A Unified Approach to Outliers, Influence, and Transformations in Discriminant Analysis

Marco RIANI and Anthony C. ATKINSON

This article extends the analysis of multivariate transformations to linear and quadratic discriminant analysis. It shows that the standard application of deletion diagnostic techniques for validating a particular transformation suffers from masking and so may fail if several outliers are present. We therefore suggest a simple and powerful method which is based on a forward search algorithm. This robust diagnostic procedure orders the observations from those most in agreement with the suggested model to those least in agreement with it. It provides a unified approach to the detection of influential observations and outliers in discriminant analysis. Simulated and real data are used to show the necessity of considering multivariate transformations in discriminant analysis. The examples demonstrate the power of the suggested approach in revealing the correct structure of the data when this is obscured by outliers.

Key Words: Box–Cox transformation; Deletion diagnostic; Forward search; Masking; Multivariate normality; Transformation to normality; Very robust methods.

1. INTRODUCTION

Most statistical techniques for the analysis of multivariate data rely on the assumption of multivariate normality. In the majority of cases the results obtained by the standard application of these techniques are not robust to departures from this assumption. These departures can be of two kinds: systematic, due to model misspecification, and isolated, due to the presence of outliers. A severe problem is that a few outliers may not only hide each other due to masking, but may also disguise systematic departures. This article introduces a forward search algorithm that provides a robust and diagnostic analysis leading to the detection of outliers and their influence and so to a consequent correct specification of the model to be used in discriminant analysis.

The use of data transformations in discriminant analysis to achieve approximate multivariate normality was discussed by McLachlan (1992, sec. 6.3), who gave several references. Section 2 provides the extension of the univariate power transformation of Box and Cox (1964) to the systematic multivariate transformation of data for linear and quadratic

Marco Riani is Associate Professor, Dipartimento di Economia, Università di Parma, Via J. Kennedy 6, 43100 Parma, Italy (E-mail: mriani@unipr.it). Anthony Atkinson is Professor, Department of Statistics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom (E-mail: a.c.atkinson@lse.ac.uk).

Journal of Computational and Graphical Statistics, Volume 10, Number 3, Pages 513–544

discriminant analysis, thus providing a method for the removal of a form of systematic misspecification.

Section 2 continues by considering isolated departures from the model. The last two decades have seen the publication of many articles about outlier detection in discriminant analysis. Campbell (1980, 1982), in the context of an M-estimation scheme, suggested iterative methods in order to downweight the influence of outliers. Critchley and Vitiello (1991) considered the sensitivity of misclassification probability estimates to the deletion of one case in two population linear discriminant analysis. The quantities suggested by Critchley and Vitiello and in a series of papers by Fung (1992, 1995a, 1995b, 1996, 1998) all depend on two fundamental diagnostic statistics, which are analogous to the residual and leverage measures in regression. Although these methods are helpful for the detection of single outliers, they can fail if masking is present.

Section 2.3 uses a simulated example to show not only the necessity of considering transformations in discriminant analysis, but at the same time the difficulties and the intricacies in the choice of the appropriate transformations when some outliers are present in the data. We exemplify these difficulties by finding a transformation and then by applying diagnostic techniques for its confirmation.

Section 3 describes our robust diagnostic method, which overcomes masking. It is based on the forward search described by Atkinson and Riani (2000) for regression data, which we extend to multivariate observations. It allows us, at the same time, to estimate the transformation of the data and to detect outliers. We use a robust method to find an initial subset which is outlier free. This subset grows in size as we move forward through the data. When we have found the correct transformation, the outliers enter last at the end of the search and we can determine their effects on the estimated transformation and on the properties of the discriminant analysis. We discuss the relationship with other robust methods in Section 3 and, in the latter part of the section, illustrate the excellent behavior of our method on the simulated dataset, showing how we are led to the correct transformation.

Section 4 is devoted to the analysis of data on muscular dystrophy. An earlier analysis used only part of the data, so we are able to perform two analyses, one on the smaller part and then a confirmatory analysis on the whole set of observations, while avoiding an arbitrary division of the data. We find a transformation and sets of observations which are influential for the transformation, as well as outliers which are not influential, and exhibit their effect on the discriminant analysis. The results of our method are displayed through plots that show which are the influential observations, in what way they are influential, and what proportion of the data is compatible with any suggested transformation. Section 5 concludes.

2. TRANSFORMATIONS IN DISCRIMINANT ANALYSIS

2.1 Discriminant Analysis

We begin with a brief definition of discriminant analysis, which serves to establish notation. We then describe some standard procedures for checking the proposed discriminant rule for the effect of outlying observations.

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 that unit i belongs to class l after observing y_i is

$$p(l|y_i) = \frac{\pi_l p(y_i|l)}{p(y_i)} \propto \pi_l p(y_i|l), \qquad i = 1, 2, \ldots, n. \tag{2.1}$$

Following the Bayes rule, we choose the class with maximum posterior probability $p(l|y_i)$. 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 (2.1) can be written as

$$-\frac{p}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_l| - \frac{1}{2}(y_i - \mu_l)^T \Sigma_l^{-1}(y_i - \mu_l) + \log \pi_l. \tag{2.2}$$

Given training sets $S^{(m_l)}_{i_1,\ldots,i_{m_l}} = S^{(m_l)}_*$ of size m_l from each population P_l, the maximum likelihood estimates of the parameters μ_l and Σ_l are the means and covariance matrices of these training sets: $\hat{\mu}_{l(S^{(m_l)}_*)}$ and $\hat{\Sigma}_{l(S^{(m_l)}_*)}$. The squared Mahalanobis distance for observation y_i from population P_l is

$$d^2_{i\left(S^{(m_l)}_*\right)} = \{y_i - \hat{\mu}_{l\left(S^{(m_l)}_*\right)}\}^T \hat{\Sigma}^{-1}_{l\left(S^{(m_l)}_*\right)} \{y_i - \hat{\mu}_{l\left(S^{(m_l)}_*\right)}\}, \tag{2.3}$$

a quantity important in the forward search as well as in the calculation of classification probabilities.

If the hypothesis of equality among covariances is true—that is, $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_g = \Sigma$—then the training sets $(m_1, \ldots, m_g)$ are pooled for estimation of Σ to give the overall training set $(S^{(m)}_*)$ of size $m = \sum_{l=1}^{g} m_l$. The estimated within groups covariance matrix is

$$\widehat{W}_{(S^{(m)}_*)} = \frac{1}{m}\sum_{l=1}^{g}\sum_{i=1}^{m_l}\{y_{il} - \hat{\mu}_{l\left(S^{(m_l)}_*\right)}\}\{y_{il} - \hat{\mu}_{l\left(S^{(m_l)}_*\right)}\}^T, \tag{2.4}$$

and the squared Mahalanobis distances (2.3) become

$$d^2_{i\left(S^{(m)}_*\right)} = \left\{y_i - \hat{\mu}_{l\left(S^{(m_l)}_*\right)}\right\}^T \widehat{W}^{-1}_{\left(S^{(m)}_*\right)} \left\{y_i - \hat{\mu}_{l\left(S^{(m_l)}_*\right)}\right\}. \tag{2.5}$$

If the covariance matrices of the groups are different we obtain quadratic discriminant analysis. But if they are the same, the discriminant function is linear. In practice, linear discriminant analysis is often preferred, because of the reduction in the number of parameters to be estimated to determine the discrimination rule and the consequent decrease in variance of the estimates.

As we will see in this article, transformations of the data can be used to help satisfy the assumptions of normality and of equality of covariance matrices. If the assumptions are not satisfied—that is, if the data are not appropriately transformed—the order of importance of the variables may significantly change and the probability of misallocation may significantly increase.

2.2 TESTS OF TRANSFORMATIONS

One purpose of the transformation is to achieve a common covariance matrix for all groups. However, we initially describe the transformation for the more general case in which each group has its own covariance matrix. Let y_{ijl} be the ith observation on response j for group l $(i = 1, \ldots, n_l;\ j = 1, \ldots, p;\ l = 1, \ldots, g)$. In the extension of the Box and Cox (1964) family to multivariate responses the normalized transformation of y_{ijl} is

$$\begin{aligned} z_{ijl}(\lambda_j) &= \frac{y_{ijl}^{\lambda_j} - 1}{\lambda_j G_j^{\lambda_j - 1}} \qquad (\lambda_j \neq 0) && (2.6)\\ &= G_j \log y_{ijl} \qquad (\lambda_j = 0), && (2.7)\end{aligned}$$

where G_j is the geometric mean of the jth variable. The value $\lambda_j = 1$ corresponds to no transformation of the jth response. If the transformed observations are normally distributed with mean μ_l and covariance matrix Σ_l ($l = 1, \ldots, g$), the log-likelihood of the observations is given by

$$l(\lambda) = -\frac{1}{2}\sum_{l=1}^{g} n_l \log|2\pi\Sigma_l(\lambda)| - \frac{1}{2}\sum_{l=1}^{g}\sum_{i=1}^{n_l}(z_{il} - \mu_l(\lambda))^T\Sigma_l^{-1}(\lambda)(z_{il} - \mu_l(\lambda)), \quad (2.8)$$

where $z_{il} = (z_{il1}, \ldots, z_{ilp})^T$ is the $p \times 1$ vector which denotes the transformed data for unit i coming from group l, $\mu_l(\lambda)$ and $\Sigma_l(\lambda)$ are, respectively, the mean vector and the covariance matrix for population l. Substituting the maximum likelihood estimates $\hat{\mu}_l(\lambda)$ and $\hat{\Sigma}_l^{-1}(\lambda)$ in Equation (2.8), the maximized log-likelihood can be written as

$$l(\lambda) = \text{constant} - \frac{1}{2}\sum_{l=1}^{g} n_l \log|\hat{\Sigma}_l(\lambda)|. \quad (2.9)$$

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

$$T_{\text{LR}} = \sum_{l=1}^{g} n_l \log\{|\hat{\Sigma}_l(\lambda_0)|/|\hat{\Sigma}_l(\hat{\lambda})|\} \quad (2.10)$$

can be compared with the χ^2 distribution on p degrees of freedom. If the hypothesis of equality of covariances is true, then Equation (2.10) becomes

$$n \log\{|\widehat{W}(\lambda_0)|/|\widehat{W}(\hat{\lambda})|\}, \quad (2.11)$$

where $n = \sum_{l=1}^{g} n_l$. In Equations (2.10) and (2.11) $\hat{\lambda}$ is found by numerical search.

If we want to use standard deletion methods to validate a particular set of transformations, we have to delete each observation in turn and every time remaximize the likelihood. Because of the necessity for remaximization, this approach becomes infeasible if multiple deletions are required. In addition, if there is a group of influential observations, the backwards approach may fail due to masking.

Table 1. Simulated Data: Correct Transformation and Index Numbers of Contaminated Units

True transformation: λ^T	0.5	−0.5	−0.5	−0.5
Outliers: original scale	1	5	10	13
Outliers: transformed scale	51	84	92	99

2.3 IMPORTANCE AND DIFFICULTIES OF TRANSFORMATIONS IN DISCRIMINANT ANALYSIS: A SIMULATED EXAMPLE

To illustrate the necessity, but at the same time the difficulties and intricacies of the choice of the appropriate transformation in discriminant analysis, we use an example with simulated data. This has been chosen to illuminate many of the inferential problems that arise in the analysis of the muscular dystrophy data in Section 4.

We generated two groups of 50 observations with the same covariance matrix, so that linear discriminant analysis is appropriate. Group one consisted of a 46 × 4 matrix from a multivariate normal population with mean equal to 7.9 for all variables. The remaining four units were generated from a multivariate normal population with a mean of 10.5 and were included as observations 1, 5, 10, and 13. The 50 × 4 matrix of observations in Group 2 were also from a multivariate normal population, but with mean equal to 6.3. The normal data were transformed by squaring the first variable and raising the remaining variables to the power -2. So the true transformation vector is $\lambda = (0.5, -0.5, -0.5, -0.5)^T$. Finally, in this transformed scale we contaminated four units of the second group by subtracting 1.7 from the values of y_3. A summary of the data structure is given in Table 1.

