Flexible Discriminant and Mixture Models

Trevor Hastie

trevor@playfair.stanford.edu

Statistics Department

and Division of Biostatistics

Stanford University

joint with Andreas Buja and Rob Tibshirani

April 28, 1997

Papers fda.ps.Z, pda.ps.Z and mda.ps.Z are available from:

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

Theme: Modular Extensions of Standard Tools

Linear Discriminant Analysis or LDA is a classic technique for discrimination and classification

Virtues of LDA:

+ Simple prototype method for multiple class classification
+ Can produce optimal low dimensional views of the data
+ Sometimes produces the best results; e.g. LDA featured in top 3 classifiers for 11/22 of the STATLOG datasets, overall winner in 3/22.

Limitations of LDA:

- Lots of data, many predictors: LDA underfits (restricts to linear boundaries)
- Many correlated predictors: LDA (noisy/wiggly coefficients)
- Dimension reduction limited by the number of classes

Example of extension: FDA

- $\hat{Y} = S_X(Y)$ where Y is an indicator response matrix and S_X a regression procedure (Linear regression, Polynomial Regression, Additive Models, MARS, Neural Network, $\cdots$)
- eigen$Y^T\hat{Y} =$ eigen$Y^T S_X Y \Rightarrow$ LDA, flexible extensions of LDA.

Typically this amounts to expanding/selecting the predictors via basis transformations chosen by regression, and then (penalized) LDA in the new space.

Example: Vowel Recognition

Vowel	Word	Vowel	Word
i	heed	o	hod
I	hid	c:	hoard
E	head	U	hood
A	had	u:	who'd
a:	hard	3:	heard
y	hud		

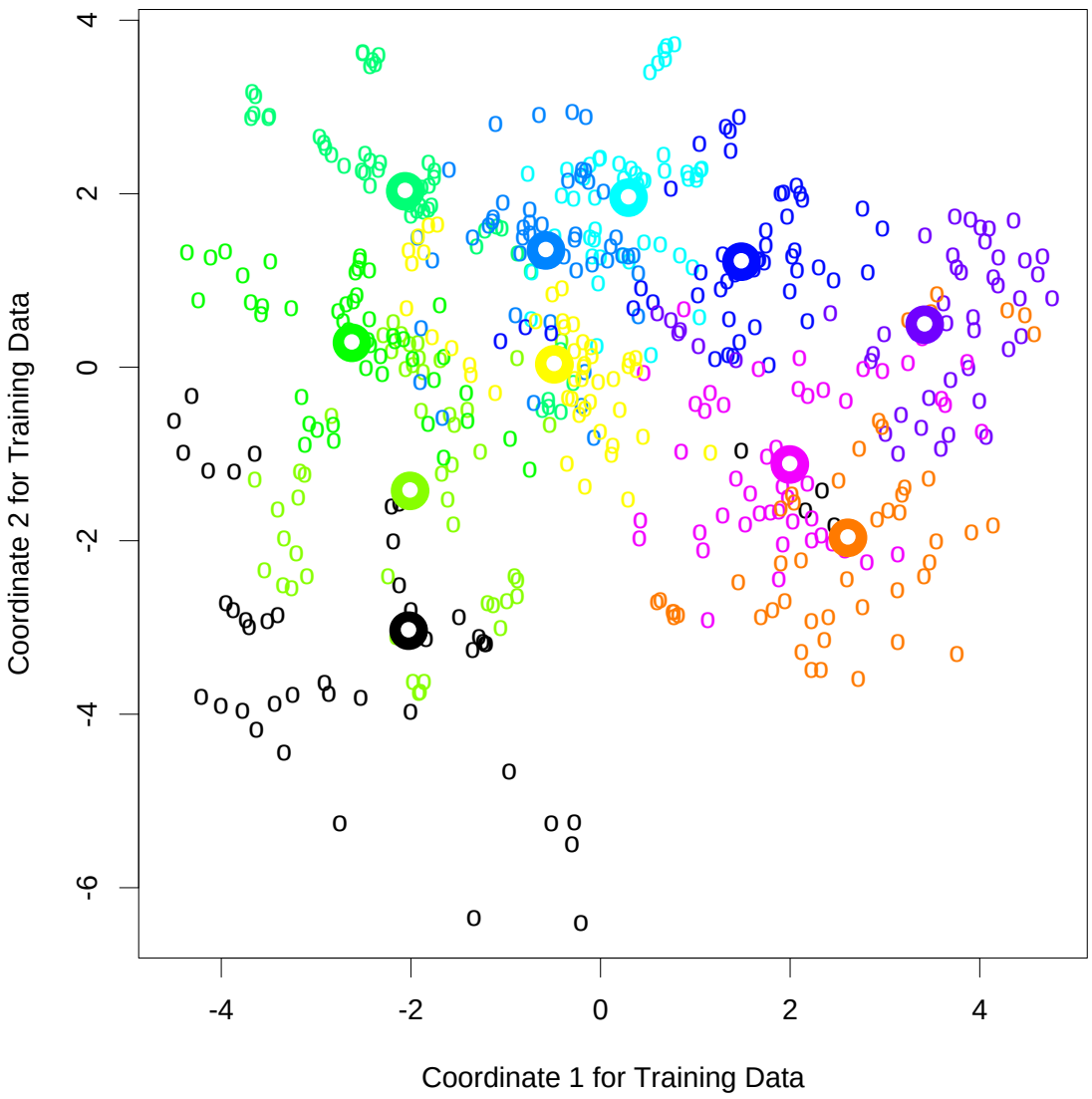
11 symbols, 8 speakers (train), 7 speakers (test), 6 replications each. 10 inputs features based on digitized utterances.

