Chapter 8
Robustness for Compositional Data

Peter Filzmoser and Karel Hron

8.1 Introduction

Many real-world multivariate data sets are of compositional nature, which means that not the absolute reported information in variables but their ratios are informative. This situation frequently occurs in geochemistry, but also in biosciences or economics and many other applications (Pawlowsky-Glahn and Buccianti 2011). When analyzing a chemical composition of a rock, not the absolute values of the mass of the compounds (which depends on the size of the sample), but ratios provide a relevant picture of the multivariate data structure. Such observations, called in the following compositional data (or compositions for short), are popularly represented in proportions or percentages, i.e., as data with a constant sum constraint. This fact lead in the past to a mismatch of the concept of compositional data with that of constrained observations. In the latter case, natural requirements valid for compositional data are not met, like scale invariance (the information in a composition does not depend on the particular units in which the composition is expressed) and sub-compositional coherence (information conveyed by a full composition should not be in contradiction with that coming from a sub-composition), see, e.g., Egozcue (2009) for details. The presented more general (and also a more natural) definition

P. Filzmoser (✉)
Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstrasse 8-10, Vienna, Austria
e-mail: P.Filzmoser@tuwien.ac.at

K. Hron
Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, 17. listopadu 12, Olomouc, Czech Republic
e-mail: hronk@seznam.cz

K. Hron
Department of Geoinformatics, Faculty of Science, Palacký University, tř. Svobody 26, Olomouc, Czech Republic

C. Becker et al. (eds.), *Robustness and Complex Data Structures*,
DOI 10.1007/978-3-642-35494-6_8, © Springer-Verlag Berlin Heidelberg 2013

was established in Aitchison (1986), where the logratio approach to the statistical analysis of compositional data was introduced. Since the specific properties of compositions naturally induce their own geometry (the Aitchison geometry) on the simplex, the sample space of compositions, the main effort is devoted to express the compositions in orthonormal coordinates, where the usual Euclidean rules already hold (Egozcue and Pawlowsky-Glahn 2006), and accommodate the standard statistical methods for their analysis.

Also in the case of compositional data, outlying observations can completely destroy results of a statistical analysis, comparing to those obtained from the homogeneous majority of observations in a data set. However, the specific geometric behavior of compositional data induces a different view of outliers compared to the usual case. For example, now obviously an observation with high absolute values on the compounds (parts) must not necessarily be an outlier, if the corresponding ratios between its parts follow the dominant data behavior. For this reason, not only the classical statistical methods, but even the robust ones cannot be applied directly to raw compositional data. This would lead (among other problems mentioned below) to a mismatch of regular and outlying observations.

In the following, a concise way how to perform multivariate statistical analysis of compositional data using the logratio approach will be presented. The next section introduces the Aitchison geometry on the simplex, together with the main inherent concepts of compositional data analysis and a way how to express (and interpret) the compositions in orthonormal coordinates. Section 8.3 shows how the classical and robust versions of standard multivariate methods like outlier detection, principal component analysis, correlation analysis and discriminant analysis can be applied to a compositional data set, together with interpreting the corresponding results. A real-world example on geochemical (compositional) data in presented in Sect. 8.4, and Sect. 8.5 concludes.

8.2 Geometrical Properties of Compositional Data

In addition to scale invariance and subcompositional coherence, another crucial property that characterizes compositional data is represented by their relative scale. The concept of relative scale naturally occurs already for most positive univariate data sets (Mateu-Figueras and Pawlowsky-Glahn 2008): although Euclidean distances within two pairs of samples taken at two rain gauges, {5; 10} and {100; 105} in mm are the same, quite probably, in the first case most observers would say there was double the total rain in the second gauge compared to the first, while in the second case they will say it rained a lot but approximately the same. Similarly, Euclidean distance between two multivariate observations does not reflect the relative growth in the compositional parts (concentrations of chemical elements, household expenditures on various commodities, etc.). Unfortunately, standard statistical methods completely ignore the relative scale concept since they rely on the usual Euclidean geometry in the real space (Eaton 1983). For this reason, the geometrical behavior of D-part compositional data $\mathbf{x} = \mathcal{C}(x_1, \ldots, x_D)'$ is characterized by the Aitchison geometry, defined for compositions $\mathbf{x}$, $\mathbf{y}$ and a real constant a by the operations perturbation, $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, \ldots, x_D y_D)'$, powering, $a \odot \mathbf{x} = \mathcal{C}(x_1^a, \ldots, x_D^a)'$ and the Aitchison inner product,

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}$$

(note that the sums run over all log-ratios of the parts in the compositions $\mathbf{x}$ and $\mathbf{y}$), resulting in the Aitchison norm

$$\|\mathbf{x}\|_a = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_a}$$

and the Aitchison distance,

$$d_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_a.$$

The symbol $\mathcal{C}$ stands for the closure operation that rescales the resulting compositions to an arbitrarily chosen, but fixed constant sum constraint κ, like 100 in case of percentages, corresponding to the actual representation of the compositions in the D-part simplex,

$$\mathcal{S}^D = \left\{ (x_1, \ldots, x_D)',\ x_i > 0,\ \sum_{i=1}^{D} x_i = \kappa \right\}.$$

Because all the relevant information in compositional data is contained in ratios between the parts, it is natural that zero compositional parts are not allowed for the analysis. According to the character of the occurrence of zeros, either as a result of an imprecise measurement of a trace element in the composition (i.e., rounding zeros) or the result of structural processes (structural zeros), special care has to be taken prior to a further processing of the observations (Aitchison and Kay 1999; Martín-Fernández et al. 2012).

