The Behavior of the Stahel-Donoho Robust Multivariate Estimator

Ricardo A. MARONNA AND Víctor J. YOHAI*

The Stahel–Donoho estimators $(\mathbf{t}, \mathbf{V})$ of multivariate location and scatter are defined as a weighted mean and a weighted covariance matrix with weights of the form $w(r)$, where w is a weight function and r is a measure of "outlyingness," obtained by considering all univariate projections of the data. It has a high breakdown point for all dimensions and order $\sqrt{n}$ consistency. The asymptotic bias of $\mathbf{V}$ for point mass contamination for suitable weight functions is compared with that of Rousseeuw's minimum volume ellipsoid (MVE) estimator. A simulation shows that for a suitable w, $\mathbf{t}$ and $\mathbf{V}$ exhibit high efficiency for both normal and Cauchy distributions and are better than their competitors for normal data with point-mass contamination. The performances of the estimators for detecting outliers are compared for both a real and a synthetic data set.

KEY WORDS: Bias robustness; Multivariate location and scatter; Projection methods.

1. INTRODUCTION

The estimator defined by Stabel (1981) and Donoho (1982) was the first robust equivariant estimator 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 that 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, then there must be some one-dimensional projection of the data for which the point is a (univariate) outlier.

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n)$ represent a set of n data points in $\Re^p$. Let $\mu(\cdot)$ and $\sigma(\cdot)$ be shift and scale equivariant (resp. shift invariant and scale equivariant) univariate location and scale statistics. Define for any $\mathbf{y} \in \Re^p$ the "outlyingness" r:

$$r(\mathbf{y}, \mathbf{X}) = \sup_{\mathbf{a}} r_1(\mathbf{y}, \mathbf{a}, \mathbf{X}), \tag{1}$$

where

$$r_1(\mathbf{y}, \mathbf{a}, \mathbf{X}) = |\mathbf{a}'\mathbf{y} - \mu(\mathbf{a}'\mathbf{X})| / \sigma(\mathbf{a}'\mathbf{X}) \tag{2}$$

and the supremum is over $\mathbf{a} \in \Re^p$ with $\mathbf{a} \neq \mathbf{0}$ or, equivalently, over $\mathbf{a} \in S_p = \{\mathbf{a} \in \Re^p : \|\mathbf{a}\| = 1\}$. Let w (the "weight function") be a function $R_+ \rightarrow R_+$ (where $R_+ = \{s : s\rangle \geq 0\}$). The Stahel–Donoho estimator (SDE) of location and scatter $(\mathbf{t}(\mathbf{X}), \mathbf{V}(\mathbf{X}))$ is defined as

$$\mathbf{t} = \mathbf{t}(\mathbf{X}) = \frac{\sum_{i=1}^n w_i \mathbf{x}_i}{\sum_{i=1}^n w_i} \tag{3}$$

and

$$\mathbf{V} = \mathbf{V}(\mathbf{X}) = \frac{\sum_{i=1}^n w_i (\mathbf{x}_i - \mathbf{t})(\mathbf{x}_i - \mathbf{t})'}{\sum_{i=1}^n w_i}, \tag{4}$$

with $w_i = w(r(\mathbf{x}_i, \mathbf{X}))$.

The value of $r(\mathbf{y}, \mathbf{X})$ is affine invariant; that is, $r(\mathbf{y}, \mathbf{X}) = r(\mathbf{A}\mathbf{y} + \mathbf{b}, \mathbf{A}\mathbf{X} + \mathbf{b})$ for any nonsingular $\mathbf{A}$ and any $\mathbf{b} \in \Re^p$. This implies that $(\mathbf{t}, \mathbf{V})$ is *affine equivariant;* that is, $\mathbf{t}(\mathbf{AX} + \mathbf{b}) = \mathbf{At}(\mathbf{X}) + \mathbf{b}$ and $\mathbf{V}(\mathbf{AX} + \mathbf{b}) = \mathbf{AV}(\mathbf{X})\mathbf{A}'$.

Note that if μ and σ are the mean and the standard deviation, then $r(\mathbf{y}, \mathbf{X}) = (\mathbf{y} - \bar{\mathbf{x}})'\mathbf{C}^{-1}(\mathbf{y} - \bar{\mathbf{x}})$, where $\bar{\mathbf{x}}$ and $\mathbf{C}$ are the sample mean and covariance matrix.

Stahel (1981) showed that $(\mathbf{t}, \mathbf{V})$ has an asymptotic breakdown point of $\frac{1}{2}$ at continuous multivariate models if μ and σ have an asymptotic breakdown point of $\frac{1}{2}$ (see Hampel, Ronchetti, Rousseeuw, and Stahel 1986, thm. 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). No further results on these estimators were published in the ensuing years, one likely reason being the seeming intractability of their properties and of their computation. Stahel (1981) proposed an algorithm based on subsampling for approximate computation of $(\mathbf{t}, \mathbf{V})$, but no attempts were made at experimenting with it. The popularity and better tractability of Rousseeuw's minimum volume ellipsoid estimator (MVEE) (see Rousseeuw 1985 and Rousseeuw and Leroy 1987) and in general of multivariate S estimators (Davies 1987) may also explain the lack of interest in the SDE's.

Recently, Tyler (1994) obtained important results on the replacement finite-sample breakdown point $\epsilon^*(\mathbf{t}, \mathbf{V}; \mathbf{X})$ of $(\mathbf{t}, \mathbf{V})$, assuming w to satisfy the following:

$$w \text{ is continuous and positive} \quad \text{and}$$

$$w(r) \text{ and } r^2 w(r) \text{ are bounded for } r \geq 0. \tag{5}$$

In particular, he derived conditions under which $\epsilon^*(\mathbf{t}, \mathbf{V}; \mathbf{X})$ attains the upper bound derived by Davies (1987) for the finite-sample replacement breakdown point of any affine-equivariant location and scatter statistics, namely

$$\epsilon_0^* = [(n - p + 1)/2]/n \tag{6}$$

(where $[\cdot]$ denotes the integer part), if $\mathbf{X}$ is in general position. The two most important cases in which maximum breakdown is attained are:

* Ricardo A. Maronna is Professor, Departamento de Matemática, Facultad de Ciencias Exactas, Universidad Nacional de la Plata, 1900 La Plata, Argentina, and Researcher at CICPBA. Víctor J. Yohai is Professor, Universidad de San Andrés, 1644 Victoria, Argentina, Professor at the University of Buenos Aires, and Researcher at CONICET. Part of this work was done while the authors were visiting the Department of Statistics of the University of Washington, supported by Office of Naval Research Grant ONR N 00014-91-J-1074.

© 1995 American Statistical Association
Journal of the American Statistical Association
March 1995, Vol. 90, No. 429, Theory and Methods

- μ is the median and σ is the average of the k_1th and the k_2th smallest absolute deviations about μ, with

$$k_1 = p - 1 + [(n+1)/2] \quad \text{and}$$

$$k_2 = p - 1 + [(n+2)/2]. \tag{7}$$

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

$$\nu = \frac{n+p}{n-p}. \tag{8}$$

The article is organized as follows. In Section 2 we show that $(\mathbf{t}, \mathbf{V})$ has order $\sqrt{n}$-consistency. In Section 3 we find w functions for which the maximum asymptotic bias of $\mathbf{V}$ under contamination of a spherical model behaves very satisfactorily as compared to that of the MVEE and of M estimators. In Section 4 we investigate the behavior of Stahel's approximate algorithm for finite-sample computation of the statistics and the breakdown point of such approximate estimators, and we run a simulation comparing the SDE with several S and M estimators, in which the SDE shows a very satisfactory behavior with respect to both scatter and location. In Section 5 we explore the SDE's performance for the detection of outliers, when applied to real and synthetic data sets. In Section 6 we summarize the main results obtained and their practical implications. Finally, in Appendixes A and B we provide proofs of results and an explicit asymptotic expression for r.

2. ASYMPTOTIC BEHAVIOR

Let $\mathbf{x}$ be a random vector in $\Re^p$ with distribution F. Assume that μ and σ are functionals defined for any distribution. Let $\mu_{\mathbf{a}} = \mu(F_{\mathbf{a}})$ and $\sigma_{\mathbf{a}} = \sigma(F_{\mathbf{a}})$, where $F_{\mathbf{a}}$ is the distribution of $\mathbf{a}'\mathbf{x}$. We define $\mathbf{t}(F)$ and $\mathbf{V}(F)$ as follows. Let, for $\mathbf{y} \in \Re^p$,

$$r(\mathbf{y}, F) = \sup_{\mathbf{a}} r_1(\mathbf{y}, \mathbf{a}, F), \tag{9}$$

where

$$r_1(\mathbf{y}, \mathbf{a}, F) = |\mathbf{a}'\mathbf{y} - \mu_{\mathbf{a}}|/\sigma_{\mathbf{a}}, \tag{10}$$

and define

$$\mathbf{t} = \mathbf{t}(F) = \frac{Ew(r(\mathbf{x}, F))\mathbf{x}}{Ew(r(\mathbf{x}, F))} \tag{11}$$

and

$$\mathbf{V} = \mathbf{V}(F) = \frac{Ew(r(\mathbf{X}, F))(\mathbf{x} - \mathbf{t})(\mathbf{x} - \mathbf{t})'}{Ew(r(\mathbf{x}, F))}. \tag{12}$$

If $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and F_n is the respective empirical distribution, then (3)–(4) are equal to $(\mathbf{t}(F_n), \mathbf{V}(F_n))$.

Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be iid random vectors in $\Re^p$ with distribution F. For simplicity, for $\mathbf{a} \in \Re^p$, denote the empirical distribution of $\mathbf{x}_1'\mathbf{a}, \ldots, \mathbf{x}_n'\mathbf{a}$ by $F_{n\mathbf{a}}$ and denote the distribution of $\mathbf{a}'\mathbf{x}$ under F by $F_{\mathbf{a}}$, as before. Define $\mu_{n\mathbf{a}} = \mu(F_{n\mathbf{a}})$ and $\sigma_{n\mathbf{a}} = \sigma(F_{n\mathbf{a}})$. The same notation will be used for both distributions and distribution functions.

