Austrian Journal of Statistics
January 2021, Volume 50, 16–37.
http://www.ajs.or.at/
doi:10.17713/ajs.v50i2.1067

Monitoring Robust Estimates for Compositional Data

Valentin Todorov
United Nations Industrial Development Organization (UNIDO)

Abstract

In a number of recent articles Riani, Cerioli, Atkinson and others advocate the technique of monitoring robust estimates computed over a range of key parameter values. Through this approach the diagnostic tools of choice can be tuned in such a way that highly robust estimators which are as efficient as possible are obtained. This approach is applicable to various robust multivariate estimates like S- and MM-estimates, MVE and MCD as well as to the Forward Search in which monitoring is part of the robust method. Key tool for detection of multivariate outliers and for monitoring of robust estimates is the Mahalanobis distances and statistics related to these distances. However, the results obtained with this tool in case of compositional data might be unrealistic since compositional data contain relative rather than absolute information and need to be transformed to the usual Euclidean geometry before the standard statistical tools can be applied. Various data transformations of compositional data have been introduced in the literature and theoretical results on the equivalence of the additive, the centered, and the isometric logratio transformation in the context of outlier identification exist. To illustrate the problem of monitoring compositional data and to demonstrate the usefulness of monitoring in this case we start with a simple example and then analyze a real life data set presenting the technological structure of manufactured exports. The analysis is conducted with the R package **fsdaR**, which makes the analytical and graphical tools provided in the MATLAB FSDA library available for R users.

Keywords: compositional data, forward search, robust estimates, outliers.

1. Introduction

In many cases the data sets are characterized by multivariate observations (vectors) containing relative contributions of parts to a whole. Examples are geochemical composition of rocks, household budget patterns, time budget, ceramic compositions. A plethora of further examples can be found in Aitchison (1986, 2005) and the hundreds of papers published on this topic. One example which was the motivation for this contribution is in order. Since 2002 the United Nations Industrial Development Organization (UNIDO) publishes the Competitive Industrial Performance (CIP) Index and accompanying report (Todorov and Pedersen, 2017), see http://stat.unido.org/cip. Through this index monitoring the industrial competitiveness of countries will, to a great extent, reflect how well they manage to adapt to these new challenges and embrace the opportunities. The CIP Index is an essential tool for countries to view and compare their industrial competitiveness with that of others. The CIP

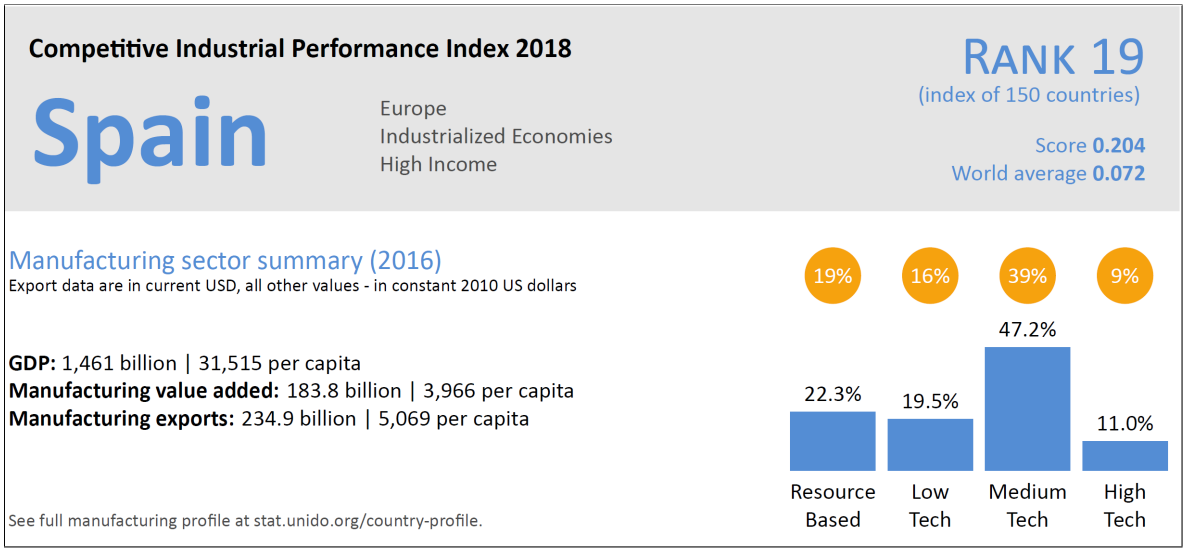

Figure 1: Example of a Competitive Industrial Performance (CIP) Index country profile. The bar chart in the top right corner represents the export structure.

Index is composed of eight sub-indicators defined within the framework of three key dimensions that capture different aspects of a country's industrial competitive performance. One of these sub-indicators is the technological structure of manufactured exports representing their "quality". There exists a well established decomposition analysis by technology level of the export structure (Lall, 2000) presenting the manufactured exports in four categories: Resource-based, Low technology, Medium technology and High technology (about the source of data and how these categories are defined see Todorov and Pedersen, 2017). Figure 1 is an excerpt from one country profile based on CIP edition 2018. The bar chart in the top right corner represents the export structure of the manufacturing exports of the country. The percentages shown sum up to 100. However, the country does not export only manufacturing goods, also agricultural, mineral, energy goods or services can compose the total exports. This is shown by the four circles above the corresponding bars — the percentages shown in the circles are the shares of the respective manufacturing category in the total exports. A vast number of works in the economics and development literature show that exports and their composition, particularly in terms of their technological intensity, are one of the most important economic growth conditions in a country or region (Crespo Cuaresma and Woerz (2005); Hausmann, Hwang, and Rodrik (2007); Raiher, Souza do Carmo, and Stege (2017)). Hausmann *et al.* (2007) show evidence that the economic growth is influenced by the composition of the exports agenda, and the more sophisticated exports (which they characterize as high productivity exports) are associated to higher economic growth rates. Crespo Cuaresma and Woerz (2005) investigate the hypothesis of qualitative differences between high and low tech exports with respect to output growth and show that the superior performance of high tech exports stems from their positive productivity differential to the domestic sector. However, to our knowledge, none of the studies on export structure consider them as compositional data although the focus is on the structure of the exports and the absolute values of the exports are not relevant for the analysis. This motivation example will be analyzed in more detail in Section 4.2 as a first step in this direction in which we will look in the following two points: (1) are there outliers in the data and (2) can we identify the presence of group patterns.

Compositional data have generally been defined as a vector of proportions, a vector with strictly positive components whose sum is a constant (Aitchison (1986); Aitchison (2005); Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado (2008); Pawlowsky-Glahn and Egozcue (2006)), however, this definition has changed and nowadays we refer to any set of multivariate observations with strictly positive components where relative rather than absolute information is relevant for the analysis (Pawlowsky-Glahn, Egozcue, and Lovell (2015a); Pawlowsky-

Glahn, Egozcue, and Tolosana-Delgado (2015b); Egozcue and Pawlowsky-Glahn (2019)). Due to the relative character of compositional data, application of standard statistical multivariate methods, which mostly rely on Euclidean geometry, might lead to misleading results, therefore, it is usual to apply a suitable transformation. A first remedy would be a log-transformation which will often reduce data skewness, but will not accommodate the compositional nature of the data. Aitchison (1986) suggested several possible transformations from the family of logratio transformations: the *additive log-ratio (alr)* and the *centered log-ratio (clr)* transformations, but these present certain inconvenience for the multivariate analysis (Argote-Espino, Lopez-García, and Facevicova (2018); Filzmoser, Hron, and Reimann (2012)). This disadvantages have led to the development of the *isometric log-ratio (ilr)* transformation which relates the geometry on the simplex directly to the Euclidean geometry (Egozcue, Pawlowsky-Glahn, Mateu-Figueras, and Barceló-Vidal, 2003).

After transforming suitably the compositional data we have a sample from a single population in which outliers are (possibly) present. To make the analysis insensitive to the influence of the outliers (if present) robust methods are called for. Many robust estimators for location and covariance have been introduced in the literature. Maybe the most popular is the Minimum Covariance Determinant (MCD), introduced by Rousseeuw (1985) for which a fast computational algorithm exists (Rousseeuw and Van Driessen, 1999). The Minimum Volume Ellipsoid (MVE) was also introduced by Rousseeuw (1985) and was even more popular than the MCD in the past. Nowadays it is recommended only as an initial estimator for S-estimation (Maronna, Martin, Yohai, and Salibian-Barrera, 2019). The multivariate S-estimates, which are a smooth version of the MVE estimator, were introduced by Davies (1987) and further studied by Lopuhaä (1989) (see also Rousseeuw and Leroy, 1987, p.263). The MM-estimator (Tatsuoka and Tyler, 2000) is an extension of the S estimator, that has high efficiency under multivariate normality. The orthogonalized Gnanadesikan-Kettenring (OGK) estimator of Maronna and Zamar (2002) is much faster than MCD, has high breakdown point and can work even with data containing more variables than observations.

If we perform a very robust analysis with 50 % breakdown point the analysis will be safeguarded against outliers, however this will result in an unnecessary low efficiency for clean data without any outliers. The usual approach to improve efficiency is to apply reweighting (for MCD) or to use MM-estimates instead of S-estimates. The *forward search (FS)* and related methods provide a tool for adaptively choosing a suitable combination of breakdown point and efficiency by *monitoring* a series of fits over a range of values of these quantities to the data, i.e. repeating the estimation process for different choices of the tuning parameters (Cerioli, Riani, and Atkinson (2009); Atkinson, Riani, and Cerioli (2004); Riani, Atkinson, and Cerioli (2009); Cerioli, Riani, Atkinson, and Corbellini (2018); Greco and Farcomeni (2015)). In FS we start from small subsets of data and observations which are close to the fitted model are added sequentially to the observations used in parameter estimation. At each step of this process, while the subset grows we monitor parameter estimates, test statistics and measures of fit such as the squared Mahalanobis distance (MD). The FS process is presented in monitoring plots which display the squared Mahalanobis distance of all observations at each step (at each intermediate, growing, subset of observations). Thus, for each observation in the data set a trajectory of MD is obtained and abrupt (visual) changes in these trajectories will indicate an abrupt change in the model (i.e. outliers are entering the model and distorting the estimates). In the case of S-estimates, in a similar manner we vary the breakdown point (*bdp*) from 0.5 (maximal breakdown point) to 0 (maximum likelihood estimation) by a step of, say, 0.01 and for each value of *bdp* we estimate the location and covariance and with these estimates compute the Mahalanobis distances for each observation—again these can be presented in a monitoring plot of squared Mahalanobis distances against *bdp* which will highlight any abrupt change in the MD trajectories. Continuing further with this analogy we obtain in the case of MM-estimates monitoring plots of squared MDs against efficiency which is varied from 0.5 to 0.99 by a suitable step. Such plots are presented in the examples, in Figure 8 and later.