A correct analysis should find the true transformation and reveal those outliers that are influential either on the transformation or on the misclassification probabilities. The next section shows that our forward method does just that. This section shows the markedly less successful outcome of an analysis using standard techniques, including some deletion diagnostics. First, we quantify the effect on the discriminant analysis of getting the right model. Some results are in Table 2. The nomenclature of the data is important. We intend:

Original or untransformed data to mean the dataset we generated which needs the transformation of Table 1 to achieve normality for 96 of the observations.

Correctly transformed data are the data after this transformation, usually consisting of the $n = 92$ uncontaminated observations.

The results of Table 2 show that finding the correct transformation leads to a three-fold reduction in the average misclassification probability. The discriminant function changes markedly, particularly in the loading on y_1. The indication of the modified Box test for equality of the covariance matrices is that there is clear evidence that they are not equal—the statistic is to be compared with χ^2_{10}. Quadratic discriminant analysis might then be indicated. The results of this table show clearly how one might be misled by failing to transform.

Before finding a transformation, we look at the data. Figure 1 shows the untransformed data—they look very non-normal, but the four outliers on this scale, units 51, 84, 92, and 99 are not at all evident. The outliers on the transformed scale, 1, 5, 10, and 13 are evident in Figure 2. We now consider the effect of the two sets of outliers on estimation of a normalizing transformation.

Table 2. Comparison of Discrimination Results Using Transformed and Untransformed Observations

Test for equality of covariances				
Transformed Data ($n = 92$)			12.58	
Untransformed Data ($n = 100$)			313.1	
Average misclassification probabilities				
		Group 1	*Group 2*	
Transformed data ($n = 92$)		0.04	0.08	
Untransformed data ($n = 100$)		0.11	0.22	
Elements of standardized canonical eigenvector				
	y_1	y_2	y_3	y_4
Transformed Data ($n = 92$)	0.08	0.45	0.38	0.48
Untransformed Data ($n = 100$)	-0.23	0.23	0.32	0.63

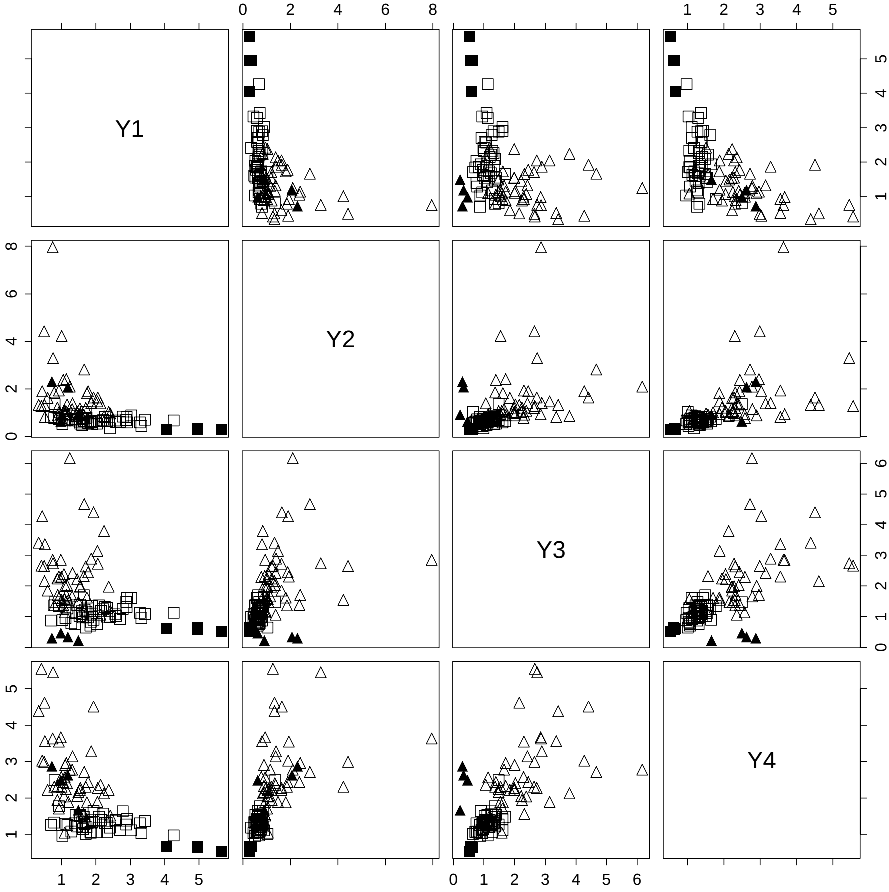


Figure 1. Scatterplot matrix of untransformed simulated data. The filled triangles show the four outliers on this scale—observations 51, 84, 92, and 99—which are not easily detected.

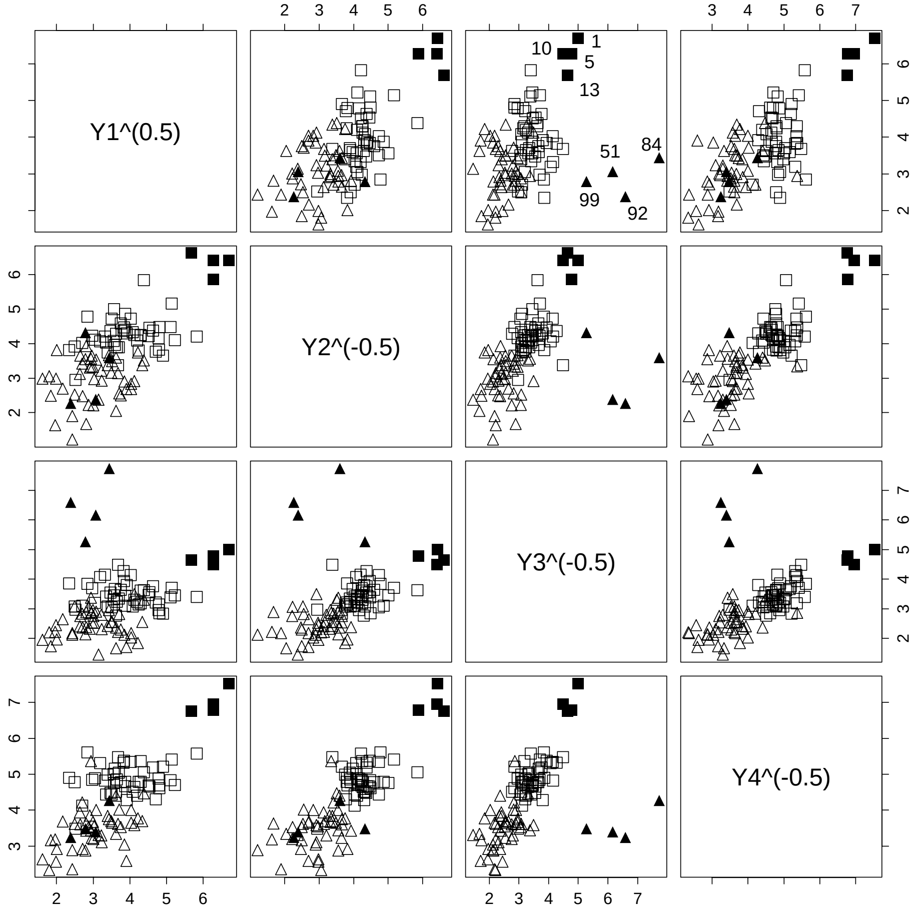


Figure 2. Scatterplot matrix of transformed simulated data. The filled squares show the four outliers on this scale—observations 1, 5, 10, and 13. All outliers are now detectable.

Throughout we assume that one purpose of the transformation is to find a scale in which the two covariance matrices are the same and so we calculate likelihoods for such a model. The maximum likelihood estimates of the transformation parameter, found by numerical search, is $\hat{\lambda} = (0.26, -0.49, 0.28, -0.24)^T$. The value of the likelihood ratio test for the hypothesis of no transformation is equal to 360 and strongly suggests that the data must be transformed. It is usual, and scientifically sensible, to try to find a simple transformation close to $\hat{\lambda}$. The values -1, -0.5, 0, 0.5, and 1 are commonly considered together with 1/3, particularly if the variable is a volume.

Table 3 gives the likelihood ratio tests for several values of λ which seem plausible and close to $\hat{\lambda}$. The tests are to be compared to the χ^2_4 distribution. To detect the effect of individual observations on this inference we consider the effect of the deletion of single observations on these tests. The table gives just the minimum and maximum values from this computationally exhausting procedure—for each deletion the new parameter estimate $\hat{\lambda}_{(i)}$ has to be found by a numerical search in four dimensions. The results presented in Table 3,

Table 3. Deletion Likelihood Ratio Analysis of Simulated Data

	min(LR$_{(i)}$)	*LR*	*max(LR$_{(i)}$)*
H$_0$: $\lambda = $ (0,0,0,0)	24.489	34.954	40.099
H$_0$: $\lambda = $ (0.5,−0.5,0.5,0)	8.706	11.996	15.518
H$_0$: $\lambda = $ (0.5,−0.5,−0.5,−0.5)	38.556	55.882	62.112
H$_0$: $\lambda = $ (1/3,−0.5,1/3,0)	3.042	4.847	6.326

therefore, required 401 maximizations. A computationally less demanding alternative to these exact calculations is the use of constructed variables to find approximate deletion statistics (Atkinson 1995). Since our main interest is not in deletion diagnostics, we do not explore this method here.

We now consider the results of Table 3. The maximum likelihood estimates suggest testing whether a common transformation is possible for all variables. The results of the first line of the table show that the log transformation for all the variables is incompatible with the data. If the maximum likelihood estimates are rounded to the five most common values of λ, one obtains the hypothesis $\lambda = (0.5, -0.5, 0.5, 0)^T$. The second line of the table shows that the minimum value of the deletion likelihood ratio test is below the $\chi^2_4(0.95)$ threshold of 9.49. One could then start a backward procedure, deleting the observation associated with the minimum value of the deletion likelihood ratio. The third row of Table 3 shows the results when one tries to validate the correct transformation, which is the hypothesis that $\lambda = (0.5, -0.5, -0.5, -0.5)^T$. The smallest value of the deletion likelihood ratio is well above the $\chi^2_4(0.99)$ threshold suggesting that this combination of values of λ must be firmly rejected. Finally, the last line shows the results when the null hypothesis is that $\lambda = (1/3, -0.5, 1/3, 0)^T$, a combination which comes from rounding the maximum likelihood estimates of the first and third transformation parameters to 1/3 and the others to one of

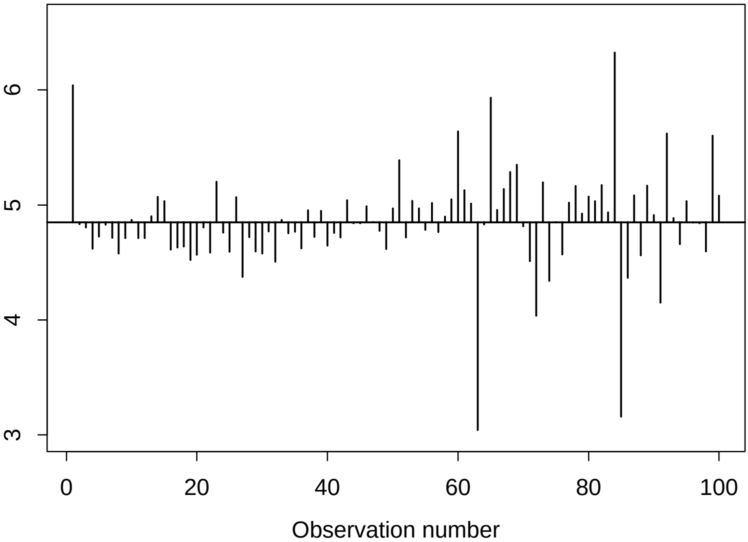


Figure 3. Incorrectly transformed simulated data: single deletion likelihood ratio test, which supports the hypothesis λ = (1/3, −0.5, 1/3, 0)^T.

