Multivariate Outlier Detection and Robust Covariance Matrix Estimation

Daniel Peña and Francisco J. Prieto

Department of Statistics and Econometrics
Universidad Carlos III de Madrid
28903 Getafe (Madrid)
Spain
(*dpena@est-econ.uc3m.es*) (*fjp@est-econ.uc3m.es*)

In this article, we present a simple multivariate outlier-detection procedure and a robust estimator for the covariance matrix, based on the use of information obtained from projections onto the directions that maximize and minimize the kurtosis coefficient of the projected data. The properties of this estimator (computational cost, bias) are analyzed and compared with those of other robust estimators described in the literature through simulation studies. The performance of the outlier-detection procedure is analyzed by applying it to a set of well-known examples.

KEY WORDS: Kurtosis; Linear projection; Multivariate statistics.

The detection of outliers in multivariate data is recognized to be an important and difficult problem in the physical, chemical, and engineering sciences. Whenever multiple measurements are obtained, there is always the possibility that changes in the measurement process will generate clusters of outliers. Most standard multivariate analysis techniques rely on the assumption of normality and require the use of estimates for both the location and scale parameters of the distribution. The presence of outliers may distort arbitrarily the values of these estimators and render meaningless the results of the application of these techniques. According to Rocke and Woodruff (1996), the problem of the joint estimation of location and shape is one of the most difficult in robust statistics.

Wilks (1963) proposed identifying sets of outliers of size j in normal multivariate data by checking the minimum values of the ratios $|A_{(I)}|/|A|$, where $|A_{(I)}|$ is the internal scatter of a modified sample in which the set of observations I of size j has been deleted and $|A|$ is the internal scatter of the complete sample. The internal scatter is proportional to the determinant of the covariance matrix and the ratios are computed for all possible sets of size j. Wilks computed the distribution of the statistic for j equal to 1 and 2. It is well known that this procedure is a likelihood ratio test and that for $j = 1$ the method is equivalent to selecting the observation with the largest Mahalanobis distance from the center of the data.

Because a direct extension of this idea to sets of outliers larger than 2 or 3 is not practical, Gnanadesikan and Kettenring (1972) proposed to reduce the multivariate detection problem to a set of univariate problems by looking at projections of the data onto some direction. They chose the direction of maximum variability of the data and, therefore, they proposed to obtain the principal components of the data and search for outliers in these directions. Although this method provides the correct solution when the outliers are located close to the directions of the principal components, it may fail to identify outliers in the general case.

An alternative approach is to use robust location and scale estimators. Maronna (1976) studied affinely equivariant M estimators for covariance matrices, and Campbell (1980) proposed using the Mahalanobis distance computed using M estimators for the mean and covariance matrix. Stahel (1981) and Donoho (1982) proposed to solve the dimensionality problem by computing the weights for the robust estimators from the projections of the data onto some directions. These directions were chosen to maximize distances based on robust univariate location and scale estimators, and the optimal values for the distances could also be used to weigh each point in the computation of a robust covariance matrix. To ensure a high breakdown point, one global optimization problem with discontinuous derivatives had to be solved for each data point, and the associated computational cost became prohibitive for large high-dimensional datasets. This computational cost can be reduced if the directions are generated by a resampling procedure of the original data, but the number of directions to consider still grows exponentially with the dimension of the problem.

A different procedure was proposed by Rousseeuw (1985) based on the computation of the ellipsoid with the smallest volume or with the smallest covariance determinant that would encompass at least half of the data points. This procedure has been analyzed and extended in a large number of articles; see, for example, Hampel, Ronchetti, Rousseeuw, and Stahel (1986), Rousseeuw and Leroy (1987), Davies (1987), Rousseeuw and van Zomeren (1990), Tyler (1991), Cook, Hawkins, and Weisberg (1993), Rocke and Woodruff (1993, 1996), Maronna and Yohai (1995), Agulló (1996), Hawkins and Olive (1999), Becker and Gather (1999), and Rousseeuw and Van Driessen (1999). Public-domain codes implementing these procedures can be found in STATLIB—for example, the code FSAMVE from Hawkins (1994) and MULTOUT from Rocke and Woodruff (1993, 1996). FAST-MCD from Rousseeuw and Van Driessen is implemented as the "mcd.cov" function of S-PLUS.

Because these procedures are based on the minimization of certain nonconvex and nondifferentiable criteria, these estimators are computed by resampling. For example, Rousseeuw (1993) proposed selecting p observations from the original sample and computing the direction orthogonal to the hyperplane defined by these observations. The maximum over this finite set of directions is used as an approximation to the exact solution. Unfortunately, the number of candidate solutions grows exponentially with the size of the problem and, as a consequence, the corresponding procedures become computationally expensive for even moderately sized problems. Hadi (1992, 1994), Atkinson (1994), Hawkins and Olive (1999), and Rousseeuw and van Driessen (1999) presented methods to compute approximations for these estimates requiring reasonable computation times.

In this article, we present an alternative procedure, based on the analysis of the projections of the sample points onto a certain set of $2p$ directions, where p is the dimension of the sample space. These directions are obtained by maximizing and minimizing the kurtosis coefficient of the projections. The proposed procedure can be seen as an empirically successful and faster way of implementing the Stahel–Donoho (SD) algorithm. The justification for using these directions is presented in Section 1. Section 2 describes the proposed procedure and illustrates its behavior on an example. Section 3 compares it to other procedures by a simulation study. It is shown that the proposed procedure works well in practice, is simple to implement, and requires reasonable computation times, even for large problems. Finally, Section 4 presents some conclusions.

1. KURTOSIS AND OUTLIERS

The idea of using projections to identify outliers is the basis for several outlier-detection procedures. These procedures rely on the fact that in multivariate contaminated samples each outlier must be an extreme point along the direction from the mean of the uncontaminated data to the outlier. Unfortunately, high-breakdown-point methods developed to date along these lines, such as the SD algorithm, require projecting the data onto randomly generated directions and need very large numbers of directions to be successful. The efficiency of these methods could be significantly improved, at least from a computational point of view, if a limited number of appropriate directions would suffice to identify the outliers. Our proposal is to choose these directions based on the values of the kurtosis coefficients of the projected observations.

In this section, we study the impact of the presence of outliers on the kurtosis values and the use of this moment coefficient to identify them. We start by considering the univariate case in which different types of outliers produce different effects on the kurtosis coefficient. Outliers generated by the usual symmetric contaminated model increase the kurtosis coefficient of the observed data. A small proportion of outliers generated by an asymmetric contaminated model also increase the kurtosis coefficient of the observed data. These two results suggest that for multivariate data outliers may be revealed on univariate projections onto directions obtained by maximizing the kurtosis coefficient of the projected data. However, a large proportion of outliers generated by an asymmetric contamination model can make the kurtosis coefficient of the data very small, close to its minimum possible value. This result suggests searching for outliers also using directions obtained by minimizing the kurtosis of the projections. Therefore, a procedure that would search for outliers by projecting the data onto the directions that maximize or minimize the kurtosis of the projected points would seem promising.

In univariate normal data, outliers have often been associated with large kurtosis values, and some well-known tests of normality are based on the asymmetry and kurtosis coefficients. These ideas have also been used to test for multivariate normality (see Malkovich and Afifi 1973). Additionally, some projection indices that have been applied in projection pursuit algorithms are related to the third and fourth moments (Jones and Sibson 1987; Posse 1995). Hampel (1985) derived the relationship between the critical value and the breakdown point of the kurtosis coefficient in univariate samples. He also showed that two-point distributions are the least favorable for detecting univariate outliers using the kurtosis coefficient.

To understand the effect of different types of outliers on the kurtosis coefficient, suppose that we have a sample of univariate data from a random variable that has a distribution F with finite moments (the uncontaminated sample). We assume without loss of generality that $\mu_F = \int x dF(x) = 0$, and we will use the notation $m_F(j) = \int x^j dF(x)$. The sample is contaminated by a fraction $\alpha < 1/2$ of outliers generated from some contaminating distribution G, with $\mu_G = \int x dG(x)$, and we will denote the centered moments of this distribution by $m_G(j) = \int (x-\mu_G)^j dG(x)$. Therefore, the resulting observed random variable X follows a mixture of two distributions, $(1-\alpha)F + \alpha G$. The signal-to-noise ratio will be given by $r^2 = \mu_G^2/m_F(2)$, and the ratio of variances of the two distributions will be $v^2 = m_G(2)/m_F(2)$. The third- and fourth-order moment coefficients for the mixture and the original and contaminating distributions will be denoted by $a_i = m_i(3)/m_i^{3/2}(2)$ and $\gamma_i = m_i(4)/m_i^2(2)$ for $i = X, F, G$, respectively. The ratio of the kurtosis coefficients of G and F will be $\theta = \gamma_G/\gamma_F$.

Some conditions must be introduced on F and G to ensure that this is a reasonable model for outliers. The first condition is that, for any values of the distribution parameters, the standard distribution F has a bounded kurtosis coefficient, γ_F. This bound avoids the situation in which the tails of the standard distribution are so heavy that extreme observations, which cannot be distinguished from outliers, will appear with significant probability. Note that the most often used distributions (normal, Student's t, gamma, beta, ...) satisfy this condition. The second condition is that the contaminating distribution G is such that

$$\mu_G m_G(3) \geq 0; \tag{1}$$

that is, if the distribution is not symmetric, the relevant tail of G for the generation of outliers and the mean of G both lie on the same side with respect to the mean of F. This second assumption avoids situations in which most of the observations generated from G might not be outliers.

To analyze the effect of outliers on the kurtosis coefficient, we write its value for the contaminated population as (see the appendix for the derivation)

$$\gamma_X = \frac{\gamma_F + \alpha(1-\alpha)(c_4 + 4rc_3 + 6r^2c_2 + r^4c_0)}{h_0 + h_1r^2 + h_2r^4}, \tag{2}$$

where $c_4 = \gamma_F(\theta v^4 - 1)/(1-\alpha)$, $c_3 = a_G v^3 - a_F$, $c_2 = \alpha + (1-\alpha)v^2$, $c_0 = \alpha^3 + (1-\alpha)^3$, $h_0 = (1+\alpha(v^2-1))^2$, $h_1 = 2\alpha(1-\alpha)h_0^{1/2}$, and $h_2 = \alpha^2(1-\alpha)^2$. We consider the two following cases:

1. The centered case, in which we suppose that both F and G have the same mean, and as a consequence $\mu_G = 0$ and $r = 0$. From (2), we obtain for this case

$$\gamma_X = \frac{\gamma_F(1+\alpha(\theta v^4 - 1))}{(1+\alpha(v^2-1))^2}.$$

Note that the kurtosis coefficient increases due to the presence of outliers; that is, $\gamma_X \geq \gamma_F$ whenever $\theta v^4 - 2v^2 + 1 \geq \alpha(v^2-1)^2$. This holds if $\theta \geq 1$, or equivalently if $\gamma_G \geq \gamma_F$, and under this condition the kurtosis will increase for any value of α. Thus in the usual situation in which the outlier model is built by using a contaminating distribution of the same family as the original distribution [as in the often-used normal scale-contaminated model; see, for instance, Box and Tiao (1968)] or with heavier tails, the kurtosis coefficient of the observed data is expected to be larger than that of the original distribution.

2. Consider now the noncentered case, in which both distributions are arbitrary and we assume that the means of G and F are different ($\mu_G \neq 0$). A reasonable condition to ensure that G will generate outliers for F is that the signal-to-noise ratio r is large enough. If we let $r \to \infty$ in (2) (and we assume that the moment coefficients in the expression remain bounded), we obtain

$$\gamma_X \to \frac{\alpha^3 + (1-\alpha)^3}{\alpha(1-\alpha)}.$$

This result agrees with the one obtained by Hampel (1985). Note that if $\alpha = .5$ the kurtosis coefficient of the observed data will be equal to 1, the minimum possible value. On the other hand, if $\alpha \to 0$ the kurtosis coefficient increases without bound and will become larger than γ_F, which is bounded. Therefore, in the asymmetric case, if the contamination is very large the kurtosis coefficient will be very small, whereas if the contamination is small the kurtosis coefficient will be large.

The preceding results agree with the dual interpretation of the standard fourth-moment coefficient of kurtosis (see Ruppert 1987; Balanda and MacGillivray 1988) as measuring tail heaviness and lack of bimodality. A small number of outliers will produce heavy tails and a larger kurtosis coefficient. But, if we increase the amount of outliers, we can start introducing bimodality and the kurtosis coefficient may decrease.

Kurtosis and Projections

The preceding discussion centered on the behavior of the kurtosis coefficient in the univariate case as an indicator for the presence of outliers. A multivariate method to take advantage of these properties would proceed through two stages—determining a set of projection directions to obtain univariate samples and then conducting an analysis of these samples to determine if any outlier may be present in the original sample. As indicated in the introduction, this is the approach developed by Stahel and Donoho. In this section we will show how to find interesting directions to detect outliers.

The study of the univariate kurtosis coefficient indicates that the presence of outliers in the projected data will imply particularly large (or small) values for the kurtosis coefficient. As a consequence, it would be reasonable to use as projection directions those that maximize or minimize the kurtosis coefficient of the projected data. For a standard multivariate contamination model, we will show that these directions are able to identify a set of outliers.

Consider a p-dimensional random variable X following a (contaminated normal) distribution given as a mixture of normals of the form $(1-\alpha)\mathrm{N}(0, I) + \alpha\mathrm{N}(\delta e_1, \lambda I)$, where e_1 denotes the first unit vector. This contamination model is particularly difficult to analyze for many outlier-detection procedures, (e.g., see Maronna and Yohai 1995). Moreover, the analysis in the preceding section indicates that this model, for noncentered contaminating distributions, may correspond to an unfavorable situation from the point of view of the kurtosis coefficient.

Since the kurtosis coefficient is invariant to affine transformations, we will center and scale the variable to ensure that it has mean 0 and covariance matrix equal to the identity. This transformed variable, Y, will follow a distribution of the form $(1-\alpha)\mathrm{N}(m_1, S) + \alpha\mathrm{N}(m_2, \lambda S)$, where

$$m_1 = -\alpha\delta S^{1/2} e_1, \quad m_2 = (1-\alpha)\delta S^{1/2} e_1$$
$$\nu_1 = 1 - \alpha(1-\lambda), \quad \nu_2 = \frac{\delta^2\alpha(1-\alpha)}{\nu_1 + \delta^2\alpha(1-\alpha)}$$
$$S = \frac{1}{\nu_1}(I - \nu_2 e_1 e_1') \tag{3}$$

and ν_1 and ν_2 denote auxiliary parameters, introduced to simplify the expressions (see the appendix for a derivation of these values).

We wish to study the behavior of the univariate projections for this variable and their kurtosis coefficient values. Consider an arbitrary projection direction u. Using the affine invariance of the kurtosis coefficient, we will assume $\|u\| = 1$. The projected univariate random variable $Z = u'Y$ will follow a distribution $(1-\alpha)\mathrm{N}(m_1'u, u'Su) + \alpha\mathrm{N}(m_2'u, \lambda u'Su)$, with $E(Z) = 0$, and $E(Z^2) = 1$. The kurtosis coefficient of Z will be given by

$$\gamma_Z(\omega) = a(\alpha, \delta, \lambda) + b(\alpha, \delta, \lambda)\omega^2 + c(\alpha, \delta, \lambda)\omega^4, \tag{4}$$

where the coefficients a, b, and c correspond to

$$a(\alpha, \delta, \lambda) = \frac{3}{\nu_1^2}(1 - \alpha + \alpha\lambda^2)$$
$$b(\alpha, \delta, \lambda) = \frac{6\nu_2}{\nu_1^2}(1-\lambda)(\alpha^2\lambda - (1-\alpha)^2)$$
$$c(\alpha, \delta, \lambda) = \nu_2^2\left(3\frac{1-\alpha+\alpha\lambda^2}{\nu_1^2} - 6\frac{\alpha + \lambda(1-\alpha)}{\nu_1} + \frac{\alpha^3 + (1-\alpha)^3}{\alpha(1-\alpha)}\right) \tag{5}$$

and $\omega \equiv u_1 = e_1'u$ (see the appendix for details on this derivation).

We wish to study the relationship between the direction to the outliers, e_1 in our model and the directions u that correspond to extremes for the projected kurtosis coefficient.