Theorem 1 ($\sqrt{n}$-consistency). Assume the following:

$$n^{1/2}\sup|\mu_{n\mathbf{a}} - \mu_{\mathbf{a}}| = O_P(1), \tag{13}$$

$$n^{1/2}\sup|\sigma_{n\mathbf{a}} - \sigma_{\mathbf{a}}| = O_P(1), \tag{14}$$

$$\sup \mu_{\mathbf{a}} = C < \infty, \tag{15}$$

$$\inf \sigma_{\mathbf{a}} > 0, \tag{16}$$

and

$$\sup \sigma_{\mathbf{a}} = E < \infty, \tag{17}$$

where all sup and inf are for $\mathbf{a} \in S_p$ and the weight function w is such that there exist constants γ and η such that

$$|w(y) - w(y')| \le \gamma|y - y'| \quad \forall\, y, y' \in \Re \tag{18}$$

and

$$|w(y) - w(y')| \le \frac{\eta|y - y'|}{\min(y, y')^3} \quad \forall\, y, y' \in \Re. \tag{19}$$

Then

$$\sqrt{n}(\mathbf{t}(F_n), \mathbf{V}(F_n)) - (\mathbf{t}(F), \mathbf{V}(F)) = O_P(1).$$

The following theorem shows that conditions (13)–(17) hold in a simple case.

Theorem 2. Let μ and σ be the median and MAD. Assume that there exist constants $c > 0$ and $\delta > 0$ such that

$$|F_{\mathbf{a}}(\mu_{\mathbf{a}} + \nu) - F_{\mathbf{a}}(\mu_{\mathbf{a}})| \ge c|\nu| \qquad \forall\, \mathbf{a} \in S_p \quad \forall\, |\nu| \le \delta \tag{20}$$

and

$$|F_{\mathbf{a}}(\mu_{\mathbf{a}} \pm \sigma_{\mathbf{a}} + \nu) - F_{\mathbf{a}}(\mu_{\mathbf{a}} \pm \sigma_{\mathbf{a}})| \ge c|\nu| \qquad \forall\, \mathbf{a} \in \Re^p \quad \forall\, |\nu| \le \delta, \tag{21}$$

and that

$$\mathbf{P}_F(\mathbf{x}'\mathbf{a} \ne 0) > .5 \quad \forall\, \mathbf{a} \ne \mathbf{0}. \tag{22}$$

Then μ and σ satisfy (13)–(17).

The theorems are proved in Appendix A.

Remarks.

1. The method used to prove Theorem 2 may be used to show that (13) holds for any M estimate of location.

2. For (18)–(19) to hold, it suffices that w has a bounded derivative w' such that $|w'(r)| \le c/r^3$. It will be seen that this holds for all w functions used in the article.

3. A sufficient condition for (20) is that $F = (1 - \varepsilon)F_0 + \varepsilon F^*$, where $\varepsilon < 1$ and F^* is any distribution, and F_0 is such that there exist $c > 0$ and $\delta > 0$ such that $F_{0,\mathbf{a}}'(u) > c$ for $|u - \mu_{\mathbf{a}}| \le \delta$. This holds in particular if F is spherical with a positive density in a neighborhood of $\mathbf{0}$. A similar condition can be established for the validity of (21).

3. BIAS AND THE WEIGHT FUNCTION

We shall choose the weight function w to attain high asymptotic bias robustness, in the following sense. Given a

Table 1. Maximum Bias of Some Estimators for Normal Data Contaminated by a Point Mass

p	ε	HR	SR	R	H	Tyler	MVEE	Proj.
2	.05	1.5	1.3	1.3	1.3	1.2	6.3	1.5
	.10	2.3	1.8	1.7	1.7	1.6	14.2	2.3
	.20	5.8	3.9	3.3	3.3	2.8	58.3	5.5
3	.05	1.7	1.4	1.3	1.3	1.3	4.6	1.5
	.10	2.7	2.0	1.8	1.8	1.8	9.5	2.3
	.20	7.4	4.7	3.9	3.9	4.3	32.2	5.5
5	.05	1.8	1.5	1.5	1.5	1.5	3.9	1.5
	.10	3.2	2.3	2.1	2.1	2.5	7.5	2.3
	.20	10.1	6.3	5.2	5.3	∞	23.1	5.5
10	.05	2.3	1.8	1.9	1.8	2.2	3.6	1.5
	.10	4.6	3.2	3.2	3.0	∞	6.7	2.3
	.20	16.5	11.0	9.6	9.3	∞	19.3	5.5
15	.05	2.8	2.2	2.3	2.1	4.5	3.7	1.5
	.10	5.9	4.1	4.4	3.9	∞	6.7	2.3
	.20	22.4	16.7	15.5	13.4	∞	18.9	5.5
20	.05	3.2	2.5	2.8	2.5	∞	3.7	1.5
	.10	7.2	5.1	5.7	4.9	∞	6.8	2.3
	.20	28.2	24.7	21.4	18.0	∞	19.0	5.5

NOTE: p is the dimension and ε the contamination rate. The estimators are: SDE with hard rejection (HR), soft rejection (SR), rational (R) and Huber-type weight functions, and Tyler's estimator, minimum volume ellipsoid and projection-based estimator.

"central" model F_0, we measure the (asymptotic) robustness of $(\mathbf{t}, \mathbf{V})$ by comparing $(\mathbf{t}_0, \mathbf{V}_0) = (\mathbf{t}(F_0), \mathbf{V}(F_0))$ with $(\mathbf{t}(F), \mathbf{V}(F))$, where $F = (1 - \varepsilon)F_0 + \varepsilon G$, G being an arbitrary distribution. An appropriate measure of "bias" is needed. We define the bias of $\mathbf{t}$ as $\text{bias}(\mathbf{t}, F) = (\mathbf{t}(F) - \mathbf{t}_0)' \mathbf{V}_0^{-1}(\mathbf{t}(F) - \mathbf{t}_0)$.

As for the bias of $\mathbf{V}$, we shall be more interested in the estimation of the "shape" of $\mathbf{V}_0$; hence we define $\text{bias}(\mathbf{V}, F) = \varphi(\mathbf{A}\mathbf{V}(F)\mathbf{A}')$, where φ is any measure of nonsphericity and $\mathbf{A}'\mathbf{A} = \mathbf{V}_0^{-1}$. These bias measures are clearly invariant under affine transformations. The simplest measure of nonsphericity of a matrix $\mathbf{W}$ is its condition number $\text{cond}(\mathbf{W})$, defined as the ratio of the largest to the smallest eigenvalues. Another such measure is the likelihood ratio test statistics for nonsphericity (Muirhead 1982),

$$\varphi_0(\mathbf{W}) = (\text{tr } \mathbf{W}/p)^p/\det(\mathbf{W}), \tag{23}$$

where tr denotes the trace.

To be able to obtain numerical results, we must specialize to simpler models. Assume that F_0 is ellipsoidal; that is, F_0 is the distribution of $\mathbf{x} = \mathbf{A}\mathbf{y} + \mathbf{c}$, where $\mathbf{A}$ is nonsingular and $\mathbf{y}$ has a spherical distribution. We shall also assume that G is a point mass, $G = \delta_{\mathbf{x}_0}$. This is the simplest case for numerical computing. It is plausible that this is also the "least favorable" situation, in the sense of being the one yielding highest biases. Adrover (1993) has shown that this is the case for M estimators of multivariate scatter.

Because of the invariance, there is no loss of generality in assuming that $\mathbf{c} = \mathbf{0}$ and $\mathbf{A} = \mathbf{I}$ (i.e., $\mathbf{x}$ is spherical) and $\mathbf{x}_0 = k\mathbf{b}_1$, where $\mathbf{b}_1' = (1, 0, \ldots, 0)$. We have $\mathbf{V}_0 = (\mathbf{A}\mathbf{A}')^{-1} = \mathbf{I}$, and the symmetry of the situation implies that $\mathbf{V}(F) = \text{diag}\{v_1, v_2, \ldots, v_p\}$, with $v_i = v_2$ for $i \geq 2$. Thus bias $(\mathbf{V}, F)$ depends only on v_1/v_2, which is the condition number of $\mathbf{V}(F)$. We shall hence deal with $\text{cond}(V)$ throughout this section.

Given ε and w, we would like to maximize the bias over k and then find the w minimizing this maximum bias. Although we have not been able to do this in general, we have done so for some selected families of w functions. We shall henceforth consider only the bias of $\mathbf{V}$.

The families we shall deal with have a scale parameter c and a shape parameter q; that is, w is of the form $w(r) = w_0(r/c, q)$. We have considered the following families: $w_{hr}(r) = I(r \leq c)$, where $I(\cdot)$ is the indicator function ("hard rejection"); $w_{sr}(r) = (1 - (r/c)^q)I(r \leq c)$ ("soft rejection"); $w_H(r) = I(r \leq c) + (c/r)^q I(r > c)$ ("Huber weights"); and $w_R(r) = 1/(1 + (r/c)^q)$ ("rational weights"). In the limit case $c = 0$, we define $w_H(r) = w_R(r) = 1/r^q$. Note that in all cases, the weight functions tend to w_{hr} when $q \to \infty$.

The assumptions (5), (18), and (19) are satisfied by w_{sr} for all q and by w_H and w_R for $q \geq 2$.

For each ε, $\mathbf{V}(F)$ depends on c, q, and k. Let

$$b(c, k, q, \varepsilon) = \text{cond}(\mathbf{V}(F)),$$

$$b(\varepsilon, q) = \inf_c \sup_k b(c, k, q, \varepsilon). \tag{24}$$

To compute $b(c, k, q, \varepsilon)$, we need to compute (9). If we put $\mu = 0$ and $\sigma(z) = \text{med}(|z|)$, an explicit expression for $r(\mathbf{y}, F)$ is obtained, which is given in Appendix B. With this simplification, we have computed $b(\varepsilon, q)$ numerically (normalizing σ to coincide with the standard deviation at the normal; i.e., taking $\sigma(z) = \text{med}(z)/\Phi^{-1}(.75)$, where Φ is the standard normal cumulative distribution function), assuming that $F_0 = \mathcal{N}_p(\mathbf{0}, \mathbf{I})$ for $p = 2$ through 20 and $\varepsilon = .05$, .10, and .20. The values $q = 2$ and 4 were used for the weight functions.

