frontiers | Frontiers in Microbiology

TYPE Methods
PUBLISHED 30 May 2024
DOI 10.3389/fmicb.2024.1394204

OPEN ACCESS

EDITED BY
Hyun-Seob Song,
University of Nebraska-Lincoln, United States

REVIEWED BY
Mohamed R. Abonazel,
Cairo University, Egypt
Aditya Mishra,
Flatiron Institute, United States
Pierluigi Polese,
University of Udine, Italy

*CORRESPONDENCE
Jimin Ye
✉ jmye@mail.xidian.edu.cn
Ying Zhou
✉ zhouying@hlju.edu.cn

RECEIVED 01 March 2024
ACCEPTED 20 May 2024
PUBLISHED 30 May 2024

CITATION
Chi JL, Ye JM and Zhou Y (2024) A GLM-based zero-inflated generalized Poisson factor model for analyzing microbiome data. *Front. Microbiol.* 15:1394204. doi: 10.3389/fmicb.2024.1394204

A GLM-based zero-inflated generalized Poisson factor model for analyzing microbiome data

Jinling Chi[1], Jimin Ye[1]* and Ying Zhou[2]*

[1]School of Mathematics and Statistics, Xidian University, Xi'an, China, [2]School of Mathematical Sciences, Heilongjiang University, Harbin, China

Motivation: High-throughput sequencing technology facilitates the quantitative analysis of microbial communities, improving the capacity to investigate the associations between the human microbiome and diseases. Our primary motivating application is to explore the association between gut microbes and obesity. The complex characteristics of microbiome data, including high dimensionality, zero inflation, and over-dispersion, pose new statistical challenges for downstream analysis.

Results: We propose a GLM-based zero-inflated generalized Poisson factor analysis (GZIGPFA) model to analyze microbiome data with complex characteristics. The GZIGPFA model is based on a zero-inflated generalized Poisson (ZIGP) distribution for modeling microbiome count data. A link function between the generalized Poisson rate and the probability of excess zeros is established within the generalized linear model (GLM) framework. The latent parameters of the GZIGPFA model constitute a low-rank matrix comprising a low-dimensional score matrix and a loading matrix. An alternating maximum likelihood algorithm is employed to estimate the unknown parameters, and cross-validation is utilized to determine the rank of the model in this study. The proposed GZIGPFA model demonstrates superior performance and advantages through comprehensive simulation studies and real data applications.

KEYWORDS

factor analysis, GLM, microbiome data, zero inflation, ZIGP model

1 Introduction

The human microbiome is the collection of all microorganisms that live in and associate with the human body, including bacteria, archaeobacteria, protists, and viruses, distributed in the nasal cavity, oral cavity, skin, gastrointestinal tract, and genitourinary tract. The growing significance of the microbiome in ecosystems is increasingly recognized. In particular, the relationship between gut microorganisms and human health has garnered widespread scientific interest. Over time, an increasing number of studies have demonstrated that dysbiosis of the gut microbial community is associated with complex diseases, such as human gastrointestinal disorders (Willing et al., 2010; Machiels et al., 2013; Knights et al., 2014), metabolic traits, diabetes (Turnbaugh et al., 2006; Wen et al., 2008; Vijay-Kumar et al., 2010), obesity (McKnite et al., 2012; Carlisle et al., 2013; Parks et al., 2013), and inflammatory bowel disease (Frank et al., 2007). These investigations significantly contribute on exploring the causes and treatments of diseases. Furthermore, complex interactions between hosts and microbiota are also observed in various ecosystems. For example, in marine ecosystems, microbial communities associated with seaweed play an vital role in the development, reproduction, function, and defense

of seaweeds (Egan et al., 2013; Singh and Reddy, 2016). Therefore, it becomes crucial to quantify the abundance of microbial taxa and investigate the association between microbiota and diseases or traits.

The development of high-throughput sequencing (HTS) technology has been widely employed in microbial research, enabling researchers to identify the composition and abundance of microbial species directly (Kuczynski et al., 2011). Microbiome data are typically generated by extracting samples from the specified environment, followed by sequencing the 16S rRNA genes of the DNA extracts using high-throughput sequencing technology. The obtained sequence reads are compared with the reference 16S rRNA database and assigned to Operational Taxonomic Units (OTUs) based on a sequence similarity threshold (e.g., 97%; Tyler et al., 2014). High-throughput sequencing data provide valuable insights for investigating the relationship between the microbiome and the host environment or clinical factors. As a motivating application, we consider the gut microbiome data in Sun et al. (2019), which explores the association between gut microbes and obesity. The authors sequenced 16S rDNA genes of 48 individuals and obtained a dataset with 895 OTUs, where the number of variables (i.e., OTUs) vastly exceeds the number of observations (i.e., the number of samples). Moreover, we found that ∼45% of the OTU counts were zero, and the variance of the data significantly exceeded the mean. These characteristics are manifestations of high dimensionality, zero inflation, and over-dispersion, which may distort downstream analysis. However, many microbiome datasets exhibit the same problems as the motivating data, posing challenges for statistical analysis. Firstly, most microbiome data are non-negative counts with a large number of zeros (i.e., zero-inflated; Xu et al., 2015; Kaul et al., 2017). Some of these observed zeros result from insufficient sequencing depth (i.e., library size, which is the total number of reads obtained by per sample from equipment) or other technical reasons that result in some taxa not being detected, and others are the fact that some taxa are very rare and not present in most samples (Silverman et al., 2020). Traditional statistical methods may not accurately estimate the parameters of the data distribution due to the preponderance of zeros, leading to biased results (Campbell, 2021). Secondly, microbial abundance data only represent relative information in observed samples and cannot describe the abundance in the entire ecosystem (Mandal et al., 2015; Gloor et al., 2017). Moreover, the sequencing depth varies among samples, and even the variation between samples is magnitude (Sims et al., 2014). Finally, microbiome data are typically over-dispersed and high-dimensional (Kurtz et al., 2015; Xu et al., 2015; Armstrong et al., 2022). The number of taxa in the OTUs table may significantly exceed the number of observed samples, which is a sign of high dimensionality. The high dimensionality of the data may strain computational resources and increase the risk of overfitting. Meanwhile, the standard model may underestimate the true variation within the data when over-dispersion exists, leading to inaccurate estimation and hypothesis testing (Robinson et al., 2009; Love et al., 2014). Detecting associations between microbes and diseases remains challenging because of the complex features of microbiome data and the limitations of current statistical methods. Therefore, it is necessary to develop novel statistical analysis methods for the characteristics of microbiome data.

Zero-inflation and over-dispersion of count data have received widespread attention from scholars recently. Wagh and Kamalja (2017) briefly reviews different zero-inflated models for handling count data and the performance of their parameter estimation, which provides suggestions for selecting parameter estimation methods for zero-inflated models. Motivated by zero-inflation and over-dispersion problems, a zero-inflated negative binomial (ZINB) mixed regression approach is proposed to analyze the data on the length of stay for pancreas disorder (Yau et al., 2003). However, in a few cases, the parameter estimation algorithm for the ZINB regression model fails to converge (Lambert, 1992). A zero-inflated generalized Poisson (ZIGP) regression model has been proposed to model domestic violence data with too many zeros (Famoye and Singh, 2006). It is a strong competitor to the Poisson and negative binomial regression model when the count data is over-dispersed. In addition, zero-inflated generalized Poisson and zero-inflated negative binomial regression models were used in QTL mapping studies for the count traits with excess zeros (Cui and Yang, 2009; Moghimbeigi, 2015; Chi et al., 2020). More recently, Tirozzi et al. (2022) used zero-inflation models to assess long-term population trends and elucidate the effects of environmental bias, over-dispersion, and zero-inflation on the population trend estimates. These studies provide some inspiration for analyzing microbiome data with complex characteristics.

In recent years, extensive research has been conducted by scholars to address the challenges associated with microbiome data, including zero inflation, high dimensionality, and over-dispersion (Zhang et al., 2018; Xu et al., 2020; Jiang et al., 2023). Two typical methods have been proposed to address the zero-inflated structure of sequencing data. One method is replacing the zeros with small non-zero positive number (pseudo count; Chen and Li, 2013; Lin et al., 2014). However, the effects of creating pseudo count has not been evaluated thoroughly when the data contain excessive zeros. Besides, the choice of pseudo count may impact subsequent analysis (Costea et al., 2014), and this approach is not statistically rigorous. Moreover, the idea of multiplicative replacement has been proposed. The non-parametric replacement method can be used to adjust the data through multiplicative modification under simple conditions (a small number of zeros; Martín-Fernández et al., 2003). In other cases, more sophisticated model-based methods can be utilized to replace zeros in the data (Martín-Fernández et al., 2012). Recently, a Bayesian-multiplicative treatment has been proposed to solve the problem of count zero, which assumes a Dirichlet prior for the proportions and replaces the zeros with posterior Bayesian estimates (Martín-Fernández et al., 2014). The other standard and widely used method is to construct a two-part model with a point probability mass at zero along with another parametric distribution, such as zero-inflated Gaussian model (Xu et al., 2015), zero-inflated lognormal model (Sohn et al., 2015), zero-inflated Poisson model (Xu et al., 2020), zero-inflated negative binomial model (Jiang et al., 2019; Zhang and Yi, 2020), and many others (Peng et al., 2016; Tang and Chen, 2018; Zeng et al., 2022; Jiang et al., 2023). The advantage of this method is that an appropriate model can be selected according to the nature of the data. For example, the zero-inflated negative binomial model can effectively address the issue of zero-inflated and over-dispersion in the data because

the negative binomial provides a standard statistical model for over-dispersed data.

The other well-known challenge for analyzing microbial data is the high dimensionality of the data. Generally, the number of taxa usually far exceeds the observed samples in the OTUs table, which is a symbol of high dimensionality (Armstrong et al., 2022). Therefore, dimensionality reduction technology is used to map high-dimensional data into a potential low-dimensional space while retaining the primary information in the data intact to facilitate the subsequent analysis, which is a desirable preprocessing step (Fan et al., 2015; Jasner et al., 2021).

