Numerical Comparison: Different Methods of Handling Zeros in Microbiome Data Analysis

A thesis submitted to the
College of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon

By
Huokai Wu

Permission to Use

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Disclaimer

Reference in this thesis to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the University of Saskatchewan. The views and opinions of the author expressed herein do not state or reflect those of the University of Saskatchewan, and shall not be used for advertising or product endorsement purposes.

Requests for permission to copy or to make other uses of materials in this thesis in whole or part should be addressed to:

Head of the Department of Mathematics and Statistics
142 McLean Hall, 106 Wiggins Road
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5E6
Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9 Canada

Abstract

With the advancement of sequencing methods, investigators now have more opportunities to understand the microbial community's role in human and plant health. For instance, studying the biological network among microbial taxa can offer researchers insights into plant breeding. Also, studying the human microbiome can help us to understand functions and illnesses. However, analyzing microbiome data presents significant challenges due to the structure of the data. A critical issue in microbiome data analysis is the presence of a large number of zeros. Although many methods for microbiome data analysis have been published in the current literature, it remains challenging for investigators to select the appropriate method. Therefore, our work focuses on exploring recent methods for handling zeros in microbiome data analysis and provides a detailed numerical comparison. First, we introduce four recent methods: the Bayesian-multiplicative replacement model, the gamma-normal mixture model, the zero-inflated Dirichlet tree multinomial model, and the zero-inflated probabilistic PCA model, detailing their advantages and limitations. Second, we design and implement simulation studies using our novel data generator, the zero-inflated logistic normal multinomial model, which makes use of phylogenetic tree distance. To the best of our knowledge, this is the first zero-inflated model that employs the phylogenetic tree distance. Finally, we evaluated these four methods using the Frobenius norm error, mean squared error for Simpson's Index, and Wasserstein distance error in this thesis. The simulation results suggest that the Zero-Inflated Dirichlet Tree Multinomial model (with pseudo counts of 0.5 used as the smoothing method) outperforms other methods with the smallest Frobenius norm error and mean squared error for Simpson's Index. Additionally, the Square Root Multiplicative Treatment model displays notable performance, evidenced by a minimal Wasserstein distance error and efficient running time in our simulation study. Conversely, the zero-inflated probabilistic PCA model does not perform as expected due to issues with parameter estimation convergence.

Acknowledgements

First and foremost, I sincerely thank my supervisor, Dr. Juxin Liu, for her careful guidance, advice, and encouragement throughout my graduate study journey. I am deeply grateful for her thoughtful suggestions on my thesis. Also, her generous financial support was the cornerstone of my focus on the thesis. Her assistance has enabled me to concentrate on and efficiently complete my graduate study. Additionally, Dr. Liu's exemplary qualities, including her confidence and work ethic, have profoundly influenced me, allowing me to develop the quality of "grasp the nettle".

Next, I want to express my heartfelt gratitude to our dedicated members, Dr. Li Xing and Dr. Lingling Jin. The kindness and guidance from both of them have been invaluable to me. Further, I would like to extend my thanks to Dr. Shahedul A. Khan. His willingness to serve at "arm-length" greatly enhanced the defense process. Additionally, I am grateful to Dr. Yunliang Li and Elijah Shurmer from the research team for providing numerous opinions on my research.

Further, I would like to express my appreciation for the financial support from the Plant Phenotyping and Imaging Research Centre (P2IRC). I am deeply grateful for the opportunities and resources they have provided, which were instrumental in making this research possible. Additionally, I wish to express my appreciation to the Department of Mathematics and Statistics at the University of Saskatchewan for the assistantship. I also want to thank the faculty and fellow students for their continuous support in my academic journey and beyond.

Especially, I would like to thank my parents and brother for their unwavering support and love. Their belief in me, constant encouragement, and the sacrifices they've made have been the foundation upon which all my achievements stand. I am eternally grateful for everything they've done and the strength they've instilled in me. Finally, thanks to my friends Jinhao Zhong and Xudong Li for their helps and supports in my daily life.

Contents

List of Tables

List of Figures

List of Abbreviations

LOF	List of Figures
LOT	List of Tables
16S rRNA	16S ribosomal RNA

1 Introduction

1.1 Microbiome in plant

The microbiome, comprising the collective genetic material of all microorganisms—including bacteria, fungi, protozoa, and viruses—is indispensable for life sustenance on earth. Research on the plant microbiome and its implications for plant health underscores this significance, as demonstrated in the study by Turner et al.[31]. Additional work by Berendsen and colleagues [3] emphasizes the integral role of the rhizosphere microbiome in plant health. The rhizosphere, an area of soil specifically altered by root exudates, accommodates over 3×10^4 distinct prokaryotic species and can harbour up to 10^{11} microbial cells per gram of root material. According to Berendsen et al., the intricate system of the rhizosphere, which includes associated microbes, genetic components, and their interactions, greatly influences plant health. The microbiome also modulates a plant's efficacy in nutrient absorption [4]. Kudoyarova et al.[13] have shown that introducing growth-promoting bacteria into the soil can alter the soil's phosphorus content. The measured phosphorus concentrations underscore the pivotal role of bacteria in plant development. Research by Simmons et al.[28] indicates that the microbiome can enhance plant tolerance to abiotic stress and diseases. Kim et al.[10] explain that environmental stressors, such as drought escape and tolerance, frequently impact plant growth. They propose that Plant Growth-Promoting Bacteria (PGPB), endophytic fungi, and mycorrhizal fungi can enhance a plant's drought tolerance, providing a comprehensive strategy to augment its resistance to various abiotic stresses.

1.2 Microbiome in Human

The importance of the microbiome is not restricted to the realm of plant health, as it has also been shown to be a pivotal factor in human health. Pflughoeft et al.[23] underscore this, stating that the systems biology paradigm has broadened to incorporate investigations of microbiomes and functional metagenomics across multiple body sites. They suggest that this shift in focus could foster further research into the human microbiome, potentially paving the way for developing novel diagnostics and treatments for a range of acute, chronic, and systemic diseases.

Delving deeper into the human microbiome, Pathak[22] explains that it comprises numerous species, most of which inhabit the gut. The most complex and dynamic gut microbiota regarding species diversity plays a central role in regulating host homeostasis. The symbiotic relationship between the gut microbiota and the host is essential for host functionality. Dysbiosis, an imbalance in the microbiota, can disrupt a multitude of processes, leading to diseases such as inflammatory bowel disease (IBD), diabetes, and rheumatoid arthritis.

Furthermore, this symbiotic relationship establishes a complex 'super-organism' wherein dysbiosis and alterations to the microbial community can precipitate disease. For the microbial equilibrium, Gupta et al.[7] emphasize its criticality in maintaining health. They indicated that dysbiosis is emerging as a significant factor in the onset and progression of numerous diseases, encompassing a broad spectrum that includes IBS, obesity, and cancer. Their research underscores the urgent need for further exploration of the microbiome to understand better and mitigate the detrimental effects of dysbiosis.

Likewise, Mahapatra et al.[17] highlight the role of bacterial microbiota in cancer. They note an expanding body of evidence implicating bacterial microbiota in initiating various cancers, including colorectal, gastric, esophageal, lung, and oral cancers. Moreover, they suggest that alterations in the gut's microbial composition could render it susceptible to pathogenic exploitation, a scenario in which potentially harmful bacteria could increase at the expense of beneficial ones, resulting in a significant imbalance and subsequent disease.

1.3 16S rRNA sequencing techniques for microbiome studies

Given the microbiome's profound impact on plant and human health, its crucial role in related fields of study is irrefutable. This understanding necessitates a more comprehensive microbiome exploration to harness its potential fully. This pivotal process demands sequencing methods. Also, large-scale microbiome studies are rapidly advancing, stimulating the development of various sequencing methods. Amplicon sequencing and shotgun sequencing are the predominant techniques used to sequence uncultured microbes. Amplicon sequencing focuses on sequencing a specific region of DNA. It's commonly used to study a specific genomic region. However, shotgun sequencing is used to sequence entire genomes or large portions of genomes.

So, this thesis emphasizes the microbiome data generated using 16S rRNA amplicon sequencing techniques. The 16S rRNA gene, universally distributed and conserved among microbes, is commonly utilized in microbiome studies for the classification and identification of uncultured microbes [34]. The 16S rRNA sequencing method offers high taxonomic resolution, allowing the identification of specific species or taxa within a sample. It is widely recognized as a "high-throughput" and "cost-low" technique for studying the microbiome due to its ability to analyze a large number of taxa in a sample concurrently [24]. However, the method is not without limitations. For instance, while 16S rRNA sequencing provides good taxonomic reso-

lution at the phylum and class levels, its precision may diminish at lower taxonomic levels, such as genus or species, due to the high degree of sequence homology within and among different taxa [11]. Furthermore, the microbiome data generated via 16S rRNA sequencing methods can potentially introduce bias owing to PCR-based amplification, leading to inaccuracies in the assessment of the diversity and composition of microbial communities [25][26].

1.4 Phylogenetic tree

A phylogenetic tree is a branching diagram that describes the evolutionary relationships among various species or organisms.[1] It delineates the descent of species from common ancestors and their subsequent divergence into distinct groups over time. Within this hierarchical structure, each branch embodies a unique lineage, commencing with the common ancestor at the trunk and culminating with contemporary species or groups at the tips of the branches.

Phylogenetic trees are constructed utilizing a variety of data types, such as genetic sequencing, morphological data, and fossil evidence. Nevertheless, genetic sequencing has emerged as the dominant method in recent times, owing to its comprehensive insights into the genetic constitution of different organisms. Researchers can discern similarities and differences by comparing the sequences of various species, thus unveiling the complexities of their evolutionary links [2].

This thesis primarily employs phylogenetic trees that are derived from genetic sequencing. The advent of genetic sequencing has paved the way for numerous methodologies to construct a phylogenetic tree, including distance-based, parsimony-based, and likelihood-based methods. The distance-based approach computes the genetic distance between different sequences, subsequently utilizing this information to construct the tree. Conversely, the parsimony-based approach contends that the simplest explanation for the observed genetic discrepancies among sequences is likely the most accurate portrayal of evolutionary events. As for likelihood-based methods, the construction of a phylogenetic tree is predicated on the likelihood of observing the genetic differences among sequences under an array of evolutionary scenarios [8].

1.5 Characteristics of microbiome data and challenges for microbiome data analyses

In this thesis, we focus on the operational taxonomic units (OTUs) count analysis because the 16S rRNA sequencing reads are usually clustered into the OTUs [9]. The study and analysis of microbiome data, particularly the genetic material originating from a diverse array of microorganisms inhabiting a host, pose

distinct challenges that demand meticulous statistical methodologies. Several characteristics of microbiome data contribute to this complexity, including its high dimensionality, compositional nature, and the temporal and spatial variations it exhibits, in addition to intricate non-linear interactions among microbes and between the microbiome and its host [16][12].

Microbiome data's high dimensionality, derived from thousands of species and millions of genes, necessitates deploying robust techniques for data reduction and variable selection. Further, the compositional nature of the data, which often depicts relative abundances rather than absolute quantities, can complicate interpretation and analysis. This underscores the need for specialized statistical methods [5].

Moreover, statistical analysis of microbiome data faces challenges such as handling zero values and data normalization. The large number of zeros in microbiome data, resulting either from the genuine absence of organisms or from limitations in sequencing depth, call for cautious statistical analysis [18]. Considering the relative nature of microbiome data, the lack of a consensus on the optimal normalization method can yield disparate results, underscoring the necessity for uniform analytical practices [32].

1.6 Motivations and Organization of the thesis

In conclusion, the statistical analysis of zero-inflated microbiome data can be challenging. So, the primary focus of this thesis is addressing the issue of zero inflation in microbiome data.

