Zero-inflated Bayesian factor analysis model with skew-normal priors for modeling microbiome data

Saurabh Panchasara[1], Hanna Jankowski[1], and Kevin McGregor[2]

[1]Department of Mathematics and Statistics, York University
[2]Department of Statistics, University of Manitoba

Abstract
Background: Advancements in next-generation sequencing have transformed our understanding of host-microbe interactions, revealing links between microbial composition and chronic conditions such as diabetes, Crohn's disease, and others. However, the analysis of microbiome data is complex due to its high-dimension as well as additional statistical characteristics. One primary objective is to achieve effective dimensionality reduction while simultaneously accounting for the data's compositional nature and zero inflation. Existing probabilistic models are often based on the assumption that the log-ratio-transformed compositions are normally distributed. This assumption is problematic because it can fail to capture the significant skewness inherent in these transformed compositions.
Results: We propose a new model called the Zero-Inflated Factor Analysis Logistic Skew-Normal Multinomial (ZIFA-LSNM) model. ZIFA-LSNM integrates a zero-inflation component to handle excess zeros, employs factor analysis for dimensionality reduction, and, critically, utilizes skew-normal priors on the latent factors to explicitly model data asymmetry. Posterior inference is performed using an efficient variational inference algorithm. Through simulation studies and real data analysis, the ZIFA-LSNM model is shown to have improved performance in parameter recovery and composition estimation compared to its Gaussian-based counterparts.
Conclusion: The ZIFA-LSNM model demonstrates that explicitly accounting for skewness in the latent factor structure can substantially improve inference in commonly-observed microbiome data. The proposed model, therefore, offers a flexible and scalable framework allowing for the improved analysis of the complex relationships between microbial communities and human health.
Corresponding author: Kevin McGregor
Keywords: Microbiome data, high dimensionality, additive log ratio transformation, skew normal priors, factor analysis, variational inference

1. Background

The communities of microorganisms residing on and within the human body, collectively known as the human microbiome (Tyler et al. [2014]), play a pivotal role in regulating host health. These microbial ecosystems influence fundamental physiological processes, from immune system development to metabolism, and have been linked to a wide range of chronic conditions (Ley [2010], Manichanh et al. [2012], Fujimura and Lynch [2015],

Lutz et al. [2022]). Advancements in high-throughput sequencing technologies, such as 16S rRNA gene sequencing and shotgun metagenomics, have enabled researchers to understand the composition of these communities in great detail by generating large-scale sample-by-taxon count datasets that are central to modern biological research. However, analyzing these datasets is complex, as they are characterized by a unique combination of statistical properties. They are (i) compositional, (ii) characterized by excess zeros (zero-inflated), and (iii) can exhibit skewness in the log-transformed compositional ratios. Addressing these issues along with the high-dimensionality of the datasets is the major challenge of modern microbiome analysis.

The first challenge in the analysis of microbiome data is its compositional nature. The total number of sequencing reads per sample (often called library size or sequencing depth) is an arbitrary technical artifact, implying that the observed counts provide information only about relative, not absolute, abundances. Failure to account for this proportion-sum constraint can lead to spurious correlations and misleading biological conclusion [Gloor et al., 2017]. The standard approach here is to apply a log-ratio transformation, which maps the composition from the constrained simplex to an unconstrained real space of one less dimension. These transformations possess the critical properties of scale invariance and sub-compositional coherence, ensuring that analytical results are not affected by observing only a subset of the composition (Van den Boogaart and Tolosana-Delgado [2013]). Early frameworks, such as the Dirichlet-multinomial (DM) and logistic normal multinomial (LNM) models were used as models for the transformed data. However, the DM model cannot adequately capture covariance among taxa due to its limited parameterization (Xia et al. [2013]). Hence, to address the correlation structure, Xia et al. [2013] combined a multinomial distribution conditional on latent log-ratio abundances with a Gaussian prior placed on the transformed compositions (Aitchison [1982]).

Next, microbiome data exhibit large numbers of zeros across taxa. Some taxa are truly absent in certain samples (structural zeros), while others go undetected due to limited sequencing depth or experimental biases (sampling zeros) (Lutz et al. [2022]). This high prevalence of zeros creates a major challenge because standard statistical methods, like Poisson or negative binomial models often produce biased results. Several zero-inflated models have been proposed for compositional inference under extreme sparsity (Paulson et al. [2013], Chen and Li [2016], Tang and Chen [2019], Xu et al. [2021], Zeng et al. [2023], Ba et al. [2024]). This challenge is compounded by the "curse of dimensionality," wherein the number of taxa (p) often far exceeds the number of samples (n). Consequently, dimensionality-reduction strategies such as probabilistic principal component analysis (PCA) (Tipping and Bishop [1999]), and factor analysis have been frequently incorporated into microbiome models to capture latent community structure.

Recent methodological advances address these challenges individually or in partial combinations. For instance, Zeng et al. [2023] proposed a Zero-Inflated Probabilistic PCA Logistic Normal Multinomial (ZIPPCA-LNM) model, which integrates a low-rank logistic normal framework with a classical probabilistic PCA. To infer the maximum likelihood estimates of latent factors and microbial compositions, they developed an iterative algorithm based on a variational approximation (Ormerod and Wand [2010], Blei et al. [2017], Tran et al. [2021]) within an empirical Bayes framework. More recently, Ba et al. [2024] extended the work of Zeng et al. [2023] to a fully Bayesian setup.

Despite these advancements, a critical limitation remains in that issue (iii) listed above has not been addressed. Real datasets frequently exhibit substantial skewness—particularly at lower taxonomic levels such as genus or species—in the transformed abundance ratios, suggesting the inadequateness of the Gaussian assumption (Tu et al. [2024]). As demonstrated by Montanari and Viroli [2010], factor models based on the Gaussian assumption can be severely mis-specified in the presence of strong asymmetry. We know that the log-ratio-transformation is a standard preprocessing step in microbiome data analysis. If the underlying microbial relative abundances are skewed, ignoring such skewness could lead to model misspecification and biased inference.

Motivated by this, we introduce the Zero-Inflated Factor Analysis Logistic Skew Normal Multinomial (ZIFA-LSNM) model, a comprehensive Bayesian framework that unifies the critical statistical challenges inherent in microbiome data. Our approach integrates a zero-inflation component to explicitly account for excessive zeros and utilizes log-ratio transformation in conjunction with Bayesian factor analysis. In order to capture the asymmetry and heavy-tailed distributions commonly observed in ALR-transformed data, our model imposes skew-normal priors on the latent factors. This is a key departure from conventional factor analysis, both classical and Bayesian, which typically assumes the latent factors to be standard Gaussian.

The article is organized as follows. Section 2 formally introduces the ZIFA-LSNM model and describes the computation of the estimators. To ensure computational efficiency and scalability for high-dimensional datasets, we develop a variational inference algorithm that reformulates the inference challenge as an optimization task, thereby avoiding the computational expense of MCMC methods which often struggle with the complex, multimodal posteriors induced by zero-inflation and skewness.

Section 3 presents several simulation studies and demonstrates the model's application to the real dataset. To motivate and understand the use of skew-normality in our ZIFA-LSNM framework, we conduct a simulation study examining how skewness in the latent space propagates through the full compositional pipeline to the observed log-ratio-transformed counts. As a result, we have showed in model simulation results (Section 3.2) and in real data analysis (Section 3.3) that employing a skew-normal distribution provides substantially improved model fit and hence, more interpretable latent structures. This flexibility enables ZIFA-LSNM to achieve a robust representation of the underlying microbial community. We summarize our findings with a discussion and conclusion in Sections 4 and 5, respectively.

