The identification of outliers in exponential samples

V. Schultze and J. Pawlitschko*

Department of Statistics, University of Dortmund, Vogelpothsweg 87, D-44221 Dortmund, Federal Republic of Germany

Abstract

In this paper, the task of identifying outliers in exponential samples is treated conceptionally in the sense of DAVIES and GATHER (1989, 1993) by means of a so-called outlier region. In case of an exponential distribution, an empirical approximation of such a region – also called an outlier identifier – is mainly dependent on some estimator of the unknown scale parameter. The worst-case behaviour of several reasonable outlier identifiers is investigated thoroughly and it is shown that only robust estimators of scale should be used to construct reliable identifiers. These findings lead to the recommendation of an outlier identifier that is based on a standardized version of the sample median.

Key Words & Phrases: Outlier identifier, breakdown point, masking, swamping, maximum asymptotic bias, robustness.

1 Introduction

It is a common problem in applied statistics that in samples which are taken from some target population some observations occur which seem to differ strongly from the bulk of the data. Such an observation is usually called an "outlier". However, there exists no formal definition of what constitutes an outlier that has been widely accepted.

In this paper we focus on outlying observations in life time data. A simple but nevertheless useful model for such data assumes that the observed life times $x_1, \ldots, x_N$ form a random sample from an exponentially distributed random variable X with

*pawl@amadeus.statistik.uni-dortmund.de

unknown scale parameter $\nu > 0$. Hence, the distribution and density function of X are given by $F_\nu(x) = 1 - \exp(-x/\nu)$, $x \geq 0$, and $f_\nu(x) = 1/\nu \exp(-x/\nu)$, $x \geq 0$, respectively.

In the statistical literature, the problem of detecting the presence of outliers in exponential samples has been investigated intensively. A comprehensive account on contributions to this topic can be found in GATHER (1995). Most of this work is based on so-called outlier generating models. Here it is assumed that potential outliers come from different distributions than the rest of the data. The problem of outlier detection is then seen as a testing problem with null hypothesis that all observed life times come from the same exponential distribution – the null model – and alternative that at least one life time comes from another distribution permitted by the chosen outlier generating model. This approach has some drawbacks: One is that it does not really take into account that the only property of outliers being commonly supposed is that their position is quite unlikely under the null model, irrespective which distribution they follow. Further, if a test rejects the null model then one can only conclude that outliers are present, but not identify them.

To overcome these drawbacks, DAVIES and GATHER (1989, 1993) (see also GATHER, 1990) introduced the notion of an outlier region. Let F be an absolutely continuous distribution function with density f. For any α, $0 < \alpha < 1$, the α-outlier region of F is defined as $out(\alpha, F) = \{x \in \mathbb{R} \mid f(x) < \delta(\alpha)\}$, where $\delta(\alpha) = \sup\{\delta > 0 \mid P(f(X) < \delta) \leq \alpha\}$ and X has distribution function F. Then, any real number x is called an α-outlier with respect to F if it lies in $out(\alpha, F)$. In case of an exponential distribution F_ν, a corresponding α-outlier region is given by

$$out(\alpha, F_\nu) = \{x \geq 0 \mid x > -\nu \ln \alpha\}. \tag{1}$$

Often, the level $\alpha = \alpha_N$ of an outlier region is chosen depending on the size N of a given sample. One possible choice is based on the requirement that under the null model for some $\tilde{\alpha}$, $0 < \tilde{\alpha} < 1$, and for an i.i.d. sample $\underline{X}_N = (X_1, \ldots, X_N)$ one has

$$P(X_i \notin out(\alpha_N, F_\nu),\ i = 1, \ldots, N) \ = \ 1 - \tilde{\alpha},$$

which leads to

$$\alpha_N \;=\; 1-(1-\tilde{\alpha})^{1/N}. \tag{2}$$

The task of identifying all outliers in an exponential sample can now be formalized in the following way: Given a realized sample $\underline{x}_N = (x_1, \ldots, x_N)$ with at least $n > N/2$ regular observations, i.e. these observations come from i.i.d. F_ν-distributed random variables, for fixed α_N, find all observations which lie in $out(\alpha_N, F_\nu)$. Since ν is unknown, one has to find an empirical approximation of $out(\alpha_N, F_\nu)$, such an approximation is usually called an outlier identifier. From (1) it is obvious that this problem can essentially be solved by estimating the unknown scale parameter ν. Then it must be taken into account that estimators of ν might be heavily distorted if outliers are contained in the sample.

The rest of this paper is organized as follows: In Section 3, four different outlier identifiers are presented which are based on different reasonable estimators of the scale parameter ν. In Sections 3 and 4, these identifiers are investigated with respect to their worst-case behaviour. For this purpose, their masking and swamping breakdown point as well as their maximum asymptotic bias are compared. It turns out that only robust estimators of scale lead to reliable outlier identifiers. In Section 5, the results of an extensive simulation study are presented. Finally, a real data example is contained in Section 6.

