



ELSEVIER

Available at

www.ElsevierMathematics.com

POWERED BY SCIENCE @ DIRECT®

Computational Statistics & Data Analysis 44 (2004) 625–638

COMPUTATIONAL
STATISTICS
& DATA ANALYSIS

www.elsevier.com/locate/csda

Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator

Johanna Hardin^{a,*}, David M. Rocke^b

^aDepartment of Mathematics, Pomona College, 610 N. College Ave., Claremont, CA 91711, USA

^bCenter for Image Processing and Integrated Computing, University of California at Davis, USA

Received 1 January 2002; received in revised form 1 August 2002

Abstract

Mahalanobis-type distances in which the shape matrix is derived from a consistent high-breakdown robust multivariate location and scale estimator can be used to find outlying points. Hardin and Rocke (<http://www.cipic.ucdavis.edu/~dmrocke/preprints.html>) developed a new method for identifying outliers in a one-cluster setting using an F distribution. We extend the method to the multiple cluster case which gives a robust clustering method in conjunction with an outlier identification method. We provide results of the F distribution method for multiple clusters which have different sizes and shapes.

© 2002 Elsevier B.V. All rights reserved.

Keywords: Minimum covariance determinant; Robust clustering; Outlier detection

1. Introduction

Good methods for identifying clusters of well-separated uncontaminated groups of data have been available for many years. The process of clustering becomes more difficult, however, when the data include points which do not belong to any cluster. (Everitt, 1993) Outlying points can skew the shape estimates of the clusters or distort optimization criterion which in turn can mask separation between clusters. A group of diabolically placed points can completely change the cluster arrangement under many clustering algorithms.

* Corresponding author. Fax: +909-607-8717.

E-mail addresses: jo.hardin@pomona.edu (J. Hardin), dmrocke@ucdavis.edu (D.M. Rocke).

Robust clustering methods often give an accurate depiction of the underlying data format, but even so, they do not usually identify particular outlying points which may be of interest or importance in their own right. Along with robust clustering methods, it is important to have a complimentary outlier identification method.

Various method for detecting outliers in the one cluster multiple dimensional setting have been studied (Atkinson, 1994; Barnett and Lewis, 1994; Gnanadesikan and Kettenring, 1972; Hadi, 1992; Hawkins, 1980; Maronna and Yohai, 1995; Penny, 1995; Roche and Woodruff, 1996; Rousseeuw and VanZomeren, 1990). One way to identify possible multivariate outliers is to calculate a distance from each point to a “center” of the data. An outlier would then be a point with a distance larger than some predetermined value. A conventional measurement of quadratic distance from a point X to a location Y given a shape S , in the multivariate setting is

$$d_S^2(X, Y) = (X - Y)^T S^{-1} (X - Y).$$

This quadratic form is often called the Mahalanobis squared distance (MSD).

In the clustering context, an outlier can be thought of as a point with a large MSD from the center of each and every one of the clusters. After a robust clustering algorithm is applied to the data, each of the clusters can be thought of as coming from a unique population, and outlier identification methods can be used individually on each cluster.

For data that come from one population, the distribution of the MSD with both the true location and shape parameters and the conventional location and shape parameters is well known (Gnanadesikan and Kettenring, 1972). However, the conventional location and shape parameters are not robust to outliers, and the distributional fit breaks down when robust measures of location and shape are used in the MSD (Rousseeuw and VanZomeren, 1990). Hardin and Roche (2002) developed a distributional fit to Mahalanobis distances which uses a robust shape and location estimate, namely the minimum covariance determinant (MCD).

Given n data points, the MCD of those data is the mean and covariance matrix based on the sample of size h ($h \leq n$) that minimizes the determinant of the covariance matrix (Rousseeuw, 1984).

$$MCD = (\bar{X}_J^*, S_J^*),$$

where

$$J = \{ \text{set of } h \text{ points} : |S_J^*| \leq |S_K^*| \forall \text{ sets } K \text{ s.t. } \#|K| = h \}$$

where $\#|\omega|$ defines the number of elements in set ω

$$\bar{X}_J^* = \frac{1}{h} \sum_{i \in J} x_i$$

$$S_J^* = \frac{1}{h} \sum_{i \in J} (x_i - \bar{X}_J^*)(x_i - \bar{X}_J^*)^t$$

The value h can be thought of as the minimum number of points which must not be outlying. The MCD has its highest possible breakdown at $h = \lfloor (n + p + 1)/2 \rfloor$ where $\lfloor \cdot \rfloor$ is the greatest integer function (Rousseeuw and Leroy, 1987; Lopuhaä and

