Hastie, T., Tibshirani, R. & Buja, A. (1994*b*), 'Flexible discriminant analysis by optimal scoring', *Journal of the American Statistical Association* **89**, 1255–1270.

Hertz, J., Krogh, A. & Palmer, R. (1991), *Introduction to the theory of neural computation*, Addison Wesley, Redwood City.

Kiiveri, H. (1989), Canonical variate analysis of high dimensional data: Smoothing canonical vectors, Technical Report WA 89/1, DMS, CSIRO.

Kohonen, T. (1989), *Self-Organization and Associative Memory (3rd edition)*, Springer-Verlag, Berlin.

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

Leurgans, S., Moyeed, R. & Silverman, B. (1993), 'Canonical correlation analysis when the data are curves', *J. Royal Statist. Soc. (Series B)* **55**, 725–740.

Mardia, K., Kent, J. & Bibby, J. (1979), *Multivariate Analysis*, Academic Press.

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

Pal, N., Bezdek, J. & Tsao, E. (1993), 'Generalized clustering networks and kohonen self-organizing scheme', *IEEE trans. on Neural Networks* **4**, 549–557.

Ripley, B. (1992), Classification and clustering in spatial and image data, To appear, proc. of '15 Jahrestagung von Gesellschaft fur Klassification', H. Goehl and M. Schader ed.

Ripley, B. (1994), 'Neural networks and related methods for classification', *J. Royal Statist. Soc. (Series B)* pp. 409–456. (with discussion).

Taxt, T., Hjort, N. & Eikvil, L. (1991), 'Statistical classification using a linear mixture of multinormal probability densities', *Pattern Recognition Letters* **12**, 731–737.

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

We now use the results of Hastie et al. (1994*a*, section 3.4) to tie this decomposition to optimal scoring. This solution corresponds to a generalized SVD of $\Sigma_{12} = (Y^T X)/N = D_p M$ with left and right metrics $\Sigma_{11} = D_p + Q$ and $\Sigma_{22} = W$ respectively. This corresponds to optimal scoring with score-metric $D_p + Q$.

Remark: An interesting duality emerges. Subtracting a penalty from the log-likelihood to shrink the left eigen-vectors Φ leads to a $D_p + Q$ normalization for the optimal scoring problem. Adding a penalty to the optimal scoring criterion to smooth the right eigen-vectors leads to a $W + \Omega$ normalization for the likelihood/LDA problem. Both these can be used simultaneously, and amount to a left- and right-penalized canonical correlation analysis(Leurgans et al. 1993, Hastie et al. 1994*a*).

References

Becker, R., Chambers, J. & Wilks, A. (1988), *The New S Language*, Wadsworth International Group.

Breiman, L. & Ihaka, R. (1984), Nonlinear discriminant analysis via scaling and ACE, Technical report, Univ. of California, Berkeley.

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

Campbell, N. (1984), 'Canonical variate analysis — a general formulation', *Australian Journal of Statistics* **26**, 86–96.

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

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

Coomans, D., Broeckaert, M., M., J. & D.L., M. (1983), 'Comparison of multivariate discriminant techniques for clinical data - application to the thyroid functional state', *Meth. Inform. Med.* **22**, 93–101.

Friedman, J. (1991), 'Multivariate adaptive regression splines (with discussion)', *Annals of Statistics* **19**(1), 1–141.

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

Hastie, T., Buja, A. & Tibshirani, R. (1994*a*), 'Penalized discriminant analysis', *Annals of Statistics*. in press.

where S_λ is the *collapsed* $N \times N$ *smoother matrix*. Thus (25) can be written

$$RSS(\Theta) = \Theta^T Y^T (\mathcal{W} - \mathcal{W}M^T S_\lambda M\mathcal{W}) Y\Theta \tag{27}$$

Now $Z = M\mathcal{W}Y$ is the $N \times R$ *blurred* indicator matrix defined in section 5, and $D_p = Y^T\mathcal{W}Y$ is a diagonal $R \times R$ weight matrix with elements the sum of the probabilities in the columns of Z. Then (27) can be written

$$RSS(\Theta) = \Theta^T (D_p - Z^T S_\lambda Z)\Theta \tag{28}$$

and we wish to minimize $\mathrm{tr}RSS(\Theta)$ subject to $\Theta^T D_p \Theta = I_K$. This is what our MDA algorithm solves.

A.3 Centroid shrinking

Here we prove proposition 3. Let W be the weighted within subclass covariance matrix. We wish to maximize

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

where $U = U_0 + \Phi V^T W$. The solution minimizes

$$\mathrm{tr}D_p(M-U)W^{-1}(M-U)^T + \mathrm{tr}(\Phi^T Q\Phi) \tag{29}$$

where D_p is $R \times R$ diagonal, with rth element the sum of all the weights associated with the rth subclass. It is easily shown that $U_0 = 1\bar{x}^T$, and without loss of generality we assume the overall mean $\bar{x} = 0$. Let $V^* = W^{\frac{1}{2}}V$, $M^* = MW^{-\frac{1}{2}}$, then (29) can be written

$$\mathrm{tr}D_p(M^* - \Phi V^{*T})(M^* - \Phi V^{*T})^T + \mathrm{tr}(\Phi^T Q\Phi) \tag{30}$$

For fixed V^* and after some simple manipulations this leads to

$$\begin{aligned}\Phi &= (D_p + Q)^{-1}D_p M^* V^* \\ &= (D_p + Q)^{-1}D_p MV \end{aligned} \tag{31}$$

Plugging this back into (29) and simplifying we get

$$\mathrm{tr}(M^T D_p M W^{-1}) - \mathrm{tr}(V^T M^T D_p (D_p + Q)^{-1} D_p MV) \tag{32}$$

The second term involves V, and it corresponds to the eigendecomposition of the *shrunken between-subclass covariance matrix*

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

(while the unshrunk version is defined $B = M^T D_p M$).

distances $(x-\hat{\mu_j})^T\hat{\Sigma}^{-1}(x-\hat{\mu_j})-(x-\hat{\mu_\ell})^T\hat{\Sigma}^{-1}(x-\hat{\mu_\ell})$ coincide with relative Euclidean distances in the reduced LDA space, and hence classification based on the fitted constrained Gaussian model and LDA coincide.