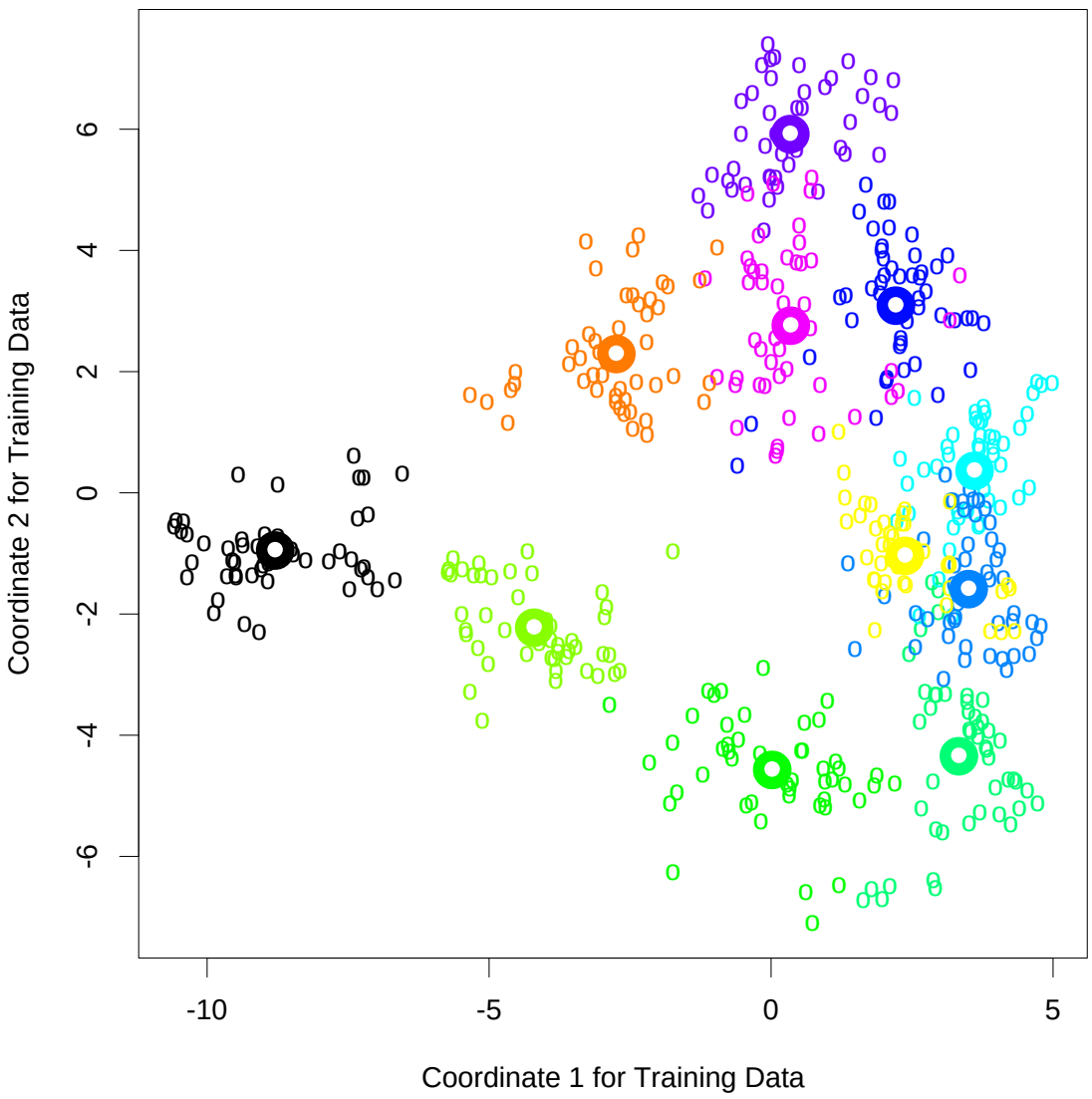
Source: Tony Robinson, via Scott Falman, CMU

GOAL: Predict Vowels

Some Results

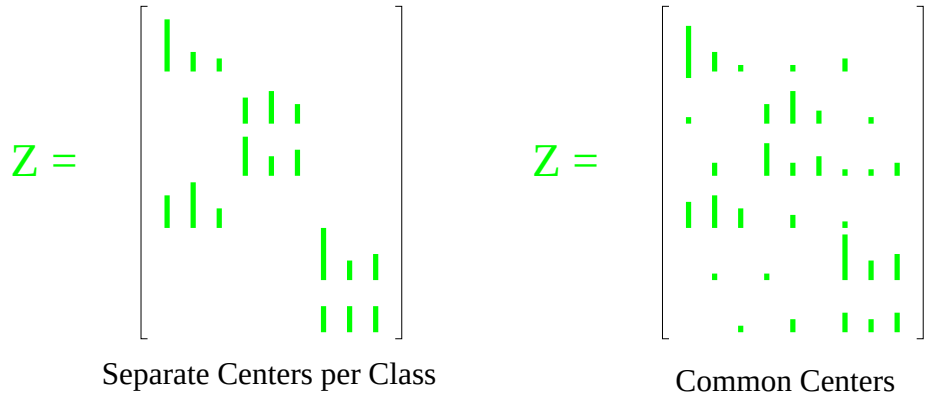
	Technique	Error rates	
		Training	Test
(1)	LDA	0.32	0.56
	Softmax	0.48	0.67
(2)	QDA	0.01	0.53
(3)	CART	0.05	0.56
(4)	CART (linear combination splits)	0.05	0.54
(5)	Single-layer Perceptron		0.67
(6)	Multi-layer Perceptron (88 hidden units)		0.49
(7)	Gaussian Node Network (528 hidden units)		0.45
(8)	Nearest Neighbor		0.44
(9)	FDA/BRUTO	0.06	0.44
	Softmax	0.11	0.50
(10)	FDA/MARS (degree = 1)	0.09	0.45
	Best reduced dimension (=2)	0.18	0.42
	Softmax	0.14	0.48
(11)	FDA/MARS (degree = 2)	0.02	0.42
	Best reduced dimension (=6)	0.13	0.39
	Softmax	0.10	0.50


Linear Discriminant Analysis
Coordinate 2 for Training Data
Coordinate 1 for Training Data
4
2
0
-2
-4
-6
-4
-2
0
2
4


Flexible Discriminant Analysis
Coordinate 2 for Training Data
6
4
2
0
-2
-4
-6
-10
-5
0
5
Coordinate 1 for Training Data

List of Extensions

- (Reduced Rank) LDA $\rightarrow$ (reduced rank) FDA via flexible regression: $\hat{Y} = S_X Y$
- (Reduced rank) LDA $\rightarrow$ (reduced rank) PDA (Penalized Discriminant Analysis) via penalized regression $\hat{Y} = S_\Omega Y = [X(X^T X + \lambda\Omega)^{-1} X^T]Y$, e.g. for image and signal classification.
- (Reduced rank) Mixture models. Each class a mixture of Gaussians. Each iteration of EM is a special form of FDA/PDA: $\hat{Z} = SZ$ where Z is a random response matrix.
- In the mixture model above the classes can share centers (radial basis functions) or own separate ones.



Separate Centers per Class

Common Centers

- In the above we use full likelihood for training: $P(X, G) = P(G|X)P(X)$; what if we use the conditional likelihood $P(G|X)$?

  [+ LDA becomes multinomial regression, and the non-parametric versions likewise.

  [+ The mixture problems again result in an E-M algorithm, where each M-step is a multinomial regression with random response Z.

References:

Breiman and Ihaka (1984) Unpublished manuscript

Campbell (1980) Applied Statistics

Kiiveri (1982) Technometrics

Ripley and Hjort (1995) Monograph in preparation

Ahmad and Tresp (1994) papers on RBFs

Bishop (1995) Monograph preprint

Details of LDA

LDA is Bayes rule for $P(X|G)$ Gaussian with density

$$\phi(X;\mu_j,\Sigma) = \frac{1}{(2\pi^p|\Sigma|)^{\frac{1}{2}}} e^{-\frac{1}{2}(X-\mu_j)^T\Sigma^{-1}(X-\mu_j)}$$

$$\begin{aligned} P(j|x) &= \frac{\phi(x;\mu_j,\Sigma)\Pi_j}{\sum_\ell \phi(x;\mu_\ell,\Sigma)\Pi_\ell} \\ &\sim \exp\left(x^T\Sigma^{-1}\mu_j - \frac{1}{2}\mu_j^T\Sigma^{-1}\mu_j + \log\Pi_j\right) \\ &= \exp(x^T\beta_j + \alpha_j) \end{aligned}$$

Note:

$$\max_{\text{rank}\{\mu_j\}=K;\Sigma} \sum_{j=1}^{J}\sum_{g_i=j} [\log\phi(x_i;\mu_j,\Sigma) + \log\Pi_j]$$

is equivalent to Fisher's rank-K LDA:

$$\max v_k^T B v_k \text{ subject to } v_k^T W v_k = 1,\ k = 1,\ldots,K$$

where B and W are the sample Between and Within covariance matrices. The latter is the usual formulation of Fisher's LDA.

