

A Subsampling Method for the Computation of Multivariate Estimators With High Breakdown Point

Jesus JUAN and Francisco J. PRIETO

All known robust location and scale estimators with high breakdown point for multivariate samples are very expensive to compute. In practice, this computation has to be carried out using an approximate subsampling procedure. In this article we describe an alternative subsampling scheme, applicable to both the Stahel–Donoho estimator and the minimum volume ellipsoid estimator, with the property that the number of subsamples required can be substantially reduced with respect to the standard subsampling procedures used in both cases. We also discuss some bias and variability properties of the estimator obtained from the proposed subsampling process.

Key Words: Minimum volume ellipsoid estimator; Outlier detection; Robust estimation; Stahel–Donoho estimator.

1. INTRODUCTION

Most classical techniques in multivariate analysis are based on the assumption that the observations follow a normal distribution $N(\mu, \Sigma)$, where μ and Σ denote the location and scale parameters of the distribution, respectively. The presence of outliers in the sample can introduce arbitrary modifications in the values of the maximum-likelihood estimators and, consequently, on the results and conclusions of any multivariate analysis technique based on their values.

A measure of the robustness of an estimator is given by its breakdown point ϵ^* (Hampel, Ronchetti, Rousseeuw, and Stahel 1986). For a given sample of size n ,

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}, \quad \mathbf{x}_i \in \mathbb{R}^p$$

assumed to be in general position; that is, having no more than p points laying on any hyperplane of dimension $p-1$, the *breakdown point* of the position estimator T is defined as

$$\epsilon_n^*(T, \mathbf{X}) = \frac{1}{n} \max\{m : \sup_{\mathbf{X}_m} \|T(\mathbf{X}_m)\| < \infty\},$$

Jesus Juan is Associate Professor, Statistics Laboratory, E.T.S.I. Industriales, Univ. Politécnica de Madrid, Spain, jjuan@ccupm.upm.es. Francisco J. Prieto is Associate Professor, Department of Statistics and Econometrics, Univ. Carlos III de Madrid, Spain, fjp@eco.uc3m.es.

©1995 American Statistical Association, Institute of Mathematical Statistics,
and Interface Foundation of North America

Journal of Computational and Graphical Statistics, Volume 4, Number 4, Pages 319–334

where \mathbf{X}_m denotes the sample obtained after replacing m observations randomly chosen from \mathbf{X} with arbitrary values. (For the finite sample size case with replacement, see Donoho and Huber [1983].) For a scatter matrix estimator \mathbf{V} we require instead that $\sup \varphi_0(\mathbf{V}(\mathbf{X})) < \infty$ (see Section 3.3).

The breakdown point for the sample mean and the sample covariance matrix is $\epsilon^* = 0$; that is, it is possible to alter by an arbitrary amount the value of both estimators by modifying just one observation in the sample. As a consequence, it would be of interest to define estimators that are less sensitive to the presence of outliers in the sample, even if that property implies a loss in efficiency. Another condition that is normally required of location and scale estimators is the property of affine equivariance.

A significant improvement in the solution of the robust estimation and outlier identification problems came as a consequence of the introduction of the M estimators (Maronna 1976). These equivariant estimators have a breakdown point smaller than $1/(p+1)$. Unfortunately, this value becomes less satisfactory as the dimension of the problem increases. Stahel (1981) and Donoho (1982) proposed the first robust location and scale estimator with high breakdown point for any dimension of the problem (asymptotically equal to .5). Later on, Rousseeuw (1985) presented the minimum volume ellipsoid estimator, having similar properties.

From a computational point of view, both estimators require a prohibitive amount of time to evaluate, even for small problems. As a consequence, in practice only approximate solutions based on subsampling procedures are computed for both cases. These procedures aim at obtaining subsamples that do not include any outliers. In this article we present a simple subsampling scheme that guarantees a higher probability of obtaining subsamples having this property, and requires a reduced computational effort.

Section 2 briefly describes the two estimators mentioned previously. Section 3 presents the subsampling method that we propose, together with its main properties. Finally, Section 4 discusses some conclusions.

2. HIGH BREAKDOWN POINT ESTIMATORS

2.1 THE STAHEL-DONOHO ESTIMATOR

For a given sample of n observations from \mathbb{R}^p , $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, the Stahel-Donoho location and scale estimator ($\mathbf{T}_{SD}(\mathbf{X}), \mathbf{V}_{SD}(\mathbf{X})$) is defined as

$$\begin{aligned} \mathbf{T}_{SD}(\mathbf{X}) &= \frac{\sum_1^n w_i \mathbf{x}_i}{\sum_1^n w_i} \\ \mathbf{V}_{SD}(\mathbf{X}) &= \frac{\sum_1^n w_i (\mathbf{x}_i - \mathbf{T}_{SD}(\mathbf{X}))(\mathbf{x}_i - \mathbf{T}_{SD}(\mathbf{X}))^T}{\sum_1^n w_i}, \end{aligned} \quad (2.1)$$

where $w_i = w(r_i)$,

$$r_i = \sup_{\mathbf{d} \in S_p} \frac{|\mathbf{d}^T \mathbf{x}_i - \text{med}_j(\mathbf{d}^T \mathbf{x}_j)|}{\text{MAD}_j(\mathbf{d}^T \mathbf{x}_j)}, \quad (2.2)$$

$S_p = \{\mathbf{d} \in \mathbb{R}^p : \|\mathbf{d}\| = 1\}$, and $w(\cdot)$ denotes a weight function (Hampel et al. 1986).

