Outlier Detection in the Multiple Cluster Setting Using the Minimum Covariance Determinant Estimator

Johanna Hardin
Department of Mathematics
Pomona College
610 N. College Ave.
Claremont, CA 91711
(909) 607-8717
jo.hardin@pomona.edu

David M. Rocke
Center for Image Processing and Integrated Computing
University of California at Davis
dmrocke@ucdavis.edu

August 2002

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 (2002) 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.

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.

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; Rocke 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 Rocke (2001) 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).

$$\begin{aligned} MCD &= (\overline{X}^*_J, S^*_J) \\ \text{where} \quad J &= \{ \text{ set of } h \text{ points}: |S^*_J| \leq |S^*_K| \quad \forall \text{ sets } K \text{ s.t. } \#|K| = h\} \\ & \quad \text{where } \#|\omega| \text{ defines the number of elements in set } \omega \\ \overline{X}^*_J &= \frac{1}{h}\sum_{i \in J} x_i \\ S^*_J &= \frac{1}{h}\sum_{i \in J}(x_i - \overline{X}^*_J)(x_i - \overline{X}^*_J)^t \end{aligned}$$

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 $h = \lfloor \frac{(n+p+1)}{2} \rfloor$ where $\lfloor \cdot \rfloor$ is the greatest integer function (Rousseeuw and Leroy, 1987; Lopuhaä and Rousseeuw, 1991). We will use $h = \lfloor \frac{(n+p+1)}{2} \rfloor$ 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 $h_i = \lfloor \frac{(n_i+p+1)}{2} \rfloor \;\; i = 1, \ldots, g$, 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 out 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 (Rocke 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 $\mathbf{y}_n \in \mathbf{R}^p$ is affine equivariant if and only if for any vector $\mathbf{v} \in \mathbb{R}^p$ and any nonsingular $p \times p$ matrix $\mathbf{M}$,

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

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 $\mathbf{M}$,

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

Stretching or rotating the data will not change an affine estimate of the data. If the location and shape estimates are affine equivariant, the Mahalanobis Squared Distances 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 MSD's 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 Minimum Covariance Determinant (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 $\overline{X}_{H_1}$ 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 $d^2_{S_{H_1}}(x_i, \overline{X}_{H_1}) = d^2_{H_1}(i)$ and sort them for some permutation π such that,

$$d^2_{H_1}(\pi(1)) \leq d^2_{H_1}(\pi(2)) \leq \ldots \leq d^2_{H_1}(\pi(n)).$$

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

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 $d^2_{S^*_j}(x_i, \overline{X^*_j})$, where S^*_j and $\overline{X^*_j}$ 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, $d^2_{S^*_j}(x_i, \overline{X^*_j})$ will be large for all j, and for points x_i that are not extreme, $d^2_{S^*_j}(x_i, \overline{X^*_j})$ 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

Mahalanobis squared distances 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 Rocke (2001), 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^2_p, 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,

$$\frac{c(m-p+1)}{pm} d^2_{S^*_X}(X_i, \overline{X^*}) \sim F_{p,m-p+1}. \tag{4.1}$$

where $\overline{X^*}$ 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,

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

where χ^2_ν is a Chi-Square random variable with ν degrees of freedom, and $\chi^2_{\nu,\epsilon}$ is the ϵ quantile for a χ^2_ν 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 Rocke, 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 Rocke, 2002).

$$m_{pred} = m_{asy} \cdot e^{(0.725 - 0.00668p - 0.0780\ln(n))} \tag{4.2}$$

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.

$$\#\text{ in cluster } j = n_j = \sum_{i=1}^{n} I(d^2_{S_j^*}(X_i, \overline{X_j^*}) \leq d^2_{S_k^*}(X_i, \overline{X_k^*}) \quad \forall\, k = 1, \ldots, g \text{ groups})$$

where $\overline{X_j^*}$ 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 =observation and j =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,

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

and

$$\frac{c_j(m_j - p + 1)}{p m_j} d^2_{S_j^*}(X_i, \overline{X_j^*}) \sim F_{p, m_j - p + 1}.$$

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