Using Angles to Identify Concentrated Multivariate Outliers

Jesus Juan

Lab. de Estadística (ETSI Industriales)
Universidad Politécnica de Madrid
28006 Madrid
Spain
(*jjuan@etsii.upm.es*)

Francisco J. Prieto

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

This article describes a procedure for the detection of multivariate outliers based on the analysis of certain angular properties of the observations. The method is simple, exploratory in nature, and particularly well suited for the detection of concentrated contamination patterns, in which the outliers appear to form a cluster, separated from the sample. It is shown that it presents good properties for the identification of contaminations on high-dimensional sample spaces and for high contamination levels, including some cases in which methods based on robust estimators (the minimum covariance determinant and minimum volume ellipsoid estimators, the Stahel–Donoho estimator, or other recent proposals) may fail. The use of the procedure is illustrated through several examples.

KEY WORDS: Exploratory data analysis; Q-Q plot; Robust estimation.

Data often include some outliers. If the mechanism generating the observations were perfectly well known, it would be possible to detect and explain those abnormal observations. Often such information is unavailable, so outliers must be determined on the basis of data analysis. The need to identify the outliers is an immediate consequence of the distortions that they introduce on the results obtained from the application of classical estimation procedures to contaminated samples.

Except for low-dimension cases (samples in dimensions 1, 2, or at most 3), in which a complete graphical representation of the data may be used to visually identify the potential outliers, detecting multivariate outliers is difficult with no completely satisfactory procedure available for the general case. The usual strategy is based on the computation of some Mahalanobis distance for each observation $\mathbf{x} \in \Re^p$, defined as

$$D_{\mathbf{V}}(\mathbf{x}, \mathbf{c}) = \left\{(\mathbf{x}-\mathbf{c})^T \mathbf{V}^{-1}(\mathbf{x}-\mathbf{c})\right\}^{1/2}, \tag{1}$$

where $\mathbf{c} \in \Re^p$ and the $p \times p$ matrix $\mathbf{V}$ denote, respectively, the estimators for the center and the covariance matrix obtained from the sample points.

However, outliers may result in unreliable distance values when $\mathbf{c}$ and $\mathbf{V}$ are the sample mean and sample covariance, respectively. In recent years many robust alternatives for these estimators have been proposed in the literature, such as the M estimators, studied by Maronna (1976) for the multivariate case, the estimator based on the minimum volume ellipsoid (MVE) (Rousseeuw 1985) and its derivations such as the minimum covariance determinant (MCD) method, or the Stahel–Donoho estimator (SDE) (Stahel 1981; Donoho 1982).

Direct implementations of M estimators may present a very low breakdown point, $1/(p+1)$, and those versions that have a high breakdown point, such as the S estimators, are very expensive to compute even for moderate sample-space dimensions. The other two estimators have a 50% breakdown point, independently of the dimension of the data, but their exact computation is also expensive. This computation requires solving a global optimization problem with a nonconvex objective function that in general presents a large number of local minimizers. Solution techniques currently available for this problem are too inefficient to be of practical use, even for low-dimension problems. As a consequence, in practice approximate solutions based on resampling procedures or heuristic procedures are used for both cases. A detailed description of the advantages and limitations of these estimators was given by Rousseeuw and van Zomeren (1990), Cook and Hawkins (1990), and Maronna and Yohai (1995).

In particular, these methods will have difficulty identifying contaminations that are not far from the original sample, even when they present other distinguishing features. An example illustrating this last situation is the case of concentrated contaminations. In this contamination pattern the outliers are closely grouped, forming clusters separated from the main sample. The effect of this contamination scheme was analyzed by Maronna and Yohai (1995), who suggested that this scheme may induce the largest bias in the estimation of location and scale for multivariate samples. Adrover (1993) showed that this is the case for M estimators. It also seems to be the most difficult case for algorithms based on the MVE (see Rocke and Woodruff 1996).

In this work, a procedure that takes into account other information, in addition to distances, is proposed. This procedure is illustrated through its application to the particular case of the detection of concentrated contaminations. It is shown to present properties that are complementary to those

of distance-based methods and in particular to be able to identify outliers in cases in which these methods may fail.

The method proceeds by projecting the standardized data onto the unit hypersphere and testing the projected data to identify any lack of uniformity that might be associated with the presence of outliers. This procedure is based on the observation that the anomalies that characterize many contamination patterns, and in particular those difficult to detect by distance-based methods, arise from the relative disposition of the contaminating observations and, more specifically, from an excessive proximity between these observations. The distortions introduced on the projected data by this proximity may be easier to detect than the presence of large values for some robust distance to the center of the data. Note that it is possible to have contaminations that have large distances without distorting any symmetry properties. As a consequence, a reasonable procedure should study both distances and angles. We propose a practical two-step method based on the combination of a distance-based algorithm with the one presented in this article.

This article will be concerned only with the details of the method related to the analysis of the angles, without further reference to the analysis of distances. This latter part has been studied extensively in the literature (e.g., see Hawkins 1980; Beckman and Cook 1983) and can be carried out using one of several procedures (e.g., see Rousseeuw and van Zomeren 1990; Rocke and Woodruff 1996; Barnett and Lewis 1994; Rousseeuw and van Driessen 1999).

In Section 1, we analyze some characteristics of concentrated contamination patterns that justify the use of the outlier-detection procedure described in the article. Section 2 introduces the procedure and justifies its validity. Section 3 studies some properties of the procedure in terms of the configuration of the sample. Finally, in Section 4, the practical behavior of this procedure is illustrated on some representative examples.

1. MOTIVATION

Before describing the proposed procedure, we illustrate some of the practical difficulties with distance-based outlier-identification methods. In particular, these procedures may fail to detect outliers when these observations appear grouped together and not very far from the uncontaminated sample.

Table 1 presents the results of a simulation experiment conducted using several available codes based on MCD techniques, and an implementation of the Stahel–Donoho estimator (SDE) procedure: FSAMCD from Hawkins (1994), MULTOUT from Woodruff and Rocke (1996), FAST-MCD from Rousseeuw and van Driessen (1999), and the SDE implementation of Maronna and Yohai (1995). The experiment consists of randomly generating a sample of n observations, whose majority subset of $(1-\epsilon)n$ observations is generated from an $N(0, \mathbf{I})$ distribution in dimension p and whose minority subsets of ϵn observations (the outliers) come from an $N(k\mathbf{e}_1, \lambda^2\mathbf{I})$ distribution, where $\mathbf{e}_1$ denotes the first unit vector. One hundred samples were generated for each set of parameter values ($p = 5, 10, 20$; $n = 10p$; $\epsilon = .05, .1, .15, .2$, and $\lambda = .1$). The distance of the outliers to the center of the uncontaminated data, k, was set to the values $2\sqrt{\chi^2_{p,.95}}$ and $4\sqrt{\chi^2_{p,.95}}$, as shown by Rocke and Woodruff (1996).

Table 1. Simulation Experiment: Number of Successes in Identifying All Concentrated Outliers Using the Codes FAST-MCD, MULTOUT, FSAMCD, and SDE

ϵ Cont.	p Dim.	k Dist.	% success FAST-MCD	MULTOUT	FSAMCD	SDE
.05	5	6.65	100	100	100	100
		13.31	100	100	100	100
	10	8.56	100	100	100	94
		17.11	100	100	100	100
	20	11.21	76	80	91	4
		22.42	100	100	100	100
.10	5	6.65	98	98	100	99
		13.31	100	100	100	100
	10	8.56	16	59	26	77
		17.11	100	99	98	100
	20	11.21	0	2	0	0
		22.42	0	13	0	100
.15	5	6.65	69	60	80	93
		13.31	100	100	100	100
	10	8.56	0	8	0	26
		17.11	5	31	1	100
	20	11.21	0	0	0	0
		22.42	0	0	0	92
.20	5	6.65	0	18	1	55
		13.31	59	92	77	100
	10	8.56	0	0	0	3
		17.11	0	5	0	100
	20	11.21	0	0	0	0
		22.42	0	0	0	24

The codes were run on each sample, and their output was compared to the actual outliers from the preceding model. Table 1 gives the number of times each code was able to identify as outliers all the sample points generated from the contaminating model, for each code and each set of parameter values. For FSAMCD, a success was declared when none of the outliers were contained in the basis returned by the code (only one solution was tracked). For the remaining codes, the decision was based on the labeling of the observations provided in their output files.