For example, Silverman et al.[27] initially identified three types of zero-generating processes: sampling zeros, biological zeros, and technical zeros. Also, they applied zero—handling models on real microbiome datasets, including a random intercept model, a zero-inflated Poisson model, and a Biological Zero model. Their work demonstrated that the selection of a model can significantly influence the identification of sequences with the most differential expression. Their simulations suggested that three guidelines for modeling sequence count data. First, biological zeros can be effectively approximated as sampling zeros if new sequencing count models are proposed. Second, it is advisable to avoid zero-inflated models as they can yield biased estimates when the data does not encompass sample-specific complete technical processes. Third, the Poisson, negative binomial or multinomial models can incorporate covariate information for modelling zero values.

Additionally, Tang and Chen developed a novel probability distribution, known as the Zero-Inflated Generalized Dirichlet Multinomial (ZIGDM), for modeling microbiome compositional data. This method can accommodate more flexible correlation structures, allowing for both positive and negative correlations among taxa when covariates are available [30].

However, my thesis primarily focuses on analyzing microbiome data without covariate information. My

thesis considers between two types of zeros: sampling zeros and biological zeros. Sampling zeros emerge due to limitations in the total number of sequencing reads counted for a given sample, potentially leading to uncounted sequences, particularly the rare ones. In contrast, biological zeros arise when a taxon is genuinely absent from a biological system.

Given the range of existing methods, my thesis aims to compare some recent techniques designed to manage zero-inflation problems in microbiome data analyses. My thesis has also conducted simulation studies of numerical comparisons using the zero-inflated logistic normal multinomial (ZILNM) model [36] incorporating the phylogenetic tree information. In particular, we make use of the phylogenetic distance in our data generation.

The remainder of the thesis is structured as follows. Chapter 2 explains four recent methods for addressing zeros in microbiome data analysis, detailing their methodologies, advantages, and limitations. Chapter 3 presents simulation studies with the ZILNM model with tree distance. We apply the four recent methods to simulated data and evaluate their performance using a variety of metrics. Finally, Chapter 4 summarizes the findings and outlines future research directions.

2 Recent methods for handling zeros

In this chapter, four recent techniques for handling zero-inflated problems in microbiome data analysis are presented. Suppose, we have N samples and K taxa in the OTUs count data set. Let $i \in \{1, \cdots, N\}$ and $j \in \{1, \cdots, K\}$ be the index for sample i and taxon j, respectively. Also, let y_{ij} be the taxon j's OTU count in sample i, $\mathbf{N}_i = \sum_{j=1}^{K} y_{ij}$ be the total OTUs count in sample i, and z_{ij} is the imputed value for the taxon j's OTU count in sample i. Lastly, we define $y^*_{ij} = \frac{y_{ij}}{\mathbf{N}_i}$ be the proportion value for taxon j in sample i

2.1 Bayesian-Multiplicative replacement method

To address the issue of zero inflation in compositional data analysis, Martín-Fernández et al. proposed a Bayesian-multiplicative treatment in 2015 [19]. They concluded that log-ratio methodologies, commonly used in compositional techniques, necessitate positive data, implying that zeros in count compositions must be properly replaced before any statistical analysis. Accordingly, they introduced a Bayesian-multiplicative method that leverages the Dirichlet prior distribution, the conjugate of the multinomial distribution. This technique involves a multiplicative adjustment of non-zero values. Different parameterizations of the prior distribution result in various zero replacement outcomes, all of which are consistent with the vector space structure of the simplex.

2.1.1 Adavantages and Limitations

There are some adavantages for using the Bayesian multiplicative methods:

- Regarding the non-zero values, the ratios of any two imputed value are the same as the original ratios (i.e. $\frac{z_{ij}}{z_{ik}} = \frac{y^*_{ij}}{y^*_{ik}}$). There is minor distortion of the association between the taxa for this method.
- This method is effective, meaning it has a fast execution time in R and can efficiently handle the imputation of the microbiome datasets with the large number of taxa and samples.

However, there are also some limitations with the this method:

- When using the Bayesian multiplicative treatment, all zeros in microbiome data set are treated as the results of the insufficiently large samples, and all of them are called “sampling zeros”.

- When using this method, the column with all zero values should be removed.

2.1.2 Methodology

Let $\hat{m}_{ij} = \frac{\sum_{u=1,u\neq i}^{n} y_{uj}}{\sum_{k=1}^{K}\sum_{u=1,u\neq i}^{n} y_{uk}}$ be the prior information. $\hat{\boldsymbol{m}}_i = (\hat{m}_{i1}, \hat{m}_{i2}, \cdots, \hat{m}_{iK})$ is the maximum likelihood estimation of the multinomial model. In this Bayesian-Multiplicative method, this prior serves as an estimate obtained through a leave-one-out scheme. When handling the vector $\boldsymbol{y}_i^* = (y_{i1}^*, y_{i2}^*, \cdots, y_{iK}^*)$, it assumes that the information from the remaining samples serves as the prior information. Also, s_i is the strength of the prior information in sample i, and it will be varied in different prior. Because the prior $\hat{\boldsymbol{m}}_i$ is very informative, the posterior probabilties tend to the prior m_{ij} when $s \to \infty$

So, the Bayesian multiplicative model is shown following:

$$z_{ij} = \begin{cases} \hat{m}_{ij} \cdot \frac{s_i}{\mathbf{N}_i + s_i}, \text{ if } y_{ij}^* = 0 \\ y_{ij}^* \cdot (1 - \sum_{\{k:y_{ik}^*=0\}} \hat{m}_{ik} \cdot \frac{s_i}{\mathbf{N}_i + s_i}), \text{ if } y_{ij}^* > 0 \end{cases} \tag{2.1}$$

where $\sum_{j=1}^{K} z_{ij} = 1$.

For Geometric Bayesian multiplicative treatment model (GBM), the strength of the prior is $s_i = \frac{1}{g_i}$, and $g_i = (\prod_{k=1}^{K} \hat{m}_{ik})^{\frac{1}{K}}$ is the geometric mean of $\hat{m}_i = (\hat{m}_{i1}, ..., \hat{m}_{iK})$

$$z_{ij} = \begin{cases} \hat{m}_{ij} \cdot \frac{1}{g_i \cdot \mathbf{N}_i + 1}, \text{ if } y_{ij}^* = 0 \\ y_{ij}^* \cdot (1 - \sum_{\{k:y_{ik}^*=0\}} \hat{m}_{ik} \cdot \frac{1}{g_i \cdot \mathbf{N}_i + 1}), \text{ if } y_{ij}^* > 0 \end{cases} \tag{2.2}$$

For square root multiplicative treatment model (SQ), the strength of the prior is $s_i = \sqrt{\mathbf{N}_i}$

$$z_{ij} = \begin{cases} \hat{m}_{ij} \cdot \frac{1}{\sqrt{\mathbf{N}_i} + 1}, \text{ if } y_{ij}^* = 0 \\ y_{ij}^* \cdot (1 - (\sum_{\{k:y_{ik}^*=0\}} \hat{m}_{ik}) \cdot \frac{1}{\sqrt{\mathbf{N}_i} + 1}), \text{ if } y_{ij}^* > 0 \end{cases} \tag{2.3}$$

For Bayes-Laplace multiplicative treatment model (BL), the strength of the prior is $s_i = K$

$$z_{ij} = \begin{cases} \hat{m}_{ij} \cdot \dfrac{K}{\mathbf{N}_i + K}, \text{ if } y^*_{ij} = 0 \\ y^*_{ij} \cdot (1 - (\sum\limits_{\{k: y^*_{ik}=0\}} \hat{m}_{ik}) \cdot \dfrac{K}{\mathbf{N}_i + K}), \text{ if } y^*_{ij} > 0 \end{cases} \tag{2.4}$$

2.2 Gamma-Normal Mixture Model

The mbImpute method is the first imputation method for microbiome data proposed by Jiang et al. in 2021 [9]. Consider $\widetilde{Y}_{ij}$ to represent the normalized OTU count for taxon j in sample i, and let $p_j \in (0, 1)$ denote the probability that taxon j is falsely undetected. Then, α_j and β_j represent the shape parameter and rate parameter in Gamma distribution, respectively, and both of them are greater than zero. $\mathrm{X} = \{X_{1\cdot}, \cdots, X_{N\cdot}\}$ represent the covariate matrix. So, γ_j denotes the covariate's parameter for taxon j, and σ_j^2 means the variance of the normal distribution.

2.2.1 Adavantages and Limitations

As the first imputation method for microbiome data, the advantages are summarized in the following:

- The gamma-normal model can distinguish the zeros into the sampling and biological zeros, and the biological zeros will not be imputed.
- In the gamma-normal model, the mean parameter of the normal distribution enables a taxon to have similarly expected abundances in those samples with close covariates' information.
- mbImpute uses a linear model that incorporates the close taxa, all samples and covariates information for the imputation in its imputation model. However, neither the sample covariate matrix nor the phylogenetic tree is required by the mbImpute method.

However, there are also some limitations when using the mbImpute method:

- In the mbImpute method, the sampling zeros and the low abundances of taxa will also be imputed. In addition, the output from mbImpute does not include the identification information for sampling zeros/biological zeros.
- Also, the input OTU counts need to be normalized to the same total OTU count for each sample. If the user does not provide a normalization method, the built-in method is used to normalize the OTU counts.

- Lastly, this method is not effective in situations where the sample covariate matrix is unavailable, which leads to slow execution time in R because metadata information needs to be randomly generated. As the number of taxa increases, the efficiency of the method gradually decreases.

2.2.2 Methodology

$$\widetilde{Y}_{ij} \sim p_j \cdot \Gamma(\alpha_j, \beta_j) + (1 - p_j) \cdot \mathcal{N}(X_{i\cdot}^T \gamma_j, \sigma_j^2) \tag{2.5}$$

where the Gamma distribution is for the taxon's likely sampling zeros and low abundances (need to impute), and the normal distribution for the taxon's actual abundances (if biological zeros, do not need to impute). Then, the estimated posterior probability $\hat{p}_j$ for $\widetilde{Y}_{ij}$ that comes from the part of Gamma distribution will be used to specify whether $\widetilde{Y}_{ij}$ need to impute:

$$d_{ij} = \frac{\hat{p}_j \cdot f_\Gamma(\widetilde{Y}_{ij}; \hat{\alpha}_j, \hat{\beta}_j)}{\hat{p}_j \cdot f_\Gamma(\widetilde{Y}_{ij}; \hat{\alpha}_j, \hat{\beta}_j) + (1 - \hat{p}_j) \cdot f_\mathcal{N}(\widetilde{Y}_{ij}; X_{i\cdot}^T \hat{\gamma}_j, \hat{\sigma}_j^2)} \tag{2.6}$$

where $f_\Gamma(\widetilde{Y}_{ij}; \hat{\alpha}_j, \hat{\beta}_j)$ and $f_\mathcal{N}(\widetilde{Y}_{ij}; X_{i\cdot}^T \hat{\gamma}_j, \hat{\sigma}_j^2)$ are the probability density functions.
Based on the results of d_{ij}, mbImpute method defines the following sets:

- A set that does not need imputation (structural zeros and non-zero values)

$$\Theta = \{(i,j) : d_{ij} < d_{\text{thre}}, i = 1, ..., N; j = 1, ..., K\}$$

- A set that needs imputation (sampling zeros and low abundance)

$$\Theta^c = \{(i,j) : d_{ij} \geq d_{\text{thre}}, i = 1, ..., N; j = 1, ..., K\}$$

For Θ^c, the imputed $\hat{Y}_{ij}$ will be:

$$\hat{Y}_{ij} = \widetilde{Y}_{i\cdot}^T \hat{\kappa}_j + \widetilde{Y}_{\cdot j}^T \hat{\tau}_i + X_{i\cdot}^T \hat{\xi}_j \tag{2.7}$$

where $\hat{\kappa}_j$ denotes the estimated K taxa's coefficients, $\hat{\tau}_i$ denotes the N samples' coefficients, and $\hat{\xi}_j$ denotes the sample covariate's coefficients for taxon j.