The Aitchison geometry forms a Euclidean vector space of dimension $D - 1$, so it is possible to express the compositions in coordinates with respect to an orthonormal basis on the simplex, i.e., to obtain orthonormal coordinates of compositional data. The corresponding mapping $h : \mathcal{S}^D \to \mathbf{R}^{D-1}$, that results in the real vector $h(\mathbf{x}) = \mathbf{z} = (z_1, \ldots, z_{D-1})'$, moves the Aitchison geometry to the standard Euclidean geometry in the real space (with vector addition, multiplication by a scalar and the Euclidean inner product), $h(\mathbf{x} \oplus \mathbf{y}) = h(\mathbf{x}) + h(\mathbf{y})$, $h(a \odot \mathbf{x}) = a \cdot h(\mathbf{x})$ and $\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle h(\mathbf{x}), h(\mathbf{y}) \rangle_e$. These properties are the reason why the mapping h is usually referred to as the isometric logratio (ilr) transformation (Egozcue et al. 2003).

Obviously, there are infinitely many possibilities how to choose the orthonormal basis on the simplex and construct the orthonormal coordinates. Unfortunately, there is no canonical basis on the simplex (D original compositional parts are represented by only $D-1$ new coordinates), so that interpretable alternatives are needed. One possible choice represents the concept of balances (Egozcue and Pawlowsky-Glahn 2005), that enables an interpretation of the orthonormal coordinates in the sense of balances between groups of compositional parts. However, their construction usually assumes a prior knowledge of the studied problem. For this reason, we present below an "automated" version of balances, as described in Filzmoser et al. (2012a), that frequently occurs in different contexts of compositional data analysis (Hron et al. 2010, 2012). Explicitely, we obtain $(D-1)$-dimensional real vectors $\mathbf{z}^{(l)} = (z_1^{(l)}, \dots, z_{D-1}^{(l)})', l = 1, \dots, D,$

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}}, \quad i = 1, \dots, D-1, \tag{8.1}$$

where $(x_1^{(l)}, x_2^{(l)}, \dots, x_l^{(l)}, x_{l+1}^{(l)}, \dots, x_D^{(l)})$ stands for such a permutation of the parts of $\mathbf{x}$ that always the l-th compositional part fills the first position, $(x_l, x_1, \dots, x_{l-1}, x_{l+1}, \dots, x_D)$. In such a configuration, the first ilr variable $z_1^{(l)}$ explains all the relative information (log-ratios) about the original compositional part x_l. The coordinates $z_2^{(l)}, \dots, z_{D-1}^{(l)}$ then explain the remaining log-ratios in the composition (Fišerová and Hron 2011). Note that the only important position is that of $x_1^{(l)}$ (because it can be fully explained by $z_1^{(l)}$). The other parts can be chosen arbitrarily, because different ilr transformations are orthogonal rotations of each other (Egozcue et al. 2003). Of course, we cannot identify $z_1^{(l)}$ with the original compositional part x_l, but it explains all the information concerning x_l. Without loss of generality, we identify $z_i^{(1)} = z_i$ and use this simplified notation throughout the paper. Finally, the inverse ilr transformation, defined as $\mathbf{x} = \mathcal{C}(x_1, \dots, x_D)' = h^{-1}(\mathbf{z})$, where

$$\begin{aligned}
x_1 &= \exp\left(\frac{\sqrt{D-1}}{\sqrt{D}} z_1\right), \\
x_i &= \exp\left(-\sum_{j=1}^{i-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j + \frac{\sqrt{D-i}}{\sqrt{D-i+1}} z_i\right), \\
&\quad i = 2, \dots, D-1, \\
x_D &= \exp\left(-\sum_{j=1}^{D-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j\right),
\end{aligned} \tag{8.2}$$

is used to express the orthonormal coordinates back on the simplex.

For most statistical methods, an interpretation of the compositional data analysis in orthonormal coordinates is fully satisfactory. An exception is the biplot of principal component analysis which is related to the centered logratio (clr) transformation, resulting for a composition $\mathbf{x} = \mathcal{C}(x_1, \ldots, x_D)'$ in a real vector

$$\mathbf{y} = (y_1, \ldots, y_D)' = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right)'.$$

