Stat Methods Appl (2018) 27:631–639
https://doi.org/10.1007/s10260-018-0424-4

ORIGINAL PAPER

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

Valentin Todorov[1]

Accepted: 25 February 2018 / Published online: 15 March 2018

Abstract The paper of Andrea Cerioli, Marco Riani, Anthony Atkinson and Aldo Corbellini is a fine review of the practical value of the forward search and the other related robust estimation methods based around monitoring of quantities of interest over a range of consecutive values of the tuning parameters. From a practical standpoint in data analysis the availability of such tools is essential, and the research reported in this paper has brought them to an wide audience. As a potential user of such tools I am particulary interested in their software implementation on one hand and their applicability to an wide range of data analysis problems. More precisely, I would like to address the following two points: (1) the software availability and computational issues related to monitoring and (2) monitoring in one special case, the case of compositional data.

Keywords Robust · Monitoring · Compositional data · R

1 Introduction

First I would like to thank Andrea Cerioli, Marco Riani, Anthony Atkinson and Aldo Corbellini for this paper (Cerioli et al. in the following) which made me aware of the practical value of the forward search and the other related robust estimation methods based around monitoring of quantities of interest over a range of consecutive values of the tuning parameters. The forward search for multivariate analysis is an algorithm for avoiding outliers by recursively constructing subsets of "good" observations and the underlying idea can be extended to many other techniques like S- and

✉ Valentin Todorov
v.todorov@unido.org

[1] United Nations Industrial Development Organization (UNIDO), Vienna, Austria

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 scatterplot matrices. 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 is essential, and the research reported in this paper has brought them to an wide audience. As a potential user of such tools I am particulary interested in their software implementation on one hand and their applicability to an wide range of data analysis problems. More precisely, I would like to address the following two points:

- The software availability and computational issues related to monitoring, considered in Sect. 2 and;
- Monitoring in one special case, the case of compositional data, considered in Sect. 3.

2 Software and computational issues

All the methods discussed in Cerioli et al. (and many more) are implemented in the 'Flexible Statistics and Data Analysis (FSDA)' toolbox, 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.

FSDA software for R (R Core Team 2017) *users* A downside of the current software implementing monitoring (FSDA) is that it is based on the commercial software MATLAB, which apart from its license costs, is not so appealing to the majority of the statistical community, where R is more widespread. The heart of the monitoring approach is the ability to present the result in a way revealing as much information as possible. While R has advanced graphical capabilities, these graphics are static, do not allow much interactivity and here is the main advantage of using MATLAB for implementing the monitoring functions. Developing the computational algorithms discussed in this paper for R would not be a problem and an R package (Atkinson et al. 2006) implementing forward search was available on CRAN in the past. However, the advantages provided by presenting the results visually in interlinked graphs allowing interaction with the user will be missing. Therefore, a possible solution for making the FSDA toolbox available to the R community would be not to port the toolbox, but to implement an R interface to a MATLAB engine running in the background. Such a technical solution is made possible by the MATLAB Runtime which allows to run compiled MATLAB applications on computers that do not have MATLAB installed. A prototype of an R package interfacing to the FSDA toolbox was presented by Sordini et al. (2016) which proved the concept and investigated the technical issues (creating a Java interface between an R package and a MATLAB toolbox running on the MATLAB Runtime). Additional technical challenge is how to extend a CRAN package with binaries, in this case the compiled Java code, but even more serious challenge is the design and implementation of the interface (the function calls) in a way acceptable for an R user. Formula interface, optional/default arguments to the functions, object orientation, documentation are just several topics presenting differences between MATLAB and R. For example an R user will prefer to call a method `plot()` on an object returned by a function, instead of passing optional arguments `(…, 'plots', 1, …)` to the function. Similarly, an R user will not be happy to follow strict positioning of the (mandatory) arguments as this is done in MATLAB and will prefer to use the formula interface where appropriate. Colors and color names, line types and other graphical parameters is also an area requiring a lot of effort to make the two languages compatible. But we hope that when all these problems are solved an R package implementing the monitoring functionality combined with advanced dynamic graphics will be available at CRAN and this will be very soon.