Table 1 displays the minimax biases corresponding to w_{hr}, w_{sr}, w_R, and w_H, with $q = 2$, for different values of p and ε. Because $q = 4$ yielded always larger biases than $q = 2$, the respective results are omitted. As a comparison, we display

Table 2. Optimal Scale c^* and Least Favorable Contamination k^* for $\varepsilon = .10$

	c^*				k^*			
p	*HR*	*SR*	*R*	*H*	*HR*	*SR*	*R*	*H*
2	2.4	3.2	.5	1.2	2.7	2.6	2.1	1.5
3	2.7	3.7	.3	1.1	3.1	3.0	2.5	2.1
5	3.4	4.4	.5	1.1	3.9	3.5	4.7	3.0
10	4.5	5.5	1.2	1.7	5.1	4.4	14.0	3.7
15	5.3	6.6	1.9	2.4	6.0	5.3	21.2	3.2
20	6.1	7.3	2.4	3.3	6.9	5.8	27.4	3.8

NOTE: The columns correspond to SDE with hard rejection (HR), soft rejection (SR), rational (R) and Huber-type (H) weight functions.

also the maximum biases of Tyler's (1987) "distribution-free" M estimator and of the MVEE, taken from Yohai and Maronna (1990). The last column contains the maximum biases of the projection-based estimator treated by Maronna, Stahel, and Yohai (1992), which do not depend on p.

It is seen that:

- w_{hr} is the weight function yielding the largest biases.
- w_{sr} yields slightly larger biases than w_H or w_R.
- the biases for w_H and w_R are almost identical for $p \leq 10$; for larger p, w_H is slightly better than w_R.
- For $p \leq 15$, the MVEE has much larger biases than the SDE with w_H.
- The bias for w_H exceeds that of the projection estimator for $p \geq 6$.

For each weight function, call $c^* = c(p, \varepsilon)$ the value of c attaining the infimum ("optimum c") and call $k^* = k^*(p, \varepsilon)$ the respective value of k attaining the supremum ("least favorable k") in (24), corresponding to $q = 2$, dimension p, and contamination ε. It turns out that c^* varies very slowly with ε. Because the function (24) is extremely flat as a function of c, $c^*(p, .10)$ yields a bias close to the minimum also for $\varepsilon = .05$ and $.20$. Table 2 displays the values of $c^*(p, .10)$ and $k^*(p, .10)$ corresponding to the four families of weight functions.

4. FINITE-SAMPLE BEHAVIOR

4.1 Numerical Computation

The practical difficulty with the SDE obviously lies in computing r. For $p = 2$, one can take a sufficiently fine grid in S_p; numerical experiments show that 50 equispaced points in $S_p \cap \{x_1 \geq 0\}$ (where x_1 is the first coordinate of $\mathbf{x}$) suffice for all practical purposes. For larger p, however, no feasible method is known. The difficulties are not due solely to the "unsmoothness" of the median and MAD. If we use instead the Student maximum likelihood estimates mentioned earlier (8), it turns out that the corresponding r possesses several local maxima, thus making gradient search methods unfeasible. Stahel (1981) proposed an approximate version of r, based on subsampling. Define $\tilde{r}$ as r in (1), but where the sup is taken for $\mathbf{a}$ in a finite set $\mathcal{A}$, defined as follows. For each subsample $\tilde{\mathbf{X}}$ of size p from $\mathbf{X}$, let $\mathbf{a}$ be the direction orthogonal to the hyperplane containing $\tilde{\mathbf{X}}$; let $\mathcal{A}$ be the set of all these $\mathbf{a}$'s. It is easy to show that $\tilde{r}$ is invariant. Because $\mathcal{A}$ will be too large unless p and n are rather small, one replaces $\mathcal{A}$ by a random subsample of size N, $\mathcal{A}_N$. Call $\tilde{r}_N$ the resulting outlyingness measure.

It is difficult to choose N rationally. The criterion of choosing it large enough to ensure a high probability of getting at least one subsample without outliers does not suffice, for if we knew that the sample contained no outliers, this criterion would choose $N = 1$! What we really need is that the behavior of the resulting estimator be similar to that of the *exact* estimator.

For the case where μ is the median and σ is defined as in (7), numerical experiments show the following. For $p = 2$, there is practically no difference between the "exact" r computed through a grid and $\tilde{r}$, even for n as small as 10. For $p = 4$, taking N between 500 and 1,000 suffices to make $\tilde{r}_N$ sufficiently close to $\tilde{r}$ to make the resulting $(\mathbf{t}, \mathbf{V})$'s rather close. Besides, taking larger N's (even up to 10,000) does not seem to guarantee a higher probability of getting better approximations. For $p = 6$, $N = 1{,}000$ also suffices. Nonetheless, the general problem of computing r efficiently remains open.

Recent results by Woodruff and Rocke (1993) suggest that heuristic methods like genetic and tabu search might offer new possibilities in the computation of high-breakdown estimators in high dimensionality.

4.2 Ensuring a High Breakdown Point

Stahel's algorithm is based on maximizing r_1 for $\mathbf{a}$ ranging not over S_p but over a finite set that may depend on $\mathbf{X}$. We will determine conditions under which the breakdown point of such procedure is the same as that of the estimator using the "true" r. For this, we need to define some notation.

Given a sample $\mathbf{X}$ of size n, and $m \in [0, n]$, denote the set of all samples in which m elements of $\mathbf{X}$ are replaced by arbitrary values by $\mathcal{X}_m$. Recall that the finite-sample replacement breakdown point of an estimator $\mathbf{T}$ at $\mathbf{X}$ is m_0/n, where m_0 is the largest m such that $\mathbf{T}(\mathbf{Z})$ remains in a compact when $\mathbf{Z} \in \mathcal{X}_m$.

Assume that for each $\mathbf{X}$ we have a set $\mathcal{A}(\mathbf{X}) \subseteq S_p$. Let $(\mathbf{t}, \mathbf{V})$ be the SDE with r defined in (1), and let $(\mathbf{t}^*, \mathbf{V}^*)$ be the estimator with r defined as

$$r(\mathbf{y}, \mathbf{X}) = \sup_{\mathbf{a} \in \mathcal{A}(\mathbf{X})} r_1(\mathbf{a}, \mathbf{y}, \mathbf{X}). \qquad (25)$$

Represent a set $\{\mathbf{a}_1, \ldots, \mathbf{a}_p\} \subseteq S_p$ by the $p \times p$ matrix $\mathbf{A}$ with the $\mathbf{a}_i$'s as its columns, and denote the smallest eigenvalue of $\mathbf{A}'\mathbf{A}$ by $\lambda(\mathbf{A})$. Then $\lambda(\mathbf{A}) > 0$ if and only if the elements of $\mathbf{A}$ are linearly independent.

Theorem 3. Let μ and σ be the median and MAD. Let $m = n\varepsilon^*(\mathbf{t}, \mathbf{V}; \mathbf{X})$. Assume that w satisfies (5) and that

$$\inf_{\mathbf{Z} \in \mathcal{X}_m} \sup_{\mathbf{A} \subseteq \mathcal{A}(\mathbf{Z})} \lambda(\mathbf{A}) = \lambda_0 > 0. \qquad (26)$$

Then $\varepsilon^*(\mathbf{t}^*, \mathbf{V}^*; \mathbf{X}) \geq \varepsilon^*(\mathbf{t}, \mathbf{V}; \mathbf{X})$.

The proof is given in Appendix A.

Hence, to ensure that the breakdown point of the approximate estimator is not less than that of the exact estimator, it would suffice that there exist p linearly independent directions in the intersection of all $\mathcal{A}(\mathbf{Z})$'s. To simplify the discussion, assume that $\mathbf{X}$ is in general position and that, moreover, p different subsamples of size p determine p linearly independent directions. Then it would suffice that

($*$) for some $m \geq n\varepsilon^*$, all $\mathbf{Z} \in \mathcal{X}_m$ have in common p subsamples of size p contained in $\mathbf{X}$.

This condition would be fulfilled if $\mathcal{A}(\mathbf{X})$ were obtained by taking all subsamples of size p from a subset of $\mathbf{X}$ of at least $m + p + 1$ elements; but this would usually require an unfeasible computational effort. Following an idea of Rousseeuw (1993), it is easy to show that if $m < n/2$, then splitting the data into blocks of size $2p$ and taking all subsets of size p in each yields at least one subset of size $p + 1$ in $\mathbf{X}$, from which $p + 1$ subsets of size p in $\mathbf{X}$ are obtained. (A simple modification suffices if n is not divisible by $2p$.) This procedure requires $N = n\binom{2p}{p}/(2p)$ directions. Thus for $n = 30$ and $p = 4$, 6, 8, and 10, we have $N = 263$, 2,310, 24,130, and 2.77×10^5.

In view of the large computing effort required, it seems more practical to use simple random subsampling, choosing N large enough to attain for a given ε a probability $> 1 - \varepsilon$ that ($*$) holds. This means that, given $\mathbf{X}$, there is a compact K such that $\Pr((\mathbf{t}(\mathbf{Z}), \mathbf{V}(\mathbf{Z})) \in K) > 1 - \varepsilon$ for all $\mathbf{Z} \in \mathcal{X}_m$ for $m = [n\varepsilon^*]$. A straightforward computation shows that for $m = .5n$, $\varepsilon = .001$, and $p = 4$, 6, 8, and 10, the values of N that imply ($*$) for all n are 210, 1,050, 5,000, and 26,260. The mean number of "good" subsamples is 13, 16, 19, and 26.

If we conform ourselves with $m = .4n$ (i.e., 80% of the maximum breakdown point), then the N's become 100, 350, 1,200, and 3.700. The mean number of good subsamples is about the same. We thus can obtain more manageable N's with only a small loss in robustness. But for a given breakdown point, N grows exponentially with p, implying unavoidable difficulties for larger p.

Note that using these values of N merely implies that the approximate estimator resembles the exact estimator *with respect to the breakdown point.* If one wants the former to be a good approximation to the latter, still-larger values may be needed.

4.3 Simulation

To study the small-sample behavior of the SDE and compare it to other location and scatter estimators, we ran a limited simulation. A smaller-scale exploratory simulation was previously run to try a number of variants.