From the results in Table 1, MCD-based methods work reasonably well for reduced contamination levels ($\epsilon = .05, .1$) and sample-space dimensions ($p = 5$), but they have increasing difficulties in identifying concentrated outliers as the contamination level and the sample-space dimension increases. In fact, the percentage of outliers for which these methods start to fail seems to decrease monotonically with the dimension p. The conclusions for the SDE method are similar, although the deterioration is less marked. The preceding results on the MCD-based methods are similar for other values of scale contamination ($\lambda = .32, .032$).

To examine further this behavior, we consider a particular sample obtained from the preceding model having low success rates. The sample has been generated using $p = 20$, $n = 200$, $\epsilon = .1$, $k = 11.21$, and $\lambda = .32$. The Mahalanobis distances to the origin (the center of the original sample) for all the observations, computed using both the sample covariance matrix and the covariance matrix for the first 180 observations,

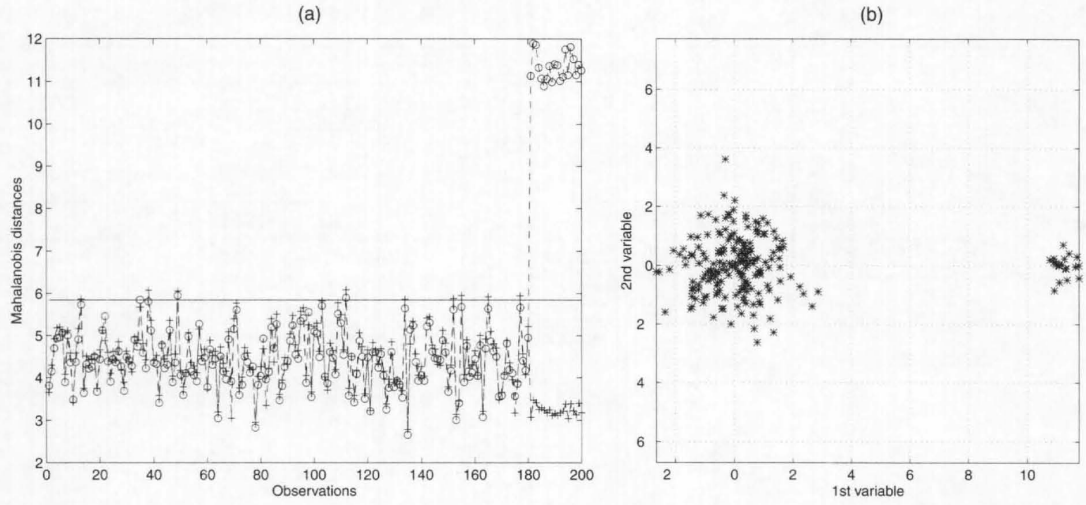

Figure 1. (a) Mahalanobis Distances for the Dataset Using the Mean and Covariance Matrix for the First 85 Observations (o) and the Mean and Covariance Matrix for all 100 Observations (+); (b) Scatterplot of the Observations Projected Onto the First Two Coordinate Axes.

are shown in Figure 1(a). The outliers (observations 181–200) are readily apparent from the values of the distances (indicated with an "o") computed using the uncontaminated covariance matrix, while the Mahalanobis distances (labeled with a "+") computed using the whole sample would not reveal any outliers. The horizontal line in the plot corresponds to the value $\sqrt{\chi^2_{20,.975}}$ (the cutoff value used in the code FAST-MCD), given as a reference for the actual distance of the outliers to the center of the sample.

Figure 1(b) presents the scatterplot of the projections onto the first two coordinates. These projections show the anomalous character of these observations due to both their distance to the center of the remaining observations and their relative concentration. The robust distances obtained using the four codes described previously for this example are shown in Figure 2. Again, the horizontal lines correspond to the FAST-MCD cutoff value $\sqrt{\chi^2_{20,.975}}$ and are included as a visual reference.

These plots show that none of the procedures is able to identify the outliers in this example, as might be expected from the simulation results in Table 1. A more remarkable result from these plots, and a consequence of the concentration in the contamination, is that in all cases the methods have failed to identify any of the outliers. This last behavior is common for those cases associated with clear failures in Table 1. It may also be of interest to comment that many regular observations would have been labeled as outliers by both FAST-MCD and FSAMCD.

The simulation results and the previous example illustrate that robust estimators with high breakdown points ensure the identification of "far" outliers but may fail in cases in which the outliers are concentrated and not too far away from the uncontaminated sample. In practical cases, we seek to determine not only the existence of outliers but also the extent to which they cluster. This latter anomaly would not be readily apparent from an analysis based exclusively on Mahalanobis distances.

Outlier detection procedures based on Mahalanobis distances could be improved in these cases if angular information on the data were taken into account, together with the robust distances computed by codes such as FAST-MCD, FSAMCD, or MULTOUT. The method that we introduce in Section 2 is based on the analysis of the distortions introduced by the contamination on the distribution of the angles between observations. In the cases illustrated previously and in other cases analyzed in Section 4, these distortions are far easier to detect than anomalies in the distribution of the distances.

2. PROJECTIONS ONTO A HYPERSPHERE: A METHOD BASED ON ANGLES

In this section, we propose an outlier-detection procedure and present the specific properties of the angles on which it is based. Let $\mathbf{X}$ denote a random vector in $\Re^p$ with distribution function F. Assume that F is the (ellipsoidal) distribution function of $\mathbf{X} = \mathbf{PY} + \boldsymbol{\mu}$, where $\mathbf{P}$ is a nonsingular $p \times p$ matrix and $\mathbf{Y}$ has a spherical (isotropic) distribution; that is, for any orthogonal $p \times p$ matrix $\boldsymbol{\Gamma}$, both $\mathbf{Y}$ and $\boldsymbol{\Gamma}\mathbf{Y}$ have the same distribution. A multivariate normal would be an example of an ellipsoidal distribution. Let $\mathcal{S}_{p-1} = \{\mathbf{x} \in \Re^p : \|\mathbf{x}\| = 1\}$ denote the unit hypersphere in $\Re^p$. The vector $\mathbf{U} = \mathbf{Y}/\|\mathbf{Y}\|$ has a uniform distribution on $\mathcal{S}_{p-1}$ (Eaton 1983).

The proposed method will be based on assuming that for the uncontaminated sample $\mathbf{U}$ follows a uniform distribution $\mathcal{S}_{p-1}$. This would be the case, for example, if the uncontaminated data came from an ellipsoidal distribution and in particular if it followed a multivariate normal distribution. The test for the uniformity of $\mathbf{U}$ will be based on a related univariate distribution, that of the angle between $\mathbf{U}$ and a given reference direction $\mathbf{u}_0$. For a given vector $\mathbf{u}_0$ the distribution function of W, the angle between $\mathbf{u}_0$ and $\mathbf{U}$ (see Fig. 3), can be obtained from the normalized surface measure for the spherical patch corresponding to the angle

