AUSTRIAN JOURNAL OF STATISTICS
Volume 34 (2005), Number 2, 127–138

Identification of Multivariate Outliers: A Performance Study

Peter Filzmoser
Vienna University of Technology, Austria

Abstract: Three methods for the identification of multivariate outliers (Rousseeuw and Van Zomeren, 1990; Becker and Gather, 1999; Filzmoser et al., 2005) are compared. They are based on the Mahalanobis distance that will be made resistant against outliers and model deviations by robust estimation of location and covariance. The comparison is made by means of a simulation study. Not only the case of multivariate normally distributed data, but also heavy tailed and asymmetric distributions will be considered. The simulations are focused on low dimensional ($p = 5$) and high dimensional ($p = 30$) data.

Keywords: Outlier Detection, MCD Estimator, Mahalanobis Distance, Robustness.

1 Introduction

The increasing size of data sets makes it more and more difficult to identify common structures in the data. Especially for high dimensional data it is often impossible to see data structures by visualizations even with highly sophisticated graphical tools (e.g. Swayne et al., 1998; Doleisch et al., 2003). Data mining algorithms as an answer to these difficulties try to fit a variety of different models to the data in order to get an idea of relations in the data, but usually another problem arises: multivariate outliers.

Many papers and studies with real data have demonstrated that data without any outliers ("clean data") are rather an exception. Outliers can–and very often do– influence the fit of statistical models, and it is not desirable that parameter estimations are biased by the outliers. This problem can be avoided by either using a robust method for model fitting or by first cleaning the data from outliers and then applying classical statistical methods for model fitting.

Removing outliers does not mean to throw away measured information. Outliers usually include important information about certain phenomena, artifacts, or substructures in the data. The knowledge about this deviating behavior is important, although it might not always be easy for the practitioner to find the reasons for the existence of outliers in the data, or to interpret them.

Multivariate outliers are not necessarily characterized by extremely high or low values along single coordinates. Rather, their univariate projection on certain directions separates them from the mass of the data (this projection approach for outlier detection was introduced by Gnanadesikan and Kettenring, 1972). Standard methods for multivariate outlier detection are based on the robust Mahalanobis distance which is defined as

$$\mathrm{MD}_i = \left((\boldsymbol{x}_i - \boldsymbol{t})^T \boldsymbol{C}^{-1} (\boldsymbol{x}_i - \boldsymbol{t})\right)^{1/2} \tag{1}$$

for a p-dimensional observation $\boldsymbol{x}_i$ and $i = 1, \ldots, n$. $\boldsymbol{t}$ and $\boldsymbol{C}$ are robust estimations of location and scatter, respectively. For normally distributed data (and if arithmetic mean and sample covariance matrix were used), the Mahalanobis distance is approximately chi-square distributed with p degrees of freedom (χ^2_p). Potential multivariate outliers $\boldsymbol{x}_i$ will typically have large values MD_i, and a comparison with the χ^2_p distribution can be made.

Garrett (1989) introduced the chi-square plot, which draws the empirical distribution function of the robust Mahalanobis distances against the the χ^2_p distribution. A break in the tail of the distributions is an indication for outliers, and values beyond this break are iteratively deleted until a straight line appears.

Rousseeuw and Van Zomeren (1990) use a cut-off value for distinguishing outliers from non-outliers. This value is a certain quantile (e.g., the 97.5% quantile) of the χ^2_p distribution. For $\boldsymbol{t}$ and $\boldsymbol{C}$ the MVE (minimum volume ellipsoid) estimator (Rousseeuw, 1985) was used. However, several years later the MVE was replaced by the MCD (minimum covariance determinant) estimator for this purpose which has better statistical properties and because a fast algorithm exists for its computation (Rousseeuw and Van Driessen, 1999).

Various other concepts for multivariate outlier detection methods exist in the literature (e.g. Barnett and Lewis, 1994; Rocke and Woodruff, 1996; Becker and Gather, 1999; Peña and Prieto, 2001) and different other robust estimators for multivariate location and scatter can be considered (e.g. Maronna, 1976; Davies, 1987; Tyler, 1991; Maronna and Yohai, 1995; Kent and Tyler, 1996).

Recently, Filzmoser et al. (2005) introduced a multivariate outlier detection method that can be seen as an automation of the method of Garrett (1989). The principle is to measure the deviation of the data distribution from multivariate normality in the tails. In Section 2 we will briefly introduce this method. A comparison with other outlier identification methods is done by means of simulated data in Section 3. Throughout the paper we restrict ourselves to the p-dimensional normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu}$ and positive definite covariance matrix $\boldsymbol{\Sigma}$, as model distribution. However, we also simulate data from other distributions in order to get an idea about the performance in different situations. Section 4 provides conclusions.