Table 1. Stahel Algorithm: Number of Subsamples N_0 to Attain the Breakdown Point of the Exact Estimator With Probability Equal to P_0

<i>Stahel-Donoho $P_0 = .95$</i>					
$p \setminus \epsilon$.1	.2	.3	.4	.5
4	9	17	30	58	122
6	17	38	87	223	670
8	28	76	225	780	3365
10	42	143	553	2594	16078
20	225	2414	34936	762520	29233500

In this context, r_i provides a measure of how reasonable it is to consider the i th observation, x_i , as an outlier. If x_i is an outlier, for some unidimensional projection, associated to a direction d , the projected observation $d^T x_i$ will also be an outlier. The median and the median of the absolute deviations (MAD) can be used as robust location and scale estimators for the projections, with breakdown points equal to .5. The multivariate robust position and scale estimators are then defined as the weighted sample mean and weighted sample covariance matrix, using weights w_i defined as nonincreasing functions of r_i .

To compute each r_i from (2.2) we would need to solve a global optimization problem with a nonconvex objective function, having in general a large number of local minimizers. The optimization techniques currently available to solve this problem are too inefficient to be of practical use, even for low dimension problems.

To avoid this difficulty, Stahel (1981) proposed to compute an approximation to r_i using the following subsampling procedure: Choose randomly p points from the sample X , and compute a direction orthogonal to the hyperplane defined by the p points, d . Repeat this procedure N_0 times and compute r from (2.2), replacing S_p with this finite set of directions.

The estimator obtained from this procedure is affine equivariant. Maronna and Yohai (1995) show that the breakdown point of the modified estimator coincides with the value for the estimator computed from the exact procedure under certain conditions. Assume that in a sample X we have replaced a number $m = n\epsilon$ of the original points with arbitrary observations; we will denote the modified sample by X_m . The subsampling method guarantees that the estimator will remain bounded for any X_m if in the process we obtain at least p different subsamples that contain no outliers. If the subsampling procedure is perfectly random, the probability of this condition holding is given by

$$P_0 = 1 - \sum_{k=0}^{p-1} \binom{N_0}{p} \left((1 - \epsilon)^p \right)^k \left((1 - (1 - \epsilon)^p) \right)^{N_0 - k}.$$

We assume the probability of generating the same sample twice is negligible.

Table 1 shows the number of subsamples N_0 needed to ensure a probability of success equal to $P_0 = .95$, for different contamination levels ϵ and different dimensions of the problem, p . The number of subsamples required is independent of n , and it grows exponentially with the dimension of the problem.

2.2 THE MINIMUM VOLUME ELLIPSOID ESTIMATOR

Rousseeuw (1985) introduced the minimum volume ellipsoid (MVE) estimator

$$(T_R(\mathbf{X}), V_R(\mathbf{X})),$$

defined as follows: $T_R(\mathbf{X})$ is obtained as the center of the minimum volume ellipsoid containing half the observations, and $V_R(\mathbf{X})$ is the matrix of coefficients of the quadratic form defining the ellipsoid, scaled by a factor to ensure consistency for normal observations. The breakdown point of the MVE estimator is $\epsilon^* = .5$ for all p .

In order to compute the minimum volume ellipsoid for a sample \mathbf{X} with n observations, it would be necessary to consider all the

$$\binom{n}{[n/2] + 1}$$

subsamples of size $[n/2] + 1$ in \mathbf{X} , and then determine the minimum volume ellipsoid for each one of them. The complexity of the computation of the minimum volume ellipsoid makes this procedure infeasible for problem dimensions larger than two. Furthermore, the growth in the number of ellipsoids to be considered makes the method impractical once n becomes sufficiently large.

An approximate solution (Rousseeuw and Leroy 1987; Rousseeuw and van Zomeren 1990) is based on computing a large number of ellipsoids that are not too expensive to generate, and then choosing the one having minimum volume. A subsampling procedure similar to the one described for the Stahel–Donoho estimator can be used to obtain these ellipsoids. This procedure generates N random subsamples of size $p + 1$ from \mathbf{X} ; for each subsample the mean vector \bar{x}_j and the variance matrix V_j are computed, and the ellipsoid defined by $\{x : (x - \bar{x}_j)^T V_j^{-1} (x - \bar{x}_j) \leq 1\}$ is scaled to ensure that it contains $h = [n/2] + 1$ observations (if $h = [(n + p + 1)/2]$ were used, the breakdown point of the estimator would be slightly improved [Davies 1987]).