The following estimators were considered:

1. The SDE, taking for μ the median and for σ the modified MAD described in (7), divided by $\Phi^{-1}((1 + k_1/n)/2)$ to make it approximately unbiased at the normal, and with r computed by subsampling. For $p = 2$, all subsamples were used; for $p > 2$, $N = 1{,}000$ random subsamples were taken. The weight function w_H with $q = 2$ was used. Besides the constant c^* defined at the end of Section 3—which is asymptotically optimal with respect to bias—several multiples of it were used, to also take variability into account. Let SD(b) denote the SDE corresponding to w_H with $q = 2$ and constant $c = bc^*(p, .1)$. The values of b chosen were 1, 1.5, 2, 2.5, and 3 for $p = 2$, plus 4 for $p = 4$ and 6, the purpose being to take a range wide enough to find the most robust estimators.

The results of the exploratory simulation showed that the performance of w_R was very similar to that of w_H and (to our disappointment) that the "redescending" SDE's (i.e., those based on w_{hr} and w_{sr}, and on w_H with $q = 4$) yielded uniformly worse results than w_H with $q = 2$. Hence only the latter was included in the simulation.

2. S estimators. Define in general the "Mahalanobis distances,"

$$d_i = d(\mathbf{x}_i; \mathbf{t}, \mathbf{V}) = (\mathbf{x}_i - \mathbf{t})'\mathbf{V}^{-1}(\mathbf{x}_i - \mathbf{t}). \qquad (27)$$

Let $\mathbf{d}(\mathbf{t}, \mathbf{V}) = (d_i : i = 1, \ldots, n)$. Let s be a scale statistic. The multivariate S estimator $(\mathbf{t}, \mathbf{V})$ is defined as a solution of

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

Given a nonnegative random variable z, define the scale M estimator $s = s(z)$ by $\operatorname{ave}\{\rho(z/s)\} = \delta$, where ρ is a bounded nondecreasing function with $\rho(0) = 0$ and $\rho(\infty) = 1$. Davies (1987) proved that for $(\mathbf{t}, \mathbf{V})$ to attain the maximum breakdown point (6), one must have $\delta = (n - p - 1)/2n$. The (modified) MVEE corresponds to taking $\rho(z) = I(z \geq 1)$; that is, s is the kth order statistic, with $k = [(n + p + 1)/2]$. It will be denoted by MVE.

We have also included a "smooth"-scale M estimator, given by the "biweight" function, defined by $\rho'(z) = (1 - z)^2 I(z \leq 1)$. (The reason that z, and not z^2, appears here, is that the d_i's are the *squared* Mahalanobis distances.) This S estimator will be denoted by SE.

The MVE was computed approximately using the subsampling algorithm (Rousseeuw and Leroy 1987), taking subsamples of size $p + 1$ and choosing all subsamples for $p = 2$ and 1,000 otherwise. The biweight S estimator was computed by first finding an approximate minimum of s by subsampling (with the same number of subsamples as the MVE) and then using it as a starting point to compute a sequence of iteratively reweighted means and covariances converging to a local minimum.

3. Reweighted S estimators. Given the S estimators ($\mathbf{t}$, $\mathbf{V}$), define the d_i's as in (27). Let W be a weight function and let $w_i = W(d_i)$. Define $(\mathbf{t}^*, \mathbf{V}^*)$ as a weighted mean and weighted covariance matrix with weights w_i, as in (3)–(4). We used "hard rejection": $W(d) = I(d \leq d_0)$. The threshold d_0 ranged over a set of values (depending on p) to find the best behavior. A convenient way to do this is to reparameterize d_0 as $d_0 = c_{p,n}\operatorname{med}\{\mathbf{d}\}\chi^2(\beta)/\chi^2(.5)$, where $\chi^2(\beta)$ is the β quantile of the chi-squared distribution with p degrees of freedom and $c_{p,n}$ is a constant, which was taken as the "finite-sample correction" proposed by Rousseeuw and van Zomeren (1991); namely $c_{p,n} = (1 + 15/(n - p))^2$. (The use of the correction does not alter the range of d_0, but only the way to describe it.)

Let RWMVE(β) and RWSE(β) denote the reweighted MVE and SE estimators. The values $\beta = .8$, .9, .95, and .975 were used for RWMVE, and $\beta = .6$, .7, .8, .9, and .95 were used for RWSE. (The reason to take different ranges was again to ensure bracketing an "optimum" β.)

It was observed that SE is more stable with respect to random subsampling than MVE, and that the reweighted estimators are more stable than their "parent" ones.

4. M estimators (Maronna 1976). The Cauchy maximum likelihood (CML) estimator was included. Its finite-sample replacement breakdown point is

$$\varepsilon^* = 1/(p+1) - 1/n \qquad (28)$$

(Tyler 1990). Hence it has an asymptotic breakdown point of $1/(1+p)$.

5. The sample mean and covariance matrix, denoted by MEAN and COV.

The projection estimator of Maronna et al. (1992) was not included because of its substantially greater computing cost.

Two other possibilities examined in the exploratory simulation were using the bisquare weight function for reweighted S estimators (instead of hard rejection) and using the reweighting procedure with SDE. Because the results did not yield any improvement, these variants were not considered any further.

We now describe the *sampling situations.* The dimensions were taken as $p = 2$, 4, and 6. The sample sizes n were 10 and 20 for $p = 2$, 20 for $p = 4$, and 30 for $p = 6$. The reason for these choices, besides time savings, is that a large proportion of high-dimensional data sets observed in practice are rather sparse, with ratios n/p ranging between 3 and 6 and ratios as high as 10 being a fortunate exception.

The distributions used were:

1. The unit normal spherical distribution
2. The Cauchy spherical distribution, chosen as an extreme case of heavy-tailed symmetric situation
3. Contaminated normal samples CN(ε, 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 $m = [n\varepsilon]$ and $\mathbf{b}_1' = (1, 0, \ldots, 0)$. Note that this is not exactly the same as a sample from a contaminated normal distribution.

The values $\varepsilon = .10$ and .20 were chosen. Several values of k were used, searching for the worst behavior of each of the estimators. They ranged between 1 and 15 for $p = 2$, between 2 and 25 for $p = 4$, and between 4 and 25 for $p = 6$. For each n and p, the normal samples used were the same for all ε and k, thus allowing more accurate comparisons across sampling situations. The number of replications was 400 in all cases.

As *performance criteria* for matrices, two measures of nonsphericity were used. One is the nonsphericity statistic φ_0 defined in (23). The other is the condition number of $\mathbf{V}$, cond($\mathbf{V}$). For each sampling situation and each estimator, the mean and median of log φ_0 and log cond were taken as measures of its behavior. Because these include both bias and variability, like the MSE, we shall call them mean (or median) error. The performance criterion for location vectors $\mathbf{t}$ was the mean (or median) of $\|\mathbf{t}\|^2$.

We display the medians (ME) only, for the empirical distributions of both φ_0 and the condition number show not only asymmetry but also sometimes high heavy-tailedness. In view of the bulkiness of the output, Table 3 contains a condensed version of the results. The maximum (over k) ME's for CN(ε, k) are given, together with the corresponding values of k ("worst k").

When $\mathbf{x}$ is spherical normal, it is proved (Muirhead 1982) that n log $\varphi_0(\mathbf{C})$ converges in law. Thus if $\mathbf{V}$ is any of the scatter estimators, then the ratio of the mean or the median error for COV to the corresponding value for $\mathbf{V}$ may be considered a measure of efficiency at the normal. Define the "efficiency" of the estimator $\mathbf{V}$ for the normal (or Cauchy) distribution as ME($\mathbf{V}_0$)/ME($\mathbf{V}$), where $\mathbf{V}_0$ is the COV (or CML) estimator. Because $n\|\bar{\mathbf{x}}\|^2$ also converges in distribution, the efficiency of location vectors is defined likewise. It was considered more clear to display the efficiencies rather than the ME's for both spherical distributions.

Let us say that a location or scatter estimator *dominates* another if it is better with respect to all four criteria: efficiencies for both spherical distributions and maximum ME with $\varepsilon = .10$ and .20. It was observed that for both location and scatter, all RWMVE's were dominated by some RWSE, and the same happened with MVE in most cases. To simplify the table, we proceeded to eliminate all estimators that were dominated by another one from the same family, considering SD's as one family and MVE, RWMVE, SE, and RWSE as another. Within SD's or RWSE's, the definition of dominance was extended to an eventual allowance of .01 in a single entry.

Table 3 displays the results. To help their interpretation, an asterisk indicates the highest efficiencies (not considering COV or CML) and smallest maximum ME's; estimators dominated by some SD are indicated by a dagger (†).

Figure 1 shows the ME's of four scatter estimators: SD(3), SE, CML, and COV, for $p = 4$ and $n = 20$, for CN(.20, k) as a function of k. Note the "redescending" behavior of SE, with a large maximum ME and a low ME for large k.

4.4 Discussion

It is indeed surprising that there are SD's attaining high efficiencies (larger than .8) for *both* the normal and Cauchy distributions.

SE and all RWSE are dominated by some SD for the situations $p = 2$ with $n = 10$ (scatter), $p = 2$ with $n = 20$ (location), and $p = 6$. For $p = 4$, RWSE(.95) attains a higher normal efficiency, at the cost of a much lower Cauchy efficiency and much higher maximum ME's. When $p > 2$, SD's attain always the lowest maximum ME's; and for $p = 2$ they are very close to the minimum.

When $\varepsilon = 2$ and $p = 4$ or 6, the maximum ME's for CML are very large; for $p = 6$ they are even much higher than those for COV or MEAN. The reason is that M estimators may break down due to contamination concentrated on a lower-dimensional hyperplane (in this case, of dimension 0), even if it is not distant (Tyler 1990). This drawback ("breakdown by implosion") is not shared by COV or MEAN.

