ELSEVIER Computational Statistics & Data Analysis 40 (2002) 741–758

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Robust estimation in very small samples

Peter J. Rousseeuw*, Sabine Verboven

Departement of Mathematics and Computer Science, Universitaire Instelling Antwerpen, Universiteitsplein 1, B-2610 Antwerpen, Belgium

Received 1 August 2001; received in revised form 1 February 2002

Abstract

In experimental science measurements are typically repeated only a few times, yielding a sample size n of the order of 3 to 8. One then wants to summarize the measurements by a central value and measure their variability, i.e. estimate location and scale. These estimates should preferably be robust against outliers, as reflected by their small-sample breakdown value. The estimator's stylized empirical influence function should be smooth, monotone increasing for location, and decreasing–increasing for scale. It turns out that location can be estimated robustly for $n \geqslant 3$, whereas for scale $n \geqslant 4$ is needed. Several well-known robust estimators are studied for small n, yielding some surprising results. For instance, the Hodges–Lehmann estimator equals the average when $n=4$. Also location M-estimators with auxiliary scale are studied, addressing issues like the difference between one-step and fully iterated M-estimators. Simultaneous M-estimators of location and scale ('Huber's Proposal 2') are considered as well, and it turns out that their lack of robustness is already noticeable for such small samples. Recommendations are given as to which estimators to use.

Keywords: Breakdown value; Stylized empirical influence function; Small data sets; Three-dimensional empirical influence function

1. Introduction

In practice, it often happens that we need to robustly estimate location and/or scale from a very small sample. The sample size n is often constrained by the cost of an observation. In many experimental settings (e.g. in chemistry) one will typically repeat each measurement only a few times, say $n \leqslant 5$. Even a small sample may contain aberrant values due to technical problems or measurement inaccuracies

* Corresponding author.
E-mail address: peter.rousseeuw@ua.ac.be (P.J. Rousseeuw).

0167-9473/02/$ - see front matter
PII: S0167-9473(02)00078-6

for example, so we want our estimates to be robust. Currently, most practitioners use less efficient procedures, such as subjective deletion of outliers or some ad hoc outlier flagging rules. In this paper we will see that the well-known medium- and large-sample behavior of the usual robust estimators does not necessarily extend to sample sizes as small as $n = 3$ or 4, and we will make some suggestions for practical use.

Consider the following data provided by the Department of Chemistry at the University of Antwerp. In many glass samples the concentration of a given compound needed to be estimated. For this purpose, 4 to 5 measurements were taken in each sample. The usual concentration estimate is then the average of these 4–5 measurements. However, this estimate can be way off due to the presence of outliers. For example, in one sample the measured concentrations of SiO_2 were 68.52, 68.23, 67.42, 68.94, and 68.34 (in units of weight percent). The usual average of these five values is 68.29, whereas their median is 68.34. If the first measurement were wrongly recorded as 18.52 the average would become 58.29, i.e. the average is strongly attracted by the outlier. On the other hand, the median becomes 68.23 so it is not changed by much. This illustrates the usefulness of a robust method, especially if the measurement process is computerized (i.e. the measurement instrument sends the data directly to the computer, which processes it without human intervention).

Throughout this paper we will consider the general case where both the location θ and the scale σ are unknown. The underlying model states that the observations x_i are i.i.d. with distribution function $F((x - \theta))/\sigma$ where F is known. (Typically, F is the standard gaussian distribution function Φ.) It is thus natural to impose *location-scale equivariance*. That is, if we replace our data set $X = (x_1, \ldots, x_n)$ by $aX + b = (ax_1 + b, \ldots, ax_n + b)$ then a location estimator T_n and a scale estimator S_n must satisfy

$$T_n(aX + b) = aT_n(X) + b, \tag{1}$$

$$S_n(aX + b) = |a|S_n(X), \tag{2}$$

where a and b are any real numbers. We also impose *permutation invariance*, i.e. the observations $x_1, \ldots, x_n$ may be put in any order without affecting the estimates.

One way to describe the robustness of T_n or S_n is by its empirical influence function (EIF). The EIF of an estimator shows what happens with the estimate when varying one observation from $-\infty$ to $+\infty$. In general, the EIF of an estimator T_n at a univariate sample $(x_1, \ldots, x_{n-1})$ is defined as the function

$$\mathrm{EIF}(x) = T_n(x_1, \ldots, x_{n-1}, x), \tag{3}$$

where x ranges from $-\infty$ to $+\infty$. In order to avoid the dependence of (3) on the particular sample $(x_1, \ldots, x_{n-1})$ we will use a constructed sample. This artificial sample consists of $m = n - 1$ quantiles of the standard normal distribution, where the x_i are defined as

$$x_i = \Phi^{-1}\left(\frac{i - 1/3}{m + 1/3}\right) \quad \text{for } i = 1, \ldots, m \tag{4}$$

(Andrews et al., 1972). Such a 'stylized' sample gives a better approximation to Φ than a random sample. The resulting function (3) will be called the *stylized empirical influence function* (SEIF) of T_n. [Note that for $n = 3$ the SEIF of T_3 completely determines T_3 itself on all data sets, due to (1)–(2) and permutation invariance.]

In the next section we will consider simple location estimators, whereas Section 3 focuses on scale estimators. Section 4 studies M-estimators of location and scale (estimated separately or simultaneously). For the sake of completeness, in Section 5 we will briefly deal with the (rare) case where either location or scale is assumed to be known in advance. Section 6 contains our conclusions and some topics for further research.

2. Robust location estimators with explicit formulas

Any permutation invariant location estimator T_n satisfying (1) equals x_1 for $n = 1$ and equals $(x_1 + x_2)/2$ for $n = 2$. Hence, location estimators can only differ when $n \geqslant 3$.

2.1. Some simple estimators

The most well-known location estimator is of course the *sample average* (also called the sample mean). Given a data set $X = (x_1, \ldots, x_n)$ we denote it as

$$\mathrm{ave}_n(X) = \operatorname*{ave}_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{5}$$

The runner-up in popularity is the *sample median*, given by

$$\mathrm{med}_n(X) = \operatorname*{med}_{i=1}^{n} x_i = \begin{cases} x_{\frac{n+1}{2}:n} & \text{when } n \text{ is odd,} \\ \frac{1}{2}(x_{\frac{n}{2}:n} + x_{\frac{n}{2}+1:n}) & \text{when } n \text{ is even,} \end{cases} \tag{6}$$

where $x_{1:n} \leqslant x_{2:n} \leqslant \cdots \leqslant x_{n:n}$ are the ordered observations. The first time people encounter the notion of robust statistics is often when comparing the behavior of the sample average (which can be moved as far as you want by moving one observation) with that of the sample median (which moves very little in that situation).