2. Methods

2.1 ZIFA-LSNM model

Let $\mathbf{X}^{n\times p} \in (\mathbb{N} \cup \{0\})^{n\times p}$ denote the $n \times p$ microbiome count matrix, where x_{ij} represents the observed read count for taxon j in sample i. The sequencing depth for the ith sample is defined as $M_i = \sum_{j=1}^{p} x_{ij}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. We model the count vector for sample i, denoted by $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})^T$ as following a multinomial distribution with total count M_i and probability vector $\boldsymbol{\rho_i} = (\rho_{i1}, \ldots, \rho_{ip})^T$. This probability vector $\boldsymbol{\rho_i}$ is

an element of the p-simplex $\mathcal{S}^p$, satisfying $0 < \rho_{ij} < 1$ and $\sum_{j=1}^{p} \rho_{ij} = 1$. In this article, we map $\boldsymbol{\rho_i}$ from the simplex to an unconstrained real space using the ALR transformation. Let $\boldsymbol{a_i} = (a_{i1}, \ldots, a_{i(p-1)})^T \in \mathbb{R}^{p-1}$ be the transformed vector for sample i. Using the pth taxon as the reference, the ALR transformation is defined as:

$$\boldsymbol{a_i} = alr(\boldsymbol{\rho_i}) = \left(\log \frac{\rho_{i1}}{\rho_{ip}}, \ldots, \log \frac{\rho_{i(p-1)}}{\rho_{ip}} \right)^T$$

This ALR transformation is bijective. Hence, the original proportions can be recovered through its inverse,

$$\boldsymbol{\rho_i} = (alr^{-1}(\boldsymbol{a_i})) = \begin{cases} \frac{\exp(a_{ij})}{1+\sum_{k=1}^{p-1} \exp(a_{ik})} & \text{if } j = 1, \ldots, p-1 \\ \frac{1}{1+\sum_{k=1}^{p-1} \exp(a_{ik})} & \text{if } j = p \end{cases}$$

Due to their compositional nature, raw microbiome data cannot be directly analyzed using standard factor analysis, which requires data to reside in a Euclidean space. To appropriately apply a (Bayesian) factor analysis framework, we first map the compositional vector $\boldsymbol{\rho_i}$ from the simplex to $\mathbb{R}^{p-1}$ using the ALR transformation. We then model each element of the transformed vector $\boldsymbol{a_i}$ as a linear combination of a k-dimensional vector of latent factors $\boldsymbol{F}_i = (F_{i1}, \ldots, F_{ik})^T$:

$$a_{ij} = \beta_{0j} + \boldsymbol{F}_i^T \boldsymbol{\beta}_j$$

where β_{0j} represents the taxon-specific intercept and $\boldsymbol{\beta_j} = (\beta_{j1}, \ldots, \beta_{jk})^T$ is the corresponding $k-$dimensional vector of factor loadings. We further place priors on the latent factors and loadings. Specifically, we assume a skew-normal distribution for each latent factor F_{it} and a normal prior for each factor loading β_{jt}.

To account for the excess of zeros in microbiome data, we introduce a latent binary variable, z_{ij}. This variable follows a taxon-specific Bernoulli distribution, i.e., $z_{ij} \sim \mathcal{B}\text{ern}(\kappa_j)$ where κ_j is the probability of zero-inflation for taxon j. Specifically,

$$\rho_{ij} = \begin{cases} \frac{\exp\left(\beta_{0j} + \boldsymbol{F}_i^T \boldsymbol{\beta}_j\right)}{\sum_{j=1}^{p} \exp\left(\beta_{0j} + \boldsymbol{F}_i^T \boldsymbol{\beta}_j\right)} & \text{if } z_{ij} = 0 \text{ with probability } 1 - \kappa_j \\ 0 & \text{if } z_{ij} = 1 \text{ with probability } \kappa_j \end{cases}$$

We implicitly assume that $x_{ij} = 0$ if $\rho_{ij} = 0$.

For $i = 1, \ldots, n,\ j = 1, \ldots, p,\ t = 1, \ldots, k$, we propose a Bayesian zero inflated factor analysis logistic skew normal multinomial (ZIFA-LSNM) model of the form,

$$\boldsymbol{x_i} \mid \boldsymbol{\rho_i}, M_i \overset{\text{ind}}{\sim} \mathcal{MN}(\boldsymbol{\rho_i}, M_i), \quad \text{where} \quad \rho_{ij} = \frac{(1 - z_{ij}) \exp\left(\beta_{0j} + \boldsymbol{F}_i^T \boldsymbol{\beta}_j\right)}{\sum_{j=1}^{p} (1 - z_{ij}) \exp\left(\beta_{0j} + \boldsymbol{F}_i^T \boldsymbol{\beta}_j\right)},$$

with priors,

$$
\begin{aligned}
z_{ij} &\overset{\text{ind}}{\sim} \mathcal{B}\text{ern}(\kappa_j), & \kappa_j &\overset{\text{iid}}{\sim} \mathcal{B}\text{eta}(\nu_1, \nu_2),\\
\beta_{0j} &\overset{\text{iid}}{\sim} \mathcal{N}(0,1), & F_{it} &\overset{\text{iid}}{\sim} \mathcal{SN}(\alpha^*, 1, \underline{\alpha}),\\
\beta_{jt} &\overset{\text{ind}}{\sim} \mathcal{N}(0, \delta_{jt}^{-1}), & \delta_{jt} &\overset{\text{iid}}{\sim} \mathcal{G}\text{amma}(g_1, g_2).
\end{aligned}
$$

Here, $\mathcal{MN}$, $\mathcal{B}$ern, $\mathcal{B}$eta, and $\mathcal{N}$ denote multinomial, Bernoulli, beta, and normal distributions, respectively. The vectors $\boldsymbol{F}_i$ represent unobserved environmental factors or latent scores and $\boldsymbol{\beta_j}$ denote the corresponding factor loadings, where k is the number of latent factors. The notation $\mathcal{SN}(\alpha^*, 1, \underline{\alpha})$ refers to the skew normal distribution with location, scale, and shape parameters given by $\alpha^*, 1$, and $\underline{\alpha}$ respectively, where

$$\alpha^* = -\sqrt{\tfrac{2}{\pi}}\tfrac{\underline{\alpha}}{\sqrt{1+\underline{\alpha}^2}}.$$

The location parameter α^* is chosen such that $\mathbb{E}[F_{it}] = 0$, consistent with the assumption of zero-mean latent factors in classical factor analysis. The key innovation is that we use a skew-normal priors on the latent factors to explicitly model data skewness. Moreover, it is worth mentioning that we use informative normal-gamma shrinkage priors on each element of factor loadings β_{jt} where δ_{jt} is the local shrinkage parameter. Moreover, the underlying compositions are given by

$$\rho_{ij} = \frac{\exp\left(\beta_{0j} + \boldsymbol{F}_i^T \boldsymbol{\beta}_j\right)}{\sum_{j=1}^{p} \exp\left(\beta_{0j} + \boldsymbol{F}_i^T \boldsymbol{\beta}_j\right)}.$$

The hierarchical architecture of the ZIFA-LSNM model, characterized by multiple layers of latent variables and non-conjugate components, results in an analytically intractable posterior distribution. While MCMC is a gold standard for asymptotically exact inference (Robert and Casella [2004]), its computational burden is often prohibitive for the high-dimensional datasets typical in microbiome research. Therefore, to ensure our method is both scalable and efficient, we adopt variational inference, which reframes the challenge of posterior approximation as an optimization problem.

2.2 Variational Inference

Let $\boldsymbol{\Theta} = \left(\boldsymbol{B}, \boldsymbol{\Delta}, \boldsymbol{\kappa}, \boldsymbol{\beta}_0, \boldsymbol{F}, \boldsymbol{z}\right)$ be the set of all model parameters and latent variables where, $\boldsymbol{B} = \{\beta_{jt}\}_{1\le j\le p,\, 1\le t\le k}$, $\boldsymbol{\Delta} = \{\delta_{jt}\}_{1\le j\le p,\, 1\le t\le k}$, $\boldsymbol{\kappa} = \{\kappa_j\}_{1\le j\le p}$, $\boldsymbol{\beta}_0 = \{\beta_{0j}\}_{1\le j\le p}$, $\boldsymbol{F} = \{F_{it}\}_{1\le i\le n,\, 1\le t\le k}$, and $\boldsymbol{z} = \{z_{ij}\}_{1\le i\le n,\, 1\le j\le p}$. The posterior $p(\boldsymbol{\Theta} \mid \boldsymbol{X})$ is analytically intractable because the marginal likelihood,

$$p(\boldsymbol{X}) = \int p(\boldsymbol{X} \mid \boldsymbol{\Theta}) p(\boldsymbol{\Theta}) d\boldsymbol{\Theta}$$

has no closed-form solution.

Variational inference provides a computational alternative to MCMC. It reframes the inference problem as an optimization problem, where the posterior distribution $p(\boldsymbol{\Theta} \mid \boldsymbol{X})$ is approximated by a tractable variational distribution $q(\boldsymbol{\Theta})$ belonging to some known family of densities, $\mathcal{F}$ (Bishop and Nasrabadi [2006], Ormerod and Wand [2010], Blei

et al. [2017]). The best approximation, $q^* \in \mathcal{F}$, is found by minimizing the reverse Kullback-Leibler (rKL) divergence between $q(\boldsymbol{\Theta})$ and $p(\boldsymbol{\Theta} \mid \boldsymbol{X})$,

$$q^* = \arg\min_{q \in \mathcal{F}} rKL(q(\boldsymbol{\Theta}) \mid\mid p(\boldsymbol{\Theta} \mid \boldsymbol{X})).$$

where,

$$\begin{aligned} rKL(q(\boldsymbol{\Theta}) \mid\mid p(\boldsymbol{\Theta} \mid \boldsymbol{X})) &= \int q(\boldsymbol{\Theta}) \log \left(\frac{q(\boldsymbol{\Theta})}{p(\boldsymbol{\Theta} \mid \boldsymbol{X})} \right) d\boldsymbol{\Theta} \\ &= -\mathbb{E}_q \left[\log p(\boldsymbol{X}, \boldsymbol{\Theta}) - \log q(\boldsymbol{\Theta}) \right] + \log p(\boldsymbol{X}) \end{aligned}$$

