Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

A Flexible Zero-Inflated Poisson-Gamma Model with Application to Microbiome Sequence Count Data

Roulan Jiang, Xiang Zhan & Tianying Wang

To cite this article: Roulan Jiang, Xiang Zhan & Tianying Wang (2023): A Flexible Zero-Inflated Poisson-Gamma Model with Application to Microbiome Sequence Count Data, Journal of the American Statistical Association, DOI: 10.1080/01621459.2022.2151447

To link to this article: https://doi.org/10.1080/01621459.2022.2151447

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 00, NO. 0, 1–13: Applications and Case Studies
https://doi.org/10.1080/01621459.2022.2151447

Taylor & Francis
Taylor & Francis Group

A Flexible Zero-Inflated Poisson-Gamma Model with Application to Microbiome Sequence Count Data

Roulan Jiang [a], Xiang Zhan [b], and Tianying Wang [a]

[a]Center for Statistical Science and Department of Industrial Engineering, Tsinghua University, Beijing, China; [b]Department of Biostatistics, School of Public Health, Beijing International Center for Mathematical Research and Center for Statistical Science, Peking University, Beijing, China

ABSTRACT

In microbiome studies, it is of interest to use a sample from a population of microbes, such as the gut microbiota community, to estimate the population proportion of these taxa. However, due to biases introduced in sampling and preprocessing steps, these observed taxa abundances may not reflect true taxa abundance patterns in the ecosystem. Repeated measures, including longitudinal study designs, may be potential solutions to mitigate the discrepancy between observed abundances and true underlying abundances. Yet, widely observed zero-inflation and over-dispersion issues can distort downstream statistical analyses aiming to associate taxa abundances with covariates of interest. To this end, we propose a Zero-Inflated Poisson Gamma (ZIPG) model framework to address these aforementioned challenges. From a perspective of measurement errors, we accommodate the discrepancy between observations and truths by decomposing the mean parameter in Poisson regression into a true abundance level and a multiplicative measurement of sampling variability from the microbial ecosystem. Then, we provide a flexible ZIPG model framework by connecting both the mean abundance and the variability of abundances to different covariates, and build valid statistical inference procedures for both parameter estimation and hypothesis testing. Through comprehensive simulation studies and real data applications, the proposed ZIPG method provides significant insights into distinguished differential variability and mean abundance. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received June 2022
Accepted November 2022

KEYWORDS
Longitudinal data;
Measurement error;
Poisson-Gamma distribution;
Sequence count data;
Zero-inflation

1. Introduction

The human microbiome consists of the collection of all microbes living in or on the human body and plays an important role in maintaining human health (Manor et al. 2020). Quantification of the microbiome usually proceeds by 16s rRNA sequencing or metagenomic shotgun sequencing, where sequence read counts are often summarized into a taxa count table. Here the word *taxa* generically refers to features such as operational taxonomic units or other taxonomic or functional groupings of bacterial sequences. A crucial task in microbiome research is to link these taxa counts to covariates of interest (e.g., disease status, health outcomes, and environmental conditions) via statistical analysis (Li 2015). To achieve this goal, one needs first to address some common challenges, such as zero inflation and over-dispersion in observed taxa counts, and more importantly, the discrepancy between observed taxa abundances in samples and true abundances in the underlying microbial ecosystem, such as the gut microbiota community, to guarantee rigor and reproducibility of the analysis results (Willis 2019).

Owing to biases introduced in sampling extraction, polymerase chain reaction (PCR) amplification, sequencing, bioinformatics prepossessing, and other possible experimental procedures, observed taxa abundances may not well reflect unobserved true abundances in the ecosystem. While multiple versatile statistical methods have been proposed to address the aforementioned issue for microbiome compositional data (Shi et al. 2022; Martin et al. 2020), measurement errors actually occur at latent count variables rather than proportions. Such a compositional transformation may lose some variation/dispersion information that is important to subsequent statistical analysis (McMurdie and Holmes 2014; Li et al. 2021; Xu et al. 2021). Moreover, recent research indicates that it is possible to quantify microbial load (and hence the absolute abundance of each taxon) using flow cytometry (Vandeputte et al. 2017). Following this research vein, we will propose valid statistical inference for microbiome count data accommodating the discrepancy between observed sample abundances and underlying true abundances. Specifically, motivated by a recent inference procedure based on multiple rarefaction-based resamplings (Hu et al. 2021), we take samples with repeated measures (or longitudinal measurements) to account for sampling fluctuations.

Like many high-throughput DNA sequencing assays exhibiting high sparsity, microbiome experiments often have about 50% or more zero measurements (Silverman et al. 2020). There are, in general, two types of approaches to handle these zeros in

CONTACT Xiang Zhan *zhanx@bjmu.edu.cn* Department of Biostatistics, School of Public Health, Beijing International Center for Mathematical Research and Center for Statistical Science, Peking University, Beijing, 100871, China; Tianying Wang *tianyingw@tsinghua.edu.cn* Center for Statistical Science and Department of Industrial Engineering, Tsinghua University, Beijing 100084, China.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

microbiome sequence count data. One is to impute zeros based on missing data scheme assumptions (Martın-Fernandez et al. 2011) or random matrix low-rank assumptions (Cao et al. 2020). The imputation approach is often coupled with downstream log-ratio-based compositional data analysis. The other approach is to propose a two-part model with a point probability mass at zero along with another parametric distribution. Examples include zero-inflated Poisson, zero-inflated negative binomial and many others (Li et al. 2018; Tang and Chen 2019; Zhang and Yi 2020; Xu et al. 2021; Zeng et al. 2022). While both approaches are popular in microbiome data analysis, a recent study demonstrates that a potential limitation of imputing zeros is that violation of underlying assumptions may distort downstream statistical analysis (Silverman et al. 2020). To this end, we will take the two-part modeling strategy to handle excessive zeros in microbiome data in this article.

Suppose samples are collected from n subjects, and every subject could have multiple measurements with counts of K taxa measured in each sample. For $i = 1, \ldots, n$, $j = 1, \ldots, n_i$ and $k = 1, \ldots, K$, denote W_{ijk} as the count of the kth taxon in the jth sample/measurement from the ith subject. For ease of presentation, we assume hereafter that sample j, with $j = 1, \ldots, n_i$, was collected longitudinally from subject i, without loss of generality. The sequencing depth, or library size, of each sample is $M_{ij} = \sum_{k=1}^{K} W_{ijk}$. To account for the aforementioned excessive zeros issue, the zero-inflated Poisson distribution has been proposed to model microbiome counts (Xu et al. 2021):

$$W_{ijk} \sim \begin{cases} 0 & \text{with probability } p_{ijk} \\ \text{Poisson}(\lambda_{ijk}) & \text{with probability } 1 - p_{ijk}, \end{cases} \quad (1)$$

where λ_{ijk} represents the mean abundance of taxon k on the jth observation from subject i, and p_{ijk} is the probability mass to model excessive zeros. A major critique of Poisson models is the failure to accommodate over-dispersion, which has been widely observed for sequence count data, including microbiome data. An alternative of Poisson is the negative binomial distribution, originally proposed for RNA sequence count data (Robinson et al. 2010; Love et al. 2014) and recently extended to microbiome data (Zhang and Yi 2020). The zero-inflated negative binomial (ZINB) distribution is given by

$$W_{ijk} \sim \begin{cases} 0 & \text{with probability } p_{ijk} \\ \text{NB}(\mu_{ijk}, \alpha_k) & \text{with probability } 1 - p_{ijk}, \end{cases} \quad (2)$$

where μ_{ijk} and α_k are the mean parameter and (over-)dispersion parameter of the negative binomial distribution, respectively. ZINB can also be expressed as a Gamma prior upon λ_{ijk} in ZIP with $E(\lambda_{ijk}) = \mu_{ijk}$ and $\text{var}(\lambda_{ijk}) = \mu_{ijk}^2 \alpha_k$. That is, $\lambda_{ijk} \sim \text{Gamma}\left(\alpha_k^{-1}, \mu_{ijk}\alpha_k\right)$, where α_k is a nuisance parameter depending on the kth taxon only.

Host factors, like disease status and dietary regimes, can impact microbiome stability, referred to as *dysbiosis* to describe the imbalance of microbiome community during some unhealthy conditions compared to normal ones (Petersen and Round 2014). Thus, it is of our interest to investigate the relationship between the taxa abundance variation and covariates, which is naturally caused by microbiome stability perturbation yet overlooked in most existing ZINB models. For illustration purposes, we use the taxon Burkholderiales bacterium from the diet-microbiome study (Johnson et al. 2019) to show potential abundance variation associated with covariates. One primary goal of the diet-microbiome study is to analyze the microbial difference between alcohol drinkers and teetotalers. According to boxplots for the raw data and the predicted distribution using the ZINB model by R package `pscl` (Zeileis et al. 2008), we observe that two groups have similar mean abundances, but the variance among alcohol drinkers is evidently larger, which cannot be captured by `pscl` (Figure 1). Therefore, it is crucial to consider variation in microbiome count data to obtain more robust analysis results.

To address potential limitations, we propose a Zero-Inflated Poisson-Gamma (ZIPG) framework, which provides flexible modeling by connecting both the mean abundance and its dispersion with different sets of covariates, respectively. First, we consider a hierarchical model by adjusting the mean parameter λ_{ijk} of ZIP (i.e., model (1)) with a multiplicative factor U_{ijk}, whose distribution is Gamma and can be viewed as a multiplicative measurement error. Further, we construct the ZIPG model and connect mean parameter λ_{ijk} and variation factor U_{ijk} to different sets of covariates. Note that our model is different from the traditional Bayesian expression of negative-binomial in the sense that we model the mean and variability of taxa abundance separately, providing a meaningful explanation from the individual-level variation perspective.