We will also look at the *Hodges–Lehmann estimator* (HL) which is well-known in the robustness literature (Hampel et al., 1986). It is defined as

$$\mathrm{HL}_n(X) = \mathrm{med}\left\{\left(\frac{x_i + x_j}{2}\right);\ 1 \leqslant i < j \leqslant n\right\}. \tag{7}$$

As our fourth location estimator we consider the (k/n)-*trimmed average*, which is the average of the data set except the k smallest and the k largest observations, i.e.

$$(k/n) - \text{trimmed average} = \mathrm{ave}\{x_{k+1:n}, \ldots, x_{n-k:n}\}. \tag{8}$$

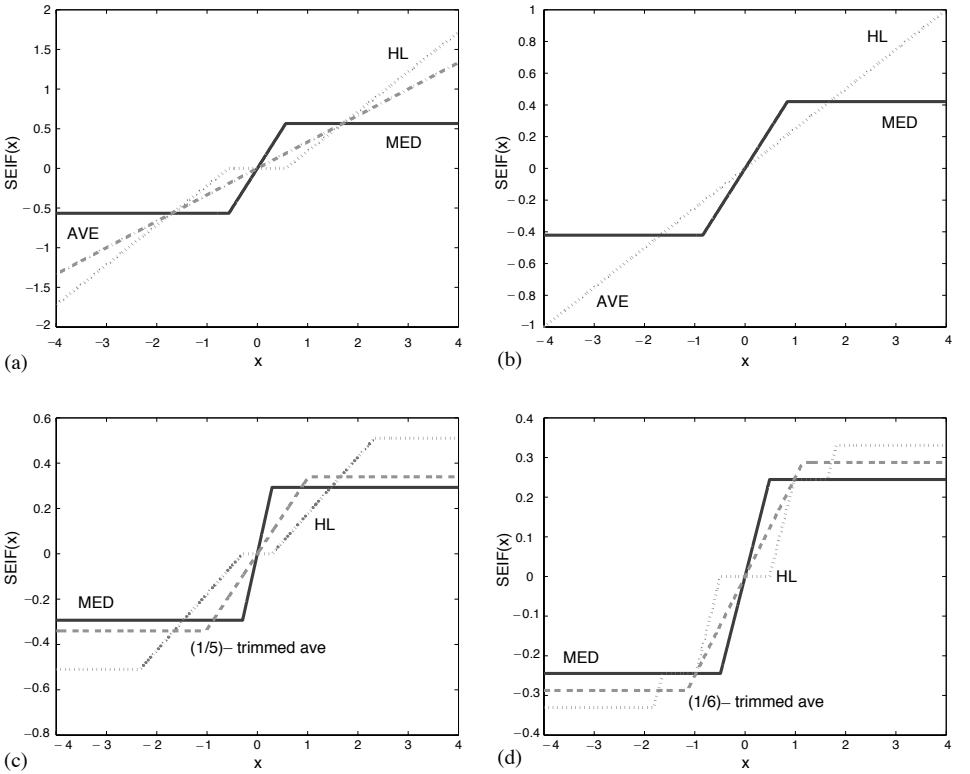

Fig. 1. Stylized empirical influence functions of the average (dot–dashed line), the median (solid line), the Hodges–Lehmann estimator (dotted line), and the $(1/n)$-trimmed average (dashed line) for (a) $n = 3$, (b) $n = 4$, (c) $n = 5$, and (d) $n = 6$.

2.2. Stylized empirical influence functions

We will first compare these location estimators by means of their SEIF's. For any n the average has $\mathrm{SEIF}(x) = x/n$ which is a straight line with slope $1/n > 0$, whereas the SEIF of a robust location estimator should be bounded.

For $n = 3$ our stylized observations (4) become $x_1 = \Phi^{-1}(2/7) = -0.57$ and $x_2 = \Phi^{-1}(5/7) = 0.57$. Fig. 1a shows the SEIF's for $n=3$. The one of the median is bounded. (Note that the $(0/3)$-trimmed average is the average, and the $(1/3)$-trimmed average is the median here.) The SEIF of HL_3 is not bounded, and in fact its slope is $\frac{1}{2}$ which makes HL_3 even more sensitive to an outlier than the average (whose slope is lower). This is due to

Remark 1. For any $X = (x_1, x_2, x_3)$ it holds that $\mathrm{HL}_3(X) = (x_{1:3} + x_{3:3})/2$.

Proof. Assume w.l.o.g. that $x_1 \leqslant x_2 \leqslant x_3$, then $\mathrm{HL}_3(X) = \frac{1}{2}\mathrm{med}\{x_1 + x_2, x_1 + x_3, x_2 + x_3\} = (x_1 + x_3)/2$. □

For $n = 4$ we see the SEIF of the average and the median in Fig. 1b. (Note that the $(k/4)$-trimmed average brings nothing new here, since $k = 0$ yields the average and $k = 1$ yields the median.) Surprisingly, the SEIF of the Hodges–Lehmann estimator coincides with that of the average. This made us realize that:

Remark 2. For any $X = (x_1, x_2, x_3, x_4)$ it holds that $\mathrm{HL}_4(X) = \mathrm{ave}_4(X)$.

Proof. Assuming $x_1 \leqslant x_2 \leqslant x_3 \leqslant x_4$ we find that

$$\mathrm{HL}_4(X) = \tfrac{1}{2}\mathrm{med}\{x_i + x_j; 1 \leqslant i < j \leqslant 4\}.$$

These six sums satisfy the inequalities

$$x_1 + x_2 \leqslant x_1 + x_3 \leqslant x_1 + x_4 \leqslant x_2 + x_4 \leqslant x_3 + x_4$$

and

$$x_1 + x_2 \leqslant x_1 + x_3 \leqslant x_2 + x_3 \leqslant x_2 + x_4 \leqslant x_3 + x_4$$

so the two middle values must be $x_1 + x_4$ and $x_2 + x_3$ (the ordering between these two is immaterial). Therefore, always

$$\mathrm{HL}_4(X) = \frac{1}{2}\left(\frac{(x_1 + x_4) + (x_2 + x_3)}{2}\right) = \mathrm{ave}_4(X). \qquad \square$$

For $n \leqslant 4$ we have seen that among our estimators only the median has a bounded SEIF. For $n \geqslant 5$ this is no longer the case, e.g. the $(1/5)$-trimmed average and HL_5 also have a bounded SEIF. In fact, for $n \geqslant 5$ the SEIF only eliminates the sample average. In order to differentiate between the robustness of $\mathrm{med}_n, \mathrm{HL}_n$ and the (k/n)-trimmed average we will use the following tool.

2.3. The breakdown value

The breakdown value is a simple but far-reaching tool, which asks *how many* outliers in the sample can make the estimate useless. (By contrast, the SEIF describes what happens for *one* outlier.) We say that a location estimator T_n breaks down (i.e. becomes useless) when $|T_n|$ can become arbitrarily large (that is, when T_n can be pulled to $+\infty$ or to $-\infty$).

