Mixed-membership (MM) models such as latent Dirichlet allocation (LDA) have been applied to microbiome compositional data to identify latent subcommunities of microbial species. These subcommunities are informative for understanding the biological interplay of microbes and for predicting health outcomes. However, microbiome compositions typically display substantial cross-sample heterogeneities in subcommunity compositions—that is, the variability in the proportions of microbes in shared subcommunities across samples—which is not accounted for in prior analyses. As a result, LDA can produce inference, which is highly sensitive to the specification of the number of subcommunities and often divides a single subcommunity into multiple artificial ones. To address this limitation, we incorporate the logistic-tree normal (LTN) model into LDA to form a new MM model. This model allows cross-sample variation in the composition of each subcommunity around some “centroid” composition that defines the subcommunity. Incorporation of auxiliary Pólya-Gamma variables enables a computationally efficient collapsed blocked Gibbs sampler to carry out Bayesian inference under this model. By accounting for such heterogeneity, our new model restores the robustness of the inference in the specification of the number of subcommunities and allows meaningful subcommunities to be identified.

KEYWORDS
Bayesian inference, compositional data, latent variable models, mixed-membership models

INTRODUCTION

The human gut microbiome is the genetic content of all bacteria, archaea, viruses, and eukaryotic microbes residing in the human gut and is commonly used to profile the composition of the gut microbiota. Advances in next-generation sequencing techniques have substantially reduced the cost of this approach and made it widely accessible. One cost-effective microbiome profiling strategy is based on targeting a single marker gene, the 16S ribosomal RNA (rRNA) gene, through amplicon-based sequencing (Li, 2015). A more expensive, but more precise, approach is whole-genome shotgun metagenomic sequencing (Weber & Myers, 1997). Traditionally, sequencing reads have been clustered into operational taxonomic units (OTUs), which serve as the basic unit of microbial taxa. Recently, amplicon sequencing variants (ASVs) have come into wider use as they can achieve more precise characterization of microbial species and resolve the sample-specificity issue of the OTU (Callahan et al., 2017). Our work is applicable to either method of characterizing microbial taxa; in the following, we shall generically refer to the basic unit as ASVs.
Gut microbiome studies often involve highly heterogeneous samples due to the multitude of factors that can influence an individual’s gut microbiota. A useful data analytical strategy for microbiome compositions is to sort microbiome samples into clusters characterized by particular compositional signatures. In the context of gut microbiome, these clusters are called “enterotypes” (Siezen & Kleerebezem, 2011) and are associated with health outcomes (Del Chierico et al., 2014). One of the most popular microbiome clustering methods is the Dirichlet-multinomial mixture (DMM) model (Holmes et al., 2012; Nigam et al., 2000), which uses a hierarchical structure to allow within-cluster cross-sample variability in subcommunity compositions. However, the DMM is too restrictive to realistically characterize the within-cluster cross-sample variance in microbiome data (Tang et al., 2018; Wang & Zhao, 2017) as it uses a single scalar parameter to characterize the entire covariance structure across all microbial taxa. More general methods have recently been introduced to alleviate, though not eliminate, this limitation through the use of Dirichlet-tree models (Dennis III, 1991; Wang & Zhao, 2017).

Such clustering analysis, however, makes the implicit assumption that each microbiome sample must belong to a single signature “community” characterized by the cluster centroid. This assumption is often unrealistic and overly restrictive for complex environments such as the gut microbiome (Holmes et al., 2012; Mao et al., 2020). Recent developments embrace the more relaxed biological hypothesis that the ASVs characterizing a microbiota sample hail from a combination of multiple microbial “clusters,” or more precisely “subcommunities.”

Mixed-membership (MM) models are generalizations of clustering models that provide a generative modeling framework for data involving subcommunity structure as they allow each sample to be composed of multiple subcommunities. Sankaran and Holmes (2019) applied the most well-known MM model, latent Dirichlet allocation (LDA), to microbiome profiling. Earlier, Shafiee et al. (2015) and Deek and Li (2019) proposed variations of LDA accounting for environmental factors and inflated zero counts, respectively, in the microbiome context. The key motivation for our paper is the observation that existing MM models such as LDA and its variations—originally developed for other contexts such as topic modeling (Blei et al., 2003) and population genetics (Pritchard et al., 2000)—do not incorporate key features of microbiome compositions. Most notably, they assume that a microbial subcommunity’s composition must remain exactly the same across all samples. This is unrealistic in the vast majority of microbiome studies collected from diverse environments such as the gut where samples often possess large heterogeneities (Jeganathan & Holmes, 2021; Tang et al., 2018). It is interesting to note that such heterogeneity has been well recognized in clustering models for microbiome data (Holmes et al., 2012; Mao et al., 2020), but has been largely ignored in existing MM models. Additionally, choosing the number of subcommunities for LDA is not trivial in the presence of cross-sample heterogeneity, and LDA-based approaches often lead to overestimates in the number of subcommunities in microbiome applications (Fukuyama et al., 2012).