Note that if $\mathbf{x}$ is normal, then $r^2(\mathbf{x})$ is asymptotically χ^2_p. Hence $\chi^2_p(bc^*)$ is the proportion of observations to which w_H attributes maximum weight. For the values of $(p, b) \in \{(2, 2), (4, 3), (6, 3)\}$, these proportions are .94, .97,

Table 3. Simulation Results

					Contam. normal										*Contam. normal*			
					max ME		*worst k*								*max ME*		*worst k*	
			Spherical efficiency		*ε*								*Spherical efficiency*		*ε*			
p	*n*	*Estimator*	*NOR*	*CAU*	*.10*	*.20*	*.10*	*.20*	*p*	*n*	*Estimator*	*NOR*	*CAU*	*.10*	*.20*	*.10*	*.20*
									scatter								
2	10	SD(2)	.84	.89*	.26	.52	3	6	4	20	SD(2.5)	.75	.91*	.74	2.09*	6	12
		SD(2.5)	.88	.88	.26*	.60	6	15			SD(3)	.82	.91*	.72*	2.15	10	12
		SD(3)	.95*	.81	.28	.72	6	15			SD(4)	.92	.83	.76	2.53	10	12
		RWSE(.8)†	.77	.53	.38	.64	4	4			SE†	.80	.56	.95	6.24	5	12
		RWSE(.9)†	.86	.52	.48	.81	6	6			RWSE(.8)†	.88	.47	1.02	5.56	5	12
		RWSE(.95)†	.94	.52	.55	.91	6	6			RWSE(.95)	.97*	.45	1.04	5.42	5	12
		CML	.72	1.00	.29	.50*	15	15			CML	.73	1.00	.79	9.41	12	25
		COV	1.00	.34	2.00	2.62	15	15			COV	1.00	.17	3.84	9.51	12	25
									location								
		SD(1.5)	.79	.87*	.21	.37*	3	2			SD(2)	.67	.90*	.28	.70*	2	4
		SD(2)	.92	.75	.20	.40	3	3			SD(2.5)	.73	.81	.25*	.78	3	4
		SD(2.5)	.94	.71	.19*	.47	4	4			SD(3)	.80	.75	.25*	.88	3	4
		SD(3)	.95	.65	.19*	.53	4	4			SD(4)	.90	.69	.27	1.03	3	4
		SE†	.87	.73	.20	.49	2	4			MVE†	.41	.19	.66	3.83	4	12
		RWSE(.6)†	.92	.68	.20	.59	4	4			SE†	.85	.63	.28	6.66	4	12
		RWSE(.8)†	.94	.67	.22	.64	5	4			RWSE(.8)	.93	.52	.32	5.77	5	12
		RWSE(.9)	.96*	.55	.26	.86	6	6			RWSE(.95)	.99*	.46	.33	5.57	5	12
		CML	.67	1.00	.25	.40	2	2			CML	.65	1.00	.30	12.6	2	12
		MEAN	1.00	.09	2.30	9.12	15	15			MEAN	1.00	.04	6.41	25.3	25	25
									scatter								
2	20	SD(1.5)	.64	.99*	.16	.46	2	6	6	30	SD(2)	.76	.94*	1.29	3.95*	12	25
		SD(2)	.73	.95	.15	.50	3	6			SD(2.5)	.81	.93	1.26*	4.04	16	25
		SD(2.5)	.81	.90	.15	.58	6	6			SD(3)	.87	.88	1.27	4.23	16	25
		SD(3)	.86*	.85	.18	.68	6	15			SD(4)	.96*	.87	1.50	4.90	16	25
		RWSE(.9)	.77	.49	.19	.44	4	4			SE†	.88	.52	2.92	16.3	8	25
		RWSE(.95)	.82	.52	.23	.51	4	4			RWSE(.95)†	.92	.37	3.06	15.1	8	25
		CML	.73	1.00	.14*	.45*	15	15			CML	.76	1.00	1.64	23.7	20	25
		COV	1.00	.16	1.93	2.57	15	15			COV	1.00	.12	11.26	14.5	25	25
									location								
		SD(1.5)	.81	.90*	.12	.31*	2	2			SD(1.5)	.75	.94*	.30	1.32*	4	4
		SD(2)	.88	.72	.11*	.38	3	3			SD(2)	.80	.85	.29	1.38	4	4
		SD(2.5)	.97	.65	.11*	.46	3	4			SD(2.5)	.86	.73	.28*	1.43	4	4
		SD(3)	.98*	.59	.13	.60	3	4			SD(3)	.90	.65	.30	1.45	4	4
		SE†	.75	.79	.13	.49	2	3			SD(4)	.99*	.53	.32	1.58	4	8
		RWSE(.8)†	.87	.68	.13	.43	3	3			MVE†	.33	.09	1.15	12.5	8	25
		RWSE(.9)†	.93	.65	.13	.55	3	4			SE†	.96	.52	.57	37.1	8	25
		RWSE(.95)†	.97	.61	.15	.61	4	4			RWSE(.95)†	.98	.39	.70	26.5	8	25
		CML	.68	1.00	.14	.33	1	1			CML	.76	1.00	.36	578.0	4	25
		MEAN	1.00	.04	2.34	9.05	15	15			MEAN	1.00	.02	6.53	25.3	25	25

NOTE: (*) indicates maximum efficiencies (excluding MEAN-COV or CML) or minimum max ME's; (†) indicates estimators dominated by some SDE. Efficiencies are given for spherical normal (NOR) and Cauchy (CAU) distributions. The last four columns give, for contaminated normals CN (ε, k), the maximum over k of median errors (max ME), and the value of k at which the maximum is attained (worst k).

and .95. This leads to the conjecture that a good choice for the scale constant would be

$$c = \sqrt{\chi_p^2(\beta)}, \qquad (29)$$

with β about .95.

Using the condition number as a criterion yielded the same qualitative results.

The simulation was run on a 386 PC with 33–55 mHz frequency, using the Gauss-386i system. The complete results may be requested from the authors. The average computing times (in minutes) for the SDE for $n = 30$ and $p = 4$, 6, and 8 are .40, .50, and .65.

5. OUTLIER DETECTION

For the purposes of detecting outlying observations in a sample, one might expect that the outlyingness measure r would be an appropriate tool. We have performed some experiments with normal and contaminated normal samples of dimension 2 and 3, computing r from a grid in the first case and from all subsamples in the second. It turns out that although r does attribute high values to outliers, it also often attributes high values to "good" points (sometimes even as high as those of outliers). The reasons are still obscure to us, but we have found an explanation for at least one important case. Consider a sample with $p = 2$, with some outliers concentrated near a point on the $(1, 0)$-axis. Then for the direction $\mathbf{a} = (0, 1)'$, these outliers will be "inliers," which lower the value of $\sigma(\mathbf{a}'\mathbf{x})$ and hence give higher r's to points with a large ordinate, even if it is not "abnormally" large. Besides, for $\mathbf{a} = (1, 0)'$, $\mu(\mathbf{a}'\mathbf{x})$ will be shifted to the right, hence giving larger r's to points with negative abscissas. This phenomenon was

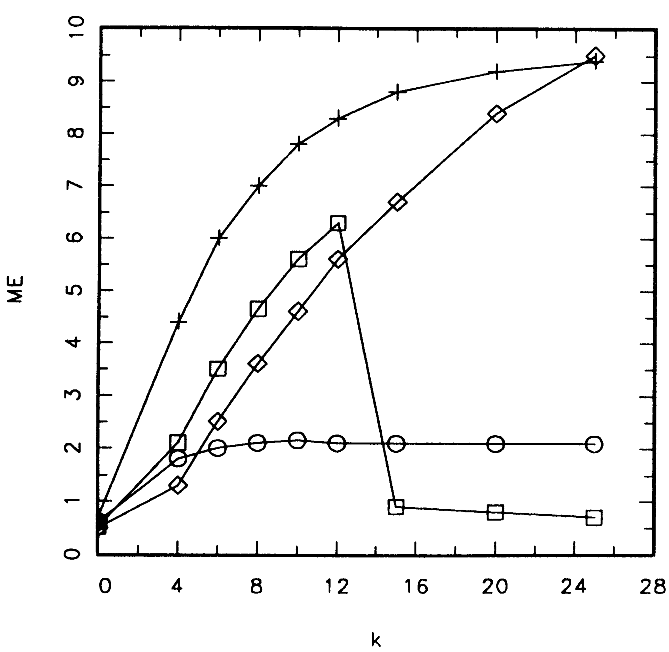

Figure 1. Median Error (ME) as a Function of k, for Contaminated Normal Distributions CN(.2, k) With p = 4 and n = 20, Corresponding to Four Scatter Estimators: CML (+), COV (◇), SE (□), and SDE(3) (○). Note that SE has the smallest ME for large k, but a large maximum ME.

observed for (μ, σ) equal to median and MAD and to the Student MLE mentioned earlier; see (8).

Hence it is better to compute the SDE ($\mathbf{t}$, $\mathbf{V}$) and to use for analysis the Mahalanobis distances d_i, in (27), which do not seem so affected by this drawback. Let $d_{(i)}$ ($i = 1, \ldots, n$) be the ordered d_i's and let $d^*_{(i)} = d_{(i)}/\text{med}\{d_{(\cdot)}\}$. We obtained by simulation the distributions of the $d_{(i)}$'s and of the $d^*_{(i)}$'s. It turned out that the latter were more stable and easier to fit by some familiar distribution, namely $d^*_{(i)} \approx F(i/(n+1); p, n-k)/F(.5; p, n-k)$, with $p \le k \le 2p$, where $F(\beta; l, m)$ is the β quantile of the $\mathcal{F}$ distribution with l and m degrees of freedom.

Define for brevity the "normalized distances" $D_i = F(.5; p, n-p)d^*_{(i)}$ and put $f_i = F(i/(n+1); p, n-p)$. We shall base our analysis on the approximation $D_i \approx f_i$ (for normal data) and on the ratio $R_i = D_i/f_i$.

5.1 An Example With Real Data

Table 4 displays one of the data sets used by Campbell (1989) to locate bushfire scars, which contains satellite measurements on five frequency bands, corresponding to each of 38 pixels. The raw data were kindly supplied by Dr. Campbell.

The following estimators were used to search the data for outlying points: the mean and covariance matrix, CML, RWSE(.95), and SDE using w_H with scale constant c given by (24) with $\beta = .95$. The number of subsamples for both RWSE and SDE was 2,000. To be sure of the validity of the results, in this and the following example the experiment was repeated for several random subsamplings. The results were practically the same.