2 Outlier identifiers

Let $S_N = S_N(\underline{x}_N)$ be an arbitrary estimator of the scale parameter ν. Then, a one-step α_N-outlier identifier based on S_N generally has the form

$$OR_{S_N}(\underline{x}_N, \alpha_N) \;=\; (S_N(\underline{x}_N)\, g(N, \alpha_N), \infty),$$

where $g(N, \alpha_N)$ is a normalizing constant, whose choice is discussed later.

The following outlier identifiers are considered in this paper:

i) Standardized median identifier (SM-Oi)

$$\begin{aligned} OR_{SM_N}(\underline{x}_N, \alpha_N) &= (SM_N(\underline{x}_N)\, g(N, \alpha_N), \infty) \\ \text{with } \ SM_N(\underline{x}_N) &= a\; Med(x_1, \ldots, x_N), \end{aligned}$$

ii) RCS identifier (RCS-Oi)

$$\begin{aligned} OR_{RCS_N}(\underline{x}_N, \alpha_N) &= (RCS_N(\underline{x}_N)\, g(N, \alpha_N), \infty) \\ \text{with } \ RCS_N(\underline{x}_N) &= b\; Med_i\{Med_j\{|x_i - x_j|;\ i, j \in \{1, \ldots, N\}\}\}, \end{aligned}$$

iii) RCQ identifier (RCQ-Oi)

$$\begin{aligned} OR_{RCQ_N}(\underline{x}_N, \alpha_N) &= (RCQ_N(\underline{x}_N)\, g(N, \alpha_N), \infty) \\ \text{with } RCQ_N(\underline{x}_N) &= c\, \{|x_i - x_j|;\ i, j \in \{1, \ldots, N\}, i < j\}_{(l)} \\ \text{where } \ l &= \lceil \frac{N\,(N-1)}{8} \rceil, \end{aligned}$$

iv) Mean identifier or Maximum likelihood identifier (ML-Oi)

$$\begin{aligned} OR_{ML_N}(\underline{x}_N, \alpha_N) &= (ML_N(\underline{x}_N)\, g(N, \alpha_N), \infty) \\ \text{with } \ ML_N(\underline{x}_N) &= \frac{1}{N} \sum_{i=1}^{N} x_i. \end{aligned}$$

The estimator SM_N has been suggested by Gather and Schultze (1998) as a robust estimator of scale specially for exponential samples. To achieve Fisher-consistency, the constant a must be set to $a = 1/\ln 2 = 1.4427$. The estimators RCS_N and RCQ_N have been proposed by Rousseeuw and Croux (1993) and are generally useful for estimating a scale parameter in samples from location-scale distributions. When applied to estimate the standard deviation of a normal distribution they are very robust and work quite well. To make them Fisher-consistent in the exponential case, one has to choose $b = 1.6982$ and $c = 3.4760$.

It remains to specify the constants $g(N, \alpha_N)$. This is done by the requirement that in samples without any outliers

$$P(OR_{S_N}(\underline{X}_N, \alpha_N) \subset out(\alpha_N, F_\nu)) = 1 - \tilde{\alpha} \tag{3}$$

$$\text{or} \quad P(X_i \notin OR_{S_N}(\underline{X}_N, \alpha_N);\ i = 1, \ldots, N) = 1 - \tilde{\alpha}. \tag{4}$$

Tables 1 – 4 contain the constants for $\tilde{\alpha} = 0.05$, α_N determined according to (2), and sample sizes $N = 10, 20, 50, 100$, under both requirements. For the RCS and the RCQ identifier, the constants have been simulated, each value is based on 10000 runs. For the other two identifiers, they have been calculated exactly.

SM-Oi	N=10	N=20	N=50	N=100
(3)	11.39	10.36	9.76	9.65
(4)	6.97	7.01	7.54	7.99

Table 1. Values of $g(N, \alpha_N)$ for SM-Oi

RCS-Oi	N=10	N=20	N=50	N=100
(3)	13.74	11.14	9.96	9.70
(4)	7.38	7.50	7.66	8.04

Table 2. Values of $g(N, \alpha_N)$ for RCS-Oi

RCQ-Oi	N=10	N=20	N=50	N=100
(3)	11.23	9.51	9.21	9.18
(4)	5.81	6.45	7.16	7.75

Table 3. Values of $g(N, \alpha_N)$ for RCQ-Oi

ML-Oi	N=10	N=20	N=50	N=100
(3)	9.72	9.00	8.83	9.00
(4)	4.45	5.41	6.57	7.38

Table 4. Values of $g(N, \alpha_N)$ for ML-Oi

3 Breakdown points

The reliability of an outlier identifier can be judged by his proneness for false decisions. There exist two possibilities of making mistakes. The first one is to fail to identify a clear outlier and the opposite mistake is to discover more outliers than are really existing.