Elements of $\mathbf{y}$ represent coefficients with respect to a generating system of compositions, i.e., the covariance matrix of a random composition $\mathbf{y}$ is positive semidefinite and thus the clr transformed data are not appropriate for a robust statistical analysis. Fortunately, there exists a linear relation between the clr coefficients and orthonormal coordinates (like those from (8.1)), $\mathbf{y} = \mathbf{V}\mathbf{z}$. The columns of the $D \times (D-1)$-matrix $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_{D-1})$ are formed by the clr transformation of the orthonormal basis vectors, resulting in coordinates $\mathbf{z}$, concretely

$$\mathbf{v}_{D-i} = \sqrt{\frac{i}{i+1}} \left(0, \ldots, 0, 1, -\frac{1}{i}, \ldots, -\frac{1}{i} \right)', \quad i = 1, \ldots, D-1.$$

Note that there is an important relation between the clr variables and the ilr transformation used in (8.1), namely $y_l = \sqrt{\frac{D}{D-1}} z_1^{(l)}$, $l = 1, \ldots, D$. This means that each clr variable captures all the relative information about the compositional part x_l, and y_l is proportional to $z_1^{(l)}$.

Although measures of location and variability of a random composition can even be expressed directly on the simplex using the mean value of the Aitchison distance as the center and the total variance,

$$\mathrm{cen}(\mathbf{x}) = \mathrm{argmin}_{\boldsymbol{\eta}} \mathsf{E}\left[d_a^2(\mathbf{x}, \boldsymbol{\eta})\right], \qquad \mathrm{totvar}(\mathbf{x}) = \mathsf{E}\left[d_a^2\big(\mathbf{x}, \mathrm{cen}(\mathbf{x})\big)\right],$$

respectively (Pawlowsky-Glahn and Egozcue 2002), it is usually preferred to capture location and variability of compositions directly in coordinates using the expectation $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. The variances and covariances of coordinates can be expressed using variances of log-ratios of parts of the original composition, $\mathrm{var}(\ln \frac{x_i}{x_j})$, $i, j = 1 \ldots, D$, that are advantageous for interpretation purposes (Fišerová and Hron 2011). If a sample $\mathbf{z}_1, \ldots, \mathbf{z}_n$ is given for the coordinate $\mathbf{z}$, one usually arrives at the arithmetic mean $\overline{\mathbf{z}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i$ and the sample covariance matrix $\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{z}_i - \overline{\mathbf{z}})(\mathbf{z}_i - \overline{\mathbf{z}})'$. Clearly, these characteristics have a breakdown point of zero, and thus they are highly sensitive to outlying observations, see the contribution by Müller, Chap. 5. Consequently, proper robust alternatives are needed, see the contribution by Rousseeuw and Hubert, Chap. 4. Because of different representations of compositions in coordinates, the affine equivariance of the corresponding estimators is crucial. In the following, we use the Minimum Covariance Determinant (MCD) estimator (Maronna et al. 2006), which is advantageous in particular for computational reasons (Rousseeuw and Van Driessen 1999). Using the relations

$$\mathrm{cen}(\mathbf{x}) = h^{-1}\big(\mathsf{E}\left[h(\mathbf{x})\right]\big), \qquad \mathrm{totvar}(\mathbf{x}) = \mathrm{trace}(\boldsymbol{\Sigma}),$$

the MCD algorithm can be directly used to estimate the center and the total variance of compositional data.

8.3 Multivariate Statistical Methods for Compositional Data

The scope of the statistical analysis of compositions does not differ from the case of standard multivariate data. Starting with an unsupervised case, the analyst is interested in identifying groups and main patterns of the data as well as in detecting outliers that depart from the main data cloud. Here, principal component analysis and outlier detection provide a first insight into the multivariate data structure. In the second step, natural groups of observations and their variables are analyzed in a supervised manner using discriminant analysis and correlation analysis, respectively, although the latter one can also be advantageously applied to support the unsupervised case. The main aspects and interpretation of the above mentioned methods are described in the following sections. Since all the methods strongly rely on a proper (robust) estimation of location and covariance, special care will be devoted to the compositional data specificity also in this context.

8.3.1 Outlier Detection

Outlier detection belongs to starting points of each exploratory data analysis, as outliers give valuable information on data quality, and they are indicative of atypical phenomena. Outlier detection methods usually assume a certain type of (theoretical) distribution of the main data cloud; in case of an elliptical distribution, the corresponding criterion can be based on the Mahalanobis distance, defined for a sample of compositions in coordinates as

$$\mathrm{MD}(\mathbf{z}_i) = \left[(\mathbf{z}_i - \mathbf{t})'\mathbf{C}^{-1}(\mathbf{z}_i - \mathbf{t})\right]^{1/2}, \quad i = 1, \ldots, n. \tag{8.3}$$

