Gather, Ursula; Davies, P. Laurie

Working Paper

Robust Statistics

Papers, No. 2004,20

Provided in Cooperation with:
CASE - Center for Applied Statistics and Economics, Humboldt University Berlin

Suggested Citation: Gather, Ursula; Davies, P. Laurie (2004) : Robust Statistics, Papers, No. 2004,20, Humboldt-Universität zu Berlin, Center for Applied Statistics and Economics (CASE), Berlin

This Version is available at:
https://hdl.handle.net/10419/22194

Robust Statistics

Laurie Davies[1] and Ursula Gather[2]

[1] Department of Mathematics, University of Essen, 45117 Essen, Germany,
laurie.davies@uni-essen.de
[2] Department of Statistics, University of Dortmund, 44221 Dortmund, Germany,
gather@statistik.uni-dortmund.de

1 Robust statistics; Examples and Introduction

1.1 Two examples

The first example involves the real data given in Table 1 which are the results of an interlaboratory test. The boxplots are shown in Fig. 1 where the dotted line denotes the mean of the observations and the solid line the median.

Table 1. The results of an interlaboratory test involving 14 laboratories

1	2	3	4	5	6	7	9	9	10	11	12	13	14
1.4	5.7	2.64	5.5	5.2	5.5	6.1	5.54	6.0	5.1	5.5	5.9	5.5	5.3
1.5	5.8	2.88	5.4	5.7	5.8	6.3	5.47	5.9	5.1	5.5	5.6	5.4	5.3
1.4	5.8	2.42	5.1	5.9	5.3	6.2	5.48	6.1	5.1	5.5	5.7	5.5	5.4
0.9	5.7	2.62	5.3	5.6	5.3	6.1	5.51	5.9	5.3	5.3	5.6	5.6	

We note that only the results of the Laboratories 1 and 3 lie below the mean whereas all the remaining laboratories return larger values. In the case of the median, 7 of the readings coincide with the median, 24 readings are smaller and 24 are larger. A glance at Fig. 1 suggests that in the absence of further information the Laboratories 1 and 3 should be treated as outliers. This is the course which we recommend although the issues involved require careful thought. For the moment we note simply that the median is a robust statistic whereas the mean is not.

The second example concerns quantifying the scatter of real valued observations $x_1, \ldots, x_n$. This example is partially taken from Huber (1981) and reports a dispute between Eddington (1914, p.147) and Fisher (1920, p.762) about the relative merits of

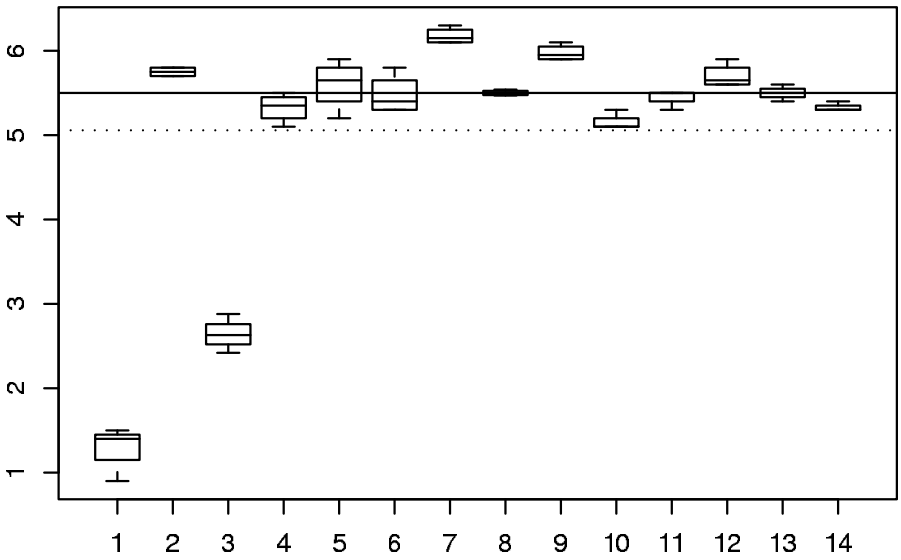

Fig. 1. A boxplot of the data of Table 1. The dotted line and the solid line denote respectively the mean and the median of the observations.

$$s_n = \left(\frac{1}{n}\sum(x_i - \bar{x})^2\right)^{\frac{1}{2}} \text{ and } d_n = \frac{1}{n}\sum|x_i - \bar{x}|.$$

Fisher argued that for normal observations the standard deviation s_n is about 12% more efficient than the mean absolute deviation d_n. In contrast Eddington claimed that his experience with real data indicates that d_n is better than s_n. In Tukey (1960) and Huber (1977) we find a resolution of this apparent contradiction. Consider the model

$$\mathcal{N}_\varepsilon = \big(1-\varepsilon\big)N(\mu,\sigma^2) + \varepsilon N\big(\mu, 9\sigma^2\big) \tag{1}$$

where $N(\mu,\sigma^2)$ denotes a normal distribution with mean μ and variance σ^2 and $0 \le \varepsilon \le 1$. For data distributed according to (1) one can calculate the asymptotic relative efficiency ARE of d_n with respect to s_n,

$$ARE(\varepsilon) = \lim_{n\to\infty} RE_n(\varepsilon) = \lim_{n\to\infty} \frac{Var(s_n)/E(s_n)^2}{Var(d_n)/E(d_n)^2}.$$

As Huber states, the result is disquieting. Already for $\varepsilon \ge 0.002$ ARE exceeds 1 and the effect is apparent for samples of size 1000. For $\varepsilon = 0.05$ we have $ARE(\varepsilon) = 2.035$ and simulations show that for samples of size 20 the relative efficiency exceeds 1.5 and increases to 2.0 for samples of size 100. This is a severe deficiency of s_n as models such as $\mathcal{N}_\varepsilon$ with ε between 0.01 and 0.1 often give better descriptions of real data than the normal distribution itself. We quote Huber (1981)

"thus it becomes painfully clear that the naturally occurring deviations from the idealized model are large enough to render meaningless the traditional asymptotic optimality theory".

1.2 General philosophy

The two examples of the previous section illustrate a general phenomenon. An optimal statistical procedure based on a particular family of models $\mathcal{M}_1$ can differ considerably from an optimal procedure based on another family $\mathcal{M}_2$ even though the families $\mathcal{M}_1$ and $\mathcal{M}_2$ are very close. This may be expressed by saying that optimal procedures are often unstable in that small changes in the data or the model can lead to large changes in the analysis. The basic philosophy of robust statistics is to produce statistical procedures which are stable with respect to small changes in the data or model and even large changes should not cause a complete breakdown of the procedure.

Any inspection of the data and the removal of aberrant observations may be regarded as part of robust statistics but it was only with Pearson (1931) that the consideration of deviations from models commenced. He showed that the exact theory based on the normal distribution for variances is highly non-robust. There were other isolated papers on the problem of robustness (Pearson, 1929; Bartlett, 1935; Geary, 1936, 1947; Gayen, 1950; Box, 1953; Box and Andersen, 1955). Tukey (1960) initiated a wide spread interest in robust statistics which has continued to this day. The first systematic investigation of robustness is due to Huber (1964) and was expounded in Huber (1981). Huber's approach is functional analytic and he was the first to investigate the behaviour of a statistical functional over a full topological neighbourhood of a model instead of restricting the investigation to other parametric families as in (1). Huber considers three problems. The first is that of minimizing the bias over certain neighbourhoods and results in the median as the most robust location functional. For large samples deviations from the model have consequences which are dominated by the bias and so this is an important result. The second problem is concerned with what Tukey calls the statistical version of no free lunches. If we take the simple model of i.i.d. $N(\mu, 1)$ observations then the confidence interval for μ based on the mean is on average shorter than that based on any other statistic. If short confidence intervals are of interest then one can not only choose the statistic which gives the shortest interval but also the model itself. The new model must of course still be consistent with the data but even with this restriction the confidence interval can be made as small as desired (Davies, 1995). Such a short confidence interval represents a free lunch and if we do not believe in free lunches then we must look for that model which *maximizes* the length of the confidence interval over a given family of models. If we take all distributions with variance 1 then the confidence interval for the $N(\mu, 1)$ distribution is the longest. Huber considers the same problem over the family $\mathcal{F} = \{F : d_{ko}(F, N(0, 1)) < \varepsilon\}$ where d_{ko} denotes the Kolmogoroff metric. Under certain simplifying assumptions Huber solves

this problem and the solution is known as the Huber distribution (see Huber, 1981). Huber's third problem is the robustification of the Neyman-Pearson test theory. Given two distributions P_0 and P_1 Neyman and Pearson (1933) derive the optimal test for testing P_0 against P_1. Huber considers full neighbourhoods $\mathcal{P}_0$ of P_0 and $\mathcal{P}_1$ of P_1 and then derives the form of the minimax test for the composite hypothesis of $\mathcal{P}_0$ against $\mathcal{P}_1$. The weakness of Huber's approach is that it does not generalize easily to other situations. Nevertheless it is the spirit of this approach which we adopt here. It involves treating estimators as functionals on the space of distributions, investigating where possible their behaviour over full neighbourhoods and always being aware of the danger of a free lunch.

Hampel (1968) introduced another approach to robustness, that based on the influence function $I(x, T, F)$ defined for a statistical functional T as follows

$$I(x, T, F) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F + \varepsilon\delta_x) - T(F)}{\varepsilon} \tag{2}$$

where δ_x denotes the point mass at the point x. The influence function has two interpretations. On the one hand it measures the infinitesimal influence of an observation situated at the point x on the value of the functional T. On the other hand if $P_n(F)$ denotes the empirical measure of a sample of n i.i.d. random variables with common distribution F then under appropriate regularity conditions

$$\lim_{n\to\infty} \sqrt{n}(T(P_n(F)) - T(F)) \stackrel{D}{=} N\left(0, \int I(x, T, F)^2 \, dF(x)\right) \tag{3}$$

where $\stackrel{D}{=}$ denotes equality of distribution. Given a parametric family $\mathcal{P}' = \{P_\theta : \theta \in \Theta\}$ of distributions we restrict attention to those functionals which are Fisher consistent that is

$$T(P_\theta) = \theta, \quad \theta \in \Theta. \tag{4}$$

Hampel's idea was to minimize the asymptotic variance of T as an estimate of a parameter θ subject to a bound on the influence function

$$\min_T \int I(x, T, P_\theta)^2 \, dP_\theta(x) \quad \text{under (4) and } \sup_x |I(x, T, P_\theta)| \le k(\theta) \tag{5}$$

where $k(\theta)$ is a given function of θ. Hampel complemented the infinitesimal part of his approach by considering also the global behaviour of the functional T. He introduced the concept of breakdown point which has had and continues to have a major influence on research in robust statistics. The approach based on the influence function was carried out in Hampel et al. (1986). The strength of the Hampel approach is that it can be used to robustify in some sense the estimation of parameters in any parametric model. The weaknesses are that (5) only bounds infinitesimally small deviations from the model and that the

approach does not explicitly take into account the free lunch problem. Hampel is aware of this and recommends simple models but simplicity is an addition to and not an integral part of his approach. The influence function is usually used as a heuristic tool and care must be taken in interpreting the results. For examples of situations where the heuristics go wrong we refer to Davies (1993).

Another approach which lies so to speak between that of Huber and Hampel is the so called shrinking neighbourhood approach. It has been worked out in full generality by Rieder (1994). Instead of considering neighbourhoods of a fixed size (Huber) or only infinitesimal neighbourhoods (Hampel) this approach considers full neighbourhoods of a model but whose size decreases at the rate of $n^{-1/2}$ as the sample size n tends to infinity. The size of the neighbourhoods is governed by the fact that for larger neighbourhoods the bias term is dominant whereas models in smaller neighbourhoods cannot be distinguished. The shrinking neighbourhoods approach has the advantage that it does not need any assumptions of symmetry. The disdavantage is that the size of the neighbourhoods goes to zero so that the resulting theory is only robustness over vanishingly small neighbourhoods.

1.3 Functional approach

Although a statistic based on a data sample may be regarded as a function of the data a more general approach is often useful. Given a data set $(x_1, \ldots, x_n)$ we define the corresponding empirical distribution P_n by

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \tag{6}$$

where δ_x denotes the unit mass in x. Although P_n clearly depends on the sample $(x_1, \ldots, x_n)$ we will usually suppress the dependency for the sake of clarity. With this notation we can now regard the arithmetic mean $\bar{x}_n = \sum_{i=1}^{n} x_i/n$ either as a function of the data or as a function T_{av} of the empirical measure P_n,

$$\bar{x}_n = \int x \, dP_n(x) = T_{av}(P_n).$$

The function T_{av} can be extended to all measures P which have a finite mean

$$T_{av}(P) = \int x \, dP(x) \tag{7}$$

and is now a functional defined on a certain subset of the family $\mathcal{P}$ of probability measures on $\mathbb{R}$. This manner of treating statistics is one whose origins go back to von Mises (1937). In the context of robust statistics it was introduced by Huber (1964) and has proved very useful (see Fernholz, 1983). Another example is given by the functional T_{sh} defined as the length of the shortest interval which carries a mass of at least $1/2$

$$T_{sh}(P) = \text{argmin}\{|I| : P(I) \geq 1/2,\, I \subset \mathbb{R}\} \tag{8}$$

where $|I|$ denotes the length of the interval I. The idea of using the shortest half interval goes back to Tukey (see Andrews et al., 1972) who proposed using the mean of the observations contained in it as a robust location functional.

