Vienna University of Technology

Deliverable 4. Final Report

Contract with the world bank (1157976)

Detecting outliers in household consumption survey data

Authors:

Peter Filzmoser

Johannes Gussenbauer

Matthias Templ

April 26, 2016

Contents

1 Introduction

In most countries, national statistical agencies conduct sample surveys to collect data on household expenditure or consumption. Such surveys may be referred to as household budget surveys, household income and expenditure surveys, welfare monitoring surveys, or other. Using different methods (recall interviews or use of diaries), information is obtained from respondents on their spending on food and non-food goods and services. Many countries will design their survey questionnaires following the international Classification of Individual Consumption by Purpose (COICOP). The information obtained from respondents can be more or less detailed (from a few dozen categories of products and services to thousands of very specific products and services). In countries where own-production of goods represents a significant share of household consumption, the value of self-produced goods consumed by households is also measured. Last, some values may be obtained by imputation (such as use-value of durable goods, or rental value of owner-occupied dwellings).

Consumption or expenditure data are used for multiple purposes. Typical use include the establishment of weighting coefficients for the calculation of consumer price indices, the compilation of national accounts, or the measurement of poverty and inequality.

Producing such data is complex – for the survey statistician, as well as for the respondents. Errors are made at various stages of data production (by the respondents, by the interviewers, and during data capture). These errors include the introduction of "impossible" expenditure values, i.e. values that are too high or too low to be plausible. These outlying values may have a significant impact on some types of analysis (e.g., inequality indicators, or regression coefficients, can be significantly impacted by a few number of extreme values in a dataset). We are thus interested in detecting them, and in fixing the issue by replacing implausible values with more realistic ones. Detecting outliers, and distinguishing those that are "errors" from those that are unusually high (or low) but correct values, is a challenge. Making these corrections in the microdata (i.e. in data at the household level) instead of at the aggregated level (in results tables), adds to the challenge.

There is a rich literature devoted to the issue. But no common "best practice" has emerged from it. Different agencies adopt different solutions, some of them too radical – at the risk of creating bias and corrupting their data files.

To assess the relevance and impact of various outliers detection and imputation methods, the World Bank and the International Household Survey Network commissioned a comprehensive review of existing algorithms, including an assessment of their implementation on actual survey datasets. This report presents the main findings of this work.

This study was sponsored by a Trust Fund from the Department for International Development of the United Kingdom, administered by the World Bank Development Data Group (TF011722).

Outline:

In Section 2 some problems are mentioned. This covers missing information in the surveys which needs to be assumed as zero value, and structural zeros. Also problems with sparse components are mentioned, the concrete strategies to aggregate components are given in 6 and 6.1.4. In addition, a discussion about the types of outliers is included.

Section 3 gives a detailed description of univariate outlier detection while Section 4 discusses multivariate outlier detection methods. Section 5 describes how the outliers are imputed/replaced. In the numerical study (Section 6), the results on consumption data are provided. There is detailed analysis for the data from the Albania Living Standards Survey 2008 included and tables that summarizes all results for the Gini for each country. A simulation study gives further insights about the quality of the methods.

All code for the methods described in this final report is provided to the World Bank in form of the unpublished R package **robout**. Next to the code for the methods, it includes examples of applications of the package on simulated and real data.

2 Some basics and challenges

In the literature the term "outlier" is not defined uniformly and many different definitions can be found. In general, it can be said that an outlier is a data point which deviates from the data structure formed by the data majority.

2.1 Robust estimation or outlier detection

Outlier detection and robust estimation are closely related (see Hampel et al. 1986; Hubert et al. 2008) and robust estimation (find an estimate which is not influenced by the presence of outliers in the sample) and outlier detection (find all outliers, which could distort the estimate) are similar tasks. A solution of the first problem allows us to identify the outliers using their robust residuals or distances while on the other hand, if we know the outliers we could remove or downweight them and then use the classical estimation methods [Hulliger, 2007, Todorov et al., 2011].

In many research areas the first approach is the preferred one but for the purposes of official statistics the second one is more appropriate. This is especially true if the organizations deliver micro-data for general and public use, but also for use in different departments in the same organization. Hereby, it is hard to imagine that a non-expert user of this data set will employ the same sophisticated robust techniques that the statistician has applied to those parts of the data set containing outliers [Hulliger, 2007]. Therefore the aim is to deliver an outlier-free data set, with outlier values appropriately modified, such that the data set is suitable for general use with standard statistical software and classical estimation methods. This can be achieved by using an outlier imputation procedure.

The focus in this work is on using robust methods to identify the outliers and to impute them so that the data can be treated in the traditional way afterwards.

2.2 Sampling weights

One of the unique features of the outlier detection problem in the analysis of survey data is the presence of sampling weights. For example, in a survey the sample design might be defined in such a way that households are sampled proportional to size in each region and all persons in a household are included in the survey. The sampling weights are (calibrated) inverses basically coming from the inclusion probabilities of units in the sampling frame considering also non-responses and, possibly, calibration on known population characteristics. These sampling weights are used in the estimation procedure and therefore may not be left unaccounted for in the phase of data cleaning and outlier detection [Todorov et al., 2011].

2.3 Imputation of outliers

Imputation methods have traditionally been used for item non-responses. The basic idea in this case is that by "filling in" the missing values in a data set, standard methods of inference are applicable.

Imputations are necessary after outlier identification, i.e. once the outliers in the survey data have been identified and classified, the outliers may be imputed. In any case, non-representative outliers are very similar in concept to missing data since from both these values are wrong and should be changed to plausible values [Hulliger, 2007, Hulliger et al., 2011]. These values can be derived from the non-outliers in the survey data set.

2.4 Structural zeros

In consumption surveys, zero values for the consumption of particular items are frequent as one cannot expect all households to consume all possible items. For example, considering components on consumptions, a person will only have consumptions in transport services if a transport service was used in the reference year.

For particular components, the amount of zeros can be quite high and multivariate outlier detection algorithms may fail if the number of zeros is above a certain threshold. Observations could even become outliers because of zeros when multivariate detection methods are applied. Assume two-dimensional bivariate normal data. If one observation has a zero in the first variable, the observation may easily become a multivariate outlier. An example for this phenomenon is illustrated in Figure 1 where one value is replaced by a zero.

There are multiple strategies to deal with zeros:

- apply only univariate outlier detection on the observed values (in Section 3);
- impute zeros and apply multivariate outlier detection (in Section 4).

Figure 1 shows the problem of structural zeros on a simple two-dimensional toy data example. The observation $\mathbf{x}_1$ is in the center of the data cloud and is surely not an outlier. However, when assuming that $\mathbf{x}_1$ includes a zero, i.e. we put a zero in $\mathbf{x}_1$ to get $\mathbf{x}_1^*$, then the observation $\mathbf{x}_1^*$ will be flagged as an outlier if multivariate outlier detection methods are applied. To adequately

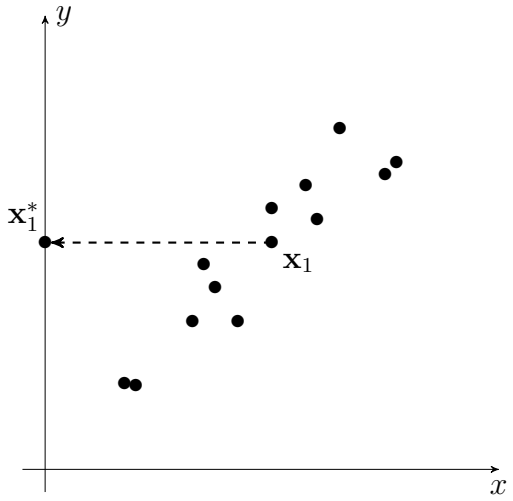


Figure 1: Two-dimensional toy example where for one observation $\mathbf{x}_1$, its value on the x-axis is replaced by a zero ($\mathbf{x}_1^*$). The observation immediately becomes a multivariate outlier only because of the zero.

deal with zeros is therefore essential for any outlier detection method. For example, a zero on expenditures for food would be unrealistic, but it could be true for expenditures on transportation.

2.5 Aggregation of sparse components

Generally, before applying multivariate outlier detection algorithms on very sparse components, one has to think of reducing the dimensions by aggregation of components, while holding the loss of information minimal. The aggregation of components is further investigated in this section.

2.6 Deterministic methods and selective editing

Techniques such as selective editing (see, e.g., Lawrence and McKenzie 2000), automatic editing (De Waal 2003; Fellegi and Holt 1976) and macro-editing [see, e.g., Granquist, 1990] can be applied instead of the traditional micro-editing approach, where all records are extensively edited manually [e.g., De Waal, 2009].

2.6.1 Automated deterministic editing

Automatic editing is done by formulating a framework for both, the automatic construction of implicit edits and a linear programming algorithm to edit the data in an optimized manner automatically [Fellegi and Holt, 1976] using heuristic approaches for large data sets. NSO's often use these kind of methods and correct the collected data in detail. This process can be error-prone, because editing rules determined by subject matter specialists may destroy the multivariate structure of the data and in case of survey data, sampling errors may cause that correct data entries are marked as incorrect by deterministic rules [see also Chambers et al., 2004, De Waal, 2009, Todorov et al., 2011] . Establishing editing rules is highly time-consuming and resource intensive and often represents a significant part of the costs of the whole survey [Chambers et al., 2004, Todorov et al., 2011].

2.6.2 Selective editing

Selective editing is closely related to the concept of the influence function in robust statistics (in both concepts, the effect of each observation on an estimator is measured), i.e. selective editing detects those statistical units which highly influence one (or a few) pre-defined measure(s). However, usually many different estimates are obtained from the same data set and those outliers which do only have high influence on the measures as used in the selective editing process are not detected.

Due to the mentioned disadvantages of automatic editing procedures, we will focus on non-deterministic methods.

2.6.3 Errors resulting from measurement units

In many surveys, some values are calculated by multiplying measures of quantities by unit values. The quantities are reported in many different units, some non-standard. Reporting and data capture errors in the quantities may cause outliers (e.g., using the unit code for kilos when the quantity is actually reported in grams would inflate the true value by 1000).

Algorithms are available to detect and repair observations that violate linear equality constraints by correcting simple typo's. However, a set of deterministic edits has to be defined first. Such edits are usually country-specific and data-related. Thus, such algorithms are not ideally suited for

detecting errors in standardized data sets from many countries since too many manual work must be invested to define the edits.

2.7 A brief description of the various types of outliers

Virtually any data set in survey statistics contains outlying values. This is especially true for consumption data. Dupriez [2007] claims that outliers in food consumption often seem to come from errors in quantity measurement units (many values are around 1,000 or 100 times the mean of valid values, indicating that grams (or ml) and kilos (or liters) may have been mixed, or that decimal points may have been missed). The detection of this kind of measurement errors is usually rather simple compared to other kinds of outliers. Since the distribution of many expenditure variables is skewed to the right, outliers are more obvious in the very right tail of the distribution rather than in the lower tail. Outlier detection methods which first symmetrize the distributions may thus be worthwhile to being considered. The outlying procedures are applied on variables depending if a variable is measured on individual (person) level or household level. For example, for some items (e.g., food, transport services, personal effects), the outliers are detected using per capita values.

2.7.1 Univariate versus multivariate outliers

Univariate outlier detection usually leads to quite different results than multivariate outlier detection. This will be illustrated in the following by the so-called bushfire data set with 38 observations in 5 variables (Campbell, 1989). The data set contains satellite measurements on five frequency bands, corresponding to each of 38 pixels and it is used to locate bushfire scars. This data set is very well studied (Maronna and Yohai, 1995; Maronna and Zamar, 2002) and this is the reason why it is considered here for illustration purposes. It is known that it contains 12 clear outliers, the observations 33-38, 32, 7-11; 12 and 13 are suspect. The data set is available in the `R` package **robustbase**. Figure 2a shows boxplots of all variables in this data set. Only three observations are detected as outliers using the boxplot approach (see Section 3.1). Similar results are obtained using any of the other univariate outlier detection methods. In Figure 2b the raw data from the second and third variable are shown. Figure 2c shows the region (within the blue rectangle) of outlier-free observations. Everything outside the rectangular region is

defined as outlier according to a univariate outlier detection rule. The result of a univariate outlier detection is always a rectangular region, not considering the multivariate structure of the data. When applying multivariate outlier detection (considering all variables) using classical estimates of the covariance matrix, we see that not even one observation is identified as an outlier in Figure 2d. However, if robust multivariate methods are used, the outliers are detected (see Figure 2e). This clearly indicates advantages of robust multivariate outlier detection methods.

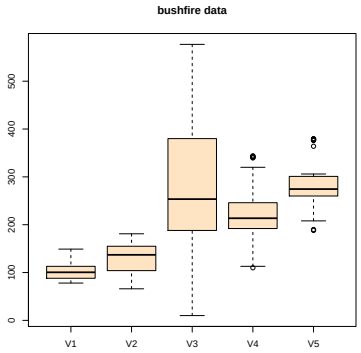
Consumption/expenditure data are multivariate, thus multivariate methods should be applied for detection of outliers.

2.7.2 Bottom outliers

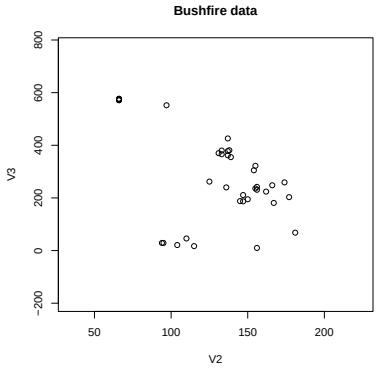
Bottom outliers are considered by many outlier detection methods as soon as the data are transformed to a symmetric distribution, e.g. by log-transformation. These methods are based on thresholds, and these thresholds are defined by lower and upper quantiles of the distribution of a single variable (univariate methods) or by distances to a multivariate center (multivariate methods). For univariate methods usually only outliers in the upper tail of the distribution are imputed, since low values have no serious effect on classical indicators such as the Gini coefficient. However, it depends on the underlying question: For other uses of consumption data, such as measurement of nutrition and food vulnerability, the bottom outliers are important. For multivariate outlier detection, observations with large distance to the multivariate data center with respect to the covariance structure are imputed, i.e. bottom and top outliers are treated in the same manner. More on imputation of multivariate outliers is given in Section 5.

2.7.3 Valid / invalid outliers

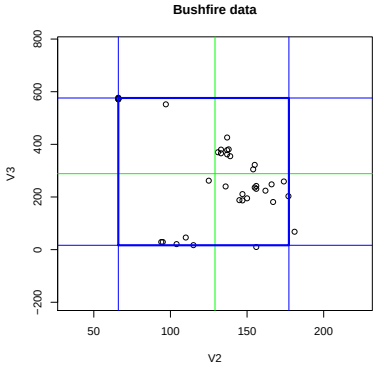
Outliers may be representative (extreme but true values) or non-representative (measurement errors) (Chambers 1986). An outlier detection procedure typically flags not only measurement errors but also correct entries, which do have an abnormal behavior compared to the main bulk of the data. In other words, outliers can be errors or true (extreme) values, but both can have a large influence on the estimates and it is reasonable to detect them. In practice it is not always possible to distinguish if an outlying value is the result of measurement errors or a genuine, but extreme observation. Therefore,



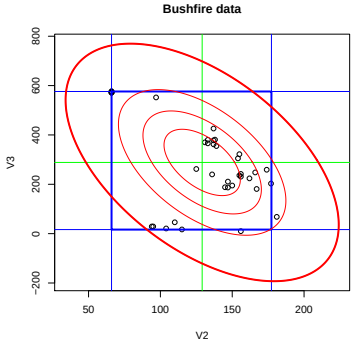
(a) Univariate outlier detection.



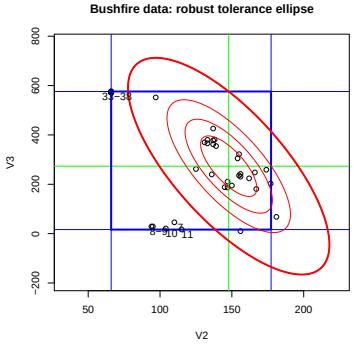
(b) The second and third variable of the bushfire data set.



(c) Univariate outlier detection.



(d) Multivariate outlier detection based on classical estimates.



(e) Multivariate outlier detection based on robust estimation.

Figure 2: Univariate and multivariate outlier detection on the bushfire data to show advantages of multivariate outlier detection methods.

methods for outlier detection can not truly detect outliers, but detect data points which have the potential of being an outlier.

It is the characteristic of a representative outlier that additional similar elements can be found in the population. The outlier is therefore "representative" for these elements. When carrying out estimations including representative outliers, the influence of the representative outliers can be reduced, for example, by reducing the sample weight of the observation.. Reducing the weights can be seen from two aspects. First it can be assumed that fewer outliers are in the population as the sampling weights suggested. Then it is natural to decrease the sampling weight of outliers. The second aspect covers the reduction of the influence of representative outliers to the variance of an indicator. On the one hand, by reducing the influence of representative outliers, bias might be introduced. On the other hand, the influence of representative outliers on the variance can be reduced. An optimal method might therefore be defined by having a minimal square error criterion (minimizing (variance + bias^2)).

3 Univariate outlier detection methods

In terms of considering single variables, potential outliers are solely those points which are "far enough" away from the main bulk of the data. In order to locate these points, one way is to measure scale and location of a data sample in a robust way. For example, all observations which are outside the range of location plus/minus multiple times the scale can be considered as potential outliers.

For this section we will suppose a sample $\mathbf{s}$ (with fixed sample size $n \leq N$) which has been drawn, according to the sampling design $p(S)$, from the finite population $U = \{1, \ldots, N\}$. Furthermore, to each sampled element in $\mathbf{s}$ a weight w_i is attached that reflects the sample inclusion probability π_i. In addition we have that $w_i = \frac{1}{\pi_i}$, where $\pi_i = \sum_{i \in \mathbf{s}, \mathbf{s} \in S} p(\mathbf{s})$, $i = 1, \ldots, N$, and we assume that $\sum_{i \in \mathbf{s}} w_i = N$.

3.1 Robust location $\pm$ constant $*$ robust scale

As mentioned above, a common practice for univariate outlier detection is to take robust measures of location and scale to check which points deviate

from the rest of the data.

Robust location

A very common, but nor robust, measure for location is the arithmetic mean $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Since the arbitrary change of only one single observation can change the mean $\overline{x}$ to any position, the arithmetic mean has a breakdown point of 0% (it is not robust to any small amount of contamination). A more suitable, robust measure of location is the **median** (med). For a random variable X with distribution F, the median is defined by

$$\mathbb{P}(X \geq med) = F(med) = 0.5.$$

For a data sample $(x_1, \ldots, x_n)$ the median is determined by

$$med = \begin{cases} x_{(\lfloor n/2 \rfloor + 1)} & \ldots n \text{ odd} \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) & \ldots n \text{ even,} \end{cases}$$

where $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ denotes the ordered sample, and $\lfloor a \rfloor$ means truncating a to the largest integer $\leq a$. The median achieves the maximum breakdown point of 50%, i.e. up to 50% of any observations in the data set could be modified without causing an arbitrary change of the estimator. Taking into account sample weights $w_i, i = 1, \ldots, n$, the pth quantile, $\hat{Q}_p$, is given by solving the estimating equation

$$\sum_{i \in s} w_i [\mathbb{1}\left\{x_i \leq \hat{Q}_p\right\} - p] = 0. \tag{1}$$

The weighted sample median is obtained by $p = 0.5$ and has a breakdown point of 0%. This is demonstrated by the following example. It is easy to see that the sample median in Table 1 equals 5, while the weighted median is influenced so heavily by the last weight, that it equals 11. Therefore the weighted median needs only be influenced by one out of n data points at any arbitrary position to break down which results in a breakdown point of 0%.

However, in practical applications with comparable large sampling weights, the weighted median turns out to be quite robust. This is especially true if the weights do not depend on the size of the data values.

data	weights
1	3
3	2
4	1
5	2
7	4
10	3
11	25

Table 1: Data set showing that the weighted median has a breakdown point of 0%.

Robust scale

The most common estimator of scale of a sample is the empirical standard deviation, defined as $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$. Since this estimator is not robust, it is a poor choice in the context of outlier detection. A more robust way to estimate the scale is by using the sample quartiles. Namely, the first and third quartiles (Q_1 and Q_3) which are defined, for a random variable X with distribution F, by

$$\mathbb{P}(X \leq Q_1) = F(Q_1) = 0.25$$
$$\mathbb{P}(X \leq Q_3) = F(Q_3) = 0.75.$$

To attain a robust measure of scale, the so called interquartile range (IQR) can be used, which is defined by the difference of the third and the first quartile

$$IQR = Q_3 - Q_1.$$

The IQR has a breakdown point of 25% and is therefore more suited in case of outliers than the empirical standard deviation.

Considering sample weights $w_i, i = 1, \ldots, n$, the first and third weighted quartiles are determined by solving Equation (1) for $p = 0.25$ and $p = 0.75$. The weighted interquartile range has, like the weighted sample median, a breakdown point of 0%. However, the more similar the sampling weights, the more robust becomes the estimator.

Another way to robustly measure the scale is the so called median absolute deviation (MAD). It is defined for a random variable X as

$$\mathbb{P}(|X - med| \leq MAD) = 0.5.$$

For a data sample $(x_1, \ldots, x_n)$ the MAD is determined by

$$MAD = \underset{i}{\text{med}}\, |x_i - \underset{j}{\text{med}}(x_j)|.$$

The MAD achieves the maximal breakdown point of 50% .