Rousseeuw, 1991). We will use $h = \lfloor (n + p + 1)/2 \rfloor$ in our calculations and refer to a sample of size h as a half sample. The MCD is computed from the “closest” half sample, and therefore, the outlying points will not skew the MCD location or shape estimates. The concept of the MCD can be modified easily to fit the multiple cluster setting. With a good initialization and a known number of clusters, g , the MCD can be found separately for each of the clusters. The size of each cluster is determined by the number of points which are closer to that cluster’s center than to any other cluster center. The sizes of the clusters and the MCD samples will be n_i and $h_i = \lfloor (n_i + p + 1)/2 \rfloor$ $i = 1, \dots, g$, respectively.

Using the MCD estimates in the MSD leads to robust distances from each of the cluster centers for each of the data points. A datum which is outlying will have a large robust distance from each of the cluster centers. However, not every data set will give rise to an obvious separation between extreme points which belong to the data set (i.e. are not outliers) and those which do not (i.e. are outliers.) In order to distinguish between these two types of extrema, outliers can be identified using the quantiles of an F distribution (Hardin and Rocke, 2002) on a cluster by cluster basis. An outlying point will be labeled as such only if it is outlying with respect to all of the clusters.

Note that throughout this paper, the number of groups is assumed to be fixed and known. This assumption is malleable in that additional small clusters will be ignored in the robust analysis, so they will not have an impact on estimating the g principal clusters. If an analyst is unsure of the number of populations present in the data, it is wise to try the analysis on a variety of values for the number of clusters.

We will apply cutoff values to multi-cluster, multivariate normal data given different values of g , n , p , and different arrangements and percentages of outlying points.

2. Clustering methods

Many algorithms exist for clustering various types of data. These algorithms use data, multivariate or univariate, as input, and as output the algorithm gives each datum a classification into a particular group. Some algorithms require that the number of clusters be pre-specified, and some algorithms allow for an unknown number of clusters. Those algorithms that do require as input a number of groups can be run multiple times with different values for the number of groups. The user can then choose the result that makes the most sense according to the problem or according to some statistical criterion. Finding an appropriate criterion may prove to be a hard problem. For the method we use, finding the correct number of groups for a particular data set is beyond the scope of this work. Our methods are applicable to data with no known structure or a priori metric, so we restrict our work to partitioning clustering methods, and we disregard hierarchical clustering methods for the time being. These methods can be used to find a best fit to a problem with a given number of groups.

2.1. Robust optimization clustering

The clustering method we used assumes only that the clusters are elliptical. (The outlier identification methods, however, are calibrated at the multivariate normal.) Since

the cluster shape is estimated from assigned points, it is required that $p + 1$ points be assigned to each of the main clusters. However, this method allows for unassigned points, so there could easily be allowed a cluster of points which is smaller than $p + 1$ included in the group of outlying points. A program, called `CLUSTER`, implementing this method is described in (Reiners and Woodruff, 2001).

3. Robust estimators in a cluster setting

Estimating cluster location and shape, even when the cluster membership is known, is a difficult problem if outliers are present (Roche and Woodruff, 1996). Since clustering methods need also to assign points to clusters as well as simultaneously estimate the cluster location and shape, the problem of cluster analysis in the presence of outliers is even more difficult.

