Resenhas IME-USP 1993, Vol. 1, No. 2/3, 305 - 319.

Recent results on robust estimation in multivariate analysis [1]

Ricardo A. Maronna

Abstract: Classic methods in multivariate analysis require the estimation of mean vectors and covariance matrices, and their results can therefore be substantially altered by a small proportion of atypical observations ("outliers"). This paper reviews for nonspecialists the current state of research on the main approaches to replacing means and covariances by a location vector and a dispersion matrix which are not affected by outliers ("robust methods"), and the relationships among these approaches. They are: estimates based on an extension of the method of Maximum Likelihood ("M–estimates"); estimates based on the minimization of a robust scale of Mahalanobis distances ("S–estimates"); and two families of estimates based on projections: "P–estimates" and the Stahel–Donoho estimates. The advantages and drawbacks of these families are compared with respect to: efficiency, breakdown point, maximum bias and computational cost.

Key words: robust estimation, multivariate location and scatter.

1 Introduction

Classical methods in Multivariate Analysis require the estimation of means and covariances. It is well–known, however, that a small proportion of atypical points in the data ("outliers") suffices to drastically alter them. This is clearly illustrated in (Devlin, Gnanadesikan and Kettenring, 1982) and (Rousseeuw and Leroy, 1987).

As an example, we generate a pseudorandom sample $(x_{1i}, x_{2i}), i = 1, \ldots, n$ of size $n = 20$ ¿from a bivariate normal distribution with zero means, unit variances, and correlation $\rho = 0.7$. Suppose we are mainly interested in estimating ρ. The sample correlation is 0.77. Now we alter the sample by changing the sign of the two smallest x_{2i}'s . The altered sample is shown in Figure 1, and its sample correlation is -0.15. Thus, modifiying 10% of the observations may yield a drastic change in the estimate. Furthermore, this shows that such changes may be caused by observations that, being atypical, cannot be detected as univariate outliers in any of the coordinates.

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ represent a set of n data points in $\Re^p$. Call $\mathbf{m}(\mathbf{X})$ and $\mathbf{C}(\mathbf{X})$ the sample mean vector and covariance matrix based on $\mathbf{X}$, respectively. If $\mathbf{c} \in \Re^p$ and $\mathbf{A}$ is any $p \times p$–matrix, and $\mathbf{Y} = \mathbf{AX} + \mathbf{c}$, then $\mathbf{m}$ and $\mathbf{C}$ satisfy the *affine equivariance* properties

[1] Part of this work was done while the author was visiting the Department of Statistics, University of Washington, supported by Grant ONR N 00014-91-J-1074 AMS 1980 subject classification: 62H12.

$$\mathbf{m}(\mathbf{Y}) = \mathbf{A}\mathbf{m}(\mathbf{X}) + \mathbf{c}, \tag{1.1}$$

$$\mathbf{C}(\mathbf{Y}) = \mathbf{A}\mathbf{C}(\mathbf{X})\mathbf{A}'. \tag{1.2}$$

We are interested in defining a location vector $\mathbf{t}(\mathbf{X})$ and a positive definite dispersion matrix $\mathbf{V}(\mathbf{X})$ satisfying the equivariance conditions (1.1)-(1.2), such that, if $\mathbf{X}$ is contains only "good" observations, then $(\mathbf{t}, \mathbf{V}) \approx (\mathbf{m}, \mathbf{C})$; and, if a small proportion of observations is modified in any way, $(\mathbf{t}, \mathbf{V})$ does not change much. These two features may be called "efficiency" and "resistance", respectively. There are several important situations in Multivariate Analysis in which only the "shape" of the covariance matrix is required; i.e., one is interested in $\mathbf{C}$ only up to a scalar multiple; e.g. Principal Components and Discriminant Analysis. Attention will be devoted mainly to the estimation of the "shape" aspect of dispersion..

This paper reviews some recent results on these topics. It reflects the author's points of view, and hence does not attempt to be a fair survey.

2 M–estimates

One aproach to the definition of robust equivariant estimates is to define $\mathbf{t}$ and $\mathbf{V}$ as solutions to

$$\sum_{i=1}^{n} u_1(d_i)(\mathbf{x}_i - \mathbf{t}) = \mathbf{0}, \tag{2.3}$$

$$n^{-1} \sum_{i=1}^{n} u_2(d_i)(\mathbf{x}_i - \mathbf{t})(\mathbf{x}_i - \mathbf{t})' = \mathbf{V}; \tag{2.4}$$

with $d_i = d(\mathbf{x}_i; \mathbf{t}, \mathbf{V})$, where the "squared Mahalanobis distances" d are defined in general by

$$d(\mathbf{x}; \mathbf{t}, \mathbf{V}) = (\mathbf{x} - \mathbf{t})'\mathbf{V}^{-1}(\mathbf{x} - \mathbf{t}). \tag{2.5}$$

M–estimates are a generalisation of maximum likelihood estimates for elliptical distributions. If the observations $\mathbf{x}_i$ are i.i.d with density $f(\mathbf{x}; \mathbf{t}, \mathbf{V}) = (\det \mathbf{V})^{-1} f_0(d(\mathbf{x}; \mathbf{t}, \mathbf{V}))$ - –with $f_0 : \Re_+ \to \Re_+$ (where $\Re_+ = \{s : s \geq 0\}$) and d is defined in (2.5)– then the maximum likelihood estimates of $\mathbf{t}$ and $\mathbf{V}$ satisfy (2.3)-(2.4), with

$$u_1(s) = s^{-1/2}\psi_0(s^{1/2}), u_2(s) = u_1(s), \tag{2.6}$$

where $\psi_0(s) = -\mathrm{d} \log h_0(s)/\mathrm{d}s$. In particular, for the multivariate normal, $f_0(t) = (2\pi)^{-n/2} \exp(-t/2)$, and this implies $u_1(s) = u_2(s) \equiv 1$.

