

Robust Diagnostics for Multivariate Analysis and Classification Problems

Metodi diagnostici robusti per l'analisi multivariata e la classificazione

Anthony Atkinson
Department of Statistics
London School of Economics, UK
a.c.atkinson@lse.ac.uk

Riassunto: In questo lavoro si descrive l'impiego d'un nuovo metodo robusto, denominato "forward search", per la ricerca di outliers multipli nell'analisi multivariata. La tecnica proposta non risente dei problemi di mascheramento e di diffusione e consente di mettere in luce, attraverso numerosi grafici, tutte le caratteristiche salienti dei dati multivariati. Le potenzialità del metodo sono qui estese ai problemi di classificazione.

Keywords. Box-Cox transformation, classification, cluster analysis, forward search, masked outliers, robustness.

1. Introduction

This paper summarises joint research with Marco Riani and Andrea Cerioli on the forward search, a powerful general method for detecting multiple masked outliers and for determining their effect on models fitted to the data. Atkinson and Riani (2000) describe its use in linear and nonlinear regression, response transformation and in generalized linear models. These examples are here extended to include the difficult problem of cluster analysis. They are described more fully in the forthcoming book Atkinson, Riani and Cerioli (2003).

In most examples we fit one multivariate normal distribution to the data, which is sufficient, in combination with the forward search, to reveal the structure. The first example, with two clusters, is analysed in §2 and again in §4, using the methods of the forward search described in §3. Unlike standard robust methods, the forward search reveals the presence of the clusters. Section 5 is an analysis of data on Swiss banknotes. In §6 an exploratory analysis of data on diabetes uses one multivariate distribution. The confirmatory analysis in §7 fits one distribution to each cluster. The final section discusses the importance of transformations in model-based multivariate data analysis.

2. The 60:80 Data and Failure of the Minimum Covariance Determinant

Figure 1 shows a scatterplot of 140 data points, 80 from a rather diffuse group and 60 from a tight cluster. The pattern is obvious to the unaided eye. The purpose of cluster analysis is to detect such groups, usually when the number of groups is unknown. This can be a difficult task for high dimensional data. But it would not seem to be particularly challenging when there are just two dimensions. However, when we fit a single model to

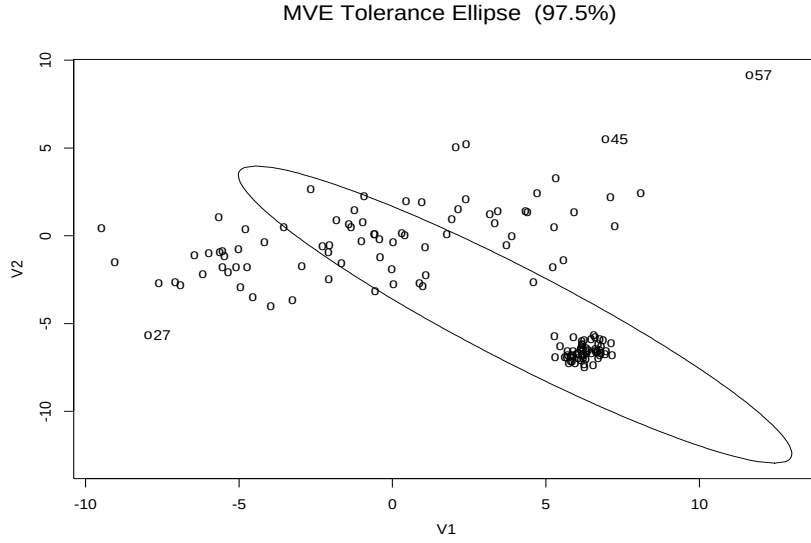


Figure 1: The 60:80 data: robust ellipse from S-Plus. The fit uses observations from both clusters

these data, neither standard classical methods nor very robust diagnostic methods unambiguously show that there are two different groups of observations.

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

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

where $\hat{\mu}$ and $\hat{\Sigma}$ are estimates of the mean and covariance matrix of the n observations.

