Some proposals to investigate zero patterns in compositional data sets

J. Palarea-Albaladejo[1], J.A. Martín-Fernández[2], S. F. M. Chastin[3]

[1]Biomathematics and Statistics Scotland, Edinburgh (UK)
[2]Dept. Computer Science, Applied Mathematics and Statistics Department, University of Girona, (Spain)
[3]Glasgow Caledonian University, Glasgow (UK)

Abstract

Compositional data sets often include compositions with zero values. When it is sensible to assume that these zeros result from rounding errors, values below a detection limit or, in the case of discrete compositions, zero counts due to limited sampling, then it is a valid approach to apply some statistical data imputation pre-processing in order to deal with them. In this regard, there is a number of proposals based on compositional principles, including univariate, multivariate, non-parametric, parametric, Bayesian, random and robust alternatives (Palarea-Albaladejo and Martín-Fernández, 2015). All of these methods treat zero values as censored values, with the censoring point or detection limit acting as a threshold for the imputation procedure.

However, in some applications, zero values correspond with truly zeros. For example this is usual in physical activity research where some individuals can spend no time on a certain activity category. That is, we need to assume here that zeros are feasible values in the sample space. These zeros correspond to parts that can be absent from the composition. This type of zeros (referred to as essential zeros) are troublesome because it is not generally realistic to replace them by small values. In this context, one important question that a researcher should face is “Are the pattern of zeros associated to the presence of subpopulations in the data set?”. In order to investigate this we propose to analyze whether the subgroups of samples defined by the pattern of zeros can be considered statistically different from one another in terms of, for example, their location and variability measures obtained from common non-zero parts.

In this work we introduce some graphical and statistical tools to, respectively, explore and testing for differences between groups defined by zero patterns. In particular, parametric and permutation tests for the elements of the variation array are developed. These statistical tests are generalized for the case of projections along log-contrasts of interest that can be determined by the user. We use real and simulated data sets to illustrate their performance.

References

Palarea-Albaladejo, J. and Martín-Fernández, J.A. (2015). zCompositions - R package for multivariate imputation of nondetects and zeros in compositional data sets. *Chemometrics and Intelligent Laboratory Systems*, 143, 85-96.