2 Methods

The method of Filzmoser et al. (2005) follows an idea of Gervini (2003) for increasing the efficiency of the robust estimation of multivariate location and scatter. Let $G_n(u)$ denote the empirical distribution function of the squared robust Mahalanobis distances MD_i^2, and let $G(u)$ be the distribution function of χ^2_p. For multivariate normally distributed samples, G_n converges to G. Therefore the tails of G_n and G can be compared to detect outliers. The tails will be defined by the quantile $\delta = \chi^2_{p,1-\beta}$ for a certain small β (e.g., $\beta = 0.025$), and

$$p_n(\delta) = \sup_{u \geq \delta} \Big(G(u) - G_n(u) \Big)^+ \tag{2}$$

is considered, where "+" indicates the positive differences. In this way, $p_n(\delta)$ measures the departure of the empirical from the theoretical distribution only in the tails, defined by the value of δ. If $p_n(\delta)$ is larger than a critical value $p_{crit}(\delta, n, p)$, it can be considered

as a measure of outliers in the sample. If this is not the case, the outlier measure it set to zero.

The critical value $p_{crit}(\delta, n, p)$ depends on the quantile δ, and on the size of the data set. It is derived by simulations as follows. Since we consider multivariate normal distribution as model distribution, samples with size n are simulated from the p-variate standard normal distribution. Then the outlier detection method is applied and for each simulated sample $p_n(\delta)$ is computed for a fixed value of δ. The critical value is then defined as a certain quantile $(1-\varepsilon)$ of all values $p_n(\delta)$ for a small value of ε, e.g. $\varepsilon = 0.05$.

Filzmoser et al. (2005) provide formulas for approximating the critical values for different n and p and for $\delta = \chi^2_{p,0.975}$.

One goal of this paper is to get an idea about the performance of the multivariate outlier detection method of Filzmoser et al. (2005). We decided to make a comparison with methods that are also based on robust Mahalanobis distances. Moreover, since the basis for the robust Mahalanobis distance is multivariate location and scatter estimation, we decided to use the MCD estimator (Rousseeuw, 1985) for this purpose.

One reference method for multivariate outlier detection is the method of Rousseeuw and Van Zomeren (1990) which uses fixed quantiles $\chi^2_{p,1-\varphi}$ as cut-off values for outliers. The other method is that of Becker and Gather (1999) which works somewhat different: A so-called α outlier with respect to $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is an element of the set

$$\text{out}(\alpha, \boldsymbol{\mu}, \boldsymbol{\Sigma}) := \{\boldsymbol{x} \in I\!R^p \ : \ (\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) > \chi^2_{p,1-\alpha}\} \tag{3}$$

which is also called α *outlier region*. The size of the outlier region is adjusted to the sample size n. This is done by including the condition that under the model, with probability $1-\alpha$, no observation lies in the outlier region $\text{out}(\alpha_n, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, and thus $\alpha_n = 1-(1-\alpha)^{1/n}$. An α_n outlier identifier is defined as a region

$$\text{OR}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \alpha_n) := \{\boldsymbol{x} \in I\!R^p \ : \ (\boldsymbol{x}-\boldsymbol{t})^\top \boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{t}) \geq c(\alpha_n, n, p)\} \tag{4}$$

The critical value $c(\alpha_n, n, p)$ is obtained by simulations due to the above mentioned condition that with probability $1-\alpha$ no observation will be identified as an outlier.

3 Simulation Study

In the previous section we mentioned three methods for comparison. Here we will use the abbreviations FGR for the method of Filzmoser et al. (2005), RZ for that of Rousseeuw and Van Zomeren (1990), and BG for the outlier detection method of Becker and Gather (1999).

It should be noted that the concept of the methods RZ and BG for outlier detection are similar because both methods are directly identifying outliers according to their distance. FGR on the other hand tries to identify values that deviate from a majority of observations. In order to make the comparison useful we will generate outliers that are far away from the "clean" data. Thus, outliers will be identified in the tail of the distribution by all three methods, leading to a comparable situation.

In the following we will study the behavior of the methods in a low dimensional ($p = 5$ and $n = 200$) and in a high dimensional ($p = 30$ and $n = 1000$) situation. Moreover,

several data configurations will be considered. The critical values needed for the methods FGR and BG result from simulations with 1000 replications for the corresponding n and p and for the parameters (δ, ε; α) being used.

