Combining Random and Specific Directions for Outlier Detection and Robust Estimation in High-Dimensional Multivariate Data

Daniel PEÑA and Francisco J. PRIETO

A powerful procedure for outlier detection and robust estimation of shape and location with multivariate data in high dimension is proposed. The procedure searches for outliers in univariate projections on directions that are obtained both randomly, as in the Stahel-Donoho method, and by maximizing and minimizing the kurtosis coefficient of the projected data, as in the Peña and Prieto method. We propose modifications of both methods to improve their computational efficiency and combine them in a procedure which is affine equivariant, has a high breakdown point, is fast to compute and can be applied when the dimension is large. Its performance is illustrated with a Monte Carlo experiment and in a real dataset.

Key Words: Kurtosis; Projections; Stahel-Donoho.

1. INTRODUCTION

Classical techniques for dimension reduction and discrimination with multivariate data, such as principal components, canonical correlation or linear discriminant analysis, depend on the estimation of the location and shape of the sample data. It is well known that a few outliers in the data may arbitrarily distort the sample mean and the sample covariance matrix, therefore, the robust estimation of location and shape is a crucial problem in multivariate statistics. Several robust estimates have been proposed, see Gnanadesikan and Kettenring (1972), Maronna (1976), Stahel (1981), Donoho (1982), Rousseeuw (1985), Davies (1987), Rousseeuw and van Zomeren (1990), Tyler (1991, 1994), Hadi (1992), Cook, Hawkins, and Weisberg (1993), Rocke and Woodruff (1993, 1996), Atkinson (1994), Hawkins (1994), Maronna and Yohai (1995), Agulló (1996), Rousseeuw and van Driessen (1999), Becker and Gather (2001), Peña and Prieto (2001a), Juan and Prieto (2001), Hawkins and Olive (2002), and Maronna and Zamar (2002) and the references therein. For high-dimensional

Daniel Peña is Professor, and Francisco J. Prieto is Professor, Department of Statistics, Universidad Carlos III de Madrid, 28903 Getafe Madrid, Spain (E-mail addresses: *Daniel.pena@uc3m.es* and *francisco-javier.prieto@uc3m.es*).

Journal of Computational and Graphical Statistics, Volume 16, Number 1, Pages 228–254
DOI: 10.1198/106186007X181236

large datasets a useful way to avoid the curse of dimensionality in data mining applications is to search for outliers in univariate projections of the data. Two procedures that use this approach are the Stahel-Donoho (SD from now on) procedure, that searches for univariate outliers in projections on random directions, and the method proposed by Peña and Prieto (PP from now on), that searches for outliers in projections obtained by maximizing and minimizing the kurtosis coefficient of the projected data. The first procedure has good theoretical properties, but fails for concentrated contaminations and requires prohibitive computer times for large dimension problems. The second procedure works very well for concentrated contaminations and it can be applied in much larger dimension than the previous one but its theoretical properties are unknown. As both procedures are based on projections, it seems sensible to explore how to combine them to avoid their particular limitations. This is the objective of this article.

The first contribution of the article is to propose a new method to generate random directions which is much more effective than the standard SD subsampling scheme. The second contribution is to present a modification of the PP procedure which is computationally more efficient when the dimension is large. The third contribution is a new procedure, which combines these modifications of the SD and the PP methods, which can be applied in large dimensions. The proposed procedure combines random and specific directions and has the following properties: (1) is affine equivariant; (2) inherits the good theoretical properties of the SD method; (3) inherits the good properties for finding high leverage concentrated outliers of the PP procedure; (4) it is fast to compute so that it can be applied to large datasets.

The rest of the article is organized as follows. Section 2 briefly reviews the SD method for generating random directions based on random sampling and proposes a more effective way to generate them by using stratified sampling. This section also reviews some limitations of the PP procedure for large dimensions and considers a simplification of this procedure which makes it faster to compute with a very small effect on its performance. Section 3 presents the proposed algorithm, combining random and specific directions. Section 4 illustrates the performance of the proposed method in a Monte Carlo study and compares it to the FASTMCD algorithm by Rousseeuw and Van Driessen (1999), the implementation of the Stahel-Donoho algorithm by Maronna and Yohai (1995), the computationally efficient algorithm recently proposed by Maronna and Zamar (2002), and the algorithm proposed by Peña and Prieto (2001a). Section 5 contains examples and Section 6 some concluding remarks.

2. FINDING INTERESTING DIRECTIONS

Suppose we have a sample $(x_1, \ldots, x_n)$ of a p-dimensional vector random variable X. We are interested in searching for outliers by projecting the data onto a set of directions d_j, $j = 1, \ldots, J$. Let $z_i^{(j)} = d_j' x_i$ be the projection of point x_i onto direction d_j and $z^{(j)} = (z_1^{(j)}, \ldots, z_n^{(j)})$. A univariate "measure of outlyingness" for each observation based

on these projections is

$$r_i = \max_{1 \le j \le J} \frac{|z_i^{(j)} - \text{median}(z^{(j)})|}{\text{MAD}(z^{(j)})}. \tag{2.1}$$

These measures can be used to both build robust estimates and identify outliers. The Stahel-Donoho (SD) robust estimate of the mean and covariance matrix (see Stahel 1981 and Donoho 1982) is defined by

$$m_r = \frac{\sum_1^n w_i x_i}{\sum_1^n w_i}, \tag{2.2}$$

$$S_r = \frac{\sum_1^n w_i (x_i - m)(x_i - m)'}{\sum_1^n w_i}, \tag{2.3}$$

where $w_i = w(r_i)$ is a function of the outlyingness measure r_i. For example, the Huber function $w(r_i) = \min(1, c/r_i)$ where c is a tuning constant can be used, or a redescending function that deletes points when r_i is large enough. This estimator is equivariant and has a high breakdown point in any dimension (see Stahel 1981, Tyler 1994, and Maronna and Yohai 1995). Once a robust estimate is obtained, outliers can be identified and deleted and the standard estimates of the mean and the covariance function can be applied to the uncontaminated data. Note that the resulting covariance matrix estimates have to be scaled for consistency. In this way we achieve robustness and high efficiency under normality.

The key step in the method is obtaining the directions d_j. The procedure proposed by Stahel (1981), which is the standard method used in the implementation of the algorithm, is to generate these directions randomly: a random sample of size p is chosen, a hyperplane is fitted to this sample and the direction d_j orthogonal to this hyperplane is chosen. Note that if we have a set of outliers and the data is standardized, the direction orthogonal to the fitted hyperplane is, a priori, a good one to search for outliers. This is illustrated in Figure 1(a) and (b). In case (a) the proportion of outliers is moderate, (10%), whereas in (b) the proportion is large (30%). In both cases the outliers are located in the direction of the first variable. Note that the most likely direction obtained when fitting a straight line to a random sample of two points will be approximately orthogonal to the outlier direction.

A procedure for obtaining specific directions that can reveal the presence of outliers was proposed by Peña and Prieto (2001a). They showed that the projection of the data on the direction of the outliers will lead to (1) a distribution with a large univariate kurtosis coefficient if the level of contamination is small and (2) a distribution with small univariate kurtosis coefficient if the level of contamination is large. For instance, in Figure 1(a) the projection of the observations in the direction of the outliers (the x axis) will lead to a distribution with heavy tails and a large kurtosis coefficient ($\sum(x_i - \overline{x})^4/(ns^4) = 7.85$). On the other hand, in case (b), the distribution of the projected data will be bimodal, and the kurtosis coefficient will be small ($\sum(x_i - \overline{x})^4/(ns^4) = 1.78$). Peña and Prieto (2001b) showed that if the data come from a mixture of two distributions $(1-\alpha) f_1(X) + \alpha f_2(X)$, with $0.5 > \alpha > 0$ and f_i, $i = 1, 2$, is an elliptical distribution with mean μ_i and covariance matrix V_i, the directions that maximize or minimize the kurtosis coefficient of the projected data are of the form given by Anderson and Bahadur (1962) for the admissible linear classification rules. In particular, if the distributions were normal with the same covariance

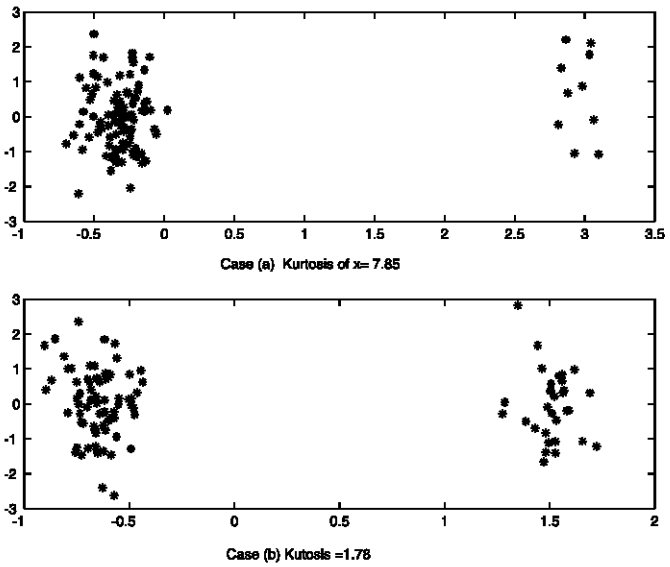

Figure 1. Two contaminated samples. In case (a) the proportion of outliers is not large (10%) and the kurtosis of the projected data in the directions of the outliers is large. In case (b) the proportion of outliers is large (30%) and the kurtosis of the projected data is very small.

