Linear Discriminant Analysis (LDA)
&
Quadratic Discriminant Analysis (QDA)
&
Regularized Discriminant Analysis (RDA)

Statistical Data Mining I

Rachael Hageman Blair

Recap

Notation

- We need to know the class of posteriors to find the optimal classifier:

$$G^*(x) = \arg\max_{k \in G} \Pr(G = k \mid X = x).$$

Linear Discriminant Analysis

Notation

- The prior probability of class *k* is $\pi_k, \sum_{k=1}^{K} \pi_k = 1.$
  - π_k is usually estimated simple by empirical frequencies of the training set:

$$\hat{\pi}_k = \frac{\#\text{ of samples in class k}}{\text{Total \# of samples}}.$$

- The class-conditional density of X in class $G = k$ is $f_k(x).$
- Compute the posterior probability (Bayes' theorem):

$$\Pr(G = k \mid X = x) = \frac{f_k(x)\pi_k}{\sum_{i=1}^{K} f_i(x)\pi_i}.$$

- By MAP (Bayes Rule for 0-1 loss):

$$\hat{G}(x) = \arg\max_k \Pr(G = k \mid X = x)$$
$$= \arg\max_k f_k(x)\pi_k.$$

Linear Discriminant Analysis

Suppose each class is modeled by a multivariate Gaussian:

$$f_{k(x)} = \frac{1}{(2\pi)^{p/2}\left|\Sigma_k\right|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}.$$

Linear Discriminant Analysis (LDA) arises when we assume all classes have a common covariance $\Sigma_k = \Sigma \ \ \forall k$.

Linear Discriminant Analysis

Optimal Classification

$$\hat{G}(x) = \arg\max_k \Pr(g = k \mid X = x)$$

$$= \arg\max_k f_k(x)\pi_k$$

$$= \arg\max_k \log(f_k(x)\pi_k)$$

$$= \arg\max_k \left[-\log\left((2\pi)^{p/2} \left| \Sigma^{1/2} \right| \right) - \frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k) + \log(\pi_k) \right]$$

Note: $$-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k - \frac{1}{2}x^T\Sigma^{-1}x$$

Linear Discriminant Analysis

- The optimal classifier:

$$\hat{G}(x) = \arg\max_k \left[x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k) \right].$$

- Define the Linear Discriminant Function:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k).$$

then

$$\hat{G}(x) = \arg\max_k \delta_k(x).$$

Linear Discriminant Analysis

- The decision boundary between class *k* and *l* is:

$$\left\{x : \delta_k(x) = \delta_l(x)\right\}.$$

- Equivalently the following holds:

$$\log\frac{\Pr\left(G = k \mid X = x\right)}{\Pr(G = l \mid X = x)} = \log\frac{f_k(x)}{f_l(x)} + \log\frac{\pi_k}{\pi_l}$$

$$= \log\frac{\pi_k}{\pi_l} - \frac{1}{2}\left(\mu_k + \mu_l\right)^T \Sigma^{-1}\left(\mu_k - \mu_l\right) + x^T\Sigma^{-1}\left(\mu_k - \mu_l\right).$$

which implies that the decision boundary is where:

$$\Pr\left(G = k \mid X = x\right) = \Pr(G = l \mid X = x).$$

Linear Discriminant Analysis

In practice, we do not know the parameters for the Gaussian distributions, and have to estimate them.

This includes:

$$\hat{\pi}_k = N_k / N, \quad \text{where } N_k \text{ is the number of class } k \text{ observations.}$$

$$\hat{\mu}_k = \sum_{g_i=k} x_i / N_k$$

$$\hat{\Sigma} = \sum_{k=1}^{K} \sum_{g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / N - K.$$

Linear Discriminant Analysis

Relationship to LS Classification:

When you have two classes only, the coefficient vector from LS is proportional to the LDA direction.

Not the case when more than two classes…. LDA actually avoids the masking problem.

(Exercise 4.2)