3.1 Normal Data with Shift Normal Outliers

In this first experiment we generate $n - nout$ data points from the p-variate standard normal distribution $N_p(\mathbf{0}, \boldsymbol{I})$ and $nout$ samples from the "outlier distribution" $N_p(\eta \cdot \mathbf{1}, \boldsymbol{I})$ (shift outliers). The proportion of outliers is varied as $nout/n = 0.05, 0.10, \ldots, 0.45$.

We compute the proportions of identified outliers on the samples generated from the outlier distribution (percentage of correct identified outliers) and the proportion of identified outliers from the "clean data" distribution (percentage of wrong identified outliers). The proportions are averaged over 100 replications of the simulation. The parameter choices are:

- for the method FGR: $\beta = 0.025$ (therefore $\delta = \chi^2_{p,0.975}$) and $\varepsilon = 0.05$
- for the method RZ: $\phi = 0.025$ (therefore cut-off $\chi^2_{p,0.975}$)
- for the method BG: $\alpha = 0.05$

The results are presented in Figure 2, using the legend of Figure 1. For the low dimensional data (left picture) the distance of the outliers was chosen by the value $\eta = 3$, and for the high dimensional data we took $\eta = 1.5$. Compared to other studies (e.g. Rousseeuw and Van Driessen, 1999; Peña and Prieto, 2001) this outlier distance is very low, and in fact there is a significant overlap of the data points from both distributions (more details below). It can be seen that the methods FGR and RZ have similar behavior, except for small outlier fractions for the low dimensional data where FGR does not work well. The method BG performs rather poor in this situation for detecting the outliers. Note that all three methods break down for high outlier percentages. This, however, is due to the properties of the algorithm for computing the MCD estimator: Rousseeuw and Van Driessen (1999) used the same setup–except the distance of the outliers was much higher with $\eta = 10$–and for $n = 1000$ and $p = 30$ the MCD gave the correct solution for a maximum of 24% outliers in the data. For the wrongly identified outliers the method BG gives the smallest percentages, followed by FGR and RZ.

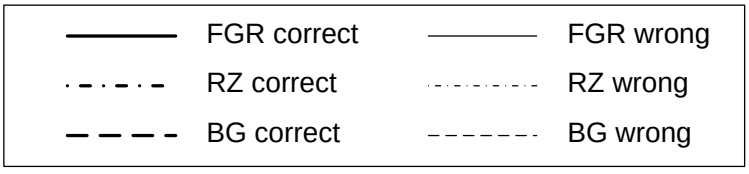

Figure 1: Legend to Figures 2, 4 and 5.

It should be noted that for a larger outlier distance, e.g. by taking $\eta = 10$, the three methods would yield essentially the same (good) results.

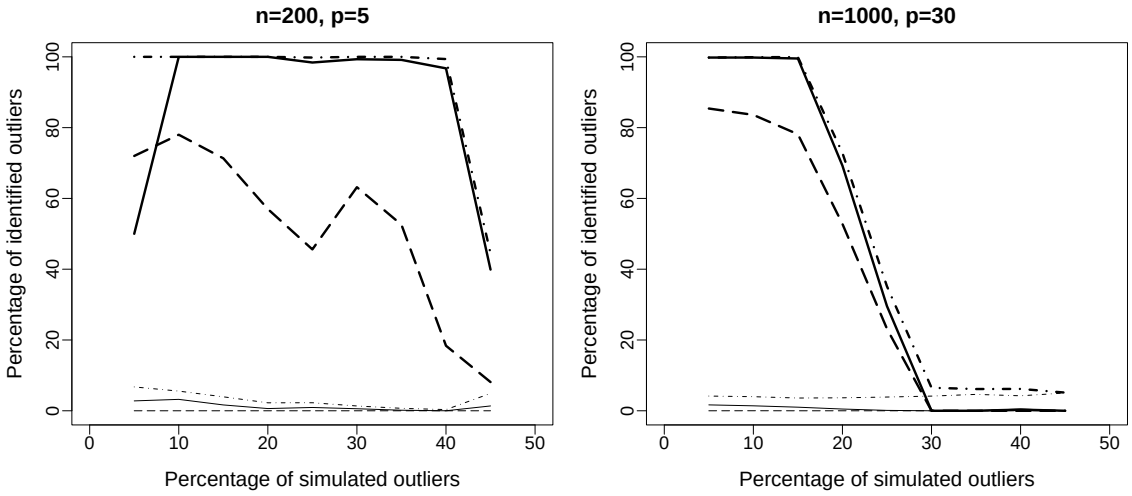

Figure 2: Multivariate standard normal distribution with normally distributed shift outliers with $\eta = 3$ (left) and $\eta = 1.5$ (right), respectively. For the legend see Figure 1.