matrix and the proportion of contamination is not large, $0 < \alpha < 0.289$, the direction obtained by maximizing the kurtosis coefficient is the Fisher linear discriminant function, whereas when the proportion of contamination is large, $0.289 < \alpha < 0.5$, the direction which minimizes the kurtosis coefficient is again the Fisher linear discriminant function. Thus, the extreme directions of the kurtosis coefficient seem to provide a powerful tool for searching for groups of masked outliers. Peña and Prieto (2001a) proposed an iterative procedure based on the projection onto a set of $2p$ orthogonal directions obtained as extremes for the kurtosis of the projected data. Note that the first set of p directions is closely related to the independent components of the data (see Hyvarinen, Karhunen, and Oja 2001), which are defined as a set of p variables obtained by linear transformations of the original data such that the new variables are as independent as possible. It can be shown that the independent components can be obtained by maximizing the absolute value or the square of the kurtosis coefficient and, as this coefficient cannot be smaller than one, these directions will be the same as those obtained by maximizing the kurtosis coefficient. The performance of these directions for outlier detection was found to be very good for concentrated contamination but, as can be expected from the previous results, it was not as good when the proportion of contamination is close to 0.3 and the contaminating distribution has the same variance as the original distribution. This behavior of the algorithm is not surprising because in that case the values of the kurtosis for the projected data are not expected to be either very large or very small.

Thus, it seems that we may have a powerful procedure by combining the specific directions obtained as extremes of the kurtosis with some random directions. However, as we are interested in a procedure that works in large datasets and it is well known (and it will be

discussed in the next section) that the Stahel-Donoho procedure requires a huge number of directions to work as the sample size increases, and the PP procedure requires $2p$ directions, which can be a large number in high-dimensional problems, both methods are modified in this implementation. The random directions are not generated by random sampling but by using some stratified sampling scheme that is found to be more useful in large dimensions. The PP directions are chosen by taking only a few out of the $2p$ directions, as explained in the next section.

2.1 GENERATING RANDOM DIRECTIONS

Figure 1 shows that directions computed from samples in which the p observations belong to the same group, either the group of good observations or the group of outliers, will in general be useful to identify the outliers, whereas those computed from a sample which includes points from both groups will be of limited usefulness. Thus, if the proportion of outliers is α we will obtain a good direction, defined as one computed from points that belong to the same population, with probability $\alpha^p + (1-\alpha)^p$. For large p and α not too small this probability will be very small and most of the directions generated by direct subsampling will not be useful. If we were able to decrease α, this probability would increase. This suggests increasing the probability of generating good directions by using a stratified sampling procedure from subsamples which contain a smaller proportion of outliers than the original sample.

In order to motivate the proposed procedure note that, as illustrated in Figures 1(a) and (b), when we have a sample contaminated by a group of outliers, the projections onto the direction defined by a "good" observation and a contaminating observation tend to order the observations so that on one side the projections correspond mostly to the "good" observations, while on the other side the outliers predominate. This suggests that if we select two observations at random, project the data onto the direction defined by them and then select the sample from the extremes of the projected data we can increase the probability of generating good directions, because the proportion of good or bad observations in the extremes is expected to be greater than in the whole sample. On the other hand, if the direction is computed from two good or bad observations the outliers will appear together, if we have concentrated contamination, or in the extremes of the projection, if the outliers have a larger variability than the good points. Thus it seems that we can increase the probability of good directions by dividing the projected points in K intervals (strata) of consecutive observations and by taking a random sampling of size p from each of these intervals.

In order to justify this intuition let us consider a simple model of concentrated contamination where n observations have been obtained as a random sample from the distribution in $\mathbb{R}^p$ defined by $(1-\alpha)F(x)+\alpha G(x)$. In what follows we derive the probability of obtaining a good subsampling direction (G_S), defined as one that is orthogonal to p "good" or "bad" observations. The p observations are selected through a stratified sampling procedure where we start by defining a direction from two randomly selected observations, we then form K intervals, or strata, each one containing n/K consecutive observations defined from the

ordered projections, and finally we take a random sample of size p from each one of the K intervals obtained in this manner.

In order to carry out this analysis we need to analyze the distribution of the projections onto a given direction, and study the probability of finding a good observation in a given interval. A second, more complex step requires also studying the distribution of the directions generated from a random pair of observations in $\mathbb{R}^p$.

Given the difficulty of this analysis we will center on the study of a simplified case, namely that of a mixture of normal distributions where F is obtained from a $N(0, I)$ distribution and G corresponds to a $N(\delta e_1, \lambda^2 I)$ distribution, where e_1 denotes the first unit vector. This is traditionally a hard case for outlier detection procedures, particularly if α is large. In our proposed algorithm we conduct an initial scaling and centering step, and obtain the distribution for $y_i = S^{-1/2}(x_i - \overline{x})$ indicated in the Appendix (p. 252) for large values of n, assuming that asymptotic independence between the observations and the projection direction holds. As additional simplifications we will assume in what follows that $n \to \infty$ and $\delta \to \infty$; and, without loss of generality, we assume that the data are translated so that the mean of the central distribution is the zero vector. Thus, the distribution of the y_i observations corresponds to

$$(1-\alpha)N(0, \sigma^2 \bar{I}) + \alpha N(\theta e_1, \sigma^2 \lambda^2 \bar{I}),$$

where

$$\theta = \frac{1}{\sqrt{\alpha(1-\alpha)}}, \qquad \sigma^2 = \frac{1}{\alpha\lambda^2 + 1 - \alpha}, \tag{2.4}$$

and $\bar{I}$ denotes an identity matrix with the first element equal to zero.

Under these assumptions we search for bounds on the probability of obtaining a sample formed by p observations from the same distribution, G_S, by following a two-step procedure: we first compute this probability when we sample from intervals obtained from a given projection direction, and then we compute this probability when the direction is obtained from randomly chosen pairs of observations.

- *Fixed projection direction*

  Consider a given projection direction $u \in \mathbb{R}^p$ with $\|u\| = 1$, the distribution of $y_i' u$ will be given by $M_u(x) \equiv (1-\alpha)F_u + \alpha G_u$, where F_u corresponds to a $N(0, \sigma^2(1 - u_1^2))$ and G_u to a $N(\theta u_1, \sigma^2\lambda^2(1-u_1^2))$ distribution.

  In what follows we study the case when the p observations are sampled from one of the two extreme intervals which contains the $1/K$ of the projected observations. The interval associated with the smallest values will be of the form $(-\infty, A)$ and the value A should verify

$$1/K = P(x \le A) = M_u(A) \equiv (1-\alpha)F_u(A) + \alpha G_u(A). \tag{2.5}$$

  Denote by $q(u)$ the probability, for given u, that one given observation in this interval comes from the central distribution of "good" observations F. Note that this probability only depends on u_1, that is, $q(u) \equiv q(u_1)$, and its value is given by

$$q(u_1) = \frac{(1-\alpha)F_u(A)}{(1-\alpha)F_u(A) + \alpha G_u(A)}. \tag{2.6}$$

If we consider now the other extreme interval of larger values, $(\bar{A}, \infty)$, where $\bar{A}$ is given by

$$1/K = P(x > \bar{A}) = 1 - M_u(\bar{A}), \tag{2.7}$$

and let $\bar{q}(u_1)$ denote the probability, for given u, that one given observation in this interval comes from the central distribution of "good" observations F, then

$$\bar{q}(u_1) = \frac{(1-\alpha)(1-F_u(\bar{A}))}{(1-\alpha)(1-F_u(\bar{A})) + \alpha(1-G_u(\bar{A}))}.$$

Let D_1 and D_K denote the events where the observations have been taken from the first and the last intervals, respectively. Then, the conditional probabilities of selecting p observations from the same group, G_S, can be obtained as $P(G_S|D_1, u_1) = q(u_1)^p + (1-q(u_1))^p$ and $P(G_S|D_K, u_1) = \bar{q}(u_1)^p + (1-\bar{q}(u_1))^p$. These values correspond to the probabilities that either all the observations are taken from the subset of good observations or all the observations for each case are taken from the bad subset.

- *Projection directions obtained from random pairs of observations*

The preceding values are functions of u, and in particular of u_1. For our proposed procedure we must analyze the distribution of u_1 for directions defined from pairs of randomly chosen observations $(\hat{y}, \tilde{y})$. Then $u = (\hat{y} - \tilde{y})/\|\hat{y} - \tilde{y}\|$, $\|u\| = 1$, the range of possible values for u_1 goes from -1 to 1 and

$$\begin{aligned} P(G_S) &= \int_{-1}^{1} P(G_S|u_1) dF_{u_1} = 1/2 \int_{-1}^{1} (P(G_S|D_1, u_1) + P(G_S|D_K, u_1))\, dF_{u_1} \\ &= 1/2 \int_{-1}^{1} \left(q(x)^p + (1-q(x))^p + \bar{q}(x)^p + (1-\bar{q}(x))^p\right) dF_{u_1}, \end{aligned} \tag{2.8}$$

where F_{u_1} denotes the distribution function for u_1, D_i denotes the event where the observations have been taken from interval i, we are assuming that we only generate directions from observations in the two extreme intervals and we condition on the observations belonging to each of these intervals, with probability $1/2$.