Another important case is the maximum likelihood estimate for the multivariate Student distribution with ν degrees of freedom, for which

$$u_1(s) = (\nu + p)/(\nu + s) \tag{2.7}$$

The case $\nu = 1$ corresponds to the Cauchy distribution. Tyler (1987) studied a "distribution-free" M-estimate which can be considered as the limit case of (2.7) when $\nu \to 0$; and recently Adrover (1993) found this estimate to have certain optimal properties.

If follows from (2.3)-(2.4) that $\mathbf{t}$ and $\mathbf{V}$ can be respectively viewed as a weighted mean and a weighted covariance matrix, with weights depending on the d_i's. Since u_1 and u_2 are usually decreasing, it follows that "distant" observations receive smaller weights.

The existence and uniqueness of solutions of (2.3)-(2.4), as well as their consistency and asymptotic normality, were first derived by Maronna (1976) under certain restrictions. A more general definition of M-estimators is given by Huber (1981). A very general result on existence and uniqueness for an important clas of M-estimates is given by Kent and Tyler (1991).

Numerical computing of M-estimates can be performed by an iterative algorithm which takes advantage of their expression as weighted means and covariances. More sophisticated algorithms have been implemented in the package ROBETH (Marazzi, 1993).

Given $\mathbf{V} = [v_{ij}]$, we can estimate correlations by $v_{ij}/(v_{ii}v_{jj})^{1/2}$. Using the Maximum Likelihood estimate for the Cauchy distribuution (henceforth CMLE) on the simulated bivariate sample yields correlation estimates of 0.75 for the original sample and 0.55 for the altered one, thus exhibiting a good behavior for both cases.

To see what happens in higher dimensions, we now generate a normal *spherical* sample of size 100 in dimension 10. The CMLE applied to this sample yields all correlations very near to those obtained ¿from $\mathbf{C}$. Now we modify the sample by replacing ten points $\mathbf{x}_i$ by $2\mathbf{u} + \mathbf{x}_i/2$, where $\mathbf{u} = (1, \ldots, 1)'$. The result is that now all correlations become larger than 0.65!.

This is an example of a drawback of M-estimates: their robustness decreases when the dimension increases. To make this assertion more precise, let us introduce a measure of robustness. The *breakdown point* δ^* of an estimate is –roughly speaking– the largest proportion of observations which may be arbitrarily altered without the estimate becoming totally meaningless. More precisely, let $\mathbf{T}(\mathbf{X})$ be an estimate defined for samples $\mathbf{X}$ of size n and taking values on a space $\mathcal{T}$. Let $m \in \{0, \ldots, n\}$, and call $\mathcal{X}_m$ the set of all samples of size n obtained by replacing m elements of $\mathbf{X}$ by arbitrary values. Let m_0 be the maximum of all m such that there exists a compact $\mathcal{K} \subset \mathcal{T}$ for which $\mathbf{T}(\mathbf{Y}) \in \mathcal{K} \ \forall \ \mathbf{Y} \in \mathcal{X}_m$. Then the breakdown point of $\mathbf{T}$ at $\mathbf{X}$ is defined as $\delta^*(\mathbf{T}, \mathbf{X}) = m_0/n$. In the jargon of robustness, $\mathbf{Y}$ is called a *contaminated sample*, and m/n is the *contamination proportion*.

This is the so-called *finite-sample replacement breakdown point* (Donoho and Huber, 1983). The asymptotic version is as follows. Given a distribution F_0, the analogue of a comtaminated sample is an "ϵ-contaminated distribution": $(1 - \epsilon)F_0 + \epsilon G$, where G is an arbitrary distribution. Assume that the estimate $\mathbf{T}$ is defined as a functional on distributions. Then the *contamination (or "gross error") breakdown point* $\epsilon^*(\mathbf{T}, F_0)$ of $\mathbf{T}$ at F_0 is the supremum of $\epsilon \in [0, 1]$ such

that $\mathbf{T}((1-\epsilon)F_0 + \epsilon G)$ remains in a compact for all G. A very general account of the concept of breakdown point is given in (Hampel, 1971).

In our case, we have $\mathbf{T} = (\mathbf{t}, \mathbf{V})$, and hence $\mathcal{T} = \Re^p \times \mathcal{SP}_p$, where $\mathcal{SP}_p$ is the space of symmetric positive definite $p \times p$–matrices. For $(\mathbf{t}, \mathbf{V})$ to remain in a compact it is necessary and sufficient that $||\mathbf{t}||$ be bounded, and that the eigenvalues of $\mathbf{V}$ be bounded away from 0 and from infinity. The last condition implies that the *condition number* of $\mathbf{V}$ –i.e., the ratio of its largest to its smallest eigenvalue– remain bounded.