The number N_1 of subsamples to be generated can be determined from probabilistic arguments. If the breakdown point of the exact estimator must be achieved, we need to have at least one subsample that contains no outliers. If the number of outliers in \mathbf{X} is m and we define $\epsilon = m/n$, the probability of having at least one subsample with this property is given by

$$P_1 = 1 - \left(1 - (1 - \epsilon)^{p+1}\right)^{N_1}.$$

Table 2 shows the value of N_1 for $P_1 = .95$ and different values of the contamination level ϵ and the dimension of the problem p .

2.3 ADDITIONAL CONSIDERATIONS

Other estimators with high breakdown point have been defined: Rousseeuw (1985, p. 291) proposed a variant of the MVE estimator, the minimum covariance matrix determinant estimator (MCD). Davies (1987) suggested some modifications for the MVE

Table 2. Rousseeuw's Algorithm: Number of Subsamples N_1 to Attain the Breakdown Point of the Exact Algorithm With Probability Equal to P_1

<i>MVE $P_1 = .95$</i>					
$p \setminus \epsilon$.1	.2	.3	.4	.5
4	4	8	17	37	95
6	5	13	35	106	382
8	7	21	73	296	1533
10	8	34	150	825	6134
20	26	324	5362	136560	6282506

estimator, while studying its convergence and breakdown point properties for finite samples. Maronna, Stahel, and Yohai (1992) presented an affine equivariant estimator based on projections, having also a breakdown point that is independent of the dimension of the data. The algorithm suggested for the computation of this estimator is based on a subsampling scheme that can also be modified to use the subsampling scheme proposed in the following section.

For robust regression and for the MVE and MCD, Rousseeuw (1993) proposed a sampling procedure that guarantees the generation of estimators with a high breakdown point. In this case the breakdown point is deterministic. That is, the probability that the estimator remains bounded is exactly 1, instead of .95, as in Tables 1 and 2. The adaptation of this procedure to the Stahel–Donoho estimator is discussed in Maronna and Yohai (1995). The procedure divides the n observations into groups of size $2p$, and then analyzes all subsamples of p observations in each group. If the number of outlier observations is smaller than $n/2$, at least $p + 1$ samples will contain no outliers. The number of subsamples generated by this procedure is

$$\frac{n}{2p} \binom{2p}{p},$$

and even for moderate values of n and p this number of subsamples is much higher than the corresponding numbers for equivalent procedures based on probabilistic bounds.

An extensive simulation study conducted by Maronna and Yohai (1995) compares the behavior of most of the methods described in this section, concluding that the Stahel–Donoho estimator has the best bias and variability properties; this estimator is also the most efficient one for outlier identification under a range of different structures in the distribution of the outliers.

The subsampling approximations described in the preceding paragraphs have been defined with the goal of replicating the breakdown point properties of the corresponding exact estimator. Any reasonable approximation to the bias and variability properties of the exact estimators would require a significantly higher number of subsamples. These remarks constitute an additional motivation for the development of subsampling methods that require a reduced number of subsamples, but are able to generate a high proportion of “good” subsamples.

3. PROPOSED SUBSAMPLING ALGORITHM

Let ϵ denote the proportion of outliers in the sample \mathbf{X} ; the probability of a subsample of size p generating a “good” direction for the Stahel–Donoho estimator; that is, the probability of the subsample containing no outliers is given by $(1 - \epsilon)^p$, and for a subsample of size $p + 1$ for the MVE estimator the probability is given by $(1 - \epsilon)^{p+1}$.

The motivation behind the proposed subsampling scheme is to increase the probability of obtaining “good” subsamples, and as a consequence “good” directions from these subsamples. For equal behavior regarding breakdown properties, a method generating a larger number of good directions should have lower computational costs; for equal computational costs it should have better bias and variability properties.

This goal can be achieved by using the following procedure: Construct subsamples of size k , remove from each subsample one observation, and take the remaining $k - 1$ observations as the final subsample to construct the desired estimator. The final subsample will be a “better” subsample than the original one if the probability of removing an outlier from the initial sample is sufficiently high.

Given that our interest is the study of the breakdown point properties for the procedure, we will be primarily concerned with the case in which the outliers are arbitrarily removed from the observations in the uncontaminated sample. In this setting, we now describe a procedure to remove one observation from the subsample having the property that, if the subsample contains just one outlier, then with large probability the outlier will be the observation excluded from the subsample.

