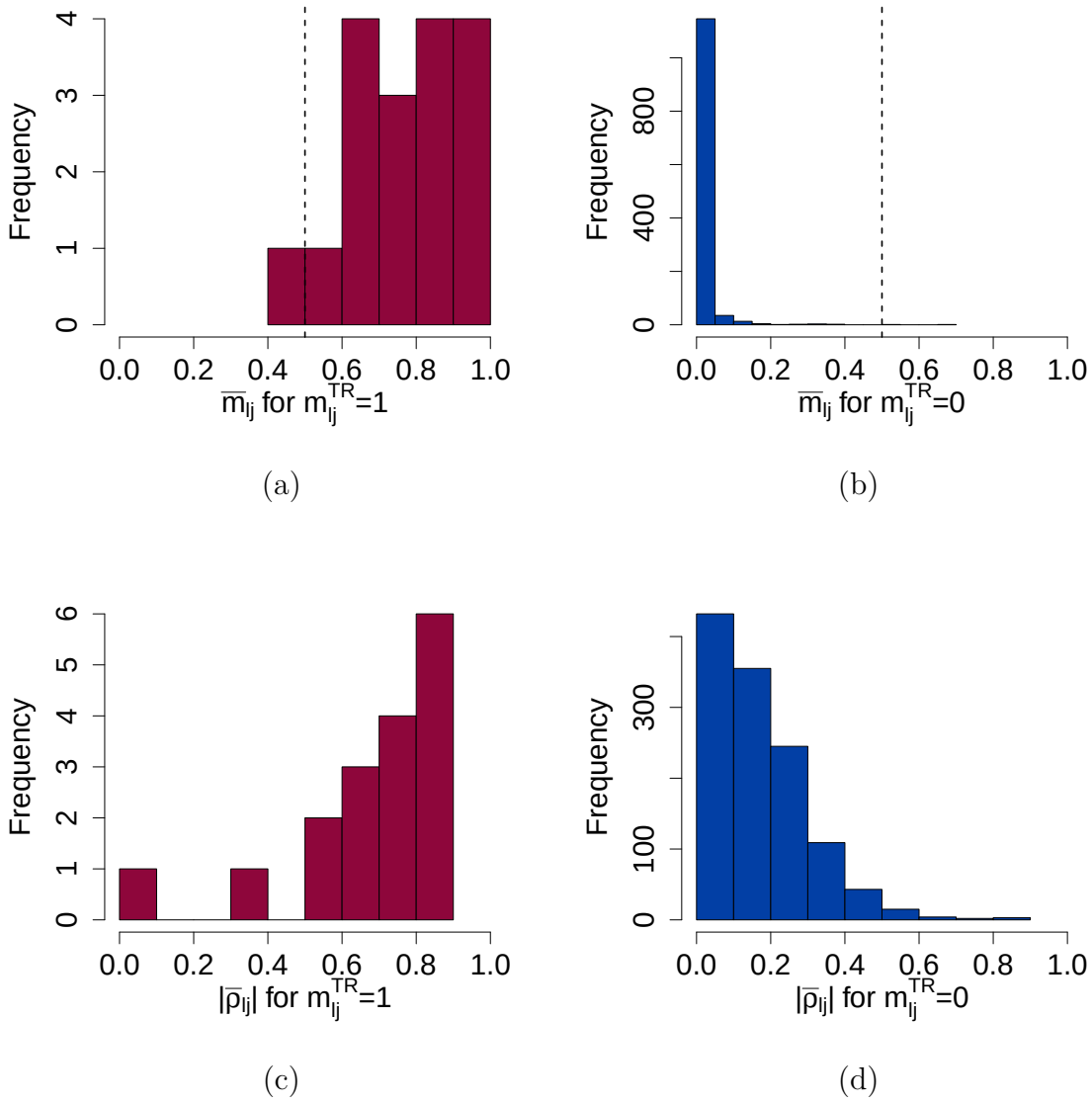
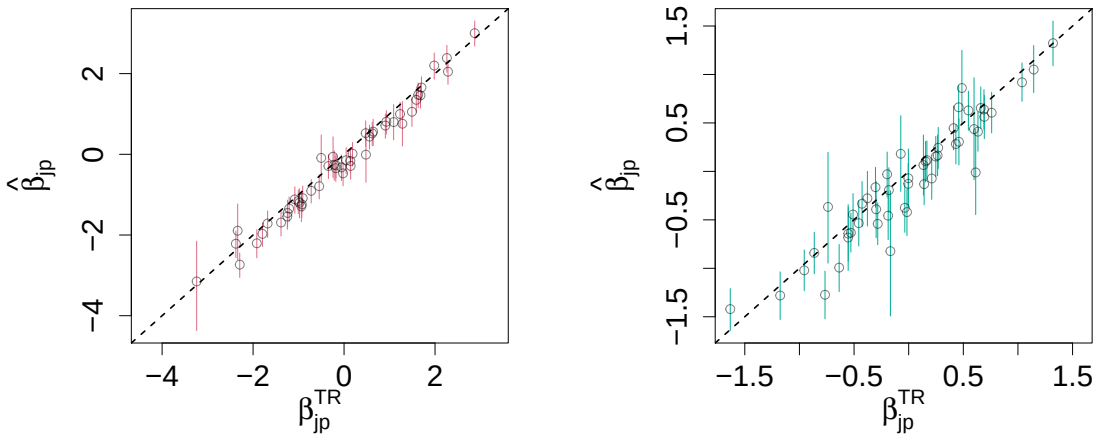


Figure 4.12: [Simulation 2] Posterior probabilities of edge inclusion from BRM-G ((a) and (b)) and pairwise correlations from BRM-Cov ((c) an (d)) for $l < j$ conditional on the true moral graph having ((a) and (c)) or not having ((b and (d)) an edge .

more complicated graph of Simulation 2 BRM-G is able to recover reasonable estimates for the regression coefficients.

Sensitivity analysis and details of the chain's convergence for Simulation 2 are



(a) $p = 1$ (b) $p = 2$

Figure 4.13: [Simulation 2] Posterior means $\hat{\beta}_{jp}$ and 95% credible intervals plotted against the simulation truth β_{jp}^{TR}.

given in §C.2.2.

As in Simulation 1, for Simulation 2 we produced 100 replicated datasets and compare the performance of BRM-G to edgeR and BRM-Cov. We considered the models' performances through averages of the point estimates for the moral graph and for the correlation matrix using BRM-G and BRM-Cov, respectively. The results for BRM-G and BRM-Cov are show in Figure 4.14. The graph structure is more well defined, on average, using the graphical analysis produced by BRM-G than using inference on the correlation matrix produced by BRM-Cov. Figure 4.15 shows the standard deviations of the absolute values of the posterior correlations for BRM-Cov and of the posterior edge probabilities for BRM-G. The results from Simulation 2 are consistent with those of Simulation 1 in that the estimates for OTU relationships using correlations from BRM-Cov are noisier than the estimates produced by BRM-G.

The RMSEs for β_{j1}, β_{j2}, and μ_{ij} under BRM-G, BRM-Cov, and edgeR are

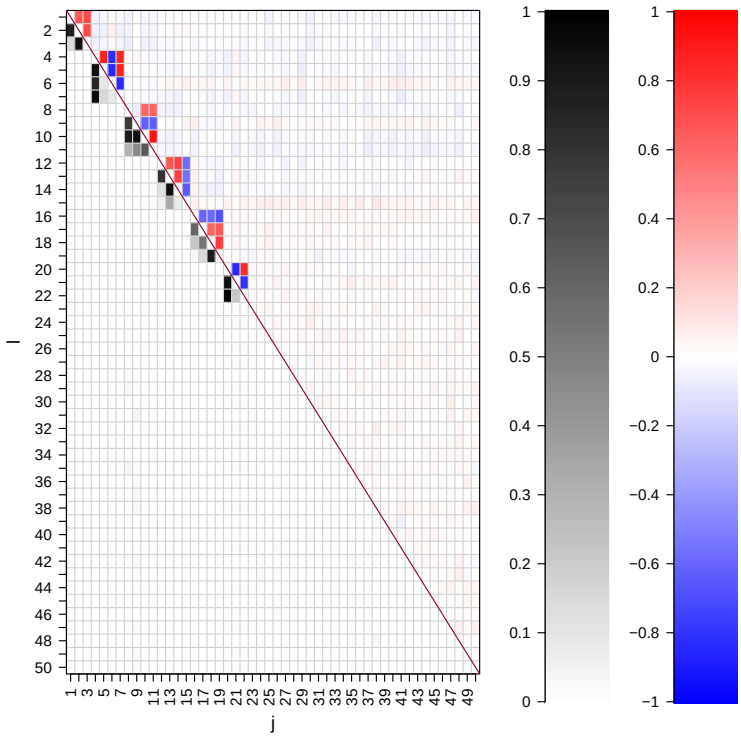


Figure 4.14: [Simulation 2] Results based on 100 simulated datasets. Proportions of edge inclusions over $\hat{G}_k^m$, $k = 1, \ldots, 100$ computed under BRM-G are in the lower triangle. Averages of $\hat{\rho}_{\ell,j,k}$, $k = 1, \ldots, 100$ computed under BRM-Cov are in the upper triangle.

shown in Table 4.2. Both BRM-G and BRM-Cov outperform edgeR when estimating the regression coefficients and the mean OTU abundances. For the Bayesian models we include DIC and LPML for model comparison. The performances of BRM-G and BRM-Cov are very similar, both in terms of RMSE and in terms of the model comparison metrics, confirming the findings of Simulation 1 where BRM-G and BRM-Cov also had similar performance.

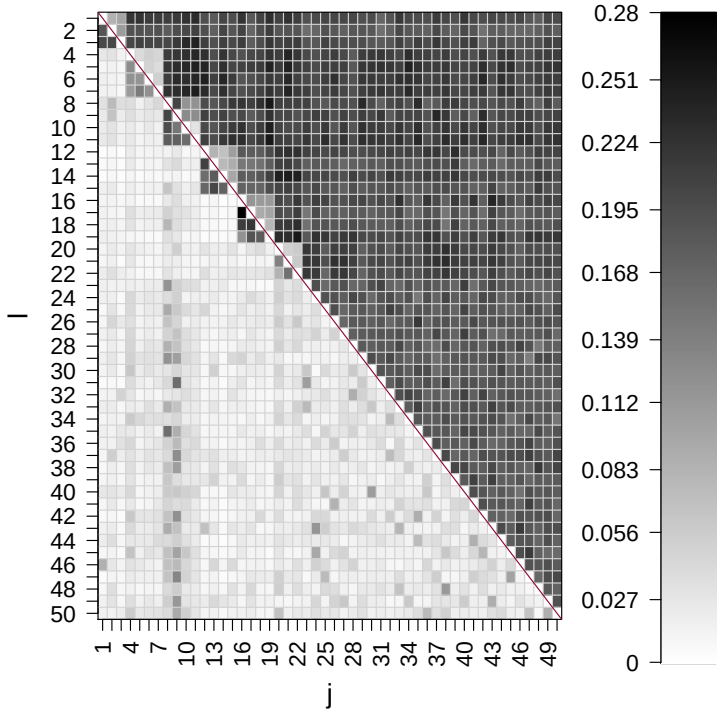


Figure 4.15: [Simulation 2] Variability of the OTU dependence structure over 100 simulated datasets. Standard deviation of posterior probabilities of edge inclusion under BRM-G and of the correlation estimates under BRM-Cov are shown in the lower and upper triangles, respectively.

4.4 Chronic Wound Microbiome Data Analysis

In this section we discuss the application of BRM-G to the microbiome dataset of Verbanic *et al.* (2019) and Shuler *et al.* (2019b). Different from the previous works, we applied BRM-G to a genus-collapsed OTU table produced by the `Phyloseq` R package (McMurdie and Holmes, 2013) to obtain reliable inferences on individual OTUs. Such agglomeration yields larger counts for individual OTUs and reduces the prevalence of small counts. The dataset consists of microbiome samples taken from $M = 20$ patients with chronic wounds. Swab samples were taken from the wounds pre- and post-debridement, as well as from sites with

Model	β_{j1}	β_{j2}	μ_{ij}	DIC	LPML
BRM-G	**0.263 (0.074)**	**0.257 (0.054)**	574,475 (2,079,977)	**-1,419 (171)**	**-11,412 (871)**
BRM-Cov	0.265 (0.080)	0.261 (0.063)	**574,347 (2,098,460)**	-1,410 (170)	-11,436 (867)
edgeR	0.613 (0.279)	0.607 (0.250)	1,943,481 (5,021,776)	–	–

Table 4.2: [Simulation 2] Performance metrics on 100 simulated datasets. Standard deviations in parenthesis. RMSE shown for β_{j1}, β_{j2}, and μ_{ij}. DIC and LPML shown for the Bayesian models.

healthy skin as a control, for a total of $n = M \times 3 = 60$ samples. We removed OTUs whose counts are zero in all experimental conditions for more than one subjects, for reliable estimates of θ_{mj}. After pre-processing a total of $J = 46$ OTUs were included for analysis. Empirical partial correlations of $\log(Y_{ij} + \epsilon)$ with $\epsilon = 0.1$ are shown in the upper triangle in Figure 4.16. We found that the empirical partial correlation estimates are sensitive to the choice of ϵ due to overdispersion. We set the hyperparameters and fit BRM-G to the chronic wound microbiome dataset using the same procedures described in §4.3. We checked the chain for convergence by inspecting traceplots and comparing the model's results to another chain using different initial conditions and a different random seed. We did not find evidence suggesting the chain failed to converge. More details about the chain's convergence and diagnostics are described in Supplementary §C.3.1.

As in the simulation studies, we computed posterior probabilities $\bar{m}_{\ell j}$ of including individual edges in the moral graph and produced a point estimate $\hat{G}^m$ for the moral graph by including an edge if $\bar{m}_{\ell j} > 0.5$. The estimates of posterior edge inclusion probabilities $\bar{m}_{\ell j}$ and its resulting moral graph estimate $\hat{G}^m$ are shown in Figures 4.16 and 4.17(a), respectively. The genus names corresponding to the OTUs in the subgraphs inferred by $\hat{G}^m$ are shown in Figure 4.17(b). Compared to the empirical partial correlations, $\bar{m}_{\ell j}$'s have a notably different pattern, underscoring the importance of appropriate modeling of the interaction structure of OTUs. The graph contains seven subgraphs, each with 2 or 3 related

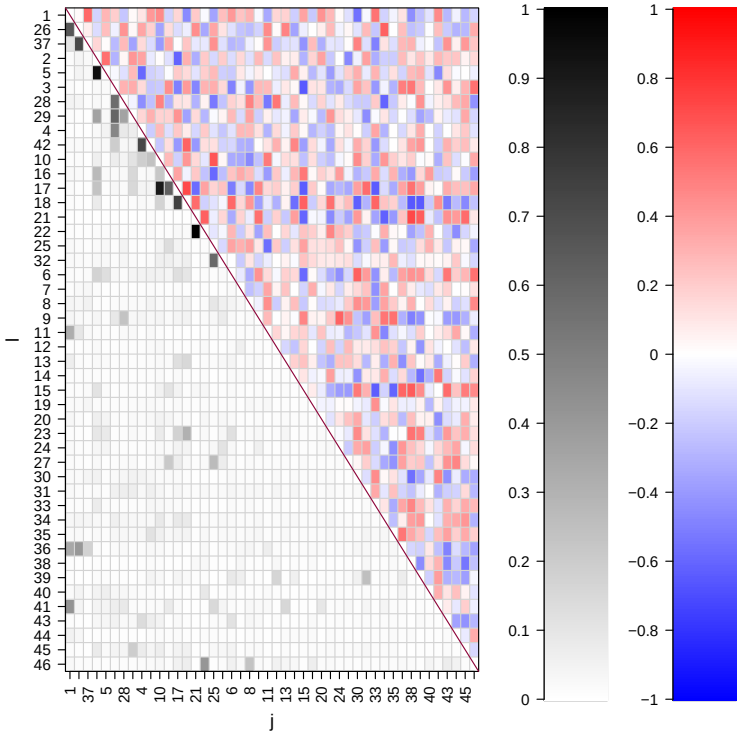
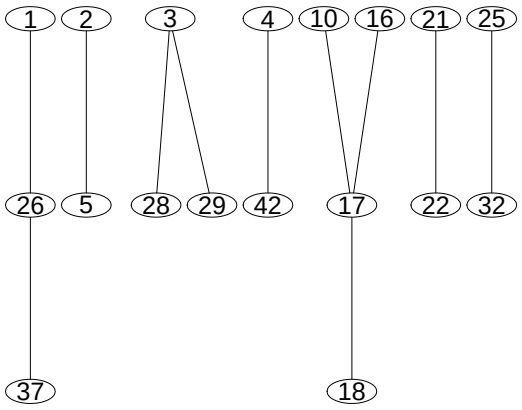


Figure 4.16: [Chronic Wound Data] Posterior probabilities of edge inclusion, $\bar{m}_{\ell j}$ under BRM-G and empirical partial correlations of $\log(Y_{ij} + 0.1)$ are shown in the lower triangle and upper triangle, respectively.

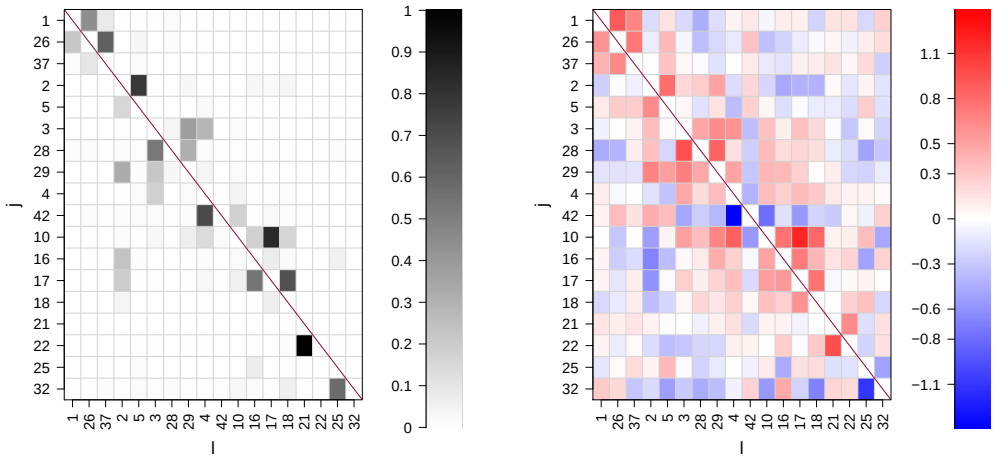


(a) $\hat{G}^m$

OTU #	Genus
1	*Micrococcus*
26	*Enhydrobacter*
37	*Brevundimonas*
2	*Pseudomonas*
5	*Bacteroides*
3	*Anaerococcus*
28	*NA*
29	*Peptoniphilus*
4	*Helcococcus*
42	*Propionibacterium*
10	*Actinomyces*
16	*NA*
17	*Campylobacter*
18	*Porphyromonas*
21	*Enterobacter*
22	*NA*
25	*Haemophilus*
32	*Ralstonia*

(b) Genus

Figure 4.17: [Chronic Wound Data] (a) Moral graph point estimate, $\hat{G}^m$ and (b) Genus names of the OTUs connected through the edges in $\hat{G}^m$.



(a) $\bar{a}_{\ell j}$ (b) $\hat{\gamma}_{\ell j}$

Figure 4.18: [Chronic Wound Data] Posterior estimates $\bar{a}_{\ell j}$ of the probabilities of including the directed edges $(\ell \to j)$ for the pairs with $\bar{m}_{\ell j} > 0.5$ are in panel (a). Panel (b) has the posterior mean estimates of $\gamma_{\ell j}$ given that $(\ell \to j)$ is included.

OTUs. $\bar{a}_{\ell j}$ and $\hat{\gamma}_{\ell j}$ are also illustrated in Figure 4.18(a) and (b), respectively, for the OTUs in any of the subgraphs in $\hat{G}^m$ to infer the signs associated with the interactions along with the probability estimates of a directed edge inclusion. The inferred graph exhibits experimentally and/or biologically relevant features. Previous work has identified the co-occurrence of *Micrococcus*, *Enhydrobacter*, and *Brevundimonas* (OTUs 1, 26, 37) in polymicrobial biofilms isolated from a variety of sources, including Timke *et al.* (2004, 2005); Vornhagen *et al.* (2013); Callewaert *et al.* (2015), though others argue they may be common contaminants in molecular biology reagents (Salter *et al.*, 2014). *Pseudomonas* and *Bacteroides* (OTUs 2 and 5) are both gram-negative, rod-shaped bacteria that form biofilms (Jang and Eom, 2019; Mulcahy *et al.*, 2014) and are implicated in chronic wound infections and necrotizing fasciitis (Sarani *et al.*, 2009). Similarly, *Anaerococcus* and *Peptoniphilus* (OTUs 3 and 29) are both clinically-relevant gram-positive anaerobic cocci (Murphy and Frick, 2013), which frequently co-colonize wounds and may

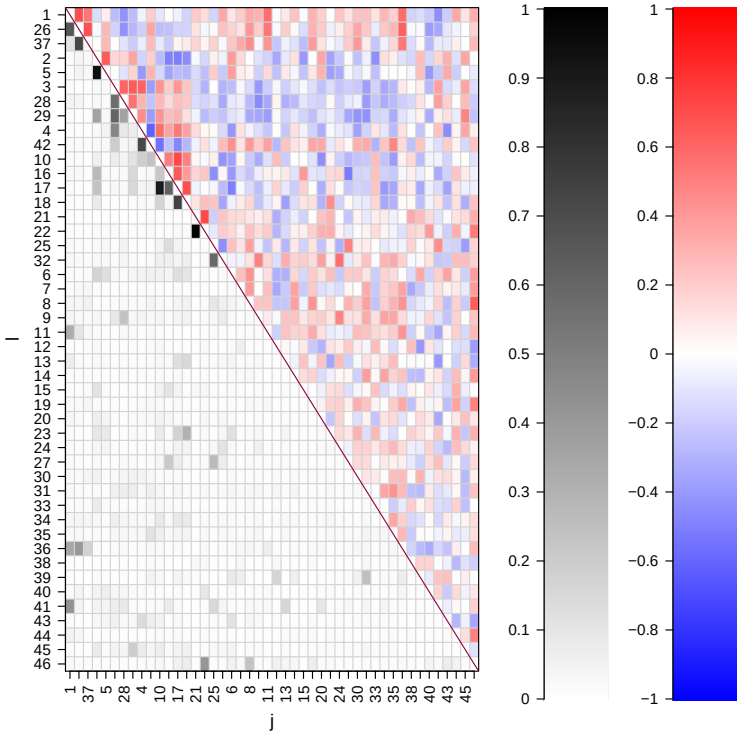
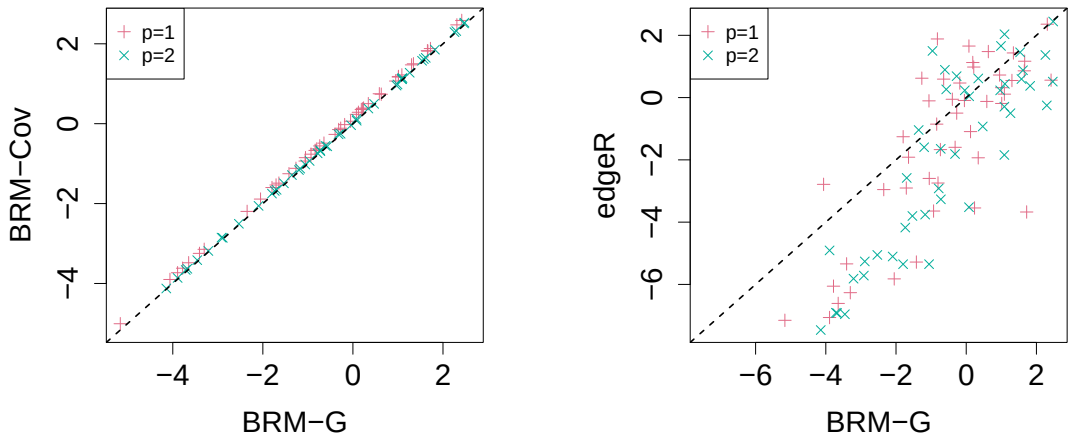


Figure 4.19: [Chronic Wound Data] Elementwise posterior means of pairwise correlations $\bar{\rho}_{\ell j}$ under BRM-Cov are shown in the upper triangle. For an easy comparison, posterior probabilities of edge inclusion, $\bar{m}_{\ell j}$ under BRM-G are shown in the lower triangle.

be associated with impaired healing of diabetic foot ulcers. Previous work has established several interactions between *Actinomyces*, *Campylobacter*, and *Porphyromonas* (OTUs 10, 17, 18), especially in the oral microbiome. *Actinomyces* is known to form biofilms with Porphyromonas on tooth enamel (Periasamy and Kolenbrander, 2009), in both healthy and diseased states, and *Porphyromonas* specifically suppresses host immune response to *Campylobacter* (Bostanci *et al.*, 2007), which may allow the pair to evade or overcome an inflammatory response to colonization or infection. Two pairs of OTUs had negative interactions. *Helcococcus* and *Propionibacterium* (OTUs 4 and 42), have been implicated in diabetic osteomyelitis (bone infection) (Van Asten *et al.*, 2016), though specific interactions between these bacteria have not been reported. Similarly, *Haemophilus* and *Ralstonia* (OTUs 25 and 32) have been identified in the lung microbiome of cystic fibrosis patients, but their interactions remain unresolved (Green *et al.*, 2017). BRM-G provides useful insights on the dependence structure between OTUs. To further investigate causal relationships between the inferred OTUs, experimental validations are needed.

We conducted sensitivity analysis with respect to the specification of κ, P_g, υ_α and υ_r. Across a range of values for these parameters we found that inference on G^m was robust. The signs associated with the OTU interactions and the probabilities of directed edge inclusion were also insensitive to the choice of these parameters. Details of the sensitivity analysis and figures showing the results of fitting BRM-G using these alternative specifications are described in Appendix §C.3.1.

For comparison, we fit BRM-Cov. Figure 4.19 shows elementwise posterior means $\bar{\rho}_{\ell j}$ of pairwise correlations produced by BRM-Cov. The estimate $\bar{\rho}_{\ell j}$ produced by BRM-Cov reasonably agrees with the moral graph estimate produced by BRM-G, though the nature of the relationships among the OTUs is less clear



(a) BRM-G for $p = 1$

(b) β_{jp}

Figure 4.20: [Chronic Wound Data] Regression coefficient estimates β_{jp} for (a) BRM-G versus BRM-Cov and (b) BRM-G versus edgeR.