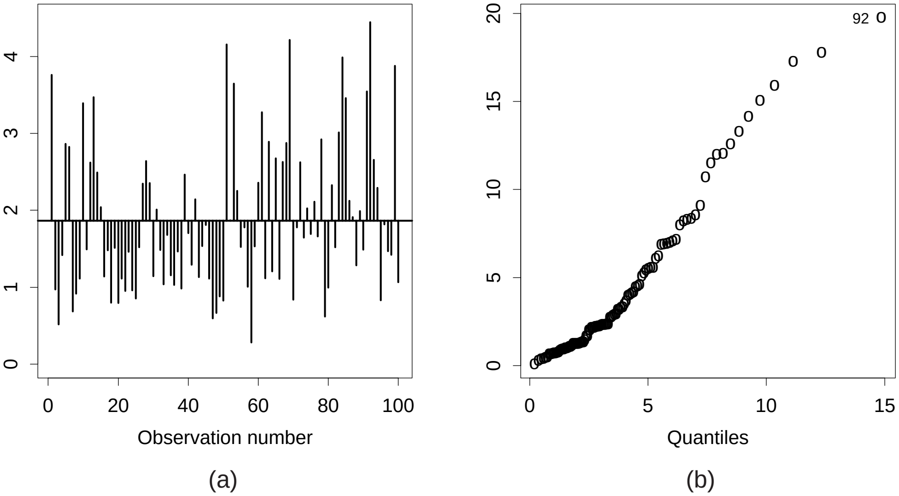


Figure 4. Incorrectly transformed simulated data. Deletion Mahalanobis distances: (a) index plot, (b) QQ-plot against chi-squared on four degrees of freedom.

the five most common values of λ. The maximum deletion value of the likelihood ratio is below the $\chi^2_4(0.95)$ threshold, while the final value (4.847) is very close to the expectation of the χ^2_4 distribution.

As a result of this diagnostic analysis one might think that λ = (1/3, -0.5, 1/3, 0)^T would be a good transformation, since the values of the statistic are always within the 95% confidence boundary. The plot of this deletion statistic is in Figure 3. As the summary in the table indicated, there are no obviously influential observations.

We now check this selected transformation for outliers, which might have an inflationary effect on the estimation of the covariance matrix. Figure 4(a) is a plot of the deletion Mahalanobis distances for the transformation $\lambda = (1/3, -0.5, 1/3, 0)^T$. There are no obviously large values, although some are clearly larger than others. Because squared Mahalanobis distances in this example should follow the χ^2_4 distribution, we give in Figure 4(b) a QQ plot of the squared distances against the chi-squared order statistics. There is no evidence of outliers from this smooth and uneventful plot.

This simple example shows how extremely difficult it is, when outliers are present, to find the correct transformation in discriminant analysis when the starting point is the maximum likelihood estimates of the parameters based on all the observations.

3. A ROBUST AND EFFICIENT APPROACH TO TRANSFORMATION IN DISCRIMINANT ANALYSIS

3.1 General Principles of the Forward Search

If the values of the parameters of the model were known, there would be no difficulty in detecting the outliers, which would have large Mahalanobis distances. The difficulty arises because the outliers are included in the data used to estimate the parameters, which can then, as we saw in the earlier example, be badly biased. Most methods for outlier detection therefore seek to divide the data into two parts, a larger "clean" part and the outliers. The clean data are then used for parameter estimation. The simplest example of this division of the data into two parts is in the use of single deletion diagnostics, where the division is into one potential outlier and the rest of the data. However, if multiple deletions are required, there is a combinatorial explosion of the number of cases that have to be considered by such backwards working.

Many methods for the detection of multiple outliers, therefore, use very robust methods to sort the data into a clean part and potential outliers. In the resampling algorithm for the detection of multivariate outliers using the minimum volume ellipsoid (MVE) (Rousseeuw and van Zomeren 1990) the model is fitted to $p+1$ observations. The resulting parameter estimates are very robust, but are defined by the algorithm which produces them, a crucial distinction with standard statistical methods such as maximum likelihood.

In the forward search, larger subsamples of outlier free observations are found by starting from small subsets and incrementing them with observations which have small Mahalanobis distances, and so are unlikely to be outliers. The method was introduced by Hadi (1992) for the detection of outliers in regression. The emphasis in this and related papers—for example, Atkinson (1994) which considers multivariate problems—is on using the forward search to find a single set of parameter estimates and of outliers. These are determined by the point at which the algorithm stops, which may be either deterministic or data dependent. Our forward search runs to the end of the data, monitoring quantities of interest at each step. If it were stopped early at a particular point, say when just over half the data had been fitted, it would provide very robust parameter estimates. The comparisons in Atkinson and Cheng (2000) show this use of the forward search in calculating least trimmed squares estimators in regression. The fast algorithm for minimum covariance determinant estimation of Rousseeuw and van Driessen (1999) likewise uses ordering of Mahalanobis distances with an initial small subset. Very robust estimators in discriminant analysis were studied by Hawkins and McLachlan (1997) and by He and Fung (2000). Stopping our search early would give estimators with similar very high breakdown points.

This article not only extends the forward search to discriminant analysis, but uses it in a completely different way: at each stage of the forward search we use information such as parameter estimates and plots of Mahalanobis distances to guide us to a suitable model. Our interest is in the evolution of the likelihood ratio test for transformations as m goes from r to n (where $p << r < n$), together with the evolution of Mahalanobis distances and of the misclassification probabilities. We monitor changes which occur, which will be associated with the introduction of a particular observation into the subset m used for fitting.

Remark 1: The search starts with a robust estimator described in the next subsection, which is based on a subset of r observations. Let this be $\hat{\theta}_r^*$ and let the usual estimator of means and covariances at the end of the search be $\hat{\theta}_n^* = \hat{\theta}$. In the absence of outliers and systematic departures from the model

$$E(\hat{\theta}_r^*) = E(\hat{\theta}) = \theta,$$

that is, both parameter estimates are unbiased estimators of the same quantity. The same property holds for the sequence of estimates $\hat{\theta}_m^*$ produced in the forward search. Therefore, in the absence of outliers, we expect parameter estimates, likelihood ratio tests, and Mahalanobis distances to remain sensibly constant during the forward search. We shall see in examples that this is so.

Remark 2: Now suppose there are k outliers. Starting from a clean subset, these will have large Mahalanobis distances and the forward procedure will include them towards the end of the search, usually in the last k steps. Until these outliers are included, we expect that the conditions of Remark 1 will hold and that plots of the Mahalanobis distances and parameter estimates will remain sensibly constant until the outliers are incorporated in the subset used for fitting.

Remark 3: Outliers in one transformed scale may not be outliers in another scale. If the data are analyzed using the wrong transformation, the k outliers may enter the search well before the end.

The forward search algorithm is made up of three steps: the first concerns the choice of an initial subset, the second refers to the way in which we progress in the forward search and the third relates to the monitoring of the statistics during the progress of the search.

3.2 STEP 1: CHOICE OF THE INITIAL SUBSET

We find an initial subset of moderate size by robust analysis of the matrix of bivariate scatterplots. The initial subset of r observations, which we denote with $S_*^{(r)}$, consists of those observations which are not outlying on any scatterplot, found as the intersection of all points lying within a robust contour containing a specified portion of the data (Riani and Zani 1997) and inside the univariate boxplot. More formally, let $S_{i_1,\ldots,i_{r_l}}^{(r_l)}$ be the initial subset of size r_l from group l $(r_l \leq n_l)$. A unit i from group l belongs to this subset if $i \in \cap_{j,k=1}^{p} E_{jk}$ where E_{jk} denotes a robust bivariate contour with $(1-\alpha)$ level if $j \neq k = 1, 2, \ldots, p$ or the traditional univariate boxplot if $j = k$. The overall initial subset is found as: $S_*^{(r)} = \cup_{l=1}^{g} S_{i_1,\ldots,i_{r_l}}^{(r_l)}$. There are two versions of the robust bivariate contour. The first uses convex hull peeling and B-spline smoothing (Zani, Riani, and Corbellini 1998). The second, more simple but less robust, is based on robust ellipses (Riani and Zani 1997). An important property of this method is that the size of the subset can easily be increased or decreased by changing the level of the contour. In the examples that follow we always use the quicker version, because we simply need to find an initial subset of a certain size which is not contaminated by masked outliers.

3.3 STEP 2: ADDING OBSERVATIONS DURING THE FORWARD SEARCH

In every step of the forward search given a subset $S_*^{(m)}$ of size m $(m = r, \ldots, n-1)$, $(m = m_1 + \cdots + m_g)$, we move to a subset of size $(m+1)$. The selection of these $(m+1)$ units can be unconstrained or constrained to obtain a balance across groups.

Unconstrained. In every step of the forward search we simply select the $(m+1)$ units with the $(m+1)$ smallest Mahalanobis distances.

Constrained. Let R_l be the ratio between the number of units of universe l in the subset and the global sample size: $R_l = m_l/n_l$, $l = 1, \ldots, g$. We first select the groups with the smallest R_l. Among these we increase by one unit the group which has the smallest m_{l+1}th ordered Mahalanobis distance (which we denote by $d_{[m_{l+1}]}$). For example, if $d_{[m_{l+1}]}$ is in group s, the new subset is formed by the units associated with the following distances:

$$\begin{aligned} d_{[m_1]}, \ldots, d_{[m_l]}, \qquad & l \neq s = 1, \ldots, g; \\ d_{[m_1]}, \ldots, d_{[m_{s+1}]}. & \end{aligned}$$

In the constrained search we therefore impose the condition that the subset in every step of the forward search must contain proportions of the units which agree, as closely as possible, with the proportions in the overall sample.

Remark 1: Since we progressively include units with small Mahalanobis distances, the units of a group with small variance will tend to be included by the unconstrained search before those from a group with large variance. The problem is particularly severe when we are considering transformations, when, before transformation, the groups may have very different variances. When an unconstrained search is used, the analysis of the group to which the last units belong provides information about differences in variance between the groups. Both constrained and unconstrained searches provide useful information about the structure of the data.

In most moves from m to $m+1$ just one new unit joins the subset. Occasionally two or more units join $S_*^{(m)}$ as one or more leave, an unusual event, which occurs when the search includes one unit which belongs to a cluster of outliers. At the next step the remaining outliers in the cluster seem less outlying and so several may be included at once. Of course, several other units then have to leave the subset.

The forward search estimator $\hat{\theta}_{\text{FS}}$ is defined as the collection of maximum likelihood estimators in each step of the forward search; that is,

$$\hat{\theta}_{\text{FS}} = \left(\hat{\theta}_r^*, \ldots, \hat{\theta}_n^*\right). \tag{3.1}$$

We use a robust method for the estimation of $\hat{\theta}_r^*$ and maximum likelihood (that is fully efficient) estimators during the remainder of the search. The zero breakdown point of maximum likelihood estimators has advantages in the forward search. The introduction of atypical (influential) observations is signalled by sharp changes in the curves which monitor parameter estimates, misclassification probabilities, or other statistics at every step. The robustness of the method comes from the progressive inclusion of units into a subset which, in the first steps, is outlier free.

The search which we use avoids the inclusion of outliers in the first steps and provides a natural ordering of the data according to the specified null model. It is therefore possible to know how many observations are compatible with a particular specification. Furthermore, this approach enables us to analyze the inferential effect of the atypical units (outliers) on the results of statistical analyses.

Remark 2: Our approach is not sensitive to the method used to select an initial subset. For example, the criterion based on robust bivariate boxplots can be replaced by estimation of the minimum volume ellipsoid to provide an initial subset of size $r = p + 1$, without affecting the last, important, part of the search. The first few steps of our searches are occasionally quite active, as potentially outlying observations are identified and removed. But the final, informative, third of the forward search is insensitive to the precise selection of the initial subset.

3.4 MAHALANOBIS DISTANCES AND DISCRIMINANT ANALYSIS IN STEP 2

In our approach we use the forward search on Mahalanobis distances to monitor the evolution of the posterior probabilities as observations are included in the subset.

We now consider the link between Mahalanobis distances and the probabilities.

From Equation (2.2) 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

$$\left|\hat{\Sigma}_{l\left(S^{(m_l-1)}_{i_1,\ldots,i_{m_l-1}}\right)}\right| = \left|\hat{\Sigma}_{l\left(S^{(m_l)}_{i_1,\ldots,i_{m_l}}\right)}\right| \left(\frac{m_l-1}{m_l-2}\right)^p \left[1 - \frac{m_l}{(m_l-1)^2} d^2_{i_{m_l}\left(S^{(m_l)}_{i_1,\ldots,i_{m_l}}\right)}\right]. \tag{3.2}$$

A large increase of Mahalanobis distance due to inclusion of unit i_{m_l}, therefore, will automatically also produce an increase in $|\hat{\Sigma}_{l(S^{(m_l)}_{i_1,\ldots,i_{m_l}})}|$, which is likely to produce a big change in the posterior probability of unit i_{m_l}. 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. Equations (2.2) and (3.2) 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 the determinants in Equation (2.2) become equal for all groups. Then we have linear discriminant analysis when the posterior probabilities depend just on the Mahalanobis distances and prior probabilities.