Linear Discriminant Analysis

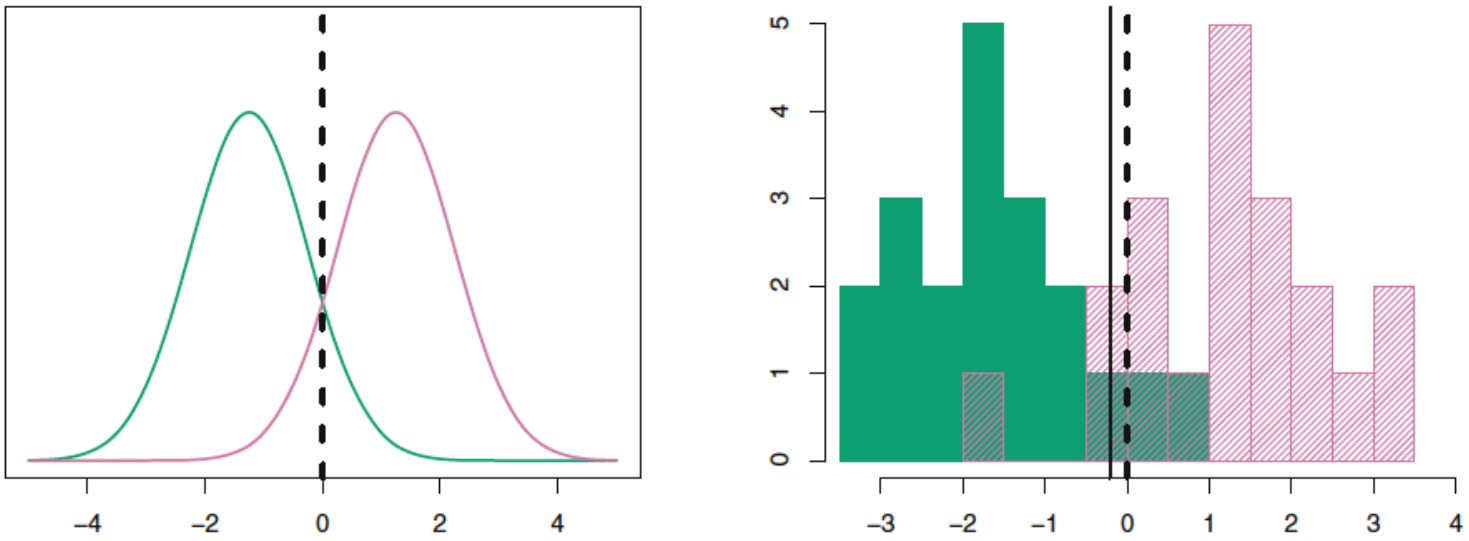


FIGURE 4.4. Left: *Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary.* Right: *20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line. The solid vertical line represents the LDA decision boundary estimated from the training data.*

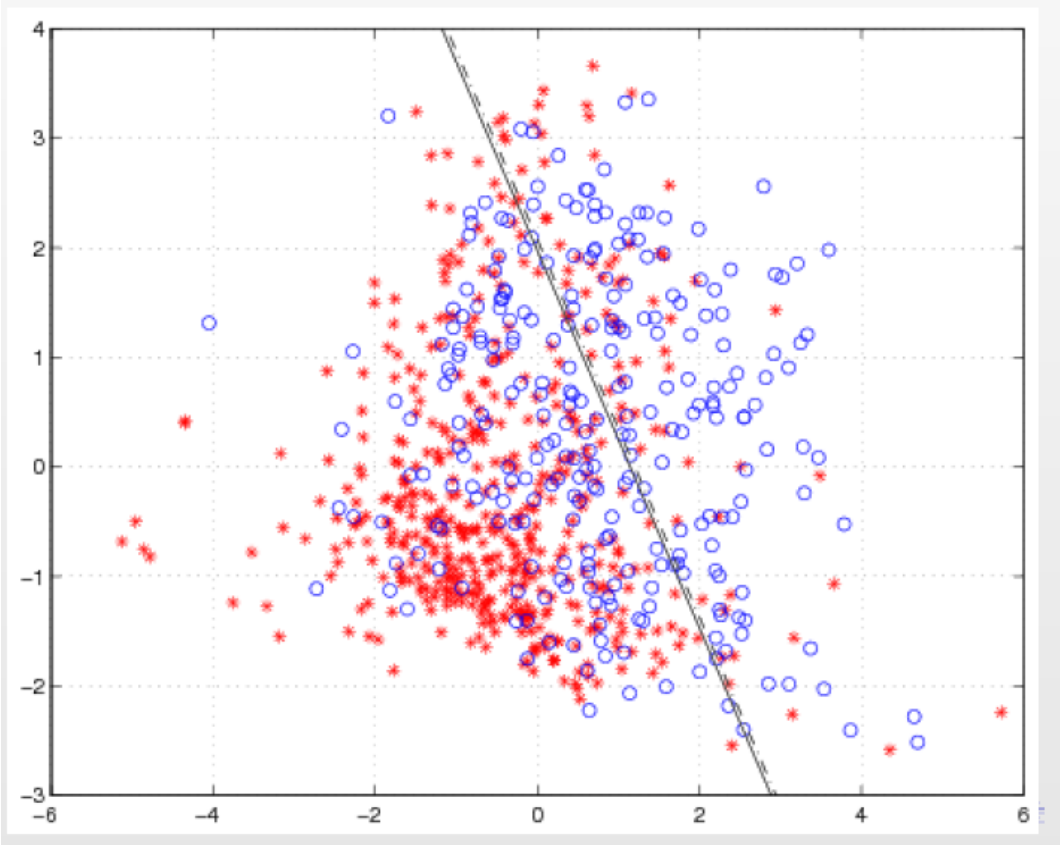
Linear Discriminant Analysis

Diabetes Data Set

- Two input variables computed from principal components of the original 8 variables.
- Prior probabilities: $\hat{\pi}_1 = 0.651,\ \hat{\pi}_1 = 0.349.$
- $\hat{\mu}_1 = (-0.4035, -0.1935)^T,\ \hat{\mu}_1 = (0.7528, 0.3611)^T.$
- $\hat{\Sigma} = \begin{pmatrix} 1.7925 & -0.1461 \\ -0.1461 & 1.6634 \end{pmatrix}.$