M–estimates are robust against contamination which is not concentrated on any hyperplane (Tyler, 1986); but are very sensitive to contamination concentrated in a small cluster. Tyler (1991) found that for any M–estimate,

$$\delta^* \leq 1/(p+1) - 1/n, \tag{2.8}$$

the upper bound being attained, among others, by the CMLE given by (2.7) with $\nu = 1$. Tyler also described the form of the breakdown. Let $(\mathbf{t}, \mathbf{V})$ be a multivariate M–estimator with breakdown point δ^*. Let m observations tend to a point –say $\mathbf{x}_0$– where $m \geq n\delta^*$. Then $\mathbf{t} \to \mathbf{x}_0$, and the smallest eigenvalue of $\mathbf{V}$ tends to 0. This implies that for large p and very asymmetric contamination, M–estimates may be even less reliable than the classical ones!

The former results on the breakdown point hold also for $\mathbf{V}$ when $\mathbf{t}$ is known, but not conversely, implying that the weakness lies in the matrix $\mathbf{V}$.

3 S–estimators

A step towards the goal of defining equivariant estimates with a high breakdown point for all dimensions was Rousseeuw's *Minimum Volume Ellipsoid Estimate* (MVEE) (Rousseeuw, 1984 and Rousseeuw and Leroy, 1987), defined as follows. Among all ellipsoids $\{\mathbf{x} : d(\mathbf{x}; \mathbf{t}, \mathbf{V}) \leq 1\}$ containing at least half of the data points, choose $(\mathbf{t}, \mathbf{V})$ such that $\det \mathbf{V}$ –i.e., the volume of the ellipsoid– is minimized. Rousseeuw showed that this estimator has asymptotic breakdown point 1/2 for all dimensions, and that its finite sample breakdown point is $([n/2] - p + 1)/2n$ (where [.] denotes the integer part), thus making a significant improvement in robustness over M–estimators. Recently Davies (1993) studied the asymptotic behavior of the MVEE.

Davies (1987) generalized this estimate as follows. Given $(\mathbf{t}, \mathbf{V})$, let $\mathbf{d}(\mathbf{t}, \mathbf{V}) = (d(\mathbf{x}_i; \mathbf{t}, \mathbf{V}) : i = 1, \ldots, n)$. Let s be a scale statistics; define the multivariate S–estimator $(\mathbf{t}, \mathbf{V})$ as a solution of

$$(\mathbf{t}, \mathbf{V}) = \arg\min\{s(\mathbf{d}(\mathbf{t}, \mathbf{V})) : \mathbf{V} \in \mathcal{SP}_p,\ \det(\mathbf{V}) = 1\}.$$

It is easy to show that if s is the mean, then $\mathbf{t}$ and $\mathbf{V}$ are the sample mean and a scalar multiple of the sample covariance matrix, respectively; and that if s is

the median, the the solution is the MVEE. By the way, we see that by their very definition S–estimators can yield only the "shape" of dispersion.

Davies specialized to the case in which s is an "M–estimate of scale", defined as follows. Given a sample of nonnegative values $\mathbf{z} = (z_1, \ldots, z_n)$, define the scale M–estimator $s = s(\mathbf{z})$ by $\text{ave}\{\rho(z/s)\} = \delta$, where "ave" is the average, and ρ is a bounded nondecreasing function with $\rho(0) = 0$ and $\rho(\infty) = 1$.

Davies (1987) found an upper bound for the finite–sample replacement breakdown point of any affine–equivariant location and scatter statistics, namely

$$\delta^* \leq [(n - p + 1)/2]/n, \tag{3.9}$$

if $\mathbf{X}$ is in general position. Davies proved that in order that an S–estimate $(\mathbf{t}, \mathbf{V})$ attain the maximum breakdown point (3.9), one must have

$$\delta = (n - p - 1)/2n. \tag{3.10}$$

A slight modification of the MVEE attains this maximum breakdown point. It is obtained by taking $\rho(z) = I(z \geq 1)$ (where I is the indicator function) and δ as in (3.10) ; this implies that s is the k-th order statistics, with $k = [(n + p + 1)/2]$ (instead of $k = [(n + 1)/2]$ as in the median). Henceforth, we shall refer to this modified estimate as the MVEE.

A higher asymptotic efficiency may be obtained by using a smooth ρ-function, such as the "biweight" function, defined by

$$\rho'(z) = (1 - z)^2 I(z \leq 1). \tag{3.11}$$

Numerical computing of S–estimates presents difficult problems, due to the existence of many local minima. An attempt to approximate the MVEE, based on subsampling, is given in (Rousseeuw and Leroy, 1987). The reliability of this algorithm has been questioned by Cook and Hawkins (1991). An attempt to improve on this procedure is given by Ruppert (1993). Rocke and Woodruff (1993) and Woodruff and Rocke (1993) experiment the use of heuristic programming for computing the MVEE. Smooth S–estimates seem to present less difficulties than the MVEE.

