A novel statistical method for handling zeros in microbiome data

Juxin Liu [1] Huokai Wu [1] Mingyang (Ming) Chen [1] Peixuan (Ashleigh) Chen [2] Ang (David) Li [1]

[1]University of Saskatchewan [2]University of Victoria

Pacific Institute for the Mathematical Sciences

VXML

Abstract

Microbiome data are extremely important for understanding the composition of microbial communities, leading to impactful applications in studying diseases in both humans and plants. Recently, 16S rRNA gene sequencing has become a popular method to perform microbial community studies. Despite the popularity of this method, there are challenges in analyzing 16S rRNA gene sequence data. Although a large body of literature develops different data-generating models, more work must be done on developing a 16S count data generator. Our project proposes a novel data generator that makes use of phylogenetic tree distances. This generator is applied to compare the four recent methods via a comprehensive simulation study.

Introduction

Challenges in Analyzing 16S rRNA Gene Sequence Data

Despite the widespread use of 16S rRNA gene sequencing data, a key tool in microbiome research[1]. In this project, challenging issues are present due to the limitations of the sequencing technique: composition, high dimensionality, and an excess amount of zeros (zero-inflation), and we want to address the problems by creating a novel data generator. We propose a data generator that considers phylogenetic tree distances, mimics real-world data, and allows advanced statistical methods to be systematically compared to one another to gain a better understanding of the complicated microbial data.

Phylogenetic Tree

Phylogenetic trees reflect ancestral information between species, which eventually leads to a common ancestor[2]. These trees are based on genetic information, with each stem representing a unique yet distinct evolutionary path[3]. A plausible assumption is that the abundances of two closely related species (taxa in OTU data) will be similar.

Recent Four Method For Handling zero

- **Bayesian-Multiplicative replacement method**
  Martín-Fernández et al.[4] proposed a Bayesian-multiplicative treatment in 2015. This method treat all zero as sampling zeros that means non of zeros are biological zeros,and this technique do not using the information of phylogenetic tree. By using Bayesian-multiplicative method and mutltinomail distribution to adjustment of non-zero values.
- **MbImpute**
  The mbImpute method is proposed by Jiang et al [5] in 2021. Firstly, mbImpute distinguish all zero values into two types that are biological zero and sampling zero. Second, by borrorrowing information from similar taxa and using information from phylogenetic tree to imputed all sampling zero values in dataset.
- **Zero inflated probabilistic PCA model**
  The ZIPPCA method was proposed by Zeng[6] et al in 2022, the ZIPPCA method processes multivariate abundance data directly. Instead of converting raw abundance data into compositions or relative abundances, this method employs an empirical Bayes approach to infer microbial compositions
- **Zero inflated Dirichlet tree multinomial model**
  Zhou et al[3]. proposed a Zero-Inflated Dirichlet Tree Multinomial distribution to handle zero-inflated microbiome data in 2021. ZIDTM is adept at handling data sparsity, compositionality, and high dimensionality by setting a multivariate distribution. A notable challenge introduced is the multiple possible orderings of child nodes for each taxa. To address this, different ZIDM models are applied to every possible outcome to identify the best fitting model.

Simulation Data Generation

Notations:

- N: Number of the sampling size, N = 98
- K: Number of the taxon, K = 62
- i: The i-th sample, i = 1, 2,..., N
- j: The j-th taxon, j = 1, 2,..., K
- π_{ij}: Relative abundance for taxa j in sample i
- $N_i = \sum_{j=1}^{K} Y_{ij}$: Total OTU counts
- $\delta_j = \begin{cases} 1, & \text{if taxon } j \text{ does not exist.} \\ 0, & \text{otherwise.} \end{cases}$

Parameters in Simulation:

In the simulation study, there are three different proportions of biological zero scenarios. The α, ρ, σ, and θ are the four parameters for generating true abundance. The first scenario α is 5, the second is 15, and the third one is 50. The mean vector of Multinomial distribution, θ follows normal distribution with vary mean are 1 and 4, and standard deviation is 1; the evolutionary rate, ρ is 1; the σ is 2

$$P(\delta_j = 1) = 1 - exp(-d_{jK}/\alpha), j = 1, ..., K$$

$$\delta_{ij} \sim Bernoulli(p_j)$$

$$\text{if } \delta_j = 1 : \pi_{ij} = 0$$
$$\text{if } \delta_j = 0 : U_i \sim MVN(\theta_i, \Sigma)$$

$$\pi_{ij_1} = \frac{exp(U_{ij_1})}{1 + \sum_{l=1}^{L-1} exp(U_{ij_l})}$$
$$\vdots$$
$$\pi_{ij_{L-1}} = \frac{exp(U_{ij_{L-1}})}{1 + \sum_{l=1}^{L-1} exp(U_{ij_l})}$$
$$\pi_{iK} = \frac{1}{1 + \sum_{l=1}^{L-1} exp(U_{ij_l})}$$
$$Y_i|\pi_i \sim Multinomial(\pi_i, N_i)$$

Note: The variance-covariance matrix $\Sigma_{lm} = \sigma^2 \exp\{-2\rho_{lm} D_{lm}\}$ [7], and the mean θ is for different taxa in Multinomial distribution. By the variance-covariance, the phylogenetic tree distances denotes as D_{lm} between taxon l and m, the evolutionary rate is ρ.

