The Forward Search and Data Visualisation

Anthony C. Atkinson* and Marco Riani†
The London School of Economics, London WC2A 2AE, UK and
Sezione di Statistica e Informatica, Dipartimento di Economia,
Università di Parma, Italy

Summary

A statistical analysis using the forward search produces many graphs. For multivariate data an appreciable proportion of these are a variety of plots of the Mahalanobis distances of the individual observations during the search. Each unit, originally a point in v-dimensional space, is then represented by a curve in two dimensions connecting the almost n values of the distance for each unit calculated during the search. Our task is now to recognise and classify these curves: we may find several clusters of data, or outliers or some unexpected, non-normal, structure. We look at the plots from five data sets. Statistical techniques include cluster analysis and transformations to multivariate normality.

Keywords: clustering; forward plot; graphics for multivariate data; Mahalanobis distance; outliers; power transformation; robustness; very robust methods.

1 Introduction

The forward search is a powerful robust statistical method for exploring the relationship between data and fitted models. It relies on the interpretation of a large number of graphs. Atkinson and Riani (2000) describe its use in linear and nonlinear regression, response transformation and in generalized linear models, where the emphasis is on the detection of unidentified subsets of the data and of multiple masked outliers and of their effect on inferences. In this paper we extend the method to the analysis of multivariate data. Our emphasis here is rather more on the data and less on the multivariate normal model.

The forward search orders the observations by closeness to the assumed model, starting from a small subset of the data and increasing the number of observations m used for fitting the model. Outliers and small unidentified subsets of observations enter at the end of the search. Even if there are a number of groups, as in cluster analysis, we start by fitting one multivariate normal distribution to the data. An important graphical tool is a variety of

*e-mail: a.c.atkinson@lse.ac.uk
†e-mail: mriani@unipr.it

plots of the Mahalanobis distances of the individual observations during the search. Each unit, originally a point in v-dimensional space, is then represented by a curve in two dimensions connecting the almost n values of the distance for that unit calculated during the search. Our task is now to classify these curves.

Section 2 defines the search. Thereafter we proceed by examples, all of which are of multivariate data. The first is of measurements on Swiss heads which we use to introduce forward plots of Mahalanobis distances. These plots indicate some outliers in the milk data analysed in §4. We then use a synthetic example in §5 to calibrate our plots, showing the effect of clusters. Forward plots of individual Mahalanobis distances and of the trace of the estimated covariance matrix are also helpful in revealing unsuspected structure. These ideas are illustrated in §6 where more structure is discovered in the milk data. Sections 7 and 8 are both concerned with data with several groups. In §8 we outline the use of the forward search in cluster analysis.

Because we use Mahalanobis distances, it is important that the data are approximately normal. We therefore introduce in §9 a multivariate form of the Box-Cox family of transformations. We use forward plots of tests for transformation in §10 to emphasize the importance of the forward search in detecting the effect of just a few observations on inferences. We conclude in §11 with some comments on computation and on extensions of our graphical presentation to larger data sets.

2 The Forward Search

The main diagnostic tools that we use are various plots of Mahalanobis distances. The squared distances for the sample are defined as

$$d_i^2 = \{y_i - \hat{\mu}\}^T \hat{\Sigma}^{-1} \{y_i - \hat{\mu}\}, \quad (i = 1, \ldots, n), \tag{1}$$

where $\hat{\mu}$ is the vector of means of the n observations and $\hat{\Sigma}$ is the unbiased estimator of the population covariance matrix.

In the forward search the parameters μ and Σ are estimated from a subset of m observations to give the estimates $\hat{\mu}(m)$ and $\hat{\Sigma}(m)$. From this subset we obtain n squared Mahalanobis distances

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

We often start the search with a small subset of m_0 observations, chosen from the robust bivariate boxplots of Zani, Riani, and Corbellini (1998) to exclude outlying observations in any two-dimensional plot of the data. The content of the contours is adjusted to give an initial subset of the required size. The search is not sensitive to the exact choice of this subset.

When m observations are used in fitting, the optimum subset $S^*(m)$ yields n squared distances $d_i^2(m^*)$. We order these squared distances and take the observations corresponding to the $m+1$ smallest as the new subset $S^*(m+1)$. Usually this process augments the subset by only one observation, but sometimes two or more observations enter as one or more leave. Due to the form of the search, outliers, if any, tend to enter as m approaches n.

In our examples we look at forward plots of the distances $d_i(m^*)$. These distances tend to decrease as n increases. If interest is in the latter part of the

search we may also look at **scaled** distances

$$d_i(m^*) \times \left\{ |\hat{\Sigma}(m^*)|/|\hat{\Sigma}(n)| \right\}^{1/2v}, \tag{3}$$

where v is the dimension of the observations y and $\hat{\Sigma}(n)$ is the estimate of Σ at the end of the search.

Such a rescaling increases emphasis on the later parts of forward plots. Thus, for detecting initial clusters, the original distances may be better, whereas for a confirmatory analysis, where we are interested in the possible presence of outliers or undetected small clusters, we might prefer scaled distances.

If there are clusters, the confirmatory part of our analysis includes a forward search fitting an individual model for each cluster. For all unassigned observations we calculate the distance from each cluster centre. Observations are included in the subset of the cluster to which they are nearest and the distances to all cluster centres are monitored.

Unfortunately we may be comparing Mahalanobis distances for compact and dispersed clusters. A forward allocation can then lead to a dispersed group "invading" a compact group with a wrong allocation of several observations. We therefore introduce standardized distances, adjusted for the variance of the individual cluster. The customary squared Mahalanobis distance for the ith observation from the gth group at step m is

$$d^2_{gi}(m) = \{y_{gi} - \hat{\mu}_g(m)\}^T \hat{\Sigma}_g^{-1}(m)\{y_{gi} - \hat{\mu}_g(m)\}, \tag{4}$$

where μ and Σ are estimated for each group. The **standardized** distance is

$$d_{gi}(m) = d_{gi}(m)|\hat{\Sigma}_{gm}|^{1/2v}, \tag{5}$$

in which the effect of differing variances between groups has been completely eliminated. These distances will produce measures close to Euclidean distance, with the consequence that compact clusters may tend to "invade" dispersed ones, the reverse of the behaviour with the usual Mahalanobis distance. We generally need to look at plots of both distances.

3 Swiss Heads

As a first example of the use of forward plots we start with data which seem to have a single multivariate normal distribution. The data, given by Flury and Riedwyl (1988, p. 218), are six readings on the dimensions of the heads of 200 twenty year old Swiss soldiers. The forward plot of scaled distances in Figure 1 shows little structure. The rising diagonal white band separates those units which are in the subset from those that are not. At the end of the search there seem to be two outliers, observations 104 and 111.

Figure 2 repeats Figure 1 with the distances for units 104 and 111 highlighted. These are the largest distances at the end of the search, even though they decrease in the final two steps, when the units concerned join the subset. This is an example of masking, which is so slight as not to be misleading. Although the distances for the two units are the largest in the last thirty steps of the search, they rank much lower on size earlier on. This atypical behaviour

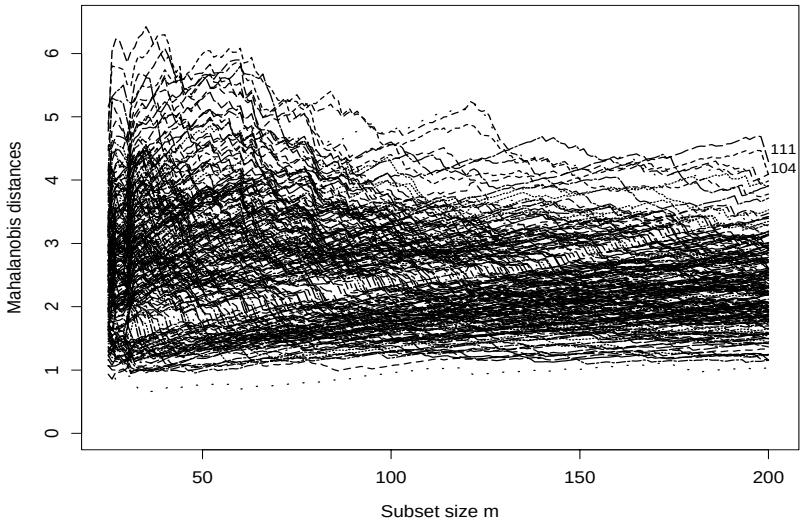

Figure 1: Swiss heads: forward plot of scaled Mahalanobis distances. Units 104 and 111 are outlying towards the end of the search