The space $\mathcal{P}$ may be metricized in many ways but we prefer the Kolmogoroff metric d_{ko} defined by

$$d_{ko}(P, Q) = \sup_{x \in \mathbb{R}} |P((-\infty,\, x]) - Q((-\infty,\, x])|. \tag{9}$$

The Glivenko-Cantelli theorem states

$$\lim_{n \to \infty} d_{ko}(P_n(P), P) = 0, \quad a.s. \tag{10}$$

where $P_n(P)$ denotes the empirical measure of the n random variables $X_1(P), \ldots, X_n(P)$ of the i.i.d. sequence $(X_i(P))_1^\infty$. In conjunction with (10) the metric d_{ko} makes it possible to connect analytic properties of a functional T and its statistical properties. As a first step we note that a functional T which is locally bounded in the Kolmogoroff metric

$$\sup\{|T(Q) - T(P)| : d_{ko}(P, Q) < \varepsilon)\} < \infty \tag{11}$$

for some $\varepsilon > 0$ offers protection against outliers. On moving from local boundedness to continuity we see that if a functional T is continuous at P then the sequence $T(P_n(P))$ is a consistent statistic in that

$$\lim_{n \to \infty} T(P_n(P)) = T(P), \quad a.s.$$

Finally we consider a functional T which is differentiable at P, that is

$$T(Q) - T(P) = \int I(x, P, T)\, d(Q - P)(x) + \mathrm{o}_P(d_{ko}(P, Q)) \tag{12}$$

for some bounded function $I(\cdot, P, T) : \mathbb{R} \to \mathbb{R}$ where, without loss of generality, $\int I(x, P, T)\, dP(x) = 0$ (see Clarke, 1983). On putting

$$Q = Q_\varepsilon = (1 - \varepsilon)P + \varepsilon\delta_x$$

it is seen that $I(x, P, T)$ is the influence function of (2). As

$$d_{ko}(P_n(P), P) = \mathrm{O}_P(1/\sqrt{n}) \tag{13}$$

the central limit theorem (3) follows immediately. Textbooks which make use of this functional analytic approach are as already mentioned Huber (1981), Hampel et al. (1986), Rieder (1994), and also Staudte and Sheather (1990) a book which can be strongly recommended to students as a well written and at the same time deep introductory text.

2 Location and scale in $\mathbb{R}$

2.1 Location, scale and equivariance

Changes in measurement units and baseline correspond to affine transformations on $\mathbb{R}$. We write

$$\mathcal{A} = \{A : \mathbb{R} \to \mathbb{R} \text{ with } A(x) = ax + b,\ a \neq 0,\ b \in \mathbb{R}\}. \tag{14}$$

For any probability measure P and for any $A \in \mathcal{A}$ we define

$$P^A(B) = P(\{x : A(x) \in B\}), \quad B \in \mathcal{B}, \tag{15}$$

$\mathcal{B}$ denoting all Borel sets on $\mathbb{R}$. Consider a subset $\mathcal{P}'$ of $\mathcal{P}$ which is closed under affine transformations, that is

$$P \in \mathcal{P}' \Rightarrow P^A \in \mathcal{P}' \quad \text{for all } P \in \mathcal{P}',\ A \in \mathcal{A}. \tag{16}$$

A functional $T_l : \mathcal{P}' \to \mathbb{R}$ will be called a location functional on $\mathcal{P}'$ if

$$T_l(P^A) = A(T_l(P)), \quad A \in \mathcal{A},\ P \in \mathcal{P}'. \tag{17}$$

Similarly we define a functional $T_s : \mathcal{P}' \to \mathbb{R}_+$ to be a scale functional if

$$T_s(P^A) = |a| T_s(P), \quad A \in \mathcal{A},\ A(x) = ax + b,\ P \in \mathcal{P}'. \tag{18}$$

2.2 Existence and uniqueness

The fact that the mean T_{av} of (7) cannot be defined for all distributions is an indication of its lack of robustness. More precisely the functional T_{av} is not locally bounded (11) in the metric d_{ko} at any distribution P. The median $\mathrm{MED}(P)$ can be defined at any distribution P as the mid-point of the interval of m-values for which

$$P((-\infty, m]) \geq 1/2 \text{ and } P([m, \infty)) \geq 1/2. \tag{19}$$

Similar considerations apply to scale functionals. The standard deviation requires the existence of the second moment of a distribution. The median absolute deviation MAD (see Andrews et al., 1972) of a distribution can be well defined at all distributions as follows. Given P we define P' by

$$P'(B) = P(\{x : |x - \mathrm{MED}(P)| \in B\}), \quad B \in \mathcal{B}.$$

and set

$$\mathrm{MAD}(P) = \mathrm{MED}(P'). \tag{20}$$

2.3 M-estimators

An important family of statistical functionals is the family of M-functionals introduced by Huber (1964) Let ψ and χ be functions defined on $\mathbb{R}$ with values in the interval $[-1, 1]$. For a given probability distribution P we consider the following two equations for m and s

$$\int \psi\left(\frac{x-m}{s}\right) dP(x) = 0 \tag{21}$$

$$\int \chi\left(\frac{x-m}{s}\right) dP(x) = 0. \tag{22}$$

If the solution exists and is uniquely defined we denote it by

$$T(P) = (T_l(P), T_s(P)) = (m, s).$$

In order to guarantee existence and uniqueness conditions have to be placed on the functions ψ and χ as well as on the probability measure P. The ones we use are due to Scholz (1971) (see also Huber, 1981) and are as follows:

($\boldsymbol{\psi 1}$) $\psi(-x) = -\psi(x)$ for all $x \in \mathbb{R}$.
($\boldsymbol{\psi 2}$) ψ is strictly increasing
($\boldsymbol{\psi 3}$) $\lim_{x\to\infty} \psi(x) = 1$
($\boldsymbol{\psi 4}$) ψ is continuously differentiable with derivative $\psi^{(1)}$.

($\boldsymbol{\chi 1}$) $\chi(-x) = \chi(x)$ for all $x \in \mathbb{R}$.
($\boldsymbol{\chi 2}$) $\chi : \mathbb{R}_+ \to [-1, 1]$ is strictly increasing
($\boldsymbol{\chi 3}$) $\chi(0) = -1$
($\boldsymbol{\chi 4}$) $\lim_{x\to\infty} \chi(x) = 1$
($\boldsymbol{\chi 5}$) χ is continuously differentiable with derivative $\chi^{(1)}$.

($\boldsymbol{\psi\chi 1}$) $\chi^{(1)}/\psi^{(1)} : \mathbb{R}_+ \to \mathbb{R}_+$ is strictly increasing

If these conditions hold and P satisfies

$$\Delta(P) = \max_x P(\{x\}) < 1/2 \tag{23}$$

then the equations (21) and (22) have precisely one solution. If we set

$$\mathcal{P}' = \{P : \Delta(P) < 1/2\}$$

then $\mathcal{P}'$ satisfies (16) and $T_l : \mathcal{P}' \to \mathbb{R}$ and $T_s : \mathcal{P}' \to \mathbb{R}_+$ are a location and a scale functional respectively. Two functions which satisfy the above conditions are

$$\psi(x) = \frac{\exp(x/c) - 1}{\exp(x/c) + 1} \tag{24}$$

$$\chi(x) = \frac{x^4 - 1}{x^4 + 1} \tag{25}$$

where $c < 0.39$ is a tuning parameter. The restriction on c is to guarantee $(\boldsymbol{\psi\chi 1})$. Algorithms for calculating the solution of (21) and (22) are given in the Fortran library ROBETH (Marazzi, 1992) which also contains many other algorithms related to robust statistics.

The main disadvantage of M-functionals defined by (21) and (22) is $(\boldsymbol{\psi\chi 1})$ which links the location and scale parts in a manner which may not be desirable. In particular there is a conflict between the breakdown behaviour and the efficiency of the M-functional (see below). There are several ways of overcoming this. One is to take the scale function T_s and then to calculate a second location functional by solving

$$\int \tilde{\psi}\left(\frac{x-m}{T_s(P)}\right) dP(x) = 0. \tag{26}$$

If now $\tilde{\psi}$ satisfies $(\boldsymbol{\psi 1})$-$(\boldsymbol{\psi 4})$ then this new functional will exist only under the assumption that the scale functional exists and is non-zero. Furthermore the functional can be made as efficient as desired by a suitable choice of $\tilde{\psi}$ removing the conflict between breakdown and efficiency. One possible choice for $T_s(P)$ is the MAD of (20) which is simple, highly robust and which performed well in the Princeton robustness study (Andrews et al., 1972).

In some situations there is an interest in downweighting outlying observations completely rather than in just bounding their effect. A downweighting to zero is not possible for a ψ-function which satisfies $(\boldsymbol{\psi 2})$ but can be achieved by using so called redescending ψ-functions such as Tukey's biweight

$$\tilde{\psi}(x) = x(1-x^2)^2\{|x| \leq 1\}. \tag{27}$$

In general there will be many solutions of (26) for such ψ-functions and to obtain a well defined functional some choice must be made. One possibility is to take the solution closest to the median, another is to take

$$\text{argmin}_m \int \rho\left(\frac{x-m}{T_s(P)}\right) dP(x) \tag{28}$$

where $\rho^{(1)} = \tilde{\psi}$. Both solutions pose algorithmic problems. The effect of downweighting outlying observations to zero can be attained by using a so called one-step functional T_{om} defined by

$$T_{om}(P) = T_m(P) + T_s(P)\frac{\int \tilde{\psi}\left(\frac{x-T_m(P)}{T_s(P)}\right) dP(x)}{\int \tilde{\psi}^{(1)}\left(\frac{x-T_m(P)}{T_s(P)}\right) dP(x)} \tag{29}$$

where T_m is as above and $\tilde{\psi}$ is redescending. We refer to Hampel et al. (1986) and Rousseeuw and Croux (1994) for more details.

So far all scale functionals have been defined in terms of a deviation from a location functional. This link can be broken as follows. Consider the functional T_{ss} defined to be the solution s of

$$\int \chi\left(\frac{x-y}{s}\right) dP(x)\, dP(y) = 0. \tag{30}$$

where χ satisfies the conditions above. It may be shown that the solution is unique with $0 < s < \infty$, if

$$\sum_{a_i} P(\{a_i\})^2 < 1/4 \tag{31}$$

where the a_i denote the countably many atoms of P. The main disadvantage of this method is the computational complexity of (30) requiring as it does $\mathrm{O}(n^2)$ operations for a sample of size n. If χ is of the form

$$\chi(x) = \begin{cases} a > 0, & |x| > 1, \\ b < 0, & |x| \leq 1 \end{cases}$$

then T_{ss} reduces to a quantile of the $|x_i - x_j|$ and much more efficient algorithms exist which allow the functional to be calculated in $\mathrm{O}(n \log n)$ operations (see Croux and Rousseeuw, 1992, Rousseeuw and Croux, 1992, and Rousseeuw and Croux, 1993).

Although we have defined M-functionals as a solution of (21) and (22) there are sometimes advantages in defining them as a solution of a minimization problem. Consider the Cauchy distribution with density

$$f(x : \mu, \sigma) = \frac{1}{\pi} \frac{\sigma}{\sigma^2 + (x-\mu)^2}. \tag{32}$$

We now define $\boldsymbol{T}_c(P) = (T_{cm}(P), T_{cs}(P))$ by

$$\boldsymbol{T}_c(P) = \mathrm{argmin}_{(m,s)} \left(- \int \log(f(x : m, s))\, dP(x) + \frac{1}{2} \log(s) \right). \tag{33}$$

This is simply the standard maximum likelihood estimate for a Cauchy distribution but there is no suggestion here that the data are so distributed. If $\Delta(P) < 1/2$ it can be shown that the solution exists and is unique. Moreover there exists a simple convergent algorithm for calculating $(T_{cm}(P), T_{cs}(P))$ for a data sample. We refer to Kent and Tyler (1991) for this and the multidimensional case to be studied below. By differentiating the right hand side of (33) it is seen that $(T_{cm}(P), T_{cs}(P))$ may be viewed as an M-functional with a redescending ψ-function.

Another class of functionals defined by a minimization problem is the class of S-functionals. Given a function $\rho : \mathbb{R} \to [0, 1]$ which is symmetric, continuous on the right and non-increasing on $\mathbb{R}_+$ with $\rho(1) = 1$ and $\lim_{x\to\infty} \rho(x) = 0$. We define $(T_{sm}(P), T_{ss}(P))$ by

$$(T_{sm}(P), T_{ss}(P)) = \mathrm{argmin}_{(m,s)} \left\{ s : \int \rho((x-m)/s)\, dP(x) \geq 1/2 \right\}. \tag{34}$$

A special case is a minor variation of the shortest-half functional of (8) which is obtained by taking ρ to be the indicator function of the interval $[0, 1)$. Although the existence of solutions of (34) is guaranteed if $\Delta(P) < 1/2$ the problem of uniqueness is not trivial and requires the existence of a density subject to certain conditions. If ρ is smooth then by differentiation it is seen that $(T_{sm}(P), T_{ss}(P))$ may be regarded as an M-functional with a redescending ψ-function given by $\tilde{\psi} = \rho^{(1)}$. The minimization problem (34) acts as a choice function. We refer to Davies (1987).

