STAT 621 Lecture Notes
Linear Discriminant Analysis

Next we another method for classification. Again, we have have a distinct response variable whose value we would like to predict based on a set of clearly defined predictor variables.

Linear Discriminant Analysis (LDA) is another approach to supervised classification. This technique is designed to allow multiple predictors, $X_1, \ldots, X_p$ for classification of the single response Y. It shares some similarities with the Naive Bayes classifier. Suppose data $(Y_i, X_{i1}, \ldots, X_{ip})$ are recorded for $i = 1, \ldots, n$. Also assume that the response Y falls into one of J categories.

Recall from our discussion of the Naive Bayes classifier that the probability that a response falls into category j can be written

$$P(Y = j | \boldsymbol{X}) = \frac{f_j(\boldsymbol{x})\pi_j}{\sum_{k=1}^{J} f_k(\boldsymbol{x})\pi_j}$$

Since the denominator in the expression above is constant in Y, the focus for estimation is on the numerator (i.e., the denominator is just a scaling factor that can be determined at the end). Recall the terms in the numerator and how we discussed estimating them.

In Linear Discriminant Analysis (LDA), it is assumed that the density in each class follows a multivariate normal distribution. That is,

$$f_j(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_j)\right\}$$

Here the parameters are

LDA seeks to partition the p-dimensional space spanned by the covariates into sections associated with each response category. Then given new values of the covariates, a predicted response is determined by the section into which the covariate values fall. To define the partition, we need to find the boundaries that define the J response categories. That is, we want to find the boundaries where

$$P(Y = j|\boldsymbol{X} = \boldsymbol{x}) = P(Y = k|\boldsymbol{X} = \boldsymbol{x}) \qquad j \neq k$$

Here's a picture from *Elements of Statistical Learning*, Chapter 4.
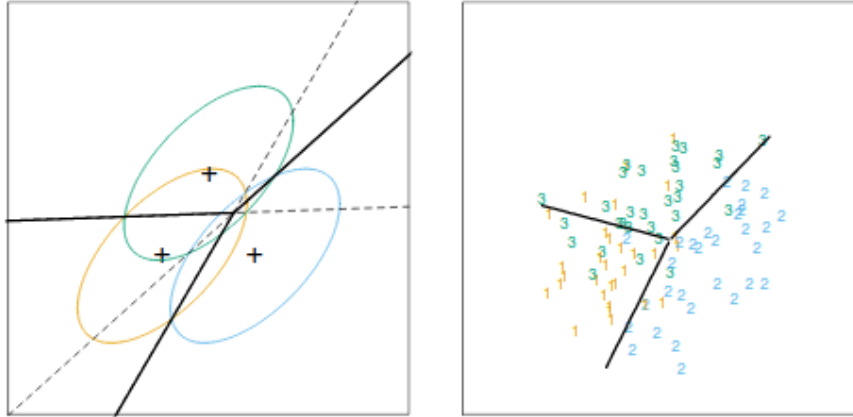
FIGURE 4.5. *The left panel shows three Gaussian distributions, with the same covariance and different means. Included are the contours of constant density enclosing 95 % of the probability in each case. The Bayes decision boundaries between each pair of classes are shown (broken straight lines), and the Bayes decision boundaries separating all three classes are the thicker solid lines (a subset of the former). On the right we see a sample of 30 drawn from each Gaussian distribution, and the fitted LDA decision boundaries.*

It turns out that under the assumption the the $f_j(\boldsymbol{x})$ are normal with equal covariance functions, that these boudaries are given by linear combinations of the X's. In other words, regions associated with different response categories are separated by points when $p = 1$, lines when $p = 2$, and hyperplanes when $p > 2$. These are the linear discriminant functions,

$$\delta_j(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j - \frac{1}{2}\boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j + \log(\pi_j), \qquad j = 1, \ldots, J$$

Estimation for LDA boils down to estimating the unknown parameters in the above equations. Specifically,

Then, given an observed values of the covariates $\boldsymbol{x}^*$, a response Y^* is assigned to the class for which $\delta_j(\boldsymbol{x}^*)$ is largest.