To calculate MAD with sample weights $w_i, i = 1, \ldots, n$, one has to solve Equation (1) for $p = 0.5$ to obtain the weighted median $\hat{Q}_{0.5}$. Using Equation (1) once again on $|x_i - \hat{Q}_{0.5}|$, the estimation equation for the weighted MAD (MAD_w) is presented by

$$\sum_{i \in s} w_i [\mathbb{1} \left\{ |x_i - \hat{Q}_{0.5}| \leq MAD_w \right\} - 0.5] = 0.$$

Both, IQR and MAD are not consistent estimators of the scale parameter σ. Assuming that the data come from a normal distribution, consistent estimators are

$$S_{IQR} = \frac{IQR}{1.35}$$
$$S_{MAD} = \frac{MAD}{0.675}.$$

Detecting potential outliers

To detect potential outliers it is quite common to use the median plus/minus a constant c, $c \in \mathbb{R}$, times a robust measure of scale to determine a range at which the data points are not considered outliers. Points that lie outside this interval are potential outliers. In our context this results in the following intervals

$$[med \ - \ c \cdot S_{IQR} \ , med \ + \ c \cdot S_{IQR}] \tag{2}$$
$$[med \ - \ c \cdot S_{MAD} \ , med \ + \ c \cdot S_{MAD}]. \tag{3}$$

Depending on the type of problem the choice of c differs. In many cases c will be chosen equal to 3, since, considering that the data sample is normally distributed, the interval $[\mu \pm 3\sigma]$, with μ the expectation and σ the standard

deviation, overlaps just over 99% of the possible realizations.

[Dupriez, 2007], for example, used 5 times IQR which is relatively conservative (high value of the constant), but the impact of fixing these outliers on the total consumption and on its distribution is significant.

Figure 3 illustrates the use of the boundaries from Equation (2) and (3), with $c = 3$, for outlier detection on a simulated data set with sample weights equal to 1 (upper plot) and with sample weights that increase in accordance to the value of the data points (lower plot). The five lowest and five highest points of the simulated data have been drawn from a distribution smaller respectively larger mean than the distribution of the main body of the data. In the upper part of Figure 3 one can see that with the use of IQR and MAD for outlier detection the four lowest and five highest points would be declared as potential outliers. Therefore, almost every data points which was generated by a different mechanism than the main bulk of the data was declared a potential outlier. The lower part of Figure 3 shows the influence of the sample weights on the estimates. The boundaries for weighted IQR and weighted MAD as well as the weighted median have been shifted slightly to the right due to the weights. This is due to the design of the sample weights, since the sample weights positively depend on the value of the data points.

Box-Cox transformation

For very skewed data the above mentioned methods for outlier detection could prove to be problematic, since the interval in which data points are not considered to be outliers is symmetric around the median. To adjust for this issue, Box and Cox [1964] proposed a transformation in order to transform the data to behave like it was generated from a normal distribution.

The Box-Cox transformation is a parametric family of transformations from x_i to $x_i^{(\lambda)}$, $i = 1, \ldots, n$, with the parameter λ defining a particular transformation. For data $x_i > 0$, the Box-Cox transformation is defined as

$$x_i^{(\lambda)} = \begin{cases} \frac{x_i^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \log(x_i) & \lambda = 0. \end{cases}$$

To determine the appropriate value of λ for a given data sample, Box and Cox [1964] presented a maximum likelihood as well as a Bayesian approach. Figure 4 shows the impact of the Box-Cox transformation on outlier detection methods on a simulated data, following a standard log-normal distribution;

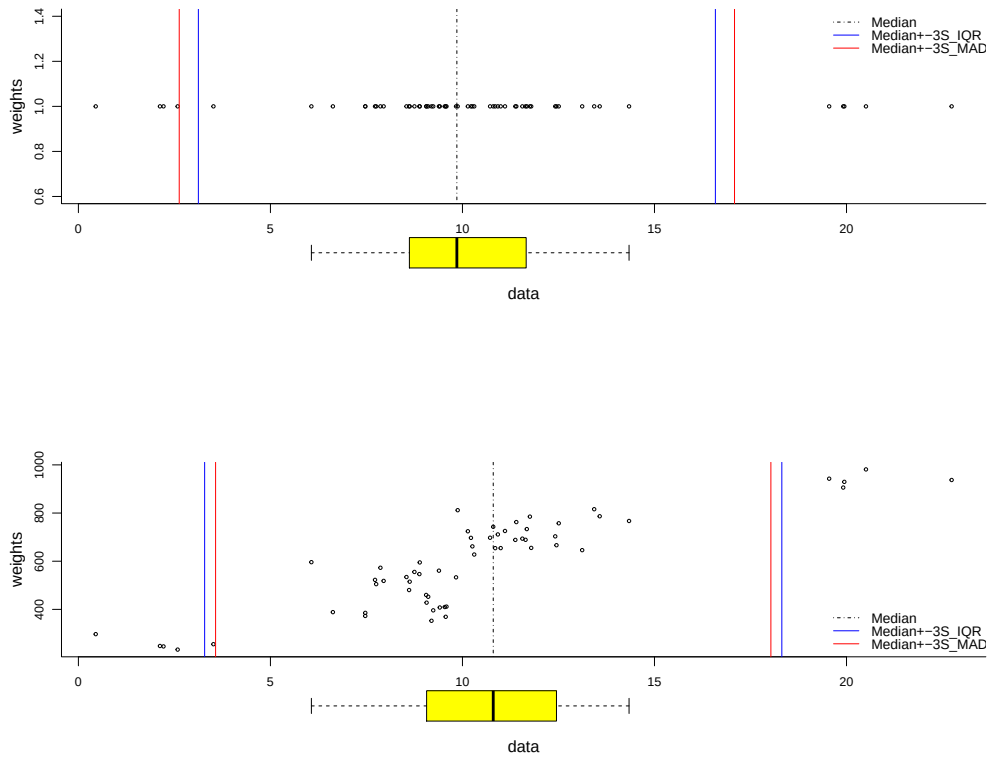


Figure 3: Outlier detection using IQR, MAD and a boxplot for a data sample with sample weights equal to 1 (upper plot) and sample weights, which positively depend on the value of the data points (lower plot).

the sample has no sample weights. The upper part of the graphic shows the use of the above mentioned outlier detection methods, using IQR and MAD, without the Box-Cox transformation. The lower part shows the back-transformed bounds of the intervals given in Equation (2) and (3) after the outlier detection has been applied on the Box-Cox transformed data.

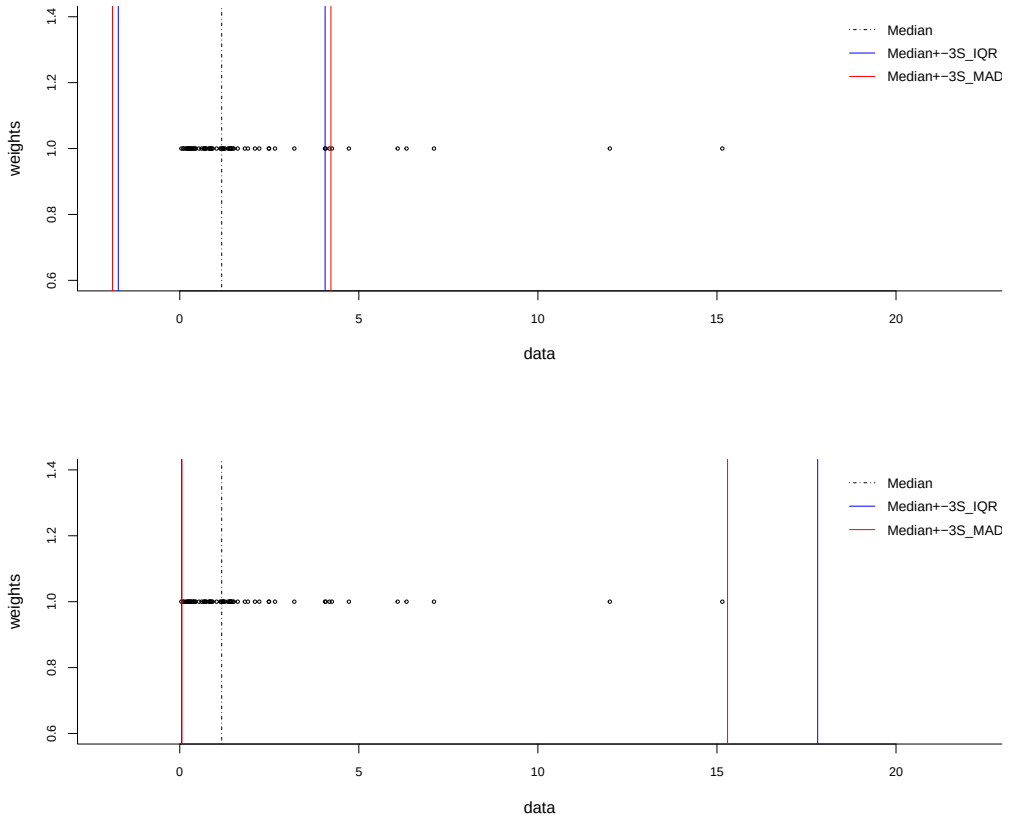


Figure 4: Outlier detection using IQR and MAD without Box-Cox transformation (upper plot) and back-transformed bounds after outlier detection was applied on the Box-Cox- transformed data (lower plot).

Calculating the parameter λ for the Box-Cox transformation using a maximum likelihood estimate does not account for extreme data values. The maximum likelihood approach is therefore not robust and outlier detection schemes could fail to detect potential outliers on the data sample after the

transformation. To address this problem one can calculate λ in a robust way by using robust linear regression. Consider a sample of data values $x_i, i = 1, \ldots, n$, $x_i > 0$, which are already sorted in ascending order, and the linear model

$$x_i^{(\lambda_0)} = \alpha + \beta z_i + u_i,$$

with α, β and λ_0 as real parameters, and z_i as the $\frac{i}{n}$th quantile of the standard normal distribution. Furthermore, the errors u_i are considered i.i.d., independent of z_i and $\mathbb{E}[u_i] = 0$. To calculate the true Box-Cox parameter λ_0 one can apply MM-estimation to the responses $y_i^{(\lambda)}$ for given λ and choose λ_0 such that the robust residual autocorrelation $\rho_n(\lambda)$ is minimized.

Let $s_n(\lambda)$ be a robust measure of the residual scale and $r_i(\lambda) = x_i^{(\lambda_0)} - \alpha(\lambda_0) + \beta(\lambda_0) z_i$ for given λ, then the robust residual autocorrelation is defined by

$$\rho_n(\lambda) \ = \frac{1}{n} \sum_{j=1}^{n-1} v_j(\lambda) v_{j+1}(\lambda)$$

with

$$v_j(\lambda) \ = \psi \left(\frac{r_j(\lambda)}{s_n(\lambda)} \right)$$

where $\psi(.)$ is the Huber's function. Note that the v_j's are always positive. For more details on robust Box-Cox transformation for linear regression, see Maronna et al. [2006].

3.2 Boxplot

The boxplot is a widely used graphical tool to visualize the distribution of continuous univariate data. Fundamental to the boxplot is the five-number summary which contains the sample minimum, the first quartile (Q_1), the median, the third quartile (Q_3) and the maximum. The *box* ranges from the first to the third quartile containing per definition 50% of the innermost data as well as the median which is usually marked by a middle line. To detect if a point has the potential of being an outlier it depends on whether the point

is within the range of the so called *fences* (lower fence LF, upper fence UF). The lower and upper fence are defined as follows:

$$LF = Q_1 - 1.5 \underbrace{(Q_3 - Q_1)}_{:=IQR} \qquad UF = Q_3 + 1.5\ IQR. \tag{4}$$

In addition to the fences the boxplot also contains the so called *whiskers* which are lines that range from the box to the point which is the farthest from the box but still inside the fences.
It is possible to incorporate sample weights into the boxplot by using Equation (1) for weighted sample median and weighted sample quartiles.

3.3 Adjusted Boxplot

By using the position of the median within the box, the length of the box and the whiskers, the boxplot gives information about the location, spread, skewness and tails of the data. However, for very skewed data the boxplot can possibly fail to properly mark potential outliers since the rule for outlier classification is solely based on location and scale measures and the fences are derived from the normal distribution. Therefore a boxplot will classify too many points as outliers if the data are sampled from a skewed distribution. To adjust the boxplot for skewed data it is possible to incorporate a robust measure of skewness into the calculations for the fences, which leads to the adjusted boxplot.

Robust measure of skewness (medcouple)
A robust measure of skewness of a continuous distribution F is the so called medcouple (MC). It is defined as

$$MC(F) = \underset{x_1 < m_F < x_2}{\text{med}} h(x_1, x_2)$$

with m_F as the median of F, x_1 and x_2 sampled independently from the data and h as the kernel function given by

$$h(x_i, x_j) = \frac{(x_j - m_F) - (m_F - x_i)}{x_j - x_i}.$$

By definition the medcouple always lies between -1 and 1 and takes on positive values for right-skewed data and negative values for left-skewed data.

Brys et al. [2004] showed that this robust measure of skewness has a bounded influence function and a breakdown point of 25%.

To adjust a boxplot for skewed data one can incorporate the medcouple in the calculation of the fences. This can be done by using functions h_l and h_r, defined in the following, to determine the fences. Thus instead of using the interval of regular observation as

$$[Q_1 - 1.5\ IQR\ ;\ Q_3 + 1.5\ IQR]$$

one can choose the boundaries of the interval to be defined as

$$[Q_1 - h_l(MC)\ IQR\ ;\ Q_3 + h_r(MC)\ IQR].$$

The functions h_l and h_r are independent from each other allowing for different lengths of whiskers. In addition one requires that $h_l(0) = h_r(0) = 1.5$ to obtain the original boxplot for symmetric data.
Vandervieren and Hubert [2008] studied three different models for the choice of the functions h_l and h_r, a linear model, a quadratic model and an exponential model. They come to the conclusion that the exponential model performed best and proposed for the interval of the fences and the functions h_l and h_r

$$[Q_1 - 1.5e^{-3.5MC}\ IQR\ ;\ Q_3 + 1.5e^{4MC}\ IQR].$$

Dealing with a data sample, including sample weights, one can use Equation (1) to calculate weighted quartiles and a weighted IQR. Sample weights will not be incorporated into the medcouple since the factor $1.5e^{-3.5MC}$ and $1.5e^{4MC}$ has performed well by the simulations by Vandervieren and Hubert [2008].

3.4 Pareto tail modeling

Data on household expenditures are typically skewed to the right which implies that potential outliers are more likely detected in the upper tail of the data distribution. As a consequence, the upper tail of the data distribution may be modeled with a Pareto distribution (Pareto tail modeling), in order to re-calibrate the sample weights or fitting data values for observations in the upper tail.

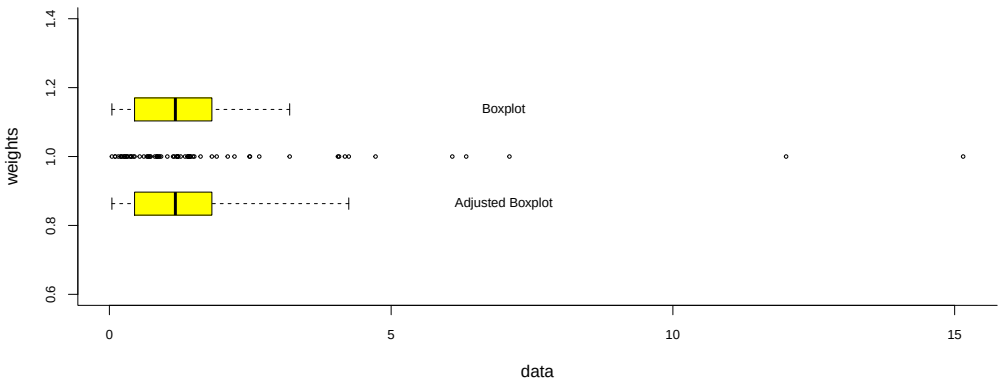


Figure 5: Boxplot vs. Adjusted Boxplot

The Pareto distribution has been well studied in the literature, see also Kleiber and Kotz [2003], and is defined as

$$F_\theta(x) = 1 - (\frac{x}{x_0})^{-\theta}, \quad x \geq x_0$$

with $x_0 > 0$ as scale parameter and $\theta > 0$ as shape parameter. In Pareto tail modeling, the cumulative distribution function of the whole data sample is modeled as

$$F(x) = \begin{cases} G(x) & \text{if } x \leq x_0 \\ G(x_0) + (1 - G(x_0))F_\theta(x) & \text{else,} \end{cases} \tag{5}$$

where G is an unknown distribution function (Dupuis and Victoria-Feser [2006]).

Values larger than the threshold x_0 (i.e. potential outliers) will then be replaced by a value corresponding to the fitted distribution. There are several methods to estimate the threshold x_0 and the shape parameter θ. Here we will focus on the Van Kerm's rule of thumb (Van Kerm [2007]) to model x_0 and the partial density component estimator to model θ.

Van Kerm's rule of thumb

The Van Kerm's rule of thumb represents a suggestion, based on the EU-

SILC data, for the threshold x_0. It is given by

$$\hat{x}_0 := min(max(2.5\overline{x}_w, Q_{0.98}), Q_{0.97})$$

with $\overline{x}_w$ as the weighted mean and $Q_{0.98}$ and $Q_{0.97}$ as weighted quantiles defined by Equation (1).

Partial density component estimator

The partial density component estimator is an extension of the integrated squared error (ISE) estimator where the Pareto distribution, for a data sample $x_1, \ldots, x_n$, is modeled in terms of the relative excess

$$y_i := \frac{x_{(n-k+i)}}{x_{(n-k)}}, \quad i = 1, \ldots, k.$$

The integrated squared error estimator is then given by minimizing the integrated squared error criterion (Terrell [1990])

$$\hat{\theta} = arg\min_{\theta} \left[\int f_\theta^2(y)dy - 2\mathbb{E}(f_\theta(Y)) \right]$$

with $f_\theta(y)$ as the approximation of the density function of the Pareto distribution given by

$$f_\theta(y) = \theta y^{-(1+\theta)}.$$

The partial density component (PDC) estimator minimizes the integrated squared error criterion using an incomplete density mixture model uf_θ. The PDC estimator is thus given by

$$\hat{\theta}_{PDC} = arg\min_{\theta} \left[u^2 \int f_\theta^2(y)dy - \frac{2u}{k} \sum_{i=1}^{k} f_\theta(y_i) \right].$$

The parameter u can be estimated by

$$\hat{u} = \frac{\frac{1}{k}\sum_{i=1}^{k} f_{\hat{\theta}}(y_i)}{\int f_{\hat{\theta}}^2(y_i)dy}$$

and can be interpreted as a measure of the uncontaminated part of the sample.

Alfons et al. [2013] proposed for the weighted partial density component estimator with sample weights $w_1, \ldots, w_n$

$$\hat{\theta}_{PDC,w} = arg\min_{\theta} \left[u^2 \int f_\theta^2(y)dy - \frac{2u}{\sum_{i=1}^{k} w_{n-k+i}} \sum_{i=1}^{k} w_{n-k+i} f_\theta(y_i) \right],$$

$$\hat{u} = \frac{\frac{1}{\sum_{i=1}^{k} w_{n-k+i}} \sum_{i=1}^{k} w_{n-k+i} f_{\hat{\theta}}(y_i)}{\int f_{\hat{\theta}}^2(y_i)dy}.$$

For more detailed information on ISE and PDC, see also Vandewalle et al. [2007] and Alfons et al. [2013].

4 Multivariate outlier detection methods

The different variables of the consumption data can be considered jointly. Thus, if p variables are available, each observation consists of p values with the corresponding consumption data. An outlier would then be an observation where the joint multivariate information is different from the multivariate data distribution of the data majority.

Most multivariate outlier detection methods are based on estimates of multivariate location and covariance. There are various possibilities to obtain robust estimates of these quantities, and they differ in their statistical properties. These estimates are then used to determine the outlyingness, e.g. in terms of a distance measure, like the Mahalanobis distance.

4.1 Mahalanobis distances

Consider p-dimensional observations $\boldsymbol{x}_i$ (column vector), for $i = 1, \ldots, n$, which are collected in the rows of the data matrix $\boldsymbol{X}$. The traditional estimators of multivariate location and scatter are the sample mean $\bar{\boldsymbol{x}}$ and the sample covariance matrix $\mathbf{S}$ given by

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$$

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^t. \tag{6}$$

These estimates have very good properties if the data come from a multivariate normal distribution. However, they can be very misleading in presence of outlying observations. Estimates of location and covariance are needed for outlier detection methods which are based on the Mahalanobis distance. The squared Mahalanobis distance MD_i^2 for an observation $\boldsymbol{x}_i$, for $i = 1, \ldots, n$, is defined as

$$MD_i^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^t \mathbf{S}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}). \tag{7}$$

In case of multivariate normality, MD_i^2 follows approximately a chi-square distribution with p degrees of freedom, χ_p^2. Thus, large values of MD_i^2 are suspicious to be potential outliers. Taking a certain quantile, like the 97.5% quantile $\chi^2_{p;0.975}$ of this distribution can be used as an outlier cutoff: observations for which the MD_i^2 exceeds this quantile can be declared as potential outliers.

