J. R. Statist. Soc. B (1996)
58, *No.* 1, *pp.* 155–176

Discriminant Analysis by Gaussian Mixtures

By TREVOR HASTIE† and ROBERT TIBSHIRANI

Stanford University, USA *University of Toronto, Canada*

[Received January 1994. Final revision December 1994]

SUMMARY

Fisher–Rao linear discriminant analysis (LDA) is a valuable tool for multigroup classification. LDA is equivalent to maximum likelihood classification assuming Gaussian distributions for each class. In this paper, we fit Gaussian mixtures to each class to facilitate effective classification in non-normal settings, especially when the classes are clustered. Low dimensional views are an important by-product of LDA — our new techniques inherit this feature. We can control the within-class spread of the subclass centres relative to the between-class spread. Our technique for fitting these models permits a natural blend with nonparametric versions of LDA.

Keywords: CLASSIFICATION; CLUSTERING; NONPARAMETRIC REGRESSION; PATTERN RECOGNITION; RADIAL BASIS FUNCTIONS

1. INTRODUCTION

In the generic classification or discrimination problem, the outcome of interest G falls into J unordered classes, which for convenience we denote by the set $\mathcal{J} = \{1, 2, 3, \ldots, J\}$. We wish to build a rule for predicting the class membership of an item based on p measurements of predictors or features $X \in R^p$. Our training sample consists of the class membership and predictors for N items. This is an important practical problem with applications in many fields. Traditional statistical methods for this problem include linear discriminant analysis (LDA) and multiple logistic regression, nearest neighbour methods and classification trees. Neural network classifiers have become a powerful alternative, with the ability to incorporate a very large number of features in an adaptive non-linear model. Ripley (1994) gives an informative review from a statistician's viewpoint.

LDA can be derived as the maximum likelihood method for normal populations with different means and common covariance matrix. It is natural therefore to generalize LDA by assuming that each observed class is a mixture of unobserved normally distributed subclasses. This approach is sometimes mentioned in the statistical literature (McLachlan (1992) and Cheng and Titterington (1994), for example) and more often in the pattern recognition literature (Taxt *et al.*, 1991) but does not seem to have generated much attention. Particular neural network models using Gaussian radial basis functions can be derived from a Gaussian mixture modelling formulation (Tresp *et al.*, 1994).

In this paper we develop the mixture approach to discrimination in several interesting and useful directions. We refer to this as mixture discriminant analysis (MDA).

†*Address for correspondence*: Department of Statistics, Sequoia Hall, Stanford University, Stanford, CA 94305-4065, USA.
E-mail: trevor@playfair.stanford.edu

0035–9246/96/58155

LDA is not only a classification procedure but also a data reduction tool. We can represent multiclass data in low dimensional projections or plots that highlight their class differences. Our MDA procedure has this feature as well—we can produce a hierarchy of co-ordinates in terms of their abilities to separate the classes (and subclasses). An interesting twist occurs here: for two-class data, LDA produces a (not so interesting) single co-ordinate, whereas MDA will produce up to one fewer than the number of subclasses in the mixture representation.

A technique known as *learning vector quantization* (LVQ) has received much attention in the pattern recognition literature (Kohonen, 1989); MDA can be viewed as a smooth version of LVQ. LVQ finds a *set* of cluster centres for each class; classification is performed by finding the closest centre and assigning the associated class. The on-line learning algorithms for LVQ are similar to the k-means algorithm, with biases built in to encourage good classification. Whereas LVQ generalizes clustering to classification problems, MDA generalizes mixture density estimation to classification problems.

An additional feature of our model is *subclass shrinkage*: we can regulate the within-class spread of the mixture centres relative to their between-class spread. This emphasizes our bias towards classification, much like LVQ biases clustering techniques towards classification. A small amount of shrinkage tends to smooth the decision boundaries and is a form of regularization. It also allows us to use many subclass centres per class, and to shrink them until the *effective number* of centres is a required smaller amount.

Other methods have been proposed to generalize LDA to allow non-linear decision boundaries. In Hastie *et al.* (1994), this is achieved by using adaptive, nonparametric regression methods. The link between nonparametric regression and discriminant analysis is provided by the *optimal scoring* approach, first suggested by Breiman and Ihaka (1984). Hastie *et al.* (1994) called this procedure *flexible discriminant analysis* (FDA). The generic version of FDA based on smoothing splines operates by expanding the predictors into a large (adaptively selected) basis set, and then performing a *penalized* discriminant analysis (PDA) in the enlarged space. Intuitively the penalization works by downweighting 'rough' directions relative to 'smooth' directions in this enlarged space, when computing Mahalanobis distances. PDA (Hastie *et al.*, 1995) is a closely related technique for classifying digitized analogue signals. PDA starts with a high dimensional feature set, such as pixels from a digitized image, or spectral values on a grid of frequencies. Again a PDA is used, but here to ensure spatial smoothness of the discriminant coefficients. Both of these techniques adapt naturally to MDA: in the (enlarged) feature space, we use a mixture of Gaussian distributions rather than a single distribution per class, and we still use a penalized metric when computing distances.

Fig. 1 illustrates some of these techniques with a simple example. Not surprisingly the MDA technique has the most satisfactory boundary (the problem is a set-up for MDA), and hence its boundary is nearly optimal. The LVQ boundary seems unnecessarily biased and wiggly. The FDA boundary is a reasonable approximation to the optimal boundary. Fig. 1(e) shows a rank 1 version of the MDA fit, which also does acceptably well. Fig. 1(f) shows the result of MDA combined with the non-linear transformations of FDA—there is no noticeable improvement.

To summarize then, our proposal for MDA has the following features:

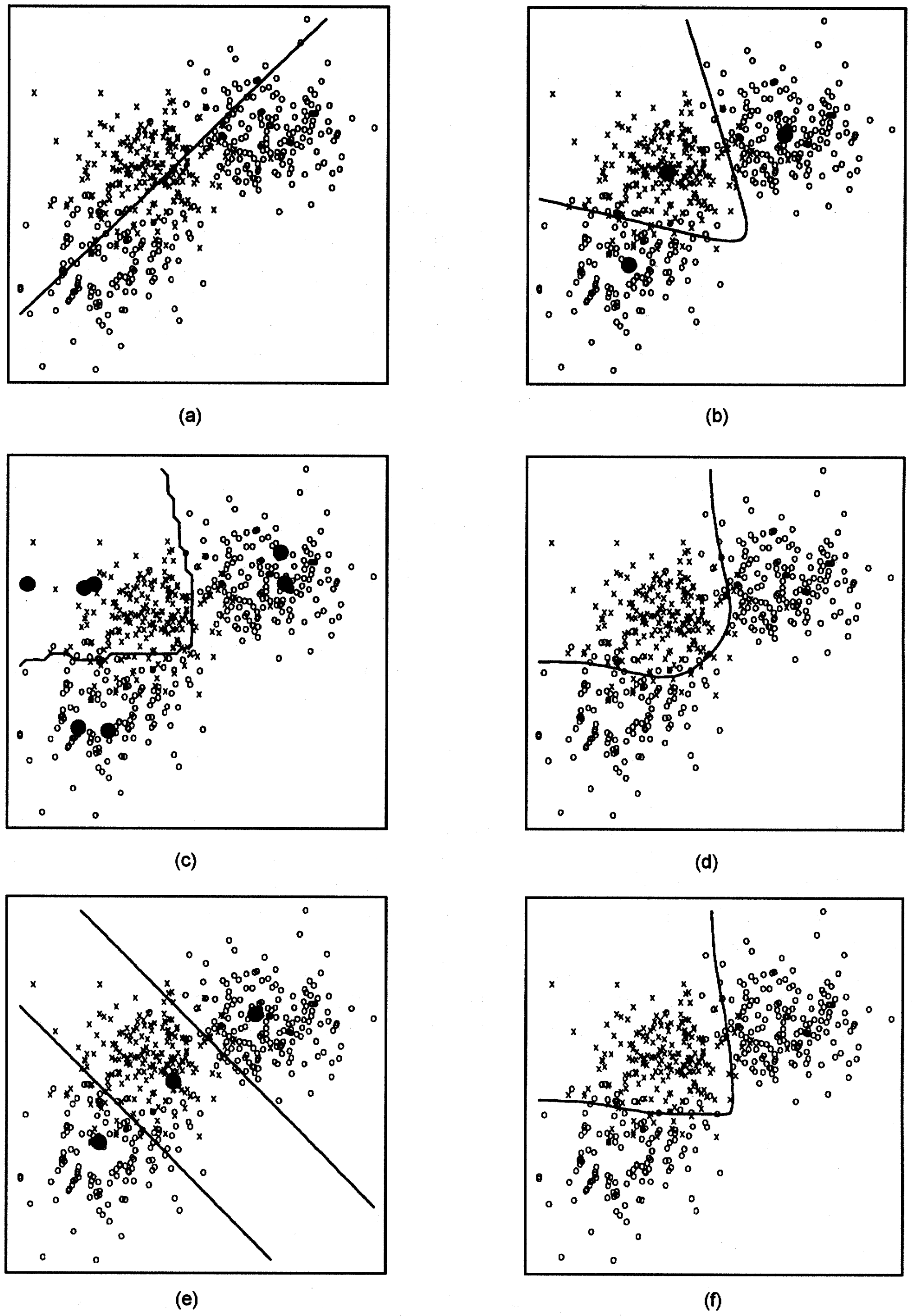

Fig. 1. Two-class problem in which one of the classes occurs in two separated circular clouds: (a) LDA; (b) MDA; (c) LVQ; (d) FDA; (e) MDA, rank 1 model; (f) MDA–FDA

(a) the classes are modelled as mixtures of Gaussian distributions, rather than a single Gaussian distribution as in LDA;
(b) optimal subspace identification is possible as in LDA, with added functionality;
(c) we can shrink the subclass centres, for example towards a common centre;
(d) the flexibility of FDA and PDA is easily and naturally accommodated.

