MIMIX: a Bayesian Mixed-Effects Model for Microbiome Data from Designed Experiments

Neal S. Grantham[1], Yawen Guan[1], Brian J. Reich[1], Elizabeth T. Borer[2], and Kevin Gross[1]

[1]Department of Statistics, North Carolina State University, Raleigh, NC
[2]Department of Ecology, Evolution, and Behavior, University of Minnesota, St. Paul, MN

May 20, 2019

Abstract

Recent advances in bioinformatics have made high-throughput microbiome data widely available, and new statistical tools are required to maximize the information gained from these data. For example, analysis of high-dimensional microbiome data from designed experiments remains an open area in microbiome research. Contemporary analyses work on metrics that summarize collective properties of the microbiome, but such reductions preclude inference on the fine-scale effects of environmental stimuli on individual microbial taxa. Other approaches model the proportions or counts of individual taxa as response variables in mixed models, but these methods fail to account for complex correlation patterns among microbial communities. In this paper, we propose a novel Bayesian mixed-effects model that exploits cross-taxa correlations within the microbiome, a model we call MIMIX (MIcrobiome MIXed model). MIMIX offers global tests for treatment effects, local tests and estimation of treatment effects on individual taxa, quantification of the relative contribution from heterogeneous sources to microbiome variability, and identification of latent ecological subcommunities in the microbiome. MIMIX is tailored to large microbiome experiments using a combination of Bayesian factor analysis to efficiently represent dependence between taxa and Bayesian variable selection methods to achieve sparsity. We demonstrate the model using a simulation experiment and on a 2x2 factorial experiment of the effects of nutrient supplement and herbivore exclusion on the foliar fungal microbiome of *Andropogon gerardii*, a perennial bunchgrass, as part of the global Nutrient Network research initiative.

Keywords: continuous shrinkage prior; factor analysis; microbiome; mixed model; Nutrient Network; OTU abundance data.

1 Introduction

A microbiome is a community of microorganisms and their genomes that belong to a particular ecological niche such as the human gut, soil, plants, or ambient dust. Samples collected from these habitats invariably contain thousands of archaea, bacteria, and fungi which may be identified through their DNA with next-generation sequencing technologies (Metzker, 2010). Understanding how these microbial communities interact with their environment holds significant implications for the fields of human health (Wu and Lewis, 2013), climate change (Bond-Lamberty et al., 2016), forensics (Grantham et al., 2015), and more. However, the tools available for characterizing microbiomes are, at present, largely limited to descriptive studies and must evolve to meet the advanced needs of the microbiome research community. To this end, the interdisciplinary Unified Microbiome Initiative (Alivisatos et al., 2015) aims to achieve "predictive understanding that allows evidence-based, model-informed microbiome management and design" by encouraging collaborative work on several promising areas of emphasis.

One such area of emphasis is the development of new statistical models for microbiome data analysis with environmental covariates, particularly in the presence of heterogeneous sources of variability. Microbiome data are difficult to model because they are high-dimensional, sparse, overdispersed, and possess complex dependence structure. Moreover, as a consequence of the next-generation sequencing technology, the data are compositional, meaning they convey relative rather than absolute information; a microbe's abundance in a sample (the number of times its DNA was read by the sequencer) depends on the sequencing depth (the total number of reads). Most standard multivariate statistical methods are designed for the analysis of absolute information and will yield spurious correlations among variables when applied indiscriminately to compositional data (Pearson, 1896).

In the face of these difficulties, contemporary analysis of microbiome data often works on metrics that summarize collective properties of the entire microbiome, such as measures of taxonomic diversity. For example, in ecology, within-sample diversity is most simply measured as the mean

number of unique taxa observed in a sample, referred to as α-diversity. Among-sample diversity, referred to as β-diversity, describes differences in taxonomic composition among samples, and may be quantified by measures like Bray-Curtis dissimilarity (Bray and Curtis, 1957) or, if full taxonomic assignments are available, UniFrac distance (Lozupone and Knight, 2005). Permutational multivariate analysis of variance (PERMANOVA) with pairwise differences between samples (McArdle and Anderson, 2001) is a popular tool to test whether environmental covariates are associated with significant differences in these summary metrics. However, PERMANOVA does not yield inferences about how the environment affects individual microbes. Additionally, implicit assumptions made by such distance-based multivariate methods may be inappropriate for ecological count data altogether (Warton et al., 2012).

More recently, and in a different vein, others have suggested analyzing the microbiome by fitting a separate generalized linear mixed model to the abundance of each taxon. For instance, a linear mixed model with arcsine square root transformation or, if sparsity and overdispersion are of particular concern, a zero-inflated beta model (E. Chen and Li, 2016) are viable methods for inferring treatment effects on the relative abundance (proportions) of taxa in the presence of random effects. Alternative approaches model the raw abundance (counts) directly, accounting for the uncertainty associated with a taxon's abundance by conditioning on the total reads per sample (McMurdie and Holmes, 2014). Hierarchical mixtures, such as beta-binomial and gamma-Poisson, are quite robust for this purpose and possess added flexibility for overdispersed data (Zhang et al., 2017).

An alternative to these univariate approaches is to model taxa within the microbiome jointly rather than individually. Unlike univariate models, multivariate models can pool information across taxa to increase power for detecting and estimating treatment effects. Towards this end, the Dirichlet-multinomial (DM) model — the multivariate extension of beta-binomial — provides a rich framework for modeling the entire vector of raw abundance data in each microbiome sample. For example, the DM has proven useful for microbiome analysis in the areas of hypothesis testing and power calculations (La Rosa et al., 2012), sparse variable selection (J. Chen and Li, 2013), inference of microbial community structure (Shafiei et al., 2015), and regression modeling (Wadsworth

et al., 2017). Despite their utility, without further hierarchical structure Dirichlet variates have the undesirable property that they must negatively co-vary, making them ill-suited for modeling microbial taxa that often have positive associations, perhaps because they share similar habitat niches or because they interact symbiotically.

Models with more flexible dependence structures among microbial taxa have recently been provided by Xia et al. (2013) and Ren et al. (2017). Xia et al. (2013) use a logistic normal multinomial (LNM) model that links the multinomial probability vector to a multivariate normal random variable, resulting in unconstrained occurrence probabilities on the linked scale. The covariance structure specified by Xia et al. (2013) captures both positive and negative associations among taxa, unlike the DM covariance. However, while suitable for small collections of taxa, their method for estimating this dependence structure is infeasible for high numbers of unique taxa produced by next-generation sequencing technologies. In a different vein, Ren et al. (2017) use dependent Dirichlet processes to develop a Bayesian nonparametric ordination that enables convenient visualization of differences among microbial communities. Mixed-model versions of any of these approaches — DM, LNM, or Bayesian nonparametric ordination — needed to analyze experiments following split-block designs have yet to be developed for microbiome data, owing to the difficulty of introducing random effects into the model hierarchy.

With these considerations in mind, we propose MIMIX (MIcrobiome MIXed model), a Bayesian mixed-effects model for analyzing microbiome data as a response variable in designed experiments. MIMIX achieves four scientific objectives: (1) global tests of whether experimental treatments affect microbiome composition; (2) local tests for treatment effects on individual taxa and estimation of such effects if present; (3) quantification of how different sources of variability contribute to the microbiome heterogeneity; and (4) characterization of latent structure in the microbiome, which may suggest ecological subcommunities. MIMIX is a LNM mixed model that uses Bayesian factor analysis (Rowe, 2002) to capture complex dependence patterns among microbial taxa. Specifically, MIMIX models high-dimensional relationships among the transformed abundance probabilities of individual taxa through a set of low-dimensional unobservable variables, or factors. MIMIX natu-

rally identifies clusters of microbes that respond similarly to experimental conditions by developing continuous shrinkage Dirichlet-Laplace priors (Bhattacharya et al., 2015) for these latent factors. We then apply Bayesian variable selection priors for the fixed effects on subpopulation abundance, reflecting the prior belief that treatments will not affect all ecological communities. In this paper, these objectives and features of MIMIX are motivated by a multi-location randomized complete block design (RCBD) experiment that seeks to identify the influence of nutrient supplement and herbivory on the foliar fungal microbiome of a common perennial prairie bunchgrass.

The paper proceeds as follows. Section 2 introduces the data that motivate our development of MIMIX in Section 3. Section 4 demonstrates our method on simulated data in comparison with competing microbiome data analysis methods. Finally, we apply MIMIX to RCBD experiment data in Section 5 and close with a discussion in Section 6. Details of posterior sampling schemes are left to the Appendix, and open-source code to reproduce the statistical analyses in this paper is available online at `https://www.github.com/nsgrantham/mimix`.