If an outlier identifer is unable to recognize an arbitrarily large outlier because of the presence of some other outliers, it is said that "the identifier breaks down by masking". A measure for the sensitivity of an identifier w.r.t. this kind of failure is its masking breakdown point (see Davies and Gather, 1993) which is defined as the minimal fraction of badly placed observations which let the identifier break down. More formally, given an outlier identifier, a sequence $\underline{\alpha} = (\alpha_N)_{N \in}$ with $\alpha_N \in (0, 1)$, $\delta \in (0, 1)$, and a sample with n regular observations $\underline{x}_n^r$, the masking breakdown point is given by

$$\epsilon^M(\underline{\alpha}, \underline{x}_n^r, \delta) \;=\; \frac{k^M}{n + k^M},$$

with $k^M = \min\{k : \beta^M(\alpha_{n+k}, \underline{x}_n^r, k, \delta) = 0\}$ and

$$\begin{aligned}\beta^M(\alpha_{n+k}, \underline{x}_n^r, k, \delta) \;=\; & \inf\{\beta > 0 : \text{there exist } \delta\text{-outliers } \underline{x}_k^o = (x_1^o, \ldots, x_k^o) \\ & \text{such that some point in } out(\beta, F_\nu) \text{ is not identified} \\ & \text{as an } \alpha_{k+n}\text{-outlier}\}.\end{aligned}$$

If the presence of outliers in a sample has the effect that an identifier classifies some non-outlying observations as outliers, then it is said that the identifier suffers from swamping. The swamping breakdown point of an identifier is the smallest fraction of badly placed observations which cause a non-outlying observation to be identified as arbitrarily large outlier. More formally, for a given identifier, a given sequence $\underline{\alpha} = (\alpha_N)_{N\in}$ with $\alpha_N \in (0,1)$, $\delta \in (0,1)$, and a sample with n regular observations $\underline{x}_n^r$, the swamping breakdown point is defined as

$$\epsilon^S(\underline{\alpha}, \underline{x}_n^r, \delta) \;=\; \frac{k^S}{n + k^S},$$

with $k^S = \min\{k : \beta^S(\alpha_{n+k}, \underline{x}_n^r, k, \delta) = 0\}$ and

$$\begin{aligned}\beta^S(\alpha_{n+k}, \underline{x}_n^r, k, \delta) \;=\; & \inf\{\beta > 0 : \text{ there exist } \delta\text{-outliers } \underline{x}_k^o \text{ such} \\ & \text{that some non-}\alpha_{n+k}\text{-outlier is identified as } \beta\text{-outlier}\}.\end{aligned}$$

Theorem 3.1. Let $\underline{x}_n^r$ be a regular sample from an exponential distribution, $\delta \in (0,1)$, and $\underline{\alpha} = (\alpha_N)_{N\in}$ with $\alpha_N \in (0,1)$, then

i) for the standardized median identifier: $\epsilon^M(\underline{\alpha}, \underline{x}_n^r, \delta) = 1/2$,

ii) for the RCS identifier: $\epsilon^M(\underline{\alpha}, \underline{x}_n^r, \delta) = 1/2$,

iii) for the RCQ identifier: $\epsilon^M(\underline{\alpha}, \underline{x}_n^r, \delta) = 1/2$,

iv) and for the mean identifier: $\epsilon^M(\underline{\alpha}, \underline{x}_n^r, \delta) = 1/(n+1)$.

Proof.

i) Consider a sample of size $N = 2n$ containing a number n of δ-outliers all being equal to some x^o which is larger than the maximal regular observation. If $x^o \to \infty$, then it follows that also $SM_N(\underline{x}_N) \to \infty$ and $OR_{SM_N}(\underline{x}_N, \alpha_N) \to \emptyset$. So $\epsilon^M(\underline{\alpha}, \underline{x}_n^r, \delta) = 1/2$.

ii) In addition to the regular sample, choose a number n of δ-outliers $\underline{x}_n^o = (x_{(1)}^o, \ldots, x_{(n)}^o)$ such that for some constant $L > 0$

$$\begin{aligned} x_{(1)}^o &= x_{(n)}^r + L \\ x_{(2)}^o &= x_{(n)}^r + 2\,L \\ &\vdots \\ x_{(n)}^o &= x_{(n)}^r + n\,L, \end{aligned}$$

where $x_{(n)}^r$ denotes the largest regular observation. If $L \to \infty$, then it follows that $RCQ_N(\underline{x}_N) \to \infty$ and $OR_{RCQ_N}(\underline{x}_N, \alpha_N) \to \emptyset$.

iii) Consider a sample of size $2n$ containing a number n of δ-outliers as in part i). If $x^o \to \infty$, then $RCS_N(\underline{x}_N) \to \infty$ and hence $OR_{RCS_N}(\underline{x}_N, \alpha_N) \to \emptyset$.

iv) Choose a single δ-outlier x^o, hence $N = n+1$. If $x^o \to \infty$, then $ML_N(\underline{x}_N) \to \infty$ and hence $OR_{ML_N}(\underline{x}_N, \alpha_N) \to \emptyset$.