LDA, FDA $\Longleftrightarrow$ Optimal Scoring

$$\sum_{i=1}^{n}[\theta(g_i) - \eta(x_i)]^2 + J(\eta) = \min$$

with normalization $\sum_i \theta^2(g_i) = 1$, and "roughness" penalty functional J.

$$\eta(x_i) = \begin{cases} x_i^T\beta \ \text{LDA} \\ f_1(x_{i1}) + f_2(x_{i2}) + \ldots + f_p(x_{ip}) \\ \text{MARS}(x_i) \\ \text{NN}(x_i) \\ \vdots \end{cases}$$

Often $\eta(x) = \sum_m h_m(x)\beta_m$ and $J(\eta) = \beta^T\Omega\beta$.

There is a 1-1 correspondence between optimal scoring and Fishers LDA, as well as the flexible extensions.

Optimal scoring algorithm for LDA/FDA

Indicator Response Matrix Y

$$
\begin{array}{c} \\ g_1 = 2 \\ g_2 = 1 \\ g_3 = 1 \\ g_4 = 5 \\ g_5 = 4 \\ \vdots \\ g_n = 3 \end{array}
\begin{array}{c}
\begin{array}{ccccc} C_1 & C_2 & C_3 & C_4 & C_5 \end{array} \\
\left(\begin{array}{ccccc}
0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 0 \\
 & & \vdots & & \\
0 & 0 & 1 & 0 & 0
\end{array} \right)
\end{array}
$$

$$
\begin{aligned}
\hat{Y} &= SY \\
Y^T \hat{Y} &= \Theta D \Theta^T
\end{aligned}
$$

where $S =$ linear regression, additive regression, MARS, NN,. . . , each giving a different version of FDA.

FDA and Penalized Discriminant Analysis

The steps in FDA are

- Enlarge the set of predictors X via a basis expansion $h(X)$, and hence inject us into a higher dimensional space.
- Use (penalized) LDA in the enlarged space, where the penalized Mahalanobis distance is given by

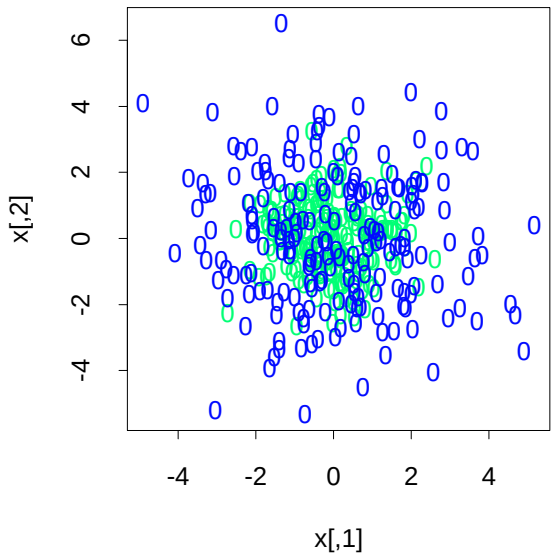
$$D(x,\mu) = (h(x)-h(\mu))^T(\Sigma_W+\Omega)^{-1}(h(x)-h(\mu))$$

  Σ_W is defined in terms of bases functions $h(x_i)$.
- Decompose the classification subspace using a penalized metric:

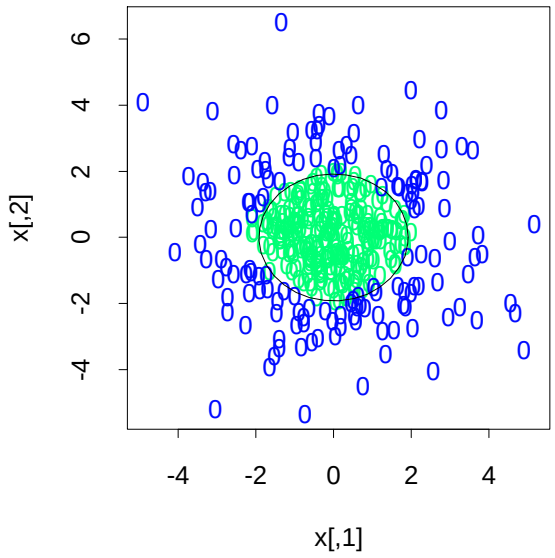
$$\max \text{tr}(U^T\Sigma_{Bet}U) \text{ subject to } U^T(\Sigma_W+\Omega)U = I$$

Skin of the Orange

Training Data



Predicted Classes



FDA vs Regression

In FDA algorithm, we decompose

$$Y^T S(\lambda) Y = Y^T \hat{Y}$$

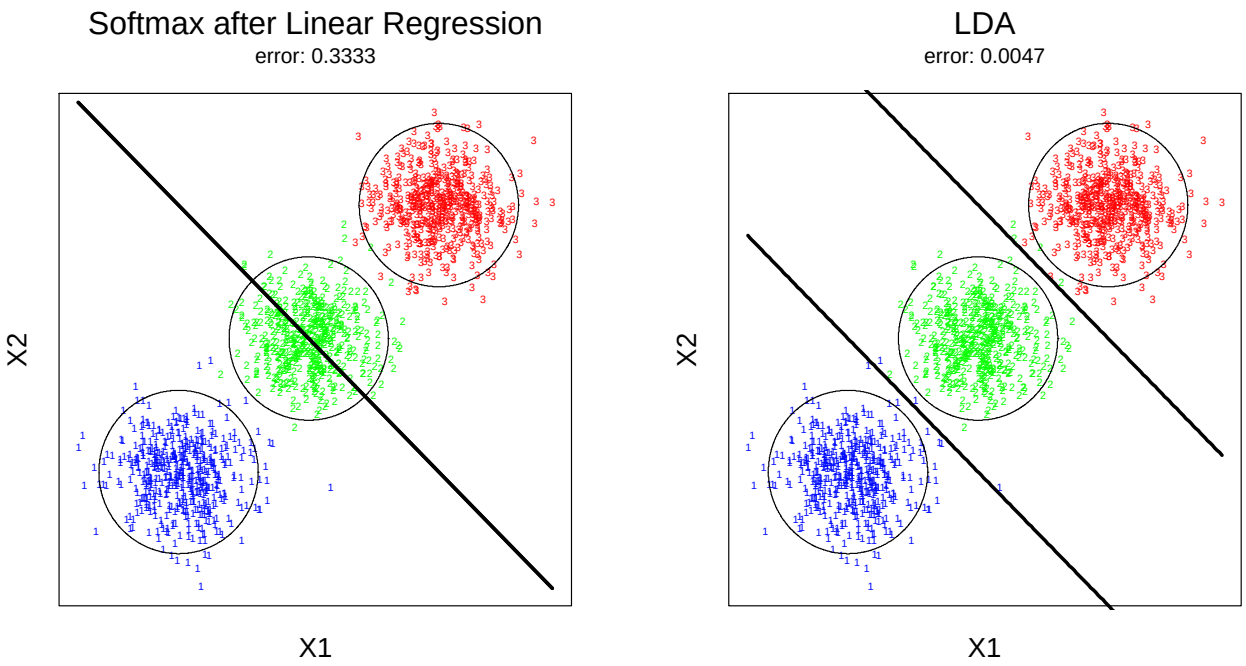
$\hat{Y}$ fitted values for dummy response matrix