$$F_W(w) = K \int_0^w \sin^{p-2} t \, dt, \qquad 0 \le w \le \pi, \tag{2}$$

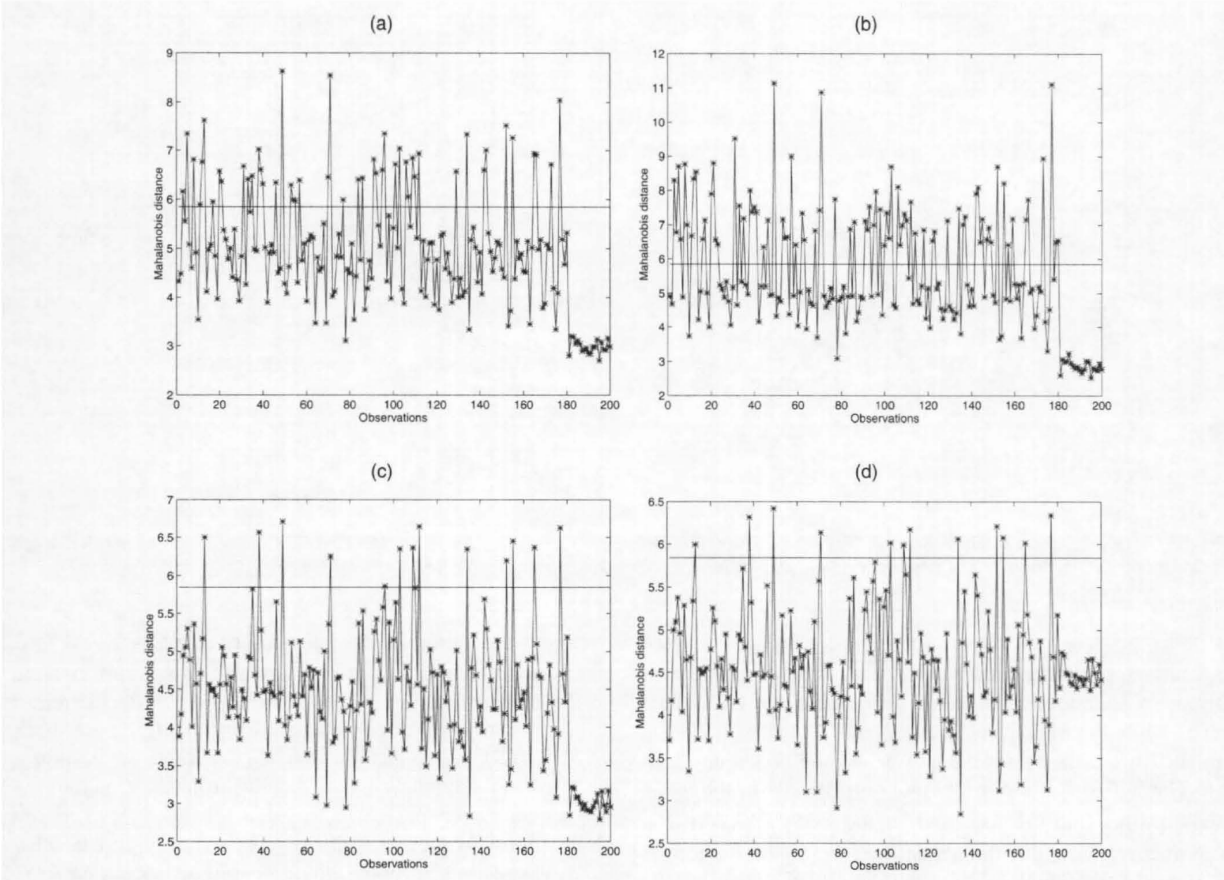

Figure 2. Robust Mahalanobis Distances for the Synthetic Dataset, Computed using (a) FAST-MCD, (b) FSAMCD, (c) MULTOUT, (d) SDE.

where K is the normalizing constant. Some authors refer to the angle W as the Mahalanobis angle; Fisher (1938) seems to have been first to use this concept. Mardia (1977) outlined its role in various techniques such as factor analysis and discriminant analysis.

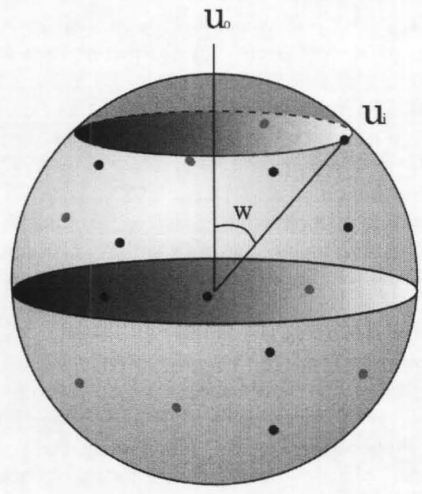

Figure 3. Geometric Representation of the Probability Distribution Associated With the Angle Between Observations for an Ellipsoidal Distribution.

Using the change of variables $u = \sin^2 t$, the preceding equation can be written as

$$F_W(w) = \begin{cases} 1/2 I(\sin^2 w; (p-1)/2, 1/2), & 0 \le w \le \pi/2, \\ 1 - 1/2 I(\sin^2 w, & \\ \quad (p-1)/2, 1/2), & \pi/2 \le w \le \pi, \end{cases} \tag{3}$$

where $I(z; a, b)$ corresponds to the beta distribution with parameters a, b. The q_β quantile of the distribution (3) can be obtained from

$$q_\beta = \begin{cases} \sin^{-1}\sqrt{z_{2\beta}}, & 0 \le \beta \le 1/2, \\ \pi - \sin^{-1}\sqrt{z_{2(1-\beta)}}, & 1/2 \le \beta \le 1, \end{cases} \tag{4}$$

where z_α is the α quantile of the beta distribution with parameters $(p-1)/2$ and $1/2$.

Let the given sample in $\Re^p$ be denoted by $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. We compute the values $\mathbf{y}_i = \mathbf{S}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}})$, where $\bar{\mathbf{x}}$ and $\mathbf{S}$ are the sample mean and the covariance matrix. The sample mean and the covariance matrix have been chosen to standardize the data as the best alternatives in the case of absence of contamination. Note that using the sample mean and the covariance matrix in the presence of outliers will in general help to detect the outliers by introducing additional asymmetries in the angles that should be readily apparent to the tests used to identify the presence of outliers.

Next, the observations are projected onto the unit hypersphere $\mathcal{S}_{p-1}$ by computing $\mathbf{u}_i = \mathbf{y}_i / \|\mathbf{y}_i\|$. Then, for a given reference direction $\mathbf{u}_0$, selected in the manner to be described, the procedure computes the angles between the observations $\mathbf{u}_i$ and the reference direction

$$w_i = \cos^{-1}(\mathbf{u}_0^T \mathbf{u}_i). \tag{5}$$

These values w_i form a univariate sample. They are tested to see if they follow the distribution defined by (3). This should be the case in particular in the absence of outliers. Procedures to conduct this test will be described.

2.1 Direction for the Projections

We need to select an adequate reference direction $\mathbf{u}_0$, as mentioned previously. The importance of this choice lies in the fact that the departure from uniformity for a given contaminated sample may be far more significant for some directions than for others. For example, with concentrated contaminations, the directions from the center toward the contamination are much better able to reveal the presence of outliers. As a consequence, in this case the reference direction should be chosen to be as close as possible to the direction of the outliers—that is, the direction from the center of the regular observations to the center of the outliers.

In practice, we have found that a very good approximation to this direction can be obtained from the following procedure:

1. Consider the normalized direction $\mathbf{u}_k = \mathbf{y}_k / \|\mathbf{y}_k\|$ from the center of the data to each observation $k = 1, \ldots, n$, where $\mathbf{y}_k = \mathbf{S}^{-1/2}(\mathbf{x}_k - \bar{\mathbf{x}})$. Compute the corresponding value of the function $z(\mathbf{u}_k)$, defined as

$$z(\mathbf{u}_k) = \sum_{i=1}^{n} (v_{(i)} - r_i)^2$$

$$v_i = \mathbf{u}_i^T \mathbf{u}_k, \quad i = 1, \ldots, n,$$

where r_i denotes the value of $\cos q_\beta$ for $\beta = (i - .5)/n$ and q_β is the quantile defined in (4); $v_{(i)}$ denotes the ith ordered value of v_i. The function $z(\mathbf{u})$ measures the lack of uniformity in the cosines of the angles formed by the observations and the reference direction $\mathbf{u}$. We have found it more efficient to look at these cosines, rather than the angles because they are linear functions of the directions $\mathbf{u}$. We determine the direction $\mathbf{u}_i$ that provides the largest value for z, $\mathbf{u}_i \in \arg\max_k z(\mathbf{u}_k)$.

2. Using this direction as the initial point, we solve the continuous optimization problem

$$\begin{aligned} &\max_{\mathbf{u}} \quad z(\mathbf{u}) \\ &\text{subject to} \quad \|\mathbf{u}\| = 1. \end{aligned} \tag{6}$$

This is a quadratic optimization problem with discontinuous first derivatives, whose solution can be computed using some nondifferentiable optimization procedure, for example. In practice, we have found that differentiable (Newton-method based) procedures also work quite well. The solution of (6) is used as the reference direction $\mathbf{u}_0$ for all subsequent computations in the proposed procedure.

The function $z(\mathbf{u})$ in Problem (6) presents many local extrema. The choice of initial direction in Step 1 of the procedure has been designed to ensure that the local minimizer chosen in Step 2 is a very good reference direction. In particular, if the contamination would be highly concentrated, step 2 could be omitted without any significant impact on the results. For the simulation study in Table 1, we have verified that the direction computed from the preceding procedure is very close to the direction of the outliers, $\mathbf{e}_1$. The average cosine between both directions for all the simulations in the table is .99, with a standard deviation smaller than .01.

