BIVARIATE BOXPLOTS, MULTIPLE OUTLIERS, MULTIVARIATE TRANSFORMATIONS AND DISCRIMINANT ANALYSIS: THE 1997 HUNTER LECTURE*

ANTHONY C. ATKINSON[1]† AND MARCO RIANI[2]

[1] *The London School of Economics, London WC2A 2AE, UK*
[2] *Istituto di Statistica, Università di Parma, Italy*

SUMMARY

Outliers can have a large influence on the model fitted to data. The models we consider are the transformation of data to approximate normality and also discriminant analysis, perhaps on transformed observations. If there are only one or a few outliers, they may often be detected by the deletion methods associated with regression diagnostics. These can be thought of as 'backwards' methods, as they start from a model fitted to all the data. However such methods become cumbersome, and may fail, in the presence of multiple outliers. We instead consider a 'forward' procedure in which very robust methods, such as least median of squares, are used to select a small, outlier free, subset of the data. This subset is increased in size using a search which avoids the inclusion of outliers. During the forward search we monitor quantities of interest, such as score statistics for transformation or, in discriminant analysis, misclassification probabilities. Examples demonstrate how the method very clearly reveals structure in the data and finds influential observations, which appear towards the end of the search. In our examples these influential observations can readily be related to patterns in the original data, perhaps after transformation. © 1997 John Wiley & Sons, Ltd.

Environmetrics, **8**, 583–602 (1997)
No. of Figures: 12. No. of Tables: 1. No of References: 23.

KEY WORDS Box–Cox transformation; deletion diagnostic; forward search; least median of squares; multivariate normality; probability of misclassification; score statistic; transformation to normality; very robust methods

1. INTRODUCTION

Multiple outliers can strongly affect the model fitted to data, as may unidentified distinct subsets. But such important observations may be hard to identify, even with deletion techniques such as those of regression diagnostics. The major difficulty, often called masking, arises because the deletion of several observations may be necessary before there is an appreciable change in the fitted model or in the pattern of residuals. These diagnostic techniques may be thought of as

* Presented at the Environmetrics Meeting, Innsbruck, Austria, 4–8 August 1997.
† Correspondence to: Professor A. C. Atkinson, London School of Economics, London WC2A 2AE, UK

CCC 1180–4009/97/060583–20$17·50
Received 9 June 1997
Revised 12 July 1997

'backwards' methods: they start from a fit to all the data and then study the effects of deletion. Instead we consider a 'forward' search through the data. As we show, this forward search clearly displays any outlying or influential observations in a way which can easily be linked to simple plots of the data.

Bivariate boxplots and methods related to robust techniques are used to identify a small outlier free subset of the observations which agrees with the model being fitted. The subset then grows by the sequential selection of observations closest to the model, for example those with the smallest residuals from the fitted subset. During the growth of the subset we monitor quantities of interest: for transformations we look at score statistics for the Box–Cox family and for discriminant analysis at misclassification probabilities. The graphs of such quantities lead to the identification of interesting observations, which nearly always occur in the last steps of the search. For transformations, which observations are influential will depend on whether we search on transformed or untransformed data.

Two bivariate boxplots are described in Section 2. In Section 3 the bivariate boxplots are superimposed on scatterplot matrices, providing information about potential outliers. For one of the boxplots, the shape of the robust contours of the bivariate distribution indicates whether the data should be transformed. The next section uses the Box–Cox family for univariate and multivariate transformations to normality. For univariate data the initial subset is found using the least median of squares criterion and the forward search is on ordered residuals. For multivariate transformations we use the contours of the bivariate boxplots to define the initial subset and search on ordered Mahalanobis distances. In the last section we apply our multivariate method to discriminant analysis.

The emphasis is on the analysis of data using many plots. The examples show how features of plots from the forward search can be informatively related to structure in the scatterplot matrices of the data. The paper demonstrates how the forward search enables us to get inside the data in a way which conventional deletion methods do not.

2. BIVARIATE BOXPLOTS

The univariate boxplot (Tukey 1977, p. 40) is a well-established technique for summarizing the distribution of a single random variable. A major advantage is that it is available in many statistical packages. In this section we describe two bivariate extensions of the boxplot. These not only provide informative summaries of the data but can be used to provide starting points for forward searches through the data.

If the data have a bivariate normal distribution the contours of the joint density will be elliptical. If the contours of the empirical distribution are far from elliptical this will indicate systematic departures from normality. Normality, and so elliptical contours, may often be achieved by the deletion of outliers and by transformation of the data. We give an example in Section 4 and show the effect on bivariate boxplots.

We first need a rough ordering of the observations from those most outlying to those closest to a bivariate normal distribution. Ruts and Rousseeuw (1996) describe a method which considers observations individually. But, for the construction of the boxplot, it is sufficient to consider the observations in groups.

The computationally more intensive of the two versions of the bivariate boxplot (Zani *et al.* 1997) uses peeling of convex hulls to establish the shape of the central part of the data. Successive convex hulls are peeled until the first one is obtained which includes less than

50 per cent of the data (and so asymptotically half the data as the sample size increases). The convex hull so found (which we call the 50 per cent hull) is smoothed using a *B*-spline, constructed from cubic polynomial pieces, which uses the vertices of the 50 per cent hull to provide information about the location of the knots. (Eilers and Marx 1996 give computational details for construction of the *B*-spline curve.)

Zani *et al.* (1997) discuss several choices of a robust bivariate median. In this paper we find the robust centre as the arithmetic mean of those observations lying within the 50 per cent contour. In this way we can exploit both the efficiency properties of the arithmetic mean and the natural trimming offered by the hulls. Other contours, to discriminate between central observations and outliers, are found by linear scaling of the distance of the smoothed 50 per cent contour from the centre. The calculations depend solely on the percentage points of the χ^2_2 distribution: for a 90 per cent contour the outer contour should be 1·82 times as far from the centre as the smoothed contour. Simulation results for small sample sizes indicate that such regions are slightly too small, as the *B*-spline lies within the convex hull which may anyway contain slightly less than half the data. The exact value of the scaling coefficient is not important if the contours are to be used solely to provide a starting point for the forward search.

As an example we take the 57 readings on five properties of soil samples given in Table I (Mulira 1992). The first two variables are measurements of pH, which are highly correlated. The other three are measures of available phosphorus, potassium and magnesium. The data are appreciably rounded. To avoid numerical problems with the S-Plus peeling algorithm, the data were jittered by adding small normal errors. Figure 1 is a plot of just two variables, concentration of potassium, K and concentration of phosphorus P. There is one very clear outlier, observation 20, and a tendency for the data to be concentrated in the lower left corner of the plot: observations 33,

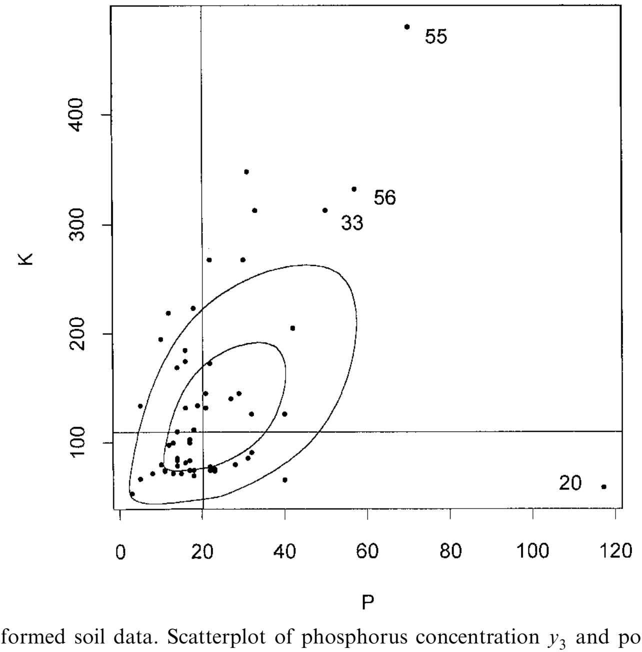

Figure 1. Untransformed soil data. Scatterplot of phosphorus concentration y_3 and potassium concentration y_4 with robust contours