2 Motivating Data

The Nutrient Network (NutNet, `www.nutnet.org`) is a global research cooperative hosted at the University of Minnesota that uses a standardized experimental protocol to study the impact of human activity at over 100 grassland sites spanning 6 continents (Borer et al., 2014). This article is motivated by data collected at 4 of these sites, all in the central US (Iowa, Kansas, Kentucky, and Minnesota). Each of these 4 sites features a 2×2 factorial experiment that crosses a nutrient-supplement treatment (i.e., fertilization) with an herbivore-exclusion treatment in a randomized complete block design (RCBD) (Figure 1). Here, we consider the effect of the two experimental treatments on the foliar fungal microbiome of *Andropogon gerardii*, a perennial bunchgrass found at each site, and native to prairie ecosystems of central North America.

In August 2014, leaf samples were collected from four *A. gerardii* individuals in each treatment plot. Samples were collected from plots in three replicated blocks, except in Iowa where *A. gerardii*

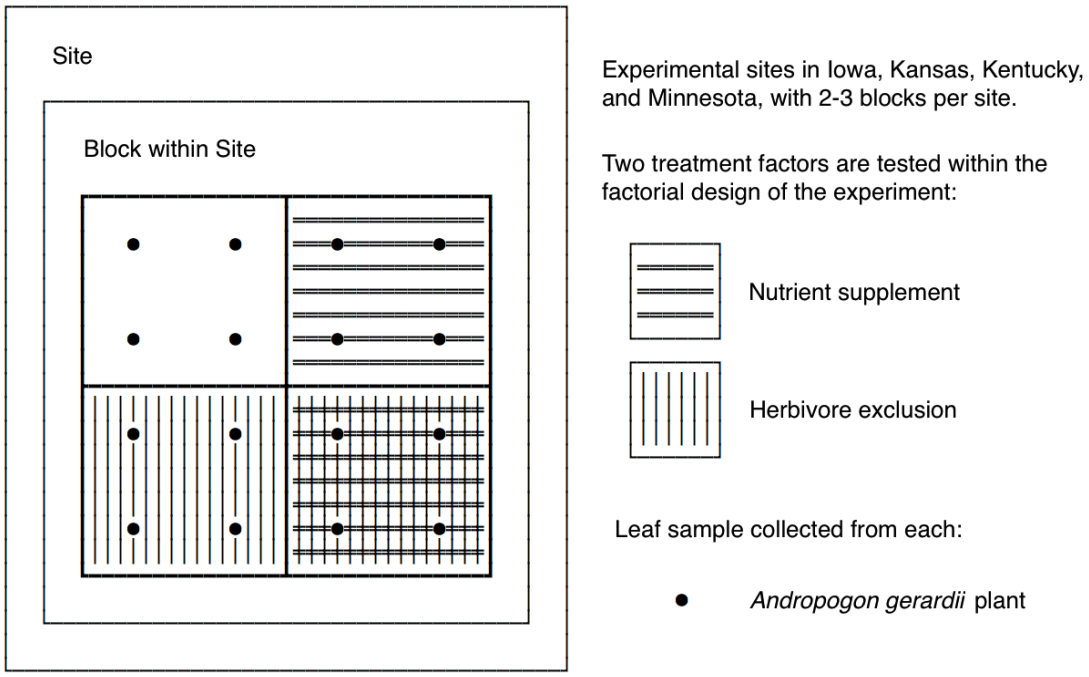


Figure 1: A schematic representation of the Nutrient Network experimental design. This experiment replicates a 2x2 RCB design across several sites. Four plants are sampled from each plot and a microbial sample is collected from each plant late in the growing season.

was present in only two blocks. Fungal rDNA was amplified and sequenced from each sample, and counts of operational taxonomic units (OTUs) within each sample were recorded. (Details of the molecular methods and bioinformatics pipeline are provided in the Supplement.) Ten samples were later removed from the study due to errors during sequencing, leaving a total of 166 leaf samples. Overall, 2,662 fungal OTUs were identified across the 166 samples. Samples contained a median of 74,099 separate reads, and harbored an average of 200 unique OTUs. Many OTUs were rare, as 85% of OTUs were identified in $<10\%$ of samples.

Given the preliminary OTU assignments, we wish to investigate each of the following using these data. First, we seek to characterize how the experimental treatments affect microbiome communities. We perform this analysis in stages: first at a global level where the response is the composition of the microbiome community as a whole, and then (if the global test identifies a significant treatment effect) at a local level that characterizes the effects on the relative abundance of individual OTUs. Second, we wish to characterize how the residual variation in the microbiome composition

varies among blocks within sites and across sites, as quantifying these sources of variation may suggest insight into the ecological processes controlling these microbial communities. Finally, we wish to characterize the dependence structure among OTUs, and identify clusters of OTUs that may suggest underlying ecological subcommunities.

3 Methods

Let Y_{ik} denote the count for sample $i = 1, \ldots, n$ and taxon $k = 1, ..., K$, and let $m_i = \sum_{k=1}^{K} Y_{ik}$ be the total counts for sample i. The value of m_i is an artifact of the sequencing depth of the high-throughput sequencing process and thus analyses are performed conditional on its value. For observation i, let $\boldsymbol{x}_i$ be a p-vector of covariates and let $z_i \in \{1, \ldots, q\}$ record the source of the random effects from one of q blocking factors. This latter notation may be generalized to accommodate arbitrarily complex blocking designs, but for notational simplicity we assume a single blocking factor in this initial development.

A multinomial likelihood is natural for multivariate count data, so we take $\mathbf{Y}_i = (Y_{i1}, ..., Y_{iK})' \sim$ Multinomial$(m_i, \boldsymbol{\phi}_i)$ where $\boldsymbol{\phi}_i = (\phi_{i1}, ..., \phi_{iK})'$ is the vector of expected proportions with $\boldsymbol{\phi}_i \in \mathbb{S}^K = \{(\phi_1, \ldots, \phi_K)' : \phi_k \geq 0, k = 1, \ldots, K, \sum_{k=1}^{K} \phi_k = 1\}$. We define sample-specific $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iK})' \in \mathbb{R}^K$ mapped to $\mathbb{S}^K$ by the inverse log-ratio transformation (Aitchison, 1986)

$$\phi_{ik} = \phi_k(\boldsymbol{\theta}_i) = \frac{\exp(\theta_{ik})}{\sum_{l=1}^{K} \exp(\theta_{il})} \quad \text{for} \quad k = 1, \ldots, K. \tag{1}$$

There is a loss of dimension in transforming from $\mathbb{R}^K$ to $\mathbb{S}^K$ due to the latter's unit-sum constraint. The likelihood is invariant to adding a constant to the parameters θ_{ik}; however, as the prior mean of the average $\sum_k \theta_{ik}$ is 0, the parameters will tend to be centered around 0 in the posterior.

In the spirit of Billheimer et al. (2001), we associate fixed and random effects with the microbiome composition through the mean of $\boldsymbol{\theta}_i$. The mixed effects decomposition is given by $\boldsymbol{\theta}_i = \boldsymbol{\mu} + \boldsymbol{\beta}\boldsymbol{x}_i + \boldsymbol{\gamma}_{z_i} + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)'$ is the overall population mean, $\boldsymbol{\beta}$ is a $K \times p$ matrix of unknown

fixed effect coefficients, $\boldsymbol{\gamma}_r$ is a K-vector of random effects from block r, and $\boldsymbol{\epsilon}_i \overset{iid}{\sim} N_K(\mathbf{0}, \boldsymbol{\Omega})$ is sample-specific random variation. Conditioned on the random effects, the regression coefficient for covariate j and taxon k, β_{jk}, has the usual logistic regression interpretation when reducing compositional data to the binary outcome that a sample comes from taxon k versus any of the other $K-1$ taxa. That is, with all else fixed, if the j^{th} covariate increases by one the log odds of a sample coming from taxon k increase by β_{jk}.

The number of taxa, K, is often very large in microbiome compositions. To address complications due to high dimensionality and to account for relationships among taxa, we use Bayesian factor analysis (Rowe, 2002) to model the fixed and random effects within a lower dimensional representation. For a number of factors L, let $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_L)$ be the $K \times L$ latent factor loading matrix, unknown and to be estimated. Suppose $\boldsymbol{\Lambda}$ is common to all fixed and random components of the model, i.e., $\boldsymbol{\beta} = \boldsymbol{\Lambda}\boldsymbol{b}$, $\boldsymbol{\gamma}_r = \boldsymbol{\Lambda}\mathbf{g}_r$, $r = 1, \ldots, q$, and $\boldsymbol{\epsilon}_i = \boldsymbol{\Lambda}\boldsymbol{e}_i + \boldsymbol{\delta}_i$, $i = 1, \ldots, n$. Then we may represent $\boldsymbol{\theta}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f}_i + \boldsymbol{\delta}_i$ where $\mathbf{f}_i = \boldsymbol{b}\boldsymbol{x}_i + \mathbf{g}_{z_i} + \boldsymbol{e}_i$ is the low-dimensional factor score vector for sample i. Under this common factor structure each latent factor captures sets of taxa correlated in their response to the model's covariates and other sources of variability. This shared-factor assumption is not inherent to our approach and if separate factors are thought to drive the fixed and random effects then these two components can be modelled separately.