2.2 Quantile-Quantile Plots and Gaps

Once the sample $\{w_i\}$ has been generated using (5), a goodness-of-fit test must be conducted to determine if there are outliers in the sample. This test can be carried out using several procedures.

Here, we use the quantile-quantile (Q-Q) plot—that is, a plot of $(f_i, w_{(i)})$, where $w_{(i)}, i = 1, \ldots, n$, denotes the ith ordered statistic, n, and f_i is the quantile $(i - .5)/n$.

Figure 4 shows the Q-Q plot for the dataset introduced in Section 1, for a reference direction obtained using the procedure described in the preceding section. Note that this reference direction corresponds very well to the direction of the outliers. As a consequence, a very large gap is clearly visible in the plot and separates the last 20 contaminating observations from the initial 180. This lack of uniformity of the projections onto the sphere illustrates the expected pattern for the case of concentrated contaminations.

We now study the spacing in the projected data from the Q-Q plot to derive a quantitative measure of the lack of fit and test for the presence of outliers. The spacings in the data are defined as the differences between consecutive ordered observations in univariate samples. Although any other standard statistic may be used, resulting in more powerful tests than the one suggested later, in our case the spacing presents some advantages. The presence of outliers introduces significant gaps between the observations and thus large gaps in the spacings. The mathematical expressions related to these

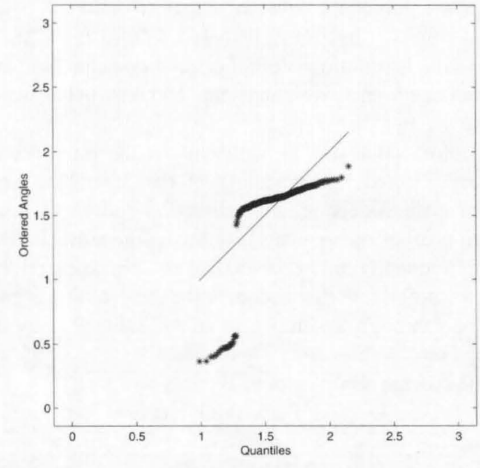

Figure 4. Q-Q Plot for the Synthetic Sample Described in Section 1.

values are analytically manageable, allowing the analysis of the behavior of the method with respect to changes in the contamination level, the concentration of the outliers, or the sample-space dimension. Additionally, this method provides useful information for separating the outliers from the regular observations.

The use of spacings for tests of fit has been suggested by several authors. An excellent treatment of this subject was given by Pyke (1965). We briefly describe its application to our particular case. Let $W_1, W_2, \ldots, W_n$ be independent univariate random variables each with distribution function F on $[0, \pi)$. (If the multivariate observations do not contain any outliers, we shall have $F = F_W$.) Let $\{W_{(1)}, W_{(2)}, \ldots, W_{(n)}\}$ denote the order statistics and set $W_{(0)} = 0$ and $W_{(n+1)} = \pi$. The spacings of the sample are defined by $D_i = W_{(i)} - W_{(i-1)}$, for $1 \le i \le n+1$. Note that these spacings or gaps can be measured directly on a Q-Q plot.

The outlier identification problem in this setting is based on testing the sizes of the spacings. Given the distribution function for the uncontaminated case, we expect to find smaller spacings in the middle of the data and relatively larger ones in the extremes. As the dimension for the multivariate observations increases, the size of the middle spacings decreases. Thus, if a gap in the middle of the distribution is as large as one in the tail, it will be much more likely to indicate the presence of outliers. As a consequence, it is necessary to normalize each angular gap.

Let $V_i = F(W_i)$, for $1 \le i \le n$. The transformed random variable V_i is a uniform random variable on $(0, 1)$, and $\{V_{(1)}, V_{(2)}, \ldots, V_{(n)}\}$ are the order statistics of a sample of n independent random variables with that distribution. The spacings $\overline{D}_i$ between the observations in this sample are called the normalized spacings, $\overline{D}_i = V_{(i)} - V_{(i-1)}$, $1 \le i \le n+1$, where $V_{(0)} = 0$ and $V_{(n+1)} = 1$. The form of the distribution of $\overline{D}_{(n)}$, the length of the longest interval between n consecutive points chosen at random on the unit interval $(0, 1)$, is well known (David 1981):

$$\Pr(\overline{D}_{(n)} \le y) = \sum_{0 \le i < 1/y} (-1)^i \binom{n+1}{i} (1 - iy)^n. \qquad (7)$$

Given a significance level α, a cutoff value $D_{n;\alpha}$ can be obtained from (7) by setting $\Pr(\overline{D}_{(n)} \le D_{n;\alpha}) = 1 - \alpha$. If the original data distribution is elliptical and does not contain outliers, we expect each weighted gap $\overline{D}_i$ constructed using F_W (3) to be smaller than $D_{n;\alpha}$.

This cutoff value will be valid only if the reference direction were selected independently of the data. However, the proposed method depends on a direction chosen by applying a selection criterion to many candidates generated from the data. Consequently this data-dependent direction requires a modified cutoff value that accounts for these multiple choices, determined through a simulation study. Table 2 presents the resulting cutoff values for a significance level $\alpha = .05$, different values of the sample-space dimension (1 to 25), and the sample size (50 to 250). Each value has been estimated from 5,000 replications of the procedure, except for those in column 1, computed directly from (7). The values corresponding to $n/p < 5$ have been omitted from the table.

Table 2. Cutoff Values $D_{n,p;\alpha}$ for a Significance Level $\alpha = .05$, Different Sample-Space Dimensions p, and Sample Sizes n

	Dimension p								
n	1	2	3	4	5	10	15	20	25
50	.131	.142	.164	.172	.181	.221	—	—	—
75	.094	.101	.116	.123	.130	.153	.181	—	—
100	.074	.080	.089	.094	.099	.117	.136	.155	—
125	.061	.066	.073	.077	.080	.094	.107	.123	.141
150	.052	.055	.061	.066	.068	.079	.089	.097	.112
175	.046	.049	.054	.057	.059	.067	.075	.085	.098
200	.041	.044	.047	.050	.051	.058	.065	.074	.082
225	.037	.039	.043	.045	.046	.052	.058	.065	.072
250	.034	.036	.039	.040	.041	.046	.051	.058	.065

We have checked that an empirical rule to derive cutoff values for arbitrary values of p and n (assuming $n/p \ge 5$) that fits very closely the preceding values is given by $D_{n,p;.05} = D_{n,1;.05}\, p^{-2}$, and the values for $D_{n,1;\alpha} \equiv D_{n;\alpha}$ can be obtained from (7). This expression also provides reasonable approximations for a significance level $\alpha = .01$.

Consider again the synthetic dataset introduced in Section 1, whose Q-Q plot is shown in Figure 4. The value of the largest normalized gap is .263, and from Table 2 the cutoff value is $D_{200,20;.05} = .074$. The outliers would be clearly identified using the preceding test.

To detect the presence of several clusters of outliers, one might iterate the proposed procedure until either the value of the largest gap is no longer significant or the number of remaining observations becomes smaller than $\lfloor (n+p+1)/2 \rfloor$. Using this dataset as an example, the procedure could be applied again after removing the last 20 observations (the ones separated by the largest gap). The largest gap is now .048, which is less than the cutoff value $D_{180,20;.05} = .085$, so the procedure stops after this point, correctly identifying all 20 outliers and mislabeling no observation.

The procedure described in this section has also been applied to the datasets used in the simulation study described in Section 1. For each set of parameter values, 1,000 datasets were generated, and a single pass of the procedure was applied to them. In all cases the procedure had 100% success in correctly identifying all the outliers, except for the following three sets of parameter values: (1) $p = 5, \epsilon = .05, k = 6.65$ (96% success); (2) $p = 5, \epsilon = .1, k = 6.65$ (99% success); (3) $p = 10, \epsilon = .05, k = 8.56$ (99% success). The experiment was repeated using larger samples, composed of $n = 50p$ observations, instead of using $n = 10p$ as in Table 1. The results were again 100% successful except for the single case $p = 5, \epsilon = .05, k = 6.65$ (95% success). As a consequence, the method seems quite efficient in the detection of concentrated outliers, as expected from the motivation presented previously. We will try to justify this behavior in a more formal manner through the theoretical analysis conducted in the following section.