The paper is organized as follows. In Section 2 we discuss the normal mixture model and an EM algorithm for its estimation. In Section 3 we discuss reduced rank versions of the procedures, which we demonstrate in Section 4 on the well-known waveform example of Breiman *et al.* (1984). In Section 5 we show how the MDA algorithm can be expressed as a repeated regression procedure using optimal scoring (Hastie *et al.*, 1994), which in turn allows useful nonparametric extensions. Section 6 further illustrates these techniques on a handwritten digit recognition problem. Section 7 discusses the centroid shrinking.

2. EXTENDING LINEAR DISCRIMINANT ANALYSIS BY NORMAL MIXTURES

The approach to classification taken here is to model the class densities of the predictors $P(X|G)$ by Gaussian mixture models. These are flipped around via Bayes theorem and class priors to give models for the class posterior probabilities $P(G|X)$—a basic ingredient for classification.

Suppose that we have training data $(x_i, g_i) \in R^p \times \mathcal{J}$, $i = 1, 2, \ldots, N$. We divide each class j into R_j artificial subclasses, denoted by c_{jr}, $r = 1, 2, \ldots, R_j$, and define

$$R = \sum_{j=1}^{J} R_j.$$

Our model assumes that each subclass has a multivariate normal distribution with its own mean vector μ_{jr} and common covariance matrix Σ. This is not the only possible mixture model; for example, we could allow each subclass to have a different covariance matrix or force the subclasses within a given class to have the same covariance matrix. The particular model chosen here is attractive because it keeps the total number of parameters under control and, as we shall see, it has the right structure to permit the other generalizations that we have in mind.

Let Π_j be the prior probability for class j, and within class j let π_{jr} be the mixing probability for the rth subclass,

$$\sum_{r=1}^{R_j} \pi_{jr} = 1.$$

Although often the Π_j are known or easily estimated from the training data, the π_{jr} are unknown model parameters. Let

$$D(x, \mu) = (x - \mu)^{\mathrm{T}} \Sigma^{-1} (x - \mu)$$

be the Mahalanobis distance between x and μ.

The mixture density for class j is

$$\begin{aligned} m_j(x) &= P(X=x|G=j) \\ &= |2\pi\Sigma|^{-1/2} \sum_{r=1}^{R_j} \pi_{jr} \exp\{-D(x, \mu_{jr})/2\}, \end{aligned} \tag{1}$$

and the conditional log-likelihood for the data is

$$l^{\text{mix}}(\mu_{jr}, \Sigma, \pi_{jr}) = \sum_{i=1}^{N} \log m_{g_i}(x_i). \tag{2}$$

The EM algorithm provides a convenient method for maximizing $l^{\text{mix}}(\theta)$. The EM steps are

$$\begin{aligned} \hat{p}(c_{jr}|x, j) &= \text{Prob}(x \in r\text{th subclass of class } j|x, j) \\ &= \frac{\pi_{jr} \exp\{-D(x, \mu_{jr})/2\}}{\sum_{k=1}^{R_j} \pi_{jk} \exp\{-D(x, \mu_{jk})/2\}}, \end{aligned} \tag{3}$$

$$\hat{\pi}_{jr} \propto \sum_{g_i=j} p(c_{jr}|x_i, j), \qquad \sum_{r=1}^{R_j} \hat{\pi}_{jr} = 1, \tag{4}$$

$$\hat{\mu}_{jr} = \frac{\sum_{g_i=j} x_i\, p(c_{jr}|x_i, j)}{\sum_{g_i=j} p(c_{jr}|x_i, j)}, \tag{5}$$

$$\hat{\Sigma} = \frac{1}{N} \sum_{j=1}^{J} \sum_{g_i=j} \sum_{r=1}^{R_j} p(c_{jr}|x_i, j)(x_i - \mu_{jr})(x_i - \mu_{jr})^{\mathrm{T}}. \tag{6}$$

The notation $\Sigma_{g_i=j}$ means summing over all observations belonging to the jth class.

In the above, expression (3) is the *estimation* step, whereas expressions (4)–(6) are the *maximization* step. These are a straightforward generalization of the EM algorithm for estimating normal mixtures (e.g. Titterington *et al.* (1985), pages 86–87). Expressions (5) and (6) have the same appearance as the maximum likelihood estimate for the complete normal discriminant problem, i.e. the situation where we observe the subclass membership. The only difference is that the subclass indicator is replaced by $\hat{p}(c_{jr}|x_i, j)$, the estimated probability that observation i falls in subclass c_{jr} given that it is observed to fall into class j. Since $\hat{p}(c_{jr}|x_i, j)$ is a function of $\hat{\mu}_{jr}$ and $\hat{\Sigma}$, equations (3)–(6) must be iterated.

The posterior class probabilities are (by Bayes theorem)

$$P(G=j|X=x) \sim \Pi_j\, \text{Prob}(x|j) \sim \Pi_j \sum_{r=1}^{R_j} \pi_{jr} \exp\{-D(x, \mu_{jr})/2\} \tag{7}$$

normalized so that

$$\sum_{j=1}^{J} P(G=j|X=x)=1.$$

The classification rule chooses j to maximize $P(j|x)$. Notice that this does not have the same form as a linear discriminant rule for the R subclasses and in particular is likely to be non-linear.

2.1. *Cluster Sizes and Starting Values*

The EM iteration above requires a choice of cluster sizes R_j, and starting values for the means μ_{jr}, the covariance matrix Σ and the cluster probabilities $p(c_{jr}|x, j)$. We currently use two different strategies.

(a) *K-means*—we choose a fixed number of clusters (say r) for each class and then use a k-means clustering algorithm to estimate a set of r subclass centroids $\hat{\mu}_{jr}$ for each class. Then, for all observations in class j, $\hat{p}(c_{jr}|x_i, j)$ is set to 1 if $\hat{\mu}_{jr}$ is the closest centroid to x_i, and to 0 otherwise.

(b) *LVQ*—we run the LVQ algorithm on the training data and let it select the R_j and $\hat{\mu}_{jr}$ (we supply R). As in the previous case we use the output to produce $\hat{p}(c_{jr}|x_i, j)$.

In either case these within-class weights determine a new set of $\hat{\mu}_{jr}$s from equation (5) and $\hat{\Sigma}$ from equation (6), and the iterations go from there.

Both k-means and LVQ require starting centres, which are typically randomly selected x_i. It is known (Pal *et al.* (1993), for example) that both these procedures suffer from variability due to random starting values. Our strategy is to try a number (10) of different starts, and to choose the best. We use either a likelihood-based criterion or, more cheaply, training sample misclassification to guide the choice. Our experience has been that MDA *always* outperforms the starting procedure (LVQ or k-means) with respect to classification of the training data.

We have no automatic way for selecting the R_j or, in the case of LVQ, the total R. We treat these as meta-parameters and can use a validation set to choose good values.

3. REDUCED RANK DISCRIMINATION

In LDA with J classes, we can choose a subspace of rank $r < J$ that maximally separates the class centroids. This is mainly useful for descriptive purposes but is also a form of regularization and often leads to improved classification performance. This standard Fisher–Rao decomposition (Mardia *et al.* (1979), for example) is derived by successively maximizing the ratio of between- to within-group variance of linear combinations of the variables: $v^T Bv/v^T Wv$, where B is the between-class covariance (of the class centroids) and W the pooled within-class covariance. In this section we show that reduced rank LDA can be viewed as a restricted Gaussian maximum likelihood solution, and we then extend this concept to the mixture model.