Model	DIC	LPML
BRM-G	**-1,361**	**-16,617**
BRM-Cov	-1,359	-16,618

Table 4.3: [Chronic Wound Data] Model fit metrics for the chronic wound microbiome dataset.

under BRM-Cov. In Figure 4.20 we compare the regression coefficient estimates produced by BRM-G, BRM-Cov, and edgeR. For the Bayesian models we use the posterior mean of β_{jp} as a point estimate. The estimates produced by BRM-G and BRM-Cov are very similar, confirming that incorporating the graphical structure into the model does not interfere with the estimation of the covariate effects. The estimates of β_{jp} produced by edgeR trend similarly to the estimates produced by BRM-G, with a correlation of 0.82, although for small values of β_{jp} BRM-G tends to produce estimates less than those produced by edgeR. BRM-G and BRM-Cov produced very similar model comparison metrics when fit to the chronic wound microbiome dataset. LPML and DIC values for the two models are listed in Table 4.3. The similar metrics mirrors the results of the simulation studies, where BRM-G and BRM-Cov also had similar LPML and DIC scores across replicated datasets. The similar scores suggest that adding the graphical component to BRM-G does not interfere with its ability to explain the chronic wound microbiome data.

4.5 Discussion

We have presented a Bayesian graphical model to infer graphical structure from count data with applications to microbiome analysis. Our simulation studies indicate BRM-G identifies groups of related OTUs and accurately estimates covariate effects. In these simulations BRM-G outperforms the popular alternative edgeR, which does not incorporate methods for inferring graphical structure. Analysis of the simulation study results indicate the additional complexity of BRM-G over BRM-Cov is warranted, as BRM-G detects network structure better than a the simpler approach of using BRM-Cov's inference on the correlation structure to identify related groups of OTUs. Application of BRM-G to

the chronic wound microbiome dataset demonstrates the model's utility for microbiome studies. BRM-G identified several groups of OTUs whose relationships were not clearly identifiable from inspection of the empirical correlation estimates of the OTU counts. BRM-G may be extended by incorporating more general graphs, for example graphs that account for spatiotemporal structure in the data or allowing the graph to vary by experimental condition. These are potential areas of future research.

Chapter 5

Conclusion

This work introduced Bayesian models for multivariate count data. New strategies for handling challenging aspects of count data analysis, such as normalization, zero inflation, overdispersion, and dependent samples were introduced. The models were created in the context of microbiome analysis, with a focus on answering questions posed by biologists conducting metagenomics studies. Bayesian regression models were developed to compare microbiome communities, relate taxa abundance to environmental factors and experimental conditions, and identify related taxa.

Chapter 2 described the development of a Bayesian regression model using novel non-local priors to identify important covariates related to taxa abundance. These priors provide superior variable selection performance over existing alternatives. The model produces convenient summaries of the effect directions of environmental factors on OTU abundances, allowing researchers to easily evaluate how covariates are related to the microbial community. Unlike other popular models, the OTU abundances were modeled jointly using a Bayesian hierarchical model, allowing information to be pooled across OTUs to improve inference. The samples' complicated temporal dependence structure was accounted for through a

process convolutions component that describes how OTU abundances evolve over time. The model's utility was confirmed by its application to an ocean microbiome dataset which found that domoic acid affects microbiome composition, and these findings were validated through further lab experiments.

The model presented in chapter 3 provided a way to compare microbiomes at the community level. The model used a Bayesian nonparametric approach to get estimates for the distributions of OTU abundance levels and of the probabilities of OTU presence across experimental conditions. These distributions provide a clearer, more nuanced way to compare different microbiomes than simple significance tests or distance metrics, and the use of a dependent Dirichlet Process (DDP) approach to infer these distributions avoids restrictive and potentially unrealistic parametric assumptions. Importantly, the model carefully handles excess zero inflation in the OTU counts which improves inference on OTU abundance levels and community structure. The model was applied to a chronic wound microbiome dataset and the results were consistent with previous work finding greater species richness and abundance in healthy skin versus the wound conditions.

Chapter 4 introduced a Bayesian graphical model to identify OTU interactions. The model used a directed acyclic graph (DAG) component to identify such relationships. The performance of the DAG approach is superior to existing methods based on marginal correlations among taxa, and provides clearer insight into the OTUs' network structure than methods using empirical correlations. A regression component gives insight into how OTU abundances vary across experimental conditions, and provides more accurate estimates and uncertainty quantification for these effects than alternative models. The model was applied to a chronic wound microbiome dataset aggregated at the genus level, and the relationships it detected were confirmed as biologically relevant by previous literature.

Although the models presented here help address some of the most challenging aspects of modeling multivariate count data there is still more work to be done. In some cases the relationships of OTU abundance may not be log-linear with the covariates, and a more flexible regression structure structure may be desirable. The graphical model of chapter 4 may be extended by explicitly modeling zero inflation to better handle sparse OTU tables. Furthermore, the graphical assumptions may be relaxed by allowing spatiotemporal variation in the graph, or allowing it to vary with experimental conditions. A Bayesian hierarchical model on the graph may allow for such flexibility while allowing strength to be borrowed across the graphs. To some degree, it is already possible to explicitly incorporate information about metabolic pathways across OTUs into the graph through its prior, but future additions may make incorporating such information into the model easier and more flexible. Bi-directional relationships among OTUs may be added as well, allowing for explicit mutualistic relationships across OTUs such as through reciprocal graphical models in the spirit of Koster and Others (1996) and Ni *et al.* (2018). These areas may be addressed in future research.

Bibliography

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. J. (2011). Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, 30–38.

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

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.

Altomare, D., Consonni, G., and La Rocca, L. (2013). Objective Bayesian Search of Gaussian Directed Acyclic Graphical Models for Ordered Variables with Non-Local Priors. *Biometrics* **69**, 2, 478–487.

Anders, S. and Huber, W. (2010). Differential Expression analysis for sequence count data. *Nature Preceedings* **11**, 10, R106.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26**, 1, 32–46.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., and Others (2011). Enterotypes of the human gut microbiome. *nature* **473**, 7346, 174–180.

Banerjee, S., Schlaeppi, K., and van der Heijden, M. G. A. (2018). Keystone taxa as drivers of microbiome structure and functioning. *Nature Reviews Microbiology* **16**, 9, 567–576.

Barker, D. J., Hill, S. M., and Mukherjee, S. (2010). MC 4: a tempering algorithm for large-sample network inference. In *IAPR International Conference on Pattern Recognition in Bioinformatics*, 431–442. Springer.

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.

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.

Berry, D. and Widder, S. (2014). Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in Microbiology* **5**, 219.

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.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research* **3**, Jan, 993–1022.

Bostanci, N., Allaker, R. P., Belibasakis, G. N., Rangarajan, M., Curtis, M. A., Hughes, F. J., and McKay, I. J. (2007). Porphyromonas gingivalis antagonises Campylobacter rectus induced cytokine production by human monocytes. *Cytokine* **39**, 2, 147–156.

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

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 1, 94.

Buza, T. M., Tonui, T., Stomeo, F., Tiambo, C., Katani, R., Schilling, M., Lyimo, B., Gwakisa, P., Cattadori, I. M., Buza, J., and Kapur, V. (2019). iMAP: an integrated bioinformatics and visualization pipeline for microbiome data analysis. *BMC Bioinformatics* **20**, 1, 374.

Callewaert, C., Van Nevel, S., Kerckhof, F.-M., Granitsiotis, M. S., and Boon, N. (2015). Bacterial exchange in household washing machines. *Frontiers in microbiology* **6**, 1381.

Carlin, B. P. and Chib, S. (1995). Bayesian Model Choice via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society* **57**, 3, 473–484.

Casella, G. (1985). An Introduction to Empirical Bayes Data Analysis. *The American Statistician* **39**, 2, 83–87.

Castelletti, F., Consonni, G., and Others (2019). Objective Bayes model selection of Gaussian interventional essential graphs for the identification of signaling pathways. *The Annals of Applied Statistics* **13**, 4, 2289–2311.

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

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.

Chickering, D. M. (2002). Learning equivalence classes of Bayesian-network structures. *Journal of machine learning research* **2**, Feb, 445–498.

Chipman, H., George, E. I., McCulloch, R. E., Clyde, M., Foster, D. P., and Stine, R. A. (2001). The practical implementation of Bayesian model selection. *Lecture Notes-Monograph Series* 65–134.

Cirri, E. and Pohnert, G. (2019). Algae bacteria interactions that balance the planktonic microbiome. *New Phytologist* **223**, 1, 100–106.

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.

Consortium, T. H. M. P. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207.

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

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

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

Deng, Y., Jiang, Y.-H., Yang, Y., He, Z., Luo, F., and Zhou, J. (2012). Molecular ecological network analyses. *BMC Bioinformatics* **13**, 1, 113.

Devroye, L. and Lugosi, G. (2001). *Total Variation*, 38–46. Springer New York, New York, NY.

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

Faust, K. and Raes, J. (2012). Microbial interactions: from networks to models. *Nature Reviews Microbiology* **10**, 8, 538–550.

Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., and Huttenhower, C. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS computational biology* **8**, 7, e1002606–e1002606.

Frank, D. N., Amand, A. L. S., Feldman, R. A., Boedeker, E. C., Harpaz, N., and Pace, N. R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences* **104**, 34, 13780–13785.

Friedman, N. and Koller, D. (2003). Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine learning* **50**, 1-2, 95–125.

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

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

Geisser, S. (1993). *Predictive Inference*, vol. 55. CRC Press.

Geisser, S. and Eddy, W. F. (1979). A Predictive Approach to Model Selection. *Journal of the American Statistical Association* **74**, 365, 153–160.

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.

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

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

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., and Knight, R. (2018). Current understanding of the human microbiome. *Nature Medicine* **24**, 4, 392–400.

Gilbert, J. A., Quinn, R. A., Debelius, J., Xu, Z. Z., Morton, J., Garg, N., Jansson, J. K., Dorrestein, P. C., and Knight, R. (2016). Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature* **535**, 7610, 94–103.

Gilbert, J. A., Steele, J. A., Caporaso, J. G., Steinbrück, L., Reeder, J., Temperton, B., Huse, S., McHardy, A. C., Knight, R., Joint, I., Somerfield, P., Fuhrman, J. A., and Field, D. (2012). Defining seasonal marine microbial community dynamics. *The ISME Journal* **6**, 2, 298–308.

Giudici, P. and Castelo, R. (2003). Improving Markov chain Monte Carlo model search for data mining. *Machine learning* **50**, 1-2, 127–158.

Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* **102**, 477, 359–378.

Goudie, R. J. B. and Mukherjee, S. (2016). A Gibbs sampler for learning DAGs. *The Journal of Machine Learning Research* **17**, 1, 1032–1070.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**, 3-4, 325–338.

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* .

Green, H. D., Bright-Thomas, R., Kenna, D. T., Turton, J. F., Woodford, N., and Jones, A. M. (2017). Ralstonia infection in cystic fibrosis. *Epidemiology & Infection* **145**, 13, 2864–2872.

Grice, E. A., Kong, H. H., Conlan, S., Deming, C. B., Davis, J., Young, A. C., Bouffard, G. G., Blakesley, R. W., Murray, P. R., Green, E. D., *et al.* (2009). Topographical and temporal diversity of the human skin microbiome. *science* **324**, 5931, 1190–1192.

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

Grzegorczyk, M. and Husmeier, D. (2008). Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning* **71**, 2-3, 265.

Higdon, D. (2002). Space and Space-Time Modeling using Process Convolutions. In *Quantitative Methods for Current Environmental Issues*, 37–56. Springer.

Holmes, E., Li, J. V., Marchesi, J. R., and Nicholson, J. K. (2012). Gut microbiota composition and activity in relation to host metabolic phenotype and disease risk. *Cell metabolism* **16**, 5, 559–564.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., FitzGerald, M. G., Fulton, R. S., and Others (2012). Structure, function and diversity of the healthy human microbiome. *nature* **486**, 7402, 207.

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

Jang, H.-I. and Eom, Y.-B. (2019). Antibiofilm and antibacterial activities of repurposing auranofin against Bacteroides fragilis. *Archives of microbiology* 1–10.

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

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, 137–142. Springer.

Johnson, H. E., Scott Mills, L., Wehausen, J. D., and Stephenson, T. R. (2010). Combining ground count, telemetry, and mark–resight data to infer population dynamics in an endangered species. *Journal of Applied Ecology* **47**, 5, 1083–1093.

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.

Jonsson, V., Österlund, T., Nerman, O., and Kristiansson, E. (2018). Modelling of zero-inflation improves inference of metagenomic gene count data. *Statistical Methods in Medical Research* 0962280218811354.

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

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

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 Review of Genomics and Human Genetics* **18**, 1, 65–86.

Koster, J. T. A. and Others (1996). Markov properties of nonrecursive causal models. *The Annals of Statistics* **24**, 5, 2148–2177.

Kostic, A. D., Gevers, D., Pedamallu, C. S., Michaud, M., Duke, F., Earl, A. M., Ojesina, A. I., Jung, J., Bass, A. J., Tabernero, J., and Others (2012). Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome research* **22**, 2, 292–298.

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS computational biology* **11**, 5, e1004226–e1004226.

Layeghifard, M., Hwang, D. M., and Guttman, D. S. (2017). Disentangling Interactions in the Microbiome: A Network Perspective. *Trends in Microbiology* **25**, 3, 217–228.

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.

Lee, J. and Sison-Mangus, M. (2018). A Bayesian Semiparametric Regression Model for Joint Analysis of Microbiome Data. *Frontiers in Microbiology* **9**, 522.

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

Levy, R. and Borenstein, E. (2013). Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *Proceedings of the National Academy of Sciences* **110**, 31, 12804–12809.

Lewis, J. D., Chen, E. Z., Baldassano, R. N., Otley, A. R., Griffiths, A. M., Lee, D., Bittinger, K., Bailey, A., Friedman, E. S., Hoffmann, C., and Others (2015). Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric Crohn's disease. *Cell host & microbe* **18**, 4, 489–500.

Ley, R. E., Turnbaugh, P. J., Klein, S., and Gordon, J. I. (2006). Human gut microbes associated with obesity. *nature* **444**, 7122, 1022–1023.

Li, H. (2015). Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis. *Annual Review of Statistics and Its Application* **2**, 1, 73–94.

Li, J., Witten, D. M., Johnstone, I. M., and Tibshirani, R. (2012). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics* **13**, 3, 523–538.

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.

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

Louca, S., Parfrey, L. W., and Doebeli, M. (2016). Decoupling function and taxonomy in the global ocean microbiome. *Science* **353**, 6305, 1272–1277.

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

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and Shafee, T. (2017). Transcriptomics technologies. *PLoS computational biology* **13**, 5.

MacEachern, S. N. (1999). Dependent Nonparametric Processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*. American Statistical Association.

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

Madigan, D., York, J., and Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique* 215–232.

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., and Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial ecology in health and disease* **26**, 1, 27663.

Mann, H. B. and Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Statist.* **18**, 1, 50–60.

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

McArdle, B. H. and Anderson, M. J. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* **82**, 1, 290–297.

McCallum, A., Nigam, K., and Others (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, vol. 752, 41–48. Citeseer.

McMurdie, P. J. and Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE* **8**, 4, e61217.

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

Mendes, R., Kruijt, M., De Bruijn, I., Dekkers, E., van der Voort, M., Schneider, J. H. M., Piceno, Y. M., DeSantis, T. Z., Andersen, G. L., Bakker, P. A. H. M., and Others (2011). Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science* **332**, 6033, 1097–1100.

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

Mulcahy, L. R., Isabella, V. M., and Lewis, K. (2014). Pseudomonas aeruginosa biofilms in disease. *Microbial ecology* **68**, 1, 1–12.

Murphy, E. C. and Frick, I.-M. (2013). Gram-positive anaerobic cocci–commensals and opportunistic pathogens. *FEMS microbiology reviews* **37**, 4, 520–553.

Ni, Y., Ji, Y., and Müller, P. (2018). Reciprocal Graphical Models for Integrative Gene Regulatory Network Analysis. *Bayesian Anal.* **13**, 4, 1095–1110.

Ni, Y., Stingo, F. C., and Baladandayuthapani, V. (2015). Bayesian nonlinear model selection for gene regulatory networks. *Biometrics* **71**, 3, 585–595.

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

O'Hara, R. B. and Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution* **1**, 2, 118–122.

Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M. H. H., Oksanen, M. J., and Suggests, M. (2007). The vegan package. *Community ecology package* **10**, 631–637.

Ozsolak, F. and Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature reviews genetics* **12**, 2, 87–98.

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

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

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Periasamy, S. and Kolenbrander, P. E. (2009). Mutualistic biofilm communities develop with Porphyromonas gingivalis and initial, early, and late colonizers of enamel. *Journal of bacteriology* **191**, 22, 6804–6811.

Peterson, C., Stingo, F. C., and Vannucci, M. (2015). Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association* **110**, 509, 159–174.

Radhakrishnan, A., Solus, L., and Uhler, C. (2018). Counting Markov equivalence classes for DAG models on trees. *Discrete Applied Mathematics* **244**, 170–185.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**, 437, 179–191.

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.

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* .

Reuter, J. A., Spacek, D. V., and Snyder, M. P. (2015). High-throughput sequencing technologies. *Molecular cell* **58**, 4, 586–597.

Richards, S. A. (2008). Dealing with overdispersed count data in applied ecology. *Journal of Applied Ecology* **45**, 1, 218–227.

Rivera-Pinto, J., Egozcue, J. J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., and Calle, M. L. (2018). Balances: a New Perspective for Microbiome Analysis. *mSystems* **3**, 4, e00053–18.

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.

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

Robinson, M. D. and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**, 21, 2881–2887.

Rodriguez, A. and Dunson, D. B. (2011). Nonparametric bayesian models through probit stick-breaking processes. *Bayesian analysis (Online)* **6**, 1.

Rodríguez, A., Lenkoski, A., and Dobra, A. (2011). Sparse covariance estimation in heterogeneous samples. *Electronic journal of statistics* **5**, 981–1014.

Rossell, D. and Telesca, D. (2017). Nonlocal Priors for High-Dimensional Estimation. *Journal of the American Statistical Association* **112**, 517, 254–265.

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J., and Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC biology* **12**, 1, 87.

Sambo, F., Finotello, F., Lavezzo, E., Baruzzo, G., Masi, G., Peta, E., Falda, M., Toppo, S., Barzon, L., and Di Camillo, B. (2018). Optimizing PCR primers targeting the bacterial 16S ribosomal RNA gene. *BMC bioinformatics* **19**, 1, 343.

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

Sarani, B., Strong, M., Pascual, J., and Schwab, C. W. (2009). Necrotizing fasciitis: current concepts and review of the literature. *Journal of the American College of Surgeons* **208**, 2, 279–288.

Scher, J. U., Sczesnak, A., Longman, R. S., Segata, N., Ubeda, C., Bielski, C., Rostron, T., Cerundolo, V., Pamer, E. G., Abramson, S. B., and Others (2013). Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. *elife* **2**, e01202.

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.

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.

Shuler, K., Sison-Mangus, M., and Lee, J. (2019a). Bayesian Sparse Multivariate Regression with Asymmetric Nonlocal Priors for Microbiome Data Analysis {(In press)}. *Bayesian Analysis* .

Shuler, K. U. S. C., Barbara), Verbanic, S. U. S., Chen, I. U. S. B., and Lee, J. U. S. C. (2019b). A Bayesian Nonparametric Analysis for Zero Inflated Multivariate Count Data with Application to Microbiome Study. Tech. rep.

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.

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.

Tang, Z.-Z. and Chen, G. (2018). Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics* .

Telesca, D., Müller, P., Kornblau, S. M., Suchard, M. A., and Ji, Y. (2012). Modeling protein expression and protein signaling pathways. *Journal of the American Statistical Association* **107**, 500, 1372–1384.

Timke, M., Wang-Lieu, N. Q., Altendorf, K., and Lipski, A. (2005). Community structure and diversity of biofilms from a beer bottling plant as revealed using 16S rRNA gene clone libraries. *Appl. Environ. Microbiol.* **71**, 10, 6446–6452.

Timke, M., Wolking, D., Wang-Lieu, N. Q., Altendorf, K., and Lipski, A. (2004). Microbial composition of biofilms in a brewery investigated by fatty acid analysis, fluorescence in situ hybridisation and isolation techniques. *Applied microbiology and biotechnology* **66**, 1, 100–107.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature* **449**, 7164, 804–810.

Van Asten, S. A. V., La Fontaine, J., Peters, E. J. G., Bhavan, K., Kim, P. J., and Lavery, L. A. (2016). The microbiome of diabetic foot osteomyelitis. *European Journal of Clinical Microbiology & Infectious Diseases* **35**, 2, 293–298.

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

Vornhagen, J., Stevens, M., McCormick, D. W., Dowd, S. E., Eisenberg, J. N. S., Boles, B. R., and Rickard, A. H. (2013). Coaggregation occurs amongst bacteria within and between biofilms in domestic showerheads. *Biofouling* **29**, 1, 53–68.

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.

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L. C., Xu, Z. Z., Ursell, L., Alm, E. J., Birmingham, A., Cram, J. A., Fuhrman, J. A., Raes, J., Sun, F., Zhou, J., and Knight, R. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME Journal* **10**, 7, 1669–1681.

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., and Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 1, 27.

Witten, D. M. (2011). Classification and clustering of sequencing data using a poisson model. *Annals of Applied Statistics* **5**, 4, 2493–2518.

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

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.

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.

Xia, Y., Sun, J., and Chen, D.-G. (2018). *Statistical analysis of microbiome data with R*. Springer.

Xiao, S. (2015). *Bayesian nonparametric modeling for some classes of temporal point processes*. Ph.D. thesis, University of California Santa Cruz, Santa Cruz.

Xu, L., Paterson, A. D., Turpin, W., and Xu, W. (2015). Assessment and Selection of Competing Models for Zero-Inflated Microbiome Data. *PLOS ONE* **10**, 7, e0129606.

Xu, L., Paterson, A. D., and Xu, W. (2017). Bayesian latent variable models for hierarchical clustered count outcomes with repeated measures in microbiome studies. *Genetic Epidemiology* **41**, 3, 221–232.

Zelezniak, A., Andrejev, S., Ponomarova, O., Mende, D. R., Bork, P., and Patil, K. R. (2015). Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proceedings of the National Academy of Sciences* **112**, 20, 6449–6454.

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

Zhang, Y., Jin, R., and Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics* **1**, 1-4, 43–52.

Zhang, Y., Zhou, H., Zhou, J., and Sun, W. (2017b). Regression Models for Multivariate Count Data. *Journal of Computational and Graphical Statistics* **26**, 1, 1–13.

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H., and Wu, M. C. (2015). Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. *The American Journal of Human Genetics* **96**, 5, 797–807.

Zhou, J., Deng, Y., Luo, F., He, Z., Tu, Q., and Zhi, X. (2010). Functional Molecular Ecological Networks. *mBio* **1**, 4.

Appendix A

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

A.1 MCMC Algorithm

We obtain a sample from the posterior distribution using an MCMC comprised of a combination of Gibbs and Metropolis-within-Gibbs steps. Recall that we have mixture-of-mixtures distributions for the priors of r_{tk} and α_{0j} in (5) and (6) of the main text, respectively. For easy posterior simulation, we introduce latent variables that indicate which mixture component r_{tk} and α_{0j} are from, and do categorical/Bernoulli draws to update these latent variables. Specifically, for r_{tk},

we let $c^r_{tk} = \ell$, $\ell = 1, \ldots, L^r$ if and only if r_{tk} came from the ℓ^{th} mixture component. Conditional on c^r_{tk}, we introduce another indicator variable $\lambda^r_{tk} \in \{0, 1\}$ to indicate the Gaussian component from which r_{tk} came, $\mathrm{N}(\eta^r_\ell, u^2_r)$ or $\mathrm{N}\left(\frac{\upsilon_r - w^r_\ell \eta^r_\ell}{1-w^r_\ell}, u^2_r\right)$. Let $\mathrm{P}(c^r_{tk} = \ell \mid \boldsymbol{\psi}^r) = \psi^r_\ell$ and $\mathrm{P}(\lambda^r_{tk} = 1 \mid c^r_{tk} = \ell, w^r_\ell) = w^r_\ell$. The joint prior distribution of $\boldsymbol{r}$, $\boldsymbol{c}^r$ and $\boldsymbol{\lambda}^r$ can be written as

$$\begin{aligned}
\mathrm{P}(\boldsymbol{c}^r, \boldsymbol{\lambda}^r, \boldsymbol{r} \mid \boldsymbol{\psi}^r, \boldsymbol{w}^r, \boldsymbol{\eta}^r) &= \prod_{t,k} \mathrm{P}(c^r_{tk} \mid \boldsymbol{\psi}^r) \mathrm{P}(\lambda^r_{tk} \mid c^r_{tk}, \boldsymbol{w}^r) \mathrm{P}(r_{tk} \mid c^r_{tk}, \lambda^r_{tk}, \boldsymbol{\eta}^r) \\
&= \prod_{\ell=1}^{L^r} (\psi^r_\ell)^{d^r_\ell} \prod_{t,k} (w^r_{c_{tk}})^{\lambda^r_{tk}} (1 - w^r_{c_{tk}})^{1-\lambda^r_{tk}} \\
&\quad \times \prod_{t,k} \mathrm{N}\left(r_{tk} \mid \eta^r_{c_{tk}}, u^2_r\right)^{\lambda^r_{tk}} \mathrm{N}\left(r_{tk} \mid \frac{\upsilon_r - w^r_{c_{tk}} \eta^r_{c_{tk}}}{1 - w^r_{c_{tk}}}, u^2_r\right)^{1-\lambda^r_{tk}},
\end{aligned}$$

where $d^r_\ell = \sum_t \sum_{k=1}^{K_i} \mathbb{I}(c_{tk} = \ell)$ denotes the number of elements in the ℓ^{th} mixture component. We update mixture locations η^r_ℓ and mixture weights $\boldsymbol{\psi}^r$ and w^r_ℓ conditional on λ^r_{tk} and c^r_{tk} using Gibbs steps. Similar to the method used to sample r_{tk}, we also introduce auxiliary variables $c^\alpha_j \in \{1, \ldots, L^\alpha\}$ and $\lambda^\alpha_j \in \{0, 1\}$ to specify the mixture components from which α_{0j} came, and we let $d^\alpha_\ell = \sum_{j=1}^{J} \mathbb{I}(c^\alpha_j = \ell)$ denote the number of elements assigned to the ℓ^{th} mixture component. Again, we do categorical/Bernoulli draws for c^α_j and λ^α_j and then update α_{0j} conditional on these assignments, and update mixture locations η^α_ℓ and mixture weights $\boldsymbol{\psi}^\alpha$ and w^α_ℓ conditional on mixture labels λ^α_j and c^α_j using Gibbs steps. Letting $\underline{\boldsymbol{\theta}}$ denote the vector of all unknown parameters and $\boldsymbol{\Omega}$ denote the model's fixed hyperparameters, the joint posterior distribution of all unknown parameters up to proportionality is

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

where $\boldsymbol{X}_{\text{miss}}$ denotes the missing covariates, $\boldsymbol{X}_{\text{obs}}$ denotes the observed covariates, and $\boldsymbol{X} = \{\boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{miss}}\}$. At each MCMC iteration we impute missing values through categorical draws for covariates whose values are missing at random in some samples. We treat each category as equally likely a priori, and the probabilities for the categories are proportional to the likelihood induced by selecting each of those categories.