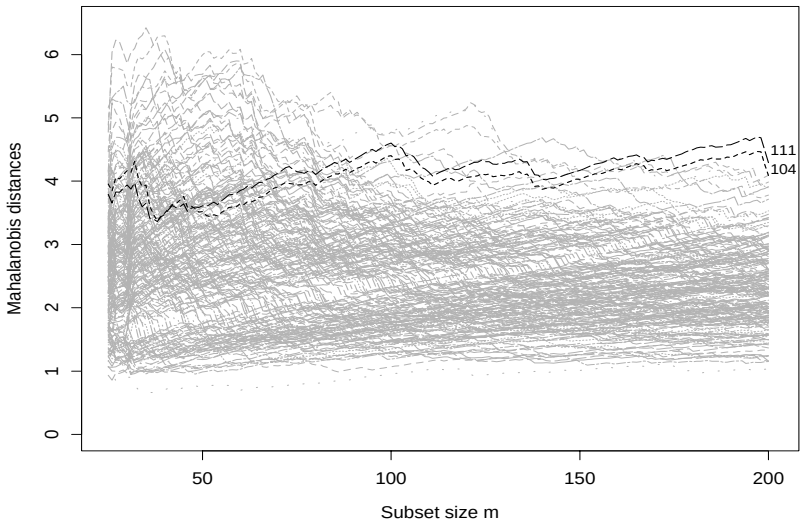

Figure 2: Swiss heads: forward plot of scaled Mahalanobis distances. The trajectories for units 104 and 111 are highlighted

during the search is a useful way of detecting outliers or units that have been classified into an incorrect group.

Figure 3 is a scatterplot matrix of the observations. As in Figure 2, we highlight units 104 and 111. It is clear that the search has identified two units with the largest values of y_4. What is not clear is whether these units affect any inferences drawn from the data. We return to this in §10 where we monitor some other quantities during the search.

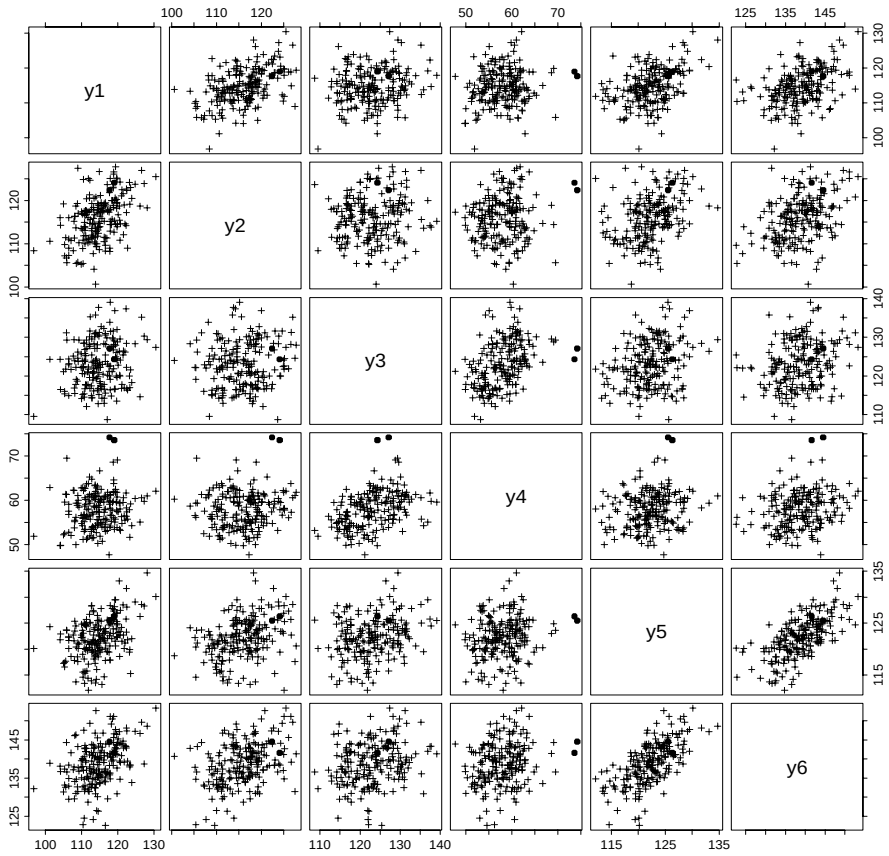

Figure 3: Swiss heads: scatterplot matrix of the six measurements on 200 heads. Units 104 and 111 are plotted as dots

4 Milk Data

We now consider a slightly more complicated example, to which we shall again return for a fuller analysis.

Daudin, Duby, and Trecourt (1988) give data on the composition of 85 containers of milk, on each of which eight measurements were made. A scatterplot matrix of the data is in Figure 4. The panel for y_5 and y_6 shows clearly that one unit is remote in this bivariate projection. Otherwise, several panels show

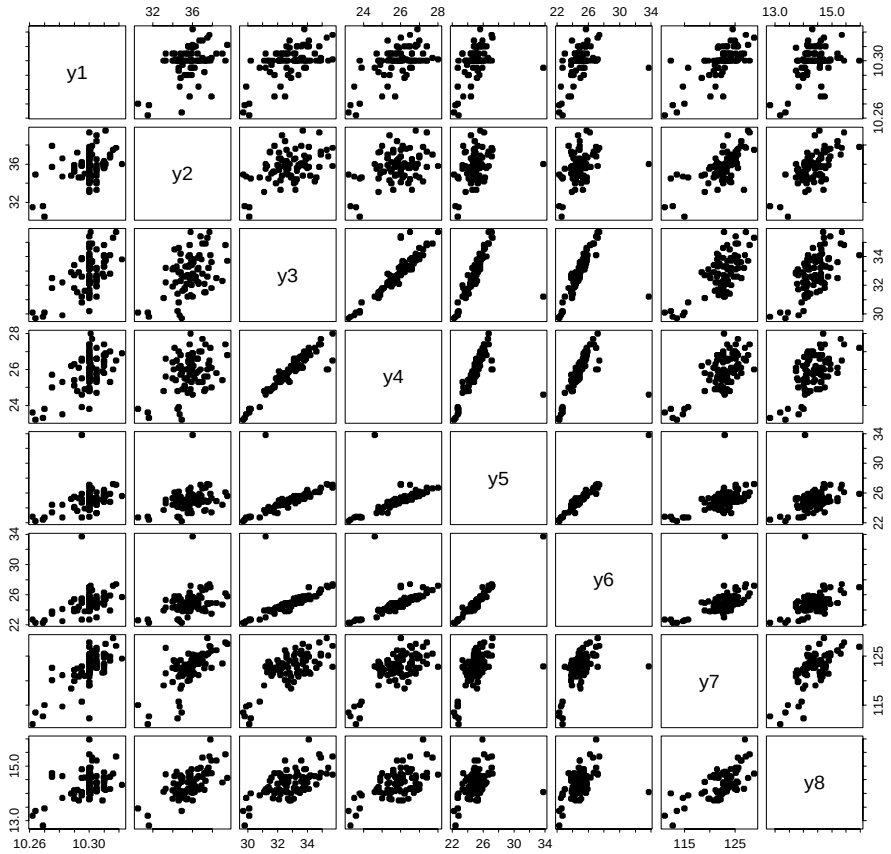

Figure 4: Milk data: scatterplot matrix. There is a strong diagonal structure in some panels

a strong rising diagonal structure. There is one gross outlier in Figure 5, the forward plot of scaled Mahalanobis distances. Until $m = n$ the distance for unit 69 is off the top of the plot. Unit 44 becomes less outlying as the search progresses and is the first of a group of four moderately outlying units to enter the subset. The others, units 1, 2 and 41 are clearly outlying until $m = 81$ when unit 44 joins. The distances then decrease until there is some masking at the end of the search, with unit 77 having the third largest distance.

We now look in more detail at the data. The two scatterplots of Figure 6 are details of Figure 4 with the five points highlighted. Unit 69, at the top of both plots, is the last to enter and is the clear outlier already mentioned. The group of four units, 1, 2, 41 and 44, are particularly evident in the right-hand panel of the figure. What also seems apparent in this figure is a separated group of seven observations in the lower left-hand corner of the plot of y_6 against y_4. This group has no obvious effect on the forward plot of distances.