The rest of the paper is structured as follows. Section 2 discusses the need of robust methods and presents briefly the forward search method, the monitoring of various methods and the available software for doing this. In Section 3 the specifics of robust methods in the case of compositional data are considered and are illustrated with a simple example. Section 4 presents the monitoring of compositional data on two examples and Section 5 concludes.

2. Forward search and monitoring of robust estimates

In statistical modeling and estimation assumptions like normal distribution or independence are used, however, the practice is usually different: practical data sets often do not follow these strict assumptions. There might be several different processes inherent in the data generating process or other effects that cannot be controlled. It is then often unclear how reliable the results are, if the model assumptions are violated. The multivariate aspect of the data used makes the task of outlier identification particularly challenging. The outliers can be completely hidden in one or two dimensional views of the data. This underlines that univariate outlier detection methods are useless, although they are often favored by researchers because of their simplicity.

Outlier detection and robust estimation are closely related (see Hampel, Ronchetti, Rousseeuw, and Stahel, 1986; Hubert, Rousseeuw, and van Aelst, 2008) and the following two problems are essentially equivalent:

- Robust estimation: develop statistical techniques which are inherently insensitive to the presence of outliers and find an estimate which is not influenced by these outliers, even if their amount is large (many good robust techniques can tolerate up to 50% contamination). The ability of the estimators to cope with large amount of outliers is measured by their *breakdown point* (*bdp*) which can reach the maximum of 50%. Estimators which can cope with this maximum amount of contamination in the sample are known as *high breakdown point estimators* (*HBDP estimators*) and examples of popular HBDP estimators are Minimum Covariance Determinant (MCD) estimator (Hubert, Debruyne, and Rousseeuw, 2017), S-and MM-estimators (Maronna *et al.*, 2019) as well as the Forward Search estimator (Cerioli, Farcomeni, and Riani, 2014).

- Outlier detection: find all outliers, which could distort the estimate. A classical approach to detecting multivariate outliers in a given sample $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^\top$ of n observations in the D-dimensional real space $\mathbb{R}^D$ would be to compute the Mahalanobis distance

$$MD(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^t \mathbf{S}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})}, i = 1, 2, \ldots, n \tag{1}$$

  for each $\boldsymbol{x}_i$. Here $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ are the sample mean and covariance matrix of the data set $\boldsymbol{X}$. Outliers may be identified by large values of $MD(\boldsymbol{x}_i)$. Since for multivariate normally distributed data, the squared Mahalanobis distances computed with the sample mean $\bar{\boldsymbol{x}}$ and covariance matrix $\boldsymbol{S}$, approximate a chi-square distribution χ^2 with D degrees of freedom (Seber, 1984), as a cut-off value can be taken the square root of a certain quantile β of the χ^2 distribution with D degrees of freedom. Then the customary adopted rule is to take $\beta = 0.975$ and points with Mahalanobis distances higher than this cut-off value $\sqrt{\chi^2_{D;0.975}}$ will be flagged as potential outliers (Maronna and Zamar, 2002). Unfortunately this approach suffers from two problems: (a) *Masking:* multiple outliers can distort the classical estimates of the mean $\bar{\boldsymbol{x}}$ and the covariance $\boldsymbol{S}$ in such a way (attracting $\bar{\boldsymbol{x}}$ and inflating $\boldsymbol{S}$) that they do not get necessarily large values of $MD(\boldsymbol{x}_i)$ and (b) *Swamping:* multiple outliers can distort the classical estimates of mean $\bar{\boldsymbol{x}}$ and covariance $\boldsymbol{S}$ in such a way that observations which are consistent with the majority of the data get large values of $MD(\boldsymbol{x}_i)$. To cope with these undesirable effects it is necessary to base the diagnostic tools on high breakdown point methods and

replace the classical Mahalanobis distances by their robust alternative

$$RD(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - \mathbf{T})^t \mathbf{C}^{-1} (\boldsymbol{x}_i - \mathbf{T})} \tag{2}$$

where $(\mathbf{T}, \mathbf{C})$ is a HBDP robust estimator of multivariate location and scatter (Rousseeuw and van Zomeren (1990); Maronna *et al.* (2019)).

A solution to the first problem allows us to identify the outliers using their robust distances while on the other hand, if we know the outliers we could remove or downweight them and then use the classical estimation methods. In many research areas the first approach is the preferred one but often the second one is more appropriate.

The exact distribution of the robust distances (2) is not known for finite sample sizes but the popular approach is to compare them to the same threshold $\sqrt{\chi^2_{D;0.975}}$ used for the classical Mahalanobis distance given in Equation (1) (Rousseeuw and Van Driessen (1999); Pison, Van Aelst, and Willems (2002)). Apart from the arbitrary choice of the 0.975 quantile, the χ^2 approximation is known to flag too many observations as outliers in data sets with small to moderate size. We will refer here to several alternatives to this approach for outlier identification. Hardin and Rocke (2005) suggested an improved approximation based on the F-distribution. Their results are based on the robust distances computed with the MCD estimator and do not consider the supposedly more efficient reweighted alternatives. A more advanced approach for choosing the cut-off value was proposed by Filzmoser, Garrett, and Reimann (2005) which accounts for the actual numbers of observations and variables in the data set, and also tries to distinguish among extremes of the data distribution and outliers coming from a different distribution. Based on this approach Filzmoser *et al.* (2012) proposed graphical tools for interpretation of multivariate outliers. Cerioli *et al.* (2009) proposed to calibrate the asymptotic cut-off values by Monte Carlo simulation. Calibration is first performed for some selected values of n and D and then extended to any n and D by parametric non-linear interpolation. Cerioli (2010) introduced an accurate approximation to the distribution of one-step reweighted robust distances. This approximation is based on a scaled *Beta*-distribution for the observations not suspected of being outliers, and on a scaled F-distribution for the units which are trimmed in the reweighting step.

When applying robust methods we want they to behave as similar as possible to the classical methods in case of clean data, therefore a first requirement of robust estimators is high *efficiency* at the specified model distribution. The loss of efficiency can be evaluated by the relative efficiency which is the ratio of variance of the robust and classical estimators. The second most important requirement for robust estimates is a high *breakdown point*: robust estimators need to resist a large amount of contamination before they become useless. Another important and desirable feature of statistical estimates is the *affine equivariance* which guarantees that the estimate will have a predictable behavior if the data were subjected to an affine transformation. For an $n \times D$ data matrix $\mathbf{X}$, a location estimate $\mathbf{T}(\mathbf{X})$ and a scatter estimate $\mathbf{S}(\mathbf{X})$ are called affine equivariant if, for any $D \times 1$ vector $\mathbf{a}$ and any non-singular square $D \times D$ matrix $\mathbf{B}$:

$$\begin{aligned} \mathbf{T}(\mathbf{X}\mathbf{B} + \mathbf{1}_n \mathbf{a}^\top) &= \mathbf{T}(\mathbf{X})\mathbf{B} + \mathbf{a} \\ \mathbf{S}(\mathbf{X}\mathbf{B} + \mathbf{1}_n \mathbf{a}^\top) &= \mathbf{B}^\top \mathbf{S} \mathbf{B} \end{aligned} \tag{3}$$

The vector $\mathbf{1}_n$ is $(1, \ldots, 1)^\top$ with n elements. If a Mahalanobis distance-like measure is computed using affine equivariant location vector $\mathbf{T}$ and covariance matrix $\mathbf{C}$ then the resulting distance measure is also affine equivariant. Most of the well known multivariate location and scatter estimates like MCD, MVE, S and MM are affine equivariant, while OGK is not.

In our study we assume that we have a single multivariate population possibly containing outliers. The straightforward approach is to use an estimator with 50% breakdown point, like MCD or S-estimator, however, in the case of clean data the results will be with very low

efficiency. A first remedy for this is to use reweighting (for MCD) or MM-estimates instead of S-estimates. For MCD the breakdown point and the efficiency are directly connected through the parameter $h, 1/2 \leq h < 1$—the number of observations on which the estimator is based (Maronna *et al.*, 2019). The larger h the lower breakdown point and the higher efficiency. For S estimates the results of Riani, Cerioli, and Torti (2014b) show asymptotic relationship between the breakdown point and efficiency: as one increases the other decreases. A solution to this dilemma should provide the MM-estimation introduced by Yohai (1987) which is a two step extension of the S-estimation. In the first step the breakdown point of the scale estimator is fixed to its maximum of 0.5 and in the second step the already obtained highly robust estimate is used for a new estimate of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for which the tuning constant of the ρ function can be chosen to provide high efficiency.