Proposition 1. Consider maximizing the Gaussian log-likelihood

$$2l(\mu_j, \Sigma, d) = -\sum_{j=1}^{J} \sum_{g_i=j} (x_i - \mu_j)^{\mathrm{T}} \Sigma^{-1} (x_i - \mu_j) - N \log|\Sigma| \tag{8}$$

subject to the constraints rank $\{\mu_j\}_{j=1}^{J} = K \leqslant \min(J-1, p)$. The solution is

$$\hat{\mu}_j = WVV^{\mathrm{T}}(\bar{x}_j - \bar{x}) + \bar{x}, \tag{9}$$

$$\begin{aligned} \hat{\Sigma} &= W + \sum_{j=1}^{J} \frac{N_j}{N} (\bar{x}_j - \hat{\mu}_j)(\bar{x}_j - \hat{\mu}_j)^{\mathrm{T}} \\ &= W + WV_{\perp}V_{\perp}^{\mathrm{T}}BV_{\perp}V_{\perp}^{\mathrm{T}}W \end{aligned} \tag{10}$$

where V is a matrix consisting of the leading K eigenvectors of $W^{-1}B$. This solution effectively coincides with the reduced rank LDA solution:

(a) $\hat{\mu}_j$ is the projection of the jth sample mean onto the discriminant subspace of rank K;
(b) classification based on $(x - \hat{\mu}_j)^{\mathrm{T}}\hat{\Sigma}^{-1}(x - \hat{\mu}_j)$ is equivalent to the reduced linear discriminant rule of rank K.

Much of this result was proved by Campbell (1984), although he appears to have overlooked the last claim—we outline a simple proof in Appendix A.

Now consider a reduced rank version of the mixture model. We take the Gaussian mixture log-likelihood l^{mix} (equation (2)) and maximize it subject to the rank K constraint on the subclass means: rank$\{\mu_{rj}\} = K$.

How might we achieve this maximization? Once again we can use the EM algorithm. Steps (3) and (4) remain the same and are conditional on the current (reduced rank) versions of the centres and the corresponding pooled covariance estimate. Steps (5) and (6) can be viewed as weighted mean and pooled covariance maximum likelihood estimates in a *weighted* and *augmented* R-class problem. We augment the data by replicating the N_j observations in class j R_j times, with the lth such replication having observation weights $p(c_{jl}|x_i, j)$. This is done for each of the J classes, resulting in an augmented and weighted training set of

$$\sum_{j=1}^{J} N_j R_j$$

observations. Note that the sum of all the weights is N. We now impose the rank restriction. By analogy with the early part of this section, this can be achieved by a weighted version of LDA. We postpone the details of how this final step is carried out until Section 5. Two details are worth noting immediately.

(a) Step (3) of the EM algorithm requires only *differences* in Mahalanobis distances from the estimated centres. This is convenient, since weighted LDA will only supply distances in the K-dimensional subspace of span$\{\mu_{jr}\}$.
(b) The solution to this problem *cannot* be obtained by a simple reduction of the full rank mixture solution. We need to perform a reduced rank weighted LDA at each of the iterations of the EM algorithm. The reason is that the weights

computed in step (3) should depend on the reduced rank rather than the full rank solution.

We collect these results in the following.

Proposition 2. Consider maximization of the constrained mixture density log-likelihood

$$l^{\text{mix}}(\mu_{jr}, \Sigma, \pi_{jr}) \propto \sum_{j=1}^{J} \sum_{g_i=j} \log\left[\sum_{r=1}^{R_j} \pi_{jr} \exp\left\{-\frac{D(x, \mu_{jr})}{2}\right\}\right] - \frac{N}{2} \log|\Sigma| \qquad (11)$$

subject to $\text{rank}\{\mu_{jr}\} = K$. This is achieved by an EM algorithm analogous to that defined in equations (3)–(6), with steps (5)–(6) replaced by the equations for maximizing the weighted and augmented log-likelihood:

$$2l^{\text{weight}}(\mu_{jr}, \Sigma) \propto -\sum_{j=1}^{J} \sum_{g_i=j} \sum_{r=1}^{R_j} p(c_{jr}|x_i, j)\, D(x, \mu_{jr}) - N \log|\Sigma|$$

subject to $\text{rank}\{\mu_{jr}\} = K$. This maximization can be achieved via a similarly augmented and weighted rank K LDA.

Fig. 1(e) shows the effect of rank reduction on this simple example. Practically speaking, the approximate reduced rank solution—obtained by a weighted rank reduction of the full mixture solution—is more attractive than the exact reduced rank solution described above. This is because the latter requires us to fit the model iteratively for each dimension K of interest. We have found in a simulation experiment (not shown here) that the two approaches tend to agree quite well.

4. EXAMPLE: WAVEFORM DATA

We now illustrate some of these ideas on a popular simulated example, taken from Breiman *et al.* (1984), pages 49–55, and used in Hastie *et al.* (1994) and elsewhere. It is a three-class problem with 21 variables and is considered to be a difficult pattern recognition problem. The predictors are defined by

$$\left.\begin{aligned} x_i &= u\,h_1(i) + (1-u)\,h_2(i) + \epsilon_i && \text{(class 1)},\\ x_i &= u\,h_1(i) + (1-u)\,h_3(i) + \epsilon_i && \text{(class 2)},\\ x_i &= u\,h_2(i) + (1-u)\,h_3(i) + \epsilon_i && \text{(class 3)}, \end{aligned}\right\} \qquad (12)$$

where $i = 1, 2, \ldots, 21$, u is uniform on (0, 1), ϵ_i are standard normal variates and the h_i are the shifted triangular waveforms: $h_1(i) = \max(6 - |i - 11|, 0)$, $h_2(i) = h_1(i - 4)$ and $h_3(i) = h_1(i + 4)$.

Table 1 extends a simulation of Hastie *et al.* (1994) to include the methods discussed here. Each training sample has 300 observations, and equal priors were used, so there are roughly 100 observations in each class. We used test samples of size 500. The two MDA models are described in the footnote to Table 1. More details on the penalization are given in Section 7.

TABLE 1
Results for the waveform data†

Technique‡	*Error rates*	
	Training	*Test*
LDA	0.121 (0.006)	0.191 (0.006)
QDA	0.039 (0.004)	0.205 (0.006)
CART	0.072 (0.003)	0.289 (0.004)
FDA–MARS (degree = 1)	0.100 (0.006)	0.191 (0.006)
FDA–MARS (degree = 2)	0.068 (0.004)	0.215 (0.002)
MDA (3 subclasses)	0.087 (0.005)	0.169 (0.006)
MDA (3 subclasses, penalized 4 degrees of freedom)	0.137 (0.006)	0.157 (0.005)
PDA (penalized 4 degrees of freedom)	0.150 (0.005)	0.171 (0.005)

†The values are averages over 10 simulations, with the standard error of the average in parentheses. The five entries above the line are taken from Hastie *et al.* (1994). The first model below the line is MDA with three subclasses per class. The model in the next line is the same, except that the discriminant coefficients are penalized via a roughness penalty to effectively 4 degrees of freedom. The third is the corresponding penalized LDA or PDA model.
‡CART, classification and regression trees; MARS, multivariate adaptive regression splines.

Fig. 2 shows the leading canonical variates for the penalized MDA model, evaluated on the test data. As we might have guessed, the classes appear to lie on the edges of a triangle. This is because the $h_j(i)$ are represented by three points in 21-space, thereby forming vertices of a triangle, and each class is represented as a convex combination of a pair of vertices, and hence lie on an edge. Also it is clear visually

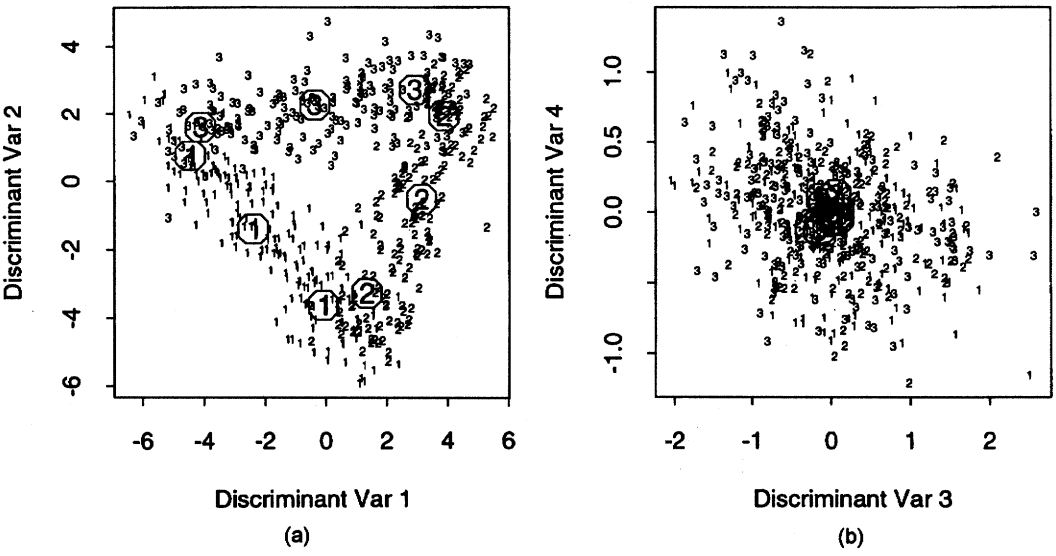

Fig. 2. Some two dimensional views of the MDA model fitted to a sample of the waveform model: the points are independent test data, (a) projected onto the leading two canonical co-ordinates and (b) projected onto the third and fourth co-ordinates (the subclass centres are indicated; three subclasses, penalized, 4 degrees of freedom)

that all the information lies in the first two dimensions; the percentage variance explained by the first two co-ordinates is 99.8%, and we would lose nothing by truncating the solution thus. The Bayes risk for this problem is about 0.14 (Breiman *et al.*, 1984); MDA comes close to the optimal rate, which is not surprising since the structure of the MDA model is similar to the generating model.