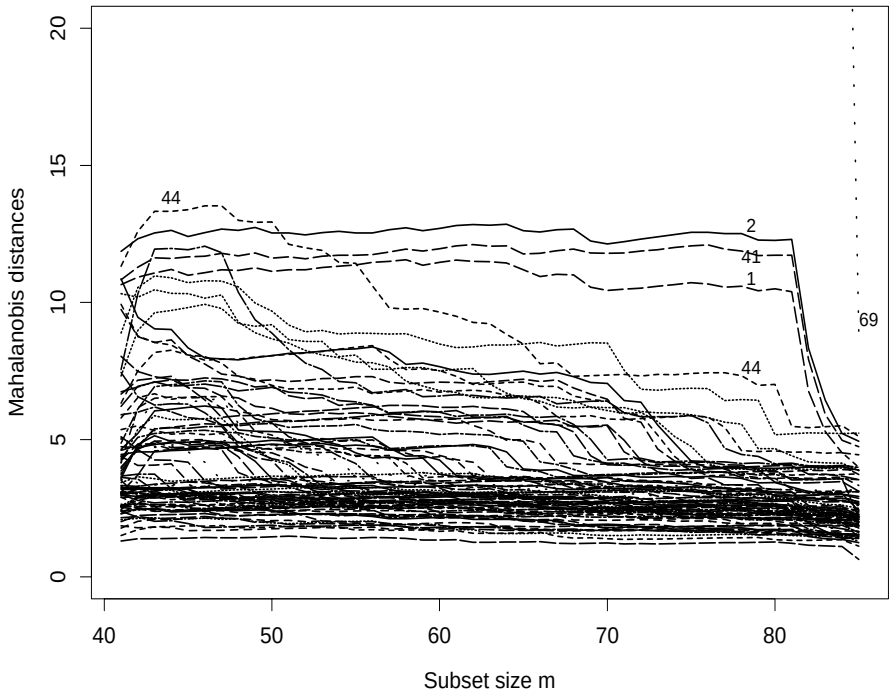

Figure 5: Milk data: forward plot of scaled Mahalanobis distances. The trace for unit 69 is off the plot until the last step

5 Three Clusters, Two Outliers

We now consider the analysis of a simulated data set, partly in order to train our eyes to the interpretation of further plots of the type we have already seen. There are three new features of this example and its analysis - one is that the data contain more than one cluster, although we initially fit a single multivariate normal distribution. The second is that we can choose our starting point to be in each cluster in turn and the third is that we introduce several powerful new plots.

There are only two variables. The data contain three clusters and two outliers. The sizes of the groups are 60 in the tighter cluster, 80 in the more dispersed cluster, 18 and 2. Figure 7 shows the data. Since there are only two variables, the structure of the data is apparent from the scatterplot. We now see how it is reflected in the plots from the forward search. We begin the forward search with $m_0 = 28$, the observations being chosen by the method of robust boxplots with elliptical contours. The starting ellipse has mostly chosen observations from the group of 60. The few observations from the diffuse group are eliminated at the first forward step and the search then continues solely with observations from the compact group until $m = 61$ when observations from the diffuse group enter. Just before all the diffuse observations are included, the two outliers, observations 159 and 160 enter the subset, although they are soon rejected, rejoining again at the very end of the search.

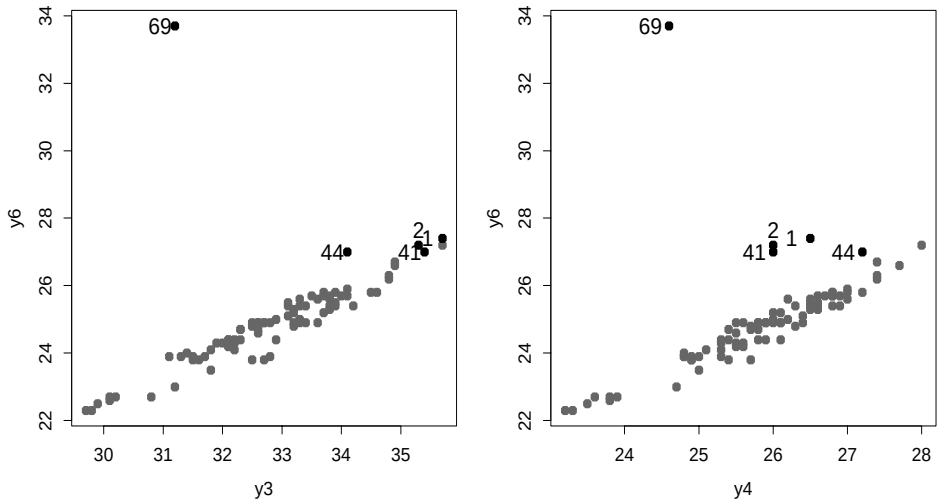

Figure 6: Milk data: scatterplots of y_6 against y_3 and y_6 against y_4. The outlying units in Figure 5 are labelled

Fig. 8 is a forward plot of the scaled Mahalanobis distances, in which we have used different line types for the three clusters and the two outliers. This plot indicates all the structure in the data, which is particularly clear up to $m = 60$: the second group is clearly separated from the first, the distances of the compact group of 18 are evident at the top of the plot and the two outliers follow an independent path. Once the second group starts to enter at $m = 61$, the smaller distances are not so readily interpreted, although the group of 18 remains distinct. The two outliers re-emerge at the very end.

We do not have to plot all the Mahalanobis distances. Figure 9 summarises the distribution of distances in Fig. 8. For each of the three clusters the figure shows the 5%, median and 95% points of the distribution of distances at each m. Group 4 consists of the two outliers. The panels clearly show that the trajectories of the distances in the different groups are distinct. Figure 10 is the forward plot of the minimum Mahalanobis distances amongst units not in the subset. This distance will be large for a single outlier. If there are several outliers the distance will be large just before the first one enters the subset. Once it has entered, the estimates of the parameters will change and the distance to the remaining outliers may be less. This is true of a cluster of outliers and is true of the sharp spike at $m = 60$ which clearly shows the end of the first cluster. The second spike at $m = 142$ is a rather less clear indication of the second group, although the spike at the end of the plot clearly shows the presence of, now, two outliers.

A third new plot, shown in Figure 11, is the forward plot of the scaled trace of the estimated covariance matrix. The scaling gives a value of one to the trace at the end of the search. This is small up to $m = 60$, corresponding to fitting

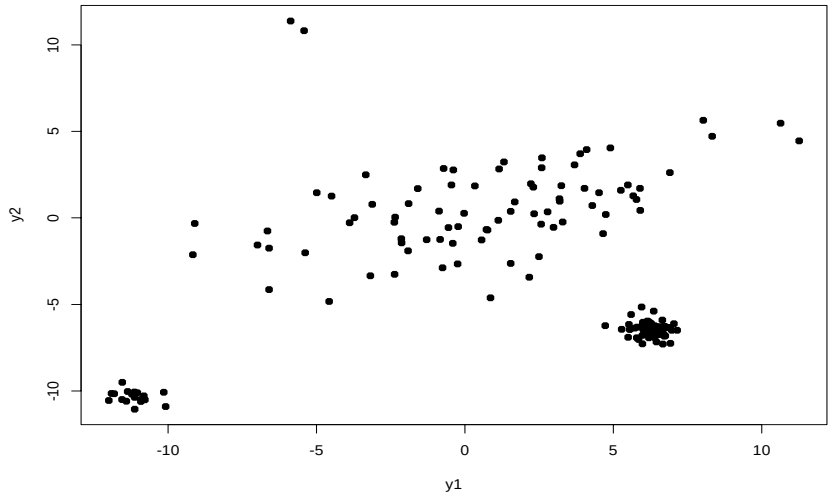

Figure 7: Three Clusters, Two Outliers: scatterplot of the data

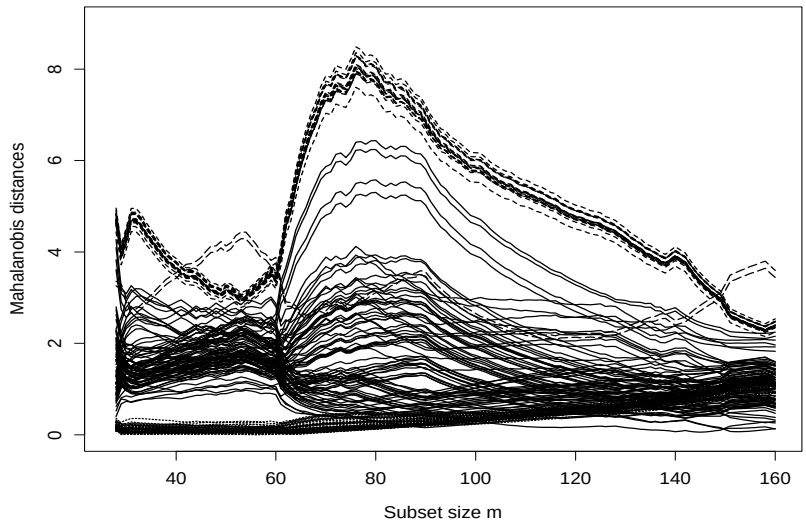

Figure 8: Three Clusters, Two Outliers: forward plot of scaled Mahalanobis distances, with line type indicating membership of the clusters and the outliers. The initial subset, found by the method of robust ellipses, consists mostly of units in the compact group