3. ANALYSIS OF THE PROCEDURE

The basic requirement for an outlier-detection method is that it should be able to detect outliers for any reasonable

contamination pattern. A high breakdown point ensures this property for very large distances to the contamination but not necessarily to moderate distances. Section 1 illustrated some difficulties in robust methods in identifying particular classes of contaminations at moderate distance. As a consequence, it would be of interest to study the behavior of the proposed procedure in those cases—that is, in the presence of concentrated contaminations located at a finite distance from the sample center.

In this section, we show that, for this particular case, the proposed procedure has properties that are complementary to those of distance-based methods, such as MCD or SDE. In particular, we will see that the procedure works better as the sample-space dimension or the concentration in the outliers increases.

We consider again a sample from a contaminated normal distribution—that is, $n(1-\epsilon)$ observations from an $N(0, \mathbf{I})$ distribution in $\Re^p$ (the regular observations), contaminated with $n\epsilon$ observations from an $N(k\mathbf{e}_1, \lambda^2\mathbf{I})$ distribution, where $\mathbf{e}_1 = (1, 0, \ldots, 0)^T$ and $\epsilon < 1/2$ denotes the proportion of outliers in the sample. Note that, because the procedure is affine equivariant, the basic assumption made (apart from using a mixture of normals as the reference distribution) is that the shape of the covariance matrix is the same both for the outliers and the uncontaminated sample.

The observed gap will depend on the characteristics of the outlier distribution—dimension of the sample space p, contamination level ϵ, distance to the reference observations k, and concentration λ. We analyze in particular the dependence of the spacings with respect to k, ϵ, and p; in some of these cases the properties of the method are markedly different from those of methods based on distances.

The analytic study will focus on a particular quantity related to the spacings in the data—an angle θ such that for some $0 < \beta < 1, \Pr(\Theta \geq \theta) \geq \beta$, where Θ denotes the angle between the extreme observations from the regular observations and the outliers. A geometric illustration of the meaning of this angle is provided in Figure 5. A realization of the random variable Θ would correspond to the gap between the two groups of observations, as long as the reference direction is chosen to be $\mathbf{e}_1$. In this case, the angle θ would provide a (probabilistic) bound on the size of this gap. The procedure presented in Section 2 provides reference directions for finite samples that are very close to $\mathbf{e}_1$ (see the results in Sec. 4).

Let z_β denote the $(1+\sqrt[n]{\beta})/2$ quantile of a standard univariate normal distribution. To simplify the derivation of an expression for the bound θ for the model introduced previously, we introduce the following "regularity" condition: We require that the parameters ϵ, λ, and k satisfy $z_\beta < \min(\epsilon, (1-\epsilon)/\lambda)k$. If this condition is not satisfied, then with probability larger than $1-\beta$ it is possible to find observations forming arbitrary angles with the reference direction (because the center of the data lies within one of the isoprobability curves), and the projections onto the unit hypersphere of the two samples may overlap.

We now derive an expression for θ. The first step in the application of the outlier-detection procedure presented in Section 2 is to introduce an affine transformation to ensure that the transformed observations have zero mean and an identity covariance matrix. After this transformation, for the preceding model, we have two groups of observations, one of them composed of $n(1-\epsilon)$ observations from an $N(-\epsilon k\mathbf{S}^{-1/2}\mathbf{e}_1, \mathbf{S}^{-1})$ distribution and another group of $n\epsilon$ observations from an $N((1-\epsilon)k\mathbf{S}^{-1/2}\mathbf{e}_1, \lambda^2\mathbf{S}^{-1})$ distribution, where

$$\mathbf{S}^{-1} = \tfrac{1}{\gamma_1}\left(\mathbf{I} - \gamma_2\mathbf{e}_1\mathbf{e}_1^T\right)$$

$$\gamma_1 = 1 - \epsilon(1-\lambda^2)$$

$$\gamma_2 = \frac{k^2\epsilon(1-\epsilon)}{\gamma_1 + k^2\epsilon(1-\epsilon)}$$

$$\mathbf{S}^{-1/2}\mathbf{e}_1 = \sqrt{\frac{1-\gamma_2}{\gamma_1}}\mathbf{e}_1. \tag{8}$$

Due to the (axial) symmetry of the problem, we need to analyze the properties of the angles for only the projections of the observations onto a plane defined by $\mathbf{e}_1$ and any direction orthogonal to it. For these projections the observations will follow the same distributions described previously but now restricted to $\Re^2$ (this is the case illustrated in Fig. 5). As mentioned previously, the reference direction affects the size of the observed gap. We consider the case in which the reference direction is the direction to the center of the outliers, $\mathbf{e}_1$. For this case, the angle θ (see Fig. 5) can be obtained as $\theta = \pi - \theta_1 - \theta_2$ from the pair of angles θ_1 and θ_2, defined as those such that an observation from each of the two samples forms an angle with $\mathbf{e}_1$ that is smaller than these angles with probability equal to $\sqrt[n]{\beta}$.

We will make use of the fact that the angle with the x axis of the tangent to an ellipse of the form $ax^2 + by^2 = c$ from a point $(r, 0)$ is given by

$$\tan\varphi = \frac{1}{\sqrt{\frac{br^2}{c} - \frac{b}{a}}}. \tag{9}$$

From the equations of the isoprobability lines for each of the samples corresponding to a probability level equal to $\sqrt[n]{\beta}$,

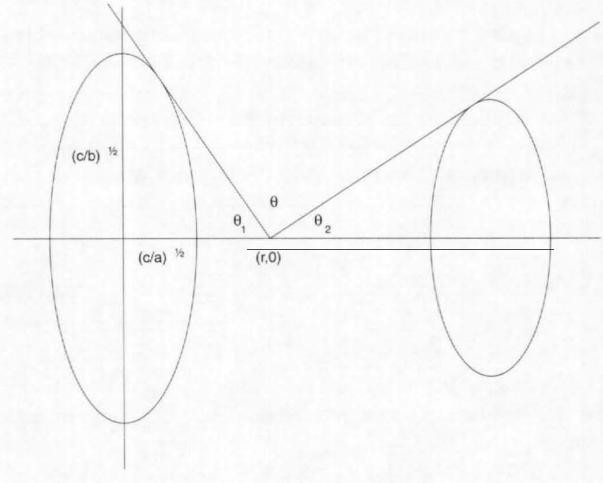

Figure 5. Angles Between Observations for the Mixture-of-Two Normals Case.

derived from (8), and (9), the preceding angles are given by

$$\tan\theta_1 = \frac{z_\beta}{\sqrt{\gamma_1}}\sqrt{\frac{\gamma_1/k^2+\epsilon(1-\epsilon)}{\epsilon^2-z_\beta^2/k^2}}$$

$$\tan\theta_2 = \frac{z_\beta}{\sqrt{\gamma_1}}\sqrt{\frac{\gamma_1/k^2+\epsilon(1-\epsilon)}{(1-\epsilon)^2/\lambda^2-z_\beta^2/k^2}}. \tag{10}$$

Using the trigonometric equivalence

$$\tan\theta = \tan(\pi-\theta_1-\theta_2) = -\frac{\tan\theta_1+\tan\theta_2}{1-\tan\theta_1\tan\theta_2}$$

and (10), we obtain the desired expression,

$$\tan\theta = z_\beta\sqrt{\Delta_0}\frac{\sqrt{\Delta_1}+\sqrt{\Delta_2}}{z_\beta^2\Delta_0^2-\sqrt{\Delta_1\Delta_2}}$$

$$\Delta_0 = \frac{1}{k^2}+\frac{\epsilon(1-\epsilon)}{\gamma_1}$$

$$\Delta_1 = \epsilon^2-\frac{z_\beta^2}{k^2}$$

$$\Delta_2 = \frac{(1-\epsilon)^2}{\lambda^2}-\frac{z_\beta^2}{k^2}. \tag{11}$$

This expression relates a bound on a particularly significant spacing to the characteristics of the observations and the contamination. Although it is fairly complex, some conclusions can be reached from it.