3.1. Affine equivariant estimators

We are particularly interested in affine equivariant estimators of the data. A location estimator $\mathbf{y}_n \in \mathbb{R}^p$ is affine equivariant if and only if for any vector $\mathbf{v} \in \mathbb{R}^p$ and any nonsingular $p \times p$ matrix \mathbf{M} ,

$$\mathbf{y}_n(\mathbf{M}\mathbf{x} + \mathbf{v}) = \mathbf{M}\mathbf{y}_n(x) + \mathbf{v}.$$

A shape estimator $\mathbf{S}_n \in PDS(p)$ (the set of $p \times p$ positive definite symmetric matrices) is affine equivariant if and only if for any vector $\mathbf{v} \in \mathbb{R}^p$ and any nonsingular $p \times p$ matrix \mathbf{M} ,

$$\mathbf{S}_n(\mathbf{M}\mathbf{x} + \mathbf{v}) = \mathbf{M}\mathbf{S}_n(x)\mathbf{M}^T.$$

Stretching or rotating the data will not change an affine estimate of the data. If the location and shape estimates are affine equivariant, the MSDs are affine invariant, which means the shape of the data determines the distances between the points. The only other real alternative to affine estimates is to make a prior assumption about the correct distance measure. It is important to have affine equivariant estimators so that the measurement scale, location, and orientation can be ignored. Since MSDs are affine invariant, the properties and procedures that use the MSD can be calculated without loss of generality for standardized distributions. For the properties under normality, we can use $N(0, I)$.

3.2. Minimum covariance determinant

The MCD location and shape estimates are used as robust estimates of the location and shape of the clusters. Points that are outliers with respect to a particular cluster will not be involved in the location and shape calculations of that cluster, and points that are outliers with respect to all clusters will not be involved in the calculations of any clusters. The difference between the single population case and the multiple cluster case is that, in the latter, MCD samples need to be computed for each cluster. This

important difference leads to a need for a good robust starting point in the clustering situation.

3.3. Estimating the MCD

The exact MCD is impossible to find except in small samples or trivial cases. So, the algorithm used to estimate the MCD will be the estimator. The algorithm used in the multiple cluster case will be similar to the single population algorithm (Hawkins, 1999; Rousseeuw and VanDriessen, 1999) with the exception that the starting point of the algorithm will no longer be a random sub-sample of the data. The reason that it is important to have a non-random starting point for robust clustering is that random starts often give rise to shapes that are more representative of the entire data metric than the individual cluster metrics. Even with random samples of only $g \times (p + 1)$ points (where g is the number of clusters and p is the dimension), it is highly unlikely that a random starting point would partition the points into their g clusters respectively. From a starting point which reflects the entire data metric, it is difficult to separate the points into the correct g clusters.

For a robust start, we used the method of Reiners and Woodruff (2001) discussed previously. The outlier detection methods described in this paper are not dependent on the particular robust clustering algorithm `CLUSTER`. Any robust initialization would presumably give similar results. Random starts could be used if a condition was added to prevent the clusters from converging to the large dataset shape.

The core of the MCD estimation algorithm is as follows:

- Let H_1 be a subset of h points.
- Find \bar{X}_{H_1} and S_{H_1} . (If $\det(S_{H_1}) = 0$ then add points to the subset until $\det(S_{H_1}) > 0$.)
- Compute the distances $d_{S_{H_1}}^2(x_i, \bar{X}_{H_1}) = d_{H_1}^2(i)$ and sort them for some permutation π such that,

$$d_{H_1}^2(\pi(1)) \leq d_{H_1}^2(\pi(2)) \leq \dots \leq d_{H_1}^2(\pi(n)).$$

- $H_2 := \{\pi(1), \pi(2), \dots, \pi(h)\}$

For each dataset, the complete procedure for calculating the MCDs for each cluster is as follows.

1. Decide from how many populations the data came.
2. Use the program `CLUSTER` (or similar clustering algorithm) to find an initial robust clustering of the data.
3. From the initial clustering, calculate the mean and covariance of each of the clusters. (Each point belongs to at most one cluster, use the points belonging to a particular cluster to calculate its mean and covariance in the usual way.)
4. Calculate the MSD to each cluster, based on the most recently calculated mean and covariance, for each point in the dataset.
5. Assign each point to the cluster for which it has the smallest MSD, thereby determining a cluster size (n_j) for each cluster based on the number of points that are closest to that cluster.

6. For each cluster, choose a “half sample” ($h_j = \lfloor (n_j + p + 1)/2 \rfloor$) of those points with the smallest MSDs from step 4.
7. For each cluster, compute the mean and covariance of the current half sample.
8. Repeat steps 4–7 until the half sample no longer changes.
9. Report estimates.