If this procedure is used, the probability that the final subsample contains no outliers is given by

$$p(k) = (1 - \epsilon)^k + k(1 - \epsilon)^{k-1}\epsilon. \tag{3.1}$$

This probability is a decreasing function of k , and it would be optimal to choose k as small as possible. The actual value of k will also depend on the procedure used to select the observation to be removed from the subsample. An additional condition on the whole procedure is that it should be computationally efficient.

Let $\bar{\mathbf{x}}_{(i)}$ and $\mathbf{V}_{(i)}$ denote the mean and covariance matrix of the modified subsample, obtained by removing observation \mathbf{x}_i from the subsample of size k . If observation \mathbf{x}_i were the only outlier in the subsample, its distance to the mean, $d_{(i)}$, defined as

$$d_{(i)}^2 = (\mathbf{x}_i - \bar{\mathbf{x}}_{(i)})^T \mathbf{V}_{(i)}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_{(i)}),$$

will typically be larger than $d_{(j)}$ for any $j \neq i$. If \mathbf{x}_i is the only outlier in the subsample, both $\bar{\mathbf{x}}_{(i)}$ and $\mathbf{V}_{(i)}$ are estimators unaffected by the contamination in the sample.

The proposed scheme proceeds by removing the observation having the largest value of $d_{(i)}$. If $\bar{\mathbf{x}}$ and \mathbf{V} denote the sample mean and the sample covariance matrix for the subsample of size k , the Mahalanobis distance for observation i , d_i , given by

$$d_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{V}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}), \tag{3.2}$$

and $d_{(i)}^2$ are related by

$$d_{(i)}^2 = \frac{(k - 2)k^2}{(k - 1)^3} \frac{d_i^2}{1 - kd_i^2/(k - 1)^2}.$$

Table 3. Proposed Method: Number of Subsamples N_2 to Attain the Breakdown Point of the Exact Algorithm With Probability Equal to P_2

Stahel-Donoho $P_2 = .95$					
$p \setminus \epsilon$.1	.2	.3	.4	.5
4	2	3	6	12	26
6	2	5	11	27	84
8	3	7	19	64	278
10	3	10	34	152	943
20	8	61	734	14527	546304

This equality implies that $d_{(i)}^2$ is a monotonically increasing function of d_i^2 ; the largest value of $d_{(i)}$ will be the one corresponding to the largest distance d_i .

For a sample with exactly one outlier, the most powerful test is the one that removes the observation having the largest Mahalanobis distance, d_i .

To apply this procedure we must have a subsample of size at least equal to $k = p + 2$.

3.1 APPLICATION TO THE STAHEL-DONOHO ESTIMATOR

The algorithm that uses the proposed subsampling method to compute the Stahel-Donoho estimator has the following form:

1. Construct N subsamples of size $p + 2$.
2. Remove from each subsample the observation having the largest Mahalanobis distance.
3. Compute the directions orthogonal to each of the $p + 1$ subsets of p observations that can be formed from the final subsample of size $p + 1$.
4. Compute r_i from (2.2), replacing S_p with the set of directions obtained in Step 3.

We now compare this procedure with the subsampling scheme described in Section 2.1, under the condition that both procedures have similar breakdown point properties. We will assume that all outliers are sufficiently removed from the uncontaminated sample, and the probability that the final subsample contains no outliers is given by (3.1) with $k = p + 2$.

If this final subsample contains no outliers, the procedure would compute $p + 1$ “good” directions from each subsample. If we generate N_2 subsamples, the probability of having at least one that contains no outliers after removing the “worst” observation is given by

$$P_2 = 1 - \left(1 - (1 - \epsilon)^{p+2} - (p + 2)(1 - \epsilon)^{p+1}\epsilon \right)^{N_2}.$$

Table 3 shows the number of subsamples N_2 required to have $P_2 = .95$ for different contamination levels ϵ and different dimensions of the data p .

The reduction in the number of subsamples with respect to the values shown in Table 1 is significant. The computations required to determine the $p + 1$ directions for each subsample in the proposed method are naturally more expensive than the computations

Table 4. Ratio of Operations Required by the Stahel Subsampling Algorithm and the Proposed Method

<i>Stahel-Donoho $P_2 = .95$</i>					
$p \backslash \epsilon$.1	.2	.3	.4	.5
4	1.1	1.3	1.2	1.2	1.2
6	1.5	1.4	1.5	1.6	1.5
8	1.4	1.7	1.9	1.9	1.9
10	1.8	1.9	2.3	2.4	2.4
20	2.7	3.9	4.8	5.3	5.4

required by the traditional method, but even if this factor is taken into account (see the Appendix), the proposed method is still more efficient than the traditional subsampling algorithm. In Table 4 we show the ratio of the computational cost required by the Stahel subsampling method and the computational cost of the proposed scheme when both procedures generate the number of subsamples needed to guarantee the breakdown point of the Stahel-Donoho method with probability .95, as shown in Tables 1 and 3. Following Maronna and Yohai (1995), we have assumed $n = 5p$ for all cases. A justification for this choice is that in practice most data sets have ratios between 3 and 6, and it is unusual to encounter cases with values larger than 6.