2.4 Bias and breakdown

Given a location functional T_l the bias is defined by

$$b(T_l, P, \varepsilon, d_{ko}) = \sup\{|T_l(Q) - T_l(P)| : d_{ko}(P, Q) < \varepsilon\} \tag{35}$$

where by convention $T_l(Q) = \infty$ if T_l is not defined at Q. For a scale functional T_s we set

$$b(T_s, P, \varepsilon, d_{ko}) = \sup\{|\log(T_s(Q)/T_s(P))| : d_{ko}(P, Q) < \varepsilon\} \tag{36}$$

where again by convention $T_s(Q) = \infty$ if T_s is not defined at Q. A popular although weaker form of bias function based on the so called gross error neighbourhood is given by

$$b(T_l, P, \varepsilon, GE) = \sup\{|T_l(Q) - T_l(P)| : Q = (1 - \varepsilon)P + \varepsilon H, H \in \mathcal{P}\} \tag{37}$$

with a corresponding definition for $b(T_s, P, \varepsilon, GE)$. We have

$$b(T_l, P, \varepsilon, GE) \leq b(T_l, P, \varepsilon, d_{ko}). \tag{38}$$

We refer to Huber (1981) for more details.

The breakdown point $\varepsilon^*(T_l, P, d_{ko})$ of T_l at P with respect to d_{ko} is defined by

$$\varepsilon^*(T_l, P, d_{ko}) = \sup\{\varepsilon : b(T_l, P, \varepsilon, d_{ko}) < \infty\} \tag{39}$$

with the corresponding definitions for scale functionals and the gross error neighbourhood. Corresponding to (38) we have

$$\varepsilon^*(T_l, P, d_{ko}) \leq \varepsilon^*(T_l, P, GE). \tag{40}$$

If a functional T_l has a positive breakdown point at a distribution P then it exhibits a certain degree of stability in a neighbourhood of P as may be seen as follows. Consider a sample $x_1, \ldots, x_n$ and add to it k further observations $x_{n+1}, \ldots, x_{n+k}$. If P_n and P_{n+k} denote the empirical measures based on $x_1, \ldots, x_n$ and $x_1, \ldots, x_{n+k}$ respectively then $d_{ko}(P_n, P_{n+k}) \leq k/(n + k)$. In particular if $k/(n + k) < \varepsilon^*(T_l, P_n, d_{ko})$ then it follows that $T_l(P_{n+k})$ remains bounded whatever the added observations . This finite sample concept

of breakdown was introduced by Donoho and Huber (1983). Another version replaces k observations by other values instead of adding k observations and is as follows. Let $x_1^k, \ldots, x_n^k$ denote a sample differing from $x_1, \ldots, x_n$ in at most k readings. We denote the empirical distributions by P_n^k and define

$$\varepsilon^*(T_l, P_n, fsbp) = \max\{k/n : |T_l(P_n^k)| < \infty\} \tag{41}$$

where P_n^k ranges over all possible $x_1^k, \ldots, x_n^k$. This version of the finite sample breakdown point is called the replacement version as k of the original observations can be replaced by arbitrary values. The two breakdown points are related (see Zuo, 2001). There are corresponding versions for scale functionals.

For location and scale functionals there exist upper bounds for the breakdown points. For location functionals T_l we have

Theorem 1.

$$\varepsilon^*(T_l, P, d_{ko}) \leq 1/2, \tag{42}$$
$$\varepsilon^*(T_l, P, GE) \leq 1/2, \tag{43}$$
$$\varepsilon^*(T_l, P_n, fsbp) \leq \lfloor n/2 \rfloor / n. \tag{44}$$

We refer to Huber (1981) It may be shown that all breakdown points of the mean are zero whereas the median attains the highest possible breakdown point in each case.The corresponding result for scale functionals is more complicated. Whereas we know of no reasonable metric in (42) of Theorem 1 which leads to a different upper bound this is not the case for scale functionals. Huber (1981) shows that for the Kolmogoroff metric d_{ko} the corresponding upper bound is $1/4$ but is $1/2$ for the gross error neighbourhood. If we replace the Kolmogoroff metric d_{ko} by the standard Kuiper metric d_{ku} defined by

$$d_{ku}(P, Q) = \sup\{|P(I) - Q(I)| : I \text{ an interval}\} \tag{45}$$

then we again obtain an upper bound of $1/2$. For scale functionals T_s we have

Theorem 2.

$$\varepsilon^*(T_s, P, d_{ku}) \leq (1 - \Delta(P))/2, \tag{46}$$
$$\varepsilon^*(T_s, P, GE) \leq (1 - \Delta(P))/2, \tag{47}$$
$$\varepsilon^*(T_s, P_n, fsbp) \leq (1 - \Delta(P))/2. \tag{48}$$

Similarly all breakdown points of the standard deviation are zero but, in contrast to the median, the MAD does not attain the upper bounds of (44). We have

$$\varepsilon^*(MAD, P_n, fsbp) = \max\{0,\, 1/2 - \Delta(P_n)\}.$$

A simple modification of the MAD, namely

$$\text{MMAD}(P) = \min\{|I| : \tilde{P}(I) \geq (1 + \Delta(I))/2\} \tag{49}$$

where $\tilde{P}(B) = P(\{x : |x - \text{MED}(P)| \in B\})$ and $\Delta(I) = \max\{P(\{x\}), x \in I\}$ can be shown to obtain the highest possible finite sample breakdown point of (48).

The M-functional defined by (21) and (22) has a breakdown point ε^* which satisfies

$$\psi^{-1}\left(\frac{\varepsilon^*}{1-\varepsilon^*}\right) = \chi^{-1}\left(\frac{-\varepsilon^*}{1-\varepsilon^*}\right) \tag{50}$$

(see Huber, 1981). For the functions defined by (24) and (25) the breakdown point is a decreasing function of c. As c tends to zero the breakdown point tends to $1/2$. Indeed, as c tends to zero the location part of the functional tends to the median. For $c = 0.2$ numerical calculations show that the breakdown point is 0.48. The calculation of breakdown points is not always simple. We refer to Huber (1981) and Gather and Hilker (1997).

The breakdown point is a simple but often effective measure of the robustness of a statistical functional. It does not however take into account the size of the bias. This can be done by trying to quantify the minimum bias over some neighbourhood of the distribution P and if possible to identify a functional which attains it. We formulate this for $P = N(0, 1)$ and consider the Kolmogoroff ball of radius ε. We have (Huber, 1981)

Theorem 3. *For every $\varepsilon < 1/2$ we have*

$$b(MED, P, \varepsilon, d_{ko}) \leq b(T_l, P, \varepsilon, d_{ko})$$

for any translation functional T_l.

In other words the median minimizes the bias over any Kolmogoroff neighbourhood of the normal distribution. This theorem can be extended to other symmetric distributions and to other situations (Riedel, 1989a,b). It is more difficult to obtain such a theorem for scale functionals because of the lack of a property equivalent to symmetry for location. Nevertheless some results in this direction have been obtained and indicate that the length of the shortest half T_{sh} of (8) has very good bias properties (Martin and Zamar, 1993b).

2.5 Confidence intervals and differentiability

Given a sample $x_1, \ldots, x_n$ with empirical measure P_n we can calculate a location functional $T_l(P_n)$ which in some sense describes the location of the sample. Such a point value is rarely sufficient and in general should be supplemented by a confidence interval, that is a range of values consistent with the data. If T_l is differentiable (12) and the data are i.i.d. random variables with distribution P then it follows from (3) (see Section 1.3) that an asymptotic α-confidence interval for $T_l(P)$ is given by

$$[T_l(P_n(P)) - z((1+\alpha)/2)\Sigma(P)/\sqrt{n}, T_l(P_n(P)) + z((1+\alpha)/2)\Sigma(P)/\sqrt{n}]. \tag{51}$$

Here $z(\alpha)$ denotes the α-quantile of the standard normal distribution and

$$\Sigma(P)^2 = \int I(x, T_l, P)^2 \, dP(x). \tag{52}$$

At first glance this cannot lead to a confidence interval as P is unknown. If however $\Sigma(P)$ is also Fréchet differentiable at P then we can replace $\Sigma(P)$ by $\Sigma(P_n(P))$ with an error of order $\mathrm{O}_P(1/\sqrt{n})$. This leads to the asymptotic α-confidence interval

$$[T_l(P_n(P)) - z((1+\alpha)/2)\Sigma(P_n(P))/\sqrt{n}, T_l(P_n(P)) + z((1+\alpha)/2)\Sigma(P_n(P))/\sqrt{n}]. \tag{53}$$

A second problem is that (53) depends on asymptotic normality and the accuracy of the interval in turn will depend on the rate of convergence to the normal distribution which in turn may depend on P. Both problems can be overcome if T_l is locally uniformly Fréchet differentiable at P. If we consider the M-functionals of Section 2.3 then they are locally uniformly Fréchet differentiable if the ψ- and χ-functions are sufficiently smooth (see Bednarski et al., 1991, Bednarski, 1993, Bednarski and Clarke, 1998, and Davies, 1998). The influence function $I(\cdot, T_l, P)$ is given by

$$I(x, T_l, P) = T_s(P) \frac{D(P)\tilde{\psi}\left(\frac{x - T_l(P)}{T_s(P)}\right) - B(P)\chi\left(\frac{x - T_l(P)}{T_s(P)}\right)}{A(P)D(P) - B(P)C(P)} \tag{54}$$

where

$$A(P) = \int \tilde{\psi}^{(1)}\left(\frac{x - T_l(P)}{T_s(P)}\right) dP(x) \tag{55}$$

$$B(P) = \int \left(\frac{x - T_l(P)}{T_s(P)}\right) \tilde{\psi}^{(1)}\left(\frac{x - T_l(P)}{T_s(P)}\right) dP(x) \tag{56}$$

$$C(P) = \int \chi^{(1)}\left(\frac{x - T_l(P)}{T_s(P)}\right) dP(x) \tag{57}$$

$$D(P) = \int \left(\frac{x - T_l(P)}{T_s(P)}\right) \chi^{(1)}\left(\frac{x - T_l(P)}{T_s(P)}\right) dP(x). \tag{58}$$

Simulations suggest that the covering probabilities of the confidence interval (53) are good for sample sizes of 20 or more as long as the distribution P is almost symmetric. For the sample $x_1, \ldots, x_n$ this leads to the interval

$$[T_l(P_n) - z((1+\alpha)/2)\Sigma(P_n)/\sqrt{n}, T_l(P_n) + z((1+\alpha)/2)\Sigma(P_n)/\sqrt{n}] \tag{59}$$

with $\Sigma(P)$ given by (52) and $I(x, T_l, P)$ by (54). Similar intervals can be obtained for the variations on M-functionals discussed in Section 2.3.

2.6 Efficiency and bias

The precision of the functional T at the distribution P can be quantified by the length $2z((1+\alpha)/2)\Sigma(P)/\sqrt{n}$ of the asymptotic confidence interval (51).

As the only quantity which depends on T is $\Sigma(P)$ we see that an increase in precision is equivalent to reducing the size of $\Sigma(P)$. The question which naturally arises is then that of determining how small $\Sigma(P)$ can be made. A statistical functional which attains this lower bound is asymptotically optimal and if we denote this lower bound by $\Sigma_{opt}(P)$, the efficiency of the functional T can be defined as $\Sigma_{opt}(P)^2/\Sigma(P)^2$. The efficiency depends on P and we must now decide which P or indeed Ps to choose. The arguments given in Section 1.2 suggest choosing a P which maximizes $\Sigma_{opt}(P)$ over a class of models. This holds for the normal distribution which maximizes $\Sigma_{opt}(P)$ over the class of all distributions with a given variance. For this reason and for simplicity and familiarity we shall take the normal distribution as the reference distribution. If a reference distribution is required which also produces outliers then the slash distribution is to be preferred to the Cauchy distribution. We refer to Cohen (1991) and the discussion given there.

If we consider the M-functionals defined by (24) and (25) the efficiency at the normal distribution is an increasing function of the tuning parameter c. As the breakdown point is a decreasing function of c this would seem to indicate that there is a conflict between efficiency and breakdown point. This is the case for the M-functional defined by (24) and (25) and is due to the linking of the location and scale parts of the functional. If this is severed by, for example, recalculating a location functional as in (26) then there is no longer a conflict between efficiency and breakdown. As however the efficiency of the location functional increases the more it behaves like the mean with a corresponding increase in the bias function of (35) and (37). The conflict between efficiency and bias is a real one and gives rise to an optimality criterion, namely that of minimizing the bias subject to a lower bound on the efficiency. We refer to Martin and Zamar (1993a).

