Robust Statistical Methods for Outlier Detection with Application to Household Expenditure Data

Ausgeführt am
Institut für Stochastik und Wirtschaftsmathematik
der Technischen Universität Wien

unter der Anleitung von
Privatdoz. Dipl.-Ing. Dr.techn. Matthias Templ
und
Ao. Univ.-Prof. Dipl.-Ing. Dr.tech. Peter Filzmoser

durch
Johannes GUSSENBAUER
Schleifmühlgasse 2/19
1040 Wien

Wien, 26 August 2015

Unterschrift
Acknowledgment

First and foremost I would like to thank my supervisors Matthias Templ and Peter Filzmoser for giving me the opportunity to write this master thesis. Their support during the process of writing this master thesis as well as their guidance when I had questions was more than I could have asked for.
I would also like to thank my family and friends for all their encouragement and aid during my academic studies. Without their support my academic career would not have been possible.
This work has been funded by the World Bank under the project "Improving the quality of sample household expenditure data and the reliability of poverty and inequality measures", selection number 1157976.

Johannes Gussenbauer
Abstract

Outlier detection can be seen as a pre-processing step for locating data points in a data sample, which do not conform with the rest of the data. Various techniques and methods for outlier detection can be found in the literature dealing with different data types. In this master thesis the data sets used for outlier detection methods are household expenditure data from five countries. Based on classical estimates of the Gini coefficient these data sets are suspected to contain outlier. In order to detect data points that deviate from the rest of the data, one- and multi-dimensional outlier detection methods are applied on the household expenditure data. The outlier detection methods are based on robust estimates and incorporate, in some cases, the use of sample weights. Important issues concerning the data and outlier detection methods are the number of missing values in each data set as well as the position of true outliers, which is completely unknown. The main focus of this thesis lies in the understanding of the outlier detection methods and their influence of the estimated Gini coefficient. Apart from applying the outlier detection methods on the various data sets and presenting the results, a recommendation on which of the outlier detection methods should be preferred when it comes to outlier detection on household expenditure data is presented in this work. In order to give a recommendation for outlier detection methods it is important to get a clearer vision of the performance of each outlier detection method on household expenditure data. To help understand the performance of the different outlier detection methods a simulation study, based on the original data from the survey, was conducted. The simulation study and all other calculations where executed using the R-programming language.
## 1 Mathematical background

1.1 Robust properties ........................................... 1  
1.2 Statistic efficiency ........................................... 2  
1.3 Affine equivariance ........................................... 4  

## 2 Methods for one-dimensional data  

2.1 Robust location ± constant * robust scale .................. 5  
2.1.1 Robust location ........................................... 5  
2.1.2 Robust scale ............................................. 6  
2.1.3 Detecting potential outliers ................................ 7  
2.1.4 Box-Cox transformation ................................... 9  
2.2 Boxplot ...................................................... 11  
2.3 Adjusted boxplot ............................................. 11  
2.4 Pareto tail modeling .......................................... 13  
2.4.1 The Pareto distribution ................................... 13  
2.4.2 Van Kerm’s rule of thumb ................................ 14  
2.4.3 Partial density component estimator ..................... 14  
2.5 Adjusting potential outliers .................................. 15  
2.5.1 Adjusting outlier for univariate outlier detection methods ... 15  
2.5.2 Adjusting outliers for semi-parametric modeling methods ... 15  

## 3 Methods for multivariate data  

3.1 M-estimates .................................................. 18  
3.2 S-estimate and MM-estimate ................................ 20  
3.3 Minimum covariance determinant (MCD) estimator .......... 24  
3.4 Minimum volume ellipsoid (MVE) estimator ............... 25  
3.5 The Stahel-Donoho estimator ................................ 26  
3.6 Orthogonalized Gnanadesikan–Kettenring estimator .......... 28  
3.7 Epidemic algorithm .......................................... 30  
3.8 BACON-EEM .................................................. 33  
3.8.1 BACON algorithm ........................................ 33  
3.8.2 EEM algorithm ........................................... 33  
3.9 Imputation of potential outliers .............................. 35  

## 4 Measures of inequality  

4.1 Lorenz curve and Gini coefficient ............................ 38  

## 5 Empirical analysis  

5.1 Provided data and data structure ............................ 42  
5.1.1 Harmonization of the data ................................ 43  
5.1.2 Categories and missing values ............................ 44  
5.1.3 Data structure ............................................ 45
5.1.4 Selected data and further data preparation 45
5.2 Univariate methods 46
5.2.1 Results of univariate outlier detection methods for different countries 50
5.2.2 Column-wise application of univariate methods 52
5.3 Multivariate methods 54
5.3.1 Imputation of missing values 54
5.3.2 Applying multivariate outlier detection methods on the Albanian data set 56
5.3.3 Results of multivariate outlier detection methods for different countries 57

6 Simulation study 59
6.1 Simulation setup 59
6.1.1 Simulated data 60
6.1.2 Simulation of “clean” data and outliers 60
6.1.3 Inclusion of missing values 61
6.1.4 Application of univariate Methods 61
6.1.5 Application of multivariate Methods 62
6.2 Simulations results 63
6.2.1 Results from univariate methods 63
6.2.2 Results from multivariate methods 66

7 Conclusion 70
Outr"enle This master thesis is structured into the following Sections. In Section 1 a short summary of statistical properties, which play an important role for outlier detection is presented. This is followed by Section 2 and Section 3 where univariate and multivariate outlier detection methods are described. Section 4 introduces the Gini coefficient, a measure of inequality, which is used especially often for income or expenditure data. In Section 5 the application and results of the outlier detection methods used on household expenditure data of five different countries is discussed. To get a better understanding of the performance of the outlier detection methods, a simulation study was conducted. The simulation design and the results are presented in Section 6. Finally a short summary and conclusion of this thesis is presented in Section 7.

1 Mathematical background

In the literature the term 'outlier' is not defined uniformly nor are the definitions for an outlier connected to one mathematical formula. Nevertheless, it can be said that an outlier is a data point which deviates from the rest of the data in such a way that it seems to be the product of a different data generating mechanism. This circumscription for outliers does not distinguish between data points with extreme erroneous values (measurement errors, nonrepresentative outliers), and data points which are true values but are too distant from the rest of the data so that they will have huge influences on classical estimators (representative outliers). In practice it is not always possible to distinguish if a data point is product of measurement errors or a genuine, but very extreme observation. Therefore, methods for outlier detection can not truly detect outliers, but detect data points which have the potential of being an outlier, so called potential outliers.

The following section introduces, in a brief manner, some basic statistical properties, which play an important role for outlier detection methods, based on robust estimates.

1.1 Robust properties

Useful in terms of outlier and outlier detection are robust estimators. Such estimators are characterized by not being strongly influenced by outliers and are therefore able to cope with such extreme values. The sensitivity of an outlier on an estimator $T$ can be expressed by the so called influence function (IF).

Definition 1. The influence function (IF) of an estimator $T$ for a model distribution $F$ is defined by

$$IF(x; T, F) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon \Delta x) - T(F)}{\epsilon}.$$  (1)

The $IF(x, T, F)$ specifies the influence of an estimator $T$ with model distribution $F$, for an infinitesimal contamination at point $x$. 

1
For estimation methods, including various outlier detection methods, one prefers estimators $T$ with bounded influence function. A further characterization in context with the IF is the so called "gross-error-sensitivity", defined by

$$\gamma^*(T, F) = \sup_x |IF(x; T, F)|.$$  

It presents the largest influence on an estimator $T$ by infinitesimal contamination. Estimators with bounded gross-error-sensitivity are called B(ias)-robust.

A third characteristic to evaluate the robustness of an estimator $T$, is the "breakdown point" (BP). Contrary to the IF, which describes the influence on the estimator $T$ by small amounts of contamination, the breakdown point specifies the minimal amount of contamination for which the estimator is no longer able to produce a useful estimation value.

**Definition 2.** Given a data sample $x_1, \ldots, x_n$, $m$ data points $x_{1}, \ldots, x_m$ are replaced with arbitrary values $y_1, \ldots, y_m$ to create a new sample $z_1, \ldots, z_n$. The sample breakdown point ($\epsilon^*_n$) of an estimate $T$ is then given by

$$\epsilon^*_n(T; x_1, \ldots, x_n) = \min \left\{ \frac{m}{n} \max \sup_{i_1, \ldots, i_m, y_1, \ldots, y_m} |T(z_1, \ldots, z_n)| = \infty \right\}. \quad (2)$$

In other words, the sample breakdown point of an estimator $T$ is the minimum share of data points which can be manipulated in such a way, that the estimator $T$ can overtake any boundary $c \in \mathbb{R}$.

For an infinite data sample the breakdown point complies (asymptotically) in general with

$$\epsilon^* = \lim_{n \to \infty} \epsilon^*_n.$$  

The maximal achievable breakdown point is 50%, since for a value higher than 50% the bigger share of the outliers could be as considered "genuine" data.

### 1.2 Statistic efficiency

Although robust estimates are more reliable in case of contaminated, or extreme data points in a data set they do have a drawback when it comes to the quality of the estimate. Contrary to their classical counterpart, robust estimates are in general not very efficient. Statistical efficiency is an important characteristic when it comes to the precision of an estimator. It is defined via the so called Cramér-Rao bound or Cramér-Rao inequality which is a lower bound for the variance of an estimator.

For the following considerations let $x_1, \ldots, x_n$ be a sample from a distribution $X$, for which the density $f(x; \theta)$ exists with unknown parameter $\theta \in \Theta$.

Before defining the Cramér-Rao inequality consider the following regularity assumptions
1. (R1) The true parameter $\theta_0$ is an inner point of $\Theta$ and from $\theta \neq \theta'$ follows $f(x; \theta) \neq f(x; \theta')$.

2. (R2) The distributions $f(x_i; \theta)$, $i = 1, \ldots, n$, have common support for all $\theta \in \Theta$.

3. (R3) $f(x; \theta)$ is twice differentiable in $\theta$ and for $k = 1, 2$ holds

$$\frac{\partial^k}{\partial \theta^k} \int_{-\infty}^{\infty} f(x; \theta) dx = \int_{-\infty}^{\infty} \frac{\partial^k f(x; \theta)}{\partial \theta^k} dx.$$ 

Thus the operations of integration with respect to $x$ and differentiation with respect to $\theta$ can be permuted.

The regularity assumptions above are necessary for the Cramér-Rao inequality and are not only sufficient for the existence of the Fisher information $I(\theta)$, defined by

$$I(\theta) = \mathbb{E} \left[ \left( \frac{\partial \ln f(X; \theta)}{\partial \theta} \right)^2 \right].$$

Now the Cramér-Rao inequality can be introduced as follows:

**Definition 3.** Consider a sample $x_1, \ldots, x_n$ from a distribution $X$, for which the density $f(x; \theta)$ exists with unknown parameter $\theta \in \Theta$. Let $T = T(x_1, \ldots, x_n)$ be an estimate for the parameter $\theta$ with $\mathbb{E}[T(x_1, \ldots, x_n)] = k(\theta)$. Given the regularity assumptions (R1)–(R3) the following inequality holds true

$$\mathbb{V}[T] \leq \frac{[k'(\theta)]^2}{nI(\theta)}.$$

For an unbiased estimate $T$ for $\theta$, meaning that $\mathbb{E}[T] = \theta$, the inequality reduces to

$$\mathbb{V}[T] \leq \frac{1}{nI(\theta)}.$$

**Definition 4.** The efficiency $e(T)$ of an unbiased estimate $T$ is defined by the quotient of the Cramér-Rao bound and the true variance of the estimator.

$$e(T) = \frac{1}{nI(\theta)} \mathbb{V}[T].$$

An unbiased estimate is called efficient if its variance reaches the Cramér-Rao bound, which implies that $e(T) = 1$ for efficient estimates. If the variance of an unbiased estimate reaches the Cramér-Rao bound only for $n \to \infty$ the estimate is called asymptotically efficient:

$$\lim_{n \to \infty} \frac{1}{nI(\theta)} \mathbb{V}[T] = 1.$$
1.3 Affine equivariance

A further desirable characteristic of an estimator is the so called affine equivariance. This characteristic is specifically interesting when it comes to multivariate estimates for location or scale. The great advantage of estimates which are affine equivariant is the invariance of the estimate when the data is translated, rotated or the scale changes. This is the reason why in the last decades a lot of effort was put into the developing of affine equivariant estimates with a high breakdown point.

Let $X$ be $n \times p$ dimensional data, then an location estimator $T$ is called affine equivariant if and only if for all $p$-dimensional vectors $b$ and all nonsingular $p \times p$ matrices $A$ holds

$$T(XA + 1_n b^t) = T(X)A + 1_n b^t,$$

with $1_n$ being a column vector with $n$ components all equal to 1.

For the estimation of the covariance of the data $X \in \mathbb{R}^{n \times p}$ an estimator $C$, which is a positive definite symmetric $p \times p$ matrix is called affine equivariant if and only if

$$C(XA + 1_n b^t) = A^t C(X) A$$

hold true again for all $p$-dimensional vectors $b$ and all nonsingular $p \times p$ matrices $A$. 
2 Methods for one-dimensional data

In terms of one-dimensional data 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 the observations which lie out of the range between the location plus/minus multiple times the scale can be considered as potential outliers.

For this section we will suppose a sample \( s \) (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 \( s \) a weight \( w_i \) is attached that reflects the sample inclusion probability \( \pi_i \). In addition we have that \( w_i = \frac{1}{\pi_i} \), where \( \pi_i = \sum_{s \in S} p(s), i = 1, \ldots, N \) and we assume that \( \sum_{i \in s} w_i = N \).

2.1 Robust location \( \pm \) constant \( \ast \) 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.

2.1.1 Robust location

A very common measure for location is the arithmetic mean \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \). Since \( \bar{x} \) breaks down already with one out of \( n \) observations placed at any arbitrary position, the arithmetic mean has a breakdown point of 0%. A more suitable, robust measure of location is the median (\( med \)). For a random variable \( X \) with distribution \( F \) the median is defined by

\[
P(X \geq med) = F(med) = 0.5
\]

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

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

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%. Taking into account sample weights \( w_i, i = 1, \ldots, n \), the \( p \)th quantile, \( \hat{Q}_p \), is given by solving the estimating equation

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

The weighted sample median is obtained by \( p = 0.5 \) and has a breakdown point of 0%. This is demonstrated by the following example of data points together with sample weights (Table 1). It is easy to see that the sample median of the data in Table 1

\[\begin{array}{c}
\hline
x_i & w_i \\
\hline
1.0 & 0.2 \\
2.0 & 0.3 \\
3.0 & 0.5 \\
4.0 & 0.2 \\
5.0 & 0.1 \\
\hline
\end{array}\]
equals 5, while the weighted median is influenced so heavily by the last weight, that it equals 11. Therefore the weighted median need 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%.