It is clear that minimizing the rKL-divergence is equivalent to maximizing the evidence lower bound (ELBO) where,

$$\text{ELBO} = \mathbb{E}_q \left[\log p(\boldsymbol{X}, \boldsymbol{\Theta}) - \log q(\boldsymbol{\Theta}) \right]. \tag{1}$$

The complexity of the variational family dictates the difficulty of the optimization. To ensure a tractable inference procedure, we adopt the mean-field approximation, which assumes the variational distribution $q(\boldsymbol{\Theta})$ fully factorizes over all parameters and latent variables as follows:

$$q(\boldsymbol{\Theta}) = \left(\prod_{j=1}^{p} \prod_{t=1}^{k} q(\beta_{jt}) q(\delta_{jt}) \right) \left(\prod_{j=1}^{p} q(\kappa_j) q(\beta_{0j}) \right) \left(\prod_{i=1}^{n} \prod_{t=1}^{k} q(F_{it}) \right) \left(\prod_{i=1}^{n} \prod_{j=1}^{p} q(z_{ij}) \right)$$

We specify a parametric form for each parameter in the variational family by selecting distributions that balance natural conjugacy with computational convenience. Specifically, we define,

$$\begin{aligned} q(\beta_{jt}) &\sim \mathcal{N}(r_{jt}, \lambda_{jt}^2), & q(\delta_{jt}) &\sim \mathcal{G}\text{amma}(g_{1jt}, g_{2jt}), \\ q(\kappa_j) &\sim \mathcal{B}\text{eta}(\tau_{j1}, \tau_{j2}), & q(\beta_{0j}) &\sim \mathcal{N}(a_{0j}, C_{0j}^2), \\ q(F_{it}) &\sim \mathcal{SN}(\xi_{it}, \omega_{it}, \alpha_{it}), & q(z_{ij}) &\sim \mathcal{B}\text{ern}(\pi_{ij}). \end{aligned}$$

The goal of the variational inference procedure is then to optimize the set of variational parameters,

$$\boldsymbol{\Theta}^{\text{var}} = \{ r_{jt}, \lambda_{jt}^2, g_{1jt}, g_{2jt}, \tau_{j1}, \tau_{j2}, a_{0j}, C_{0j}^2, \xi_{it}, \omega_{it}, \alpha_{it}, \pi_{ij} : 1 \leq i \leq n, 1 \leq j \leq p, \, 1 \leq t \leq k \}$$

Now, calculating the ELBO in (1) requires computing the expectations $\mathbb{E}_q \left[\log p(\boldsymbol{X}, \boldsymbol{\Theta}) \right]$ and $\mathbb{E}_q \left[\log q(\boldsymbol{\Theta}) \right]$, which becomes analytically challenging. Therefore, we use Taylor ex-

pansions to evaluate these expressions,

$$
\begin{aligned}
\mathbb{E}_q[\log p(\boldsymbol{X},\boldsymbol{\Theta})] \approx T_2 = & \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}a_{0j} + \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{t=1}^{k} x_{ij}r_{jt}\left(\xi_{it} + \frac{\omega_{it}\alpha_{it}}{\sqrt{1+\alpha_{it}^2}}\sqrt{\frac{2}{\pi}}\right) \\
& - \sum_{i=1}^{n} M_i \log\left(\sum_{j=1}^{p}(1-\pi_{ij})\exp\left(a_{0j} + \frac{1}{2}C_{0j}^2 + \boldsymbol{L}_{ij}\right)\right) \\
& + \sum_{j=1}^{p}\sum_{t=1}^{k}\left(\left(g_1 - \frac{1}{2}\right)(\psi(g_{1jt}) - \log g_{2jt}) - g_2\frac{g_{1jt}}{g_{2jt}} - \frac{1}{2}\left(\lambda_{jt}^2 + r_{jt}^2\right)\frac{g_{1jt}}{g_{2jt}}\right) \\
& - \sum_{i=1}^{n}\sum_{t=1}^{k}\left(\frac{\omega_{it}^2}{2} + \frac{\xi_{it}^2}{2} + \sqrt{\frac{2}{\pi}}\frac{\xi_{it}\omega_{it}\alpha_{it}}{\sqrt{1+\alpha_{it}^2}} - \alpha^*\left(\xi_{it} + \frac{\omega_{it}\alpha_{it}}{\sqrt{1+\alpha_{it}^2}}\sqrt{\frac{2}{\pi}}\right)\right. \\
& \left. - \Psi_1^{AP}(\xi_{it},\omega_{it},\alpha_{it})\right) - \sum_{j=1}^{p}\frac{1}{2}\left(a_{0j}^2 + C_{0j}^2\right) \\
& + \sum_{j=1}^{p}\left[(\nu_1 - 1)\{\psi(\tau_{j1}) - \psi(\tau_{j1}+\tau_{j2})\} + (\nu_2 - 1)\{\psi(\tau_{j2}) - \psi(\tau_{j1}+\tau_{j2})\}\right] \\
& + \sum_{i=1}^{n}\sum_{j=1}^{p}\left[\pi_{ij}\{\psi(\tau_{j1}) - \psi(\tau_{j1}+\tau_{j2})\} + (1-\pi_{ij})\{\psi(\tau_{j2}) - \psi(\tau_{j1}+\tau_{j2})\}\right],
\end{aligned}
$$

and,

$$
\begin{aligned}
\mathbb{E}_q[\log q(\boldsymbol{\Theta})] \approx T_1 = & \sum_{j=1}^{p}\sum_{t=1}^{k}\left(-\frac{1}{2}\log\lambda_{jt}^2 - g_{1jt} + \log g_{2jt} - \log\Gamma g_{1jt} - (1-g_{1jt})\psi(g_{1jt})\right) \\
& + \sum_{j=1}^{p}\Big((\tau_{j1}-1)\{\psi(\tau_{j1}) - \psi(\tau_{j1}+\tau_{j2})\} + (\tau_{j2}-1)\{\psi(\tau_{j2}) - \psi(\tau_{j1}+\tau_{j2})\} \\
& - \log\mathcal{B}(\tau_{j1},\tau_{j2})\Big) - \frac{1}{2}\sum_{j=1}^{p}\log C_{0j}^2 - \sum_{i=1}^{n}\sum_{t=1}^{k}\left(\log\omega_{it} - \Psi_2^{AP}(\alpha_{it})\right) \\
& + \sum_{i=1}^{n}\sum_{j=1}^{p}\left(\pi_{ij}\log\pi_{ij} + (1-\pi_{ij})\log(1-\pi_{ij})\right)
\end{aligned}
$$

Therefore,

$$
\text{ELBO} \approx \text{ELBO}^{\text{AP}} = T_2 - T_1 \tag{2}
$$

where, $\mathcal{B}(,)$ is the beta function, ψ is the digamma function, and

$$
\begin{aligned}
\boldsymbol{L}_{ij} := \sum_{t=1}^{k}\Bigg[\log 2 - \frac{1}{2}\log(1-\lambda_{jt}^2\omega_{it}^2) + \frac{\lambda_{jt}^2\xi_{it}^2 + r_{jt}^2\omega_{it}^2 + 2r_{jt}\xi_{it}}{2(1-\lambda_{jt}^2\omega_{it}^2)} \\
+ \log\left(\Phi\left[\frac{(r_{jt} + \lambda_{jt}^2\xi_{it})\omega_{it}\alpha_{it}}{\sqrt{1-\lambda_{jt}^2\omega_{it}^2}\sqrt{1+\alpha_{it}^2-\lambda_{jt}^2\omega_{it}^2}}\right]\right)\Bigg].
\end{aligned}
$$

Here and throughout the paper, we drop all constants that do not depend on the parameters. The term $\boldsymbol{L_{ij}}$ is sum over $t = 1$ to k, log of moment generating function of product of normal ($q(\beta_{jt})$) and skew-normal ($q(F_{it})$) random variable. Further details

on the derivations and explanations of ELBO^{AP}, $\boldsymbol{L}_{ij}$, $\Psi_1^{AP}(\xi_{it}, \omega_{it}, \alpha_{it})$, and $\Psi_2^{AP}(\alpha_{it})$ can be found in Supplementary Material. Hence, Maximizing ELBO^{AP} with respect to $\boldsymbol{\Theta}^{\text{var}}$ forms the basis of our inference algorithm.

2.3 Parameter estimation

Direct maximization of (2) is difficult due to parameter interactions and the sum of the terms inside the logarithm. We use a variational algorithm (VA) that maximizes (2) by alternating updates of $\{\pi_{ij}\}$ and $\boldsymbol{\Theta}^{\text{var}}/\{\pi_{ij}\}$ in turn, and iterating these steps until convergence of ELBO^{AP}. This is outlined in Algorithm 1.

