1 Introduction

The objective is to identify different sources of zero (structural zeros vs sampling zeros) by making use of the phylogenetic tree information.

It has been long recognized the challenges to analyzing microbiome data due to the following reasons:

- a large proportion of zero ($\geq 90\%$);
- compositional nature induced by the measurement tool (i.e., fixed total read counts in 16s rRNA sequencing methods);
- high-dimensionality, i.e., a large number of taxa (typically larger than the sample size).

Different from the existing methods, we consider the following data structure. Covariates are *not* available because we are trying to mimic the real dataset collected by Professor Siciliano's team. Please be noted that the model frame can be easily extended to a data structure including covariates.

2 Data Structure

Let n denote the total number of samples and K denote the total number of taxa.

- OTU count matrix, denoted by $\mathbf{Y} = (Y_{ij}), i = 1, \ldots, n; j = 1, \ldots, K$.
- Distance matrix, denoted by $\mathbf{D} = (d_{jk}), j, k = 1, \ldots, K$, where d_{jk} denotes the phylogenetic distance between taxa j and k.

3 Proposed models

3.1 Model for generating synthetic data

Define

$$\delta_j = \begin{cases} 1, & \text{if taxon } j \text{ does not exist in the ecological community,} \\ 0, & \text{otherwise.} \end{cases}$$

Suppose we know taxon L must exist (say based on the emiprical observed OTU counts). We use the following model for generating structural zeros:

$$P(\delta_j = 1) = 1 - \exp(-d_{jK}/T), \quad j = 1, \ldots, K,$$

where d_{jK} stands for the phylogenetic distance between taxon j and the reference taxon K. Without loss of generality, we assign the label K to the reference taxon that must exist. The

existence of the reference taxon is based on both the empirical observed OTU counts and background knowledge.

Let $\Omega = \{j : \delta_j = 0\}$. Let L denote the number of elements in Ω. Therefore, one can write $\Omega = \{j_1, j_2, \ldots, j_L\}$. Note that $\delta_K = 0$ as described above, i.e., $j_L = K$.

Define $\mathbf{U}_i = (\pi_{ij}/\pi_{iK})_{j \neq K: j \in \Omega}$. Following the logit normal distribution, we have

$$\log \mathbf{U}_i \sim N(\boldsymbol{\theta}, \Sigma), \tag{2.3}$$

where $\sigma_{lm} = e^{-2\rho D_{lm}^2}\sigma^2 (l, m \in \Omega)$ are the entries for the variance-covariance matrix Σ (Xiao et al. 2018). Please note $\boldsymbol{\theta}$ is $L-1$ by 1 vector and Σ is $L-1$ square matrix.Here we make use of the phylogenetic distance for parameterizing Σ. Intuitively if two taxa are closely related in the phylogenetic tree, we expect their abundances are closed related as well. One can easily see that the model for Σ only involves two parameters, i.e., ρ and σ^2.

Let $\boldsymbol{\pi}_i = \{\pi_{ij} : j \in \Omega\}$. One can easily derive the following facts:

$$\begin{cases} \pi_{ij_1} = \dfrac{exp(U_{ij_1})}{\sum_{k=1}^{L-1} exp(U_{ij_k}) + 1}, \\ \vdots \\ \pi_{ij_{L-1}} = \dfrac{exp(U_{ij_{L-1}})}{\sum_{k=1}^{L-1} exp(U_{ij_k}) + 1}, \\ \pi_{ij_L} = \dfrac{1}{\sum_{k=1}^{L-1} exp(U_{ij_k}) + 1} \end{cases} \tag{2.4}$$

The OTU counts are generated from multinomial distributions, that is,

$$(Y_{ij} : j \in \Omega) \sim \text{Multinomial}(n, \pi_{i\Omega}),$$