Supplementary Materials

Quang P. Nguyen

November 11, 2020

1 Simulation Design

We simulated microbiome relative abundance data with flexible correlation structures using the Normal to Anything (NorTA) approach [1]. This technique can generate arbitrary multivariate continuous distributions from a target correlation structure and marginal univariate distribution. Given an n by p matrix of values $\mathbf{U}$ sampled from multivariate normal distribution with correlation matrix ρ, we can generate target microbiome count vector $\mathbf{X_i}$ for taxa i following distribution $\mathbf{P}$ characterized by the cumulative distribution $\mathbb{F}$:

$$X_i = \mathbb{F}^{-1}(\Phi_{U_i})$$

The negative binomial distribution was chosen as it has been shown empirically to replicate the overall distribution of non-zero microbiome count data [2]. In order to simulate data as close as possible to those collected in the field, we performed maximum likelihood fit (using *fitdistrplus* in R) of the negative binomial distribution to each OTU from stool 16S rRNA sequencing data from the Human Microbiome Project (HMP), accquired through via the *HMP16S* package in R [3]. Figure 2 shows the distribution of the size and mean parameters for the negative binomial distribution.
To control for overall sparsity, we randomly sample from the count matrix with probability $1 - p_0$ where p_0 is the overall desired level of sparsity. Differentially abundant taxa were generated with elevated means with the effect size as the multiplier. We assumed that the overall inter-taxa correlation follows an exchangable structure with correlation ρ_{ij}.

1.1 Hypothesis testing at the sample level

To test for hypothesis testing at the sample level, we simulate microbiome counts with marginals generated from a negative binomial distribution with mean 3.05 and size 1.67 which are the median values of fitted mean and size parameters from the HMP 16S stool data. For each experiment, the enrichment of one set will be evaluated.

- For simulations to evalute type I error control, we simulated 1 data set of 20,000 samples each with 1,000 taxa per sample with no taxa being significantly enriched for each simulation condition. Models were evaluated on a candidate set with varying sizes (50, 100, 150, 200). We also varied the overall added sparsity (0.2, 0.4, 0.6, 0.8) and the degree of inter-taxa correlation within the set (0, 0.2, 0.5).
- For simulations to evaluate power, we simulated 1 data set of 20,000 samples each with 1,000 taxa per sample for each simulation condition. Models were evaluated on a candidate set with set size of 100 taxa, enriched across all samples under different effect sizes (1.5, 2, 2.5, 3). We also varied the overall added sparsity (0.2, 0.4, 0.6, 0.8) and the degree of inter-taxa correlation within the set (0, 0.2, 0.5).
- For simulations of power and AUC classification capacity, we simulated 10 data sets of 1,000 samples each with 1,000 taxa for each simulation condition. We evaluated enrichment for one candidate set with size of 50 taxa per set. per sample with one set of 50 taxa being significantly enriched in either all samples (power evaluation) or half the samples (AUC evaluation). We varied the overall sparsity (0.2, 0.4, 0.8), the correlation (0, 0.2, 0.5) and the effect size (1.5, 2, 2.5, 3).

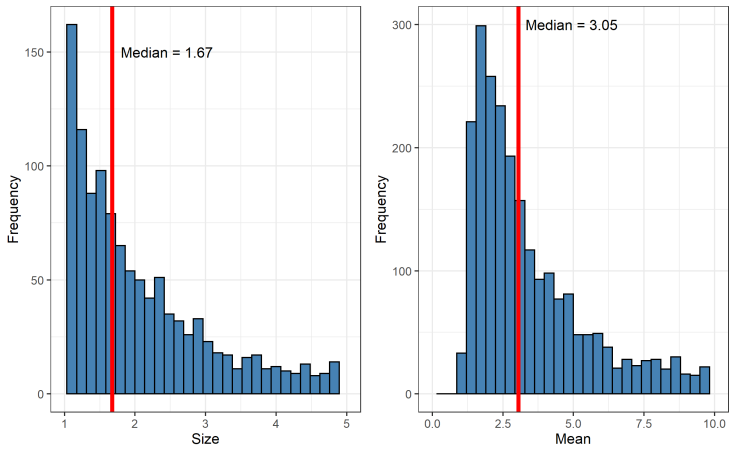

Figure 1. Distribution of each parameter of the zero inflated negative binomial distribution fitted to non-zero values for each OTU in stool samples from HMP data set. The parameters are size (panel A) and mean (panel B).