<table>
<thead>
<tr>
<th>data</th>
<th>weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>25</td>
</tr>
</tbody>
</table>

Table 1: Data set showing that the weighted median has 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 values for weights have no dependency on the data values.

2.1.2 Robust scale

The most common estimator of scale of a sample is the empirical standard deviation, defined as $S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$. Since this estimator is not robust it cannot be useful 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

$$
P(X \leq Q_1) = F(Q_1) = 0.25$$
$$
P(X \leq Q_3) = F(Q_3) = 0.75$$

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 estimating Equation (3) 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

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

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

$$\text{MAD} = \text{med}_i |x_i - \text{med}(x_j)|$$

The MAD achieves the maximal breakdown point of 50%.

To calculate MAD in context with sample weights $w_i, i = 1, \ldots, n$ one has to solve Equation (3) for $p = 0.5$ to gain the weighted median $\hat{Q}_{0.5}$. Using Equation (3) once again on $|x_i - \hat{Q}_{0.5}|$ the estimation equation for the weighted MAD ($\text{MAD}_w$) is presented by

$$\sum_{i \in s} w_i [1 \{ |x_i - \hat{Q}_{0.5}| \leq \text{MAD}_w \} - 0.5] = 0$$

Both, IQR and MAD are not consistent estimators of the scale parameter $\sigma$. Assuming that the data comes from a normal distribution, consistent estimators are

$$S_{IQR} = \frac{\text{IQR}}{1.35}$$

$$S_{MAD} = \frac{\text{MAD}}{0.675}$$

### 2.1.3 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 within the data points are not considered outliers. Points that lie outside this interval are flagged as potential outliers. In our context this results in the following intervals

$$[\text{med} - c \cdot S_{IQR}, \text{med} + c \cdot S_{IQR}]$$

$$[\text{med} - c \cdot S_{MAD}, \text{med} + c \cdot S_{MAD}]$$

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 $\mu$ the expectation and $\sigma$ the standard deviation, overlaps just over 99% of the possible realizations.

Figure 1 illustrates the use of the boundaries from Equation (4) and (5), 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 with lower respectively higher location. In the upper part of Figure 1 one can see that with the use of IQR and MAD for outlier detection the four lowest and
Figure 1: 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).
five highest points would be declared potential outliers. Therefore almost every data points, which was generated by a different mechanism then the main bulk of the data was declared a potential outlier. The lower part of Figure 1 shows the influence of the sample weights to 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 sample weight positively depend on the value of the data points.

2.1.4 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 \( \lambda \) defining a particular transformation. For data \( x_i > 0 \) the Box-Cox transformation is defined by

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

To determine the appropriate value of \( \lambda \) for a given data sample, Box and Cox [1964] presented a maximum likelihood as well as a Bayesian approach. Figure 2 shows the impact of the Box-Cox transformation on outlier detection methods on a simulated data, following a standard log-normal distribution, sample without 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 (4) and (5) after the outlier detection has been applied on the Box-Cox transformed data. Calculating the parameter \( \lambda \) 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 \( \lambda \) in a robust way by using robust linear regression. Considering \( x_i, i = 1, \ldots, n \) already, depending on their values, sorted data points with \( x_i > 0 \) and the linear model

\[
x_i^{(\lambda_0)} = \alpha + \beta z_i + u_i
\]

with \( \alpha, \beta \) and \( \lambda_0 \) as real parameters \( x_i > 0 \) 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 \( E[u_i] = 0 \). To calculate the true Box-Cox parameter \( \lambda_0 \) one can apply MM-estimation to the responses \( x_i^{(\lambda)} \) for given \( \lambda \) and choose \( \lambda_0 \) such that the robust residual autocorrelation \( \rho_n(\lambda) \) is minimized.
Figure 2: 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).
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 $\lambda$, 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 Marazzi and Yohai [2006].

2.2 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:

$$
LF = Q_1 - 1.5\left(Q_3 - Q_1\right) \quad UF = Q_3 + 1.5IQR
$$

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 (3) for weighted sample median and weighted sample quartiles. Figure 1 displays the boxplots for a data sample, generated from a log-normal distribution, beneath the $x$-axis.

2.3 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

\[
MC(F) = \text{med}_{x_1 < m_F < x_2} h(x_1, x_2)
\]

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 \( h_r \) to determine the fences. Thus instead of using the interval of regular observation as

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

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 function \( h_l \) 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 \( h_r \)

\[
[Q_1 - 1.5e^{-3MC} IQR; Q_3 + 1.5e^{4MC} IQR]
\]

Dealing with a data sample, including sample weights, one can use Equation (3) to calculate weighted quartiles and weighted \( IQR \). Sample weights will not be incorporated into the medcouple since the factor \( 1.5e^{-3MC} \) and \( 1.5e^{4MC} \) has performed all around well by the simulations by Vandervieren and Hubert [2008].
2.4 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 recalibrate the sample weights or fitting data values for observations in the upper tail.

2.4.1 The Pareto distribution

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

\[ F_\theta(x) = 1 - \left( \frac{x}{x_0} \right)^{-\theta}, \quad x \geq x_0 \]

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

\[ F(x) = \begin{cases} G(x) & \text{if } x \leq x_0 \\ G(x_0) + (1 - G(x_0))F_\theta(x) & \text{else} \end{cases} \]

(7)

where \( G \) is an unknown distribution function [Dupuis and Victoria-Feser, 2006]. There are several methods to estimate the threshold \( x_0 \) and the shape parameter \( \theta \). 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 \( \theta \).
2.4.2 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

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

with $\bar{x}_w$ as the weighted mean and $Q_{0.98}$ and $Q_{0.97}$ as weighted quantiles defined by Equation (3).

2.4.3 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

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

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

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

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 $uf_\theta$. PDC estimator in is thus given by

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

The parameter $u$ can be estimated by

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

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 $w_1, \ldots, w_n$

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

$$\hat{u} = \frac{1}{\sum_{i=1}^k w_{n-k+i}} \sum_{i=1}^k \frac{w_{n-k+i} f_\theta(y_i)}{\int f^2_\theta(y) dy}.$$
2.5 Adjusting potential outliers

2.5.1 Adjusting outlier for univariate outlier detection methods

After locating potential outlier the data points will not be discarded for further calculations since this would result in a heavy loss of information. Therefore the values or weights of the potential outliers will be replaced to adjust for their extreme values. When using methods regarding robust location plus/minus constant times robust measure of scale potential outliers will be placed to the upper or lower boundaries defined by Equations (4) and (5). In the case of boxplot or adjusted boxplot potential outliers will be put to the upper or lower fences defined by (6).

2.5.2 Adjusting outliers for semi-parametric modeling methods

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 get 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. The order of the original values will be preserved.
3 Methods for multivariate data

For multivariate data, the task in finding potential outliers yields various problems. Potential outliers, which can be circumscribed by deviating too much from the main bulk of the data, are not as easily determined as in the one-dimensional case. Figure 4 shows simulated data points generated from multivariate standard normal distribution with very high correlation between the two dimensions including one outlier marked as point A. Furthermore, the dashed lines in red and blue indicate the borders of the interval \([\text{med} - 2 \times \text{MAD}, \text{med} + 2 \times \text{MAD}]\), when one-dimensional methods for detection of outliers are performed on each dimension of the data separately. The plot shows, that it is not sufficient to declare potential outliers as data points which are simply far away from the center of the data. The outlying point A is closer to the theoretical center of the data \((0, 0)\) than for example the point B. In addition the plot shows, that the point A would have been detected as potential outlier if one-dimensional methods would have been applied on each dimension of the data. The contrary is true for point B for which the value in the 2\textsuperscript{nd} dimension is outside of the interval \([\text{med} - 2 \times \text{MAD}, \text{med} + 2 \times \text{MAD}]\). The above example shows that for detecting potential outliers one needs a distance measure which takes into account the multidimensional structure of the data. Contrary to the one-dimensional case, where potentially outliers where classified by being "far enough" away from the center, in the multivariate case potential outliers will primarily be data points which are not in correspondence with the structure of the main bulk of the data. A very prominent measure which fulfills this criterion is the so called squared Mahalanobis distance \(MD_i^2\).

Figure 4: Simulated data from multivariate standard normal distribution including one outlier (point A)
Definition 5. Consider a data matrix $X \in \mathbb{R}^{n \times p}$ with $n$ observations $x_i$, for $i = 1, \ldots, n$, containing $p$ dimensions. Then the squared Mahalanobis distance for the $i$-th observation $MD_i^2$ is defined by

$$MD_i^2 = (x_i - \bar{x})^t S^{-1} (x_i - \bar{x}) ,$$

with $\bar{x}$ as the sample mean and $S$ as the sample covariance matrix

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

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^t .$$

In case of data following a multivariate normal distribution the squared Mahalanobis distance $MD_i^2$ is approximately chi-square distributed with $p$ degrees of freedom, $\chi^2_p$. Therefore, observations with high squared Mahalanobis distance are possible candidates for potential outliers. A common rule is to declare data points as potential outliers if they exceed the 97.5% quantile of the chi-square distribution, $\chi^2_{0.025}$. Using the squared Mahalanobis distance as defined above can yield the following two problems [see also Rousseeuw and Leroy, 1987]:

- **Masking:** Multiple outliers can influence the estimates for $\bar{x}$ and $S$ in such a way that their corresponding squared Mahalanobis distances are not large enough to identify these points as potential outliers.

- **Swamping:** Multiple outliers can influence the estimates for $\bar{x}$ and $S$ in such a way that data points which are consistent with the majority of the data get large squared Mahalanobis distances.

So the non-robustness of the classical estimates for location and scale, which have a breakdown point of 0%, distort the calculations when outliers or extreme values are present in the data and result in above mentioned problems. One way to take these two problems into account is the use of robust estimates for location $T$ and covariance $C$. The squared robust distance of a data point $x_i$, $i = 1, \ldots, n$, is then defined by

$$RD_i^2 = (x_i - T)^t C^{-1} (x_i - T) .$$

In the following this distance measure will be denoted by $d(x_i, T, C)$ or $d_i$, meaning that given a vector $T \in \mathbb{R}^p$ and a positive symmetric matrix $C \in \mathbb{R}^{p \times p}$ the distance $d(x_i, T, C)$ of an observation $x_i$, depending on $T$ and $C$ is defined by

$$d_i = d(x_i, T, C) = (x_i - T)^t C^{-1} (x_i - T) .$$

In the literature one can find many different robust estimates for location $T$ and covariance $C$ which not only differ in breakdown points but also in statistical efficiency. More developed robust estimates can be tuned to achieve high breakdown point combined with high efficiency.
3.1 M-estimates

The multivariate M-estimate for location and covariance can be derived by generalizing the Maximum Likelihood estimate (MLE) for \( \mu \) and \( \Sigma \).

Consider a \( p \)-dimensional random variable \( X \) with distribution function \( F \) and density function \( f(x, \mu, \Sigma) \) of the form

\[
f(x, \mu, \Sigma) = \frac{1}{\sqrt{|\Sigma|}} h(d(x, \mu, \Sigma)),
\]

where \( d(x, \mu, \Sigma) \) is the Mahalanobis distance of \( x \) with respect to \( \mu \) and \( \Sigma \),

\[
d(x, \mu, \Sigma) = (x - \mu)^t \Sigma^{-1} (x - \mu)
\]

and \( h(.) \) is a non-negative function such that \( \int_{\mathbb{R}^p} f(x, \mu, \Sigma) dx_1 \ldots dx_p = 1 \).

Note that for the choice of \( h(s) = (2\pi)^{-p/2} \exp(-s/2) \) the function \( f \) will be a multivariate normal density function. Another example is given by \( h(s) = c(s + v)/(p + v) \), with \( c \) as a constant, which will lead to the \( p \)-variate Student distribution with \( v \) degrees of freedom.

Let \( x_1, \ldots, x_n \) be an i.i.d. sample from a \( p \)-dimensional random variable \( X \) with distribution function \( F \) and density function \( f \) as mentioned above. The likelihood function to calculate \( \mu \) and \( \Sigma \) is then given by

\[
L(\mu, \Sigma) = \frac{1}{|\Sigma|^{n/2}} \prod_{i=1}^{n} h(d(x_i, \mu, \Sigma)).
\]

To maximize the likelihood function it is sufficient to maximize the log–likelihood function

\[
l(\mu, \Sigma) := \log(L(\mu, \Sigma))
\]

\[
l(\mu, \Sigma) = -\frac{n}{2} \log(|\Sigma|) + \sum_{i=1}^{n} \log(h(d(x_i, \mu, \Sigma))).
\]

Differentiating the log–likelihood function with respect to \( \mu \) and \( \Sigma \) yields the following system of equations

\[
\sum_{i=1}^{n} \frac{\partial \rho(d_i)}{\partial \mu} (x_i - \mu) = 0
\]

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \rho(d_i)}{\partial \Sigma} (x_i - \mu)(x_i - \mu)^t = \Sigma,
\]

with \( \rho(d_i) = -2 \log(h(d_i)) \) and \( d_i = d(x_i, \mu, \Sigma) \). In general, M-estimates \( \mu_M \) and \( \Sigma_M \) are defined as solutions of equations

\[
\sum_{i=1}^{n} W_1(d_i)(x_i - \mu) = 0
\]

\[
\frac{1}{n} \sum_{i=1}^{n} W_2(d_i)(x_i - \mu)(x_i - \mu)^t = \Sigma,
\]

18
with appropriate weight functions $W_1$ and $W_2$ [see also Maronna et al., 2006].
For existence and uniqueness of the solution for M-estimates, the function $dW_2(d)$ must be nondecreasing in $d$ [Maronna, 1976, see also]. In this case the M-estimate is also called monotone M-estimate and otherwise redescending M-estimate. While monotone M-estimate are solutions of Equation (11), redescending M-estimate are defined by minimization of some objective function. Monotone M-estimates are also of special interest since solutions to the M-estimating Equation (11) are consistent estimates. For a proof of consistency we refer to Huber [1981]. It can be shown that for the asymptotic breakdown point of an monotone M-estimate the inequality

$$\epsilon^* \leq \frac{1}{1 + p}$$

holds true.

In this work a so called constrained M-estimate, also abbreviated with CM-estimate, as described in Rocke [1996] will be used for empirical calculations. It is based on the translated bi-weight function and uses the MVE-estimate as high breakdown point initial estimate, which will be discussed later on.

