Supervised learning: Linear Discriminant Analysis

Linear Discriminant Analysis

Discriminant analysis techniques are used to classify individuals into one of two or more alternative groups (or populations) on the basis of a set of measurements, say p.

The populations are known to be distinct, and each individual belongs to one of them. These techniques can also be used to identify which variables contribute to making the classification. Thus, as in regression analysis, we have two uses, prediction and description.

Bayes classifier

1. Suppose $f_k(x)$ is the class-conditional density of X in class $G = k$, and let π_k be the prior probability of class k, with $\sum_{k=1}^{K} \pi_k = 1$.
2. A simple application of Bayes Theorem gives us

$$\delta_k(x) \equiv \Pr(G = k \mid X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

3. We see that in terms of ability to classify, having the $f_k(x)$ is almost equivalent to having the quantity $\Pr(G = k|X = x)$.
4. Rule: Classify an observation to a class for which $\delta_k(x)$ is largest.

Classifier based on class density functions

1. Naive Bayes models assume that each of the class densities are products of marginal densities; that is, they assume that the inputs are conditionally independent in each class.
2. linear and quadratic discriminant analysis use Gaussian densities;

Suppose that we model each class density as multivariate Gaussian

$$f_k(x) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k{}^{-1}(x-\mu_k)} \text{....(\#)}$$

Linear discriminant analysis (LDA) arises in the special case when we assume that the classes have a common covariance matrix $\Sigma_k = \Sigma$, for all k.

In comparing two classes *k* and *l*, it is sufficient to look at the log-ratio, and we see that

$$\log \frac{\Pr(G=k \mid X=x)}{\Pr(G=l \mid X=x)}$$

$$= \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l} = \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)^T \Sigma^{-1}(\mu_k - \mu_l) + x^T \Sigma^{-1}(\mu_k - \mu_l) \text{....(*)}$$

which is an equation in *x*.

Linear decision boundaries

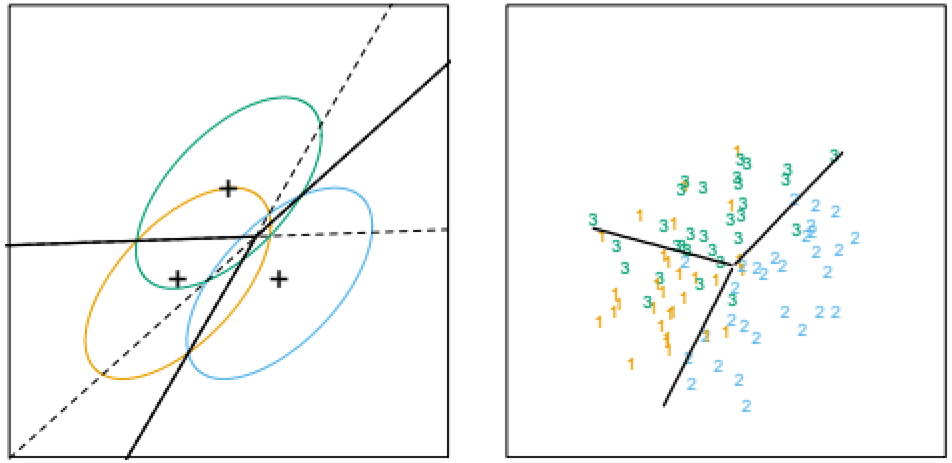
The equal covariance matrices cause the normalization factors to cancel, as well as the quadratic part in the exponents.

This linear log-odds function implies that the decision boundary between classes k and l -- the set where $\Pr(G = k|X = x) = \Pr(G = l|X = x)$ -- is linear in x ; in p dimensions a hyperplane.

This is of course true for any pair of classes, so all the decision boundaries are linear.

If we divide $\mathbb{R}^p$ into regions that are classified as class 1, class 2, etc., these regions will be separated by hyperplanes.

The figure shows an idealized example with three classes and p = 2.



Here the data do arise from three Gaussian distributions with a common covariance matrix.

We have included in the figure the contours corresponding to 95% highest probability density, as well as the class centroids.

LDA rule

- From equation (*), we see that the linear discriminant functions

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

are an equivalent description of the decision rule, with $G(x) = \mathit{arg}\ \max_k \delta_k(x)$.