For each cluster, the MCD sample will then be the final half sample (step 6). For each cluster (j), a robust distance like $d_{S_j^*}^2(x_i, \bar{X}_j^*)$, where S_j^* and \bar{X}_j^* are the MCD shape and location estimates for cluster j , is likely to detect outliers because outlying points will not affect the MCD shape and location estimates. For points x_i that are extreme, $d_{S_j^*}^2(x_i, \bar{X}_j^*)$ will be large for all j , and for points x_i that are not extreme, $d_{S_j^*}^2(x_i, \bar{X}_j^*)$ will not be large for a particular j .

We note that this algorithm may not converge for clusters with significant overlap. In this case, a model based likelihood algorithm should be used with the identified outliers classified as noise (or removed) and an iteration step that includes the model based algorithm each time instead of simply the partitioning given by the distances.

4. Distance distributions

4.1. Single population distance distributions

MSDs give a one-dimensional measure of how far a point is from a location with respect to a shape. Using MSDs we can find points that are unusually far away from a location and call those points outlying. Unfortunately, using robust estimates gives MSDs with unknown distributional properties.

In [Hardin and Roche \(2002\)](#), an approximate distributional result for MSDs based on location and shape derived from an MCD sample is given. Although the robust distances are asymptotically χ_p^2 , an F distribution fits the extreme points much more accurately across all sample sizes but especially in small samples. The distances based on the MCD metric can be expressed as

$$\frac{c(m-p+1)}{pm} d_{S_X^*}^2(X_i, \bar{X}^*) \sim F_{p, m-p+1}, \quad (4.1)$$

where \bar{X}^* and S_X^* are the location and shape estimates of the MCD sample, p is the dimension of the sample, and m and c are both parameters based on the size and shape of the MCD sample. The unknown parameters, m and c , can be estimated in three ways: using simulations, using an asymptotic result, or using an adjustment to the asymptotic result. The simulation results are the most accurate but also the most time consuming.

The parameter c can be estimated by

$$c = \frac{P(\chi_{p+2}^2 < \chi_{(p, h/n)}^2)}{h/n},$$

where χ_v^2 is a Chi-Square random variable with v degrees of freedom, and $\chi_{v,\varepsilon}^2$ is the ε quantile for a χ_v^2 random variable (Croux and Haesbroeck, 1999). We treat the extreme points as effectively independent of the MCD estimates, since they have little or no effect on the estimates themselves (Hardin and Roche, 2002).

For m there exists an asymptotic expression that is good in large samples and only moderately accurate in small samples (Croux and Haesbroeck, 1999). For small samples, an adjustment to the parameter is provided using a linear equation to estimate m more accurately. The following interpolation formula is used to modify the theoretical parameter value of the degrees of freedom (Hardin and Roche, 2002).

$$m_{\text{pred}} = m_{\text{asy}} \cdot e^{(0.725 - 0.00663p - 0.0780 \ln(n))}, \tag{4.2}$$

where m_{pred} is the predicted degrees of freedom from adjusting the asymptotic degrees of freedom, m_{asy} , given by Croux and Haesbroeck (1999). Croux and Haesbroeck used influence functions to determine an asymptotic expression for the variance elements of the MCD sample. Details are given in Appendix A. In this paper we use only the theoretical and adjusted parameter estimates in the interest of computing time. The empirical results for m hold for cluster sample sizes in the hundreds. For cluster sample sizes in the thousands, the asymptotic value of m should be used.

4.2. Multiple population distance distributions

Using the same arguments from the single population setting in the cluster setting, an F distribution can be used to approximate distances which are large with respect to a cluster location and shape. However, in this setting there are new factors to consider such as how many points are in each cluster and whether extreme points simply belong to another cluster.