The optimization problem defining these extreme directions would be either

$$\begin{array}{ll} \max\limits_{\omega} & \gamma_Z(\omega) \\ \text{s.t.} & -1 \le \omega \le 1 \end{array} \qquad (6)$$

or the equivalent minimization problem.

From the first-order optimality conditions for (6), the extremes may correspond to either a point in the interval $(-1, 1)$ such that $\gamma_Z'(\omega) = 0$ or to the extreme points of the interval, $\omega = \pm 1$. Since $\gamma_Z' = 4c\omega^3 + 2b\omega$, the points that make this derivative equal to 0 are $\omega = 0$ and $\omega = \pm\sqrt{-b/(2c)}$. We now analyze in detail the nature of each of these three possible extreme points.

1. $\omega = \pm 1$—that is, the direction of the outliers. This direction corresponds to a local maximizer whenever $4c + 2b > 0$ (the derivative at $\omega = \pm 1$) and to a minimizer whenever $4c + 2b < 0$ (the case in which $4c + 2b = 0$ will be treated when considering the third candidate to an extreme point). The expression for $4c + 2b$ from (5) is quite complex to analyze in the general case, but for the case of small contamination levels ($\alpha \to 0$) it holds that $\nu_1 \to 1$, $\nu_2/\alpha \to \delta^2$, and

$$\lim_{\alpha\to 0} \frac{4c+2b}{\alpha} = 4\delta^4 + 12\delta^2(\lambda - 1),$$

implying that, since this value is positive except for very small values of δ and $\lambda (\lambda \le 1 - \delta^2/3)$, for small values of α the direction of the outliers will be a maximizer for the kurtosis coefficient. Moreover, for large contamination levels ($\alpha \to 1/2$), after some manipulation of the expressions in (5) it holds that

$$\lim_{\alpha\to 1/2} (4c + 2b) = -8\delta^2 \frac{3(\lambda-1)^2 + \delta^2(\lambda+1)}{(\lambda+1)(2+2\lambda+\delta^2)^2} < 0,$$

and, as a consequence, if α is large we always have a minimizer along the direction of the outliers.

2. $\omega = 0$, a direction orthogonal to the outliers. Along this direction, as $\gamma_Z''(0) = 2b$, the kurtosis coefficient has a maximizer whenever $b < 0$. For small contaminations, from $\lim_{\alpha\to 0} b/\alpha = 6\delta^2(\lambda - 1)$, the kurtosis coefficient has a maximizer for $\lambda < 1$ and a minimizer for $\lambda > 1$. Comparing the kurtosis coefficient values when the direction to the outliers and a direction orthogonal to it are both local maximizers (when $\lambda < 1$), from (4) and

$$\lim_{\alpha\to 0} \frac{\gamma_Z(\pm 1) - \gamma_Z(0)}{\alpha} = \delta^2(\delta^2 - 6(1-\lambda)),$$

it follows that $\omega = \pm 1$ corresponds to the global maximizer, except for very small values of δ. For large contaminations, from

$$\lim_{\alpha\to 1/2} b = -6\delta^2 \left(\frac{\lambda-1}{\lambda+1}\right)^2 \frac{1}{\delta^2 + 2\lambda + 2},$$

the kurtosis coefficient has a maximizer at $\omega = 0$.

3. $\omega = \pm\sqrt{-b/2c}$, an intermediate direction, if this value lies in the interval $[-1, 1]$—that is, if $0 < -b/2c < 1$. For small contamination levels ($\alpha \to 0$), it holds that

$$\lim_{\alpha\to 0} -\frac{b}{2c} = 3\frac{1-\lambda}{\delta^2},$$

and this local extreme point exists whenever $1 - \delta^2/3 < \lambda < 1$—that is, basically when the dispersion of the contamination is smaller than 1. Additionally, since in this case $\gamma_Z'' = -4b$ for $\lambda < 1$, it holds that $b < 0$ whenever $\lambda < 1$, implying $\gamma_Z'' > 0$. Consequently, for small concentrated contaminations this additional extreme point exists and is a minimizer. For large contamination levels, it holds that

$$\lim_{\alpha\to 1/2} -\frac{b}{2c} = 3\left(\frac{\lambda-1}{\delta}\right)^2 \frac{2+2\lambda+\delta^2}{1-10\lambda+\lambda^2}.$$

For $\lambda \ge 0$ and any δ, this expression is either negative or larger than 1. As a consequence, no extreme intermediate direction exists if the contamination is large.

Table 1. Extreme Directions for the Concentrated Contamination Model

	Small contamination		
Direction	$\lambda < 1$	$\lambda > 1$	*Large cont.*
$\omega = \pm 1$	Global max.	Global max.	Global min.
$\omega = 0$	Local max.	Global min.	Global max.
$\omega = \pm\sqrt{-b/2c}$	Global min.	—	—

Table 1 provides a brief summary of the preceding results. Entries "Global max." and "Global min." indicate if a given direction is the global maximizer or the global minimizer of the kurtosis coefficient, respectively. "Local max." indicates the case in which the direction orthogonal to the outliers is a local maximizer for the kurtosis coefficient.

To detect the outliers, the procedure should be able to compute the direction to the outliers ($\omega = \pm 1$) from Problem (6). However, from Table 1, to obtain this direction we would need both the global minimizer and the global maximizer of (6). In practice this cannot be done efficiently; as an alternative solution, we compute one local minimizer and one local maximizer. These computations can be done with reasonable computational effort, as we describe later on. This would not ensure obtaining the direction to the outliers because in some cases these two directions might correspond to a direction orthogonal to the outliers and the intermediate direction, for example, but we are assured of obtaining either the direction to the outliers or a direction orthogonal to it. The computation procedure is then continued by projecting the data onto a subspace orthogonal to the computed directions, and additional directions are obtained by solving the resulting optimization problems. In summary, we will compute one minimizer and one maximizer for the projected kurtosis coefficient, project the data onto an orthogonal subspace, and repeat this procedure until $2p$ directions have been computed. For the case analyzed previously, this procedure should ensure that the direction to the outliers is one of these $2p$ directions.

Note that, although we have considered a special contamination pattern, this suggested procedure also seems reasonable in those cases in which the contamination patterns are different—for example, when more than one cluster of contaminating observations is present.

2. DESCRIPTION OF THE ALGORITHM

We assume that we are given a sample $(x_1, \ldots, x_n)$ of a p-dimensional vector random variable X. The proposed procedure is based on projecting each observation onto a set of $2p$ directions and then analyzing the univariate projections onto these directions in a similar manner to the SD algorithm. These directions are obtained as the solutions of $2p$ simple smooth optimization problems, as follows:

1. The original data are rescaled and centered. Let $\bar{x}$ denote the mean and S the covariance matrix of the original data; then the points are transformed using

$$y_i = S^{-1/2}(x_i - \bar{x}), \qquad i = 1, \ldots, n. \tag{7}$$

2. Compute p orthogonal directions and projections maximizing the kurtosis coefficient.

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

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

$$\begin{array}{ll} d_j = \arg\max\limits_{d} & \dfrac{1}{n}\sum\limits_{i=1}^{n}\left(d' y_i^{(j)}\right)^4 \\ \text{s.t.} & d'd = 1 \end{array}. \tag{8}$$

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

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

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

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

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 $j < p$, we go back to step 2b. Otherwise, we let $z_i^{(p)} = y_i^{(p)}$.

3. We compute another set of p orthogonal directions and projections minimizing the kurtosis coefficient.

a. Reset $y_i^{(p+1)} = y_i$ and $j = p + 1$.

b. The preceding steps 2b and 2c are repeated, but now instead of (8) we solve the minimization problem

$$\begin{array}{ll} d_j = \arg\min\limits_{d} & \dfrac{1}{n}\sum\limits_{i=1}^{n}\left(d' y_i^{(j)}\right)^4 \\ \text{s.t.} & d'd = 1 \end{array} \tag{9}$$

to compute the projection directions.

4. To determine if $z_i^{(j)}$ is an outlier in any one of the $2p$ directions, we compute a univariate "measure of outlyingness" for each observation as

$$r_i = \max_{1 \le j \le 2p} \frac{|z_i^{(j)} - \text{median}(z^{(j)})|}{\text{MAD}\,(z^{(j)})}. \tag{10}$$

5. These measures r_i are used to test if a given observation is considered to be an outlier. If $r_i > \beta_p$, then observation i is suspected of being an outlier and labeled as such. The cutoff values β_p are chosen to ensure a reasonable level of Type I errors and depend on the sample space dimension p.

Table 2. Cutoff Values for Univariate Projections

Sample space dimension p	5	10	20
Cutoff value β_p	4.1	6.9	10.8

6. If the condition in Step 5 were satisfied for some i, a new sample composed of all observations i such that $r_i \le \beta_p$ is formed, and the procedure is applied again to the reduced sample. This is repeated until either no additional observations satisfy $r_i > \beta_p$ or the number of remaining observations would be less than $\lfloor (n + p + 1)/2 \rfloor$.

7. Finally, 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

$$\tilde{m} = \frac{1}{|U|}\sum_{i \in U} x_i,$$

$$\widetilde{S} = \frac{1}{|U| - 1}\sum_{i \in U}(x_i - \tilde{m})(x_i - \tilde{m}'),$$

and

$$v_i = (x_i - \tilde{m})'\widetilde{S}^{-1}(x_i - \tilde{m}) \qquad \forall\ i \notin U.$$

Those observations $i \notin U$ such that $v_i < \chi^2_{p,.99}$ are considered not 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).

The values of β_p in Step 5 of the algorithm have been obtained from simulation experiments 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 $\log p$.

2.1 Computation of the Projection Directions

The main computational effort in the application of the preceding algorithm is associated with the determination of local solutions for either (8) or (9). This computation can be conducted in several ways:

1. Applying a modified version of Newton's method.

2. Obtaining the solution directly from the first-order optimality conditions. The optimality conditions for both problems are

$$4\sum_{i=1}^{n}(d' y_i^{(j)})^3 y_i^{(j)} - 2\lambda d = 0$$

$$d'd = 1.$$

Multiplying the first equation by d and replacing the constraint, we obtain the value of λ. The resulting condition is

$$\left(\sum_{i=1}^{n}(d' y_i^{(j)})^2 y_i^{(j)} y_i^{(j)'}\right) d = \sum_{i=1}^{n}(d' y_i^{(j)})^4 d. \tag{11}$$

This equation indicates that the optimal d will be a unit eigenvector of the matrix

$$M(d) \equiv \sum_{i=1}^{n}(d' y_i^{(j)})^2 y_i^{(j)} y_i^{(j)'},$$

Table 3. Cutoff Values for Univariate Projections

Sample space dimension p	5	10	20
Scaling factor k_p	.98	.95	.92

that is, of a weighted covariance matrix for the sample, with positive weights (depending on d). Moreover since the eigenvalue at the solution is the value of the fourth moment, we are interested in computing the eigenvector corresponding to the largest or smallest eigenvalue.

In summary, an iterative procedure to compute the direction d proceeds through the following steps:

1. Select an initial direction $\bar{d}_0$ such that $\|\bar{d}_0\|^2 = 1$.
2. In iteration $l+1$, compute $\bar{d}_{l+1}$, as the unit eigenvector associated with the largest (smallest) eigenvalue of $M(\bar{d}_l)$.
3. Terminate whenever $\|\bar{d}_{l+1} - \bar{d}_l\| < \epsilon$, and set $d_j = \bar{d}_{l+1}$.

Another relevant issue is the definition of the initial direction $\bar{d}_0$. Our choice has been to start with the direction corresponding to the largest (when computing maximizers for the kurtosis) or the smallest (when minimizing) principal components of the normalized observations $y_i^{(j)}/\|y_i^{(j)}\|$. These directions have the property that, once the observations have been standardized, they are affine equivariant. They would also correspond to directions along which the observations projected onto the unit hypersphere seem to present some relevant structure; this would provide a reasonable starting point when the outliers are concentrated, for example. A more detailed discussion on the motivation for this choice of initial directions was given by Juan and Prieto (1997).

2.2 Robust Covariance Matrix Estimation

Once the observations have been labeled either as outliers or as part of the uncontaminated sample, following the procedure described previously, it is possible to generate robust estimates for the mean and covariance of the data as the mean and covariance of the uncontaminated observations. This approach, as opposed to the use of weight functions, seems reasonable given the reduced Type I errors associated with the procedure. Nevertheless, note that it is necessary to correct the covariance estimator to account for the bias associated with these errors.

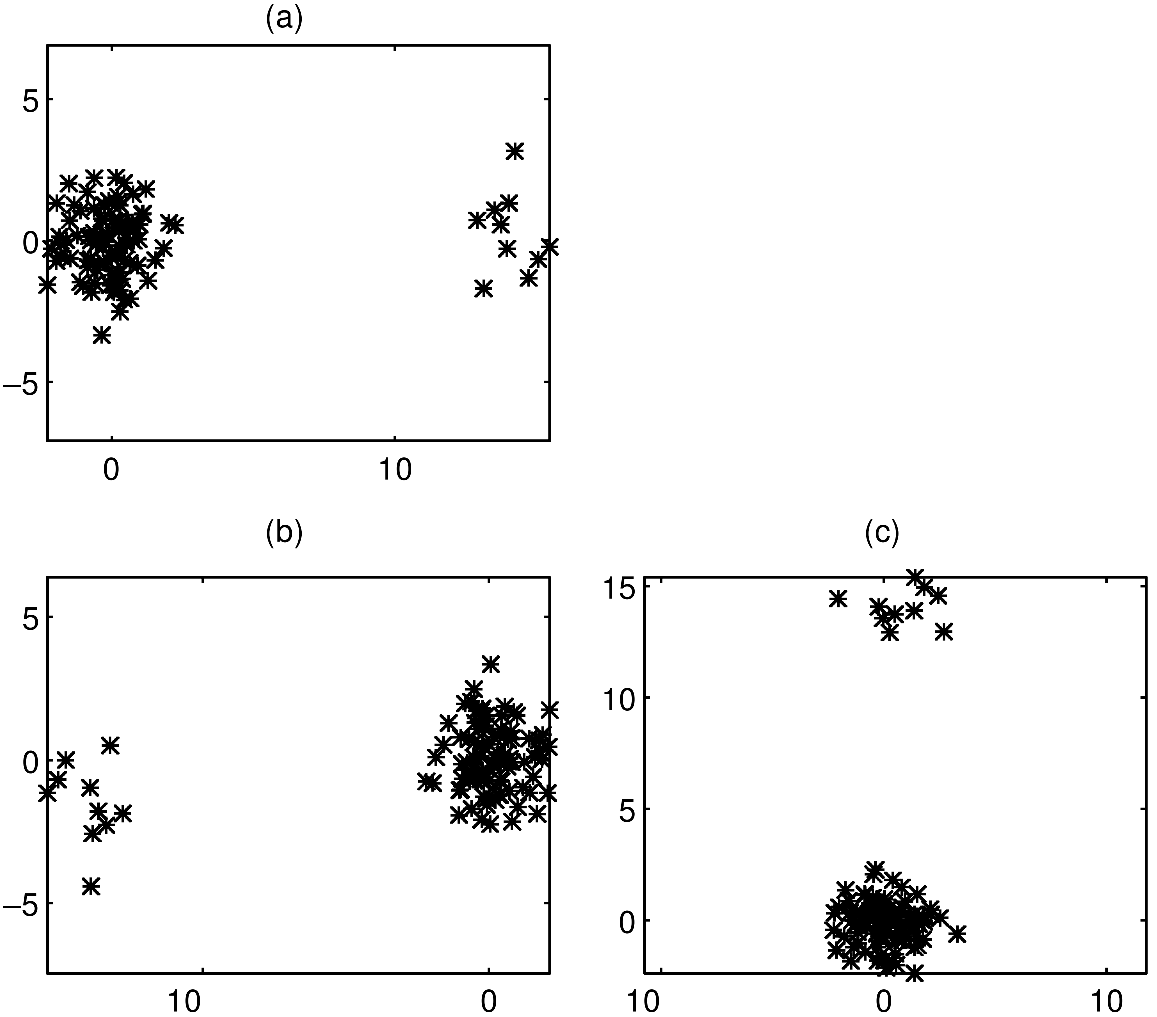

Figure 1. Scatterplots for a Dataset With $\alpha = .1$. The axes correspond to projections onto (a) the direction to the outliers (x axis) and orthogonal direction (y axis), (b) the direction maximizing the kurtosis coef' cient (x axis) and an orthogonal direction (y axis), (c) the direction minimizing the kurtosis coef' cient (x axis) and an orthogonal direction (y axis).