Linear Discriminant Analysis

Diabetes Data Set



Linear Discriminant Analysis

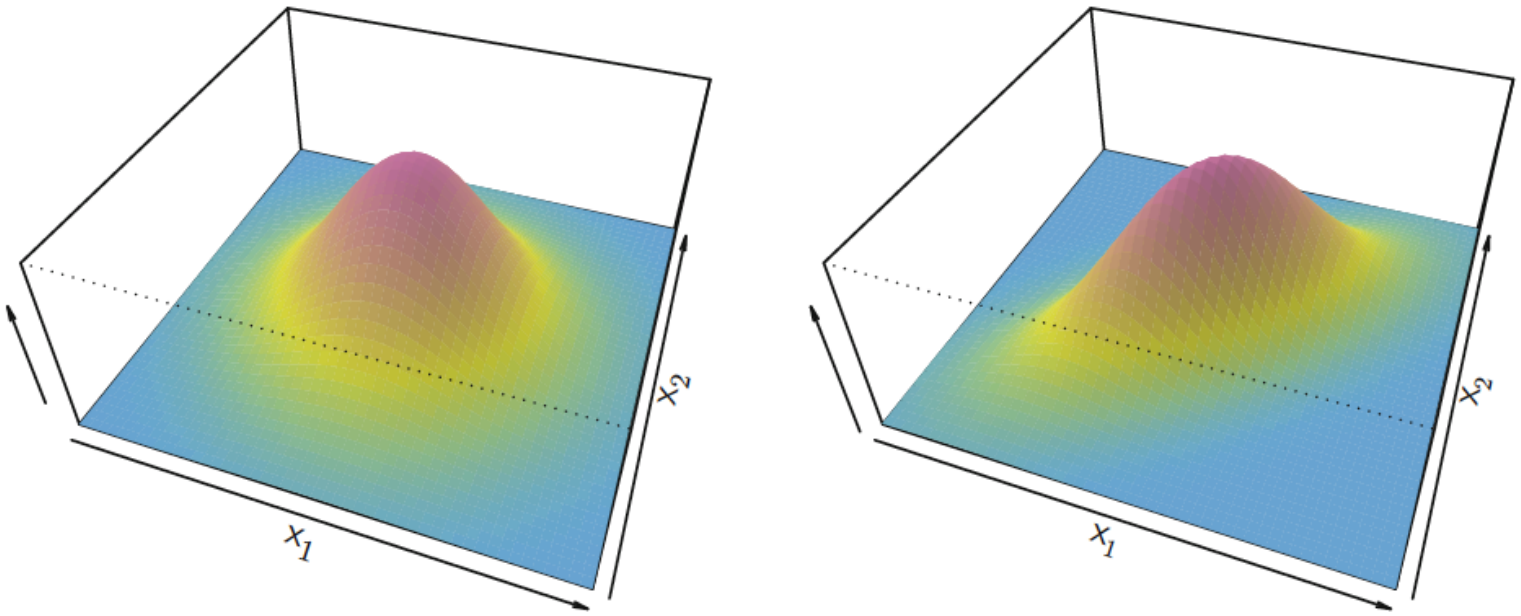


FIGURE 4.5. *Two multivariate Gaussian density functions are shown, with* $p = 2$. Left: *The two predictors are uncorrelated.* Right: *The two variables have a correlation of* 0.7.

Quadratic Discriminant Analysis

- Estimate the covariance matrix k separately for each class, $k = 1, 2, \ldots, K.$ (can be expensive!)
- Quadratic discriminant function:

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) + \log(\pi_k).$$

- Classification Rule:

$$\hat{G}(x) = \arg\max_k \delta_k(x).$$

- Decision boundaries are quadratic equations in x.
- QDA fits the data better than LDA, but there are more parameters to estimate.

