9-7-2023

Clustering microbiome data using mixtures of logistic normal multinomial models

Yuan Fang
Binghamton University–SUNY

Sanjeena Subedi

scientific reports

Check for updates

OPEN

Clustering microbiome data using mixtures of logistic normal multinomial models

Yuan Fang[1] & Sanjeena Subedi[2]✉

Discrete data such as counts of microbiome taxa resulting from next-generation sequencing are routinely encountered in bioinformatics. Taxa count data in microbiome studies are typically high-dimensional, over-dispersed, and can only reveal relative abundance therefore being treated as compositional. Analyzing compositional data presents many challenges because they are restricted to a simplex. In a logistic normal multinomial model, the relative abundance is mapped from a simplex to a latent variable that exists on the real Euclidean space using the additive log-ratio transformation. While a logistic normal multinomial approach brings flexibility for modeling the data, it comes with a heavy computational cost as the parameter estimation typically relies on Bayesian techniques. In this paper, we develop a novel mixture of logistic normal multinomial models for clustering microbiome data. Additionally, we utilize an efficient framework for parameter estimation using variational Gaussian approximations (VGA). Adopting a variational Gaussian approximation for the posterior of the latent variable reduces the computational overhead substantially. The proposed method is illustrated on simulated and real datasets.

The human microbiome comprises complex communities of microorganisms including but not limited to bacteria, fungi, and viruses, that inhabit in and on a human body[1,2]. It is estimated that there are approximately 10^{14} microbial cells associated with the human body, which is around 10 times the number of human cells[3,4]. The human microbiome plays a significant role in human health and disease status. There is evidence indicating that microbial dysbiosis may lead to diseases such as cardiovascular diseases[5], diabetes[6], inflammatory bowel disease[7], obesity[8], and many others. Next-generation sequencing techniques, such as the 16S ribosomal RNA (rRNA) amplicon sequencing or shotgun metagenomics sequencing, provide an effective way for quantification and comparison of the bacterial composition, including types and abundance of different bacteria within biological samples[9–12]. In 16S rRNA sequencing, the 16S rRNA, which is ubiquitous in all bacterial organisms but also has distinct variable regions that can be used to discriminate between different bacteria is first PCR-amplified and then sequenced[10]. Shotgun sequencing on the other hand is an untargeted sequencing of all microbial genomes in a sample[13]. In either case, short reads are preprocessed through steps of quality control and filtering steps. The processed raw sequence reads are then clustered into operational taxonomic units (OTUs) at a certain similarity level[14] where each OTU is characterized by a representative DNA sequence that could be assigned to a taxonomic lineage by comparing to a known database[2]. Resulting read counts at different taxonomic levels for n samples over $K+1$ taxa are stored as a $n \times (K+1)$ matrix $\mathbf{W}$, with the entry $W[i, k]$ representing the counts recorded for the k^{th} taxon in the i^{th} sample.

Statistical analysis of microbiome data is complicated. The microbiome count data can only reveal relative abundance, i.e., the abundance for each taxa is constrained by the total sum of the microbes in that particular sample and the total sum of microbes could vary among the samples depending on the sequencing depth. Different individuals could share various communities of microorganisms, with only a few major ones in common, and even for one person, the microbial composition could be totally different in different body sites. The heterogeneity of the microbiome samples also leads to over-dispersion. See Hamady and Knight[15] for a detailed review of challenges related to analyzing microbiome data. Standard multivariate analysis usually fails to capture these properties of the microbiome data. Different models have been proposed for the microbiome counts in the literature that captures one or more of the above intrinsic characteristics such as the negative binomial model[16], zero-inflated negative binomial model[17], zero-inflated Poisson model[18,19], Dirichlet-multinomial model[20–22], and

[1]School of Pharmacy and Pharmaceutical Sciences, Binghamton University, State University of New York, 4400 Vestal Parkway East, Binghamton, NY 13902, USA. [2]School of Mathematics and Statistics, 4302 Herzberg Laboratories, Carleton University, 1125 Colonel By Drive, Ottawa, ON K1S 5B6, Canada. ✉email: sanjeena.dang@carleton.ca

the logistic normal multinomial model[23]. While modelling such count data, a negative binomial (NB) model can allow for the variance to be larger than the mean using a dispersion parameter, thus handling over-dispersion better than a simple Poison model. The zero-inflated negative binomial (ZINB) and zero-inflated Poisson (ZIP) have been proposed to account for the excessive number of zeros[18]. Xu et al.[24] provides a comparison among the zero-inflated models. However, the NB and ZINB models ignore the compositional nature of these microbial counts. Chen and Li[21], Holmes et al.[20], Wadsworth et al.[25], and Subedi et al.[22] utilized the Dirichlet-multinomial model for microbial counts that takes into account the compositional nature of these data. Alternately, Xia et al.[23] employed the logistic normal multinomial model, mapping the relative abundance from a simplex to a latent variable that exists on the real Euclidian space using the additive log-ratio transformation. Cao et al.[26] exploited a Poisson-multinomial model and performed a multi-sample estimation of microbial composition in positive simplex space from a high-dimensional sparse count table. Caporaso et al.[27] quantified variations of microbial composition across time by projecting the dynamics using low-dimensional embedding. Äijö et al.[12] proposed a temporal probabilistic model for the microbiome composition using a hierarchical multinomial model. Silverman et al.[28] also developed a dynamic linear model based on the logistic normal multinomial model to study the artificial human guts microbiome.

Clustering microbiome samples into groups that share similar microbial compositional patterns is of great interest[20]. Clustering algorithms are usually categorized into hierarchical clustering and distance-based clustering. Hierarchical clustering has been applied for clustering microbiome data, yet it requires the choice of a cut-off threshold, according to which samples can be divided into groups[20]. On the other hand, k—means clustering, a distance-based method, might not be appropriate for microbiome compositions because it is typically used for continuous data and obtains spherical clusters. Hence, model-based clustering approaches that utilize a finite mixture model have been widely used in the last decade to cluster microbiome data[20,22]. A finite mixture model assumes that the population consists of a finite mixture of subpopulations (or clusters), each represented by a known distribution[29–32]. Due to the flexibility in choosing component distributions to model different types of data, several mixture models based on discrete distributions have been developed to study count data, especially, for gene expression data. Rau et al.[33] proposed a clustering approach for RNA-seq data using mixtures of univariate Poisson distributions; Papastamoulis et al.[34] proposed a mixture of Poisson regression models; Si et al.[35] studied model-based clustering for RNA-seq data using a mixture of negative binomial (NB) distributions; Silva et al.[36] proposed a multivariate Poisson-log normal mixture model for clustering gene expression data. However, due to the compositional nature of microbiome data, none of the above discrete mixture models can be employed directly for clustering microbiome data. Holmes et al.[20] adopted the Dirichlet-multinomial (DM) model, where the underlying compositions are modeled as a Dirichlet prior to a multinomial distribution that describes the taxa counts, and proposed a mixture of DM models to cluster samples.

In this paper, we develop a model-based clustering approach using the logistic normal multinomial model proposed by Xia et al[23] to cluster microbiome data. In the logistic normal multinomial model, the observed counts are modeled using a multinomial distribution, and the relative abundance is regarded as a random vector on a simplex, which is further mapped to a latent variable that exists on the real Euclidean space through an additive log-ratio transformation. While this approach captures the additional variability compared to a multinomial model, it does not possess a closed form expression of the log-likelihood functions and of the posterior distributions of the latent variables. Therefore, the expected complete-data log-likelihoods needed in the E-step of a traditional EM algorithm are usually intractable. In such a scenario, one commonly used approach is a variant of the EM algorithm that relies on Bayesian techniques using Markov chain Monte Carlo (MCMC); however, this would typically bring in high computational cost. Here, we develop a variant of the EM algorithm, here on referred to as a variational EM algorithm for parameter estimation that utilizes variational Gaussian approximations (VGA). In Variational Gaussian approximations (VGA)[37], a complex posterior distribution is approximated using computationally convenient Gaussian densities by minimizing the Kullback-Leibler (KL) divergence between the true and the approximating densities[38,39]. Adopting a variational Gaussian approximation delivers accurate approximations of the complex posterior while reducing computational overhead substantially. Hence, this approach has become extremely popular in many different fields of machine learning[37,38,40–43].

The contribution of the paper is two folds - first, we develop a computationally efficient framework for parameter estimation for a logistic normal multinomial model through the use of variational Gaussian approximations and second, we utilize this framework to develop a model-based clustering framework for clustering microbiome data. Through simulations and applications to microbiome data, the utilities of the proposed approach are illustrated. The paper is structured as follows: First two subsections in the Methods section describe the logistic normal multinomial model for microbiome count data and detail the variational Gaussian approximations. The third and fourth subsections in the Methods section provide a mixture model framework based on the model described above together with a variational EM algorithm for parameter estimation. In the Results section, clustering results are illustrated by applying the proposed algorithm to both simulated and real data. Finally, a discussion on the advantages and limitations along with some future directions are provided in the Discussion section.

Methods

The logistic normal multinomial model for microbiome compositional data. Suppose we have $K+1$ bacterial taxa for a sample denoted as a random vector $\mathbf{W} = (W_1, \ldots, W_{K+1})^\top$. Here, the taxa could represent any level of the bacterial phylogeny such as OTU, species, genus, phylum, etc. Due to the fact that taxa count from 16S sequencing can only reveal relative abundance, let's suppose there is a vector $\boldsymbol{\Theta} = (\Theta, \ldots, \Theta_{K+1})$ such that $\sum_{k=1}^{K+1} \Theta_k = 1$, which represents the underlying composition of the bacterial taxa. Then, the microbial taxa count $\mathbf{W}$ can be modeled as a multinomial random variable with the following conditional density function:

$$p(\mathbf{w}|\boldsymbol{\Theta}) \propto \prod_{k=1}^{K+1} (\Theta_k)^{w_k}.$$

Several models have been proposed in the literature that capture the relative abundance nature of microbiome data and analyze the compositional data[20,23]. Here we use the model by Xia et al.[23] that utilizes an additive log-ratio (ALR) transformation $\phi(\boldsymbol{\Theta})$ proposed by Aitchison[44] such that:

$$\mathbf{Y} = \phi(\boldsymbol{\Theta}) = \left(\log\left(\frac{\Theta_1}{\Theta_{K+1}}\right), \ldots, \log\left(\frac{\Theta_K}{\Theta_{K+1}}\right)\right)^\top. \tag{1}$$

This transformation ϕ maps the vector $\boldsymbol{\Theta}$ from a K-dimensional simplex to the K-dimensional real space $\mathbb{R}^K$. The prior distribution for $\mathbf{Y}$ is assumed to be a multivariate normal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ with the density function

$$p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\}.$$

As this additive log-ratio transformation is a one-to-one map, the inverse operator of ϕ exists and is given by

$$\boldsymbol{\Theta} = \phi^{-1}(\mathbf{Y}) = \begin{cases} \dfrac{\exp(Y_k)}{1+\sum_{k=1}^{K}\exp(Y_k)} & k = 1, \ldots, K \\ \dfrac{1}{1+\sum_{k=1}^{K}\exp(Y_k)} & k = K+1 \end{cases}.$$