3.5 STEP 3: MONITORING THE SEARCH

Outliers and influential observations can be detected by simple graphical displays of statistics involved in the forward search. It is extremely useful to monitor particular squared Mahalanobis distances such as:

$$d^2_{[m_l]\left(S^{(m_l)}_*\right)} \qquad m = r, \ldots, n \qquad l = 1, \ldots, g \tag{3.3}$$

and

$$d^2_{[m_l+1]\left(S^{(m_l)}_*\right)} \qquad m = r, \ldots, n-1 \qquad l = 1, \ldots, g. \tag{3.4}$$

Statistics in Equations (3.3) and (3.4), respectively, refer to the maximum Mahalanobis distance in the subset and the minimum Mahalanobis distance among the units not belonging to the subset.

If the dispersion among the groups is markedly different, the curve of $d^2_{[m_t+1](S_*^{(m_t)})}$ never overlaps that of $d^2_{[m_l+1](S_*^{(m_l)})}$, $l \neq t = 1, \ldots, g$. Moreover, these curves give, for each group, a series of outlier tests comparing the observation about to be introduced with those already in. If one or more atypical observations are present, the plot of $d^2_{[m_l+1](S_*^{(m_l)})}$ shows a peak in the step prior to the inclusion of the first outlier, while the plot monitoring $d^2_{[m_l](S_*^{(m_l)})}$ shows a sharp increase when the first outlier joins $S_*^{(m)}$. This curve may also show a subsequent decrease due to masking.

Another way to examine the differences in variability between the two groups is to monitor

$$\log |\widehat{W}_{\left(S_*^{(m)}\right)}| \qquad m = r, \ldots, n; \tag{3.5}$$

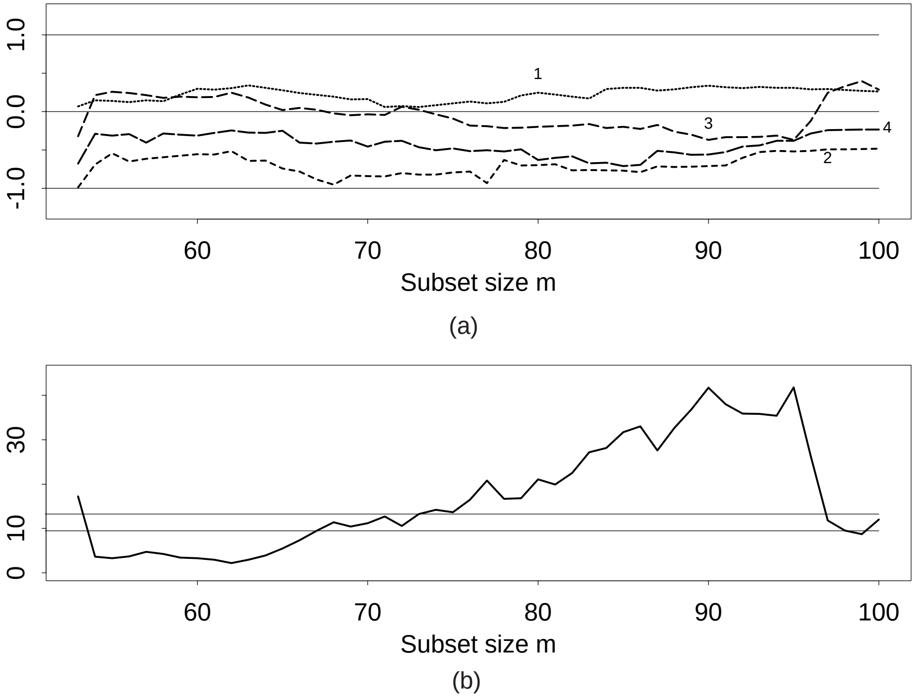
that is, the logarithm of the determinant of the estimated within groups covariance matrix defined in Equation (2.4). Whether or not a balanced search is used, this plot will be an approximately straight line when the groups have the same variability: each observation, regardless of group, will make much the same contribution to $\widehat{W}$. However, if the variability in the groups is different and a balanced search is used, the plot will have a zig-zag form.

In the following sections we will refer to Equations (3.3), (3.4), and (3.5) as monitoring the "maximum distance," the "minimum distance," and the "pooled determinant."

3.6 FINDING A TRANSFORMATION WITH THE FORWARD SEARCH

With just one variable for transformation it is comparatively easy to use our forward search to find satisfactory transformations, if such exist, and the observations that are influential in their choice. Riani and Atkinson (2000) performed a forward search using five standard values of λ $(-1, -0, 5, 0, 0.5$, and $1)$ and monitor the score statistic for transformations for each search. However, with four variables for transformation there would be 625 combinations of the standard values. Whether or not the calculations are time consuming, trying to absorb and sort the information would be difficult. We therefore suggest three steps to help structure the search for a transformation:

1. Run a forward search through the data, ordering the observations at each m by Mahalanobis distances calculated from untransformed observations. Estimate λ at each value of m. Use the results to select a set of transformation parameters.
2. Rerun the forward search using distances calculated with the parameters selected in the first step, again estimating λ for each m. If some change is suggested in λ, repeat this step until a reasonable set of transformations has been found. Let this be λ_R.
3. Test the suggested transformation. We expand each transformation parameter in turn around the five common values of λ $(-1, -0.5, 0, 0.5, 1)$, using the values of the vector λ_R for transforming the other variables. In this way we turn a multivariate



(a)

(b)

Figure 5. Simulated data. Forward search ordering Mahalanobis distances based on λ = (0.5,−0.5,0.5,0): (a) MLE of transformation parameters. The numbers by the curves are the labels of the y; (b) likelihood ratio test for the parameter value used in the search. The horizontal lines are the 5% and 1% points of the associated χ^2 distribution.

problem into a series of univariate ones. In each search we can test the transformation by comparing the likelihood ratio test with χ^2 on 1 degree of freedom. But we use the signed square root of the likelihood ratio in order to learn whether lower or higher values of λ are indicated.

In Step 2 of the procedure we suggest using an unconstrained search because, if the data are appropriately transformed, a cluster of k outliers from one group will enter the subset in the last k steps of the forward search. We now exemplify this procedure for the simulated data discussed in Section 2.3 using unconstrained searches. We notice however, that the conclusions we reach are not at all affected by this choice.

Step 1. We start with a forward search using for ordering the Mahalanobis distances based on untransformed observations. The curves for the parameter estimates, which we do not show, are stable apart from that for λ_3 which shows a jump when the subset size m is between 70 and 80. If we consider only the five most common values of λ, this plot suggests taking $\lambda = (0.5, -0.5, 0.5, 0)^T$ in the second search.

Step 2. Figure 5 shows the results of this new search. The maximum likelihood estimates in Figure 5(a) are much more stable during the central part of the search than they were for the search in Step 1. The jump in values for λ_3 now occurs between $m =$ 95 and

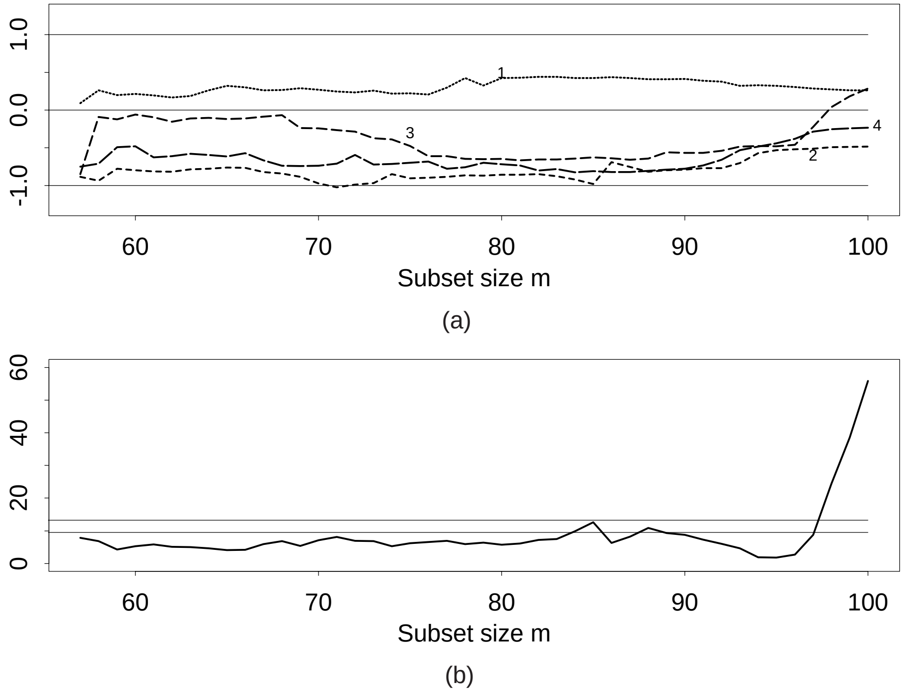


Figure 6. Simulated data. Forward search ordering Mahalanobis distances based on λ =(0.5,−0.5,−0.5,−0.5): (a) MLE of transformation parameters; (b) likelihood ratio test for the parameter value used in the search. The outliers now enter at the end of the search. The numbers by the curves are the labels of the y.

$m = 98$. This means that we have begun to find a transformation in which the outliers are clearly identified and enter near the end of the search. If the transformation were correct the outliers would enter at the end. But the plot of the likelihood ratio test for the null hypothesis that the value used in ordering is correct, shows in Figure 5(b) that evidence against these values accumulates steadily during the forward search until the very end, when the introduction of the last few observations causes the transformation to become seemingly acceptable.

The next combination of values of λ which is suggested by the stable part of the search in Figure 5(a) up to $m = 95$ is $\lambda = (0.5, -0.5, -0.5, -0.5)^T$. The forward search using these new values is shown in Figure 6. In this case all the estimates along the forward search are as specified in our vector λ until $m = 97$. The plot of the likelihood ratio in Figure 6(a) shows that it is the last three observations to enter which cause rejection of our initial estimate. We therefore take $\lambda_R = (0.5, -0.5, -0.5, -0.5)^T$ for further diagnostic calculations.

Table 4. Order of Inclusion of Observations for Search With $\lambda_R = (0.5, -0.5, -0.5, -0.5)^T$

Step (*m*)	92	93	94	95	96	97	98	99	100
Unit Included	69	5	13	10	1	99	51	92	84

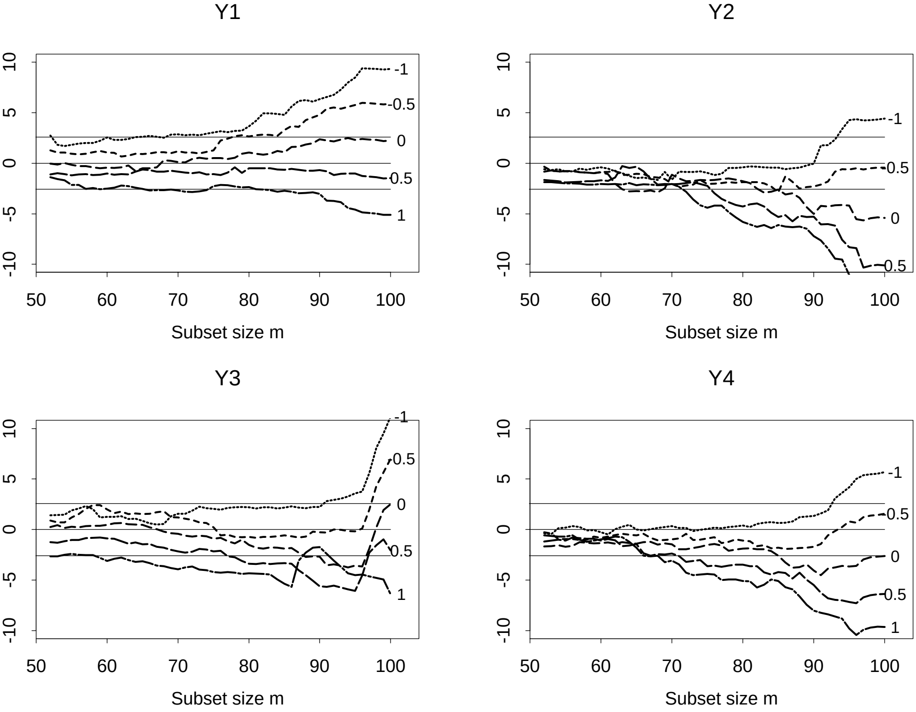


Figure 7. Simulated data: signed square root of the likelihood ratio test for transformation, confirming $\lambda_R =$ (0.5,−0.5,−0.5,−0.5). The numbers by the curves are the values of λ.