Remark 1: With respect to the value η of the shift outliers it is interesting to know the "overlap" of the non-outlier and the outlier distribution for growing dimension. We have computed the overlap for $\eta = 1.5$ by simulations as follows. 10^5 data points have been generated according to $N_p(1.5 \cdot \mathbf{1}, \boldsymbol{I})$, and their Mahalanobis distances with respect to center $\mathbf{0}$ and covariance $\boldsymbol{I}$ have been computed. Then we counted the number of samples with distance smaller than $\chi^2_{p,0.975}$. In this way we can estimate the probability mass of the outlier distribution intersecting the p-variate standard normal distribution at the quantile 0.975. The result is shown in the left picture of Figure 3 for dimensions $p = 2, \ldots, 30$. Since this value $\eta = 1.5$ was used in the previous simulation for $p = 30$, it is surprising that the overlap of the outlier distribution is indeed very small. This leads to the situation that with increasing dimension the classification problem of identifying shift outliers should in principle become easier, but due to several studies (e.g. Rousseeuw and Van Driessen, 1999; Rocke and Woodruff, 1999) the computational problems in higher dimensions become larger.

On the other hand we can fix the overlap and ask for the distance of the outlier distribution. The result is shown in the right picture of Figure 3 for a fixed overlap of 10%. We used a log scale on both axes, and it turns out that the relation between (log-)dimension and (log-)distance is almost linear.

3.2 T_3 Distributed Data with Shift Normal Outliers

We take the same simulation setup as in the first experiment, only the "clean" data distribution is changed from multivariate standard normal distribution to multivariate t distribution with 3 degrees of freedom (T_3) (see e.g. Genz and Bretz, 1999). The T_3 distribution has heavier tails and we thus expect more overlap with the outlier distribution. Here we choose the distance of the shift normal outliers due to $\eta = 3$ in both cases $p = 5$ and $p = 30$. The results are shown in Figure 4. Compared to Figure 2 it is clearly visible that the percentage of wrongly identified outliers is much higher in general which is a conse-

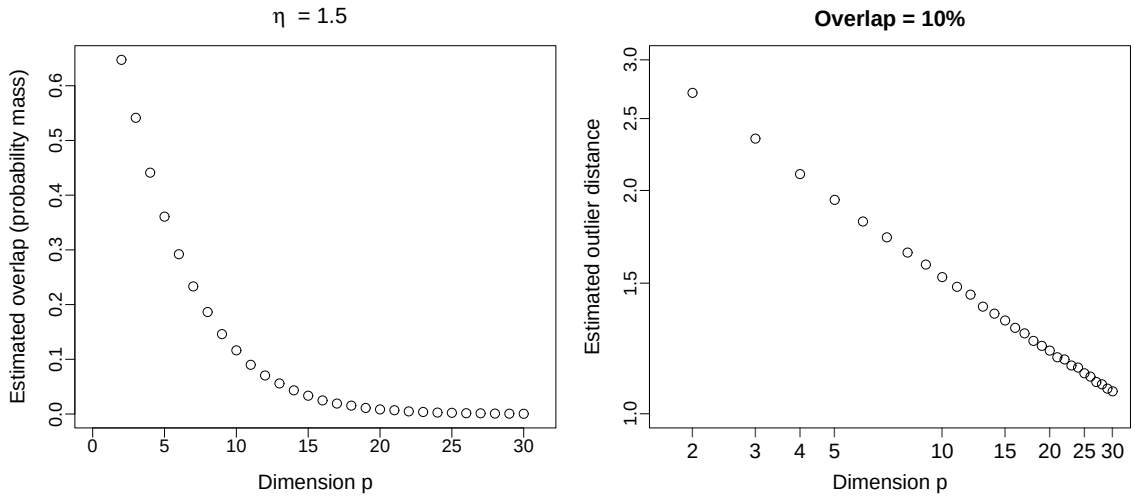

Figure 3: Effect of the dimension on the overlap (left) and of the choice of η (right) of a shifted distribution.

quence of the heavier tails of the T_3 distribution. Method BG gives the best results with this respect. The breakdown in the curve of correctly identified outliers occurs already at a lower percentage of simulated outliers. Method BG gives rather poor results in the low dimensional situation (left picture) but very good results for $p = 30$ (right picture). If we take the same value $\eta = 1.5$ of the shift outliers as in the previous simulation, the results for the correctly identified outliers are comparable to the right picture of Figure 2.