Table I. Soil data: pH and available nutrients in 57 soil samples from fields in England and Wales

Observation	y_1 pH$_1$	y_2 pH$_2$	y_3 P	y_4 K	y_5 Mg
1	6·3	5·8	31	88	130
2	6·6	6·0	40	68	76
3	6·6	6·0	32	93	79
4	6·3	5·8	40	128	106
5	6·6	5·9	22	77	61
6	6·9	6·3	11	77	45
7	5·8	5·1	22	175	91
8	5·5	5·0	12	221	106
9	6·2	5·7	17	77	103
10	6·2	5·6	18	114	225
11	6·6	6·1	14	86	275
12	6·5	6·1	30	270	245
13	7·0	6·5	18	72	180
14	5·8	5·1	5	136	118
15	6·5	5·7	17	86	193
16	6·3	5·7	16	134	158
17	8·0	7·4	21	134	109
18	7·0	6·3	18	77	61
19	8·3	7·7	13	102	70
20	8·0	7·5	117	61	70
21	5·8	5·1	13	102	165
22	6·8	6·0	5	69	214
23	7·2	6·4	28	82	176
24	6·8	6·0	3	56	138
25	6·2	5·6	10	82	275
26	6·6	6·0	10	197	325
27	6·6	6·0	12	100	308
28	5·6	4·9	14	88	224
29	6·5	5·8	23	76	138
30	6·0	5·5	16	187	96
31	5·9	5·3	22	80	68
32	5·8	5·3	23	78	79
33	7·6	7·0	50	315	370
34	7·1	6·5	16	177	686
35	6·5	6·0	42	207	358
36	7·0	6·4	29	147	348
37	6·2	5·6	8	74	150
38	6·2	5·5	33	315	148
39	6·3	5·6	17	102	125
40	5·5	4·8	17	105	180
41	5·6	5·0	14	171	144
42	5·9	5·3	22	270	239
43	5·8	5·2	15	74	330
44	6·3	5·9	31	350	574
45	6·8	6·2	19	136	353
46	7·2	6·7	21	147	506
47	6·9	6·3	18	225	551
48	6·2	5·9	27	142	89
49	5·5	5·0	14	112	110
50	5·5	5·0	14	112	110
51	6·3	5·7	16	84	77
52	5·8	5·1	14	81	91
53	6·9	6·2	11	76	73
54	6·6	6·1	32	128	46
55	7·5	6·8	70	481	88
56	7·1	6·4	57	334	68
57	6·2	5·6	13	74	62

55 and 56 also lie away from the main body of data. The resulting boxplot contours are not elliptical and call attention to the skewed distribution of values, which can perhaps be reduced by transformation.

A computationally less intensive form of boxplot can be found by fitting one or more ellipses to the data, using robust estimates of the parameters. Goldberg and Iglewicz (1992) describe two methods. The more complicated requires fitting quadrants of four different ellipses. We exemplify the simpler form, called the 'Relplot', in which the marginal medians, as opposed to means, of the observations are used as locational estimates. The required covariance matrix of the observations is then estimated by sums of squares and products about these medians. The central 50 per cent of the observations is defined by the ellipse passing through the median Mahalanobis distance and the F distribution on 2 and $n - 2$ degrees of freedom is used to scale up the outer contours, which are now elliptical.

Figure 2 reproduces the data of Figure 1 but now with elliptical contours. In the plot the variables have been scaled by division by their marginal standard deviations about the medians. The contours are now not informative about the form of any systematic departure of the plot from normality. However the tentative outliers are still clearly displayed. More importantly for our application, even though the estimate of the covariance matrix is not robust, the data within the 50 per cent contour clearly contain no outliers.

3. SCATTERPLOT MATRICES

The scatterplot matrix is a very useful tool for obtaining a preliminary impression of the structure of data. It is probably most helpful when, as here, regression structure is absent. Cook and Weisberg (1994, p. 85) show how scatterplot matrices may be difficult to interpret if there are

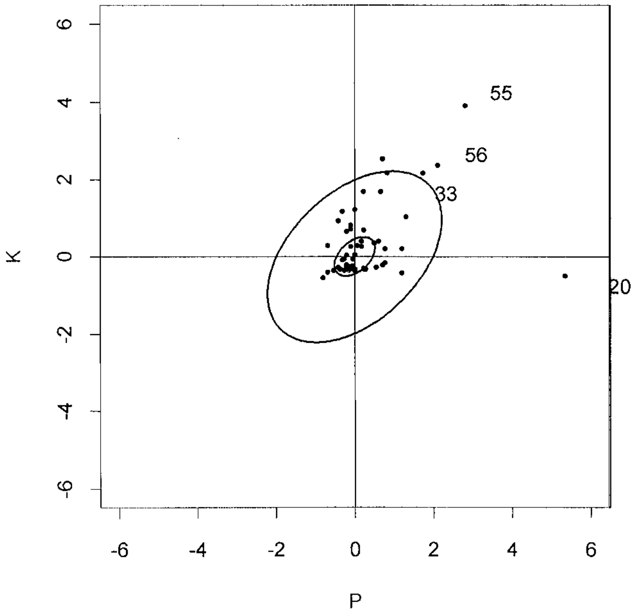

Figure 2. Untransformed soil data. Scatterplot of phosphorus concentration y_3 and potassium concentration y_4 with elliptical contour

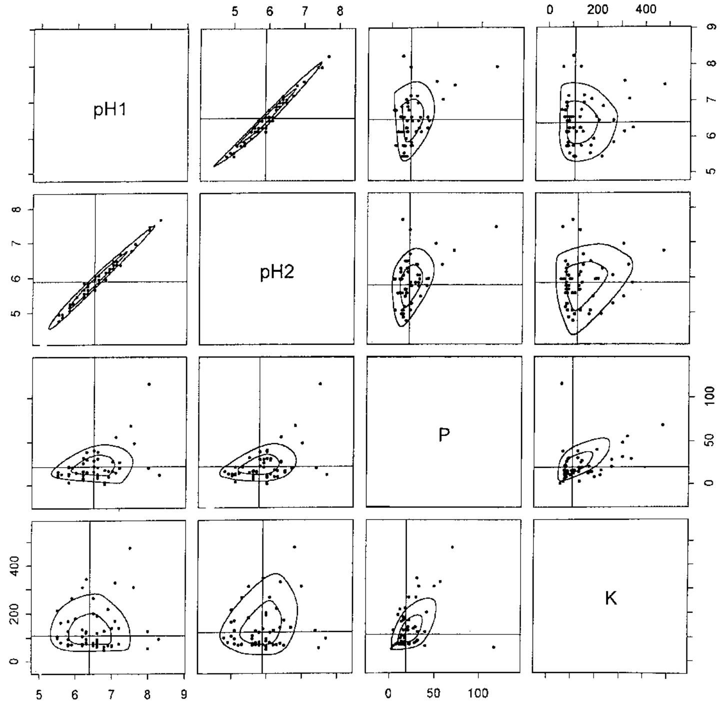

Figure 3. Untransformed soil data. Scatterplot matrix of first four variables: the non-elliptical robust contours suggest the data should be transformed

non-linear relationships between regression variables. In this section we illustrate the properties of scatterplot matrices formed from the bivariate boxplots of Section 2.