In the single cluster case, the sample size of the dataset is used in the F cutoff calculation. Therefore, in the multiple cluster case, a sample size must be known or estimated for each cluster. The sample size also determines the “ h ” parameter used in the MCD calculation. The sizes from the final MCD iteration (step 5 in Section 3.3) will be used as the sizes of each of the clusters. The last MCD iteration also provides a robust location and shape for each cluster, these estimates are used to compute distances from each cluster. For a particular point, the distance from each cluster center will be found, and a point will be counted in the cluster for which its distance is the smallest.

$$\# \text{ in cluster } j = n_j = \sum_{i=1}^n I(d_{S_j^*}^2(X_i, \bar{X}_j^*) \leq d_{S_k^*}^2(X_i, \bar{X}_k^*) \forall k = 1, \dots, g \text{ groups}),$$

where \bar{X}_j^* and S_j^* are the location and shape estimates of the MCD sample, and m_j and c_j are both parameters based on the size and shape of the MCD sample from the j th cluster. With these constructs in mind, the distances of interest are those associated with the cluster to which a point is closest. Let \tilde{d}_i be the distance from point x_i to the closest cluster. An outlying point, x_i , will be one with \tilde{d}_i greater than some cutoff value.

Consider g groups of multivariate normal data in dimension p , and let $X_{ij} \sim N_p(\mu_j, \Sigma_j)$ where $i = \text{observation}$ and $j = \text{cluster}$. Let S_j be an estimate of Σ_j such that, $m_j S_j \sim \text{Wishart}_p(\Sigma_j, m_j)$. For the multiple cluster case,

$$c_j = \frac{P(\chi_{p+2}^2 < \chi_{p, h_j/n_j}^2)}{h_j/n_j}$$

and

$$\frac{c_j(m_j - p + 1)}{pm_j} d_{S_j^*}^2(X_i, \bar{X}_j^*) \sim F_{p, m_j - p + 1}.$$

Distributional cutoff results for distances based on the above type of clustered data with four different types of outlier arrangements: none, cluster, radial, and diffuse are given.

5. Results

Outliers can be identified as points with robust distances that exceed some cutoff value. The cutoffs are computed from distributional quantiles of χ^2 and F distributions. Because finding estimates for the parameters m and c by simulation is so computationally intensive, results are provided only for the F cutoff with theoretical and adjusted estimates for m and for the χ_p^2 cutoff. To compare the accuracy of these three estimates, we ran experiments on clean multivariate normal data and also on contaminated multivariate normal data. The contamination was done in three ways:

1. As a distinct cluster of outliers.
2. As radial outliers.
3. As diffuse outliers generated using a covariance structure equal to the entire dataset.

The Monte Carlo experiments were done at $p = 4, 7, 10$ with $g = 2, 3$ for the clean data and $g = 2$ for the contaminated data. Let n_g^p be the size of the groups simulated for a particular dimension (p) and number of groups (g). (In the interest of brevity, only results for the balanced case are reported. The unbalanced sample size results are equivalent to those reported for the equal sample sizes.)

$$n_2^4 = 300, 300 \text{ and } 200, 400$$

$$n_3^4 = 300, 300, 300 \text{ and } 200, 300, 400$$

$$n_2^7 = 500, 500 \text{ and } 300, 700$$

$$n_3^7 = 500, 500, 500 \text{ and } 300, 500, 700$$

$$n_2^{10} = 500, 500 \text{ and } 300, 700$$

$$n_3^{10} = 500, 500, 500 \text{ and } 300, 500, 700$$

Table 1
Clean data (2 and 3 clusters)

		2 Clusters Balanced	3 Clusters Balanced
1% F	4	0.63	0.64
P	7	0.81	0.82
	10	0.80	0.81
1% Adj F	4	0.91	0.90
P	7	0.99	0.99
	10	0.96	1.00
1% Chisq	4	21.66	21.57
P	7	17.94	17.88
	10	17.59	17.30

Each entry represents the percent of simulated data that was misclassified according to a specified cutoff and percentage. The first column reports data that come from two populations, and the second column reports data that come from three populations. Balanced refers to data that consist of equal sized clusters; results were similar for unbalanced clusters. These data were generated as clusters of multivariate normal data with no contamination. The analysis was done using the MCD estimates. Results were equivalent for cutoffs at the 5% and 0.1% level.

The cluster centers were separated by a distance of $2D$ where $D = \sqrt{\chi^2_{p,99}/p}$. The value $\sqrt{\chi^2_{p,99}}$ is the asymptotic radius of a 99% containment ellipsoid around a cluster of standard normal data. By separating the clusters at a distance of $2D$, we have clusters that do not overlap with high probability.