Factor analysis, an extensively employed technique, serves as a prominent method for dimensionality reduction of high-dimensional data. Pierson and Yau (2015) proposed a zero-inflated factor analysis (ZIFA) model to explicitly consider excess zeros in Single-cell RNA-seq data. However, the ZIFA model preprocesses count data via a normal transformation, which may overlook its inherent count nature and potentially result in information loss during the preprocessing step. Lee et al. (2013) developed a Poisson factor model with offsets to explicitly incorporate the special features that count nature and heterogeneous library size (the total reads per sample). Subsequently, the negative binomial factor regression model was proposed to reduce the dimensionality of microbial abundance data, and then model the associations of microbial abundance and host-associated features by including only a subset of the predictors for a few latent factors (Mishra and Müller, 2022). However, these two methods (Poisson factor model and negative binomial factor model) are only suitable for data that does not contain excessive zeros, which fail to consider zero inflation. Sohn and Li (2017) proposed a GLM-based ordination method for microbiome samples (GOMMS), which employs a zero-inflated quasi-Poisson factor model to dimensionality reduction and overcome the challenge of zero-inflation. However, this method assumes that each taxa has a fixed probability of zero, which is generally not easily satisfied. More recently, Xu et al. (2020) proposed a factor analysis model based on the zero-inflated Poisson distribution (ZIPFA), which can more flexibly adapt to some characteristics of microbial data, such as count value, excessive zeros, and high dimensionality. A significant critique of Poisson models is the failure to accommodate over-dispersion, which has been widely observed for microbiome data. Following this line of research, we combined the zero-inflated generalized Poisson distribution with factor analysis under the framework of the generalized linear model to propose a GLM-based zero-inflated generalized Poisson factor analysis (GZIGPFA) model, which provides a valuable dimensionality reduction tool for microbiome data. The GZIGPFA model can also address the issues of over-dispersion and handle the zero-inflated structure. Furthermore, our method models absolute abundance directly, avoiding the information loss attributable to data transformation.

The rest of this paper is organized below. Section 2 presents a new GZIGPFA model for handling microbiome data and introduces methods for parameter estimation and rank selection. A simulation and comparison study are conducted in Section 3 to demonstrate the performance of the proposed method. In Section 4, we apply our method to the gut microbial data to explore the association between gut microbes and obesity. In Section 5, a conclusion of this paper is drawn with a discussion of extensions and areas for subsequent work.

2 Method

For $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$, let y_{ij} denote the count of the j-th taxon from the i-th individual, then, an $n \times m$ microbial abundance matrix can be expressed as $\boldsymbol{Y} = (y_{ij})_{n\times m}$. Denote the ith row of matrix $\boldsymbol{Y}$ as $\boldsymbol{y}_{(i)} = (y_{i1}, \ldots, y_{im})$, refer to as the i-th sample of sequencing data.

2.1 Zero-inflated generalized Poisson factor model

The microbiome dataset typically presents as a highly skewed non-negative count matrix with numerous zeros, often characterized by over-dispersion. Therefore, we build statistical models to address these issues for microbiome data.

The presence of zeros in microbiome data may be true absences or undetected taxa. Considering the over-dispersion characteristics of the data, we assume that the sequencing count y_{ij} follows the zero-inflated generalized Poisson (ZIGP) distribution (Famoye and Singh, 2006):

$$y_{ij} \sim \begin{cases} 0, & \text{with probability } \phi_{ij}, \\ GP(T_i\lambda_{ij}, \alpha), & \text{with probability } 1-\phi_{ij}, \end{cases} \tag{1}$$

where ϕ_{ij} is the zero-inflation parameter describing the probability of excess zero; $GP(T_i\lambda_{ij}, \alpha)$ is the generalized Poisson distribution (Consul and Famoye, 1992; Famoye, 1993), with the probability function

$$p(y_{ij};\, T_i\lambda_{ij}, \alpha) = \frac{1}{y_{ij}!}\left(\frac{T_i\lambda_{ij}}{1+\alpha T_i\lambda_{ij}}\right)^{y_{ij}} (1+\alpha y_{ij})^{y_{ij}-1} \exp \left\{-\frac{T_i\lambda_{ij}(1+\alpha y_{ij})}{1+\alpha T_i\lambda_{ij}}\right\}, \tag{2}$$

where λ_{ij} and α are the mean and dispersion parameters of the generalized Poisson part, respectively; T_i is the relative library size of the i-th sample, which is utilized to regulate λ_{ij}. Generally, there are many representations of T_i (Anders and Huber, 2010; Eddy, 2011; Badri et al., 2020; Mishra and Müller, 2022). In this paper, we take

$$T_i = \sum_{j=1}^{m} y_{ij} \Big/ \text{median}\left(\sum_{j=1}^{m} y_{1j}, \ldots, \sum_{j=1}^{m} y_{nj}\right).$$

Next, the link between the zero-inflation probability ϕ_{ij} and the mean parameter λ_{ij} is established according to Lambert (1992). Typically, an increase in the number of zeros in the data results in a smaller overall mean. Therefore, a negative relationship between ϕ_{ij} and λ_{ij} is established, i.e.,

$$\text{logit}(\phi_{ij}) = -\tau \log(\lambda_{ij}), \tag{3}$$

where τ is the shape parameter; $\text{logit}(\phi_{ij})$ and $\log(\lambda_{ij})$ are the link functions for the probability of zero-inflation and the mean of generalized Poisson in the generalized linear model (GLM), respectively. Let $\boldsymbol{\Lambda} = (\lambda_{ij})_{n\times m} \in \mathbb{R}^{n\times m}$ and $\boldsymbol{\Phi} = (\phi_{ij})_{n\times m} \in \mathbb{R}^{n\times m}$ be the matrix forms of λ_{ij} and ϕ_{ij}, respectively. Therefore, the matrix form of Equation (3) can be expressed as $\text{logit}(\boldsymbol{\Phi}) = -\tau \log(\boldsymbol{\Lambda})$.

The ZIGP model (Equation 1) described above can accommodate simultaneously zero inflation and over-dispersion count data. Furthermore, upon review of the existing literature about microbiome data analysis, the ZIGP model represents the inaugural utilization of the zero-inflated generalized Poisson model in microbiome datasets. In the following, we intend to solve the prevalent issue of high dimensionality in the microbiome data with a factor analysis model. Therefore, we propose a GLM-based zero-inflated generalized Poisson factor analysis (GZIGPFA) model to provide a suitable model for zero-inflated, over-dispersed, and high-dimensional microbiome data.

Assume that matrix $\log(\boldsymbol{\Lambda})$ has a low-rank structure $\log(\boldsymbol{\Lambda}) = \boldsymbol{F}\boldsymbol{L}^T$ with rank K (Lee et al., 2013), where $\boldsymbol{F} \in \mathbb{R}^{n\times K}$ is the factor score matrix and $\boldsymbol{F} = (\boldsymbol{f}_{(1)}^T, \ldots, \boldsymbol{f}_{(n)}^T)^T$ with $\boldsymbol{f}_{(i)} = (f_{i1}, \ldots, f_{iK})$, $i = 1, 2, \ldots, n$; $\boldsymbol{L} \in \mathbb{R}^{m\times K}$ is the loading matrix and $\boldsymbol{L} = (\boldsymbol{l}_{(1)}^T, \ldots, \boldsymbol{l}_{(m)}^T)^T$ with $\boldsymbol{l}_{(j)} = (l_{j1}, l_{j2}, \ldots, l_{jK})$, $j = 1, 2, \ldots, m$. Then, we consider the following zero-inflated generalized Poisson factor model:

$$\begin{cases} y_{ij} \sim \text{ZIGP}(T_i\lambda_{ij}, \alpha, \phi_{ij}), \\ \text{logit}(\phi_{ij}) = -\tau \log(\lambda_{ij}), \\ \log(\lambda_{ij}) = f_{i1}l_{j1} + f_{i2}l_{j2} + \ldots + f_{iK}l_{jK}, \end{cases} \tag{4}$$

where f_{ik} is an element of the matrix $\boldsymbol{F}$, denoting the kth factor score for the i-th sample; l_{jk} is an element of matrix $\boldsymbol{L}$, denoting the loading of the jth taxon on the kth factor, where $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$, and $k = 1, 2, \ldots, K$. In this model, the logarithm is the canonical link function in the generalized linear model (GLM) framework (McCullagh and Nelder, 1989).

After the rank K is determined and the unknown parameters $\alpha, \tau, \boldsymbol{F}, \boldsymbol{L}$ in the model (Equation 4) are estimated, we reduce the dimensionality of the microbiome dataset from m to K. The score matrix $\boldsymbol{F}$ possesses an equivalent sample size to the original microbiome dataset $\boldsymbol{Y}$ but only has K variables. In subsequent work, it is easier to perform association analysis between disease phenotypes and the low-dimensional score matrix, providing a brief tool for investigating the relationship between microbiome and disease.

2.2 An alternating maximum likelihood algorithm

To estimate the unknown parameters $\alpha, \tau, \boldsymbol{F}, \boldsymbol{L}$ in model (Equation 4), we adopt a method that maximizes the ZIGP likelihood function:

$$L(\alpha, \tau, \boldsymbol{F}, \boldsymbol{L}) = \prod_{i=1}^{n}\prod_{j=1}^{m}\left[\phi_{ij}I_{\{y_{ij}=0\}} + (1-\phi_{ij})p(y_{ij}; T_i\lambda_{ij}, \alpha)\right], \tag{5}$$

where $p(y_{ij}; T_i\lambda_{ij}, \alpha)$ is the probability function of GP distribution (Equation 2); $\log(\lambda_{ij}) = \sum_{k=1}^{K} f_{ik}l_{jk}$ and $\text{logit}(\phi_{ij}) = -\tau \log(\lambda_{ij})$. In Equation (5), $\alpha, \tau, \boldsymbol{F}, \boldsymbol{L}$ are the unknown parameters, and it is challenging to maximize the likelihood directly. Therefore, we consider an alternating maximum likelihood algorithm in the GLM framework to estimate the parameters.

In order to obtain the initial $\boldsymbol{F}$ and $\boldsymbol{L}$, we apply the singular value decomposition (SVD) to the log-transformed matrix $\widetilde{Y}$ and obtain the singular, i.e., $\log(\widetilde{Y}) = U\Sigma V^T$. Set $\boldsymbol{L}^{old} = V^T$ and $\boldsymbol{F}^{old} = (U_{(,1)}\Sigma_{11}, U_{(,2)}\Sigma_{22}, \ldots, U_{(,K)}\Sigma_{KK})$, where $\Sigma_{kk}, k = 1, \ldots, K$ is the kth diagonal element of Σ.

Step 1: Assuming that factor score matrix $\boldsymbol{F}$ is known as $\boldsymbol{F}^{old}$, a ZIGP regression model is fitted with the jth column of matrix $\boldsymbol{Y}$ (denoted by $\boldsymbol{y}_j$) as the response and $\boldsymbol{F}^{old}$ as a covariate matrix, the regression model can be written as

$$\begin{cases} \boldsymbol{y}_j \sim \begin{cases} 0, & \text{with probability } \boldsymbol{\phi}_j, \\ GP(\boldsymbol{T}\boldsymbol{\lambda}_j, \alpha), & \text{with probability } 1-\boldsymbol{\phi}_j, \end{cases} \\ \log(\boldsymbol{\lambda}_j) = \boldsymbol{F}^{old}\boldsymbol{l}_{(j)}^T, \\ \text{logit}(\boldsymbol{\phi}_j) = -\tau \log(\boldsymbol{\lambda}_j), \end{cases}$$

where the vector $\boldsymbol{T} = (T_1, T_2, \ldots, T_n)$ is the relative library size vector; the regression coefficient vector $\boldsymbol{l}_{(j)} = (l_{j1}, l_{j2}, \ldots, l_{jK})$ is the jth row of the factor loading matrix $\boldsymbol{L}^{new} = (\boldsymbol{l}_{(1)}^T, \boldsymbol{l}_{(2)}^T, \ldots, \boldsymbol{l}_{(m)}^T)^T$; the vectors $\boldsymbol{\lambda}_j$ and $\boldsymbol{\phi}_j$ are the jth column of the matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Phi}$, respectively.