To improve mixing, at each MCMC iteration we use multiple steps to update β_{jp}, γ_{jp} and ι_p. We update β_{jp} conditional on all other parameters including γ_{jp} and ι_p. We also do a joint update of γ_{jp} and β_{jp} as follows; We use a method similar to that in Carlin and Chib (1995) and generate the joint proposal of $(\gamma_{jp}, \beta_{jp})$ using a linking density,

$$J(\hat{\beta}_{jp} \mid \gamma^0_{jp},\, \beta^0_{jp}) = \begin{cases} \text{TruncNorm}(0,\, \sigma_p^2,\, -\iota_p\sigma_p,\, \iota_p\sigma_p) & \text{if } \gamma^0_{jp} = 0 \\ \beta^0_{jp} & \text{if } \gamma^0_{jp} \neq 0 \end{cases} \tag{A.1}$$

where β^0_{jp} and γ^0_{jp} are the current values of β_{jp} and γ_{jp}. We then set $\hat{\beta}'_{jp} = \hat{\beta}_{jp} + e^{\beta}_{jp}$, where e^{β}_{jp} is a perturbation drawn from a normal distribution with mean 0 and fixed variance. The joint proposal $\beta^\star_{jp}$ and $\gamma^\star_{jp}$ is generated as a function of $\hat{\beta}'_{jp}$,

$$F(\beta^\star_{jp}, \gamma^\star_{jp} | \hat{\beta}'_{jp}) = \begin{cases} \beta^\star_{jp} = 0,\, \gamma^\star_{jp} = 0 & \text{if } \iota_p\sigma_p \geq \left|\hat{\beta}'_{jp}\right| \\ \beta^\star_{jp} = \hat{\beta}'_{jp},\, \gamma^\star_{jp} = 1 & \text{if } \iota_p\sigma_p < \hat{\beta}'_{jp} \\ \beta^\star_{jp} = \hat{\beta}'_{jp},\, \gamma^\star_{jp} = 2 & \text{if } -\iota_p\sigma_p > \hat{\beta}'_{jp} \end{cases} \tag{A.2}$$

The proposal is accepted with probability $\min(\text{m}_{MH}, 1)$, with the algorithm's

Metropolis-Hastings acceptance ratio m_{MH} is given by

$$\begin{aligned}\mathrm{m}_{MH} &= \frac{\mathrm{P}(\boldsymbol{Y}_j \mid \gamma_{jp}^{\star}, \beta_{jp}^{\star}, \cdots)\mathrm{P}(\gamma_{jp}^{\star})\mathrm{P}(\beta_{jp}^{\star} \mid \gamma_{jp}^{\star})}{\mathrm{P}(\boldsymbol{Y}_j \mid \gamma_{jp}^{0}, \beta_{jp}^{0}, \cdots)\mathrm{P}(\gamma_{jp}^{0})\mathrm{P}(\beta_{jp}^{0} \mid \gamma_{jp}^{0})} \qquad \text{(A.3)} \\ &\times \frac{J(\hat{\beta}_{jp}|\gamma_{jp}^{\star},\, \beta_{jp}^{\star})G(\hat{\beta}'_{jp}|\hat{\beta}_{jp})F(\beta_{jp}^{0},\, \gamma_{jp}^{0} \mid \hat{\beta}'_{jp})}{J(\hat{\beta}_{jp}|\gamma_{jp}^{0},\, \beta_{jp}^{0})G(\hat{\beta}'_{jp}|\hat{\beta}_{jp})F(\beta_{jp}^{\star},\, \gamma_{jp}^{\star} \mid \hat{\beta}'_{jp})} \\ &= \frac{\mathrm{P}(\boldsymbol{Y}_j \mid \gamma_{jp}^{\star}, \beta_{jp}^{\star}, \cdots)\mathrm{P}(\gamma_{jp}^{\star})\mathrm{P}(\beta_{jp}^{\star} \mid \gamma_{jp}^{\star})}{\mathrm{P}(\boldsymbol{Y}_j \mid \gamma_{jp}^{0}, \beta_{jp}^{0}, \cdots)\mathrm{P}(\gamma_{jp}^{0})\mathrm{P}(\beta_{jp}^{0} \mid \gamma_{jp}^{0})} \frac{J(\hat{\beta}_{jp}|\gamma_{jp}^{\star},\, \beta_{jp}^{\star})}{J(\hat{\beta}_{jp}|\gamma_{jp}^{0},\, \beta_{jp}^{0})} \qquad \text{(A.4)}\end{aligned}$$

where $G(\hat{\beta}'_{jp}|\hat{\beta}_{jp})$ denotes the Gaussian probability density induced by e_{jp}^{β} of going from $\hat{\beta}_{jp}$ to $\hat{\beta}'_{jp}$. The simplification in the second line can be seen by noting that $F(\beta_{jp},\, \gamma_{jp}|\hat{\beta}'_{jp})$ is degenerate. In addition to doing this update individually for each combination of j and p, we do a joint update of β_{jp} and γ_{jp} across all j as well using the straightforward extension of the same algorithm. Lastly, we do a joint update of β_{jp} and γ_{jp} using a method similar to the 'swap' step of the add/swap/delete algorithm in Li *et al.* (2017). In this approach, we first choose some covariate p_1 at random, and then randomly choose a different covariate p_2 with $\gamma_{j,p_2} \neq \gamma_{j,p_1}$. In our proposal, we swap γ for the two selected covariates so that $\gamma_{j,p_1}^{\star} = \gamma_{j,p_2}^{0}$ and $\gamma_{j,p_2}^{\star} = \gamma_{j,p_1}^{0}$. Then, conditional on the proposed $\gamma_{j,p_1}^{\star}$ and $\gamma_{j,p_2}^{\star}$, we generate $\beta_{j,p_1}^{\star}$ and $\beta_{j,p_2}^{\star}$ from their respective priors. We accept the proposal $(\gamma_{j,p_1}^{\star}, \beta_{j,p_1}^{\star})$ and $(\gamma_{j,p_2}^{\star}, \beta_{j,p_2}^{\star})$ with probability $\min(\mathrm{m}_{MH}, 1)$ where

$$\begin{aligned}\mathrm{m}_{MH} &= \frac{\mathrm{P}(\boldsymbol{Y}_j \mid \gamma_{j,p_1}^{\star}, \beta_{j,p_1}^{\star}, \gamma_{j,p_2}^{\star}, \beta_{j,p_2}^{\star}, \cdots)\mathrm{P}(\gamma_{j,p_1}^{\star})p(\gamma_{j,p_2}^{\star})\mathrm{P}(\beta_{j,p_1}^{\star} \mid \gamma_{j,p_1}^{\star})\mathrm{P}(\beta_{j,p_2}^{\star} \mid \gamma_{j,p_2}^{\star})}{\mathrm{P}(\boldsymbol{Y}_j \mid \gamma_{j,p_1}^{0}, \beta_{j,p_1}^{0}, \gamma_{j,p_2}^{0}, \beta_{j,p_2}^{0}, \cdots)\mathrm{P}(\gamma_{j,p_1}^{0})\mathrm{P}(\gamma_{j,p_2}^{0})\mathrm{P}(\beta_{j,p_1}^{0} \mid \gamma_{j,p_1}^{0})\mathrm{P}(\beta_{j,p_2}^{0} \mid \gamma_{j,p_2}^{0})} \\ &\times \frac{Q(\{(\gamma_{j,p_1}^{\star}, \beta_{j,p_1}^{\star}), (\gamma_{j,p_2}^{\star}, \beta_{j,p_2}^{\star})\} \to \{(\gamma_{j,p_1}^{0}, \beta_{j,p_1}^{0}), (\gamma_{j,p_2}^{0}, \beta_{j,p_2}^{0})\})}{Q(\{(\gamma_{j,p_1}^{0}, \beta_{j,p_1}^{0}), (\gamma_{j,p_2}^{0}, \beta_{j,p_2}^{0})\} \to \{(\gamma_{j,p_1}^{\star}, \beta_{j,p_1}^{\star}), (\gamma_{j,p_2}^{\star}, \beta_{j,p_2}^{\star})\})} \\ &= \frac{\mathrm{P}(\boldsymbol{Y}_j \mid \gamma_{j,p_1}^{\star}, \beta_{j,p_1}^{\star}, \gamma_{j,p_2}^{\star}, \beta_{j,p_2}^{\star}, \cdots)\mathrm{P}(\gamma_{j,p_1}^{\star})\mathrm{P}(\gamma_{j,p_2}^{\star})}{\mathrm{P}(\boldsymbol{Y}_j \mid \gamma_{j,p_1}^{0}, \beta_{j,p_1}^{0}, \gamma_{j,p_2}^{0}, \beta_{j,p_2}^{0}, \cdots)\mathrm{P}(\gamma_{j,p_1}^{0})\mathrm{P}(\gamma_{j,p_2}^{0})}\end{aligned}$$

where Q is the proposal distribution. We repeat this swap step 10 times within

each MCMC iteration.

The details of the remaining MCMC simulation steps are described below.

1. s_j, h, κ^2 (parameters related to overdispersion parameters): We perform individual metropolis udpates for the over-dispersion parameters $\tilde{s}_j = \log(s_j)$. The updates for h and κ^2 are conjugate conditional on $\tilde{s}_j$. These updates are the standard Gibbs steps of a normal-normal model with known variance and normal-inverse-gamma with known mean.

   - s_j

   $$\mathrm{P}(\tilde{s}_j \mid -) \propto \mathrm{N}\left(\tilde{s}_j \middle| h, \kappa^2\right) \times \prod_t \prod_{k=1}^{K_i} \frac{\Gamma(y_{tkj} + s_j^{-1})}{y_{tkj}!\Gamma(s_j^{-1})} \left(\frac{\mu_{tkj} s_j}{1 + \mu_{tkj} s_j}\right)^{y_{tkj}} \left(\frac{1}{1 + \mu_{tkj} s_j}\right)^{s_j^{-1}},$$

   where '$-$' represents all other parameters and data. Update via random walk Metropolis-Hastings.

   - h

   $$h \mid - \sim \mathrm{N}\left(\frac{\frac{1}{b_h^2} a_h + \frac{J}{\kappa^2}\bar{s}}{\frac{1}{b_h^2} + \frac{J}{\kappa^2}}, \left(\frac{1}{b_h^2} + \frac{J}{\kappa^2}\right)^{-1}\right)$$

   where $\bar{s} = \sum_{j=1}^{J} \tilde{s}_j$.

   - κ^2

   $$\kappa^2 \mid - \sim \mathrm{IG}\left(a_\kappa + \frac{J}{2}, b_\kappa + \frac{1}{2}\sum_{j=1}^{J}(\tilde{s}_j - h)^2\right)$$

2. $\boldsymbol{\psi}^r$, $\boldsymbol{w}^r$, $\boldsymbol{\eta}^r$ (parameters related to library size adjustment):

- $\boldsymbol{\psi}^r$

$$\mathrm{P}(\boldsymbol{\psi}^r \mid -) \propto \prod_{\ell=1}^{L^r} (\psi_\ell^r)^{a_\ell^{\psi^r} + d_\ell^r - 1}$$

Draw from Dirichlet distribution with concentration parameters $a_\ell^{\psi^r} + d_\ell^r$.

- w_ℓ^r

$$\begin{aligned}
\mathrm{P}(w_\ell^r \mid -) \propto & \prod_{t,k|c_{tk}^r=\ell} (w_\ell^r)^{\lambda_{tk}^r} (1-w_\ell^r)^{1-\lambda_{tk}^r} \\
& \times \prod_{t,k|c_{tk}^r=\ell \text{ and } \lambda_{tk}^r=0} \mathrm{N}\left(r_{tk} \middle| \frac{\upsilon_r - w_\ell^r \eta_\ell^r}{1-w_\ell^r},\, u_r^2\right) \\
& \quad \times (w_\ell^r)^{a_w^r - 1} (1-w_\ell^r)^{b_w^r - 1} \\
& \propto (w_\ell^r)^{a_w^r + \left[\sum_{tk|c_{tk}^r=\ell} \lambda_{tk}^r\right] - 1} (1-w_\ell^r)^{b_w^r + \left[\sum_{tk|c_{tk}^r=\ell} (1-\lambda_{tk}^r)\right] - 1} \\
& \quad \times \prod_{t,k|c_{tk}^r=\ell \text{ and } \lambda_{tk}^r=0} \mathrm{N}\left(r_{tk} \middle| \frac{\upsilon_r - w_\ell^r \eta_\ell^r}{1-w_\ell^r},\, u_r^2\right)
\end{aligned}$$

Update via random walk Metropolis-Hastings.

- η_ℓ^r

$$\eta_\ell^r \mid - \sim \mathrm{N}\left(\frac{\rho_\ell^{r_1} m_\ell^{r_1} + \rho_\ell^{r_2} m_\ell^{r_2} + \rho^{r_3} m^{r_3}}{\rho_\ell^{r_1} + \rho_\ell^{r_2} + \rho^{r_3}},\, (\rho_\ell^{r_1} + \rho_\ell^{r_2} + \rho^{r_3})^{-1}\right),$$

where

$$\rho_\ell^{r_1} = \frac{1}{u_r^2} |A_\ell^r|, \quad \rho_\ell^{r_2} = \frac{1}{u_r^2} \left(\frac{w_\ell^r}{1-w_\ell^r}\right)^2 |B_\ell^r|, \quad \rho^{r_3} = \frac{1}{b_{\eta^r}^2},$$

and

$$m_\ell^{r_1} = \frac{1}{|A_\ell^r|}\sum_{A_\ell^r} r_{tk}, \quad m_\ell^{r_2} = \frac{1}{|B_\ell^r|}\sum_{B_\ell^r} \frac{\upsilon_r - (1-w_\ell^r)r_{tk}}{w_\ell^r}, \quad m^{r_3} = \upsilon_r,$$

with $A_\ell^r = \{t,\, k | c_{tk}^r = \ell \text{ and } \lambda_{tk}^r = 1\}$; and $|A_\ell^r|$ is the cardinality of this set, $|A_\ell^r| = \sum_{t,\,k} \mathbb{I}(c_{tk}^r = \ell)\mathbb{I}(\lambda_{tk}^r = 1)$. Similarly,
$B_\ell^r = \{t,\, k | c_{tk}^r = \ell \text{ and } \lambda_{tk}^r = 0\}$; and $|B_\ell^r|$ is the cardinality of this set, $|B_\ell^r| = \sum_{t,\,k} \mathbb{I}(c_{tk}^r = \ell)\mathbb{I}(\lambda_{tk}^r = 0)$.

- c_{tk}^r

$$p(c_{tk}^r = \ell \mid -) \propto \psi_\ell^r \left[w_\ell^r \,\mathrm{N}\left(r_{tk} \middle| \eta_\ell^r,\, u_r^2\right) + (1-w_\ell^r)\,\mathrm{N}\left(r_{tk} \middle| \frac{\upsilon_r - w_\ell^r \eta_\ell^r}{1-w_\ell^r},\, u_r^2\right)\right]$$

Update by drawing from the multinomial distribution with probabilities $p(c_{tk}^r = \ell \mid -)$.

- λ_{tk}^r

$$\begin{aligned}
\mathrm{P}(\lambda_{tk}^r = 1 | -) &\propto w_{c_{tk}}\,\mathrm{N}\left(r_{tk} \middle| \eta_{c_{tk}^r,\, u_r^2}\right) \\
\mathrm{P}(\lambda_{tk}^r = 0 | -) &\propto (1-w_{c_{tk}})\,\mathrm{N}\left(r_{tk} \middle| \frac{\upsilon_r - w_{c_{tk}^r}^r \eta_{c_{tk}^r}^r}{1-w_{c_{tk}^r}^r},\, u_r^2\right)
\end{aligned}$$

Update by drawing from the Bernoulli distribution with probability $\mathrm{P}(\lambda_{tk}^r = 1 | -)$.

- r_{tk}

$$\begin{aligned}
\mathrm{P}(r_{tk} \mid c_{tk} = \ell, -) \propto \mathrm{N}\left(r_{tk} \middle| \eta_\ell^r,\, u_r^2\right)^{\lambda_{tk}^r} \mathrm{N}\left(r_{tk} \middle| \frac{\upsilon_r - w_\ell^r \eta_\ell^r}{1-w_\ell^r},\, u_r^2\right)^{1-\lambda_{tk}^r} \\
\times \prod_{j=1}^{J} \left(\frac{\mu_{tkj} s_j}{1+\mu_{tkj} s_j}\right)^{y_{tkj}} \left(\frac{1}{1+\mu_{tkj} s_j}\right)^{s_j^{-1}}
\end{aligned}$$

Update via random walk Metropolis-Hastings.

3. $\boldsymbol{\psi}^\alpha$, $\boldsymbol{w}^\alpha$, $\boldsymbol{\eta}^\alpha$ (OTU baseline abundance parameters)

   - $\boldsymbol{\psi}^\alpha$

$$\mathrm{P}(\boldsymbol{\psi}^\alpha|-) \propto \prod_{\ell=1}^{L^\alpha} (\psi_\ell^\alpha)^{a_\ell^\alpha + d_\ell^\alpha - 1}$$

     Draw from the Dirichlet distribution with concentration parameters $a_\ell^\alpha + d_\ell^\alpha$.

   - w_ℓ^α

$$\begin{aligned}
\mathrm{P}(w_\ell^\alpha|-) &\propto \prod_{j|c_j^\alpha=\ell} (w_\ell^\alpha)^{\lambda_j^\alpha}(1-w_\ell^\alpha)^{1-\lambda_j^\alpha} \\
&\quad \times \prod_{j|c_j^\alpha=\ell \text{ and } \lambda_j^\alpha=0} \mathrm{N}\left(\alpha_{0j}\middle|\frac{\upsilon_\alpha - w_\ell^\alpha\eta_\ell^\alpha}{1-w_\ell^\alpha},\, u_\alpha^2\right) \\
&\quad\quad \times (w_\ell^\alpha)^{a_w^\alpha - 1}(1-w_\ell^\alpha)^{b_w^\alpha - 1} \\
&\propto (w_\ell^\alpha)^{a_w^\alpha + \left[\sum_{j|c_j^\alpha=\ell}\lambda_j^\alpha\right] - 1}(1-w_\ell^\alpha)^{b_w^\alpha + \left[\sum_{j|c_j^\alpha=\ell}(1-\lambda_j^\alpha)\right] - 1} \\
&\quad\quad \times \prod_{j|c_j^\alpha=\ell \text{ and } \lambda_j^\alpha=0} \mathrm{N}\left(\alpha_{0j}\middle|\frac{\upsilon_\alpha - w_\ell^\alpha\eta_\ell^\alpha}{1-w_\ell^\alpha},\, u_\alpha^2\right)
\end{aligned}$$

   - η_ℓ^α

$$\eta_\ell^\alpha \mid - \sim \mathrm{N}\left(\frac{\rho_\ell^{\alpha_1} m_\ell^{\alpha_1} + \rho_\ell^{\alpha_2} m_\ell^{\alpha_2} + \rho^{\alpha_3} m^{\alpha_3}}{\rho_\ell^{\alpha_1} + \rho_\ell^{\alpha_2} + \rho^{\alpha_3}},\, (\rho_\ell^{\alpha_1} + \rho_\ell^{\alpha_2} + \rho^{\alpha_3})^{-1}\right),$$

     where

$$\rho_\ell^{\alpha_1} = \frac{1}{u_\alpha^2}|A_\ell^\alpha|, \quad \rho_\ell^{\alpha_2} = \frac{1}{u_\alpha^2}\left(\frac{w_\ell^\alpha}{1-w_\ell^\alpha}\right)^2 |B_\ell^\alpha|, \quad \rho^{\alpha_3} = \frac{1}{b_{\eta^\alpha}^2},$$

and

$$m_\ell^{\alpha_1} = \frac{1}{|A_\ell^\alpha|}\sum_{A_\ell^\alpha}\alpha_{0j}, \quad m_\ell^{\alpha_2} = \frac{1}{|B_\ell^\alpha|}\sum_{B_\ell^\alpha}\frac{\upsilon_\alpha - (1-w_\ell^\alpha)\alpha_{0j}}{w_\ell^\alpha}, \quad m^{\alpha_3} = \upsilon_\alpha,$$

with $A_\ell^\alpha = \left\{j \middle| c_j^\alpha = \ell \text{ and } \lambda_j^\alpha = 1\right\}$; and $|A_\ell^\alpha|$ is the cardinality of this set, $|A_\ell^\alpha| = \sum_{j=1}^J \mathbb{I}(c_j^\alpha = \ell)\mathbb{I}(\lambda_j^\alpha = 1)$. Similarly, $B_\ell^\alpha = \left\{j \middle| c_j^\alpha = \ell \text{ and } \lambda_j^\alpha = 0\right\}$; and $|B_\ell^\alpha|$ is the cardinality of this set, $|B_\ell^\alpha| = \sum_{j=1}^J \mathbb{I}(c_j^\alpha = \ell)\mathbb{I}(\lambda_j^\alpha = 0)$.

- c_j^α

$$\mathrm{P}(c_j^\alpha = \ell|-) \propto \psi_\ell^\alpha \left[w_\ell^\alpha \,\mathrm{N}\left(\alpha_{0j}\middle|\eta_\ell^\alpha,\, u_\alpha^2\right) + (1-w_\ell^\alpha)\,\mathrm{N}\left(\alpha_{0j}\middle|\frac{\upsilon_\alpha - w_\ell^\alpha\eta_\ell^\alpha}{1-w_\ell^\alpha},\, u_\alpha^2\right)\right]$$

Update by drawing from the multinomial distribution with probabilities $p(c_j^\alpha = \ell \mid -)$.

- λ_j^α

$$\begin{aligned}\mathrm{P}(\lambda_j^\alpha = 1 \mid c_j^\alpha = \ell, -) &\propto w_\ell^\alpha \,\mathrm{N}\left(\alpha_{0j}\middle|\eta_\ell^\alpha,\, u_\alpha^2\right)\\ \mathrm{P}(\lambda_j^\alpha = 0|c_j^\alpha = \ell, -) &\propto (1-w_\ell^\alpha)\,\mathrm{N}\left(\alpha_{0j}\middle|\frac{\upsilon_\alpha - w_\ell^\alpha\eta_\ell^\alpha}{1-w_\ell^\alpha},\, u_\alpha^2\right)\end{aligned}$$

Update by drawing from the Bernoulli distribution with probability $\mathrm{P}(\lambda_j^\alpha = 1|-)$.

- α_{0j}

$$\mathrm{P}(\alpha_{0j} \mid c_j^\alpha = \ell, -) \propto \mathrm{N}\left(\alpha_{0j} \mid \eta_\ell^\alpha, u_\alpha^2\right)^{\lambda_j^\alpha} \mathrm{N}\left(\alpha_{0j} \mid \frac{v_\alpha - w_\ell^\alpha \eta_\ell^\alpha}{1 - w_\ell^\alpha}, u_\alpha^2\right)^{1-\lambda_j^\alpha} \times \prod_t \prod_{k=1}^{K_i} \left(\frac{\mu_{tkj} s_j}{1 + \mu_{tkj} s_j}\right)^{y_{tkj}} \left(\frac{1}{1 + \mu_{tkj} s_j}\right)^{s_j^{-1}}$$

  Update via random walk Metropolis-Hastings.

4. θ_{mj} and τ_j^2 (parameters for process convolution)

- θ_{mj}

$$\mathrm{P}(\theta_{mj}|-) \propto \mathrm{N}(\theta_{mj}|0,\, \tau_j^2) \prod_t \prod_{k=1}^{K_i} \left(\frac{\mu_{tkj} s_j}{1 + \mu_{tkj} s_j}\right)^{y_{tkj}} \left(\frac{1}{1 + \mu_{tkj} s_j}\right)^{s_j^{-1}}$$

  Update via random walk Metropolis-Hastings.

- τ_j^2

$$\tau_j^2|- \sim \mathrm{IG}\left(a_\tau + \frac{M}{2},\, b_\tau + \frac{1}{2} \sum_{m=1}^{M} \theta_{mj}^2\right)$$

5. $\pi_{p0}^\star$, π_{p1}, β_{jp}, γ_{jp}, ι_p and σ_j^2 (parameters related to covariate effects)

- $\pi_{p0}^\star$

$$\pi_{p0}^\star \mid - \sim \mathrm{Be}\left(a_{\pi 0} + \sum_{j=1}^{J} \mathbb{I}\left(\gamma_{jp} = 0\right),\, b_{\pi 0} + \sum_{j=1}^{J} \mathbb{I}\left(\gamma_{jp} \neq 0\right)\right)$$

- π_{p1}

$$\pi_{p1}|- \sim \mathrm{Be}\left(a_{\pi 1} + \sum_{j=1}^{J} \mathbb{I}\left(\gamma_{jp} = 1\right),\, b_{\pi 1} + \sum_{j=1}^{J} \mathbb{I}\left(\gamma_{jp} = 2\right)\right)$$

- σ_p^2

$$\begin{aligned}\mathrm{P}(\sigma_p^2|-) \propto (\sigma_p^2)^{-a_\sigma-1}\exp\left(-\frac{b_\sigma}{\sigma_p^2}\right) \\ \times \prod_{j=1}^{J}\left[\mathrm{N}(\beta_{jp}|0,\,\sigma_p^2)\right]^{\mathbb{I}(\gamma_{jp}=2)}\left[\mathrm{N}(\beta_{jp}|0,\sigma_p^2)\right]^{\mathbb{I}(\gamma_{jp}=1)} \\ \times \mathbb{I}(\sigma_p^2 < U_{\sigma_p^2})\end{aligned}$$

where $U_{\sigma_p^2}$ is an upper bound given by $\min\{(\beta_{jp}/\iota_p)^2,\ \text{for } j \text{ with } \gamma_{jp}\neq 0\}$. Update by drawing σ_p^2 from $\mathrm{IG}\left(a_\sigma + \sum_{j=1}^{J} I(\gamma_{jp}\neq 0)/2,\quad b_\sigma + \sum_{j=1}^{J}\beta_{jp}^2/2\right)$, but truncated at $U_{\sigma_p^2}$.