The CM-estimate was originally introduced by Kent and Tyler [1996] as an answer to the low breakdown point of M-estimates. The CM-estimate differs from the M-estimate in the way that $\mu_{CM}$ and $\Sigma_{CM}$ minimize the log likelihood function from Equation (10) under the constraint

$$\frac{1}{n} \sum_{i=1}^{n} \rho \left( (x_i - \mu)^t \Sigma^{-1} (x_i - \mu) \right) \leq \epsilon \rho(\infty),$$

where $\rho(s)$ is a bounded and nondecreasing function for $s > 0$ and $\epsilon$ is between 0 and 1. Kent and Tyler [1996] show that the CM-estimate is affine equivariant and has a breakdown point of $\min(\epsilon, 1 - \epsilon)$ which results in a breakdown point of 50% for $\epsilon = 0.5$. Furthermore, the CM-estimate is consistent and asymptotically normal and can be tuned to have good efficiency even for choosing a high breakdown point.

To demonstrate the influence of outliers onto classical estimates for location and scatter compared to the M-estimate, 110 data points were simulated from multivariate normal distributions. From this data set 100 points have been simulated from a multivariate normal distribution with $\mu = (0, 0)^t$ and $\Sigma = \left( \begin{array}{cc} 1 & 0.95 \\ 0.95 & 1 \end{array} \right)$. The other 10 data points were simulated from a multivariate normal distribution with $\mu = (-1.5, 1.5)^t$ and $\Sigma = \left( \begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right)$. Figure 5 shows the influence of outliers onto classical estimates for location and scale as well as the result of the M-estimate. Displayed are 110 data points from multivariate normal distribution, from which Figure 5 displays the corresponding 0.975% tolerance ellipse for the classical and robust estimates. Compared to the robust estimate, the classical method is not able to detect all artificial outliers. Furthermore, it is clear visible that in the classical case the calculations for covariance are heavily influenced by the contamination.
Figure 5: 0.975% tolerance ellipse for classical and M-estimate for location and covariance applied to simulated data from 2-dimensional normal distribution with added contaminated data points.

3.2 S-estimate and MM-estimate

The S-estimate for location and scale, $\mu_S$ and $\Sigma_S$ was first introduced by Davies [1987] and later studied by Lopuhaä [1989]. To introduce the S-estimate, one first has to introduce the notion of an M-scale estimate. Consider $p$-dimensional observations $\{x_1, \ldots, x_n\}$ then multivariate estimates for location and scale can be chosen such that the distances $MD^2_i = MD^2(x_i, \hat{\mu}, \hat{\Sigma})$, $i = 1, \ldots, n$ are small. Therefore robust location and scale estimates can be derived by using a robust univariate measure of scale $\hat{\sigma}$ and solving the minimization problem

$$\min_{\hat{\mu}, \hat{\Sigma}} \hat{\sigma}(MD^2_1, \ldots, MD^2_n)$$

s.t

$$\hat{\mu} \in \mathbb{R}^p, \: \hat{\Sigma} \in \mathbb{R}^{p \times p} \: , \text{ positive definite and symmetric and } |\hat{\Sigma}| = 1 \: .$$

The constraint $|\hat{\Sigma}| = 1$ ensures that minimizing the scale of the distances $MD^2_i$ does yield a more meaningful solution, otherwise a trivial solution of the problem would be found in letting the smallest eigenvalue of $\hat{\Sigma}$ tend to zero.

With the notion of an M-scale estimator we can now introduce the so called S-estimator. Given $p$-dimensional observations $\{x_1, \ldots, x_n\}$ the S-estimator is defined via an M-scale
estimate by choosing $\hat{\sigma}$ such that it satisfies

$$\frac{1}{n} \sum_{i=1}^{n} \rho \left( \frac{MD_i^2}{\hat{\sigma}} \right) = \delta$$

with $\rho$ as a non increasing function with $\rho(0) = 0$, $\rho(\infty) = 1$ and $\delta \in (0, 1)$.

Another way to define an S-estimate is by finding the vector $\mu_S$ and the positive definite symmetric matrix $\Sigma_S$ which solve

$$\min_{\Sigma; \text{positive definite, symmetric}} \quad \text{det}(\Sigma)$$

s.t.

$$\frac{1}{n} \sum_{i=1}^{n} \rho(MD_i^2) = \epsilon \rho(\infty) ,$$

where $\epsilon$ is a constant between 0 and 1 and $\rho(s)$ is a bounded and nondecreasing function for $s > 0$, with $\rho(0) = 0$. This definition of the S-estimate requires $\rho$ to be continuous. For discontinuous $\rho$ the S-estimate is defined in the same way but with the constraint

$$\frac{1}{n} \sum_{i=1}^{n} \rho(MD_i^2) \geq \epsilon \rho(\infty) .$$

The asymptotic breakdown point of the S-estimator depends on the choice of $\epsilon$ and is given by $\min(\epsilon, 1 - \epsilon)$ as shown in Davies [1987]. Furthermore Davies [1987] shows that the S-estimator is an affine equivariant estimator which is consistent and asymptotically normal if the function $\rho$ is smooth enough. A drawback of the S-estimate is the increasing inefficiency if $\epsilon$ increases from 0 to $\frac{1}{2}$. For high breakdown point the S-estimate can be quite inefficient, as noted in Lopuhaä [1989].

S-estimates are closely related to M-estimates, as shown by Lopuhaä [1989]. For a differentiable $\rho$-function, the S-estimates $\mu_S$ and $\Sigma_S$ also solve the following system of equations,

$$\sum_{i=1}^{n} W \left( \frac{MD_i}{\hat{\sigma}} \right) (x_i - \mu) = 0 \quad (12)$$

$$\frac{1}{n} \sum_{i=1}^{n} W \left( \frac{MD_i}{\hat{\sigma}} \right) (x_i - \mu)(x_i - \mu)^t = \Sigma , \quad (13)$$

with $W = \rho'$ and $\hat{\sigma}(MD_1, \ldots, MD_n)$. Note that $dW(d)$ cannot be monotone if $\rho$ is bounded. Therefore the S-estimate is a redescending M-estimate and the system of equations above only yields local minima for $\hat{\sigma}$.

Many algorithms have been developed to find the solution of the S-estimate. For example SURREAL proposed by Ruppert [1992], the Bisquare S-estimation with HBDP start as described in Maronna et al. [2006], the Rocke type S-estimates by Rocke [1996], and Fast S estimates proposed by Salibian-Barrera and Yohai [2006]. A detailed description
about the computation of these estimates can be found in Todorov and Filzmoser [2009].
In this work an algorithm similar to the Fast S-estimates proposed by Salibian-Barrera
and Yohai [2006] will be used for empirical calculations.
As with the MCD estimate the MVE estimate is in general not very efficient.
Figure 6 displays the 0.975% tolerance ellipsoid of the S-estimate and classical estimate
applied on the same data as in Figure 5.

![Figure 6: 0.975% tolerance ellipse for classical and S-estimate for location and covariance
applied to simulated data from 2-dimensional normal distribution with added
contaminated data points.](image)

The MM-estimator was initially introduced by Yohai [1987] in the context of linear
regression. The idea of the MM-estimate is to start with an S-estimate with high break-
down point, which is as mentioned before quite inefficient. This preliminary scale esti-
mate is then used in combination with a "better" tuned $\rho$-function to gain an efficient
estimate with high breakdown point.
The MM-estimate for robust multivariate estimation of covariance and location was first
introduced by Tatsuoka and Tyler [2000]. They introduce a new class of estimators, the
so called *M-estimates with auxiliary scale* in which the S-estimates and CM-estimates
can be embedded. The M-estimate with auxiliary scale is defined as follows.
Let $\{x_1, \ldots, x_n\}$ be $p$-dimensional observations $\sigma()$ a given scale statistic with $\sigma(.) > 0$
and $\rho(s)$ a continuous, nondecreasing function for $s \geq 0$ with $0 = \rho(0) < \rho(\infty) < \infty$.
Then the multivariate M-estimates with auxiliary scale $\sigma(x_1, \ldots, x_n)$ for location $\mu_M^{aux}$
and scatter $\Sigma_{M;\text{aux}} = \sigma^2(x_1, \ldots, x_n) \Gamma_{M;\text{aux}}$, where the pair $(\mu_{M;\text{aux}}, \Gamma_{M;\text{aux}})$ minimizes

$$
\frac{1}{n} \sum_{i=1}^{n} \rho \left( \frac{(x_i - \mu)^T \Gamma^{-1} (x_i - \mu)}{\sigma^2(x_1, \ldots, x_n)} \right)
$$

(14)

over all $(\mu, \Gamma)$ for which $\Gamma \in \mathbb{R}^{p \times p}$ is a positive definite matrix with $\det(\Gamma) = 1$ and $\mu \in \mathbb{R}^p$. The S- and CM-estimates can be derived from this definition of the M-estimates with auxiliary scale as shown in Tatsuoka and Tyler [2000].

Tatsuoka and Tyler [2000] point out that taking $\sigma(x_1, \ldots, x_n)$ as a preliminary scale statistic results in a new class of estimates which they call the MM-estimates. For example choosing $\sigma(x_1, \ldots, x_n) = \det(\Sigma_S)^{1/(2p)}$ where $\Sigma_S$ is an S-estimate for scatter with high breakdown point. Using this scale statistic and a different $\rho$-function as for the S-estimate the resulting M-estimate with auxiliary scale $\sigma(x_1, \ldots, x_n)$ will be an MM-estimate. The resulting MM-estimate is an affine equivariant estimator with high breakdown point and high efficiency.

Lopuhaä [1992] also introduced an MM-estimate with the difference that he used the whole scatter matrix as preliminary auxiliary statistic and then uses it to solve Equation (14) for $\mu$.

Figure 7 displays the 0.975% tolerance ellipsoid of the S-estimate and classical estimate. The underlying data is the same as in Figure 5.
3.3 Minimum covariance determinant (MCD) estimator

The minimum covariance determinant (MCD) estimator was first introduced by Rousseeuw [1985] and is an affine equivariant estimate for location and scale, which can achieve the maximal breakdown point of 50%.

Given a data set \( \{x_1, \ldots, x_n\} \) with \( x_i \in \mathbb{R}^p, \ i = 1, \ldots, n \), the minimum covariance determinant estimate is defined by the subset of \( h \) data points \( \{x_{i_1}, \ldots, x_{i_h}\} \) with the smallest determinant of the sample covariance matrix among all subsets of size \( h \). The MCD estimate for location \( T_{MCD} \) and covariance \( C_{MCD} \) is then defined by:

\[
T_{MCD} = \frac{1}{h} \sum_{j=1}^{h} x_{i_j}
\]

\[
C_{MCD} = \frac{1}{h-1} \sum_{j=1}^{h} (x_{i_j} - T_{MCD})(x_{i_j} - T_{MCD})^t
\]

The coefficients \( c_{ccf} \) and \( c_{sscf} \) are correction factors. The consistency correction factor \( c_{ccf} \) is chosen such that \( C_{MCD} \) is consistent at the multivariate normal model, however this factor is only reliable for subset size \( h \) close to \( n \). In other cases the use of this factor produces overestimation. The small sample correction factor \( c_{sscf} \) is chosen such that the estimate is unbiased at samples with small number of observations. For more information on the correction factors see also Rocke [1996], Croux and Haesbroeck [1999], Pison et al. [2002], Todorov [2008].

With the parameter \( h \) one can regulate the breakdown point of the MCD estimate. For \( h \) equal to the sample size \( n \), the MCD estimate reduces to the classical estimate for location and scale, which inherit a breakdown point of 0%. Choosing \( h \) equal \( \lfloor (n + p + 1)/2 \rfloor \), the maximal possible breakdown point of 50% can be achieved. Since in many dataset the amount of contamination is not very high, \( h = 0.75n \) is often used in practice to get a sufficient breakdown point and still good statistical properties.

Figure 8 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 as described in Figure 5. Figure 8 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.

The MCD estimate is in general not very efficient, especially if \( h \) is chosen to get the maximal breakdown point. A way to tackle the low efficiency of the MCD estimate is by introducing an adaptation of the MCD estimate including sample weights. The weights \( w_i, \ i = 1, \ldots, n \) are introduced in such a way, that

\[
w_i = \begin{cases} 
1 & \text{if } (x_i - T_{MCD})^t C_{MCD}^{-1}(x_i - T_{MCD}) \leq \chi^2_{p,0.975} \\
0 & \text{otherwise}
\end{cases}
\]

with \( T_{MCD} \) and \( C_{MCD} \) as the original MCD estimates for location and scale. The
Figure 8: 0.975% 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.

Reweighted MCD estimates, $T_R$ and $C_R$ are the given by

$$T_R = \frac{1}{v} \sum_{j=1}^{n} x_{ij}$$

$$C_R = c_{r,ccf}c_{r,sscf} \frac{1}{v - 1} \sum_{j=1}^{n} w_i (x_{ij} - T_R)(x_{ij} - T_R)^t,$$

where $v$ equals the sum of all weights, $v = \sum_{i=1}^{n} w_i$. As with the original MCD estimate, the factors $c_{r,ccf}$ and $c_{r,sscf}$ are correction factors to achieve consistency ($c_{r,ccf}$) and unbiasedness for small samples ($c_{r,sscf}$).

The breakdown point of the reweighted MCD estimates is identical with the initial MCD estimate, but with better efficiency.

A great advantage of the MCD estimate and also one of the reasons why it is a popular robust estimate is the fast algorithm for its computation.

### 3.4 Minimum volume ellipsoid (MVE) estimator

The minimum volume ellipsoid estimate was, like the MCD estimate, first introduced by Rousseeuw [1985], and is also an affine equivariant estimate for location and scale which can achieve a breakdown point of 50%.

