Computational Statistics and Data Analysis 55 (2011) 324–337

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

# A smoothing principle for the Huber and other location M-estimators

Frank Hampel[a], Christian Hennig[b,*], Elvezio Ronchetti[c]

[a] Seminar für Statistik, ETH Zürich, Switzerland
[b] Department of Statistical Science, UCL, Gower Street, WC1E 6BT London, United Kingdom
[c] Department of Econometrics, University of Geneva, Switzerland

ARTICLE INFO

*Article history:*
Received 2 July 2009
Received in revised form 30 April 2010
Accepted 3 May 2010
Available online 10 May 2010

*Keywords:*
Pitman estimator
ML-estimator
Median
MAD
Breakdown point
Small sample asymptotics
Cauchy distribution
Huber's least favourable distribution
Double exponential distribution
Robust estimation

ABSTRACT

A smoothing principle for M-estimators is proposed. The smoothing depends on the sample size so that the resulting smoothed M-estimator coincides with the initial M-estimator when $n \to \infty$. The smoothing principle is motivated by an analysis of the requirements in the proof of the Cramér–Rao bound. The principle can be applied to every M-estimator. A simulation study is carried out where smoothed Huber, ML-, and Bisquare M-estimators are compared with their non-smoothed counterparts and with Pitman estimators on data generated from several distributions with and without estimated scale. This leads to encouraging results for the smoothed estimators, and particularly the smoothed Huber estimator, as they improve upon the initial M-estimators particularly in the tail areas of the distributions of the estimators. The results are backed up by small sample asymptotics.

## 1. Introduction

The parametric estimation of the location of a one-dimensional symmetric distribution is among the easiest and most comprehensively worked on problems in statistics. It is a benchmark to illustrate and investigate new ideas in estimation which may generalize to more complicated situations in order to gain a better understanding about estimation problems in general. Here we focus on small to moderate sample sizes. Small samples are relevant in many applications (Lischer, 1996; Rousseeuw and Verboven, 2002), particularly if, in the context of modelling of complex technical experiments, a few measurements of the same situation are to be summarized to make fitting of a more sophisticated model easier. Other examples include medical studies with budget restrictions or limited numbers of patients.

The present paper is about an idea introduced by Hampel (1996), which he called "Huber without corners" (in Hampel (1996) only the definition is given). The idea of that paper was to smooth the corners of the $\psi$-function defining the well-known Huber M-estimator (Huber, 1964). The smoothing depends on the distribution of the initial estimator for $n$ observations, so that the $\psi$-function is much smoother than that of the initial estimator for small $n$, but asymptotically equivalent. In Hampel (1996) the initial estimator is the Huber M-estimator, but the principle can be applied to any M-estimator. It can even be applied to $\psi$-functions that are already smooth, and it can still lead to improvements.

The aim of this paper is to introduce and motivate the smoothing principle, and to investigate the finite sample properties of some smoothed M-estimators (including the smoothed Huber estimator) in comparison to their initial M-estimators,

[*] Corresponding author.
*E-mail addresses:* hampel@stat.math.ethz.ch (F. Hampel), chrish@stats.ucl.ac.uk (C. Hennig), Elvezio.Ronchetti@unige.ch (E. Ronchetti).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.05.001

but also to ML-estimators and Pitman estimators (Pitman, 1939), which have the minimal variance among all equivariant estimators.

We carried out an extensive simulation study in which smoothed M-estimators (including the smoothed Huber M-estimator and a smoothed median) were compared to the ML-estimators and Pitman estimators for small samples from the normal, the Huber least favourable, the double exponential and the Cauchy distribution with known scale. Other simulated setups concern a more realistic situation with unknown scale, in which the MAD is used as a preliminary estimator of scale. ML-, smoothed ML-estimators and Pitman estimators for the distributions given above are compared with a smoothed Huber M-estimator and a smoothed Bisquare M-estimator (e.g., Maronna et al., 2006) and their initial (non-smoothed) M-estimators. The Bisquare was included as a popular example of an M-estimator with redescending $\psi$-function. It was of interest to see how the Huber and Bisquare M-estimator and their smoothed versions perform without assuming knowledge of the underlying distribution. We did not only compare the MSEs, but we also examined the shapes of the distribution of squared errors (s.e.). This leads to some surprising insights that cannot be explained by asymptotic theory.

Smoothed M-estimators are defined in Section 2. A motivation why the smoothing principle may lead to an improvement for small samples is given in Section 3. Finite sample breakdown points of the smoothed M-estimators compared with the original and other M-estimators are briefly discussed in Section 4. The simulation study is described in Section 6 and the results are given and discussed in Section 7. The smoothing principle leads to good results. Particularly the smoothed Huber estimator exhibits excellent small sample properties in a reasonable range of situations. The results are supplemented and confirmed in Section 8 by computations of tail probabilities of the estimator's distributions based on small sample asymptotics (Hampel, 1973; Field and Ronchetti, 1990) as explained in Section 5. A conclusion is given in Section 9.

## 2. Smoothing the Huber and other M-estimators

Consider an i.i.d. sample of $n$ observations from a distribution $P_{\mu,\sigma}$ with unimodal symmetric density

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right).$$

To simplify the setup even more, we first assume $\sigma$ to be known and we set $\sigma = 1$.

An M-estimator for the location parameter $\mu$ is defined as a solution $t$ of

$$\sum_{i=1}^{n} \rho\left(\frac{x_i - t}{\sigma}\right) = \min_t \quad \text{or} \quad \sum_{i=1}^{n} \psi\left(\frac{x_i - t}{\sigma}\right) = 0, \tag{1}$$

where $\psi = \rho'$. The maximum likelihood (ML)-estimator is defined by $\rho_f = -\log f$ or $\psi_f = -f'/f$. For a given positive constant $k$, the Huber estimator is defined by the following function $\psi$ in (1):

$$\psi_k(x) = \begin{cases} k: & x > k \\ x: & -k \le x \le k \\ -k: & x < -k. \end{cases}$$

It is the ML-estimator for the location parameter of Huber's least favourable distribution with density

$$f_k(x) = \begin{cases} (1-\epsilon)\varphi(k)\exp(-k(x-k)): & x > k \\ (1-\epsilon)\varphi(x): & |x| \le k \\ (1-\epsilon)\varphi(k)\exp(k(x+k)): & x < -k, \end{cases}$$

where $\epsilon$ and $k$ are linked by $\frac{2\varphi(k)}{k} - 2\Phi(-k) = \frac{\epsilon}{1-\epsilon}$, and $\varphi$ and $\Phi$ denote the pdf and cdf of the standard normal distribution. Huber (1964) showed that the above distribution has smallest Fisher information among the symmetric distributions of the form $(1-\epsilon)\varphi(x) + \epsilon h(x)$, $h$ being a symmetrical density, and that the Huber estimator has minimax asymptotic variance for this class of distributions. In our simulations we used $k = 0.862$, which corresponds to $\epsilon = 0.2$.

The smoothed Huber estimator introduced by Hampel (1996) is a smoothed version of the Huber estimator, where the degree of smoothness depends on the sample size so that the new estimator keeps the asymptotic optimality properties of the Huber estimator whereas performing better on small sample sizes. We formulate the principle for a general $\psi$-function of an M-estimator. Its smoothed version is defined by the score function

$$\tilde{\psi}(x) = \int \psi(x+u)\mathrm{d}Q_n(u), \tag{2}$$

where $Q_n$ may be chosen as the distribution of the initial M-estimator for $n$ i.i.d. observations from an assumed underlying distribution. The natural choice for an ML-estimator would be the corresponding distribution under which it is asymptotically optimal. The exact distribution of the ML-estimator under this distribution is usually difficult to obtain. Therefore, as a default choice, we approximate it by the normal distribution with expectation 0 and variance $V/n$, where $V$ is the asymptotic variance. This principle can generally be used for M-estimators, for which it is needed to specify a distribution under which the estimator's asymptotic variance is computed. Note that, strictly speaking, the term "smoothed" is often not justified because the principle can also be applied to $\psi$-functions that are already smooth.

The $\psi$-function of the smoothed Huber estimator defined by $\psi = \psi_k$ and $Q_n$ obtained from the asymptotic normality of the Huber estimator under Huber's least favourable distribution can be easily written in closed form as

$$\begin{aligned}\tilde{\psi}_k(x) &= k\Phi\left(\frac{x-k}{\sigma_n}\right) - k\left(1-\Phi\left(\frac{x+k}{\sigma_n}\right)\right)\\ &\quad + x\left(\Phi\left(\frac{x+k}{\sigma_n}\right) - \Phi\left(\frac{x-k}{\sigma_n}\right)\right) + \sigma_n\left(\varphi\left(\frac{x+k}{\sigma_n}\right) - \varphi\left(\frac{x-k}{\sigma_n}\right)\right),\end{aligned}$$

where $\sigma_n = \sqrt{V/n}$, which equals $\sqrt{2.046/n}$ for $\epsilon = 0.2$. Since $Q_n$ tends to a Dirac measure at 0 for $n \to \infty$, the smoothed Huber is asymptotically equivalent to the Huber estimator.

## 3. Motivation of the smoothing principle

We give a motivation that leaves some theoretical gaps and is therefore heuristic. Most of this section focuses on ML-estimators.

