Marco Riani, Anthony C. Atkinson, Andrea Cerioli and Aldo Corbellini

Efficient robust methods via monitoring for clustering and multivariate data analysis

Article (Accepted version) (Refereed)

Efficient Robust Methods via Monitoring for Clustering and Multivariate Data Analysis

Marco Riani[a], Anthony C. Atkinson[b,*], Andrea Cerioli[a], Aldo Corbellini[a]

[a] *Dipartimento di Scienze Economiche e Aziendali and Interdepartmental Centre for Robust Statistics, Università di Parma, 43100 Parma, Italy*
[b] *Department of Statistics, The London School of Economics, London WC2A 2AE, UK*

Abstract

Monitoring the properties of single sample robust analyses of multivariate data as a function of breakdown point or efficiency leads to the adaptive choice of the best values of these parameters, eliminating arbitrary decisions about their values and so increasing the quality of estimators. Monitoring the trimming proportion in robust cluster analysis likewise leads to improved estimators. We illustrate these procedures on a sample of 424 cows with bovine phlegmon. For clustering we use a method which includes constraints on the eigenvalues of the dispersion matrices, so avoiding thread shaped clusters. The "car-bike" plot reveals the stability of clustering as the trimming level changes. The pattern of clusters and outliers alters appreciably for low levels of trimming.

Keywords: bovine phlegmon, "car-bike" plot, clustering, eigenvalue constraint, forward search, MCD, MM-estimation, modified BIC, outliers

*Corresponding author
Email addresses: mriani@unipr.it (Marco Riani), a.c.atkinson@lse.ac.uk (Anthony C. Atkinson), andrea.cerioli@unipr.it (Andrea Cerioli), aldo.corbellini@unipr.it (Aldo Corbellini)

1. Introduction

Robustness of statistical methods is the ability to provide correct answers about the generating mechanism of the main body of the data when contamination occurs. Historically, there have been two alternative ways to look at the outlier contamination problem and to achieve immunization against it.

The first approach consists in the development of statistical techniques that are inherently insensitive to the presence of even a substantial fraction of outliers. The largest fraction of contamination that the robust method can tolerate is called the *breakdown point* (bdp), and methods that can deal with up to 50% contamination are usually referred to as high-breakdown procedures [1]. Some well-known examples in this class include, for multivariate data, the Minimum Covariance Determinant estimator [2], S and MM estimators [3], and the Forward Search estimator [4]. All of them will be introduced in §2 below. Methods with breakdown point larger than 50% involve additional issues [5] and will not be considered in this paper. Robust tools also exist for more structured multidimensional tasks, such as the clustering problems addressed in our work [6, 7, 8, 9], and in other contexts not considered here, such as Principal Component Analysis [10, 11, 12], multivariate ranking [13, 14], multidimensional scaling [15], Support Vector Machines [16, 17] and feature extraction [18].

The second path to protecting against contamination is diagnostic and consists in the explicit identification of deviating observations before the main statistical analysis is performed. In recent years outlier (or anomaly) detection has gained considerable popularity also outside the statistical community; see, e.g., [19, 20, 21]. However, it is important to recall that any outlier detection technique must satisfy a crucial statistical requirement: it must guarantee against *masking* and *swamping*. The former arises when the aberrant observations attract the estimates in such a way that they do not appear anomalous anymore; conversely, the latter occurs when the estimation bias leads uncontaminated observations to be mistakenly labeled as outliers. Safeguard against these undesirable effects can be obtained by the use of diagnostic tools based

on high-breakdown estimators and by a careful design of the related statistical testing procedures [22, 23, 24, 25, 26].

Cerioli *et al.* [27, 28] give a brief history of robust statistical methods from the hopeful dawn at the time of the Princeton Robustness Study [29]. They suggest that a major disincentive to the routine use of standard robust methods is the dependence of the efficient application of these methods on the proportion of outliers expected in the particular set of data being analysed, that determine the desired efficiency or, equivalently, breakdown point. They also argue that monitoring the results of a statistical analysis, i.e. repeating the estimation process for different choices of the tuning parameters, can greatly enhance the applicability of robust statistical methods. The monitoring approach for multivariate data that we then exploit in this work, and that we extend to clustering, is developed by [30] as a fruitful reconciliation of the two alternative paths to robustness outlined above. Indeed, it can lead to robust estimators that are as statistically efficient as possible.

Our first analyses assume that we have a sample from a single population, contaminated by outliers. Clearly in such cases, a very robust analysis with a breakdown point of 50% can always be used, but this results in an unnecessarily low efficiency for data that are virtually outlier free. Standard approaches to increase efficiency are reweighting and the use of MM, rather than S, estimation. The contribution of our paper is to exhibit the use of adaptive methods based on monitoring a series of fits to the data that indicate good choices of efficiency or bdp.

Our major example, of measurements on 424 cows, shows a strong cluster structure with several clusters. We extend our method to the constrained cluster analysis of [31], in which constraints are placed on the shape of clusters through specification of the ratio of the largest to the smallest eigenvalues of dispersion matrices. This ratio is again a quantity which has to be chosen before the data are analysed. Their method achieves robustness via a specified level of trimming, which requires pre-specification. Dotto *et al.* [32] illustrate the use of reweighting to increase the efficiency of this procedure. We instead

use monitoring as an adaptive method that avoids the *a priori* choice of the trimming level.

We start with robustly fitting a single multivariate normal distribution to our data set. In §2 we define three sets of robust methods. The best values of bdp or efficiency are found by monitoring the behaviour of the robust fits over a range of values of these quantities. For hard trimming methods, such as the MCD, we monitor performance over a range of trimming values. These procedures are briefly illustrated in §3 where the data fall into two clusters. This structure is indicated by the patterns in plots of Mahalanobis distances resulting from monitoring. However, the emphasis in the examples in the section is on analyses using the Forward Search in which monitoring is part of the robust method.

Our major example is in §4 where the plots from monitoring are more complicated, indicating at least a three-part structure to the data. The robust cluster analysis of these data is presented in §5, beginning, in §5.1, with the use of random start forward searches to diagnose the presence of clusters. Cluster analyses for a variety of trimming levels are presented in §5.2, culminating in the use of monitoring to choose an appropriate amount of trimming. In this case we monitor the Adjusted Rand Index to determine the change in cluster membership as the amount of trimming decreases; monitoring the "car-bike" plot reveals the stability of solutions with differing numbers of clusters. This analysis finds five clusters and around 50 unclassified units. In §5.3 we investigate how the clustering structure and patterns of residuals change at low trimming levels. Our monitored clustering of the data illuminates the properties of the measurements, which come from cows on seven farms.

Conclusions are in §6 followed by a two-part Appendix, the first part presenting theoretical results for the Forward Search and the second illustrating the use of the Search in analyses of multivariate data with increasing contamination.

2. A Taxonomy of Robust Methods and Their Monitoring

We can identify three classes of robust estimators for multivariate data:

1. Soft Trimming (downweighting). The intention is that observations near the centre of the distribution essentially retain their value, but a suitable weight function ensures that increasingly remote observations have an effect on fitting that decreases with distance from the centre. We look at two such methods, S estimation and MM estimation for both of which we employ the Tukey biweight as the downweighting function [3, 33].

2. Hard (0,1) Trimming: the minimum covariance determinant (MCD) and the minimum volume ellipsoid (MVE) [34]. In both methods h out of the n observations are used to estimate the parameters. The value of h is often taken as $\lfloor 2\lfloor (n+v+1)/2 \rfloor - n + (n - \lfloor (n+v+1)/2 \rfloor) \rfloor$, where v is the number of variables. Larger values give more efficient estimates of the parameters but with lower bdp.

3. Adaptive Hard Trimming. In the Forward Search (FS), the observations are again hard trimmed, but the amount of trimming is determined by the data, being found adaptively by the search. See [22] for multivariate data and [35] for a general survey of the FS, with discussion.

We obtain adaptive versions of these procedures through the use of "monitoring"; we calculate a series of robust fits as the parameter determining the properties of the fit varies over its whole range. For S estimation we vary the bdp from 0.5 (most robust) to 0.01, virtually the non-robust maximum likelihood (ML) solution. In our MM estimation we start with the most robust scale estimate found using S estimation and then monitor the fits obtained as the efficiency varies from 0.5 to 0.99.

In the hard trimming methods we monitor the fits obtained as h varies for $n/2$ to n. We do not need to adapt the FS since this already provides a series of fits as the subset size m increases from very small to all the data.

Producing such a number of robust fits is no longer a computational bur-

den. Partly this is due to the continually improving performance of computers. However a major factor is the efficient programming in the FSDA toolbox [36], which allows very fast computation of robust procedures and related graphical methods.

The idea of monitoring robust procedures was introduced for regression by [27]. More recently [30] provided a thorough exploration of monitoring single population robust methods for multivariate data, where further details of the soft trimming procedures of §§3 and 4 may be found.

3. A Straightforward Example: Eruption of Old Faithful

To demonstrate the use of monitoring in the analysis of a straightforward data set, we start with a brief analysis of data on the eruptions of the Old Faithful geyser in Yellowstone National Park, Montana. First we provide a summary of results for monitoring MM estimation. An extended analysis using monitoring to compare the properties of several robust estimators is in §5 of [30]. We then use the forward search both to identify outliers and, through the random start forward search, to identify the cluster structure of the data. These analyses serve as an introduction to the procedures for the analysis of the more complicated data set in §4.

The data are taken from the MASS library [37]. There are 272 observations with y_{1i} the duration of the ith eruption and y_{2i} the waiting time to the start of that eruption from the start of eruption $i-1$.