A prior on $\boldsymbol{\Lambda}$ should ensure that the factor loading matrix captures common, cross-species covariance that lends itself to post-hoc inference of collective taxa responses. For instance, setting entire columns of $\boldsymbol{\Lambda}$ to zero is a means of selecting the number of active factors and setting individual elements within $\boldsymbol{\Lambda}$ to zero allows the factors to represent subsets of taxa (Carvalho et al., 2008). To achieve both forms of sparsity, we place continuous shrinkage priors on the high-dimensional factor loadings $\boldsymbol{\lambda}_l, l = 1, \ldots, L$ comprising $\boldsymbol{\Lambda}$. In particular, we select a Dirichlet-Laplace prior (Bhattacharya et al., 2015) for its ability to detect sparse signals in high-dimensional linear regression, which we modify here for factor analysis. For factors $l = 1, \ldots, L$, let $\boldsymbol{\lambda}_l \sim \mathrm{DL}_{a_l}$ represented by

$$\lambda_{kl} \mid \xi_{kl}, \tau_l \sim \mathrm{Lap}\left(\xi_{kl}\tau_l\right), \ \boldsymbol{\xi}_l = (\xi_{1l}, \ldots, \xi_{Kl})' \sim \mathrm{Dir}(a_l, \ldots, a_l), \text{ and } \tau_l \sim \mathrm{Gam}(Ka_l, \nu)$$

for $k = 1, \ldots, K$, where $\nu \sim \mathcal{G}(c_0, d_0)$ and each a_l is given a discrete uniform prior over $(0, 1)$ with smaller values favoring aggressive shrinkage of terms toward zero. The Laplace distribution may be equivalently represented as a scale mixture of normals with exponential mixing density,

$$\lambda_{kl} \mid \psi_{kl}, \xi_{kl}, \tau_l \sim N\left(0, \psi_{kl}\xi_{kl}^2\tau_l^2\right) \quad \text{and} \quad \psi_{kl} \sim \text{Exp}(1/2),$$

a form that allows for straightforward Gibbs sampling of the associated parameters.

Another aim of this model is to test and quantify treatment effects on each taxon, so we place a spike-and-slab prior on $\boldsymbol{b}$ for the purposes of stochastic variable selection (Mitchell and Beauchamp, 1988). Unlike the DL prior, the spike-and-slab prior places probability on the coefficients being exactly zero. This allows us to compute posterior probabilities that effects are zero leading to a Bayesian test, satisfying one of MIMIX's objectives. Let ω_{jl} be an indicator variable taking the value 1 (0) when b_{jl} is included (excluded) from the model. The spike-and-slab prior is given by $Pr(\omega_{jl} = 0) = 1 - \pi_j$ and $b_{jl} \mid \omega_{jl} = 1 \sim N(0, \sigma_b^2)$, where $\pi_j \sim \text{Beta}(a_0, b_0)$ is the inclusion probability for covariate j in the model. We select a_0 and b_0 such that the prior inclusion probability for each covariate is set at some $c \in [0, 1]$. In particular, the number of factors for which fixed effect j is active, $S_j = \sum_{l=1}^{L} \omega_{jl}$, follows a beta-binomial distribution such that $Pr(S_j > 0) = 1 - Pr(S_j = 0) = 1 - \frac{\Gamma(L+b_0)\Gamma(a_0+b_0)}{\Gamma(L+a_0+b_0)\Gamma(b_0)}$, where $\Gamma(\cdot)$ is the gamma function. Fixing this quantity at c and selecting $a_0 = 1$ for convenience, we use the property that $\Gamma(n + 1) = n\Gamma(n)$ for any positive integer n to arrive at $b_0 = (\frac{1-c}{c})L$.

Placing priors on the remaining parameters and hyperparameters of this model completes the Bayesian specification. We use continuous priors for the intercept and random effects terms as we do not intend to test whether or not they are zero. Let $\mu_k \overset{iid}{\sim} N(0, \sigma_\mu^2)$, $g_{rl} \overset{iid}{\sim} N(0, \sigma_g^2)$, $e_{il} \overset{iid}{\sim} N(0, \sigma_e^2)$ where $\sigma_e^2 = 1$ to identify the scale of $\boldsymbol{\Lambda}$, and $\delta_{ik} \overset{ind}{\sim} N(0, \sigma_k^2)$ where σ_k^2 can capture over-dispersion

in the sequence reads of taxon k. These choices induce covariance

$$Cov(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i'} | \boldsymbol{\mu}, \boldsymbol{b}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \sigma_g) = \begin{cases} (\sigma_g^2 + 1)\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Sigma} & i = i' \\ \sigma_g^2 \boldsymbol{\Lambda}\boldsymbol{\Lambda}' & i \neq i' \text{ and } z_i = z_{i'} \\ \mathbf{0} & z_i \neq z_{i'} \end{cases}$$

where $\boldsymbol{\Sigma}$ is the $K \times K$ diagonal matrix with diagonal elements $(\sigma_1^2, \ldots, \sigma_K^2)$. The expression of the covariance of $\boldsymbol{\theta}_i$ illustrates that the product $\boldsymbol{\Lambda}\boldsymbol{\Lambda}'$ but not the individual elements of $\boldsymbol{\Lambda}$ are identified; we therefore use the posterior of $\boldsymbol{\Lambda}\boldsymbol{\Lambda}'$ to summarize the posterior of the covariance structure. We suppose the variance terms of the model follow independent inverse gamma priors with shape u_0 and scale v_0. The most important tuning parameters are the inclusion probability, c, and the number of latent factors, L. We set $c = 0.5$ so that each hypothesis has the same prior probability which is reasonable in our studies with a small number of fixed effects. The number of latent factors, L, is set to the minimum of the number of samples and the number of taxa (i.e., the maximum number of identifiable factors) and allows the Dirichlet-Laplace shrinkage prior to eliminate unnecessary factors.

Posterior sampling is conducted using Markov chain Monte Carlo (MCMC). Most terms in this model formulation are conjugate and are updated via Gibbs sampling. One exception is $\boldsymbol{\theta}_i$ which we update with Hamiltonian Monte Carlo (HMC) (Neal et al., 2011). The details of these sampling schemes are found in the Online Supplement. Code to perform MCMC is written in Julia (Bezanson et al., 2017) and available online at `https://www.github.com/nsgrantham/mimix`.

4 Simulation Study

We test our model on a simulated experiment with $K = 100$ taxa, $p = 1$ fixed effect, one blocking factor that takes $q = 5$ levels, and $n = 40$ observations with 8 assigned to each block. Within each block exists a balanced experiment with two levels of a single experimental factor, where $x_i = 1$

if observation i receives one level of the experimental factor, and $x_i = 0$ otherwise, and $z_i = r$ if observation i belongs to block r.

The fixed treatment effect, $\boldsymbol{\beta}$, is a sparse K-vector with a varying percentage of non-zero elements: 0% dense ($\boldsymbol{\beta} = \mathbf{0}$, i.e., no signal), 10% dense, or 20% dense. To generate $\boldsymbol{\beta}$ that is 10% dense, we partition the taxa into 20 clusters of 5 taxa each, select two of the twenty groups at random, and within each group draw a value from $v \sim \text{Unif}([-3,-1) \cup (1,3])$ to signify the group's collective response to the fixed treatment. That is, $\beta_k = v$ if taxon k belongs to the selected group and $\beta_k = 0$ otherwise. For 20% dense, we do this for four total groups. These groups are designed to represent taxa with shared phylogenetic ancestry or taxa that react similarly to the fixed effect. In the Online Supplement, we also investigate a scenario in which OTUs respond to the treatment individually, instead of in groups.

