Bayesian Statistical Framework for High-Dimensional Count Data and its Application in Microbiome Studies

Bayesian statistical framework for high-dimensional count data and its application in microbiome studies

A dissertation presented

by

Boyu Ren

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

May 2017

Dissertation Advisors: Professor Lorenzo Trippa and Curtis Huttenhower Boyu Ren

Bayesian statistical framework for high-dimensional count data and its application in microbiome studies

Abstract

High-dimensional count data arising from multinomial sampling is ubiquitous in microbiome studies. This dissertation aims to develop flexible Bayesian framework to model high-dimensional count data, which provides reliable and automatic inference for biological questions in microbiome studies.

In Chapter 1, we present a nonparametric Bayesian model for dependent distributions to depict simultaneously multiple species sampling sequences. Our marginal prior for each sampling sequence is a normalized Gamma process and the dependence between the sequences is represented by a low-dimensional latent factors. The resulting posterior samples of model parameters can be used to evaluate uncertainty in analyses routinely applied in microbiome studies such as ordination.

In Chapter 2, we extend the latent factor model in Chapter 1 to enable estimating of effect of covariates. We proved analytically and numerically that this augmented model is identifiable and it separates the effect of covariates and that of latent factors accurately. We provides techniques to transform model parameters to interpretable results. An application of this model on a longitudinal microbiome dataset illustrates the use of this model in microbiome studies.

Chapter 3 focuses more on a bioinformatics tool that simulates realistic microbiome data and benchmarks statistical tools for microbiome studies. We model the count as over-dispersed Poisson outcome by a hierarchical lognormal distribution. We then propose a heuristic algorithm which generates data that resemble real microbiome data. A benchmark of a previously published method illustrates the simulated data provide accurate characterization of the method.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisors, Lorenzo Trippa and Curtis Huttenhower, for being so patient with me and teaching me so much in statistics and biology. I can't imagine how I could have tackled the mysterious Latin names of those bacetria and the even mysterious behaviors of my nonparametric priors without the help from them.

I would also like to thank Sergio Bacallado and Giovanni Parmigiani for their invaluable advices throughout my working on the dissertation. Sergio's ideas never fail to illuminate me and Giovanni's questions never fail to remind me of how little I have reflected on my own projects.

I would also like to thank all my friends in Boston. There are unfortunately not too many people to name here, but it also means everyone I mentioned is so special to me. I want to thank Siyuan, for sharing his time to discuss with me all the random stuff that cross my mind and sporadically, some serious statistics and existential crises. Thank you Jeremiah, for all those enlightening lunches and dinners. Thank you Emma and Himel, for all the small group meetings we suffered and enjoyed togther. Thank you Galeb and I shall return your favor as long as I still know statistics. Thank you Ilaria, for soothing me before my defense and being a great collegue to chat with. Last but not the least, I want to thank all my students, especially all those in the Inference II class, for their understanding on my slow progress in grading and scribbled lab materials.

I would especially like to thank my wife, Yutong He for being an amazing person and partner. You have been incredibly supportive of me and you always inspire me to be a better husband and a better person. The past six years have been extremely special in many ways and that is only possible because of you.

Finally, I would like to thank my parents, for their constant support on all the decisions I made and all the difficulties I faced with in my PhD years. I consider myself very lucky to call you my family.

Bayesian nonparametric ordination for the analysis of microbial communities

Boyu Ren

Department of Biostatistics

Harvard Graduate School of Arts and Sciences

Sergio Bacallado

Department of Pure Mathematics and Mathematical Statistics

University of Cambridge

Stefano Favaro

Departimento di Scienze Economico-Sociali e Matematico-Statistiche

Università di Torino

Susan Holmes

Department of Statistics

Stanford University

Lorenzo Trippa

Department of Biostatistics

Harvard Chan School of Public Health

1.1 Introduction

Next generation sequencing (NGS) has transformed the study of microbial ecology. Through the availability of cheap efficient amplification and sequencing, marker genes such as 16S rRNA are used to provide inventories of bacteria in many different environments. For instance soil and waste water microbiota have been inventoried (DeSantis et al., 2006) as well as the human body (Dethlefsen et al., 2007). NGS also enables researchers to describe the *metagenome* by computing counts of DNA reads and matching them to the genes present in various environments.

Over the last ten years, numerous studies have shown the effects of environmental and clinical factors on the bacterial communities of the human microbiome. These studies enhance our understanding of how the microbiome is involved in obesity (Turnbaugh et al., 2009a), Crohn's disease (Quince et al., 2013), or diabetes (Kostic et al., 2015). Studies are currently underway to improve our understanding of the effects of antibiotics (Dethlefsen and Relman, 2011), pregnancy (DiGiulio et al., 2015), and other perturbations to the human microbiome.

Common microbial ecology pipelines either start by grouping the 16S rRNA sequences into known Operational Taxonomic Units (OTUs) or taxa as done in Caporaso et al. (2010), or denoising and grouping the reads into more refined strains sometimes referred to as oligotypes, phylotypes, or ribosomal variants (RSV) (Rosen et al., 2012; Eren et al., 2014; Callahan et al., 2016). We will call all types of groupings OTUs to maintain consistency. In all cases the data are analyzed in the form of contingency tables of read counts per sample for the different OTUs , as exemplified in Table 2.1. Associated to these contingency tables are clinical and environmental covariates such as time, treatment, and patients' BMI, information collected on the same biological samples or environments. These are sometimes misnamed "metadata"; this contiguous information is usually fundamental in the analyses. The data are often assembled in multi-type structures, for instance `phyloseq` (McMurdie and Holmes, 2013) uses lists (S4 classes) to capture all the different aspects of the data at once.

Currently bioinformaticians and statisticians analyze the preprocessed microbiome data

using linear ordination methods such as Correspondence Analysis (CA), Canonical or Constrained Correspondence Analysis (CCA) , and Multidimensional Scaling (MDS) (Caporaso et al., 2010; Oksanen et al., 2015; McMurdie and Holmes, 2013). Distance-based ordination methods use measures of between-sample or Beta diversity, such as the Unifrac distance (Lozupone and Knight, 2005). These analyses can reveal clustering of biological samples or taxa, or meaningful ecological or clinical gradients in the community structure of the bacteria. Clustering, when it occurs indicates a latent variable which is discrete, whereas gradients correspond to latent continuous variables. Following these exploratory stages, confirmatory analyses can include differential abundance testing (McMurdie and Holmes, 2014a), two-sample tests for Beta diversity scores (Anderson et al., 2006), ANOVA permutation tests in CCA (Oksanen et al., 2015), or tests based on generalized linear models that include adjustment for multiple confounders (Paulson et al., 2013a).

The interaction between these tasks can be problematic. In particular, the uncertainty in the estimation of OTUs' prevalence is often not propagated to subsequent steps (Peiffer et al., 2013). Moreover, unequal sequencing depths generate variations of the number of OTUs with zero counts across biological samples. Finally, the hypotheses tested in the inferential step are often formulated after significant exploration of the data and are sensitive to earlier choices in data preprocessing.

These issues motivate a Bayesian approach that enables us to integrate the steps of the analytical pipeline. Holmes et al. (2012a); La Rosa et al. (2012); Ding and Schloss (2014) have suggested the use of a simple Dirichlet-Multinomial model for these data; however, in those analyses the multinomial probabilities for each biological sample are independent in the prior and posterior, which fails to capture underlying relationships between biological samples. The simple Dirichlet-Multinomial model is also not able to account for strong positive correlations (high co-occurrences (Faust et al., 2012a)) or negative correlations (checker board effect (Koenig et al., 2011)) that can exist between different species (Gorvitovskaia et al., 2016).

We propose a Bayesian procedure, which jointly models the read counts from different OTUs and sample-specific latent multinomial distributions, allowing for correlations be-

tween OTUs. The prior assigned to these multinomial probabilities is highly flexible, such that the analysis learns the dependence structure from the data, rather than constraining it *a priori*. The method can deal with uncertainty coherently, provides model-based visualizations of the data, and is extensible to describe the effects of observed clinical and environmental covariates.

Bayesian analysis with Dirichlet priors is a convenient starting point for microbiome data, since the OTU distributions are inherently discrete. Moreover, Bayesian nonparametric priors for discrete distributions, suitable for an unbounded number of OTUs, have been the topic of intense research in recent years. General classes of priors such as normalized random measures have been developed, and their properties in relation to classical estimators of species diversity are well-understood (Ferguson, 1973; Lijoi and Prünster, 2010). The problem of modeling dependent distributions has also been extensively studied since the proposal of the Dependent Dirichlet Process (MacEachern, 2000) by Müller et al. (2004), Rodríguez et al. (2009), and Griffin et al. (2013)).

In this paper, we try to capture the variation in the composition of microbial communities as a result of a group of unobserved samples' characteristics. With this goal we introduce a model which expresses the dependence between OTUs abundances in different environments through vectors embedded in a low dimensional space. Our model has aspects in common with nonparametric priors for dependent distributions, including a generalized Dirichlet type marginal prior on each distribution, but is also similar in spirit to the multivariate methods currently employed in the microbial ecology community. Namely, it allows us to visualize the relationship between biological samples through low dimensional projections.

The paper is organized as follows. Section 2 describes a prior for dependent microbial distributions, first constructing the marginal prior of a single discrete distribution through manipulation of a Gaussian process and then extending this to multiple correlated distributions. The extension is achieved through a set of continuous latent factors, one for each biological sample, whose prior has been frequently used in Bayesian factor analyses. Section 3 derives an MCMC sampling algorithm for posterior inference and a fast algorithm to estimate biological samples' similarity. Section 4 discusses a method for visualizing the

uncertainty in ordinations through conjoint analysis. Section 5 contains analyses of simulated data, which serve to demonstrate desirable properties of the method, followed by applications to real microbiome data in Section 6. Section 7 discusses potential improvement and concludes. The code for implementing the analyses discussed in this article is included in the Supplementary Materials.

1.2 Probability Model

In Table 2.1, we illustrate an example of a typical OTU table with 10 biological samples, where half are healthy subjects, and half are Inflammatory Bowel disease (IBD) patients. This contingency table is a subset of the data in Morgan et al. (2012a) and records the observed frequencies of five most abundant genus level OTUs in all biological samples based on 16S rRNA sequencing results. Let Z_i be the ith observed OTU (e.g. Z_1 is Bacteroides) and $n_{i,j}$ be the observed frequency of OTU Z_i in biological sample j. As an example, $n_{11} = 1822$ is the observed frequency of Bacteroides in the biological sample Ctrl1. We will denote an OTU table as $(n_{i,j})_{i \leq I, j \leq J}$, where I is the number of observed OTUs and J the number of biological samples.

For the biological sample j, we will assume the vector $(n_{1,j}, \ldots, n_{I,j})$ follows a multinomial distribution, noting that our analysis extends easily to the case in which the total count $\sum_{i=1}^{I} n_{ij}$ is a Poisson random variable.The unobserved multinomial probabilities of OTUs present in biological sample j determine the distribution of the frequencies $n_{i,j}$. These probabilities form a discrete probability measure, which we call a microbial distribution, on the space $\mathcal{Z}$ of all OTUs.

We denote this discrete measure as P^j and $P^j(\{Z_i\})$ gives the probability of sampling Z_i from biological sample j. If we consider all J biological samples, we expect there will be variation in the respective P^j's. This variation usually can be explained by specific characteristics of the biological sample. For instance, in Table 2.1, we can see the empirical multinomial probability of Enterococcus is higher in healthy controls than in IBD patients on average. This variation has been discovered in prior publication (Morgan et al., 2012a) and is attributed to the IBD status. Microbiome studies aim to elucidate the characteristics

Table 1.1: An example of OTU table derived from data published in Morgan et al. (2012a).

OTU	Ctrl1	Ctrl2	Ctrl3	Ctrl4	Ctrl5	IBD1	IBD2	IBD3	IBD4	IBD5
Bacteroides	1822	913	147	2988	4616	172	3516	657	550	1423
Bifidobacterium	0	162	0	0	84	0	85	1927	0	286
Collinsella	1359	0	0	206	0	327	0	0	160	122
Enterococcus	621	0	0	3	40	0	0	0	0	0
Streptococcus	75	139	2161	110	97	1820	85	58	5	294

that explain these types of variations.

Our method focuses on modeling the distributions P^j's and the variations among them. For biological samples labelled in $\mathcal{J} = \{1, \dots, J\}$, we assume they have the same infinite set of OTUs $Z_1, Z_2, \dots \in \mathcal{Z}$. We let the number of OTUs present in a biological sample be infinity to make our model nonparametric in consideration of the fact that there might be an unknown number of OTUs that are not observed in the experiment. We specify the probability mass assigned to a group of OTUs $A \subset \mathcal{Z}$ as

$$\begin{aligned} P^j(A) &= M^j(A)/M^j(\mathcal{Z}), \\ M^j(A) &= \sum_{i=1}^{\infty} \mathbb{I}(Z_i \in A)\sigma_i \langle \mathbf{X}_i, \mathbf{Y}^j \rangle^{+2}, \end{aligned} \tag{1.1}$$

where $\sigma_i \in (0, 1)$, $\mathbf{X}_i, \mathbf{Y}^j \in \mathbb{R}^m$, $\mathbb{I}(\cdot)$ is the indicator function, and $x^+ = x \times \mathbb{I}(x > 0)$. In addition, $\langle \cdot, \cdot \rangle$ is the standard inner product in $\mathbb{R}^m$.

In this model specification, σ_i is related to the average abundance of OTU i across all biological samples. When σ_i is large, the average probability mass assigned to OTU Z_i will also be large. We refer to $\mathbf{X}_i$ and $\mathbf{Y}^j$ as OTU vector and biological sample vectors respectively. The variation of the P^j's is determined by the vectors $\mathbf{Y}^j$, which can be treated as latent characteristics of the biological samples that associate with microbial composition; for example, an unobserved feature of the subject's diet, such as vegetarianism, could affect the abundance of certain OTUs. We assume there are m such characteristics, and the lth component in $\mathbf{Y}^j$ is the measurement of the lth latent characteristic in biological sample j. The vector $\mathbf{X}_i$ denotes the effects of each of the m latent characteristics on the abundance of the OTU Z_i. Therefore $\mathbf{X}_i$ has m entries.

In subsection 1.2.1 we consider a single microbial distribution P^j with fixed parameter $\mathbf{Y}^j$ and define a prior on $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \dots)$ and $(\mathbf{X}_i)_{i \geq 1}$ which makes P^j a Dirichlet process

(Ferguson, 1973). The degree of similarity between the discrete distributions $\{P^j; j \in \mathcal{J}\}$ is summarized by the Gram matrix $(\phi(j, j') = \langle \mathbf{Y}^j, \mathbf{Y}^{j'} \rangle; j, j' \in \mathcal{J})$. Subsection 1.2.2 discusses the interpretation of this matrix. Subsection 1.2.3 proposes a prior for the parameters $\{\mathbf{Y}^j, j \in \mathcal{J}\}$ which has been previously used in Bayesian factor analysis, and which has the effect of shrinking the dimensionality of the Gram matrix $(\phi(j, j'))$ and is used to infer the number of latent characteristics m. The parameters $\{\mathbf{Y}^j, j \in \mathcal{J}\}$ or $(\phi(j, j'))$ can be used to visualize and understand variations of microbial distributions across biological samples.

1.2.1 Construction of a Dirichlet Process

The prior on $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots)$ is the distribution of ordered points ($\sigma_i > \sigma_{i+1}$) in a Poisson process on $(0, 1)$ with intensity

$$\nu(\sigma) = \alpha \sigma^{-1} (1 - \sigma)^{-1/2}, \tag{1.2}$$

where $\alpha > 0$ is a concentration parameter. Denote the index of component of $\mathbf{Y}^j$ and $\mathbf{X}_i$ as l. Fix j, and let $\mathbf{Y}^j = (Y_{l,j}, l \leq m)$ be a fixed vector in $\mathbb{R}^m$ such that $\langle \mathbf{Y}^j, \mathbf{Y}^j \rangle = 1$. We let $\mathbf{X}_i = (X_{l,i}, l \leq m)$ be a random vector for $i = 1, 2, \ldots$ and $X_{l,i}$ be independent and $N(0, 1)$ *a priori* for $l = 1, 2, \ldots, m$ and $i = 1, 2, \ldots$ Finally, let G be a nonatomic probability measure on the measurable space $(\mathcal{Z}, \mathcal{F})$, where $\mathcal{F}$ is the sigma-algebra on $\mathcal{Z}$, and $Z_1, Z_2, \ldots$ is a sequence of independent random variables with distribution G. We claim that the probability distribution P^j defined in Equation (1.1) is a Dirichlet Process with base measure G.

We note that the point process $\boldsymbol{\sigma}$ defines an infinite sequence of positive numbers, the products $\langle \mathbf{X}_i, \mathbf{Y}^j \rangle$, $i = 1, 2, \ldots$, are independent Gaussian $N(0, 1)$ variables, and that the intensity ν satisfies the inequality $\int_0^1 \sigma d\nu < \infty$. These facts directly imply that with probability 1, $0 < M^j(A) < \infty$ when $G(A) > 0$. It also follows that for any sequence of disjoint sets $A_1, A_2, \ldots \in \mathcal{F}$ the corresponding random variables $M^j(A_i)$'s are independent. In different words, M^j is a completely random measure (Kingman, 1967). The marginal Lévy intensity can be factorized as $\mu_M(ds) \times G(dz)$, where

$$\mu_M(ds) \propto \int_0^1 \nu(\sigma) \left(\frac{1}{\sigma}\right)^{1/2} s^{-1/2} \exp\left(-\frac{s}{2\sigma}\right) d\sigma \; ds$$

$$\propto \frac{\exp(-s/2)}{s} ds, \quad \text{for } s \in (0, \infty).$$

The above expression shows that M^j is a Gamma process. We recall that the Lévy intensity of a Gamma process is proportional to the map $s \mapsto \exp(-c \times s) \times s^{-1}$, where c is a positive scale parameter. In Ferguson (1973) it is shown that a Dirichlet process can be defined by normalizing a Gamma process. It directly follows that P^j is a Dirichlet Process with base measure G.

Remark. *Our construction can be extended to a wider class of normalized random measures (James, 2002; Regazzini et al., 2003) by changing the intensity ν that defines the Poisson process $\boldsymbol{\sigma}$. If we set*

$$\nu(\sigma) = \alpha \sigma^{-1-\beta}(1-\sigma)^{-1/2+\beta},$$

$\beta \in [0, 1)$, in our definition of M^j , then the Lévy intensity of the random measure in (1.1) becomes proportional to

$$s^{-1-\beta} \exp(-s/2).$$

In this case the Lévy intensity indicates that M^j is a generalized Gamma process (Brix, 1999). We recall that by normalizing this class one obtains normalized generalized Gamma processes (Lijoi et al., 2007), which include the Dirichlet process and the normalized Inverse Gaussian process (Lijoi et al., 2005) as special cases.

A few comments capture the relation between our definition of $P^j(A)$ in (1.1) and alternative definitions of the Dirichlet Process. If we normalize h independent Gamma$(\alpha/h, 1/2)$ variables, we obtain a vector with Dirichlet$(\alpha/h, \ldots, \alpha/h)$ distribution. To interpret our construction we can note that, when $\alpha/h < 1/2$, each of the Gamma$(\alpha/h, 1/2)$ components can be obtained by multiplying a Beta$(\alpha/h, 1/2 - \alpha/h)$ variable and an independent Gamma$(1/2, 1/2)$. The distribution of the $\langle \mathbf{X}_i, \mathbf{Y}^j \rangle^{+2}$ variables in (1.1) is in fact a mixture with a Gamma$(1/2, 1/2)$ component and a point mass at zero. Finally if we let h increase to ∞, the law of the ordered Beta$(\alpha/h, 1/2 - \alpha/h)$ converges weakly to the law of ordered points of a Poisson point process on $(0, 1)$ with intensity ν (see Supplementary Document S1).

1.2.2 Dependent Dirichlet Processes

We use the representation for Dirichlet processes from Equation (1.1) to define a family of dependent Dirichlet processes labelled by a general index set $\mathcal{J}$. The dependency structure of this family is related to $(\phi(j,j') = \langle \mathbf{Y}^j, \mathbf{Y}^{j'} \rangle)_{j,j' \in \mathcal{J}}$. Geometrically $\phi(j,j')$ is the cosine of the angle between $\mathbf{Y}^j$ and $\mathbf{Y}^{j'}$. The dependent Dirichlet processes is defined by setting

$$P^j(A) = \frac{\sum_i \mathbb{I}(Z_i \in A) \times \sigma_i \langle \mathbf{X}_i, \mathbf{Y}^j \rangle^{+2}}{\sum_i \sigma_i \langle \mathbf{X}_i, \mathbf{Y}^j \rangle^{+2}}, \qquad \forall j \in \mathcal{J}, \tag{1.3}$$

for every $A \in \mathcal{F}$. Here the sequence $(Z_1, Z_2, \ldots)$ and the array $(\mathbf{X}_1, \mathbf{X}_2, \ldots)$, as in Section 1.2.1, contain independent and identically distributed random variables, while $\boldsymbol{\sigma}$ is our Poisson process on the unit interval defined in (1.2). We will use the notation $Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}^j \rangle$. This construction has an interpretable dependency structure between the P^j's that we state in the next proposition.

Proposition 1.1. *There exists a real function $\eta : [0,1] \to [0,1]$ such that the correlation between $P^j(A)$ and $P^{j'}(A)$ is equal to $\eta(\phi(j,j'))$ for every A that satisfies $G(A) > 0$. In different words, the correlation between $P^j(A)$ and $P^{j'}(A)$ does not depend on the specific measurable set A, it is a function of the angle defined by $\mathbf{Y}^j$ and $\mathbf{Y}^{j'}$.*

The proof is in the Supplementary Document S2. The first panel of Figure 1.1 shows a simulation of P^j's. In this figure $\mathcal{J} = \{1,2,3,4\}$. When $\phi(j,j')$, the cosine of the angle between two vectors $\mathbf{Y}^j$ and $\mathbf{Y}^{j'}$, corresponding to distinct biological samples j and j', decreases to -1 the random measures tend to concentrate on two disjoint sets. The second panel shows the function η that maps the $\phi(j,j')$'s into the correlations $\text{corr}(P^j(A), P^{j'}(A)) = \eta(\phi(j,j'))$. As expected the correlation increases with $\phi(j,j')$.

We want to point out that the construction in (1.3) extends easily to the setting where we are given any positive semi-definite kernel $\phi : \mathcal{J} \times \mathcal{J} \to (-1,1)$ capturing the similarity between biological samples labelled by $\mathcal{J}$. Mercer's theorem (Mercer, 1909) guarantees the kernel is represented by the inner product in an $\mathcal{L}^2$ space, whose elements are infinite-dimensional analogues of the vectors $\mathbf{Y}^j$. The analysis presented in this section is unchanged in this general setting.

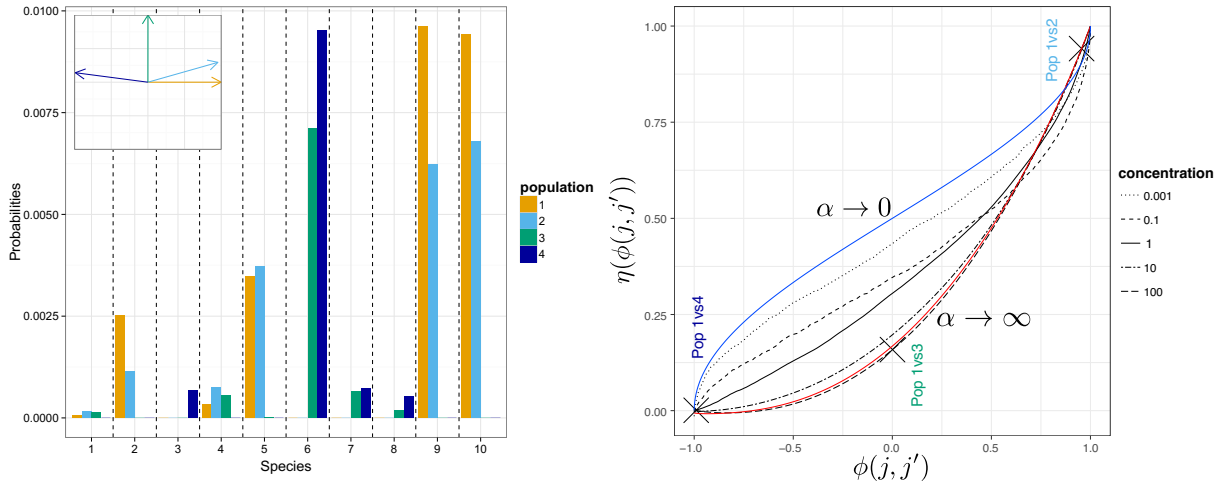


Figure 1.1: (**Left**) Realization of 4 microbial distributions from our dependent Dirichlet processes. We illustrate 10 representative OTUs and set $\alpha = 100$. The miniature figure at the top-left corner shows the relative positions of the four biological sample vectors $\mathbf{Y}^j$. The OTUs are those associated to the 10 largest σ's. As suggested by this panel, the larger the angle between two $\mathbf{Y}^j$'s, the more the corresponding random distributions tend to concentrate on distinct sets. (**Right**) Correlation of two random probability measures when the cosine $\phi(j, j')$ between $\mathbf{Y}^j$ and $\mathbf{Y}^{j'}$ varies from -1 to 1. We consider five different values of the concentration parameter α. The blue and the red curves indicate the limiting cases when $\alpha \to 0$ and $\alpha \to \infty$ respectively. We also mark with crosses the correlations between $P^j(A)$ and $P^{j'}(A)$ for pairs of biological samples j, j' considered in the left panel.

The next proposition provides mild conditions that guarantee a large support for the dependent Dirichlet processes that we defined.

Proposition 1.2. *Consider a collection of probability measures $(F_j, j = 1, \ldots, J)$ on $\mathcal{Z}$ and a positive definite kernel ϕ. Assume that $\mathcal{J} = \{1, \ldots, J\}$ and the support of G coincides with $\mathcal{Z}$. The prior distribution in (1.3) assigns strictly positive probability to the neighborhood $\{(F'_1, \ldots, F'_J) : |\int f_i dF'_j - \int f_i dF_j| < \epsilon, i = 1, \ldots, L,\ j = 1, \ldots, J\}$, where $\epsilon > 0$ and f_i, $i = 1, \ldots, L$, are bounded continuous functions.*

In what follows we will replace the constraint $\langle \mathbf{Y}^j, \mathbf{Y}^j \rangle = 1$ with the requirement $\langle \mathbf{Y}^j, \mathbf{Y}^j \rangle < \infty$. The two constraints are equivalent for our purpose, because we normalize $M^j(\cdot) = \sum_i \mathbb{I}(Z_i \in \cdot) \times \sigma_i \langle \mathbf{X}_i, \mathbf{Y}^j \rangle^{+2}$, and $\langle \mathbf{Y}^j, \mathbf{Y}^j \rangle$ can be viewed as a scale parameter.

1.2.3 Prior on biological sample parameters

This subsection deals with the task of estimating the parameters $\mathbf{Y}^j, j \in \mathcal{J} = \{1, \ldots, J\}$, that capture most of the variability observed when comparing J biological samples with different OTU counts. We define a joint prior on these factors which makes them concentrate on a low dimensional space; equivalently, the prior tends to shrinks the nuclear norm of the Gram matrix $(\phi(j_1, j_2))_{j_1, j_2 \in \mathcal{J}}$. The problem of estimating low dimensional factor loadings or a low-rank covariance matrix is common in Bayesian factor analysis, and the prior defined below has been used in this area of research.

The parameters $\mathbf{Y}^j$ can be interpreted as key characteristics of the biological samples that affect the relative abundance of OTUs. As in factor analysis, it is difficult to interpret these parameters unambiguously (Press and Shigemasu, 1989; Rowe, 2002); however, the angles between their directions have a clear interpretation. As observed in Figure 1.1, if the kernel $\phi(j_1, j_2) \approx \sqrt{\phi(j_1, j_1)\phi(j_2, j_2)}$, the two microbial distributions P^{j_1} and P^{j_2} will be very similar. If $\phi(j_1, j_2) \approx 0$, then there will be little correlation between OTUs' abundances in the two samples. If $\phi(j_1, j_2) \approx -\sqrt{\phi(j_1, j_1)\phi(j_2, j_2)}$, then the two microbial distributions are concentrated on disjoint sets. This interpretation suggests Principal component analysis (PCA) of the Gram matrix $(\phi(j_1, j_2))_{j_1, j_2 \in \mathcal{J}}$ as a useful exploratory data analysis technique.

It is common in factor analysis to restrict the dimensionality of factor loadings. In our model, this is accomplished by assuming $\mathbf{Y}^j$ to be in $\mathbb{R}^m$ and adding an error term ϵ in the definition of $Q_{i,j}$, the OTU-specific latent weights,

$$Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}^j \rangle + \epsilon_{i,j}, \tag{1.4}$$

where the $\epsilon_{i,j}$ are independent standard normal variables. Recall that each sample-specific random distribution P^j is obtained by normalizing the random variables $\sigma_i(Q_{i,j}^+)^2$. If we denote the covariance matrix of $(Q_{i,1}, \ldots, Q_{i,J})$ as $\mathbf{\Sigma}$, this factor model specification indicates $\mathbf{\Sigma} = \mathbf{Y}^\intercal\mathbf{Y} + \mathbf{I}$ conditioning on $\mathbf{Y}$, where $\mathbf{I}$ is the identity matrix and $\mathbf{Y} = (\mathbf{Y}^1, \ldots, \mathbf{Y}^J)$. As a result, the correlation matrix $\mathbf{S}$ induced by $\mathbf{\Sigma}$ only depends on $\mathbf{Y}$.

In most applications the dimensionality m is unknown. Several approaches to estimate m have been proposed (Lopes and West, 2004; Lee and Song, 2002; Lucas et al., 2006; Carvalho et al., 2008; Ando, 2009). However, most of them involve either calculation of Bayes Factors or complex MCMC algorithms. Instead we use a normal shrinkage prior proposed by Bhattacharya and Dunson (2011). This prior includes an infinite sequence of factors ($m = \infty$), but the variability captured by this sequence of latent factors rapidly decreases to zero. A key advantage of the model is that it does not require the user to choose the number of factors. The prior is designed to replace direct selection of m with the shrinkage toward zero of the unnecessary latent factors. In addition, this prior is nearly conjugate, which simplifies computations. The prior is defined as follows,

$$\begin{aligned} &\gamma_l \sim \text{Gamma}(a_l, 1), \qquad \gamma'_{l,j} \sim \text{Gamma}(v/2, v/2), \\ &Y_{l,j}|\boldsymbol{\gamma} \sim N\left(0, (\gamma'_{l,j})^{-1} \prod_{k \le l} \gamma_k^{-1}\right), \ \ l \ge 1, \ j \in \mathcal{J}, \end{aligned} \tag{1.5}$$

where the random variables $\boldsymbol{\gamma} = (\gamma_l, \gamma'_{l,j};\ l, j \ge 1)$ are independent and, conditionally on these variables, the $Y_{l,j}$'s are independent.

When $a_l > 1$, the shrinkage strength *a priori* increases with the index l, and therefore the variability captured by each latent factor tends to decrease with l. We refer to Bhattacharya and Dunson (2011) for a detailed analysis of the prior in (1.5). In practice, the assumption of infinitely many factors is replaced for data analysis and posterior compu-

tations by a finite and sufficiently large number m of factors. The choice of m is based on computational considerations. It is desirable that posterior variability of the last components ($l \sim m$) of the factor model in (1.4) is negligible. This prior model is conditionally conjugate when paired with the dependent Dirichlet processes prior in subsection 1.2.2, a relevant and convenient characteristic for posterior simulations. We summarize the full model with a plate diagram, shown in Figure 1.2.

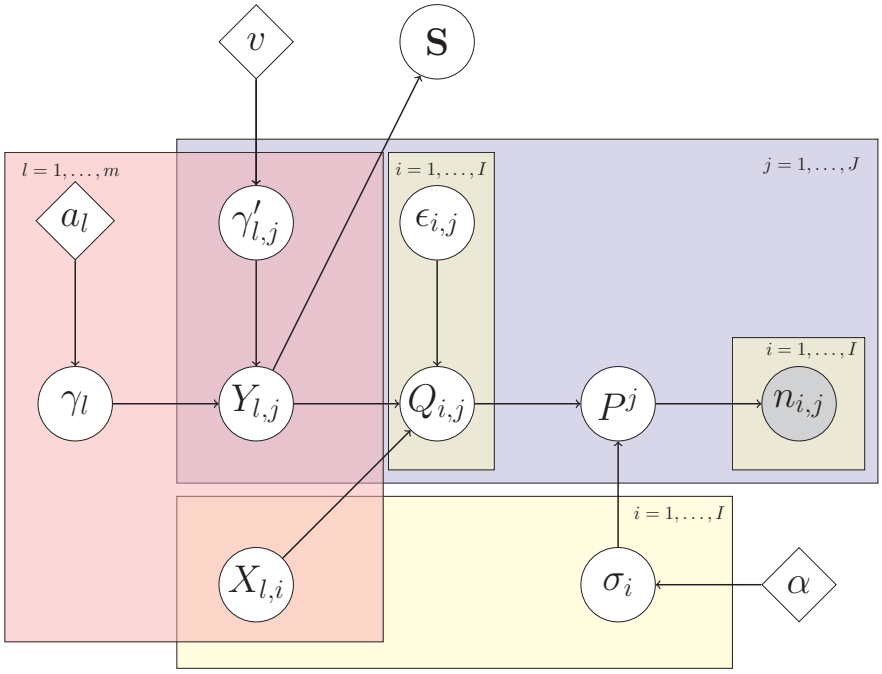


Figure 1.2: Plate diagram. We include the factor model for the latent variables $Q_{i,j}$ as well as the matrix $\mathbf{S}$. Nodes encompassed by a rectangle are defined over the range of indices indicated at the corner of the rectangle, and the connections shown within the rectangle are between nodes with the same index. We use j to index biological samples, i to index microbial species and l to index the components of latent factors.