The left-hand panel of Figure 1 shows the effect on estimation of changing the stipulated efficiency of the MM procedure. The plot shows the values of all n squared Mahalanobis distances as efficiency varies using a "heat map". In the coloured .pdf version areas with many overlapping trajectories are shown in bright blue with the remaining individual trajectories in dark blue. The horizontal line is the 99% point of χ^2_v. From values of efficiency from 0.5 to 0.7 a robust analysis is obtained in which the outlying observations correspond to the smaller of the two clusters into which the data fall, and the standard analysis

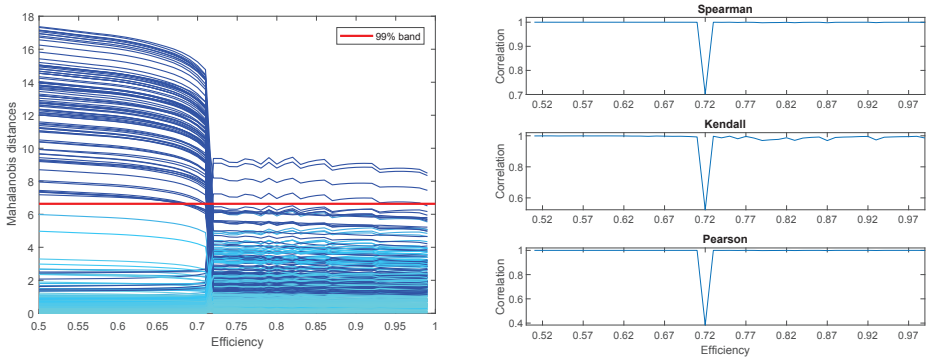

Figure 1: Eruptions of Old Faithful. Left-hand panel, Mahalanobis distances from monitoring MM estimation. Right-hand panel, monitoring correlation between consecutive distances.

is recovered. On the other hand, for high values of efficiency, from one down to just above 0.7, the maximum likelihood solution is obtained in which there is no indication of any clustered structure. The right-hand panel of the plot provides a method of obtaining the maximum empirical efficiency for these data, using various measures of the correlation between the n Mahalanobis distances at adjacent values of efficiency. These plots clearly indicate a maximum empirical efficiency of 0.71.

The standard advice to use a high value of efficiency such as 0.99 is overambitious. The monitoring of Mahalanobis distances leads to the adaptive choice of the highest possible efficiency for these data and so to the most efficient robust MM estimator.

Further information can be extracted from Figure 1 by "brushing" the units with large Mahalanobis distances in the left-hand part of the figure. In this process the cursor is used to select a region on the screen in which the trajectories of interest lie. The units corresponding to these trajectories are then highlighted in a linked scatterplot, in this case showing that they form a cluster of observations distinct from the greater part of the population. All distances in the right-hand part of the plot in the left-hand panel of Figure 1 are relatively small because a single multivariate normal model is being fitted which has its centre between the two cluster centres. An example of this for a different set of

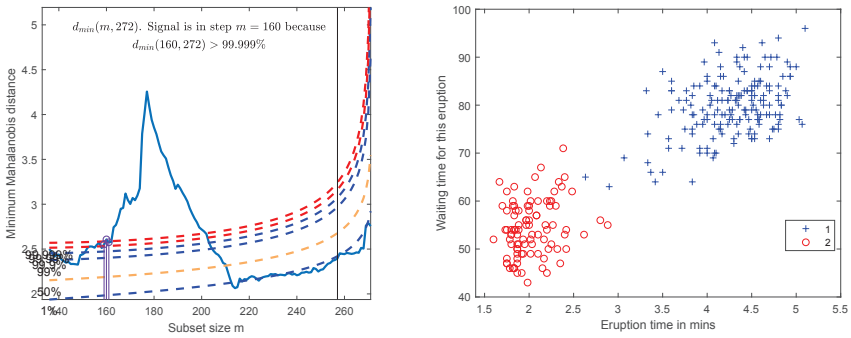

Figure 2: Eruptions of Old Faithful. Left-hand panel, minimum Mahalanobis distances from the Forward Search. Right-hand panel, + the cluster of non-outlying observations.

data is in [38, §7.3].

We now give a brief description of the analysis of these data using the FS in which, as described in §Appendix A.1, monitoring is built into the robust method. We start by fitting a single multivariate model, the analysis producing figures similar to those in §Appendix A.2.

The left-hand panel of Figure 2 shows the plot of the minimum Mahalanobis distance of the observations not in the subset used in fitting against subset size, as in Figure A.25. As described in §Appendix A.1 we test for outliers for a variety of sample sizes $n^\dagger$ until we find the largest sample containing no outliers. For testing we extend the notation for the minimum Mahalanobis distance to $d_{\min}(m, n^\dagger)$. Initially $n^\dagger = n$ and the signal for outliers occurs earlier at $m = 160$ because $d_{\min}(160, 272)$ is greater than the threshold in [22]. In order to detect which observations are indeed outlying, we illustrated in Figures A.23 and A.26 the use of the superimposition of envelopes. In these figures the envelopes are appreciably curved for values of m near n, since more remote observations enter towards the end of the search, giving rise to larger distances. However, the check of whether the value of $d_{\min}(m, n^\dagger)$ is above or below a threshold is pointwise for each m. We can therefore transform the vertical scale at each m without changing the rule. One possibility for ease of reading the graphs is to use the normal probability transformation to straighten the envelopes. Let the level of an envelope be γ. Then the normal probability transformation yields an

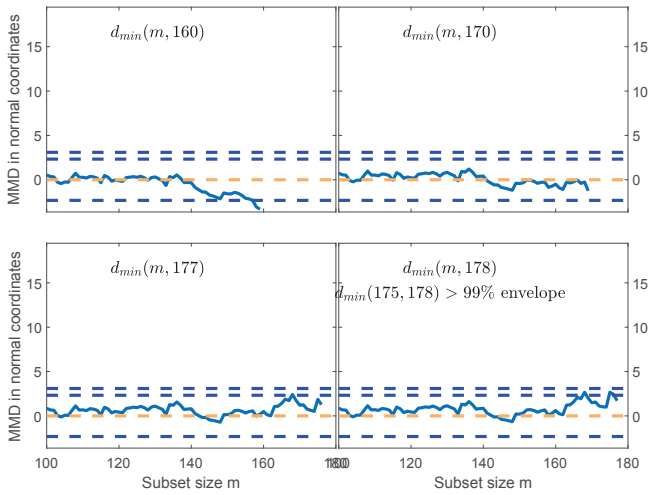

Figure 3: Eruptions of Old Faithful. Resuperimposition of envelopes for minimum Mahalanobis distances (MMD) in normal coordinates. The outlier-free cluster contains 177 observations.

envelope value $\Phi^{-1}(\gamma)$.

Figure 3 shows several resuperimposition curves in normal coordinates for the Old Faithful data. The top left-hand plot is for $n^\dagger = 160$, the value at which a signal occurred. There is no evidence of any outliers in a sample of this size, nor for that with $n^\dagger = 170$ in the right-hand panel in this row. The bottom left-hand panel, for $n^\dagger = 177$ likewise shows no outlier but the final panel, for $n^\dagger = 178$ reveals that one observation now lies above the 99% bound. Since the procedure is executed automatically the added clarity from the use of normal coordinates is solely for illustration of the FS.

The data have therefore been divided into two groups, one with a multivariate normal structure containing 177 observations, and the rest of the data. From the right-hand panel of Figure 2 it is clear that these form a second cluster. We now illustrate the use of the random start FS introduced at the end of §Appendix A.2 to determine the membership of the clusters. The trajectories of minimum Mahalanobis distances from 200 random starts are in Figure 4. The structure is similar to that of Figure A.28 but now the two peaks, indicating

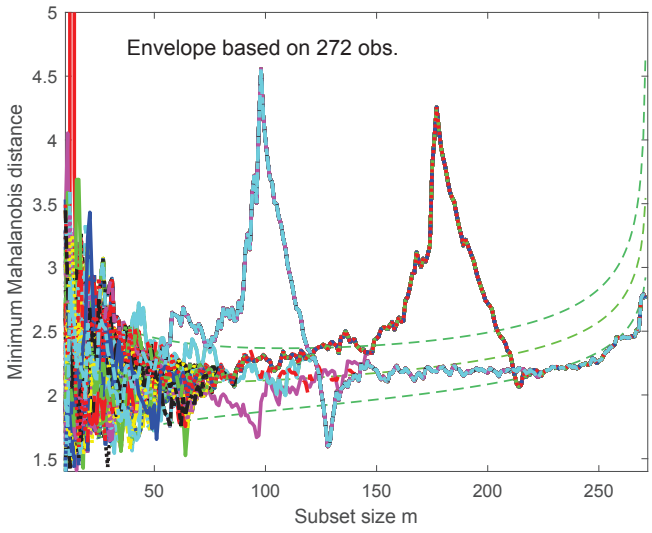

Figure 4: Eruptions of Old Faithful. Minimum Mahalanobis distances from 200 random start Forward Searches indicating the presence of two clusters.

the two clusters, are at 98 and 177. To identify the two clusters we re-run the FS twice, starting with the initial subset of observations from a randomly selected trajectory that gave each peak. The left-hand panel of Figure 5 shows the results of the FS in normal coordinates and the right-hand panel shows the final classification after resuperimposition. The cluster contains 97 units. Since there are 272 units in all, this analysis shows that there are two units that could belong in either cluster. In contrast, in the robust clustering method of §5.2 a firm decision is made about the allocation of each unit; it is either allocated to a specific cluster or is treated as a outlier.

4. Cows with Bovine phlegmon

In this section we consider an example in which the structure is shown to be more than a main sample and a second distinct cluster.

The data are 424 readings on four properties of cows suffering from Phlegmon, a form of foot rot. The four variables are numerical properties calculated from photographic measurements of the cows. The left-hand panel of Figure 6

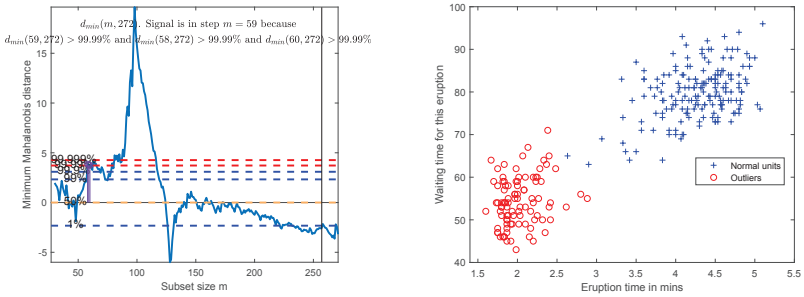

Figure 5: Eruptions of Old Faithful. Left-hand panel, minimum Mahalanobis distances in the normal scale from the Forward Search when initializing the search in the lower group. Right-hand panel, + the cluster of non-outlying observations.

