Latent variable modeling for the microbiome

KRIS SANKARAN*, SUSAN P. HOLMES

Department of Statistics, Stanford University, 390 Serra Mall, Stanford, CA 94305, USA

krissankaran@stanford.edu

Summary

The human microbiome is a complex ecological system, and describing its structure and function under different environmental conditions is important from both basic scientific and medical perspectives. Viewed through a biostatistical lens, many microbiome analysis goals can be formulated as latent variable modeling problems. However, although probabilistic latent variable models are a cornerstone of modern unsupervised learning, they are rarely applied in the context of microbiome data analysis, in spite of the evolutionary, temporal, and count structure that could be directly incorporated through such models. We explore the application of probabilistic latent variable models to microbiome data, with a focus on Latent Dirichlet allocation, Non-negative matrix factorization, and Dynamic Unigram models. To develop guidelines for when different methods are appropriate, we perform a simulation study. We further illustrate and compare these techniques using the data of Dethlefsen and Relman (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), a study on the effects of antibiotics on bacterial community composition. Code and data for all simulations and case studies are available publicly.

Keywords: Microbiome; Microbial ecology; Latent Dirichlet allocation; Non-negative matrix factorization; Posterior predictive checks; Bayesian data analysis.

1. Introduction

Microbiome studies attempt to characterize variation in bacterial abundance profiles across different experimental conditions (Gilbert *and others*, 2014). For example, a study may attempt to describe differences in bacterial communities between diseased and healthy states or after deliberately induced perturbations (Dethlefsen and Relman, 2011; Fukuyama *and others*, 2017).

In the process, two complementary difficulties arise. Firstly, the data are often high dimensional, measured over several hundreds or thousands types of bacteria. Studying patterns at the level of particular bacteria is typically uninformative. Secondly, it can be important to study bacterial abundances in the context of existing biological knowledge. For example, it is scientifically meaningful when a collection of bacteria that are known to be evolutionarily related have either similar or opposed abundance profiles. These challenges motivate methodological work in the microbiome, including phylogenetically-informed techniques for dimensionality-reduction and high-dimensional regression.

Towards phylogenetically-informed dimensionality reduction, Chen *and others* (2013) applied a structured regularization penalty in sparse canonical correlation analysis to incorporate phylogenetic

---

*To whom correspondence should be addressed.

information, encouraging scores for similar species to be placed close to one another. A similar effect was obtained by placing an appropriate prior in a Bayesian model, as explored in Fukuyama (2017). By varying the strength of the prior, it is possible to encourage different degrees of smoothness with respect to phylogeny, and Empirical Bayes estimation allows for a certain type of adaptivity.

In addition to dimensionality reduction, high-dimensional classification is popular in the microbiome literature. Often interesting species can be identified by searching for important features in models that predict sample characteristics—treatment vs. control (e.g. from species abundances). Segata *and others* (2011) provided an approach for accounting for phylogenetic structure in this problem by initially prescreening according to independent biological interest. Alternatively, Chen and Li (2013) applied a Dirichlet multinomial regression model to study the relationship between bacterial counts and sample characteristics in a fully generative fashion, and then added an ℓ^1-penalty to induce sparsity and facilitate interpretability. The development of structured modeling techniques tailored to the microbiome remains an active area of current research.

2. Methods

We review a few of the statistical modeling techniques that are the focus of this work. Many of these techniques have been borrowed from text analysis, thinking of the bacterial counts matrix as a biological analog of the document-term matrix. The idea of transferring these techniques to the microbiome is not new, though its appropriateness and usefulness has only been explored in relatively limited settings (Schloss and Handelsman, 2007; Chen *and others*, 2012; Holmes *and others*, 2012; Chen and Li, 2013; Shafiei *and others*, 2015; Jiang *and others*, 2017).

2.1. *Dirichlet multinomial mixture model*

While the multinomial distribution is the fundamental probability mechanism for sampling counts, multinomial models are only appropriate for relatively homogeneous datasets, where categories are nearly independent. An extension, Dirichlet multinomial mixture modeling, allows for count modeling in the presence of increased heterogeneity, and its relevance to the text and microbiome settings were key contributions of Nigam *and others* (2000) and Holmes *and others* (2012), respectively. In this section, we adopt text analysis terminology, where the count matrices of interest are counts of terms across documents. In Section 2.5, we clarify the connection to microbiome analysis.

Suppose there are D documents across V terms, and that these documents are assumed to belong to one of K underlying topics, where a topic is defined as a distribution over words. The Dirichlet multinomial mixture model draws each topic from a distribution over probabilities. Then, for each document, a topic is chosen by flipping a K-sided coin with probability θ_k of coming up on side k. Conditional on the selected topic, all words are drawn independently from the probabilities associated with the selected topic.

More formally, represent the topic for the dth document by $z_d \in \{1, \ldots, K\}$ and the term in the n_{th} word of this document by w_{dn}. Suppose the kth topic places weight β_{vk} on the vth term, so that $\beta_k \in \mathcal{S}^{V-1}$. Suppose there are N_d words in the dth document. Then, the generative mechanism for the observed dataset is

$$w_{dn} | (\beta_k)_{k=1}^K, z_d \stackrel{iid}{\sim} \text{Mult}\left(1, \beta_{z_d}\right) \text{ for } d = 1, \ldots, D \text{ and } n = 1, \ldots, N_d$$

$$z_d | \theta \stackrel{iid}{\sim} \text{Mult}\left(1, \theta\right) \text{ for } d = 1, \ldots, D$$

$$\beta_k \stackrel{iid}{\sim} \text{Dir}\left(\gamma\right) \text{ for } k = 1, \ldots, K.$$

Equivalently, we could write $w_d | (\beta_k)_{k=1}^K, z_d \sim \text{Mult}(N_d, \beta_{z_d})$, though the form above makes comparison with Latent Dirichlet allocation (LDA) more straightforward. Further, while we treat θ, and γ as fixed parameters, it is possible to place priors on them as well. Geometrically, this model identifies each document with one of the β_k on the V-dimensional simplex.

In practice, interpretation revolves around the posterior topic memberships, $\mathbb{P}(z_d = k | (w_{dn}))$, and probabilities, $\mathbb{P}(\beta_{kv} | (w_{dn}))$. While these estimates can be useful in guiding scientific analysis (Nigam *and others*, 2000; Holmes *and others*, 2012), the assumption that each document belongs completely to one topic is sometimes unnecessarily restrictive. For example, in learning a Dirichlet multinomial mixture model on a collection of newspaper articles, we may recover separate topics related to science and personal health, but there would be no way to express the mixture of topics in an article about health research.

2.2. *Latent Dirichlet allocation*

LDA is a generalization of Dirichlet multinomial mixture modeling where documents are allowed to have fractional membership across a set of topics (Blei *and others*, 2003). This addresses the key limitation of Dirichlet multinomial mixture modeling, and one goal of the case study in Section 4 is to demonstrate the usefulness of this additional flexibility in microbiome analysis.

