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

M.Sc. Defense, August 28th, 2023

Huokai Wu

Supervisor

Dr. Juxin Liu

University of Saskatchewan

Outline

Introduction

Microbiome

- The microbiome, comprising the collective genetic material of all microorganisms—including bacteria, fungi, protozoa, and viruses—is indispensable for life sustenance on earth.

Introduction

Microbiome in plant

- The rhizosphere microbiome determines plant health [2].
- Modulates a plant's efficacy in nutrient absorption [3].
- Enhance plant tolerance to abiotic stress and diseases [13].

Introduction

Microbiome in human

- Cause illnesses like gut problems, diabetes, obesity and joint pain [11, 5].
- Initiating various cancers: colorectal, gastric, lung, and oral cancers [7].

Introduction

Challenging features of microbiome data

- Compositional
- Over-dispersed
- Zero-inflated

Motivation

The importance of handling zeros cannot be overstated because almost all different types of downstream analyses (e.g., network analysis and differential abundance analysis) are based on the quality of imputed data. So, not having a proper way to adjust and analyze those zero values can lead to misleading conclusions. This highlights the need for a proper method of addressing zero-inflated problems.

Objectives

My thesis primarily focuses on analyzing microbiome data without covariate information, which is rarely studied in the existing literature.

Types of zero in microbiome data [12]

- Sampling zero: below the detection limit, particularly the rare ones
- Biological zero: a taxon is actually absent

Objectives

Phylogenetic tree [1]

- In evolutionary biology, a phylogenetic tree serves as a visual representation that charts the evolutionary pathways among diverse species or organisms.

Objectives

My thesis aims to compare four recent methods designed to handle zero-inflation problems in microbiome data analysis. The comparison is conducted on a simulation study. In particular, we make use of the phylogenetic distance in our data generation.

Four recent methods

- Martín-Fernández et al.: Bayesian-multiplicative replacement method [8] (zComposition)
- Jiang et al.: Gamma-Normal mixture model [6] (mbImpute)
- Zhou et al.: Zero-inflated Dirichlet tree multinomial model [17] (ZIDTM)
- Zeng et al.: Zero-inflated probabilistic PCA model [16] (ZIPPCA)

zComposition

Advantages:

- Regarding the non-zero values, the ratios of any two imputed value are the same as the original ratios.
- This method can be quickly performed on large microbiome datasets in R.

Limitations:

- All zeros are treated as "sampling zeros."
- The column with all zero values should be removed.

mbImpute

Advantages:

- Distinguish the zeros into the sampling and biological zeros; the biological zeros will not be imputed.
- Enables a taxon to have similarly expected abundances in those samples with close covariates' information.
- Uses the close taxa, all samples and covariates information for the imputation.
- Neither the sample covariate matrix nor the phylogenetic tree is required by the mbImpute method.

mblmpute

Limitations:

- The low abundances of taxa will also be imputed.
- Does not include the identification information for sampling zeros/biological zeros.
- The input microbiome count data must be normalized to the same library sizes.
- Ineffective when the sample covariate matrix is unavailable, leading to slow execution time in R.

ZIDTM

Advantages:

- This method explicitly makes use of the phylogenetic tree information in the modelling.
- Zero-inflated Dirichlet tree multinomial distribution can simultaneously accommodate negative and positive correlations between counts on tree nodes.

ZIDTM

Limitations:

- The phylogenetic tree must be a 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 with the built-in function, this method can not handle the microbiome count data with a large number of taxa.

ZIPPCA

Advantages:

- Without knowing the covariate information.
- The output includes the identification information for sampling and biological zeros.

Limitations:

- Ineffective in situations where the number of taxa is large, leading to slow execution time in R.
- The parameter estimation does not converge when the dataset contains many taxa.

