Analysis of Compositional Data using Robust Methods. The R-package robCompositons

M. TEMPL[1,3], P. FILZMOSER[1] and K. HRON[2]

[1]Department of Statistics and Probability Theory - Vienna University of Technology, Austria templ@tuwien.ac.at
[2]Department of Mathematical Analysis and Applications of Mathematics - Palacký University, Czech Republic
[3]Statistics Austria, Vienna, Austria

1 Few Words about R and CoDa

The free and open-source programming language and software environment `R` (R Development Core Team, 2010) is currently both, the most widely used and most popular software for statistics and data analysis. In addition, `R` becomes quite popular as a (programming) language, ranked currently (February 2011) on place 25 at the TIOBE Programming Community Index (e.g., Matlab: 29, SAS: 30, see http://www.tiobe.com).

The basic `R` environment can be downloaded from the comprehensive `R` archive network (http://cran.r-project.org). `R` is enhanceable via *packages* which consist of code and structured standard documentation including code application examples and possible further documents (so called *vignettes*) showing further applications of the packages.

Two contributed packages for compositional data analysis comes with `R`, version 2.12.1.: the package `compositions` (van den Boogaart et al., 2010) and the package `robCompositions` (Templ et al., 2011).

Package `compositions` provides functions for the consistent analysis of compositional data and positive numbers in the way proposed originally by John Aitchison (see van den Boogaart et al., 2010).

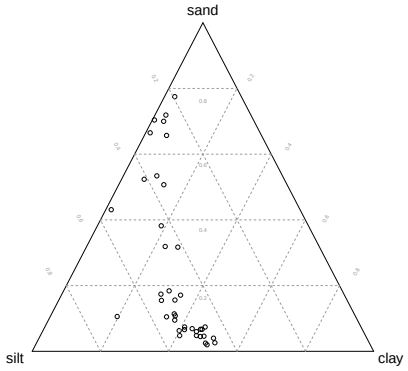
In addition to the basic functionality and estimation procedures in package `compositions`, package `robCompositions` provides tools for a (classical) and robust multivariate statistical analysis of compositional data together with corresponding graphical tools. In addition, several data sets are provided as well as useful utility functions.

2 Motivation to Robust Statistics

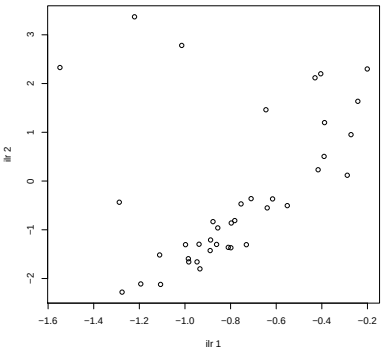
Both measurement errors and population outliers can have a high influence on classical estimators. Arbitrary results may be the consequence, because outliers may have a large influence and wrong interpretation of estimations may result. In addition, checking model assumptions is then often not possible since outliers may disturb the applied model itself. All these problems may be avoided when applying methods based on robust estimators.

To be more specific, a simple analysis is done in the following by applying principal component analysis - using function `pcaCoDa()` of package `robCompositions` - to the *Arctic Lake sediment data set* (Aitchison, 1986). We show the effect of outliers on a **simplyfied example for demostration purposes**. However, the same problems occur in higher dimensions where usually principal component analysis is applied mostly for dimension reduction purposes.

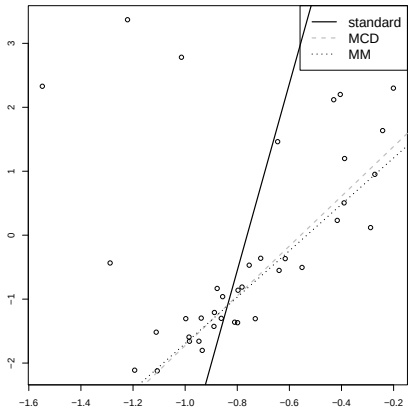
Figure 1(a) shows the 3-part compositions of the Arctic Lake Sedimanet Data in a ternary diagram. Few outliers are clearly visible, like the two ones with higher percentages in the *silc*-part. After transforming the parts by using the isometric log-ratio transformation (Egozcue et al., 2003), outliers are still visible (see Figure 1(b)). For obtaining the principal components, the eigenvalues of the covariance matrix need to be derived. A robust estimation of the underlying covariance matrix leads to robust principal components. Figure 1(c) shows the direction of the first principal component when using different covariance estimators: classical estimation (black solid line), and robust estimation using the MM estimator (see, e.g., Maronna et al., 2006) (dotted line in grey), and the (fast) MCD estimator (Rousseeuw and von Driessen, 1999) (black coloured dashed line) with high degree of



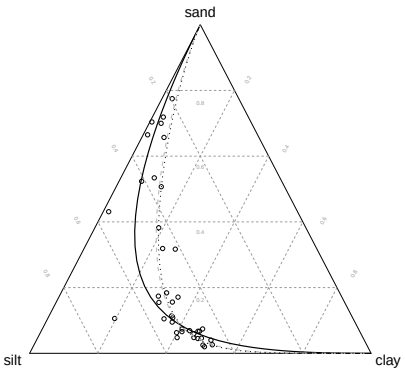
(a) Ternary diagram of the arcticLake data



(b) Ilr transformed arcticLake data



(c) First principal component for ilr transformed data.