Formally, we start with a data set $X = (x_1, \ldots, x_n)$ in which no two observations coincide (this is called *general position*). Next, consider all data sets Z^m obtained by replacing any m data points $x_{i_1}, \ldots, x_{i_m}$ by arbitrary values $y_1, \ldots, y_m$. Then the breakdown value of the location estimator T_n at X is defined as

$$\varepsilon_n^*(T_n; X) = \min\left\{\frac{m}{n};\ \sup_{Z^m} |T_n(Z^m)| = \infty\right\}. \tag{9}$$

The breakdown value of the average is $\varepsilon_n^*(\mathrm{ave}_n, X) = 1/n$ for any $n \geqslant 1$, so the average is never robust. One can prove that the breakdown value of any equivariant location

estimator (i.e. (1) must hold) satisfies

$$\varepsilon_n^*(T_n,X) \leqslant \frac{\lceil n/2 \rceil}{n} \tag{10}$$

and that the sample median attains this upper bound. (Here, the ceiling $\lceil t \rceil$ is the smallest integer $\geqslant t$.) The other estimators in this section have a lower breakdown value, e.g. that of the (k/n)-trimmed average is $(k+1)/n$ and that of HL_n tends to $1-2^{-1/2} \approx 29\%$.

2.4. Discussion

For $n=1$ the only equivariant location estimator is $T_1 = x_1$, and for $n=2$ we must have $T_2 = \mathrm{ave}(x_1,x_2)$. These estimators are not robust because they cannot withstand even one outlier. For $n=3$ only one of our location estimators is robust, namely the sample median. All the other location estimators we consider in this paper have a breakdown value $\varepsilon_3^* = \frac{1}{3}$ which means that they cannot withstand a single outlier. Moreover the SEIF of the median is monotone, in the sense that for any $x < y$ it holds that $\mathrm{SEIF}(x) \leqslant \mathrm{SEIF}(y)$. Therefore, for $n=3$ we will only use the sample median.

For $n \geqslant 4$, the breakdown considerations in Section 2.3 have eliminated $\mathrm{ave}_n, \mathrm{HL}_n$ and the (k/n)-trimmed average. Among the four estimators considered here, the sample median thus emerges as the clear winner. Its SEIF is always monotone, but it is not smooth. (Making use of the median, we will construct an estimator with smooth SEIF for $n \geqslant 4$ in Section 4.)

3. Robust scale estimators with explicit formulas

3.1. Some simple scale estimators

By Eq. (2) any equivariant scale estimator is zero for $n=1$ and is a multiple of $|x_2-x_1|$ for $n=2$, so we will focus on $n \geqslant 3$. The scale of $X=(x_1,\ldots,x_n)$ is typically estimated by the standard deviation

$$\mathrm{SD}_n(X) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mathrm{ave}_n(X))^2} \tag{11}$$

which is very sensitive to outliers. Here we will consider scale estimators that are more robust. A first approach is to replace the squares by absolute values and to remove the square root, leading to the average distance to the average (ADA) which is sometimes used. We prefer the average distance to the median (ADM) given by

$$\mathrm{ADM}_n(X) = \operatorname*{ave}_{i=1}^{n} |x_i - \mathrm{med}_n(X)| \tag{12}$$

which agrees with the L^1 definition of $\mathrm{med}_n(X)$ and the definition of a medoid in the k-medoid clustering method (Kaufman and Rousseeuw, 1990). Of course we can also replace the remaining average in (12) by a median, yielding the median distance to the

median (MDM). Traditionally this estimator goes by the unfortunate acronym MAD (for *median absolute deviation*), and is given by

$$\mathrm{MAD}_n(X) = b_n 1.4826 \operatorname*{med}_{i=1}^{n} |x_i - \mathrm{med}_n(X)|. \tag{13}$$

The estimator Q of Rousseeuw and Croux (1993) is given by

$$Q_n(X) = c_n 2.2219\{|x_i - x_j|; i < j\}_{(l)} \tag{14}$$

where the last part is the lth ordered value among this set of $\binom{n}{2}$ values, where $l = \binom{h}{2} \approx \binom{n}{2}/4$ with $h = \lfloor n/2 \rfloor + 1$. (The floor $\lfloor t \rfloor$ is the largest integer $\leqslant t$.) The constants b_n and c_n are small-sample correction factors which make MAD_n and Q_n unbiased.

We also consider another type of scale estimator which we will call a *trimmed range*. For any integer k with $0 \leqslant k < n/2$ we define the (k/n)-trimmed range as the range of the data set except the k smallest and the k largest observations. Formally

$$(k/n)\text{-trimmed range} = |x_{n-k:n} - x_{k+1:n}|. \tag{15}$$

For instance, for $k = 0$ we obtain the $(0/n)$-trimmed range which is the (non-robust) *range* $|x_{n:n} - x_{1:n}|$, and for some combinations of k and n we obtain the interquartile range. (We will not concern ourselves here with correction factors to make the (k/n)-trimmed range unbiased.)

3.2. Stylized empirical influence functions

The solid line in Fig. 2a is the SEIF of the MAD, given by $\mathrm{SEIF}(x) = \mathrm{MAD}_3(-0.57, 0.57, x)$ according to (3). This SEIF has several peaks. Most importantly, $\mathrm{SEIF}(x)$ attains the value zero at $x = \pm 0.57$ because in those x we have $\mathrm{MAD}_3 = \mathrm{med}\{0, 0, 2(0.57)\} = 0$. We say that the MAD *implodes* at these samples. To see why implosion creates problems, consider the fact that standardized observations have a scale estimate in their denominator.

On the other hand, MAD_3 does not explode (i.e. become arbitrarily large) because its SEIF is bounded. An example of the opposite behavior is the $(0/3)$-trimmed range, i.e. the range, which is also plotted in Fig. 2a: its SEIF stays away from zero but is unbounded. The same is true for the ADM (12), whose SEIF goes to infinity more slowly than that of the range (this effect becomes more pronounced for larger n).

In Fig. 2a we note that the SEIF of the MAD coincides with that of Q. This is due to the following property.

Remark 3. For any $X = (x_1, x_2, x_3)$ it holds that $Q_3(X) = \mathrm{MAD}_3(X)$.

Proof. Assume without loss of generality that $x_1 \leqslant x_2 \leqslant x_3$. Then

$$\begin{aligned} \mathrm{MAD}_3(X) &= \operatorname*{med}_{i=1}^{3} |x_i - \mathrm{med}_n(X)| \\ &= \mathrm{med}\{|x_1 - x_2|, |x_2 - x_2|, |x_3 - x_2|\} \\ &= \min\{x_2 - x_1, x_3 - x_2\} \end{aligned}$$