In practice we do not know the parameters of the Gaussian distributions, and will need to estimate them using our training data:

- $\hat{\pi}_k(x) = N_k \,/\, N,$ where is the number of class-k observations;
- $\hat{\mu}_k = \sum_{g_i = k} x_i \,/\, N_k;$
- $\hat{\Sigma} = \sum_{k=1}^{K} \sum_{g_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T \,/\, (N - K)$

QDA

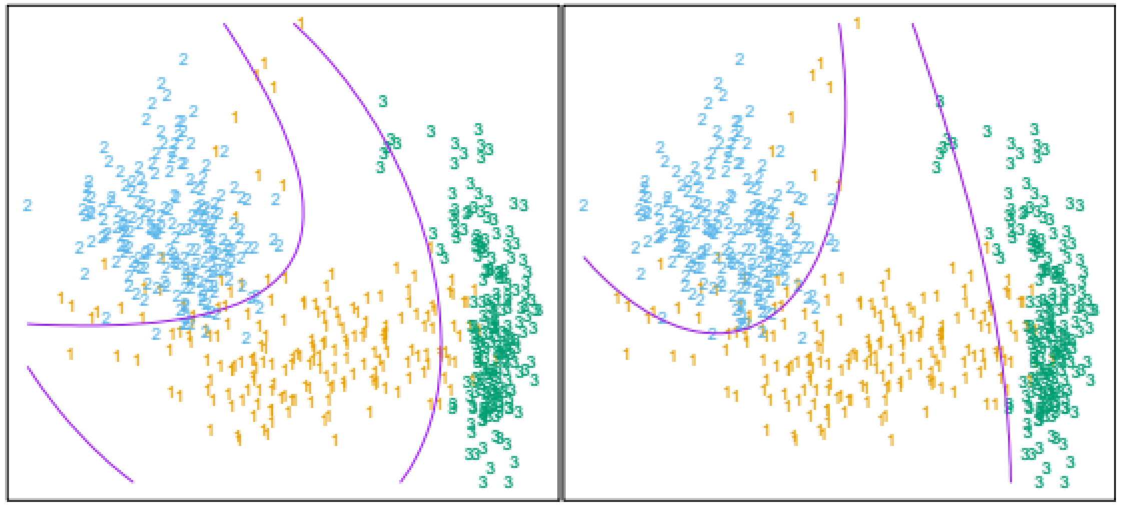
Getting back to the general discriminant problem (#), if the Σ_k are not assumed to be equal, then the convenient cancellations in equation (*) do not occur; in particular the pieces quadratic in *x* remain.

We then get quadratic discriminant functions (QDA),

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

The decision boundary between each pair of classes k and l is described by a quadratic equation $\{x : \delta_k(x) = \delta_l(x)\}$.

The figure shows an example where the three classes are Gaussian mixtures and the decision boundaries are approximated by quadratic equations in *x*.



1. Here we illustrate two popular ways of fitting these quadratic boundaries.
2. The right plot uses QDA as described here, while the left plot uses LDA in the enlarged five-dimensional quadratic polynomial space.
3. The differences are generally small; QDA is the preferred approach, with the LDA method a convenient substitute.

LDA or QDA? A bias-variance trade-off

- QDA: when there are p predictors, then estimating a covariance matrix requires estimating $p(p+1)/2$ parameters. QDA estimates a separate covariance matrix for each class, for a lot of $Kp(p+1)/2$ parameters. Low Bias.
- LDA: By assuming that the K classes share a common covariance matrix, there are only Kp parameters to estimate. Low variances.
- Recommendation: LDA tends to be a better bet than QDA if there are relatively few training observations. QDA is recommended if the training set is very large.

Case: Swiss Banknotes Data

- The dataset contains six measurements made on 100 genuine and 100 counterfeit old 1000-Franc Swiss banknotes.