Algorithm 1: VA for ZIFA-LSNM

Initialize $\{a_{0j}^{(0)}, C_{0j}^{2(0)}, r_{jt}^{(0)}, \lambda_{jt}^{2(0)}, g_{1jt}^{(0)}, g_{2jt}^{(0)}, \tau_{j1}^{(0)}, \tau_{j2}^{(0)}, \xi_{it}^{(0)}, \omega_{it}^{(0)}, \alpha_{it}^{(0)}, \pi_{ij}^{(0)}\}$, and $s = 0$;
while *ELBO^{AP} not converged* **do**
  $s \leftarrow s + 1$;
  given $\{a_{0j}^{(s-1)}, C_{0j}^{2(s-1)}, r_{jt}^{(s-1)}, \lambda_{jt}^{2(s-1)}, \tau_{j1}^{(s-1)}, \tau_{j2}^{(s-1)}, \xi_{it}^{(s-1)}, \omega_{it}^{(s-1)}, \alpha_{it}^{(s-1)}\}$, update $\pi_{ij}^{(s)} = I(\hat{\pi}_{ij} \geq \pi_0)$;
  given $\{r_{jt}^{(s-1)}, \lambda_{jt}^{2(s-1)}\}$, update $g_{1jt}^{(s)}, g_{2jt}^{(s)}$;
  given $\{a_{0j}^{(s-1)}, C_{0j}^{2(s-1)}, g_{1jt}^{(s)}, g_{2jt}^{(s)}, \xi_{it}^{(s-1)}, \omega_{it}^{(s-1)}, \alpha_{it}^{(s-1)}, \pi_{ij}^{(s)}\}$, update $\tau_{j1}^{(s)}, \tau_{j2}^{(s)}, r_{jt}^{(s)}, \lambda_{jt}^{2(s)}$;
  given $\{a_{0j}^{(s-1)}, C_{0j}^{2(s-1)}, r_{jt}^{(s)}, \lambda_{jt}^{2(s)}, \pi_{ij}^{(s)}\}$, update $\xi_{it}^{(s)}, \omega_{it}^{(s)}, \alpha_{it}^{(s)}$;
  given $\{r_{jt}^{(s)}, \lambda_{jt}^{2(s)}, \xi_{it}^{(s)}, \omega_{it}^{(s)}, \alpha_{it}^{(s)}, \pi_{ij}^{(s)}\}$, update $a_{0j}^{(s)}, C_{0j}^{2(s)}$;
end while

Detailed derivations for updating the variational parameters are provided in the Supplementary Material. Note that we obtain the final point estimates for the model parameters by taking the posterior mean of their respective variational distributions,

$$\hat{\kappa}_j = \frac{\hat{\tau}_{j1}}{\hat{\tau}_{j1} + \hat{\tau}_{j2}}, \quad \hat{\beta}_{0j} = \hat{a}_{0j}, \quad \hat{\beta}_{jt} = \hat{r}_{jt}, \quad \hat{F}_{it} = \hat{\xi}_{it} + \sqrt{\frac{2}{\pi}} \frac{\hat{\omega}_{it}\hat{\alpha}_{it}}{\sqrt{1 + \hat{\alpha}_{it}^2}}.$$

Moreover, given these estimates, the underlying compositions ρ_{ij} are then estimated as,

$$\hat{\rho}_{ij} = \frac{\exp\left(\hat{a}_{0j} + \frac{1}{2}\hat{C}_{0j}^2 + \boldsymbol{L}_{ij}\right)}{\sum_{j=1}^{p} \exp\left(\hat{a}_{0j} + \frac{1}{2}\hat{C}_{0j}^2 + \boldsymbol{L}_{ij}\right)}.$$

It is worth noting that we've used the relative difference criteria for the convergence of ELBO^{AP}, i.e.,

$$\Delta^{(s)} = \left| \frac{\text{ELBO}_{(New)}^{\text{AP}} - \text{ELBO}_{(Old)}^{\text{AP}}}{\text{ELBO}_{(Old)}^{\text{AP}}} \right| < 5 \times 10^{-5}$$

Now, given the analytical form of ELBO^{AP} in equation (2), it becomes notoriously difficult to find the update for π_{ij}, primarily due to the log-of-sum terms. Therefore, we employ two computational strategies.

2.3.1 Multinomial-Poisson equivalence

First, we leverage the well-established relationship between the multinomial and Poisson distributions to simplify the update of π_{ij}. As noted by Baker [1994], for every multinomial regression model, one can define an equivalent Poisson regression model by introducing a latent parameter. Therefore, we provide an equivalent Zero Inflated Factor Analysis Logistic Skew Normal Poisson (ZIFA-LSNP) model with the sample-specific latent parameter, Υ_{i0}, which has the following update

$$\Upsilon_{i0}^{(s)} = \log \left\{ \frac{M_i}{\sum_{j=1}^{p} \left(1 - \pi_{ij}^{(s-1)}\right) \exp\left(a_{0j}^{(s-1)} + \frac{1}{2}\left(C_{0j}^2\right)^{(s-1)} + \boldsymbol{L}_{ij}\right)} \right\}.$$

The variational update for π_{ij} is then consequently derived from this more convenient Poisson-based ELBO,

$$\pi_{ij}^{(s)} = \begin{cases} \frac{\exp\left(\psi\left(\tau_{j1}^{(s-1)}\right)\right)}{\exp\left(\psi\left(\tau_{j1}^{(s-1)}\right)\right) + \exp\left(\psi\left(\tau_{j2}^{(s-1)}\right) - \exp\left(\Upsilon_{i0}^{(s)} + a_{0j}^{(s-1)} + \frac{1}{2}\left(C_{0j}^2\right)^{(s-1)} + \boldsymbol{L}_{ij}\right)\right)} & \text{if } x_{ij} = 0 \\ 0 & \text{if } x_{ij} > 0 \end{cases}$$

Note that these updates $\left(\Upsilon_{i0}^{(s)} \text{ and } \pi_{ij}^{(s)}\right)$ are derived from the ELBO of the ZIFA-LSNP model, which is detailed in the Supplementary Material.

2.3.2 Classification Variational Inference Step

Second, inspired by the classification EM algorithm (Celeux and Govaert [1992]), we introduce a classification step for π_{ij}. In the early stages of optimization, when parameter estimates are unreliable, this can lead to slow or unstable convergence (Celeux and Govaert [1992]). The classification step addresses this by converting it into "hard" assignments (0 or 1) at each iteration,

$$\hat{\pi}_{ij}^{(s)} = I(\hat{\pi}_{ij} > \pi_0)$$

where π_0 is a pre-specified threshold for this classification. Following standard practice in binary classification, we set $\pi_0 = 0.5$ as the default value.

2.4 Simulation design

The primary objectives of our simulation study are to assess the ZIFA-LSNM model's ability to: (i) recover the latent structure (loadings and scores); (ii) effectively incorporate zero-inflation and skewness; and (iii) provide reliable estimates of the true underlying microbial compositions.

We compared our model's performance against the recent model ZIPPCA-LPNM (Ba et al. [2024]), which was selected for its demonstrated improvements over its predecessor, ZIPPCA-LNM (Zeng et al. [2023]). We assess performance using the root mean squared error (RMSE). For model fitting, the prior hyperparameters were set to $\nu_1 = 1$, $\nu_2 = 3$, $g_1 = 200$, and $g_2 = 1/200$.

A key challenge in evaluating factor analysis models is the rotational ambiguity of the latent space. Factor loadings and scores are only identifiable up to an orthogonal transformation. Consequently, a direct RMSE calculation is not a fair measure of accuracy. To ensure a fair comparison, we first resolved this rotational ambiguity using a Procrustes transformation. This procedure finds the optimal orthogonal rotation that aligns the estimated factor matrix with the true matrix. All reported RMSE values for factor loadings and scores were therefore calculated post-alignment, providing a robust measure of each model's ability to recover the underlying latent structure.

We simulated 1,000 count datasets designed to mimic the key characteristics of real-world microbiome sequencing data. For each scenario of the number of latent factor (k) and prior value for $\underline{\alpha}$ ($\underline{\alpha} = k = 2$ and $\underline{\alpha} = k = 5$), we consider each combination of sample size $n \in \{50, 100, 1000\}$ and number of taxa $p \in \{50, 100\}$. The specifics of the data-generating processes are detailed in Table 1.

$\underline{\alpha}$	k	M_i	β_{jt}	κ_j	β_{0j}	F_{it}
2	2	$(8 \times 10^4, 10^5)$	$\mathcal{U}(-2.5, 2.5)$	$\mathcal{U}(0, 1)$	$\mathcal{U}(-2, 2)$	$\mathcal{SN}(\alpha^*, 1, \alpha)$
5	5	$(8 \times 10^4, 10^5)$	$\mathcal{U}(-2.5, 2.5)$	$\mathcal{U}(0, 1)$	$\mathcal{U}(-2, 2)$	$\mathcal{SN}(\alpha^*, 1, \alpha)$

Table 1: Specifications of parameters in zero-inflated models

3. Results

3.1 Skewness simulation

Skewness could naturally occur in ALR-transformed count data whose underlying proportions were truly drawn from a Gaussian distribution on the ALR-scale. This is due to multinomial variation and varying numbers of sequences between samples. In this subsection, we present a motivating simulation that demonstrates the presence of significant skewness in the ALR-transformed data beyond what could be explained by multinomial variation.