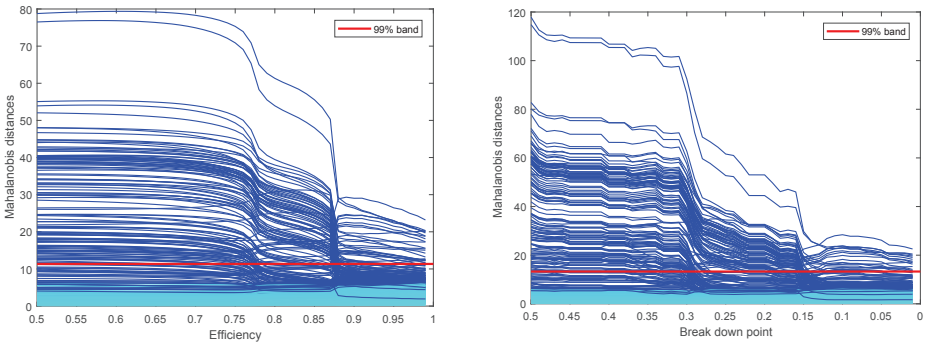

Figure 6: Cows with Phlegmon. Monitoring Mahalanobis distances. Left-hand panel MM estimation, right-hand panel, reweighted MCD.

shows the monitoring plot of squared Mahalanobis distances for the MM estimator. The behaviour is more interesting and complex than the monitoring plots of Figure 1. There are now three stable regions; the plot of Kendall's τ analogously to that in Figure 1, now indicates changes at efficiencies of 0.78 and 0.88.

We now turn to the monitoring of hard trimming methods. The right-hand panel of Figure 6 shows the monitoring plot for Mahalanobis distances for the reweighted MCD, with a pointwise boundary of the 99% point of the χ^2_4 distribution for giving an observation zero weight. The distances form three groups in which they are roughly parallel, with transitions occurring at bdp values around 0.3 and 0.15. The stable structure of this plot with roughly

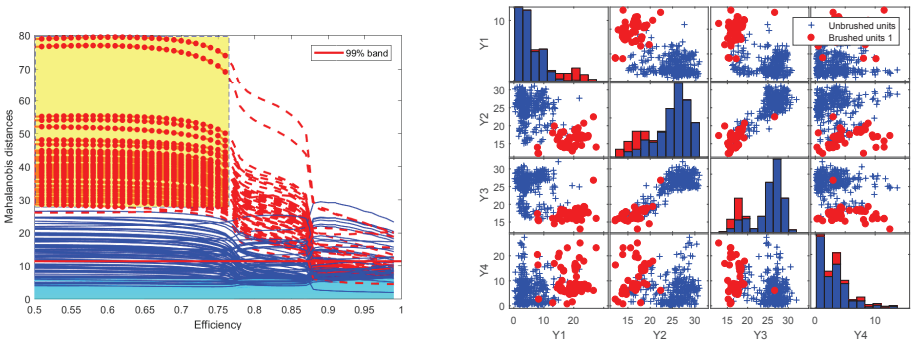

Figure 7: Cows with Phlegmon. Left-hand panel, brushing the most outlying distances with efficiencies between 0.5 and 0.77 in the MM analysis of Figure 6. Right-hand panel, clustering of brushed units.

parallel sets of distances arises when the reweighting has no effect over some range of bdp values.

The two panels of Figure 7 help interpret this structure for the MM estimator. The left-hand panel shows a brush for the most outlying distances between an efficiency of 0.5 and 0.77. The right-hand panel of the figure shows that the brushed units form a neat cluster, particularly evident in the plot of y_1 against y_2. However, the scatterplot of y_3 against y_4 suggests that the unbrushed group of units may also form a compact group, but with a scattering of outliers.

We now briefly report an analysis of the data in which we use the Forward Search. Figure 8 shows a plot of the minimum Mahalanobis distances during the search. There is a signal at $m = 246$ and 127 outliers are identified. What is interesting in view of the results from monitoring the MM and reweighted MCD estimates is the trajectory of the distance in the left-hand panel of Figure 8 which exhibits two large peaks, perhaps indicative of two groups. The scatterplot matrix in the right-hand panel shows that the FS has found an ellipsoidal group of central observations. As with the results of the MM analysis shown in Figure 7, there does seem to be some further clustering in the outliers.

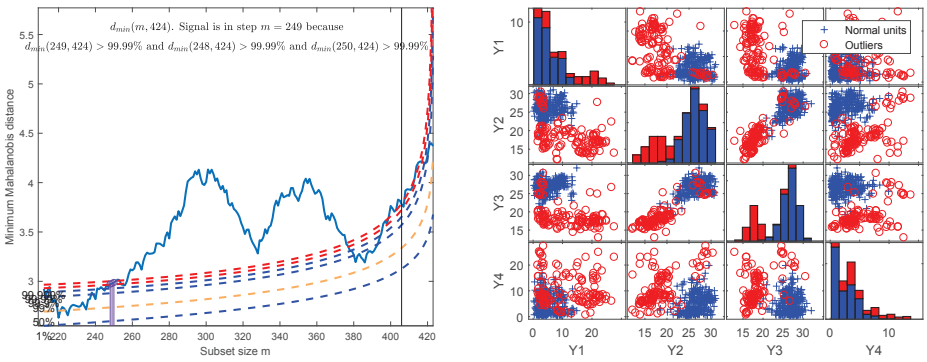

Figure 8: Cows with Phlegmon. Left-hand panel, minimum Mahalanobis distances from the Forward Search. 127 outliers are identified. Right-hand panel, the cluster of non-outlying observations.

5. Clustering the Data on Bovine Phlegmon

5.1. Random Start Forward Searches

The left-hand panel of Figure 9 shows the results of 200 random start forward searches. Particular interest was in the searches that gave a peak at $m = 103$. These have been plotted in a darker colour than the remaining searches.

The obvious feature of the plot is that from $m = 241$ this plot is the same as that in the left-hand panel of Figure 8; as the search progresses trajectories from various starting points converge. Once they have converged there is no possibility of divergence. Figure 8 indicated one cluster of observations. Brushing the peak at $m = 103$ in Figure 9 indicates a second cluster, shown in the right-hand panel. Thus two clusters have been tentatively identified. The plot of y_2 against y_3 shows the clusters in an exemplary way. The third peak in the forward plot is formed from units in these two cluster. However, inspection of the other panels of the scatterplot suggests that these two clusters could perhaps be further divided. For this we turn to robust cluster analysis.

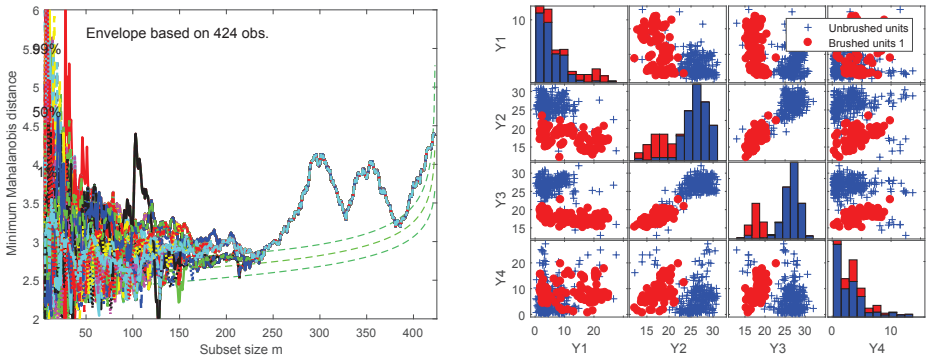

Figure 9: Cows with Phlegmon; random start Forward Searches. Left-hand panel, minimum Mahalanobis distances from 200 random start Forward Searches. There is a peak at $m = 103$. Right-hand panel, the cluster of observations from brushing this peak.

5.2. Robust Clustering with Constraints

We base our method of robust clustering on a trimmed version of the constrained likelihood clustering procedure of [39], using monitoring to determine the amount of trimming. The fitted model is a mixture of multivariate normal distributions.

This method, in the absence of trimming, overcomes the problem of unbounded likelihood associated with fitting an unconstrained mixture of normal distributions. For a specified number of clusters K, the fitting procedure for a v-dimensional problem starts by randomly choosing, without replacement, $v+1$ observations to form each of K cluster centres. The clusters are then grown from these centres. The procedure is repeated, in our example 2,000 times. Since interest is not only in finding the best clustering for a given K, but also in finding stable solutions as the other parameters change, we need to look at a few best solutions. A problem is that these may be virtually identical. The Adjusted Rand Index (ARI) provides a method of identifying partitions which are "essentially the same", all except the best of which are discarded as being "spurious" [40]. We use a value of 0.7 as the threshold above which clusters are considered the same.

We follow [31] and use a constraint $c \geq 1$ on the ratio of the largest to smallest

eigenvalues of the dispersion matrices of the clusters. For a given trimming level α we find clustering solutions over a range of values of c and K. In standard clustering, such as [41], the Bayesian Information Criterion (BIC) [42] is used to select the value of K. If $L_K(\theta)$ is the loglikelihood of the observations for a particular K, this criterion minimizes $-2L_K(\theta) + P_K$, where

$$P_K = \{(Kv + K - 1) + Kv(v + 1)/2\} \log n,$$

is a penalty term for the number of free parameters; there are Kv means, $K - 1$ mixture proportions and K dispersion matrices each with $v(v+1)/2$ parameters. But the application of the constraints from the value of c reduces the number of free parameters in the model and should be allowed for in the model selection criterion. We use the modified BIC criterion introduced by [39] when the penalty term becomes

$$P_K^c = \{(Kv + K - 1) + Kv(v - 1)/2 + (Kv - 1)(1 - 1/c) + 1\} \log n.$$

Now the second term is the number of parameters for orthogonal rotations of the dispersion matrices, unaffected by the constraints, and the third those related to the eigenvalues. This term moves smoothly from the most constrained case, that is $c = 1$, to complete freedom in the choice of all eigenvalues when the criterion become the standard BIC.

Since our earlier analysis of the cows data has indicated the presence of numerous outliers, we start with the high value of 0.3 for the trimming parameter α. Our aim in monitoring is to see how sensitive the clustering solution is to the value of α, in the hope that we can find a stable solution with a lower amount of trimming.

Figure 10 shows the plot of the modified BIC, when $\alpha = 0.3$, for the number of clusters K going from 1 to 8 and c a power of 2 from 1 to 128. For this high level of trimming we find that 5 clusters is optimum with $c = 128$. The values of c that give the minimum modified BIC for each K are listed at the top of the figure; apart from that for $K = 1$, high values of c are optimum. It is clear that trimming in this range provides a strong cluster structure.