Hence, the joint density of $\mathbf{W}$ and $\mathbf{Y}$ up to a constant is as follows:

$$p(\mathbf{w}, \mathbf{y}) \propto p\big(\mathbf{w}|\phi^{-1}(\mathbf{y})\big)p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{k=1}^{K+1}\left(\phi^{-1}(\mathbf{y})_k\right)^{w_k} \times |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\} \quad .$$

A variational Gaussian lower bound. For the microbiome data, only the count vector $\mathbf{W}$ is observed while the latent variable $\mathbf{Y}$ is unobserved. The marginal density of $\mathbf{W}$ can be written as

$$p(\mathbf{w}) = \int_{\mathbb{R}^K} p(\mathbf{w}, \mathbf{y}) d\mathbf{y} \propto \int_{\mathbb{R}^K} \prod_{k=1}^{K+1}\left(\phi^{-1}(\mathbf{y})_k\right)^{w_k} \times |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\} d\mathbf{y}.$$

Note that this marginal distribution of $\mathbf{W}$ involves multiple integrals and cannot be further simplified. Here, in the presence of missing data, an expectation-maximization (EM) algorithm[45] or some variant of it is typically utilized for parameter estimation. An EM algorithm comprises two steps: an E-step in which the expected value of the complete data (i.e. observed and missing data) log-likelihood is computed given the observed data and current parameter estimate and an M-step in which the complete data log-likelihood is maximized. These steps are repeated until convergence to obtain the maximum likelihood estimate of the parameters. To compute the expected value of the complete data log-likelihood, $\mathbb{E}(\mathbf{Y} \mid \mathbf{w})$ and $\mathbb{E}(\mathbf{Y}\mathbf{Y}^T \mid \mathbf{w})$ needs to be computed for which we need $p(\mathbf{y}|\mathbf{w})$. Mathematically,

$$p(\mathbf{y}|\mathbf{w}) = \frac{p(\mathbf{w}, \mathbf{y})}{p(\mathbf{w})} = \frac{\prod_{k=1}^{K+1}\left(\phi^{-1}(\mathbf{y})_k\right)^{w_k} \times |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\}}{\int_{\mathbb{R}^K} \prod_{k=1}^{K+1}\left(\phi^{-1}(\mathbf{y})_k\right)^{w_k} \times |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\} d\mathbf{y}}.$$

However, the denominator involves multiple integrals and cannot be further simplified. One could employ a Markov chain Monte Carlo (MCMC) approach to explore the posterior state space; however, these methods are typically computationally expensive, especially for high-dimensional problems. Here, we propose the use of variational Gaussian approximation (VGA)[37] for parameter estimation. A VGA aims to find an optimal and tractable approximation that has a Gaussian parametric form to approximate the true complex posterior by minimizing the Kullback-Leibler divergence between the true and the approximating densities. It has been successfully used in many practical applications to overcome this challenge[38–42,46]. In order to utilize VGA, we define a new latent variable $\boldsymbol{\eta}$ by transforming $\mathbf{Y}$ such that

$$\boldsymbol{\eta} = B\mathbf{Y}, \quad \text{where } B = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \\ 0 & 0 & \ldots & 0 \end{pmatrix}, \tag{2}$$

is a $(K+1) \times K$ matrix which takes the form as an identity matrix attached by a row of K zeros. Given that $\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the new latent variable $\boldsymbol{\eta} \sim \mathrm{N}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ where

$$\tilde{\boldsymbol{\mu}} = B\boldsymbol{\mu} = (\boldsymbol{\mu}, 0)^\top; \quad \tilde{\boldsymbol{\Sigma}} = B\boldsymbol{\Sigma}B^\top = \left(\begin{array}{c|c} \boldsymbol{\Sigma} & \mathbf{0}_{K\times 1} \\ \hline \mathbf{0}_{1\times K} & 0 \end{array}\right). \tag{3}$$

Then, the underlying composition variable $\boldsymbol{\Theta}$ can be written as a function of $\boldsymbol{\eta}$:

$$\boldsymbol{\Theta} = \tilde{\phi}^{-1}(\boldsymbol{\eta}) = \frac{\exp \eta_k}{\sum_{k=1}^{K+1} \exp \eta_k} k = 1 \ldots, K+1. \tag{4}$$

Suppose we have an approximating density $q(\boldsymbol{\eta})$, then the marginal log density of $\mathbf{W}$ can be written as:

$$\begin{aligned} \log p(\mathbf{w}) &= \int \log p(\mathbf{w})\, q(\boldsymbol{\eta})\, d\boldsymbol{\eta} = \int \log \frac{p(\mathbf{w}, \boldsymbol{\eta})/q(\boldsymbol{\eta})}{p(\boldsymbol{\eta} \mid \mathbf{w})/q(\boldsymbol{\eta})}\, q(\boldsymbol{\eta})\, d\boldsymbol{\eta} \\ &= \int \left[\log\, p(\mathbf{w}, \boldsymbol{\eta}) - \log q(\boldsymbol{\eta})\right] q(\boldsymbol{\eta})\, d\boldsymbol{\eta} + \int \log \frac{q(\boldsymbol{\eta})}{p(\boldsymbol{\eta}|\mathbf{w})}\, q(\boldsymbol{\eta})\, d\boldsymbol{\eta} \\ &= F(q(\boldsymbol{\eta}), \mathbf{w}) + D_{KL}(q||p), \end{aligned}$$

where the first part $F(q(\boldsymbol{\eta}), \mathbf{w}) = \int q(\boldsymbol{\eta}) \log \frac{p(\mathbf{w}, \boldsymbol{\eta})}{q(\boldsymbol{\eta})} d\boldsymbol{\eta}$ is called the evidence lower bound (ELBO)[37] and the second part $D_{KL}(q||p) = \int \log \frac{q(\boldsymbol{\eta})}{p(\boldsymbol{\eta}|\mathbf{w})}\, q(\boldsymbol{\eta})\, d\boldsymbol{\eta}$ is the Kullback-Leibler divergence from $p(\boldsymbol{\eta}|\mathbf{w})$ to $q(\boldsymbol{\eta})$. Hence, minimizing the Kullback-Leibler divergence is equivalent to maximizing the following evidence lower bound (ELBO). In VGA, we assume $q(\boldsymbol{\eta})$ is a Gaussian distribution, such that

$$q(\boldsymbol{\eta}) = \mathrm{N}(\boldsymbol{\eta}|\mathbf{m}, V) \propto |V|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\eta} - \mathbf{m})^\top V^{-1}(\boldsymbol{\eta} - \mathbf{m})\right\}.$$

Given the fact that $q(\boldsymbol{\eta})$ is fully characterized by its mean vector and covariance matrix, the above lower bound is a function of the variational parameters $\mathbf{m}$ and V and we aim to find the optimal set of $(\mathbf{m}, V)$ such that it maximizes $F(q(\boldsymbol{\eta}, \mathbf{w}))$. $F(q(\boldsymbol{\eta}, \mathbf{w}))$ can be separated into three parts:

$$F(q(\boldsymbol{\eta}), \mathbf{w}) = F(\mathbf{m}, V) = -\int q(\boldsymbol{\eta}) \log q(\boldsymbol{\eta}) d\boldsymbol{\eta} + \int q(\boldsymbol{\eta}) \log p(\boldsymbol{\eta}) d\boldsymbol{\eta} + \int q(\boldsymbol{\eta}) \log p(\mathbf{w}|\boldsymbol{\eta}) d\boldsymbol{\eta}.$$

Up to a constant, the last integral, which is denoted as γ, in the above decomposition is given as follows:

$$\begin{aligned} \gamma &= \int q(\boldsymbol{\eta}) \log p(\mathbf{w}|\boldsymbol{\eta}) d\boldsymbol{\eta} = \mathbb{E}_{q(\boldsymbol{\eta}|\mathbf{m},V)}\left[\mathbf{w}^\top \boldsymbol{\eta} - \sum_{k=1}^{K+1} w_k \log\left(\sum_{k=1}^{K+1} \exp \eta_k\right)\right] \\ &= \mathbf{w}^\top \mathbf{m} - \left(\sum_{k=1}^{K+1} w_k\right) \mathbb{E}_{q(\boldsymbol{\eta}|\mathbf{m},V)}\left[\log\left(\sum_{k=1}^{K+1} \exp \eta_k\right)\right]. \end{aligned}$$

Similar to Blei and Lafferty[47], we use an upper bound for the expectation of log sum exponential term with a Taylor expansion,

$$\mathbb{E}_{q(\boldsymbol{\eta}|\mathbf{m},V)}\left[\log\left(\sum_{k=1}^{K+1} \exp \eta_k\right)\right] \leq \xi^{-1}\left\{\sum_{k=1}^{K+1} \mathbb{E}_{q(\boldsymbol{\eta}|\mathbf{m},V)}\left[\exp(\eta_k)\right]\right\} - 1 + \log(\xi),$$

where $\xi \in \mathbb{R}$ is introduced as a new variational parameter.

Here, we further assume that V is a diagonal matrix with the first K diagonal element of V as v_k^2 and the $K+1^{th}$ diagonal element is set to 0 such that

$$v_k^2 = \begin{cases} v_k^2, & k = 1, \ldots, K \\ 0, & k = K+1. \end{cases}$$

We also denote the $k-$th element of $\mathbf{m}$ as m_k such that

$$m_k = \begin{cases} m_k, & k = 1, \ldots, K \\ 0, & k = K+1. \end{cases}$$

Hence, the expectation

$$\mathbb{E}_{q(\boldsymbol{\eta}|\mathbf{m},V)}\left[\exp(\eta_k)\right] = \exp\left(m_k + \frac{v_k^2}{2}\right), \text{ for } k = 1, \ldots, K+1.$$

Based on this upper bound, we obtain a concave lower bound to γ and to the ELBO. The new concave variational Gaussian lower bound to the model evidence $\log p(\mathbf{w})$ is given as follows

$$\begin{aligned}\tilde{F}\left(\mathbf{m}, V, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}, \xi\right) =& \mathbf{w}^\top \mathbf{m} - \left(\sum_{k=1}^{K+1} w_k\right)\left\{\xi^{-1}\left[\sum_{k=1}^{K+1}\exp\left(m_k + \frac{v_k^2}{2}\right)\right] - 1 + \log(\xi)\right\} \\ & - \frac{1}{2}\log|B^\top \tilde{\boldsymbol{\Sigma}} B| - \frac{1}{2}(\mathbf{m}-\tilde{\boldsymbol{\mu}})^\top \tilde{\boldsymbol{\Sigma}}^*(\mathbf{m}-\tilde{\boldsymbol{\mu}}) - \frac{1}{2}\mathrm{Tr}(\tilde{\boldsymbol{\Sigma}}^* V) + \frac{1}{2}\sum_{k=1}^{K}\log(v_k^2) + \frac{K}{2},\end{aligned} \tag{5}$$

where $\tilde{\boldsymbol{\Sigma}}^* = \left(\begin{array}{c|c} \boldsymbol{\Sigma}^{-1} & \mathbf{0}_{K\times 1} \\ \hline \mathbf{0}_{1\times K} & 0 \end{array}\right)$ is the generalized inverse of $\tilde{\boldsymbol{\Sigma}}$. Details on the derivation of this lower bound can be found in the Supplementary material Mathematical Detail section. Given fixed $\mathbf{w}$, $\tilde{\boldsymbol{\mu}}$, and $\tilde{\boldsymbol{\Sigma}}$, this lower bound only depends on the variational parameter set $(\mathbf{m}, V, \xi)$.