Here, $\mathbf{t} = \mathbf{t}(\mathbf{z}_1, \ldots, \mathbf{z}_n)$ and $\mathbf{C} = \mathbf{C}(\mathbf{z}_1, \ldots, \mathbf{z}_n)$ are location and covariance estimators, respectively. The choice of the estimators is crucial for the quality of multivariate outlier detection. Taking the classical estimators arithmetic mean and sample covariance matrix often leads to useless results, because these estimators themselves are influenced by deviating data points. For this reason, robust counterparts, like the above mentioned MCD estimator, need to be taken that downweight the influence of outliers on the resulting location and covariance estimation statistics, see the contributions by Mosler (Chap. 2) and Rousseeuw and Hubert (Chap. 4). Under the assumption of multivariate normal distribution on the simplex, i.e., normal distribution of the orthonormal coordinates (Mateu-Figueras and Pawlowsky-Glahn 2008), the (classical) squared Mahalanobis distances follow a chi-square distribution with $D - 1$ degrees of freedom, see, e.g., Maronna et al. (2006). This distribution might also be considered for the robust case, and a quantile, e.g., 0.975, can be used as a cut-off value separating regular observations from outliers. In case of compositional data, the values of the Mahalanobis distances do not depend on the chosen coordinates, i.e., also $\mathrm{MD}(\mathbf{z}_i) = \mathrm{MD}(\mathbf{z}_i^{(l)}),\ l = 2, \ldots, D$, if affine equivariant estimators of location and covariance are used (Filzmoser and Hron 2008).

A more advanced approach for the cut-off value was used in Filzmoser et al. (2012a). This accounts for the actual numbers of observations and variables in the data set, and it tries to distinguish among extremes of the data distribution and outliers coming from a different distribution. The paper also provides an overview of graphical methods for an interpretation of the multivariate outliers, available in the R package `mvoutlier`. In particular, the coordinates $z_1^{(l)}$ can be advantageously used for univariate presentations in order to reveal outliers, connected with one or more concrete compositional parts (see Sect. 8.4 for an example).

8.3.2 Principal Component Analysis and the Compositional Biplot

Also for the well-known method principal component analysis (PCA), the proper estimation of location ($\mathbf{t}$) and covariance ($\mathbf{C}$) plays an important role. Let $\mathbf{C} = \mathbf{G_z L G_z'}$ denote a spectral decomposition of the estimated covariance matrix $\mathbf{C}$, with the diagonal matrix $\mathbf{L}$ of eigenvalues and the matrix $\mathbf{G_z}$ of eigenvectors of $\mathbf{C}$. Then PCA results in a linear transformation

$$\mathbf{z}_i^* = \mathbf{G}_\mathbf{z}'(\mathbf{z}_i - \mathbf{t}), \tag{8.4}$$

of the coordinates into new variables (principal components) such that the first principal component has the largest possible variance (accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. Although both scores $\mathbf{z}_i^*$ and loadings (columns of the matrix $\mathbf{G_z}$) of the principal components could also be interpreted directly in orthonormal coordinates, it is rather common to transform the loadings back to the clr space, $\mathbf{G_y} = \mathbf{V G_z}$, where the affine equivariance property of the MCD estimator was utilized (Filzmoser et al. 2009). The scores in the clr space are identical to the scores of the ilr space, except that the additional last column of the clr score matrix has entries of zero. Finally, the transformed loadings and scores are used to construct the biplot of clr transformed compositional data (Aitchison and Greenacre (2002), also called "compositional biplot"). Although the purpose of the compositional biplot is the same as for the standard one (Gabriel 1971), i.e., to provide a planar graph that represents a rank-two approximation of both the observations (PCA scores, plotted as points) and variables (loadings, rays) of multivariate data, its interpretation is different: The main interest is in the links (distances between vertices of the rays); concretely, for the rays i and j ($i, j = 1, \ldots, D$) the link approximates the log-ratio variance $\mathrm{var}(\ln \frac{x_i}{x_j})$. Hence, when the vertices coincide, or nearly so, then the ratio between x_i and x_j is constant, or nearly so. Consequently, this characteristic replaces the thinking in terms of correlation coefficients between two coordinates (standard variables). In addition, directions of the rays signalize where observations with dominance of the corresponding compositional part are located.

8.3.3 Correlation Analysis

Correlation analysis is not applicable to the original compositional parts. An exception is the case where compositional data (or constrained data) are modeled with a Dirichlet distribution. There, correlation analysis it is still a popular choice for expressing the strength of a linear relation between the parts of a positive vector with constant sum constraint. However, the interpretation of the resulting correlation coefficients is misleading due to negative bias of the covariance structure of constrained observations (Aitchison 1986). In fact, a correlation analysis of compositions is meaningful only if it is applied to orthonormal coordinates. Here the concept of balances provides a possibility of interpretable results, when the groups of compositional parts are clearly stated, like based on the results of the compositional biplot. However, practical experiences show that also in such cases a deeper insight to the studied problem is usually required (Filzmoser and Hron 2009).