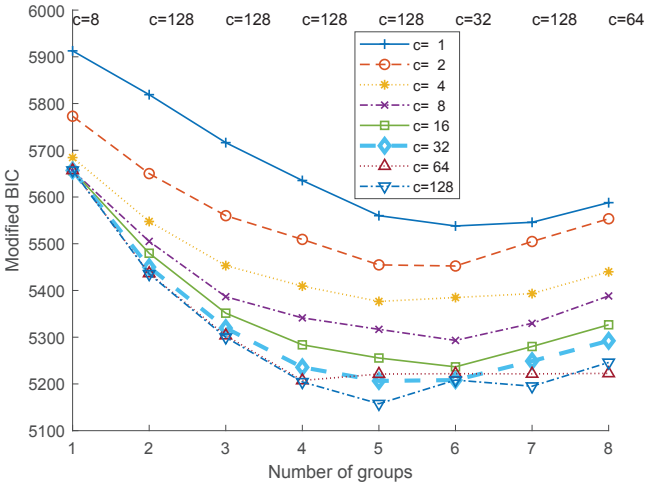

Figure 10: Cows with Phlegmon. Modified BIC as a function of cluster number K and eigenvalue ratio c. Trimming level $\alpha = 0.30$.

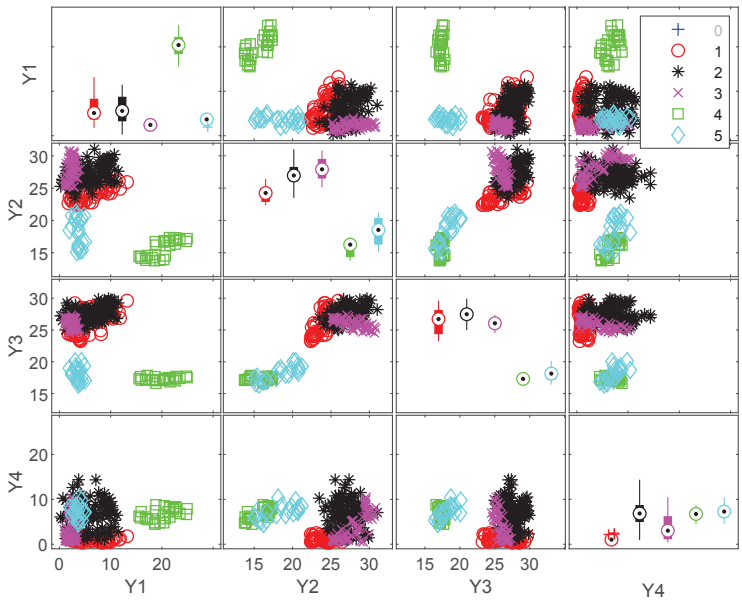

Figure 11: Cows with Phlegmon. Scatterplot matrix of the five clusters identified when $c = 128$ and $\alpha = 0.30$.

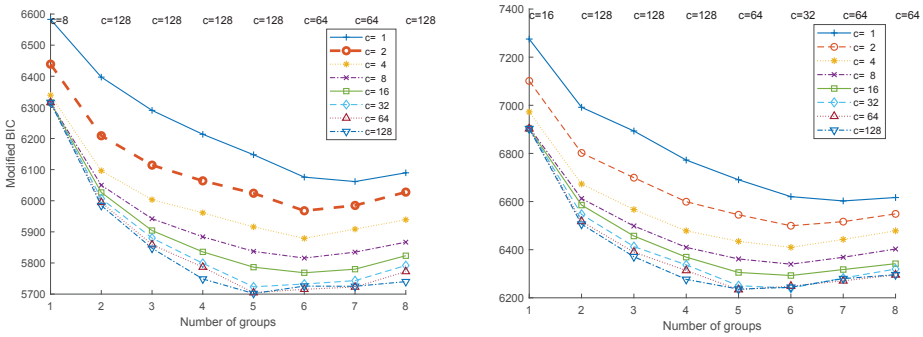

Figure 12: Cows with Phlegmon. Modified BIC as a function of cluster number K and eigenvalue ratio c. Left-hand panel, trimming level $\alpha = 0.25$. Right-hand panel, $\alpha = 0.2$.

The clusters that are indicated are shown in the scatterplot matrix of Figure 11 with the 30% trimmed observations removed from the plot. The plot shows how the two clusters found earlier have been split. The left-hand half of the plot of y_1 against y_3 shows how the elongated cluster in the left-hand half of the right-hand panel of Figure 9 has been divided into two. The division of the larger cluster into three parts is less clear. The diagonal panels of the matrix give boxplots of the observations in the clusters in that row. The x co-ordinate of the boxplots is the number of the cluster. These four panels of the boxplots fail to reveal appreciable differences between the co-ordinate wise means or scatters of the five groups.

We now repeat the analyses for smaller levels of α, that is for lower levels of trimming. Figure 12 shows plots of the modified BIC for $\alpha = 0.25$ and 0.20. In both cases the three highest values of c indicate five clusters. Lower values of c indicate more clusters. The scatterplot matrix of the clusters for $\alpha = 0.2$ is in Figure 13. As in Figure 11 for $\alpha = 0.3$, the trimmed observations have been removed from the plot. Comparison of the two figures shows particular growth in groups 4 and 5, which become less clearly separated from the other groups.

Plots of the modified BIC against K, such as Figure 12, only present information about the best partition of the data for each value of K and c. In order to exhibit the stability of the solutions to changes in c, [39] introduced the "car-bike" plot. An example is shown in Figure 14 for $\alpha = 0.2$. Of the five best

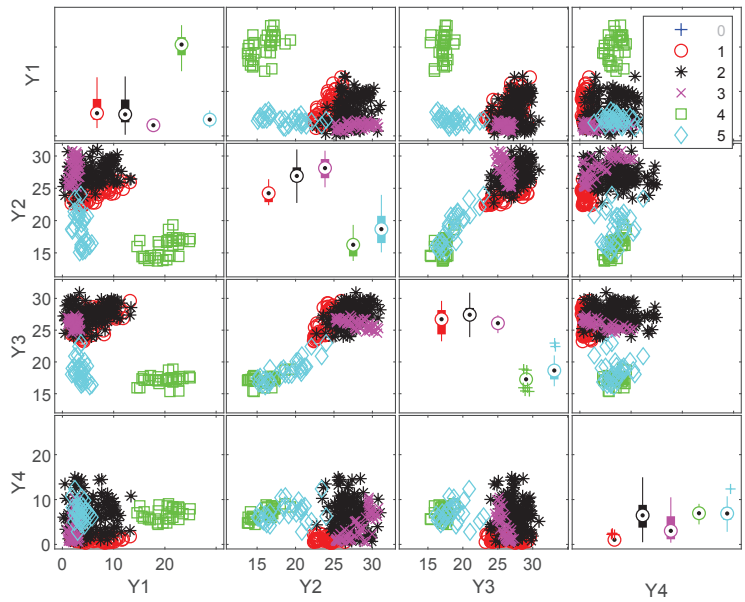

Figure 13: Cows with Phlegmon. Scatterplot matrix of the five clusters identified when $c = 128$ and $\alpha = 0.2$.

solutions found, calculation of the ARI shows that two are very close to some of the other clusterings of the data; they are therefore discarded as spurious. The figure shows that, for $c = 64$ and 128, the best solution has five clusters. The two numbers in the circle are the ranking after and before spurious solutions have been removed. The bar indicates the values of c for which these are the best solutions. The line, in this case for lower values of c, shows that partitions into five clusters, similar to these (as measured by the ARI), are obtained for c = 32 and 16. There are also good solutions for $K = 7$. The second best solution overall, ignoring the solutions for five clusters in the figure, is for $c = 64$; the line shows that a similar solution is found for $c = 128$. The third best solution is for $c = 32$ and $K = 7$. The conclusion is that five group partitions are stable over a range of c values whereas the other solutions, for seven clusters, are second best, also being less stable to changes in c.

So far we have considered relatively heavy levels of trimming. As a final detailed cluster analysis we look at partitions with a 10% trimming level. The

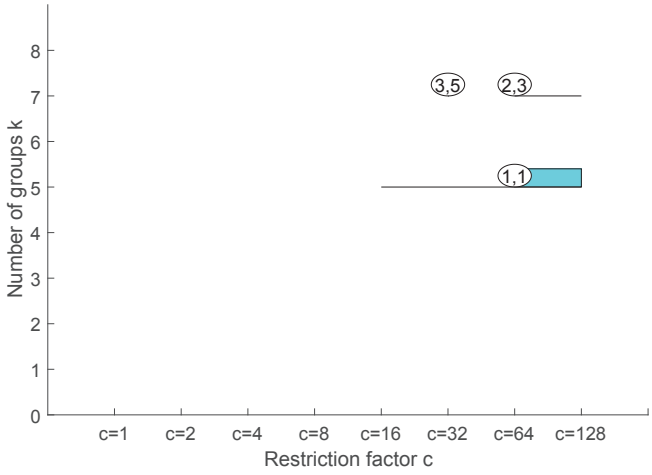

Figure 14: Cows with Phlegmon. "Car-bike" plot for $\alpha = 0.2$ showing the stability of the five cluster solution to the value of the eigenvalue ratio c. The bar shows that the five cluster solution is best for c from 64 to 128.

results are shown in Figure 15. The left-hand panel shows the plot of modified BIC. The minimum of these curves is for a partition with $K = 7$ when $c = 128$. However, the car-bike plot in the right-hand panel shows that this solution is sensitive to the value of c; the second best partition has six groups, likewise for $c = 128$. The only solution stable over a range of values of c is again that for five groups, which is the third best solution. The bar, thinner than that in Figure 14, shows that this is also the third best solution for $c = 32$ and 64. The line shows that a similar partition is obtained for $c = 16$. The conclusion is that, if highly non-spherical groups are allowed with low levels of trimming, the seven cluster solution is optimum, but unstable to values of c. As c decreases the five cluster solution is preferred. These two values of K are important in the remainder of our analyses.