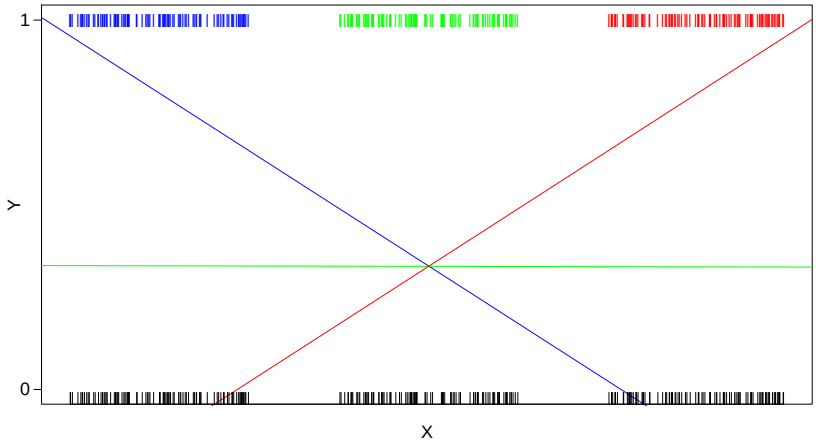
Why not stop at $\hat{Y} \approx E(Y|X)$?

i.e. for new x, compute

$$\hat{y}_1(x), \hat{y}_2(x), \ldots, \hat{y}_J(x)$$

and assign x to the class j with the largest $\hat{y}_j(x)$.





With many classes, high order (polynomial) interaction terms might be needed to avoid masking.

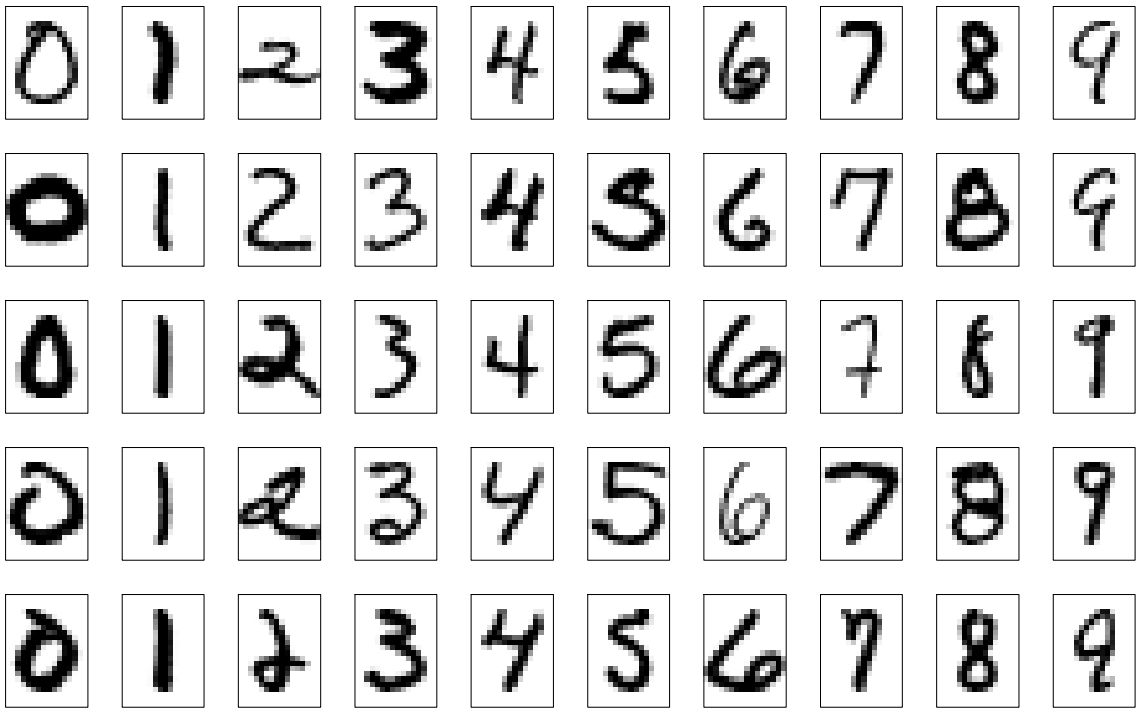
FDA vs PDA

There are two complimentary situations:

FDA $h(x)$ is an expansion of x into $m \gg p$ basis functions.
$\beta^T \Omega \beta$ makes $\eta(x) = \beta^T h(x)$ smooth in x.

PDA $h(x) = x$, e.g. digitization of an analog signal—$x = (x_1, \ldots, x_p)$, $x_j = x(t_j)$.
$\beta^T \Omega \beta$ makes β smooth in t

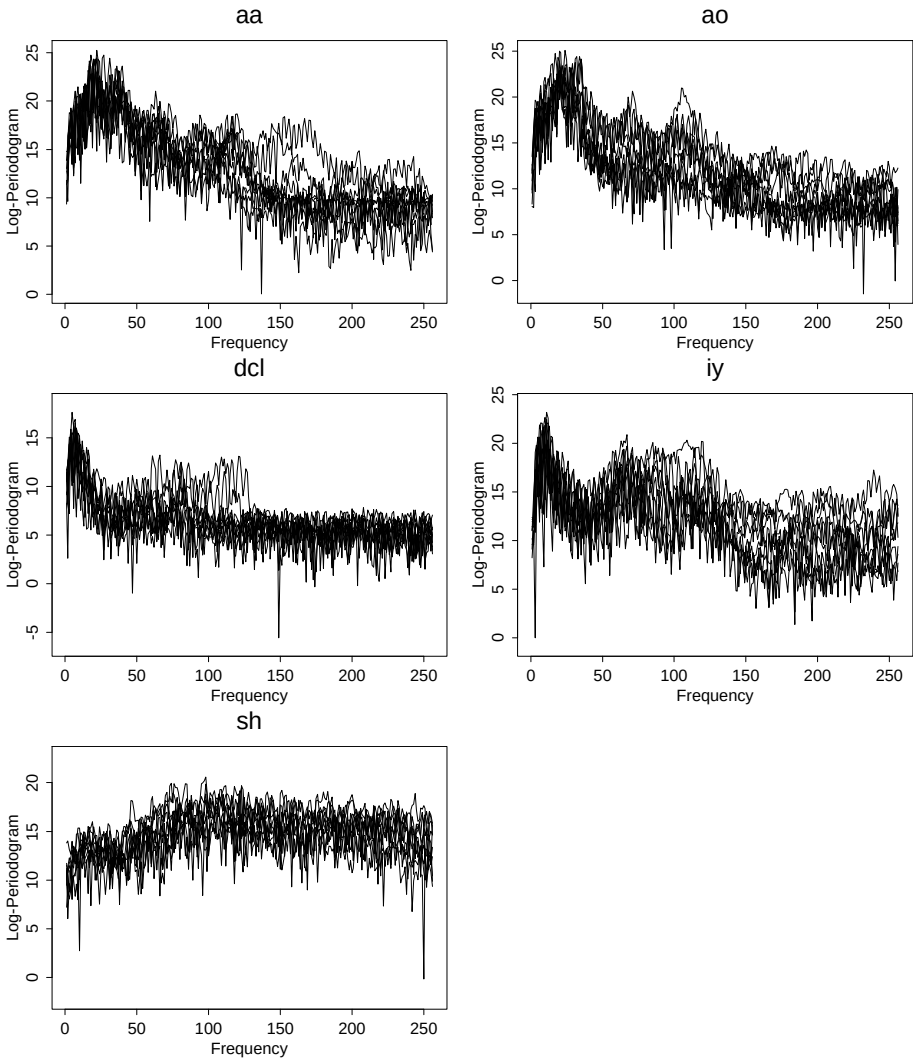
Handwritten Digit Identification



A sample of 5 handwritten digits within each digit class.