□

The theorem shows clearly that only such outlier identifiers should be used which are based on robust estimators of the unknown scale parameter ν. For n tending to infinity, the masking breakdown point of the mean identifier tends to zero. Hence, the mean identifier works bad especially in large samples.

Theorem 3.2. Let $\underline{x}_n^r$ be a regular sample from an exponential distribution, such that $x_i^r \neq x_j^r$ for $i \neq j$ with $i, j \in \{1, \ldots, n\}$. Further, let $\underline{\alpha} = (\alpha_N)_{N \in}$ with $\alpha_N \in (0,1)$, and $\delta \in (0,1)$. Then it follows

i) for the standardized median identifier: $\epsilon^S(\underline{\alpha}, \underline{x}_n^r, \delta) = 1$,

ii) for the RCQ identifier: $\epsilon^S(\underline{\alpha}, \underline{x}_n^r, \delta) = (n+1)/(2n+1)$,

iii) for the RCS identifier: $\epsilon^S(\underline{\alpha}, \underline{x}_n^r, \delta) = (n+1)/(2n+1)$,

iv) and for the mean identifier: $\epsilon^S(\underline{\alpha}, \underline{x}_n^r, \delta) = 1$.

Proof.

i) Since $\delta > 0$ it follows that no δ-outlier added to the sample can be equal to zero, and hence $SM_N(\underline{x}_N) > 0$ irrespective of how many outliers occur. Hence, for no finite N we have $OR_{SM_N}(\underline{x}_N, \alpha_N) = \mathbb{R}^+$ and hence $\epsilon^S(\underline{\alpha}, \underline{x}_n^r, \delta) = 1$.

ii) Consider a sample of size $N = 2\,n + 1$ containing a number $n+1$ of δ-outliers all being equal to some $x^o > 0$. Then one has $RCQ_N(\underline{x}_N) = 0$, and hence $OR_{RCQ_N}(\underline{x}_N, \alpha_N) = \mathbb{R}^+$. It follows that $\epsilon^S(\underline{\alpha}, \underline{x}_n^r, \delta) = (n+1)/(2n+1)$.

Part iii) of the theorem can be proven similary to part ii), and part iv) similary to part i). □

Davies and Gather (1993) have pointed out that masking and swamping breakdown point of an identifier behave contrary if the regular observations come from a normal distribution. This means that if an identifier has a small masking breakdown point, it usually has a high swamping breakdown point. In samples where the regular observations come from an exponential distribution, this is not necessarily true: e.g. the median identifier has both a high masking and a high swamping breakdown point. The reason is that here, in opposite to the normal distribution, outlier regions only extend over the upper tail of the distribution. Hence, extremely small observations are never considered as outliers, so the median identifier cannot break down by swamping.

4 Asymptotic bias and large outliers

Another interesting problem is the search for the largest outlier which cannot be discovered by an identifier. It has already been pointed out that the reason why outlier identifiers fail to detect outliers in a sample are the outliers themselves. They distort the scale estimators on which the identifiers are based. The following definition of Davies and Gather (1993) quantifies this distortion.

Let $\eta \in (0,1)$ and $\underline{\delta} = (\delta_N)_{N\in}$, $\delta_N \in (0,1)$, be given. Consider a sequence $(x_i)_{i\in}$ of regular observations from an exponential distribution F_ν. For each N, let $\underline{x}_N$ be a sample of size N which contains the first n regular observations of the given sequence and $k = \lfloor \eta n \rfloor$ nonregular observations $\underline{x}_k^o$ which lie in $out(\delta_N, F_\nu)$. Now, let S_N be an estimator of the scale parameter ν. Then the maximum asymptotic bias of S_N is defined as

$$b_S(S, \eta, \underline{\delta}) \;=\; \limsup_{N\to\infty} \sup_{\underline{x}_k^o \in out(\delta_N, F_\nu)} \ln \frac{S_N(\underline{x}_N)}{\nu} \quad .$$

For all estimators considered here, the maximum asymptotic bias is independent of the sequence $(x_i)_{i\in}$.