Also in Figure 1 is an ellipse, estimated to contain 97.5% of the data. It was found by use of the “fast” algorithm of Rousseeuw and Van Driessen (1999), which finds the 50% of the data for which the covariance determinant is a minimum. This half includes elements from both groups, although the determinant is smaller than that which can be found containing the same number of observations solely from the larger group. Figure 2(a) shows a QQ plot of Mahalanobis distances from the fit to all the data. Neither it, nor Figure 2(b), based on the robust fit, reveal the two clusters. Perhaps there is some slight indication of a set of 60 small distances in the plot of robust distances, but it is not blindingly clear. We now look at diagnostics based on the forward search which do reveal the two groups.

3. The Forward Search

The parameters μ and Σ are estimated by maximum likelihood applied to 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

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

We 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

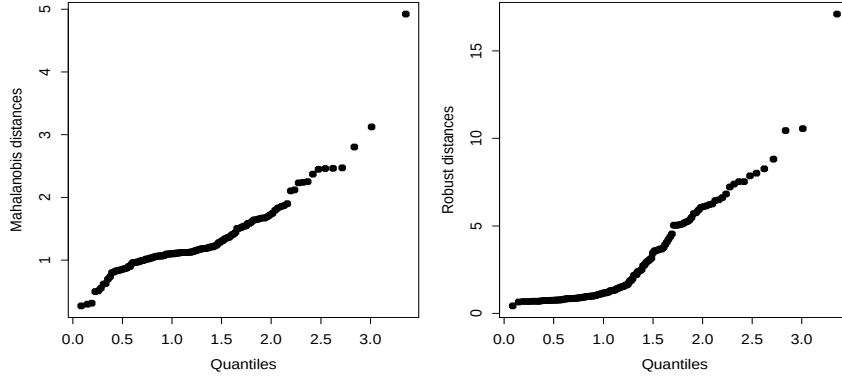


Figure 2: The 60:80 data: QQ plots of Mahalanobis distances: (a) non-robust; (b) robust from the minimum covariance determinant. Neither plot clearly shows the two groups

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 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

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

where p 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 right hand end of the forward plot. 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.

In the confirmatory part of our analysis we run 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. An observation that is on the edge of a tight cluster may have a relatively high Mahalanobis distance from that cluster, due to the low dispersion of the cluster, and have a comparable Mahalanobis distance from a dispersed cluster. 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, which are adjusted for the variance of the individual cluster. The customary squared Mahalanobis distance

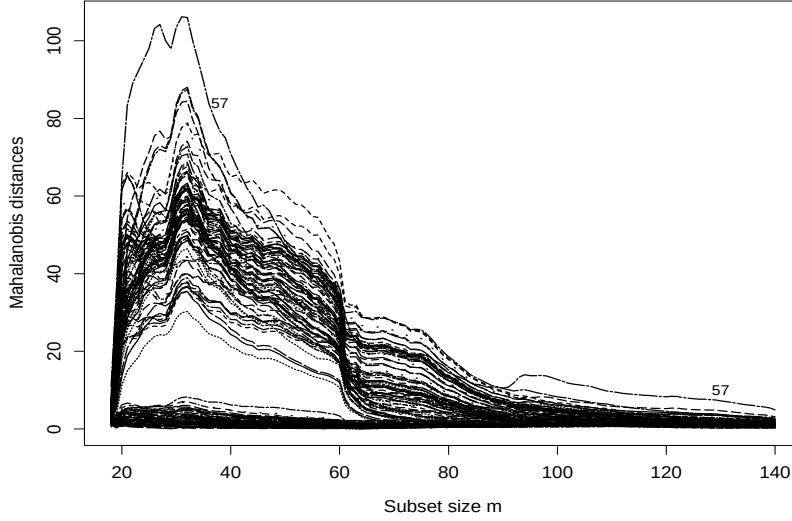


Figure 3: The 60:80 data: forward plot of Mahalanobis distances

for the i th observation from the g th group at step m is

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

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

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

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 therefore need to look at plots of both distances.

4. The Forward Search and the 60:80 Data

Figure 3 is a plot of the Mahalanobis distances from a forward search. The amount of information given by the figure is impressive. The starting point for this search includes units from both groups. At the very beginning of the search there is appreciable interchange as the units from the larger, more dispersed, group are rejected by the algorithm, which first selects those from the tight cluster. Up to $m = 60$ the plot shows a clear division of the distances into two groups. There are small distances for the tight group from which the parameters are estimated and a distinct set of much larger distances for the more dispersed group. When $m = 61$ the first unit from the dispersed group is included in the subset. Immediately there is a change in the plot, the distances for units from the larger group decreasing. As the forward plot shows, these distances remain large, while those for the small group increase slightly.