To estimate the unknown parameter vector $\boldsymbol{\theta} = (\tau, \alpha, \boldsymbol{l}_{(j)})^T$ of the regression model, we should maximize the likelihood function. However, the explicit solution of each parameter cannot be obtained by directly using the maximum likelihood estimation method. Therefore, we perform parameter estimation of the regression model with the EM algorithm. The detailed procedure of the EM algorithm is given in Appendix A.

Since matrix $\boldsymbol{Y}$ has m columns, we need to fit m GLMs to obtain m rows of $\boldsymbol{L}^{new}$. However, the proposed model assumes that the τ and α remain the same across all m different GLMs. To accommodate this, we combine $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ into a column vector and solve all m models simultaneously to obtain the globally optimal τ and α values.

After estimating τ, α and $\boldsymbol{L}$, we continue to update $\boldsymbol{F}$. The process is similar to Step 1.

Step 2: Fit a ZIGP regression model with the ith row of matrix $\boldsymbol{Y}$ (denoted by $\boldsymbol{y}_{(i)}$) as the response and the estimated loading matrix $\boldsymbol{L} = \boldsymbol{L}^{new}$ from the previous step as a covariate, the regression model can be written as

$$\begin{cases} \boldsymbol{y}_{(i)} \sim \begin{cases} 0, & \text{with probability } \boldsymbol{\phi}_{(i)}, \\ GP(T_i\boldsymbol{\lambda}_{(i)}, \alpha), & \text{with probability } 1-\boldsymbol{\phi}_{(i)}, \end{cases} \\ \log(\boldsymbol{\lambda}_{(i)}) = \boldsymbol{f}_{(i)}\boldsymbol{L}^{newT}, \\ \text{logit}(\boldsymbol{\phi}_{(i)}) = -\tau \log(\boldsymbol{\lambda}_{(i)}), \end{cases}$$

where T_i is the relative library size of the ith sample; the regression coefficient vector $\boldsymbol{f}_{(i)} = (f_{i1}, f_{i2}, \ldots, f_{iK})$ is the ith row of the factor score matrix $\boldsymbol{F}^{new} = (\boldsymbol{f}_{(1)}^T, \boldsymbol{f}_{(2)}^T, \ldots, \boldsymbol{f}_{(n)}^T)^T$; the vectors $\boldsymbol{\lambda}_{(i)}$ and $\boldsymbol{\phi}_{(i)}$

are the ith row of the matrices $\mathbf{\Lambda}$ and $\mathbf{\Phi}$, respectively. Next, the parameters α, τ, and $\boldsymbol{f}_{(i)}$ in the regression model are estimated by the EM algorithm. The specific process is similar to Step 1.

Since matrix $\boldsymbol{Y}$ has n rows, we need to fit n GLMs to obtain n rows of $\boldsymbol{F}^{new}$. However, the proposed model assumes that the τ and α remain the same across all n different GLMs. Therefore, similar to step 1, we combine $\boldsymbol{y}_{(1)}, \ldots, \boldsymbol{y}_{(n)}$ into a column vector and solve all n models simultaneously to obtain the globally optimal τ and α values.

Step 3: Apply the singular value decomposition (SVD) method to the $\boldsymbol{F}^{new}\boldsymbol{L}^{newT}$ to obtain a new $\boldsymbol{F}^{old}$, and repeat the above alternating algorithm until convergence.

When the percentage of total likelihood difference between two iterations is less than a certain small value, the algorithm terminates; otherwise, we continue to update $\boldsymbol{F}, \boldsymbol{L}, \tau$ and α until convergence. In the ZIGP regression step, we will use the EM algorithm to estimate the parameters. Therefore, the likelihood increases due to the nature of the EM algorithm used in regression estimation (Dempster et al., 1977; Wu, 1983). The likelihood remains the same in the SVD step. Overall, the algorithm is guaranteed to converge. In the Step 3, we apply SVD to $\boldsymbol{F}^{new}\boldsymbol{L}^{newT}$, which ensures the uniqueness and the orthogonality of the updated components. We briefly summarize the alternating maximum likelihood algorithm under the GLM framework in the "Algorithm 1" box.

Initialize:

- Replace all zeros in matrix $\boldsymbol{Y}$ with column means, and the replaced matrix is represented as $\widetilde{\boldsymbol{Y}}$.
- Apply the SVD to the $\log(\widetilde{\boldsymbol{Y}})$ to obtain the singular vectors, i.e., $\log(\widetilde{\boldsymbol{Y}}) = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$; Set $\boldsymbol{F}^{old} = (U_{(,1)}\Sigma_{11}, U_{(,2)}\Sigma_{22}, \ldots, U_{(,K)}\Sigma_{KK})$.

Update:

(1) Fit m ZIGP regression models with $\boldsymbol{y}_j, j = 1, 2, \ldots, m$ as the response and the score matrix $\boldsymbol{F}^{old}$ as the covariates to obtain the estimates for $\boldsymbol{l}_j$ through the EM algorithm. Denote that loading matrix is $\boldsymbol{L}^{new} = (\boldsymbol{l}_{(1)}^T, \boldsymbol{l}_{(2)}^T, \ldots, \boldsymbol{l}_{(m)}^T)^T$ with $\boldsymbol{l}_{(j)} = (l_{j1}, l_{j2}, \ldots, l_{jK})$.

(2) Fit n ZIGP regression models with $\boldsymbol{y}_{(i)}, i = 1, 2, \ldots, n$ as the response and the loading matrix $\boldsymbol{L}^{new}$ as the covariates to obtain the estimates for $\boldsymbol{f}_{(i)}$ through the EM algorithm. Denote that the factor score matrix is $\boldsymbol{F}^{new} = (\boldsymbol{f}_{(1)}^T, \boldsymbol{f}_{(2)}^T, \ldots, \boldsymbol{f}_{(n)}^T)^T$ with $\boldsymbol{f}_{(i)} = (f_{i1}, f_{i2}, \ldots, f_{iK})$.

(3) Apply SVD method to $\boldsymbol{F}^{new}\boldsymbol{L}^{newT}$ to obtain a new $\boldsymbol{F}^{old}$

(4) Repeat from step 1 to step 3 until convergence.

Algorithm 1. GZIGPFA algorithm.

2.3 Rank estimation

We use the N-fold cross-validation suggested by Li et al. (2018) to determine the optimal number of factors, i.e., the rank K of model (Equation 4). The idea is to randomly divide the entries of a data matrix into N non-overlapping parts. We systematically exclude one block at a time and utilize the remaining data to estimate the unknown parameters with varying ranks. Subsequently, we compute the likelihood of the model using the data of the excluded block. Finally, we sum up the likelihood of all N folds to obtain the total cross-validation (CV) likelihood of the model with rank k and calculate the CV likelihood for every rank k. The rank that provides the maximum CV likelihood is chosen as the optimal rank. The procedure of rank selection is briefly summarized in the "Algorithm 2" box.

Set the candidate rank set $\mathbb{K} = \{1, \ldots, m\}$;
Randomly split $\boldsymbol{Y}$ into N folds, with indicates contained in $I^{[1]}, I^{[2]}, \ldots, I^{[N]}$;
for $k \in \mathbb{K}$ **do**
  for $t = 1, \ldots, N$ **do**
    • Eliminate the elements with index $I^{[t]}$ in $\boldsymbol{Y}$ and estimate the unknown parameter with rank k (i.e., $\boldsymbol{\theta}_k^{[-t]}$) using the GZIGPFA algorithm;
    • Calculate the likelihood of the model with rank k in the t-th fold using the elements with index $I^{[t]}$ in $\boldsymbol{Y}$;
  end for
  Sum up the likelihood of all N folds to obtain the CV likelihood of the model with rank k;
end for
Calculate the CV likelihood of every rank across N folds;
Choose the rank that provides the maximum CV likelihood as the optimal rank.

Algorithm 2. N-fold cross-validation for rank estimation.

3 Simulation studies

The performance of our proposed GZIGPFA method is demonstrated through a simulation study. We compare the GZIGPFA method with four other methods:

- ZIPFA (Zero-inflated Poisson factor analysis): This method uses a zero-inflated Poisson factor analysis model for reducing the dimension of the microbiome data while accommodating the zero-inflated nature of the data (Xu et al., 2020).
- log-PCA (log-principal component analysis): The data is preprocessed by replacing all zeros with a small value and then taking a logarithm of the transformed data. After that, the data is processed by performing a principal component analysis (PCA).
- PSVDOS (Poisson Singular Value Decomposition with Offset): The method is an efficient algorithm for estimating

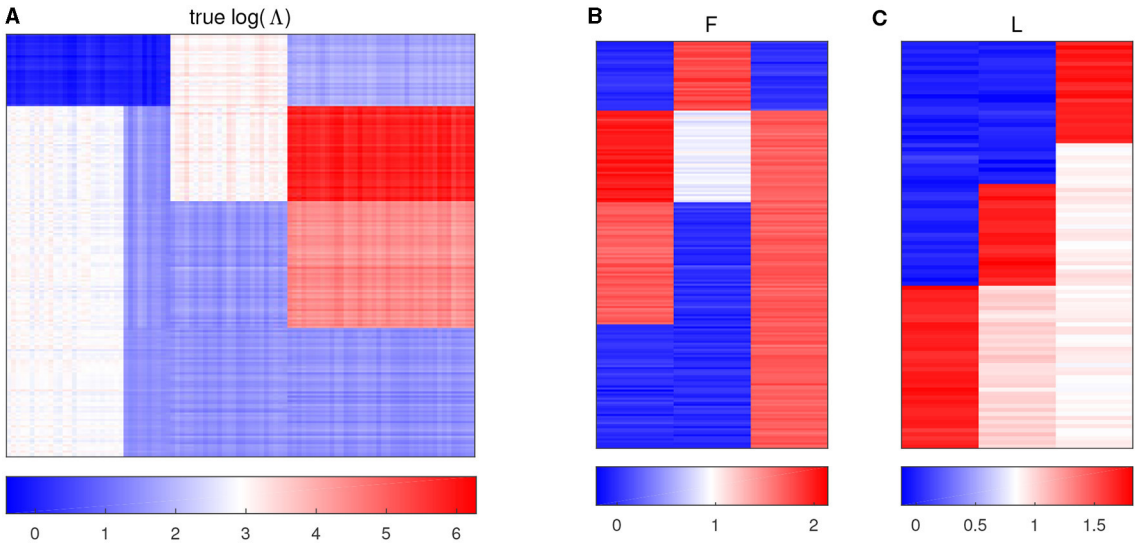

FIGURE 1
Plots of simulation parameters. **(A)** The heatmap of true log($\boldsymbol{\Lambda}$) matrix. The rows represent samples, and the columns represent taxa. **(B)** True left singular vector matrix $\boldsymbol{F}$. The rows correspond to the samples, and the columns denote factors. **(C)** True right singular vector matrix $\boldsymbol{L}$. The rows correspond to the taxa, and the columns denote factors.

the Poisson factor model, which addresses the issue of sample normalization through the use of unknown offset parameters (Lee et al., 2013).