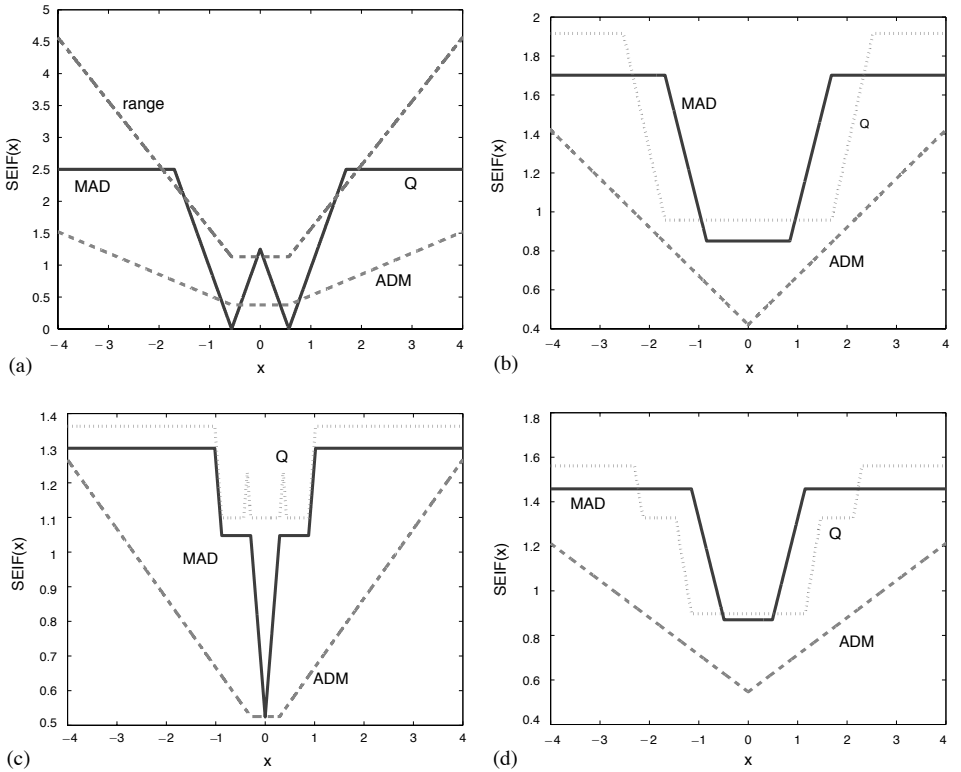

Fig. 2. Stylized empirical influence functions of (a) MAD and Q (solid line) and range and ADM (dashed lines) for $n = 3$; (b) the MAD (solid), Q (dotted), and ADM (dashed) for $n = 4$; (c) for $n = 5$; and (d) for $n = 6$.

and

$$\begin{aligned} Q_3(X) &= \{|x_i - x_j|; 1 \leqslant i < j \leqslant 3\}_{(l)} \\ &= \{|x_1 - x_2|, |x_1 - x_3|, |x_2 - x_3|\}_{(1)} \\ &= \min\{x_2 - x_1, x_3 - x_2\} \end{aligned}$$

since $h = \lfloor n/2 \rfloor + 1$ hence $l = \binom{h}{2} = 1$. □

In Fig. 2b we see that the MAD and Q no longer coincide for $n=4$, and their SEIF's are bounded and stay away from zero. Unfortunately the SEIF's are not smooth, since both have four corner points. The SEIF's for $n = 5$ up to $n = 10$ (not all shown here) are not smooth either. In order to further distinguish between the scale estimators (11) –(15) we will consider their breakdown values below.

3.3. The breakdown value

Scale estimators can break down in two ways: when they become arbitrarily large (explosion) or when they become arbitrarily close to zero (implosion). Formally, we define the *explosion breakdown value* of a scale estimator S_n at X as

$$\varepsilon_n^+(S_n,X) = \min\left\{\frac{m}{n}; \sup_{Z^m} S_n(Z^m) = \infty\right\} \tag{16}$$

with the notation Z^m of (9). Analogously, the *implosion breakdown value* is defined as

$$\varepsilon_n^-(S_n,X) = \min\left\{\frac{m}{n}; \inf_{Z^m} S_n(Z^m) = 0\right\}. \tag{17}$$

Naturally, we want to protect ourselves against explosion as well as implosion, so we want both $\varepsilon_n^+(S_n;X)$ and $\varepsilon_n^-(S_n;X)$ to be high enough. Therefore, we define the (overall) breakdown value of S_n at X as

$$\varepsilon_n^*(S_n,X) = \min\{\varepsilon_n^+(S_n;X), \varepsilon_n^-(S_n;X)\}. \tag{18}$$

For the SD, the ADM and the range it holds that $\varepsilon_n^+(S_n;X) = 1/n$ and $\varepsilon_n^-(S_n;X) = (n-1)/n$ hence $\varepsilon_n^*(S_n;X) = 1/n$, which means that even a single outlier can break them down. On the other hand, Croux and Rousseeuw (1992) showed that the breakdown value of any equivariant scale estimator satisfies

$$\varepsilon_n^*(S_n,X) \leqslant \frac{\lfloor n/2 \rfloor}{n}. \tag{19}$$

This upper bound is actually attained for the MAD and Q.

For samples of size $n = 3$ any scale estimator must satisfy $\varepsilon_3^*(S_3,X) \leqslant \frac{1}{3}$ by (19), so it is always possible to make S_3 either explode or implode by replacing a single data point. Therefore, we cannot protect against explosion and implosion at the same time. It is possible to protect against explosion alone, e.g. by using the MAD for which $\varepsilon_3^+(\mathrm{MAD}_3,X) = \frac{2}{3}$ but $\varepsilon_3^-(\mathrm{MAD}_3,X) = \frac{1}{3}$. On the other hand, the SD, the ADM and the range protect against implosion since $\varepsilon_3^-(S_3,X) = \frac{2}{3}$, but $\varepsilon_3^+(S_3,X) = \frac{1}{3}$.

It is only possible to estimate scale robustly for $n \geqslant 4$. In that case, both the MAD and Q have the maximal breakdown value $\varepsilon_n^*(S_n,X) = \frac{\lfloor n/2 \rfloor}{n} > 1/n$ so they can withstand at least one outlier.

Note that for $n \geqslant 4$ the breakdown value of the (k/n)-trimmed range is always less than that of the MAD. For instance, let $n = 4$. Then the $(0/4)$-trimmed range, i.e. the usual range, explodes if we move one data point far away. On the other hand, the $(1/4)$-trimmed range can implode by moving one data point. To see this, suppose $x_1 < x_2 < x_3 < x_4$ and replace x_3 by $y_3 := x_2$; then the $(1/4)$-trimmed range becomes $|x_2 - x_2| = 0$.

3.4. Discussion

When $n = 1$ we obviously cannot estimate scale, and for $n = 2$ the only scale estimate is $|x_2 - x_1|$ which can be made to implode as well as explode. Also for $n = 3$ it

is impossible to estimate scale robustly, in the sense that we cannot protect against implosion and explosion at the same time. If the context is such that explosion should be avoided, we recommend to use the MAD. If implosion should be avoided, the ADM is our best choice because its SEIF has the lowest slope among the estimators considered here. (When the goal is to standardize data 'as robustly as possible', as in Kaufman and Rousseeuw (1990), one can subtract the median first and then divide by the MAD if it is nonzero, and divide by the ADM otherwise.)

For $n \geqslant 4$ the breakdown arguments above have eliminated the SD, ADM, and (k/n)-trimmed range, hence only the MAD and Q remain. Since the SEIF of the MAD is more smooth than that of Q for small n, we prefer to use the MAD. In the next section we will encounter more advanced estimators, which use the median and the MAD in their construction.