Theorem 4.1.

i) The maximum asymptotic bias of $SM_N(\underline{x}_N)$ is

$$b_S(SM, \eta, \underline{\delta}) \;=\; \ln \;\; -1.4427 \ln \;\; \frac{1-\eta}{2} \quad .$$

ii) The maximum asymptotic bias of $RCQ_N(\underline{x}_N)$ is

$$\begin{aligned} b_S(RCQ, \eta, \underline{\delta}) &= \ln(RCQ(\eta)) \\ \text{with } RCQ(\eta) &= 3.476\, F_*^{-1} \;\; \frac{5\,(1+\eta)^2 - 8\,\eta\,(1+\eta) + 4\,\eta^2}{8} \end{aligned}$$

$$\text{and } F_*^{-1}(x) \;=\; \begin{matrix} \ln(2\,x), & 0 < x \le 1/2 \\ -\ln(2-2\,x), & 1/2 < x < 1. \end{matrix}$$

iii) The maximum asymptotic bias of $RCS_N(\underline{x}_N)$ is

$$b_S(RCS, \eta, \underline{\delta}) \;=\; \ln(1.6982\, x_S(\eta)),$$

where $x_S(\eta)$ is the smallest positive solution of

$$(1-\eta)\, e^{2x} - (1+\eta)\,(e^{x} - e^{-x}) - 2 \;=\; 0.$$

iv) The maximum asymptotic bias of the sample mean is

$$b_S(ML, \eta, \underline{\delta}) \;=\; \infty.$$

Proof. Since we only allow distortion of the estimators due to $k \leq n$ badly positioned δ_N-outliers, the expressions for the maximum asymptotic bias in parts i) – iii) can easily be deduced from the corresponding explosion bias curves developed in Gather and Schultze (1998) – note that ϵ there must be replaced by $\eta/(1+\eta)$. Part iv) is clear form the proof of Theorem 3.1 iv). $\square$.

Independently of (3) or (4), one has

$$\lim_{N\to\infty} \; g(N, \alpha_N) - (-\ln \alpha_N) \;\;=\; 0, \tag{5}$$

because the four estimators are consistent in i.i.d. samples. Now, let S_N be equal to the standardized median or one of the two estimators proposed by Rousseeuw and Croux (1993). Then the size of the largest nonidentifiable α_N-outlier in large samples with a given fraction $\eta/(1+\eta)$ of δ_N-outliers can easily be approximated by

$$ALO(OR_{S_N}) \;=\; -\nu \ln \alpha_N \;\exp\; b_S(S, \eta, \underline{\delta}) \;.$$

This approach does not work for the mean identifier, because its maximum asymptotic bias is infinite. However, at least for η sufficiently small, a different approach is possible. Consider samples with $k = \lfloor \eta n \rfloor$ outliers all being equal to some $x^o \in out(\delta_N, F_\nu)$. If δ_N is large enough, x^o will also be the maximum of the entire sample. Consider now the case that x^o lies on the left border of the approximated outlier region so that it is the largest value that the mean identifier cannot identify as α_N-outlier, that is

$$x^o = \frac{1}{N} \sum_{i=1}^{N} x_i \; g(N, \alpha_N).$$

If N is large enough, by (5) one has

$$\frac{x^o}{1/N \sum_{i=1}^{N} x_i} \; = \; -\ln \alpha_N.$$

Approximating the mean of the $n = N - k$ regular observations by ν leads to

$$\frac{x^0}{n/(n+k)\, \nu + k/(n+k)\, x^o} \; = \; -\ln \alpha_N,$$

and replacing k by $\eta\, n$ yields

$$x^o \; = \; \frac{-\nu \, \ln \alpha_N}{1 + \eta + \eta \, \ln \alpha_N}.$$

Hence, if $1 + \eta + \eta \ln \alpha_N > 0$, then the largest nonidentifiable α_N-outlier of the mean identifier can be approximated by

$$ALO(OR_{ML_N}) \; = \; \frac{-\nu \, \ln \alpha_N}{1 + \eta + \eta \, \ln \alpha_N}.$$

In other cases, the corresponding largest nonidentifiable outlier is not bounded.

A comparison of the other three identifiers which are based on robust estimators of scale shows that the median identifier behaves best and that the approximated largest nonidentifiable outlier is generally smaller for the RCS than for the RCQ identifier.

5 Simulations

To give an idea of the sample sizes which are necessary for a good approximation of the largest nonidentifiable outlier, the asymptotic results are supported by some simulations. We consider sample sizes of $N = 10$, 20, 50 and 100, with the number of δ-outliers chosen as $k = 1, 2, 3, 5, 7, 10, 15, 20, 25, 30, 49,$ but only if $k/N < 1/2$.

All identifiers are standardized according to either (3) or (4) and are designed to detect α_N-outliers, where α_N is chosen according to (2) with $\tilde{\alpha} = 0.05$.

For the simulation, for each combination of k and N, 2000 samples were generated as follows: First $n = N - k$ observations were taken from a standard exponential distribution (i.e. $\nu = 1$). Then the remaining k observations were placed such that the identifier could not detect them as outliers, but their values were as large as possible. For δ small enough, the resulting samples had n regular observations and k observations in $out(\delta, F_1)$. Now, for each sample, the size of the largest nonidentifiable outlier was determined, and their average was calculated.

