Logistic Normal Multinomial Factor Analyzers for Clustering Microbiome Data

Wangshu Tu · Sanjeena Subedi

Accepted: 12 October 2023 / Published online: 7 November 2023
© The Author(s) under exclusive licence to The Classification Society 2023

Abstract
The human microbiome plays an important role in human health and disease status. Next-generating sequencing technologies allow for quantifying the composition of the human microbiome. Clustering these microbiome data can provide valuable information by identifying underlying patterns across samples. Recently, Fang and Subedi (2023) proposed a logistic normal multinomial mixture model (LNM-MM) for clustering microbiome data. As microbiome data tends to be high dimensional, here, we develop a family of logistic normal multinomial factor analyzers (LNM-FA) by incorporating a factor analyzer structure in the LNM-MM. This family of models is more suitable for high-dimensional data as the number of free parameters in LNM-FA can be greatly reduced by assuming that the number of latent factors is small. Parameter estimation is done using a computationally efficient variant of the alternating expectation conditional maximization algorithm that utilizes variational Gaussian approximations. The proposed method is illustrated using simulated and real datasets.

Keywords Cluster analysis · Microbiome data · Model-based clustering · High-dimensional data · Mixture model · Logistic normal multinomial model

1 Introduction
The human microbiota is a complex collection of microbes including but not limited to bacteria, fungi, and viruses that reside in the human body. It is estimated that nearly 30 trillion bacterial cells are living in or on each human body, which is about one bacterium for every cell in the human body (Sender et al., 2016). These organisms play an important role in human health and diseases (Huttenhower et al., 2012). For example, changes in the gut microbiota have been linked to inflammatory bowel disease (Becker et al., 2015), obesity (Davis, 2016), type 2 diabetes (Cho & Blaser, 2012), and cancer (Pfirsichke et al., 2015). Using next-generating sequencing technologies, the abundance and composition of
these microbes can be quantified. Microbiome data is treated as compositional because the read counts for each taxon are constrained by the capacity of the sequencing instrument (Gloor et al., 2017). A detailed discussion on the compositional nature of microbiome data is provided in Gloor et al. (2017). Using Euclidean geometry on such compositional datasets is not appropriate—when the compositional vectors are added or multiplied, the results may end up outside the simplex and Euclidean difference does not reveal the true difference in relative abundance (Pawlowsky-Glahn et al., 2007). Therefore, it is important to take into account the compositional nature of the data.

Cluster analysis has been widely used to gain insights from microbiome data. Cluster analysis is used to group observations into homogeneous subpopulations with similar characteristics. Enterotype refers to groups of individuals with similar gut microbial communities (Arumugam et al., 2011). Wu et al. (2011) used a partitioning around medoids (PAM) approach with various distance measures to cluster the gut microbiota samples of 98 healthy volunteers and found that the number of enterotypes varied between two and three. Abdel-Aziz et al. (2021) utilized hierarchical clustering to cluster the sputum microbiome datasets and identified two distinct robust phenotypes of severe asthma. In other work, $k$-means clustering has also been used to cluster microbiome data (Taie et al., 2018; Hotterbeekx et al., 2016). Although $k$-means, PAM, and hierarchical clustering are well-established clustering techniques and are frequently used in many fields, when applied to observed counts directly, these approaches fail to take into account the compositional nature of the microbiome data and cannot be applied directly to the counts. A common practice is to utilize count normalization either before using such techniques or to incorporate a normalization factor in the modelling framework when using univariate models like a negative binomial model (Zhang et al., 2017). Commonly used normalization approaches rely on assumptions such that most features remain unchanged which may hold for RNA-seq data but does not hold for microbiome data (Gloor et al., 2017). Moreover, commonly used normalization approaches are not as suitable for microbiome datasets which are highly asymmetrical or sparse datasets (Gloor et al., 2017).

A natural choice for modelling these compositional data is a multinomial model with parameter $\mathbf{p} = (p_1, \ldots, p_{K+1})'$. However, the observed variability in the microbiome composition data is greater than predicted by the multinomial model. To account for this additional variability, one of two common approaches may be used. The first approach is to treat the probability vector $\mathbf{p}$ as a random variable and impose a Dirichlet distribution such that for each observation $i$, $Y_i | \mathbf{p}_i \sim \text{Multinomial}(\mathbf{p}_i)$ and $\mathbf{p} \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_{K+1})$. The resulting compound distribution is known as the Dirichlet-multinomial distribution. A Dirichlet-multinomial model has been widely used for modelling microbiome data (La Rosa et al., 2012; Chen & Li, 2013; Wadsworth et al., 2017; Koslovsky & Vannucci, 2020). In the cluster analysis literature, Holmes et al. (2012) proposed a Dirichlet-multinomial mixture (DMM) model to cluster microbiome data. Subedi et al. (2020) proposed mixtures of Dirichlet-multinomial regression models to cluster microbiome data which can incorporate the effects of covariates. However, due to the limited number of parameters in the Dirichlet distribution, the covariance of the microbiome data cannot be modelled adequately using a Dirichlet-multinomial distribution (Xia et al., 2013).

The Dirichlet-tree multinomial model by Wang and Zhao (2017) is an alternate distribution for modelling microbiome data that utilizes the hierarchical structure among the taxa (i.e., phylogenetic tree) in the modelling paradigm. This is done through the use of a Dirichlet-
tree (DT) distribution via $p$ to provide a more flexible covariance structure. However, Wang and Zhao (2017) note that correctly specifying the phylogenetic tree structure is crucial and misspecification of the tree structure from the shotgun sequencing data could be a major risk. Furthermore, Mao and Ma (2022) states that while the DT provides a more flexible covariance structure, it is still limited in comparison to approaches based on log-ratio transformations.

A log-ratio transformation is another commonly used approach for dealing with compositional data in which $p$ is mapped from a restricted simplex to real space through the use of a log-ratio transformation. To account for additional heterogeneity, priors are imposed on the transformed variable (Xia et al., 2013; Äijö et al., 2018; Silverman et al., 2018). Three log-ratio transformations that are used for microbiome data are the additive log-ratio (ALR) transformation (Xia et al., 2013; Äijö et al., 2018), centered log-ratio (CLR) transformation (Fernandes et al., 2014), and the isometric log-ratio (ILR) transformation (Silverman et al., 2018). With an ALR transformation, $p$ is mapped from the restricted simplex $\mathbb{S}^K$ to a $K$-dimensional open real space $\mathbb{R}^K$ such that $W = \phi(p) = \begin{bmatrix} \log \left( \frac{p_1}{p_{K+1}} \right), \ldots, \log \left( \frac{p_K}{p_{K+1}} \right) \end{bmatrix}^T$, where $p_{K+1}$ is used as a reference. On the other hand, with a CLR transformation, $p$ is mapped from the restricted simplex $\mathbb{S}^K$ to a $K+1$-dimensional open real space $\mathbb{R}^{K+1}$ such that $W = \begin{bmatrix} \log \left( \frac{p_1}{g(p)} \right), \ldots, \log \left( \frac{p_K}{g(p)} \right) \end{bmatrix}^T$, where $g(p)$ is the geometric mean of the composition. The ILR transformation used by Silverman et al. (2018) for microbiome data is a series of sequential log-ratios between two groups of features (Pawlowsky-Glahn & Buccianti, 2011; Gloor et al., 2017). While both ALR and CLR transformations provide better interpretability compared to ILR in terms of individual features in the datasets, the ALR transformation does not preserve distances. In contrast, the CLR transformation preserves distances but leads to a singular covariance matrix (Pawlowsky-Glahn & Buccianti, 2011; Pawlowsky-Glahn et al., 2007). Due to this, the CLR transformation is often used for visualization (Pawlowsky-Glahn et al., 2007). The ILR transformation overcomes the issue of the singular covariance matrix from the CLR transformation (Calle, 2019); however, it does not allow us to infer the relationship between individual features in the dataset (Pawlowsky-Glahn et al., 2007; Quinn et al., 2019).

Here, we will focus on the additive logistic normal multinomial (LNM) model (Xia et al., 2013) that uses the ALR transformation and develop a model-based clustering approach that utilizes a finite mixture model. A finite mixture model assumes that the data come from a finite collection of subpopulations or components where each subpopulation can be represented by a distribution function and the appropriate distribution is chosen depending on the nature of the data. Several model-based clustering frameworks have been proposed for microbiome data (Holmes et al., 2012; Subedi et al., 2020; Fang & Subedi, 2023).

In the LNM model, the observed counts are modelled using a multinomial distribution conditional on the compositions $p$, and a Gaussian prior is imposed on the log-ratio transformed compositions. While this approach brings flexibility in modelling the data, the posterior distribution of the transformed variable does not have a closed-form solution. A Markov chain Monte Carlo (MCMC) approach is typically utilized for parameter estimation (Xia et al., 2013; Äijö et al., 2018), which comes with a heavy computational cost. Recently, Fang and Subedi (2023) proposed a mixture of additive logistic normal multinomial (LNM) models to cluster microbiome data and proposed an alternate approach for parameter estimation that utilized variational Gaussian approximations (VGA; Wainwright & Jordan, 2008). VGA is an alternative parameter estimation framework that utilizes a computationally convenient Gaussian density to approximate a more complex but “true” posterior density. The complex
posterior distribution is approximated by minimizing the Kullback-Leibler (KL; Kullback & Leibler, 1951) divergence between the true and the approximating densities.

In the LNM model, the log-ratio transformed composition variable is assumed to be a multivariate Gaussian distribution, and hence, the number of free parameters in the covariance matrix of the transformed variable grows quadratically with the dimensionality. Various decompositions of the covariance matrices have been proposed for the Gaussian mixture models where constraints have been imposed on these components of the covariance matrices to reduce the number of free parameters (Celeux & Govaert, 1995; McNicholas & Murphy, 2008; Bouveyron & Brunet, 2012). One such approach incorporates a factor analyzer structure in the Gaussian mixture model framework (Ghahramani & Hinton, 1997; McLachlan & Peel, 2000b). The factor analysis model assumes that the observed high-dimensional data can be modelled using a small number of latent factors. In mixtures of factor analyzers, the number of free parameters in the covariance matrix is linear with dimensionality, and by choosing the number of latent factors to be sufficiently small, the number of free parameters in the covariance matrix can be greatly reduced. McNicholas and Murphy (2008) imposed further constraints on the components of the covariance matrices of the mixtures of factor analyzers resulting in a family of parsimonious Gaussian mixture models (PGMM). In this paper, we extend the mixture of logistic normal multinomial models for high-dimensional data by incorporating a factor analyzer structure in the latent space. We develop a variational variant of the alternating expectation conditional maximization for parameter estimation. The paper is structured as follows: Sect. 2 provides details of the logistic normal multinomial model and the finite mixture of logistic normal multinomial factor analyzers (LNM-FAs) along with details on parameter estimation; in Sects. 3 and 4, these models are applied to simulated and real datasets, respectively; and Sect. 5 concludes the paper.