A main point of our paper is the importance of monitoring to provide adaptive values for the parameters required in a robust analysis. In the form of robust clustering we have been investigating we need to specify both c and α. The two car-bike plots show that, except for low levels of trimming, a value of 32 for c provides a stable five-cluster solution. We now use monitoring of the

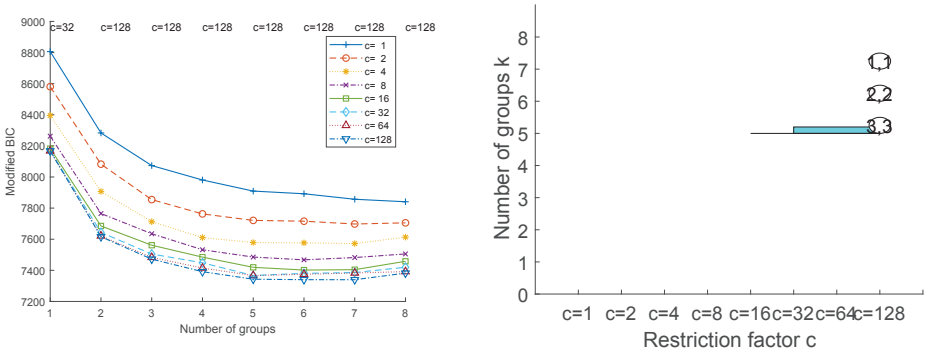

Figure 15: Cows with Phlegmon. Left-hand panel: modified BIC as a function of cluster number K and eigenvalue ratio c. Trimming level $\alpha = 0.1$. Right-hand panel: "Car-bike" plot for $\alpha = 0.1$ showing the stability of the five cluster solution as c changes.

trimming level to determine the best value of α. For a series of values of α between 0.3 and 0, that is no trimming, we calculate the ARI between partitions for adjacent values of α, in all cases for $K = 5$. As Figure 16 indicates, we obtain stable solutions up to $\alpha = 0.04$; sometimes the solutions are so similar that the ARI is close to its maximum value of one. For $\alpha = 0.03$ and lower values the clustering structure starts to change appreciably with the trimming level and 0.04 is the optimum trimming level. This determination of the optimum value of α is analogous to the monitoring used in §4 to find data dependent values of efficiency and bdp.

In order to avoid excessive random fluctuations in Figure 16 we used a set of 20,000 seeds for the initial cluster centres and mixing proportions. These sets of points in 4-dimensional space were calculated once and used to initialise the clustering for each value of α. The alternative, independent random sampling of the starts for each value of α, gave a rougher plot than that shown here.

Figure 16 shows rapid change in cluster structure for $\alpha < 0.04$. As a final illustration of monitoring we present a plot that monitors two aspects of the cluster structure as a function of α, throughout for five clusters.

For the jth value of the trimming parameter, α_j, let the estimated mean in the kth group be $\hat{\mu}_{jk}$. The change in this mean in moving from α_{j-1} to α_j is $\delta(\mu_{jk})$. As one diagnostic measure we monitor the squared Euclidean distance of

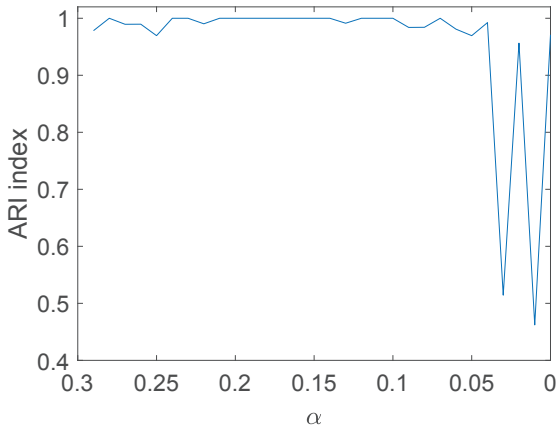

Figure 16: Cows with Phlegmon. Monitoring the ARI between consecutive cluster allocations as a function of the trimming proportion α; $K = 5, c = 32$.

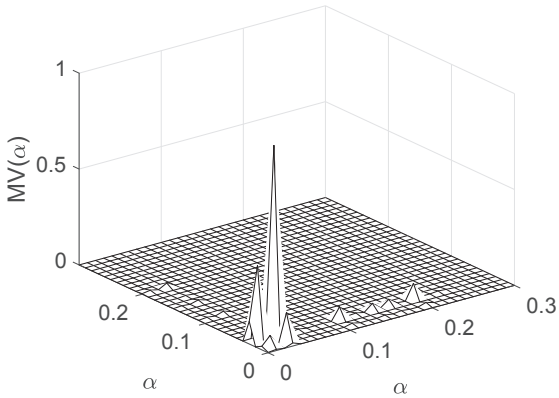

Figure 17: Cows with Phlegmon. The matrix measure $MV(\alpha)$ of changes in means and covariance matrix with trimming level; $K = 5, c = 32$.

these changes $\Delta(\mu_j) = \sum_{k=1}^{v} \delta^2(\mu_{jk})$. To monitor the changes in the estimated covariance matrices $\hat{\Sigma}_j$ we look at the diagonal elements $\hat{\sigma}_{jkk}$. Taking the sum of the squared differences of these elements in going from α_{j-1} to α_j gives a second vector of measures $\Delta(\Sigma)_j$. Let

$$\Delta(\mu) = \{\Delta(\mu_j)\} \quad \text{and} \quad \Delta(\Sigma) = \{\Delta(\Sigma_j)\}$$

be $r \times 1$ vectors. As a measure we form the outer product of the two vectors. After standardization by the maxima of each vector we obtain the measure of change in the means and covariance as a function of α

$$MV(\alpha) = \frac{\Delta(\mu)\Delta(\Sigma)^T}{\sup_j \Delta(\mu_j) \sup_j \Delta(\Sigma_j)}, \tag{1}$$

a matrix of dimension $r \times r$.

The bivariate plot of this matrix monitoring measure is given in Figure 17. It confirms that the large changes in the structure of the groups, as reflected through the differences in means and the diagonals of the covariance matrices of the five groups occur for the low values of α that we have already noted. The small peaks for values of α between 0.1 and 0.2 are caused by changes in the variance measure $\Delta(\Sigma)$.

5.3. Interpretation of the Five Clusters

The data come from measurements at seven different farms, a piece of information we did not use in the analysis described here. In order to see how our clustering agrees with the allocation to farms we again plot the ARI as a function of α, with the index calculated for the agreement between the trimmed clustering allocation with $K = 5$ and $c = 32$ and the identically trimmed set of observations from the farms. As Figure 18 shows, there is good agreement between the two allocations up again to an α value of 0.04. Thereafter, as might be expected when comparing a five group allocation with one with seven groups, the two allocations become less close.

There is no reason why all seven farms should be distinct. If two of the farms are close in properties to some of the other five, the plot of Figure 18

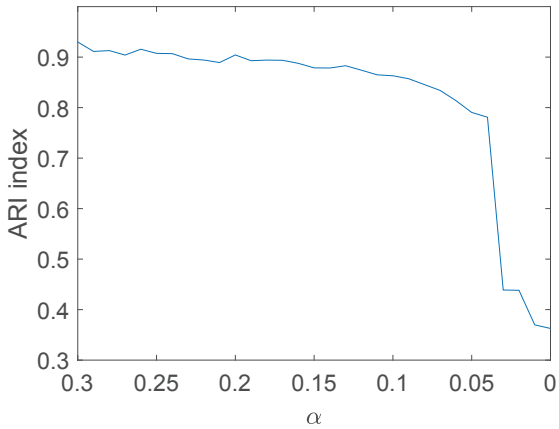

Figure 18: Cows with Phlegmon. Monitoring the ARI between cluster allocations and farm number as a function of the trimming proportion α; $K = 5, c = 32$.

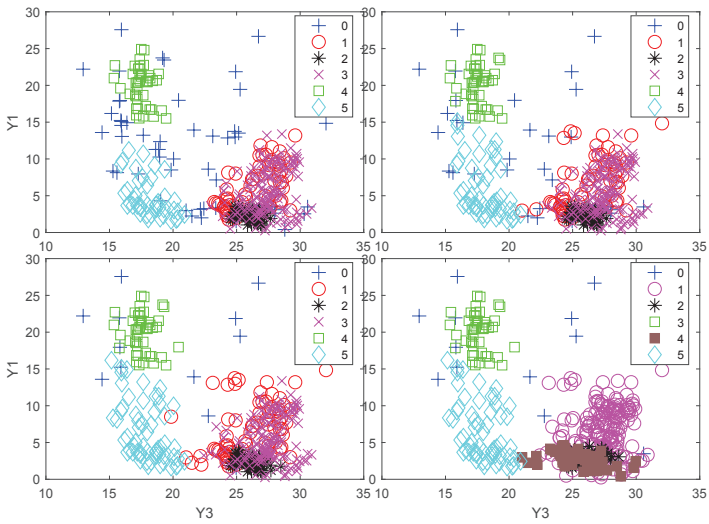

Figure 19: Cows with Phlegmon. Scatterplots of y_1 against y_3, including unclassified units (Group 0). Top left $\alpha = 0.14$, top right $\alpha = 0.09$, bottom left $\alpha = 0.04$, bottom right $\alpha = 0.03$.

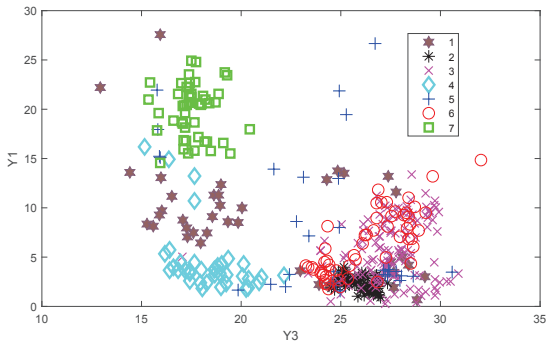

Figure 20: Cows with Phlegmon. Scatterplot of y_1 against y_3, including unclassified units, grouped by farm.

would continue near one for all α. The high values for much of the range of α suggest that we have indeed found clusters that correspond to the different farms when there is some trimming. The low values from low trimming may suggest the presence of groups we have not found or perhaps the presence of some unstructured outliers. A third possibility is a highly non-normal cluster.

To explore these possibilities Figure 19 shows scatterplots of y_1 against y_3 for a range of values of α in which unclassified units are included as Group 0. The top left-hand plot shows the cluster allocation for $\alpha = 0.14$. The clustering, in line with the indication of Figure 16, is similar to that for $\alpha = 0.2$ (Figure 13) and $\alpha = 0.3$ (Figure 11). The differences are that the clusters contain more units, since α is smaller, and that the unclassified units are included in Figure 19. Moving to a value of 0.09 for α in the top right-hand panel of the figure shows appreciable growth in Group 5, which has moved towards Group 4 by absorbing some previously unclustered units. The changes in going from $\alpha = 0.09$ to 0.04 (in the bottom left-hand panel) are slight; Groups 1 and 5 have both gained some units. As would be expected from the ARI plot of Figure 12, there is an appreciable change in going to $\alpha = 0.03$. The chief change is in Groups 1 and 3 for $\alpha = 0.04$ with Group 1 expanding and Group 2 being replaced by a different structure (shown by brown squares in the .pdf).