We use the same notation as before, but instead of fixing a global θ parameter, let $\theta_d \in \mathcal{S}^{K-1}$ represent the dth document's mixture over the K underlying topics. Represent the term in the n_{th} word of this document by w_{dn}, and the associated topic by z_{dn}. Suppose the kth topic places weight β_{vk} on the vth term, so that $\beta_k \in \mathcal{S}^{V-1}$. Suppose there are N_d words in the dth document. Then, the generative mechanism for the observed dataset is

$$w_{dn} | (\beta_k)_{k=1}^K, z_{dn} \overset{iid}{\sim} \text{Mult}(1, \beta_{z_{dn}}) \text{ for } d = 1, \ldots, D \text{ and } n = 1, \ldots, N_d$$

$$z_{dn} | \theta_d \overset{iid}{\sim} \text{Mult}(1, \theta_d) \text{ for } d = 1, \ldots, D \text{ and } n = 1, \ldots, N_d$$

$$\theta_d \overset{iid}{\sim} \text{Dir}(\alpha) \text{ for } d = 1, \ldots, D$$

$$\beta_k \overset{iid}{\sim} \text{Dir}(\gamma) \text{ for } k = 1, \ldots, K.$$

In microbiome applications, we will find a formulation that marginalizes over the z_{dn} more convenient. Setting $x_{dv} = \sum_{n=1}^{N_d} \mathbb{I}(w_{dn} = v)$, we can write the marginal distribution as

$$x_{d\cdot} | (\beta_k)_1^K \overset{iid}{\sim} \text{Mult}(N_d, B\theta_d) \text{ for } d = 1, \ldots, D$$

$$\theta_d \overset{iid}{\sim} \text{Dir}(\alpha) \text{ for } d = 1, \ldots, D$$

$$\beta_k \overset{iid}{\sim} \text{Dir}(\gamma) \text{ for } k = 1, \ldots, K,$$

where we have introduced the $V \times K$ matrix concatenating all topics column-wise, $B = (\beta_1, \ldots, \beta_K)$. Geometrically, LDA identifies samples with points in the convex hull of K topics (β_k) on the V-dimensional simplex, rather than the individual corners, as in the Dirichlet multinomial mixture model.

2.3. *Dynamic Unigram model*

Upon examining this geometric interpretation, we might consider in some situations a model that identifies samples with a continuous curve on this V-dimensional simplex. This reasoning leads naturally to the Dynamic Unigram model (Blei and Lafferty, 2006). The underlying curve reflects the gradual evolution

of probabilities over time, and is implemented by passing a random walk through a multilogit link. That is, the Dynamic Unigram model posits the generative model

$$x_{d\cdot}|\mu_{t(d)} \overset{iid}{\sim} \text{Mult}\left(N_d, S\left(\mu_{t(d)}\right)\right) \text{ for } d = 1, \ldots, D$$

$$\mu_t|\mu_{t-1} \overset{iid}{\sim} \mathcal{N}\left(\mu_{t-1}, \sigma^2 I_V\right) \text{ for } t = 1, \ldots, T$$

$$\mu_0 \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2 I_V\right),$$

where S is the multilogit link

$$[S\left(\mu\right)]_v = \frac{\exp\left(\mu_v\right)}{\sum_{v'} \exp\left(\mu_{v'}\right)},$$

and $t\left(d\right)$ maps document d to the time it was sampled. The $\left(\mu_{t()}\right)$ define a Gaussian random walk in $\mathbb{R}^V$ with step-size σ, and S transforms the walk into a sequence of probability distributions. In our experiments, we place a vague inverse-gamma prior on σ^2, since this hyperparameter is rarely known in practice.

2.4. *Non-negative matrix factorization*

In LDA, count matrices are modeled by sampling from multinomials with total counts coming from the word count of each document and probabilities coming from the rows of ΘB^T where $\Theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_D \end{pmatrix}$ and $B = \left(\beta_1 \ldots \beta_K\right)$ are $D \times K$ and $V \times K$ matrices representing document and topic distributions, respectively, and where each $\theta_d \in \mathcal{S}^{K-1}$ and $\beta_k \in \mathcal{S}^{V-1}$.

Alternatively, it is possible to model the non-negative matrix X by the product of low rank matrices, $X \approx \Theta B^T$, where now the only constraints on Θ and B are that $\Theta \in \mathbb{R}_+^{D \times K}$ and $B \in \mathbb{R}_+^{V \times K}$. This is the point of departure for a variety of algorithms in the non-negative matrix factorization (NMF) literature (Wang and Zhang, 2013).

We focus on the Gamma-Poisson (GaP) factorization model (Canny, 2004) which posits the hierarchical model

$$X \sim \text{Poi}\left(\Theta B^T\right)$$

$$\Theta \sim \text{Gam}\left(a_0 1_{D \times K}, b_0 1_{D \times K}\right)$$

$$B \sim \text{Gam}\left(c_0 1_{V \times K}, d_0 1_{V \times K}\right),$$

where our notation means that each entry in these matrices is sampled independently, with parameters given by the corresponding entry in the parameter matrix. As a consequence of the representation of the negative binomial distribution as a Gamma mixture of Poissons, this is a natural model of overdispersed counts, which arise frequently in genomic and microbiome settings. In practice, the hyperparameters a_0, b_0, c_0, and d_0 are unknown. In all of our experiments, we optimize over them, though it would also be possible to add a level to the hierarchy and place vague nonnegative priors on these parameters.

In our simulation studies, we also consider a slight variant of this model, similar to the proposal of (Romero *and others*, 2014), which independently sends entries of X to zero with probability p_0. In our experiments, we assume that this zero-inflation probability is known. This procedure is denoted by Z-NMF. Throughout, we use the probabilistic programming language Stan (Carpenter *and others*, 2016).

2.5. *Microbiome vs. text analysis*

One of the primary contributions of our work is to develop the observation that methods popular in text analysis can be adapted to the microbiome setting in a way that produces useful summaries. Before applying these methods, we develop the analogy between these text and microbiome analysis and also draw attention to points where the parallels break down.

In the abstract, it becomes clear that the semantic differences between the units of statistical analysis are often superficial. For example, we can map between the most common field-specific terms as follows:

- **Document $\iff$ Biological Sample**: The basic sampling units, over which conclusions are generalized, are documents (text analysis) and biological samples (microbiome analysis). It is of interest to highlight similarities and differences across these units, often through some variation on clustering or dimensionality reduction.

- **Term $\iff$ Bacterial species**: The fundamental features with which to describe samples are the counts of terms (text analysis) and bacterial species (microbiome analysis). More formally, by bacterial species, we mean Amplicon Sequence Variants (Callahan *and others*, 2017), though we avoid this terminology for simplicity of exposition.