2 Methodology

2.1 Additive Logistic Normal Multinomial Model

Consider microbiome count data on $K + 1$ taxa, where one taxon represents a group of populations of organisms that are phylogenetically related. Let $\mathbf{W} = (W_1, \ldots, W_{K+1})^T$ denote the random vector of counts of $K + 1$ microbial taxa, and $\mathbf{p} = (p_1, \ldots, p_{K+1})^T$ be the underlying composition of the microbial taxa such that $\sum_{k=1}^{K+1} p_k = 1$. In the additive logistic normal multinomial (LNM) model by Xia et al. (2013), the observed counts $\mathbf{w}$ are modelled using a multinomial distribution such that

$$f(\mathbf{w}|\mathbf{p}) \propto \prod_{k=1}^{K+1} (p_k)^{w_k},$$

and an additive log-ratio (ALR) transformation is utilized to map $\mathbf{p}$ from the restricted simplex $S^K$ to a $K$-dimensional open real space $\mathbb{R}^K$ such that

$$\mathbf{Y} = \phi(\mathbf{p}) = \left[ \log \left( \frac{p_1}{p_{K+1}} \right), \ldots, \log \left( \frac{p_K}{p_{K+1}} \right) \right]^T,$$

(1)
where $p_{K+1}$ is used as a reference and a multivariate Gaussian distribution is imposed with mean $\mu$ and covariance $\Sigma$ on $Y$. Here, $\phi : (0, 1)^K \rightarrow \mathbb{R}^K$ is a one-to-one function, and therefore,

$$p = \phi^{-1}(Y) = \left[ \frac{\exp(Y_1)}{\sum_{k=1}^{K} \exp(Y_k) + 1}, \ldots, \frac{\exp(Y_K)}{\sum_{k=1}^{K} \exp(Y_k) + 1}, 1 \right]^\top.$$  

Thus, the conditional probability function of $W \mid Y$ is

$$f(w \mid y) \propto \prod_{k=1}^{K} \left\{ \frac{\exp(y_k)}{\sum_{k=1}^{K} \exp(y_k) + 1} \right\}^{w_k} \left\{ \frac{1}{\sum_{k=1}^{K} \exp(y_k) + 1} \right\}^{w_{K+1}},$$

and the marginal probability function of $W$ becomes

$$f(w \mid \mu, \Sigma) = \int_{\mathbb{R}^K} f(w \mid y) f(y \mid \mu, \Sigma) \, dy$$

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

Note that this marginal probability function of $W$ involves multiple integrals and cannot be further simplified. Although the LNM model provides flexibility in the modelling structure, parameter estimation thus far has mostly relied on Bayesian MCMC-based approaches that come with a heavy computational burden (Xia et al., 2013). Recently, Fang and Subedi (2023) proposed mixtures of the logistic normal multinomial models (LNM-MM) for clustering microbiome data where a computationally efficient framework for parameter estimation was developed using variational Gaussian approximations.

Using an approximating density $q(y)$, the marginal log density of $W$ can be written as follows:

$$\log f(w) = \int q(y) \log \frac{q(y)}{f(y \mid w)} \, dy + \int q(y) \log \frac{f(w, y)}{q(y)} \, dy$$

$$= D_{KL} [q(y) \| f(y \mid w)] + F(q(y), w),$$

where $D_{KL} [q(y) \| f(y \mid w)]$ is the KL divergence from $f(y \mid w)$ to $q(y)$ and $F(q(y), w)$ is known as the evidence lower bound (ELBO; (Blei et al., 2017)). Then, minimizing the KL divergence from $f(y \mid w)$ to $q(y)$ is equivalent to maximizing the ELBO. In a VGA framework, $q(y)$ is taken to be a Gaussian distribution. If we assume $q(y)$ to be a Gaussian distribution with mean $m$ and diagonal covariance matrix $V$, the lower bound of $F(q(y), w)$ becomes

$$\tilde{F}(m, V, \mu, \Sigma) = C + \underline{w}^\top m - \left( \sum_{k=1}^{K+1} w_k \right) \left[ \log \left( \sum_{k=1}^{K} \exp \left( m_k + \frac{\nu_k^2}{2} \right) + 1 \right) \right] +$$

$$\frac{1}{2} \log |V| + \frac{K}{2} - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (m - \mu)^\top \Sigma^{-1} (m - \mu) - \frac{1}{2} \text{tr}(\Sigma^{-1} V),$$

where $\underline{w}$ is a $K$-dimensional vector with the first $K$ elements of $w$, $\nu_k^2$ is the $k^{th}$ diagonal entry of $V$, and $C$ is a constant. Details of the derivation of this lower bound are provided in Appendix A. This lower bound can be maximized with respect to the model parameters and the variational parameters using an iterative approach. Thus, the use of the VGA eliminates the need for an MCMC-based approach for parameter estimation and drastically reduces the computational burden.
computational overhead making it feasible to extend these models for clustering in a high-dimensional setting. Several studies have shown that VGA delivers accurate approximations (Archambeau et al., 2007; Arridge et al., 2018; Challis & Barber, 2013; Subedi & Browne, 2020).

2.2 Mixtures of Logistic Normal Multinomial Factor Analyzers

A $G$-component finite mixture of LNM models can be written as

$$f(w_i | \vartheta) = \sum_{g=1}^{G} \pi_g f(w_i | \mu_g, \Sigma_g),$$

where $f(w_i | \mu_g, \Sigma_g)$ represents the marginal probability mass function of the logistic normal multinomial model of $i^{th}$ ($i = 1, \ldots, n$) observation in the $g^{th}$ subpopulation, each subpopulation has sample size $n_g (\sum_{g=1}^{G} n_g = n)$, $\pi_g > 0$ is the mixing proportion of the $g^{th}$ subpopulation such that $\sum_{g=1}^{G} \pi_g = 1$, and $\vartheta$ represents all the model parameters. The likelihood of the mixture is

$$L(\vartheta) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g f(w_i | \mu_g, \Sigma_g).$$

(5)

We define a group membership indicator variable $z_i = (z_{i1}, \ldots, z_{iG})$ such that $z_{ig} = 1$ if observation $i$ belongs to group $g$ ($g = 1, \ldots, G$) and 0 otherwise. In the context of clustering, these group memberships are treated as unobserved or missing data and the likelihood function in (5) is considered an incomplete-data likelihood function.

Then, the complete-data likelihood with observed data ($w_1, \ldots, w_n$) and missing data ($z_1, \ldots, z_n$) can be written as

$$L(\vartheta) = \prod_{i=1}^{n} \prod_{g=1}^{G} \{\pi_g f(w_i | \mu_g, \Sigma_g)\}^{z_{ig}}.$$

Now, the complete-data log-likelihood becomes

$$l(\vartheta) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \left\{ \log \pi_g + \log f(w_i | \mu_g, \Sigma_g) \right\}.$$

For incorporating a factor analyzer structure (Ghahramani & Hinton, 1997; McLachlan & Peel, 2000b) in the mixtures of LNM models, we utilize the following structure on $Y_i$ from the $g^{th}$ component:

$$Y_i = \mu_g + \Lambda_g U_{ig} + \epsilon_{ig},$$

where $\mu_g$ is a $K$-dimensional mean vector, $U_{ig} \sim N(0, I_q)$ is $q$-dimensional vector of latent factors where $q \ll K$, $\Lambda_g$ is a $K \times q$ matrix of factor loadings, $\epsilon_{ig} \sim N(0, D_g)$ is a $K$-dimensional vector of errors where $D_g$ is diagonal matrix, and $U_{ig} \perp \epsilon_{ig}$. Thus, for the $g^{th}$ component, $Y_i | z_{ig} = 1 \sim N(\mu_g, \Lambda_g \Lambda_g^T + D_g)$ and $Y | z_{ig} = 1, U_{ig} \sim N(\mu_g + \Lambda_g u_{ig}, D_g)$. The number of free parameters in the covariance matrix $\Sigma_g$ in a LNM-MM model is $K(K+1)/2$ while the number of free parameters in the covariance matrix of $Y$ in LNM-FA is $Kq - q(q-1)/2 + K$. Hence, by choosing $q \ll K$, the number of free parameters in the covariance of $Y$ can be greatly reduced in the LNM-FA framework. To introduce
further parsimony, McNicholas and Murphy (2008) proposed constraints on $\Lambda_g$ and $D_g$ and imposition on various combinations of these constraints resulted in a family of eight models (discussed in detail in Sect. 2.4).

2.3 Parameter Estimation

Parameter estimation of the mixtures of factor analyzers is typically done using an alternating expectation conditional maximization (AECM) algorithm (Meng & Van Dyk, 1997), which is an extension of the expectation-maximization (EM) algorithm (Dempster et al., 1977). The AECM algorithm uses different specifications of missing data at different cycles and the maximization step comprises a series of conditional maximizations. McLachlan and Krishnan (2007) showed that the rate of convergence of the EM algorithm is dependent on the proportion of the missing data—the rate of convergence decreases as the proportion of missing information increases. Through cycle-specific missing data specification, the AECM algorithm provides a faster rate of convergence and maintains the stability of the EM algorithm (McLachlan & Krishnan, 2007). In the mixtures of LNM-FA, we have missing data from two sources: the unobserved latent variables $U$ and $Y$ and the unobserved component indicator variable $Z$. Each cycle of the AECM algorithm consists of an E-step in which the expected value of the complete-data log-likelihood is computed with respect to cycle-specific missing data, which is then followed by a conditional maximization step where a subset of the model parameters is updated. Parameters that are updated by conditional maximization depend on each other, so we will use superscript $(t)$ to denote the $t^{th}$ iteration. We will develop a variational version of the AECM algorithm that uses different specifications of the missing data at different cycles.

**First Cycle**

In the first cycle, we utilize the following hierarchical structure:

$$W_i | Y_i \sim \text{Multinomial}(p_i) \quad \text{and} \quad Y_i \sim N(\mu_g, \Lambda_g \Lambda_g^\top + D_g).$$

where $p_i$ can be obtained from $Y_i$ using (2). Then, the component-specific marginal probability function of the observed data $w_i$ is

$$f(w_i | \mu_g, \Lambda_g, D_g) = \int_{\mathbb{R}^K} f(w_i | y_i) f(y_i | \mu_g, \Lambda_g \Lambda_g^\top + D_g) \, dy$$

$$\propto \int_{\mathbb{R}^K} \prod_{k=1}^{K+1} |\phi^{-1}(y_i)_k|^{w_{ik}} |\Lambda_g \Lambda_g^\top + D_g|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (y_i - \mu_g)^\top (\Lambda_g \Lambda_g^\top + D_g)^{-1} (y_i - \mu_g) \right\} dy.$$  

Assuming $Z$ and $Y$ as missing variables, the complete-data log-likelihood using the marginal probability function of $W$ is

$$l(\vartheta | w_i) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \left\{ \log \pi_g + \log f(w_i | \mu_g, \Lambda_g \Lambda_g^\top + D_g) \right\}$$

$$= \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \left\{ \log \pi_g + \log \int f(w_i | y_i) f_g(y_i | \mu_g, \Lambda_g \Lambda_g^\top + D_g) \, dy \right\}.$$  