Results Of The Four Method Comparison

alpha	zComp-SQ	zCom-GBM	mbImpute	phyloMDA-MIX	ZIPPCAlnm
5	37.98	0.85	0.034	0.033	27.5
10	125.21	2.66	0.07	0.082	44.14
50	131.99	3.13	0.09	0.098	41.85

Table 1. Running Time Comparison For 200 Replication

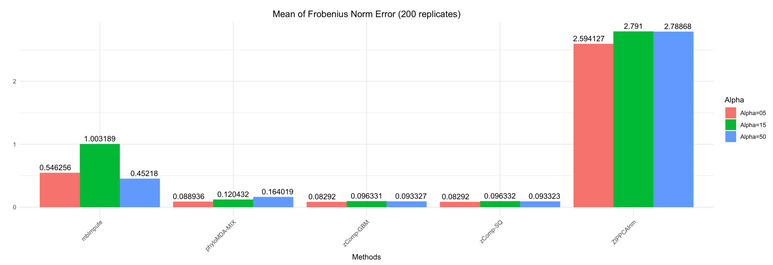

Figure 1. Mean of Frobenius Norm Error

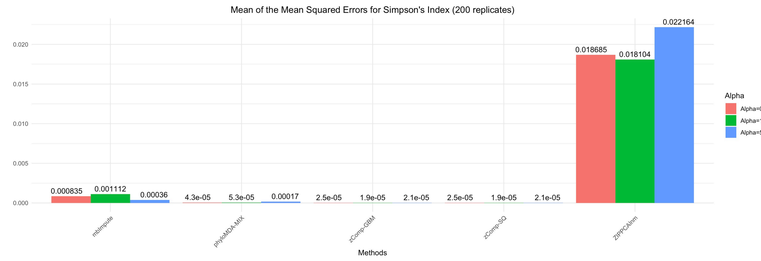

Figure 2. Mean of the Mean SQ Error of Simpson's Index

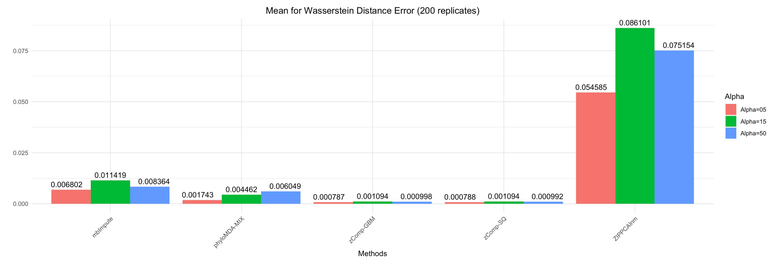

Figure 3. Mean for Wasserstein Distance Error

In our data generation section, we adjust the value of alpha to 5, 15, and 50, which correspondingly generates proportions of biological zeros at approximately 85%, 35%, and 20% in the data. To quantify the difference between the estimated abundance ratios and the true abundance ratios, the following metrics are used: Frobenius Norm Error, Mean Squared Error of Simpson's Index, and Wasserstein Distance Error, primarily focusing on the differences between estimated and true abundances. Overall, the methods Z-composition and Phylo-MDA perform the best in all three scenarios, while ZIPPCA performs the worst compared to the others. The performance of Mbimpute lies between the best and the worst methods.

References

[1] D.-G. C. Yinglin Xia, Jun Sun, *Statistical Analysis of Microbiome Data with R*. Springer Singapore, 2018, vol. 847.

[2] S. D. S. David A. Baum, "Tree thinking," *An Introduction to Phylogenetic Biology. Roberts and Company Publishers*, 2013.

[3] C. Zhou, H. Zhao, and T. Wang, "Transformation and differential abundance analysis of microbiome data incorporating phylogeny," *Bioinformatics*, vol. 37, no. 24, pp. 4652–4660, 2021.

[4] M. T. P. F. J. P.-A. Josep-Antoni Mart´in Fern´andez, Karel Hron, *Bayesian-multiplicative treatment of count zeros in compositional data sets*, 2015.

[5] R. Jiang, W. V. Li, and J. J. Li, "mbimpute: an accurate and robust imputation method for microbiome data," *Genome biology*, vol. 22, no. 1, p. 192, 2021.

[6] Y. Zeng, D. Pang, H. Zhao, and T. Wang, "A zero-inflated logistic normal multinomial model for extracting microbial compositions," *Journal of the American Statistical Association*, vol. 118, no. 544, pp. 2356–2369, 2023.

[7] S. J. Y. Y. X. Z. Jian Xiao, Li Chen and J. Chen, "Predictive modeling of microbiome data using a phylogeny-regularized generalized linear mixed model," *Frontiers in microbiology*.

[8] S. Mandal, W. Van Treuren, R. A. White, M. Eggesbø, R. Knight, and S. D. Peddada, "Analysis of composition of microbiomes: a novel method for studying microbial composition," *Microbial ecology in health and disease*, vol. 26, no. 1, p. 27663, 2015.