Data Description

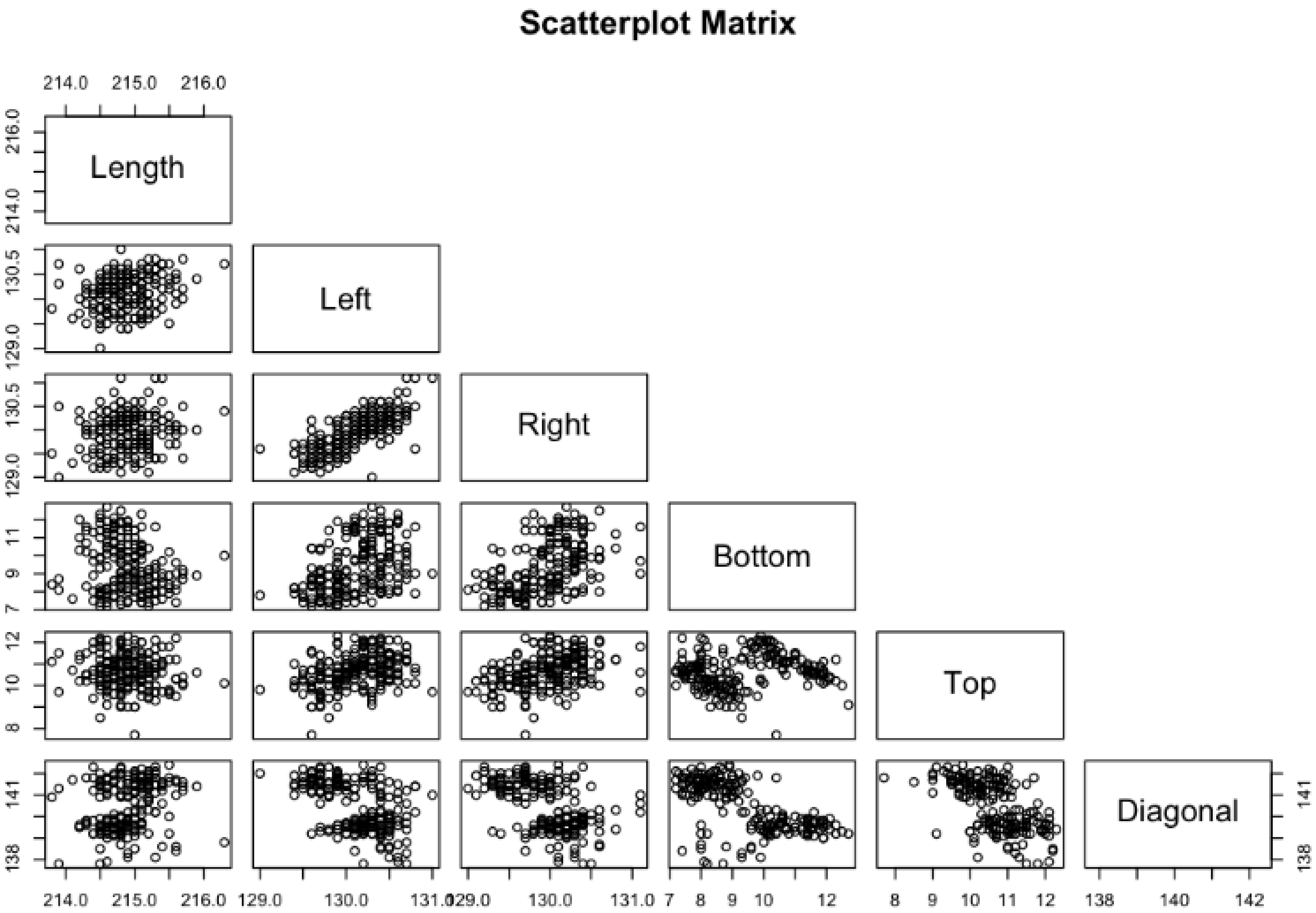
Name	Details
Length	Length of bill (mm)
Left	Width of left edge (mm)
Right	Width of right edge (mm)
Bottom	Bottom margin width (mm)
Top	Top margin width (mm)
Diagonal	Length of diagonal (mm)
Type	1 for genuine; 0 for counterfeit

Data Description

- Group Means:

	Type = 0	Type = 1
Length	214.823	214.969
Left	130.300	129.943
Right	130.193	129.720
Bottom	10.530	8.305
Top	11.133	139.450
Diagonal	139.450	141.517

Data Description



Methodology

- In total 200 banknote observations.
- Randomly divide the whole dataset into halves.
- Name them train_set and test_set.
- Implement LDA using the train_set and use the test_set to evaluate the performance of the resultant linear discriminant function.
- Iterate the above procedures.

Prior Information

Assume

- Pr(A banknote is counterfeit) = Pr(Type = 0) = 0.01;
- Pr(A banknote is genuine) = Pr(Type = 1) = 0.99.

One Iteration

- Coefficients of Linear Discriminants:

Name	**LD1**
Length	-0.2626489
Left	0.8026030
Right	-0.5637819
Bottom	-0.9848281
Top	-1.1832084
Diagonal	1.6870771

One Iteration

- Confusion Matrix
- The apparent error rate (APER) = (1+0)/100 = 1%.