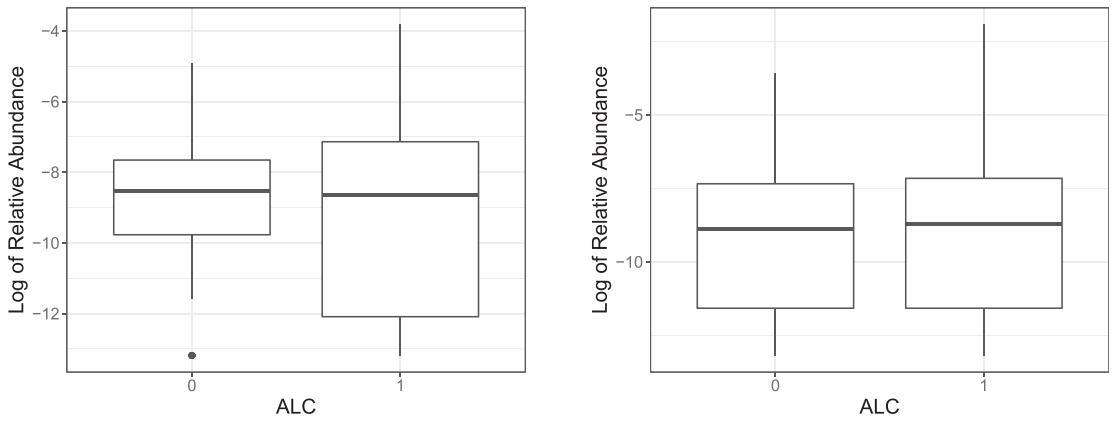
Our contributions are 2-fold. First, our ZIPG framework provides flexible modeling of microbiome sequence counts with repeated measures and allows us to analyze how different sets of variables affect both mean taxa abundance and its dispersion. Second, within the ZIPG framework, we develop inference procedures, including point and interval estimation and hypothesis testing, to examine the relationship between microbial taxa abundances and covariates of interest. By introducing the variation factor as a multiplicative measurement error term, our ZIPG method is able to capture higher-order moment information of taxa abundance and has been shown to be more powerful than ZINB-based methods. Through extensive simulations, we illustrate that existing ZINB-based methods could have severely inflated Type I error when differential variability exists, whereas ZIPG can control Type I error around the nominal level. When analyzing two real microbiome datasets, ZIPG identified more significant taxa than two ZINB models under the same nominal false discovery rate level, and also distinguished differential variability and differential abundance, providing more insights for further biological or biomedical functional investigations.

The rest of this article is organized as follows. In Section 2, we introduce our ZIPG model and discuss some parameters of interest. ZIPG model fitting and hypothesis testing procedures are proposed in Section 3. In Section 4, we demonstrate the superior performance of our approach under different simulation settings. We apply our method to datasets from a vaginal microbiome study and a diet-microbiome study in Section 5, and conclude it with a brief discussion in Section 6.

2. Model and Notation

2.1. Zero-Inflated Poisson Gamma Model

For taxon count W_{ijk}, we decompose the mean of Poisson distribution in Equation (1) into a true abundance level λ_{ijk} and a



(a) Raw data (b) pscl predicted data

Figure 1. Boxplot for log of relative abundance of Burkholderiales bacterium in alcohol (ALC = 1) and nonalcohol (ALC = 0) groups.

multiplicative factor U_{ijk} and consider the following hierarchical Zero-Inflated Poisson-Gamma (ZIPG) model:

$$W_{ijk} \mid U_{ijk} \sim \begin{cases} 0 & \text{with probability } p_k \\ \text{Poisson}(\lambda_{ijk}U_{ijk}) & \text{with probability } 1-p_k, \end{cases} \quad (3)$$
$$U_{ijk} \sim \text{Gamma}(\theta_{ik}^{-1}, \theta_{ik}),$$

where λ_{ijk} represents the true abundance level for taxon k on the jth observation from subject i, and p_k denotes the zero-inflation parameter describing the probability of true zero occurrence of taxon k. U_{ijk} follows Gamma distribution with the same rate parameter and shape parameter θ_{ik}^{-1}. This factor does not change the average abundance level given the fact that $E(U_{ijk}) = 1$. On the other hand, $\text{var}(U_{ijk}) = \theta_{ik}$ allows extra variation of the observed abundance W_{ijk} around the average level λ_{ijk}, a phenomenon described as the deviation of observed abundance to unobserved true abundance. It is of note that the variability term, θ_{ik}, remains the same for all measurements (i.e., $j = 1, \ldots, n_i$) across individual i. Thus, it reflects the stability of taxon k in the microbial system of individual i. This is motivated by the assumption about metagenomic sequencing bias being taxon-specific but not sample-specific made in literature (McLaren et al. 2019). Finally, we assume the zero-inflation parameter p_k is only taxon-specific and is common across samples (j) and individuals (i). This is because many experimental factors in sequencing can introduce the measurement of zeros (Silverman et al. 2020) and hence it is less appealing to link zero inflation parameter p_k to other covariates (e.g., biological or environmental conditions) possessed by individuals or samples. We have checked the sensitivity of ZIPG under model misspecification when this assumption is violated by comparing the performance of the current ZIPG model (3) to a full ZIPG model (denoted as ZIPG-full), which replaces p_k by p_{ijk} and links p_{ijk} to covariates. Results in Section 4.4 indicates that the current ZIPG model tends to have better model-fitting performance than ZIPG-full, which is consistent with previous empirical conclusions that modeling zero inflation in sequence count data should be careful (with respect to underlying zero generating process) and numerical evidence tends to favor simpler models (Silverman et al. 2020).

To further explore the new ZIPG framework, Let $W_{ijk}^{\text{Pois}} \sim \text{Poisson}(\lambda_{ijk})$ be the random variable generated from the Poisson distribution and W_{ijk}^{PG} be the random variable generated from the Poisson-Gamma part in Equation (3). We have

$$E(W_{ijk}^{\text{PG}}) = E(W_{ijk}^{\text{Pois}}) = \lambda_{ijk},$$
$$\text{var}(W_{ijk}^{\text{PG}}) = \lambda_{ijk}(1+\lambda_{ijk}\theta_{ik}) = \text{var}(W_{ijk}^{\text{Pois}})(1+\lambda_{ijk}\theta_{ik}).$$

Thus, the mean of Poisson-Gamma distribution is the same as the regular Poisson distribution, but its variance is multiplied by $(1+\lambda_{ijk}\theta_{ik})$ to account for the over-dispersion caused by the multiplicative measurement error factor U_{ijk}. We also observe the similar phenomenon of using a more sophisticated hierarchical model to account for over-dispersion in microbiome data analysis, such as the Beta-Binomial distribution (Martin et al. 2020) and Dirichlet-multinomial distribution (La Rosa et al. 2012). In this article, we refer to λ_{ijk} as the abundance mean parameter and θ_{ik} as the abundance dispersion parameter.

2.2. Parameters of Interest

A critical task in microbiome research is to explore the relationship between taxa abundances and covariates of interest. Compared to noisy observed counts W_{ijk}, it is more interesting to investigate the association between key parameters (i.e., λ_{ijk} and θ_{ik}) of the underlying taxa abundance distribution and covariates of interest. To this end, we connect the mean parameter and dispersion parameter with different sets of covariates, respectively. In the repeated measures or longitudinal study design considered in the current article, some covariates vary across different samples within the same subject, such as dietary intake, and we refer to them as "time-dependent" covariates. Other covariates, which do not change during the study, such as the treatment group assigned at the beginning of the study, are referred to as "time-independent" covariates. Since the dispersion parameter θ_{ik} describes the deviation of short-term abundance from long-term mean abundance λ_{ijk}, we propose to link it to time-independent covariates, supported by the evidence of microbiome stability perturbation in Morgan et al. (2012) and Couch et al. (2021). For the mean abundance

λ_{ijk}, it can be linked to either time-dependent covariates or time-independent covariates, or both. Therefore, we define the following link functions:

$$g(\lambda_{ijk}) = \beta_{k,0} + \boldsymbol{X_{ij}}^T \boldsymbol{\beta_k} + \log(M_{ij}),$$
$$g^*(\theta_{ik}) = \beta^*_{k,0} + \boldsymbol{X_i^{*T}} \boldsymbol{\beta^*_k}, \tag{4}$$

where $\boldsymbol{X_{ij}} \in \mathbb{R}^{d_1}$ is a vector of covariates associated with λ_{ijk}, $\boldsymbol{X_i^*} \in \mathbb{R}^{d_2}$ include covariates associated with θ_{ik}, $\boldsymbol{\beta_k} = (\beta_1, \ldots, \beta_{d_1})^T$ and $\boldsymbol{\beta^*_k} = (\beta^*_1, \ldots, \beta^*_{d_2})^T$ are regression coefficients of interest. We allow overlapped covariates in two models. For ease of presentation, we term the two models in (4) as ZIPG mean model and ZIPG dispersion model, respectively. The $\log(M_{ij})$ term in the mean model accounts for the effect of sequencing depth variation on mean abundances. The same offset is used in previous ZINB models (Robinson et al. 2010; Zhang and Yi 2020), and the log of median-of-ratios is another possible candidate for offset used in literature (Love et al. 2014; Xu et al. 2021). While most existing ZINB-based methods (2) models μ_{ijk} but treat α_k as a nuisance parameter, our approach allows additional time-independent covariates linked to the dispersion of abundance, which would lead to better model fit and more powerful association analysis as will be shown later in this article. According to previous discussions, we suggest including time-independent covariates such as demographic and lifestyle-related variables in $\boldsymbol{X_i^*}$. All covariates of interest, regardless of time-dependent or not, shall be included in $\boldsymbol{X_{ij}}$. By testing the coefficients $\boldsymbol{\beta_k}$ and $\boldsymbol{\beta^*_k}$, we can detect differential abundance and differential variability impacted by physiological status or host environment, respectively. Finally, if we do not include any covariates in $\boldsymbol{X_i^*}$, our ZIPG model will degenerate to ZINB (i.e., model (2)) with $\theta_{ik} = \alpha_k$ for any subject. Throughout this article we choose a logarithmic link function $g(x) = g^*(x) := \log(x)$ to ensure $\lambda_{ijk} > 0$ and $\theta_{ik} > 0$.

The proposed ZIPG model has several key advantages. First, to handle over-dispersion in microbiome count data, we decompose abundances into the long-term true abundance and its individual-level variation through a multiplicative factor. Second, we allow different sets of variables associated with the mean and the variation of abundance and provide explanations for variations in the individual-level microbial system. Thus, the proposed method is not only able to test the change of the mean abundance but also the microbiome stability affected by covariates, which cannot be achieved by existing ZIP models. In addition, parameter θ_{ik} also controls skewness and kurtosis of the Gamma distribution. That is, we allow the higher-order moments (or the shape of the distribution) to be linked to covariates, which is another feature that is typically missed in existing models.