Figure 3 shows the matrix of bivariate boxplots for the data of Table I with the outer contour at 90 per cent. For legibility on the printed page only the first four variables are plotted. On the computer screen the use of brushing makes it possible to interpret plots with more variables. What is most noticeable in the figure is the shape of the various contours. If all were elliptical the data could be treated as having a multivariate normal distribution, and this does seem to be the case for the two pH measurements. But the variety of shapes for the other plots suggests that we should try transforming K and P and that different transformations may be needed for the two variables.

The plots with elliptical contours are shown in Figure 4. In this scatterplot matrix the univariate boxplots for each variable are on the diagonal of the matrix. These indicate one outlier for pH_2 and skewed distributions for K and P. The bivariate plots show patterns of outliers in the upper right hand corners which may be reconciled with the data through transformation, although observation 20 seems outlying in several plots.

4. TRANSFORMATIONS

4.1 Univariate transformation

We consider first the transformation of just one of the variables in the soil data, using the concentration of phosphorus, y_3, which nicely illustrates several points. We use the Box and Cox

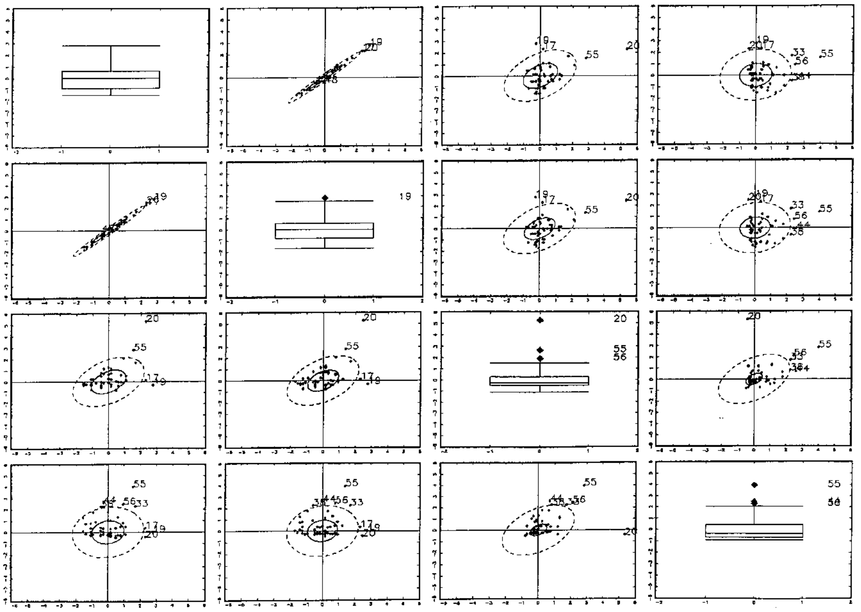

Figure 4. Untransformed soil data. Scatterplot matrix of first four variables, showing potential outliers

(1964) parametric family of power transformations, written in normalized form as

$$z(\lambda) = \begin{cases} (y^{\lambda} - 1)/(\lambda \dot{y}^{\lambda-1}) & \lambda \neq 0 \\ \dot{y} \log y & \lambda = 0 \end{cases}, \tag{1}$$

where $\dot{y} = \exp(\Sigma \log y_i/n)$ is the geometric mean of the observations. For a regression model with residual sum of squares of the $z(\lambda)$ equal to $R(\lambda)$, the profile log-likelihood of the observations, maximized over β and σ^2, is

$$L_{\max}(\lambda) = \text{const} - (n/2)\log\{R(\lambda)/n\} \tag{2}$$

so that $\hat{\lambda}$ minimises $R(\lambda)$.

For inference about the transformation parameter λ, Box and Cox use likelihood ratio tests derived from (2), that is the statistic

$$T_{\mathrm{LR}} = 2\{L_{\max}(\hat{\lambda}) - L_{\max}(\lambda_0)\} = n[\log\{R(\lambda_0)/R(\hat{\lambda})\}]. \tag{3}$$

A disadvantage of this likelihood ratio test is that numerical maximization is required to find the value of $\hat{\lambda}$. A computationally simpler alternative is the approximate score statistic $T_p(\lambda)$ (Atkinson 1985, Chapter 6) which is the t test for regression on the constructed variable $\partial z(\lambda)/\partial\lambda$, derived from Taylor series expansion of (1). We exemplify the use of both tests.

4.2 The forward search

These tests for transformation are aggregate statistics, based on all the data. To find the effect of single observations on the statistics, deletion methods can be used, which yield the effect of each observation in turn, given that the other $n-1$ observations are still used for fitting. However, if there are several interesting observations their effects may not be detected by only deleting one observation at a time, a condition known as masking. To avoid this problem we instead look at the effect of adding observations. We start with a small subset and allow it to grow in size by selecting observations closest to the assumed model. For each subset size we calculate the statistic of interest: for transformations this is one of the tests given above. For the discriminant analysis of the next section, the evolution of misclassification probabilities is of interest.

To get inside the data in this way we order the observations from those nearest to the normal theory model to those furthest from it. Our methods are described in detail in Riani and Atkinson (1998).

The ordering is in two stages. First we need to find a subset of the data which is free of outliers. We then conduct a forward search (Hadi 1992; Atkinson 1994) based on residuals for univariate data or on Mahalanobis distances for multivariate data. The comparison of subsets uses measures from very robust analyses. For regression this is least median of squares (LMS).

For the linear regression model $E(Y) = X\beta$, with X of rank p, let b be any estimate of β. With n observations the residuals from this estimate are $e_i(b) = y_i - x_i^{\mathrm{T}} b$ $(i = 1, \ldots, n)$. The LMS estimate $\tilde{\beta}$ minimizes the median value of $e_i^2(b)$. Rousseeuw (1984) finds an approximation to $\tilde{\beta}$ by searching only over elemental sets, that is subsets of p observations, taken at random. Depending on the dimension of the problem we find the starting point for the forward search either by sampling 1000 subsets or by exhaustively evaluating all subsets.

We require a subset which is outlier free. The best subset is that which gives the smallest value of the LMS criterion. For a particular subset $\mathcal{M}$ of size m let the least squares estimate of β be $\hat{\beta}(\mathcal{M})$ and let the median, allowing for estimation, be

$$\mathrm{med} = [(n + p + 1)/2], \tag{4}$$

the integer part of $(n + p + 1)/2$. The LMS criterion for $\hat{\beta}(\mathcal{M})$ requires ordering the residuals to obtain the variance estimate

$$\tilde{\sigma}^2(\mathcal{M}) = e^2_{[\mathrm{med}]}\{\hat{\beta}(\mathcal{M})\}, \tag{5}$$

where $e^2_{[k]}$ is the kth ordered squared residual. We take as our initial subset that for which $\tilde{\sigma}^2(\mathcal{M})$ is a minimum, so obtaining an outlier free start for our forward search.

The forward search for regression moves from fitting m observations to $m + 1$ by choosing the $m + 1$ observations with the smallest least squares residuals, with β estimated from the subset of size m. The observations are chosen by ordering all n residuals. Because n distances are calculated and ordered for each move from m to $m + 1$, observations can leave the subset used for fitting as well as joining it as m increases. Forward searches allowing for the variances of the residuals are used by Hadi and Simonoff (1993) and by Atkinson (1994). Our comparisons show that although the choice of residual has a slight effect on the forward search, it has no substantial effect on the plots and inferences derived from the search. In most moves from m to $m + 1$ observations, one new observation joins the subset. However there are times when one leaves as two join. This usually happens when we include one observation which belongs to a cluster of outliers. As our examples show, it is the last third or so of the search that contains the information about transformations. The ordering of this part of the data does not seem to be sensitive to the particular search strategy employed.