2.3 Zero-Inflated Dirichlet Tree Multinomial distribution

Zhou et al. proposed a Zero-Inflated Dirichlet Tree Multinomial distribution to handle zero-inflated microbiome data in 2021 [38]. This method marks the first approach to multivariate modelling of microbial counts

that both addresses data sparsity and integrates correlation and phylogeny among bacterial taxa. Upon fitting the count data into the Zero-Inflated Dirichlet Tree Multinomial model, Zhou et al. utilize a Bayesian formulation to apply a posterior mean transformation to the raw counts. This procedure effectively transforms these counts into non-zero relative abundances that collectively sum to one, thereby accommodating the compositional nature of microbiome data.

2.3.1 Adavantages and Limitations

As the first method that incorporate phylogenetic information, there are some advantages for using the zero-inflated Dirichlet tree multinomial distribution:

- This method explicitly makes use of the phylogenetic tree information in the modeling.
- Zero-Inflated Dirichlet Tree Multinomial (ZIDTM) distribution can accommodate both negative and positive correlations between counts on tree nodes simultaneously.

However, there are also some limitations for using this method:

- The phylogenetic tree must be binary tree and is assumed to be perfectly constructed.
- The count distribution over a subtree is assumed to be conditionally independent (given the total counts of the subtree) across internal nodes. Such an assumption is imposed for computational convenience without scientific support.
- Due to the technical issues for the built-in function, this method can not handle the OTUs count data with large number of taxa.

2.3.2 Methodology

Suppose $\boldsymbol{y} = (y_1, y_2, \cdots, y_K)^T$ is the vector of OTU counts for a sample with K taxa, $\boldsymbol{p} = (p_1, p_2, \cdots, p_K)^T$ is the vector of probabilities that taxa exist with $p_j > 0$ and $\sum_{j=1}^{K} p_j = 1$. Then, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_K)^T$ is the vector of concentration parameters in Dirichlet distribution.

Multinomial distribution

Then, we have the probability mass function for multinomial distribution:

$$f_{\text{Multi}}(\boldsymbol{y}|\boldsymbol{p}) = \frac{\Gamma(\sum_{j=1}^{K} y_j + 1)}{\prod_{j=1}^{K} \Gamma(y_j + 1)} \prod_{j=1}^{K} p_j^{y_j} \tag{2.8}$$

Dirichlet distribution

The probability density function for Dirichlet distribution:

$$f_{\text{Diri}}(\boldsymbol{p}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^{K}\alpha_j)}{\prod_{j=1}^{K}\Gamma(\alpha_j)}\prod_{j=1}^{K}p_j^{\alpha_j-1} \tag{2.9}$$

Dirichlet-Multinomial distribution

The probability density function for Dirichlet Multinomial distribution:

$$f_{\text{Diri-Mult}}(\boldsymbol{y}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^{K}y_j+1)\Gamma(\sum_{j=1}^{K}\alpha_j)}{\Gamma(\sum_{j=1}^{K}y_j+\sum_{j=1}^{K}\alpha_j)}\prod_{j=1}^{K}\frac{\Gamma(y_j+\alpha_j)}{\Gamma(y_j+1)\Gamma(\alpha_j)} \tag{2.10}$$

Zero-Inflated Dirichlet Multinomial distribution

Suppose, we have a vector $\boldsymbol{Z} = (Z_1, Z_2, \cdots, Z_K)^T$ and $Z_j \sim \text{Gamma}(\alpha_j, \lambda)$ independently. Let $\boldsymbol{W} = (W_1, W_2, \cdots, W_K)^T$ be the vector with independent beta variables $W_j \sim \text{Beta}(\alpha_j, \sum_{j+1}^{K}\alpha_j)$. Also, we let $X_j = \frac{Z_j}{\sum_{m=1}^{K}Z_m}$, $W_j = \frac{Z_j}{\sum_{m=1}^{K}Z_m}$ for $j = 1, 2, \cdots, K-1$. Then, $X_K = 1 - \sum_{j=1}^{K-1}X_j$ and the joint $\boldsymbol{X} = (X_1, X_2, \cdots, X_K)^T$ follow the Dirichlet distribution with parameter vector $\boldsymbol{\alpha}$.

Then, we have the Beta density function:

$$f_{\text{Beta}}(W_j|\alpha_j, \sum_{j+1}^{K}\alpha_j) = \frac{\Gamma(\alpha_j\sum_{j+1}^{K}\alpha_j)}{\Gamma(\alpha_j)\Gamma(\sum_{j+1}^{K}\alpha_j)}W_j^{\alpha_j-1}(1-W_j)^{\sum_{j+1}^{K}\alpha_j-1} \tag{2.11}$$

In addition, we have the transformation function:

$$\begin{cases} X_1 = W_1 \\ X_j = W_j\prod_{m=1}^{j-1}(1-W_m), \text{ for } j = 2, 3, \cdots, K-1 \end{cases} \tag{2.12}$$

Thus, the density functions for the Dirichlet multinomial distribution and zero-inflated Dirichlet multinomial distribution will become:

$$f_{\text{Diri-Mult}}(\boldsymbol{y}|\boldsymbol{\alpha}) = \int\left\{\frac{\Gamma(\sum_{j=1}^{K}y_j+1)}{\prod_{j=1}^{K}\Gamma(y_j+1)}\prod_{j=1}^{K-1}(f_{\text{tranform}})^{y_j}f_{\text{Beta}}(W_j|\alpha_j, \sum_{j+1}^{K}\alpha_j)\right\}d\boldsymbol{W} \tag{2.13}$$

If we let $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_j)$ be the vector of probabilities of zero-inflation, and $\delta(\cdot)$ be the Dirac delta function.

$$f_{\text{ZI-Diri-Mult}}(\boldsymbol{y}|\boldsymbol{\pi}, \boldsymbol{\alpha}) = \int \left\{ \frac{\Gamma(\sum_{j=1}^{K} y_j + 1)}{\prod_{j=1}^{K} \Gamma(y_j + 1)} \prod_{j=1}^{K-1} (f_{\text{tranform}})^{y_j} \left[\pi_j \delta(0) + (1 - \pi_j) f_{\text{Beta}}(W_j|\alpha_j, \sum_{j+1}^{K} \alpha_j) \right] \right\} d\boldsymbol{W} \tag{2.14}$$

Dirichlet tree multinomial distribution

Suppose we have the phylogenetic tree that including all the taxa, the internal node set is defined as $\mathcal{Q}$ and each internal node $\mathfrak{q} \in \mathcal{Q}$. Then, we let $\mathcal{C}_\mathfrak{q}$ be the set of child nodes of $\mathfrak{q}$, $\boldsymbol{y}_\mathfrak{q}$ be the vector of OTUs count corresponding to $\mathcal{C}_\mathfrak{q}$. So, the density function for the Dirichlet tree multinomial distribution will be:

$$f_{\text{Diri-Tree-Mult}}(\boldsymbol{y}|\boldsymbol{\alpha}_\mathfrak{q}, \mathfrak{q} \in \mathcal{Q}) = \prod_{\mathfrak{q} \in \mathcal{Q}} f_{\text{Diri-Mult}}(\boldsymbol{y}_\mathfrak{q}|\sum_{u \in \mathcal{C}_\mathfrak{q}} y_u, \boldsymbol{\alpha}_\mathfrak{q}) \tag{2.15}$$

zero-inflated Dirichlet tree multinomial distribution

For the zero-inflated Dirichlet tree multinomial distribution, the density function will be:

$$f_{\text{ZI-Diri-Tree-Mult}}(\boldsymbol{y}|\boldsymbol{\pi}_\mathfrak{q}, \boldsymbol{\alpha}_\mathfrak{q}, \mathfrak{q} \in \mathcal{Q}) = \prod_{\mathfrak{q} \in \mathcal{Q}} f_{\text{ZI- Diri-Mult}}(\boldsymbol{y}_\mathfrak{q}|\sum_{u \in \mathcal{C}_\mathfrak{q}} y_u, \boldsymbol{\pi}_\mathfrak{q}, \boldsymbol{\alpha}_\mathfrak{q}) \tag{2.16}$$

2.4 Zero-Inflated Probabilistic PCA Model

Given the multifaceted influences such as high dimensionality, over-dispersion, and complex co-occurrence relationships that impact the estimation of microbial compositions, Zeng et al. introduced a novel approach termed zero-inflated probabilistic PCA (ZIPPCA) in 2022 [36]. This method thoughtfully considers the compositional aspect of microbiome data. It employs an empirical Bayes approach in estimating microbial compositions, thereby enhancing the precision and reliability of the results. Furthermore, the unique classification variational approximation algorithm embedded in this method simplifies the execution of maximum likelihood estimations, paving the way for more efficient and accurate analyses.

2.4.1 Adavantages and Limitations

As the most recent method for zero-inflated microbiome data analysis, there is advantage for this method:

- This method can be used without knowing the covariate information.

- In this method, the output includes the identification information for sampling zeros and biological zeros.

However, there is also limitation for this method:

- This method is not effective in situations where the number of taxa is large, which leads to slow execution time in R.
- Also, the parameter estimation does not converge.

2.4.2 Methodology

First, we define $\boldsymbol{p} = (p_1, p_2, \cdots, p_K)^T$ as the vector of latent probabilities that indicate the nonexistence of taxa. Then, we introduce $\boldsymbol{\delta}_i = (\delta_{i1}, \delta_{i2}, \cdots, \delta_{iK})^T$ as the vector of indicators for the presence of different taxa in each sample. Here, if $\delta_{ij} = 1$, it signifies that taxon j is absent in sample i. Conversely, if $\delta_{ij} \neq 1$, taxon j is present in sample i. Second, we assume that the latent environmental factors $f_{i1}, \cdots, f_{iF}$ are independently distributed according to $\mathcal{N}(0, 1)$, and the corresponding factor coefficients are contained in the vector $\boldsymbol{B} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K)^T$ where $\boldsymbol{\beta}_j = (\beta_{j1}, \cdots, \beta_{jF})$. Third, we define $\boldsymbol{\pi}_i = (\pi_{i1}, \cdots, \pi_{iK})^T$ as the true abundance in sample i. We also represent the operational taxonomic units (OTUs) count as $\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{iK})$ and the total OTUs count in sample i as $\boldsymbol{N}_i = \sum_{j=1}^{K} Y_{ij}$.

Thus, the ZIPPCA model will be:

$$\delta_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(p_j),$$

$$f_{i1}, \cdots, f_{iF} \overset{\text{ind}}{\sim} \mathcal{N}(0, 1)$$

$$\pi_{ij} = \frac{(1 - \delta_{ij}) \exp(\beta_{0_j} + \boldsymbol{f}_i^T \boldsymbol{\beta}_j)}{\sum_{m=1}^{K} (1 - \delta_{im}) \exp(\beta_{0_m} + \boldsymbol{f}_i^T \boldsymbol{\beta}_m)}$$

$$\boldsymbol{Y_i} | \boldsymbol{\pi_i} \overset{\text{ind}}{\sim} \text{Multinomial}(\boldsymbol{\pi_i}, \mathbf{N}_i) \text{ , where } \boldsymbol{\pi_i} = (\pi_{i1}, ..., \pi_{iK})$$

3 Simulation Studies

This chapter focuses on simulated studies with the aim of comparing the four methods discussed in Chapter 2. In the following, Section 3.1 introduces the generative model; Section 3.2 provides details of the data generation process; Section 3.3 presents the metrics used for evaluating model fit performance; Section 3.4 reports the findings of the simulation studies.