We introduce a generalization of LDA that aims to appropriately incorporate cross-sample heterogeneity, or “random effects,” in microbial subcommunity compositions due to unmeasured sources, thereby leading to more accurate identification of subcommunities in MM models. Our approach takes advantage of the availability of a natural tree structure relating the microbial taxa—the phylogenetic tree—which allows us to decompose the compositional vector into a collection of binomial observations on the tree nodes. This transform serves two purposes. First, it allows us to model the heterogeneity by modeling the vector of log-odds transforms of the binomial probabilities at each node as Gaussian. By modeling the subcommunity compositions as realizations from this logistic-tree normal (LTN) (Wang et al., 2021) distribution, we are able to impose constraints on the underlying covariance structure to ensure the identifiability of the subcommunities. A second purpose of the tree-based transform is computational. By utilizing the Pólya-Gamma (PG) data augmentation technique (Polson et al., 2013), Bayesian inference under the resulting MM model can be readily accomplished through fully conjugate collapsed blocked Gibbs sampling. We term our new model logistic-tree normal latent Dirichlet allocation (LTN-LDA).

Several other relevant prior works are worth mentioning. Graph-sparse LDA (Doshi-Velez et al., 2015) also incorporates random effects from subcommunity-to-subcommunity using a tree structure. However, in the context of microbiome compositions, it would assume that every node of the tree is an ASV, which can occur in a sample and is thus incompatible with the phylogenetic tree. Other tree-based MM methods include Tam and Schultz (2007), which uses trees to model the abundance of subcommunities in samples, and Andrzejewski et al. (2009), which uses mixtures of trees to model subcommunity composition by explicitly modeling, which ASVs must co-occur and which cannot.

In the following, we will briefly review the LDA and LTN models before introducing the LTN-LDA model. We will augment the LTN-LDA model using a class of auxiliary PG variables (Polson et al., 2013) and present a collapsed blocked Gibbs sampler for carrying out fully Bayesian inference. We will demonstrate in simulations...
that, in the presence of cross-sample heterogeneity, inference by LTN-LDA is robust with respect to overspecifying the number of subcommunities while inference by LDA can be highly sensitive to the choice of the number of subcommunities. We apply LTN-LDA to the data set of Dethlefsen and Relman (2011), which has been used for demonstrating MM models in the microbiome settings (Sankaran & Holmes, 2019), and compare our results to LDA.

2 METHODS

2.1 LDA

Let there be $D$ samples consisting of counts of $V$ unique ASVs indexed by $1, 2, ..., V$. For sample $d$, let $x_d = (x_{d,1}, ..., x_{d,V})$ be the vector of ASV counts such that $x_{d,v}$ is the total count for ASV $v$ in sample $d$. Let $N_d = \sum_{v=1}^{V} x_{d,v}$ be the sum of counts in sample $d$, which is determined by the sequencing depth. Subcommunities are defined to be collections of ASVs that co-occur in samples at given relative proportions. An ASV can occur in multiple subcommunities at various abundances and the key assumption underlying an MM model, in contrast to a clustering model, is that different instances (i.e., different sequencing reads) of the same ASV in a sample can arise from the participation of that ASV in multiple microbial subcommunities. Key parameters of interest in MM models are subcommunity abundance, that is, the proportions of the various subcommunities in each sample, and subcommunity composition, that is, the proportions of the ASVs in each subcommunity.

To describe LDA, it is convenient to introduce categorical indicators for each read and its associated subcommunity identity. For $d = 1, 2, ..., D$, let $w_d = (w_{d,1}, ..., w_{d,V})$ be the vector of ASV categories such that $w_{d,v} \in \{1, 2, ..., V\}$ is the categorical indicator of the ASV associated with the $n$th read in the sample. We refer to the elements $w_{d,n}$ in this vector as “tokens” to draw analogy with topic modeling. There, each token is a word in a document; here, each token corresponds to a read in a sample. We also note that $x_{d,v} = \sum_{n=1}^{N_d} \mathbf{1}_{\{w_{d,n}=v\}}$.

Let $\phi_d = (\phi_{d,1}, \phi_{d,2}, ..., \phi_{d,K})' \in \Delta^{K-1}$, where $\Delta^S$ is the $S$-dimensional simplex, be the subcommunity abundance vector. That is, $\phi_{d,k}$ represents the relative abundance of subcommunity $k$ in sample $d$, and so $\phi_d$ specifies the categorical distribution of each token over the $K$ underlying subcommunities in sample $d$. Let $z_{d,n}$ represent the subcommunity from which the $n$th token in sample $d$ arises from and let $z_d$ be the vector of all such assignments for sample $d$. Also, let $\beta_k = (\beta_{k,1}, \beta_{k,2}, ..., \beta_{k,V})' \in \Delta^{V-1}$ be the subcommunity composition for subcommunity $k$. That is, $\beta_k$ gives the relative proportions of the $V$ unique ASVs in subcommunity $k$. For $d = 1, ..., D$ and $n = 1, ..., N_d$ and while $\alpha$ and $\gamma$ are hyperparameters, the LDA model (Figure 1A) (Blei et al., 2003) is then

$$w_{d,n} | z_{d,n}, \beta_{z_{d,n}} \overset{\text{iid}}{\sim} \text{Cat}(\beta_{z_{d,n}}) \quad z_{d,n} | \phi_{d} \overset{\text{iid}}{\sim} \text{Cat}(\phi_{d}) \quad \phi_{d} | \alpha \overset{\text{iid}}{\sim} \text{Dir}(\alpha) \quad \beta_{k} | \gamma \overset{\text{iid}}{\sim} \text{Dir}(\gamma).$$

Though LDA can be applied in the microbiome context (Sankaran & Holmes, 2019), it does not account for cross-sample heterogeneity in subcommunity composition. In particular, it assumes that the $\beta_k$ are the exact same across all samples. This is inconsistent with the empirical behavior of the microbiome where large cross-sample heterogeneities exist (Holmes et al., 2012). LDA thus tends to interpret cross-sample heterogeneity as the presence of additional subcommunities.