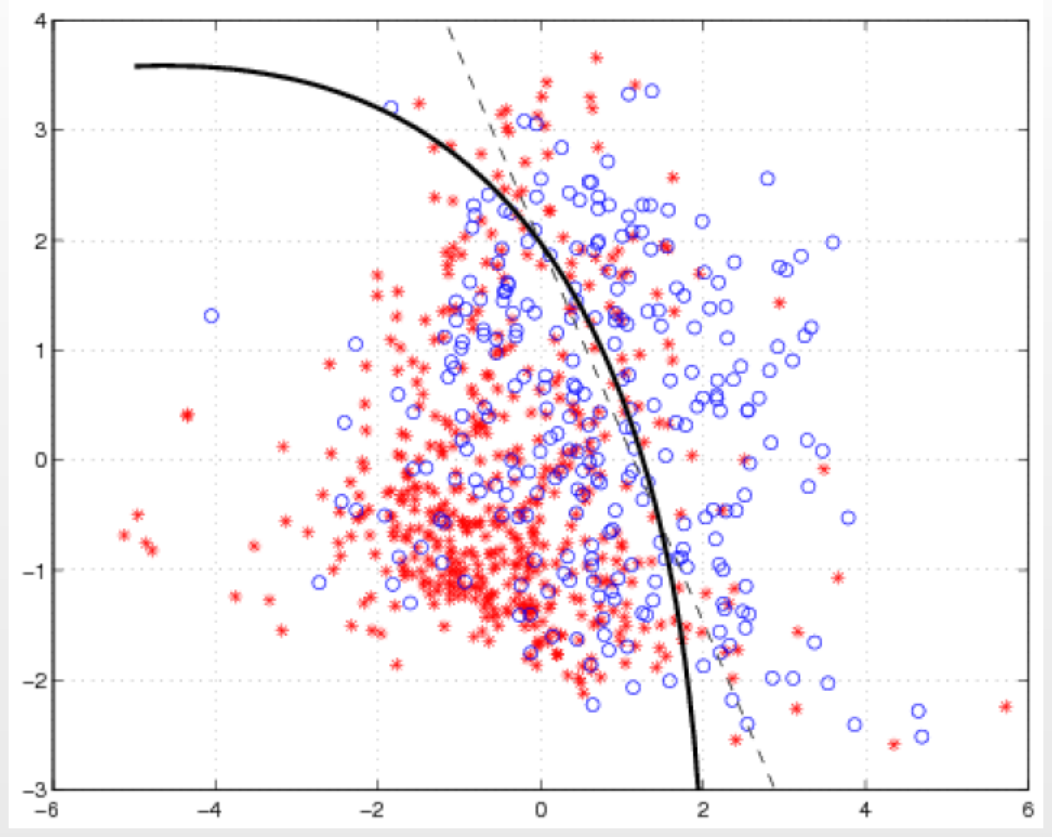
Quadratic Discriminant Analysis

Diabetes Data Set

- Prior probabilities: $\hat{\pi}_1 = 0.651,\ \hat{\pi}_1 = 0.349.$
- $\hat{\mu}_1 = (-0.4035, -0.1935)^T,\ \hat{\mu}_1 = (0.7528, 0.3611)^T.$
- $\hat{\Sigma}_1 = \begin{pmatrix} 1.6769 & -0.0461 \\ -0.0461 & 1.5964 \end{pmatrix},\ \hat{\Sigma}_2 = \begin{pmatrix} 2.0087 & -0.3330 \\ -0.3330 & 1.7887 \end{pmatrix}.$

Quadratic Discriminant Analysis

Diabetes Data Set – sensitivity is the same as LDA, but the specificity is slightly lower.