Simulation Study- Notations I

- 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.
- $\boldsymbol{\delta}_i = (\delta_{i1}, \delta_{i2}, \cdots, \delta_{ik})$ to signify the presence of taxa in sample i
- $\delta_{ij_1} = \delta_{ij_2} = \cdots = \delta_{ij_L} = 0$. Here, $\boldsymbol{\pi}_i^* = (\pi_{ij_1}, \pi_{ij_2}, \cdots, \pi_{ij_L})^T$ 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.

Simulation Study- Notations II

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

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

- $\boldsymbol{D}_{k\times k} = (D_{lm})$ represents the phylogenetic distance matrix, where D_{lm} denotes the phylogenetic distance between taxon l and taxon m.
- $\boldsymbol{\theta}_i$ signifies the vector of mean parameters in the multivariate normal distribution
- Variance-covariance matrix for multivariate normal distribution as $\boldsymbol{\Sigma} = (\Sigma_{lm})$, where $\Sigma_{lm} = \sigma^2 \exp\{-2\rho_{lm} D_{lm}\}$
- σ^2 represents the variance component

Simulation Study- Notations III

- $\rho_{lm} \in (0, \infty)$ 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.
- $\mathbf{N}_i$ represents library size in sample i

Simulation Study- Data Generator

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

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

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

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

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

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

Simulation Study- Assumption

- The existence probability p of each taxon is constant across all samples
- Limited to the phylogenetic tree that was constructed using reference data. This implies that our analysis is restricted to taxa annotated on the phylogenetic tree.

Simulation Study- Parameters Setting

- $N = 98$, $K = 62$ and the library sizes for each sample $\mathbf{N}_i$, are borrowed from the "COMBO" dataset [15].
- Phylogenetic distance matrix: "cophenetic.phylo" [10]
  - Phylogenetic distance matrix $\boldsymbol{D}_1$: Sum of actual edge length
  - Phylogenetic distance matrix $\boldsymbol{D}_2$: Number of edges linking two taxa
- $\alpha_j \sim \mathcal{N}(0, 1)$
- $\boldsymbol{\theta}$ are sampled 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
- $\rho = 0.25$ and $\sigma = 1$

Simulation Study- Data Fitting I

- For the Bayesian-Multiplicative replacement method: "cmultRepl" function from the "zComposition" package.
  - The square root multiplicative treatment
  - The geometric Bayesian multiplicative treatment
- For the Gamma-Normal mixture model: "mbImpute" function from the "mbImpute" package.
  - the number of nearest taxa in a phylogenetic tree that used to impute a missing value: 5 (default value)
- Regarding the ZIDTM method: "eBay_comps" function in the "phyloMDA" package.
  - Prior: zero-inflated Dirichlet tree multinomial model
  - Smooth methods: 0.5 and MIX
- Zero-Inflated Probabilistic PCA method: "ZIPPCAlnm" function from the "ZIPPCAlnm" package
  - number of latent variables: 2

Simulation Study- Evaluation Metrics

- Frobenius norm error
  - The Frobenius norm, also known as the Euclidean norm, measures a matrix's size. It's often used to measure the disparity between two matrices [4].
- Mean squared error of Simpson's index
  - Simpson's Index measures habitat biodiversity in ecology, and its mean squared error quantifies the average squared discrepancies between predicted and actual values over various predictions or simulations [14].
- Wasserstein distance error
  - The Wasserstein distance measures the differences between two distributions, and its error is used to measure the accuracy of data recovery between imputed methods and true abundance [9].