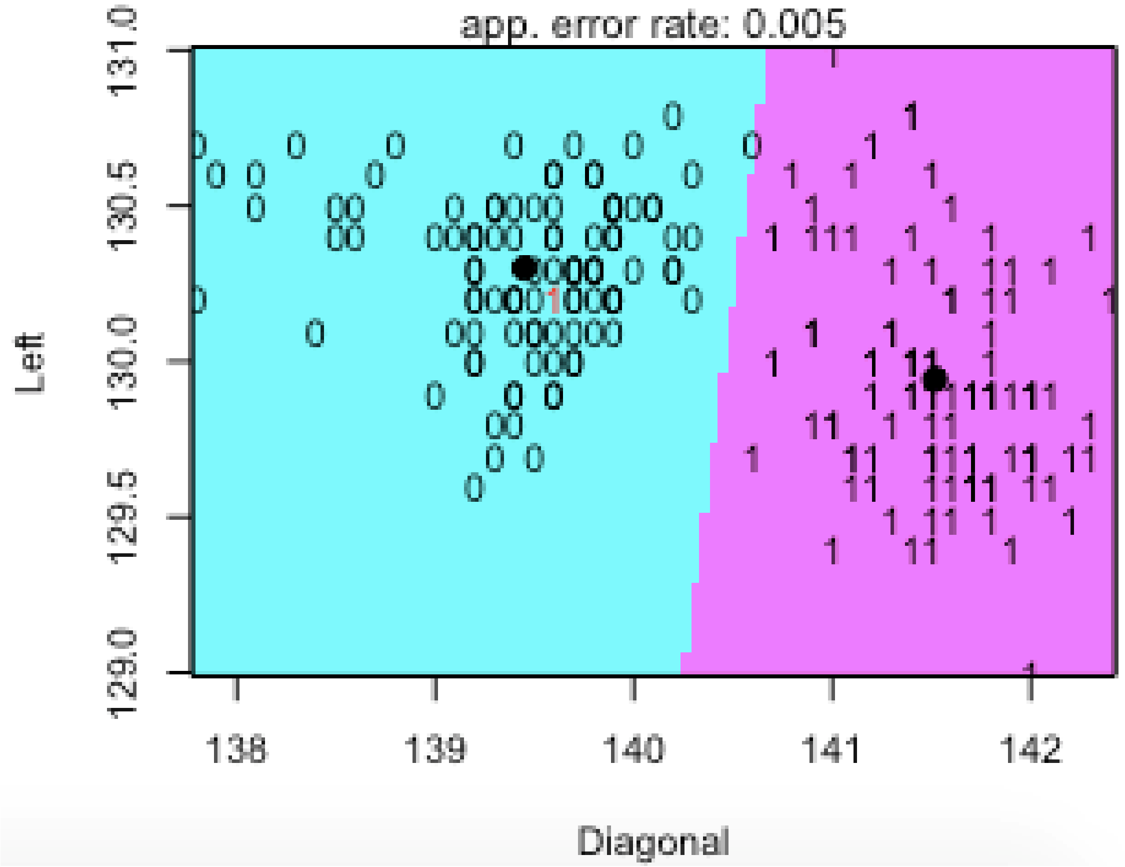
	Predicted 0	**Predicted 1**	**Total**
Actual 0	48	0	48
Actual 1	1	51	52
Total	49	51	100

Ten replication

- Average APER = (1 + 0 + 1 + 1 + 2 + 1 + 1 + 1 + 4 + 0)% / 10 = 1.2%
- High Accuracy!
- Next, train a LD function using the whole dataset.