1. If limits are taken in (11) as $k\to\infty$—that is, when contaminations arbitrarily removed from the original sample are considered—it follows that

$$\tan\theta \to z_\beta\sqrt{\frac{\epsilon(1-\epsilon)}{\gamma_1}}\frac{\epsilon+(1-\epsilon)/\lambda}{z_\beta^2(\epsilon(1-\epsilon)/\gamma_1)-(\epsilon(1-\epsilon)/\lambda)}$$

$$= z_\beta\sqrt{\frac{\gamma_1}{\epsilon(1-\epsilon)}}\frac{1-\epsilon+\lambda\epsilon}{z_\beta^2\lambda-\gamma_1}.$$

From this expression, as $k\to\infty$ the gap becomes larger than $\pi/2$ (a value that can be trivially identified in a Q-Q plot, for example) whenever $z_\beta^2\lambda-\gamma_1<0$. This is equivalent to $\lambda\geq 1$ and $\epsilon\geq(z_\beta^2\lambda-1)/(\lambda^2-1)$, or $\lambda\leq 1$ and $\epsilon\leq(1-z_\beta^2\lambda)/(1-\lambda^2)$. The first condition holds for all sufficiently large values of λ, while the second one always holds for λ sufficiently small.

2. Consider now the behavior of θ with respect to the contamination level ϵ for the particular case of a concentrated contamination, $\lambda\to 0$. From (11), as $\lambda\to 0$ it follows that

$$\tan\theta\to\psi(\epsilon)\equiv-\sqrt{\frac{z_\beta^2\epsilon+z_\beta^2/k^2}{\epsilon^2-z_\beta^2/k^2}},$$

$$\psi'(\epsilon) = \frac{z_\beta^2}{2}\frac{z_\beta^2/k^2+2\epsilon/k^2+\epsilon^2}{(\epsilon^2-z_\beta^2/k^2)^2}\sqrt{\frac{\epsilon^2-z_\beta^2/k^2}{z_\beta^2\epsilon+z_\beta^2/k^2}}.$$

This derivative is positive for any values of z_β and k satisfying the regularity condition $z_\beta<\epsilon k$. Thus, for concentrated contaminations with sufficiently small values of λ, the gap between the two groups of observations increases with the contamination level ϵ. This behavior differs from that for most distance-based methods.

The analysis of the behavior of the gaps with respect to p cannot be based on (11) because the size of the gap does not depend on p for the model presented in this section and the reference direction we have considered. If p is increased, but the remaining parameters in the contamination model do not change, the values of $W_{(i)}$ and $W_{(i+1)}$ (using the notation from Sec. 2.2) are not affected. The only impact of these changes appears through F_W in the value of $D_i = F_W(W_{(i+1)}) - F_W(W_{(i)})$. We now show that this value increases with p (for fixed $W_{(i)}$ and $W_{(i+1)}$) if $W_{(i)}<\pi/2<W_{(i+1)}$.

To simplify the notation, define

$$J(a,b;p)\equiv\int_a^b \sin^{p-2}t\,dt, \qquad a,b\in[0,\pi].$$

We have $J(a,b;p)\geq 0$, $J(0,a;p)=J(\pi-a,\pi;p)$, and, from (2),

$$D_i(p) = \frac{J(W_{(i)},W_{(i+1)};p)}{J(0,\pi;p)}.$$

We wish to study the sign of $D_i(p+1)-D_i(p)$; equivalently, we may analyze the sign of

$$\begin{aligned}\Delta &= \big(D_i(p+1)-D_i(p)\big)J(0,\pi;p+1)J(0,\pi;p)\\ &= J(0,\pi;p)J\big(W_{(i)},W_{(i+1)};p+1\big)\\ &\quad -J(0,\pi;p+1)J(W_{(i)},W_{(i+1)};p).\end{aligned}$$

From the mean value theorem, for any $0<a<\pi/2$,

$$J(0,a;p+1) = \sin\varphi J(0,a;p), \qquad \varphi\in(0,a),$$

$$J(0,\pi/2;p+1) = \eta_p J(0,\pi/2;p), \qquad \eta_p>\sin\varphi.$$

Let $a\equiv W_{(i)}$ and $b\equiv W_{(i+1)}$, and assume that $0<a<\pi/2<b<\pi$ holds; then

$$\begin{aligned}\Delta &= J(0,\pi;p)J(a,b;p+1)-J(0,\pi;p+1)J(a,b;p)\\ &= J(0,\pi;p)\big(2J(0,\pi/2;p+1)-J(0,a;p+1)\\ &\quad -J(0,\pi-b;p+1)\big)-J(0,\pi;p+1)J(a,b;p)\\ &= J(0,\pi;p)\big(2\eta_pJ(0,\pi/2;p)-\sin\varphi_aJ(0,a;p)\\ &\quad -\sin\varphi_bJ(0,\pi-b;p)\big)-\eta_pJ(0,\pi;p)\big(2J(0,\pi/2;p)\\ &\quad -J(0,a;p)-J(0,\pi-b;p)\big)\\ &= J(0,\pi;p)\big((\eta_p-\sin\varphi_a)J(0,a;p)\\ &\quad +(\eta_p-\sin\varphi_b)J(0,\pi-b;p)\big)>0,\end{aligned}$$

and we have the desired bound.

From (10), the preceding condition will be satisfied whenever $\theta_1,\theta_2<\pi/2$, but this will hold as long as $z_\beta<\min(\epsilon,(1-\epsilon)/\lambda)k$. These conditions are sufficient, but not necessary and are trivially satisfied if $\lambda=0$ and $z_\beta<\epsilon k$. Whenever these conditions are satisfied, the distortion associated with the presence of outliers increases with p, and as a consequence the probability of observing a given normalized gap between the reference observations and the outliers increases with the dimension of the problem.

This behavior of the method makes the proposed procedure particularly useful for those cases in which either p or ϵ are large, corresponding to situations in which the procedures based on high breakdown-point estimators are less effective.

4. EXAMPLES

In this section we describe the practical behavior of the proposed procedure through several examples, most of them taken from the literature. Our goal is to illustrate the way the procedure works in different cases, based both on synthetic and real data. Although many of the test cases considered have already been successfully analyzed using different robust procedures, these examples are intended to show how the proposed method is able to handle a wide range of contamination patterns.

We have analyzed the dataset MULCROSS, available in STATLIB jointly with the code MULTOUT. This dataset has 200 observations in dimension 10, with 150 observations generated from a normal distribution, and 50 outliers from a different normal distribution, displaced with respect to the initial observations. The outliers form a single cluster, with dispersion similar to that of the main set of 150 observations. When the proposed procedure is applied, the maximum gap appears between the ordered observations 150 and 151; it is the only one lying above the relevant signification levels and separates the regular observations from the outliers. The corresponding Q-Q plot is shown in Figure 6, where the two groups of observations are readily apparent; one of them contains the 50 outliers, and the other corresponds to the remaining 150 observations. The values for the gap statistics in this example are $\overline{D}_{(200)} = .387$ and, from Table 2, $D_{200,\,10;\,.05} = .058$.

The "wood gravity" dataset (Rousseeuw and Leroy 1987), a set of 20 observations in dimension 5 that has been studied in several works related to multivariate outlier detection, has also been analyzed. Previous studies have identified four outliers, corresponding to observations 4, 6, 8, and 19, from the two- and three-dimensional scatterplots. Nevertheless, some identification methods based on the MVE [MULTOUT; see also the comments of Cook and Hawkins (1990) to Rousseeuw and van Zomeren (1990)] and those based on the SDE may fail to identify these outliers. The procedure described in this article generates the Q-Q plot shown in Figure 7(a), where the outliers are readily apparent. The normalized gap takes the value .490, above the cutoff value $D_{20,\,5;\,.05} = .373$.

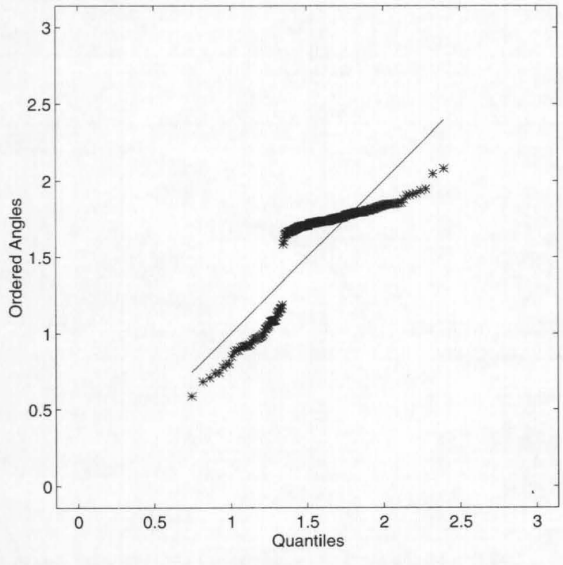

Figure 6. Q-Q Plot for the MULCROSS Dataset.