(d) First principal component back-transformed to original scale.

Figure 1: The upper left graphic (a) shows a ternary diagram of the *Arctic Lake Sediment Data*. In the upper right graphic (b), the ilr-transformed data are shown and the first principal component is displayed in Figure (c) while the first principal component is shown in the ternary diagram in Figure (d).

robustness. It is easy to see that the first principal component is attracted by the few outliers in the lower right plot, while the principal components obtained from robust estimates are not. Finally, in Figure 1(d) the first principal components of the classical and the robust estimators are shown in the ternary diagram. Again, it is easy to see that the first principal component from classical estimation is highly influenced, especially by the two outliers having higher concentration in silt. The line does not follow the main part of the data.

This example shows that robust estimation is important to get reliable estimates for multivariate analysis of compositional data, especially when using more complex data than this simple 3-part composition.

3 Available Functionality

In the following the data sets and most important functions of package `robCompositions` are briefly described. Note, that almost all print and summary functions are not listed here, but their description is available in Templ et al. (2011).

3.1 Data sets

Several compositional data sets are included in the package, like:

`arcticLake`	The Artic Lake Sediment Data from the Aitchison book (Aitchison, 1986).
`coffee`	The Coffee Data contain 27 commercially available coffee samples of different origins (see Korhonová et al., 2009).
`expenditures`	The Household Expenditures Data on five commodity groups of 20 single men from the Aitchison book (Aitchison, 1986).
`expendituresEU`	Mean consumption expenditure of households at EU-level (2005) provided by Eurostat.
`haplogroups`	Distribution of European Y-chromosome DNA (Y-DNA) haplogroups by region in percentages, from Eupedia.
`machineOperators`	This data set from (Aitchison, 1986, p. 382) contains compositions of eight-hour shifts of 27 machine operators.
`phd`	PhD students in Europe based on the standard classification system splitted by different kind of studies (given as percentages), provided by Eurostat 2009.
`skyeLavas`	AFM compositions of 23 aphyric Skye lavas (Aitchison, 1986, p. 360).

3.2 Basic functions

Basic utility functions like log-ratio transformations but also functions which specially written in `C` (for e.g. to compute distances between compositions) are implemented in the package. The most important are:

`aDist(x, y)`	Computes the Aitchison distance between two observations or between two data sets. The underlying code is written in `C` and allows a fast computation also for large data sets.
`constSum(x, const=1)`	Closes compositions to sum up to a given constant (default 1).
`robVariation(x, robust=TRUE)`	Estimates the variation matrix with robust or classical methods.
`ternaryDiag(x, ...)`	Ternary diagram, optionally with grid.
`alr(x, ivar=ncol(x))`	The alr transformation moves D-part compositional data from the simplex into a $(D-1)$-dimensional real space.
`invalr(x, ...)`	Inverse additive log-ratio transformation, often called additive logistic transform. The function allows also to preserve absolute values when parameter class info is provided.
`clr(x)`	The clr transformation moves D-part compositional data from the simplex into a D-dimensional real space.
`invclr(x, useClassInfo = TRUE)`	The inverse clr transformation. Absolute values are preserved optionally.
`ilr(x)`	An isometric log-ratio transformation with a special choice of the balances according to Hron et al. (2010).
`invilr(x.ilr)`	The inverse transformation of `ilr()`.

3.3 Exploratory Tools

Multivariate outlier detection can give a first impression about the general data structure and quality (Maronna et al., 2006). This is also true for compositional data. The compositions are firstly transformed to the real space before robust methods are applied for outlier detection (Filzmoser and Hron, 2008).

The (robust) compositional biplot displays both samples and variables of a data matrix graphically in the form of scores and loadings of a principal component analysis, preferably - because of interpretation of the biplot - after clr transformation of the data (Filzmoser et al., 2009a).