Let $T_n$ be a consistent and unbiased estimator for $\mu$, $\mathbf{X} = (X_1, \dots, X_n)'$ be an $\mathbb{R}^n$-valued random variable, $\mathbf{x} = (x_1, \dots, x_n)' \in \mathbb{R}^n, f_\mu(\mathbf{x}) = \prod_{i=1}^n f_{\mu,1}(x_i)$, $\psi_f(x) = -\frac{f'(x)}{f(x)}$ (the score function of the ML-estimator under $f$), and

$$\psi_{f,n}(\mu, \mathbf{x}) = -\frac{\partial}{\partial\mu}\log f_\mu(\mathbf{x}) = \sum_{i=1}^n \psi_f(x_i - \mu).$$

Schwarz's inequality yields

$$\mathrm{Var}_\mu(T_n) \geq \frac{\mathrm{Cov}_\mu(T_n, \psi_{f,n}(\mu, \mathbf{X}))^2}{\mathrm{Var}_\mu(\psi_{f,n}(\mu, \mathbf{X}))}.$$

Assuming that all necessary derivatives exist and behave regularly and under the conditions of the Cramér–Rao inequality, the covariance is 1 independently of $T_n$. Therefore, $\mathrm{Var}_\mu(T_n)$ can be minimized if $T_n$ can be chosen so that $\mathrm{Corr}_\mu(T_n, \psi_{f,n}(\mu, \mathbf{X}))$ is maximized.

For an M-estimator $T_n$ with symmetric score function $\psi$ we obtain by a standard Taylor expansion about $\mu$

$$T_n = \mu - \frac{\frac{1}{n}\sum\limits_{i=1}^n \psi(x_i - \mu)}{\frac{1}{n}\sum\limits_{i=1}^n \psi'(x_i - \xi)}, \tag{3}$$

where $\xi$ lies between $T_n$ and $\mu$. In particular, $\psi = \psi_f$ defines the ML-estimator. When $n \to \infty$, $\frac{1}{n}\sum_{i=1}^n \psi_f'(x_i - \xi)$ converges to $\mathrm{E}[\psi_f'(X_1)] > 0$, where by equivariance the expectation can be taken under $\mu = 0$. Therefore, the variations in $\left|\frac{1}{n}\sum_{i=1}^n \psi_f'(x_i - \xi)\right|$ become negligible compared to the variations in $\left|\frac{1}{n}\sum_{i=1}^n \psi_f(x_i - \mu)\right|$, which converges to 0. Thus, $T_n$ is asymptotically a linear function in $\psi_{f,n}(\mu, \mathbf{X})$ and the ML-estimator reaches the Cramér–Rao bound. For general (non-ML) M-estimators it follows in the same way that $T_n$ is asymptotically a linear function in $\sum_{i=1}^n \psi(x_i - \mu)$ under regularity conditions ensuring consistency, though this only yields minimum asymptotic variance among those estimators for which

$$\mathrm{Cov}_\mu\left(T_n, \sum_{i=1}^n \psi(x_i - \mu)\right) = \text{const.}$$

This argument requires $T_n - \mu \approx 0$ and therefore works only asymptotically. For fixed (small) $n$, the expansion (3) is not very useful when $T_n$ is at some distance from $\mu$. In the spirit of small sample asymptotics (Hampel, 1973; Field and Ronchetti, 1990) we may recenter the expansion at $u \approx T_n - \mu$ not close to 0 and obtain:

$$T_n = u + \mu - \frac{\frac{1}{n}\sum\limits_{i=1}^n \psi(x_i - u - \mu)}{\frac{1}{n}\sum\limits_{i=1}^n \psi'(x_i - \nu)}, \tag{4}$$

where now $\nu$ lies between $T_n$ and $u + \mu$, so $T_n$ may be far away from $\mu$. If we choose $\psi(x) = \psi_f(x + u)$ in this situation and consider again the denominator to be more or less constant compared to the absolute value of the numerator (which is reasonable taking into account that $T_n \approx u + \mu$ is a location estimator for the $x_i$), we have again that $T_n$ is approximately linear in $\psi_{f,n}(\mu, \mathbf{X})$. In this way, $\psi_f(x+u)$ can be interpreted as an approximatively optimal $\psi$-function for an M-estimator $T_n$, given $T_n - \mu$ is in a neighborhood of $u$.

Knowing $u$ would imply knowing $T_n - \mu$, which is impossible. Therefore, the corresponding score function $\psi_f(x + u)$ is not available. However, if the distribution of $T_n - \mu$ were available, we could average with respect to this distribution by taking the expectation and we could choose $\tilde{\psi}(x) = \mathrm{E}\psi_f(x+U)$. This leads to the smoothing principle (2) for ML-estimators (for general M-estimators, the argument requires $\mathrm{Cov}_\mu(T_n, \sum_{i=1}^n \psi(x_i-\mu)) = \text{const}$, see above). The argument can be interpreted as leading to a fixed point iteration, because $Q_n(u)$ should ideally be the distribution of the estimator $\tilde{T}_n$ corresponding to its score function $\tilde{\psi}$. The resulting $\tilde{T}_n$ will be asymptotically equivalent to the ML-estimator. Hence, if we derive $Q_n$ from an asymptotic normal approximation, there is no difference between $Q_n$ for the ML-estimator and for the smoothed ML-estimator.

Note, however, that this argument is incomplete because it is not entirely clear why the good correlation properties of some $T_n^*$ defined by $\psi_f(x+u)$ with $\psi_{f,n}(\mu, \mathbf{X})$ for $T_n - \mu$ in a neighborhood of $u$ imply a large overall correlation between $\tilde{T}_n$ defined by $\mathrm{E}\psi_f(x+U)$ and $\psi_{f,n}(\mu, \mathbf{X})$. This claim is not proved, but it seems to be confirmed by our simulation results.

Apart from the motivation of the precise form of $\tilde{\psi}$, it can be seen as an aim in itself to smooth the $\psi$-function of a general "not smooth enough" M-estimator depending on $n$ keeping its asymptotic properties. For example, as opposed to the median, the smoothed median has a finite local-shift sensitivity (Hampel et al., 1986, p. 88), and also this will improve the change-of-variance sensitivity of the estimator (Hampel et al., 1986, p. 130).

## 4. Finite sample breakdown points

The finite sample breakdown point of an estimator measures the minimum proportion of points that have to be added (or changed; there are different definitions of the finite sample breakdown point, see Donoho and Huber (1983)) to a dataset so that an estimator can be driven infinitely far away from its value for the original dataset.

There are well-known results for M-estimators of location under some conditions on $\psi$. For bounded, monotone and symmetric $\psi$-functions, the finite sample breakdown point is $\frac{1}{n}\lfloor\frac{n-1}{2}\rfloor$ (Huber, 1981). For redescending M-estimators, the situation is more complicated and depends on the dataset. If a preliminary scale estimator such as the MAD is introduced, the breakdown point cannot be larger than that of the scale estimator (note that in case of implosion of the scale estimator to zero, plugging it in for $\sigma$ in (1) does not yield a well defined estimator), but for bounded, monotone and symmetric $\psi$-functions, and MAD scale it is still $\frac{1}{n}\lfloor\frac{n-1}{2}\rfloor$. For redescending M-estimators it can be the same but this depends on the dataset, see Chen and Tyler (2004).

The smoothing principle only affects the $\psi$-function, and only in such a way that the conditions for the results cited above still hold for the smoothed estimators if they hold for the initial ones. Therefore, the smoothing principle does not introduce additional problems with the finite sample breakdown point.

## 5. Accurate small sample approximations of tail areas

Small sample asymptotic techniques provide very accurate approximations of densities and tail probabilities down to very small sample sizes. In Section 8 we will use these approximations to supplement the Monte Carlo simulations to evaluate the performance of several estimators including the smoothed Huber M-estimator under several distributions $F$ of the observations. A measure of quality will be the tail probability $\bar{F}_n(t) = P_F[T_n > t]$ of the estimators for different values of $t$ and different sample sizes $n$.

Let $f_n(t)$ be the density of an M-estimator of location $T_n$ defined by (1). By expanding the logarithmic derivative $f_n'(t)/f_n(t)$ locally around each point $t$ separately, Field and Hampel (1982) derived a very accurate approximation for this quantity and, by integration, for the density $f_n(t)$. To obtain tail probabilities $\bar{F}_n(t) = \int_t^\infty f_n(s)\mathrm{d}s$ for the estimator $T_n$, we would need a numerical integration. However, it turns out that this can be approximated analytically to get the following tail area approximation:

$$\begin{aligned}\bar{F}_n(t) &= P_F[T_n > t]\\ &\approx 1 - \Phi\left(\sqrt{2n\,\log C(t)}\right) + \frac{C(t)^{-n}}{\sqrt{2\pi n}}\left(\frac{1}{\sigma(t)\alpha(t)} - \frac{1}{\sqrt{2\,\log C(t)}}\right),\end{aligned} \tag{5}$$

where $t > \mu$ and

$$C(t) = \left(\int \mathrm{e}^{\alpha(t)\psi(x-t)}\mathrm{d}F(x)\right)^{-1},$$

$$\alpha(t) \text{ solves } \int \psi(x-t)\,\mathrm{e}^{\alpha(t)\psi(x-t)}\mathrm{d}F(x) = 0,$$

$$\sigma^2(t) = C(t)\int \psi^2(x-t)\,\mathrm{e}^{\alpha(t)\psi(x-t)}\mathrm{d}F(x);$$

see Lugannani and Rice (1980) in the case of the arithmetic mean ($\psi(x) = x$), and Daniels (1983) and Field and Ronchetti (1990) for M-estimators.