The proposed estimators become

$$\tilde{m} = \frac{1}{|U|}\sum_{i\in U} x_i$$

and

$$\widetilde{S}_c = \frac{1}{(|U|-1)k_d}\sum_{i\in U}(x_i - \tilde{m})(x_i - \tilde{m})',$$

where U is the set of all observations not labeled as outliers, $|U|$ denotes the number of observations in this set, and k_d is a constant that has been estimated to ensure that the trace of the estimated matrix is unbiased.

The values of k_d have been obtained through a simulation experiment for several sample space dimensions and are given in Table 3. The values for other dimensions could be obtained by interpolating $\log k_p$ linearly in $\log p$.

2.3 Examples

To illustrate the procedure and the relevance of choosing projection directions in the manner described previously, we show the results from the computation of the projection directions for a few simple cases. The first ones are based on generating 100 observations from a model of the form $(1-\alpha)\mathrm{N}(0, I) + \alpha\mathrm{N}(10e, I)$ in dimension 2, where $e = (1\ 1)'$, for different values of α.

Consider Figure 1, corresponding to the preceding model with $\alpha = .1$. This figure shows the scatterplots of the data, where the axes have been chosen as (1) in Figure 1(a), the direction to the outliers (e) and a direction orthogonal to it; (2) in Figure 1(b), the direction giving a maximizer for the kurtosis coefficient (the x axis) and a direction orthogonal to it; and (3) in Figure 1(c), the direction corresponding to a minimizer for the kurtosis coefficient (also for the x axis) and a direction orthogonal to it. In this case, the direction maximizing the kurtosis coefficient allows the correct identification of the outliers, in agreement with the results in Table 1 for the case with small α.

Figure 2 shows another dataset, this time corresponding to $\alpha = .3$, in the same format as Figure 1. As the analysis in the preceding section showed and the figure illustrates, here the relevant direction to identify the outliers is the one minimizing the kurtosis coefficient, given the large contamination present.

Finally, Figure 3 presents a dataset generated from the model using $\alpha = .2$ in the same format as the preceding

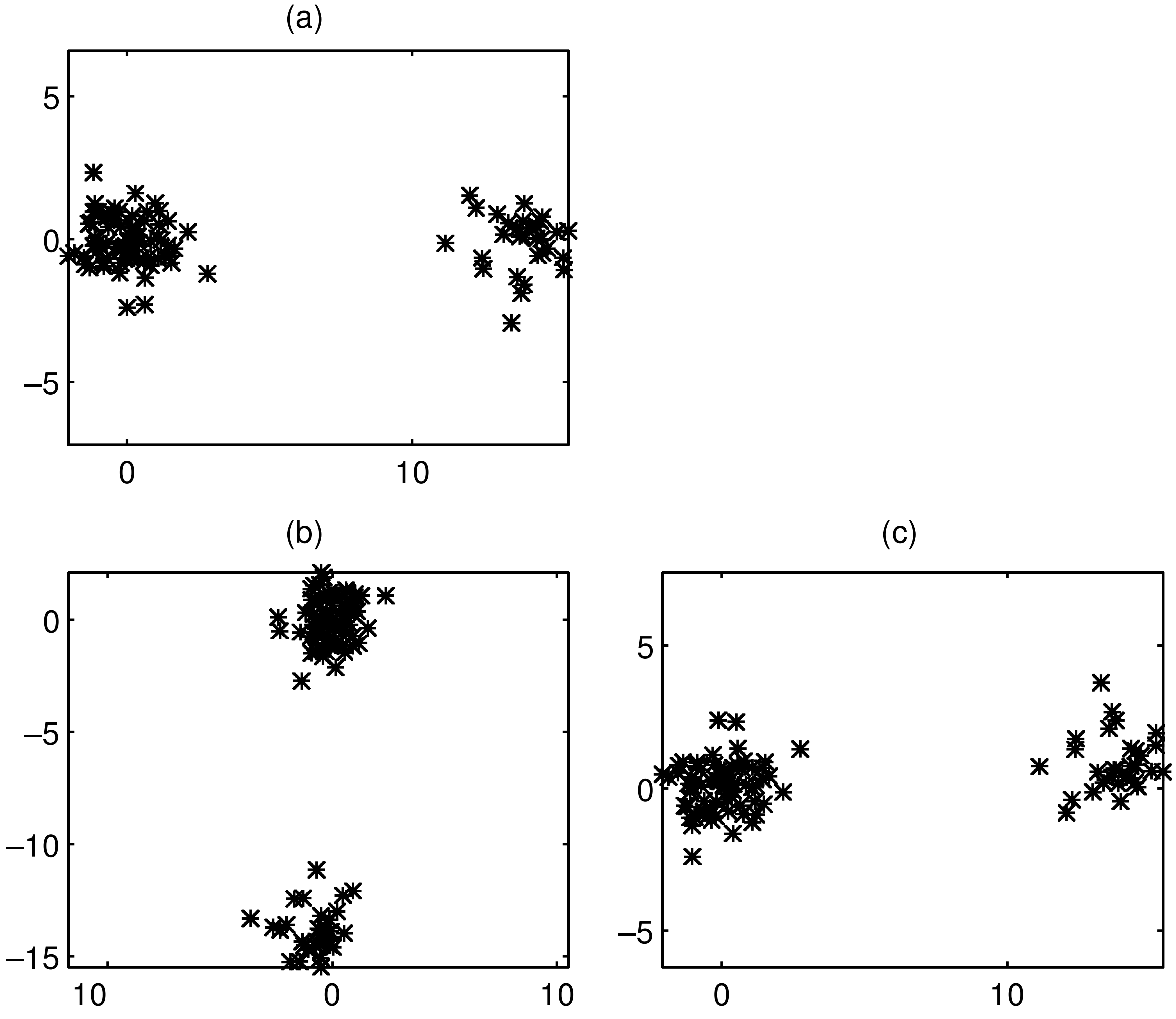

Figure 2. Scatterplots for a Dataset With $\alpha = .3$. The axes correspond to projections onto (a) the direction to the outliers (x axis) and an orthogonal direction (y axis), (b) the direction maximizing the kurtosis coefficient (x axis) and an orthogonal direction (y axis), (c) the direction minimizing the kurtosis coefficient (x axis) and an orthogonal direction (y axis).

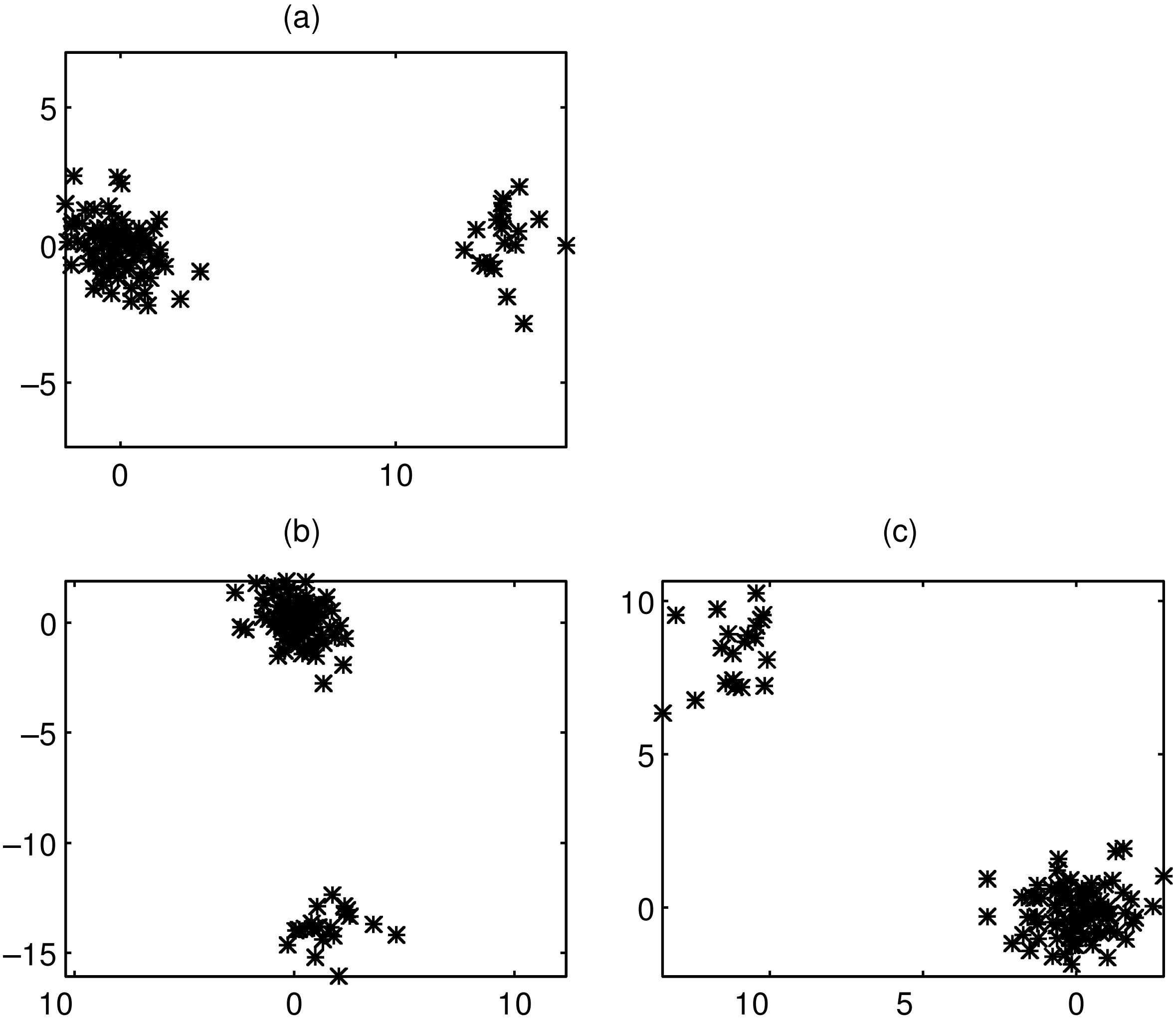

Figure 3. Scatterplots for a Dataset With $\alpha = .2$. The axes correspond to projections onto (a) the direction to the outliers (x axis) and an orthogonal direction (y axis), (b) the direction maximizing the kurtosis coefficient (x axis) and an orthogonal direction (y axis), (c) the direction minimizing the kurtosis coefficient (x axis) and an orthogonal direction (y axis).

figures. It is remarkable in this case that both the direction maximizing the kurtosis coefficient and the direction minimizing it are not the best ones for identifying the outliers; instead, the direction orthogonal to that maximizing the kurtosis corresponds now to the direction to the outliers. The optimization procedure has computed a direction orthogonal to the outliers as the maximizer and an intermediate direction as the minimizer. As a consequence, the direction to the outliers is obtained once Problem (6) is solved for the observations projected onto the direction maximizing the kurtosis. This result justifies that in some cases (for intermediate contamination levels) it is important to compute directions orthogonal to those corresponding to extremes in the kurtosis coefficient, and this effect becomes even more significant as the sample-space dimension increases.

Consider a final example in higher dimension. A sample of 100 observations has been obtained by generating 60 observations from an $\mathrm{N}(0, I)$ distribution in dimension 10, and 10 observations each from $\mathrm{N}(10d_i, I)$ distributions for $i = 1, \ldots, 4$, where d_i were distributed uniformly on the unit hypersphere. Figure 4 shows the projections of these observations onto four of the directions obtained from the application of the proposed procedure. Each plot gives the value of the projection onto one of the directions for each observation in the sample. The outliers are the last 40 observations and have been plotted using the symbols "+," "o," "#," and "×" for each of the clusters, while the uncontaminated observations are the first 60 in the set and have been plotted using the symbol "$*$."

Figure 4(a) shows the projections onto the kurtosis maximization direction. This direction is able to isolate the observations corresponding to one of the clusters of outliers in the data (the one labeled as "#") but not the remaining outliers. The next direction, which maximizes the kurtosis on a subspace orthogonal to the preceding direction, reveals the outliers indicated as "+," as shown in Figure 4(b). This process is repeated until eight additional directions, maximizing the kurtosis onto the corresponding orthogonal subspaces, are generated. The next direction obtained in this way (the third one maximizing the kurtosis) is not able to reveal any outliers, but the fourth, shown in Figure 4(c), allows the identification of the outliers shown as "o." The remaining kurtosis maximization directions

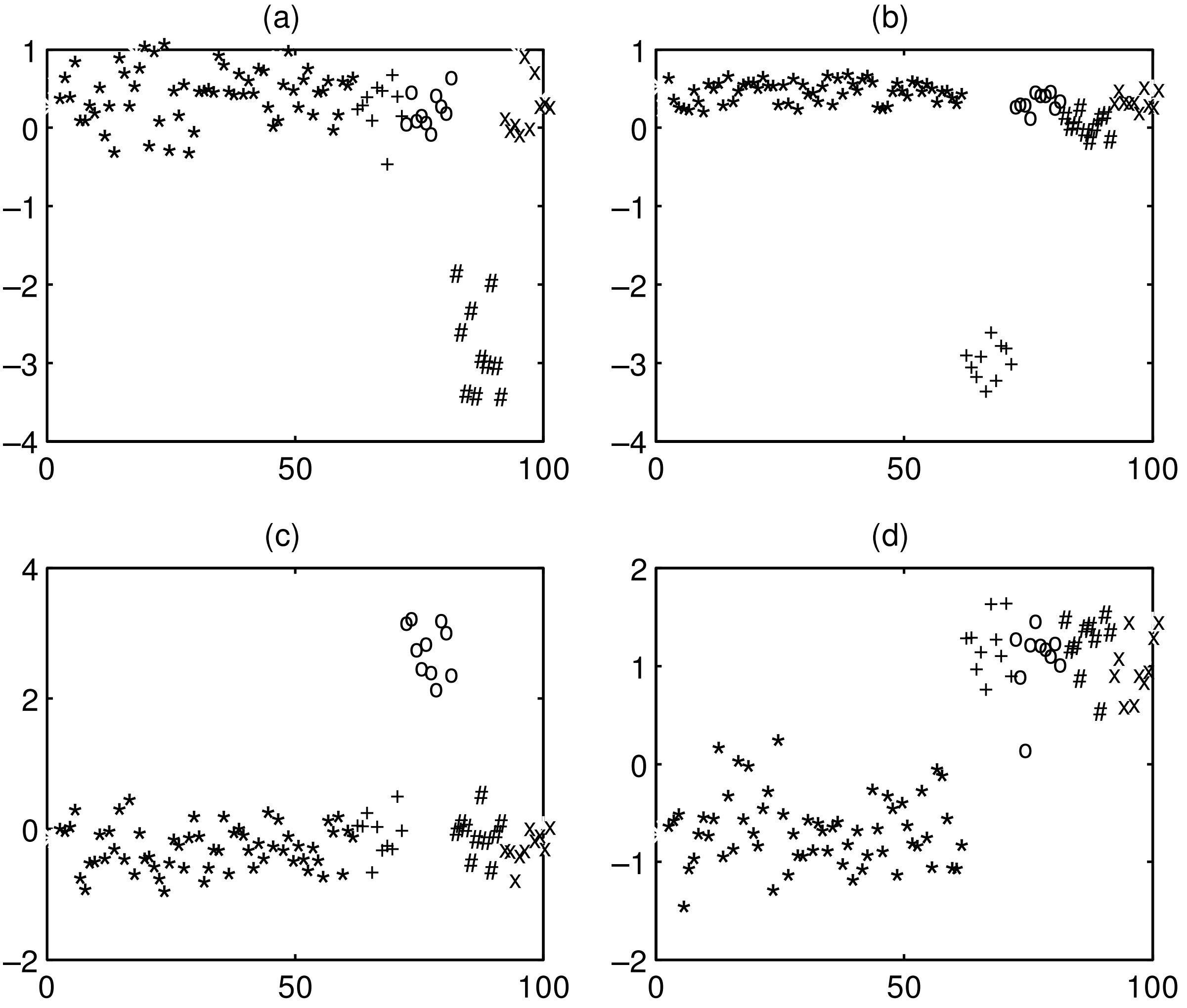

Figure 4. Univariate Projections Onto Directions Generated by the Algorithm for a Dataset in Dimension 10. The x axis represents the observation number, while the y axis corresponds to the projections of each observation onto (a) the 'rst maximization direction, (b) the second maximization direction, (c) the fourth maximization direction, (d) the third minimization direction.