In a cluster situation the size of the entire dataset is known, but the size of the individual clusters is not known. Therefore, the clustering algorithm must be applied before outliers can be identified. In the clustering algorithm n_j , the number of points belonging to cluster j , is determined. From n_j we apply the theoretical formulas given for m_j and c_j , and we calculate the adjustment for m_j . The cutoff values for 5%, 1% and 0.1% rejection are calculated for F_{p,m_j-p+1} (with both variants of m) and χ^2_p .

5.1. Clean data

For each combination of size, dimension, and number of clusters 100 sets of independent data were simulated, and the number of points the cutoffs identified as outlying was counted. (The number of independent data sets is small due to the computational intensity of the initial clustering algorithm.)

The tabulated results in Table 1 report the percentage of data identified as outlying for each nominal level at $p = 4, 7, 10$ and cluster sizes as mentioned above. The F cutoff results are much closer to the target significance level than the χ^2_p cutoffs, and the adjusted cutoff is the most accurate. Also, because the cluster sizes are relatively large, the asymptotic F cutoff is only slightly conservative.

5.2. Contaminated data

Three methods were employed to generate outliers in each dataset, and both type I (i.e. “masked” data) and type II (i.e. “swamped” data) errors were measured. The type II error measures what percentage of true outlying points would be identified as non-outlying. The type I error measures the percentage of points which were generated under the null hypothesis that were not allocated to a cluster under the identification method.

The outliers which were generated as a separate cluster were placed along the axis that was used to separate the clean clusters at the same distance the clean clusters were separated. E.g. cluster 1 is distributed $N(0, I)$, cluster 2 was distributed $N(2D, I)$, and the cluster of outliers was distributed $N(4D, I)$. (Where $D = \sqrt{\chi_{p, .99}^2/p}$).

The radial outliers were generated with the same center as their respective clean clusters but with a covariance of 5 times the cluster’s covariance. We are interested in points that are truly outlying so that we can calculate the type II error of our method, but some points that are generated with a large covariance matrix may still be close to the cluster’s center. In dimension 4, 38% of points generated with $N(0, 5I)$ will be within the $\chi_{4, .99}^2$ bound of a $N(0, I)$ dataset. Using an acceptance-rejection algorithm, each outlying radial point that was generated was accepted if and only if it was outside the ellipse of clean data. The outliers were constructed to form an annulus around the clean data such that the average squared distance of an outlying point was $2\chi_{p, .99}^2$ away from the center of the clean data. By creating an annulus of outliers, we can correctly measure our type II error as those outlying points that get clustered.

The diffuse outliers were generated using the location and shape of the entire dataset. Again, we are only interested in points that are truly outlying, so we use the same acceptance-rejection algorithm to reject points that fall among the clusters of good data. In dimension 4, $n_2^4 = 300, 300$, 82.7% of the disperse points would have fallen within $\chi_{4, .99}^2$ of one of the two clusters.

The number of outliers simulated for each combination of size and dimension was constructed relative to the smallest cluster. For dimension 4, $n_2^4 = 200, 400$, if the number of outliers is 20% of the larger cluster (80 points), it would be difficult to distinguish between the good cluster of 200 points and the bad cluster of 80 points. So, the percentage of outliers of each pair of size and dimension is relative to the smaller of the two clusters. For each dimension, size, and outlier type we simulated outliers of size 20% of the smallest cluster. (In preliminary work different outlier percentages were tried and results were virtually identical.)

For each combination of size, dimension, and type of outliers, 100 sets of independent data were simulated, and the number of points the asymptotic F , the adjusted F , and the χ^2 cutoffs identified as outlying was counted.

5.2.1. Cluster outliers

Table 2 gives the results for both type I and type II errors for contaminated data. These simulations were done with two different cluster configurations, but the total sample size of the clean data stayed constant across dimension (600 total clean points

Table 2
Contaminated data (2 clusters)

	Type I Error				Type II Error				
	Cluster	Radial	Diffuse		Cluster	Radial	Diffuse		
1% F	4	0.42	0.37	0.37	1% F	4	0.17	1.73	0.32
P	7	0.57	0.47	0.51	P	7	0.00	0.06	0.00
	10	0.53	0.48	0.49		10	0.00	0.00	0.00
1% Adj F	4	0.60	0.54	0.52	1% Adj F	4	0.12	1.43	0.10
P	7	0.68	0.56	0.61	P	7	0.00	0.06	0.00
	10	0.64	0.57	0.60		10	0.00	0.00	0.00
1% Chisq	4	18.03	18.02	17.59	1% Chisq	4	0.00	0.12	0.00
P	7	15.02	14.72	14.53	P	7	0.00	0.00	0.00
	10	14.39	14.31	14.20		10	0.00	0.00	0.00

