Robustness for Compositional Data Using the R Package robCompositions

M. Templ[1,2], P. Filzmoser[1], and K. Hron[3]

[1] Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstr. 8-10, 1030 Wien, Austria.
[2] Department of Methodology, Statistics Austria, Guglgasse 13, 1110 Wien, Austria.
[3] Dept. of Mathematical Analysis and Applications of Mathematics, Tomkova 40, 77100 OLOMOUC, Czech Republic.

Keywords: Missing Values, Compositional Data, Imputation, R.

1 Compositional Data

Compositional data occur frequently in official statistics (tax components in tax data, income components, wage components, expenditures, etc.), in environmental and technical sciences, and in many other fields. An observation $\boldsymbol{x} = (x_1, \ldots, x_D)^t$ is by definition a D-part composition if, and only if, all its components are strictly positive real numbers, and if all the relevant information is contained in the ratios between them (Aitchison, 1986). As a consequence of this formal definition, $(x_1, \ldots, x_D)^t$ and its $c > 0$ multiple $(cx_1, \ldots, cx_D)^t$ contain essentially the same information. One can thus define the *simplex*, which is the sample space of D-part compositions, as

$$\boldsymbol{x} = (x_1, \ldots, x_D)^t, \quad x_i > 0, \quad i = 1, \ldots, D, \quad \sum_{i=1}^{D} x_i = \kappa. \tag{1}$$

The constant κ represents the sum of the parts.

The application of standard statistical methods such as correlation analysis, principal component analysis or factor analysis directly to compositional data can thus lead to meaningless results (Filzmoser and Hron, 2008, 2009). This is also the case for imputation methods. Thus, this type of data needs to be transformed first with an appropriate transformation before any statistical analysis can be applied. In addition to that, various standard methods must be adapted to be able to obtain valid results.

2 The R Package robCompositions

The R package robCompositions includes various imputation methods for compositional data, such as k-nn imputation based on Aitchison distances (Aitchison, 1986) using a re-scaling of the neighbors, an iterative model-based procedure (Hron et al., 2008), and an imputation method using an adapted EM-algorithm for compositional data (Palarea-Albaladejo and Martín-Fernández, 2008). Furthermore, some plot methods to highlight the quality of the imputation are included (Templ et al., 2009).

It also includes multivariate methods such as (robust) factor analysis, (robust) principal component analysis (Filzmoser et al., 2009) and procedures for outlier detection in compositional data (Filzmoser and Hron, 2008).

Various data sets from the literature are included in the package, but also print, summary and plot methods are implemented for almost every method.

Some of the implemented features will be presented with focus on imputation of missing values. In addition to that, the methods will be described briefly and the advantages of iterative model-based methods will be shown using results obtained from real and simulated data.

The open-source package can be downloaded from the CRAN (`http://cran.r-project.org`).

References

J. Aitchison (1986). *The Statistical Analysis of Compositional Data.* Chapman & Hall, London. Reprinted in 2003 by Blackburn Press.

P. Filzmoser and K. Hron (2008). Outlier detection for compositional data using robust methods. *Mathematical Geosciences* 40 (3), 233-248.

P. Filzmoser and K. Hron (2009). Correlation analysis for compositional data. *Mathematical Geosciences*, in press.

P. Filzmoser, K. Hron, and C. Reimann (2009). Principal component analysis for compositional data with outliers. *Environmetrics*, in press.

K. Hron, M. Templ, and P. Filzmoser (2008). Imputation of compositional data using robust methods. *Research Report SM-2008-4*, Department of Statistics and Probability Theory, Vienna University of Technology.

J. Palarea-Albaladejo and J.A. Martín-Fernández (2008). A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. *Computer & Geosciences* 34(8), 902–917.

M. Templ, P. Filzmoser, K. Hron (2009). Robust Imputation of Missing Values in Compositional Data Using the R-Package robCompositions. *In Proceedings of New Technologies and Techniques for Statistics*, Eurostat, 11 pages.

M. Templ, P. Filzmoser, K. Hron (2009). robCompositions: Robust Estimation for Compositional Data. R package version 1.3. `http://cran.r-project.org/web/packages/robCompositions/index.html`