Data Generator

At the beginning, we would call out the real data from COMBO from the phyloMDA package. As we all know that alpha is one of the most important parameters that affect the proportion of structural zeros. By running the scenarios for different settings of parameters, I noticed that a higher value of alpha produces a higher proportion of structural zeros.

In the true abundance generator, we are following the new model for generating structural zeros. In the formula p\_delta1[1, j] = 1 - exp((-D\_phylo[j,k])/(alpha)), which is the phylogenetic-distance-tree-based calculation. One thing worth our attention is when the j and k are the same in the D\_phylo[j,k]. Then -D\_phylo[j,k] will be zero and it makes the p\_delta1 become 0 because there is no distance when the taxon is compared to itself. For the delta\_param (where the structural zeros at), they generate randomly under the binomial distribution with 1 observation. And we only consider one row that has the same value due to the homogeneity.

After getting the delta space 1 and 0 from Bernoulli distribution, we get the index of zeros from delta\_param (ind), and the related abundance (y\_ind) values generated through the random multivariate normal distribution function. To simulate the condition of the environment, we set the mean as theta and the sigma is our variance-covariance matrix (it is generated after delta\_param) that the function is given in the pdf file.

Finally, we use the mu\_prob (mentioned in the meeting with the notes on the whiteboard) which is the log of the ratio for pi ji/ pi jL to fit in true\_abundance[i, ind] to get the true abundance rate.