Assuming the component-specific $q(y)$ to be a Gaussian distribution with mean $m_g$ and diagonal covariance matrix $V_g$ and replacing the log of the marginal of the component prob-
ability function by the component-specific $\tilde{F}(m_{ig}, V_{ig}, \mu_g, \Sigma_g)$, the variational Gaussian lower bound of the complete-data log-likelihood can be written as

$$
\tilde{L}_1 = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \left\{ \log \pi_g - (1_{(K+1)} w_i) \left[ \log \left( 1_{(K)} \exp \left( \frac{m_{ig} + \text{diag}(V_{ig})}{2} \right) + 1 \right) \right] + C_i + w_i^T m_{ig} + \frac{1}{2} \log |V_{ig}| + \frac{K}{2} - \frac{1}{2} \log |\Sigma_g| - \frac{1}{2} \text{tr}(\Sigma_g^{-1} V_{ig}) - \frac{1}{2} (m_{ig} - \mu_g)^T \Sigma_g^{-1} (m_{ig} - \mu_g) \right\},
$$

where $1_{(K)}$ stands for column vector of 1’s with dimension $K$, $C_i$ stands for $\log \prod_{k=1}^{K} w_{ik}$, $\text{diag}(V_{ig}) = (v_{ig,11}^2, v_{ig,22}^2, \ldots, v_{ig, KK}^2)$ puts the diagonal elements of the $K \times K$ matrix $V_{ig}$ into a $K$-dimensional vector, and $\Sigma_g = \Lambda_g \Lambda_g^T + D_g$. In this cycle, for the parameter updates in the $(t+1)^{th}$ iteration, the following steps are conducted:

1. Update the variational Gaussian lower bound of the complete-data log-likelihood from the first cycle $\tilde{L}_1$ by updating $m_{ig}$ and $V_{ig}$. For updating $V_{ig}^{(t+1)}$, we use the Newton-Raphson method. We take the derivative with respect to standard deviation vector $v_{ig}^{(t+1)}$ and find the solution to the following score function:

$$
\frac{\partial \tilde{L}_1}{\partial v_{ig}} = v_{ig}^{(t)} - v_{ig}^{(t) \text{diag}(\Sigma_g^{(t)-1})} - (1_{(K+1)} w_i) v_{ig}^{(t)} \frac{\exp(m_{ig}^{(t)} + (v_{ig}^{(t)})^2)}{1_{(K)} \exp(m_{ig}^{(t)} + (v_{ig}^{(t)})^2) + 1}.
$$

For updating $m_{ig}^{(t+1)}$, we again use the Newton-Raphson method to find the solution to the following score function:

$$
\frac{\partial \tilde{L}_1}{\partial m_{ig}} = w_i^* - \Sigma_g^{(t)-1} (m_{ig}^{(t)} - \mu_g^{(t)}) - (1_{(K+1)} w_i) \frac{\exp(m_{ig}^{(t)} + (v_{ig}^{(t)})^2)}{1_{(K)} \exp(m_{ig}^{(t)} + (v_{ig}^{(t)})^2) + 1}.
$$

2. Update the component indicator variable $Z_{ig}$. Conditional on the variational parameters $m_{ig}^{(t+1)}$, $V_{ig}^{(t+1)}$ and on $\mu_g^{(t)}$, $\Lambda_g^{(t)}$, and $D_g^{(t)}$, the expected value of $Z_{ig}$ can be computed as

$$
E(Z_{ig}^{(t+1)}) = \frac{\pi_g^{(t)} f(w_i \mid \mu_g^{(t)}, \Lambda_g^{(t)}, D_g^{(t)})}{\sum_{h=1}^{G} \pi_h^{(t)} f(w_i \mid \mu_h^{(t)}, \Lambda_h^{(t)}, D_h^{(t)})}.
$$

As this involves the marginal distribution of $W$, which is difficult to compute, we use an approximation of $E(Z_{ig}^{(t+1)})$ using the ELBO:

$$
\hat{Z}_{ig}^{(t+1)} = \frac{\pi_g^{(t)} \exp\{\tilde{F}(\mu_g^{(t)}, \Lambda_g^{(t)}, D_g^{(t)}; m_{ig}^{(t+1)}, V_{ig}^{(t+1)})\}}{\sum_{g=1}^{G} \pi_g^{(t)} \exp\{\tilde{F}(\mu_g^{(t)}, \Lambda_g^{(t)}, D_g^{(t)}; m_{ig}^{(t+1)}, V_{ig}^{(t+1)})\}}.
$$

This approximation has been previously used in the literature for approximating $E(Z_{ig}^{(t+1)})$ in the variational framework (Tang et al., 2015; Gollini & Murphy, 2014; Subedi & Browne, 2020).
3. Given the variational parameters and $\tilde{z}^{(t+1)}_{ig}$, we then update the parameters $\pi_g$ and $\mu_g$ as follows:

$$\hat{\pi}^{(t+1)}_g = \frac{\sum_{i=1}^n z^{(t+1)}_{ig}}{n}, \quad \text{and} \quad \hat{\mu}^{(t+1)}_g = \frac{\sum_{i=1}^n z^{(t+1)}_{ig} m^{(t+1)}_{ig}}{\sum_{i=1}^n z^{(t+1)}_{ig}}.$$ 

**Second Cycle**

In the second cycle, we utilize the following hierarchical structure:

$$W_i | Y_i \sim \text{Multinomial}(p_i), \quad Y_i | U_i = u_i \sim N(\mu_g + \Lambda_g u_i, D_g), \quad \text{and} \quad U_i \sim N(0, I_q),$$

where $p_i$ can be obtained from $Y_i$ using (2). Assuming $Z, Y$ and $U$ as missing variables, the complete-data log-likelihood using the marginal probability function of $W$ has the following form:

$$l_2(W, Z) = \sum_{i=1}^n \sum_{g=1}^G z_{ig} \left\{ \log \pi_g + \log \left[ \int f(w_i | y_i) f_g(y_i | \mu_g + \Lambda_g u_i, D_g) f_g(u_i | 0, I_q) \, dy \, du \right] \right\}.$$  

In this cycle, we derive an approximate lower bound for the log of the marginal probability function of $W$ using the approximating density $q(y, u)$

$$\log f(w) = \int q(y, u) \log \frac{q(y, u)}{f(y, u | w)} \, dy \, du + \int q(y, u) \log \frac{f(w, y, u)}{q(y, u)} \, dy \, du$$

$$= D_{KL} [q(y, u) || f(y, u | w)] + F(q(y, u), w), \quad (6)$$

where $F(q(y, u), w)$ is the ELBO and $D_{KL} [q(y, u) || f(y, u | w)]$ is the KL divergence from $f(y, u | w)$ to $q(y, u)$. Furthermore, assuming $q(y, u) = q(y)q(u)$, $q(u) = N(\tilde{m}_g, \tilde{V}_g)$, and $q(y) = N(m_{ig}, V_{ig})$, we can show that

$$F(q(y, u), w) \geq C + w_i^T m_{ig} - \left( I_{(K+1)} \right) \left[ \log \left( \frac{1}{1} \exp \left( m_{ig} + \text{diag}(V_{ig}) \right) + 1 \right) \right]$$

$$+ \frac{1}{2} \log |V_{ig}| + \log |\tilde{V}_g| + q + K \log |D_g| - m_{ig}^T \tilde{m}_g - \text{tr}(\tilde{V}_g)$$

$$- \text{tr}(D_g^{-1} (V_{ig} + (m_{ig} - \mu_g)^T (m_{ig} - \mu_g))) + 2(m_{ig} - \mu_g)^T D_g^{-1} \Lambda_g \tilde{m}_g$$

$$- m_{ig}^T \Lambda_g^T D_g^{-1} A_g \tilde{m}_g - \text{tr}(\Lambda_g^T D_g^{-1} A_g \tilde{V}_g))$$

$$= F_2(\mu_g, \Lambda_g, D_g, m_{ig}, V_{ig}, \tilde{m}_g, \tilde{V}_g).$$

Here, $m_{ig}$ and $V_{ig}$ are the variational parameters of $q(y_i)$ from first cycle, and $\tilde{m}_{ig}$ and $\tilde{V}_{ig}$ are the variational parameters of $q(u_i)$. Details of the derivation of the lower bound are provided in Appendix B. In this cycle, for the parameter updates in the $(t + 1)^{th}$ iteration, the following steps are conducted:

1. Update the variational Gaussian lower bound of complete-data log-likelihood of the second cycle $\tilde{L}_2$ by updating $\tilde{m}_{ig}^{(t+1)}$ and $\tilde{V}_g^{(t+1)}$ as

$$\tilde{m}_{ig}^{(t+1)} = (A_g^T D_g^{-1} A_g + I_q)^{-1} A_g^T D_g^{-1} (m_{ig}^{(t+1)} - \mu_g^{(t+1)}), \quad \text{and}$$

$$\tilde{V}_g^{(t+1)} = (A_g^T D_g^{-1} A_g + I_q)^{-1}.$$
2. Update the group indicator variable $Z$. Similar to the first cycle, we compute an approximation of $E(Z_{ig})$ using the ELBO from the second cycle:

\[
\bar{z}^{(t+1)}_{ig} = \frac{\pi_g^{(t+1)} \exp(F_2(\mu_g^{(t+1)} + \Lambda_g^{(t)} D_g^{(t)-1} m_{ig}^{(t+1)} + V_{ig}^{(t+1)} + \bar{m}_{ig}^{(t+1)} + \bar{\nu}_{ig}^{(t+1)}))}{\sum_{h=1}^{G} \pi_h^{(t+1)} \exp(F_2(\mu_h^{(t+1)} + \Lambda_h^{(t)} D_h^{(t)-1} m_{ih}^{(t+1)} + V_{ih}^{(t+1)} + \bar{m}_{ih}^{(t+1)} + \bar{\nu}_{ih}^{(t+1)}))}.
\]

3. Update $D_g^{(t+1)}$ and $\Lambda_g^{(t+1)}$ as

\[
\hat{D}_g^{(t+1)} = \text{diag}(\hat{\Sigma}_g^{(t+1)}) - 2A_g^{(t)} (A_g^{(t)} D_g^{(t)-1} A_g^{(t)} + I_q)^{-1} A_g^{(t)} D_g^{(t)-1} \hat{S}_g^{(t+1)} + A_g^{(t)} \hat{\theta}_g^{(t+1)} A_g^{(t)}),
\]

\[
\hat{\Lambda}_g^{(t+1)} = \hat{S}_g^{(t+1)} \hat{\beta}_g^{(t+1)} \hat{\theta}_g^{(t+1)}
\]

where

\[
\hat{S}_g^{(t+1)} = \frac{\sum_{i=1}^{n} \bar{z}_{ig}^{(t+1)} (m_{ig}^{(t+1)} - \mu_g^{(t+1)}) (m_{ig}^{(t+1)} - \mu_g^{(t+1)})}{\sum_{i=1}^{n} \bar{z}_{ig}^{(t+1)}},
\]

\[
\hat{\Sigma}_g^{(t+1)} = \frac{\sum_{i=1}^{n} \bar{z}_{ig}^{(t+1)} \left[ V_{ig}^{(t+1)} + (m_{ig}^{(t+1)} - \mu_g^{(t+1)})(m_{ig}^{(t+1)} - \mu_g^{(t+1)})^\top \right]}{\sum_{i=1}^{n} \bar{z}_{ig}^{(t+1)}},
\]

\[
\hat{\theta}_g^{(t+1)} = (A_g^{(t)} \hat{D}_g^{(t)-1} A_g^{(t)} + I_q)^{-1} + \hat{\beta}_g^{(t+1)} \hat{S}_g^{(t+1)} \hat{\beta}_g^{(t+1)} \hat{\theta}_g^{(t+1)}
\]

and

\[
\hat{\beta}_g^{(t+1)} = (A_g^{(t)} \hat{D}_g^{(t)-1} A_g^{(t)} + I_q)^{-1} A_g^{(t)} \hat{D}_g^{(t)-1}.
\]

Overall, our algorithm consists of the following steps:

I. Specify the number of clusters $G$, latent dimension $q$, and initial starting value for $A_g$, $D_g$, and $Z_{ig}$.

II. First cycle:

1) Update the variational Gaussian lower bound of complete-data log-likelihood of the first cycle by estimating $V_{ig}$ and $m_{ig}$.
2) Update $Z_{ig}$.
3) Update $\pi_g$ and $\mu_g$.