It is clear that classical estimators of location (arithmetic mean) and covariance (sample covariance matrix) are not useful for this kind of outlier detection since they are themselves influenced by outliers. Thus, robust counterparts are needed.

Before listing such robust versions, we go in more detail into the problem of the non-robustness of classical estimators. The outlier identification procedure based on $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ will suffer from the following two problems [Rousseeuw and Leroy, 1987]:

1. *Masking:* multiple outliers can distort the classical estimates of mean $\bar{\boldsymbol{x}}$ and covariance $\boldsymbol{S}$ in such a way (attracting $\bar{\boldsymbol{x}}$ and inflating $\boldsymbol{S}$) that they do not get necessarily large values of the Mahalanobis distance, and

2. *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 for the Mahalanobis distance.

The problems of masking and swamping are the reasons why diagnostic tools based on classical estimators are unreliable. Also "robustified" procedures that omit one observation in turn (leave-one-out) and the application of classical estimators on the remainder will not work, since such a procedure is not protecting against the effects of multiple outliers.

A reliable procedure for outlier detection based on Mahalanobis distance thus needs to use robust estimates of location $\boldsymbol{T}$ and covariance $\boldsymbol{C}$. The squared *robust distance* is then defined for a data point $\boldsymbol{x}_i, i = 1, \ldots, n$, as

$$RD_i^2 = (\boldsymbol{x}_i - \boldsymbol{T})^t \boldsymbol{C}^{-1} (\boldsymbol{x}_i - \boldsymbol{T}). \tag{8}$$

It is quite common to still use the cutoff value $\chi^2_{p;0.975}$, although the exact distribution of RD_i^2 will depend on the choice of $\boldsymbol{T}$ and $\boldsymbol{C}$. In Maronna and Zamar [2002] it was proposed to use a transformation of the cutoff value which should help the distribution of the squared robust distances RD_i^2 to resemble χ^2 for non-normal original data:

$$D_0 = \frac{\chi^2_{p,0.975} med(RD_1^2, \ldots, RD_n^2)}{\chi^2_{p,0.5}}. \tag{9}$$

For other alternatives which could lead to more accurate cutoff values, see Filzmoser et al. [2005], Hardin and Rocke [2005], Cerioli et al. [2009], Riani et al. [2009].

4.2 Affine equivariance

In the last several decades much effort was devoted to the development of affine equivariant estimators possessing a high breakdown point. The *affine equivariance* is a desirable property of a multivariate estimator, which makes the analysis independent of translations and rotations of the data. We say that a location estimator $\boldsymbol{T}$ of an $n \times p$ data set $\boldsymbol{X}$ is affine equivariant if

$$\boldsymbol{T}(\boldsymbol{X}\boldsymbol{A} + \mathbf{1}_n \boldsymbol{b}^t) = \boldsymbol{T}(\boldsymbol{X})\boldsymbol{A} + \boldsymbol{b} \tag{10}$$

for all p-dimensional vectors $\boldsymbol{b}$ and all nonsingular $p \times p$ matrices $\boldsymbol{A}$. The vector $\mathbf{1}_n$ is a column vector with all n components equal to 1. A scatter estimator $\boldsymbol{C}$ being a positive-definite symmetric $p \times p$ matrix is affine equivariant if

$$\boldsymbol{C}(\boldsymbol{X}\boldsymbol{A} + \mathbf{1}_n \boldsymbol{b}^t) = \boldsymbol{A}^t \boldsymbol{C}(\boldsymbol{X})\boldsymbol{A} \tag{11}$$

holds, again for all p-dimensional vectors $\boldsymbol{b}$ and all nonsingular $p \times p$ matrices $\boldsymbol{A}$. If an estimator is affine equivariant it transforms properly when the data are translated, rotated or the scale changes and thus the analysis is independent of the measurement scales of the variables or their translations or rotations. When plugging in affine equivariant robust location and scatter estimators into multivariate statistical methods, like discriminant analysis, these methods also inherit this property. For some methods, like principal component analysis (PCA) a weaker form of equivariance, namely *orthogonal equivariance* is sufficient. Orthogonal equivariance means that the estimator transforms properly under orthogonal transformations (but not affine transformations) of the data.

4.3 Statistical efficiency and computational feasibility

A very important performance criterion of any statistical procedure is its *statistical efficiency*, i.e. the precision of the estimate, and robust estimators are known in general as not very efficient. One way to increase the statistical efficiency of a high breakdown point estimator is to sacrifice the maximal breakdown point of 50% and work with lower, say 25% which in most of the cases is quite reasonable.

All the desirable features of a robust estimator listed above are useless if the estimator cannot be computed in a reasonable amount of time, also in high dimensions and with large amounts of data. Therefore the *computational*

feasibility is one of the most important features for the practical application of any estimator or procedure.

An early approach for multivariate location and scatter estimation was that of M-estimation introduced by Maronna [1976] which provides robust, affine equivariant and easy to compute estimates, but unfortunately these estimates have an unacceptably low breakdown point of $1/p$. Only later, this concept has been refined to the concept of MM estimators of Yohai [1987], which achieve high breakdown point and tunable efficiency.

Currently one of the most widely used estimators are the Minimum Covariance Determinant (MCD), S-estimators and the Stahel-Donoho estimator (see later in this section). These estimators can be configured in such a way as to achieve the theoretically maximal possible breakdown point of 50% which gives them the ability to detect outliers even if their number is as much as almost half of the sample size. If we give up the requirement for affine equivariance, estimators like the orthogonalized Gnanadesikan-Kettenring (OGK) estimator are available and the reward is an extreme gain in speed. For definitions, algorithms and references to the original papers it is suitable to use Maronna et al. [2006]. Most of these methods are implemented in the R statistical environment [R Development Core Team, 2009] and are available in the object-oriented framework for robust multivariate analysis [Todorov and Filzmoser, 2009].

4.4 Minimum covariance determinant (MCD) and minimum volume ellipsoid (MVE) estimators

The MCD estimator for a data set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ in $\Re^p$ is defined by that subset $\{\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_h}\}$ of h observations whose covariance matrix has the smallest determinant among all possible subsets of size h. The MCD location and scatter estimate $\mathbf{T}_{MCD}$ and $\mathbf{C}_{MCD}$ are then given as the arithmetic mean and a multiple of the sample covariance matrix of that subset

$$\begin{aligned} \mathbf{T}_{MCD} &= \frac{1}{h}\sum_{j=1}^{h} \boldsymbol{x}_{i_j} \\ \mathbf{C}_{MCD} &= c_{ccf}c_{sscf}\frac{1}{h-1}\sum_{j=1}^{h}(\boldsymbol{x}_{i_j} - \mathbf{T}_{MCD})(\boldsymbol{x}_{i_j} - \mathbf{T}_{MCD})^t. \end{aligned} \qquad (12)$$

The multiplication factors c_{ccf} (consistency correction factor) and c_{sscf} (small sample correction factor) are selected so that $\mathbf{C}$ is consistent at the multivariate normal model and unbiased at small samples [see Butler et al., 1993, Croux and Haesbroeck, 1999, Pison et al., 2002, Todorov, 2008]. Note, however, that only for a very small amount of trimming (h close to n), the proposed constant c_{ccf} allows to get a consistent estimator for the determinant and, therefore, for the whole covariance matrix. Otherwise, the use of this constant will produce overestimation. Since usually the amount of contamination in the data is unknown, a value like $h = [0.75\ n]$ is used in practice, resulting in an estimator with good statistical properties of the shape matrix, but not of the covariance matrix.

The breakdown point of the estimator is controlled by the parameter h. To achieve the maximal possible BP of the MCD, the choice for h is $\lfloor (n+p+1)/2 \rfloor$, but any integer h within the interval $[(n+p+1)/2, n]$ can be chosen, see Rousseeuw and Leroy [1987]. Here $\lfloor z \rfloor$ denotes the integer part of z which is not less than z. If $h = n$ then the MCD location and scatter estimate $\mathbf{T}_{MCD}$ and $\mathbf{C}_{MCD}$ reduce to the sample mean and covariance matrix of the full data set.

The MCD estimator is not very efficient at normal models, especially if h is selected so that maximal BP is achieved. To overcome the low efficiency of the MCD estimator, a reweighed version can be used. For this purpose a weight w_i is assigned to each observation $\boldsymbol{x}_i$, defined as $w_i = 1$ if $(\boldsymbol{x}_i - \mathbf{T}_{MCD})^t \mathbf{C}_{MCD}^{-1} (\boldsymbol{x}_i - \mathbf{T}_{MCD}) \leq \chi^2_{p,0.975}$ and $w_i = 0$ otherwise, relative to the raw MCD estimates $(\mathbf{T}_{MCD}, \mathbf{C}_{MCD})$. Then the reweighted estimates are computed as

$$\begin{aligned} \mathbf{T}_R &= \frac{1}{\nu} \sum_{i=1}^{n} w_i \boldsymbol{x}_i, \\ \mathbf{C}_R &= c_{r.ccf} c_{r.sscf} \frac{1}{\nu - 1} \sum_{i=1}^{n} w_i (\boldsymbol{x}_i - \mathbf{T}_R)(\boldsymbol{x}_i - \mathbf{T}_R)^t, \end{aligned} \tag{13}$$

where ν is the sum of the weights, $\nu = \sum_{i=1}^{n} w_i$. Again, the multiplication factors $c_{r.ccf}$ and $c_{r.sscf}$ are selected so that $\mathbf{C}_R$ is consistent at the multivariate normal model and unbiased at small samples [see Pison et al., 2002, Todorov, 2008, and the references therein]. The reweighted estimates $(\mathbf{T}_R, \mathbf{C}_R)$ have the same breakdown point as the initial (raw) *MCD* estimates but better statistical efficiency. The reweighted estimator should not be used

when contaminating observations are close to the boundary of the regular observations, because the outliers could then get masked.

The MCD estimator is popular also because of a fast algorithm for its computation [Rousseeuw and Van Driessen, 1999].

Figure 6 shows the influence of outliers onto classical estimates for location and scale as well as the result of the MCD estimate. Displayed are the same 110 data points from multivariate normal distribution. This figure (Figure 6) displays the corresponding 0.975% tolerance ellipse for the classical and robust estimates. Again it is clearly visible that the robust method can cope with outliers, whereas the classical estimates are highly affected by them.

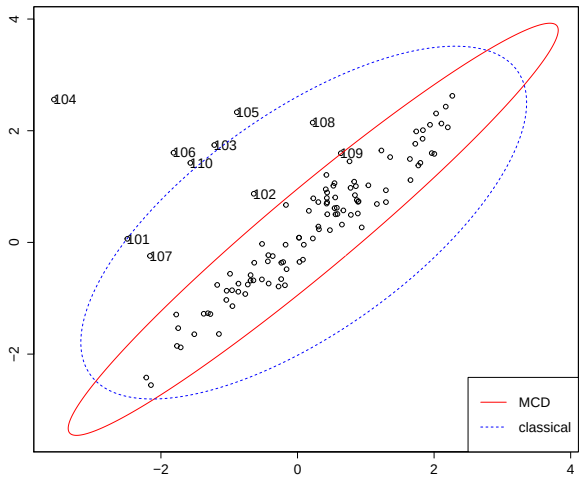


Figure 6: 97.5% tolerance ellipse for classical and MCD estimate for location and covariance applied to simulated data from 2-dimensional normal distribution with added contaminated data points.

The minimum volume ellipsoid (MVE) estimator introduced by Rousseeuw [1985] together with the MCD estimator searches for the ellipsoid of minimal volume containing at least half of the points in the data set $\mathbf{X}$. Then the location estimate is defined as the center of this ellipsoid and the covariance estimate is provided by its shape. Formally the estimate is defined as those

$\mathbf{T}_{MVE}, \mathbf{C}_{MVE}$ that minimize $\det(\mathbf{C})$ subject to

$$\#\{i : (\boldsymbol{x}_i - \mathbf{T})^t \mathbf{C}^{-1} (\boldsymbol{x}_i - \mathbf{T}) \leq c^2\} \geq \left\lfloor \frac{n+p+1}{2} \right\rfloor, \tag{14}$$

where # denotes the cardinality. The constant c is chosen as $\chi^2_{p,0.5}$.

The search for the approximate solution is made over ellipsoids determined by the covariance matrix of $p+1$ of the data points and by applying a simple but effective improvement of the sub-sampling procedure as described in Maronna et al. [2006], p. 198. Although there exists no formal proof of this improvement (as for MCD and LTS), simulations show that it can be recommended as an approximation of the MVE.

As with the MCD estimate, the MVE estimate is in general not very efficient.

Figure 7 displays the 0.975% tolerance ellipsoid of the MVE estimate and the classical estimate applied on the same data as in Figure 6.

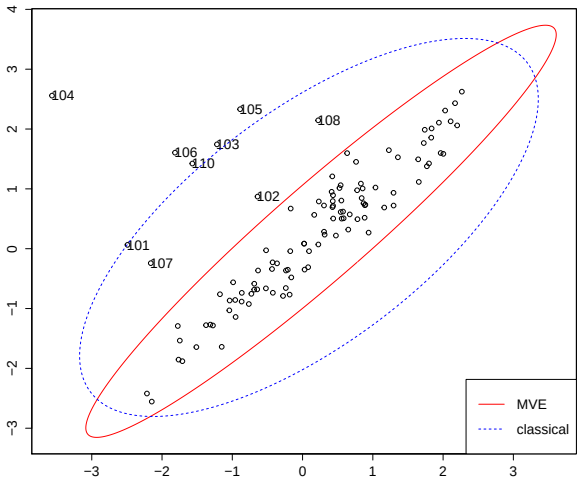


Figure 7: 97.5% tolerance ellipse for classical and MVE estimate for location and covariance applied to simulated data from 2-dimensional normal distribution with added contaminated data points.

4.5 The Stahel-Donoho estimator

The first affine equivariant estimator of location and scatter with high breakdown point was proposed by Stahel [1981b,a] and Donoho [1982] but became better known after the analysis of Maronna and Yohai [1995]. For a data set $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ in $\Re^p$ it is defined as a weighted mean and covariance matrix of the form given by Equation (13) where the weight w_i of each observation is inverse proportional to the "outlyingness" of the observation. Let the univariate outlyingness of a point $\boldsymbol{x}_i$ with respect to the data set $\mathbf{X}$ along a vector $\mathbf{a} \in \Re^p, ||\mathbf{a}|| \neq \mathbf{0}$, be given by

$$r(\boldsymbol{x}_i, \mathbf{a}) = \frac{|\boldsymbol{x}^t\mathbf{a} - m(\mathbf{a}^t\mathbf{X})|}{s(\mathbf{a}^t\mathbf{X})} \quad i = 1, \ldots, n, \tag{15}$$

where $(\mathbf{a}^t\mathbf{X})$ is the projection of the data set $\mathbf{X}$ on $\mathbf{a}$ and the functions $m()$ and $s()$ are robust univariate location and scale statistics, for example the median and MAD, respectively. When applying this idea to all possible univariate projections, one obtains in a natural way a measure for multivariate outlyingness of $\boldsymbol{x}_i$, which is defined by

$$r_i = r(\boldsymbol{x}_i) = \max_{\mathbf{a}} r(\boldsymbol{x}_i, \mathbf{a}). \tag{16}$$

The weights are computed by $w_i = w(r_i)$ where $w(r)$ is a non-increasing function of r, and $w(r)$ and $w(r)r^2$ are bounded. Maronna and Yohai [1995] use the weights

$$w(r) = \min\left(1, \left(\frac{c}{t}\right)^2\right) \tag{17}$$

with $c = \sqrt{\chi^2_{p,\beta}}$ and $\beta = 0.95$, that are known in the literature as "Huber weights".

Exact computation of the estimator is not possible and an approximate solution is found by subsampling a large number of directions $\mathbf{a}$ and computing the outlyingness measures $r_i, i = 1, \ldots, n$, for them. For each subsample of p points the vector $\mathbf{a}$ is taken as the norm 1 vector orthogonal to the hyperplane spanned by these points.

4.6 Orthogonalized Gnanadesikan/Kettenring

The MCD estimator and all other known affine equivariant high-breakdown point estimates are solutions to a highly non-convex optimization problem

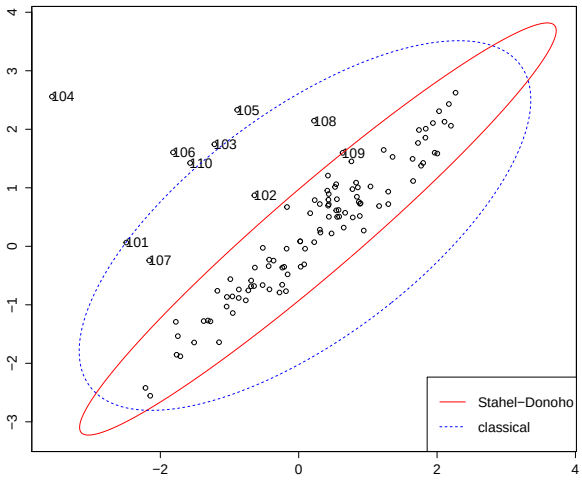


Figure 8: 97.5% tolerance ellipse for classical and Stahel-Donoho estimate for location and scale applied to simulated data from 2-dimensional normal distribution with added contaminated data points.

and as such pose a serious computational challenge. Much faster estimates with high breakdown point can be computed if one gives up the requirements of affine equivariance of the covariance matrix. Such an algorithm was proposed by Maronna and Zamar [2002] which is based on the simple robust bivariate covariance estimator s_{jk} proposed by Gnanadesikan and Kettenring [1972] and studied by Devlin et al. [1981]. For a pair of random variables Y_j and Y_k and a standard deviation function $\sigma()$, s_{jk} is defined as

$$s_{jk} = \frac{1}{4}\left(\sigma\left(\frac{Y_j}{\sigma(Y_j)} + \frac{Y_k}{\sigma(Y_k)}\right)^2 - \sigma\left(\frac{Y_j}{\sigma(Y_j)} - \frac{Y_k}{\sigma(Y_k)}\right)^2\right). \tag{18}$$

If a robust function is chosen for $\sigma()$ then s_{jk} is also robust and an estimate of the covariance matrix can be obtained by computing each of its elements s_{jk} for each $j = 1, \ldots, p$ and $k = 1, \ldots, p$ using Equation (18). This estimator does not necessarily produce a positive definite matrix (although symmetric) and it is not affine equivariant. Maronna and Zamar [2002] overcome the lack of positive definiteness by the following steps:

1. Define $\boldsymbol{y}_i = \boldsymbol{D}^{-1}\boldsymbol{x}_i, i = 1, \ldots, n$, with $\mathbf{D} = \text{diag}(\sigma(X_1), \ldots, \sigma(X_p))$, where $X_l, l = 1, \ldots, p$, are the columns of the data matrix $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. Thus a normalized data matrix $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ is computed.

2. Compute the matrix $\boldsymbol{U} = (u_{jk})$ as $u_{jk} = s_{jk} = s(Y_j, Y_k)$ if $j \neq k$ or $u_{jk} = 1$ otherwise. Here $Y_l, l = 1, \ldots, p$, are the columns of the transformed data matrix $\boldsymbol{Y}$ and $s(.,.)$ is a robust estimate of the covariance of two random variables like the one in Equation (18).

3. Obtain the "principal component decomposition" of $\boldsymbol{Y}$ by decomposing $\boldsymbol{U} = \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}^t$ where $\boldsymbol{\Lambda}$ is a diagonal matrix $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$ with the eigenvalues λ_j of $\boldsymbol{U}$ and $\boldsymbol{E}$ is a matrix with columns the eigenvalues $\boldsymbol{e}_j$ of $\boldsymbol{U}$.

4. Define $\boldsymbol{z}_i = \boldsymbol{E}^t\boldsymbol{y}_i = \boldsymbol{E}^t\boldsymbol{D}^{-1}\boldsymbol{x}_i$ and $\boldsymbol{A} = \boldsymbol{D}\boldsymbol{E}$. Then the estimator of $\boldsymbol{\Sigma}$ is $\boldsymbol{C}_{OGK} = \boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{A}^t$ where $\boldsymbol{\Gamma} = \text{diag}(\sigma(Z_j)^2), j = 1, \ldots, p$, and the location estimator is $\boldsymbol{T}_{OGK} = \boldsymbol{A}\boldsymbol{m}$ where $\boldsymbol{m} = m(\boldsymbol{z}_i) = (m(Z_1), \ldots, m(Z_p))$ is a robust mean function.

This can be iterated by computing $\boldsymbol{C}_{OGK}$ and $\boldsymbol{T}_{OGK}$ for $\boldsymbol{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$ obtained in the last step of the procedure and then transforming back to the original coordinate system. Simulations [Maronna and Zamar, 2002] show that iterations beyond the second did not lead to an improvement.

Similar as for the MCD estimator, a one-step reweighting can be performed using Equations (13), but the weights w_i are based on the 0.9 quantile of the χ^2_p distribution (instead of 0.975) and the correction factors $c_r.ccf$ and $c_r.sscf$ are not used.

In order to complete the algorithm we need a robust and efficient location function $m()$ and scale function $\sigma()$, and one proposal is given in Maronna and Zamar [2002]. Further, the robust estimate of covariance between two random vectors $s()$ given by Equation (18) can be replaced by another one. This *OGK* algorithm preserves the positive definiteness of the covariance matrix and is "almost affine equivariant". Even faster versions of this algorithm were proposed by Alqallaf et al. [2002].