The reduction shown in the tables is significant for problems of high dimension, and it increases with the dimension, p .

In addition to this improvement in computational performance, another significant advantage of the proposed algorithm is that, by being able to compute $p + 1$ directions from each sample, the average number of “good” directions,

$$N_2(p + 1)((1 - \epsilon)^{p+2} + (p + 2)(1 - \epsilon)^{p+1}\epsilon),$$

is also greatly increased. Stahel’s method generates just one direction per sample, and its expected number of good directions is given by $N_0(1 - \epsilon)^p$. The increase in the expected number of good directions suggests that the estimator obtained after applying the proposed scheme should have better properties than the traditional one.

Table 5 compares the expected number of “good” directions for both methods when $\epsilon = .5$ and the number of subsamples taken for each method are the ones given in Tables 1 and 3, respectively. For values of P larger than .95 the comparison results are even more favorable to the proposed algorithm.

Table 5. Expected Number of “Good” Directions When $\epsilon = .5$ for Stahel’s Method and the Proposed Algorithm

p	<i>Stahel</i>	<i>Proposed</i>
4	8	14
6	10	21
8	13	27
10	16	33
20	28	63

Table 6. Expected Number of Subsamples With No Outliers When $\epsilon = .5$ for Stahel's Method and the Proposed Algorithm. Equal computational effort.

p	Stahel	Proposed
4	8	17
6	10	31
8	13	52
10	16	79
20	28	338

We could also compare the expected number of “good” directions that can be obtained for both methods for the same computational cost. Assume that we compute the number of subsamples given in Table 1 for the Stahel procedure, and that for the proposed algorithm we generate a number of subsamples such that the computational cost is the same. Table 6 gives the average number of good directions generated by Stahel’s method and the proposed algorithm for that fixed computational cost (see the Appendix).

For the case when n is large with respect to p , most of the computational effort is devoted to obtain the projections on the computed directions. Thus, if we neglect the contribution from all other computations, computing time is proportional to the number of generated directions. For this case, if the computational cost is taken to be equal for both procedures, then $N_0 = N_2(p + 1)$, where N_0 and N_2 denote the number of subsamples required by Stahel’s method and the proposed algorithm respectively and, for a breakdown point of 50%, for each good direction obtained via Stahel’s procedure the proposed procedure computes $(p + 3)/4$ directions. Moreover, the confidence level of the estimators obtained by the new method is higher ($P_2 > P_0$).

3.2 APPLICATION TO THE MVE ESTIMATOR

This scheme can also be applied to the MVE estimator in the following manner: Obtain subsamples of size $p + 2$, remove the observation with the largest Mahalanobis distance and compute the elemental ellipsoid corresponding to the remaining $p + 1$ observations. The number of subsamples that are needed to ensure with probability .95 that at least one of them contains no outliers coincide with the values shown in Table 3. Table 7 shows the ratio of the computational costs required by the Rouseeuw and

Table 7. Ratio of Operations Required by the Rouseeuw and van Zomeren Subsampling Algorithm and the Proposed Method

Ratio computational cost MVE. $P = .95$					
$p \setminus \epsilon$.1	.2	.3	.4	.5
4	1.4	1.9	2.0	2.2	2.6
6	1.8	1.9	2.3	2.9	3.3
8	1.8	2.3	2.9	3.5	4.2
10	2.0	2.6	3.4	4.2	5.0
20	2.6	4.2	5.8	7.5	9.2

van Zomeren (1990) method and the proposed method. The computational cost for each subsample is very similar for both procedures (see the Appendix), implying that the gain in computational efficiency when using the proposed algorithm is even more significant than in the case of Stahel's method, see Table 7.

For the number of subsamples required by both methods to attain a given breakdown point with given probability, the expected number of ellipsoids obtained from subsamples with no outliers is similar for both methods and very small (≈ 3 for a probability of .95). This fact may explain the high bias and variability of the MVE estimator, as mentioned in Cook and Hawkins (1990), Maronna, Stahel, and Yohai (1992), and Maronna and Yohai (1995). The proposed subsampling method could be very effective in this sense, as for a given computational cost the expected number of "good" ellipsoids would be increased in the proportion shown in Table 7.

3.3 SIMULATIONS

When the procedure described in this section is applied to the computation of the Stahel–Donoho estimator, it generates $p+1$ directions for each subsample. Each direction is obtained from p points, and any pair of directions from a given subsample shares $p-1$ common points, implying a certain "dependence" structure between the directions. Although the breakdown point is not affected by this fact, it might have some influence on other properties of the estimator, such as its bias or variability.