- GOMMS (GLM-based ordination method for microbiome samples): This method uses a zero-inflated quasi-Poisson factor model, which accounts for characteristics of microbiome data (e.g., highly skewed non-negative counts with excessive zeros) while reducing dimensionality (Sohn and Li, 2017).

3.1 Simulation design

We followed the design of Xu et al. (2020) to simulate microbiome datasets. A sequence data of n samples and m taxa is generated according to model (Equation 4). We simulate $n = 200$ different samples measured on $m = 100$ taxa. The rate matrix $\boldsymbol{\Lambda}$ follows: $\log(\boldsymbol{\Lambda}) = \boldsymbol{F}\boldsymbol{L}^T$, where the $\boldsymbol{F} \in \mathbb{R}^{n\times 3}$ is a left singular vector matrix, and $\boldsymbol{L} \in \mathbb{R}^{m\times 3}$ is a right singular vector matrix. To generate matrix $\boldsymbol{F}$, we create a 200-by-3 matrix $\boldsymbol{F}$ such that:

Column 1: $F(36:80, 1) = 2$, $F(81:140, 1) = 1.7$,
Column 2: $F(1:35, 2) = 1.8$, $F(36:80, 2) = 0.9$,
Column 3: $F(1:35, 3) = 1.7$, $F(36:200, 2) = 0$,

with all the other entries being 0, and then jitter all the entries by adding random noises generated from $N(0, 0.06^2)$. Similarly, to generate matrix $\boldsymbol{L}$, we create a 100-by-3 matrix $\boldsymbol{L}$ such that:

Column 1: $L(1:60, 1) = 0$, $L(61:100, 1) = 1.7$,
Column 2: $L(36:60, 2) = 1.7$, $L(61:100, 2) = 1$,
Column 3: $L(1:25, 3) = 1.7$, $L(26:100, 2) = 0.9$,

with all the other entries being 0, and then jitter all the entries by adding random noises generated from $N(0, 0.05^2)$. Figure 1A displays the heatmap of the true log($\boldsymbol{\Lambda}$) matrix, and the three columns of $\boldsymbol{F}$ and $\boldsymbol{L}$ are shown in the columns of Figures 1B, C, respectively. Each row in Figure 1B corresponds to one sample, and Figure 1C shows the heatmap of the right singular vector matrix $\boldsymbol{L}$, in which each row indicates one taxon profile.

After the matrices $\boldsymbol{F}$ and $\boldsymbol{L}$ are generated, $\boldsymbol{\Lambda}$ can be obtained according to $\boldsymbol{\Lambda} = \exp(\boldsymbol{F}\boldsymbol{L}^T)$. Next, a zero-inflated sequencing matrix $\boldsymbol{Y}$ was generated from the following ZIGP model,

$$f(y_{ij}; \alpha, \lambda_{ij}, \phi_{ij}, T_i) = \phi_{ij} I_{\{y_{ij}=0\}} + (1-\phi_{ij}) GP(y_{ij}; T_i\lambda_{ij}, \alpha),$$

where λ_{ij} is an element of the matrix $\boldsymbol{\Lambda}$; the scaling parameter T_i and the dispersion parameter α were set to 1 and 0.2, respectively; the probability of excess zero ϕ_{ij} is obtained by establishing the link between ϕ_{ij} and λ_{ij}. Firstly, we consider the scenario with the relationship between ϕ_{ij} and λ_{ij} established in Section 2, that is,

- *Scenario* 1: $\text{logit}(\phi_{ij}) = -\tau \log(\lambda_{ij})$.

Furthermore, to more comprehensively evaluate the robustness of the proposed method, we further considered generating sequencing data under several misspecified scenarios. First, consider two common links to ϕ_{ij} and λ_{ij}, which are mentioned in Lambert (1992) besides Scenario 1:

- *Scenario* 2: $\log\{-\log(\phi_{ij})\} = \tau \log(\lambda_{ij})$.
- *Scenario* 3: $\log\{-\log(1-\phi_{ij})\} = \tau \log(\lambda_{ij})$.

In addition, we set up a scenario along the lines in Sohn and Li (2017), namely, each taxon has a fixed probability ϕ_j independent of the λ_{ij}:

- *Scenario* 4: $\phi_j \sim \text{Uniform}(\tau - 0.10, \tau + 0.10)$.

Finally, considering that actual microbiome data may come from different distributions, we set up two misspecified scenarios

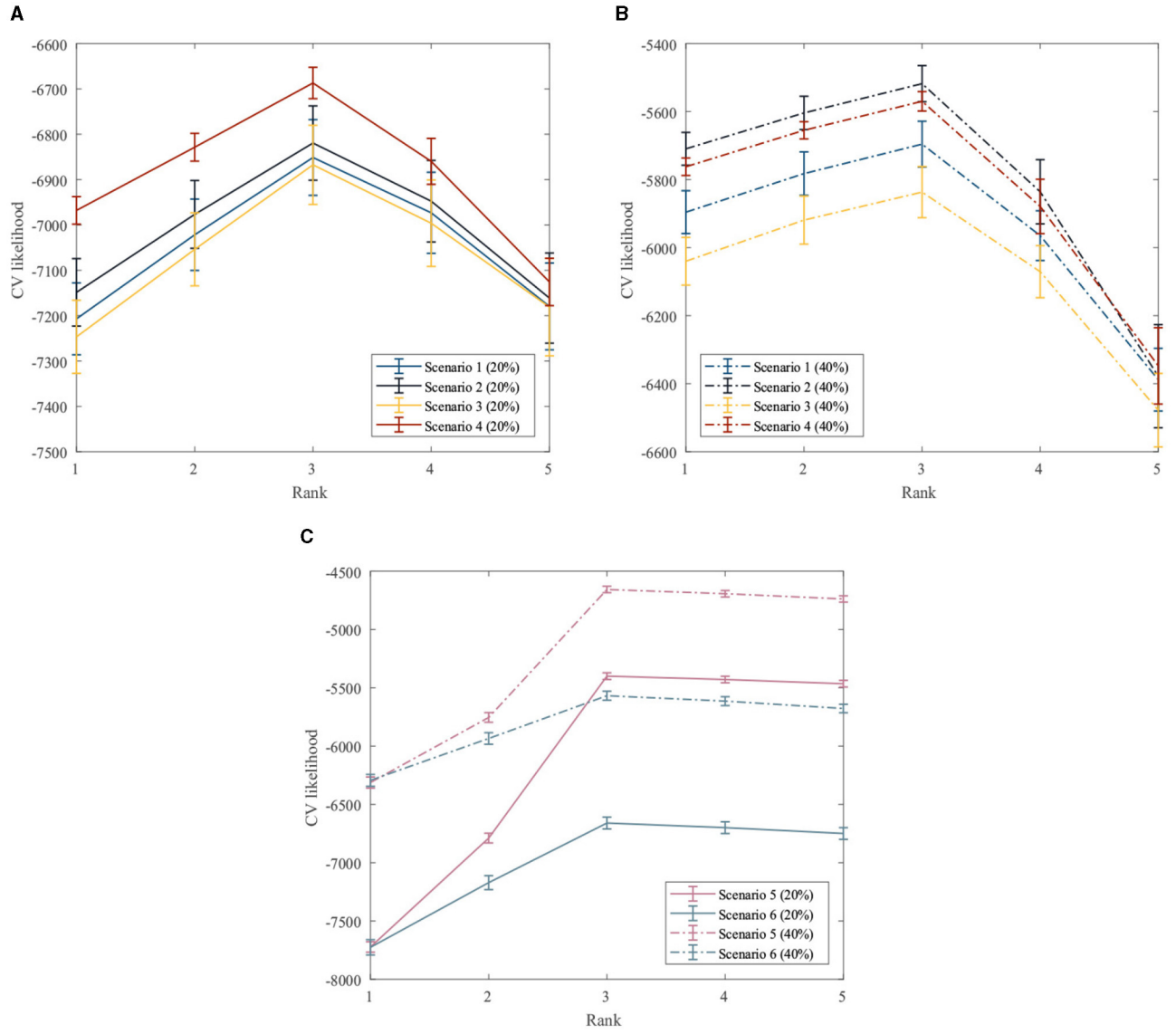

FIGURE 2
Cross-validation to choose the rank in the GZIGPFA model. **(A)** The CV likelihood of the GZIGPFA model under scenarios 14 with a 20% zero-inflated probability. **(B)** The CV likelihood of the GZIGPFA model under scenarios 14 with a 40% zero-inflated proportion. **(C)** The CV likelihood of the GZIGPFA model under scenarios 5 and 6 with 20% and 40% zero-inflated proportions. The proposed method provides maximum CV likelihoods with rank 3 under all simulation scenarios.

for generating microbiome data from other distributions, that is, the data come from the ZIP and ZINB distributions, and the relationship between the ϕ_{ij} and λ_{ij} follows the setup in Scenario 1:

- *Scenario* 5: $y_{ij} \sim$ ZIP distribution, and $\text{logit}(\phi_{ij}) = -\tau \log(\lambda_{ij})$.
- *Scenario* 6: $y_{ij} \sim$ ZINB distribution, and $\text{logit}(\phi_{ij}) = -\tau \log(\lambda_{ij})$.

We evaluated the simulation results for all the scenarios above at light zero inflation (20%) and higher zero inflation (40%), respectively.

3.2 Simulation results

First, the performance of the proposed method for rank estimation in all scenarios is examined. A 10-fold cross-validation is performed on the data generated in each scenario separately to compute the cross-validation likelihood for different ranks. The rank estimation results of the GZIGPFA method for all simulation scenarios are displayed in Figure 2. It can be seen from Figure 2 that the proposed method provides the maximum CV likelihood with rank 3 in all simulation scenarios. Figure 2 shows that the proposed method is accurate in the rank estimation under the given model [Model (Equation 4)] and performs well under the misspecified scenarios, indicating the robustness of the GZIGPFA method.