An "unsupervised version" of correlation analysis is given by the multiple correlation coefficient applied to the special choice of balances according to (8.1). The multiple correlation coefficient ϱ_l between a balance $z_1^{(l)}$ and the remaining balances $\mathbf{z}_l = (z_2^{(l)}, \ldots, z_{D-1}^{(l)})'$ can be expressed as

$$\varrho_l^2 = 1 - \frac{|\boldsymbol{\Sigma}|}{\mathrm{var}(z_1^{(l)}) \cdot |\boldsymbol{\Sigma}_l|}, \quad l = 1, \ldots, D, \tag{8.5}$$

where $\boldsymbol{\Sigma}_l$ denotes the covariance matrix of $\mathbf{z}_l$ (Johnson and Wichern 2007). The estimation of the theoretical characteristics can again be performed using the MCD estimator. Note that any rotation of $\mathbf{z}_l$ (another choice of the corresponding balances) would lead to the same value of the coefficient ϱ_l (Filzmoser and Hron 2009). The interpretation of the multiple correlation coefficient for the above chosen coordinates can be directly derived from the standard case and the properties of balances. As the balance $z_1^{(l)}$ explains all the relative information (log-ratios) about the compositional part x_l and the vector $\mathbf{z}_l$ expresses orthonormal coordinates for the subcomposition $\mathcal{C}(x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})'$, the coefficient $\varrho_l \in [0, 1]$ can be interpreted as a measure of strength of the linear relation between relative information on x_l and the rest of the composition. Consequently, small values of ϱ_l signalize an exceptional behavior of the part x_l with respect to the other compositional parts. In addition, a big difference between the classical and robust version of the coefficient indicates that the possible relation is driven by outliers.

8.3.4 Discriminant Analysis

Discriminant analysis is a tool for supervised classification. It can be directly applied after any orthonormal coordinates are chosen to express the compositions in the real space (Filzmoser et al. 2012b). This holds for both well established rules, the Bayesian and the Fisher discriminant rule (Johnson and Wichern 2007). For the purpose of exploratory analysis, the Fisher rule seems more convenient. Having g groups of observations with p_j as their prior probabilities ($\sum_{j=1}^{g} p_j = 1$), their expectations $\boldsymbol{\mu}_j$ and the covariance matrices $\boldsymbol{\Sigma}_j$, $j = 1, \ldots, g$, the Fisher rule is based on the matrix $\mathbf{B}$ describing the variation between the groups, and the matrix $\mathbf{W}$ that stands for the within groups covariance matrix. If the notation $\boldsymbol{\mu} = \sum_{j=1}^{g} p_j \boldsymbol{\mu}_j$ for the overall weighted mean of all populations is used, the two matrices are defined as

$$\mathbf{B} = \sum_{j=1}^{g} p_j(\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})', \qquad \mathbf{W} = \sum_{j=1}^{g} \boldsymbol{\Sigma}_j.$$

Under the assumption of equal group covariance matrices, it can be shown that the best separation of the group means can be achieved by maximizing

$$\frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{W}\mathbf{a}} \quad \text{for } \mathbf{a} \in \mathbf{R}^{D-1}, \ \mathbf{a} \neq \mathbf{0}, \tag{8.6}$$

in the sample version preferably by using robust estimates of $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$. The solution of this maximization problem is given by the eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_l$ of the matrix $\mathbf{W}^{-1}\mathbf{B}$, scaled so that $\mathbf{v}_i'\mathbf{W}\mathbf{v}_i = 1$ for $i = 1, \ldots, l$. The number l of strictly positive eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ turns out to be $l \leq \min(g-1, D-1)$. The main advantage of the Fisher discriminant rule is its ability for dimension reduction. Concretely, by projecting the data in the space of the first two eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$, one obtains a data presentation in the plane that best captures the differences among the groups. Note that the resulting scores of these projections are invariant to the choice of the orthonormal coordinates, see Filzmoser et al. (2012b) for details.

8.4 Example

The methods described above are illustrated with a data set from geochemistry, the so-called *Oslo transect data*, see Reimann et al. (2007), and references therein. The data set is available in the R package `rrcov` as data `OsloTransect` (R Development Core Team 2012). Samples of different plant species (birch, mountain ash, fern and spruce) were collected along a 120 km transect running through Oslo (Norway), and the concentration of 25 chemical elements for the sample materials are reported. Since different parts were taken from the four sample materials (like leaves and bark), in total nine groups are available, where each group consists of 40 samples.

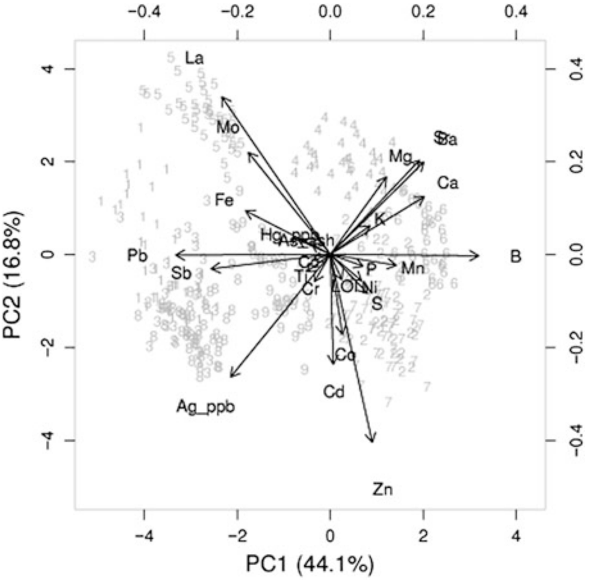

Fig. 8.1 Biplot of the first two robust principal components for all observations and variables. The numbers correspond to the data groups: birch bark (*1*), birch leaves (*2*), birch wood (*3*), fern (*4*), moss (**5**), European mountain ash leaves (*6*), spruce needles (**7**), spruce tree wood (*8*), spruce tree twigs (*9*)

The goal of this section is not a complete analysis of the *Oslo transect data*, but rather an illustration of the robust multivariate methods applied to subsets of this compositional data set. Concentrations of chemical elements are indeed compositions, because an increase of one part will automatically result in a decrease of the other parts, and thus ratios rather than absolute concentrations are of interest.