The tables in the Appendix contain the simulated (SLO) as well as the approximated (ALO) largest nonidentifiable outlier, further the quality of the approximation is described by $Pr = 100\, ALO/SLO$. It turns out that the approximation works quite well if standardisation according to (4) is choosen. Further, the following results can be stated: Independently of N, k, and condition (3) or (4), for SLO one has:

$$SLO(OR_{SM_N}) \; < \; SLO(OR_{RCS_N}).$$

Hence, for the standardized median identifier the largest nonidentifiable outlier is always smaller than for the RCS identifier. For samples with only few outliers and large sample sizes, one has

$$SLO(OR_{RCQ_N}) < SLO(OR_{RCS_N}) \quad \text{and} \quad SLO(OR_{RCQ_N}) < SLO(OR_{SM_N}).$$

If the number of outliers increases, one has

$$SLO(OR_{RCS_N}) \; < \; SLO(OR_{RCQ_N}).$$

6 Example and Conclusions

As an example for the application of the outlier identifiers discussed in this paper, we consider a data set taken from Nelson (1982, p. 104). This data set containes the times to breakdown of an insulating fluid between two electrodes, recorded at a voltage of 34 kV. The recorded breakdown times in ascending order are 0.19, 0.78, 0.96, 1.31, 2.78, 3.16, 4.15, 4.67, 4.85, 6.50, 7.35, 8.01, 8.27, 12.06, 31.75, 32.52, 33.91, 36.71, 72.89. The following table contains α_N-outlier identifiers for these data set, where α_N is chosen as in (2) with $\tilde{\alpha} = 0.05$, and with standardization according to either (3) or (4).

	S_N	Outlier identifier, standardization according to (3)	(4)
SM-Oi	9.38	$(99.71, \infty)$	$(66.69, \infty)$
RCS-Oi	9.32	$(109.42, \infty)$	$(72.70, \infty)$
RCQ-Oi	11.12	$(106.97, \infty)$	$(71.17, \infty)$
ML-Oi	14.36	$(129.67, \infty)$	$(76.68, \infty)$

Table 5. Outlier identifiers for the insulating fluid example

As is seen from Table 5, no observation is identified as α_N-outlier if standardization is made according to (3). However, it is also seen that the median identifier has the smallest lower border of the four competing identifiers. When standardized according to (4), all identifiers based on robust estimators of scale detect the largest observation 72.89 as α_N-outlier, even though for the RCS and RCQ identifier it lies very close to their lower border. The mean identifier, however, does not find any outlying observation in this case, too.

To come to a final conclusion, it can be stated that with respect to their worst-case behaviour, for samples from an exponential distribution, only robust estimators of scale lead to reliable one-step outlier identifiers. The mean identifier is only suitable in case of one single outlying observation, because of its very small masking breakdown point. In summary, the use of the standardized median identifier is recommended, because the median is easy to calculate, the corresponding identifier has optimal breakdown points, and especially in large samples, for the largest nonidentifiable outlier one has $SLO(OR_{SM_N}) < SLO(OR_{RCS_N}) < SLO(OR_{RCQ_N})$.

Appendix: Tables of the largest nonidentifiable outliers

a) Standardization according to (3):

$$N = 10: \quad out(\alpha_{10}, F_1) = (5.28, \infty)$$

k		ML-Oi	Pr	RCQ-Oi	Pr	RCS-Oi	Pr	SM-Oi	Pr
1	ALO	10.04	3.23	6.77	41.26	6.44	37.42	6.18	42.47
	SLO	310.69		16.41		17.21		14.55	
2	ALO	.	.	9.09	38.58	8.12	36.30	7.47	42.40
	SLO	$> 10^6$		23.56		22.37		17.62	
3	ALO	.	.	13.10	37.27	10.78	34.84	9.53	42.62
	SLO	$> 10^6$		35.15		30.94		22.36	

$$N = 20: \quad out(\alpha_{20}, F_1) = (5.97, \infty)$$

k		ML-Oi	Pr	RCQ-Oi	Pr	RCS-Oi	Pr	SM-Oi	Pr
1	ALO	8.08	51.56	6.73	58.67	6.57	53.24	6.44	55.28
	SLO	15.67		11.47		12.34		11.65	
2	ALO	13.31	16.31	7.66	58.16	7.28	53.02	6.98	55.22
	SLO	81.63		13.17		13.73		12.64	
3	ALO	48.24	.	8.81	57.51	8.13	52.69	7.46	55.36
	SLO	$> 10^6$		15.32		15.43		13.80	
5	ALO	.	.	12.20	55.25	10.49	51.47	9.46	55.26
	SLO	$> 10^6$		22.08		20.38		17.12	
7	ALO	.	.	18.59	52.35	14.50	50.24	12.63	55.59
	SLO	$> 10^6$		35.51		28.86		22.72	