A.2 The MDA algorithm fits a reduced-rank Gaussian-mixture discriminant model

In proposition 2 we have that the reduced rank Gaussian mixture model can be fit by an EM algorithm, with the M-step a reduced-rank augmented/weighted LDA problem. In Hastie et al. (1994*b*) we show that optimal scoring solves the LDA problem. It is immediate that the augmented, weighted LDA problem is solved by a similarly augmented and weighted optimal scoring procedure, but that is not what is presented in the MDA algorithm in section 5. For example, in MDA we still have N observations, while the augmented problem has $N^*=\sum N_jR_j>N$ pseudo observations.

For more generality we justify a penalized version of the algorithm. The augmented/weighted/penalized optimal scoring criterion for a single score θ is

$$\sum_{j=1}^{J}\sum_{g_i=j}\sum_{r=1}^{R_j}p(c_{jr}|x_i,j)(\theta_{jr}-\eta(x_i))^2+\lambda J(\eta) \tag{23}$$

where $\eta(x)=x^T\beta$ and $J(\eta)=\beta^T\Omega\beta$ and Ω is a penalty matrix.

It is convenient to convert to a matrix notation. Let Y be the augmented $N^*\times R$ indicator matrix representing the observations, and $\mathcal{W}$ a diagonal $N^*\times N^*$ weight matrix with entries the values $p(c_{jr}|x_i,j)$. M is a $N\times N^*$ *margin* matrix with a row for each (original) observation g_i consisting of a 1 for each of the R_{g_i} augmented observations associated with it. Notice that since the conditional probabilities sum to 1 for each observation, $M\mathcal{W}M^T=I_N$.

Let Θ be the $R\times K$ matrix of score vectors, and $X^*=M^TX$ be the augmented model matrix (with replicated rows). Then the version of (23) summed over all K scores θ_k is the trace of

$$RSS(\Theta,\mathrm{B})=(Y\Theta-X^*\mathrm{B})^T\mathcal{W}(Y\Theta-X^*\mathrm{B})+\lambda\mathrm{B}^T\Omega\mathrm{B} \tag{24}$$

which is to be minimized wrt B and Θ subject to $\Theta^TY^T\mathcal{W}Y\Theta=I_K$. We can solve for B for fixed Θ (Hastie et al. 1994*b*) to get

$$RSS(\Theta)=\Theta^TY^T(\mathcal{W}-\mathcal{W}S_\lambda^*)Y\Theta \tag{25}$$

where

$$\begin{aligned}S_\lambda^* &= X^*(X^*\mathcal{W}X^*+\lambda\Omega)^{-1}X^{*T}\mathcal{W}\\ &= M^TX(X^TX+\lambda\Omega)^{-1}X^TM\mathcal{W}\\ &= M^TS_\lambda M\mathcal{W}\end{aligned} \tag{26}$$

μ_j **known, $\boldsymbol{\Sigma}$ unknown** Although this case is not explicitly stated in MKB, one deduces (and easily checks) from their equation (4.2.7) on page 104 that

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

These are each obtained by solving the partial score equations for μ_j or Σ, assuming 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 (20) which gives

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

where $V_\perp^T WV = 0$ and $V_\perp^T$ 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 of all note that

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

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

Note that

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

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 wrt $\hat{\Sigma}$, with eigenvalues $(I_{p-K} + D_{p-K})^{-1}D_{p-K} \leq D_{p-K}$, 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 $\hat{\Sigma}^{-1} = VV^T + V_\perp(D_{p-K}^{-1}(I_{p-K} + D_{p-K})V_\perp^T$, is not hard to show that *relative* Mahalanobis provided for FDA are polynomial regression, ridge regression, BRUTO, and MARS. For PDA a generalized form of ridge regression is included. Users supply a penalty matrix and target *df*, and the procedure derives the appropriate penalty constant. The `mda()` function has additional arguments for controlling the number of subclasses, their initialization, and *df* parameters to control the shrinkage. A variety of functions are 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 software is publicly available from the statistics archive at Carnegie-Mellon University — `lib.stat.cmu.edu`. Either send the email message `send fda-mda from S` to `statlib@lib.stat.cmu.edu`, or else login to `statlib` via ftp and `cd S`.

Acknowledgements

The authors thank Andreas Buja, Richard Lippmann, Geoff McLachlan, Daryl Pregibon, Muni Srivastava and Scott VanderWiel for helpful discussion. The second author was supported by the Natural Sciences and Engineering Research Council of Canada.

A Appendix

A.1 Reduced rank models

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

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

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, hereafter MKB) 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 $\Sigma^{-1}B$.

Σ known, μ_j unknown In section 12.5.2 (page 338) of MKB a solution to (8) is given, assuming Σ is known. This has the form of the usual LDA solution, except with W replaced by Σ. One can write the estimated means as

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

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

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

Table 5: *Thyroid data results*

Technique	Error rates	
	Training	Test
LDA	0.091	0.083
MDA	0.028	0.042
MDA/FDA	0.049	0.014
FDA	0.049	0.042
Neural net (10 hidden units)	0.000	0.027

- Reduced rank versions allow valuable low dimensional views of the data, even in the two class case, and provide a natural means for regularization.
- 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.
- This allows natural generalizations by replacing the linear regression by more exotic forms:
  - Adaptive nonparametric regressions enrich the procedure by expanding and transforming the predictor set;
  - Penalized regressions regularize the coefficients in cases where the predictors are sampled analog signals over some spatial domain.
- 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.
- Our procedure provides a smooth enhancement to the LVQ algorithm, and in fact uses LVQ for initialization.

Further comparisons with other discrimination techniques, in a variety of different problems, would be useful future work.

Software

We have written a set of functions for fitting FDA, PDA and MDA models in the S/Splus language (Becker, Chambers & Wilks 1988, Chambers & Hastie 1991). The function `fda()` fits FDA and PDA models. A `method` argument allows the user to specify the multi-response regression method to be used; the default is linear regression and thus Fisher's LDA. Other regression methods

Table 4: *Results for Simulation 2. The values are averages over 10 simulations, with the standard error of the average in parentheses.*

Technique	—Error rates—			
			With 8 noise variables	
	Training	Test	Training	Test
LDA	0.473 (0.030)	0.474 (0.030)	0.422 (0.029)	0.476 (0.033)
MDA	0.034 (0.011)	0.043 (0.011)	0.048 (0.007)	0.099 (0.014)
MDA/FDA	0.025 (0.003)	0.040 (0.002)	0.028(0.005)	0.045 (0.007)
FDA	0.114 (0.021)	0.127 (0.019)	0.083 (0.009)	0.090 (0.009)
Neural net	0.032 (0.011)	0.053 (0.009)	0.094 (0.028)	0.270 (0.040)
LVQ2 (12 prototypes)	0.093 (0.024)	0.100 (0.024)	0.284 (0.017)	0.433 (0.018)

dimensions and has the lowest error rate.