1.3 Posterior Analysis

Given an exchangeable sequence $W_1, \ldots, W_n$ from $P^j = M^j \times M^j(\mathcal{Z})^{-1}$ as defined in subsection 1.2.1, we can rewrite the likelihood function using variable augmentation as in James et al. (2009),

$$\prod_{i=1}^{n} P^j(\{W_i\}) = \int_0^\infty \frac{\exp[-M^j(\mathcal{Z})\, T] \times T^{n-1}}{\Gamma(n)} \prod_{i=1}^{I} M^j(\{W_i^*\})^{n_i} dT. \tag{1.6}$$

Here $W_1^*, \ldots, W_I^*$ is the list of distinct values in $(W_1, \ldots W_n)$ and $n_1, \ldots, n_I$ are the occurrences in $(W_1, \ldots W_n)$, so that $\sum_{i=1}^{I} n_i = n$. We use expression (1.6) to specify an algorithm that allows us to infer microbial abundances $P^1, \ldots, P^J$ in J biological samples.

We proceed, similarly to Muliere and Tardella (1998) and Ishwaran and James (2001), using truncated versions of the processes in subsection 1.2.2. We replace $\boldsymbol{\sigma} = \{\sigma_i, i \geq 1\}$ with a finite number I of independent Beta$(\epsilon_I, 1/2 - \epsilon_I)$ points in $(0, 1)$. Supplementary Document S1 shows that when I diverges, and $\epsilon_I = \alpha/I$, this finite dimensional version converges weakly to the process in (1.2). Each point σ_i is paired with a multivariate normal $\mathbf{Q}_i = (Q_{i,1}, \ldots, Q_{i,J})$ with mean zero and covariance $\boldsymbol{\Sigma}$. The distribution of $M_{i,j} = \sigma_i (Q_{i,j}^+)^2$ is a mixture of a point mass at zero and a Gamma distribution. In this section $\mathbf{Q}$ and $\boldsymbol{\sigma}$ are finite dimensional, and the normalized vectors P^j, which assign random probabilities to I OTUs in J biological samples, are proportional to $(M_{1,j}, \ldots, M_{I,j})$, $j = 1, \ldots, J$. Note that P^j conditional on $\mathbb{I}(Q_{1,j} > 0), \ldots, \mathbb{I}(Q_{I,j} > 0)$ follows a Dirichlet distribution with parameters proportional to $\mathbb{I}(Q_{1,j} > 0), \ldots, \mathbb{I}(Q_{I,j} > 0)$.

The algorithm is based on iterative sampling from the full conditional distributions. We first provide a description assuming that $\boldsymbol{\Sigma}$ is known. We then extend the description to allow sampling under the shrinkage prior in Section 1.2.3 and to infer $\boldsymbol{\Sigma}$.

With I OTUs and J biological samples, the typical dataset is $\mathbf{n} = (\mathbf{n}_1, \ldots, \mathbf{n}_J)$, where $\mathbf{n}_j = (n_{1,j}, \ldots, n_{I,j})$ and $n_{i,j}$ is the absolute frequency of the ith OTU in the jth biological sample. We use the notation $n^j = \sum_{i=1}^{I} n_{i,j}$, $n_i = \sum_{j=1}^{J} n_{i,j}$, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_I)$, $\mathbf{Y} = (\mathbf{Y}^j, j = 1, \ldots, J)$ and $\mathbf{Q} = (Q_{i,j}, 1 \leq i \leq I, 1 \leq j \leq J)$. By using the representation in (1.6) we introduce the latent random variables $\mathbf{T} = (T_1, \ldots, T_J)$ and rewrite the posterior distribution of $(\boldsymbol{\sigma}, \mathbf{Q})$:

$$p(\boldsymbol{\sigma}, \mathbf{Q}|\mathbf{n}) \propto \left(\prod_{j=1}^{J} \prod_{i=1}^{I} \left(\sigma_i Q_{i,j}^{+2}\right)^{n_{i,j}} \right) \times \prod_{j=1}^{J} \left(\sum_{i=1}^{I} \sigma_i Q_{i,j}^{+2} \right)^{-n^j} \times \pi(\boldsymbol{\sigma}, \mathbf{Q}) \tag{1.7}$$

$$\propto \int \pi(\boldsymbol{\sigma}, \mathbf{Q}) \prod_{j=1}^{J} \left\{ \left(\prod_{i=1}^{I} \left(\sigma_i Q_{i,j}^{+2}\right)^{n_{i,j}} \right) \frac{T_j^{n^j - 1} \exp\left(-T_j \sum_i \sigma_i Q_{i,j}^{+2}\right)}{\Gamma(n^j)} \right\} d\mathbf{T}, \tag{1.8}$$

where π is the prior. In order to obtain approximate $(\boldsymbol{\sigma}, \mathbf{Q})$ sampling we specify a Gibbs

sampler for $(\boldsymbol{\sigma}, \mathbf{Q}, \mathbf{T})$ with target distribution

$$p(\boldsymbol{\sigma}, \mathbf{Q}, \mathbf{T}|\mathbf{n}) \propto \pi(\boldsymbol{\sigma}, \mathbf{Q}) \prod_{j=1}^{J} \left\{ \left(\prod_{i=1}^{I} \left(\sigma_i Q_{i,j}^{+2}\right)^{n_{i,j}} \right) \frac{T_j^{n^j - 1} \exp\left(-T_j \sum_i \sigma_i Q_{i,j}^{+2}\right)}{\Gamma(n^j)} \right\}. \tag{1.9}$$

The sampler iterates the following steps:

[Step 1] Sample T_j independently, one for each biological sample $j = 1, \ldots, J$,

$$T_j|\mathbf{Q}, \boldsymbol{\sigma}, \mathbf{n} \sim \text{Gamma}(n^j, \sum_i \sigma_i Q_{i,j}^{+2}).$$

[Step 2] Sample $\mathbf{Q}_i$ independently, one for each OTU $i = 1, \ldots, I$. The conditional density of $\mathbf{Q}_i = (Q_{i,1} \ldots Q_{i,J})$ given $\boldsymbol{\sigma}, \mathbf{T}, \mathbf{n}$ is log-concave, and the random vectors $\mathbf{Q}_i$, $i = 1, \ldots, I$, given $\boldsymbol{\sigma}, \mathbf{T}, \mathbf{n}$ are conditionally independent.

We simulate, for $j = 1, \ldots, J$, from

$$p(Q_{i,j}|\mathbf{Q}_{i,-j}, \boldsymbol{\sigma}, \mathbf{T}, \mathbf{n}) \propto Q_{i,j}^{+2n_{i,j}} \times \exp\left(-T_j \sigma_i Q_{i,j}^{+2}\right) \times \exp\left(-\frac{(Q_{i,j} - \mu_{i,j})^2}{2s_j^2}\right), \tag{1.10}$$

where $\mathbf{Q}_{i,-j} = (Q_{i,1}, \ldots, Q_{i,j-1}, Q_{i,j+1}, \ldots, Q_{i,J})$, $\mu_{i,j} = E[Q_{i,j}|\mathbf{Q}_{i,-j}]$, $s_j^2 = \text{var}[Q_{i,j}|\mathbf{Q}_{i,-j}]$, with the proviso $0^0 = 1$. Since $\mathbf{Q}_i$ is a multivariate normal, both $\mu_{i,j}$ and s_j have simple closed form expressions.

When $n_{i,j} = 0$ the density in (1.10) reduces to a mixture of truncated normals:

$$(1 - p_1) N(Q_{i,j}; \frac{\mu_{i,j}}{\Delta_{i,j}}, \frac{s_j^2}{\Delta_{i,j}}) I(Q_{i,j} > 0) + p_1 N(Q_{i,j}; \mu_{i,j}, s_j^2) I(Q_{i,j} \leq 0),$$

$$p_1 = \frac{\Phi(0; \mu_{i,j}, s_j^2) N(0; \frac{\mu_{i,j}}{\Delta_{i,j}}, \frac{s_j^2}{\Delta_{i,j}})}{\Phi(0; \mu_{i,j}, s_j^2) N(0; \frac{\mu_{i,j}}{\Delta_{i,j}}, \frac{s_j^2}{\Delta_{i,j}}) + N(0; \mu_{i,j}, s_j^2) \left(1 - \Phi(0; \frac{\mu_{i,j}}{\Delta_{i,j}}, \frac{s_j^2}{\Delta_{i,j}})\right)},$$

and $\Delta_{i,j} = 1 + 2\sigma_i T_j s_j^2$. Here $N(\cdot; \mu, s^2)$ and $\Phi(\cdot; \mu, s^2)$ are the density and cumulative density functions of a normal variable with mean μ and variance s^2.

When $n_{i,j} > 0$ the density $p[Q_{i,j}|\mathbf{Q}_{i,-j}, \boldsymbol{\sigma}, \mathbf{T}, \mathbf{n}]$ remains log-concave, and the support becomes $(0, +\infty)$. We update $Q_{i,j}$ using a Metropolis-Hastings step with proposal identical to the Laplace approximation $N(\widehat{\mu}_{i,j}, \widehat{s}_{i,j}^2)$ of the density in (1.10),

$$\widehat{\mu}_{i,j} = \frac{\mu_{i,j}/s_j^2 + \sqrt{\mu_{i,j}^2/s_j^4 + 8n_{i,j}(2\sigma_i T_j + 1/s_j^2)}}{2(2\sigma_i T_j + 1/s_j^2)}, \quad \widehat{s}_{i,j}^2 = \left(\frac{2n_{i,j}}{\widehat{\mu}_{i,j}^2} + 2T_j\sigma_i + \frac{1}{s_j^2}\right)^{-1}. \tag{1.11}$$

Here $\widehat{\mu}_{i,j}$ maximizes the density (1.10), and $\widehat{s}^2_{i,j}$ is obtained from the second derivative of the log-density at $\widehat{\mu}_{i,j}$. We found the approximation accurate. In Supplementary Document S4 we provide bounds of the total variation distance between the target (1.10) and the approximation (1.11). When $n_{i,j}$ increases, the bound of the total variation decreases to zero. See also Figure S1 in the Supplementary Document.

[Step 3] Sample σ_i independently, one for each OTU $i = 1, \ldots, I$, from the density $p(\sigma_i|\mathbf{Q}, \mathbf{T}, \mathbf{n}) \propto \pi(\sigma_i)\sigma_i^{n_i} \exp(-\sigma_i \sum_{j=1}^{J} T_j Q_{i,j}^{+2})$. The σ_i's are a priori independent Beta$(\alpha/I, 1/2 - \alpha/I)$ variables. We use piecewise constant bounds for $\sigma \to \exp(-\sigma_i \sum_{j=1}^{J} T_j Q_{i,j}^{+2})$, $\sigma \in [0, 1]$ and an accept/reject step to sample from $p(\sigma_i|\mathbf{Q}, \mathbf{T}, \mathbf{n})$.

We now consider inference on $\mathbf{\Sigma}$ using the prior on $\mathbf{Y}$ in subsection 1.2.3. The goal is to generate approximate samples of $\mathbf{Y}$ from the posterior. We exploit the identity of the conditional distributions of $\mathbf{Y}$ given $(\boldsymbol{\sigma}, \mathbf{T}, \mathbf{Q}, \mathbf{n})$ and $\mathbf{Q}$. In order to sample $\mathbf{Y}$ from the posterior we can therefore directly apply the MCMC transitions in Bhattacharya and Dunson (2011), with $\mathbf{Q}$ replacing the observable variables in their work.

1.3.1 Self-consistent estimates of biological samples' similarity

We discuss an EM-type algorithm to estimate the correlation matrix $\mathbf{S}$ of the vectors $(Q_{i,1}, \ldots, Q_{i,J})$, $i = 1, \ldots, I$. Under our construction in subsection 1.2.3, we interpret $\mathbf{S}$ as the normalized version of Gram matrix $(\phi(j, j'))_{j,j' \in \mathcal{J}}$ between biological samples. In this subsection we describe an alternative estimating procedure, distinct from the Gibbs sampler, which does not require tuning of the prior probability model. The algorithm can be used for MCMC initialization and for exploratory data analyses. It assumes that the observed OTU abundances are representative of the microbial distributions, i.e. $P^j = (n_{1,j}/n^j, \ldots, n_{I,j}/n^j)$. Under this assumption, for each biological sample j,

$$\sigma_i Q_{i,j}^{+2} \times \mathbb{I}(n_{i,j} > 0) \propto n_{i,j},\ i = 1, \ldots, I,$$
$$\text{and } Q_{i,j} \leq 0 \text{ when } n_{i,j} = 0. \tag{1.12}$$

For σ_i, $i = 1, \ldots, I$, we use a moment estimate $\widehat{\sigma}_i = (1/J) \sum_j \left(n_{i',j} / \sum_{i \neq i'} n_{i,j} \right)$. The procedure uses these estimates and at iteration $t + 1$ generates the following results:

[Expectation] Impute repeatedly $\mathbf{Q}$, $\ell = 1, \ldots, D$ times, consistently with the constraints

(1.12) and using a $N(0, \boldsymbol{\Sigma}_t)$ joint distribution. Here $\boldsymbol{\Sigma}_t$ is the estimate of $\boldsymbol{\Sigma}$, the covariance matrix of $(Q_{i,1}, \ldots, Q_{i,J})$, after the t-th iteration. For each replicate $\ell = 1, \ldots, D$, we fix $Q_{i,j}^{\ell}$ for all (i, j) pairs with strictly positive $n_{i,j}$ counts at $\sqrt{n_{i,j}/\widehat{\sigma}_i}$ and sample jointly, conditional on these values, negative $Q_{i,j}^{\ell}$ values for the remaining (i, j) pairs with $n_{i,j} = 0$. We use these $Q_{i,j}^{\ell}$ values to approximate $\mathcal{L}(\boldsymbol{\Sigma})$, the full data log-likelihood, our target function as in any other EM algorithm.

[Maximization] Set $\boldsymbol{\Sigma}_{t+1}$ equal to the empirical covariance matrix of the $(Q_{i,1}^{\ell}, \ldots, Q_{i,J}^{\ell})$ vectors, thus maximizing the $\mathcal{L}(\boldsymbol{\Sigma})$ approximation.

We iterate until convergence of $\boldsymbol{\Sigma}_t$. Then, after the last iteration, the inferred covariance matrix of $(Q_{i,1}, \ldots, Q_{i,J})$ directly identifies an estimate of $\mathbf{S}$. We evaluated the algorithm using in-silico datasets from the simulation study in Section 1.5. Overall it generates estimates that are slightly less accurate compared to posterior estimation based on MCMC simulations. We use the datasets considered in Figure 1.3(a), with number of factors fixed at three and n^j at 100,000, for a representative example. In this case the average RV-coefficient between the true $\mathbf{S}$ and the estimated matrix is 0.93 for the EM-type algorithm and 0.95 for posterior simulations. In our work the described procedure reduced the computing time to approximately 10% compared to the Gibbs sampler. More details on this procedure are provided in the Supplementary Document S5.

1.4 Visualizing uncertainty in ordination plots

Ordination methods such as Multidimensional Scaling of ecological distances or Canonical Correspondence Analysis are central in microbiome research. Given posterior samples of the model parameters, we use a procedure to plot credible regions in visualizations such as Fig 1.3(f). The methods that we consider here are all related to PCA and use the normalized Gram matrix $\mathbf{S}$ between biological samples. We recall that in our model $\mathbf{S}$ is the correlation matrix of $(Q_{i,1}, \ldots, Q_{i,J})$. Based on a single posterior instance of $\mathbf{S}$, we can visualize biological samples in a lower dimensional space through PCA, with each biological sample projected once. Naively, one could think that simply overlaying projections of the principal component loadings generated from different posterior samples

of $\mathbf{S}$ on the same graph would show the variability of the projections. However, these super-impositions could be spurious if we carry out PCA for each $\mathbf{S}$ sample separately. One possible problem is principal component (PC) switching, when two PCs have similar eigenvalues. Another problem is the ambiguity of signs in PCA, which would lead to random signs of the loadings that result in symmetric groups of projections of the same biological sample at different sides of the axes. More generally PCA projections from different posterior samples of $\mathbf{S}$ are difficult to compare, as the different lower dimensional spaces are not aligned.

We alternatively identify a consensus lower dimensional space for all posterior samples of $\mathbf{S}$ (Escoufier, 1973; Lavit et al., 1994; Abdi et al., 2005). We list the three main steps used to visualize the variability of $\mathbf{S}$.

1. Identify a normalized Gram matrix $\mathbf{S}_0$ that best summarizes K posterior samples of normalized Gram matrix $\mathbf{S}_1, \ldots, \mathbf{S}_K$. One simple criterion is to minimize L_2 loss element-wise. This leads to $\mathbf{S}_0 = (\sum_i \mathbf{S}_i)/K$. Alternatively, we can define $\mathbf{S}_0$ as the normalized Gram matrix that maximizes similarity with $\mathbf{S}_1, \ldots, \mathbf{S}_K$. One possible similarity metric between two symmetric square matrices $\mathbf{A}$ and $\mathbf{B}$ is the RV-coefficient (Robert and Escoufier, 1976), $\text{RV}(\mathbf{A}, \mathbf{B}) = \text{Tr}(\mathbf{AB})/\sqrt{\text{Tr}(\mathbf{AA})\text{Tr}(\mathbf{BB})}$. We refer to Holmes (2008) for a discussion on RV-coefficients.

2. Identify the lower dimensional consensus space V based on $\mathbf{S}_0$. Assume we want $\dim(V) = 2$; the basis of V will be the orthonormal eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ of $\mathbf{S}_0$ corresponding to the largest eigenvalues λ_1 and λ_2. The configuration of all biological samples in V is visualized by projecting rows of $\mathbf{S}_0$ onto V: $(\boldsymbol{\psi}_1^0, \boldsymbol{\psi}_2^0) = \mathbf{S}_0(\mathbf{v}_1\lambda_1^{-1/2}, \mathbf{v}_2\lambda_2^{-1/2})$. As in a standard PCA, this configuration best approximates the normalized Gram matrix in the L_2 sense: $(\boldsymbol{\psi}_1^0, \boldsymbol{\psi}_2^0) = \text{argmin}_{\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}_2 \rangle = 0} \|\mathbf{S}_0 - (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)'\|^2$.

3. Project the rows of posterior sample $\mathbf{S}_k$ onto V by $(\boldsymbol{\psi}_1^k, \boldsymbol{\psi}_2^k) = \mathbf{S}_k(\mathbf{v}_1\lambda_1^{-1/2}, \mathbf{v}_2\lambda_2^{-1/2})$. Overlaying all the $\boldsymbol{\psi}^k$ displays uncertainty of $\mathbf{S}$ in the same linear subspace. Posterior variability of the biological samples' projections is visualized in V by plotting each row of the matrices $(\boldsymbol{\psi}_1^k, \boldsymbol{\psi}_2^k)$, $k = 1, \ldots, K$, in the same figure. A contour plot

is produced for each biological sample (see for example Fig 1.3(f)) to facilitate visualization of the posterior variability of its position in the consensus space V.

1.5 Simulation Study

In this section, we evaluate the procedure described in Section 1.3 and explore whether the shrinkage prior allows us to infer the number of factors and the normalized Gram matrix between biological samples $\mathbf{S}$. We also consider the estimates $E(P^j|\mathbf{n})$ obtained with our joint model, one for each biological sample j, and compare their precision with the empirical estimator. Throughout the section, we assumed the number of factors is $m = 10$ when running the posterior simulations.

We first defined a scenario with distributions P^j generated from the prior (1.1), with $I = 68$ OTUs and $J = 22$ biological samples. The true number of factors is m_0, and for biological samples $j = 1, \ldots, m_0/2$, the vector $\mathbf{Y}^j = (Y_{l,j}, 1 \leq l \leq m_0)$ has elements $l = m_0/2+1, \ldots, m_0$ equal to zero, while symmetrically, for $j = J/2+1, \ldots, J$, the vectors $\mathbf{Y}^j$ have the elements $l = 1, \ldots, m_0/2$ equal to zero. The underlying normalized Gram matrix $\mathbf{S}$ is therefore block-diagonal. After generating the distributions P^j, we sampled with fixed total counts (n^j) per biological sample n^j= 1,000. We produced 50 replicates with m_0 =3, 6, and 9. In our simulations the non-zero components $Y_{l,j}$'s are independent standard normal.

We use PCA-type summaries for the posterior samples of $\mathbf{Y}$ generated from $p(\mathbf{Y}|\mathbf{n})$. Computations are based on the $J \times J$ normalized Gram matrix $\mathbf{S}$. At each MCMC iteration we generate approximate samples $\mathbf{Y}$ from the posterior, compute $\mathbf{S}$ by normalizing the Gram matrix $\mathbf{Y}'\mathbf{Y}$, and operate standard spectral decomposition on $\mathbf{S}$. This allows us to estimate the ranked eigenvalues, i.e. the principal components' variance of our $\mathbf{Q}$ latent vectors (after normalization), by averaging over the MCMC iterations. Figure 1.3(a) shows the variability captured by the first 10 principal components, with the box-plots illustrating posterior means' variability across our 50 replicates. The proportion of variability associated to each *principal component* decreases rapidly after the *true* number of factors $m_0 = 3, 6, 9$. This suggests that the shrinkage model (Bhattacharya and Dunson,

2011) tends to produce posterior distributions for our $\mathbf{Y}$ latent variables that concentrates around a linear subspace.

Figure 1.3(c) illustrates the accuracy of the estimated normalized Gram matrix $\widehat{\mathbf{S}}$ with n^j equal to 1,000, 10,000, and 100,000. We estimated the unknown $J \times J$ normalized Gram matrix $\mathbf{S}$ with the posterior mean of the normalized Gram matrix, which we approximate by averaging over MCMC iterations. We summarized the accuracy using the RV coefficient between $\widehat{\mathbf{S}}$ and $\mathbf{S}$, see Robert and Escoufier (1976) for a discussion on this metric. The box-plots illustrate variability of estimates' accuracy across 50 simulation replicates. As expected, when the total counts per sample increases from 10,000 to 100,000, we only observe limited gain in accuracy. Indeed the overall number of observed OTUs with positive counts per biological sample remains comparable, with expected values equal to 30 and 33 when the total counts per biological sample are fixed at 10,000 and 100,000 respectively. We also note that when m_0 increases, the accuracy decreases.

We investigate interpretability of our model by using distributions P^j generated from a probability model that slightly differs from the prior. More precisely, the ith random weight in P^j, conditionally on $\mathbf{Y}$ and $\mathbf{X}$, is defined proportional to a monotone function of $\langle \mathbf{X}_i, \mathbf{Y}^j \rangle^+$. We considered for example

$$P^j(A) = \frac{\sum_i \sigma_i \langle \mathbf{X}_i, \mathbf{Y}^j \rangle^{+a} I_{Z_i}(A)}{\sum_i \sigma_i \langle \mathbf{X}_i, \mathbf{Y}^j \rangle^{+a}}, \quad a > 0. \tag{1.13}$$

When the monotone function is quadratic the probability model becomes identical to our prior. In Figure 1.3(b) and Figure 1.3(d) we used model (1.13) with $a = 1$ to generate datasets. We repeated the same simulation study summarized in the previous paragraphs.

We evaluated the effectiveness of borrowing information across biological samples for estimating the vectors P^j. The accuracy metric that we used is the total variation distance. We compared the Bayesian estimator $E(P^j|\mathbf{n})$ and the empirical estimator $\tilde{P}^j$ which assigns mass $n_{i,j}/n^j$ to the i^{th} OTUs. The advantage of pooling information varies with the similarity between biological samples. To reflect this, we generated P^j with non-zero components of $\mathbf{Y}$ sampled from a zero mean multivariate normal with $\text{cov}(Y_{l,j}, Y_{l,j'})$ equal to θ. We considered the case when P^j is generated either from our prior or model (1.13)

with $a = 0.5, 1, 3$. In addition, we considered $\theta = 0.5, 0.75, 0.95$, $I = 68, J = 22$, and $m_0 = 3$, while n^j varies from 10 to 100.

The results are summarized in Figure 1.3(e) which shows the average difference in total variation, contrasting the Bayesian and empirical estimators. The results, both when the model is correctly specified, and when mis-specified, quantify the advantages in using a joint Bayesian model.

We complete this section with one illustration of the method in Section 1.4. We simulate a dataset with two clusters by generating $Y_{l,j}$ for $l = 1, \ldots, m_0$ from $N(-3, 1)$ when $j = 1, \ldots, J/2$ and from $N(3, 1)$ when $j = J/2 + 1, \ldots, J$. All $Y_{l,j}$ are different from zero. We expected a low n^j to be sufficient for detecting the clusters. We sampled P^j from the prior and set $J = 22$, $I = 68$, $m_0 = 3$, and $n^j = 100$. The PC plot and the biological sample specific credible regions are shown in Figure 1.3(f). In the PC plot the two clusters are illustrated with different colors. In this simulation exercise the posterior credible regions leave little ambiguity both on the presence of clusters and also on samples-specific cluster membership. To compare this with the Principal Coordinates Analysis (PCoA) method used in microbiome studies, we plot the ordination results using PCoA based on the Bray-Curtis dissimilarity matrix derived from the empirical microbial distributions (See Figure S3). We can see that the PCoA point estimate is similar to the centroids identified by the proposed Bayesian ordination method.

1.6 Application to microbiome datasets

In this section, we apply our Bayesian analysis to two microbiome datasets. We show that our method gives results that are consistent with previous studies, and we show our novel visualization of uncertainty in ordination plots. We start with the Global Patterns data (Caporaso et al., 2011) where human-derived and environmental biological samples are included. We then considered data on the vaginal microbiome (Ravel et al., 2011a).

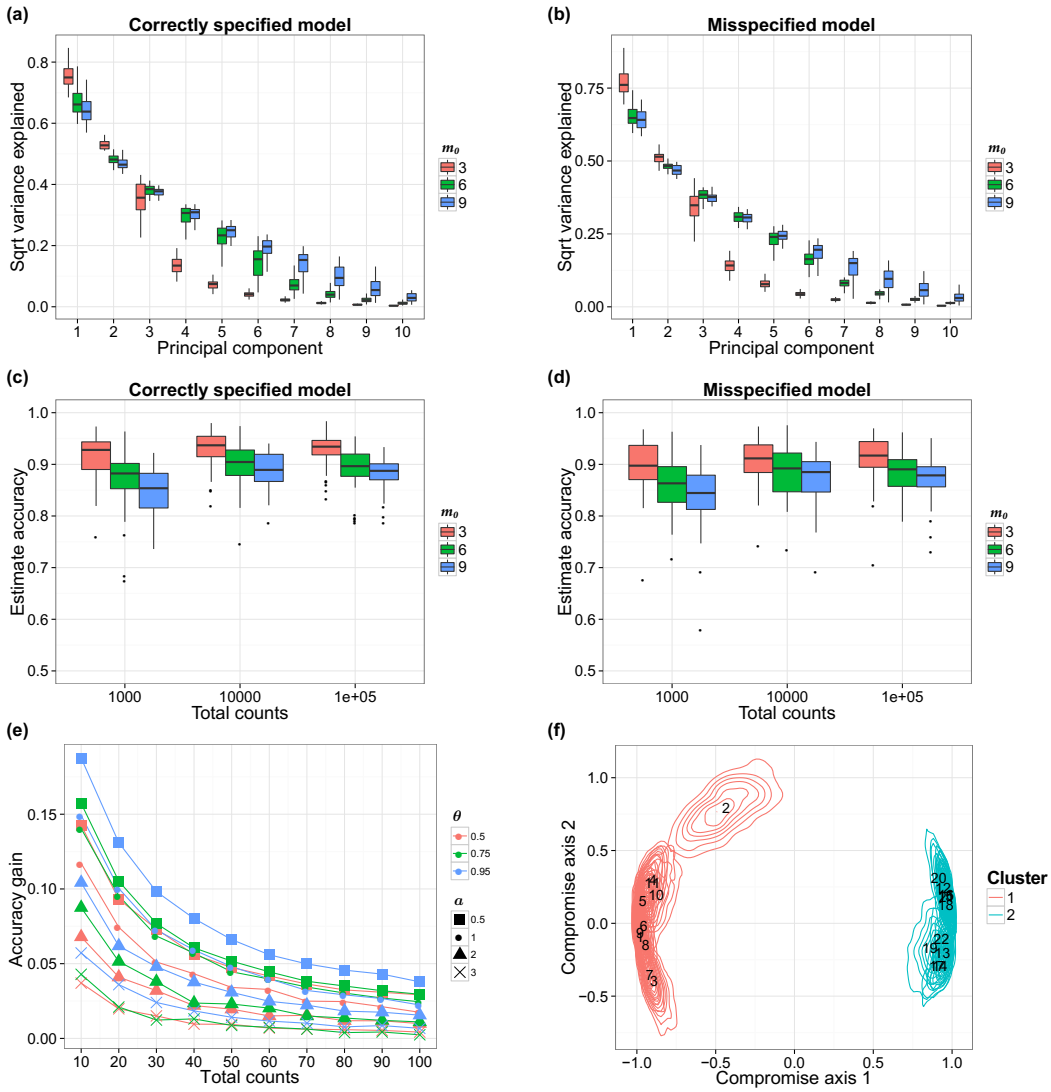


Figure 1.3: (**a-b**) Estimated proportion of variability captured by the first 10 PCs. Each box-plot here shows the variability of the estimated proportion across 50 simulation replicates. We show the results when the data are generated from the prior (Panel **a**) and from the model in (1.13) with $a = 1$ (Panel **b**). (**c-d**) Accuracy of the correlation matrix estimates $\widehat{\mathbf{S}}$. The box-plots show the variability of the accuracy in 50 simulation replicates, with data generated from the prior (Panel **c**) and from model (1.13) with $a = 1$ (Panel **d**). We vary the true number of factors m_0 (colors) and n^j and show the corresponding accuracy variations. (**e**) Comparison between Bayesian estimates of the underlying microbial distributions P^j and the empirical estimates. We consider the average total variation difference, averaging across all J biological samples. Each curve shows the relationship between n^j and average accuracy gain. We set $m_0 = 3$ and the parameter a varies from 0.5 to 3 (shapes). The similarity parameter θ is equal to 0.5, 0.75 or 0.95 (colors). (**f**) PCoA plot with confidence regions. We visualize the confidence regions using the method in Section 1.4. Each contour illustrates the uncertainty of a single biological sample's position. Colors indicate cluster membership and annotated numbers are biological samples' IDs.

1.6.1 Global Patterns dataset

The Global Patterns dataset includes 26 biological samples derived from both human and environmental specimens. There are a total of 19,216 OTUs, and the average total counts per biological sample is larger than 100,000. We collapsed all taxa OTUs to the genus level—a standard operation in microbiome studies—and yielded 996 distinct genera. We treated these genera as OTUs' and fit our model to this collapsed dataset. We ran one MCMC chain for 50,000 iterations and recorded posterior samples every 10 iterations.

We first performed a cluster analysis of biological samples based on their microbial compositions. For each posterior sample of the model parameters, we computed P^j for $j = 1, \ldots, J$ and calculated the Bray-Curtis dissimilarity matrix between biological samples. We then clustered the biological samples using this dissimilarity matrix with Partitioning Among Medoids (PAM) (Tibshirani et al., 2002). By averaging over the MCMC iterations for the clustering results from each dissimilarity matrix, we obtained the posterior probability of two biological samples being clustered together. Figure 1.4(a) illustrates the clustering probabilities. We can see that biological samples belonging to a specific specimen type are tightly clustered together while different specimens tend to define separate clusters. This is consistent with the conclusion in Caporaso et al. (2011), where the authors suggest, that within specimen microbiome variations are limited when compared to variations across specimen types. We also observed that biological samples from the skin are clustered with those from the tongue. This is to some extent an expected result, because both specimens are derived from humans, and because the skin microbiome has often OTUs frequencies comparable to other body sites (Grice and Segre, 2011).

We then visualized the biological samples using ordination plots and applying the method described in Section 1.4. We fixed the dimension of the consensus space V at three. We plotted all biological samples' projections onto V along with contours to visualize their posterior variability. The results are shown in Figure 1.4(b-d). We observe a clear separation between human-derived (tongue, skin, and feces) biological samples and biological samples from free environments. This separation is mostly identified by the first two compromise axes. The third axis defines a saline/non-saline samples separa-

tion. Biological samples derived from saline environment (e.g. Ocean) are well separated when projected on this axis from those derived from non-saline environment (e.g. Creek freshwater). We observed small 95% credible regions for all biological samples projections. This low level of uncertainty captured by the small credible regions in Figure 1.4(b-d) is mainly explained by the large total counts n^j for all biological samples. Finally, to compare the ordination results to those given by standard methods used in microbiome studies, we generated ordination results using PCoA. Figure S4 shows that the relative positions of different types of biological samples in PCoA plots and in the Bayesian ordination plots are similar.