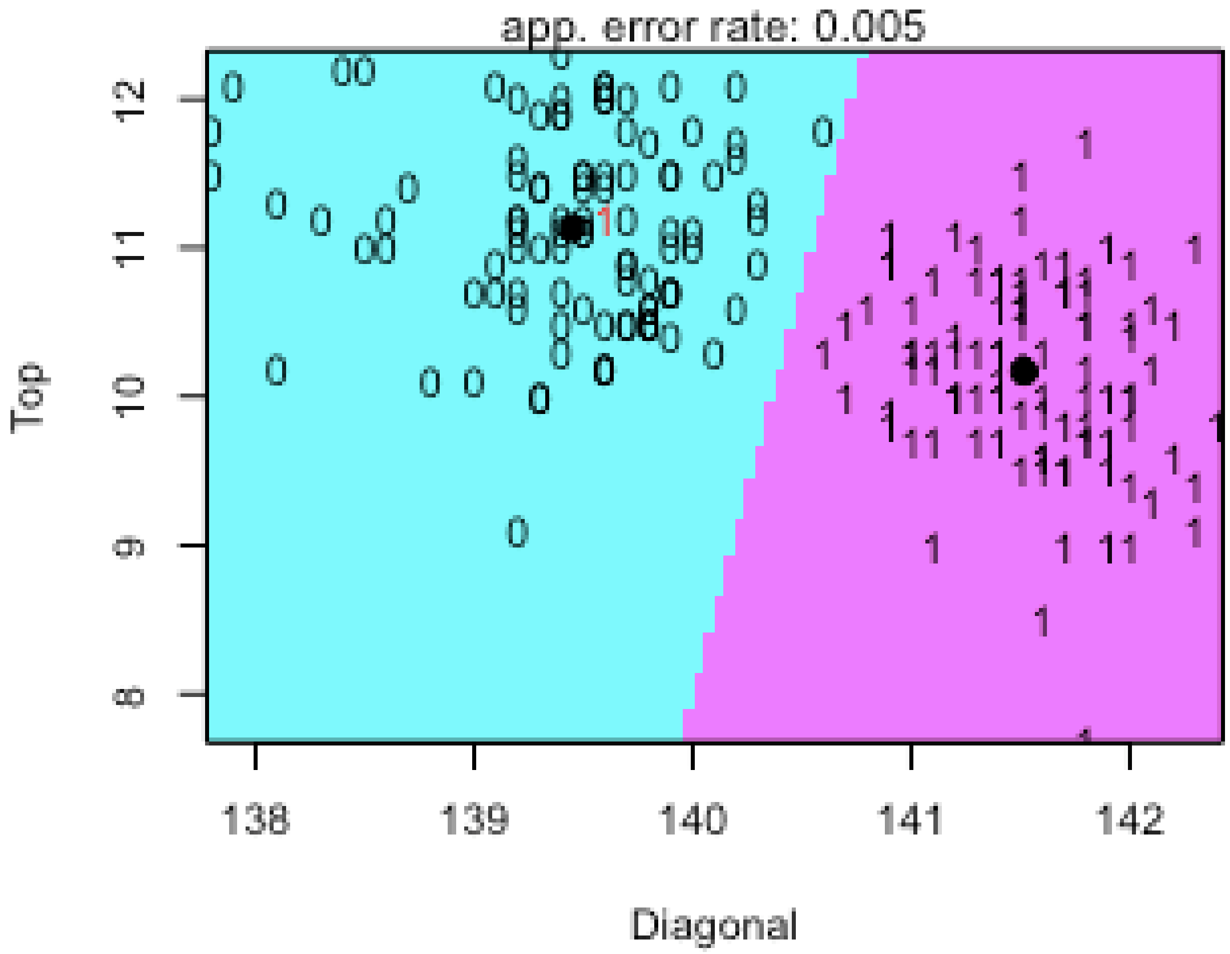
2D Partition Plots

- Best partition plot, Left v.s. Diagonal:

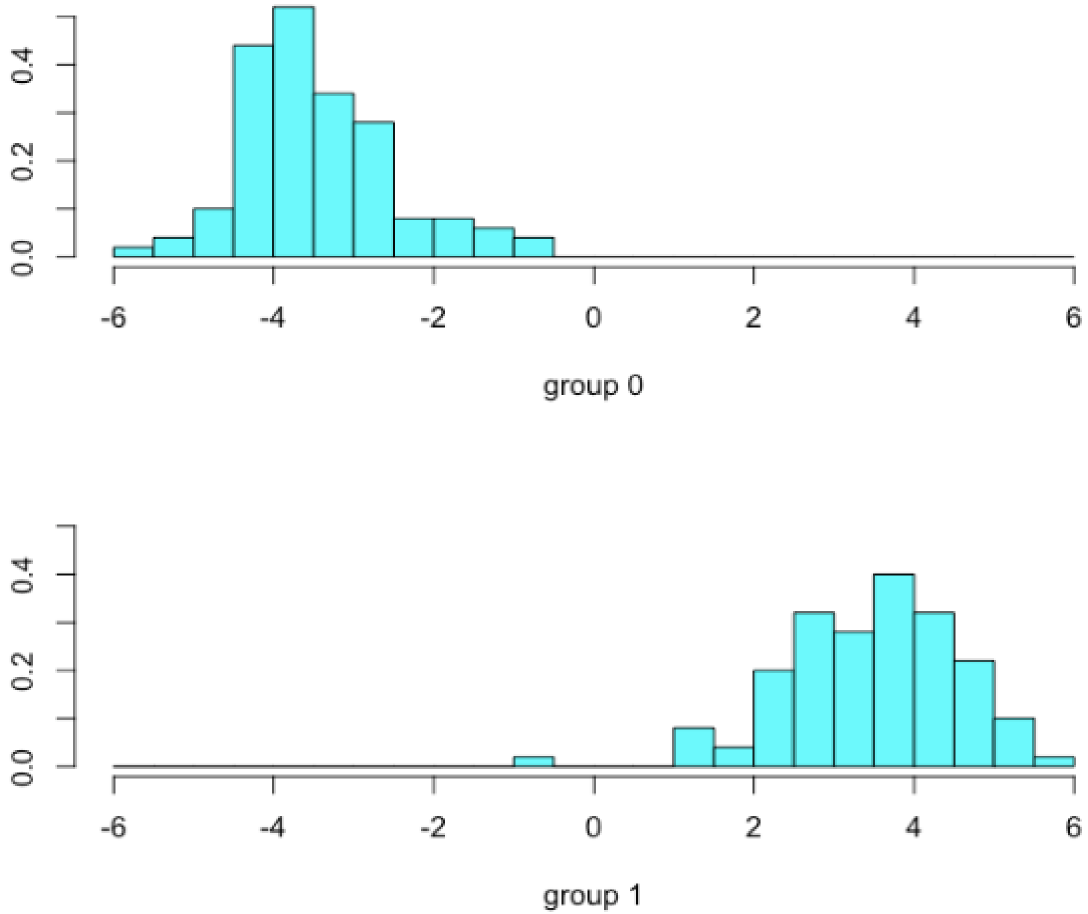


2D Partition Plots

- Best partition plot, Top v.s. Diagonal:



Histogram of LDA Values



Regularized Discriminant Analysis

The regularized covariance matrices have the form

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

where $\hat{\Sigma}$ is the pooled covariance matrix as used in LDA.