Each entry represents the percent of simulated data that was misclassified according to a specified cutoff and percentage. The first column of tables reports the type I error for the procedure, and the second column reports the type II error. Cluster refers to data that was contaminated by generating clusters of multivariate normal data with a cluster of contamination of size 20% of the smallest clean cluster. Radial refers to data that was contaminated by generating radial outliers of size 20% of the smallest clean cluster. Diffuse refers to data that was contaminated by generating diffuse outliers of size 20% of the smallest clean cluster. The analysis was done using the MCD estimates. Results were equivalent for cutoffs at the 5% and 0.1% level.

in dimension 4 and 1000 total clean points in dimensions 7 and 10.) The theoretical *F* cutoff is a little conservative, but it is an improvement over the χ^2 cutoff. Also, the adjusted *F* cutoff offers a slight improvement over the theoretical *F*. For type II error, there is virtually no clustering of the outlying points.

5.2.2. *Radial outliers*

Like the cluster outliers, the total sample size of the clean data stayed constant across dimension, and the outliers were always generated as 20% of the smallest cluster size. We see again that the theoretical *F* is an improvement over the χ^2 which is far too liberal at identifying outliers. Also, again the adjusted *F* cutoff offers a slight improvement over the theoretical *F* both in terms of type I and type II error. There is more type II error with the radial outliers, but the amount is negligible except for dimension 4 with an extreme cutoff (0.1% or less.)

5.2.3. *Diffuse outliers*

The results for diffuse outliers are similar to those for radial outliers. Once again, the theoretical *F* cutoff gives an improved but conservative bound for the type I error while the adjusted *F* gives the best results for type I error and improved results over the theoretical *F* for type II error. The type II error is seen as a problem again only in dimension 4 with extreme cutoffs (0.1% or less.)

5.3. Summary of results

The results for both the clean and contaminated data appear to depend strongly on p and only nominally on the sample or cluster sizes. Further simulations would probably show the dependence to be primarily on the sample size. We are encouraged to notice that the results are similar for balanced and unbalanced clusterings. The moderate type II error may be disconcerting, but it is important to remember that as the level of significance gets smaller, eventually the type II error will get larger. These results lead to the following recommendations:

1. With a robust initial starting value, the MCD estimator can be modified to find clusters of data and their corresponding MCD shape and location.
2. With the MCD estimates, the theoretical F cutoff is superior to the χ_p^2 cutoff, and the adjustment can be used to find m more accurately.
3. The adjusted F cutoff method works well under the null hypothesis that all the data come from g multivariate normal populations (where g is the number of clusters). When the data are contaminated the adjusted F cutoff method still performs adequately in measures of both the type I and type II error.

6. Conclusion

A method for finding outliers and a refined robust clustering method is given. The method relies on an initial robust clustering, and reasons are given for the importance of a robust method for finding clusters. In determining outliers within clusters, the F cutoff is clearly superior to the commonly used χ_p^2 cutoff.

One robust clustering method is given here, other robust methods could be applied in the same manner (e.g., Raftery's Mclust or McLachlan's EMMIX (Fraley and Raftery, 1999; McLachlan et al., 1999)). Additionally, M-estimates and S-estimates could be used in place of the MCD estimate.

Also, this method could be further explored by using more clusters, different outlier configurations, and the simulated values of m and c as parameters for the F cutoff.

Appendix A.