3.1 Data Generated from the Zero-Inflated Logistic Normal Multinomial Model based on Phylogenetic Tree

Before we introduced the data generation model, we first define the following notations:

- Let $i \in (1, 2, \cdots, N)$ be the index of the sample and $j \in (1, 2, \cdots, K)$ be the index of taxa.
- $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \cdots, Y_{iK})^T$ indicates the observed OTU counts in sample i.
- $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \cdots, \pi_{iK})^T$ means the relative abundance in sample i.

In the Zero-Inflated Logistic Normal Multinomial Model (ZILNM model) ([36], [14], [37]), we define the latent parameter space $\boldsymbol{\delta}_i = (\delta_{i1}, \delta_{i2}, \cdots, \delta_{ik})$ to signify the presence of taxa in sample i. If $\delta_{ij} = 1$, taxon j is absent in sample i; otherwise, if $\delta_{ij} = 0$, taxon j is present.

Our simulation utilizes the phylogenetic tree and assumes that the existence probability of each taxon is constant across all samples. We designate p_j as the probability that $\delta_{ij} = 1$, with α_j being the parameter that controls p_j. We assume that when $\delta_{ij_1} = \delta_{ij_2} = \cdots = \delta_{ij_L} = 0$, $\boldsymbol{\pi}_i^* = (\pi_{ij_1}, \pi_{ij_2}, \cdots, \pi_{ij_L})^T$ represents the non-zero true abundance in sample i, where $1 < j_1 < \cdots < j_L \leq K$. Importantly, the total number of j_l will vary between samples.

We define the log-ratio transformation from $\boldsymbol{\pi}_i^*$ to $\boldsymbol{U}_i$:

$$(\log \frac{\pi_{ij_1}}{\pi_{ij_L}}, \log \frac{\pi_{ij_2}}{\pi_{ij_L}}, \cdots, \log \frac{\pi_{ij_{L-1}}}{\pi_{ij_L}})^T = (u_{ij_1}, u_{ij_2}, \cdots, u_{ij_{L-1}})^T$$

Distinct from existing generative models, our model employs the phylogenetic distance to address the issue of taxa interaction. Consequently, we use $\boldsymbol{D}_{k\times k} = (D_{lm})$ to represent the phylogenetic distance matrix,

where D_{lm} denotes the phylogenetic distance between taxon l and taxon m.

In this model, $\boldsymbol{U}_i$ follows a multivariate normal distribution, i.e. $\boldsymbol{U}_i \sim \text{MVN}(\boldsymbol{\theta}_i, \boldsymbol{\Sigma})$. Here, $\boldsymbol{\theta}_i$ signifies the vector of mean parameters in the multivariate normal distribution. We define the variance-covariance matrix for various taxa as $\boldsymbol{\Sigma} = (\Sigma_{lm})$, where $\Sigma_{lm} = \sigma^2 \exp\{-2\rho_{lm} D_{lm}\}$ is proposed by Xiao et al.[35]. In this thesis, σ^2 represents the variance component, and $\rho_{lm} \in (0, \infty)$ indicates the evolutionary rate between taxon l and taxon m. When $\rho_{lm} = 0$, no evolution occurs between taxa l and m. Conversely, a ρ_{lm} approaching ∞ implies rapid evolution between these taxa. Lastly, $\mathbf{N}_i$ represents the total number of the OTU counts in sample i.

The ZILNM model based on phylogenetic distance will be expressed as:

$$p_j = \frac{1}{1 + \exp\{-\alpha_j\}},$$

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

$$\text{if } \delta_{ij} = 1 : \pi_{ij} = 0$$

$$\text{if } \delta_{ij_1} = \delta_{ij_2} = \cdots = \delta_{ij_L} = 0 : \boldsymbol{U}_i \sim MVN(\boldsymbol{\theta}_i, \boldsymbol{\Sigma})$$

$$\pi_{ij_l} = \frac{\frac{\pi_{ij_l}}{\pi_{ij_L}}}{1 + \sum_{l=1}^{j_L - 1} \frac{\pi_{ij_l}}{\pi_{ij_L}}} \text{ , and } \pi_{ij_L} = \frac{1}{1 + \sum_{l=1}^{j_L - 1} \frac{\pi_{ij_l}}{\pi_{ij_L}}}$$

$$\boldsymbol{Y_i} | \boldsymbol{\pi_i} \sim \text{Multinomial}(\boldsymbol{\pi}_i^*, \mathbf{N}_i)$$

3.2 Procedures for Data Generation

Regarding the phylogenetic tree, certain assumptions were made for the simulation study. In light of the numerous methods for constructing such a tree, this thesis is limited to the tree constructed using reference data. This implies that our analysis is restricted to those taxa that are annotated on the phylogenetic tree.

Our simulation studies utilize the original "COMBO" dataset first published by Wu et al. [33]. This dataset was devised to investigate the correlations between dietary variables and gut microbiota. Wu et al. conducted a cross-sectional study on 98 healthy volunteers, colloquially termed "COMBO." They collected dietary information through two distinct questionnaires: one targeting recent diet ("Recall") and another probing into long-term dietary habits ("FFQ" or Food Frequency Questionnaire). The original study used 16S ribosomal DNA (rDNA) sequence data to compute pairwise UniFrac distances amongst the microbial

communities. Following a preprocessing stage managed by Liu et al. [15], the restructured "COMBO" dataset comprises an OTU count table with 98 samples and 62 taxa. The corresponding phylogenetic tree is also incorporated within the "COMBO" dataset.

Since we used the phylogenetic tree in the "COMBO" dataset as our reference phylogenetic tree in our simulation study, the phylogenetic relation between the taxa will not varied across different samples.

Lastly, the procedures for the simulation study is shown below:

1. To simulate conditions that closely resemble the real data example, we set the sample size, denoted as N, to 98, reflecting the number of individuals in the "COMBO" dataset. Likewise, the total taxa, represented by K, is fixed at 62. Consequently, the total OTU counts for each sample, indicated by $\mathbf{N}_i$, are derived from the existing data in the "COMBO" dataset.

2. To compute the phylogenetic distance, we utilized the "cophenetic.phylo" function [21], which calculates pairwise distances between pairs of tips on a phylogenetic tree using their branch lengths. However, we generated two distinct phylogenetic distance matrices by modifying the edge lengths. Firstly, we conceptualized the phylogenetic distance as the sum of the edge lengths between two taxa. By applying the "cophenetic.phylo" function directly to the phylogenetic tree, we obtained the distance matrix $\boldsymbol{D}_1$. Secondly, we defined the phylogenetic distance as the number of edges linking two taxa. To accomplish this, we adjusted all the edge lengths in the phylogenetic tree to be equal to one, resulting in the phylogenetic tree distance matrix $\boldsymbol{D}_2$.

3. To reflect the reality that different taxa may have varying probabilities of being biologically zero. For instance, a taxon A, with high relative abundance in most samples, would be less likely to be absent. Accordingly, we sampled the values of $\alpha_j, j = 1, \cdots, k$ from the standard normal distribution. This approach allows for a realistic representation of the varied likelihood of biological zero status across different taxa.

$$\alpha_j \sim \mathcal{N}(0, 1)$$

4. To account for the proportion of sampling zeros in the existing taxa, we sampled $\boldsymbol{\theta}$ from a normal distribution with varying mean parameters. The mean parameters were set at 0.1 and 5, both with a standard deviation of 1. A smaller mean value indicates a higher likelihood of a smaller relative abundance value, which in turn increases the probability of obtaining a sampling zero.

5. In the covariate matrix $\boldsymbol{\Sigma}$, we assign a value of 0.25 to ρ. Here, $\rho_{lm} \in (0, \infty)$ signifies the evolutionary rate between taxon l and taxon m. If $\rho_{lm} = 0$, it suggests no evolution between taxon l and taxon m. Conversely, a ρ value nearing ∞ indicates a rapid pace of evolution between the two taxa. Given that

0.25 is relatively close to zero, we interpret this as a slower rate of evolution. Additionally, we designate the variance component σ to be equal to 1.

6. After we generated the relative abundance dataset $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \cdots, \boldsymbol{\pi}_N)^T$, we repeatly generate the OTU counts datasets by using the same $\boldsymbol{\pi}$.

3.3 Comparison of the 4 methods

3.3.1 Data Fitting

Once we generate the OTU counts, we fit the data by using the exisitng methods in Chapter 2.

- For the Bayesian-Multiplicative replacement method, we utilized the "cmultRepl" function from the "zComposition" package to obtain the estimated values. This thesis applies the square root multiplicative treatment and geometric Bayesian multiplicative treatment.

- For the Gamma-Normal mixture model, the "mbImpute" function from the "mbImpute" package is employed to derive the estimated values with all the default arguments' value.

- Regarding the empirical Bayes normalization approach, the "eBay_comps" function in the "phyloMDA" package was utilized to yield the estimated values. This simulation study sets the prior as a zero-inflated Dirichlet tree model.

- In terms of the Zero-Inflated Probabilistic PCA method, we employ the "ZIPPCAlnm" function from the "ZIPPCAlnm" package to obtain the estimated values. For this simulation study, the latent environmental factors are set to their default value, which is two.

3.3.2 Evaluation Metrics

First, we let $\hat{\pi}_{ij}$ be the estimated abundance and π_{ij} be the true abundance for taxon j in sample i. Then, we employ the Frobenius norm error and the mean square error of Simpson's index to determine the similarity between the compositional estimates from the imputation methods and the true abundance.

The Frobenius norm, also known as the Euclidean norm, serves as a measure of a matrix's size. It is calculated as the square root of the sum of the absolute squares of its entries. The Frobenius norm error is a technique frequently employed to quantify the difference between two matrices [6]. In this thesis, our focus lies in comparing actual and imputed compositions. A smaller Frobenius norm indicates closer proximity of the elements of the two matrices, thus signifying a smaller error or difference between them. The formula

associated with the Frobenius norm error is presented as follows:

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

Simpson's Index, also known as Simpson's Diversity Index or the Gini-Simpson Index, is a metric employed in ecology to measure the biodiversity of a habitat. In terms of error measurement, the mean squared error of the Simpson's Index represents the average of the squared differences between the predicted and the actual Simpson's Index over a number of predictions or simulations [29]. A smaller mean squared error for the Simpson's Index indicates that the predictions or calculations are closer to the actual values, signifying better performance. The formula for the mean squared error of the Simpson's Index is defined as follows:

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

The Wasserstein distance [20] error is utilized to measure the accuracy of data recovery between the imputed methods and the true abundance. The Wasserstein distance error between two distributions essentially refers to the Wasserstein distance itself, given that this metric is inherently a measure of the disparity or 'error' between the two distributions. This distance error serves to quantify the similarity between two data distributions, where a smaller Wasserstein distance implies a more remarkable similarity between the distributions. The following steps are used in this simulation study to compute the Wasserstein distance error:

1. Mean of the true abundance and imputed value:

$$\bar{\pi}_{\cdot j}=\frac{1}{n}\sum_{i=1}^{n}\pi_{ij} \text{ and } \bar{\hat{\pi}}_{\cdot j}=\frac{1}{n}\sum_{i=1}^{n}\hat{\pi}_{ij}$$

2. Sample standard deviation ($\hat{\sigma}$) of the true abundance and standard deviation ($\hat{\sigma}^*$) of imputed value:

$$\hat{\sigma}_{\cdot j}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\pi_{ij}-\bar{\pi}_{\cdot j})^2}$$
$$\hat{\sigma}^*_{\cdot j}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\pi}_{ij}-\bar{\hat{\pi}}_{\cdot j})^2}$$

3. Ratio between mean and standard deviation of true abundance ($r=\{r_1,r_2,\cdots,r_K\}$). Ratio between

mean and standard deviation of imputed value ($r^* = \{r_1^*, r_2^*, \cdots, r_K^*\}$).

$$r_j = \frac{\bar{\pi}_{\cdot j}}{\hat{\sigma}_{\cdot j}} \text{ and } r_j^* = \frac{\bar{\hat{\pi}}_{\cdot j}}{\hat{\sigma}_{\cdot j}^*}$$

4. Then, we transform the r and r^* into the order statistics $\zeta = \{\zeta_1, \zeta_2, \cdots, \zeta_K\}$ and $\zeta^* = \{\zeta_1^*, \zeta_2^*, \cdots, \zeta_K^*\}$, respectively.

