Overview of the robCompositions package. Compositional Data Analysis using Robust Methods.

M. TEMPL\textsuperscript{1,3}, P. FILZMOSER\textsuperscript{1} and K. HRON\textsuperscript{2}

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

\section{1 Few Words about R and CoDa}

The free and open-source programming language and software environment R \cite{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 \cite{van den Boogaart et al., 2010} and the package robCompositions \cite{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 \cite{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.

\section{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 \cite{Aitchison, 1986}. We show the effect of outliers on a simplyfied example for demonstration 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 Sediment 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 \cite{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 \cite{Maronna et al., 2006} (dotted line in grey), and the (fast) MCD estimator \cite{Rousseeuw and von Driessen, 1999} (black coloured dashed line) with high degree of 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:

<table>
<thead>
<tr>
<th>Data set</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>arcticLake</td>
<td>The Artic Lake Sediment Data from the Aitchison book (Aitchison, 1986).</td>
</tr>
<tr>
<td>coffee</td>
<td>The Coffee Data contain 27 commercially available coffee samples of different origins (see Korhonová et al., 2009).</td>
</tr>
<tr>
<td>expendituresEU</td>
<td>Mean consumption expenditure of households at EU-level (2005) provided by Eurostat.</td>
</tr>
<tr>
<td>haplogroups</td>
<td>Distribution of European Y-chromosome DNA (Y-DNA) haplogroups by region in percentages, from Eupedia.</td>
</tr>
<tr>
<td>machineOperators</td>
<td>This data set from (Aitchison, 1986, p. 382) contains compositions of eight-hour shifts of 27 machine operators.</td>
</tr>
<tr>
<td>phd</td>
<td>PhD students in Europe based on the standard classification system splitted by different kind of studies (given as percentages), provided by Eurostat 2009.</td>
</tr>
</tbody>
</table>

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:
### 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")** or **plot.pcaCoDa()**: This function applies robust principal component analysis for compositional data (Filzmoser et al., 2009a). Provided 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


(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 arcticLake 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).