(not shown in the figure) are not able to reveal any additional groups of outliers.

To detect the outliers labeled as "×," the kurtosis minimization directions must be used. The first two of these are again unable to reveal the presence of any outliers. On the other hand, the third minimization direction, shown in Figure 4(d), allows the identification of (nearly) all the outliers at once (it is a direction on the subspace generated by the four directions to the centers of the outlier clusters). The remaining directions are not very useful.

This example illustrates the importance of using both minimization and maximization directions and in each case relying not just on the first optimizer but on computing a full set of orthogonal directions.

3. PROPERTIES OF THE ESTIMATOR

The computation of directions maximizing the kurtosis coefficient is affine equivariant. Note that the standardization of the data in Step 1 of the algorithm ensures that the resulting data are invariant to affine transformations, except for a rotation. The computation of the projection directions preserves this property, and the values of the projections are affine invariant. Note also that the initial point for the optimization algorithm is not affected by affine transformations.

As a consequence of the analysis in Section 1.1, we conclude that the algorithm is expected to work properly if the directions computed are those to the outliers or orthogonal to them since additional orthogonal directions will be computed in later iterations. It might fail if one of the computed directions is the one corresponding to the intermediate extreme direction (whenever it exists). This intermediate direction will correspond either to a maximizer or to a minimizer, depending on the values of α, δ, and λ. Because the projection step does not affect the values of α or λ, if we assume that $\delta \to \infty$, this intermediate direction would be found either as part of the set of p directions maximizing the kurtosis coefficient or as part of the p minimizers, but it cannot appear on both sets. As a consequence, if this intermediate direction appears as a maximizer (minimizer), the set of minimizing (maximizing) directions will include only the directions corresponding to $\omega = 0$ or $\omega = \pm 1$ and, therefore, the true direction to the outliers will always be a member of one of these two sets.

Table 4. Results Obtained by the Proposed Algorithm, Using Both Maximization and Minimization Directions, on Some Small Datasets

Dataset	*Dimension*	*# Observations*	*# Outliers*	*Outliers*	*Time (s.)*
Heart	2	12	5	2, 6, 8, 10, 12	.05
Phosphor	2	18	7	1, 4, 6, 7, 10, 16, 18	.16
Stackloss	3	21	8	1, 2, 3, 4, 13, 14, 20, 21	.27
Salinity	3	28	8	5, 10, 11, 15, 16, 17, 23, 24	.32
Hawkins–Bradu–Kass (1984)	3	75	14	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14	.17
Coleman	5	20	7	1, 6, 9, 10, 11, 13, 18	.22
Wood	5	20	4	4, 6, 8, 19	.22
Bushfire	5	38	15	7, 8, 9, 10, 11, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38	.11

Simulation Results

We have conducted a number of computational experiments to study the practical behavior of the proposed procedure. Since the use of minimization directions for the kurtosis coefficient is not a very intuitive choice, we have implemented two versions of the proposed algorithm—kurtosis1 corresponds to the description given in Section 2; kurtosis2 uses only the set of p maximization directions, while preserving the remaining implementation details in the algorithm.

Our first experiment has analyzed the outlier-detection behavior of the algorithm on a collection of eight small datasets. The first seven were taken from Rousseeuw and Leroy (1987) and were studied by Rousseeuw and Van Driessen (1999), for example. The last one is from Campbell (1989) and was analyzed by Maronna and Yohai (1995) and Becker and Gather (1999), among others. Table 4 gives the corresponding results for algorithm kurtosis1, indicating the dataset, its dimension and number of observations, the number of observations that have been labeled as suspected outliers, the specific observations that have been so labeled, and the running times in seconds. The cutoff points used to label the observations as outliers have been those indicated in the description of the algorithm in Section 2 (Steps 5 and 7 and Table 2). All values are based on a Matlab implementation of the proposed procedure, and the running times have been obtained using Matlab 4.2 on a 450 MHz Pentium PC.

These results are similar to those reported in the literature for other outlier-detection methods, and they indicate that the proposed method behaves reliably on these test sets. These same test problems have been analyzed using kurtosis2. The results are given in Table 5, and for these small problems are nearly identical (except for the "phosphor" dataset) to the ones obtained using both minimization and maximization directions and presented in Table 4.

To explore further the properties of the method, we have performed an extensive set of simulation experiments for larger sample sizes and observation dimensions. The experiments compare the performance of both proposed algorithms, regarding the identification of the outliers and the estimation of covariance matrices, with the results from two other codes:

1. A recent and efficient algorithm for the implementation of the minimum covariance determinant (MCD) procedure proposed by Rousseeuw (1985). The FAST-MCD algorithm based on the splitting of the problem into smaller subproblems, is much faster than previous algorithms; it was proposed by Rousseeuw and Van Driessen (1999).
2. A version of the SD algorithm, corresponding to the implementation described by Maronna and Yohai (1995). The choice of parameters has been the same as in this reference. In particular, the number of subsamples has been chosen as 1,000 for dimension 5. For dimensions 10 and 20, not included in the Monte Carlo study by Maronna and Yohai (1995), we have used 2,000 and 5,000 subsamples, respectively.

For a given contamination level α, we have generated a set of $100(1-\alpha)$ observations from an $\mathrm{N}(0, I)$ distribution in dimension p. We have added 100α additional observations from an $\mathrm{N}(\delta e, \lambda I)$ distribution, where e denotes the vector $(1, \ldots, 1)'$. This model is analogous to the one used by Rousseeuw and van Driessen (1999). This experiment has been conducted for different values of the sample-space dimension p ($p = 5, 10, 20$), the contamination level α ($\alpha = .1, .2, .3, .4$), the distance of the outliers δ ($\delta = 10, 100$), and the standard deviation of these outliers $\sqrt{\lambda}$($\sqrt{\lambda} = .1, 1, 5$). For each set of values, 100 samples have been generated.

Table 5. Results Obtained by the Proposed Algorithm, Using Only Maximization Directions, on Some Small Datasets

Dataset	*# Outliers*	*Outliers*	*Time (s.)*
Heart	5	2, 6, 8, 10, 12	.05
Phosphor	2	1, 6	.05
Stackloss	8	1, 2, 3, 4, 13, 14, 20, 21	.16
Salinity	8	5, 10, 11, 15, 16, 17, 23, 24	.11
Hawkins–Bradu–Kass (1984)	14	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14	.06
Coleman	7	1, 6, 9, 10, 11, 13, 18	.11
Wood	4	4, 6, 8, 19	.11
Bushfire	16	7, 8, 9, 10, 11, 12, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38	.11

Table 6. Success Rates for the Detection of Outliers Forming One Cluster

			$\delta = 10$				$\delta = 100$			
p	α	$\sqrt{\lambda}$	*FAST-MCD*	*SD*	*Kurtosis1*	*Kurtosis2*	*FAST-MCD*	*SD*	*Kurtosis1*	*Kurtosis2*
5	.3	.1	0	100	100	83	100	100	100	88
		1	100	100	95	38	100	100	94	31
	.4	.1	0	0	53	0	0	100	100	0
		1	100	99	91	0	100	100	93	0
		5	100	93	100	100	100	100	100	100
10	.2	.1	0	100	100	100	100	100	100	100
		1	100	100	60	83	100	100	61	84
	.3	.1	0	100	100	0	0	100	100	1
		1	100	100	23	2	100	100	21	0
	.4	.1	0	0	52	0	0	0	100	0
		1	74	0	82	0	67	0	81	0
		5	100	53	100	100	100	73	100	100
20	.1	.1	86	100	100	100	100	100	100	100
		1	100	100	87	88	100	100	84	82
	.2	.1	0	72	100	8	0	100	100	7
		1	98	61	1	2	100	100	0	0
		5	100	67	100	100	100	100	100	100
	.3	.1	0	0	98	0	0	0	100	0
		1	19	0	0	0	20	0	0	0
		5	100	0	100	100	100	0	100	100
	.4	.1	0	0	1	0	0	0	5	0
		1	0	0	9	0	1	0	8	0
		5	100	0	99	95	100	0	99	90

Table 6 gives the number of samples in which all the outliers have been correctly identified for each set of parameter values and both the proposed algorithms (kurtosis1 and kurtosis2) and the FAST-MCD and SD algorithms. To limit the size of the table, we have shown only those cases in which one of the algorithms scored less than 95 successes.

The proposed method (kurtosis1) seems to perform much better than FAST-MCD for concentrated contaminations, while its behavior is worse for those cases in which the shape of the contamination is similar to that of the original data ($\lambda = 1$). From the results in Section 1, this case tends to be one of the most difficult ones for the kurtosis algorithm because the objective function is nearly constant for all directions, and for finite samples it tends to present many local minimizers, particularly along directions that are nearly orthogonal to the outliers. Nevertheless, this behavior, closely associated with the value $\lambda = 1$, disappears as λ moves away from 1. For example, for $p = 10$ and $\alpha = .3$ the number of successes in 100 trials goes up from 23 for $\lambda = 1$ to 61 for $\sqrt{\lambda} = .8$ and 64 for $\lambda = 1.25$. In any case, we have included the values for $\lambda = 1$ to show the worst-case behavior of the algorithm.

Regarding the SD algorithm, the proposed method behaves better for large space dimensions and large contamination levels, showing that it is advantageous to study the data on a small set of reasonably chosen projection directions, particularly in those situations in which a random choice of directions would appear to be inefficient.

The variant of the algorithm that uses only maximization directions (kurtosis2) presents much worse results than kurtosis1 when the contamination level is high and the contamination is concentrated. As the analysis in Section 1 suggested, in those cases the minimization directions are important.

The case analyzed in Table 6 covers a particular contamination model, the one analyzed in Section 1. It is interesting to study the behavior of the algorithm on other possible contamination models, for example when the outliers form several clusters. We have simulated cases with two and four clusters of outliers, constructed to contain the same number of observations ($\lfloor 100\alpha/k \rfloor$), with centers that lie at a distance $\delta = 10\sqrt{p}$ from the origin (the center of the uncontaminated observations) along random uniformly distributed directions. The variability inside each cluster is the same λ for all of them. Table 7 gives the results of these simulations, in the same format as Table 6.

The results are similar to those in Table 6. The proposed method works much better than FAST-MCD for small values of λ and worse for values of λ close to 1. Regarding the SD algorithm, the random choice of directions works better as the number of clusters increases. Nevertheless, note that, as the sample space dimension and the contamination level increase, the preceding results seem to indicate that the SD algorithm may start to become less efficient.

The results in Tables 6 and 7 show that the $2p$ directions obtained as extremes of the kurtosis coefficient of the projections can be computed in a few seconds and perform in most cases better than the thousands of directions randomly generated by the SD estimator, requiring a much larger computational time. Moreover, the minimization directions play a significant role for large concentrated contaminations. These results suggest that the SD estimator can be easily improved while preserving its good theoretical properties by including these $2p$ directions in addition to the other randomly selected directions. From the same results, we also see, that the FAST-

Table 7. Success Rate for the Detection of Outliers Forming Two and Four Clusters

			2 clusters				*4 clusters*			
p	α	$\sqrt{\lambda}$	*FAST-MCD*	*SD*	*Kurtosis1*	*Kurtosis2*	*FAST-MCD*	*SD*	*Kurtosis1*	*Kurtosis2*
5	.4	.1	65	100	16	0	100	100	89	94
		1	100	100	100	81	100	100	100	100
10	.3	.1	18	100	100	78	100	100	100	100
	.4	.1	0	100	51	0	72	100	15	5
		1	83	100	60	0	99	100	97	95
20	.2	.1	15	100	100	100	93	100	100	100
		1	100	100	90	88	100	100	100	100
	.3	.1	0	100	100	0	22	100	100	99
		1	37	98	1	0	99	100	99	98
		5	100	60	100	100	100	100	100	100
	.4	.1	0	60	5	0	0	100	6	0
		1	0	28	3	0	1	99	2	0
		5	100	0	100	93	100	23	98	100

MCD code performs very well in situations in which the kurtosis procedure fails and vice versa. Again, a combination of these two procedures can be very fruitful.

The preceding tables have presented information related to the behavior of the procedures with respect to Type II errors. To complement this information, Type I errors have also been studied. Table 8 shows the average number of observations that are labeled as outliers by both procedures when 100 observations are generated from an $\mathrm{N}(0, I)$ distribution. Each value is based on 100 repetitions.

The kurtosis algorithm is able to limit the size of these errors through a proper choice of the constants β_p in Step 5 of the algorithm. The SD algorithm could also be adjusted in this way, although, in the implementation used, the cutoff for the observations has been chosen as $\sqrt{\chi^2_{p,.95}}$, following the suggestion of Maronna and Yohai (1995).

A second important application of these procedures is the robust estimation of the covariance matrix. The same simulation experiments described previously have been repeated but now measuring the bias in the estimation of the covariance matrix. The chosen measure has been the average of the logarithms of the condition numbers for the robust covariance matrix estimators obtained using the three methods—FAST-MCD, SD, and kurtosis. Given the sample generation process, a value close to 0 would indicate a small bias in this condition number. Tables 9 and 10 show the average values for these estimates for the settings in Tables 6 and 7, respectively. To limit the size of the tables, only two values for the contamination level ($\alpha = .1, .3$) have been considered.

The abnormally large entries in these tables correspond to situations in which the algorithm is not able to identify the outliers properly. An interesting result is that the kurtosis procedure does a very good job regarding this measure of performance in the estimation of the covariance matrix, at least whenever it is able to identify the outliers properly. Note in particular how well it compares to FAST-MCD, a procedure that should perform very well, particularly for small contamination levels or large dimensions. Its performance is even better when compared to SD, showing again the advantages of a nonrandom choice of projection directions.

Table 8. Percentage of Normal Observations Mislabeled as Outliers

Dimension	*FASTMCD*	*SD*	*Kurtosis1*	*Kurtosis2*
5	9.9	8.4	6.9	6.9
10	22.9	.2	9.9	11.2
20	36.2	.0	7.6	7.2

Regarding computational costs, comparisons are not simple to carry out because the FAST-MCD code is a FORTRAN code, while the kurtosis procedure has been written in Matlab. Tables 4 and 5 include running times for some small datasets. Table 11 presents some running times for larger datasets, constructed in the same manner as those included in Tables 6–10. All cases correspond to $\alpha = .2$, $\delta = 10$, and $\sqrt{\lambda} = .1$. The SD code used $15,000$ replications for $p = 30$ and $30,000$ for $p = 40$. All other values have been fixed as indicated for Table 6. The times correspond to the analysis of a single dataset and are based on the average of the running times for 10 random datasets. They have been obtained on a 450 MHz Pentium PC under Windows 98.

These times compare quite well with those of SD and FAST-MCD. Since the version of FAST-MCD we have used is a FORTRAN code, this should imply additional advantages if a FORTRAN implementation of the proposed procedure were developed. A Matlab implementation of the proposed procedures is available at *http://halweb.uc3m.es/fjp/download.html*.

4. CONCLUSIONS

A method to identify outliers in multivariate samples, based on the analysis of univariate projections onto directions that correspond to extremes for the kurtosis coefficient, has been motivated and developed. In particular, a detailed analysis has been conducted on the properties of the kurtosis coefficient in contaminated univariate samples and on the relationship between directions to outliers and extremes for the kurtosis in the multivariate case.