- ι_p

$$\begin{aligned}\mathrm{P}(\iota_p|-) \propto \iota_p^{a_\iota-1}\exp\left(-b_\iota\iota_p\right) \\ \times\left(\frac{1}{\Phi\left(-\iota_p\right)}\right)^{\sum_{j=1}^{J}\mathbb{I}(\gamma_{jp}=2)}\times\left(\frac{1}{1-\Phi\left(\iota_p\right)}\right)^{\sum_{j=1}^{J}\mathbb{I}(\gamma_{jp}=1)}\times\mathbb{I}(\iota_p > U_{\iota_p}),\end{aligned}$$

where $U_{\iota_p} = \min\{|\beta_{jp}|/\sigma_p,\ \text{for } j \text{ with } \gamma_{jp}\neq 0\}$. Update via random walk Metropolis-Hastings.

- β_{jp}

We do not update β_{jp} if $\gamma_{jp}=0$. For β_{jp} with γ_{jp}=1:

$$\begin{aligned}\mathrm{P}(\beta_{jp}|\gamma_{jp}=1,\,-) \propto \frac{\phi(\beta_{jp}|0,\,\sigma_p^2)}{1-\Phi(\iota_p)}\mathbb{I}(\frac{\beta_{jp}}{\sigma_p}>\iota_p) \\ \times\prod_{i=1}^{n}\prod_{k=1}^{K_i}\left(\frac{\mu_{tkj}s_j}{1+\mu_{tkj}s_j}\right)^{y_{tkj}}\left(\frac{1}{1+\mu_{tkj}s_j}\right)^{s_j^{-1}}\end{aligned}$$

Use a Metropolis-Hastings algorithm to update using proposal

$$\text{TruncNorm}(\beta^{\star}_{jp}, (\sigma^2)', \iota_p\sigma_p, \infty),$$

where $(\sigma^2)'$ is a fixed proposal variance. For β_{jp} with γ_{jp}=2:

$$\begin{aligned}\text{P}(\beta_{jp}|\gamma_{jp}=2,\,-) \propto \frac{\phi(\beta_{jp}|0,\,\sigma_p^2)}{\Phi(-\iota_p)}\mathbb{I}(\frac{\beta_{jp}}{\sigma_p} < -\iota_p) \\ \times \prod_{i=1}^{n}\prod_{k=1}^{K_i}\left(\frac{\mu_{tkj}s_j}{1+\mu_{tkj}s_j}\right)^{y_{tkj}}\left(\frac{1}{1+\mu_{tkj}s_j}\right)^{s_j^{-1}}\end{aligned}$$

Do a metropolis update using proposal $\text{TruncNorm}(\beta^{\star}_{jp}, (\sigma^2)', -\infty, -\iota_p\sigma_p)$, where $(\sigma^2)'$ is a fixed proposal variance.

	a_σ	b_σ	a_ι	b_ι
Case 1	1	1	2.5	10
Case 2	0.1	0.1	2.5	10
Case 3	1	1	1	10
Case 4	0.1	0.1	1	10
Case 5	1	1	5	20
Case 6	0.1	0.1	5	20

Table A.1: Prior specifications for ι_p and σ_p^2. $\iota_p \overset{iid}{\sim} \text{Gamma}(a_\iota, b_\iota)$ and $\sigma_p^2 \overset{iid}{\sim} \text{IG}(a_\sigma, b_\sigma)$ are assumed.

A.2 Additional Results for Simulation 1

In this section we present additional results from Simulation 1 in §2.3 of the main text. Figure A.1 has histograms of posterior estimates $\hat{d}_{jp} = \hat{\text{P}}(\gamma_{jp} = \gamma_{jp}^{\text{TR}} \mid \boldsymbol{Y})$, the probabilities that β_{jp} is correctly selected with its true direction. Figure

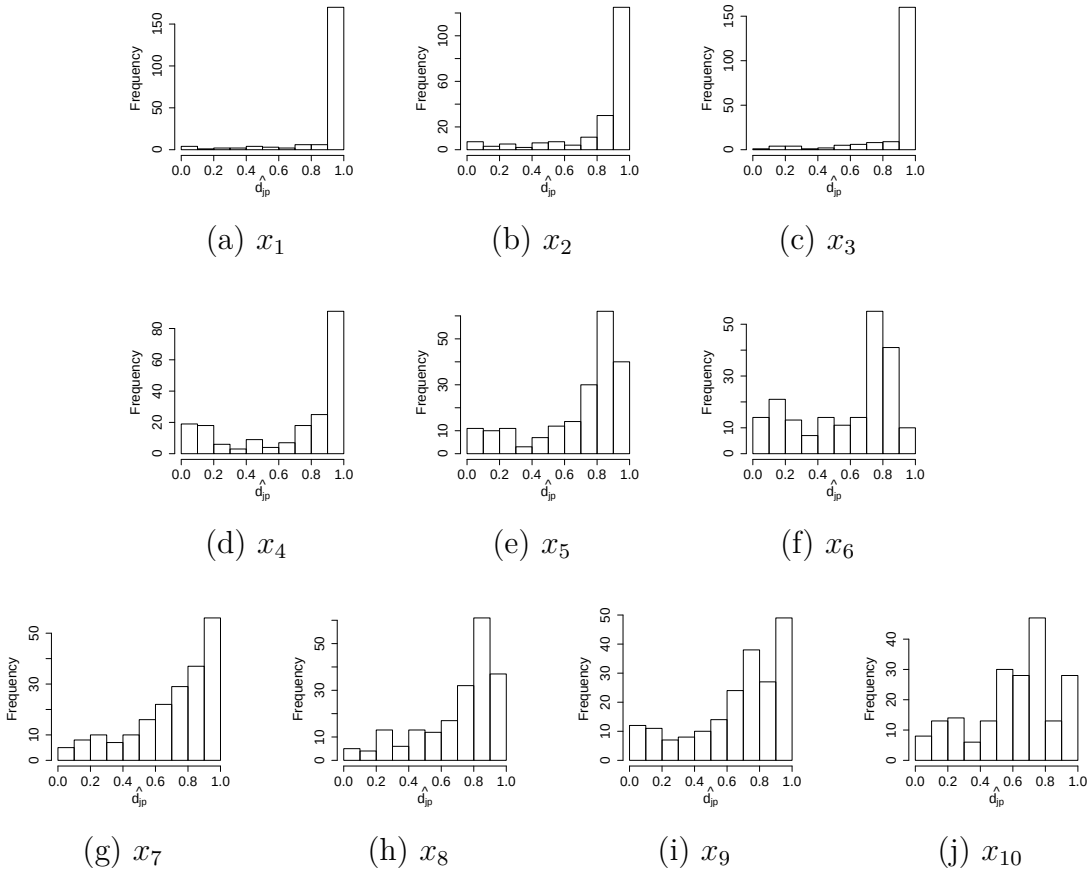


Figure A.1: [Simulation 1] Histograms of $\hat{d}_{jp} = \hat{\mathrm{P}}(\gamma_{jp} = \gamma_{jp}^{\mathrm{TR}} \mid \boldsymbol{Y})$.

A.2 compares the posterior mean estimates $\hat{\beta}_{jp}$ to their true values β_{jp}^{TR} for all covariates. Note that covariates $x_1 - x_3$ are continuous covariates and $x_4 - x_{10}$ are binary indicators to represent different concentration levels of *Alexandrium* (Ax) and domoic acid (DA). Figure A.3(a) compares posterior estimates $\hat{g}_{tkj}$ of baseline mean counts to their true values. The difference of $g_{tkj}^{\mathrm{TR}} - \hat{g}_{tkj}$ is distributed tightly around zero, indicating the baseline counts are well estimated. Figure A.3(b) and (c) compare posterior estimates $\hat{r}_{tk}$ and $\hat{\alpha}_{0j}$ of library size adjustment factors r_{tk} and OTU specific baseline abundance α_{0j} to their true values, respectively. We observe vertical shifts from the true values in opposite directions; that is,

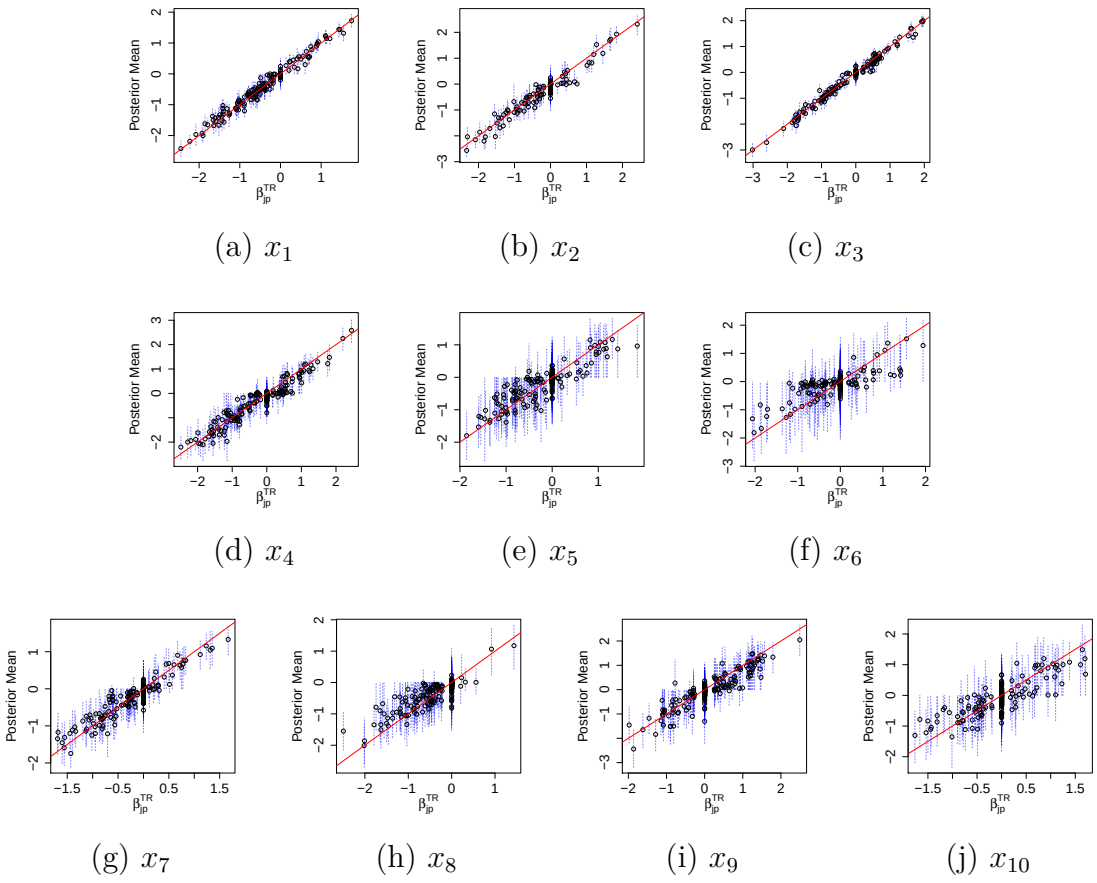
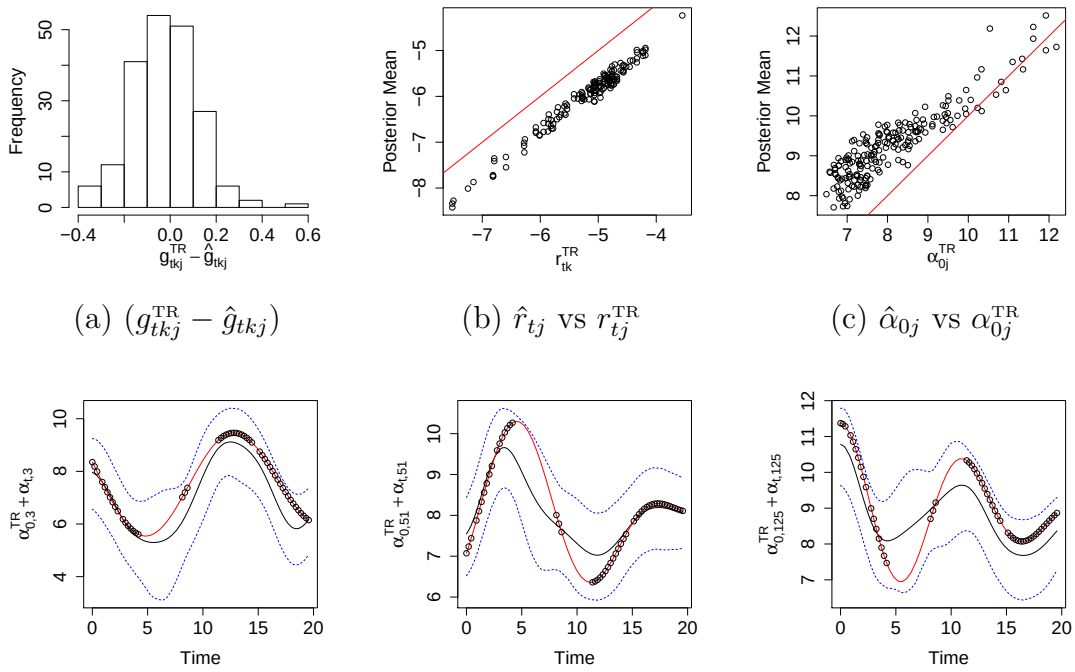


Figure A.2: [Simulation 1] Posterior means of the regression coefficients $\hat{\beta}_{jp}$ versus their true values β_{jp}^{TR}. The dashed blue lines show 95% posterior credible intervals, and the solid red lines are 45 degree reference lines.

r_{tk}s are underestimated for all OTUs and α_{0j} are overestimated due to the lack of identifiability in the construction of g_{tkj}. Figure A.3(d)-(f) show posterior estimates $\hat{\alpha}_{tj}$ of the temporal structure. In the figure, we plot $\alpha_{0j}^{\mathrm{TR}} + \hat{\alpha}_{tj}$ over t for some selected OTUs. For easy comparison to the truth, $\alpha_{0j}^{\mathrm{TR}}$ is used instead of posterior estimates of α_{0j}. The red line in the figure represents the simulation truth. Sample time points are shown by open black circles, and the posterior mean is shown by a black line with pointwise 95% credible intervals (blue dashed



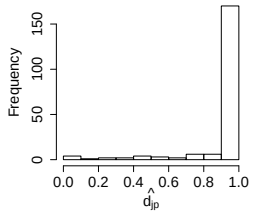
(a) $(g^{\mathrm{TR}}_{tkj} - \hat{g}_{tkj})$ (b) $\hat{r}_{tj}$ vs r^{TR}_{tj} (c) $\hat{\alpha}_{0j}$ vs $\alpha^{\mathrm{TR}}_{0j}$

(d) $\hat{\alpha}_{tj}$ with $j = 3$ (e) $\hat{\alpha}_{tj}$ with $j = 51$ (f) $\hat{\alpha}_{tj}$ with $j = 125$

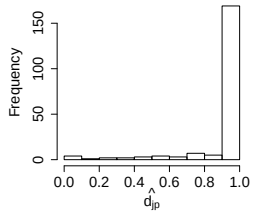
Figure A.3: [Simulation 1] Panel (a): Histogram of the differences between the posterior means of the baseline mean counts $\hat{g}_{tkj}$ and their true values g^{TR}_{tkj}. Panel (b): Posterior means of the sample scale factors $\hat{r}_{tk}$ versus their true values r^{TR}_{tk}. Panel (c): Posterior means of the OTU-specific baseline abundance $\hat{\alpha}_{0j}$ versus their true values $\alpha^{\mathrm{TR}}_{0j}$. Panels (d) through (f): Posterior means $\hat{\alpha}_{tj}$ of α_{tj} (black, solid) compared to the simulation truth (red, solid) for some selected OTUs with 95% credible intervals (blue, dotted).

lines). Overall, the model does a good job of capturing the temporal dependence introduced by the sampling procedure.

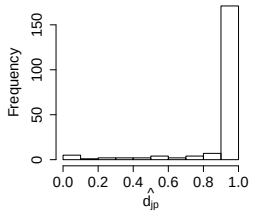
We examined the sensitivity of the estimation of β_{jp} and γ_{jp} to the prior specification of ι_p and σ_p for ANLP-SB. Recall that $\iota_p \overset{iid}{\sim} \mathrm{Gamma}(a_\iota, b_\iota)$ and $\sigma_p^2 \overset{iid}{\sim} \mathrm{IG}(a_\sigma, b_\sigma)$ are assumed. Six different specifications of $(a_\iota, b_\iota, a_\sigma, b_\sigma)$ are given in Table A.1, including the specification used in §3 of the main text to facilitate easy comparison. Results are illustrated in Figure A.4 for x_1 and x_3, the



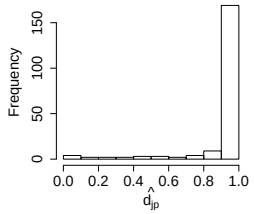
(a) Case 1: (1, 1, 2.5, 10)



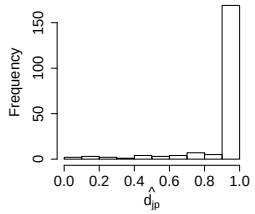
(b) Case 2: (0.1, 0.1, 2.5, 10)



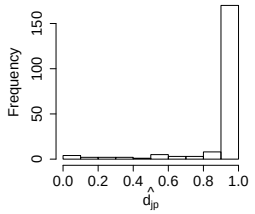
(c) Case 3: (1, 1, 1, 10)



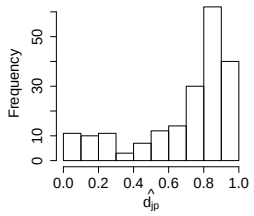
(d) Case 4: (0.1, 0.1, 1, 10)



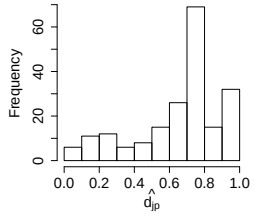
(e) Case 5: (1, 1, 5, 20)



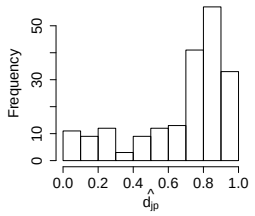
(f) Case 6: (0.1, 0.1, 5, 10)



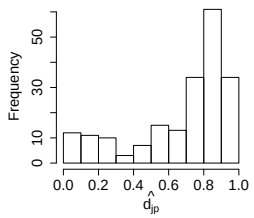
(g) Case 1: (1, 1, 2.5, 10)



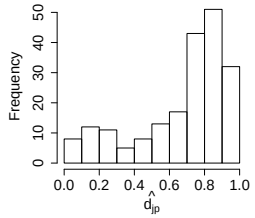
(h) Case 2: (0.1, 0.1, 2.5, 10)



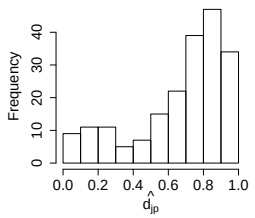
(i) Case 3: (1, 1, 1, 10)



(j) Case 4: (0.1, 0.1, 1, 10)

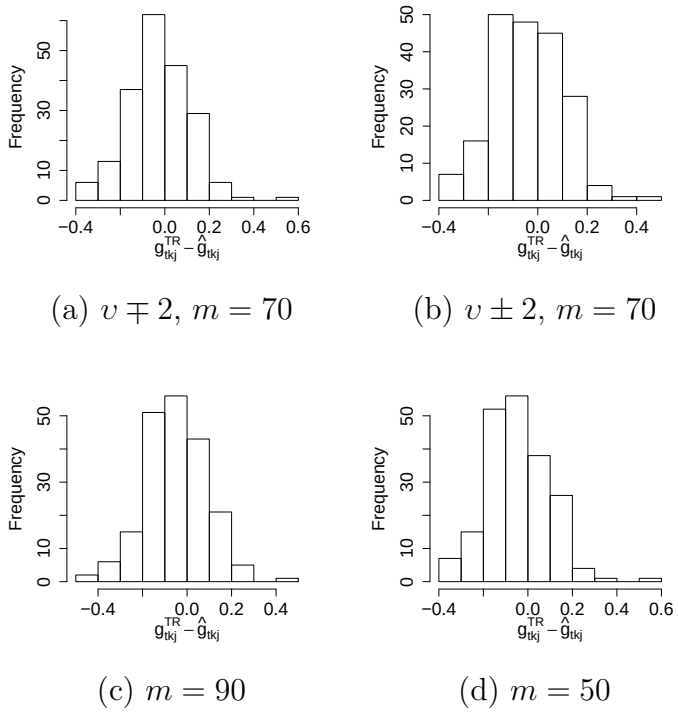


(k) Case 5: (1, 1, 5, 20)



(ℓ) Case 6: (0.1, 0.1, 5, 10)

Figure A.4: [Simulation 1] Histograms of posterior estimates of $\hat{d}_{jp} = \hat{\mathrm{P}}(\gamma_{jp} = \gamma_{jp}^{\mathrm{TR}})$ for selected covariates x_1 and x_5 under the six different specifications of $(a_\iota, b_\iota, a_\sigma, b_\sigma)$ in Table A.1. Panels (a)-(f) show results from x_1 (Silicate), and panels (g)-(ℓ) show results from x_5 (low concentration of Alexandrium).



(a) $\upsilon \mp 2$, $m = 70$ (b) $\upsilon \pm 2$, $m = 70$

(c) $m = 90$ (d) $m = 50$

Figure A.5: [Simulation 1] Histograms of differences between the true baseline mean counts and their estimates, $g^{\text{TR}}_{tkj} - \hat{g}_{tkj}$, under different specifications for υ_r, υ_α and M.

same covariates in Figure 2.3 in the main text. The figures show that the model is not overly sensitive to the prior specification of ι_p and σ^2_p within a reasonable range. We found that overdispersed priors for ι_p may lead to poor convergence and/or inference (results are not shown). We also examined the sensitivity of the estimation of g_{tkj} by varying the number of knot points M and the values of the prespecified mean constraints υ_r and υ_α. We tried $M = 50$ and $M = 90$ and compared the results to those with $M = 70$. We found relatively little impact on the posterior inference. The model is more sensitive to changes in the range parameter, as the range parameter can be chosen in such a way that the temporal dependence between sample points is too strong. We also specified υ_r and υ_α

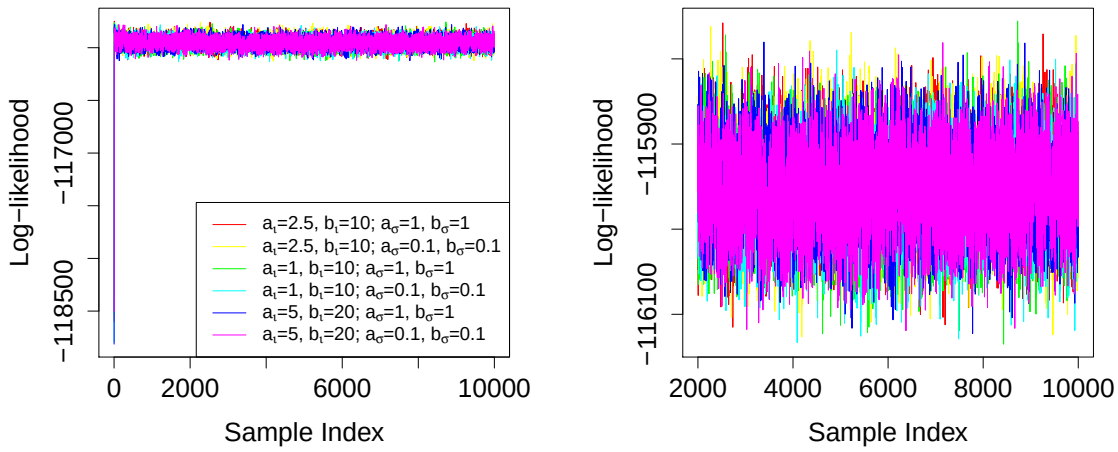


Figure A.6: [Simulation 1] Trace plots of the log-likelihood under different prior specifications. The plots are over the course of the entire MCMC (left), and after 2,000 samples (right).

by ± 2 and ∓ 2 to the values empirically specified as described in the main text. We found that these changes had little impact on the posterior estimates of g_{tkj}. Histograms of the differences between the true baseline mean counts g_{tkj}^{TR} and their posterior estimates $\hat{g}_{tkj}$ under the different simulation conditions are shown in Figure A.5.

We diagnosed the convergence of the posterior MCMC simulation using trace plots of the log-likelihood. Figure A.6 shows trace plots of the log-likehood based on imputed parameters under the six different hyperparameter specifications in Table A.1. The chains converge to similar log-likelihood ranges regardless of the hyperparameter specification, which provides no evidence of poor convergence or poor mixing. Examination of trace plots and autoccorrelation plots for some parameters also show practical evidence of posterior convergence and good mixing (not shown).

A.3 Additional Results for Ocean Microbiome Data Analysis

In this section, we present additional results from the ocean microbiome data analysis in §2.4 of the main text. Table A.2 shows the names of the covariates. Figure A.7 has simplex plots of the posterior probability vectors $\hat{\boldsymbol{z}}_{jp} = (\hat{z}_{jp0},\ \hat{z}_{jp1}, \hat{z}_{jp2})$, where $\hat{z}_{jpq}$ is an posterior estimate of the probability that $\gamma_{jp} = q$, $q \in \{0, 1, 2\}$ for the complete set of covariates under ANLP-SB. Each point on the simplex plot represents the posterior probabilities of no effect (bottom left tip)/positive effect (apex)/negative effect (bottom right tip) of that covariate for an OTU. These plots provide insights into how the covariates are associated with community dynamics. For some covariates the effect directions on the OTU abundances are relatively homogeneous, while other covariates have more varied effects. Figure A.8 illustrates posterior inference for the OTUs belonging to class *Gamma-proteobacteria*. The figure has histograms of posterior estimates of regression coefficients ($\hat{\beta}_{jp}$) and probabilities of $\gamma_{jp} = 2$ ($\hat{z}_{jp2}$) for the different DA concentration levels. Figure A.9 shows results on bacterial growth of a lab experiment using a cultured *Gamma-proteobacteria* strain. The results show that the bacteria was significantly affected by DA at concentrations ranging from 25 to 50 $\mu g/ml$, which validates our findings from Figure A.8. To assess the convergence of the posterior MCMC simulation, we re-ran the data analysis under the six different hyperparameter specifications given in Table A.1. The trace plots in Figure A.10 illustrate that the log-likehood converges to similar states under the different prior specifications and different random seeds. The results provide practical evidence of posterior convergence and indicate robustness of ANLP-SB to the prior specification.