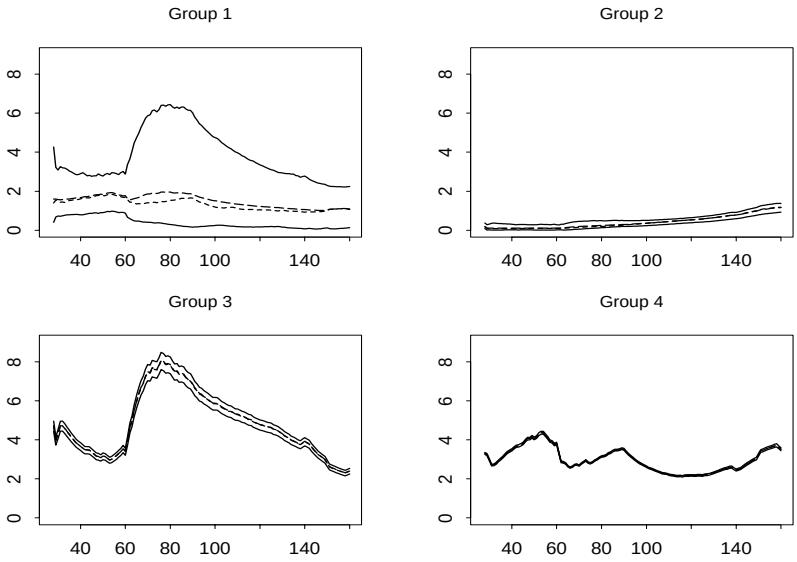

Figure 9: Three Clusters, Two Outliers: forward plot of summary of scaled Mahalanobis distances for the groups in Fig. 9; 5%, median and 95% points of the distribution of distances for each m. Group 4 consists of the two outliers

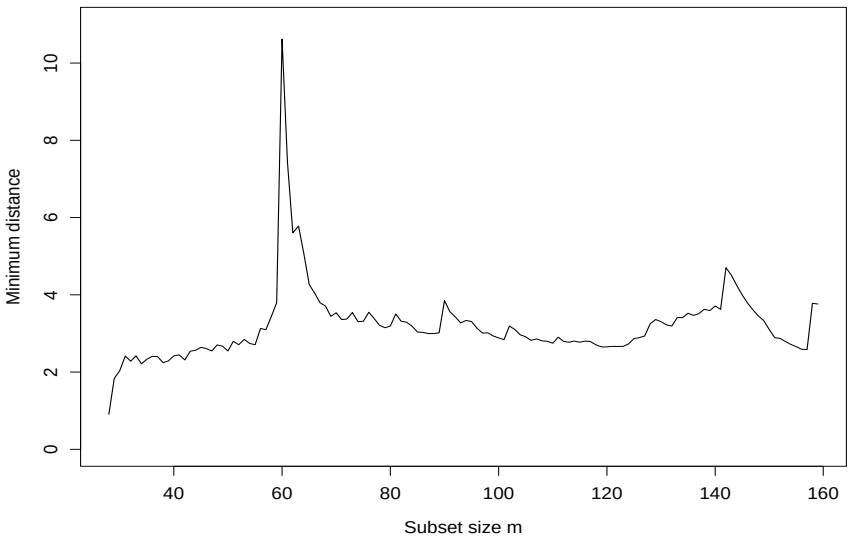

Figure 10: Three Clusters, Two Outliers: forward plot of the minimum Mahalanobis distances amongst units not in the subset: again $m_0 = 28$, found by the method of robust ellipses

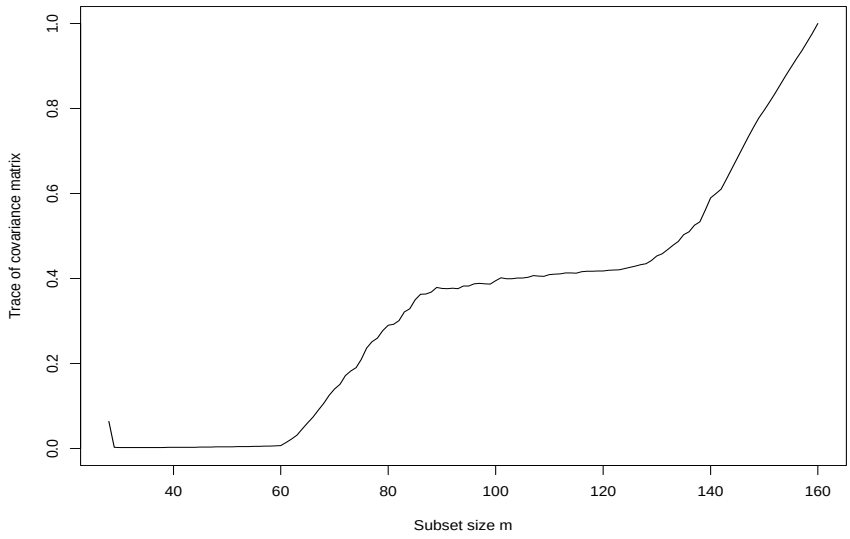

Figure 11: Three Clusters, Two Outliers: forward plot of the scaled trace of the covariance matrix for the search also shown in Figures 8 and 10

units in the tight cluster. It then increases steadily up to $m = 80$. At this point units from both groups are in the subset, with the mean of the observations in between the two groups. The value of the trace increases only gradually until the third group and outliers become important after $m = 140$.

To illustrate the effect of the starting point of the search we now begin with 20 observations from the cluster of 80 indicated in Figure 8. The forward plot of scaled Mahalanobis distances is in Figure 12. It is amazingly informative. The first part of the plot, up to $m = 80$ seems to show two groups and two outliers. Something is clearly wrong with fitting a single model around $m = 100$: there are perhaps two clusters, of different sizes than before. The transition between these two parts of the plot is fascinating. Above $m = 80$ the seemingly larger cluster splits into two parts. One part becomes the group of 18 outliers in the latter part of the plot. The other becomes the group of 60 with the smallest distances in the latter part of the plot. Thus the plot reveals the three groups and two outliers, which reappear at the end of the search. This one plot shows all the structure that we have built into the data. The indication is that we may need to run more than one forward search, constraining the starting point by the information that we have obtained from earlier searches.

6 Milk Data Again

So far we have detected the effect of one gross outlier and found a further cluster of four outliers from the forward plot of Mahalanobis distances in Figure 5. Figure 13 shows the trace of the covariance matrix. There is an appreciable increase starting at $m = 71$ when observation 11 enters. The other observations entering during this rising period are 76, 15, 14, 12 and 13. This behaviour is reminiscent of that we saw in Figure 11 as a second cluster entered the subset

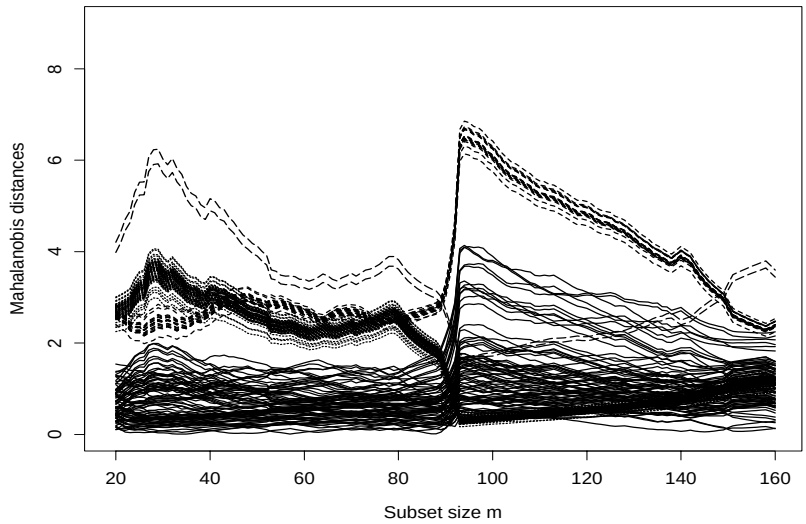

Figure 12: Three Clusters, Two Outliers: forward plot of scaled Mahalanobis distances starting in the largest group, with line type indicating membership of the clusters and the outliers, as in Figure 8. All the structure is revealed

in the example with three clusters and two outliers. The scatterplot of the data, with unit 69 removed, is in Figure 14. We have highlighted the six observations entering during this rising period. In all panels they cause an extension of the range of the data and so of the variances of the individual responses. They are six of the seven we previously noted in the lower left-hand corner of the plot of y_6 against y_4 of Figure 6. The panel for y_1 against y_7 in Figure 14 clearly separates the seventh observation from the other six. Although this group of six observations has no obvious effect on the forward plot of distances, we have been able to identify its effect on the estimated covariance matrix.