2.7 Outliers in $\mathbb{R}$

One of the main uses of robust functionals is the labelling of so called outliers (see Barnett and Lewis, 1994, Hawkins, 1980, Atkinson, 1994, Gather, 1990, Gather et al., 2003, and Simonoff, 1984, 1987). In the data of Table 1 the laboratories 1 and 3 are clearly outliers which should be flagged. The discussion in Section 1.1 already indicates that the mean and standard deviation are not appropriate tools for the identification of outliers as they themselves are so strongly influenced by the very outliers they are intended to identify. We now demonstrate this more precisely. One simple rule is to classify all observations more than three standard deviations from the mean as outliers. A simple calculation shows that this rule will fail to identify 10% arbitrarily large outliers with the same sign. More generally if all observations more than λ standard deviations from the mean are classified as outliers then this rule will fail to identify a proportion of $1/(1+\lambda^2)$ outliers with the same sign. This is known as the masking effect (Pearson and Chandra Sekar, 1936) where the outliers mask their presence by distorting the mean and, more importantly,

the standard deviation to such an extent as to render them useless for the detection of the outliers. One possibility is to choose a small value of λ but clearly if λ is too small then some non-outliers will be declared as outliers. In many cases the main body of the data can be well approximated by a normal distribution so we now investigate the choice of λ for samples of i.i.d. normal random variables. One possibility is to choose λ dependent on the sample size n so that with probability say 0.95 no observation will be flagged as an outlier. This leads to a value of λ of about $\sqrt{2\log(n)}$ (Davies and Gather, 1993) and the largest proportion of one-sided outliers which can be detected is approximately $1/(1+2\log(n))$ which tends to zero with n. It follows that there is no choice of λ which can detect say 10% outliers and at the same time not falsely flag non-outliers. In order to achieve this the mean and standard deviation must be replaced by functionals which are less effected by the outliers. In particular these functionals should be locally bounded (11). Considerations of asymptotic normality or efficiency are of little relevance here. Two obvious candidates are the median and MAD and if we use them instead of the mean and standard deviation we are led to the identification rule (Hampel, 1985) of the form

$$|x_i - \mathrm{MED}(\boldsymbol{x}_n)| \geq \lambda \mathrm{MAD}(\boldsymbol{x}_n). \tag{60}$$

Hampel (1975) proposed setting $\lambda = 5.2$ as a general all purpose value. The concept of an outlier cannot in practice be very precise but in order to compare different identification rules we require a precise definition and a precise measure of performance. To do this we shall restrict attention to the normal model as one which is often reasonable for the main body of data. In other situations such as waiting times the exponential distribution may be more appropriate. The following is based on Davies and Gather (1993). To define an outlier we introduce the concept of an α- outlier. For the normal distribution $N(\mu, \sigma^2)$ and $\alpha \in (0,1)$ we define the α-outlier region by

$$out(\alpha, N(\mu, \sigma^2)) \;=\; \{x \in \mathbb{R}\colon\; |x - \mu| > \sigma\, z_{1-\alpha/2}\}. \tag{61}$$

which is just the union of the lower and the upper $\alpha/2$-tail regions. Here $z_{1-\alpha/2}$ denotes the $1-\alpha/2$-quantile of the standard normal distribution. For the exponential distribution $Exp(\lambda)$ with parameter λ we set

$$out(\alpha, Exp(\lambda)) \;=\; \{x \in \mathbb{R}\colon\; x > -\lambda \ln \alpha\} \tag{62}$$

which is the upper α-tail region (Gather and Schultze, 1999). The extension to other distributions P is clear. Each point located in the outlier region is called an α-outlier, otherwise it is called an α-inlier. This definition of an outlier refers only to its position in relation to the statistical model for the good data. No assumptions are made concerning the distribution of these outliers or the mechanism by which they are generated.

We can now formulate the task of outlier identification for the normal distribution as follows: For a given sample $\boldsymbol{x}_n = (x_1, \ldots, x_n)$ which contains

at least $[n/2]+1$ i.i.d. observations distributed according to $N(\mu, \sigma^2)$, we have to find all those x_i that are located in $out(\alpha, N(\mu, \sigma^2))$. The level α can be chosen to be dependent on the sample size. If for some $\tilde{\alpha} \in (0, 1)$ we set

$$\alpha \;=\; \alpha_n \;=\; 1 - (1 - \tilde{\alpha})^{1/n}, \tag{63}$$

then the probability of finding at least one observation of a $N(\mu, \sigma^2)-$sample of size n within $out(\alpha_n, N(\mu, \sigma^2))$ is not larger than $\tilde{\alpha}$. Consider now the general Hampel identifier which classifies all observations x_i in

$$OR^H(\mathbf{x}_n, \alpha_n) \;=\; \{x \in \mathbb{R}\colon\; |x - Med(\mathbf{x}_n)| \;>\; g_n(\alpha_n)\; MAD(\mathbf{x}_n)\} \tag{64}$$

as outliers. The region $OR^H(\mathbf{x}_n, \alpha_n)$ may be regarded as an empirical version of the outlier region $out(\alpha_n, N(\mu, \sigma^2))$. The constant $g_n(\alpha_n)$ standardizes the behaviour of the procedure for i.i.d. normal samples which may be done in several ways. One is to determine the constant so that with probability at least $1 - \tilde{\alpha}$ no observation X_i is identified as an outlier, that is

$$P\big(X_i \notin OR(\mathbf{X}_n, \alpha_n),\; i = 1, \ldots, n\big) \;\geq\; 1 - \tilde{\alpha}. \tag{65}$$

A second possibility is to require that

$$P\big(OR(\mathbf{X}_n, \alpha_n) \subset out(\alpha_n, P)\big) \;\geq\; 1 - \tilde{\alpha}. \tag{66}$$

If we use (65) and set $\tilde{\alpha} = 0.05$ then for $n = 20, 50$ and 100 simulations give $g_n(\alpha_n) = 5.82,\; 5.53$ and 5.52 respectively. For $n > 10$ the normalizing constants $g_n(\alpha_n)$ can also be approximated according to the equations given in Section 5 of Gather (1990).
To describe the worst case behaviour of an outlier identifier we can look at the largest nonidentifiable outlier, which it allows. From Davies and Gather (1993) we report some values of this quantity for the Hampel identifier (HAMP) and contrast them with the corresponding values of a sophisticated high breakdown point outwards testing identifier (ROS), based on the non-robust mean and standard deviation (Rosner, 1975; Tietjen and Moore, 1972). Both identifiers are standardized by (65) with $\tilde{\alpha} = 0.05$. Outliers are then observations with absolute values greater than $3.016(n = 20)$, $3.284(n = 50)$ and $3.474(n = 100)$. For $k = 2$ outliers and $n = 20$ the average sizes of the largest non-detected outlier are 6.68 (HAMP) and 8.77 (ROS), for $k = 5$ outliers and $n = 50$ the corresponding values are 4.64 (HAMP) and 5.91 (ROS) and finally for $k = 15$ outliers and $n = 100$ the values are 5.07 (HAMP) and 9.29 (ROS).

3 Location and scale in $\mathbb{R}^k$

3.1 Equivariance and metrics

In Section 2.1 we discussed the equivariance of estimators for location and scale with respect to the affine group of transformations on $\mathbb{R}$. This carries

over to higher dimensions although here the requirement of affine equivariance lacks immediate plausibility. A change of location and scale for each individual component in $\mathbb{R}^k$ is represented by an affine transformation of the form $\Lambda(x) + b$ where Λ is a diagonal matrix. A general affine transformation forms linear combinations of the individual components which goes beyond arguments based on units of measurement. The use of affine equivariance reduces to the almost empirical question as to whether the data, regarded as a cloud of points in $\mathbb{R}^k$, can be well represented by an ellipsoid. If this is the case as it often is then consideration of linear combinations of different components makes data analytical sense. With this proviso in mind we consider the affine group $\boldsymbol{\mathcal{A}}$ of transformations of $\mathbb{R}^k$ into itself,

$$\boldsymbol{\mathcal{A}} = \{\mathcal{A} : \mathcal{A}(x) = A(x) + b\} \tag{67}$$

where A is a non-singular $k \times k$-matrix and b is an arbitrary point in $\mathbb{R}^k$. Let $\mathcal{P}'_k$ denote a family of distributions over $\mathbb{R}^k$ which is closed under affine transformations

$$P \in \mathcal{P}'_k \Rightarrow P^{\mathcal{A}} \in \mathcal{P}'_k, \text{ for all } \mathcal{A} \in \boldsymbol{\mathcal{A}}. \tag{68}$$

A function $T_l : \mathcal{P}'_k \to \mathbb{R}^k$ is called a location functional if it is well defined and

$$T_l(P^{\mathcal{A}}) = \mathcal{A}(T_l(P)), \text{ for all } \mathcal{A} \in \boldsymbol{\mathcal{A}},\ P \in \mathcal{P}'_k. \tag{69}$$

A functional $T_s : \mathcal{P}'_k \to \boldsymbol{\Sigma}_k$ where $\boldsymbol{\Sigma}_k$ denotes the set of all strictly positive definite symmetric $k \times k$ matrices is called a scale or scatter functional if

$$T_s(P^{\mathcal{A}}) = AT_l(P)A^{\top}, \text{ for all } \mathcal{A} \in \boldsymbol{\mathcal{A}},\ P \in \mathcal{P}'_k \text{ with } \mathcal{A}(x) = A(x) + b. \tag{70}$$

The requirement of affine equivariance is a strong one as we now indicate. The most obvious way of defining the median of a k-dimensional data set is to define it by the medians of the individual components. With this definition the median is equivariant with respect to transformations of the form $\Lambda(x)+b$ with Λ a diagonal matrix but it is not equivariant for the affine group. A second possibility is to define the median of a distribution P by

$$\text{MED}(P) = \text{argmin}_{\mu} \int (\|x - \mu\| - \|x\|)\, dP(x).$$

With this definition the median is equivariant with respect to transformations of the form $x \to O(x) + b$ with O an orthogonal matrix but not with respect to the affine group or the group $x \to \Lambda(x) + b$ with Λ a diagonal matrix. The conclusion is that there is no canonical extension of the median to higher dimensions which is equivariant with respect to the affine group.

In Section 2 use was made of metrics on the space of probability distributions on $\mathbb{R}$. We extend this to $\mathbb{R}^k$ where all metrics we consider are of the form

$$d_{\mathcal{C}}(P,Q) = \sup_{C\in\mathcal{C}} |P(C) - Q(C)| \tag{71}$$

where $\mathcal{C}$ is a so called Vapnik-Cervonenkis class (see for example Pollard (1984)). (see for example Pollard, 1984). The class $\mathcal{C}$ can be chosen to suit the problem. Examples are the class of all lower dimensional hyperplanes

$$\mathcal{H} = \{H : H \text{ lower dimensional hyperplane}\} \tag{72}$$

and the class of all ellipsoids

$$\mathcal{E} = \{E : E \text{ an ellipsoid}\}. \tag{73}$$

These give rise to the metrics $d_{\mathcal{H}}$ and $d_{\mathcal{E}}$ respectively. Just as in $\mathbb{R}$ metrics $d_{\mathcal{C}}$ of the form (71) allow direct comparisons between empirical measures and models. We have

$$d_{\mathcal{C}}(P_n(P), P) = \mathrm{O}(1/\sqrt{n}) \tag{74}$$

uniformly in P (see Pollard, 1984).

3.2 M-estimators of location and scale

Given the usefulness of M-estimators for one dimensional data it seems natural to extend the concept to higher dimensions. We follow Maronna (1976). For any positive definite symmetric $k \times k$-matrix Σ we define the metric $d(\cdot,\cdot : \Sigma)$ by

$$d(x, y : \Sigma)^2 = (x-y)^\top \Sigma^{-1}(x-y), \quad x,\, y \in \mathbb{R}^k.$$

Further, let u_1 and u_2 be two non-negative continuous functions defined on $\mathbb{R}_+$ and be such that $su_i(s), s \in \mathbb{R}_+$, $i = 1, 2$ are both bounded. For a given probability distribution P on the Borel sets of $\mathbb{R}^k$ we consider in analogy to (21) and (22) the two equations in μ and Σ

$$\int (x-\mu) u_1(d(x,\mu;\Sigma))\, dP = 0 \tag{75}$$

$$\int u_2(d(x,\mu;\Sigma)^2)(x-\mu)(x-\mu)^\top\, dP = 0. \tag{76}$$

Assuming that at least one solution (μ, Σ) exists we denote it by $T_M(P) = (\mu, \Sigma)$. The existence of a solution of (75) and (76) can be shown under weak conditions as follows. If we define

$$\Delta(P) = \max\{P(H) : H \in \mathcal{H}\} \tag{77}$$

with $\mathcal{H}$ as in (73) then a solution exists if $\Delta(P) < 1 - \delta$ where δ depends only on the functions u_1 and u_2 (Maronna, 1976). Unfortunately the problem of uniqueness is much more difficult than in the one-dimensional case. The conditions placed on P in Maronna (1976) are either that it has a density

$f_P(x)$ which is a decreasing function of $\|x\|$ or that it is symmetric $P(B) = P(-B)$ for every Borel set B. Such conditions do not hold for real data sets which puts us in an awkward position. Furthermore without existence and uniqueness there can be no results on asymptotic normality and consequently no results on confidence intervals. The situation is unsatisfactory so we now turn to the one class of M-functionals for which existence and uniqueness can be shown. The following is based on Kent and Tyler (1991) and is the multidimensional generalization of (33). The k-dimensional t-distribution with density $f_{k,\nu}(\cdot : \mu, \Sigma)$ is defined by

$$f_{k,\nu}(x : \mu, \Sigma) = \frac{\Gamma(\frac{1}{2}(\nu + k))}{(\nu k)^{k/2}\Gamma(\frac{1}{2}\nu)}|\Sigma|^{-\frac{1}{2}}\left(1 + \frac{1}{\nu}(x-\mu)^{top}\Sigma^{-1}(x-\mu)\right)^{-\frac{1}{2}(\nu+k)} \tag{78}$$

and we consider the minimization problem

$$T_M(p) = (T_l(P), T_s(P)) = \text{argmin}_{\mu,\Sigma}\int f_{k,\nu}(x : \mu, \Sigma)\, dP(x) + \frac{1}{2}\log(|\Sigma|) \tag{79}$$

where $|\Sigma|$ denotes the determinant of the positive definite matrix Σ. For any distribution P on the Borel sets of $\mathbb{R}^k$ we define $\Delta(P)$ which is the k-dimensional version of (23). It can be shown that if $\Delta(P) < 1/2$ then (79) has a unique solution. Moreover for data sets there is a simple algorithm which converges to the solution. On differentiating the right hand side of (79) it is seen that the solution is an M-estimator as in (75) and (76). Although this has not been proven explicitly it seems clear that the solution will be locally uniformly Fréchet differentiable, that is, it will satisfy (12) where the influence function $I(x, T_M, P)$ can be obtained as in (54) and the metric d_{ko} is replaced by the metric $d_{\mathcal{H}}$. This together with (74) leads to uniform asymptotic normality and allows the construction of confidence regions. The only weakness of the proposal is the low gross error breakdown point $\varepsilon^*(T_M, P, GE)$ defined below which is at most $1/(k+1)$. This upper bound is shared with the M-functionals defined by (75) and (76) (Maronna, 1976). The problem of constructing high breakdown functionals in k dimensions will be discussed below.

