Exploratory compositional data analysis using the R-package robCompositions

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

[(1)] *Dept. of Mathematical Analysis and Applications of Mathematics, Palacký University, Olomouc, CZECH REPUBLIC*
e-mail: `hronk@seznam.cz`
[(2)] *Dept. of Statistics and Probability Theory, Vienna University of Technology, Vienna, AUSTRIA*
[(3)] *Methods Unit, Statistics Austria, AUSTRIA.*

Abstract

Compositional data are multivariate observations that carry only relative information. This means that not the absolute values but the ratios between the variables are of interest. This is important also for an exploratory analysis of such data. We present two basic methods for the exploratory compositional data analysis (ECDA), namely multivariate outlier detection and the compositional biplot. The methods are illustrated at a small data example using the R package `robCompositions`.

1 Compositional data

In practice, data frequently consist of percentages or, more general, not the absolute values but the ratios between the variables are of interest. Usually, this kind of observations is characterized with a positive constant sum constraint of variables (usually 1 or 100 in the case of proportions or percentages, respectively), however, this condition is obviously not necessary. Nowadays, multivariate observations that represent quantitative descriptions of the parts of some whole, conveying exclusively relative information, are known under the term compositional data or compositions for short. Obviously, the D-part composition $\mathbf{x} = (x_1, \dots, x_D)'$ and its positive real multiple $c\mathbf{x}$, $c > 0$, convey essentially the same information. The sample space of compositions is a D-part simplex, a $(D-1)$-dimensional subset of $\mathbf{R}^{D-1}$ that contain all D-part compositions that sum up to a prescribed constant sum constraint. The nature of compositions claim for a special geometry, called nowadays the Aitchison geometry with special operations of perturbation, power transformations and the Aitchison inner product with the usual Hilbert space properties [5]. The name of the geometry comes according to John Aitchison, a British statistician that proposed the first comprehensive theory for statistical analysis of compositional data [1]. A special treatment for compositions is necessary because of the different sample space. Thus, the usual statistical methods cannot be applied directly to compositions as they are designed for the Euclidean sample space, where the information is absolute and not relative. As a way out, J. Aitchison proposed the family of log-ratio transformations from the simplex to the real space, known as additive logratio (alr) and centered logratio (clr) transformations. In fact, the new (transformed) variables represent coefficients to a non-orthonormal basis and a generating system on the simplex (with respect to the corresponding geometry), respectively. Nevertheless, only the latter one, defined for a composition $\mathbf{x}$ as

$$clr(\mathbf{x}) = \left(\frac{x_1}{\prod_{i=1}^{D} x_i}, \ldots, \frac{x_D}{\prod_{i=1}^{D} x_i} \right)' \tag{1}$$

is isometric, and thus it moves the Aitchison geometry to the standard Euclidean one. Due to its symmetry, the clr variables are used to construct the popular compositional biplot [2]. On the other hand, the clr transformed data are singular because they sum up to zero. Thus, any statistical analysis based on the assumption of regularity (like, e.g., robust methods) cannot be used. For this reason, nowadays the so called isometric logratio (ilr) transformation [3], that is represented by coefficients to a chosen orthonormal basis on the simplex, became popular. For one such choice we get

$$\mathbf{z} = (z_1, \ldots, z_{D-1})', \; z_i = \sqrt{\frac{i}{i+1}} \ln \frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}}, \; i = 1, \ldots, D-1. \tag{2}$$

All the advantageous theoretical properties of the clr transformation are preserved. On the other hand, the new variables are not easy to interpret, thus the ilr transformation is rather used for the analysis of objects than of compositional parts, although an interpretation for groups of compositional parts is possible [4]. For the latter purpose also the geometric relation between clr and ilr transformations will be useful: the ilr variables are coefficients of an orthonormal basis on the hyperplane formed by the clr transformation. In matrix notation we can write

$$clr(\mathbf{x}) = \mathbf{V}\, ilr(\mathbf{x}), \tag{3}$$

where the columns of the $D \times (D-1)$ matrix $\mathbf{V}$ consist of the mentioned orthonormal vectors. Nowadays, for most statistical methods for compositional data the ilr transformation seems to be convenient for the evaluation of the results. However, there are some exceptions, like the compositional biplot (see the next section) where the clr transformation allows easier interpretation.