1.6.2 The Vaginal Microbiome

We also consider a dataset previously presented in Ravel et al. (2011a) which contains a larger number of biological samples (900) and a simpler bacterial community structure. These biological samples are derived from 54 healthy women. Multiple biological samples are taken from each individual, ranging from one to 32 biological samples per individual. Each woman has been classified, before our microbiome sequencing data were generated, into vaginal community state subtypes (CST). This dataset contains only species level taxonomic information, and we filtered OTUs by occurrence. We only retain species with more than five reads in at least 10% of biological samples. This filtering resulted in 31 distinct OTUs. We ran one MCMC chain with 50,000 iterations.

We performed the same analyses as in the previous subsection. The results are shown in Figure 1.5. Clustering probabilities indicate strong within CST similarity (panel a). There is one exception, CST IV-A samples, in some cases, presenting low levels of similarities when compared to each other and tend to cluster with CST I, CST III, and CST IV-B samples. This is because CST IV-A is characterized as a highly heterogeneous subtype (Ravel et al., 2011a). The ordination plots are consistent with the discoveries in Ravel et al. (2011a). A tetrahedron shape is recovered, and CST I, II, III, IV-B occupy the four vertices. CST II is well separated from other CSTs by the third axis. This pattern is similar to the one observed in the plots generated using PCoA (Figure S5). We also observed a sub-clustering in CST II which has not been detected and discussed in Ravel et al. (2011a).

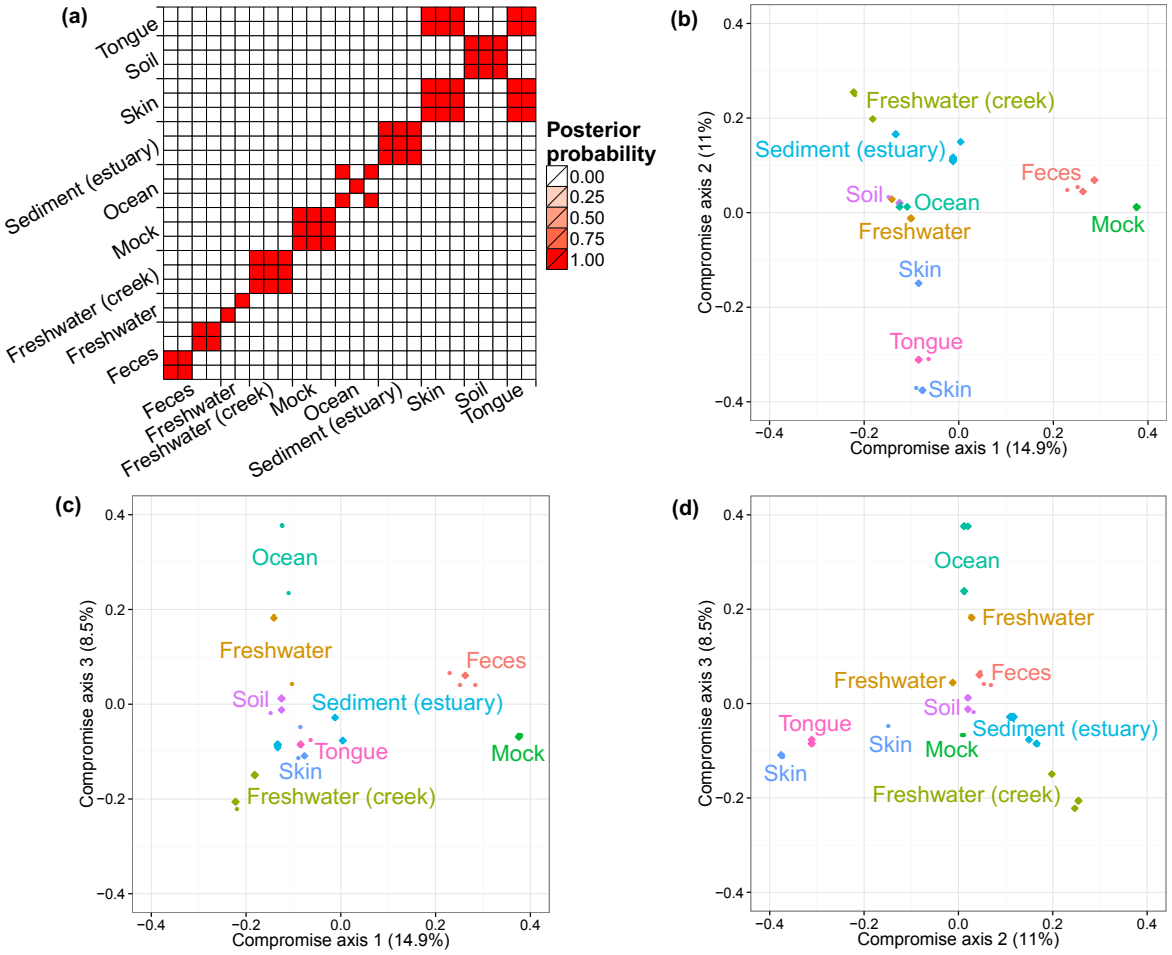


Figure 1.4: (**a**) Posterior Probability of each pair of biological samples (j, j') being clustered together. The labels on axes indicate the environment of origin for each biological sample. (**b-d**) Ordination plots of biological samples and 95% posterior credible regions. We illustrate the first three compromise axes with three panels. Panel (b) plots projections on the first and second axes. Panel (c) plots projections on the first and third axes. Panel (d) plots projections on the second and third axes. The percentages on the three axes are the ratios of the corresponding S_0 eigenvalues and the trace of the matrix. The credible regions for some biological samples are so small that appears as single points. Colors and annotated text indicate the environments.

This difference in the results can be due to distinct clustering metrics in the analyses.
Note that there are two biological samples with large credible regions, indicating high uncertainty of the corresponding positions. This uncertainty propagates on their cluster membership. Both biological samples have small total counts compared to the others. The lack of precision when using biological samples with small sequencing depth leads to high uncertainty in ordination and classification. It is therefore important to account for uncertainty in the validation of subgroups biological differences—in our case CST subtypes—based on microbiome profiling. Our example suggests also the importance of uncertainty summaries when microbiome profiles are used to classify samples. Uncertainty summaries allow us to retain all samples, including those with low counts, without the risk of overinterpreting the estimated locations and projections. This also argues for the retention of raw counts in microbiome studies (McMurdie and Holmes, 2014a). By using raw counts, we can evaluate the uncertainty of our estimates and exploit the information and statistical power carried by the full dataset; whereas if we downsample the data we lose information and increase uncertainty on the projections.
It is ubiquitous to have biological samples with relevant differences in their total counts, and in some cases the number of OTUs and the total number of reads can be comparable. In this cases, the empirical estimates of microbial distributions are not reliable, and an assessment of the uncertainty is necessary for downstream analyses. The two biological samples with low total counts in the vaginal microbiome dataset are examples. For biological samples with a scarce amount of data our model provides measures of uncertainty and allows uncertainty visualizations with ordination plots.

1.7 Conclusion

We propose a joint model for multinomial sampling of OTUs in multiple biological samples. We apply a prior from Bayesian factor analysis to estimate the similarity between biological samples, which is summarized by a Gram matrix. Simulation studies give evidence that this parameter is recovered by the Bayes estimate, and in particular, the inherent dimensionality of the latent factors is effectively learned from the data. The simu-

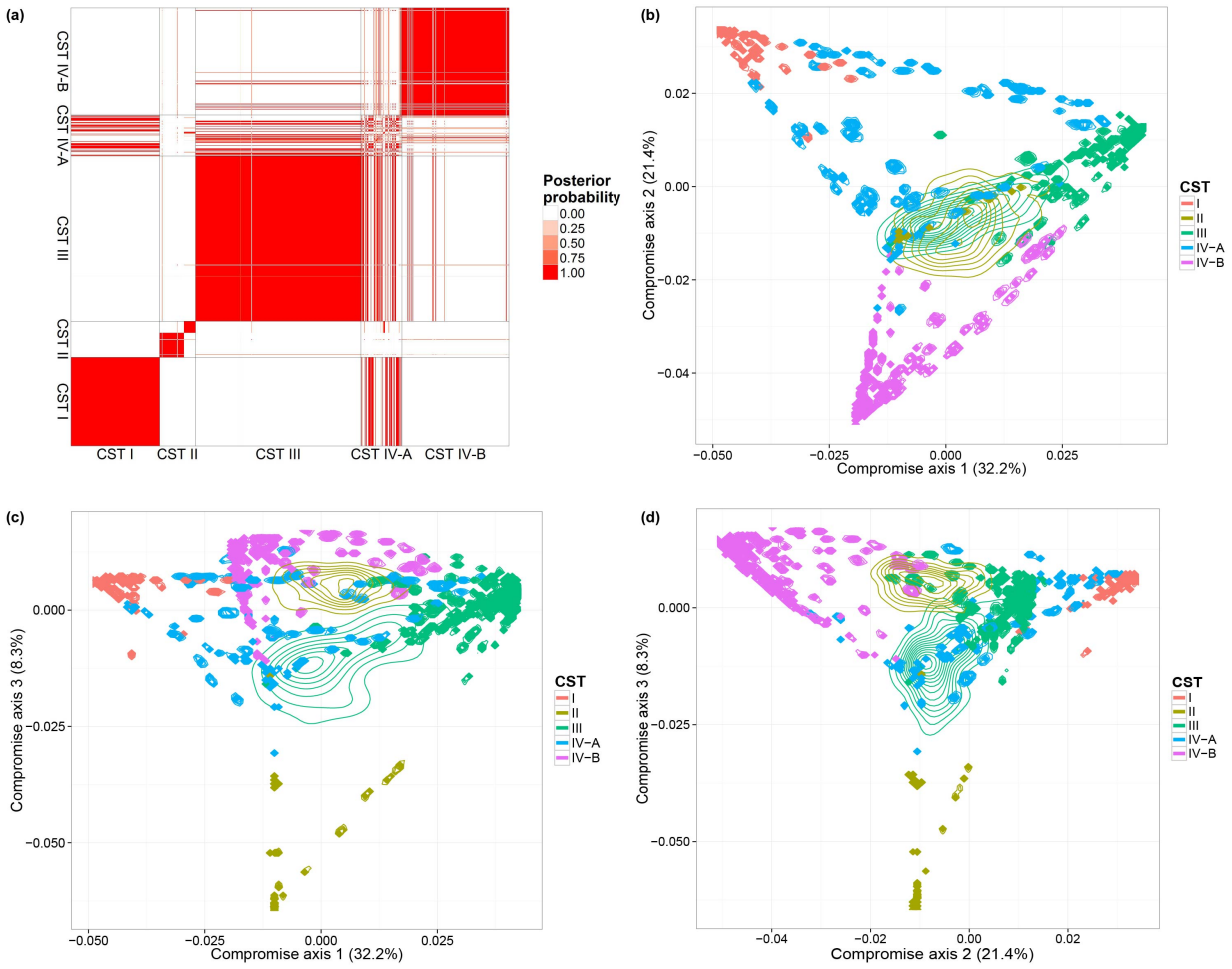


Figure 1.5: (**a**) Posterior Probability of each pair of biological samples (j, j') being clustered together. The labels on axes indicate the CST for each biological sample. (**b-d**) Ordination plots of biological samples and posterior credible regions. We illustrate the first three compromise axes with three panels. The percentages on the three axes are the ratios of the corresponding $\mathbf{S}_0$ eigenvalues and the trace of the matrix. Colors and indicate CSTs.

lation also demonstrated that the analysis yields more accurate estimates of the microbial distributions by borrowing information across biological samples.

In addition, we provide a robust method to visualize the uncertainty in ecological ordinations, furnishing each point in the plot with a credible region. Two published microbiome datasets were analyzed, and the results are consistent with previous findings. The second analysis demonstrates that the level of uncertainty can vary across biological samples due to differences in sampling depth, which underlines the importance of modeling multinomial sampling variations coherently. We believe our analysis will mitigate artifacts arising from rarefaction, thresholding of rare species, and other preprocessing steps.

There are several directions for development which are not explored here. We highlight the possibility of incorporating prior knowledge about the biological samples, such as the subject or group identifier in a clinical study. To achieve this, we can augment the latent factors $\mathbf{Y}^j$ by a vector of covariates $(b_1 w_1^j, \ldots, b_p w_p^j)$, whose coefficients b could be given a normal prior, for example. The posterior distribution of the coefficients could be used to infer the magnitude of covariates' effects. A less straightforward extension involves moving away from the assumption of *a priori* exchangeability between OTUs to include prior information about phylogenetic or functional relationships between them. In our present analysis, these relationships are not taken into account in the definition of the prior for microbial distributions.

Bayesian nonparametric mixed effect latent variable model in microbiome data analysis

Boyu Ren

Department of Biostatistics

Harvard Graduate School of Arts and Sciences

Sergio Bacallado

Department of Pure Mathematics and Mathematical Statistics

University of Cambridge

Stefano Favaro

Departimento di Scienze Economico-Sociali e Matematico-Statistiche

Università di Torino

Tommi Vatanen

Broad Institute of MIT and Harvard

Curtis Huttenhower

Department of Biostatistics

Harvard Chan School of Public Health

Lorenzo Trippa

Department of Biostatistics

Harvard Chan School of Public Health

2.1 Introduction

Large scale survey of microbial composition of host- and environmentally-associated communities becomes readily available by the rapid advancement of Next generation sequencing (NGS) technology. By targeting at the unique marker genes such as 16S rRNA, researchers are able to measure the abundances of microbes by counting the frequencies them in multiple samples. These measurements of microbial abundances are usually organized into a contingency table known as operational taxonomic unit (OTU) table, which serves as the start point of many microbiome studies. This type of data enables quantitative characterization of microbial communities. Particularly, it can be used to estimate the association between sample metadata and microbial abundances, which are relevant for a wide range of biological and clinical questions.

Many statistical methods for association studies with microbial survey data borrow the basic framework of association studies in gene expression data (Morgan et al., 2012b; Paulson et al., 2013a; McMurdie and Holmes, 2013). These methods start with normalizing the raw counts of all microbes within each biological sample to remove artificial variability due to technical variation. The normalized counts are then used for the hypothesis testing of individual microbes' association with sample covariates. Typically, this test is carried out one microbe at a time by using a generalized linear model (GLM) with a distribution family that are over-dispersed and sparse to accommodate the characteristics of microbial abundance data (Li, 2015a). Common choices of such families include zero-inflated log-normal distribution and zero-inflated negative binomial distribution (Xu et al., 2015). This univariate framework is inefficient as it cannot borrow information across different microbes when estimating the effect of sample metadata. Moreover, it ignores the correlation structure between microbes, which may cause problems when adjusting for multiple hypothesis testing.

Recently, people begin to consider multivariate statistical models for microbial survey data (Holmes et al., 2012b; Xia et al., 2013a). Because sequencing data can be treated as multinomial sampling results, it is natural to target on the underlying multinomial distributions of microbes for the association studies. In Holmes et al. (2012b) and Xia et al.

(2013a), the authors model the multinomial probabilities using parametric families on simplex. Both models link the sample metadata to the parameters of the assumed distribution through a multivariate regression. In Holmes et al. (2012b), a Dirichlet distribution is used and in Xia et al. (2013a), a multivariate logistic normal distribution. Although these two models can overcome the difficulties mentioned above in the univariate analyses and are computationally feasible, neither of them is flexible enough to fully capture the characteristics of the underlying multinomial distributions of microbes. Specifically, components of a Dirichlet distribution can only be negatively correlated, which is not consistent with the fact that some OTUs can only exist in tight-knit communities. And the multivariate logistic normal distribution cannot recapitulate the sparseness observed in real microbial survey data.

The limitations in parametric families on simplex can be resolved by considering nonparametric models for discrete distributions. Current development in normalized random measures (Regazzini et al., 2003; James et al., 2009) opened up the possibility of applying nonparametric models for microbial abundance data in Bayesian paradigm. Attempts of using nonparametric methods in microbiome analyses usually employ the most common Bayesian nonparametric model, Dirichlet process, to depict the distributions of microbes. In O'Brien et al. (2016), the authors assume a marginally exchangeable structure of biological samples. Therefore, a series of independent Dirichlet processes are used *a priori* to model each of the biological samples. This exchangeability assumption renders incorporating information of covariates into the model impossible and therefore is not suitable for the association studies in microbiome research. Arbel et al. (2014) allows for dependence structure between biological samples but only focuses on a univariate summary of each microbial distribution.

We discovered a recent Bayesian nonparametric model for microbiome data (Ren et al., 2016) is very relevant to the task of association studies for microbiome data. This model respects the fact that microbial distributions are non-exchangeable by connecting a group of sample-specific latent variables to the collection of microbial distributions. This construction induces a Dependent Dirichlet Processes prior on the microbial distributions, which possesses desirable properties. For example, it assigns positive prior probability to

every possible collection of microbial distributions and tends to produce a parsimonious set of latent variables. The framework lends a convenient extension for association analyses in microbiome studies. If we add the observed sample covariates on top of the latent variables as the potential factors that affect the probabilities of microbes, we can draw conclusion about whether certain covariates are associated with microbial abundances significantly by estimating the magnitude of those effects. We will discuss this extension in details in the remainder of this paper.

The paper is organized as follows: Section 2 will be focused on the description of the extension and some discussion about the model identifiability. Section 3 deals with the computational aspect of the model, offering an overview of the sampling algorithm. Section 4 will be results from simulation studies, in which we are mainly interested in evaluating the performance of our model in terms of estimating parameters and translating the parameters to interpretable results. Section 5 demonstrates an application of the model to a longitudinal microbiome dataset collected from infants. Section 6 concludes and discusses several future directions.

2.2 Method

In this section, we will first review the construction of the Dependent Dirichlet process in Ren et al. We then introduce the extension of this model which incorporates sample covariates and discuss the identifiability of the model parameters, especially the parameters that correspond to the effect of covariates. The data we are working with is the OTU table $(n_{i,j})_{i\in\mathbb{N}^+,j\in\mathcal{J}}$ where $n_{i,j}$ is the observed frequency of species i in biological sample j and $\mathcal{J}$ is the set of biological samples. We want to leverage the information of microbial abundances in heterogeneous samples to discover the relationships between microbial compositions and observed sample characteristics.

2.2.1 Dependent Dirichlet process

To make the notations and definitions more concrete, we illustrate in Table 2.1 an example of a typical OTU table with 12 biological samples, where half are healthy subjects and half

Table 2.1: An example OTU table derived from data published in Vatanen et al. (2016).

Species	G69147	G69149	G69152	G69155	G69156	G69158
Bifidobacterium longum	0	73222	3014074	14294	7291	9228
Bifidobacterium bifidum	3594189	49223	0	11177	11656816	14759
Escherichia coli	4210380	23025	635855	29700	7508	556208
Bifidobacterium breve	0	136	245827	19312	7223273	0
Bacteroides fragilis	0	88751	0	6257732	343	75506
Bacteroides vulgatus	0	7454	0	4745	0	25859
Bacteroides dorei	0	0	0	0	0	0
Bifidobacterium adolescentis	0	111248	1626357	735715	1194	0
Bacteroides uniformis	0	3901	0	5859	1633	28638
Ruminococcus gnavus	145485	33004	92101	253830	29	1186774

are Inflammatory Bowel disease (IBD) patients. This type of contingency table records the observed frequencies of 10 genus level OTUs in a collection of biological samples based on 16S sequencing results. Let Z_i be the ith observed OTU (e.g. Z_1 is Bacteroides) and $n_{i,j}$ be the observed frequency of OTU Z_i in biological sample j. As an example, $n_{11} = 1822$ is the observed frequency of Bacteroides in the biological sample Control.1. We will denote an OTU table as $(n_{i,j})_{i \leq I, j \leq J}$, where I is the number of observed OTUs and J the number of biological samples.

For the biological sample j, we will assume the vector $(n_{1,j}, \ldots, n_{I,j})$ follows a multinomial distribution, noting that our analysis extends easily to the case in which the total count $\sum_{i=1}^{I} n_{ij}$ is a Poisson random variable.The unobserved multinomial probabilities of OTUs present in biological sample j determine the distribution of frequencies completely. These probabilities form a discrete probability measure, which we call a microbial distribution, on the set of all OTUs. We denote this discrete measure as P^j and $P^j(\{Z_i\})$ gives the probability of sampling Z_i from biological sample j. If we consider all J biological samples, we expect there will be variation in the respective P^j's. This variation usually can be explained by specific characteristics of the biological sample. For instance, in Table 2.1, we can see the empirical probability of Enterococcus is higher in healthy controls than in IBD patients. This variation is attributed to the IBD status (Morgan et al., 2012b). Microbiome studies aim to elucidate the characteristics that explain this kind of variation. We focus on modeling the P^j's and the variation among them. For biological samples labelled in $\mathcal{J} = \{1, \ldots, J\}$, we assume they have the same infinite set of OTUs

$Z_1, Z_2, \ldots \in \mathcal{Z}$. We let the number of OTUs present in a biological sample be infinity to make our model nonparametric in consideration of the fact that there might be an unknown number of OTUs that are not observed in the experiment. We specify the probability mass assigned to a group of OTUs $A \subset \mathcal{Z}$ as

$$\begin{aligned} P^j(A) &= M^j(A)/M^j(\mathcal{Z}), \\ M^j(A) &= \sum_{i=1}^{\infty} \mathbb{I}(Z_i \in A)\sigma_i \langle \mathbf{X}_i, \mathbf{Y}^j \rangle^{+2}, \end{aligned} \tag{2.1}$$

where $\sigma_i \in (0,1)$, $\mathbf{X}_i, \mathbf{Y}^j \in \mathbb{R}^K$, $\mathbb{I}(\cdot)$ is the indicator function and $x^+ = x \times \mathbb{I}(x > 0)$. In addition, $\langle \cdot, \cdot \rangle$ is the standard inner product in $\mathbb{R}^K$.

In this model specification, σ_i is related to the average abundance of OTU i across all biological samples. When σ_i is large, the average probability mass assigned to OTU Z_i will also be large. We refer to $\mathbf{X}_i$ and $\mathbf{Y}^j$ as OTU vector and biological sample vectors respectively. The variation of the P^j's is determined by the vectors $\mathbf{Y}^j$, which can be treated as latent characteristics of the biological samples that associate with microbial composition; for example, an unobserved feature of the subject's diet, such as vegetarianism, could affect the abundance of certain OTUs. We assume there are at most K such characteristics. $\mathbf{X}_i$ has the same dimension as $\mathbf{Y}^j$ and it denotes the effect of each latent characteristic of the biological samples on the abundance of the OTU Z_i.

If we assume $\sigma_1 > \sigma_2 > \sigma_3 \ldots$, are ordered points from a Poisson process *a priori* on $(0,1)$ with intensity $\nu(\sigma) = \alpha\sigma^{-1}(1-\sigma)^{-1/2}$, priors on $X_{l,i}$ are iid $N(0,1)$ for $i = 1, 2, \ldots$, $l = 1, 2, \ldots, K$, and $\langle \mathbf{Y}^j, \mathbf{Y}^j \rangle < \infty$ for all $j \in \mathcal{J}$, the induced prior on P^j is a Dirichlet process marginally (Ren et al., 2016). To see this heuristically, we note that this prior implies that fixing j, $Q_{i,j}^{+2} (i = 1, 2, \ldots)$ are independent and are mixtures of a point mass at zero and a Gamma distribution with shape parameter $1/2$. In addition, for a fixed m, the prior distribution of $(\sigma_1, \ldots, \sigma_m)$ can be approximated by Beta$(\alpha/m, 1/2 - \alpha/m)$. Therefore, when $\alpha/m < 1/2$, the prior distribution of normalized $(\sigma_1 Q_{1,j}^{+2}, \ldots, \sigma_m Q_{m,j}^{+2})$ follows approximately Dirichlet$(\alpha/m, \ldots, \alpha/m)$. As a result, if m goes to infinity, we expect the prior distribution of P^j becomes a Dirichlet process.

When applied to real data, we need to specify the dimension of $\mathbf{Y}^j$. Usually, we want to find a set of $\mathbf{Y}^j$ with reasonably small dimension but can approximate the angles between

samples well. We applied a shrinkage prior (Bhattacharya and Dunson, 2011) on the components of $\mathbf{Y}^j$ to achieve this goal. This prior tries to make the first several components of $\mathbf{Y}^j$ capture as much variation as possible and shrink $Y_{l,j}$ towards zero for each $j \in \mathcal{J}$ when l is large.

2.2.2 Dependent Dirichlet process with fixed effect

Microbiome studies usually seek to discover relationships between microbial compositions and sample characteristics. For example, in the studies of Inflammatory Bowel Disease(IBD), researchers are interested to find microbes that are correlated with the onset of IBD and try to devise the treatment accordingly. Such problems are usually solved by using a regression model with the abundances of microbes as the outcome. The estimates of regression coefficients are interpreted as the measurement of associations between sample covariates and microbial abundances. Following the same strategy, we expanded the model in Section 2.2.1 to perform a multivariate regression analysis to jointly estimate the effect of covariates on all observed species. This extended model is very flexible in that it allows for correlation structures in the regression residuals.

Assume there are $K \geq 1$ observed covariates. Denote the observed covariates for sample j as $\mathbf{w}^j = (w_{1,j}, \ldots, w_{K,j})^\intercal$ and the effect of this set of covariates on species i as $\mathbf{v}_i = (v_{1,i}, \ldots, v_{K,i})^\intercal$. Our extended model directly modify the term $Q_{i,j}$ in (2.1) by adding a linear function of $\mathbf{w}_j$:

$$Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}^j \rangle + \mathbf{v}_i^\intercal \mathbf{w}^j + \epsilon_{i,j}, \tag{2.2}$$

where $\epsilon_{i,j} \overset{iid}{\sim} \mathcal{N}(0,1)$ are pure noise.

In this construction, $\mathbf{v}_i$ and $\mathbf{w}^j$ can be seen as additional dimensions of $\mathbf{X}_i$ and $\mathbf{Y}^j$. It follows that the concatenated vector $(\mathbf{w}^j, \mathbf{Y}^j)^\intercal$ indicates the position of sample j in a subspace and the angle between $(\mathbf{w}^j, \mathbf{Y}^j)^\intercal$ and $(\mathbf{w}^{j'}, \mathbf{Y}^{j'})^\intercal$ measures the similarity of distributions of microbes in biological sample j and j'. The concatenated vector $(\mathbf{v}_i, \mathbf{X}_i)^\intercal$ is interpreted similarly. We added the noise term $\epsilon_{i,j}$ to capture the variability that cannot be explained by the linear term $\langle \mathbf{X}_i, \mathbf{Y}^j \rangle + \mathbf{v}_i^\intercal \mathbf{w}^j$. The variance of this error is fixed at one due to the fact that any scaling of $Q_{i,j}$ will result in the same model on P^j. Specifying a

variance parameter for $\epsilon_{i,j}$ is thus redundant. We put an independent $MVN(\mathbf{0}, \mathbb{I})$ prior on each $\mathbf{v}_i$. When fixing the latent factors $\mathbf{Y}^j$ and preserving the prior specification of $\mathbf{X}_i$ and σ_i in Section 2.2.1, the marginal prior of P^j induced by (2.2) remains to be a Dirichlet process. This can be proved using the same argument as in Section 2.2.1.

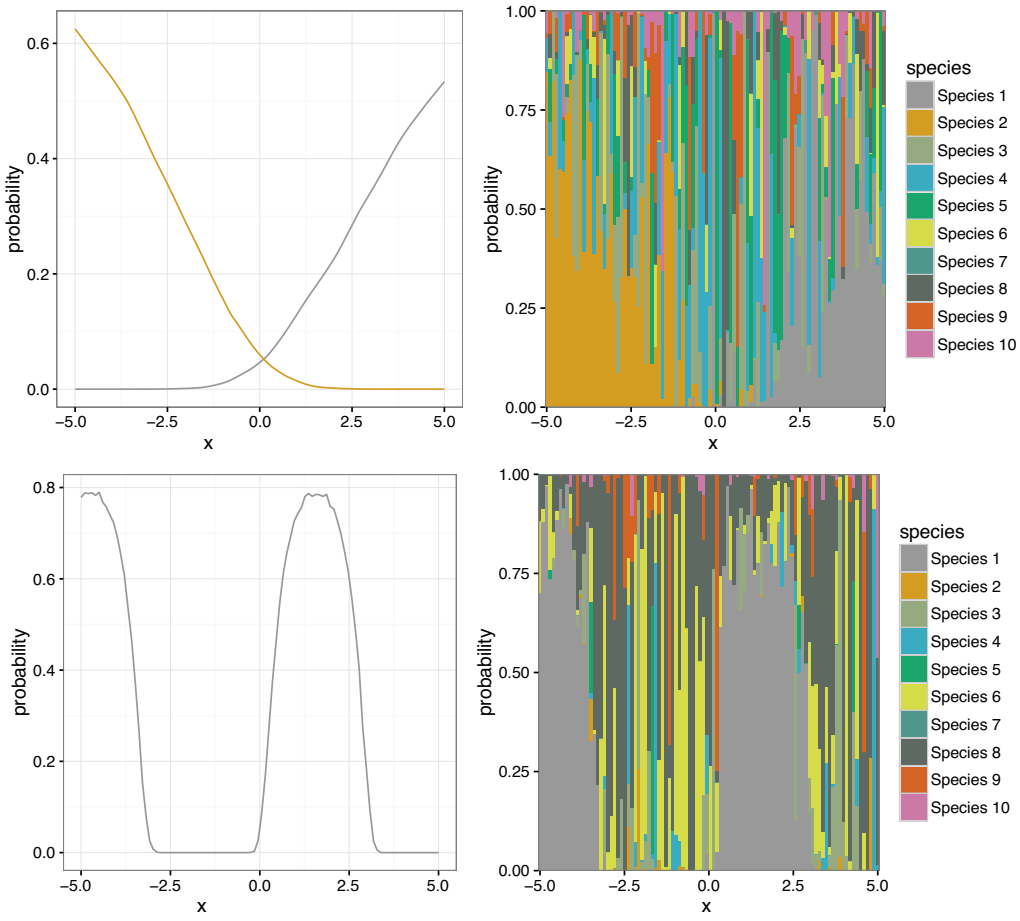


Figure 2.1: Effect of a one-dimensional covariate on the probability of a species. We considered the case where the covariate is w, $\sin(w)$ and $\mathbb{I}(w > 0)$. For each scenario, we plot the the expected (Left panel) and observed (Right panel) probabilities of a species as a function of w when w ranges from -5 to 5.

The model in (2.2) is effectively a generalized multivariate linear regression model with unobserved covariates, in which $\mathbf{X}_i$ and $\mathbf{v}_i$ quantify the effect of the unobserved and observed covariates respectively. Due to the randomness in the model specification, the observed microbial abundance is not necessarily a monotone function of the covariate values (Figure 2.1). We are mainly interest in $\mathbf{v}_i$ as it is immediately related to the biological questions in real applications. (2.2) is more flexible than the typical multivariate linear

regression in that when estimating $\mathbf{v}_i$, we can non-parametrically adjust for latent covariates $\mathbf{Y}^j$ that contribute to the variations of the multinomial distributions. Since there is no clear interpretation of the estimated latent factor $\mathbf{Y}^j$, $\mathbf{X}_i$ is not of major interest in our analyses. The more important information carried by $\mathbf{Y}^j$ is the Gram matrix $(\langle \mathbf{Y}^j, \mathbf{Y}^{j'} \rangle)_{j,j' \in \mathcal{J}}$, which represents the similarity structure between the residuals of the regression.

2.2.3 Identification of model parameters

In this section, we consider the identifiability of the model parameters σ_i, $\mathbf{Y}^j$ and $\mathbf{v}_i$ in the likelihood. We integrate out $\mathbf{X}_i$ since our major interest is on $\mathbf{v}_i$ and $\mathbf{Y}^\intercal\mathbf{Y}$. Throughout the section, we assume that $\sum_i n_{i,j}$ is large enough such that $n_{i,j}/(\sigma_i Q_{i,j}^{+2}) = c_j > 0$ for every i and j. This assumption holds approximately when the number of species are much smaller than the total number of cells that are counted in each biological sample. It is reasonable as modern high-throughput sequencing platforms can easily achieve large read-depth (n^j).

We notice that the observed data are finite species sampling sequences data. The labelling of species based on this type of data is usually arbitrary since there is no inherent order of species. Usually, we label the first observed species in the first biological samples as species 1, the second as species 2 and so on. Indeed, if we assume the species have label $\{1, 2, \ldots\}$, the labels given by this assignment rule is a permutation π of $1, 2, \ldots$. When we refer to species i and $Z_{i,j}$, σ_i, $\mathbf{v}_i$ in the rest of this section, we are actually referring to $Z_{\pi_i,j}$, σ_{π_i} and $\mathbf{v}_{\pi_i}$, where π_i is the ith element of π. The joint distribution of all $Z_{\pi_i,j}$ is different from the joint distribution $Z_{i,j}$ only in the normalizing constant since species are exchangeable. Therefore, if the distribution of $(Z_{\pi_i,j})_{i\geq 1, j\leq J}$ are different under two sets of model parameters, the distributions of $(Z_{i,j})_{i\geq 1, j\leq J}$ are also different. As a result, analyzing the distribution of $Z_{\pi_i,j}$, σ_{π_i} and $\mathbf{v}_{\pi_i}$ in the argument of model identifiability is sufficient.