An alternative approach is the use of adaptive methods based on monitoring a series of fits to the data that indicate good choices of efficiency or breakdown point (Cerioli *et al.*, 2018; Riani, Atkinson, Cerioli, and Corbellini, 2019). The forward search for multivariate analysis is an algorithm for avoiding outliers by recursively constructing subsets of "good" observations, thus providing an automatic form of monitoring. Formal description of the forward search method can be found in many papers, see for example Riani *et al.* (2019). We start by choosing (by some robust criterion) a small subset of observations with size m_0 (usually $m_0 = D + 1$ or slightly larger) and then repeatedly extend it in such a way that outliers and other influential observations enter only toward the end of the search, arriving to the final fit that corresponds to the classical statistical estimates. During this process we monitor a suitable diagnostic measure and the inclusion of outliers is typically signalled by a sharp increase in this measure. Denote $S^{(m)}$ the subset used by the forward search at step m with $m = m_0, \ldots, n$. At step m the outlyingness of the observation $\mathbf{x}_i$ is evaluated by its squared Mahalanobis distance computed by Equation 1 with the current location and covariance estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ instead of $\bar{\boldsymbol{x}}$ and $\mathbf{S}$. $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are the sample estimates computed from the observations in $S^{(m)}$. The squared distances $d_1^2(m), \ldots, d_n^2(m)$ are ordered and the first $m+1$ observations are taken to form the subset $S^{(m+1)}$ for the next step. As long as $S^{(m)}$ does not contain any outlier neither masking nor swamping will impact the distances $d_i^2(m)$ and they can be taken as robust estimates of the population Mahalanobis distances. To detect outliers at step m we will look at the minimum Mahalanobis distance amongst observations not in the subset $S^{(m)}$: $d_{min}(m) = \min d_i(m), i \notin S^{(m)}$. To conduct the corresponding tests a reference distribution for $d_i(m)$ and $d_{min}(m)$ is required. Since $\hat{\boldsymbol{\Sigma}}$ is estimated from a subset of all observations the variability is underestimated and a scaling factor $c(m, n)$ is used in order to obtain an asymptotically unbiased estimate of $\boldsymbol{\Sigma}$ (Cerioli *et al.*, 2018). The *scaled* Mahalanobis distances are obtained by replacing $\hat{\boldsymbol{\Sigma}}(m)$ by $c(m, n)\hat{\boldsymbol{\Sigma}}(m)$ in Equation 1. The distributional results derived in Riani *et al.* (2009) lead to the distribution of $d_{min}(m)$ for a given m and allow to create forward plots of this quantity during the search and to compare with the *envelopes* (confidence bands) formed by the forward plots of several quantiles. If at some step m^* during the search the observation nearest to the observations already in the subset appears to be an outlier according to an appropriate envelope of the distribution of the test statistics, this event is called *"signal"*. If a "signal" occurs at observation m^*, this means that the observation m^* and all observations following it may be outliers. To precisely identify the outliers (after a "signal" occurs at step m^*), envelopes for a series of smaller sample sizes (starting with $m^* - 1$ onwards until an outlier is recognized) are superimposed. This process is automatic and all details are given in Riani *et al.* (2009). Further, in Section 4.2 these steps will be illustrated in the case of compositional data.

These underlying ideas can be extended to many other techniques like S- and MM-estimates. The subsequent estimations are presented in monitoring plots of all n squared Mahalanobis distances which can be combined with brushing to relate Mahalanobis distances to data points exhibited in the scatterplot matrix. In this way a straight relationship between statistical results and individual observations is established.

From a practical standpoint in data analysis the availability of such tools and their soft-

ware implementation is essential to make their applicability to a wide range of data analysis problems. All the methods discussed in a number of papers on forward search and monitoring are implemented in the *Flexible Statistics and Data Analysis (FSDA)* toolbox (Riani, Perrotta, and Torti, 2012), freely available for users with a MATLAB license at hand from `http://rosa.unipr.it`. It features robust and efficient statistical analysis of data sets, not only in multivariate context but also in regression and cluster analysis problems. An R package interfacing to the FSDA toolbox, **fsdaR** is available at CRAN (Todorov and Sordini, 2020). Todorov (2018) presents a brief study of the computational efficiency of the different monitoring methods and states that it is still an open issue and further work is necessary to make the monitoring of S-estimation for different consecutive values of breakdown point efficient for large data sets.

3. Compositional data and robust methods

Compositional data were defined traditionally as multivariate data with positive values that sum up to a constant (1 or 100 per cent or any other constant), i.e. constrained data (Aitchison, 1986). Nowadays this definition is generalized in more practical terms to any set of multivariate observations with strictly positive components where relative rather than absolute information is relevant for the analysis (Pawlowsky-Glahn *et al.* (2015a); Pawlowsky-Glahn *et al.* (2015b); Egozcue and Pawlowsky-Glahn (2019)). Mathematically a $D-$part compositional data set is a matrix $\mathbf{X}$ of n compositional vectors $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})^T$ in $\mathbb{R}^D$, $i = 1, \ldots, n$. A $D-$part simplex is defined as

$$S^D = \{\mathbf{x} = (x_1, \ldots, x_D)^T, x_i > 0, \sum_{i=1}^{D} x_i = \kappa\} \tag{4}$$

where $\kappa > 0$ is an arbitrary constant. This vector space is characterized by its own geometry called Aitchison geometry (Egozcue, Barceló-Vidal, Martín-Fernández, Jarauta-Bragulat, Díaz-Barrero, and Mateu-Figueras, 2011). Each compositional vector $\mathbf{x}$ can be rescaled by a constant c and then the compositions $\mathbf{x}$ and $\mathbf{y} = c\mathbf{x}$ are compositionally equivallent. This rescaling can be defined formally by the so called *closure* operator given by

$$\mathcal{C}(\mathbf{x}) = \frac{\kappa}{\Sigma_{i=1}^{D} x_i}(x_1, \ldots, x_D). \tag{5}$$

Even if the data are not subject to a constant sum constraint, standard statistical methods could lead to doubtful results if they are directly applied to the original data (Filzmoser, Hron, and Templ, 2018). A common practice to analyze compositions is to first project them onto the unconstrained real space by expressing them in logarithms of ratios between parts. Several types of transformations have been proposed in the literature and the choice of the most appropriate transformation depends on the data at hand and on the type of envisaged statistical analysis. Then standard statistical methods can be applied to the transformed data and the results are back transformed to the original space. The *additive logratio transformation (alr)* introduced by Aitchison (1986) maps a D-part composition in the simplex non-isometrically to a $D-1$ dimensional vector in $\mathbb{R}^D$, treating one part (usually the last one) as a common denominator of the others. While very easy to interpret, the disadvantage of this transformation is that distances are not preserved (because the corresponding basis on the simplex is not orthonormal with respect to the Aitchison geometry (Pawlowsky-Glahn *et al.*, 2008)). Another transformation, also proposed by Aitchison (1986) is the *centered logratio (clr)* transformation which maps a $D-$part composition from the simplex isometrically to $\mathbb{R}^D$, using the geometric mean as a common denominator. While very useful in terms of interpretation, this transformation produces observations with components which sum up to zero and thus the obtained data matrix has not full rank. The covariance matrix of the transformed data is singular which does not allow application of most of the multivariate

robust statistical methods. The desired orthonormality is fulfilled by the *isometric logratio (ilr)* transformations from S^D to R^{D-1} which, for a given basis can be defined as (Egozcue *et al.*, 2003):

$$z_i = \sqrt{\frac{i}{i+1}} ln \frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}} \text{ for } i = 1, \ldots, D-1. \tag{6}$$

Key tool for detection of multivariate outliers and for monitoring in the approach described in the previous Section are the Mahalanobis distances (see Equations 1 and 2) and statistics related to these distances. However, the results obtained with this tool in the case of compositional data might be unrealistic. If a transformation is applied, it is not clear how it will affect the Mahalanobis distances used for ranking the data points according to their outlyingness. Filzmoser and Hron (2008) considered these three well known transformations and showed how the additive, the centered and the isometric logratio transformations, will affect the Mahalanobis distances computed by classical and robust methods. They show that in case of classical location and covariance estimators all three transformations lead to the same Mahalnobis distances, however, only *alr* and *ilr* extend this property to any affine equivariant estimator. The former is simple and could be used in the context of outlier detection. In many of the examples in the next Sections *alr* produces the same results as *ilr*. However, the use of *alr* is not recommended because it does not result in an orthogonal basis system, which might be necessary for diagnostic tools following outlier detection or any further analysis which involves distances (Filzmoser *et al.*, 2012). If multivariate normal distribution on the simplex is assumed, i.e., the orthonormal coordinates are normally distributed (Mateu-Figueras and Pawlowsky-Glahn, 2008), the distribution of the (classical) squared Mahalanobis distances can be approximated by a chi-square distribution with D degrees of freedom as discussed in Section 2 and the same distribution might be used for the robust distances.

To illustrate the problem of applying robust methods to compositional data we start with a simple example based on the data set `Vegetables`. The data set, available in the R package **easyCODA** (Greenacre, 2019) contains the compositions of protein, carbohydrate and fat as a percentage of their respective totals for ten different vegetables and is shown in Table 1. Based on different types of logratios Greenacre (2019, p. 22, 31) shows that the vegetables fall into two groups: {Potatoes, Onions, Peas, Asparagus, Spinach, Broccoli} and {Beans (soya), Carrots, Corn, Mushrooms}. This is visible in a number of plots as well as can be confirmed by formal tests. As pointed out by Filzmoser and Hron (2008) for other compositional data sets,

Table 1: `Vegetables data set` from the R package **easyCODA**: composition of three nutrients in ten different vegetables

	Protein	Carbohydrate	Fat
Asparagus	35.66	61.07	3.27
Beans(soya)	42.05	35.88	22.07
Broccoli	48.69	43.78	7.53
Carrots	8.65	89.12	2.23
Corn	11.32	85.70	2.98
Mushrooms	16.78	77.25	5.97
Onions	10.44	88.61	0.95
Peas	26.74	71.29	1.97
Potatoes(boiled)	7.84	91.70	0.46
Spinach	41.57	52.76	5.67

we cannot apply standard outlier detection based on Mahalanobis distances, neither classical nor robust, directly to the data set, because the covariance matrix estimated from the original data is singular due to the fact that the data are subject to a constant sum constraint. Applying the outlier detection methods from the R package **rrcov** (Todorov and Filzmoser, 2009) as well as the methods from the MATLAB toolbox FSDA (Riani *et al.*, 2012) results in an

error. After applying *ilr* transformation to the data bivariate structure is revealed as shown in the distance-distance plot (Rousseeuw and Van Driessen, 1999) in Figure 2 (robust Mahalanobis distances computed by MCD are plotted against classical Mahalanobis distances). Four observations, namely mushrooms, carrots, corn and beans which form the second group, are identified as potential outliers (using the cutoff $\sqrt{\chi^2_{2:0.975}} = 2.7162$). The classical Mahalanobis distances, computed with the sample mean vector and covariance matrix do not identify any outlier.