4.7 S estimates and MM estimates

S estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ were introduced by Davies [1987] and further studied by Lopuhaä [1989] [see also Rousseeuw and Leroy, 1987, p. 263]. For a data

set of p-variate observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ an S estimate $(\mathbf{T}, \mathbf{C})$ is defined as the solution of $\sigma(d_1, \ldots, d_n) = \min$ where $d_i = (\boldsymbol{x} - \mathbf{T})^t \mathbf{C}^{-1} (\boldsymbol{x} - \mathbf{T})$ and $\det(\mathbf{C}) = 1$. Here $\sigma = \sigma(\boldsymbol{z})$ is the M-scale estimate of a data set $\boldsymbol{z} = \{z_1, \ldots, z_n\}$ defined as the solution of $\frac{1}{n}\Sigma\rho(z/\sigma) = \delta$ where ρ is nondecreasing, $\rho(0) = 0$ and $\rho(\infty) = 1$ and $\delta \in (0, 1)$. An equivalent definition is to find the vector $\mathbf{T}$ and a positive definite symmetric matrix $\mathbf{C}$ that minimizes $\det(\mathbf{C})$ subject to

$$\frac{1}{n} \sum_{i=1}^{n} \rho(d_i) = b_0 \tag{19}$$

where d_i and ρ are defined as above and the constant b_0 is chosen for consistency of the estimator.

As shown by Lopuhaä [1989], S estimators have a close connection to the M estimators and the solution $(\mathbf{T}, \mathbf{C})$ is also a solution to an equation defining an M estimator as well as a weighted sample mean and covariance matrix:

$$\begin{aligned} d_i^j &= [(\boldsymbol{x}_i - \mathbf{T}^{(j-1)})^t (\mathbf{C}^{(j-1)})^{-1} (\boldsymbol{x} - \mathbf{T}^{(j-1)})]^{1/2} \\ \mathbf{T}^{(j)} &= \frac{\Sigma w(d_i^{(j)}) \boldsymbol{x}_i}{\Sigma w(d_i^{(j)})} \\ \mathbf{C}^{(j)} &= \frac{\Sigma w(d_i^{(j)})(\boldsymbol{x}_i - \mathbf{T}^{(j)})(\boldsymbol{x}_i - \mathbf{T}^{(j)})^t}{\Sigma w(d_i^{(j)})} \end{aligned} \tag{20}$$

There are several algorithms for computing the S estimates: (i) *SURREAL* proposed by Ruppert [1992] as an analog to the algorithm proposed by the same author for computing S estimators of regression; (ii)*Bisquare S estimation with HBDP start*: as described in Maronna et al. [2006]; (iii) *Rocke type S estimates* [Rocke, 1996] and (iv) *Fast S estimates* proposed by Salibian-Barrera and Yohai [2006]. For more details about the computation of these estimates, see Todorov and Filzmoser [2009]. Rocke [1996] warns that when using S estimators in high dimension, they can fail to reject as outliers points that have large distances from the main mass of points, although attaining a breakdown point approaching 50%,

Tatsuoka and Tyler [2000] introduced the multivariate MM estimators together with the broader class of estimators which they call "multivariate M-estimators with auxiliary scale". They estimate the scale by means of a very robust S estimator, and then estimate the location and covariance using a different ρ-function with better efficiency at the normal model. The location

and covariance estimates inherit the breakdown point of the auxiliary scale and can be seen as a generalization of the regression MM estimators of Yohai [1987].

Figure 9 displays the 97.5% tolerance ellipsoid of the OGK-estimate and classical estimate applied on 110 data points simulated from multivariate normal distribution as in Figure 6.

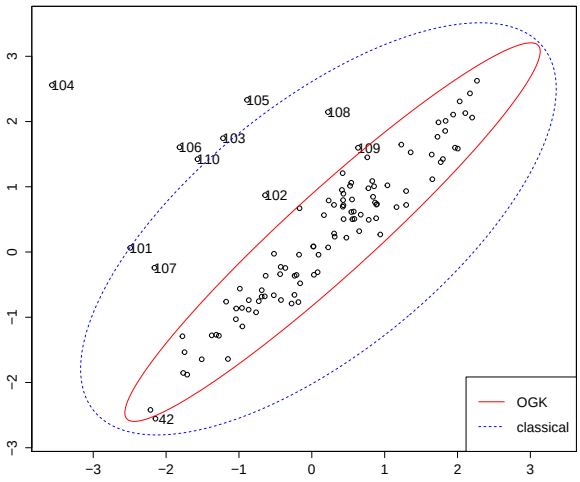


Figure 9: 97.5% tolerance ellipse for classical and OGK estimate for location and scale applied to simulated data from 2-dimensional normal distribution with added contaminated data points.

4.8 The PCOut Algorithm

The *PCOut* algorithm proposed by Filzmoser et al. [2008] combines two complementary measures of outlyingness in two subsequent steps. In the first step the target are the location outliers (mean-shift outliers, described by a different location vector) and in the second step the aim is to detect scatter outliers (variance inflation/deflation outliers, which possess a different scatter matrix than the rest of the data). The algorithm thus provides a final score that allows the ranking of the observations according to their deviation from the bulk of the data. A brief sketch of the algorithm will be presented

in the following sections, all details and many examples are available in the original paper.

4.8.1 Preprocessing

The algorithm starts by several preprocessing steps, the first of which is robust rescaling each component by the coordinate-wise median and MAD,

$$x^*_{ij} = \frac{x_{ij} - med(x_{1j}, \ldots, x_{nj})}{MAD(x_{1j}, \ldots, x_{nj})}, j = 1, \ldots, p. \tag{21}$$

In order to be able to perform this rescaling it is necessary either to omit the dimensions with MAD equal to zero or to use another measure. From this rescaled data the covariance matrix is calculated and eigen-decomposition is performed which results in a semi robust PCA. Only those p^* components which amount to at least 99% of the total variance are retained. Skipping out the components which contribute only useless noise, a representation in a lower dimensional p^* space, $p^* < p$ is obtained. This step solves also the problem with $p \gg n$ since we can select $p^* < n$. This decomposition can be represented by

$$\mathbf{Z} = \mathbf{X}^*\mathbf{V} \tag{22}$$

where $\mathbf{V}$ is the matrix of eigenvectors and $\mathbf{X}^*$ is the matrix with components x^*_{ij}.

These principal components are rescaled by the median and the MAD similarly as above,

$$z^*_{ij} = \frac{z_{ij} - med(z_{1j}, \ldots, z_{nj})}{MAD(z_{1j}, \ldots, z_{nj})}, j = 1, \ldots, p^*. \tag{23}$$

The resulting matrix $\mathbf{Z}^* = (z^*_{ij})$ is the input for the next two steps of the algorithm.

4.8.2 Detection of location outliers

The detection of location outliers, data points generated from a distribution with shifted mean but same covariance as the main data distribution, starts by calculation of component-wise robust kurtosis measure according to:

$$w_j = \left| \frac{1}{n} \sum_{i=1}^{n} \frac{(z^*_{ij} - med(z^*_{1j}, \ldots, z^*_{nj}))^4}{MAD(z^*_{1j}, \ldots, z^*_{nj})^4} - 3 \right|, j = 1, \ldots, p^*, \tag{24}$$

where z_{ij}^* are the rescaled principal components from Equation (23). Equation (24) assigns higher weights to the components where outliers clearly stand out. If no outliers are present in a given component then the principal component is expected to be approximately normally distributed and the kurtosis will be close to zero. Therefore each dimension $q = 1, \ldots, p^*$ is weighted proportionally to the absolute value of the kurtosis given by Equation (24). Since the components are uncorrelated it is possible to calculate a robust Mahalanobis distance utilizing the distance from the median (as scaled by MAD),

$$RD_i = \sqrt{\sum_{j=1}^{p^*} (z_{ij}^* w_j^*)^2}, \tag{25}$$

which then are translated to

$$d_i = RD_i \frac{\sqrt{\chi^2_{p^*,0.5}}}{med(RD_1, \ldots, RD_n)} \text{ for } i = 1, \ldots, n \tag{26}$$

in order to bring the empirical distances $\{d_i\}$ closer to $\chi^2_{p^*}$. These distances are used in the translated biweight function [Rocke, 1996] to assign weights to each observation. These weights are used as a measure of outlyingness. Filzmoser et al. [2008] claim that the translated biweight function (shown below) has certain advantages over other weighting schemes they have experimented with. The weights for the observations are calculated as follows,

$$w_i = \begin{cases} 0, & d_i \geq c \\ \left(1 - \left(\frac{d_i - M}{c - M}\right)^2\right)^2, & M < d_i < c, \\ 1, & d_i \leq M, \end{cases} \tag{27}$$

where $i = 1, \ldots, n$ and c is given by

$$c = med(d_1, \ldots, d_n) + 2.5.MAD(d_1, \ldots, d_n). \tag{28}$$

M is found by sorting the distances $\{d_1, \ldots, d_n\}$ in ascending order and taking M equal to the distance at position $\lfloor n/3 \rfloor$. These weights $w_{1i}, i = 1, \ldots, n$, are kept to combine with the result from step two to obtain the final weights.

4.8.3 Detection of scatter outliers

The scatter outliers, outliers generated by a distribution with an inflated covariance matrix, are searched for in the space defined by $\mathbf{Z}^*$ in Equation (23) by calculating the Euclidean norm for data in principal component space (which is equivalent to the Mahalanobis distance in the original data space but is much faster to compute). The weights $w_{2i}, i = 1, \ldots, n$, of the second step are calculated using again the translated biweight function from Equation (27) and setting $M^2 = \chi^2_{p^*,0.25}$ and $c = \chi^2_{p^*,0.99}$.

4.8.4 Computation of final weights

The weights resulting from the two phases of the algorithm are combined into final weights according to

$$w_i = \frac{(w_{1i} + s)(w_{2i} + s)}{(1 + s)^2}, i = 1, \ldots, n, \tag{29}$$

where typically $s = 0.25$. Outliers are then identified as points having weights $w_i < 0.25$.

4.9 Epidemic algorithm

Proposed by Béguin and Hulliger [2004], this method does not assume a certain model distribution of the data. The algorithm simulates an epidemic which starts at a multivariate robust center (sample spatial median) and propagates through the point cloud. The infection time is used to judge on the outlyingness of the points. The latest infected points or those not infected at all are considered outliers. The adaption of the algorithm to missing values is straightforward by leaving out missing values from the calculations of the univariate statistics (medians and median absolute deviations) as well as from the distance calculations. If too many items are missing in a pair of observations the corresponding distance is set to infinity (in the practical implementation a simplified criterion is applied excluding observations with less than p/2 observed items from the epidemic). The difficulty in using this algorithms is the crucial importance of the choice of the transmission function, which determines the probability that a point is infected given another point and the distance between them, the reach of the infection and the selection of the deterministic mode of infection. In other words, the

algorithm is based on many parameters and in practice it is difficult to find optimal parameter values. In addition, the computation time of the epidemic algorithm is high.

4.10 Bacon-EEM

This algorithm [Beguin and Hulliger, 2008] developed in the framework of the EUREDIT project is based on the algorithm proposed by [Billor et al., 2000], which in turn is an improvement over an earlier "forward search" based algorithm by one of the authors. It is supposed to be a balance between affine equivariance and robustness. The algorithm starts from a data set that is supposed to be outlier free and moves forward by inspecting the rest of the observations. Good points (non-outlying observations) are added as long as possible. The adaptation for incomplete data consists in replacing the computation of the location and covariance at each step by an EM-algorithm [see also Beguin and Hulliger, 2008, Todorov et al., 2011].

In the following, the BACON algorithm is described in more detail. The BACON algorithm, proposed by Billor et al. [2000], is a step-wise algorithm for robust estimates of location and covariance. The description of the algorithm in this work is the one used by the BACON-EEM algorithm, see Beguin and Hulliger [2008]. For a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ denote $\boldsymbol{\mu}_{\mathbf{X}}$, $\boldsymbol{\Sigma}_{\mathbf{X}}$ and $MD_{\mathbf{X}}(\mathbf{x})$ as the classical estimates for location and covariance as well as the corresponding Mahalanobis distance of an observation $\mathbf{x}$.
The first step of the BACON algorithm consists of estimating an initial "good" subset G. There are two ways to determine such a subset. For the calculations in this work the set G is determined by the $c \cdot p$ points, $c = 3$, with smallest Mahalanobis distance $MD_{\mathbf{X}}(x_i),\ i = 1, \ldots, n$. Starting with the subset G, the BACON algorithm performs the following steps:

1. Compute the Mahalanobis distances, corresponding to the subset G, $MD_{\mathbf{G}}(\mathbf{x}_i) = (\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{G}})^t \boldsymbol{\Sigma}_{\mathbf{G}}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{G}}),\ i = 1, \ldots, n$, for every observation in $\mathbf{X}$.

2. Determine a new subset G', containing all the points with $MD_{\mathbf{G}}(\mathbf{x}_i) < c_{npr} \chi^2_{p,\alpha/n}$, with $c_{npr} = c_{np} + c_{hr}$ as correction factor, where $c_{np} = 1 + (p+1)/(n-p) + 1/(n-h-p),\ c_{hr} = \max\{0, (h-r)/(h+r)\},\ h = \lceil (n+p+1)/2 \rceil$ and $r = |G|$.

3. If $G' = G$ stop, otherwise set G to G' and go to (1).

Observations which are not contained in the final set G are declared outliers.

EEM algorithm: The EEM algorithm is an extension to the Expectation/Maximization algorithm, which can be used to estimate location and covariance in a data set with incomplete observation. The EM algorithm contains an E-step and a M-step which are iterated sequentially until convergence.
Let given data $\mathbf{X} \in \mathbb{R}^{n \times p}$ be made up of observed and missing values $\mathbf{X} = \mathbf{X}_o \cup \mathbf{X}_m$. Furthermore the missingness mechanism is ignorable and the missingness is independent from the sample. Assuming that the observations where generated by multivariate normal distribution with density $f(\mathbf{x}, \theta)$, the complete log-likelihood can be written as

$$l(\theta|\mathbf{X}) = \boldsymbol{\nu}(\theta)^t \cdot \mathbf{T}(\mathbf{X}) + Ng(\theta) + c,$$

with $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_k)$ as the canonical form of the parameter $\boldsymbol{\theta}$ and $\mathbf{T}(\mathbf{X}) = (\mathbf{T}_1(\mathbf{X}), \ldots, \mathbf{T}_k(\mathbf{X}))$ as the vector of complete-data sufficient statistics. Since the data was generated by a multivariate normal distribution the sufficient statistics are composed of the sums $\sum_{i=1}^n \mathbf{x}_i^k$ and sums of products $\sum_{i=1}^n \mathbf{x}_i^k \mathbf{x}_i^l$, $1 \leq k, l \leq p$. In the E-step the conditional expectations of these sums are calculated, given the preliminary parameter $\theta^{(t)}$ and the observed data $\mathbf{X}_o$.
For these conditional expectations it can be shown that

$$\mathbb{E}\left(\sum_{i=1}^n \mathbf{x}_i^k | \mathbf{X}_o, \theta^{(t)}\right) \;=\; \sum_{i=1}^n \mathbb{E}\left(\mathbf{x}_i^k | \mathbf{x}_i^{obs}, \theta^{(t)}\right) \;,\; 1 \leq k \leq p,$$

and

$$\mathbb{E}\left(\sum_{i=1}^n \mathbf{x}_i^k \mathbf{x}_i^l | \mathbf{X}_o, \theta^{(t)}\right) \;=\; \sum_{i=1}^n \mathbb{E}\left(\mathbf{x}_i^k \mathbf{x}_i^l | \mathbf{x}_i^{obs}, \theta^{(t)}\right) \;,\; 1 \leq k, l \leq p.$$

To estimate these conditional expectations a Horvitz-Thompson estimator is used. The resulting estimates are given by

$$T^{k0} = \sum_s w_i \mathbb{E}(\mathbf{x}_i^k | \mathbf{x}_i^{obs}, \theta^{(t)}) \;,\; 1 \leq k \leq p,$$

and

$$T^{kl} = \sum_s w_i \mathbb{E}(\mathbf{x}_i^k \mathbf{x}_i^l | \mathbf{x}_i^{obs}, \theta^{(t)}) \;,\; 1 \leq k, l \leq p.$$

By using an estimate for the conditional expectations the E-step is called the estimated expectation step (EE-step), see Beguin and Hulliger [2008].

Using these estimates for the complete log-likelihood function yields a so called average population likelihood, which is then maximized with regard to θ (M-step). The solution, $\theta^{(t+1)}$, is given by

$$\theta^{(t+1)} \;=\; SWP\,[0]\left(\frac{(T^{kl})_{0\leq k,l\leq p}}{\sum_s w_i}\right),$$

where $(T^{kl})_{0\leq k,l\leq p}$ is the symmetric $(p+1)\times(p+1)$ matrix from the EE-step, with T^{00} set to 1 and $SWP\,[0]$ is the sweep operator on the first line/column of the matrix.

Combining the BACON algorithm with the EEM-algorithm yields a robust estimate for location and covariance which can handle missing values in the given data. The steps of the BACON-EEM algorithm are as follows:

1. Calculate a starting set G by using the $c \cdot p$ observations with minimal squared marginal Mahalanobis distance, MD^2_{marg}. The marginal Mahalanobis distance can be used if observations ($\mathbf{x}$) have an unobserved ($\mathbf{x}_m$) and observed part ($\mathbf{x}_o$) and is defined by

   $$MD^2_{marg} \;=\; \frac{p}{q}(\mathbf{x}_o - \boldsymbol{\mu}_o)^t \boldsymbol{\Sigma}^{-1}_{oo}(\mathbf{x}_o - \boldsymbol{\mu}_o),$$

   where $\boldsymbol{\mu}_o$ and $\boldsymbol{\Sigma}_{oo}$ are part of the location vector and covariance matrix corresponding to $\mathbf{x}_o$. The factor p/q, with p as the number of variables and $q = \sum_k r_{ik}$ as the number of non-missing variables, are meant as scaling factor. The subset G can also be determined by using the coordinate-wise median, but for the calculations in this work the former method is used.

2. Compute $\hat{\boldsymbol{\mu}}_G$ and $\hat{\boldsymbol{\Sigma}}_G$ using the EEM-algorithm.

3. Calculate the squared marginal Mahalanobis distances $MD^2_G(\mathbf{x}_i)$ for $i = 1, \ldots, n$ and determine a new subset G' by those observations for which $MD^2_G(\mathbf{x}_i) < c_{\hat{N}p\hat{r}}\chi^2_{p,\alpha}$.

4. If $G = G'$ stop, otherwise set G to G' and go to (2).

Observations which are not in the final subset G are declared outliers. For more information on the BACON-EEM algorithm, see Beguin and Hulliger [2008].

5 Imputation of outliers

5.1 Adjusting potential outliers for univariate methods

After locating potential outliers, the data points will not be discarded since this would result in a heavy loss of information. Instead, the values or weights of the potential outliers will be adjusted. The adjustment of the outliers depends on the rule that is used for outlier detection:

- When using methods that are based on the rule "robust location $\pm$ constant times robust measure of scale", potential outliers will be placed to the upper or lower boundaries defined by Equations (2) and (3).
- In the case of the boxplot or the adjusted boxplot rule, potential outliers will be replaced with the upper or lower fences defined by (4).
- For the Pareto tail modeling, the potential outliers will be dealt with in two different ways (see also Alfons and Templ [2013]:
  - **Calibration of potential outliers:** Values larger than a certain quantile of the fitted distribution will receive a sample weight equal to 1 and the weights of the remaining observations are adjusted accordingly by calibration.
  - **Replacement of potential outliers:** Values larger than a certain quantile of the fitted distribution will be replaced by values drawn from the fitted distribution. This is shown in Figure 10. The solid black line represents the distribution of the original values, the grey line corresponds to the Pareto fit, and the red line symbolizes the distribution of values drawn from the Pareto distribution. The order of the original values will be preserved.

5.2 Adjusting potential outliers for multivariate methods

As mentioned in the previous section, dealing with potential outliers after they have been detected is an important task. Discarding them would result

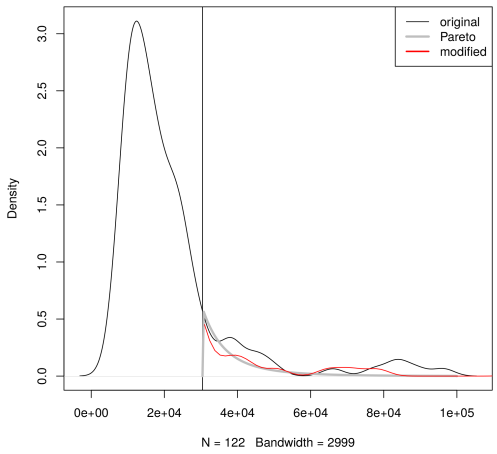


Figure 10: Replacement of the whole tail by draws from a fitted Pareto distribution.

in heavy loss of information as well as a heavy influence on further calculations on the data set. Therefore, imputing the data points would be more appropriate. In the case of one-dimensional outlier detection one of the presented ways of dealing with outliers was to move them to the upper or lower boundaries of the interval which ranges over the "clean" bulk of the data set. When it comes to multivariate outliers, this adjustment of potential outliers to preserve meaningful observations is not as straightforward as in the one-dimensional case. Nevertheless, multivariate outliers can be imputed in a similar way. In this work potential outliers will be imputed by projecting them onto the boundaries of a 97.5% tolerance ellipse. This procedure is described in the following.