For the random blocking effects, we draw $\boldsymbol{\gamma}_r \overset{iid}{\sim} N_K(\mathbf{0}, \boldsymbol{\Sigma}_\gamma)$ with autoregressive covariance $(\boldsymbol{\Sigma}_\gamma)_{kk'} = \sigma^2_\gamma \rho_\gamma^{|k-k'|}$, $\rho_\gamma = 0.9$. Block-to-block variability, σ^2_γ, is set at 1 (medium) or 4 (high). For each observation i, define $\boldsymbol{\theta}_i = \boldsymbol{\mu} + \boldsymbol{\beta} x_i + \boldsymbol{\gamma}_{z_i} + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\mu}$ is a vector of length K with equally-spaced steps from 1 to -1 and $\boldsymbol{\epsilon}_i \overset{iid}{\sim} N_K(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$ with autoregressive covariance $(\boldsymbol{\Sigma}_\epsilon)_{kk'} = \sigma^2_\epsilon \rho_\epsilon^{|k-k'|}$. We fix $\rho_\epsilon = 0.9$ and examine sample-to-sample variability, σ^2_ϵ, over 1 (medium), 4 (high), and 9 (very high). Finally, we arrive at the final data by drawing each vector of counts $\mathbf{Y}_i$ from a multinomial distribution with total counts m_i chosen at-random from $2,500$ to $5,000$ and taxa proportions $\boldsymbol{\phi}_i$ calculated according to (1). We do not generate data directly from the MIMIX model to test for robustness to model assumptions (e.g., fixed and random effect correlation); data generated assuming a low-dimensional dependence structure would unduly favor MIMIX over the competitors.

In total, we examine each of 18 factor combinations (0, 10, and 20 % dense, medium/high block variance, medium/high/very high error variance) with 50 replications and compare the performance of three competing microbiome data analysis methods:

1. **PERMANOVA**: Permutational multivariate analysis of variance (McArdle and Anderson, 2001), or PERMANOVA, with Bray-Curtis dissimilarity (BC), a common analysis procedure

in ecology. This method tests the null hypothesis that there is no difference between the centroids of the treatment and control groups in BC space, but is unequipped to perform parameter estimation. It is implemented by `adonis` in the `R` package `vegan 2.3-5` `R`, among other available software options.

2. **MIMIX**: Our Bayesian mixed-effects model as presented in Section 3 with $L = 40$ factors. 10,000 posterior samples are collected with 5,000 removed for burn-in.

3. **MIMIX w/o Factors**: Bayesian mixed-effects model with no factors. This formulation mimics the available mixed model approaches to microbiome data analysis that do not account for dependence patterns among taxa, i.e., $\boldsymbol{\Lambda} = \mathbf{I}_K$ and $\boldsymbol{e}_i = \mathbf{0}$ for all $i = 1, \ldots, n$. 10,000 posterior samples are collected with 5,000 removed for burn-in.

Additional simulation results are given in the Online Supplement.

The methods are first evaluated on their power/type I error of a global test for treatment effect where PERMANOVA rejects for p-value < 0.05 and MIMIX and MIMIX w/o Factors reject if $Pr(\boldsymbol{\beta} \neq \mathbf{0} \mid \mathbf{Y}) > 0.9$. *We reject if* $Pr(\boldsymbol{\beta} \neq \mathbf{0} \mid \mathbf{Y}) > 0.9$ *to construct a test that has roughly the same size as the PERMANOVA test. Adopting this rule,* both MIMIX and MIMIX w/o Factors regularly outperform PERMANOVA in detecting the presence of a significant signal (Figure 2). In situations with medium error variance, MIMIX and MIMIX w/o Factors achieve similar power regardless of the block variance. As error variance increases, MIMIX is more likely than MIMIX w/o Factors to correctly identify the presence of a significant treatment effect.

We further compare MIMIX and MIMIX w/o Factors on their local tests and estimation of treatment effects for each OTU. Several metrics are considered: root mean squared error, RMSE $= \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\hat{\beta}_k^{\text{mean}} - \beta_k)^2}$ where $\hat{\beta}_k^{\text{mean}}$ is the posterior mean of β_k, coverage of 95% credible intervals, C95 $= \frac{1}{K}\sum_{k=1}^{K} I(\hat{\beta}_k^{0.025} < \beta_k < \hat{\beta}_k^{0.975})$ where $\hat{\beta}_k^q$ is the posterior qth quantile of β_k, and the true positive rate (TPR) and true negative rate (TNR) of local tests that reject if the 95% credible interval for β_k excludes zero. Table 1 gives the values of these metrics averaged over 50 replications at each factor combination.

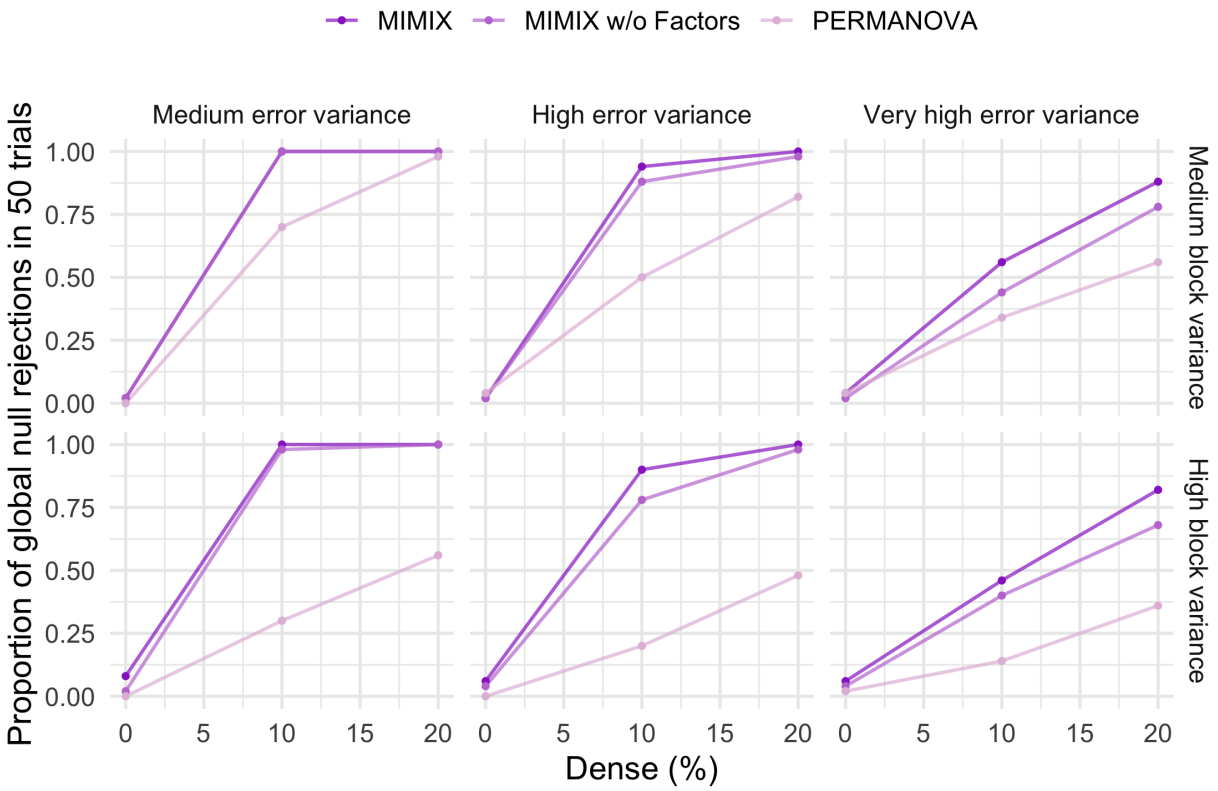


Figure 2: Results of a global test for treatment effect by MIMIX, MIMIX w/o Factors, and PERMANOVA, under a variety of simulation conditions. When $\boldsymbol{\beta} = \mathbf{0}$ (0% Dense) the line gives the test's type I error. For $\boldsymbol{\beta} \neq \mathbf{0}$ (>0% Dense), the line values depict the statistical power of each test.

When there is no signal present in the fixed effects (i.e., $\boldsymbol{\beta} = \mathbf{0}$, or 0% dense), MIMIX w/o Factors achieves lower RMSE on average than MIMIX in estimating all treatment effects to be zero, and both methods yield credible intervals that nearly always correctly include zero. In practice, these estimates are inconsequential if the global test appropriately fails to identify a significant treatment effect. When a proportion of OTUs are affected by the treatment (10% dense and 20% dense), MIMIX regularly outperforms MIMIX w/o Factors in the detection and estimation of these non-zero fixed effects. Specifically, when error variance is medium, the TPR of MIMIX is very high (85.7% to 94.0%) and beats MIMIX w/o Factors (68.6% to 79.5%), whereas the RMSE of the two methods are comparable. For high error variance, the TPR drops to about 50% (MIMIX) and 20% (MIMIX w/o Factors), with higher RMSE achieved by both methods, as expected, but lower RMSE obtained by MIMIX on average. This trend continues with very high error variance, resulting in

Table 1: Local test and estimation performance for MIMIX and MIMIX w/o Factors under a variety of simulation conditions as measured by root mean squared error (RMSE), coverage of 95% credible intervals (C95), true positive rate (TPR), and true negative rate (TNR). All values are multiplied by 100.