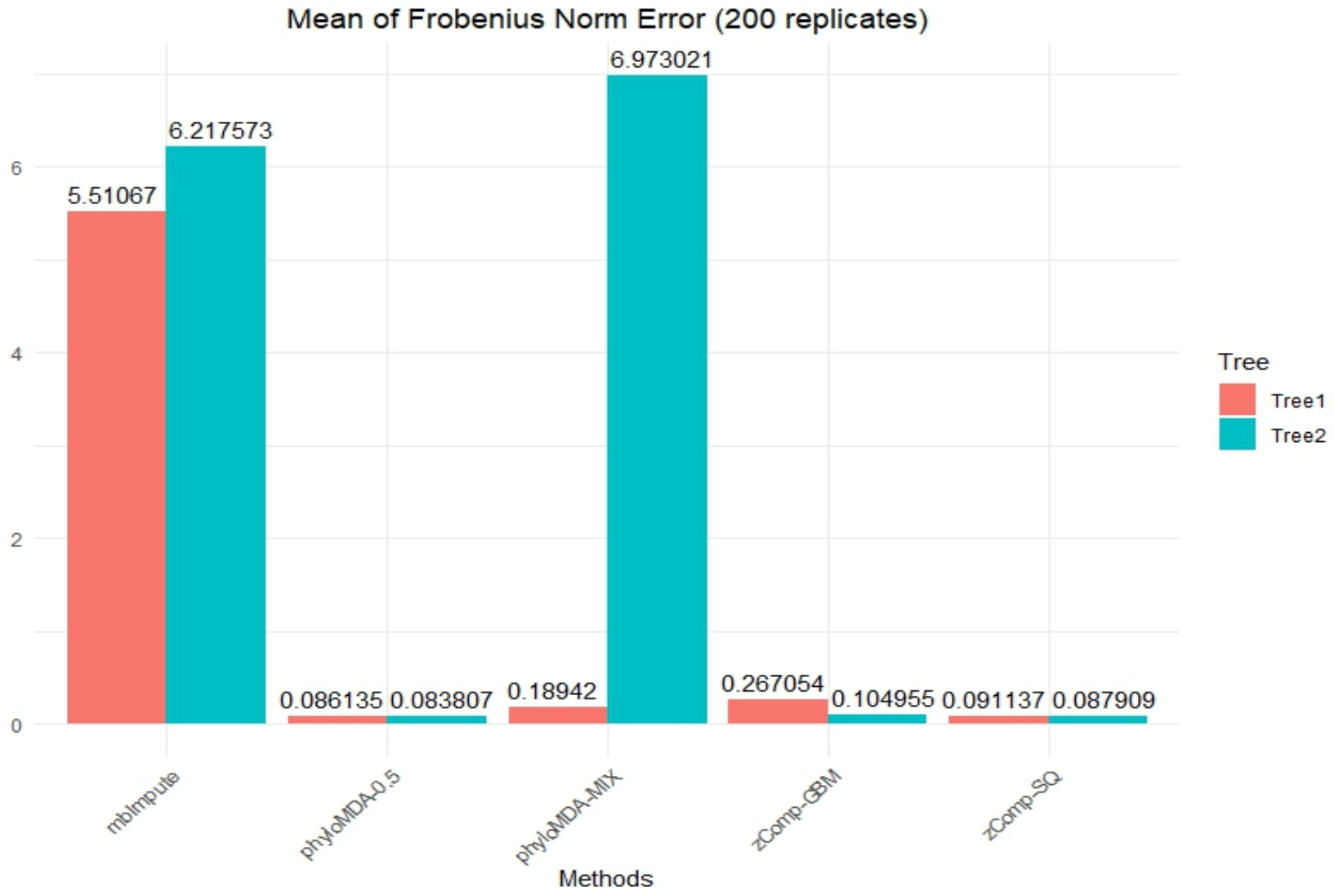
Conclusion I

- 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 settings.
- 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.