5. MIXTURE DISCRIMINANT ANALYSIS BY OPTIMAL SCORING

We can use 'optimal scoring'—multiple linear regression followed by an eigenanalysis—to fit both the LDA and the MDA models, as we show in this section. This has significant computational advantages and facilitates some useful generalizations of both techniques.

Consider first the LDA model. If $Y_{N\times J}$ is an *indicator response matrix* representing the classes, then we can regress Y against the predictor matrix X to derive fitted values $\hat{Y}=HY$. This is followed by a suitably normalized eigendecomposition of $Y^{\mathrm{T}}HY=Y^{\mathrm{T}}\hat{Y}$, which contains all the ingredients for an LDA of any rank (Breiman and Ihaka, 1984; Hastie *et al.*, 1994).

This simple procedure carries over to the M-step of the MDA algorithm. Instead of using a response indicator matrix Y, we use a *blurred* response matrix $Z_{N\times R}$, whose rows consist of the current subclass probabilities for each observation. At each M-step, this Z is used in a multiple linear regression followed by an eigendecomposition, just as in the LDA case above.

Hastie *et al.* (1994) called this procedure optimal scoring, which we now describe in more detail for MDA.

5.1. *Mixture Discriminant Analysis by Optimal Scoring*

5.1.1. *Initialize*

Start with a set of R_j subclasses c_{jr} for each class, and associated subclass probabilities $\hat{p}(c_{jr}|x_i,\, j)$. These can be derived, for example, from an LVQ or K-means preprocessing of the data. Let $R=\Sigma\, R_j$.

5.1.2. *Iterate*

Step 1—compute the blurred response: define the blurred response matrix $Z_{N\times R}$ as follows. If $g_i=j$, then fill the jth block of R_j entries in the ith row with the values $\hat{p}(c_{jr}|x_i,\, j)$, $r=1, \ldots, R_j$, and the rest with 0s. Z is the mixture analogue of an indicator–response matrix, except that observations can belong to several (sub-) classes with associated probabilities.

Step 2—multivariate linear regression: fit a multiresponse, linear regression of Z on X. Let $\hat{Z}$ be the fitted values and $\eta(x)$ be the vector of fitted regression functions.

Step 3—optimal scores: let Θ be the largest K non-trivial eigenvectors of $Z^{\mathrm{T}}\hat{Z}$, with normalization $\Theta^{\mathrm{T}}D_p\Theta=I_K$. Here D_p is a diagonal $R\times R$ matrix of weights, with rth entry the sum of the elements of the rth column of Z (the total weight for subclass r).

Step 4—update the fitted model from step 2 using the optimal scores: $\eta(x)\leftarrow\Theta^{\mathrm{T}}\,\eta(x)$.

Step 5—update $\hat{p}(c_{jr}|x_i,\, j)$ and $\hat{\pi}_{jr}$ using formulae (3) and (4).

To show that this MDA algorithm corresponds to the EM algorithm in proposition 2, we need to show that steps 1–4 fit the augmented and weighted rank K discriminant analysis. For brevity we omit the proof, which is given in Hastie and Tibshirani (1993) (see the file transfer information in Section 8.1).

We do not actually obtain estimates of the means and covariance from this MDA algorithm. The K-dimensional fit $\eta(x)$ produced by the algorithm is the co-ordinate of the discriminant projection of x onto the subspace spanned by the reduced rank subclass centroids (up to known scale factors). Relative Euclidean distances in this subspace are Mahalanobis distances in the implicitly estimated covariance. The projected means themselves are scaled rows of Θ. Since relative distances are all that are needed in equations (3) and (7), we have all the ingredients that are needed for fitting the model, and for classification. Details of these equivalences are given in Hastie *et al.* (1994, 1995).

One advantage of this approach is immediately clear: the M-step in proposition 2 involved a weighted LDA of $\Sigma_j N_j R_j$ observations, whereas here we only have N observations!

There is an even more compelling reason for using optimal scoring to fit LDA and MDA models. By expressing the problem as a multiple regression of a response matrix Y or Z, we can immediately generalize it by using in step 2 regression methods that are more exotic than linear regression. Two classes of such generalizations emerge.

(a) *FDA*: here multivariate adaptive non-linear regression replaces the linear regression, and the linear map $\eta(X)$ becomes a non-linear map. Hastie *et al.* (1994) describe the class of non-linear regression methods that can be used, including adaptive additive models, the multivariate adaptive regression spline (MARS) model of Friedman (1991), projection pursuit regression and neural networks. They called the resulting procedure ‘FDA’. These procedures amount to expanding the original predictors x in an adaptive way into a set of transformed predictors $h(x)$, and then performing LDA or MDA in the new space. In some cases the nonparametric regressions consist of a basis expansion followed by a penalized regression (e.g. additive splines). In this case the transformed space $h(x)$ is large, and the penalized regression translates into a penalized version of LDA or MDA.

(b) *PDA*: here we fit a more restrictive linear map via penalized least squares (Hastie *et al.*, 1995). The idea is that the feature vectors x_i arise as digitized analogue signals and are spatially correlated (pixels in an image, log-spectra at different frequencies). The penalization amounts to regularizing the large, nearly singular, covariance matrix $\hat{\Sigma}$ by adding a penalty $\hat{\Sigma} + \Omega$, which in turn ensures that the coefficients v_k are spatially smooth (and hence borrow strength from neighbouring values).

Our experience in many examples and simulations (Hastie and Tibshirani, 1993) in using this FDA extension of MDA suggests the following.

(a) Both FDA and standard MDA are capable of producing non-linear decision boundaries. Adding the non-linear capabilities of FDA to MDA puts these components in competition with each other and typically does not yield a much improved decision boundary. Fig. 1 supports this claim. One can imagine that in difficult situations the two components can support each other:

(i) if too few centres are provided, non-linearities can compensate;

(ii) extreme non-linearities, which are costly in terms of fitted degrees of freedom, can be avoided by using mixtures.

(b) The adaptive aspect of FDA plays the role of a variable selector, and it is convenient to have a regression procedure playing this role.

The PDA extension to MDA is justified in exactly the same way as for PDA itself. Two of our examples support penalization.

(a) The waveform predictors in Section 4 represent noisy functions sampled at 21 evenly spaced abscissae. The discriminant coefficients were penalized by using a second-derivative roughness penalty, reducing the effective number of parameters from 21 down to four, and gave on average an 8% improvement in the classification performance. Fig. 3 shows the fitted penalized discriminant functions $(\hat{\Sigma}+\Omega)^{-1}\hat{\mu}_{jr}$ for each subclass—and compares them with the unpenalized versions. We see that penalization plays two roles.

  (i) Variance reduction: the unpenalized MDA discriminant functions appear unnecessarily wiggly, for which we pay a price in generalization errors.

  (ii) Interpretation: by superimposing the known generating functions, we see that the discriminant functions are appropriately placed for picking out the peaks and valley of the mixture distribution for each class. It would be much more difficult and less convincing to attempt the same interpretations by using the unpenalized discriminant functions.

(b) The handwritten digit example in the next section is natural for both mixture models and for penalization:

  (i) there are several characteristic ways for writing each digit, calling for a mixture model;

  (ii) pixel grey scale values have strong spatial correlation, calling for regularization for the same reasons as given in the waveform example.

6. EXAMPLE: HANDWRITTEN DIGIT RECOGNITION

Hastie *et al.* (1995) illustrated PDA on a handwritten digit recognition task. Here we focus on a difficult subtask, that of distinguishing handwritten 3s, 5s and 8s. We use the same training data as Le Cun *et al.* (1990), who normalized binary images for size and orientation, resulting in 8-bit, 16 × 16 grey scale images. We have 658 3s, 556 5s and 542 8s in our training set and roughly a quarter the number of test examples. Fig. 4 shows some examples from each, and Table 2 shows the classification results.