As already described in Section 2 the MCD is not the only robust estimator which can be used for outlier detection. There exist other estimators which can be applied to data where the number of variables is larger than the number of observations and thus the covariance matrix is singular, like for example the Orthogonalized Gnanadesikan-Kettenring (OGK) estimator (Maronna and Zamar (2002); Todorov and Filzmoser (2009)). We can apply this estimator directly to the original data and obtain a robust estimate of the covariance matrix, however, this will not help for solving the outlier detection problem because the estimated robust covariance will again be singular and cannot be used for computing the Mahalanobis distances.

Another approach to outlier detection, especially when the number of variables is larger than the number of observations (and the covariance matrix of the data is singular) is to use robust principal component analysis (PCA). In the literature on robust compositional data analysis usually PCA based on MCD is used (Filzmoser *et al.* (2018)) but nothing could stop us from using ROBPCA (Hubert, Rousseeuw, and Vanden Branden (2005); Todorov and Filzmoser (2009)) which combines ideas from projection pursuit and MCD estimation and thus can work on data with singular covariance matrix. It should be noted that a disadvantage of the PCA for outlier detection is that it is not affine equivariant, only rotation equivariant. Robust PCA on the original data is shown in the right hand panel of Figure 2. It presents the so called orthogonal distances against the score distances computed from the PCA (see Hubert *et al.* (2005) for details about the definition of these distances and the cutoff values represented by the horizontal and vertical lines in the plot). Although ROBPCA works on data with singular covariance matrix and we could carry out the analysis on the original data in spite of the fact that the data are subject to constant sum constraint, the identified outliers are not the correct ones (i.e. ROBPCA identifies four vegetables as outliers but these observations actually belong to the larger group). It should be noted, however that if ROBPCA is carried out on the *ilr*-transformed data, it correctly identifies the four outlier.

One could be tempted to look at each variable separately and apply robust univariate methods for outlier detection. However, first of all, the compositional data can never be seen as truly univariate data because when we are looking at a single variable we actually observe not absolute values but relative information of this variable to all the rest in the composition. Therefore we talk not about variables but about parts of the composition. Discussion of the problems and possibilities of univariate statistical analysis of compositional data can be found in Filzmoser, Hron, and Reimann (2009).

Since the original data in this example are three-dimensional they can conveniently be presented in a ternary diagram (right hand panel of Figure 3). To better visualize the multivariate data structure we superimpose 0.975 tolerance ellipses of the Mahalanobis distances computed by the sample mean and covariance (blue) and by MCD (red) respectively. The ellipses are back-transformed to the original space using the inverse *ilr* transformation as proposed in Filzmoser and Hron (2008). The ellipse corresponding to the classical estimates (blue) covers all data points, while the robust one (red), based on MCD, excludes the four points identified as potential outliers.

4. Monitoring for compositional data

Now we will illustrate the methods and ideas presented in Sections 2 and 3 on two extensive

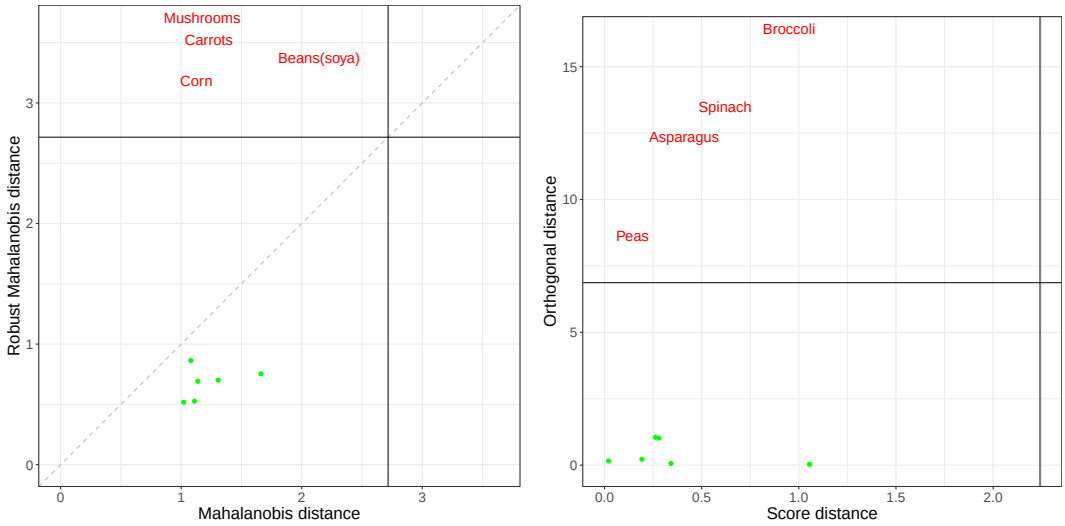

Figure 2: Vegetables data set. Robust Mahalanobis distance based on MCD versus classical Mahalanobis distance of the *ilr*-transformed data in the left hand panel. The horizontal and vertical lines are at the cut-off used (square root of the 0.975 quantile of the χ^2_2 distribution). The identified by the robust Mahalanobis distance outliers (in red) are the vegetables from the smaller group. Robust PCA (ROBPCA) on the original data in the right hand panel. Although ROBPCA works on data with singular covariance matrix the identified outliers are not the correct ones.

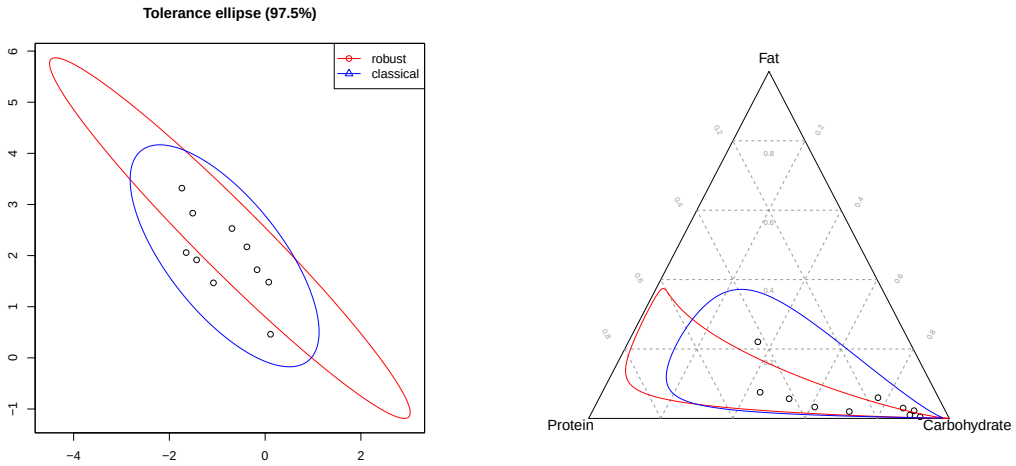

Figure 3: Vegetables data set, *ilr*-transformed. 0.975 tolerance ellipses based on the robust (MCD) and classical Mahalanobis distances in the left hand panel. The identified by the robust Mahalanobis distances outliers (outside the red ellipse) are covered by the classical (blue) ellipse. A ternary diagram with transformed Mahalanobis distances tolerance ellipses, classical and robust in the right hand panel.

examples. Both data sets were not analyzed previously in the literature in the context of outlier detection and we do not have any information about the presence of outliers. Therefore we start by the standard outlier detection methods in the R package **rrcov** based on MCD and robust Mahalanobis distances. Since both data sets are not subject to a constant sum constraint, the MCD (as well as other robust estimators) can be computed on the original data but the results might be doubtful as often pointed out in the literature. Therefore we first suitably transform the data using the *ilr* transformation. After demonstrating the outlier detection with MCD on the *ilr*-transformed data, we continue with S- and MM-estimation, as well as the Forward search, and the corresponding monitoring functions from the R package **fsdaR**. Finally we demonstrate the brushing and linking functionalities for establishing a straight relationship between statistical results obtained and the individual observations.

4.1. Example 1: Fish morphology data set

For our first example of illustrating the problem of monitoring compositional data we use the `FishMorphology` data set from the package **easyCODA** (Greenacre (2019), see also Greenacre and Primicerio (2010)). The data set consists of 26 morphometric measurements, in millimeters, on a sample of 75 fish of the species Arctic charr (*Salvelinus alpinus*). Additionally, to each observation, sex (male or female), habitat (littoral, close to the shore and pelagic, in deeper water far from the shore) and the body mass are recorded. For the sake our example we select only the former habitat (59 observations) and take the first 10 (out of the 26) morphometric measures because we want a single population of moderate size. It is a rule of thumb for the methods we are considering that for each dimension should be at least five observations. Of course we could choose any other set of variables and then the results might be different from the shown here but the principle of the illustration would not change. Thus we remain with a data set of 59 observations on the following 10 variables: {Bg, Bd, Bcw, Jw, Jl, Bp, Bac, Bch, Fc, Fdw}. It should be noted that in this data set there is one observation which is "obvious" outlier, the observation with $ID = 51$ which in our data set is number 16, i.e. it can be identified with any outlier detection method and it will appear in all graphs shown below. The data set is not subject to a constant sum constrained but nevertheless it is compositional in nature according to the more general definition (see Section 3). Greenacre (2019), the source of our data set, first closed the data, before starting the analysis, by applying the closure operator (Equation 5) but for our example this is not necessary.

Compositional data are by definition multivariate data, therefore it does not make sense to apply univariate outlier detection methods on the original data. We illustrate this by the boxplot of the original data set shown in the left-hand panel of Figure 4. A single outlier is visible, in variable Fdw, and this is observation #5. Univariate methods for outlier detection in compositional data can be used only on log-ratios or on coordinates, after transforming the data with a suitable transformation (Filzmoser *et al.*, 2018, p. 94).