Example: This example uses a data set of measurements from prehistoric bones of different dogs. The species measured were Modern dog, Golden jackal, Chinese wolf, Indian wolf, Cuon, Dingo and Prehistoric dog. Measurements included mandible bredth, molar length, and other morphological characteristics. A little of the data.

```
> head(dogs)
  species  X1    X2 X3 X4 X5  X6 X7 X8  X9
1  modern 123 10.1 23 23 19 7.8 32 33 5.6
2  modern 137  9.6 19 22 19 7.8 32 40 5.8
3  modern 121 10.2 18 21 21 7.9 35 38 6.2
```
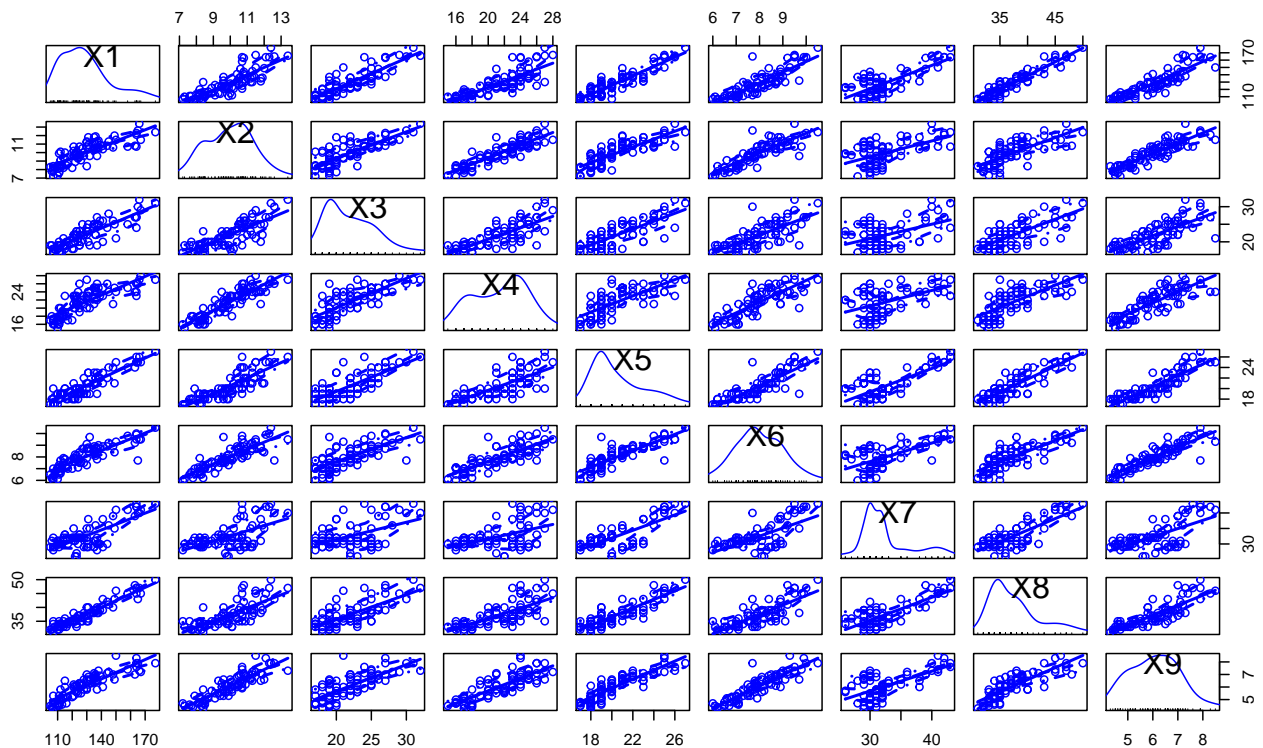

The `lda` function in the `MASS` package fits the LDA model. Here's the basic code and output.

```
> dog.lda=lda(species~., CV=F, data=dogs)    # add prior= to change prior (default=sample props)
> dog.lda
Call:
lda(species ~ ., data = dogs, CV = F)

Prior probabilities of groups:
     cuon    indwolf     jackal     modern     prehist
0.2207792 0.1818182 0.2597403 0.2077922 0.1298701

Group means:
             X1       X2       X3       X4       X5       X6       X7       X8       X9
cuon    133.2353 10.72353 24.05882 23.64706 21.47059 8.488235 29.00000 37.70588 6.611765
indwolf 157.3571 11.57857 26.21429 24.71429 24.71429 9.335714 40.21429 44.78571 7.407143
jackal  111.0000  8.18000 18.60000 17.00000 18.20000 6.815000 30.35000 33.35000 4.805000
modern  125.9375  9.72500 21.37500 21.12500 19.37500 7.675000 32.06250 36.62500 5.868750
prehist 122.8000 10.34000 20.00000 22.90000 19.30000 8.190000 32.80000 35.90000 6.170000
```