			Medium block variance				High block variance			
Dense	Error var.	Method	RMSE	C95	TPR	TNR	RMSE	C95	TPR	TNR
0%	Medium	MIMIX	0.3	100.0		100.0	0.6	99.6		99.6
		w/o Factors	0.0	100.0		100.0	0.1	100.0		100.0
	High	MIMIX	0.9	100.0		100.0	1.9	99.9		99.9
		w/o Factors	0.1	100.0		100.0	0.5	100.0		100.0
	Very High	MIMIX	2.0	100.0		100.0	3.3	99.9		99.9
		w/o Factors	0.2	100.0		100.0	0.5	100.0		100.0
10%	Medium	MIMIX	2.2	98.8	92.0	99.8	3.7	98.3	86.8	99.6
		w/o Factors	2.8	98.6	76.4	100.0	3.9	98.7	68.6	100.0
	High	MIMIX	10.3	97.3	53.6	99.6	14.4	96.1	42.6	99.4
		w/o Factors	14.6	96.5	24.8	100.0	17.9	96.0	18.6	100.0
	Very High	MIMIX	23.6	95.5	22.2	99.8	27.4	95.0	14.0	99.7
		w/o Factors	28.9	93.9	6.2	100.0	31.7	93.5	2.2	100.0
20%	Medium	MIMIX	3.9	98.0	94.0	99.4	6.3	96.7	85.7	99.0
		w/o Factors	3.6	98.7	79.5	100.0	6.0	98.0	70.7	100.0
	High	MIMIX	16.2	95.6	58.9	99.4	20.3	94.5	54.6	99.3
		w/o Factors	21.1	95.3	30.1	100.0	26.8	94.5	22.8	100.0
	Very High	MIMIX	36.1	93.0	26.7	99.5	41.7	91.7	18.7	99.6
		w/o Factors	50.5	90.0	6.3	100.0	56.9	88.8	3.4	100.0

lower TPR at around 20% and 5% respectively, and greater difference in RMSE in favor of MIMIX. Both methods are strongly conservative, achieving TNRs that are overwhelmingly 100%, and their C95 is often greater than the expected 95%, except in extreme variance situations (very high error variance, high block variance) with 20% dense fixed effects.

From the global and local simulation results, we draw three broad conclusions about the performance of MIMIX, MIMIX w/o Factors, and PERMANOVA for microbiome data analysis in designed experiments. First, at a global level, MIMIX and MIMIX w/o Factors achieve far greater power and comparable type I error (about 0.05) to PERMANOVA. Moreover, the global tests for all three methods appear more adversely affected by higher overdispersion in taxa counts than higher variability introduced by blocking factors in the experimental design. Second, at the local level, MIMIX is better suited for both the detection and estimation of sparse treatment effects compared to MIMIX w/o Factors. In this case, MIMIX w/o Factors achieves lower TPR and higher RMSE on average because it does not account for correlation patterns among taxa. Finally, TNR and C95

are very high and relatively consistent between the two methods under all simulation conditions, suggesting MIMIX and MIMIX w/o Factors are conservative in detecting significant OTU-specific fixed effects.

5 Analysis of the NutNet Experiment

We first compare the performance of MIMIX and MIMIX w/o Factors on the NutNet data through five-fold cross-validation, setting the maximum number of latent factors (L) for MIMIX equal to the number of samples (166). Specifically, for each $\mathbf{Y}_i$ with total counts m_i we construct $\mathbf{Y}_i^{\text{test}f}$, $f = 1, \ldots, 5$ by assigning each of m_i observations to one of the five folds at random and let $\mathbf{Y}_i^{\text{train}f} = \sum_{g \neq f} \mathbf{Y}_i^{\text{test}g}$. For each fold $f = 1, \ldots, 5$, we fit both models to training data f, drawing 20,000 posterior samples and discarding the first 10,000 for burn-in. Next, we examine the difference of their log-likelihoods (MIMIX minus MIMIX w/o Factors) evaluated on testing data f where the multinomial probability vector is estimated by the normalized posterior mean vector of occurrence probabilities $\hat{\boldsymbol{\phi}}_i$. Over all five folds, 69% of differences are positive on average, favoring MIMIX. Thus, MIMIX appears to be a more apt model for these data than MIMIX w/o Factors.

To assess whether MIMIX adequately captures the sparsity in the data, we fit a preliminary model to the data and perform posterior predictive checks (Gelman et al., 2014). These checks examine the proportion of OTUs within each sample with zero counts (sparsity). This is done by predicting new $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ from every posterior sample and comparing the sparsity and overdispersion in these predicted samples with the observed data. With respect to sparsity, we also consider the proportion of OTUs within each sample with two or fewer counts, as these singletons and doubletons are thought by biologists to be generated by errors in the sequencing process. Figure 3 depicts posterior predictive checks on sparsity of MIMIX after 20,000 posterior samples with the first 10,000 removed for burn-in. MIMIX does not accurately estimate the proportion of OTUs within each sample with exactly zero counts, but when singleton and doubleton counts are further considered the model recovers the observed near-sparsity of the original data. The distinction between zero

counts and two or fewer counts is likely of little consequence. Additional posterior predictive checks on the maximum proportion of total counts within a sample from a single OTU (overdispersion) and the average Bray-Curtis similarity between samples from the same site, block, and treatment group also suggest that the marginal and joint distributions of OTU counts are captured faithfully by the MIMIX fit (results in the Online Supplemental).

We now use MIMIX to characterize the effects of the nutrient-supplement and herbivore-exclusion treatments on the fungal foliar microbiome of *A. gerardii*. For the purposes of comparison, we also present analyses from Bray-Curtis PERMANOVA, which represents the current state-of-the-art in ecological analysis, and MIMIX w/o Factors. We present MIMIX w/o Factors to illustrate the consequences of using factor analysis to account for dependence patterns among taxa in the microbiome, but we emphasize that our simulation studies and preliminary analysis point towards MIMIX as the most trustworthy analysis. For the Bayesian models, we collect 20,000 posterior samples and discard the first 10,000 for burn-in.