Step 3. As a result of our forward analysis we have easily recovered the transformation in Table 1 which leads back to normality. Equally important, we can now identify the outliers that were causing difficulty in finding these transformations. Table 4 gives the order of inclusion of the last nine units in the forward search with λ_R. The last four are the units that are outliers on the scale in which we started to analyze the data. The four before are outliers once we have correctly transformed the data.

We now identify the effect of the individual influential observations by running searches for the five standard values for each parameter in turn. Figure 7 shows the signed square root of the likelihood ratio test for 20 forward searches (5 for each variable) around the vector λ_R. This plot (using 99% confidence bands) shows that for the first variable the square root transformation is the best. Even if the null hypothesis of log transformation cannot be rejected at the 99% level the corresponding curve is very close to the rejection line. For the second and fourth variables only $\lambda = -0.5$ seems to be compatible with the data. Finally, the most interesting plot is the one for variable 3. The hypothesis $\lambda = -0.5$ is perfectly in agreement with the data up to the inclusion of the last four units which are observations 99, 51, 92, and 84. The effects of these four units in the five searches are very different. When $\lambda = -1$ or $\lambda = -0.5$ they are included in the last four steps and cause a big jump in the value of the statistic. When $\lambda = 0$ and $\lambda = 0.5$ they are included around the end of the search and cause the value of the statistic to re-enter inside the confidence bands. Using $\lambda = 1$ these 4 observations are included in the middle of the search and

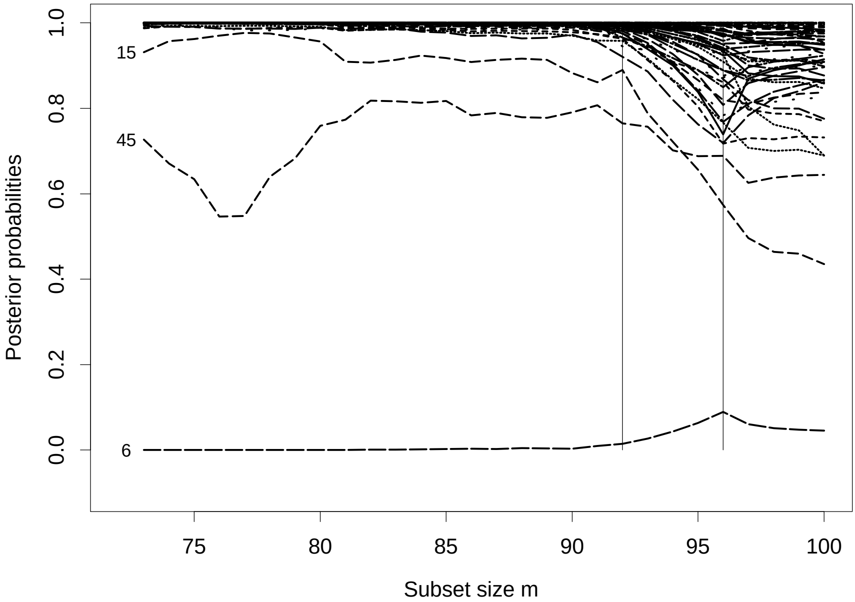


Figure 8. Simulated data: posterior probabilities of correct classification of units of group one after correct transformation λ_R .

cause an upward jump. In this scale, however, the steady downward trend caused by all the other observations again brings the value of the statistic below the lower threshold. A fuller discussion of similar behavior for a different example with four masked outliers was given by Atkinson and Riani (2000, p. 107).

The plot in Figure 7 is easy to understand and at the same time very powerful because it clearly shows which and how many observations are responsible for a particular transformation and what is the inferential effect of each unit on the results of statistical analyses.

3.7 Discriminant Analysis and Confirmation of the Transformation

We now look at the effect of our transformation on the behavior of the discriminant analysis. Since there is no explicit connection between the Box–Cox transformation and discrimination, we repeated our analysis of transformations by finding values of λ which minimized the misclassification probabilities, rather than maximizing the likelihood (2.8). The answers were not very clear. The transformations of some variables were reasonably defined, but for those with a small weighting in the canonical eigenvector the transformation was very poorly defined. Transformation of such a variable has little effect on the discrimination, even if the small value is caused by outliers. We therefore did not pursue this procedure.

The results of Table 2 stress the importance, for discriminant analysis, of correctly transforming the data. We now consider the much more detailed information on discrimination

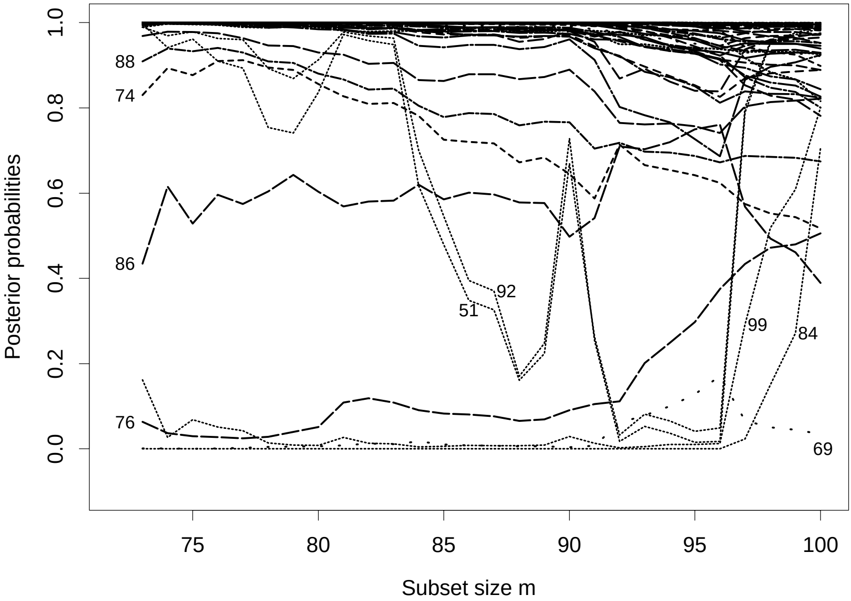


Figure 9. Simulated data: posterior probabilities of correct classification of units of group two after correct transformation λ_R.

which is gathered from the forward search.

Figure 8 gives the probabilities of correct classification of units in Group 1 during the last third of the search. There are three units—6, 45, and 15—which are not well classified. The classification of the other units remains stable and good until the search reaches $m = 92$. Then the four outliers from Group 1 enter, one after the other, and cause the probabilities to worsen. At the end of the search units from Group 2 enter and do not have much effect on the probabilities in Group 1.

The plot for Group 2 is much more dramatic. We see from Figure 9 that the outliers on the original scale of our data are even more outlying after the data have been transformed—as would be expected from their large influence on the transformation. When these observations are introduced, the probabilities change appreciably and the units move to being correctly classified. But there remain two observations, 74 and 86, on the boundary of the two groups, for which the probabilities oscillate during the search. Also the naturally outlying 69 is continually misclassified.

Finally in Figure 10 we look at the monitoring of the Mahalanobis distances and Box statistic during the search. Figure 10(a) shows the maximum distance monitoring plot. In the first stages the plot shows that the maxima are very close—there is no evidence of any difference in variance in the two groups. When $m = 92$, observation 69 enters Group 2 (the order of inclusion of the last observations is given in Table 4) and the distance jumps up. After that four outlying units enter Group 1, causing a jump in the maximum distance for that group. As successive units enter there is a slight decrease due to masking, but observation 1 causes a slight increase. After this the last four units enter Group 2—there is

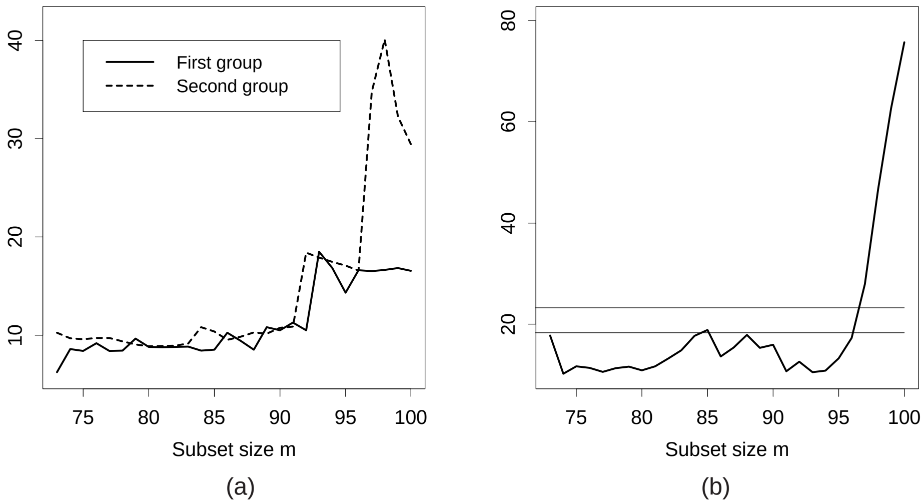


Figure 10. Correctly transformed simulated data. Monitoring, during the forward search, (a) the maximum Mahalanobis distance of units included in the subset and (b) Box test of equality of covariance matrices.

again a big increase in the distance for Group 2, which then drops back a bit due to masking. The distances for Group 1 are hardly altered by the introduction of these four units, despite the common estimate of the covariance matrix used. Two conclusions from this plot are that initially the within group variances are very similar and that we do not need to constrain the search to be balanced—a situation very different from that of the example in the next section. And, second, that the outliers, influential or not, are entering at the end of the search.

We finish with Figure 10(b) in which we monitor the Box statistic for equality of variances. This is constant in the earlier stages of the search, below the 5% point, showing no evidence of any inequality. But, at the end of the search, starting with observation 97, the evidence for nonequality of covariance matrices builds steadily. The starting point of $m = 97$ for this information agrees with the plot of Figure 10(a) where, although there is an increase in distances from $m = 92$, the distances for both groups initially increase together. Only from $m = 97$ are the two curves markedly different.

Our exemplary analysis of these data has enabled us, in a straightforward manner, to recover all the features that we built into the data and to assess their effects on discrimination. We now use this experience to analyse a set of data about which we have no such prior knowledge.

4. MUSCULAR DYSTROPHY DATA

4.1 The Data

We now apply our method to a real dataset. Duchenne Muscular Dystrophy (DMD) is a genetically transmitted disease, passed from a mother to her children. Affected male offspring may unknowingly carry the disease but male offspring with the disease die at a young age. Although carriers of DMD usually have no physical symptoms, they tend to exhibit elevated levels of serum markers. In addition, the levels of these enzymes may also depend on age and season. Levels of the enzymes were measured in noncarriers and in a group of carriers using standard laboratory procedures. The variables used are: Age $(\mathrm{AGE} = y_1)$, month of the year $(\mathrm{M} = y_2)$, creatine kinase $(\mathrm{CK} = y_3)$, hemopexin $(\mathrm{H} = y_4)$, lactate dehydrogenase $(\mathrm{LD} = y_5)$, and pyruvate kinase $(\mathrm{PK} = y_6)$. The first two serum markers (CK and H) may be measured rather inexpensively from frozen serum. The second two (LD and PK) require fresh serum. An important scientific problem is whether use of the expensive second pair of readings causes an appreciable increase in detection. A further feature of the data is that the water supply to the laboratory was changed in the course of the study, although when is not recorded. It is therefore likely that some outliers are present in the data.