Quadratic Discriminant Analysis

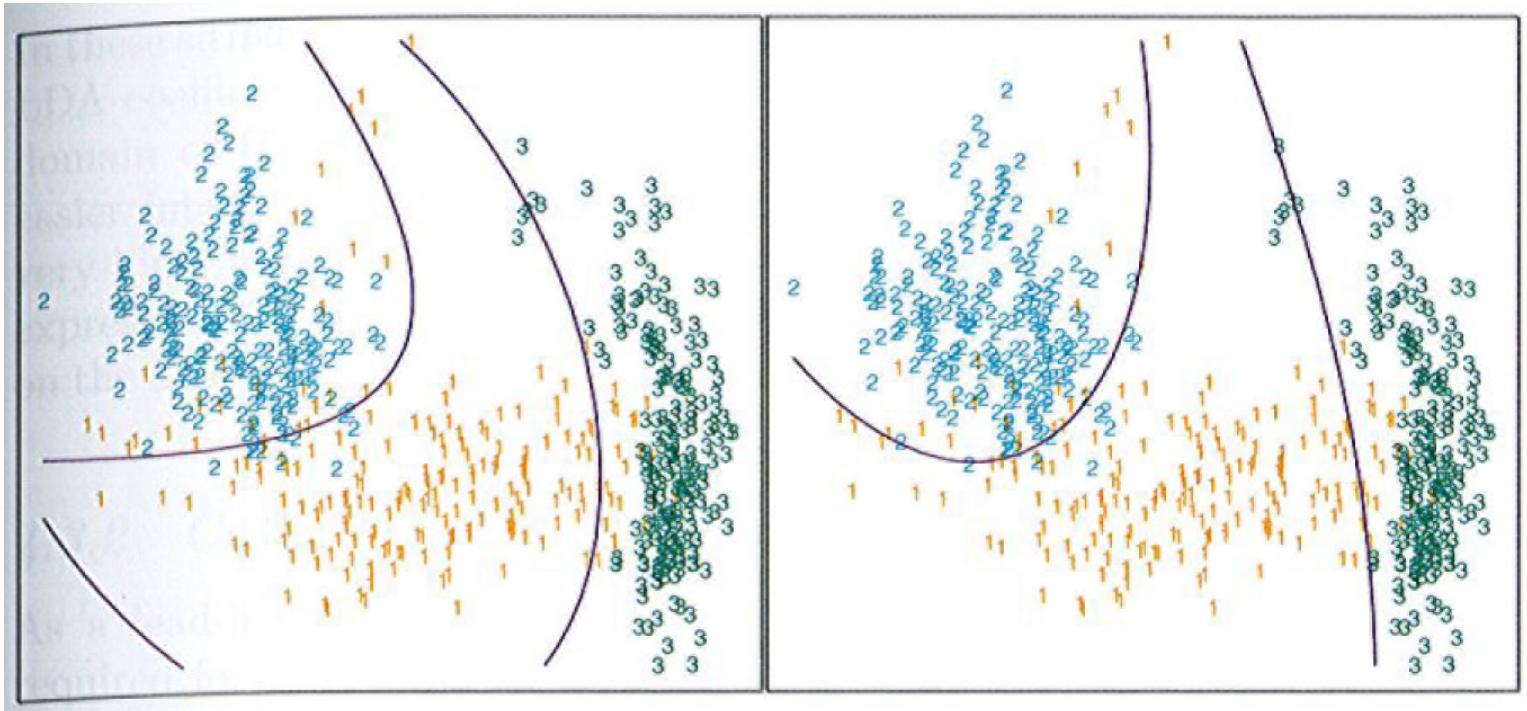


FIGURE 4.6. *Two methods for fitting quadratic boundaries. The left plot shows the quadratic decision boundaries for the data in Figure 4.1 (obtained using LDA in the five-dimensional space $X_1, X_2, X_1X_2, X_1^2, X_2^2$). The right plot shows the quadratic decision boundaries found by QDA. The differences are small, as is usually the case.*

LDA and QDA

- Often both perform well, with marginal differences.

Why?

- Probably NOT b/c the data is approximately Gaussian with equal covariance.
- Most likely, the data can only support simple decision boundaries, and the estimates provided are stable.

Analogous to the bias variance tradeoff – we can put up with the bias of a linear decision boundary because it can be estimated with much lower variance.

Regularized Discriminant Analysis

- A compromise between LDA and QDA.
- Shrink the separate covariance matrices for QDA toward a common covariance as in LDA.

- Regularized covariance matrices:

$$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}.$$

- The quadratic discriminant function $\delta_k(x)$ is defined using the shrunken covariance matrices $\hat{\Sigma}_k(\alpha)$.
- The parameter α controls the complexity of the model, can be chosen using cross validation.

Regularized Discriminant Analysis

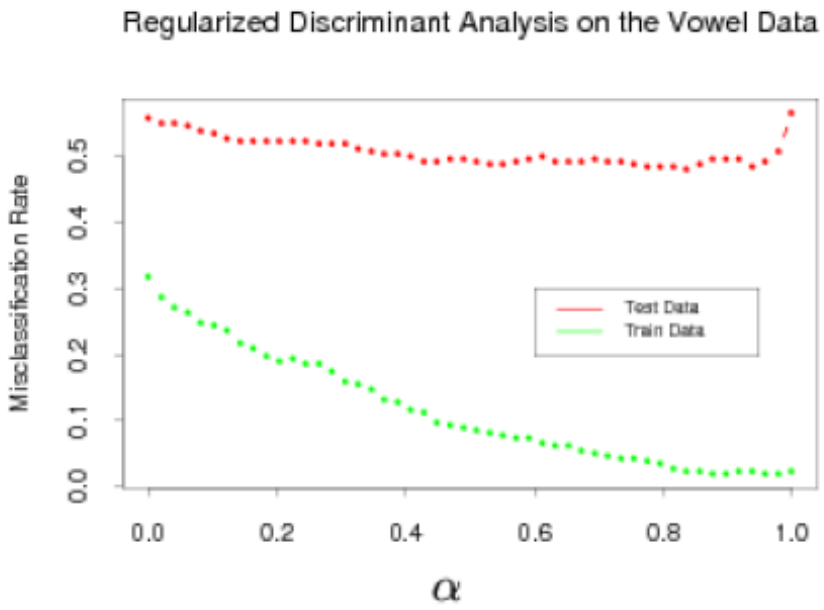


Figure 4.7: *Test and training errors for the vowel data, using regularized discriminant analysis with a series of values of* $\alpha \in [0,1]$. *The optimum for the test data occurs around* $\alpha = 0.9$, *close to quadratic discriminant analysis.*

$$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}.$$

Reduced-Rank LDA

- The idea: we can **view the data** using low dimensional projections and lose no information.

