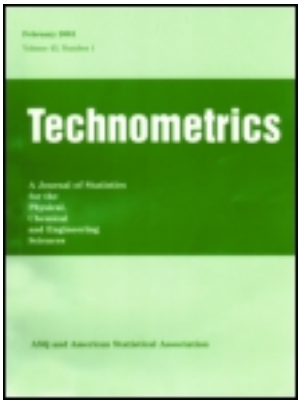

Technometrics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utch20

Robust Estimates of Location and Dispersion for High-Dimensional Datasets

Ricardo A Maronna[a] & Ruben H Zamar[b]
[a] Mathematics Department of the Faculty of Exact Sciences, Universidad Nacional La Plata and Principal Researcher at C.I.C.P.B.A Argentina
[b] Department of Statistics, University of British Columbia, Canada
Published online: 01 Jan 2012.

To cite this article: Ricardo A Maronna & Ruben H Zamar (2002) Robust Estimates of Location and Dispersion for High-Dimensional Datasets, Technometrics, 44:4, 307-317, DOI: 10.1198/004017002188618509

To link to this article: http://dx.doi.org/10.1198/004017002188618509

Robust Estimates of Location and Dispersion for High-Dimensional Datasets

Ricardo A. Maronna

Mathematics Department of the Faculty of Exact Sciences
Universidad Nacional La Plata and
Principal Researcher at C.I.C.P.B.A
Argentina
(*rmaronna@mail.retina.ar*)

Ruben H. Zamar

Department of Statistics
University of British Columbia
Canada
(*ruben@stat.ubc.ca*)

The computing times of high-breakdown point estimates of multivariate location and scatter increase rapidly with the number of variables, which makes them impractical for high-dimensional datasets, such as those used in data mining. We propose an estimator of location and scatter based on a modified version of the Gnanadesikan–Kettenring robust covariance estimate. We compare its behavior with that of the Stahel–Donoho (SD) and Rousseeuw and Van Driessen's fast MCD (FMCD) estimates. In simulations with contaminated multivariate normal data, our estimate is almost as good as SD and clearly better than FMCD. It is much faster than both, especially for large dimension. We give examples with real data with dimensions between 5 and 93, in which the proposed estimate is as good as or better than SD and FMCD at detecting outliers and other structures, with much shorter computing times.

KEY WORDS: Data mining; Minimum covariance determinant; Robust covariances; Stahel–Donoho estimate.

1. INTRODUCTION

It is well known that the sample mean and covariance matrix, which are basic elements of many multivariate procedures, are sensitive to outlying observations. There are several approaches to deal with this problem. M estimates (Maronna 1976) are relatively simple to compute, but their breakdown point (i.e., the maximum proportion of outliers that the estimate can safely tolerate) is at most $1/p$, where p is the dimension of the data. Different approaches have been proposed to overcome this difficulty. Some of them are based on the minimization of a robust scale of Mahalanobis distances: the minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) estimates (Rousseeuw 1984, 1985), S estimates (Davies 1987), and τ estimates (Lopuhaä 1991). Others are based on projections: the Stahel–Donoho estimate (SDE) proposed by Stahel (1982) and Donoho (1981) and studied by Maronna and Yohai (1995); P estimates (Maronna, Stahel, and Yohai 1992); and a recent proposal by Peña and Prieto (2001).

All of these estimates have a high breakdown point for all p; in fact, if conveniently tuned, they may attain the maximum breakdown point for affine-equivariant estimates (Davies 1987). However, their computation requires a heavy effort. Exact computation of the MCD may be performed through heuristic procedures (Agulló 1996), but nevertheless remains feasible only for small datasets. Feasible sets (Hawkins 1994) ensure attaining the solution with probability 1, but are very time-consuming for large p.

Approximate computing is usually based on taking a number N_s of subsamples—generally of size $p+1$—to obtain an initial set of solutions, which are the starting point for the search for a (hopefully global) extremum. Ruppert (1992) developed a heuristic procedure for S estimates.

To ensure a given breakdown point, the value of N_s must increase exponentially with p. A sufficiently high value of N_s is also necessary to ensure stability of the result. In general, all of these methods are feasible for moderate p, but computing them for large p in a reasonable time requires using values of N_s that imply giving up a high breakdown point. Woodruff and Rocke (1993, 1994) proposed procedures to deal with this problem. Recently, Rousseeuw and van Driesen (1999) proposed the "fast MCD" (FMCD), a procedure much more effective than naive subsampling for minimizing the objective function of the MCD, which seems capable of yielding "good" solutions without requiring huge values of N_s. But FMCD still requires substantial running times for large p. Recently, Peña and Prieto (2001) proposed a fast algorithm based on the kurtosis of projections, which does not require subsampling.

Much faster estimates can be computed if one drops the requirements of positive definiteness and affine equivariance. Early proposals of robust procedures are of this type (see Bickel 1964, Sen and Puri 1971). A straightforward approach for multivariate location is to simply calculate a robust location estimate to each individual variable. In the case of multivariate scatter, one can similarly apply a robust covariance or correlation estimate to each pair of variables. Estimates of this type are called "coordinatewise" and "pairwise."

There are many proposals for robust univariate location estimates (see, e.g., Hampel, Ronchetti, Rousseeuw, and Stahel 1986), and also several proposals for the robust estimation of covariance or correlation of a pair of variables. The simplest methods are based on (a) ranks, such as the Spearman's ρ and Kendall's τ (Abdullah 1990); (b) winsorization of the data, such as the quadrant correlation and the "Huberized" covariance estimates (Huber 1981, p. 204); and (c) robustification of

TECHNOMETRICS, NOVEMBER 2002, VOL. 44, NO. 4
DOI 10.1198/004017002188618509

the relationship between variances and covariances, initially proposed by Gnanadesikan and Kettenring (1972) and studied by Devlin, Gnanadesikan, and Kettenring (1981).

Unfortunately, the resulting multivariate location and scatter matrix estimates are not affine equivariant, and the scatter matrix is not guaranteed to be positive definite. Rousseeuw and Molenberghs (1993) proposed several methods to deal with the problem of negative eigenvalues. Note that although the scatter matrices obtained by approaches (a) and (b) are positive definite, they require a correction to make them consistent for normal data, and the correction destroys their positive definiteness.

In this article we present a general method to obtain positive-definite and approximately affine-equivariant robust scatter matrices starting from any pairwise robust scatter matrix. We apply our method to estimates obtained by the aforementioned method (c) to define multivariate location and scatter estimates that are shown to be as good as the equivariant ones reviewed before, while requiring much less computing effort. Although our estimates are not affine equivariant, they are shown to perform well even under very high collinearity. We give some numerical evidence indicating that the lack of equivariance is not a serious concern in our estimates.

We define the estimate in Section 2. In Section 3 we show the results of a simulation study comparing it to the SDE and FMCD under contaminated normal distributions. In Section 4 we treat some high-dimensional real datasets. In Section 5 we deal with the lack of equivariance of the estimates. In Section 6 we compare the computing times of the different estimates, and finally, in Section 7 we discuss the results.

2. THE ESTIMATE

The estimate defined by Gnanadesikan and Kettenring (1972) is based on the identity

$$\operatorname{cov}(X, Y) = \frac{1}{4}\left(\sigma(X+Y)^2 - \sigma(X-Y)^2\right), \tag{1}$$

where σ is the standard deviation and X, Y is a pair of random variables. These authors proposed to define a "robust covariance matrix" by using a robust scale as σ; they used a trimmed standard deviation. The resulting matrix is symmetric, but not necessarily positive semidefinite, and is not affine-equivariant either. Genton and Ma (1999) calculated its influence function and asymptotic efficiency.

Recall that if $\mathbf{V}$ is the covariance matrix of the p-dimensional random vector $\mathbf{x}$ and σ denotes the standard deviation, then

$$\sigma(\mathbf{a}'\mathbf{x})^2 = \mathbf{a}'\mathbf{V}\mathbf{a} \tag{2}$$

for all $\mathbf{a} \in R^p$. The Gnanadesikan–Kettenring estimate forces (2) for a robust scale σ and a small set of directions $\mathbf{a}$. The P estimates of Maronna et al. (1992) attempt to fulfill (2) approximately for *all* directions.