8.3 Example: Thyroid data

In this example, the data consist of observations on 215 patients. The object is to predict the thyroid status (1=normal, 2=hyper, 3=hypo) based on 5 measurements:

1. T3-resin uptake test
2. Total Serum thyroxin
3. Total serum triiodothyronine
4. Basal thyroid-stimulating hormone
5. Maximal absolute difference of TSH value after injection of 200 micro grams of thyrotropin-releasing hormone as compared to the basal value.

They are described in Coomans, Broeckaert, M. & D.L. (1983). We took a random training sample of size 143, and the balance (72 observations) acted as the test sample. The classification results are shown in Table 5. The nonlinear methods all perform very well, with MDA/FDA winning by a small margin.

9 Discussion

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

Table 3: *Results for simulation 1. The values are averages over 10 simulations, with the standard error of the average in parentheses.*

Technique	—Error rates—			
			With 8 noise variables	
	Training	Test	Training	Test
LDA	0.335 (0.002)	0.335 (0.001)	0.315 (0.007)	0.367 (0.006)
MDA	0.098 (0.008)	0.111 (0.005)	0.156 (0.017)	0.287 (0.029)
MDA/FDA	0.096 (0.011)	0.132 (0.005)	0.107 (0.017)	0.188 (0.023)
FDA	0.115 (0.007)	0.161 (0.007)	0.107 (0.009)	0.165 (0.005)
Neural net	0.105 (0.069)	0.161 (0.027)	0.118 (0.030)	0.388 (0.012)
LVQ2	0.109 (0.006)	0.130 (0.005)	0.283 (0.001)	0.376 (0.001)

8.1 Simulation 1: separated clouds

Here we examine in more detail an example like that of Figure 1. There are two classes and two variables. The first class has 40 observations with a bivariate normal distribution with mean $(0,0)$, and 40 more with mean $(4,4)$. The second class has mean $(2,2)$. The covariance matrix is the identity is all cases. The left part of Table 3 shows a summary of the performance of a number of procedures over 10 realizations. The MDA procedure performs the best, and all of the adaptive procedures outperform LDA. FDA works here by transforming the predictor space so that the two subclasses of class 1 lie on top of one another. The right half of the table shows the results when eight $N(0,9)$ noise variables were added the predictor set. The deterioration in performance of MDA and the neural network is quite striking. FDA and MDA/FDA show the best performance, probably because of the role of FDA as a variable selector.

8.2 Simulation 2: unstructured data

In this example we simulated data with extremely disconnected class structure. There are 4 classes each with 3 spherical bivariate normal subclasses, having standard deviation 0.25. The means of the 12 subclasses were chosen at random (without replacement) from the integers $[1,2,\cdots 5]\times[1,2,\cdots 5]$. Each training sample had 20 observations per subclass, for a total of 240 observations.

For testing, we used test samples of size 720. Table 4 shows the error rates over 10 simulations. MDA, MDA/FDA and the neural net clearly outperform the other procedures. FDA and LVQ2 also offer large improvements over LDA. In the right half of the table we have added 8 independent $N(0,1)$ noise variables. The performance of MDA degrades a little, while the neural net, and especially LVQ2 are adversely affected. MDA/FDA is able to select the important input 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 (16)(Hastie et al. 1994*a*), 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 analyze the simpler case of penalized LDA via a penalized Gaussian likelihood. Unfortunately, it does not lead to a simple (non-iterative) MLE 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. (1994*a*) show that optimal scoring corresponds to LDA with a fixed, penalized within-class covariance $W + \Omega$. In other words, 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, Moyeed & Silverman (1993).

For MDA, things get more complicated since the corresponding version of W is a weighted within-covariance, which changes at each iteration. We are thus unable to 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 More examples

Here we use two simulated examples and some real data to compare a number of classification procedures. The methods include

1. LDA — linear discriminant analysis, using all of the variables.

2. MDA — mixture discriminant analysis as described in section 2.

3. FDA — flexible discriminant analysis using adaptive backfitting as described in Hastie et al. (1994*b*)

4. MDA/FDA — mixture discriminant analysis combined with flexible discriminant analysis, as described in section 5.

5. Neural network — a single layer perceptron with sigmoidal outputs, cross-entropy cost function, weight decay and variable metric optimizer(Ripley 1992).

6. LVQ2 — a version of Kohonen's learning vector quantization, as described in Hertz, Krogh & Palmer (1991).

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 get 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 choose γ in some objective way, for example by cross-validation or evaluation on a test data set. Here we describe a strategy that helps the analyst choose γ_j in a more subjective way.

In the proof of proposition A.3, it emerges that the shrunken means $\tilde{M} = (D_p + Q_\gamma)^{-1}D_pM$ 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 & Tibshirani 1990). Given D_p, it is easy to work backwards and derive γ such that $\mathrm{tr}(S) = df$ for some nominal value of df. A subjective strategy is to select a (large) number of subclass centers R_j for each class, and then fit the shrunken model such that:

- The effective overall number of centers is df, with $J \leq df \leq R = \sum_j R_j$; or
- The effective number of centers per class is df_j, with $1 \leq df_j \leq R_j$. Here we use the fact that Q and hence S is block diagonal, so one can compute df_j for each class.

Figure 2 illustrates this latter case, where 3 subclasses where shrunk to an effective 2.5 per class. Similarly, in figure 6 we shrunk from a total of 12 centers to an effective total of 8. Notice how the leading two eigenvectors focus more on between-class separation (than for the unshrunk case), while the third starts picking up within-subclass variation.

7.3 MDA 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 $\eta_k(x) = x^T\beta_k$, $J(\eta_k) = \beta_k^T\Omega\beta_k$, Ω is a penalty matrix, and the scoring functions θ_k are suitable normalized. Here x might represent the original predictors, and the penalty enforces smoothness of the coefficients vectors β_k

Gaussian log-likelihood analogous to that in proposition 2:

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

subject to rank$(U) = K$. Unfortunately, the maximizer of (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 (18) and (20) in appendix A.1. 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 version, where we fix the estimated covariance at the weighted, pooled within-subclass covariance W as in (6). For this restricted case we have:

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

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

subject to rank$(U) = K$. *Here* $D(\cdot, \cdot)$ *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 fit using a modified MDA algorithm, in which the only change is that $Z^T \hat{Z}$ *is decomposed with normalization* $\Theta^T (D_p + Q) \Theta = I_K$.