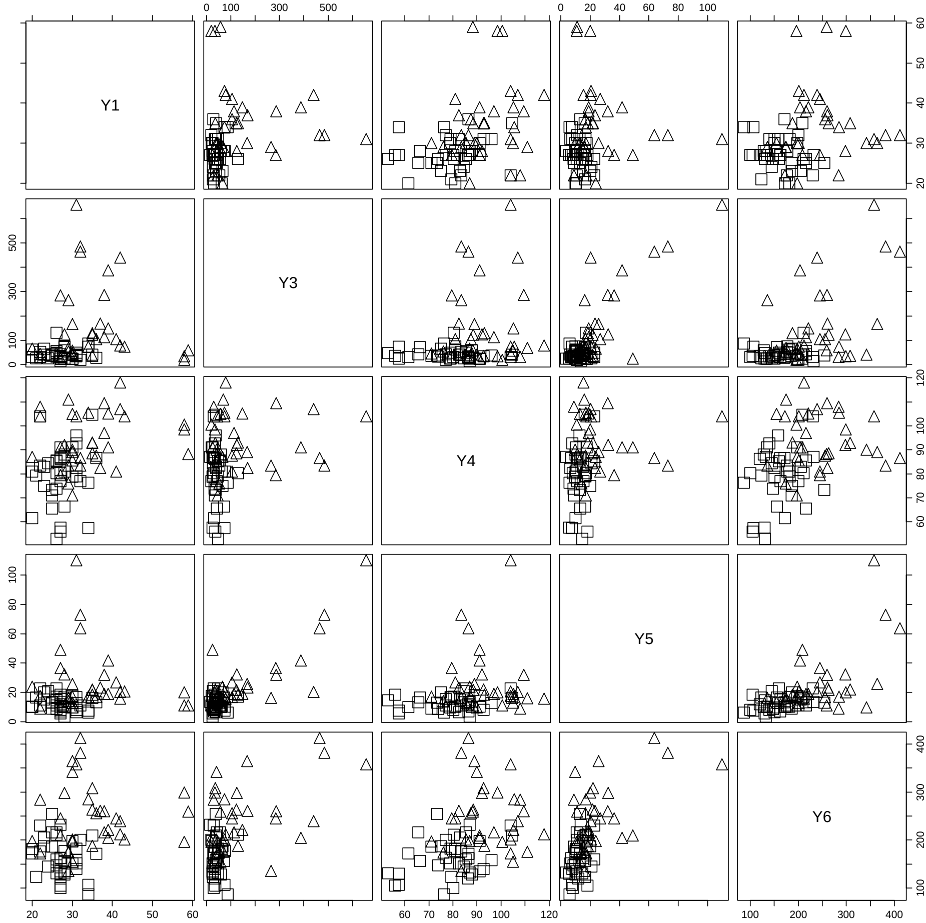


Figure 11. Small muscular dystrophy dataset. Scatterplot matrix of untransformed data. Triangles are carriers, squares are noncarriers.

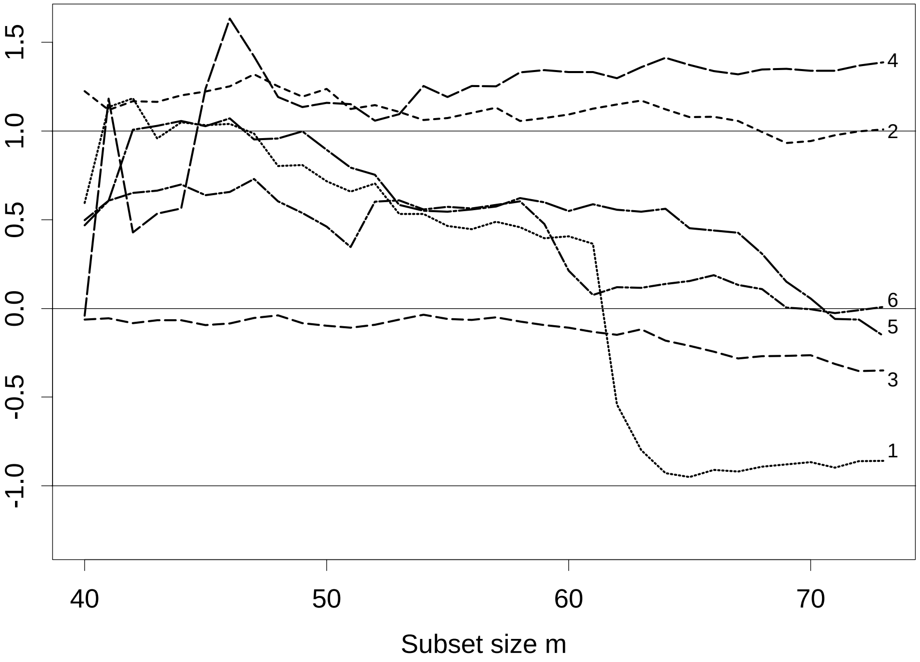


Figure 12. Small muscular dystrophy dataset. Estimates of transformation parameters from an unbalanced search on untransformed data. The numbers by the curves are the labels of the y.

In the analysis of these data we make use of the expertise and insights gained in analyzing the simulated data in Section 3. We start with the 73 observations used by Rencher (1995, p. 170), find a transformation, look for and identify outliers and consider discrimination. We then move to the complete dataset given by Andrews and Herzberg (1985, pp. 223–228) and see how well our model fits. We have thus the advantages of a confirmatory sample splitting, without having, ourselves, to make an arbitrary split of the data.

4.2 FINDING THE TRANSFORMATION

We first, as ever, look at a plot of the data, which is given in Figure 11. For clarity on the printed page we include only variables 1, 3, 4, 5, and 6. It certainly looks as if the data should benefit from transformation—some marginal distributions are skew, the variances in the two groups appear different, and there may be some outliers.

Step 1. As before we start with a forward search using untransformed data to obtain a preliminary idea of a good transformation. The resulting forward plot of parameter estimates, using an unbalanced search, is in Figure 12. For most of the plot the estimates of all λ's except λ_1 are reasonably stable, with some downward drift, but no sudden jumps. The exception is the estimate for λ_1 which decreases gradually to 0.5 when at $m = 62$ it drops abruptly to -1. Our previous experience suggests that a group of outliers may be entering at this point. To determine whether this is so, we use λ_1 = 0.5 in Step 2.

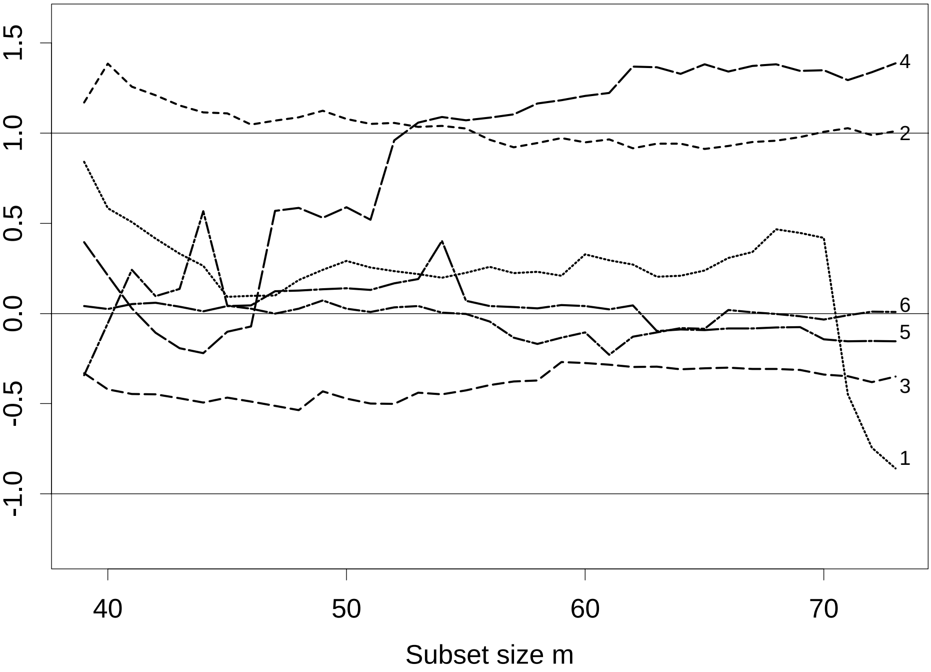


Figure 13. Small muscular dystrophy dataset. Estimates of transformation parameters from a search with $\lambda =$ (0.5,1,−0.5,1,0,0). The numbers by the curves are the labels of the y.

Remark. Figure 12 shows the effect of performing an unbalanced search with untransformed data. There are two groups with the smaller observations predominantly in one group, so that observations from that group tend to enter the search earlier. Whether or not we use a balanced search, the later observations to enter the search will be those that are more informative about the transformation. The curves of the estimates will then tend to drift away from one if a transformation is needed. When we search on the correct scale, if there is such, the two groups will have similar variances and observations from one group will not enter the search preferentially in the earlier stages.

In Figure 12 we show more of the search than is needed to determine the transformation, in order to exhibit the amount of movement that may be found in the initial stages of the search. At the beginning, one or two unmasked outliers may have been included, but are rapidly removed by the forward search. If it is felt that the results of the search are being influenced by the starting conditions, several searches can be run from a variety of starting points. We found here, as usual, that for all starts the last third of the plot was unchanged.

Step 2. We now run the forward search from the starting point suggested in Step 1, that is, with $\lambda = (0.5, 1, -0.5, 1,0,0)^T$. The resulting forward plot of estimates is in Figure 13. The change in $\hat{\lambda}_1$ is now right at the end of the search, starting with $m = 71$. The last three observations to enter—46, 67, and 68—have a large effect on the estimated transformation.

The effect of the last three observations on the evidence for the transformation is visible in Figure 14(a), which shows the manner in which the likelihood ratio statistic for

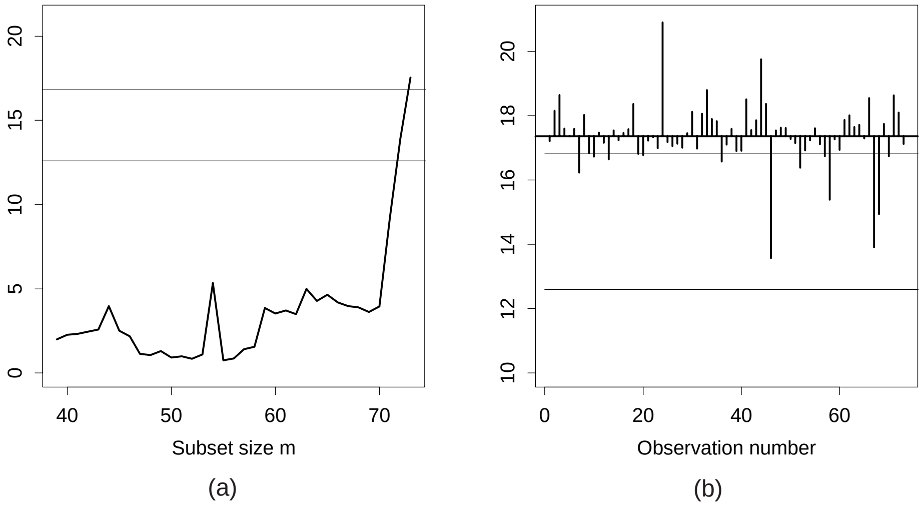


Figure 14. Small muscular dystrophy dataset. Results of a search with λ = (0.5,1,−0.5,1,0,0): (a) likelihood ratio test of the transformation during the forward search; (b) deletion likelihood ratio test, which fails to identify the importance of the outliers. The horizontal lines give the 95 and 99% points of χ^2_6.

this transformation jumps up at the end of the search, so that the transformation is rejected by the test. Our initial conclusion is that we have found the correct transformation and that there are three outliers.

The masked nature of these outliers is demonstrated in Figure 14(b), which shows the deletion version of the likelihood ratio test of Figure 14(a). Although individual deletion of our three outliers (46, 67, and 68) causes the three largest decreases in the statistic, the decrease is nothing like significant. One would be hard put to it to identify the importance of these three observations from this plot.

Step 3. We now confirm the suggested transformation. Figure 15 shows a signed square root likelihood ratio expansion for each variable around $\lambda_R = (0.5, 1, -0.5, 1, 0, 0)^T$. Since we are interested in the fine structure of the plot, even for wrong transformations, we used a balanced search.