2000 training and 2000 test images.

Speech Recognition



A sample of 10 log-periodograms within each phoneme class. 1633 training (11 speakers) and 1661 test (12 different speakers) frames.

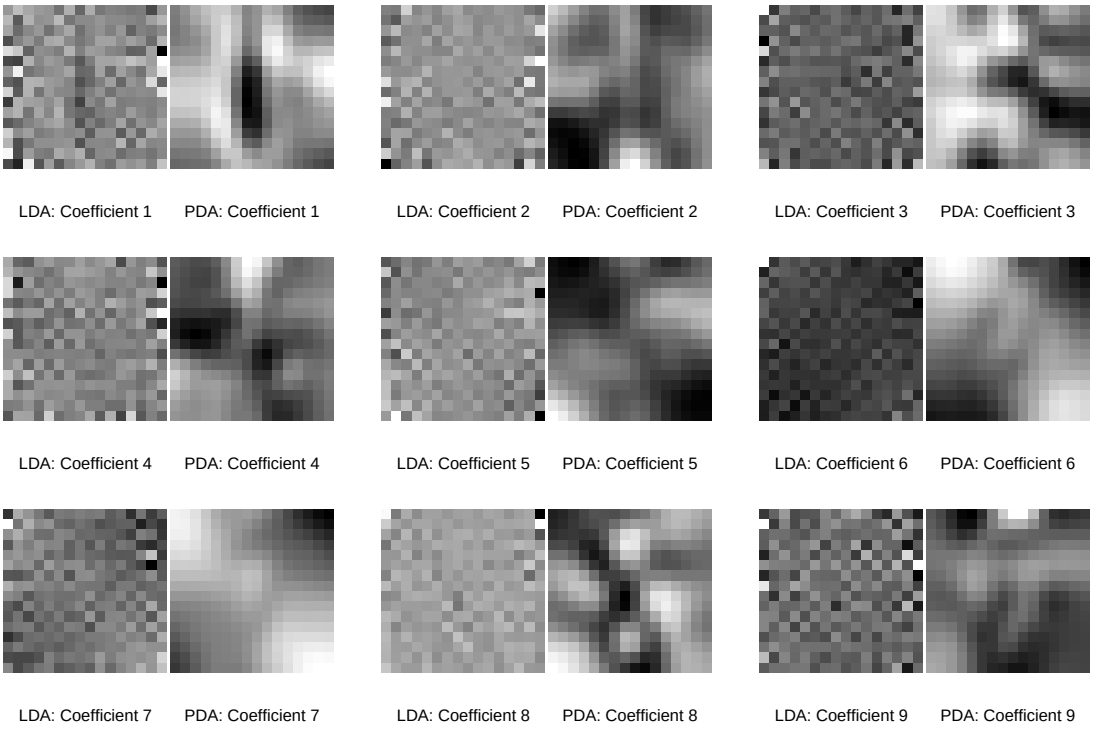
Reasons for Regularization

There are two different reasons why regularization is necessary:

Estimation: With digitized analog signals, the variables-to-observation ratio can be small. Regularization is needed for consistency in estimating a population LDA model [Leurgans, Moyeed & Silverman, JRSSB, 1993]

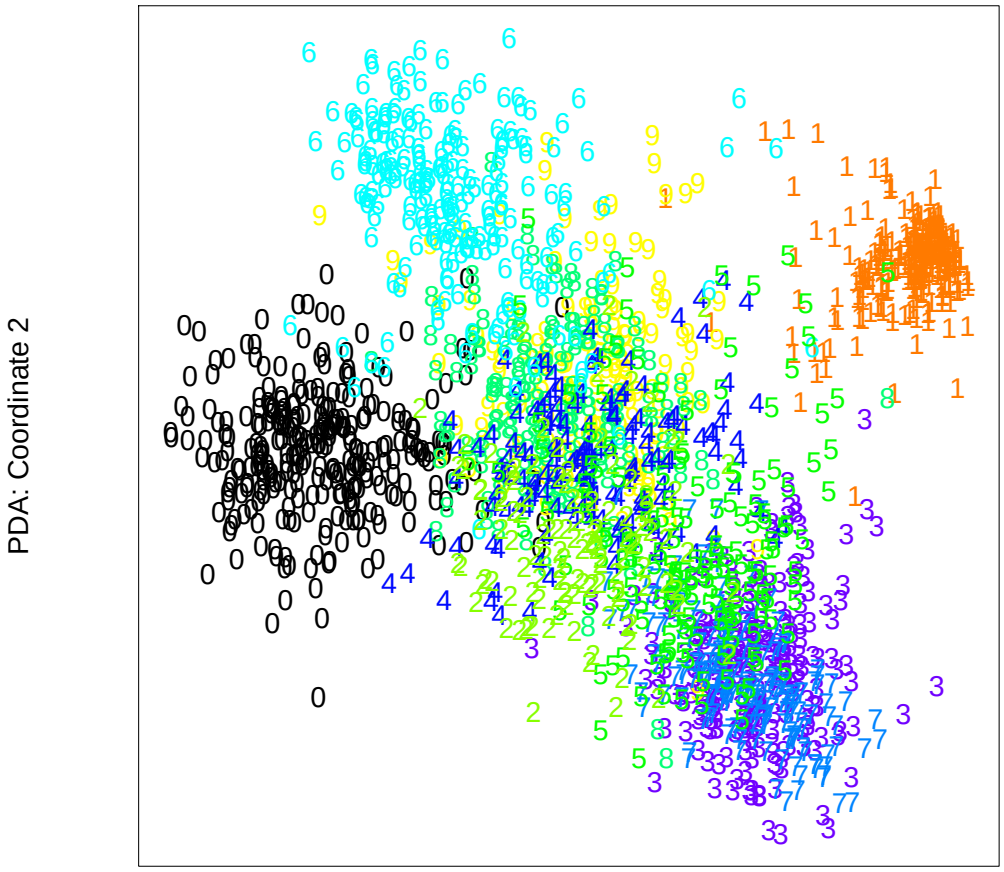
Interpretation: The LDA coefficients are given by $\beta = \Sigma^{-1}\mu_j$. If the eigenstructure of Σ_W concentrates on low-frequency signals, then β will tend to be noisy, and hence hard to interpret.