The parameter m , the degrees of freedom for the estimated Wishart distribution, is calculated using a series of steps. This estimation is due to Croux and Haesbroeck (1999).

$$\alpha = \frac{n - h}{n}, \tag{A.1}$$

where n is the sample size and $h = \lfloor (n + p + 1)/2 \rfloor$.

$$q_\alpha \text{ is such that } 1 - \alpha = P(\chi_p^2 \leq q_\alpha), \tag{A.2}$$

$$c_\alpha = \frac{1 - \alpha}{P(\chi_{p+2}^2 \leq q_\alpha)}, \tag{A.3}$$

$$c_2 = \frac{-P(\chi_{p+2}^2 \leq q_\alpha)}{2}, \tag{A.4}$$

$$c_3 = \frac{-P(\chi_{p+4}^2 \leq q_\alpha)}{2}, \tag{A.5}$$

$$c_4 = 3 \cdot c_3, \tag{A.6}$$

$$b_1 = \frac{c_\alpha(c_3 - c_4)}{1 - \alpha}, \tag{A.7}$$

$$b_2 = 0.5 + \frac{c_\alpha}{(1 - \alpha)} \left(c_3 - \frac{q_\alpha}{p} \left(c_2 + \frac{(1 - \alpha)}{2} \right) \right), \tag{A.8}$$

$$v_1 = (1 - \alpha)b_1^2 \left(\alpha \left(\frac{c_\alpha q_\alpha}{p} - 1 \right)^2 - 1 \right) - 2c_3c_\alpha^2(3(b_1 - pb_2)^2 + (p + 2)b_2(2b_1 - pb_2)), \tag{A.9}$$

$$v_2 = n(b_1(b_1 - pb_2)(1 - \alpha))^2c_\alpha^2, \tag{A.10}$$

$$v = \frac{v_1}{v_2}, \tag{A.11}$$

$$\hat{m} = \frac{2}{c_\alpha^2v}. \tag{A.12}$$

References

Atkinson, A., 1994. Fast very robust methods for the detection of multiple outliers. *J. Amer. Statist. Assoc.* 89, 1329–1339.

Barnett, V., Lewis, T., 1994. *Outliers in Statistical Data*. Wiley, New York.

Croux, C., Haesbroeck, G., 1999. Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *J. Multivariate Anal.* 71, 161–190.

Everitt, B.S., 1993. *Cluster Analysis*. Wiley, New York.

Fraley, C., Raftery, A., 1999. MCLUST: Software for model-based clustering and discriminant analysis. *J. Classification* 16, 297–306.

Gnanadesikan, R., Kettenring, J., 1972. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* 28, 81–124.

Hadi, A., 1992. Identifying multiple outliers in multivariate data. *J. Roy. Statist. Soc. B* 54, 761–771.

Hardin, J., Roche, D., 2002. The distribution of robust distances. <http://www.cipic.ucdavis.edu/~dmrocke/preprints.html>.

Hawkins, D., 1980. *Identification of Outliers*. Chapman & Hall, London.

Hawkins, D., 1999. Improved feasible solution algorithms for high breakdown estimation. *Comput. Statist. Data Anal.* 30, 1–11.

- Lopuhaä, H., Rousseeuw, P., 1991. Breakdown of points of affine equivariant estimators of multivariate location and covariance matrices. *Ann. Statist.* 19, 229–248.
- Maronna, R., Yohai, V., 1995. The behavior of the Stahel-Donoho robust multivariate estimator. *J. Amer. Statist. Assoc.* 90, 330–341.
- McLachlan, G., Peel, D., Basford, K., Adams, P., 1999. The emmix software for the fitting of mixtures of normal and t-components. *J. Statist. Software* 4 (2).
- Penny, K., 1995. Appropriate critical values when testing for a single multivariate outlier by using the mahalanobis distance. *Appl. Statist.* 45, 73–81.
- Reiners, T., Woodruff, D., 2001. A blackboard architecture applied to maximum likelihood clustering. Technical Report, CIPIC, UC Davis, Davis, CA 95616.
- Rocke, D., Woodruff, D., 1996. Identification of outliers in multivariate data. *J. Amer. Statist. Assoc.* 91, 1047–1061.
- Rousseeuw, P., 1984. Least median of squares regression. *J. Amer. Statist. Assoc.* 79, 871–880.
- Rousseeuw, P., Leroy, A., 1987. *Robust Regression and Outlier Detection*. Wiley, New York.
- Rousseeuw, P., VanDriessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41, 212–223.
- Rousseeuw, P., VanZomeren, B., 1990. Unmasking multivariate outliers and leverage points. *J. Amer. Statist. Assoc.* 85, 633–639.