The plot shows that only -0.5 is reasonable for y_3. For y_2 and y_4 $\lambda = 1$ seems to be best even if the data do not provide significant evidence against the square root transformation for y_4. The plot for y_6 confirms that the best transformation for this variable is the logarithm, as it is for y_5. Notice a small jump at the end of the curve for y_5 when $\lambda = -0.5$. This is caused by unit 24 (the smallest for y_5). It is only for y_1 that the three outliers we have identified are highly influential for the transformation. The panel for y_1 shows that the presence of units 46, 67, and 68 is incompatible with $\lambda = 0$ (at the 5% level), 0.5 and 1. The other two values appear to be compatible with all the data but the inclusion of the three observations earlier in the search causes breaks in the levels of the curves

A final comment on these plots, especially that for y_1, is that the three outliers are all in Group 2, the carriers of the disease. With the balanced search they cannot enter the subset consecutively, so that their inclusion is spread over the last five values of m.

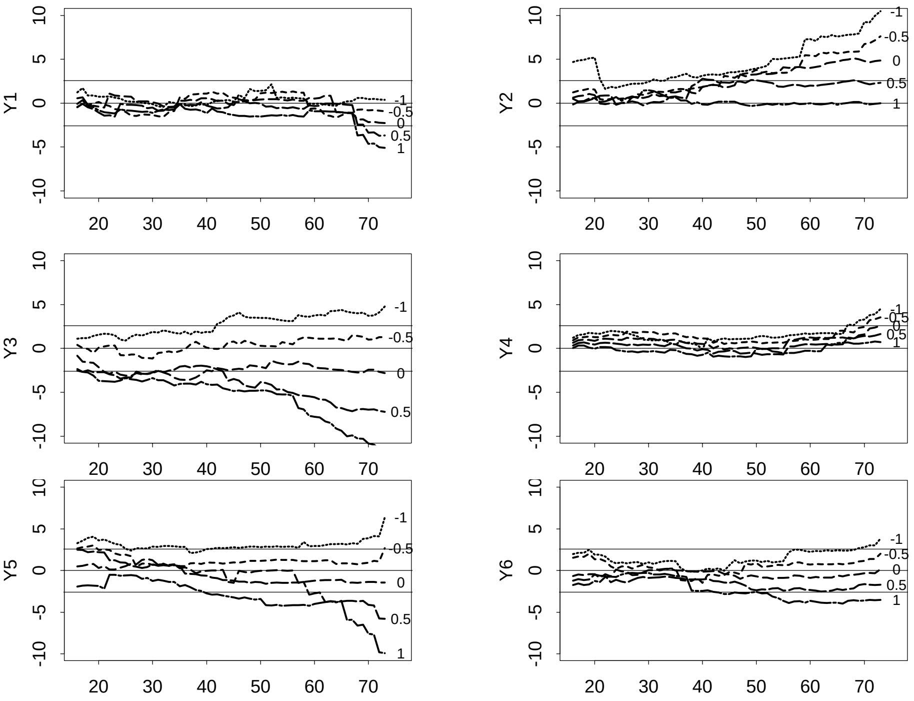


Figure 15. Small muscular dystrophy dataset: signed square root of the likelihood ratio test for transformation, confirming λ_R = (0.5,1,−0.5,1,0,0). The numbers by the curves are the values of λ.

A benefit of the forward search is that it can clearly determine the number of units compatible with a particular transformation without suffering from masking effects.

4.3 Outliers and Discriminant Analysis

Our procedure has suggested three outliers. It is important to refer these back to the original data. Their influence was on the transformation of variable 1, age. The three outliers are by far the oldest people present. For the other 70 observations the ages range from 29–42. The three outliers have ages at measurement of 58, 58, and 59. Once our attention has been drawn to them by the forward analysis, not only are they clearly evident in the scatterplot matrix of transformed data (Figure 16) but they are also evident at the top of the top row of plots in the scatterplot matrix of untransformed data (Figure 11). Finally, if we compare Figure 16 with Figure 11 we can clearly see how much more normal the data have become.

The effect of the transformation on the discriminant analysis is appreciable. For our comparisons we use the transformation $\lambda_R = (0.5, 1\,, -0.5, 1, 0, 0)^T$ for all 73 observations. Some results are in Table 5, both for all 73 observations and for the 70 observations when the outliers are deleted.

Failure to use the transformation results in an approximately 25% increase in misclassification, whether or not the outliers are excluded. Under similar conditions, use of the

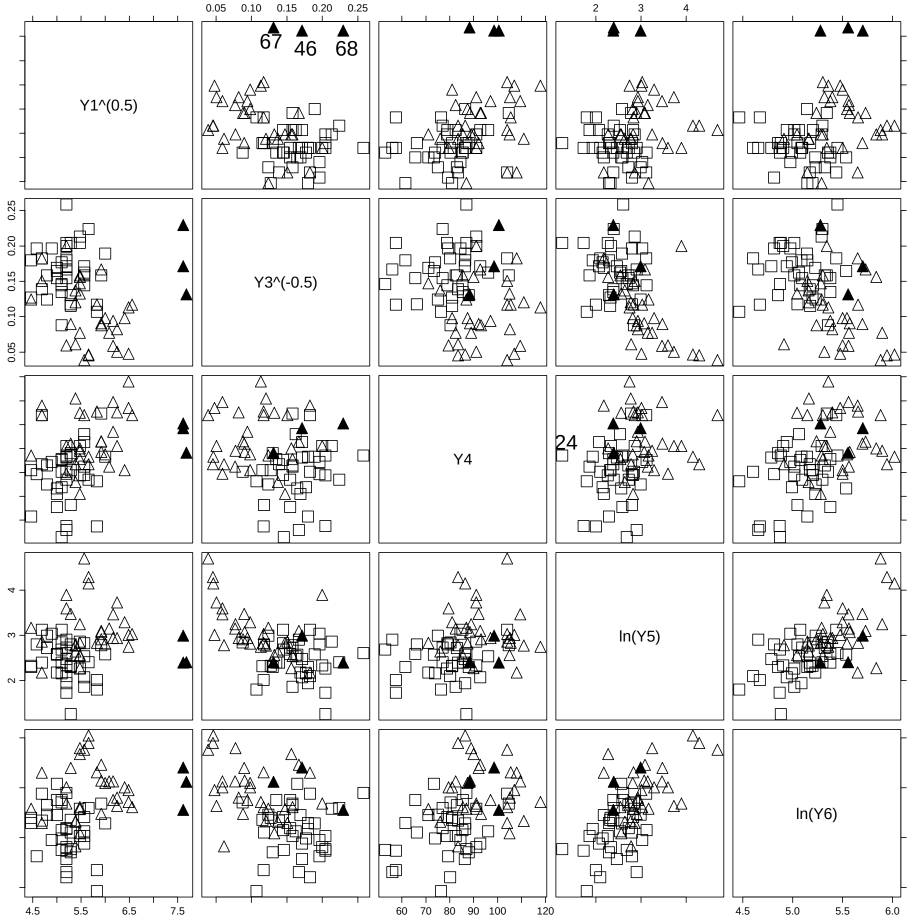


Figure 16. Small muscular dystrophy dataset: scatterplot matrix of transformed data showing revealing observations 46, 67, and 68. Squares are noncarriers, triangles are carriers, and filled triangles are outliers.

transformation gives almost 80% of the units a decreased probability of misclassification. These results on probabilities can be visualized by ordering the probabilities of misclassification for each group and then plotting the logits of the probabilities against order number. Figure 17 shows that the improvement is over almost all units.

4.4 MORE DATA

We now extend our analysis from the 73 units studied by Rencher to the 194 complete observations originally given by Andrews and Herzberg (1985). An interesting question is the stability of our conclusions in the extended set. If observations 46, 67, and 68 are treated as outliers, 0.5 is a good estimate of λ_1. If they are treated as part of the data, then -0.5 is a better value. These three individuals appeared as outliers because they are much older than the remaining 70. But the results of Table 5 show how little effect these observations have

Table 5. Dystrophy Data, Small Set (73 observations). Overall misclassification results for discrimination using transformed and untransformed observations: λ_R =(0.5, 1 , −0.5, 1, 0, 0)^T: $n = 70$, observations 46, 67, and 68 deleted

	$n = 73$	$n = 70$
Transformed data	0.153	0.159
Untransformed data	0.188	0.199
Percentage of units improved	79.5	77.1

on the discriminant analysis.

The data with 194 units are no longer balanced—there are 127 noncarriers but Group 2 contains only 67 carriers. With the new data we may expect that there may be some additional outliers. Also there are 14 new units with ages at least 43. Although these are all carriers, they can be expected to tell us something about the three outliers identified in the smaller dataset.

We do not display here the results of the three stages in finding a suitable multivariate transformation, but move straight to the confirmatory expansion in five values of λ. The conclusion is that the previous transformation holds except that now $\lambda_1 = -0.5$. The 14 newly included units for subjects at least 43 years old thus provide information to support the stronger transformation of y_1 indicated by the three units out of the initial 73. We therefore want to confirm the value $\lambda_R = (-0.5, 1\ , -0.5, 1, 0, 0)^T$. The signed square-root tests from the expansion around this value are plotted in Figure 18, in which we have used a balanced search as some of the 30 forward searches are for values far from λ_R. The order of inclusion of the units for the search with λ_R is given in Table 6.

The plots for variables 1, 2, and 3 are uneventful. The inclusion of observation 78 at the end of the search causes a jump in the value of the statistic for λ_4 as it does for that for

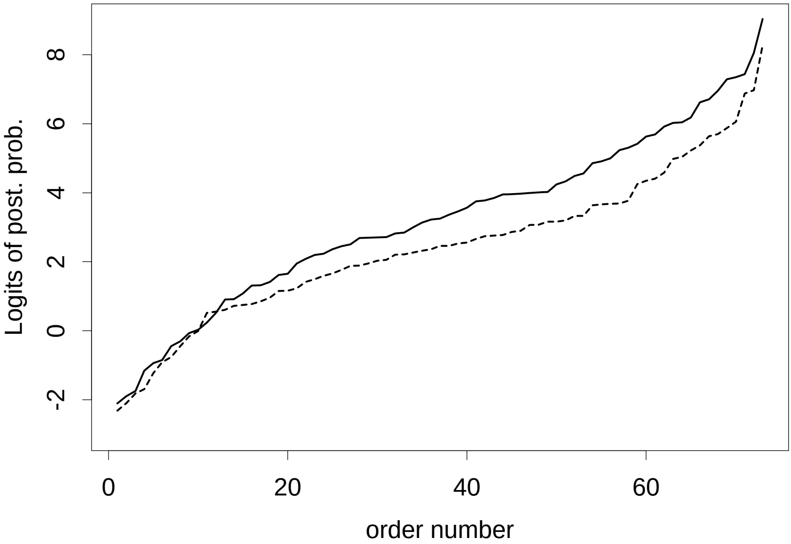


Figure 17. Small muscular dystrophy dataset: the upper curve is the logit of the posterior probability of correct classification after transformation, the lower curve that before.

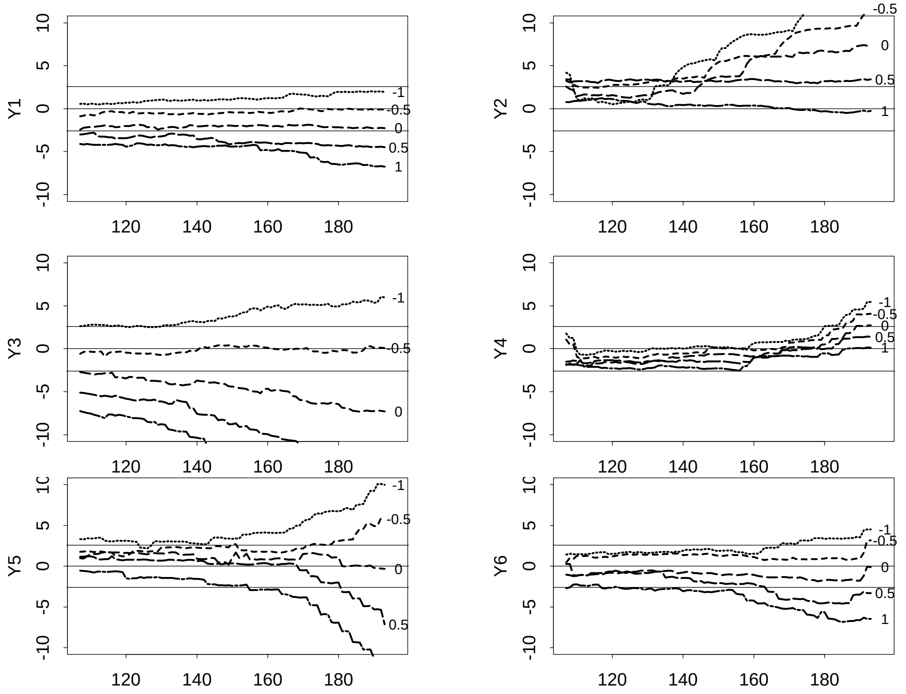


Figure 18. Large muscular dystrophy dataset: signed square root of the likelihood ratio test for transformation, confirming λ_R = (−0.5,1,−0.5,1,0,0). The numbers by the curves are the values of λ.