To overcome the lack of positive semidefiniteness, we propose a modification that forces (2) for a set of "principal directions" and is based on the observation that the eigenvalues of the covariance matrix are the variances along the directions given by the respective eigenvectors. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^p$ be a dataset. As a general notation, call $\mathbf{X} = [x_{ij}]$ the $n \times p$ matrix with rows $\mathbf{x}_i'$ $(i = 1, \ldots, n)$ and columns X_j $(j = 1, \ldots, p)$. Let $\sigma(\cdot)$ and $\mu(\cdot)$ be robust univariate dispersion and location statistics, and let $v(\cdot, \cdot)$ be a robust estimate of the covariance of two random variables. We define a scatter matrix $\mathbf{V}(\mathbf{X})$ and a location vector $\mathbf{t}(\mathbf{X})$ as follows:

1. Let $\mathbf{D} = \operatorname{diag}(\sigma(X_1), \ldots, \sigma(X_p))$ and $\mathbf{y}_i = \mathbf{D}^{-1}\mathbf{x}_i$, $i = 1, \ldots, n$.
2. Compute the "correlation matrix" $\mathbf{U} = [U_{jk}]$, applying v to the columns of $\mathbf{Y}$, that is

$$U_{jj} = 1, \qquad \text{and} \qquad U_{jk} = v(Y_j, Y_k), \qquad j \neq k.$$

3. Compute the eigenvalues λ_j and eigenvectors $\mathbf{e}_j$ of $\mathbf{U}$ $(j \equiv 1, \ldots, p)$, and call $\mathbf{E}$ the matrix whose columns are the $\mathbf{e}_j$'s, so that $\mathbf{U} \equiv \mathbf{E}\mathbf{\Lambda}\mathbf{E}'$, where $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$.
4. Let

$$\mathbf{A} = \mathbf{D}\mathbf{E}, \qquad \text{and} \qquad \mathbf{z}_i = \mathbf{E}'\mathbf{y}_i = \mathbf{A}^{-1}\mathbf{x}_i, \tag{3}$$

so that $\mathbf{x}_i = \mathbf{A}\mathbf{z}_i$, and define

$$\mathbf{V}(\mathbf{X}) = \mathbf{A}\mathbf{\Gamma}\mathbf{A}' \qquad \text{and} \qquad \mathbf{t}(\mathbf{X}) = \mathbf{A}\boldsymbol{\nu}, \tag{4}$$

where $\mathbf{\Gamma} = \operatorname{diag}(\sigma(Z_1)^2, \ldots, \sigma(Z_p)^2)$ and $\boldsymbol{\nu} = (\mu(Z_1), \ldots, \mu(Z_p))'$.

The first step makes the estimate scale-equivariant. The other steps are a kind of "principal components," replacing the λ's—which may be negative—by the "robust variances" of the corresponding directions. Another way to view the estimate is to consider that if $\mathbf{U}$ approximates the covariance matrix of $\mathbf{Y}$, then $Z_1, \ldots, Z_p$ should be approximately uncorrelated and hence should have a diagonal covariance matrix (i.e., $\mathbf{\Gamma}$). Likewise, it is better to apply a coordinatewise location estimate to the (approximately uncorrelated) Z_j's, and then transform back to the $\mathbf{X}$ coordinates, than to apply it directly to the X_j's.

We take as v the Gnanadesikan–Kettenring estimator defined in (1), which in step 2 yields

$$U_{jk} = \frac{1}{4}\left[\sigma(Y_j + Y_k)^2 - \sigma(Y_j - Y_k)^2\right], \qquad j \neq k.$$

The resulting estimate is called an *orthogonalized Gnanadesikan–Kettenring* (OGK) estimate.

The procedure can be iterated, computing $\mathbf{V}$ and $\mathbf{t}$ for $\mathbf{Z}$ obtained in step 4, and then expressing them in the original coordinate system, that is

$$\mathbf{V}_{(2)}(\mathbf{X}) = \mathbf{A}\mathbf{V}(\mathbf{Z})\mathbf{A}', \qquad \text{and} \qquad \mathbf{t}_{(2)}(\mathbf{X}) = \mathbf{A}\mathbf{t}(\mathbf{Z}), \tag{5}$$

with $\mathbf{Z}$ and $\mathbf{A}$ defined in (3). Further iterations are defined likewise.

The definition can be extended to include zero scales. If $\sigma(X_j) = 0$, then define $Y_j = 0$ in step 1.

The estimate can be improved on by a reweighting step. Denote in general the Mahalanobis distances by

$$d_i = d(\mathbf{x}_i) = (\mathbf{x}_i - \mathbf{t})'\mathbf{V}^{-1}(\mathbf{x}_i - \mathbf{t}), \tag{6}$$

with $\mathbf{t} = \mathbf{t}(\mathbf{X})$ and $\mathbf{V} = \mathbf{V}(\mathbf{X})$. Let W be a weight function, and define $\mathbf{t}_w$ and $\mathbf{V}_w$ as the weighted mean and covariance matrix, where each $\mathbf{x}_i$ has weight $w_i = W(d_i)$, that is,

$$\mathbf{t}_w = \frac{\sum_i w_i \mathbf{x}_i}{\sum_i w_i} \quad \text{and} \quad \mathbf{V}_w = \frac{\sum_i w_i (\mathbf{x}_i - \mathbf{t}_w)(\mathbf{x}_i - \mathbf{t}_w)'}{\sum_i w_i}. \tag{7}$$

The simplest W is "hard rejection," with $W(d) = I(d \le d_0)$, where $I(\cdot)$ is the indicator function. We take

$$d_0 = \frac{\chi^2_p(\beta)\,\mathrm{med}(d_1, \ldots, d_n)}{\chi^2_p(.5)}, \tag{8}$$

where $\chi^2_p(\beta)$ is the β-quantile of the chi-squared distribution with p degrees of freedom, and "med" denotes the median.

Note that to compute (6) from (4), no matrix inversion is required, because

$$d_i = \sum_j \left(\frac{z_{ij} - \mu(Z_j)}{\sigma(Z_j)} \right)^2.$$

As a general notation, $\mathrm{OGK}_{(l)}$ henceforth denotes the OGK estimate with l iterations, so that $\mathrm{OGK}_{(1)}$ corresponds to the initial estimate (4); $\mathrm{OGK}_{(l)}(\beta)$ denotes the reweighted version (7)–(8), and OGK remains the generic name of the family of estimates.

2.1 Properties

It follows from the definition that $\mathbf{t}$ is shift-equivariant. It is easy to prove that if σ and μ are consistent, then $\mathbf{t}$ and $\mathbf{V}$ in (4) are consistent for the location and shape of elliptical distributions. This is described more precisely in the following proposition.

Proposition 1. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n, \ldots$ be iid with $\mathbf{x}_i = \mathbf{B}\mathbf{u}_i + \mathbf{t}_0$ where $\mathbf{u}_i$ has a spherical distribution. Put for $\mathbf{a} \in R^p$: $X_{\mathbf{a}n} = \{\mathbf{a}'\mathbf{x}_1, \ldots, \mathbf{a}'\mathbf{x}_n\}$. Assume that for all $\mathbf{a}$, the limits in probability of $\mu(X_{\mathbf{a}n})$ and of $\sigma(X_{\mathbf{a}n})$ exist. Then when $n \to \infty$, $\mathbf{t}$ converges in probability to $\mathbf{t}_0$ and $\mathbf{V}$ to $c\mathbf{B}\mathbf{B}'$, where c is a scalar.

It is also easy to show that if the breakdown points of μ and σ (for both implosion and explosion) are not less than ε, then so is the breakdown point of $(\mathbf{t}, \mathbf{V})$ if the data are not collinear. More precisely, let $X = \{x_1, \ldots, x_n\}$ be a univariate sample. For $m \in \{0, \ldots, n\}$ define the "contamination neighborhood" of X as the set of samples of size n having $n - m$ elements in common with X, that is,

$$\mathcal{X}_m = \{\widetilde{X} : \#(\widetilde{X}) = n, \#(\widetilde{X} \cap X) = n - m\}, \tag{9}$$

where $\#(\cdot)$ denotes the cardinality. Then the contamination breakdown point of μ at X is

$$\varepsilon^*(\mu, X) = \frac{1}{n} \max\left\{ m : \sup_{\widetilde{X} \in \mathcal{X}_m} |\mu(\widetilde{X})| < \infty \right\},$$

and the explosion and implosion breakdown points of σ are

$$\varepsilon^*_+(\sigma, X) = \frac{1}{n} \max\left\{ m : \sup_{\widetilde{X} \in \mathcal{X}_m} \sigma(\widetilde{X}) < \infty \right\}$$

and

$$\varepsilon^*_-(\sigma, X) = \frac{1}{n} \max\left\{ m : \inf_{\widetilde{X} \in \mathcal{X}_m} \sigma(\widetilde{X}) > 0 \right\}.$$

Now let $\mathbf{X}$ be a sample of size n in R^p. The breakdown points of $\mathbf{t}$ and $\mathbf{V}$ at $\mathbf{X}$ are

$$\varepsilon^*(\mathbf{t}, \mathbf{X}) = \frac{1}{n} \max\left\{ m : \sup_{\widetilde{\mathbf{X}} \in \mathcal{X}_m} \|\mathbf{t}(\widetilde{\mathbf{X}})\| < \infty \right\}$$

and

$$\varepsilon^*(\mathbf{V}, \mathbf{X}) = \frac{1}{n} \max\left\{ m : 0 < \inf_{\widetilde{\mathbf{X}} \in \mathcal{X}_m} \lambda_1(\mathbf{V}(\widetilde{\mathbf{X}})) < \sup_{\widetilde{\mathbf{X}} \in \mathcal{X}_m} \lambda_p(\mathbf{V}(\widetilde{\mathbf{X}})) < \infty \right\},$$

where $\lambda_1(\mathbf{V})$ and $\lambda_p(\mathbf{V})$ are the smallest and largest eigenvalues of $\mathbf{V}$ and $\mathcal{X}_m$ is defined as in (9) but with $\mathbf{X}$ instead of X. Then for $\mathbf{t}$ and $\mathbf{V}$ in (4), we have the following.