TABLE 1 The mean of loss values and standard errors (in the parenthesis) for the five methods under different scenarios.

Scenario	Zero (%)	GZIGPFA	ZIPFA	log-PCA	PSVDOS	GOMMS
Scenario (1)	20%	**6.9878** (1.4968)	10.7800 (1.8965)	10.2970 (0.1332)	26.4999 (0.0594)	92.5108 (207.740)
	40%	**12.7542** (7.1569)	13.2471 (2.9255)	16.7528 (0.1516)	26.6175 (0.0843)	102.689 (238.263)
Scenario (2)	20%	**8.9129** (3.2038)	10.8945 (2.0912)	10.7148 (0.1366)	26.5064 (0.0581)	87.5358 (166.145)
	40%	**10.2343** (4.8788)	13.5586 (3.3809)	17.7686 (0.1339)	26.7447 (0.1101)	91.5380 (144.401)
Scenario (3)	20%	**7.1095** (1.2251)	10.8162 (1.8302)	10.0293 (0.1294)	26.5015 (0.0644)	99.1790 (183.318)
	40%	21.5684 (8.9917)	**12.1517** (2.5438)	15.8540 (0.1635)	26.5817 (0.0819)	114.108 (243.236)
Scenario (4)	20%	**10.2995** (0.9843)	11.5437 (2.1585)	12.0547 (0.2186)	26.5496 (0.0694)	78.4083 (114.947)
	40%	**9.8594** (4.7380)	13.0524 (3.0187)	17.5302 (0.2297)	26.7168 (0.0957)	100.821 (241.874)
Scenario (5)	20%	**1.9069** (0.0882)	2.0074 (0.0686)	5.7243 (0.1271)	26.2521 (0.0052)	4.7897 (0.0036)
	40%	3.9383 (0.2266)	**3.2860** (0.2841)	13.0891 (0.1697)	27.2165 (0.2778)	10.5189 (0.0140)
Scenario (6)	20%	**2.9830** (0.1168)	3.3560 (0.2495)	6.4711 (0.1243)	26.2505 (0.0030)	6.0003 (0.0013)
	40%	**4.4550** (0.3200)	4.5013 (0.5164)	13.7113 (0.1683)	26.6065 (0.3625)	10.0941 (0.0107)

The best results in each setting are in boldface.

Next, a comprehensive comparison of the GZIGPFA method with other approaches (ZIPFA, log-PCA, PSVDOS, and GOMMS) is presented to illustrate the superior performance of the proposed method in depth. For each simulation scenario, 200 replicates are performed. The Frobenius norm of the error matrix (denoted as loss value) is utilized to evaluate the effectiveness of each method. The loss values of several methods in all simulation scenarios are listed in Table 1. Table 1 shows that the GZIGPFA method has a small loss in most simulation scenarios, indicating that the proposed method is effective. In Scenarios (1)–(4), the performance of all four methods is significantly better than the GOMMS method. In addition, we find that the convergence effect of the GOMMS method is poor when there are more zeros in the data. In scenario (5), ZIPFA outperforms GZIGPFA when the zero percentage is high (40%) because this scenario essentially favors ZIPFA by using the ZIP model. In Scenario (6), the data is generated by the ZINB model, and the PSVDOS method performs the worst among the five methods because this method cannot consider over-dispersed and zero-inflated data. In addition, the GOMMS method performs second only to GZIGPFA and ZIPFA in Scenario (6) because GOMMS is based on the zero-inflated quasi-Poisson, which is intrinsically closer to ZINB. Overall, our method performs better than competing methods, even in the misspecified scenarios.

Finally, we show the heatmaps of the true $\log(\boldsymbol{\Lambda})$ and the estimated $\log(\boldsymbol{\Lambda})$ of several methods in Figure 3 to visualize the performance of the GZIGPFA method, and we also display the clustering effects of several methods at the taxa (top) and sample (left side) levels. Since the GOMMS method performs poorly in Table 1, only the estimation and clustering results of the four methods GZIGPFA, ZIPFA, log-PCA, and PSVDOS are presented in Figure 3. Panel (a) in Figures 3A, B displays the true $\log(\boldsymbol{\Lambda})$ used in the simulation. The phylogenetic tree on the left side of the heatmap shows the clustering of the sample, which falls into four clusters. Similarly, the phylogenetic tree above the heatmap shows the clustering of the taxa. The clustering pattern is obtained by applying the complete linkage hierarchical clustering analysis to $\boldsymbol{F}$ and $\boldsymbol{L}$ (Wilkinson and Friendly, 2009). Figures 3A, B show the estimation and clustering effects of the four methods when the zero-inflated proportion is low (20%) and high (40%) in Scenario (1), respectively. GZIGPFA method (Panel b in Figures 3A, B) offers the best approximation to the true signal, and it gives the accurate clustering result, which is as expected, as the dataset was designed in a way that takes advantage of the unique features of GZIGPFA. The $\log(\boldsymbol{\Lambda})$ estimated by the log-PCA method (Panel d in Figures 3A, B) is far from the true value (Panel a) and is the worst performer among several methods. Meanwhile, log-PCA fails to capture the right sample clustering when the zero-inflated proportion goes from 20 to 40%. The log-PCA performs poorly overall because it does not consider the underlying distribution and excessive zeros.

4 Application to the gut microbiome data

Empirical research has demonstrated that mice and humans harbor similar microbiota at high taxonomic levels (Ley et al., 2008; Krych et al., 2013). Therefore, laboratory mice can be used to simulate the human gut environment for experiments and to explore the mechanisms of host-microbial interactions in a data-driven manner when studying human gut microbes. In this section, the GZIGPFA model is applied to the mouse gut microbial dataset (Sun et al., 2019) to explore the association between gut microbes and obesity. Microbial datasets were extracted from 48 male mice. The mice were divided into the blank group, high-fat control group, and probiotic experimental group, in which the blank group was fed normal chow, the high-fat control group and the probiotic experimental group were fed a high-fat diet for 4 weeks to establish an obesity model for the mice, and the probiotic experimental group was fed high-fat chow plus probiotic capsules starting from the 5th week of the successful modeling, while the high-fat control group continued to be fed high-fat chow. At the end of the 8th week, various indicators

10.3389/fmicb.2024.1394204

of the mice were measured, including weight, body length, total cholesterol, endotoxin, etc. The samples were first amplified with a set of primers targeting the 16S rDNA V4 region. Then, the original data were subjected to operational taxonomic unit (OTU) clustering and species classification analysis based on valid data. According to the results of OTU clustering, species annotation was performed for the representative sequences of each OTU, and the corresponding species information and species-based abundance were obtained. Then, we reduced the dimensionality of the dataset with the proposed GZIGPFA method to extract the common factors and further explore the association between the common factors and obesity.

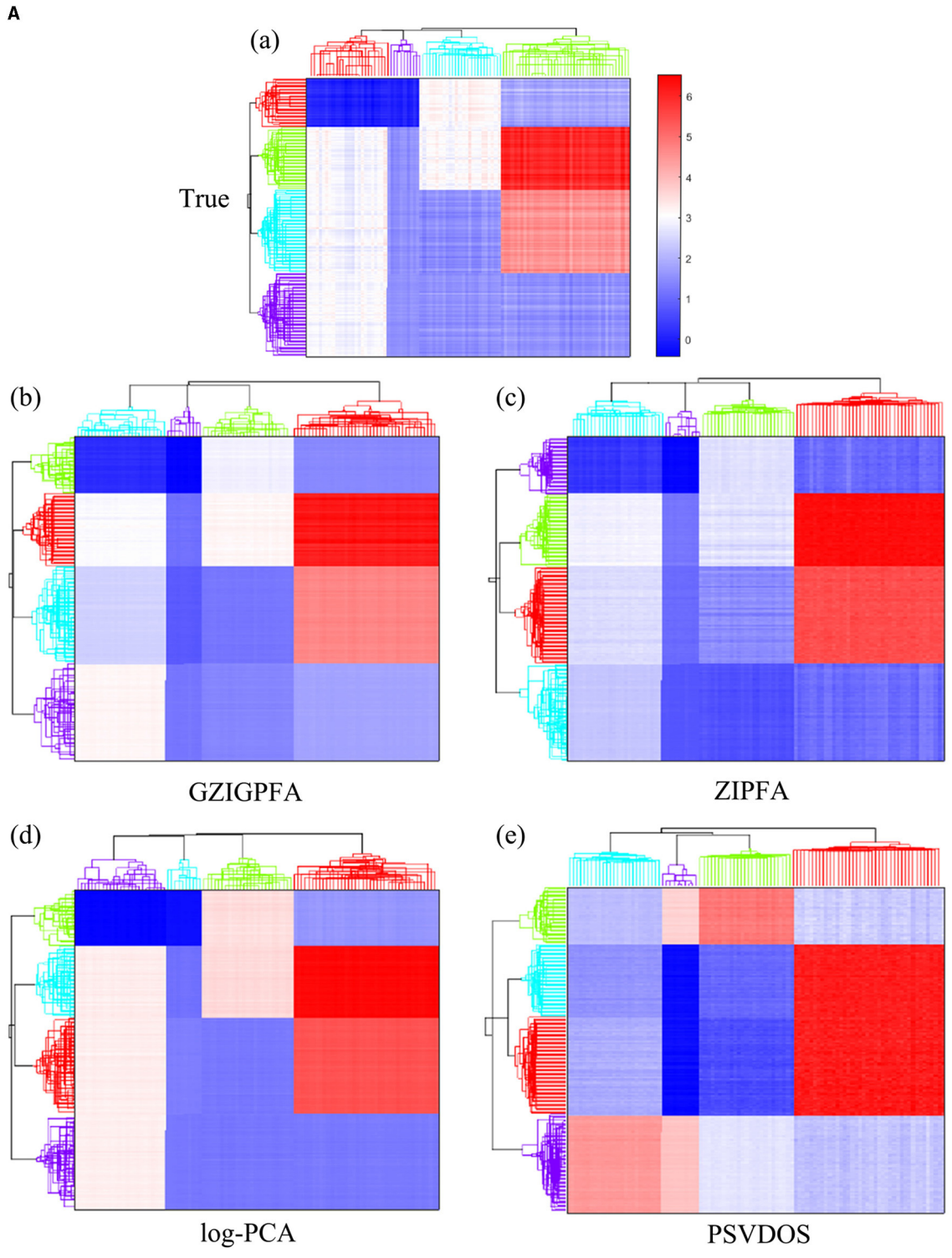

FIGURE 3 (Continued)