Figure 4 is a QQ plot of Mahalanobis distances during the forward search for two values of m . The left-hand panel, for $m = 37$, shows the two groups with a gap between the 60 small distances and the 80 large, just as there is in Figure 3 for this value of m .

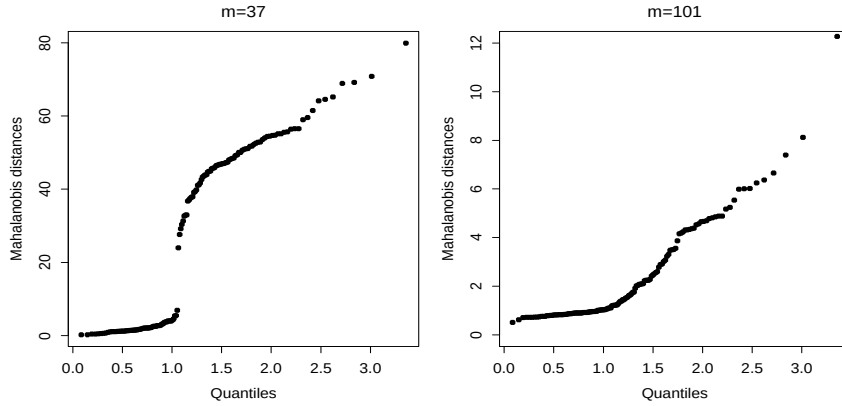


Figure 4: The 60:80 data: QQ plots of Mahalanobis distances: (left) $m = 37$; (right) $m = 101$

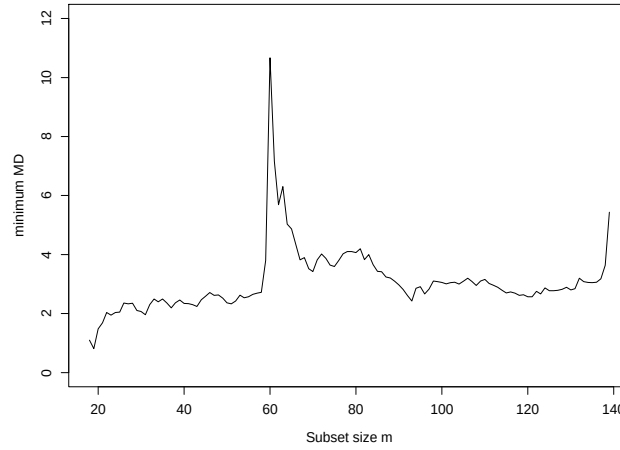


Figure 5: The 60:80 data: forward plot of the minimum Mahalanobis distance amongst units not in the subset: the cluster of 60 observations is manifest

The plot in the right-hand panel, for $m = 101$, shows, in the left-hand tail, the absence of very small distances: the centroid of the fitted observations is not near any data point: the smallest distance plots well above zero.

The forward plot shows that there are two groups of observations, one with small dispersion. The exact point of introduction of units from the second group is shown in Figure 5, the forward plot of the minimum Mahalanobis distances amongst units not in the subset. There is a needle sharp peak at $m = 60$ indicating that the next unit to be introduced is remote from the group of observations so far fitted. The search has fitted observations from just one of the groups and the structure of the data is clear.

