Simulating Realistic Correlation Patterns in Microbiome Data with Different Statistical Methods

Chen Yutong
Department of Math and Statistics, York University
Summer 2021 DURA
Professor Kevin McGregor

January 4, 2022

Abstract

Due to the closely linkage between the various disease and gut microbiome data, there is an increasing research on how to generate simulated microbiome data in order to realistically simulate the correlation pattern in microbiome data. This report will focus on reviewing different statistical methods, namely Normal to anything, MetaSPARSim, Sparse_DOSSA2 in simulating realistic correlation patterns, which are known to be related to metabolic interactions between taxa.

1 Introduction

The human microbiome consists of the collection of naturally occurring microorganisms inhabit in the human body. The composition of the human microbiome is known to have a close link to various diseases such as asthma, obesity, and inflammatory bowel disease. One way to measure the composition of the microbiome is by sequencing the 16S ribosomal subunit of the microorganisms. The development of statistical methodology for microbiome 16S sequencing data relies on algorithms designed to simulate realistic microbiome datasets.

One important challenge in simulating microbiome data is that this type of data is compositional, meaning that we need to apply appropriate transformations in order to interpret the simulated microbiome data. In addition to the compositional nature of the data, there are also difficulties encountered in the statistical distribution of the data; namely, zero inflation and overdispersion in the counts of different taxa. In this project, we review three recently developed statistical methods for simulating realistic microbiome data. In particular, we focus on simulating realistic correlation patterns, which are known to be related to metabolic interactions between taxa.

2 Methodology

In this report, we mainly using three statistical methods in practicing simulating microbiome data and testify which one gives you the best correlation pattern. We are using the same real microbiome data from the 2014's research paper done by Nielsen as our input. Then we will learning this true data through our three statistical methods and see which simuluated dataset gives you the best correlation pattern between the different columns and which has the best marginal species distributions compared to the true input dataset.

Here are the break down of the three methods we have.

2.1 Normal to Anything

Normal to anything (NORTA) is a very well-known, one of the most bench-standard methods in simulating data. It generating the correlation matrices of normal random vectors from a multivariate normal distribution and transforms to the zero-inflated negative binomial distribution. (Kurtz et al. 2014)

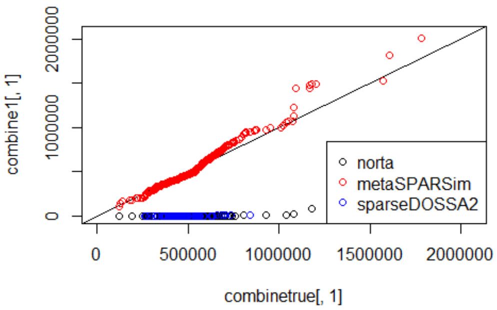


Figure 1: Qq plot for species 1 with comparing three methods' simulated data

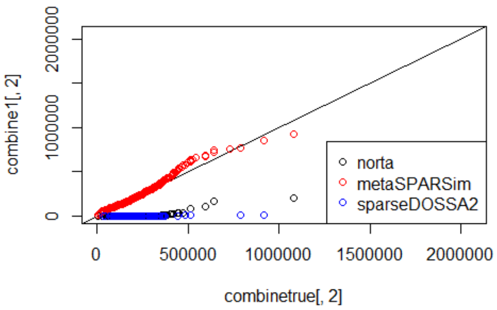


Figure 2: Qq plot for species 4 with comparing three methods' simulated data

2.2 MetaSPARSim

MetaSPARSim is a recent discovery method discovered for simulating dataset in 2019. The dataset is simulatesd based on the multivariate hypergeometric distribution. However, it does not explicitly address correlation structure according to research by Patuzzi et al in 2019.

2.3 SparseDOSSA2

SparseDOSSA2 is the most updated methods discovered in 2021 by Ma et al. This methods simulates based on the zero-inflated log-normal distribution which also address correlation structure in the simulated dataset.

3 Results

Here are the results we have for how well these simulated microbiome dataset turn out to be through these three statistical methods. We have evaluated by two approaches: first one is by using R to simulating marginal species count distributions. The second one is to compare and see the correlation table to find which method's simulated correlation is the most alike to the true correlation.