III. Second cycle:

1) Update the variational Gaussian lower bound of complete-data log-likelihood of the second cycle by estimating $\bar{V}_{ig}$ and $\bar{m}_{ig}$.
2) Update $Z_{ig}$ again.
3) Update $S_g$, $\Sigma_g$, $D_g$, and $A_g$.

IV. Compute the likelihood $\sum_{i=1}^{n} \log \sum_{g=1}^{G} \pi_g f(w_i | \theta_g)$ using the current estimates and check for convergence. If it is converged, then stop; otherwise, go to step II.

The convergence of the algorithm is determined using a criterion based on Aitken’s acceleration. Aitken’s acceleration (Aitken, 1926) is defined as follows:

\[
d^{(r)} = \frac{l^{(t+1)} - l^{(t)}}{l^{(t)} - l^{(t-1)}},
\]
where \( I^{(t+1)} \) stands for the log-likelihood values at \( t + 1 \) iteration. Then, the asymptotic estimate for log-likelihood at iteration \( t + 1 \) is as follows:

\[
I^{(t+1)} = I^{(t)} + \frac{I^{(t+1)} - I^{(t)}}{1 - a^{(t)}}.
\]

The algorithm can be considered as converged when

\[
|I^{(t+1)} - I^{(t)}| < \epsilon,
\]

where \( \epsilon \) is a small number (Böhning et al., 1994). Here, similar to McNicholas and Murphy (2008), we set \( \epsilon = 10^{-2} \).

Note that \( \Lambda_g \) is unidentifiable. This can be seen if we let \( \Lambda^*_g = \Lambda_g T \) be a new factor loading matrix where \( T \) be an orthonormal matrix, i.e., \( TT^T = I \), then \( \Lambda^*_g \Lambda^*_g^T + D_g = \Lambda_g TT^T \Lambda_g^T + D_g = \Lambda_g \Lambda_g^T + D_g \). Hence, we focus on the recovery of \( \Sigma_g = \Lambda_g \Lambda_g^T + D_g \) which is identifiable. Additionally, by incorporating a factor structure, McNicholas and Murphy (2010) show that the Woodbury identity (Woodbury, 1950) can be utilized to compute \( \Sigma_g^{-1} \):

\[
\Sigma_g^{-1} = (\Lambda_g \Lambda_g^T + D_g)^{-1} = D_g^{-1} - D_g^{-1} \Lambda_g (I_g + \Lambda_g^T D_g^{-1} \Lambda_g)^{-1} \Lambda_g^T D_g^{-1},
\]

and thus, the matrix inversion is \( O(q^3) \) as opposed to \( O(K^3) \). Therefore, when \( q \ll K \), inverting \( \Sigma_g \) is computationally efficient.

### 2.4 A Family of Mixture Models for Clustering

Introducing the factor analyzer structure and assuming the latent number of components \( q \) is small reduces the number of free parameters in the covariance matrix of \( Y \). However, in a clustering context, we estimate \( G \) different covariance matrices. Thus, to introduce further parsimony, we allowed the imposition of constraints on the parameters of the covariance matrix of \( Y \) to be equal across groups (i.e., \( \Lambda_1 = \Lambda_2 = \cdots = \Lambda_g = \Lambda \) and/or \( D_1 = D_2 = \cdots = D_g = D \)) and imposition of isotropic constraints on \( D_g \) (i.e., \( D_g = d_g I \)). Various combinations of these constraints result in a family of eight parsimonious LNM-FAs (Table 1). Similar constraints on the components of the covariance matrices were utilized by McNicholas and Murphy (2008) and Subedi et al. (2013, 2015).

| Model | \( \Lambda_g \) Group | \( D_g \) Group | Diagonal | Total Par |
|-------|----------------|----------------|-----------|----------|
| “UUU” | U | U | U | \( G(Kq - q(q - 1)/2 + KG + G - 1 + K \) |
| “UUC” | U | U | C | \( G(Kq - q(q - 1)/2 + G + G - 1 + K \) |
| “UCU” | U | C | U | \( G(Kq - q(q - 1)/2 + K + G - 1 + K \) |
| “UCC” | U | C | C | \( G(Kq - q(q - 1)/2 + 1 + G - 1 + K \) |
| “CUU” | C | U | U | \( Kq - q(q - 1)/2 + KG + G - 1 + K \) |
| “CUC” | C | U | C | \( Kq - q(q - 1)/2 + G + G - 1 + K \) |
| “CCU” | C | C | U | \( Kq - q(q - 1)/2 + K + G - 1 + K \) |
| “CCC” | C | C | C | \( Kq - q(q - 1)/2 + 1 + G - 1 + K \) |
In Table 1, the column “Group” refers to constraints across groups, the column “Diagonal” refers to the matrix having the same diagonal elements, and the letters refer to whether or not the constraints were imposed on the parameters such that “U” stands for unconstrained and “C” stands for constrained.

For example, the model “UCU” refers to unconstrained $A_g$ but constrained $D_g = D$ whereas model “UCC” refers to constraints on both the “Group” and the “Diagonal” such that $D_1 = D_2 = \cdots = D_G = d_{IK}$. The advantages of imposing such constraints include a dramatic reduction of parameters in some models and better stability of the fitted model. McNicholas and Murphy (2008) mentioned that the constrained model (CUU, CUC, CCC, CCU) could dramatically reduce the number of covariance parameters. An equal noise model (i.e., UCU) could lead to more stable results (McLachlan & Peel, 2000b; McLachlan et al., 2003). Details of the parameter estimates for the LNM-FA family are provided in Appendix C, and details of initialization of the parameters are provided in Appendix D.

### 2.5 Model Selection and Performance Assessment

Note that in a clustering context, the true number of clusters ($G$) is unknown. When implementing approaches based on the mixtures of factor analyzers, the number of latent factors ($q$) is also unknown. We suggest running the algorithm for a range of $G$ and range of $q$ and selecting the optimal ($G, q$) pair using model selection criteria. Here, we use the Bayesian information criterion (BIC; Schwarz, 1978), which is considered to be consistent and efficient in practice under certain regularity conditions (Keribin, 2000; Fraley & Raftery, 1998). The BIC is the most popular choice for model selection in model-based clustering and is defined as

$$BIC = \psi \log(n) - 2l(w, \hat{\theta}),$$

where $l(w, \hat{\theta})$ is the log-likelihood evaluated using the estimated $\hat{\theta}$, $\psi$ is the number of free parameters, and $n$ is the number of observations. The model with the lowest BIC is selected as the best-fitting model.

In selecting the maximum number of clusters to be fitted ($G_{\text{max}}$), one may want to consider the sample size of the dataset as well. Fitting a model with too many components on a small dataset might result in empty clusters or spurious clusters with very few observations. On the other hand, if the best-fitting model is selected to be a $G_{\text{max}}$-component model, we recommend increasing the value of $G_{\text{max}}$ and running the algorithm again. For the maximum number of latent variables to be fitted ($q_{\text{max}}$), McLachlan and Peel (2000a) suggest selecting $q_{\text{max}}$ such that the difference $C = \frac{1}{2} \{(K - q_{\text{max}})^2 - (K + q_{\text{max}})\}$ is positive, where $C$ is the difference in the number of free parameters in the covariance of the latent variable with and without assuming a factor analyzer structure. If we set $q_{\text{max}} < 0.5 \times K$ for $K \geq 4$, this difference is always positive, and for $K < 4$, the difference is non-negative.

When the true labels are known, the adjusted Rand index (ARI; Hubert & Arabie, 1985) is used for performance assessment. For perfect agreement, the ARI is 1 while the expected value of ARI is 0 under random classification.

### 3 Simulation Studies

In this section, we use simulation studies to demonstrate the clustering performance and parameter recovery of the proposed LNM-FA models. For Simulation Study 1 and 2, we first generate $Y$ from a multivariate normal distribution, then transform the data into composition
pusing the additive log-ratio transformation. Count data were then generated based on a multinomial distribution with composition p with the total count for each observation being generated from a uniform distribution U (5000, 10, 000). For Simulation Study 1, we generated 100 datasets from the most constrained model, and for Simulation Study 2, we generated 100 datasets from the most flexible model. We fitted all eight LNM-FA models for a range of values for G and q, and the best-fitting model was chosen a posteriori using BIC. In Simulation Study 3, we generated datasets from a higher-dimensional three-component multinomial model with 100 dimensions. For comparison, we ran the LNM-MM and DMM models on all 100 datasets from all three simulation studies. We fitted models with G = 1, . . . , 5. We also compared our approach with non-parametric methods (k-means and hierarchical clustering) to the simulation study. For both k-means and hierarchical clustering, we simply fix G to be the true G for the optimal result and use Canberra distance as a distance measure for hierarchical clustering.

Visualizations of the cluster structures in the latent space of one of the 100 datasets are provided in Appendix E.1 and Appendix E.2. The true values of the parameters π and μ are provided in Appendix F. As Λ is not identifiable but Σ = ΛΛ−1 + D is identifiable, we demonstrate the recovery of Σ. The true value of Λ and D for Σ is provided in the Appendix F, and the mean square error (MSE, defined as the average of the square difference between estimated and true value) is provided in Table 2. The overall average ARI calculated using the selected models for all 100 datasets and using only the datasets/runs in which the correct number of components have been selected are both provided in Table 2.

• Simulation 1
Here, each dataset is generated with size n = 1000 and K = 10 from the most constrained model “CCC” with G = 3 and q = 3. We ran all 8 models in the LNM-FA family for G = 1, . . . , 5 and q = 1, . . . , 5 and selected the best model using the BIC. The average ARI across all selected models was 0.998 (standard deviation (sd) of 0.01). In 96 out of 100 times, the BIC selected the true “CCC” model with G = 3 and q = 3 with an average ARI of 0.999 (sd of 0.003). In 81 out of the 100 datasets, the BIC selected a three-component LNM-MM model with an average ARI of 0.99 (sd of 0.00) and a four-component model in 13 of the datasets. The LNM-MM model encountered computational issues in 6 out of the 100 datasets. On the other hand, a five-component DMM was selected by the BIC in all 100 datasets.

• Simulation 2
Here, each dataset is generated with size n = 1000 and K = 10 from the most flexible model “UUU” with G = 3 and q = 3. We ran all 8 models in the LNM-FA family for G = 1, . . . , 4 and q = 1, . . . , 6 and selected the best model using the BIC. In all 100 out of the 100 datasets, the BIC selected the true “UUU” model with G = 3 and q = 3, with an average ARI of 1 (sd of 0). From the LNM-MM family, the BIC selected a three-component model in 12 out of the 100 datasets with perfect classification but four and five-component models in 70 and 9 datasets, respectively. The LNM-MM implementation encountered singularities in the remaining 9 datasets. On the other hand, a five-component model is selected by the DMM for all 100 datasets.