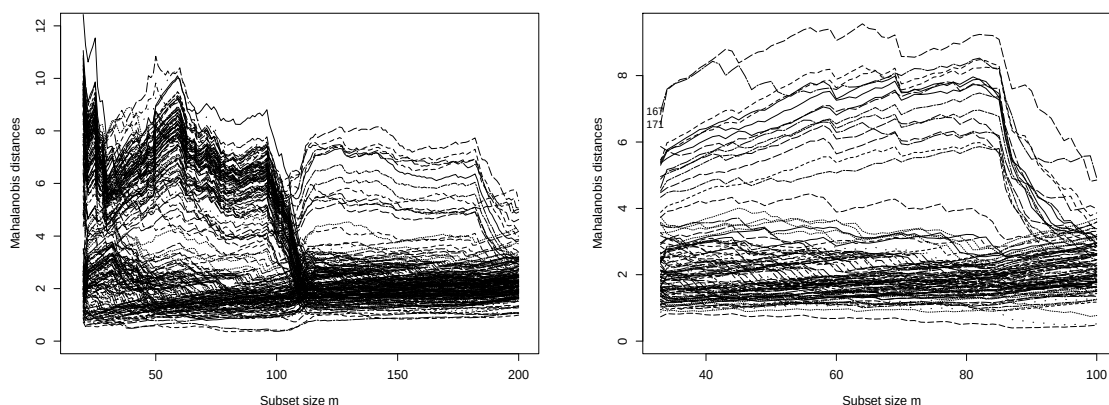


Figure 6: Swiss Banknote Data, forward plots of Mahalanobis distances: both groups and, on the right, the forgeries, showing evidence of a third group

5. Swiss Banknote Data

As a second example 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).

The left-hand panel of Figure 6 shows the scaled Mahalanobis distances from a forward search starting with 20 observations on notes believed genuine. 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 two groups are apparent, the forward search yielding the same 100 forgeries as Flury and Riedwyl (1988).

The structure of the group of forgeries is also readily revealed by the forward search. The right-hand panel of Figure 6 is a forward plot of the scaled Mahalanobis distances just for the forgeries. 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.

6. Diabetes Data: Exploratory Analysis

Our third example is of 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

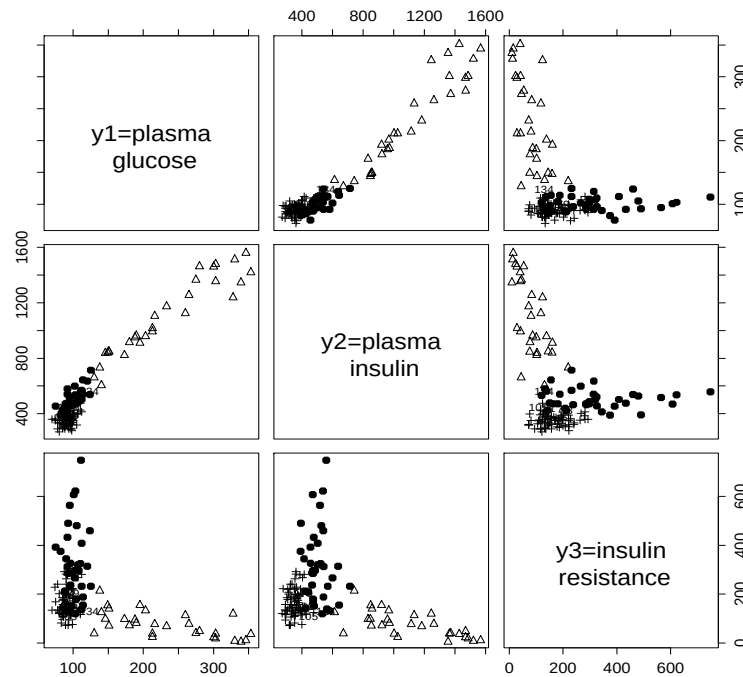


Figure 7: Diabetes data: scatterplot matrix, including final classification

Miller (1979). There are three measurements on each patient: x_1 and x_2 are respectively plasma glucose and plasma insulin response to oral glucose; x_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 7 is a scatterplot matrix of the data. Because of the shortage of space, our final clustering is shown. There seems to be a central 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.

The forward plot of Mahalanobis distances in Figure 8 has some strange features. There is a shortage of very small distances and overall the distances arguably fall into three groups: there is a gap between the largest distances and the rest around $m = 70$, these largest distances being rather uniformly distributed. Around $m = 45$ it looks as if there is a gap between the smallest distances and those which are somewhat larger, which again have a rather uniform distribution. These impressions are confirmed by a QQ plot of Mahalanobis distances.

The plot of the variance explained by the first two principal components, Figure 9(a), also suggests three large groups, with an early peak in the curve indicating some further structure. Two principal components seem to explain virtually 100% of the variability. These and similar analyses lead to a first division into three groups, which can then be refined using our confirmatory techniques.