As the name implies, the MVE estimate is defined by the ellipsoid of minimal volume
containing at least half the data points of a data set $X \in \mathbb{R}^{n \times p}$. To be more precise, for a data sample $X \in \mathbb{R}^{n \times p}$, the MVE-estimates for location and scale are defined by the estimates $T_{MVE}$ and $C_{MVE}$ which fulfill

$$\min_{C \in C} \det(C)$$

s.t

$$\# \mathcal{I}_{T,C} \geq \left\lfloor \frac{n + p + 1}{2} \right\rfloor, \text{ with } \mathcal{I}_{T,C} := \{ x_i : (x_i - T)^t C^{-1}(x_i - T) \leq c^2 \}$$

$$C = \left\{ C : C = \frac{1}{\# \mathcal{I}_{T,C} - 1} \sum_{i \in \mathcal{I}_{T,C}} (x_i - T)(x_i - T)^t \right\},$$

with $\#$ as the cardinality of a set. Note that the subsets $\mathcal{I}_{T,C}$ depend on $T$ and $C$ and are made up of all data points $x_i$ for which their Mahalanobis distance based on $T$ and $C$ is less or equal a constant $c^2$. For choosing $c^2$ as $\chi^2_{p,0.5}$ this set is made up of the points which create a 50% tolerance ellipsoid. It should be noted that contrary to the MCD-estimate for which the subsets need to be of size $h$, the valid subsets for the MVE-estimate are determined by the mahalanobis distance and the constant $c^2$. To solve this problem in practice one usually generates an approximated solution. This is done by using ellipsoids determined by the covariance matrix of $p + 1$ data points. For those ellipsoids an effective improvement of the sub-sampling procedure is applied. Even though it is not known that this sub-sampling procedure improves the MVE it can still be recommended since this procedure is all around successful. A detailed description of this procedure can be found in Maronna et al. [2006]. Agulló [1996] proposed an exact method for the MVE estimate with the downside of being not computationally feasible unless $p$ and $n$ are small.

As with the MCD estimate the MVE estimate is in general not very efficient. Figure 9 displays the 0.975% tolerance ellipsoid of the MVE estimate and classical estimate applied on the same data as in Figure 5.

### 3.5 The Stahel-Donoho estimator

The Stahel-Donoho estimator was first introduced by Stahel [1981a,b] and Donoho [1982] and is the first multivariate affine equivariant estimator for location and scale which achieves the maximal breakdown point of 50%. After the analysis by Maronna and Yohai [1995] the estimator became better known.

The Stahel-Donoho estimator takes into account weights which correspond to the "outlyingness" of each data point. Given a data set $X \in \mathbb{R}^{n \times p}$ the univariate outlyingness of the data point $x_i$ from $X$ along a vector $a \in \mathbb{R}^p$, with $\|a\| = 1$ is given by

$$r(x_i, a) = \frac{|x_i^t a - m(a^t X)|}{s(a^t X)},$$

with $(a^t X)$ as the projection of the data matrix $X$ onto the vector $a$. The functions $m(.)$ and $s(.)$ are robust univariate estimates for location and scale respectively. To expand
Figure 9: 0.975% 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.

this measure of outlyingness for a data point $x_i$ to the multivariate case, one can apply the function $r(x_i, \cdot)$ to all possible directions $a$ and define a measure for multivariate outlyingness of $x_i$ by

$$r_i = r(x_i) = \sup_{\|a\|=1} r(x_i, a) \quad .$$

(18)

Using as robust univariate estimates the median and MAD, the multivariate measure of outlyingness becomes

$$r_i = \sup_{\|a\|=1} \frac{|x_i^t a - med_j(x_j^t a)|}{med_i(|x_i^t a - med_j(x_j^t a)|)} \quad .$$

(19)

The Stahel-Donoho estimator uses this measure of outlyingness by incorporating it into weights $w_i = w(r_i), \ i = 1, \ldots, n$, such that the weight $w(r_i)$ is indirectly proportional to the outlyingness of the data point $x_i$. Following this, the Stahel-Donoho estimator for location $T$ and covariance $C$ is defined by

$$T_{SD}(X) = \frac{\sum_{i=1}^{n} w(r_i)x_i}{\sum_{i=1}^{n} w(r_i)}$$

$$C_{SD}(X) = \frac{1}{\sum_{i=1}^{n} w(r_i)} \sum_{i=1}^{n} w(r_i)(x_i - T)(x_i - T)^t \quad .$$
The weight function \( w(.) \) is a non-increasing function, and \( w(r) \) as well as \( w(r)r^2 \) are bounded. The weight function proposed by Maronna and Yohai [1995] is given by the so-called "Huber weights"

\[
w(r) = \min \left( 1, \left( \frac{c}{r} \right)^2 \right),
\]

with \( c = \sqrt{\chi^2_{p,0.95}} \).

To calculate the Stahel-Donoho estimator in an exact way one has to compute \( r(x_i, a) \) for all directions \( a \). Since this is computationally not feasible, only approximate solutions exist for the Stahel-Donoho estimator. The approximate solutions are calculated by subsampling a large number of directions \( a \) and computing the multivariate measure of outlyingness for this subsample of directions. Figure 10 displays the 0.975% tolerance ellipsoid of the Stahel Donoho estimate and classical estimate applied on data simulated from multivariate normal distribution. Again, the data contain 110 data points from which 10 data points are simulated from a multivariate normal distribution with different \( \mu \) and \( \mu \) than the rest of the data as used in Figure 5.

![Figure 10: 0.975% 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.](image)

**3.6 Orthogonalized Gnanadesikan–Kettenring estimator**

The orthogonalized Gnanadesikan–Kettenring estimator, also OGK-estimator, is based on the robust bivariate covariance estimator \( s_{jk} \) proposed by Gnanadesikan and Ketten-
Let $X_j$ and $X_k$ be a pair of random variables and $\sigma(.)$ an estimate for the standard deviation, then $s_{jk}$ is defined by

$$s_{jk} = \frac{1}{4} \left( \sigma \left( \frac{X_j}{\sigma(X_j)} + \frac{X_k}{\sigma(X_k)} \right)^2 - \sigma \left( \frac{X_j}{\sigma(X_j)} - \frac{X_k}{\sigma(X_k)} \right)^2 \right).$$

(20)

Given random variables $X_1, \ldots, X_p$ an estimate of the covariance matrix can be obtained by computing $s_{jk}$ for each $j = 1, \ldots, p$ and $k = 1, \ldots, p$. When taking a robust estimate for the standard deviation $\sigma(.)$ the estimate for $s_{jk}$ will also be robust which will result in a robust estimate for the covariance matrix. Since $s_{jk} = s_{kj}$, this estimator produces a symmetric matrix. However the resulting matrix is in general not positive definite and also lacks the property of affine equivariance. This estimate for covariance was studied by Devlin et al. [1981] and later on used by Maronna and Zamar [2002] for a modified version of the OGK estimator. This modified version overcomes the lack of positive definiteness.

Let $\sigma(.)$ and $m(.)$ be robust univariate estimators for scale and location. Given a data matrix $X \in \mathbb{R}^{n \times p}$ with rows $x_i, i = 1, \ldots, n$ and columns $x^j, j = 1, \ldots, p$, a robust estimate for covariance $C_{OGK}$ and location $T_{OGK}$ is defined by the following steps:

1. Compute the normalized data matrix $Y$ from the original data matrix $X$ by

$$y_i = D^{-1}x_i, i = 1, \ldots, n$$

with

$$D = diag(\sigma(x^1), \ldots, \sigma(x^p))$$

2. Use the matrix $Y$ to compute the matrix $U = [u_{jk}]$ with

$$u_{jk} = \begin{cases} s_{jk} = s(y^j, y^k), & \text{for } j \neq k \\ 1 & \text{else} \end{cases} .$$

Here $s(., .)$ is a robust estimate for the covariance of two random variables as in (20) and $y^j, j = 1, \ldots, p$ are the columns of $Y$.

3. Compute the eigenvalues $\lambda_j$ and corresponding eigenvectors $e_j, j = 1, \ldots, p$, of $U$. It follows that $U = E\Lambda E^t$ with $E$ the orthogonal matrix containing the eigenvectors $e_j, j = 1, \ldots, p$, as columns and $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$, the diagonal matrix containing the eigenvalues $\lambda_j j = 1, \ldots, p$. This procedure is better known as the principal component decomposition of $Y$.

4. Compute the principal components of $Y$ by calculating the matrix $Z$ with

$$z_i = E^t y_i = E^t D^{-1} x_i \quad \text{for } i = 1, \ldots, n .$$

Then the principal components of $Y$ are given by $(z^1, \ldots, z^p)$. 29
5. Compute $\sigma(z^j)$ and $m(z^j)$ for $j = 1, \ldots, p$ and define $\Gamma$ and $\boldsymbol{\nu}$ as

$$
\Gamma := \text{diag}(\sigma(z^1)^2, \ldots, \sigma(z^p)^2), \quad \boldsymbol{\nu} := (m(z^1), \ldots, m(z^p))^t
$$

Then the estimator for covariance and location are given by

$$
C_{OGK} = \mathbf{A} \Gamma \mathbf{A}^t \quad \text{with} \quad \mathbf{A} = \mathbf{D} \mathbf{E}
$$

$$
T_{OGK} = \mathbf{A} \boldsymbol{\nu}
$$

It is possible to iterate this procedure by performing these steps on the data matrix $\mathbf{Z}$ calculated in the fourth step. However it was shown by simulations from Maronna and Zamar [2002] that iterations of the procedure beyond the second did not improve the results.

Finally, it is desirable to increase the efficiency of the estimate. To increase efficiency it is possible to introduce a single reweighting step as in Equations (17). Instead of using the 0.975 quantile of the $\chi^2_p$ distribution in the case of the OGK-estimate the weights will be based on the 0.9 quantile of the $\chi^2_p$. Correction factors for small samples or consistency at the normal model will not be used.

A big role in the calculations of the OGK-estimator is played by the choice of the robust univariate estimates for location $m(.)$ and $\sigma(.)$. Maronna and Zamar [2002] give an example for the choice of $m(.)$ and $\sigma(.)$ where both estimates are bias robust and efficient at the normal distribution. They suggest for $\sigma(.)$ an estimate similar to the $\tau$-scale estimate and as location estimate a weighted mean.

The breakdown point of the OGK estimator is determined by the breakdown point of $m(.)$ and $\sigma(.)$. Maronna and Zamar [2002] show that the breakdown point of $C_{OGK}$ and $T_{OGK}$ is not less than the one of $m(.)$ and $\sigma(.)$.

Contrary to the MCD or MVE estimators where the calculations of the estimators is computationally very intensive, the OGK-estimator has comparably very low computation times. Therefore it is, in this way, superior when it comes to very large data sets and was also discussed for applications on data mining in Alqallaf et al. [2002]. The drawback of the OGK estimator is that it is not an affine equivariant estimator.

Figure 11 displays the 0.975% tolerance ellipsoid of the OGK-estimate and classical estimate applied on 110 data points simulated from multivariate normal distribution as in Figure 5.

### 3.7 Epidemic algorithm

The Epidemic algorithm (EA) builds on the notion, that in the case of an epidemic, resulting from a virus, bacteria or others, regions or individuals which are far away from the main population are usually infected last. To utilize this behavior for outlier detection an epidemic is simulated, starting from the center of the available data and progressing stepwise until most or all of the data points are infected. The infection of a clean data point is subject to the transmission probability, depending on the distance between the clean and infected data points.
Figure 11: 0.975% 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.

Béguin and Hulliger [2004] present a description of the EA in context with outlier detection which is also able to deal with incomplete survey samples. The distances are calculated by using the Euclidean distance and data points are standardized by the median and the median absolute deviation beforehand. Meaning that for a data matrix \( X \in \mathbb{R}^{n \times p} \) the point \( x_{ik} \) is replaced by

\[
\tilde{x}_{ik} = x_{ik} - \frac{\text{med}_i(x_{ik})}{\text{mad}_i(x_{ik})}
\]

and the distances between an observation \( x_i \) and \( x_j \) are then defined by

\[
d(x_i, x_j) := \left( \sum_{k=1}^{p} (\tilde{x}_{ik} - \tilde{x}_{jk})^2 \right)^{1/2}.
\]

For the starting point of the epidemic the sample spatial median \( x_{i(c)} \) is chosen with

\[
i(c) = \arg \min_i \left( \sum_{j=1}^{n} d(x_i, x_j) \right).
\]

If the sample spatial median is not unique the observation with the smallest index is chosen.
The probability that the observation \( i \) infects the observation \( j \) is defined by the transmission function \( h(d_{ij}) \) defined as

\[
h(d_{ij}) = h(d(x_i, x_j)) := \mathbb{P}[j|i] = \mathbb{P}[i|j] = h(d_{ji})
\]

with \( h(d) \) monotone decreasing and \( h(0) = 0 \leq h(d) \leq 1 \).

The EA starts at the point \( x_{ik} \) at time \( t = 1 \) with a set of infected points \( I_1 = \{i(c)\} \) for which the infection time \( t_{i(c)} = 1 \) and proceeds as follows.

1. increase the running time by 1, \( t := t + 1 \)

2. calculate the total infection probability \( \mathbb{P}[j|I_{t-1}] \) for all non-infected points \( j \not\in I_{t-1} \), with \( I_{t-1} \) as the set of points which have been infected until time \( t - 1 \).

3. calculate the expected number of infections by \( v = \sum_{j \not\in I_{t-1}} \mathbb{P}[j|I_{t-1}] \). Then infect the \( v \) non-infected points with largest total infection probability and set their infection time to \( t \).

4. if \( |I_t| = n \) or if the algorithm does not infect additional observations for \( l \) consecutive time steps, e.q \( t - \max_i \{t_i : i \in I_t\} > l \), then stop. Otherwise go to (1).

After the EA is completed the infection times indicate which of the observations is an outlier or not. The boxplot rule is applied on those infection times to determine if an observation is an outlier.

If the data inherits incomplete observations the calculation of the median and the median absolute deviation are calculated without the missing values and the distance between an observation \( x_i \) and \( x_j \) are calculated by

\[
d'_{ij} = d(x_i, x_j, r_i, r_j) = \left( \frac{p}{\sum_{k=1}^{p} r_{ik} r_{jk}} \sum_{k=1}^{p} r_{ik} r_{jk} (\tilde{x}_{ik} - \tilde{x}_{jk})^2 \right)^{1/2}
\]

with

\[
r_{ik} = \begin{cases} 
1 & \text{if } x_{ik} \text{ is given} \\
0 & \text{if } x_{ik} \text{ is missing}
\end{cases}
\]

For the EA the choice of the transmission function \( h() \) and for the calculation of the distances has a great impact on the outcome of the algorithm. Ideally these parameters should be chose separately for every data set to utilize the epidemic algorithm to its full potential. For the calculation in this work however the EA algorithm was used in connection with the Euclidean distance and a linear transmission function. For more details on the Epidemic algorithm and its application on outlier detection see Béguin and Hulliger [2004].
3.8 BACON-EEM

The BACON-EEM algorithm was proposed by Béguin and Hulliger [2008] and presents a robust estimate for location and covariance given multivariate data, which is unimodal and roughly elliptical symmetric. The BACON-EEM algorithm is a combination of the Blocked Adaptative Computationally-efficient Outlier Nominators (BACON) algorithm, proposed by Billor et al. [2000], and the Estimated-Expectation/Maximization (EEM) algorithm, which is a modified version of the Expectation/Maximization algorithm. An important aspect of the BACON-EEM algorithm is its ability to handle data with incomplete observations.