Although we tried some preliminary experiments on the larger 10-class problem (roughly 8000 observations, 256 predictors and 40 subclasses), the computer time in our current S implementation encouraged us to focus on this smaller subset of the data.

The inputs are highly correlated because of their spatial arrangement, and some kind of smoothing or filtering always helps. We used an M-dimensional orthonormal basis of smooth functions on the 16 × 16 spatial domain of the image. These were derived from the roughness penalty matrix of a two-dimensional, third-degree thin

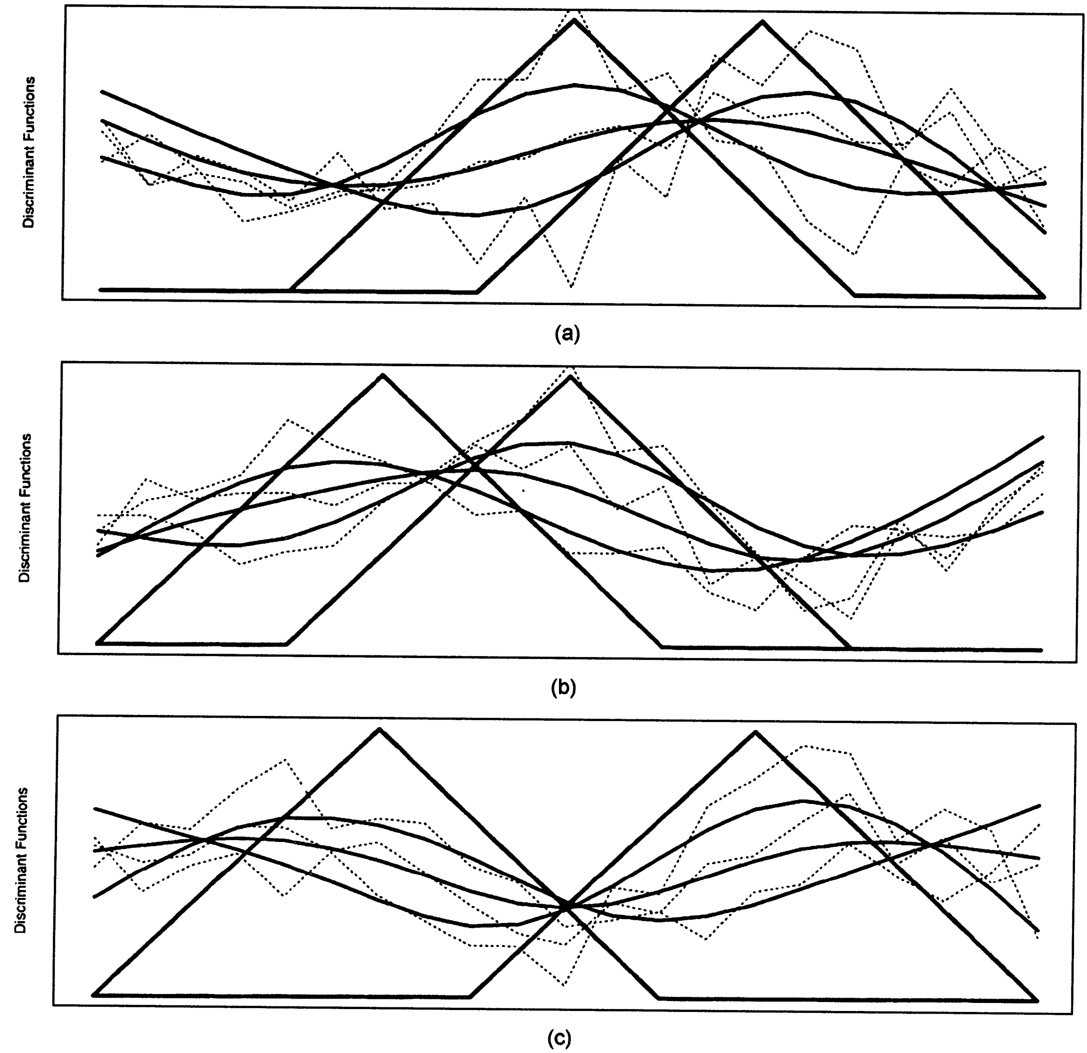

Fig. 3. Discriminant functions for the subclasses in the waveform example: subclasses for (a) class 1, (b) class 2 and (c) class 3 (the smooth functions are $\hat{\Sigma} + \Omega)^{-1}\hat{\mu}_{jr}$, produced by the MDA algorithm using a penalized regression method; the wiggly functions are the unpenalized versions $\hat{\Sigma}^{-1}\hat{\mu}_{jr}$; superimposed are the generating waveforms, shifted and scaled to fit into the plots)

plate smoothing spline; they are the M trailing eigenvectors, which are the smoothest and penalized the least. Given this $256 \times M$ basis matrix P, an input pixel vector x is replaced by $x^* = P^{\mathrm{T}}x$. The performance of LDA improved by nearly 20% because of the filtering alone, with $M = 64$. MDA with four subclasses per digit gave no better performance, but further regularization helped:

(a) reducing M from 64 to 49 yielded a 12% improvement;
(b) with $M = 64$ and shrinking to a total of eight effective centres yielded a 20% improvement.

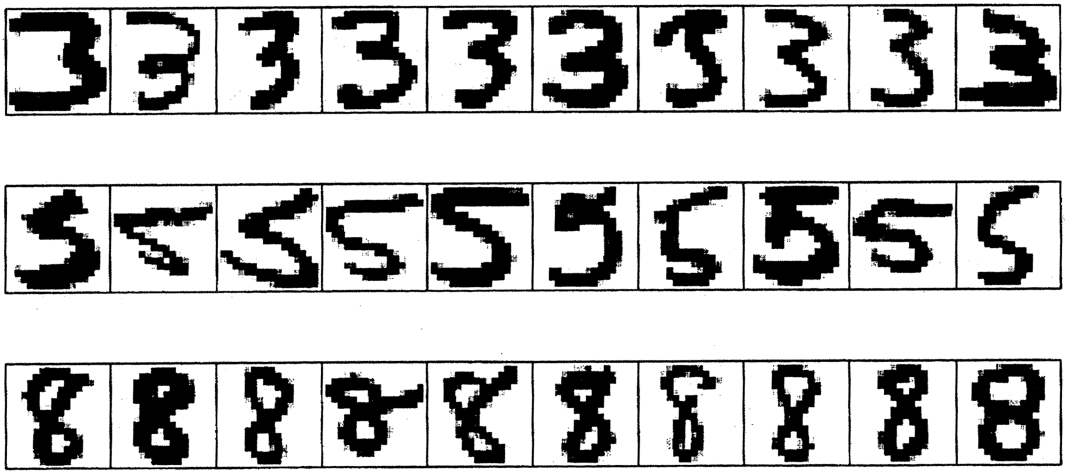

Fig. 4. Random selection of digitized handwritten 3s, 5s and 8s: each image is an 8-bit grey scale version of the original binary image, size and orientation normalized to 16 × 16 pixels

TABLE 2
Digit classification results — 3s, 5s and 8s†

Technique	*Error rates (%)*	
	Training	*Test*
LDA	1.6	8.7
LDA (filtered — 64 degrees of freedom)	3.1	7.1
LDA (filtered — 49 degrees of freedom)	3.6	7.7
MDA (filtered — 64 degrees of freedom, 4 subclasses)	2.7	7.1
MDA (filtered — 49 degrees of freedom, 4 subclasses)	2.3	6.3
MDA (filtered — 64 degrees of freedom, 4 subclasses, shrunk)	2.7	5.8

†The filtered models correspond to a hierarchical basis of smooth two-dimensional functions, derived from the thin plate smoothing spline penalty functional. The shrunken MDA model shrunk the 12 centroids to an effective total of eight.

We tried other shrinking strategies, such as using seven subclasses per class and shrinking down to eight as above; the test errors were 6.6%. Shrinking down to 18 gave 6.4% errors. An empirical conclusion we draw is that a small amount of shrinking helps by smoothing the decision boundaries; a large amount of shrinking can hurt by losing the flexibility of having different centres.

Fig. 5 displays the fitted subclass centroids as images: some spatial differences are apparent within a class. Fig. 6 illustrates the effect of shrinkage on the eigenvectors, discussed further in Section 7.

7. SHRINKING AND PENALIZATION

In this section we make precise our notion of shrinking the subclass centres, as well as the interpretation of penalized optimal scoring in the context of the present model.