The comparison with the groups of farms in Figure 20 is instructive. It is

clear that trimming has an effect on the identification of Farm 4 with Group 5. At least in this projection, the units for Farm 4 are divided from those from Farm 7 by some from Farm 1. As α decreases from 0.14 to 0.04, Group 5 expands to include some units from Farm 1. The plot also shows that Group 2 mostly includes units from Farm 2. Farm 1 mostly lies between Groups 4 and 5, but includes a scattering of units in or near other groups. Farm 5 likewise has a scattering of units. Trimming is needed to avoid these units from distorting the groups corresponding to other farms. Farm 3 is included in Group 3 and Farm 6 in Group 1, close together in the three plots excluding that for $\alpha = 0.3$.

The least regular behaviour is that for readings from Farm 1, which provide many of the outliers in the panels of Figure 19. It is interesting that this was the first farm on which the photographic procedure was tried; the data structure is a clear indication of learning and calibration difficulties. Overall, our cluster analysis of the data on cows with phlegmon shows that trimming of as little as 0.04 achieves an efficient partition of the data at the cost of some misallocations of units from two very non-normal populations.

6. Conclusions

Data often contain outliers. Appropriate robust methods, correctly tuned, can reveal the outliers and provide fitted models with highly efficient parameter estimates. In this paper we have considered the clustering of data into groups following multivariate normal distributions. In order to bound the likelihood for this clustering problem we have used a constraint c on the shapes of the covariance matrices of the clusters. Robustness has been obtained by trimming a proportion α of the observations, those that are furthest from the centre of the clusters to which they are assigned. For a fixed level of trimming, a modification of the information criterion BIC can be used to select K, the number of clusters. Such plots are informative about the best partition of the data for each value of K and c.

The stability of the solutions to changes in c is shown by the car-bike plot,

calculated for a specified level of trimming α. To provide an efficient and robust clustering procedure, we monitor the behaviour of the partitions of the data as α changes. The Adjusted Rand Index (ARI) provides a measure of the similarity of two partitions of the data. Monitoring this index as α decreases from a high value to zero (no trimming, appropriate in the absence of outliers) leads to the estimate of α as the smallest value for which the clustering is stable; smaller values of α lead to changes in the ARI, indicating changes of cluster structure as outliers are introduced into the data being fitted. Our example on cows in Figure 16 shows how sharp the inference on trimming level can be.

The analyses in our paper extend the single-sample monitoring procedure presented in [30] to cluster analysis. Monitoring the effect of varying the level of α leads to a data-adaptive choice of trimming level. The discussion to [30] contains three contributions ([45], [46] and [47]) which describe other ways of monitoring cluster analyses. These further illustrate the power of monitoring, combined with informative plots, in establishing the structure of the data and determining the best values of the parameters defining a variety of robust methods.

Although the focus of our paper is on data with multivariate normal distributions, the scope is much wider. Use of the approximate normalising transformation of [48] makes normal theory clustering appropriate for many data sets with skewed observations. Examples of the use of the Box and Cox transformation in the analysis of multivariate data are in Chapter 4 and successive chapters of [38].

Acknowledgements

This research benefits from the HPC (High Performance Computing) facility of the University of Parma. M.R. gratefully acknowledges support from the CRoNoS project, reference CRoNoS COST Action IC1408. M.R. and A.C.A . would like to thank the European Unions Horizon 2020 Research and Innovation Programme for its financial support of the Prime Fish project, Grant Agreement No. 635761.

Appendix A. The Forward Search

The purpose of this two-part appendix is to give a short summary of the Forward Search, providing background for the FS analyses in §§4 and 5. Theoretical results are in the first part with numerical examples in the second.

Appendix A.1. Key Ideas and Mahalanobis Distances

The forward search (FS) provides an automatic form of monitoring. We start by fitting a small and supposedly homogenous subset of observations, often chosen through some robust criterion. The fitting subset is then repeatedly augmented in such a way that outliers and other influential observations enter toward the end of the search. Their inclusion is typically signalled by a sharp increase in suitable diagnostic measures, the values of which are monitored as the search progresses from the small starting subset to the final fit that corresponds to the classical statistical summary of the data.

The search for a single population starts from a subset of m_0 observations, say $S^*(m_0)$, robustly chosen. The size of the fitting subset is increased from m to $m+1$ by forming the new subset $S^*(m+1)$ from those observations with the $m+1$ smallest squared Mahalanobis distances when the parameters are estimated from $S^*(m)$. Thus, some observations in $S^*(m)$ may not be included in $S^*(m+1)$. For each m $(m_0 \leq m \leq n-1)$, the test for the presence of outliers is based on the observation outside the subset with the smallest squared Mahalanobis distance.

The parameters μ and Σ of the v-dimensional multivariate normal distribution of y are estimated in the FS by the standard unbiased estimators from a subset of m observations, providing estimates $\hat{\mu}(m)$ and $\hat{\Sigma}(m)$. Using these estimates we calculate n squared Mahalanobis distances

$$d_i^2(m) = \{y_i - \hat{\mu}(m)\}' \hat{\Sigma}^{-1}(m) \{y_i - \hat{\mu}(m)\}, \qquad i = 1, \ldots, n. \tag{A.1}$$

To detect outliers we use the minimum Mahalanobis distance amongst observations not in the subset

$$d_{\min}(m) = \min d_i(m) \quad i \notin S^*(m). \tag{A.2}$$

Testing for outliers requires a reference distribution for $d_i^2(m)$ in (A.1) and hence for $d_{\min}(m)$ in (A.2). When Σ is estimated from all n observations, the squared statistics have a scaled beta distribution. However, the estimate $\hat{\Sigma}(m)$ in the search uses the central m out of n observations, so that the variability is underestimated. Results of [43] on truncated distributions provide a consistency factor

$$c(m,n) = \frac{n}{m} C_{v+2}\{\chi^2_{v,m/n}\}, \tag{A.3}$$

where $C_r(y)$ is the c.d.f. of the χ^2 distribution on r degrees of freedom evaluated at y and $\chi^2_{r,\zeta} = C_r^{-1}(\zeta)$, for $0 < \zeta < 1$, is the ζth quantile of the same distribution. Then the scaled and asymptotically unbiased estimate of Σ is

$$\hat{\Sigma}^{\text{SC}}(m) = c(m,n)\hat{\Sigma}(m).$$

The scaled minimum Mahalanobis distance $d^{\text{SC}}_{\min}(m)$ follows from (A.2) when $\hat{\Sigma}(m)$ in (A.1) is replaced by $\hat{\Sigma}^{\text{SC}}(m)$.

Further distributional results in [22] lead to the distribution of $d_{\min}(m)$ (A.2) for a given m. As we show, it is extremely helpful to look at forward plots of quantities of interest such as $d_{\min}(m)$ during the search and to compare them with the envelopes formed by the forward plots of several quantiles. Such monitoring plots, drawn for a range of values of m, provide information about departures, if any, of the data from the assumed structure.

For precise outlier identification we perform a series of tests, one for each $m \geq m_0$. To allow for the multiple testing involved, we use a rule which depends on the sample size n to determine the relationship between the envelopes calculated for the distribution of $d_{\min}(m)$ and the significance of the observed values. But, if there are outliers, we need to judge the values of the statistics against envelopes from appropriately smaller population sizes that exclude potential outliers. To achieve this we introduce the idea of a "signal". If at some point $m^\dagger$ in the search the nearest observation to those already in the subset appears to be an outlier, as judged by an appropriate envelope of the distribution of the test statistic, we call this a signal. Appearance of a signal indicates that observation $m^\dagger$, and the remaining observations not in the subset, may be outliers. The second

stage of the analysis consists of superimposing envelopes for a series of smaller sample sizes $n^\dagger$, starting from $m^\dagger - 1$ onwards, until the first introduction of an observation recognised as an outlier. The details of the procedure are described in [22] and exemplified in the next part of this appendix.

In the procedure described so far, a single population multivariate normal model is fitted to the data belong to $S^*(m)$, the purpose being robust estimation and the detection of outliers. If, however, the outliers are clustered this structure may be determined by starting a FS near each cluster centre, when observations in other clusters are revealed as outlying. Since the clusters are unknown, [44] suggest starting the forward search with a number of randomly selected initial subsets. Once two searches converge to the same subset, they cannot diverge; as the search progresses the number of distinct trajectories reduces and peaks in the forward plots of minimum Mahalanobis distances indicate the presence of clusters, provided they are not too many.

Appendix A.2. Numerical Examples

We start the series of examples on simulated data with one in which there are no outliers. In all examples the data are 300 observations simulated from a five-dimensional normal distribution, constructed from independent standard normal observations. Figure A.21 shows the forward plot of the minimum Mahalanobis distances of the data together with the 1%, 50% 99%, 99.9%, 99.99% and 99.999% points of the null distribution of the distances calculated as in [22]. Over the range $m = 150$ to 300, all observed distances lie between the 1% and 99% limits and there is no evidence of any outlying observations. The vertical line in the plot indicates the change in outlier detection rules given by [22]; the change is necessary because of the increasingly steep shape of the envelopes towards the end of the search, when the more remote observations, with larger Mahalanobis distances, enter the subset.