Remark 2: Note that the critical values for the methods BG and FGR were computed for the multivariate normal distribution as model distribution. Since we used T_3 distribution here as "clean data" distribution it would be correct to compute the critical values under this model. However, an aspect of this simulation was to see the effect of deviations from the model.

3.3 Skewed Data with Shift Normal Outliers

Deviations from normality often occur in practical applications, and here we will study the effect of asymmetric data. The simulation setup is similar as before with the difference that we take the absolute values of the data generated from the T_3 distribution. The normal shift outliers are at a value of $\eta = 3$ (for $p = 5$, $n = 200$) and $\eta = 1.5$ (for $p = 30$, $n = 1000$), respectively. The results (Figure 5) are coherent with the results of the previous experiments. The percentages of wrongly identified outliers are comparable to the previous experiment with T_3 distribution, but they decrease with increasing outlier percentage. Again, method BG gives the best results. Methods FGR and RZ have a very good performance for identifying the outliers whereas BG has difficulties.

3.4 Sensitivity with Respect to the Choice of the Parameters

The parameters for the different outlier detection methods were fixed in the previous experiments. Of course it is of interest if this choice has severe influence to the performance

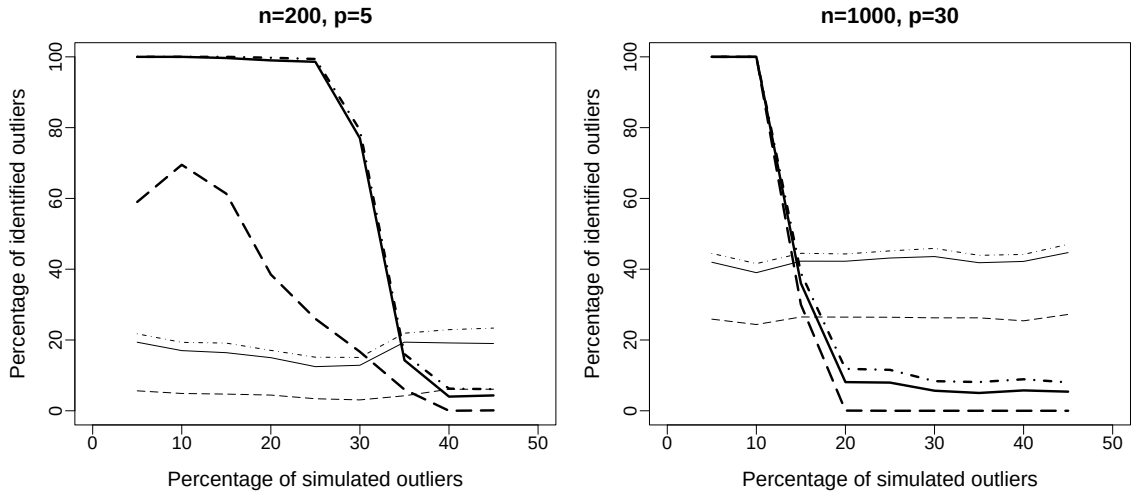

Figure 4: T_3 distribution with normally distributed shift outliers with $\eta = 3$ (left) and $\eta = 3$ (right), respectively. For the legend see Figure 1.