Figure 8.1 shows a compositional biplot of a robust PCA carried out with the complete data set. The first two principal components explain about 60 % of the total variability. Several variable groups can be identified, pointing at different processes in the plants. In addition, the data groups are well visible as score groups in the plot (each group is represented as an own number). Consequently, some groups show a higher dominance of certain variables, like a dominance of boron (B) in group 6 (European Mount Ash leaves).

Multivariate outlier detection is illustrated with the variable set Ba, Ca, Cu, Mn, Pb, Sr, Zn and the data group birch bark (the samples are now numbered from 1 to 40). In addition, 5 samples from the group birch wood (numbered as 41–45) are taken. Outlier detection should be helpful to identify the 5 samples from the different sample material. Indeed, applying the function `mvoutlier.CoDa` from the R package `mvoutlier` reveals the observations 5, 9, 25, 29, 40, 41, 42, 43, 44 and 45 as multivariate outliers. Several diagnostic plots are provided for finding an interpretation of the outliers. Figure 8.2 shows a parallel coordinate plot with the outliers shown by black lines. The presented axes are the ilr variables constructed with (8.1), and so they represent all the relative information of the corresponding chemical elements. It can be seen that the outliers have a very different behavior than the regular observations. For example, they are much less dominated by Cu and Mn.

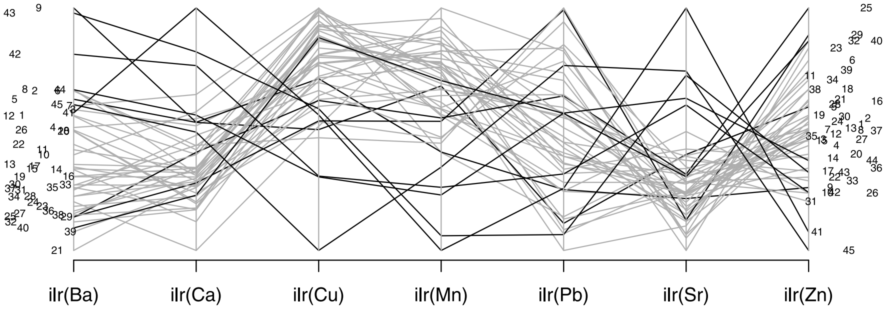

Fig. 8.2 Parallel coordinate plot for the ilr variables: multivariate outliers are shown by *black lines*; the numbers are the sample numbers: 1–40 for birch bark samples, and 41–45 for birch wood samples

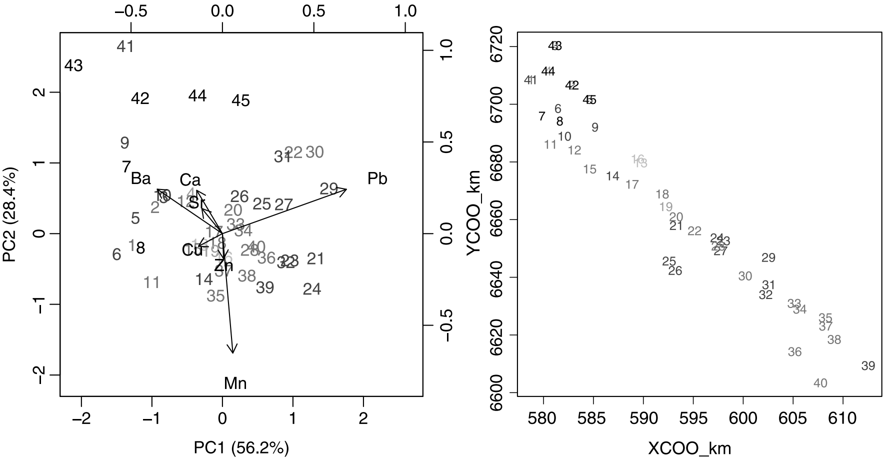

Fig. 8.3 Biplot (*left*) and plot of the spatial coordinates (*right*) of the data subset; the darker the symbols for the samples, the higher is their "degree of outlyingness"; the birch bark samples are numbered from 1–40, the birch wood samples from 41–45

Further plots for interpreting the multivariate outliers are shown in Figure 8.3. The biplot (left) reveals the group 41–45 as dominated by Ba and Ca, and with a shortage of Mn, compared to the other observations. The grey scale of the symbols corresponds to the "level of outlyingness", and it is also used in the further plots. On the right-hand side we see the samples plotted at their spatial coordinates. The birch wood samples originate from the North–West part of the region, and they are taken from the same locations as samples 1–5 (the symbols are over-plotted). Obviously, the sample material has more effect on the chemical composition than the origin of the samples.