Since the data are not subject to a constant sum constraint, as in the example in the previous Section, outlier detection methods based on robust estimates of location and covariance can be applied. A plot of the robust Mahalanobis distance based on MCD versus classical Mahalanobis distance is shown in the middle panel Figure 4—it identifies six outliers. A plot of the outlier detection using robust PCA is shown in the right-hand panel of Figure 4—only a single outlier is detected, observation #16. And finally, using OGK for outlier detection (not shown here) identifies eight observations as potential outliers. The drastic differences shown in these three plots are already cause of concern. As often mentioned in the literature, even when standard statistical analysis methods can be computed on the original compositional data (when the required covariance matrices are not singular) the results can be doubtful.

After applying *ilr* transformation, the multivariate structure is revealed as shown in the distance-distance plot in Figure 5 (robust Mahalanobis distances computed by MCD are plotted against classical Mahalanobis distances). Observations 20 and 55 are identified as

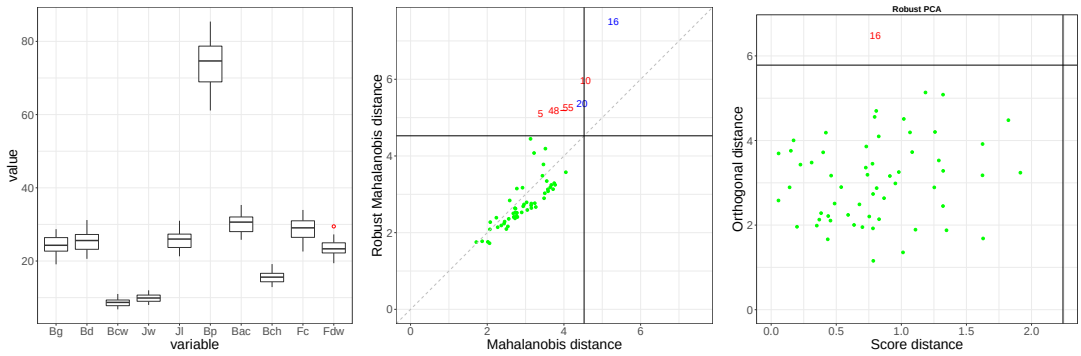

Figure 4: Fish Morphology data set (littoral habitat), original data. Parallel boxplots in the left-hand panel show a single outlier in the *Fwd* variable. Robust Mahalanobis distance based on MCD versus classical Mahalanobis distance in the middle panel—too many outliers. Robust PCA in the right-hand panel shows a single outlier.

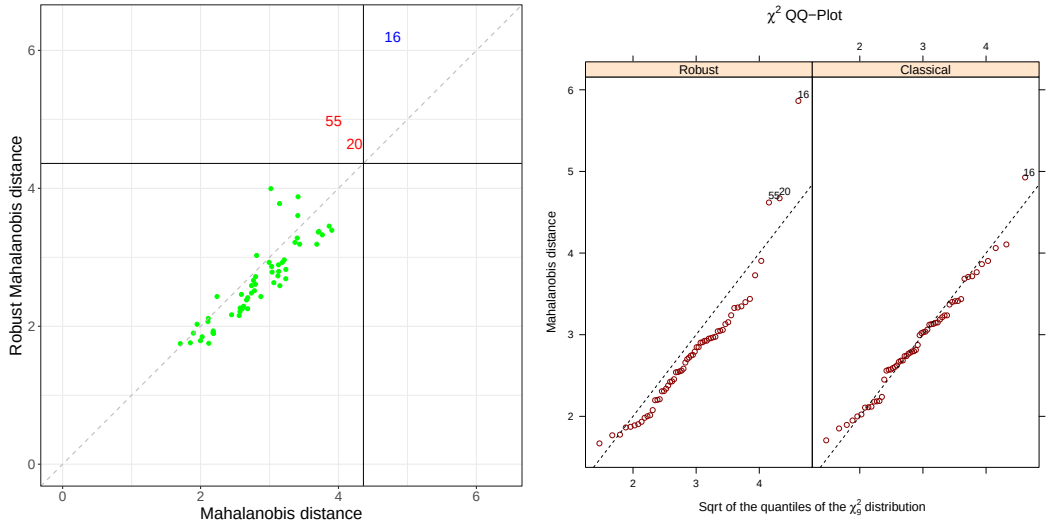

Figure 5: Fish Morphology data set (littoral habitat), *ilr* transformed. Robust Mahalanobis distance based on MCD versus classical Mahalanobis distance in the left hand panel. The horizontal and vertical lines are at the cut-off used (square root of the 0.975 quantile of the χ^2_9 distribution). A χ^2 QQ-plot, classical and robust, in the right hand panel.

potential outliers by the robust distances while observation 16 is flagged by both the classical and robust distances (using the 0.975 quantile of the χ^2_9 distribution as a cut-off in both cases). The right hand panel of Figure 5 shows the robust and classical χ^2 QQ-plots which present the squared (robust) Mahalanobis distances against the quantiles of the χ^2_9 distribution. The χ^2 QQ-plot is useful for visualizing the deviation of the data distribution from multivariate normality in the tails and is often used for outlier detection. For example Garrett (1989) and later Filzmoser *et al.* (2005) use these ideas to define adaptive outlier detection methods. In our case the result is identical to the plot shown in the left-hand panel, observation 16 is identified by both classical and robust distances and 20 and 55 are visible only in the robust method. Except for these observations, all the rest lie more or less on the diagonal line, confirming the χ^2-approximation. Since the original data are in more than three dimensions they cannot be conveniently presented in a ternary diagram as the `Vegetables` data set was presented in the right hand panel of Figure 3.

Computing the S-estimates of the multivariate location and covariance matrix of the *ilr* transformed data with 50% breakdown point and Tukey's biweight function (Fig. 6, left panel) produces similar result as the reweighted MCD, identifying the three outliers, however the S-estimates with reduced breakdown point (with the hope to obtain better efficiency) does

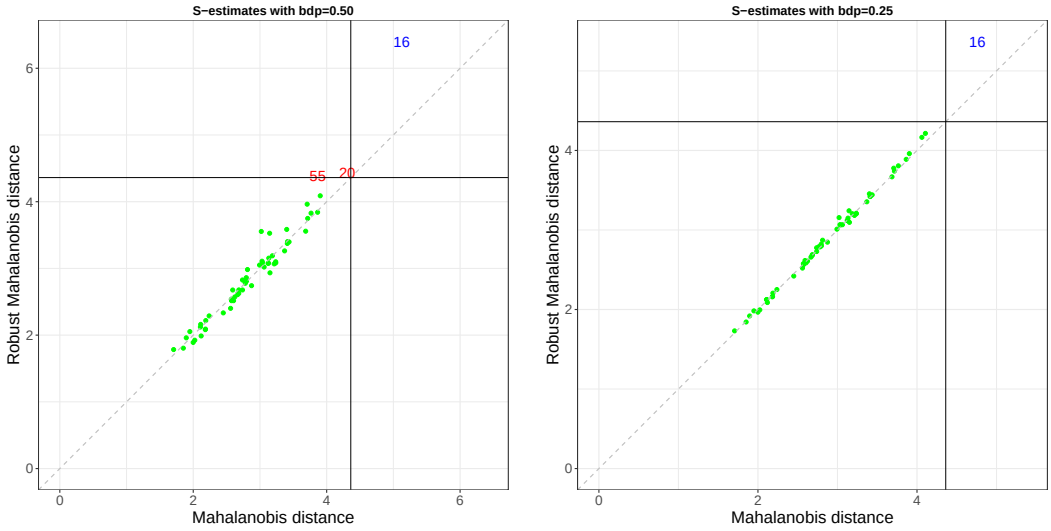

Figure 6: Fish Morphology data set (littoral habitat), *ilr* transformed. S-estimation with breakdown point 50% in the left-hand panel. In the right-hand panel—S-estimation with 25% breakdown point.

not identify any outliers (right panel in Fig. 6). Similarly, computing MM-estimates with 80% efficiency and Tukey's biweight function (Fig. 7, left panel) produces similar result as the reweighted MCD, however the MM-estimates with the efficiency 95% (which is often advocated and is the default value in most software packages) does not identify any outliers except observation 16, which, as we already noted is an obvious outlier and is identified by any method (right panel in Fig. 7). As Cerioli *et al.* (2018) point out, the recommended default efficiency of 95% or 99% for the MM-estimates might be too optimistic, also in our case. Following their approach for data driven balance between robustness and efficiency, which we apply to the case of compositional data, we present in the following the monitoring of the estimation parameters (breakdown point and efficiency) resulting in plots of the squared Mahalanobis distances of the *ilr* transformed data.

Figure 8 presents the monitoring of the MM-estimation. A series of robust MM fits are conducted by varying the efficiency from 0.5 to 0.99 by step of 0.01 (this is the default setup, but other configurations can be selected) and for each of them the robust distances (2) of all observations in the data set are computed. The estimated location and covariance matrix, the distances and other related statistics are stored for each step in the process. The left-hand panel shows the squared Mahalanobis distances presented against the values of efficiency at each step. Each line running horizontally from 0.5 to 0.99 represents the trajectory of one observation. The color of the lines varies from light blue to dark blue. The color becomes darker as the maxima of the individual trajectories increase. Thus the eye is drawn to the behaviour of the most outlying units represented by darker color. The red horizontal lines show the 0.975 and 0.99 quantiles of the squared χ^2 distribution with 9 degrees of freedom. The trajectories are stable until the efficiency reaches 0.54 when the fit changes abruptly and remains so until 0.77 when it changes again and becomes similar to the maximum likelihood and remains stable until the efficiency reaches 0.99. It reveals why the index plot of the MM-estimates in the right hand panel of Figure 7 did not show any outliers—for efficiency higher than 0.77 the fit is identical to the maximum likelihood. This is also clearly seen from the correlation monitoring in the right hand panel of Figure 8. It shows the monitoring of the three correlation measures which summarize the structure of the plot in the left-hand side by the correlation between the squared Mahalanobis distances at adjacent monitoring values. The three standard measures of correlation are:

1. *Spearman*: This is the correlation between the ranks of the two sets of observations.