Maximization of the lower bound $\tilde{F}\left(\mathbf{m}, V, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}, \xi\right)$ with respect to ξ has a closed form solution and is given by

$$\hat{\xi} = \sum_{k=1}^{K+1}\exp\left(m_k + \frac{v_k^2}{2}\right). \tag{6}$$

However, maximization with respect to $\mathbf{m}$ and $v_k, k = 1, \ldots, K$ do not have analytical solutions. We use Newton's method to search for roots to the following derivatives:

$$\frac{\partial \tilde{F}}{\partial \mathbf{m}} = \mathbf{w} - \tilde{\boldsymbol{\Sigma}}^*(\mathbf{m}-\tilde{\boldsymbol{\mu}}) - \left(\sum_{k=1}^{K+1} w_k\right)\xi^{-1}\exp\left(\mathbf{m} + \frac{\mathbf{v}^2}{2}\right), \tag{7}$$

with $\mathbf{v}^2 = (v_1^2, \ldots, v_K^2, 0)$ denoting the diagonal element of V as a vector; and

$$\frac{\partial \tilde{F}}{\partial v_k} = v_k^{-1} - v_k\tilde{\boldsymbol{\Sigma}}^*_{k,k} - \left(\sum_{k=1}^{K+1} w_k\right)\xi^{-1}\exp\left(m_k + \frac{v_k^2}{2}\right)v_k. \tag{8}$$

Details can be found in the Supplementary material Mathematical Detail section.

Mixture of logistic normal multinomial models. Assume there are G subgroups in the population, with π_g denoting the mixing weight of the $g-$th component such that $\sum_{g=1}^{G}\pi_g = 1$. Then, a $G-$component finite mixtures logistic normal multinomial models can be written as

$$f(\mathbf{w} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G}\pi_g f_g(\mathbf{w} \mid \vartheta_g),$$

where $f_g(\mathbf{w} \mid \boldsymbol{\vartheta}_g)$ represents the density function of the observation $\mathbf{W} = \mathbf{w}$, given that $\mathbf{W}$ comes from the $g-$th component with parameters ϑ_g.

Provided n observed counts, $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)$ with a transformed underlying the composition $\mathbf{Y}_i, i = 1, \ldots, n$, the likelihood of a $G-$component finite mixture is given as

$$\mathscr{L}(\boldsymbol{\vartheta} \mid \mathbf{w}) = \prod_{i=1}^{n} f(\mathbf{w}_i \mid \boldsymbol{\vartheta}) = \prod_{i=1}^{n}\sum_{g=1}^{G}\pi_g f_g(\mathbf{w}_i \mid \vartheta_g) = \prod_{i=1}^{n}\sum_{g=1}^{G}\pi_g \int p(\mathbf{w}_i \mid \mathbf{y}_i)p(\mathbf{y}_i \mid \vartheta_g)d\mathbf{y}_i.$$

In clustering, the unobserved component membership is denoted by an indicator variable $z_{ig}, i = 1, \ldots, n, g = 1, \ldots, G$ that takes the form

$$z_{ig} = \begin{cases} 1 & \text{if the } i-th \text{ observation is from the } g-th \text{ group,} \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, conditional on z_{ig}, we have

$$\mathbf{Y}_i | z_{ig} = 1 \sim \mathrm{N}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g).$$

In order to utilize the variational approach for parameter estimation, we again define a new latent variable $\boldsymbol{\eta}$ such that $\boldsymbol{\eta} = B\mathbf{Y}$ and

$$\boldsymbol{\eta}_i \mid z_{ig} = 1 \sim \mathrm{N}(\tilde{\boldsymbol{\mu}}_g, \tilde{\boldsymbol{\Sigma}}_g), \quad \text{where} \quad \tilde{\boldsymbol{\mu}}_g = B\boldsymbol{\mu}_g = (\boldsymbol{\mu}_g, 0)^\top \quad \text{and} \quad \tilde{\boldsymbol{\Sigma}}_g = B\boldsymbol{\Sigma}_g B^\top = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_g & \mathbf{0}_{K\times 1} \\ \hline \mathbf{0}_{1\times K} & 0 \end{array}\right).$$

Therefore, the complete data (i.e., observed counts $\mathbf{W}$ and unobserved class label indicator variable) log-likelihood using the marginal density of $\mathbf{W}$ is

$$\ell = \log\left[\prod_{i=1}^{n}\prod_{g=1}^{G}\pi_g f_g(\mathbf{w}_i \mid \vartheta_g)\right]^{z_{ig}} = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left\{\log\pi_g + \log\left[\int p(\mathbf{w}_i|\boldsymbol{\eta}_i)p(\boldsymbol{\eta}_i|\tilde{\boldsymbol{\mu}}_g, \tilde{\boldsymbol{\Sigma}}_g)d\boldsymbol{\eta}_i\right]\right\}.$$

To perform variational inference on the mixture model, we substitute $\log\left[\int p(\mathbf{w}_i|\boldsymbol{\eta}_i)p(\boldsymbol{\eta}_i|\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)d\boldsymbol{\eta}_i\right]$ by the variational Gaussian lower bound $\tilde{F}\left(\mathbf{m}, V, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}, \xi\right)$ derived in Section "A variational Gaussian lower bound". Hence, the variational Gaussian lower bound of complete data log likelihood can be written as:

$$\begin{aligned}\tilde{\mathscr{L}} =& \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\log\pi_g + \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\mathbf{w}_i^\top \mathbf{m}_{ig} - \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left(\sum_{k=1}^{K+1} w_{ik}\right)\\ &\left\{\xi_i^{-1}\left[\sum_{k=1}^{K+1}\exp\left(m_{igk}+\frac{v_{igk}^2}{2}\right)\right] - 1 + \log(\xi_i)\right\}\\ &+\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left\{\frac{1}{2}\log|B^\top\tilde{\boldsymbol{\Sigma}}_g B| - \frac{1}{2}(\mathbf{m}_{ig}-\tilde{\boldsymbol{\mu}}_g)^\top\tilde{\boldsymbol{\Sigma}}_g^*(\mathbf{m}_{ig}-\tilde{\boldsymbol{\mu}}_g) - \frac{1}{2}\mathrm{Tr}(\tilde{\boldsymbol{\Sigma}}_g^* V_{ig}) + \frac{1}{2}\sum_{k=1}^{K}\log(v_{igk}^2) + \frac{K}{2}\right\}.\end{aligned} \tag{9}$$

Hence, we need to find optimal solutions to variational parameters $(\mathbf{m}_{ig}, V_{ig}, \xi_i)$ that are associated with each observation $\mathbf{w}_i, i = 1,\ldots,n$, as well as the model group-specific Gaussian parameters $(\tilde{\boldsymbol{\mu}}_g, \tilde{\boldsymbol{\Sigma}}_g), g = 1,\ldots,G$, such that the complete data variational Gaussian lower bound $\tilde{\mathscr{L}}$ is maximized. The use of VGA provides great reduction in the computational time.

The variational EM algorithm. Parameter estimation can be done in an iterative EM-type approach, from here on referred to as variational EM such that the following steps are iterated until convergence. For the parameters that do not have a closed form solution to the optimization, we perform one step of Newton's method to approximate the root to their first derivatives.

Step 1: Conditional on the variational parameters $(\mathbf{m}_{ig}, V_{ig}, \xi_i)$ and model group-specific Gaussian parameters $(\tilde{\boldsymbol{\mu}}_g, \tilde{\boldsymbol{\Sigma}}_g)$, $\mathbb{E}(Z_{ig}\mathbf{W}_i)$ is computed. Given $(\tilde{\boldsymbol{\mu}}_g, \tilde{\boldsymbol{\Sigma}}_g)$,

$$\mathbb{E}\left(Z_{ig} \mid \mathbf{w}_i\right) = \frac{\pi_g f_g(\mathbf{w}_i \mid \tilde{\boldsymbol{\mu}}_g, \tilde{\boldsymbol{\Sigma}}_g)}{\sum_{h=1}^{G}\pi_h f_h(\mathbf{w}_i \mid \tilde{\boldsymbol{\mu}}_h, \tilde{\boldsymbol{\Sigma}}_h)}.$$

This involves the marginal distribution of $\mathbf{W}$ and hence, we use an approximation of $\mathbb{E}\left(Z_{ig} \mid \mathbf{w}_i\right)$ where we replace the marginal density $\mathbf{W}$ by the exponent of ELBO such that

$$\hat{z}_{ig} := \frac{\pi_g \exp\left\{\tilde{F}\left(\mathbf{w}_i, \mathbf{m}_{ig}, V_{ig}, \tilde{\boldsymbol{\mu}}_g, \tilde{\boldsymbol{\Sigma}}_g, \xi_i\right)\right\}}{\sum_{j=1}^{G}\pi_j \exp\left\{\tilde{F}\left(\mathbf{w}_i, \mathbf{m}_{ij}, V_{ij}, \tilde{\boldsymbol{\mu}}_j, \tilde{\boldsymbol{\Sigma}}_j, \xi_i\right)\right\}}.$$

Step 2: Update $\hat{\xi}_i, \hat{\mathbf{m}}_{ig}, \hat{V}_{ig}$:

- update $\hat{\xi}_i$ according to Eq. (6);
- update $\hat{\mathbf{m}}_{ig}$ by performing one step of Newton's method for approximating the root to the derivative in Eq. (7), then let $\hat{m}_{ig(K+1)} = 0$;
- for $k = 1,\ldots,K$, update $\hat{v}_{igk}^2$ by performing one step of Newton's method searching root to the derivative in Eq. (8), let $\hat{v}_{ig(K+1)}^2 = 0$, then $\hat{V}_{ig} = \mathrm{diag}(\hat{v}_{ig1}^2,\ldots,\hat{v}_{ig(K+1)}^2)$.