## 6. Simulation design

We simulated data with $n = 3, 4, 5, 8, 20$ from four distributions: the normal distribution, Huber's least favourable distribution with $k = 0.862$ ("Huber distribution" in the following), the double exponential distribution (in some literature referred to as Laplace distribution), for which the median is the ML-estimator, and the Cauchy distribution. We run 100,000 simulations for each setup.

### 6.1. Simulations with scale assumed to be known

The simulations with the scale assumed to be known are mainly of theoretical interest in order to show how the smoothed estimators compare to their initial non-smoothed M-estimators, the mean, the median and the optimal estimators for the simulated distributions.

The following estimators were computed: mean, median, ML-estimator, smoothed ML-estimator (with $Q_n$ chosen as the asymptotic normal approximation, see Section 2) and the Pitman estimator. For the Huber, Cauchy and double exponential distribution, the ML, Pitman and smoothed ML-estimator were computed with respect to the correct underlying distribution (including the variance of $Q_n$). The Huber and smoothed Huber M-estimator (with the same underlying variance of $Q_n$) were computed for all setups. Under the normal distribution, we also included the ML-estimators and smoothed ML-estimators with respect to the three other distributions.

For every setup and every estimator, we consider the following statistics of the distribution of the squared errors: mean (which is almost equal to the variance of all estimators because of the unbiasedness and the large number of simulation runs), median, first and third quartile, 0.9-, 0.95- and 0.99-quantiles. We computed other measures than just the MSE because it is doubtful that the latter is the most reasonable quantity to compare estimators. Indeed, especially in situations where more than half of the data may be outlying with respect to the correct parameter (Cauchy, small $n$), it is more relevant to know that an estimator has very often a small or moderate squared error (measured by the 0.9- or 0.99-quantile, say) than how bad it is exactly in the situations where it is determined by 50% or more outliers (which may dominate the MSE of the estimator). It also turns out that the estimators differ considerably with respect to the shape of their squared error distribution, so different estimators are optimal with respect to different criteria.

We estimated the standard deviation for all these measures. The results of two estimators were judged as "clearly" different if the intervals obtained by adding or subtracting twice the estimated standard deviation did not intersect. Note, however, that this rule does not lead to a proper significance test, because in the study all estimators were computed for the same samples, and therefore the results for the various estimators were dependent.

We also carried out paired $t$- and Wilcoxon tests in some situations between results of pairs of estimators (results not shown; we were particularly interested in comparing estimators with rather similar results) and found that 100,000 simulations are enough to make almost all differences between estimators highly significant, even between those that look almost equal.

### 6.2. Simulations with unknown scale

Most M-estimators for location (though not the mean and the median) depend on the scale $\sigma$ (see Section 2). In reality, this is not known. One method to deal with this is to estimate a highly robust scale estimator first (the median absolute deviation from the median MAD is the most popular choice) and plug the estimated value of $\sigma$ into (1).

We carried out four simulations in which, for the Pitman estimator and those M-estimators that depend on the scale, this principle was applied with the MAD multiplied by 1.4826 scaled for consistency for $\sigma$ at the normal distribution.

We restricted ourselves to a single value of $n$ for each distribution, namely $n = 4$ for the normal distribution, $n = 5$ for the Huber distribution, $n = 8$ for the Cauchy distribution and $n = 20$ for the double exponential distribution.

Apart from the mean and median, the Pitman, ML and smoothed ML-estimator for the simulated distribution, we included the Huber and smoothed Huber estimator with tuning constant and variance of $Q_n$ derived from the Huber distribution with $k = 0.862$ and the Bisquare M-estimator (Maronna et al., 2006) with redescending $\psi$-function, tuned to 95% efficiency under the normal distribution, and a "smoothed Bisquare", i.e., the smoothing principle applied to the Bisquare M-estimator (with the variance of $Q_n$ derived from its asymptotics under the normal distribution) for all distributions. Note that the same tuning for the latter four estimators was applied for all distributions, so that these are here used in a universal fashion that does not require the knowledge of the distribution.

### 6.3. Computational aspects

The necessary numerical integrations were carried out by means of the function `integrate` of the statistics freeware `R` with default settings. Note that the Pitman estimators based on numerical integration are compared with exact Pitman estimators under the Cauchy distribution in Cohen Freue (2007), and are very similar.

The standard deviation for the $q$-quantiles of the squared error distribution was estimated by means of the formula $q * (1 - q)/(100{,}000 * \hat{h}(v_q))$, where $v_q$ denotes the $q$-quantile and $h$ denotes the density of the distribution of the $q$-quantile. The latter was estimated by a kernel density estimator computed with the `R`-function `density` using the default

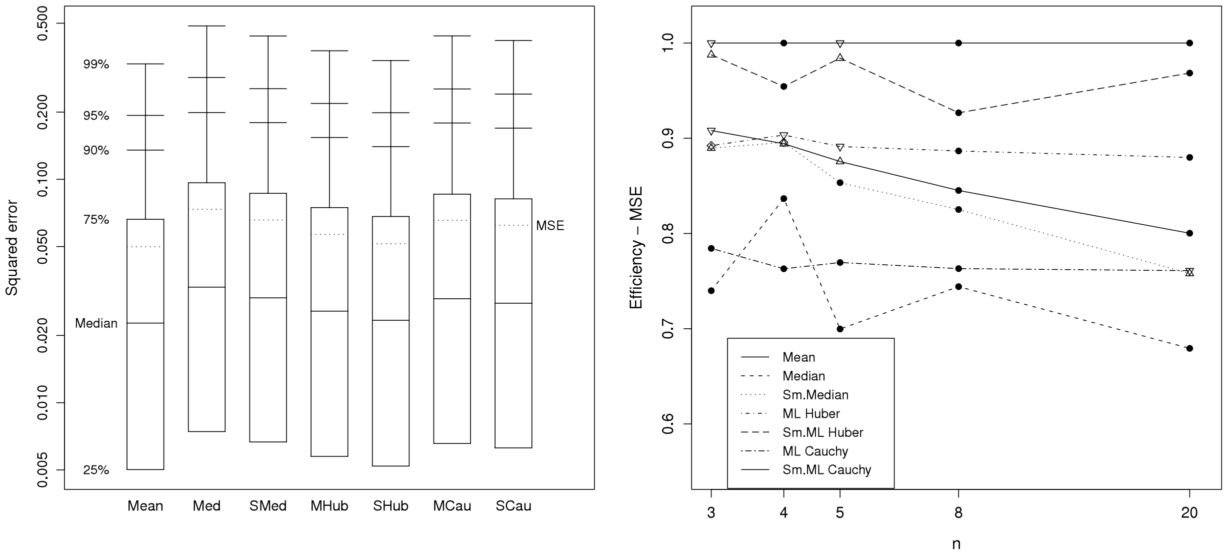

**Fig. 1.** Quantiles and MSE for $n = 20$ (left) and efficiencies of the MSEs (right) for the normal distribution with known scale.

settings. M-estimators, except for the mean and the median, were iterated by the algorithm of the function `huber` of the `MASS` package of `R` (Newton algorithm). The median was used as starting value for all estimators. This led to good results except in some cases for the Cauchy estimators. Thus, for the ML and smoothed ML-estimator for the Cauchy distribution, we started first from the median, then from the two neighboring order statistics, then from two further neighboring order statistics and so on, until an iteration result was found with a larger value of the log-likelihood function than that of the median. In each step, we started from two order statistics and the better result was chosen to guarantee the symmetry of the distribution of the estimator.

## 7. Simulation results

### *7.1. Presentation of the results*

The simulation results are presented graphically. There are two types of plots, namely quantile plots and efficiency plots. A quantile plot shows all the quantiles and the mean of the s.e. distributions of all estimators for a single setup defined by $n$ and the underlying distribution. An efficiency plot shows the relative efficiencies compared to the best estimator with respect to a single statistic (three statistics are used: MSE, median s.e. and 0.99-quantile of the s.e. distribution) for a single underlying distribution for all $n$. The relative efficiencies were computed for all statistics as it is usually done for the MSE. The second type of plot shows whether the estimators differ “clearly” as defined in Section 6: a full circle indicates that the interval of the statistic ± twice the estimated standard deviation of the simulated value is disjunct from the corresponding intervals of all other estimators (same $n$). A triangle with the peak pointing up indicates that the interval intersects with the interval of an estimator with a higher efficiency value. A triangle with the peak pointing down indicates that the interval intersects with the interval of an estimator with a lower efficiency value. A diamond indicates that both types of intersection occur. Efficiency plots include only the estimators that were simulated for all $n$.

In the quantile plots, “M” indicates the ML-estimator corresponding to the simulated distribution, unless indicated explicitly (e.g., “MHub” is the ML-estimator for Huber’s distribution). “S” stands for “smoothed ML”, “P” for the Pitman estimator (e.g., “PCau” for the Cauchy and “PDE” for the double exponential distribution; note, however, that median and smoothed median are denoted by “Med” and “SMed”).

For the Cauchy distribution, the mean is not included in any plot because it is so much worse than the other estimators that its inclusion would have resulted in an unfavourable plot range. The number of observations $n$ and the squared errors are plotted on a logarithmic scale.

To save some space, only one out of the five quantile plots is shown for every distribution for the simulations with known scale. They were chosen so that all typical features can be seen (which often differ between even and odd $n$, as can be seen from the efficiency plot), namely $n = 20$ for the normal distribution, $n = 8$ for Huber’s distribution, $n = 3$ for the double exponential and $n = 5$ for the Cauchy distribution.