Computational efficiency Almost all robust estimation methods are computationally intensive and the computational effort increases with increasing number of observations n and number of variables p towards the limits of feasibility. Since the key idea in the discussed paper is to monitor quantities of interest, such as parameter estimates, measures of discrepancy and test statistics, as the model is fitted to data subsets of increasing size, it is inevitable that the computational effort grows exponentially and it is obvious that none of these procedures would be feasible, if special care was not taken in their implementation in FSDA. Riani et al. (2015) describe the efficient routines for fast updating of the model parameter estimates in forward search and show that the new algorithms enable a reduction of the computation time by more than 80% and allow the running time to increase almost linearly with the sample size. In Riani et al. (2014) are given computational advances, suggesting efficient procedures for calculation of consistency factors of robust S-estimators of location and scale. However, it is still an open issue and further work is necessary to make the monitoring of S-estimation for different consecutive values of bdp efficient for large data sets. In Fig. 1, which shows the computational time for monitoring of forward search, S- and MM-estimation it is visible that the S-estimation is by far the slowest one.

3 Monitoring of compositional data

Key tool for detection of multivariate outliers and for monitoring in the approach presented in the discussed paper are the scaled Mahalanobis distances and statistics related to these distances. However, the results obtained with this tool in case of compositional data might be unrealistic. Compositional data are closed data, i.e. they sum up to a constant value (1 or 100% or any other constant), (see Aitchison 1986). This constraint makes it necessary to first transform the data from the so called simplex sample space to the usual real space. Then standard statistical methods can be applied to the transformed data and the results are back transformed to the original space. One of the most convenient transformation is the family of logratio transformations but it is not clear how the different transformations will affect the Mahalanobis distances used for ranking the data points according to their outlyingness. Filzmoser and Hron (2008) considered three well known transformations and showed how these transformations, namely the additive, the centered and the isometric logratio transformations, will affect the Mahalanobis distances computed by classical and robust methods. They show that

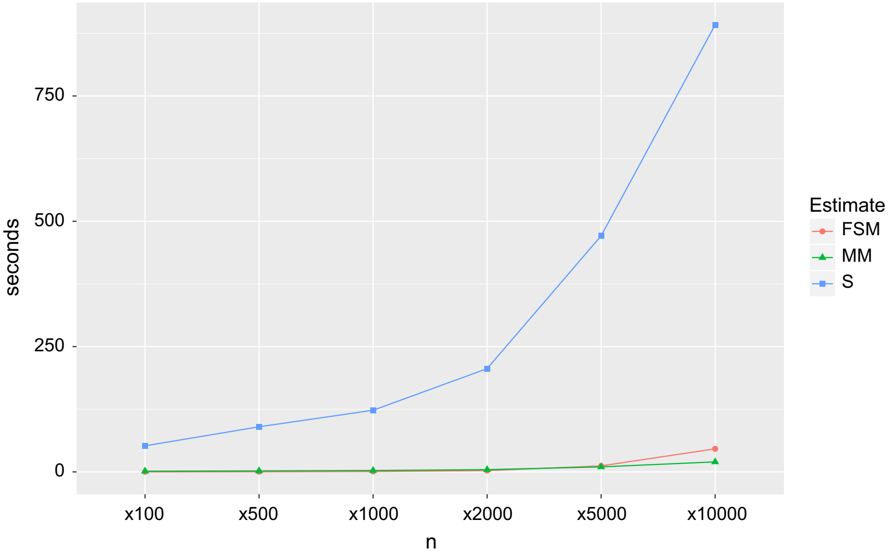

Fig. 1 Comparison of computation time of forward search, S- and MM-estimation monitoring as implemented in the FSDA functions `FSMeda()`, `Smulteda()` and `MMmulteda()` for a data set with `p=10` variables and varying sample size

in case of classical location and covariance estimators all three transformations lead to the same Mahalanobis distances, however, only *alr* and *ilr* extend this property to any affine equivariant estimator.