Step 3: Update $\pi_{ig}, \tilde{\boldsymbol{\mu}}_g$ and $\tilde{\boldsymbol{\Sigma}}_g$ as

$$\hat{\pi}_{ig} = \frac{\sum_{n=1}^{n}\hat{z}_{ig}}{n}; \quad \hat{\tilde{\boldsymbol{\mu}}}_g = \frac{\sum_{i=1}^{n}\hat{z}_{ig}\hat{\mathbf{m}}_{ig}}{\sum_{i=1}^{n}\hat{z}_{ig}}; \quad \hat{\tilde{\boldsymbol{\Sigma}}}_g = \frac{\sum_{i=1}^{n}\hat{z}_{ig}\left[\hat{V}_{ig} + (\hat{\mathbf{m}}_{ig}-\hat{\tilde{\boldsymbol{\mu}}}_g)(\hat{\mathbf{m}}_{ig}-\hat{\tilde{\boldsymbol{\mu}}}_g)^\top\right]}{\sum_{i=1}^{n}\hat{z}_{ig}}.$$

Note that the original parameters $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ can be obtained by the transformation

$$\hat{\boldsymbol{\mu}}_g = B^\top\hat{\tilde{\boldsymbol{\mu}}}_g; \quad \hat{\boldsymbol{\Sigma}}_g = B^\top\hat{\tilde{\boldsymbol{\Sigma}}}_g B.$$

An Aitken acceleration criterion[48] is employed to stop the iterations. More specifically, at t th iteration, when $t > 2$, calculate

$$a^{(t-1)} = \frac{\ell^{(t)} - \ell^{(m-1)}}{\ell^{(t-1)} - \ell^{(t-2)}}; \quad \ell_\infty^{(t)} = \ell^{(t-1)} + \frac{1}{1-a^{(t-1)}}\left(\ell^{(t)} - \ell^{(t-2)}\right),$$

where $\ell^{(t)} = \tilde{F}\left(\mathbf{w}_i, \mathbf{m}_{ig}, V_{ig}, \tilde{\boldsymbol{\mu}}_g, \tilde{\boldsymbol{\Sigma}}_g, \xi_i\right)$ is the variational Gaussian lower bound who approximates the log likelihood at t th iteration. Then, the algorithm will be stopped when $\left|\ell_\infty^{(t)} - \ell_\infty^{(t-1)}\right| < \varepsilon$ for a given ε[49]. In our analysis, ε is set to be 1×10^{-3}.

Hybrid approach. While the VGA based approach only approximates the posterior distribution and it does not guarantee exact posterior[50], it is computationally efficient. On the other hand, a fully Bayesian MCMC based

approach can generate exact results, fitting such models can take substantial computational time. For example, fitting one iteration using a fully Bayesian MCMC model for a five dimensional dataset (from Simulation study 1) with $n = 1000$ takes on average of 45 minutes. In a clustering context, the number of iterations required for the analysis is typically in hundreds. Thus, we provide a computationally efficient hybrid approach in which

- Step 1: Fit the model using the VGA based approach.
- Step 2: Estimate the component indicator variable Z_{ig} conditional on the parameter estimates from the VGA based approach.
- Step 3: Using the parameter estimates from Step 1 as the initial values for the parameters and using the classification from Step 2, compute the MCMC based expectation for the latent variable $\tilde{\boldsymbol{\eta}}_{ig}$ as:

$$\mathbb{E}\left(\tilde{\eta}_{ig}|\mathbf{W}_i\right) \simeq \frac{1}{R}\sum_{k=1}^{R}\tilde{\eta}_{ig}^{(k)}.$$

  And $\boldsymbol{\theta}_{ng}^{(k)}$ is a random sample from the posterior distribution of $\tilde{\boldsymbol{\eta}}_{ig}$ simulated via the `RStan` package[51] for iterations $k = 1, \ldots, R$ (after discarding the burn-in).
- Step 4: Obtain the final estimates of the model parameters as:

$$\hat{\pi}_{ig} = \frac{\sum_{n=1}^{n}\hat{z}_{ig}}{n}; \quad \hat{\tilde{\boldsymbol{\mu}}}_g = \frac{\sum_{n=1}^{n}\hat{z}_{ig}\mathbb{E}\left(\tilde{\boldsymbol{\eta}}_{ig}\right)}{\sum_{n=1}^{n}\hat{z}_{ig}}; \quad \hat{\tilde{\boldsymbol{\Sigma}}}_g = \frac{\sum_{i=1}^{n}\hat{z}_{ig}\mathbb{E}\left[(\hat{\tilde{\boldsymbol{\eta}}}_{ig} - \hat{\tilde{\boldsymbol{\mu}}}_g)(\hat{\tilde{\boldsymbol{\eta}}}_{ig} - \hat{\tilde{\boldsymbol{\mu}}}_g)^{\top}\right]}{\sum_{i=1}^{n}\hat{z}_{ig}}.$$

The hybrid approach comes with a substantial reduction in computational overhead compared to a traditional MCMC based approach but it can generate samples from the exact posterior posterior distribution. Detailed comparison on computational time among the VGA based approach, the hybrid approach, and the MCMC-EM approach could be found in the results section. When the primary goal is to detect the underlying clusters (which is the case for our the real data analysis), the VGA based approach is sufficient. However, when the primarily goal is posterior inference, we recommend the hybrid approach as it can better yield an exact posterior similar to the MCMC-EM approach but is computationally efficient. For simulation studies 1 and 2 in which we show parameter recovery, we show parameter estimation using both VGA and the hybrid approach.

Initialization. For initialization of $\hat{z}_{ig}$, we used k-means clustering[52,53] on the estimate of the underlying latent variable η_i obtained by first calculating the underlying composition using $\mathbf{w}_i / \sum_{k=1}^{K+1}\mathbf{w}_{ik}$ for each observation; mapping this composition to the latent variable $\mathbf{y}_i$ using the additive log-ratio transformation in Eq. (1), and transforming the variable to get $\boldsymbol{\eta}_i$ through Eq. (2). For initializing the variational parameters for each observation $\mathbf{w}_i$, we obtain $\boldsymbol{\eta}_i$ first, same as in the $\hat{z}_{ig}$ initialization step. We use this calculated latent variable $\boldsymbol{\eta}_i$ as initialization of $\mathbf{m}_{ig}$. V_{ig} for each i are initialized as $K+1$ diagonal matrix such that $\tilde{V}_{kk} = 1$ for $k = 1, \ldots, K$ and $V_{kk} = 0$ for $k = K+1$. ξ_i's are initialized using 1. According to the initialization on the group label $\hat{z}_{ig}$, $\tilde{\boldsymbol{\mu}}_g$ and $\tilde{\boldsymbol{\Sigma}}_g$ are initialized as group-specific mean and covariance of $\boldsymbol{\eta}_i$, respectively.

Model selection and performance assessment. In the clustering context, the number of components G is unknown. Hence, one typically fits models for a large range of possible G and the number of clusters is then chosen *a posteriori* using a model selection criteria. The Bayesian information criterion (BIC)[54] is one of the most popular criteria in the model-based clustering literature[32]. Here, we use an approximation to BIC defined as

$$\text{BIC} \approx -2\tilde{\mathscr{L}} + d\log(n),$$

where $\tilde{\mathscr{L}}$, defined in Eq. (9), is the variational Gaussian lower bound of the complete data log likelihood, and d is the number of free parameters in the model. Specifically, when fitting a $G-$component model, $d = \frac{(K+1)K}{2} \times G + K \times G + G - 1$.

When the true class labels are known (e.g., in simulation studies), we assess the performance of our proposed model using the adjusted Rand index (ARI)[55]. It is a measure of the pairwise agreement between the predicted and true classifications such that an ARI of 1 indicates perfect classification and 0 indicates that the classification obtained is no better than by chance.

Results

Main simulation studies. To illustrate the performance of our proposed clustering framework, we conducted two sets of simulation studies. For both studies, the $i-$th observed counts data $\mathbf{W}_i$ are generated as:

1. First, we generate the total counts $\sum_{k=1}^{K+1} W_{ik}$ as a random number from a uniform distribution $U[5000, 10000]$.
2. Given pre-specified group specific parameters $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$, we transform using Eq. (3) to get $\tilde{\boldsymbol{\mu}}_g$ and $\tilde{\boldsymbol{\Sigma}}_g$ and generate $\boldsymbol{\eta}_i$ from $\text{N}(\tilde{\boldsymbol{\mu}}_g, \tilde{\boldsymbol{\Sigma}}_g)$.
3. Based on $\boldsymbol{\eta}_i$, we calculate $\boldsymbol{\Theta}_i$ using the inverse additive log-ratio transformation ϕ^{-1} using Eq. (4).
4. Using $\boldsymbol{\Theta}_i$ as the underlying composition, together with the total counts generated at the first step, we generate discrete random numbers $\mathbf{W}_i$ from multinomial distributions.

5. To initialize the variational parameters, we need to use the additive log-ratio transformation which takes the log transformation of the observed count for taxa k divided by total count for all taxa for sample i. If there are any 0 in the generated count data, we substitute the 0 with 1 for initialization.

We also compared the performance of our proposed model to Dirichlet-multinomial mixture models (DMMs)[20] which is widely used to cluster microbiome data. Implementation of the Dirichlet mixture model is available in the R package `DirichletMultinomial`[56]. We also applied Gaussian mixture models (GMMs) on the ALR-transformed compositions derived from these datasets with BIC for model selection. The GMMs were fitted using the `Mclust` function in the R package `mclust`. A family of finite mixture models with different covariance structures are implemented in `mclust`. The GMM model with unconstrained covariance structure "VVV" (the one that is most comparable to our proposed unrestricted covariance structure) encountered computational error for all simulated datasets. Only models assuming a spherical shape converged.

Simulation study 1. In this simulation study, we generated 100 datasets where the underlying latent variable $\mathbf{Y}$ came from two-component, three-dimensional multivariate Gaussian distributions with mixing proportions $\boldsymbol{\pi} = (0.6, 0.4)$; see Fig. 1a. The first three dimensions of the observed counts $\mathbf{W}$ are shown in Fig. 1b. The first component consists of $n_1 = 600$ observations and the second component consists of $n_2 = 400$ observations. The parameters used to generate the datasets are summarized in the Supplementary Table 1. We fitted the models with $G = 1, \ldots, 5$ on all 100 datasets. In 100 out of 100 datasets, BIC selected a two-component model. The models selected by BIC yielded an average ARI = 0.94 with a standard deviation of 0.02. The average and standard deviation of the estimated parameters for all 100 datasets using the VGA approach are summarized