A proof is given in Appendix A.3.

The form of the penalty in (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 has to add these up in a sensible way, since the predictors are correlated. The rank-K mean model can be written

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

with $\Phi_{R \times K}$ and $V_{p \times K}$ ($V^T \Sigma V = I_K$) to be determined. This is the form in which the unshrunken solutions emerge. Now $\mathrm{tr}(\Sigma^{-1} U^T Q U) = \mathrm{tr}(\Phi^T Q \Phi)$, 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

Figure 5 displays the fitted subclass centroids as images: some spatial differences are apparent within a class. Figure 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 sub-class centers, as well as the interpretation of penalized optimal scoring in the context of the present model.

7.1 MDA with centroid shrinking

The subspace reduced model in proposition 2 is fit iteratively 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. We do this by penalizing the mixture likelihood for between-subclass variability. In the limit (infinite penalty), we expect the standard single-Gaussian-per-class LDA model to emerge.

Let $\pi_j = \{\pi_{jk}\}$ be the vector of R_j subclass probabilities for the Jth class (summing to 1), and $\Delta_j = (I - 1\pi_j^T)$, where I is the identity matrix and 1 is a column vector of ones. 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 $u^T \Delta_j^T \Delta_j u$ is a positive, scalar measure of the spread of u about its mean. If $Q_j = \Delta_j^T \Delta_j$ 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 sub-components 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.

We now consider maximizing a penalized version of (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:*

$$\ell_Q^{mix}(U, \Sigma, \pi_{jr}) = \ell^{mix}(U, \Sigma, \pi_{jr}) - tr(\Sigma^{-1} U^T Q U)/2 \tag{13}$$

subject to $rank(U) = K$.

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

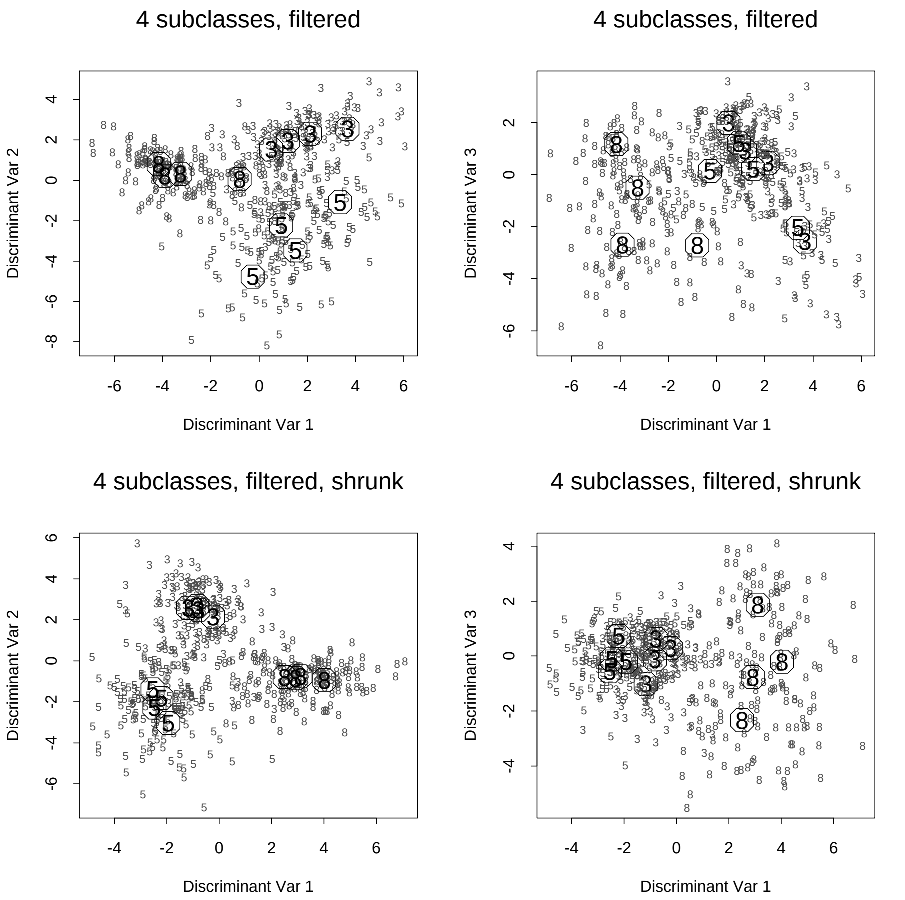

Figure 6: *The top two panels show the the first 3 discriminant coordinates for the MDA fit, with the centroids indicated. The lower two panels show the corresponding coordinates for the shrunken MDA model. We see that in the latter case, the leading two coordinates concentrate on group separation, while separation in the subclass centroids plays a stronger role for the 3s in the 3rd coordinate. For the top left plot, subclass separation is given as much weight as class separation, and we see more class overlap.*

Table 2: *Digit classification results — 3s, 5s and 8s. The filtered models correspond 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 8.*

Technique	Error rates (%)	
	Training	Test
LDA	1.6	8.7
LDA (filtered – 64df)	3.1	7.1
MDA (filtered – 64df, 4 subclasses)	2.7	7.1
MDA (filtered – 49df, 4 subclasses)	2.3	6.3
MDA (filtered – 64df, 4 subclasses, shrunk)	2.7	5.8