Let $\mathbf{w} = (\mathbf{w}^1, \ldots, \mathbf{w}^J)$ and $\mathbf{Y} = (\mathbf{Y}^1, \ldots, \mathbf{Y}^J)$. We first notice the fixed effect model in (2.2) implies that

$$(Q_{i,1}, \ldots, Q_{i,J})^\intercal \sim MVN(\mathbf{Y}^\intercal \mathbf{v}_i, \mathbf{\Sigma}),$$

where $\mathbf{\Sigma} = \mathbf{Y}^\intercal\mathbf{Y} + I_J$ and I_J is the $J \times J$ identity matrix. Since $n_{i,j}/(\sigma_i Q_{i,j}^{+2}) = c_j$, we have $n_{i,j} = 0 \Leftrightarrow Q_{i,j} < 0$. Consider a transformed dataset, denoted as $(Z_{i,j})_{i\geq 1, j\in\mathcal{J}}$, where $Z_{i,j} = \mathbb{I}(Q_{i,j} = 0)$. $z_{i,j} = \mathbb{I}(n_{i,j} > 0)$ is a realization of the random variable $Z_{i,j}$. If the model parameters are identifiable under the reduced dataset $(Z_{i,j})_{i\geq 1, j\in\mathcal{J}}$, it follows that those parameters are also identifiable using the full dataset $(n_{i,j})_{i\geq 1, j\in\mathcal{J}}$. In the following discussion, we will mainly use the reduced dataset to investigate the identifiability of parameters. But we will also consider the full dataset if necessary.

The probability mass function of $(Z_{i,j})_{i\geq 1, j\in\mathcal{J}}$ can be written as

$$\begin{aligned} &P(Z_{i,j} = z_{i,j}, i \geq 1, j \in \mathcal{J}; \{\sigma_i\}, \{\mathbf{Y}^j\}, \{\mathbf{v}_i\}) \\ =&\prod_i \left[\int_{A_i} (2\pi)^{-J/2} |\mathbf{\Sigma}|^{-1/2} \times \exp\left(-\frac{1}{2} (\mathbf{Q}_i - \boldsymbol{\mu}_i)^\intercal \mathbf{\Sigma}^{-1} (\mathbf{Q}_i - \boldsymbol{\mu}_i) \right) d\mathbf{Q}_i \right]. \end{aligned} \tag{2.3}$$

Here $\mathbf{Q}_i = (Q_{i,1}, \ldots, Q_{i,J})^\intercal$, $\mu_i = \mathbf{w}^\intercal \mathbf{v}_i$ and $\mathbf{\Sigma} = \mathbf{Y}^\intercal\mathbf{Y} + I_J$. $A_i = \times_{j=1}^J A_{i,j}$ and $A_{i,j} = (-\infty, 0]$ if $z_{i,j} = 0$, $A_{i,j} = [0, \infty)$ if $z_{i,j} = 1$. To illustrate the identifiability of the model parameters, we start with two simplified cases and then give a proposition for the general case.

1. *No ranodm effect* ($\mathbf{Y} = \mathbf{0}$). In this scenario, we only need to consider two sets of parameters, $\{\sigma_i\}$ and $\{\mathbf{v}_i\}$. Using the reduced dataset $(Z_{i,j})_{i\geq 1, j\in\mathcal{J}}$, we can argue that all $\mathbf{v}_i$ are identifiable. Indeed, for a fixed i, the likelihood of $(Z_{i,1}, \ldots, Z_{i,J})$ is exactly the likelihood of a standard probit model, where the mean vector and covariance matrix of the underlying normal random variables is $\mathbf{w}^\intercal\mathbf{v}_i$ and I_J respectively. Based on the asymptotic theory of generalized linear model, the MLE of $\mathbf{v}_i$ is consistent when $J \to \infty$, if $\mathbf{w}^j$ are iid and $\mathbb{E}\mathbf{w}^j(\mathbf{w}^j)^\intercal$ is positive definite. The consistency of an estimator of $\mathbf{v}_i$ implies immediately the identifiability of it. Using argument similar to the proof of Proposition 2.1, we can also prove that $\sigma_i/\sigma_{i'}$ is identifiable for every pair of $i \neq i'$ based on the full dataset $(n_{i,j})_{i\geq 1, j\leq J}$.

2. *No fixed effect* ($\mathbf{v}_i = \mathbf{0}$). We now only need to identify $\{\sigma_i\}$ and $\mathbf{Y}$ (or $\mathbf{\Sigma}$ equivalently). Consider the distribution of $(Z_{1,j}, Z_{1,j'})$, it can be written as

$$P(Z_{1,j} = z_{1,j}, Z_{1,j'} = z_{1,j'}) \propto \int_{A_1} |\mathbf{\Sigma}_{j:j'}|^{-1/2} \exp\left(-\frac{1}{2} \mathbf{Q}^\intercal \mathbf{\Sigma}_{j:j'}^{-1} \mathbf{Q} \right) d\mathbf{Q},$$

where $\mathbf{\Sigma}_{j:j'} = \begin{pmatrix} \Sigma_{j,j} & \Sigma_{j,j'} \\ \Sigma_{j',j} & \Sigma_{j',j'} \end{pmatrix}$ and $A_1 = A_{1,j} \times A_{1,j'}$, $A_{1,j} = (-\infty, 0]$ if $z_{1,j} = 0$ and $A_{1,j} = [0, \infty)$ if $z_{1,j} = 1$. Theorem 3.1 in Ledoux and Talagrand (2013) shows

that $P(Z_{1,j} = z_{1,j}, Z_{1,j'} = z_{1,j'})$ is an increasing function of the correlation between j and j', $\Sigma_{j,j'}/\sqrt{\Sigma_{j,j}\Sigma_{j',j'}}$. This implies that the correlation matrix $\mathbf{S}$ induced by $\mathbf{Y}^\intercal\mathbf{Y} + I_J$ is identifiable from the data. In fact, this conclusion cannot be improved by considering the full data. This is because the likelihood function of the full data is also invariant under the standardization of the covariance matrix $\mathbf{Y}^\intercal\mathbf{Y} + I_J$. The identifiability of $\sigma_i/\sigma_{i'}$ can be established using the same argument in Proposition 2.1.

3. *Both fixed and random effect.* This general case can be treated as the integration of the previous two scenarios. To have identifiability, we put a constraint on the sample covariates $\mathbf{w}^j$ such that $\mathbf{w}^j \overset{iid}{\sim} f$, where f is a continuous distribution on $\mathbb{R}^k$. We give the following proposition for the identifiability of the model parameters based on the absence-presence data $(Z_{i,j} = \mathbb{I}(Q_{i,j} > 0))_{i\geq 1, j\leq J}$ as well as the full dataset $(n_{i,j})_{i\geq 1, j\leq J}$. The proof is in the appendix.

   Proposition 2.1. *Assume $\mathbb{E}\mathbf{w}^j(\mathbf{w}^j)^\intercal$ is positive definite and $\sum_j \Sigma_{j,j} = 1$. Consider two different sets of model parameters $\{(\sigma_i)_{i\geq 1}, (\mathbf{v}_i)_{i\geq 1}, \mathbf{\Sigma}\}$ and $\{(\sigma'_i)_{i\geq 1}, (\mathbf{v}'_i)_{i\geq 1}, \mathbf{\Sigma}'\}$. The distribution of the data $(Z_{i,j})_{i\geq 1, j\leq J}$ is the same under these two sets of parameter values if and only if $\mathbf{v}_i = \mathbf{v}'_i$ and $\mathbf{\Sigma} = \mathbf{\Sigma}'$ for every $i \geq 1$. Moreover, if there exists a pair of $i \neq i'$ such that $\sigma_i/\sigma_{i'} \neq \sigma'_i/\sigma'_{i'}$, the distribution of $(n_{i,j})_{i\geq 1, j\leq J}$ will be different under these two sets of parameter values.*

2.3 Posterior simulation and visualization

In this section, we focus on the computational aspects of our model. In Section 2.3.1, we introduce a subject-specific latent factor model and the posterior simulation algorithm we used for this model. This special case of the model in (2.2) utilizes information of subjects in longitudinal dataset and will be used in our real data application. In Section 2.3.2, we discuss a summary statistic that transforms the estimates of parameter $\mathbf{v}_i$ into interpretable results regarding the effect of covariates on species. This statistic is necessary as the value of $\mathbf{v}_i$ does not directly indicate the relationship between covariates and relative abundances of species.

2.3.1 Gibbs sampler for subject-specific latent factor model

We first define the subject-specific latent factor model. Denote s_j as the subject index of the jth sample. The subject-specific latent factor model puts an additional restriction on the mixed effect model in (2.2) by forcing the latent factor $\mathbf{Y}^j$ to be identical for all samples that are derived from the same subject. To be more specific, we require $\mathbf{Y}^j = \mathbf{Y}^{j'}$ when $s_j = s_{j'}$. Assume there are N subjects in the dataset, and denote the latent factor corresponds to subject s as $\widetilde{\mathbf{Y}}^s$, we can rewrite (2.2) as

$$Q_{i,j} = \langle \mathbf{X}_i, \widetilde{\mathbf{Y}}^{s_j} \rangle + \mathbf{v}_i^\intercal \mathbf{w}^j + \epsilon_{i,j}. \tag{2.4}$$

The indicators of subjects of all samples can be collected in a $J \times N$ matrix $\mathbf{B}$ as in linear mixed effect model. Specifically, $B_{j,s} = 1$ when sample j belongs to subject s and zero otherwise. Using this matrix, we can write the model specification in a more compact way:

$$\mathbf{Q}_i = \mathbf{X}_i^\intercal \widetilde{\mathbf{Y}} \mathbf{B}^\intercal + \mathbf{v}_i^\intercal \mathbf{w} + \boldsymbol{\epsilon}_i,$$

where $\widetilde{\mathbf{Y}} = (\widetilde{\mathbf{Y}}^1, \ldots, \widetilde{\mathbf{Y}}^N)$.

Based on this model formulation, it is straightforward to derive the full conditional distribution of the species factor $\mathbf{X}_i$ and species regression coefficient $\mathbf{v}_i$ based on the results in Ren et al. (2017):

$$\begin{aligned}
\mathbf{X}_i | \widetilde{\mathbf{Y}}, \mathbf{Q}_i, \mathbf{v}_i &\sim MVN((\mathbf{B}\widetilde{\mathbf{Y}}^\intercal \widetilde{\mathbf{Y}} \mathbf{B}^\intercal + \mathbb{I})^{-1} \mathbf{B}\widetilde{\mathbf{Y}}^\intercal (\mathbf{Q}_i - \mathbf{v}_i^\intercal \mathbf{w}), (\mathbf{B}\widetilde{\mathbf{Y}}^\intercal \widetilde{\mathbf{Y}} \mathbf{B}^\intercal + \mathbb{I})^{-1}), \\
\mathbf{v}_i | \widetilde{\mathbf{Y}}, \mathbf{X}_i, \mathbf{Q}_i &\sim MVN((\mathbf{w}^\intercal \mathbf{w} + \mathbb{I})^{-1} \mathbf{w}^\intercal (\mathbf{Q}_i - \mathbf{X}_i^\intercal \widetilde{\mathbf{Y}} \mathbf{B}), (\mathbf{w}^\intercal \mathbf{w} + \mathbb{I})^{-1}).
\end{aligned}$$

The only part that is different will be the sampling of $\widetilde{\mathbf{Y}}^s$ as it is now shared by multiple samples. Denote the number of repeated measures for subject s as m_s, the full conditional posterior distribution of $\widetilde{\mathbf{Y}}^s$ is

$$\widetilde{\mathbf{Y}}^s | \mathbf{X}_i, \mathbf{Q}_i, \mathbf{v}_i, \boldsymbol{\Sigma}_s \sim MVN((\mathbf{X}_i^\intercal \mathbf{X}_i + 1/m_s \boldsymbol{\Sigma}_s)^{-1} \mathbf{X}_i^\intercal \sum_{j:s_j = s} (Q_{i,j} - \mathbf{v}_i^\intercal \mathbf{w}^j), (m_s \mathbf{X}_i^\intercal \mathbf{X}_i + \boldsymbol{\Sigma}_Y)^{-1}),$$

where $\boldsymbol{\Sigma}_Y$ is derived from the shrinkage prior we assumed for $\mathbf{Y}^j$. Notice now this shrinkage prior is effectively applied to $\widetilde{\mathbf{Y}}^s$ and thus we do not need to rewrite the sampling algorithm for updating the hyper-parameters in the shrinkage prior.

2.3.2 Converting the model parameters to interpretable results

In this section, we discuss a summary statistics we used to demonstrate relationship between covariates and microbial abundances based on estimates of $\mathbf{v}_i$. In the scenario where one component of $\mathbf{w}^j$ is continuous (assume it is $w_{k,j}$), we usually want to know how the microbial abundance changes when $w_{k,j}$ changes for each biological sample. We propose to estimate this relationship by calculating the first order derivative of the estimated species abundance at the observed biological samples. It will be sensitive to rapid changes in microbial abundance and can be potentially useful when detecting disruptive events of microbial communities. Notice the derivative here captures the model-based rate of change specific to each biological sample, which is affected by between-subject variability.

Using the model specification, we can derive a closed-form expression for the derivatives. Recall the probability of species i in biological sample j is $P^j(\{Z_i\}) = \sigma_i Q_{i,j}^{+2}/(\sum_i \sigma_i Q_{i,j}^{+2})$. It follows that the derivative of this quantity over a continuous covariate $w_{k,j}$ can be written as

$$\begin{aligned}\frac{\partial P^j(\{Z_i\})}{\partial w_{k,j}} =& \partial[\frac{\sigma_i(\mathbf{X}_i^\intercal \mathbf{Y}^j + \mathbf{v}_i\mathbf{w}^j + \epsilon_{i,j})^{+2}}{\sum_l \sigma_l(\mathbf{X}_l^\intercal \mathbf{Y}^j + \mathbf{v}_l^\intercal \mathbf{w}^j + \epsilon_{l,j})^{+2}}]/\partial w_{k,j} \\ =& \frac{2\sigma_i v_{k,i} Q_{i,j}^{+} \sum_l \sigma_l Q_{l,j}^{+2} - 2\sigma_i Q_{i,j}^{+2} \sum_l \sigma_k v_{k,l} Q_{l,j}^{+}}{(\sum_l \sigma_l Q_{l,j}^{+2})^2},\end{aligned}$$

where $Q_{i,j} = \mathbf{X}_i^\intercal \mathbf{Y}^j + \mathbf{v}_i\mathbf{w}^j + \epsilon_{i,j}$. We can calculate the posterior distribution of this quantity based on the MCMC simulation results. In the calculation, we average over multiple realization of $\epsilon_{i,j}$ to get rid of effect of pure noise.

2.4 Simulation results

In this section, we focus on the subject-specific model specification introduced in Section 2.3.1 and illustrate that we can estimate the parameters of interest accurately, which verifies numerically our results on model identifiability. We then use the posterior samples of the model parameters to estimate the rate of change of relative abundances over a continuous covariate in individual level and the overall trends in population level. This

simulation demonstrates that we can transform the model parameters into interpretable and biologically interesting results, which recapitulate the ground truth accurately.

In all the simulation studies we carried out, we assume there are 100 species ($I = 100$) and 300 biological samples ($J = 300$). The 300 biological samples are repeated measures from 50 subjects ($N = 50$), with each subject measured 6 times. We assume the total number of reads for each biological sample is $n^j = 10^5$. This large read depth makes the approximation in Proposition 2.1 reasonable. We included in the simulation a continuous covariate $w_{1,j}$, generated independently from $N(0, 1)$, and a subject-specific binary covariate $w_{2,j}$, generated from a Bernoulli distribution with probability 0.5. We also added the interaction term $w_{1,j} \times w_{2,j}$ to mimic the scenario where trends over time differ in the two groups of subjects. For the latent factor $\widetilde{\mathbf{Y}}$, we assumed the true dimension of it is four ($\widetilde{\mathbf{Y}}^s \in \mathbb{R}^4$) and for the first half of the subjects, $s = 1, \ldots, 25$, we set $\widetilde{Y}_{3,s} = \widetilde{Y}_{4,s} = 0$ and for the other half, $s = 26, \ldots, 50$, we set $\widetilde{Y}_{1,s} = \widetilde{Y}_{2,s} = 0$. The non-zero components in $\widetilde{\mathbf{Y}}^s$ were simulated from $N(0, 1)$ independently. This specification of $\widetilde{\mathbf{Y}}$ makes the underlying correlation matrix $\mathbf{S}$ to be block diagonal. In the latent scale, we specify the first eight species to have increasing trends over $\mathbf{w}_1$ and the following eight species to have decreasing trend. Within each eight species, we further assume half of them are more abundant in subjects with $w_{2,j} = 1$ and the other half less abundant. Moreover, we assume the trends over $\mathbf{w}_1$ are either amplified, evened out or flipped when changing from one of the groups to the other in these 16 species. The rest of the species will have no relationship with $\mathbf{w}$. This results in the following specification of $\mathbf{v}$:

$$\mathbf{v} = \begin{pmatrix} 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & -5 & -5 & -5 & -5 & -5 & -5 & -5 & -5 & 0 & \ldots & 0 \\ 5 & 5 & 5 & 5 & -5 & -5 & -5 & -5 & 5 & 5 & 5 & 5 & -5 & -5 & -5 & -5 & 0 & \ldots & 0 \\ 10 & -5 & -5 & -10 & 10 & -5 & -5 & -10 & -10 & 5 & 5 & 10 & -10 & 5 & 5 & 10 & 0 & \ldots & 0 \end{pmatrix}.$$

2.4.1 Estimate the correlation matrix S and regression coefficients v

We consider first the estimation of $\mathbf{S}$ and $\mathbf{v}$, which are the most interesting parameters we want to recover from the data. We only focus on the between-subject correlation matrix $\widetilde{\mathbf{S}}$, which is given by normalizing $\widetilde{\mathbf{Y}}^\intercal\widetilde{\mathbf{Y}} + \mathbb{I}$. Since $\mathbf{v}$ is identifiable up to a constant determined by the ratio between estimated $\boldsymbol{\Sigma}$ and the true $\boldsymbol{\Sigma}$ (see Proposition 2.1), we scaled the posterior samples of $\mathbf{v}$ by the posterior mean of this ratio, which makes them comparable

to the true values. In Figure 2.2, we show the results from the above simulation. The plots suggest in general, the estimates of $\mathbf{v}$ are accurate. Several exceptions (e.g. Species 15) are species with low overall abundances. The estimate of $\mathbf{S}$ is also highly accurate, with negligible level of non-zero correlation between two groups of subjects. Both results confirm the conclusion in Proposition 2.1, which validates the usefulness of our model specification.

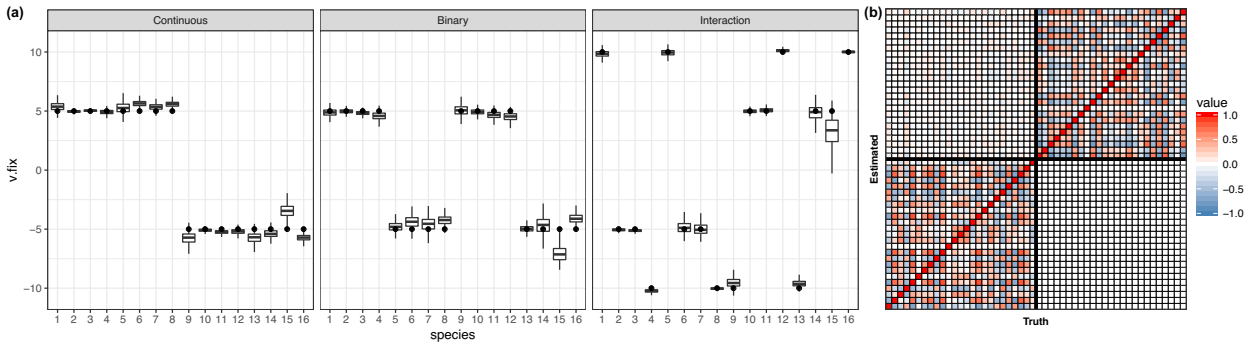


Figure 2.2: Bayesian estimates accurately recover the truth. (**a**) Rescaled estimates of regression coefficients $\mathbf{v}$. The boxplots show the posterior distribution of each component of $\mathbf{v}$ for the first 16 species. The black dots indicate the corresponding true values. (**b**) Posterior mean of between-subject correlation matrix $\widetilde{\mathbf{S}}$ compared to the truth. The lower triangle of the matrix shows the truth and the upper triangle shows the corresponding posterior mean.

2.4.2 Estimating the relationship between the continuous covariate and the probabilities of species

As we have mentioned in Section 2.3.2, the values of $\mathbf{v}$ are not directly related to the trend of relative abundances due to normalization. We thus want to check if we can estimate the trend of species abundances over the continuous covariate $\mathbf{w}_1$. We consider two different kinds of estimates in this section. The first is generated using the algorithm described in Section 2.3.2 and it provides individual-level information about the relationship between $\mathbf{w}_1$ and relative abundances of species. The second provides instead the population level trend of relative abundances over $\mathbf{w}_1$. For each posterior sample of all model parameters, we generate a curve of relative abundances over $\mathbf{w}_1$ for each subject. By averaging all those curves, we can get a population average trend over $\mathbf{w}_1$. We generated all estimates

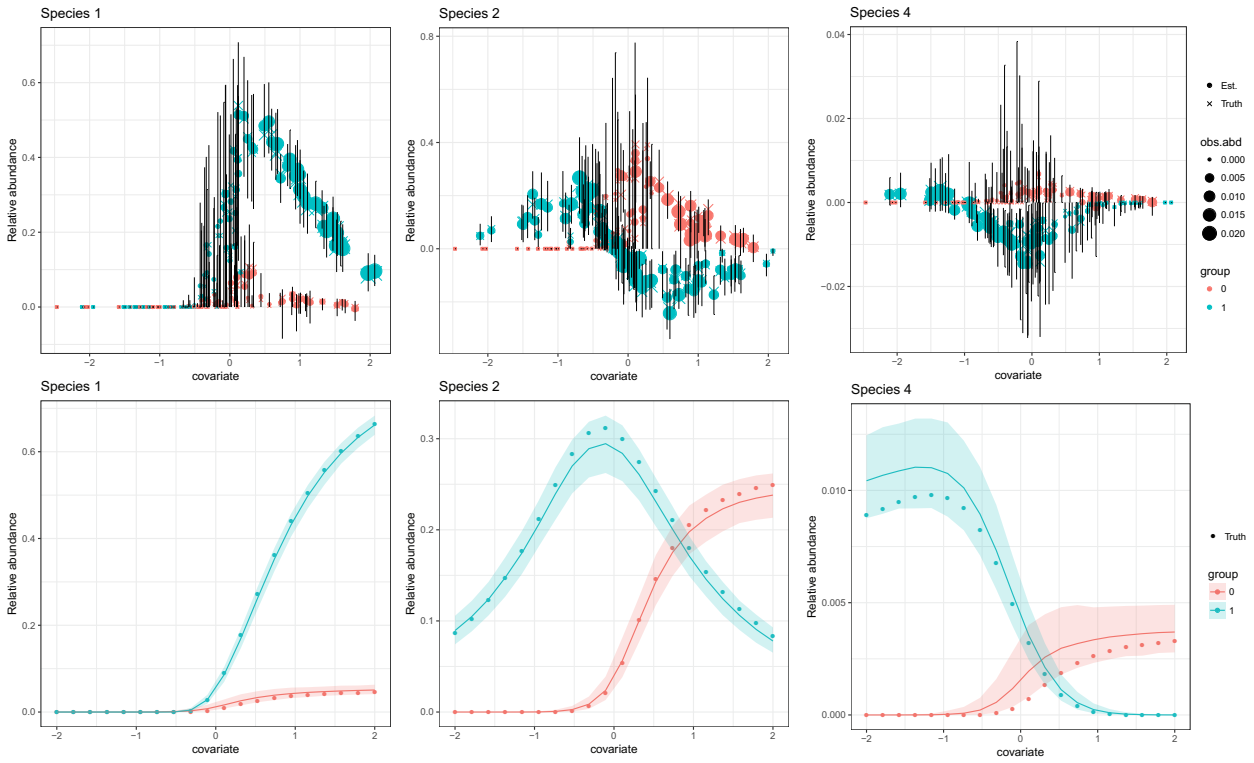


Figure 2.3: Posterior estimates of individual- and population- level relationship between $\mathbf{w}_1$ and relative abundances, stratified by $\mathbf{w}_2$. (**Left**) Increasing trend for group 0 and faster increasing trend for group 1. (**Middle**) Increasing trend for group 0 and no trend for group 1. (**Right**) Increasing trend for group 0 and decreasing trend for group 1. Higher abundance in group 1 overall for all three species.

stratified by $\mathbf{w}_2$. Results for three representative species are summarized in Figure 2.3. From the results we can see both the estimated derivatives of species abundances and the estimates of the population average trend over $\mathbf{w}_1$ are close to the corresponding truth. In addition, the 95% posterior credible intervals/bands cover the truth well. For the population average trend, the credible bands tend to shrink when the observed species abundance is high, which is concordant with the fact that more information is available for those species. For species that are with low abundance (e.g. species 4), the relative discrepancy between the estimated curve and the truth tends to get larger, where the relative width of the credible bands are also larger. Overall, the plots suggest our model captures accurately both the individual- and population-level information about the effect of the continuous covariate $\mathbf{w}_1$. It also indicates the evaluation of the uncertainty on the estimates is sensible.

2.5 Diabimmune data analysis

In this section, we apply our model to a gut microbiome dataset (DIABIMMUNE) collected on 157 newly-born infants over a period extended to 1600 days after birth, yielding 762 samples and 55 genera (Vatanen et al., 2016). These infants are from Finland, Estonia and Russia and seven of them are seroconverted, an indicator of onset of type I diabetes. This dataset also includes dietary information as well as demographic and clinical information. Vatanen et al. discovered a large collection of potential associations between relative abundances of taxa and sample covariates. Among these associations, the most significant ones are related to nationality and age of the infants. Although the goal of collecting this dataset is to examine the relationship between seroconversion and gut microbiome, there is only limited evidence of changes in microbiome profile that are correlated with seroconversion. Based on the previous findings and the aim of this study, we include nationality, age, seroconversion and the interaction between age and nationality into our Bayesian model as fixed effects. We want to first check whether we can recover the results in the literature about nationality and age. Furthermore, we want to see if truly there is not enough information to determine the relationship between seroconversions and microbial distributions in human gut. We ran our MCMC sampler for 100,000 iterations with a burn-in of 20,000 and collected every 50th sample to thin the chain.

2.5.1 Estimating the effect of age

We first estimated the trend over age of the microbiome profile. In particular, we check the derivatives of species abundances over age for each individual subjects and the population average trend over age when controlling the seroconversion status to be negative. In Figure 2.4, we showed two examples with the most significant association with age, Bifidobacterium and Bacteroides. We plotted the estimated derivatives and population average curves along with the observed relative abundances for both genera.

The derivative curve for Bifidobacterium is significantly smaller than zero for most of the samples, indicating a consistent decreasing trend of this genus in infants' gut microbiome, which is in accordance with the underlying biology since bacteria from this genus

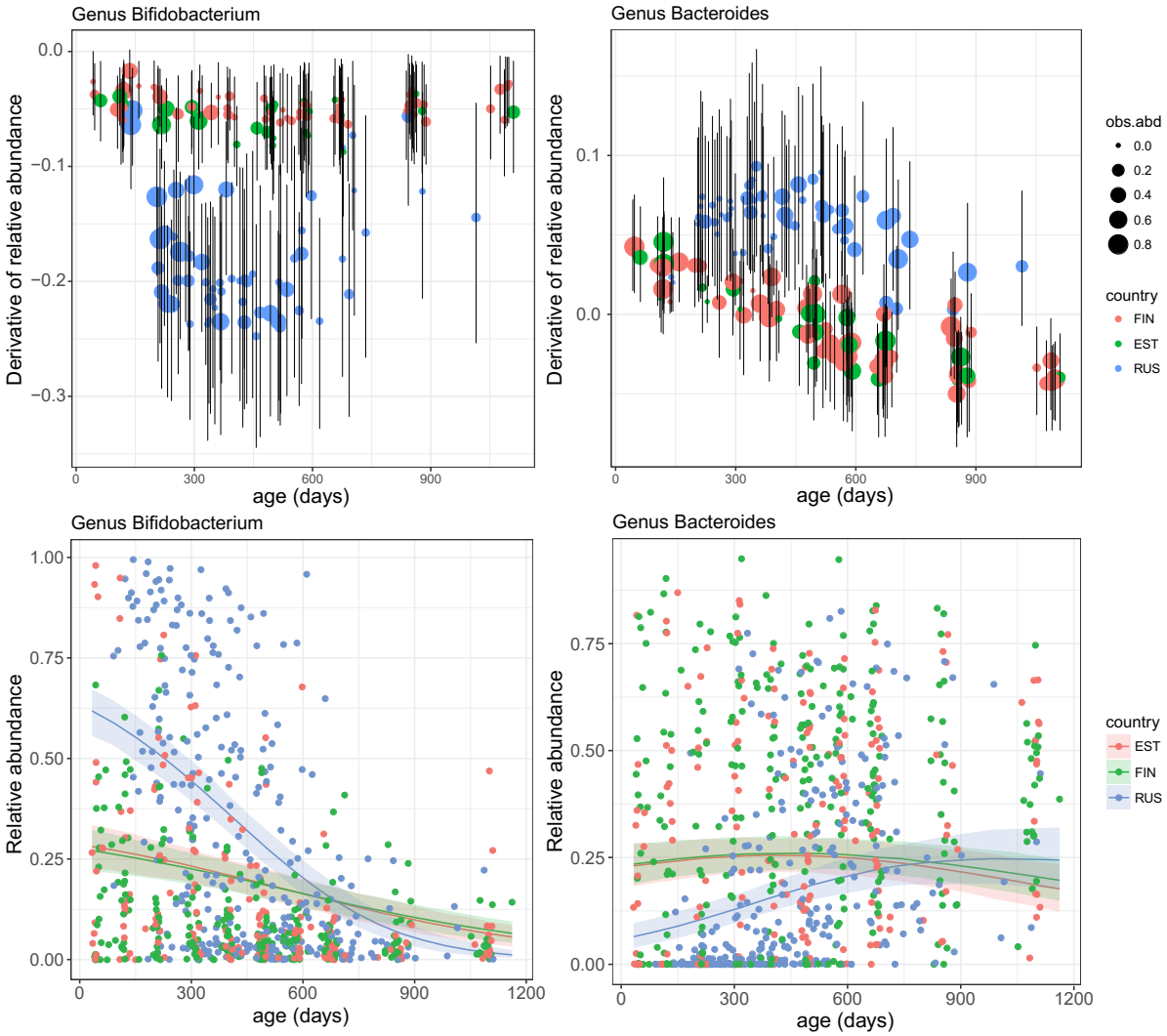


Figure 2.4: (**Top**) Derivatives of probabilities of two genera with respect to age, colored by nationality. We only show a random subset of 150 samples. The error bars indicate the point-wise 95% credible intervals of the estimates. (**Bottom**) Estimated population average trend of species abundance, colored by nationality. We hold the seroconversion status to be negative for all three curves. The shaded areas illustrate the point-wise 95% posterior confidence bands of all three curves.

are associated with breastfeeding. This decreasing trend is confirmed by the population average curves, which show very similar trajectories for Finnish and Estonian populations. The curve for Russian infants exhibits a more extreme decreasing trend at early age (before 600 days old), which is consistent with the significantly larger derivatives observed in Russian samples. Indeed, this difference between Russian and the other two populations is already reported in Vatanen et al. (2016) and can be recognized directly from the relative abundance data.

The trend associated with genus Bacteroides is not as strong as in the case of Bifidobactrium. The derivatives of this genus are positive with high credibility when age is smaller than 300 days for all three populations and become slightly negative when infants are older in only Estonian and Finnish populations. The Russian infants instead have slight increasing trend even at older age. The population average curves replicate these behaviors and it is clear that the increasing trend is more extreme in Russian overall. The real data also supports what we found here. Specifically, when age is larger than 900 days, there are observable decrease for both the Estonian and Finnish samples, although the decrease is relatively small. In addition, Russian samples tend to have very low abundance Bacteroides at early age (¡150 days) and pick up the genus rapidly after 300 days. We would like to point out that there are only a few Russian samples after age 900 days. This means the leveled-out trend in the population curve is mainly an extrapolation based on the curve before 900 days and might not reflect the underlying population trend accurately.

2.5.2 Estimating the effect of nationality

The nationality of the infants supposedly introduces a strong effect on the gut microbiome. Association between nationality and microbiome profile is reported in Vatanen et al. (2016) and for some genera, the effect is strong. In our model specification, the difference in countries is captured by a constant shift as well as a linear function of age due to the interaction term in the latent scale $\mathbf{Q}$. Therefore the changes of probability of species associated with nationality will depend on age. As a result, we check the average changes of probabilities over five consecutive age windows and explore possible age win-

dows corresponding to strong country effect. In Figure 2.5, we plot the estimated changes associated with nationality for two genera, Bacterioides and Bifidobactirum, which are previous identified as significantly less abundance and more abundance in Russian respectively.