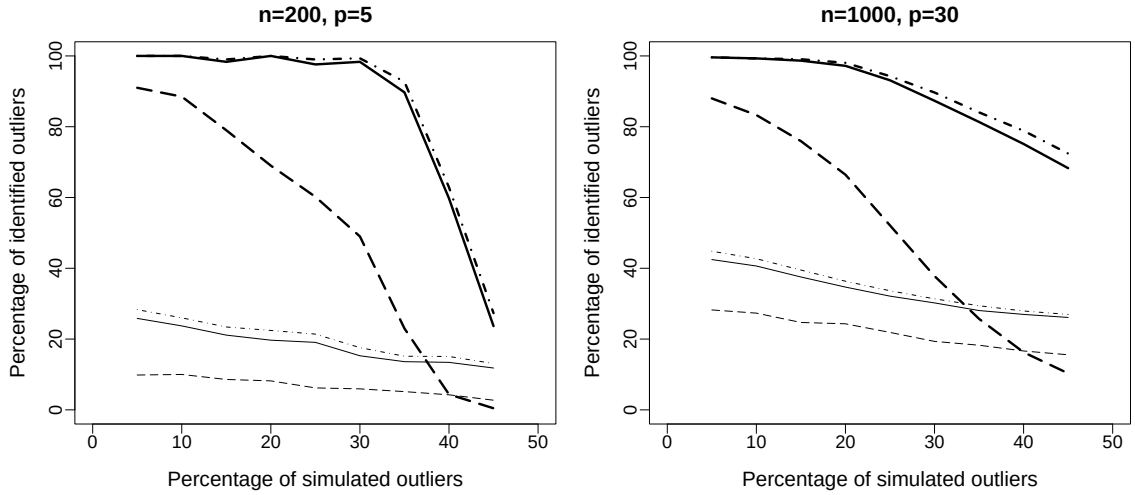

Figure 5: Asymmetric distribution with normally distributed shift outliers with $\eta = 3$ (left) and $\eta = 1.5$ (right), respectively. For the legend see Figure 1.

of the methods. Table 1 presents the results for several parameter choices for the first simulation experiment (p-variate normal data with shift normal outliers) with 10% outliers for $n = 1000$ and $p = 30$ (compare right picture of Figure 1). The left part of the table shows the percentages of correctly identified outliers, and the right part refers to wrongly identified non-outliers. The rows correspond to different choices of the parameter β (defining the tail by $\delta = \chi^2_{p,1-\beta}$) for the method FGR, and to different values of φ (outlier cut-off $\chi^2_{p,1-\varphi}$) for the method RZ. The columns of the table refer to values of $1 - \varepsilon$ for FGR and to the parameter $1 - \alpha$ for method BG (last row). The results for method FGR are very stable, except for some values of $1 - \varepsilon = 0.975$ for the simulated outliers. RZ is sensitive for the parameter choice to identify outliers in the non-outlying group, and BG is rather unstable for the correct outlier identification.

Table 1: Multivariate normally distributed data with 10% shift normal outliers ($n = 1000$, $p = 30$): Percentages of correctly (left) and wrongly (right) identified outliers for different choices of the parameters. The rows correspond to β (for FGR) and φ (for RZ), and the columns to $1 - \varepsilon$ (for FGR) and $1 - \alpha$ (for BG), respectively.

	Outliers							**Non-outliers**						
	FGR						RZ	FGR						RZ
	0.975	0.95	0.9	0.8	0.7	0.6		0.975	0.95	0.9	0.8	0.7	0.6	
0.025	100	100	100	100	100	100	100	1	2	3	3	3	3	4
0.05	100	100	100	100	100	100	100	1	2	3	3	3	3	7
0.1	100	100	100	100	100	100	100	1	2	3	3	3	3	13
0.2	97	99	100	100	100	100	100	1	2	3	3	3	3	24
0.3	82	99	100	100	100	100	100	1	2	3	3	3	3	34
0.4	82	99	100	100	100	100	100	1	2	3	3	3	3	41
BG	81	84	88	92	94	95		0	0	0	0	0	0	

We should mention that for the low dimensional data ($p = 5$, $n = 200$) the results for FGR for detecting the simulated outliers are much more unstable as for the high dimensional data. This can already be expected when looking at the left picture of Figure 1 where the method FGR is unstable for a proportion of up to 10% of outliers.

Note that we also used the correct critical values for FGR and BG in the sense that the simulation for these values was done for the parameter choices used in the table.

The second simulation experiment with the T_3 distribution was repeated for 10% outliers and different parameter choices. The result for the low dimensional data is presented in Table 2. FGR is very stable, RZ has problems for the non-outliers, and BG is sensitive for the correct identification of the outliers. The results for the high dimensional data (not shown) are more stable.

Table 2: T_3 distributed data with 10% shift normal outliers ($n = 200$, $p = 5$): Percentages of correctly (left) and wrongly (right) identified outliers for different choices of the parameters. The rows correspond to β (for FGR) and φ (for RZ), and the columns to $1-\varepsilon$ (for FGR) and $1 - \alpha$ (for BG), respectively.

	Outliers							**Non-outliers**						
	FGR						RZ	FGR						RZ
	0.975	0.95	0.9	0.8	0.7	0.6		0.975	0.95	0.9	0.8	0.7	0.6	
0.025	100	100	100	100	100	100	100	18	19	19	19	19	19	20
0.05	100	100	100	100	100	100	100	18	19	19	19	19	19	24
0.1	100	100	100	100	100	100	100	18	19	19	19	19	19	29
0.2	100	100	100	100	100	100	100	18	19	19	19	19	19	35
0.3	100	100	100	100	100	100	100	18	19	19	19	19	19	41
0.4	100	100	100	100	100	100	100	18	19	19	19	19	19	43
BG	51	57	69	78	84	88		5	5	6	7	8	9	

Essentially the same conclusions can be drawn from repeating the third simulation experiment with asymmetric data ($n = 200$ and $p = 5$), now with 20% shift outliers (Table 3). Method BG is quite stable here also for the detection of the simulated outliers.