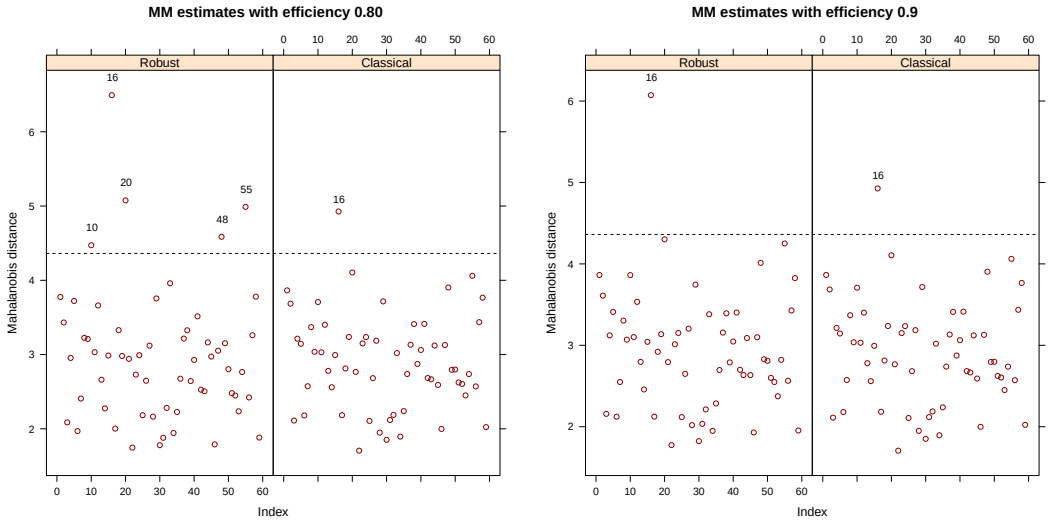

Figure 7: Fish Morphology data set (littoral habitat), *ilr* transformed. MM-estimation with efficiency 80% in the left hand panel. In the right hand panel - MM-estimation with 95% efficiency.

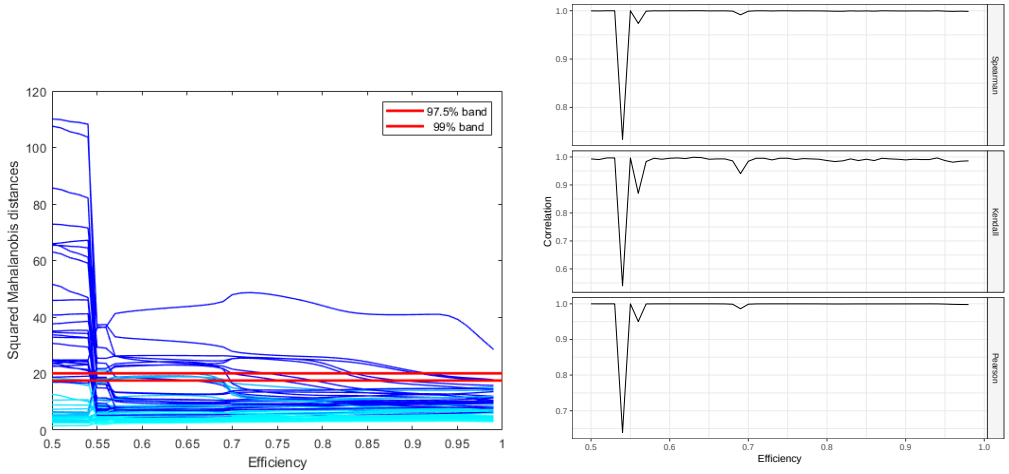

Figure 8: Fish Morphology data set (littoral habitat), *ilr* transformed. The left-hand panel shows the squared Mahalanobis distances from monitoring MM-estimation and the right-hand panel—the correlation between distances for consecutive values of *eff* (efficiency)

2. *Kendall*: The concordance of the pairs of ranks.

3. *Pearson*: The product-moment correlation coefficient.

All three correlations indicate the abrupt change at 54% and then a change at 77%. As we have already seen in Figure 7, for efficiency = 80% the analysis is still robust but increasing the efficiency to 95% and more results in a non-robust analysis.

Using the brushing functionality of the package, we can identify the outlying units, as shown in Figure 9: in the right hand panel the outliers are shown as red circles. A more advanced version of the brushing function allows to do the brushing in several steps, in each selecting different points. The points selected at each step are added to the points selected in the previous step but are presented in different pattern/color. In the first step we scan elect the outliers which affect the model results before 90% efficiency and they will be shown as red circles. In the second step the "obvious" outlier which can be identified also by the maximum likelihood estimates (MLE) can be selected and it will be shown as a light blue star.

Computing S-estimates with 50% (asymptotic) breakdown point and Tukey's biweight func-

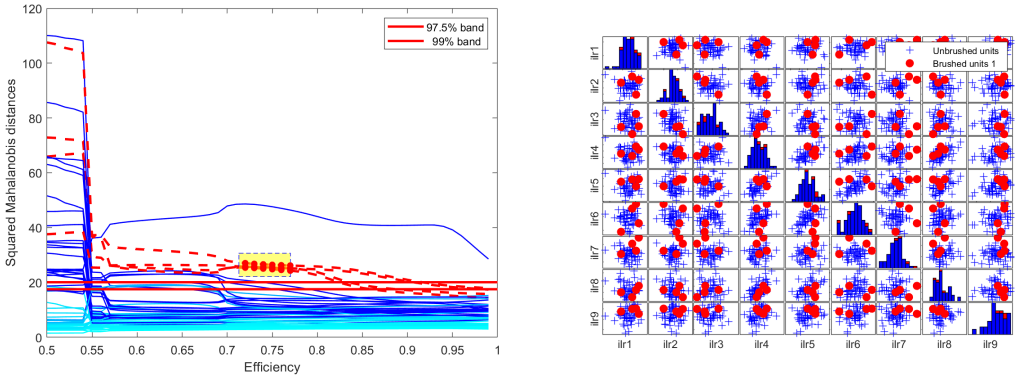

Figure 9: Fish Morphology data set (littoral habitat), *ilr* transformed. The left-hand panel shows brushing of the monitoring plot of MM-estimation and the right-hand panel—the scatter plot matrix of the ilr-transformed data, identifying the four outliers.

tion produces similar result as the reweighted MCD, however, this is not the case if the breakdown point is reduced to say 25%. As we have already seen in Figure 6, for breakdown point of 50% the analysis is robust but reducing the breakdown point to say 25% (with the hope to increase the efficiency), the result is identical to the maximum likelihood analysis. That is, with *bdp* 50% three outliers are detected and when we reduce the *bdp*, we expect that these three outliers will still be identified. However, this does not happen and thus, by reducing the *bdp* we destroyed the robustness of the estimator. This problem of S-estimates, that with increasing dimension, although having high breakdown point, the ability to detect outliers decreases, was pointed out already by Rocke (1996), see also (Maronna *et al.*, 2019, p. 190). The forward search methodology and its extension to S-estimates provides a solution: in this case it tells us that we cannot reduce the breakdown point without loosing the robustness of the estimator. Alternatively, to achieve higher efficiency, we should use MM-estimates but tuning their efficiency as shown above in our example.

4.2. Example 2: Technology intensity of exports

We turn now to the motivation example about the technological structure of manufactured exports presented in the introduction. The data set is available in the R package **rrcov3way** (Todorov, Simonacci, Di Palma, and Gallo, 2020). For our example we select only one year, 2017 and remove any country with missing data, remaining with 129 observations. Needles to say that applying the outlier detection methods from the R package **rrcov** or the methods from the MATLAB toolbox FSDA to the original data are meaningless: the reweighted MCD, for example, identifies 63 outliers out of 129 observations (49%). The data set is not subjected to a constant sum constraint and the covariance matrix is not singular; thus a robust covariance matrix (MCD, S or MM) can be computed, however, it is ill-conditioned and the outlier detection results in flagging almost half of the data as outliers. After applying *ilr* transformation the structure is revealed as shown in the distance-distance plot in Figure 10. Now 18 observations are identified as outliers by the reweighted MCD estimator.

This is definitely a compositional data set (the four categories are parts of one whole) but the closure is not visible when inspecting the row sums. This is due to the fact that we consider only the manufactured exports while the countries also export agricultural, mining and other products. This demonstrates the problem of the so called *subcompositions* (Aitchison, 1986)—we cannot hope that the effect of the closure will disappear if not all parts are included in the analysis and an appropriate transformation is needed.