2.2 Incorporating cross-sample heterogeneity

We shall enrich the LDA framework to allow the subcommunity compositions to vary across samples. There are several hierarchical models for microbiome compositions such as the Dirichlet-multinomial (DM) model (Holmes et al., 2012; Nigam et al., 2000) and Aitchinson’s log-ratio–based normal (LN) models (Aitchison, 1982), which could be embedded into LDA for this purpose. However, the DM is highly restrictive in its ability to characterize the underlying cross-sample variability as the Dirichlet distribution has only one scalar variance parameter, while the LN models are computationally challenging due to lack of conjugacy to the multinomial sampling model. To resolve these difficulties, we adopt the recently introduced LTN model (Wang et al., 2021). In particular, we will show that the LTN model can be embedded into the LDA model to accommodate cross-sample heterogeneity and that posterior inference can be accomplished through simple collapsed blocked Gibbs sampling using a data-augmentation technique called PG augmentation. Moreover, since the adoption of the LTN model requires specifying a dyadic partition tree on the ASVs, the phylogenetic tree relating the taxa is a natural choice, though we will also examine robustness to the choice of the tree.

2.2.1 The phylogenetic tree

Let $T$ denote a phylogenetic tree capturing genetic similarities between the observed ASVs. The leaf nodes in the tree correspond to the observed ASVs in the data set. Each interior node is the inferred common ancestral taxon for
the ASVs lying in the corresponding descendant subtree at
the node. Each node (or taxon) $A$ in the phylogenetic tree
$\mathcal{T}$ can be represented by the collection of its descendant
ASVs. In particular, each leaf node $A$ contains a single ASV,
whereas each internal node $A$ contains multiple ASVs.
In the following, we let $I$ be the set of internal nodes.
Throughout this work, we shall assume that the phyloge-
netic tree is rooted and binary in the sense that each $A \in I$
has exactly two child nodes (i.e., direct descendants); let $A_l$
and $A_r$ be the left and right children of $A$, respectively.

2.2.2 The LTN model

We shall adopt the LTN model introduced in Wang et al.
(2021) as the sampling model for the ASV count distri-
bution within each subcommunity. LTN is a distribution
on a tree-based log-odds transform of the categorical
probabilities $\beta = (\beta_1, \beta_2, \ldots, \beta_\mathcal{V})' \in \Delta^{\mathcal{V}-1}$. Specifically,
given the phylogenetic tree $\mathcal{T}$, for each interior node,
we define $\bar{\theta}(A) = \sum_{v \in A_l} \beta_v \sum_{v \in A} \beta_v$: the probability that a token
belongs to an ASV in $A_l$ given that it belongs to an
ASV in $A$. The collection of $\bar{\theta}(A)$ on all $A \in I$ gives an
equivalent reparameterization of $\beta$. In Figure 2, we plot an
example phylogenetic tree over six ASVs with labeled
nodes (Figure 2A) and with labeled $\beta_v$ and $\theta(A)$
(Figure 2B) to demonstrate the link between the $\beta_v$
and the $\theta(A)$.

After taking the logit transform of these binomial prob-
abilities on the tree nodes, $\psi(A) = \log \frac{\theta(A)}{1-\theta(A)}$, let $\psi$ be

![Diagram of LDA and LTN-LDA](image)

**Figure 1** Graphical model representations for LDA and LTN-LDA

![Diagram of phylogenetic tree and parameters](image)

**Figure 2** An example phylogenetic tree for six ASVs and the
graphical relationship between $\beta_v$ and $\theta(A)$.
overall average profile of the count distribution and the cross-sample variability.

Posterior computation under LTN, which we will describe later, relies on an equivalent representation of the categorical sampling on the leaves of the tree as a collection of sequential binomial experiments on the internal nodes of the tree. Specifically, generating a categorical draw from the probability vector $\beta$ can be achieved by sequentially “dropping” the token from top-to-bottom along the phylogenetic tree: at each node, determine whether the token belongs to the left or right child node with probabilities $\theta(A)$ and $1 - \theta(A)$, respectively. More formally, for each node $A \in I$, we use $y(A)$ to denote the total counts associated with the ASVs descended from node $A$. That is, $y(A) = \sum_{n=1}^{N} w_{n} \in A$, where $w_{n}$ represents the $n$th count. Generating a multinomial count vector with probability $\beta$ can be achieved by sequentially drawing $y(A)$ given $y(A)$ from $\text{Bin}(y(A), \theta(A))$. Putting the pieces together, and letting $\expit(\psi) = 1/(1 + e^{-\psi})$, LTN is the following generative model: for all internal nodes $A \in \mathcal{T}$,

$$y(A) \mid y(A), \psi(A) \sim \text{Bin}(y(A), \theta(A) = \expit(\psi(A))),$$

$$\psi \mid \mu, \Sigma \sim \text{MVN}(\mu, \Sigma).$$

### 2.3 | LTN-LDA

We incorporate the LTN model into LDA to allow cross-sample heterogeneity in subcommunity compositions. The resulting model is termed LTN-LDA. Specifically, for $d = 1, \ldots, D$, $k = 1, \ldots, K$, $n = 1, \ldots, N_{d}$, and $A \in I$, where the subscripts $d, k, n$ indicate the corresponding quantities associated with the $d$th sample, $k$th subcommunity, and $n$th read, the model is as follows