5. Mean error of Wasserstein distance

$$\frac{1}{K}\sum_{j=1}^{K} |\zeta_j - \zeta_j^*|$$

3.4 Results of the Simulation Studies

In this section, we present the results of the simulation studies conducted to evaluate the performance of the methods introduced in Chapter 2. The first set of results was conducted to explore the Frobenius norm error between the true relative abundance $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \cdots, \boldsymbol{\pi}_N)^T$ and the imputed abundance $\hat{\boldsymbol{\pi}} = (\hat{\boldsymbol{\pi}}_1, \hat{\boldsymbol{\pi}}_2, \cdots, \hat{\boldsymbol{\pi}}_N)^T$. The second set of results was comparing the mean squared error of the Simpson's index. The last set of results explored the Wasserstein distance error. First, we define the following:

- zComposition-SQ: The zComposition package in combination with the square root Bayesian multiplicative method.
- zComposition-GBM: The zComposition package using the geometric Bayesian multiplicative method
- mbImpute: The Gamma-Normal Mixture model with all default setting.
- PhyloMDA-0.5: This denotes the zero-inflated Dirichlet-tree multinomial model with smoothing methods that involve the addition of a pseudo count of 0.5.
- phyloMDA-MIX: This represents the zero-inflated Dirichlet-tree multinomial model, employing a smoothing method that first fits a beta-binomial, then tests for over-dispersion. If there is no over-dispersion for the nodes, a pseudo count of 0.5 is added to the OTU counts and then transformed into a proportion. However, if the nodes exhibit over-dispersion, the counts are refitted to the zero-inflated beta binomial and tested for zero-inflation. Subsequently, the OTU count is converted into the posterior mean of the zero-inflated beta binomial.
- ZIPPCAlnm: The zero-inflated probabilistic PCA model.

- $\boldsymbol{D}_1$: The Operational Taxonomic Unit (OTU) count datasets are generated based on the phylogenetic tree distance matrix. For this distance matrix ($\boldsymbol{D}_1$), the distances are calculated by summing the true lengths of the edges between two taxa.

- $\boldsymbol{D}_2$: The Operational Taxonomic Unit (OTU) count datasets are generated based on the phylogenetic tree distance matrix. For this distance matrix ($\boldsymbol{D}_2$), the distances are calculated by summing the number of the edges connecting two taxa.

Secondly, due to certain technical issues with its built-in functions, the 'phyloMDA' package is not compatible with higher versions of R (i.e., R $\geq$ 3.0). Consequently, we utilized a modified version in the 'phyloMDA-MIX' method. In addition, the ZIPPCAlnm method has been excluded from the general comparison, as the majority of the parameter estimations within this method do not converge. We will summarize the ZIPPCAlnm method's results in the last part of the section.

Lastly, due to "error" messages arising from the 'phyloMDA-MIX' and 'ZIPPCAlnm' methods after 200 replications in the simulation, we will primarily use the first 200 replications as the main results. The outcomes from the 500 replications will be provided in Appendix A (Figure A.1, Figure A.2, Figure A.3, Figure A.4).

3.4.1 Frobenius Norm Error

As discussed in Section 3.3.2, the minimal value of the Frobenius norm error implies that, on average, each entry in the imputed abundance matrix closely approximates its corresponding entry in the true abundance matrix. Figure 3.1 presents the mean values of the Frobenius norm errors from five different methods, compiled from 200 iterations and evaluated under two distinct phylogenetic tree settings. The vertical line represents the methods we used for the comparison, and the horizontal line is the mean values of the Frobenius norm errors. This simulation shows that the "phyloMDA-0.5" method consistently yields the smallest Frobenius norm error under both phylogenetic tree distance settings. Similarly, the "zComp-SQ" method also exhibits smaller Frobenius norm errors. However, the Frobenius norm error for "phyloMDA-MIX" is significantly larger in the Tree 2 setting compared to its performance in the Tree 1 setting. For the "mbImpute" method, the Frobenius norm errors are substantial under both phylogenetic tree distance settings.

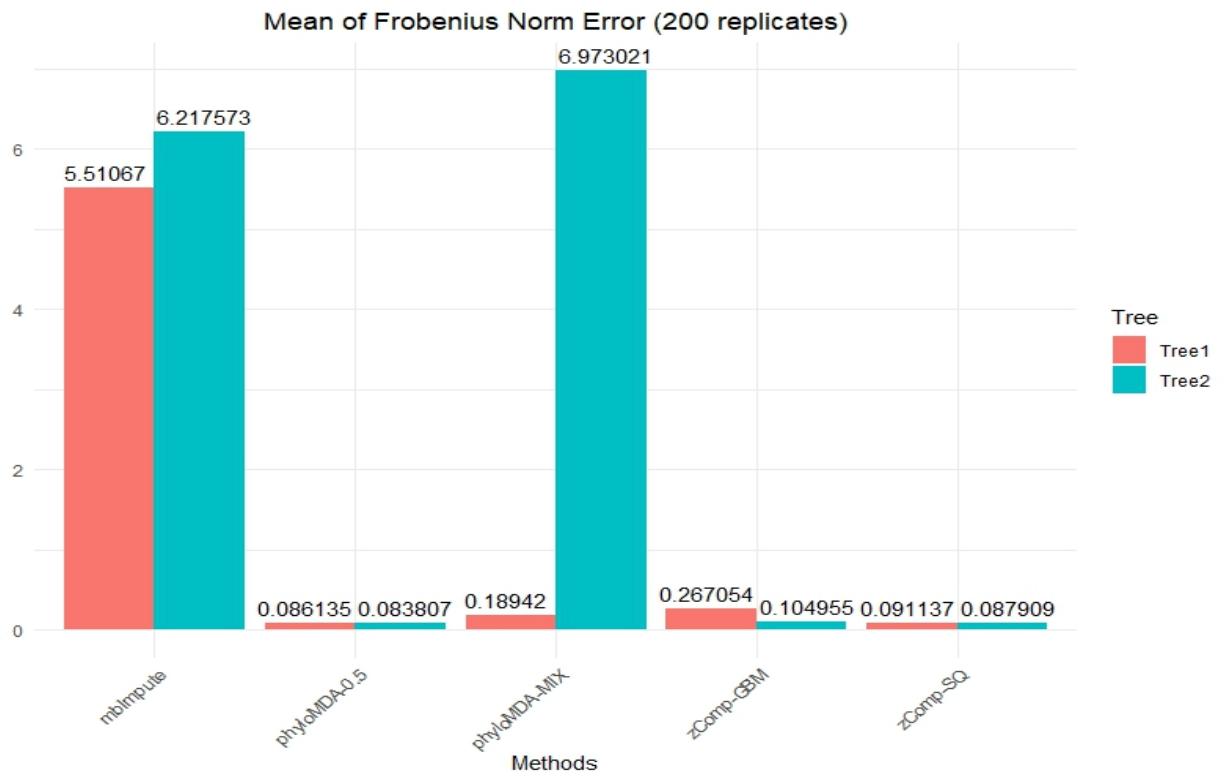

Figure 3.1: Means of the Frobenius Norm Error (200 replicates)

3.4.2 Mean Squared Error of Simpson's Index

Given that a smaller mean squared error (MSE) of the Simpson's Index indicates a lower average of squared differences between the predicted and actual values over simulations, we present the means of the MSE for Simpson's Index from 200 replicates using various methods in Figure 3.2. The graph demonstrates that "zComp-SQ" has the smallest MSE for Simpson's Index under both phylogenetic tree distance settings. The "phyloMDA-0.5" and "zComp-GBM" methods also exhibit smaller MSEs for Simpson's Index in these settings. However, the MSE for Simpson's Index associated with "phyloMDA-MIX" is significantly larger in the Tree 2 setting compared to its performance in the Tree 1 setting. For the "mbImpute" method, the MSEs for Simpson's Index are notably large under both tree settings.

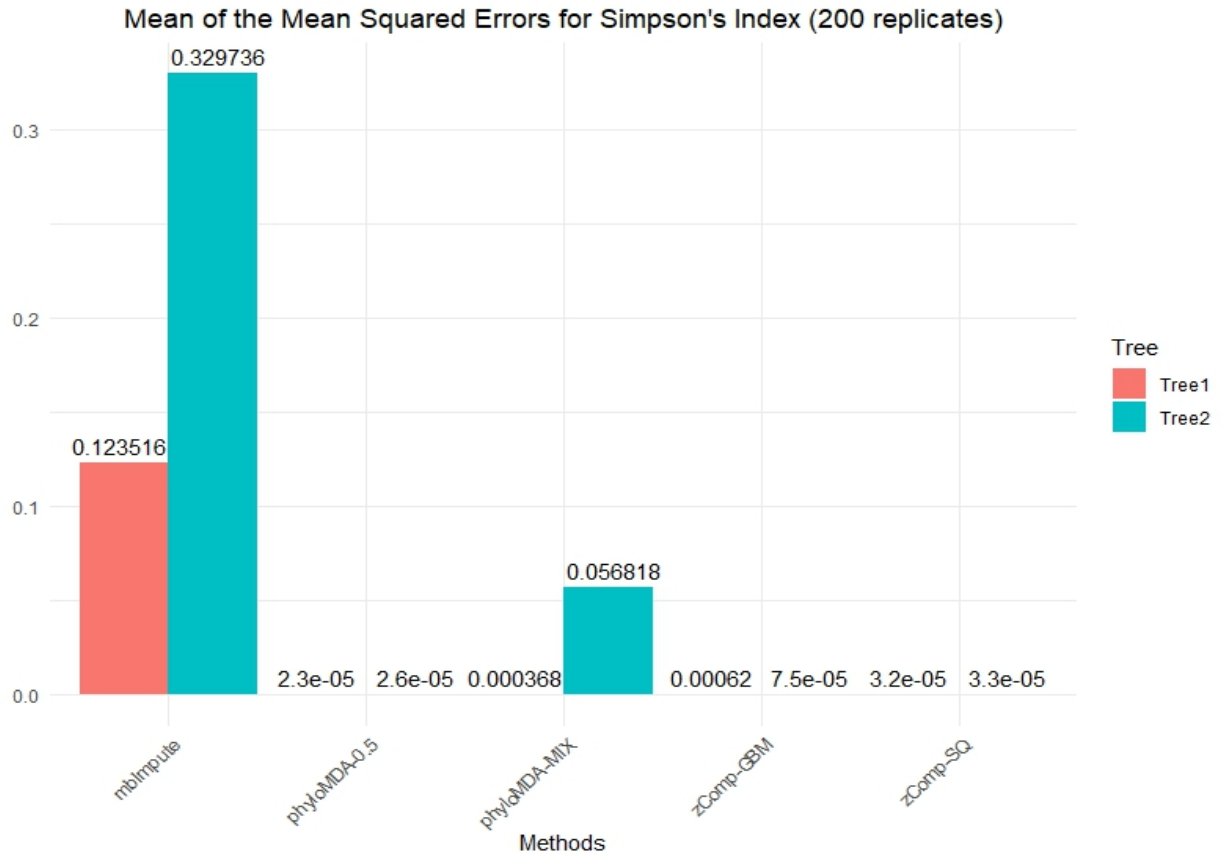

Figure 3.2: Means of MES for Simpson's Index (200 replicates)

3.4.3 Wasserstein Distance Error

In our analysis, we applied the Wasserstein Distance Error to ascertain the effectiveness of our model in accurately capturing the underlying distribution of the data. If a method exhibits a smaller Wasserstein Distance Error, it suggests that the imputed relative abundance is closer to the actual relative abundance. In Figure 3.3, the "phyloMDA-0.5" and "zComp-SQ" methods exhibit similar and comparatively more minor Wasserstein distance errors. However, the Wasserstein distance errors associated with "phyloMDA-MIX" is more significant in the Tree 2 setting than in the Tree 1 setting. Even though the "mbImpute" method has the largest Wasserstein distance errors among these methods, it still demonstrates solid performance with values of 0.023 and 0.035 in both tree settings.