Proposition 2. Assume that

$$\gamma = \frac{1}{n} \sup\{\#\{i : \mathbf{a}'\mathbf{x}_i = c\} : \mathbf{a} \ne \mathbf{0}, \mathbf{c} \in \mathbf{R}\} < 1.$$

Let μ and σ satisfy $\varepsilon^*(\mu, X) \ge \varepsilon$ and $\varepsilon^*_+(\sigma, X) \ge \varepsilon$ for all (univariate) X, and $\varepsilon^*_-(\sigma, X) \ge \varepsilon$ for all X such that $\#\{i : X_i = c\} \le n\gamma$ for all $c \in R$. Then $\varepsilon^*(\mathbf{t}, \mathbf{X}) \ge \varepsilon$ and $\varepsilon^*(\mathbf{V}, \mathbf{X}) \ge \varepsilon$.

The proofs of Propositions 1 and 2 are straightforward and are not given here. Ma and Genton (2001, sec. 4.1) dealt only with the breakdown points of individual covariances computed through (1).

It should be noted that having a high breakdown point is not always an important merit for a nonequivariant estimate. For example, the "robust covariance matrix" defined as $\mathrm{diag}(\mathrm{MAD}(X_1)^2, \ldots, \mathrm{MAD}(X_p)^2)$, where MAD stands for mean absolute deviation, has breakdown .5!

The maximum bias under pointwise contamination has been computed for the MVE and the SDE (Yohai and Maronna 1990; Maronna and Yohai 1995). The lack of equivariance of the OGK makes the study of its bias extremely difficult.

3. SIMULATION

We have run a simulation comparing the SDE, FMCD, and OGK estimates. To evaluate their statistical behavior, we need situations in which the "true values" are known; we have chosen the contaminated multivariate normal model. Because exploring a full neighborhood is infeasible, we focus on point mass contamination; that is, for a sample $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, the first $n - m$ elements are iid multivariate normal, and the remaining m are equal to a fixed vector.

We used the SDE with "Huber weight function" following Maronna and Yohai (1995, p. 334), with threshold

$$b = \sqrt{\chi^2_p(\beta)} \text{ with } \beta = .50. \tag{10}$$

The FMCD was computed through the algorithm of Rousseeuw and van Driesen (1999), followed by a step of hard rejection with $\beta = .975$.

We used the estimates $\mathrm{OGK}_{(l)}$ with $l = 1, 2$ and their reweighted versions $\mathrm{OGK}_{(l)}(\beta)$ with $\beta = .90, .95$, and .975. Because $\beta = .9$ generally yielded the best results, only this case is shown, but $\beta = .95$ was almost as good. Exploratory simulations showed that iterations beyond the second did not lead to improvement. Numerical experiments do not show any convergence when iterating a large number of times.

Because we need robust and efficient scale and location estimates, we chose for σ the "τ scale" of Yohai and

Zamar (1988), which is a truncated standard deviation, and a weighted mean for μ. Define the functions

$$W_c(x) = \left(1 - \left(\frac{x}{c}\right)^2\right)^2 I(|x| \le c) \quad \text{and} \quad \rho_c(x) = \min(x^2, c^2).$$

Let $X = \{x_1, \ldots, x_n\}$ be a univariate sample and put

$$\sigma_0 = \mathrm{MAD}(X) = \mathrm{med}(|X - \mathrm{med}(X)|) \qquad \text{and}$$

$$w_i = W_{c_1}\left(\frac{x_i - \mathrm{med}(X)}{\sigma_0}\right).$$

Then the location and scale statistics are defined as

$$\mu(X) = \frac{\sum_i x_i w_i}{\sum_i w_i} \qquad \text{and}$$

$$\sigma(X)^2 = \frac{\sigma_0^2}{n} \sum_i \rho_{c_2}\left(\frac{x_i - \mu(X)}{\sigma_0}\right). \tag{11}$$

To combine robustness and efficiency, we took $c_1 = 4.5$ and $c_2 = 3$, which yield approximately 80% efficient univariate location and scale for both normal and Cauchy data. Simply using the median and the MAD clearly worsened the simulation results, especially for collinear data. Ma and Genton (2001) advocated using the scale estimate Q_n proposed by Croux and Rousseeuw (1992) and Rousseeuw and Croux (1993), but we prefer (11) for reasons of speed. In the pure normal situation, the results for the sample mean and covariance are also shown.

Unfortunately, the procedure proposed by Peña and Prieto (2001) was not available to us when the simulation study was conducted. A comparison with this method would be of interest.

The sampling situations were p-variate normal ε-contaminated distributions, with p taking the values 5 and 10, and $n = 10p$. In view of the lack of equivariance of the OGK estimate, its behavior may depend on the covariance structure; hence we generated correlated data as follows. Let $m = [n\varepsilon]$ (where $[\cdot]$ denotes the integer part); generate $\mathbf{y}_i$ as p-variate normals $\mathrm{N}_p(\mathbf{0}, \mathbf{I})$ for $i = 1, \ldots, n - m$, and as $\mathrm{N}_p(\mathbf{y}_0, \delta^2\mathbf{I})$ for some $\mathbf{y}_0$ and $i > n - m$; we chose $\delta = .1$. The choice of a normal distribution with a small dispersion, rather than exact point-mass contamination, is due to the fact that exactly repeated points may cause problems with the subsampling algorithms used to compute the SDE and FMCD.

Put $\mathbf{x}_i = \mathbf{R}\mathbf{y}_i$, where $\mathbf{R}$ is the matrix with

$$R_{jj} = 1, \qquad \text{and} \qquad R_{jk} = \rho \qquad \text{for } i \ne j. \tag{12}$$

Then for $\varepsilon = 0$, $\mathbf{X}$ has covariance matrix $\mathbf{R}^2$, and the multiple correlation ρ_{mult} between any coordinate of $\mathbf{X}$ and all of the others is easily calculated as a function of ρ. We chose ρ so that ρ_{mult} took on chosen values. If ρ_{mult} is high, then $\mathbf{X}$ is concentrated around the line with direction $\mathbf{a}_1 = (1, 1, \ldots, 1)'$, the eigenvector of $\mathbf{R}$ corresponding to its largest eigenvalue. We took $\mathbf{y}_0 = k\mathbf{a}_0$, where $\mathbf{a}_0$ is a unit vector. Preliminary simulations suggested that the least favorable direction for OGK is orthogonal to $\mathbf{a}_1$. Given $\mathbf{b}$, take $\mathbf{a}_0 = \mathbf{b} - \mathbf{b}'\mathbf{a}_1/p$ and then normalize it to unit norm. We tried two options, one using a fixed $\mathbf{b}$ with $b_j = (-1)^j$ and the other taking $\mathbf{b}$ at random with a spherical distribution. They yielded similar results, and we report those corresponding to the first option. The value of k ranged over a set of values to search for the least favorable ones for location and scatter, which appear in the table as $k_{\mathbf{t}}$ and $k_{\mathbf{V}}$.

Exploratory simulations were run with different values of ρ_{mult}: 0, .5, .7, .9, and .999. For $\rho_{\mathrm{mult}} = 0$ and .5, OGK behaved surprisingly well (similarly to SDE); but as could be expected, its behavior deteriorated with increasing ρ_{mult}. The reweighted versions were more stable. We show only the results corresponding to the least favorable case, $\rho_{\mathrm{mult}} = .999$. This is a very collinear situation, the ratio of variances of projections orthogonal to $\mathbf{a}_0$ to those along $\mathbf{a}_0$ is .0003 and .0002 for $p = 5$ and 10. This collinearity is much higher than that in the simulations of Devlin et al. (1981) and Ma and Genton (2001). Of course, the value of ρ_{mult} does not affect the other estimates, because they are equivariant.