$$y_{d,k}(A) \mid y_{d,k}(A), \psi_{d,k}(A) \sim \text{Bin}(y_{d,k}(A), \expit(\psi_{d,k}(A))),$$

$$y_{d,k}(A) = \sum_{n=1}^{N_{d}} 1_{x_{d,n} = k} 1_{w_{d,n} \in A},$$

$$z_{d,n} \mid \phi_{d} \sim \text{Cat}(\phi_{d}),$$

$$\phi_{d} \mid \alpha \sim \text{Dir}(\alpha),$$

$$\psi_{d,k} \mid \mu_{k}, \Sigma_{k} \sim \text{MVN}(\mu_{k}, \Sigma_{k}),$$

$$\mu_{k} \mid \mu_{0}, \Lambda_{0} \sim \text{MVN}(\mu_{0}, \Lambda_{0}),$$

$$\Sigma_{k} \mid G \sim G.$$

Note that we also endowed the subcommunity mean $\mu_{k}$ and covariance $\Sigma_{k}$, with corresponding priors $\text{MVN}(\mu_{0}, \Lambda_{0})$ and $G$, which will be specified later. Figure 1B provides the graphical model representation for this full hierarchical model. The key distinction between LTN-LDA and LDA is that LTN-LDA uses a hierarchical kernel, namely, LTN, to model cross-sample heterogeneity for each subcommunity. In particular, the composition in sample $d$ of subcommunity $k$ is determined by $\psi_{d,k}$ and is explicitly allowed to vary across samples.

Without additional constraints on the high-dimensional covariance matrices for each subcommunity, $\Sigma_{k}$, the model is too flexible (Haffari & Teh, 2009), and can become unidentifiable. Additional structural constraints serving the purpose of regularization on the covariance structure are thus necessary and so we assume that $\Sigma_{k}$ is a diagonal covariance matrix. An LTN distribution with diagonal covariance is similar in distributional properties to a Dirichlet-tree multinomial (DTM) distribution (Dennis III, 1991; Wang & Zhao, 2017) but is computationally more efficient because there are no known conjugate priors for the mean and variance parameters under the DTM model. While this limitation is manageable when the DTM is used as a standalone model or the top layer in a hierarchical model, when embedded as a kernel within an MM model such as LDA, the incurred numerical computational cost becomes prohibitive. For more details, see Supporting Information S1.

While the covariance constraint may appear strong, we note that the dependence among the tree-based log-odds ratios is generally much weaker than the complex dependence structure among the ASV counts themselves. In a sense, the tree-based log-odds transform of the abundance vectors “decorrelates” the data. For the interested reader, this decorrelation phenomenon is analogous to the so-called “whitening” effects in wavelet analysis (Nason, 2008), as the dyadic tree transform we incorporate here is the counterpart of Haar-wavelet transform on functions. In Supporting Information S2, we investigate the effects of relaxing the diagonal covariance to a blocked diagonal covariance, and the results show that the additional sophistication does not lead to noticeable improvement in the inference.

Aside from the diagonal covariance, we also assume that the amount of variability for each node depends on that node’s distance to the bottom (i.e., leaf) level of the tree. In particular, we assume that taxa close to the bottom of the phylogenetic tree have larger cross-sample variability in the corresponding log-odds ratio than those that are distant. This is motivated by the biological intuition that taxa close to each other on deep levels of the phylogenetic
tree tend to have comparable functionality; the relative proportions of such taxa thus often display elevated levels of variance (Jeganathan & Holmes, 2021).

Specifically, let |A| measure the distance of A from the leaf level by denoting the number of leaves descended from node A. For i = 1, ..., p, k = 1, ..., K, C ∈ N (a tuning parameter), and τ_k = (τ^1_k, ..., τ^p_k), the prior we adopt has the form \( \tau_k \mid A_i \sim IG(a_1, a_2, b) \) where

\[
\tau_k \mid A_i, a_1, a_2, b \sim \begin{cases} 
IG(a_1, b) & |A_i| \geq C, \\
IG(a_2, b) & |A_i| < C.
\end{cases}
\]

We default to \((a_1, a_2, b) = (10^4, 10^2, 10)\) and note that while we still refer to the \(\psi_{d,k}\) as being drawn from a multivariate normal distribution, we have \(\psi_{d,k} \mid \mu_k, \tau_k \sim N(\mu_k, \tau_k^-)\).

This choice of priors ensures conjugate updating and avoids identifiability issues. Further, it partitions the internal nodes of the tree in two: We shall refer to these sets as the upper tree \(U = \{A \in I : |A| \geq C\}\) and the lower tree \(L = \{A \in I : |A| < C\}\). In \(U\), the hyperparameters \(a_1\) and \(b\) are such that the \(\tau_k^1\) will be small and the \(\psi_{d,k}\) will vary little around \(\mu_k^1\); in \(L\), the hyperparameters \(a_2\) and \(b\) are such that the \(\tau_k^2\) are allowed to be large and the \(\psi_{d,k}\) can vary significantly across samples. This implies that if \(A_k\) is the child of \(A_i\), and \(A_i \in L\) but \(A \in U\), then all ASVs descended from \(A_c\) can substitute for each other across samples in a given subcommunity. We call sets of ASVs, which are allowed to substitute for each other, substitution sets. All ASVs are either part of a substitution set or singletons. The tree structure is critical to how LTN-LDA models cross-sample heterogeneity, and we include an analysis on the robustness to misspecified trees in Supporting Information S3.