Table 5 displays for each estimator the cases with largest D_i's, together with the corresponding R_i's. Figure 2 is the plot of $\sqrt{D_i}$ versus $\sqrt{f_i}$ for MEAN-COV and CML. (Square roots rather than raw values are displayed, due to the large differences in magnitudes of D_i's corresponding to different estimators.) Clearly, MEAN-COV does not detect any anomalous points, whereas CML shows two very outstanding cases—8 and 9—followed by 10, 7, and 11, which also have high R_i's (almost 4.0), and by 12 and 32, which may seem suspect. Figure 3 shows the plot corresponding to RWSE and SDE. For the former, cases 8 and 9 are again very outstanding; 7, 10, and 11 seem also outlying; and 12 and 32 could be suspect. SDE does not show exactly the same picture. Cases 8 and 9 are again outstanding, but they are now followed by a patch of cases with high and similar D_i's, cases 32–38. They are followed by 10, 7, and 11, with smaller but still large R_i's, and finally by 31 and 12.

Inspection of the data shows that cases 32–38 have very similar values. They have the seven smallest coordinates on variable 2 and the seven largest in variables 4 and 5; they are thus in a "corner" of the data set. Cases 7, 10, and 11 are also clustered, as are 8 and 9. Figure 4 (p. 339) is a plot of variables 2 and 3.

The patch of outliers 32–38 is a real life instance of point mass contamination, with a proportion of $\frac{7}{38} = .18$. As this fraction is larger than the breakdown point of CML, we cannot expect that estimator to detect this cluster as anomalous.

We can think that the appearance of 32–38 in the plot for SDE shows some real characteristic of the data, rather than an artifact. To explore this matter further, we repeat the analysis, this time omitting the cases that had been pinpointed by RWSE as outlying, namely 7–11.

Table 6 (p. 339) gives the largest D_i's. Although COV and CML do not detect anything suspicious, RWSE shows 32–38 as very outlying and 31 as suspect. SDE shows also 32–38 as very outlying and 31 and 12 as somewhat suspect. It appears that cases 7–11 have "masked" the cluster 32–38 to RWSE, whereas SDE has not suffered from this drawback.

Table 4. Bushfire Data

	Variables						Variables				
Case	1	2	3	4	5	Case	1	2	3	4	5
1	111	145	188	190	260	20	89	133	380	246	302
2	113	147	187	190	259	21	92	137	362	244	300
3	113	150	195	192	259	22	94	139	355	240	299
4	110	147	211	195	262	23	115	156	231	215	276
5	101	136	240	200	266	24	128	167	181	201	265
6	93	125	262	203	271	25	113	156	242	212	273
7	92	110	46	165	235	26	112	155	236	209	271
8	94	95	29	113	190	27	121	162	224	205	268
9	94	94	29	110	188	28	126	166	248	212	273
10	100	104	21	133	208	29	136	174	259	217	278
11	108	115	17	144	215	30	146	177	203	212	272
12	134	156	10	163	233	31	136	155	322	246	301
13	149	181	68	180	247	32	103	97	552	320	364
14	108	154	305	222	285	33	80	66	576	340	377
15	81	137	426	251	306	34	79	66	572	340	376
16	86	138	381	246	300	35	79	66	577	341	379
17	89	137	378	246	301	36	78	66	574	342	377
18	88	133	366	244	298	37	78	66	571	343	379
19	88	131	370	243	298	38	78	66	572	344	380

Table 5. Largest D_i's for Bushfire Data

MEAN-COV			CML			RWSE			SDE		
Case	D_i	R_i	Case	D_i	R_i	Case	D_i	R_i	Case	D_i	R_i
38	4.7	.9	29	9.3	1.9	1	5.0	1.0	13	7.7	1.5
35	4.7	.9	30	9.4	1.8	5	5.1	1.0	29	10.4	2.0
16	5.0	.9	13	9.8	1.8	15	5.4	1.0	30	10.8	2.0
36	5.0	.9	34	10.8	1.9	30	5.6	1.0	12	13.3	2.3
15	5.1	.9	33	10.9	1.8	22	6.0	1.0	31	20.8	3.5
10	5.2	.8	36	11.1	1.8	29	6.2	1.0	11	34.3	5.5
1	5.6	.8	37	11.4	1.7	14	6.5	1.0	7	37.7	5.7
11	5.7	.8	35	11.4	1.7	6	7.4	1.1	10	42.8	6.2
31	6.3	.9	38	11.8	1.6	31	9.3	1.3	32	54.7	7.5
30	6.7	.9	31	14.4	1.9	13	11.5	1.5	36	56.4	7.3
13	6.9	.8	12	16.7	2.0	32	19.9	2.4	34	56.5	6.9
6	7.3	.8	32	21.3	2.4	12	26.8	3.1	33	57.9	6.7
12	7.8	.8	11	34.6	3.7	7	64.7	7.0	37	58.3	6.3
32	9.2	.9	7	39.8	4.0	11	128.4	12.8	35	58.9	5.9
8	12.0	1.1	10	41.5	3.8	10	155.2	14.1	38	59.3	5.4
7	12.6	1.0	8	102.3	8.2	8	335.7	27.1	8	104.9	8.5
9	13.1	.9	9	109.2	7.4	9	350.3	23.7	9	112.2	7.6

5.2 An Example With Synthetic Data

To gain a better understanding of the differences in behavior between the SDE and a redescending estimator like RWSE, a synthetic data set was generated. First, 40 independent pseudorandom four-dimensional vectors with distribution $\mathcal{N}_4(\mathbf{0}, \mathbf{I})$ were generated. Then they were transformed to have sample mean $\mathbf{0}$ and sample covariance matrix $\mathbf{I}$. Finally, the last eight observations (20%) were replaced as $\mathbf{x} \rightarrow k + .1\mathbf{x}$, thus forming a cluster centered at a point $(k, 0, 0, 0)'$ on the x_1-axis.

Then RWSE(.95) and SDE with w_H and $q = 2$ were applied to the data, with scale constant c given by (4.5) with $\beta = .95$. The results are shown in Table 7.

When $k = 15$, both estimators are able to detect the outliers, because the D_i's and R_i's corresponding to outliers ($i > 32$) are much larger than those for the "good" data points. RWSE shows a much better performance, because it yields a sharper distinction between good points and outliers (expressed in larger jumps in D_i and R_i at $i = 32$) and lower values of $\|\mathbf{t}\|$ and cond($\mathbf{V}$).

When $k = 14$, we see that although SDE assigns to the outliers high enough values of D_i's and R_i's to make them recognizable, RWSE does not. Thus it suffices to shift the outliers one unit toward the center to make the performance of RWSE change drastically. Besides, although the outliers are not distant enough to let RWSE recognize them, they have a devastating effect on its ability to estimate the location

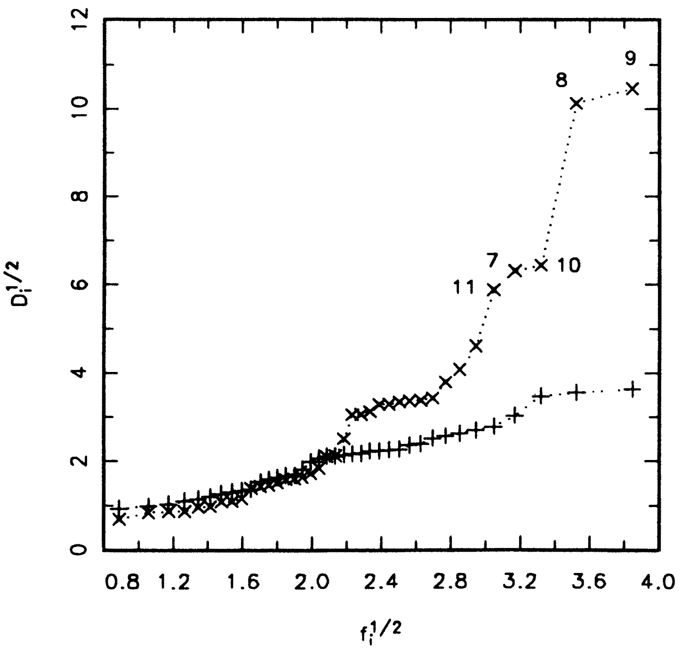

Figure 2. Bushfire Data: Q-Q Plot of Square Roots of Mahalanobis Distances for MEAN-COV (+) and CML (×) (see Table 5). Labeled points correspond to potential outliers.

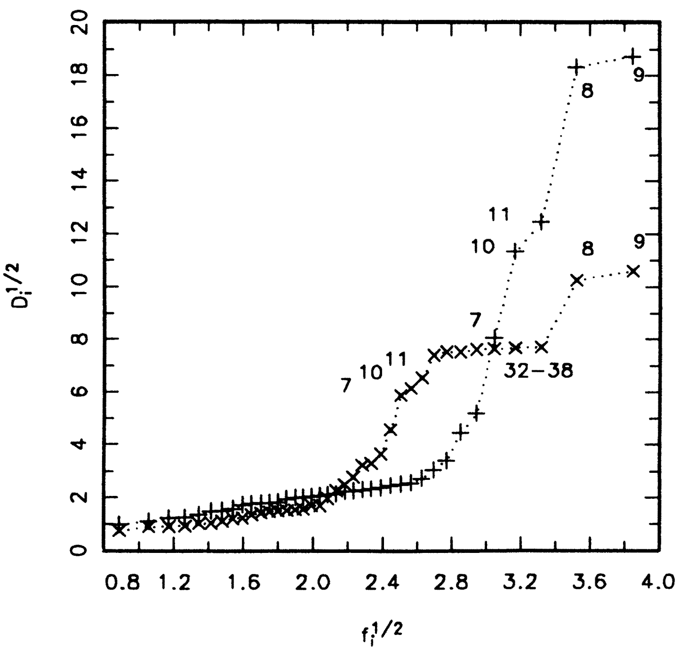

Figure 3. Bushfire Data: Q-Q Plot of Square Roots of Mahalanobis Distances for RWSE (+) and SDE (×) (see Table 5). Labeled points correspond to potential outliers.