The distribution function F_{u_1} is different when the two observations used for generating the direction come from the same distribution, or when they come from different distributions. In the first case $u_1 = 0$ with probability one, as the variability of the observations along the first coordinate is zero (see the Appendix, p. 252). Denoting by $\bar{F}_{u_1}$ the distribution in the second case, we can write (2.8) as

$$\begin{aligned} P(G_S) &= \frac{\alpha^2 + (1-\alpha)^2}{2} \left(q(0)^p + (1-q(0))^p + \bar{q}(0)^p + (1-\bar{q}(0))^p\right) \\ &\quad + \alpha(1-\alpha) \int_{-1}^{1} \left(q(x)^p + (1-q(x))^p + \bar{q}(x)^p \right. \\ &\quad \left. + (1-\bar{q}(x))^p\right) d\bar{F}_{u_1}(x). \end{aligned} \tag{2.9}$$

Note that $q(u_1) = \bar{q}(-u_1)$ and also that u_1 computed when the first observation belongs to the central group and the second one to the outliers follows the same

distribution as $-u_1$ when the first observation is an outlier and the second observation belongs to the central group. As a consequence,

$$\bar{F}_{u_1}(x) = \frac{1}{2}\left(P(u_1 \leq x|S) + P(u_1 \geq -x|S)\right), \tag{2.10}$$

where S denotes the event where the first observation belongs to the central group and the second one is an outlier.

As $u_1 = (\hat{y}_1 - \tilde{y}_1)/(\|\hat{y} - \tilde{y}\|)$ and calling $v_i \equiv \hat{y}_i - \tilde{y}_i$, we have that $\|\hat{y} - \tilde{y}\|^2 = \sum_i v_i^2$ and under S it holds that $v_1 = \hat{y}_1 - \tilde{y}_1 = \theta$ with probability 1, as along the first coordinate the distance between the centers of the subgroups is $\theta = (\alpha(1-\alpha))^{-1/2}$ (see (2.4) and the Appendix) and the variability is zero, implying

$$u_1 = \frac{\theta}{\sqrt{\theta^2 + \sum_{i>1} v_i^2}}.$$

Note that for $i > 1$, $\hat{y}_i$ follows a $N(0, \sigma^2)$ distribution, where $\sigma^2 = (\lambda^2\alpha + 1 - \alpha)^{-1}$ (see (2.4)) while $\tilde{y}_i$ follows a $N(0, \lambda^2\sigma^2)$ distribution, and all variables are independent (see the Appendix). As a result, $v_i = \hat{y}_i - \tilde{y}_i$ follows a $N(0, \sigma^2(1+\lambda^2))$ distribution, and $\sum_{i>1} v_i^2/(\sigma^2(1+\lambda^2))$ follows a χ^2_{p-1} distribution. From this result and (2.10),

$$\bar{F}_{u_1}(x) = \frac{1}{2}\left(P\left(\theta \leq x\sqrt{\theta^2 + \sum_{i>1} v_i^2}\right) + P\left(\theta \geq -x\sqrt{\theta^2 + \sum_{i>1} v_i^2}\right)\right),$$

and as for $x \leq 0$ it holds that $P(u_1 \leq x|S) = 0$ and for $x \geq 0$ it holds that $P(u_1 \geq -x|S) = 1$, we obtain

$$\bar{F}_{u_1}(x) = \begin{cases} \frac{1}{2}P\left(\chi^2_{p-1} \leq \theta^2(1-x^2)/x^2\right) & \text{if} \quad x \leq 0, \\ \frac{1}{2}\left(1 + P\left(\chi^2_{p-1} \geq \theta^2(1-x^2)/x^2\right)\right) & \text{if} \quad x \geq 0, \end{cases}$$

and

$$d\bar{F}_{u_1}(x) = \begin{cases} -\theta^2 f_{\chi^2_{p-1}}\left(\theta^2(1-x^2)/x^2\right)/x^3\, dx & \text{if} \quad x < 0, \\ \theta^2 f_{\chi^2_{p-1}}\left(\theta^2(1-x^2)/x^2\right)/x^3\, dx & \text{if} \quad x > 0, \end{cases} \tag{2.11}$$

where $f_{\chi^2_{p-1}}$ denotes the density function for a χ^2_{p-1} random variable. Using the symmetry properties of q and $\bar{q}$, (2.11) and (2.9) we finally have

$$\begin{aligned} P(G_S) = &\left(\alpha^2 + (1-\alpha)^2\right)\left(q(0)^p + (1-q(0))^p\right) \\ &+ 2\int_0^1 \left(q(x)^p + (1-q(x))^p \right. \\ &\left. + \bar{q}(x)^p + (1-\bar{q}(x))^p\right) f_{\chi^2_{p-1}}\left(\frac{\theta^2(1-x^2)}{x^2}\right)\frac{dx}{x^3}. \end{aligned} \tag{2.12}$$

Table 1. Probability of Generating a Good Direction

α	λ	p	K	Probability for SD, p_{SD}	Probability bound (2.12)	Efficiency ratio
0.3	1	10	5	0.0283	0.2235	7.91
0.3	1	20	5	$7.98\ 10^{-4}$	0.0642	80.42
0.4	1	20	5	$3.66\ 10^{-5}$	0.0221	605.42
0.3	0.1	20	5	$7.98\ 10^{-4}$	0.7901	990.21
0.4	0.1	20	5	$3.66\ 10^{-5}$	0.7599	20778.41
0.3	1	30	5	$2.25\ 10^{-5}$	0.0140	622.25
0.4	1	30	5	$2.21\ 10^{-7}$	$1.49\ 10^{-3}$	6753.64
0.3	0.1	30	5	$2.25\ 10^{-5}$	0.7901	35054.42
0.4	0.1	30	5	$2.21\ 10^{-7}$	0.7598	$3.43\ 10^{7}$

The previous analysis has been made under the assumption that we obtain one direction by selecting with probability 1/2 one of the two extreme intervals and then obtaining one sample of size p at random from the observations in the selected interval. Thus, the value in (2.12) is exact only if we sample from these extreme intervals. For the algorithm described in the following section, we obtain directions from each of the K intervals and the corresponding probability $P(\bar{G}_S)$ satisfies

$$P(\bar{G}_S) = \frac{1}{K}\sum_{i=1}^{K} P(\bar{G}_S|D_i) = \frac{2}{K}P(G_S) + \frac{1}{K}\sum_{i\neq 1,K} P(\bar{G}_S|D_i) \Rightarrow P(\bar{G}_S) \geq \frac{2}{K}P(G_S).$$

To illustrate the behavior of the method, the Table 1 includes the values obtained from (2.12) for some particular cases, computed using numerical quadrature methods, as well as the corresponding values when taking one direction at random by the standard Stahel-Donoho algorithm, which will lead to a good direction with probability $p_{\mathrm{SD}} = (1-\alpha)^p + \alpha^p$. These values correspond to situations that are difficult both for Stahel-Donoho and the kurtosis algorithm, and show a marked improvement, particularly as the dimension increases, in the proposed stratified sampling scheme.

2.2 GENERATING SPECIFIC DIRECTIONS

The algorithm proposed by PP generates $2p$ orthogonal directions obtained by maximizing and minimizing the kurtosis coefficient of the projections. Suppose that we have just one group of similar outliers. This group will usually appear in the direction of either the largest or smallest projected kurtosis, and the rest of the directions will not useful. If we have several groups of similar outliers, it seems better to find a group, remove it from the sample, and start the search again instead of going through the process of computing the $2p$ directions.

There are two possible solutions to speed up the process. The first one is to compute only two directions, those with largest and smallest kurtosis coefficients. The second is to compute $n_1 < p$ directions by maximizing, where n_1 is determined by monitoring the value obtained for the kurtosis of the projections. When this value is close to 3, we stop

the process. In the same way we can compute $n_2 < p$ values by minimizing the kurtosis with the same objective of having directions which are useful. We have also explored the alternative of stopping the computation of the d direction when outliers are not found in the $d-1$st previous direction. After many Monte Carlo simulations we have found that the best and simplest solution is to compute a small number of directions, as we will discuss in the next sections.

3. DESCRIPTION OF THE ALGORITHM

The details of the computation of the directions and the analysis of the projections are presented in the following. Note that the procedure is affine equivariant. The algorithm requires four parameters: the numbers of maximization and minimization directions n_1, the number of random directions, L, the number of intervals for each random direction, K and the correction factor β_p to identify outliers. These parameters will be discussed after presenting the algorithm. First, we assume that the original data are scaled and centered, that is, letting $\bar{x}$ be the mean and S the covariance matrix of the original data, the observations are transformed by using

$$y_i = S^{-1/2}(x_i - \bar{x}), \quad i = 1, \ldots, n. \tag{3.1}$$

- **Stage I:** Specific directions. Compute n_1 orthogonal directions and projections maximizing the kurtosis coefficient ($1 \leq n_1 \leq p$) and n_1 directions minimizing this coefficient.