3.3 Bias and breakdown

The concepts of bias and breakdown developed in Section 2.4 carry over to higher dimensions. Given a metric d on the space of distributions on $\mathbb{R}^k$ and a location functional T_l we follow (37) and define

$$b(T_l, P, \varepsilon, d) = \sup\{\|T_l(Q)\| : d(P, Q) < \varepsilon\} \tag{80}$$

and

$$b(T_l, P, \varepsilon, GE) = \sup\{\|T_l(Q)\| : Q = (1-\varepsilon)P + \varepsilon G,\ G \in \mathcal{P}\} \tag{81}$$

where by convention $\|T_l(Q)\| = \infty$ if T_l is not defined at Q. The extension to scale functionals is not so obvious as there is no canonical definition of bias. We require a measure of difference between two positive definite symmetric matrices. For reasons of simplicity and because it is sufficient for our purposes the one we take is $|\log(|\Sigma_1|/|\Sigma_2|)|$. Corresponding to (36) we define

$$b(T_s, P, \varepsilon, d) = \sup\{|\log(|T_s(Q)|/|T_s(P)|)| : d(P, Q) < \varepsilon\} \tag{82}$$

and

$$b(T_s, P, \varepsilon, GE) = \sup\{|\log(|T_s(Q)|/|T_s(P)|)| : Q = (1-\varepsilon)P + \varepsilon G,\ G \in \mathcal{P}\}. \tag{83}$$

Most work is done using the gross error model (81)and (83). The breakdown points of T_l are defined by

$$\varepsilon^*(T_l, P, d) = \sup\{\varepsilon : b(T_l, P, \varepsilon, d) < \infty\} \tag{84}$$

$$\varepsilon^*(T_l, P, GE) = \sup\{\varepsilon : b(T_l, P, \varepsilon, GE) < \infty\} \tag{85}$$

$$\varepsilon^*(T_l, P_n, fsbp) = \max\{k/n : |T_l(P_n^k)| < \infty\} \tag{86}$$

where (86) corresponds in the obvious manner to (41). The breakdown points for the scale functional T_s are defined analogously using the bias functional (82). We have

Theorem 4. *For any translation equivariant functional T_l*

$$\varepsilon^*(T_l, P, d_{\mathcal{H}}) \leq 1/2 \text{ and } \varepsilon^*(T_l, P_n, fsbp) \leq \lfloor n/2 \rfloor / n \tag{87}$$

and for any affine equivariant scale functional

$$\varepsilon^*(T_s, P, d_{\mathcal{E}}) \leq (1 - \Delta(P))/2 \text{ and } \varepsilon^*(T_s, P_n, fsbp) \leq (1 - \Delta(P_n))/2. \tag{88}$$

In Section 2.4 it was shown that the M-estimators of Section 2.3 can attain or almost attain the upper bounds of Theorem 1. Unfortunately this is not the case in k dimensions where as we have already mentioned the breakdown points of M-functionals of Section 3.2 are at most $1/(k+1)$. In recent years much research activity has been directed towards finding high breakdown affinely equivariant location and scale functionals which attain or nearly attain the upper bounds of Theorem 4. This is discussed in the next section.

3.4 High breakdown location and scale functionals in $\mathbb{R}^k$

The first high breakdown affine equivariant location and scale functionals were proposed independently of each other by Stahel (1981) and Donoho (1982). They were defined for empirical data but the construction can be carried over to measures satisfying a certain weak condition. The idea is to project the data points onto lines through the origin and then to determine which points

are outliers with respect to this projection using one-dimensional functions with a high breakdown point. More precisely we set

$$o(x_i, \theta) = |x_i^\top \theta - \mathrm{MED}(x_1^\top \theta, \ldots, x_n^\top \theta)| / \mathrm{MAD}(x_1^\top \theta, \ldots, x_n^\top \theta) \tag{89}$$

and

$$o(x_i) = \sup\{o(x_i, \theta) : \|\theta\| = 1\}. \tag{90}$$

This is a measure for the outlyingness of the point x_i and it may be checked that it is affine invariant. Location and scale functionals may now be obtained by taking for example the mean and the covariance matrix of those $\lfloor n/2 + 1 \rfloor$ observations with the smallest outlyingness measure. Although (90) requires a supremum over all values of θ this can be reduced for empirical distributions as follows. Choose all linearly independent subsets $x_{i_1}, \ldots, x_{i_k}$ of size k and for each such subset determine a θ which is orthogonal to their span. If the sup in (90) is replaced by a maximum over all such θ then the location and scale functionals remain affine equivariant and retain the high breakdown point. Although this requires the consideration of only a finite number of directions namely at most $\binom{n}{k}$ this number is too large to make it a practicable possibility even for small values of n and k. The problem of calculability has remained with high breakdown methods ever since and it is their main weakness. There are still no high breakdown affine equivariant functionals which can be calculated exactly except for very small data sets. Huber (1995) goes as far as to say that the problem of calculability is the breakdown of high breakdown methods. This is perhaps too pessimistic but the problem remains unsolved.

Rousseeuw (1985) introduced two further high breakdown location and scale functionals as follows. The first, the so called minimum volume ellipsoid (MVE) functional, is a multidimensional version of Tukey's shortest half-sample (8) and is defined as follows. We set

$$E = \mathrm{argmin}_{\tilde{E}}\{|\tilde{E}| : |\{i : x_i \in \tilde{E}\}| \geq \lfloor n/2 \rfloor\} \tag{91}$$

where $|E|$ denotes the volume of E and $|\{\,\}|$ denotes the number of elements of the set $\{\,\}$. In other words E has the smallest volume of any ellipsoid which contains more than half the data points. For a general distribution P we define

$$E(P) = \mathrm{argmin}_{\tilde{E}}\{|\tilde{E}| : \int_{\tilde{E}} dP \geq 1/2\}. \tag{92}$$

Given E the location functional $T_l(P)$ is defined to be the centre $\mu(E)$ of E and the covariance functional $T_s(P)$ is taken to be $c(k)\Sigma(E)$ where

$$E = \{x : (x - \mu(E))^\top \Sigma^{-1}(x - \mu(E)) \leq 1\}. \tag{93}$$

The factor $c(k)$ can be chosen so that $c(k)\Sigma(E) = I_k$ for the standard normal distribution in k dimensions.

The second functional is based on the so called minimum covariance determinant (MCD) and is as follows. We write

$$\mu(B) = \int_B x\, dP(x)/P(B) \tag{94}$$

$$\Sigma(B) = \int_B (x - \mu(B))(x - \mu(B))^\top\, dP(x)/P(B) \tag{95}$$

and define

$$\mathrm{MCD}(P) = \mathrm{argmin}_B \{|\Sigma(B)| : P(B) \geq 1/2\} \tag{96}$$

where $|\Sigma(B)|$ is defined to be infinite if either of (94) or (95) does not exist. The location functional is taken to be $\mu(\mathrm{MCD}(B))$ and the scatter functional $c(k)\Sigma(\mathrm{MCD}(B))$ where again $c(k)$ is usually chosen so that $c(k)\Sigma(\mathrm{MCD}(B)) = I_k$ for the standard normal distribution in k-dimensions. It can be shown that both these functionals are affinely equivariant.

A smoothed version of the minimum volume estimator can be obtained by considering the minimization problem

$$\text{minimize } |\Sigma| \text{ subject to } \int \rho\left((x-\mu)^\top \Sigma^{-1}(x-\mu)\right)\, dP(x) \geq 1/2 \tag{97}$$

where $\rho : \mathbb{R}_+ \to [0, 1]$ satisfies $\rho(0) = 1, \lim_{x\to\infty} \rho(x) = 0$ and is continuous on the right (see Davies, 1987). This gives rise to the class of so called S-functionals. The minimum volume estimator can be obtained by specializing to the case $\rho(x) = \{0 \leq x < 1\}$.

On differentiating (97) it can be seen that an S-functional can be regarded as an M-functional but with redescending functions u_1 and u_2 in contrast to the conditions placed on u_1 and u_2 in (75) and (76) (Lopuhaä, 1989). For such functions the defining equations for an M-estimator have many solutions and the minimization problem of (97) can be viewed as a choice function. Other choice functions can be made giving rise to different high breakdown M-estimators. We refer to Lopuhaä (1991) and Kent and Tyler (1996). A further class of location and scatter functionals have been developed from Tukey's concept of depth (Tukey, 1975). We refer to Donoho and Gasko (1992), Liu at al. (1999) and Zuo and Serfling (2000a,b). Many of the above functionals have breakdown points close to or equal to the upper bound of Theorem 4 . For the calculation of breakdown points we refer to Davies (1987), Lopuhaä and Rousseeuw (1991), Donoho and Gasko (1992), Davies (1993) and Tyler (1994).

The problem of determining a functional which minimizes the bias over a neighbourhood was considered in the one-dimensional case in Section 2.4. The problem is much more difficult in $\mathbb{R}^k$ but some work in this direction has been done (see Adrover, 1998). The more tractable problem of determining the size of the bias function for particular functionals or classes of functionals has also been considered (Yohai and Maronna, 1990; Maronna et al., 1992).

All the above functionals can be shown to exist but there are problems concerning the uniqueness of the functional. Just as in the case of Tukey's shortest half (8) restrictions must be placed on the distribution P which generally include the existence of a density with given properties (see Davies, 1987 and Tatsuoka and Tyler, 2000) and which is therefore at odds with the spirit of robust statistics. Moreover even uniqueness and asymptotic normality at some small class of models are not sufficient. Ideally the functional should exist and be uniquely defined and locally uniformly Fréchet differentiable just as in Section 2.5. It is not easy to construct affinely equivariant location and scatter functionals which satisfy the first two conditions but it has been accomplished by Dietel (1993) using the Stahel-Donoho idea of projections described above. To go further and define functionals which are also locally uniformly Fréchet differentiable with respect to some metric $d_{\mathcal{C}}$ just as in the one-dimensional case considered in Section 2.5 is a very difficult problem. The only result in this direction is again due to Dietel (1993) who managed to construct functionals which are locally uniformly Lipschitz. The lack of locally uniform Fréchet differentiability means that all derived confidence intervals will exhibit a certain degree of instability. Moreover the problem is compounded by the inability to calculate the functionals themselves. To some extent it is possible to reduce the instability by say using the MCD functional in preference to the MVE functional, by reweighting the observations or by calculating a one-step M-functional as in (29) (see Davies, 1992a). However the problem remains and for this reason we do not discuss the research which has been carried out on the efficiency of the location and scatter functionals mentioned above. Their main use is in data analysis where they are an invaluable tool for detecting outliers. This will be discussed in the following section.

A scatter matrix plays an important role in many statistical techniques such as principal component analysis and factor analysis. The use of robust scatter functionals in some of these areas has been studied by among others Croux and Haesbroeck (2000), Croux and Dehon (2001) and Willems et al. (2002).

As already mentioned the major weakness of all known high breakdown functionals is their computational complexity. For the MCD functional an exact algorithm of the order of $n^{k(k+3)/2}$ exists and there are reasons for supposing that this cannot be reduced to below n^k (Bernholt and Fischer, 2001). This means that in practice for all but very small data sets heuristic algorithms have to be used. We refer to Rousseeuw and Van Driesen (1999) for a heuristic algorithm for the MCD-functional.