To illustrate the problem of monitoring compositional data we start with a simple example which was used by Filzmoser and Hron (2008) to introduce outlier detection methods for compositional data. The data *Aphyric Skye Lavas* is from Aitchison (1986, p. 360) and represents percentages of $\text{Na}_2 O + \text{K}_2 O$(A), Fe_2O_3(F) and MgO(M) in 23 aphyric Skye Lavas which sum up to 100%. It is available as data set `skyeLavas` in the R package **robCompositions** (Templ et al. 2011). As pointed out by these authors, we cannot apply standard outlier detection based on Mahalanobis distances, neither classical nor robust, directly to the data set, because, since it is closed, its covariance matrix is singular. Applying the outlier detection methods from the R package **rrcov** as well as the methods from the MATLAB toolbox FSDA described Cerioli et al. result in an error. After applying *ilr* transformation the data will be open and the bivariate structure is revealed as shown in the distance–distance plot in Fig. 2 (robust Mahalanobis distances computed by MCD are plotted against classical Mahalanobis distances). Observations 2 and 3 are identified as potential outliers and observation 1 is a border case (using the 0.975 quantile of the χ^2 distribution as a cut off). Since the (closed) data are three-dimensional they can conveniently be presented in a ternary diagram (right hand panel of Fig. 2). 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).

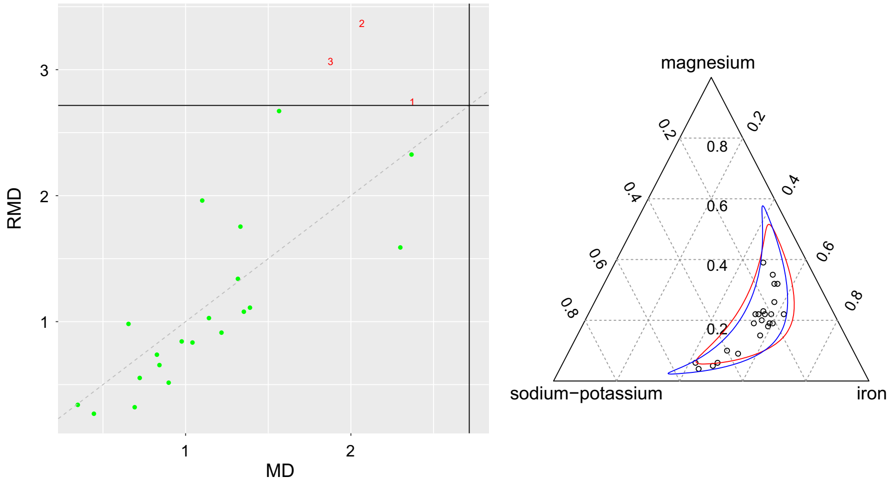

Fig. 2 Aphyric Skye Lavas data set, *ilr* transformed. MCD distance–distance plot in the left hand panel. A ternary diagram with transformed Mahalanobis distances tolerance ellipses, classical and robust (color figure online)

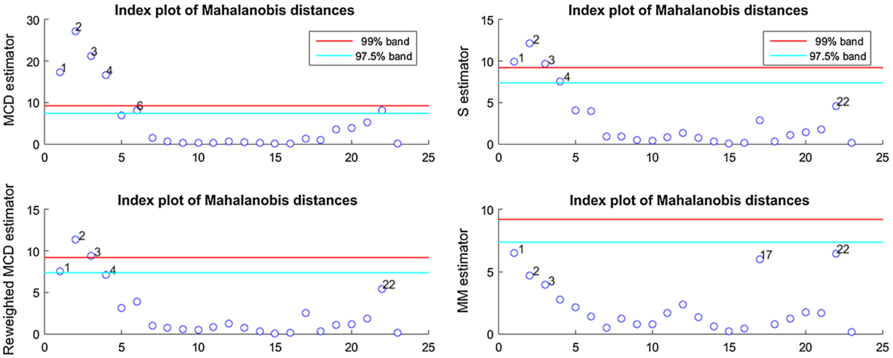

Fig. 3 Aphyric Skye Lavas data set, *ilr* transformed. MCD estimation with $\alpha = 0.5$ and reweighted MCD with cutoff 0.975 in the left hand panel. In the right hand panel—S-estimation with 50% bdp and MM estimation with 99% efficiency

Computing S-estimates with 50% (asymptotic) breakdown point and Tukey's biweight function (Fig. 3, right-upper panel) produces similar results to the reweighted MCD (lower-left panel in the same figure), however the MM estimates with the default efficiency does not identify any outliers (lower-right panel in Fig. 3). As Cerioli et al. 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 in 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. In Fig. 4 is presented the monitoring of the S-estimation. As we have already seen in Fig. 3, for `bdp=50%` the analysis is robust but reducing the breakdown to less