To analyze the influence of this "dependence" between directions we have conducted a limited simulation study, comparing both subsampling schemes. For a given normal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (this study can be easily extended to any ellipsoidal model) we analyze the effect of an ϵ -contamination, generated from an arbitrary distribution G , on the estimators $(\mathbf{T}_{SD}, \mathbf{V}_{SD})$. Maronna and Yohai (1994) defined as a measure of the bias in the position estimator, $\text{bias}(\mathbf{T}_{SD}, G) = (\mathbf{T}_{SD} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{T}_{SD} - \boldsymbol{\mu})$, and for the variance estimator \mathbf{V}_{SD} , $\text{bias}(\mathbf{V}_{SD}, G) = \varphi(\mathbf{L} \mathbf{V}_{SD} \mathbf{L}^T)$, where φ denotes some measure of nonsphericity and $\mathbf{L}^T \mathbf{L} = \boldsymbol{\Sigma}^{-1}$ (the Cholesky factor of $\boldsymbol{\Sigma}^{-1}$). The most common measure of nonsphericity for a matrix \mathbf{A} is the condition number $\text{cond}(\mathbf{A})$, defined as the square root of the ratio between the largest and smallest singular values of \mathbf{A} . Another measure, used in this simulation study, is

$$\varphi_0(\mathbf{A}) = \frac{(\text{tr}(\mathbf{A})/p)^p}{\det(\mathbf{A})};$$

that is, the ratio between the arithmetic and geometric means of the eigenvalues of \mathbf{A} , raised to the p th power. The lower bound for φ_0 is 1, corresponding to the case in which all eigenvalues are equal (sphericity).

Following Maronna and Yohai (1995) we have chosen:

- The most unfavorable contamination model (all outlier observations are concentrated in one point); a sample of n observations with $n-m$ observations taken from an $N_p(0, \mathbf{I})$ distribution (the affine equivariance property of the estimator implies no lack of generality in taking $\boldsymbol{\mu} = 0$ and $\boldsymbol{\Sigma} = \mathbf{I}$), and m observations concentrated in be_1 , with $m = [n\epsilon]$ and $e_1^T = (1 \ 0 \ \dots \ 0)$.

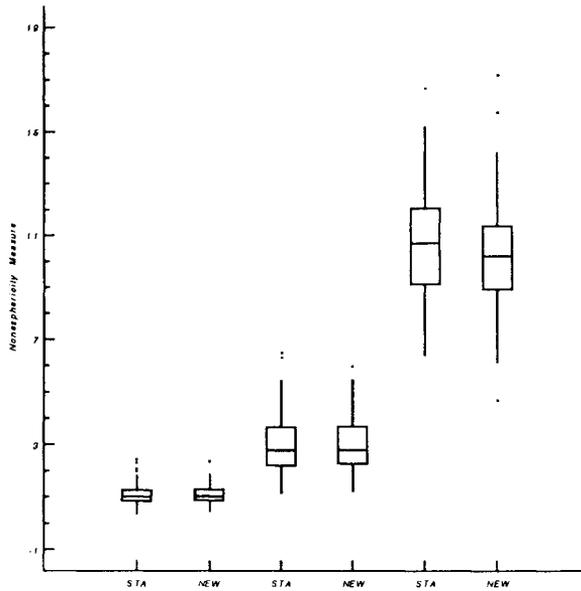


Figure 1. Log of Nonsphericity Measure for the Standard and Proposed Subsampling Schemes. STA=Stahel, NEW=Proposed.

- The Huber function

$$w(r) = I_{\{r \leq c\}} + \frac{c^2}{r^2} I_{\{r > c\}},$$

where $c = \sqrt{\chi_p^2(0.95)}$, as the weight function in (2.1).

Figure 1 shows the boxplot of $\log \varphi_0(V_{SD})$ corresponding to $p = 6$, $n = 30$, and $b = 50$; the contamination level ϵ for the first group of two columns is $\epsilon = .1$, for the second group it is $\epsilon = .2$, and for the third group we used $\epsilon = .3$. The plot was generated from the results of 100 replications of the estimation procedure; each replication was based on the computation of 1,000 directions.

Other values of p , n , ϵ , and b give results similar to the ones shown in Figure 1, both for the position and the scale estimators. This seems to indicate that the close relationship between the directions obtained from a given subsample implies no significant loss in the “quality” of the directions generated by the proposed subsampling method.

4. CONCLUSIONS

Several robust estimators for the position and scale parameters of a multivariate normal sample, with good theoretical properties regarding convergence, efficiency, bias, and breakdown point for highly contaminated samples, have been proposed in the literature. None of these estimators can be computed in exactly the form they have been defined, and all of them must be approximated by procedures based on subsampling schemes. In this article we have presented a new subsampling procedure that requires a significantly

smaller number of subsamples. By taking advantage of this property, it would be possible to obtain a much better estimator at a lower computational cost. The estimators obtained in this manner are able to detect complex contamination patterns in the sample.