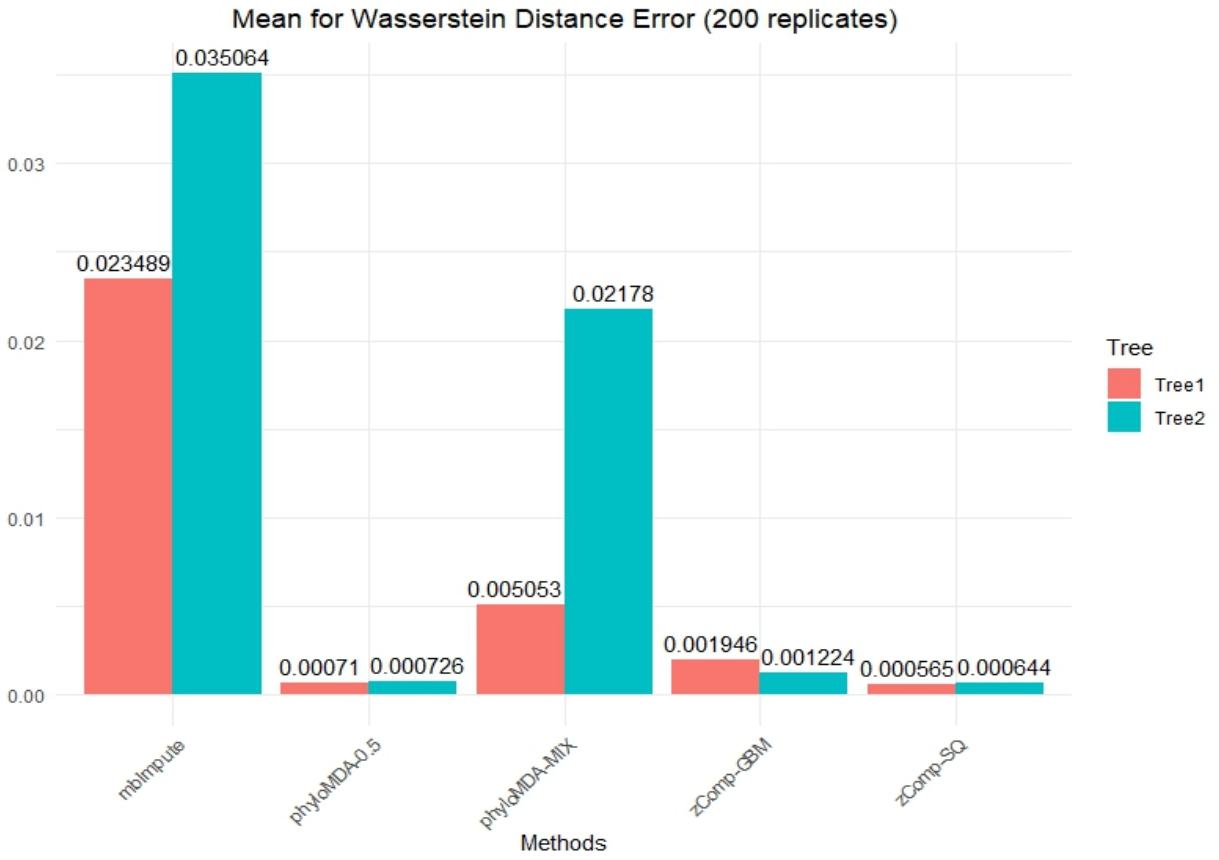

Figure 3.3: Means of Wasserstein Distance Error (200 replicates)

3.4.4 Running Time for 200 Replications

In Figure 3.4, we extensively compared different methods with regard to their processing time. As the figure shows, "phyloMDA-0.5", "zComp-GBM", and "zComp-SQ" are highly efficient methods in terms of processing time. The "phyloMDA-MIX" method also demonstrates efficiency by completing 200 replications in less than 10 minutes across both phylogenetic tree distance settings. However, the "mbImpute" method is not quite efficient as it requires more than one hour for both tree distance settings.

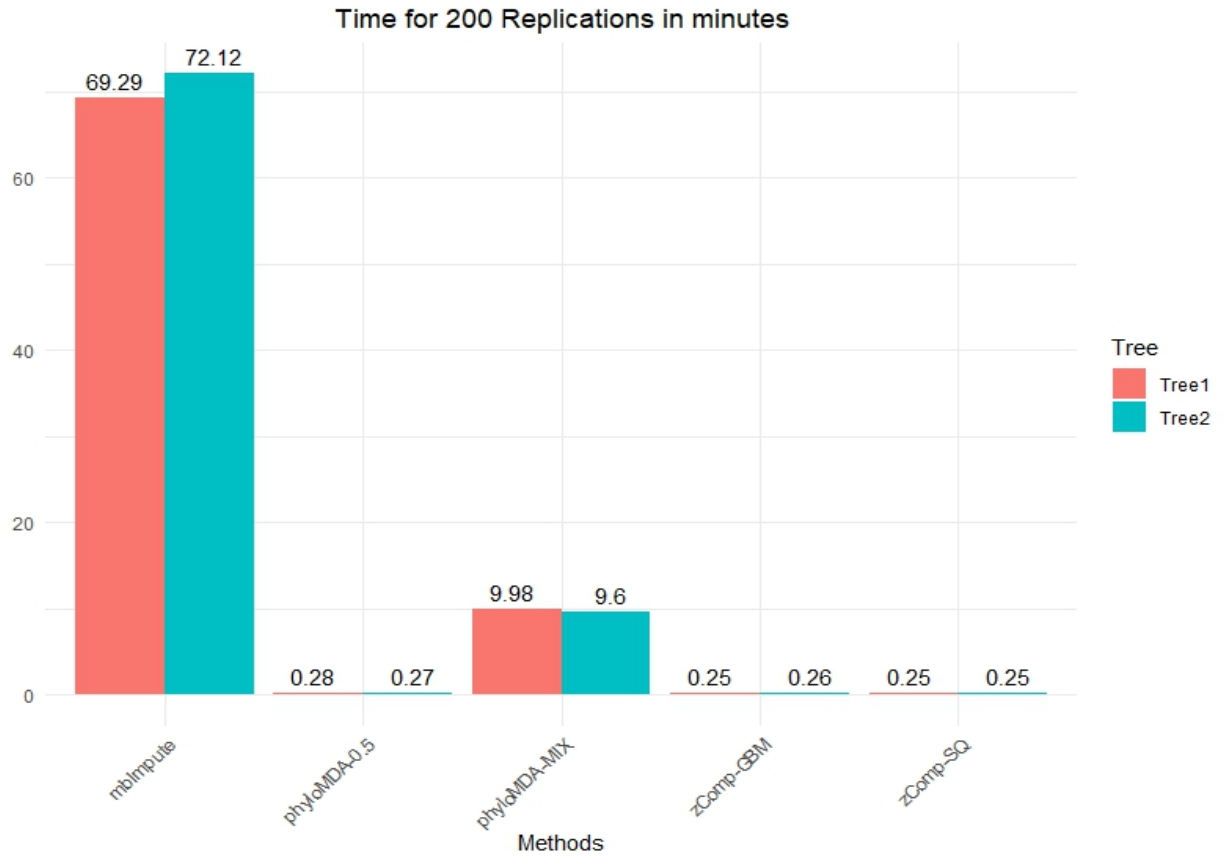

Figure 3.4: Running Time (200 replicates)

3.4.5 Results of ZIPPCAlnm for 200 replicates

	Frobenius Norm Error	MSE for Simpson's Index	Wasserstein Distance Error	Running Time (minutes)
$\boldsymbol{D}_1$	6.26042	0.07427	0.02198	519.07
$\boldsymbol{D}_2$	7.67705	0.10861	0.02191	862.19

Table 3.1: Results of ZIPPCAlnm for 200 replicates

According to Table 3.1, the results show that the "ZIPPCAlnm" method has overall better performance in the Tree 1 setting than in the Tree 2 setting. However, most of the replicates produced the message "ZILNM does not converge", so we do not believe it is fair to compare "ZIPPCAlnm" with the other methods. Additionally, the results for the 500 replications are provided in Appendix B (Table B.1).

4 Summary and Future Work

In this thesis, we first introduce four recent methods for handling zero-inflated microbiome data: the Bayesian-Multiplicative Replacement Method, the Gamma-Normal Mixture Model, the Zero-Inflated Dirichlet Tree Multinomial Model, and the Zero-Inflated Probabilistic PCA Model. Subsequently, we evaluate these methods using simulation studies. Importantly, we generated the simulated data from the zero-inflated logistic normal multinomial model, based on the phylogenetic tree distance. Following this, we assess these methods using metrics such as the Frobenius norm error, the Mean Squared Error (MSE) for Simpson's Index, and the Wasserstein distance error.

Our simulation study's result first shows that the "phyloMDA-0.5" and "zComp-SQ" methods are the most recommended for the applied practitioners. These two methods consistently performed well across several metrics including Frobenius norm errors, Mean Squared Errors (MSE) for Simpson's Index, and Wasserstein Distance Errors. The consistently small values for these metrics illustrate the ability of "phyloMDA-0.5" and "zComp-SQ" to closely approximate the imputed abundance matrices to the true relative abundance. Additionally, these two methods were efficient in terms of computational time. Therefore, they offer an advantageous blend of accuracy and efficiency, making them suitable for handling zero-inflated microbiome data.

Secondly, the simulation study results demonstrate that the "mbImpute" method underperforms with high values in Frobenius norm error, MSE for Simpson's Index, and Wasserstein distance error. Moreover, the "mbImpute" method is not efficient in terms of computational time.

Third, the "phyloMDA" package is not compatible with higher versions of R (i.e., R $\geq$ 3.0) due to certain technical issues with its built-in functions. Consequently, when employing the "MIX" option as the smoothing method, it is recommended to modify built-in function in the phyloMDA package.

Lastly, it is not recommended for practical use of the "ZIPPCAlnm" package due to its issues in parameter estimation when drawing samples from the posterior distribution of the proposed models.

Our future work is going to focus on the following aspects:

1. There is no clear reason why "phyloMDA-0.5" and "zComp-SQ" outperformed other methods at this point, and further investigation is needed.