In our example, the correlation estimates based on the MVEE are 0.85 for the original sample and 0.45 for the modified one, showing a much worse behavior than the CMLE. Note however that, according to (2.8) and (3.9), the breakdown points of CMLE and MVEE are 0.28 and 0.45 respectively. To understand why the latter had a worse behavior than the former, despite a higher breakdown point, we need a more general measure of behavior: the *bias under contamination*. Let $\Delta(.,.)$ be a measure of dissimilarity on $\mathcal{T}$ (if $\mathcal{T} = \Re^p$, usually $\Delta(\mathbf{t}_1, \mathbf{t}_2) = \|\mathbf{t}_1 - \mathbf{t}_2\|$). Then the bias of $\mathbf{T}$ at G is $\text{bias}(\mathbf{T}; \epsilon, G) = \Delta(\mathbf{T}((1 - \epsilon)F_0 + \epsilon G), \mathbf{T}(F_0))$. Consideration is often restricted to *pointwise contamination*, which corresponds to G of the form $G = \delta_{\mathbf{x}_0}$, i.e. the point mass at $\mathbf{x}_0$. In this case the bias is expressed by the *bias function*: $b(\mathbf{T}; \epsilon, \mathbf{x}_0) = \text{bias}(\mathbf{T}; \epsilon, \delta_{\mathbf{x}_0})$.

For the location vector, a suitable measure of dissimilarity is $\Delta(\mathbf{t}_1, \mathbf{t}_0) = (\mathbf{t}_1 - \mathbf{t}_0)'\mathbf{V}(F_0)^{-1}(\mathbf{t}_1 - \mathbf{t}_0)$. If we are interested in the shape of scatter, then a measure is

$$\Delta(\mathbf{V}_1, \mathbf{V}_0) = \varphi(\mathbf{A}_0\mathbf{V}_1\mathbf{A}_0'),$$

where φ is any measure of nonsphericity, and $\mathbf{A}_0$ is such that $\mathbf{A}_0'\mathbf{A}_0 = \mathbf{V}_0^{-1}$. These bias measures are clearly invariant under affine transformations. The simplest measure of nonsphericity of a matrix $\mathbf{W}$ is its condition number $\mathrm{cond}(\mathbf{W})$. Another one is the likelihood ratio test statistics for nonsphericity (Muirhead 1982), namely

$$\varphi_0(\mathbf{W}) = (\mathrm{tr}\,\mathbf{W}/p)^p / \det(\mathbf{W}), \tag{3.12}$$

where tr denotes the trace.

Unfortunetely, despite their high breakdown point, S–estimates may have a very high bias under pointwise contamination, as proved by Yohai and Maronna (1990). This reveals that the breakdown point cannot be the sole criterion used to evaluate robustness : if an estimate has breakdown point ϵ^*, this means that its bias under a contamination proportion $\epsilon < \epsilon^*$ is *bounded*; but this does not imply that the bound is *small*!.

In order to make S-estimators more efficient, Rousseeuw proposed *reweighted S–estimators*. Given the S–estimators $(\mathbf{t}, \mathbf{V})$, define the d_i's as in (2.5). Let W be a weight function, and $w_i = W(d_i)$. Define $(\mathbf{t}^*, \mathbf{V}^*)$ as a weighted mean and a weighted covariance matrix with weights w_i. The most usual choice is "hard rejection": $W(d) = I(d \leq d_0)$, where the threshold d_0 is conveniently chosen (depending on p) in order to find the best behavior.

4 P–estimates

The unpleasant features of S-estimates show that it does not suffice for an estimator to have a high breakdown point, but rather that the behavior of its bias function must be taken into account. Maronna, Stahel and Yohai (1992) developed an idea that had been successfully used by Maronna and Yohai (1993) to find regression estimates with low maximum bias, and which took into account all univariate *projections* of the data. Note first that the covariance matrix $\mathbf{C}$ has the property that, if $\mathbf{A}$ is such that $\mathbf{AA} = \mathbf{C}^{-1}$, then

$$\mathrm{var}(\mathbf{a}'\mathbf{AX}) = 1 \;\forall\; \mathbf{a} \in S_p = \{\mathbf{a} \in \Re^p : \|\mathbf{a}\| = 1\}; \tag{4.13}$$

i.e., the dispersion of the transformed data is constant in all directions. The idea is to replace the variance by a robust (univariate) dispersion estimate s. A simple possibility is the *median absolute deviation*: $\mathrm{MAD}(\mathbf{z}) = \mathrm{med}(|\mathbf{z} - \mathrm{med}(\mathbf{z})|$ (where "med" stands for the median). Since in general there will not exist a transformation for which the dispersion is constant, we look for one making it

"as constant as possible". Thus, we define the *P-estimator* $\mathbf{V}$ of multivariate dispersion as $\mathbf{V} = (\mathbf{A}'\mathbf{A})^{-1}$, where

$$\mathbf{A} = \arg\min\{\log \sup_{\mathbf{a}\in S_p} |s(\mathbf{a}'\mathbf{A}\mathbf{X}) - 1|\}; \tag{4.14}$$

or, alternatively

$$\mathbf{A} = \arg\min \frac{\sup_{\mathbf{a}\in S_p} s(\mathbf{a}'\mathbf{A}\mathbf{X})}{\inf_{\mathbf{a}\in S_p} s(\mathbf{a}'\mathbf{A}\mathbf{X})}, \tag{4.15}$$

with the restriction

$$\inf_{\mathbf{a}\in S_p} s(\mathbf{a}'\mathbf{A}\mathbf{X}) = 1. \tag{4.16}$$

It is proved in (Maronna et al., 1992) that solutions of (4.14) and of (4.15)–(4.16) differ only in a scalar multiple.