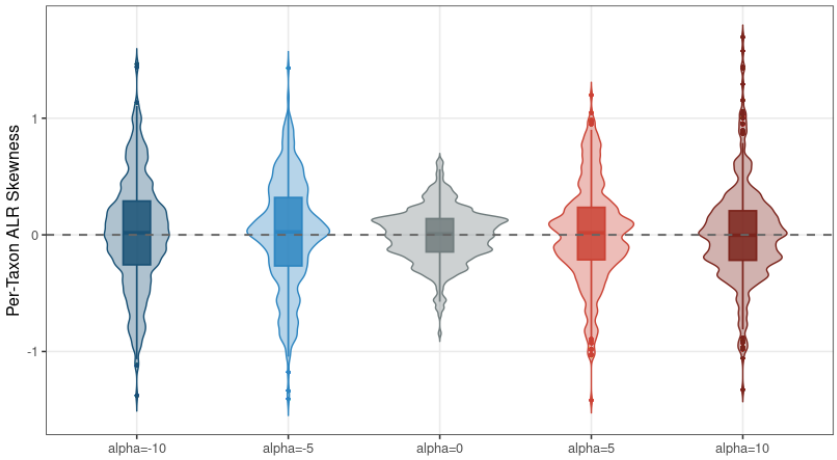


Figure 1: Distribution of Per-Taxon ALR-transformed Skewness

Figure 1 summarizes a simulation study designed to assess how skewness propagates through the compositional pipeline. For each of 100 replicates, we draw an $n \times (p-1)$ matrix from a skew-normal distribution $\mathcal{SN}(0, 1, \alpha)$ with shape parameter $\alpha \in \{-10, -5, 0, 5, 10\}$, sample size $n = 100$, and number of taxa $p = 100$. Note that $\alpha = 0$ reduces to standard normal. Each row of the matrix is mapped to the simplex via the inverse additive log-ratio (ALR) transformation to produce a composition, from which multinomial counts are drawn with sequencing depths $M_i \sim \mathcal{U}(10^6, 2 \times 10^6)$. The ALR transformation is then recomputed on the observed counts, and the sample skewness is recorded for every taxon–replicate pair. Under the Gaussian ($\alpha = 0$), the resulting distribution of per-taxon ALR skewness is concentrated around zero. However, as $|\alpha|$ increases, the distribution becomes more dispersed, indicating that even moderate departures from normality produce detectable skewness in the observed ALR-tansformed counts.

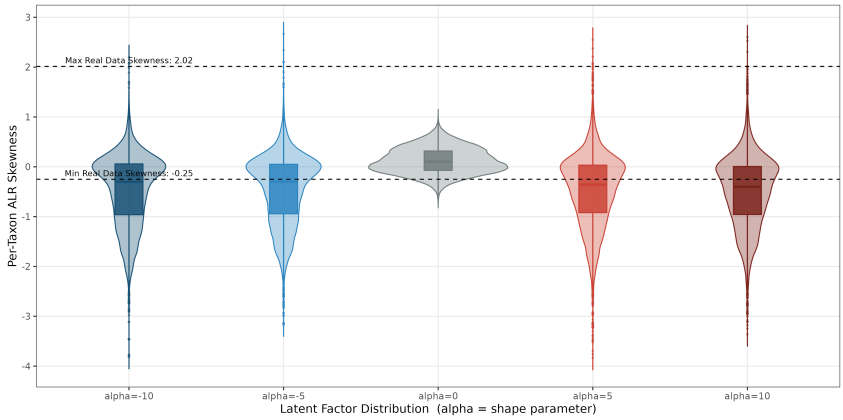


Figure 2: Distribution of Per-Taxon ALR-transformed Skewness from full model

In Figure 2 we repeat the simulation under the complete ZIFA-LSNM generative model with skew-normal latent factors, again varying the shape parameter $\alpha \in \{-10, -5, 0, 5, 10\}$ across 100 replicates. For each generated dataset, we apply the ALR transformation to the observed counts and record the per-taxon skewness. Under this full model-based generative process, the contrast between Gaussian and skew-normal latent factors is even more pronounced. When $\alpha = 0$ (the standard Gaussian latent factors), the ALR skewness values remain concentrated around zero, whereas nonzero values of α yield greater skewness in the ALR-transformed compositions.

Taken together with the empirical evidence of ALR skewness in real microbiome data, which varies from -0.25 to 2.02 (Figure 4 Section 3.3), these simulation results provide strong justification for placing skew-normal priors on the latent factors in the ZIFA-LSNM model.

3.2 Model simulation results

Across all evaluated simulation scenarios from Table 1, our proposed ZIFA-LSNM model outperforms the Gaussian-based ZIPPCA-LPNM model, achieving lower root mean squared errors (RMSE) for all key parameter estimates.

In the scenario with $k = 2$ (Table 2), ZIFA-LSNM maintains its superior performance

across all combinations of sample size (n) and number of taxa (p). Furthermore, as n increases from 50 to 1000, the RMSE for all ZIFA-LSNM parameters consistently decreases. This behavior demonstrates that the model effectively leverages larger sample sizes to improve estimation accuracy, with the estimates empirically converging towards the true parameter values.

In $k = 5$ scenario (Table 3), ZIFA-LSNM continues to outperform ZIPPCA-LPNM. This estimation task is substantially more complex, as the number of variational parameters to be estimated increases significantly. For instance, with $n = 100$, $p = 50$, and $k = 5$, the algorithm estimates 7700 variational parameters. Note that the total number of variational parameters to estimate (for n samples, p taxa, and k factors) is $4pk+4p+3nk+np$.

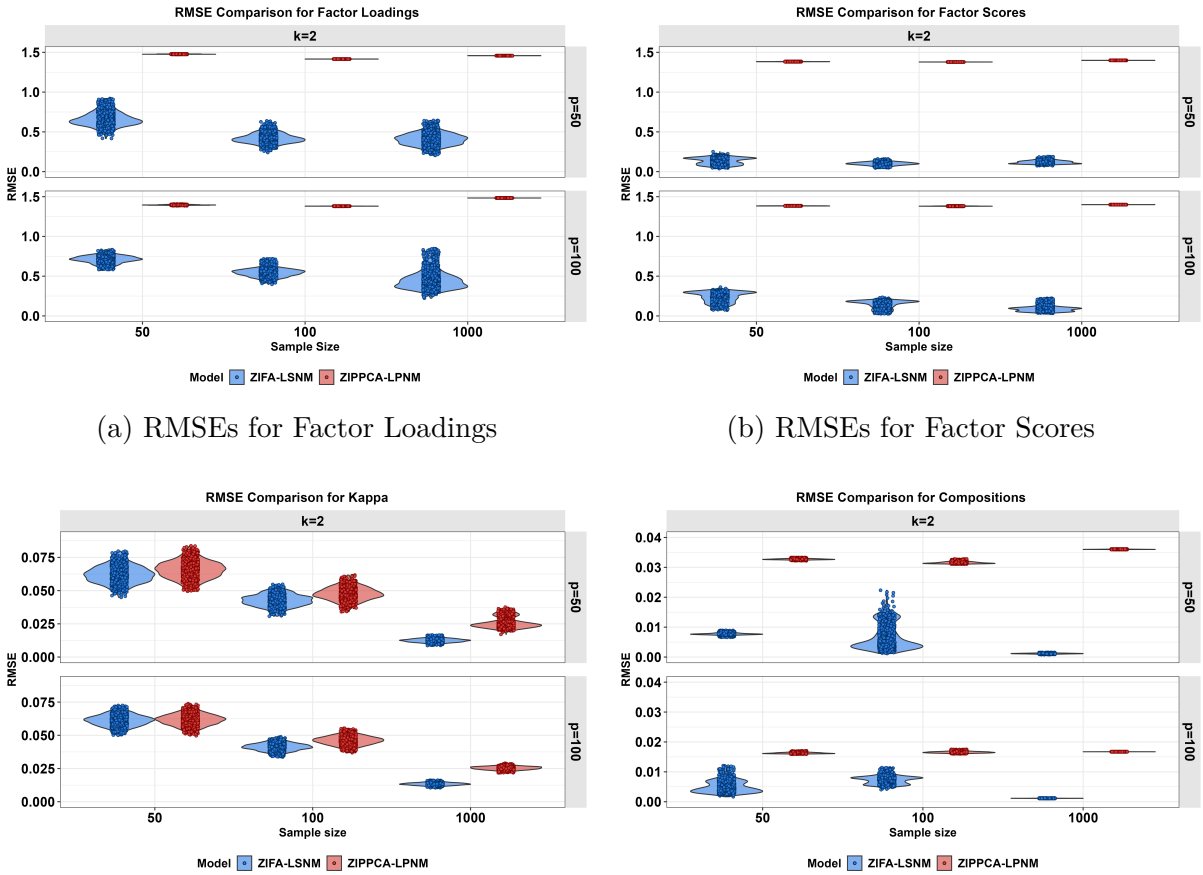
Figure 3 provides a visual summary of the performance for $k = 2$ across all simulation settings. The corresponding plots for $k = 5$ are presented in the Supplementary Material. On average, ZIFA-LSNM consistently achieves lower RMSE for all the parameters of interest. The improvements are even more pronounced in the recovery of the latent factor scores $\boldsymbol{F}$ (Figure 3b). The observed improvement highlights the benefit of modeling skewness. Furthermore, the accuracy extends to the underlying microbial compositions $\boldsymbol{\rho}$ (Figure 3d), where ZIFA-LSNM again achieves notably lower RMSE, indicating a more faithful representation of the microbial community.