4. *M*-estimators of location and scale

4.1. Location *M*-estimators with auxiliary scale

The location estimators we saw in Section 2 did not use any scale estimate. For the location described below this will be different, since we first need to compute a robust scale estimate $S_n(x_1,\ldots,x_n)$. Since scale cannot be estimated robustly for $n \leqslant 3$ (see Section 3), we can only use this approach when $n \geqslant 4$. We then define the *M-estimator* T_n (Huber, 1981) as solution of the equation

$$\operatorname*{ave}_{i=1}^{n} \psi\left(\frac{x_i - T_n}{S_n}\right) = 0. \tag{20}$$

(Note that without S_n in the denominator, T_n would not be equivariant.) For S_n we will take $\mathrm{MAD}_n(X)$. We will use odd, continuous and monotone functions ψ. We can compute T_n by the Newton–Raphson algorithm, starting from the initial location estimate $T_n^{(0)} = \mathrm{med}_n(X)$. From each $T_n^{(k-1)}$ we compute the next $T_n^{(k)}$ by

$$T_n^{(k)} = T_n^{(k-1)} + S_n \frac{\mathrm{ave}_{i=1}^{n}\psi((x_i - T_n^{(k-1)})/S_n)}{\mathrm{ave}_{i=1}^{n}\psi'((x_i - T_n^{(k-1)})/S_n)} \tag{21}$$

and we keep doing this until $T_n^{(k)}$ converges. Note that sometimes the denominator of (21) could be close to zero. It is well-known in the robustness folklore that it is safer to replace the denominator by the constant $\int \psi'(u)\,\mathrm{d}\Phi(u)$. Throughout this paper we will use the latter version.

Instead of the fully iterated M-estimators given by (21), it is possible to take only a single iteration step. This yields the *one-step M-estimator*

$$T_n^{(1)} = T_n^{(0)} + S_n \frac{\mathrm{ave}_{i=1}^{n}\psi((x_i - T_n^{(0)})/S_n)}{\int \psi'(u)\,\mathrm{d}\Phi(u)}. \tag{22}$$

Note that $T_n^{(1)}$ is not an exact solution of (20), but it has an explicit formula and is easier to compute. Moreover, asymptotically (for $n \to \infty$) and in simulations with

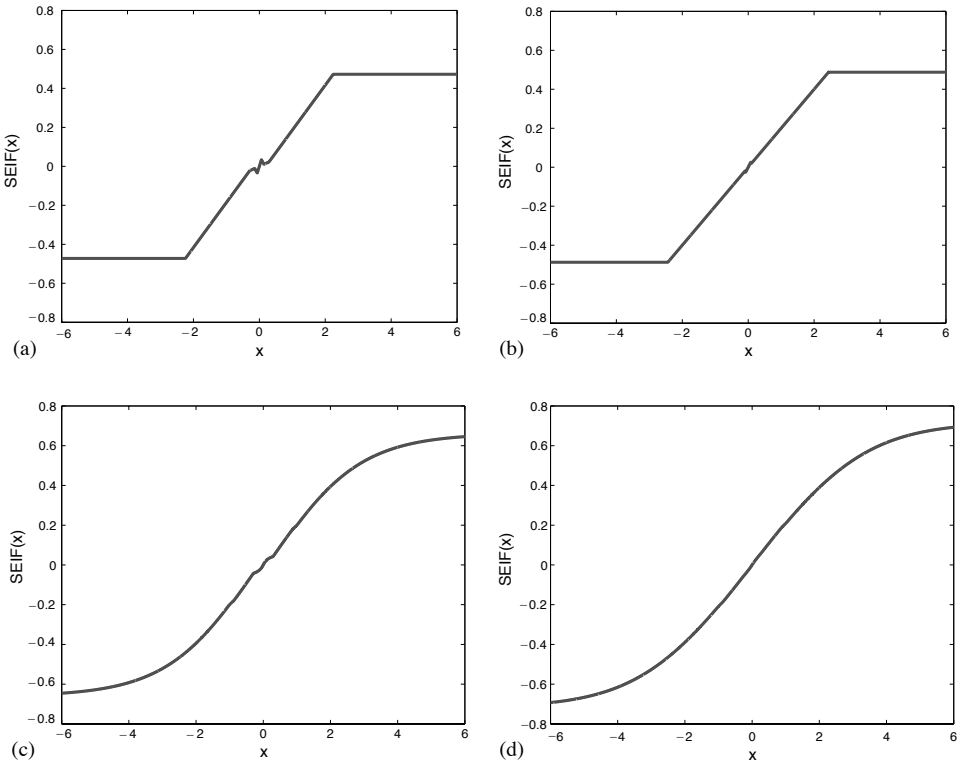

Fig. 3. SEIF's of location M-estimation with auxiliary scale, all for $n = 5$: (a) one-step with Huber's ψ_b; (b) fully iterated with Huber's ψ_b; (c) one-step with logistic ψ; (d) fully iterated with logistic ψ.

medium-sized n (Andrews et al., 1972) it turns out that $T_n^{(1)}$ behaves as well as T_n. Therefore, many people prefer $T_n^{(1)}$ over T_n. Here, however, we ask ourselves which of them behaves best for very small n.

We cannot choose between $T_n^{(1)}$ and T_n on the basis of their breakdown value, since for any monotone and bounded ψ-function their breakdown values coincide and attain the upper bound in (10). But their SEIF's do not coincide, and we want the SEIF to be as smooth and as monotone as possible. Naturally, the SEIF also depends on the function ψ. Huber (1981) proposed the function $\psi_b(x) := \min\{b, \max\{-b, x\}\}$ which is now named after him, where typically $b = 1.5$. Fig. 3a shows the SEIF of the one-step M-estimator based on ψ_b. We note that the SEIF has sharp corners near $x = \pm 2$, which are due to the corners in ψ_b itself. Moreover there are some strange bumps near $x = 0$, which destroy the monotonicity of the SEIF. As already noted in Rousseeuw (1994) this lack of monotonicity is due to the auxiliary scale estimate, since the MAD becomes small for x near zero and its own SEIF has corners there. In Fig. 3b we note that the SEIF of the fully iterated M-estimator based on ψ_b has smaller bumps near zero than the one-step version. The figures for other small values of n (not given here) show the same effect. We conclude that for small n and monotone ψ the fully

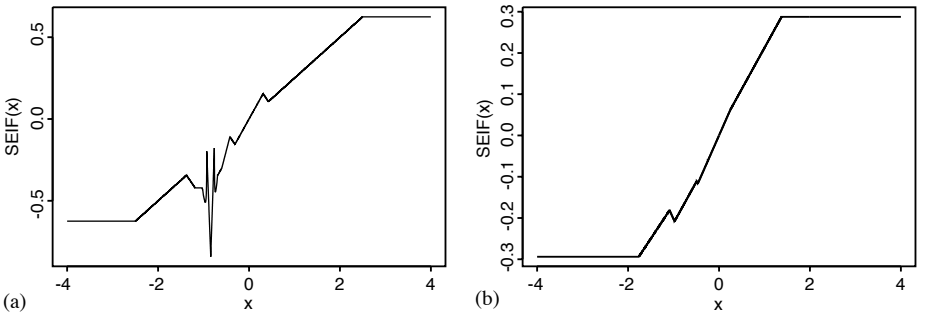

Fig. 4. SEIF's of the fully iterated location M-estimator with Huber's ψ_b as obtained from the S-Plus function *location.m*, for (a) $n=4$ and (b) $n=6$.

iterated version is a lot smoother than the one-step version, which is opposite to the usual recommendation in favor of the one-step version for large n.

