PACIFIC INSTITUTE FOR THE MATHEMATICAL SCIENCES VIRTUAL EXPERIMENTAL MATHEMATICS LAB (PIMS VXML) FINAL REPORT: A NOVEL STATISTICAL METHOD FOR HANDLING ZEROS IN MICROBIOME DATA

FACULTY MENTORS: JUXIN LIU [1]
GRADUATE TUTOR: HUOKAI WU [1]
MEMBERS: MINGYANG CHEN [1]; PEIXUAN CHEN[2]; ANG LI[1]

1. Introduction

1.1. Challenges in Analyzing 16S rRNA Gene Sequence Data.

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 considerable difficulties with analyzing 16S rRNA gene sequence data. Firstly, the number and diversity of species and genes present require concentrating on what is important. Moreover, the data shows the relative abundance of each species[1]. Secondly, a large portion of zeros in these datasets poses challenges for data analysis[2]. Lastly, a lack of a universally recognized way to handle 16S rRNA data prevents meaningful comparisons[3, 4].

In this project, we develop a novel data generator that makes use of phylogenetic tree distance information. Systematic simulation studies are conducted using this proposed data generator to compare four advanced methods for handling zero-inflation microbiome data.

1.2. Operational Taxonomic Units (OTUs).

Our project focuses on Operational Taxonomic Units (OTUs) counts. OTUs are groups of closely related 16S-rRNA gene sequences. These sequences resemble fingerprints for several bacteria. Researchers utilize 16S rRNA sequencing to investigate bacteria[5]. This approach generates a large amount of data, with each sequence representing a single

[1]University of Saskatchewan
[2]University of Victoria
Date: 2023-24.

bacterium. Grouping comparable sequences into OTUs facilitates the management of a large number of species. We can assume that identical 16S rRNA sequences indicate that the microbes are related. However, the definitions of OTUs can change. A modest difference in similarity level can result in distinct groupings of sequences.

1.3. **Phylogenetic Tree.** phylogenetic trees represent ancestor information between the species that finally leads to one joint ancestor[6]. These trees were based on genetic information, where each stem represents a different but separate evolutionary path[7]. The evolutionary past is assessed with the help of sequence similarities and differences, using sequence distance, parsimony, and maximum likelihood[8].

2. Recent Four Methods for Handling Zeros

Due to the challenges of analyzing microbiome data collected via 16S rRNA sequencing, researchers have developed various tools.[9]In this section, four recent methods for handling zeros will be introduced.

2.1. **Bayesian-Multiplicative Replacement Method.**

Martín-Fernández et al. proposed a Bayesian-multiplicative treatment in 2015.[10] This method primarily addresses the challenge of compositional data in microbiome datasets collected via 16S rRNA sequencing. It involves creating a compositional vector of counts that includes some zero values, and then using a posterior Bayesian estimate to replace each zero with an expected value. This approach utilizes information from the total zero counts and prior data to derive these Bayesian estimates. However, it does not incorporate information from the phylogenetic tree.

2.2. **MBImpute.**

The mbImpute method is the first imputation method for microbiome data proposed by Jiang et al[11]. In 2021.This method distinguishes between biological zeros, sampling zeros, and technical zeros in microbiome data. Biological zeros represent the true absence of taxa in microbiome samples. In contrast, technical and sampling zeros are not truly absent; they may not appear due to limited sample size or technical measurement errors. The MbImpute method does not assume that all zeros are truly absent. Instead, it uses information from the phylogenetic tree and sample covariates to impute zeros to non-zero values. This imputation aims to achieve a 'True' microbiome dataset,

where 'True' indicates that the data accurately reflect the actual sample situation, with only biological zeros remaining.

2.3. **Zero inflated probabilistic PCA model.**

Zeng et al. introduced a novel approach termed zero-inflated probabilistic PCA (ZIPPCA) in 2022[12].Many samples contain numerous zeros, indicating the non-existence of certain taxa. This method defines a Bernoulli random variable for each taxa, representing the probability that a taxa truly exists. By counting the total number of all existing taxa in every sample, we can calculate the true abundance. Using correlations across taxa and the phylogenetic tree, 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.