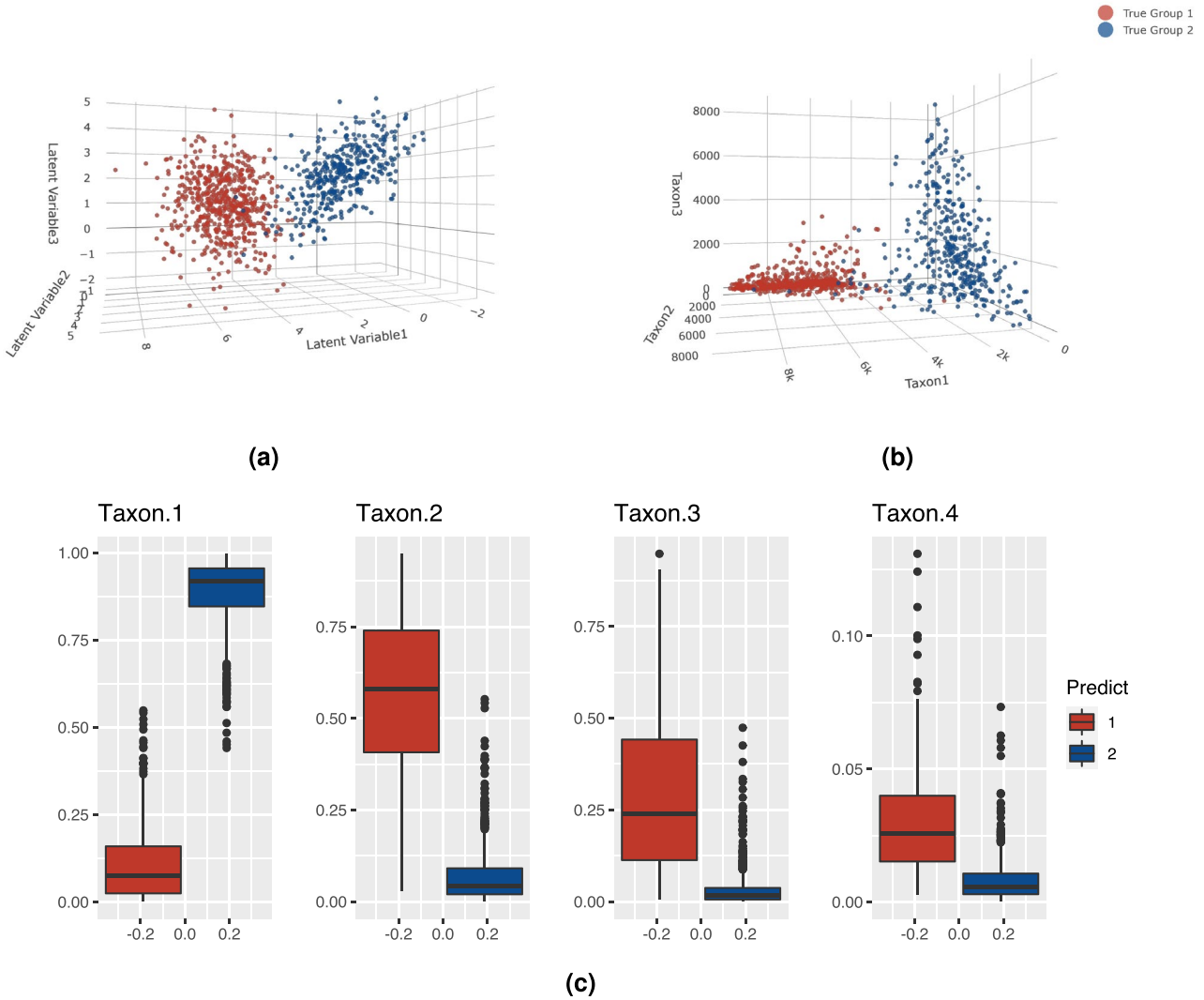

Figure 1. True and predicted cluster structure for one of the 100 datasets in Simulation Study 1. Panel (**a**) is the three-dimensional scatter plot of the underlying latent variable highlighted in true labels. Panel (**b**) is the first three dimensions of the observed count highlighted in true labels. Panel (**c**) is the relative abundance of observed counts of the four taxa for the predicted clusters. For this dataset, ARI was 0.95.

in Supplementary Table 1 and the hybrid approach are summarized in Supplementary Table 2. Note that the parameter estimations using both approaches are very close to the true value of the parameters.

The hybrid approach comes with a substantial reduction in computational overhead compared to a traditional MCMC-based approach but it can generate samples from the exact posterior distribution. The average computation time for Simulation Study 1 using the proposed VGA approach was 2.64 (sd of 0.61) minutes. The mean computation time using the hybrid approach was 47.78 (sd of 16.45) minutes. On the other hand, it took on average 45.14 (sd of 15.84) minutes for one iteration of the full Bayesian, and the number of iterations required for clustering is typically in the hundreds.

Figure 1c illustrates a clear difference in the distribution of the relative abundance of taxa in the two predicted groups. We also ran the DMMs and the GMMs on the ALR transformed compositions for $G = 1 : 5$ and selected the best model using BIC. In all 100 out of 100 datasets, a $G = 4$ or 5 model was selected for DMM with an average ARI of 0.46 (sd of 0.05). Similarly, in all 100 out of 100 datasets, a $G = 4$ or 5 model was selected for the GMMs with an average ARI of 0.39 (sd of 0.03). Both the DMMs and GMMs overestimated the number of components by splitting the true clusters into multiple clusters with some misclassifications among them.

Simulation study 2. In this simulation study, we generated 100 datasets with the underlying latent variable $\mathbf{Y}$ from three component five-dimensional multivariate Gaussian distributions (see Fig. 2a for the three-dimensional scatter plot of the first three dimensions of the underlying latent variable $\mathbf{Y}_i$).

There are $n_1 = 300$ observations in Group 1, $n_2 = 400$ observations in Group 1, and $n_3 = 200$ observations in Group 3. The true parameters are summarized in Supplementary Table 3. Figure 2b shows the first three

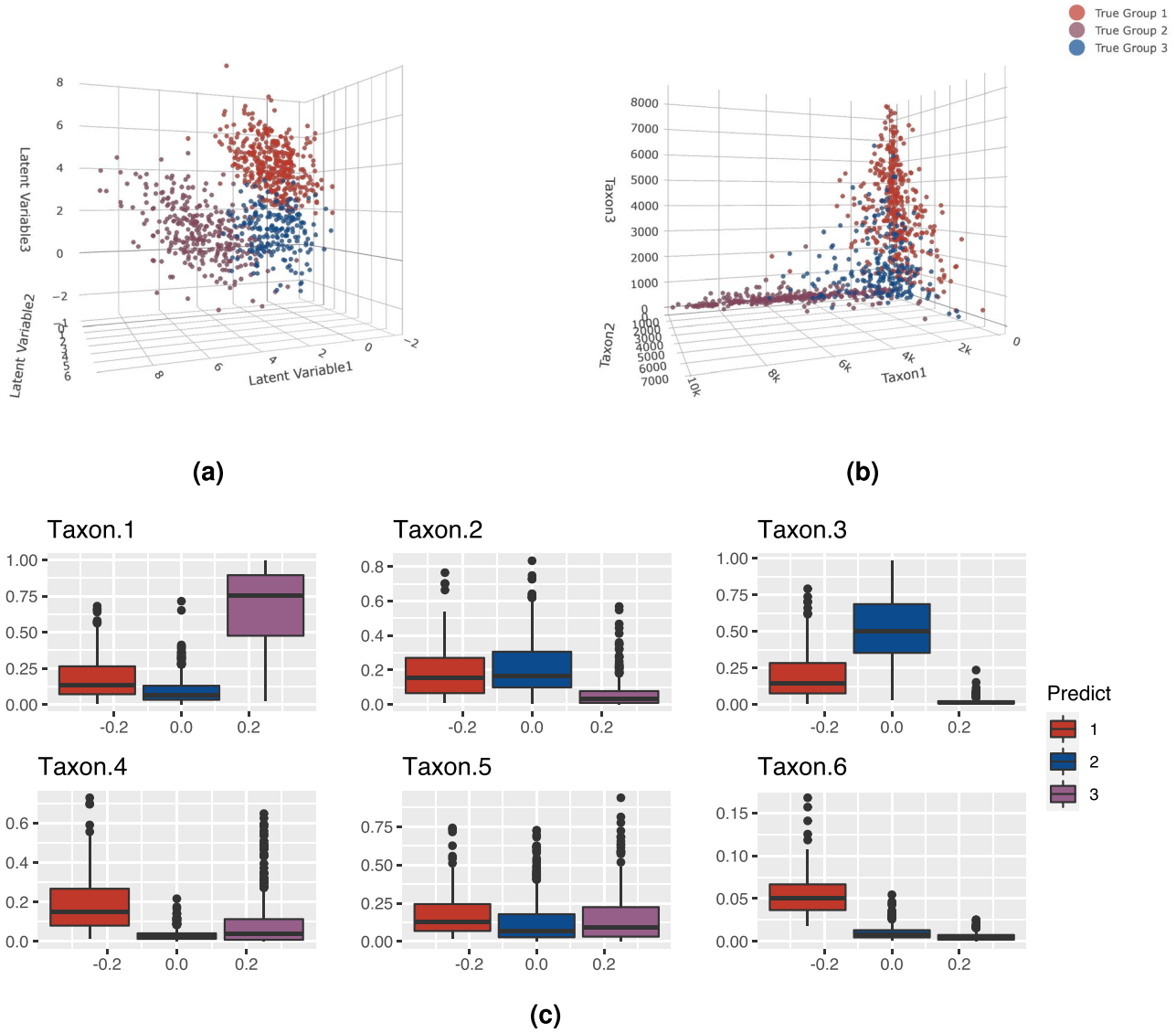

Figure 2. True and predicted cluster structure for one of the 100 datasets in Simulation Study 2. Panel (**a**) is the first three dimensions of the underlying latent variable highlighted in true labels. Panel (**b**) is the first three dimensions of the observed count highlighted in true labels. Panel (**c**) is the relative abundance of observed counts of the four taxa for the predicted clusters. For this dataset, ARI was 0.95.

dimensions of the observed counts $\mathbf{W}_i$'s. There is a clearer separation between the groups when visualizing the latent variables as opposed to the observed counts.

The proposed algorithm was applied on all 100 datasets where for each dataset, we fitted the models with for $G = 1, \ldots, 5$. In all 100 datasets, a $G = 3$ model was selected using the BIC and an overall mean ARI of 0.93 (sd of 0.02). The average and standard deviation of the estimated parameters for all 100 datasets using the VGA approach are summarized in Supplementary Table 3 and the hybrid approach are summarized in Supplementary Table 4. Note that the parameter estimation using both approaches are very close to the true value of the parameters. The average computation time for Simulation Study 2 using the proposed VGA approach was 3.03 (sd of 0.94) minutes. It took on average 40.84 (sd of 16.64) minutes for only one iteration of the MCMC-EM algorithm using the fully Bayesian approach (and the number of iterations required until convergence can be in the hundreds); whereas the mean computation time for fitting the hybrid approach until convergence (i.e., the sum of the computational times for all iterations) was 43.87 (sd of 17.58) minutes. Figure 2c illustrates a clear difference in the distribution of the relative abundance of taxa in the predicted groups. We also ran the DMMs on the observed abundance matrix and the GMMs on the ALR-transformed compositions for $G = 1 : 5$ and selected the best model using BIC for both approaches. In all 100 out of 100 datasets, both the DMMs and the GMMs overestimated the number of components. The DMMs selected a $G = 4$ model for 7 datasets and selected a $G = 5$ model for the remaining 93 datasets with an average ARI of 0.31 (sd of 0.03). The GMM selected a $G = 5$ model for all 100 datasets with an average ARI of 0.57 (sd of 0.03).

Additional simulation studies. To test the performance of the proposed algorithm on higher dimensional datasets, as well as datasets generated from a mixture of Dirichlet-multinomial models, we performed a series of 10 additional simulation studies, each containing 100 datasets, as described below:

- Generate 100 datasets from a two-component mixture of logistic normal multinomial models with each of the following:
  - K, the dimension of the latent variable, being 5, 10, and 20;
  - n, the sample size, being 100, 200, and 500.
  - True parameters are the same for different n but the same K.
- Generate 100 datasets from a mixture of two-component Dirichlet-multinomial models with dimension 6, and a sample size of 200.
- Generate 100 datasets from a mixture of two-component high dimensional logistic normal multinomial models with K=50 and $n = 500$.