Of course Fig. 3b still has sharp corners near ± 2 due to the formula of ψ_b. Therefore we prefer to use a smooth ψ-function. We recommend

$$\psi_{\log}(x) = \frac{e^x - 1}{e^x + 1}, \tag{23}$$

which can also be written as $\psi_{\log} = 2F(x) - 1$ where $F(x) = \text{expit}(x) = 1/(1+e^{-x})$ is the cumulative distribution function of the logistic distribution. Note that F is also known as the sigmoid function in the neural net literature. The function $\psi_{\log}(x)$ is clearly odd, monotone and bounded. Compared to ψ_b it has the advantage that its derivative $\psi'_{\log}(x)$ exists in all x and is always strictly positive. Since $\psi_{\log}$ is continuous and strictly increasing, the solution T_n to (20) always exists and is unique. The denominator of (21) and (22) can now be replaced by the value $\int \psi'_{\log}(u)\,d\Phi(u) = 0.4132$.

In Figs. 3c and d we see that the logistic ψ-function (23) yields a much smoother SEIF than the Huber ψ_b in Figs. 3a and b. Moreover, the MAD-induced bumps in the one-step logistic M (see Fig. 3c) become nearly invisible in the fully iterated version (Fig. 3d). Among location M-estimators with monotone ψ we thus prefer the logistic ψ-function (23) and full iteration (21). The result has the maximal breakdown value (10) and a bounded and nearly smooth SEIF.

Remark. The location M-estimator with Huber's ψ_b-function is already available as the intrinsic function *location.m* in S-Plus. This function contains some errors however, as we found out since we started our computations in S-Plus rather than Matlab. Fig. 4a is the SEIF of the output of the function location.m (in S-Plus 2000, release 3) for $n=4$, where the auxiliary scale was the MAD and we selected 50 iteration steps to emulate full iteration. We took $b=1.5$ as before. In addition to the two corner points near ± 2 as in Fig. 1b, we now see a lot of sharp peaks. The SEIF in Fig. 4a is not symmetric, which it should be by affine equivariance: take $a=-1$ and $b=0$ in (1). For $n=5$ the SEIF looked normal, but the one for $n=6$ (Fig. 4b) is again impossible. The error turned out to be that $T_n^{(0)}$ in (22) was not $\text{med}_n(X)$ but the *low median*,

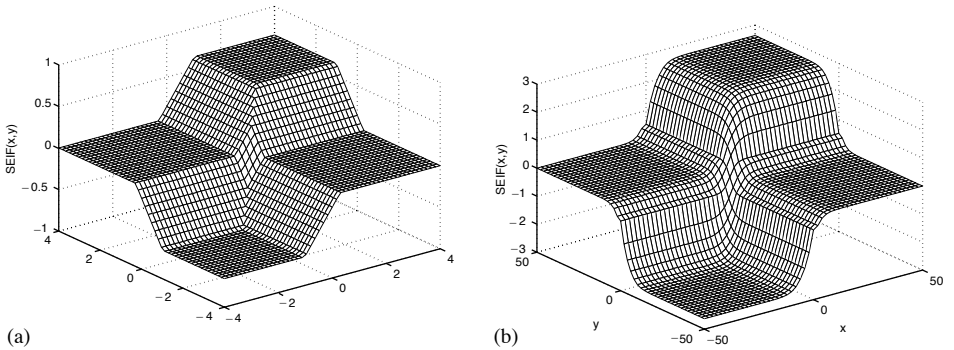

Fig. 5. The 3D-SEIF of (a) the sample median, and (b) the location M-estimator with logistic ψ, both for $n = 6$.

which is defined as $x_{\lfloor (n+1)/2 \rfloor :n}$ for both odd n and even n, and that S_n was not the real MAD but a version in which both medians in (13) had been replaced by low medians as well. These errors have little effect for large n, but make a lot of difference in Fig. 4. This illustrates that statisticians have often paid too little attention to small sample sizes. On the positive side, it also illustrates that the SEIF can be a useful tool to check programs.

An interesting extension of the SEIF, which has not been used before to our knowledge, is to allow *two* new values x and y in the sample. More precisely, we define the three-dimensional SEIF as

$$\text{3D-SEIF}(x, y) = T_n(x_1, \ldots, x_{n-2}, x, y), \tag{24}$$

where $(x_1, \ldots, x_{n-2})$ is a stylized sample (4) with $m = n - 2$. The plot of (24) as a function of x and y is three-dimensional and shows what happens when x and y are outlying in the same direction or in opposite directions. The 3D-SEIF of the median (Fig. 5a) is bounded, since two outliers out of $n = 6$ cannot cause breakdown. It is piecewise planar, with several edges that can be interpreted. The 3D-SEIF of our recommended M-estimator (Fig. 5b) is still bounded but much smoother.

4.2. Scale M-estimators with auxiliary location

For M-estimators of scale we first need an auxiliary location estimate $T_n(x_1, \ldots, x_n)$. Following Section 2 we will take $T_n(X) = \text{med}_n(X)$. A scale M-estimator S_n is then defined as solution of

$$\underset{i=1}{\overset{n}{\text{ave}}}\, \rho\left(\frac{x_i - T_n}{S_n}\right) = \beta, \tag{25}$$

where $\rho(0) = 0$, $\rho(-x) = \rho(x)$, $\rho(x)$ is monotone for $x \geqslant 0$, ρ is bounded, and the constant β is set equal to $\int \rho(u)\, \mathrm{d}\Phi(u)$. Huber's favorite choice for ρ was $\rho_b(x)=$

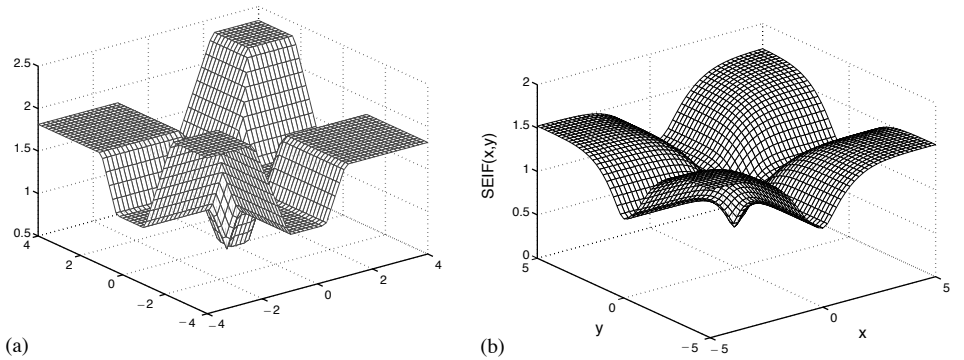

Fig. 6. The 3D-SEIF of (a) the MAD and (b) the scale M-estimator with logistic ρ, both for $n = 6$.