2.4. **Zero inflated Dirichlet tree multinomial model.**

Zhou et al. proposed a Zero-Inflated Dirichlet Tree Multinomial distribution to handle zero-inflated micro biome data in 2021[13].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. This step requires using the phylogenetic tree to determine the child taxa for each taxa. An empirical Bayes approach is then employed to transform counts into non-zero relative abundances, which can enhance the quality of the posterior mean transformation.

3. A novel data generator

The proposed data generative model aims to identify different sources of zero values based on phylogenetic tree distance. The rationale for introducing the phylogenetic tree distance is closely related species are expected to have similar abundance rates.

Before introducing the data generated model, firstly we will explain the following notations:

3.1. **Notation.**

- Let $j \in (1, 2, 3, ..., K)$ be the index of taxa and let the $i \in (1, 2, 3, ..., N)$ be the index of sample.

- OTU count matrix, denoted by $\mathbf{Y} = (Y_{ij}), i = 1, ..., N; j = 1, ..., K$.
- π_{ij} indicates the relative abundance of taxon j in sample i.
- Distance matrix is $D = (d_{jk}), j, k = 1, 2, ..., K$, where d_{jk} is phylogenetic tree distance between taxa j and k.

3.2. **Model.**

First, we define the latent parameter space δ_j to signify the presence of taxa. If $\delta_j = 1$, then taxa j is absent; otherwise, if $\delta_j = 0$, then taxa j is present. We utilize the Zero-Inflated Logistic Normal Multinomial Model (ZILNM model) [14][15][16].
Our simulation assumes a constant existence probability for each taxon, utilizing the phylogenetic tree distance. We denote P once $\delta_j = 1$, with α being parameter control probability, and π represents the non-zero true abundance.

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

where d_{jK} define as the phylogenetic tree distance between taxon j and a reference taxon K, where the label K must correspond to an existing reference taxon. Let $j_1, \cdots, j_L$ denote the indices of those taxa that exist. Obviously $j_L = K$. Let

$$U_{ij_l} = \log(\pi_{ij_l}/\pi_{iK}),\ l = 1, \ldots, L-1$$

Let $U_i \sim MVN(\theta_i, \Sigma)$, where entries of variance-covariance matrix Σ_{lm} produces by $\sigma^2 \exp\{-2\rho_{lm} D_{lm}\}$ [17].

The parameter σ^2 represents the variance component, while $\rho_{lm} \in (0, \infty)$ indicates the evolutionary rate between taxon l and taxon m. A ρ_{lm} approaching ∞ suggests rapid evolution between these taxa.[18] $\mathbf{N}_i$ represents the total number of the OTU counts.

One can easily derive the one-to-one mapping between U_i and π. That is,

$$\begin{aligned}
\pi_{ij_1} &= \frac{exp(U_{ij_1})}{1 + \sum_{l=1}^{L-1} exp(U_{ij_l})} \\
&\vdots \\
\pi_{ij_{L-1}} &= \frac{exp(U_{ij_{L-1}})}{1 + \sum_{l=1}^{L-1} exp(U_{ij_l})} \\
\pi_{iK} &= \frac{1}{1 + \sum_{l=1}^{L-1} exp(U_{ij_l})}
\end{aligned}$$

Then the observed OTU counts

$$\mathbf{Y}_i|\pi_i \sim Multinomial(\pi_i, N_i).$$

3.3. **Processing for Simulation Data.**

In this simulation study, we utilized a real-world designed to investigate the correlations between dietary variables and gut microbiota, provided by Wu et al. This experiment was conducted on a study group of 98 healthy volunteers, referred to as "COMBO" in the original dataset. [19] Following management and processing by Liu et al., the original data employed 16S ribosomal DNA sequence data to compute distances between microbial communities. [20] The "COMBO" dataset with 98 samples and 62 taxa, with corresponding phylogentic tree. "COMBO" dataset as the reference phylogenetic tree in simulation study, the phylogenetic relation between the taxa will not varied across different samples.
The produce for the simulation study is shown below:

(1) In a simulation study of data, it is necessary to specify the sample size and the total number of taxa. Here, the sample size, denoted by N, is fixed at 98, and the total number of taxa, denoted by K, is set to 62.