The original data in this example are four-dimensional and if we want to present them in a ternary diagram we need to get rid of one of the four dimensions. We have two options: (i) to remove the resource based category and remain with LT, MT and HT—actually the

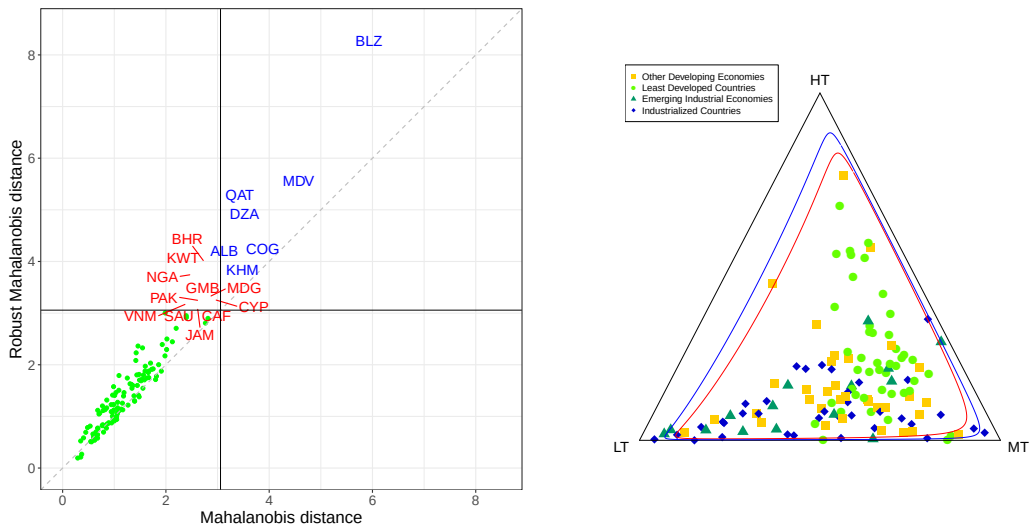

Figure 10: Technological structure of manufactured exports, *ilr* transformed. Robust Mahalanobis distance based on MCD versus classical Mahalanobis distance in the left hand panel. The horizontal and vertical lines are at the cut-off used (square root of the 0.975 quantile of the χ^2_3 distribution). A ternary diagram of the sub-composition {LT, MT, HT} with back-transformed Mahalanobis distance tolerance ellipses, classical (blue) and robust (red).

first version of the technology classification of the exports (Lall (1998); Lall (2000)) contained only these three categories and the processed foods like sugar, cheese, vegetable preparations were classified as primary products, not as resource based manufactures as in the present classification; (ii) alternatively, we can combine RB and LT into one category, at some risk of simplification, calling this category "easy" technologies, with the main drivers of competitiveness being natural resource endowments in the former case and low wages in the latter. We show in the right hand panel of Figure 10 the first option, with LT, MT and HT but the picture with option (ii) is similar, with the difference that the "LT+RB" corner looks a bit "heavier". As in Figure 3, to better visualize the multivariate data structure we superimpose 0.975 tolerance ellipses of the Mahalanobis distances computed by the sample mean and covariance (blue) and by MCD (red) respectively. The ellipses are back-transformed to the original space using the inverse *ilr* transformation as proposed in Filzmoser and Hron (2008). A legend shows the classification of the countries according to their development (Upadhyaya, 2013). The three Least developed countries identified by both classical and robust methods are Belize, Maldives and Cambodia. There are three industrialized countries identified as outliers - from the left hand panel of Figure 10 we see that these are Qatar, Bahrain and Kuwait, all oil producing countries. Only Qatar is identified by both the classical and the robust method. The two outlying observations from the category of emerging industrial economies visible by the lower left corner (LT) are Vietnam and Saudi Arabia. With Nigeria and Algeria we have six outlying countries which are among the top 20 oil producing countries, and two more, Congo and Vietnam come close.

We continue by running the automatic outlier detection procedure based on forward search and the top left panel of Figure 11 represents the corresponding plot. The magenta-colored line is the trajectory of the minimum Mahalanobis distance $d_{min}(m)$ amongst observations not in the subset $S^{(m)}$ for $m = m_0, \ldots, n$. The dashed lines present the *envelopes* computed for different quantiles at each step m of the search as described in Section 2, see also Riani *et al.* (2009) where a two stage procedure is described in detail. Through monitoring the bounds of all n observations we look if a "signal" will be obtained, say at step m^*, meaning that the value of the statistic lies beyond our threshold given by the envelope with significance level 99.99%. This will indicate that observation m^*, and therefore all succeeding observations, may be outliers. As visible in our plot in Figure 11 the "signal" for outliers occurs at observation 106,

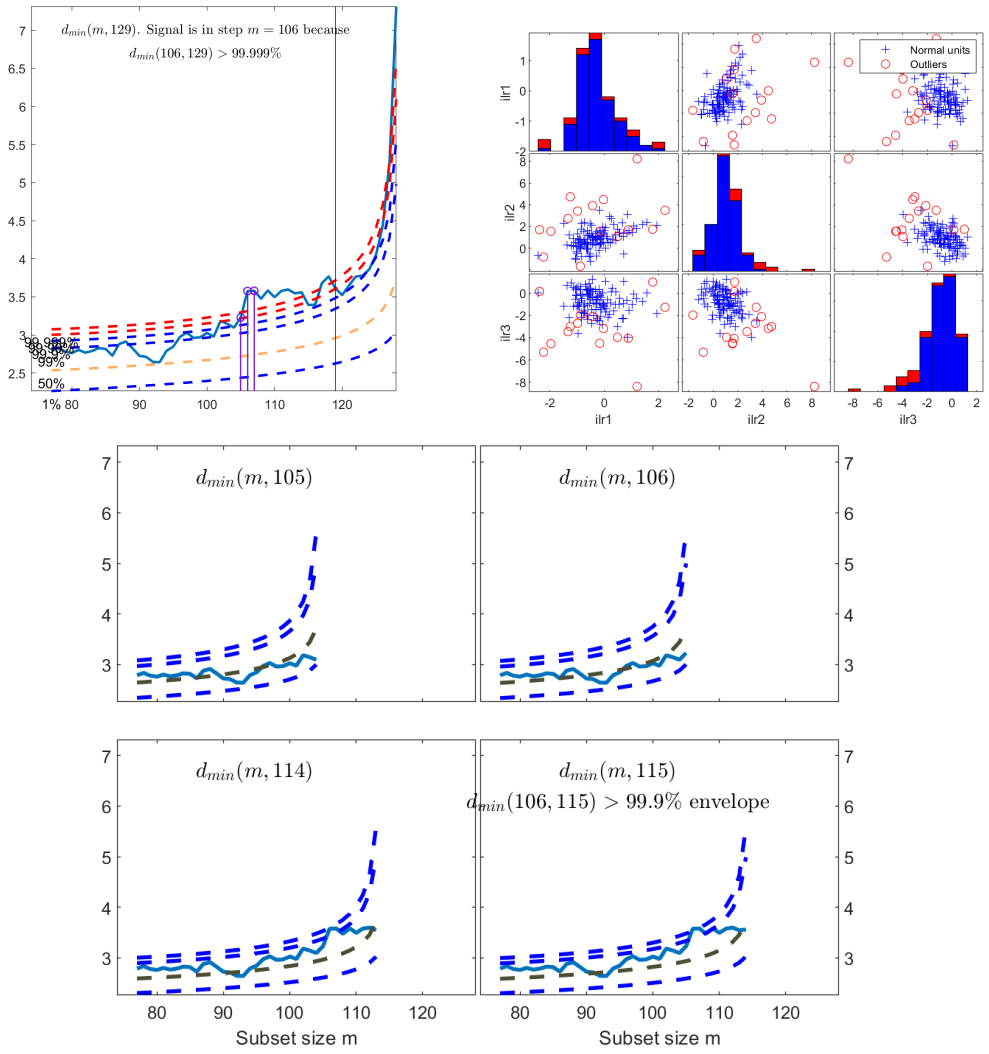

Figure 11: Technological structure of manufactured exports, *ilr* transformed. The top left-hand panel shows the forward search plot of minimum Mahalanobis distance $d_{min}(m)$, with a "signal" for the presence of outliers around step 106. The bottom for panels show resuperimpositions of envelopes computed for different sample size. The top right-hand panel shows the scatter plot of the data with the 15 observations identified as outliers by FS as red circles.

indicating that it and the succeeding observations might be outliers. In the second stage of the procedure we superimpose envelopes for values of n (the subsample size) from the point of the signal until the first time that we introduce an observation we recognize as an outlier. In the four bottom panels of Figure 11 we show the superimposition of envelopes for $n = 105$ (just before the signal) followed by $n = 106$ and $n = 114$ in which no outlier is detected. However, for $n = 115$ we see that $d_{min}(106)$ goes beyond the 99.9% envelope computed with $n = 115$. This means that there is a outlier-free subset with 114 observations and the remaining 15 observations are outliers. The identified outliers are shown in the top right panel of Figure 11 as red circles.

It turns out that these 15 outliers are quite similar but still different from the outliers detected by the MCD (Saudi Arabia, Vietnam, Pakistan, Madagascar and the Central African Republic are now not outliers but Namibia and Mongolia are identified additionally). Since in this automatic outlier detection procedure multiple outlier tests are conducted and we do not rely on the quite arbitrary threshold based on the 0.975 quantile as in the MCD outlier detection, we can conclude that the outlier list obtained is much more precise than the one obtained by MCD and shown in Figure 10. Performing the same analysis on the original, not transformed

data (not shown here) indicates a "signal" at observation 79 and identifies 51 observations as outliers (40%) which is similar to the result of MCD applied on the original data.

5. Conclusions

Robust methods are not only useful but also a required tool when analyzing real life data which very often are plagued by the presence of outliers. It does not matter if the robust methods are used directly to fit models or indirectly to identify outliers, some arbitrarily chosen parameters can have a destructive effect on the results. In a number of recent articles Riani, Cerioli, Atkinson and others advocate the technique of monitoring robust estimates computed over a range of key parameter values. Through this approach the diagnostic tools of choice can be tuned in such a way that highly robust estimators which are as efficient as possible are obtained. This approach is applicable to different robust multivariate estimates like S- and MM-estimates, MVE and MCD as well as to the Forward Search in which monitoring is part of the robust method. We show that in order to apply these adaptive methods to compositional data which are parts of some whole and often they are recorded as data subject to a constant-sum constraint, it is necessary to transform the compositional data. The affine equivariance of the key measure for detecting outliers, the Mahalanobis distance (when computed with affine equivariant robust estimates of location and scatter), ensures the invariance of the identified irregularities from the choice of transformation used (Filzmoser and Hron, 2008). We demonstrate on several examples how the monitoring can be conducted, providing highly efficient estimates and demonstrate the role of advanced dynamic graphics like brushing and linking for establishing a straight relationship between statistical results and individual observations. All computations were performed with the **fsdaR** package available at CRAN which brings almost all the functions of the MATLAB toolbox FSDA to the R user. Since the scope of these functions covers also robust regression (Riani, Cerioli, Atkinson, and Perrotta, 2014a) and robust clustering (Riani *et al.*, 2019) a natural extension of this study is to consider monitoring for robust regression and clustering for compositional data in the future.

Acknowledgements

The views expressed herein are those of the authors and do not necessarily reflect the views of the United Nations Industrial Development Organization (UNIDO).

