High-dimensional compositional microbiota data: state-of-the-art of methods and software implementations

Perrine Soret*[1,2,3], Marta Avalos[1,2], Cheng Soon Ong[5], and Rodolphe Thiébaut[1,2,3,4]

[1]Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, UMR 1219, F-33000 Bordeaux, France
[2]INRIA SISTM Team, F-33405 Talence, France
[3]Vaccine Research Institute (VRI), F-94000 Créteil, France
[4]CHU Bordeaux, Department of Public Health, F-33000 Bordeaux, France
[5]Data61, CSIRO, 7 London Circuit, Canberra ACT 2601, Australia

Prioritized topic: High-dimensional data in clinical research

Compositional data (CoDa) consist of a collection of nonnegative measurements that sum to a constant value, typically, proportions that sum to 1. Because knowing the sum, one component can be determined from the sum of the remainder, the parts that make up the composition are mathematically and statistically dependent. This distinct structure complicates analysis and does not allow standard statistical analyses. Aitchison (JRSS-B, 1982) and Egozcue and colleagues (Math. Geol., 2003), among others, provided a framework to analyze CoDa by mapping data from the constrained simplex space to the Euclidian space using nonlinear transforms such as the log-odds or the isometric log-ratio transforms.

The increasing quality/reducing cost of high-throughput sequencing technology, in particular, 16S rRNA gene sequencing of the bacterial component of the human microbial community (microbiota), has enabled researchers to investigate human diseases. Subsequently, microbiota has been associated with numerous diseases, including inflammatory bowel disease, diabetes, cancer and cystic fibrosis. Because of the compositional structure and the high-dimensional data generated by microbiota sequencing, there is also a parallel development of specific statistical analysis methods and computational tools.

Microbiota are usually measured as relative abundance of species and analyzed as CoDa. The objectives of this work are the following:

- First, to review theory and usage of CoDa analysis in the microbiota setting, with particular emphasis on recent proposals adapted to high-dimensional problems (e.g. supervised –constrained Lasso, hierarchical Lasso, kernel methods, sPLS, or unsupervised – PCoA, PCA, Sparse inverse covariance estimation).

- Second, to investigate the current state-of-the-art software implementations (basically, R packages: compositions, vegan, ALDex2, PERMANOVA, MiRKAT, MixMC ...)

- Third, using toy examples and publicly available data (the 16S data from the Koren and colleagues' study in March 2011's PNAS, available in the MixMC R package), to implement and evaluate those methods with publicly available codes. Evaluation criteria are mainly based on computational and practical aspects.

*Corresponding Author: perrine.soret@u-bordeaux.fr