PDA coefficients: Handwritten Digits

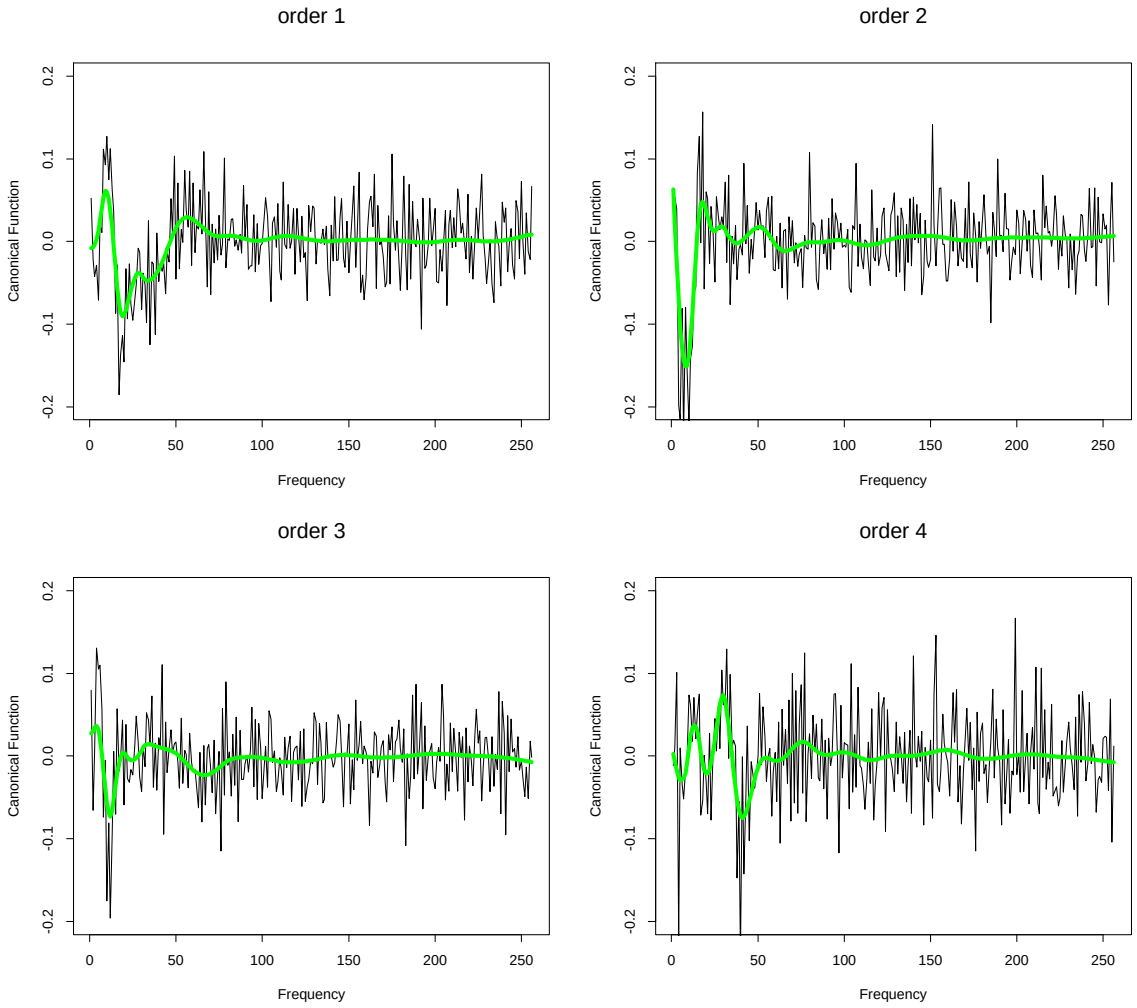


Left images: LDA coefficient image

Right images: PDA coefficient image


Canonical Variate Plot --- Digit Test Data
PDA: Coordinate 2
PDA: Coordinate 1

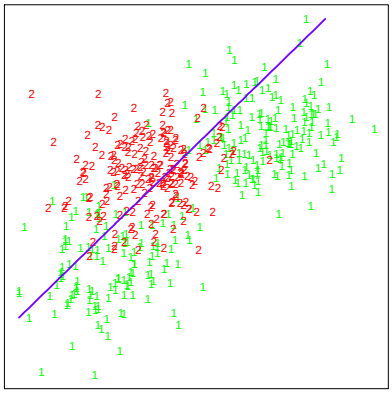
PDA coefficients: Phoneme Data



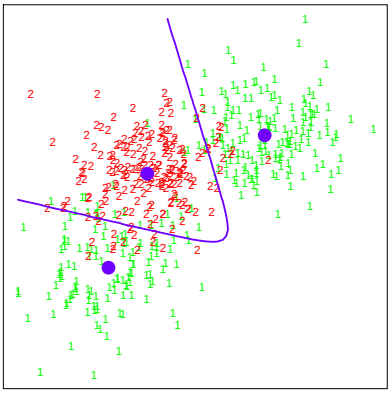
Ordinary LDA coefficient functions for the phoneme data, and regularized versions.

Mixture Discriminant Analysis: MDA

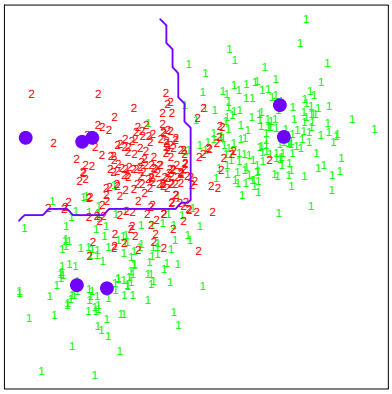
Linear Discriminant Analysis



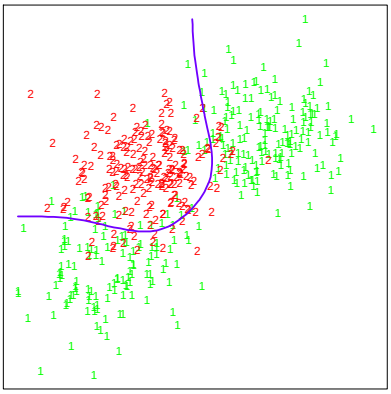
Mixture Discriminant Analysis



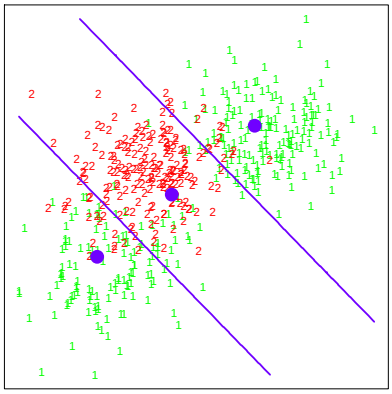
Learning Vector Quantization



Flexible Discriminant Analysis

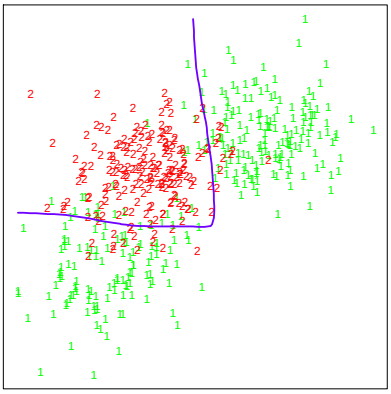


Mixture Discriminant Analysis



Rank 1 Model

MDA/FDA



Gaussian Mixture Model

$$P(X|G=j) = \sum_{r=1}^{R_j} \pi_{jr}\phi(X;\mu_{jr},\Sigma), \text{ Mixture of Gaussians}$$

Then

$$P(G=j|X=x) = \frac{\sum_{r=1}^{R_j} \pi_{jr}\phi(X;\mu_{jr},\Sigma)\Pi_j}{\sum_{\ell=1}^{J}\sum_{r=1}^{R_\ell} \pi_{\ell r}\phi(X;\mu_{\ell r},\Sigma)\Pi_\ell}$$

Estimate parameters by maximum likelihood of $P(X,G)$ (possibly subject to rank constraints!)

$$\max_{\text{rank}\{\mu_{jr}\}=K;\Sigma} \sum_{j=1}^{J}\sum_{g_i=j} \log\Big(\sum_{r=1}^{R_j} \pi_{jr}\phi(x_i;\mu_{jr},\Sigma)\Pi_j\Big)$$

Note: reduced rank amounts to dimension reduction in predictor space!