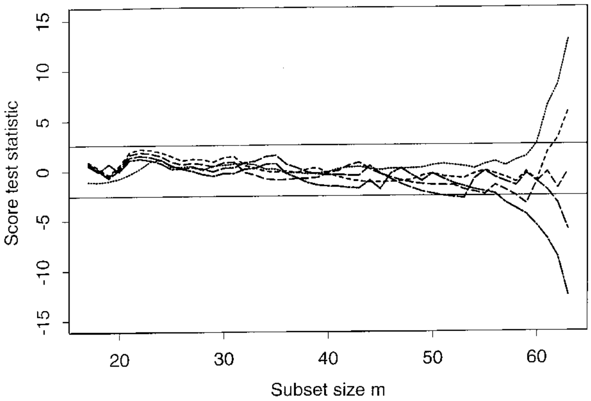

Figure 5. Soil data. Score statistic for power transformation $T_p(\lambda_0)$ for y_3 as the subset size m increases. The parameter values are: $\lambda = 1$ (—-—-), $\lambda = 0{\cdot}5$ (———), $\lambda = 0$ (–––), $\lambda = -0{\cdot}5$ (-----), $\lambda = -1$ (······). The log transformation is acceptable

The resulting analysis for the transformation of y_3 is summarized in Figure 5. The test statistic $T_p(\lambda)$ was calculated for five values of λ (-1, $-0{\cdot}5$, 0, 0·5 and 1). For each of the five values the data were appropriately transformed and a search was made for the basic subset giving the smallest median squared residual. (Since there is no regression in this example, $p = 1$ and there are only 57 basic subsets corresponding to trying each observation in turn as the estimate of location.) Forward searches were then performed on the transformed data, giving the five curves of Figure 5. The plot shows that the log transformation $\lambda = 0$ is acceptable, but that for other values of λ the last few observations to be included in the search cause rejection of the other transformations: the horizontal lines in the plot correspond to the 99 per cent points of the standard normal distribution. These boundaries give a rather too strong impression of the significance of the test as the variance of the approximate score test tends to be larger than one. Some simulation results on the comparison of several tests for transformations are given by Atkinson and Lawrance (1989).

As in all forward searches we have analysed, there is a strong link between the forward search and the data. For $\lambda = 1$, the untransformed data, the last four observations to enter, working backwards, are 20, 55, 56 and 33. These, the four largest observations, are labelled in the bivariate boxplots of Figures 1 and 2. Three of them are also shown in the univariate boxplot on the diagonal of Figure 4. For $\lambda = 0$ the boxplot, not reproduced here, exhibits a symmetrical and normal-seeming distribution. For further transformation, such as the reciprocal, the smallest observations are overtransformed and become the outliers.

4.3 Multivariate transformation

The forward search procedure for multivariate transformation described by Riani and Atkinson (1998) is similar. Let y_{ij} be the ith observation on the jth response out of v. In the extension of the Box and Cox (1964) family to multivariate responses the normalized transformation of y_{ij} is

$$\begin{aligned} z_{ij}(\lambda_j) &= (y_{ij}^{\lambda_j} - 1)/(\lambda_j \dot{y}_j^{\lambda_j - 1}) && \lambda \neq 0 \\ &= \dot{y}_j \log y_{ij} && \lambda = 0, \end{aligned} \tag{6}$$

where $\dot{y}_j$ is the geometric mean of the jth response. The value $\lambda_j = 1$ ($j = 1, \ldots, v$) corresponds to no transformation of any of the responses. If the transformed observations are normally distributed with mean μ_i for the ith observation and covariance matrix Σ, twice the profile log-likelihood of the observations is given by

$$\begin{aligned} 2L_{\max}(\lambda) &= \text{const} - n\log|\hat{\Sigma}(\lambda)| - \sum_{i=1}^{n}\{z_i(\lambda) - \hat{\mu}_i(\lambda)\}^{\mathrm{T}}\hat{\Sigma}^{-1}(\lambda)\{z_i(\lambda) - \hat{\mu}_i(\lambda)\} \\ &= \text{const} - n\log|\hat{\Sigma}(\lambda)| - \sum_{i=1}^{n} e_i(\lambda)^{\mathrm{T}}\hat{\Sigma}^{-1}(\lambda)e_i(\lambda). \end{aligned} \tag{7}$$

In (7) $\hat{\mu}_i(\lambda)$ and $\hat{\Sigma}(\lambda)$ are derived from least squares estimates for fixed λ and $e_i(\lambda)$ is the $v \times 1$ vector of residuals.

The calculations of $\hat{\mu}_i(\lambda)$ and $\hat{\Sigma}(\lambda)$ is simplified when, as in this paper, the $n \times p$ matrix of explanatory variables X is the same for all responses. As a result, the least squares estimates are found by independent regressions for each response, yielding the $p \times v$ matrix of parameter estimates $\hat{\beta}(\lambda) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}z(\lambda)$. Then, in the usual way,

$$\begin{aligned} (n-p)\hat{\Sigma}(\lambda) &= \sum_{i=1}^{n} e_i(\lambda)e_i(\lambda)^{\mathrm{T}} \\ &= \{z(\lambda) - X\hat{\beta}(\lambda)\}^{\mathrm{T}}\{z(\lambda) - X\hat{\beta}(\lambda)\}. \end{aligned} \tag{8}$$

When these estimates are substituted in (7), the profile log-likelihood reduces to

$$2L_{\max}(\lambda) = \text{const}' - n\log|\hat{\Sigma}(\lambda)|. \tag{9}$$

So, to test the hypothesis $\lambda = \lambda_0$, the statistic

$$\mathrm{T}_{\mathrm{LR}} = n\log\{|\hat{\Sigma}(\lambda_0)|/|\hat{\Sigma}(\hat{\lambda})|\} \tag{10}$$

is compared with the χ^2 distribution on v degrees of freedom, the generalization of the univariate statistic (3).

In (10) $\hat{\lambda}$ is the vector of v parameter estimates maximizing (7), which is found by numerical search. As in Section 2, constructed variables can be used to define diagnostic tests which avoid numerical maximization of the likelihood. However, owing to the presence of the constructed variables, the explanatory variables are no longer the same for all responses and the covariance Σ between the v responses has to be allowed for in estimation: independent least squares is replaced by generalized least squares. The special structure, that of seemingly unrelated regression (Zellner 1962), is used by Atkinson (1995) to obtain deletion diagnostics for multivariate transformations. We do not explore these methods here, simplicity being lost when straightforward regression can no longer be used.

To use the forward search to order the observations we make the appropriate multivariate transformation for the hypothesis to be tested and then use the Mahalanobis distance in place of the squared residuals used for the univariate transformation. Suppose that a subset $\mathcal{M}$ of m observations is used to estimate the regression (if any) and covariances. Let the estimates be $\hat{\mu}(\mathcal{M})$ and $\hat{\Sigma}(\mathcal{M})$, yielding the set of squared Mahalanobis distances

$$d_i^2(\mathcal{M}) = \{y_i - \hat{\mu}_i(\mathcal{M})\}^{\mathrm{T}}\hat{\Sigma}^{-1}(\mathcal{M})\{y_i - \hat{\mu}_i(\mathcal{M})\}, \tag{11}$$

for $i = 1, \ldots, n$. Ordering these distances and then taking the observations with the $m + 1$ smallest distances takes the forward search from m to $m + 1$. To start the forward search the initial subset of outlier free observations is found as the intersection of the observations lying within a suitably chosen elliptical contour on all panels of the scatterplot matrix similar to Figure 4, but including all five variables. The factor for scaling the contour is chosen to give an initial subset of suitable size. If the subset is too small, there tend initially to be seemingly random fluctuations in the quantities being monitored until the forward search has established a subset in agreement with the model. If the subset is too large, outliers may be masked and, having been included, will never be excluded.