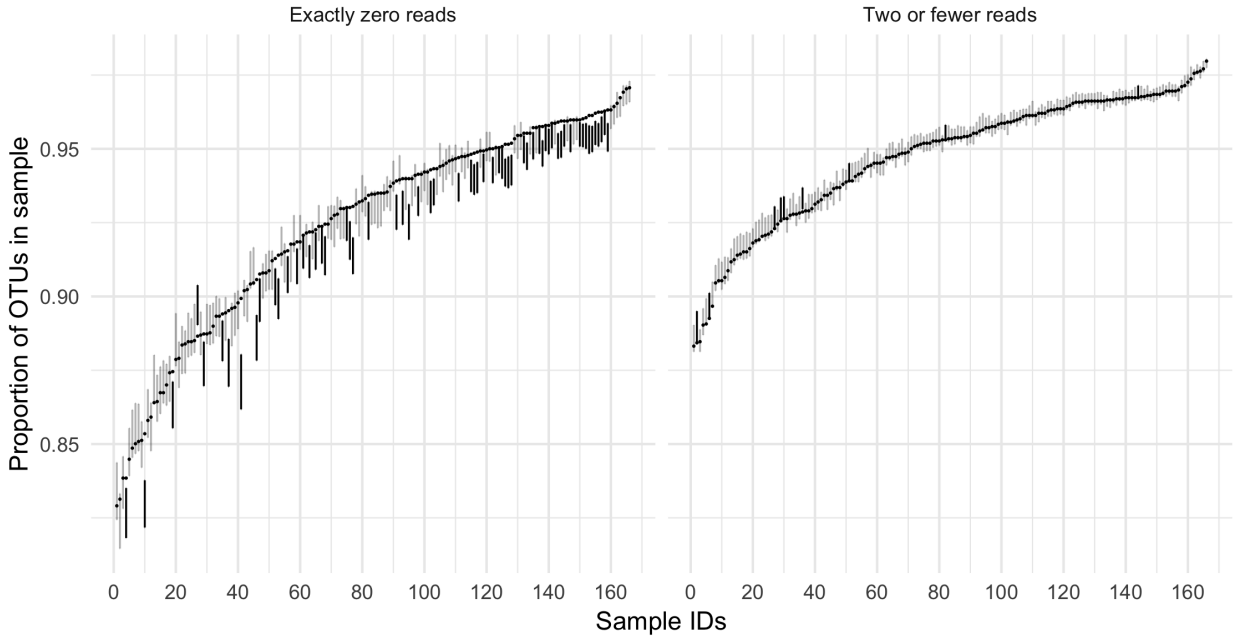
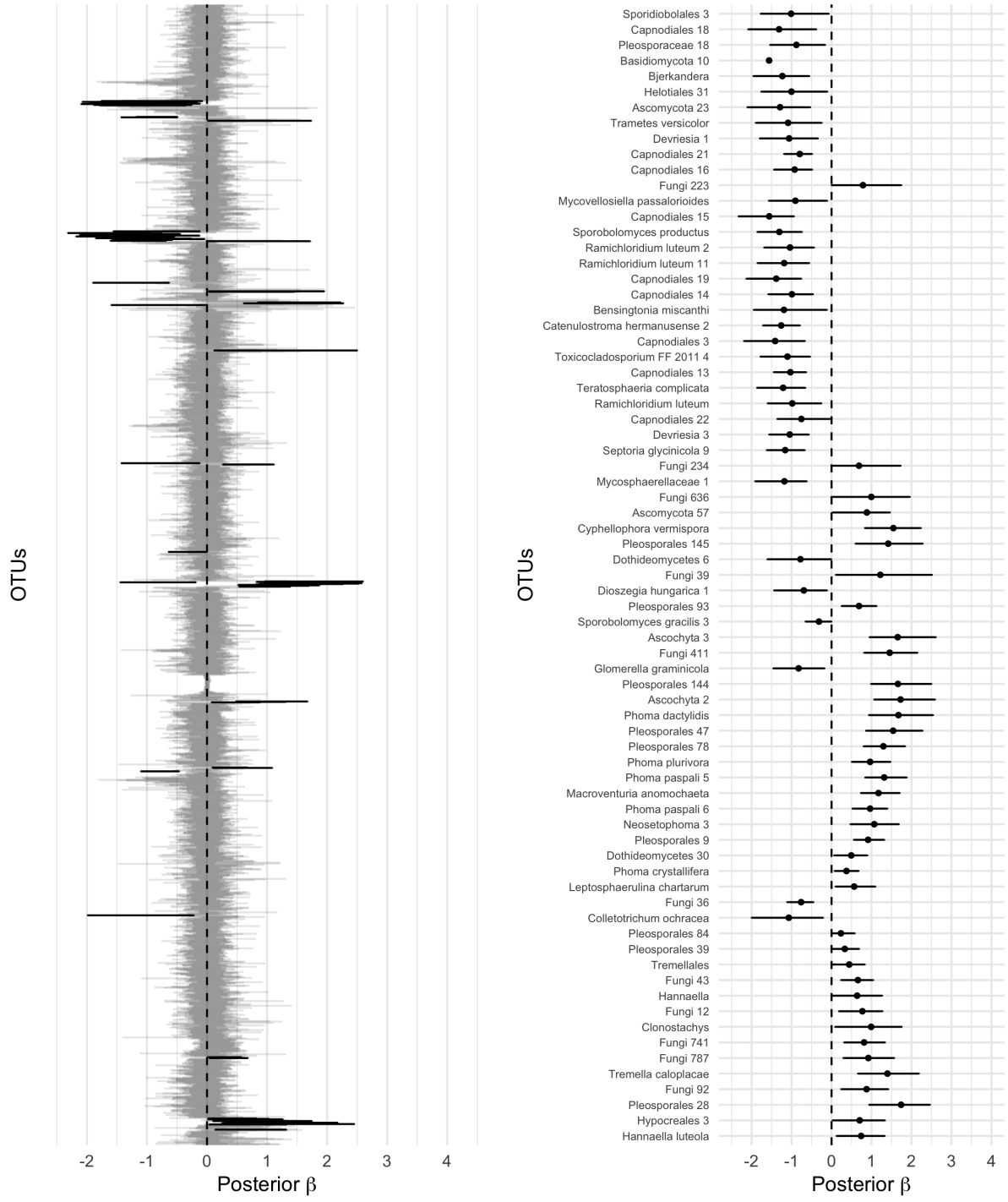


Figure 3: Posterior predictive checks on sparsity for MIMIX applied to the NutNet data. For each sample, the dot indicates the proportion of OTUs in the sample with 0 (left) or ≤ 2 (right) reads. The vertical line is the 95% posterior predictive distribution, shaded black if the interval excludes the observed value and gray otherwise.

First, we conduct a global test of whether the experimental treatments affect the overall composition of the microbiome. No method identifies a significant interaction effect between the two treatments, with PERMANOVA $p = 0.120$, MIMIX posterior probability 0.236, and MIMIX w/o Factors posterior probability 0.413. However, PERMANOVA and MIMIX find strong evidence that the fungal microbiome composition is affected by nutrient supplement, with $p = 0.003$ and posterior probability 1.0, respectively, while MIMIX w/o Factors does not, with posterior probability 0.757. No method detects an effect of herbivore exclusion, with PERMANOVA $p = 0.787$, MIMIX posterior probability 0.196, and MIMIX w/o Factors posterior probability 0.523. These results suggest that the composition of the foliar fungal microbiome of *A. gerardii* is impacted by the resources available to the plant host.

We next use MIMIX to estimate the effects of nutrient supplement on individual fungal OTUs. Because the OTU assignments for this particular data set are only preliminary, we focus here on the distribution of OTU-level effects, and reserve the characterization of effects on specific OTUs for later work. The fixed effect for nutrient supplement estimated by MIMIX has 95% credible intervals that exclude zero for 73 OTUs (Figure 4). Thus, while this analysis finds overwhelming evidence that environmental nutrient supply alters the composition of these microbiomes, this effect appears to be driven by only a few constituent microbes. Moreover, it appears accounting for correlation among OTUs is essential to detecting these individual microbes. In the Online Supplement, we show that estimates of OTU-level treatment effects are robust to the choice of variance hyperparameters c_0 and d_0.



(a) All OTUs

(b) Only affected OTUs

Figure 4: Posterior 95% credible intervals for the effect of nutrient supplement on each OTU. Across all OTUs (a), most are not significantly affected by the treatment (gray lines), but MIMIX identifies 73 of 2,662 OTUs (2.7%) that show a significant response (black lines). Among these affected OTUs (b), the taxonomy of each fungal OTU is given up to species, if known, or at a higher taxonomic rank, such as genus or order, with trailing numbers identifying distinct strains. The OTUs are ordered along the y-axis according to complete linkage hierarchical clustering of the estimated factor correlation matrix from MIMIX.

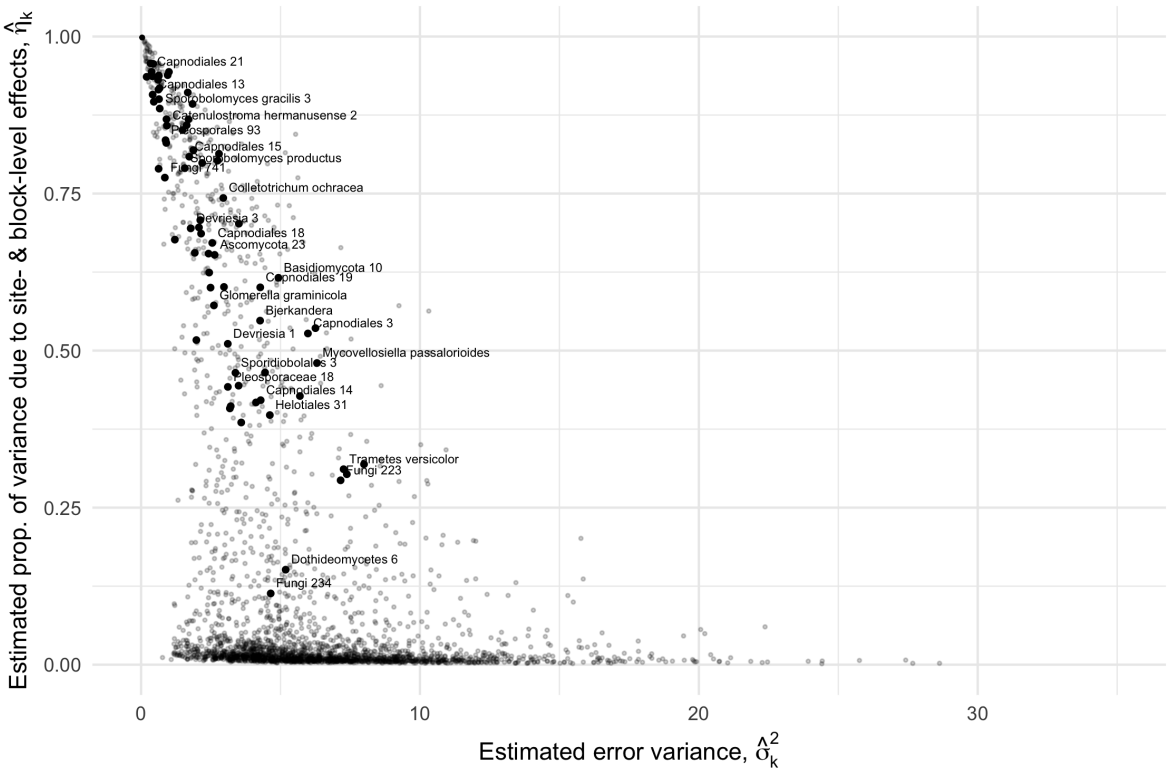


Figure 5: The proportion of variance for each OTU that is explained by the contribution of site and block vs. unexplained residual variation. Black dots correspond to OTUs identified by MIMIX as being significantly affected by nutrient supplement in Figure 4b, though only a subset of names are displayed to avoid overlapping labels.