3. Statistical Inference in ZIPG

In this section, we develop statistical inference procedures in ZIPG, including parameter point estimation, interval estimation, and hypothesis testing. For ease of presenting, we omit the subscript k and simply denote the parameter set associated with taxon k as $\boldsymbol{\Omega} = (\beta_0, \boldsymbol{\beta}^T, \beta^*_0, \boldsymbol{\beta^*}^T, \gamma)^T$, where $\gamma = \log\{p/(1-p)\}$ is the logit transformation of zero-inflated parameter p in ZIPG model (3).

3.1. Model Fitting

Given covariates $\boldsymbol{X}$ and $\boldsymbol{X^*}$, observed count data $\boldsymbol{W}$ and sequencing depth $\boldsymbol{M}$, we write the log-likelihood of $\boldsymbol{\Omega}$ as follows:

$$\begin{aligned} L(\boldsymbol{\Omega} \mid \boldsymbol{W}) = \sum_{i=1}^{n}\sum_{j=1}^{n_i} \big[& I(W_{ij}=0)\log\{\exp(\gamma) + P_{\mathrm{PG}}(W_{ij} \mid \boldsymbol{\Omega})\} \\ & + I(W_{ij}>0)\log\{P_{\mathrm{PG}}(W_{ij} \mid \boldsymbol{\Omega})\} \\ & - \log\{\exp(\gamma)+1\}\big], \end{aligned} \tag{5}$$

$$\text{with } P_{\mathrm{PG}}(W_{ij} \mid \boldsymbol{\Omega}) = \frac{\Gamma(W_{ij}+\theta_i^{-1})}{\Gamma(W_{ij}+1)\Gamma(\theta_i^{-1})} \frac{(\lambda_{ij}\theta_i)^{W_{ij}}}{(1+\lambda_{ij}\theta_i)^{\theta_i^{-1}+W_{ij}}},$$

where λ_{ij} and θ_i are functions of $\boldsymbol{\Omega}$ defined in Equation (4) and $\Gamma(x) = \int_0^{+\infty} t^{x-1}\exp(-t)\mathrm{d}t$ is the Gamma function. The log-likelihood Equation (5) is nonconcave in $\boldsymbol{\Omega}$ (see Section 1.1 of supplementary materials). In practice, we found that directly maximizing Equation (5) can cause trouble in distinguishing zeros from the Poisson-Gamma part and the other zero-inflation part of the ZIPG model, leading to an unreasonably low estimator of γ. A similar phenomenon has been observed for pscl, with more discussions provided in Section 4. Therefore, we use the EM algorithm for a reliable estimator of $\boldsymbol{\Omega}$.

Let z_{ij} be the latent variable, where $z_{ij} = 1$ indicates W_{ij} is generated from zero-inflated part with probability $p = \exp(\gamma)/(\exp(\gamma)+1)$, and $z_{ij} = 0$ indicates W_{ij} is generated from the Poisson-Gamma distribution with probability $1-p$. The log-likelihood with complete data $\{W_{ij}, M_{ij}, X_{ij}, X^*_{ij}, z_{ij}\}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$ is written as

$$\begin{aligned} & L(\boldsymbol{\Omega} \mid \boldsymbol{W}, \boldsymbol{z}) \\ & \quad = \sum_{i,j} \big[z_{ij}\log(p) + (1-z_{ij})\log\{(1-p)P_{\mathrm{PG}}(W_{ij} \mid \boldsymbol{\Omega})\}\big]. \end{aligned} \tag{6}$$

The detailed procedure of the EM algorithm is provided in Algorithm 1. We first initialize $\boldsymbol{\Omega}^{(0)}$ by the results of `zeroinfl` in `pscl` with $\boldsymbol{\beta^*} = \boldsymbol{0}$. $p_{ij}^{(0)}$ is adjusted to the proportion of observed zeros of $\boldsymbol{W}$ to avoid the local maximum at the start point. Then we can repeat the E-step and M-step until convergence or the maximum number of iterations $t_{\max}$ is reached. For the tth iteration, in M-step, we update $\boldsymbol{\Omega}$ by maximizing Equation (6) given latent variable $\boldsymbol{z}^{(t-1)}$:

$$\boldsymbol{\Omega}^{(t)} = \arg\max_{\boldsymbol{\Omega}} L\left(\boldsymbol{\Omega} \mid \boldsymbol{W}, \boldsymbol{z}^{(t-1)}\right). \tag{7}$$

We use BFGS in R function `optim` (Broyden 1970; Fletcher 1970; Goldfarb 1970; Shanno 1970), and the gradient of Equation (6) is applied to improve the computational efficiency (see Section 1.2 of supplementary material). In E-step, we update latent variables z_{ij} by their conditional expectations given $\boldsymbol{\Omega}^{(t)}$ estimated from M-step:

$$\begin{aligned} z_{ij}^{(t)} &= E\left\{z_{ij} \mid W_{ij}, \boldsymbol{\Omega}^{(t)}\right\} \\ &= \frac{I(W_{ij}=0)p^{(t)}}{I(W_{ij}=0)p^{(t)} + P_{\mathrm{PG}}(W_{ij} \mid \boldsymbol{\Omega}^{(t)})(1-p^{(t)})}. \end{aligned} \tag{8}$$

Substituting latent variables $z_{ij}^{(t)} = z(W_{ij}, \boldsymbol{\Omega}^{(t)})$ into Equation (6), we can write the expectation of log-likelihood on a single observation as a new function

$$\begin{aligned} Q_{ij}(\boldsymbol{\Omega} \mid \boldsymbol{\Omega}^{(t)}) &= E\left\{L(\boldsymbol{\Omega} \mid W_{ij}, z_{ij}) \mid W_{ij}, \boldsymbol{\Omega}^{(t)}\right\} \\ &= z(W_{ij}, \boldsymbol{\Omega}^{(t)}) \log p + \left\{1 - z(W_{ij}, \boldsymbol{\Omega}^{(t)})\right\} \\ &\quad \log\left\{(1-p) P_{\mathrm{PG}}(W_{ij} \mid \boldsymbol{\Omega})\right\}, \end{aligned}$$

then $Q(\boldsymbol{\Omega} \mid \boldsymbol{\Omega}^{(t)}) = \sum_{ij} Q_{ij}(\boldsymbol{\Omega} \mid \boldsymbol{\Omega}^{(t)})$ is the quantity maximized in M-step equivalently.

According to Theorem 6 in Wu (1983), since $\partial Q(\boldsymbol{\Omega} \mid \boldsymbol{\Omega}^*)/\partial \boldsymbol{\Omega}$ is continuous in $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}^*$, and our algorithm ensures $\partial Q(\boldsymbol{\Omega} \mid \boldsymbol{\Omega}^{(t)})/\partial \boldsymbol{\Omega}\,|_{\boldsymbol{\Omega}=\boldsymbol{\Omega}^{(t+1)}} = \mathbf{0}$ in each iteration, then EM estimator will converge to a stationary point of $L(\boldsymbol{\Omega} \mid \boldsymbol{W})$. Under mild conditions, EM algorithm converges to a local maximum (Wu 1983, Theorem 3). Based on our suggested initialization, namely the MLE assuming $\boldsymbol{\beta}^* = 0$ with adjusted $p^{(0)}$, numerical studies suggest that our estimators are nearly unbiased.

Algorithm 1 ZIPG Expectation Maximization Algorithm

Require: $\boldsymbol{W}, \boldsymbol{M}, \boldsymbol{X}, \boldsymbol{X}^*$, the maximum iterations $t_{\max}$, a tolerance ϵ_{tol}.
Initialize $\boldsymbol{\Omega}^{(0)}$ by adjusted pscl estimation regardless of $\boldsymbol{X}^*$, set $t = 0$.
Initialize $\boldsymbol{z}^{(0)}$ with $\boldsymbol{\Omega}^{(0)}$ based on Equation (8).
Calculate $L^{(0)} = L(\boldsymbol{\Omega}^{(0)} \mid \boldsymbol{W}, \boldsymbol{z}^{(0)})$ as Equation (6).
while $t < t_{\max}$ and $|L^{(t)} - L^{(0)}|/|L^{(0)}| < \epsilon_{\mathrm{tol}}$ **do**
  Given $\boldsymbol{z}^{(t-1)}$, estimate $\boldsymbol{\Omega}^{(t)}$ by Equation (7) (M-step).
  Get the maximized $L^{(t)} = L(\boldsymbol{\Omega}^{(t)} \mid \boldsymbol{W}, \boldsymbol{z}^{(t-1)})$ as in Equation (6).
  Given $\boldsymbol{\Omega}^{(t)}$, update $\boldsymbol{z}^{(t)}$ by Equation (8) (E-step).
  $t = t + 1$.
end while
Return $\boldsymbol{\Omega}^{(t)}$.

3.2. Hypothesis Testing

In this article, proving the asymptotic normality of the EM estimator is less of our interest, and despite its lack of rigorous theoretical justification, the EM estimator has been routinely treated as MLE in literature. Thus, we treat EM estimator $\hat{\boldsymbol{\Omega}}$ as MLE and construct the Wald test statistics and confidence interval based on the asymptotic normality of MLE. Further, we carefully consider the potential practical issues and evaluate different bootstrap methods and interval construction strategies with extensive numerical studies.

Consider the null hypothesis $H_0 : \boldsymbol{A\Omega} = \boldsymbol{b}$ as the general form to test arbitrary subsets and linear combinations of parameters within $\boldsymbol{\Omega}$, where $\boldsymbol{A} \in \mathbb{R}^{r \times (d_1+d_2+3)}$ has full rank, $r < d_1 + d_2 + 3$, and $\boldsymbol{b} \in \mathbb{R}^r$. The classical Wald test statistic can be constructed as $T_{\mathrm{Wald}} = N(\boldsymbol{A\hat{\Omega}} - \boldsymbol{b})^T (\boldsymbol{AVA}^T)^{-1} (\boldsymbol{A\hat{\Omega}} - \boldsymbol{b})$, where $\boldsymbol{V}$ is the asymptotic covariance matrix of $\boldsymbol{\Omega}$, N is the sample size, T_{Wald} is asymptotic χ_r^2 under the null hypothesis. In practice, we found that directly using the inverse of observed information derived from the last M-step might underestimate the variance matrix (see Section 3.1 of the supplementary material). Thus, we propose to use nonparametric bootstrap to estimate the covariance matrix and construct the test statistic as in Algorithm 2.