7 Swiss Bank Note Data

As a second example with at least two clusters we look at readings on six dimensions of 200 Swiss bank notes, 100 of which may be genuine and 100 forged. All notes have been withdrawn from circulation, so some of the notes in either group may have been misclassified. Also, the forged notes may not form a homogeneous group. For example, there may be more than one forger at work. The data, and a reproduction of the bank note, are given by Flury and Riedwyl (1988, p. 4-8).

Figure 15 shows the scaled Mahalanobis distances from a forward search starting with 20 observations on notes believed genuine. The traces of distances for the forgeries are shown with dotted lines. In the first part of the search, up to $m = 93$, the observations seem to fall into two groups. One has small distances and is composed of observations within or shortly to join the subset. Above these there are some outliers and then, higher still, a concentrated band of outliers, all of which are behaving similarly. The structure of this plot has

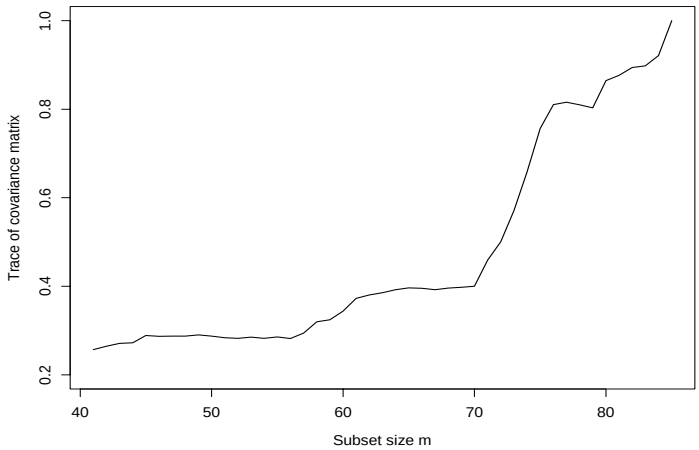

Figure 13: Milk data: forward plot of the scaled trace of the covariance matrix

much in common with that of Figure 12, although there are three clusters. Here the two groups are apparent, the forward search yielding the same 100 forgeries as Flury and Riedwyl (1988). However, the latter part of the search in Figure 15 does suggest there is something further to explain: units 1 and 40, shown by continuous lines, are outlying from Group 1 and fall among a much larger number of outlying units all from Group 2.

The structure of the group of forgeries is also readily revealed by the forward plot of scaled Mahalanobis distances just for the forgeries shown in Figure 16. In the centre of the plot, around $m = 70$ this shows a clear structure of a central group, one outlier from that group and a second group of 15 outliers. As successive units from this cluster enter after $m = 85$, they become less remote and the distances decrease.

In this example the forward search clearly indicates not only the presence of two groups of notes, but also that the group of forgeries is not homogeneous, itself consisting of two subgroups. Once we know what we are looking for, this third group can be identified on the scatterplot matrix of the observations. Figure 17 is one panel, that of y_6 against y_4, in which the three groups are most clearly seen.

8 Cluster Analysis and the Diabetes Data

In cluster analysis we seek to divide the data into homogeneous groups, or clusters, without knowing how many clusters there are. As an example of how the forward search can help we look at 145 observations on diabetes patients, which have been used in the statistical literature as a difficult example of cluster analysis. A discussion is given, for example, by Fraley and Raftery (1998). The data were introduced by Reaven and Miller (1979). There are three measurements on each patient: y_1 and y_2 are respectively plasma glucose and plasma insulin response to oral glucose; y_3 is the degree of insulin resistance. The observations were classified into three clusters by doctors, but we ignore this information in our analysis.

Figure 18 is a scatterplot matrix of the data. There seems to be a central

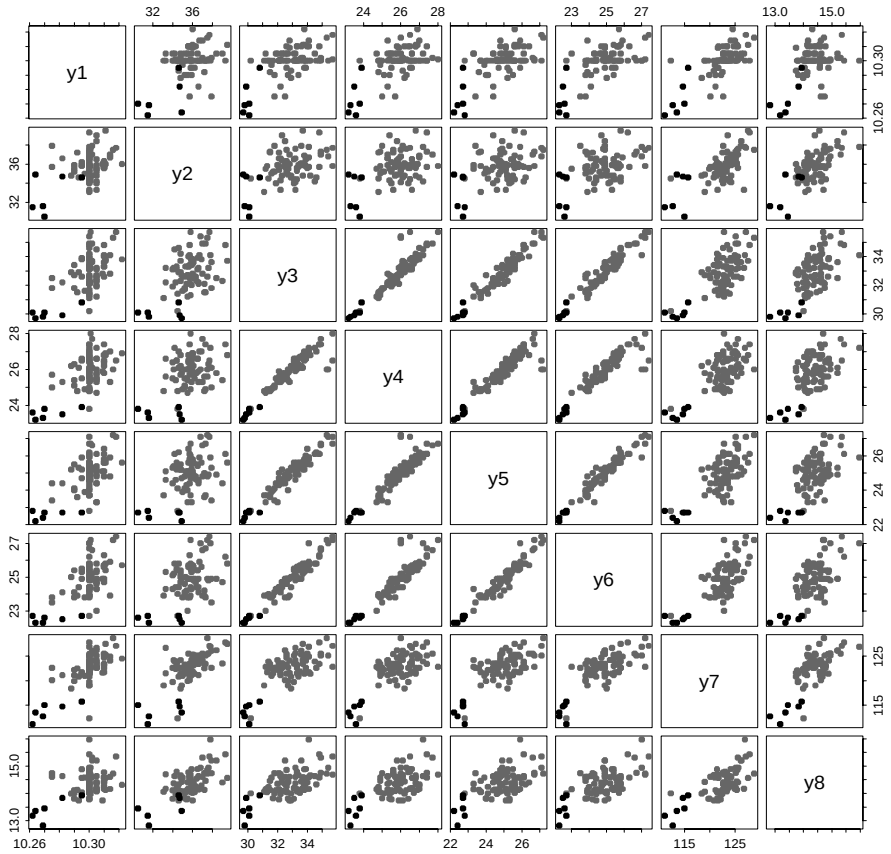

Figure 14: Milk data: scatterplot matrix. The six observations highlighted enter during the rising period from $m = 71$ in Figure 13

cluster and two 'arms' forming separate clusters. The first cluster is appreciably more compact than the other two. There would seem to be no obvious breaks between clusters, so that we can expect our plots to yield less sharp answers than those for the previous examples with two or more groups.

We proceed as before, first fitting a single distribution to all the data. The forward plot of Mahalanobis distances contains some gaps, although not as clear as those in Figures 8 and 12, which do suggest three groups. If there are groups of observations, their distances will tend to increase and decrease together. The information in such forward plots of distances can then be summarized by looking at the changes in distances. Figure 19 shows such changes in the forward plot of Mahalanobis distances, ordered by first appearance in the subset, with black representing an increase. The third group are the last to enter this forward search, and show clearly at the top of the plot. The second group is certainly different, entering immediately before the third group and so coming lower in the plot. The division between the first and second groups is not so clear - it might be anywhere between 55 and 75 on the scale of ordered units.

These preliminary groups can be refined by further forward searches, for example starting with units believed to be correctly classified. Group 2 is rather heterogeneous, being formed from units we did not assign to Groups 1 or 3 in

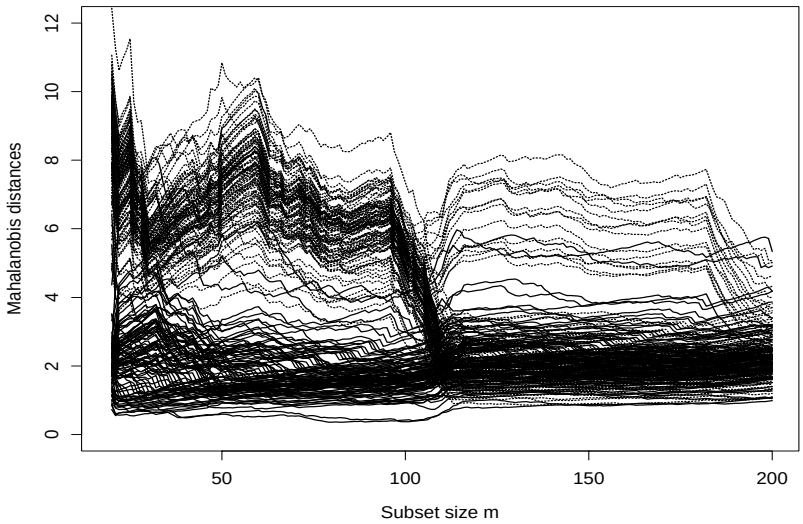

Figure 15: Swiss bank note data, forward plot of scaled Mahalanobis distances for both groups: the dotted lines are for the forgeries

the preliminary analysis. Despite this, the forward plot of scaled Mahalanobis distances in Figure 20 does show a common structure for all units. Initially there is much activity as the units from Group 2 are introduced. Then Group 1 enters, providing a period of stable growth in nearly all distances. At the end, Group 3 enters and the behaviour is once again less homogeneous. The trajectories of five units have been highlighted on the plot. These seem rather different from the other trajectories and we return to these units in a moment.