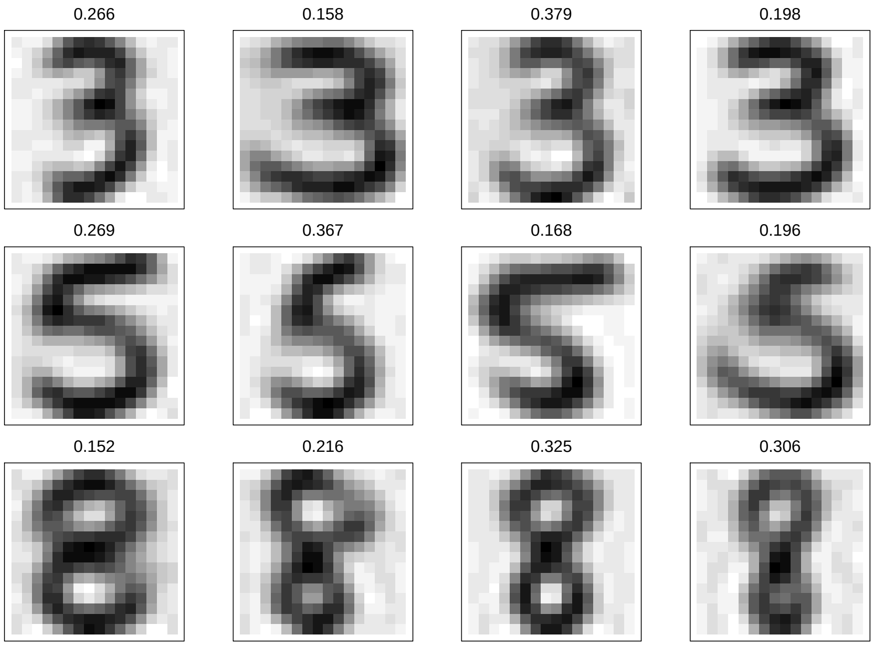

Figure 5: *The subclass centroids for the filtered MDA fit, displayed in unfiltered form. Above each image, the within-class mixing parameter π_{jr} is indicated. Not all the centroids are very different, which suggests that different numbers could be used with different classes.*

6 Example: Handwritten Digit Recognition

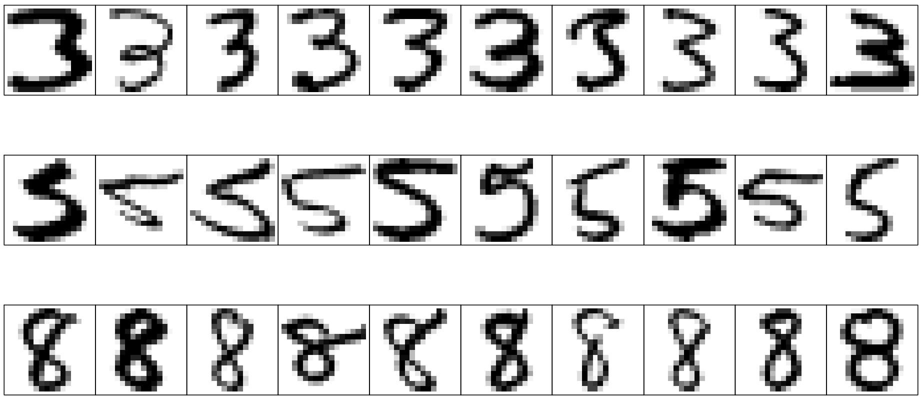

Figure 4: *A random selection of digitized handwritten 3s, 5s and 8s. Each image is a 8 bit, grayscale version of the original binary image, size and orientation normalized to* 16×16 *pixels.*

In Hastie et al. (1994*a*) we illustrated PDA on a handwritten digit recognition task. Here we focus on a difficult sub-task, 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 grayscale images. We have 658 threes, 556 fives and 542 eights in our training set, and roughly a quarter the number of test examples. Figure 4 shows some examples from each.

The inputs are highly correlated because of their spatial arrangement, and some kind of smoothing or filtering always helps. We used a 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, 3rd degree thin-plate spline smoother; 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 $x^* = Px$. 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:

- reducing M from 64 to 49 yielded a 12% improvement;
- with $M = 64$, and shrinking to a total of 8 effective centers yielded a 20% improvement.

In figure 1 and section 8 we demonstrate the use of the FDA extension to MDA on some examples. Our experience there, and on other examples tried so far, suggests:

- 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. Figure 1 supports this claim. One can imagine that in difficult situations the two components can support each other:
  - if too few centers are provided, nonlinearities can compensate;
  - extreme nonlinearities, which are costly in terms of fitted degrees of freedom, can be avoided by using mixtures.
- the adaptive aspect of FDA plays the role of a variable selector; the examples in section 8 support this claim.

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

- The waveform predictors in section 4 represent noisy functions sampled at 21 evenly spaced abscissae. The discriminant coefficients were penalized using a 2nd-derivative roughness penalty, reducing the effective number of parameters from 21 down to 4, and gave on average an 8% improvement in the classification performance. Figure 3 shows the fitted penalized discriminant functions $(\hat{\Sigma} + \Omega)^{-1}\hat{\mu}_{jr}$ for each subclass — and compares them to the unpenalized versions. We see that penalization plays two roles:
  - Variance reduction: the unpenalized MDA discriminant functions appear unnecessarily wiggly, for which we pay a price in generalization errors.
  - 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 harder and less convincing to attempt the same interpretations using the unpenalized discriminant functions.
- the handwritten digit example in the next section is a natural for both mixture models and for penalization:
  - There are several characteristic ways for writing each digit, calling for a mixture model;
  - pixel grayscale values have strong spatial correlation, calling for regularization for the same reasons as given in the waveform example.

