Literature Review: Methods for Handling Zero-Inflated Problem in Sequencing Data

Huokai Wu [1], Juxin Liu [1], Yunliang Li [1], Daiqing Huang [2], Kevin Stanley [1], Steve Siciliano [1]

[1] University of Saskatchewan, [2] National Research Council Canada

ABSTRACT

Due to the limitations of modern sequencing technologies (e.g., 16S rRNA sequencing), the sequencing count data are compositional, over-dispersed, and zero-inflated. There has been an extensive amount of work on how to tackle those challenges. Our review focus on methods for handling zeros. The importance of handling zeros cannot be overstated because almost all different types of downstream analyses (e.g., network analysis) are based on the quality of imputed data.

To our best knowledge, there is no most up-to-date review on how to handle zero-inflated problem in sequencing data analysis. Our poster aims to fill this gap.

Here is the outline of the remaining parts of the poster. Section 2 provides a summary of the most popular methods in chronological order. Section 3 presents the data generative model used in the simulation study. Section 4 reports the comparison results. Finally, Section 5 concludes the main findings and discuss some future work.

POPULAR METHODS

There has been a large body of published work on developing methods for handling zeros in sequencing data analyses. Nonetheless, most of the existing methods assume either a common distribution for the true relative abundance of the taxa or the availability of covariates. To name a few, Tang and Chen[6] developed a zero-inflated generalized Dirichlet multinomial (ZIGDM) regression method to accommodate more flexible correlation structures among taxa when covariates are available. Silverman et al. [5] introduced three types of zero generating process: sampling zeros, biological zeros and technical zeros.. The data we are considering can have sample-specific compositions (true relative abundance).

In the following, three recent imputation methods are listed with detailed information.

Martín-Fernández et al. (2015) - zComposition — Bayesian-multiplicative treatment of count zeros in composition data sets [4]

This paper proposes the Bayesian-multiplicative treatment for preserving the ratios between taxa. All zeros are treated as the results of the insufficiently large samples, and all of them are "sampling zeros".

Advantages: (i). Regarding the non-zero values, the ratios of any two imputed value are the same as the original ratios. (ii) There are four methods provided in the packages, and they are geometric Bayesian multiplicative (GBM), square root Bayesian multiplicative (SQ), Bayes-Laplace Bayesian multiplicative (BL) and count zero Bayesian multiplicative (CZM).

Limitations:.(i) When using GBM, practitioners must remove those taxa with less than two positive counts in all samples. (ii) When the number of samples becomes larger, there are negative imputed values by adding a fixed value to the zeros (e.g. using CZM methods).

Jiang et al. (2021) - mbImpute — mbImpute: an accurate and robust imputation method for microbiome data [1]

This paper proposed a gamma-normal mixture model to identify and impute sampling zeros.

Advantages: (i) The method distinguishes sampling zeros from biological zeros and only imputes the former ones. (ii) The imputation borrows information from similar samples, closest taxa based on phylogenetic distances, and metadata information.

Limitations: This method cannot effectively handle cases where meta information is not available. In addition, the output from mbImpute does not include the identification information for sampling zeros/biological zeros.

Liu et al. (2022) - phyloMDA — phyloMDA: an R package for phylogeny-aware microbiome data analysis [3]

This package includes three different modules that can perform modeling for multivariate abundance data, computing the relative abundances using an empirical Bayes method, and tree-based regression models.

Advantages: (i) This method explicitly makes use of the phylogenetic tree information in the modeling. (ii) This package also provides a transformation method incorporating phylogenetic-tree-based prior.

Limitations: (i) The phylogenetic tree must be binary tree and is assumed to be perfectly constructed. (ii) The method only makes use of the tree structure but not the phylogenetic distances. (iii) 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.

SIMULATION STUDY: DATA GENERATION

Notations

- n: Total number of the samples, $n = 53$
- K: Total number of the taxa, $K = 344$
- i: The i-th sample, $i = 1, 2, ..., n$
- j: The j-th taxa, $j = 1, 2, ..., K$
- Y_{ij}: value of observed abundance for taxa j in sample i
- $N_i = (\sum_{j=1}^{K} Y_{ij})$: The library size for sample i.
- $\delta_{ij} = \begin{cases} 1, & \text{if taxon } j \text{ does not exist in the sample } i \\ 0, & \text{otherwise} \end{cases}$
- p_{ij}: The probability that $\delta_{ij} = 1$
- α_{0_j} and α_1: Parameters that can vary the proportion of structural zeros