Similar analyses of individual Mahalanobis distances from the other groups lead to a division into certainly assigned units and those about which we are less sure. We then switch to forward searches fitting three ellipsoids. Since the variances of the groups are not equal, we use standardized distances with all seemingly certainly allocated observations fitted before any of the unassigned observations are allocated. Many of these units do not change their group membership and so can be allocated with certainty.

Our final forward search is represented in Figure 21, for $m = 117$ to 145. As a result of our earlier analyses, 116 units were considered classified with certainty. The bottom three lines of the figure are the key and the first column our provisional classification. The lower section of the figure above the key shows the behaviour of units we considered well classified, but which were subject to reclassification during the search. Using standardized distances, units are attracted to less dispersed groups. The figure shows several units from Group 3 attracted to group 2, which has a smaller dispersion, although larger than that of Group 1. An example is Unit 131, highlighted in Figure 20, which remained throughout in Group 2, although our other analyses indicated membership of Group 3. The top part of the figure shows units about whose membership we are uncertain. Three of the other units highlighted in Figure 20 are classified in Group 1. Despite all our analyses, we are unable to decide about the first five

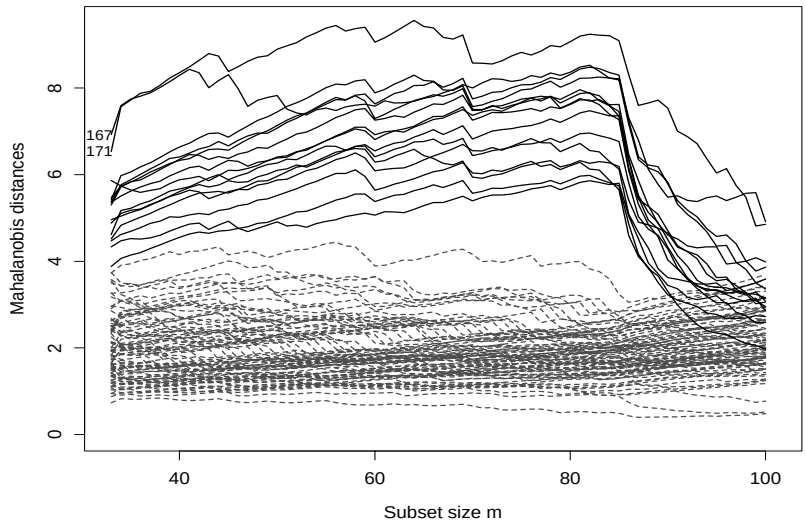

Figure 16: Swiss bank note data, forward plot of scaled Mahalanobis distances: the forgeries, showing evidence of a third group of 15 units plotted with continuous lines

units in the figure.

Figure 22 is the scatterplot matrix of the final allocation from the forward search, with uncertain units highlighted: these units fall at the intersections of clusters. Comparison with the doctors' original allocation shows that our clustering provides groups which are more coherent and compact.

9 Finding a Transformation with the Forward Search

Transformations of the data can be used to help satisfy the assumptions of multivariate normality. We have found that extension of the Box and Cox (1964) family to multivariate responses often leads to a significant increase in normality, and so to a simplified analysis of the data. We let y_{ij} be the i-th observation on response j; $j = 1, \ldots, v$. The normalized transformation of y_{ij} is

$$\begin{aligned} z_{ij}(\lambda_j) &= \frac{y_{ij}^{\lambda_j} - 1}{\lambda_j G_j^{\lambda_j - 1}} \qquad (\lambda \neq 0) \\ &= G_j \log y_{ij} \qquad (\lambda = 0), \end{aligned}$$

where G_j is the geometric mean of the j-th variable. The values $\lambda_j = 1$, $j = 1, \ldots, v$, correspond to no transformation of any of the responses. If the transformed observations are normally distributed with mean $\mu(\lambda)$ and covari-

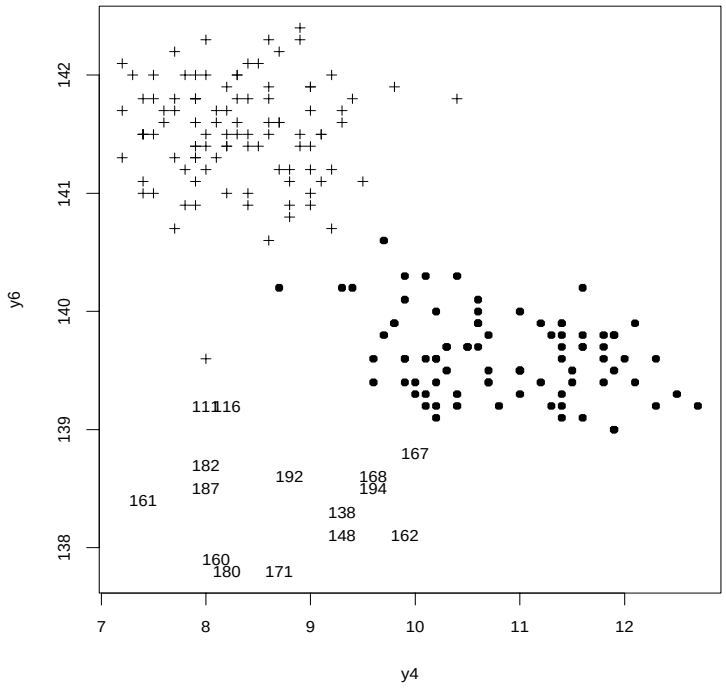

Figure 17: Swiss bank note data. Scatterplot of y_6 against y_4 which reveals most of the structure of all 200 observations: there are three groups and an outlier from Group 1, the crosses

ance matrix $\Sigma(\lambda)$, the loglikelihood of the observations is given by

$$l(\lambda) = -\frac{n}{2}\log 2\pi|\Sigma(\lambda)| - \frac{1}{2}\sum_{i=1}^{n}\{z_i - \mu(\lambda)\}^T\Sigma^{-1}(\lambda)\{z_i - \mu(\lambda)\}, \qquad (6)$$

where $z_i = (z_{i1}, \ldots, z_{iv})^T$ is the $v \times 1$ vector which denotes the transformed data for unit i. Substituting the maximum likelihood estimates $\hat{\mu}(\lambda)$ and $\hat{\Sigma}(\lambda)$ in equation (6), the maximized loglikelihood can be written as

$$l(\lambda) = \text{constant} - n/2\log|\hat{\Sigma}(\lambda)|. \qquad (7)$$

To test the hypothesis $\lambda = \lambda_0$, the likelihood ratio test

$$T_{LR} = n\log\{|\hat{\Sigma}(\lambda_0)|/|\hat{\Sigma}(\hat{\lambda})|\} \qquad (8)$$

can be compared with the χ^2 distribution on v degrees of freedom. In equation (8) the maximum likelihood estimate $\hat{\lambda}$ is found by numerical search.