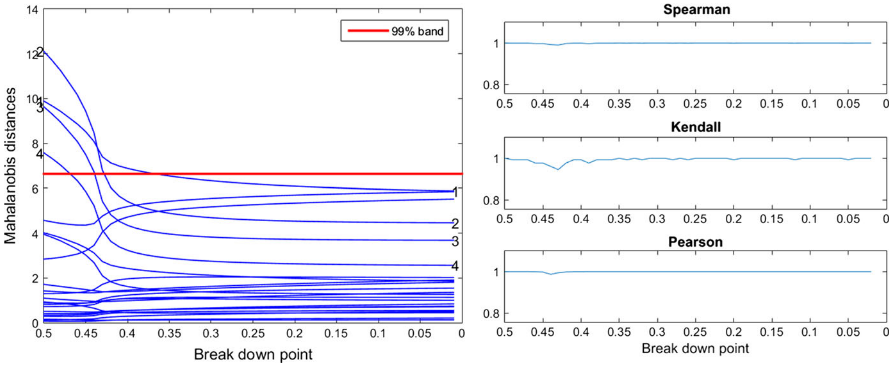

Fig. 4 Aphyric Skye Lavas data set, *ilr* transformed. The left-hand panel shows the squared Mahalanobis distances from monitoring S-estimation and the right-hand panel—the correlation between distances for consecutive values of bdp

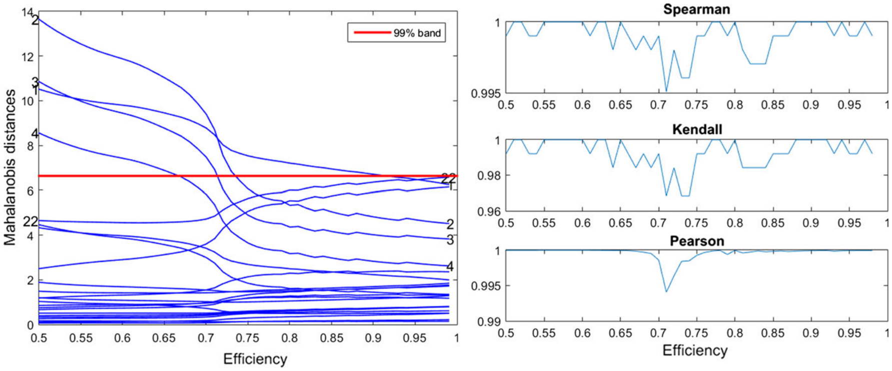

Fig. 5 Aphyric Skye Lavas data set, *ilr* transformed. The left-hand panel shows the squared Mahalanobis distances from monitoring MM-estimation and the right-hand panel—the correlation between distances for consecutive values of eff

than 45% (with the hope to increase the efficiency) results in a non-robust analysis. This is clearly seen in the left hand panel of Fig. 4 but also in the correlation plot on the right side. Monitoring the efficiency of MM-estimates is shown in Fig. 5. It reveals why the index plot of the MM-estimates in Fig. 2 did not show any outliers—for efficiency higher than 0.71 the fit is the same as the maximum likelihood. This is also clearly seen from the correlation monitoring in the right hand panel. Using the brushing functionality of the toolbox, we can identify the outlying units, as shown in Fig. 6: in the right hand panel the outliers are shown as red circles.

Technology intensity of exports The technological structure of manufactured exports as an indicator of their "quality" is an important criteria for understanding the relative position of countries measured by their industrial competitiveness and the determinants of the competitive ability, which are particularly reflected in changes to manufacturing value added and manufactured exports (Todorov and Pedersen 2017). There exists an