The biplot in Figure 8.3(left) suggests that the variables are not highly related to each other. For example, the element Pb seems to be quite different from the other

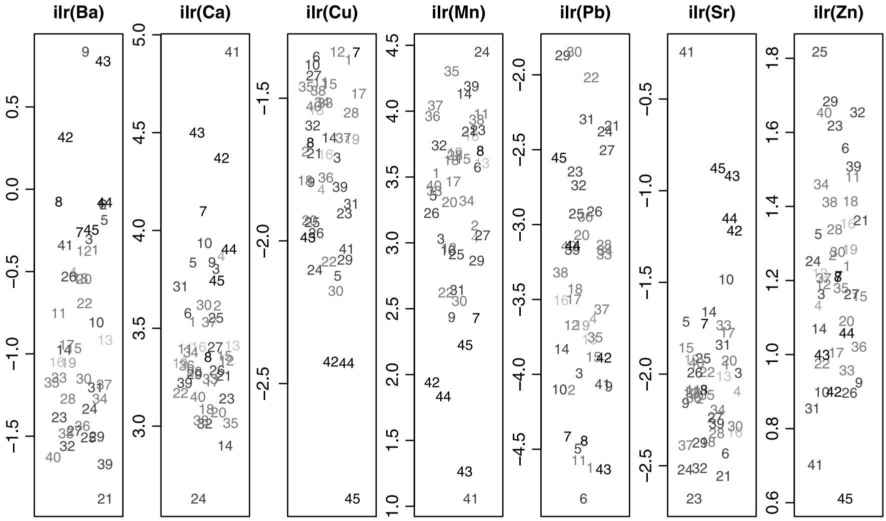

Fig. 8.4 Univariate scatter plots for the single ilr variables, including all the relative information of the corresponding chemical element; the outlier group 41–45 has a quite different chemical composition

elements. Computing usual correlations would not be appropriate here, but it is possible to compute the squared multiple correlation between Pb and the remaining elements using (8.5). The resulting robust (MCD-based) squared multiple correlation coefficient between the balance representing Pb and the remaining balances is 0.82, so not particularly high. In comparison, using classical estimators leads to a coefficient of 0.65, which reflects the influence of the outliers.

In a final outlier diagnostic in form of univariate scatter plots in Figure 8.4, we show again the ilr coordinates including all the relative information of an element, similar to Figure 8.2. This presentation can be used to see in detail the composition of the samples. The group 41–45 of birch wood samples is more dominant in Ba, Ca, and Sr, and much less dominant in Cu and Mn. Also some of the remaining outliers 5, 9, 25, 29, 40 can be found on extreme positions in this plot, justifying the reason for their outlyingness.

Robust discriminant analysis is applied using all observations of the *Oslo transect* data set, but only the variables Ba, Ca, Cu, Mn, Pb, Sr, Zn (they have no problem with rounding artefacts). Figure 8.5(left) shows the projection of the data onto the space of the first two Fisher discriminant directions. The different data groups (shown by different symbols) are quite well separated. However, single outliers are visible, and thus a robust analysis is useful. Figure 8.5(right) is an attempt to learn about the discrimination ability of the established discriminant rules. Here we use 5-fold cross-validation: The data set is randomly split into five parts of about equal size, and in each step 4 parts are used as training set and the remaining part as validation set. The discriminant rule is established from the training data and it is evaluated for the validation data. This is done until each of the five parts once had the

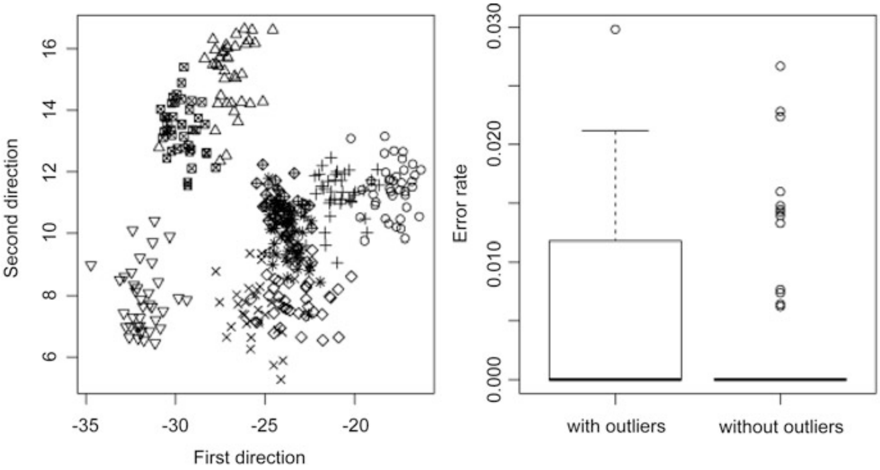

Fig. 8.5 Robust discriminant analysis using all samples (nine groups) and selected variables; the groups are well visible in the projection on the first two Fisher directions (*left*); the discriminant rules are evaluated with 5-fold cross-validation (*right*), by using all data in the validation sets (*left boxplot*) and removing the outliers from the validation sets (*right boxplot*)

role of the validation set. In each trial, we calculate the proportion of misclassified observations. In order to have a more general picture, the whole process is repeated 100 times, each time with new random assignments for 5-fold cross-validation. The left boxplot in the figure shows the resulting proportions of misclassification (error rate). Robust discriminant analysis is very effective here: the median error rate is zero, and at most 3 % error rate were reached. For generating the right boxplot we remove the outlying observations from the validation sets (they are still included in the training sets) by applying outlier identification in advance to each group separately. This gives a clearer picture of a robust discrimination rule, because outliers are not aimed to be classified correctly, in contrast to regular observations. The boxplot reveals that nearly all regular observations are correctly classified, and that the outliers in the training sets have no effect on the quality of the discrimination rules.