λ_6, which is also slightly influenced by observation 53. But there is no reason to question the suggested value of λ_R.

We now examine the data for any other outliers and the effect of transformation on their presence. If there are any further outliers, the plots in Figure 18 show that they will not have had an influence on the choice of transformation.

We start with the untransformed data. Because we are interested in the differences between the groups we use a balanced search so that we can monitor the Mahalanobis distances from the two groups as the search evolves. Figure 19(a) (minimum distance monitoring plot) shows that before transforming the groups seem to have a completely different structure in terms of Mahalanobis distances. The different structure of the two groups is also immediately evident from the zig-zag in Figure 19(b) (pooled determinant monitoring). The addition of a first and second unit from Group 1 does not have as big an effect on this statistic as adding a single unit from Group 2. It is clear from these two panels that, in the absence of transformation, one would be forced to consider quadratic discriminant analysis or nonparametric procedures.

Table 6. Full Data. Order of inclusion of observations for balanced forward search with $\lambda_R = (-0.5, 1, -0.5, 1, 0, 0)^T$

Step (m)	184	185	186	187	188	189	190	191	192	193	194
Unit Included	156	101	117	155	95	27	140	118	53	130	78
Group	2	1	1	2	1	1	2	1	1	2	1

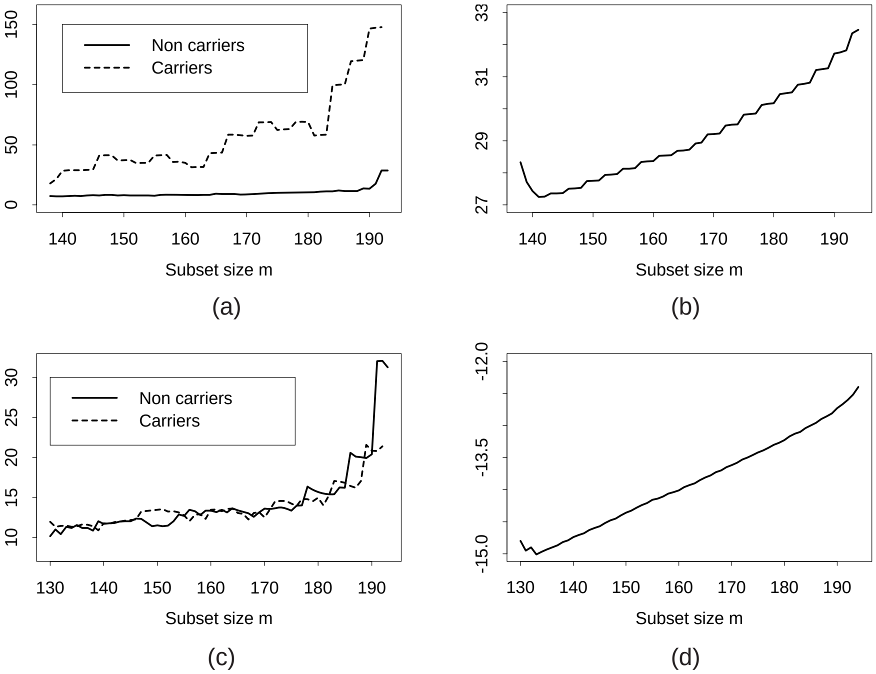


Figure 19. Large muscular dystrophy dataset: (a) Mahalanobis distances from a balanced search on untransformed data—the difference in variance in the two groups is evident; (b) determinant of the estimated common covariance matrix—the zig-zag pattern is caused by the constraint imposed by balance; (c) distances as for (a), but from an unbalanced search on transformed data—the outliers are now revealed; (d) covariance matrix as in (b), but now the data have been transformed and there is no longer an effect of different variances.

The other half of the plot is for a search using our transformation λ_R. The contrast with Figure 19(a) and Figure 19(b) is fascinating. Figure 19(d) shows the new pooled determinant monitoring. Now that the data have been transformed this curve has become a straight line. The initial decrease of this curve in the first steps of the forward search reflects the initial exchange of units in and out of the subset. As an aid to outlier detection Figure 19(c) shows the minimum distance monitoring plot now using an unbalanced search. The order of inclusion of the observations is given in Table 7. The plot shows that, for most of the search, the transformation has made the variances in the two groups comparable, but that there are five outliers. For Group 1 observations 118, 53, and 78 are shown to be outlying by the upward jump in the plot. The two outliers for Group 2 are 130 and 140. One interesting feature is that these outliers are for subjects with ages between 22 and 39. The inclusion of the additional 14 units with ages at least 43 has rendered the previous three outliers extreme

Table 7. Full Data. Order of inclusion of observations for unbalanced forward search with $\lambda_R = (-0.5, 1, -0.5, 1, 0, 0)^T$

Step (*m*)	184	185	186	187	188	189	190	191	192	193	194
Unit Included	117	27	95	156	155	146	118	53	140	130	78
Group	1	1	1	2	2	2	1	1	2	2	1

Table 8. Dystrophy Data, large set (194 observations). Results for discrimination using transformed and untransformed observations: $\lambda_R = (-0.5, 1, -0.5, 1, 0, 0)^T$: $n = 189$, observations 53, 78, 118, 130, and 140 deleted

Overall misclassification rates

	$n = 194$	$n = 189$
Transformed data	0.132	0.119
Untransformed data	0.170	0.166
Percentage of units improved	72.7	78.3

Tests of equality of covariance matrices

	$n = 194$	$n = 189$
Untransformed data	605.6	602.9
Transformed data	48.059	51.401

but not highly atypical. We now consider the effect of transformations and outlier detection on the properties of the discriminant analyses.

We compare all 194 observations and the 189 observations left after outlier detection, both on the original and on the transformed scales. Table 8, to be compared with Table 5, shows the overall misclassification probabilities for the four analyses. The general difference between the two tables is that, with more observations used in estimation, the overall rates have gone down. The highest average probability of misclassification now is 0.170 for the original data, dropping to 0.119 for the transformed data without the outliers, a 43% increase in misclassification from failure to use the transformation. The percentage of units for which the probability of misclassification decreases because of the transformation is around 75% and is slightly increased when the outliers are removed.

The table also gives the results of the test for equality of covariances. It is clear from the table, as it was from the figures, that the transformation has an enormous effect in improving the equality of the variances of the two groups, although, with the 99% point of χ^2_{21} being 38.93, equality has not quite been obtained.

The final table gives information on the discriminant function. We list the coefficients of the standardized canonical eigenvector in Table 9. Apart from y_1 and y_2, age and month, the other four variables are serum markers: y_3 and y_4 are inexpensive to measure, y_5 and y_6

Table 9. Comparison of the Coefficients of the Canonical Eigenvector Using Transformed and Untransformed Observations

Elements of standardized canonical eigenvector	*AGE* y_1	*M* y_2	*CK* y_3	*H* y_4	*LD* y_5	*PK* y_6
Untransformed data ($n = 194$)	0.593	0.001	0.160	0.319	0.304	0.488
Transformed data ($n = 194$)	0.490	−0.052	0.582	−0.387	−0.250	−0.276
Transformed data ($n = 189$)	0.542	−0.073	0.650	−0.310	−0.202	−0.211

are expensive to measure. For the untransformed data the inexpensive variables are second and fourth in importance among the markers whereas, after transformation, they are first and second. Removal of the outliers further increases the weighting on y_3 and y_4. There is thus an indication that use of the transformed data could lead to the development of a cheaper medical test.

5. CONCLUSIONS

This article applies a form of the forward search in which diagnostic quantities are calculated and plotted at every stage to the calculation of estimates and score tests for multivariate transformations. For this we have had to invent a new systematic method of diagnostic searching to find our transformation estimator λ_R. We have then applied the new methodology to discriminant analysis. The examples show the power of our method in the presence of multiple outliers. We note an affinity with the results of Box and Cox (1964) who found that a transformation in regression often not only improved the normality of the errors but led to a simpler additive systematic model. Likewise we have found that our multivariate transformation not only leads to a more symmetrical distribution of the data, it also leads to the conditions of near equality of covariance matrices under which linear discriminant analysis can be applied.

[Received March 2000. Revised October 2000.]

REFERENCES

Andrews, D. F., and Herzberg, A. M. (1985), *Data,* New York: Springer Verlag.

Atkinson, A. C. (1994), "Fast Very Robust Methods for the Detection of Multiple Outliers," *Journal of the American Statistical Association,* 89, 1329–1339.

——— (1995), "Multivariate Transformations, Regression Diagnostics and Seemingly Unrelated Regression," in *MODA 4—Advances in Model-Oriented Data Analysis,* eds. C. P. Kitsos and W. G. Müller, Heidelberg: Physica-Verlag, pp. 181–192.

Atkinson, A. C., and Cheng, T.-C. (2000), "Computing Least Trimmed Squares Regression With the Forward Search," *Statistics and Computing,* 3, 251–263.

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

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.

Campbell, N. A. (1980), "Shrunken Estimators in Discriminant and Canonical Variate Analysis," *Applied Statistics,* 29, 5–14.

——— (1982), "Robust Procedures in Multivariate Analysis ii: Robust Canonical Variate Analysis," *Applied Statistics,* 31, 1–8.

Critchley, F., and Vitiello. F. (1991), "The Influence of Observations on Misclassification Probability Estimates in Linear Discriminant Analysis," *Biometrika,* 78, 677–690.

Fung, W. K. (1992), "Some Diagnostic Measures in Discriminant Analysis," *Statistics & Probability Letters,* 13, 279–285.

——— (1995a), "Diagnostics in Linear Discriminant Analysis," *Journal of the American Statistical Association*, 90, 952–956.

——— (1995b), "Influence on Classification and Probability of Misclassification," *Sankhya*, 57, 377–384.

——— (1996), "The Influence of Observations for Local Log-Odds in Linear Discriminant Analysis," *Communications in Statistics, Theory, and Methods*, 25, 257–268.

——— (1998), "On the Equivalence of Two Diagnostic Measures in Discriminant Analysis," *Communications in Statistics, Theory and Methods*, 27, 1923–1935.

Hadi, A. S. (1992), "Identifying Multiple Outliers in Multivariate Data," *Journal of the Royal Statistical Society*, Series B, 54, 761–771.

Hawkins, D. M., and McLachlan, G. J. (1997), "High-Breakdown Linear Discriminant Analysis," *Journal of the American Statistical Association*, 92, 136–143.

He, X., and Fung, W. K. (2000), "High Breakdown Estimation for Multiple Populations With Applications to Discriminant Analysis," *Journal of Multivariate Analysis*, 72, 151–162.

McLachlan, G. M. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, New York: Wiley.

Rencher, A. C. (1995), *Methods of Multivariate Analysis*. New York: Wiley.

Riani, M., and Atkinson, A. C. (2000), "Robust Diagnostic Data Analysis: Transformations in Regression," *Technometrics*, 42, 384–394.

Riani, M., and Zani, S. (1997), "An Iterative Method for the Detection of Multivariate Outliers," *Metron*, 55, 101–117.

Rousseeuw, P. J., and van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, 41, 212–223.

Rousseeuw, P. J., and van Zomeren, B. C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, 633–639.

Zani, S., Riani, M., and Corbellini, A. (1998), "Robust Bivariate Boxplots and Multiple Outlier Detection," *Computational Statistics and Data Analysis*, 28, 257–270.