(2) We utilized the 'cophenetic.phylo' function to compute pairwise distances between pairs of branch lengths in the phylogenetic tree [21]. In this simulation study, we defined phylogenetic distance as the number of edges linking two taxa, with each edge length representing a branch length in the phylogenetic tree.

(3) Different alpha values have varying probabilities of being biologically zero. In our simulation study, we aim to compare the performance of three scenarios, each with a different proportion of biological zero experiments. In the first scenario, we have a higher proportion of biological zeros with alpha equal to 5. In the second scenario, alpha is set to 15, indicating the median proportion of biological zeros. The last scenario has the lowest proportion of biological zeros, with alpha set to 50.

(4) A smaller value of θ corresponds to a higher probability of obtaining a zero sample. The parameter θ follows a normal distribution with varying mean parameters and a standard deviation

of 1. The mean parameters are set at 1 and 4.

(5) To compute the covariate matrix Σ, we set the variance component σ to be equal to 2, and we assume a fixed evolutionary rate between taxa, ρ, equal to 1.

4. Results Of The Four Method Comparison

4.1. **Frobenius Norm Error.**

The Frobenius norm error is presented as follow:

$$\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{k}(\pi_{ij}-\hat{\pi}_{ij})^2}$$

The Frobenius norm is to found out the difference between two matrices that is true abundance matrix and estimaye abundance matrix, Let $\hat{\pi}_{ij}$ represent the estimated abundance and π_{ij} represent the true abundance. By summing the differences between the estimated and true abundances for each sample and each taxa, and then taking the square root of the result, we obtain a measure of how closely the imputation results approximate the actual results.In other words, a lower value of Frobenius norm error[22] implies that, on average, each entry in the imputed matrix closely approximates it's corresponding entry in the true abundance matrix.

4.2. **Mean squared error of Simpson's Index.**

The mean squared error of the Simpson's index[23] is defined as follows:

$$\frac{1}{n}\left(\sum_{i=1}^{n}(\sum_{j=1}^{k}\pi_{ij}^2-\sum_{j=1}^{k}\hat{\pi}_{ij}^2)^2\right)$$

The Simpson's index is represents the error of biodiversity of a habitat, there is not much difference between the mean squared error of Simpson's Index and the Frobenius norm error, as both equations primarily involve the estimated abundance and true abundance. The mean squared error of Simpson's Index involves squaring each main term before calculating the differences and summing across all samples. One significant difference from the Frobenius norm error is that, after summing the differences between the estimated and true abundances, the total is divided by the sample size n. This represents, on average, how far the predicted values deviate from the true values in a single sample. A lower mean squared error for Simpson's Index indicates better performance of that method.

4.3. **Wasserstein Distance Error.**

The waserstein distance error is data recovery between the estimate abundance and true abundance. The goal of Wasserstein Distance Error[24] is still to found out prediction that is estimate abundance how far from true abundance. The first step is to found out the mean of true abundance that is $\overline{\pi}_{ij}$ by summing up all sample true abundance and divide by sample size n, then using some procedure to compute mean of estimate abundance that is $\overline{\hat{\pi}_{ij}}$.

The second step is to compute the two standard deviation $\hat{\sigma}_j$,$\hat{\sigma}_j^*$.The formulas are :

$$\hat{\sigma}_j = \sqrt{1/(n-1)\sum_{i=1}^{n}(\pi_{ij} - \overline{\pi}_{ij})^2}$$

$$\hat{\sigma}_j^* = \sqrt{1/(n-1)\sum_{i=1}^{n}(\hat{\pi_{ij}} - \overline{\hat{\pi}_{ij}})^2}$$

The third step is to taking the ratio between mean and standard deviation that can be represents by :

$$(r = \{r_1, r_2, \ldots, r_K\})$$

$$(r^* = \{r_1^*, r_2^*, \ldots, r_K^*\})$$

where $r_j = \frac{\pi_{ij}}{\sigma_j}$ and $r_j^* = \frac{\hat{\pi}_{ij}}{\hat{\sigma}_j^*}$