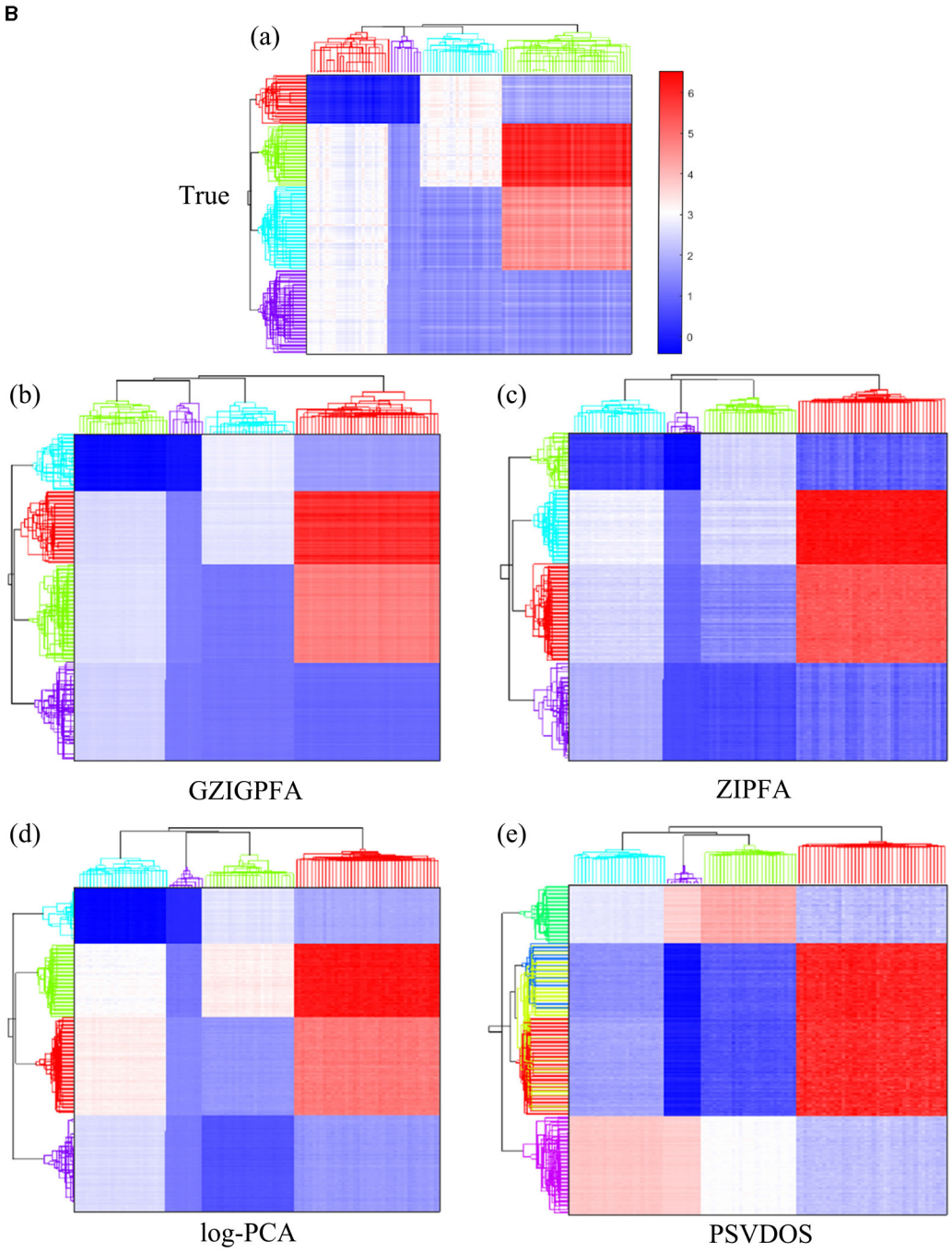

FIGURE 3 (Continued)
The heatmap of true log($\mathbf{\Lambda}$) and the estimated log($\mathbf{\Lambda}$) from different methods in Scenario 1. The phylogenetic tree on the top and left show the clustering of taxa and samples, respectively. **(A)** The zero-inflated proportion is 20%. **(B)** The zero-inflated proportion is 40%. The rows and columns of all heatmaps represent samples and taxa, respectively.

We selected body weight, total cholesterol, and endotoxin as three responses from the measured indicators of mice, where the weight of mice can intuitively reflect the degree of obesity. Obesity caused by a high-fat diet is often accompanied by hyperlipidemia, and total cholesterol (TC) is widely employed clinically as an indicator for measuring blood lipids. Endotoxin, also known as lipopolysaccharide (LPS), is a critical factor in the systemic inflammatory reaction. When the intestinal microbiota is imbalanced and harmful bacteria increase, the body is susceptible to endotoxemia, and sustained low-level endotoxemia is the leading cause of obesity and metabolic disorders. Therefore, we will focus on the relationship between the gut microbial community and three responses (weight, TC, and LPS).

We applied 10-fold cross-validation on microbial abundance data. Figure 4 shows that CV likelihood reaches the maximum point at a rank equal to 3, so we will use three factors in the

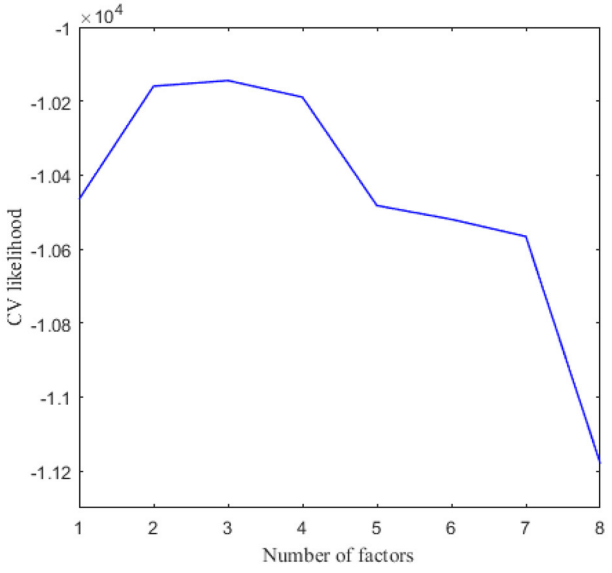

FIGURE 4
Cross validation to choose the number of factors in real data analysis.

following analysis. We performed GZIGPFA fitting with a rank of 3 on the microbiome data. The algorithm converged after 6 iterations and obtained score matrix estimates ($\boldsymbol{F}$) and loading matrix estimates ($\boldsymbol{L}$). We can compute $\log(\boldsymbol{\Lambda}) = \boldsymbol{F}\boldsymbol{L}^T$ according to the estimates of $\boldsymbol{F}$ and $\boldsymbol{L}$, and the zero-inflated probability matrix $\boldsymbol{\Phi}$ can be obtained through the relationship between $\boldsymbol{\Lambda}$ and $\boldsymbol{\Phi}$ assumed in Section 2.1 [i.e., $\text{logit}(\boldsymbol{\Phi}) = -\tau \log(\boldsymbol{\Lambda})$]. The total probability of zero for each count is estimated as $\hat{\phi}_{ij} + e^{-T_i\hat{\lambda}_{ij}/(1+\hat{\alpha}T_i\hat{\lambda}_{ij})}$. We reorder the total zero probability matrix and plot the corresponding heatmap (Figure 5A). In Figure 5A, the bottom right indicates the large values of total zero probability (red points), and the small total zero probability values are sorted to the top left (blue points). Meanwhile, the rearrangement of the true data is plotted in Figure 5B, where non-zero values are shown in the top left (blue points) and zeros in the bottom right (white points). We compare the predicted probability of zeros with the distribution of zeros in the real data. The significant similarity between the red part in Figure 5A and the white part in Figure 5B shows that the proposed method captures the structure of excess zeros well.

To determine the association between the three factors obtained through GZIGPFA dimensionality reduction and the three responses (weight, TC, and LPS), a linear model was fitted in which each response was regressed on all three factors, respectively. The p-values corresponding to different factors and responses are listed in Table 2. In addition, we demonstrate the results of the other comparison methods (ZIPFA, log-PCA, PSVDOS, and GOMMS) introduced in Section 3. It can be observed in Table 2 that the GZIGPFA and log-PCA methods can identify factors significantly associated with each response, while ZIPFA and PSVDOS failed to provide significant factors for TC and LPS. In addition, GOMMS is also unable to find factors associated with LPS. In particular, all five methods identified factors (factors 2 or 3) associated with weight, indicating that gut microbiome composition may be an essential factor influencing weight. Furthermore, factor 2 was a significant predictor of all responses in our proposed method, suggesting a potential link between obesity diseases and gut microbiome.

We consulted the literature to explain the factors obtained by GZIGPFA. By searching for gut microbes and obesity keywords on PubMed, some review articles were screened to identify microbes related to obesity. To identify microbes associated with obesity by searching for gut microbes and obesity keywords on Pubmed and filtering some review articles. After a full-text review of 116 papers, Pinart et al. (2021) concluded that *Firmicutes* and *Bacteroidetes* are the two microorganisms that mainly affect obesity at the phylum level. Therefore, factors 2 and 3 significantly associated with the obesity phenotype in the association analysis may be summarized as *Firmicutes* and *Bacteroidetes*. In conclusion, the proposed method can help the experimenter to determine the approximate factors affecting the experiment in advance.

Finally, in order to demonstrate that the proposed model does the absence of over-prediction problems, we additionally selected a response for analysis, i.e., tumor necrosis factor-alpha (TNF-α), which is not directly related to obesity. As can be seen from Table 2, all these methods did not identify factors significantly related to TNF-α, indicating that there is no over-prediction problem.

5 Discussion

Dimensionality reduction is a prevalent preprocessing step in high dimensional microbiome analysis. In this paper, we propose a new GLM-based zero-inflated generalized Poisson factor analysis model to analyze high-dimensional microbiome count data. This method explores the correlation between microbial taxa and response variables, and focuses on selecting a few common factors that summarize the majority of variable information, thus one can mitigate the high dimensionality problem and the computational expenses. The GZIGPFA model can simultaneously consider the zero-inflation, over-dispersion, and high-dimensional characteristics of microbial data. Meanwhile, the model directly models absolute abundance data, avoiding the problem of information loss during data conversion. We establish a link function between generalized Poisson expectation and true zero probability within the GLM framework, and perform parameter estimation using the alternating maximum likelihood algorithm. The rank of the model was determined via cross-validation method. In addition, we performed simulation studies under different scenarios and compared the GZIGPFA method with existing methods to validate the performance of the proposed method. In the analysis of gut microbiome data, the proposed method identified microorganisms significantly associated with obesity.