For each estimate, the location vector $\mathbf{t}$ and scatter matrix $\mathbf{V}$ were computed, and then "back-transformed," $\mathbf{t}_1 = \mathbf{R}^{-1}\mathbf{t}$, $\mathbf{V}_1 = \mathbf{R}^{-1}\mathbf{V}\mathbf{R}^{-1}$, with $\mathbf{R}$ defined in (12). They were evaluated through the distributions of the "errors" $e_t = \|\mathbf{t}_1\|^2$ and $e_V = \log(\mathrm{cond}(\mathbf{V}_1))$ (the decimal logarithm). Their mean and α-quantiles were computed, with $\alpha = .5, .75$, and .90. Only the values corresponding to $\alpha = .75$ are shown; the others yield qualitatively similar results. The condition numbers are more easily displayed in the log scale, because they range between about 3 and 20,000.

The number of subsamples corresponding to $p = 5$ and 10 was 1,000 and 2,000 for SDE and 500 and 1,000 for FMCD. This is probably much larger than needed, but we wanted to see the behavior of these estimates at their best. The number of Monte Carlo replications was 1,000 in all cases. For each combination of n and p, the samples were the same for all estimates and all ε and k. The results are displayed in Table 1.

Discussion. The SDE appears to be the overall best estimate for point contamination, and FMCD appears to be the worst. Among the four variants of OGK, $\mathrm{OGK}_{(1)}(.9)$ seems the best. $\mathrm{OGK}_2(.9)$ is better than $\mathrm{OGK}_{(2)}$ for scatter, but worse for location. The failure of FMCD at $\varepsilon = .2$, $p = 10$ is surprising, as is the high $k = 70$ at which it occurs; at $k = 75$, the values for scale and location drop to .89 and .25.

Reweighting slightly improves on the efficiency of OGK when $\varepsilon = 0$. The efficiency of the reweighted OGK is similar or greater than that of the SDE, and both the reweighted OGK and the SDE are much more efficient than FMCD.

We must remember, however, that normal data with point mass contamination is but one simplified version of the many possibilities reality has to offer. In the next section we explain that with real data, the comparisons may yield results different than those of the simulations.

4. REAL DATA

We analyzed several datasets with p between 5 and 93. Here we show the results for the most interesting ones. For each dataset, we computed the same estimates as in the simulation. Because the reweighted versions of OGK always showed more structure than the raw ones, only the results for $\mathrm{OGK}_{(1)}(.9)$ and $\mathrm{OGK}_{(2)}(.9)$ are displayed, and the "(.9)" is omitted for brevity in this section. The number N_s of subsamples is the default 500 for FMCD and depends on p for SDE. The threshold b of SDE is taken as in (10) for $p \le 10$; because for larger p this

Table 1. Simulation Results

		$p=5, n=50$				$p=10, n=100$			
ε	*Estimate*	e_V	e_t	k_V	k_t	e_V	e_t	k_V	k_t
0	SD(.5)	.55	.16			.54	.14		
	FMCD	1.03	.25			.90	.20		
	OGK	.59	.18			.56	.17		
	OGK(.9)	.57	.17			.57	.15		
	$OGK_{(2)}$	.54	.17			.54	.17		
	$OGK_{(2)}(.9)$	.59	.16			.56	.15		
	Mean-covariance	.48	.13			.53	.13		
.1	SD	.73	.31	5.5	2.5	.90	.36	50	5
	FMCD	1.56	.67	4	4	2.10	2.56	10	10
	OGK	2.49	.54	200	200	2.52	.61	200	200
	OGK(.9)	.81	.32	4	4	.95	.48	5	5
	$OGK_{(2)}$	1.48	.36	200	4	1.68	.42	200	7
	$OGK_{(2)}(.9)$	.87	.38	4	4	1.09	.59	6	6
.2	SD	1.27	1.51	17	3	1.56	2.78	35	5
	FMCD	3.1	22.2	12	15	4.32	515.5	70	70
	OGK	3.35	6.22	200	200	3.46	35.37	200	200
	OGK(.9)	1.58	3.63	9	9	1.70	4.66	10	10
	$OGK_{(2)}$	2.50	5.18	200	15	2.67	3.87	200	9
	$OGK_{(2)}(.9)$	1.99	9.93	15	15	2.24	17.42	20	20

NOTE: e_t and e_V are the error measures for $\mathbf{t}$ and $\mathbf{V}$, equal to the .75 quantiles of $\|\mathbf{t}\|^2$ and of $\log_{10}\text{cond}(\mathbf{V})$. k_t and k_V are the respective contamination locations yielding the highest errors for each estimate.

may yield excessively large values, and hence a less robust estimate, we used

$$b = \min\left(\sqrt{\chi^2_p(.25)}, 4\right).$$

For each estimate $(\mathbf{V}, \mathbf{t})$, call d_i the Mahalanobis distances as in (6), put

$$D_i = \chi^2_p(.5)\frac{d_i}{\text{med}(d_.)},$$

call $D_{(i)}$ the ordered D_i's; and let $f_i = \chi^2_p(i/(n+1))$. Then for normal data, we should have $D_{(i)} \approx f_i$. For each dataset, we plotted D_i versus case i and $D_{(i)}$ versus f_i.

4.1 Bushfire Data (Campbell 1989)

This dataset containing satellite measurements on five frequency bands, corresponding to each of $n = 38$ pixels, was analyzed by Maronna and Yohai (1995). Here $N_s = 500$ for SDE. Figures 1–2 display D_i versus i. All estimates show the same structure, but with different degrees of emphasis. Pixels 32–38 appear as clear outliers, and also 31 to a lesser extent. But $OGK_{(1)}$ gives only faint indications of 7–9, whereas the other estimates clearly point out 7–11 and give some indications for 29 and 30. N. A. Campbell (personal communication) pointed out that the pixels may be classified as "burnt," "unburnt," and "water" and that the suspect ones lie on boundary areas between the classes.

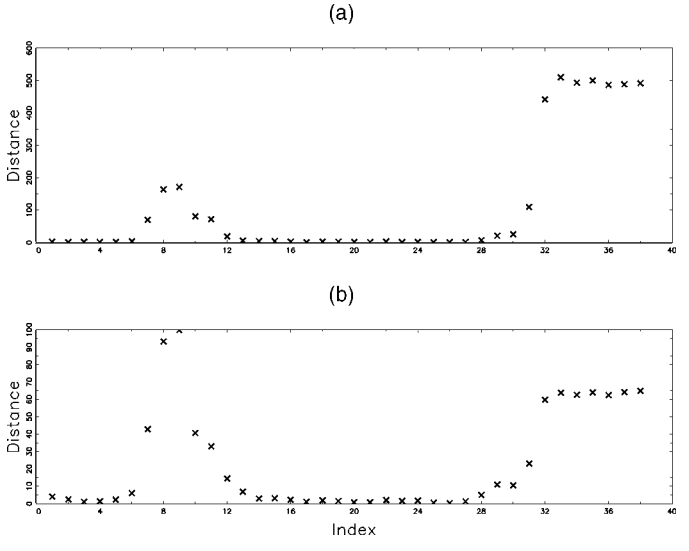

Figure 1. Bushfire Data: Distance D_i Versus Index i for (a) FMCD and (b) SDE.

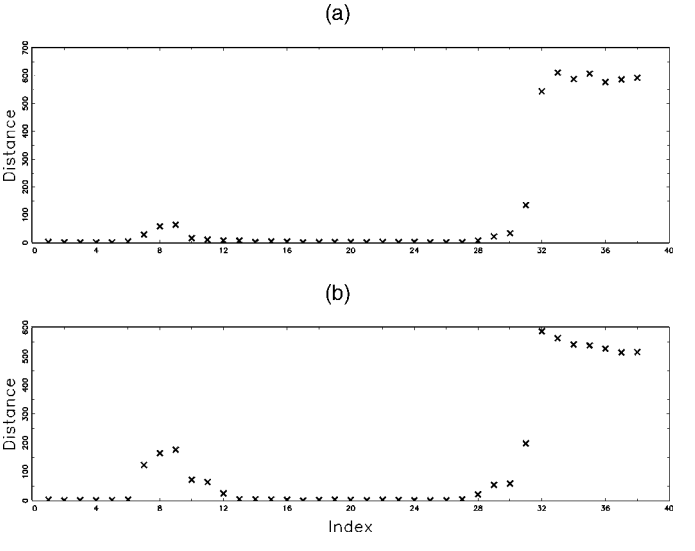

Figure 2. Bushfire Data: Distance D_i Versus Index i for (a) OGK and (b) $OGK_{(2)}$.