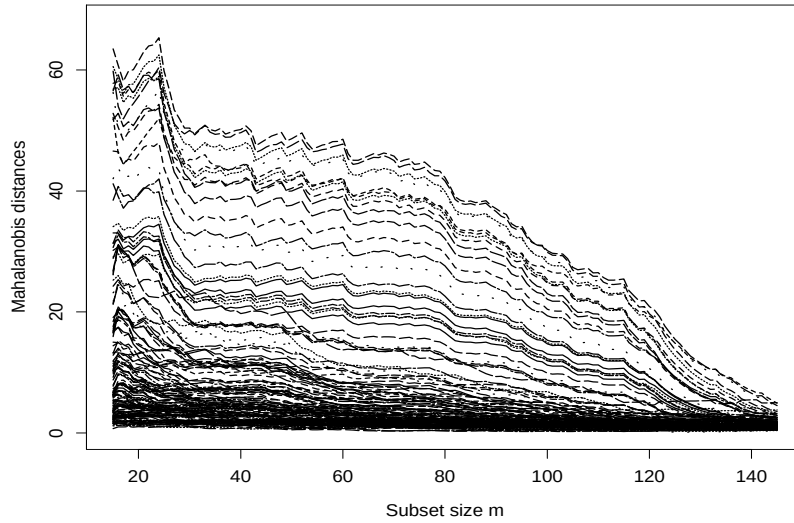


Figure 8: Diabetes data: forward plot of Mahalanobis distances, indicating three groups

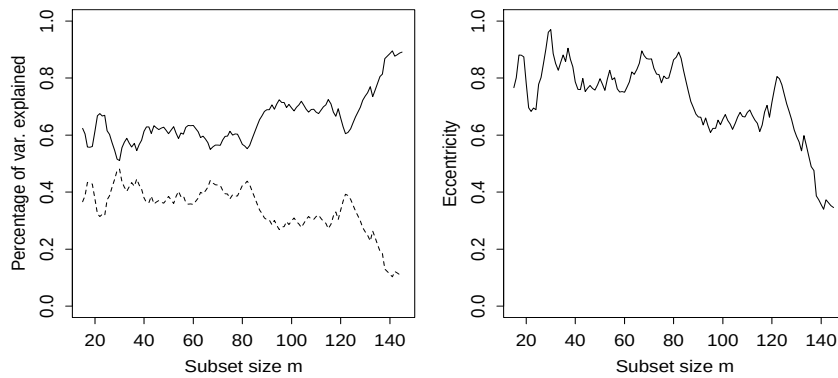


Figure 9: Diabetes data: forward plots of (a) percentage of variance explained by first two principal components and (b) eccentricity of fitted ellipse

7. Diabetes Data: Confirmatory Analysis

We start our confirmatory analysis by once more looking at forward searches based on our preliminary three groups. Figure 10 shows the scaled Mahalanobis distances for Group 1 for the search in which that group is included first. Overall these distances are roughly horizontal until all the group has been fitted. Thereafter they increase towards the end of the search. An exception to this is unit 79, which has the smallest distance at the end of the search.

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.

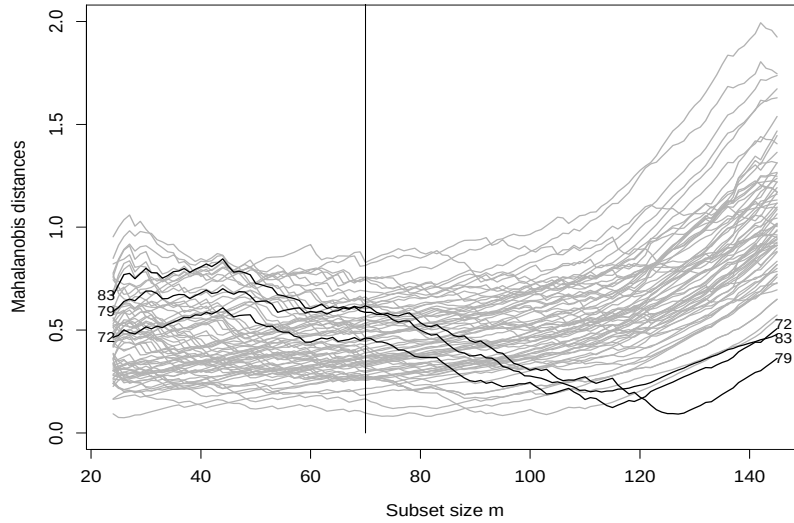


Figure 10: Diabetes data: forward plot of Mahalanobis distances for Group I

Table 1: Diabetes Data: final clustering of units

Group	I	II	III
Doctors	1:76	77:112	113:145
Certain	96	64, 75, 115 131, 136	
Uncertain	79, 83	81, 105	134