### *7.2. Results with scale assumed to be known*

- For the normal distribution (Figs. 1 and 2), the ML-estimators corresponding to the three non-normal distributions are with respect to all statistics and for all $n$ clearly worse than their smoothed counterparts. The smoothed Huber estimator

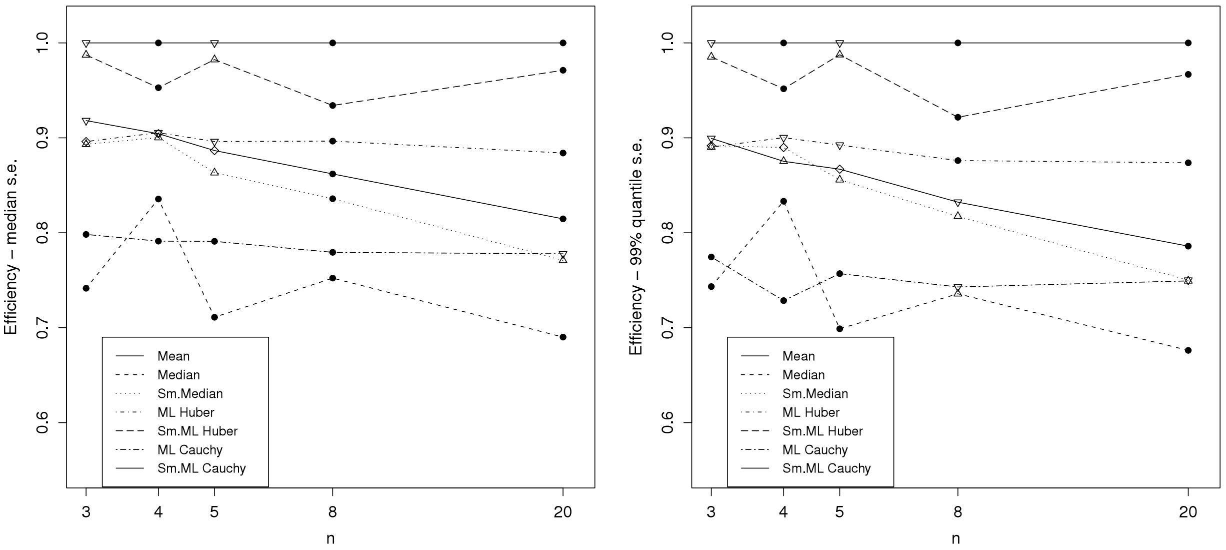

**Fig. 2.** Efficiencies of the median squared errors (left) and of 0.99-quantiles of the squared errors (right) for the normal distribution with known scale.

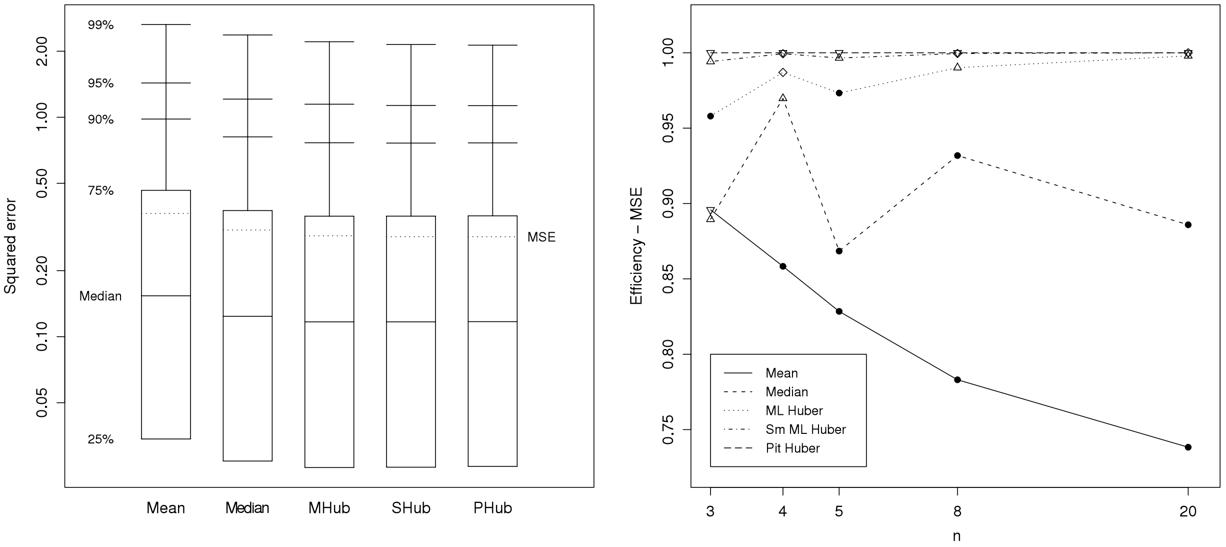

**Fig. 3.** Quantiles and MSE for $n = 8$ (left) and efficiencies of the MSEs (right) for Huber's least favourable ($k = 0.862$) distribution with known scale.

is almost as good as the mean, reaching an efficiency of about 99% for all statistics with $n = 3$ and $n = 5$, where the next best estimators are about 90% efficient. As in many other setups, there seems to be a strong effect of the sample size being odd or even, and for $n = 8$, the efficiency of the smoothed Huber (all statistics) is only about 93%. However, it is still the second best estimator.

- For the three non-normal distributions, the shape of the s.e. distribution for the ML-estimator differs from that of the Pitman estimator and also from that of the smoothed ML-estimator (which can be seen as a compromise of the former two; Figs. 3–8). Whereas the Pitman estimator is optimal with respect to the MSE, the ML-estimator is better with respect to the lower quantiles (0.25 and median, though often not clearly). This corresponds nicely to the discussion in Section 3: the ML-estimator correlates very well with $\psi_{f,n}(\mu, \mathbf{X})$ and this is what a good estimator should do if it is close to $\mu$. In the tails of the distribution, it becomes worse, and this is the area where the Pitman estimator performs better. The smoothed ML-estimator is defined as a direct compromise between "good linearity" in $\psi_{f,n}(\mu, \mathbf{X})$ in the center and the tail areas, and this yields worse lower quantiles and better higher quantiles compared to the ML-estimator, as desired.

  There are even some quantiles, for which the smoothed ML-estimator is better than both the Pitman and the ML-estimator. This happens for the Huber distribution with the 0.75-quantile ($n \leq 8$, left panel of Fig. 3), the 0.9-quantile ($n \leq 8$) and the 0.95-quantile ($n = 4$), for the double exponential distribution with the 0.75-quantile (all $n$, left panel of Fig. 5) and the 0.9-quantile ($n = 4, 8$), and for the Cauchy distribution with the 0.75-quantile ($n \leq 8$) and the 0.9-quantile (all $n$, left panel of Fig. 7), though differences are not very clear in most cases.

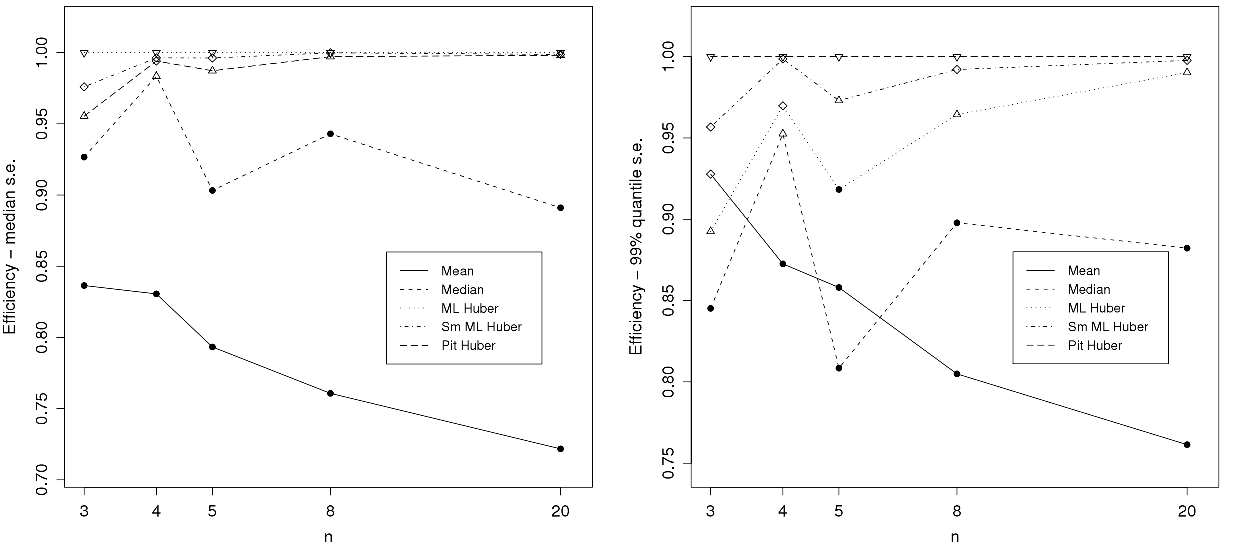

**Fig. 4.** Efficiencies of the median squared errors (left) and of the 0.99-quantiles of the squared errors (right) for Huber's least favourable ($k = 0.862$) distribution with known scale.