For our simulated sample, the estimated correlation based on a P–estimate is 0.75 for the original sample, and 0.72 for the modified one, showing an excellent behavior for both cases.

The maximum bias of these estimates is computed in (Maronna et al., 1992) and shown to be much better than that of M– and S–estimates for $p \geq 5$. Simulations in that paper also show that they behave better than the MVEE for finite sample sizes.

However, some simulations show P–estimates to behave rather erratically as a function of data –even for normal data– when the ratio n/p is not large enough (e.g. =5); and this drawback should be fixed to make the estimator reliable.

Numerical computing of P–estimates is much harder than that of S–estimates, because of the double optimization process involved (over matrices and over directions). Maronna et al. (1992) found a subsampling algorithm with a shortcut which saves much effort; but even so, computing for large p seems still impractical.

5 The Stahel–Donoho estimate

Another estimate based on projections was the one defined independently by Stahel (1981) and Donoho (1982). It was the first robust equivariant estimate of multivariate location and scatter having a high breakdown point for any dimension. The estimator is defined as a weighted mean and a weighted covariance matrix, where each point has a weight which is a function of an "outlyingness" measure, with points having large outlyingness receiving small weights. The outlyingness measure is based on the idea that if a point is a multivariate outlier, there must be some one–dimensional projection of the data for which it is a (univariate) outlier.

Let $\mu(.)$ and $\sigma(.)$ be shift and scale equivariant (resp. shift invariant and scale equivariant) univariate location and dispersion statistics. Note that if $z = \{z_1, \ldots, z_n\}$ is a univariate sample, one may detect suspicious observations by

looking at high values of $|z_i - \mu(\mathbf{z})|/\sigma(\mathbf{z})$. Define now for any $\mathbf{y} \in \Re^p$ the "multivariate outlyingness" r:

$$r(\mathbf{y}, \mathbf{X}) = \sup_{\mathbf{a}} r_1(\mathbf{y}, \mathbf{a}, \mathbf{X}), \tag{5.17}$$

where the "univariate outlyingness" r_1 is

$$r_1(\mathbf{y}, \mathbf{a}, \mathbf{X}) = |\mathbf{a}'\mathbf{y} - \mu(\mathbf{a}'\mathbf{X})|/\sigma(\mathbf{a}'\mathbf{X}), \tag{5.18}$$

and the supremum is over $\mathbf{a} \in \Re$ with $\mathbf{a} \neq \mathbf{0}$ or equivalently over $\mathbf{a} \in S_p$.

Let the "weight function" $w : R_+ \to R_+$ be bounded and continuous, with $r^2 w(r)$ being bounded. The Stahel–Donoho estimator (SDE) of location and scatter $(\mathbf{t}(\mathbf{X}), \mathbf{V}(\mathbf{X}))$ is defined as

$$\mathbf{t} = \mathbf{t}(\mathbf{X}) = \frac{\sum_{i=1}^n w_i \mathbf{x}_i}{\sum_{i=1}^n w_i} \tag{5.19}$$

and

$$\mathbf{V} = \mathbf{V}(\mathbf{X}) = \frac{\sum_{i=1}^n w_i (\mathbf{x}_i - \mathbf{t})(\mathbf{x}_i - \mathbf{t})'}{\sum_{i=1}^n w_i}. \tag{5.20}$$

with $w_i = w(r(\mathbf{x}_i, \mathbf{X}))$. The value of $r_1(\mathbf{y}, \mathbf{X})$ is affine invariant, i.e. $r_1(\mathbf{y}, \mathbf{X}) = r_1(\mathbf{A}\mathbf{y} + \mathbf{b}, \mathbf{A}\mathbf{X} + \mathbf{b})$. for any nonsingular $\mathbf{A}$ and any $\mathbf{b} \in \Re^p$; and this implies that $(\mathbf{t}, \mathbf{V})$ are affine equivariant. Note that if μ and σ are respectively the mean and the standard deviation, then $r(\mathbf{y}, \mathbf{X}) = (\mathbf{y} - \mathbf{m})'\mathbf{C}^{-1}(\mathbf{y} - \mathbf{m})$, where $\mathbf{m}$ and $\mathbf{C}$ are the sample mean and covariance matrix, respectively. Stahel (1981) showed that $(\mathbf{t}, \mathbf{V})$ has asymptotic breakdown point $1/2$ at continous multivariate models if μ and σ have asymptotic breakdown point $1/2$ (see Hampel et al, (1986), Theorem 5.5.3). Donoho (1982) derived the finite sample breakdown point of $(\mathbf{t}, \mathbf{V})$ for the case in which μ and σ are the median and the median absolute deviation (MAD) respectively.

No further results on these estimators were published in the ensuing years, one likely ground being the seeming intractability of their properties and of their computation. Stahel (1981) himself had proposed an algorithm based on subsampling for the approximate computing of $(\mathbf{t}, \mathbf{V})$, but no attempts were made at experimenting it. The popularity and better tractability of the MVEE and in general of multivariate S–estimators may also explain the lack of interest in the Stahel–Donoho estimators.

Recently, Tyler (1993) obtained important results on the replacement finite–sample breakdown point of $(\mathbf{t}, \mathbf{V})$. In particular, he derived conditions under which it attains the upper bound (3.9). The two most important cases in which maximum breakdown is attained are:

- μ is the median, and σ the average of the k_1–th and the k_2–th smallest absolute deviations about μ, with

$$k_1 = p - 1 + [(n+1)/2] \text{ and } k_2 = p - 1 + [(p+2)/2]. \tag{5.21}$$

  This is a slight modification of the MAD.