For comparison, we applied the competing models to the ocean microbiome

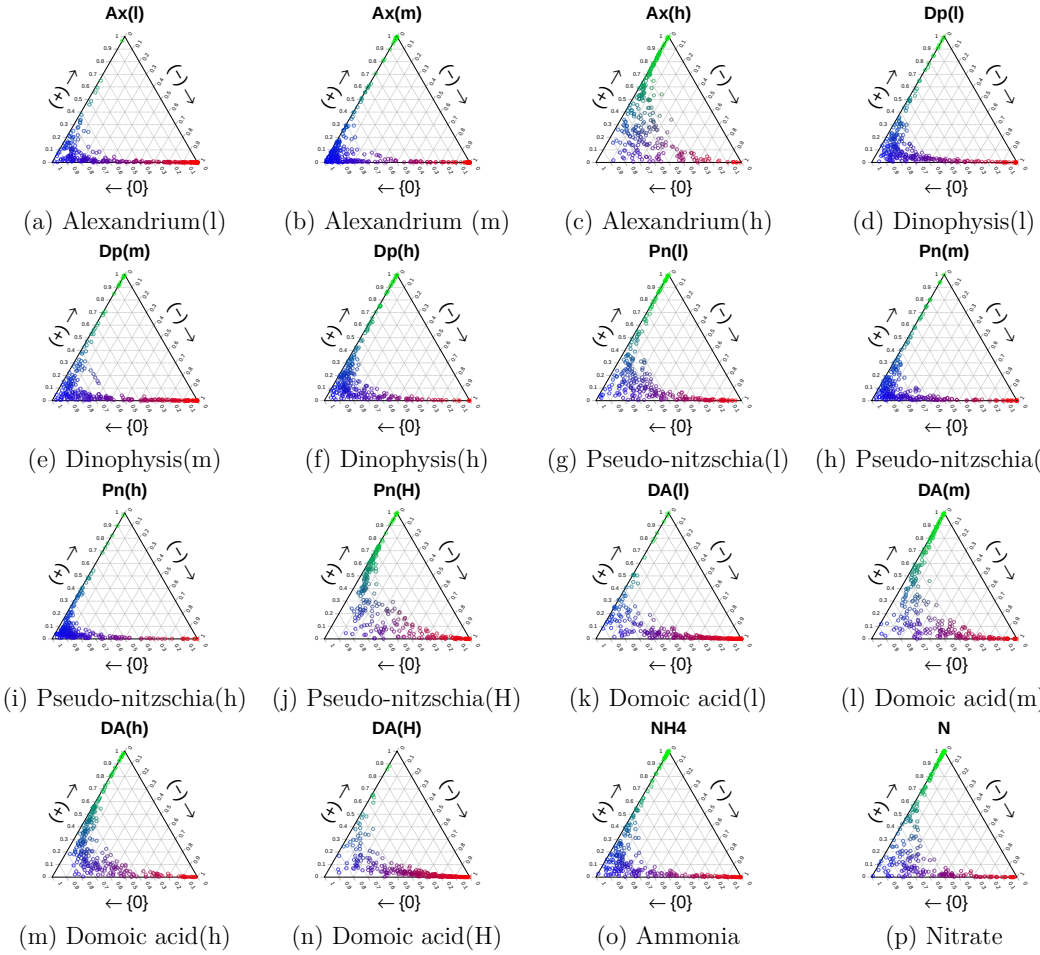
Covariate Name	Short Name	Levels
Alexandrium ($x_1 - x_3$)	Ax	low/medium/high
Dinophysis ($x_4 - x_6$)	Dp	low/medium/high
Pseudo-nitzschia ($x_7 - x_{10}$)	Pn	low/medium/high/very high
Domoic acid ($x_{11} - x_{14}$)	DA	low/medium/high/very high
Ammonia (x_{15})	NH_4	continuous
Nitrate (x_{16})	N	continuous
Phosphate (x_{17})	P	continuous
Silicate (x_{18})	Si	continuous
Water Temperature (x_{19})	T	continuous
Chlorophyll (x_{20})	Chl	continuous

Table A.2: Covariate names in the ocean microbiome dataset. Categories are listed for the discretized covariates.

Model	DIC	LPML
ANLP-SB	**256,189**	**-3.252**
ALP-SB	256,238	-3.254
SLP-SB	256,330	-3.254
BayesReg	267,010	-3.609

Table A.3: Performance metrics of the Bayesian models applied to the ocean microbiome dataset. Best performances are in bold.

data. Figure A.11(a)-(b) have the posterior distributions of $\pi^\star_{p0}$ probabilities that a covariate has no effect on OTU abundance and of π_{p1} conditional probabilities that a covariate has a positive effect given it has a significant effect under ALP-SB, respectively. Figure A.11(c) shows posterior distributions of $\pi^\star_{p0}$ under SLP-SB. The posterior distributions of $\pi^\star_{p0}$ and π_{p1} under ANLP-SB are shown in Figure 2.4 of the main text. The models with local priors have posterior mean estimates of $\pi^\star_{p0}$ between 0.2 and 0.4 for all covariates, while those values are above or around 0.5 under ANLP-SB. The ANLP induces more sparsity than the ALP and the SLP, resulting in more parsimonious models. Table A.3 has DIC and LPML for the Bayesian models. ANLP-SB has the best fit according to both criteria, followed by ALP-SB, and then SLP-SB. Figure A.12 shows the proportion of OTUs that



(a) Alexandrium(l) (b) Alexandrium (m) (c) Alexandrium(h) (d) Dinophysis(l)

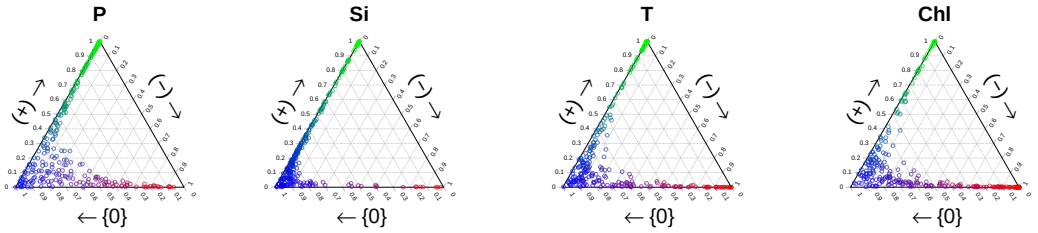
(e) Dinophysis(m) (f) Dinophysis(h) (g) Pseudo-nitzschia(l) (h) Pseudo-nitzschia(m)

(i) Pseudo-nitzschia(h) (j) Pseudo-nitzschia(H) (k) Domoic acid(l) (l) Domoic acid(m)

(m) Domoic acid(h) (n) Domoic acid(H) (o) Ammonia (p) Nitrate

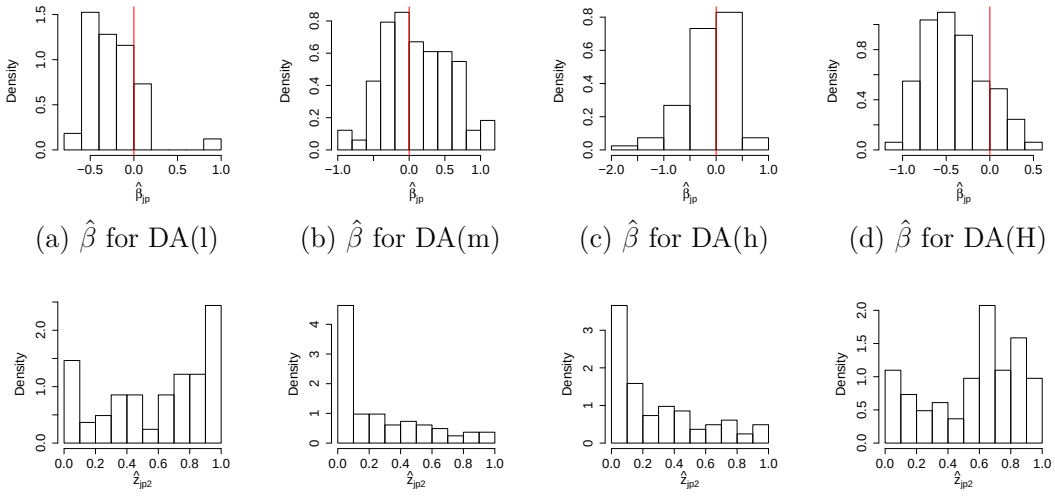
Figure A.7: [Ocean Microbiome Data] Simplex plots of the posterior means $\hat{\boldsymbol{z}}_{jp} = (\hat{z}_{jp0}, \hat{z}_{jp1}, \hat{z}_{jp2})$ of $\gamma_{jp} = 0$, (no effect), $\gamma_{jp} = 1$, (positive effect) and $\gamma_{jp} = 2$, (negative effect). The colors, blue, red and green, indicate no relationship, a negative relationship, and a positive relationship with OTU abundances, respectively.

have a significant relationship with each covariate, where the criteria used to define a variable as 'selected' is the same as was described in §2.3 of the main text. For ANLP-SB, a variable is 'selected' when $\mathrm{P}(\gamma_{jp} \neq 0 \mid \boldsymbol{Y}) > 0.5$. For BayesReg, the 95% posterior credible intervals were used to select a variable. For the frequentist models, we selected a variable when the adjusted p-value for that regression coefficient was less than 0.05.



(q) Phosphate (r) Silicate (s) Water Temperature (t) Chlorophyll

[Ocean Microbiome Data - Supplementary Figure A.7 continued] 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 abundances, respectively.



(a) $\hat{\beta}$ for DA(l) (b) $\hat{\beta}$ for DA(m) (c) $\hat{\beta}$ for DA(h) (d) $\hat{\beta}$ for DA(H)

(e) $\hat{z}_2$ for DA(l) (f) $\hat{z}_2$ for DA(m) (g) $\hat{z}_2$ for DA(h) (h) $\hat{z}_2$ for DA(H)

Figure A.8: [Ocean Microbiome Data] Histograms of $\hat{\beta}_{jp}$ and $\hat{z}_{jp2}$ for different DA concentration levels for OTUs belonging to the class *Gamma-proteobacteria*.

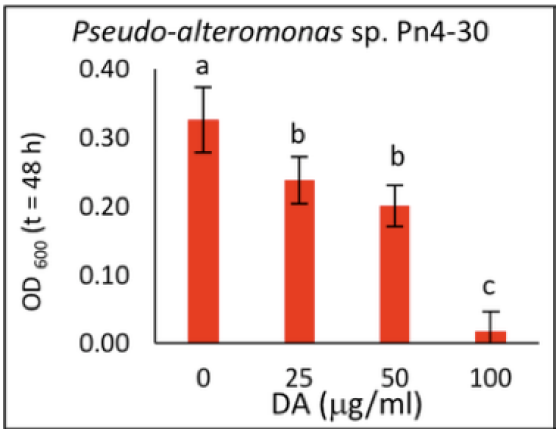


Figure A.9: [Ocean Microbiome Data] Plots of growth measurements (Optical Density at 600 nm) of a bacterial cultured isolate belonging to Gammaproteobacteria measured after 48 hours of exposure to different concentration levels of domoic acid (DA).

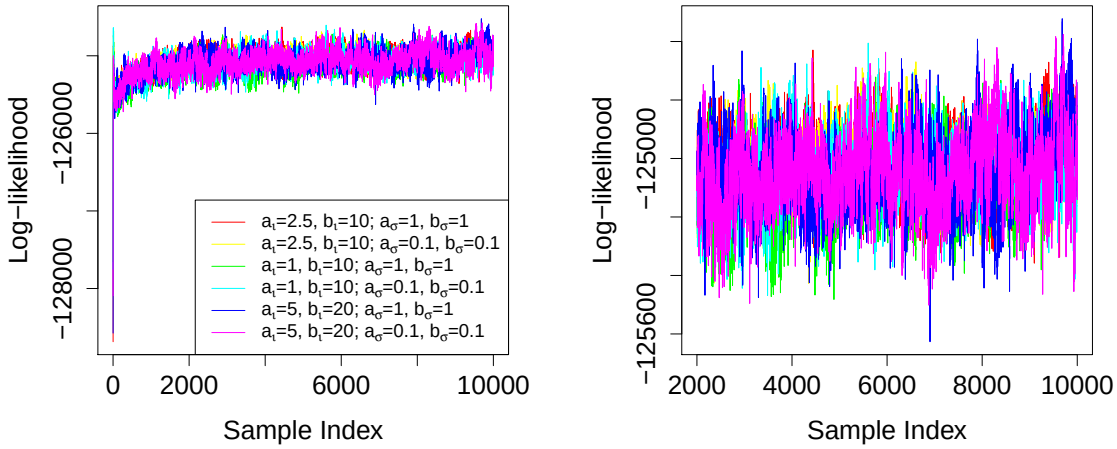
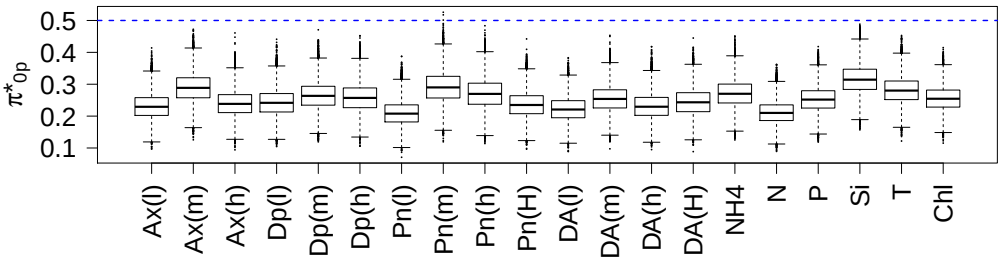
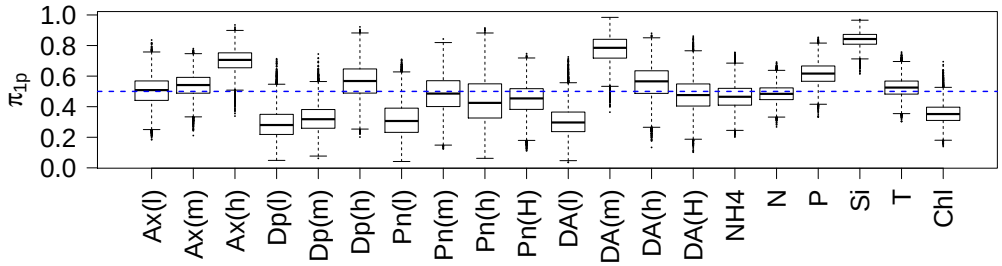


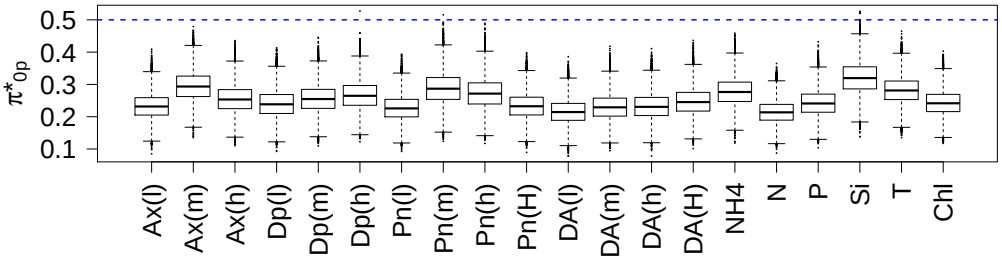
Figure A.10: [Ocean Microbiome Data] Trace plots of the log-likelihood under different prior specifications. The plots are over the course of the entire MCMC (left), and after 2,000 samples (right).



(a) Posterior distributions of $\pi^{\star}_{p0}$ under ALP-SB



(b) Posterior distributions of π_{p1} under ALP-SB



(c) Posterior distributions of $\pi^{\star}_{p0}$ under SLP-SB

Figure A.11: [Ocean Microbiome Data] Panels (a) and (c): Boxplots of the posterior distributions of $\pi^{\star}_{p0}$, the probability of a non-zero effect on OTU abundance, under ALP-SB and SLP-SB, respectively. Panel (b): Boxplots of the posterior distributions of π_{p1}, the conditional probability of a positive effect direction given the covariate has a non-zero effect under ALP-SB.

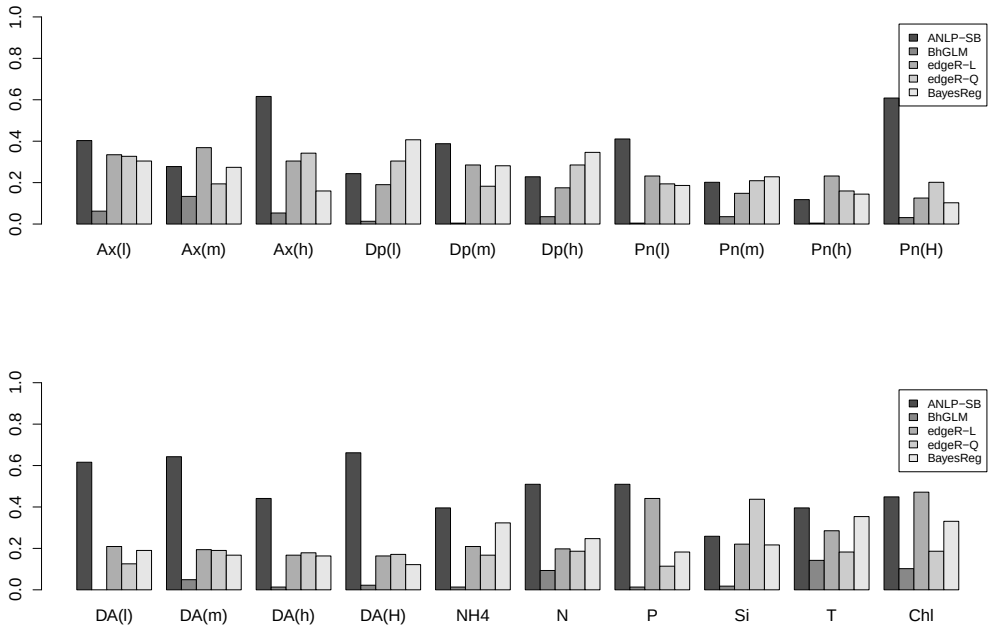


Figure A.12: [Ocean Microbiome Data] Proportions that each covariate is selected under ANLP-SB, BayesReg, BhGLM, edgeR-L and edgeR-Q.

Appendix B

A Bayesian Nonparametric Analysis for Zero Inflated Multivariate Count Data with Application to Microbiome Study Supplementary Material

B.1 MCMC Algorithm

We obtain samples from the posterior distribution using an MCMC algorithm. Let $\underline{\boldsymbol{\theta}} = [s_j, \delta_{ij}, r_i, \alpha_{jm}, \xi_{jk}, \theta_{jk}, (\chi^{\star}_{k\ell}, v^{\chi}_{\ell}, \sigma^2_{\chi k}, \chi \in \{\theta, \xi\}), (\psi^{\chi}_{\ell}, w^{\chi}_{\ell}, \eta^{\chi}_{\ell}, \chi \in \{r, \alpha\})]$ denote the vector of all unknown parameters. The joint posterior distribution is

given by

$$
\begin{aligned}
\mathrm{P}\left(\underline{\boldsymbol{\theta}} \mid \boldsymbol{Y}, \boldsymbol{x}, \boldsymbol{u}\right) &\propto \mathrm{P}\left(\boldsymbol{Y} \mid \underline{\boldsymbol{\theta}}, \boldsymbol{x}, \boldsymbol{u}\right) \mathrm{P}\left(\underline{\boldsymbol{\theta}}\right) \\
&\propto \prod_{i=1}^{n} \prod_{j=1}^{J} \left\{ p(y_{ij} \mid \delta_{ij}, \mu_{ij}(x_i, u_i), s_j) \cdot \pi(\delta_{ij} \mid \epsilon_{jx_i}) \right\} \prod_{j=1}^{J} \pi(s_j \mid a_s, b_s^2) \\
&\quad \times \prod_{j=1}^{J} \prod_{m=1}^{M} \pi(\alpha_{jm} \mid \boldsymbol{\delta}, \boldsymbol{\psi}^{\alpha}, \boldsymbol{w}^{\alpha}, \boldsymbol{\eta}^{\alpha}) \cdot \pi(\boldsymbol{\psi}^{\alpha}, \boldsymbol{w}^{\alpha}, \boldsymbol{\eta}^{\alpha} \mid \upsilon_{\alpha}, u_{\alpha}^2) \\
&\quad \times \prod_{i=1}^{n} \pi(r_i \mid \boldsymbol{\delta}, \boldsymbol{\psi}^{r}, \boldsymbol{w}^{r}, \boldsymbol{\eta}^{r}) \cdot \pi(\boldsymbol{\psi}^{r}, \boldsymbol{w}^{r}, \boldsymbol{\eta}^{r} \mid \upsilon_r, u_r^2) \\
&\quad \times \prod_{k=1}^{K} \left\{ \prod_{j=1}^{J} \pi(\xi_{jk} \mid \boldsymbol{\psi}^{\xi}, \boldsymbol{\xi}_k^{\star}, \sigma_{\xi k}^2) \right\} \pi(\sigma_{\xi k}^2 \mid a_{\sigma}^{\xi}, b_{\sigma}^{\xi}) \\
&\quad \times \prod_{k=1}^{K} \left\{ \prod_{j=1}^{J} \pi(\theta_{jk} \mid \boldsymbol{\delta}, \boldsymbol{\psi}^{\theta}, \boldsymbol{\theta}_k^{\star}, \sigma_{\theta k}^2) \right\} \pi(\sigma_{\theta k}^2 \mid a_{\sigma}^{\theta}, b_{\sigma}^{\theta}) \\
&\quad \times \prod_{\ell=1}^{L^{\xi}} \left\{ \prod_{k=1}^{K} \left\{ \pi(\xi_{k\ell}^{\star} \mid \bar{\xi}^{\star}, \tau_{\xi}^2) \right\} \pi(\upsilon_{\ell}^{\xi} \mid \rho^{\xi}) \right\} \\
&\quad \times \prod_{\ell=1}^{L^{\theta}} \left\{ \prod_{k=1}^{K} \pi(\theta_{k\ell}^{\star} \mid \bar{\theta}^{\star}, \tau_{\theta}^2) \pi(\upsilon_{\ell}^{\theta} \mid \rho^{\theta}) \right\}.
\end{aligned}
$$

The majority of the MCMC steps consist of straightforward Gibbs and Metropolis-within-Gibbs steps. Recall that parameters α_{mj}, r_i, θ_{jk} and ξ_{jk} are from the mixture models. For easy posterior simulation, we introduce latent variables that indicate which mixture component those parameters are from. We easily draw those latent indicators from categorical distributions. The mixture weights and locations can be easily updated conditioning on the indicators.

The sampling procedure for δ_{ij}, the OTU presence/absence indicators, is complicated because the indicator may have implications for the existence of θ_{jk} and α_{jm}. More specifically, θ_{jk} is only meaningful if there exists at least one $\delta_{ij} = 1$ for $\{i : x_i = k\}$. In words, estimating how an OTU's abundance differs in level k as compared to that OTU's baseline abundance requires that the OTU be present

in at least one sample having covariate level k. Otherwise, we conclude simply that the OTU is absent in that condition. Similarly, α_{jm} can only be estimated or interpreted if there exists at least one $\delta_{ij} = 1$ for $\{i : u_i = m\}$, that is, estimating the group-specific random effect of an OTU requires the OTU to be present for at least one sample from that group. In addition, we perform a joint update of θ_{jk} and α_{jm} with δ_{ij} in the MCMC using a Metropolis-Hastings step because the full conditionals of θ_{jk} and α_{jm} greatly depend on δ_{ij}, . Let $\Omega = \{\delta_{ij}, \theta_{jk}, \alpha_{jm}\}$, and let Ω_0 denote the parameter set from the previous MCMC iteration and Ω_1 the proposed parameter set. The proposal is accepted/rejected in the usual way, with the Metropolis-Hastings acceptance ratio given by

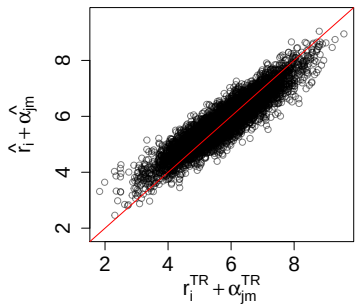
$$\frac{\mathrm{P}(\boldsymbol{Y}_j \mid \Omega_1, \ldots)\mathrm{P}(\Omega_1)}{\mathrm{P}(\boldsymbol{Y}_j \mid \Omega_0, \ldots)\mathrm{P}(\Omega_0)} \frac{Q(\Omega_0 \mid \Omega_1)}{Q(\Omega_1 \mid \Omega_0)}, \tag{B.1}$$

where $\boldsymbol{Y}_j$ is the n-length vector of counts from each sample for OTU j, and $Q(b \mid a)$ a conditional distribution of proposing b given a. Let δ_{ij}^0 and δ_{ij}^1 denote the current and proposed values of δ_{ij}, respectively. We define δ_{ij}^1 conditional on δ_{ij}^0 by letting

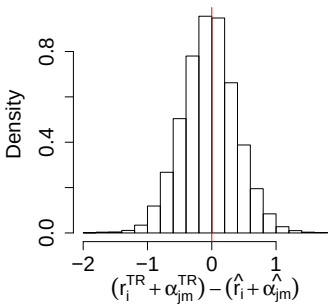
$$\delta_{ij}^1 = \begin{cases} 1, & \text{if } \delta_{ij}^0 = 0, \\ 0, & \text{if } \delta_{ij}^0 = 1. \end{cases}$$

As indicated previously, changing the value of δ_{ij} has implications for α_{jm} and θ_{jk}. Letting $\delta_{ij}^1 = 1$ may require new values for θ_{jk}, α_{jm}, or both. If α_{jm} was defined in the previous MCMC iteration (i.e. α_{jm}^0) we use that value for the proposal, otherwise we draw α_{jm}^1 from its prior distribution, as defined in equation (3.4) of the main text. This draw consists of first drawing the auxiliary variables indicating which mixture components α_{jm}^1 belongs to (conditional on the sets of

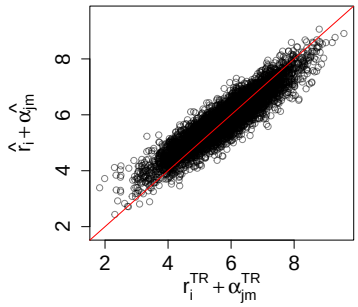
mixture weights $\{\psi_\ell^\alpha\}$ and $\{w_\ell^\alpha\}$), and then drawing α_{jm}^1 from the appropriate Gaussian distribution. Similarly, when proposing $\delta_{ij}^1 = 1$ demands a value for θ_{jk}, we draw θ_{jk}^1 from the prior if θ_{jk}^0 was undefined in the previous MCMC iteration. Otherwise we let $\theta_{jk}^1 = \theta_{jk}^0$ for the proposal. If a prior draw is necessary, $\theta_{jk}^1 = 0$ if an OTU is present for only one level of k, otherwise θ_{jk}^1 is drawn from the distribution defined in equation (3.3) of the main text. This draw is simple when using the finite truncation of the DDP. First an auxiliary variable indicating θ_{jk}'s mixture membership is drawn conditional on the set of mixture weights $\left\{\Psi_\ell^\theta\right\}$, and then θ_{jk}^1 is drawn from the indicated Gaussian distribution. Because $\theta_{jk}^1 = 0$ if an OTU is present for only one level of k, changing δ_{ij} may also require a new value for $\theta_{jk'}^1$, $k' \neq k$. Proposing $\delta_{ij}^1 = 1$ may indicate that an OTU is present for multiple levels of k where previously it was only present in k', in which case the proposal for $\theta_{jk'}^1$ must be some non-zero value. As we do for θ_{jk}^1, we handle this case by drawing $\theta_{jk'}^1$ from the prior $F_{k'}^1$. Likewise, proposing $\delta_{ij}^1 = 0$ may imply $\theta_{jk'}^1 = 0$ if the OTU was previously present for multiple levels of k but now is present for only k'. In many cases there is substantial cancellation in the Metropolis-Hastings ratio in (B.1), which can be used to reduce the computation time when sampling δ_{ij}. Using the prior distributions to draw proposals for θ_{jk} and α_{jm} often results in $\mathrm{P}(\Omega)$ canceling with the transition probabilities. When the proposals θ_{jk}^1 and α_{jm}^1 can be set to θ_{jk}^0 and α_{jm}^0, as is typically the case when the data is not overly sparse, there is considerable cancellation in the likelihood. These cancellations can reduce the computational burden of updating the set of $\{\delta_{ij}\}$, which is high-dimensional in most microbiome settings.