- **Topic $\iff$ Community**: For interpretation, it is common to imagine "prototypical" units which can be used as a point of reference for observed samples. In text analysis, these are called topics—for example, "business" or "politics" articles have their own specific vocabularies. On the other hand, in microbiome analysis, these are called "communities" – different communities have different bacterial signatures.

- **Word $\iff$ Sequencing Read**: A "word" in text analysis refers to a single instance of a term in a document, not its total count. The analog in microbiome analysis is an individual read that has been mapped to a unique sequence variant, though this is rarely an object of intrinsic interest.

- **Corpus $\iff$ Environment**: Sometimes a grouping structure is known apriori among sampling units. In this case, it can be informative to describe whether topics are or are not shared across these groups (Teh *and others*, 2005). In the text analysis literature, a known group of documents (e.g. all articles coming from one newspaper) is called a corpus. In the microbiome literature, the associated concept is the environment (e.g. skin, ocean, or soil) from which a sample was obtained.

Now that we have established the semantic connections between text and microbiome analyses, we compare the types of data and analysis goals that are typical within the respective fields. In both, a central element of study is the sample-by-feature (either document-by-term or sample-by-bacteria) matrix. Besides count structure, a striking similarity between these data matrices is sparsity: most entries are zero. Further, observed counts tend to be highly-skewed—some terms are far more common than others, and in the same way, some microbes are much more abundant than others. Finally, in both fields, contextual information beyond the raw sample-by-feature matrix is typically available. For example, timestamps are often available in both domains, *n*-grams have an analog in terms of small subnetworks of co-occurring bacteria, and phylogenetic similarity between species parallels known linguistic characteristics of terms.

Nonetheless, in practice, the structure of these data can be quite different. First, text data can be on a much larger scale. For example, the Wikipedia corpus studied by Hoffman *and others* (2013) includes 3.8 million articles. In contrast, even large microbiome datasets, like the one studied in (Gilbert *and others*, 2014), typically only have on the order of tens of thousands of samples. Similarly, the total number of

terms in such large-scale text analysis problems can be substantially larger than the number of bacterial species under consideration.

On the other hand, in a different sense, microbiome studies are larger in scale, there tend to be tens of thousands of reads per sample in microbiome studies, but only hundreds to thousands of words within any article. This means that techniques that rely on the representation of documents as sequences of individual words, rather than vectors of word counts, require too much memory to be practical. This makes many useful text analysis techniques—for example, some methods for model inference and evaluation (Wallach and others, 2009)—out of reach for standard microbiome problems. This does, however, suggest potential opportunities for future research.

Lastly, we compare the prevailing analytical goals within text and microbiome analyses. In both fields, data reduction can be informative for developing models of system behavior. However, an essential difference is that even unsupervised text analysis techniques are often embedded within automatic systems, for text classification or information retrieval for example, which do not require the intervention of a scientific investigator. In contrast, in microbiome studies, researchers often have control over specific experimental design structure, and collect and analyze data on a per-study basis. In this setting, success is defined somewhat amorphously as an ability to describe the structure and function underlying a biological system of interest.

3. SIMULATION STUDY

It can be liberating to have easy access to such a variety of modeling strategies for any given microbiome analysis problem. However, with this increased flexibility comes the difficulty of determining when to use which methods. To build some intuition about estimation accuracy across combinations of data settings and model types, we conduct a series of simulation studies. These are meant to complement the model-checking that should follow parametric analysis—since we know the truth in simulations, it is easier to develop more unambiguous guidelines.

More specifically, our plan is to divide our experiments into simulations generating data from the true LDA, unigram, and NMF/Z-NMF models. In each, we vary the sample size and dimension, performing model estimation using either Markov Chain Monte Carlo (MCMC) sampling, variational Bayes (VB), or a bootstrap. The only misspecification, we consider is a failure to account for zero inflation when the true data were generated according to the Z-NMF model—though not pursued here, it could be interesting to study robustness of study conclusions to misspecification in the number of topics or deliberate contamination. Due to space constraints, we relegate a discussion of the unigram simulation experiment to "Dynamic Unigram Simulation Experiment" section of supplementary material available at *Biostatistics* online.

For the LDA experiment, we vary the number of samples $D \in \{20, 100\}$, the number of features $V \in \{325, 650\}$, and the total count per sample $N \in \{1625, 3250, 6500\}$, in order to approximately match dimensions typical in real microbiome datasets. On the other hand, we fix the number of topics to $K = 2$ and the Dirichlet hyperparameter to $\alpha = \gamma = 1$. For each simulated dataset, we perform estimation using MCMC sampling, VB, and a parametric bootstrap. In more detail, this bootstrap procedure fits VB to the original data, simulates $B = 500$ new datasets X_b^* according to the LDA model using VB-estimated parameters $\{\hat{\theta}_b^*, \hat{\beta}_b^*\}$, and re-estimates parameters $\{\hat{\theta}_b^{**}, \hat{\beta}_b^{**}\}$ on each simulated data set, again using VB. The motivation for this procedure is the desire to strike an easily-parallelizable compromise between MCMC sampling, which can be time consuming but has reliable uncertainty estimates, and VB, which is fast, but can underestimate uncertainty (Wang and Titterington, 2005).