References

Aitchison J (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK), London (UK).

Aitchison J (2005). *A Concise Guide to Compositional Data Analysis*. CoDaWeb. Lecture Notes. Available online at CoDaWeb.

Argote-Espino D, Lopez-García P, Facevicova K (2018). "Statistical Processing of Compositional Data. The Case of Ceramic Samples from the Archaeological Site of Xalasco, Tlaxcala, Mexico." *Journal of Archaeological Science: Reports*, **19**, 100–114.

Atkinson AC, Riani M, Cerioli A (2004). *Exploring Multivariate Data with the Forward Search*. Springer series in statistics. Springer-Verlag, New York.

Cerioli A (2010). "Multivariate Outlier Detection With High-Breakdown Estimators." *Journal of the American Statistical Association*, **105**(489), 147–156.

Cerioli A, Farcomeni A, Riani M (2014). "Strong Consistency and Robustness of the Forward Search Estimator of Multivariate Location and Scatter." *Journal of Multivariate Analysis*, **126**, 167–183.

Cerioli A, Riani M, Atkinson AC (2009). "Controlling the Size of Multivariate Outlier Tests with the MCD Estimator of Scatter." *Statistics and Computing*, **19**(3), 341–353.

Cerioli A, Riani M, Atkinson AC, Corbellini A (2018). "The Power of Monitoring: How to Make the Most of a Contaminated Multivariate Sample (with Discussion)." *Statistical Methods and Applications*, **27**, 559–587.

Crespo Cuaresma J, Woerz J (2005). "On Export Composition and Growth." **141**, 33–49.

Egozcue JJ, Barceló-Vidal C, Martín-Fernández JA, Jarauta-Bragulat E, Díaz-Barrero JL, Mateu-Figueras G (2011). "Elements of Simplicial Linear Algebra and Geometry." *Compositional Data Analysis: Theory and Applications*, pp. 141–157.

Egozcue JJ, Pawlowsky-Glahn V (2019). "Compositional Data: The Sample Space and Its Structure." *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, **28**(3), 599–638.

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003). "Isometric Logratio Transformations for Compositional Data Analysis." *Mathematical Geology*, **35**, 279–300.

Filzmoser P, Garrett RG, Reimann C (2005). "Multivariate Outlier Detection in Exploration Geochemistry." *Computers & Geosciences*, **31**, 579–587.

Filzmoser P, Hron K (2008). "Outlier Detection for Compositional Data Using Robust Methods." *Mathematical Geosciences*, **40**, 233–248.

Filzmoser P, Hron K, Reimann C (2009). "Univariate Statistical Analysis of Environmental (Compositional) Data: Problems and Possibilities." *Science of the Total Environment*, **407**, 6100–6108.

Filzmoser P, Hron K, Reimann C (2012). "Interpretation of Multivariate Outliers for Compositional Data." *Computational Geosciences*, **39**, 77–85.

Filzmoser P, Hron K, Templ M (2018). *Applied Compositional Data Analysis: With Worked Examples in R*. Springer series in statistics. Springer, Cham, Switzerland.

Garrett RR (1989). "The chi-square Plot: A Tool for Multivariate Outlier Recognition." *Journal of Geochemical Exploration*, **32**(1–3), 319–341.

Greco L, Farcomeni A (2015). *Robust Methods for Data Reduction*. Chapman and Hall/CRC, Boca Raton, London, New York.

Greenacre M (2019). *Compositional Data Analysis in Practice*. Chapman & Hall / CRC Press.

Greenacre M, Primicerio R (2010). *Multivariate Analysis of Ecological Data*. BBVA Foundation, Bilbao.

Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986). *Robust Statistics. The Approach Based on Infuence Functions*. John Wiley & Sons.

Hardin J, Rocke DM (2005). "The Distribution of Robust Distances." *Journal of Computational and Graphical Statistics*, **14**, 910–927.

Hausmann R, Hwang J, Rodrik D (2007). "What You Export Matters." *ournal of Economic Growth*, **12**, 1–25.

Hubert M, Debruyne M, Rousseeuw PJ (2017). "Minimum Covariance Determinant and Extensions." *WIREs computational statistics*, **10**.

Hubert M, Rousseeuw PJ, van Aelst S (2008). "High-Breakdown Robust Multivariate Methods." *Statistical Science*, **23**, 92–119.

Hubert M, Rousseeuw PJ, Vanden Branden K (2005). "ROBPCA: A New Approach to Robust Principal Component Analysis." *Technometrics*, **47**, 64–79.

Lall S (1998). "Exports of Manufactures by Developing Countries: Emerging Patterns of Trade and Location." *Oxford Review of Economic Policy*, **11**(2), 54–73.

Lall S (2000). "The Technological Structure and Performance of Developing Country Manufactured Exports, 1985-98." *Oxford Development Studies*, **28**(3), 337–369.

Lopuhaä HP (1989). "On the Relation Between S-Estimators and M-estimators of Multivariate Location and Covariance." *The Annals of Statistics*, **17**, 1662–1683.

Maronna RA, Martin D, Yohai V, Salibian-Barrera M (2019). *Robust Statistics: Theory and Methods (with R): Second edition*. John Wiley & Sons, New York.

Maronna RA, Zamar RH (2002). "Robust Estimation of Location and Dispersion for High-Dimensional Datasets." *Technometrics*, **44**, 307–317.

Mateu-Figueras G, Pawlowsky-Glahn V (2008). "A Critical Approach to Probability Laws in Geochemistry." *Mathematical Geosciences*, **40**, 489–502.

Pawlowsky-Glahn V, Egozcue JJ (2006). "Compositional Data and Their Analysis: An Introduction." In A Buccianti, G Mateu-Figueras, V Pawlowsky-Glahn (eds.), *Compositional Data Analysis in the Geosciences: From Theory to Practice*, pp. 1–10. London: Geological Society.

Pawlowsky-Glahn V, Egozcue JJ, Lovell D (2015a). "Tools for Compositional Data with a Total." *Statistical Modelling*, **2**(15), 175–190.

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2008). "Lecture Notes on Compositional Data Analysis." *Report*, Universitat de Girona, Gironna.

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015b). *Modeling and Analysis of Compositional Data*. John Wiley & Sons, New York.

Pison G, Van Aelst S, Willems G (2002). "Small Sample Corrections for LTS and MCD." *Metrika*, **55**, 111–123.

Raiher AP, Souza do Carmo AS, Stege AL (2017). "The Effect of Technological Intensity of Exports on the Economic Growth of Brazilian Microregions: A Spatial Analysis with Panel Data." *EconomiA*, **18**(3), 310–327.

Riani M, Atkinson AC, Cerioli A (2009). "Finding an Unknown Number of Multivariate Outliers." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**(2), 447–466.

Riani M, Atkinson AC, Cerioli A, Corbellini A (2019). "Efficient Robust Methods via Monitoring for Clustering and Multivariate Data Analysis." *Pattern Recognition*, **88**, 246–260.

Riani M, Cerioli A, Atkinson A, Perrotta D (2014a). "Monitoring Robust Regression." *Electronic Journal of Statistics*, **8**, 646–677.

Riani M, Cerioli A, Torti F (2014b). "On Consistency Factors and Efficiency of Robust S-estimators." *TEST*, **23**, 356–387.

Riani M, Perrotta D, Torti F (2012). "FSDA: A MATLAB Toolbox for Robust Analysis and Interactive Data Exploration." *Chemometrics and Intelligent Laboratory Systems*, **116**, 17–32.

Rocke DM (1996). "Robustness Properties of S-estimators of Multivariate Location and Shape in High Dimension." *The Annals of Statistics*, **24**, 1327–1345.

Rousseeuw PJ (1985). "Multivariate Estimation with High Breakdown Point." In W Grossmann, G Pflug, I Vincze, W Wertz (eds.), *Mathematical Statistics and Applications Vol. B*, pp. 283–297. Reidel Publishing, Dordrecht.

Rousseeuw PJ, Leroy AM (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons, New York.

Rousseeuw PJ, Van Driessen K (1999). "A Fast Algorithm for the Minimum Covariance Determinant Estimator." *Technometrics*, **41**, 212–223.

Rousseeuw PJ, van Zomeren BC (1990). "Unmasking Multivariate Outliers and Leverage Points." *Journal of the American Statistical Association*, **85**, 633–651.

Seber GAF (1984). *Multivariate Observations*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

Tatsuoka KS, Tyler DE (2000). "The Uniqueness of S and M-functionals under Nonelliptical Distributions." *The Annals of Statistics*, **28**, 1219–1243.

Todorov V (2018). "Discussion of "The Power of Monitoring: How to Make the Most of a Contaminated Multivariate Sample" by Andrea Cerioli, Marco Riani, Anthony C. Atkinson and Aldo Corbellini." *Statistical Methods & Applications*, **27**(4), 631–639.

Todorov V, Filzmoser P (2009). "An Object Oriented Framework for Robust Multivariate Analysis." *Journal of Statistical Software*, **32**(3), 1–47.

Todorov V, Pedersen AL (2017). "Competitive Industrial Performance Report 2016. Volumes I and II." *Report*, United Nations Industrial Development Organization (UNIDO), Vienna.

Todorov V, Simonacci V, Di Palma MA, Gallo M (2020). ***rrcov3way**: Robust Methods for Multiway Data Analysis, Applicable also for Compositional Data*. R package version 0.2.

Todorov V, Sordini E (2020). ***fsdaR**: Robust Data Analysis through Monitoring and Dynamic Visualization*. R package version 0.4-9.

Upadhyaya S (2013). "Country Grouping in UNIDO Statistics." *UNIDO Staff Working Paper*, Vienna.

Yohai VJ (1987). "High Breakdown-Point and High Efficiency Robust Estimates for Regression." *The Annals of Statistics*, **15**, 642–656.

Affiliation:

Valentin Todorov
United Nations Industrial Development Organization (UNIDO)
Vienna International Center
1140 Vienna, Austria
E-mail: v.todorov@unido.org

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 50
January 2021

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2019-12-01
Accepted: 2020-11-03