• Simulation 3
Similar to Simulation 1 and 2, each dataset is generated with size n = 1000 and K = 100 from the most flexible “UUU” with G = 3 and q = 5. We ran all 8 models in the LNM-FA family for G = 1, . . . , 4 and q = 1, . . . , 6 and selected the best model using the BIC. The average ARI across all selected models was 0.998 (sd of 0.00). In 91 out of the 100 datasets, the BIC selected the true “UUU” model with G = 3 and q = 5 with an average
Table 2  Summary of parameter recovery and clustering performance on all three simulated datasets

|                | Simulation 1 | Simulation 2 (n₁ = 500) | Simulation 3 |
|----------------|-------------|------------------------|--------------|
| **Component 1** |             |                        |              |
| Sum of MSE(μ₁) | 0.0081      | 0.0081                 | 0.54         |
| π₁             | 0.0002      | 0.0002                 | 0.0002       |
| For Simulations 2 and 3 | -           | 0.0364                 | 7.79         |
| Sum of MSE(Σ₁) |             |                        |              |
| Sum of MSE(μ₂) | 0.0121      | 0.0094                 | 0.62         |
| π₂             | 0.0002      | 0.0002                 | 0.0002       |
| For Simulations 2 and 3 | -           | 0.0390                 | 0.40         |
| Sum of MSE(Σ₂) |             |                        |              |
| Sum of MSE(μ₃) | 0.0187      | 0.0028                 | 0.11         |
| π₃             | 0.0001      | 0.0001                 | 0.0002       |
| For Simulations 2 and 3 | -           | 0.0018                 | 0.36         |
| Note that for “CCC” model, all components have the same Σ | -            | -                      | -            |
| For Simulation 1 |             |                        |              |
| Sum of MSE(Σ)  | 0.0159      | -                      | -            |
| **Component 2** |             |                        |              |
| Sum of MSE(μ₂) | 0.0121      | 0.0094                 | 0.62         |
| π₂             | 0.0002      | 0.0002                 | 0.0002       |
| For Simulations 2 and 3 | -           | 0.0390                 | 0.40         |
| Sum of MSE(Σ₂) |             |                        |              |
| Sum of MSE(μ₃) | 0.0187      | 0.0028                 | 0.11         |
| π₃             | 0.0001      | 0.0001                 | 0.0002       |
| For Simulations 2 and 3 | -           | 0.0018                 | 0.36         |
| Note that for “CCC” model, all components have the same Σ | -            | -                      | -            |
| For Simulation 1 |             |                        |              |
| Sum of MSE(Σ)  | 0.0159      | -                      | -            |
| **Component 3** |             |                        |              |
| Sum of MSE(μ₃) | 0.0187      | 0.0028                 | 0.11         |
| π₃             | 0.0001      | 0.0001                 | 0.0002       |
| For Simulations 2 and 3 | -           | 0.0018                 | 0.36         |
| Note that for “CCC” model, all components have the same Σ | -            | -                      | -            |
| For Simulation 1 |             |                        |              |
| Sum of MSE(Σ)  | 0.0159      | -                      | -            |
| **Summary of cluster recovery and clustering performance** |             |                        |              |
| LNM-FA         |             |                        |              |
| Number of times correct G selected | 96 out of 100 | 100 out of 100 | 91 out of 100 |
| Overall average ARI | 0.99 | 1 | 0.99 |
| Average ARI when correct G is selected | 0.99 | 1 | 0.99 |
| Method  | Simulation 1 | Simulation 2 | Simulation 3 |
|---------|--------------|--------------|--------------|
|         | Component 1 ($n_1 = 500$) |               |              |
| LNM-MM  | Number of times correct G selected | 81 out of 100 | 12 out of 100 | - |
|         | Overall average ARI | 0.98 | 0.77 | - |
|         | Average ARI when correct G is selected | 0.99 | 1 | - |
| DMM     | Number of times correct G selected | 0 out of 100 | 0 out of 100 | 0 out of 100 |
|         | Overall average ARI | 0.00 | 0.27 | 0.20 |
|         | Average ARI when correct G is selected | NA | NA | NA |
| $k$-means | Average ARI (true $G$ was provided) | 0.11 | 0.66 | 0.15 |
| Hierarchical | Average ARI (true $G$ was provided) | 0.06 | 0.15 | 0.09 |
ARI of 0.998 (standard deviation (sd) of 0.003). LNM-MM encountered singularities in all 100 datasets, and DMM selected 4 components model for all 100 datasets

4 Real Data Analysis

We applied our method to two publicly available microbiome datasets.

**Martinez dataset:** The microbiome dataset Martinez (Martínez et al., 2015) is available from R (R Core Team, 2023) package MicrobiomeCluster (Shi, 2020). In order to have a better understanding of the diversity and composition of the gut microbiome, Martínez et al. (2015) compared the fecal microbiota of adults between Papua New Guinea (PNG) and the USA. The results show that compared to the US residents, PNG has greater bacterial diversity, lower inter-individual variation, and vastly different abundance profiles. The data consists of 62 samples comprising 40 participants from PNG and 22 subjects from the USA, and counts of 10,227 OTUs were available.

**Dietswap dataset:** The microbiome dataset Dietswap (O’Keefe et al., 2015) is available from R package Microbiome (Lahti & Shetty, 2012-2019). Colorectal cancer is the third most prevalent cancer worldwide (Garrett, 2019). The rate of colon cancer in Americans of African descent is much higher than compared to rural Africans (O’Keefe et al., 2015). Recent findings indicate that the risk of colon cancer may be associated with dietary habits that affect the gut microbiota (Garrett, 2019). To investigate diet-associated cancer risk, O’Keefe et al. (2015) collected fecal samples from healthy middle-aged 20 Africans (AFR) and 20 African Americans (AAM). Fecal samples were taken at 6 different time points: the first three measurements (i.e., day 0, day 7, and day 14) were taken in their home environment with their regular dietary habits, and the last three measurements (i.e., day 15, day 22, and day 29) were taken after an intervention diet. As repeated measurements at different time points are taken on the same individuals and our model currently cannot model that (violates the independence assumption), we only utilize the measurements at day 0. Hence, the resulting dataset comprises 38 individuals from day 0, and we focus our analysis at the genus level resulting in 130 genera.

As microbiome data can be highly sparse at a lower taxonomic level (e.g., genus level), we recommend first filtering rare taxa (i.e., taxa present in less than 10 or 20% of the samples) to reduce sparsity and use the resulting taxa for cluster analysis. Depending on the size of the dataset, the number of taxa can be further reduced by focusing on the top “x” most abundant taxa. First, for both datasets, we removed taxa that were present in less than 20% of the samples which resulted in 20 and 99 genera for Martinez and Dietswap datasets, respectively. Then, we use the top 20 most abundant genera from the 99 genera from Dietswap and all genera from Martinez. To preserve the relative abundance, the remaining genera are aggregated in the category “Others,” which is then used as the reference level for the additive log-ratio transformation. Then, we ran all 8 models from our mixtures of LNM-FA family for \(G = 1, \ldots, 3\) and \(q = 1, \ldots, 5\) for both Martinez and Dietswap datasets. The BIC was used to select a best-fitting model. For comparison, we also ran the mixtures of LNM models without the factor structure and the DMM model on both datasets for \(G = 1, \ldots, 3\). The classification results from all three approaches are summarized in Table 3.

For both datasets, our proposed LNM-FA was able to recover the underlying groups. Our proposed approach outperformed DMM in both the Martinez and Dietswap datasets. While fitting the LNM-MM model to the real datasets with 20 taxa and small sample sizes \((n = 62\) and \(n = 38\) for Martinez and Dietswap, respectively), \(\Sigma\) became singular,
while the LNM-FA could be fitted due to the computational advantage that comes with the incorporation of factor analyzer structure. Note that the DMM model could be fitted to both datasets; however, the clustering performance of the DMM model was inferior compared to the LNM-FA model. The DMM model accounts for overdispersion by utilizing a Dirichlet prior to the multinomial parameter $p$. However, as noted by Aitchison (1982) and Xia et al. (2013), the logistic normal multinomial distribution allows for a more flexible covariance structure than the DMM model.

### 5 Conclusion

Here, we extended the additive logistic normal multinomial mixture model for high-dimensional data by incorporating a factor analyzer structure. A family of eight mixture models was proposed by imposing constraints on the components of the covariance matrix of the latent variable. Due to the incorporation of the factor analyzer structure, the number of free parameters is now linear in the dimensionality of the latent variable as opposed to the additive logistic normal multinomial mixture model where the number of free parameters grows quadratically. Through simulation studies, we demonstrated that our proposed approach provides excellent clustering performance and parameter recovery. Imposing a factor analyzer structure allows us to work on a lower dimension $q$ compared to $K$, and thus, the number of free parameters in the covariance matrix is greatly reduced when $q$ is chosen to be sufficiently smaller than $K$. Additionally, the use of Woodbury identity provides additional computational advantages. For real data analysis, our approach outperforms the DMM model while the LNM-MM by Fang and Subedi (2023) (i.e., the additive logistic multinomial mixture model without a factor analyzer structure) could not be fitted due to computational issues.
Gloor et al. (2017) state that the choice of the invariant taxon (i.e., reference taxon) when using ALR transformation can have a large effect on the result, if the presumed invariant taxon is not invariant or is correlated with other taxa. To ensure that the reference taxon has little to no correlation, in real data analysis, we used the taxa “Others” (i.e., the category comprising of the sum of all taxa that were not among the top 20 abundant taxa) as the reference category. Some future work will focus on the impact of the choice of reference taxa. Additionally, some feature selection techniques (for example, differential abundance analysis) could be applied prior to cluster analysis to eliminate redundant or irrelevant taxa and improve clustering performance. Several studies have shown that such an approach can provide a better clustering performance (Sørlie et al., 2001; Chipman et al., 2003; Baek & McLachlan, 2011).

The AECM algorithm and variational approximations are both highly dependent on their initial values. An unstable initial guess may lead the algorithm to converge to a local maximum or increase computation time. While in our simulations, the use of variational approximations provided good clustering performance and parameter recovery, the assumption that the approximating densities are independent can underestimate the posterior variance of the latent variables (Blei et al., 2017). A comprehensive investigation of the optimization strategy with respect to initialization and the impact of the underestimation of the posterior variances is warranted. When working on a deeper taxonomic level, modelling excessive 0 counts in microbiome data is also necessary as microbiome datasets are very sparse at a deeper taxonomic level. While our approach can deal with the high-dimensional nature of the microbiome data, caution is warranted while selecting a taxa reference group. Our simulation studies suggest that choosing an arbitrary taxon as the reference group could have a large impact on the clustering performance of the model especially when the sample size is small. Some future research is warranted on how to select the optimal reference group. Microbiome composition is very dynamic and is affected by time-variant covariates such as diet, environmental exposures, and gender. Understanding how various biological/environmental factors affect the changes in the microbiome compositions might be valuable in gaining valuable biological insight into disease diagnosis and prognosis. Therefore, extending our LNM-FA model to account for covariates would be valuable.