The exploratory analyses of the soil data described by Riani and Atkinson (1998) indicate values of the individual parameters for confirmatory checking. There is no indication of the need to transform y_1 and y_2. Since these are measurements of pH, and so are already logged hydrogen ion concentrations, any further transformation would seem unlikely. However Richardson and Green (1997), following earlier log-normal analyses, use the logarithmic transformation in the analysis of data on the pH of lakes. Riani and Atkinson (1998) fail to find any evidence for this further transformation for these data. They do find that the log transformation is needed for y_3 (as was indicated by the univariate analysis) and y_5, but that the reciprocal is needed for y_4. This is surprising as the last three responses are all measurements of amounts of chemical elements in the soil.

To see whether these conclusions are affected by particular groups of observations we check the individual transformation of each response using the multivariate statistic, which is on one degree of freedom since the values of four out of the five parameters are held at specified values. For each search we use the four relevant values from the vector $\lambda = (1, 1, 0, -1, 0)$, checking the value of the transformation by use of the likelihood ratio for five values of each parameter. The plots of the 25 forward searches are given in Figure 6, where there is a panel for each variable. Within each panel we give a plot of the signed square root of the likelihood ratio statistic for the five values of λ. Use of the signed square root gives plots similar to that of Figure 5 which cogently illustrate whether lower or higher values of λ are preferred. The value of 1 is acceptable for λ_1 and λ_2 and, as the smoothness of the curves indicates, does not depend on any particular observations. This is very different from variable 3, for which the log transformation is the only possibility. The last stage of the forward search is the addition of either observation 24 or 20, which respectively cause rejection of $\lambda = -0{\cdot}5$ and $\lambda = 0{\cdot}5$, the values on either side of zero. For variable 5 either $-0{\cdot}5$ or 0 are possible values. Finally, for variable 4, either -1 or $-0{\cdot}5$ are possible. These plots clearly show that we cannot find a common transformation for y_3, y_4 and y_5 for m greater than 53. The four observations to be deleted to achieve this are 19, 20, 24 and 55, the last four to be added, in various orders, in all searches leading to acceptable transformations. The source of these observations should thus be checked for anomalies and transcription errors. Whether or not they are deleted, $\lambda = 1$ is acceptable for y_1 and y_2.

A last comment on Figure 6 is that the values of the likelihood ratio statistics for transformation of y_3 are less extreme than those for the approximate score statistic in Figure 5. This is in line with the comparison of statistics by Atkinson and Lawrance (1989) mentioned above.

To conclude this analysis we return to the bivariate boxplots, but now for the transformed data. Figure 7 is a plot of $1/y_4$ against $\log y_3$. Comparison with Figure 1 shows how the joint transformation has achieved nearly elliptical contours and how observations 20 and 55, which seemed to be outlying on the original scale, seem much less so on the transformed scale.

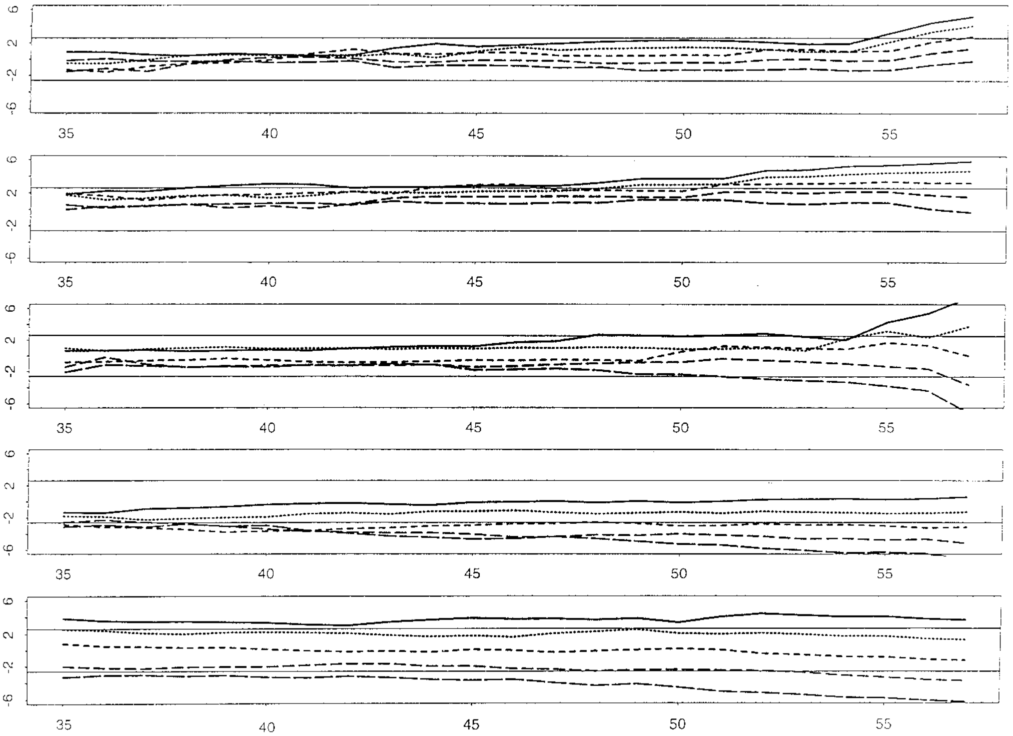

Figure 6. Soil data. Signed square roots of likelihood ratio tests for the standard five values for each component of λ when the other four values are the relevant members of $(1, 1, 0, -1, 0)$. The top panel is for λ_1. In each panel the parameter values are: $\lambda = 1$ (———), $\lambda = 0{\cdot}5$ (– – –), $\lambda = 0$ (- - - - -), $\lambda = -0{\cdot}5$ (· · · · · ·) and $\lambda = -1$ (——). In all 25 searches were required

Figure 8 is the scatterplot matrix for the first four variables after transformation. Comparison of these contours with those of Figure 3 shows how much more elliptical they have become as a result of the transformations and so much closer to the multivariate normal model.

5. DISCRIMINANT ANALYSIS

5.1 An example

As a final demonstration of the power of the forward search for revealing the structure of data, we turn to discriminant analysis. Analogues of regression diagnostics for two-population linear discriminant analysis are derived by Fung (1995). Very robust methods are in Hawkins and McLachlan (1997). Both papers give references to earlier work. There is however surprisingly little graphical interpretation of diagnostic methods for discriminant analysis. We believe our analyses demonstrate how graphics can be fruitfully combined with statistical analysis.