2. Since the "ZIPPCAlnm" method did not perform well in the simulation studies outlined in this thesis, we attempted to use RJAGS for parameter estimation. However, the posterior samples did not converge to the true values of parameters, even for larger sample sizes like 5000. The potential cause for this parameter estimation issue could be non-identification of the proposed model. This will be further investigated in the future.

3. We will attempt to use RJAGS to estimate the parameters of the zero-inflated logistic normal multinomial model based on the phylogenetic tree distance proposed in Section 3.1.

4. Finally, there is no recent imputation method based on the phylogenetic tree distance, we aim to consider the imputation method based on the phylogenetic tree distance.

References

[1] David Baum et al. Reading a phylogenetic tree: the meaning of monophyletic groups. *Nature Education*, 1(1):190, 2008.

[2] David A Baum and Stacey D Smith. Tree thinking: an introduction to phylogenetic biology. In *Tree thinking: An introduction to phylogenetic biology*, pages 476–476. 2012.

[3] Roeland L Berendsen, Corné MJ Pieterse, and Peter AHM Bakker. The rhizosphere microbiome and plant health. *Trends in plant science*, 17(8):478–486, 2012.

[4] Lilia C Carvalhais and Paul G Dennis. *Plant Microbiome*. Springer, 2021.

[5] Gregory B Gloor, Jean M Macklaim, Vera Pawlowsky-Glahn, and Juan J Egozcue. Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology*, 8:2224, 2017.

[6] Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

[7] Aeshna Gupta, Vijai Singh, and Indra Mani. Dysbiosis of human microbiome and infectious diseases. *Progress in Molecular Biology and Translational Science*, 192(1):33–51, 2022.

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

[9] Ruochen Jiang, Wei Vivian Li, and Jingyi Jessica Li. mbimpute: an accurate and robust imputation method for microbiome data. *Genome biology*, 22(1):1–27, 2021.

[10] Young-Cheol Kim, Bernard R Glick, Yoav Bashan, and Choong-Min Ryu. Enhancement of plant drought tolerance by microbes. In *Plant responses to drought stress*, pages 383–413. Springer, 2012.

[11] Anna Klindworth, Elmar Pruesse, Timmy Schweer, Jörg Peplies, Christian Quast, Matthias Horn, and Frank Oliver Glöckner. Evaluation of general 16s ribosomal rna gene pcr primers for classical and next-generation sequencing-based diversity studies. *Nucleic acids research*, 41(1):e1–e1, 2013.

[12] Rob Knight, Alison Vrbanac, Bryn C Taylor, Alexander Aksenov, Chris Callewaert, Justine Debelius, Antonio Gonzalez, Tomasz Kosciolek, Laura-Isobel McCall, Daniel McDonald, et al. Best practices for analysing microbiomes. *Nature Reviews Microbiology*, 16(7):410–422, 2018.

[13] Guzel R Kudoyarova, Lidiya B Vysotskaya, Tatiana N Arkhipova, Ludmila Yu Kuzmina, Nailya F Galimsyanova, Ludmila V Sidorova, Ilusa M Gabbasova, Alexander I Melentiev, and Stanislav Yu Veselov. Effect of auxin producing and phosphate solubilizing bacteria on mobility of soil phosphorus, growth rate, and p acquisition by wheat plants. *Acta physiologiae plantarum*, 39(11):1–8, 2017.

[14] Zhigang Li, Katherine Lee, Margaret R Karagas, Juliette C Madan, Anne G Hoen, and Hongzhe Li. A multivariate zero-inflated logistic model for microbiome relative abundance data. *ArXiv e-prints*, 1709, 2017.

[15] Tiantian Liu, Chao Zhou, Huimin Wang, Hongyu Zhao, and Tao Wang. phylomda: an r package for phylogeny-aware microbiome data analysis. *BMC bioinformatics*, 23(1):1–6, 2022.

[16] Jason Lloyd-Price, Galeb Abu-Ali, and Curtis Huttenhower. The healthy human microbiome. *Genome medicine*, 8(1):1–11, 2016.

[17] Soumendu Mahapatra, Smrutishree Mohanty, Rasmita Mishra, and Punit Prasad. An overview of cancer and the human microbiome. *Human Microbiome in Health and Disease-Part A*, 191:83, 2022.

[18] Siddhartha Mandal, Will Van Treuren, Richard A White, Merete Eggesbø, Rob Knight, and Shyamal D Peddada. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial ecology in health and disease*, 26(1):27663, 2015.

[19] Josep-Antoni Martín-Fernández, Karel Hron, Matthias Templ, Peter Filzmoser, and Javier Palarea-Albaladejo. Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling*, 15(2):134–158, 2015.

[20] Victor M Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6:405–431, 2019.

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

[22] Prerna Pathak. Diversity and dynamics of the gut microbiome and immune cells. *Microbiome in Human Health and Disease*, pages 53–67, 2021.

[23] Kathryn J Pflughoeft and James Versalovic. Human microbiome in health and disease. *Annual Review of Pathology: Mechanisms of Disease*, 7:99–122, 2012.

[24] Rachel Poretsky, Luis M Rodriguez-R, Chengwei Luo, Despina Tsementzi, and Konstantinos T Konstantinidis. Strengths and limitations of 16s rrna gene amplicon sequencing in revealing temporal microbial community dynamics. *PloS one*, 9(4):e93827, 2014.

[25] Melanie Schirmer, Umer Z Ijaz, Rosalinda D'Amore, Neil Hall, William T Sloan, and Christopher Quince. Insight into biases and sequencing errors for amplicon sequencing with the illumina miseq platform. *Nucleic acids research*, 43(6):e37–e37, 2015.

[26] Patrick D Schloss, Dirk Gevers, and Sarah L Westcott. Reducing the effects of pcr amplification and sequencing artifacts on 16s rrna-based studies. *PloS one*, 6(12):e27310, 2011.

[27] Justin D. Silverman, Kimberly Roche, Sayan Mukherjee, and Lawrence A. David. Naught all zeros in sequence count data are the same. *Computational and Structural Biotechnology Journal*, 18:2789–2798, 2020.

[28] Tuesday Simmons, Daniel F Caddell, Siwen Deng, and Devin Coleman-Derr. Exploring the root microbiome: extracting bacterial community data from the soil, rhizosphere, and root endosphere. *JoVE (Journal of Visualized Experiments)*, (135):e57561, 2018.

[29] Edward H Simpson. Measurement of diversity. *nature*, 163(4148):688–688, 1949.

[30] Zheng-Zheng Tang and Guanhua Chen. Zero-inflated generalized dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*, 20(4):698–713, 2019.

[31] Thomas R Turner, Euan K James, and Philip S Poole. The plant microbiome. *Genome biology*, 14(6):1–10, 2013.

[32] Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5:1–18, 2017.

[33] Gary D Wu, Jun Chen, Christian Hoffmann, Kyle Bittinger, Ying-Yu Chen, Sue A Keilbaugh, Meenakshi Bewtra, Dan Knights, William A Walters, Rob Knight, et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108, 2011.

[34] Yinglin Xia, Jun Sun, Ding-Geng Chen, et al. *Statistical analysis of microbiome data with R*, volume 847. Springer, 2018.

[35] Jian Xiao, Li Chen, Stephen Johnson, Yue Yu, Xianyang Zhang, and Jun Chen. Predictive modeling of microbiome data using a phylogeny-regularized generalized linear mixed model. *Frontiers in microbiology*, 9:1391, 2018.

[36] Yanyan Zeng, Daolin Pang, Hongyu Zhao, and Tao Wang. A zero-inflated logistic normal multinomial model for extracting microbial compositions. *Journal of the American Statistical Association*, pages 1–14, 2022.

[37] Jingru Zhang and Wei Lin. Scalable estimation and regularization for the logistic normal multinomial model. *Biometrics*, 75(4):1098–1108, 2019.

[38] Chao Zhou, Hongyu Zhao, and Tao Wang. Transformation and differential abundance analysis of microbiome data incorporating phylogeny. *Bioinformatics*, 37(24):4652–4660, 2021.