$\psi_b^2(x) = \min\{x^2, b^2\}$ which has corners at $x = \pm b$ and leads to a breakdown value below 50% (for $b = 1.5$ we find $\varepsilon_n^*(S_n) \to \beta/\rho(\infty) = 35\%$).

We prefer a smooth ρ-function, which at the same time leads to a 50% breakdown value. For this we propose

$$\rho_{\log}(x) = \psi_{\log}^2\left(\frac{x}{0.3739}\right) \tag{26}$$

with $\psi_{\log}$ as in (23), hence $\rho(\infty) = 1$. The constant 0.3739 was obtained from the condition that $\beta = \frac{1}{2}$ which yields the 50% breakdown value. Since $\rho_{\log}$ is continuous and strictly increasing for positive arguments, the solution S_n to (25) always exists and is unique.

To compute S_n we start from the initial scale estimate $S_n^{(0)} = \mathrm{MAD}_n(X)$ and then take iteration steps given by

$$S_n^{(k)} = S_n^{(k-1)} \sqrt{2 \operatorname*{ave}_{i=1}^{n} \rho_{\log}((x_i - T_n)/S_n^{(k-1)})} \tag{27}$$

until convergence. (As in the location case, taking only one step yields a more 'bumpy' estimator when n is small.)

Fig. 6a shows the 3D-SEIF of the MAD for $n = 6$, which is piecewise planar with many edges. In contrast, the 3D-SEIF of our scale M-estimator in Fig. 6b is smooth (and still bounded). Note that the MAD is the starting value for the iterations (27), but that also the auxiliary location $T_n = \mathrm{med}_n(X)$ plays an essential role in (25) and (27). We have thus combined the best estimators of Sections 2 and 3, which have bumpy 3D-SEIF's (Figs. 5a and 6a), to yield the smooth M-estimator of scale. The same is true for the smooth M-estimator of location in Section 4.1, which also uses both the median and the MAD.

4.3. Huber's proposal 2

Instead of defining T_n by (20) with auxiliary MAD scale, and defining S_n by (25) with auxiliary median location, Huber (1981, Section 6.4) proposed to define T_n and

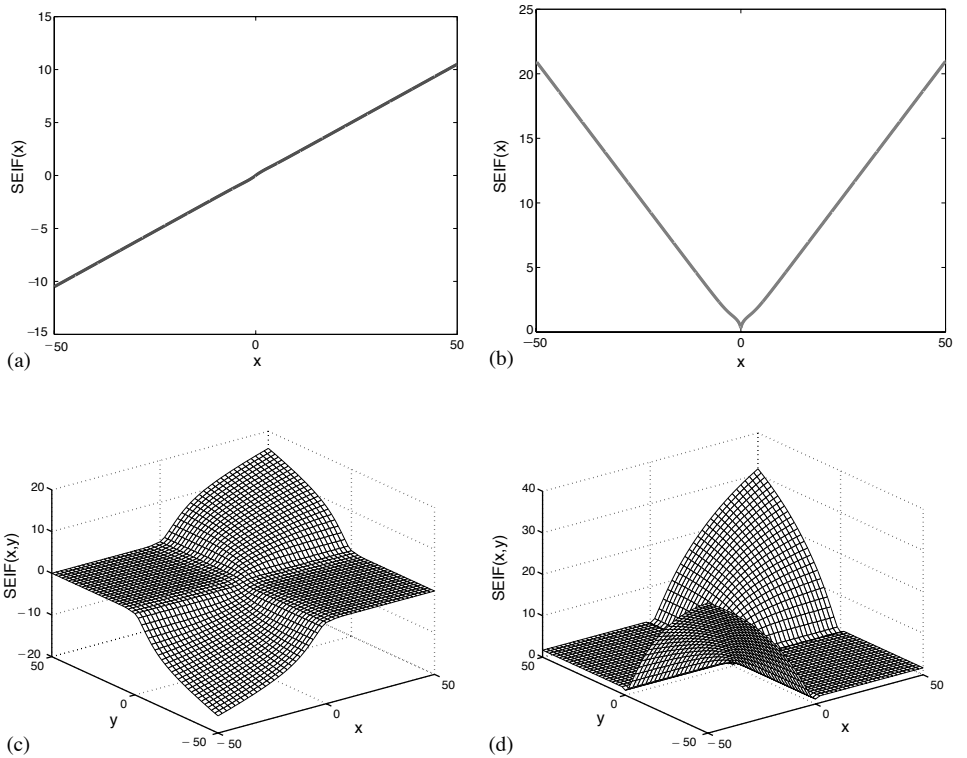

Fig. 7. Simultaneous M-estimation of location and scale according to (28) using the functions $\psi_{\log}$ and $\rho_{\log}$: (a) SEIF of T_4; (b) SEIF of S_4; (c) 3D-SEIF of T_6; and (d) 3D-SEIF of S_6.

S_n as simultaneous solutions of the equations

$$\underset{i=1}{\overset{n}{\text{ave}}}\,\psi\left(\frac{x_i - T_n}{S_n}\right) = 0 \quad \text{and} \quad \underset{i=1}{\overset{n}{\text{ave}}}\,\rho\left(\frac{x_i - T_n}{S_n}\right) = \beta \tag{28}$$

in which he used ψ_b and ρ_b but other choices of ψ and ρ are also possible. The resulting T_n and S_n are not the same as in Sections 4.1 and 4.2. Huber's proposal (28) is mathematically elegant because it treats T_n and S_n in the same way, and does away with the need for auxiliary estimates of scale or location. Under the same conditions on ψ and ρ as before, the solution (T_n, S_n) to (28) again exists and is unique, and it can be computed by an iterative algorithm that alternates steps like (21) and (27). In spite of this, Huber's proposal is being used less and less nowadays because the resulting T_n and S_n have a lower breakdown value than in Sections 4.1 and 4.2. For instance, S-Plus has deprecated its function for (28) and instead recommends its users to switch to *location.m* which implements (20).

We have found that the lack of robustness of Huber's proposal persists even for very small n. Figs. 7a and b show the (two-dimensional) SEIF's of T_n and S_n for $n=4$ based on (28) using our $\psi_{\log}$ and $\rho_{\log}$. Both are unbounded, indicating that they cannot

withstand a single outlier (although the separate M-estimators can). The same is true for the 3D-SEIF's in Figs. 7c and d for $n=6$, whereas the separate M-estimators were able to resist two outliers (compare with the bounded 3D-SEIF's in Figs. 5a and 6b).