Riani and Atkinson (2000) use the forward search to find satisfactory transformations for a single variable, if such exist, and the observations that are influential in their choice. For multivariate transformations Riani and Atkinson

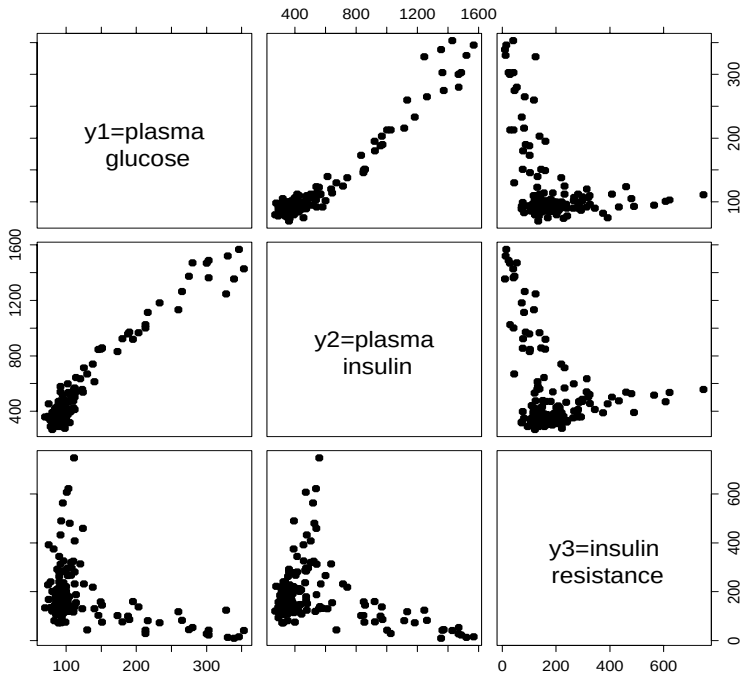

Figure 18: Diabetes data: scatterplot matrix. There may be three overlapping clusters

(2001) monitor a sequence of forward plots of the square root of the likelihood ratio statistic (8) to obtain transformations which described most of the data, with the outliers entering at the end of the searches. Results on the distribution of the test statistic for univariate transformations in regression are in Atkinson and Riani (2002). These show that asymptotic results are a good guide to significance, provided there is a strong relationship between the mean and the explanatory variables.

10 Swiss Heads Again

One purpose of an analysis using the forward search is to establish the connection between individual units and inferences. We conclude with a dramatic example of the way in which overall statistics can be misleading about the greater part of the data, using the data on Swiss heads introduced in §3.

In §3 we showed that there were two outliers, units 104 and 111, which were the last two to enter the forward search. They did not seem to have any effect on inferences. However, Figure 23 is a forward plot of the likelihood ratio test for testing that all six values of λ are equal to one. This is based on a search on untransformed data, so that the order of entry of the units is the same as in our earlier analysis. In particular, units 104 and 111 are the last two to enter. The figure shows the enormous impact these two observations have on the evidence for transforming the data. At $m = 198$ the value of the statistic is 7.218, only slightly above the expected value for a χ^2_6 random variable. This

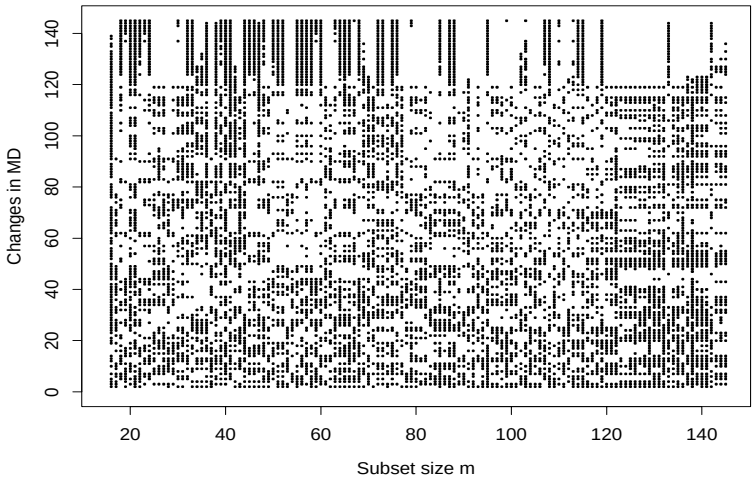

Figure 19: Diabetes data: forward plot of changes in Mahalanobis distances (black is an increase), suggesting membership of the three groups

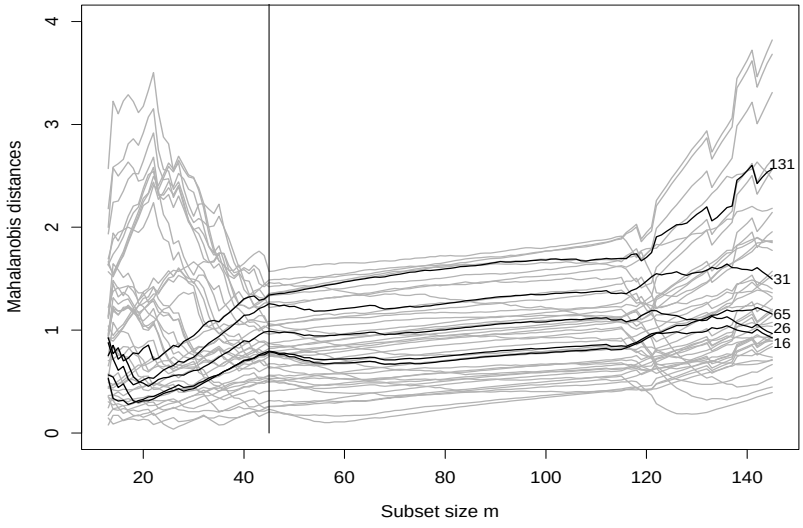

Figure 20: Diabetes data, starting from Group 2: forward plot of scaled Mahalanobis distances for units provisionally in Group 2; the highlighted units are not certainly classified

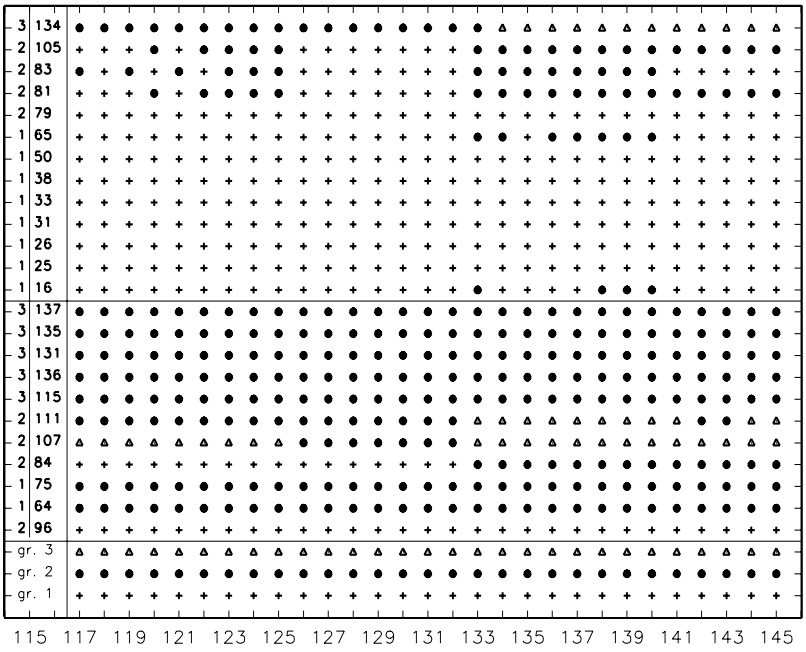

Figure 21: Diabetes data: final forward search. The classification of units in the top panel is uncertain

rises to 15.84 after the two outliers have entered, a value above the 95% point of the distribution, which is indicated on the plot. Without the information provided by the forward plot it would be easy to be misled into believing that the data need transformation.

Evidence for a transformation is provided by a skewed distribution. The only skew distribution in the scatterplots of the data such as those in Figure 3 is the marginal distribution of y_4, caused by the outlying values of units 104 and 111. To test whether all the evidence for the transformation is due to y_4 we repeat the calculation of the likelihood ratio, but now only testing whether $\lambda_4 = 1$. The other five values of λ are kept at one, both in the null parameter vector λ_0 and in the m.l.e. $\hat{\lambda}$. The search is therefore the same as before, but now gives rise to Figure 24, showing a test statistic to be compared with χ^2_1. It is now even clearer that all evidence for transformation of y_4 is provided by the inclusion of units 104 and 111. At the end of the search the test statistic has a value of 8.789, compared with 15.84 in Figure 23 for transforming all six variables. The difference, 7.05, is not significant for a χ^2_5 random variable, so that the evidence of the tests at the end of the search is that $\lambda_4 \neq 1$, whereas all other variables are equal to one.

This example shows the importance of the forward search in discovering influential observations. It also shows the power of combining statistical modelling with graphical presentation.

