Bayesian ordination of species sampling data from microbiome studies

Sergio Bacallado*
Stanford University, Stanford, CA, United States - sergiob@gmail.com

Boyu Ren
Harvard University, Brookline, MA, United States - bor158@mail.harvard.edu

Lorenzo Trippa
Dana-Faber Cancer Center, Brookline, MA, United States - ltrippa@jimmy.harvard.edu

Stefano Favaro
University of Torino, Torino, Italy - stefano.favaro@unito.it

Studies of the human microbiome via high-throughput sequencing of ribosomal DNA produces count data in the form of species sampling sequences from diverse environments. The goal is to characterize the effect of the environment on microbial abundances, and the environments are frequently structured according to multiple covariates, both categorical (patient identity, location in the body, pathologies) and quantitative (time, drug dose). We introduce a nonparametric prior for multiple discrete distributions which allows us to model such data. The dependence between distributions is represented by factors embedded in a low-dimensional Euclidean space. These factors can represent random effects or fixed effects. The analysis yields useful visualizations of the data with analogs in multivariate statistics, which have the advantage of a Bayesian approach when it comes to assessing credibility and dealing with missing data, a recurring problem in microbiome studies.

Keywords: Bayesian nonparametrics, Dependent random measures, Species sampling models.