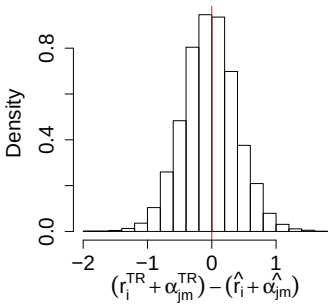
(a) No mean constraint offset



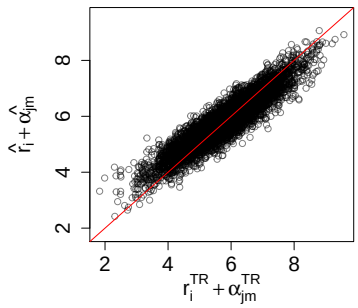
(b) No mean constraint offset



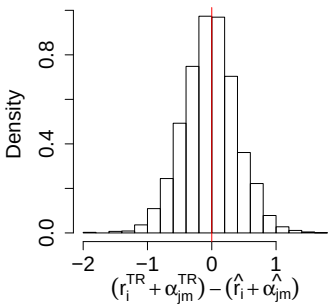
(c) $\upsilon_r + 2$, $\upsilon_\alpha - 2$



(d) $\upsilon_r + 2$, $\upsilon_\alpha - 2$



(e) $\upsilon_r - 2$, $\upsilon_\alpha + 2$



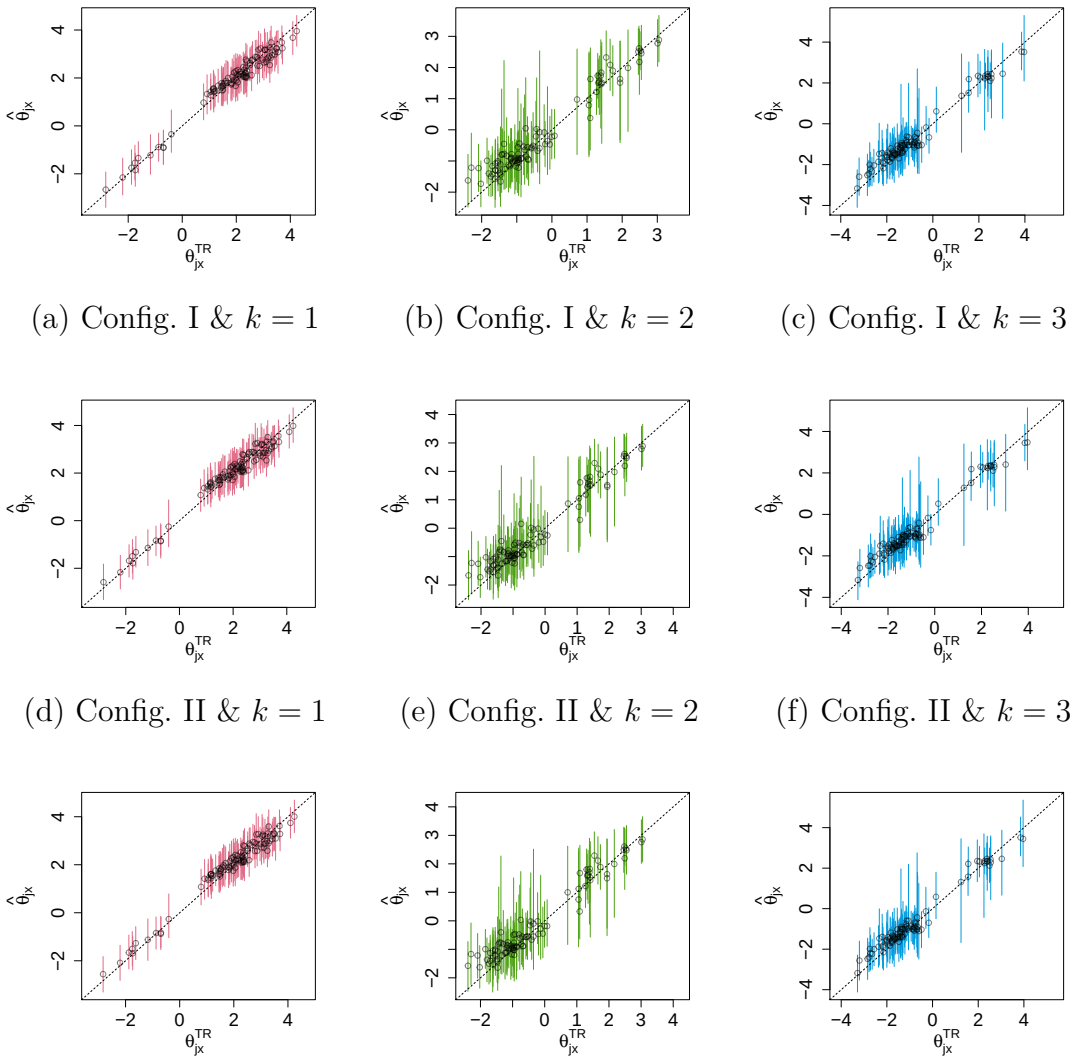
(f) $\upsilon_r - 2$, $\upsilon_\alpha + 2$

Figure B.1: [Simulation 1] Panel (a): Posterior medians of $r_i + \alpha_{jm}$ plotted against the simulation truth. Panel (b): Histogram of the residuals of $r_i + \alpha_{jm}$. Panels (c)-(d): Results when $+2$ is added to υ_α and -2 is added to υ_r. Panels (e)-(f): Results when -2 is added to υ_α and $+2$ is added to υ_r.

B.2 Additional Simulation 1 Results

In this section we present additional results from Simulation 1, described in §3.3 of the main text. Figure B.1 shows posterior estimates of the baseline counts $\hat{r}_i + \hat{\alpha}_{jm}$ of OTU j in sample i compared to the simulation truth. We use the posterior median as their point estimates. The figure illustrates that the quantity $r_i + \alpha_{jm}$ is identifiable, with the residuals between the model estimates and the simulation truth centered roughly at zero. The identifiability of the baseline counts enables differential abundance parameters θ_{jk} to be estimated accurately. We also analyzed the model's sensitivity to specification of the mean constraints υ_α and υ_r. For this sensitivity analysis we set υ_α and υ_r using the procedure described in §3.2.2 of the main text, but with added offsets of ± 2 and ∓ 2. The baseline counts $\hat{r}_i + \hat{\alpha}_{jm}$ compared to the simulation truth under these alternative prior specifications are also shown in Figure B.1(c)-(f). The model is relatively robust to these alternative prior specifications, with the baseline counts well estimated and the residuals centered roughly at zero. Estimates of θ_{jk} under these alternative specifications were similar to the estimates obtained under the original prior specification (not shown).

We also examined the model's convergence and sensitivity to different mixture truncation levels. For the results shown in the main text the truncation specification was $L^\alpha = 150$, $L^r = 20$, $L^\theta = 50$, and $L^\xi = 50$ (Config. I). In addition to this specification we also considered two additional truncation specifications: (1) $L^\alpha = 75$, $L^r = 10$, $L^\theta = 25$, $L^\xi = 25$ (Config. II) and (2) $L^\alpha = 300$, $L^r = 40$, $L^\theta = 100$, $L^\xi = 100$ (Config. III). We found minimal difference in the inference on the parameters of interest produced by the model with these alternative specifications; Figure B.2 shows inference on θ_{jk} is very similar under the different

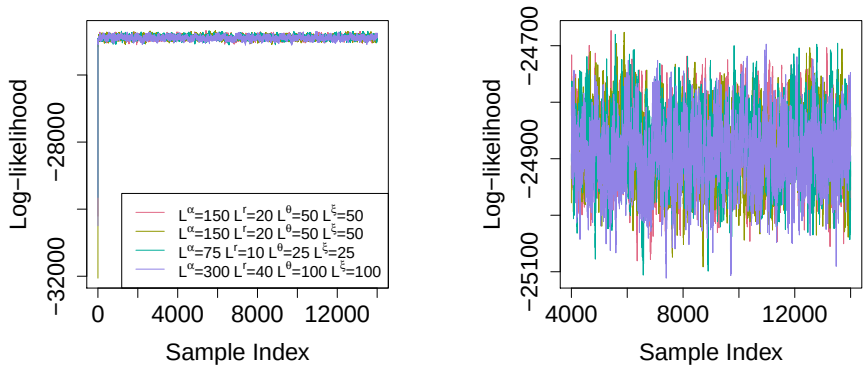


(a) Config. I & $k = 1$ (b) Config. I & $k = 2$ (c) Config. I & $k = 3$

(d) Config. II & $k = 1$ (e) Config. II & $k = 2$ (f) Config. II & $k = 3$

(g) Config. III & $k = 1$ (h) Config. III & $k = 2$ (i) Config. III & $k = 3$

Figure B.2: [Simulation 1] Posterior means of differential abundances θ_{jk} for $k = 1, 2, 3$, along with 95% credible intervals and reference lines. Panels (a)-(c): Original configuration of the truncation levels $L^{\alpha} = 150$, $L^{r} = 20$, $L^{\theta} = 50$ and $L^{\xi} = 50$ (Config. I). Panels (d)-(f): a configuration of truncation levels halved $L^{\alpha} = 75$, $L^{r} = 10$, $L^{\theta} = 25$ and $L^{\xi} = 25$ (Config. II). Panels (g)-(i): a configuration $\boldsymbol{L}_3$ of truncation levels doubled $L^{\alpha} = 300$, $L^{r} = 40$, $L^{\theta} = 100$ and $L^{\xi} = 100$ (Config. III).



(a) (b)

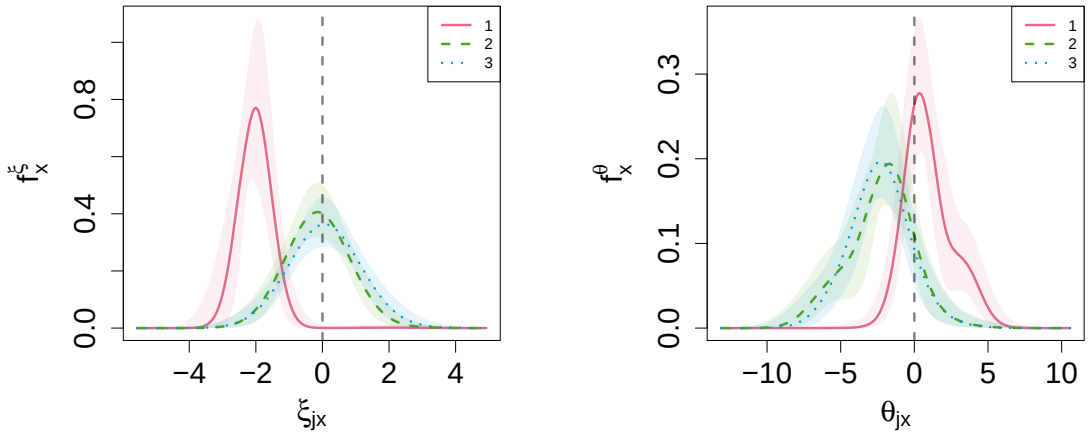
Figure B.3: [Simulation 1] Traceplots of the log-likelihood before burn-in (a) and after burn-in (b). The model specification from the main text (red line) as well as alternative specifications with different random seeds and initializations (other colors) are shown.

truncation specifications. The model was run under different initializations and random seeds for each configuration of the truncation levels. Traceplots of the log-likelihood shown in Figure B.3 provide practical evidence of the model's convergence under these different specifications. As the log-likelihood plots suggest, we found that the model converged to a similar state under these alternative specifications.

B.3 Additional Chronic Wound Microbiome Results

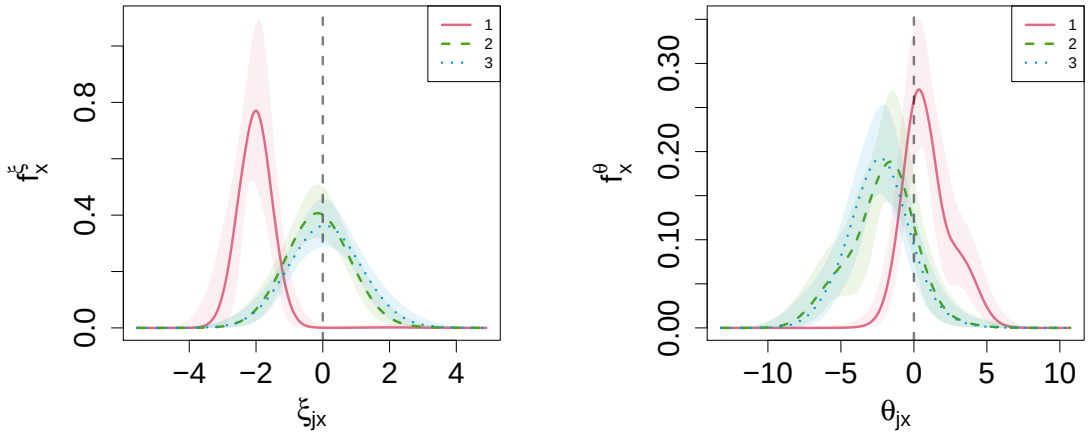
We examined the model's sensitivity to the specification of the mean constraints υ_r and υ_α. Similar to the simulation studies, we added offsets of ± 2 and ∓ 2 to υ_r and υ_α, respectively, and reanalyzed the chronic wound data. Estimates

of f_k^{ξ} and f_k^{θ} with the different specifications of υ_r and υ_α are shown in Figure B.4. Compared to the estimates with the previous specification in Figure 3.11 of the main text, the inference remains almost unchanged. We found that changes in estimation of ϵ_{jk} and θ_{jk} for individual OTUs are also minimal. We conducted a sensitivity analysis to the specification of mixture truncation levels. In addition to the original specification of $L^{\alpha} = 150$, $L^{r} = 20$, $L^{\theta} = 50$, and $L^{\xi} = 50$, we considered $L^{\alpha} = 75$, $L^{r} = 10$, $L^{\theta} = 25$, $L^{\xi} = 25$ and $L^{\alpha} = 300$, $L^{r} = 40$, $L^{\theta} = 100$, and $L^{\xi} = 100$. We found that the inference produced by the model was robust with respect to these alternative specifications. We also ran the model on the chronic wound microbiome dataset with different initializations and random seeds for the MCMC chain and did not find evidence suggesting the Markov chain failed to converge. Traceplots of the log-likelihood under the different truncation specifications and different random seeds and initializations are shown in Figure B.5. The figure shows the MCMC converges to similar log-likelihood ranges under these alternative specifications.



(a) $\hat{f}_k^{\xi}$ with $\upsilon_r + 2$ & $\upsilon_\alpha - 2$

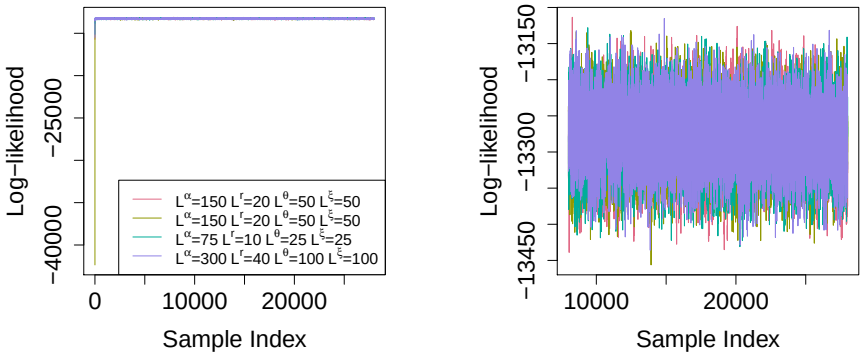
(b) $\hat{f}_k^{\theta}$ with $\upsilon_r + 2$ & $\upsilon_\alpha - 2$



(c) $\hat{f}_k^{\xi}$ with $\upsilon_r - 2$ & $\upsilon_\alpha + 2$

(d) $\hat{f}_k^{\theta}$ with $\upsilon_r - 2$ & $\upsilon_\alpha + 2$

Figure B.4: [Chronic Wound Data - Sensitivity to the specification of υ_r and υ_α] Panels (a) and (b) illustrate estimates of f_k^{ξ} and f_k^{θ}, respectively, when $+2$ is added to υ_α and -2 is added to υ_r. In panels (c) and (d), estimates of f_k^{ξ} and f_k^{θ} are shown when -2 is added to υ_α and $+2$ is added to υ_r.



(a) (b)

Figure B.5: [Chronic Wound Data] Traceplots of the log-likelihood before burn-in (a) and after burn-in (b). The model specification from the main text (red line) as well as alternative specifications with different random seeds and initializations (other colors) are shown.

Appendix C

Bayesian Graphical Modeling of Microbial Community Composition Supplementary Material

C.1 MCMC Algorithm

In this section we outline the steps used to draw samples from the joint posterior distribution via MCMC. Let

$\underline{\boldsymbol{\theta}} = [s_j, r_i, \alpha_j, (\psi_\ell^\chi, w_\ell^\chi, \eta_\ell^\chi, \chi \in \{r, \alpha\}), \beta_{jp}, \tau_p^2, \theta_{mj}, \sigma^2, \gamma_{\ell j}, G]$ be the vector of all unknown parameters. The target distribution is the joint posterior, given by Bayes' rule, $\mathrm{P}(\underline{\boldsymbol{\theta}} \mid \boldsymbol{Y}, \boldsymbol{X}) \propto \mathrm{P}(\underline{\boldsymbol{\theta}})\mathrm{P}(\boldsymbol{Y} \mid \boldsymbol{X}, \underline{\boldsymbol{\theta}})$. We use a combination of Gibbs and Metropolis-within-Gibbs steps to obtain draws from the target distribution. The MCMC steps for the parameters and the graph are described below. Let $\mathrm{Pa}(j)$ denote the set of parents of the j^{th} OTU, and $\mathrm{Ch}(j)$ denote the set of

children.

- s_j

  Let $\tilde{s}_j = \log(s_j)$.

  $$\mathrm{P}(\tilde{s}_j \mid -) \propto \mathrm{N}\left(\tilde{s}_j \mid a_s,\, b_s^2\right) \prod_{i=1}^{n} \frac{\Gamma(Y_{ij} + 1/s_j)}{\Gamma(1/s_j)} \left(\frac{\mu_{ij} s_j}{1 + \mu_{ij} s_j}\right)^{Y_{ij}} \left(\frac{1}{1 + \mu_{ij} s_j}\right)^{1/s_j}$$

  s_j is updated via random-walk Metropolis-Hastings steps.

- $\chi \in \{r, \alpha\}$, w_ℓ^χ, η_ℓ^χ

  The prior distribution for the normalization parameters comes from a mixture-of-mixtures distribution as described in the main text. As is common in finite mixture models, we introduce latent variables indicating from which mixture components the parameter belongs. The mixture component indicators are updated via categorical draws, and r_i and α_j are updated using random-walk Metropolis-Hastings steps conditional on their mixture memberships. For more specifics we refer the reader to Shuler *et al.* (2019a) which describes the algorithm in more detail.

- β_{jp}

  $$\mathrm{P}(\beta_{jp} \mid -) \propto \mathrm{N}(\beta_{jp} | 0,\, \tau_p^2) \prod_{i=1}^{n} \left(\frac{\mu_{ij} s_j}{1 + \mu_{ij} s_j}\right)^{Y_{ij}} \left(\frac{1}{1 + \mu_{ij} s_j}\right)^{1/s_j}$$

  β_{jp} is updated via random-walk Metropolis-Hastings steps.

- τ_p^2

$$\tau_p^2 \sim \text{IG}\left(a_\tau + J/2, b_\tau + \sum_{j=1}^{J} \beta_{jp}^2/2\right)$$

- θ_{mj}

$$\text{P}(\theta_{mj} \mid -) \propto \text{N}\left(\theta_{mj} \mid \sum_{l \in \text{Pa}(j)} \gamma_{lj}\theta_{ml},\, \sigma^2\right) \prod_{j' \in \text{Ch}(j)} \text{N}\left(\theta_{mj'} \mid \sum_{l' \in \text{Pa}(j')} \gamma_{l'j'}\theta_{ml'},\, \sigma^2\right)$$
$$\times \prod_{i|u_i=m} \left(\frac{\mu_{ij}s_j}{1+\mu_{ij}s_j}\right)^{Y_{ij}} \left(\frac{1}{1+\mu_{ij}s_j}\right)^{1/s_j}$$

  θ_{mj} is updated via random-walk Metropolis-Hastings steps.

- σ_j^2

$$\sigma_j^2|- \sim \text{IG}\left(a_\sigma + \frac{M}{2} + \frac{1}{2}\sum_{l=1}^{J} a_{lj},\, b_\sigma + \frac{1}{2}\sum_{i|u_i=m}\left(\theta_{mj} - \sum_{l \in \text{Pa}(j)} \gamma_{lj}\theta_{ml}\right)^2 + \frac{1}{2\kappa}\sum_{l \in \text{Pa}(j)} \gamma_{lj}^2\right)$$

- γ_{lj}

  If OTU l is not a parent of OTU j, no update is necessary. Otherwise:

$$\gamma_{lj} \mid - \sim \text{N}\left(\left(\sigma^2 + \kappa\sigma^2 \sum_{m=1}^{M} \theta_{ml}^2\right)^{-1} \kappa\sigma^2 \sum_{m=1}^{M} \theta_{ml}t_{mj},\, \left(\sigma^2 + \kappa\sigma^2 \sum_{m=1}^{M} \theta_{ml}^2\right)^{-1} \kappa\left(\sigma^2\right)^2\right)$$

  where

$$t_{mj} = \theta_{mj} - \sum_{\substack{l' \in \text{Pa}(j),\\ l' \neq l}} \gamma_{l'j}\theta_{ml'}$$

- G: Update edges individually.

  Repeat this step multiple times within one MCMC iteration.

Select an edge $(j' \to j)$ at random. Let 'A_r' be the Metropolis acceptance ratio.

1. **Birth:** If $(j' \to j) \notin E$ & $(j \to j') \notin E$, propose addition to E.

   Draw a new $\gamma^{\star}_{j'j}$ from the prior $\mathrm{P}(\gamma_{j'j})$ and let $\gamma^{\star}_{lj} = \gamma^{0}_{lj}$ for all $l \in \mathrm{Pa}^{0}(j)$.

   Check that the proposed graph is acyclic, if it is not, reject the proposal.

   Otherwise,

$$A_r = \frac{P_g \prod_{m=1}^{M} \mathrm{N}\left(\theta_{mj} \mid \sum_{l \in \mathrm{Pa}^{\star}(j)} \gamma^{\star}_{lj} \theta_{ml},\, \sigma^2\right)}{\prod_{m=1}^{M} \mathrm{N}\left(\theta_{mj} \mid \sum_{l \in \mathrm{Pa}^{0}(j)} \gamma^{0}_{lj} \theta_{ml},\, \sigma^2\right)}$$

   where $\mathrm{Pa}^{\star}(j)$ denotes the parents in the proposed graph, and $\mathrm{Pa}^{0}(j)$ the parents in the current graph. If the birth move is accepted, we update γ_{lj} from its full conditionals for $l \in Pa(j)$ of the updated G.

2. **Death:** If $(j' \to j) \in E$, propose removal.

$$A_r = \frac{\prod_{m=1}^{M} \mathrm{N}\left(\theta_{mj} \mid \sum_{l \in \mathrm{Pa}^{\star}(j)} \gamma^{\star}_{lj} \theta_{ml},\, \sigma^2\right)}{P_g \prod_{m=1}^{M} \mathrm{N}\left(\theta_{mj} \mid \sum_{l \in \mathrm{Pa}^{0}(j)} \gamma^{0}_{lj} \theta_{ml},\, \sigma^2\right)}$$

   Let $\gamma^{\star}_{lj} = \gamma_{lj}$ for all $l \in \mathrm{Pa}^{\star}(j)$. If the death move is accepted, we update γ_{lj} from its full conditionals for $l \in pa(j)$ of the updated G.

3. **Switch:** If $(j' \to j) \notin E$ & $(j' \leftarrow j) \in E$, propose to remove $(j' \leftarrow j)$ and add $(j' \to j)$:

   Draw a new $\gamma^{\star}_{j'j}$ from the prior $\mathrm{P}(\gamma_{j'j})$ and let $\gamma^{\star}_{lj} = \gamma^{0}_{lj}$ for all $l \in \mathrm{Pa}^{0}(j)$ and $\gamma^{\star}_{lj'} = \gamma^{0}_{lj'}$ for all $l \in \mathrm{Pa}^{0}(j')$.