Our final forward search is represented in Figure 11, with a summary in Table 1. This table gives the doctors' allocation and then the way in which our clustering differs from it. The "certain" line includes all units which have the same allocation, with only occasional changes, throughout the search in Figure 11 where this differs from the doctors' allocation. In some cases the doctors' allocation seems to be wrong. The "uncertain" line contains all units for which the allocation changed appreciably during the search, classified by their final allocation. A general feature of Table 1 is that Group 2 appears to have "stolen" several units from the other groups and indeed we would expect that, using standardized distances, units would be attracted to a less dispersed group. This has happened to units 135 and 137, for which we keep the classification from our conservative confirmatory analysis on single groups.

Figure 12 is the scatterplot matrix of the final allocation from the forward search, with uncertain units represented by their number: these units fall at the intersections of clusters. Comparison with the scatterplot matrix of the doctors' allocation, shows that our clustering provides groups which are more coherent and compact.

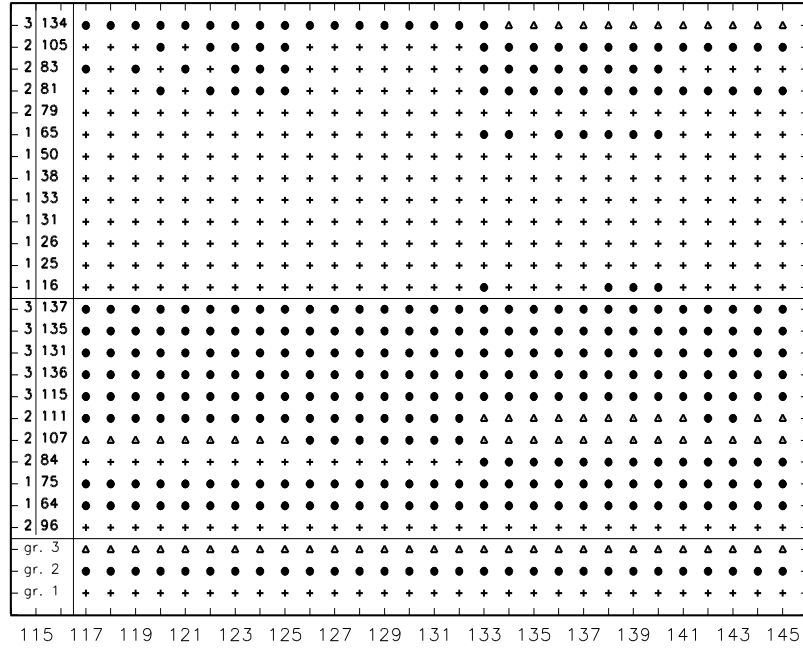


Figure 11: Diabetes data: final forward search

8. 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 ; $j = 1, \dots, p$. The normalized transformation of y_{ij} is

$$z_{ij}(\lambda_j) = \frac{y_{ij}^{\lambda_j} - 1}{\lambda_j G_j^{\lambda_j - 1}} \quad (\lambda \neq 0) \quad (6)$$

$$= G_j \log y_{ij} \quad (\lambda = 0) \quad (7)$$

where G_j is the geometric mean of the j -th variable. The value $\lambda_j = 1$ corresponds to no transformation of any of the responses. If the transformed observations are normally distributed with mean $\mu(\lambda)$ and covariance matrix $\Sigma(\lambda)$, the loglikelihood of the observations is given by:

$$l(\lambda) = -\frac{n}{2} \log 2\pi |\Sigma(\lambda)| - \frac{1}{2} \sum_{i=1}^n \{z_i - \mu(\lambda)\}^T \Sigma^{-1}(\lambda) \{z_i - \mu(\lambda)\}, \quad (8)$$

where $z_i = (z_{i1}, \dots, z_{ip})^T$ is the $p \times 1$ vector which denotes the transformed data for unit i . Substituting the maximum likelihood estimates $\hat{\mu}(\lambda)$ and $\hat{\Sigma}(\lambda)$ in equation (8), the maximized log likelihood can be written as:

$$l(\lambda) = \text{constant} - n/2 \log |\hat{\Sigma}(\lambda)|. \quad (9)$$

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

$$T_{LR} = n \log \{ |\hat{\Sigma}(\lambda_0)| / |\hat{\Sigma}(\hat{\lambda})| \} \quad (10)$$