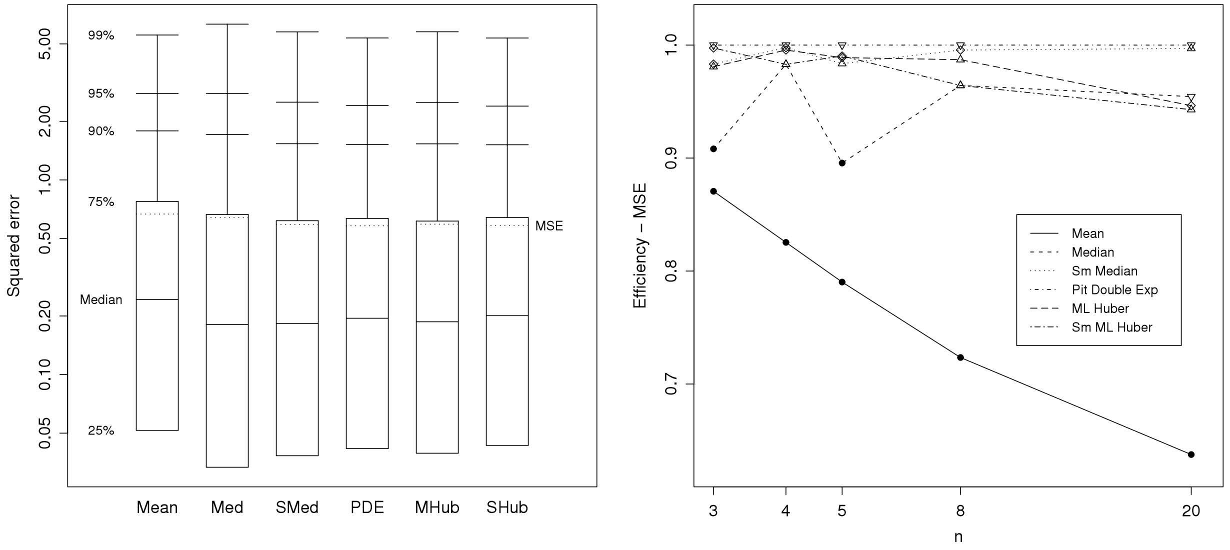

**Fig. 5.** Quantiles and MSE for $n = 3$ (left) and efficiencies of the MSEs (right) for the double exponential distribution with known scale.

The smoothed ML-estimator is always better than the ML-estimator with respect to the MSE. The differences are sometimes significant, and the efficiency gain is up to 10% (double exponential distribution, $n = 3, 5$, right side of Fig. 5). For the Huber distribution, the smoothed ML-estimator is almost indistinguishable from the Pitman estimator (right panel of Fig. 3), whereas the ML-estimator is up to 4% worse. For the Cauchy distribution, the difference between smoothed ML and ML is again up to 4% ($n = 8$, right panel of Fig. 7), but usually small.

- The results with respect to the 0.99-quantile are similar, except that the differences between the estimators are a bit larger and the smoothed ML-estimator is almost 10% better than the ML for the Cauchy distribution for $n = 8$ (right panel of Fig. 8).
- The comparison between the Huber and the smoothed Huber estimator for the double exponential and Cauchy distribution is more ambiguous. The smoothed Huber estimator performs worse with respect to the median s.e. (left panels of Figs. 6, 8). With respect to the MSE and the 0.99-quantile, both estimators perform very similarly for the Cauchy distribution (right panels of Figs. 7, 8). For the double exponential distribution, the behaviour depends strongly on $n$, with the Huber estimator being optimal among all estimators with respect to the 0.99-quantile and $n = 20$, and better than the smoothed Huber estimator with respect to the MSE for $n = 4, 8, 20$, whereas the smoothed Huber estimator is superior with respect to the 0.99-quantile for most $n$, where it even dominates the median, smoothed median and the Pitman estimator (right panels of Figs. 5, 6).

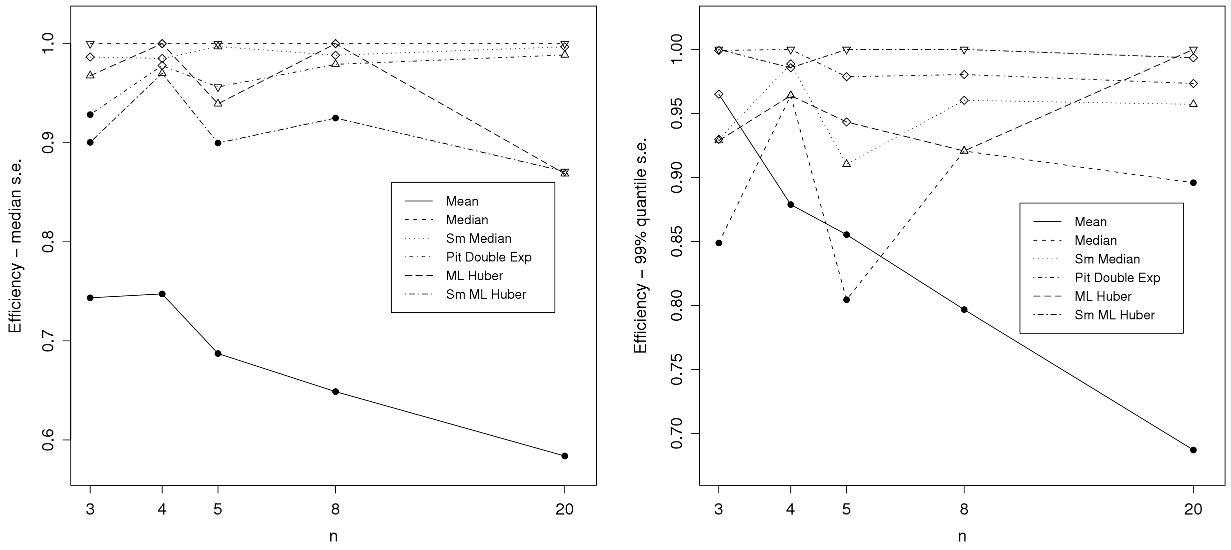

**Fig. 6.** Efficiencies of the median squared errors (left) and of the 0.99-quantiles of the squared errors (right) for the double exponential distribution with known scale.

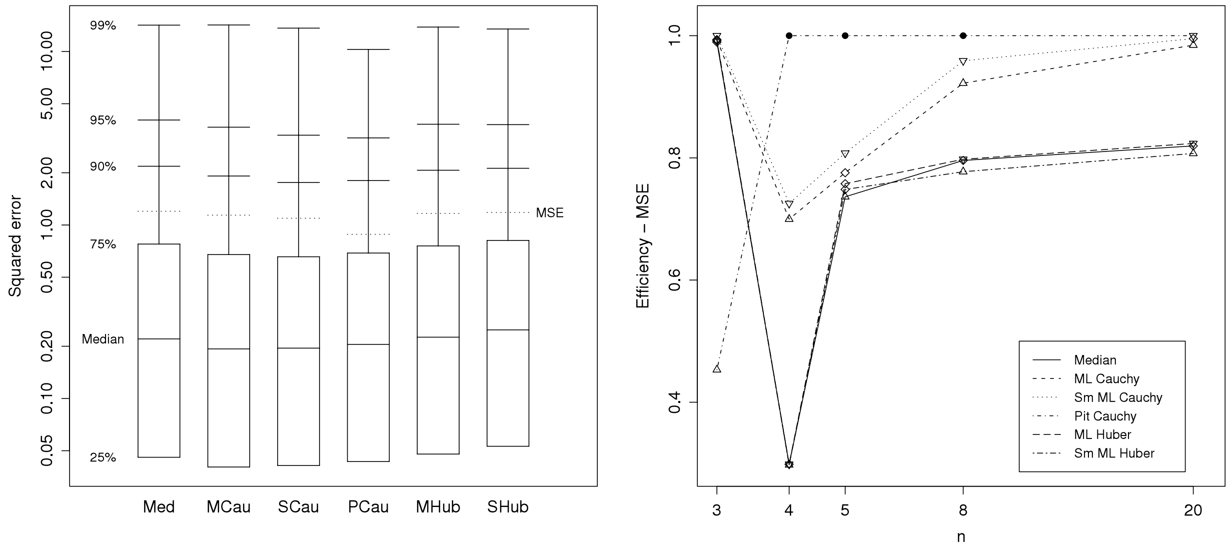

**Fig. 7.** Quantiles and MSE for $n = 5$ (left) and efficiencies of the MSEs (right) for the Cauchy distribution with known scale.