2 Aspects of ECDA for compositional data: outlier detection and compositional biplots

To start a statistical analysis of a compositional data set, one is usually interested in patterns of the main structure of the data set (groups in the data, relations between variables) as well as in deviations, represented by outlying observations. As the corresponding statistical tools themselves can be strongly influenced by these irregularities, instead of the classical statistical methods their robust counterparts need to be taken. This means that for the sample $\mathbf{z}_1, \ldots, \mathbf{z}_n$ of n ilr-transformed compositions, the classical sample mean value (arithmetic mean) $\overline{\mathbf{z}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i$ and the sample covariance matrix $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{z}_i - \overline{\mathbf{z}})^2$ as estimators of the true location and covariance should be replaced by their robust versions, being resistant to deviations from the compositional multivariate data structure. In addition, such an estimator should follow the usual properties of the estimates under affine transformations of the sample, what is known as the affine invariance property. Due to both theoretical and practical advances, the MCD (Minimum Covariance Determinant) estimator represents a good choice, being also effectively computable [10].

Multivariate ***outlier detection*** should be the first step in each exploratory compositional data analysis (ECDA). Identified outliers are candidates for aberrant data that may otherwise adversely lead to model misspecification, biased parameter estimation and incorrect results. It is therefore important to identify them prior to modeling and analysis. Outlier detection for compositional data is based on robust Mahalanobis distances, defined for regular $(D-1)$-dimensional data as

$$\mathrm{MD}(\mathbf{z}_i) = [(\mathbf{z}_i - T)' C^{-1} (\mathbf{z}_i - T)]^{1/2},\ i = 1, \dots, n,$$

with robust (e.g. MCD) estimators T of location and C of covariance, respectively. Here, the estimated covariance structure is used to assign a distance to each observation indicating how far the observation is from the center of the data cloud with respect to the covariance structure. For the computed (squared) robust Mahalanobis distance it is usual to use a certain quantile (e.g., the quantiles 0.95 or 0.975) of the χ^2 distribution with $D-1$ degrees of freedom as a cut-off value for outlier identification, see [6]; observations with larger (squared) robust Mahalanobis distance are considered as potential outliers. However, compositional data first needs to be moved to the real space using an suitable transformation, see [6].

The ***compositional biplot*** is nowadays one of the most widely used tools for ECDA. It displays both samples and variables of a data matrix graphically in the form of scores and loadings of a principal component analysis [8]. Usually, samples are displayed as points while variables are displayed either as vectors or rays. For compositional data, one would intuitively construct the biplot for ilr-transformed data, however, due to the complex interpretation of the new variables, it is common to construct the compositional biplot for clr-transformed compositions as proposed in [2]. The scores represent the structure of the compositional data set in the Euclidean space, so they can be used to see patterns and groups in the data. The loadings (rays) represent the corresponding clr-variables. Accordingly, their interpretation is different from the usual case. Namely, the main interest is concentrated to links (distances between vertices of the rays); concretely, for the rays i and j $(i, j = 1, \dots, D)$ the link approximates the (usual) variance $\mathrm{var}(\ln \frac{x_i}{x_j})$ of the logratio between the compositional parts x_i and x_j. Hence, when the vertices coincide, or nearly so, then the ratio between x_i and x_j is constant, or nearly so. In addition, directions of the rays signalize where observations with dominance of the corresponding compositional part are located. Again, outliers can substantially affect results of the underlying principal component analysis and depreciate the predicative value of the biplot. For this reason, again the robust version of the biplot is needed. However, as the robust methods cannot work with singular data, the robust scores and loadings must be computed from ilr-transformed compositions and the result needs to be back-transformed using (3) to the clr plane, see [7] for details. Afterwards, the robust compositional biplot (with above interpretation) can be constructed.