The last step is transform these 2 ratio vector into order statistics and the mean error of Waserstein distance is :

$$1/K \sum_{j=1}^{K} \left| r_j - r_j^* \right|$$

Note here r_j^* and r_j are order statistics

4.4. **Results In Three scenario.**

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. Our goal is to evaluate which methods perform best under different proportions of biological and sampling zeros across 200 iterations. We use three standards for

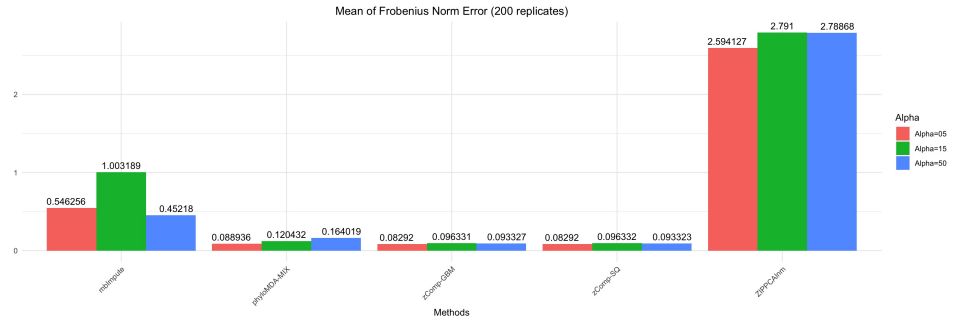

FIGURE 1. Frobenius Norm error

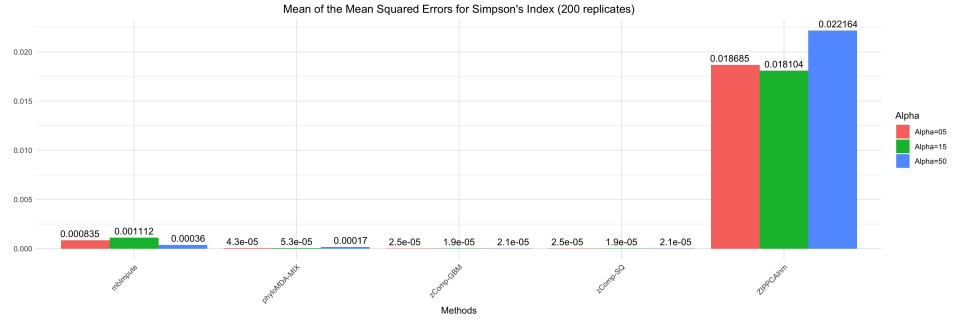

FIGURE 2. Mean Squared Error of Simpson's index

assessment: Frobenius Norm Error, Mean Squared Error of Simpson's Index, and Wasserstein Distance Error, primarily focusing on the differences between estimated and true abundances. The horizontal line in the visualization represents four methods. For each method, there are three bars representing the three scenarios adjusted by the alpha parameter. 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.The last graph is about the time consuming in every method, the method of Mbimpute cost more time than others method and next follow by method of ZIPPCA.As we can see, The most accurate and effeiency methods are Z-compositions and Phylo-MDA in all three scenarios cross 200 iterations.