Table 3: Asymmetric data with 20% shift normal outliers ($n = 200$, $p = 5$): Percentages of correctly (left) and wrongly (right) identified outliers for different choices of the parameters. The rows correspond to β (for FGR) and φ (for RZ), and the columns to $1 - \varepsilon$ (for FGR) and $1 - \alpha$ (for BG), respectively.

	Outliers							**Non-outliers**						
	FGR						RZ	FGR						RZ
	0.975	0.95	0.9	0.8	0.7	0.6		0.975	0.95	0.9	0.8	0.7	0.6	
0.025	100	100	100	100	100	100	100	6	6	7	7	7	7	8
0.05	100	100	100	100	100	100	100	6	6	7	7	7	7	11
0.1	100	100	100	100	100	100	100	6	6	7	7	7	7	16
0.2	100	100	100	100	100	100	100	6	6	7	7	7	7	23
0.3	100	100	100	100	100	100	100	6	6	7	7	7	7	30
0.4	100	100	100	100	100	100	100	6	6	7	7	7	7	35
BG	95	96	98	99	99	99		0	0	0	0	0	1	

4 Conclusions

The performance of three methods for identifying multivariate outliers was compared. All considered methods are based on the robust Mahalanobis distance, so they rely on a robust estimation of location and covariance. In our simulations we used the MCD estimator where the determinant was minimized over subsets of size $(n + p + 1)/2$ (maximum breakdown value, see Rousseeuw and Van Driessen, 1999). The method RZ (Rousseeuw and Van Zomeren, 1990) uses a quantile of the χ^2_p distribution as outlier cut-off. Method BG (Becker and Gather, 1999) is based on a similar idea, but uses a critical value obtained by simulations for separating outliers. The method FGR (Filzmoser et al., 2005) compares the difference between the empirical distribution of the squared robust Mahalanobis distances and the distribution function of the chi-square distribution. Large differences in the tails indicate outliers, and a critical value obtained by simulations is used for comparison.

The simulations show that the performance of the three methods is mainly determined by the performance of the MCD estimator. Especially the experiments with high dimensional data reflect the limitations of the MCD to identify higher percentages of outliers. As a way out we could use other estimators of multivariate location and scatter (see Section 1). In fact, as was shown in Becker and Gather (2001) the MCD estimator leads in general to the worst results among the methods compared there. In our study the MCD estimator was chosen because it is available in standard statistical software packages and thus frequently used.

In the simulations we observed that the performance of the methods FGR and RZ is comparable. Approximately the same percentages of simulated outliers and non-outliers were detected by both methods. Method BG is preferable for its low rate of wrong outlier classification. However, the behavior as an outlier identifier was rather poor for low dimensional data, and much better for data in higher dimension.

An important aspect is the sensitivity of the methods with respect to the choice of parameters since for real data the percentage of outliers is usually unknown. In our simulations we used a broad range of parameter choices, and it turned out that method FGR is very stable. Since the behavior of the method is also quite good in different data configurations (heavy tails, skewed data) and dimensions, we believe that it is indeed very useful for real data analysis.

A final note should be made on the performance of the methods for data with low sample size n. There are some well-known datasets in the literature which have been analyzed several times for multivariate outliers (see e.g. Rousseeuw and Van Driessen, 1999; Peña and Prieto, 2001). We applied the three methods with the parameter choices $\beta = 0.025$ and $\varepsilon = 0.1$ (FGR), $\varphi = 0.025$ (RZ), and $\alpha = 0.1$ (BG). For each size n and p the corresponding critical values for FGR and BG have been computed by simulations. The indexes of the identified outliers are shown in Table 4. For some datasets the methods

Table 4: Resulting outliers by the methods FGR ($\beta = 0.025$, $\varepsilon = 0.1$), RZ ($\varphi = 0.025$), and BG ($\alpha = 0.1$), for some small datasets.