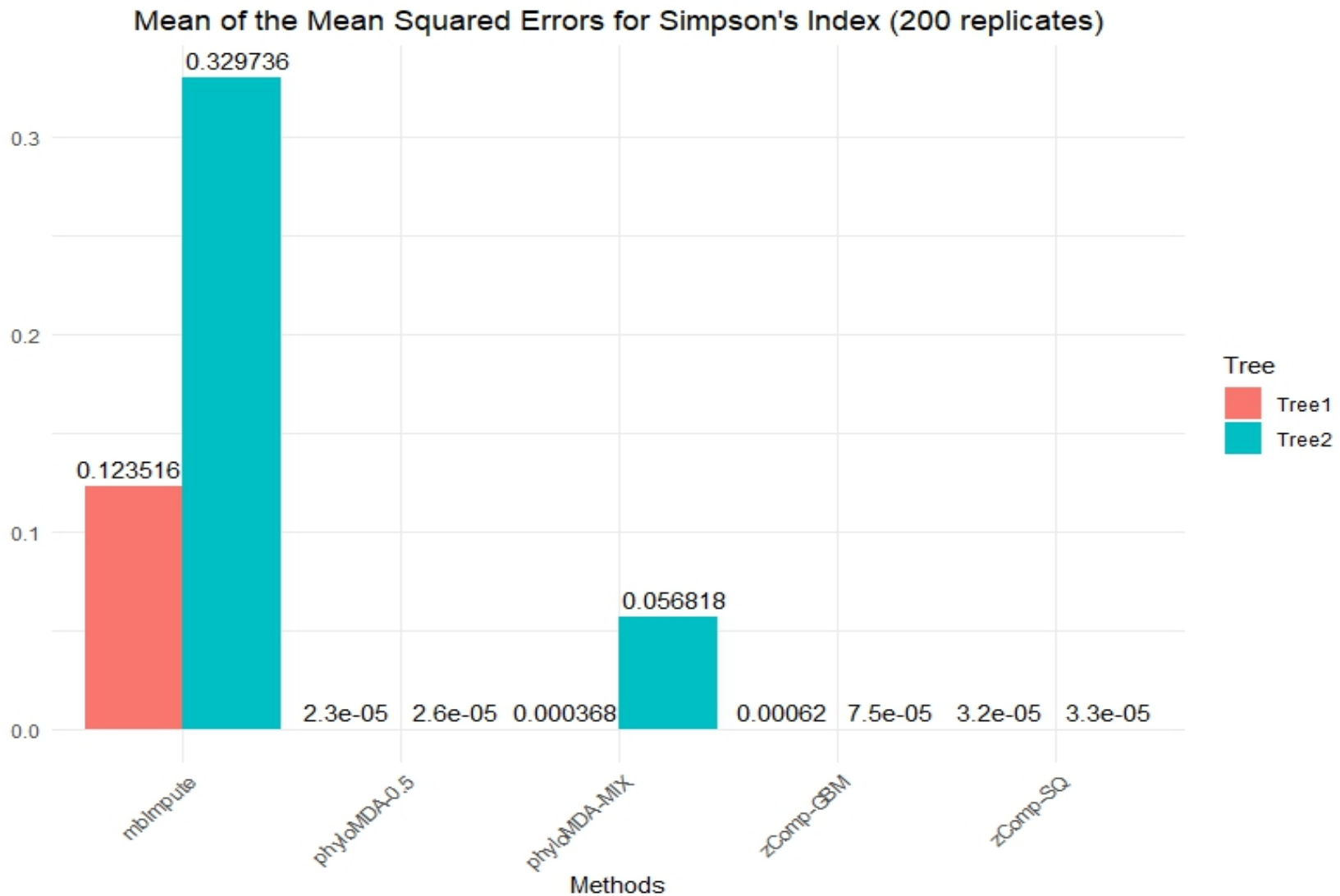
Conclusion II

- 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.