- For the Cauchy distribution, $n = 3$, the Pitman estimator has by far the largest MSE (apart from the mean; right panel of Fig. 7). However, the result can be explained by a few simulated data configurations with two very large outliers. This behaviour highlights that comparing estimators by the MSE suffers in principle from the same robustness problems that affect the mean as an estimator of location. However, measuring the quality of estimators is somewhat different from the estimation problem itself, because an adequate quality measure should not ignore or downweight the tails of the s.e. distribution in the same way. A good estimator should be reliable for 90% or more datasets, and therefore it is reasonable for a quality measure to have a breakdown point smaller than 10% (but not 0). In fact, the MSE results for the Cauchy distribution motivated us to include the quantiles of the s.e. distribution in the simulations, and at least for small $n$, the 0.90-, 0.95- and 0.99-quantiles seem to be much more reasonable as quality measures than the MSE.
- For similar reasons, the mean outperforms the median for the double exponential and Huber distribution, $n = 3$, with respect to the 0.99-quantile and, for the Huber distribution, with respect to the MSE (right panels of Figs. 3–5). If two of three observations are far away in the same direction from the true location, the mean weights the single good observation by $\frac{1}{3}$, which is better than choosing one of the outliers as the median does.
- The effect of the sample size parity (even or odd) is surprisingly large and differs between the setups. In most situations, the median, ML and smoothed ML-estimators are better for even $n$ (Figs. 1–6). For the Cauchy distribution, the opposite

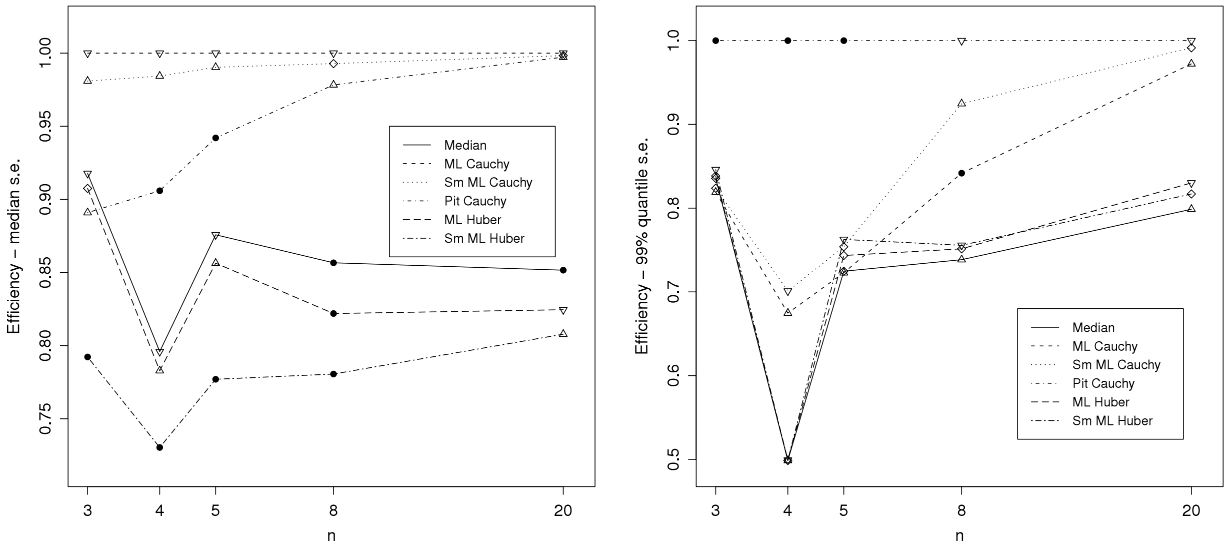

**Fig. 8.** Efficiencies of the median squared errors (left) and of the 0.99-quantiles of the squared errors (right) for the Cauchy distribution with known scale.

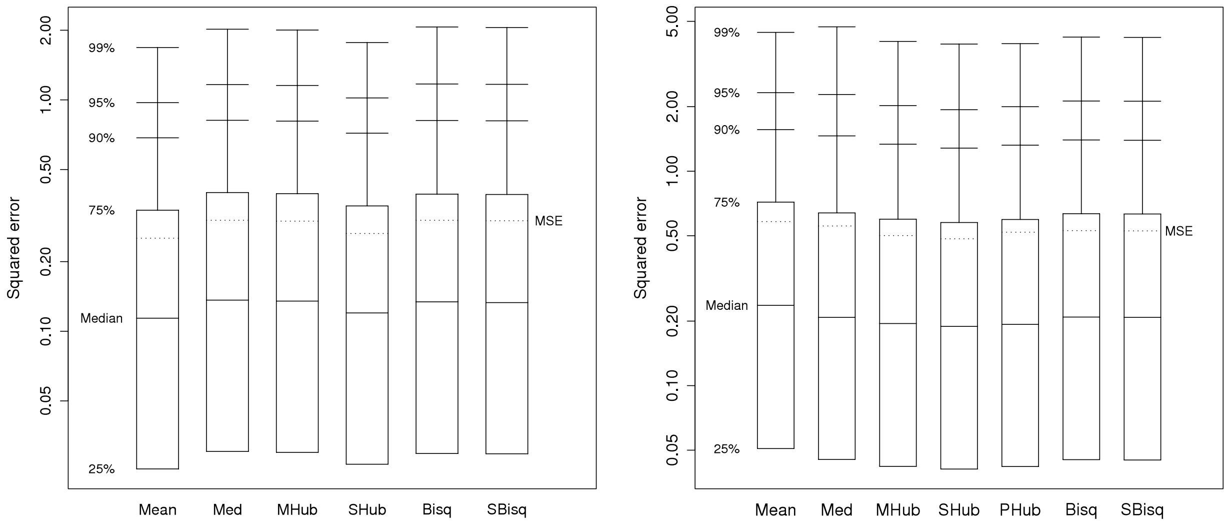

**Fig. 9.** Quantiles and MSE for the normal distribution, $n = 4$ (left) and Huber's least favourable ($k = 0.862$) distribution, $n = 5$ (right) with scale estimated by the MAD.

seems to be true (Figs. 7, 8), and for the normal distribution (Figs. 1, 2), some smoothed ML-estimators are worse for even $n$, where especially the median is relatively good.

### 7.3. Results with scale assumed to be unknown

The simulations with estimation of $\sigma$ by the MAD (Figs. 9, 10) confirm the good properties of the smoothing principle. For the normal distribution, the smoothed Huber estimator clearly remains the second best estimator (about 95% efficient with respect to the MSE, compared to, e.g., 84% of the Huber estimator). For the Cauchy and double exponential distribution, the smoothed ML-estimators are better than the Pitman estimator even for the MSE and for all quantiles. The Pitman estimator seems to be more sensitive against a misspecification of the distribution of the scale. The smoothed Huber estimator outperforms the median and the Huber estimator at least with respect to the MSE and the higher quantiles clearly (and sometimes with respect to all quantiles) and is usually almost as good as the Pitman estimator.

The Bisquare M-estimator and its "smoothed" version deliver always very similar results, but paired Wilcoxon tests reveal that the smoothed Bisquare is always significantly better. Both, however, are clearly outperformed by the smoothed Huber estimator.

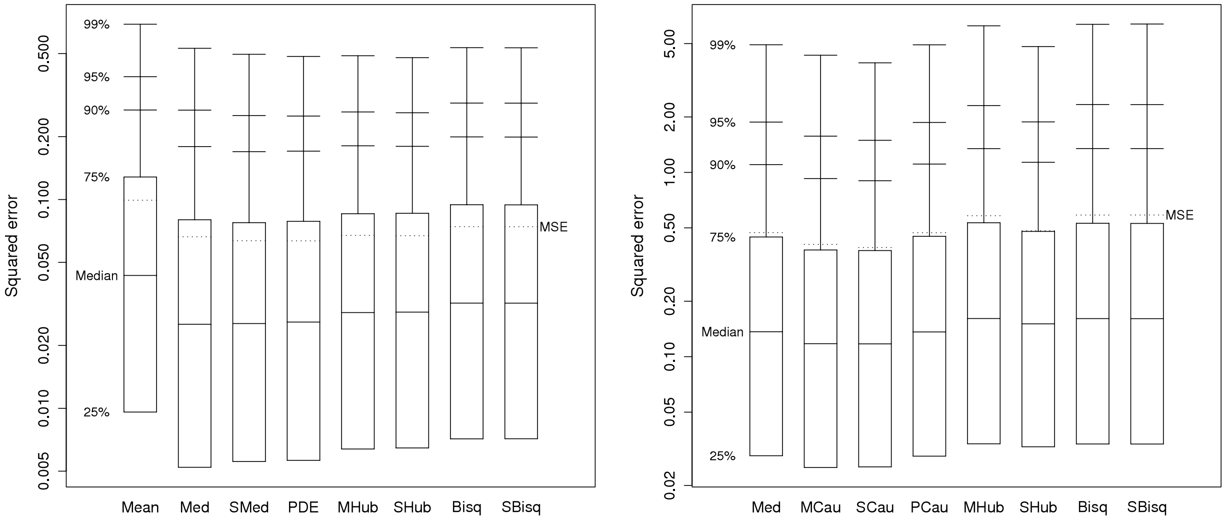

**Fig. 10.** Quantiles and MSE for the double exponential distribution, $n = 20$ (left) and the Cauchy distribution, $n = 8$ (right) with scale estimated by the MAD.

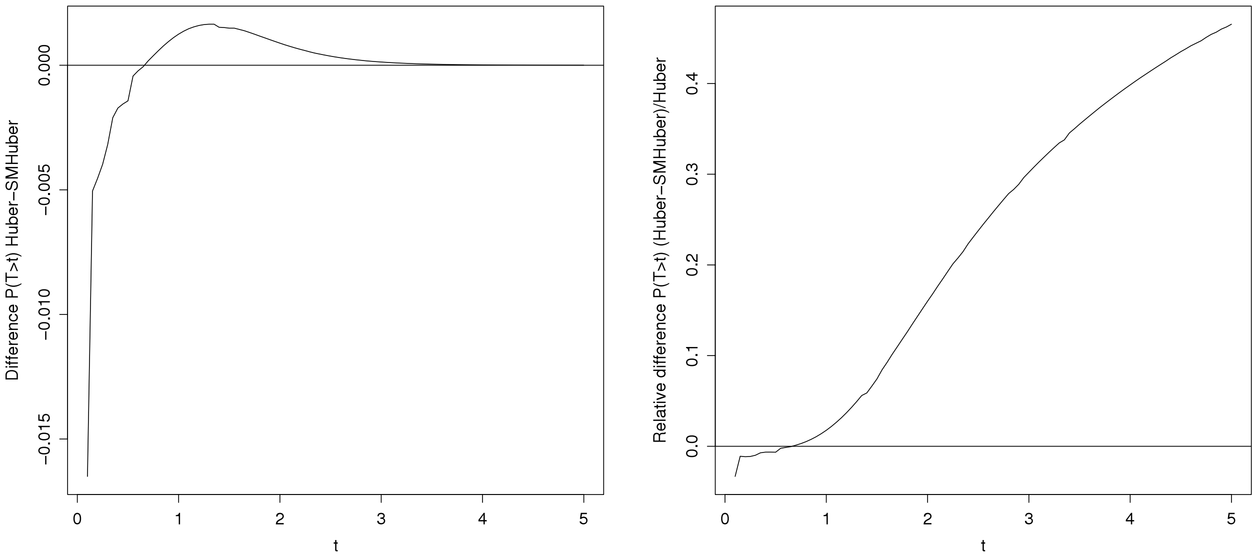

**Fig. 11.** Left: difference in tail probabilities, approximated by (5), between the Huber and the smoothed Huber estimator under Huber's least favourable distribution ($k = 0.862,\ n = 5$; values larger than 0 mean that the smoothed Huber estimator is better). Right: relative difference (difference divided by the tail probability of the Huber estimator).