The method is affine equivariant, and it shows a very satisfactory practical performance, especially for large sample

Table 9. Average Logarithm of the Condition Numbers for Covariance Matrix Estimates, Outliers Forming One Cluster

			$\delta = 10$				$\delta = 100$			
p	α	$\sqrt{\lambda}$	*FAST-MCD*	*SD*	*Kurtosis1*	*Kurtosis2*	*FAST-MCD*	*SD*	*Kurtosis1*	*Kurtosis2*
5	.1	.1	.97	1.26	.90	.88	1.01	1.22	.91	.90
		1	.07	1.15	.90	.94	1.05	1.10	.84	.90
		5	1.02	.99	.88	.93	.99	1.06	.90	.90
	.3	.1	7.79	4.09	.97	1.71	.92	4.08	.95	1.87
		1	.91	2.95	1.21	3.48	.88	2.95	1.48	6.96
		5	.89	2.08	1.05	1.05	.93	2.56	1.03	1.07
10	.1	.1	1.87	2.00	1.60	1.63	1.84	1.97	1.56	1.58
		1	1.85	1.76	1.59	1.61	1.84	1.84	1.61	1.64
		5	1.86	1.60	1.53	1.59	1.85	1.71	1.55	1.59
	.3	.1	9.53	5.55	1.57	8.38	14.00	5.59	1.59	12.89
		1	1.63	4.41	5.11	6.21	1.59	4.45	8.85	10.89
		5	1.69	3.32	1.72	1.73	1.68	4.21	1.75	1.75
20	.1	.1	3.87	3.52	2.60	2.53	3.01	3.56	2.48	2.49
		1	2.99	3.54	3.00	2.88	3.10	3.49	3.85	4.05
		5	3.06	3.19	2.45	2.42	3.09	3.51	2.42	2.43
	.3	.1	10.97	7.97	2.45	9.96	15.56	12.57	2.44	14.59
		1	7.32	7.11	7.33	7.32	10.96	11.70	11.94	11.95
		5	2.87	5.26	2.61	2.56	2.77	9.85	2.60	2.60

space dimensions and concentrated contaminations. In this sense, it complements the practical properties of MCD-based methods such as the FAST-MCD procedure. The method also produces good robust estimates for the covariance matrix, with low bias.

The associate editor of this article suggested a generalization of this method based on using the measure of multivariate kurtosis introduced by Arnold (1964) and discussed by Mardia (1970) and selecting $h \leq p$ directions at a time to maximize (or minimize) the h-variate kurtosis. A second set of h directions orthogonal to the first can then be obtained and the procedure can be repeated as in the proposed algorithm. This idea seems very promising for further research on this problem.

There are also practical problems in which the affine equivariance property may not be very relevant. For example, in many engineering problems arbitrary linear combinations of the design variables have no particular meaning. For these cases, and especially in the presence of concentrated contaminations, we have found that adding those directions that maximize the fourth central moment of the data results in a more powerful procedure.

The results presented in this article emphasize the advantages of combining random and specific directions. It can be expected that, if we have a large set of random uniformly distributed outliers in high dimension, a method that computes a very large set of random directions will be more powerful than another one that computes a small number of specific

Table 10. Average Logarithm of the Condition Numbers for Covariance Matrix Estimates, Outliers Forming Two and Four Clusters

			2 clusters				*4 clusters*			
p	α	$\sqrt{\lambda}$	*FAST-MCD*	*SD*	*Kurtosis1*	*Kurtosis2*	*FAST-MCD*	*SD*	*Kurtosis1*	*Kurtosis2*
5	.1	.1	1.01	.95	.93	.91	1.06	.79	.86	.83
		1	1.03	.90	.90	.92	1.04	.77	.89	.91
		5	.97	.84	.87	.90	1.03	.79	.92	.90
	.3	.1	.92	2.42	.94	.98	.92	1.48	.99	1.01
		1	.92	1.95	1.02	1.08	.90	1.37	1.06	1.04
		5	.89	1.56	1.04	1.02	.91	1.13	1.00	1.02
10	.1	.1	1.81	1.54	1.56	1.58	1.84	1.26	1.53	1.56
		1	1.85	1.44	1.57	1.64	1.87	1.23	1.51	1.60
		5	1.90	1.38	1.61	1.64	1.86	1.21	1.56	1.54
	.3	.1	8.00	3.40	1.56	2.41	1.68	2.27	1.57	1.59
		1	1.70	2.87	1.83	1.78	1.66	2.13	1.72	1.75
		5	1.67	2.51	1.75	1.71	1.69	1.89	1.73	1.76
20	.1	.1	3.04	2.74	2.51	2.45	3.11	2.14	2.42	2.38
		1	3.08	2.79	2.42	2.41	3.15	2.16	2.39	2.32
		5	3.08	2.61	2.40	2.39	3.16	2.07	2.37	2.24
	.3	.1	10.91	5.05	2.47	9.10	6.50	3.53	2.50	2.50
		1	5.87	4.93	6.96	6.98	2.76	3.82	2.63	2.70
		5	2.76	4.40	2.55	2.61	2.81	3.58	2.63	2.61

Table 11. Running Times (in s.) on Large Synthetic Datasets

p	*n*	*FAST-MCD*	*SD*	*Kurtosis1*	*Kurtosis2*
10	100	5.5	4.0	1.2	.6
	200	9.8	8.0	2.6	1.8
20	100	20.6	11.7	3.3	1.5
	200	36.0	22.1	7.9	4.0
30	300	114.8	109.6	28.0	18.9
	500	183.6	182.8	54.1	46.6
40	400	270.5	338.9	74.1	38.4

directions. On the other hand, when the outliers appear along specific directions, a method that searches for these directions is expected to be very useful. These results emphasize the advantages of combining random and specific directions in the search for multivariate outliers. In particular, the incorporation of the kurtosis directions in the standard SD procedure can improve it in many cases with small additional computational time.

ACKNOWLEDGMENTS

We thank Peter Rousseeuw and Katrien van Driessen for making their code FAST-MCD available to us. We also thank the referees, the associate editor, and *Technometrics* Editor Karen Kafadar for their suggestions and comments. They have been very helpful in improving the content and presentation of the article.

APPENDIX: DETAILS OF THE DERIVATION OF THEORETICAL RESULTS IN SECTION 1.

A.1 An Expression for γ_X

To derive (2), we need expressions for $m_X(4)$ and $m_X(2)$ in terms of the moments of the distributions F and G. Note that, since $\mu_F = 0$, we have that $E(X) = \alpha\mu_G$. Moreover

$$\begin{aligned} E(X^2) &= (1-\alpha)m_F(2) + \alpha(m_G(2)+\mu_G^2) \\ &= m_F(2)(1-\alpha+\alpha v^2+\alpha r^2) \end{aligned}$$

and

$$m_X(2) = m_F(2)(1+\alpha(v^2-1)+\alpha(1-\alpha)r^2).$$

For the fourth moment,

$$\begin{aligned} m_X(4) = (1-\alpha)\int(x-\alpha\mu_G)^4 dF(x) \\ +\alpha\int(x-\alpha\mu_G)^4 dG(x), \end{aligned}$$

where

$$\begin{aligned} &\int(x-\alpha\mu_G)^4 dF(x) \\ &\quad = m_F(4) - 4\alpha\mu_G m_F(3) + 6\alpha^2\mu_G^2 m_F(2) + \alpha^4\mu_G^4, \end{aligned}$$

and

$$\begin{aligned} &\int(x-\alpha\mu_G)^4 dG(x) \\ &\quad = m_G(4) + 4(1-\alpha)\mu_G m_G(3) + 6(1-\alpha)^2\mu_G^2 m_G(2) \\ &\qquad + (1-\alpha)^4\mu_G^4. \end{aligned}$$

Combining these results and rearranging terms, we have

$$\begin{aligned} &m_X(4)/m_F(2)^2 \\ &\quad = \gamma_F + \alpha(1-\alpha)\big(4r(a_G v^3 - a_F) + (\gamma_G v^4 - \gamma_F)/(1-\alpha) \\ &\qquad + 6r^2(\alpha+(1-\alpha)v^2) + r^4(\alpha^3+(1-\alpha)^3)\big). \end{aligned}$$

The desired result follows from $\gamma_X = m_X(4)/m_X(2)^2$ and these expressions.

A.2 Parameters in the Distribution of Y

The mean of a random variable X following a distribution of the form $(1-\alpha)\mathrm{N}(0, I) + \alpha\mathrm{N}(\delta e_1, \lambda I)$ is $\mu_X = \alpha\delta e_1$ and its covariance matrix is

$$\begin{aligned} \overline{S} &= (1-\alpha)I + \alpha(\lambda I + \delta^2 e_1 e_1') - \alpha^2\delta^2 e_1 e_1' \\ &= (1-\alpha+\alpha\lambda)I + \alpha(1-\alpha)\delta^2 e_1 e_1' \\ &= \nu_1\left(I + \frac{\delta^2\alpha(1-\alpha)}{\nu_1} e_1 e_1'\right). \end{aligned}$$

The inverse of $\overline{S}$ will also be a rank-one modification of the identity. It is easy to check that

$$S \equiv \overline{S}^{-1} = \frac{1}{\nu_1}(I - \nu_2 e_1 e_1). \tag{A.1}$$

Note that S is diagonal with all entries equal to $1/\nu_1$ except for the first one. Its square root, $S^{1/2}$, is also a diagonal matrix with all entries equal to $1/\sqrt{\nu_1}$ except for the first one, which equals $\sqrt{(1-\nu_2)/\nu_1}$. In particular,

$$S^{1/2} e_1 = \sqrt{\frac{1-\nu_2}{\nu_1}} e_1 = \frac{1}{\sqrt{\nu_1 + \delta^2\alpha(1-\alpha)}} e_1. \tag{A.2}$$

The distribution of $Y = S^{1/2}(X - \mu_X)$ follows from these results.

A.3 An Expression for γ_Z

The kurtosis coefficient of Z will be equal to its fourth moment. $E(Z^4) = (1-\alpha)E(Z_1^4) + \alpha E(Z_2^4)$, where Z_1 is $\mathrm{N}(m_1'u, u'Su)$, Z_2 is $\mathrm{N}(m_2'u, \lambda u'Su)$, and

$$\begin{aligned} E(Z_i^4) &= E((Z_i - \bar{z}_i)^4) + 6E((Z_i - \bar{z}_i)^2)\bar{z}_i^2 + \bar{z}_i^4 \\ &= 3\sigma_i^4 + 6\sigma_i^2\bar{z}_i^2 + \bar{z}_i^4, \end{aligned} \tag{A.3}$$

where $\bar{z}_i$ and σ_i denote the mean and standard deviation of Z_i. Letting $\omega = e_1'u$, from (A.2) and (3) it follows that

$$\begin{aligned} \bar{z}_1 &= m_1'u = -\frac{\alpha\delta\omega}{\sqrt{\nu_1 + \delta^2\alpha(1-\alpha)}} \\ \bar{z}_2 &= m_2'u = \frac{(1-\alpha)\delta\omega}{\sqrt{\nu_1 + \delta^2\alpha(1-\alpha)}}, \end{aligned}$$

and from (A.1)

$$\sigma_1^2 = \sigma_2^2/\lambda = u'Su = \frac{1-\nu_2\omega^2}{\nu_1}.$$

Replacing these values in (A.3), we have

$$\begin{aligned} \gamma_Z = 3&\frac{(1-\nu_2\omega^2)^2}{\nu_1^2}(1-\alpha+\alpha\lambda^2) \\ &+ 6\frac{1-\nu_2\omega^2}{\nu_1}\frac{\delta^2\alpha(1-\alpha)\omega^2}{\nu_1+\delta^2\alpha(1-\alpha)}(\alpha+\lambda(1-\alpha)) \\ &+ \frac{\delta^4\alpha^2(1-\alpha)^2\omega^4}{(\nu_1+\delta^2\alpha(1-\alpha))^2}\left(\frac{\alpha^3+(1-\alpha)^3}{\alpha(1-\alpha)}\right). \end{aligned}$$

Grouping all the terms that correspond to the same powers of ω and using $\nu_1(\alpha+\lambda(1-\alpha))-(1-\alpha+\alpha\lambda^2)=(1-\lambda)(\alpha^2\lambda-(1-\alpha)^2)$, the result in (4) is obtained.

[Received March 1999. Revised June 2000.]

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.

Arnold, H. J. (1964), "Permutation Support for Multivariate Techniques," *Biometrika*, 51, 65–70.

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

Balanda, K. P., and MacGillivray, H. L. (1988), "Kurtosis: A Critical Review," *The American Statistician*, 42, 111–119.

Becker, C., and Gather, U. (1999), "The Masking Breakdown Point of Multivariate Outlier Identification Rules," *Journal of the American Statistical Association*, 94, 947–955.

Box, G. E. P., and Tiao, G. C. (1968), "A Bayesian Approach to Some Outlier Problems," *Biometrika*, 55, 119–129.

Campbell, N. A. (1980), "Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation," *Applied Statistics*, 29, 231–237.

——— (1989), "Bushfire Mapping Using NOAA AVHRR Data," technical report, CSIRO, North Ryde, Australia.

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," unpublished 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*, Ser. B, 54, 761–771.

——— (1994), "A Modification of a Method for the Detection of Outliers in Multivariate Samples," *Journal of the Royal Statistical Society*, Ser. B, 56, 393–396.

Hampel, F. R. (1985), "The Breakdown Point of the Mean Combined With Some Rejection Rules," *Technometrics*, 27, 95–107.

Hampel, F. R., Ronchetti, E. M., 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.

Hawkins, D. M., Bradu, D., and Kass, G. V. (1984), "Location of Several Outliers in Multiple Regression Data Using Elemental Sets," *Technometrics*, 26, 197–208.

Hawkins, D. M., and Olive, D. J. (1999), "Improved Feasible Solution Algorithms for High Breakdown Estimation," *Computational Statistics and Data Analysis*, 30, 1–11.

Jones, M. C., and Sibson, R. (1987), "What Is Projection Pursuit?" *Journal of the Royal Statistical Society*, Ser. A, 150, 29–30.

Juan, J., and Prieto, F. J. (1997), "Identification of Point-Mass Contaminations in Multivariate Samples," Working Paper 97–13, Statistics and Econometrics Series, Universidad Carlos III de Madrid.

Malkovich, J. F., and Afifi, A. A. (1973), "On Tests for Multivariate Normality," *Journal of the American Statistical Association*, 68, 176–179.

Mardia, K. V. (1970), "Measures of Multivariate Skewness and Kurtosis with Applications," *Biometrika*, 57, 519–530.

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.

Posse, C. (1995), "Tools for Two-Dimensional Exploratory Projection Pursuit," *Journal of Computational and Graphical Statistics*, 4, 83–100.

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.

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

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

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.

Ruppert, D. (1987), "What Is Kurtosis," *The American Statistician*, 41, 1–5.

Stahel, W. A. (1981), "Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen," unpublished Ph.D. thesis, Eidgenössische Technische Hochschule, 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, pp. 327–336.

Wilks, S. S. (1963), "Multivariate Statistical Outliers," *Sankhya*, Ser. A, 25, 407–426.

Discussion

David M. Rocke

Department of Applied Science
University of California
Davis, CA 95616
(*dmrocke@ucdavis.edu*)

David L. Woodruff

Graduate School of Management
University of California
Davis, CA 95616
(*dlwoodruff@ucdavis.edu*)

Peña and Prieto present a new method for robust multivariate estimation of location and shape and identification of multivariate outliers. These problems are intimately connected in that identifying the outliers correctly automatically allows excellent robust estimation results and vice versa.

Some types of outliers are easy to find and some are difficult. In general, the previous literature concludes that problems are more difficult when the fraction of outliers is large. Moreover, widely scattered outliers are relatively easy, whereas

TECHNOMETRICS, AUGUST 2001, VOL. 43, NO. 3

concentrated outliers can be very difficult (Rocke and Woodruff1996). This article is aimed squarely at the case in which the outliers form a single cluster, separated from the main data and of the same shape (but possibly different size) as the main data, an especially difficult case for many outlier-detection methods (Rocke and Woodruff 1996). We believe that it is entirely appropriate that special methods like those in the present article be developed to handle this case, which is often too difficult for general-purpose outlier-detection and robust-estimation methods.

In this discussion, we make some comparisons of the estimators kurtosis1, FAST-MCD, SD, and also certain M and S estimators; then we point out the connection to cluster analysis (Rocke and Woodruff 2001). The MCD, MVE, SD, and S estimators with hard redescending influence functions are known to have maximum breakdown. Kurtosis1 probably also has maximum breakdown, although this article has no formal proof. M estimators are sometimes thought to be of breakdown $1/(p+1)$, but this is actually incorrect. Work by Maronna (1976), Donoho (1982), and Stahel (1981) showed that whenever the amount of contamination exceeds $1/(p+1)$ a root of the estimating equations that can be carried over all bounds exists. In fact, a root may exist that remains bounded. S estimators are a subclass of M estimators, as shown by Lopuhaä (1989), and have maximal breakdown when hard redescending ψ functions are used and the parameters are correctly chosen. This provides an example of a class of M estimators that have maximal breakdown.