The novelty of the GZIGPFA method is reflected in the combination of the ZIGP model and the factor analysis model, which provides more possibilities for future microbial-related analysis work. Furthermore, upon review of the existing literature pertaining to microbiome data analysis, our proposed approach represents the inaugural utilization of the zero-inflated generalized Poisson model in microbial datasets, which expands the methodological options of researchers for addressing complex microbiome datasets. In addition to microbiome data, the proposed method can be used for other count data such as micro RNA

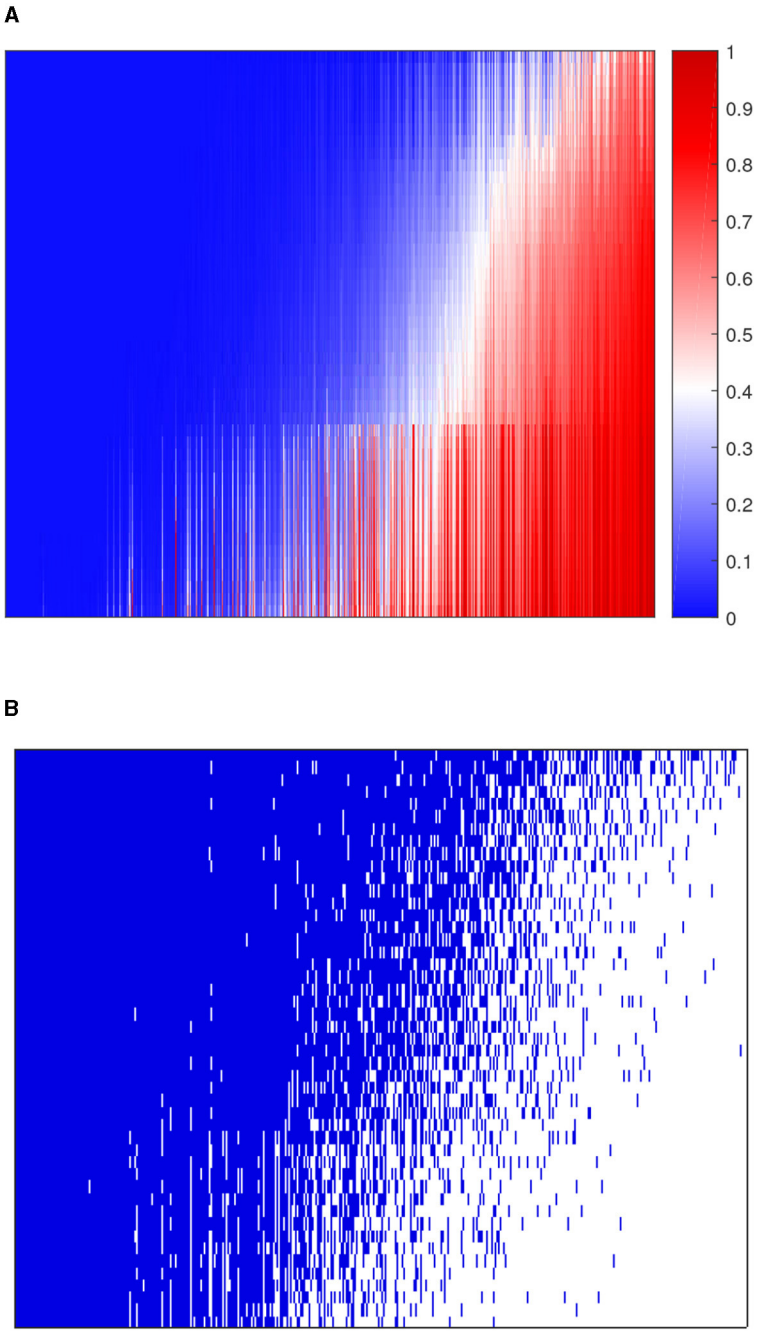

FIGURE 5
Comparison of predicted probability of zeros and real zero distribution in the dataset. **(A)** The heatmap of predicted zero probability. **(B)** The heatmap of the binary real data value. Blue points are non-zero values and white points are zeros. The rows and columns of both heatmaps represent samples and taxa, respectively.

data, single-cell RNA-seq data, etc. In addition, other suitable models can be extended to the framework of this article to provide more statistical methods for the analysis of high-dimensional microbiome data in the future.

The work presented in this paper remains subject to certain limitations. In this paper, a cross-validation method is used for rank estimation, which is accompanied by a high computational cost, although the results have high accuracy. In future work, the process of rank estimation can be further optimized to improve computational efficiency. In addition, the GZIGPFA model proposed in this article can only extract common factors associated with obesity phenotypes from numerous microbial taxa. The meaning of common factors needs to be determined based on existing prior information, and the interpretation of the actual meaning of each factor is not absolute. We can further extend our approach to provide a more comprehensive tool for the analysis of microorganisms in the future. Finally, although the performance of the log-PCA method in real data analysis closely resembles that

TABLE 2 The *P*-values corresponding to different factors and response variables in different models.

Response	Factor	GZIGPFA	ZIPFA	log-PCA	PSVDOS	GOMMS
Weight	Factor 1	0.2874	0.3727	0.3551	0.5600	0.0810
	Factor 2	0.0004***	0.4077	0.0003***	0.0482*	0.9480
	Factor 3	0.0281*	0.0008***	0.0493*	0.1706	0.0260*
LPS	Factor 1	0.1243	0.2850	0.5406	0.8740	0.0684
	Factor 2	0.0171*	0.2530	0.0092**	0.1460	0.4635
	Factor 3	0.7297	0.1050	0.7876	0.660	0.2885
TC	Factor 1	0.4014	0.7030	0.8232	0.3290	0.0099**
	Factor 2	0.0076**	0.5972	0.0035**	0.1430	0.9905
	Factor 3	0.7657	0.0504	0.5464	0.6070	0.1704
TNF-α	Factor 1	0.6075	0.0705	0.6760	0.7870	0.7837
	Factor 2	0.7743	0.0873	0.6760	0.4050	0.5351
	Factor 3	0.9345	0.4159	0.5480	0.3520	0.2076

***, **, and * denote the significance level takes 0.001, 0.01, 0.05, respectively.

of our method, it employs a strategy of replacing zeros in the data with pseudo counts. However, there is no consensus on how to choose the pseudo count, and it has been shown that the choice of pseudo count can affect the conclusions of a microbiome analysis (Costea et al., 2014; Paulson et al., 2014). The gut microbiome data that we used in real data analysis contains ~45% zeros, which is moderately zero-inflated. Perhaps the strategy of replacing zeros has less impact on the results, which may be the main reason why we did not show a clear advantage. Once a dataset shows a serious zero-inflated trend, the log-PCA method may become unstable. In the field of microbiology, it is common for microbiome data to be severely zero-inflated (Paulson et al., 2013; Silverman et al., 2020). Due to sharing restrictions on these data, we do not conduct a practical demonstration in this article.

Data availability statement

The mouse gut microbiome dataset for the real data analysis section was obtained with the support of Professor Qingshen Sun. The datasets presented in this article are not readily available because they they have not been made publicly available by Sun et al. (2019). Requests to access these datasets should be directed to corresponding author JC, jinlingchi_edu@163.com.

Ethics statement

The manuscript presents research on animals that do not require ethical approval for their study.

Author contributions

JC: Software, Visualization, Writing—original draft, Writing—review & editing, Methodology. JY: Funding acquisition, Writing—review & editing. YZ: Data curation, Funding acquisition, Writing—review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was supported by the National Natural Science Foundation of China (Grant No. 12071114) and Natural Science Basic Research Program of Shaanxi (Program No. 2024JC-YBMS-043).

Acknowledgments

The authors would like to thank the Qingshen Sun for providing the gut microbiome data.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2024.1394204/full#supplementary-material

References

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Nat. Prec.* 2010:1. doi: 10.1038/npre.2010.4282.1

Armstrong, G., Rahman, G., Martino, C., McDonald, D., Gonzalez, A., Mishne, G., et al. (2022). Applications and comparison of dimensionality reduction methods for microbiome data. *Front. Bioinformat.* 2:821861. doi: 10.3389/fbinf.2022.821861

Badri, M., Kurtz, Z. D., Bonneau, R., and Müller, C. L. (2020). Shrinkage improves estimation of microbial associations under different normalization methods. *NAR Genom. Bioinformat.* 2, 1–14. doi: 10.1101/406264

Campbell, H. (2021). The consequences of checking for zero–inflation and overdispersion in the analysis of count data. *Methods Ecol. Evol.* 12, 665–680. doi: 10.1111/2041-210x.13559

Carlisle, E. M., Poroyko, V., Caplan, M. S., Alverdy, J., Morowitz, M. J., and Liu, D. (2013). Murine gut microbiota and transcriptome are diet dependent. *Ann. Surg.* 257, 287–294. doi: 10.1097/sla.0b013e318262a6a6

Chen, J., and Li, H. (2013). Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Stat.* 7, 418–442. doi: 10.1214/12-aoas592

Chi, J., Zhou, Y., Chen, L., and Zhou, Y. (2020). Bayesian interval mapping of count trait loci based on zero–inflated generalized poisson regression model. *Biometr. J.* 62, 1428–1442. doi: 10.1002/bimj.201900274

Consul, P. C., and Famoye, F. (1992). Generalized poisson regression model. *Commun. Stat.* 21, 89–109.

Costea, P. I., Zeller, G., Sunagawa, S., and Bork, P. (2014). A fair comparison. *Nat. Methods* 11:359. doi: 10.1038/nmeth.2897

Cui, Y., and Yang, W. (2009). Zero-inflated generalized poisson regression mixture model for mapping quantitative trait loci underlying count trait with many zeros. *J. Theoret. Biol.* 256, 276–285. doi: 10.1016/j.jtbi.2008.10.003

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc. Ser. B* 39, 1–22.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comput. Biol.* 7:e1002195. doi: 10.1371/journal.pcbi.1002195

Egan, S., Harder, T., Burke, C., Steinberg, P., Kjelleberg, S., and Thomas, T. (2013). The seaweed holobiont: understanding seaweed–bacteria interactions. *FEMS Microbiol. Rev.* 37, 462–476. doi: 10.1111/1574-6976.12011

Famoye, F. (1993). Restricted generalized poisson regression model. *Commun. Stat.* 22, 1335–1354.

Famoye, F., and Singh, K. P. (2006). Zero-inflated generalized poisson regression model with an application to domestic violence data. *J. Data Sci.* 4, 117–130. doi: 10.6339/jds.2006.04(1).257

Fan, Y., Jiang, X., Hu, X., Song, B., Ling, Y., and Wu, W. (2015). "A novel dimensionality reduction algorithm based on laplace matrix for microbiome data analysis," in *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Washington, DC: IEEE.