- μ and σ are the maximum likelihood estimates for location and scale corresponding to a sample from a location–scale family of distributions based on Student's t–distribution with ν degrees of freedom, with

$$\nu = \frac{n+p}{n-p} \tag{5.22}$$

Maronna and Yohai (1993) showed that SDE has order $\sqrt{n}$–consistency. They also computed numerically the bias for μ and σ chosen as in (5.21) and different weight functions. The best results were obtained for functions of the form $w(r) = I(r \leq c) + (c/r)^2 I(r > c)$ ("Huber weights"); where the apropriate value of c depends on p. The maximum bias of the SDE was better than those of M–estimates and MVEE.

For our simulated bivariate data, the estimated correlations are 0.7 and 0.6 for the original and the altered sample respectively, thus showing a very satisfactory behavior. Numerical computing of the estimate is difficult for $p > 2$, due to the maximization in (5.17), which involves functions with many local extrema. However, Maronna and Yohai (1993) experimented Stahel's subsampling algorithm, which turned out to yield satisfactory results at least for $p \leq 10$. For samples of size $n = 30$, the computer times needed, using the GAUSS system on a PC with 55 Mhz frequency for $p = 4$, 6 and 10 were 1.1, 2.3 and 5.4 minutes respectively. S–estimates require about the same time, while the P–estimate required 2, 7 and 15 minutes respectively.

6 Comparisons

Maronna and Yohai (1993) performed a simulation to compare several estimates, namely:

- The Maximum Likelihood estimate for the Cauchy distribuution (CMLE), chosen among M–estimates for its maximum breakdown point.
- The MVEE.
- The S–estimate with biweight function (3.11) ("S–E"). with the parameter δ chosen as (3.10).
- Reweighted versions of both types of S–estimates, with the "hard rejection" weight function, trying different values of the cutoff threshold in order to find the best behavior.
- The Stahel–Donoho estimate (SDE)with μ and σ given as in (5.21), and "Huber weight function", with different values of c.
- The mean and covariance matrix ("COV").

P–estimates were not included because of their much higher computational cost.

The dimensions chosen were: $p = 2$ (with $n = 10$ and 20), $p = 4$ with $n = 20$, and $p = 6$ with $n = 30$. The distributions employed were:

- The unit normal spherical distribution.
- The Cauchy spherical distribution, chosen as an extreme case of heavy–tailed symmetric situation.
- Contaminated normal samples $\mathrm{CN}(\epsilon, k)$, chosen as an extreme case of asymmetric contamination. They consisted of $n - m$ observations distributed as $\mathcal{N}_p(\mathbf{0}, \mathbf{I})$, and m observations concentrated at $k\mathbf{b}_1$ with $m = [n\epsilon]$ and $\mathbf{b}_1' = (1, 0, \ldots, 0)$. The values $\epsilon = 0.10$ and 0.20 were chosen. Several values of k were used, searching for the worst behavior of each of the estimators.

For each estimate $\mathbf{V}$ and each distribution, the measure of error (the substitute of mean squared error) with respect to the spherical form was chosen as the "median error" $\mathrm{ME}(\mathbf{V}) = \mathrm{med} \log \varphi_0(\mathbf{V})$, where φ_0 is defined in (3.12). Medians rather than means were used, because of the skewness and heavy-tailedness of the empirical distributions of φ_0.

In view of the bulkyness of the output of the simulation, four criteria were displayed for each estimator: the ME's for normal and Cauchy distributions, and the maximal (over k) ME's for $\mathrm{CN}(\epsilon, k)$ for $\epsilon = 0.10$ and 0.20.

When $\mathbf{x}$ is spherical normal, it is proved (Muirhead 1982) that $n \log \varphi_0(\mathbf{C})$ converges in law to a linear combination of chi–squared distributions. Thus, if $\mathbf{V}$ is any of the estimators, the ratio of the median error for COV to the corresponding value for $\mathbf{V}$ may be considered as a measure of efficiency. Define the "efficiency" of the estimator $\mathbf{V}$ for the normal (resp. Cauchy) distribution as $\mathrm{ME}(\mathbf{V}_0)/\mathrm{ME}(\mathbf{V})$ where $\mathbf{V}_0$ is the COV (resp. CMLE). It was considered more clear to display the efficiencies rather than the ME's for both spherical distributions. The Stahel–Donoho estimate showed in general the best behavior. Details may be found in (Maronna and Yohai, 1993).

To give an comparative idea of the behavior of the different types of estimates, we plot the ME's corresponding to $\mathrm{CN}(\epsilon, k)$ as a function of k for $p = 4$ (the values corresponding to $k = 0$ in the plot ae actually for $\epsilon = 0$). Figure 2 displays for $\epsilon = 0.20$ the behavior of the Cauchy Maximum Likelihood, the best choice of the Stahel–Donoho, S– and Covariance estimates (labeled as CMLE, SDE, the S–E and COV, respectively). A remarkable feature is the "redescending" behavior of the S–E: its ME is the lowest one for large k, but its maximum ME is the largest. The SDE has both the smallest maximum ME and the smallest ME for $\epsilon = 0$. We see that the CMLE behaves worse than COV, the reason being that ϵ is larger than its brakdown point, causing the phenomenon described below (2.8).