$$p_{ij} = \frac{1}{1 + \exp\{-(\alpha_{0_j} + \alpha_1 \log N_i)\}},$$

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

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

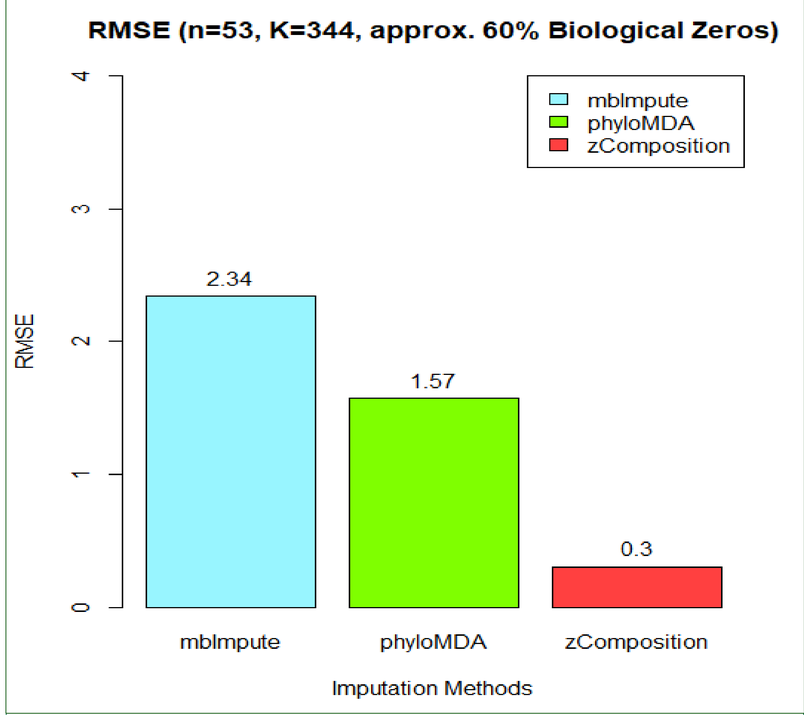
$$\text{if } \delta_{ij} = 0 : (\log \frac{\pi_{ij}}{\pi_{ip}}) \sim MVN(\boldsymbol{\theta_i}, \boldsymbol{\Sigma})$$

$$\pi_{ij} = \frac{\frac{\pi_{ij}}{\pi_{iK}}}{1 + \sum_{j=1}^{K-1} \frac{\pi_{ij}}{\pi_{iK}}} \text{ , and } \pi_{iK} = \frac{1}{1 + \sum_{j=1}^{K-1} \frac{\pi_{ij}}{\pi_{iK}}}$$

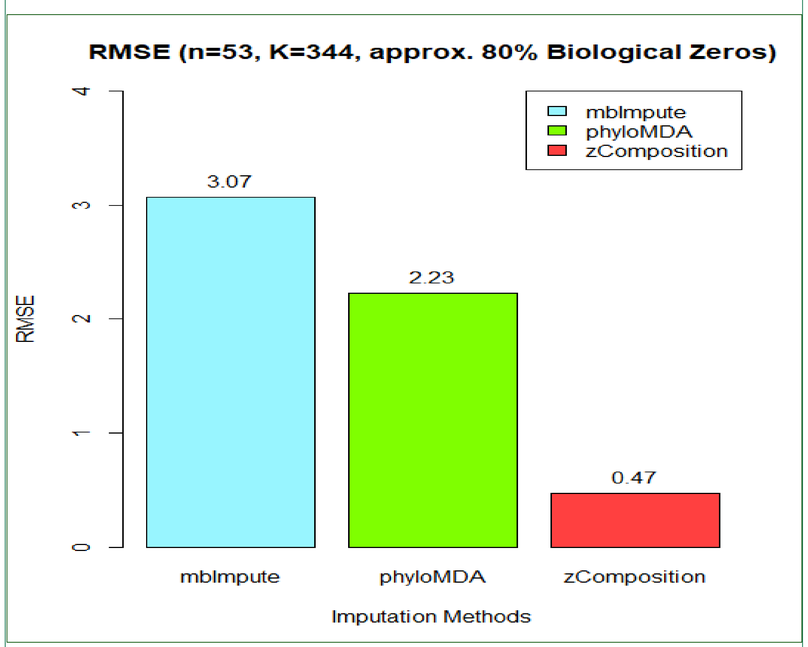
$$\boldsymbol{Y_i} | \boldsymbol{\pi_i} \sim \text{Multinomial}(\boldsymbol{\pi_i}, \boldsymbol{N_i}) \text{ , where } \boldsymbol{\pi_i} = (\pi_{i1}, ..., \pi_{iK})$$