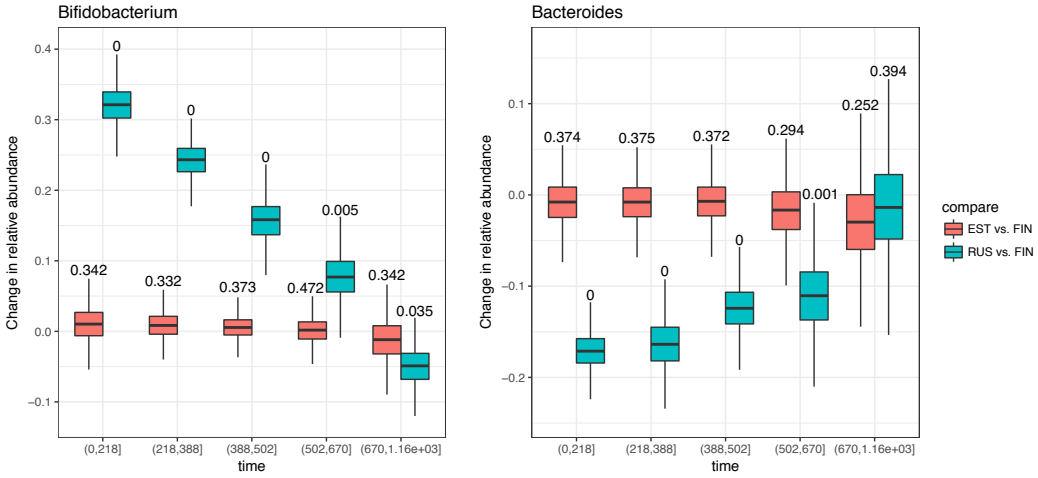


Figure 2.5: Estimated average difference in probabilities of two genera in five consecutive age groups when comparing Russian and Estonian to Finnish infants. The boxplots indicate the posterior distributions of the corresponding differences and the numbers on top of each boxplot is the one-sided posterior probability of zero.

From the result we can see clearly there is a strong depletion of Bacteroides in Russian infants comparing to Finnish infants. This depletion diminishes when infants get older. At the last age window (670-1160 days), the posterior credible interval covers zero with high probability, indicating the two populations have very similar abundance profiles for Bacteroides. There is no significant difference between Estonian and Finnish infants throughout the five age windows. On the other hand, we observe a large increase in the abundance of Bifidobacterium in Russian infants while Estonian and Finnish infants still possess similar abundance profiles. The increase observed in Russian infants again diminishes when the infants get older and completely disappear at the last age window. These results are aligned with the discoveries in Vatanen et al. (2016). Since our Bayesian method evaluate the uncertainty of the estimates using all data available, the significant results given by the model should reflect the underlying patterns in this cohort and are interesting for further validations and investigations from biological perspectives.

2.5.3 Estimating the effect of seroconversion

We further explored microbiome associations with seroconversion, specifically the presence of at least one autoantibody, as reported in Vatanen et al. (2016). We calculated the posterior distributions of the difference of average abundances of seroconverted samples and controls when holding other covariates to be constant. We also stratified the analysis by age groups, since even though there is no interaction between seroconversion and age on the latent scale, normalization to relative abundances could introduce potential interactions. Among the 55 genera analyzed by the resulting model, the most significant was Lachnospiraceae at age window (0, 218) days with one-sided posterior probability of zero at 0.176. The results thus indicate no evidence of genus-level associations with seroconversion on a population scale. This may be due to associations at a more specific level within the microbiome (e.g. species, strain, or functional elements), or even this more nuanced model may be under-powered to recover associations given the extensive inter-individual variation (only 37 out of 762 samples were seroconverted).

2.5.4 Relationship between species

As a by-product of our model, we can examine the cross-sectional relationship between species by using the correlation between species vector $\mathbf{X}$. The interpretation of this correlation matrix is the similarity between species in samples with fixed covariate values. Since a natural way to describe the relationship between microbial genera is by phylogenetic tree. We sorted the estimated correlation matrix based on the phylogenetic distances such that rows or columns that are close to each other belong to genera with small phylogenetic distance. We further evaluated the uncertainty of the patterns observed in the posterior mean estimate of the correlation matrix. An simultaneous ordination analysis of multiple correlation matrices is used (Escoufier, 1973) to visualize the posterior uncertainty of the correlation matrix between species. The results are summarized in Figure 2.6.

From the figure we can see the correlation matrix estimated using the latent factors $\mathbf{X}$ reveals more structure than the raw correlation. It also removes the apparent correla-

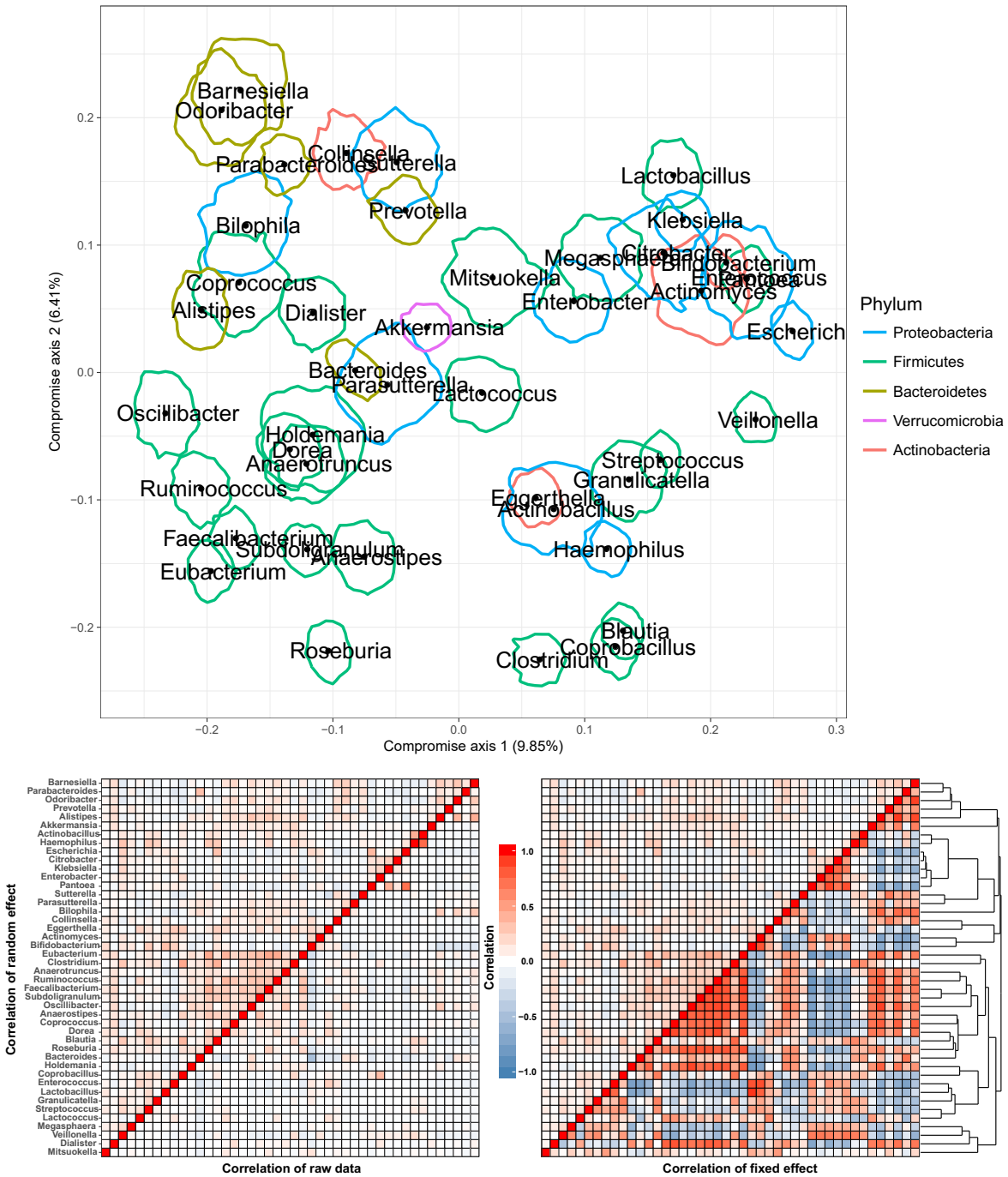


Figure 2.6: Cross-sectional relationship between species. (**Top**) Ordination of species based on the correlation matrix given by $\mathbf{X}$. The contour lines indicate the boundaries of the posterior credible regions of the ordination configuration. (**Bottom**) Correlation matrix between species estimated by $\mathbf{X}$ compared to raw correlations between observed relative abundances (left) and correlation between fixed effects (right). The rows and columns are reordered such that adjacent rows or columns are close in the phylogenetic tree. The phylogenetic tree for these genera are plotted at the right side of the figure.

tion introduced by fixed effects. We observed three clusters of genera in the estimated correlation matrix and the ordination results suggests genera in the same cluster tend to have overlapped credible region of their ordination coordinates. All of these clusters are formed by phylogenetically related genera as confirmed by the phylogenetic tree we showed in the figure. The first and largest cluster is formed by 13 genera from phylum Firmicutes (Clostridium, Ruminococcus, etc). This cluster corresponds to the cluster at the bottom-left corner of the ordination plot. The second cluster is formed by seven genera in Proteobacteria. They are mostly accounted for by the cluster at the upper-right corner of the ordination. The last cluster includes five genera in Bacteroidetes, which occupies the cluster at the upper-left corner of the ordination plot. We want to point out that the correlation estimated here is not only restricted to demonstrate relationships between microbial taxa. In general, since our model can be applied to any count tables that arise from multinomial sampling, we can visualize the relationships between genes, proteins and metabolites by using the appropriate types of data.

2.6 Conclusion

We propose a mixed effect Bayesian factor model to perform multivariate regression analysis for microbiome data. This regression analysis estimates the effect of covariate on microbial composition while allowing for flexible correlation structure between the residuals. Under regularity conditions, we proved the model parameters are identifiable, which indicates with enough amount of data, the posterior estimates can be close to the truth. We verified this by numerical experiments. By appropriate transformation, the model parameters can be converted into interpretable and biologically interesting results. These transformations are further proved to be well-defined by simulation studies as the truth can be accurately recovered by the estimates. We finally applied this model on a longitudinal microbiome dataset and replicated results that are reported in existing literatures.
We would like to point out at least two major limitations of this model. The first one has to do with the computation cost. The current posterior computation is implemented via a Gibbs-sampler, which can be inefficient especially when the number of parameters are

large. This is usually the case when the numbers of species and biological samples are large. In this scenario, the mixing of the posterior samples deteriorates and the computational cost for each iteration increases. A more efficient algorithm, such as Hamiltonian Monte Carlo, which jointly samples all parameters, can help with the mixing speed whereas variational inference algorithm can be useful to reduce the computational cost per iteration. The other limitation is more about the scenario where the model might behave poorly. Specifically, if the underlying truth dictates most of the species to have similar trends over a continuous covariate, the inference given by the model might be unstable since the observed data, which reflects the normalized latent variables, might exhibit little to none trend. The inference will then be determined mostly by the prior and might give meaningless results.

There are also several future directions which we want to explore. We would like to investigate appropriate variable selection techniques for the fixed effect. This is especially helpful when there are a large collection of covariates and no prior knowledge is known with regard to the important ones. A more flexible model for the fixed effect is also desirable. Currently, the relationship between the abundances of species and the covariates are depicted by a linear function in the latent scale, which indicates the model only captures continuous trends over continuous covariates. In microbiome studies, there are scenarios where disruption of microbial profiles happens. In these circumstances, a step function or a wavelet bases might be more relevant even in the latent scale. Finally, we notice that the number of regression coefficients $\mathbf{v}_i$ increases as the number of species increases and the current prior on $\mathbf{v}_i$ does not regularize these coefficients. This prior specification ignores potential relationship between $\mathbf{v}_i$ imposed by prior knowledge of the relationship between the corresponding species. A systematic way to incorporate such information on $\mathbf{v}_i$ will be helpful in efficiency of the estimates.

A Hierarchical probabilistic model of microbial community structure

Boyu Ren

Department of Biostatistics

Harvard Graduate School of Arts and Sciences

Yo Sup Moon

Department of Biostatistics

Harvard Chan School of Public Health

Emma Schwager

Department of Biostatistics

Harvard Chan School of Public Health

Timothy L. Tickle

Broad Institute of MIT and Harvard

Yiren Lu

Department of Biostatistics

Harvard Chan School of Public Health

Eric A. Franzosa

Department of Biostatistics

Harvard Chan School of Public Health

Curtis Huttenhower

Department of Biostatistics

Harvard Chan School of Public Health

3.1 Introduction

Understanding the structure and interactions of host- and environmentally-associated microbial communities is a fast-growing focus in biological and biomedical research. While work in this area has benefitted immensely from advances in high-throughput DNA sequencing, extracting biologically relevant trends from these data requires carefully designed statistical methods. To date, a number of statistical methods have been proposed to derive biological trends from metagenomic sequencing data, including ecological associations among community taxa (Faust et al., 2012b; Friedman and Alm, 2012; Fang et al., 2015), as well as associations between community features and sample metadata (Xia et al., 2013b; Chen and Li, 2013; Lin et al., 2014; Paulson et al., 2013b). However, it is challenging to benchmark these methods in a systematic manner: a critical limitation for researchers who must identify and choose among appropriate tools for executing their particular microbial community studies. Moreover, the historical absence of a standard evaluation framework for statistical analyses of the microbiome has likely hindered methods development in the area.

One means for carrying out statistical methods benchmarking is data simulation. When working with simulated data, the true signals and structures of the data are known, thus enabling objective evaluation of a method's performance (for example, true positive and false positive rates). However, for these evaluations to be meaningful, the simulation must provide a very close approximation of the underlying biology. This frequently requires the use of a sophisticated statistical model. Such models have been previously described in the context of gene expression profiles analysis (Smyth, 2005; Anders and Huber, 2010; Robinson et al., 2010). Indeed, many models specific to gene expression data simulation are currently available (Van den Bulcke et al., 2006; Hoops et al., 2006; Long and Roth, 2008), including stochastic, deterministic, and network-guided approaches (Van den Bulcke et al., 2006; Hoops et al., 2006), as well as alternative approaches focused on introducing predefined variability within simulated expression data [which are particularly well-suited to sensitivity analysis (Long and Roth, 2008)]. Collectively, these methods have proven useful for exploring gene expression patterns and mechanisms, as

well as validating statistical approaches to the analysis of gene expression data (Meyer et al., 2007; Li et al., 2010; Carrera et al., 2009).

Much as biological models of gene expression data benefit statistical methods development in transcriptomics, we expect a standard model of microbial community profile structure to be similarity useful in microbiome analysis. Moreover, some features of gene expression data and their analysis are shared with metagenomic profiles and analyses, and thus aspects of existing gene expression models may be directly transferable to the microbiome field. For instance, for both metagenomic and gene expression data, there is an interest to simulate measurements for a large number of features across a large number of synthetic samples, while simultaneously 1) insuring that the distributional properties of the features and samples are biologically realistic, and 2) enforcing specific relationships among individual features. However, metagenomic sequencing data also possess unique properties that complicate direct application of gene expression models to microbiome analysis. For example, relative to gene expression data, metagenomic sequencing data tends to be more variable and considerably more sparse (Li, 2015b): properties that must be taken into account when building a microbiome-specific standard for statistical analysis.

A small number of statistical models have been proposed to specifically model the properties of taxonomic profiles derived from metagenomic sequencing, including models based on the Dirichlet-Mutinomial (DM) and Zero-inflated Gaussian (ZIG) distributions. One DM-based approach (Chen and Li, 2013) utilized a hierarchical structure to share information between taxonomic features and generate jointly multinomial count data, but failed to capture the sparsity of real-world microbiome datasets. An alternative ZIG-based approach (Paulson et al., 2013b) addressed the sparsity issue and was less computational demanding to fit, but did not consider dependence among microbes. Moreover, while the ZIG-based model remains useful for simulating normalized microbiome data (i.e. with effects of sample-specific sequencing depth removed), it does not prescribe a fully interpretable generative process for raw metagenomic count data. Notably, the information lost in the process of converting raw to normalized microbiome data has been proposed to contribute to suboptimal downstream analysis under some circumstances (McMurdie

and Holmes, 2014b), which suggests further limitations to the existing ZIG-based model as an aid in benchmarking metagenomic analysis methods and results.

Motivated by the lack of comprehensive statistical models for explaining microbial taxonomic profiling data, we developed a statistical model that we implement and evaluate here as SparseDOSSA (Sparse Data Observations for the Simulation of Synthetic Abundance). The method models the marginal distributions of metagenomic features using zero-inflated log-normal distributions, the parameters of which are hierarchically linked to a parent log-normal distribution (which governs the global structure of the dataset). This approach allows us to recapitulate both the variability and sparsity of real-world metagenomic sequencing datasets. Moreover, SparseDOSSA serves as a generative model for raw metagenomic count data, making its outputs appropriate for the analysis and benchmarking of a wide variety of experimental designs and statistical methods. We demonstrate that SparseDOSSA can be fit using both fully Bayesian and nave empirical approaches. Both approaches are successful in producing synthetic metagenomic datasets that recapitulate the structural properties of target real-world datasets. In addition, SparseDOSSA can induce a correlation structure 1) among pairs of metagenomic features and 2) between metagenomic features and sample metadata. This option makes SparseDOSSA a powerful aid for evaluating microbiome-focused statistical methods, which we demonstrate in a series of literature-based benchmarking applications. SparseDOSSA is open source and available for download from http://huttenhower.sph.harvard.edu/sparsedossa.

3.2 Methods

The SparseDOSSA generative model has three main components: 1) estimation of model parameters for a null feature matrix, 2) generation of the null feature matrix, and 3) spiking associations into the null feature matrix. Below, we introduce the terminology used throughout the remainder of this section, and then describe these methodological steps in detail. A feature matrix with I microbial community features (represented as rows) and J biological samples (represented as columns) is denoted as $\mathbf{Y}$ and each of its entries take

an integer (count-like) value; i.e. $Y_{i,j} \in \mathbb{N}$. The row corresponding to feature i is represented as $\mathbf{Y}_i$; similarly, the column corresponding to biological sample j is as $\mathbf{Y}^j$. A null feature matrix $\mathbf{Y}^{(0)}$ is defined as a feature matrix that lacks statistically significant correlation structure among its component features and (optionally) with a given metadata matrix. A metadata matrix is a feature matrix with J columns (corresponding to the J biological samples from the null matrix) and K rows corresponding to metadata features; metadata features may be continuously valued or discrete. A final spiked feature matrix $\mathbf{Y}^{(1)}$ is generated by inducing correlation structure among the rows of the null matrix or between rows of the null matrix and metadata features.

3.2.1 Model for the null feature matrix

We employ a two-layer hierarchical model to capture the general patterns in microbial community measurements. As a result, the null feature matrix can be simulated by this model as it guarantees independence between microbiome features and between microbiome features and sample metadata. The model is specified as following

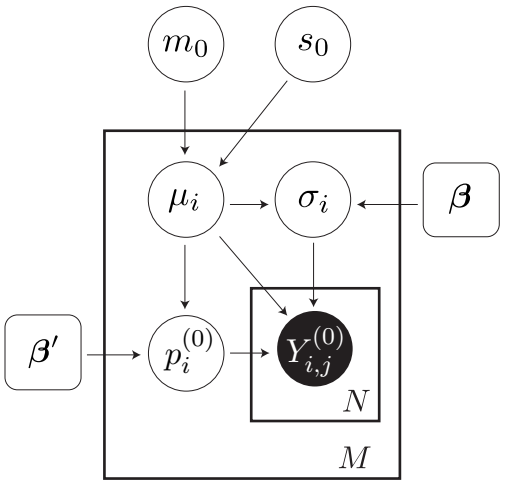
$$\mu_i \overset{iid}{\sim} \text{Lognormal}(m_0, s_0^2)$$
$$Y_{i,j}^{(0)} \overset{iid}{\sim} (1 - p_i) \left[\text{Lognormal}(\mu_i, \sigma_i^2)\right] + p_i \delta_0.$$

Here μ_i is the parameter for the average abundance of feature i, whose distribution is determined by the hyperparameters m_0 and s_0. The distribution of $Y_{i,j}^{(0)}$ is assumed to be a mixture of a lognormal and a point mass at zero with mixture probabilty p_i. We further assume σ_i and p_i are both functions of μ_i:

$$\log(\sigma_i) = \beta_0 + \beta_1 \log(\mu_i)$$
$$\text{logit}(p_i) = \beta_0' + \beta_1' \log(\mu_i).$$

In this model, the first layer controls the distribution of the means of all features. The second layer controls the distribution of abundances of a given feature across samples. Based on observations from microbial community sequencing data, we adopted a lognormal distribution for the first layer and a zero-inflated lognormal distribution for the

second layer. The lognormal distribution captures the high dynamic range of microbial community and can be served as a computationally efficient approximation to the negative-binomial model when zero-inflation in added. Note the negative-binomial distribution is widely accepted for counts data in genomic studies (Anders and Huber, 2010; Robinson et al., 2010). The zero-inflated flavor of this distribution helps to distinguish two biologically important interpretations of a zero measurement: 1) a feature that is truly absent versus 2) a feature that is present, but with abundance below the limit of detection (i.e. insufficient to produce a count of at least 1). By visualizing the relationship of log-transformed marginal mean and marginal standard deviation of species abundances, we considered a simple linear model is sufficient. This assumption is further supported by formal tests on three real datasets. A similar goodness-of-fit test is performed for the logit model for the relationship between marginal sparsity and marginal mean abundance and the results supported the choice. The model specification is also visualized as a plate graph in Figure 3.1.



m_0, s_0: parameters of distribution of μ_i
μ_i: mean parameter of ith feature
σ_i: standard deviation parameter of ith feature
$p_i^{(0)}$: sparsity parameter of ith feature
$\boldsymbol{\beta}$: parameters linking σ_i to μ_i
$\boldsymbol{\beta}'$: parameters linking $p_i^{(0)}$ to μ_i
$Y_{i,j}^{(0)}$: counts for feature i in sample j
N, M: number of samples and features

$$\mu_i \overset{i.i.d}{\sim} LN(m_0, s_0)$$
$$\log(\sigma_i) = \beta_0 + \beta_1 \log(\mu_i)$$
$$\text{logit}(p_i^{(0)}) = \beta_0' + \beta_1' \log(\mu_i)$$
$$Y_{i,j}^{(0)} \overset{i.i.d.}{\sim} p_i^{(0)}\delta_0 + (1 - p_i^{(0)})\lfloor LN(\mu_i, \sigma_i)\rfloor$$

Figure 3.1: **SparseDOSSA provides a generative hierarchical Bayesian model for microbial community taxonomic profiles**. Individual microbial features are assumed to be drawn from a zero-inflated lognormal distribution with marginal mean μ_i and standard deviation σ_i with feature-specific sparsity (i.e. expected fraction of zeros) p_i. μ_i also follows log-normal distribution with parameter m_0 and s_0. σ_i and p_i are determined by μ_i through parameters $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$. Rectangles denote replication of the model, i.e. plate structure, with the number of the replication labeled at the bottom-right corner.

3.2.2 Calibration on real microbial community measurements

Our goal is to simulate realistic microbial community count-based abundance data, similar to those typically derived from marker gene-based surveys. To that end, we calibrate the parameters of the model described in the previous section by fitting to real microbial community datasets (referred to as template datasets). We have implemented two methods to perform the fitting: a naive stepwise maximum likelihood estimation (MLE) method and a fully Bayesian method. The nave method estimates μ_i, σ_i, p_i using MLE based only upon observations of individual feature i (see Figure 3.1 for more details). It is only used for large datasets with many samples. Specifically, the method can only achieve reasonable performance when the total number of reads per sample is much larger than the number of OTUs and the OTUs are grouped to genus or higher taxonomic levels. $\boldsymbol{\beta}, \boldsymbol{\beta}'$ are then estimated using linear regression and logistic regression with estimated σ_i and p_i as responses and μ_i as covariate. The hyperparameters m_0 and s_0 are directly estimated by fitting the density function of lognormal distribution to the collection of all estimated μ_i. The detailed steps are listed below:

1. Calculate the marginal means and standard deviations of log-transformed non-zero counts for each feature ($\widehat{\mu}_i$, $\widehat{\sigma}_i$) and the fraction of zeros of each feature ($\widehat{p}_i$) directly from the data. Denote the vectors formed by feature-specific $\widehat{\mu}_i$, $\widehat{\sigma}_i$, and $\widehat{p}_i$ values as $\widehat{\boldsymbol{\mu}}$, $\widehat{\boldsymbol{\sigma}}$ and $\widehat{\mathbf{p}}$.

2. Fit a log-normal distribution on $\widehat{\boldsymbol{\mu}}$ to obtain the estimates of the mean and standard deviation parameters ($\widehat{m}_0$ and $\widehat{s}_0$) for the distribution of feature-specific marginal mean parameters.

3. Perform a linear regression of $\widehat{\boldsymbol{\sigma}}$ on $\log(\widehat{\boldsymbol{\mu}})$ to obtain the ordinary least square (OLS) estimates of regression coefficients $\widehat{\boldsymbol{\beta}}$.

4. Perform a logistic regression of $\widehat{\mathbf{p}}$ on $\log(\widehat{\boldsymbol{\mu}})$ to obtain the OLS estimates of regression coefficients $\widehat{\boldsymbol{\beta}}'$.

The alternative, fully Bayesian method is performed using RStan (Carpenter et al., 2016) and this is the default fitting procedure for the SparseDOSSA model. We assume a

$MVN(\mathbf{0}, \mathbf{I}_{2\times 2})$ prior for $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ and a non-informative prior on m_0 and s_0. RStan implements a no-U-turn Hamiltonian Monte Carlo Markov Chain algorithm to sample from the posterior distribution and is highly efficient for hierarchical models (Hoffman and Gelman, 2014). We run three parallel chains with 5,000 total iterations and a burn-in period of 1,000 iterations. The convergence of the HMC algorithm is examined first by the traceplots and then by the Rhat statistics (Gelman and Rubin, 1992a). We conclude the convergence is achieved when the traceplots of the three parallel chains mix well and the Rhat statistics is smaller than 1.1. The final estimates of the parameters are taken as the means of respective posterior medians over three parallel chains. We preserve the full set of posterior samples to estimate the posterior predictive distribution of the data.

3.2.3 Validation datasets

To validate SparseDOSSA's model for null matrices, we explored microbiome data from two human body sites (stool and vagina) using shotgun metagenomes downloaded from the HMP DACC (http://hmpdacc.org) and taxonomically profiled using MetaPhlAn2 v2.5.0 (Truong et al., 2015), plus one additional published dataset from patients with inflammatory bowel disease [the PRISM dataset (Morgan et al., 2012c)]. HMP species were summarized at the genus level, resulting in 553 stool profiles spanning 312 genera and 234 vaginal samples with 156 genera. The PRISM dataset contained 250 samples and 158 genera (after summarizing OTUs).

When comparing real and simulated samples, it was necessary to pair real microbial features (i.e. taxa, genera, or OTUs) with simulated features (which do not have any intrinsic taxonomy). To do this, we matched the ranks of the real and simulated features' marginal means: the real taxon with the ith largest marginal mean was paired with the simulated feature with the ith largest marginal mean. To validate this pairing in e.g. Figure 3.2(b), we split the real dataset into training and testing sets containing equal numbers of samples, and produced a new synthetic dataset fit to the training data only. We applied the pairing scheme for both the synthetic dataset and the testing set. We then jointly ordinated all three datasets. As expected, the small region of deviation occurred only in the deviated region, while the synthetic and (independent) validation samples showed ex-

hibited closer agreement.

3.2.4 Generating null matrices

Null feature matrix generation. If using the naive fitting method, we set the values of the parameters at their estimated values. If using the Bayesian fitting method, we take a random draw of all model parameters from the estimated posterior distribution. Once a set of parameter values is established, we follow the generative process defined by the hierarchical model in Figure 3.1 to generate the null feature matrix.

Metadata matrix generation. To generate the metadata matrix $\mathbf{X}$ with K metadata features, half of the K features are set as continuous and half as categorical. For the continuous case where $k \in \{1, \ldots, \lfloor K/2 \rfloor\}$, we draw $X_{k,j} \overset{iid}{\sim} N(0,1)$ for $j = 1, \ldots, J$. For the discrete case where $p \in \{\lfloor K/2 \rfloor + 1, \ldots, K\}$, we define metadata features as quadrant variables and draw $X_{k,j} \overset{iid}{\sim}$ Multinomial$(1, \Theta)$, where Θ is set as (0.25,0.25,0.25,0.25). This probability vector can be changed by user.

3.2.5 Building association patterns

One run of SparseDOSSA can only produce a dataset with either feature-feature correlation or feature-metadata correlation, but not both. We followed this convention to accommodate the fact that most of the statistical tools developed in microbiome analysis only focus on one of the correlation structures.

Feature-feature correlation. Correlation structure between features of the is introduced by modifying the null feature matrix generation algorithm. Specifically, assume the correlation matrix between a set of M_c features is denoted by $\mathbf{S}_c$, the mean parameters μ_i for these M_c features will be simulated using a multivariate lognormal distribution with mean parameter $m_0 \mathbf{1}_{M_c}$ and covariance matrix $s_0^2 \mathbf{S}_c$. The rest of the μ_i's will be simulated independently from Lognormal(m_0, s_0^2). Matrix $\mathbf{S}_c$ is user-specified.

Feature-metadata correlation. Introduction of correlation structure between features and metadata is done by additively spiking in signal from the metadata features into the data (microbial) features. We carry out a standardization procedure for both features and metadata to ensure the counts of the modified feature are not dominated by the values of the

target metadata but rather distributed similarly to real data. With the exception of samples with zero reads for the given feature, the standardized abundances of other samples are added by a linear combination of all correlated standardized metadata. We repeat the following procedure M_f (user-defined variable) times to introduce M_f modified features:

1. Randomly choose a feature vector $\mathbf{Y}_i^{(0)}$ from the simulated null matrix and pick K_f metadata without replacement.

2. For each of the K_f metadata, calculate the marginal mean $\mu_X^{(k)}$ and standard deviation $\sigma_X^{(k)}$.

3. Calculate the mean $\mu_Y^{(i)}$ and standard deviation $\mu_Y^{(i)}$ of the non-zero samples from the chosen feature vector $\mathbf{Y}_i^{(0)}$. 4. For the non-zero samples of vector $\mathbf{Y}_i$, set $Y_{i,j}^{(1)} = \left\lfloor (Y_{i,j}^{(0)} + \phi \sum_k \omega_{k,i,j})/(K_f\phi + 1) \right\rfloor$, where $\omega_{k,i,j} = (X_{k,j} - \mu_X^{(k)})\sigma_Y^{(i)}/\sigma_X^{(k)} + \mu_Y^{(k)}$. The index k is iterated over all covariates generated by K_c selected metadata, where for categorical metadata, multiple dummy variables will be included and the constant ϕ is a real-valued strength parameter. The zero-read samples of vector $\mathbf{Y}_i$ are unchanged.

3.3 Results

We developed SparseDOSSA as a hierarchical Bayesian model capable of both describing existing metagenomic sequencing data using a small number of parameters andonce fitsimulating new raw (count-based) taxonomic profiles that recapitulate real-world datasets (with optional spiked-in correlation structure; Figure 3.1). We evaluated two approaches for fitting this model: a fully Bayesian approach and a nave empirical approach (see Methods). We validated SparseDOSSA's performance under the two model-fitting schemes by comparing feature- and sample-level ecological properties between simulated and real datasets. Finally, we demonstrated the method's utility as a statistical benchmarking tool in applications involving microbial biomarker discovery.

3.3.1 SparseDOSSA accurately models global microbial abundance patterns

We first assessed the degree to which SparseDOSSA's fitted two-layer model (Figure 3.1) captured the marginal variation of microbial community taxonomic profiling data across simulated microbial features. Specifically, we focused on a real-world dataset consisting of 158 microbial genera quantified across 250 human gut microbiome samples derived from patients with Inflammatory Bowel Disease [the "PRISM" dataset (Morgan et al., 2012c)]. Using SparseDOSSA, we fit two models to this dataset: one using a fully Bayesian approach and a second using a nave empirical approach. Each model was then used to simulate a synthetic dataset with the same dimensions as the PRISM dataset (158 features in 250 samples). Subsequent comparisons of the real and simulated datasets serve to both benchmark the appropriateness of SparseDOSSA's underlying hierarchical model, as well as to compare the relative performance of the two model-fitting methods.

Rank-average abundance of microbial taxa is frequently applied as a quantitative description of microbial community structure (Börnigen et al., 2013); hence, to be useful, a model fitted to real metagenomic data should produce simulated communities with rank-abundance distributions similar to those of the template community. Indeed, we observed strong agreement in rank-average abundance for the real versus SparseDOSSA-simulated PRISM datasets, as inferred from a high degree of overlap between the estimated trends [Figure 3.2(a)]. Furthermore, the 95% point-wise posterior credible interval given by fully Bayesian fitting of the same model contains both curves, with the Bayesian fitting being a more comprehensive but computationally expensive approach (see Methods). This credible interval suggests that the model SparseDOSSA assumed accurately reproduces the marginal distribution of microbe-specific abundances of a real microbial community. Moreover, the simulated data from the nave method agree with the posterior predictive results from the full Bayesian method, suggesting that the nave fitting approach is an accurate and efficient approximation of the fully Bayesian method. The minor differences between the two rank-abundance distributions lie largely in the left tail of the distributions (i.e. among rare, low-abundance organisms that are likely to be

stochastically sampled even during the original biological sampling process). We observed similar patterns of relative performance when fitting on two additional real-world template datasets: human gut and vaginal metagenomes sampled during the Human Microbiome Project (HMP) (Turnbaugh et al., 2007) (Figure A.7). These results speak to the generality of SparseDOSSA's underlying model, as well as the effectiveness of the nave fitting approach.

We further evaluated SparseDOSSA's ability to recapitulate population-level patterns of ecological similarity (beta diversity) between samples [Figure 3.2(b)]. We generated a synthetic microbiome dataset by fitting SparseDOSSA to a template dataset of 234 HMP vaginal samples using the nave fitting method and the Bayesian method. The vaginal microbiome is known to have a highly structured ecological ordination configuration (Ravel et al., 2011b). We choose to compare with this dataset since it is visually more clear to check if the simulated data produces the same ordination configuration. In order to compare real and simulated samples required first aligning the real and simulated features between the datasets, which we accomplished by comparing the ranks of the real and simulated features' marginal means (i.e. the real taxon with the ith largest marginal mean was paired with the simulated feature with the ith largest marginal mean). Joint ordination of the real and simulated datasets revealed similar structural patterns ([Figure 3.2(b)], suggesting that, in addition to generating synthetic microbiome datasets with realistic feature-level behavior, SparseDOSSA's model also captures population-level structural phenomena among samples.