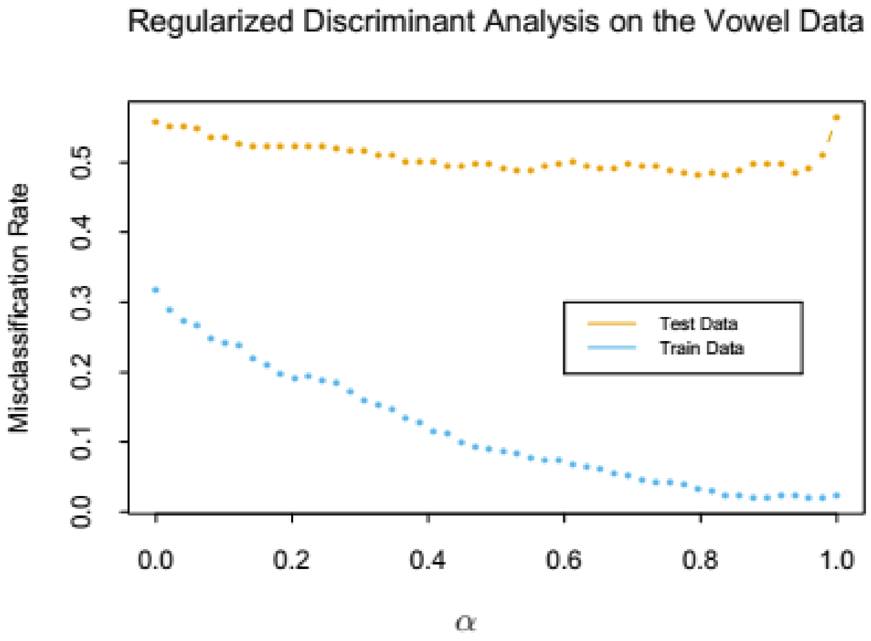
Here α ∈ [0, 1] allows a continuum of models between LDA and QDA, and needs to be specified.

In practice α can be chosen based on the performance of the model on validation data, or by cross-validation.

The figure shows the results of RDA applied to the vowel data.

Both the training and test error improve with increasing α, although the test error increases sharply after α = 0.9.

The large discrepancy between the training and test error is partly due to the fact that there are many repeat measurements on a small number of individuals, different in the training and test set.



Similar modifications allow $\hat{\Sigma}$ itself to be shrunk toward the scalar covariance,

$$\hat{\Sigma}(\gamma) = \gamma\hat{\Sigma} + (1-\gamma)\hat{\sigma}^2 I.....(2)$$

for γ ∈ [0, 1].

Replacing $\hat{\Sigma}$ in equation (1) by $\hat{\Sigma}(\gamma)$ leads to a more general family of covariances $\hat{\Sigma}(\alpha, \gamma)$ indexed by a pair of parameters.

In these situations the features are high-dimensional and correlated, and the LDA coefficients can be regularized to be smooth or sparse in the original domain of the signal.

This leads to better generalization and allows for easier interpretation of the coefficients.

Computations for LDA

The computation is simplified by diagonalizing $\hat{\Sigma}$ or $\hat{\Sigma}_k$.

Suppose we compute the eigen-decomposition for each $\hat{\Sigma}_k = U_k D_k U_k^T$, where U_k is p × p orthonormal, and D_k a diagonal matrix of positive eigenvalues d_{kl}.

Then the ingredients for $\delta_k(x)$ are

- $(x-\hat{\mu}_k)^T \hat{\Sigma}_k^{-1}(x-\hat{\mu}_k) = [U_k^T(x-\hat{\mu}_k)]^T D_k^{-1}[U_k^T(x-\hat{\mu}_k)]$;
- $\log|\hat{\Sigma}_k| = \Sigma_l \log d_{kl}$

In light of the computational steps outlined above, the LDA classifier can be implemented by the following pair of steps:

- Sphere the data with respect to the common covariance estimate $\hat{\Sigma} : X^* \leftarrow D^{-\frac{1}{2}} U^T X$, where $\hat{\Sigma} = UDU^T$. The common covariance estimate of X^* will now be the identity.
- Classify to the closest class centroid in the transformed space, modulo the effect of the class prior probabilities π_k.

Reduced-Rank Linear Discriminant Analysis

- The K centroids in p-dimensional input space lie in an affine subspace of dimension ≤ K - 1, and if p is much larger than K, this will be a considerable drop in dimension.
- Moreover, in locating the closest centroid, we can ignore distances orthogonal to this subspace, since they will contribute equally to each class.

- Thus we might just as well project the X^* onto this centroid-spanning subspace H_{K-1}, and make distance comparisons there.

- Thus there is a fundamental dimension reduction in LDA, namely, that we need only consider the data in a subspace of dimension at most K - 1.