$k = 2$	ZIPPCA-LPNM					ZIFA-LSNM				
(n, p)	$\boldsymbol{\kappa}$	$\boldsymbol{B}$	$\boldsymbol{F}$	β_0	$\boldsymbol{\rho}$	$\boldsymbol{\kappa}$	$\boldsymbol{B}$	$\boldsymbol{F}$	β_0	$\boldsymbol{\rho}$
(50,50)	0.0656	1.4772	1.3832	0.9889	0.0327	0.0621	0.6633	0.1328	0.2587	0.0077
(100,50)	0.0477	1.4156	1.3787	1.0967	0.0316	0.0428	0.4200	0.1003	0.3031	0.0068
(1000,50)	0.0260	1.4581	1.3991	1.2993	0.0360	0.0124	0.4107	0.1198	0.5157	0.0011
(50,100)	0.0616	1.3960	1.3842	0.9186	0.0163	0.0612	0.7087	0.2346	0.3573	0.0051
(100,100)	0.0461	1.3808	1.3809	1.0808	0.0165	0.0412	0.5526	0.1423	0.2869	0.0073
(1000,100)	0.0253	1.4835	1.3994	0.9746	0.0167	0.0133	0.4596	0.0989	1.1238	0.0011

Table 2: RMSEs of estimates of the parameters for $\alpha = 2$.

$k = 5$	ZIPPCA-LPNM					ZIFA-LSNM				
(n, p)	$\boldsymbol{\kappa}$	$\boldsymbol{B}$	$\boldsymbol{F}$	β_0	$\boldsymbol{\rho}$	$\boldsymbol{\kappa}$	$\boldsymbol{B}$	$\boldsymbol{F}$	β_0	$\boldsymbol{\rho}$
(50,50)	0.0686	1.4562	2.9507	1.1579	0.0471	0.0600	0.8287	0.1421	0.4740	0.0349
(100,50)	0.0508	1.4339	3.0276	0.9989	0.0485	0.0424	0.8636	0.1790	0.4395	0.0334
(1000,50)	0.0295	1.4397	3.0674	1.1880	0.0478	0.0143	1.4389	0.3035	0.7168	0.0055
(50,100)	0.1090	1.4399	2.9499	1.2331	0.0334	0.0614	0.8818	0.2599	0.4909	0.0106
(100,100)	0.0507	1.4615	3.0274	1.2478	0.0292	0.0420	0.7499	0.1209	0.4396	0.0127
(1000,100)	0.0225	1.4685	3.0661	1.1403	0.0284	0.0164	0.8838	0.2299	0.6732	0.0015

Table 3: RMSEs of estimates of the parameters for $\alpha = 5$.



(a) RMSEs for Factor Loadings

(b) RMSEs for Factor Scores

(c) RMSEs for Kappas

(d) RMSEs for Compositions

Figure 3: Root Mean Squared Error Plots for $k = 2$

3.3 Data application and results

In this section, we applied the ZIFA-LSNM model to a publicly available dataset from a family-based study of inflammatory bowel disease (IBD) (Muller et al. [2022], Jacobs et al. [2016]). The dataset comprises 16S rRNA sequencing data from 90 participants: 54 healthy controls and 36 IBD patients (26 with Crohn's disease [CD] and 10 with ulcerative colitis [UC]).

The raw count data were aggregated to the genus level, with unclassified taxa removed. To reduce extreme sparsity, we applied three standard prevalence filters: genera were retained if the proportion of zeros was below 97%, if they were present in at least 5% of samples, and if the cumulative read count exceeded 20. The filtered dataset comprised 178 genera across 90 samples.

A key motivation for our model is the pronounced skewness in ALR-transformed counts for each taxon. To examine that, we computed the sample skewness of ALR-transformed counts for each taxon (Figure 4). For this exploratory visualization, a pseudo-count of 10^{-8} was added to accommodate zero counts in the log-ratio transformation. We emphasize that this ad-hoc adjustment is not required by the ZIFA-LSNM model itself, which handles zeros explicitly through its zero-inflation component. As shown in Figure 4, the majority of genera exhibit positive skewness in their ALR-transformed counts, with 58

% of taxa exceeding a skewness of 0.5 and 30% of taxa exceeding a skewness of 1. This empirical observation provides motivation for the skew-normal latent factor specification. Note that, we calculate the skewness using the `skewness()` function in `moments` package in `R`.

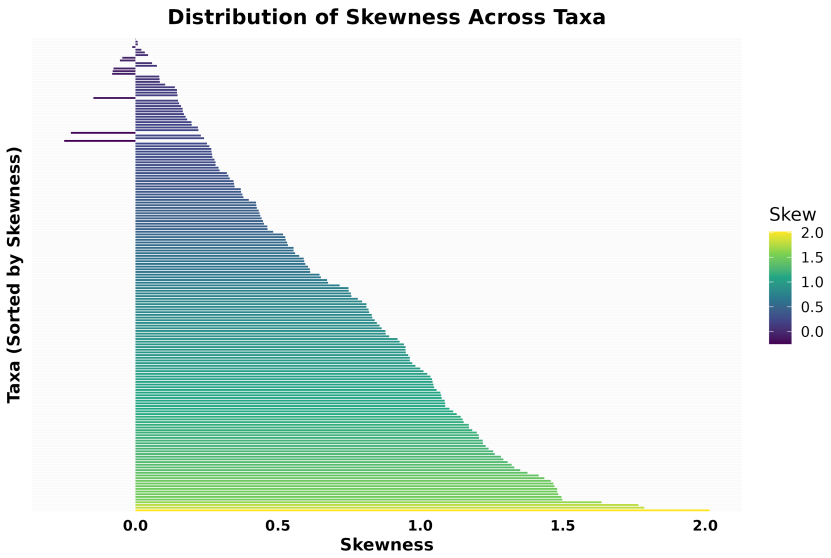


Figure 4: Per-Taxon Skewness of ALR transformed Counts of real data

We fit the ZIFA-LSNM model with $k = 3$ latent factors. The resulting pairwise scatter plots of estimated factor scores are shown in Figure 5, with samples colored by diagnosis. Note that the estimated factor scores are obtained after the `varimax` rotation. Visual inspection suggests that the second latent factor (V2) exhibits the most apparent differences across diagnostic groups: in the V2-versus-V1 and V3-versus-V2 plots, the healthy control samples appear to cluster more tightly near the center of the factor space, while IBD samples (both CD and UC) tend to be displaced along the V2 axis. In contrast, the V3-versus-V1 plot shows considerable overlap among all three groups. We also fit the model with $k = 2, 4, 5$, and the factor scores plots are provided in Supplementary Material.

For comparison, we also fit the ZIPPCA-LPNM model with k = 3 factors, and the corresponding factor scores scatter plots are displayed in Figure 6. Visual comparison of Figures 5 and 6 suggests that the ZIFA-LSNM model yields more compact clustering of healthy controls and more discernible displacement of IBD samples. Hence, these plots indicate that the skew-normal latent factor specification provides a substantially improved representation of the latent structure compared to the Gaussian-based alternative.

Furthermore, to evaluate the discriminatory power of our methodology in distinguishing between the two groups (Normal vs. Diseased [UC + CD]) in the IBD dataset, we performed logistic regression for models ZIFA-LSNM and ZIPPCA LPNM, using the three factors obtained from the respective model fitting as independent variables. The predictive performance was then quantified by computing the corresponding area under the ROC curve (AUC) values. The results are presented in Figure 7. Our model yields higher AUC value (77.42% for ZIFA-LSNM and 74.18% for ZIPPCA-LPNM), suggesting

strong discriminatory performance between the two groups.

We interpret the biological relevance of the latent structure as Figure 8 displays the estimated factor loadings for the top 25 genera contributing to each factor. Of particular interest are the taxa with loadings on V2, as this factor was most associated with IBD status. Several of these genera have been previously implicated in IBD pathogenesis (Kostic et al. [2014], Singh et al. [2023], Tharu et al. [2024]). These associations are consistent with the hypothesis that V2 captures a biologically meaningful axis of microbial variation related to gut inflammation.