   Check that the proposed graph is acyclic, if it is not, reject the proposal.

Otherwise,

$$A_r = \frac{\prod_{m=1}^{M} \mathrm{N}\left(\theta_{mj'} \mid \sum_{l \in \mathrm{Pa}^\star(j')} \gamma^\star_{lj'} \theta_{ml},\, \sigma^2\right) \mathrm{N}\left(\theta_{mj} \mid \sum_{l \in \mathrm{Pa}^\star(j)} \gamma^\star_{lj} \theta_{ml},\, \sigma^2\right)}{\prod_{m=1}^{M} \mathrm{N}\left(\theta_{mj'} \mid \sum_{l \in \mathrm{Pa}^0(j')} \gamma^0_{lj'} \theta_{ml},\, \sigma^2\right) \mathrm{N}\left(\theta_{mj} \mid \sum_{l \in \mathrm{Pa}^0(j)} \gamma^0_{lj} \theta_{ml},\, \sigma^2\right)}$$

If the switch move is accepted, we update γ_{lj} from its full conditionals for $l \in pa(j)$ of the updated G and $\gamma_{lj'}$ for $l \in pa(j')$.

Acyclic check:

Let $|E|$ be the number of edges in the graph. If $\mathrm{diag}\left(A^k\right) \neq 0$, for any $k \in \{1, \ldots, \min(J, |E|)\}$, where A^k is the matrix exponent, then G is *not* acyclic (i.e., has a cycle).

- G: Update by edge swap.

  1. Select an OTU having at least one parent, $\mathrm{Pa}(j) \neq \emptyset$
  2. Choose j' from $\mathrm{Pa}(j)$ at random (probability $1/\left|\mathrm{Pa}(j)\right|$)
  3. Choose j'' from $\mathrm{Pa}(j)^c$ at random (probability $1/(J - |\mathrm{Pa}(J)|)$ and add $(j'' \to j)$
  4. If G is acyclic, draw $\gamma_{j'',j}$ from $\mathrm{P}(\gamma_{\ell j})$ and accept $G^\star$ with probability

$$A_r = \frac{\prod_{m=1}^{M} \mathrm{P}(\theta_{ij} \mid G^\star,\, \gamma^\star_{j'',j})}{\prod_{m=1}^{M} \mathrm{P}(\theta_{ij} \mid G^0,\, \gamma^0_{j',j})} = \frac{\prod_{m=1}^{M} \mathrm{N}\left(\theta_{mj} \mid \sum_{l \in \mathrm{Pa}^\star(j)} \gamma^\star_{lj} \theta_{ml},\, \sigma^2\right)}{\prod_{m=1}^{M} \mathrm{N}\left(\theta_{mj} \mid \sum_{l \in \mathrm{Pa}^0(j)} \gamma^0_{lj} \theta_{ml},\, \sigma^2\right)}$$

  5. Update $\gamma_{j'\prime,j}$ several times if $G^\star$ is accepted

C.2 Additional Simulation 1 Results

C.2.1 Sensitivity and Convergence

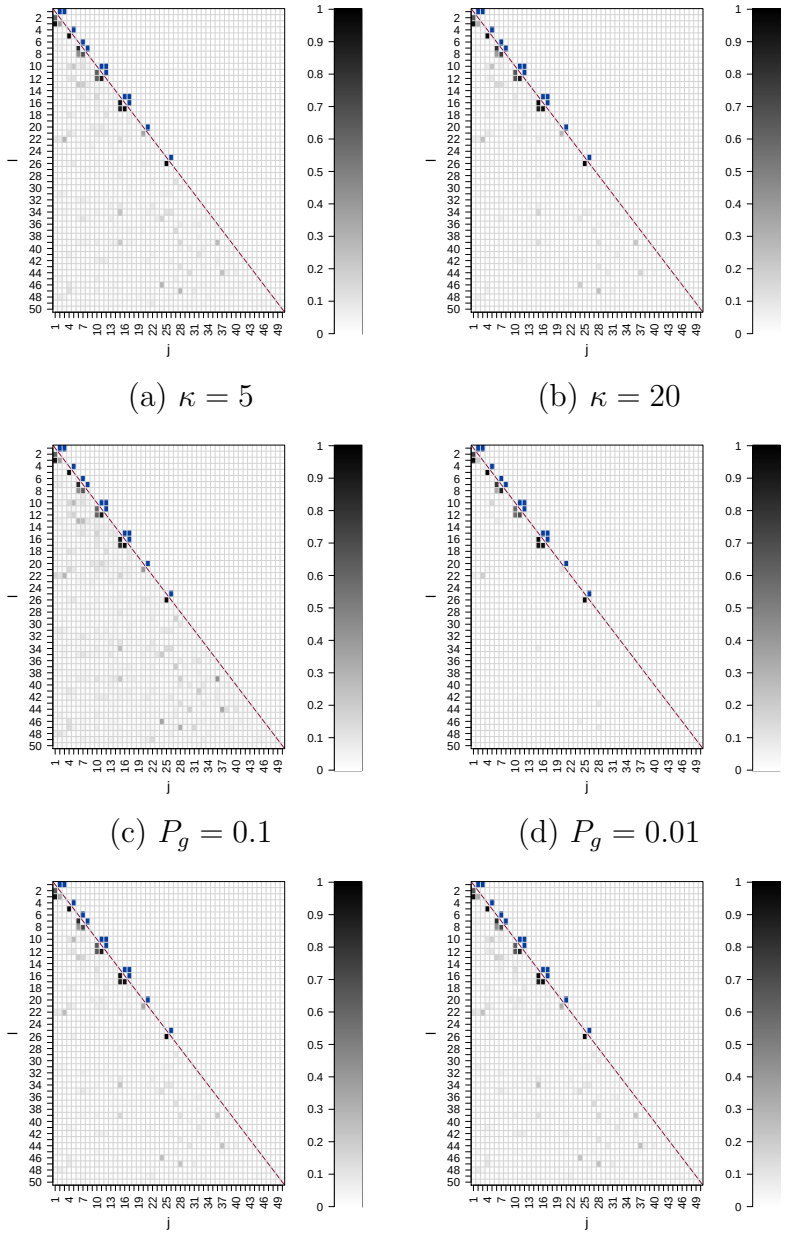
We analyzed the model's sensitivity to different specifications of the fixed scaling factor κ and the edge inclusion probability P_G, as well as to misspecification of υ_α and υ_r. In the results presented in the main text we set $\kappa = 10$ and $P_g = 0.05$. Here, we consider alternative specifications using $\kappa = 5$ and $\kappa = 20$, and using $P_g = 0.1$ and $P_g = 0.01$. The remainder of the hyperparameters were set to the same values as in the main text. For the mean constraints we considered alternative specifications where offsets of ± 2 and ∓ 2 were added to υ_α and υ_r. We found that the alternative specifications had minimal impact on the inference on the graph and regression coefficients. Figure C.1 shows the moral graph edge inclusion probabilities $\bar{m}_{lj}$ using the alternative specifications. The values of $\bar{m}_{lj}$ change very little using alternative specifications of κ and when the mean constraints are misspecified. Figure C.2 shows how the estimates of α_j and r_i change in response to the alternative mean constraint specifications. In each case estimates for the baseline abundances $r_i + \alpha_j$ are on average unbiased, showing the model is robust to misspecification of the mean constraints. P_g has a larger impact on the edge inclusion probabilities, with lower values of P_g yielding lower probabilities of edge inclusion and larger values yielding larger probabilities of edge inclusion, as would be expected from the model specification. Nonetheless, the point estimates for the moral graph obtained from BRM-G remain robust. For all of the specifications we tried the moral graph point estimate $\hat{G}^m$ remained the same as the point estimate obtained using the model specification described in the main text. Estimates obtained for the regression coefficients β_{jp} likewise are robust, with similar estimate obtained across the different specifications we

considered, as shown in Figure C.3 and Figure C.4.

We assessed the chain's convergence by looking at traceplots of the model parameters and comparing the model results to an alternative MCMC chain using a different random seed and having a different initialization. For this alternative chain, rather than using the empirical partial correlations, no edges were included in the graph for the initialization. Also, the regression coefficients β_{jp} were all initialized to 0, and s_j was initialized using $s_j \overset{iid}{\sim}$ Log-Normal(10^{-4}, 10^{-4}) instead of $s_j \overset{iid}{\sim}$ Log-Normal(0.3, 10^{-4}). Using this alternative initialization BRM-G produces the same moral graph point estimate as the original initialization, and the regression coefficients are well recovered. The two chain results are very similar across all of the model's parameters, suggesting good convergence. Because it is not possible to include traceplots for all of the model's many parameters, we show traceplots of the posterior log-likelihood before and after burn-in as a proxy as evidence of the chain's convergence. These traceplots are shown both for the original initialization and the alternative initialization in Figure C.5. The number of edges in the DAG for the two chains is also shown in the figure. The chains quickly converge to including a similar number of edges in the DAG despite using very different initialization procedures. Overall we did not find evidence suggesting the chains did not converge.

C.2.2 Sensitivity and Convergence

As in Simulation 1 we conducted sensitivity analysis for Simulation 2 via alternative specifications of κ and P_g, and to misspecification of the mean constraints υ_r and υ_α. As before we found the model was robust to alternative specifications, with the largest impact coming from different values of P_g which lead to different



(a) $\kappa = 5$ (b) $\kappa = 20$

(c) $P_g = 0.1$ (d) $P_g = 0.01$

(e) $\upsilon_\alpha - 2$, $\upsilon_r + 2$ (f) $\upsilon_\alpha + 2$, $\upsilon_r - 2$

Figure C.1: [Simulation 1 Sensitivity] Lower-diagonals: Posterior probabilities of edge inclusion using different model specifications. Upper-diagonals: Edges of the true moral graph M^{TR}.

edge inclusion probabilities. Figure C.11 shows estimates for r_i and α_j using the alternative mean constraint specifications. Again we find estimates for the base-

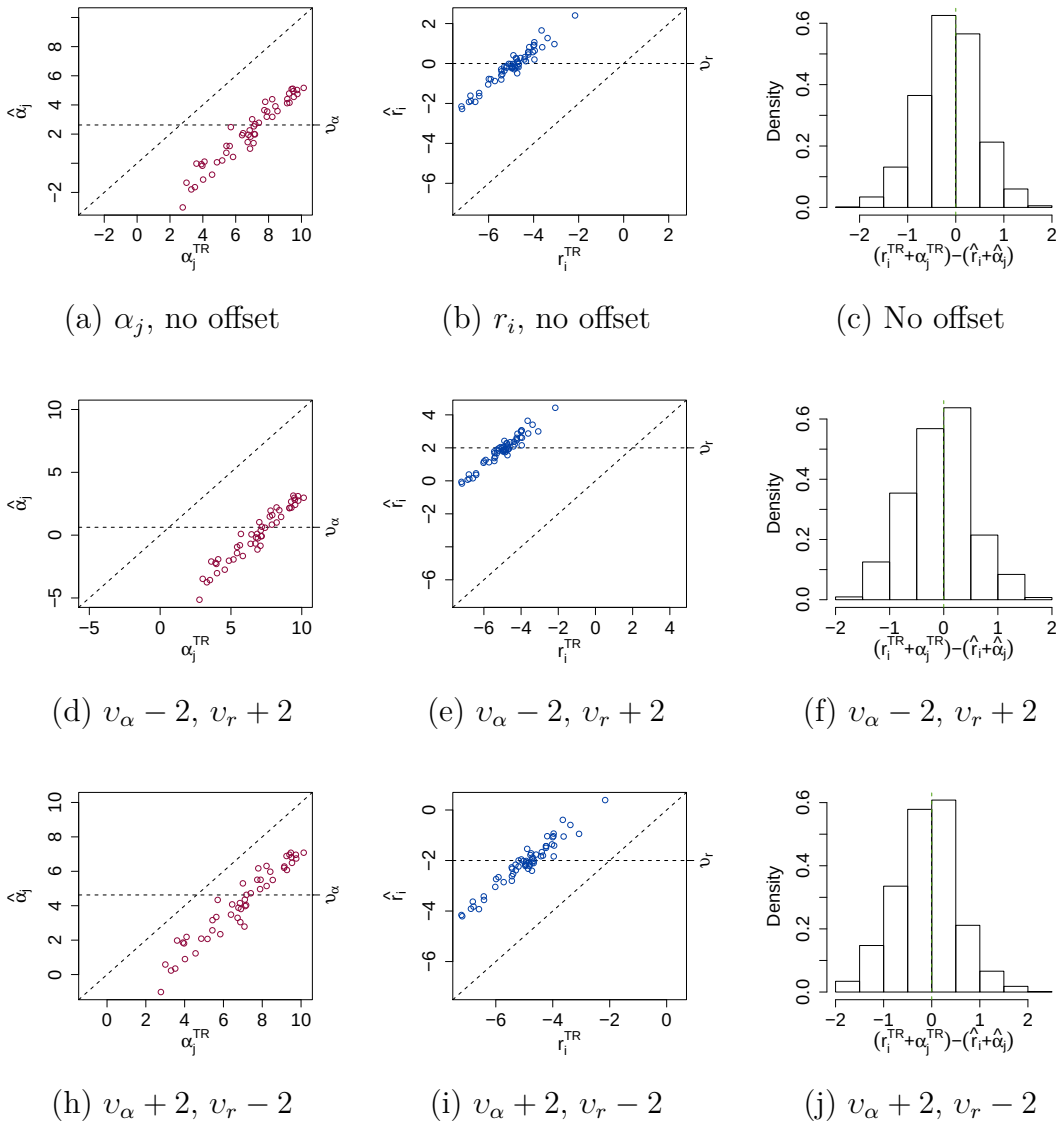
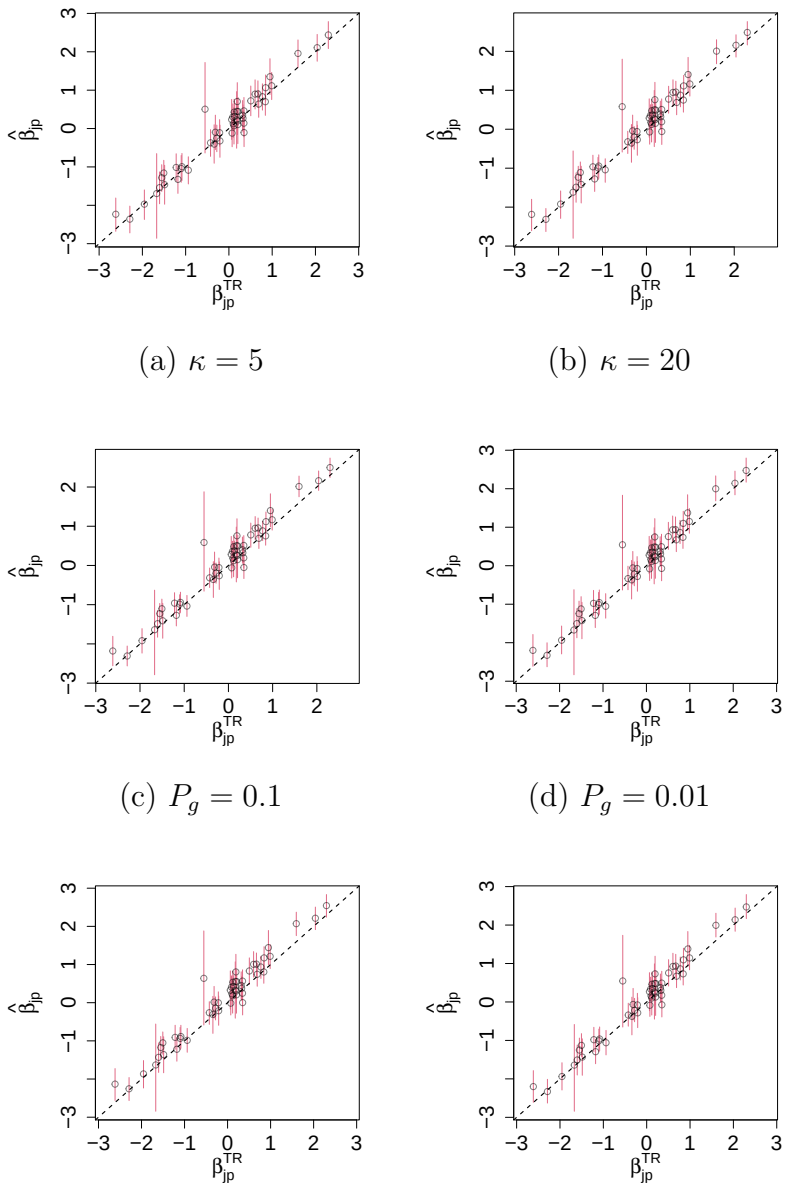


Figure C.2: [Simulation 1] Estimated baseline abundance levels under different specifications for the mean constraints υ_α and υ_r. Columns 1 and 2 show posterior means $\hat{\alpha}_j$ and $\hat{r}_i$ plotted against the simulation truth. Column 3 shows differences of the baseline abundance from the simulation truth compared to the baseline abundance estimated by $\hat{\alpha}_j + \hat{r}_i$. Row 1 is the original mean constraint specification. Row 2 shows results using $\upsilon_\alpha - 2$, $\upsilon_r + 2$. Row 3 shows results using $\upsilon_\alpha + 2$, $\upsilon_r - 2$.

line abundances are, on average, unbiased using these alternative specifications for the mean constraints. Figures showing the edge inclusion probabilities under



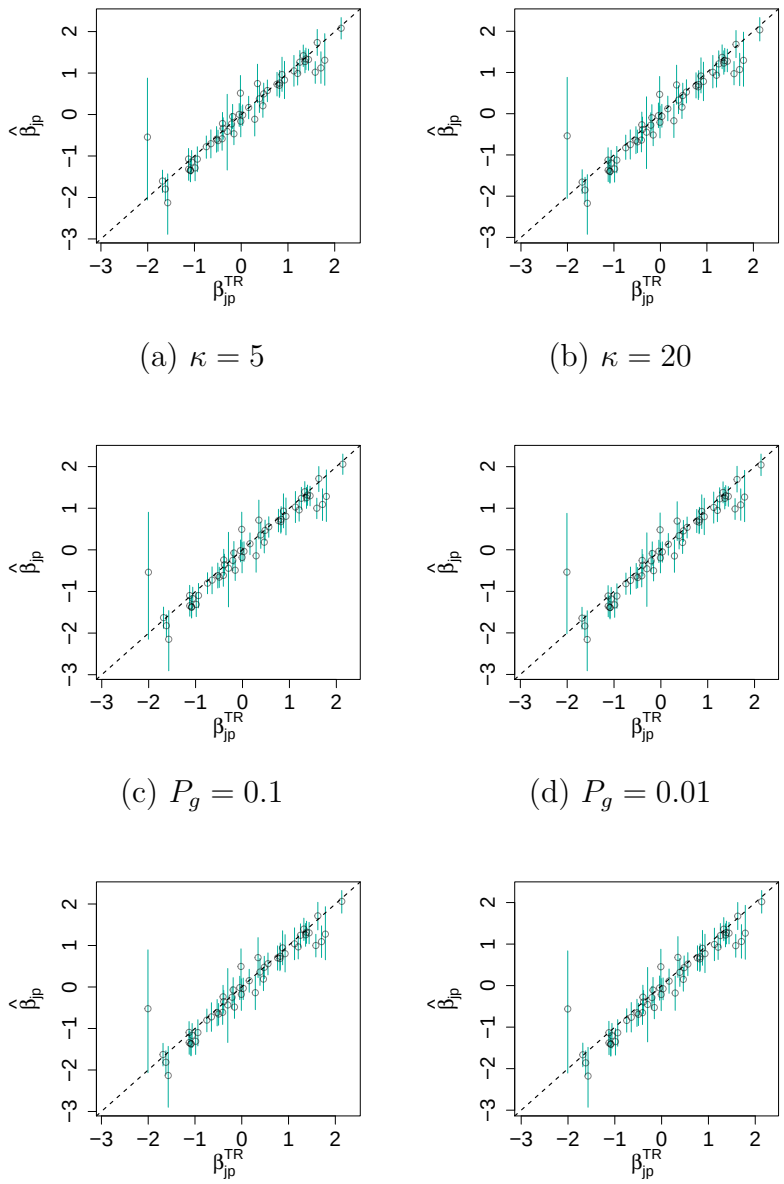
(a) $\kappa = 5$ (b) $\kappa = 20$

(c) $P_g = 0.1$ (d) $P_g = 0.01$

(e) $\upsilon_\alpha - 2$, $\upsilon_r + 2$ (f) $\upsilon_\alpha + 2$, $\upsilon_r - 2$

Figure C.3: [Simulation 1 Sensitivity] Posterior means $\hat{\beta}_{j1}$ and 95% credible intervals plotted against the simulation truth β_{j1}^{TR} using different model specifications.

the alternative specifications are shown in Figure C.6, and figures showing the resulting estimates for the regression coefficients β_{jp} are shown in Figures C.8 and



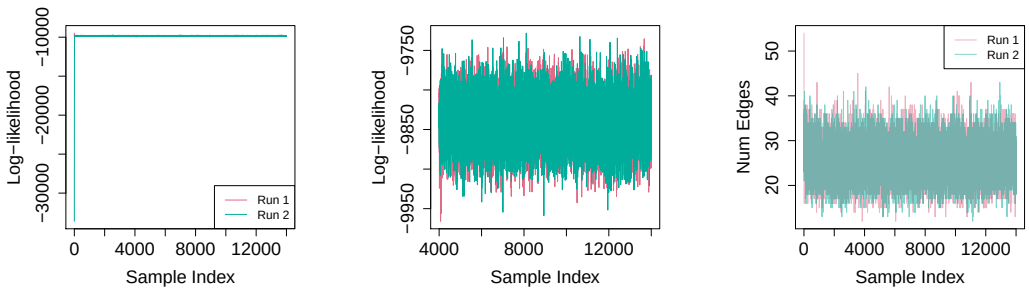
(a) $\kappa = 5$ (b) $\kappa = 20$

(c) $P_g = 0.1$ (d) $P_g = 0.01$

(e) $\upsilon_\alpha - 2$, $\upsilon_r + 2$ (f) $\upsilon_\alpha + 2$, $\upsilon_r - 2$

Figure C.4: [Simulation 1 Sensitivity] Posterior means $\hat{\beta}_{j2}$ and 95% credible intervals plotted against the simulation truth β_{j2}^{TR} using different model specifications.

C.9. The impact of P_g on point estimates of the moral graph produced by including edges with posterior probability > 0.5 can be see in Figure C.7, which shows

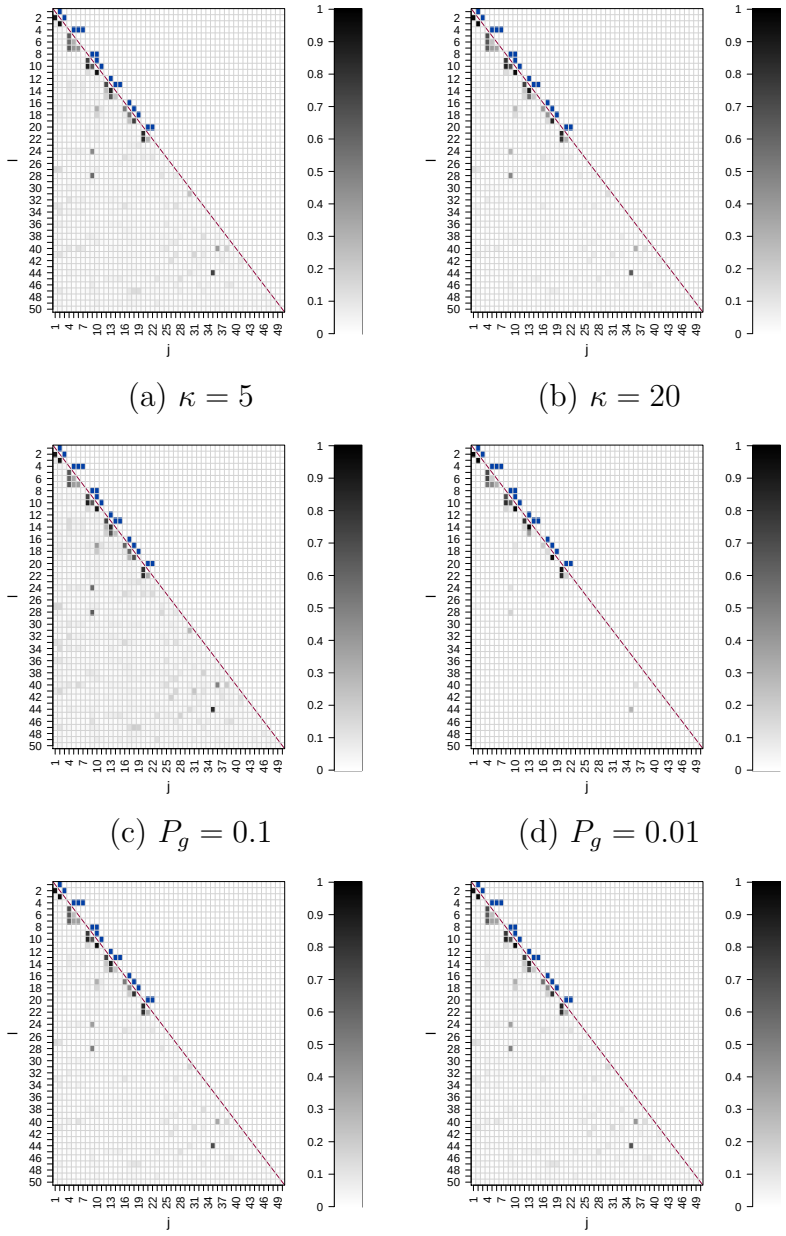


(a) Before burn-in (b) After burn-in (c) Number of edges

Figure C.5: [Simulation 1] (a) and (b) Traceplots of the posterior log-likelihood using two different initializations and random seeds for the MCMC chain. (c) Number of edges in the DAG

the point estimates $\hat{G}^m$ under the alternative specifications. The estimate using the specification with $P_g = 0.1$ contains an additional edge, while using $P_g = 0.01$ produces a slightly sparser graph. Overall, however, we find the estimated graph fairly robust to alternative choices of P_g. We observed very little impact of the model specification on β_{jp}; plots showing their estimates against their true values across the different model specifications are shown in Figures C.8 and C.9.