2.4 Bayesian inference by collapsed blocked Gibbs sampling

While the LTN-LDA model is not conditionally conjugate by itself, one can restore conjugacy by introducing a class of PG latent variables (Polson et al., 2013) \(v_{d,k}(A)\)—one for each interior node A—which are independent of \(y_{d,k}(A)\) conditioned on \(y_{d,k}(A)\) and \(\psi_{d,k}(A): v_{d,k}(A) \mid y_{d,k}(A), \psi_{d,k}(A) \sim PG(y_{d,k}(A), \psi_{d,k}(A))\). The full conditional for \(\psi_{d,k}(A)\) is then proportional to

\[
\exp \left( \frac{1}{2} \left( \psi_{d,k}(A) - y_{d,k}(A) \right) - \frac{1}{2} \psi_{d,k}(A)^2 \right),
\]

which takes a quadratic form in the exponent and thus is conjugate to the Gaussian model on \(\psi_{d,k}(A)\). The graphical model for LTN-LDA with the PG variables is presented in Figure 1C. To speed up the sampling of PG variables, we adopt an approximate sampler proposed by Glynn et al. (2019) for \(y_{d,k}(A) \geq 30\). Further, we integrate \(\phi_d\) out of the sampling model to improve convergence as in Griffiths and Steyvers (2004). The algorithm scales linearly with D, K, V, and \(N_d\); for details, see Supporting Information S4.

3 | NUMERICAL EXPERIMENTS

3.1 Robustness in choosing the number of subcommunities

The true number of subcommunities \(K\) in a given data set is typically unknown and it is common to treat \(K\) as a tuning parameter. However, for data with large cross-sample heterogeneity such as microbiome data, intuition suggests that a model assuming zero heterogeneity will confuse sample-specific variation around a subcommunity with the presence of additional subcommunities. This results in difficulty estimating \(K\) and inference sensitive to \(K\); indeed, LDA encounters both of these difficulties (Fukuyama et al., 2022).

To verify this intuition, we generated data from a known LTN-LDA model, which induces cross-sample heterogeneity. In particular, we simulated \(D = 50\) samples, and \(N_d = 10,000\) reads per sample; we set \(\alpha = 1, \mu = 0, \Lambda = I, a_1 = 10^4, a_2 = b = 10,\) and \((K, C) = (4, 5)\). The underlying phylogenetic tree is presented in Supporting Information S5. There are \(V = 49\) ASVs. We then contrasted LDA and LTN-LDA by running Gibbs samplers on the data generated above with \(K \in \{4, 5, 7, 10\}\) and \(C = 5\). In the left part of Figure 3, we plot the posterior means of the subcommunity abundances \(\phi_d\) for both LDA and LTN-LDA. We corrected for label switching and estimated the \(\phi_d\) as in Griffiths and Steyvers (2004).

With \(K\) set to truth, LDA performs comparably to LTN-LDA in estimating the true values of \(\phi_d\); however, as we increase \(K\), the inference provided by LDA worsens. While it still recovers the abundances for subcommunities 1 and 2, it does a worse job at recovering subcommunities 3 and 4. Moreover, LDA detects the presence of additional subcommunities, which do not exist in the true generative model. LTN-LDA, in contrast, is remarkably stable when \(K\) is overspecified. No matter the modeled value of \(K\), it detects the four true subcommunities with approximately the same abundances while estimating that additional subcommunities have little abundance. For \(K = 10\), we plotted the subcommunity compositions on the right part of Figure 3. (This figure appears in color in the electronic version of this article, and any mention of color refers to that version.) For LTN-LDA, distributions for the \(\beta_d\) are in blue and the \(\beta_k\) are in red; the LDA \(\beta_d\) distributions are in black. LTN-LDA finds moderate levels of cross-sample heterogeneity in subcommunity 2, and a high levels in samples 3 and 4.
These figures imply that LDA is able to recover the subcommunity abundances only for those subcommunities with low cross-sample heterogeneity. LDA fails to recover the subcommunity abundances for those subcommunities with high cross-sample heterogeneity, mistaking heterogeneity for additional subcommunities. In effect, LDA splits true heterogeneous subcommunities into many smaller subcommunities with no heterogeneity and ASVs, which ought to belong in the same subcommunity that are separated. LTN-LDA, on the other hand, provides stable and accurate inference as the modeled $K$ increases. This confirms our intuition about the behavior of LDA in the presence of cross-sample heterogeneity.

### 3.2 Predictive scoring as a device for choosing tuning parameters

While incorporating cross-sample heterogeneity enhances the robustness of LTN-LDA to overspecifying the number of subcommunities, it is still useful to have a generally applicable strategy for setting the tuning parameters for LTN-LDA: $K$ and $C$. One option is to use out-of-sample predictive performance to identify suitable choices of the tuning parameters. A popular performance measure for MM models is perplexity (Wallach et al., 2009): a transform of out-of-sample predictive likelihood such that lower perplexity is preferred.

We thus implement the simple strategy of computing the average out-of-sample perplexity score for different choices of $(K, C)$ and examine whether that can lead to a practical way of choosing these parameters. We will also examine whether this strategy could be adopted for models without cross-sample heterogeneity, namely, LDA, to alleviate their limitations. We follow the procedure in Section 5.1 of Wallach et al. (2009) for computing the perplexity for LDA, and generalize that strategy to LTN-LDA.