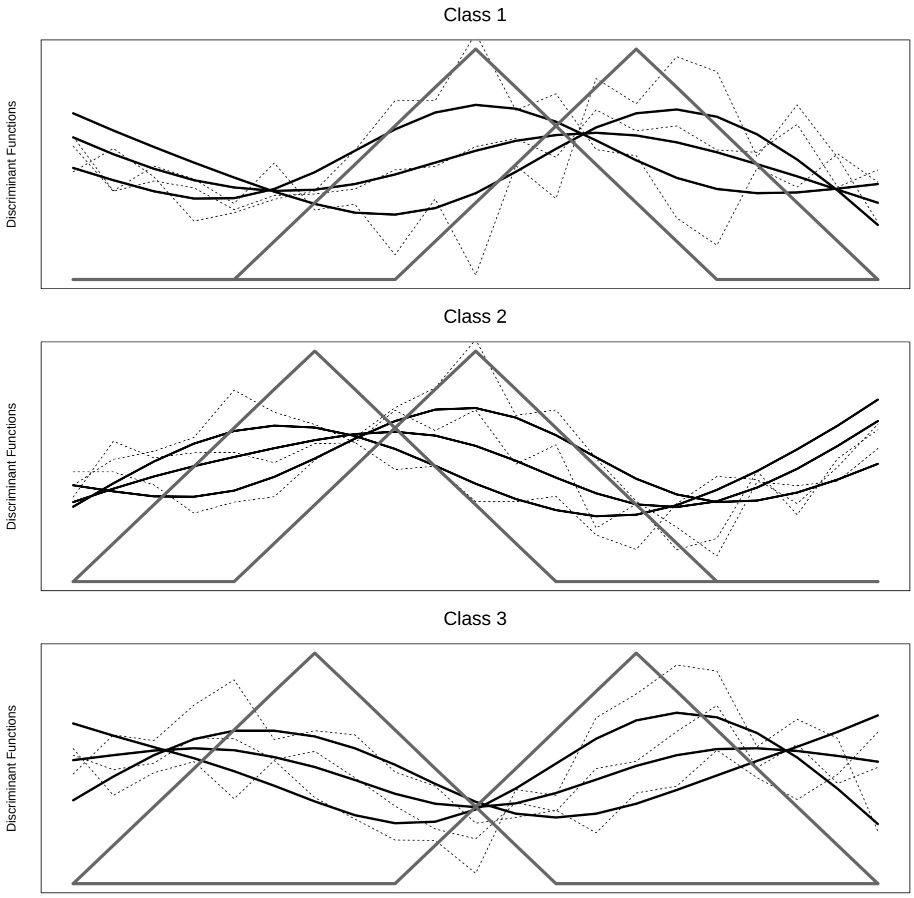

Figure 3: *Discriminant functions for the subclasses in the waveform example; each panel represents the subclasses for a single class. The smooth functions are* $(\hat{\Sigma} + \Omega)^{-1}\hat{\mu}_{jr}$*, produced by the MDA algorithm using a penalized regression method. The wiggly functions are the unpenalized versions* $\hat{\Sigma}^{-1}\hat{\mu}_{jr}$*. Superimposed on the figures are the generating waveforms, shifted and scaled to fit into the plots.*

K-dimensional fit $\eta(x)$ produced by the algorithm is actually the coordinate 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 (3) and (7), we have all the ingredients needed for fitting the model, and for classification. Details of these equivalences are given in Hastie et al (1992a, b).

One advantage of this approach is immediately clear: the M-step in proposition 2 involved a weighted LDA of $\sum_j N_j R_j$ observations, while 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 more exotic than linear regression. In the context of LDA, these ideas were originally proposed by Breiman & Ihaka (1984), and developed further by Hastie et al (1992a, b). Two classes of such generalizations emerge:

FDA: Here multivariate adaptive nonlinear regression replaces the linear regression, and the linear map $\eta(X)$ becomes a nonlinear map. Hastie et al. (1994*b*) describe the class of nonlinear 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 call the resulting procedure "FDA" for flexible discriminant analysis. 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.

PDA: Here we fit a more restrictive linear map via penalized least squares (Hastie et al. 1994*a*). The idea is that the feature vectors x_i arise as digitized analog signals, and are spatially correlated (pixels in an image, log-spectra at different frequencies). The penalization amounts to regularizing the large, near 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 neighboring values.)

Again we don't explicitly compute the means and covariance in the (transformed) feature space; all that is needed are the regression fits $\eta(x)$. In the case of FDA, this feature space is often a high-dimensional space of basis functions, so this simplification is very important computationally.

representing the classes, then we can regress Y against the predictor matrix X to get fitted values $\hat{Y} = HY$. This is followed by a suitably normalized eigen-decomposition of $Y^T HY = Y^T \hat{Y}$, which contains all the ingredients for an LDA of any rank (Breiman & Ihaka (1984), Hastie et al. (1994*b*)).

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 eigen-decomposition, just as in the LDA case above.

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

Mixture Discriminant Analysis by Optimal Scoring: MDA

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 a LVQ or K-means preprocessing of the data. Let $R = \sum R_j$.

Iterate:

1. *Compute 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 $\hat{p}(c_{jr}|x_i, j)$, $r = 1, \dots, R_j$, and the rest with zeros. Z is the mixture analog of an indicator-response matrix, except observations can belong to several (sub-)classes with associated probabilities.
2. *Multivariate linear regression:* Fit a multi-response, linear regression of Z on X. Let $\hat{Z}$ be the fitted values, and $\eta(x)$ be the vector of fitted regression functions.
3. *Optimal scores:* Let Θ be the largest K nontrivial eigenvectors of $Z^T \hat{Z}$, with normalization $\Theta^T D_p \Theta = I_K$. 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).
4. *Update* the fitted model from step 2 using the optimal scores: $\eta(x) \leftarrow \Theta^T \eta(x)$.
5. *Update* $\hat{p}(c_{jr}|x_i, j)$ and $\hat{\pi}_{jr}$ using formulas (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. A proof is given in Appendix A.2. We don't actually obtain estimates of the means and covariance from this MDA algorithm. The

Table 1: *Results for waveform data. 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*b*). The first model below the line is MDA with 3 subclasses per class. The next line is the same, except that the discriminant coefficients are penalized via a roughness penalty to effectively 4df. The third is the same as the second, with the centroids shrunk to effectively 2.5 centroids per class.*

Technique	Error Rates	
	Training	Test
LDA	0.121(.006)	0.191(.006)
QDA	0.039(.004)	0.205(.006)
CART	0.072(.003)	0.289(.004)
FDA/MARS (degree = 1)	0.100(.006)	0.191(.006)
FDA/MARS (degree = 2)	0.068(.004)	0.215(.002)
MDA (3 subclasses)	0.087(.005)	0.169(.006)
MDA (3 subclasses, penalized 4df)	0.137(.006)	0.157(.005)
MDA (3 subclasses shrunk to 2.5, penalized 4df)	0.139(.005)	0.155(.005)

Table 2 extends a simulation of Hastie et al. (1994*b*) 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 three MDA models are described in the caption. More details on the penalization and centroid shrinkage is given in section 7.