$$N = 50: \quad out(\alpha_{50}, F_1) = (6.88, \infty)$$

k		ML-Oi	Pr	RCQ-Oi	Pr	RCS-Oi	Pr	SM-Oi	Pr
1	ALO	7.82	74.12	7.21	72.61	7.14	68.65	7.08	68.54
	SLO	10.55		9.93		10.40		10.33	
2	ALO	9.17	68.23	7.57	72.65	7.42	68.77	7.30	68.74
	SLO	13.14		10.42		10.79		10.62	
3	ALO	11.02	62.33	7.95	72.54	7.72	68.74	7.53	68.83
	SLO	17.68		10.96		11.23		10.94	
5	ALO	19.87	29.29	8.83	72.44	8.39	68.66	8.05	69.04
	SLO	67.84		12.19		12.22		11.66	
7	ALO	162.18	.	9.87	72.10	9.16	68.41	8.64	68.95
	SLO	$> 10^6$		13.69		13.39		12.53	
10	ALO	.	.	11.85	71.77	10.58	68.26	9.74	69.13
	SLO	$> 10^6$		16.51		15.50		14.09	
15	ALO	.	.	17.07	65.25	14.02	67.57	12.44	69.30
	SLO	$> 10^6$		26.16		20.75		17.95	
20	ALO	.	.	28.36	68.06	20.69	66.92	17.79	69.87
	SLO	$> 10^6$		41.67		30.92		25.46	

$N = 100: \quad out(\alpha_{100}, F_1) = (7.58, \infty)$

k		ML-Oi	Pr	RCQ-Oi	Pr	RCS-Oi	Pr	SM-Oi	Pr
1	ALO	8.11	82.17	7.76	81.09	7.72	77.59	7.69	77.21
	SLO	9.87		9.57		9.95		9.96	
2	ALO	8.75	80.79	7.94	81.02	7.87	77.61	7.81	77.25
	SLO	10.83		9.80		10.14		10.11	
3	ALO	9.51	78.99	8.14	81.16	8.02	77.71	7.92	77.27
	SLO	12.04		10.03		10.32		10.25	
5	ALO	11.58	74.04	8.55	81.12	8.34	77.65	8.17	77.29
	SLO	15.64		10.54		10.75		10.57	
7	ALO	15.00	65.99	8.99	81.14	8.68	77.57	8.44	77.43
	SLO	22.73		11.08		11.19		10.90	
10	ALO	28.19	34.67	9.73	81.02	9.24	77.58	8.87	77.40
	SLO	81.31		12.01		11.91		11.46	
15	ALO	.	.	11.19	81.00	10.33	77.61	9.70	77.54
	SLO	$> 10^6$		13.81		13.31		12.51	
20	ALO	.	.	13.05	80.71	11.65	77.51	10.73	77.70
	SLO	$> 10^6$		16.17		15.03		13.81	
25	ALO	.	.	15.49	80.51	13.22	76.86	12.01	77.78
	SLO	$> 10^6$		19.24		17.20		15.44	
30	ALO	.	.	18.81	80.25	15.47	77.31	13.70	77.80
	SLO	$> 10^6$		23.44		20.01		17.61	
49	ALO	.	.	85.60	61.75	50.61	68.11	43.00	76.87
	SLO	$> 10^6$		138.63		74.31		55.94	

b) Standardization according to (4):

$N = 10: \quad out(\alpha_{10}, F_1) = (5.28, \infty)$

k		ML-Oi	Pr	RCQ-Oi	Pr	RCS-Oi	Pr	SM-Oi	Pr
1	ALO	10.04	134.58	6.77	77.37	6.44	67.79	6.18	72.03
	SLO	7.46		8.75		9.50		8.58	
2	ALO	.	.	9.09	74.88	8.12	66.89	7.47	73.09
	SLO	32.37		12.14		12.14		10.22	
3	ALO	.	.	13.10	71.86	10.78	64.63	9.53	74.05
	SLO	$> 10^6$		18.23		16.68		12.87	

$N = 20 : \quad out(\alpha_{20}, F_1) = (5.97, \infty)$

k		ML-Oi	Pr	RCQ-Oi	Pr	RCS-Oi	Pr	SM-Oi	Pr
1	ALO	8.08	109.93	6.73	84.13	6.57	77.00	6.44	79.21
	SLO	7.35		8.00		8.54		8.13	
2	ALO	13.31	123.93	7.66	84.64	7.28	77.45	6.98	80.05
	SLO	10.74		9.05		9.40		8.72	
3	ALO	48.24	195.86	8.81	84.31	8.13	77.58	7.64	80.85
	SLO	24.63		10.45		10.48		9.45	
5	ALO	.	.	12.20	81.33	10.49	76.24	9.46	81.34
	SLO	$> 10^6$		15.00		13.76		11.63	
7	ALO	.	.	18.59	77.14	14.50	74.55	12.63	82.01
	SLO	$> 10^6$		24.10		19.45		15.40	