3.3.2 Modeling correlation structure between taxonomic features and sample metadata

When used as a model for simulating realistic microbial community data, an important feature of SparseDOSSA is the ability to impose known correlations between simulated microbial features and sample metadata. This is done by first capturing the overall diversity pattern of a community and then subsequently inducing the artificial correlations. Our hierarchical model allows the introduction of associations between simulated microbial features and 1) categorical sample properties (e.g. health/disease status or soil type)

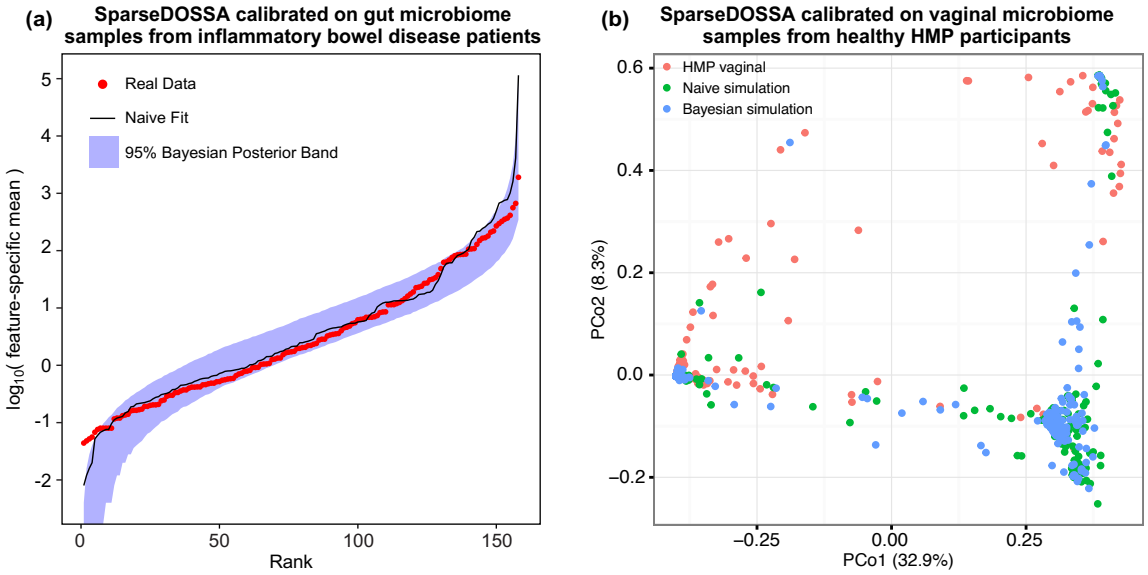


Figure 3.2: **The SparseDOSSA model accurately captures feature mean distributions and beta diversities of microbial communities**. (a) Rank abundance of log-transformed feature-specific means for a template dataset (PRISM) and simulated datasets based on fitting models to the template using nave versus fully Bayesian approaches. (b) MDS analysis on Bray-Curtis distance of real taxonomic profiles from 234 vaginal microbial communities from the Human Microbiome Project (HMP) and two sets of 234 simulated samples from the resulting SparseDOSSA model using the nave fitting method and the Bayesian method respectively (all containing 156 microbial features).

and/or 2) continuous sample properties (e.g. subject BMI or soil pH). This is accomplished by adding a linear combination of associated metadata values to the raw values of a target feature (see Methods).

We verified SparseDOSSA's ability to include detectable categorical feature-metadata associations in its simulated output by artificially associating nine randomly selected features with a binary sample property, focusing on the previously described PRSIM dataset as a training template. For the convenience of visualization, we plot the difference in average relative abundance across all samples between the two levels of the binary metadatum [Figure 3.3(a)]. The nine (true positive) synthetic associations are detected among the 17 features of greatest effect size (as measured by the standardized by the overall standard deviation). Notably, for each spiked-in feature, the approximated 95% confidence interval for the difference in means does not cross zero, implying a statistically significant effect size (p-value < 0.05). We also observed several false positives, which we attribute not to the SparseDOSSA model but rather the known effects of compositionality (i.e. relative abundance normalization) in ecological data (Pearson, 1896).

Simulated associations involving continuous sample properties were similarly successful [Figure 3.3(b)]. Starting from a synthetic dataset trained on the PRISM dataset, one feature was modified to be correlated with a randomly generated continuous sample metadatum. By design, SparseDOSSA only manipulates samples with non-zero counts when creating artificial correlation patterns, thus correctly preserving phenotypic associations in a zero-inflated manner. This is useful as zero-inflation reflects the sparsity of real microbiome datasets. Statistical analysis algorithms that fail to capture this behavior will have substantially lower statistical power in detecting associations.

3.3.3 Simulating controlled correlation structure among modeled microbial features

In addition to modeling associations between microbial features and sample properties (metadata), SparseDOSSA provides a way to model ecological associations among microbial features. In order to validate synthetic manipulation of feature-feature associations using our hierarchical model, we produced a synthetic dataset trained on the real-world

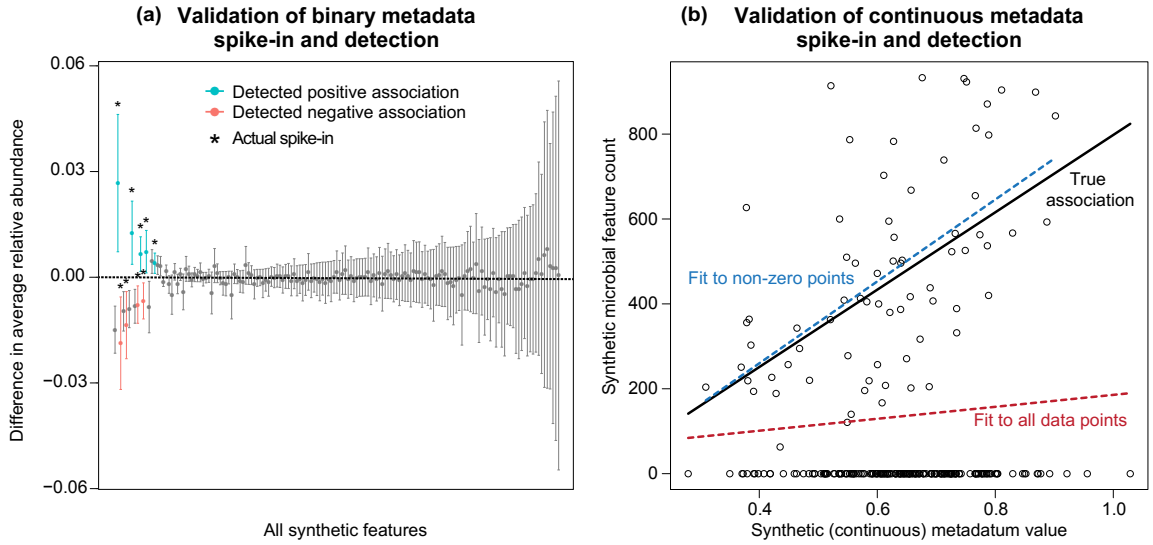


Figure 3.3: **Simulating categorical or continuously valued population variability among microbial community samples**. (a) Differences in mean relative abundances between two classes of a simulated binary sample property (metadatum) along with the empirical inter-quartile range of all features as contrasted between metadatum levels. (b) Correlation of one feature into which an association to a continuously varied sample metadatum has been spiked (Y-axis) with that metadatums value (X-axis).

250-sample PRISM dataset. Notably, the PRISM dataset includes several strong correlations among its component features [absolute Pearson's correlation larger than 0.5; Figure 3.4(a), above the diagonal]. We randomly selected 50 synthetic features and verified that, among the raw synthetic data, no two were strongly correlated [Figure 3.4(a), below the diagonal]. Next, we targeted ten random pairs of simulated microbial features and introduced synthetic positive correlations between these pairs [Figure 3.4(b), above the diagonal]. The pattern of measured correlation among the modified synthetic data was concordant with the intended correlation structure [Figure 3.4(b), below the diagonal], with background association levels similar to those observed in the template PRISM dataset. These results support SparseDOSSA's ability to induce correlation structure within synthetic microbiome datasets: a useful feature for evaluating statistical approaches to reconstructing microbial ecological interaction networks (Faust et al., 2012b; Friedman and Alm, 2012; Fang et al., 2015).

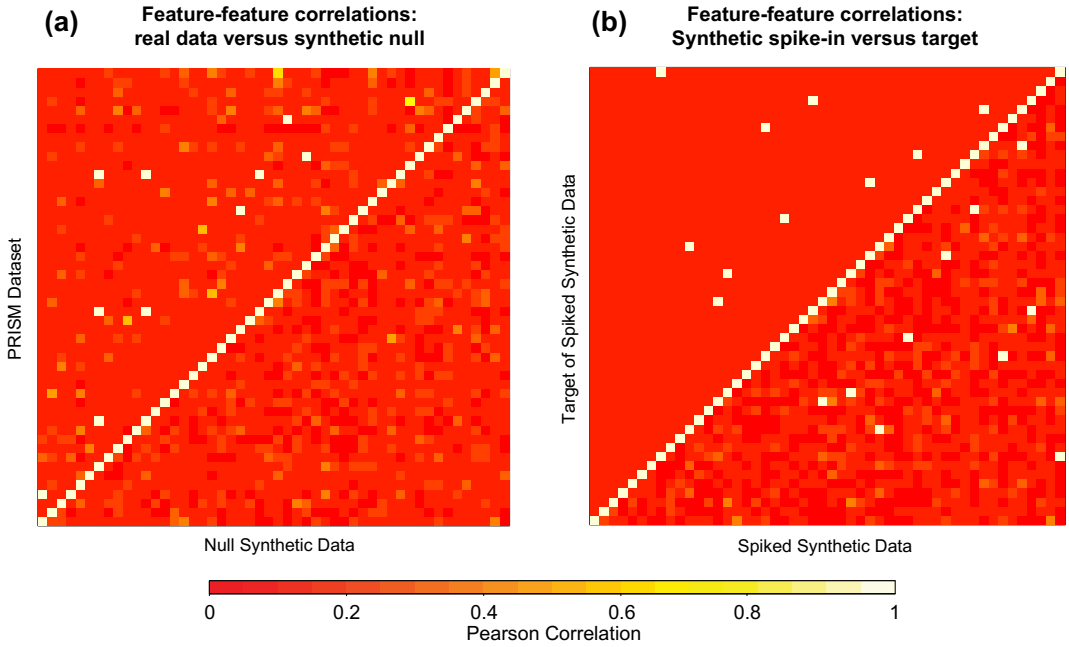


Figure 3.4: **Simulating correlation structure among microbial features.** (a) The pairwise (absolute) Pearson correlations based on raw counts between microbial features in the PRISM cohort (above the diagonal) and in the SparseDOSSA fit to this dataset (below the diagonal). (b) Pairs of features that are targeted to be correlated with each other (above the diagonal) and pairwise Pearson correlations in the resulting modified dataset (below the diagonal). Spiked-in correlations between features were constrained to have Pearsons correlation larger than 0.1.

3.3.4 SparseDOSSA accurately reproduces quantitative microbial community analysis results

Finally, we demonstrated that taxonomic count data simulated with SparseDOSSA can be used to recapitulate published statistical analyses based on real data. Specifically, we focused on Paulson et al.'s 2013 demonstration of the cumulative-sum scaling (CSS) technique for normalizing differences in sequencing depth across samples upstream of differential feature abundance testing (Paulson et al., 2013b). Paulson et al. compared the taxonomic composition of mouse microbiomes adapted to high- versus low-fat diets (Turnbaugh et al., 2009b) and found that their CSS technique identified a greater separation between the two diet groups than was observed using the more common total-sum scaling (TSS) technique.

We trained a SparseDOSSA model on the same raw taxonomic count data from mice used in the Paulson et al. study (Turnbaugh et al., 2009b), which consisted of 139 samples and 10,172 operational taxonomic units (OTUs). We collapsed OTUs into 484 genus-level taxonomic features and used this reduced dataset as our simulation template. To simulate the effects of adaption to two diet types, we generated a binary sample metadatum that divided the samples into two groups (Group 1 and Group 2) and then induced a correlation between this property and 10% of the 484 simulated features. We then repeated the analyses of the Paulson et al. study using our simulated dataset (Figure 3.5). First, we performed multi-dimensional scaling (MDS) to assess visual separation between Group 1 and Group 2 samples after normalizing for differences in sequencing depth using four different techniques [analogous to Figure 1(a-d) in (Paulson et al., 2013b)]: CSS [Figure 3.5(a)], size-factor normalization [as implemented in DESeq (Anders and Huber, 2010); Figure 3.5(b)], trimmed mean of M-values (TMM) normalization [as implemented in edgeR (Robinson et al., 2010); Figure 3.5(c)], and TSS [Figure 3.5(d)]. Under each normalization scheme, SparseDOSSA's simulated data exhibited similar ordination structure to that observed among the real (template) data in the Paulson et al. study.

In addition to comparing the real vs. simulated sample separation in the context of unsupervised ordination, we also applied supervised learning techniques to assess whether

differentially abundant features could accurately distinguished Group 1 and Group 2 samples [Figure 3.5(e), analogous to Figure 1(e) in (Paulson et al., 2013b)]. Specifically, we trained a sparse logistic classifier on the simulated data following application of each of the four normalization methods (CSS, DESeq, TMM, and TSS). We then calculated the leave-one-out odds ratio of belonging to Group 2 versus Group 1 for all samples across normalization methods. Our findings based on simulated data mirrored those of the original analysis based on the mouse samples: namely that CSS and TSS lead to nearly the same classification power, while CSS outperforms the DESeq and TMM normalization methods. Collectively, the results of these supervised and unsupervised validation experiments demonstrate that SparseDOSSA's model-simulated data can be used to recapitulate results of statistical analyses originally based on real sequencing data.

3.4 Discussion

We have developed a hierarchical Bayesian model, implemented in the R package SparseDOSSA, for fitting and/or simulating new microbial community taxonomic count data. The model accurately captures the fundamental characteristics of real microbial communities, including the distribution of counts or relative abundance across community members and the diversity of microbial composition across study populations. In addition, to support quantitative benchmarking of new methods, SparseDOSSA is able to reliably induce user-specified correlation structures involving feature-metadata or feature-feature associations in simulated datasets. The underlying generative model thus efficiently and effectively summarizes real microbial communities and recapitulates their latent structure in a statistically viable and principled manner.

The SparseDOSSA model assumes that the characteristics of a template (real) microbial community are well-captured by the baseline null distributions. More specifically, this requires assuming 1) that features are independent and 2) that population substructure among samples and intrinsic associations among features are absent in the template dataset (i.e. before optional spike-in of such structure). The former assumption will be true for most ecologically diverse communities, which observationally follow power-law

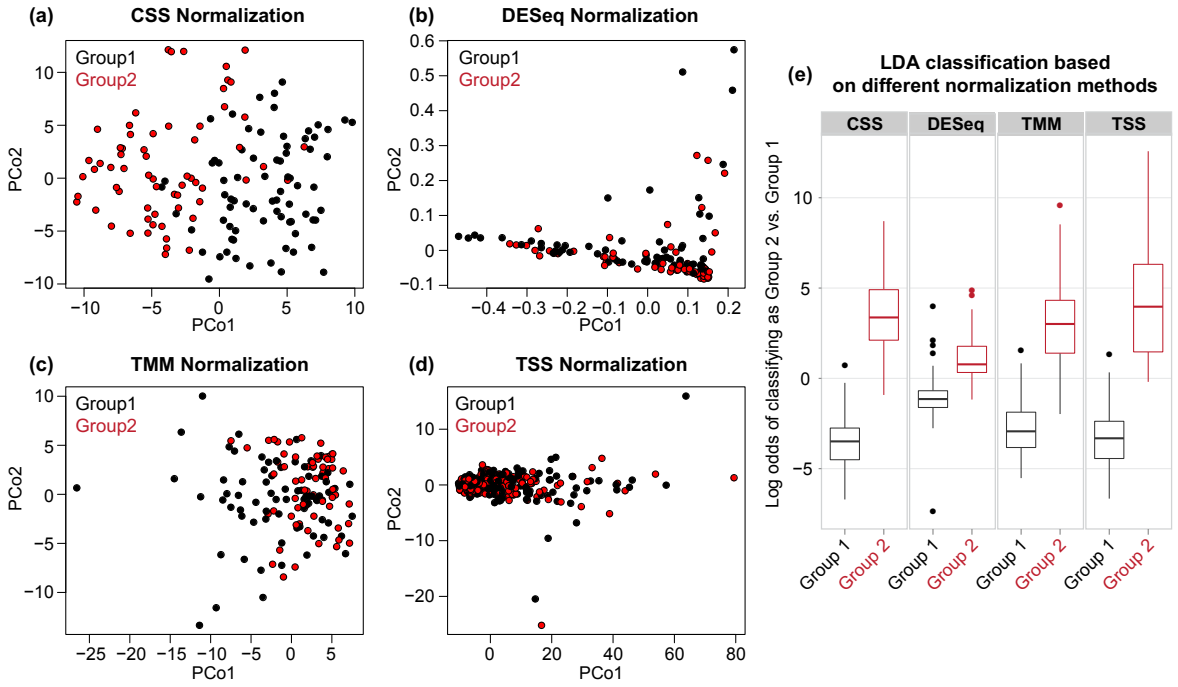


Figure 3.5: **SparseDOSSA reproduces biological diversity patterns among simulated microbial communities and permits comparative evaluation of statistical analysis techniques.** MDS analysis based on (a) cumulative sum scaling (CSS), (b) DESeq size factor normalization, (c) edgeRs trimmed mean of M-values (TMM), and (d) total sum scaling (TSS). (e) Linear discriminant analysis (LDA) posterior probability log-ratios for a synthetic binary class label (Group 1 vs. Group 2) based on a sparse logistic classifier. Each box corresponds to the distribution of leave-one-out posterior probability of assignment to Group 2 following different upstream normalization methods.

or log-normal behaviors (with a few abundant organisms and a long tail representing the increasingly rare biosphere). The latter assumption holds reasonably well even when any correlation structure originally present is weak or rare relative to overall microbial variance or affects only a small proportion of features. Inasmuch as the read count of each feature depends on its own observed mean, variance, and sparsity, SparseDOSSA's simulated data will replicate the marginal distribution of the originating template community. This guarantee on the null distribution of subsequently generated communities allows correlation structure (with samples or among features) to be optionally added in isolation for evaluation of microbial community analysis methods.

Our implementation of SparseDOSSA also provides two methods for fitting the hierarchical model to underlying data: one fully Bayesian method and a second nave method based on capturing parameters' simple stepwise maximum likelihoods. Since the nave method is far more computationally efficient in practice, we suggest using it as the default, and we have shown that this yields model fits (and thus simulated data) consistent with the posteriors of the fully Bayesian approach (e.g. Figure 3.2). Note that this does not, however, mean that the actual estimates of the parameters given by the nave method are also comparable to those given by the fully Bayesian approach. We assume that most users' interests will lie in simulating realistic taxonomic profiles based on observed data, in which case consistency of the predictive distribution is sufficient to guarantee the validity of the nave method. For users interested instead in dimensionality reduction of a microbial environment using the model's parameterization, we suggest applying the more computationally intensive but accurate Bayesian fit instead.

One limitation of our current probabilistic generative model is that we do not explicitly consider the effect of read depth. In the current model, variation in read depth is captured as a consequence of variability in feature-specific marginal distributions. In reality, the sampling scheme of microbial marker gene survey data suggests that variation in read depth is caused by an independent process and in fact contributes to the variability of the microbial abundances within a given feature across samples. One potential way to incorporate this would be to assume a sample-specific scaling factor that is dependent on read depth. This would then modify the actual mean parameter of the log-normal distribution

for each feature within a given sample. In pilot studies of such a modified model, the resulting fits did not show significant improvement over SparseDOSSA's default model (which lacks the explicit scaling factor for read depth; Figure A.7 and A.9).

Another property of real datasets that we do not explicitly model is the full extent of sparsity of features (i.e. microbes) with high estimated marginal means. That is, in some ecologies, a subset of microbes occur that are abundant in a few samples but near- or completely absent in most samples. Such behavior is evident in two HMP datasets, for example (Figure A.10). Our model assumes a monotone relationship between sparsity and marginal mean, which is insufficient to describe this bimodality of some extreme features. By fitting a simple logistic regression, we potentially over-estimate such features' sparsities. This may be the reason that the tails of simulated rank abundances for some environments (e.g. Figure 3.1 and A.7) are uniformly slightly sparser than the underlying data. Use of a mixture model in the sparsity structure for future versions of SparseDOSSAwhere the sparsity is derived either from an extreme value or is determined by the marginal meansmay resolve this issue.

Apart from limitations in capturing the general structure of microbiome data, our way to introduce correlation structures is also restrictive. The linear correlations we considered serve only as a simple start point. We would like to explore the possibility of generating more flexible correlation structure in our future work. For the correlation between metadata and microbiomes, instead of modifying the abundance of OTUs in the null community, we can explicitly make the marginal means of OTUs to be functions of chosen metadata, which is strictly compatible with the standard assumption in (generalized) linear model. For the correlation between species, we can consider a multivariate log-normal distribution for the abundance of all species in each sample. By specifying the covariance matrix of the multivariate log-normal distribution, we can construct a large class of dependency structure between species. An even more flexible method to introduce correlation between species is through copula (Weiss et al., 2016), which can be applied to arbitrary pre-specified marginal distributions of species.

In practice, we expect SparseDOSSA to primarily be useful as a model for benchmarking statistical methods that assess correlation structure in microbial taxonomic profiles. It

can potentially also extend such systems by providing a set of marginal parameters with lower dimensionality and potentially reduced noise relative to raw data, allowing sample metadata covariates to be more accurately tested for association with microbial features. In addition to the areas discussed above, future expansions of the model might include longitudinal structure or other interdependencies among samples (i.e. population substructure), as well as diversifying the application areas for the model (e.g. for power calculations during microbial community study design). As currently implemented, SparseDOSSA provides an end-to-end system that enables reproducible and efficient validation of quantitative methods applied to microbial community taxonomic profiles, allowing fair comparisons to be made between different methods or studies to establish a consistent baseline for statistical validation.

Appendix

A.1 Supplementary materials of Chapter 1

A.1.1 Approximating a Poisson Process using Beta random variables

Consider approximating a Poisson process on $(0,1)$ with intensity $\nu(\sigma) = \alpha\sigma^{-1}(1-\sigma)^{-1/2}$ by a finite counting process formed by n iid samples drawn from Beta$(\epsilon_n, 1/2-\epsilon_n)$ where $\epsilon_n < 1/2$. Denote the Poisson process as $N(t)$ and the approximating process as $N'_n(t)$, we first calculate the probability of having m points in interval $(\delta, t]$, where $m \leq n$, $t < 1$ and $0 < \delta \ll 1$,

$$
\begin{aligned}
P\left[N((\delta,t]) = m\right] =& \frac{\left[\int_\delta^t \alpha\sigma^{-1}(1-\sigma)^{-1/2}d\sigma\right]^m}{m!}\exp\left(-\int_\delta^t \alpha\sigma^{-1}(1-\sigma)^{-1/2}d\sigma\right), \\
P\left[N'_n((\delta,t]) = m\right] =& \binom{n}{m}\left(\frac{1}{\text{Beta}(\epsilon_n, 1/2-\epsilon_n)}\int_\delta^t \sigma^{-1+\epsilon_n}(1-\sigma)^{-1/2-\epsilon_n}d\sigma\right)^m \times \\
& \left(1 - \frac{1}{\text{Beta}(\epsilon_n, 1/2-\epsilon_n)}\int_\delta^t \sigma^{-1+\epsilon_n}(1-\sigma)^{-1/2-\epsilon_n}d\sigma\right)^{n-m}.
\end{aligned}
$$

The moment generating functions (MGFs) of $N((\delta,t])$ and $N'_n((\delta,t])$ are

$$
\begin{aligned}
M_N(\lambda) =& \exp\left[\left(e^\lambda - 1\right)\int_\delta^t \alpha\sigma^{-1}(1-\sigma)^{-1/2}d\sigma\right], \\
M_{N'_n}(\lambda) =& \left[\frac{e^\lambda - 1}{\text{Beta}(\epsilon_n, 1/2-\epsilon_n)}\int_\delta^t \sigma^{-1+\epsilon_n}(1-\sigma)^{-1/2-\epsilon_n}d\sigma + 1\right]^n.
\end{aligned}
$$

These two MGFs will be the same asymptotically if

$$
\lim_{n\to\infty} \frac{n}{\text{Beta}(\epsilon_n, 1/2-\epsilon_n)}\int_\delta^t \sigma^{-1+\epsilon_n}(1-\sigma)^{-1/2-\epsilon_n}d\sigma = \alpha\int_\delta^t \sigma^{-1}(1-\sigma)^{-1/2}d\sigma. \tag{A.1}
$$

This will be satisfied when $\epsilon_n = \alpha/n$. Indeed, under this assumption, we have

$$
\lim_{n\to\infty} \frac{n\left(\sigma/(1-\sigma)\right)^{\epsilon_n}}{\text{Beta}(\epsilon_n, 1/2-\epsilon_n)} = \alpha.
$$

In addition, since when n is large enough, the map $n \mapsto \frac{n(\sigma/(1-\sigma))^{\epsilon_n}}{\text{Beta}(\epsilon_n, 1/2-\epsilon_n)}$ is a non-increasing function, by Lebesgue's monotone convergence theorem, we can establish the convergence of the left hand side of (A.1) to the right hand side. Using this result, we can prove the weak convergence of the finite dimension distribution: $(N'(\delta, t_1], \ldots, N'(\delta, t_n]) \xrightarrow{d} (N(\delta, t_1], \ldots, N(\delta, t_n])$. This follows by a direct application of the multinomial theorem. Now we need to verify the tightness condition, this is automatically satisfied as $N_n(t)'$ is a càdlàg process (Daley and Vere-Jones, 1988) (Theorem 11.1. VII and Proposition 11.1. VIII, iv, Volume 2). Therefore we prove the weak convergence of the process $N_n'(t)$ to the Poisson process $N(t)$ when $n \to \infty$ and $\epsilon_n = \alpha/n$.

A.1.2 Proof of Proposition 1.1

We use the notation $P^j(\cdot) = \frac{\sum_i I(Z_i \in \cdot)\sigma_i Q_{i,j}^{+2}}{\sum_i \sigma_i Q_{i,j}^{+2}}$ where $Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}^j \rangle$. Denote $((Q_{i,j}, Q_{i,j'}), i \geq 1)$ as $\mathbf{Q}$. The joint distribution of $(Q_{i,j}, Q_{i,j'})$ is a multivariate normal with mean $\mathbf{0}$ and covariance $\phi(j, j')$, and the vectors $(Q_{k,j}, Q_{k,j'})$, $k = 1, 2, \ldots,$ are independent. We derive an expression for the covariance

$$\begin{aligned}\text{cov}[P^j(A), P^{j'}(A)] =& E[E[P^j(A)P^{j'}(A)|\sigma, \mathbf{Q}]] - E[P^j(A)]E[P^{j'}(A)] \\ =& (G(A) - G^2(A))E\left[\frac{\sum_i \sigma_i^2 Q_{i,j}^{+2} Q_{i,j'}^{+2}}{\sum_i \sigma_i Q_{i,j}^{+2} \sum_k \sigma_k Q_{k,j'}^{+2}}\right].\end{aligned}$$

Similarly, we can get the expression for the variance,

$$\text{var}[P^j(A)] = (G(A) - G^2(A))E\left[\frac{\sum_i \sigma_i^2 Q_{i,j}^{+4}}{\sum_i \sigma_i Q_{i,j}^{+2} \sum_k \sigma_k Q_{k,j}^{+2}}\right].$$

It follows that

$$\text{corr}[P^j(A), P^{j'}(A)] = E\left[\frac{\sum_i \sigma_i^2 Q_{i,j}^{+2} Q_{i,j'}^{+2}}{\sum_i \sigma_i Q_{i,j}^{+2} \sum_k \sigma_k Q_{k,j'}^{+2}}\right] \times \left(E\left[\frac{\sum_i \sigma_i^2 Q_{i,j}^{+4}}{\sum_i \sigma_i Q_{i,j}^{+2} \sum_k \sigma_k Q_{k,j}^{+2}}\right]\right)^{-1}.$$

Therefore the correlation is independent of the set A.

A.1.3 Proof of Proposition 1.2

We follow the framework of proofs for Theorem 1 and Theorem 3 in Barrientos et al. (2012). Let $\mathcal{P}(\mathcal{Z})$ be the set of all Borel probability measures defined on $(\mathcal{Z}, \mathcal{F})$ and $\mathcal{P}(\mathcal{Z})^J$

the product space of $J\,\mathcal{P}(\mathcal{Z})$. Assume $\Theta \subset \mathcal{Z}$ is the support of G. To show the prior assigns strictly positive probability to the neighborhood in Proposition 2, it is sufficient to show such neighborhood contains certain subset-neighborhoods with positive probability. As in Barrientos et al. (2012), we consider the subset-neighborhoods U:

$$U(G_1, \ldots, G_J, \{A_{i,j}\}, \epsilon^*) = \prod_{i=1}^{J} \{F_i \in \mathcal{P}(\Theta) : |F_i(A_{i,j}) - G_i(A_{i,j})| < \epsilon^*, j = 1, \ldots, m_i\},$$

where G_i is a probability measure absolutely continuous w.r.t. G for $i = 1, \ldots, J$, $A_{i,1}, \ldots, A_{i,m_i} \subset \Theta$ are measurable sets with G_i-null boundary and $\epsilon^* > 0$. The existence of such subset-neighborhoods is proved in Barrientos et al. (2012). We then define sets $B_{\nu_{1,1}\ldots\nu_{m_J,J}}$ for each $\nu_{i,j} \in \{0, 1\}$ as

$$B_{\nu_{1,1}\ldots\nu_{m_J,J}} = \bigcap_{i=1}^{J} \bigcap_{j=1}^{m_i} A_{i,j}^{\nu_{i,j}},$$

where $A_{i,j}^1 = A_{i,j}$ and $A_{i,j}^0 = A_{i,j}^c$. Set

$$J_\nu = \{\nu_{1,1} \ldots \nu_{m_J,J} : G(B_{\nu_{1,1},\ldots,\nu_{m_J,J}}) > 0\},$$

and let $\mathcal{M}$ be a bijective mapping from J_ν to $\{0, \ldots, k\}$ where $k = |J_\nu| - 1$. We can simplify the notation using $A_{\mathcal{M}(\nu)} = B_\nu$ for every $\nu \in J_\nu$. Define a vector $\mathbf{s}_i = (w_{i,0}, \ldots, w_{i,k}) = (Q_i(A_0), \ldots, Q_i(A_k))$ that belongs to the $k-$simplex Δ_k. Set

$$B(\mathbf{s}_i, \epsilon) = \{(w_0, \ldots, w_k) \in \Delta_k : |Q_i(A_j) - w_j| < \epsilon, j = 0, \ldots, k\},$$

where $\epsilon = 2^{-\sum_{i=1}^{J} m_i} \epsilon^*$. The derivation in Barrientos et al. (2012) suggests a sufficient condition for assigning positive mass to $U(G_1, \ldots, G_J, \{A_{i,j}\}, \epsilon^*)$ is

$$\Pi([P^i(A_0), \ldots, P^i(A_k)] \in B(\mathbf{s}_i, \epsilon), i = 1, \ldots, J) > 0. \tag{A.2}$$

Here Π is the prior.

Now consider the following conditions

C.1 $w_{i,l} - \epsilon_0 < \sigma_{l+1} Q_{l+1,i}^{+2} < w_{i,l} + \epsilon_0$ for $i = 1, \ldots, J$ and $l = 0, \ldots, k$.

C.2 $0 < \sum_{l>k+1} \sigma_l Q_{l,i}^{+2} < \epsilon_0$.

C.3 $Z_{l+1} \in A_l$ for $l = 0, \ldots, k$.