- If K = 3, for instance, this could allow us to view the data in a two dimensional plot, color-coding the classes. In doing so we would not have relinquished any of the information needed for LDA classification.

- If K > 3, we might then ask for a $L < K - 1$ dimensional subspace $H_L \subseteq H_{K-1}$ optimal for LDA in some sense.

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

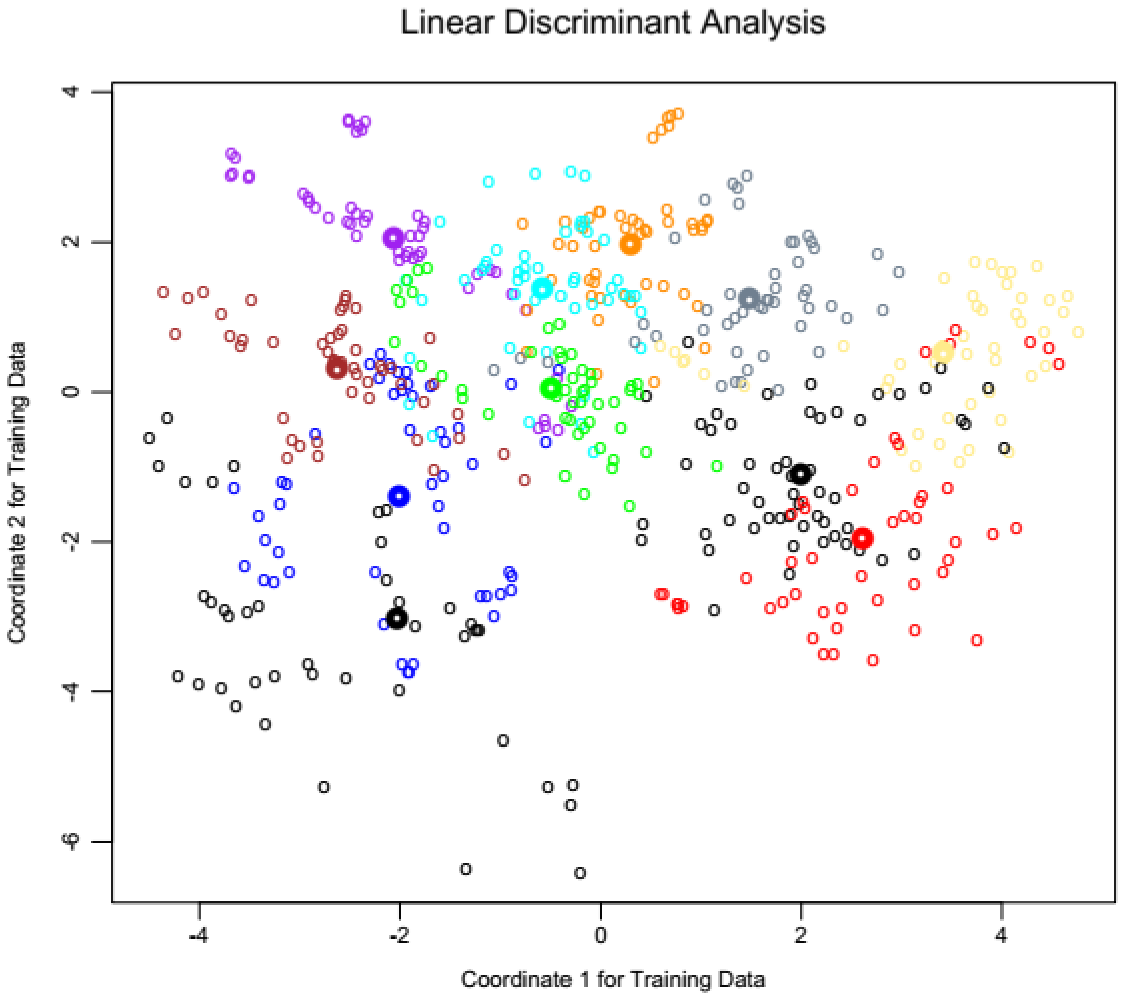
- This amounts to finding principal component subspaces of the centroids themselves.

The figure shows such an optimal two-dimensional subspace for the vowel data.

Here there are eleven classes, each a different vowel sound, in a ten-dimensional input space.

The centroids require the full space in this case, since K - 1 = p, but we have shown an optimal two-dimensional subspace.

The dimensions are ordered, so we can compute additional dimensions in sequence.