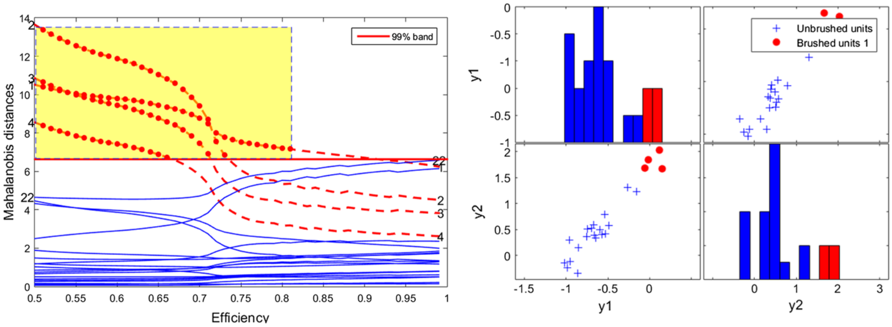

Fig. 6 Aphyric Skye Lavas data set, *ilr* transformed. The left-hand panel shows brushing of the monitoring plot of MM-estimation and the right-hand panel—the scatter plot matrix of the units, identifying the four outliers (color figure online)

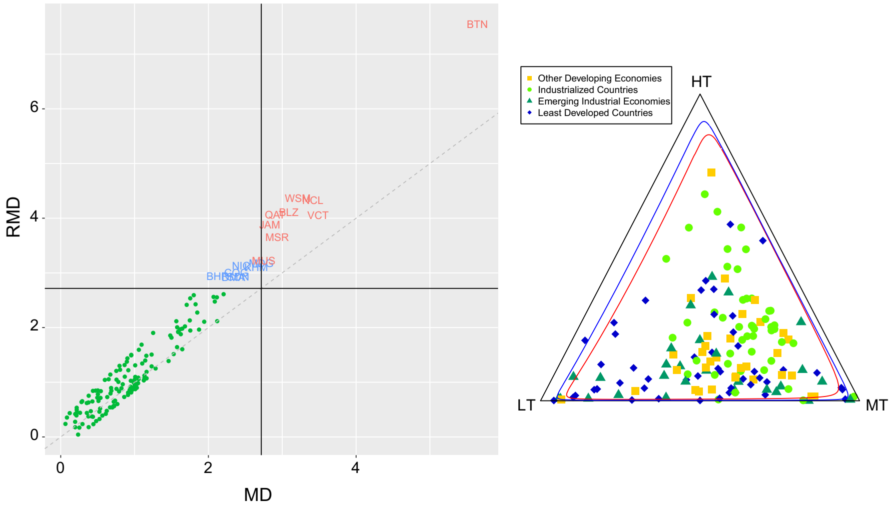

Fig. 7 Technological structure of manufactured exports, *ilr* transformed. MCD distance–distance plot in the left hand panel. A ternary diagram with transformed Mahalanobis distances tolerance ellipses, classical and robust

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). The data set is available in the R package **rrcov3way** (Todorov 2017).

For our example we select only one year, 2012, and remove any countries with missing data, remaining with 153 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 79 outliers out of 153 observations. After applying *ilr* transformation the data will be open and the structure is revealed as shown in the distance–distance plot in Fig 7. Now 22 observations are identified as outliers by the reweighted MCD estimator.

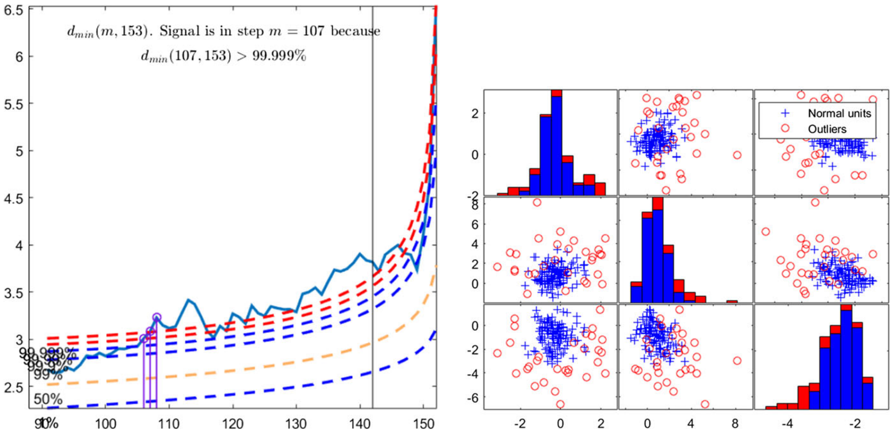

Fig. 8 Technological structure of manufactured exports, *ilr* transformed. The left-hand panel shows the forward search plot of minimum Mahalanobis distance, with a signal for the presence of outliers. The right hand panel shows the scatter plot of the data with the 29 observations identified as outliers by FS as red circles (color figure online)