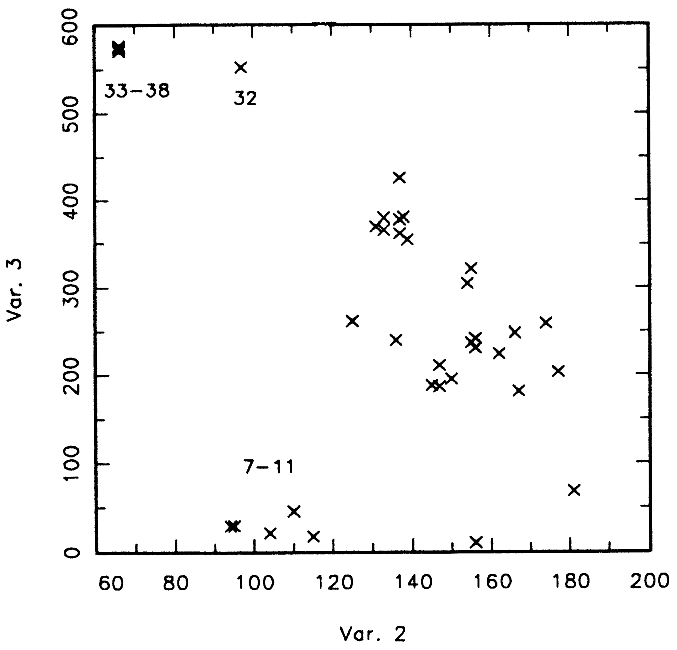

Figure 4. Bushfire Data: Scatterplot of Variables 2 and 3. Points 8 and 9 and 33-38 appear as pasted together.

vector and the covariance matrix of the "good" data, as is shown by the large values of $\|\mathbf{t}\|$ and cond($\mathbf{V}$).

It seems that the redescending character of RWSE, which makes it so successful for very distant contamination, makes it unreliable when the outliers are not so far away. It may be prudent to use *both* types of estimators to gain a better insight on the data. It would be optimum to have estimators sharing the favorable features of the SDE and the redescending character of S estimators. As mentioned in the description of the simulation, neither SDE with redescending w functions nor reweighted SDE yielded good results.

6. CONCLUSIONS

The SDE with "Huber-type" weight function and scale constant (29) is capable of attaining a high efficiency for multivariate normal data, as well as for heavy-tailed distributions like the multivariate Cauchy, while yielding for very asymmetrically contaminated data much lower median errors than S estimators or CML. It is able to detect difficult patterns of multivariate outliers. Its computation is at least not heavier than that of S estimators. We think that it is a good choice for robust and efficient multivariate inference and data analysis.

Table 7. Synthetic Data: p = 4, n = 40, 8 Outliers

k	Estimator	D_{32}	D_{33}	R_{32}	R_{33}	$\|\mathbf{t}\|$	cond(V)
15	SD	11	34	1.8	5.3	.4	15.0
	RWSE	8	180	1.4	28.0	.1	1.7
14	SD	11	28	1.8	4.4	.4	16.0
	RWSE	7	11	1.2	1.7	3.2	166.0

APPENDIX A: PROOFS OF THE THEOREMS

Proof of Theorem 1

We first show that there exist two sequences of random variables A_n and B_n, bounded in probability such that

$$n^{1/2}|r(\mathbf{y}, F_n) - r(\mathbf{y}, F)| \le A_n\|\mathbf{y}\| + B_n \quad \forall\, y \in \Re^p, \qquad (A.1)$$

where r is defined in (9)-(10).

Define

$$A_n = \sup \frac{n^{1/2}|\sigma_{\mathbf{a}} - \sigma_{n\mathbf{a}}|}{\sigma_{n\mathbf{a}}\sigma_{\mathbf{a}}}$$

and

$$B_n = n^{1/2}\sup \frac{|\mu_{\mathbf{a}}||\sigma_{\mathbf{a}} - \sigma_{n\mathbf{a}}| + \sigma_{\mathbf{a}}|\mu_{\mathbf{a}} - \mu_{n\mathbf{a}}|}{\sigma_{n\mathbf{a}}\sigma_{\mathbf{a}}},$$

Assumptions (13)-(17) imply that they are bounded in probability. It is easy to show that

$$n^{1/2}|r_1(\mathbf{y}, \mathbf{a}, F_n) - r_1(\mathbf{y}, \mathbf{a}, F)| \le A_n\|\mathbf{y}\| + B_n,$$
$$\forall\, \mathbf{a} \in S_p, \qquad \forall\, \mathbf{y} \in \Re^p. \qquad (A.2)$$

Given $\varepsilon > 0$, there exists for each $\mathbf{y} \in \Re^p$ and n, $\mathbf{a}_n^*$ such that $r(\mathbf{y}, F_n) \le r_1(\mathbf{y}, \mathbf{a}_n^*, F_n) + \varepsilon n^{-1/2}$ and, by (A.2),

$$n^{1/2}r(\mathbf{y}, F_n) \le n^{1/2}r_1(\mathbf{y}, \mathbf{a}_n^*, F) + A_n\|\mathbf{y}\| + B_n + \varepsilon$$
$$\le n^{1/2}r(\mathbf{y}, F) + A_n\|\mathbf{y}\| + B_n + \varepsilon.$$

Because this inequality holds for all $\varepsilon > 0$, we get

$$n^{1/2}(r(\mathbf{y}, F_n) - r(\mathbf{y}, F)) \le A_n\|\mathbf{y}\| + B_n \quad \forall\, \mathbf{y} \in \Re^p.$$

The same argument shows that

Table 6. Largest D_i's for Bushfire Data Without Cases 7-11

MEAN-COV			CML			RWSE			SDE		
Case	D_i	R_i	Case	D_i	R_i	Case	D_i	R_i	Case	D_i	R_i
29	5.2	.8	35	5.2	.9	29	10.9	1.8	29	9.4	1.5
1	5.4	.8	38	5.3	.8	30	17.6	2.7	6	10.2	1.6
14	5.6	.8	36	5.4	.8	12	25.1	3.7	31	15.8	2.3
5	5.8	.8	14	5.7	.8	31	94.2	13.0	12	17.9	2.5
22	6.3	.8	6	7.5	1.0	32	493.3	63.6	34	32.7	4.2
30	6.4	.8	30	7.6	.9	36	650.7	78.4	33	32.8	4.0
31	7.7	.9	29	8.5	.9	34	654.6	73.2	36	33.1	3.7
6	8.5	.9	13	10.1	1.0	37	660.6	67.9	35	33.5	3.5
13	9.2	.9	31	12.3	1.2	35	662.7	61.7	37	33.8	3.1
32	10.5	.9	32	17.9	1.5	38	666.9	54.7	38	34.6	2.8
12	13.0	.9	12	18.7	1.3	33	668.0	45.4	32	34.6	2.3

$$n^{1/2}(r(\mathbf{y},F_n)-r(\mathbf{y},F)) \ge -A_n\|\mathbf{y}\| - B_n \quad \forall\, \mathbf{y}\in\Re^p,$$

and this proves (A.1).

Next we show that

$$n^{1/2}\sup_{\mathbf{y}\in\Re^p} |w(r(\mathbf{y},F_n)) - w(r(\mathbf{y},F))|\,\|\mathbf{y}\|^k = O_P(1) \qquad 0\le k\le 2. \quad (A.3)$$

Let $C_n = \sup \mu_{n\mathbf{a}}$, $D_n = \inf \sigma_{n\mathbf{a}}$, and $E_n = \sup \sigma_{n\mathbf{a}}$. Assumptions (13)-(17) imply

$$C_n = O_P(1), \qquad \frac{1}{D_n} = O_P(1), \qquad E_n = O_P(1).$$

It is easy to show that

$$r(\mathbf{y},F_n) \ge \frac{\|\mathbf{y}\| - C_n}{E_n} \quad \text{and} \quad r(\mathbf{y},F) \ge \frac{\|\mathbf{y}\| - C}{E}. \quad (A.4)$$

Put $C_n^* = \max(C_n, C)$, $E_n^* = \max(E_n, E)$, and $H_n = \max(2C_n^*, 1)$. Then it follows from (18) and (A.1) that

$$n^{1/2}\sup_{\|\mathbf{y}\|\le H_n} |w(r(\mathbf{y},F_n)) - w(r(\mathbf{y},F))|\,\|\mathbf{y}\|^k \le \gamma H_n^k(H_nA_n + B_n), \qquad 0\le k\le 2. \quad (A.5)$$

Using (19), (A.1) and (A.4) yields

$$\begin{aligned} n^{1/2}&\sup_{\|\mathbf{y}\|\ge H_n} |w(r(\mathbf{y},F_n)) - w(r(\mathbf{y},F))|\,\|\mathbf{y}\|^k \\ &\le \sup_{\|\mathbf{y}\|\ge H_n} \eta E_n^{*3}\frac{(\|\mathbf{y}\|A_n + B_n)\|\mathbf{y}\|^k}{(\|\mathbf{y}\| - C_n^*)^3} \\ &\le \sup_{\|\mathbf{y}\|\ge H_n} \eta E_n^{*3}\frac{(\|\mathbf{y}\|A_n + B_n)\|\mathbf{y}\|^k}{(\|\mathbf{y}\|/2)^3} \\ &\le 8\eta E_n^{*3}(A_n + B_n). \qquad (A.6)\end{aligned}$$

Because A_n, B_n, H_n, and E_n^* are bounded in probability, (A.5) and (A.6) imply (A.3).

To prove the assertions of the theorem, is enough to show that

$$n^{1/2}\left[\frac{1}{n}\sum_{i=1}^n w(r(\mathbf{x}_i,F_n)) - Ew(r(\mathbf{x},F))\right] = O_P(1), \quad (A.7)$$

$$n^{1/2}\left[\frac{1}{n}\sum_{i=1}^n w(r(\mathbf{x}_i,F_n))\mathbf{x}_i - Ew(r(\mathbf{x},F))\mathbf{x}\right] = O_P(1), \quad (A.8)$$

and

$$n^{1/2}\left[\frac{1}{n}\sum_{i=1}^n w(r(\mathbf{x}_i,F_n))\mathbf{x}_i\mathbf{x}_i' - Ew(r(\mathbf{x},F))\mathbf{x}\mathbf{x}'\right] = O_P(1). \quad (A.9)$$

By the central limit theorem, to prove (A.7)-(A.9), it is enough to show that

$$n^{1/2}\left[\frac{1}{n}\sum_{i=1}^n (w(r(\mathbf{x}_i,F_n)) - w(r(\mathbf{x}_i,F)))\right] = O_P(1), \quad (A.10)$$

$$n^{1/2}\left[\frac{1}{n}\sum_{i=1}^n (w(r(\mathbf{x}_i,F_n))\mathbf{x}_i - w(r(\mathbf{x}_i,F))\mathbf{x}_i)\right] = O_P(1), \quad (A.11)$$

and

$$n^{1/2}\left[\frac{1}{n}\sum_{i=1}^n (w(r(\mathbf{x}_i,F_n)) - w(r(\mathbf{x}_i,F))\mathbf{x}_i\mathbf{x}_i')\right] = O_P(1), \quad (A.12)$$

which follow immediately from (A.3).