ϵ_0 in the above conditions satisfies the following inequality

$$\frac{w_{(i,l)} - \epsilon_0}{1 + (k+2)\epsilon_0} \geq w_{(i,l)} - \epsilon$$
$$\frac{w_{(i,l)} + 2\epsilon_0}{1 - (k+1)\epsilon_0} \leq w_{(i,l)} + \epsilon$$

for $i = 1, \ldots, J$ and $l = 0, \ldots, k$. This system of inequalities can be satisfied when k is large enough. If conditions $(C.1)$ to $(C.3)$ hold, it follows that $[P^i(A_0), \ldots, P^i(A_k)] \in B(\mathbf{s}_i, \epsilon)$ for $i = 1, \ldots, J$. Therefore, we have

$$\begin{aligned}
&\Pi([P^i(A_0), \ldots, P^i(A_k)] \in B(\mathbf{s}_i, \epsilon), i = 1, \ldots, J) \geq \\
&\prod_{l=0}^{k} \Pi(w_{(i,l)} - \epsilon_0 < \sigma_{l+1} Q^{+2}_{l+1,i} < w_{(i,l)} + \epsilon_0, i = 1, \ldots, J) \times \\
&\Pi(\sum_{l>k+1} \sigma_l Q^{+2}_{l,i} < \epsilon_0, i = 1, \ldots, J) \times \\
&\prod_{l=0}^{k} \Pi(Z_{l+1} \in A_l) \times \Pi(Z_l \in \mathcal{Z}, l = k+2, \ldots).
\end{aligned}$$

Since $(Q_{l,1}, \ldots, Q_{l,J})$ are multivariate normal random vectors with strictly positive definite covariance matrix and σ_l are always positive, the vector $(\sigma_{l+1} Q^{+2}_{l+1,i}, i = 1, \ldots, J)$ has full support on $\mathbb{R}^{+J}$ and will assign positive probability to any subset of the space. If follows that

$$\Pi(w_{i,l} - \epsilon_0 < \sigma_{l+1} Q^{+2}_{l+1,i} < w_{i,l} + \epsilon_0, i = 1, \ldots, J) > 0 \text{ for } l = 0, \ldots, k.$$

Using the Gamma process argument, we know $\sum_{l>k+1} \sigma_l Q^{+2}_{l,i}$ is the tail probability mass for a well-defined Gamma process and thus will always be positive and continuous for all i. It follows that

$$\Pi(\sum_{l>k+1} \sigma_l Q^{+2}_{l,i} < \epsilon_0, i = 1, \ldots, J) > 0.$$

Since $\mathcal{Z}$ is the topological support of G, it follows that $P(Z_{i+1} \in A_i) > 0$ and $P(Z_i \in \mathcal{Z}) = 1$. Combining these facts, we prove that Equation (A.2) holds.

A.1.4 Total variation bound of Laplace approximate of $p(Q_{i,j}|\mathbf{Q}_{i,-j}, \boldsymbol{\sigma}, \mathbf{T}, \mathbf{n})$

We consider the class of densities $g(x; k, \mu, s^2)$

$$g(x; k, \mu, s^2) \propto I(x \geq 0) x^{2k} f(x; \mu, s^2), k \in \mathbb{N}^+$$

where $f(x; \mu, s^2)$ is the density function of $N(\mu, s^2)$. The Laplace approximation of $g(x; k, \mu, s^2)$ is written as $f(x; \widehat{\mu}, \widehat{s}^2)$. Here $\widehat{\mu} = \text{argmax}_x g(x; k, \mu, s^2)$ and $\widehat{s}^2 = -\left(\left(\partial^2 \log(g)/\partial x^2\right)|_{\widehat{\mu}}\right)^{-1}$. We want to calculate the total variation distance between density $f(x; \widehat{\mu}, \widehat{s}^2)$ and $g(x; k, \mu, s^2)$, denoted as $d_{TV}(f(x; \widehat{\mu}, \widehat{s}^2), g(x; k, \mu, s^2))$.

Define class of functions $V(x; k, \mu)$ for $k \in \mathbb{N}^+, \mu > 0$:

$$V(x; k, \mu) = \begin{cases} 2k\left[\log(x/\mu) - (x/\mu - 1) + \frac{1}{2}(x/\mu - 1)^2\right] & x > 0 \\ -\infty & x \leq 0 \end{cases}$$

This function is non-decreasing and when $x = \mu$, $V(x; k, \mu) = 0$, $dV/dx = 0$ and $d^2V/dx^2 = 0$.

It follows that

$$\log g(x; k, \mu, s^2) - \log f(x; \widehat{\mu}, \widehat{s}^2) = V(x; k, \widehat{\mu}) + a_0 + a_1 x + a_2 x^2.$$

Moreover, since the $\widehat{\mu}$ is the mode of both $g(x; k, \mu, s^2)$ and $f(x; \widehat{\mu}, \widehat{s}^2)$, and the second derivative of $\log g(x; k, \mu, s^2)$ and $\log f(x; \widehat{\mu}, \widehat{s}^2)$ are identical at $x = \widehat{\mu}$, we can find that $a_1 = a_2 = 0$. Hence,

$$\log g(x; k, \mu, s^2) - \log f(x; \widehat{\mu}, \widehat{s}^2) = V(x; k, \widehat{\mu}) + a_0$$

and $g(x; k, \mu, s^2) = \exp\left(V(x; k, \widehat{\mu}) + a_0\right) f(x; \widehat{\mu}, \widehat{s}^2)$.

Since $V(x; k, \widehat{\mu})$ is monotone increasing, the total variation distance between $g(x; k, \mu, s^2)$ and $f(x; \widehat{\mu}, \widehat{s}^2)$ can be expressed as

$$\begin{aligned} d_{TV}(g(x; k, \mu, s^2), f(x; \widehat{\mu}, \widehat{s}^2)) &= \int_{x_0}^{+\infty} \left[\exp\left(V(x; k, \widehat{\mu}) + a_0\right) - 1\right] f(x; \widehat{\mu}, \widehat{s}^2) dx \\ &= \int_{-\infty}^{x_0} \left[1 - \exp\left(V(x; k, \widehat{\mu}) + a_0\right)\right] f(x; \widehat{\mu}, \widehat{s}^2) dx \end{aligned}$$

where $V(x_0; k, \widehat{\mu}) = -a_0$. If $a_0 \leq 0$, we have $x_0 \geq \widehat{\mu}$ and

$$\int_{x_0}^{+\infty} \left[\exp\left(V(x; k, \widehat{\mu}) + a_0\right) - 1\right] f(x; \widehat{\mu}, \widehat{s}^2) dx$$

$$\leq \int_{x_0}^{+\infty} \left[\exp\left(V(x;k,\widehat{\mu})\right) - 1\right] f(x;\widehat{\mu},\widehat{s}^2)dx$$
$$\leq \int_{\widehat{\mu}}^{+\infty} \left[\exp\left(V(x;k,\widehat{\mu})\right) - 1\right] f(x;\widehat{\mu},\widehat{s}^2)dx$$

Similarly, if $a_0 \geq 0$, we have

$$\int_{-\infty}^{x_0} \left[1 - \exp\left(V(x;k,\widehat{\mu}) + a_0\right)\right] f(x;\widehat{\mu},\widehat{s}^2)dx \leq \int_{-\infty}^{\widehat{\mu}} \left[1 - \exp\left(V(x;k,\widehat{\mu})\right)\right] f(x;\widehat{\mu},\widehat{s}^2)dx$$

To summarize, we have

$$d_{TV}(g(x;k,\mu,s^2), f(x;\widehat{\mu},\widehat{s}^2)) \leq \max\left(\int_{\widehat{\mu}}^{+\infty} \left[\exp\left(V(x;k,\widehat{\mu})\right) - 1\right] f(x;\widehat{\mu},\widehat{s}^2)dx, \right.$$
$$\left. \int_{-\infty}^{\widehat{\mu}} \left[1 - \exp\left(V(x;k,\widehat{\mu})\right)\right] f(x;\widehat{\mu},\widehat{s}^2)dx\right)$$

As we have shown in Equation (12) of the main manuscript, $\widehat{s}^2 = \left(\frac{2k}{\widehat{\mu}^2} + C\right)^{-1}$, where $C > 0$. This suggests that $\widehat{s} \leq \widehat{\mu}/\sqrt{2k}$. Therefore

$$d_{TV}(g(x;k,\mu,s^2), f(x;\widehat{\mu},\widehat{s}^2)) \leq \max\left(\int_{\widehat{\mu}}^{+\infty} \left[\exp\left(V(x;k,\widehat{\mu})\right) - 1\right] f(x;\widehat{\mu},\widehat{\mu}/2k)dx, \right.$$
$$\left. \int_{-\infty}^{\widehat{\mu}} \left[1 - \exp\left(V(x;k,\widehat{\mu})\right)\right] f(x;\widehat{\mu},\widehat{\mu}/2k)dx\right)$$

Since $V(x;\mu,s^2)$ and $f(x;\mu,s^2)$ are location-scale families, the above expression can be made free of $\widehat{\mu}$ and thus μ and s^2:

$$d_{TV}(g(x;k,\mu,s^2), f(x;\widehat{\mu},\widehat{s}^2)) \leq \max\left(\int_{1}^{+\infty} \left[\exp\left(V(x;k,1)\right) - 1\right] f(x;1,1/2k)dx, \right.$$
$$\left. \int_{-\infty}^{1} \left[1 - \exp\left(V(x;k,1)\right)\right] f(x;1,1/2k)dx\right) \tag{A.3}$$

This upper bound on the total variation distance decreases as k increases and it goes to 0 as $k \to \infty$. This suggests the convergence of the approximating normal distribution to the density family g in total variation sense. We also plot this upper bound as a function of k to verify the conclusion. It is shown in the supplemental Figure A.1.

A.1.5 Details of self-consistent estimates in Section 3.1

First we estimate $\boldsymbol{\sigma}$ and then we transform the data $n_{i,j}$ into $\sqrt{n_{i,j}/\sigma_i}$. If $n_{i,j}$ is representative and $\boldsymbol{\sigma}$ is estimated accurately, we have $\sqrt{n_{i,j}/\sigma_i} = c_j Q^+_{i,j}$. If the covariance matrix

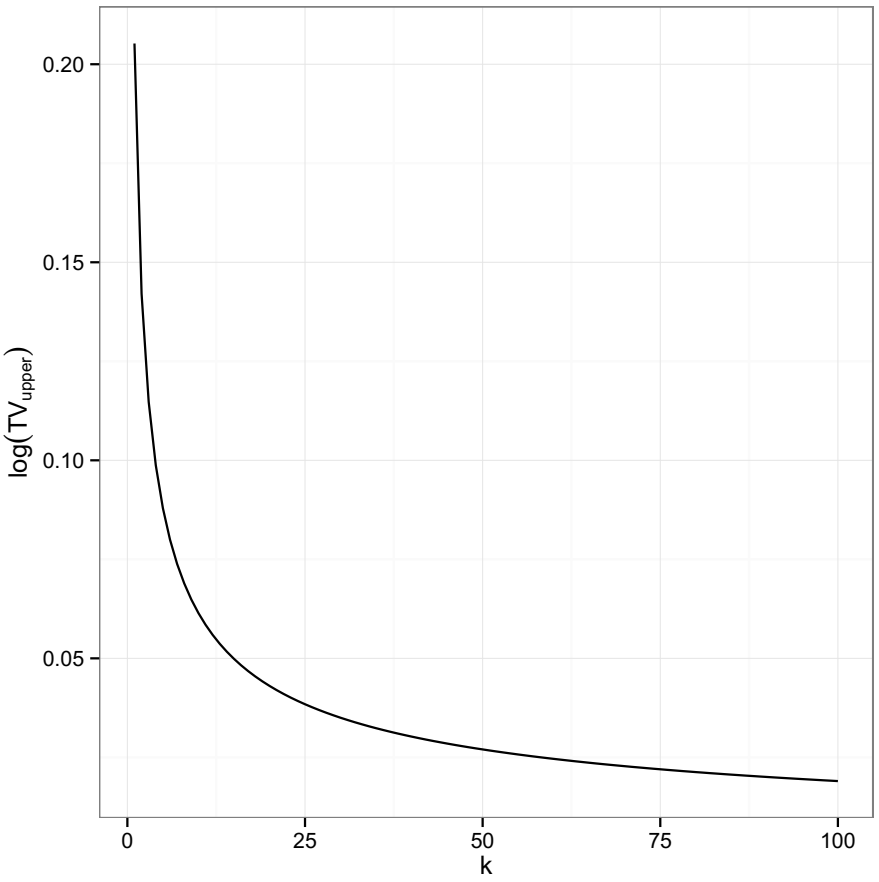


Figure A.1: Upper bound of the total variation distance of Laplace approximation in (12) to the density in (11) as given in (A.3) when frequency k increases.

of $\mathbf{Q}_i$ is $\mathbf{\Sigma}$, then the covariance matrix of $(\sqrt{n_{i,j}/\sigma_i}, j = 1, \ldots, J)$ will be $\widetilde{\mathbf{\Sigma}} = \mathbf{\Lambda}\mathbf{\Sigma}\mathbf{\Lambda}$ where $\mathbf{\Lambda} = diag\{c_1, \ldots, c_J\}$.

It is obvious that $(\sqrt{n_{i,j}/\sigma_i}, j = 1, \ldots, J)$ is MVN and the correlation matrix will be the same as the induced correlation matrix from $\mathbf{\Sigma}$. Methods on identifying the covariance matrix using this truncated dataset are abundant and well-studied. One way to do it is the EM algorithm. This estimated covariance matrix will by no means to be the same as $\mathbf{\Sigma}$, but the induced correlation matrix will be very close to the true correlation matrix induced by $\mathbf{\Sigma}$. Hence if our interest is on estimating correlation matrix, we can just treat $(\sqrt{n_{i,j}/\sigma_i}, j = 1, \ldots, J)$ as the truncated version of the true $\mathbf{Q}_i$ and proceed.

The EM algorithm should then be derived for the following settings. Let $\mathbf{Q}_i \overset{iid}{\sim} MVN(\mathbf{0}, \mathbf{\Sigma})$. Instead of observing I independent $\mathbf{Q}_i$, we only observe the positive entries in each $\mathbf{Q}_i$ and know the rest of the entries are negative. Denote the observed data vector as $\widetilde{\mathbf{Q}}_i$. We want to estimate $\mathbf{\Sigma}$ from the data $\widetilde{\mathbf{Q}}_i, i = 1, \ldots, I$. A standard EM algorithm can be easily formulated as following:

E-step Get the conditional expectation of full data log likelihood, given the observed data. Define two index sets, $\mathcal{A}_i = \{j|\widetilde{Q}_{i,j} > 0\}$ and $\mathcal{B}_i = \{j|\widetilde{Q}_{i,j} = 0\}$. For an arbitrary index set $\mathcal{I}$, denote $Q_{\mathcal{I}} = (Q_{i,j}|j \in \mathcal{I})$. Denote $\mathcal{A} = \{(i,j)|j \in \mathcal{A}_i, i = 1, \ldots, I\}$ and $\mathcal{B} = \{(i,j)|j \in \mathcal{B}_i, i = 1, \ldots, I\}$. The E-step function at $t+1$ iteration is,

$$L(\Sigma|\Sigma_t) = \mathbb{E}\left[-\frac{I}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\text{Tr}(\mathbf{\Sigma}^{-1}\sum_i \mathbf{Q}_i\mathbf{Q}_i')|\mathbf{\Sigma}_t, Q_{\mathcal{A}} = \widetilde{Q}_{\mathcal{A}}, Q_{\mathcal{B}} < 0\right].$$

Notice this expectation is not easy to calculate in general. We use instead Monte Carlo method to approximate it. We sample K copies of $\mathbf{Q}_i$ from the conditional distribution $(\mathbf{Q}_i|Q_{\mathcal{A}_i} = \widetilde{Q}_{\mathcal{A}_i}, Q_{\mathcal{B}_i} < 0)$ where $\mathbf{Q}_i \sim MVN(\mathbf{0}, \mathbf{\Sigma}_t)$. The conditional distribution is a truncated multivariate normal distribution and we use the R package `tmvtnorm` (Wilhelm, 2015) to sample from it. If we denote by $\mathbf{Q}_i^1, \ldots, \mathbf{Q}_i^K$ the K samples of Q_i, L can be approximated as

$$\widehat{L}(\Sigma|\Sigma_t) = -\frac{1}{K}\sum_{k=1}^{K}\left[\text{Tr}(\mathbf{\Sigma}^{-1}\sum_i \mathbf{Q}_i^k(\mathbf{Q}_i^k)')\right] - \frac{I}{2}\log|\mathbf{\Sigma}|.$$

M-step We seek to maximize $\widehat{L}$ with respect to $\mathbf{\Sigma}$. Due to a well-known fact on the maximum likelihood estimate of covariance matrix of multivariate normal, it is straightforward to get

$$\mathbf{\Sigma}_{t+1} = \frac{1}{IK} \sum_{i,k} \mathbf{Q}_i^k (\mathbf{Q}_i^k)'.$$

We applied this algorithm to the simulated datasets generated for Figure 3(a) to estimate the normalized Gram matrix $\mathbf{S}$. A summary of the RV-coefficients between the estimates from the above algorithm and the truth is shown in Figure A.2. We also compared the estimates from this algorithm with those from MCMC simulations in Figure A.2. The estimates of $\mathbf{S}$ from MCMC simulation are always better than those given by the self-consistent algorithm but both perform very well.

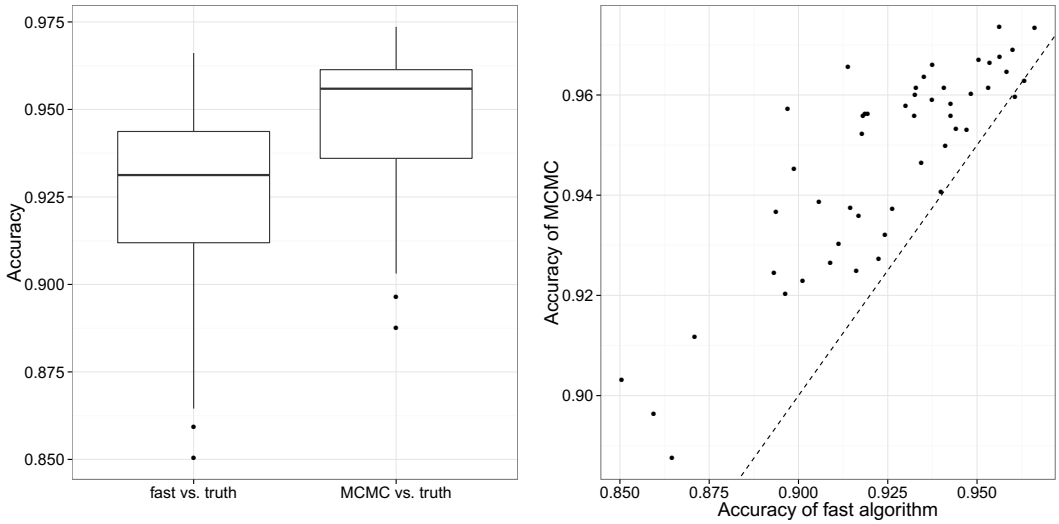


Figure A.2: (**Left**) Box-plots compare the distributions of RV-coefficients between estimates from our self-consistent algorithm and between estimates from MCMC simulation and truth. (**Right**) Scatter plot to show per simulation comparison of RV coefficients for the self-consistent algorithm and MCMC sampling. Dashed line indicates where the two algorithms have identical accuracy.

A.1.6 Standard PCoA for ordination of simulated dataset, Global Patterns dataset and Ravel's vaginal microbiome dataset

In this section, we include three sets of ordination figures generated using the standard PCoA method in microbiome studies. We first calculate the dissimilarity matrix of bio-

logical samples by applying Bray-Curtis dissimilarity metric on the empirical microbial distributions. We then perform classic Multi-dimensional Scaling (MDS) to ordinate biological samples based on the dissimilarity matrix. In Figure A.3, we show the PCoA result for the simulated dataset generated for Figure 3(f). In Figure A.4 and A.5, we illustrate the PCoA results for the Global Patterns dataset and Ravel's vaginal microbiome dataset respectively. To be consistent with the main results, we show the ordination results based on the first three principal coordinates for the Global Patterns dataset and Ravel's vaginal microbiome dataset.

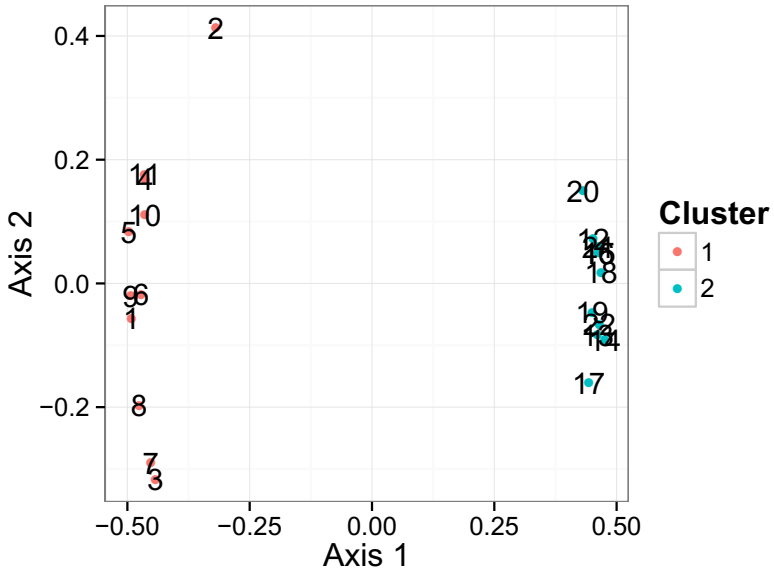


Figure A.3: PCoA result for the simulated dataset generated for Figure 1.3(f).

A.1.7 Benchmarking the MCMC sampler

In this section, we focus on evaluating the computational performance of our MCMC sampler. We first consider the computational time of the sampler under different scenarios. We then illustrated a convergence diagnosis to check whether the sampler has reached mixing in the setting of our simulation study in the main manuscript. In addition, we created two larger datasets to verify the number of iterations needed to reach mixing will not be compromised if the underlying latent structure remains low dimensional.

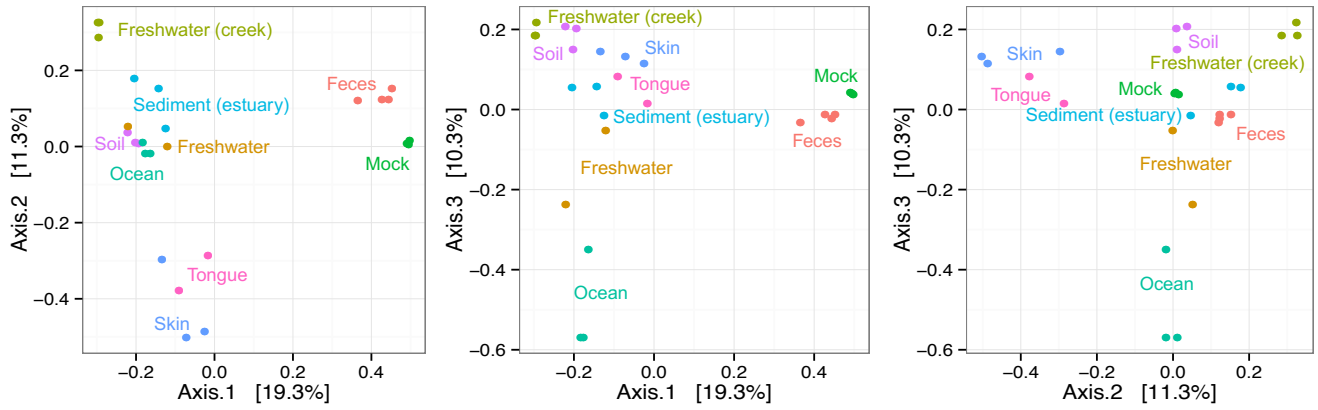


Figure A.4: PCoA results for the Global Patterns dataset. We show the three two-dimensional representations of the ordination given by the first three principal coordinates.

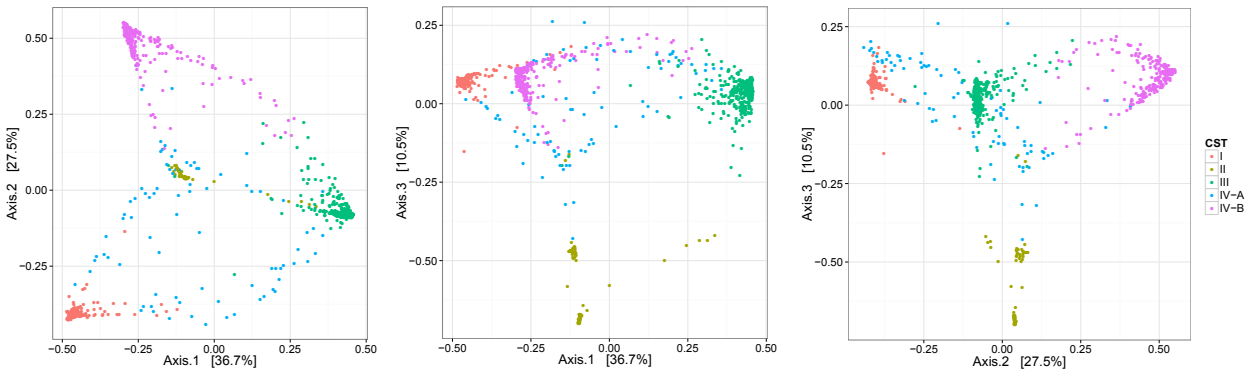


Figure A.5: PCoA results for Ravel's vaginal microbiome dataset. We show the three two-dimensional representations of the ordination given by the first three principal coordinates.

Computation time of the MCMC sampler

In Table A.1 we listed the elapsed time in seconds for the MCMC sampler to finish $1,000$ iterations under different scenarios. All the scenarios are run with a single thread on a MacBook Pro with 2.7GHz Intel Core i5 and 8 GB 1867 MHz DDR3 RAM. In particular, we evaluated the effect of the number of biological samples (J), the number of species (I), the dimension of the latent factors (m), and the total counts per biological sample (n^j).

Table A.1: Computation time (in seconds) of 1,000 iterations for the MCMC sampler

		$I = 68$			$I = 500$			$I = 1000$		
		$m = 5$	$m = 10$	$m = 20$	$m = 5$	$m = 10$	$m = 20$	$m = 5$	$m = 10$	$m = 20$
$J = 22$	$n^j = 10^3$	2.3	2.8	2.4	5.7	5.8	7.0	11.4	10.4	12.6
	$n^j = 10^4$	1.3	1.6	1.9	5.7	5.5	6.4	8.7	8.8	11.3
	$n^j = 10^5$	1.1	1.4	1.5	4.7	3.9	6.3	7.2	8.2	11.5
$J = 100$	$n^j = 10^3$	3.6	3.7	5.5	11.5	14.6	17.1	21.8	21.0	30.2
	$n^j = 10^4$	3.3	3.7	5.4	11.5	12.1	20.4	18.1	21.1	29.5
	$n^j = 10^5$	3.4	4.0	5.5	12.3	18.9	17.8	19.2	21.5	31.1
$J = 1000$	$n^j = 10^3$	31.4	34.3	49.6	121.2	118.4	152.1	152.1	173.8	251.0
	$n^j = 10^4$	28.2	33.4	53.1	96.3	144.3	159.7	143.7	164.8	254.2
	$n^j = 10^5$	40.1	38.2	52.2	129.1	111.5	138.2	163.2	171.7	246.0

Increasing the total number of reads per biological sample (n^j) does not affect the computation time. On the other hand, there is a weak effect associated with the dimension of the latent factors (m). In general, the computation time tends to increase with m. The number of species (I) and the number of biological samples (J) affect the speed of computation significantly. These results illustrate that the MCMC sampler can finish $50,000$ iterations for a dataset with 100 samples and 1000 species in less than 20 minutes.

The table illustrates that it is possible to apply our model to microbiome datasets with comparable numbers of biological samples. It is rare to have datasets with more than a thousand confidently assigned OTUs (Callahan et al., 2016).

Convergence diagnosis of the MCMC sampler

We evaluate the convergence of the MCMC sampler in the setting of Section 5 (simulation study). The number of biological samples is fixed at $J = 22$. We ran three parallel chains for three scenarios $I = 68$, $I = 500$ and $I = 1,000$. For each different I, we obtain the

posterior samples of the first three eigenvalues of the normalized Gram matrix $\mathbf{S}$ in all three chains and use $\widehat{R}$ statistics (Gelman and Rubin, 1992b) to check if the chains reached mixing. We chose to visualize the eigenvalues of $\mathbf{S}$ since in our model $\mathbf{S}$ is identifiable. The results are shown in Figure A.6.

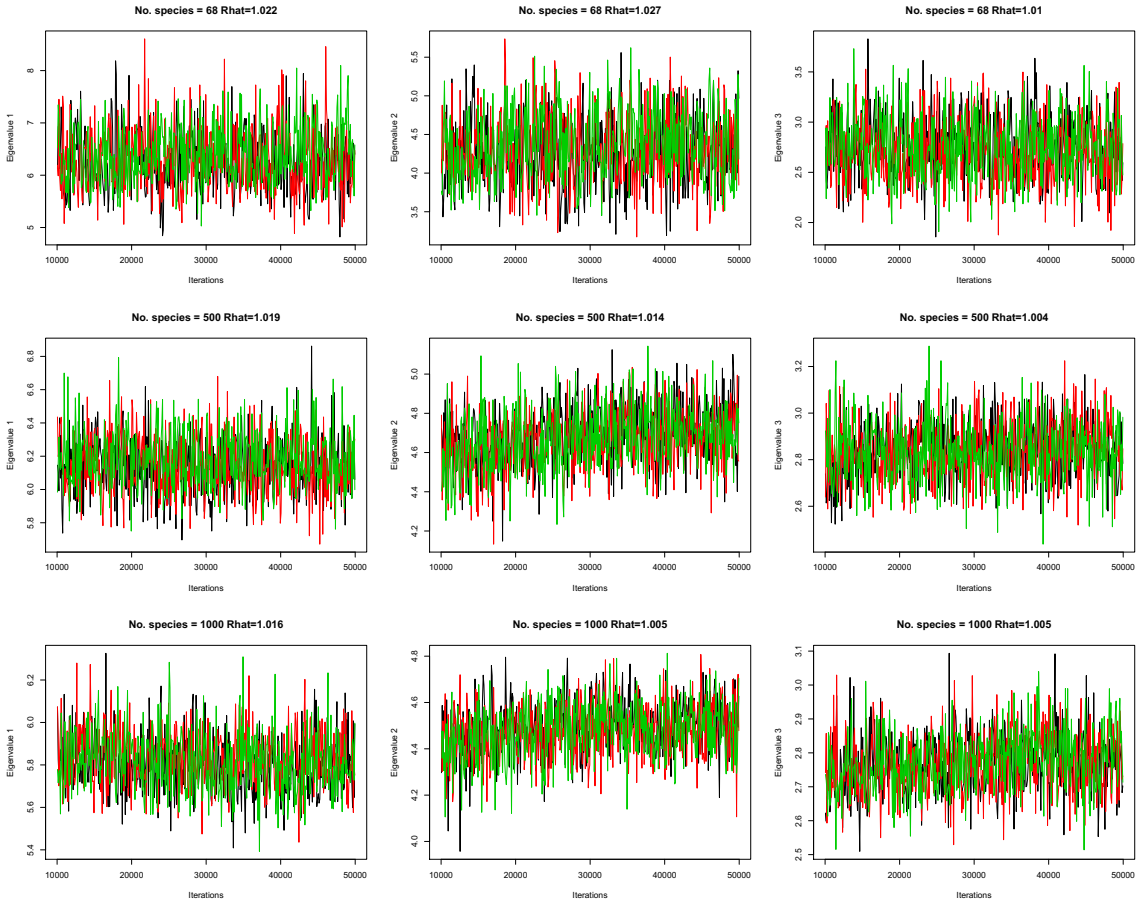


Figure A.6: Traceplots for the posterior samples of the first three eigenvalues of $\mathbf{S}$. Each row corresponds to a different I and each column to a different eigenvalue. The $\widehat{R}$ statistics are shown in the title of each figure.

The $\widehat{R}$ statistics are all close to one supporting good MCMC mixing after 20,000 iterations, so our choice of 50,000 total iterations seems reasonable for providing posterior inference.

A.2 Supplementary materials of Chapter 2

A.2.1 Proof of Proposition 2.1

Before we prove Proposition 2.1, we want to introduce a lemma that will be used frequently in the proof.

Lemma A.1. *Assume $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n$ are iid sample from a distribution $F(z; \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is the parameter of the distribution. If there exists an estimator $\widehat{\boldsymbol{\beta}}(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ such that $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ when $n \to \infty$, then the distribution of $\mathbf{Z}_i$ is the same under $F(z; \boldsymbol{\beta})$ and $F(z; \boldsymbol{\beta}')$ if and only if $\boldsymbol{\beta} = \boldsymbol{\beta}'$.*

Proof of Lemma A.1. If $\boldsymbol{\beta} = \boldsymbol{\beta}'$, the distribution of $\mathbf{Z}_i$ will be identical by definition. Assume there exist $\boldsymbol{\beta} \neq \boldsymbol{\beta}'$ such that $F(z; \boldsymbol{\beta}) = F(z; \boldsymbol{\beta}')$ for every z. If $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are iid samples from $F(z; \boldsymbol{\beta})$, by the assumption we have $\widehat{\boldsymbol{\beta}}(\mathbf{Z}_1, \ldots, \mathbf{Z}_n) \to \boldsymbol{\beta}$. Because $F(z; \boldsymbol{\beta}) = F(z; \boldsymbol{\beta}')$, we also have $\widehat{\boldsymbol{\beta}}(\mathbf{Z}_1, \ldots, \mathbf{Z}_n) \to \boldsymbol{\beta}'$. Since $\boldsymbol{\beta} \neq \boldsymbol{\beta}'$, this leads to a contraction. Therefore the distribution of $\mathbf{Z}_i$ will be the same only if $\boldsymbol{\beta} = \boldsymbol{\beta}'$. □

We now consider Proposition 2.1. We know $\mathbf{Q}_i = (Q_{i,1}, \ldots, Q_{i,J})^\intercal$ is distributed independently as $MVN(\mathbf{w}^\intercal \mathbf{v}_i, \boldsymbol{\Sigma})$ conditioning on $\mathbf{w} = (\mathbf{w}^1, \ldots, \mathbf{w}^J)$ for $i = 1, 2, \ldots$ and $\mathbf{w}^j \overset{iid}{\sim} f(\cdot)$. Denote $\mathbf{Z}_i = (\mathbb{I}(Q_{i,1} > 0), \ldots, \mathbb{I}(Q_{i,J} > 0))^\intercal$. Let $\mathbf{U} = \text{diag}(\Sigma_{1,1}^{-1/2}, \ldots, \Sigma_{J,J}^{-1/2})$, the correlation matrix induced by $\boldsymbol{\Sigma}$ is written as $\mathbf{S} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}$.