**Figure 4A** shows the average perplexity scores for LDA and LTN-LDA as a function of $K$ and $C$. There are three main observations: (1) LTN-LDA significantly outperforms LDA for $K$ near truth, (2) the perplexity curve for LTN-LDA decreases until it stabilizes at the true value of $K$, and (3) the perplexity curve for LDA continues to decrease as the modeled $K$ is increased past its true value.
value. The main reason for the difference is that LDA interprets the presence of cross-sample heterogeneity as extra subcommunities and so finds as many subcommunities as are modeled. While this improves out-of-sample predictive performance, it does not improve inference on the underlying truth. Thus, using perplexity to select the modeled number of subcommunities for LDA is a poor method if there is significant cross-sample heterogeneity.

LTN-LDA is more robust and parsimonious in its representation of the data because it incorporates cross-sample heterogeneity in subcommunity compositions.

Fixing $K$ to truth, we computed average perplexity for LTN-LDA as we varied $C$ in Figure 4B. The perplexity curve decreases until it stabilizes at the true value of $C$. In addition to perplexity, we also computed the $L_2$ distances between the posterior mean estimates and the true values for the $\phi_d, \beta_{d,k}$, and $\beta_k$ distributions (Figure 4C). Unlike the perplexity curves, the $L_2$ distances are lowest around $C = 5$ and increase as $C$ increases. Thus, if the modeled value of $C$ is increased too far above truth, inference becomes unreliable.

The above results suggest a simple two-stage strategy for choosing $(K, C)$ using perplexity. First, let $(K, C)$ vary jointly on a grid and use cross-validation to compute the average perplexity, giving $K$ perplexity curves over $C$. Set $C$ to be the inflection point in these curves. Second, vary $K$ and set the value of $K$ to be the inflection point of the resulting perplexity curve. Note that this strategy may fail for LDA: As our numerical examples show below, due to the lack of cross-sample heterogeneity in LDA, the perplexity score generally continues to improve as one increases the number of subcommunities beyond truth. This in turn leads to misleading inference on subcommunity abundance and composition.

4 | EVALUATION ON A MICROBIOME STUDY

We apply LTN-LDA to identify subcommunity dynamics in the data set of Dethlefsen & Relman (2011), which has been previously investigated by Sankaran & Holmes (2019) using LDA. The data include gut microbiome samples of three patients who were administered two 5-day courses of ciprofloxacin over a 10-month span. We focus on the 54 samples from patient F, each consisting of
approximately 10,000 reads. Ciprofloxin was administered during samples 12-23 and 41-51. There are 2852 unique ASVs in the data set; we merged ASVs into taxa at the finest known level and pruned all taxa, which did not total at least 100 sequencing reads. This left 44 taxa comprising 99.86% of the original counts. The resulting phylogenetic tree is included in Supporting Information S7.

We implemented the strategy outlined above to choose tuning parameters. In particular, we implement a four-fold cross-validation letting $K$ vary in $\{2, 3, \ldots, 8\}$ and $C$ in $\{1, 2, \ldots, 21\}$. The resulting $K$ perplexity curves over $C$ are presented in Figure 4D. The inflection point in the curve appears at $C = 8$. Setting $C = 8$ and varying $K$ gives the results in Figure 4E; for comparison, we also applied LDA to the data over varying $K$. LTN-LDA has strictly lower perplexity than LDA, indicating that there are significant levels of cross-sample heterogeneity in the data set. Moreover, LTN-LDA experiences a noticeable inflection point (near $K = 5$) in contrast to LDA whose perplexity decays slowly.

We now present more detailed analysis for LTN-LDA and LDA with $C = 8$. For $K \in \{3, 4, 7\}$, we plotted the subcommunity abundance on the left side of Figure 5, after manually correcting for label switching. The gray regions indicate periods of ciprofloxin treatment. The subcommunities found by LTN-LDA are remarkably stable as $K$ changes. Subcommunities 1, 2, and 3 have almost the exact same abundance, and additional subcommunities have minimal abundance. LDA, however, finds as many subcommunities as are modeled: It will split a heterogeneous subcommunity into multiple subcommunities with no heterogeneity. For $K = 7$, we plotted the ASV-subcommunity distributions on the right side of Figure 5. Distributions for the $\beta_{d,k}$ are in blue, the $\beta_k$ in red, and the LDA distributions in black. The three most prevalent ASVs in each subcommunity are presented in Figure 6 for LDA and LTN-LDA. These demonstrate that LTN-LDA finds significant levels of cross-sample heterogeneity and subcommunities with meaningfully different compositions than LDA.

LTN-LDA thus provides two major advantages. First, LTN-LDA is more robust with respect to modeling differing numbers of subcommunities than LDA. This is similar to our simulations and indicates that LTN-LDA better accounts for the cross-sample heterogeneity in the data than does LDA. Moreover, the three subcommunities found by LTN-LDA are biologically interpretable. The first subcommunity is composed mostly of Lachnospiraceae
and Ruminococcaceae and displays significant levels of cross-sample heterogeneity, indicating that LTN-LDA has found these two ASVs can substitute for each other. Haak et al. (2018) found this phenomena in humans undergoing ciprofloxin treatment. LTN-LDA can thus learn when two ASVs substitute for each other across samples from the data, with no prior knowledge. The second subcommunity, composed mainly of Bacteroides, increases in abundance during the antibiotic treatments. Studies in mice (Zhu et al., 2020) and humans (Stewardson et al., 2015) indicate that the abundance of Bacteroides increases during ciprofloxin treatment. The third subcommunity has a small spike in abundance only on the first day of the second antibiotic course, and is composed mostly of Dialister and Veillonella. Ciprofloxin has been shown to be effective against Dialister (Morio et al., 2007), which may explain the decrease in this subcommunity after treatment began.