EM and Optimal Scoring

E-Step:

Compute memberships Prob$\big($obs $\in r$th subclass of class j $\big|x, j\big)$

$$W(c_r|x,j) = \frac{\pi_r \phi(x;\mu_{jr},\Sigma)}{\sum_{k=1}^{R_j} \pi_{jk}\phi(x;\mu_{jk},\Sigma)}$$

M-Step:

Construct Random Response Matrix Z with elements $W(c_{jr}|x,j)$:

$$\begin{array}{c} \\ g_1 = 2 \\ g_2 = 1 \\ g_3 = 1 \\ g_4 = 3 \\ g_5 = 2 \\ \vdots \\ g_n = 3 \end{array} \begin{array}{c} \begin{array}{ccccccccc} c_{11} & c_{12} & c_{13} & c_{21} & c_{22} & c_{23} & c_{31} & c_{32} & c_{33} \end{array} \\ \left(\begin{array}{ccccccccc} 0 & 0 & 0 & .3 & .5 & .2 & 0 & 0 & 0 \\ .9 & .1 & .0 & 0 & 0 & 0 & 0 & 0 & 0 \\ .1 & .8 & .1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .5 & .4 & .1 \\ 0 & 0 & 0 & .7 & .1 & .2 & 0 & 0 & 0 \\ & & & \vdots & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & .1 & .1 & .8 \end{array} \right) \end{array}$$

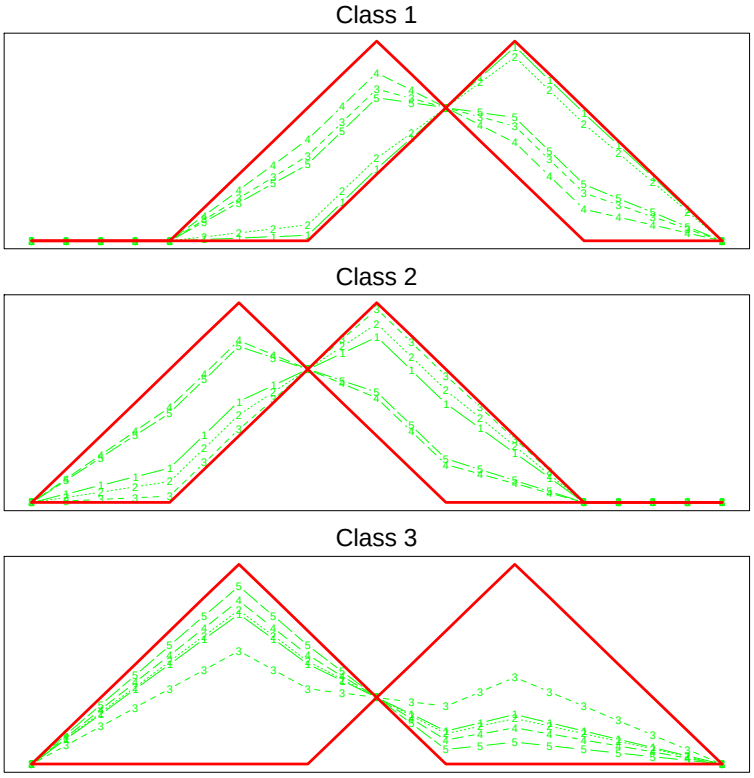
$$\left.\begin{array}{l} \hat{Z} = SZ \\ Z^T\hat{Z} = \Theta D \Theta^T \\ \text{Update } \pi\text{s and } \Pi\text{s} \end{array}\right\} \Leftrightarrow \text{M-step of MDA}$$

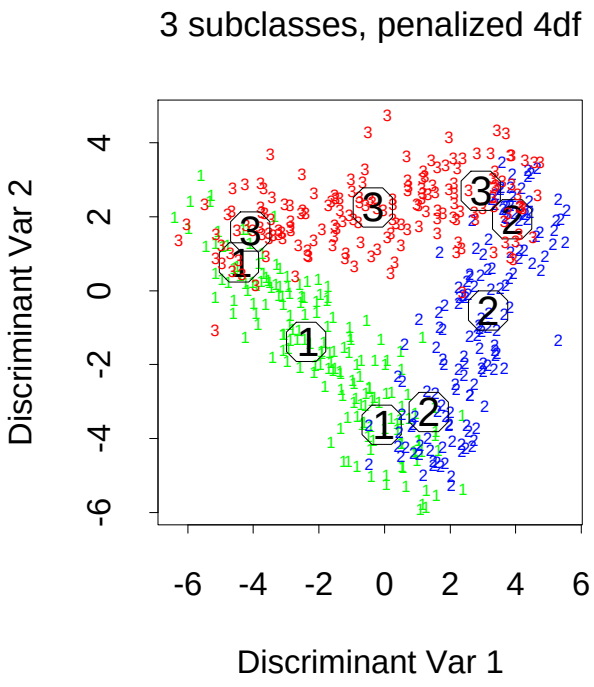
Waveform Example

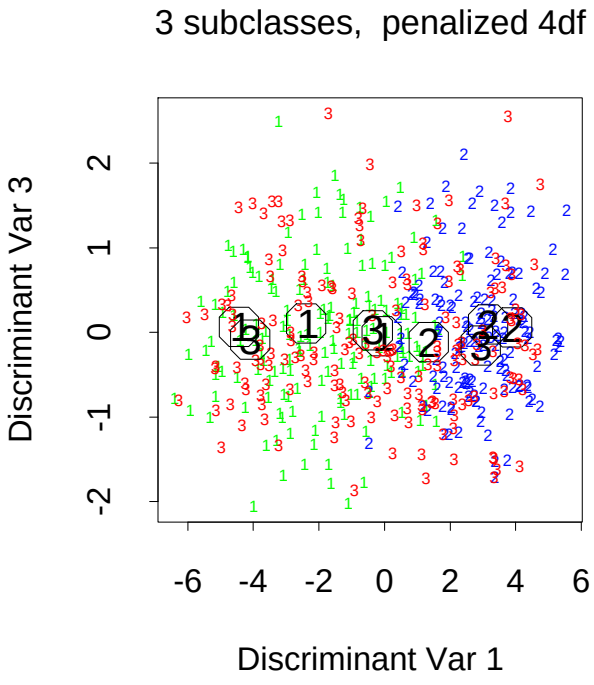
$$x_i = uh_1(i) + (1-u)h_2(i) + \epsilon_i \quad \text{Class 1}$$