Now we look at a series of similar simulations, but with an increasing number of outliers, in each case generated by adding three to the values of y in all five dimensions. In the first example with outliers, shown in Figure A.22, there are

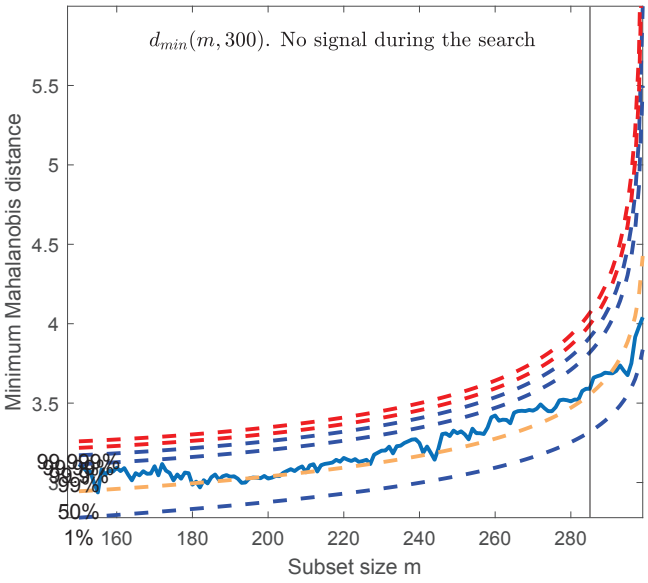

Figure A.21: 300 simulated observations, no outliers. Minimum Mahalanobis distances from the Forward Search together with 1%, 50% 99%, 99.9%, 99.99% and 99.999% pointwise envelopes. No evidence of any outliers.

six outlying observations. The forward plot of minimum Mahalanobis distances in the left-hand panel of the plot shows a signal at $m = 294$ when the plot of observed values goes outside the uppermost envelope. It stays outside thereafter. We start the resuperimposition of envelopes from $m^{\dagger} - 1 = 293$. The three panels of Figure A.23 confirm that the first outlier is identified when $n = 295$. We have thus correctly identified the six outliers which are shown in the scatterplot matrix of Figure A.24.

The process becomes less straightforward as the number of outliers increases. Figure A.25 gives the forward plot of minimum Mahalanobis distances when there are 30 outliers (10% contamination). This is a completely different plot from those we have seen before. There is a signal at $m^{\dagger} = 235$. The trajectory then rises to a sharp peak at $m = 271$ before returning to lie near the 50% quantile at the end of the search. It is clear that the presence of outliers would be completely missed if deletion methods of 2 or 3 observations were applied to

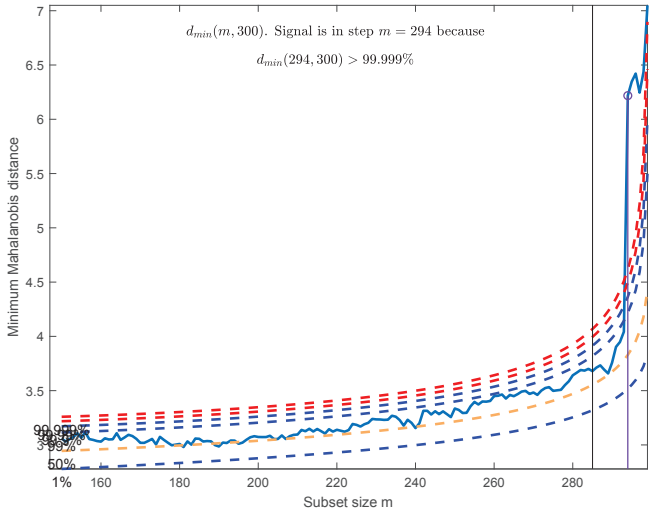

Figure A.22: 300 simulated observations, six outliers. Minimum Mahalanobis distances from the Forward Search. There is a "signal" at $m = 294$.

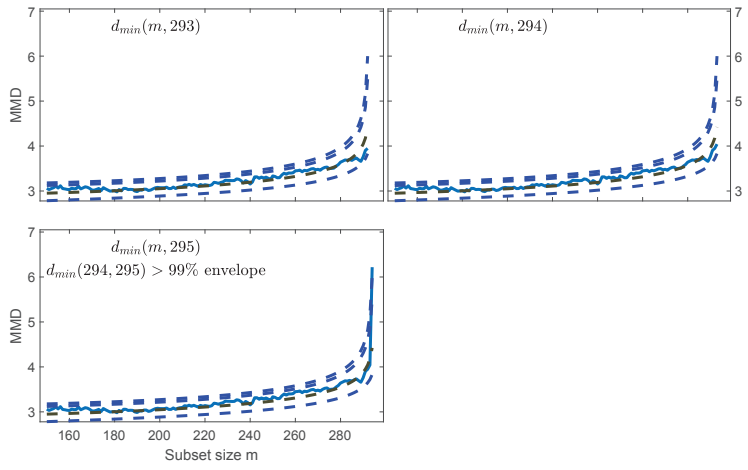

Figure A.23: 300 simulated observations, six outliers. Resuperimposition of envelopes for minimum Mahalanobis distances (MMD); there are no outliers for $m = 293$ and 294, but one is identified when $m = 295$. All six outliers are identified.

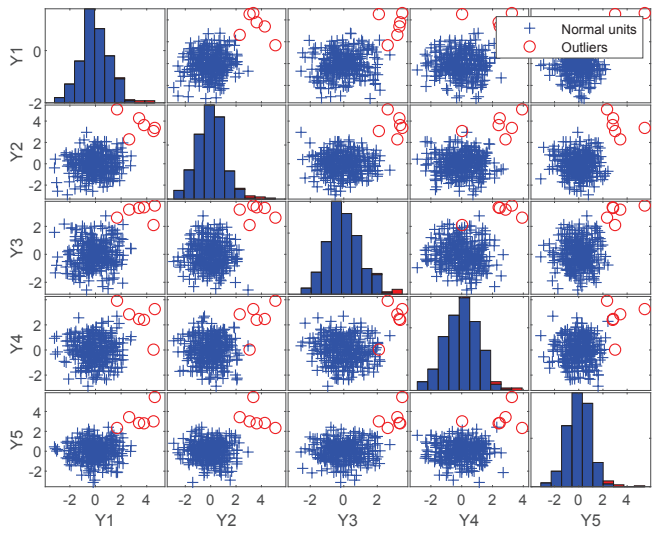

Figure A.24: 300 simulated observations, six outliers, scatterplot matrix showing all six outliers correctly identified by the Forward Search.

the fit to all the data, the phenomenon known as masking. The sharp peak is caused by the distance between the outliers and the uncontaminated data when the subset contains no outliers; as soon as a few are introduced, the parameter estimates change and the remaining outliers seem less remote. The trajectory up to the peak illustrates the necessity of a signal. It is similar in shape to the calculated quantiles in the plot, but for a smaller value of n; resuperimposition allows comparison of the trajectory with envelopes from a series of sample sizes. We start in Figure A.26 with $m = 234$. Here the trajectory lies below the envelopes. Similar patterns persist until $m = 270$, when the trajectory is well within the central band. As the remaining two panels of the figure show, there is no evidence of an outlier for $m = 271$, but there is for $m = 272$. Thus we have found 29 out of the 30 outliers and obtained efficient estimates of the parameters.

Figure A.27 gives a scatterplot matrix of the 29 outlying observations. As would be excepted from the way in which these are simulated, they form a cluster. We conclude this section with an example of the use of the random

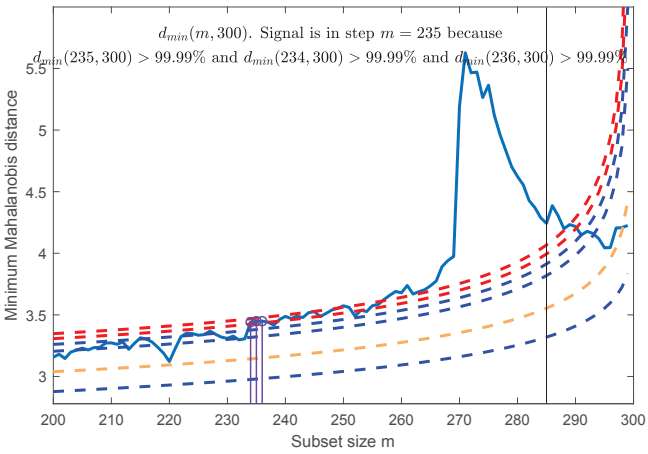

Figure A.25: 300 simulated observations, 10% outliers. Minimum Mahalanobis distances from the Forward Search. There is a "signal" at $m = 235$.

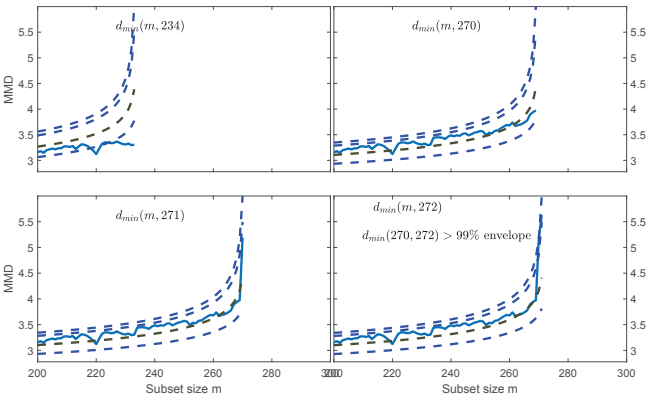

Figure A.26: 300 simulated observations, 10% outliers. Resuperimposition of envelopes for minimum Mahalanobis distances (MMD). There are no outliers for $m = 234$ and 270 and 271, but one is identified when $m = 272$; 29 outliers are identified and efficient parameter estimates obtained.

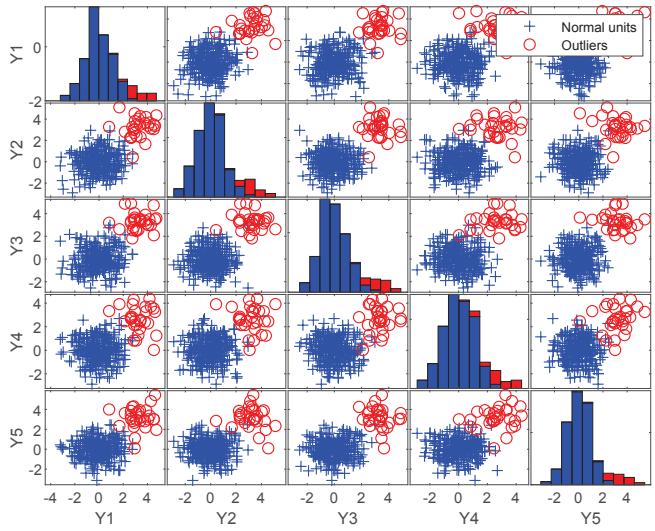

Figure A.27: 300 simulated observations, 10% outliers, scatterplot matrix showing the 29 identified outliers.

start forward search in detecting a cluster structure.