Using one of the above mentioned methods to robustly estimate location and covariance it is possible to detect potential outliers by using robust distances. After these outliers have been identified, they will be moved towards the 97.5% tolerance ellipse of the previously calculated location and covariance estimates, in the direction of the robust center of the data set. To be more precise, let $\mathbf{T}_R$ and $\mathbf{C}_R$ be the robust estimates of location and covariance of the p-dimensional observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Furthermore, let $\mathbf{x}_j$ be a potential outlier, i.e. $d(\mathbf{x}_j, \mathbf{T}_R, \mathbf{C}_R) > \chi^2_{p;0.975}$. To replace this point by a the point on the 97.5% tolerance ellipse into the direction of the location $\mathbf{T}_R$

implies that the imputed observation $\tilde{\mathbf{x}}_j$ must be on the straight line between $\mathbf{x}_j$ and $\mathbf{T}_R$, which means $\tilde{\mathbf{x}}_j \in \{\alpha\mathbf{x}_j + (1-\alpha)\mathbf{T}_R, 0 \leq \alpha \leq 1\}$. In addition to that, the imputed point $\tilde{\mathbf{x}}_j$ must lie on the tolerance ellipse which results in $d(\tilde{\mathbf{x}}_j, \mathbf{T}_R, \mathbf{C}_R) = \chi^2_{p;0.975}$. To conclude, the imputed observation must solve the following equalities,

$$\begin{aligned}(\tilde{\mathbf{x}}_j - \mathbf{T}_R)^t \mathbf{C}_R^{-1} (\tilde{\mathbf{x}}_j - \mathbf{T}_R) &= \chi^2_{p;0.975} \\ \tilde{\mathbf{x}}_j &= \alpha\mathbf{x}_j + (1-\alpha)\mathbf{T}_R \\ \alpha &\in [0,1] \quad .\end{aligned}$$

To determine α, one can simply use the first two equations to obtain

$$\begin{aligned}((\alpha\mathbf{x}_j + (1-\alpha)\mathbf{T}_R) - \mathbf{T}_R)^t \mathbf{C}_R^{-1} ((\alpha\mathbf{x}_j + (1-\alpha)\mathbf{T}_R) - \mathbf{T}_R) &= \chi^2_{p;0.975} \\ (\alpha(\mathbf{x}_j - \mathbf{T}_R))^t \mathbf{C}_R^{-1} (\alpha(\mathbf{x}_j - \mathbf{T}_R)) &= \chi^2_{p;0.975} \\ \alpha^2 \underbrace{(\mathbf{x}_j - \mathbf{T}_R)^t \mathbf{C}_R^{-1} (\mathbf{x}_j - \mathbf{T}_R)}_{d(\mathbf{x}_j, \mathbf{T}_R, \mathbf{C}_R)} &= \chi^2_{p;0.975} \quad .\end{aligned}$$

Finally we get that $\alpha = \sqrt{\frac{d(\mathbf{x}_j, \mathbf{T}_R, \mathbf{C}_R)}{\chi^2_{p;0.975}}}$ and the imputed observation $\tilde{\mathbf{x}}_j = \alpha\mathbf{x}_j + (1-\alpha)\mathbf{T}_R$.

Figure 11 displays the imputation of outliers from data generated by multivariate normal distribution. For this data set, 100 observations have been simulated from a multivariate normal distribution with $\boldsymbol{\mu} = (0,0)^t$ and $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} 1 & 0.95 \\ 0.95 & 1 \end{smallmatrix}\right)$. In addition, 10 data points, were generated from a multivariate normal distribution with $\boldsymbol{\mu} = (-1.5, 1.5)^t$ and $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. These 10 outliers with different mean and covariance as the data majority are replaced using the above strategy. The blue symbols "+" indicate the positions of the data points after imputation.

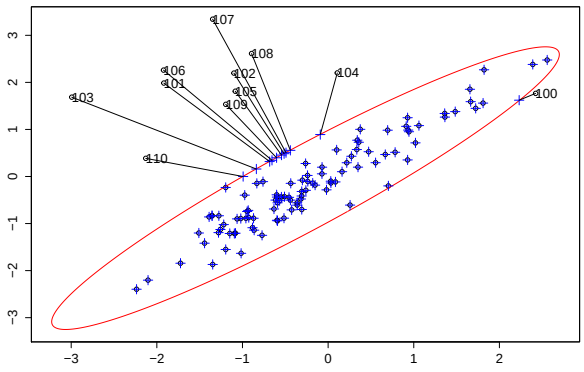


Figure 11: Imputation of potential outliers of data from 2-dimensional normal distribution with added contaminated data points.

6 Numerical study

In this section, univariate and multivariate outlier detection methods are applied on household expenditure data from various countries. The data contain information of the households; annual consumption expenditure, in local currency, by product and service (the list of which is specific to each survey). One way to analyze these household data is by estimating the Gini coefficient over the yearly consumption. In this context, the Gini coefficient would measure the inequality of consumption in terms of monetary value between the surveyed households. Since the data was generated through large surveys, extreme values or measurement errors can occur which can influence the estimation of the Gini coefficient [Alfons et al., 2013]. Therefore it will be beneficial to detect and impute potential outliers beforehand. In the following, the outlier detection methods discussed in the previous sections are used to detect potential outliers in the household expenditure data. Afterwards, potential outliers will be imputed and the Gini coefficient will be calculated. It is not known how many extreme values or measurement errors are contained in the household expenditure data. Moreover, the "correct" or "true" value of the Gini coefficient is also not known. These circumstances make it difficult to see which of the outlier detection methods provide the most reliable results. To get a deeper understanding of how well these methods perform, a sensitivity analysis as well as simulations are conducted. The results should provide evidence on which methods are most suitable for outlier detection on large household expenditure data sets. It should also be noted that the outlier detection schemes presented in this work do not account for country specific information or characteristics, because the attempt is to present an outlier detection scheme which performs well on household expenditure data regardless of the origin of the data. The methods should also be applicable on household data from various countries, therefore it would be an unfavorable approach to the problem if country specific details were taken into account.

The calculations in the work have all been done using the programming language R and a variety of existing R-packages.

Before the calculations of the outlier detection methods can be presented, it is necessary to understand how the household expenditure data was extracted, what kind of information is presented in it, and how the data is structured.

6.1 Provided data and data structure

The household expenditure data used here comprises household surveys from five countries, namely Albania, Mexico, India, Malawi and Tajikistan conducted in the year 2008, 2010, 2009, 2010 and 2007, respectively. These data sets were provided by the World Bank. These data sets were originally taken from the national data producers and are the result of large household surveys. The surveys include sociodemographic characteristics of each household as well as information on the household structure, including household size, education and age structure of the household members. Furthermore, the participants were asked to state how much the household consumes in local currency over a given time horizon in various spending categories. These categories range from different kinds of food-products over general living expenditures like gas, electricity or water to expenses on education, health and others. The number and type of categories of products and services differ for each survey but have in common that the combined categories reflect most of the consumption of a household for a given time horizon.

Ideally, we would like to detect outliers at product/service level, not in values for aggregated categories of products and services. However, since many zero observations are present in the single variables, outlier detection may become unreliable or even infeasible. This is an issue in particular for the discussed multivariate methods, which require at least twice as many observations without zeros as variables (for good stability even much more). But also for univariate estimation it is recommended to work with a representative amount of non-zero data.

6.1.1 Harmonization of the data

Since the surveys and resulting data sets differ in methodology and terminology, the World Bank started to harmonize the resulting data into a common framework with common data dictionary. For this harmonization process a series of steps were carried out on each data set. This starts with the extraction of household characteristics and the calculation of annual consumption for all goods and services. Regarding the annualizing of consumption values this process is in many cases not trivial since multiplying the consumption of a household for a specific food or service by a factor would often not make sense. For instance, in the case of durable goods the annual consumption was calculated with the use of depreciation rates, if the necessary information was

present in the survey. In some categories this annualization was not always possible, like in case of expenditures for health over a time horizon of a week or month.

Another part of the harmonization process which led to difficulties was the mapping of the household expenditures of each survey onto a standardized framework for goods and services, namely the basic headings used by the International Comparison Program (ICP) 2005. This heading consists of 107 different categories. Since the local surveys did not always use this categorization it was not easy to map the given data sets onto these basic headings. In some cases there was a perfect match in the survey questionnaire to one of the basic headings; in other cases an item in the survey questionnaire corresponded to many basic headings and vice versa. None of the household surveys covered all ICP basic headings.

As a last step of the harmonization, a series of quality control tables were calculated to validate the harmonized data sets. Apart from the ICP basic headings the harmonization also provided some grouping which condensed the ICP basic headings. This results in the ICP class, the ICP group and ICP category code for which every code represents a rougher grouping of the former. This means that the ICP category code is a regrouping of the ICP group code and the ICP group code a regrouping of the ICP class code which is a regrouping of the ICP basic headings.

Since there are considerable differences between the original surveys the harmonized data sets are not fully comparable. For a more detailed description of this harmonization process, see Dupriez [2007].

6.1.2 Categories and missing values

The data sets for each country provided by the World Bank are divided into three files. One file corresponds to the household characteristics of the questionnaired households. Next to household ID's and household weights, the household data set includes 50 variables such as geographical region, household size, civil status, highest education of the household head, having a car, etc. The second data set includes individual information of the residences of each household, like sex, age, the relation in the household (e.g. grandchild), martial status and the household ID that is needed to merge this data set with the other data sets. The third data set corresponds to the consumption of each household. Some variables are (unnecessarily) repeated (such as geographical information). The most important variables are the ICP cate-

gories [see The World Bank Group, 2015, page 229, and further discussion below] and total consumption (for each ICP class). The data sets also contain household and population weights. The calculations in this work will mainly focus on the household expenditures and household characteristics. For the latter, the information used is limited to geographical characteristics, household size and household weights. The consumption data of each household contain for each household the annual value of goods or services consumed in local currency. The total consumption for a specific good or service in local currency is furthermore divided into three subgroups. These groups consist of the value of good or service purchased in local currency, the estimated value of good or services which are home-produced and the value of goods or services which are received as a gift. Furthermore, the annual spendings of a household for a specific good or service are only listed if the total consumption of this good or service in local currency is greater than zero. This means that for goods or services for which there is no information of the total consumption for a specific household, it could be that the household has no annual spendings for this good or service, or that the information is missing. In both cases the corresponding value needs to be considered as zero.

In the context of not existing data entries it is important to note that the original surveys did not always use the ICP basic headings and that the ICP basic headings have such a variety that one can not expect that households have expenditures in each of these categories. Therefore, for many households their expenditures are only listed in some of the ICP basic headings leading to a lot of missing values or real zeros. This kind of incompleteness of the data set can result in some problems depending on the type of analysis that has do be done. To overcome these problems it is possible to use not the ICP basic headings but for example the ICP category code. This category code groups the different expenditures of each household into 13 different categories. Using these 13 different categories instead of the ICP basic headings results in a loss of information since fewer categories are present, but it also reduces the amount of missing values. Because of this, analyzing the expenditure data based on these 13 categories will be the preferred approach here. Other approaches are of course also possible, depending on the task of the analysis.

Even when only analyzing the 13 main expenditure categories, the amount of zero entries can be quite high for some categories. For the data sets from the five different countries mentioned above, the amount of zero entries for some categories is more than 50% of the corresponding sample size; in some

$$\begin{array}{cccccccc} \overbrace{\text{household ID}} & \overbrace{\text{household weights}} & \overbrace{\text{geographic characteristics}} & & \overbrace{\text{expenditure categories}} & & \\ \left(\begin{array}{c} 1 \\ \vdots \\ n \end{array} \right. & \begin{array}{c} 50 \\ \vdots \\ 4 \end{array} & \begin{array}{c} locationA \\ \vdots \\ locationZ \end{array} & \begin{array}{c} \ldots \\ \vdots \\ \ldots \end{array} & \begin{array}{c} 13 \\ \vdots \\ 25 \end{array} & \begin{array}{c} \ldots \\ \vdots \\ \ldots \end{array} & \begin{array}{c} 785 \\ \vdots \\ 353 \end{array} & \left. \begin{array}{c} \\ \\ \\ \end{array} \right) \end{array}$$

Table 2: Schematic structure of the data set after restructuring.

cases even more than 80% or 90%. For some of the outlier detection methods, these amounts are too large and the methods are not able to produce any result. To overcome such problems, an R routine was implemented to combine certain categories, and therefore reduce the number of zero entries. The main objective of this routine lies in the comparison of the interquartile range (IQR) for the expenditures of each category. If the IQRs of two or more categories overlap 'considerably' then those categories are combined. This regrouping scheme does not take into account a regrouping which results in a minimal fraction of zeros. Furthermore, this regrouping scheme represents solely a suggestion and is not proven to work in every case.

6.1.3 Data format

As mentioned before, household characteristics as well as consumption data are used for the calculations. Since the provided data sets are not in a favorable format to perform calculations, the data are extracted and restructured to a matrix format prior to the calculations. In this format the columns contain the household ID, household characteristics, the consumption categories and the sample weights, and the rows represent each survey participant. Table 2 shows the structure of the data set after restructuring the original data into the matrix format. If a household does not have expenditure information in a specific category or specific good or service, the corresponding entry in the restructured matrix format is zero. Furthermore, the expenditure categories display the annual consumption in local currency, and therefore, depending on the country the values can be reported in millions or higher and are transcended for the calculations by a factor 1000.

6.1.4 Selected data and further data preparation

In the following univariate and multivariate outlier detection methods are applied on the household surveys of Albania, Mexico, India, Malawi and Tajikistan. First and foremost, the results for the outlier detection methods are demonstrated on the Albanian household survey. One reason for this is that the Albanian household survey has a relatively small sample size with 3600 household, including 14785 individuals, allowing for quick computations. For the other surveys, the corresponding results are listed in a table later on. In the case of univariate outlier detection, the corresponding household sample weights are always considered for any calculations.

Before applying the outlier detection algorithms, the data set is transformed into the matrix format as discussed previously. Table 3 shows the number of zero entries for each category in the original data set. The amount of zeros for the Albanian household survey is similar for the other four surveys. Note that the high number of zeros in the category "Education" is due to the fact that public schools and universities are free of charge.

Category	**Zero entries**
Food and non-alcoholic beverages	2
Alcoholic beverages, tobacco and narcotic	1476
Clothing and footwear	347
Housing, water, electricity, gas and other fuels	25
Furnishings, household equipment, household maintenance	2
Health	1264
Transport	1468
Communication	407
Recreation and culture	19
Education	3278
Restaurants and hotels	1814
Miscellaneous goods and services	114

Table 3: Number of zero entries per category for the Albanian household survey data

Zeros will be treated in different ways whether univariate or multivariate outlier detection methods are applied. Univariate outlier detection methods

will either be applied on each of the expenditure categories separately or on the aggregated expenditures for every household.

When applying univariate outlier detection methods on each of the categories separately, zero values of each category will be discarded for outlier detection. Note that excluding these households from the calculations also influences the corresponding household weights. If some of the households are discarded, the household weights given in the data set do no longer add up to the whole population size. This might influence further calculations including the household sample weights. Whether the sample weights add up to the population size or not does not influence the presented outlier detection methods. Therefore, household sample weights will not be adjusted for outlier detection.

6.2 Univariate methods

The following univariate outlier detection methods are tested on the data set: rules defined by the boxplot and the adjusted boxplot [Vandervieren and Hubert, 2008], methods based on robust location $\pm$ constant $\times$ scale in conjunction with the Box-Cox transformation [Box and Cox, 1964, Maronna et al., 2006], and Pareto tail modeling [Alfons et al., 2013]. As mentioned in Section 3, the constant for the use of IQR or MAD for outlier detection is chosen as $c = 3$. In case of underlying normal distribution, this choice would lead to a very similar behavior as the boxplot rule.
Figure 12 shows the total annual consumption of each household in the Albanian household survey on the x-axis. The corresponding household weights are placed on the y-axis. Figure 12 also displays the weighted median and the Pareto threshold beyond which the Pareto distribution is fit to the data. Note that the household weights are, except for some few data points, comparably large and calculations for the weighted quantiles will therefore be quite robust. The other vertical lines represent the upper and lower bounds for the outlier detection schemes using weighted IQR and weighted MAD. Besides the conventional use of weighted IQR and MAD for univariate outlier detection these methods have also been used in combination with the Box-Cox transformation and the robustification of the Box-Cox transformation. In these cases, the corresponding outlier detection schemes have been conducted after data transformation, the calculated bounds for potential outliers were transformed back and the potential outliers were identified. The boxplot and the adjusted boxplot are displayed below the x-axis.

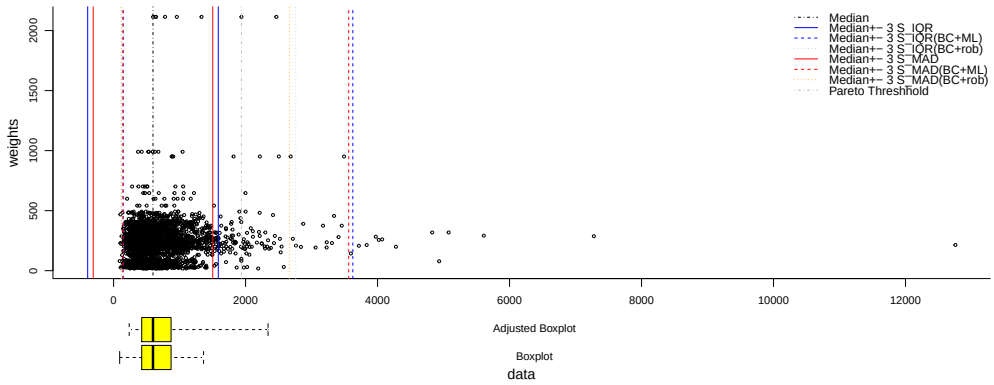


Figure 12: Univariate outlier detection methods applied to the total annual consumption data of the Albanian household survey.

For a better understanding of the calculated bounds, Table 4 shows the upper and lower bounds beyond which potential outliers are detected for the above mentioned univariate outlier detection schemes. In the context of Pareto tail modeling, the displayed value represents the Pareto threshold which is the point beyond which the Pareto distribution is fitted.

Figure 12 shows that the expenditure data is skewed to the right. Therefore, outlier detection schemes that are not adapted for skewness in the data will most likely show poor performance or can be expected to be not suitable for the problem. These methods are the boxplot and the calculations using IQR or MAD without the use of the Box-Cox transformation. Also Table 4 indicates that these methods are not suitable for the problem since their lower bounds are negative, which is in this context an invalid result. Not only the lower bounds are negative but also a large chunk of the data points with higher values are declared as potential outliers. The calculations using IQR or MAD in connection with the Box-Cox transformation seem to produce more reasonable results. This is purely based on the notion that these methods can deal with the skewness of the data and the calculated boundaries are all positive and do not seem to be so strict, in the case of the upper boundary. The same holds for the adjusted boxplot.

	upper bound	lower bound
IQR	1585.5592	-392.1509
Box-Cox with IQR	3619.7083	143.3289
robust Box-Cox with IQR	2758.3991	113.6876
MAD	1502.0627	-308.6544
Box-Cox with MAD	3557.0694	144.9035
robust Box-Cox with MAD	2663.8560	118.4252
Boxplot	1365.37964	-68.46019
adjusted Boxplot	2342.039	235.527
Pareto Modeling	1937.27	—

Table 4: Upper and lower bounds for univariate outlier detection schemes applied to the total annual consumption data of the Albanian household survey.

One reason to apply outlier detection for this kind of data is to measure inequality in the consumption distribution, e.g., by estimating the Gini coefficient. A classical calculation of the Gini coefficient would be highly influenced by outliers and produce arbitrary results. Therefore, locating and adjusting potential outliers would produce more reliable results for the Gini coefficient. On the other hand, the above presented univariate outlier detection schemes lead to quite different results for the identified number of outliers. Since the number and position of the true outliers in the data set is not available, it is not straightforward to determine which of the methods delivers the most reliable results. In addition, adjusting the potential outliers results in different 'corrected' data sets for each scheme which will then result in different values for the Gini coefficient. Also in case of the Gini coefficient, the true value, or a value which is expected to be true, is unknown. Nevertheless, it will be interesting to see and investigate the effect of the outlier detection methods and adjustments on the resulting Gini coefficients.