After examining all the rows and columns, we discover the most abundant species and form these graphs below with scattered plot of the simulated dataset by these three statistical methods against the true dataset. The most abundant species are species 1, species 4, species 5, species 15, species 16, species 33, species 34, species 57, species 60, species 97. Thus, we only examine these columns as this gives us the most significant result. The diagonal line represents the true dataset, so the closer to the diagonal line, the better marginal distribution and marginal counts of this simulated dataset by that statistical method. The results of QQ plots of simulated data by these three methods are shown in the graphs for different species number below.

Besides the QQ plots, we also compose a correlation table below to show directly the comparison of the true correlation value against the correlation value of the simulated data by the three statistical methods. We compose the table by selecting the top 5 most abundant species among the 10 species mentioned early to show the significance of the findings. These species are species 1, species 4, species 5, species 16, species 34. So we group any 2 species among these 5 species and create this 5 by 5 compositions of different pairs of species. The highlighted part among each row of the correlation table shows the most realistic value that closest to the true correlation. We can see, overall out of 25 test cases, MetaSPARSim method has generated 13 sets that most similar to the true correlation whereas both normal to anything and Sparse_DOSSA2 have 11 sets that corresponding to true correlation.

4 Conclusion

1. We find that MetaSPARSim performs best for simulating marginal species count distributions. We saw from the scatter plot, the simulated data generate from the MetaSPARSim are the most