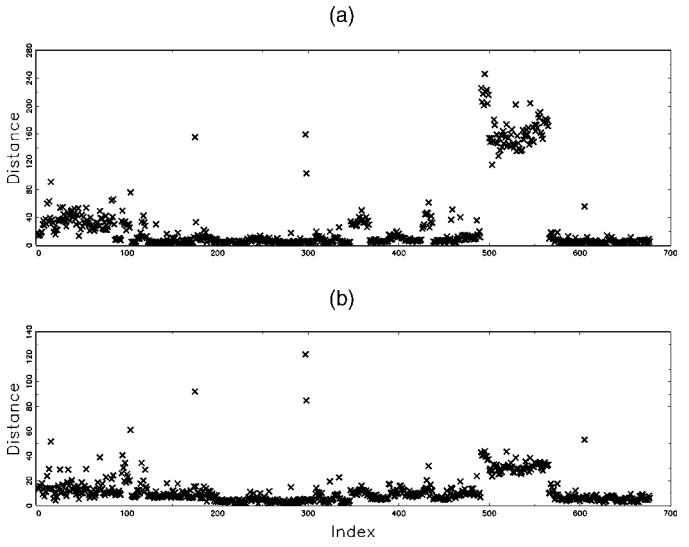

Figure 3. Engineering Data: Distance D_i Versus Index i for (a) FMCD and (b) SDE.

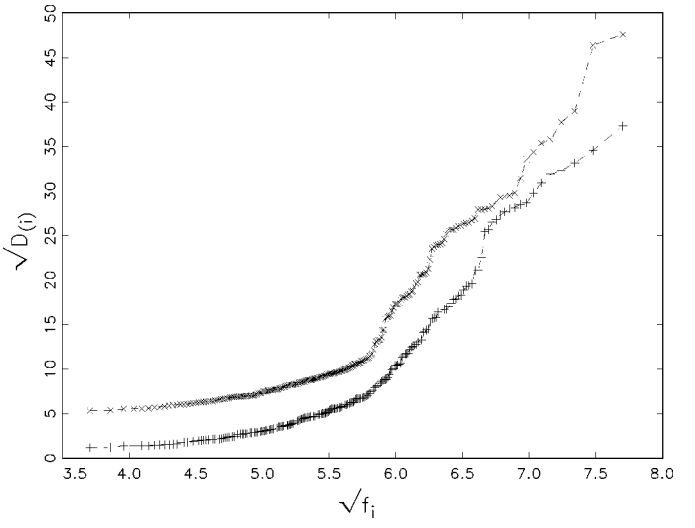

Figure 5. Q-Q Plots of Ionospheric Data for OGK (+) and $OGK_{(2)}$ (×).

4.2 Engineering Data

Rousseuw and van Driesen kindly supplied the data used in their article: nine characteristics measured on $n = 677$ diaphragm parts for TV sets. Here $N_s = 2{,}000$ for SDE. Figures 3–4 show D_i versus i. It is seen that all estimates identify essentially the same structure: some isolated outliers, plus points 491–565, but FMCD and $OGK_{(2)}$ do so more strongly than SDE and $OGK_{(1)}$. The plot for mean-covariance (not shown here) identifies only the isolated outliers.

4.3 Ionospheric Data

This dataset from the Johns Hopkins University Ionosphere database was taken from the "Data Repository" of Bay (1999) and has been used by Sigillito, Wing, Hutton, and Baker (1989). It consists of 351 radar measurements on 34 continuous characteristics, which are the real and imaginary parts of the complex responses corresponding to each of 17 pulse numbers. The measurements are classified as "good" radar returns (those showing evidence of some type of structure in the ionosphere) or "bad" ones. We analyze the $n = 225$ "good" ones. Variables 1, 2, and 27 were omitted from the analysis because they had MAD = 0, so that here $p = 31$. These are very collinear data; the condition numbers of the covariance matrix and of $OGK_{(1)}(.9)$ are about 4,000 and 14,000. We took $N_s = 2{,}000$ for the SDE. Ploting D_i versus i shows no structure. To plot $D_{(i)}$ versus f_i we found the problem of the large range of the former, which prevents us from seeing details in the lower values; hence we plotted the *square roots* of $D_{(i)}$ and f_i. In each plot the curves were slightly displaced to avoid superimposing them. Figures 5–6 show that the data structure is more complex than just "normal data with outliers." The plots for both FMCD and $OGK_{(2)}$ show an almost straight part for $\sqrt{f_i} <$ about 5.7 (the smallest 128 distances), which may describe a "central part"

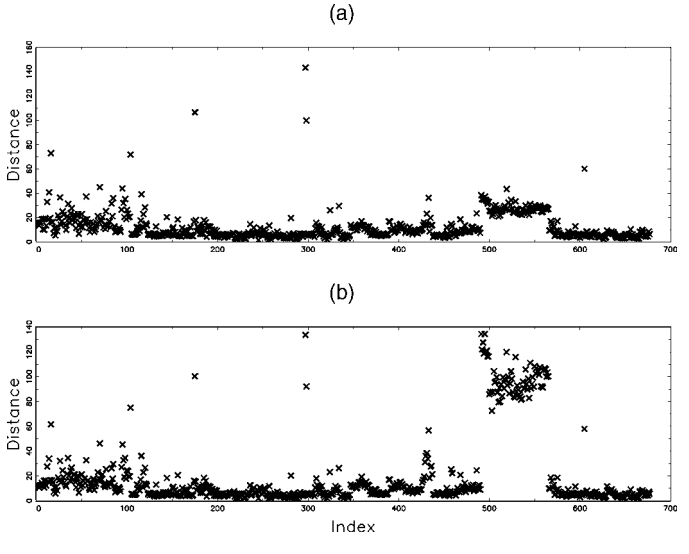

Figure 4. Engineering Data: Distance D_i Versus Index i for (a) OGK and (b) $OGK_{(2)}$.

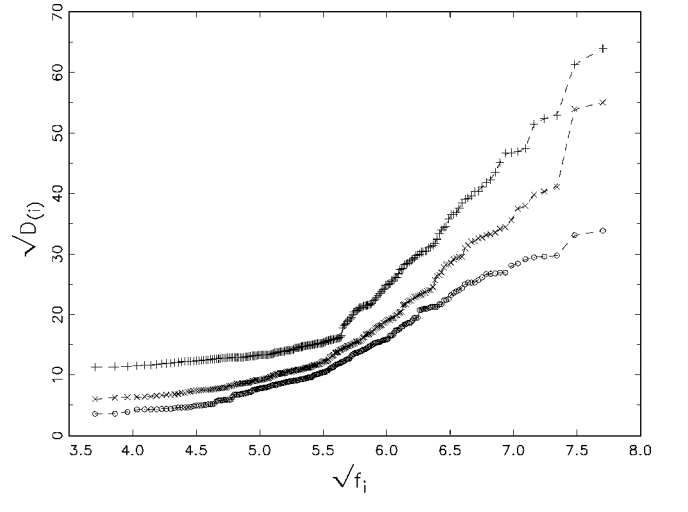

Figure 6. Q-Q Plot of Ionospheric Data for FMCD (+), SDE (×), and cov (○).

Table 2. Ionospheric Data: Points With the Largest Mahalanobis Distances

Estimate	*Points with largest D_i (inverse order)*														
FMCD ($N_s = 500$)	96	95	18	62	26	14	33	27	202	56	116	41	29	119	129
SDE ($N_s = 2{,}000$)	95	96	27	18	62	116	14	26	56	85	41				
OGK(.9)	85	95	84	96	81	83	202	109	214	14	18	203	94	62	130
$OGK_{(2)}(.9)$	95	96	62	14	18	85	202	27	26	41	64	215	81		
Mean-covariance	95	96	62	27	18	116	40	14	26	85	108				

of the data, followed by an abrupt increase. The points with largest distances are given in Table 2.

For a more detailed analysis of the data, we plotted for each observation the sequence of coordinates, but first placing the odd and then the even numbered ones (the real and imaginary parts of the signal). The following features emerged:

a. 138 of the 225 observations have 1 of 4 characteristic forms. Figure 7 plots observations 4, 32, 58, and 79, which are "pure specimens"; most specimens are noisier. Forms (d) and (a) are the most and the least abundant, with 70 and 10 points. Lacking subject matter knowledge, we ignore the physical meaning of the forms.

b. 22 observations look like a mixture of form (b) with (c) or (d).

c. 39 observations look like very noisy versions of type a or b.

d. 26 do not seem to belong to any of the former; these are subjective classifications.