3.8.1 BACON algorithm

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 Béguin and Hulliger [2008]. For a data matrix \( \mathbf{X} \in \mathbb{R}^{n \times p} \) denote \( \mu_X, \Sigma_X \) and \( MD_X(x) \) as the classical estimates for location and covariance as well as the corresponding Mahalanobis distance of an observation \( x \).

The first step of the BACON algorithm contains 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 \( cp \) points, \( c = 3 \), with smallest Mahalanobis distance \( MD_X(x_i), i = 1, \ldots, n \). Starting with the subset \( G \) and the BACON algorithm contains the following steps:

1. Compute the Mahalanobis distances, corresponding to the subset \( G \), \( MD_G(x_i) = (x_i - \mu_G)^t \Sigma_G^{-1} (x_i - \mu_G), i = 1, \ldots, n \), for every observation in \( X \).

2. Determine a new subset \( G' \), containing all the points with \( MD_G(x_i) < c_{np} \chi^2_{p, \alpha / n} \), with \( c_{np} = c_{np} + c_{hr} \) 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.

3.8.2 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 \( f(x, \theta) \), the complete log-likelihood can be written as

\[
l(\theta|X) = \nu(\theta)^T \cdot T(X) + Ng(\theta) + c,
\]

with \( \nu = (\nu_1, \ldots, \nu_k) \) as the canonical form of the parameter \( \theta \) and \( T(X) = (T_1(X), \ldots, T_k(X)) \) 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 \( \sum_{i=1}^{n} x_i^k \) and sums of products \( \sum_{i=1}^{n} x_i^k x_i^l, \ 1 \leq k, l \leq p \). In the E-step the conditional expectations of these sums are calculated, given the preliminary parameter \( \theta^{(t)} \) and the observed data \( X_o \).

For these conditional expectations it can be shown that

\[
\mathbb{E}
\left(
\sum_{i=1}^{n} x_i^k|X_o, \theta^{(t)}
\right)
= \sum_{i=1}^{n} \mathbb{E}
\left(
 x_i^k| x_i^{obs}, \theta^{(t)}
\right), \ 1 \leq k \leq p
\]

and

\[
\mathbb{E}
\left(
\sum_{i=1}^{n} x_i^k x_i^l|X_o, \theta^{(t)}
\right)
= \sum_{i=1}^{n} \mathbb{E}
\left(
 x_i^k x_i^l| x_i^{obs}, \theta^{(t)}
\right), \ 1 \leq k, l \leq p
\].

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

\[
T^{k0} = \sum_s w_s \mathbb{E}(x_i^k|x_i^{obs}, \theta^{(t)}), \ 1 \leq k \leq p
\]

and

\[
T^{kl} = \sum_s w_s \mathbb{E}(x_i^k x_i^l|x_i^{obs}, \theta^{(t)}), \ 1 \leq k, l \leq p
\].

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

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

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

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 $cp$ observations with minimal squared marginal Mahalanobis distance, $MD^2_{\text{marg}}$. The marginal Mahalanobis distance can be used if observations ($x$) have an unobserved ($x_m$) and observed part ($x_o$) and is defined by

$$MD^2_{\text{marg}} = \frac{p}{q}(x_o - \mu_o)'\Sigma^{-1}_{oo}(x_o - \mu_o),$$

where $\mu_o$ and $\Sigma_{oo}$ are part of the location vector and covariance matrix corresponding to $x_o$. 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 $\hat{\mu}_G$ and $\hat{\Sigma}_G$ using the EEM-algorithm.

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

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 please have a look at Béguin and Hulliger [2008]. Figure 12 displays the 0.975% tolerance ellipsoid of the estimates from the final step of the BACON-EEM and classical estimates for location and covariance applied on 110 data points simulated from multivariate normal distribution as in Figure 5.

3.9 Imputation of potential outliers

As mentioned in the previous section dealing with potential outliers after they have been detected is an important task. Discarding them would result 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 favorable. 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 winsorising them onto the boundaries of a 0.975% tolerance ellipse.

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 potential outliers have been identified, they will be moved onto the 0.975% tolerance ellipse of the previously calculated location and covariance estimate, in direction of the robust center of the data set. To be more precise, let $T_R$ and $C_R$ be the robust estimates.
of location and covariance of the $p$-dimensional observations $\{x_1, \ldots, x_n\}$. Furthermore let $x_j$ be a potential outlier i.e. $d(x_j, T_R, C_R) > \chi^2_{p,0.975}$. To impute this point onto the 0.975% tolerance ellipse into the direction of the location $T_R$ implies that the imputed observation $\tilde{x}_j$ must be on the straight line between $x_j$ and $T_R$, which means $\tilde{x}_j \in \{\alpha x_j + (1 - \alpha) T_R, 0 \leq \alpha \leq 1\}$. In addition to that, the imputed point $\tilde{x}_j$ must lie on the tolerance ellipse which results in $d(\tilde{x}_j, T_R, C_R) = \chi^2_{p,0.975}$. To conclude, the imputed observation must solve the following equalities

$$
(\tilde{x}_j - T_R)^t C_R^{-1} (\tilde{x}_j - T_R) = \chi^2_{p,0.975}
$$

$$
\tilde{x}_j = \alpha x_j + (1 - \alpha) T_R
$$

$$
\alpha \in [0, 1]
$$

To determine $\alpha$, one can simply use the first two equations to obtain

$$
((\alpha x_j + (1 - \alpha) T_R) - T_R)^t C_R^{-1} ((\alpha x_j + (1 - \alpha) T_R) - T_R) = \chi^2_{p,0.975}
$$

$$
(\alpha (x_j - T_R))^t C_R^{-1} (\alpha (x_j - T_R)) = \chi^2_{p,0.975}
$$

$$
\alpha^2 (x_j - T_R)^t C_R^{-1} (x_j - T_R) = \chi^2_{p,0.975}
$$

$$
d(x_j, T_R, C_R)
$$

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

Figure 13 displays the imputation of outliers from data generated by multivariate normal distribution. The data exist of 110 data points from which 100 observations have
been simulated from a multivariate normal distribution with $\mu = (0,0)^t$ and $\Sigma = \begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix}$. The other 10 data points were simulated from a multivariate normal distribution with $\mu = (-1.5, 1.5)^t$ and $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. The blue symbols “+” indicate the positions of the data points after imputation.

Figure 13: Imputation of potential outliers of data from 2-dimensional normal distribution with added contaminated data points.
4 Measures of inequality

Measures of inequality based on income or welfare distributions play an important role in economics. It is possible to get a deeper understanding in the historic development of the distribution of income and on the possible influences of political instruments, such as taxes. Furthermore, measures of inequality are useful to compare different countries and to understand the impact of inequality on the welfare and wealth of these countries. Besides the distribution of income, measures of inequality can be used on the distribution of social welfare, household expenditures and many other population related quantities. Measures of inequality are also used in natural sciences, see for example Dixon et al. [1987] and Damgaard and Weiner [2000].

A very popular measure of inequality is the so called Gini coefficient which can be derived by the Lorenz curve [see also Gini, 1912 and Lorenz [1905]].

4.1 Lorenz curve and Gini coefficient

The Lorenz curve was initially introduced in context with total income of a population. When ordering the population by income earned, the Lorenz curve is the percentage of earned income plotted by various portions of the population size.

For a more general definition let $X$ be a 1-dimensional random variable with distribution function $F(x)$. Consider the quantile function $F^{-1}(p)$ of the probability distribution $F(x)$ defined by

$$F^{-1}(p) = \inf \{x : F(x) \geq p\}. $$

Note that the quantile function is well defined for discrete and continuous distributions. Then the Lorenz curve $L(p)$ corresponding to the random variable $X$ with distribution function $F(x)$ and finite mean $\mu = \int x dF(x)$, $\mu \neq 0$ is defined by

$$L(p) = \mu^{-1} \int_0^p F^{-1}(t) dt, \quad 0 \leq p \leq 1.$$ 

By definition the Lorenz curve takes values between 0 and 1 with $L(0) = 0$ and $L(1) = 1$. In addition the Lorenz curve cannot rise over the identity line, in this context also labeled line of equality, and is invariant under positive scaling, meaning that the Lorenz curve for $c \cdot X$ is the same as for $X$ for positive numbers $c$. Considering an ordered sample $x(1) \leq x(2) \leq \ldots \leq x(n)$ the empirical Lorenz curve is generated by the points

$$L(0) = 0$$

$$L\left(\frac{i}{n}\right) = \frac{s_i}{s_n}, \quad i = 1, \ldots, n,$$ 

with

$$s_i = \sum_{j=1}^{i} x(j).$$
The empirical Lorenz curve is defined by the linear interpolation of the points \((\frac{i}{n}, L(\frac{i}{n}))\), \(i = 1, \ldots, n\), and represents the fraction of the sum over the data that the smallest \(\frac{i}{n}\) fraction of the data points expresses. For more detailed information on the Lorenz curve see also Gastwirth [1972] and Gastwirth [1971].

Figure 14 shows the Lorenz curve for simulated data generated from a log-normal distribution. Interpreting the data sample as income or any other hold good of a population sample one can use the Lorenz curve to directly describe the distribution of this good over the sample population. For example, in Figure 14 the lower 40% of the data only hold 12.5% of the total good. Furthermore, the lower 90% hold 65% of the total good which implie that the upper 10% hold 35% of the total good.

The Lorenz curve can be put in relation to a very common measure of inequality, the Gini coefficient. A possible way to define the Gini coefficient is by the ratio between the area between the line of equality and the Lorenz curve, area \(A\) in Figure 15 and the are
Figure 15: Lorenz curve for data, generated from a log normal distribution, with marked area of concentration beneath the line of equality, which equals $\frac{1}{2}$. The area between the line of equality and the Lorenz curve is also called area of concentration.

An alternative definition of the Gini coefficient is proposed by Kendall and Stuart [1963] which is based on the mean difference of the underlying distribution function $F(x)$. Let $X$ be a random variable with distribution function $F(x)$. Then the Gini coefficient is defined by

$$Gini = \frac{\Delta}{2\mu} \quad \text{with}$$

$$\Delta = \int_{\mathbb{R}} \int_{\mathbb{R}} |x - y| dF(x) dF(y)$$

$$= 4 \int x \left[ F(x) - \frac{1}{2} \right] dF(x).$$
Given a data sample $x_1, \ldots, x_n$ with sample weights $w_1, \ldots, w_n$ the Gini coefficient can be estimated by [see also Alfons and Templ, 2013]

$$\hat{Gini} = \frac{2 \sum_{i=1}^{n} (w_i x_i \sum_{j=1}^{i} w_j) - \sum_{i=1}^{n} w_i^2 x_i}{(\sum_{i=1}^{n} w_i) \sum_{i=1}^{n} w_i x_i} - 1.$$  

The Gini coefficient is still well defined for negative values in the data, but in practice this case is often left out.

The Gini coefficient ranges between 0 and 1. A value of 0 stand for perfect equality, meaning that every data point in the sample has the same value or every individual has the same income or volume of certain good. With a value of 1 the Gini coefficient would indicate perfect inequality, meaning that all but one data point are equal to zero or that one individual has all the income or volume of certain good.

Apart from the Gini coefficient many other measures of inequality are used like the **quintile share ratio (QRS)** [see Eurostat, 2004], the **Atkinson Index** [see Atkinson, 1970] and many more. Since the Gini coefficient is well studied and has a long history of measuring distribution inequality it can be seen as the most popular inequality measure.
5 Empirical analysis

In this section univariate and multivariate outlier detection methods are applied on household expenditure data from various countries. These household expenditure data contain the yearly consumption, expressed in monetary value, of any good of the surveyed households. 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 that influence the Gini coefficient [see also Alfons et al., 2013]. Therefore it would 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 estimated. It is not known how many extreme values or measurement errors are contained in the household expenditure data. Moreover, the correct 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 on how well these methods perform, a sensitivity analysis as well as simulations are conducted. The results should provide evidence on which methods are more suited 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 should be taken into account.

The calculations in this work have all been done using the programming language R and a variety of already 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.

5.1 Provided data and data structure

The household expenditure data used in this thesis is comprised of household surveys from five different 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 and are freely accessible on their web page\(^1\).

These data sets were originally taken from the national data producers and are the product of large household surveys. The surveys include sociodemographic characteristics of

\(^1\)http://data.worldbank.org
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 for education, health and others. The number and type of categories differ from each survey but have in common that the categories combined reflect most of the consumption of a household for a given time horizon.

5.1.1 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, for instance expenditures for health over a time horizon of a week or month.

Another part of the harmonization process which brought difficult decisions was the mapping of the household expenditures of each survey onto a standardized framework for goods and services namely the ICP basic headings. 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. For most surveys the questionnaire did not cover so many different goods and services such that all the ICP basic headings were covered. This results in missing values for which their share is rather large in some cases. 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, meaning 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].
5.1.2 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 participants, the other corresponds to individual information of the members of each household and the last file corresponds to the consumption of each household. The data sets also contain household and population weights. The calculations in this work mainly focus on the household expenditures and household characteristics.

Regarding household characteristics the information used is limited to geographical characteristics and household weights. The data contains the annual value of goods or services consumed in local currency for each household. 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. Therefore it is not possible to differentiate if those are real zeros or missing values. How this is considered for the calculations is discussed in Section 5.1.4 and Section 5.3.1.

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, as discussed beforehand. This kind of incompleteness of the data set can result in some problems depending on the type of analysis that has to 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 is our preferred approach.

Even when only analyzing the 13 main expenditure categories the amount of incomplete entries can be quite high for some categories, meaning that for the data sets from the five different countries mentioned above, the amount of missing entries for some categories is more than 50% of the corresponding sample size, in some cases even over 80% or 90%. For some of the outlier detection methods, which are presented, these shares are too large and they are not able to produce any result. To not render these methods useless an R routine was implemented to combine certain categories and therefore reduce the number of missing entries. The main objective of this routine lies in the comparison of the interquantile range (IQR) for the expenditures of each category. If the IQRs of two
or more categories overlap ‘well enough’ then those categories are combined. This re-
grouping scheme does not take into account a regrouping which results in minimal share
of missing values. Furthermore this regrouping scheme represents solely a suggestion
and is not proven to work in every case.