2 Distribution of cILR

We investigated the distribution of the cILR statistic under the taxa-permuted null. Under parametric simulations, we observed that the

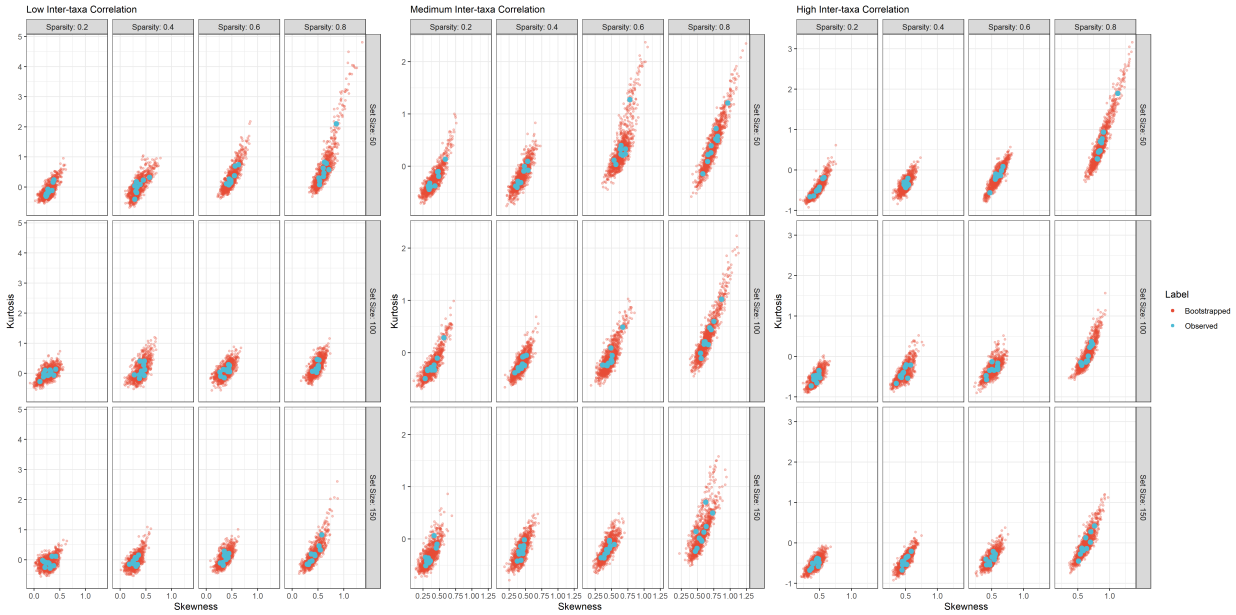

Figure 2. Kurtosis and Skewness

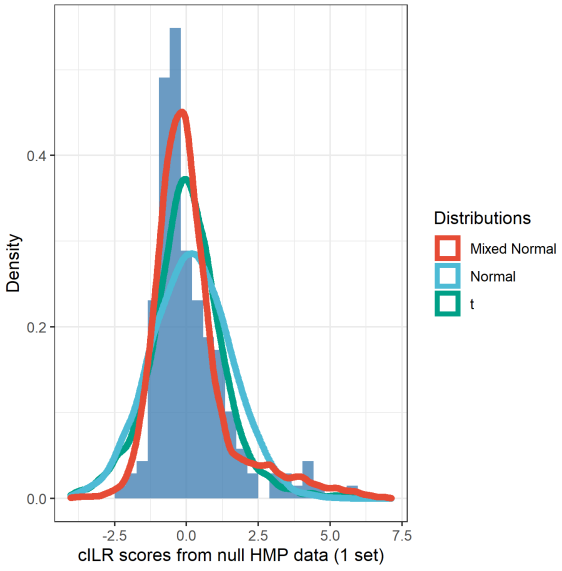

Figure 3. The distribution of clLR statistic under the taxa-permuted null. We compared the different fit for the mixture Gaussian distribution, the Gaussian distribution and the t-distribution. The t and Gaussian distribution was fitted using maximum likelihood while the mixture Gaussian was fitted using Expectaiton Maximization (EM)

References

[1] Marne C Cario. Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix. page 19.

[2] Zachary D. Kurtz, Christian L. Müller, Emily R. Miraldi, Dan R. Littman, Martin J. Blaser, and Richard A. Bonneau. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Computational Biology*, 11(5):e1004226, May 2015.

[3] Lucas Schiffer, Rimsha Azhar, Lori Shepherd, Marcel Ramos, Ludwig Geistlinger, Curtis Huttenhower, Jennifer B Dowd, Nicola Segata, and Levi Waldron. HMP16SData: Efficient access to the human microbiome project through bioconductor. *American Journal of Epidemiology*, 2019.