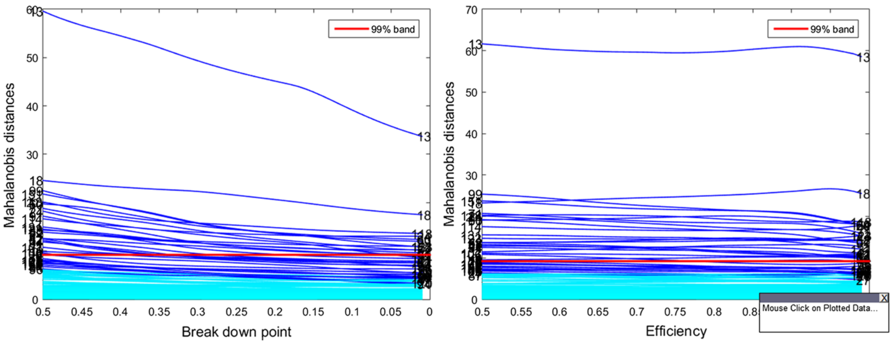

Fig. 9 Technological structure of manufactured exports, *ilr* transformed. The left-hand panel shows the squared Mahalanobis distances from monitoring S-estimation and the right-hand panel—from monitoring the MM-estimation

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.

We continue by running the automatic outlier detection procedure based on forward search. As visible in the left hand panel of Fig. 8 the signal is at observation 107, indicating that it and the succeeding observations might be outliers. Resuperimposition of envelopes leads to the identification of 29 outliers [which turn out to be identical to the outliers detected by the raw (not reweighted) MCD]. Performing the same analysis on the original data (not shown here) indicates a signal at observation 93 and identifies 61 observations as outliers.

Figure 9 shows the monitoring of the Mahalanobis distances of the S- and MM-estimation. The S-estimator with 0.5 bdp is similar to the maximum likelihood. Not much difference is shown in the monitoring plot of the MM-estimation.

References

Aitchison J (1986) The statistical analysis of compositional data. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (Reprinted in 2003 with additional material by The Blackburn Press), London (UK), 416

Atkinson A, Cerioli A, Riani M (2006) **Rfwdmv**: forward search for multivariate data. R package version 0.72-2. https://cran.r-project.org/src/contrib/Archive/Rfwdmv/. Accessed Mar 2018

Filzmoser P, Hron K (2008) Outlier detection for compositional data using robust methods. Math Geosci 40:233–248

Lall S (2000) The technological structure and performance of developing country manufactured exports, 1985–98. Oxford Dev Stud 28(3):337–369

R Core Team (2017) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org/. Accessed Mar 2018

Riani M, Cerioli A, Torti F (2014) On consistency factors and efficiency of robust S-estimators. TEST 23:356–387. https://www.jstatsoft.org/v067/c01

Riani M, Perrotta D, Cerioli A (2015) The forward search for very large datasets. J Stat Softw Code Snippets 67(1):1–20. https://www.jstatsoft.org/v067/c01

Sordini E, Todorov V, Corbellini A (2016) FSDA4R: porting the FSDA toolbox to R. In: Blanco-Fernandez A, Gonzalez-Rodriguez G (eds) International conference of the ERCIM WG on computational and methodological statistics (ERCIM2016). CFE and CMStatistics networks, London

Templ M, Hron K, Filzmoser P (2011) robCompositions: an R-package for robust statistical analysis of compositional data. In: Pawlowsky-Glahn V, Buccianti A (eds) Compositional data analysis: theory and applications. Wiley, New York, pp 341–355

Todorov V (2017) **rrcov3way**: robust methods for multiway data analysis, applicable also for compositional data. R package version 0.1-10. http://CRAN.R-project.org/package=rrcov3way

Todorov V, Pedersen AL (2017) Competitive industrial performance report 2016. Volumes I and II. Report, United Nations Industrial Development Organization (UNIDO), Vienna. http://stat.unido.org. Accessed Mar 2018