As in Simulation 1 we assessed the chain's convergence using traceplots and compared the model results to an alternative MCMC chain initialization with a different random seed. We initialized the alternative chain using the same manner as in Simulation 1. Traceplots of the resulting posterior log-likelihoods under the two chains are shown in Figure C.10. The number of edges in the DAG across the MCMC iterations is shown as well. Again we did not find evidence suggesting poor convergence. The alternative chain produced the same point estimate for the moral graph as the original chain, and the regression coefficients were well recovered.

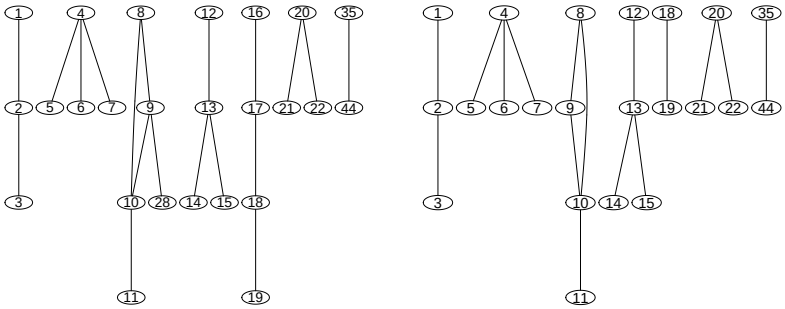


(a) $\kappa = 5$ (b) $\kappa = 20$

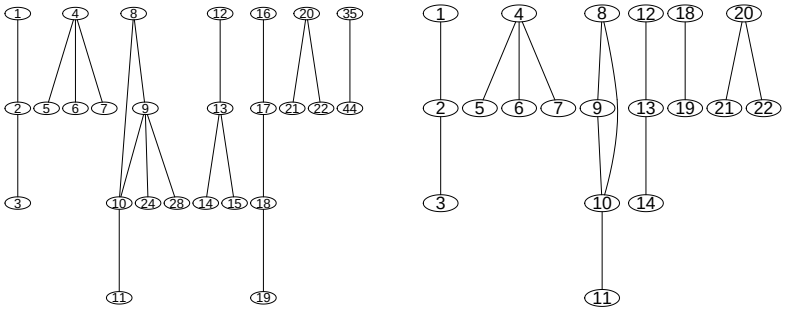
(c) $P_g = 0.1$ (d) $P_g = 0.01$

(e) $\upsilon_\alpha - 2$, $\upsilon_r + 2$ (f) $\upsilon_\alpha + 2$, $\upsilon_r - 2$

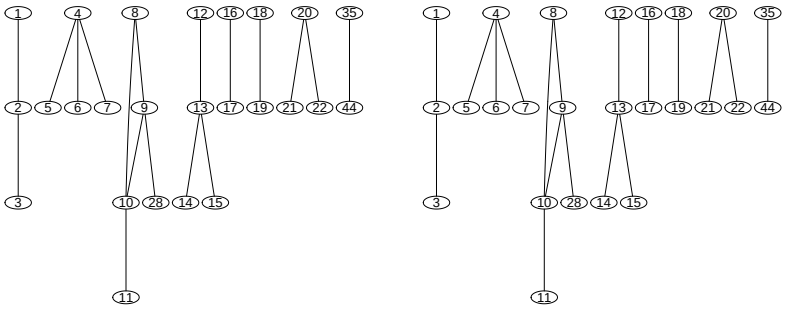
Figure C.6: [Simulation 2 Sensitivity] Lower-diagonals: Posterior probabilities of edge inclusion using different model specifications. Upper-diagonals: Edges of the true moral graph M^{TR}.



(a) $\kappa = 5$ (b) $\kappa = 20$



(c) $P_g = 0.1$ (d) $P_g = 0.01$



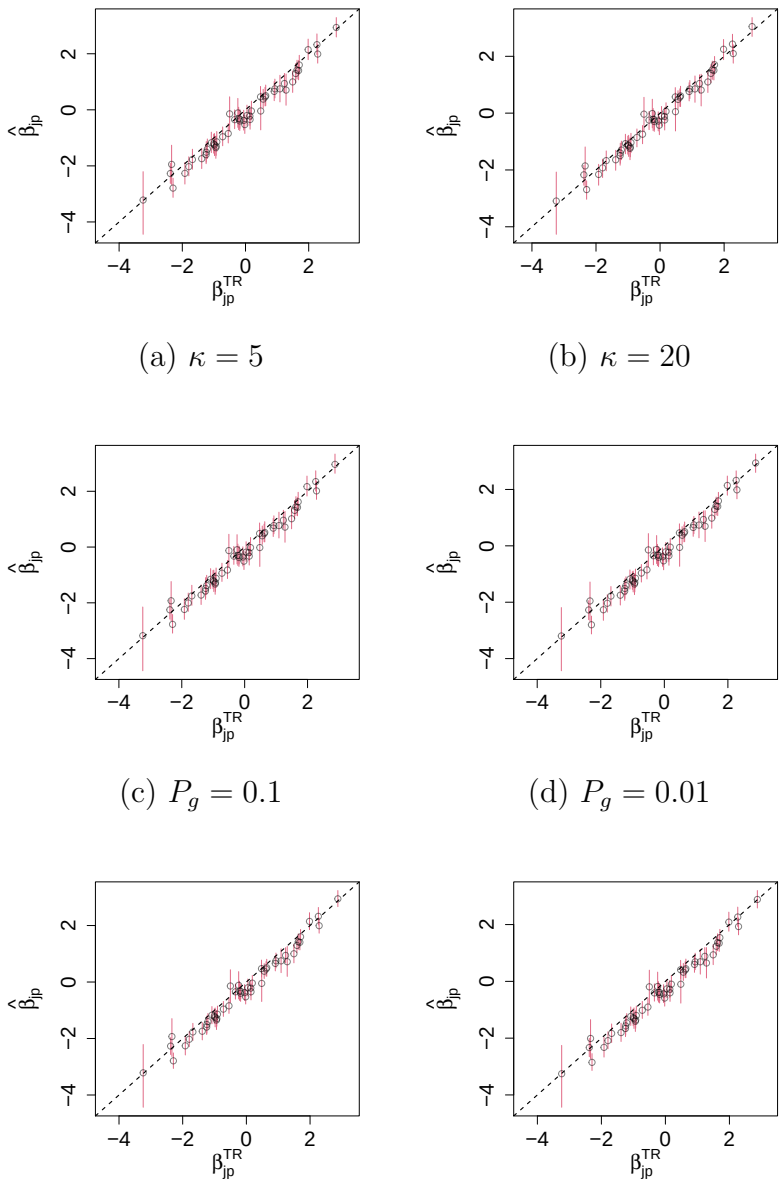
(e) $\upsilon_\alpha - 2$, $\upsilon_r + 2$ (f) $\upsilon_\alpha + 2$, $\upsilon_r - 2$

Figure C.7: [Simulation 2 Sensitivity] Moral graph point estimates $\hat{G}^m$ under different model specifications.

C.3 Additional Chronic Wound Microbiome Results

C.3.1 Sensitivity and Convergence

We conducted sensitivity analysis with respect to the graph inferred by BRM-G when applied to the chronic wound dataset by considering alternative specifica-



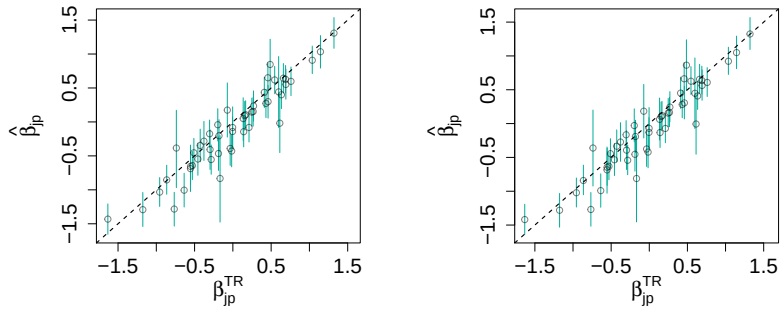
(a) $\kappa = 5$ (b) $\kappa = 20$

(c) $P_g = 0.1$ (d) $P_g = 0.01$

(e) $\upsilon_\alpha - 2$, $\upsilon_r + 2$ (f) $\upsilon_\alpha + 2$, $\upsilon_r - 2$

Figure C.8: [Simulation 2 Sensitivity] Posterior means $\hat{\beta}_{j1}$ and 95% credible intervals plotted against the simulation truth β_{j1}^{TR} using different model specifications.

tions of the fixed scaling factor κ and the edge inclusion probability P_G, as well as to alternative choices for υ_α and υ_r. In the run presented in the main text



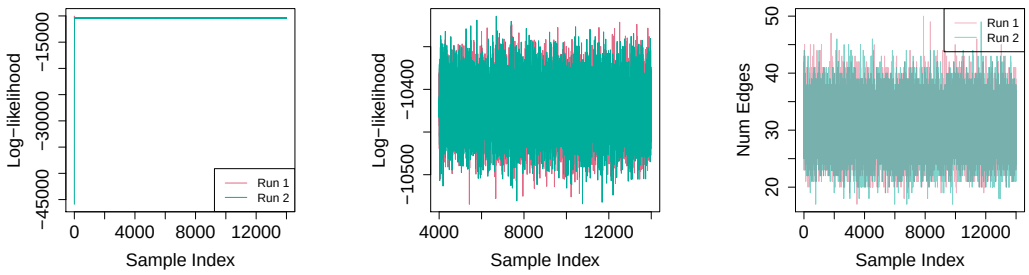
(a) $\kappa = 5$ (b) $\kappa = 20$

(c) $P_g = 0.1$ (d) $P_g = 0.01$

(e) $\upsilon_\alpha - 2$, $\upsilon_r + 2$ (f) $\upsilon_\alpha + 2$, $\upsilon_r - 2$

Figure C.9: [Simulation 2 Sensitivity] Posterior means $\hat{\beta}_{j2}$ and 95% credible intervals plotted against the simulation truth β_{j2}^{TR} using different model specifications.

we set $\kappa = 10$ and $P_g = 0.05$. Here, we consider alternative specifications using $\kappa = 5$ and $\kappa = 20$, and to $P_g = 0.1$ and $P_g = 0.01$. The other hyperparameters

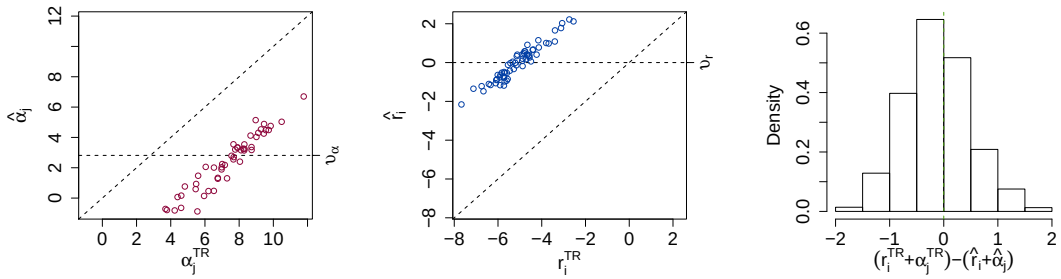


(a) Before burn-in (b) After burn-in (c) Number of edges

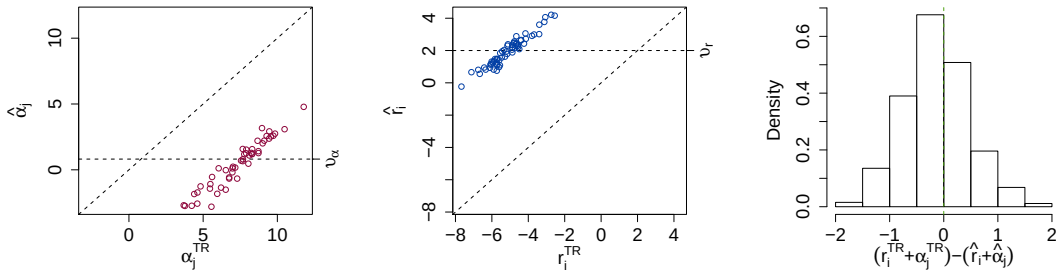
Figure C.10: [Simulation 2] (a) and (b) Traceplots of the posterior log-likelihood using two different initializations and random seeds for the MCMC chain. (c) Number of edges in the DAG

were set to the same values as before. For the mean constraints we used alternative specifications where offsets of ± 2 and ∓ 2 were added to υ_α and υ_r. Point estimates for the moral graph under these alternative specifications are shown in Figure C.12. The results produced by BRM-G are fairly robust across these alternative specifications. Four of the six point estimates for the moral graph exactly match the graph produced by the original specification. The graphs produced using $\kappa = 5$ and $P_g = 0.1$ are the same and are very similar to the other four graphs. These two graphs have two additional edges as compared to the others – one edge between OTU 3 and 4, and one between OTU 4 and 42. The directed edge probabilities and effect directions also appear to be robust to the choice of these hyperparameters. Figures C.13 and C.14 show the posterior probabilities of the directed edges $\bar{a}_{\ell j}$, and the posterior mean estimates of $\gamma_{\ell j}$ given $(\ell \to j)$, for the OTUs having $\bar{m}_{\ell j} > 0.5$ in the original MCMC run in the main text. The results agree well the results obtained using the original specifications of κ, P_g, υ_α and υ_r.

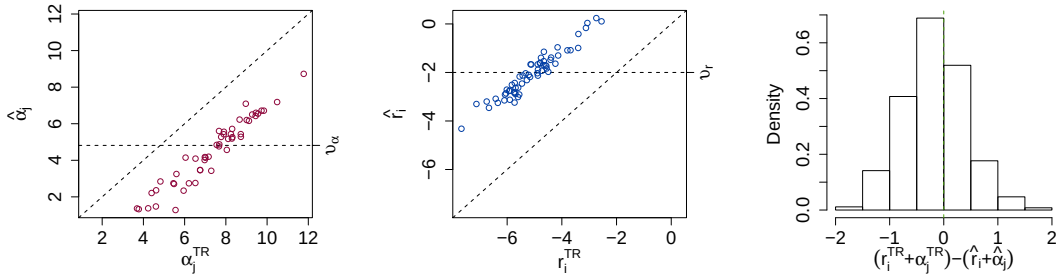
As we did in the simulation studies we assess the chain's convergence by comparing it to another chain using an alternative initialization and different random



(a) α_j, no offset (b) r_i, no offset (c) No offset



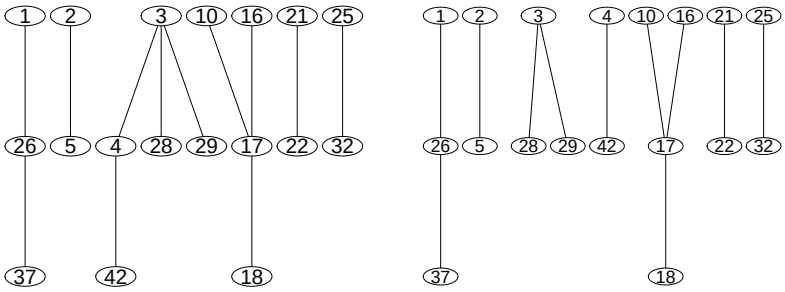
(d) $\upsilon_\alpha - 2$, $\upsilon_r + 2$ (e) $\upsilon_\alpha - 2$, $\upsilon_r + 2$ (f) $\upsilon_\alpha - 2$, $\upsilon_r + 2$



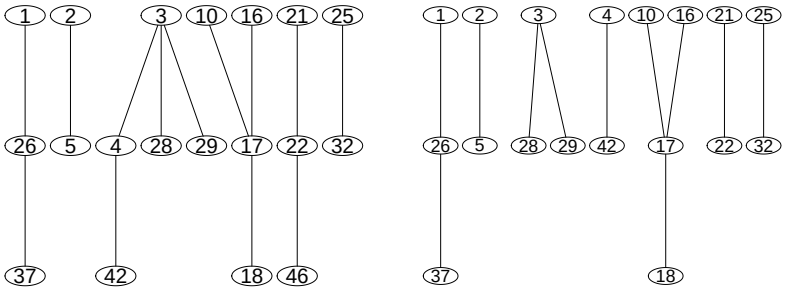
(h) $\upsilon_\alpha + 2$, $\upsilon_r - 2$ (i) $\upsilon_\alpha + 2$, $\upsilon_r - 2$ (j) $\upsilon_\alpha + 2$, $\upsilon_r - 2$

Figure C.11: [Simulation 2] Estimated baseline abundance levels under different specifications for the mean constraints υ_α and υ_r. Columns 1 and 2 show posterior means $\hat{\alpha}_j$ and $\hat{r}_i$ plotted against the simulation truth. Column 3 shows differences of the baseline abundance from the simulation truth compared to the baseline abundance estimated by $\hat{\alpha}_j + \hat{r}_i$. Row 1 is the original mean constraint specification. Row 2 shows results using $\upsilon_\alpha - 2$, $\upsilon_r + 2$. Row 3 shows results using $\upsilon_\alpha + 2$, $\upsilon_r - 2$.

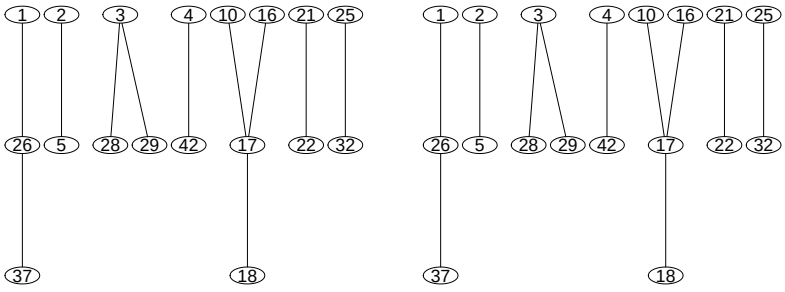
seed. We initialized this alternative chain using the same procedure that we used for Simulation 1 and Simulation 2. Using the alternative chain we obtained the



(a) $\kappa = 5$ (b) $\kappa = 20$



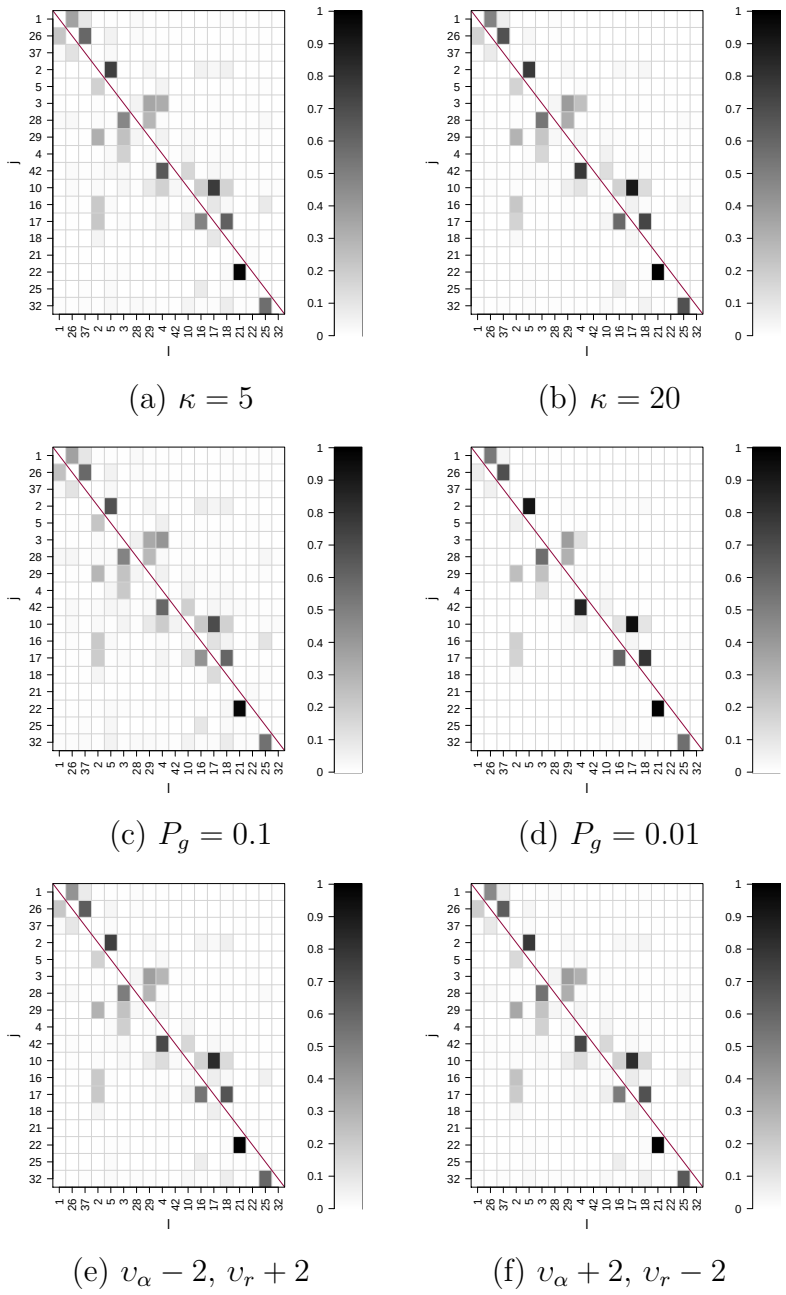
(c) $P_g = 0.1$ (d) $P_g = 0.01$



(e) $\upsilon_\alpha - 2$, $\upsilon_r + 2$ (f) $\upsilon_\alpha + 2$, $\upsilon_r - 2$

Figure C.12: [Chronic Wound Data] Moral graph point estimates $\hat{G}^m$ under different model specifications.

same point estimate for the moral graph as we obtained using the original chain. Traceplots of the posterior log-likelihood and the number of edges in the graph using both the original and alternative chains are shown in Figure C.15. Both chains



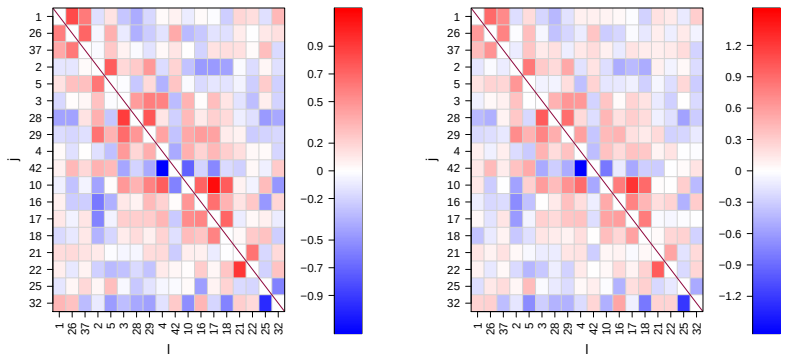
(a) $\kappa = 5$ (b) $\kappa = 20$

(c) $P_g = 0.1$ (d) $P_g = 0.01$

(e) $\upsilon_\alpha - 2,\ \upsilon_r + 2$ (f) $\upsilon_\alpha + 2,\ \upsilon_r - 2$

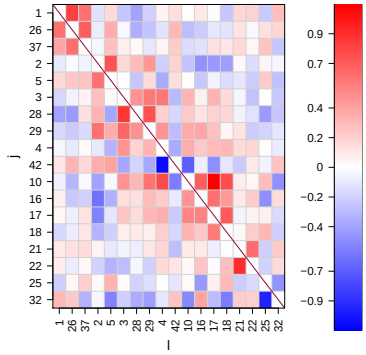
Figure C.13: [Chronic Wound Data] Posterior estimates of $\bar{a}_{lj}$ of the probabilities of including the directed edges $(\ell \leftarrow j)$ for selected OTUs.

quickly converge to similar values despite their very different initializations. As in the simulation studies we did not find evidence suggesting the chains failed to converge.

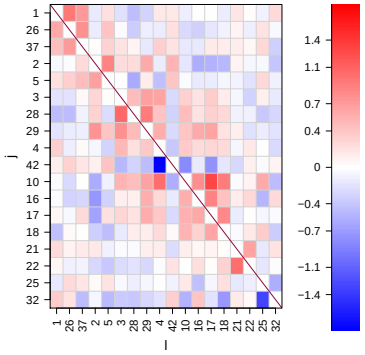


(a) $\kappa = 5$

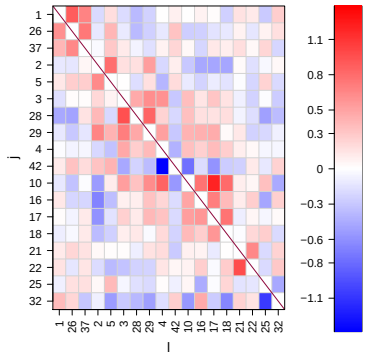
(b) $\kappa = 20$



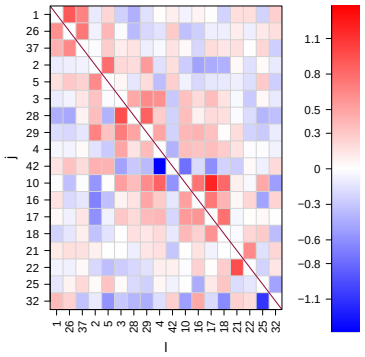
(c) $P_g = 0.1$



(d) $P_g = 0.01$

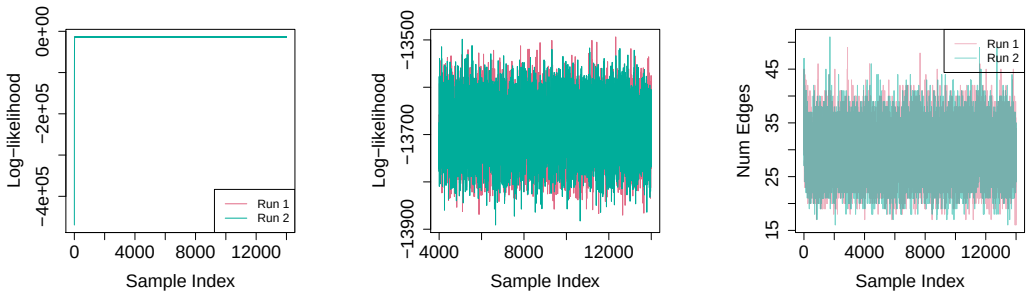


(e) $\upsilon_\alpha - 2,\ \upsilon_r + 2$



(f) $\upsilon_\alpha + 2,\ \upsilon_r - 2$

Figure C.14: [Chronic Wound Data] Posterior mean estimates of $\gamma_{\ell j}$ given $(\ell \to j)$ for selected OTUs



(a) Before burn-in (b) After burn-in (c) Number of edges

Figure C.15: [Chronic Wound Data] (a) and (b) Traceplots of the posterior log-likelihood using two different initializations and random seeds for the MCMC chain. (c) Number of edges in the DAG