The following auxiliary result is needed to prove Theorem 2.

Lemma 1. Let $Z_n = \sup_{\mathbf{a}\in S, t\in\Re}|F_{n\mathbf{a}}(t) - F_{\mathbf{a}}(t)|$. Then $Z_n = O_p(1)$.

The proof is an immediate corollary of proposition 1 of Beran and Millar (1986, sec. 4).

Proof of Theorem 2

It is easy to prove (15) and (17); (16) is proved using (22). To prove (13), denote the left side of (13) by Q_n. Then we have to show that $Q_n = O_P(1)$.

Lemma 1 implies that given $\eta > 0$, there exist K_0 and n_0 such that $P(Z_n \ge K_0) \le \eta$ for all $n \ge n_0$. Put $K = K_0/c$, where c is the constant in (20). We shall show that

$$Q_n \ge K \Rightarrow Z_n \ge K_0 \quad \text{for} \quad n \ge n_0. \quad (A.13)$$

If $Q_n \ge K$, then $n^{1/2}(\mu_{n\mathbf{a}} - \mu_{\mathbf{a}}) \ge K$ for some $\mathbf{a}\in S_p$. Using the definition of $\mu_{n\mathbf{a}}$, the definition of Z_n, and assumption (20), it follows that

$$\begin{aligned} .5 \ge F_{n\mathbf{a}}(\mu_{\mathbf{a}} + Kn^{-1/2}) &\ge F_{\mathbf{a}}(\mu_{\mathbf{a}} + Kn^{-1/2}) - Z_n n^{-1/2} \\ &\ge F_{\mathbf{a}}(\mu_{\mathbf{a}}) + cKn^{-1/2} - Z_n n^{-1/2}. \quad (A.14)\end{aligned}$$

Because $F_{\mathbf{a}}(\mu_{\mathbf{a}}) \ge .5$, (A.13) follows. Hence $P(Q_n \ge K) \ge P(Z_n \ge K_0)$ for $n \ge n_0$, which implies that $Q_n = O_P(1)$.

Proof of Theorem 3

It suffices to show that

$$\sup_{\mathbf{Z}\in\mathcal{X}_m}\sup_{\mathbf{a}\in S_p} |\mathbf{a}'\mathbf{t}^*(\mathbf{Z})| < \infty, \qquad \sup_{\mathbf{Z}\in\mathcal{X}_m}\sup_{\mathbf{a}\in S_p} \mathbf{a}'\mathbf{V}^*(\mathbf{Z})\mathbf{a} < \infty,$$

$$\inf_{\mathbf{Z}\in\mathcal{X}_m}\inf_{\mathbf{a}\in S_p} \mathbf{a}'\mathbf{V}^*(\mathbf{Z})\mathbf{a} > 0. \quad (A.15)$$

Proceeding as in the proof of theorem (3.1) of Tyler (1994), it may be shown that

$$\sup_{\mathbf{Z}\in\mathcal{X}_m}\sup_{\mathbf{a}\in\mathcal{A}(\mathbf{Z})} |\mathbf{a}'\mathbf{t}^*(\mathbf{Z})| < \infty,$$

$$\sup_{\mathbf{Z}\in\mathcal{X}_m}\sup_{\mathbf{a}\in\mathcal{A}(\mathbf{Z})} \mathbf{a}'\mathbf{V}^*(\mathbf{Z})\mathbf{a} < \infty. \quad (A.16)$$

It is easy to show that

$$1/\lambda(\mathbf{A}) = \sup_{\|\mathbf{a}\|=1}\{\|\mathbf{c}\| : \mathbf{Ac} = \mathbf{a}\}. \quad (A.17)$$

By (26) and (A.17), given $\mathbf{Z}\in\mathcal{X}_m$, there exist $\mathbf{a}_1, \ldots, \mathbf{a}_p \in \mathcal{A}(\mathbf{Z})$ such that for all $\mathbf{a}\in S_p$, there exists $\mathbf{c}$ with $\|\mathbf{c}\| \le 1/\lambda_0$, such that $\mathbf{a} = \mathbf{Ac}$. This implies the first two assertions in (A.15). The last one is proved as in the proof of Tyler's theorem 3.1.

APPENDIX B: EXPLICIT CALCULATION OF *r*

In this section we give an explicit expression for the outlyingness measure $r(\mathbf{y}, F)$ defined in (9) for the case in which F is spherical, μ is identically 0, and σ is the median of absolute values.

Theorem 4. Let $\mu \equiv 0$ and $\sigma(H) = \text{med}(|z|)$, where z has distribution H. Let F be a spherical distribution in $\Re^p$ and let $F_{\varepsilon,k} = (1-\varepsilon)F + \varepsilon\delta_{k\mathbf{b}_1}$, where $\mathbf{b}_1' = (1, 0, \ldots, 0)$. Denote the distribution function of $|x_1|$ where $(x_1, \ldots, x_p)'$ has distribution F by G, and assume that G^{-1} exists. Let

$$\varepsilon_1 = \frac{1 - 2\varepsilon}{2(1-\varepsilon)} \quad \text{and} \quad \varepsilon_2 = \frac{1}{2(1-\varepsilon)},$$

and let for $i = 1, 2$: $b_i = G^{-1}(\varepsilon_i)$. Define for $t\in[0, 1]$ and $\mathbf{y} = (y_1, \ldots, y_p)'$,

$$R(t, \mathbf{y}) = \frac{t|y_1| + (1 - t^2)^{1/2}\|\mathbf{y}_2\|}{\psi(kt, b_1, b_2)},$$

where $\mathbf{y}_2 = (y_2, \ldots, y_p)'$ and $\psi(s, a, b) = \max(a, \min(s, b))$. Then

$$\begin{aligned} r(\mathbf{y}, F_{\varepsilon,k}) &= \|\mathbf{y}\|/b_1 && \text{if } k \leq b_1, \\ &= \max(R(|y_1|/\|\mathbf{y}\|), R(b_1/k)) && \text{if } k > b_1. \qquad (\text{A}.17) \end{aligned}$$

Proof. If $\mathbf{a} = (a_1, a_2, \ldots, a_p)' \in S_p$, put $\mathbf{a}_2 = (a_2, \ldots, a_p)'$. If $\mathbf{x}$ has distribution $F_{\varepsilon,k}$, then $|\mathbf{a}'\mathbf{x}|$ has distribution function $(1 - \varepsilon)G(s) + \varepsilon I(s > |a_1|k)$. It follows that $\sigma_{\mathbf{a}} = \psi(|a_1|k, b_1, b_2)$. Now we have to maximize $r_1(\mathbf{y}, \mathbf{a}, F_{\varepsilon,k}) = |a_1 y_1 + \mathbf{a}_2'\mathbf{y}_2| / \psi(|a_1|k, b_1, b_2)$ for $a_1^2 + \|\mathbf{a}_2\|^2 = 1$. For each a_1, this is maximized over $\|\mathbf{a}_2\| = (1 - a_1^2)^{1/2}$ by $\mathbf{a}_2 = \text{sign}(a_1 y_1)(1 - a_1^2)^{1/2}\mathbf{y}_2/\|\mathbf{y}_2\|$, yielding $R(|a_1|, \mathbf{y})$. This now must be maximized over $|a_1| \in [0, 1]$. A tedious but straightforward analysis of the sign of $dR(t, \mathbf{y})/dt$ yields (A.17).

[Received May 1993. Revised March 1994.]

REFERENCES

Adrover, J. (1993), "Minimax Bias-Robust Estimation for Multivariate Dispersion Matrices," unpublished manuscript.

Beran, R., and Millar, P. W. (1986), "Confidence Sets for a Multivariate Distribution," *The Annals of Statistics,* 14, 431–443.

Campbell, N. A. (1989), "Bushfire Mapping Using NOAA AVHRR Data," technical report, CSIRO.

Davies, P. L. (1987), "Asymptotic Behavior of S-Estimates of Multivariate Location Parameters and Dispersion Matrices," *The Annals of Statistics,* 15, 1269–1292.

Donoho, D. L. (1982), "Breakdown Properties of Multivariate Location Estimators," Ph.D. qualifying paper, Harvard University.

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

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.

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

Rousseeuw, P. J. (1985), "Multivariate Estimators With High Breakdown Point," in *Mathematical Statistics and Its Applications* (Vol. B), eds. W. Grossman, G. Pflug, I. Vincze, and W. Wertz, Dordrecht, The Netherlands: Reidel, pp. 283–297.

——— (1993), "A Resampling Design for Computing High-Breakdown Regression," *Statistics and Probability Letters,* 18, 125–128.

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

Rousseeuw, P. J., and van Zomeren, B. C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association,* 85, 633–639.

Stahel, W. A. (1981), "Breakdown of Covariance Estimators," Research Report 31, Fachgruppe für Statistik, E.T.H. Zürich.

Tyler, D. E. (1987), "A Distribution-Free M-Estimator of Multivariate Scatter," *The Annals of Statistics,* 15, 234–251.

——— (1990), "Breakdown Properties of the M-Estimators of Multivariate Scatter," technical report, Rutgers University, Dept. of Statistics.

——— (1994), "Finite-Sample Breakdown Points of Projection-Based Multivariate Location and Scatter Statistics," to appear in *The Annals of Statistics,* 22, 1024–1044.

Woodruff, D. L., and Rocke, D. M. (1993), "Heuristic Search Algorithms for the Minimum Volume Ellipsoid," *Journal of Computational and Graphical Statistics,* 2, 69–95.

Yohai, V. J., and Maronna, R. A. (1990), "The Maximum Bias of Robust Covariances," *Communications in Statistics, Part A—Theory and Methods,* 19, 3924–3933.