5.1.3 Data structure

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 is 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 sample weight, 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 are denoted as 'not available’ (NA).

5.1.4 Selected data and further data preparation

In the following, univariate and multivariate outlier detection methods are applied to 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 house-
hold survey. One reason for this is that the Albanian household survey has a reasonable
sample size with 3600 household, including 14785 individuals and the data set is fairly
complete. 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 which was discussed in the previous section. To give a better understand-
ing on the incompleteness of the data Table 3 shows the number of missing entries for
each category in the original data set. The amount of missing values for the Albanian
household survey is similar to the other four surveys.

Missing values are treated in different ways whether univariate or multivariate outlier de-
tection methods are applied. Univariate outlier detection methods will either be applied
on each of the expenditure categories separately or on the accumulated expenditures for
every household. In order to calculate the sum over all expenditures per household miss-
Table 3: Number of missing entries per category for the Albanian household survey

<table>
<thead>
<tr>
<th>Category</th>
<th>Missing entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food and non-alcoholic beverages</td>
<td>2</td>
</tr>
<tr>
<td>Alcoholic beverages, tobacco and narcotic</td>
<td>1476</td>
</tr>
<tr>
<td>Clothing and footwear</td>
<td>347</td>
</tr>
<tr>
<td>Housing, water, electricity, gas and other fuels</td>
<td>25</td>
</tr>
<tr>
<td>Furnishings, household equipment, household maintenance</td>
<td>2</td>
</tr>
<tr>
<td>Health</td>
<td>1264</td>
</tr>
<tr>
<td>Transport</td>
<td>1468</td>
</tr>
<tr>
<td>Communication</td>
<td>407</td>
</tr>
<tr>
<td>Recreation and culture</td>
<td>19</td>
</tr>
<tr>
<td>Education</td>
<td>3278</td>
</tr>
<tr>
<td>Restaurants and hotels</td>
<td>1814</td>
</tr>
<tr>
<td>Miscellaneous goods and services</td>
<td>114</td>
</tr>
<tr>
<td>Net purchases abroad</td>
<td>3600</td>
</tr>
</tbody>
</table>

Entries for incomplete observations will be set to zero. For the Albanian household survey the annual total consumption for each household is greater than zero, meaning that every household in the survey presented information on their own expenditure. For other surveys the number of households which participated in the survey and the number of household which specified their own expenditure habits may not be the same. In that case it would be recommended to exclude those households from the outlier detection for which there is no expenditure information in the data set and not include them as a household with zero spendings per year.

When applying univariate outlier detection methods on each of the categories separately, missing values of each category will be discarded for outlier detection only.

Note that excluding these households from the calculations also influences the corresponding household weights, meaning that the household weights, given in the data set, do no longer add up to the whole population size, if some of the household are discarded. 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.

The treatment of missing values for multivariate outlier detection methods is described in Section 5.3.1.

5.2 Univariate methods

The univariate outlier detection methods which are tested on the data are the boxplot, adjusted boxplot [Vandervieren and Hubert, 2008], methods which use the rule of robust location ± constant × scale in conjunction with the Box-Cox transformation [Box and
Cox, 1964, Marazzi and Yohai, 2006] and the Pareto tail modeling [Alfons et al., 2013] as described in Section 2. As mentioned in Section 2, the constant, for the use of IQR or MAD for outlier detection is $c = 3$. Figure 16 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 16 also displays the weighted median and the Pareto threshold beyond which the Pareto distribution is fitted onto 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 ($BC+ML$) and the robustification of the Box-Cox transformation ($BC+rob$). Meaning that the outlier detection schemes have been conducted after the data has been transformed. After that the calculated bounds for potential outliers were transformed back and the potential outliers were identified. Beneath the $x$-axis the boxplot and adjusted boxplot are displayed.

To give a better understanding of the calculated bounds, Table 4 shows the upper and lower bounds beyond which potential outliers can be detected for the above mentioned univariate outlier detection schemes. In context the Pareto tail modeling the displayed value represents the Pareto threshold which indicates the point beyond where the Pareto distribution is fitted. In Figure 16 it is easy to see, 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 perform poorly or can be expected to be not suited for the problem.

Figure 16: Univariate outlier detection methods used on the total annual consumption of the Albanian household survey.
Table 4: Upper and lower bounds for univariate outlier detection schemes applied on the total annual consumption of the Albanian household survey.

<table>
<thead>
<tr>
<th>Method</th>
<th>Upper Bound</th>
<th>Lower Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>IQR</td>
<td>1586292</td>
<td>-392883.90</td>
</tr>
<tr>
<td>Box-Cox with IQR</td>
<td>3625201</td>
<td>143191.19</td>
</tr>
<tr>
<td>robust Box-Cox with IQR</td>
<td>2761414</td>
<td>113541.96</td>
</tr>
<tr>
<td>MAD</td>
<td>1502748</td>
<td>-309339.27</td>
</tr>
<tr>
<td>Box-Cox with MAD</td>
<td>3562527</td>
<td>144762.24</td>
</tr>
<tr>
<td>robust Box-Cox with MAD</td>
<td>2666762</td>
<td>118274.35</td>
</tr>
<tr>
<td>Boxplot</td>
<td>1365380</td>
<td>-68460.19</td>
</tr>
<tr>
<td>adjusted Boxplot</td>
<td>2342039</td>
<td>235526.95</td>
</tr>
<tr>
<td>Pareto Modeling</td>
<td>1937270</td>
<td></td>
</tr>
</tbody>
</table>

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 suited for the problem since their lower bounds are negative which is, in context with the type of data, a result that is not valid. 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 don’t seem to be so strict, in the case of the upper boundary. The same can be said about the adjusted boxplot.

One of the main reasons to apply outlier detection for this kind of data is the estimation of the Gini coefficient afterwards. Since the original data set is suspected to contain outliers, a classical calculation of the Gini coefficient would be highly influenced and produce arbitrary results. Therefore locating and adjusting potential outliers would be a desired method to produce more reliable results for the Gini coefficient. The above presented univariate outlier detection schemes declare a different number of data points to be potential outliers and since the number or 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, looking at the different values for the Gini coefficient after potential outlier detection and adjustment gives inside on how strongly these detection schemes influence the level of the Gini coefficient.

Figure 17 shows on the top left side the estimated values for the Gini coefficient after outlier detection schemes have been applied and potential outliers have been adjusted. *IQR, MAD, box* and *adjbox* indicate the use of the interquantile range, the median
Figure 17: Top: Estimates for Gini coefficient (left) and variance of Gini coefficient (right) of Albanian data set after univariate outlier detection methods as well as outlier imputation have been applied.
Bottom: Share of upper and lower outlier for each outlier detection scheme applied on Albanian data set.

absolute deviation, the boxplot or the adjusted boxplot for outlier detection. The abbreviations \(bc\) and \(bcrob\) indicate that in these cases the Box-Cox transformation and the robustification of the Box-Cox transformation, respectively, were applied together with the IQR or MAD. 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 for the other observations have been recalibrated 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 17 the variance of the estimated Gini coefficients are displayed for every univariate outlier detection scheme. The confidence interval of the Gini coefficient and the variance were calculated.
using a bootstrapping routine, with 100 bootstrap replicas, in which the sample weights are recalibrated using geographical information provided by the data sets. 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 to it is also displayed. The bottom of Figure 17 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 17 it is clear that the corresponding Gini values are heavily influenced after the adjustment of these potential outliers. Interesting to see is that, leaving out the 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 rather seems 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 outlier and since the data is skewed to the right upper outlier 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 inherits 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.

**A note on sensitivity analysis**

To better evaluate the performance of the univariate outlier detection methods a sensitivity analysis was conducted, meaning that the weighted upper 1% of the Albanian data on expenditure per household were shifted further away from the rest of the data over a series of steps. In every step the outlier detection methods were applied and the Gini coefficient is estimated after the imputation of potential outliers. The resulting values show the influence on the detection methods and on the estimated Gini coefficient after imputation. Since the outlier detection methods are based on robust estimates it should be suspected that the step-wise shift of a small portion of the data does not have an effect on the robust estimates and therefore no effect on outlier detection and estimation of Gini coefficient afterwards. The calculated values of the Gini coefficient were in accordance to this assumption and showed little to no influence of the step wise shift onto the estimated Gini coefficient. The results of this sensitivity analysis are not presented in this work, since they don’t yield usefull information, which should be suspected since the outlier detection methods are based on robust estimates.

**5.2.1 Results of univariate outlier detection methods for different countries**

Univariate outlier detection methods were also applied on the household expenditure data from the other countries namely, India, Mexico, Malawi and Tajikistan. The results, including estimates of Gini coefficient and variance of Gini coefficient as well as number of potential outliers and upper and lower bound are listed in Table 5.
<table>
<thead>
<tr>
<th>Country</th>
<th>Number of households</th>
<th>Gini</th>
<th>Variance Gini</th>
<th>Upper bound</th>
<th>Lower bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albania(2008)</td>
<td>3600</td>
<td>31.9516</td>
<td>0.5634</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31.3368</td>
<td>0.1328</td>
<td>(121)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30.7868</td>
<td>0.4144</td>
<td>(12)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>27.6850</td>
<td>0.3456</td>
<td>(25)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31.3132</td>
<td>0.1167</td>
<td>(141)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31.6646</td>
<td>0.3154</td>
<td>(0)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>26.8415</td>
<td>0.1278</td>
<td>(27)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30.1016</td>
<td>0.2865</td>
<td>(0)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31.7374</td>
<td>0.4891</td>
<td>(0)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31.6646</td>
<td>0.4891</td>
<td>(0)</td>
<td>(0)</td>
</tr>
<tr>
<td>India(2009)</td>
<td>100852</td>
<td>39.8225</td>
<td>0.0780</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31.5778</td>
<td>0.0137</td>
<td>(9131)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>38.8646</td>
<td>0.0294</td>
<td>(842)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32.7119</td>
<td>0.0358</td>
<td>(349)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>38.2566</td>
<td>0.0127</td>
<td>(1143)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>38.8819</td>
<td>0.0292</td>
<td>(608)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31.8037</td>
<td>0.0358</td>
<td>(35)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>36.7449</td>
<td>0.0120</td>
<td>(1334)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>39.4401</td>
<td>0.0154</td>
<td>(2298)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>39.4401</td>
<td>0.0541</td>
<td>(32)</td>
<td>(0)</td>
</tr>
<tr>
<td>Mexico(2010)</td>
<td>27655</td>
<td>44.1972</td>
<td>0.1394</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>37.6173</td>
<td>0.0224</td>
<td>(229158.88)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>43.4220</td>
<td>0.0000</td>
<td>(710869.91)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>43.7338</td>
<td>0.1098</td>
<td>(948703.75)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>36.5904</td>
<td>0.0197</td>
<td>(207669.64)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>43.3171</td>
<td>0.0898</td>
<td>(700041.53)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>43.7228</td>
<td>0.1095</td>
<td>(495857.40)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35.8057</td>
<td>0.0184</td>
<td>(195190.25)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>41.6716</td>
<td>0.0559</td>
<td>(440889.78)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>44.0889</td>
<td>0.1327</td>
<td>(328053.9)</td>
<td>(0)</td>
</tr>
<tr>
<td>Malawi(2010)</td>
<td>12096</td>
<td>48.5196</td>
<td>0.5180</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>36.1359</td>
<td>0.0352</td>
<td>(54619.7)</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>47.2401</td>
<td>0.2386</td>
<td>(340824.98)</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>45.5800</td>
<td>0.2185</td>
<td>(199245.0)</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>45.2068</td>
<td>0.0322</td>
<td>(82419.7)</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>47.2501</td>
<td>0.3209</td>
<td>(322054.6)</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>45.5905</td>
<td>0.2195</td>
<td>(2003072.6)</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>34.2140</td>
<td>0.0299</td>
<td>(454713.5)</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>41.3443</td>
<td>0.0095</td>
<td>(443007.78)</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48.5196</td>
<td>0.5180</td>
<td>(328053.9)</td>
<td>–</td>
</tr>
<tr>
<td>Tajikistan(2007)</td>
<td>4860</td>
<td>33.1176</td>
<td>0.3812</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>28.3864</td>
<td>0.0718</td>
<td>(27682.19)</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31.4923</td>
<td>0.1799</td>
<td>(52411.20)</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32.7892</td>
<td>0.3693</td>
<td>(142057.94)</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>28.3344</td>
<td>0.0713</td>
<td>(26682.26)</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31.3102</td>
<td>0.1270</td>
<td>(52655.35)</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32.7730</td>
<td>0.3397</td>
<td>(134913.97)</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>27.2635</td>
<td>0.0607</td>
<td>(21610.47)</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>29.8537</td>
<td>0.1034</td>
<td>(36601.04)</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32.3064</td>
<td>0.2753</td>
<td>(37286.99)</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 5: Results of univariate outlier detection methods applied on the household expenditure data of various countries.
5.2.2 Column-wise application of univariate methods

Applying univariate outlier detection methods on the data of total annual expenditure is a very straightforward and simple approach to detect potential outliers. One of the drawbacks with this approach is the disregard to the multidimensional structure of the data set. Due to that, households which have listed erroneous information for a specific good or service might be overlooked since their total annual expenditure do 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 missing values will be left out, meaning that if a household has 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. In this context missing values will be treated as zeros. This approach as been applied on the Albanian data set with the use of the ICP category code described in Section 5.1. The resulting Gini coefficients and statistics for potential outliers are displayed in Figure 18. The left hand side of Figure 18 shows the resulting Gini coefficient and corresponding 95% confidence interval, scaled on the left axis and the variance of the Gini coefficient scaled on the right axis. The right side of the plot shows the share of observations which were flagged as potential outliers in the whole data set. The color coding indicates if an observation was flagged once or more during the outlier detection scheme. This occurs if an observation has extreme values in more than one category which are then each declared as potential outliers.

As with the use of the univariate outlier detection methods used on the total annual expenditures of each household, the results displayed in Figure 18 suspect that methods which do not account for skewed data are not suited for this kind of data. The share of observations which have been flagged as potential outliers in at least one category are beyond 50%. Therefore it is questionable if these univariate methods, applied on each column of the data, produce reliable results at all. Also the outlier detection method using the adjusted boxplot does declare a suspiciously high amount of observations as potential outliers. The methods using the Box-Cox transformation and the Pareto tail modeling produce, based on the share of detected potential outliers, more reasonable results. The share of detected potential outliers are within the range of 10% and for the Pareto modelling 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 of one column must not correspond to the same observation as potential outliers of an other column. In a worst case scenario this column-wise approach could result in flagged potential outliers in at least one cell of every observation. To declare more then 50% of observations in a data set as potential outliers contradicts with the definition of an outlier as well as the definition of the breakdown point, which is a fundamental charac-
characteristic for statistical outlier detection methods. This leads to reason, that column-wise application of univariate methods on a data set, as it is presented in this work, does inherit major flaws.