Algorithm 2 ZIPG Bootstrap Wald Test

Require: $\boldsymbol{W}, \boldsymbol{M}, \boldsymbol{X}, \boldsymbol{X}^*$, bootstrap replicates B.
Estimate $\hat{\boldsymbol{\Omega}}^{\boldsymbol{0}}$ by EM (Algorithm 1).
for $b = 1, \ldots, B$ **do**
  Randomly draw samples regarding all measurements from original data at the same sample size with replacement, thus, we get $\boldsymbol{W}^{\boldsymbol{b}}, \boldsymbol{M}^{\boldsymbol{b}}, \boldsymbol{X}^{\boldsymbol{b}}, \boldsymbol{X}^{*\boldsymbol{b}}$.
  Estimate $\hat{\boldsymbol{\Omega}}^{\boldsymbol{b}}$ using $\boldsymbol{W}^{\boldsymbol{b}}, \boldsymbol{M}^{\boldsymbol{b}}, \boldsymbol{X}^{\boldsymbol{b}}, \boldsymbol{X}^{*\boldsymbol{b}}$ by EM (Algorithm 1).
end for
Compute the covariance matrix of $\hat{\boldsymbol{\Omega}}^{\boldsymbol{b}}$ as $\hat{\boldsymbol{V}} = \mathrm{var}(\hat{\boldsymbol{\Omega}}^{\boldsymbol{b}})$.
Compute

$$\hat{T}_{\mathrm{Wald}} = (\boldsymbol{A\hat{\Omega}^0} - \boldsymbol{b})^T (\boldsymbol{A\hat{V}A}^T)^{-1} (\boldsymbol{A\hat{\Omega}^0} - \boldsymbol{b}).$$

The $100(1-\alpha)\%$ confidence interval for any single parameter $\beta \in \boldsymbol{\Omega}$ can also be obtained through nonparametric bootstrap Wald test as $\left(\hat{\beta} - z_{\alpha/2}\mathrm{SD}(\beta^b), \hat{\beta} + z_{\alpha/2}\mathrm{SD}(\beta^b)\right)$ based on the standard error (SD) of $\beta^b \in \boldsymbol{\Omega}^{\boldsymbol{b}}$ from B bootstrap samples. When the sample size is small or when the collected data is unbalanced (e.g., Romero data in Section 5), parametric bootstrap could be applied for more robust results as in Martin et al. (2020). Thus, we also developed the Wald test based on parametric bootstrap (i.e., ZIPG-pbWald). That is, we simulate bootstrap samples from the ZIPG model with its parameters estimated under H_0. A detailed algorithm is provided in Section 2 of the supplementary material.

There are multiple ways to construct test statistics and confidence intervals. We conducted extensive simulations studies evaluating the performance of (a) the Wald test without bootstrap (ZIPG-Wald) and the likelihood ratio test (ZIPG-LRT); (b) nonparametric and parametric bootstrap through different construction strategies, such as normality-based/quantile-based/$\mathrm{BC_a}$(Efron and Tibshirani 1994) intervals; (c) resampling schemes for ZIPG-bWald, such as resampling based on measurements or subjects. Simulation results suggest that bootstrap-based Wald tests (i.e., ZIPG-bWald and ZIPG-pbWald) with normality-based confidence intervals are desired, and resampling based on measurements yields satisfactory results. More discussions are provided in Section 4.5.

4. Simulation Studies

We have conducted comprehensive numerical studies to evaluate the performance of ZIPG in terms of both hypothesis testing and point/interval estimation. We evaluated both ZIPG with bootstrap Wald test (denoted as ZIPG-bWald) and parametric bootstrap Wald test (denoted as ZIPG-pbWald), and then we compared them with ZINB-based methods, including NBZ-IMM (Zhang and Yi 2020) and pscl (Zeileis et al. 2008). We also assessed the Poisson-Gamma (PG) model implemented by `glmmtmb` (Brooks et al. 2017) with the Wald test (no bootstrap). PG can link covariates to λ_{ij} and θ_i as in ZIPG, but it is not adjusted for inflated zeros. We summarized all methods compared in Section 4 in Table 1.

Table 1. Summary of methods evaluated in simulation studies.

Method	λ	θ	Zero-inflation	Notes
ZIPG (proposed)	✓	✓	✓	(parametric) bootstrap-based test
pscl	✓	×	✓	
NBZIMM	✓	×	✓	consider random effect on λ
PG	✓	✓	×	

Besides hypothesis testing and estimation, we further check the sensitivity of proposed ZIPG inference procedures under misspecified models. In particular, we evaluated the performance of ZIPG when the true zero proportion p is also associated with covariates (Section 4.4), and when data is generated from Poisson-Gamma distribution or zero-inflated Beta-Binomial distribution (see Section 3.5 of the supplementary material).

4.1. Simulation Settings

We simulated $n = 20$ subjects with $m = \{5, 25\}$ measurements for each subject, with a total sample size of $N = \{100, 500\}$. For $i = 1, \ldots, n$ and $j = 1, \ldots, m$, we generated covariates $\boldsymbol{X_{ij}} = (X_{i,1}, X_{ij,2})$ associated with the mean parameter λ_{ij} and $\boldsymbol{X_i^*} = X_{i,1}$ associated with the dispersion parameter θ_i, where $X_{i,1}$ was a time-independent indicator sampled from a Bernoulli distribution with equal probabilities and served as the group indicator for each subject, such as pregnant or nonpregnant, alcohol drinker or nonalcohol drinker. $X_{ij,2}$ was a time-dependent longitudinal measurement generated from $X_{ij,2} = X_{i,2} + \epsilon_{ij}$, where $X_{i,2} \sim \mathcal{N}(0, 1)$ and $\epsilon_{ij} \sim \mathcal{N}(0, 0.1)$, representing covariates with variation in different measurements, such as calorie intake. Sequencing depths M_{ij}'s were generated based on the empirical distribution of observed sequencing depths in the Romero dataset (Romero et al. 2014) analyzed later in Section 5. Finally, the observed count data $\boldsymbol{W}$ was generated based on models (3)–(4) with $\boldsymbol{\beta}$ and $\boldsymbol{\beta^*}$ specified as below.

To imitate real-world microbial data, we set $(\beta_0, \beta_0^*) = (-4.23, 0.6)$, guided by the ZIPG estimates in Romero data. We investigate performance of ZIPG under different model parameter configurations (Table 2). In each simulation setting, we explored three different values of the zero-inflation proportion $p = \{0.3, 0.5, 0.7\}$, which is equivalent to $\gamma = \{-0.847, 0, 0.847\}$. For power analysis, we set $p = 0.5$ and vary the parameter of interest under each null hypothesis from 0 to 1.8. We compare ZIPG-bWald, ZIPG-pbWald, pscl, NBZIMM, and PG for the inference of β_1 and only ZIPG-bWald, ZIPG-pbWald, and PG for the inference on β_1^*, because inference of the dispersion parameter is not applicable in pscl and NBZIMM. Results are presented based on $L = 1000$ Monte Carlo replicates for each scenario. The performances of ZIPG-bWald and ZIPG-pbWald are evaluated using $B = 200$ bootstrap samples, which are numerically sufficient and stable based on our experience.

4.2. Hypothesis Testing Results

Type I error analysis for β_1 and β_1^*. For $H_0 : \beta_1 = 0$ in the mean model, we observe that both pscl and NBZIMM have inflated Type I errors under all simulation scenarios, and PG is too conservative regardless of the proportion p of inflated zeros (Figure 2(a)). In particular, the Type I error of pscl and NBZIMM increase significantly with the increase of β_1^*, indicating that ignoring differential variability could lead to severely increased false positives for the mean model. Of note, the Type I error of ZIPG with the bootstrap-based Wald test (ZIPG-bWald) is slightly inflated in a few cases with a small total number of observations ($N = 100$), while its results are satisfactory with a larger sample size ($N = 500$). In general, the Type I error of ZIPG is robustly controlled at the nominal level $\alpha = 0.05$, regardless of the change of β_1^* and p.

For $H_0 : \beta_1^* = 0$ in the dispersion model, we observe that ZIPG-bWald maintains a controlled Type I error for β_1^*, yet a little conservative when $N = 100$, while ZIPG-pbWald controls Type I error to 0.05 more robustly (Figure 2(b)). Therefore, we suggest using ZIPG-pbWald as an alternative for hypothesis testing in small-sample scenarios. However, when we have a larger sample size ($N = 500$), both ZIPG-bWald and ZIPG-pbWald perform similarly.

Power analysis for β_1 and β_1^*. We present power results of testing $H_0 : \beta_1 = 0$ (Figure 3(a)) and $H_0 : \beta_1^* = 0$ (Figure 3(b)). Since pscl, NBZIMM, and PG fail to preserve the nominal Type I error, we do not evaluate their empirical power in the power analysis. For both null hypotheses, the power curves increase with the increase of sample size and true effect size. ZIPG-pbWald and ZIPG-bWald perform similarly in larger sample cases, while ZIPG-pbWald is relatively more powerful for detecting differential variability ($H_0 : \beta_1^* = 0$) with a small sample size ($N = 100$). We only report the common covariates in both the mean model and dispersion model in the main text, and hypothesis testing results on covariates only in the mean model are reported in Section 3.3 of the supplementary material.

4.3. Point and Interval Estimation Results

To demonstrate the advantage of ZIPG regarding point estimators and confidence intervals, we report the average bias (e.g., $\{\sum_l(\hat{\beta}_l - \beta)\}/L$) with its standard error over $l = 1, \ldots, L = 1000$ Monte Carlo replicates, average bootstrap standard error (avg-SE, e.g., $\{\sum_l \widehat{SE}(\hat{\beta}_l)\}/L$), root mean squared error (RMSE, e.g., $\sqrt{\{\sum_l(\hat{\beta}_l - \beta)^2\}/L}$), and converge rate of confidence intervals (CR) for the settings with $\beta_1 = 0, \beta_1^* = 1$ (Figure 2(a)) and $\beta_1 = 1, \beta_1^* = 0$ (Figure 2(b)) with $p = \{0.5, 0.7\}$ and $N = 500$. Results under other settings are similar and hence not reported.