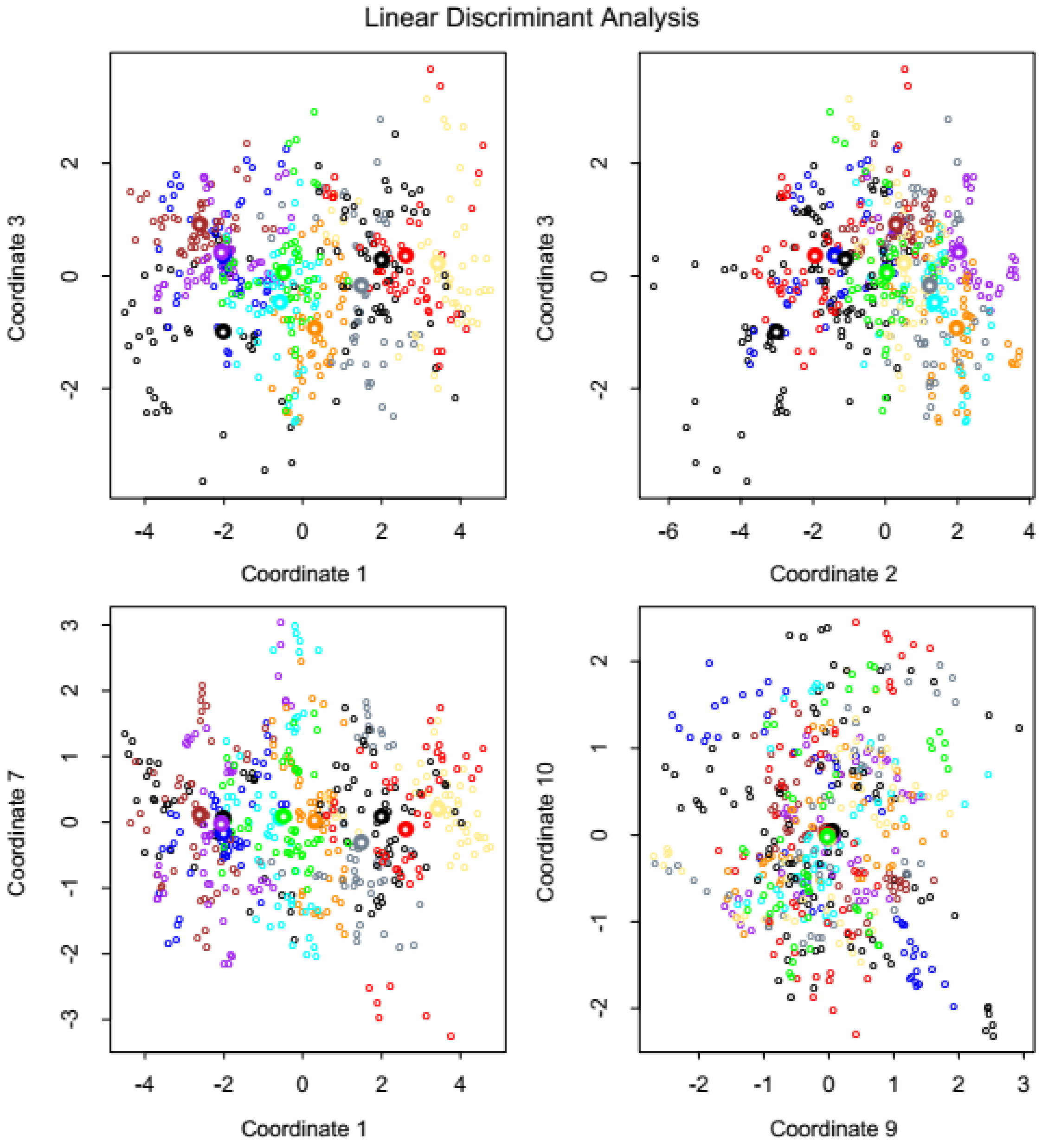
The figure shows four additional pairs of coordinates, also known as canonical or discriminant variables.

In summary then, finding the sequences of optimal subspaces for LDA involves the following steps:

- compute the $K \times p$ matrix of class centroids $\mathbf{M}$ and the common covariance matrix $\mathbf{W}$ (for within-class covariance);
- compute $\mathbf{M}^* = \mathbf{M}\mathbf{W}^{-1/2}$ using the eigen-decomposition of $\mathbf{W}$;
- compute $\mathbf{B}^*$, the covariance matrix of $\mathbf{M}^*$ ($\mathbf{B}$ for *between-class* covariance), and its eigen-decomposition $\mathbf{B}^* = \mathbf{V}^* \ \mathbf{D}_\mathbf{B}\mathbf{V}^{*\,\mathrm{T}}$. The columns v_l^* of $\mathbf{V}^*$ in sequence from first to last define the coordinates of the optimal subspaces.

Combining all these operations the *l*-th discriminant variable is given by

$$Z_l = v_l^{\mathrm{T}} X \text{ with } v_l = \mathbf{W}^{-1/2} v_l^*.$$


Linear Discriminant Analysis
Coordinate 3
Coordinate 1
Coordinate 3
Coordinate 2
Coordinate 7
Coordinate 1
Coordinate 10
Coordinate 9