Literature Review

The thesis submission explores various methods of coping with zeros in the microbiome data analysis. This study mainly elaborates on four approaches: the Bayesian-multiplicative replacement model, the gamma-normal mixture model, the zero-inflated Dirichlet tree multinomial model, and the zero-inflated probabilistic PCA model. The simulation study uses a zero-inflated logistic normal multinomial model combined with phylogenetic tree distances. The performance of each method is evaluated using the Frobenius norm error, the mean squared error of Simpson’s exponent, and the Wasserstein distance error.

The whole research is based on biotech called 16S rRNA, which plays a significant role during the translation by providing a structural and functional framework for ribosome assembling and the interaction between mRNA and tRNA. This technology is cheap and fast to make through the polymerase chain reaction (PCR). From the article Huokai writes, that the thesis starts with operational taxonomic units (OTUs) count analysis of microbiome data and compares four recent methods to handle zero-inflation issues through sampling zero and biological zero with phylogenetic tree.

As we all know, zero inflation is still a big challenge in microbiome data analysis. From Sliverman’s study, I noticed that sampling zero happened because the taxa cannot be detected due to the sample's limitation of sequence depth or diversity. However, biological zero means the taxa are truly absent from the studying sample system. In general, the thesis used reference data to create a phylogenetic tree, and the phylogenetic tree will be the frame to analyze the simulated data. The distance between taxa in the phylogenetic tree is the concept that needs to be used to find the biological zero. [[1]](#silverman)

In the first place, Martín-Fernández and his colleagues’ study on the Bayesian-Multiplicative Replacement Method and notice that the Dirichlet Prior Distributions are used as the covariates of multinomial (conjugate) distributions of the multinomial distribution in the Bayesian framework for count data analysis, which as prior distributions of parameters, and to accommodate prior knowledge about count distributions. The multiplicative modifications of the non-zero values, based on CoDa (compositional data analysis) principle are applied to deal with zeros in count data, in particular through Bayesian multiplicative (BM) processing using Dirichlet prior distribution. To align the CoDa principle, different zero replacement results are provided for various parameterizations of Dirichlet Prior Distribution. By combining the Dirichlet Prior Distribution with the multiplicative modification, the Bayesian multiplicative methods provide count values that efficiently handle zero values in count data sets by organizing prior knowledge and data in accordance with the CoDa principles. Among the benefits of using the Bayesian-multiplicative replacement method to solve count zeros in compositional data sets are its ability to use multiplicative modification to avoid distortion caused by other replacement methods, align with CoDa principles to preserve the relative structure of data during zero replacement, and incorporate prior information to improve accuracy. Its drawbacks include its sensitivity to prior knowledge, which can alter study outcomes and interpretation, and its assumption of independence, which may make it inappropriate for data sets that actually could conduct biased estimates. [[2]](#martin)

Secondly, a statistical model for modelling data with both gamma and a normal distribution is the Gamma-Normal Mixture Model. Its typical applications include handling zero-inflation data and capturing the probable distribution of abundance. A method for handling missing values from microbiome abundance and zero-inflation datasets is described in Jiang et al.'s publication. It is dubbed mbImputaion. For example, zero inflation, sparsity, and missing values. The gamma-normal mixture model is also utilized in mbImputation for handling microbiome abundance matrices and computing zero-inflation data with missing values. It provides accurate and robust imputation for missing values, to achieve more reliable downstream analyses (statistical and computational methods used to analyze and interpret microbiome data after preprocessing steps such as differential abundance analysis, quality control, and imputation) which aims to identify differentially abundant taxa between groups, clustering samples based on their microbiome composition. Details of mbImpute method involve running underlying statistical models, using the estimated parameters to generate imputed values for missing or zero-inflated entries in the abundance matrix; and mbImpute would evaluate the performance of imputing by using various metrics to assess the accuracy and robustness of the imputed value. For imputation, mbImpute employs a linear model that incorporates samples and covariates, and the gamma-normal model can discern between sampling and biological zeros. The output of this method does not include identifying information for sampling zeros or biological zeros, and it can impute low abundance taxa and sampling zeros. [[3]](#jiang)

The Zero-Inflated Dirichlet Tree Multinomial (ZIDTM) distribution is a model for the multivariate analysis of microbial abundance data, is used to address zero inflation in microbiome data by modelling the probability of observing zero counts for a taxon in a sample. In Zhou et al.’s paper, they offered a model-based technique for handling zero inflation in microbiome data by extending the DTM to ZIDTM incorporating the phylogenetic tree information. This extension addresses the sparsity and zero-inflation commonly observed in microbiome data. To estimate the parameter of ZIDTM distribution, an expectation-maximization algorithm for maximum likelihood estimation is developed which can be implemented efficiently on a parallel architecture computer. The algorithm helps to solve the high-dimensional matching problem for DTM and ZIDM by incorporating phylogenetics. The fitting algorithm for ZIDTM adaptively determines the level of zero inflation based on the data. Also, they introduce a Bayesian formulation called posterior mean transformation, which converts raw counts into non-zero relative abundances that have sum equal to one and accounts for the compositional nature of microbiome data. This transformation helps to handle zero inflation and allows a more meaningful explanation of the data. [[4]](#zhou)

In Zeng’s paper, they offer a statistical model called zero-inflation probabilistic PCA (ZIPPCA) that focuses on handling the excess zeros in microbiome data. ZIPPCA extends probabilistic PCA from the Gaussian setting to multivariate abundance data and accounts for zero inflation. In ZIPPCA, the model assumes that some of the microbial counts in a sample are structural zeros, and all the others in this sample jointly follow a multinomial distribution with index and vector of probabilities. Excessed zeros are modelled using potential indicator and corresponding probability of zero inflation. To evaluate the composition of microbiome, they also mention an empirical Bayes approach that uses a classification variational approximation algorithm for maximum likelihood estimation. This approach considers the compositional nature of microbiome data and the complex co-occurrence relationships among microbes. [[5]](#zeng)

The simulation study in the thesis can be considered as a comprehensive evaluation of the four methods for handling zero-inflated microbiome data. The study is designed to assess the performance of these methods using various metrics, including the Frobenius norm error, Mean Squared Error (MSE) for Simpson’s Index, and Wasserstein distance error. Based on phylogenetic tree distance, the thesis generates simulated data from the zero-inflated logistic normal multinomial model, offering a reliable foundation for comparing the approaches in controlled conditions. The simulation study offers insightful information about the advantages and shortages of each technique, as well as how effectively they work to solve zero inflation in the interpretation of microbiome data.

Overall, the purpose of the project is to explore the feasibility of handling zero in microbiome data through the simulation of four methods from zero-imputation with the phylogenetic distance by reference data that contains sampling zero and biological zero. Also, our future research could focus on refining the simulating models to more accurately differentiate between sampling zero and biological zero. Therefore, improving the overall analysis of microbiome data, evaluating their comparison of performance, and trying to make it in a more practical scenario.

Reference

[3] 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.

[2] 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.

[1] 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.

[5] 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.

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