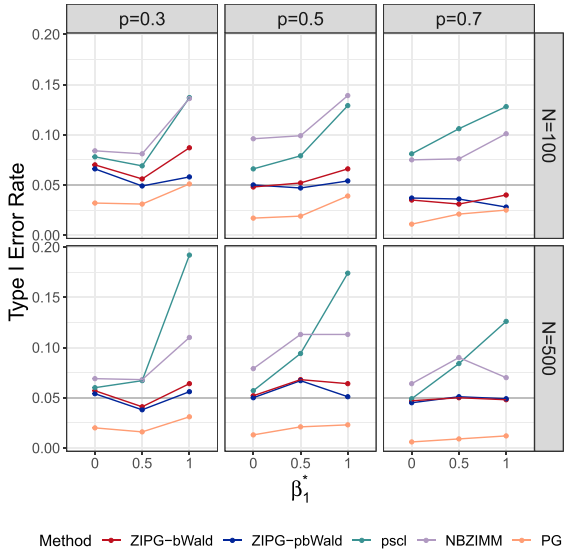
In Table 3, we observe that ZIPG-bWald has the smallest bias of β_1, β_1^* and γ among all methods when $\beta_1^* = 1$. In addition, ZIPG is often more efficient than the other two methods, providing a smaller RMSE. For β_1 and β_1^*, ZIPG-bWald always maintains a valid confidence interval with its coverage rate close to nominal level 0.95, whereas pscl and NBZIMM provide underestimated confidence intervals in most cases, and PG estimate β_1^* with strong bias and provide a more conservative CI for β_1. Additional results with setting $\beta_1 = 1, \beta_1^* = 0$ corresponds to Figure 2(b) are presented in Section 3.4 of the supplementary material.

4.4. Model Sensitivity Analysis

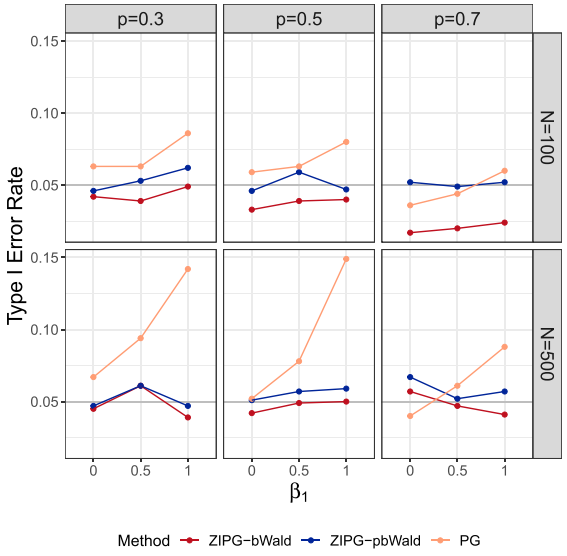
While ZIPG assumes that all differential variability comes from the Poisson-Gamma part (i.e., θ) and the true zero proportion p is only taxon-specific, one may wonder how the model fits when

Table 2. Summary for simulation settings.

H_0	β_0	β_1	β_2	β_0^*	β_1^*	p
Type I error settings						
$\beta_1 = 0$	−4.23	0	0.45	0.6	{0,0.5,1}	{0.3,0.5,0.7}
$\beta_1^* = 0$	−4.23	{0,0.5,1}	0.45	0.6	0	{0.3,0.5,0.7}
$\beta_2 = 0$	−4.23	1	0	0.6	{0,0.5,1}	{0.3,0.5,0.7}
Power settings						
$\beta_1 = 0$	−4.23	{0,0.2,…,1.8}	0.45	0.6	1	0.5
$\beta_1^* = 0$	−4.23	1	0.45	0.6	{0,0.2,…,1.8}	0.5
$\beta_2 = 0$	−4.23	1	{0,0.2,…,1.8}	0.6	1	0.5

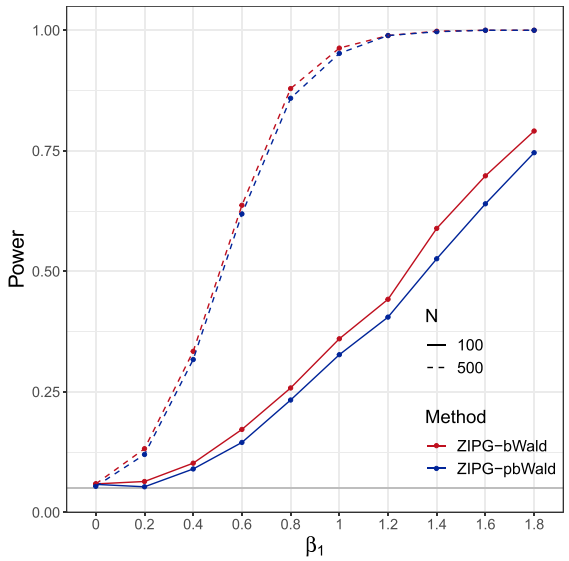


(a) Type I error of $H_0 : \beta_1 = 0$

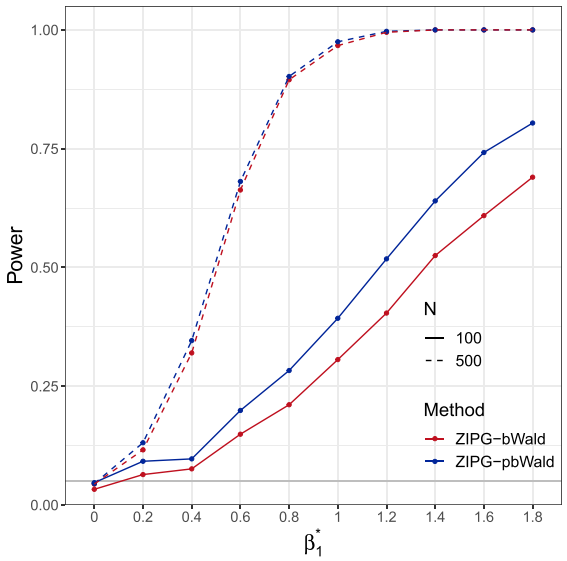


(b) Type I error of $H_0 : \beta_1^* = 0$

Figure 2. Type I error results for (a) β_1 and (b) β_1^*. The significance level is $\alpha = 0.05$.



(a) Power of rejecting $H_0 : \beta_1 = 0$



(b) Power of rejecting $H_0 : \beta_1^* = 0$

Figure 3. Power curves of rejecting null hypothesis with $N = 100$ (solid lines) and $N = 500$ (dash lines). With $p = 0.5$, the proportion of observed zeros decreased from 0.606 to 0.582 with the increase of β_1 in (a), while it increased from 0.536 to 0.653 with the increase of β_1^* in (b).

these assumptions are violated. Here, we evaluate ZIPG under the misspecified model, in which the true zero proportion p is also associated with covariates (denoted as ZIPG-full). Of note, ZIPG-full is equivalent to Omnibus (Chen et al. 2018) from a hypothesis testing perspective, whereas the Omnibus test does not provide point/interval estimation and hypothesis test for

Table 3. Average bias and its SE, average standard error, RMSE, and the empirical coverage rate (CR) of β_1 and β_1^* estimators.

	Method	avg-bias(SE)	avg-SE	RMSE	CR
		$N = 500, \beta_1 = 0, \beta_1^* = 1, p = 0.5, p_{obs} = 0.61$			
β_1	**ZIPG-bWald**	−0.013(0.254)	0.254	0.254	0.936
	pscl	0.216(0.252)	0.23	0.332	0.826
	NBZIMM	0.084(0.274)	0.235	0.286	0.877
	PG	−0.016(0.257)	0.307	0.258	0.977
β_1^*	**ZIPG-bWald**	−0.009(0.246)	0.256	0.246	0.954
	PG	−0.449(0.169)	0.165	0.48	0.231
γ	**ZIPG-bWald**	−0.004(0.145)	0.158	0.145	0.968
	pscl	0.108(0.152)	0.154	0.186	0.828
	NBZIMM	−0.245(0.454)	–	0.516	–
		$N = 500, \beta_1 = 0, \beta_1^* = 1, p = 0.7, p_{obs} = 0.77$			
β_1	**ZIPG-bWald**	−0.027(0.326)	0.341	0.327	0.952
	pscl	0.220(0.311)	0.299	0.381	0.874
	NBZIMM	−0.07(0.408)	0.363	0.414	0.917
	PG	−0.037(0.34)	0.432	0.342	0.988
β_1^*	**ZIPG-bWald**	0.006(0.341)	0.373	0.341	0.959
	PG	−0.536(0.21)	0.211	0.576	0.282
γ	**ZIPG-bWald**	0.007(0.158)	0.165	0.158	0.958
	pscl	0.047(0.872)	0.153	0.873	0.864
	NBZIMM	−0.294(0.604)	–	0.671	–

NOTE: Simulation parameters are set as $\beta_1 = 0, \beta_1^* = 1$ corresponding to Figure 2(a); $p = \{0.5, 0.7\}$ or $\gamma = \{0, 0.847\}$. NBZIMM does not report inference and CI on γ. The proportion of zeros observed in each simulation setting is denoted as p_{obs}.

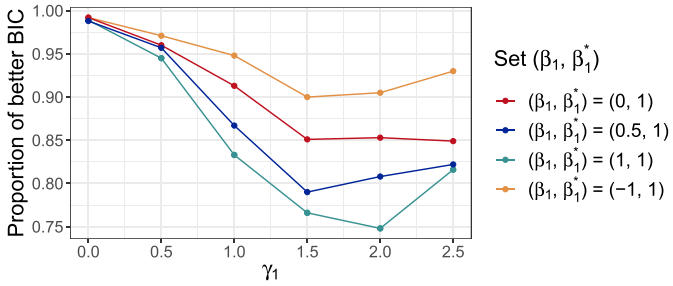


Figure 4. Proportion of ZIPG having smaller BIC than ZIPG-full in each simulation setting when $N = 500$.

each parameter separately. We consider the group covariates X_1 only and sample size $N = 500$. We use a logistic link $\text{logit}(p) = \gamma_0 + \gamma_1 X_1$ with $(\beta_0, \beta_0^*, \gamma_0) = (-4.23, 0.6, -0.847)$, and then we report the performance of ZIPG with γ_1 increasing from 0 to 2.5, which is equivalent to increasing p from 0.30 to 0.84 in the group with $X_1 = 1$.