M estimators can be highly statistically efficient and are easy to compute by iteratively reweighted least squares but need a high-breakdown initial estimator to avoid converging on the "bad" root (this also applies to S estimators, which are computed by solving the related constrained M estimation problem). MULTOUT (Rocke and Woodruff 1996) combines a robust initial estimator with an M estimator to yield a high-breakdown, statistically efficient methodology. Note also that identifying outliers by distance and reweighting points with weight 0 if the point is declared an outlier and weight 1 otherwise is a type of M estimator with a ψ function that is constant until the outlier rejection cutoff point. This is used by many methods (e.g., FAST-MCD, kurtosis1) as a final step, which improves statistical efficiency.

Thus, the key step in outlier identification and robust estimation is to use an initial procedure that gives a sufficiently good starting place. The gap can be quite large between the theoretical breakdown of an estimator and its "practical breakdown," the amount and type of contamination such that success is unlikely. Consider, for example, the case in Table 6 in which $n = 100$, $p = 20$, $\alpha = .3$, $\lambda = 1$, and $\delta = 100$. The amount of contamination is well below the breakdown point (40%), and the contamination is at a great distance from the main data, but none of the methods used in this study are very successful.

The study reported in Table 6 has some difficulties:

1. The number of data points is held fixed at 100 as the dimension rises, which leads to a very sparse data problem in dimension 20. Many of these methods could perhaps do much better with an adequate amount of data. This is particularly true of FAST-MCD and MULTOUT, which are designed to handle large amounts of data.

2. When $n = 100$ and $p = 20$, the maximum breakdown of any equivariant estimator is .4. Of course, this does not mean that the estimator must break down whatever the configuration of the outliers, but it does show that at $\alpha = .4$ and $p = 20$ and for every dataset generated there is an outlier configuration such that no equivariant estimator will work.

3. The case $\lambda = 5$ is one in which the outliers are actually scattered rather than concentrated. This is especially true when $\delta = 10$. With standard normal data in 20 dimensions, almost all observations lie at a distance less than the .001 point of a χ_{20} distribution, which is 6.73. The outliers are placed at a distance of 10, and the .999 sphere for these has a radius of $(5)(6.73) = 33.65$. Thus, the outliers actually surround the good data, rather than lying on one side. This explains why FAST-MCD gets all of the $\lambda = 5$ cases regardless of other parameter settings.

4. When $n = 100$ and $p = 20$, the preceding computation shows that standard normal data will almost all lie in a sphere of radius 6.73 around the center. If the outliers are displaced by 10 with the same covariance ($\lambda = 1$), these spheres overlap considerably, and in some instances no method can identify all "outliers" because some of them lie within the good data. Shrunken outliers ($\lambda = .1$) do not have this problem since the radius of the outliers is only .67, so a displacement of 10 prevents overlap. Expanded outliers ($\lambda = 5$) are relatively easy because the outliers surround the main data, and the chance of one falling in the relatively small sphere of the good data is small.

The results are better for two and four clusters (Table 8) than for one and would be even better for individually scattered outliers. The important insight is that many methods work well except in one hard case—when the outliers lie in one or two clusters. Methods such as the FAST-MCD and MULTOUT must be supplemented by methods specifically designed to find clustered or concentrated outliers. These clustered outlier methods are not replacements for the other methods because they may perform poorly when there are significant nonclustered outliers.

The remaining question concerns appropriate methods to supplement the more general estimation techniques and allow detection of clustered/concentrated outliers. The methods of Peña and Prieto are aimed at that task, by trying to find the direction in which the clustered outliers lie. The performance of these methods is documented in the article under discussion. Another attempt in this direction was given by Juan and Prieto (2001), who tried to find outliers by looking at the angles subtended by clusters of points projected on an ellipsoid. They showed reasonable performance of this method but provided computations only for very concentrated outliers ($\lambda = .01$).

Rocke and Woodruff (2001) presented another approach. If the most problematic case for methods like FAST-MCD and MULTOUT is when the outliers form clusters, why not apply methods of cluster analysis to identify them? There are many important aspects to this problem that cannot be treated in the limited space here, but we can look at a simple version of our procedure. We search for clusters using a model-based clustering framework and heuristic search optimization methods, then apply an M estimator to the largest identified cluster (for details, see Coleman and Woodruff 2000; Coleman et al. 1999; Reiners 1998; Rocke and Woodruff 2001).

To illustrate our point, we ran a slightly altered version of the simulation study in Table 6 from Peña and Prieto. First, we placed the outliers on a diagonal rather than on a coordinate axis. If $\boldsymbol{u} = (1, 1, \ldots, 1)$, then the mean of the outliers was taken to be $p^{-1/2}\delta\boldsymbol{u}$, which lies at a distance of δ from $\mathbf{0}$. After generating the data matrix $\boldsymbol{X}$ otherwise in the same way as did Peña and Prieto, we centered the data to form a new matrix $\boldsymbol{X}^*$ and then generated "sphered" data in the following way: Let $\boldsymbol{X}^* = \boldsymbol{UDV}^\top$, the singular value decomposition (SVD) in which $\boldsymbol{D}$ is the diagonal matrix of singular values. Then the matrix $\widetilde{\boldsymbol{X}} = \boldsymbol{X}^*\boldsymbol{VD}^{-1}\boldsymbol{V}^\top = \boldsymbol{UV}^\top$ is an affine transformed version of the data with observed mean vector $\mathbf{0}$ and observed covariance matrix $\boldsymbol{I}$. The purpose of these manipulations is to ensure that methods that may use nonaffine-equivariant methods do not have an undue advantage. Pushing the outliers on the diagonal avoids giving an advantage to componentwise methods. The SVD method of sphering, unlike the usual Cholesky factor approach, preserves the direction of displacement of the outliers.

We ran all the cases from the simulation of Peña and Prieto. Time constraints prevented us from running the 100 repetitions, but out of the five repetitions we did run, we succeeded in 100% of the cases in identifying all of the outliers, except for the cases in which $p = 20$ and $\lambda = 1$; these required more data for our methods to work. We could identify all of the outliers in 100% of the cases when $n = 500$, for example, instead of $n = 100$. The successful trials included all of the other cases in which no other method reported in Table 6 was very successful.

This insight transforms part of the outlier-identification problem into a cluster-analysis problem. However, the latter is not necessarily easy (Rocke and Woodruff 2001). For example, we tried the same set of simulation trials using the standard clustering methods available in S-PLUS. In this case, we ran 10 trials of each method. The methods used were two hierarchical agglomeration methods, mclust (Banfield and Raftery 1993) and agnes (Kaufman and Rousseeuw 1990; Struyf, Hubert, and Rousseeuw 1997); diana, a divisive hierarchical clustering method; fanny, a fuzzy clustering method; pam, clustering around medoids (Kaufman and Rousseeuw 1990; Struyf et al. 1997); and k-means (Hartigan 1975). First, it should be noted that all of these methods succeed almost all of the time for separated clusters ($\lambda = .1$ or $\lambda = 1$) if the data are not sphered. All of these methods except mclust use Euclidean distance and make no attempt at affine equivariance. Mclust uses an affine equivariant objective function based on a mixture model, but the initialization steps that are crucial to its performance are not equivariant.

Over the 72 cases considered (with sphered data), none of these methods achieved as much as 50% success. The overall success rates were agnes 36%, diana 48%, fanny 41%, k-means 10%, mclust 37%, and pam 12% (compared to MCD 75%, SD 77%, kurtosis1 83%, and our clustering method 89%). For shrunken outliers ($\lambda = .1$), the most successful were fanny 81% and mclust 37%, and the least successful were agnes 1% and k-means 2% (compared to MCD 41%, SD 70%, kurtosis1 88%, and our clustering method 100%). For expanded outliers ($\lambda = 5$), the most successful were agnes 97%, diana 96%, mclust 73%, and fanny 39%, and the least successful was pam 0% (compared to MCD 100%, SD 88%, kurtosis1 100%, and our clustering method 100%). This case is quite easy for robust multivariate estimation methods; FAST-MCD and MULTOUT each get 100% of these. Again, the most difficult case is shift outliers (Rocke and Woodruff 1996; Hawkins 1980, p. 104), in which $\lambda = 1$. The best performance among these clustering methods is diana 34%, with the next best being k-means at 16% (compared to MCD 82%, SD 73%, kurtosis1 62%, and our clustering method 67%).

The poor performance of these clustering methods is probably due to two factors. First, use of the Euclidean metric is devastating when the shape of the clusters is very nonspherical. Since this can certainly occur in practice, such methods should at least not be the only method of attack. Second, some of these problems require more extensive computation than these methods allow, at least in the implementation in S-PLUS. Performance of many descents of the algorithm from a wide variety of starting points, including random ones, can obtain better solutions than a single descent from a single plausible starting point. This is particularly true when the data are extremely sparse, as they are in these simulations. Many of these methods could, of course, be modified to use multiple starting points and might show enormously improved performance.

We confirmed this by using two additional model-based clustering methods that permit control of the amount of computation. The first of these was EMMIX (McLachlan and Basford 1988; McLachlan and Peel 2000). The second was EMSamp (Rocke and Dai 2001). For both programs, the performance was similar to that of our clustering method; success was usual if enough random starting points were used except for shift outliers in dimension 20 in the present $n = 100$ case.

Since different methods are better at different types of outliers and since once the outliers have been identified by any method they can be said to stay identified, an excellent strategy when using real data instead of simulated data (where the structure is known) is to use more than one method. Among the methods tested by Peña and Prieto, the best combination is FAST-MCD plus kurtosis1. Together these can improve the overall rate of identification from 75% and 83%, respectively, to 90%, since FAST-MCD is better for shift outliers and kurtosis1 is better for shrunken outliers. An even better combination is our clustering method plus FAST-MCD, which gets an estimated 95% of the cases.

We can summarize the most important points we have tried to make here as follows:

1. General-purpose robust-estimation and outlier-detection methods work well in the presence of scattered outliers or multiple clusters of outliers.

2. To deal with one or two clusters of outliers in difficult cases, methods specifically meant for this purpose such as those of Peña and Prieto and Juan and Prieto (2001) are needed to supplement the more general methods.

3. An alternative approach for this case is to use clustering methods to supplement the general-purpose robust methods.

4. Choice of clustering method and computational implementation are important determinants of success. We have found that model-based clustering together with heuristic search technology can provide high-quality methods (Coleman and Woodruff 2000; Coleman et al. 1999; Reiners 1998; Rocke and Dai 2001; Rocke and Woodruff 2001).

5. The greatest chance of success comes from use of multiple methods, at least one of which is a general-purpose method such as FAST-MCD and MULTOUT, and at least one of which is meant for clustered outliers, such as kurtosis1, the angle method of Juan and Prieto (2001), or our clustering method (Rocke and Woodruff 2001).

REFERENCES

Banfield, J. D., and Raftery, A. E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 49, 803–821.

Coleman, D. A., Dong, X., Hardin, J., Rocke, D. M., and Woodruff, D. L. (1999), "Some Computational Issues in Cluster Analysis With no a Priori Metric," *Computational Statistics and Data Analysis*, 31, 1–11.

Coleman, D. A., and Woodruff, D. L. (2000), "Cluster Analysis for Large Datasets: Efficient Algorithms for Maximizing the Mixture Likelihood," *Journal of Computational and Graphical Statistics*, 9, 672–688.

Hartigan, J. A. (1975), *Clustering Algorithms*, New York: Wiley.

Hawkins, D. M. (1980), *The Identification of Outliers*, London: Chapman and Hall.

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

Kaufman, L., and Rousseeuw, P. J. (1990), *Finding Groups in Data*, New York: Wiley.

Lopuhaä, H. P. (1989), "On the Relation Between S-Estimators and M-Estimators of Multivariate Location and Covariance," *The Annals of Statistics*, 17, 1662–1683.

McLachlan, G. J., and Basford, K. E. (1988), *Mixture Models: Inference and Application to Clustering*, New York: Marcel Dekker.

McLachlan, G. J., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

Reiners, T. (1998), "Maximum Likelihood Clustering of Data Sets Using a Multilevel, Parallel Heuristic," unpublished Master's thesis, Technische Universität Braunschweig, Germany, Dept. of Economics, Business Computer Science and Information Management.

Rocke, D. M., and Dai, J. (2001), "Sampling and Subsampling for Cluster Analysis in Data Mining With Application to Sky Survey Data," unpublished manuscript.

Rocke, D. M., and Woodruff, D. L. (2001), "A Synthesis of Outlier Detection and Cluster Identification," unpublished manuscript.

Struyf, A., Hubert, M., and Rousseeuw, P. J. (1997), "Integrating Robust Clustering Techniques in S-Plus," *Computational Statistics and Data Analysis*, 26, 17–37.

Discussion

Mia Hubert

Department of Mathematics and Computer Science
University of Antwerp
Antwerp
Belgium
(*mia.hubert@ua.ac.be*)

Peña and Prieto propose a new algorithm to detect multivariate outliers. As a byproduct, the population scatter matrix is estimated by the classical empirical covariance matrix of the remaining data points.

The interest in outlier-detection procedures is growing fast since data mining has become a standard analysis tool in both industry and research. Once information ("data") is gathered, marketing people or researchers are not only interested in the behavior of the regular clients or measurements (the "good data points") but they also want to learn about the anomalous observations (the "outliers").

As pointed out by the authors, many different procedures have already been proposed over the last decades. However, none of them has a superior performance at all kinds of contamination patterns. The high-breakdown MCD covariance estimator of Rousseeuw (1984) is probably the most well known and most respected procedure. There are several reasons for this. First, the MCD has good statistical properties since it is affine equivariant and asymptotically normally distributed. It is also a highly robust estimator, achieving a breakdown value of 50% and a bounded influence function at any elliptical distribution (Croux and Haesbroeck 1999). Another advantage is the availability of a fast and efficient algorithm, called FAST-MCD (Rousseeuw and Van Driessen 1999), which is currently incorporated in S-PLUS and SAS. Therefore, the MCD can be used to robustify many multivariate techniques such as discriminant analysis (Hawkins and McLachlan 1997), principal-component analysis (PCA) (Croux and Haesbroeck 2000), and factor analysis (Pison, Rousseeuw, Filzmoser, and Croux 2000).

Whereas the primary goal of the MCD is to robustly estimate the multivariate location and scatter matrix, it also can be used to detect outliers by looking at the (squared) robust distances $\mathrm{RD}(x_i) = (x_i - T_{\mathrm{MCD}})^t S_{\mathrm{MCD}}^{-1}(x_i - T_{\mathrm{MCD}})$. Here, T_{MCD} and S_{MCD} stand for the MCD location and scatter estimates. One can compare these robust distances with the quantiles of the χ_p^2 distribution. Since this rejection rule often leads to an inflated Type II error, Hardin and Rocke (2000) developed more precise cutoff values to improve the MCD outlier-detection method.

If one is mainly interested in finding the outliers, it is less important to estimate the shape of the good points with great accuracy. As the authors explain, it seems natural to try to find (for each data point) the direction along which the relative distance to the uncontaminated observations is largest. This was the idea behind the Stahel–Donoho estimator. Since the center of the good data is unknown, this estimator projects the dataset on all possible directions and then considers a robust univariate distance [Formula (10)] for each data point. The maximal distance of the point over all directions is then defined as its outlyingness.

This projection pursuit (PP) strategy has the advantage of being applicable even when $p > n$. This situation often occurs in chemometrics, genetics, and engineering. It is therefore not surprising that many multivariate methods in high dimensions have been robustified using the PP principle (e.g., see Li and Chen 1985; Croux and Ruiz-Gazen 1996, 2000; Hubert, Rousseeuw, and Verboven 2000). The computation time of these estimators would be huge if all possible directions were to be scanned. To reduce the search space, Stahel (1981) proposed taking the directions orthogonal to random p subsets. In PCA, where only orthogonal equivariance is required, Croux and Ruiz-Gazen (1996) and Hubert et al. (2000) looked at the n directions from the spatial median of the data to each data point. However, it is clear that the precision of these algorithms would improve if one were able to find the most interesting directions. This is exactly what the authors try to do in this article, which I have therefore read with great interest.

However, several key aspects of their proposal can be criticized.