  1. Set $y_i^{(1)} = y_i$ and the iteration index $j = 1$.

  2. The direction that maximizes the coefficient of kurtosis is obtained as the solution of the problem

$$\begin{aligned} d_j = \quad & \arg\max_d \quad \frac{1}{n}\sum_{i=1}^{n}\left(d' y_i^{(j)}\right)^4 \\ & \text{s.t.} \qquad d'd = 1. \end{aligned} \tag{3.2}$$

  3. The sample points are projected onto a lower dimension subspace, orthogonal to the direction d_j. Define

$$v_j = d_j - e_1, \quad Q_j = \begin{cases} I - \dfrac{v_j v_j'}{v_j' d_j} & \text{if } v_j' d_j \neq 0 \\ I & \text{otherwise,} \end{cases}$$

  where e_1 denotes the first unit vector. The resulting matrix Q_j is orthogonal, and we compute the new values

$$u_i^{(j)} \equiv \begin{pmatrix} z_i^{(j)} \\ y_i^{(j+1)} \end{pmatrix} = Q_j y_i^{(j)}, \quad i = 1, \ldots, n,$$

where $z_i^{(j)}$ is the first component of $u_i^{(j)}$, which satisfies $z_i^{(j)} = d_j' y_i^{(j)}$ (the univariate projection values), and $y_i^{(j+1)}$ corresponds to the remaining $p - j$ components of $u_i^{(j)}$.

We set $j = j + 1$, and if $j < n_1$ we go back to step 1(b). Otherwise, we let $z_i^{(p)} = y_i^{(p)}$.

4. The same process is applied to the computation of the directions d_j (and projections $z_i^{(j)}$), for $j = n_1 + 1, \ldots, 2n_1$ minimizing the kurtosis coefficient.
5. The normalized univariate distances r_i^j, related to (2.1), are computed as

$$r_i^j = \frac{1}{\beta_p} \frac{|z_i^{(j)} - \text{median}(z^{(j)})|}{\text{MAD}(z^{(j)})}, \tag{3.3}$$

for each direction $j = 1, \ldots, n_1 + n_2$, where β_p is a predefined reference value.

- **Stage II:** Random directions, obtained from a stratified sampling procedure as follows:

1. In iteration l, two observations are chosen randomly from the sample and the direction $\hat{d}_l$ defined by these two observations is computed. The observations are then projected onto this direction, to obtain the values $\hat{z}_i^l = \hat{d}_l' y_i$. Then the sample is partitioned into K intervals of size n/K, where K is a prespecified number, based on the ordered values of the projections $\hat{z}_i^l$, so that interval k, $1 \le k \le K$, contains those observations i satisfying

$$\hat{z}^l_{(\lfloor (k-1)n/K \rfloor + 1)} \le \hat{z}_i^l \le \hat{z}^l_{(\lfloor kn/K \rfloor)}.$$

2. From each interval k, $1 \le k \le K$, a subsample of p observations is chosen without replacement. The direction orthogonal to these observations, $\tilde{d}_{kl}$, is computed, as well as the corresponding projections $\tilde{z}_i^{kl} = \tilde{d}_{kl}' y_i$ for all observations i. These projections are used to obtain the corresponding normalized univariate distances r_i^j,

$$r_i^j = \frac{1}{\beta_p} \frac{|\tilde{z}_i^{kl} - \text{median}(\tilde{z}^{kl})|}{\text{MAD}(\tilde{z}^{kl})}, \tag{3.4}$$

where $j = 2p + \lfloor (k-1)n/K \rfloor + l$, and β_p the prespecified reference value.

3. This procedure is repeated a number of times L, until $l = L$.

- **Stage III**: Checking.

1. For each observation i its corresponding normalized outlyingness measure r_i is obtained from the univariate distances r_i^j defined in (3.3) and (3.4), as

$$r_i = \max_{1 \le j \le 2p + \lfloor Ln/K \rfloor} r_i^j.$$

Those observations having values $r_i > 1$ are labeled as outliers and removed from the sample, if their number is smaller than $n - \lfloor (n + p + 1)/2 \rfloor$. Otherwise, only those $n - \lfloor (n + p + 1)/2 \rfloor$ observations having the largest values of r_i are labeled as outliers.

2. A Mahalanobis distance is computed for all observations labeled as outliers in the preceding steps, using the data (mean and covariance matrix) from the remaining observations. Let U denote the set of all observations not labeled as outliers. The algorithm computes

$$\begin{aligned} \tilde{m} &= \frac{1}{|U|} \sum_{i \in U} x_i, \\ \tilde{S} &= \frac{1}{|U| - 1} \sum_{i \in U} (x_i - \tilde{m})(x_i - \tilde{m})', \\ v_i &= (x_i - \tilde{m})' \tilde{S}^{-1} (x_i - \tilde{m}), \qquad \forall i \notin U. \end{aligned}$$

3. Those observations $i \notin U$ such that $v_i < \chi^2_{p-1,0.99}$ are not considered to be outliers, and are included in U. The process is repeated until no more such observations are found (or U becomes the set of all observations).

As indicated before, this algorithm includes several parameters. The values assigned to them in the implementation have been chosen to ensure adequate theoretical and efficiency properties. Next we describe these choices and their motivation.

1. The number of maximization and minimization directions n_1 was selected as equal to 1 in one of the experiments and equal to p in a second experiment. In the first case we call the algorithm RASP(1) and in the second RASP(p). These two alternatives will be compared in the next section in a Monte Carlo study.

2. The use of parameter β_p in (3.3) of Stage I, jointly with the test on r_i to label the outliers, implies that β_p is acting as a cutoff value to detect outliers from projections of the observations onto the directions that minimize or maximize the kurtosis coefficient. Its value is chosen to ensure a reasonable level of Type I errors, and depends on the sample space dimension p. In particular, a set of simulation experiments were carried out to ensure that, in the absence of outliers, the percentage of correct observations mislabeled as outliers is approximately equal to 5%. Table 2 shows the values used for several sample space dimensions. The values for other dimensions could be obtained by interpolating $\log \beta_p$ linearly in $\log p$.

Table 2. Cutoff Values for Univariate Projections

Sample space dimension p	5	10	20
Cutoff value β_p	3.46	3.86	4.67

3. The number of intervals considered in each iteration of Stage II, K, was fixed so that each interval had a size of $2p$. In practice $K = 3$ or 5 seems to work well in the applications.

4. The number of iterations L for Stage II was selected so that the total number of subsampling directions was equal to $10p$. A larger number of directions provided only a limited increase in the performance of the algorithm.

The main computational effort in the application of the preceding algorithm is associated with the determination of local solutions for (3.2) and this computation has been carried out as described by Peña and Prieto (2001a).

The procedure is affine equivariant and shares some of the good theoretical properties of the Stahel-Donoho estimate obtained using a subsampling approximation. This estimate has a high breakdown point in finite samples and it has been found to exhibit high efficiency for both Normal and Cauchy distributions (see Maronna and Yohai 1995). The second part of Stage II in the proposed procedure is a modification of the Stahel-Donoho subsampling scheme with modified sample weights. For a sample including outliers arbitrarily removed from the uncontaminated observations, to identify (some of) the outliers it is enough to generate directions from hyperplanes defined by subsets of p uncontaminated observations, as in the Stahel-Donoho subsampling scheme. These directions are obtained with positive probability by the proposed scheme, implying that if the number of subsamples were sufficiently large, any outliers at infinity would be detected. In the next section we will see that our proposal is also a powerful procedure for outlier detection at moderate distances from the uncontaminated sample.

4. SIMULATION RESULTS

We have conducted a number of computational experiments to compare the performance of the proposed algorithm, RASP(1), in the identification of the outliers, with the results from other codes: (1) An efficient algorithm for the implementation of the Minimum Covariance Determinant (MCD) procedure, the FASTMCD algorithm proposed by Rousseeuw and Van Driessen (1999), which is based on the splitting of the problem into smaller subproblems. (2) An implementation of the Stahel-Donoho algorithm, as described by Maronna and Yohai (1995). The choice of parameters was the same as in this reference, except for the number of subsamples, chosen equal to $200p$ for $p = 5$, 10 and 20. These numbers of subsamples yield running times comparable with (in fact, larger than) those of the proposed algorithms. (3) A computationally efficient method recently proposed by Maronna and Zamar (2002), based on the analysis of the principal components of an adjusted covariance matrix computed from information on pairwise covariances. Two iterations of the algorithm have been carried out, as suggested by the authors. (4) An algorithm based on the directions computed from the minimization and maximization of the kurtosis coefficient, as described in Peña and Prieto (2001,a). (5) A stratified sampling procedure, SRand, corresponding to the second part of the RASP algorithm described in Section 3, using the same numbers of directions and

Table 3. Success Rates for the Detection of Outliers Forming One Cluster When One of the Algorithms Scored Fewer than 95 Successes

p	α	δ	n	$\sqrt{\lambda}$	FASTMCD	SD	MZ	kurtosis	SRand	RASP(1)	RASP(p)
5	0.3	10	100	0.1	0	100	0	100	100	100	100
5	0.4	10	100	0.1	0	97	0	100	98	100	100
5	0.4	10	100	1	100	100	0	100	95	99	100
5	0.4	100	100	0.1	0	100	0	100	100	100	100
10	0.2	10	100	0.1	0	100	48	100	100	100	100
10	0.3	10	100	0.1	0	100	0	98	100	100	100
10	0.3	100	100	0.1	0	100	2	96	100	100	100
10	0.4	10	100	0.1	0	0	0	98	65	99	99
10	0.4	10	100	1	90	33	0	97	63	99	97
10	0.4	100	100	0.1	0	100	0	100	100	100	100
10	0.4	100	100	1	70	100	0	98	99	100	100
20	0.2	10	200	0.1	0	100	9	97	100	100	100
20	0.2	10	200	1	100	100	100	0	100	99	100
20	0.2	100	200	0.1	0	100	100	91	100	100	100
20	0.3	10	200	0.1	0	33	0	89	100	100	100
20	0.3	10	200	1	0	39	1	1	49	49	47
20	0.3	100	200	0.1	0	88	0	82	100	100	100
20	0.3	100	200	1	30	87	100	1	84	78	77
20	0.4	10	200	0.1	0	0	0	100	52	100	100
20	0.4	10	200	1	0	0	0	60	5	45	53
20	0.4	100	200	0.1	0	4	0	100	100	100	100
20	0.4	100	200	1	0	6	0	73	25	55	66

parameter values indicated in that Section. (6) An implementation of the proposed RASP(1) algorithm described in Section 3. (7) An implementation of RASP(p), that is, a modification of the proposed algorithm using now the full $2p$ directions maximizing and minimizing the kurtosis coefficient.