Interesting to see is 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 reason 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.

Figure 18: Top: Estimates for Gini coefficient (left) and variance of Gini coefficient (right) of Albanian data set after column-wise outlier detection and imputation of outliers 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. Colour codes signal the frequency of which an observation was declared as potential outlier.
5.3 Multivariate methods

The multivariate outlier detection methods, which are 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 are tested on the provided data sets. To summarize, the used estimates or methods consist of the M-estimate ($\text{Mest}$), the MM-estimate ($\text{MMest}$), the S-estimate ($\text{Sest}$), the MCD estimator ($\text{Mcd}$), the MVE estimator ($\text{Mve}$), the Stahel-Donoho estimator ($\text{Sde}$), the OGK estimator ($\text{Ogk}$), the epidemic algorithm ($\text{EA}$) and the BACON-EEM ($\text{BEM}$). A description of all these algorithms can be found in Section 3. For the calculations of the methods which estimate location and covariance in a robust way the $\text{R}$-package $\text{rrcov}$ was utilized [see Todorov and Filzmoser, 2009]. As mentioned in Section 3 these methods can be tuned by choice of breakdown point and some methods by efficiency and sample correction factors. For the empirical calculations the default values for breakdown point and tuning parameters for efficiency or sample correction which are implemented in the $\text{R}$-package $\text{rrcov}$ were adopted. To calculate the BACON-EEM and the epidemic algorithm function the $\text{R}$-package $\text{modi}$ was used. The parameters for which the BACON-EEM and the epidemic algorithm were calculated are listed in Section 3 under the corresponding subsections.

Before applying the multivariate outlier detection methods onto the expenditure data matrices, from the various countries, it is important to consider, that most of these methods need the data to originate from multivariate normal distribution or at least be of elliptical shape. Therefore the data matrix on household expenditure will always be transformed by the natural logarithm prior to the outlier detection. After the outlier detection has been calculated and the potential outliers are imputed, the data matrix is transformed back using the exponential function for further calculations.

Another issue is the existence of missing values in the data sets. The majority of the multivariate outlier detection methods stated before can not deal with missing values in the data matrix. To overcome this issue, these missing values are imputed before calculating the estimates.

5.3.1 Imputation of missing values

Before the methods can be applied the data are transformed into matrix format, by use of the ICP category code, as discussed in Section 5.1. In this format the data contains many missing values which arise from not existing entries for expenditure on various goods and services for each household. Since the number of non existing entries is well over 50% for some categories the columns, containing the expenditure categories, were combined using the method mentioned in Section 5.1 which compares the IQR of each expenditure column to add two or more columns for lesser number of non existing entries. This was done to ensure that the algorithms for all the multivariate outlier detection methods run without any issues. This procedure is not essential to run the different methods and the routine does not need to produce a data set with less missing values in every case. It simply presents a suggestion to run the multivariate outlier detection
algorithms with possibly less complications.

In context with univariate outlier detection methods the missing entries where dealt with by treating them as zeros, for the detection of potential outliers on the total expenditures per household, or by discarding them in case of column-wise application of univariate outlier detection methods only. For the calculation of multivariate outlier detection, incomplete observations could be discarded, but as stated in the previous sections this would lead to heavy losses of information and is not recommendable. Among the multivariate outlier detection methods only the epidemic algorithm and the BACON-EEM can deal with missing entries in the data set. For the other multivariate outlier detection methods the missing entries are imputed prior to the outlier detection by 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 [Troyanskaya et al., 2001]. Let \( X \in \mathbb{R}^{n \times p} \) be a given data sample with \( p \) variables then the \( k \)-nearest neighbors of a new observation \( x_{n+1} \in \mathbb{R}^p \) are those \( \{x_{i_1}, \ldots, x_{i_k}\} \) for which the distances \( d(x_{n+1}, x_{j}), j = 1, \ldots, k \) are smallest. This approach makes it possible to classify data points by determining their class label by the labels of its \( k \)-nearest neighbors. As stated beforehand this methods was also used in many different ways to impute missing values in a data set. The kNN method used in this work is the one implemented in the \texttt{R}-package \texttt{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 distances between two observation \( x_i, x_j \in \mathbb{R}^p \) is defined as

\[
d(x_j, x_i) = \frac{\sum_{k=1}^{p} w_k \delta_{i,j,k}}{\sum_{k=1}^{p} w_k},
\]

with \( w_k \) as weight, which represent 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 contribution of each variable.

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

\[
\delta_{i,j,k} = \frac{|x_{i,k} - x_{j,k}|}{r_k},
\]

where \( x_{i,k} \) 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 \( x_i \in \mathbb{R}^p \) have one or more missing cells then for the imputation by the kNN algorithm, the \( k \)-nearest neighbors are calculated by using the previously mentioned

55
distance measure. The distances are only calculated between \( \mathbf{x}_i \) and observations without missing cells and for the calculations only variables which are non missing in \( \mathbf{x}_i \) are considered. Finally a missing cell of \( \mathbf{x}_i \) is imputed by using the \( k \) values of the nearest neighbors of the missing variable. For continuous variables the default option in Templ et al. [2012] is the median, although other statistics are possible.

### 5.3.2 Applying multivariate outlier detection methods on the Albanian data set

The multivariate outlier detection methods are applied onto the expenditure data from the Albanian data set and if the method can not deal with missing entries these entries are imputed by the kNN algorithm beforehand. Potential outliers are afterwards windsorized onto the 95\% tolerance ellipse, for which the robust estimates for covariance and location are used. In the case of the BACON-EEM this is the final covariance estimate from the data for which incomplete observations have been imputed by the EEM algorithm. The epidemic algorithm does not calculate a covariance during the task of detecting potential outliers. Nevertheless the imputation of potential outliers should be comparable between the different methods. Therefore, in the case of the epidemic algorithm, the covariance matrix, which is used for the imputation, is calculated by applying the classical estimate for the covariance onto 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 imputed for every detection method separately the estimate for the Gini coefficient is estimated. As in the case of univariate methods the Gini coefficient of the total annual expenditures per household is estimated. For this purpose data entries denoted by "NA" in the case of the BACON-EEM and epidemic algorithm are treated as zero. For the other multivariate outlier detection methods, those entries where imputed using the kNN algorithm prior to outlier detection. Since this imputation was done in order to calculate location and covariance in a robust way, these imputed values are replaced by zero for the estimation of the Gini coefficient. For the estimation of the Gini coefficient the household sample weights from the data set are included. Figure 19 shows the results for the estimates for the Gini coefficient and share of detected potential outliers for each multivariate outlier detection methods applied on the Albanian household expenditure data. In contrast to the case of univariate outlier detection methods (Figure 17 and Figure 18) the results for the Gini coefficients do not differ so much between the applied methods. Only the results for the OGK estimator and the epidemic algorithm differ slightly more from the results of the other methods. Also the share of detected potential outliers does not differ a lot between the methods, except for the OGK estimator and the epidemic algorithm. Compared to the column-wise application of univariate outlier detection methods the number of detected potential outlier is far lower for the multivariate outlier detection methods.

The results for the multivariate outlier detection methods are quite similar and it is therefore not clear which of the methods performed the 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 was conducted to see how the different methods performed on data, which was generated based on the Albanian household expenditure data (see Section 6).

5.3.3 Results of multivariate outlier detection methods for different countries

Multivariate outlier detection methods were also applied on the household expenditure data from the other countries, India, Mexico, Malawi and Tajikistan. The results, including estimates of Gini coefficient and variance of Gini coefficient as well as number of potential outliers are listed in Table 6. The Indian data on household expenditure presented some problems due to the size of the data set. Since the kNN algorithm for imputation can consume quite a lot of RAM this method for imputation might create memory issues for very large data sets. For the largest data set, India, with over 100000 households, the hot-deck algorithm for imputation was used [Templ et al., 2012].
<table>
<thead>
<tr>
<th>Country</th>
<th># households</th>
<th>Original Gini</th>
<th>MCD M estimate</th>
<th>MM estimate</th>
<th>MVE estimate</th>
<th>OGK estimate</th>
<th>Stahel Donoho S estimate</th>
<th>BACON EEM</th>
<th>Epidemic algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.0780</td>
<td>0.4369</td>
<td>0.4323</td>
<td>0.4346</td>
<td>0.4353</td>
<td>0.4365</td>
<td>0.4065</td>
<td>0.4300</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–</td>
<td>329</td>
<td>354</td>
<td>281</td>
<td>298</td>
<td>316</td>
<td>293</td>
<td>328</td>
</tr>
<tr>
<td>India (2009)</td>
<td>100852</td>
<td>44.1972</td>
<td>42.8459</td>
<td>42.7514</td>
<td>42.9073</td>
<td>42.7683</td>
<td>42.0050</td>
<td>42.7459</td>
<td>42.8977</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1394</td>
<td>0.1114</td>
<td>0.1100</td>
<td>0.1108</td>
<td>0.1069</td>
<td>0.0964</td>
<td>0.1101</td>
<td>0.1114</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–</td>
<td>2062</td>
<td>2218</td>
<td>1904</td>
<td>2029</td>
<td>3545</td>
<td>2241</td>
<td>1951</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5180</td>
<td>0.1213</td>
<td>0.1175</td>
<td>0.1301</td>
<td>0.1265</td>
<td>0.0913</td>
<td>0.1174</td>
<td>0.1242</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–</td>
<td>893</td>
<td>1015</td>
<td>874</td>
<td>885</td>
<td>1668</td>
<td>1023</td>
<td>884</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1854</td>
<td>0.1845</td>
<td>0.1908</td>
<td>0.1900</td>
<td>0.1617</td>
<td>0.1843</td>
<td>0.1907</td>
<td>0.1877</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–</td>
<td>483</td>
<td>500</td>
<td>406</td>
<td>441</td>
<td>776</td>
<td>505</td>
<td>413</td>
</tr>
</tbody>
</table>

Table 6: Results of multivariate outlier detection methods applied on the household expenditure data of various countries.

*Results not available, due to RAM limitations.*
6 Simulation study

It was already observed in the previous subsections and demonstrated using the Albanian data set that different outlier detection schemes deliver quite different results. In the case of univariate outlier detection schemes it is possible to rule out some of the methods based on unfavorable preconditions of the distribution of the data, like skewness. Nevertheless, for the rest of the used methods it is not clear which of them performed better or worse. This is also true for multivariate outlier detection methods. Most of them need the data to originate from multivariate normal distribution or at least be of elliptical shape. Therefore, the data set was transformed with the natural logarithm to make its distribution more conform with a multivariate normal distribution. Applying the multivariate outlier detection methods to this transformed data set delivers different results for every outlier detection method used. As with the univariate outlier detection methods it is not clear which of the multivariate methods perform well or produce unreasonable results.

The main reason for not being able to evaluate the performance of the different methods lies within the lack of knowledge of the number and position of the "real" outliers in the data set. In addition, the value for the "real" Gini coefficient is also not known. Therefore it is not possible to assess how well one or the other outlier detection schemes in combination with outlier imputation, adjusts the data set to at least produce reliable results for various calculations afterwards. Even if one would apply the outlier detection schemes on many different data sets from different countries there would be no way to evaluate the different outcomes, simply because the values or position of "true" outliers is not known in either of those data sets and information on reliable results for estimates after outlier detection and imputation does also not always exist.

To address this issue the following section presents a simulation study based on the data information from the Albanian data set to help assess the performance of the different detection schemes. Todorov et al. [2011] already looked into the performance of multivariate outlier detection methods and conducted a simulation study based on business survey data with incomplete observations.

6.1 Simulation setup

The aim of this simulation is to get a clearer picture on which of these outlier detection methods perform better or worse for large expenditure data sets. Therefore it is crucial for the simulation study to:

(a) Simulate such kind of data sets which can be comparable, regarding the data on household expenditure, with the ones provided by the World Bank.

(b) To know the number and position of "true" outliers beforehand.

Only then can the problem with assessing the performance of the different detection schemes be considered. For this reason the simulated data are generated without inheriting any outliers and afterwards this data set will be contaminated by artificial outliers.
6.1.1 Simulated data

We 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 are transformed using the logarithm since the data in each expenditure category is skewed to the right and many of the outlier detection schemes depend on data which is more or less conform with normal distribution respectively multivariate normal distribution.

6.1.2 Simulation of "clean" data and outliers

After transformation the data set is split into a "clean" data set which inherits most likely no outliers and a "contaminated" data set which inherits 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 inherits 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 outliers by the majority of outlier detection method. To be more precise, these are all 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 set the location and covariance are estimated in a classical way and are used as basis for the distribution of the simulated data set. If the clean or contaminated data set inherit any missing values, they will be imputed by the \( k \)-nearest neighbors algorithm, so that calculation of classical estimates for location and covariance is possible. To summaries this approach: first the data is simulated from a multivariate normal distribution, using the classical estimates from the clean data set. Afterwards a share of the observations is swapped with contaminated values, generated using the classical estimates for location and covariance of the contaminated data set. This resulting simulated data set follows a classical contamination scheme. More precisely, let \((\mu_{cl}, \Sigma_{cl})\) and \((\mu_{co}, \Sigma_{co})\) be classical estimates for location and covariance of the clean and contaminated data set, respectively, then the simulated data set \(X\) follows the following distribution

\[
X \sim (1 - \epsilon) MVN(\mu_{cl}, \Sigma_{cl}) + \epsilon MVN(\mu_{co}, \Sigma_{co})
\]

with \(\epsilon \in (0, 1)\) determining the share of contaminated data points. Since the Albanian data set was taken as basis for this simulation, the simulated data set consists of 3600 observations as well. For the contamination, observations were picked at random and replaced by the data simulated from a multivariate normal distribution with \((\mu_{co}, \Sigma_{co})\). In Hulliger et al. [2011] outlier mechanism were also discussed but in a different and more detailed way. Besides replacing whole observations with contaminated data, the contamination scheme also involved contaminating only a cell of an observation, meaning that for a share of
all the randomly picked observations which should be contaminated, not the whole ob-
ervation but just one cell, chosen randomly, is contaminated. Given the \(i^{th}\) observa-
tion for which only the \(j^{th}\) cell will be contaminated, the cell is replaced by \(y_{ij}\), with \(Y \sim MVN(\mu_{co}, \Sigma_{co})\). Doing so this contamination scheme simulates data which also
contains outliers in only a cell of an observation. This addition to the contamination
scheme was also implemented since such structures can be expected for real expenditure
data sets and there is no real reason why such kind of contamination should not occur.