1. *The standardization in Step 1.* First a standardization is performed [Formula (7)] using the classical mean and covariance matrix. It is well known that these estimators are extremely sensitive to outliers, which often leads to masking large Type II errors by labeling outliers as good data points and swamping large Type I errors by flagging good points as outliers. Rousseeuw and Leroy (1987, pp. 271–273) gave an example of how classical prestandardization can destroy the robustness of the final results.

2. *Maximizing and minimizing the kurtosis in Steps 2 and 3.* The theoretical discussion in Section 1.1 indicates that the kurtosis is maximal (respectively, minimal) in the direction of the outliers when the contamination is concentrated and small (respectively, large). However, this is not always true for an intermediate level of contamination. To exemplify this, I have generated 100 univariate data points according to the asymmetric contamination model of Section 1.1: $(1-\alpha)\mathrm{N}(0,1)+\alpha\mathrm{N}(\delta,\lambda)$. Figure 1, (a) and (b), confirms the authors' conclusions. For $\alpha = 5\%$, the kurtosis increases rapidly, whereas at $\alpha = 40\%$ the kurtosis decreases to 1.17. Figure 1(c) shows the kurtosis for $\alpha = 20\%$, $\lambda = .2$, and δ ranging from 1 to 20. The kurtosis first attains a minimum of 2.45 at $\delta = 3$ and then increases. But it stays close to 3, the kurtosis of the standard normal distribution. To construct a confidence interval for the kurtosis at a normal sample of size 100, I generated 10,000 samples of size 100 and computed the .025 and .975 quantiles of their kurtosis, resulting in the 95% confidence interval [2.27, 4.05]. We see that none of the contaminated samples had a kurtosis outside this confidence interval.

Moreover, Section 1.1 deals with only concentrated outliers. What if the outliers are not concentrated in one place? As an example, I generated symmetric contamination, according to the model

$$(1-\alpha)\mathrm{N}(0,1)+\frac{\alpha}{2}\mathrm{N}(\delta,\lambda)+\frac{\alpha}{2}\mathrm{N}(-\delta,\lambda).$$

Figure 1(d) shows the kurtosis as a function of δ for $\alpha = 20\%$ and $\lambda = .2$. Here again it can be seen that the kurtosis does not become extremely large or small, except for δ close to 2.

3. *Taking $2p$ orthogonal directions in Steps 2 and 3.* In Section 1.1, only a very particular kind of contamination is studied. For this pattern, it indeed seems reasonable to consider orthogonal directions, but why should this also work at other contamination patterns?

Moreover, I do not understand why first only p directions that maximize the kurtosis are considered and then p directions that minimize the kurtosis. Why is this better than for example, alternating between a direction that maximizes and one that minimizes the kurtosis? The actual implementation gives me the impression that, although inspired by the behavior of the kurtosis, the chosen directions are still rather arbitrary.

To verify this, I conducted the following experiment. Using the author's Matlab code, I changed their orthogonal directions obtained in Steps 2 and 3 into two random sets of p orthogonal directions. (A random set of p orthogonal directions was obtained by taking the principal components of 100 randomly generated p-variate standard normal observations.) Let us call this new estimator rand2p. Then I repeated some of the simulations that are described in Section 3.1 and that are reported in Table 6. My Table 1 lists the results obtained with the authors' proposed method kurtosis1, as well as with rand2p and FAST-MCD. For each set of parameters, the first line indicates the success rate of the algorithms, which (following the authors' convention) is defined as the number of samples in which all the outliers have been correctly identified. (In my opinion, it would be better to count *how many* points were flagged as outliers.) The second line shows the results obtained by the authors.

The third line lists the average number of iterations needed in Step 6 of the kurtosis1 and the rand2p algorithm. Although the authors claim that they consider projections in only $2p$ directions, this is somewhat exaggerated, since in Step 6 of the algorithm it is said that whenever some outliers are found the procedure is applied again to the reduced sample, so typically several sets of $2p$ directions are considered. Table 1 thus indicates how many times $2p$ directions were constructed on average.

Table 1 shows that the rand2p algorithm based on random directions unrelated to the data has the same or even better performance than the authors' kurtosis1. The number of iterations that occur in rand2p is larger than in kurtosis1, but since the former algorithm does not need any optimization step, its computation time is still lower than that of kurtosis1. This shows that selecting directions based on the kurtosis does not add value. On the other hand, note that, in spite of the iterations, the total number of directions considered is still very low compared with the large number of random directions

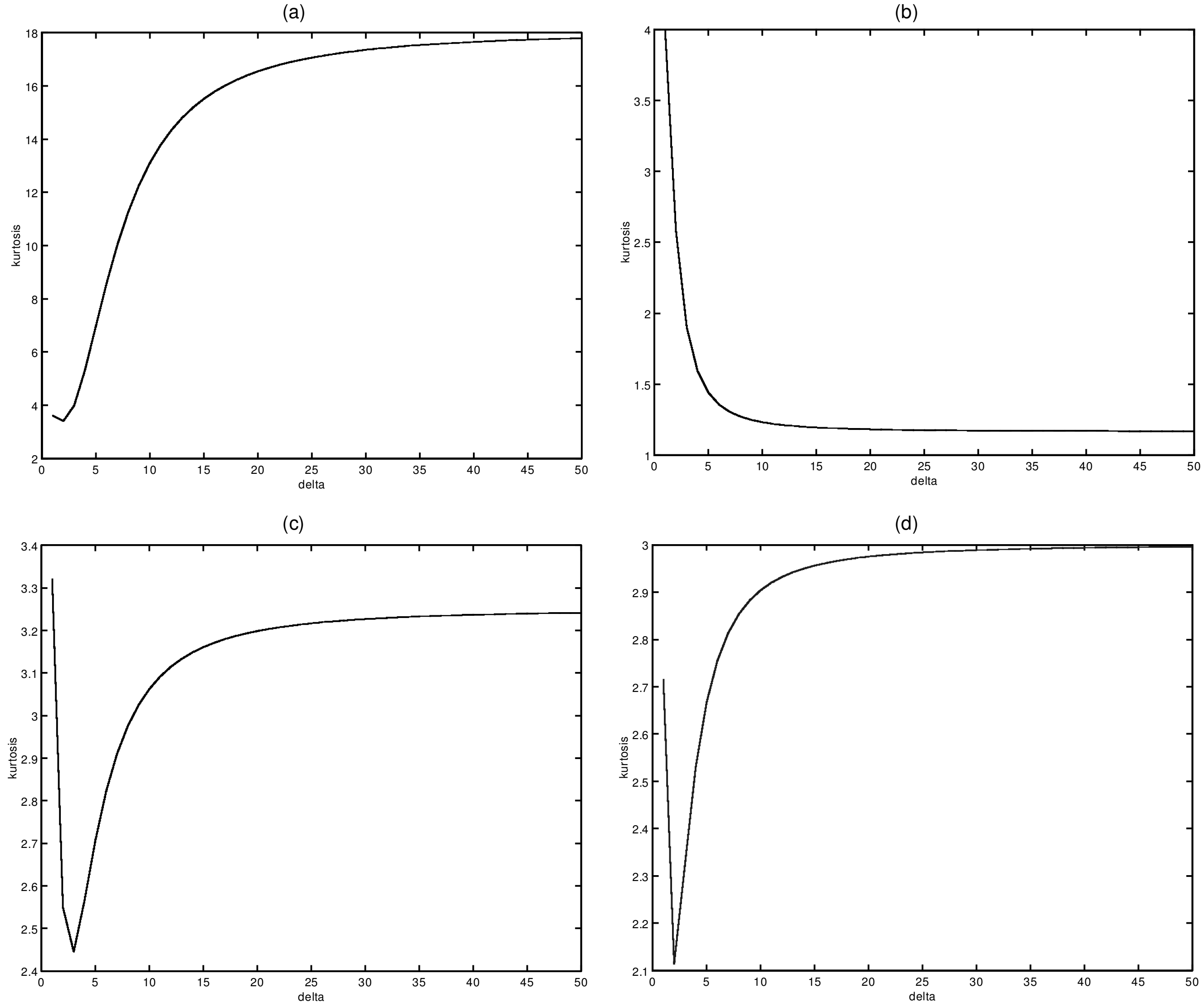

Figure 1. The Behavior of the Kurtosis at Different Contamination Patterns: (a) 5% Asymmetric Contamination, (b) 40% Asymmetric Contamination, (c) 20% Asymmetric Contamination, (d) 20% Symmetric Contamination.

used by the original Stahel–Donoho algorithm (1,000, 2,000, and 5,000 for $p = 5$, 10, and 20).

In the next column, note a large discrepancy between my results obtained for FAST-MCD and the author's results. I do not know how to explain this difference. Did the authors not use the default algorithm with 50% breakdown value? For $p = 20$, $\alpha = .2$, and $\lambda = .1$, we note that FAST-MCD has much better performance when n is 500 rather than 100. This is due to the curse of dimensionality that occurs when n/p is too small. Therefore I also considered $n = 500$ for the other simulations in $p = 20$ dimensions. Note that the author's results printed in parentheses are still based on samples of size $n = 100$.

4. *Choice of the cutoff values β_p in Step 5.* The authors did not explain how they obtained the cutoff values β_p in Table 2. Since these values are not distribution-free, I suppose they were determined by generating many samples from the normal distribution and applying the first four steps from the kurtosis1 algorithm. If so, I am very surprised that the estimated Type I errors listed in Table 7 are quite far from 5%.

5. *Sequential determination of outliers in Step 7.* The authors' version of outlyingness is only used as a first step in the outlier detection. In Step 7 the mean and covariance matrix of the good data points are computed and used to decide which outliers can still be reclassified as good observations. This procedure is repeated until no more outliers can be reallocated. My first question is why the cutoff value $\chi^2_{p-1,.99}$ is used for p-dimensional data. Second, it seems very dangerous to repeat this procedure in case the outliers are forming a chain. This is illustrated in Figure 2. This is a rather artificial example, but clearly shows what happens when the outliers are determined sequentially. The smallest ellipse is the 97.5% tolerance region based on the mean and covariance matrix of the good data points obtained after Step 6. Applying the whole kurtosis1 algorithm to this dataset yields the larger tolerance ellipse, which has been attracted by the chain of outliers. I therefore suggest that Step 7 be applied only once.

Table 1. A Comparison Among Kurtosis1 (the authors' method), rand2p (based on random sets of directions), and FAST-MCD.

p	*α*	*λ*	*δ*	*n*	*Kurtosis1*	*Rand2p*	*FAST-MCD*
5	.3	.1	10	100	100 (100) 4.88	100 10.14	57 (0)
10	.3	.1	100	100	97 (100) 1.85	100 4.34	88 (0)
		1	100	100	25 (21) 3.12	100 4.57	100 (100)
20	.2	.1	100	100	97 (100) 1.4	100 3.01	39 (0)
				500	100 (100) 3.85	100 5.09	73 (0)
		1	10	500	30 (1) 5.88	100 5.03	100 (98)
	.4	.1	10	500	100 (1) 1	100 3.93	0 (0)

NOTE: For each configuration, the first line indicates the success rate of the algorithm, the second line the success rate reported by the authors, and the third line the average number of iterations.

6. *Robustness of the estimated covariance matrix*. The final covariance matrix is said to be robust and, as the simulations indicate, this seems to be the case for certain contamination patterns. But what is the breakdown value of this estimator? It would be very useful to know, since the breakdown value is an objective measure of the robustness of the estimator toward *all* kinds of contamination. If the breakdown value is nonzero, this would give a more solid justification of this proposal, and it would allow a comparison with other robust methods that is not based only on simulations. I therefore strongly encourage the authors to work further on this topic.

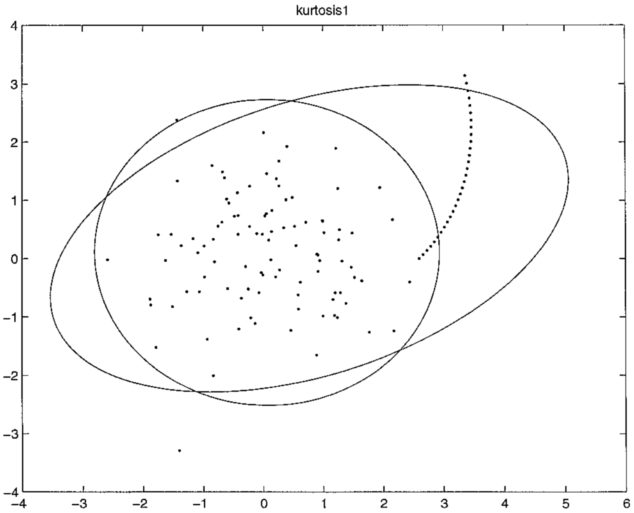

Figure 2. An Artificial Dataset Illustrating the Chaining Effect of the Covariance Matrix Computed in Step 7 of the Kurtosis1 Algorithm.

ADDITIONAL REFERENCES

Croux, C., and Haesbroeck, G. (1999), "Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator," *Journal of Multivariate Analysis*, 71, 161–190.

——— (2000), "Principal Components Analysis Based on Robust Estimators of the Covariance or Correlation Matrix: Influence Functions and Efficiencies," *Biometrika*, 87, 603–618.

Croux, C. and Ruiz-Gazen, A. (1996), "A Fast Algorithm for Robust Principal Components Based on Projection Pursuit," in *COMPSTAT: Proceedings in Computational Statistics 1996*, Heidelberg: Physica-Verlag. pp. 211–217.

——— (2000), "High Breakdown Estimators for Principal Components: The Projection-Pursuit Approach Revisited," unpublished manuscript.

Hardin, J., and Rocke, D. (2000), "The Distribution of Robust Distances," technical report, University of California-Davis, Dept. of Applied Science.

Hawkins, D. M., and McLachlan, G. J. (1997), "High-Breakdown Linear Discriminant Analysis," *Journal of the American Statistical Association*, 92, 136–143.

Hubert, M. Rousseeuw, P. J., and Verboven, S. (2000), "A Fast Method for Robust Principal Components With Applications to Chemometrics," submitted to *Chemometrics and Intelligent Laboratory Systems*.

Li, G., and Chen, Z. (1985), "Projection-Pursuit Approach to Robust Dispersion Matrices and Principal Components: Primary Theory and Monte Carlo," *Journal of the American Statistical Association*, 80, 759–766.

Pison, G., Rousseeuw, P. J., Filzmoser, P., and Croux, C. (2000), "Robust Factor Analysis," technical report, University of Antwerp, Dept. of Mathematics and Computer Science.

Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871–880.

Response

Daniel Peña and Francisco J. Prieto

We first want to thank the editor for organizing this discussion and the two discussants for their stimulating comments and valuable insights. We are grateful to them for giving us an opportunity to think harder about the problem and for clarifying several interesting issues.

Hubert indicates that, in her opinion, the MCD covariance estimator of Rousseeuw (1984) is the most respected procedure for multivariate outlier detection. We agree with her that this estimator has some excellent properties, but it also has some problems. For instance, Adrover and Yohai (2001) computed numerically the maximum bias of this estimate and showed that, when the dimension increases, the maximum bias of the MCD grows almost exponentially. This is in agreement with the result in our article in which we found that the code FAST-MCD performs badly for concentrated contamination.

Hubert raises several specific criticisms about our method, andwe thank her for this opportunity to clarify some of these issues as follows.

1. Hubert worries that the standardization using nonrobust estimators may lead to difficulties with the resulting procedure, as illustrated, for example, by Rousseeuw and Leroy (1987). This problem cannot appear in our method. First, note that the algorithm we propose is affine equivariant, independently of the initial standardization. The kurtosis coefficient is invariant to translations and scaling of the data, and a rotation will not affect the maximizers or minimizers. Moreover, we have tried to be careful when defining the operations to generate the successive directions, as well as in the choice of an initial direction for the optimization problems. As a consequence, any (invertible) linear transformation of the data, including the one used in the proposed algorithm, should not affect its results. Second, the problem illustrated by Rousseeuw and Leroy (1987) cannot happen to the class of estimators based on weighted versions of the sample mean and covariance matrix in which the suspected outliers are assigned a zero weight.The one used in the article belongs to this class. If x_i denotes the original data and $y_i = Ax_i + b$ denotes the data after a (any) linear transformation, we would define location and scale estimators for the modified data as

$$m_y = \sum_i w_i y_i, \quad S_y = \sum_i w_i (y_i - m_y)(y_i - m_y)',$$

where $\sum_i w_i = 1$. Undoing the transformations, we have

$$m_x = A^{-1}(m_y - b) = \sum_i w_i x_i$$

and

$$S_x = A^{-1} S_y (A^{-1})' = \sum_i w_i (x_i - m_x)(x_i - m_x)'.$$

Thus, if the outliers are properly identified in the transformed data, the use of the sample mean and covariance to standardize it would never cause a breakdown of the estimators, even if these sample moments (the coefficients in the transformation) become arbitrarily large due to the presence of outliers.