- LDA – we only need to consider the data in a subspace at most K-1 dimensions.

- ***For example:***

  If K=3 allows us to view the data in a two- dimensional subspace, and use color coding to depict classes.

  If K>3 then we may look for a L<K-1 dimensional subspace $H_L \subseteq H_{K-1}$ *that is "optimal" for LDA.*

Reduced-Rank LDA

- Fisher defined “optimal” to mean that the projected centroids are spread out as much as possible in terms of variance.

  This amounts to finding the principal component subspaces of the centroids themselves.

- Discriminant coordinates aka
  Discriminant variables aka
  Canonical variables.

Reduced-Rank LDA

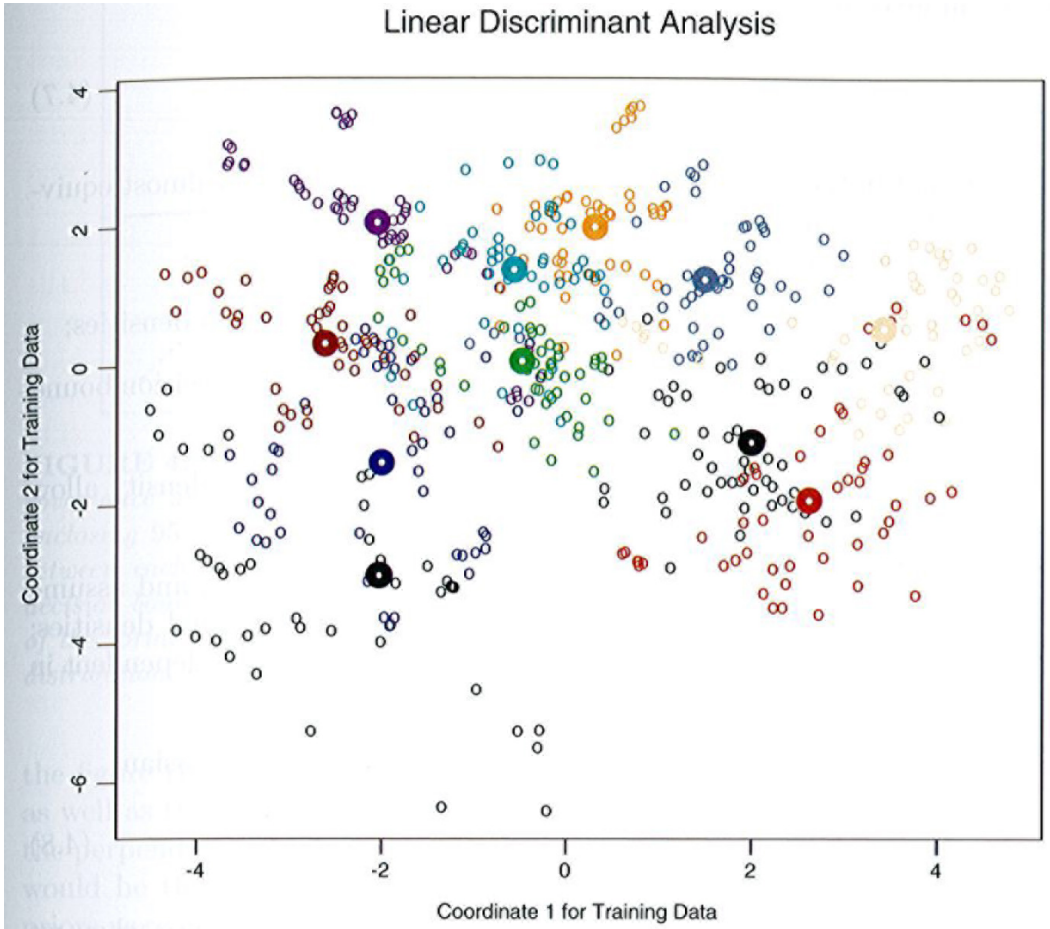
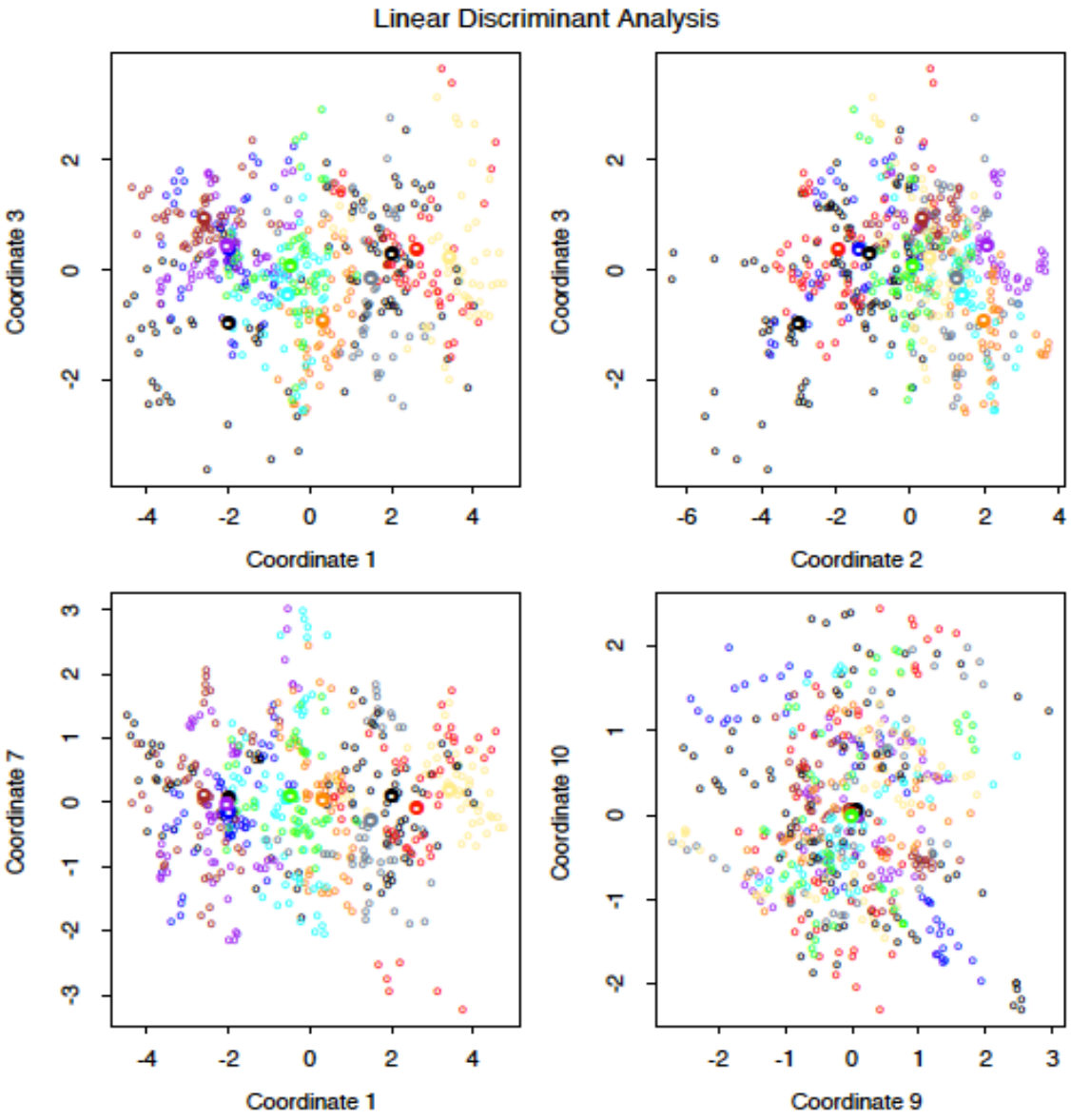