3.5 Outliers in $\mathbb{R}^k$

Whereas for univariate, bivariate and even trivariate data outliers may often be found by visual inspection, this is not practical in higher dimensions (Caroni and Prescott, 1992; Hadi, 1994; Barme-Delcroix and Gather, 2000; Gnanadesikan and Kettenring, 1972; Hadi and Simonoff, 1997). This makes

it all the more important to have methods which automatically detect high dimensional outliers. Much of the analysis of the one-dimensional problem given in Section 2.7 carries over to the k-dimensional problem. In particular outlier identification rules based on the mean and covariance of the data suffer from masking problems and must be replaced by high breakdown functionals (see also Rocke and Woodruff, 1996, 1997). We restrict attention to affine equivariant functionals so that an affine transformation of the data will not alter the observations which are identified as outliers. The identification rules we consider are of the form

$$(x_i - T_l(P_n))^\top T_s(P_n)^{-1}(x_i - T_l(P_n)) \geq c(k,n) \tag{98}$$

where P_n is the empirical measure, T_l and T_s are affine equivariant location and scatter functionals respectively and $c(k,n)$ is a constant to be determined. This rule is the k-dimensional counterpart of (60). In order to specify some reasonable value for $c(k,n)$ and in order to be able to compare different outlier identifiers we require, just as in Section 2.7, a precise definition of an outlier and a basic model for the majority of the observations. As our basic model we take the k-dimensional normal distribution $\mathcal{N}(\mu, \Sigma)$. The definition of an α_n-outlier corresponds to (62) and is

$$out(\alpha_n, \mu, \Sigma) = \left\{x \in \mathbb{R}^k : (x-\mu)^\top \Sigma^{-1}(x-\mu) > \chi^2_{k;1-\alpha_n}\right\}, \tag{99}$$

where $\alpha_n = 1 - (1-\tilde{\alpha})^{1/n}$ for some given value of $\tilde{\alpha} \in (0,1)$. Clearly for an i.i.d. sample of size n distributed according to $\mathcal{N}(\mu, \Sigma)$ the probability that no observation lies in the outlier region of (99) is just $1-\alpha$. Given location and scale functionals T_l and T_s and a sample $\tilde{x}_n$ we write

$$OR^H(\tilde{x}_n, \alpha_n) = \left\{x \in \mathbb{R}^k : (x - T_l(P_n))^\top T_s(P_n)^{-1}(x - T_l(P_n)) \geq c(k,n,\alpha_n)\right\} \tag{100}$$

which corresponds to (64). The region $OR^H(\tilde{x}_n, \alpha_n)$ is the empirical counterpart of $out(\alpha_n, \mu, \Sigma)$ of (99) and any observation lying in $OR^H(\tilde{x}_n, \alpha_n)$ will be identified as an outlier. Just as in the one-dimensional case we determine the $c(k,n,\alpha_n)$ by requiring that with probability $1-\tilde{\alpha}$ no observation is identified as an outlier in i.i.d. $\mathcal{N}(\mu, \Sigma)$ samples of size n. This can be done by simulations with appropriate asymptotic approximations for large n. The simulations will of course be based on the algorithms used to calculate the functionals and will not be based on the exact functionals assuming these to be well defined. For the purpose of outlier identification this will not be of great consequence. We give results for three multivariate outlier identifiers based on the MVE- and MCD-functionals of Rousseeuw (1985) and the S-functional based on Tukey's biweight function as given in Rocke (1996). There are good heuristic algorithms for calculating these functionals at least approximately (Rocke, 1996; Rousseeuw and Van Driesen, 1999; Rousseeuw and van Zoomeren, 1990). The following is based on Becker and Gather (2001). Table 2 gives the values of $c(k,n,\alpha_n)$ with $\alpha = 0.1$ The results are based on 10 000 simulations for each combination of k and n.

Table 2. Normalizing constants $c(k, n, \alpha_n)$ for $\mathbf{OR}_{MVE}, \mathbf{OR}_{MCD}, \mathbf{OR}_{BW}$ for $\alpha = 0.1$

n	k	c_{MVE}	c_{MCD}	c_{BW}
20	2	19.14222	85.58786	21.35944
20	3	23.47072	167.61310	26.87044
20	4	33.72110	388.84680	33.17018
50	2	17.54896	28.51695	16.93195
50	3	20.61580	41.83594	19.78682
50	4	24.65417	64.18462	23.14061

Becker and Gather (2001) show by simulations that although none of the above rules fails to detect arbitrarily large outliers it still can happen that very extreme observations are not identified as outliers. To quantify this we consider the identifier OR_{MVE} and the constellation $n = 50, k = 2$ with $m = 5$ observations replaced by other values. The mean norm of the most extreme nonidentifiable outlier is 4.17. The situation clearly becomes worse with an increasing proportion of replaced observations and with the dimension k (see Becker and Gather, 1999). If we use the mean of the norm of the most extreme non-identifiable outlier as a criterion then none of the three rules dominates the others although the biweight identifier performs reasonably well in all cases and is our preferred choice.

4 Linear regression

4.1 Equivariance and metrics

The linear regression model may be written in the form

$$Y_i = \boldsymbol{x}_i^\top \beta + \varepsilon_i, \quad i = 1, \ldots, n \tag{101}$$

where $\boldsymbol{x}_i$, $i = 1 \ldots, n$ and $\beta \in \mathbb{R}^k$. The assumptions of the standard model are that the x_i are fixed and that the ε_i are i.i.d. random variables with the default distribution being the normal distribution $N(0, \sigma^2)$. There are of course many other models in the literature including random x_i-values and a covariance structure for the errors ε_i. For the purpose of robust regression we consider probability distributions P on $\mathbb{R}^{k+1}$ where the first k components refer to the covariates $\boldsymbol{x}$ and the last component is the corresponding value of y. We restrict attention to the family $\mathcal{P}_{k+1}$ of probability measures given by

$$\mathcal{P}_{k+1} = \{P : P(H \times \mathbb{R}) < 1 \text{ for all lower dimensional subspaces } H \subset \mathbb{R}^k\}. \tag{102}$$

The metric we use on $\mathcal{P}_{k+1}$ is $d_{\mathcal{H}}$ with $\mathcal{H}$ given by (73).

Consider the regression group G of transformations $g : \mathbb{R}^{k+1} \to \mathbb{R}^{k+1}$ of the form

$$g(\boldsymbol{x}, y) = (A(\boldsymbol{x}), sy + \boldsymbol{x}^\top \gamma) \tag{103}$$

where A is a non-singular $k \times k$-matrix, $s \in \mathbb{R}$, $s \neq 0$, and $\gamma \in \mathbb{R}^k$. A functional $T : \mathcal{P}_{k+1} \to \mathbb{R}^k \times \mathbb{R}_+$ is called a regression functional if for all $g \in G$ and $P \in \mathcal{P}_{k+1}$

$$T(P^g) = h_g(T(P)) \tag{104}$$

where

$$h_g(\beta, \sigma) = (s(A^{-1})^\top(\beta + \gamma), s\sigma). \tag{105}$$

with A and γ as in (103). The first k components of $T(P)$ specify the value of $\beta \in \mathbb{R}^k$ and the last component that of σ. The restriction to models $P \in \mathcal{P}_{k+1}$ of (102) is that without such a restriction there is no uniquely defined value of β.

4.2 M-estimators for regression

Given a distribution $P \in \mathcal{P}_{k+1}$ we define an M-functional by $T(P) = (\beta^*, \sigma^*)$ where (β^*, σ^*) is a solution of the equations

$$\int \phi(\boldsymbol{x}, (y - \boldsymbol{x}^\top \beta)/\sigma)\boldsymbol{x}\, dP(\boldsymbol{x}, y) = 0 \tag{106}$$

$$\int \chi((y - \boldsymbol{x}^\top \beta)/\sigma)\, dP(\boldsymbol{x}, y) = 0 \tag{107}$$

for given functions $\phi : \mathbb{R}^{k+1} \to \mathbb{R}$ and $\chi : \mathbb{R} \to \mathbb{R}$. Just as in Section 3.2 for M-functionals of location and scatter there are problems concerning the existence and uniqueness. Maronna and Yohai (1981) give sufficient conditions for existence which depend only on the properties of ϕ and χ and the values of $\sup_\theta\{P(\theta^\top \boldsymbol{x} = 0) : \theta \neq 0\}$ and $\sup_{\alpha,\theta}\{P(\alpha y + \theta^\top \boldsymbol{x} = 0) : |\alpha| + \|\theta\| \neq 0\}$. Uniqueness requires additional strong assumptions such as either symmetry or the existence of a density for the conditional distribution of $y - \theta_0^\top \boldsymbol{x}$ for each fixed $\boldsymbol{x}$. Huber (1981) considers the minimization problem

$$(\beta^*, \sigma^*) = \operatorname{argmin} \left(\int \rho((y - \boldsymbol{x}^\top \beta)/\sigma)\, dP(\boldsymbol{x}, y) + a \right) \sigma \tag{108}$$

where $\rho : \mathbb{R} \to \mathbb{R}_+$ is convex with $\rho(0) = 0$ and $a > 0$. Under appropriate conditions on ρ it can be shown that the solution is unique and that there exists a convergent algorithm to calculate it. On differentiating (108) we obtain (106) and (107) with

$$\phi(\boldsymbol{x}, u) = \rho^{(1)}(u) \text{ and } \chi(u) = u\rho^{(1)}(u) - \rho(u) - a. \tag{109}$$

Even if the solution of (106) and (107) exists and is unique it is not necessarily regression equivariant. To make it so we must introduce a scatter

functional T_Σ on the marginal distributions $P', P'(B) = P(B \times \mathbb{R})$ of the covariate $\boldsymbol{x}$. Such a functional satisfies $T_\Sigma(P'^A) = AT_\Sigma(P')A^\top$ for any non-singular $k \times k$-matrix A and is required not only for equivariance reasons but also to downweight outlying $\boldsymbol{x}$-values or so called leverage points. For this latter purpose the functional T_Σ must also be robust. We now replace (106) by

$$\int \phi(\boldsymbol{x}^\top T_\Sigma(P)^{-1}\boldsymbol{x}, (y - \boldsymbol{x}^\top\beta)/\sigma)\boldsymbol{x}\, dP(\boldsymbol{x}, y) = 0. \tag{110}$$

The resulting functional is now regression equivariant but its analysis is more difficult requiring as it does an analysis of the robustness properties of the scatter functional T_Σ.

Finally we note that in the literature most ϕ functions of (106) are of the form

$$\phi(\boldsymbol{x}, u) = \pi(\boldsymbol{x})\psi(u) \tag{111}$$

and the resulting functionals are known as GM-functionals. We refer to Hampel et al. (1986).

4.3 Bias and Breakdown

Given a regression functional $T_r = (T_b, T_s)$ where T_b refers to the β-components and T_s is the scale part it is usual to define breakdown just by the behaviour of T_b and to neglect T_s. The breakdown point of T_r at the distribution P is defined by

$$\varepsilon^*(T_r, P, d_\mathcal{H}) = \sup\{\varepsilon : b(T_r, P, \varepsilon, d_\mathcal{H}) < \infty\} \tag{112}$$

where

$$b(T_r, P, \varepsilon, d_\mathcal{H}) = \sup\{\|T_b(Q) - T_b(P)\| : d_\mathcal{H}(P, Q) < \varepsilon\} \tag{113}$$

with corresponding definitions for the gross error neighbourhood $\varepsilon^*(T_r, P, GE)$ and for the finite sample breakdown point $\varepsilon^*(T_r, P_n, fsbp)$. To state the next theorem we set

$$\Delta(P) = \sup\{P(H \times \mathbb{R}) : H \text{ a plane in } \mathbb{R}^k \text{ of dimension at most } k-1\}$$

which is the regression equivalent of (77). We have

Theorem 5. *For any regression equivariant functional*

$$\varepsilon^*(T_r, P, d_h) \le (1 - \Delta(P))/2 \text{ and } \varepsilon^*(T_r, P_n, fsbp) \le (1 - \Delta(P_n))/2. \tag{114}$$

If one considers L_1-regression

$$\beta^* = \operatorname{argmin} \sum_{i=1}^{n} |y_i - \boldsymbol{x}_i^\top\beta| \tag{115}$$

it can be shown if one $\boldsymbol{x}_i$ is sufficiently outlying then the residual at this point will be zero and hence the finite sample breakdown point is a disappointing

$1/n$. This turns out to apply to most M-functionals of the last section whose breakdown point is at most $1/(k+1)$ irrespective of their exact definition. The literature on this point is unsatisfactory. Although some M-functionals have been shown to have a positive breakdown point this has only been done under the assumption that the scale part T_s is known. As obtaining the correct magnitude of the scale of the errors is in some sense the most difficult problem in robust regression such results are of limited value. They do not however alter the fact that M-functionals have a disappointing breakdown point. We now turn to the problem of constructing high breakdown regression functionals.