Figure 13 shows on the top left side the estimated values for the Gini coefficient after the outlier detection schemes have been applied and potential outliers have been adjusted. In the case of Pareto modeling, the detected outliers have in one case been replaced by values drawn from the fitted Pareto distribution (denoted by Pareto.rn), and in the other case the corresponding weights for the potential outliers have been set to 1 and the weights

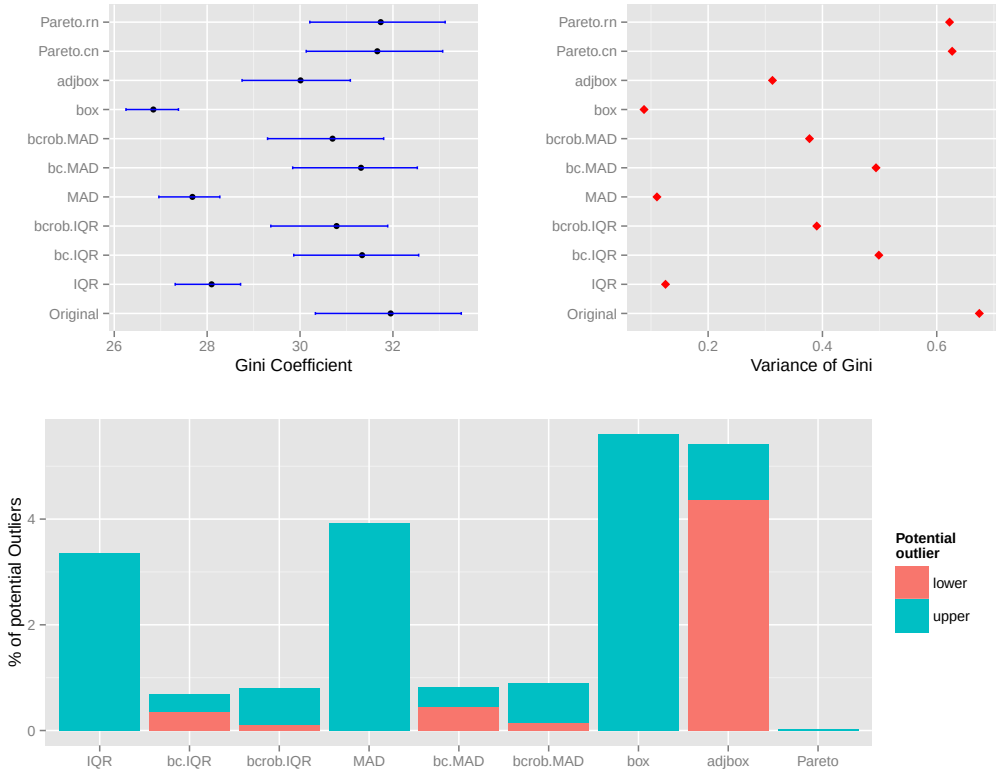


Figure 13: Top: Estimates for Gini coefficient (left) and Variance of Gini coefficient (right) of Albanian data set after outlier detection and adjustment for various outlier detection schemes.
Bottom: Share of upper and lower outlier for each outlier detection scheme applied on Albanian data set.

for the other observations have been re-calibrated accordingly (denoted by Pareto.cn). The blue horizontal lines indicate the 95% confidence interval for the estimated Gini coefficients. On the top right side of Figure 13, the variances of the estimated Gini coefficients are displayed for every univariate outlier detection scheme. The confidence intervals of the Gini coefficients and the variances were calculated using a bootstrapping routine. For a better understanding on the impact of the outlier detection schemes on the Gini coefficient, the estimated Gini coefficient for the original data set without any changes is also displayed. The bottom panel of Figure 13 shows the percentage of detected potential outliers divided by color into upper and lower outliers. It is easy to see that outlier detection methods which do not account for skewed data detect only upper outliers. Furthermore, the number of flagged potential outliers by those detection schemes is rather substantial. In combination with the top part of Figure 13 it is clear that the corresponding Gini values are heavily influenced after using the adjustment of these potential outliers. Interesting to see is that, aside from Pareto tail modeling, methods which adjust for the skewness of the data detect quite many lower potential outliers. The number of lower outliers in the case of the adjusted boxplot is especially high which indicates that the adjusted boxplot might not perform too well in this case. Reason for this is the fact that the majority of outliers in the data are expected to be upper outliers, and since the data are skewed to the right, upper outliers have a much larger influence on the spread of the data and therefore on the uniformity of the data values. Because of that, detecting lower outliers might not be unreasonable but their influence on the Gini coefficient is comparably small. This leads to the fact that, considering the data contain true outliers and the Gini coefficient is suspected to deliver biased results because of them, the bias caused by outliers will be primarily generated through upper outliers.

Univariate outlier detection methods were also applied to the household expenditure data from the countries India, Mexico, Malawi and Tajikistan. The results, including estimates of Gini coefficient and variance of Gini coefficient as well as resulting outlier information are shown in Table 5.

6.2.1 Column-wise implementation of univariate methods

Applying univariate outlier detection methods on the data of total annual expenditure is a straightforward and simple approach. One of the drawbacks with this approach is that the multidimensional structure of the data set is

Country	Number of households		Original	IQR	Box-Cox with IQR	robust Box-Cox with IQR	MAD	Box-Cox with MAD	robust Box-Cox with MAD	Boxplot	adjusted Boxplot	Pareto Modeling	
												replace	reweight
Albania(2008)	3600	Gini	31.9516	28.0961	31.3368	30.7868	27.6830	31.3132	30.7002	26.8415	30.0106	31.6639	31.7374
		Variance Gini	0.6745	0.1254	0.4987	0.3901	0.1105	0.4938	0.3773	0.0879	0.3126	0.6270	0.6224
		Upper bound	–	1585.5592	3619.7083	2758.3991	1502.0627	3557.0694	2663.8560	1365.37964	2342.039	1937.27	
		(outlier)	–	(121)	(12)	(25)	(141)	(13)	(27)	(202)	(38)	(1)	
		Lower bound	–	-392.1509	143.3289	113.6876	-308.6544	144.9035	118.4252	-68.46019	235.527	–	
		(outlier)	–	(0)	(13)	(4)	(0)	(16)	(5)	(0)	(157)	(0)	
India(2009)	100852	Gini	39.8225	33.5527	38.2812	38.8811	32.7065	38.2525	38.8784	31.8037	36.7449	39.4340	
		Variance Gini	0.0813	0.0138	0.0291	0.0357	0.0130	0.0289	0.0356	0.0126	0.0201	0.0495	
		Upper bound	–	158.2175	358.5651	488.8627	146.2504	354.9305	488.1119	135.4646	258.3157	223.2658	
		(outlier)	–	(9141)	(847)	(351)	(11153)	(878)	(353)	(13344)	(2288)	(32)	
		Lower bound	–	-52.9521	8.0248	9.9336	-40.9851	8.1038	9.9421	-17.6334	16.7187	–	
		(outlier)	–	(0)	(328)	(725)	(0)	(338)	(728)	(0)	(2666)	(0)	
Mexico(2010)	27655	Gini	44.1972	37.6122	43.3185	43.7307	36.5747	43.3136	43.7267	35.8657	41.6716	44.0815	44.0689
		Variance Gini	0.1291	0.0263	0.0948	0.1150	0.0232	0.0947	0.1147	0.0217	0.0594	0.1284	0.1273
		Upper bound	–	229.0388	709.5715	946.5002	207.5034	707.7214	943.6185	195.1592	443.0978	328.0539	
		(outlier)	–	(1669)	(104)	(48)	(2061)	(105)	(49)	(2342)	(379)	(2)	
		Lower bound	–	-94.7963	7.6987	9.2200	-73.2609	7.7156	9.2358	-39.6213	17.5928	–	
		(outlier)	–	(0)	(161)	(270)	(0)	(163)	(270)	(0)	(1359)	(0)	
Malawi(2010)	12096	Gini	48.5196	36.1303	47.2318	45.5714	35.2028	47.2416	45.5907	34.2410	41.3443	48.5196	48.5196
		Variance Gini	0.5361	0.0377	0.3581	0.2326	0.0336	0.3592	0.2336	0.0302	0.1020	0.5361	0.5361
		Upper bound	–	534.3448	3396.8879	1988.9154	492.1710	3410.3993	1998.6225	454.7135	998.1302	980.7857	
		(outlier)	–	(1004)	(41)	(105)	(1162)	(41)	(105)	(1319)	(392)	(0)	
		Lower bound	–	-207.4177	29.0747	24.4189	-165.2439	29.0385	24.3504	-83.0643	46.3913	–	
		(outlier)	–	(0)	(43)	(14)	(0)	(42)	(14)	(0)	(308)	(0)	
Tajikistan(2007)	4860	Gini	33.1176	28.5829	31.4903	32.7882	28.3309	31.5061	32.7741	27.2635	29.8577	32.2548	32.4164
		Variance Gini	0.3976	0.0814	0.1637	0.3634	0.0785	0.1648	0.3624	0.0698	0.1010	0.2548	0.2607
		Upper bound	–	27.5696	52.3478	141.5597	26.6701	52.6002	134.4467	23.6105	36.0510	33.2870	
		(outlier)	–	(244)	(51)	(2)	(260)	(51)	(2)	(379)	(108)	(8)	
		Lower bound	–	-6.4379	2.2667	3.1698	-5.5384	2.2566	3.2029	-1.0450	3.5938	–	
		(outlier)	–	(0)	(53)	(113)	(0)	(52)	(116)	(0)	(153)	(0)	

Table 5: Results of univariate outlier detection methods applied to the household expenditure data of various countries.

not taken into account. Due to that, households which have listed erroneous information for a specific good or service might be overlooked since their total annual expenditure does not appear extreme enough to be marked as potential outlier. On the other hand, wealthy households which have moderately high expenditures for a lot of goods or services might have rather extreme total annual expenditures and will be flagged as potential outliers. To address this problem while using univariate outlier detection methods one can apply these methods and adjust potential outliers, according to the used methods, on every expenditure category separately. For this application, the expenditures are divided into the ICP category code. With this approach, zeros will be left out, meaning that if a household has zero information (or no information) on expenditures in a category, they will be discarded for outlier detection on this category only. In order to calculate the Gini after potential outliers have been detected, the expenditures for the different categories will be summed up for each household. This approach has been used for the Albanian data set with the use of the ICP category code described in Section 6.1. The resulting Gini coefficients and statistics for potential outliers are displayed in Figure 14. The upper left panel shows the resulting Gini coefficients and corresponding 95% confidence intervals, and the upper right panel shows the variance of the Gini coefficient. The bottom panel shows the share of observations which were flagged as potential outliers in the whole data set. The color coding indicates for how many categories an observation was flagged as outlier for a specific outlier detection scheme.

The results displayed in Figure 14 indicate that methods which do not account for skewness in the data are not appropriate. The share of observations which have been flagged as potential outliers in at least one category is beyond 50%. Therefore it is questionable if these univariate methods, applied on each column of the data, produce reliable results at all. Also the adjusted boxplot declare a high amount of observations as potential outliers. The methods using the Box-Cox transformation and the Pareto Tail modeling produce more reasonable results. For these methods, the share of detected potential outliers is within the range of 10%, and for the Pareto modeling even far less. The high share of flagged observations results from the fact that by applying univariate methods in a column-wise manner, potential outliers in one column must not correspond to the same observation as potential outliers of another column. In a worst case scenario, this column-wise approach could result in flagged potential outliers in at least one cell of every observation. This indicates that a column-wise implementation of univariate

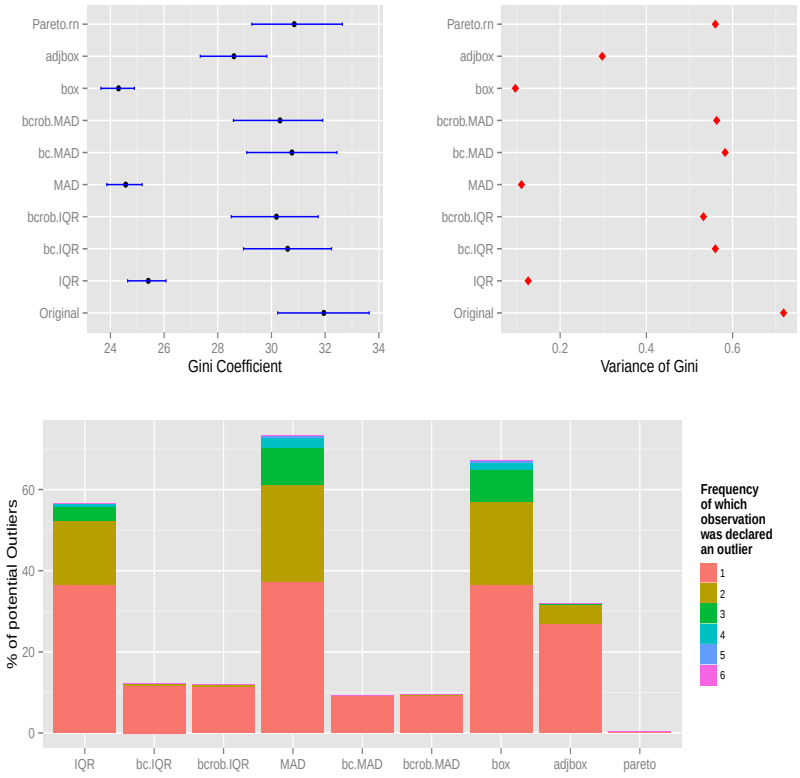


Figure 14: Top: Estimates for Gini coefficient (left) and Variance of Gini coefficient (right) of Albanian data set after column-wise outlier detection and adjustment for various outlier detection schemes.
Bottom: Share of potential outliers detected by column-wise application of different outlier detection methods onto the Albanian household data. Different colors indicate for how many categories an observation was declared as potential outlier.

methods on a data set, as it is presented in this work, has severe drawbacks.

It is interesting to see that although the outlier detection methods using the Box-Cox transformation detect far more potential outliers as the Pareto modeling, the resulting Gini coefficients after imputation of the detected outliers do not differ so much. This can possibly be explained by the fact that many potential outliers detected by the methods using Box-Cox transformation do not have a significant influence on the total annual expenditures of the corresponding individual. Therefore, the calculation of the Gini coefficient will not be heavily influenced by such potential outliers.

6.3 Multivariate Methods

The multivariate outlier detection methods, which will be applied on the household expenditure data sets from the World Bank, include mainly methods which calculate location and covariance of a multivariate data set in a robust way. In addition, the epidemic algorithm and the BACON-EEM algorithm will be tested. A description of all these algorithms can be found in Section 4. For the calculations of the methods which estimate location and covariance in a robust way, the **R**-package **rrcov** was used utilized, see Todorov and Filzmoser [2009]. As mentioned in Section 4, these methods can be tuned to achieve a certain breakdown point, and some methods have a tuning parameter for efficiency and small sample correction factors. For the following calculations, the default values for breakdown point and tuning parameters for efficiency or small sample correction which are implemented in the Rpackage **rrcov** were adopted and are listed in Table 6.

In the following, the multivariate outlier detection methods will be applied to the expenditure data of the Albanian data set.

6.3.1 Imputation to replace zeros

Before the methods can be applied, the data are transformed into matrix format by using the ICP category code, as discussed in Section 6.1. In this format, the data contain many zeros. Since the number of zero entries is well over 50% for some categories, the columns containing these expenditure categories were combined using the method mentioned in Section 6.1. This method compares the IQR of each expenditure column and adds two or more columns to reduce the number of zeros. This was done to ensure that the algorithms for the multivariate outlier detection methods are executable without

	breakdown point	efficiency	small sample correction
M-estimate	0.45	1	–
MM-estimate	0.5	0.95	–
S-estimate	0.5	can be quite inefficient	–
MCD	0.5	considerably increased by the use of weights	Yes
MVE	0.5	–	–
Stahel-Donoho	0.5	–	–
OGK-estimate	0.5	increased by reweighting	–
EA			
BAECON-EEM			

Table 6: Specifications for breakdown point, efficiency and small sample correction for the multivariate outlier detection methods.

error. This procedure does not necessarily reduce the number of zeros, but it may lead to less complications for the multivariate outlier detection methods.

Among the multivariate outlier detection methods considered here, only the epidemic algorithm and the BACON-EEM can deal with zeros (which are internally treated as missing entries). For all remaining multivariate outlier detection methods, too many zero values will cause an error in the execution, and thus the zero entries will be replaced by imputed values using the k-nearest neighbor algorithm.

K-nearest neighbor algorithm The k-nearest neighbor (kNN) algorithm is a classification algorithm which has been proven useful for imputation of multivariate normal data [Templ et al., 2012]. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a given data sample with p variables. Then the k-nearest neighbors of a new observation $\mathbf{x}_{n+1} \in \mathbb{R}^p$ are those $\{\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}\}$ for which the distances $d(\mathbf{x}_{n+1}, \mathbf{x}_{i_j})$, $j = 1, \ldots, k$, are smallest. This approach makes it possible to classify data points by determining their class label by the labels of its k-nearest neighbors. The kNN method used in this work is implemented in the `R`-package **VIM**, see Templ et al. [2012]. For defining the nearest neighbors, the distance computation is based on an extension of the Gower

distance [Gower, 1971]. This extension is able to handle distance variables of the type binary, categorical, ordered, continuous and semi-continuous. The distance between two observations $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^p$ is defined as

$$d(\mathbf{x}_j, \mathbf{x}_i) = \frac{\sum_{k=1}^{p} w_k \delta_{i,j,k}}{\sum_{k=1}^{p} w_k},$$

with w_k as weight which represents the importance of the k-th variable, and $\delta_{i,j,k}$ as the contribution of the k-th variable. This distance between two observation is therefore the weighted mean of the contributions of each variable.

Depending on the variable type, $\delta_{i,j,k}$ is defined differently. In this work the used variables are all continuous, for which $\delta_{i,j,k}$ is defined as

$$\delta_{i,j,k} = \frac{|\mathbf{x}_{i,k} - \mathbf{x}_{j,k}|}{r_k},$$

where $\mathbf{x}_{i,k}$ is the value of the k-th variable of the i-th observation and r_k is the range of the k-th variable. For the definition of other types of variables, see Templ et al. [2012].

Let an observation $\mathbf{x}_l \in \mathbb{R}^p$ have one or more zero cells. For the imputation with the kNN algorithm, the distances are only calculated between $\mathbf{x}_l$ and observations without zero cells, and for the calculations only variables which are not zero in $\mathbf{x}_l$ are considered. Finally, a zero cell of $\mathbf{x}_l$ is imputed by using the k values of the nearest neighbors of the zero variable. For continuous variables, the default option in Templ et al. [2012] is the median, although other statistics are possible.

After replacing the zero values by sensible values, the multivariate outlier detection procedures can be applied. The potential multivariate outliers are then projected onto the 95% tolerance ellipse, for which the covariance matrix was robustly estimated for the outlier detection. In the case of the BACON-EEM this is the final covariance estimate from the data for which zero observations have been imputed by the EEM algorithm. The epidemic algorithm does not compute a covariance during the task of detecting potential outliers. In order to make the epidemic algorithm comparable to other methods, the covariance matrix which is used for the imputation, will be

computed by applying the classical estimate for the covariance on the data without the potential outliers detected by the epidemic algorithm.

After the multivariate outlier detection methods have been applied and potential outliers have been adjusted, the Gini coefficient is estimated. As in the case of univariate methods, the Gini coefficient is estimated for the total annual expenditures per household, where the household sample weights from the data set are included.

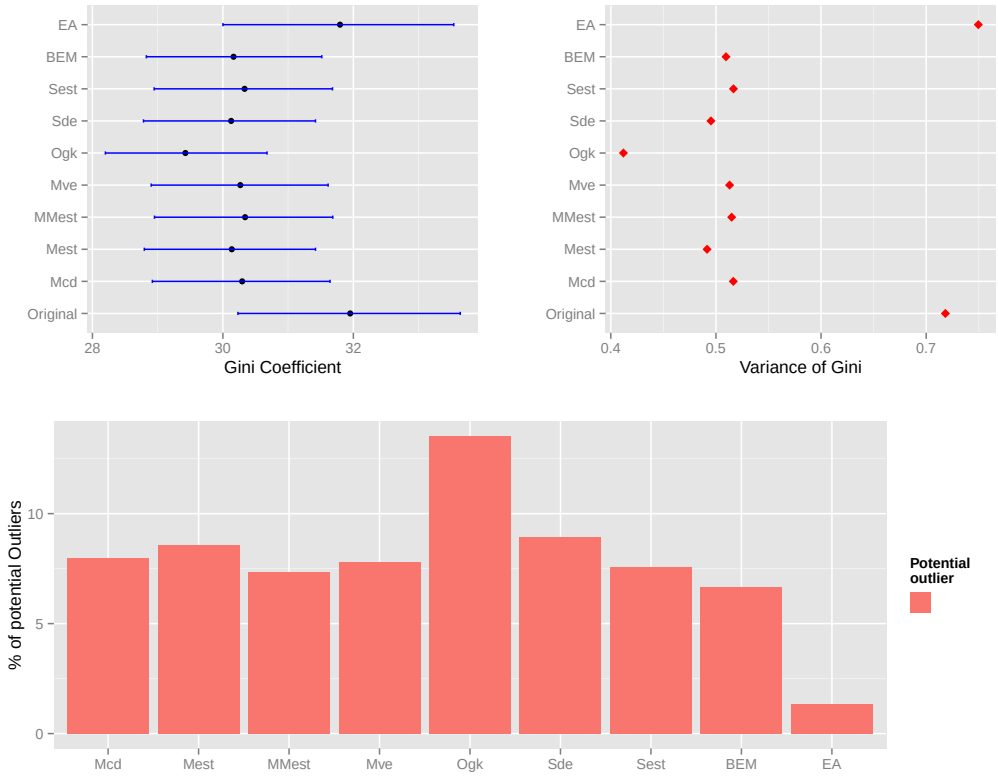


Figure 15: Top: Gini coefficients and variance of the Gini coefficients of the Albanian data set after outlier detection and adjustment.
Bottom: Fraction of identified potential outliers on all observations.