For a given contamination level α, we have generated a set of $n(1-\alpha)$ observations from a $N(0, I)$ distribution in dimension p. We have added $n\alpha$ additional observations from a $N(\delta e, \lambda I)$ distribution, where e denotes the vector $(1 \ \ldots \ 1)'$. This model is analogous to the one used by Rousseeuw and van Driessen (1999). In the method by Maronna and Zamar (2002) (MZ from now on) we have introduced a linear transformation to ensure that the resulting datasets have mean zero and covariance matrix equal to the identity, as the corresponding procedure is not affine equivariant. This experiment was conducted for different values of the sample size n ($n = 100, 200$), the sample space dimension p ($p = 5, 10, 20$), the contamination level α ($\alpha = 0.1, 0.2, 0.3, 0.4$), the distance of the outliers δ ($\delta = 10, 100$), and the standard deviation of these outliers $\sqrt{\lambda}$ ($\sqrt{\lambda} = 0.1, 1, 5$). For each set of values 100 samples were generated. Table 3 gives the number of samples in which all the outliers have been correctly identified, for each set of parameter values and the different algorithms indicated above: FASTMCD, SD, MZ, "kurtosis," "SRand," "RASP(1)," and "RASP(p)." In SRand, RASP(1) and RASP(p) the value for the number of strata used, K, was chosen so that all strata contained $2p$ observations. To limit the size of

Table 4. Overall Success Rates for the Detection of Outliers Forming One Cluster

FASTMCD	*SD*	*MZ*	*kurtosis*	*SRand*	*RASP(1)*	*RASP(p)*
74.9	90.1	70.2	88.0	94.9	97.5	98.0

the table, we have shown only those cases where at least one of the algorithms scored fewer than 95 successes.

As a summary of this experiment, we also present in Table 4 the percentage of successes for the whole simulation experiment and all the procedures, obtained as the average of the success rates over all the cases included in the experiment. The modification of the Stahel-Donoho procedure proposed in this article behaves uniformly better than the original procedure, particularly for larger dimensions and higher contamination levels, that is, the most difficult cases. The proposed combined method of random and specific projections ("RASP") seems to perform equivalently or better than the other alternatives in nearly all cases. In particular, it is clearly better than FASTMCD for concentrated contaminations, and it improves on the Stahel-Donoho implementation for large contaminations and increasing space dimensions. Furthermore, the proposed procedure improves on both the original kurtosis procedure and the stratified modification for the Stahel-Donoho resampling scheme.

Table 5 also provides the average percentages of nonoutliers detected as outliers by the different procedures in the preceding simulation experiment. Note that the values for the proposed procedures RASP(1) and RASP(p) are particularly low.

To provide some indication of the computational effort required to implement the different procedures, Table 6 shows the average running times for the algorithms to carry out the computations for sets of 100 replications and the same combinations of values for α, δ, and λ used in the experiment described in this section and $n = 100, 200, 300$. The times have been measured on an AMD 3000+ computer with 512 MB of internal memory. All codes were written in Matlab except for FASTMCD, a FORTRAN code. Note that although the best outlier detection results were obtained for RASP(p), those for RASP(1) are also significantly better than the rest and are attained with much lower running times. In fact, for large dimensional problems ($p = 20$) the running times for RASP(1) are the second lowest (after SRand). Moreover, for large n and small dimension p the best running times are those of MZ, but these times increase rapidly with p.

Table 7 shows the average Type I errors for the whole problem sets. The values for the proposed methods are very close to the target 5.0% value.

Table 5. Average Percentages of Nonoutliers Detected as Outliers

FASTMCD	*SD*	*MZ*	*kurtosis*	*SRand*	*RASP(1)*	*RASP(p)*
7.8	3.9	8.3	2.7	1.8	0.8	0.7

Table 6. Average Running Times (in seconds) for Problem Set

Sample size n	*Dimension p*	*FASTMCD*	*SD*	*MZ*	*Kurtosis*	*SRand*	*RASP(1)*	*RASP(p)*
100	5	27.3	11.4	2.6	6.7	0.9	2.5	5.6
100	10	79.4	39.4	6.6	14.7	1.6	3.4	13.6
100	20	299.8	394.0	24.4	33.8	3.9	6.9	37.0
200	5	45.4	16.3	2.9	12.5	1.6	4.5	9.7
200	10	142.3	49.6	8.4	25.5	2.5	6.1	24.0
200	20	515.9	427.7	29.3	79.7	6.4	11.9	79.4
300	5	64.1	20.7	3.2	18.5	2.4	7.0	14.2
300	10	205.9	58.6	9.3	38.0	3.4	8.9	34.5
300	20	731.1	452.0	33.6	114.6	8.7	17.4	117.8

Finally, we have also compared these methods for the robust estimation of the covariance matrix. We have generated 100 samples from the mixture model $n(1-\alpha)\mathrm{N}(0, I) + n\alpha\mathrm{N}(\delta e, \lambda I)$ explained before and we have computed in each sample the robust estimates considered in the previous simulation experiment. In order to compare the results with a non robust estimator we have also computed the sample covariance matrix (column Cov). The median of the condition number of the estimated covariance matrix in these 100 samples is reported in Table 8. The parameter values are the same than in Table 3. The performance of RASP(p) is the best of all the algorithms in the experiment, while RASP(1) is the second best, but still significantly better than the other alternatives. From these results we may conclude that if a robust estimate for the covariance matrix is needed and computational efficiency is not too relevant, RASP(p) may present some advantages compared to RASP(1).

The estimates for the covariance matrix have to be scaled for consistency. Table 9 provides scaling factors for different values of p and n obtained from a simulation study for the proposed procedure, RASP(1).

In conclusion, RASP(1) seems to offer a good compromise among reduced running times, good outlier identification and robust covariance matrix estimation properties.

5. EXAMPLES

In this section we illustrate the performance of the proposed algorithm for outlier detection with two types of examples. First, we verify that it finds the outliers that other procedures have also found in well known examples of multivariate data. Second, we apply

Table 7. Average Type I Errors for Problem Set

FASTMCD	*SD*	*MZ*	*Kurtosis*	*SRand*	*RASP(1)*	*RASP(p)*
18.0	1.1	6.4	5.1	5.3	5.1	5.4

Table 8. Medians of Condition Numbers for the Covariance Matrix Estimates

p	α	δ	n	$\sqrt{\lambda}$	*FASTMCD*	*SD*	*MZ*	*Kurtosis*	*SRand*	*RASP(1)*	*RASP(p)*	*Cov*
5	0.3	10	100	0.1	2229.6	2.5	439.9	2.6	2.5	2.4	2.4	219.4
5	0.4	10	100	0.1	5981.0	2.5	916.8	2.6	2.5	2.6	2.6	305.5
5	0.4	10	100	1	2.5	2.6	189.4	2.6	2.6	2.7	2.7	162.7
5	0.4	100	100	0.1	692634.6	2.5	131207.3	2.4	2.5	2.6	2.6	29376.8
10	0.2	10	100	0.1	4344.5	4.3	4.0	4.4	4.5	4.3	4.3	407.1
10	0.3	10	100	0.1	12679.0	5.0	1177.3	4.8	4.7	4.8	4.7	613.8
10	0.3	100	100	0.1	1423003.1	5.2	4.1	4.6	4.9	4.6	4.9	64228.0
10	0.4	10	100	0.1	25736.2	13343.6	2524.3	5.0	5.8	4.9	5.3	906.9
10	0.4	10	100	1	5.4	442.3	504.4	5.0	5.8	5.0	5.0	438.3
10	0.4	100	100	0.1	2640179.7	4.8	287946.3	5.0	4.9	4.9	4.7	91794.8
10	0.4	100	100	1	5.1	4.3	45530.4	5.2	5.1	4.9	5.0	44377.4
20	0.2	10	200	0.1	5173.7	3.8	848.6	4.7	4.7	4.5	4.6	856.3
20	0.2	10	200	1	5.1	3.8	4.0	686.4	4.7	4.7	4.8	631.6
20	0.2	100	200	0.1	515204.2	3.9	4.3	4.6	4.5	4.5	4.6	83161.7
20	0.3	10	200	0.1	17026.9	1962.2	2199.7	5.4	5.1	5.2	5.0	1389.5
20	0.3	10	200	1	1245.9	806.6	891.8	927.9	6.4	7.8	7.0	809.2
20	0.3	100	200	0.1	1786922.1	5.4	4.7	5.3	5.2	5.3	5.1	134403.8
20	0.3	100	200	1	124193.9	4.4	4.4	92069.9	5.7	6.0	6.0	82990.8
20	0.4	10	200	0.1	43521.4	33043.0	4365.1	5.4	14366.6	5.5	5.3	1955.5
20	0.4	10	200	1	1528.8	953.1	1055.3	6.2	1080.2	1002.4	6.9	937.2
20	0.4	100	200	0.1	4354604.7	3080196.3	478372.9	5.3	5.4	5.2	5.3	198653.9
20	0.4	100	200	1	153160.9	93468.6	100556.6	6.0	104381.1	44259.3	6.0	94404.7

Table 9. Scaling Factors for Consistent Covariance Estimators, RASP(1)

p/n	*100*	*200*	*300*
5	1.075	1.042	1.031
10	1.064	1.031	1.031
20	1.111	1.042	1.020

the procedure to financial data which is known to be far from normal and illustrate how the cutoff value for finding outliers can be modified when we have heavy tail distributions.