3 ECDA and the R-package robCompositions

The statistical software `R` [9] is a powerful computer environment for statistics and data analysis. It is available for all computer platforms and can be downloaded from `http://cran.r-project.org`. Nowadays, two contributed packages for compositional data analysis are available, `compositions` [12] and `robCompositions` [11]. However, only the latter one provides a comprehensive tool for robust statistical analysis of compositional data, including outlier detection, principal component analysis, factor analysis, missing values imputation, etc, together with the corresponding graphical tools. A comprehensive overview is available using the command

```
help(package="robCompositions").
```

The above described tools of ECDA can be easily applied with `robCompositions`. To make the explanation illustrative, we give practical examples using the well known data set `expenditures` from [1], p. 395, which contains household expenditures on five commodity groups of 20 single men (in former Hong Kong dollars). These variables (compositional parts) represent housing (including fuel and light), foodstuff, alcohol and tobacco, other goods (including clothing, footwear and durable goods) and services (including transport and vehicles). Thus, they represent the ratios of the men's income spent on the mentioned expenditures. Although any constant sum constraint does not occur here, the nature of the data is obviously compositional.

Once the package is loaded, we can load the expenditures data which are included in the package:

```
> library(robCompositions)
> data(expenditures)
```

Next we start to search for potential outliers by computing robust Mahalanobis distances using the function `outCoDa()`. Function `outCoDa()` internally applies a isometric log-ratio transformation to the compositions to search for outliers in the real space. The function inlcudes four function arguments, the data `x`, the significance level `alpha` (1-quantile) and the `method` used (either 'standard' or 'robust'(default)) and, in the latter case, `h` as the size of the subsets for the robust covariance estimation according to the MCD estimator. The latter three function arguments have sensible defaults, but they can also be set by the user.

With the following command, setting the parameters `alpha` and `method` for better illustration[1], we apply robust outlier detection of compositional data:

```
> outRob <- outCoDa(expenditures, alpha=0.05, method="robust")
> outlierRob
--------------------
[1] "2 out of 20 observations are detected as outliers."
--------------------
```

Almost all functions in package `robCompositions` make use of function overloading and the method dispatch of `R`. For example, the function `outCoDa()` returns an object from class `'outdect'`. Print, summary and plot methods are then implemented for objects of certain classes. By typing the corresponding object (here: `outlierRob`), in the R console, the print method (`print.outCoDa`) is selected automatically. Within our example, the print result reports that 2 out of 20 observations are detected as outliers.

For comparison, if outlier detection with the classical estimators is applied only one observation is detected as outlier:

```
> outlierCla <- outCoDa(expenditures, alpha=0.05, method="standard")
> outlierCla
--------------------
[1] "1 out of 20 observations are detected as outliers."
--------------------
```

Thus, when using classical estimates one of the outliers would be masked. In addition, also the resulting Mahalanobis distances (`mahalDist`) as well as a logical vector indicating outliers and non-outliers (`outlierIndex`) can be displayed[2]:

```
> outlierRob$mahalDist
[1] 1.1914708 1.0473757 3.8197815 1.6294140 1.0226241 1.4917820
[7] 1.7143016 2.5129201 1.1268929 3.1124932 1.6244433 1.1502261
[13] 0.8202414 1.6210270 1.0015227 1.9218942 1.7339904 2.1696844
[19] 1.5773077 0.8071032

> outlierRob$outlierIndex
[1] FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE
[12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

The resulting Mahalanobis distances (ordered according to the index of the observations), together with the corresponding cut-off value, can be displayed using the plot

[1] Writing only `outCoDa(expenditures)` is equivalent, because the default values have been used in the example.

[2] type `names(object)` into the `R` console to get information about the list of objects included, with `object` equals to `outlierRob` or `outlierCla` in our example.

function `plot.outCoDa()`. `R` first searches for objects of class `outCoDa` if a function called `'plot.outCoDa'` exists. Therefore, the users only need to know the name of the generic function, `plot()`, which is simple to keep in mind. The outliers (observations 3, 10) are marked using the symbol '+'. The corresponding Figure 1 was obtained as follows:

```
> plot(outlierRob)
```

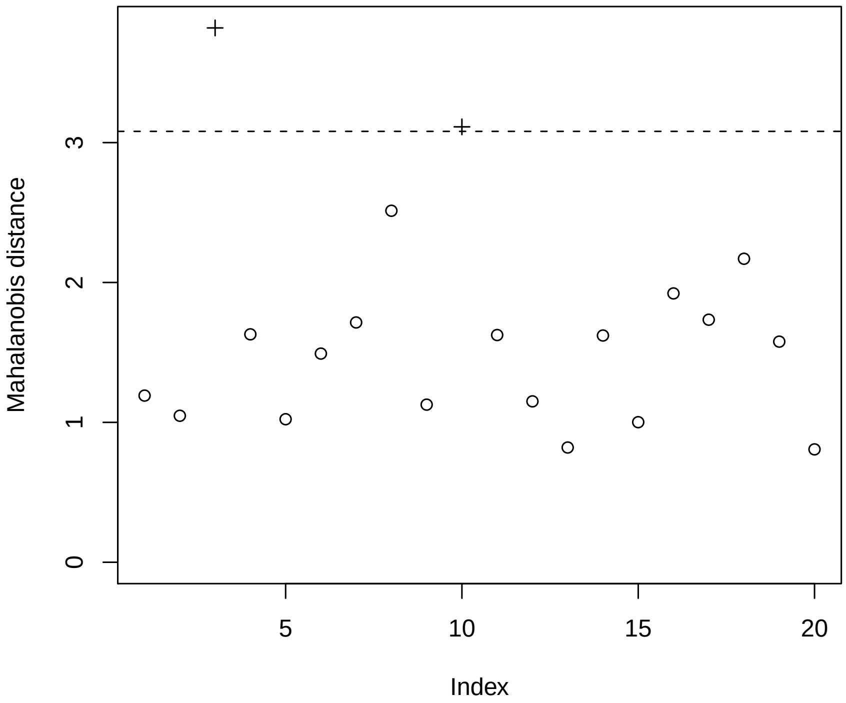

Figure 1: Graphical output by outlier detection for the expenditures data set.

A robust compositional biplot using the package `robCompositions` is obtained by applying the plot function `plot.pcaCoDa()` for (robust) principal component analysis of compositional data. Thus, in the corresponding function `pcaCoDa()` that computes scores and loadings in the clr space, the parameter `method` ('standard' or 'robust' (default)) should be set.

```
> PrinCompRob <- pcaCoDa(expenditures, method="robust")
> plot(PrinCompRob)
```

The result is displayed in Figure 2. The observations (approximated by the scores) are nicely ordered almost on a line, where the deviating observation 3 is clearly visible. On the other hand, the second outlier (10) is masked in the data structure. On the first sight, a dominant influence of any compositional part is not visible, maybe with the exception of the ray 'services'. When inspecting the data, expressed in percentages, the variables 'foodstuff', 'others', and 'services' essentially represent the ordering of the objects as it is visible in the biplot from the scores. Accordingly, observation 3 spends the smallest relative amount on 'foodstuff', and observation 8 the largest amount. The original data set is expressed in percentages by using the function