Appendix

A ELBO for LNM Model

First, we decompose \( F(q(y), w) \) into 3 parts:

\[
F(q(y), w) = \int q(y) \log f(w|y)dy + \int q(y) \log f(y)dy - \int q(y) \log q(y)dy.
\]

The second and third integral (i.e., \( E_q(y)(\log f(y)) \) and \( E_q(y)(\log q(y)) \)) have explicit solutions such that

\[
E_q(y)(\log f(y)) = -\frac{K}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (m - \mu)^\top \Sigma^{-1} (m - \mu) - \frac{1}{2} \text{tr}(\Sigma^{-1}V)
\]

and

\[
-E_q(y)(\log q(y)) = \frac{1}{2} \log |V| + \frac{K}{2} + \frac{K}{2} \log(2\pi).
\]
Note that $V$ is a diagonal matrix. As for the first integral, it has no explicit solution because of the expectation of log sum exponential term:

$$E_q(y) \log f(w|y) = C + w^\top m - \left( \sum_{k=1}^{K+1} w_k \right) E_q(y) \left[ \log \sum_{k=1}^{K+1} \exp y_k \right],$$

where $w^*$ represents a $K$ dimension vector with first $K$ elements of $w$, $y_{K+1}$ is set to 0, and $C$ stands for $\log \frac{1}{\prod_{k=1}^{K+1} w_k}$. Blei and Lafferty (2007) proposed an upper bound for $E_q(y) \left[ \log \sum_{k=1}^{K+1} \exp y_k \right]$ as

$$E_q(y|m,V) \left[ \log \sum_{k=1}^{K+1} \exp y_k \right] \leq \xi^{-1} \left\{ \sum_{k=1}^{K+1} E_q(y|m,V) \left[ \exp(y_k) \right] \right\} - 1 + \log(\xi), \tag{7}$$

where $\xi \in \mathbb{R}$ is introduced as a new variational parameter. Fang and Subedi (2023) utilized this upper bound to find a lower bound for $E_q(y) \log f(w|y)$.

Combining all 3 parts together, we have the approximate lower bound for log $f(w)$:

$$\tilde{F}(q(y),w) = C + w^\top m - \left( \sum_{k=1}^{K+1} w_k \right) \left\{ \log \left[ \sum_{k=1}^{K} \exp \left( m_k + \frac{v_k^2}{2} \right) \right] + \frac{1}{2} \log |V| + \frac{K}{2} - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (m - \mu)^\top \Sigma^{-1} (m - \mu) - \frac{1}{2} \text{tr}(\Sigma^{-1} V) \right\},$$

where $m_k, v_k^2$ stands for $k^{th}$ entry of $m$ and the $k^{th}$ diagonal entry of $V$. The two upper bounds are equal when minimize (7) with respect to $\xi$.

B ELBO for Cycle 2

Here, in the second cycle, we have

$$F(q(u,y),w) = \int q(u,y) \log \frac{f(w,u,y)}{q(u,y)} dy du$$

$$= \int q(u,y) \log f(w|u,y) dy du + \int q(u,y) \log f(u,y) dy du$$

$$- \int q(u,y) \log q(u,y) dy du.$$

Furthermore, we assume that $q(u,y) = q(u)q(y)$, $u \sim N(\bar{m}, \bar{V})$ and $y \sim N(m, V)$. Thus, the first term can be written as follows:

$$\int q(u,y) \log f(w|u,y) dy du = \int q(u)q(y) \log f(w|y) dy du$$

$$= \int q(y) \log f(w|y) dy.$$
This is identical to the first term in the ELBO in the first cycle, and thus, its lower bound is

\[ \int q(u, y) \log f(w|u, y) dydu \geq C + w^\top m - \left( \sum_{k=1}^{K+1} w_k \right) \left\{ \log \left( \sum_{k=1}^K \exp \left( m_k + \frac{v_k^2}{2} \right) + 1 \right) \right\}. \]

The third term is

\[-\int q(u, y) \log q(u, y) dydu = \frac{1}{2} \left( \log |V| + \log |\tilde{V}| + q + K + (K + q) \log 2\pi \right).\]

The second term is

\[
\int q(u, y) \log f(u, y) dydu = \int q(u)q(y) \log [f(y|u)f(u)] dydu \\
= E_{q(u)}E_{q(y)}(\log f(y|u)f(u)) \\
= \frac{1}{2} \left\{ (q + K) \log(2\pi) - \log |D| - \tilde{m}^\top \tilde{m} - \text{tr}(\tilde{V}) - \text{tr}(A^\top D^{-1} A \tilde{V}) \\
- \text{tr}(D^{-1}(V + (m - \mu)^\top (m - \mu))) + 2(m - \mu)^\top D^{-1} A \tilde{m} \\
- \tilde{m}^\top A^\top D^{-1} A \tilde{m} - \text{tr}(A^\top D^{-1} A \tilde{V}) \right\}.
\]

Overall, the ELBO in the second cycle is as follows:

\[ F(q(u, y), w) \geq C + w^\top m - \left( \sum_{i=1}^{K+1} w_i \right) \left\{ \log \left( \sum_{k=1}^K \exp \left( m_k + \frac{v_k^2}{2} \right) + 1 \right) \right\} + \]

\[
\frac{1}{2} \left( \log |V| + \log |\tilde{V}| + q + K - \log |D| - \tilde{m}^\top \tilde{m} - \text{tr}(\tilde{V}) - \text{tr}(D^{-1}(V + (m - \mu)^\top (m - \mu))) + 2(m - \mu)^\top D^{-1} A \tilde{m} - \tilde{m}^\top A^\top D^{-1} A \tilde{m} - \text{tr}(A^\top D^{-1} A \tilde{V}) \right) \]

where \( m \) and \( V \) are calculated from first stage.

In addition to variational parameter in second stage, it is worth noticing that \( \tilde{m}_{ig} = E(u_{ig}|y_i, z_{ig}) \), and \( \tilde{V}_g = \text{Cov}(u_{ig}|y_i, z_{ig}) \). Because the following relationship

\[
\begin{bmatrix} y_i \\ u_{ig} \end{bmatrix} | z_{ig} \sim \text{MVN} \left[ \begin{pmatrix} \mu_g \\ 0 \end{pmatrix}, \begin{pmatrix} \Lambda_g A_g^\top + D_g A_g & \Lambda_g^\top I_q \\ A_g^\top \Lambda_g + D_g \Lambda_g^\top & I_q \end{pmatrix} \right].
\]

Therefore,

\[
E(u_{ig}|y_i, z_{ig} = 1) = \Lambda_g^\top (\Lambda_g A_g^\top + D_g)^{-1} (m_{ig} - \mu_g), \text{ and}
\]

\[
\text{Cov}(u_{ig}|y_i, z_{ig}) = I_q - \Lambda_g^\top (\Lambda_g A_g^\top + D_g)^{-1} \Lambda_g.
\]

Then, because

\[
(\Lambda_g^\top D_g^{-1} A_g + I_q)^{-1} = I_q - \Lambda_g^\top D_g^{-1/2} (I + D_g^{-1/2} A_g \Lambda_g^\top D_g^{-1/2})^{-1} D_g^{-1/2} A_g,
\]

and because \( D_g \) is always invertible by design, we have the following:

\[
\tilde{V} = (\Lambda_g^\top D_g^{-1} A_g + I_q)^{-1} = I_q - \Lambda_g^\top (D_g + A_g \Lambda_g^\top)^{-1} \Lambda_g.
\]
The above shows $\tilde{V} = Cov(u_{ig}|y_i, z_{ig})$. Similarly, for $\tilde{m}$, we have the following:

$$\tilde{m} = (\Lambda_g^T D_g^{-1} \Lambda_g + \Theta_g)^{-1} \Lambda_g^T D_g^{-1} (m_{ig} - \mu_g)$$

$$= (I_g - \Lambda_g^T (D_g + \Lambda_g \Lambda_g^T)^{-1} \Lambda_g) \Lambda_g^T D_g^{-1} (m_{ig} - \mu_g)$$

$$= \Lambda_g^T \left( D_g^{-1} - (D_g + \Lambda_g \Lambda_g^T)^{-1} \Lambda_g \Lambda_g^T D_g^{-1} \right) (m_{ig} - \mu_g)$$

$$= \Lambda_g^T \left( \Lambda_g \Lambda_g^T + D_g \right)^{-1} (m_{ig} - \mu_g).$$

The last equality is followed by the following:

$$I = (\Lambda_g \Lambda_g^T + D_g) (D_g^{-1} - (D_g + \Lambda_g \Lambda_g^T)^{-1} \Lambda_g \Lambda_g^T D_g^{-1})$$

$$= (D_g^{-1} - (D_g + \Lambda_g \Lambda_g^T)^{-1} \Lambda_g \Lambda_g^T D_g^{-1}) (\Lambda_g \Lambda_g^T + D_g)^T$$

$$= (D_g^{-1} - (D_g + \Lambda_g \Lambda_g^T)^{-1} \Lambda_g \Lambda_g^T D_g^{-1}) (\Lambda_g \Lambda_g^T + D_g).$$

Furthermore, we showed that $(D_g^{-1} - (D_g + \Lambda_g \Lambda_g^T)^{-1} \Lambda_g \Lambda_g^T D_g^{-1}) = (\Lambda_g \Lambda_g^T + D_g)^{-1}$. Hence, we conclude that the variational parameter is essentially the conditional expectation and covariance of $u_{ig}|y_i$.

### C Parameter Estimates for the Family of Models

From here, we will derive the family of 8 models by setting different constraints on $\Sigma$. Notice the following identities are easy to verify:

$$\sum_{i=1}^n z_{ig} = n_g, \quad log |(d_g I_K)^{-1}| = log(d_g^{-K}), \quad \text{and} \quad \theta_g = \sum_{i=1}^n \frac{z_{ig}(\tilde{m}_{ig} \tilde{m}_{ig}^T + \tilde{V}_{ig})}{n_g}.$$

1. “UUU”: We do not put any constraint on $\Lambda_g, D_g$. The solution is the same as the above derivation.
2. “UUC”: We assume $D_g = d_g I_K$, and no constraint for $\Lambda_g$. Apart from $D_g$, the estimation is the same as for model “UUU.”

$$\hat{d}_g = \frac{1}{K} \text{tr} \{ \Sigma_g - 2 \Lambda_g \beta_g S_g + \Lambda_g \theta_g \Lambda_g^T \}.$$

3. “UCU”: We assume $D_g = D$, and no constraint for $\Lambda_g$. Apart from $D_g$, the rest estimation is exactly the same as for model “UUU.” Taking derivative with respect to $D^{-1}$, we get the following:

$$\hat{D} = \frac{1}{n} \sum_{g=1}^G n_g \text{diag} \{ \Sigma_g - 2 \Lambda_g \beta_g S_g + \Lambda_g \theta_g \Lambda_g^T \}.$$

4. “UCC”: We assume $D_g = d I_K$, and no constraint for $\Lambda_g$. Apart from $D_g$, the rest estimation is exactly the same as for model “UUU.” Following the same procedure as model “UUC” and “UCU,” we get the following:

$$\hat{d} = \frac{1}{Kn} \sum_{g=1}^G n_g \text{tr} \{ \Sigma_g - 2 \Lambda_g \beta_g S_g + \Lambda_g \theta_g \Lambda_g^T \}.$$
5. “CUU”: We assume \( \Lambda_g = \Lambda \), and no constraint for \( D_g \). Aside from \( \Lambda \), the estimation is the same as for model “UUU.” Taking derivative of \( l_2 \) with respect to \( \Lambda \) gives us the following:

\[
\frac{\partial l_2}{\partial \Lambda} = \sum_{g=1}^{G} n_g (D_g^{-1} S_g \beta_g^T - D_g^{-1} \Lambda g),
\]

which must be solved for \( \Lambda \) in a row-by-row manner. Let \( \lambda_i \) to represent the ith row of \( \Lambda \), and \( r_i \) to represent ith row of \( \sum_{g=1}^{G} n_g (D_g^{-1} S_g \beta_g^T) \). Then,

\[
\lambda_i = r_i \left( \sum_{g=1}^{G} \frac{n_g}{d_g(i)} \theta_g \right)^{-1},
\]

where \( d_g(i) \) is the ith entry of \( D_g \).