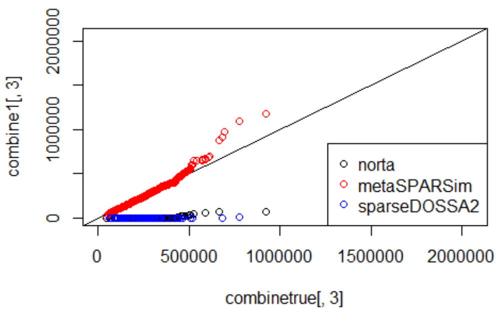


Figure 3: Qq plot for species 5 with comparing three methods' simulated data

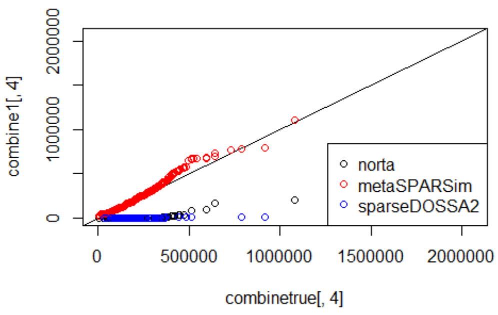


Figure 4: Qq plot for species 15 with comparing three methods' simulated data

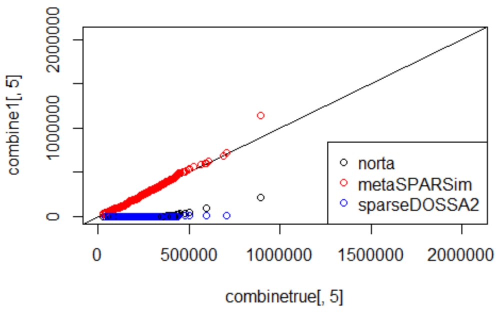


Figure 5: Qq plot for species 16 with comparing three methods' simulated data

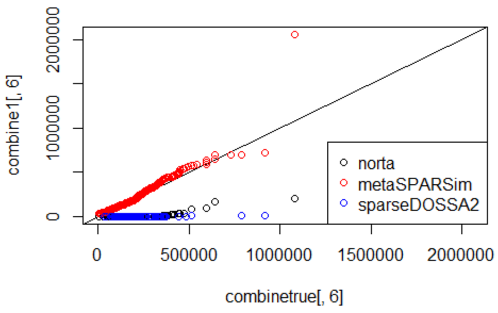


Figure 6: Qq plot for species 33 with comparing three methods' simulated data

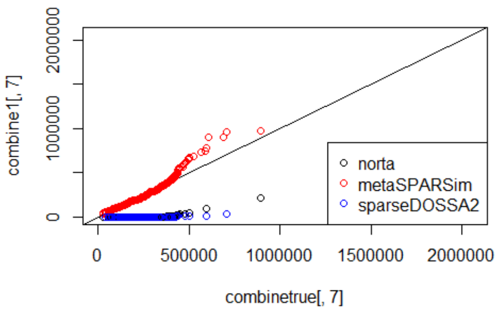


Figure 7: Qq plot for species 34 with comparing three methods' simulated data

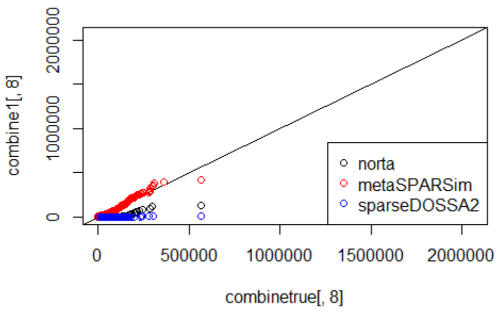


Figure 8: Qq plot for species 57 with comparing three methods' simulated data

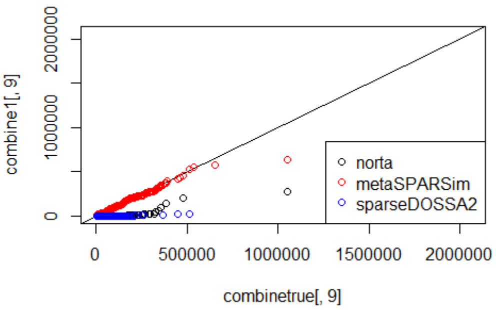


Figure 9: Qq plot for species 60 with comparing three methods' simulated data

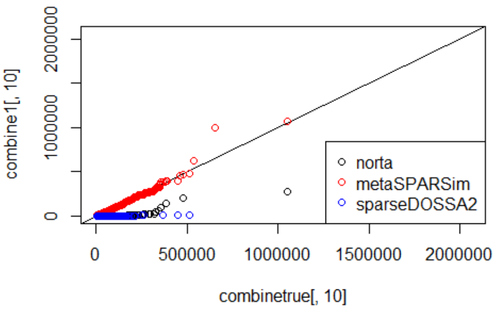
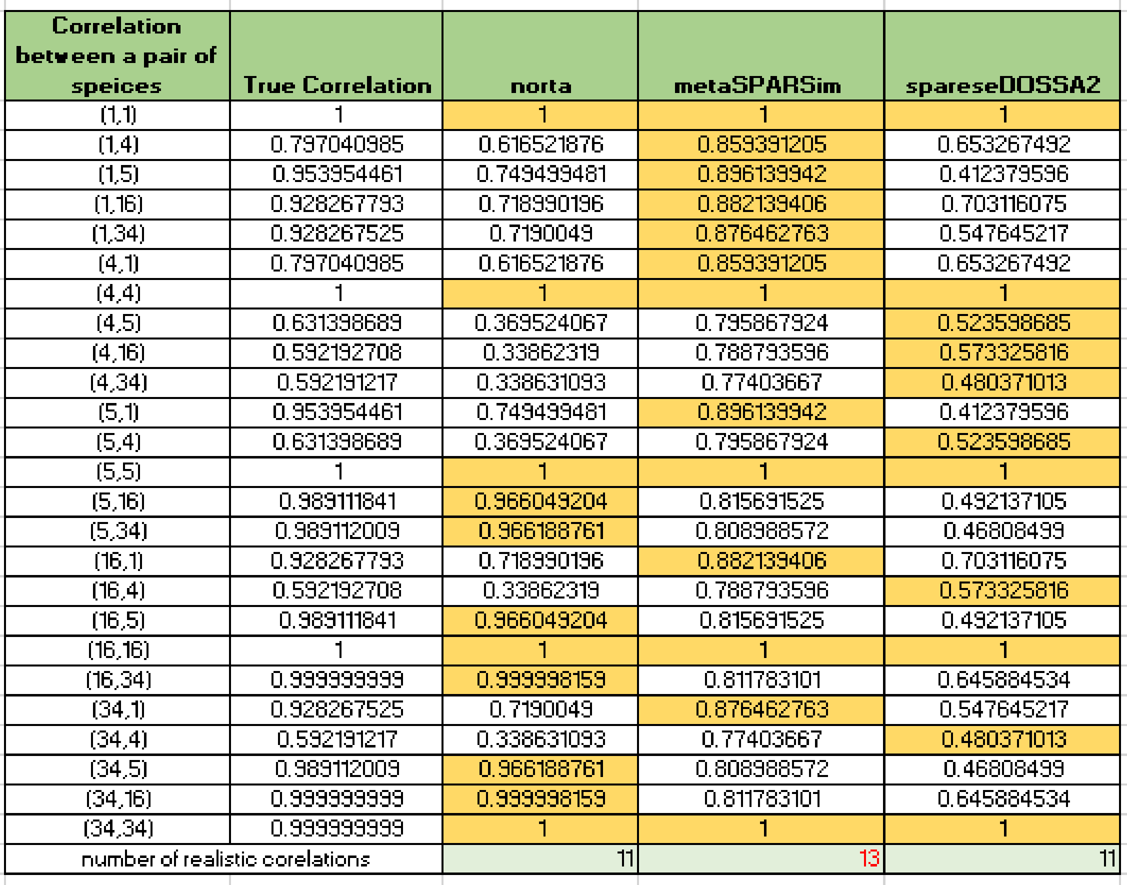