			Index of identified outliers by method ...		
Dataset	n	p	FGR	RZ	BG
Heart	12	2		2	
Phosphor	18	2		1, 6, 10	
Stackloss	21	3	1, 2, 3, 15, 16, 17, 18, 19, 21	1, 2, 3, 15, 16, 17, 18, 19, 21	
Salinity	28	3		5, 11, 16, 23, 24	16
Hawkins-Bradu-Kass	75	3	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14
Coleman	20	5		1, 6, 9, 10, 11, 15, 18	
Wood	20	5		4, 6, 7, 8, 11, 16, 19	
Bushfire	38	5	7, 8, 9, 10, 11, 12, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38	7, 8, 9, 10, 11, 12, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38	8, 9, 32, 33, 34, 35, 36, 37, 38

FGR and BG could not identify any outlier, although visual inspection gives a different impression. Obviously, for these methods the critical values are too high, and for FGR the approximation of the χ^2_p distribution by the empirical distribution function of the squared robust Mahalanobis distances can be poor for small n.

The method FGR is implemented in R (`http://cran.r-project.org`) as a contributed package called *mvoutlier*. Special symbols for the value of the robust Mahalanobis distance and colors for the coordinate values are suggested to visualize the structure of the multivariate outliers (for details see Filzmoser et al., 2005).

Acknowledgment

I am grateful to Prof. Yuriy Kharin for the excellent cooperation and for all his work to organize this wonderful conference. Moreover, I wish to thank the referees for interesting and helpful comments.

References

V. Barnett and T. Lewis. *Outliers in Statistical Data*. Wiley & Sons, New York, 3rd edition, 1994.

C. Becker and U. Gather. The masking breakdown point of multivariate outlier identification rules. *J. Am. Statist. Assoc.*, 94(447):947–955, 1999.

C. Becker and U. Gather. The largest nonidentifiable outlier: A comparison of multivariate simultaneous outlier identification rules. *Computational Statistics & Data Analysis*, 36:119–127, 2001.

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

H. Doleisch, M. Gasser, and H. Hauser. Interactive feature specification for focus+context visualization of complex simulation data. In *Proc. of the Joint IEEE TCVG – EG Symp. on Vis.*, pages 239–248, 2003.

P. Filzmoser, R.G. Garrett, and C. Reimann. Multivariate outlier detection in exploration geochemistry. *Computers and Geosciences*, 2005. In press.

R.G. Garrett. The chi-square plot: A tool for multivariate outlier recognition. *Journal of Geochemical Exploration*, 32:319–341, 1989.

A. Genz and F. Bretz. Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts. *Journal of Statistical Computation and Simulation*, 63:361–378, 1999.

D. Gervini. A robust and efficient adaptive reweighted estimator of multivariate location and scatter. *Journal of Multivariate Analysis*, 84:116–144, 2003.

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

J.T. Kent and D.E. Tyler. Constrained M-estimation for multivariate location and scatter. *The Annals of Statistics*, 24(3):1346–1370, 1996.

R.A. Maronna. Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4(1):51–67, 1976.

R.A. Maronna and V.J. Yohai. The behavior of the Stahel-Donoho robust multivariate estimator. *J. Am. Statist. Assoc.*, 90:330–341, 1995.

D. Peña and F.J Prieto. Multivariate outlier detection and robust covariance matrix estimation (with discussion). *Technometrics*, 43(3):286–310, 2001.

D.M. Rocke and D.L. Woodruff. Identification of outliers in multivariate data. *J. Am. Statist. Assoc.*, 91:1047–1061, 1996.

D.M. Rocke and D.L. Woodruff. A synthesis of outlier detection and cluster identification. Technical report, University of California, Davis, Davis CA 95616, 1999. http://handel.cipic.ucdavis.edu/ dmrocke/Synth5.pdf.

P.J. Rousseeuw and B.C. Van Zomeren. Unmasking multivariate outliers and leverage points. *J. Am. Statist. Assoc.*, 85(411):633–651, 1990.

P.J. Rousseeuw. Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, editors, *Mathematical Statistics and Applications*, volume B, pages 283–297, Budapest, 1985. Akadémiai Kiadó.

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

D. Swayne, D. Cook, and A. Buja. XGobi: Interactive dynamic data visualization in the X Windows system. *Journal of Computational and Graphical Statistics*, 7(1):113–130, 1998.

D.E. Tyler. Some issues in the robust estimation of multivariate location and scatter. In W. Stahel and S. Weisberg, editors, *Directions in Robust Statistics and Diagnostics 2*, pages 327–336. Springer, New York, 1991.

Author's address:

Prof. Peter Filzmoser
Department of Statistics and Probability Theory
Vienna University of Technology
Wiedner Hauptstraße 8-10
A-1040 Vienna, Austria

Tel. +43 1 58801 10733
Fax +43 1 58801 10799
E-mail: P.Filzmoser@tuwien.ac.at
http://www.statistik.tuwien.ac.at/public/filz/