2. As we describe in the article, and in accordance with Hubert's comments, the behavior of the kurtosis coefficient is particularly useful to reveal the presence of outliers in the cases of small and large contaminations, and this agrees with the standard interpretation of the kurtosis coefficient as measuring both the presence of outliers and the bimodality of the distribution.

What is remarkable is that in intermediate cases with $\alpha = .3$ the procedure does not break down completely and its performance improves with the sample size, as we show in Table 3. Although we cannot claim to understand perfectly well the behavior of the algorithm in these cases, we think that some explanations for the observed results can be found in the iterative application of the technique. If a few of the outliers are detected in some iterations (by chance or because they form a small cluster), the remaining outliers will constitute a smaller percentage of the remaining sample, and they will be easier to detect by the procedure. This seems to indicate that a divide-and-conquer strategy (that we have not implemented) could be of help to improve the practical behavior of the algorithm.

We do not see the relevance to our procedure of the univariate confidence interval for the kurtosis coefficient at this point of the discussion. A multivariate kurtosis test could be built by using the results of Machado (1983), who derived the distribution of the maxima and minima of the kurtosis of the projected data along u, where u is a unit norm vector, for multivariate normal data, and Baringhaus and Henze (1991), who derived the asymptotic distribution of the maximum of the projected kurtosis for elliptically symmetric distributions.

3. The $2p$ maximization and minimization directions generated by the algorithm are considered simultaneously. In this sense, Hubert's proposal to alternate between directions would be of interest if we used a procedure that stops once a significant direction is computed. The algorithm we describe does not make use of this feature, and in this sense it is a simpler one to describe and understand, although it may be more expensive to implement.

Another question raised by Hubert is the choice of orthogonal directions. Our motivation to use these orthogonal directions is twofold. On the one hand, we wish the algorithm to be able to identify contamination patterns that have more than one cluster of outliers. For example, this would be the case if we have individually scattered outliers. If these clusters are apart in space, a single projection direction would be able to find only one of the clusters, at most. Clearly, it would be of interest to try to identify several of these clusters in a single pass, by using more than one direction. To ensure that these directions do not provide overlapping information, we have chosen them to be orthogonal. The second motivation arises from a property of the kurtosis, described in Section 1.1 of the article, that implies that in some cases the directions of interest are those orthogonal to the maximization or minimization directions.

Hubert carried out an experiment to determine the relevance of different choices for the projection directions, based on the use of random orthogonal directions. Unfortunately, the generation of random directions, as she suggested, is not an affine equivariant procedure. Rather than discussing the importance of outlier-identification procedures being affine equivariant, we prefer to illustrate the impact of using a procedure that does not satisfy this property. We carried out the same experiment described by Hubert, but we standardized the data in advance. Since our procedure is affine equivariant, the results we obtain are the same (apart from statistical variability) in both cases, whereas the results of the procedure used by

Table 1. Success Rates for Outliers in One Cluster: Random Directions

p	α	$\sqrt{\lambda}$	δ	n	Kurtosis	Rand2p
5	.3	.1	10	100	100	2
10	.3	.1	100	100	97	0
		1	100	100	35	25
20	.2	.1	100	100	94	0
				500	100	0
		1	10	500	30	41
	.4	.1	10	500	100	0

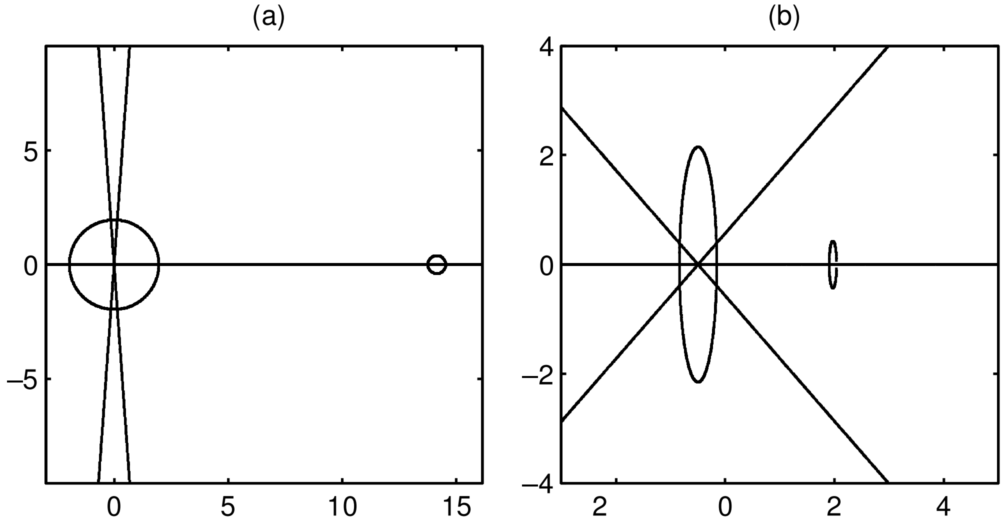

Figure 1. Cones of Random Directions for the Identification of Outliers: (a) Data Used in the Simulations; (b) Standardized Data.

Hubert change completely. Table 1 shows that the use of random directions is a very poor choice for this data.

The reason the random directions are so efficient for the data used in the simulation made by Hubert is that the outliers were located quite far out from the uncontaminated data. Figure 1 attempts to illustrate this point. In Figure 1(a), we represent data generated as in the simulation experiment and the cones of directions that would be able to separate the outliers in the projections. The useful directions are those outside the narrow cones. Figure 1(b) presents the same data after a linear transformation to standardize it. It is readily apparent that after the transformation the set of useful directions is much smaller than before.

It could be shown that the probability of the projection of the center of the outliers being beyond the $z_{1-\alpha}$ quantile of the standard normal for the data used in the simulation experiment, will be given by

$$P\left(|Y| \geq \frac{z_{1-\alpha}\sqrt{p-1}}{\sqrt{\delta^2 p - z^2_{1-\alpha}}}\right),$$

where Y follows a Student's t_{p-1} distribution. For $\alpha = .01$, $\delta = 10$, and $p = 5$, this probability is approximately equal to .83, and for $\delta = 10$ and $p = 20$, the corresponding value is .80. These values change significantly after a linear transformation is applied to the data. We prefer to use an affine equivariant algorithm based on the maximization and minimization of the kurtosis coefficient, even if in some cases it may behave worse than other alternatives.

The different results obtained with FAST-MCD are due to the different values used for the variance of the contaminating distribution. We must apologize to Hubert for introducing a source of confusion in the previous version of the article, in which in Section 1.1 λ was the variance of the contaminating distribution, whereas in the simulations λ was the standard deviation. This has been corrected in the published version of the article. Thanks to her careful checking of our results, which we appreciate very much, a possible source of confusion for the readers of the article has been removed. Table 2 gives the values we obtain for FAST-MCD for several values of λ, which basically correspond to those indicated by Hubert and show that, in agreement with previous results, FAST-MCD works better for less concentrated contamination.

4. The procedure used to compute the values in Table 3 of the article is in fact the one mentioned by Hubert in her comments. The results are unfortunately not totally satisfactory; the reason is the large variability in these values in the simulations. This variability has two main effects—it is difficult to find correct values (huge numbers of replications would be required) and for any set of 100 replications there is a high probability that the resulting values will be far from the expected one. Nevertheless, we agree that these values could be estimated with greater detail, although they do not seem to be very significant for the behavior of the algorithm, except for contaminations very close to the original sample.

5. We again thank Hubert for detecting a misprint in the article that we have corrected in this printed version. In the version she discusses, it was wrongly written that the cutoff value was χ^2_{p-1} instead of χ^2_p. Actually, in the code that we have made public the percentiles are obtained from the Hotelling-T^2 approximation instead of the χ^2_p, which behaves better for small sample sizes.

Table 2. Success Rates for Outliers in One Cluster: FAST-MCD

p	α	$\sqrt{\lambda}$	δ	n	Fast-MCD
5	.3	.1	10	100	0
		.32	10	100	55
10	.3	.1	100	100	0
		.32	100	100	94
		1	100	100	100
20	.2	.1	100	100	0
		.32	100	100	31
		.1	100	500	0
		.32	100	500	57
		1	10	500	100
	.4	.1	10	500	0
		.32	10	500	0

Table 3. Success Rates for One Cluster of Outliers in p = 20 Dimensions: Different Sample Sizes

α	$\sqrt{\lambda}$	δ	n	Fast-MCD	Kurtosis
.2	.1	100	200	0	99
			500	0	99
			1,000	0	98
			2,000	0	98
.3	1	100	200	29	1
			500	40	52
			1,000	52	99
			2,000	50	100
.4	1	100	200	2	59
			500	1	100
			1,000	2	100
			2,000	0	100

Regarding the suggestion of applying Step 7 only once, we have found that this requires a very careful calibration of the cutoff values, and as a consequence it becomes very sensitive to deviations from normality in the data. We are a little surprised by her general criticism of procedures that determine the outliers sequentially. The statistical literature is full of examples of very successful sequential procedures and, to indicate just one, Peña and Yohai (1999) presented a sequential procedure for outlier detection in large regression problems that performs much better than other nonsequential procedures.

6. Regarding breakdown properties, we agree with Hubert that it would be of great interest to prove a high-breakdown-point property of the proposed algorithm to confirm the good behavior that it shows in the simulations.

Rocke and Woodruff mention some difficulties arising from the specific contamination patterns used in the simulation experiment:

We agree that in dimension 20 a sample of size 100 is small. To check the effect of increasing the sample size on FAST-MCD and on our proposed algorithm, we have performed a small simulation experiment, reported in Table 3. As expected, the success rate improves with the sample size. It is interesting to stress that in the worst case for the kurtosis algorithm, $\alpha = .3$ and $\lambda = 1$, the performance improves very fast with the sample size. This suggests that our algorithm can be a fast and simple alternative for detecting outliers and computing robust estimates in large datasets in high dimensions.

We regret that Rocke and Woodruff may have been misled by the notation we used in Section 1.1 and in the description of the simulation experiment. In this experiment, the outliers have been generated to have mean δe, where e denotes a vector of ones—that is, the same vector that these authors denote as u in their comments. As a consequence, the distance between the means of both distributions is $\delta\sqrt{p}$ (rather than δ).

This situation modifies some of the conclusions from the Rocke–Woodruff analysis. For example, the first case has a standard deviation of the outliers equal to $5, \delta = 10$, and $p = 20$, and the .999 sphere with radius 33.65 containing the outliers would be centered on a point at a distance $10\sqrt{20} = 44.72$ from the center of the uncontaminated data. As a consequence, the outliers in this case would lie on one side of the original data, and would not surround it with high probability. The same analysis would also imply that for $\lambda = 1$ the spheres around the original data and the contaminating cluster would not overlap.

We agree with Rocke and Woodruff that methods like FAST-MCD or MULTOUT must be supplemented by other methods for concentrated contaminations. The proposal of Juan and Prieto (1997) would be very useful in this regard because it is designed just for this specific objective. However, we do not think that kurtosis1 is useful only for concentrated outliers. As we have shown in the simulations, it works as well as FAST-MCD when $\sqrt{\lambda} = 5$, whereas it leads to a much smaller proportion of good observations mislabeled as outliers when the dimension grows; see Table 6 in the article. We have conducted another limited simulation experiment for the worst-case amount of contamination, $\alpha = .3$, and different concentrations of the contamination presented in Table 4.

From Table 4 we see that the reduction in performance tends to concentrate closely around $\lambda = 1$, and for values of $\sqrt{\lambda}$ on the order of .5–.7 and 1.5, the performance seems to be back to reasonable levels. However, this reduction of performance disappears for large sample size, as was shown in Table 3.

However, since kurtosis1 does not work well for values very close to $\lambda = 1$ and small sample sizes, it would be interesting to find an optimal combination of this code and some of the other available codes. Since we have not included comparisons with MULTOUT in the article, we will use only the results from Tables 6 and 8 for FAST-MCD and SD. The best option from these tables seems to be a combination of SD and kurtosis1. Thus the directions obtained by the kurtosis algorithm can be used to compute a fast version of the Donoho–Stahel estimate. It is also possible to combine these directions with others obtained by resampling, searching for specific as well as random directions. Peña and Yohai (1999)

Table 4. Success Rates for Outliers in One Cluster: Different Concentrations

p	α	$\sqrt{\lambda}$	δ	n	Kurtosis
10	.3	.3	100	100	99
		.5	100	100	95
		.7	100	100	54
		.9	100	100	26
		1	100	100	32
		1.2	100	100	69
		1.5	100	100	98
		2	100	100	100
20	.3	.3	100	500	99
		.5	100	500	100
		.7	100	500	100
		.9	100	500	88
		1	100	500	75
		1.2	100	500	86
		1.5	100	500	100
		2	100	500	100

suggested that this type of combination could be very successful in robust regression, and we think that the empirical evidence available suggests it would also be very promising for detecting multivariate outliers.

We agree completely with Rocke and Woodruff in the interest of using clustering algorithms when a concentrated contamination is suspected to be present (a case very closely related to the standard cluster identification problem) or as a complement to other robust outlier-identification procedures. The results that they mention seem very promising, and we are looking forward to reading the two manuscripts referenced in their report with their new results. Note also that their comments can be interpreted to work in the opposite direction—that is, that an algorithm that is efficient for the detection of concentrated contaminations may be a reasonable basis for a clustering algorithm. Along these lines we have carried out some work (Peña and Prieto 2001) to improve on the performance of the kurtosis-based method in a cluster-analysis setting. Basically, we have combined the idea of projections onto interesting directions with an analysis of the spacings of the projections to identify groups in the data. Both the theoretical analysis and the simulation results seem to indicate that this approach may be of interest, at least for linearly separable clusters. One theoretical motivation for the procedure (Peña and Prieto 2000) is that if we have two clusters from two normal populations with different means but the same covariance matrix, the direction that minimizes the kurtosis of the projection is the Fisher linear discriminant function. Thus we can obtain the optimal discriminant direction without knowing the means or the common covariance matrix, which allows its application to cluster analysis. We have also proved (Peña and Prieto 2001) that if we have a sample generated by two elliptical distributions with different means and covariance matrices, the directions that maximize or minimize the kurtosis coefficient belong to the admissible linear classification rules as defined by Anderson and Bahadur (1962). We conduct a search for groups along each of these directions using spacings. The performance of the resulting cluster procedure is very encouraging.

We would like to finish this response by thanking both discussants for their thought-provoking comments that have helped us to understand the procedure better. We agree with Rocke and Woodruff that a combination of different procedures and approaches seems to be the most promising strategy to find multivariate outliers. This suggests exploring ideas from cluster analysis, robust statistics, and outlier-detection methods using both classical and Bayesian inference. The mixture of different cultures and approaches has been the main source of change and improvement in our world, and we believe that this is also true for statistics.

REFERENCES

Adrover, J., and Yohai, V. (2001), "Projection Estimates of Multivariate Location," unpublished manuscript.

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

Baringhaus, L., and Henze, N. (1991), "Limit Distributions for Measures of Multivariate Skewness and Kurtosis Based on Projections," *Journal of Multivariate Analysis,* 38, 51–69.

Machado, S. G. (1983), "Two Tests for Testing Multivariate Normality," *Biometrika,* 3, 713–718.

Peña, D., and Prieto, F. J. (2000), "The Kurtosis Coefficient and the Linear Discriminant Function," *Statistics and Probability Letters,* 49, 257–261.

——— (2001), "Cluster Identification Using Projections," working paper, Universidad Carlos III de Madrid, Dept. of Statistics and Econometrics.

Peña, D., and Yohai, V. (1999), "A Fast Procedure for Outlier Diagnostics in Large Regression Problems," *Journal of the American Statistical Association,* 94, 434–445.

Rosseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871–880.