4.4 High breakdown regression functionals

The first high breakdown regression functional was proposed by Hampel (1975) and is as follows.

$$T_{lms}(P) = \text{argmin}_{(\beta,\sigma)} \left\{ \sigma : \int \{|y - \boldsymbol{x}^\top \beta| \leq \sigma\} \, dP(\boldsymbol{x}, y) \geq 1/2 \right\}. \quad (116)$$

The idea goes back to Tukey's shortest half-sample of which it is the regression counter part. It can be shown that it has almost the highest finite sample breakdown point given by Theorem 5. By slightly altering the factor $1/2$ in (116) to take into account the dimension k of the $\boldsymbol{x}$-variables it can attain this bound. Rousseeuw (1984) propagated its use and gave it the name by which it is now known, the least median of squares LMS. Rousseeuw calculated the finite sample breakdown point and provided a first heuristic algorithm which could be applied to real data sets. He also defined a second high breakdown functional known as least trimmed squares LTS defined by

$$T_{lts}(P) = \text{argmin}_{(\beta,\sigma)} \left\{ \int (y - \boldsymbol{x}^\top \beta)^2 \{|y - \boldsymbol{x}^\top \beta| \leq \sigma\} \, dP(\boldsymbol{x}, y) : \int \{|y - \boldsymbol{x}^\top \beta| \leq \sigma\} \, dP(\boldsymbol{x}, y) \geq 1/2 \right\}. \quad (117)$$

There are now many high breakdown regression functionals such as S-functionals (Rousseeuw and Yohai, 1984), MM-functionals (Yohai, 1987), τ-functionals (Yohai and Zamar, 1988), constrained M-functionals (Mendes and Tyler, 1996), rank regression (Chang et al., 1999) and regression depth (Rousseeuw and Hubert, 1999). Just as in the location and scale problem in $\mathbb{R}^k$ statistical functionals can have the same breakdown points but very different bias functions. We refer to Martin et al. (1989), Maronna and Yohai (1993) and Berrendero and Zamar (2001). All these high breakdown functionals either attain or by some minor adjustment can be made to attain the breakdown points of Theorem 5 with the exception of depth based methods where the maximal breakdown point is $1/3$ (see Donoho and Gasko, 1992).

All the above high breakdown regressional functionals can be shown to exist under weak assumptions but just as in the case of high breakdown location

and scatter functionals in $\mathbb{R}^k$ uniqueness can only be shown under very strong conditions which typically involve the existence of a density function for the errors (see Davies, 1993). The comments made about high breakdown location and scale functionals in $\mathbb{R}^k$ apply here. Thus even if a regression functional is well defined at some particular model there will be other models arbitrarily close in the metric $d_{\mathcal{H}}$ where a unique solution does not exist. This points to an inherent local instability of high breakdown regression functionals which has been noted in the literature (Sheather et al., 1997; Ellis, 1998). Dietel (1993) has constructed regression functionals which are well defined at all models P with $\Delta(P) < 1$ and which are locally uniformly Lipschitz, not however locally uniformly Fréchet differentiable. For this reason all confidence regions and efficiency claims must be treated with a degree of caution. An increase in stability can however be attained by using the LTS-functional instead of the LMS-functional, by reweighting the observations or using some form of one-step M-functional improvement as in (29).

Just as with high breakdown location and scatter functionals in $\mathbb{R}^k$ the calculation of high breakdown regression functionals poses considerable difficulties. The first high breakdown regression functional was Hampel's least median of squares and even in the simplest case of a straight line in $\mathbb{R}^2$ the computational cost is of order n^2. The algorithm is by no means simple requiring as it does ideas from computational geometry (see Edelsbrunner and Souvaine, 1990). From this and the fact that the computational complexity increases with dimension it follows that one has to fall back on heuristic algorithms. The one recommended for linear regression is that of Rousseeuw and Van Driesen (1999) for the LTS-functional.

4.5 Outliers

To apply the concept of α-outlier regions to the linear regression model we have to specify the distribution P_Y of the response and the joint distribution $P_{\boldsymbol{X}}$ of the regressors assuming them to be random. For specificness we consider the model

$$P_{Y|\boldsymbol{X}=\boldsymbol{x}} \;=\; N(\boldsymbol{x}^\top\beta, \sigma^2), \tag{118}$$

and

$$P_{\boldsymbol{X}} \;=\; \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{119}$$

Assumption (118) states that the conditional distribution of the response given the regressors is normal and assumption (119) means that the joint distribution of the regressors is a certain p-variate normal distribution. If both assumptions are fulfilled then the joint distribution of $(Y, \boldsymbol{X})$ is a multivariate normal distribution.

We can define outlier regions under model (101) in several reasonable ways. If only (118) is assumed then a response-α-outlier region could be defined as

$$out(\alpha, P_{Y|\boldsymbol{X}=\boldsymbol{x}}) \;=\; \big\{y \in \mathbb{R}\colon\; u = |y - \boldsymbol{x}^\top \beta| \;>\; \sigma\; z_{1-\alpha/2}\big\}, \tag{120}$$

which is appropriate if the regressors are fixed and only outliers in y-direction are to be identified. If the regressors are random, which will be the more frequent case in actuarial or econometric applications, outliers in $\boldsymbol{x}$-direction are important as well. Under assumption (119) a regressor-α-outlier region is a special case of the α-outlier region (99). This approach leads to a population based version of the concept of leverage points. These are the points in a sample $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, from model (101) "for which $\boldsymbol{x}_i$ is far away from the bulk of the $\boldsymbol{x}_i$ in the data" (Rousseeuw and van Zoomeren, 1990).

For the identification of regressor-outliers (leverage points) the same identification rules can be applied as in the multivariate normal situation. For the detection of response-outliers by resistant one-step identifiers, one needs robust estimators of the regression coefficients and the scale σ. Examples of high breakdown estimators that can be used in this context are the Least Trimmed Squares estimator and the corresponding scale estimator (Rousseeuw, 1984; Rousseeuw and Leroy, 1987), S-estimators Rousseeuw and Yohai (1984), MM-estimators (Yohai, 1987) or the REWLS-estimators (Gervini and Yohai, 2002).

5 Analysis of variance

5.1 One-way table

The one-way analysis of variance is concerned with the comparison of the locations of k samples $x_{ij}, j = 1, \ldots, n_i$, $i = 1, \ldots, k$. The term "analysis of variance" goes back to the pioneering work of Fisher (1935) who decomposed the variance of the combined samples as follows

$$\sum_{ij}(x_{ij} - \bar{\boldsymbol{x}})^2 = \sum_i \sum_j (x_{ij} - \bar{\boldsymbol{x}}_i)^2 + \sum_i n_i(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})^2. \tag{121}$$

The first term of (121) is the total sum of squares, the second is the sum of squares within samples and the third is the sum of squares between samples. If the data are modelled as i.i.d. normal random variables with a common variance σ^2 but with the ith sample mean μ_i then it is possible to derive a test for the null hypothesis that the means are equal. The single hypothesis of equal means is rarely of interest in itself. All pairwise comparisons

$$\mu_i = \mu_l, \quad 1 \le i < l \le k,$$

as well as contrasts $\sum_i c_i \mu_i = 0$ may also be of interest and give rise to the problem of multiple testing and the associated difficulties. The use of the L_2-norm as in (121) is widespread perhaps because of the elegant mathematics. The peculiarities of data analysis must however have priority over mathematical theory and as real data sets may contain outliers, be skewed to some

extent and have different scales it becomes clear that an L_2-norm and Gaussian based theory is of limited applicability. We sketch a robustified approach to the one-way table (see Davies, 2004).

As a first step gross outliers are eliminated from each sample using a simplified version of the outlier identification rule based on the median and MAD of the sample. Using the robust location and scale functionals T_l and T_s an α_k confidence or approximation interval I_i for location for the ith sample is calculated. To control the error rate for Gaussian and other samples we set $\alpha_k = \alpha^{1/k}$ with for example $\alpha = 0.95$. This choice guarantees that for Gaussian samples

$$P(\mu_i \in I_i, i = 1, \ldots, k) = \alpha. \tag{122}$$

Simulations show that this holds accurately for other symmetric distributions such as the slash, Cauchy and the double exponential. All questions relating to the locations of the samples are now reduced to questions concerning the intervals. For example, the samples i and l can be approximated by the same location value if and only if $I_i \cap I_l \neq \emptyset$. Similarly if the samples are in some order derived from a covariable it may be of interest as to whether the locations can be taken to be non-decreasing. This will be the case if and only if there exist $a_i, i = 1, \ldots, k$ with $a_1 \leq a_2 \leq \ldots \leq a_k$ and $a_i \in I_i$ for each i. Because of (122) all such questions when stated in terms of the μ_i can be tested simultaneously and on Gaussian test beds the error rate will be $1-\alpha$ regardless of the number of tests. Another advantage of the method is that it allows a graphical representation. Every analysis should include a plot of the boxplots for the k data sets. This can be augmented by the corresponding plot of the intervals I_i which will often look like the boxplots but if the sample sizes differ greatly this will influence the lengths of the intervals but not the form of the boxplots.

5.2 Two-way table

Given IJ samples

$$(x_{ijk})_{k=1}^{n_{ij}}, \quad i = 1, \ldots, I, \, j = 1, \ldots, J$$

the two-way analysis of variance in its simplest version looks for a decomposition of the data of the form

$$x_{ijk} = m + a_i + b_j + c_{ij} + r_{ijk} \tag{123}$$

with the the following interpretation. The overall effect is represented by m, the row and column effects by the a_i and b_j respectively and the interactions by the c_{ij}. The residuals r_{ijk} take care of the rest. As it stands the decomposition (123) is not unique but can be made so by imposing side conditions on the a_i, b_j and the c_{ij}. Typically these are of the form

$$\sum_i a_i = \sum_j b_j = \sum_i c_{ij} = \sum_j c_{ij} = 0 \tag{124}$$

where the latter two hold for all j and i respectively. The conditions (124) are almost always stated as technical conditions required to make the decomposition (123) identifiable. The impression is given that they are neutral with respect to any form of data analysis. But this is not the case as demonstrated by Tukey (1993) and as can be seen by considering the restrictions on the interactions c_{ij}. The minimum number of interactions for which the restrictions hold is four which, in particular, excludes the case of a single interaction in one cell. The restrictions on the row and column effects can also be criticized but we take this no further than mentioning that the restrictions

$$\text{MED}(a_1, \ldots, a_I) = \text{MED}(b_1, \ldots, b_J) = 0 \tag{125}$$

may be more appropriate. The following robustification of the two-way table is based on Terbeck and Davies (1998). The idea is to look for a decomposition which minimizes the number of non-zero interactions. We consider firstly the case of one observation per cell, $n_{ij} = 1$, for all i and j, and look for a decomposition

$$x_{ij} = m + a_i + b_j + c_{ij} \tag{126}$$

with the smallest number of c_{ij} which are non-zero. We denote the positions of the c_{ij} by a $I \times J$-matrix C with $C(i,j) = 1$ if and only if $c_{ij} \neq 0$, the remaining entries being zero. It can be shown that for certain matrices C the non-zero interactions c_{ij} can be recovered whatever their values and, moreover, they are the unique non-zero residuals of the L_1-minimization problem

$$\min_{a_i, b_j} \sum_{ij} |x_{ij} - a_i - b_j|. \tag{127}$$

We call matrices C for which this holds unconditionally identifiable. They can be characterized and two such matrices are

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{128}$$

as well as matrices obtained from any permutations of rows and columns. The above considerations apply to exact models without noise. It can be shown however that the results hold true if noise is added in the sense that for unconditionally identifiable matrices sufficiently large (compared to the noise) interactions c_{ij} can be identified as the large residuals from an L_1-fit. Three further comments are in order. Firstly Tukey's median polish can often identify interactions in the two-way-table. This is because it attempts to

approximate the L_1-solution. At each step the L_1-norm is reduced or at least not increased but unfortunately the median polish may not converge and, even if it does, it may not reach the L_1 solution. Secondly L_1 solutions in the presence of noise are not unique. This can be overcome by approximating the moduls function $|x|$ by a strictly convex function almost linear in the tails. Thirdly, if there is more than one observation per cell it is recommended that they are replaced by the median and the method applied to the medians. Finally we point out that an interaction can also be an outlier. There is no a priori way of distinguishing the two.

References

Adrover, J. (1998). Minimax bias-robust estimation of the dispersion matrix of multivariate distributions. *Annals of Statistics*, 26:2301–2320.

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

Atkinson, A. C. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association*, 89:1329–1339.

Barme-Delcroix, M.-F. and Gather, U. (2000). An isobar-surfaces approach to multidimensional outlier-proneness. Technical Report 20, Sonderforschungsbereich 475, University of Dortmund, Dortmund, Germany.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*. Wiley, New York, third edition.

Bartlett, M. S. (1935). The effect of non-normality on the t-distribution. *Proceedings of the Cambridge Philosophical Society*, 31:223–231.

Becker, C. and Gather, U. (1999). The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association*, 94:947–955.

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

Bednarski, T. (1993). Fréchet differentiability and robust estimation. In Mandl, P. and Husková, M. (eds), *Asymptotic Statistics: Proceedings of the Fifth Prague Symposium*, Springer Lecture Notes, pp.49–58. Springer.

Bednarski, T. and Clarke, B. R. (1998). On locally uniform expansions of regular functionals. *Discussiones Mathematicae: Algebra and Stochastic Methods*, 18:155–165.

Bednarski, T., Clarke, B. R., and Kolkiewicz, W. (1991). Statistical expansions and locally uniform Fréchet differentiability. *Journal of the Australian Mathematical Society*, 50:88–97.

Bernholt, T. and Fischer, P. (2001). The complexity of computing the mcd-estimator. Technical Report 45, Sonderforschungsbereich 475, University of Dortmund, Dortmund, Germany.

Berrendero, J. R. and Zamar, R. H. (2001). Maximum bias curves for robust regression with non-elliptical regressors. *Annals of Statistics*, 29:224–251.

Box, G. E. P. (1953). Non-normality and test on variance. *Biometrika*, 40:318–335.