To fix ideas we analyse Fisher's data on three species of iris. There are three groups of multivariate observations, each of 50 observations on four variables. The data are given, for example, by Krzanowski (1988, pp. 46–47) and by Mardia *et al.* (1979, pp. 6–7). The data are often

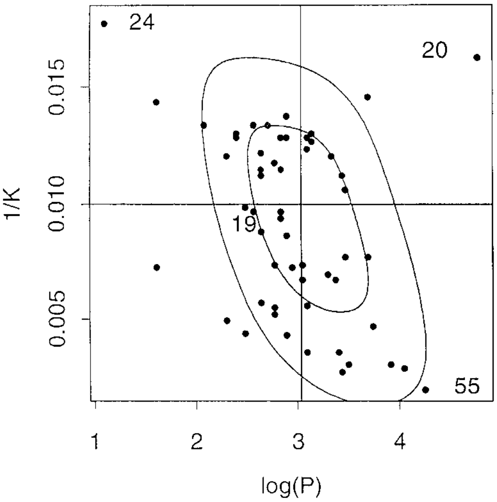

Figure 7. Transformed soil data. Scatterplot of log y_3 and $1/y_4$ with robust contours, which are more elliptical than those of Figure 1

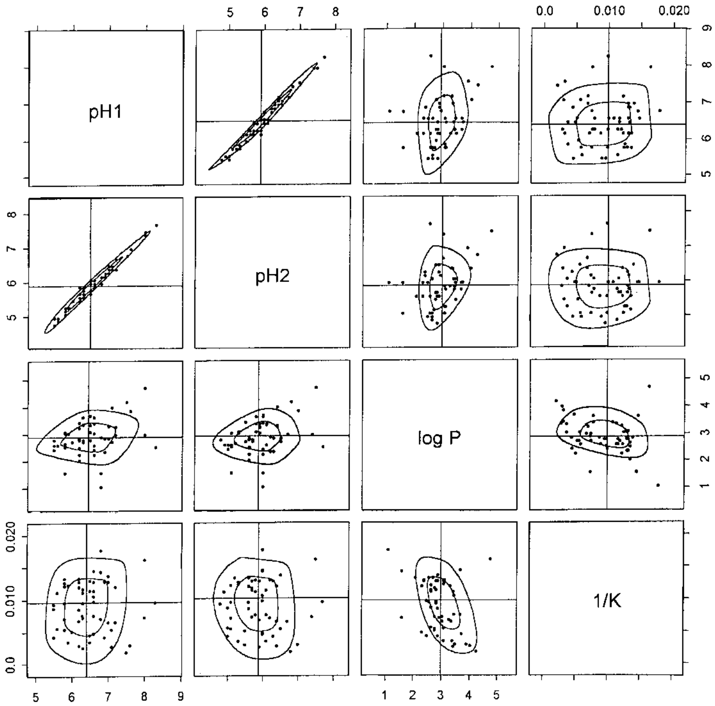

Figure 8. Transformed soil data. Scatterplot matrix of first four variables: y_1, y_2, log y_3 and $1/y_4$. The robust contours are more nearly elliptical than those of Figure 3, showing the effect of transformation

analysed on the original scale (Venables and Ripley 1994, p. 307) but sometimes logs are taken (Venables and Ripley 1994, p. 316). It is customary to use linear discriminant analysis which assumes that the three groups have equal covariance matrices. To test this hypothesis we use the likelihood ratio test for the equality of the matrices given on p. 140 of Mardia *et al.* (1979). For the untransformed data the test has the value 712·7, whereas for the log transformed data it is 782·7. These values are to be compared with χ^2_{20}: clearly on both scales the hypothesis of equality is rejected, so that quadratic discriminant analysis is indicated. We have used both linear and quadratic analyses on the original data and on the log transformed data. The main conclusions remain the same: that group 1 is very different from the other two and that there is a slight overlap between groups 2 and 3, the overlap being slightly greater after the logarithmic transformation. Here we report only our linear analysis of the untransformed data.

Traditional approaches to discriminant analysis evaluate the performance of the allocation rule through cross-validation or sample splitting. The first method, reviewed by Krzanowski and Hand (1997), consists of determining the allocation rule using the sample data minus one observation and then using the consequent rule to classify the omitted observation. Of course this method may suffer from the problem of masking if more than one outlier is present. In the second approach the training set is split randomly into two portions. One portion is used for estimation of the allocation rule itself and the other portion to assess the performance of the rule. This method has been strongly criticized, because future allocations will be made according to a rule based on the whole of the training set, not just on a random portion of it.

In our approach we use the forward search to monitor the evolution of the posterior probabilities as observations are included in the subset. We can then both detect influential observations and determine the effect of each unit on the posterior probabilities, so monitoring the performance of the allocation rule. The forward search is on the Mahalanobis distances, which are strongly linked to changes in the posterior probabilities. In order to have a better understanding of the relation between these two quantities we need some algebra.

Let π_l denote the prior probability of an individual coming from population (group) P_l, $l = 1, \ldots, g$, where g is the number of populations considered. If we indicate by $p(y|l)$ the density of the distribution of the observations for class l, then the posterior probability of class l after observing unit y_k is

$$p(l|y_k) = \frac{\pi_l p(y_k|l)}{p(y_k)} \propto \pi_l p(y_k|l), \quad k = 1, 2, \ldots, n. \tag{12}$$

Following the Bayes rule, we choose the class with maximal posterior probability $p(l|y_k)$. If we assume that P_l is a multivariate normal population with mean μ_l and dispersion matrix Σ_l, the log of the numerator of equation (12) can be written as:

$$-\frac{p}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_l| - \frac{1}{2}(y_k - \mu_l)^{\mathrm{T}}\Sigma_l^{-1}(y_k - \mu_l) + \log \pi_l. \tag{13}$$

Given training sets of size m_l from each population P_l, the parameters μ_l and Σ_l can be estimated by the means and the covariance matrices of these training sets: $\hat{\mu}_l(\mathcal{M}_l)$ and $\hat{\Sigma}_l(\mathcal{M}_l)$. From equation (13) it is clear that the posterior probabilities are positively correlated with the prior probabilities but are negatively related both to the Mahalanobis distances from the various populations and to the determinant of the covariance matrix. The term $|\hat{\Sigma}_l|$ is linked to the Mahalanobis distance by the deletion relation:

$$|\hat{\Sigma}_{l(k)}(\mathcal{M}_l)| = |\hat{\Sigma}_l(\mathcal{M}_l)|\left(\frac{m_l - 1}{m_l - 2}\right)^p \left[1 - \frac{m_l}{(m_l - 1)^2} d_k^2(\mathcal{M}_l)\right], \qquad (14)$$

where $|\hat{\Sigma}_{l(k)}(\mathcal{M}_l)|$ is the determinant of the covariance matrix of a sample size m_l excluding unit k. A large increase of Mahalanobis distance due to inclusion of unit k, therefore, will automatically also produce an increase of $|\hat{\Sigma}_l(\mathcal{M}_l)|$, which is likely to produce a big change in the posterior probability of unit k. Thus a forward search on the Mahalanobis distance of every observation from its own population leads to inclusion in the last steps of the search of those units which most affect the posterior probabilities. That is equations (13) and (14) show that the units which have the largest Mahalanobis distances (potential outliers) are also those which are likely to produce jumps in the plot of the posterior probabilities. If the covariance matrices for all groups are the same, so that we may write $\Sigma m_l = m$, the determinants in equation (13) become equal for all groups. Then we have linear discriminant analysis when the posterior probabilities depend just on the Mahalanobis distances and prior probabilities.

We now analyse some aspects of the iris data to demonstrate how the relationship between the search and the posterior probabilities works in practice. We start by finding an initial basic subset for each group, as we did for one group in Section 3.3, as the intersection of observations within a suitably chosen set of elliptical contours on scatterplot matrices similar to Figure 4. The combination of the three groups of units gives a subset which may not have equal numbers from each group. We first use the forward search to equalize the numbers present in each group, the search again being on Mahalanobis distances, with a common covariance matrix, but, of course, with different means for each group. Once equality of numbers from each group has been achieved, we maintain approximate equality by adding no more than one unit from each group until equality is again achieved. The order in which observations from the three groups are added depends on the Mahalanobis distances, which, since there is a common covariance matrix, are recalculated after each addition. We also experimented with the inclusion of units without paying attention to their group. This tended to result in the inclusion of a set of units from one group, followed by a set of units from another group. The results from such forward searches were more difficult to interpret than those in which we maintained balance.