8.5 Conclusions

Compositional data, resulting from many real-world phenomena, require special transformations before the standard statistical tools can be used. Their characteristic is the fact that ratios between the variables carry the relevant information, and not the variable values directly. The suggested ilr transformation of (8.1) not only allows to analyze this relative information, but it is also helpful for the interpretation. Moreover, the resulting coordinates avoid the singularity problem of the clr transformation, and thus multivariate estimators like MCD can be applied. Since the MCD estimator is affine equivariant, the results from a robust multivariate analysis are not depending on the choice of the specific ilr transformation.

Tools for classical and robust compositional data analysis are available for example in the R packages `compositions`, `robCompositions`, and `mvoutlier`.

Acknowledgements The authors gratefully acknowledge the support by the Operational Program Education for Competitiveness—European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic).

References

Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman and Hall.

Aitchison, J., & Greenacre, M. (2002). Biplots of compositional data. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, *51*, 375–392.

Aitchison, J., & Kay, J. W. (1999). *Possible solutions of some essential zero problems in compositional data analysis*. Available at http://dugi-doc.udg.edu/bitstream/10256/652/1/Aitchison_Kay.pdf. Cited in March 12, 2012.

Eaton, M. (1983). *Multivariate statistics. A vector space approach*. New York: Wiley.

Egozcue, J. J. (2009). Reply to "On the Harker Variation Diagrams;..." by J.A. Cortés. *Mathematical Geosciences*, *41*, 829–834.

Egozcue, J. J., & Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, *37*, 795–828.

Egozcue, J. J., & Pawlowsky-Glahn, V. (2006). Simplicial geometry for compositional data. In A. Buccianti, G. Mateu-Figueras, & V. Pawlowsky-Glahn (Eds.), *Compositional data in the geosciences: from theory to practice* (pp. 145–160). London: Geological Society.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, *35*, 279–300.

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

Filzmoser, P., & Hron, K. (2009). Correlation analysis for compositional data. *Mathematical Geosciences*, *41*, 905–919.

Filzmoser, P., Hron, K., & Reimann, C. (2009). Principal component analysis for compositional data with outliers. *Environmetrics*, *20*, 621–635.

Filzmoser, P., Hron, K., & Reimann, C. (2012a). Interpretation of multivariate outliers for compositional data. *Computational Geosciences*, *39*, 77–85.

Filzmoser, P., Hron, K., & Templ, M. (2012b). Discriminant analysis for compositional data and robust parameter estimation. *Computational Statistics* doi:10.1007/s00180-011-0279-8.

Fišerová, E., & Hron, K. (2011). On interpretation of orthonormal coordinates for compositional data. *Mathematical Geosciences*, *43*, 455–468.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, *58*, 453–467.

Hron, K., Filzmoser, P., & Thompson, K. (2012). Linear regression with compositional explanatory variables. *Journal of Applied Statistics*, *39*, 1115–1128.

Hron, K., Templ, M., & Filzmoser, P. (2010). Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics & Data Analysis*, *54*, 3095–3107.

Johnson, R., & Wichern, D. (2007). *Applied multivariate statistical analysis* (6th ed.). London: Prentice-Hall.

Maronna, R., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics: theory and methods*. New York: Wiley.

Martín-Fernández, J. A., Hron, K., Templ, M., Filzmoser, P., & Palarea-Albaladejo, J. (2012). Model-based replacement of rounded zeros in compositional data: classical and robust approaches. *Computational Statistics & Data Analysis*, *56*, 2688–2704.

Mateu-Figueras, G., & Pawlowsky-Glahn, V. (2008). A critical approach to probability laws in geochemistry. *Mathematical Geosciences*, *40*, 489–502.

Pawlowsky-Glahn, V., & Buccianti, A. (2011). *Compositional data analysis: theory and applications*. Chichester: Wiley.

Pawlowsky-Glahn, V., & Egozcue, J. J. (2002). BLU estimators and compositional data. *Mathematical Geology*, *34*, 259–274.

R Development Core Team (2012). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reimann, C., Arnoldussen, A., Boyd, R., Finne, T. E., Koller, F., Nordgullen, O., & Englmair, P. (2007). Element contents in leaves of four plant species (birch, mountain ash, fern and spruce) along anthropogenic and geogenic concentration gradients. *Science of the Total Environment*, *377*, 416–433.

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