Figure 10: Qq plot for species 97 with comparing three methods' simulated data



Correlation between a pair of speices	True Correlation	norta	metaSPARSim	spareseDOSSA2
(1,1)	1	1	1	1
(1,4)	0.797040985	0.616521876	0.859391205	0.653267492
(1,5)	0.953954461	0.749499481	0.896139942	0.412379596
(1,16)	0.928267793	0.718990196	0.882139406	0.703116075
(1,34)	0.928267525	0.7190049	0.876462763	0.547645217
(4,1)	0.797040985	0.616521876	0.859391205	0.653267492
(4,4)	1	1	1	1
(4,5)	0.631398689	0.369524067	0.795867924	0.523598685
(4,16)	0.592192708	0.33862319	0.788793596	0.573325816
(4,34)	0.592191217	0.338631093	0.77403667	0.480371013
(5,1)	0.953954461	0.749499481	0.896139942	0.412379596
(5,4)	0.631398689	0.369524067	0.795867924	0.523598685
(5,5)	1	1	1	1
(5,16)	0.989111841	0.966049204	0.815691525	0.492137105
(5,34)	0.989112009	0.966188761	0.808988572	0.46808499
(16,1)	0.928267793	0.718990196	0.882139406	0.703116075
(16,4)	0.592192708	0.33862319	0.788793596	0.573325816
(16,5)	0.989111841	0.966049204	0.815691525	0.492137105
(16,16)	1	1	1	1
(16,34)	0.999999999	0.999998159	0.811783101	0.645884534
(34,1)	0.928267525	0.7190049	0.876462763	0.547645217
(34,4)	0.592191217	0.338631093	0.77403667	0.480371013
(34,5)	0.989112009	0.966188761	0.808988572	0.46808499
(34,16)	0.999999999	0.999998159	0.811783101	0.645884534
(34,34)	0.999999999	1	1	1
number of realistic corelations		11	13	11

Figure 11: corelation table for the top 5 abundant species

closest around the diagonal line which represent the true data.

2. We also discover MetaSPARSim also performs best overall in simulating realistic correlations between species. We saw from the correlation table that MetaSPARSim has the highest frequency that its simulated data is closer to the true correlation, the second ranking which is also a tie to the third ranking is simulated data generated from Normal to anything and Sparse_DOSSA2. Normal to anything has generated better correlations between species for true correlation closer to 1. MetaSPARSim has generated better correlations between species for true correlation between 0.95 to 0.7. Moreover, Sparse_DOSSA2 has generated better correlations between species for

true correlation less than 0.7.

5 Explanation and further research

Normal to Anything has been popular for the past decade for its accuracy to generate realistic simulated dataset, we can still see it is still relevant in our research especially for the true correlation that close to 1. Thus, normal to anything is still a good validation for us to use to generate for simulated dataset.

Moreover, it is not surprising to see that we do have MetaSPARSim as the statistical methods for having the best simulated data. The reason we concluded is the other two methods may have over estimate the number of zeros to be simulated which over focused on the zero inflation parameter that simulate too many zero or small values than needed. Thus, causing the simulated data generated by MetaSPARSim be the methods that generate the most realistic simulated data.

still thinking some reason why Sparse_DOSSA2 has generated better correlations between species for true correlation less than 0.7.?

Furthermore, for the future, we plan to testify more statistical methods that recently discussed more such as the Mb_Gan methods that used in python to simulating realistic dataset to see whether maybe that method also a good method to simulated realistic data for correlation of certain range.