MIMIX quantifies the relative contribution of different sources of residual variation to fungal composition. Site- and block-level variances are estimated with posterior means $\hat{\sigma}^2_{\text{Site}} = 3.279$ and $\hat{\sigma}^2_{\text{Block}} = 0.296$. Posterior means of OTU-specific variances not attributed to the study design $(\hat{\sigma}^2_1, \ldots, \hat{\sigma}^2_K)$ are strongly skewed, ranging from 0.032 to 28.63 with mean 5.817 (Figure 5). The proportion of residual variance in OTU k that is captured by site- and block-level effects is estimated by

$$\hat{\eta}_k = 1 - \frac{\hat{\sigma}^2_k}{\hat{\sigma}^2_k + (1 + \hat{\sigma}^2_{\text{Site}} + \hat{\sigma}^2_{\text{Block}}) \sum_{l=1}^{L} \hat{\lambda}^2_{lk}},$$

where $\hat{\lambda}_{lk}$ is the posterior mean of λ_{lk}. Figure 5 shows $\hat{\eta}_k$ vs. $\hat{\sigma}^2_k$ for each OTU.

Two loose groups of points emerge, one in which residual variation is almost entirely explained by site and block variation ($\hat{\eta}_k > 0.4$), and another in which these random effects explain a relatively small amount of residual variation in OTUs ($\hat{\eta}_k < 0.4$). OTUs identified by MIMIX as being

significantly affected by nutrient supplement, indicated by black dots in Figure 5, appear to be strongly represented in the former of these two groups, although there are a few OTUs that are significantly affected by nutrient supplement that do not appear to be greatly influenced by site and block effects. Overall, site and block effects explain over half the residual variation ($\hat{\eta}_k > 0.5$) for

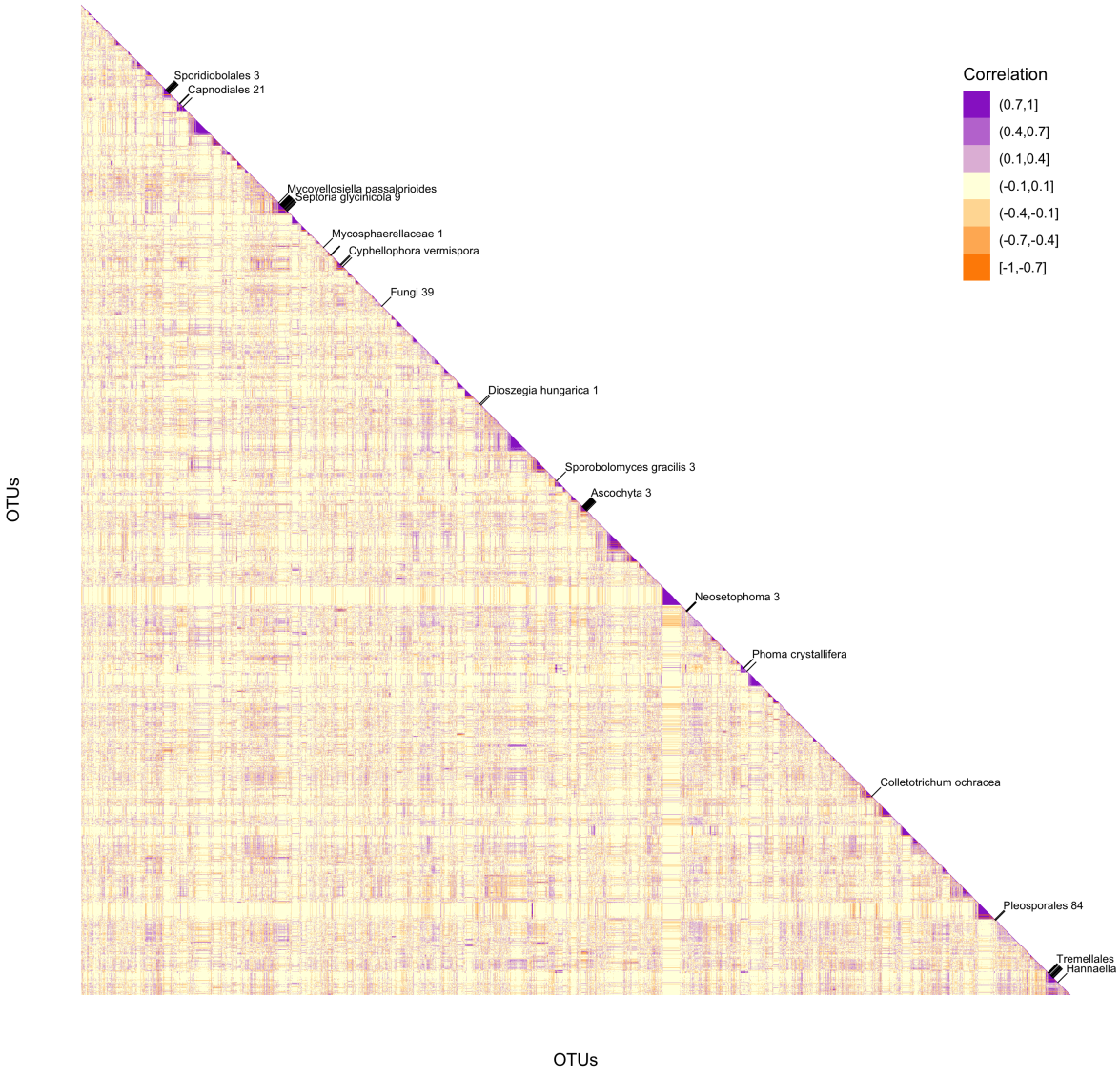


Figure 6: Estimated factor correlation matrix among OTUs, with OTUs ordered by hierarchical clustering. Clusters of strongly correlated OTUs, represented by purple triangles along the diagonal, indicate small communities that respond similarly to the fixed and random effects of the designed experiment. The OTUs highlighted along the diagonal are those identified by MIMIX as being significantly affected by nutrient supplement in Figure 4b, though only a subset of names are displayed to avoid overlapping labels.

approximately 16% of OTUs.

Finally, we take a closer look at the estimated factor correlation matrix $\mathbf{\Lambda}\mathbf{\Lambda}'$ using the posterior mean of $\mathbf{\Lambda}$ (Figure 6). MIMIX identifies a large number of potential clusters of fungal OTUs (grouped along the diagonal of Figure 6), with myriad positive and negative correlations among them. These clusters may indicate latent ecological subcommunities, or they may reflect collections of taxa that occupy similar ecological niches. MIMIX w/o Factors does not adequately account for these relationships among OTUs, which explains its poorer fit to the data.

6 Discussion

In this paper, we introduce MIMIX (MIcrobiome MIXed effects), a Bayesian mixed-effects model to analyze microbiome data as a response variable in designed experiments. MIMIX has several attractive features for the analyis of high-dimensional, sparse, microbiome count data. It performs spike-and-slab variable selection to identify treatment effects on individual Operational Taxonomic Units (OTUs). Moreover, its Bayesian factor analysis formulation with a continuous shrinkage Dirichlet-Laplace prior clusters OTUs into different factors based on how they respond to the fixed and random effects in the experiment. This allows for post-hoc analysis of the model to identify behaviorally similar clusters of OTUs within the larger microbiome community. In a simulation study, these features allow MIMIX to outperform both PERMANOVA with Bray-Curtis dissimilarity and MIMIX that does not include Bayesian factors (MIMIX w/o Factors) in identifying and estimating sparse treatment effects.