Frank, D. N., Amand, A. L. S., Feldman, R. A., Boedeker, E. C., Harpaz, N., and Pace, N. R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci. U. S. A.* 104, 13780–13785. doi: 10.1073/pnas.0706625104

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 8:2224. doi: 10.3389/fmicb.2017.02224

Jasner, Y., Belogolovski, A., Ben-Itzhak, M., Koren, O., and Louzoun, Y. (2021). Microbiome preprocessing machine learning pipeline. *Front. Immunol.* 12:677870. doi: 10.3389/fimmu.2021.677870

Jiang, R., Zhan, X., and Wang, T. (2023). A flexible zero-inflated poisson-gamma model with application to microbiome sequence count data. *J. Am. Stat. Assoc.* 118, 792–804. doi: 10.1080/01621459.2022.2151447

Jiang, S., Xiao, G., Koh, A. Y., Kim, J., Li, Q., and Zhan, X. (2019). A bayesian zero-inflated negative binomial regression model for the integrative analysis of microbiome data. *Biostatistics* 22, 522–540. doi: 10.1093/biostatistics/kxz050

Kaul, A., Mandal, S., Davidov, O., and Peddada, S. D. (2017). Analysis of microbiome data in the presence of excess zeros. *Front. Microbiol.* 8:2114. doi: 10.3389/fmicb.2017.02114

Knights, D., Silverberg, M. S., Weersma, R. K., Gevers, D., Dijkstra, G., Huang, H., et al. (2014). Complex host genetics influence the microbiome in inflammatory bowel disease. *Genome Med.* 6, 1–11. doi: 10.1186/s13073-014-0107-1

Krych, L., Hansen, C. H. F., Hansen, A. K., van den Berg, F. W. J., and Nielsen, D. S. (2013). Quantitatively different, yet qualitatively alike: a meta-analysis of the mouse core gut microbiome with a view towards the human gut microbiome. *PLoS ONE* 8:e62578. doi: 10.1371/journal.pone.0062578

Kuczynski, J., Lauber, C. L., Walters, W. A., Parfrey, L. W., Clemente, J. C., Gevers, D., et al. (2011). Experimental and analytical tools for studying the human microbiome. *Nat. Rev. Genet.* 13, 47–58. doi: 10.1038/nrg3129

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* 11:e1004226. doi: 10.1371/journal.pcbi.1004226

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics* 34, 1–14.

Lee, S., Chugh, P. E., Shen, H., Eberle, R., and Dittmer, D. P. (2013). Poisson factor models with applications to non-normalized microRNA profiling. *Bioinformatics* 29, 1105–1111. doi: 10.1093/bioinformatics/btt091

Ley, R. E., Hamady, M., Lozupone, C., Turnbaugh, P. J., Ramey, R. R., Bircher, J. S., et al. (2008). Evolution of mammals and their gut microbes. *Science* 320, 1647–1651. doi: 10.1126/science.1155725

Li, G., Huang, J. Z., and Shen, H. (2018). Exponential family functional data analysis via a low-rank model. *Biometrics* 74, 1301–1310. doi: 10.1111/biom.12885

Lin, W., Shi, P., Feng, R., and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* 101, 785–797. doi: 10.1093/biomet/asu031

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:8. doi: 10.1186/s13059-014-0550-8

Machiels, K., Joossens, M., Sabino, J., Preter, V. D., Arijs, I., Eeckhaut, V., et al. (2013). A decrease of the butyrate-producing species *Roseburia hominis* and *Faecalibacterium prausnitzii* defines dysbiosis in patients with ulcerative colitis. *Gut* 63, 1275–1283. doi: 10.1136/gutjnl-2013-304833

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., and Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* 26:27663. doi: 10.3402/mehd.v26.27663

Martín-Fernández, J. A., Barceló-Vidal, C., and Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Math. Geol.* 35, 253–278. doi: 10.1023/a:1023866030544

Martín-Fernández, J. A., Hron, K., Templ, M., Filzmoser, P., and Palarea-Albaladejo, J. (2012). Model-based replacement of rounded zeros in compositional data: classical and robust approaches. *Comput. Stat. Data Anal.* 56, 2688–2704. doi: 10.1016/j.csda.2012.02.012

Martín-Fernández, J. A., Hron, K., Templ, M., Filzmoser, P., and Palarea-Albaladejo, J. (2014). Bayesian-multiplicative treatment of count zeros in compositional data sets. *Stat. Model.* 15, 134–158. doi: 10.1177/1471082x14535524

McCullagh, P., and Nelder, J. (1989). *Generalized Linear Models, 2nd Edn*. London: Chapman and Hall.

McKnite, A. M., Perez-Munoz, M. E., Lu, L., Williams, E. G., Brewer, S., Andreux, P. A., et al. (2012). Murine gut microbiota is defined by host genetics and modulates variation of metabolic traits. *PLoS ONE* 7:e39191. doi: 10.1371/journal.pone.0039191

Mishra, A. K., and Müller, C. L. (2022). Negative binomial factor regression with application to microbiome data analysis. *Stat. Med.* 41, 2786–2803. doi: 10.1002/sim.9384

Moghimbeigi, A. (2015). Two-part zero-inflated negative binomial regression model for quantitative trait loci mapping with count trait. *J. Theoret. Biol.* 372, 74–80. doi: 10.1016/j.jtbi.2015.02.016

Parks, B. W., Nam, E., Org, E., Kostem, E., Norheim, F., Hui, S. T., et al. (2013). Genetic control of obesity and gut microbiota composition in response to high-fat, high-sucrose diet in mice. *Cell Metab.* 17, 141–152. doi: 10.1016/j.cmet.2012.12.007

Paulson, J. N., Bravo, H. C., and Pop, M. (2014). Reply to: "a fair comparison." *Nat. Methods* 11, 359–360. doi: 10.1038/nmeth.2898

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* 10, 1200–1202. doi: 10.1038/nmeth.2658

Peng, X., Li, G., and Liu, Z. (2016). Zero-inflated beta regression for differential abundance analysis with metagenomics data. *J. Comput. Biol.* 23, 102–110. doi: 10.1089/cmb.2015.0157

Pierson, E., and Yau, C. (2015). ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* 16, 1–10. doi: 10.1186/s13059-015-0805-z

Pinart, M., Dötsch, A., Schlicht, K., Laudes, M., Bouwman, J., Forslund, S. K., et al. (2021). Gut microbiome composition in obese and non-obese persons: a systematic review and meta-analysis. *Nutrients* 14, 1–12. doi: 10.3390/nu14010012

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). <tt>Edger</tt>: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Silverman, J. D., Roche, K., Mukherjee, S., and David, L. A. (2020). Naught all zeros in sequence count data are the same. *Comput. Struct. Biotechnol. J.* 18, 2789–2798. doi: 10.1016/j.csbj.2020.09.014

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* 15, 121–132. doi: 10.1038/nrg3642

Singh, R. P., and Reddy, C. R. K. (2016). Unraveling the functions of the macroalgal microbiome. *Front. Microbiol.* 6, 1–8. doi: 10.3389/fmicb.2015.01488

Sohn, M. B., Du, R., and An, L. (2015). A robust approach for identifying differentially abundant features in metagenomic samples. *Bioinformatics* 31, 2269–2275. doi: 10.1093/bioinformatics/btv165

Sohn, M. B., and Li, H. (2017). A GLM-based latent variable ordination method for microbiome samples. *Biometrics* 74, 448–457. doi: 10.1111/biom.12775

Sun, Q., Liu, X., Zhang, Y., Song, Y., Ma, X., Shi, Y., et al. (2019). *L. plantarum*, *L. fermentum*, and *B. breve* beads modified the intestinal microbiota and alleviated the inflammatory response in high-fat diet–fed mice. *Probiot. Antimicrob. Prot.* 12, 535–544. doi: 10.1007/s12602-019-09564-3

Tang, Z. Z., and Chen, G. (2018). Zero-inflated generalized dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics* 20, 698–713. doi: 10.1093/biostatistics/kxy025

Tirozzi, P., Orioli, V., Dondina, O., Kataoka, L., and Bani, L. (2022). Population trends from count data: handling environmental bias, overdispersion and excess of zeroes. *Ecol. Informat.* 69:101629. doi: 10.1016/j.ecoinf.2022.101629

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027–1031. doi: 10.1038/nature05414

Tyler, A. D., Smith, M. I., and Silverberg, M. S. (2014). Analyzing the human microbiome: a "how to" guide for physicians. *Am. J. Gastroenterol.* 109, 983–993. doi: 10.1038/ajg.2014.73

Vijay-Kumar, M., Aitken, J. D., Carvalho, F. A., Cullender, T. C., Mwangi, S., Srinivasan, S., et al. (2010). Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. *Science* 328, 228–231. doi: 10.1126/science.1179721

Wagh, Y. S., and Kamalja, K. K. (2017). Zero-inflated models and estimation in zero-inflated poisson distribution. *Commun. Stat.* 47, 2248–2265. doi: 10.1080/03610918.2017.1341526

Wen, L., Ley, R. E., Volchkov, P. Y., Stranges, P. B., Avanesyan, L., Stonebraker, A. C., et al. (2008). Innate immunity and intestinal microbiota in the development of Type 1 diabetes. *Nature* 455, 1109–1113. doi: 10.1038/nature07336

Wilkinson, L., and Friendly, M. (2009). The history of the cluster heat map. *Am. Stat.* 63, 179–184. doi: 10.1198/tas.2009.0033

Willing, B. P., Dicksved, J., Halfvarson, J., Andersson, A. F., Lucio, M., Zheng, Z., et al. (2010). A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. *Gastroenterology* 139, 1844–1854. doi: 10.1053/j.gastro.2010.08.049

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Stat.* 11, 95–103.

Xu, L., Paterson, A. D., Turpin, W., and Xu, W. (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PLoS ONE* 10:e0129606. doi: 10.1371/journal.pone.0129606

Xu, T., Demmer, R. T., and Li, G. (2020). Zero-inflated poisson factor model with application to microbiome read counts. *Biometrics* 77, 91–101. doi: 10.1111/biom.13272

Yau, K. K. W., Wang, K., and Lee, A. H. (2003). Zero–inflated negative binomial mixed regression modeling of over–dispersed count data with extra zeros. *Biometr. J.* 45, 437–452. doi: 10.1002/bimj.200390024

Zeng, Y., Pang, D., Zhao, H., and Wang, T. (2022). A zero-inflated logistic normal multinomial model for extracting microbial compositions. *J. Am. Stat. Assoc.* 2022:2044827. doi: 10.1080/01621459.2022.2044827

Zhang, X., Pei, Y. F., Zhang, L., Guo, B., Pendegraft, A. H., Zhuang, W., et al. (2018). Negative binomial mixed models for analyzing longitudinal microbiome data. *Front. Microbiol.* 9:1683. doi: 10.3389/fmicb.2018.01683

Zhang, X., and Yi, N. (2020). Fast zero-inflated negative binomial mixed modeling approach for analyzing longitudinal metagenomics data. *Bioinformatics* 36, 2345–2351. doi: 10.1093/bioinformatics/btz973