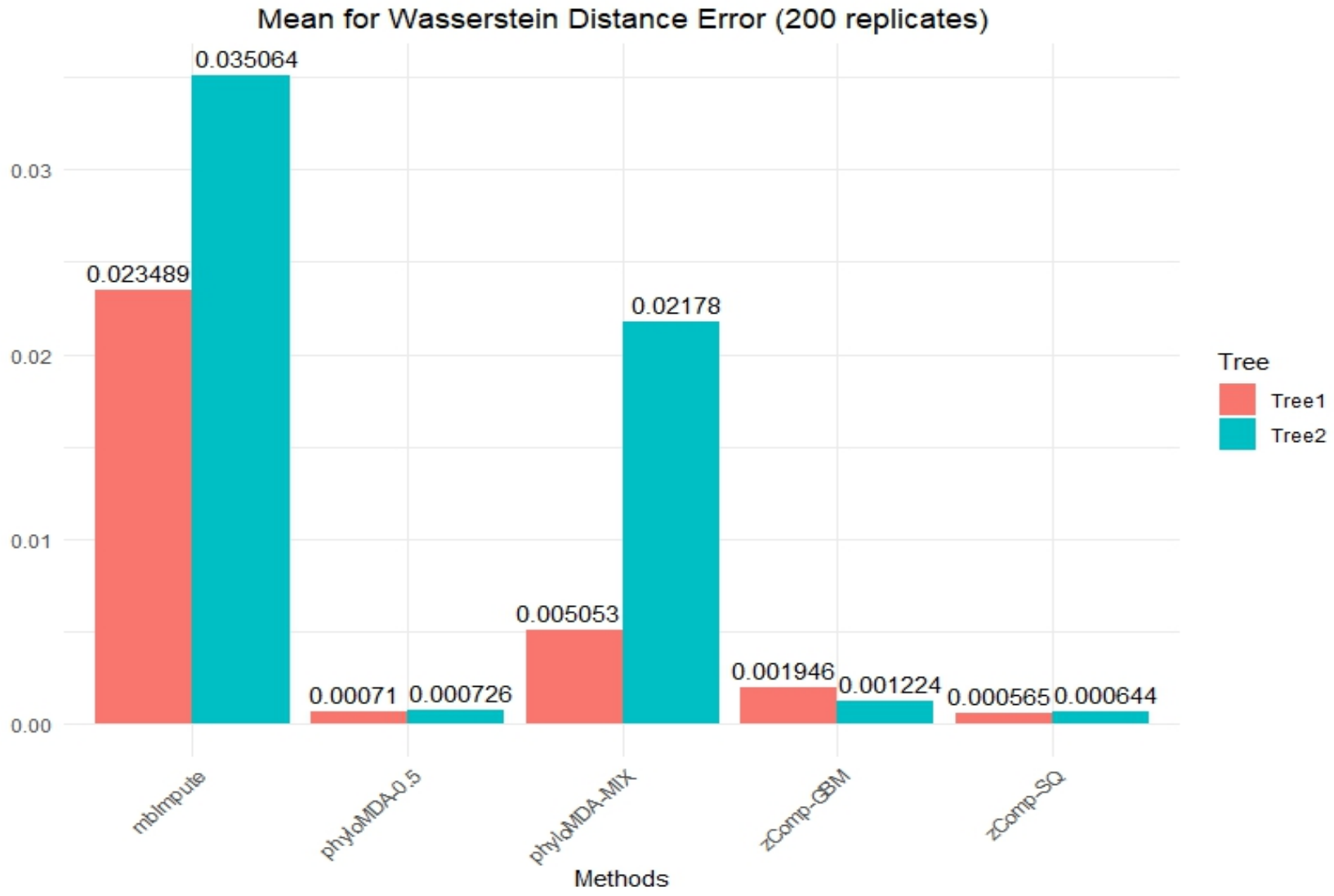
Conclusion-Frobenius norm error



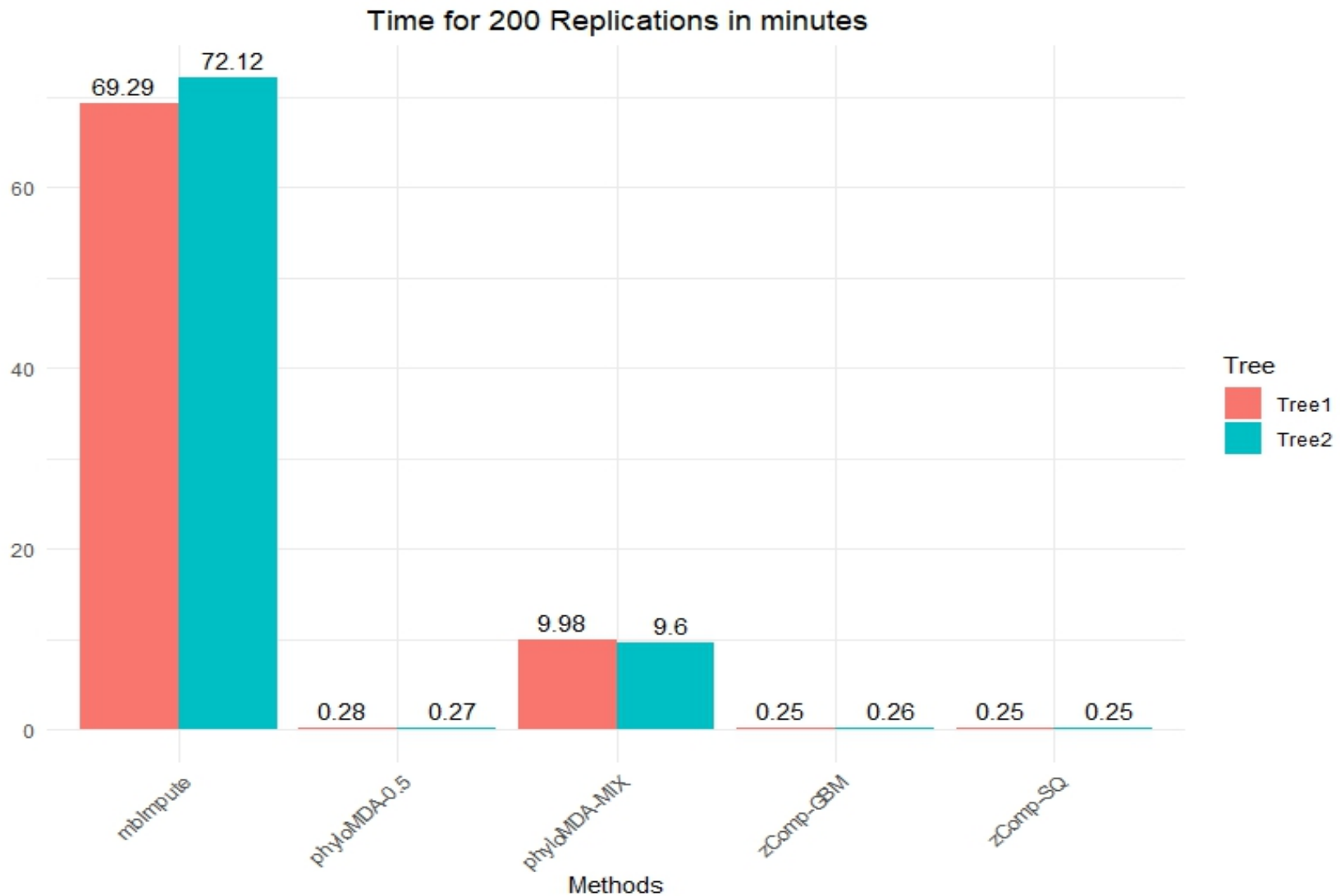
Conclusion-MSE of Simpson's index



Conclusion-Wasserstein distance error



Conclusion-Running Time



Conclusion-ZIPPCAlnm

	Frobenius Norm Error	MSE for Simpson's Index
$\boldsymbol{D}_1$	6.26042	0.07427
$\boldsymbol{D}_2$	7.67705	0.10861

Table: Results of ZIPPCAlnm for 200 replicates

	Wasserstein Distance Error	Running Time (minutes)
$\boldsymbol{D}_1$	0.02198	519.07
$\boldsymbol{D}_2$	0.02191	862.19

Table: Results of ZIPPCAlnm for 200 replicates

Conclusion

- phyloMDA-0.5: Outperforms other methods with the smallest Frobenius norm error and mean square error for Simpson's Index.
- zComp-SQ: Displays promising performance, evidenced by minimal Wasserstein distance error and efficient running time in our simulation study
- ZIPPCAlnm: Does not perform well due to issues with parameter estimation convergence

Acknowledgements

- First and foremost, I extend my sincere gratitude to my supervisor, Dr. Juxin Liu. Her unwavering guidance and generous financial support have been cornerstones of my academic journey.