Note that, due to the label-switching problem, it is not possible to directly compare the estimated topics across simulation configurations. To address this issue, we attempt to identify the permutation that aligns topics across all experiments. For each experiment, we identify the true-estimated topics pair with highest correlation, then find the next highest pair among the remaining, and so forth. Judging from the aligned
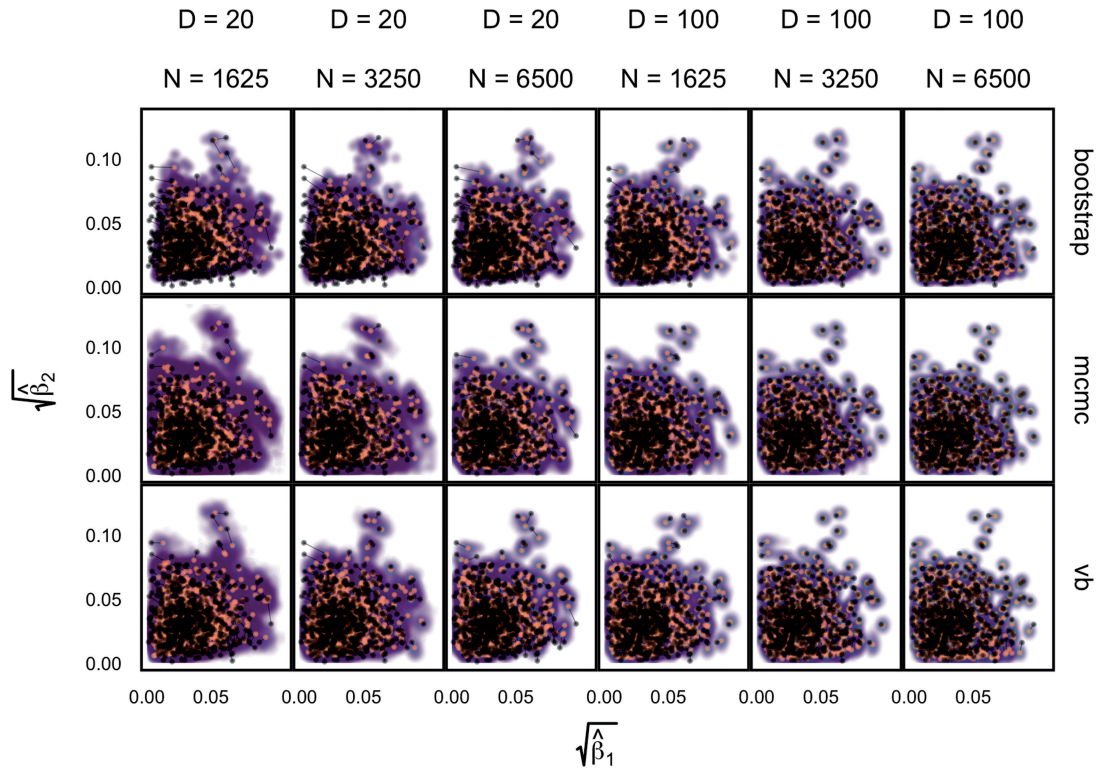
Fig. 1. Different inference algorithms for LDA provide comparable posterior estimates under the simulation config-urations we have considered. Within each panel, we display the true pairs $\left(\sqrt{\beta_{v1}}, \sqrt{\beta_{v2}}\right)$ as black points v. The purple clouds are posterior samples from the inference procedure, which are given as row labels. The posterior medians are orange points that are linked to the true v. Different columns index different D (top column label) vs. N (bottom column label) pairs. Here, we have subsetted to $V = 650$ – the corresponding figure for $V = 325$ is given by Figure 6 of supplementary material available at *Biostatistics* online. Refer online version for color figure.

densities in Figure 1 and Figures 6–8 of supplementary material available at *Biostatistics* online this *ad hoc* alignment procedure seems sufficient.

Figure 1 displays the true and posterior β_k for the LDA experiment with $V = 650$ features, while varying other simulation characteristics. Each panel represents a single experimental configuration, with each axis associated with an underlying topic. Each black point is the square-root transformed value of feature v across topics: $\left(\sqrt{\beta_{v1}}, \sqrt{\beta_{v2}}\right)$. The shaded clouds are sampled posteriors, and the linked orange point give the posterior median for the associated feature. Across rows, different inferential procedures are compared. The top row of column labels refers to the total count N within each sample, while the second refers to the number of samples D. The analogous figure when $V = 325$ is provided in Figure 6 of supplementary material available at *Biostatistics* online.

As expected, when D increases, the posterior for β, whose dimension does not increase with D, begins to concentrate around its true value. As the number of samples or total count within samples increases, the posteriors further concentrate around the truth. In each of the settings considered, the three methods seem comparable, suggesting that for microbiome studies of this approximate scale, VB may be a reasonable choice, considering that it can be run much more quickly than either MCMC or the bootstrap.

While the Kernel-smoothed posterior densities display the complete results from this simulation study, a few summary statistics from these densities can facilitate comparison across models and data generation
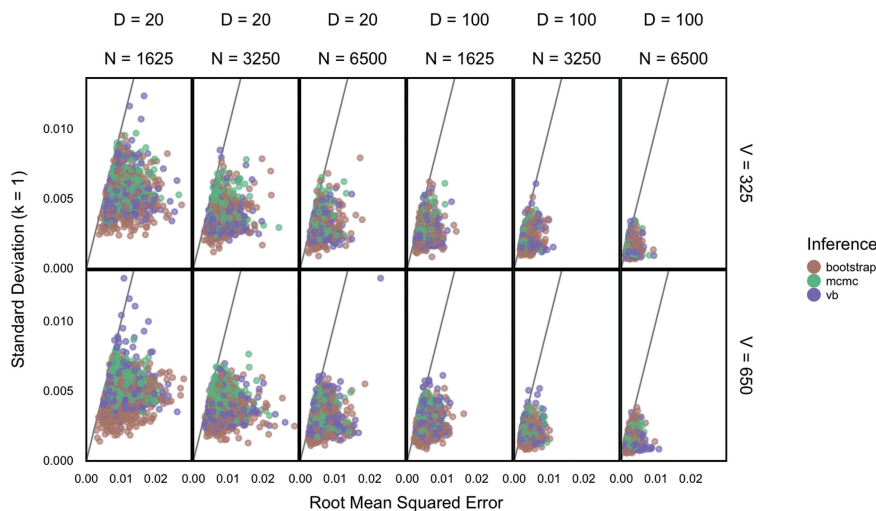
Fig. 2. A summary of the errors and uncertainty across models and regimes for the β_{vk} in Figure 1 and Figure 6 of supplementary material available at *Biostatistics* online. On the *x*-axis, we plot the difference between posterior median and the true value, after having square-root transformed. On the *y*-axis, we provide the standard deviation of the posterior samples for each v, along the first dimension $K = 1$. Columns are indexed as in Figure 1, but now rows provide different values of V. Refer online version for color figure.

regimes. The key features of interest are (i) the distance of the posterior medians for each parameter from their true values, after alignment and (ii) the concentration of these posteriors around their medians. Figure 2 addresses these questions directly—the *x*-axis gives the RMSE of the square-root transformed posterior medians for the β_{vk}, for each v, across k, and the *y*-axis gives the associated standard deviation, along $K = 1$, for each v. As in Figure 1, the first row of column numbers gives the total count N in each sample, and the second gives D, the number of samples. The grey line within each panel is the identity line, where the error equals one standard deviation.

The drift of points to the bottom left as N and D increase reflects the improved accuracy and concentration of all inference techniques when the number of samples and total count within samples increase. Note that few points lie above the identity standard deviation line—this suggests that the posteriors may be misleadingly narrow. Finally, we find that the scale of errors and concentration in the $V = 325$ and $V = 650$ cases are comparable. Considering that there is the same number of documents available for estimating each individual β_{vk}, this is not surprising.