For two model fittings, that is, ZIPG and ZIPG-full, we report the proportion of simulations suggesting better BIC from ZIPG (Figure 4). Over 1000 replicated simulations in each setting, ZIPG has a smaller BIC than ZIPG-full in most cases. Moreover, we also use Kolmogorov-Smirnov test (Kolmogorov 1933; Smirnoff 1939) to compare the ZIPG predicted distribution with the simulated observations: 100% replicates report insignificant differences between the two distributions ($p > 0.05$), suggesting that ZIPG-predicted distribution has no difference to the observed samples.

We also evaluate ZIPG's performance under other misspecified models: (a) Poisson-Gamma without zero inflation and (b) zero-inflated Beta-Binomial model (see Section 3.5 of the supplementary material). In both scenarios, ZIPG preserves nominal Type I error and retains its superior power in detecting differential abundance/variability, especially for large-sample cases.

4.5. Additional Simulation Results

We conduct additional simulations to compare multiple ways of constructing test statistics and confidence intervals. For hypothesis testing, we investigate different test statistics for the proposed ZIPG (Section 3.1 of the supplementary material), including the Wald test without bootstrap (ZIPG-Wald) and the likelihood ratio test (ZIPG-LRT). Simulation results suggest that ZIPG with bootstrap-based Wald tests (i.e., ZIPG-bWald and ZIPG-pbWald) are desired. For confidence intervals, we compare the coverage rate for both nonparametric and parametric bootstrap through different construction strategies, such as normality-based/quantile-based/BC_a(Efron and Tibshirani 1994) intervals. Results show that the normality-based confidence interval often has close-to-nominal coverage with a low computational cost (Section 3.2 of the supplementary material).

We further evaluate different resampling schemes for ZIPG-bWald, such as resampling based on measurements or subjects. Given the total sample size $N = 200$, results suggest no significant difference between the two strategies (Section 3.6 of the supplementary material).

Additional simulations about how measurement times affect ZIPG performance are provided in Section 3.6 of the supplementary material. We consider the following two scenarios: (a) when the numbers of measurements per subject are very small (i.e., $m = 2$) and (b) when the numbers of measurements per subject are unequal. Results show that the proposed ZIPG method is valid for both cases above.

5. Data Analysis

In this section, we analyze two microbiome datasets, Romero (Romero et al. 2014) and Dietary (Johnson et al. 2019), to investigate how physical conditions impact microbiome stability in specific taxa. The proposed ZIPG method with bootstrap-based Wald test is compared to pscl (Zeileis et al. 2008) and DESeq2 (Love et al. 2014) from two perspectives: identification

of taxa related to covariates and goodness of fit for prediction models. In addition, we also report hypothesis testing results from Omnibus (Chen et al. 2018) to validate the results of ZIPG. We also performed NBZIMM, but it detected only a few taxa with their default subject-level random effect, and hence we present the corresponding results in Section 4.3 of the supplementary material.

5.1. Data Description

Romero is a longitudinal case-control study including 16s rRNA gene sequence-based vaginal microbiota from 22 pregnant and 32 non-pregnant women with samples collected from each subject over intervals of weeks, resulting in 143 taxa and $N = 900$ longitudinal samples (139 measurements from pregnant women and 761 measurements from nonpregnant women). To investigate how taxa are impacted by pregnant status and age in this data, we set the covariates matrix $\boldsymbol{X}^* = \boldsymbol{X} = (X_1\ X_2\ X_3)$, where X_1 is a binary indicator of pregnant status, X_2 is the observational age, and X_3 is an indicator for the race (white or others) but not of our main interest. Note that both pregnant status and age are not changed for each person during data collection.

Dietary is diet-microbiome data with shotgun metagenomic sequencing results of fecal samples and daily dietary records of 34 subjects on 17 consecutive days. There are total $N = 475$ samples with both microbiome data and dietary records available. In this data, the main analysis of interest is how alcohol affects the microbiome variability. We created a binary indicator for 25 alcohol drinkers and 9 teetotalers, and included this variable in both $\boldsymbol{X}^*$ and $\boldsymbol{X}$. To account for the impact of other dietary intakes, we also include the first two principal components of the macronutrient matrix in $\boldsymbol{X}$.

In microbiome studies, it is common to filter out taxa with extremely low abundance (i.e., $p_{\text{obs}} > 0.9$) for more stable and reliable results (Wadsworth et al. 2017; Zhang and Yi 2020; Jiang et al. 2021). Further, taxa with $p_{\text{obs}} < 0.1$ are likely to have little or no zero inflation and can be modeled by other existing methods. Though ZIPG can still be performed with satisfactory results (see Section 3.5 of the supplementary material), this group of taxa is not of our main interest. For both Romero and Dietary data, we analyze the taxa with $0.1 < p_{\text{obs}} < 0.9$, which results in 25 taxa in Romero and 52 taxa in Dietary. More details about p_{obs} in Romero and Dietary can be found in Section 4.1 of the supplementary material. To account for multiple testing, we report the results with the controlled false discovery rate (FDR < 0.05) using the method of Benjamini and Hochberg (1995).

5.2. Results on Hypothesis Testing

We first present numbers of identified taxa regarding the covariates of interest in the two studies (Figure 5). For ZIPG, we show the set of taxa associated with $\boldsymbol{X}$ in the mean model (denoted as ZIPG β) and $\boldsymbol{X}^*$ in the dispersion model (denoted as ZIPG β^*), separately. In Romero, we observe that pregnant subjects are clustered under age 35, while nonpregnant subjects are collected in a much wider range of ages. Owing to the unbalanced sample that pregnant women have fewer measurements and are often younger than nonpregnant women, we use parametric bootstrap (i.e., ZIPG-pbWald) for more stable results. ZIPG identified 18 taxa with differential abundance ($H_0 : \beta_1 = 0$) and 17 taxa with differential variability ($H_0 : \beta_1^* = 0$) associated with pregnancy, while 13 taxa are associated with pregnant status with both abundance and variability (Figure 5(a)). Most of the taxa found by pscl and DESeq2 are also identified by ZIPG, while ZIPG also identified 3 additional taxa with significant differential variability, which are not detected by other methods. Further, ZIPG identified totally additionally 6 taxa associated to age with differential abundance ($H_0 : \beta_2 = 0$) or differential variability ($H_0 : \beta_2^* = 0$), compared to other methods (Figure 5(b)).

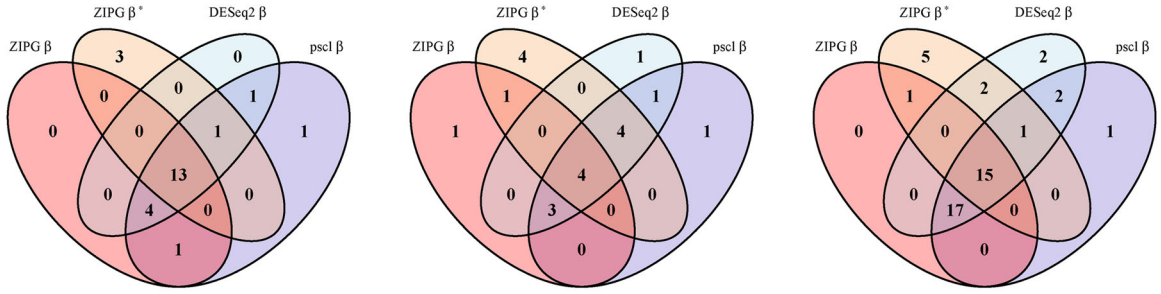
In Dietary, ZIPG is performed using nonparametric bootstrap Wald test (i.e., ZIPG-bWald), because the data is balanced and the sample size is sufficient. ZIPG identified 33 taxa with differential abundance ($H_0 : \beta_1 = 0$) and 24 taxa with differential variability ($H_0 : \beta_1^* = 0$) associated to the alcohol intake (Figure 5(c)). Compared to other methods, ZIPG discovered 5 extra taxa only associated with differential variability and 1 extra taxon associated with both differential abundance and differential variability.

We further use the Omnibus test (Chen et al. 2018) as verification for ZIPG detected taxa. The Omnibus test links covariates of interest to all three parameters in the negative-binomial distribution and rejects the null hypothesis if any covariate is associated with any of the parameters. Thus, taxa identified by both ZIPG and Omnibus are less likely to be false positives. As expected, in Romero, all taxa identified by ZIPG are also detected by Omnibus. In Dietary, 35 out of 41 taxa identified by ZIPG are also detected by Omnibus. Though the Omnibus test detected more taxa than ZIPG, it is worth pointing out that the Omnibus test cannot distinguish differential abundance and differential variability and does not provide point/interval estimation as ZIPG does. Details of taxa detected by each method and estimation results for those taxa regarding parameters of interest are shown in Section 4.3 of the supplementary material.

5.3. Analysis on Model Fitting

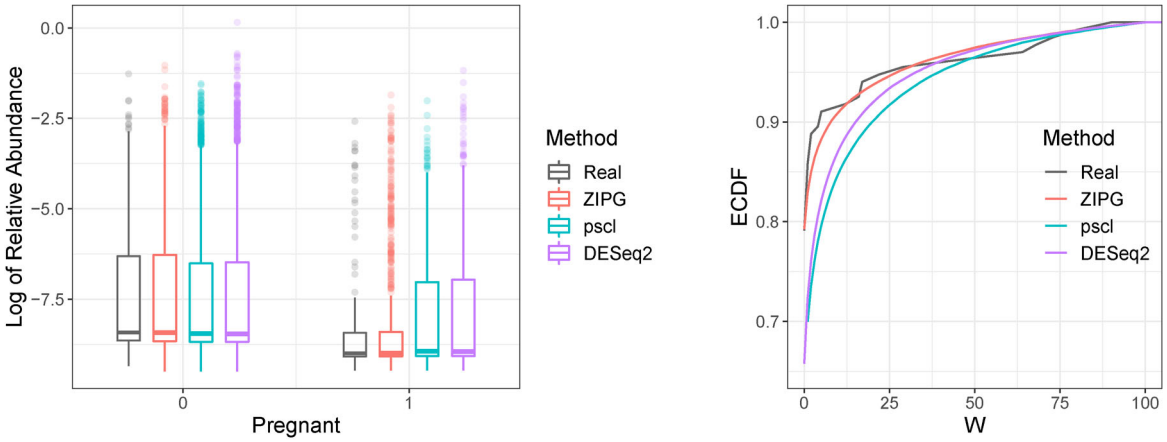
To visualize the differential variability tested by ZIPG (i.e., $H_0 : \beta^* = 0$), we further analyze the results of Bifidobacteriaceae and Lactobacillus.vaginalis from Romero, and Burkholderiales bacterium and Alistipes indistinctus from Dietary as examples. Other taxa with differential variability identified by ZIPG have similar conclusions.