We ran the proposed algorithm for $G = 1 : 5$ on all datasets and used BIC for model selection. We also applied the DMMs and GMMs on the ALR-transformed compositions derived from these datasets with BIC for model selection.

When data were generated from a mixture of logistic normal multinomial models, in all simulation scenarios, the proposed algorithm identified the correct number of components for all 100 datasets with average ARI $\geq$ 0.98. Also, it is observed that, in general, when sample size increases, the average ARI also increases and the standard deviation of ARI decreases. However, the Dirichlet-multinomial mixture model did not perform as well on data simulated from the logistic normal multinomial mixture models. Even in the case of $K = 20, n = 200$, where it correctly selected the two-component model 99 out of 100 times, the average ARI was only 0.63 with a standard deviation of 0.13. The GMMs on the ALR transformed data did not perform well either. While all models with different covariance structures available in the `mclust` package were fitted, the most comparable one with unrestricted covariance structure, specifying `modelName` = "`VVV`", encountered computational errors for all simulated datasets. Only models assuming a spherical shape converged and the GMM tended to overestimate the number of clusters (see Supplementary Table 5). Table 1 provides the number of correct G selected across 100 datasets fitting the proposed algorithm, the DMM, and the GMM on ALR transformed data, and the average ARI with standard deviation computed across all 100 datasets in each simulation scenario. Note that $G > 2$ encountered computational issues when fitting GMM with unrestricted spherical cluster model ("VII") on ALR transformed data for high dimensional $K = 50$ scenario for all datasets. Thus, only $G = 1$ and $G = 2$ were fitted and $G = 1, \ldots, 5$ could only be fitted for the model with equal spherical covariance across components ("EII") for the GMM. In 71 out of the 100 datasets, a $G = 2$ model with"VII" covariance structure was selected as the best fitting model. Although ARI here is high compared to fitting GLM on ALR-transformed data from other simulation settings, it must be noted that models with $G > 2$ encountered computational issues for the "VII" covariance structure.

When the data was generated from the Dirichlet-multinomial mixture models, our proposed model was able to recover the underlying cluster in 61 out of the 100 datasets with an average ARI of 0.78 and standard deviation of 0.18 whereas the Dirichlet-multinomial mixture model was able to recover the underlying cluster structure in all 100 datasets with an average ARI of 0.95 (sd=0.07). When performing each simulation study, the computational job was distributed onto a computer cluster, where the proposed algorithm applied on each one of the 100 datasets was run on a one-core slot. Table 1 summarizes the average elapsed time for running the proposed algorithm in, with standard deviation. In most cases, it takes the proposed algorithm less than 60 seconds. As the number of observations and the dimensionality of data increases, the time to convergence increases as well.

Simulation setting	Proposed algorithm			DMM		GMM on ALR transformed data	
	Average time in sec. (sd)	Correct G	ARI (sd)	Correct G	ARI (sd)	Correct G	ARI (sd)
K=5, n=100	3.20 (1.46)	100	0.98 (0.03)	1	0.00 (0.01)	2	0.47 (0.10)
K=5, n=200	5.76 (1.39)	100	0.99 (0.01)	5	0.00 (0.04)	0	0.42 (0.05)
K=5, n=500	16.90 (5.26)	100	0.99 (0.01)	3	0.11 (0.06)	0	0.40 (0.02)
K=10, n=100	6.18 (1.47)	100	1.00 (0.02)	64	0.42 (0.26)	0	0.51 (0.11)
K=10, n=200	14.28 (2.03)	100	1.00 (0.00)	88	0.57 (0.10)	0	0.45 (0.06)
K=10, n=500	55.22 (10.77)	100	1.00 (0.00)	0	0.37 (0.11)	0	0.45 (0.05)
K=20, n=100	12.00 (3.72)	100	1.00 (0.01)	20	0.14 (0.29)	28	0.77 (0.18)
K=20, n=200	39.45 (11.21)	100	1.00 (0.00)	99	0.63 (0.13)	0	0.47 (0.06)
K=20, n=500	151.12 (22.52)	100	1.00 (0.00)	0	0.48 (0.10)	0	0.43 (0.03)
DMM (k=5, n=200)	11.12 (5.02)	61	0.78 (0.18)	100	0.95 (0.07)	0	0.43 (0.09)
High Dimensional - K=50, n=500	1128.73 (312.71)	100	1.00 (0.00)	29	0.50 (0.39)	71	0.88 (0.19)

Table 1. Summary of the number of times the correct model is selected along with the average ARI (with standard deviation, across all 100 datasets) and average time per simulation (in seconds; with standard deviation) for completion for the 100 datasets for each of the 11 simulation studies described in the Additional Simulation Studies section fitting the proposed algorithm, the Dirichlet-multinomial mixture (DMM) models, and fitting the Gaussian mixture model (GMM) on additive log-ratio (ALR) transformed composition data.

The number of times each $G = 1:5$ were selected by the proposed algorithm are summarized in Supplementary Table 5. In nine out of the ten studies, our approach was able to identify the correct number of components for all 100 datasets. We also summarized the average of L_1 norm between the true parameters and the estimated values along with the standard errors for the simulations with data generated from a mixture of logistic normal multinomial models in Supplementary Table 6. It shows that, when the dimensionality is low, the proposed algorithm can not only identify the correct underlying group structure but also is able to recover the true parameters well. As the dimensionality increases, the proposed algorithm can still capture the true number of components in the data with high classification accuracy and the estimated central location parameter ($\boldsymbol{\mu}$) is also close to the true value. However, the estimation of the spread parameter ($\boldsymbol{\Sigma}$) become less precise as dimensionality becomes higher; however, the distance between the true and the estimated parameters decreases as the sample size becomes larger.

Real data analysis. *Scenario 1: Clustering microbiome data at a lower taxonomic level.* Here, we utilized our proposed algorithm to cluster the microbiome dataset at a lower taxonomic level. We applied our proposed algorithms to two previously published microbiome datasets.

- **The martínez dataset**: The study compares the fecal microbiota of individuals (40 adults) from two non-industrialized regions (20 participants from each of the Asaro and Sausi communities) in Papua New Guinea (PNG) with the individuals (22 adults) from the United States (US). The individuals from the Asaro and Sausi communities live a traditional agriculture-based lifestyle. The study found a greater bacterial diversity and lower inter-individual variations in the microbiome compositions of PNG individuals that were distinctly different from the individuals from industrialized US societies but no difference in bacterial diversity between the two PNG communities. The dataset was previously used for cluster analysis by Shi et al.[57] and is available through the R package `MicrobiomeCluster` via https://github.com/YushuShi/MicrobiomeCluster.git. Here, we conducted the analysis at the OTU level.
- **The ferretti 2018 dataset**: The Ferretti 2018 study[58] aims to understand the acquisition and development of the infant microbiome and assess the impact of the maternal microbiomes on the development of an infant's microbial communities from birth to 4 months of life. Twenty five mother- infant pairs who vaginally delivered healthy newborns at full term were recruited for the study. For each mother, stool (a proxy for gut microbiome), dorsum tongue swabs (for oral microbiome), vaginal introitus swabs (for vaginal microbiome), intermammary cleft swabs (skin microbiome) and breast milk were obtained. Here, we applied our algorithms to a subset of the dataset to compare the oral microbiome of the infants with their mothers. Oral samples of infants were taken at two different time points: Day 1 and Day 3. Here, we used measurements from Day 1. The resulting dataset consists of 39 individuals (23 adults and 16 infants). The dataset available through the R package `curatedMetagenomicData`[59] as `FerrettiP_2018` dataset. We conducted the analysis at the genus level.

As our approach is currently not designed for high dimensional data, we utilize two different approaches for dimension reduction prior to clustering:

- In `Case I`, we first extracted the top ten most abundant taxa using the R package `HMP`[60] and used it for the clustering analysis. This approach requires no prior information on the cluster structure and can be utilized in a true clustering scenario. Here, we used the top ten most abundant taxa for both datasets. To preserve the compositional nature of the data, the remaining taxa were all grouped into a taxa category "Others". This "Other" taxa was then used as the reference level for conducting the additive log-ratio transformation.
- In `Case II`, we first utilized the R package `ALDEx2`[61,62] for differential abundance analysis on the observed taxa counts to identify the taxa that are different among different groups in the datasets. This step is analogous to conducting differential expression analysis in RNA-seq studies before performing cluster analysis to identify variables that are group differentiating.

The motivation behind proposing two different cases is to illustrate that while `Case I` requires no prior information on cluster structure, using the top most abundant taxa may not be always appropriate. When the top most abundant taxa contain group differentiating information, the proposed approach provides a good clustering performance. However, when the taxa with lower abundance are group differentiating features, in such case, not including those features may result in a decrease in the clustering performance. When the taxa that are group differentiating are identified in `Case II` using differential abundance analysis, the proposed approach can provide a better clustering performance on the same datasets.

We applied our algorithm to all datasets for $G = 1$ to 4. We repeated the analysis 10 times with different *k*-means initialization and selected the final model using BIC. We also ran the Dirichlet-multinomial mixture model with the same set of taxa and the GMM on the transformed latent variable for $G = 1$ to 4 on all datasets and utilized BIC for model selection. A summary of the clustering performances is provided in Table 2.

For both datasets, our approach outperformed the Dirichlet-multinomial mixture models and the GMM applied to the latent variables under both scenarios. When there is a good overlap between the group differentiating taxa and most abundant (i.e., in the case of the Martínez dataset), the proposed approach, Dirichlet multinomial mixture model, and the Gaussian mixture model with the transformed variables all provide good clustering performance for Case I and Case II. On the other hand, when the group differentiating taxa differ

	Proposed (ARI: 1)		**DMM model (ARI: 0.61)**			**GMM model (ARI: 0.46)**			
	Estimated clusters		**Estimated clusters**			**Estimated clusters**			
	1	**2**	**1**	**2**	**3**	**1**	**2**	**3**	**4**
The `Martínez` Dataset									
Case I									
US	22	–	22	–	–	22	–	–	–
PNG	–	40	-	25	15	–	11	14	15
	Proposed (ARI: 1)		**DMM model (ARI: 0.76)**			**GMM model (ARI: 0.69)**			
	Estimated clusters		**Estimated clusters**			**Estimated clusters**			
	1	**2**	**1**	**2**	**3**	**1**	**2**	**3**	**4**
Case II									
US	22	–	22	–	–	12	10	–	–
PNG	–	40	–	33	7	–	–	35	5
	Proposed (ARI: 0.71)		**DMM model (ARI: 0.62)**			**GMM model (ARI: 0.61)**			
	Estimated clusters		**Estimated clusters**			**Estimated clusters**			
	1	**2**	**1**	**2**		**1**	**2**	**3**	**4**
The `Ferretti` Oral microbiome subset									
Case I									
Infant	13	3	12	4		14	1	1	–
Adult	–	23	-	23		–	18	2	3
	Proposed (ARI: 1)		**DMM model (ARI: 0.80)**			**GMM model (ARI: 0.37)**			
	Estimated clusters		**Estimated clusters**			**Estimated clusters**			
	1	**2**	**1**	**2**		**1**	**2**	**3**	**4**
Case II									
Infant	16	–	14	2		7	3	6	–
Adult	–	23	–	23		–	–	7	16

Table 2. Cross tabulation of the clusters obtained by our proposed algorithm and Dirichlet-multinomial mixture model on all three real datasets.

from the most abundant taxa (i.e., in the case of the Ferretti dataset), in such scenario, all models tested perform better for Case II.