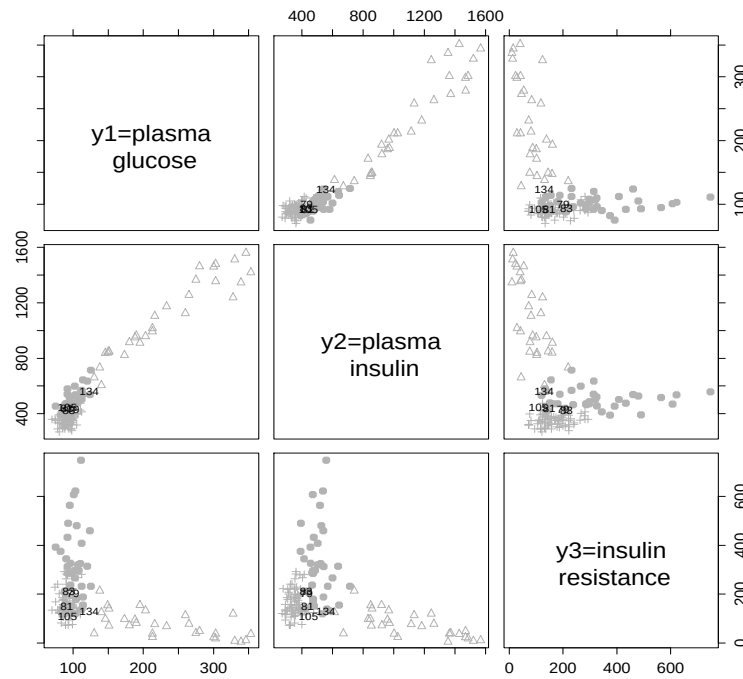


Figure 12: Diabetes data: scatterplot matrix, showing units not certainly classified

can be compared with the χ^2 distribution on p degrees of freedom. In equation (10) 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 (2001) monitor a sequence of forward plots of the square root of the likelihood ratio statistic (10) to obtain transformations which described most of the data, with the outliers entering at the end of the searches. Their example is discriminant analysis, so that the means and variances of the groups had to be estimated individually, unless a common covariance structure can be assumed. 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.

References

- Atkinson A.C. and Riani M. (2000) *Robust Diagnostic Regression Analysis*, Springer-Verlag, New York.
- Atkinson A.C. and Riani M. (2002) Tests in the fan plot for robust, diagnostic transformations in regression, *Chemometrics and Intelligent Laboratory Systems*, 60, 87–100.
- Atkinson A.C., Riani M. and Cerioli A. (2003) *Robust Diagnostic Multivariate Data Analysis*, Springer-Verlag, New York, (In preparation).
- Box G.E.P. and Cox D.R. (1964) An analysis of transformations (with discussion), *Journal of the Royal Statistical Society, Series B*, 26, 211–246.

- Flury B. and Riedwyl H. (1988) *Multivariate Statistics: A Practical Approach*, Chapman and Hall, London.
- Fraley C. and Raftery A.E. (1998) How many clusters? Which clustering method? – Answers via model-based cluster analysis, *Computer Journal*, 41, 578–588.
- Reaven G.M. and Miller R.G. (1979) An attempt to define the nature of chemical diabetes using a multidimensional analysis, *Diabetologia*, 16, 17–24.
- Riani M. and Atkinson A.C. (2000) Robust diagnostic data analysis: Transformations in regression (with discussion), *Technometrics*, 42, 384–398.
- Riani M. and Atkinson A.C. (2001) A unified approach to outliers, influence and transformations in discriminant analysis, *Journal of Computational and Graphical Statistics*, 10, 513–544.
- Rousseeuw P.J. and Van Driessen K. (1999) A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, 41, 212–223.
- Zani S., Riani M. and Corbellini A. (1998) Robust bivariate boxplots and multiple outlier detection, *Computational Statistics and Data Analysis*, 28, 257–270.