$N = 50 : \quad out(\alpha_{50}, F_1) = (6.88, \infty)$

k		ML-Oi	Pr	RCQ-Oi	Pr	RCS-Oi	Pr	SM-Oi	Pr
1	ALO	7.82	103.30	7.21	90.69	7.14	86.65	7.08	86.03
	SLO	7.57		7.95		8.24		8.23	
2	ALO	9.17	108.52	7.57	91.65	7.42	87.50	7.30	86.80
	SLO	8.45		8.26		8.48		8.41	
3	ALO	11.02	110.87	7.95	92.12	7.72	87.93	7.53	87.35
	SLO	9.94		8.63		8.78		8.62	
5	ALO	19.87	121.38	8.83	92.75	8.39	88.60	8.05	88.36
	SLO	16.37		9.52		9.47		9.11	
7	ALO	162.18	283.83	9.87	92.59	9.16	88.67	8.64	88.80
	SLO	57.14		10.66		10.33		9.73	
10	ALO	.	.	11.85	92.29	10.58	88.61	9.74	96.44
	SLO	$> 10^6$		12.84		11.94		10.10	
15	ALO	.	.	17.07	90.85	14.02	87.79	12.44	89.63
	SLO	$> 10^6$		18.79		15.97		13.88	
20	ALO	.	.	28.36	87.50	20.69	87.00	17.79	90.40
	SLO	$> 10^6$		32.41		23.79		19.68	

$$N = 100: \quad out(\alpha_{100}, F_1) = (7.58, \infty)$$

k		ML-Oi	Pr	RCQ-Oi	Pr	RCS-Oi	Pr	SM-Oi	Pr
1	ALO	8.11	100.12	7.76	93.72	7.72	91.25	7.69	91.00
	SLO	8.10		8.28		8.46		8.46	
2	ALO	8.75	102.46	7.94	94.19	7.87	91.72	7.81	91.24
	SLO	8.54		8.43		8.58		8.56	
3	ALO	9.51	103.59	8.14	94.76	8.02	92.08	7.92	91.56
	SLO	9.18		8.59		8.71		8.65	
5	ALO	11.58	104.80	8.55	95.42	8.34	92.67	8.17	92.11
	SLO	11.05		8.96		9.00		8.87	
7	ALO	15.00	106.76	8.99	95.74	8.68	93.03	8.44	92.65
	SLO	14.05		9.39		9.33		9.11	
10	ALO	28.19	114.08	9.73	95.77	9.24	93.33	8.87	93.10
	SLO	24.71		10.16		9.90		9.53	
15	ALO	.	.	11.19	95.80	10.33	93.57	9.70	93.54
	SLO	$> 10^6$		11.68		11.04		10.37	
20	ALO	.	.	13.05	95.53	11.65	93.50	10.73	93.71
	SLO	$> 10^6$		13.66		12.46		11.45	
25	ALO	.	.	15.49	95.32	13.22	92.64	12.01	93.83
	SLO	$> 10^6$		16.25		14.27		12.80	
30	ALO	.	.	18.81	95.00	15.47	93.25	13.70	94.00
	SLO	$> 10^6$		19.80		16.59		14.59	
49	ALO	.	.	85.60	73.14	50.61	82.16	43.00	92.80
	ALO	$> 10^6$		117.04		61.60		46.34	

Acknowledgement

This research has partially been supported by the Deutsche Forschungsgemeinschaft, SFB 475 "Reduction of Complexity for Multivariate Data Structures".

References

DAVIES, L. and U. GATHER (1989), The identification of multiple outliers, Technical report No. 89/1, Department of Statistics, University of Dortmund.

DAVIES, L. and U. GATHER (1993), The identification of multiple outliers, *Journal of the American Statistical Association* **88**, 782-792.

GATHER, U. (1990), Modelling the occurrence of multiple outliers, *Allgemeines Statistisches Archiv* **74**, 413-428.

GATHER, U. (1995), Outlier models and some related inferential issues, in: N. BALAKRISHNAN and A.P. BASU (eds.), *The exponential distribution: Theory, methods and applications*, Gordon and Breach Publishers, Amsterdam, 221-239.

GATHER, U. and V. SCHULTZE (1998), Robust estimation of scale of an exponential distribution, *Statistica Neerlandica*, to appear.

NELSON, W. (1982), *Applied life data analysis*, John Wiley and Sons, New York.

ROUSSEEUW, P.J. and C. CROUX (1993), Alternatives to the median absolute deviation, *Journal of the American Statistical Association* **88**, 1273-1283.