5 | DISCUSSION

We have proposed a novel MM model, which seeks to appropriately incorporate cross-sample heterogeneity in subcommunity compositions: a characteristic of the data prevalent in most microbiome studies. By incorporating the LTN model for the sample-specific compositions of each subcommunity, we explicitly allow the composition of subcommunities to vary across samples. We have shown that incorporating cross-sample heterogeneity into MM models can lead to substantially improved inference over models, which assume zero cross-sample heterogeneity. LTN-LDA is substantially more robust than LDA with respect to overspecifying $K$ and significantly outperforms LDA in terms of predictive performance. Moreover, perplexity can be a useful device to set the tuning parameters for LTN-LDA but not for LDA. Posterior computation on LTN-LDA can proceed through collapsed blocked Gibbs-sampling with the assistance of PG augmentation, and as such implementation for LTN-LDA is convenient. Moreover, LTN-LDA is a fully Bayesian model and the Gibbs sampler allows for posterior uncertainty quantification.

In comparison to LDA, LTN-LDA incorporates two new features: the tree structure and the random effects allowing cross-sample heterogeneity. The tree structures provide guidance on how to parsimoniously model the random effects without causing nonidentifiability. We carried out an additional numerical experiment that shows that using the tree structure as a way to parameterize the model without adding random effects does not lead to improved inference. For a more detailed discussion, see Supporting Information S8. While LTN-LDA relies on the tree structure to incorporate random effects, we note there are several alternative approaches to incorporating random effects in microbiome compositions (Grantham et al., 2017; Ren et al., 2020; Zhang & Lin, 2019). In principle, it is possible to incorporate random effects without a tree structure in the MM model.
Like other unsupervised learning methods, LTN-LDA is unable to differentiate between different scenarios giving rise to the same sampling distributions. That is, LTN-LDA, or any other models for that matter, cannot distinguish between multiple subcommunities and a single overdispersed one if the two give rise to the same sampling distributions. Domain knowledge is necessary to identify such possibilities; traditionally, there are two strategies to incorporate such domain knowledge. The first is through modeling assumptions, such as modeling how large the single-subcommunity dispersion is through the hyperpriors on the $\tau_k^1$. The other strategy is using a decision-theoretic formulation that introduces certain loss functions to carry out post hoc merging of the identified topics.

Moreover, we believe that the idea of incorporating cross-sample heterogeneity in MM models could be valuable beyond the context of microbiome compositions. In topic models, for example, one might expect different authors to write on the same topic using different vocabulary. LTN-LDA has the potential to be applicable to these other contexts as well, though the immediate challenge is finding an appropriate tree structure.

ACKNOWLEDGMENTS
LM’s research is partly supported by NIGMS grant R01-GM135440 and NSF grant DMS-1749789 Part of the research was completed when PL was supported by an NSF-sponsored IBIEM fellowship at Duke University.

OPEN RESEARCH BADGES
This article has earned Open Data and Open Materials badges. Data and materials are available at https://www.re3data.org/.

DATA AVAILABILITY STATEMENT
The original data that support the findings of this paper are openly available in Proceedings of the National Academy of the Sciences at https://doi.org/10.1073/pnas.1000087107 (Dethlefsen & Relman, 2011). The version of the data that support the findings of this paper are openly available in Biostatistics at https://doi.org/10.1093/biostatistics/kxy018 (Sankaran & Holmes, 2019).

ORCID
Li Ma https://orcid.org/0000-0002-0159-3296

REFERENCES
Aitchison, J. (1982) The statistical analysis of compositional data. Journal of the Royal Statistical Society, Series B (Methdological), 44, 139–177.

Andrzejewski, D., Zhu, X. & Craven, M. (2009) Incorporating domain knowledge into topic modeling via Dirichlet forest priors. Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, pp. 25–32.

Blei, D.M., Ng, A.Y. & Jordan, M.I. (2003) Latent Dirichlet allocation. Journal of Machine Learning Research, 2, 993–1022.

Callahan, B.J., McMurdie, P.J. & Holmes, S.P. (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal, 11, 2639–2643.

Deek, R.A. & Li, H. (2019) A zero-inflated latent Dirichlet allocation model for microbiome studies. Frontiers in Genetics, 11, 599–644.

Del Chierico, F., Vernocci, P., Dallapiccola, B. & Putignani, L. (2014) Mediterranean diet and health: food effects on gut microbiota and disease control. International Journal of Molecular Sciences, 15, 11678–11699.

Dennis III, S.Y. (1991) On the hyper-Dirichlet type I and hyper-Liouville distributions. Communications in Science - Theory and Methods, 20, 4069–4081.

Dethlefsen, L. & 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 the Sciences of the United States of America, 108, 4554–4561.

Doshi-Velez, F., Wallace, B.C. & Adams, R. (2015) Graph-sparse LDA: a topic model with structured sparsity. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI 2015, pp. 2575–2581.

Fukuyama, J., Sankaran, K. & Symul, L. (2022) Multiscale analysis of count data through topic alignment. Biostatistics, kxao18. https://doi.org/10.1093/biostatistics/kxao18.

Glynn, C., Tokdar, S.T. & Banks, D.L. (2019) Bayesian analysis of dynamic linear topic models. Bayesian Analysis, 14, 1–53.

