Log-ratio analysis of microbiome data with many zeroes is library size dependent

Dennis te Beest[1], Els Nijhuis[2], Tim Mohlmann[1], and Caro Ter braak[2]

[1]Wageningen Universiteit en Research
[2]Wageningen University & Research

November 30, 2020

Abstract

Microbiome composition data collected through amplicon sequencing are count data on taxa in which the total count per sample (the library size) is an artifact of the sequencing platform and as a result such data are compositional. To avoid library size dependency, one common way of analyzing multivariate compositional data is to perform a principal component analysis (PCA) on data transformed with the centered log-ratio, hereafter called a log-ratio PCA. Two aspects typical of amplicon sequencing data are the large differences in library size and the large number of zeroes. In this paper we show on real data and by simulation that, applied to data that combines these two aspects, log-ratio PCA is nevertheless heavily dependent on the library size. This leads to a reduction in power when testing against any explanatory variable in log-ratio redundancy analysis. If there is additionally a correlation between the library size and the explanatory variable, then the type 1 error becomes inflated. We explore putative solutions to this problem.

Hosted file

`te Beest et al - Log-ratio analysis of microbiome data.pdf` available at https://authorea.com/users/379938/articles/496125-log-ratio-analysis-of-microbiome-data-with-many-zeroes-is-library-size-dependent