Appendix A

Figures for Evaluation Metrics with 500 Replicates

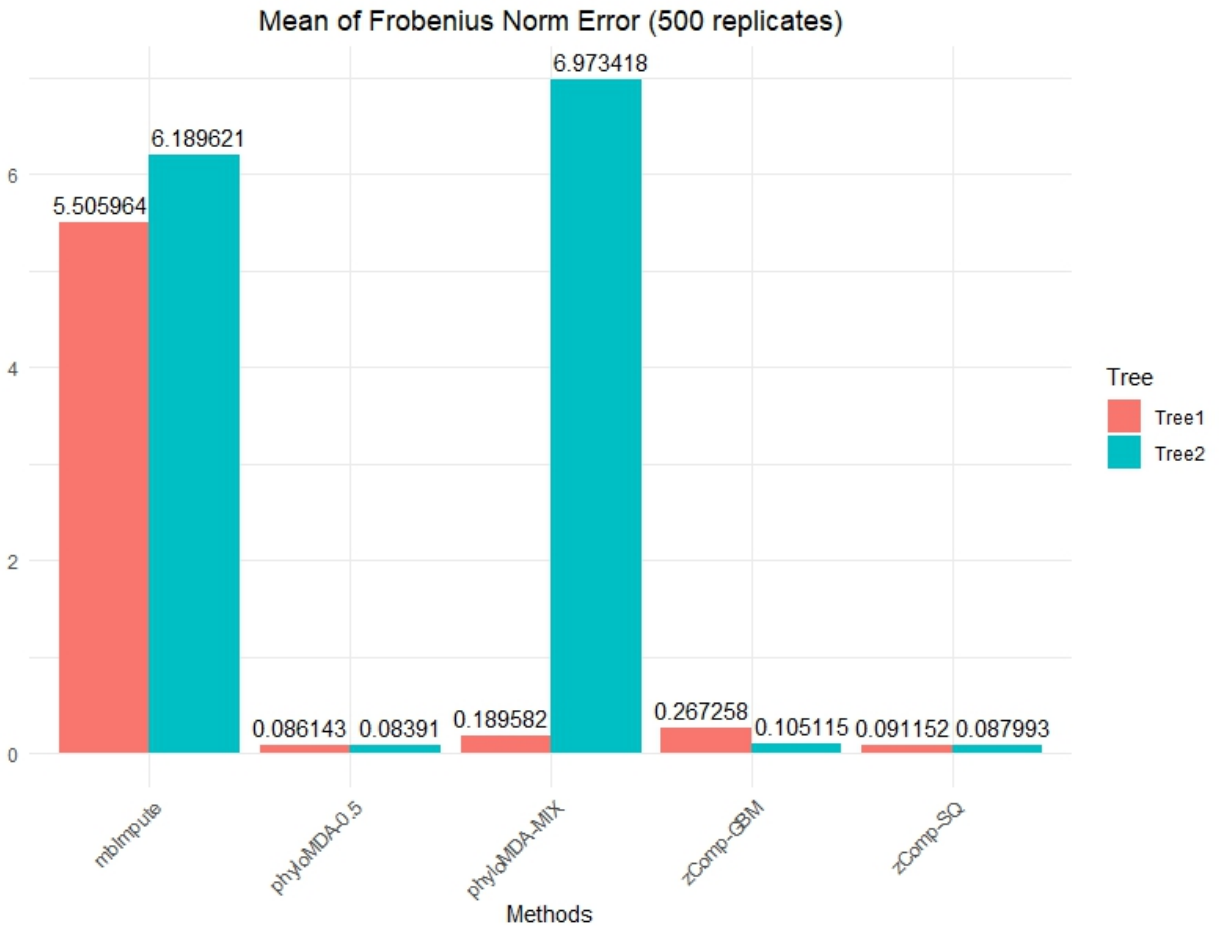

Figure A.1: Means of the Frobenius Norm Error

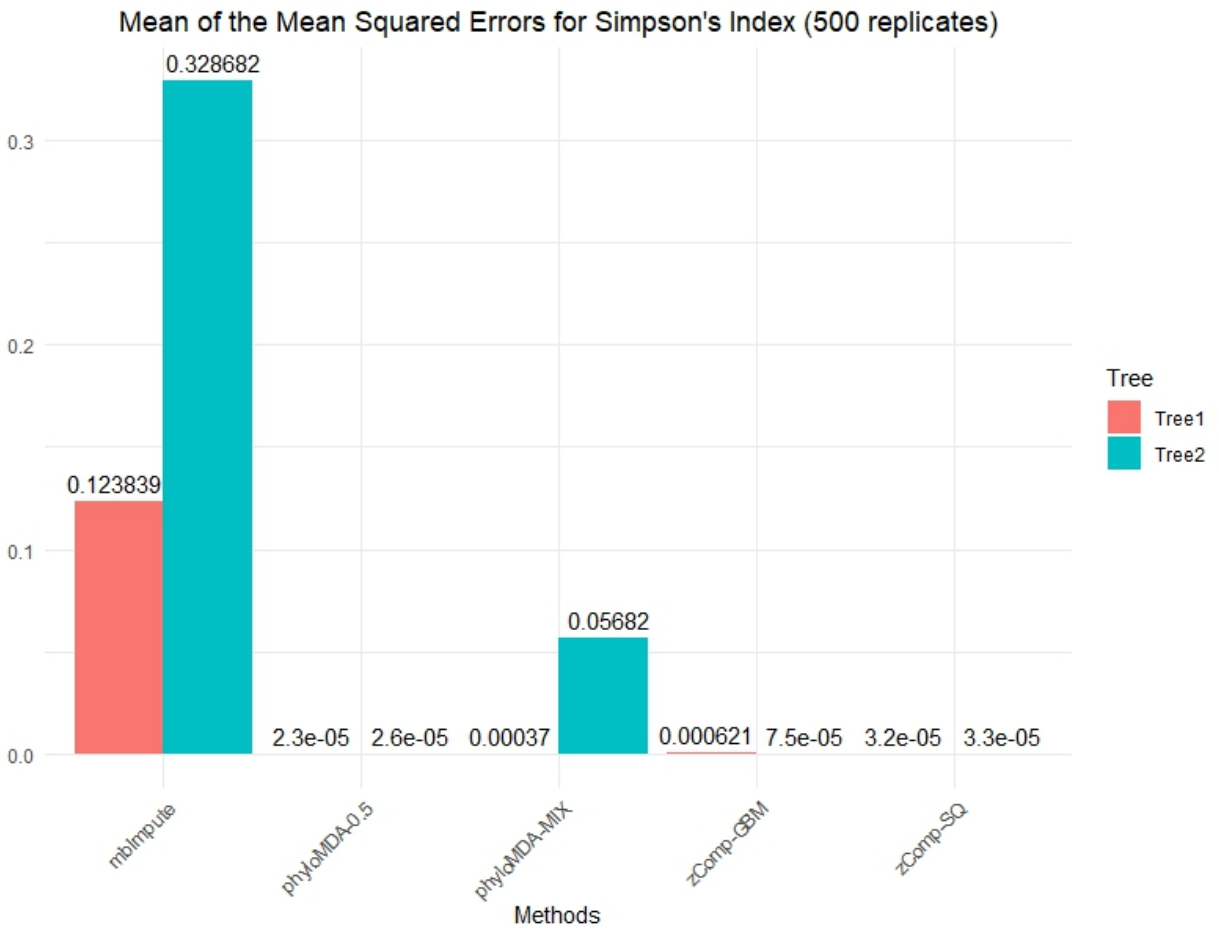

Figure A.2: Means of the MSE for the Simpson's Index

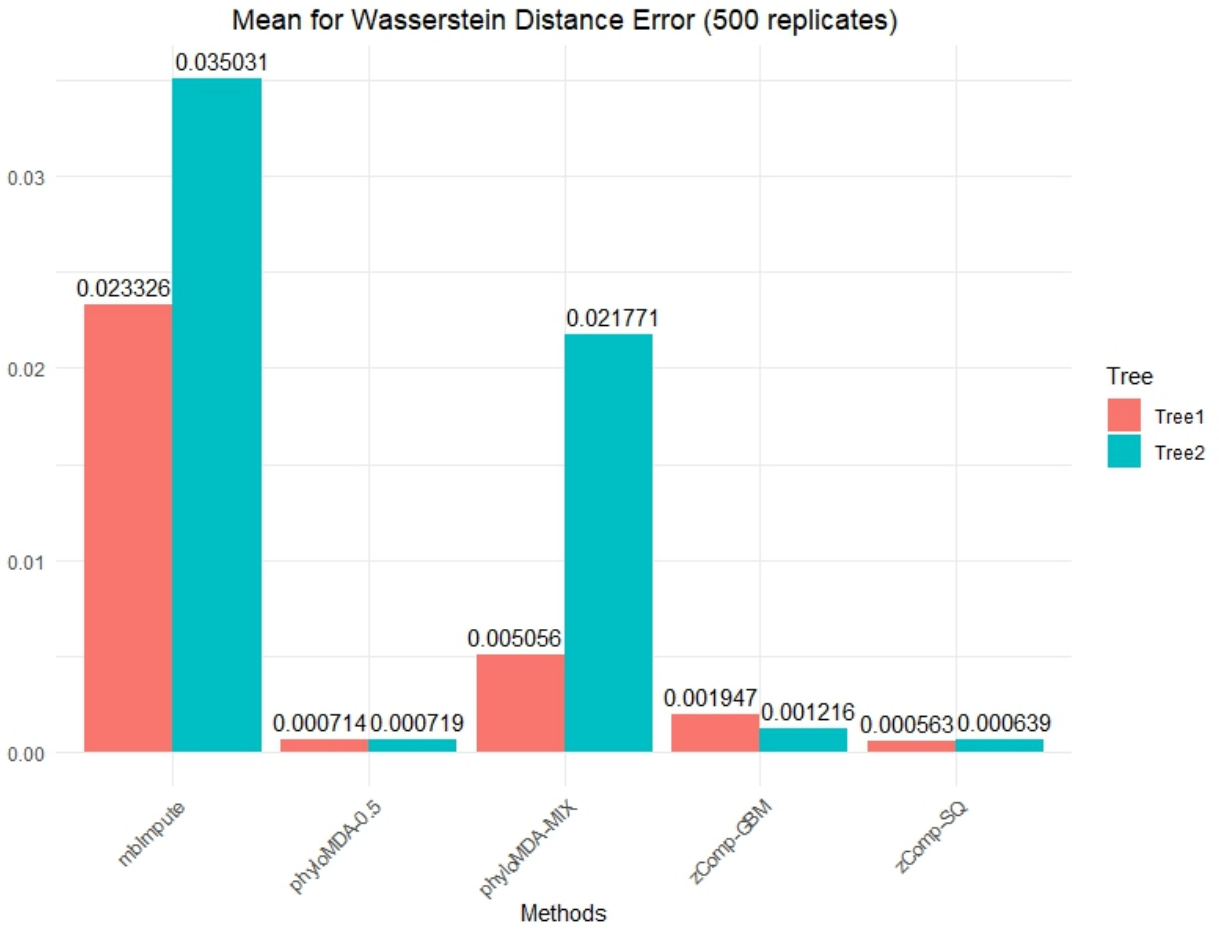

Figure A.3: Means of the Wasserstein Distance

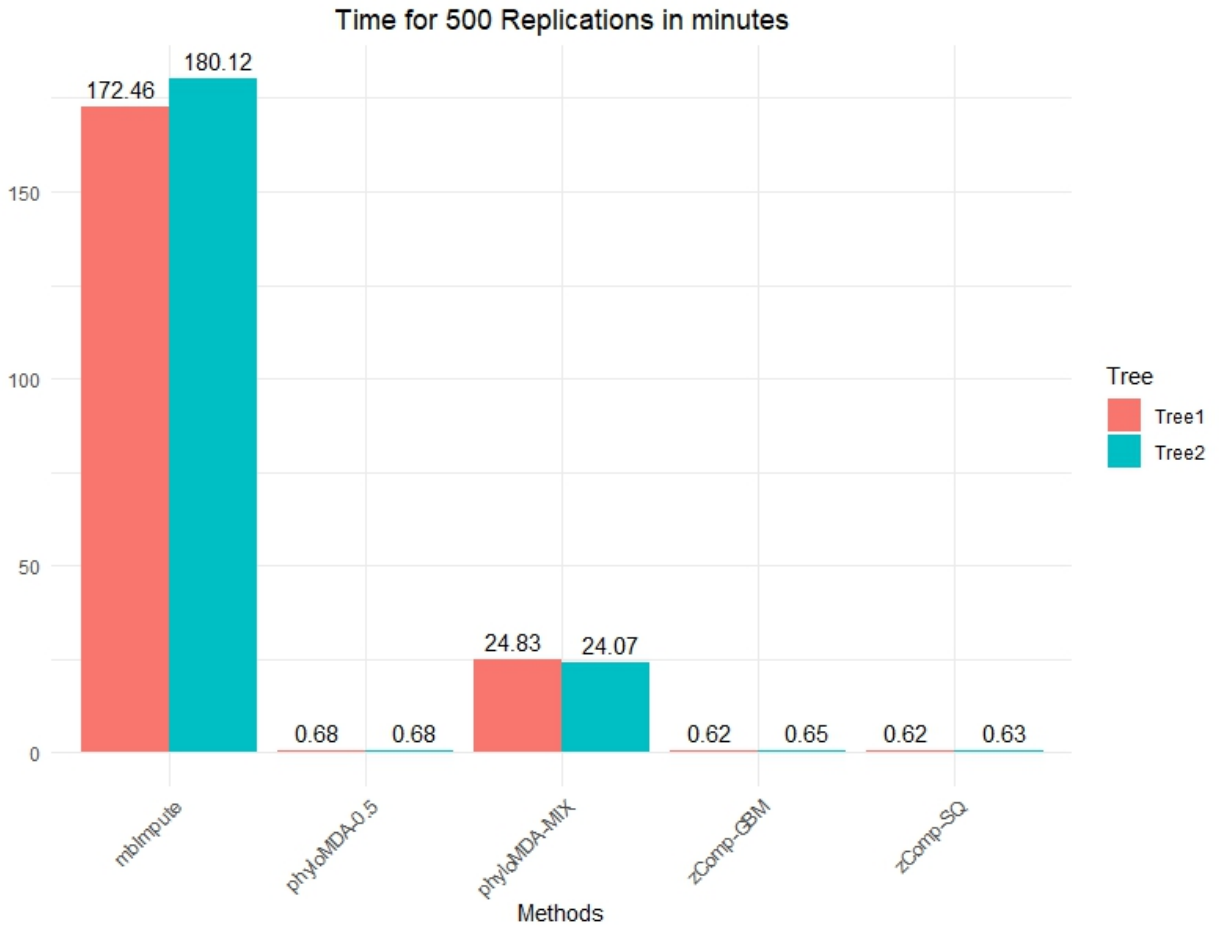

Figure A.4: Time for 500 replications in minutes

Appendix B

Results of ZIPPCAlnm for 500 replicates

	Frobenius Norm Error	MSE for Simpson's Index	Wasserstein Distance Error	Running Time (minutes)
$\boldsymbol{D}_1$	6.20720	0.07214	0.02212	1319.00
$\boldsymbol{D}_2$	7.70557	0.11087	0.02193	2225.83

Table B.1: Results of ZIPPCAlnm for 500 replicates