For the NMF experiment, we use similar parameters, except now we introduce a zero-inflation probability, $p_0 \in \{0, 0.2\}$, and have control only over the expected total count per sample $\mathbb{E}[N]$, rather than N. When $p_0 > 0$, we perform inference when the true p_0 is provided and also when p_0 is (incorrectly) assumed to be zero. To modulate $\mathbb{E}[N]$, we vary a_0 and b_0 so that $\mathbb{E}[N] = \frac{b_0}{a_0}KV \in \{1625, 3250, 6500\}$, matching the LDA experiment. As before, we vary the number of samples $D \in \{20, 100\}$ but fix the number of latent factors K at 2. Figure 9 of supplementary material available at *Biostatistics* online summarizes the estimation error across regimes, and is read in the same way as Figure 2, with two exceptions:

1. Instead of providing the fixed values of N, we display $\mathbb{E}[N]$.
2. We must distinguish between different p_0 regimes, and whether or not the inferential procedure assumed p_0 was 0 or was given the true p_0. The true value of p_0 is marked by the second row of

column labels, and inference where p_0 was provided is denoted by the row label Z-GaP, while those rows marked GaP assumed $p_0 = 0$.

The associated posterior density display is given in Figures 7 and 8 of supplementary material available at *Biostatistics* online.

The results in Figure 9 of supplementary material available at *Biostatistics* online are more problematic than those in Figure 2. First, the standard errors are no longer comparable across methods, with those for the MCMC sampling procedure appearing much larger across many simulation configurations. Second, the error rates are substantially larger than those in LDA. Examining Figures 7 and 8 of supplementary material available at *Biostatistics* online, it appears that lack of identifiability in the NMF model may be affecting our ability to evaluate the resulting fits. The only way to distinguish between models $(c\Theta)\left(\frac{1}{c}B\right)^T$ and ΘB^T is through the prior, and the prior may not be strong enough to guide inference to the right scaling.

Further, examining Figure 7 of supplementary material available at *Biostatistics* online, we note many cases where estimation fails to converge or concentrates around the incorrect value, though this may be related to rescaling. Generally, we do not find a direct probabilistic-programming implementation of the GaP or Z-GaP models reliable in the regimes under consideration, even when p_0 is known.

4. Data analysis

In applying probabilistic methods to microbiome data analysis, we concentrate on two questions:

1. Do these models fit the data well, and what techniques are available for performing this evaluation?
2. Supposing these models fit well, do they lend themselves to informative summaries of the original data?

To develop answers to these questions, we reanalyze the data of Dethlefsen and Relman (2011), a study of bacterial dynamics in response to antibiotic treatment. This study monitored the microbiomes of three patients over 10 months, with two antibiotics time courses introduced in between, in order to study the effect of antibiotic perturbations within the context of natural long-term dynamics. By applying principal components analysis (PCA), the study concluded that antibiotics cause substantial changes in short-term community composition, with certain species being substantially more resilient than others, and also detected long-term effects in one patient. The purpose of our case study is to compare these conclusions with those obtained through probabilistic latent variable models.

Variation in bacterial signatures tends to be dominated by strong inter-subject effects, and with only three subjects, there is little reason for a model which clusters across subjects. Hence, we choose to study one individual at a time. In this section, we focus on Subject F, who had been reported to exhibit incomplete recovery of the pre-antibiotic treatment bacterial community. However, analogous figures for the other two subjects are available as Figures 17 through 20 of supplementary material available at *Biostatistics* online. There are 2582 distinct species across these samples, and we perform no pre-filtering, though most species are present in very low numbers.

It is possible that future studies may consider a similar design, but involve many more subjects. In this case, it would be possible to do better than model each subject separately. The simplest way to share information across subjects would be to still have separate topics β_k^s for each individual s, but to place a common prior across all $\left(\beta_k^s\right)_{k,s}$. This approach has the disadvantage that topics are not directly comparable across subjects. An alternative that mitigates this issue would be to share topics $(\beta_k)_k$ across all individuals, choosing K to be larger than would be appropriate for any single subject. In this case, it would be expected that some θ_{ik} would be near zero for some topics for all samples within a subject—that

subject may never have samples drawn from a particular topic. It is possible to directly incorporate this behavior, in which some but not all topics are shared across subjects, into the model form by considering a hierarchical Dirichlet prior, rather than a standard one (Teh *and others*, 2005).

In this case study, we focus on LDA and the Dynamic Unigram model. A similar study using GaP and Z-GaP is omitted, in light of the difficulties in estimation during the simulation studies. Throughout, we apply VB, though considering the results of the simulation study, we exercise caution when interpreting estimated uncertainties. We set $K = 4$, based on the heuristic that a larger K would be less meaningful, since there are only 56 timepoints. In cases where it is of interest to choose K automatically, it would be possible to evaluate the likelihood of fitted models on a hold-out set, for a range of values of K, by fixing $\left(\hat{\beta}_k \right)_{k=1}^{K}$ and estimating new $\hat{\theta}_i$ for the hold-out samples (Blei *and others*, 2003). Indeed, this process reflects an advantage of probabilistic modeling: the Bayesian Occam's razor ensures that the posterior probability of a model won't always go up as K increases—contrast this with ordinary K-means, where tailored methods like the Gap or Silhouette statistics must be studied. Alternatively, the proposals of (Wallach *and others*, 2009) could be carried out. However, in this study, we guide the choice of K based on what we find most useful for scientific interpretation, rather than that which necessarily gives the best fit over the entire population.

Note that, we plot the fitted probabilities on a logit scale—for a raw vector of probabilities $\mathbf{p} = (p_1, \ldots, p_D) \in \mathcal{S}^{D-1}$, we plot $g\left(\mathbf{p} \right) := \left(\log p_1 - \overline{\log \mathbf{p}}, \ldots, \log p_K - \overline{\log \mathbf{p}} \right)$, which are similar to log-odds, but centered according to the average log probability, rather than any reference class.

Both LDA and the Dynamic Unigram models are more computationally intensive than more common approaches, like PCA or multidimensional scaling MDS), but are still amenable to routine application. For example, in the case studies below, LDA and the Dynamic Unigram model can both be compiled and run on a standard laptop (1.4GHz Intel Core i5 processor and 4GB RAM) in 5.62 and 30.0 min, respectively, and the longer runtime of the Dynamic Unigram model is due to its larger number of parameters. Further, both approaches can be scaled to much larger datasets without requiring substantially more memory (but requiring somewhat longer training time) by applying stochastic variational inference (Hoffman *and others*, 2013). For a comparison with the standard principal co-ordinates approach, see "Comparison with Principal Co-ordinates" section of supplementary material available at *Biostatistics* online.

4.1. *Latent Dirichlet allocation*

The fitted parameter values are summarized in Figure 3 and Figure 12 of supplementary material available at *Biostatistics* online. In Figure 3, rows represent topics, the *x*-axis represents time, and the *y*-axis gives the boxplots of posterior quantiles for each θ_{dk}.