- I want to express my heartfelt gratitude to the committed members, Dr. Li Xing and Dr. Lingling Jin. Both of your kindness and guidance have been invaluable to me. Further, I wish to extend my thanks to Dr. Shahedul A. Khan. His willingness to serve as the examiner significantly enriched the defence process.

- Lastly, I appreciate Dr. Steven Rayan and Kyla's diligent support in organizing this thesis defence in a tight timeframe.

Thank-you!

Reference I

[1] David Baum et al. "Reading a phylogenetic tree: the meaning of monophyletic groups". In: *Nature Education* 1.1 (2008), p. 190.

[2] Roeland L Berendsen, Corné MJ Pieterse, and Peter AHM Bakker. "The rhizosphere microbiome and plant health". In: *Trends in plant science* 17.8 (2012), pp. 478–486.

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

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

Reference II

[5] Aeshna Gupta, Vijai Singh, and Indra Mani. "Dysbiosis of human microbiome and infectious diseases". In: *Progress in Molecular Biology and Translational Science* 192.1 (2022), pp. 33–51.

[6] Ruochen Jiang, Wei Vivian Li, and Jingyi Jessica Li. "mbImpute: an accurate and robust imputation method for microbiome data". In: *Genome biology* 22.1 (2021), pp. 1–27.

[7] Soumendu Mahapatra et al. "An overview of cancer and the human microbiome". In: *Human Microbiome in Health and Disease-Part A* 191 (2022), p. 83.

Reference III

[8] Josep-Antoni Martın-Fernández et al. "Bayesian-multiplicative treatment of count zeros in compositional data sets". In: *Statistical Modelling* 15.2 (2015), pp. 134–158.

[9] Victor M Panaretos and Yoav Zemel. "Statistical aspects of Wasserstein distances". In: *Annual review of statistics and its application* 6 (2019), pp. 405–431.

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

[11] Prerna Pathak. "Diversity and Dynamics of the Gut Microbiome and Immune Cells". In: *Microbiome in Human Health and Disease* (2021), pp. 53–67.

Reference IV

[12] Justin D. Silverman et al. "Naught all zeros in sequence count data are the same". In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 2789–2798. ISSN: 2001-0370. DOI: https://doi.org/10.1016/j.csbj.2020.09.014. URL: https://www.sciencedirect.com/science/article/pii/S2001037020303986.

[13] Tuesday Simmons et al. "Exploring the root microbiome: extracting bacterial community data from the soil, rhizosphere, and root endosphere". In: *JoVE (Journal of Visualized Experiments)* 135 (2018), e57561.

[14] Edward H Simpson. "Measurement of diversity". In: *nature* 163.4148 (1949), pp. 688–688.

Reference V

[15] Gary D Wu et al. "Linking long-term dietary patterns with gut microbial enterotypes". In: *Science* 334.6052 (2011), pp. 105–108.

[16] Yanyan Zeng et al. "A zero-inflated logistic normal multinomial model for extracting microbial compositions". In: *Journal of the American Statistical Association* (2022), pp. 1–14.

[17] Chao Zhou, Hongyu Zhao, and Tao Wang. "Transformation and differential abundance analysis of microbiome data incorporating phylogeny". In: *Bioinformatics* 37.24 (2021), pp. 4652–4660.

Appendix A

- 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.
- The ZIPPCAlnm method has been excluded from the general comparison, as the majority of the parameter estimations within this method do not converge.
- Also, 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.

Appendix B I

The formula associated with the Frobenius norm error is presented as follows:

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

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 following steps are used in this simulation study to compute the Wasserstein distance error:

Appendix B II

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:

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

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

Appendix B III

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

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