APPENDIX

A. EVALUATION OF COMPUTATIONAL COSTS

In Section 3 it was mentioned that the computational costs of the different subsampling schemes should be taken into account when comparing the performance of the procedures. For example, this computational cost must be determined in order to generate the results shown in Tables 4 and 6. In this appendix we evaluate these computational costs for both the Stahel-Donoho estimator and the MVE estimator.

A detailed evaluation should take into account the hardware to be used and details of the implementation of the algorithm; for example, as we are interested only in approximate measures of efficiency, we will only consider in what follows an estimate of the numbers of arithmetic operations (sums and products) required for efficient implementations of the different methods, ignoring the cost of control instructions, comparisons, etc. The numbers of operations for basic numerical procedures can be obtained from standard references on numerical linear algebra (Golub and Van Loan 1989).

We will assume throughout that we have been given a sample X of size n in a space of dimension p .

A.1 THE STAHEL-DONOHO ESTIMATOR

A.1.1 Proposed Procedure

The subsampling procedure proposed in the article would obtain the estimator from the following steps:

1. Select a subsample of $p + 2$ observations.
2. Compute the subsample mean \bar{x} and covariance matrix V .
3. Compute the Mahalanobis distance for each observation in the subsample using (3.2). We first compute the Cholesky factor of the covariance matrix V , L , then solve the system $L^T u_i = x_i - \bar{x}$, and finally form $u_i^T u_i$.
4. Remove from the subsample the observation with the largest Mahalanobis distance.
5. Compute the projections of all points in the sample along the directions orthogonal to each subset of p points from the subsample, d_l , $l = 1, \dots, p + 1$. Let W_{jk} denote the matrix whose rows are the vectors $x_i - x_k$ for some observation k in subsample j and all observations $i \neq k$. The orthogonal direction d_l , $l = 1, \dots, p$, can be obtained as the solution of the system of equations $W_{jk} d_l = e_l$, where e_l is the l th unit vector. We can compute p orthogonal directions as the columns of the matrix D_j solution of the system of equations $W_{jk} D_j = I$. The projections of sample point x_i along these p directions corresponding to subsample j can be obtained as the components of the solution of the system

Table A.1. Operational Costs for the Proposed Procedure (Stahel–Donoho)

Step	Operation	Cost
2	$x_i - \bar{x}$	$2p(p + 2)$
	Covariance matrix	$(p + 2)(p + 1)p$
3	Choleski factorization	$p^3/3$
	Computation of u_j	$(p + 2)p^2$
	Computation of $\ u_j\ ^2$	$2(p + 2)p$
5	LU factorization of W_{jk}	$2p^3/3$
	Solution of $W_{jk}^T q_{ji} = x_i$	$2(p^2 - p)(n - p)$
	$p + 1$ st projection	$p(n - p)$
6	Computation of r_i	$2n$
7	$T_{SD}(X)$	$2np + n$
	$V_{SD}(X)$	$np(p + 1) + 2np$

of equations $W_{jk}^T q_{ji} = x_i$. The $p + 1$ st orthogonal direction is given by $d_k = -\sum_j d_j$, and the corresponding projection can be obtained as $-e^T q_{ji}$. Note that only one observation in the subsample needs to have its projection computed.

6. For each set of projections, compute the median and the MAD, and form the weights r_i from (2.2).
7. Finally, obtain the values of $(T_{SD}(X), V_{SD}(X))$ from (2.1).

Table A.1 summarizes the costs of these steps.

The total cost is given by

$$N_2(2np^2 - np + 2n + p^3 + 10p^2 + 8p) + np^2 + 5np + n,$$

where N_2 denotes the number of subsamples generated by the algorithm.

A.1.2 Stahel’s Procedure

This procedure is similar to the one described previously, except that now the subsample has only p observations, Steps 2, 3, and 4 are not needed, and Step 5 is replaced by

5. Compute the direction orthogonal to all pairs of observations in the subsample. As in the proposed algorithm, let W_{jk} denote the matrix whose rows are the vectors $x_i - x_k$ for some observation k and all observations $i \neq k$ in subsample j . The orthogonal direction d_j can be obtained as a nonzero solution for the system of equations $W_{jk} d_j = 0$, computed from an LU factorization of W_{jk} . Obtain the projections of all sample points onto this direction, $d_j^T x_i$.

The costs of these steps are shown in Table A.2.