This figure draws attention to the two antibiotic time courses, which took place between days 12–23 and 41–51. Topic 1 seems to become prominent during the interim period between the two time courses, suggesting that this is the bacterial community that fills in the niches left empty during the first time course. Conversely, Topic 3 seems to represent those bacteria that were present initially but are eliminated during the first time course, though there is a hint of a recovery at the end of sampling. This topic seems most closely related to the finding reported in Dethlefsen and Relman (2011) that Subject F experienced long-term antibiotic effects on bacterial community composition. Topics 2 and 4 seem to be overrepresented during the antibiotic treatments. Topic 2 is elevated immediately after the antibiotic treatment. Topic 4 seems to summarize those bacteria that are initially negatively impacted by the antibiotic treatments, but which recover relatively quickly and have higher abundance at the end of the trial than at the beginning. The fact that Topics 2 and 4 are learned from the data without enforcing any temporal evolution structure on the samples suggests that under an antibiotic perturbation, the system exhibits differential recovery.
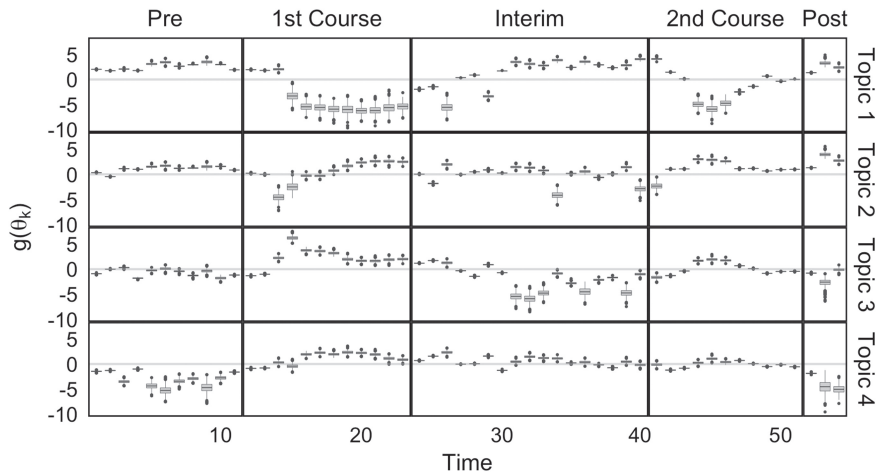
Fig. 3. Boxplots represent approximate posteriors for estimated mixture memberships θ_d, and their evolution over time. That is, each row of panels provides a different sequence of θ_{dk} for a single k, and different columns distinguish different phases of sampling. Note that the y-axis is on the g-scale, which is defined as a translated logit, $g\left(\mathbf{p}\right) := \left(\log p_1 - \overline{\log \mathbf{p}}, \ldots, \log p_K - \overline{\log \mathbf{p}}\right)$. Note that the first and second antibiotic time courses result in meaningful shifts in these sequences, and that there appear to be long-term effects of treatment among bacteria in Topic 3.

To interpret these topics in terms of their bacterial community fingerprint, we study the estimated topic distributions β_k. This is displayed in Figure 12 of supplementary material available at *Biostatistics* online. The four rows correspond to the $K = 4$ estimated topics. Within a row, we show 95% credible intervals associated with the posterior samples for each bacterium. Different colors identify different taxonomic families, and bacteria are sorted according to evolutionary relatedness. For clarity, we have displayed only the 750 most abundant bacteria.

Considering the mixture probabilities in Figure 3, those bacteria with large probabilities in the second and fourth rows of Figure 12 of supplementary material available at *Biostatistics* online constitute a large fraction of the samples taken during antibiotics time courses, reflecting those whose abundances increase rapidly (Topic 2) or are initially drop but then increase gradually (Topic 4) during the antibiotics time courses. These distributions are relatively more concentrated on a small subset of bacteria with high probabilities, reflected by the drop in logitted probabilities far below zero. This corresponds to a decrease in community diversity during antibiotic time courses. We also note groups of neighboring bacteria with similarly elevated topic probabilities. It is encouraging that, even without specifying smoothness along the phylogenetic tree in the prior for the β_ks, such smoothness emerges in the fitted model.

One disadvantage of plotting the posteriors for the β_k in this way is that it is difficult to determine the identify of any particular bacterial species associated with a specific credible interval. It also hides any interesting variation that may be occurring among species not among the 750 most abundant. To address these issues, we have designed an interactive version, available online at (link blinded for review). Alternatively, individual species that are primarily assigned to individual topics can be used to characterize the topics. For example, in Figure 4 we have screened all species for those that seem primarily associated with a single topic, and plotted them along with the topic for they are representative. Specifically, representatives of the kth topic were found by choosing the 50 species v with the largest values of $\beta_{kv} - \sum_{k' \neq k} \beta_{k'v}$.

This allows an interpretation of the topics in terms of the raw data, and the display reinforces the findings of Figure 3 and Figure 12 of supplementary material available at *Biostatistics* online. In particular, this
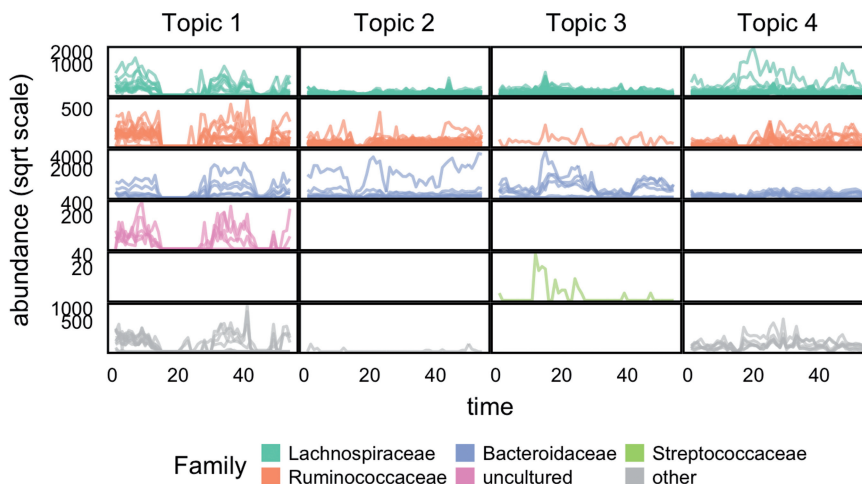
Fig. 4. Individual species that have been identified as prototypical of single topics are highlighted here, to provide a characterization of topics linked to the original bacterial abundance data. Rows and columns of panels correspond to taxonomic families and estimated topics, respectively, and each trajectory gives the abundance of a single species over time. Note that the y-axis is on a square root scale. This view provides more evidence for the interpretation of the four topics as (1) decreased abundance during antibiotic time courses, (2) increased abundance during antibiotic time courses, (3) disappearance after first time course, and (4) delayed but sustained increase after first time course. Refer online version for color figure.

view makes clear how much variation there is in species abundances within taxonomic families (e.g. compare Topics 1 and 2 among representative Lachnospiraceae and Bacteroides). Instead of displaying all species together, Figures 21 through 24 of supplementary material available at *Biostatistics* online sort individual species by the topic representativeness measure.