Figure 9 gives, as a function of subset size, the calculated posterior probabilities that observations in groups 2 and 3 belong to those groups. The plots start with a subset size of 102, out of the total 150 observations. During this period only four units from group 2 ever have posterior probabilities less than 0·6: two, units 71 and 84, are finally misclassified. For group 3 only unit 134 is misclassified. We do not show the plot for group 1 as all units in every step of the search are correctly classified with posterior probabilities of at least 0·99.

The pattern in Figure 9 is stable to the contour used to choose the initial subset. We have experimented with the addition of outliers, which cause noticeable jumps at the end of the plot of posterior probabilities, but will report these results elsewhere. The jumps occur at the end of the series because the outliers are the last units to be included by the forward search.

We also monitored a number of other quantities during the progress of the forward search. Two are plotted in Figure 10. The first panel shows the maximum Mahalanobis distance of those units which belong to the subset for each group. The second panel shows the minimum distance for each group of those units not in the subset: apart from the constraint caused by the need to

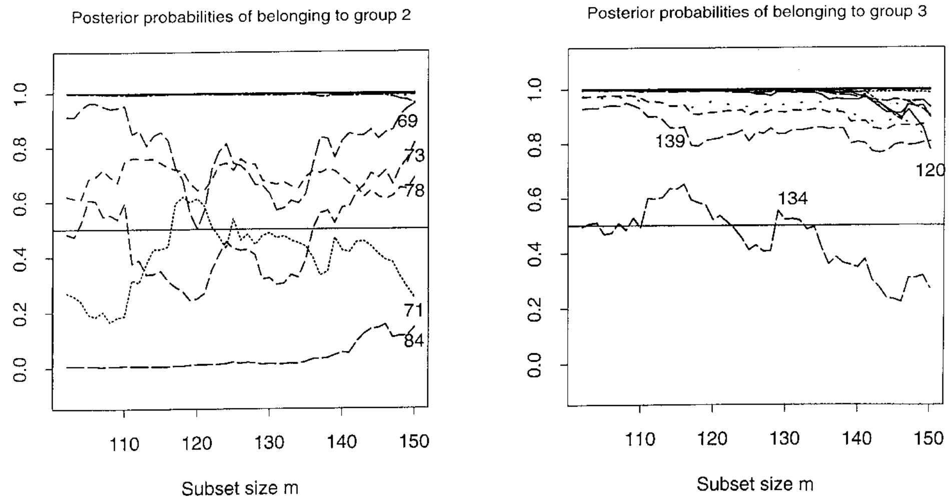

Figure 9. Iris data. Posterior probabilities, as a function of subset size, that observations in groups 2 and 3 respectively belong to those groups

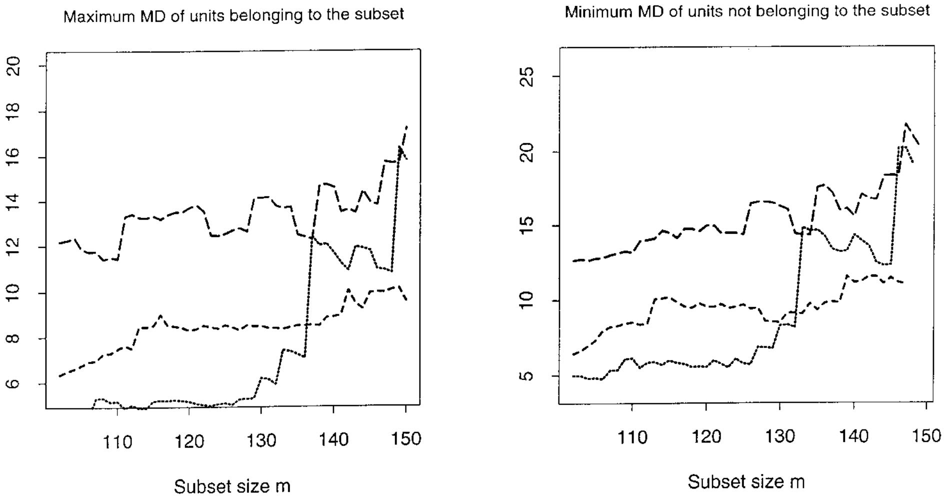

Figure 10. Iris data. Maximum Mahalanobis distances, for the three groups, of units belonging to the subset and minimum distances for those units not belonging: reading upwards in the left half of the plots, groups 1, 2 and 3

keep group sizes equal, these would be the next units to be included in the subset. The first thing to notice is the very different sizes of the distances for the groups. If a different covariance matrix were used for each group, we know that the distances for the m_l units in the subset for the lth group would sum to $v(m_l - 1)$. The pattern shown here is evidence that, in general, the covariance matrices of the three groups should not be treated as equal. Most important, however, is the behaviour of the distances for group 1. The observations entering group 1 from 137 onwards (and which gave large distances in panel 1 from 133 on as the next to enter) are 33, 34, 15, 16 and 42. If these were outliers and the covariances were estimated independently for each group, the effect of these additions on the Mahalanobis distances would rapidly die down: inclusion of these

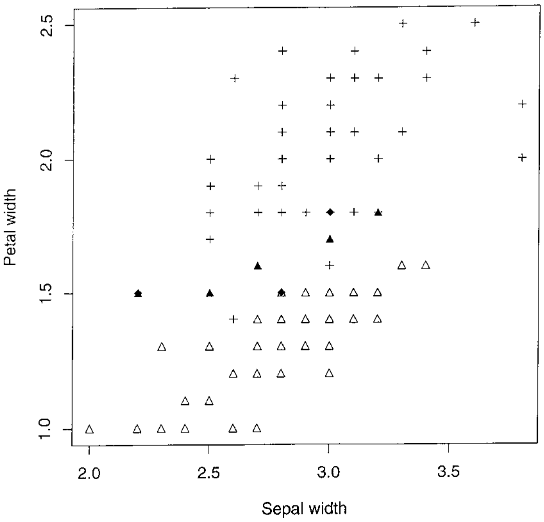

Figure 11. Iris data. Scatterplot of two variables showing, by filled symbols, the units sometimes misclassified. Triangles are for group 2, pluses and diamonds for group 3

units in the estimation of the covariance matrix would rapidly lead to masking. But here a succession of outliers enter from one group, so only have a partial effect on the common covariance matrix. They therefore remain visible in the plot.

In line with our contention of the importance and usefulness of returning from the forward analysis to further inspection of the data, we now give interpretations of these findings in the space of the original data. Figure 11 shows the scatterplot for petal width and sepal length for units in groups 2 and 3. Those which were sometimes misclassified are represented by filled symbols. For two of the three units (71, 84 and 134) which are misclassified at the end of the forward search, there are units in the other group which have identical observed values for these two variables. More precisely unit 71 (coordinates 3·2 and 1·8) presents the same values as unit 126 and unit 134 (2·8 and 1·5) overlaps with 55. Examination of the scatterplot matrix with brushing shows that observations 134 and 55 are very close to each other in all bivariate scatterplots. On the left side of Figure 11, unit 69 (coordinates 2·2 and 1·5) overlaps with observation 120. Among the units of group 3, observation 120 is, apart from 134, the one which shows the smallest posterior probability (0·779) in the last step of the forward search.

The discriminant line dividing the two groups is not shown in Figure 11, but passes close to units 73 (2·5, 1·5), 78 (3, 1·7) and 139 (3, 1·8). As Figure 9 shows the posterior probability of observation 73 fluctuates appreciably even though this unit is always categorized correctly from $m = 136$ onwards. In all steps of the forward search, unit 78 always has a posterior probability around 0·65. Unit 139 is, apart from 134, the one in group 3 showing the smallest posterior probability in almost all steps of the forward search.