Figure 15 shows the results for the estimated Gini coefficients and share of detected potential outliers for each multivariate outlier detection method applied to the Albanian household expenditure data. In contrast to univariate outlier detection (Figure 13 and Figure 14) the results for the Gini coefficients

do not differ so much between the methods. Only the results for the OGK estimate and the epidemic algorithm differ slightly. Also the share of detected potential outliers does not differ a lot between the methods, except for the OGK estimate and the epidemic algorithm. Compared with the column-wise application of univariate outlier detection methods, the number of detected potential outliers is far lower for the multivariate outlier detection methods.

The results of the multivariate outlier detection methods for Albania and other countries are presented in Table 7, and they are quite similar for all used methods. Therefore it is not clear in general, which of the methods performs best. Even in comparison with the use of univariate methods it is not clear if univariate or multivariate outlier detection methods should be preferred.

To address this problem, a simulation study is conducted to see how the different methods perform on data which are generated based on the Albanian household expenditure data.

6.4 Simulation Study

The number and positions of the "real" outliers in the data sets are unknown. Moreover, there is no knowledge about a "true" Gini value. It is thus difficult to decide which of the estimators is "better" in the sense of "more reliable", and which has poorer behavior. It is clear from theoretical properties of the estimators that differences in performance are to be expected. Moreover, it is somehow clear that a multivariate consideration should be in general more suitable than a univariate approach. However, all this might be different when we are interested for example in a reliable value for the Gini coefficient.

In order to get deeper insight, a simulation study is conducted using the Albanian data set. The results of the study can be generalized also to other data sets.

6.4.1 Simulation setup

The basic ideas for a simulation are as follows.

1. Simulate such kind of data sets which are comparable, regarding the data on household expenditure, with the ones provided by the World Bank.

2. To know the number and position of "true" outliers beforehand.

Country	Number of households		Original	MCD	M estimate	MM estimate	MVE	OGK	Stahel Donoho	S estimate	BACON EEM	Epidemic algorithm
Albania(2008)	3600	Gini	31.9516	30.5302	30.4184	30.6088	30.5676	29.8527	30.3923	30.6061	30.4417	31.9291
		Variance Gini	0.7187	0.5625	0.5514	0.5592	0.5657	0.5179	0.5489	0.5608	0.5654	0.7209
		Number outlier	–	330	357	281	326	550	371	285	332	1
India(2009)	100852	Gini	39.8225	38.0052	37.8877	38.1325	38.0377	37.1938	37.8830	38.1269	37.4457	NA[a]
		Variance Gini	0.0813	0.0336	0.0324	0.0344	0.0342	0.0287	0.0323	0.0344	0.0299	NA
		Number outlier	–	9513	10077	8344	9208	15083	10195	8448	9404	NA
Mexico(2010)	27655	Gini	44.1972	42.8460	42.7508	42.9073	42.8918	42.0050	42.7447	42.8977	42.7515	44.1968
		Variance Gini	0.1716	0.1225	0.1195	0.1222	0.1244	0.0990	0.1197	0.1232	0.1233	0.1716
		Number outlier	–	2060	2215	1904	1935	3545	2236	1951	2429	1
Malawi(2010)	12096	Gini	48.5196	41.7416	41.4384	42.1464	41.7872	39.8168	41.4754	41.7956	41.5374	48.4159
		Variance Gini	0.5143	0.1095	0.1058	0.1171	0.1091	0.0838	0.1064	0.1110	0.1049	0.5243
		Number outlier	–	753	846	745	741	1400	857	761	588	4
Tajikistan(2007)	4860	Gini	33.1176	30.9804	30.8291	31.0579	30.9835	30.0309	30.7993	31.0544	30.9219	32.7758
		Variance Gini	0.3976	0.2005	0.1961	0.1995	0.2016	0.1745	0.1954	0.2012	0.2008	0.3027
		Number outlier	–	379	421	346	369	643	426	361	288	8

Table 7: Results of multivariate outlier detection methods applied to the household expenditure data of various countries.

[a]Results not available, due to RAM limitations.

This will allow for a concise performance evaluation of the different estimators.

Simulated data
We will generate data for which the distribution is based on the distribution of the expenditure data from the Albanian data set. First of all, the expenditure data from the Albanian data set, aggregated by category of products/services, are transformed using the logarithm since the data in each expenditure category are skewed to the right and many of the outlier detection schemes depend on underlying normal distribution or even multivariate normal distribution.

Simulation of "clean" data and outliers
After transformation, the data set is split into a "clean" data set which contains most likely no outliers and a "contaminated" data set which contains mostly outliers. This is based on the results of the univariate and multivariate outlier detection schemes applied on the household expenditure data of the Albanian data set, meaning that the clean data set contains only observations that have not been flagged by any of the outlier detection schemes as potential outliers. This resulting data set consists of 2752, out of 3600, most likely uncontaminated observations. The contaminated data set contains observations that were flagged as potential outliers by at least 5 univariate outlier detection methods or at least 6 multivariate outlier detection methods. The contaminated data set consists of 390 observations. From the contaminated and clean data sets, location and covariance are estimated in a classical way. Denote the corresponding estimates by $(\bar{\boldsymbol{x}}_{cl}, \boldsymbol{S}_{cl})$ for the clean data, and by $(\bar{\boldsymbol{x}}_{co}, \boldsymbol{S}_{co})$ for the contaminated data. These estimates are then used as a basis for the distribution of the simulated data sets. If the clean or contaminated data sets contain any zero values, they will be replaced by imputation with the k-nearest neighbors algorithm, so that the multivariate outlier detection algorithms can be applied.

The simulated data are then generated from a mixture distribution according to

$$\mathbf{X} \sim (1-\epsilon)MVN(\bar{\boldsymbol{x}}_{cl}, \boldsymbol{S}_{cl}) + \epsilon MVN(\bar{\boldsymbol{x}}_{co}, \boldsymbol{S}_{co}), \tag{30}$$

with $\epsilon \in (0, 1)$ determining the proportion of contaminated data points. Here, "MVN" denotes multivariate normal distribution. The number of observations of $\mathbf{X}$ is 3600, the same as in the original Albanian data set.

Practically, the number of contaminated observations corresponds to a chosen proportion ϵ, and the corresponding rows of $\mathbf{X}$ were picked randomly. It was also of interest to investigate the behavior if not the complete observation but only some cells of the observation were contaminated. Given the i^{th} observation for which only the j^{th} cell will be contaminated, the cell is replaced by y_{ij}, with

$$\mathbf{Y} \sim MVN(\bar{\boldsymbol{x}}_{co}, \boldsymbol{S}_{co}). \tag{31}$$

Inclusion of zero values

Another aspect for the data simulation is the inclusion of zero values, since the number of zero observations can be quite high and play quite a big role for the analysis on such data sets. The placement of these zero values is the same as observed in the original Albanian data set and they will be included after the data have been generated. By replacing values in the simulated data set with zeros it can occur that previously generated artificial outliers will be replaced by those values. Since the data are simulated many times and since the placement of zeros overlapping with the randomly chosen contamination is not very likely, it is expected to have not a large impact on the simulation study. Since sample weights also play quite a role for the presented outlier detection methods as well as for the Gini calculation, the simulated data sets receive the same sample weights as the household weights in the Albanian data set.

Application of univariate methods

Since the simulated data were generated by multivariate normal distribution based on log-transformation of the original Albanian data set, the simulated data will be transformed back with the exponential function to generate skewed data that have nearly the same distribution as the original data sets. Afterwards, the univariate outlier detection methods are applied on each column of the data set. In addition, univariate outlier detection methods will make use of the household sample weights provided by the Albanian data set. After the univariate outlier detection methods have been applied and potential outliers are flagged by each method, the potential outliers are adjusted. Note that for the univariate outlier detection methods the zero values are discarded for calculating and adjusting potential outliers. Note that the positions of the artificial outliers is known beforehand, since in one scenario the whole observation is contaminated, in another one only specific cells are

contaminated. For each outlier detection method, the number of successfully identified outliers is counted afterwards. Moreover, the number of falsely declared potential outliers is also counted. Adjusting the potential outliers for every outlier detection scheme creates new data sets corresponding to each of the univariate outlier detection schemes. For each of these data sets, the estimate for the weighted Gini coefficient of the total sum of each observation is calculated. In addition, the sample weights are used to calculate the weighted Gini coefficient. For comparison, the weighted Gini coefficient will also be calculated on the data without prior application of outlier detection methods. This is done to see the impact of the contamination mechanism as well as the outlier detection methods on the estimated values of the Gini coefficient. Furthermore, the estimated Gini coefficient for a generated data set without contamination and without applying any outlier detection methods presents a useful baseline estimate which is going to be used as reference value for the estimated Gini coefficients after outlier detection and adjustment have been applied on a contaminated data set.

Application of multivariate Methods

In the case of multivariate outlier detection, the simulated data are not transformed by the exponential function prior to outlier detection, simply because the presented multivariate outlier detection methods require data which are essentially multivariate normally distributed. In contrast to the univariate outlier detection methods, not all multivariate outlier detection methods can deal with a higher proportion of zero values in the data set. Only the epidemic algorithm and the BEM are able to handle zeros, by treating them internally as missing values. For the other multivariate outlier detection methods the zeros will be imputed prior to outlier detection by using the k-nearest-neighbor algorithm. The imputed zero values are only used for the outlier detection methods, and after outlier detection the imputed values are replaced again with the zeros. As for univariate outlier detection, the number of correctly identified artificial outliers are counted, and also the number of falsely declared potential outliers is counted.

As mentioned previously, applying the outlier detection methods and adjusting potential outliers generates new data sets for each of the used outlier detection algorithms. These data sets are transformed with the natural logarithm and the columns of the data sets are summed up to calculate the weighted Gini coefficient of the total accumulated observations. The data

are transformed prior to the Gini calculations since the resulting Gini is then more comparable to the case of univariate outlier detection methods as well as to the original data set of household expenditures. As in the univariate case, the Gini coefficient will also be calculated for generated data without applying any outlier of the multivariate outlier detection methods. This estimate shows the impact of the outlier on the estimated Gini coefficient. Furthermore, estimating the Gini coefficient without applying any outlier detection methods on a data set without contamination provides a reference value to better compare the estimated Gini coefficients after applying outlier detection methods and adjusting potential outliers on contaminated data sets.

6.4.2 Simulations results

For the following results, the procedures outlined above were repeated 50 times with different levels of ϵ, i.e. $\epsilon \in \{0, 0.01, 0.025, 0.05\}$. As discussed above, for a part of the contaminated data only one cell of each observation, and for the rest of the contaminated data, the whole observation is contaminated. Concretely, 1/3 of the contamination will be cell wise, and for 2/3 of the contaminated data the whole observation is contaminated, see contamination schemes in Equation (30) and (31).

Results from univariate methods
Figure 16 shows boxplots of the resulting Gini coefficients for each method and each level ϵ of contamination. The outlier detection methods are Pareto modeling (*pareto.rn*), adjusted boxplot (*adjbox*), boxplot (*boxplot*), robust Box-Cox with MAD (*bcrob.MAD*), Box-Cox with MAD (*bc.MAD*), median ± 3 MAD (*MAD*), robust Box-Cox with IQR (*bcrob.IQR*), Box-Cox with IQR (*bc.IQR*), median ± 3 IQR (*IQR*). In the case where no outlier detection method was applied (Original) the boxplot was only plotted for $\epsilon = 0$, since for other levels of ϵ the boxplots would cover a far greater range which makes the results for the applied outlier detection methods harder to read. From Figure 16 it is interesting to see that for higher values of ϵ the values for the Gini coefficient increase. This would seem strange since the outlier detection schemes are supposed to identify the potential outliers and in connection with outlier adjustment, the effect of the potential outlier should be negated. However, for the outlier imputation, except in the case of the Pareto modeling, the potential outliers are adjusted to the interval boundaries, whereas

these boundaries are calculated during the detection methods, adjusted outliers still have an influence on the Gini. Nevertheless, this trend is rather small and for outlier detection schemes which take into account skewness of the data, the resulting Gini is still quite close to the one with no contamination and no outlier detection scheme applied. There is a qualitative difference of the rules *box*, *MAD*, and *IQR*, which clearly underestimate the Gini.

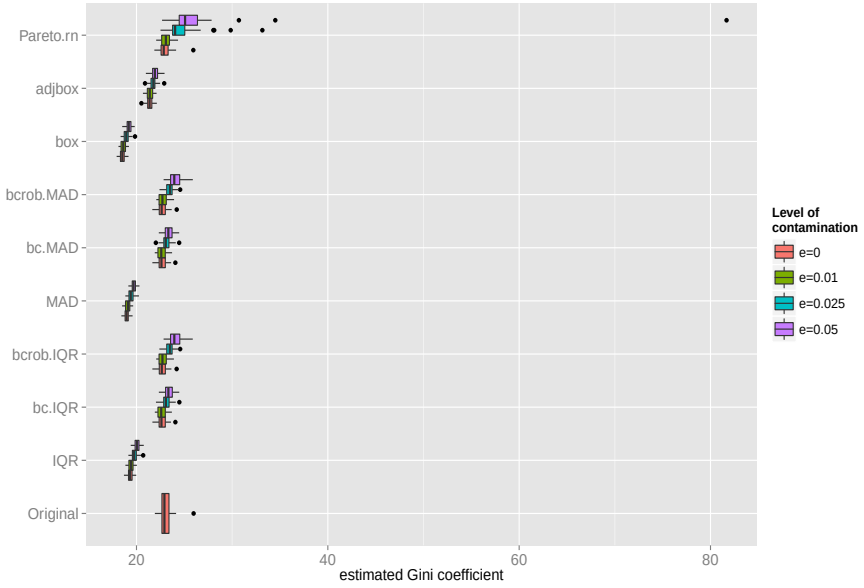


Figure 16: Boxplots of calculated Gini coefficients for different outlier detection methods and different levels of ϵ.

Apart from the resulting Gini coefficient it is of great interest to see how many artificial outliers have been successfully detected. Figure 17 shows the boxplots of the number of successfully detected artificial outliers, for which the whole observation was contaminated, for each outlier detection method and different levels of ϵ. The legend on the right side of the plot indicates how many artificial outliers were generated in total. The plot shows that the methods were not at all successful in detecting outliers and that the results got worse for higher values of ϵ. Furthermore, outlier detection schemes which used the Box-Cox transformation were not as successful at identifying outliers, for the scenario where the whole observation was contaminated, as their counterparts which do not use this transformation. In the case of contamination, where only a single cell was contaminated, the outlier detection

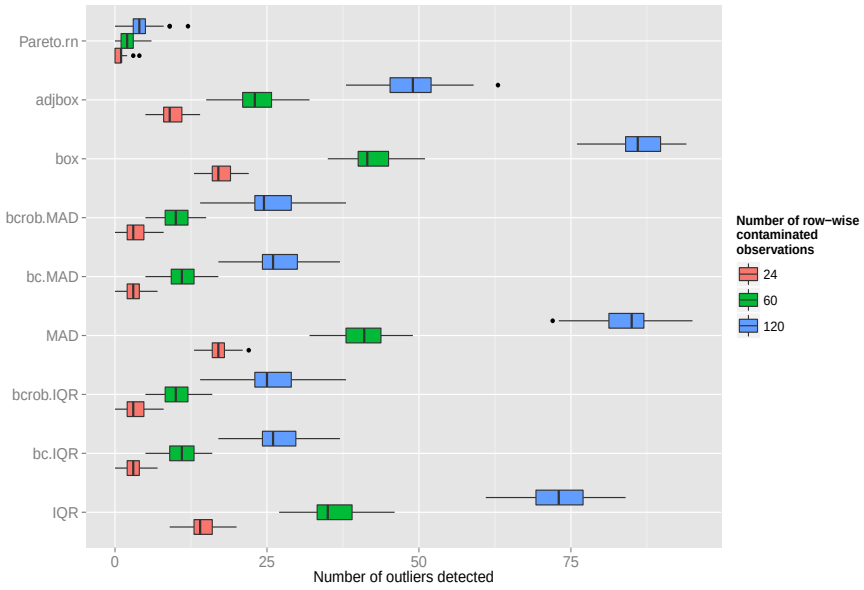


Figure 17: Boxplots for successfully detected artificial outliers, where the whole observation was contaminated, for different outlier detection methods and different levels of ϵ.

methods where not very successful as well. The results are plotted in Figure 18. In contrast to the row-wise contaminations the results for the cell-wise show not so drastic differences between the univariate measures. In the case of Pareto modeling, the algorithm seems to have not performed well. The reason for this could be the use of the Van Kerm's rule of thumb which determines after which point the Pareto distribution is fitted. This rule of thumb was a suggestion based on the EU-SILC data. It is reasonable to argue that this suggestion might not be suitable for this simulation and the results for the Pareto modeling are therefore not satisfactory. Regarding the successful detection of outliers it can be said that the boxplot, adjusted boxplot and the methods using IQR or MAD without Box-Cox transformation were able to identify comparatively more artificial outliers than the other detection methods.

Another interesting statistic is the number of flagged potential outliers which are not artificially contaminated observations. Figure 19 shows the corresponding boxplots for the different outlier detection methods and different levels of ϵ. The x-axis corresponds to the share of flagged outliers to

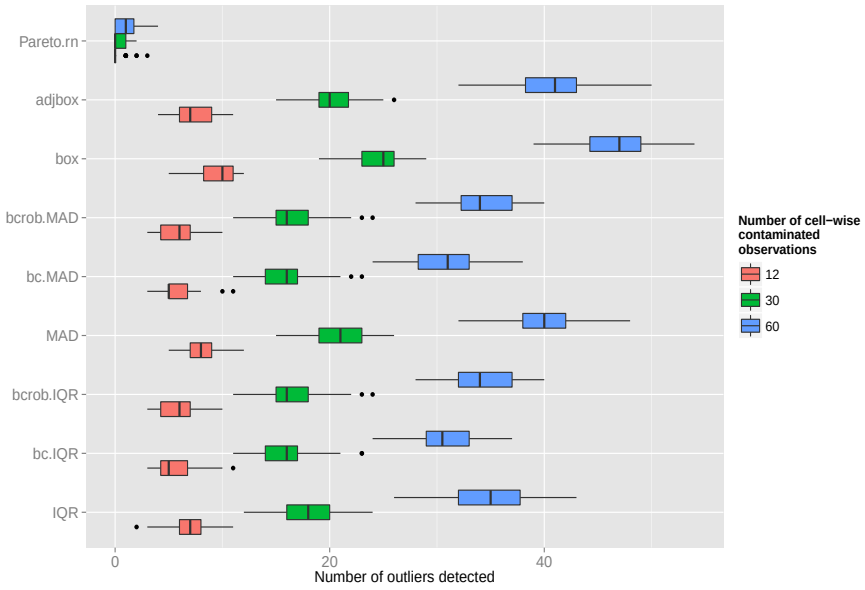


Figure 18: Boxplots for successfully detected artificial outliers, where only single cells were contaminated, for different outlier detection methods and different levels of ϵ.

the total amount of clean data in the simulated data set. Except for Pareto modeling or methods which incorporate the use of the Box-Cox transformation, the number of falsely flagged outliers is high. One could even argue that the numbers for methods which use the Box-Cox transformation are too high. The high amount of falsely flagged outliers in the case of the boxplot, adjusted boxplot and the methods using IQR or MAD without Box-Cox transformation put in perspective their relatively better performance regarding the ability to correctly identify artificial outliers. Although these methods were able to detect comparatively more outliers, the high number of falsely flagged potential outliers suggests that these methods are not precise for outlier detection. From the results shown above, the use of univariate outlier detection schemes, or at least the column-wise use of those methods does not seem appropriate for this kind of data. The ability to detect artificial outliers was not convincing and the number of falsely flagged potential outliers was far too high in almost all cases. Note that by adjusting the parameters for univariate outlier detection methods (like the constant for mean/spread rules), one could reduce the number of false outliers, but this goes hand in

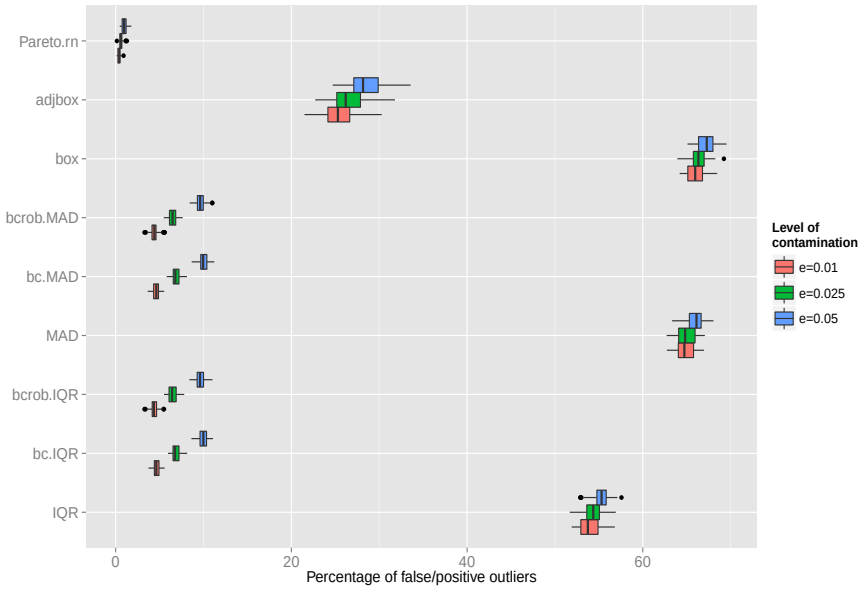


Figure 19: Boxplots for share of false/positive outliers to number of clean data points for different outlier detection methods and different levels of ϵ.

hand with a poorer performance of correctly identifying outliers.