Figure 2 shows the leading two canonical variates for the last two models, evaluated at 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 3 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. 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 MDA by optimal scoring

One can use "optimal scoring"— multiple linear regression followed by an eigen analysis— to fit both the LDA and 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* more attractive than the exact reduced rank solution described above. This is because the latter requires one to iteratively fit the model 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

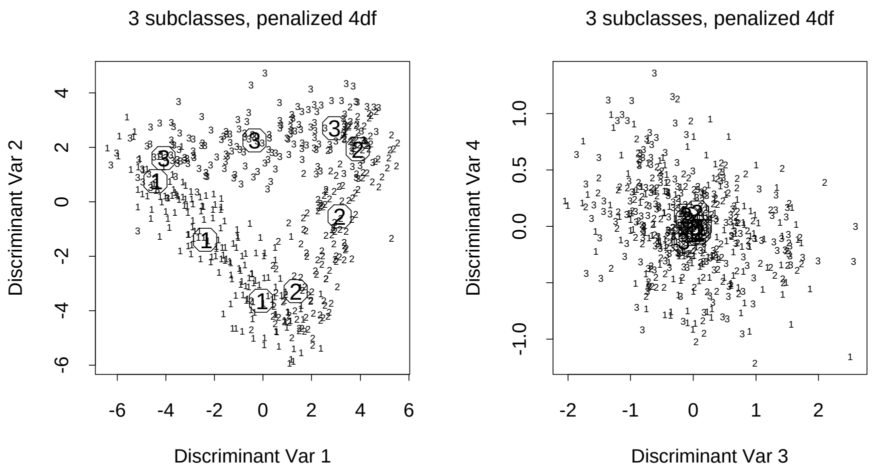

Figure 2: *A two dimensional view of the MDA model fit to a sample of the waveform model. The points are independent test data, projected onto the leading two canonical coordinates. The subclass centers are indicated.*

We now illustrate some of these ideas on a popular simulated example, taken from Breiman et al. (1984, pg 49-55), and used in Hastie et al. (1994*b*) 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

$$\begin{aligned} x_i &= uh_1(i) + (1-u)h_2(i) + \epsilon_i \qquad \text{Class 1} \\ x_i &= uh_1(i) + (1-u)h_3(i) + \epsilon_i \qquad \text{Class 2} \\ x_i &= uh_2(i) + (1-u)h_3(i) + \epsilon_i \qquad \text{Class 3} \end{aligned} \tag{12}$$

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 $h_3(i) = h_1(i + 4)$.

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 centers 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 $p(c_{j\ell}|x_i, j)$. 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 right away:

- Step (3) of the EM algorithm requires only *differences* in Mahalanobis distances from the estimated centers. This is convenient, since weighted LDA will only supply distances in the K dimensional subspace of span$\{\mu_{jr}\}$.
- 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*

$$\ell^{mix}(\mu_{jr}, \Sigma, \pi_{jr}) \propto \sum_{j=1}^{J} \sum_{g_i=j} \log \sum_{r=1}^{R_j} \pi_{jr} e^{-D(x,\mu_{jr})/2} - \frac{N}{2} \log |\Sigma| \qquad (11)$$

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

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

subject to rank$\{\mu_j\} = K$*. This latter maximization can be achieved via a similarly augmented and weighted rank-*K *LDA.*

The lower left panel in figure 1 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

In fact, the procedure implemented in our software and used in this paper, is the EM algorithm above, but with the M step carried out by "optimal scoring" as described in section 5. As we will show, this has computational advantages and facilitates powerful generalizations of the model.

3 Reduced rank discrimination

In linear discriminant analysis with J classes, one 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, Kent & Bibby 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 then extend this concept to the mixture model.

Proposition 1 *Consider maximizing the Gaussian log-likelihood*

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

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

$$\begin{aligned}
\hat{\mu}_j &= WVV^T(\bar{x}_j - \bar{x}) + \bar{x} && (9)\\
\hat{\Sigma} &= W + \sum_{j=1}^{J}\frac{N_j}{N}(\bar{x}_j - \hat{\mu}_j)(\bar{x}_j - \hat{\mu}_j)^T &&\\
&= W + WV_\perp V_\perp^T B V_\perp V_\perp^T W && (10)
\end{aligned}$$

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:

- *$\hat{\mu}_j$ is the projection of the jth sample mean onto the discriminant subspace of rank K.*
- *classification based on $(x-\hat{\mu}_j)^T\hat{\Sigma}^{-1}(x-\hat{\mu}_j)$ 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.1.

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: $\text{rank}\{\mu_{rj}\} = K$.

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

In the above (3) and (4) are the *Estimation* step, while (5) and (6) are the *Maximization* step. These are a straightforward generalization of the EM algorithm for estimating normal mixtures (e.g. Titterington, Smith and Makov, 1985, page 86-87). (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 $\hat{p}(c_{jr}|x_i, j)$, 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 $\hat{\Sigma}$, the equations (3–6) must be iterated.

The posterior class probabilities are (by Bayes theorem)

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

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 nonlinear.

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:

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 0 otherwise.

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 (5) and $\hat{\Sigma}$ from (6), and the iterations go from there.

Both k-means and LVQ require starting centers, which are typically randomly selected x_i. It is known (Pal, Bezdek & Tsao 1993, for example) that both these procedures suffer from variability due to random starting values. Our strategy is to try a number of different starts, and 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 kmeans) with respect to classification of the training data.

Bayes theorem and class priors to give models for the class posterior probabilities $P(G\,|\,X)$ — a basic ingredient for classification.

Suppose 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, one 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 will see, it has the right structure to permit the other generalizations we have in mind.

Let Π_j be the prior for class j, and within class j let π_{jr} be the mixing probability for the rth subclass, $\sum_{r=1}^{R_j} \pi_{jr} = 1$. Note that while often the Π_j are known or easily estimated from the training data, the π_{jr} are unknown model parameters. Let $D(x, \mu) = (x - \mu)^T \Sigma^{-1}(x - u)$ be the Mahalanobis distance between x and μ.

The mixture density for class j is

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

and the conditional log-likelihood for the data is

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

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

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

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

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

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

criminant analysis 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, Buja & Tibshirani 1994*a*) is a closely related technique for classifying digitized analog 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 penalized discriminant analysis 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 Gaussians rather than a single one per class, and still use a penalized metric when computing distances.

Figure 1 illustrates some of these techniques with a simple example. Not surprisingly the MDA technique has the most satisfactory boundary (the problem is a setup for MDA), and hence its boundary is optimal. The LVQ boundary seems unnecessarily biased and wiggly. The FDA boundary is a reasonable approximation to the optimal boundary. The lower-left panel shows a rank-1 version of the MDA fit, which also does acceptably well. The lower right panel shows the result of MDA combined with the nonlinear transformations of FDA — there is no noticeable improvement.

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

- The classes are modelled as mixtures of Gaussians, rather than a single Gaussian as in LDA;
- Optimal subspace identification is possible as in LDA, with added functionality;
- We can shrink the subclass centers, for example towards a common center;
- 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, Friedman, Olshen & Stone (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*b*), which in turn allows useful nonparametric extensions. Section 6 further illustrates these techniques on a handwritten digit-recognition problem. Section 7 is about the centroid shrinking, and section 8 describes some further comparative examples.