Box, G. E. P. and Andersen, S. L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society Series B*, 17:1–34.

Caroni, C. and Prescott, P. (1992). Sequential application of wilk's multivariate outlier test. *Applied Statistics*, 41:355–364.

Chang, H., McKean, J. W., Narjano, J. D., and Sheather, S. J. (1999). High-breakdown rank regression. *Journal of the American Statistical Association*, 94(445):205–219.

Clarke, B. R. (1983). Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations. *Annals of Statistics*, 11:1196–1205.

Cohen, M. (1991). The background of configural polysampling: a historical perspective. In Morgenthaler, S. and Tukey, J. W. (eds), *Configural Polysampling: A Route to Practical Robustness*, chapter 2. Wiley, New York.

Croux, C. and Dehon, C. (2001). Robust linear discriminant analysis using S-estimators. *Canadian Journal of Statistics*, 29:473–492.

Croux, C. and Haesbroeck, G. (2000). Principal components analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87:603–618.

Croux, C. and Rousseeuw, P. J. (1992). Time-efficient algorithms for two highly robust estimators of scale. In Dodge, Y. and Whittaker, J. C. (eds), *Computational Statistics*, volume 1, pp.411–428, Heidelberg. Physica-Verlag.

Davies, P. L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *Annals of Statistics*, 15:1269–1292.

Davies, P. L. (1992a). The asymptotics of Rousseeuw's minimum volume ellipsoid. *Annals of Statistics*, 20:1828–1843.

Davies, P. L. (1993). Aspects of robust linear regression. *Annals of Statistics*, 21:1843–1899.

Davies, P. L. (1995). Data features. *Statistica Neerlandica*, 49:185–245.

Davies, P. L. (1998). On locally uniformly linearizable high breakdown location and scale functionals. *Annals of Statistics*, 26:1103–1125.

Davies, P. L. (2004). The one-way table. *Journal of Statistical Planning and Inference*, 122:3–13.

Davies, P. L. and Gather, U. (1993). The identification of multiple outliers (with discussion). *Journal of the American Statistical Association*, 88:782–801.

Dietel, G. (1993). *Global location and dispersion functionals*. PhD thesis, University of Essen.

Donoho, D. L. (1982). *Breakdown properties of multivariate location estimators*. PhD thesis, Department of Statistics, Harvard University, Harvard, Mass.

Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estimated based on halfspace depth and project outlyingness. *Annals of Statistics*, 20:1803–1827.

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

Eddington, A. S. (1914). *Stellar Movements and the Structure of the Universe*. Macmillan, New York.

Edelsbrunner, H. and Souvaine, D. (1990). Computing median-of-squares regression lines and guided topological sweep. *Journal of the American Statistical Association*, 85:115–119.

Ellis, S. P. (1998). Instability of least squares, least absolute deviation and least median of squares linear regression. *Statistical Science*, 13(4):337–350.

Fernholz, L. T. (1983). *Von Mises Calculus for Statistical Functionals*. Number 19 in Lecture Notes in Statistics. Springer-Verlag, New York.

Fisher, R. A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error and the mean square error. *Monthly Notices of the Royal Astronomical Society*, 80:758–770.

Fisher, R. A. (1935). *The Design of Experiments*, Oliver and Boyd, Edinburgh and London.

Gather, U. (1990). Modelling the occurrence of multiple outliers. *Allgemeines Statistisches Archiv*, 74:413–428.

Gather, U. and Hilker, T. (1997). A note on tyler's modification of the mad for the stahel-donoho estimator. *Annals of Statistics*, 25:2024–2026.

Gather, U., Kuhnt, S., and Pawlitschko, J. (2003). Concepts of outlyingness for various data structures. In Misra, J. C. (ed), *Industrial Mathematics and Statistics*. Narosa Publishing House, New Delhi, 545–585.

Gather, U. and Schultze, V. (1999). Robust estimation of scale of an exponential distribution. *Statistica Neerlandica*, 53:327–341.

Gayen, A. K. (1950). The distribution of the variance ratio in random samples of any size drawn from non-normal universe. *Biometrika*, 37:236–255.

Geary, R. C. (1936). The distribution of 'student's' ratio for non-normal samples. *Journal of the Royal Statistical Society Supplement*, 3:178–184.

Geary, R. C. (1947). Testing for normality. *Biometrika*, 34:209–242.

Gervini, D. and Yohai, V. J. (2002). A class of robust and fully efficient regression estimators. *Annals of Statistics*, 30(2):583–616.

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

Hadi, A. S. (1994). A modification of a method for the detection of outliers in multivariate samples. *Journal of the Royal Statistical Society, Series B*, 56:393–396.

Hadi, A. S. and Simonoff, J. S. (1997). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, 88:1264–1272.

Hampel, F. R. (1968). *Contributions to the theory of robust estimation*. PhD thesis, University of California, Berkeley.

Hampel, F. R. (1975). Beyond location parameters: Robust concepts and methods (with discussion). In *Proceedings of the 40th Session of the ISI*, volume 46, Book 1, pp.375–391.

Hampel, F. R. (1985). The breakdown points of the mean combined with some rejection rules. *Technometrics*, 27:95–107.

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

Hawkins, D. M. (1980). *Identification of outliers*, Chapman and Hall, London.

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

Huber, P. J. (1977). Robust statistical procedures. In *Regional Conference Series in Applied Mathematics No. 27*, Society for Industrial and Applied Mathematics, Philadelphia, Penn.

Huber, P. J. (1981). *Robust Statistics*, Wiley, New York.

Huber, P. J. (1984). Finite sample breakdown points of m- and p-estimators. *Annals of Statistics*, 12:119–126.

Huber, P. J. (1995). Robustness: Where are we now? *Student*, 1:75–86.

Kent, J. T. and Tyler, D. E. (1991). Redescending M-estimates of multivariate location and scatter. *Annals of Statistics*, 19:2102–2119.

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

Liu, R. Y., Parelius, J. M., and Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference. *Annals of Statistics*, 27:783–840.

Lopuhaä, H. P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *Annals of Statistics*, 19:229–248.

Lopuhaä, H. P. (1991). Multivariate τ-estimators for location and scatter. *Canadian Journal of Statistics*, 19:307–321.

Lopuhaä, H. P. and Rousseeuw, P. J. (1991). Breakdown properties of affine equivariant estimators of multivariate location and covariance matrices. *Annals of Statistics*, 19:229–248.

Marazzi, A. (1992). *Algorithms, Routines, and S- Functions for Robust Statistics*. Chapman and Hall, New York.

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

Maronna, R. A., Stahel, W. A., and Yohai, V. J. (1992). Bias-robust estimators of multivariate scatter based on projections. *Journal of Multivariate Analysis*, 42:141–161.

Maronna, R. A. and Yohai, V. J. (1981). Asymptotic behavior of general M-estimates for regression and scale with random carriers. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 58:7–20.

Maronna, R. A. and Yohai, V. J. (1993). Bias-robust estimates of regression based on projections. *Annals of Statistics*, 21(2):965–990.

Martin, R. D., Yohai, V. J., and Zamar, R. H. (1989). Min-max bias robust regression. *Annals of Statistics*, 17(4):1608–1630.

Martin, R. D. and Zamar, R. H. (1993a). Bias robust estimation of scale. *Annals of Statistics*, 21(2):991–1017.

Martin, R. D. and Zamar, R. H. (1993b). Efficiency constrained bias robust estimation of location. *Annals of Statistics*, 21(1):338-354.

Mendes, B. and Tyler, D. E. (1996). Constrained M-estimates for regression. In Rieder, H. (ed), *Robust Statistics; Data Analysis and Computer Intensive Methods*, number 109 in Lecture Notes in Statistics, pp.299–320. Springer-Verlag.

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society (London), Series A*, 231:289–337.

Pearson, E. S. (1929). The distribution of frequency constants in small samples from non-normal symmetrical and skew populations. *Biometrika*, 21:259–286.

Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika*, 23:114–133.

Pearson, E. S. and Chandra Sekar, S. (1936). The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika*, 28:308–320.

Pollard, D. (1984). *Convergence of stochastic processes*. Springer-Verlag, New York.

Riedel, M. (1989a). On the bias-robustness in the location model i. *Statistics*, 2:223–233.

Riedel, M. (1989b). On the bias-robustness in the location model ii. *Statistics*, 2:235–246.

Rieder, H. (1994). *Robust Asymptotic Statistics*. Springer, Berlin.

Rocke, D. M. (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension. *Annals of Statistics*, 24:1327–1345.

Rocke, D. M. and Woodruff, D. L. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91(435):1047–1061.

Rocke, D. M. and Woodruff, D. L. (1997). Robust estimation of multivariate location and shape. *Journal of Statistical Planning and Inference*, 91:245–255.

Rosner, B. (1975). On the detection of many outliers. *Technometrics*, 17:221–227.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In Grossmann, W., Pflug, C. G., Vincze, I. and Wertz, W. (eds), *Mathematical Statistics and Applications (Proceedings of the 4th Pannonian Symposium on Mathematical Statistics)*, volume B, Dordrecht. Reidel.

Rousseeuw, P. J. and Croux, C. (1992). Explicit scale estimators with high breakdown point. In Dodge, Y. (ed), *L_1-Statistical Analysis and Related Methods*, pp.77–92, Amsterdam. North Holland.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88:1273–1283.

Rousseeuw, P. J. and Croux, C. (1994). The bias of k-step M-estimators. *Statistics and Probability Letters*, 20:411–420.

Rousseeuw, P. J. and Hubert, M. (1999). Regression depth. *Journal of the American Statistical Association*, 94:388–402.

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

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

Rousseeuw, P. J. and Van Driesen, K. (2000). An algorithm for positive-breakdown methods based on concentration steps. In Gaul, W., Opitz, O., and Schader, M. (eds), *Data Analysis: Scientific modelling and Practical Application*, pp.335–346. Springer-Verlag, New York.

Rousseeuw, P. J. and van Zoomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85:633–639.

Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S-estimators. In Franke, J. e. a. (ed), *Robust and Nonlinear Time Series Analysis*, pp.256–272, New York. Springer.

Scholz, F. W. (1971). *Comparison of optimal location estimators*. PhD thesis, Department of Statistics, University of California, Berkley.

Sheather, S. J., McKean, J. W., and Hettmansperger, T. P. (1997). Finite sample stability properties of the least median of squares estimator. *Journal of Statistical Computing and Simulation*, 58(4):371–383.

Simonoff, J. S. (1984). A comparison of robust methods and detection of outlier techniques when estimating a location parameter. *Communications in Statistics, Series A*, 13:813–842.

Simonoff, J. S. (1987). The breakdown and influence properties of outlier rejection-plus-mean procedures. *Communications in Statistics, Series A*, 16:1749–1760.

Stahel, W. A. (1981). Breakdown of covariance estimators. Research Report 31, Fachgruppe für Statistik, ETH, Zurich.

Staudte, R. G. and Sheather, S. J. (1990). *Robust Estimation and Testing*, Wiley, New York.

Tatsuoka, K. S. and Tyler, D. E. (2000). On the uniqueness of S-functionals and M-functionals under non-elliptic distributions. *Annals of Statistics*, 28(4):1219–1243.

Terbeck, W. and Davies, P. L. (1998). Interactions and outliers in the two-way analysis of variance. *Annals of Statistics*, 26:1279–1305.

Tietjen, G. L. and Moore, R. H. (1972). Some grubbs-type statistics for the detection of several outliers. *Technometrics*, 14:583–597.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In Olkin, I. (ed), *Contributions to Probability and Statistics*. Stanford University Press, Stanford, California.

Tukey, J. W. (1975). Mathematics and picturing data. In *Proceedings of International Congress of Mathematicians, Vancouver*, volume 2, pp.523–531.

Tukey, J. W. (1993). Exploratory analysis of variance as providing examples of strategic choices. In Morgenthaler, S., Ronchetti, E., and Stahel, W. A. (eds), *New Directions in Statistical Data Analysis and Robustness*, Basel. Birkhäuser.

Tyler, D. E. (1994). Finite sample breakdown points of projection based multivariate location and scatter statistics. *Annals of Statistics*, 22:1024–1044.

von Mises, R. (1937). Sur les fonctions statistiques. In *Conférence de la Réunion Internationale des Mathématiciens*. Gauthier-Villars.

Willems, S., Pison, G., Rousseeuw, P. J., and Van Aelst, S. (2002). A robust hotelling test. *Metrika*, 55:125–138.

Yohai, V. J. (1987). High breakdown point and high efficiency robust estimates for regression. *Annals of Statistics*, 15:642–656.

Yohai, V. J. and Maronna, R. A. (1990). The maximum bias of robust covariances. *Communications in Statistics - Theory and Methods*, 19:3925–3933.

Yohai, V. J. and Zamar, R. H. (1988). High breakdown point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association*, 83:406–413.

Zuo, Y. (2001). Some quantitative relationships between two types of finite sample breakdown points. *Statistics and Probability letters*, 51:369–375.

Zuo, Y. and Serfling, R. (2000a). General notions of statistical depth function. *Annals of Statistics*, 28:461–482.

Zuo, Y. and Serfling, R. (2000b). Structural properties and convergence results for contours of sample statistical depth functions. *Annals of Statistics*, 28:483–499.

Index