The points with the largest Mahalanobis distances belong to type c or d. Figure 8 shows observations 95, 96, 41, and 27, which are among the first listed in Table 2.

The rank orders of the Mahalanobis distances for all estimates (except mean-covariance) follow same pattern:

- Most points of types c and d are well above rank 128, where the break for FMCD and $OGK_{(2)}$ occurs.
- Most points of form (a) in Figure 7 are just above the break.
- Most points of forms (b), (c), and (d) in Figure 7 and type c are below rank 128.

We can thus conclude that the break in the plots for FMCD and $OGK_{(2)}$ correspond to a real feature of the data and not to an artifact. The other estimates give no hint of this feature.

We remark that this analysis has been made only to demonstrate the behavior of the estimates, and that further analysis and subject matter knowledge are needed to really understand this dataset.

4.4 Spectral Data

This dataset was also taken from Bay (1999). It is part of the Low-Resolution Spectrometer Database in the Infra-Red Astronomy Satellite Project and contains $n = 531$ high-quality spectra measured on $p = 93$ frequency bands. We used $N_s = 3{,}000$ for SDE. The results are displayed in Figures 9 and 10.

The mean and covariances point out only points 210 and maybe 307. Increasing N_s to 10,000 does not change the SDE results very much. FMCD with $N_s = 3{,}000$ yields results similar to $OGK_{(1)}$. Table 3 displays the points with the largest D_i's.

Here, too, $OGK_{(2)}$, $OGK_{(1)}$, and FMCD point out a break. Of the 302 points with $\sqrt{f_i} \le 9.8$, $OGK_{(2)}$ shares 293 with

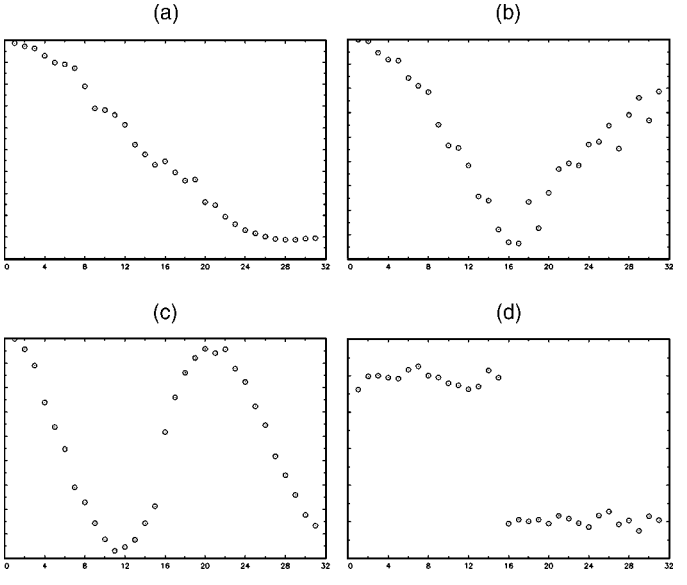

Figure 7. Ionospheric Data: "Pure Specimens." (a) Observation 4; (b) observation 32; (c) observation 58; (d) observation 79.

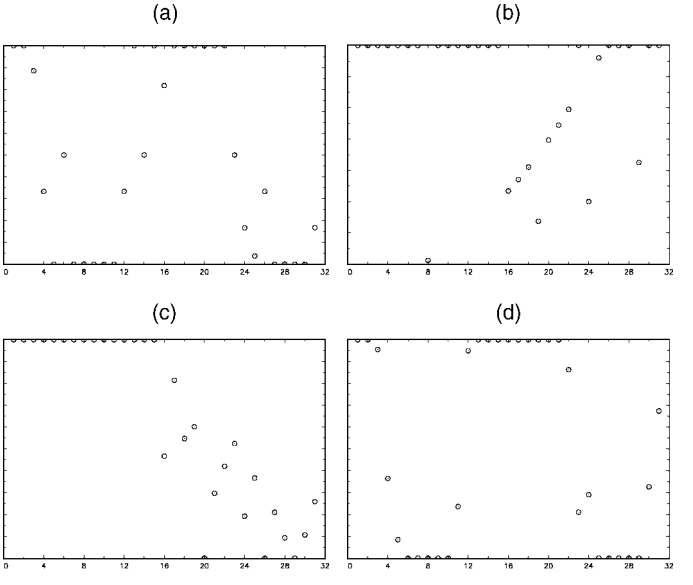

Figure 8. Ionospheric Data: Outliers. (a) Observation 95; (b) observation 96; (c) observation 41; (d) observation 27.

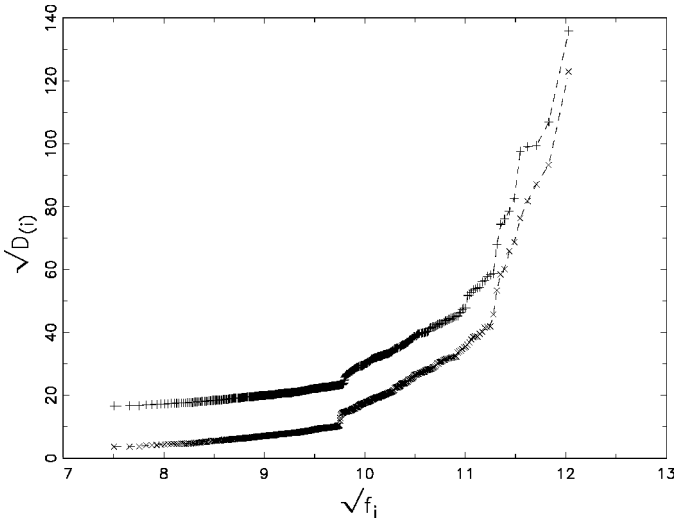

Figure 9. Q-Q Plots of LRS Data for OGK (+) and $OGK_{(2)}$ (×).

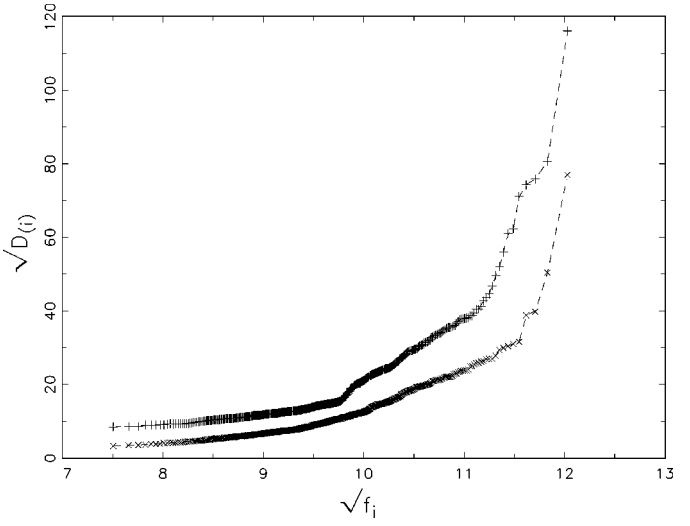

Figure 10. Q-Q Plots of LRS Data for FMCD (+) and SDE (×).

$OGK_{(1)}$, 274 with FMCD, and 291 with SDE. All estimates share 262 points. Of the 20 points with the largest D_i, $OGK_{(2)}$, $OGK_{(1)}$, and FMCD share 16.

For a more detailed analysis, we plotted the sequence of coordinates for each observation. Figures 11(a) and (b) are two typical forms; (c) is a point "just above the break" (with rank order 306), and (d) is an outlier.

Points above the break are clearly different from (a) and (b) like (d), or noisy versions of (a) and (b). We can again conclude that the observed break reveals a real feature of the data.

4.5 Other Datasets

Several other datasets from Bay (1999) were also analyzed, namely Glass ($n = 76$, $p = 7$), Wine ($n = 59$, $p = 13$), VDBC ($n = 357$, $p = 30$), Segment ($n = 330$, $p = 16$), Pima ($n = 500$, $p = 8$) and Sat ($n = 961$, $p = 36$). In all cases, $OGK_{(2)}$ and FMCD yielded similar results, both finding more structure than SDE and $OGK_{(1)}$.