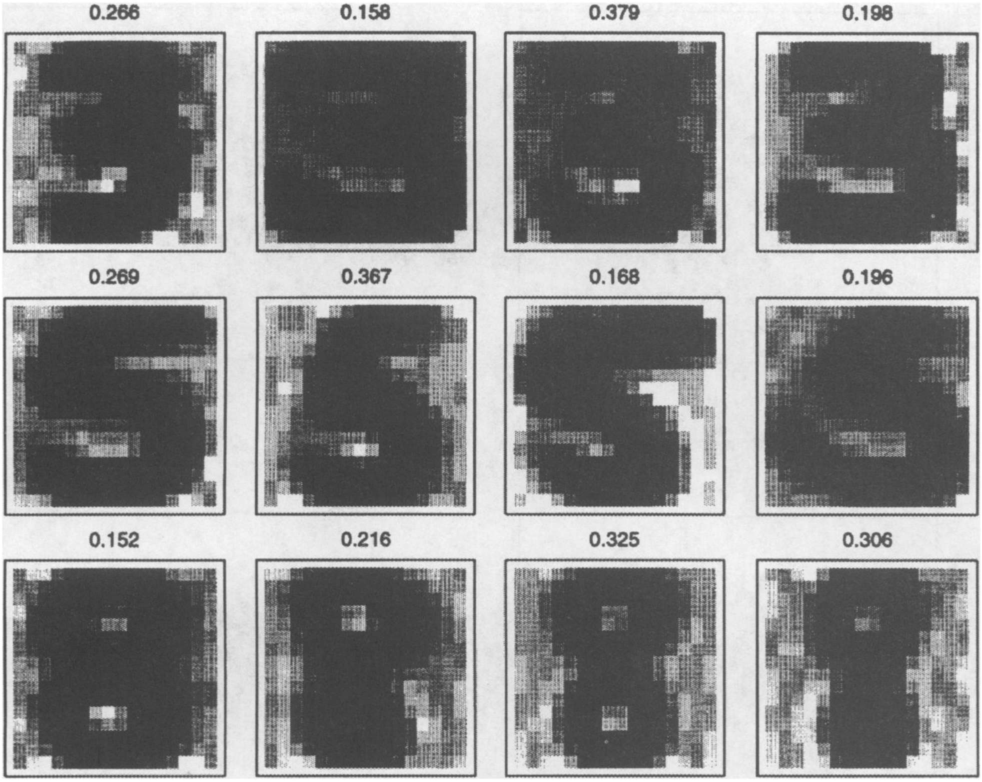

Fig. 5. Subclass centroids for the filtered MDA fit, displayed in unfiltered form: above each image, the within-class mixing parameter π_{jr} is indicated; not all the centroids are very different, which suggests that different numbers of subclasses could be used with different classes

7.1. *Mixture Discriminant Analysis with Centroid Shrinking*

The subspace-reduced model in proposition 2 is fitted iteratively by using weighted LDA, which essentially treats all classes and subclasses interchangeably. Our proposal here is to bias this decomposition, and the positions of the means themselves, in such a way that within each class the variance of the subclass centroids is damped relative to between-class variance. This will have the effect of pulling the centroids away from the decision boundaries, and hence making them smoother. We do this by penalizing the mixture likelihood for between-subclass variability. In the limit (infinite penalty), the standard single Gaussian per class LDA model emerges.

Let $\pi_j = \{\pi_{jk}\}$ be the vector of R_j subclass probabilities for the Jth class (summing to 1), and $\Delta_j = (I - \mathbf{1}\pi_j^{\mathrm{T}})$, where I is the identity matrix and $\mathbf{1}$ is a column vector of 1s. For any vector u with R_j elements, $\Delta_j u$ is a vector of deviations of the elements of u about their weighted mean, and likewise $u^{\mathrm{T}}\Delta_j^{\mathrm{T}}\Delta_j u$ is a positive scalar measure of the spread of u about its mean. If $Q_j = \Delta_j^{\mathrm{T}}\Delta_j$ is the corresponding penalty matrix, then $Q = \mathrm{diag}(\gamma_1 Q_1, \gamma_2 Q_2, \ldots, \gamma_J Q_J)$ is an appropriate, composite penalty matrix for penalizing the subcomponents of an R-vector for deviations about their class means. The relative strengths of the penalties on the diagonal can be controlled by the parameters γ_j.

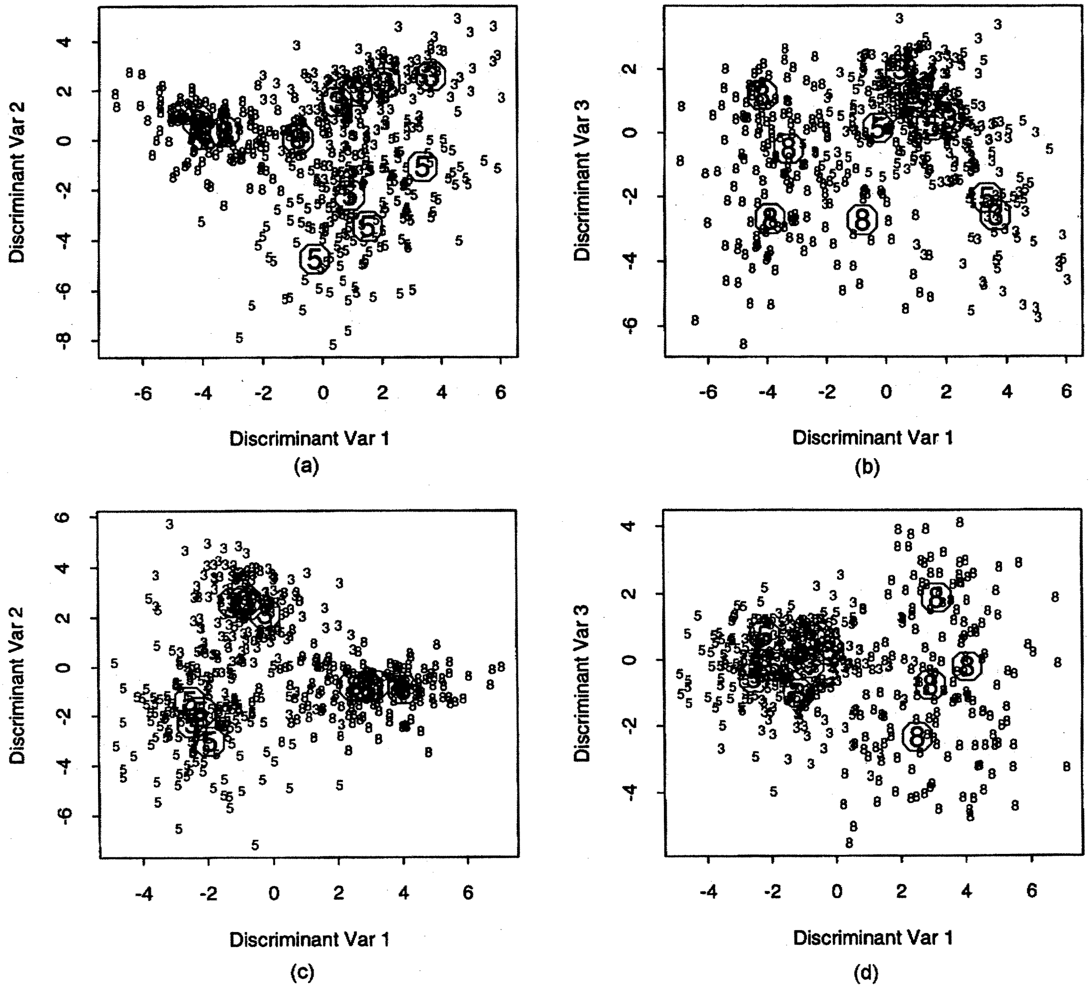

Fig. 6. (a), (b) First three discriminant co-ordinates for the MDA fit, with the centroids indicated (four subclasses, filtered); (c), (d) the corresponding co-ordinates for the shrunken MDA model (four subclasses, filtered)—in the shrunken model case, the leading two co-ordinates concentrate on group separation, whereas separation in the subclass centroids plays a stronger role for the 3s in the third co-ordinate; for (a), subclass separation is given as much weight as class separation, and we see more class overlap

We now consider maximizing a penalized version of log-likelihood (11).

Definition 1. Let U be the $R \times p$ rank K matrix of means for the rank K MDA problem. The shrunken MDA estimate of rank K maximizes the penalized mixture log-likelihood:

$$l_Q^{\text{mix}}(U, \Sigma, \pi_{jr}) = l^{\text{mix}}(U, \Sigma, \pi_{jr}) - \text{tr}(\Sigma^{-1} U^{\text{T}} Q U)/2 \qquad (13)$$

subject to $\text{rank}(U) = K$.

Our use in equation (13) of Σ^{-1} to weight the components of the penalty is explained in proposition 3 below.