$$x_i = uh_1(i) + (1-u)h_3(i) + \epsilon_i \quad \text{Class 2}$$

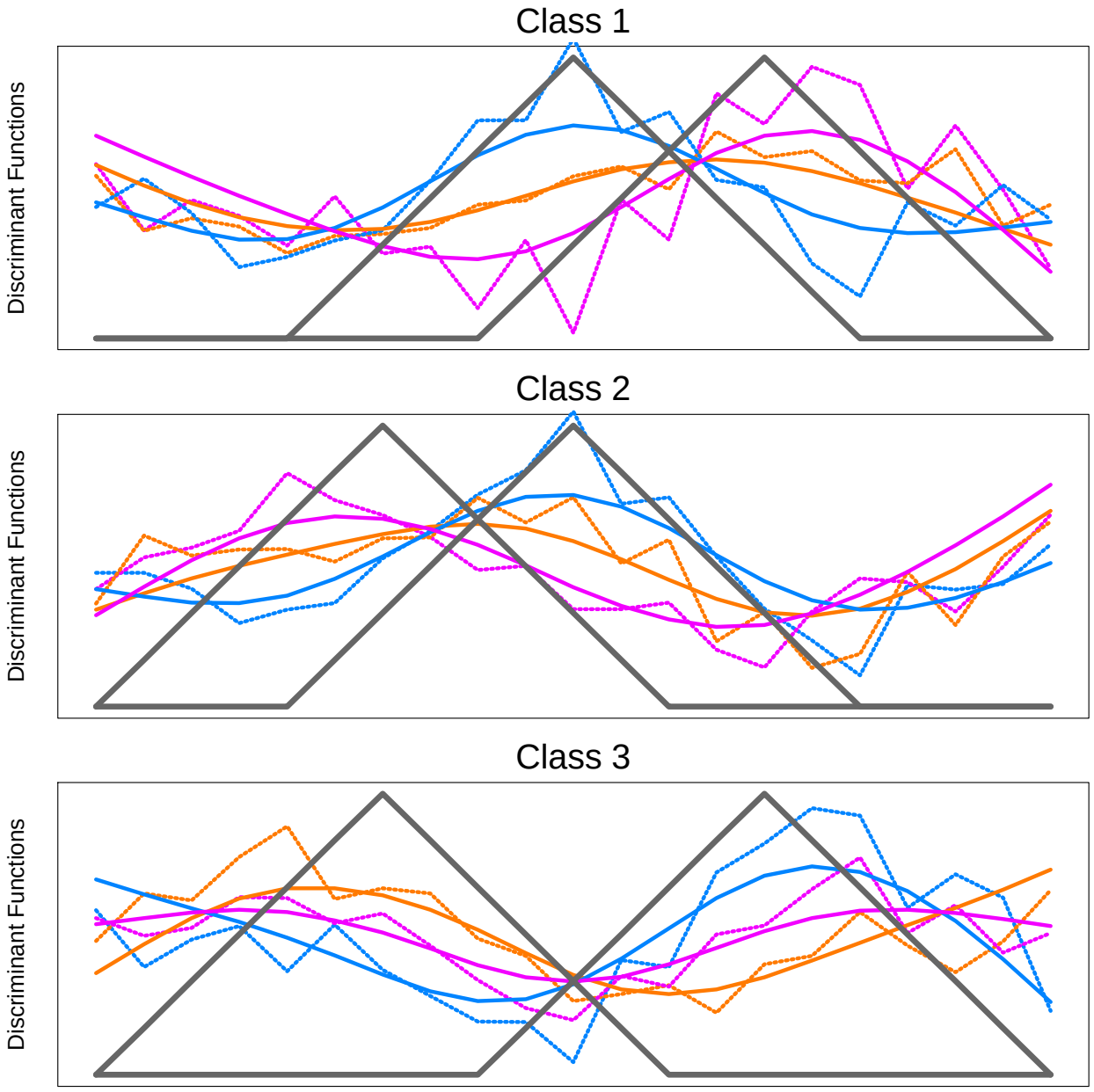
$$x_i = uh_2(i) + (1-u)h_3(i) + \epsilon_i \quad \text{Class 3}$$




3 subclasses, penalized 4df
Discriminant Var 2
Discriminant Var 1
-6
-4
-2
0
2
4
6


3 subclasses, penalized 4df
Discriminant Var 3
Discriminant Var 1
-2
-1
0
1
2
-6
-4
-2
0
2
4
6

Penalized MDA coefficients



Simulation Results for Waveform Example

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)
PDA (penalized 4df)	0.150(.005)	0.171(.005)

Gaussian Mixtures and RBFs

$$P(X, G = j) = \sum_{r=1}^{R} \pi_r \phi(X; \mu_r, \Sigma) P_r(j)$$

This is a mixture of joint-densities $P_r(X, G)$ with R shared mixture components. Then

$$P(G = j | X = x) = \frac{\sum_{r=1}^{R} \pi_r \phi(x; \mu_r, \Sigma) P_r(j)}{\sum_{r=1}^{R} \pi_r \phi(x; \mu_r, \Sigma)}$$

This closely resembles (renormalized) RBFs.

Estimate parameters by maximum likelihood of $P(X, G)$ (possibly subject to rank constraints!)

$$\max_{\text{rank}\{\mu_r\}=K;\Sigma} \sum_{j=1}^{J} \sum_{g_i=j} \log(\sum_{r=1}^{R} \pi_r \phi(x_i; \mu_r, \Sigma) P_r(j)$$

E-M algorithm again yields optimal scoring $\hat{Z} = SZ$, where now Z is the full random response matrix described earlier.

EM and Optimal Scoring

E-Step:

Compute memberships Prob$(\text{obs} \in r\text{th subclass} \,|x, j)$

$$W(c_r|x,j) = \frac{\pi_r \phi(x;\mu_r,\Sigma) P_r(j)}{\sum_{k=1}^{R} \pi_k \phi(x;\mu_k,\Sigma) P_k(j)}$$

i.e. membership depends on $||x-\mu_k||^2_\Sigma - 2\log P_k(j)$.

M-Step:

Construct Random Response Matrix Z with elements $W(c_r|x,j)$:

$$\begin{array}{c} \\ g_1 = 2 \\ g_2 = 1 \\ g_3 = 1 \\ g_4 = 3 \\ \vdots \\ g_n = 3 \end{array} \begin{array}{c} \begin{array}{ccccccccc} c_1 & c_2 & c_3 & c_4 & c_5 & c_6 & c_7 & c_8 & c_9 \end{array} \\ \begin{pmatrix} 0 & .1 & 0 & .2 & .4 & .2 & 0 & .1 & 0 \\ .8 & .1 & .0 & 0 & .1 & 0 & 0 & 0 & 0 \\ .1 & .6 & .1 & 0 & 0 & 0 & 0 & .1 & .1 \\ 0 & 0 & 0 & 0 & 0 & 0 & .5 & .4 & .1 \\ & & & \vdots & & & & & \\ 0 & 0 & 0 & .1 & 0 & 0 & 0 & .1 & .8 \end{pmatrix} \end{array}$$

$$\left.\begin{array}{l} \hat{Z} = SZ \\ Z^T\hat{Z} = \Theta D \Theta^T \\ \text{Update } \pi\text{s and } \Pi\text{s} \end{array}\right\} \Leftrightarrow \text{M-step of MDA}$$