The efficiencies of the SDE, S–E and CMLE are 0.82, 0.80 and 0.73 respectively for the normal distribution; and 0.91, 0.56 and 1.0 for the Cauchy distribution. Thus, the SDE is seen to combine a high efficiency for spherical distributions with a relatively low ME for asymmetric contamination.

It remains an open problem to find some variant of the SDE keeping all of these advantages and at the same time exhibiting the "redescending" behavior of S-estimates.

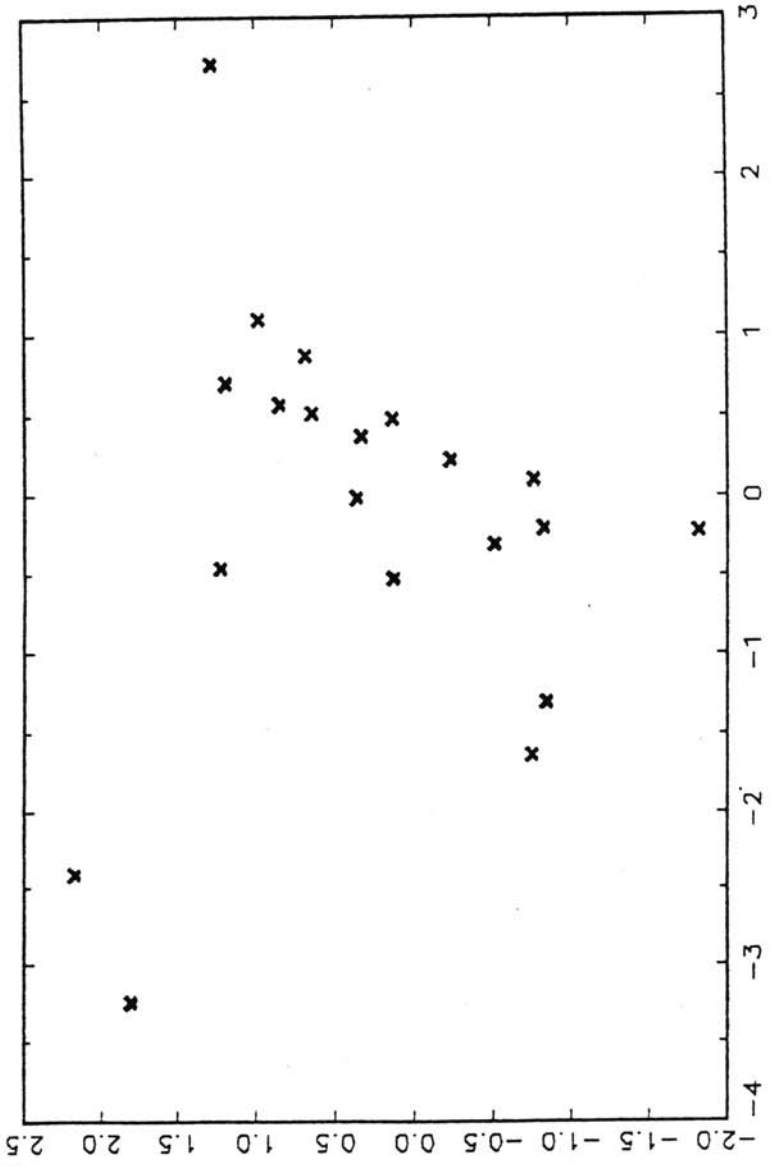

Figure 1. Simulated data with atypical points

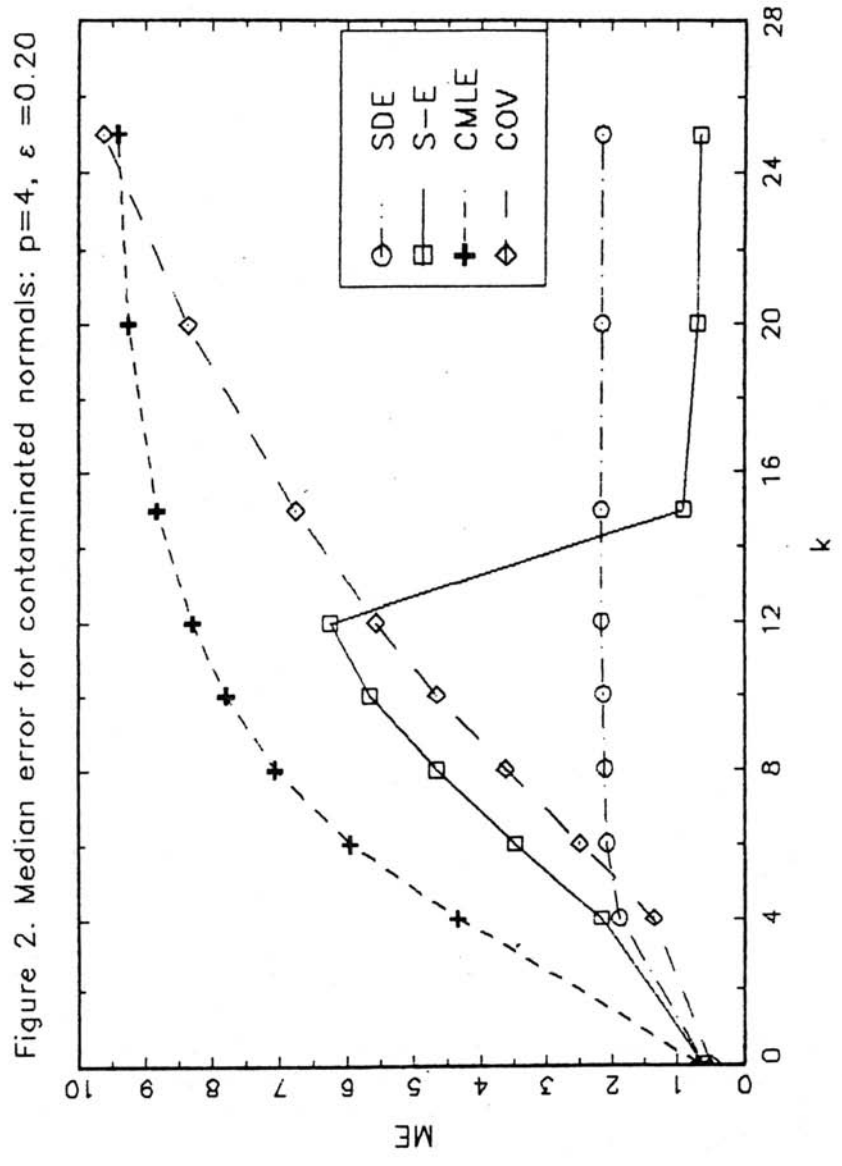

Figure 2. Median error for contaminated normals: p=4, ε =0.20

References

Adrover, J. (1993). Minimax estimates for the linear model and multivariate analysis (in Spanish). Ph. D. Thesis. University of Buenos Aires.

Cook, R.D. and Hawkins, D.M. (1990). Comments to (Rousseeuw and van Zomeren, 1990). *Jr. Amer. Statist. Assoc.* **85**, 640-644.

Davies, P.L. (1987). Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices.*Ann. Statist.* **15**, 1269-1292.

Davies, P.L. (1993) The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *Ann. Statist.*

Devlin, S.J., Gnanadesikan, R. and Kettenring, J.R. (1981). Robust estimation of dispersion matrices and principal components. *Jr. Amer. Statist. Assoc.* **76**, 354-362.

Donoho, D.L. (1982). Breakdown properties of multivariate location estimators. Ph. D. Qualifying paper. Harvard University. Donoho, D.L. and Huber, P.J. (1983). The notion of breakdown point. In: *Festschrift in honor of Erich Lehmann*, K. Doksum and J.L. Hodges (eds.). Belmont, Wadsworth.

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

Huber, P.J. (1981) *Robust Statistics.* John Wiley and Sons. New York. Kent, J.T. and Tyler, D.E. (1991). Redescending M-estimates of multivariate location and scatter. *Ann. Statist.* **19**, 2102-2119.

Marazzi, A. (1993). *Algorithms, Routines and S functions for robust statistics.* Wadsworth and Brooks, Cole. Maronna, R.A. and Yohai, V.J. (1993). Bias-robust estimates of regression based on projections. To appear in *Ann. Statist.*.