Grantham, N., Guan, Y., Reich, B., Borer, E. & Gross, K. (2017) MIMIX: a Bayesian mixed-effects model for microbiome data from designed experiments. Journal of American Statistical Association, 115, 599–609.

Griffiths, T.L. & Steyvers, M. (2004) Finding scientific topics. Proceedings of the National Academy of Sciences, 101, 5228–5235.

Haak, B.W., Lankelma, J.M., Hugenbolt, F., Belzer, C., de Vos, W.M. & Wiersinga, W.J. (2018) Long-term impact of oral vancomycin, ciprofloxin and metronidazole on the gut microbiota in healthy humans. Journal of Antimicrobial Chemotherapy, 74, 782–786.

Haffari, G. & Teh, Y.W. (2009) Hierarchical Dirichlet trees for information retrieval. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL 2009, pp. 173–181.

Holmes, I., Harris, K. & Quince, C. (2012) Dirichlet multinomial mixtures: generative models for microbial metagenomics. PloS One, 7, e20126.

Jeganathan, P. & Holmes, S.P. (2021) A statistical perspective on the challenges in molecular microbial biology. Journal of Agricultural, Biological, and Environmental Statistics, 26, 131–160.

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

Mao, J., Chen, Y. & Ma, L. (2020) Bayesian graphical compositional regression for microbiome data. Journal of the American Statistical Association, 115, 610–624.
Morio, F., Jean-Pierre, H., Dubreuil, L., Jumas-Bilak, E., Calvet, L., Mercier, G., Devine, R. & Marchandin, H. (2007) Antimicrobial susceptibilities and clinical sources of Dialister species. *Antimicrobial Agents and Chemotherapy*, 51, 4498–4501.

Nason, G.P. (2008) *Wavelet methods in statistics with R*, 1st edition. Springer Publishing Company, Incorporated.

Nigam, K., McCallum, A.K., Thrun, S. & Mitchell, T. (2000) Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39, 103–134.

Polson, N.G., Scott, J.G. & Windle, J. (2013) Bayesian inference for logistic models using Pólya-gamma latent variables. *Journal of the American Statistical Association*, 108, 1339–1349.

Pritchard, J.K., Stephens, M. & Donnelly, P. (2000) Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Genetics*, 155, 945–959.

Ren, B., Bacallado, S., Favaro, S., Vatanen, T., Huttenhower, C. & Trippa, L. (2020) Bayesian mixed effects models for zero-inflated compositions in microbiome data analysis. *The Annals of Applied Statistics*, 14, 494–517.

Sankaran, K. & Holmes, S.P. (2019) Latent variable modelling for the microbiome. *Biostatistics*, 20, 599–614.

Shafilee, M., Dunn, K.A., Boon, E., MacDonal, S.M., Walsh, D.A., Gu, H. et al. (2015) BioMiCo: a supervised Bayesian model for inference of microbial community structure. *Microbiome*, 3.

Siezen, R.J. & Kleerebezem, M. (2011) The human gut microbiome: are we our enterotypes? *Microbial Biotechnology*, 4, 550–553.

Stewardson, A.J., Gaia, N., Francois, P., Malhotra-Kumar, S., Delemont, C., de Tejada, B.M. et al. (2015) Collateral damage from oral ciprofloxacin versus nitrofurantoin in outpatients with urinary tract infections: a culture-free analysis of gut microbiota. *Clinical Microbiology and Infection*, 21, 344.e1–344.e11.

Tam, Y. & Schultz, T. (2007) Correlated latent semantic model for unsupervised LM adaptation. 2007 *IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP 2007, pp. 41–44.

Tang, Y., Ma, L. & Nicolae, D.L. (2018) A phylogenetic scan test on a Dirichlet-tree multinomial model for microbiome data. *The Annals of Applied Statistics*, 12, 1–26.

Wallach, H.M., Murray, I., Salakhutdinov, R. & Mimno, D. (2009) Evaluation methods for topic models. *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML 2009, pp. 1105–1112.

Wang, T. & Zhao, H. (2017) A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. *Biometrics*, 73, 792–801.

Wang, Z., Mao, J. & Ma, L. (2021) Microbiome compositional analysis with logistic-tree normal models. Available at: https://arxiv.org/abs/2106.15051.

Weber, J.L. & Myers, E.W. (1997) Human whole-genome shotgun sequencing. *Genome Research*, 7, 401–409.

Zhang, J. & Lin, W. (2019) Scalable estimation and regularization for the logistic normal multinomial model. *Biometrics*, 75, 1098–1108.

Zhu, S., Li, H., Liang, J., Lv, C., Zhao, K, Niu, M. et al. (2020) Assessment of oral ciprofloxacin impaired gut barrier integrity on gut bacteria in mice. *International Immunopharmacology*, 83, 106460.

**SUPPORTING INFORMATION**

Web Appendices and Figures referenced in Sections 2.3, 2.4, 3.1, 3.2, 4, and 5 are available with this paper at the Biometrics website on Wiley Online Library. Reproducible code and data for this paper are posted online with this paper. They are also available at https://github.com/PatrickLeBlanc/ReproduceLTNLDAPaper. R code for implementing the LTN-LDA model is available in the LTNLDA package: https://github.com/PatrickLeBlanc/LTNLDA.

**How to cite this article:** LeBlanc, P. & Ma, L. (2023) Microbiome subcommunity learning with logistic-tree normal latent Dirichlet allocation. *Biometrics*, 79, 2321–2332. https://doi.org/10.1111/biom.13772