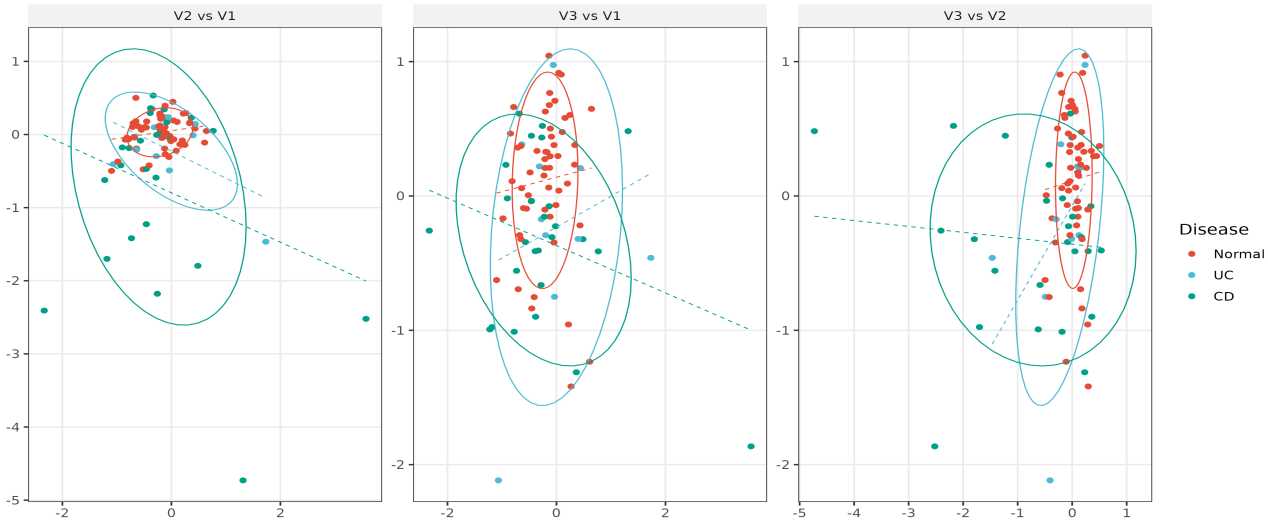


Figure 5: **Latent Microbial Factors Distinguish Healthy Controls from IBD Patients.** Pairwise scatter plots of factor scores from the three latent variables (V1, V2, V3) via ZIFA-LSNM model. Each point corresponds to an individual sample and is colored by diagnosis.

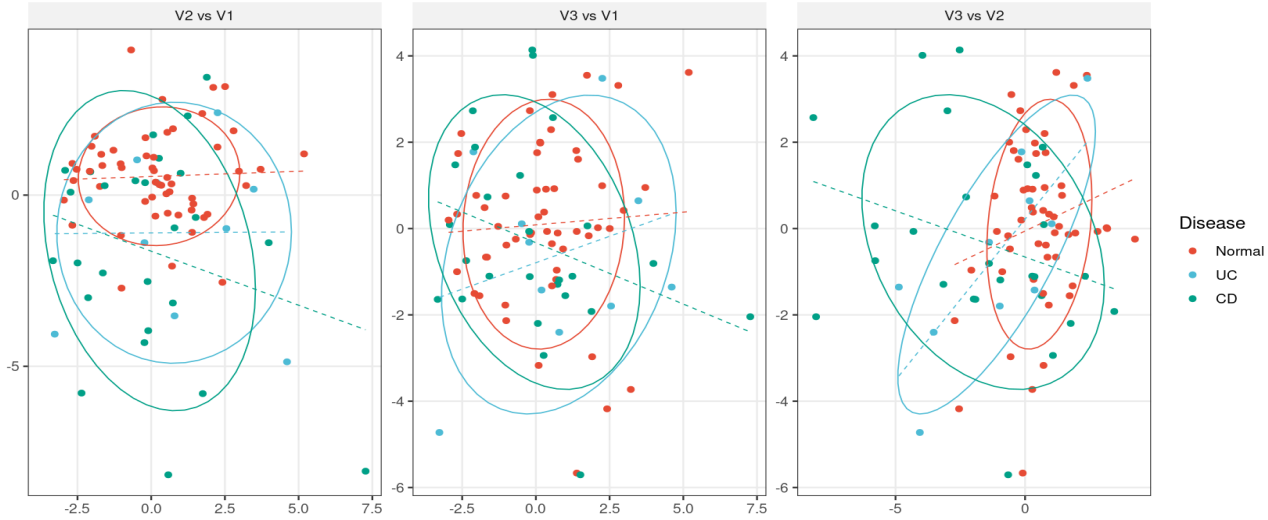


Figure 6: **Latent Microbial Factors Distinguish Healthy Controls from IBD Patients.** Pairwise scatter plots of factor scores from the three latent variables (V1, V2, V3) via ZIPPCA-LPNM model. Each point corresponds to an individual sample and is colored by diagnosis.

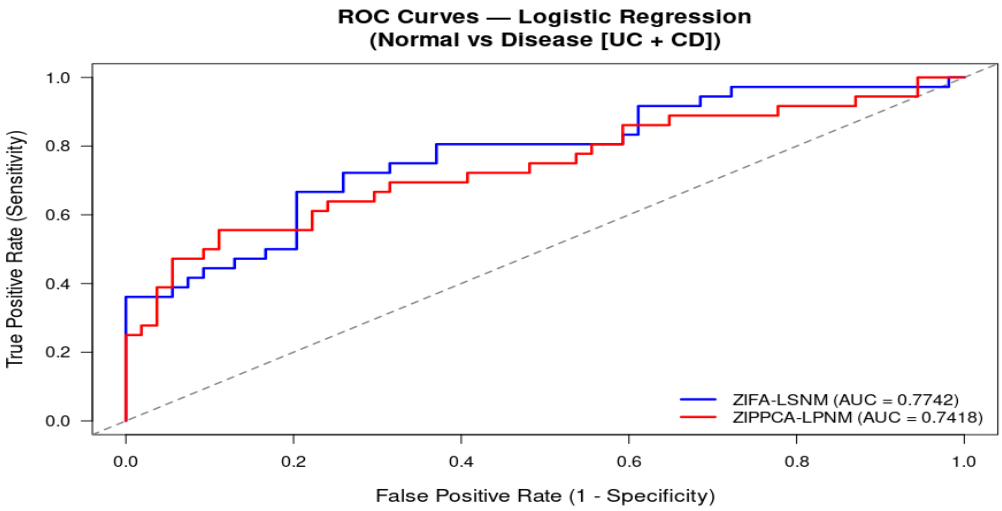
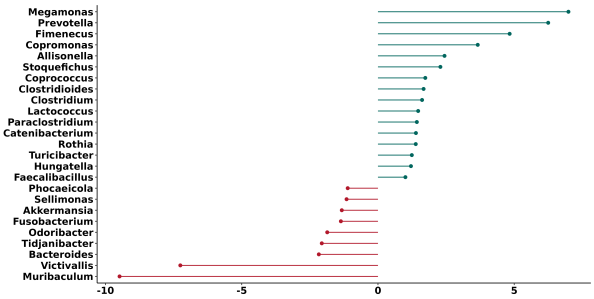
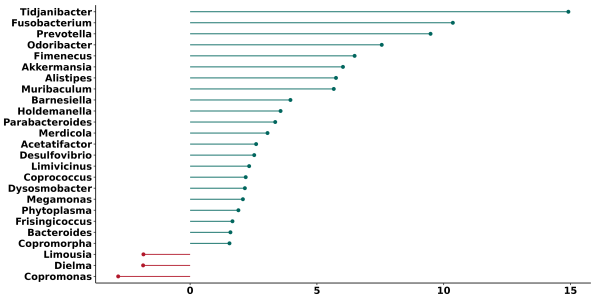


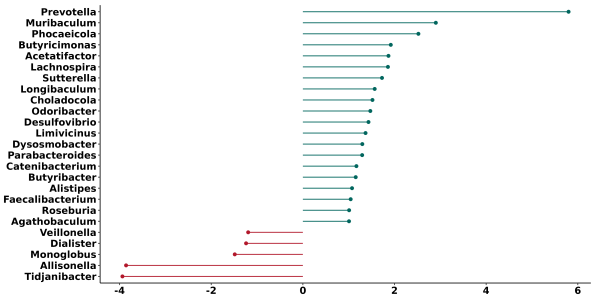
Figure 7: **AUC plot for ZIFA-LSNM and ZIPPCA LPNM**



(a) Factor 1 Loading Weights



(b) Factor 2 Loading Weights



(c) Factor 3 Loading Weights

Figure 8: **Top 25 microbial contributors to the three latent factors via ZIFA-LSNM model.** These plots display the factor loadings for the top 25 microbial genera with the largest weights for Factor 1 (top), Factor 2 (bottom left), and Factor 3 (bottom right). The length of each bar indicates the magnitude of the loading, representing the strength of association between a genus and the factor. Loadings are colored by their sign, indicating a positive (teal) or negative (red). These plots identify the key taxa that define each microbial association.