Note:
$(\log \frac{\pi_{ij}}{\pi_{ip}})$ is the log transformation for the proportion of taxon $j \in \{1, .., p-1, p+1, ..., K\}$ given taxa p in sample i.
$\boldsymbol{\theta_i}$ is a set of mean parameters for multivariate normal distribution in sample i.
$\boldsymbol{\Sigma} = (\Sigma_{lm} = \sigma^2 \exp\{-2\rho D_{lm}^2\})$ is the covariance matrix for different taxa, D_{lm} denotes phylogenetic distance between taxa l and taxa m, ρ is the evolution rate

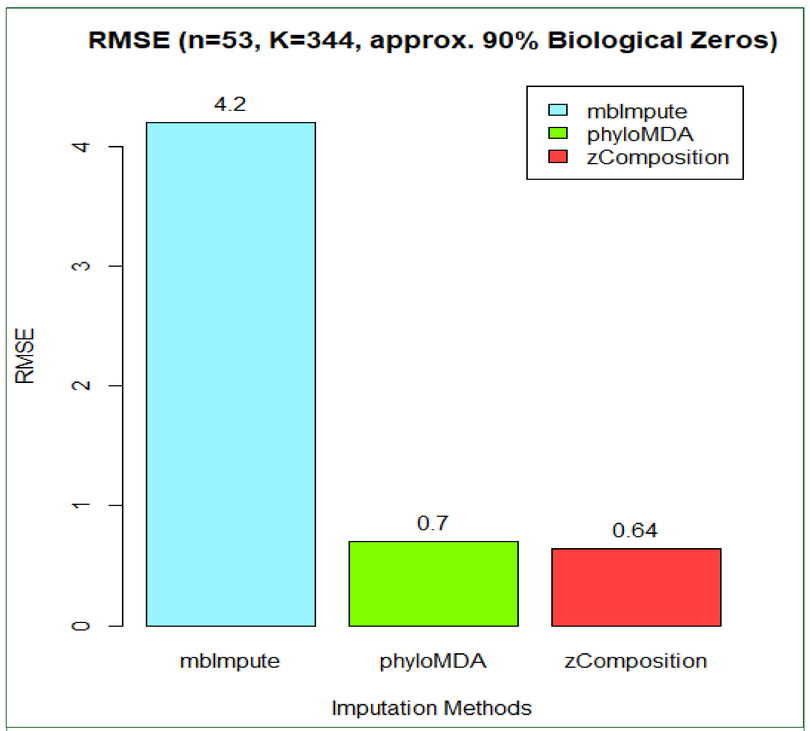
SYNTHETIC DATA: RESULTS



Simulation 1: 53 samples, 344 taxa, and approximately 60% biological zeros. Repeat the data generation for 500 times and get the average root mean square error (RMSE)



Simulation 2: 53 samples, 344 taxa, and approximately 80% biological zeros. Repeat the data generation for 500 times and get the average root mean square error (RMSE)



Simulation 3: 53 samples, 344 taxa, and approximately 90% biological zeros. Repeat the data generation for 500 times and get the average root mean square error (RMSE)

SUMMARY AND FUTURE WORK

Based on our simulation study, zComposition seems to be the winner that produces the smallest RMSE across all scenarios. As a quick note, the data-generating process proposed in our poster can serve as a data simulator for generating microbiome data.

The current simulation study is a starting point and thus is subject to some limitations. We will conduct more systematic simulation experiments by varying the sample size, the number of taxa, and the number of repeatedly generated data sets. For a more comprehensive comparison, more metrics will be considered to compare the results.

REFERENCES

[1] Jiang, R., Li, W. V., & Li, J. J. (2021). mbImpute: an accurate and robust imputation method for microbiome data. *Genome biology*, *22*(1), 1-27.

[2] Kaul, A., Mandal, S., Davidov, O., & Peddada, S. D. (2017). Analysis of microbiome data in the presence of excess zeros. *Frontiers in microbiology*, *8*, 2114.

[3] Liu, T., Zhou, C., Wang, H., Zhao, H., & Wang, T. (2022). phyloMDA: an R package for phylogeny-aware microbiome data analysis. *BMC bioinformatics*, *23*(1), 1-6.

[4] Martín-Fernández, J. A., Hron, K., Templ, M., Filzmoser, P., & Palarea-Albaladejo, J. (2015). Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling*, *15*(2), 134-158.

[5] Silverman, J. D., Roche, K., Mukherjee, S., & David, L. A. (2020). Naught all zeros in sequence count data are the same. *Computational and structural biotechnology journal*, *18*, 2789-2798.

[6] Tang, Z. Z., & Chen, G. (2019). Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*, *20*(4), 698-713.

ACKNOWLEDGEMENT

The first co-author would like to thank for the scholarship from P2IRC Flagship 2.

All the co-authors are thankful for the contribution from Elijah Shurmer at the early stage of the project.

This research was undertaken thanks in part to funding from the Canada First Research Excellence Fund.