The package comes with the following functionality for exploratory compositional data analysis:

`outCoDa(x, ...)`	Outlier detection for compositional data using classical and robust statistical methods (Filzmoser and Hron, 2008).
`plot.outCoDa` or `plot()`	Plots the Mahalanobis distances to detect potential outliers.
`pcaCoDa(x, method = "robust")`	This function applies robust principal component analysis for compositional data (Filzmoser et al., 2009a).
`plot.pcaCoDa()` or `plot()`	Provides robust compositional biplots.

3.4 Model-based Multivariate Estimation and Tests

Outliers may lead to model misspecification, biased parameter estimation and incorrect results.

The main functionality of package `robCompositions` is provided on model-based estimations, namely factor analysis (Filzmoser et al., 2009b), discriminant analysis (Filzmoser et al., 2009c) and imputation of rounded zeros (Palarea-Albaladejo, and Martin-Fernandez , 2008) or missing values (Hron et al., 2010).

The package provides the following functions:

`adtest(x, R = 1000, locscatt = "standard")`	This function provides three kinds of Anderson-Darling normality Tests.
`adtestWrapper(x, alpha = 0.05, R = 1000, robustEst = FALSE)`	A set of Anderson-Darling tests are applied as proposed by Aitchison (1986).
`summary.adtestWrapper` or `summary()`	Summary of the adtestWrapper results.
`alrEM(x, pos = ncol(x), dl = ...)`	A modified EM alr-algorithm for replacing rounded zeros in compositional data sets (Palarea-Albaladejo, and Martin-Fernandez , 2008).
`daFisher(x, grp, ...)`	Discriminant analysis by Fishers rule (as described in Filzmoser et al., 2009c).
`impCoda(x, method = "ltsReg", ...)`	Iterative model-based imputation of missing values using special balances (Hron et al., 2010).
`impKNNa(x, method = "knn", k = 3, ...)`	This function offers several k-nearest neighbor methods for the imputation of missing values in compositional data (Hron et al., 2010).
`plot.imp()` or `plot()`	This function provides several diagnostic plots for the imputed data set in order to see how the imputed values are distributed in comparison with the original data values (Templ, Filzmoser, and Hron , 2009).
`pfa(x, factors, ...)`	Computes the principal factor analysis of the input data which are clr transformed first.

4 Conclusion and Outline

In this contribution we started with a short motivation why robustness is of major concern in compositional data analysis.

We then briefly introduced and listed the methods implemented in package `robCompositions`. More details about each function can be found in the manual of the package Templ et al. (2011) and in the book chapter about `robCompositions` in the forthcoming book *Compositional Data Analysis: Theory and Applications* (Pawlowsky-Glahn, and Buccianti , 2011).

The package comes with the *General Public Licence, version 2*, and can simple be downloaded at http://cran.r-project.org/package=robCompositions .

Future developments include further methods on replacing rounded zeros in the data as well as to handle structural zeros. Furthermore, a graphical user interface is currently developed.

Comments and collaborations regarding the development of the package are warmly welcome.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

van den Boogaart, K. G., R. Tolosana, and M. Bren (2010). *compositions: Compositional Data Analysis*. R package version 1.10-1.

Egozcue, J.J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology 35*(3), 279–300.

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

Filzmoser, P., K. Hron, and C. Reimann (2009a). Principal component analysis for compositional data with outliers. *Environmetrics 20*(6), 621–632.

Filzmoser, P., K. Hron, C. Reimann, and R.G. Garrett (2009b). Robust factor analysis for compositional data. *Computers and Geosciences 35*, 1854–1861.

Filzmoser, P., K. Hron, and M. Templ (2009c). Discriminant analysis for compositional data and robust parameter estimation. Technical Report SM-2009-3, Vienna University of Technology, Austria. Submitted for publication.

Hron, K., M. Templ, and P. Filzmoser (2010). Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics and Data Analysis 54*(12), 3095–3107. DOI:10.1016/j.csda.2009.11.023.

Korhoňová, M., K. Hron, D. Klimčíková, L. Müller, P. Bednář, and P. Barták (2009). Coffee aroma - statistical analysis of compositional data. *Talanta 80*(82), 710–715.

Maronna, R., D. Martin, and V. Yohai (2006). *Robust Statistics: Theory and Methods*. John Wiley & Sons Canada Ltd., Toronto, ON.

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

Pawlowsky-Glahn, V., and A. Buccianti (2011). *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons Canada Ltd., Toronto, ON. Accepted for publication.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rousseeuw, P.J., and K. von Driessen (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223, 1999.

Templ, M., K. Hron, and P. Filzmoser (2011). *robCompositions: Robust Estimation for Compositional Data*. Manual and package, version 1.4.4.

Templ, M., P. Filzmoser, and K. Hron (2009). Imputation of item non-responses in compositional data using robust methods. *Work Session on Statistical Data Editing*, Neuchatel, Switzerland, 11 pages.