In a slightly different setting, the procedure was also applied to the well-known Anderson iris data (Anderson 1935; Fisher 1936). In our case we have used only those observations corresponding to varieties virginica and versicolor to obtain a sample composed of 100 observations in dimension 4. Although this is no longer an outlier detection problem, because $\epsilon = .5$, it serves to illustrate the possibilities of the proposed method. Figure 7(b) shows the results from the procedure and the large gap between the groups corresponding to each variety. The normalized gap is .208, much larger than the cutoff $D_{100,\,4;\,.05} = .094$. It might be difficult for a procedure based on distances to identify both groups correctly.

We consider next the situation in which the outliers may form several clusters. In these situations, MCD- and SD-based methods tend to perform better than with just one cluster. We wish to show that the proposed procedure (with very minor modifications) is also able to perform reasonably well.

Consider first a synthetic example, corresponding to a sample of 100 observations in dimension 10, with 80 observations from an $N(0, \mathbf{I})$ distribution, 10 observations from an $N(k_1\mathbf{e}_1, \lambda^2\mathbf{I})$, and the last 10 observations were generated from an $N(k_2\mathbf{e}_2, \lambda^2\mathbf{I})$, where $\mathbf{e}_1$ and $\mathbf{e}_2$ denote the first two unit vectors in $\Re^{10}$, $k_1 = 7.5$, $k_2 = 10$, and $\lambda = .1$. Figure 8(a) shows the scatterplot corresponding to the projections of the dataset onto the first two coordinate directions, clearly revealing the two clusters of outliers.

After the proposed procedure has been applied once, the resulting Q-Q plot is the one shown in Figure 8(b). Note that the reference direction $\mathbf{u}_0$ is very close to $\mathbf{e}_2$. The maximum gap is .247 and the cutoff value obtained from Table 2 is $D_{100,\,10;\,.05} = .118$. As a consequence the last 10 observations in the sample would be labeled as outliers.

This first application of the algorithm has not detected all the outliers. To complete the process, we iterate the procedure, after removing the suspected outliers, until the maximum gap is no longer significant. If the proposed procedure is applied again to observations 1–90 (after removing the last 10), the resulting Q-Q plot is shown in Figure 8(c). The reference direction is very close to $\mathbf{e}_1$, the maximum gap is .291, and the cutoff value is $D_{90,\,10;\,.05} = .133$. As a consequence, observations 81–90 are also labeled as outliers. After removing them, the procedure is applied again to the remaining 80 observations (the first ones), providing the Q-Q plot shown in Figure 8(d). In this case, the maximum gap is .088 and the cutoff value is $D_{80,\,10;\,.05} = .145$, no additional outliers are detected, and the procedure ends successfully. Note that the success of the procedure depends on the ability to identify as reference directions $\mathbf{u}_0$ the directions to the outliers. The lack of fit apparent in Figure 8(d) is due to the fact that the reference direction has been chosen to maximize this lack of fit.

Finally, we analyze a dataset presented by Campbell (1989), obtained in the process of locating bush-fire scars, and composed of 38 observations in dimension 5. This dataset was studied by Maronna and Yohai (1995) regarding the presence of outlying observations. It should be noted that, as opposed to the preceding example, these data correspond to a real situation, and as a consequence the evaluation of the results from

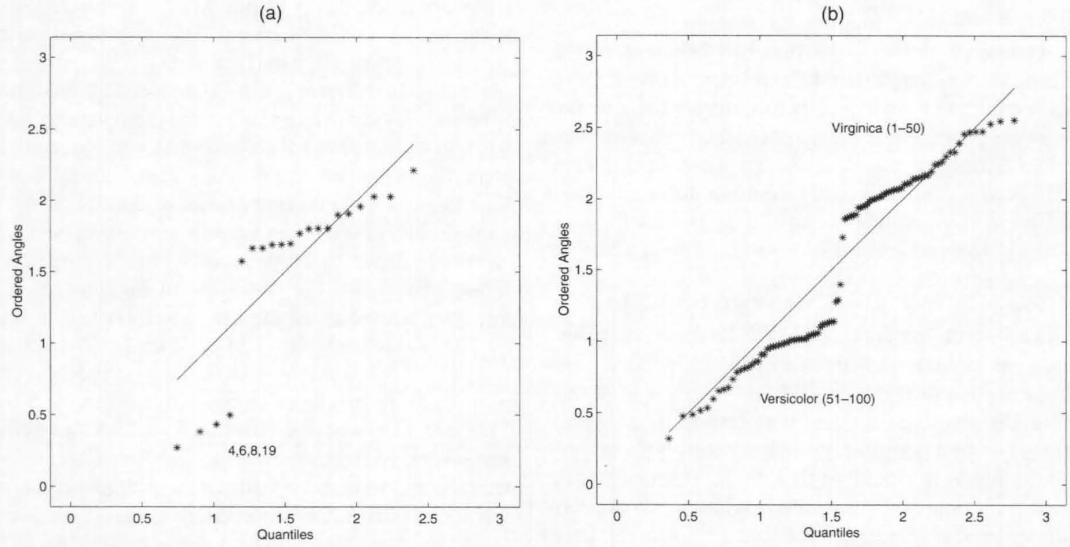

Figure 7. Q-Q Plots for (a) the Wood Gravity Dataset, (b) the Iris Data.

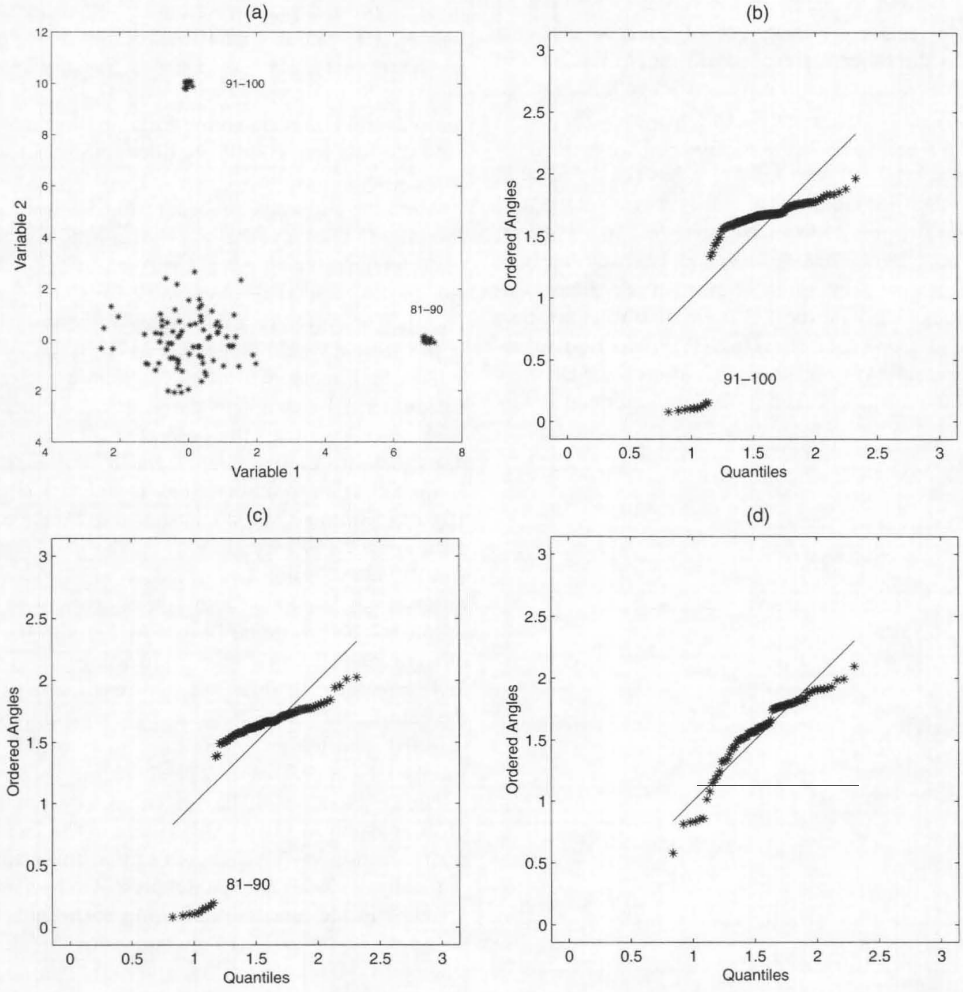

Figure 8. (a) Scatterplot for the Synthetic Example With Two Clusters, (b) First Q-Q Plot for the Synthetic Example, (c) Q-Q Plot After Removing Observations 91–100, (d) Q-Q Plot After Removing Observations 81–100.