2 Extending LDA by normal mixtures

The approach to classification taken here is to model the class densities of the predictors $P(X \mid G)$ by Gaussian mixture models. These are flipped around via

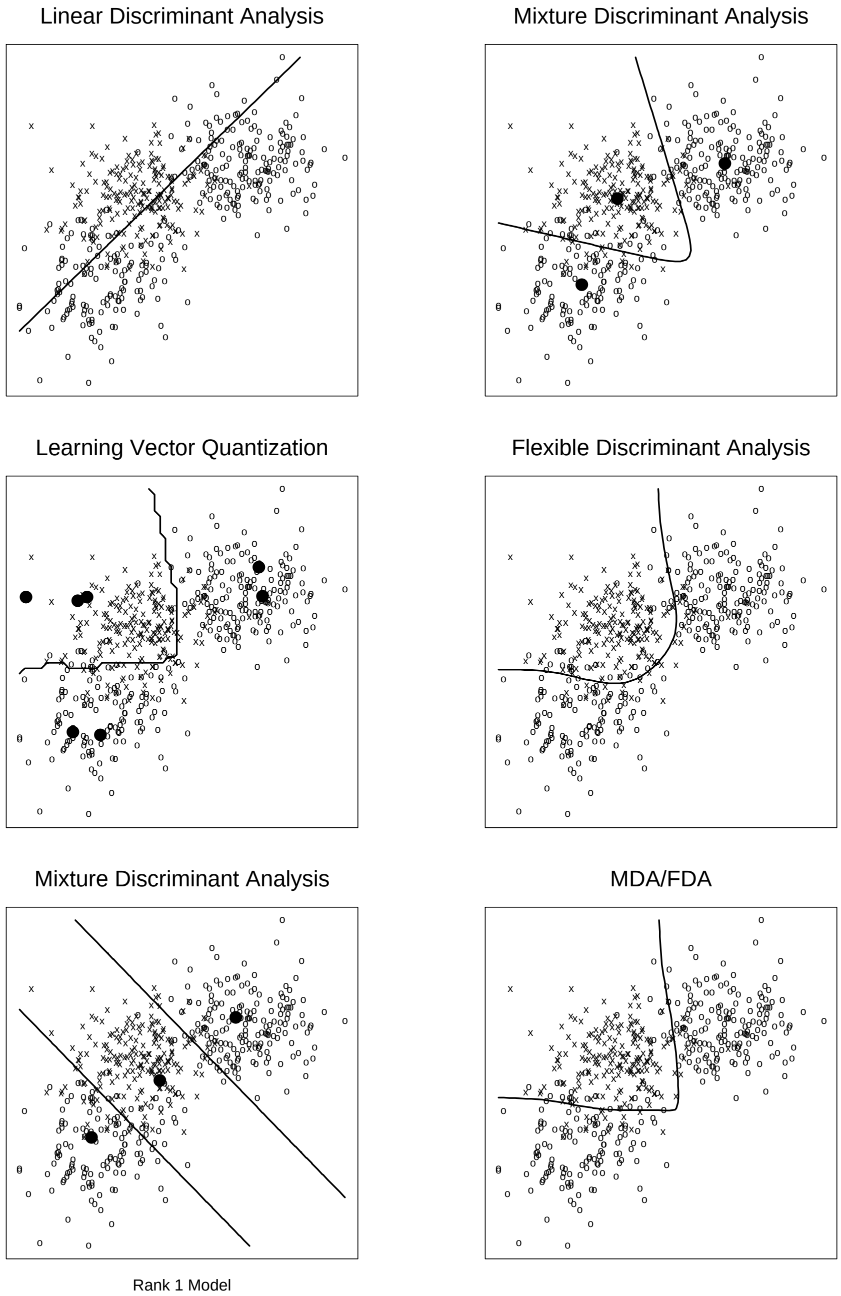

Figure 1: *A two class problem in which one of the classes occurs in two separated circular clouds. The panels are labeled according to the technique that produced the decision boundary.*

statistical methods for this problem include linear discriminant analysis (LDA) and multiple logistic regression, nearest neighbor 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 nonlinear model. Ripley (1994) gives an informative review from a statisticians' 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 in fact a mixture of unobserved normally distributed subclasses. This approach is sometimes mentioned in the statistical literature (McLachlan 1992, Cheng & Titterington 1994, for example) and pattern recognition literature (Taxt, Hjort & Eikvil 1991), but does not seem to have generated much attention.

In this paper we develop the mixture approach to discrimination in a number of interesting and useful directions. For brevity we refer to this mixture discriminant analysis technique as MDA.

LDA is not only a classification procedure, but also a data-reduction tool. We can represent multi-class 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 coordinates 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 coordinate, while MDA will produce up to one less than the number of subclasses in the mixture representation.

A technique known as *Learning Vector Quantization* or LVQ has received a lot of attention in the pattern recognition literature (Kohonen 1989); MDA can be viewed as a smooth version of LVQ. LVQ finds a *set* of cluster centers for each class; classification is performed by finding the closest center, and assigning the associated class. The online learning algorithms for LVQ are similar to the k-means algorithm, with biases built in to encourage good classification. While 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 centers relative to their between-class spread. This emphasizes our bias towards classification, much like LVQ biases clustering techniques towards classification. It also allows us to use many subclass centers per class, and shrink them until the *effective number* of centers is a required smaller amount.

Other methods have been proposed to generalize LDA to allow nonlinear decision boundaries. In Hastie, Tibshirani & Buja (1994*b*), 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 & Ihaka (1984). Hastie et al. (1994*b*) call 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* dis-

Discriminant Analysis by Gaussian Mixtures

Trevor Hastie
Statistics and Data Analysis Research Group
AT&T Bell Laboratories
Murray Hill, New Jersey

Robert Tibshirani
Department of Preventive Medicine and Biostatistics
and Department of Statistics
University of Toronto
Toronto, Ontario

February 8, 1994

Abstract

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 are able to control the within-class spread of the subclass centers relative to the between-class spread. Our technique for fitting these models permits a natural blend with nonparametric versions of LDA.

Keywords: Classification, Pattern Recognition, Clustering, Nonparametric, Penalized.

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 $\mathcal{J} = \{1, 2, 3, \cdots J\}$. 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