An EM algorithm emerges just as in the unshrunk case, and the M-step corresponds to a penalized, reduced rank, augmented and weighted R-class Gaussian log-likelihood analogous to that in proposition 2:

$$2l_Q^{\text{weight}}(\mu_{rj}, \Sigma) = -\sum_{j=1}^{J} \sum_{g_i=j} \sum_{r=1}^{R_j} p(c_{jr}|x_i, j)\, D(x, \mu_{jr}) - N \log|\Sigma| - \operatorname{tr}(\Sigma^{-1} U^{\mathrm{T}} Q U) \tag{14}$$

subject to rank(U) = K. Unfortunately, the maximizer of log-likelihood (14) does not have a simple solution like the unshrunk version. The solution is characterized by an expression for the means and covariance just as in equations (18) and (20) in Appendix A. But, unlike there, the solution now requires iteration between the mean and the covariance estimators. Apart from the added computational complexity, this destroys the valuable link with optimal scoring.

We prefer to use a restricted maximum likelihood version, where we fix the estimated covariance at the weighted, pooled within-subclass covariance W as in equation (6). For this restricted case we have the following proposition.

Proposition 3. Consider maximizing the restricted, M-step, penalized log-likelihood

$$2l_Q^{\text{weight}}(\mu_{rj}|W) = -\sum_{j=1}^{J} \sum_{g_i=j} \sum_{r=1}^{R_j} p(c_{jr}|x_i, j)\, D(x, \mu_{jr}) - \operatorname{tr}(W^{-1} U^{\mathrm{T}} Q U) \tag{15}$$

subject to rank(U) = K. Here $D(\ ,\ )$ is defined in terms of W. The solution corresponds to a shrunken LDA. If M is the $R \times p$ weighted subclass centroid matrix, and D_p the diagonal weight matrix with rth element the sum of the weights for the rth subclass, then the shrunken LDA is based on the shrunken between-subclass matrix

$$B_Q = M D_p (D_p + Q)^{-1} D_p M.$$

This shrunken LDA can be fitted by using a modified MDA algorithm, in which the only change is that $Z^{\mathrm{T}}\hat{Z}$ is decomposed with normalization $\Theta^{\mathrm{T}}(D_p + Q)\Theta = I_K$.

For brevity we omit the proof, which is given in Hastie and Tibshirani (1993).

The form of the penalty in equations (13)–(15) requires some explanation. The modelled mean matrix U has R rows for subclasses and p columns for predictors. Our penalty matrix Q works on individual columns of U; any composite penalty must add these up in a sensible way, since the predictors are correlated. The rank K mean model can be written

$$U = U_0 + \Phi V^{\mathrm{T}} \Sigma \tag{16}$$

with $\Phi_{R\times K}$ and $V_{p\times K}$ ($V^{\mathrm{T}}\Sigma V = I_K$) to be determined. This is the form in which the unshrunken solutions emerge. Now $\operatorname{tr}(\Sigma^{-1}U^{\mathrm{T}}QU) = \operatorname{tr}(\Phi^{\mathrm{T}}Q\Phi)$, where $\Phi = (U - U_0)V$, the projection of the means onto the subspace spanned by the Σ-orthonormal basis V. Hence we simply add the penalties for the uncorrelated columns of Φ, which is equivalent to adding the Σ^{-1}-weighted penalties for the columns of U.

From the form of Q, it is clear that, as the γ_j becomes large, the only vectors permitted will be constant within a class, and the problem will degenerate to the standard LDA model.

7.2. *Choosing the Amount of Shrinking*

In practice it is difficult to guess suitable values for γ_j. One strategy is to let $\gamma_j = \gamma$ $\forall j$ and then to choose γ in some objective way, e.g. by cross-validation or evaluation on a test data set. Here we describe a strategy that helps the analyst to choose γ_j in a more subjective way.

In the proof of proposition 3, it emerges that the shrunken means $\tilde{M} = (D_p + Q_\gamma)^{-1} D_p M$ are used in the subsequent LDA. The shrinkage operator $S = (D_p + Q_\gamma)^{-1} D_p$ is like a smoother, and it makes sense to use its trace as the effective number of parameters after the shrinking (Hastie and Tibshirani, 1990). Given D_p, it is easy to work backwards and to derive γ such that $\mathrm{tr}(S) = \mathrm{df}$ for some nominal value of df. A subjective strategy is to select a (possibly large) number of subclass centres R_j for each class, and then to fit the shrunken model such that

(a) the effective overall number of centres is df, with

$$J \leqslant \mathrm{df} \leqslant R = \sum_j R_j.$$

or

(b) the effective number of centres per class is df_j, with $1 \leqslant \mathrm{df}_j \leqslant R_j$. Here we use the fact that Q and hence S is block diagonal, so we can compute df_j for each class.

Fig. 6 illustrates the former case, where we shrunk from a total of 12 centres to an effective total of eight. Notice how the leading two eigenvectors focus more on between-class separation (than for the unshrunken case), whereas the third starts to pick up within-subclass variation.

7.3. *Mixture Discriminant Analysis and Penalized Optimal Scoring*

In general steps 1–4 of the MDA–FDA or MDA–PDA optimal scoring algorithm can be seen to minimize

$$\sum_{k=1}^{K} \left[\sum_{j=1}^{J} \sum_{g_i=j} \sum_{r=1}^{R_j} p(c_{jr}|x_i,\, j) \{\theta_k(c_{rk}) - \eta_k(x_i)\}^2 + \lambda\, J(\eta_k) \right] \tag{17}$$

where $\eta_k(x) = x^{\mathrm{T}} \beta_k$, $J(\eta_k) = \beta_k^{\mathrm{T}} \Omega \beta_k$, Ω is a penalty matrix and the scoring functions θ_k are suitably normalized. Here x might represent the original predictors, and the penalty enforces smoothness of the coefficients vectors β_k themselves or else might be a basis expansion of the original predictors, and the penalty ensures that the composition $\eta_k(x)$ is smooth in the original domain.

A reasonable goal, consistent with our approach so far, would be to derive this penalized optimal scoring criterion as the M-step of a suitably regularized Gaussian (mixture) likelihood. Since in the unpenalized case these β_k are scaled versions of the right eigenvectors v_k in equation (16) (Hastie *et al.*, 1995), it might seem natural to maximize the mixture likelihood subject to an appropriate penalty on the v_k. This was the approach taken by Kiiveri (1989) in the context of penalized LDA.

We first analyse the simpler case of penalized LDA via a penalized Gaussian likelihood. Unfortunately, it does not lead to a simple (non-iterative) maximum likelihood estimate if Σ is unknown, and if Σ is assumed known and equal to W it still differs from the optimal scoring procedure. Hastie *et al.* (1995) show that optimal scoring corresponds to LDA with a fixed, penalized within-class covariance $W+\Omega$, i.e. Gaussian maximum likelihood for the rank-reduced means with Σ assumed known and equal to $W+\Omega$. This is also the approach taken by Leurgans *et al.* (1993).

It is more complicated for MDA since the corresponding version of W is a weighted within-covariance, which changes at each iteration. We thus cannot write down an explicit likelihood criterion. We nevertheless use the penalized MDA optimal scoring algorithm, which corresponds to using $W+\Omega$ in place of W in the M-step.

8. DISCUSSION

Classification by Gaussian mixtures is not a new idea. In this paper we have added to the functionality of this approach as follows.

(a) Reduced rank versions allow valuable low dimensional views of the data, even in the two-class case, and provide a natural means for regularization.

(b) The M-step of the EM algorithm can be solved via a weighted optimal scoring algorithm, which amounts to a multiple linear regression of a blurred response matrix Z.

(c) This allows natural generalizations by replacing the linear regression by more exotic forms:

   (i) adaptive nonparametric regressions enrich the procedure by expanding and transforming the predictor set;

   (ii) penalized regressions regularize the coefficients in cases where the predictors are sampled analogue signals over some spatial domain.

(d) We can shrink the between-subclass variability relative to the between-class variability. This often leads to more sensible low dimensional views and is a natural way to regularize in the presence of many subclasses.

(e) Our procedure provides a smooth enhancement to the LVQ algorithm, and in fact uses LVQ for initialization.

Hastie and Tibshirani (1993) illustrate the MDA procedure on some simulated examples; it performed favourably against a range of competitors, including neural networks. Further comparisons with other discrimination techniques, in a variety of problems, would be useful future work.

8.1. *Software*

We have written a set of functions for fitting FDA, PDA and MDA models in the S or S-PLUS language (Becker *et al.*, 1988; Chambers and Hastie, 1991). The function `fda( )` fits FDA and PDA models. A `method` argument allows the user to specify the multiresponse regression method to be used; the default is linear regression and thus Fisher's LDA. Other regression methods provided for FDA are polynomial regression, ridge regression, BRUTO (adaptive additive splines) and MARS. For PDA a generalized form of ridge regression is included. Users supply a penalty matrix and target degrees of freedom and the procedure derives the appropriate penalty constant. The mda( ) function has additional arguments for controlling the number of subclasses, their initialization and degrees of freedom parameters to control the shrinkage.