5. EQUIVARIANCE

In this section we investigate the effects of the lack of equivariance of our estimates on their performance. Given $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and a nonsingular $p \times p$ matrix $\mathbf{A}$, let $\mathbf{X_A} = \{\mathbf{Ax}_1, \ldots, \mathbf{Ax}_n\}$. If the estimates are equivariant, then we should have

$$\mathbf{t}(\mathbf{X_A}) = \mathbf{A}\mathbf{t}(\mathbf{X}) \qquad \text{and} \qquad \mathbf{V}(\mathbf{X_A}) = \mathbf{A}\mathbf{V}(\mathbf{X})\mathbf{A}',$$

and hence to explore equivariance we should compare $\mathbf{t}(\mathbf{X})$ and $\mathbf{V}(\mathbf{X})$ with

$$\mathbf{t_A}(\mathbf{X}) = \mathbf{A}^{-1}\mathbf{t}(\mathbf{X_A}) \qquad \text{and} \qquad \mathbf{V_A}(\mathbf{X}) = \mathbf{A}^{-1}\mathbf{V}(\mathbf{X_A})\mathbf{A}^{-1\prime}.$$

Because exploring all transformations is infeasible, we generated random matrices as $\mathbf{A} = \mathbf{TD}$, where $\mathbf{T}$ is a random orthogonal matrix and $\mathbf{D} = \text{diag}(u_1, \ldots, u_p)$, where the u_i's are independent and uniformly distributed in (0,1).

The simulation of Section 3 was repeated for several of the sampling situations. For each generated $\mathbf{X}$ a random $\mathbf{A}$ was generated as described earlier, and the performance of $\mathbf{t_A}$ and $\mathbf{V_A}$ was evaluated. In general, the results were very similar to those for the untransformed estimates. Table 4 shows the results for $p = 5$, $n = 50$, and $\varepsilon = .2$, choosing the "least favorable situations" $k = 200$ and $k = 9$, corresponding to OGK with one and two iterations, and with or without reweighting. The columns "$\mathbf{V}$" and "$\mathbf{t}$" repeat the results of Table 1, and "$\mathbf{t_A}$" and "$\mathbf{V_A}$" correspond to the random transformation as described earlier.

It is seen that the effect of the transformation is stronger on $\mathbf{V}$ than on $\mathbf{t}$. As a general pattern, the reweighted estimators were "more equivariant" in the sense that their performances were much less affected by the transformations.

To investigate the effect of transformations on an individual sample, we define measures of "lack of equivariance" for location and scatter, namely

$$d_t = \|\mathbf{t_A}(\mathbf{X}) - \mathbf{t}(\mathbf{X})\| \qquad \text{and} \qquad d_V = \text{cond}(\mathbf{U}^{-1}\mathbf{V_A}(\mathbf{X})\mathbf{U}^{-1\prime}),$$

where $\mathbf{U}$ is any matrix such that $\mathbf{V}(\mathbf{X}) = \mathbf{UU}'$. Experiments were performed with real and simulated data. As an example,

Table 3. LRS Data: Points With the Largest Mahalanobis Distances

Estimate	*Points with largest D_i (inverse order)*										
FMCD ($N_s = 500$)	210	173	112	90	307	2	281	193	451	67	370
SDE ($N_s = 3{,}000$)	210	307	281	173	90	112	245	472	67	2	271
OGK(.9)	210	173	112	307	90	2	281	193	451	67	147
$OGK_{(2)}$(.9)	210	173	112	90	307	2	281	193	451	67	370

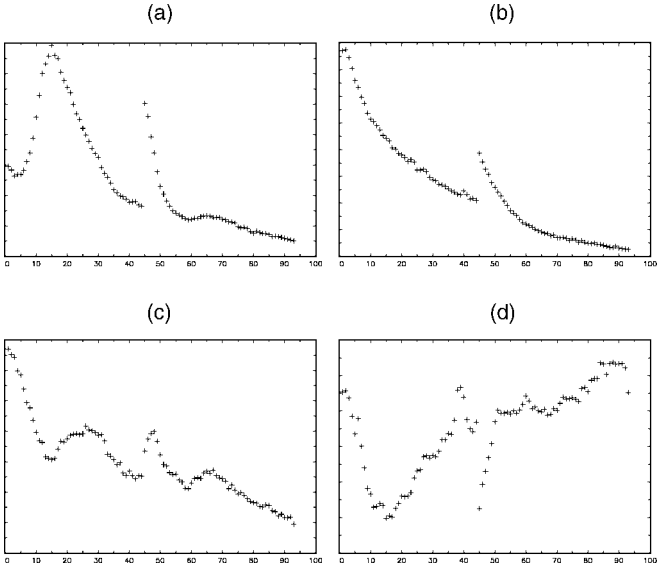

Figure 11. LRS Data. (a) and (b) Two "typical observations," 262 and 104; (c) one intermediate observation, 122; and (d) one outlier, 90.

we show the results corresponding to the ionospheric data of Section 4.3. The number of random transformations was 200. The number of iterations ranged between 1 and 4, with and without reweighting. Because the data range between 1 and -1, no scaling was used for d_t. Table 5 gives the maximum and the α-quantiles of d_V and d_t for $\alpha = .5, .7, .8$, and .9. No improvement was found beyond the second iteration. Because the values of d_V for the estimators without reweighting were about 100 times higher than those with reweighting, only the latter are shown in Table 5.

Table 5 reveals that here the effect of transformations is much stronger on $\mathbf{V}$ than on $\mathbf{t}$. Figure 12 shows the plots of Mahalanobis distances corresponding to different transformations: the untransformed data (as in Fig. 5) and the transformations corresponding to the .80 and .90 quantiles and to the maximum of d_V. For the .80 quantile, the plot is almost indistinguishable from that of the original data, and the ordering of the d_i's is essentially the same as in Table 2. For the .90 quantile, the basic features still remain, and some are still visible in the maximum case.

The following features were observed for all of the examined datasets:

- Location is much less affected than scatter.
- Reweighting makes the estimates much more equivariant.

Table 4. Simulation for $p = 5, n = 50, \epsilon = .2$ With Fixed and Random Coordinates

k		V	V_A	t	t_A
200	$OGK_{(1)}$	3.35	3.29	6.22	3.01
	$OGK_{(1)}(.9)$	.65	.69	.20	.23
	$OGK_{(2)}$	2.50	1.98	4.82	1.63
	$OGK_{(2)}(.9)$	.62	.66	.21	.24
9	$OGK_{(1)}$	1.68	1.85	.97	1.78
	$OGK_{(1)}(.9)$	1.58	1.62	3.63	3.74
	$OGK_{(2)}$	1.56	1.61	3.73	3.92
	$OGK_{(2)}(.9)$	1.62	1.65	3.83	3.94

Table 5. Measures of Lack of Equivariance for Ionospheric Data

		.5	.7	.8	.9	Max
d_V	$OGK_{(1)}(.9)$	203	264	304	373	757
	$OGK_{(2)}(.9)$	239	390	465	555	888
d_t	$OGK_{(1)}(.9)$	.39	.42	.43	.45	.51
	$OGK_{(2)}(.9)$	.47	.53	.55	.58	.67

- Further iterations do not improve on the behavior.
- Although the worst case may differ from the original data, for most transformations the results are very similar.

These results suggest that the consequences of the lack of equivariance of the estimates are not serious.

This experiment has been conducted only to demonstrate the behavior of the estimates. For this dataset, the original coordinate system is clearly the most natural one.

6. COMPUTING TIMES

To compare the computing times of the different estimates, we generated random samples with different values of n and p. We ran the experiments on a PC with a 550-MHz Intel Pentium processor with 128 Mb RAM. We first ran them in Fortran, using for FMCD the code kindly supplied by Rousseeuw and van Driessen. It turned out that the running times for FMCD were at least 100 times those for OGK, which may be due to paging. Because we could not overcome this problem, we decided to run the experiment in Gauss (version 3.2.32). This should be more favorable to SDE and FMCD, because their computing effort consists mainly of vector and matrix operations, which a matrix language like Gauss performs very quickly, whereas almost half of the time for OGK is spent computing medians.

To make our method run faster, we did not use the built-in Gauss command "median," which uses sorting. Rather, we

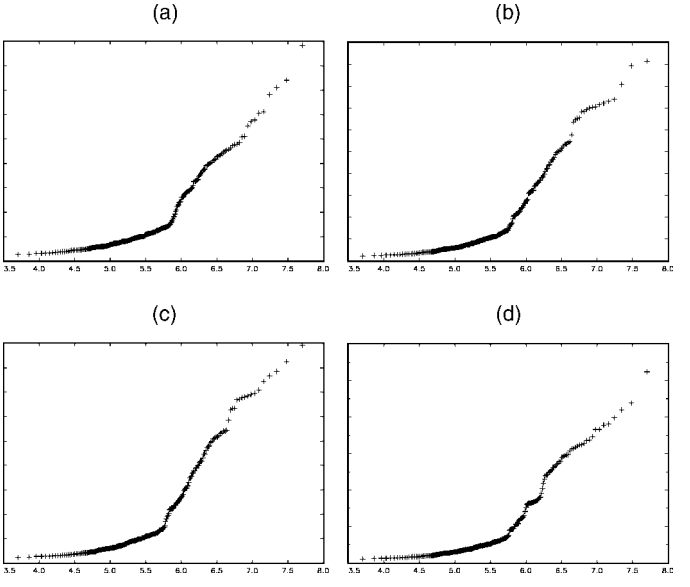

Figure 12. Q-Q Plots for Transformed Ionospheric Data. (a)–(d) correspond to original data, .80- and .90-quantiles, and maximum of d_V.