The proposed code, RASP(1), was applied to a collection of seven standard small datasets used by previous authors to detect outliers in multivariate data. The first six were studied by Rousseeuw and Van Driessen (1999), among others, and the last was analyzed by Maronna and Yohai (1995), among others. Table 10 gives the corresponding results, indicating the dataset, its dimension and number of observations, the number of outliers and their labels.

The results for the number of identified outliers are similar to the ones reported in the literature and those obtained using the Kurtosis algorithm, except for the "Salinity" dataset, where the proposed algorithm finds a slightly smaller number of outliers, and "Coleman," where it finds a slightly larger number.

We have also explored the identification of outliers in data from a heavy-tail distribution. The data matrix has 1,272 rows and 18 columns of daily return stock data from the Madrid stock market. The variable measured is the daily return of a stock, computed as $\Delta \log P_t$ where P_t is the price of the stock. The columns in this matrix are the 18 stocks which were always included in the five year period 2000–2004 in the index ibex35, which combines the stocks with the largest trading volume of the Madrid stock exchange. The rows are the value of these 18 stocks in the 1,272 trading days in the five year sample period.

We have checked first the autocorrelation structure of these time series by computing the correlogram of the 18 series. Only two of them, stocks IBE and SCG, show a small, though significant, first order autocorrelation coefficient, with values -0.1326 and 0.1275,

Table 10. Results Obtained by the Proposed Algorithm on Small Datasets

Dataset	*Dim.*	*# Obs.*	*# Outliers*	*Outliers*
Heart	2	12	5	2,6,8,10,12
Phosphor	2	18	6	1,4,6,7,10,16
Stackloss	3	21	4	1,2,3,21
Salinity	3	28	4	5,16,23,24
HBK	3	75	14	1,2,3,4,5,6,7,8,9,10,11,12,13,14
Coleman	5	20	7	1,2,6,9,10,11,18
Wood	5	20	6	4,6,7,8,9,19
Bushfire	5	38	15	7,8,9,10,11,12,30,31,32,33,34,35,36,37,38

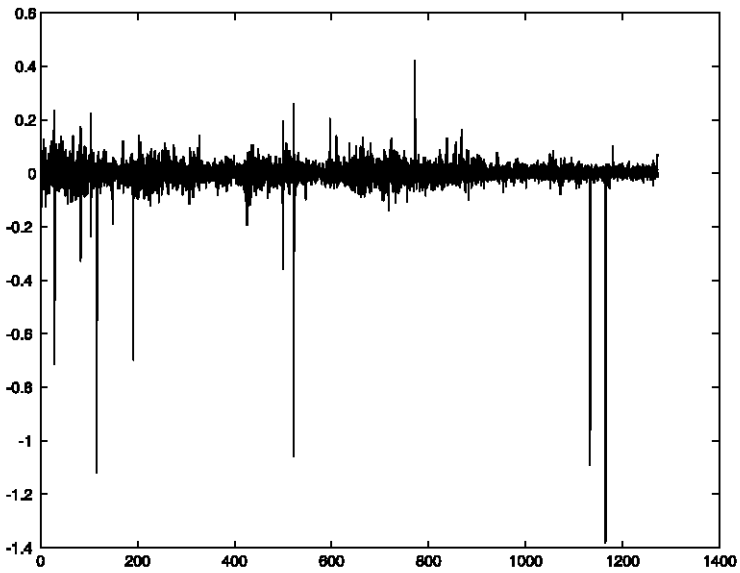

Figure 2. The 18 series of daily stock returns in the period 2000–2004.

respectively, and the rest of the autocorrelation coefficients were very small. In order to get rid of this autocorrelation we fitted an AR(1) model to these two time series and used the residuals from this fit in the analysis. However, as the results were almost identical to those obtained by using the original return series, for simplicity we present here the results for the original data.

The plot of these 18 daily returns time series is shown in Figure 2. As the values of the 18 time series are similar in most cases, the plot seems to correspond to a single time series. However, this plot is useful to show the most important outliers in any of the 18 time series. The figure shows that in 10 days we have returns which are extremely low in one or several of the stocks. We checked that these extreme values corresponded to well known changes, such as stock splits, and these changes produce a proportional drop in price and the corresponding large negative return for the next trading day.

Additionally the kurtosis coefficients of these return series, which are shown in Table 11 together with the label of the stock, indicate that the distribution of these stock returns is far from normal. The large values of these kurtosis coefficients for some of the stocks are in agreement with the large outliers, which can be seen in Figure 2.

Table 11. Kurtosis Coefficients of the Original Daily Return Data

ACS	*ACX*	*ALT*	*AMS*	*ANA*	*BBVA*	*BKT*	*ELE*	*FCC*
666.74	729.86	30.18	19.02	12.08	9.92	457.83	6.08	8.37
FER	IBE	IDR	NHH	POP	REP	SGC	TEF	TPI
57.36	10.78	189.68	20.82	430.68	8.68	7.29	22.54	259.89

Table 12. Skewness and Kurtosis Coefficients of Daily Returns in Group A of Good Data

ACS	*ACX*	*ALT*	*AMS*	*ANA*	*BBVA*	*BKT*	*ELE*	*FCC*
−0.168	−0.003	0.046	0.103	0.006	−0.023	0.184	−0.208	0.093
3.525	3.154*	3.592	3.895	3.517	3.675	3.899	3.524	3.566
FER	*IBE*	*IDR*	*NHH*	*POP*	*REP*	*SGC*	*TEF*	*TPI*
0.327	−0.036	0.018	−0.132	0.177	−0.056	0.076	−0.026	0.079
3.775	3.295*	3.596	3.826	3.566	4.168	3.806	3.217*	3.671

If we search for outliers in the individual time series, x_{it}, where $i = 1, \ldots, 18$ and $t = 1, \ldots, 1272$, and identify as outliers values larger than

$$r_{it} = \frac{|x_{it} - \text{median}(x_{it})|}{\text{MAD}(x_{it})} > 2.9$$

we find a proportion of outliers of 11%. The numbers of outliers in each time series are similar, with the smallest proportion of outliers in series IBE (9.2%) and the maximum in series BKT (13.05%). Often the outliers appear at the same time in several of the time series.

The application of the proposed procedure leads to a much larger group of outliers, indicating that the joint analysis is more powerful than the individual analysis of the series. In fact the procedure implies a split of the sample into two groups. The first group, the largest one, contains 645 good observations; we will refer to them as the A group. The second group includes 624 observations which were considered outliers (48.83% of the data), and will be called the B group. Table 12 shows the skewness and kurtosis coefficient in group A and the result of the Bera-Jarque test of univariate normality. This hypothesis is rejected at the 0.05 level in 15 out of the 18 stocks. Only in the three cases indicated by an $*$ univariate normality cannot be rejected. All the other stocks have univariate distributions with kurtosis values between 3.5 and 4.2. If we assume that the daily returns follow Student t distributions and estimate the degrees of freedom from the kurtosis coefficient we obtain t distributions with between 10 and 16 degrees of freedom.

On the one hand, group B has a set of 10 extreme points, all of which can be seen in Figure 2, which are clearly outliers due to well known events. When we delete these 10 points from group B we obtain group B* which contains data that follow univariate distributions with greater variance that those in group A. Table 13 shows the skewness and kurtosis of group B*. It can be seen that the kurtosis of all the univariate distributions are larger than those in group A, and they seem to agree with a Student t distribution with small degrees of freedom. Also, the skewness is larger in this group and some of these distributions are not symmetric. We have used the symbol $*$ to indicate that in this stock the hypothesis of a symmetric distribution is rejected. The variability in group B* is larger than in group A, and Table 14 shows that the standard deviations are twice as large as those in group A.

Table 13. Skewness and Kurtosis Coefficients of Daily Returns in Group B* of Outliers With the 10 Largest Values Deleted

ACS	*ACX*	*ALT*	*AMS*	*ANA*	*BBVA*	*BKT*	*ELE*	*FCC*
0.343	−0.209	−0.350	−0.230	0.040	0.303	0.561*	−0.088	0.415*
5.774	5.619	4.494	4.876	4.668	3.914	9.258	4.043	4.344
FER	*IBE*	*IDR*	*NHH*	*POP*	*REP*	*SGC*	*TEF*	*TPI*
0.115	0.602*	0.319	−0.118	0.243	0.373*	0.231	0.344	0.346
3.507	7.772	4.215	12.492	4.004	5.165	3.404	3.340	5.603

Figure 3 presents a plot of the observations for all the stocks in both groups, A (top) and B*(bottom), and they seem to correspond to two regimes with different variability.