First, we compare the goodness of fit of results from fitted models to the empirical distribution of the relative abundance. The log of relative abundance observed in real data (i.e., Bifidobacteriaceae from Romero) is compared to the predicted distribution according to the estimated model by ZIPG, pscl, and DESeq2 (Figure 6). For boxplots, we generated samples from each predicted distribution with five times the observed sample sizes for a better visualization at the tail (e.g., Figure 6(a)). All three methods can estimate the median of two groups (pregnant and nonpregnant) accurately, but pscl and DESeq2 cannot distinguish the overdispersion between the two groups, as the box length for the pregnant and nonpregnant groups are similar. On the contrary, ZIPG identified the differential variability of this taxon associated with the factor pregnant with $p = 0.00256$ regarding the null hypothesis $H_0 : \beta_1^* = 0$, and thus its



(a) Romero Pregnant (b) Romero Age (c) Dietary ALC

Figure 5. The numbers of taxa with significant difference detected by ZIPG, DESeq2 and pscl after controlling FDR < 0.05 regarding $H_0 : \beta = 0$ and $H_0 : \beta^* = 0$ in Romero (a and b) and Dietary (c).



(a) Box plot of predicted distribution (b) ECDF (Pregnant)

Figure 6. Bifidobacteriaceae in Romero: (a) boxplot for the predicted distribution and the real observed counts (the log of relative abundance is presented with zero count samples adjusted to 0.5 in the pregnant and nonpregnant group), (b) ECDF in the pregnant group.

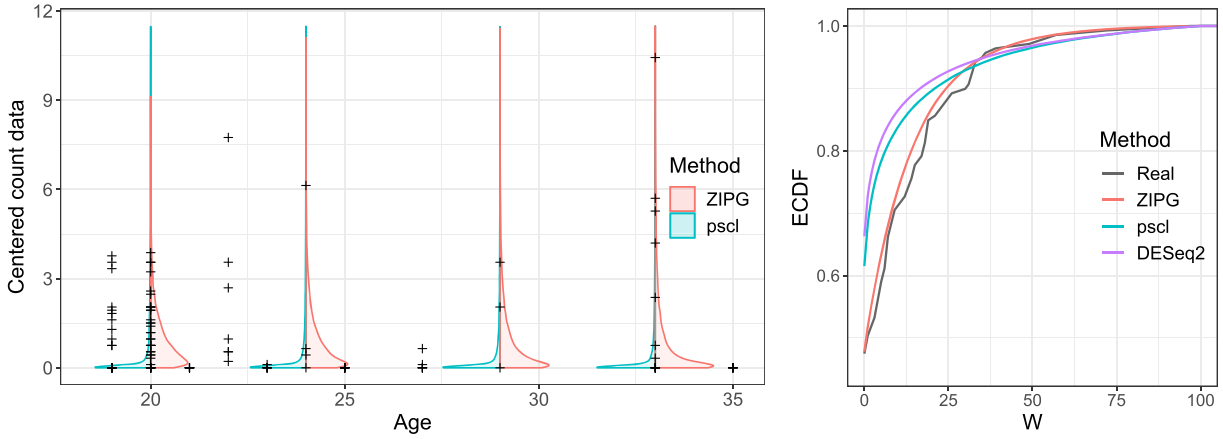
simulated data matches the real data better, providing a shorter interquartile range with a long tail in the pregnant group. We also present the empirical cumulative distribution functions (ECDF) of the log of the relative abundance for the pregnant group, using sampled data from fitted ZIPG, pscl, and DESeq2 models (Figure 6(b)). It has been shown that ZIPG also fits the real data better than other methods. Quantile-quantile plots for real data versus predicted distribution also show that ZIPG models the entire distribution better (Section 4.2 of the supplementary material).

For Lactobacillus.vaginalis in Romero, we present the relative abundance based on the simulated distributions of the fluctuation factor $U \sim \text{Gamma}(\theta_i^{-1}, \theta_i)$ from ZIPG and pscl at four representative ages (Figure 7). ZIPG identified that the differential variability of this taxon is associated to age with $p < 0.001$ regarding $H_0 : \beta_2^* = 0$. Accordingly, we observe that the shape of the fitted distribution changes with the increase of age. However, pscl is not able to model the change in the entire shape of distribution as the average relative abundance remained the same in this group. The ECDF plot of the pregnant group again shows that ZIPG fitted model is closer to the empirical distribution.

In Dietary, we present the boxplots for Burkholderiales bacterium and Alistipes indistinctus to show the differential variability between the alcohol and nonalcohol drinkers (Figure 8). For Burkholderiales bacterium, the differential abundance in the two groups (i.e., alcohol and nonalcohol drinkers) are similar, but the overdispersion in alcohol drinkers is obviously larger than that in teetotal subjects. ZIPG can distinguish the differential variability in two groups with $p = 0.00293$ ($H_0 : \beta_1^* = 0$). ZIPG also provides a shorter interquartile in the nonalcohol drinker group, which is consistent with the raw data. On the contrary, pscl failed to detect any differential abundance ($p = 0.636$), while DESeq2 provided $p = 0.049$ which is significant but much larger than the p-value of ZIPG. For Alistipes indistinctus, though pscl and DESeq2 identified the differential abundance between two groups, both of them did not approximate the data from a distributional perspective because of the ignorance of differential variability. In contrast, ZIPG can distinguish it with p = 8.54e−6 ($H_0 : \beta_1^* = 0$), and provide a long box for the non-alcohol group showing their small overdispersion and a median more closer to real data.

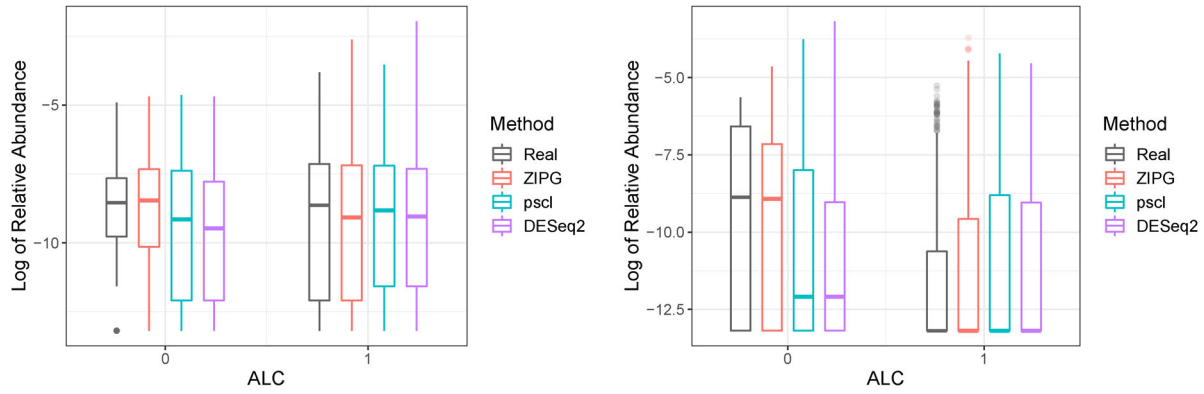
6. Discussion

In this article, we propose a Zero-Inflated Poisson-Gamma model for microbiome count data analysis. We decompose the zero-inflated Poisson model and factor the Poisson mean as λ_{ijk}



(a) Predicted distribution (Pregnant, White) (b) ECDF (Pregnant)

Figure 7. Lactobacillus vaginalis in Romero: (a) the half-violin curves of the fluctuation factor U for the pregnant, white women generating from parameters estimated by ZIPG and pscl, respectively, comparing to the raw count data divided by its mean ("+"). (b) ECDF of predicted and real observed counts in the pregnant group.



(a) Burkholderiales bacterium (b) Alistipes indistinctus

Figure 8. Boxplot for the predicted distribution and the real observed counts. We plot the log of relative abundance with zero count samples adjusted to 0.5 in the alcohol and nonalcohol groups for (a) Burkholderiales bacterium and (b) Alistipes indistinctus in Dietary.

for the average abundance level and a multiplicative factor U_{ijk} following gamma distribution controlled by variation parameter θ_{ik}, which accounts for individual-level microbiome abundance variation around λ_{ijk}. In traditional ZINB regression, the dispersion parameter is often treated as a nuisance parameter. Our model allows different sets of covariates to be linked to λ_{ijk} and θ_{ik} and provides a valid test, outperforming other negative-binomial-based models such as pscl and NBZIMM. To our knowledge, the ZINB-based Omnibus method (Chen et al. 2018) may be one of few papers that links dispersion to covariates. However, the Omnibus test cannot distinguish differential abundance and differential variability. In comparison, we test differential abundance and differential variability separately for longitudinal data and provide valid confidence intervals for each parameter. Moreover, other potential distributions for modeling the multiplicative factor U_{ijk} are worth future exploring, including the mixing distribution of log-normals. However, the difficulty in distinguishing two sources of zeros always exists when the overdispersion is large.

Though linking γ_k to covariates is proposed in pscl and NBZIMM, it is not suggested in our ZIPG model based on two reasons. First, the mechanism of zero-inflation γ_k does not have explicit biological interpretation, while θ_{ik} can be explained as individual-level microbiome stability in longitudinal data. Through simulations, we have shown that linking γ_k to covariates is not preferred from the model selection perspective, even if both θ_{ik} and γ_k are covariates-dependent. Second, the increment in either γ_k or θ_{ik} will lead to the increment of zeros in observed data, and thus linking both parameters to covariates simultaneously will make the inference more challenging and unreliable.

