Robust Principal Component Analysis for Compositional Tables

J. Rendlová[1], K. Fačevicová[1], K. Hron[1] and P. Filzmoser[2]
[1]Palacký University, Olomouc, Czech Republic; *julie.rendlova@gmail.com*
[2]Vienna University of Technology, Vienna, Austria

Abstract

Many practical examples contain relative information about the distribution according to two factors which leads to a $(I \times J)$-dimensional extension of compositional data carrying information about a relationship between these factors. Egozcue and others (2008) showed that such structure, called a compositional table $\boldsymbol{x}$,

$$\boldsymbol{x} = \begin{pmatrix} x_{11} & \cdots & x_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1} & \cdots & x_{IJ} \end{pmatrix}, x_{ij} > 0, i = 1, ..., I, j = 1, ..., J, \tag{1}$$

can be decomposed into its independent and interactive parts where only the interaction table is of further interest. The respective coordinate representation of both independence and interaction tables was proposed in Fačevicová and others (2016).

One of the primary tasks in multivariate statistics is to reduce the dimensionality of the data at hand, done using principal component analysis (PCA). To eliminate the influence of outliers in PCA, Filzmoser and Hron (2009) proposed to estimate the covariance matrix for robust PCA by the Minimum Covariance Determinant estimator. Since clr coefficients lead to singularity and are generally not appropriate for robust methods, loadings and scores for PCA need to be computed from pivot coordinates of the interaction table (introduced by Fačevicová and others (2016)) and then transformed back to clr coefficients for better intepretation of the resulting compositional biplot.

Accordingly, the aim of this contribution is to propose a robust approach to principal component analysis of compositional tables and to illustrate the introduced theory on a real data set from OECD Statistics using the statistical software R, namely the robCompositions package. Data from several different countries containing information about gender distribution and level of education, respectively, of 300 thousand students in eight different fields of studies were processed respectively as a set of 2×8 and 3×8 compositional tables. Therefore, a robust compositional biplot is a possible tool to analyze study tendencies among these countries as well as gender, or bachelor, master and doctoral educational levels differences.

References

Egozcue, J. J., Díaz-Barrero, J. L., Pawlowsky-Glahn, V. (2008). Compositional Analysis of Bivariate Discrete Probabilities. In Daunis-i-Estadella, J., Martín-Fernández, J. A. (Eds.), *Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop*. University of Girona, Spain.

Fačevicová, K., Hron, K., Todorov, V., Templ, M. (2016). Compositional Tables Analysis in Coordinates. *Scandinavian Journal of Statistics 43*(4), pp. 962–977.

Filzmoser, P., Hron, K., Reimann, C. (2009). Principal Component Analysis for Compositional Data with Outliers. *Environmetrics 20*(6), pp. 621–632.