6. “CUC”: We assume \( \Lambda_g = \Lambda \), \( D_g = d_g I_K \). Estimation of \( \Lambda_g \) is exactly the same as for model “CUU.” Estimation of \( D_g \) is exactly the same as for model “UUC.”

7. “CCU”: We assume \( \Lambda_g = \Lambda \), \( D_g = D \). Estimation of \( \Lambda_g \) is exactly the same as for model “CUU.” Estimation of \( D_g \) is exactly the same as for model “UCC.”

8. “CCC”: We assume \( \Lambda_g = \Lambda \), \( D_g = d I_K \). Estimation of \( \Lambda_g \) is exactly the same as for model “CUU.” Estimation of \( D_g \) is exactly the same as for model “UCC.”

### D Initialization

For estimation, we need to first initialize the model parameters, variational parameters, and the component indicator variable \( Z_{ig} \). The EM algorithm for finite mixture models is known to be heavily dependent on starting values. Let \( z_{ig}^*, \pi_g^*, \mu_g^*, D_g^*, \Lambda_g^*, m_{ig}^* \) and \( V_{ig}^* \) be the initial values for \( Z_{ig}, \pi_g, \mu_g, D_g, \Lambda_g, m_{ig} \) and \( V_{ig} \) respectively. The initialization is conducted as follows:

1. \( z_{ig}^* \) can be obtained by random allocation of observation to different clusters, where this initial cluster assignment can be obtained from \( k \)-means clustering or a model-based clustering algorithm. Since our algorithm is based on a factor analyzer structure, we initialize \( Z_{ig} \) using the cluster membership obtained by fitting parsimonious Gaussian mixture models (PGMM; McNicholas & Murphy, 2008) to the transformed variable \( Y \) obtained using (1). For computational purposes, any 0 in the \( W \) was replaced by 0.001 for initialization. The implementation of PGMM is available in R package “pgmm” (McNicholas et al., 2022).

2. Using this initial partition, \( \mu_g^* \) is the sample mean of the \( g^{th} \) cluster, and \( \pi_g^* \) is the proportion of observations in the \( g^{th} \) cluster in this initial partition.

3. Similar to McNicholas and Murphy (2008), we estimate the sample covariance matrix \( S_g^* \) for each group and then use eigen-decomposition of \( S_g^* \) to obtain \( D_g^* \) and \( \Lambda_g^* \). Suppose \( \lambda_g \) is a vector of the first \( q \) largest eigenvalues of \( S_g \) and the columns of \( L_g \) are the corresponding eigenvectors, then

\[
\Lambda_g^* = L_g \Lambda_g^\frac{1}{2}, \quad \text{and} \quad D_g^* = \text{diag}(S_g^* - \Lambda_g^* \Lambda_g^*^\top).
\]

4. As the Newton-Raphson method is used to update the variational parameters, we need \( m^* \) and \( V^* \). For \( m^* \), we apply an additive log-ratio transformation on the observed taxa compositions \( \hat{p} \) and set \( m^* = \phi(\hat{p}) \) using (1). For \( V^* \), we use a diagonal matrix where all diagonal entries
are 0.1. During our simulation studies, we found 0.1 worked well, and it is important to choose a small value for $V^*$ to avoid overshooting in the Newton-Raphson method.

E Visualization of the Cluster Structures from Simulation Studies 1 and 2

E.1 Simulation Study 1

![Scatter plot of latent variable Y in one of the hundred datasets from Simulation Study 1. The observations are colored using their true class label. For this dataset, ARI of 1 was obtained by LNM-FA](image)
Figure 1 shows a visualization of the cluster structure in the latent space for one of the hundred datasets.

E.2 Simulation Study 2

Figure 2 shows visualization of the cluster structure in the latent space in one of the hundred datasets.

F True Parameters in Simulation Studies

In Simulation Study 1:

\[ \mu_1 = [-0.17, 0.03, 0.08, 0.24, 0.24, -0.06, -0.03, 0.14, -0.11, 0.14] \]
\[ \mu_2 = [0.33, 0.63, 0.44, 0.60, 0.32, 0.52, 0.39, 0.50, 0.51, 0.45] \]
\[ \mu_3 = [-0.59, -0.66, -0.55, -0.45, -0.60, -0.68, -0.53, -0.41, -0.65, -0.46] \]

\[ \Lambda^T = \begin{bmatrix} -0.003 & -0.278 & -0.131 & 0.424 & 0.038 & 0.275 & -0.222 & -0.100 & 0.284 & 0.030 \\ 0.386 & 0.090 & 0.187 & 0.092 & -0.796 & 0.062 & 0.204 & 0.116 & 0.422 & -0.353 \\ -0.242 & 0.128 & 0.375 & -0.983 & -0.423 & 0.242 & -0.574 & -0.265 & -0.205 & 0.153 \end{bmatrix} \]

\[ D = 0.01 \times I_{10}. \]

In Simulation Study 2:

\[ \mu_1 = [0.16, -0.13, 0.06, 0.13, 0.00, -0.06, -0.02, -0.11, 0.00, 0.03] \]
\[ \mu_2 = [0.79, 1.01, 0.66, 0.76, 0.86, 0.83, 0.66, 0.68, 0.85, 0.84] \]
\[ \mu_3 = [-0.77, -0.89, -0.88, -0.78, -0.71, -0.89, -0.86, -0.82, -0.86, -0.80] \]

\[ \Lambda_1 = \begin{bmatrix} -0.003 & 0.386 & -0.242 \\ -0.278 & 0.090 & 0.128 \\ -0.131 & 0.187 & 0.375 \\ 0.424 & 0.092 & -0.983 \\ 0.038 & -0.796 & -0.423 \\ 0.275 & 0.062 & 0.242 \\ -0.222 & 0.204 & -0.574 \\ -0.100 & 0.116 & -0.265 \\ 0.284 & 0.422 & -0.205 \\ 0.030 & -0.353 & 0.153 \end{bmatrix}, \quad \Lambda_2 = \begin{bmatrix} -0.426 & -0.289 & 0.050 \\ -0.070 & 0.267 & 0.120 \\ 0.126 & -0.184 & -0.140 \\ 0.276 & -0.690 & 0.394 \\ 0.085 & -0.243 & -0.400 \\ 0.137 & 0.104 & -0.305 \\ 0.400 & 0.491 & -0.434 \\ 0.199 & 0.334 & 0.054 \\ 0.167 & 0.022 & -0.167 \\ 0.299 & -0.133 & -0.338 \end{bmatrix}, \quad \Lambda_3 = \begin{bmatrix} 0.082 & -0.167 & 0.050 \\ 0.146 & 0.123 & -0.033 \\ 0.164 & -0.075 & -0.142 \\ -0.107 & -0.062 & 0.002 \\ 0.086 & 0.054 & -0.143 \\ -0.078 & -0.051 & 0.155 \\ -0.074 & -0.252 & -0.048 \\ -0.059 & 0.112 & 0.076 \\ 0.047 & 0.054 & -0.019 \\ 0.220 & -0.122 & 0.026 \end{bmatrix} \]

\[ D_1 = \text{diag} [0.03, 0.004, 0.028, 0.015, 0.005, 0.029, 0.003, 0.016, 0.014, 0.015] \]
\[ D_2 = \text{diag} [0.004, 0.03, 0.015, 0.003, 0.029, 0.015, 0.028, 0.03, 0.005, 0.03] \]
\[ D_3 = \text{diag} [0.022, 0.006, 0.03, 0.018, 0.011, 0.002, 0.004, 0.015, 0.025, 0.005] \]

In Simulation Study 3:

\[ \mu_1 \sim N(0.8, 0.1), \quad \mu_2 \sim N(-0.8, 0.1), \quad \mu_3 \sim N(0, 0.1) \]
\[ \Lambda_1 \sim N(0.5, 0.1), \quad \Lambda_2 \sim N(-0.5, 0.1), \quad \Lambda_3 \sim N(0, 0.1), \]
\[ D_8 \sim \text{diag} [\text{Uniform}(0, 0.05)] \]
G Additional Simulations

To show the performance of our model on a dataset not generated from the LNM-FA model, we generated a two-component mixture model from 50-dimensional multinomial distributions. Although microbiome data are high dimensional, few taxa have high abundance and most taxa have low abundance. Additionally, the number and type of abundant taxa can vary among

Fig. 2 Scatter plot of latent variable Y in one of the hundred datasets from Simulation Study 2. The observations are colored using their true class label. For this dataset, ARI of 1 was obtained by LNM-FA
clusters (or groups). To create a similar structure, we generated the compositions of the 50 taxa for each component from two different beta distributions. For component 1, we generated 10 randomly selected taxa from a beta distribution with a mean of 0.25 and the remaining 40 taxa from a beta distribution with a mean of 0.001. The resulting vector is then normalized to sum to 1. Similarly, for component 2, we generated 15 randomly selected taxa from a beta distribution with a mean of 0.25 and the remaining 35 taxa from a beta distribution with a mean of 0.001. The resulting vector is again normalized to sum to 1. We generated 100 datasets under each of the five scenarios. The same sets of parameters were used to generate the data under all scenarios, but the sample size $n$ varied among scenarios ranging from $n = 50$ to $n = 1000$. And for each scenario, we fit LNM-FA in two different ways: first, an arbitrary column is chosen as a reference level, and second, the column which has the highest total read counts is chosen as a reference level. We ran all 8 models in the LNM-FA family for $G = 1, \ldots, 4$ and $q = 1, \ldots, 3$ for both cases and selected the best model using the BIC. Table 4 summarizes the clustering performance of the proposed approach under all 5 scenarios.

In Scenario 1 with the smallest sample size $n = 50$, by choosing an arbitrary reference level, the correct model was only selected in 10 out of the 100 datasets, and in the remaining 90 datasets, a one-component model was selected. However, when switching to the most abundant reference level, the performance becomes almost perfect. Although the arbitrary reference level did not work well when $n = 50$, as the number of observations increases, the performance becomes better. When $n = 100, 300$, $G = 2$ model is selected in more than 80% of the datasets and $G = 1$ for the rest. Note that for $n = 50, 100$, while the overall ARI is less than 0.9, the average ARI where the correct number of components is selected is 1 (i.e., perfect classification). When $n = 500, 1000$, the performance between arbitrary and most abundant reference levels seems very similar. In terms of the rate of selected $G = 2$, models that use the most abundant reference level still have a higher ratio compared to the arbitrary reference level. However, the overall average ARIs are all 0.99 (and perfect classification when the correct number of components is selected). For these two scenarios, when the correct number of components was not selected, a three-component model was selected where the third component only had a small number of observations (i.e., around 2%). While DMM correctly identified the correct number of components for all scenarios,

| Table 4 | Model selection performance for real microbiome data simulation |
|---------|---------------------------------------------------------------|
| $n$     | # of $G = 2$ is selected | ARI (when $G = 2$ is selected) | Overall average ARI |
| Arbitrary Ref | 50 | 10/100 | 1 | 0.1 |
| Most abundant Ref | 50 | 99/100 | 1 | 0.99 |
| Arbitrary Ref | 100 | 80/100 | 1 | 0.81 |
| Most abundant Ref | 100 | 90/100 | 1 | 0.99 |
| Arbitrary Ref | 300 | 92/100 | 1 | 0.99 |
| Most abundant Ref | 300 | 92/100 | 1 | 0.99 |
| Arbitrary Ref | 500 | 81/100 | 1 | 0.99 |
| Most abundant Ref | 500 | 87/100 | 1 | 0.99 |
| Arbitrary Ref | 1000 | 76/100 | 1 | 0.99 |
| Most abundant Ref | 1000 | 83/100 | 1 | 0.99 |
the LNM-MM encountered computational issues in all scenarios. It is not surprising that when the dataset is generated from a mixture of multinomial models, the DMM performs well as a mixture of multinomial models can be obtained as a special case of DMM. $k$-means and hierarchical clustering had perfect performance when fitting a two-component model on the ALR-transformed data on all datasets in this simulation study. However, for $k$-means and hierarchical clustering, the number of clusters was set to 2 (i.e., true value).