Similar approaches can be employed to identify taxa that disproportionately contain species with high membership in particular topics. For example, we can compute the average topic representativeness statistic defined above across all species within a taxonomic family in order to characterize that family. Families that are highly associated with individual topics are displayed in Figures 25 and 26 of supplementary material available at *Biostatistics* online.

4.2. *Dynamic Unigram model*

While we can interpret the LDA-estimated $\hat{\theta}_d$ according to their temporal context, this information was never directly provided to the algorithm. In contrast, we can apply the Dynamic Unigram model to the same data, which explicitly models temporal evolution. Unlike LDA, however, this model does not seek latent mixture structure. Our primary results are displayed in Figure 13.

Each row in this figure is interpreted similarly to a row in Figure 12 of supplementary material available at *Biostatistics* online, except now they correspond to estimated proportions over time $g(\mu_t)$ rather than transformed topics $g(\beta_k)$. Only four of the 54 total timepoints is displayed, highlighting a time window around the first antibiotic time course. Such a display implicitly assumes a relatively smooth interpolation between timepoints, which is enforced by the form of the Dynamic Unigram model.

This analysis yields conclusions similar to those obtained through LDA, though reaching them requires somewhat more effort. For example, on Day 10, most $g(\mu_{tv})$'s have most of their mass concentrated above zero, and not many are positioned exceptionally far from the bulk. This is consistent with higher community diversity before the antibiotic time course. On the other hand, at time 15, 1 day after the time course began,
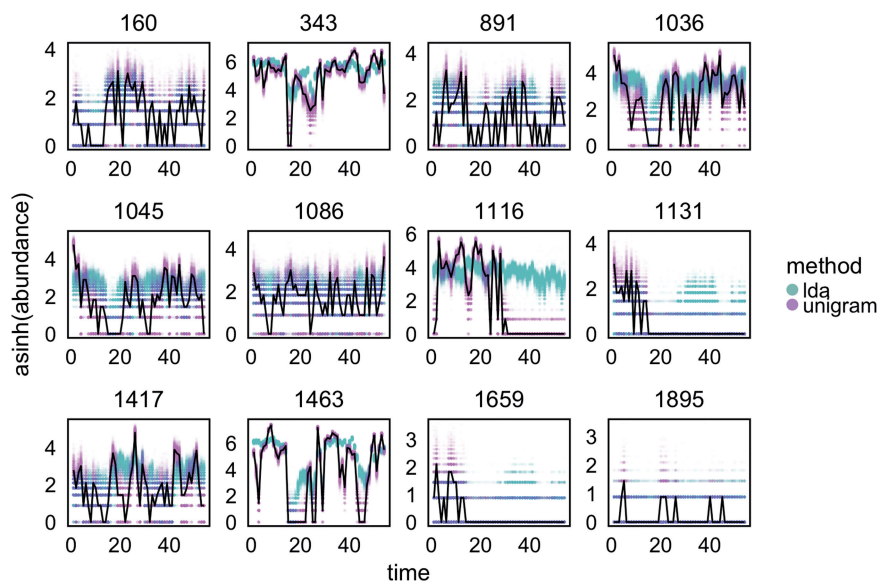
Fig. 5. We can visualize the simulated time series for a subset of species and compare them with the observed ones, as a posterior check. Each panel represents one species. The black lines represent the observed asinh-transformed abundances for subject F over time. The blue and purple dots give the posterior predictive realizations for these species over time, according to LDA and the Dynamic Unigram model, respectively. Refer online version for color figure.

most species have quantiles lower than zero, while a few are positioned much higher than the rest. This corresponds to a less diverse community, whose membership is concentrated on those species with outlying intervals. This decrease in diversity seems most profound at time 20; by time 25, during the interim, much of the community seems to have recovered.

Further, while we continue to see differential recovery across bacteria, the effect is not as obvious as in LDA, where this effect was decomposed across topics. For example, many subintervals of Lachnospiraceae seem to return to their pre-antibiotics levels by time 25, while the Ruminococceae continue to have low values of μ_{tv}.

4.3. *Posterior predictive checks*

While we have found both LDA and the Dynamic Unigram model qualitatively useful, it is still important to seek more formal diagnostics of model fit. Herein, we consider several posterior predictive checks, as reviewed in "Posterior Predictive Checks" section of supplementary material available at *Biostatistics* online.

To this end, in Figure 5, we plot observed time series for a random subset of the 350 most abundant bacteria and contrast them with samples from the posterior predictive according to the four topic LDA model of Section 4.1 and the Unigram model of Section 4.2. Each subpanel corresponds to a single species. The black lines represent observed time series. Note that the *y*-axis scales vary, as some species are much more abundant than others. Each dot is a simulated timepoint from a posterior predictive time series. Formally, this corresponds to the choice $T(x) = (x_{\cdot v})_{v \in \mathcal{V}}$ where the random subset $\mathcal{V}$ indexes the displayed species.

For LDA, the posterior predictive time series are on the appropriate scale with approximately the correct shape. However, we observe two substantial types of departures between simulated and observed data.

First, for series with larger counts, the posterior predictive tends to oversmooth. For example, the drop to 0 in species 343 is not captured in any posterior predictive samples. Similar oversmoothing is visible in species 1036, 1116, and 1463. A more startling type of departure occurs in the second half of series 1116. Herein , the posterior predictive distribution places most mass on the event that the bacterial time series rebounds to its initial abundance when in reality the species vanishes during the second antibiotic time course, never to return. A potential explanation for LDA's failure to capture this pattern is that few highly abundant species disappear after the second time course, and hence they are not captured by the global LDA summary. This suggests a technique for highlighting outlier species: we can look at the average discrepancy between observed series and their posterior predictive samples.

On the other hand, for the Dynamic Unigram model, the posterior predictive distribution places most of its support close to the observed species series. Usually, this is desirable behavior, indicating good model fit. However, here, there is reason for concern—the Unigram model may not do much more than fit empirical proportions at each timepoint, and there may be potential to produce more succinct summaries that still preserve the essential structure of the data. That is, the unigram model seems over-parameterized, simply memorizing the input.

An alternative posterior predictive check compares PCA scores and loadings in the true and posterior predictive data. Our motivation is that many microbiome studies base their findings on views generated by PCA, so it would be encouraging if our probabilistic summaries typically agree with the reductions produced by PCA. Formally, we construct $T(x) = (u_1, u_2, \lambda_1, \lambda_2, v_1, v_2)$ where (u_i), (λ_i) and (v_j) are left singular vectors, eigenvalues, and right singular vectors of x, respectively.