In order to explore this possibility we studied the proportion of outliers, defined as observations which belong to group B*, in subgroups of eight consecutive observations. That is, we split the 1,272 observations into 159 subgroups of eight consecutive data points and computed the proportion of outliers in each subgroup. Figure 4 shows this proportion with respect to the order of the subgroup, which indicates time. It can be seen that at the beginning of the sample period the proportion of outliers is very large in most of the subgroups: 100% in the first 17 subgroups and usually larger than 50% in the first 100 subgroups, whereas in the last part of the sample the proportion of outliers is very small.

This suggests a general decrease of variability of the stocks in the last part of the sample, which can be observed in Figure 2 (p. 246). Thus, the two groups of data found are consistent with two periods of different variability in the return of the stocks.

In order to understand better the distribution of the data in group A, we made a Q-Q plot of the percentiles of the Mahalanobis distances in this group with respect to the percentiles of a Chi-square with 17 degrees of freedom, the expected distribution of these Mahalanobis distances under the hypothesis of multivariate normal data. This plot, see Figure 5, shows that the distribution of the Mahalanobis distance deviates strongly from the distribution expected under normality. We then generated a sample of the same dimension as the data from a multivariate Student t distribution with 13 degrees of freedom and the same covariance matrix. Figure 5 also shows the Q-Q plot of the Mahalanobis distances

Table 14. Ratio Between the Standard Deviations in Groups B* and A for the Stock Returns

ACS	*ACX*	*ALT*	*AMS*	*ANA*	*BBVA*	*BKT*	*ELE*	*FCC*
1.951	2.067	1.884	2.084	1.821	1.837	2.375	2.150	2.067
FER	*IBE*	*IDR*	*NHH*	*POP*	*REP*	*SGC*	*TEF*	*TPI*
1.751	2.109	2.189	1.834	1.849	2.056	2.084	2.142	2.182

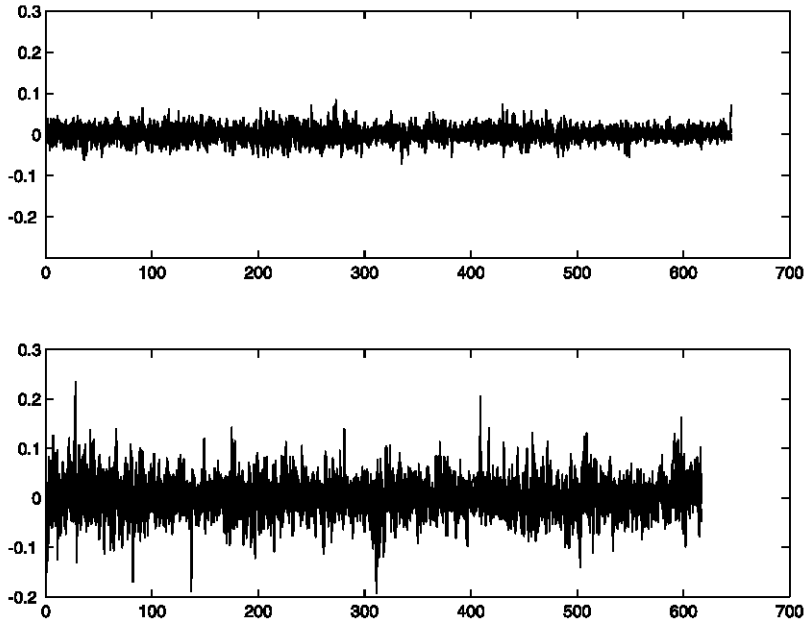

Figure 3. Observations in each of the two groups: A (top) and B (bottom).*

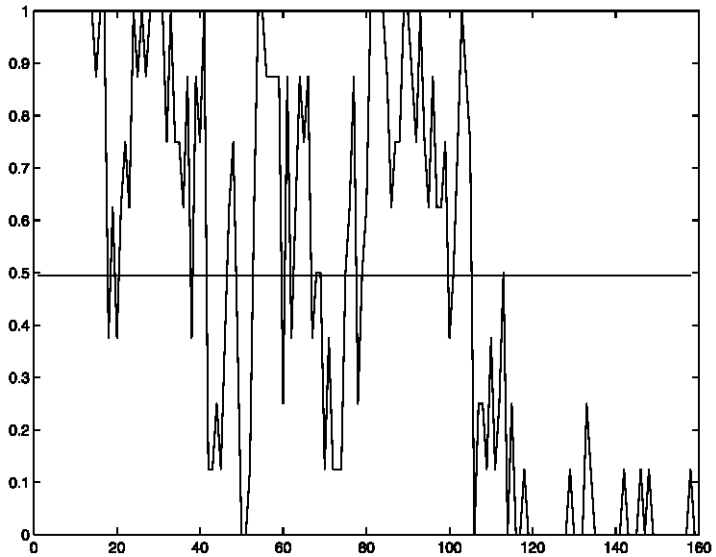

Figure 4. Proportion of observations from group B^* in subgroups of eight consecutive observations versus order of the subgroup.

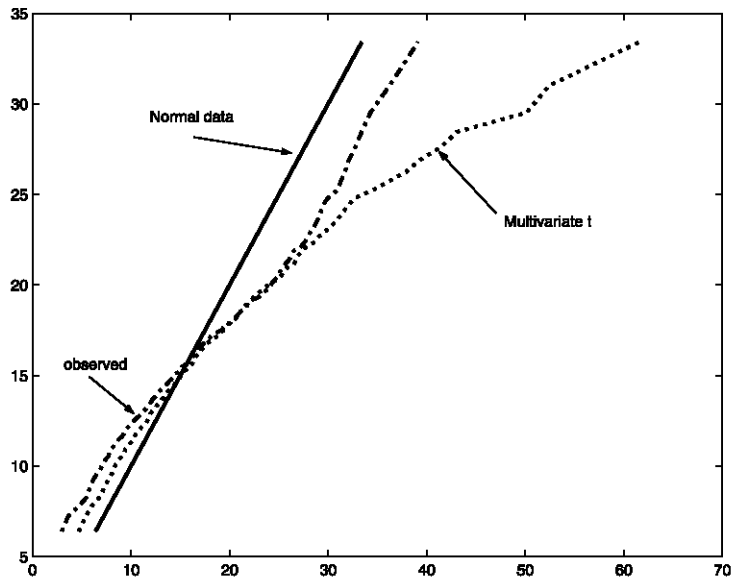

Figure 5. Quantile-Quantile plot of the observed Mahalanobis distance in the first group of data and of the Mahalanobis distance computed from a sample of Multivariate t variables with 13 degrees of freedom with respect to a chi-square.

computed in this simulated sample with respect to the chi-square distribution. We conclude that data in group A are not multivariate normal and that they are more consistent with a multivariate t distribution with 13 degrees of freedom, but truncated at about 30, as the plot of the data is very similar to the plot of the multivariate t in the interval (0, 30). Note that the .99 percentile of a chi-square with 17 degrees of freedom is 33.4, which is the value used as cutoff for outlier detection. Thus the large number of outliers found for the procedure in this dataset can be due to the fact that the data follow approximately a multivariate t, instead of a multivariate normal.

From this plot we conclude that a sensible cutoff for outliers from the multivariate t distribution is about 60. Thus we apply the detection procedure to the whole dataset with this cutoff value and now only 72 outliers are found, which correspond to 5.6% of the sample. Table 15 shows the proportion of outliers in each period. Note that the proportion of outliers decreases over time, which is consistent with the decrease in variability previously found. Figure 7 shows the Q-Q plot of the Mahalanobis distances in the bulk of the data without the 72 outliers against the distances in a sample of the same size and parameters generated from a multivariate t distribution with 13 degrees of freedom; it can be seen that the approximation is reasonable. Finally, Figure 6 shows a plot of the two groups of data finally detected: the main group that seems to follow a multivariate t, and the 72 observations detected as outliers. It can be seen that this later group can be split into the 10 large outliers due to well-known reasons and a set of data which seem to come from a distribution with larger variability than the first group.

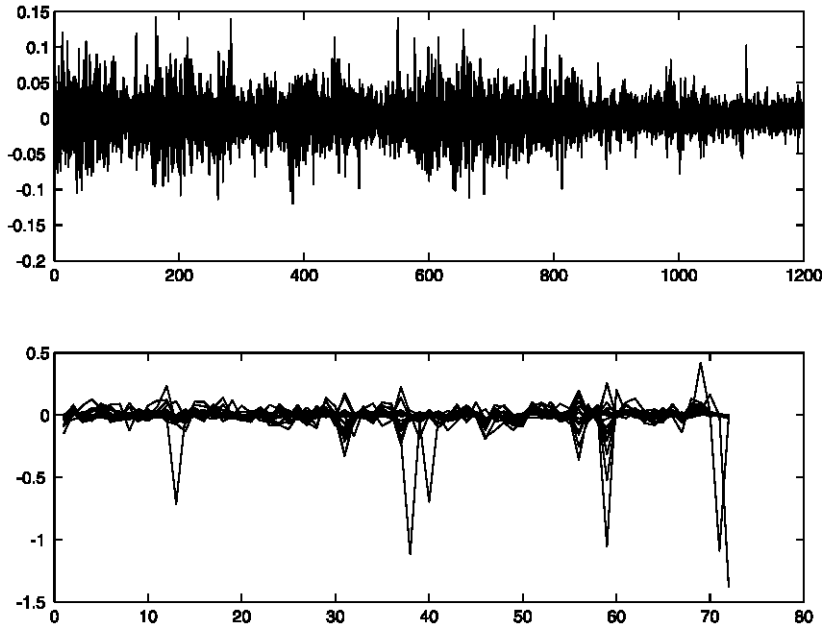

Figure 6. Group of homogeneous data which seem to follow a multivariate t distribution (top) and outliers with respect to this group (bottom).

