Exploring outliers in compositional data with structural zeros

M. Templ[1], K. Hron[2], and P. Filzmoser[3]
[1]Zurich Univeristy of Applied Sciences, Winterthur, Switzerland; *matthias.templ@zhaw.ch*
[2]Palacký University of Olomouc, Czech Republic
[3]Vienna University of Technology, Austria

Abstract

The analysis of compositional data using the log-ratio approach is based on ratios between compositional parts. Zeros in the parts thus cause serious difficulties for the analysis. This is a particular problem in presence of structural zeros, resulting from a structural process rather than from imprecision of a measurement device. Examples of structural zeros in compositional data are, e.g. expenditures in tobacco for non-smokers, tax payments on shares for non-share holders, time budget on sports for people who do not do any sports at all, etc. Therefore, structural zeros cannot be simply replaced by a non-zero value as it is done, e.g. for values below detection limit or missing values. Instead, zeros have to be incorporated into further statistical processing.

We lay the focus on exploratory tools for identifying outliers in compositional data sets with structural zeros. For this purpose, robust Mahalanobis distances are estimated; they are computed either directly for subcompositions determined by their zero patterns or by using imputation as an auxiliary step to improve the efficiency of the estimates. Consequently, we proceed to the subcompositional and subgroup level. For this approach, new theory is formulated that allows to estimate covariances for imputed compositional data and to apply estimations on subgroups using blocks of this covariance matrix (Templ, Hron and Filzmoser, 2016). Moreover, the zero pattern structure is analyzed using PCA for binary data (de Leeuw, 2016) to achieve a comprehensive view of the overall multivariate structure of zeros.

The proposed tools are applied to large-scale data from official statistics, where the need for an appropriate treatment of zeros is obvious.

References

Templ, M. and Hron, K. and Filzmoser, P. (2016). Exploratory tools for outlier detection in compositional data with structural zeros. *Journal of Applied Statistics*, online first, doi 10.1080/02664763.2016.1182135.

de Leeuw, J. (2006). Principal Component Analysis of Binary Data by Iterated Singular Value Decomposition. *Computational Statistics and Data Analysis*, pp. 21–39, 50(1).