A semiparametric approach for differential abundance analysis in microbiome experiments

Leyla Kodalci[1], Olivier Thas[1,2,3],

[1]Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Data Science Institute, Hasselt University, Hasselt, Belgium
[2]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Gent, Belgium
[3]National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Wollongong, Australia

E-mail for correspondence: `leyla.kodalci@uhasselt.be`

Abstract: Microbiome data obtained from high-throughput sequencing are considered as compositional data, which is characterised by a sum-constraint. Hence, only ratios of observations are informative. Furthermore, microbiome data are overdispersed and have many zero abundances. Many compositional data analysis methods make use of log ratios of the components of the observation vector. However, the many zero abundances cause problems when calculating ratios and logarithms.
In this work, we focus on the the identification of taxa that are differentially abundant between two groups. We have developed a semiparametric method targeting the probability that the outcome of one taxon is smaller than the outcome of another taxon (a *probabilistic index*). The estimation of this probability only requires information about the pairwise ordering of the taxa, and hence zero observations cause no problems. Testing for differential abundance then reduces to testing that the probabilistic indexes are the same in the two treatment groups. We have constructed the semiparametric efficient estimator of the effect size parameter in the model, and a hypothesis test based on this estimator. Results from a simulation study indicate that our methods control the FDR at the nominal level and have good sensitivity compared to competitors.

Key words: differential abundance analysis; high-dimensional; large scale hypothesis testing, probabilistic index; rank method; semiparametric;