Some promising future work could be incorporating auxiliary information from other taxa. One possible way is to assume the taxon-specific dispersion parameters θ_{ik}'s of closely related taxa (e.g., taxa in the same phylogenetic branch) are impacted by covariates $\boldsymbol{X}_i^*$ identically and share the same coefficient $\boldsymbol{\beta}^*$. In addition, inference on a group of taxa in a joint multivariate measurement error model is also worth future investigation.

Supplementary Materials

Supplements: The supplemental materials (ZIPG-appendix.pdf) include mathematical details of the non-concavity of the log-likelihood, analytical expressions of gradient, details of the parametric bootstrap algorithm, and supplementary figures and tables for additional simulation and real-world data analysis.

R code for ZIPG: The ZIPG method is implemented in R and available on GitHub (*https://github.com/roulan2000/ZIPG*).

Acknowledgments

We thank the editor, the associate editor, and two anonymous reviewers for their helpful comments and suggestions.

Data Availability Statement

All datasets we used are published online. *Dietary* data is available at *https://github.com/knights-lab/dietstudy_analyses*. *Romero* data is available in their paper (DOI:10.1186/2049-2618-2-4) or directly from the R package NBZIMM.

Disclosure Statement

The authors report there are no competing interests to declare.

ORCID

Roulan Jiang https://orcid.org/0000-0003-4156-8447
Xiang Zhan https://orcid.org/0000-0001-9650-143X
Tianying Wang https://orcid.org/0000-0002-2826-5364

References

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Series B, 57, 289–300. [9]

Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Machler, M., and Bolker, B. M. (2017), "glmmtmb Balances Speed and Flexibility among Packages for Zero-Inflated Generalized Linear Mixed Modeling," *The R Journal*, 9, 378–400. [5]

Broyden, C. G. (1970), "The Convergence of a Class of Double-Rank Minimization Algorithms 1. General Considerations," *IMA Journal of Applied Mathematics*, 6, 76–90. [4]

Cao, Y., Zhang, A., and Li, H. (2020), "Multisample Estimation of Bacterial Composition Matrices in Metagenomics Data," *Biometrika*, 107, 75–92. [2]

Chen, J., King, E., Deek, R., Wei, Z., Yu, Y., Grill, D., and Ballman, K. (2018), "An Omnibus Test for Differential Distribution Analysis of Microbiome Sequencing Data," *Bioinformatics*, 34, 643–651. [7,9,11]

Couch, C. E., Stagaman, K., Spaan, R. S., Combrink, H. J., Sharpton, T. J., Beechler, B. R., and Jolles, A. E. (2021), "Diet and Gut Microbiome Enterotype are Associated at the Population Level in African Buffalo," *Nature Communications*, 12, 1–11. [3]

Efron, B., and Tibshirani, R. J. (1994), *An Introduction to the Bootstrap*, Boca Raton, FL: CRC Press. [5,8]

Fletcher, R. (1970), "A New Approach to Variable Metric Algorithms," *The Computer Journal*, 13, 317–322. [4]

Goldfarb, D. (1970), "A Family of Variable-Metric Methods Derived by Variational Means," *Mathematics of Computation*, 24, 23–26. [4]

Hu, Y.-J., Lane, A., and Satten, G. A. (2021), "A Rarefaction-Based Extension of the LDM for Testing Presence–Absence Associations in the Microbiome," *Bioinformatics*, 37, 1652–1657. [1]

Jiang, S., Xiao, G., Koh, A. Y., Kim, J., Li, Q., and Zhan, X. (2021), "A Bayesian Zero-inflated Negative Binomial Regression Model for the Integrative Analysis of Microbiome Data," *Biostatistics*, 22, 522–540. [9]

Johnson, A. J., Vangay, P., Al-Ghalith, G. A., Hillmann, B. M., Ward, T. L., Shields-Cutler, R. R., Kim, A. D., Shmagel, A. K., Syed, A. N., Personalized Microbiome Class Students, Walter, J., Menon, R., Koecher, K., and Knights, D. (2019), "Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans," *Cell Host & Microbe*, 25, 789–802. [2,8]

Kolmogorov, A. (1933), "Sulla Determinazione Empirica di una lgge di Distribuzione," *Inst. Ital. Attuari, Giorn.*, 4, 83–91. [8]

La Rosa, P. S., Brooks, J. P., Deych, E., Boone, E. L., Edwards, D. J., Wang, Q., Sodergren, E., Weinstock, G., and Shannon, W. D. (2012), "Hypothesis Testing and Power Calculations for Taxonomic-based Human Microbiome Data," *PloS One*, 7, e52078. [3]

Li, H. (2015), "Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis," *Annual Review of Statistics and Its Application*, 2, 73–94. [1]

Li, Z., Lee, K., Karagas, M. R., Madan, J. C., Hoen, A. G., O'malley, A. J., and Li, H. (2018), "Conditional Regression based on a Multivariate Zero-Inflated Logistic-Normal Model for Microbiome Relative Abundance Data," *Statistics in Biosciences*, 10, 587–608. [2]

Li, Z., Tian, L., O'Malley, A. J., Karagas, M. R., Hoen, A. G., Christensen, B. C., Madan, J. C., Wu, Q., Gharaibeh, R. Z., Jobin, C., and Li, H. (2021), "IFAA: Robust Association Identification and Inference for Absolute Abundance in Microbiome Analyses," *Journal of the American Statistical Association*, 116, 1595–1608. [1]

Love, M. I., Huber, W., and Anders, S. (2014), "Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2," *Genome Biology*, 15, 550–570. [2,4,8]

Manor, O., Dai, C. L., Kornilov, S. A., Smith, B., Price, N. D., Lovejoy, J. C., Gibbons, S. M., and Magis, A. T. (2020), "Health and Disease Markers Correlate with Gut Microbiome Composition Across Thousands of People," *Nature Communications*, 11, 1–12. [1]

Martin, B. D., Witten, D., and Willis, A. D. (2020), "Modeling Microbial Abundances and Dysbiosis with Beta-Binomial Regression," *The Annals of Applied Statistics*, 14, 94–115. [1,3,5]

Martın-Fernandez, J. A., Palarea-Albaladejo, J., and Olea, R. A. (2011), "Dealing with Zeros," in *Compositional Data Analysis: Theory and Applications*, eds. V. Pawlowsky-Glahn and A. Buccianti, pp. 43–58, Chichester: Wiley. [2]

McLaren, M. R., Willis, A. D., and Callahan, B. J. (2019), "Consistent and Correctable Bias in Metagenomic Sequencing Experiments," *Elife*, 8, e46923. [3]

McMurdie, P. J., and Holmes, S. (2014), "Waste Not, Want Not: Why Rarefying Microbiome Data is Inadmissible," *PloS Computational Biology*, 10, e1003531. [1]

Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., Reyes, J. A., Shah, S. A., LeLeiko, N., Snapper, S. B., Bousvaros, A., Korzenik, J., Sands, B. E., Xavier, R. J., and Huttenhower, C. (2012), "Dysfunction of the Intestinal Microbiome in Inflammatory Bowel Disease and Treatment," *Genome Biology*, 13, 1–18. [3]

Petersen, C., and Round, J. L. (2014), "Defining Dysbiosis and its Influence on Host Immunity and Disease," *Cellular Microbiology*, 16, 1024–1033. [2]

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010), "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data," *Bioinformatics*, 26, 139–140. [2,4]

Romero, R., Hassan, S. S., Gajer, P., Tarca, A. L., Fadrosh, D. W., Nikita, L., Galuppi, M., Lamont, R. F., Chaemsaithong, P., Miranda, J., Chaiworapongsa, T., and Ravel, J. (2014), "The Composition and Stability of the Vaginal Microbiota of Normal Pregnant Women is Different from that of Non-pregnant Women," *Microbiome*, 2, 1–19. [6,8]

Shanno, D. F. (1970), "Conditioning of Quasi-Newton Methods for Function Minimization," *Mathematics of Computation*, 24, 647–656. [4]

Shi, P., Zhou, Y., and Zhang, A. R. (2022), "High-Dimensional Log-Error-in-Variable Regression with Applications to Microbial Compositional Data Analysis," *Biometrika*, 109, 405–420. [1]

Silverman, J. D., Roche, K., Mukherjee, S., and David, L. A. (2020), "Naught All Zeros in Sequence Count Data are the Same," *Computational and Structural Biotechnology Journal*, 18, 2789–2798. [1,2,3]

Smirnoff, N. (1939), "Sur les écarts de la courbe de distribution empirique," *Matematicheskii Sbornik*, 48, 3–26. [8]

Tang, Z.-Z., and Chen, G. (2019), "Zero-Inflated Generalized Dirichlet Multinomial Regression Model for Microbiome Compositional Data Analysis," *Biostatistics*, 20, 698–713. [2]

Vandeputte, D., Kathagen, G., D'hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., Wang, J., Tito, R. Y., De Commer, L., Darzi, Y., Vermeire, S., Falony, G., and Raes, J. (2017), "Quantitative Microbiome Profiling

Links Gut Community Variation to Microbial Load," *Nature*, 551, 507–511. [1]

Wadsworth, W. D., Argiento, R., Guindani, M., Galloway-Pena, J., Shelburne, S. A., and Vannucci, M. (2017), "An Integrative Bayesian Dirichlet-Multinomial Regression Model for the Analysis of Taxonomic Abundances in Microbiome Data," *BMC Bioinformatics*, 18, 94–105. [9]

Willis, A. D. (2019), "Rigorous Statistical Methods for Rigorous Microbiome Science," *MSystems*, 4, e00117-19. [1]

Wu, C. J. (1983), "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics*, 11, 95–103. [5]

Xu, T., Demmer, R. T., and Li, G. (2021), "Zero-Inflated Poisson Factor Model with Application to Microbiome Read Counts," *Biometrics*, 77, 91–101. [1,2,4]

Zeileis, A., Kleiber, C., and Jackman, S. (2008), "Regression Models for Count Data in R," *Journal of Statistical Software*, 27. [2,5,8]

Zeng, Y., Pang, D., Zhao, H., and Wang, T. (2022), "A Zero-Inflated Logistic Normal Multinomial Model for Extracting Microbial Compositions," *Journal of the American Statistical Association*, 1–14. [2]

Zhang, X., and Yi, N. (2020), "Fast Zero-Inflated Negative Binomial Mixed Modeling Approach for Analyzing Longitudinal Metagenomics Data," *Bioinformatics*, 36, 2345–2351. [2,4,5,9]