Results from multivariate methods

Similar to the univariate outlier detection methods, Figure 20 shows the boxplots of the Gini values for the different outlier detection methods and different levels of ϵ. The methods shown are epidemic algorithm (*EA*), Bacon EEM algorithm (*BEM*), S-estimators (*Sest*), Stahel-Donoho estimator (*Sde*), OGK estimator (*OGK*), MVE estimator (*Mve*), MM-estimator (*MMest*), M-estimator (*Mest*), MCD estimator (*Mcd*), and the Gini for the uncontaminated data (*Original*). It is immediate that the epidemic algorithm performs rather poor. This can be explained by the fact that this algorithm needs quite a lot of tuning for parameter calibration until it is really applicable to a problem. We used the default parameter setting in our simulation study. Under those circumstances the algorithm is not bad per se but it is not very versatile without meaningful calibration which differs depending on the underlying data. For the other outlier detection schemes one can see, as in the case of univariate outliers, an increase in the Gini for rising levels of contamination. As it was argued for the univariate case, this is caused by the

imputation, which does not perfectly replace an outlier by projecting it to the 95% tolerance ellipse. Thus a rising number of outliers leads to a rising number of observations lying on the boundary of the 95% tolerance ellipse. The data points of the resulting data set are therefore wider spread from the center of the data than the data points in the uncontaminated data set. This difference in the distribution of the data can finally be seen in values of the Gini coefficients for different levels of ϵ. Apart from that, the results for the multivariate outlier detection methods do, even for higher levels of ϵ, not differ too much from the case where the data were not contaminated and no outlier method was applied.

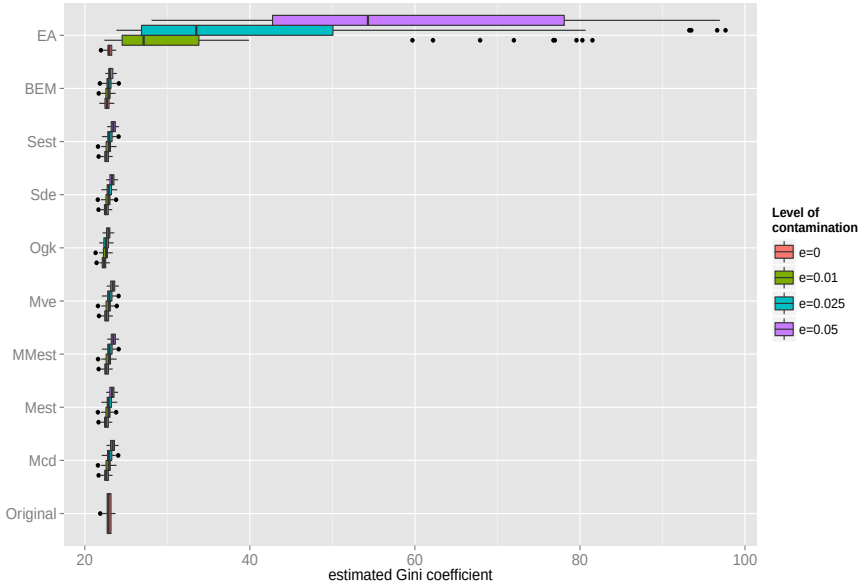


Figure 20: Boxplots of calculated Gini coefficients for different outlier detection methods and different levels of ϵ.

Figures 21 and 22 show the numbers of correctly identified artificial outliers by the multivariate outlier detection methods. Figure 21 corresponds to artificial outliers for which the whole observation was contaminated and Figure 22 corresponds to those where only one cell was contaminated. Both plots show, apart from the epidemic algorithm, that the multivariate outlier detection methods where much more successful than the univariate methods. In many cases the algorithms were able to detect every artificial outlier and even for rising values of ϵ the numbers are still high. The epidemic algorithm

did not perform too well, but as stated earlier this is due to poor calibration of the parameters, which is in practice a cumbersome task.

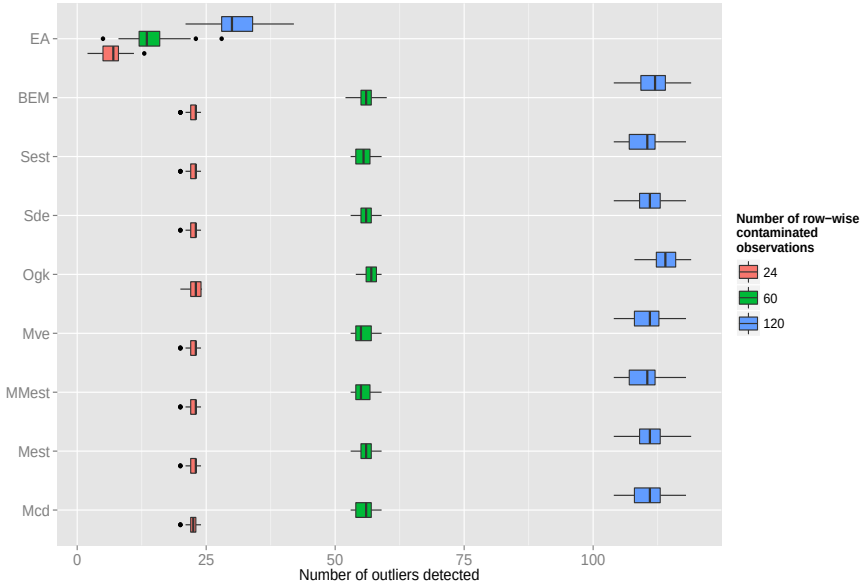


Figure 21: Boxplots for successfully detected artificial outliers, where whole observation was contaminated, for different outlier detection methods and different levels of ϵ.

For the case of falsely flagged potential outliers, Figure 23 shows the resulting boxplots for different outlier detection methods and different levels of contamination. Compared with the univariate case, the numbers of falsely flagged potential outliers are generally lower. The OGK estimate seems to perform not so well and has a far higher number of falsely flagged potential outliers than the other methods. The majority of the multivariate outlier detection schemes have for $\epsilon = 0.01$ roughly the same amount of falsely flagged potential outliers. Increasing levels of ϵ generally lead to the smaller amounts of falsely flagged potential outliers. Overall, except for the epidemic algorithm that was ruled out as valid method beforehand, the BEM delivers the smallest amount of falsely flagged potential outliers. This and the fact that the BEM was also successful in identifying artificial outliers, leads to the conclusion that the BEM performed best in this simulation study.

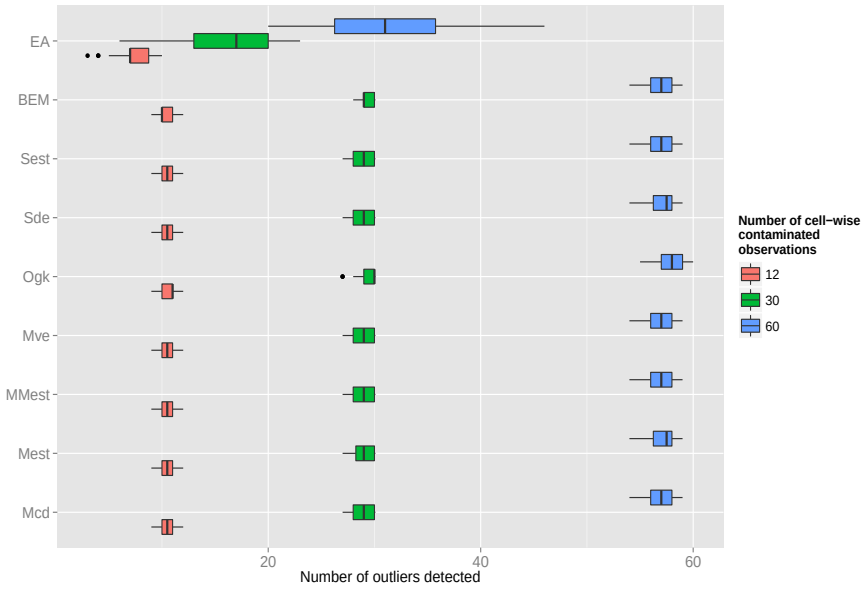


Figure 22: Boxplots of successfully detected artificial outliers, where only single cells were contaminated, for different outlier detection methods and different levels of ϵ.

7 Suggestions and recommendations

This report focuses on outlier detection in household expenditures, and on the effect on the Gini index after adjusting the outliers. For both univariate and multivariate outlier detection it is essential how the categories of the household expenditures are grouped. Sections 6.1.2 and 6.1.4 provide details about these issues.

Note that outlier detection for per capita expenditures was not considered here. Although technically the procedures could be used in the same manner, it is not so clear content-wise how to proceed. The simplest possibility, dividing the household expenditures by the household size, would lead to a bias since the age structure might be an important factor that needs to be taken into account. For example, expenditures for education or health care will strongly depend on age.

The outlined methods for univariate and multivariate outlier detection can be considered as the set of possible methods available in the statistics literature. The methods differ in their need for preprocessing (e.g. imputa-

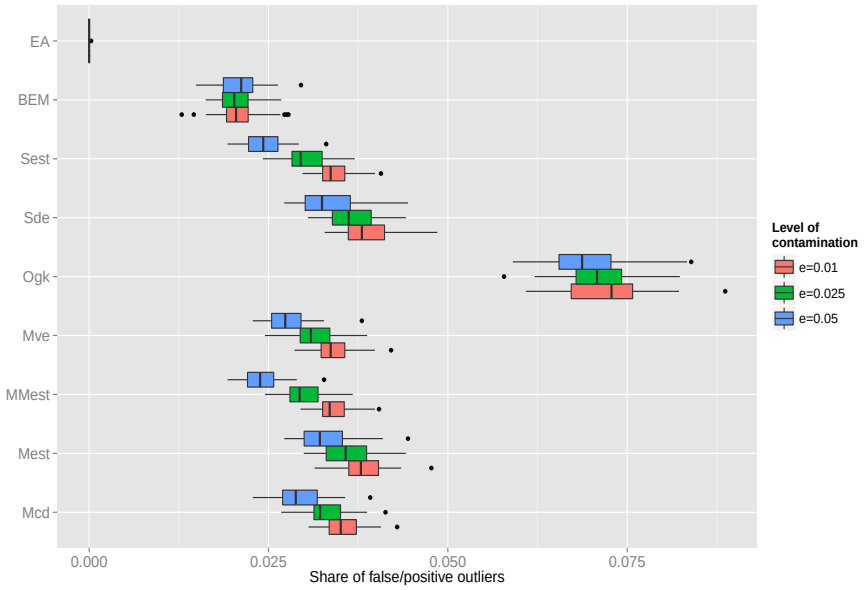


Figure 23: Boxplots for share of false/positive outliers to number of clean data points for different outlier detection methods and different levels of ϵ.

tion of missing values), distributional assumptions, but also the sensitivity and specificity to identify outliers. The conclusions made in this section are mainly based on the simulation study conducted in Section 6.4, where a setting was used to include realistic outliers following the patterns of the data structure.

Univariate outlier detection: There are several issues, like data transformation, different estimators being available to define the outlier identification rules, etc. Independent from the rule used, the univariate methods in general are unable to identify all the simulated outliers, and, depending on the method used, they flag many regular observations as outliers. For the latter case, the (robust) Box-Cox transformed values, together with a rule median plus/minus 3 times MAD or IQR perform best. This can also be seen in the resulting estimated Gini, where these methods are close to the Gini values for the "uncontaminated" original data.

So, the recommendation for univariate outlier detection is:

- Transform the data to approach normality, using the Box-Cox or a robust Box-Cox transformation;

- Use the median as a robust location estimator, and the IQR or the MAD (consistency corrected) as robust scale estimator;
- Use the rule: robust location $\pm 3\times$ robust scale.

Note that one could try to optimize the value "3" in the above rule. Although this choice is somehow a standard choice in statistical practice, a different choice it might lead to an improvement for outlier detection in this context. An evaluation could be done in a similar way as it has been done in the above simulation study.

Multivariate outlier detection: Rules based on multivariate outlier detection have the advantage over univariate rules that the correlation structure among the variables is considered. On the other hand, a multivariate outlier rule flags the whole observation as an outlier, even if some cells are non-outlying. Therefore, in the simulations also cell-wise contamination was considered in order to get an idea about this effect.

Overall, the methods were more precise than univariate rules, in terms of correctly identifying outliers, and in terms of smaller amounts of incorrectly flagged outliers. Out of the tested methods, only the Epidemic Algorithm (EA) delivered poor results, which was mainly based on suboptimal tuning parameters. One could play with these parameters, but the danger is that they need to be adjusted for every new data set. The OGK estimator gave slightly worse results, but all other methods are very much comparable.

Note that for all multivariate methods, the replacement of the outliers as suggested is important. For the multivariate methods also the choice of the granularity for the categories is important. The more categories are used, the more samples are required for gaining stability of the multivariate estimates. On top of that, more granularity for the categories usually leads to more zero values, which is another problem for most of the methods considered here.

References

A. Alfons and M. Templ. Estimation of social exclusion indicators from complex surveys: The r package laeken. *Journal of Statistical Software*, 54 (15):1–25, 2013.

A. Alfons, M. Templ, and P. Filzmoser. Robust estimation of economic indicators from survey samples based on pareto tail modelling. *Journal of*

the Royal Statistical Society: Series C (Applied Statistics), 62(2):271–286, 2013.

F.A. Alqallaf, K. Konis, P. Martin, R. Douglas, and R.H. Zamar. Scalable robust covariance and correlation estimates for data mining. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 14–23, New York, NY, USA, 2002. ACM.

C. Béguin and B. Hulliger. Multivariate outlier detection in incomplete survey data: the epidemic algorithm and transformed rank correlations. *JRSS-B*, 127(2):275–294, 2004.

C. Beguin and B. Hulliger. The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data. *Survey Methodology*, 34(1): 91–103, 2008.

N. Billor, A. S. Hadi, and P. F. Vellemann. Bacon: Blocked adaptative computationally-efficient outlier nominators. *CSDA*, 34(3):279–298, 2000.

G.E.P Box and D.R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 1964.

G. Brys, M. Hubert, and A. Struyf. A robust measure of skewness. *Journal of Computational and Graphical Statistics*, 13(4):996–1017, 2004.

R.W. Butler, P.L. Davies, and M. Jhun. Asympthotic for the minimum covariance determinant estimator. *The Annals of Statistics*, 21:1385–1401, 1993.

A. Cerioli, M. Riani, and A.C. Atkinson. Controlling the size of multivariate outlier tests with the MCD estimator of scatter. *Statistics and Computing*, 19(3):341–353, 2009.

R. Chambers, A. Hentges, and X. Zhao. Robust automatic methods for outlier and error detection. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 167(2):323–339, 2004.

C. Croux and G. Haesbroeck. Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*, 71:161–190, 1999.

P.L. Davies. Asymptotic behaviour of S-estimators of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 15:1269–1292, 1987.

T. De Waal. Statistical data editing. In D. Peffermann and C.R. Rao, editors, *Handbook of Statistics 29A. Sample Surveys: Design, Methods and Applications*, pages 187–214. Elsevier B. V., Amsterdam, The Netherlands, 2009.

S. J. Devlin, R. Gnanadesikan, and J. R. Kettenring. Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, 76:354–362, 1981.

D. L. Donoho. Breakdown properties of multivariate location estimators. Technical report, Harvard University, Boston, 1982. URL `http://www-stat.stanford.edu/ donoho/Reports/Oldies/BPMLE.pdf`.

O. Dupriez. Building a household consumption database for the calculation of poverty ppps. Technical report, world bank, Washington, 2007. Technical Note, Draft 1.0.

D. Dupuis and M.-P. Victoria-Feser. A robust prediction error criterion for Pareto modelling of upper tails. *The Canadian Journal of Statistics*, 34 (4):639–658, 2006.

I.P. Fellegi and D. Holt. A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association*, 71(353):17–35, 1976.

P. Filzmoser, R.G. Garrett, and C. Reimann. Multivariate outlier detection in exploration geochemistry. *Computers & Geosciences*, 31:579–587, 2005.

P. Filzmoser, R. Maronna, and M. Werner. Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 52(3):1694–1711, 2008.

R. Gnanadesikan and J. R. Kettenring. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28:81–124, 1972.

J.C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 1971.

L. Granquist. A review of some macro-editing methods for rationalizing the editing process. In *Proceedings of the Statistics Canada Symposium*, pages 225–234. Ottawa, Canada, 1990.

J. Hardin and D. M. Rocke. The distribution of robust distances. *Journal of Computational and Graphical Statistics*, 14:910–927, 2005.

B. Hulliger. Multivariate outlier detection and treatment in business surveys. In *Proceedings of the III International Conference on Establishment Surveys*, pages 1283–1289, Montréal, Canada, 2007. Statistics Canada.

B. Hulliger, A. Alfons, P. Filzmoser, A. Meraner, T. Schoch, and M. Templ. Robust methodology for laeken indicators. Research Project Report WP4 – D4.2, FP7-SSH-2007-217322 AMELI, 2011. URL `http://ameli.surveystatistics.net`.

C. Kleiber and S. Kotz. *Statistical Size Distributions in Economics and Actuarial Sciences*. John Wiley and Sons, 2003. ISBN 0-471-15064-9.

H. P. Lopuhaä. On the relation between S-estimators and M-estimators of multivariate location and covariance. *The Annals of Statistics*, 17:1662–1683, 1989.

R. A. Maronna. Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 1:51–67, 1976.

R. A. Maronna and V. J. Yohai. The behavoiur of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90:330–341, 1995.

R. A. Maronna and R. H. Zamar. Robust estimation of location and dispersion for high-dimensional datasets. *Technometrics*, 44:307–317, 2002.

R. A. Maronna, D. Martin, and V. Yohai. *Robust Statistics: Theory and Methods*. John Wiley & Sons, New York, 2006.

G. Pison, S. Van Aelst, and G. Willems. Small sample corrections for LTS and MCD. *Metrika*, 55:111–123, 2002.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

M. Riani, A.C. Atkinson, and A. Cerioli. Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):447–466, 2009.

D. M. Rocke. Robustness properties of S-estimators of multivariate location and shape in high dimension. *The Annals of Statistics*, 24:1327–1345, 1996.

P. J. Rousseeuw. Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, editors, *Mathematical Statistics and Applications Vol. B*, pages 283–297. Reidel Publishing, Dordrecht, 1985.

P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, New York, 1987.

P. J. Rousseeuw and K Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223, 1999.

D. Ruppert. Computing S-estimators for regression and multivariate location/dispersion. *Journal of Computational and Graphical Statistics*, 1:253–270, 1992.

M. Salibian-Barrera and V. J. Yohai. A fast algorithm for S-regression estimates. *Journal of Computational and Graphical Statistics*, 15:414–427, 2006.

W. A. Stahel. Robuste schätzungen: Infinitesimale optimalität und schätzungen von kovarianzmatrizen. Ph.d. thesis no. 6881, Swiss Federal Institute of Technology (ETH), Zürich, 1981a. URL `http://e-collection.ethbib.ethz.ch/view/eth:21890`.

W. A. Stahel. Breakdown of covariance estimators. Research Report 31, ETH Zurich, 1981b. Fachgruppe für Statistik.

K.S. Tatsuoka and D.E. Tyler. The uniqueness of S and M-functionals under nonelliptical distributions. *The Annals of Statistics*, 28:1219–1243, 2000.

M. Templ, A. Alfons, and P. Filzmoser. Exploring incomplete data using visualization techniques. *Advances in Data Analyis and Classification*, 2012. URL `http://dx.doi.org/10.1007/s11634-011-0102-y`. DOI 10.1007/s11634-011-0102-y, to appear.

G. Terrell. Linear density estimates. *Proceedings of the Statistical Computing Section*, pages 297–302, 1990. American Statistical Association.

The World Bank Group. Purchasing power parities and the real size of world economies. a comprehensive report of the 2011 international comparison program. Technical report, International Bank for Reconstruction and Development / The World Bank, Washington DC, 2015. URL `http://siteresources.worldbank.org/ICPEXT/Resources/ICP-2011-report.pdf`.

T. Todorov and P. Filzmoser. An object oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32(3):1–47, 2009. URL `http://www.jstatsoft.org/v32/i03/`.

V. Todorov. A note on the MCD consistency and small sample correction factors, 2008. Unpublished manuscript, in preparation.

V. Todorov, M. Templ, and P. Filzmoser. Detection of multivariate outliers in business survey data with incomplete information. *Advances in Data Analysis and Classification*, 5(1):37–56, 2011.

P. Van Kerm. Extreme incomes and the estimation of poverty and inequality indicators from EU-SILC. *IRISS Working Paper Series*, 2007-01, 2007. CEPS/INSTEAD.

E. Vandervieren and M. Hubert. An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis*, 52(12):5186–5201, 2008.

B. Vandewalle, J. Beirlant, A. Christmann, and M. Hubert. A robust estimator for the tail index of Pareto-type distributions. *Computational Statistics and Data Analysis*, 51(12):6252–6268, 2007.

Victor J. Yohai. High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15:642–656, 1987.