4. Discussion

The analysis of microbiome data presents a unique set of statistical challenges, primarily the need to achieve effective dimension reduction within a framework that rigorously accounts for both compositionality and zero inflation. While progress has been made, the prevalent issue of skewness in log-ratio-transformed compositions has been largely overlooked. This paper introduces the ZIFA-LSNM model, a Bayesian hierarchical framework designed to address these statistical challenges while addressing skewness through the use of skew-normal priors on the latent factors.

Our simulation studies provide evidence of improved performance of ZIFA-LSNM compared to its Gaussian-based counterparts. Across a range of scenarios, our model demonstrated uniformly lower RMSEs in parameter recovery and composition estimation. These findings suggest that explicitly modeling skewness can enhance both the statistical robustness and biological interpretability of microbiome data. Furthermore, our adoption of a variational inference framework ensures computational scalability, making the approach feasible for modern, large-scale microbiome studies.

Our approach uses variational inference to find the estimates. The criterion function that is optimized here, ELBO^{AP}, is non-decreasing by construction at each iteration and is bounded above by the log marginal likelihood or log evidence, i.e., $\log p(\boldsymbol{X})$. Consequently, the sequence of ELBO^{AP} values is assured to converge to some limiting value, however, the algorithm is not guaranteed to converge to the global optimum. There are multiple strategies that could be explored to avoid getting stuck in local optima. First, one may utilize multiple random initializations for different variational parameters (Bishop and Nasrabadi [2006], Blei et al. [2017]). Alternatively, initializing variational parameters for factor loadings and factor scores using a low-rank decomposition such as PCA or singular value decomposition (SVD) often improves convergence and reduces sensitivity to poor local optima (Tipping and Bishop [1999]). Here, we have used the SVD initialization approach to address the potential convergence issue.

While the proposed model is effective, it possesses limitations that highlight opportunities for future research. Although our variational inference framework is scalable, its computational complexity can be substantial when fitting count-based models to high-dimensional microbiome data with large sample sizes (n) and many taxa (p). This can present challenges in ultra-high-dimensional settings. Although our informative local shrinkage priors provide effective regularization, more advanced shrinkage schemes could be explored. For instance, incorporating Gamma process global-local shrinkage priors (Bhattacharya and Dunson [2011], Hansen et al. [2025]), could further enhance factor recovery. Moreover, in the current article, the number of latent factors, k, is treated as a fixed value that must be specified in advance. In the absence of prior knowledge, a more sophisticated selection methods can be used to find k, such as cross-validation or information criteria like the BIC or Variational BIC (You et al. [2014]) that allow k to be inferred directly from the data.

5. Conclusions

We have established that the ZIFA-LSNM provides improved performance over previous models when skewness is present. The new model therefore solves an established issue of microbiome data analysis (Tu et al. [2024]).

6. Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: `zifalsnm` is implemented in a freely available `R` package and one can find all the reproducible R codes, real data, simulation results on GitHub: `https://github.com/SaurabhP-MS/zifalsnm.git`. Note that we have used publicly available real microbiome data and proper citations are included.

Competing Interests: The authors declare that they have no competing interest.

Funding: K.M. acknowledges support from the Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant: RGPIN-2021-03634). The authors gratefully acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Authors' contributions: SP contributed in methodological theory and application, coded, wrote an R package, conducted simulation studies, created and synthesized all tables and figures, performed real data analysis, primary manuscript preparation; HJ & KM contributed in methodological theory and application, supervised, manuscript preparation. All authors read and approved the final manuscript.

Acknowledgements: Not applicable.

References

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

I. Ba, M. Turgeon, S. Veniamin, J. Joel, R. Miller, M. Graham, C. Bonner, C. N. Bernstein, D. L. Arnold, A. Bar-Or, et al. A Bayesian factor analysis model for high-dimensional microbiome count data. *arXiv preprint arXiv:2404.03127*, 2024.

S. G. Baker. The multinomial-Poisson transformation. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 43(4):495–504, 1994.

A. Bhattacharya and D. B. Dunson. Sparse Bayesian infinite factor models. *Biometrika*, 98(2):291–306, 2011.

C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

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

G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data analysis*, 14(3):315–332, 1992.

E. Z. Chen and H. Li. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, 32(17):2611–2617, 2016.

K. E. Fujimura and S. V. Lynch. Microbiota in allergy and asthma and the emerging relationship with the gut microbiome. *Cell Host & Microbe*, 17(5):592–602, 2015.

G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. Microbiome datasets are compositional: and this is not optional. *Frontiers in Microbiology*, 8:2224, 2017.

B. Hansen, A. Avalos-Pacheco, M. Russo, and R. De Vito. Fast variational inference for Bayesian factor analysis in single and multi-study settings. *Journal of Computational and Graphical Statistics*, 34(1):96–108, 2025.

J. P. Jacobs, M. Goudarzi, N. Singh, M. Tong, I. H. McHardy, P. Ruegger, M. Asadourian, B.-H. Moon, A. Ayson, J. Borneman, et al. A disease-associated microbial and metabolomics state in relatives of pediatric inflammatory bowel disease patients. *Cellular and Molecular Gastroenterology and Hepatology*, 2(6):750–766, 2016.

A. D. Kostic, R. J. Xavier, and D. Gevers. The microbiome in inflammatory bowel disease: current status and the future ahead. *Gastroenterology*, 146(6):1489–1499, 2014.

R. E. Ley. Obesity and the human microbiome. *Current Opinion in Gastroenterology*, 26 (1):5–11, 2010.

K. C. Lutz, S. Jiang, M. L. Neugent, N. J. De Nisco, X. Zhan, and Q. Li. A survey of statistical methods for microbiome data analysis. *Frontiers in Applied Mathematics and Statistics*, 8:884810, 2022.

C. Manichanh, N. Borruel, F. Casellas, and F. Guarner. The gut microbiota in IBD. *Nature Reviews Gastroenterology & Hepatology*, 9(10):599–608, 2012.

A. Montanari and C. Viroli. A skew-normal factor model for the analysis of student satisfaction towards university courses. *Journal of Applied Statistics*, 37(3):473–487, 2010.

E. Muller, Y. M. Algavi, and E. Borenstein. The gut microbiome-metabolome dataset collection: a curated resource for integrative meta-analysis. *npj Biofilms and Microbiomes*, 8(1):79, 2022.

J. T. Ormerod and M. P. Wand. Explaining variational approximations. *The American Statistician*, 64(2):140–153, 2010.

J. N. Paulson, O. C. Stine, H. C. Bravo, and M. Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, 10(12):1200–1202, 2013.

C. Robert and G. Casella. *Monte Carlo Statistical Methods*, volume 2. Springer, 2004.

S. B. Singh, A. Carroll-Portillo, and H. C. Lin. Desulfovibrio in the gut: the enemy within? *Microorganisms*, 11(7):1772, 2023.

Z.-Z. Tang and G. Chen. Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*, 20(4):698–713, 2019.

R. Tharu, G. M. Ahlawat, S. Kushwaha, and P. Khanna. Gut microbiota disparities between active crohn's disease and healthy controls: A global systematic review. *Clinical Epidemiology and Global Health*, 25:101497, 2024.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622, 1999.

M.-N. Tran, T.-N. Nguyen, and V.-H. Dao. A practical tutorial on variational Bayes. *arXiv preprint arXiv:2103.01327*, 2021.

W. Tu, R. Browne, and S. Subedi. A mixture of logistic skew-normal multinomial models. *Computational Statistics & Data Analysis*, 196:107946, 2024.

A. D. Tyler, M. I. Smith, and M. S. Silverberg. Analyzing the human microbiome: a "how to" guide for physicians. *Official Journal of the American College of Gastroenterology—ACG*, 109(7):983–993, 2014.

K. G. Van den Boogaart and R. Tolosana-Delgado. *Analyzing compositional data with R*, volume 122. Springer, 2013.

F. Xia, J. Chen, W. K. Fung, and H. Li. A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, 69(4):1053–1063, 2013.

T. Xu, R. T. Demmer, and G. Li. Zero-inflated Poisson factor model with application to microbiome read counts. *Biometrics*, 77(1):91–101, 2021.

C. You, J. T. Ormerod, and S. Mueller. On variational Bayes estimation and variational information criteria for linear regression models. *Australian & New Zealand Journal of Statistics*, 56(1):73–87, 2014.

Y. Zeng, D. Pang, H. Zhao, and T. Wang. A zero-inflated logistic normal multinomial model for extracting microbial compositions. *Journal of the American Statistical Association*, 118(544):2356–2369, 2023.