The software is reasonably efficient for moderately sized data sets, although ultimately it is limited by the data management strategies in S. For example the digit classification problem has 1756 training observations with 256 variables (pixels); a 12-subclass model took roughly 1 min on an SGI Indigo workstation. The main computational burden is the large regression. For smaller problems the computation time is not a problem, and typically fewer than 10 EM steps are required.

A variety of functions is provided for making predictions of varying types from these models (class predictions, posterior probability estimates, canonical coordinates), based on models of specified dimensions. Other functions are provided for producing plots, misclassification (confusion) matrices and for extracting coefficients. The mda software is publicly available from the statistics archive at Carnegie Mellon University with URL: http://lib.stat.cmu.edu/S/. The software and technical report are available from the first author's ftp site:

ftp://playfair.stanford.edu/pub/hastie

and are called mda.shar.Z and mda.tr.ps.Z.

ACKNOWLEDGEMENTS

The authors thank Andreas Buja, Richard Lippmann, Geoff McLachlan, Daryl Pregibon, Muni Srivastava and Scott VanderWiel for helpful discussions. In addition we thank the Editor, Associate Editor and referees for extremely helpful comments on an earlier draft. The second author was supported by the Natural Sciences and Engineering Research Council of Canada. This work was completed while the first author was a member of the technical staff at AT&T Bell Laboratories.

APPENDIX A: REDUCED RANK MODELS

In proposition 1 in Section 3 we claimed that maximizing the log-likelihood (8),

$$2l(\mu_j, \Sigma, K) = -\sum_{j=1}^{J} \sum_{g_i=j} (x_i - \mu_j)^{\mathrm{T}} \Sigma^{-1} (x_i - \mu_j) - N \log|\Sigma|$$

subject to rank$\{\mu_j\} = K$, is equivalent to reduced rank LDA. In the following outline of a proof, we draw heavily on Mardia *et al.* (1979) for some basic multivariate statistics results. Let B be the between-class covariance matrix and, for fixed Σ, let V denote the matrix of leading K eigenvectors of $\Sigma^{-1}B$.

A.1. Σ *Known,* μ_j *Unknown*

In section 12.5.2 (p. 338) of Mardia *et al.* (1979) a solution to equation (8) is given, assuming that Σ is known. This has the form of the usual LDA solution, except with W replaced by Σ. We can write the estimated means as

$$\hat{\mu}_j = \Sigma V V^{\mathrm{T}}(\bar{x}_j - \bar{x}) + \bar{x} \tag{18}$$

and thus the *estimated* (rank K) between-matrix as

$$\hat{B}_{(K)} = \Sigma V V^{\mathrm{T}} B V V^{\mathrm{T}} \Sigma. \tag{19}$$

A.2. μ_j *Known*, Σ *Unknown*

Although the case μ_j known, Σ unknown is not explicitly stated in Mardia *et al.* (1979), we deduce (and easily check) from their equation (4.2.7) on p. 104 that

$$\hat{\Sigma} = W + \sum_{j=1}^{J} \frac{N_j}{N} (\bar{x}_j - \mu_j)(\bar{x}_j - \mu_j)^{\mathrm{T}}. \tag{20}$$

These are each obtained by solving the partial score equations for μ_j or Σ, assuming that the other is known. The full maximum likelihood solution requires their simultaneous solution and suggests iteration. However, the solution is easier. We plug the estimated means (18) (using W for Σ) into equation (20), which gives

$$\begin{aligned}\hat{\Sigma} &= W + \sum_{j=1}^{J} \frac{N_j}{N} (\bar{x}_j - \hat{\mu}_j)(\bar{x}_j - \hat{\mu}_j)^{\mathrm{T}} \\ &= W + B - \hat{B}_{(K)} \\ &= W + B - W V V^{\mathrm{T}} B V V^{\mathrm{T}} W \\ &= W + W V_\perp V_\perp^{\mathrm{T}} B V_\perp V_\perp^{\mathrm{T}} W \end{aligned} \tag{21}$$

where $V_\perp^{\mathrm{T}} W V = 0$ and $V_\perp^{\mathrm{T}}$ spans the complementary $(p - K)$-dimensional subspace of R^p. To complete the proof, we show that the same V is optimal using the new metric $\hat{\Sigma}$.

First note that

(a) $V^{\mathrm{T}} \hat{\Sigma} V = V^{\mathrm{T}} W V + 0 = I_K$ and
(b) $BV = WVD_K = \hat{\Sigma} V D_K$, where the first equality is the definition of V, and $D_K = \mathrm{diag}(\gamma_1, \ldots, \gamma_K)$.

We have thus established that V is also an eigenmatrix of B with respect to $\hat{\Sigma}$; we have still to show that it has remained optimal.

Note that

$$\begin{aligned} V_\perp^{\mathrm{T}} \hat{\Sigma} V_\perp &= I_{p-K} + V_\perp^{\mathrm{T}} B V_\perp \\ &= I_{p-K} + D_{p-K} \end{aligned} \tag{22}$$

where D_{p-K} are the eigenvalues of B corresponding to $V_\perp$ and W. So, in the metric $\hat{\Sigma}$, the columns of V are orthonormal eigenvectors of B, $V_\perp$ is orthogonal and orthogonal to V. Thus the columns of $V_\perp$ remain eigenvectors of B with respect to $\hat{\Sigma}$, with eigenvalues $(I_{p-K} + D_{p-K})^{-1} D_{p-K} \leqslant D_{p-K}$, and thus the order does not change.

This shows that the constrained maximum likelihood estimated means coincide with the rank K LDA means. Using the fact that

$$\hat{\Sigma}^{-1} = V V^{\mathrm{T}} + V_\perp (D_{p-K}^{-1} (I_{p-K} + D_{p-K})) V_\perp^{\mathrm{T}},$$

it is not difficult to show that *relative* Mahalanobis distances

$$(x-\hat{\mu}_j)^{\mathrm{T}}\hat{\Sigma}^{-1}(x-\hat{\mu}_j)-(x-\hat{\mu}_l)^{\mathrm{T}}\hat{\Sigma}^{-1}(x-\hat{\mu}_l)$$

coincide with relative Euclidean distances in the reduced LDA space, and hence classification based on the fitted constrained Gaussian model and LDA coincide.

REFERENCES

Becker, R., Chambers, J. and Wilks, A. (1988) *The New S Language*. Belmont: Wadsworth.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984) *Classification and Regression Trees*. Belmont: Wadsworth.

Breiman, L. and Ihaka, R. (1984) Nonlinear discriminant analysis via scaling and ACE. *Technical Report*. University of California, Berkeley.

Campbell, N. (1984) Canonical variate analysis—a general formulation. *Aust. J. Statist.*, **26**, 86–96.

Chambers, J. and Hastie, T. (1991) *Statistical Models in S*. Pacific Grove: Wadsworth/Brooks Cole.

Cheng, B. and Titterington, D. (1994) Neural networks and statistical perspectives. *Statist. Sci.*, **9**, 2–54.

Friedman, J. (1991) Multivariate adaptive regression splines (with discussion). *Ann. Statist.*, **19**, 1–141.

Hastie, T., Buja, A. and Tibshirani, R. (1995) Penalized discriminant analysis. *Ann. Statist.*, **23**, 73–102.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. New York: Chapman and Hall.

——(1993) Discriminant analysis by mixture estimation. *Technical Report*. AT&T Bell Laboratories, Murray Hill.

Hastie, T., Tibshirani, R. and Buja, A. (1994) Flexible discriminant analysis by optimal scoring. *J. Am. Statist. Ass.*, **89**, 1255–1270.

Kiiveri, H. (1989) Canonical variate analysis of high dimensional data: smoothing canonical vectors. *Technical Report WA 89/1*. Commonwealth Scientific and Industrial Research Organisation.

Kohonen, T. (1989) *Self-organization and Associative Memory*, 3rd edn. Berlin: Springer.

Le Cun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R., Hubbard, W. and Jackel, L. (1990) Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems* (ed. D. Touretzky), vol. 2. Denver: Morgan Kaufman.

Leurgans, S. E., Moyeed, R. and Silverman, B. W. (1993) Canonical correlation analysis when the data are curves. *J. R. Statist. Soc.* B, **55**, 725–740.

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

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

Pal, N., Bezdek, J. and Tsao, E. (1993) Generalized clustering networks and Kohonen self-organizing scheme. *IEEE Trans. Neural Netwks*, **4**, 549–557.

Ripley, B. D. (1994) Neural networks and related methods for classification (with discussion). *J. R. Statist. Soc.* B, **56**, 409–456.

Taxt, T., Hjort, N. and Eikvil, L. (1991) Statistical classification using a linear mixture of multinormal probability densities. *Pattn Recogn Lett.*, **12**, 731–737.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985) *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

Tresp, V., Hollatz, J. and Ahmad, S. (1994) Representing probabilistic rules with networks of Gaussian basis functions. To be published.