The two remaining pluses in Figure 11 which appear close to triangles refer to units 135 (2·6 and 1·4) and 130 (3·0 and 1·6). Unit 135 is the last of the third group to be included in the forward search: it has a final posterior probability of 0·934. During the forward search unit 130 always shows a posterior probability around 0·90 (the final value is 0·896).

There remains unit 84 (2·7, 1·6), the third to be misclassified at the end of the forward search. It is included when $m = 142$. Thereafter the posterior probability that this unit belongs to group 2

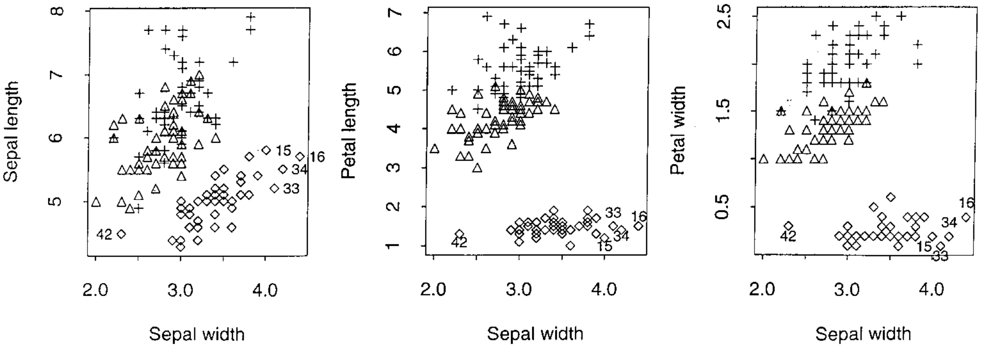

Figure 12. Iris data. Three scatterplots of pairs of variables showing, for group 1 (diamonds), the last units to be included in the forward search. These are the observations yielding the large Mahalanobis distances in Figure 10

tends generally to increase. Its final posterior probability is 0·143. An analysis of the scatterplot matrix reveals that, in almost all the bivariate plots, this unit is surrounded by some observations belonging to group 3.

Finally, in Figure 12, we plot the units which, towards the end of the search gave the large increases in Mahalanobis distances for group 1. The last of all to enter is unit 42, very much an outlier from group 1. The other four units form a cluster at the end of group 1. If these five units are excluded, the bivariate scatters of group 1 are more like those of the other groups. This effect is strongest in the leftmost panel of the figure.

5.2 Conclusions

Our analysis of the iris data shows, we believe, that the forward search technique in discriminant analysis is an extremely useful tool. As a result we can:

1. Highlight the units which are always classified correctly with high posterior probability in each step of the search. These can be separated from those units which are declared correctly only when they are included in the allocation rule.
2. See the evolution of the degree of separation or overlapping among the groups as the subset size increases and determine the relationship with those units which have a posterior probability close to 0·5.
3. Monitor the stability of the allocation rule with respect to different sample sizes.
4. Determine the influence of observations by separating the units with the biggest Mahalanobis distances into two groups: those which have an effect on the posterior probabilities and those which leave them unaltered.

In our example, monitoring the posterior probabilities enabled us to distinguish the units whose posterior probabilities tended to increase as the sample size grew (e.g. units 69 and 73), those whose posterior probability was close to 0·5 (e.g. units 78, 71 and 134) and those which were always completely misclassified (e.g. unit 84). If we have to classify a new unit we can monitor its posterior probability at each step of the forward search. In this way we can have an idea about the stability of the associated allocation and therefore which and how many observations are responsible for its allocation to a particular group.

6. DISCUSSION

We have demonstrated the diagnostic use of the forward search for transformations to normality and for discriminant analysis. Plots as the search progresses of quantities of inferential importance, such as Figures 5, 9 and 10, provide clear and informative indications of the importance of individual observations. Similarly informative plots can be generated for many other statistical procedures. The simplest is probably multiple regression, where outliers and influence on parameter estimates can be monitored as the search progresses. Our results on discriminant analysis indicate how Mahalanobis distances may be usefully monitored in the simpler problem of finding outliers in one multivariate sample. We will report on this work elsewhere, as we will on that on time series: the use of structural modelling combined with efficient modern algorithms for the Kalman filter with missing values makes it possible to fit time series models to very few observations and so to conduct an informative forward search.

ACKNOWLEDGEMENTS

The first author writes: ‘It is a great honour to have been invited to give the John Stuart Hunter lecture. I am most grateful to Dr El-Shaarawi and the organizing committee for the opportunity to visit Innsbruck to present our work.’ We are beholden to Aldo Corbellini for programming help, especially in the preparation of the bivariate boxplots. The assistance of the LSE Staff Research Fund is also acknowledged.

REFERENCES

Atkinson, A. C. (1985). *Plots, Transformations, and Regression*. Oxford University Press, Oxford.

Atkinson, A. C. (1994). ‘Fast very robust methods for the detection of multiple outliers’, *Journal of the American Statistical Association*, **89**, 1329–1339.

Atkinson, A. C. (1995). ‘Multivariate transformations, regression diagnostics and seemingly unrelated regression’, in Kitsos, C. P. and Müller, W. C. (eds), *MODA 4 – Advances in Model-Oriented Data Analysis*. Physica-Verlag, Heidelberg, pp. 181–192.

Atkinson, A. C. and Lawrance, A. J. (1989). ‘A comparison of asymptotically equivalent tests of regression transformation’, *Biometrika*, **76**, 223–229.

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.

Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*, Wiley, New York.

Eilers, P. H. C. and Marx, B. D. (1996). ‘Flexible smoothing with *B*-splines and penalties’, *Statistical Science*, **11**, 89–121.

Fung, W. K. (1995). ‘Diagnostics in linear discriminant analysis’, *Journal of the American Statistical Association*, **90**, 952–956.

Goldberg, K. M. and Iglewicz, B. (1992). ‘Bivariate extensions of the boxplot’, *Technometrics*, **34**, 307–320.

Hadi, A. S. (1992). ‘Identifying multiple outliers in multivariate data’, *Journal of the Royal Statistical Society, Series B*, **54**, 761–771.

Hadi, A. S. and Simonoff, J. S. (1993). ‘Procedures for the identification of multiple outliers in linear models’, *Journal of the American Statistical Association*, **88**, 1264–1272.

Hawkins, D. M. and McLachlan, G. J. (1997). ‘High-breakdown linear discriminant analysis’, *Journal of the American Statistical Association*, **92**, 136–143.

Krzanowski, W. J. (1988). *Principles of Multivariate Analysis*. Clarendon Press, Oxford.

Krzanowski, W. J. and Hand, D. J. (1997). ‘Assessing error rate estimators: the leave-one-out method reconsidered’, *Australian Journal of Statistics*, **39**, 35–46.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, London.

Mulira, H.-M. (1992). *Computational Methods for Transformations to Multivariate Normality*. PhD thesis, Department of Statistical and Mathematical Sciences, London School of Economics.

Riani, M. and Atkinson, A. C. (1998). 'A unified approach to multivariate transformations and multiple outliers', submitted.

Richardson, S. and Green, P. J. (1997). 'On Bayesian analysis of mixtures with an unknown number of components (with discussion)', *Journal of the Royal Statistical Society, Series B*, **59** (in press).

Ruts, I. and Rousseeuw, P. J. (1996). 'Computing depth contours of bivariate point clouds', *Computational Statistics and Data Analysis*, **23**, 153–168.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

Venables, W. N. and Ripley, B. D. (1994). *Modern Applied Statistics with S-Plus*. Springer-Verlag, New York.

Zani, S., Riani, M. and Corbellini, A. (1997). 'Robust bivariate boxplots and multiple outlier detection', to appear.

Zellner, A. (1962). 'An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias', *Journal of the American Statistical Association*, **57**, 348–368.