We demonstrate MIMIX on experimental data from four sites within the Nutrient Network cooperative to quantify the effects of nutrient supplement and herbivore exclusion on the fungal microbiome of the grass species *Andropogon gerardii*. We identify a significant effect of nutrient supplement (but not herbivore exclusion) on these microbiomes, while accounting for random effects due to both site and blocks within site. We also identify a significant treatment effect of nutrient supplement on about 2.7% of OTUs. Although the OTU assignments in this particular data set

are preliminary, our results illustrate how MIMIX enables OTU-level inferences that may allow for deeper and sharper understanding of how environmental conditions impact the abundance of specific taxa in a microbiome.

Ecologically, this analysis of the Nutrient Network data suggests the following insights. First, ecologists are frequently interested in how resource supply and grazing combine to influence the structure of ecological communities (the so-called "bottom-up" vs. "top-down" dichotomy). The results of this experiment suggest that resource supply, or "bottom-up" factors, play a larger role in structuring a host's microbiome than predation. Second, the paucity of large OTU-specific responses (Figure 4) suggests that only a handful of microbial taxa respond to the nutrient supplementation, and that these responses can be sufficient to reshape the microbiome when considered as an ecological whole. Third, the residual variation of some, though not all OTUs can be explained by site- and block-level random effects (Figure 5), suggesting that these OTUs may either be strongly influenced by regional environmental correlates, or may be limited by reduced dispersal at regional (km) scales. Finally, the estimated factor correlation matrix (Figure 6) suggests that this foliar microbiome is composed of many modestly sized clusters of similarly behaving OTUs. This pattern may either suggest many moderately sized subcommunities, aggregation of taxa into many separate ecological niches, or both.

The initial results from MIMIX are encouraging, but its features will need to scale as microbiome experiments grow in complexity. For example, MIMIX is not currently suited for handling data from longitudinal studies with repeated measures over time. Furthermore, while the dimensionality of the microbiome data analyzed here is quite high at $K \approx 2,500$, the dimensionality can grow rapidly, especially when multiple domains of life (bacteria, archaea, fungi, etc.) are studied. In such instances, computation time and memory management will become a more pressing concern which may require a reconstruction of the posterior sampling scheme. We set the number of latent factors to be the maximum number of identifiable factors and allow the Bayesian shrinkage prior to eliminate excess factors. Another approach that may be more suitable for massive datasets is to use a smaller number of latent factors with Gaussian priors on the factor matrix. This approach is

easier to explain and implement, and faster for a given number of factors. However, this approach would likely require expensive cross-validation to pick the number of factors and would not account for uncertainty in the number of factors.

Funding

This work was supported by National Science Foundation award EF-1241794.

References

Aitchison, John (1986). *The statistical analysis of compositional data*. Chapman & Hall, London.

Alivisatos, A Paul, MJ Blaser, Eoin L Brodie, Miyoung Chun, Jeffrey L Dangl, Timothy J Donohue, Pieter C Dorrestein, Jack A Gilbert, Jessica L Green, Janet K Jansson, et al. (2015). "A unified initiative to harness Earth's microbiomes". *Science* 350.6260, pages 507–508.

Bezanson, Jeff, Alan Edelman, Stefan Karpinski, and Viral B. Shah (2017). "Julia: A Fresh Approach to Numerical Computing". *SIAM Review* 59.1, pages 65–98.

Bhattacharya, Anirban, Debdeep Pati, Natesh S Pillai, and David B Dunson (2015). "Dirichlet–Laplace priors for optimal shrinkage". *Journal of the American Statistical Association* 110.512, pages 1479–1490.

Billheimer, Dean, Peter Guttorp, and William F Fagan (2001). "Statistical interpretation of species composition". *Journal of the American statistical Association* 96.456, pages 1205–1214.

Bond-Lamberty, Ben, Harvey Bolton, Sarah Fansler, Alejandro Heredia-Langner, Chongxuan Liu, Lee Ann McCue, Jeffrey Smith, and Vanessa Bailey (2016). "Soil respiration and bacterial structure and function after 17 years of a reciprocal soil transplant experiment". *PloS one* 11.3, e0150599.

Borer, Elizabeth T, W Stanley Harpole, Peter B Adler, Eric M Lind, John L Orrock, Eric W Seabloom, and Melinda D Smith (2014). "Finding generality in ecology: a model for globally distributed experiments". *Methods in Ecology and Evolution* 5.1, pages 65–73.

Bray, J Roger and John T Curtis (1957). "An ordination of the upland forest communities of southern Wisconsin". *Ecological monographs* 27.4, pages 325–349.

Carvalho, Carlos M, Jeffrey Chang, Joseph E Lucas, Joseph R Nevins, Quanli Wang, and Mike West (2008). "High-dimensional sparse factor modeling: Applications in gene expression genomics". *Journal of the American Statistical Association* 103.484, pages 1438–1456.

Chen, Eric and Hongzhe Li (2016). "A two-part mixed-effects model for analyzing longitudinal microbiome compositional data". *Bioinformatics*, btw308.

Chen, Jun and Hongzhe Li (2013). "Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis". *The annals of applied statistics* 7.1.

Gelman, Andrew, John B Carlin, Hal S Stern, and Donald B Rubin (2014). *Bayesian data analysis*. Volume 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Grantham, Neal S, Brian J Reich, Krishna Pacifici, Eric B Laber, Holly L Menninger, Jessica B Henley, Albert Barberán, Jonathan W Leff, Noah Fierer, and Robert R Dunn (2015). "Fungi identify the geographic origin of dust samples". *PloS one* 10.4, e0122605.

La Rosa, Patricio S, J Paul Brooks, Elena Deych, Edward L Boone, David J Edwards, Qin Wang, Erica Sodergren, George Weinstock, and William D Shannon (2012). "Hypothesis testing and power calculations for taxonomic-based human microbiome data". *PloS one* 7.12, e52078.

Lozupone, Catherine and Rob Knight (2005). "UniFrac: a new phylogenetic method for comparing microbial communities". *Applied and environmental microbiology* 71.12, pages 8228–8235.

McArdle, Brian H and Marti J Anderson (2001). "Fitting multivariate models to community data: a comment on distance-based redundancy analysis". *Ecology* 82.1, pages 290–297.

McMurdie, Paul and Susan Holmes (2014). "Waste not, want not: why rarefying microbiome data is inadmissible". *PLoS Comput Biol* 10.4, e1003531.

Metzker, Michael L (2010). "Sequencing technologies —the next generation". *Nature reviews genetics* 11.1, pages 31–46.

Mitchell, Toby J and John J Beauchamp (1988). "Bayesian variable selection in linear regression". *Journal of the American Statistical Association* 83.404, pages 1023–1032.

Neal, Radford M et al. (2011). "MCMC using Hamiltonian dynamics". *Handbook of Markov Chain Monte Carlo* 2, pages 113–162.

Pearson, Karl (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* 60.359-367, pages 489–498.

Ren, Boyu, Sergio Bacallado, Stefano Favaro, Susan Holmes, and Lorenzo Trippa (2017). "Bayesian nonparametric ordination for the analysis of microbial communities". *Journal of the American Statistical Association* 112.520, pages 1430–1442.

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

Shafiei, Mahdi, Katherine A Dunn, Eva Boon, Shelley M MacDonald, David A Walsh, Hong Gu, and Joseph P Bielawski (2015). "BioMiCo: a supervised Bayesian model for inference of microbial community structure". *Microbiome* 3.1, page 8.

Wadsworth, W Duncan, Raffaele Argiento, Michele Guindani, Jessica Galloway-Pena, Samuel A Shelburne, and Marina Vannucci (2017). "An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data". *BMC Bioinformatics* 18.1, page 94.

Warton, David I, Stephen T Wright, and Yi Wang (2012). "Distance-based multivariate analyses confound location and dispersion effects". *Methods in Ecology and Evolution* 3.1, pages 89–101.

Wu, Gary D and James D Lewis (2013). "Analysis of the human gut microbiome and association with disease". *Clinical gastroenterology and hepatology: the official clinical practice journal of the American Gastroenterological Association* 11.7.

Xia, Fan, Jun Chen, Wing Kam Fung, and Hongzhe Li (2013). "A logistic normal multinomial regression model for microbiome compositional data analysis". *Biometrics* 69.4, pages 1053–1063.

Zhang, Xinyan, Himel Mallick, Zaixiang Tang, Lei Zhang, Xiangqin Cui, Andrew K Benson, and Nengjun Yi (2017). "Negative binomial mixed models for analyzing microbiome count data". *BMC Bioinformatics* 18.1, page 4.