Scenario 2: clustering healthy microbiome samples from Human microbiome project at a higher taxonomic level. We also applied our algorithm to the `HMP2012` dataset[63] available from the R package `curatedMetagenomicData`[59]. The dataset comprises microbiome compositions of 129 males and 113 females. "Healthy" individuals (i.e., individuals without any evidence of diseases) were recruited and samples were collected from one or more of the five different body sites (nasal cavity, oral cavity, skin, stool, and vagina). In total, we have $n = 748$ microbiome sample profiles. Here, we focused on analyzing the dataset at the Phylum level. First, the top ten most abundant phyla were extracted and phyla with at least 5% non-zero counts were retained which resulted in 8 phyla that were retained. The remaining phyla were all grouped into a phylum category "Others" which was used as a reference level for conducting log-ratio transformation.

We utilized our algorithm to the `HMP2012` dataset for $G = 1$ to 10 and we repeated the analysis 10 times with different k-means initialization and selected the final model using BIC. A seven-component LNM-MM was selected. Figure 3 provides a visualization of the relative abundances of the top most abundant phyla across the seven components and the cross-tabulation of the estimated cluster membership against the five body sites is provided in Table 3.

Cluster 1 comprised only of stool samples and clusters 2 and 3 comprised only oral cavity samples. Cluster 6 is also comprised primarily of oral cavity samples. Cluster 4 comprised a mix of samples primarily from the nasal cavity and skin; cluster 5 comprised a mix of samples primarily from the nasal cavity and vagina; and cluster 7 comprised a mix of samples primarily of the stool and vagina. It is interesting to note that samples from the same body sites are clustered into multiple clusters, in some cases, with samples from other body sites. For example, samples from the nasal cavity were clustered into two clusters: cluster 4 and cluster 5 where the cluster 5 also comprised samples from the skin. Visualization of the relative abundance of samples from the skin and nasal cavity assigned to clusters 4 and 5 in Fig. 4 reveals that in fact, the microbiome profiles of samples from

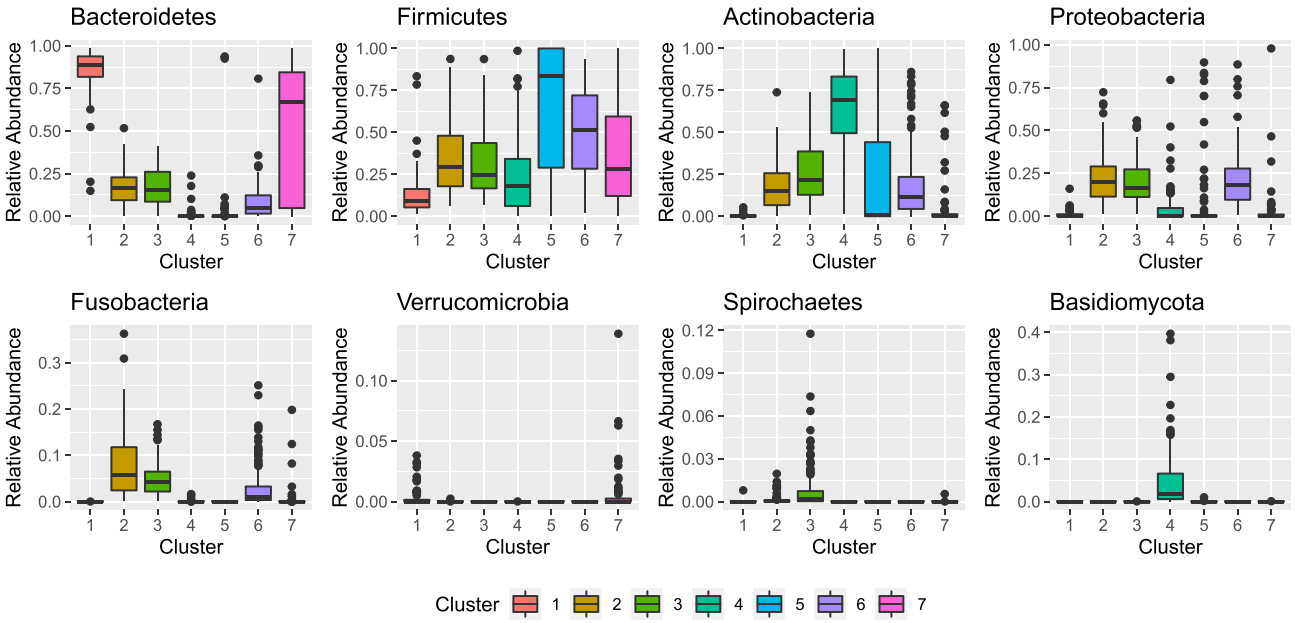

Figure 3. Boxplot of relative abundances of the top most abundant phyla for all seven components.

Clusters	Nasal cavity	Oral cavity	Skin	Stool	Vagina
1	–	–	–	73	–
2	–	124	–	–	–
3	–	116	–	–	–
4	59	2	26	–	1
5	29	6	1	3	39
6	3	160	–	1	–
7	2	6	–	70	27

Table 3. Cross tabulation of the estimated cluster membership against the five body sites and the compositions of the estimated clusters.

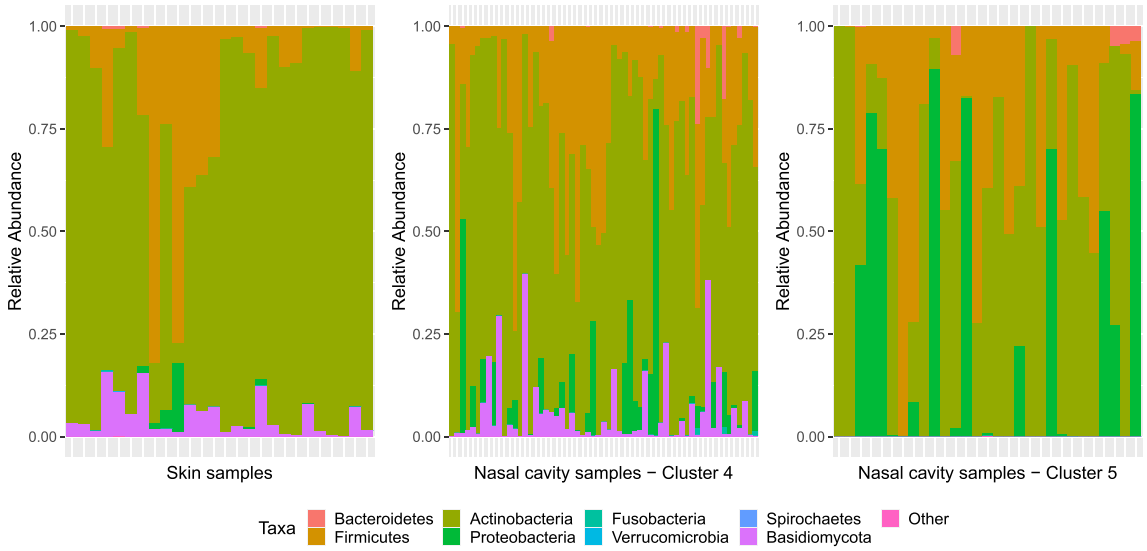

Figure 4. Visualization of the relative abundances of skin samples and samples from nasal cavity assigned to clusters 4 and 5.

the nasal cavity in cluster 4 are more similar to the microbiome profiles of samples from the skin than to the samples from the nasal cavity assigned to cluster 5. This is in alignment with the findings of the original study[63] where a high overlap between the nasal cavity samples and skin samples is observed in the principal coordinates plot of the samples.

Similarly, samples from the oral cavity were assigned to three clusters: 2, 3 and 6. Boxplots of the relative abundances of the samples in Fig. 3 reveal that the relative abundances of taxa in the three clusters are different. Samples in cluster 3 have a much higher relative abundance of Spirochaetes compared to clusters 2 and 3 and samples in cluster 6 have a higher relative abundance of Firmicutes compared to samples in clusters 2 and 3 and a lower relative abundance of Fusobacteria compared to samples in cluster 2. The DMM model and the GMM applied on latent variable were also fitted to the same dataset for $G = 1$ to 10 and BIC was utilized for model selection. A six-component model was selected by DMM and a similar trend as our proposed algorithm was observed. The samples from the oral cavity were assigned to distinct three clusters, and part of the samples from the nasal cavity were clustered together with samples from the skin. Similar to what we observed from the simulation studies and the other real data results, GMM on the ALR transformed data overestimated the number of components here as well. A ten-component model was selected and a similar trend as the proposed algorithm result was observed, where the two oral cavity sample clusters and one of the stool sample clusters were further split into two smaller clusters each separately.

Discussion

A model-based clustering framework for microbiome compositional data is developed using a mixture of logistic normal multinomial models. The novelty of this work is multi-fold. Previous work[23] has indicated that the logistic normal multinomial models can model the dependency of the bacterial composition in a microbiome compositional data in a more flexible way than the commonly used Dirichlet-multinomial models. The latent variables in the logistic normal multinomial model are assumed to follow a multivariate Gaussian distribution and a closed form expression of the log-likelihood or posterior distributions of the latent variables do not exist. Hence, prior work on model fitting relied on Markov chain Monte Carlo (MCMC) sampling techniques that come with a heavy computational burden. This is compounded in the clustering context where MCMC sampling needs to be utilized at every iteration of the variant of the EM algorithm that is typically utilized for parameter estimation. Here, we employed a variational Gaussian approximation to the posterior distribution of the latent variable and implemented a generalized EM algorithm that does not rely on MCMC sampling thus making it feasible to extend these models for clustering. This also opens up the possibility of efficiently scaling and extending these models to a high-dimensional setting.

Through simulation studies, we have shown that the proposed algorithm delivers accurate parameter recovery and good clustering performance. The proposed method is also illustrated on three real datasets in Section "Real data analysis" where we demonstrate that the proposed models can recover the interesting cluster (group) structure in the real data. While in the datasets with small sample sizes, we focus on small dimensional data by data aggregation to most differentially abundant genera in real data analysis, for larger datasets, more taxa can be used. Because of adopting an underlying Gaussian distribution, the number of parameters in the covariance matrix alone grows quadratically with K. Thus, in high dimensional datasets with small sample size, estimating $\mathbf{\Sigma}^{-1}$ becomes more challenging as it can lead to degenerate solutions and a host of other issues related to model convergence and fitting while using a traditional maximum likelihood-based expectation-maximization approach. This a well-known issue with Gaussian mixture models and is typically dealt with either variable/feature selection or dimension reduction. Feature selection typically eliminates the redundant or irrelevant

variables and reduces computational cost, provides a better understanding of data and improves predictions[64]. ALDEx2 utilized here is a widely used variable/feature selection technique specifically designed for microbiome data that identifies taxa that are differentially abundant in different conditions. Through a comparative study of ALDEx2 with other approaches commonly used for differential abundance analysis, Quinn et al.[65] showed that ALDEx2 has high precision (i.e., few false positives) across different scenarios. However, information on the group structure or conditions may not be available a-priori. In such cases, one may conduct feature selection by selecting the top few most abundant taxa and collapsing low-abundant taxa into one category "Others" to preserve the compositional nature of the data. Alternately, mixtures of logistic multinomial models can be extended to high-dimensional data by introducing subspace clustering techniques through the latent variable[66–68]. This will be the topic of some future work. Additionally, it has been well-established that different environmental or biological covariates can affect the microbiome compositions. Some future work will also focus on developing a mixture of logistic normal multinomial regression models to investigate the relationship of biological/environmental covariates with the microbiome compositions within each cluster.