## 8. Small sample asymptotic results

Formula (5) can be used to back up the simulation results, which it generally did in the cases in which it was applied. The only remarkable feature of the simulation results that is not appropriately reflected in (5) is the difference between even and odd $n$ for some setups.

On the other hand, (5) is more informative about the behaviour in the extreme tail areas, which cannot be simulated accurately, and it can add credibility to some of the non-significant results of the simulation.

Again, selected results are presented in a graphical way. All plots refer to $n = 5$ (for other values of $n$, very similar patterns are obtained). The graphs compare the tail probabilities $P(T_n > t)$ for ML and smoothed ML-estimators under Huber's least favourable, the double exponential and the Cauchy distribution (Figs. 11, 12). Furthermore, the Huber estimator and the smoothed Huber estimator are compared under the normal, the double exponential and the Cauchy distribution (Figs. 13, 14). On the right side of the Figs. 11 and 13, relative differences, divided by the tail probability for the non-smoothed (Huber) estimator, are shown. The corresponding plots for the double exponential and Cauchy distribution are similar in the sense

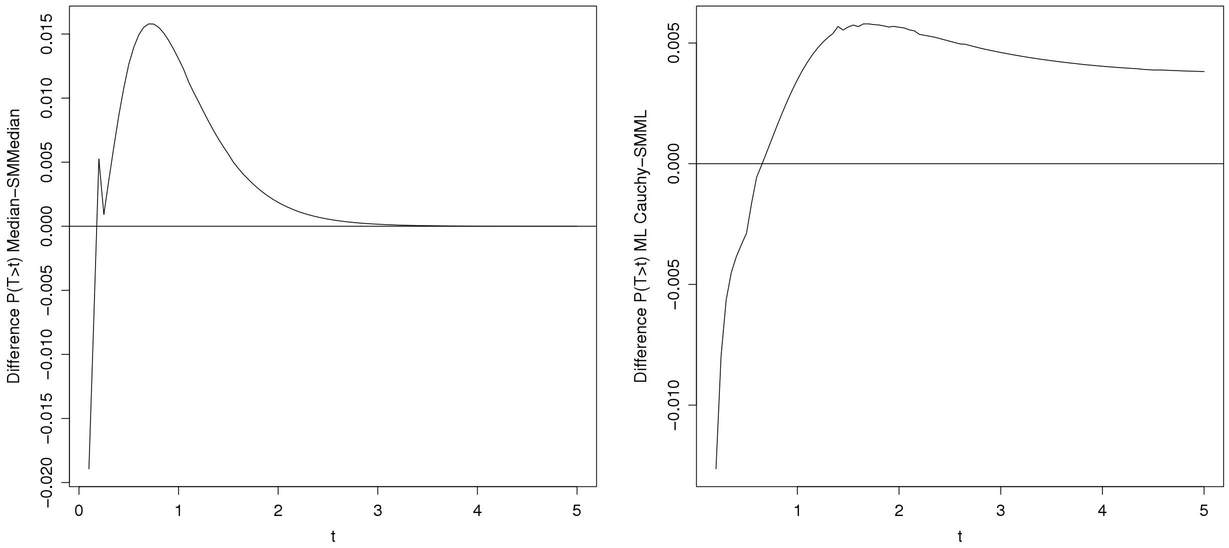

**Fig. 12.** Left: difference in tail probabilities, approximated by (5), between the ML (median) and the smoothed ML-estimator under the double exponential distribution ($n = 5$; values larger than 0 mean that the smoothed estimator is better). Right: same under the Cauchy distribution (using and smoothing the Cauchy ML-estimator).

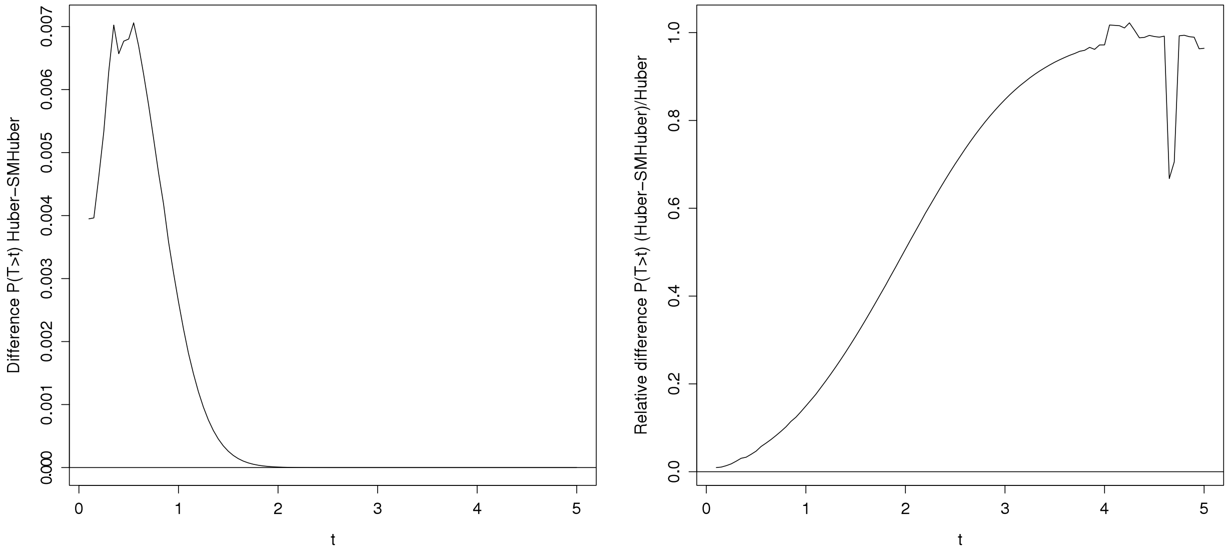

**Fig. 13.** Left: difference in tail probabilities, approximated by (5), between the Huber and the smoothed Huber estimator ($k = 0.862$) under the normal distribution. Right: relative difference (difference divided by the tail probability of the Huber estimator).

that they also show that the relative advantage of the smoothed estimators increases monotonically with $t$ in the region where the smoothed estimators are better (i.e., for $t$ large enough that the difference in tail probabilities is larger than 0). There are some small peculiarities in these plots and not all tail probabilities for $t \approx 0$ are displayed. This is due to numerical instabilities of either the involved numerical integration or some ratios close to 0/0.

The consistent pattern of all of these results is that the smoothed estimators are better than the non-smoothed ones in the tail areas, i.e., the probabilities that the smoothed estimators are very far away from the true values are lower. In terms of the relative difference, the advantage increases with $t$ (Table 1 gives the maximum relative difference for $t \leq 5$).

Another way to look at the results is to consider the values $P(|T_n| > t_0) = 2P(T_n > t_0)$ for $t_0$ so that $P(T_n > t_0)$ is equal for the smoothed and the non-smoothed estimator. This is a standardized way to measure the "size" of the tail area for which the smoothed estimator is better. The results are given in Table 1. Note that most of these values are smaller than 0.5, but this does not mean that the non-smoothed estimators are better, because in most applications it is more important that estimators are not *very* bad (i.e., in the far tail), whereas small deviations are tolerable.

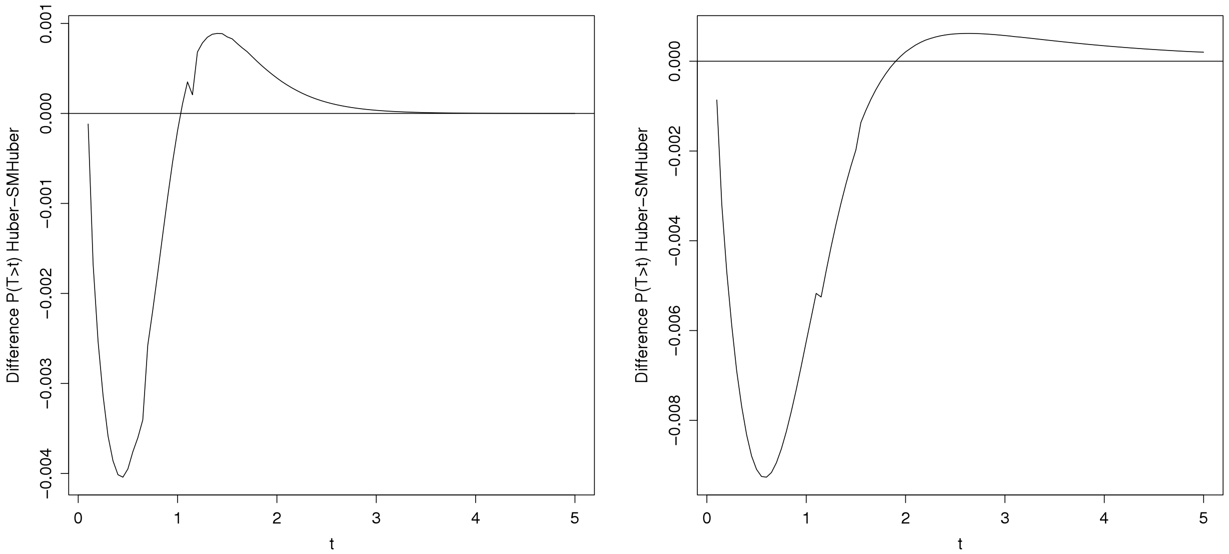

**Fig. 14.** Left: difference in tail probabilities, approximated by (5), between the Huber and the smoothed Huber estimator ($k = 0.862$) under the double exponential distribution. Right: same under the Cauchy distribution.