6.1.3 Inclusion of missing values

Another aspect for the data simulation is the inclusion of missing values, since, as dis-
cussed beforehand, the number of missing observations can be quite high and play quite
a big role for the analysis on such data sets. The simulated data will therefore inherit
missing values. The placement of these missing values are the same as observed in the
Albanian data set and they will be included after the data has been generated. By re-
placing values in the simulated data set with missing values, it can occur that previously
generated artificial outliers will be replaced by those missing values. Since the data are
simulated many times and that the placement of missing values overlapping with the
randomly chosen contaminations 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.

6.1.4 Application of univariate Methods

After data have been simulated, as described above, univariate outlier detection methods
are applied to the data set. Since the simulated data was generated by multivariate
normal distribution based on the log transformation of the original Albanian data set,
the simulated data is transformed with the exponential function to generate skewed
data, that have nearly the same distribution as the original data sets. After that, the
univariate outlier detection methods are applied in the way that each of the outlier
detection schemes as well as imputation of potential outliers are applied on each column
of the data set. The main reason for this approach lies in the fact that this way the results
are more comparable between the use of univariate and multivariate outlier detection
methods. 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 the potential outliers are flagged by each method the
potential outliers will be imputed. Note that for the univariate outlier detection methods
the missing values are discarded for calculating and imputing potential outliers. Since
the position of the artificial outliers is known beforehand the number of successfully
identified outliers, for each outlier detection method, is counted afterwards. Moreover,
the number of falsely declared potential outliers are also counted. Imputing the potential
outlier 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 Gini coefficient is calculated. To be in accordance with the approach of the previous
section the weighted Gini coefficient of the total sum of each observation is calculated.
For summing over the columns of the generated data the previously included missing
values will be treated as zeros. In addition the sample weight are used to calculate the
weighted Gini coefficient. Besides calculating the weighted Gini coefficient for each of the
data sets which result from applying one of the univariate outlier detection schemes 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 imputation have been applied on a contaminated
data set.

6.1.5 Application of multivariate Methods

In the case of multivariate outlier detection methods the simulated data are not trans-
formed by the exponential function prior to outlier detection, simply because the pre-
sented multivariate outlier detection methods require data which agrees with multivariate
normal distribution, i.e., the data would be otherwise transformed with the logarithmic
transformation in any case. In contrast to the univariate outlier detection methods, not
all multivariate outlier detection methods can deal with missing values in the data set.
Only the epidemic algorithm and the BEM are able to handle missing values. For the
other multivariate outlier detection methods the missing values will be imputed, prior
to outlier detection, by using the \( k \)-nearest-neighbor algorithm. The imputed missing
values will only be used for the outlier detection methods and after outlier detection
the imputed values will be replaced with missing values. After the missing values are
imputed, or in the case of the epidemic algorithm and BEM without imputing missing
values, the multivariate outlier detection methods are applied on the simulated data
sets and observations which are flagged as potential outliers in this process will be im-
puted. As in the case of univariate outlier detection methods the number of correctly
identified artificial outliers are counted after the multivariate outlier detection methods
have been applied. Also the number of falsely declared potential outliers are counted.
As mentioned previously applying the outlier detection methods and imputing potential
outliers generates new data sets which correspond the each of the used outlier detec-
tion methods. Finally 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 with the exponential
function 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 mentioned for the application of univariate methods the
Gini coefficient will also be calculated for generated data without applying any 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 imputing potential outliers on contaminated data sets.

6.2 Simulations results

For the simulation study, multivariate data were generated and outlier detection methods were applied, as described previously. For the following results this procedure was repeated 50 times with different levels of $\epsilon$, $\epsilon \in \{0, 0.01, 0.025, 0.05\}$. As discussed for the outlier simulation 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. For the simulation 1/3 of the contamination is cell-wise and for 2/3 of the contaminated data the whole observation is contaminated. At first the results for the univariate outlier detection methods are discussed.

6.2.1 Results from univariate methods

Figure 20 shows boxplots of the resulting Gini coefficients for each method and each level of $\epsilon$ after the simulation. In the case where no outlier detection method was applied (original) the boxplot was only plotted for $\epsilon = 0$ since for other levels of $\epsilon$ the boxplots would cover a far greater range which makes the results for the applied outlier detection methods harder to read. From Figure 20 it is interesting to see that for higher values of $\epsilon$ 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 imputation the effect of the potential outliers should be negated. However, for the outlier imputation, except in the case of the Pareto modeling, the potential outliers are winsorised onto the interval boundaries, whereas these boundaries are calculated during the detection methods, imputed 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 one with no contamination and no outlier detection scheme applied.

Apart from the resulting Gini coefficient it is of great interest to see how many artificial outliers were successfully detected. Figure 21 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 $\epsilon$. The legend to the right side of the plot indicates how many artificial outliers were generated in total. The plot indicates that the methods were not at all successful in detecting outliers and that the results got worse for higher values of $\epsilon$. 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 counter parts which do not use this transformation. In the case of contamination, where only a single cell
Figure 20: Boxplots of calculated Gini coefficients for different outlier detection methods and different levels of $\epsilon$.

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

was contaminated, the outlier detection methods were not very successful as well. The results are plotted in Figure 22. In contrast to the row-wise contaminations the results for the cell-wise show not so drastic differences in regard to the performance between
the univariate methods. In the case of Pareto modeling the algorithm seems to have not performed to well when it comes to detection these artificial outliers. 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 artificial contaminated observations. Figure 22 shows the corresponding boxplots for different outlier detection methods and different levels of $\epsilon$. The $x$-axis corresponds to the share of flagged outliers to the total amount of clean data in the simulated data set. Except for methods like the Pareto modeling or methods which incorporate the use of the Box-Cox transformation the number of falsely flagged outliers is especially 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 outlier. 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 when it comes to outlier detection. From the results shown

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

65
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 very well and the number of falsely flagged potential outliers was far too high in almost all cases.

6.2.2 Results from multivariate methods

Similar to the univariate outlier detection methods, Figure 24 shows the boxplots of the Gini values for the different outlier detection methods and different levels of $\epsilon$. It is easy to see 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 divers depending 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 winsorising it onto 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 then 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 $\epsilon$. Apart from that the results for the multivariate outlier detection methods does, even for higher levels of $\epsilon$, not differ too much from the case where the data were not contaminated and
no outlier method was applied. Figure 25 and 26 show the number of correctly identified artificial outliers by the multivariate outlier detection methods. Figure 25 corresponds to artificial outliers for which the whole observation was contaminated and Figure 26 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 as the univariate methods. In many cases the algorithms were able to detect every artificial outlier and even for rising values of $\epsilon$ the numbers are still very high. The epidemic algorithm did not perform too well, but as stated earlier this is due to poor calibration of parameters, which is in practice a very cumbersome task. For the case of falsely flagged potential outliers for multivariate outlier detection methods, Figure 27 shows the resulting boxplots for different outlier detection methods and different levels of contamination. Compared to the univariate case the number of falsely flagged potential outliers is far less for the multivariate outlier detection schemes. The OGK estimator seems to perform not so well and has a higher number of falsely flagged potential outliers then the other methods. The majority of the multivariate outlier detection schemes have for $\epsilon = 0.01$ roughly the same amount of falsely flagged potential outlier. Under these six methods, increasing levels of $\epsilon$ result for the MM-estimate or the S-estimate to produce the least amount falsely flagged potential outlier. Looking at all multivariate outlier detection methods, except the epidemic algorithm that was ruled out as valid method beforehand, the BEM delivers the least amount of falsely flagged potential outliers. This and the fact that the BEM was also successful at identifying artificial outliers, leads to reason that the BEM performed best in this simulation study.
Figure 25: Boxplots of successfully detected artificial outliers, where whole observation was contaminated, for different outlier detection methods and different levels of $\epsilon$.

Figure 26: Boxplots of successfully detected artificial outliers, where only single cells were contaminated, for different outlier detection methods and different levels of $\epsilon$. 

68
Figure 27: Boxplots of share of false/positive outliers to number of clean data points for different outlier detection methods and different levels of $\epsilon$. 


7 Conclusion

This master thesis discusses the use of various outlier detection methods, which can deal with either univariate or multivariate data. In particularly these methods are applied on household expenditure data. The underlying data was provided by the World Bank and is comprised of extensive household surveys from 5 different countries. Besides detailed information on the sociodemographic parameters of the survey participants, the data also contained yearly household expenditures categorized into a multitude of classes, which cover every expense that can occur for the households. Since those surveys are quite large and the data collected is subject to human error, outlier detection seems reasonable to ensure better data quality. Apart from that, the data is suspected to contain true but extreme values and classical estimates then may deliver unreasonable results. A major topic in this thesis is the question which of the outlier detection methods are most suitable to detect outliers in the presented data sets. This issue was especially problematic since the true position of outliers in the underlying data sets is completely unknown. Another challenge was presented by the large part of missing data for the household expenditure data. It is not known if the missing data points in the household expenditure data are actually missing values or true zeros. Depending on the outlier detection method these missing values are treated differently.

Another aspect in this thesis is the estimate of the Gini coefficient of the total annual household expenditures prior and after outlier detection methods as well as imputation of potential outliers. The Gini coefficient presents a popular measure of inequality and is typically applied to income or expenditure data. Estimating the Gini coefficient prior and after outlier detection gives inside on how the outlier detection and imputation methods influence the estimate and presents a base of comparison for the different methods.

The used outlier detection methods are comprised of univariate and multivariate outlier detection methods. Most of the outlier detection methods are based on robust estimates of location and scale of the underlying data.

The univariate outlier detection methods are, due to their nature, applied onto the vector of total annual expenditures of each household in each of the data sets respectively. Those outlier detection methods include the use of household sample weights and are comprised of methods which utilize the IQR or MAD together with the median. Since household expenditure data is typically skewed to the right these methods are extended by the use of the Box-Cox transformation and a robustification of the Box-Cox transformation. In addition the univariate outlier detection methods include the boxplot, the adjusted boxplot and Pareto Tail modeling. For most of the univariate methods the detected potential outliers are winsorised onto the boundaries which surround the "good" part of the data. In the case of the Pareto Tail modeling the potential outliers are either replaced or their sample weights are set to 1. When using univariate outlier detection methods the missing values in the data set where treated as zeros and if a household would have no expenditures listed in the data it is left out of the outlier detection. Although the univariate outlier detection methods are quite straight forward and easy to implement they do discard the multidimensional structure of the expenditure data if...
they are applied onto the total annual household expenditures. Therefore these outlier
detection methods and as well as imputation are also applied on each of the expenditure
classes separately. Subsequently the Gini coefficient of the total annual expenditures
was calculated. The results showed, however, that this kind of application for univari-
ate outlier detection method is not really suitable since detected potential outliers after
outlier detection can occur in more than 50% of the households. In the worst case poten-
tial outliers can be detected in the household expenditure categories for every household.
The multivariate outlier detection methods discussed in this thesis include the M-estimate,
the S- and MM-estimate, the MCD-estimate, the MVE-estimate, the Stahel-Donoho es-
timate, the OGK-estimate, the epidemic algorithm and the BACON-EEM. Apart from
the epidemic algorithm and the BACON-EEM the outlier detection methods can not
handle missing values in the data. Therefore the missing values in the data are imputed
via the $k$-nearest neighbor algorithm prior to outlier detection with these methods. Inter-
preting the missing values as true $0$ is not recommended since it would heavily influence
the results of the outlier detection methods. Detected potential outliers are, in the case
of multivariate outlier detection methods, winsorised onto the 97.5% tolerance ellpisoide
which is created by the resulting estimates for location and covariance of the outlier
detection methods. Subsequently the Gini coefficient is calculated for the total annual
household expenditures for a better comparison between the multivariate outlier detec-
tion methods.

Although applying various outlier detection methods onto the data sets and comparing
estimates for the Gini coefficient as well as the share of detected potential outliers it
is not clear which of the univariate or multivariate methods performs best. To address
this issue, a simulation study was conducted based on the information of the Albanian
household expenditure data set. The results from this simulation study favored the
BACON-EEM to be the most suitable method for outlier detection on household expendi-
tures, since it detected in the one hand a very high share of artificial and true outliers
and in the other hand only few false/positive potential outliers.

It is however important to point out that for the simulation study, due to the nature of
the data and outlier simulation, the multivariate methods are more favored in contrast
to univariate methods. Furthermore the simulation study did not take into account
sociodemographic criteria like region, household size or age structure of the household.
Therefore the results of this simulation study should be looked at in a critical way.
For an even better understanding of the performance of the different outlier detection
methods the simulation study should include more information to simulate more realistic
households and population groups. Also the simulation of the artificial outliers in the
simulation needs finer tuning to deliver more convincing results.
References


R-Packages used

- R version 3.2.1 (2015-06-18), x86_64-w64-mingw32
- Base packages: base, datasets, graphics, grDevices, grid, methods, parallel, stats, utils
- Other packages: boot 1.3-17, car 2.0-25, colorspace 1.2-6, data.table 1.9.4, foreign 0.8-65, geor 1.7-5.1, ggplot2 1.0.1, laeken 0.4.6, lattice 0.20-31, MASS 7.3-42, modi 1.5, nnet 7.3-10, norm 1.0-9.5, perry 0.2.0, POT 1.1-3, reshape2 1.4.1, robustbase 0.92-4, robustHD 0.5.0, rrcov 1.3-8, simPopulation 0.4.1, vcd 1.4-1, VIM 4.3.0
- Loaded via a namespace (and not attached): chron 2.3-47, class 7.3-13, cluster 2.0.2, DEoptimR 1.0-2, digest 0.6.8, e1071 1.6-4, gtable 0.1.2, lme4 1.1-8, lmtest 0.9-34, magrittr 1.5, Matrix 1.2-1, mgcv 1.8-6, minqa 1.2.4, munsell 0.4.2, mvtnorm 1.0-2, nls 3.1-121, nloptr 1.0.4, pbkrtest 0.4-2, pcaPP 1.9-60, plyr 1.8.3, proto 0.3-10, quantreg 5.11, RandomFields 3.0.62, Rcpp 0.11.6, scales 0.2.5, sp 1.1-1, SparseM 1.6, splancs 2.01-37, splines 3.2.1, stats4 3.2.1, stringi 0.5-5, stringr 1.0.0, tcltk 3.2.1, tools 3.2.1, zoo 1.7-12