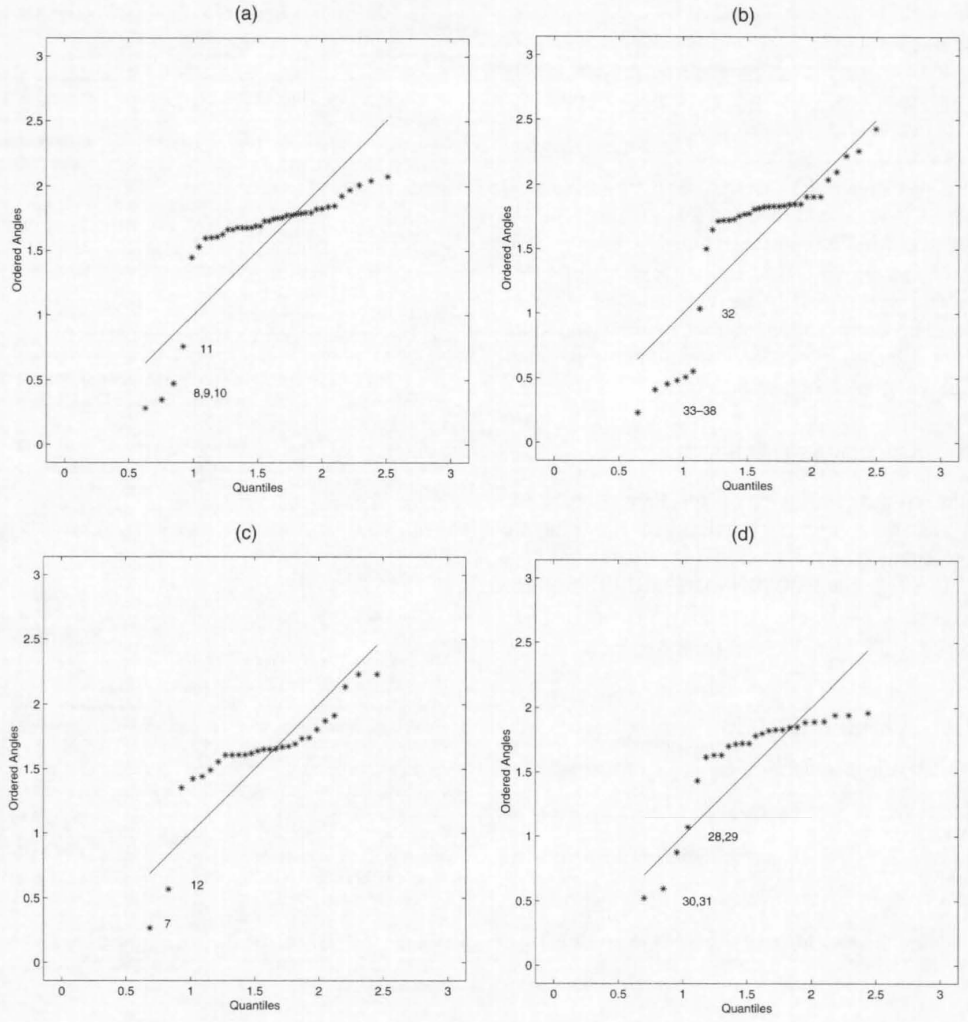

Figure 9. Q-Q Plots for the Bush-Fire Scar Data: (a) First Q-Q Plot, (b) Q-Q Plot After Removing Observations 8–11, (c) Q-Q Plot After Removing 10 Observations, (d) Q-Q Plot After Removing 12 Observations.

the identification procedure is not as straightforward as in the synthetic case. Maronna and Yohai (1995), using the SDE, found that observations 8 and 9 are the ones furthest removed from the sample center, followed by observations 32 to 38. Different results were obtained by these authors using other estimators.

The result of the application of the proposed procedure yields the Q-Q plot presented in Figure 9(a). The largest gap has a value of .355 and separates observations 8, 9, 10, and 11 from the rest. The cutoff value from Table 2 is $D_{38,.5;.05} = .226$, and as a consequence these observations are labeled as outliers. Following the same approach as in the preceding case, we again apply the procedure to the remaining 34 observations. The corresponding Q-Q plot is given in Figure 9(b). Now the largest gap is .297, and the cutoff value is $D_{34,.5;.05} = .247$. Observations 33–38 are accordingly labeled as outliers, and the procedure is repeated on the remaining 28 observations. The new Q-Q plot is shown in Figure 9(c). The largest gap is .323 and the cutoff value is $D_{28,.5;.05} = .296$. Observations 7 and 12 are labeled as outliers. Finally, for the remaining 26 observations the resulting Q-Q plot is shown in Figure 9(d). Now the largest gap is .230 and separates observations 28–31 from the rest; the cutoff value is $D_{26,.5;.05} = .315$. As a consequence, no additional observations would be labeled as outliers. Nevertheless, the lack of fit shown in the Q-Q plot, Figure 9(d), might provoke some doubts on the nature of observations 28–31. In fact, FAST-MCD labels these last four observations as outliers, while both Maronna and Yohai (1995) and Rocke and Woodruff (1996) did not consider them to be anomalous.

5. CONCLUSIONS

This work attempts to illustrate the difficulties faced by many robust procedures, and in particular those based on the use of robust Mahalanobis distances, for the detection of concentrated contaminations. Following the remark by Gnanadesikan and Kettenring (1972), cited by Barnett and Lewis (1994), "The complexity of the multivariate case

suggests that it would be fruitless to search for a truly omnibus outlier-protection procedure. A more reasonable approach seems to be to tailor detection procedures to protect against specific types of situations," a simple procedure is proposed to detect this contamination pattern, based on the analysis of the gaps associated with certain univariate projections of the observations. As opposed to other robust procedures, its behavior improves with the dimension of the problem and with the proportion of outliers in the sample.

The procedure can be considered as an exploratory tool, simple to use, and very effective on concentrated contamination patterns. The combination of this method and other traditional outlier-detection procedures should allow the identification of highly complex outlier patterns.

ACKNOWLEDGMENTS

We thank the editors and referees for their suggestions and comments that have been very helpful in clarifying the contents and presentation of the article. This research has been supported by CICYT grants BEC2000-0167 and PB98-0728.

[Received August 1999. Revised September 2000.]

REFERENCES

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

Anderson, E. (1935), "The Irises of the Gaspe Peninsula," *Bulletin of the American Iris Society*, 59, 2–5.

Barnett, V., and Lewis, T. (1994), *Outliers in Statistical Data* (3rd ed.), New York: Wiley.

Beckman, R. J., and Cook, R. D. (1983), "Outlier. . . s," *Technometrics*, 25, 119–163.

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

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

David, H. A. (1981), *Order Statistics*, New York: Wiley.

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

Eaton, M. L. (1983), "Isotropic Distributions," in *Encyclopedia of Statistical Sciences* (Vol. 4), eds. S. Kotz, N. L. Johnson and C. B. Read, New York: Wiley, pp. 265–267.

Fisher, R. A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7, 179–188. (Reprinted in Fisher, R. A. (1950), *Contributions to Mathematical Statistics*, New York: Wiley.)

———(1938), "The Statistical Utilisation of Multiple Measurements," *Annals of Eugenics*, 8, 376–386.

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

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

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

Mardia, K. V. (1977), "Mahalanobis Distances and Angles," in *Multivariate Analysis IV*, ed. P. R. Krishnaiah, Amsterdam: North-Holland, pp. 495–511.

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.

Pyke, R. (1965), "Spacings," *Journal of the Royal Statistical Society*, Ser. B, 27, 395–449.

Rocke, D. M., and Woodruff, D. L. (1996), "Identification of Outliers in Multivariate Data," *Journal of the American Statistical Association*, 91, 1047–1061.

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

Rousseeuw, P. J., and Leroy, A. (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.

Stahel, W. A. (1981), "Breakdown of Covariance Estimators," Research Report 31, Fachgruppe fur Statistik, Eidgenössische Technische Hochschule, Zurich.