The failure of (28) can be explained as follows. Consider the example in Fig. 7a where by construction $X=(-0.84,0,0.84,x)$ and let x be a large value, say $x=40$. We first compute the initial location estimate $T_4^{(0)}=\mathrm{med}_4(X)=0.42$, which is as far away from zero as it can be for any x. Consequently, also the initial scale estimate $S_4^{(0)}=\mathrm{MAD}_4(X)$ becomes as large as it can be for any x. The next step computes $T_4^{(1)}$ according to (21) where $S_4=S_4^{(0)}$ is relatively large, hence $(x_i-T_4^{(0)})/S_4^{(0)})$ comes closer to the center of ψ where it is more linear, so $T_4^{(1)}$ moves further to the right. In the next step, $S_4^{(1)}$ obtained from (27) is increased as well because $(x_i-T_4^{(1)})/S_4^{(0)}$ comes closer to the center of ρ where it is more quadratic. When computing $T_4^{(2)}$ in the next step this effect repeats itself. So, at each location step the increased $S_4^{(k)}$ causes $T_4^{(k+1)}$ to move further to the right, and at each scale step the fact that $T_4^{(k+1)}$ is further to the right inflates $S_n^{(k+1)}$ in turn. If we let x grow, the final location estimate approaches the average and the final scale estimate tends to the standard deviation.

4.4. Discussion

Apart from the separate M-estimators (Sections 4.1 and 4.2) and Huber's Proposal 2 (Section 4.3) we have also tried many other, 'hybrid' versions. For instance, we have replaced the auxiliary median and MAD by 1-step or fully iterated M-estimators, yielding two-stage procedures, and even three-stage procedures, etc. But in each version the robustness degraded in the same direction as Huber's Proposal 2 (although sometimes to a lesser extent). We thus prefer to use the separate M-estimators, even when we need to estimate location as well as scale.

5. And what if either μ or σ is assumed known?

In some cases one may assume that the true scale σ is known in advance. The typical example is when σ is the inaccuracy of a measurement instrument. Another example is in a repeated measures setup, when scale can be estimated in advance by pooling the variabilities at all sites/for all compounds. Whenever σ is assumed to be known, (1) reduces to the equivariance property $T_n(X+b)=T_n(X)+b$ with the same maximal breakdown value (10). We can now estimate location by an M-estimator if we replace the S_n in (20) and (21) by σ. It turns out that the advantages of $\psi_{\log}$ and full iteration (21) still hold in the sense of attaining the maximal breakdown value and a completely smooth SEIF. Note that we can even apply the logistic M-estimator when $n=3$ (because we do not need to estimate scale), yielding the smooth SEIF in Fig. 8a.

If on the other hand the location θ is known, we can first standardize our data set by substracting θ from each x_i. So, $\theta=0$ without loss of generality. The equivariance property (2) for scale estimators now reduces to $S_n(aX)=|a|S_n(X)$. In this context

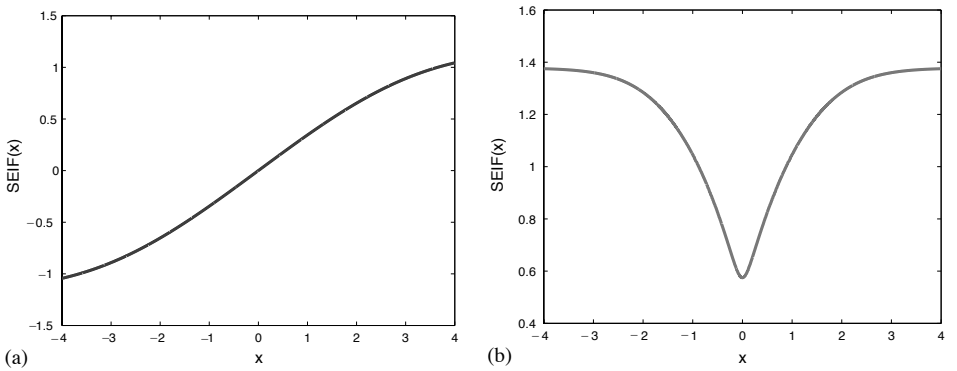

Fig. 8. The SEIF of (a) the logistic M-location with known scale, and (b) the logistic M-scale with known location, both for $n = 3$.

the maximal breakdown value of scale estimators is increased from (19) to (10). For instance, we can now use the *median distance to zero* (MDZ) defined as

$$\mathrm{MDZ}_n(X) = 1.483 \underset{i=1}{\overset{n}{\mathrm{med}}}(|x_i|), \tag{29}$$

which attains the maximal breakdown value. We recommend to use the MDZ only as the initial $S_n^{(0)}$ and to compute a fully iterated M-estimator of scale with the logistic ρ of (26). To compute S_n, iterate (27) without the term T_n. The resulting S_n has the maximal breakdown value and a smooth SEIF as in Fig. 8b, which also illustrates that now scale can be estimated robustly even when $n = 3$.

6. Discussion and recommendations

This paper focuses on the situation, common in the experimental sciences, where there are very few observations and one requires outlier-robust estimators of location and scale. In this context, an experimenter would not accept a method which gives a substantially different result when a measurement is rounded, nor would she/he continue to use it after noticing that moving a measurement to the right can cause the location estimate to move to the left. We have therefore restricted attention to estimators with a smooth and monotone SEIF. By design, this eliminates M-estimators with a non-monotone ψ-function (see, e.g. Section 2.6 of Hampel et al., 1986), as well as the univariate least median of squares, least trimmed squares, and S-estimators (see Chapter 4 of Rousseeuw and Leroy, 1987) whose SEIF is not monotone either. Moreover, it is already known that the latter location estimators need not be unique for small sample sizes (Rousseeuw, 1994), unlike all estimators considered here.

The study reported in this paper leads us to formulate the following conclusions and recommendations. When $n = 1$ or 2, no robustness is possible. When $n = 3$ and both location and scale are unknown we recommend to estimate location by the sample median, whereas the scale cannot be estimated robustly. However, if the context is

such that we only need to protect against scale implosion we can apply the ADM of Section 3, whereas if we only need to protect against explosion we can use the MAD.

When $n \geqslant 4$, we propose to estimate location by the M-estimator (20) with $\psi_{\log}$ using MAD_n as the auxiliary scale. Analogously, we estimate scale by the M-estimator (25) with $\rho_{\log}$ using med_n as the auxiliary location.

When either location or scale is known, we estimate the remaining parameter by the corresponding fully iterated logistic M-estimator, which works for any n and is robust for $n \geqslant 3$.

The MATLAB code for the methods described in this paper can be downloaded from the website http://win-www.uia.ac.be/u/statis.

Acknowledgements

We thank Tina Carrasco for insisting that the 'simple' case of very small n was worth studying because of its importance in everyday experimental science.

References

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., Tukey, J.W., 1972. Robust Estimates of Location: Survey and Advances. Princeton University Press, Princeton, NJ.

Croux, C., Rousseeuw, P.J., 1992. A class of high-breakdown scale estimators based on subranges. Comm. Statist.—Theory Meth. 21, 1935–1951.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust Statistics: The Approach Based on Influence Functions. Wiley, New York.

Huber, P.J., 1981. Robust Statistics. Wiley, New York.

Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

Rousseeuw, P.J., 1994. Unconventional features of positive-breakdown estimators. Statist. Probab. Lett. 19, 417–431.

Rousseeuw, P.J., Croux, C., 1993. Alternatives to the median absolute deviation. J. Amer. Statist. Assoc. 88, 1273–1283.

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