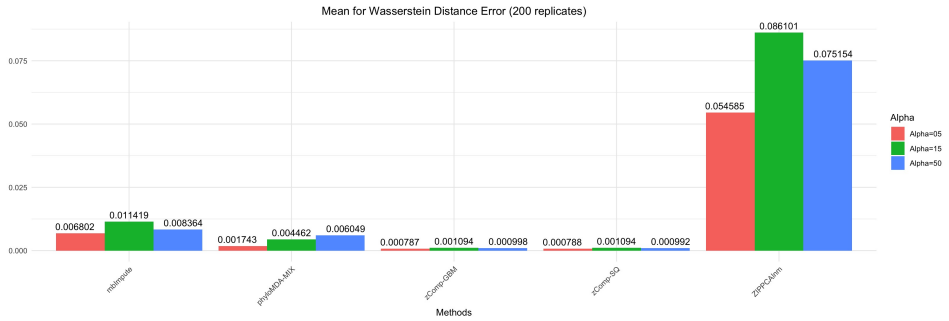

FIGURE 3. Wasserstein Distance Error

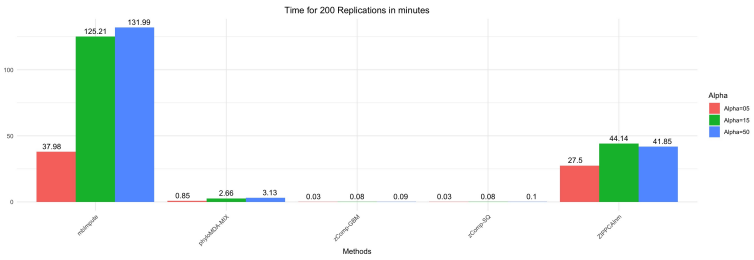

FIGURE 4. Time

5. REFERENCE

REFERENCES

[1] J. Lloyd-Price, G. Abu-Ali, and C. Huttenhower, "The healthy human microbiome," *Genome medicine*, vol. 8, pp. 1–11, 2016.

[2] 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.

[3] C. L. D. T. K. T. K. Rachel Poretsky, Luis M Rodriguez-R, "Strengths and limitations of 16s rrna gene amplicon sequencing in revealing temporal microbial community dynamics," *PloS one*, p. 9(4):e93827, 2014.

[4] S. Weiss, Z. Z. Xu, S. Peddada, A. Amir, K. Bittinger, A. Gonzalez, C. Lozupone, J. R. Zaneveld, Y. Vázquez-Baeza, A. Birmingham *et al.*, "Normalization and microbial differential abundance strategies depend upon data characteristics," *Microbiome*, vol. 5, pp. 1–18, 2017.

[5] W. V. L. Ruochen Jiang and J. J. Li, "mbimpute: an accurate and robust imputation method for microbiome data," *Genome biology*, p. 22(1):1–27, 2021.

[6] D. Baum, "Reading a phylogenetic tree: The meaning of monophyletic groups," *Nature Education*, p. 1(1):190, 2008.

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

[8] B. G. Hall, *Phylogenetic trees made easy.* WH Freeman, 2004.

[9] H. Li, "Microbiome, metagenomics, and high-dimensional compositional data analysis," *Annual Review of Statistics and Its Application*, vol. 2, pp. 73–94, 2015.

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

[11] 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.

[12] 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.

[13] 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.

[14] H. Z. Yanyan Zeng, Daolin Pang and T. Wang, "A zero-inflated logistic normal multinomial model for extracting microbial compositions," *Journal of the American Statistical Association*, p. pages 1–14.

[15] M. R. K. J. C. M. A. G. H. Zhigang Li, Katherine Lee and H. Li., "A multivariate zero-inflated logistic model for microbiome relative abundance data." *ArXiv e-prints*, 2017.

[16] J. Zhang and W. Lin., "Scalable estimation and regularization for the logistic normal multinomial model." *Biometrics*, p. 75(4):1098–1108, 2019.

[17] 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*.

[18] H. Wu *et al.*, "Numerical comparison: Different methods of handling zeros in microbiome data analysis," Ph.D. dissertation, University of Saskatchewan, 2023.

[19] G. D. Wu, J. Chen, C. Hoffmann, K. Bittinger, Y.-Y. Chen, S. A. Keilbaugh, M. Bewtra, D. Knights, W. A. Walters, R. Knight *et al.*, "Linking long-term dietary patterns with gut microbial enterotypes," *Science*, vol. 334, no. 6052, pp. 105–108, 2011.

[20] T. Liu, C. Zhou, H. Wang, H. Zhao, and T. Wang, "phylomda: an r package for phylogeny-aware microbiome data analysis," *BMC bioinformatics*, vol. 23, no. 1, p. 213, 2022.

[21] E. Paradis, *Analysis of Phylogenetics and Evolution with R*. Springer, 2012, vol. 2.

[22] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU press, 2013.

[23] T. Simmons, D. F. Caddell, S. Deng, and D. Coleman-Derr, "Exploring the root microbiome: extracting bacterial community data from the soil, rhizosphere, and root endosphere," *JoVE (Journal of Visualized Experiments)*, no. 135, p. e57561, 2018.

[24] V. M. Panaretos and Y. Zemel, "Statistical aspects of wasserstein distances," *Annual review of statistics and its application*, vol. 6, pp. 405–431, 2019.