1. *Identifiability of* $\mathbf{v}_i/\|\mathbf{v}_i\|$. We first want to prove the identifiability of $\mathbf{v}_i/\|\mathbf{v}_i\|$ based on Lemma A.1. Consider the single observation $(Z_{i1}, \mathbf{w}^1)$, the model parameters linked with it are $f(\cdot)$, $\Sigma_{1,1}$ and $\mathbf{v}_i$. If we have n iid replicates of $(Z_{i1}, \mathbf{w}^1)$ generated from the model, denoted as $(Z_1, \mathbf{w}_1^1), \ldots, (Z_n, \mathbf{w}_n^1)$, it is straightforward to see that

$$P(Z_i = z_i | \mathbf{w}_i^1) = \Phi((2z_i - 1)\mathbf{v}_i^\intercal \mathbf{w}_i^1 / \sqrt{\Sigma_{1,1}}),$$

where $z_i \in \{0, 1\}$. Based on the standard theory of GLM with probit link, the MLE $\widehat{\mathbf{v}}_i$, defined as

$$\widehat{\mathbf{v}}_i = \text{argmax}_{\mathbf{v}} \sum_{i=1}^{n} \log(\Phi((2z_i - 1)\mathbf{v}^\intercal \mathbf{w}_i^1)),$$

is consistent to $\mathbf{v}_i/\sqrt{\Sigma_{1,1}}$. This implies that $\widehat{\mathbf{v}}_i/\|\widehat{\mathbf{v}}_i\|$ is an consistent estimator of $\mathbf{v}_i/\|\mathbf{v}_i\|$. Using Lemma A.1, we know that $\mathbf{v}_i/\|\mathbf{v}_i\|$ is identifiable based on the data.

2. *Identifiability of* $\mathbf{S}$. Consider the data points $(Z_{i,1}, \mathbf{w}^1)$ and $(Z_{i,2}, \mathbf{w}^2)$. The joint distribution of $Z_{i,1}$ and $Z_{i,2}$ conditioning on $\mathbf{w}^1$ and $\mathbf{w}^2$ can be written as

$$P(Z_{i,1} = z_1, Z_{i,2} = z_2 | \mathbf{w}^1, \mathbf{w}^2) = \int_A \frac{1}{2\pi}(1-\rho_{1,2}^2)^{-1/2} \exp\left(-\frac{1}{2}\mathbf{q}^\intercal \mathbf{S}_{1:2}^{-1}\mathbf{q}\right) d\mathbf{q},$$

where $A = (-\infty, (2z_1-1)\mathbf{v}_i\mathbf{w}^1\Sigma_{1,1}^{-1/2}] \times (-\infty, (2z_2-1)\mathbf{v}_i\mathbf{w}^2\Sigma_{2,2}^{-1/2}]$, $\rho_{1,2} = \text{corr}(Q_{i,1}, Q_{i,2})$ and $\mathbf{S}_{1:2} = \begin{pmatrix} 1 & \rho_{1,2} \\ \rho_{1,2} & 1 \end{pmatrix}$.

If two different sets of parameter values $\{f, \mathbf{\Sigma}, \{\mathbf{v}_i\}, J\}$ and $\{f, \mathbf{\Sigma}', \{\mathbf{v}_i'\}, J\}$ induce the same joint distribution of $(Z_{i,j}, \mathbf{w}^j)$ for $j = 1, \ldots, J$, it follows that the conditional distribution of $(Z_{i,1}, Z_{i,2})$ given $\mathbf{w}^1, \mathbf{w}^2$ is the same and also the conditional distribution $Z_{i,1}|\mathbf{w}^1$ and $Z_{i,2}|\mathbf{w}^2$ remain the same.

We know that $P(Z_{i,1} = z|\mathbf{w}^1) = \Phi((2z-1)\mathbf{v}_i^\intercal \mathbf{w}^1 \Sigma_{1,1}^{-1/2})$, which is a monotone function of $\mathbf{v}_i^\intercal \mathbf{w}^1 \Sigma_{1,1}^{-1/2}$. Therefore if $\{f, \mathbf{\Sigma}, \{\mathbf{v}_i\}, J\}$ and $\{f, \mathbf{\Sigma}', \{\mathbf{v}_i'\}, J\}$ induce the same distribution of $Z_{i,1}|\mathbf{w}^1$, it is necessary that $\mathbf{v}_i^\intercal \mathbf{w}^1 \Sigma_{1,1}^{-1/2} = (\mathbf{v}_i')^\intercal \mathbf{w}^1 (\Sigma_{1,1}')^{-1/2}$. Similarly, $Z_{i,2}|\mathbf{w}^2$ remains the same only if $\mathbf{v}_i^\intercal \mathbf{w}^2 \Sigma_{2,2}^{-1/2} = (\mathbf{v}_i')^\intercal \mathbf{w}^2 (\Sigma_{2,2}')^{-1/2}$.

Hence, if the joint conditional distribution $Z_{i,1}, Z_{i,2}|\mathbf{w}^1, \mathbf{w}^2$ remains the same in the two different sets of parameters, it is necessary that $\mathbf{v}_i^\intercal \mathbf{w}^1 \Sigma_{1,1}^{-1/2} = (\mathbf{v}_i')^\intercal \mathbf{w}^1 (\Sigma_{1,1}')^{-1/2}$ and $\mathbf{v}_i^\intercal \mathbf{w}^2 \Sigma_{2,2}^{-1/2} = (\mathbf{v}_i')^\intercal \mathbf{w}^2 (\Sigma_{2,2}')^{-1/2}$. Moreover, based on Theorem 3.1 in Ledoux and Talagrand (2013), it is also necessary that $\rho_{1,2} = \rho_{1,2}'$. If we apply this result for every $Z_{i,j}, Z_{i,j'}|\mathbf{w}^j, \mathbf{w}^{j'}$ where $j \neq j'$, we get that $\{f, \mathbf{\Sigma}, \{\mathbf{v}_i\}, J\}$ and $\{f, \mathbf{\Sigma}', \{\mathbf{v}_i'\}, J\}$ induce the same distribution of $\{Z_{i,j}\}$ only if $\mathbf{S} = \mathbf{S}'$, which implies the identifiability of $\mathbf{S}$.

3. *Identifiability of* $\Sigma_{jj}/\Sigma_{j'j'}$. We assume there are n replicates of $(Z_{i,j}, \mathbf{w}^j, Z_{i,j'}, \mathbf{w}^{j'})$. Using the result in 1, it follows that there are two estimators that are consistent to $\mathbf{v}_i/\Sigma_{j,j}^{1/2}$ and $\mathbf{v}_i/\Sigma_{j',j'}^{1/2}$ respectively. Therefore there is one consistent estimator of $\Sigma_{j,j}/\Sigma_{j',j'}$ provided that the underlying true $\mathbf{v}_i \neq \mathbf{0}$. This suggests the identifiability of $\Sigma_{j,j}/\Sigma_{j',j'}$ based on $(Z_{i,j}, \mathbf{w}^j, Z_{i,j'}, \mathbf{w}^{j'})$.

4. *Identifiability of* $\mathbf{v}_i$. Results in 3 suggests if we constrain $\Sigma_{1,1} = c$ or $\sum_{j=1}^J \Sigma_{j,j} = c$, each individual $\Sigma_{j,j}$ is also identifiable. Using the result in 1, we can show that $\mathbf{v}_i$ is also identifiable.

5. *Identifiability of* $\sigma_i/\sigma_{i'}$. To analyze the identifiability of $\sigma_i/\sigma_{i'}$, we need to use the full data $\{n_{i,j}\}_{i\geq 1, j\leq J}$. We still proceed under the assumption that $n_{i,j}/(\sigma_i Q_{i,j}^{+2}) = c_j$. Consider N iid replicates of $(n_{i,j}, n_{i',j}, \mathbf{w}^j)$, denoted as $(n_k, n'_k, \mathbf{w}_k^j)_{k\leq N}$. By model definition, we know that

$$n_k/n'_k \times \mathbb{I}(n'_k > 0) = \frac{\sigma_i}{\sigma_{i'}} \frac{Q_k^{+2}}{(Q'_k)^{+2}} \mathbb{I}(Q'_k > 0),$$

where $(Q_k, Q'_k)^\intercal \overset{iid}{\sim} MVN((\mathbf{v}_i^\intercal \mathbf{w}_k^j, \mathbf{v}_{i'}^\intercal \mathbf{w}_k^j)^\intercal, \text{diag}(\Sigma_{j,j}, \Sigma_{j,j}))$. By weak law of large number, we have

$$\frac{1}{N}\sum_i \frac{n_k}{n'_k} \xrightarrow{p} \frac{\sigma_i}{\sigma_{i'}} \times \mathbb{E}\left(\frac{Q_k^{+2}}{(Q'_k)^{+2}} \mathbb{I}(Q'_k > 0)\right).$$

This means we can construct a consistent estimator of $\frac{\sigma_i}{\sigma_{i'}}$ based on N replicates of observed data. By invoking Lemma A.1, we proved that $\sigma_i/\sigma_{i'}$ is identifiable. Notice the finite first moment of $\frac{Q_k^{+2}}{(Q'_k)^{+2}} \mathbb{I}(Q'_k > 0)$ relies on the conclusions for the ratio between two non-centered Chi-square distributions (Hawkins and Han, 1986).

A.3 Supplementary materials of Chapter 3

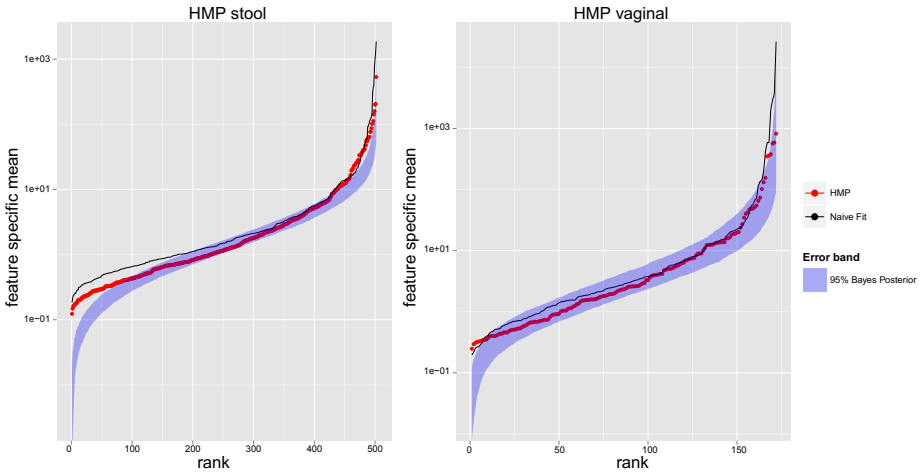


Figure A.7: **Distribution of rank relative abundances for simulated data using naive versus fully Bayesian methods**. We show the performance of the model-fitting method applied in SparseDOSSA for two additional datasets: HMP stool and HMP vaginal samples.

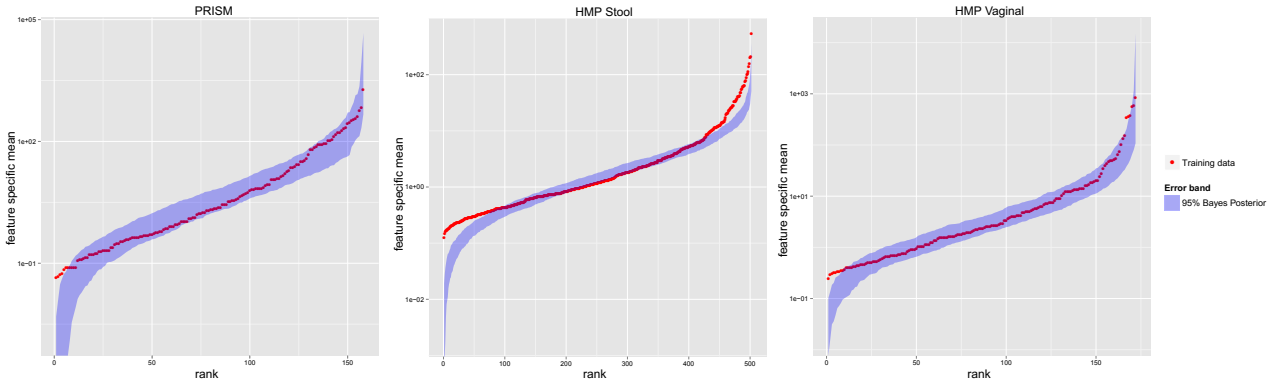


Figure A.8: **Distribution of rank relative abundances for simulated data using a model that incorporates read depth**. We repeated the experiment shown in Figure 3.2(a) from the main text using an augmented model trained on the PRISM, HMP stool, and HMP vaginal datasets.

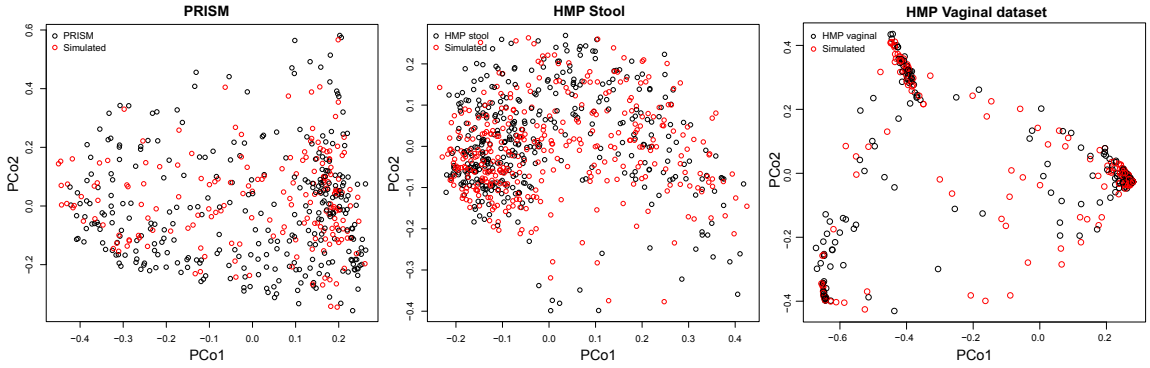


Figure A.9: **PCoA analysis for simulated data using a model that incorporates read depth**. We repeated the experiment shown in Figure 3.2(b) from the main text using an augmented model trained on the PRISM, HMP stool, and HMP vaginal datasets.

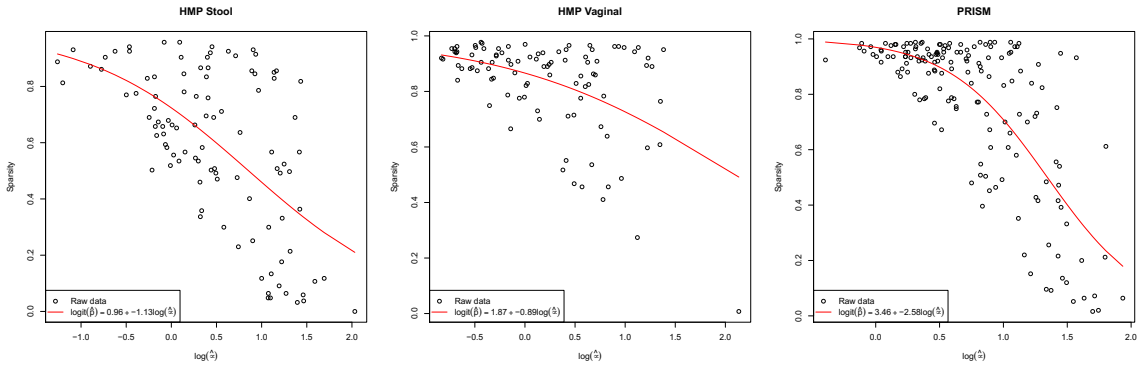


Figure A.10: **Relationship between observed feature-specific sparsity and mean abundance**. We plot the proportions of zero-count samples for all features against the corresponding log-transformed mean log-transformed abundances. Red curves are results of logistic regression for the relationship between sparsity and mean abundance. We repeated this procedure on the PRISM, HMP stool, and HMP vaginal datasets.

References

ABDI, H., O'TOOLE, A. J., VALENTIN, D. and EDELMAN, B. (2005). Distatis: The analysis of multiple distance matrices. In *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*. IEEE.

ANDERS, S. and HUBER, W. (2010). Differential expression analysis for sequence count data. *Genome biology* **11** R106.

ANDERSON, M. J., ELLINGSEN, K. E. and MCARDLE, B. H. (2006). Multivariate dispersion as a measure of beta diversity. *Ecology Letters* **9** 683–693.

ANDO, T. (2009). Bayesian factor analysis with fat-tailed factors and its exact marginal likelihood. *Journal of Multivariate Analysis* **100** 1717–1726.

ARBEL, J., MENGERSEN, K. and ROUSSEAU, J. (2014). Bayesian nonparametric dependent model for partially replicated data: the influence of fuel spills on species diversity. *arXiv preprint arXiv:1402.3093* .

BARRIENTOS, A. F., JARA, A., QUINTANA, F. A. ET AL. (2012). On the support of maceacherns dependent dirichlet processes and extensions. *Bayesian Analysis* **7** 277–310.

BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291.

BÖRNIGEN, D., MORGAN, X. C., FRANZOSA, E. A., REN, B., XAVIER, R. J., GARRETT, W. S. and HUTTENHOWER, C. (2013). Functional profiling of the gut microbiome in disease-associated inflammation. *Genome medicine* **5** 65.

BRIX, A. (1999). Generalized gamma measures and shot-noise cox processes. *Advances in Applied Probability* 929–953.

CALLAHAN, B. J., MCMURDIE, P. J., ROSEN, M. J., HAN, A. W., JOHNSON, A. J. A. and HOLMES, S. P. (2016). Dada2: High-resolution sample inference from illumina amplicon data. *Nature methods* **13** 581–583.

CAPORASO, J. G., KUCZYNSKI, J., STOMBAUGH, J., BITTINGER, K., BUSHMAN, F. D., COSTELLO, E. K., FIERER, N., PENA, A. G., GOODRICH, J. K., GORDON, J. I. and KNIGHT, R. (2010). Qiime allows analysis of high-throughput community sequencing data. *Nature methods* **7** 335–336.

CAPORASO, J. G., LAUBER, C. L., WALTERS, W. A., BERG-LYONS, D., LOZUPONE, C. A., TURNBAUGH, P. J., FIERER, N. and KNIGHT, R. (2011). Global patterns of 16s rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences* **108** 4516–4522.

CARPENTER, B., GELMAN, A., HOFFMAN, M., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M. A., GUO, J., LI, P. and RIDDELL, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software* **20**.

CARRERA, J., RODRIGO, G. and JARAMILLO, A. (2009). Model-based redesign of global transcription regulation. *Nucleic acids research* gkp022.

CARVALHO, C. M., CHANG, J., LUCAS, J. E., NEVINS, J. R., WANG, Q. and WEST, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association* **103**.

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**.

DALEY, D. J. and VERE-JONES, D. (1988). An introduction to the theory of point processes.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P. and Andersen, G. L. (2006). Greengenes, a chimera-checked 16s rRNA gene database and workbench compatible with arb. *Applied and environmental microbiology* **72** 5069–5072.

Dethlefsen, L., McFall-Ngai, M. and Relman, D. A. (2007). An ecological and evolutionary perspective on human–microbe mutualism and disease. *Nature* **449** 811–818.

Dethlefsen, L. and Relman, D. A. (2011). Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the National Academy of Sciences* **108** 4554–4561.

DiGiulio, D., Callahan, B. J., McMurdie, P. J., Costello, E. K., Lyell, D. J., Robaczewska, A., Sun, C. L., Goltsman, D. S. A., Wong, R. J., Shaw, G., Stevenson, D. K., Holmes, S. and R., R. D. A. (2015). Temporal and spatial variation of the human microbiota during pregnancy To appear.

Ding, T. and Schloss, P. D. (2014). Dynamics and associations of microbial community types across the human body. *Nature* **509** 357.

Eren, A. M., Borisy, G. G., Huse, S. M. and Welch, J. L. M. (2014). Oligotyping analysis of the human oral microbiome. *Proceedings of the National Academy of Sciences* **111** E2875–E2884.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics* 751–760.

Fang, H., Huang, C., Zhao, H. and Deng, M. (2015). Cclasso: correlation inference for compositional data through lasso. *Bioinformatics* btv349.

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

FAUST, K., SATHIRAPONGSASUTI, J. F., IZARD, J., SEGATA, N., GEVERS, D., RAES, J. and HUTTENHOWER, C. (2012b). Microbial co-occurrence relationships in the human microbiome. *PLoS Comput Biol* **8** e1002606.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The annals of statistics* 209–230.

FRIEDMAN, J. and ALM, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput Biol* **8** e1002687.

GELMAN, A. and RUBIN, D. B. (1992a). Inference from iterative simulation using multiple sequences. *Statistical science* 457–472.

GELMAN, A. and RUBIN, D. B. (1992b). Inference from iterative simulation using multiple sequences. *Statistical science* 457–472.

GORVITOVSKAIA, A., HOLMES, S. P. and HUSE, S. M. (2016). Interpreting prevotella and bacteroides as biomarkers of diet and lifestyle. *Microbiome* **4** 1.

GRICE, E. A. and SEGRE, J. A. (2011). The skin microbiome. *Nature Reviews Microbiology* **9** 244–253.

GRIFFIN, J. E., KOLOSSIATIS, M. and STEEL, M. F. J. (2013). Comparing distributions by using dependent normalized random-measure mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75** 499–529.

HAWKINS, D. and HAN, C.-P. (1986). Bivariate distributions of some ratios of independent noncentral chi-square random variables. *Communications in Statistics-Theory and Methods* **15** 261–277.

HOFFMAN, M. D. and GELMAN, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research* **15** 1593–1623.

HOLMES, I., HARRIS, K. and QUINCE, C. (2012a). Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PloS one* **7** e30126.

HOLMES, I., HARRIS, K. and QUINCE, C. (2012b). Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PloS one* **7** e30126.

HOLMES, S. (2008). Multivariate data analysis: the french way. In *Probability and statistics: Essays in honor of David A. Freedman*. Institute of Mathematical Statistics, 219–233.

HOOPS, S., SAHLE, S., GAUGES, R., LEE, C., PAHLE, J., SIMUS, N., SINGHAL, M., XU, L., MENDES, P. and KUMMER, U. (2006). Copasia complex pathway simulator. *Bioinformatics* **22** 3067–3074.

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

JAMES, L. F. (2002). Poisson process partition calculus with applications to exchangeable models and bayesian nonparametrics. *arXiv preprint math/0205093* .

JAMES, L. F., LIJOI, A. and PRÜNSTER, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics* **36** 76–97.

KINGMAN, J. (1967). Completely random measures. *Pacific Journal of Mathematics* **21** 59–78.

KOENIG, J. E., SPOR, A., SCALFONE, N., FRICKER, A. D., STOMBAUGH, J., KNIGHT, R., ANGENENT, L. T. and LEY, R. E. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences* **108** 4578–4585.

KOSTIC, A. D., GEVERS, D., SILJANDER, H., VATANEN, T., HYÖTYLÄINEN, T., HÄMÄLÄINEN, A., PEET, A., TILLMANN, V., PÖHÖ, P. and MATTILA, I. (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell host & microbe* **17** 260–273.

LA ROSA, P. S., BROOKS, J. P., DEYCH, E., BOONE, E. L., EDWARDS, D. J., WANG, Q., SODERGREN, E., WEINSTOCK, G. and SHANNON, W. D. (2012). Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PloS one* **7** e52078.

LAVIT, C., ESCOUFIER, Y., SABATIER, R. and TRAISSAC, P. (1994). The ACT (statis method). *Computational Statistics & Data Analysis* **18** 97–119.

LEDOUX, M. and TALAGRAND, M. (2013). Probability in banach spaces: isoperimetry and processes.

LEE, S. and SONG, X. (2002). Bayesian selection on the number of factors in a factor analysis model. *Behaviormetrika* **29** 23–39.

LI, C., DONIZELLI, M., RODRIGUEZ, N., DHARURI, H., ENDLER, L., CHELLIAH, V., LI, L., HE, E., HENRY, A., STEFAN, M. I. ET AL. (2010). Biomodels database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC systems biology* **4** 92.

LI, H. (2015a). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application* **2** 73–94.

LI, H. (2015b). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application* **2** 73–94.

LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2005). Hierarchical mixture modeling with normalized inverse-gaussian priors. *Journal of the American Statistical Association* **100** 1278–1291.

LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69** 715–740.

LIJOI, A. and PRÜNSTER, I. (2010). Models beyond the dirichlet process. In *Bayesian nonparametrics* (N. L. Hjort, C. Holmes, P. Müller and S. G. Walker, eds.), chap. 3. Cambridge University Press, 80–136.

LIN, W., SHI, P., FENG, R., LI, H. ET AL. (2014). Variable selection in regression with compositional covariates. *Biometrika* **101** 785–797.

LONG, J. and ROTH, M. (2008). Synthetic microarray data generation with range and nemo. *Bioinformatics* **24** 132–134.

LOPES, H. F. and WEST, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14** 41–68.

LOZUPONE, C. and KNIGHT, R. (2005). Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology* **71** 8228–8235.

LUCAS, J., CARVALHO, C., WANG, Q., BILD, A., NEVINS, J. R. and WEST, M. (2006). Sparse statistical modelling in gene expression genomics. *Bayesian Inference for Gene Expression and Proteomics* **1**.

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

MCMURDIE, P. J. and HOLMES, S. (2013). phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data. *PLOS one* **8** e61217.

MCMURDIE, P. J. and HOLMES, S. (2014a). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol* **10** e1003531.

MCMURDIE, P. J. and HOLMES, S. (2014b). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol* **10** e1003531.

MERCER, J. (1909). Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character* 415–446.

MEYER, P. E., KONTOS, K., LAFITTE, F. and BONTEMPI, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. *EURASIP journal on bioinformatics and systems biology* **2007** 79879.

MORGAN, X. C., TICKLE, T. L., SOKOL, H., GEVERS, D., DEVANEY, K. L., WARD, D. V., REYES, J. A., SHAH, S. A., LELEIKO, N., SNAPPER, S. B. ET AL. (2012a). Dysfunction of

the intestinal microbiome in inflammatory bowel disease and treatment. *Genome biology* **13** 1.

MORGAN, X. C., TICKLE, T. L., SOKOL, H., GEVERS, D., DEVANEY, K. L., WARD, D. V., REYES, J. A., SHAH, S. A., LELEIKO, N., SNAPPER, S. B. ET AL. (2012b). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome biology* **13** 1.

MORGAN, X. C., TICKLE, T. L., SOKOL, H., GEVERS, D., DEVANEY, K. L., WARD, D. V., REYES, J. A., SHAH, S. A., LELEIKO, N., SNAPPER, S. B. ET AL. (2012c). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome biology* **13** R79.

MULIERE, P. and TARDELLA, L. (1998). Approximating distributions of random functionals of ferguson-dirichlet priors. *Canadian Journal of Statistics* **26** 283–297.

MÜLLER, P., QUINTANA, F. and ROSNER, G. (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66** 735–749.

O'BRIEN, J. D., RECORD, N. and COUNTWAY, P. (2016). The power and pitfalls of dirichlet-multinomial mixture models for ecological count data. *bioRxiv* 045468.

OKSANEN, J., BLANCHET, F. G., KINDT, R., LEGENDRE, P., MINCHIN, P. R., O'HARA, R. B., SIMPSON, G. L., SOLYMOS, P., STEVENS, M. H. H. and WAGNER, H. (2015). vegan: Community Ecology Package.
URL `https://cran.r-project.org/web/packages/vegan/index.html`

PAULSON, J. N., STINE, O. C., BRAVO, H. C. and POP, M. (2013a). Differential abundance analysis for microbial marker-gene surveys. *Nature methods* **10** 1200–1202.

PAULSON, J. N., STINE, O. C., BRAVO, H. C. and POP, M. (2013b). Differential abundance analysis for microbial marker-gene surveys. *Nature methods* **10** 1200–1202.

PEARSON, K. (1896). Mathematical contributions to the theory of evolution.–on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the royal society of london* **60** 489–498.

PEIFFER, J. A., SPOR, A., KOREN, O., JIN, Z., TRINGE, S. G., DANGL, J. L., BUCKLER, E. S. and LEY, R. (2013). Diversity and heritability of the maize rhizosphere microbiome under field conditions. *Proceedings of the National Academy of Sciences* **110** 6548–6553.

PRESS, S. J. and SHIGEMASU, K. (1989). Bayesian inference in factor analysis. In *Contributions to probability and statistics*. Springer, 271–287.

QUINCE, C., LUNDIN, E. E., ANDREASSON, A. N., GRECO, D., RAFTER, J., TALLEY, N. J., AGREUS, L., ANDERSSON, A. F., ENGSTRAND, L. and D'AMATO, M. (2013). The impact of crohn's disease genes on healthy human gut microbiota: a pilot study. *Gut* 952–954.

RAVEL, J., GAJER, P., ABDO, Z., SCHNEIDER, G. M., K., S. S. K., MCCULLE, S. L., KARLEBACH, S., GORLE, R., RUSSELL, J., TACKET, C. O. and BROTMAN, R. M. (2011a). Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences* **108** 4680–4687.

RAVEL, J., GAJER, P., ABDO, Z., SCHNEIDER, G. M., KOENIG, S. S., MCCULLE, S. L., KARLEBACH, S., GORLE, R., RUSSELL, J., TACKET, C. O. ET AL. (2011b). Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences* **108** 4680–4687.

REGAZZINI, E., LIJOI, A. and PRÜNSTER, I. (2003). Distributional results for means of normalized random measures with independent increments. *Annals of Statistics* 560–585.

REN, B., BACALLADO, S., FAVARO, S., HOLMES, S. and TRIPPA, L. (2016). Bayesian nonparametric ordination for the analysis of microbial communities. *arXiv preprint arXiv:1601.05156* .

ROBERT, P. and ESCOUFIER, Y. (1976). A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Applied statistics* 257–265.

ROBINSON, M. D., MCCARTHY, D. J. and SMYTH, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26** 139–140.

RODRÍGUEZ, A., DUNSON, D. B. and GELFAND, A. E. (2009). Bayesian nonparametric functional data analysis through density estimation. *Biometrika* **96** 149–162.

ROSEN, M. J., CALLAHAN, B. J., FISHER, D. S. and HOLMES, S. (2012). Denoising pcr-amplified metagenome data. *BMC bioinformatics* **13** 283.

ROWE, D. B. (2002). *Multivariate Bayesian statistics: models for source separation and signal unmixing*. CRC Press.

SMYTH, G. (2005). Limma: linear models for microarray data. *Bioinformatics and computational biology solutions using R and Bioconductor* 397–420.

TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B. and CHU, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences* **99** 6567–6572.

TRUONG, D. T., FRANZOSA, E. A., TICKLE, T. L., SCHOLZ, M., WEINGART, G., PASOLLI, E., TETT, A., HUTTENHOWER, C. and SEGATA, N. (2015). Metaphlan2 for enhanced metagenomic taxonomic profiling. *Nature methods* **12** 902–903.

TURNBAUGH, P. J., HAMADY, M., YATSUNENKO, T., CANTAREL, B. L., DUNCAN, A., LEY, R. E., SOGIN, M. L., JONES, W. J., ROE, B. A., AFFOURTIT, J. P., EGHOLM, M., HENRISSAT, B., HEATH, A. C., KNIGHT, R. and GORDON, J. I. (2009a). A core gut microbiome in obese and lean twins. *Nature* **457** 480–484.

TURNBAUGH, P. J., LEY, R. E., HAMADY, M., FRASER-LIGGETT, C., KNIGHT, R. and GORDON, J. I. (2007). The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature* **449** 804.

TURNBAUGH, P. J., RIDAURA, V. K., FAITH, J. J., REY, F. E., KNIGHT, R. and GORDON, J. I. (2009b). The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Science translational medicine* **1** 6ra14–6ra14.

VAN DEN BULCKE, T., VAN LEEMPUT, K., NAUDTS, B., VAN REMORTEL, P., MA, H., VERSCHOREN, A., DE MOOR, B. and MARCHAL, K. (2006). Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC bioinformatics* **7** 43.

VATANEN, T., KOSTIC, A. D., DHENNEZEL, E., SILJANDER, H., FRANZOSA, E. A., YASSOUR, M., KOLDE, R., VLAMAKIS, H., ARTHUR, T. D., HÄMÄLÄINEN, A.-M. ET AL. (2016). Variation in microbiome lps immunogenicity contributes to autoimmunity in humans. *Cell* **165** 842–853.

WEISS, S., VAN TREUREN, W., LOZUPONE, C., FAUST, K., FRIEDMAN, J., DENG, Y., XIA, L. C., XU, Z. Z., URSELL, L., ALM, E. J. ET AL. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME journal* .

WILHELM, B., G. S. WITH CONTRIBUTIONS FROM MANJUNATH (2015). tmvtnorm: Truncated Multivariate Normal and Student t Distribution.
URL `https://cran.r-project.org/web/packages/tmvtnorm/index.html`

XIA, F., CHEN, J., FUNG, W. K. and LI, H. (2013a). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics* **69** 1053–1063.

XIA, F., CHEN, J., FUNG, W. K. and LI, H. (2013b). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics* **69** 1053–1063.

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** e0129606.