Data availability

The datasets used in this manuscript are publicly available from the R package `curatedMetagenomicData` (https://bioconductor.org/packages/curatedMetagenomicData/) and `MicrobiomeCluster` (https://github.com/YushuShi/MicrobiomeCluster).

Received: 19 April 2023; Accepted: 24 August 2023
Published online: 07 September 2023

References

1. Morgan, X. C. & Huttenhower, C. Human microbiome analysis. *PLOS Comput. Biol.* **8**, e1002808 (2012).
2. Li, H. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Ann. Rev. Stat. Appl.* **2**, 73–94 (2015).
3. Ley, R. E., Peterson, D. A. & Gordon, J. I. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124**, 837–848 (2006).
4. Fraher, M. H., Ótoole, P. W. & Quigley, E. M. Techniques used to characterize the gut microbiota: A guide for the clinician. *Nat. Rev. Gastroenterol. Hepatol.* **9**, 312 (2012).
5. Koeth, R. A. *et al.* Intestinal microbiota metabolism of l-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat. Med.* **19**, 576 (2013).
6. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).
7. Greenblum, S., Turnbaugh, P. J. & Borenstein, E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc. Natl. Acad. Sci.* **109**, 594–599 (2012).
8. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484 (2009).
9. Streit, W. R. & Schmitz, R. A. Metagenomics-the key to the uncultured microbes. *Curr. Opin. Microbiol.* **7**, 492–498 (2004).
10. Kuczynski, J. *et al.* Experimental and analytical tools for studying the human microbiome. *Nat. Rev. Genet.* **13**, 47–58 (2012).
11. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
12. Äijö, T., Müller, C. L. & Bonneau, R. Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing. *Bioinformatics* **34**, 372–380 (2018).
13. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
14. Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308**, 1635–1638 (2005).
15. Hamady, M. & Knight, R. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome. Res.* **19**, 1141–1152 (2009).
16. Zhang, X. *et al.* Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinform.* **18**, 4 (2017).
17. Zhang, X. & Yi, N. Fast zero-inflated negative binomial mixed modeling approach for analyzing longitudinal metagenomics data. *Bioinformatics* **36**, 2345–2351 (2020).
18. Joseph, N., Paulson, C., Corrada Bravo, H. & Pop, M. Robust methods for differential abundance analysis in marker gene surveys. *Nat. Methods* **10**, 1200–1202 (2013).
19. Xu, T., Demmer, R. T. & Li, G. Zero-inflated Poisson factor model with application to microbiome read counts. *Biometrics* (2020).
20. Holmes, I., Harris, K. & Quince, C. Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLOS One* **7**, e30126 (2012).
21. Chen, J. & Li, H. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Statist.***7** (2013).
22. Subedi, S., Neish, D., Bak, S. & Feng, Z. Cluster analysis of microbiome data by using mixtures of Dirichlet-multinomial regression models. *J. Royal Statist. Soc. Ser. C* **69**, 1163–1187 (2020).
23. Xia, F., Chen, J., Fung, W. K. & Li, H. A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics* **69**, 1053–1063 (2013).
24. Xu, L., Paterson, A. D., Turpin, W. & Xu, W. Assessment and selection of competing models for zero-inflated microbiome data. *PLOS One* **10**, e0129606 (2015).
25. Wadsworth, W. D. *et al.* An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinformatics* **18**, 94 (2017).
26. Cao, Y., Zhang, A. & Li, H. Multi-sample estimation of bacterial composition matrix in metagenomics data. arXiv preprint arXiv:1706.02380 (2017).
27. Caporaso, J. G. *et al.* Moving pictures of the human microbiome. *Genome Biol.* **12**, R50 (2011).
28. Silverman, J. D., Durand, H. K., Bloom, R. J., Mukherjee, S. & David, L. A. Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. *Microbiome* **6**, 1–20 (2018).
29. McLachlan, G. & Peel, D. *Finite Mixture Models* (Wiley, 2000).
30. Zhong, S. & Ghosh, J. A unified framework for model-based clustering. *J. Mach. Learn. Res.* **4**, 1001–1037 (2003).
31. Frühwirth-Schnatter, S. *Finite Mixture and Markov Switching Models* (Springer, 2006).
32. McNicholas, P. D. *Mixture Model-Based Classification* (Chapman and Hall/CRC, 2016).
33. Rau, A., Celeux, G., Martin-Magniette, M.-L. & Maugis-Rabusseau, C. *Clustering high-throughput sequencing data with Poisson mixture models* (Tech. Rep, INRIA, Saclay, Ile-de-France, 2011).
34. Papastamoulis, P., Martin-Magniette, M.-L. & Maugis-Rabusseau, C. On the estimation of mixtures of Poisson regression models with large number of components. *Comput. Statist. Data Anal.* **93**, 97–106 (2016).

35. Si, Y., Liu, P., Li, P. & Brutnell, T. P. Model-based clustering for RNA-seq data. *Bioinformatics* **30**, 197–205 (2014).
36. Silva, A., Rothstein, S. J., McNicholas, P. D. & Subedi, S. A multivariate Poisson-log normal mixture model for clustering transcriptome sequencing data. *BMC Bioinform.* **20**, 394 (2019).
37. Barber, D. & Bishop, C. M. Ensemble learning in Bayesian neural networks. *Nato ASI Ser. F Comput. Syst. Sci.* **168**, 215–238 (1998).
38. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).
39. Arridge, S. R., Ito, K., Jin, B. & Zhang, C. Variational Gaussian approximation for Poisson data. *Inverse Prob.* **34**, 025005 (2018).
40. Archambeau, C., Cornford, D., Opper, M. & Shawe-Taylor, J. Gaussian process approximations of stochastic differential equations. *J. Mach. Learn. Res.* **1**, 1–16 (2007).
41. Khan, E., Mohamed, S. & Murphy, K. P. Fast Bayesian inference for non-conjugate Gaussian process regression. *In Adv. Neural Inform. Process. Syst.* **25**, 3140–3148 (2012).
42. Challis, E. & Barber, D. Gaussian Kullback–Leibler approximate inference. *J. Mach. Learn. Res.* **14**, 2239–2286 (2013).
43. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* **112**, 859–877 (2017).
44. Aitchison, J. The statistical analysis of compositional data. *J. Roy. Stat. Soc.: Ser. B (Methodol.)* **44**, 139–160 (1982).
45. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc.: Ser. B (Methodol.)* **39**, 1–22 (1977).
46. Wainwright, M. J. *et al.* Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1**, 1–305 (2008).
47. Blei, D. & Lafferty, J. Correlated topic models. *Adv. Neural. Inf. Process. Syst.* **18**, 147 (2006).
48. Aitken, A. C. A series formula for the roots of algebraic and transcendental equations. *Proc. R. Soc. Edinb.* **45**, 14–22 (1926).
49. Böhning, D., Dietz, E., Schaub, R., Schlattmann, P. & Lindsay, B. G. The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Ann. Inst. Stat. Math.* **46**, 373–388 (1994).
50. Ghahramani, Z. & Beal, M. Variational inference for Bayesian mixtures of factor analysers. *Adv. Neural Inform. Process. Syst.***12** (1999).
51. Stan Development Team. RStan: the R interface to Stan (2023). R package version 2.21.8.
52. MacQuuen, J. Some methods for classification and analysis of multivariate observation, in *Proceedings of the 5th Berkley Symposium on Mathematical Statistics and Probability*, 281–297 (1967).
53. Hartigan, J. A. & Wong, M. A. A k-means clustering algorithm. *Appl. Stat.* **28**, 100–108 (1979).
54. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
55. Hubert, L. & Arabie, P. Comparing partitions. *J. Classif.* **2**, 193–218 (1985).
56. Morgan, M. *DirichletMultinomial: Dirichlet-Multinomial Mixture Model Machine Learning for Microbiome Data* (2020). R package version 1.32.0.
57. Shi, Y., Zhang, L., Peterson, C. B., Do, K.-A. & Jenq, R. R. Performance determinants of unsupervised clustering methods for microbiome data. *Microbiome* **10**, 1–12 (2022).
58. Ferretti, P. *et al.* Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. *Cell Host Microbe* **24**, 133–145 (2018).
59. Pasolli, E. *et al.* Accessible, curated metagenomic data through experimenthub. *Nat. Methods* **14**, 1023–1024. https://doi.org/10.1038/nmeth.4468 (2017).
60. Rosa, P., Deych, E., Shands, B. & Shannon, W. HMP: hypothesis testing and power calculations for comparing metagenomic samples from HMP (2013).
61. Fernandes, A., Macklaim, J., Linn, T., Reid, G. & Gloor, G. ANOVA-like differential gene expression analysis of single-organism and meta-RNA-seq. *PLoS ONE* **8**, e67019 (2013).
62. Fernandes, A. D. *et al.* Unifying the analysis of high-throughput sequencing datasets: Characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2**, 1–13 (2014).
63. Consortium HMP. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
64. Haq, A. U., Zhang, D., Peng, H. & Rahman, S. U. Combining multiple feature-ranking techniques and clustering of variables for feature selection. *IEEE Access* **7**, 151482–151492 (2019).
65. Quinn, T. P., Crowley, T. M. & Richardson, M. F. Benchmarking differential expression analysis tools for RNA-Seq: Normalization-based vs. log-ratio transformation-based methods.. *Bioinformatics* **19**, 1–15 (2018).
66. Mcnicholas, P. D. & Murphy, T. B. Parsimonious Gaussian mixture models. *Stat. Comput.* **18**, 285–296 (2008).
67. McNicholas, P. D. & Murphy, T. B. Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics* **26**, 2705–2712 (2010).
68. Bouveyron, C. & Brunet-Saumard, C. Model-based clustering of high-dimensional data: A review. *Comput. Stat. Data Anal.* **71**, 52–78 (2014).

Acknowledgements

This work was supported by Collaboration Grants for Mathematicians by Simons Foundation (Subedi); NSERC Discovery Grant (Subedi); and fundings from Canada Research Chair Program. This research was enabled in part by support provided by Research Computing Services (https://carleton.ca/rcs) at Carleton University.

Author contributions

S.S. conceptualized the framework. Y.F. derived the math and coded the scheme and performed all simulation studies. S.S. optimized the code and performed all real data analyses. Both authors wrote and revised the manuscript. Both authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-41318-8.

Correspondence and requests for materials should be addressed to S.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.