FIGURE 4.4. *A two-dimensional plot of the vowel training data. There are eleven classes with $X \in \mathbb{R}^{10}$, and this is the best view in terms of a LDA model (Section 4.3.3). The heavy circles are the projected mean vectors for each class. The class overlap is considerable.*

Reduced-Rank LDA



Visualizing the Criterion

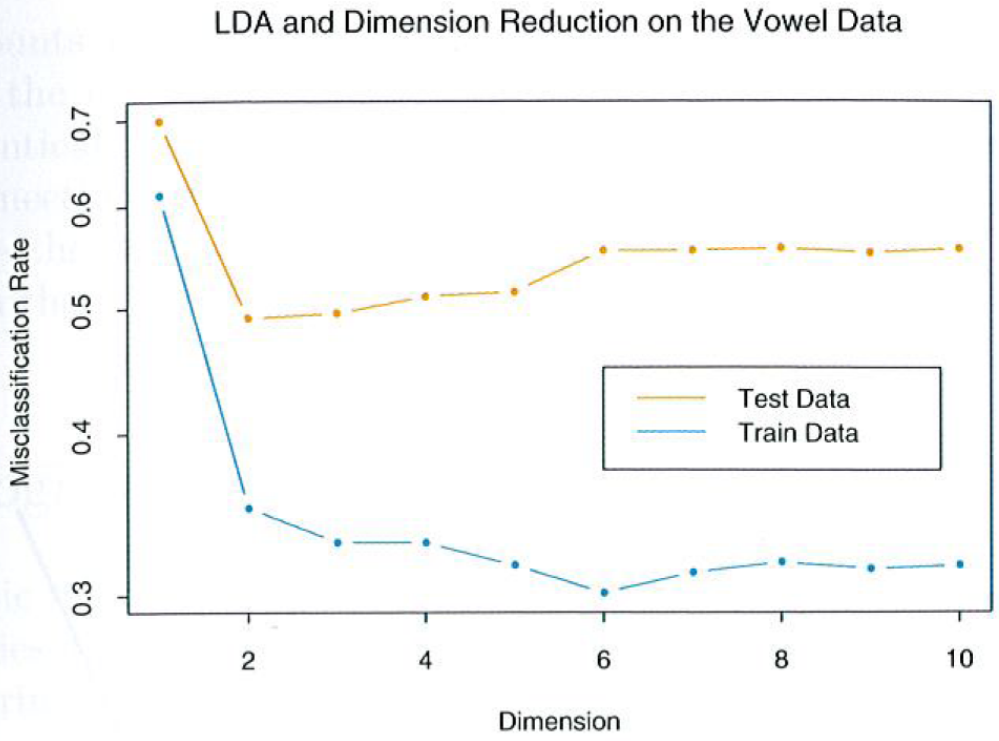
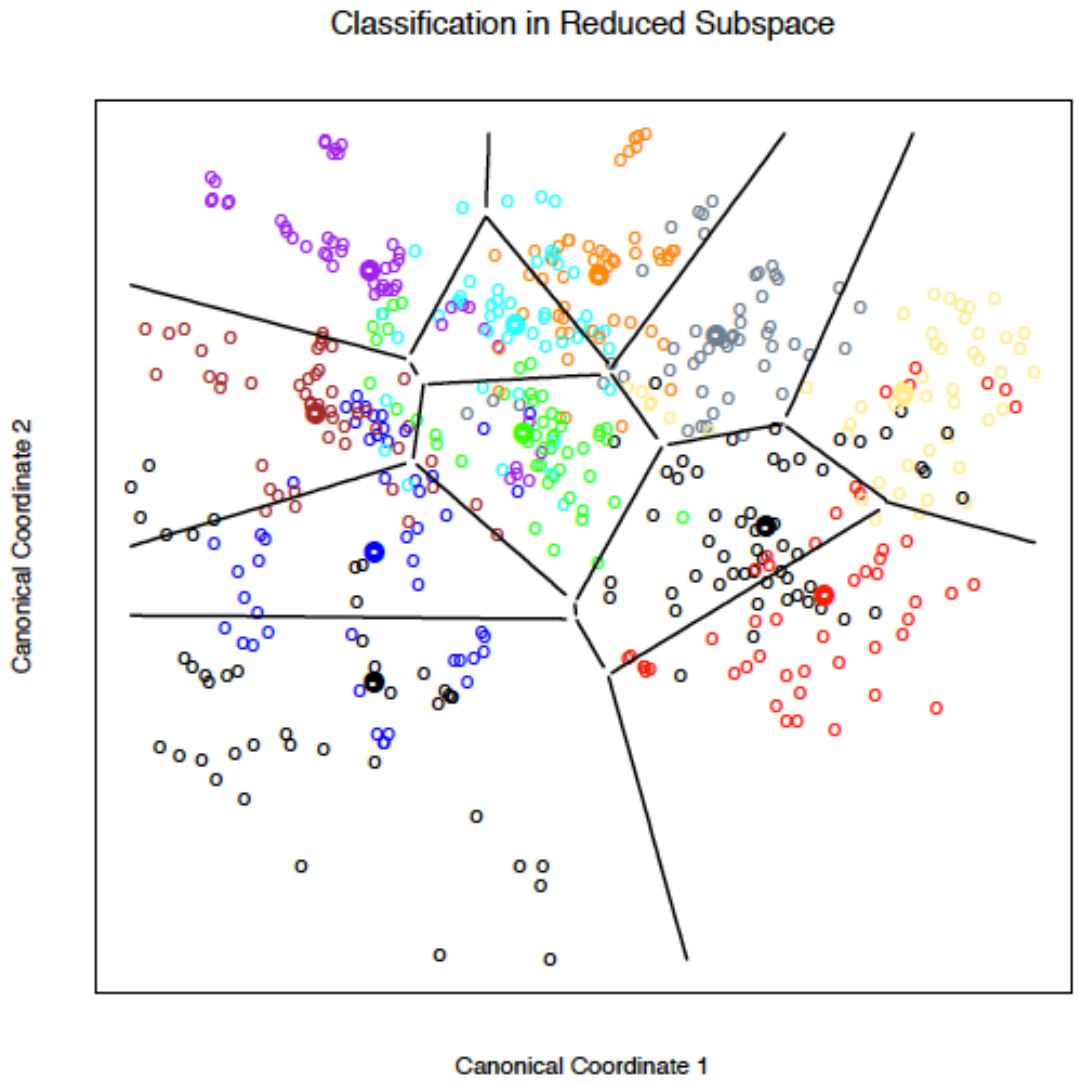


FIGURE 4.10. *Training and test error rates for the vowel data, as a function of the dimension of the discriminant subspace. In this case the best error rate is for dimension 2. Figure 4.11 shows the decision boundaries in this space.*

Vowel Data


Classification in Reduced Subspace
Canonical Coordinate 2
Canonical Coordinate 1