Maronna, R.A. and Yohai, V.J. (1993). The behavior of the Stahel-Donoho estimator. Technical Report. Department of Statistics. University of Washington.

Maronna, R.A., Stahel, W.A. and Yohai, V.J. (1982). Bias-robust estimators of multivariate scatter based on projections. *Jr. Mult. Anal.* **42**, 141-161.

Muirhead, R.J. (1982) *Aspects of Multivariate Statistical Theory.* John Wiley and Sons. New York.

Rocke, D.M and Woodruff, D.L. (1993). Computation of robust estimates of multivariate location and shape. *Statist. Neerland.* **47**, 27–42.

Rousseeuw, P.J. (1984). Multivariate estimators with high breakdown point. In *Mathematical Statistics and its Applications*, vol. B (W. Grossmann, G. Pflug, I. Vincze and W. Wertz, eds.). Reidel, Dordrechts, Netherland.

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

Rousseeuw, P.J and van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points. *Jr. Amer. Statist. Assoc.* **85**, 633-639.

Ruppert, D (1993). Computing S–estimators for regression and multivariate location/dispersion. *Jr. Comp. Graph. Statist.*

Stahel, W.A. (1981). Breakdown of covariance estimators. Research report 31, Fachgruppe für Statistik, E.T.H. Zürich.

Tyler, D.E. (1986). Breakdown properties of the M–estimators of multivariate scatter. Unpublished manuscript.

Tyler, D.E. (1987). A distribution–free M–estimator of multivariate scatter. *Ann. Statist.* **15**, 234-251.

Tyler, D.E. (1991). Personal communication.

Tyler, D.E. (1993) Finite Sample Breakdown Points of Projection Based Multivariate Location and Scatter Statistics. To appear in *Ann. Statist.*

Woodruff,D.L. and Rocke, D.M. (1993) Heuristic search algorithms for the minimum volume ellipsoid. *Jr. Comp. Graph. Statist.* **2**, 69–95.

Yohai, V.J. and Maronna, R.A. (1990). The maximum bias of robust covariances. *Comm. Stat. Theor. Meth.* **19**, 3925-3933.

R. Maronna
Departamento de Matematicas
Casilla Correo 172
1900 La Plata
Argentina