In a real dataset, where a reference group needs to be selected, one needs to be cautious regarding which group is selected as the reference group. This is especially important for the high-dimensional setting where the data is sparse and the sample size is small. Here, we have a 50-dimensional dataset. When an arbitrary taxon is selected as a reference group, for a dataset with a small sample size, the reference group could be sparse and the mean relative proportion of the reference group could be small. In our example, the mean relative abundance for the arbitrary reference group for the two components is 0.002 and 0.007. When $n$ was small, our approach did drop in performance, but when $n$ was large, our approach was able to recover the underlying cluster structure. However, choosing the most abundant taxon as the reference group performed well even when the sample size was small. When the most abundant taxa were chosen as the reference group, this ensured that the reference group did not have a relative abundance that was close to 0. This illustrates that the optimal reference group warrants further investigation.

**Funding** This work was supported by the Collaboration Grants for Mathematicians from the Simons Foundation, the Discovery Grant from the Natural Sciences and Engineering Research Council of Canada, and the Canada Research Chair Program.

**Data Availability** The datasets used in this manuscript are all publicly available in the R packages MicrobiomeCluster and Microbiome.

**Declarations**

**Conflict of Interest** The authors declare no competing interests.

**References**

Abdel-Aziz, M. I., Brinkman, P., Vijverberg, S. J., Neerinctx, A. H., Riley, J. H., Bates, S., Hashimoto, S., Kermani, N. Z., Chung, K. F., Djukanovic, R., et al. (2021). Sputum microbiome profiles identify severe asthma phenotypes of relative stability at 12–18 months. *Journal of Allergy and Clinical Immunology, 147*(1), 123–134.

Äijö, T., Müller, C. L., & Bonneau, R. (2018). Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing. *Bioinformatics, 34*(3), 372–380.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological), 44*(2), 139–160.

Aitken, A. C. (1926). A series formula for the roots of algebraic and transcendental equations. *Proceedings of the Royal Society of Edinburgh, 45*(1), 14–22.

Archambeau, C., Cornford, D., Opper, M., & Shawe-Taylor, J. (2007). Gaussian process approximations of stochastic differential equations. *Journal of Machine Learning Research - Proceedings Track, 1*, 1–16.

Arridge, S. R., Ito, K., Jin, B., & Zhang, C. (2018). Variational Gaussian approximation for Poisson data. *Inverse Problems, 34*(2), 025005.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., et al. (2011). Enterotypes of the human gut microbiome. *Nature, 473*(7346), 174–180.

Baek, J., & McLachlan, G. J. (2011). Mixtures of common t-factor analyzers for clustering high-dimensional microarray data. *Bioinformatics, 27*(9), 1269–1276.
Becker, C., Neurath, M., & Wirtz, S. (2015). The intestinal microbiota in inflammatory bowel disease. *ILAR Journal, 56*(2), 192–204.

Blei, D., & Lafferty, J. (2007). A correlated topic model of science. *The Annals of Applied Statistics, 1*(1), 17–35.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association, 112*(518), 859–877.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., & Lindsay, B. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics, 46*(2), 373–388.

Bouveyron, C., & Brunet, C. (2012). Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Statistics and Computing, 22*(1), 301–324.

Calle, M. L. (2019). Statistical analysis of metagenomics data. *Genomics & Informatics, 17*(1), e6.

Celeux, G., & Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition, 28*(5), 781–793.

Challis, E., & Barber, D. (2013). Gaussian Kullback-Leibler approximate inference. *The Journal of Machine Learning Research, 14*(8), 2239–2286.

Chen, J., & Li, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *The Annals of Applied Statistics, 7*(1), 418–442.

Chipman, H., Hastie, T. J., & Tibshirani, R. (2003). Clustering microarray data. *Statistical analysis of gene expression microarray data, 1*, 159–200.

Cho, I., & Blaser, M. J. (2012). The human microbiome: At the interface of health and disease. *Nature Reviews Genetics, 13*(4), 260–270.

Davis, C. (2016). The gut microbiome and its role in obesity. *Nutrition Today, 51*(4), 167–174.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B, 39*(1), 1–38.

Fang, Y., & Subedi, S. (2023). Clustering microbiome data using mixtures of logistic normal multinomial models. *Scientific Reports, 13*(1), 14758.

Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurry, T. A., Edgell, D. R., & Gloor, G. B. (2014). 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), 1–13.

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal, 41*(8), 578–588.

Garrett, W. S. (2019). The gut microbiota and colon cancer. *Science, 364*(6446), 1133–1135.

Ghahramani, Z., & Hinton, G. E. (1997). The EM algorithm for mixtures of factor analyzers, technical report. University of Toronto: Technical Report CRG-TR-96-1.

Gloor, G., Macklaim, J., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology, 8*, 2224.

Gollini, I., & Murphy, T. B. (2014). Mixture of latent trait analyzers for model-based clustering of categorical data. *Statistics and Computing, 24*(4), 569–588.

Holmes, I., Harris, K., & Quince, C. (2012). Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLOS One, 7*, e30126.

Hotterbeekx, A., Xavier, B. B., Bielen, K., Lammens, C., Moons, P., Schepens, T., Ieven, M., Jorens, P. G., Goossens, H., Kumar-Singh, S., et al. (2016). The endotrichal tube microbiome associated with Pseudomonas aeruginosa or Staphylococci epidermidis. *Scientific Reports, 6*(1), 1–11.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193–218.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., FitzGerald, M. G., Fulton, R. S., et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature, 486*(7402), 207–214.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A* pp. 49–66.

Koslovsky, M. D., & Vannucci, M. (2020). MicroBVS: Dirichlet-tree multinomial regression models with Bayesian variable selection-an R package. *BMC Bioinformatics, 21*(1), 1–10.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79–86.

La Rosa, P. S., Brooks, J. P., Deych, E., Boone, E. L., Edwards, D. J., Wang, Q., Sodergren, E., Weinstock, G., & Shannon, W. D. (2012). Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLOS One, 7*(12), e52078.

Lahti, L., Shetty, S. (2012-2019). *microbiome R package*.

Mao, J., & Ma, L. (2022). Dirichlet-tree multinomial mixtures for clustering microbiome compositions. *The Annals of Applied Statistics, 16*(3), 1476–1499.
Martínez, I., Stegen, J. C., Maldonado-Gómez, M. X., Eren, A. M., Siba, P. M., Greenhill, A. R., & Walter, J. (2015). The gut microbiota of rural Papua New Guineans: Composition, diversity patterns, and ecological processes. *Cell reports*, 11(4), 527–538.

McLachlan, G. J., & Krishnan, T. (2007). *The EM algorithm and extensions*. John Wiley & Sons.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. John Wiley & Sons.

McLachlan, G. J., Peel, D., & Bean, R. W. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis*, 41(3–4), 379–388.

McLachlan, G. Peel, D. (2000b). Mixtures of factor analyzers. In: *Proceedings of the seventeenth international conference on machine learning*. Morgan Kaufmann, pp. 599–606.

McNicholas, P. D., ElSherbiny, A., McDaid, A. F. & Murphy, T. B. (2022). pgmm: Parsimonious Gaussian mixture models. *R package version 1.2.6*. https://CRAN.R-project.org/package=pgmm

McNicholas, P. D., & Murphy, T. B. (2008). Parsimonious Gaussian mixture models. *Statistics and Computing*, 18(3), 285–296.

McNicholas, P. D., & Murphy, T. B. (2010). Model-based clustering of microarray expression data via latent gaussian mixture models. *Bioinformatics*, 26(21), 2705–2712.

Meng, X.-L., & Van Dyk, D. (1997). The EM algorithm-an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3), 511–567.

O’Keefe, S. J., Li, J. V., Lahti, L., Ou, J., Carbonero, F., Mohammed, K., Posma, J. M., Kinross, J., Wahl, E., Ruder, E., et al. (2015). Fat, fibre and cancer risk in African Americans and rural Africans. *Nature Communications*, 6(1), 1–14.

Pawlowsky-Glahn, V., & Buccianti, A. (2011). *Compositional data analysis: Theory and applications*. John Wiley & Sons.

Pawlowsky-Glahn, V., Egozcue, J. J. & Tolosana-Delgado, R. (2007). *Lecture notes on compositional data analysis*.

Pfirschke, C., Garris, C., & Pittet, M. J. (2015). Common TLR5 mutations control cancer progression. *Cancer Cell*, 27(1), 1–3.

Quinn, T., Erb, I., Gloor, G., Notredame, C., Richardson, M. & Crowley, T. (2019). A field guide for the compositional analysis of any-omics data. *GigaScience* 8.

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.

Sender, R., Fuchs, S., & Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLOS Biology*, 14, e1002533.

Shi, Y. (2020). *Microbiomecluster. R package*.

Silverman, J. D., Durand, H. K., Bloom, R. J., Mukherjee, S., & David, L. A. (2018). Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. *Microbiome*, 6(1), 1–20.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19), 10869–10874.

Subedi, S., & Browne, R. (2020). A parsimonious family of multivariate Poisson-lognormal distributions for clustering multivariate count data. Stat, 9(1), e310.

Subedi, S., Neish, D., Bak, S., & Feng, Z. (2020). Cluster analysis of microbiome data via mixtures of Dirichlet-multinomial regression models. *Journal of Royal Statistical Society: Series C*, 69(5), 1163–1187.

Subedi, S., Punzo, A., Ingrassia, S., & McNicholas, P. D. (2013). Clustering and classification via cluster-weighted factor analyzers. *Advances in Data Analysis and Classification*, 7(1), 5–40.

Subedi, S., Punzo, A., Ingrassia, S., & McNicholas, P. D. (2015). Cluster-weighted t-factor analyzers for robust model-based clustering and dimension reduction. *Statistical Methods & Applications*, 24(4), 623–649.

Taie, W. S., Omar, Y. & Badr, A. (2018). Clustering of human intestine microbiomes with k-means. *Cell reports*, 11(4), 527–538.

Wadsworth, W. D., Argiento, R., Guindani, M., Galloway-Pena, J., Shelburne, S. A., & Vannucci, M. (2017). An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinformatics*, 18(1), 1–12.

Wainwright, M. J., & Jordan, M. I. (2008). *Graphical models, exponential families, and variational inference*. Hanover, MA, USA: Now Publishers Inc.

Wang, T., & Zhao, H. (2017). A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. *Biometrics*, 73(3), 792–801.

Woodbury, M. A. (1950). Inverting modified matrices. *Memorandum Report*, 42(106), 336.
Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science, 334*(6052), 105–108.
Xia, F., Chen, J., Fung, W. K., & Li, H. (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics, 69*(4), 1053–1063.
Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A., & Yi, N. (2017). Negative binomial mixed models for analyzing microbiome count data’. *BMC Bioinformatics, 18*, 4.

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.