```
> ConstSum(expenditures)
```

Finally, when comparing the links, the ratios between the parts to 'footstuff' are in general not very stable. However, some relation seems to be more stable, e.g., between 'alcohol' and 'other'.

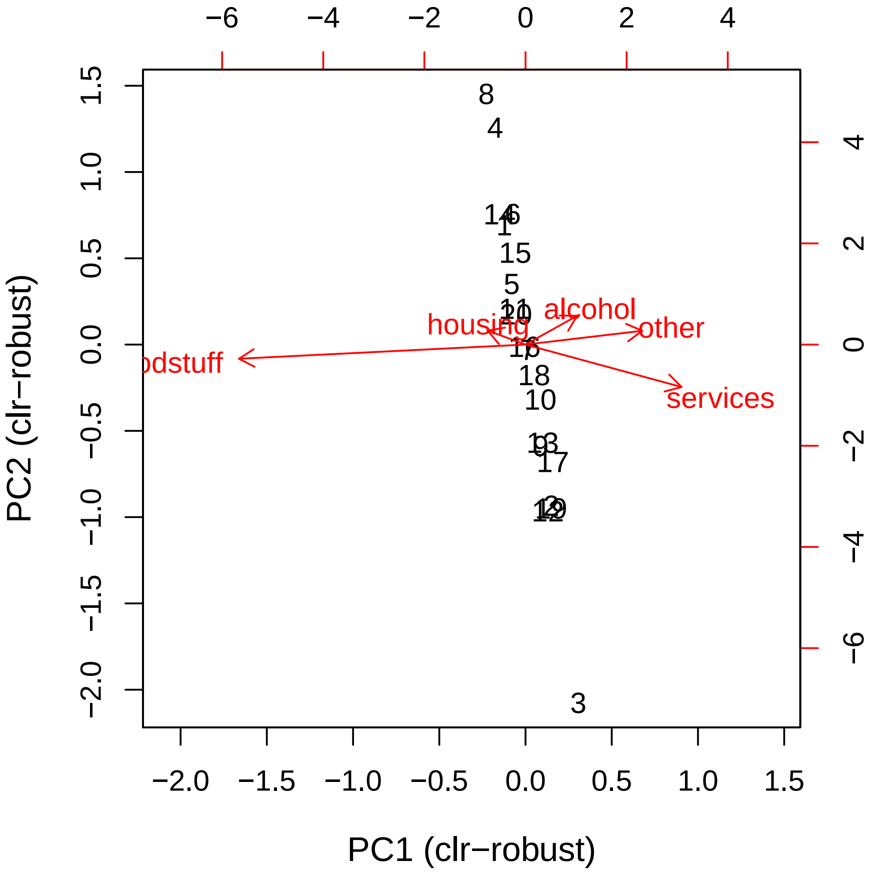

Figure 2: Robust compositional biplot for the expenditures data set.

References

[1] Aitchison J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall, London.

[2] Aitchison J., Greenacre M. (2002). Biplots of Compositional Data. *Applied Statistics*. Vol. **51**, pp. 375-392.

[3] Egozcue J.J., Pawlowsky-Glahn V., Mateu-Figueras G. (2002). Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*. Vol. **35**, pp. 279-300.

[4] Egozcue J.J., Pawlowsky-Glahn V. (2005). Groups of Parts and Their Balances in Compositional Data Analysis. *Mathematical Geology*. Vol. **37**, pp. 795-828.

[5] Egozcue J.J., Pawlowsky-Glahn V. (2006). Simplicial Geometry for Compositional Data. In: Buccianti A., Mateu-Figueras G. and Pawlowsky-Glahn V. (Eds.). *Compositional Data Analysis in the Geosciences: From Theory to Practice*. Geological Society, London, pp. 795-828.

[6] Filzmoser P., Hron K. (2008). Outlier detection for compositional data using robust methods. *Mathematical Geosciences*. Vol. **40**, pp. 233-248.

[7] Filzmoser P., Hron K., Reimann C. (2009). Principal component analysis for compositional data with outliers. *Environmetrics*. Vol. **20**, pp. 621-632.

[8] Gabriel K.R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*. Vol. **58**, pp. 453-467.

[9] R development core team (2009). R: A language and environment for statistical computing, Vienna. *http://www.r-project.org.*

[10] Rousseeuw P., Van Driessen K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*. Vol. **41**, pp. 212-223.

[11] Templ M., Hron K., Filzmoser P. (2010) robCompositions: Robust Estimation for Compositional Data. *Manual and package*, version 1.3.3. *http://cran.r-project.org/package=robCompositions*

[12] Van den Boogaart G., Tolosana R., Bren M. (2008). compositions: Compositional Data Analysis. *Manual and package*, version 1.01-1. *http://www.stat.boogaart.de/compositions*