We continue with the same data structure but now with 135 outliers (45%). The automatic procedure for outlier detection based on a signal and resuperimposition of envelopes identifies 131 outliers, but does not offer any indication of their structure. However, the forward plot of minimum Mahalanobis distances from 200 random starts in Figure A.28 shows two clear peaks, one around $m = 135$ and the other around $m = 170$. The plot thus indicates the presence of two clusters, rather than of a single population with many unstructured outliers. As Figure 9 shows, this procedure is sometimes particularly useful for identifying clustering structure when there are more than two clusters which cannot be easily identified from scatterplot matrices. Here the two clusters become evident on fitting a single multivariate model to the data.

[1] M. Hubert, P. J. Rousseeuw, S. van Aelst, High-breakdown robust multivariate methods, Statistical Science 23 (2008) 92–119.

[2] M. Hubert, M. Debruyne, P. J. Rousseeuw, Minimum covariance

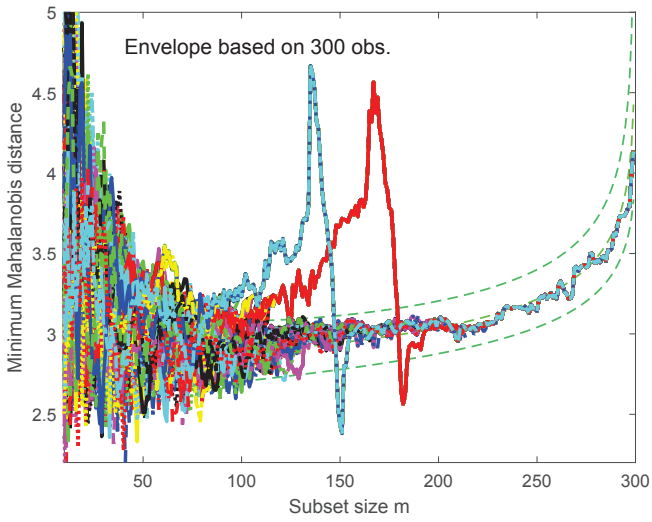

Figure A.28: 300 simulated observations, 45% outliers. Minimum Mahalanobis distances from 200 random start Forward Searches indicating the presence of two clusters.

determinant and extensions, WIREs Computational Statistics (2017) https://doi.org/10.1002/wics.1421.

[3] R. A. Maronna, R. D. Martin, V. J. Yohai, Robust Statistics: Theory and Methods, Wiley, Chichester, 2006.

[4] A. Cerioli, A. Farcomeni, M. Riani, Strong consistency and robustness of the Forward Search estimator of multivariate location and scatter, Journal of Multivariate Analysis 126 (2014) 167–183.

[5] A. Cerioli, A. Farcomeni, M. Riani, Wild adaptive trimming for robust estimation and cluster analysis, Scandinavian Journal of Statistics (2018). https: //onlinelibrary.wiley.com/doi/abs/10.1111/sjos.12349.

[6] L. A. García-Escudero, A. Gordaliza, C. Matrán, A. Mayo-Iscar, A review of robust clustering methods, Advances in Data Analysis and Classification 4 (2010) 89–109.

[7] M. Gallaugher, P. McNicholas, Finite mixtures of skewed matrix variate distributions, Pattern Recognition 76 (2018) 491–505.

[8] J. Myhre, K. Mikalsen, S. Løkse, R. Jenssen, Robust clustering using a kNN mode seeking ensemble, Pattern Recognition 80 (2018) 83–93.

[9] B. Chen, K. Ting, T. Washio, Y. Zhu, Local contrast as an effective means to robust clustering against varying densities, Machine Learning (2018) https://doi.org/10.1007/s10994–017–5693–x.

[10] C. Croux, G. Haesbroeck, Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies, Biometrika 87 (2000) 603–618.

[11] M. Hubert, P. Rousseeuw, K. Vanden Branden, ROBPCA: A new approach to robust principal component analysis, Technometrics 47 (2005) 64–79.

[12] J. Oh, N. Kwak, Generalized mean for robust principal component analysis, Pattern Recognition 54 (2016) 116–127.

[13] B. Chen, K. Ting, T. Washio, G. Haffari, Half-space mass: a maximally robust and efficient data depth method, Machine Learning 100 (2015) 697–699.

[14] R. Grbić, D. Grahovac, R. Scitovski, A method for solving the multiple ellipses detection problem, Pattern Recognition 60 (2016) 824–834.

[15] F. Mandanas, C. Kotropoulos, M-estimators for robust multidimensional scaling employing $\ell_{2,1}$ norm regularization, Pattern Recognition 73 (2018) 235–246.

[16] N. Vretos, A. Tefas, I. Pitas, Using robust dispersion estimation in support vector machines, Pattern Recognition 46 (2013) 3441–3451.

[17] A. Christmann, D.-X. Zhou, On the robustness of regularized pairwise learning methods based on kernels, Journal of Complexity 37 (2016) 1–33.

[18] A. Nurunnabi, Y. Sadahiro, D. Laefer, Robust statistical approaches for circle fitting in laser scanning three-dimensional point cloud data, Pattern Recognition 81 (2018) 417–431.

[19] H. Paulheim, R. Meusel, A decomposition of the outlier detection problem into a set of supervised learning problems, Machine Learning 100 (2015) 509–531.

[20] C. Aggarwal, Outlier Analysis. 2nd Edition, Springer International, Cham, 2017.

[21] R. Domingues, M. Filippone, P. Michiardi, J. Zouaoui, A comparative evaluation of outlier detection algorithms: Experiments and analyses, Pattern Recognition 74 (2018) 406–421.

[22] M. Riani, A. C. Atkinson, A. Cerioli, Finding an unknown number of multivariate outliers, Journal of the Royal Statistical Society, Series B 71 (2009) 447–466.

[23] A. Cerioli, Multivariate outlier detection with high-breakdown estimators, Journal of the American Statistical Association 105 (2010) 147–156.

[24] A. Cerioli, A. Farcomeni, Error rates for multivariate outlier detection, Computational Statistics and Data Analysis 55 (2011) 544–553.

[25] P. Filzmoser, V. Todorov, Robust tools for the imperfect world, Information Sciences (NY) 245 (2013) 4–20.

[26] P. Rousseeuw, W. Van Den Bossche, Detecting deviating data cells, Technometrics 60 (2018) 135–145.

[27] M. Riani, A. Cerioli, A. C. Atkinson, D. Perrotta, Monitoring robust regression, Electronic Journal of Statistics 8 (2014) 642–673.

[28] A. Cerioli, A. C. Atkinson, M. Riani, How to marry robustness and applied statistics, in: T. Di Battista, E. Moreno, W. Racugno (Eds.), Topics on

Methodological and Applied Statistical Inference, Springer International, Cham, Switzerland, 2016, pp. 51–65.

[29] D. F. Andrews, P. J. Bickel, F. R. Hampel, W. J. Tukey, P. J. Huber, Robust Estimates of Location: Survey and Advances., Princeton University Press, Princeton, NJ, 1972.

[30] A. Cerioli, M. Riani, A. C. Atkinson, A. Corbellini, The power of monitoring: How to make the most of a contaminated multivariate sample (with discussion), Statistical Methods and Applications (2017) https://doi.org/10.1007/s10260–017–0409–8.

[31] L. A. García-Escudero, A. Gordaliza, C. Matrán, A. Mayo-Iscar, A general trimming approach to robust cluster analysis, Annals of Statistics 36 (2008) 1324–1345.

[32] F. Dotto, A. Farcomeni, L. A. García-Escudero, A. Mayo-Iscar, A reweighting approach to robust clustering, Statistics and Computing 27 (2017) 1–17.

[33] P. J. Huber, E. M. Ronchetti, Robust Statistics, Second Edition, Wiley, New York, 2009.

[34] P. J. Rousseeuw, B. C. van Zomeren, Unmasking multivariate outliers and leverage points, Journal of the American Statistical Association 85 (1990) 633–9.

[35] A. C. Atkinson, M. Riani, A. Cerioli, The forward search: theory and data analysis (with discussion), Journal of the Korean Statistical Society 39 (2010) 117–134, doi:10.1016/j.jkss.2010.02.007.

[36] M. Riani, D. Perrotta, A. Cerioli, The forward search for very large datasets, Journal of Statistical Software 67 (1) (2015) 1–20.

[37] W. N. Venables, B. D. Ripley, Modern Applied Statistics with S (4th Edition), Springer-Verlag, New York, 2002.

[38] A. C. Atkinson, M. Riani, A. Cerioli, Exploring Multivariate Data with the Forward Search, Springer–Verlag, New York, 2004.

[39] A. Cerioli, L. A. García-Escudero, A. Mayo-Iscar, M. Riani, Finding the number of normal groups in model-based clustering via constrained likelihoods, Journal of Computational and Graphical Statistics 27 (2018) 404–416.

[40] W. M. Rand, Objective criteria for the evaluation of clustering methods, Journal of the American Statistical Association 66 (1971) 846–850.

[41] C. Fraley, A. E. Raftery, How many clusters? Which clustering method? – Answers via model-based cluster analysis, Computer Journal 41 (1998) 578–588.

[42] G. Schwarz, Estimating the dimension of a model, Annals of Statistics 6 (1978) 461–464.

[43] G. M. Tallis, Elliptical and radial truncation in normal samples, Annals of Mathematical Statistics 34 (1963) 940–944.

[44] A. C. Atkinson, M. Riani, A. Cerioli, Cluster detection and clustering with random start forward searches, Journal of Applied Statistics 45 (2018) 777–798, doi http://dx.doi.org/10.1080/02664763.2017.1310806.

[45] C. Agostinelli, L. Greco, Discussion of "The power of monitoring: how to make the most of a contaminated multivariate sample" by Andrea Cerioli, Marco Riani, Anthony C. Atkinson and Aldo Corbellini, Statistical Methods and Applications (2017). https://doi.org/10.1007/s10260–017–0416–9.

[46] A. Farcomeni, F. Dotto, The power of (extended) monitoring in robust clustering. Discussion of "The power of monitoring: how to make the most of a contaminated multivariate sample", Statistical Methods and Applications (2017). https://doi.org/10.1007/s10260–017–0417–8.

[47] L. A. García-Escudero, A. Gordaliza, C. Matrán, A. Mayo-Iscar, Comments on "The power of monitoring: How to make the most of a contaminated multivariate sample", Statistical Methods and Applications (2017). https://doi.org/10.1007/s10260–017–0415–x.

[48] G. E. P. Box, D. R. Cox, An analysis of transformations (with discussion), Journal of the Royal Statistical Society, Series B 26 (1964) 211–246.