**Table 1**

Maximum relative difference $\frac{P(T_n>t)-P(\tilde{T}_n>t)}{P(T_n>t)}$ for the non-smoothed estimator $T_n$ compared to the smoothed estimator $\tilde{T}_n$ and $P(|T_n| > t_0)$ for $t_0$ so that $P(|T_n| > t_0) = P(|\tilde{T}_n| > t_0)$, i.e., probability of the area in which the smoothed estimator is better.

| Distribution | Estimators | Max. rel. difference | $P(\vert T_n\vert > t_0)$ |
|---|---|---|---|
| Huber's l.f. | Huber/sm.Huber | 0.47 | 0.31 |
| Double exp. | Median/sm.Median | 0.72 | 0.72 |
| Cauchy | Cau.ML/sm.Cau.ML | 0.34 | 0.32 |
| Normal | Huber/sm.Huber | 1.00 | 1.00 |
| Double exp. | Huber/sm.Huber | 0.53 | 0.097 |
| Cauchy | Huber/sm.Huber | 0.075 | 0.055 |

## 9. Conclusion

The simulation and small sample asymptotic results show that the idea of smoothing M-estimators can be worthwhile. Given that the higher quantiles and the MSE are judged as more adequate quality measures than the lower quantiles, the smoothed M-estimators performed better than their initial counterparts in all setups (the difference in efficiency being 10% and smaller, though).

The smoothed Huber estimator behaved very well not only under Huber's least favourable distribution, but also under the normal distribution, where it dominated the non-smoothed Huber estimator uniformly, and it was not much worse than the Huber estimator under the heavier tailed distributions. In the extreme tails of the error distribution, the smoothed Huber estimator was always better. However, this advantage is quite small under the Cauchy distribution. In the setups where the scale was unknown and estimated by the MAD, the smoothed Huber estimator was always better than the Huber estimator, and not much worse than the Pitman estimator for the specific simulated distribution. The latter was, with MAD scale, outperformed by the smoothed ML-estimator, but this estimator, as well as the Pitman estimator, of course require to assume the knowledge of the underlying parametric model.

Since the Huber estimator is widely used as a standard estimator and all its robustness properties (most of them of asymptotic nature) hold also for the smoothed Huber estimator, it would be a reasonable suggestion to replace the Huber estimator by the smoothed Huber, which is similarly easy to compute because of its explicit $\psi$-function (involving $\Phi$). Note that we do not claim that the use of the MAD as scale estimator is generally optimal. However, there are very many approaches to the robust estimation of location with unknown scale and there is no uniformly optimal method, and the MAD is the most popular choice because of its simplicity and the good robustness properties of location $M$-estimation with preliminary robust estimation of scale (Andrews et al., 1972; Huber, 1981; Hampel et al., 1986; Maronna et al., 2006). In order to improve small sample properties, smoother scale estimators may be helpful, perhaps even applying the same smoothing principle, but this can be expected to depend on the underlying distribution and is a topic for future research. Comparing this with Pitman estimators for location with unknown scale as derived e.g. in Bell Krystinik and Morgenthaler (1991) could also be of interest.

By showing that for small samples the ML-estimator is still good in the lower quantiles of the error distribution, but that the behaviour for higher quantiles is different, the simulation results further confirmed the heuristic (and incomplete) motivation given in Section 3. This may add something to the intuitive understanding of the central limit theorem and its relevance for small samples. It would be interesting to explore further the potential of the proposed smoothing principle for more complicated setups such as (generalized) linear regression or multivariate location.

There is a lot of literature concerning small sample behaviour of location estimators. The Pitman estimator for the Cauchy distribution has recently been investigated for small samples by Cohen Freue (2007). Ventura (1998) gives an approximation of the Pitman estimator (though it does not perform very well in the simulation study of Cohen Freue (2007)). Ionides (2005) and Seo and Lindsay (2010) define estimators by maximizing a smoothed version of the likelihood. This idea goes back to Daniels (1960). Sugiura and Naing (1989) discuss improvements of the median for the double exponential distribution. Barndorff-Nielsen (1983, 1986) give approximations to the conditional density of the ML-estimator and an adjustment of the signed log-likelihood ratio respectively to get higher order accuracy of asymptotic approximations. Comparing these developments with the smoothing principle discussed in the present paper (which we believe is more straightforward and versatile) may be interesting, but is outside the scope of the present study.

An R-package “smoothmest” computing the smoothed M-estimators discussed in the present paper is in preparation.

## Acknowledgements

The authors would like to thank the anonymous referees and the Associate Editor for constructive comments, which helped to improve the paper. The second author thanks the Department of Econometrics, University of Geneva, and the third author thanks the Forschungsinstitut für Mathematik, ETH Zürich for partial support during this research.

## 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.

Barndorff-Nielsen, O.E., 1983. On a formula for the conditional distribution of the maximum likelihood estimator. Biometrika 70, 343–365.

Barndorff-Nielsen, O.E., 1986. Inference on full or partial parameters based on the standardized signed log likelihood ratio. Biometrika 73, 307–322.

Bell Krystinik, K., Morgenthaler, S., 1991. Point estimation: location-and-scale configurations. In: Morgenthaler, S., Tukey, J.W. (Eds.), Configural Polysampling. Wiley, New York, pp. 37–48.

Chen, Z., Tyler, D.E., 2004. On the finite sample breakdown points of redescending M-estimates of location. Statistics and Probability Letters 69, 233–242.

Cohen Freue, G.V., 2007. The Pitman estimator of the Cauchy location parameter. Journal of Statistical Planning and Inference 137, 1900–1913.

Daniels, H.E., 1960. The asymptotic efficiency of a maximum likelihood estimator. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley, pp. 151–163.

Daniels, H.E., 1983. Saddlepoint approximations for estimating equations. Biometrika 70, 89–96.

Donoho, D.L., Huber, P.J., 1983. The notion of breakdown point. In: Bickel, P.J., Doksum, K., Hodges Jr., J.L. (Eds.), A Festschrift for Erich L. Lehmann. Wadsworth, Belmont, CA, pp. 157–184.

Field, C.A., Hampel, F.R., 1982. Small-sample asymptotic distributions of M-estimators of location. Biometrika 69, 29–46.

Field, C.A., Ronchetti, E., 1990. Small sample asymptotics. In: Institute of Mathematical Statistics — Monograph Series, Hayward, CA.

Hampel, F., 1973. Some small sample asymptotics. Hajek, J.(Ed.), Proceedings of the Prague Symposium on Asymptotic Statistics. Charles University of Prague.

Hampel, F., 1996. On the philosophical foundations of statistics: bridges to Hubers work and recent results. In: Rieder, H. (Ed.), Robust Statistics, Data Analysis and Computer Intensive Methods; In Honor of Peter Hubers 60th Birthday. Springer, New York, pp. 185–196.

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

Huber, P.J., 1964. Robust estimation of a location parameter. Annals of Mathematical Statistics 35, 73–101.

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

Ionides, E.L., 2005. Maximum smoothed likelihood estimation. Statistica Sinica 15, 1003–1014.

Lischer, P., 1996. Robust statistical methods in interlaboratory analytical studies. In: Rieder, H. (Ed.), Robust Statistics, Data Analysis and Computer Intensive Methods; In Honor of Peter Huber's 60th Birthday. Springer, New York, pp. 251–264.

Lugannani, R., Rice, S.O., 1980. Saddle point approximation for the distribution of the sum of independent random variables. Advances in Applied Probability 12, 475–490.

Maronna, A.R., Martin, D.R., Yohai, V.J., 2006. Robust Statistics: Theory and Methods. Wiley, New York.

Pitman, E.J., 1939. The estimation of the location and scale parameters of a continuous population of any given form. Biometrika 30, 391–421.

Rousseeuw, P.J., Verboven, S., 2002. Robust estimation in very small samples. Computational Statistics & Data Analysis 40, 741–758.

Seo, B., Lindsay, B.G., 2010. A computational strategy for doubly smoothed MLE exemplified in the normal mixture model. Computational Statistics & Data Analysis 54, 1930–1941.

Sugiura, N., Naing, M.T., 1989. Improved estimators for the location of double exponential distribution. Communications in Statistics — Theory and Methods 18, 541–554.

Ventura, L., 1998. Higher-order approximations for pitman estimators and for optimal compromise estimators. Canadian Journal of Statistics 26, 49–55.