6. CONCLUSIONS

The analysis presented in the previous sections shows that the combination of random and specific direction leads to a powerful procedure for robust estimation and outlier detection. The random directions are generated by a stratified sampling scheme, which works better than the random sampling of the Stahel-Donoho procedure, especially with high-dimensional data. However, the random directions cannot completely cope with the deficiencies for concentrated contamination. On the other hand, the specific directions obtained by the kurtosis coefficient seem to be very powerful for detecting concentrated contamination. We have shown that if we just compute the two directions corresponding to the extremes of the kurtosis coefficient we have a powerful procedure. The combination of

Table 15. Location of the Outliers With Respect to the Observations Coming from the Multivariate *t* Distribution

Year	*Sample size*	*Outliers*	*%*
2000	250	42	16.8
2001	272	16	5.9
2002	250	11	4.4
2003	250	1	0.4
2004	251	2	0.8

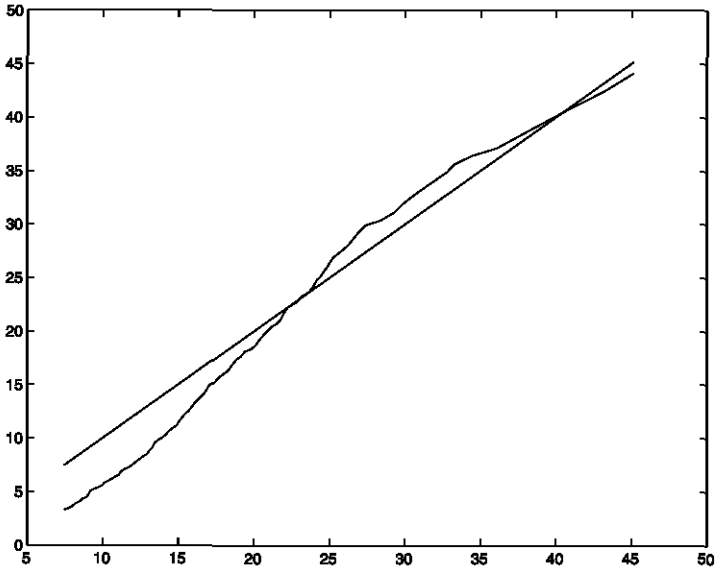

Figure 7. Empirical two-sample QQ plot of the Mahalanobis distance in the sample against those from a sample generated by a multivariate t distribution with 13 degrees of freedom.

both methods in the RASP algorithm seems to be a useful alternative for the routine analysis of multivariate data in high dimensions. It can be applied to reasonably large datasets in many variables and, as it is based on projections, it is not severely affected by the curse of dimensionality. Although we believe this procedure can be applied to large p problems a limitation of our method is that we assume $n > p$ in order to compute the covariance matrix of the observations. Thus, the present version cannot be applied when we have more variables than observations, as in microarray and image analysis. As the SD and the PP directions can be computed just as well when $p > n$, we believe that the procedure can be extended to this situation, although its properties and relative advantages over other methods when $p > n$ will be the subject of further research.

We have emphasized in this article the outlier detection capabilities of the procedure, but the same good properties are found in the robust estimation of the covariance matrix, which has the high breakdown point property of the Stahel-Donoho estimate in finite samples. Many standard multivariate procedures are based on the analysis of the covariance matrix of the data and thus using the robust covariance matrix obtained by this procedure provides a simple way to obtain robust principal components, robust canonical analysis or robust discrimination.

7. APPENDIX

In this Appendix we obtain the distribution of the standardized data $y_i = S^{-1/2}(x_i - \overline{x})$ when $x \sim (1-\alpha)N(0, I) + \alpha N(\delta e_1, \lambda^2 I)$ and give the limiting distribution when $n \to \infty$ and $\delta \to \infty$. Assuming that the distribution of x is $(1-\alpha)N(0, I) + \alpha N(\delta e_1, \lambda^2 I)$, then $E(x) = \alpha\delta e_1$ and the covariance matrix is $V_x = aI + be_1e_1'$ with $a = (1-\alpha) + \alpha\lambda^2$

and $b = (1-\alpha)\alpha\delta^2$. Then $V_x^{-1} = a^{-1}(I - b/(a+b)e_1e_1')$ and $V_x^{-1/2} = a^{-1/2}(I - ce_1e_1')$, where $c = 1 - (a/(a+b))^{1/2}$ and $y = V_x^{-1/2}(x - \alpha\delta e_1)$ has a distribution $(1-\alpha)N(-\alpha\delta(a+b)^{-1/2}e_1, V_x^{-1}) + \alpha N((1-\alpha)\delta(a+b)^{-1/2}e_1, \lambda^2 V_x^{-1})$. Now making $\delta \to \infty$ we have that $y \sim (1-\alpha)N(-\alpha\theta e_1, \sigma^2\bar{I}) + \alpha N((1-\alpha)\theta e_1, \lambda^2\sigma^2\bar{I})$ where $\theta = (\alpha(1-\alpha))^{-1/2}$, $\sigma^2 = (\alpha\lambda^2 + 1 - \alpha)^{-1}$, and $\bar{I}$ denotes an identity matrix with the first element equal to zero. If we transform the data by $z = y + \alpha\theta e_1$, so that the mean of the central distribution is the zero vector, the distribution of the transformed data will be $(1-\alpha)N(0, \sigma^2\bar{I}) + \alpha N(\theta e_1, \lambda^2\sigma^2\bar{I})$.

ACKNOWLEDGMENTS

We are grateful to the referees for helpful comments. This research has been sponsored by MEC grants SEJ2004-03303 and MTM2004-02334.

[Received June 2005. Revised May 2006.]

REFERENCES

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.

Anderson, T. W., and Bahadur, R. R. (1962), "Classification into Two Multivariate Normal Distributions with Different Covariance Matrices," *Annals of Mathematical Statistics*, 33, 420–431.

Atkinson, A. C. (1994), "Fast Very Robust Methods for the Detection of Multiple Outliers," *Journal of the American Statistical Association*, 89, 1329–1339.

Becker, C., and Gather , U. (2001), "The Largest Nonidentifiable Outlier: A Comparison of Simultaneous Outlier Detection Rules," *Computational Statistics and Data Analysis*, 36, 119–127.

Cook, R. D., Hawkins, D. M., and Weisberg, S. (1993), "Exact Iterative Computation of the Robust Multivariate Minimum Volume Ellipsoid Estimator," *Statistics and Probability Letters*, 16, 213–218.

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

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

Gnanadesikan, R., and Kettenring, J. R. (1972), "Robust Estimates, Residuals, and Outliers detection with Multiresponse Data," *Biometrics*, 28, 81–124.

Hadi, A. S. (1992), "Identifying Multiple Outliers in Multivariate Data," *Journal of the Royal Statistical Society*, Series B, 54, 761–771.

Hawkins, D. M. (1994), "The Feasible Solution for the Minimum Covariance Determinant Estimator in Multivariate Data," *Computational Statistical and Data Analysis*, 30, 1–11.

Hawkins, D. M., and Olive, D. J. (2002), "Inconsistency of Resampling Algorithms for High Breakdown Regression Estimators and A New Algorithm," *Journal of the American Statistical Association*, 97, 136–147.

Hyvarinen, A., Karhunen, J., and Oja, E. (2001) *Independent Component Analysis*, New York: Wiley.

Juan, J., and Prieto, F. J. (2001), "Using Angles to Identify Concentrated Multivariate Outliers," *Technometrics*, 43, 311–322.

Maronna, R. A. (1976), "Robust M-Estimators of Multivariate Location and Scatter," *The Annals of Statistics*, 4, 51–67.

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.

Maronna, R. A., and Zamar R, (2002) ”Robust Estimates of Location and Dispersion for High-Dimensional Datasets” *Technometrics*, 44, 307–317.

Peña, D., and Prieto, F. J. (2001a), “Robust Covariance Matrix Estimation and Multivariate Outlier Detection,” *Technometrics*, 43, 286–310.

——— (2001b), “Cluster Identification Using Projections,” *Journal of the American Statistical Association*, 96, 1433–1445.

Rocke, D. M., and Woodruff, D. L. (1993), “Computation of Robust Estimates of Multivariate Location and Shape,” *Statistica Neerlandica*, 47, 27–42.

——— (1996), “Identification of Outliers in Multivariate Data,” *Journal of the American Statistical Association*, 91, 1047–1061.

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, Dordrecht: Reidel, pp. 283–297.

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

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

Stahel, W. A. (1981), “Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen,” Ph.D. Thesis, ETH Zurich.

Tyler, D. E. (1991), “Some Issues in the Robust Estimation of Multivariate Location and Scatter,” in *Directions in Robust Statistics and Diagnostics, Part II*, eds. W. Stahel and S. Weisberg, New York: Springer Verlag.

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