If N_0 denotes the total number of subsamples, the number of operations for all steps will be approximately equal to

$$N_0(2np + 2n + \frac{2}{3}p^3 - p^2 - 3p) + np^2 + 5np + n.$$

Table A.2. Operational Costs for Stahel's Procedure (Stahel-Donoho)

Step	Operation	Cost
5	LU factorization of W_{jk}	$p(p - 1)^2 - (p - 1)^3/3$
	Computation of d_j	$2(p - 1)^2 - (p - 1)$
	Computation of $d_j^T x_i$	$2(n - p + 1)p$
6	Computation of r_j	$2n$
7	$T_{SD}(X)$	$2np + n$
	$V_{SD}(X)$	$np(p + 1) + 2np$

A.2 THE MVE ESTIMATOR

A.2.1 Proposed Procedure

The proposed subsampling procedure would have to perform the following operations:

1. Select a subsample of $p + 2$ observations.
2. Compute the subsample mean \bar{x} and covariance matrix V .
3. Compute the Mahalanobis distance for each observation in the subsample using (3.2). Use the Cholesky factor of V .
4. Remove from the subsample the observation with the largest Mahalanobis distance.
5. Compute the mean and covariance matrix for the modified subsample. Update the Cholesky factor.
6. Compute the value of d_i^2 , using (3.2) with \bar{x} and V the values for the subsample, for all points in the sample, and obtain the median of these values d_m .
7. Compute the volume of the ellipsoid from d_m and the determinant of V , from its Cholesky factor.
8. Finally, obtain the values of $(T_R(X), V_R(X))$ from the ellipsoid having minimum volume from all the ones generated in the subsamples.

Table A.3 summarizes the costs of these steps:

If N_3 denotes the number of subsamples considered, the total number of operations for all steps will be approximately equal to

$$N_3(np^2 + 3np + \frac{4}{3}p^3 + 10p^2 + 10p).$$

A.2.2 Rousseeuw and Van Zomeren Procedure

This procedure is very similar to the preceding one, except that now we only have $p + 1$ points in the subsample, and Steps 2, 3, and 4 are no longer needed.

If N_1 denotes the number of subsamples to be taken, after removing the cost of Steps 2, 3, and 4 from the preceding total we obtain

$$N_1(np^2 + 3np + \frac{1}{3}p^3 + 3p).$$

Table A.3. Operational Costs for the Proposed Procedure (MVE)

Step	Operation	Cost
2	$x_i - \bar{x}$	$2p(p + 2)$
	Covariance matrix	$(p + 2)(p + 1)p$
3	Cholesky factorization	$p^3/3$
	Computation of u_i	$(p + 2)p^2$
	Computation of $\ u_i\ ^2$	$2(p + 2)p$
5	Update \bar{x}	$2p$
	Update Cholesky factor	$5p^2$
6	Computation of d_j^2	$(n - p - 1)(p^2 + 3p)$
7	Computation of $\det(V)$	p

ACKNOWLEDGMENTS

We are grateful to the referees for their suggestions and comments, that have contributed to significant improvements in the presentation of the paper. The authors' work was partially supported by CICYT grant ROB91-0244 and DGICYT grant PB93-0232.

[Received July 1994. Revised February 1995.]

REFERENCES

Atkinson, A. C., and Mulira, H.-C. (1993), "The Stalactite Plot for the Detection of Multivariate Outliers," *Statistics and Computing*, 3, 27-35.

Cook, R. D., and Hawkins, D. M. (1990), Comment on "Unmasking Multivariate Outliers and Leverage Points," by P. J. Rousseeuw and B. C. van Zomeren, *Journal of the American Statistical Association*, 85, 640-644.

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," unpublished Ph.D. dissertation, Harvard University, Dept. of Statistics.

Donoho, D. L., and Huber, P. J. (1983), "The Notion of Breakdown Point," in *A Festschrift for Erich L. Lehmann*, eds. P. J. Bickel, K. A. Dorksum, and J. L. Hudges, Jr., Belmont, CA: Wadsworth, pp. 157-184.

Golub, G. H., and Van Loan, C. F. (1989), *Matrix Computations*, Baltimore, MD: The Johns Hopkins University Press.

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. (1976), "Robust M-estimators of Multivariate Location and Scatter," *The Annals of Statistics*, 4, 51-67.

Maronna, R. A., Stahel, W. A., and Yohai, V. J. (1992), "Bias-Robust Estimators of Multivariate Scatter Based on Projections," *Journal of Multivariate Analysis*, 42, 141-161.

Maronna, R. A., and Yohai, V. J. (1995), "The Behavior of the Stahel-Donoho Robust Multivariate Estimator," *Journal of the American Statistical Association*, 90, 330-341.

Rousseeuw, P. J. (1985), "Multivariate Estimators With High Breakdown Point," in *Mathematical Statistics and its Applications* (vol. B), eds. W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, Boston: Reidel, 283-297.

- (1993), "A Resampling Design for Computing High-Breakdown Point Regression," *Statistics and Probability Letters*, 18, 125–128.
- Rousseeuw, P., and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York: John Wiley.
- Rousseeuw, P., and van Zomeren, B. C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, 633–651.
- Stahel, W. A. (1981), "Breakdown of Covariance Estimators," Research Report 31, Fachgruppe für Statistik, E.T.H. Zurich.