Table 6. Times for Simulated Data in Seconds $a = b + c \times d$

Estimate	n	p = 20	p = 40	p = 60	p = 80
FMCD	200	13.9	42.6	89.8	202.7
($N_s = 500$)	400	33.6	86.3	171.5	417.6
	800	74.9	178.9	333.5	726.0
SD	200	3.0	9.7	25.7	57.9
($N_s = 500$)	400	4.7	12.5	27.6	65.3
	800	8.3	17.0	37.6	70.1
OGK	200	.46	1.5	3.6	7.3
	400	.87	3.9	5.4	11.9
	800	1.6	7.0	12.3	17.6

used a selection algorithm (the procedure "select" in section 8.5 of Press et al. (1992), which is linear in n.

We implemented steps 1–4 of the algorithm in section 5 of Rousseeuw and van Driessen (1999). The running times of SDE, FMCD, and $\text{OGK}_{(1)}$ were measured for 20% contaminated normal samples with $p = 20, 40, 60$, and 80 and $n = 200, 400$, and 800. The number of subsamples was $N_s = 500$ in all cases. Whereas the running times of SDE and OGK are practically independent of the dataset, this is not so for FMCD, which seems to require more time (i.e., more iterations) for contaminated data than for pure normal data.

We have not tried larger n's for several reasons. First, we were concerned with the problem of large p than large n. Second, when n is larger than a certain n_0 (the default is 600), Rousseeuw and van Driessen's FMCD algorithm applies an ingenious splitting procedure to reduce the number of evaluations. For OGK, a time-saving procedure may be as follows. When n is larger than some n_0, take a random subsample of size n_1 and use it to perform steps 1, 2, and 3 of the definition in Section 2; then use the whole sample for (4) and (7); n_1 probably should depend on p. It is difficult to determine theoretically how much the statistical performance of FMCD and OGK deteriorates with this savings, so that further experiments would be necessary to determine an adequate choice of n_0 and n_1.

Table 6 gives the running times in seconds. It is seen that those for FMCD are between 22 and 46 times those for $\text{OGK}_{(1)}$. Note that the values of N_s actually required by SDE are much larger than the 500 used for testing. Actually, the number of subsamples required to ensure an average of five "good" ones for $\varepsilon = .2$ and $p = 20, 40, 60$, and 80 are around $400, 4 \times 10^4, 3 \times 10^6$, and 3×10^8. Table 7 shows the running times for the real datasets in the preceding section, in seconds.

Table 7. Times for Real Datasets

Dataset	n	p	N_{SD}	N_{FM}	OGK	SDE	FMCD
Bushfire	38	5	500	500	.04	.4	.8
Engineering	677	9	2,000	500	.32	14.9	20.3
Ionospheric	225	31	3,000	500	1.1	29.3	21.2
Spectral	531	93	3,000	500	19.4	458.3	615.6

7. DISCUSSION

There is probably no estimate that is fully satisfactory. FMCD is equivariant, but—although the empirical results with $N_s = 500$ are satisfactory—it is difficult to determine for a given p which N_s ensures a given breakdown point. Moreover, the simulations show that it may behave poorly under point mass contamination. SDE is equivariant, and for moderate p it does a good job under point mass contamination, but with real data, it seems to fail to detect interesting structures, and for large p, it requires impractically large values of N_s to ensure a high breakdown point. Finally, OGK is not equivariant, but it performs well in simulations with point mass contamination and performs similarly to FMCD with high-dimensional real data, all at a computational cost much lower than that of its competitors. The weighted versions are better and are "more equivariant," as demonstrated in Section 5. Iterating seems advantageous; $\text{OGK}_{(2)}(.9)$ is better than $\text{OGK}_{(1)}(.9)$ for the real datasets in Sections 4.2, 4.3, and 4.5. It must be added that even for moderate datasets, a very fast procedure has the advantage of allowing the use of computer-intensive methods, such as the bootstrap and cross-validation.

ACKNOWLEDGMENT

Ruben Zamar's research was partially funded by NSERC, Canada.

[Received May 2001. Revised February 2002.]

REFERENCES

Abdullah, M. B. (1990), "On a Robust Correlation Coefficient," *The Statistician*, 39, 455–460.

Agulló, J. (1996), "Exact Iterative Computation of the Multivariate Minimum Volume Ellipsoid Estimator With a Branch and Bound Algorithm," in *Proceedings in Computational Statistics*, ed. A. Prat, Heidelberg: Physica-Verlag, pp. 175–180.

Bay, S. D. (1999), "The UCI KDD Archive" *[http://kdd.ics.uci.edu]*, University of California, Irvine, Dept. of Information and Computer Science.

Bickel, P. J. (1964), "On Some Alternative Estimates for Shift in the p-Variate One-Sample Problem," *Annals of Mathematical Statistics*, 35, 1079–1090.

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

Croux, C., and Rousseeuw, P. J. (1992), "Time-Efficient Algorithms for Two Highly Robust Estimators of Scale," *Computational Statistics*, 2, 411–428.

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

Devlin, S. J., Gnanadesikan, R., and Kettenring, J. R. (1981), "Robust Estimation of Dispersion Matrices and Principal Components," *Journal of the American Statistical Association*, 76, 354–362.

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

Genton, M. G., and Ma, Y. (1999), "Robustness Properties of Dispersion Estimators," *Statistics and Probability Letters*, 44, 343–350.

Gnanadesikan, R., and Kettenring, J. R. (1972), "Robust Estimates, Residuals, and Outlier Detection With Multiresponse Data," *Biometrics*, 28, 81–124.

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

Hawkins, D. M. (1994), "The Feasible Solution Algorithm for the Minimum Covariance Determinant Estimator in Multivariate Data," *Computational Statistics and Data Analysis*, 17, 197–210.

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

Lopuhaä, H. P. (1991), "Multivariate τ-Estimators for Location and Scatter," *Canadian Journal of Statistics*, 19, 307–321.

Ma, Y., and Genton, M. G. (2001), "Highly Robust Estimation of Dispersion Matrices," *Journal of Multivariate Analysis*, 78, 11–36.

Maronna, R. A. (1976), "Robust M-Estimates of Multivariate Location and Scatter," *The Annals of Statistics*, 4, 51–56.

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.

Peña, D., and Prieto, F. J. (2001), "Multivariate Outlier Detection and Robust Covariance Matrix Estimation," *Technometrics*, 43, 286–301.

Press, W. H., Teukolsky, S. A. Vetterling, W. T., and Flannery, B. P. (1992), *Numerical Recipes in Fortran*, New York: Cambridge University Press.

Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871–881.

——— (1985), "Multivariate Estimation With High Breakdown Point," in *Mathematical Statistics and Aplications*, Vol. B, eds. G. S. Maddala and C. R. Rao. Amsterdam: Elsevier, pp. 101–121.

Rousseeuw, P. J., and Croux, C. (1993), "Alternatives to the Median Absolute Deviation," *Journal of the American Statistical Association*, 88, 1273–1283.

Rousseeuw, P. J., and Molenberghs, G. (1993), "Transformation of Nonpositive Semidefinite Correlation Matrices," *Communications in Statistics, Part A—Theory and Methods*, 22, 965–984.

Rousseeuw, P. J., and van Driesen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, 41, 212–223.

Ruppert, D. (1992), "Computing *S*-Estimators for Regression and Multivariate Location/Dispersion," *Journal of Computational and Graphical Statistics*, 1, 253–270.

Sen, P. K., and Puri, M. L. (1971), *Nonparametric Methods in Multivariate Analysis*, New York: Wiley.

Sigillito, V. G., Wing, S. P., Hutton, L. V., and Baker, K. B. (1989), "Classification of Radar Returns From the Ionosphere Using Neural Networks," *Johns Hopkins APL Technical Digest*, 10, 262–266.

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

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.

——— (1994), "Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimators," *Journal of the American Statistical Association*, 89, 888–896.

Yohai, V. J., and Maronna, R. A. (1990), "The Maximum Bias of Robust Covariances," *Communications in Statistics, Part A—Theory and Methods*, 19, 3925–3933.

Yohai, V. J., and Zamar, R. (1988), "High Breakdown Point Estimates of Regression by Means of the Minimization of an Efficient Scale," *Journal of the American Statistical Association*, 86, 403–413.