Figure 15 of supplementary material available at *Biostatistics* online gives the PCA eigenvalues between the true and posterior predictive samples, after applying an asinh-transformation and filtering to the 1000 highest-variance bacteria. The associated scores and loadings figures are available as Figure 16. In Figure 15, each black point provides the log-eigenvalue observed in the original data. The blue and purple clouds consistent of small, semi-transparent, and horizontally-jittered points associated with eigenvalues from posterior predictive samples generated by LDA and the Dynamic Unigram model, respectively. For LDA, the posterior predictive samples have comparable top four eigenvalues, but rapidly drop-off between the fourth and fifth eigenvalues. On the other hand, the observed data have a more steady decline. This is likely a consequence of using $K = 4$ topics in the LDA model, which would be consistent with the matrix factorization view of LDA described in Section 2.4. Considering these scree plots, it may be safe to increase K in follow-up analysis, as long as topics remain interpretable. In contrast, the eigenvalues for the Dynamic Unigram model closely match those in the observed data (even overestimating the size of small eigenvalues) lending further evidence to the claim that this model is over-parameterized.

5. DISCUSSION

We have described the utility of taking a probabilistic modeling perspective in the analysis of microbiome data. We have provided a detailed implementation of benchmark analysis approaches, along with exploratory visualization of fitted parameters and model assessment through posterior predictive checks. Through simulation, we have established heuristics for determining the appropriateness of applying different models and inference mechanisms, depending on the overall data generation regime. On a real microbiome data analysis problem, we have characterized the advantages and limitations of two probabilistic modeling techniques. Rather than focusing on any single model, like most earlier work, we have emphasized the practice of contrasting, critiquing, and learning from multiple alternatives. Throughout, we have emphasized both insights, in terms of estimated parameters, as well as uncertainty, in the form of full approximate posterior distributions. We hope our efforts help to widen the biostatistician's toolbox and clarify the practical advantages and limitations of different approaches to microbiome data.

Microbiome studies are a source of richly structured, high-dimensional data, coupled with novel scientific problem setups. For example, in the antibiotics data set described here, we have already encountered structure in the form of zero-inflated counts, time series with changepoints, and apriori known phylogenetic relationships between features. Future microbiome studies will likely collect an increasing number of samples as well as more data sources per sample – metabolomic and genomic, in addition bacterial abundance, for example. Further, the investigations often revolve around a combination of ecological community characterization and medically-relevant identification of treatment effects. We believe we have only begun to see the potential for probabilistic methods to guide careful scientific reasoning—which emphasizes both insights and the degree of uncertainty about them—in these complex scenarios.

6. REPRODUCIBILITY

Code for all simulations, data analysis, and figures is available at https://github.com/krisrs1128/microbiome_plvm. Detailed instructions are available in the repository README.md. Further, a docker image with all software requirements preinstalled is available from https://hub.docker.com/r/krisrs1128/microbiome_plvm/, and the corresponding Dockerfile is provided in the github repository.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

REFERENCES

BLEI, D. M. AND LAFFERTY, J. D. (2006). Dynamic topic models. In: *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, Pennsylvania, United States: ACM. pp. 113–120.

BLEI, D. M, NG, A. Y. AND JORDAN, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research* **3**, 993–1022.

CALLAHAN, B. J., MCMURDIE, P. J. AND HOLMES, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal* **11**, 2639–2643.

CANNY, J. (2004). Gap: a factor model for discrete data. In: *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Sheffield, South Yorkshire, UK: ACM. pp. 122–129.

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**, 1–32.

CHEN, J., BUSHMAN, F. D., LEWIS, J. D., WU, G. D. AND LI, H. (2013). Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics* **14**, 244–258.

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**, 418–442.

CHEN, X., HE, T., HU, X., ZHOU, Y., AN, Y. AND WU, X. (2012). Estimating functional groups in human gut microbiome with probabilistic topic models. *IEEE Transactions on Nanobioscience* **11**, 203–215.

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.

FUKUYAMA, J. (2017). Adaptive gPCA: a method for structured dimensionality reduction. *arXiv preprint arXiv:1702.00501*.

FUKUYAMA, J., RUMKER, L., SANKARAN, K., JEGANATHAN, P., DETHLEFSEN, L., RELMAN, D. A. AND HOLMES, S. P. (2017). Multidomain analyses of a longitudinal human microbiome intestinal cleanout perturbation experiment. *PLoS Computational Biology* **13**, e1005706.

GILBERT, J. A., JANSSON, J. K. AND KNIGHT, R. (2014). The earth microbiome project: successes and aspirations. *BMC Biology* **12**, 69.

HOFFMAN, M. D., BLEI, D. M., WANG, C. AND PAISLEY, J. W. (2013). Stochastic variational inference. *Journal of Machine Learning Research* **14**, 1303–1347.

HOLMES, I., HARRIS, K. AND QUINCE, C. (2012). Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One* **7**, e30126.

JIANG, X., HU, X. AND XU, W. (2017). Microbiome data representation by joint nonnegative matrix factorization with Laplacian regularization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **14**, 353–359.

NIGAM, K., MCCALLUM, A. K., THRUN, S. AND MITCHELL, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine Learning* **39**, 103–134.

ROMERO, R., HASSAN, S. S., GAJER, P., TARCA, A. L., FADROSH, D. W., NIKITA, L., GALUPPI, M., LAMONT, R. F., CHAEMSAITHONG, P., MIRANDA, J. scemphAND OTHERS. (2014). The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome* **2**, 4.

SCHLOSS, P. D. AND HANDELSMAN, J. (2007). The last word: books as a statistical metaphor for microbial communities. *Annual Review of Microbiology* **61**, 23–34.

SEGATA, N., IZARD, J., WALDRON, L., GEVERS, D., MIROPOLSKY, L., GARRETT, W. S. AND HUTTENHOWER, C. (2011). Metagenomic biomarker discovery and explanation. *Genome Biology* **12**, R60.

SHAFIEI, M., DUNN, K. A., BOON, E., MACDONALD, S. M., WALSH, D. A., GU, H. AND BIELAWSKI, J. P. (2015). BioMiCo: a supervised Bayesian model for inference of microbial community structure. *Microbiome* **3**, 8.

TEH, Y. W., JORDAN, M. I., BEAL, M. J. AND BLEI, D. M. (2005). Sharing clusters among related groups: Hierarchical Dirichlet processes. In: *Advances in Neural Information Processing Systems*. Vancouver, Canada: NIPS 2014. pp. 1385–1392. https://papers.nips.cc/book/advances-in-neural-information-processing-systems-17-2004.

WALLACH, H. M., MURRAY, I., SALAKHUTDINOV, R. AND MIMNO, D. (2009). Evaluation methods for topic models. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. Montreal, Canada: ACM. pp. 1105–1112.

WANG, B. AND TITTERINGTON, D. M. (2005). Inadequacy of interval estimates corresponding to variational Bayesian approximations. In: *AISTATS*, Barbados. http://www.gatsby.ucl.ac.uk/aistats/AIabst.htm

WANG, Y.-X. AND ZHANG, Y.-J. (2013). Nonnegative matrix factorization: a comprehensive review. *IEEE Transactions on Knowledge and Data Engineering* **25**, 1336–1353.