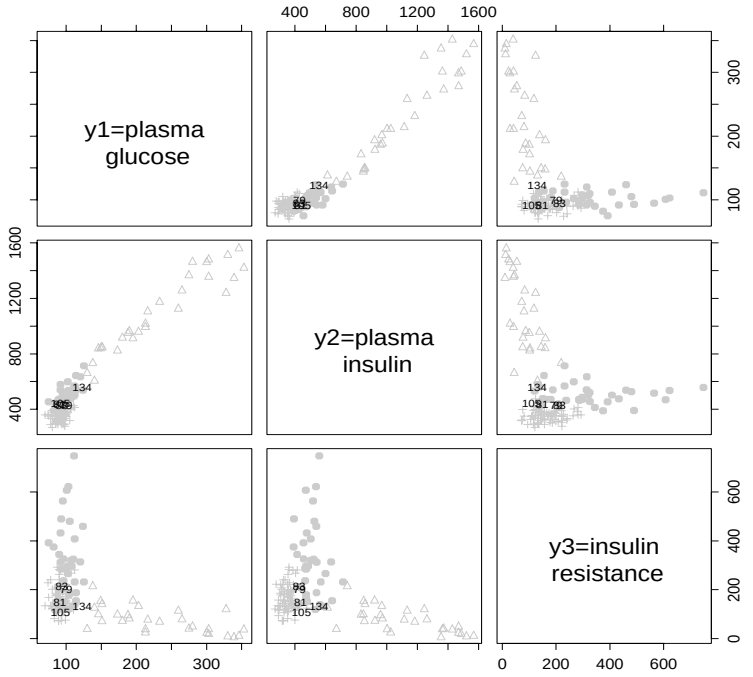

Figure 22: Diabetes data: scatterplot matrix, showing units not certainly classified

11 Discussion

11.1 Computing

Our examples are calculated using Gauss, which provides a combination of numerical programming and working graphics that we have found invaluable in developing the methods of the forward search. Most of our production plots are produced in S-Plus. The methods for univariate data (regression and generalized linear models) have been programmed in a Graphical User Interface (GUI) for S-Plus, which is available from the website for Atkinson and Riani (2000). A second book, Atkinson, Riani, and Cerioli (2004) on the analysis of multivariate data will give details of the methods sketched in this paper. Software will be made available on the same site `http://www.riani.it/ar`

11.2 Graphics

Forward plots of Mahalanobis distances, such as Figure 1, are taken directly from the output of our Gauss program. On the screen the variety of line types and colours, together with the ability to zoom, makes it possible to follow the trajectories of individual units in a way which is impossible on the printed page. Procedures for enhancing the plots include highlighting the trajectories of units of interest, as in Figure 2, and in using the same line style for units from conjectured groups as in Figure 8. Our goal is to be able to brush linked plots from the forward search, although this is still in the future.

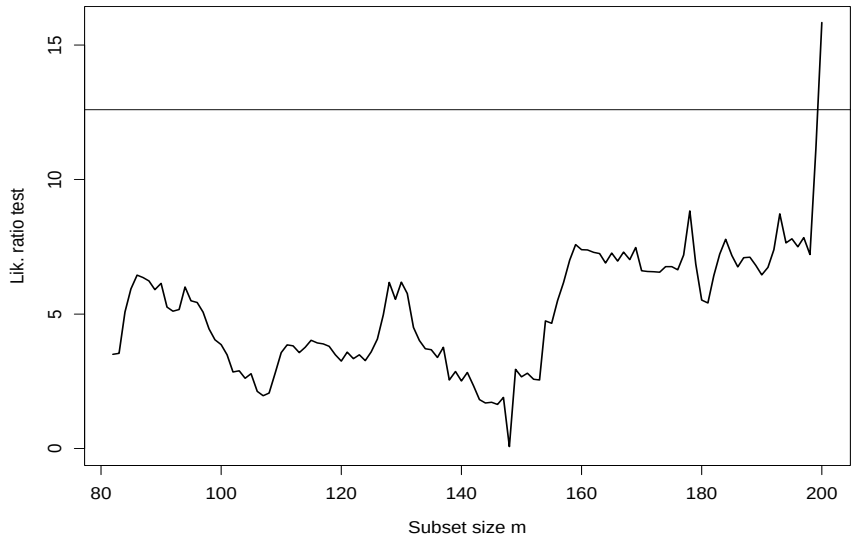

Figure 23: Swiss heads: forward plot of the likelihood ratio test for the hypothesis of no transformation. The horizontal line is the 95% point of χ^2_6. The last two units to enter provide all the evidence for a transformation

The forward search , as described in §2, moves from a subset of m observations to one of $m+1$. With larger sets of data the search can move forward in steps of size $s > 1$ from a subset of size m to one of size $m+s$, with proportional saving in computational time. Graphics can also be simplified to avoid uninformative overplotting, for example by using summary quantities as in Figure 9.

11.3 Data Analysis

Our examples show some of the ways in which the forward search is a powerful tool for exploring the structure of multivariate data. The largest data set in Atkinson, Riani, and Cerioli (2004) is of 28 responses from 341 municipalities in which we use the forward search to find appropriate transformations for each variable. As a result the structure of the data becomes much clearer. Often, for example, a transformation of the data leads to a greatly improved separation of the data into a main, multivariate normal, cluster and a few outliers. These outliers however mask the correct transformation and so have to be detected during the process of finding the transformation. Another example is the analysis of the diabetes data in §8, where our analysis led to a more coherent clustering and a clearer identification of borderline units than have other analyses, for example, that of Fraley and Raftery (1998).

Although, in the examples in this paper, we have often been able to relate our findings about the properties of particular units to the scatterplot matrix of the data, such a two-dimensional tool can fail to reveal outliers. But, much more importantly, the discovery of "interesting" observations by searching the Euclidean space of the observations can tell us nothing about the importance of each unit on inferences drawn from the data. If models are to be fitted and

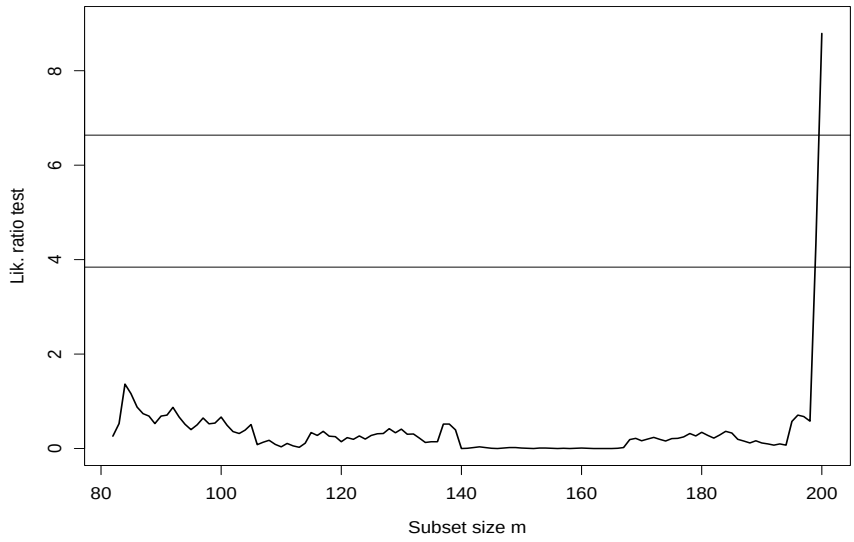

Figure 24: Swiss heads: forward plot of the likelihood ratio test for the hypothesis of no transformation of y_4. The horizontal lines are the 95% and 99% points of χ^2_1. The last two units to enter also provide all the evidence for this transformation

hypotheses tested, the forward search provides a unique way of visualising the effect of each unit

References

Atkinson, A. C. and M. Riani (2000). *Robust Diagnostic Regression Analysis*. New York: Springer–Verlag.

Atkinson, A. C. and M. Riani (2002). Tests in the fan plot for robust, diagnostic transformations in regression. *Chemometrics and Intelligent Laboratory Systems 60*, 87–100.

Atkinson, A. C., M. Riani, and A. Cerioli (2004). *Exploring Multivariate Data with the Forward Search*. New York: Springer–Verlag. (In preparation).

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

Daudin, J. J., C. Duby, and P. Trecourt (1988). Stability of principal component analysis studied by the bootstrap method. *Statistics 19*, 241–258.

Flury, B. and H. Riedwyl (1988). *Multivariate Statistics: A Practical Approach*. London: Chapman and Hall.

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

Reaven, G. M. and R. G. Miller (1979). An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia 16*,

17–24.

Riani, M. and A. C. Atkinson (2000). Robust diagnostic data analysis: Transformations in regression (with discussion). *Technometrics 42*, 384–398.

Riani, M. and A. C. Atkinson (2001). A unified approach to outliers, influence and transformations in discriminant analysis. *Journal of Computational and Graphical Statistics 10*, 513–544.

Zani, S., M. Riani, and A. Corbellini (1998). Robust bivariate boxplots and multiple outlier detection. *Computational Statistics and Data Analysis 28*, 257–270.