3

```
Coefficients of linear discriminants:
          LD1         LD2         LD3         LD4
X1 -0.1262457 -0.02141331 -0.07415646 -0.09390916
X2  0.1080702  0.02953726  0.56106589 -0.07130602
X3  0.2910043 -0.03424176 -0.10744062 -0.14141364
X4 -0.2270426  0.04637768  0.43534619  0.09499400
X5 -0.8891596 -0.74578775 -1.12909329  0.68431718
X6 -0.8115373  0.10875761  0.42482620  1.09992590
X7  1.3427053 -0.17820451  0.33210847  0.04612700
X8  0.2269895 -0.09534583  0.01837503 -0.07252568
X9 -1.6386618  0.44639830  1.20670414 -0.78224140


Proportion of trace:
   LD1    LD2    LD3    LD4
0.6505 0.2592 0.0860 0.0042
```
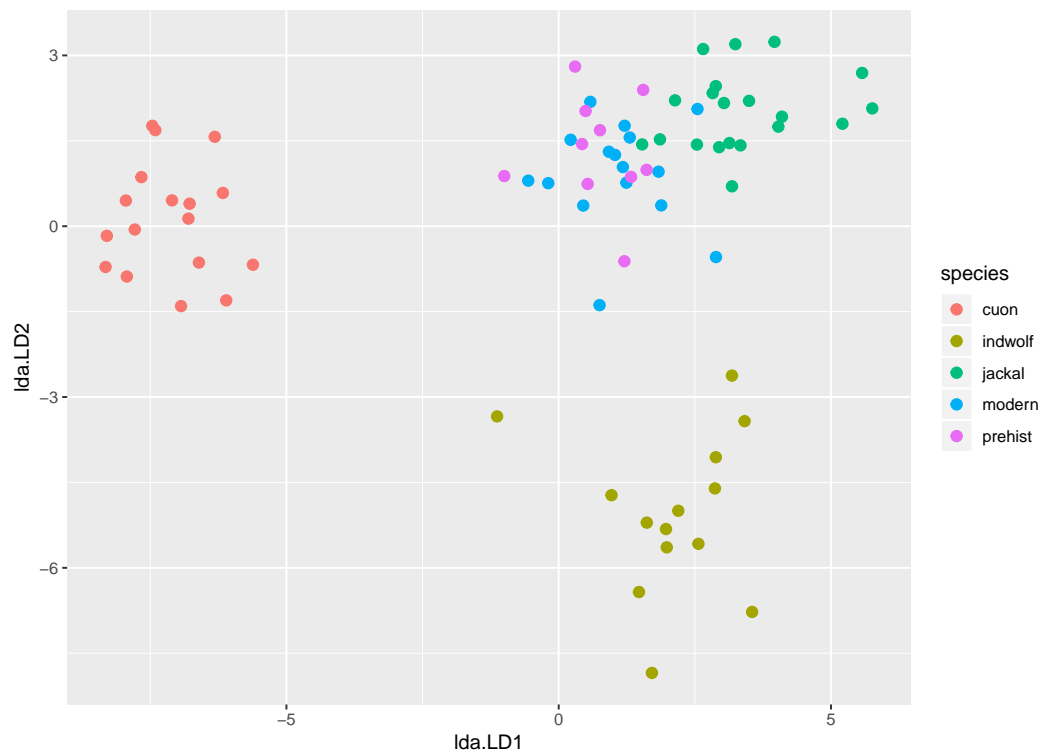
`plot(dog.lda)` will make a scatterplot matrix of the values of each discriminant function. Since the first two explain almost all of the between group variation, I'll just plot those.

```
lda.values=predict(dog.lda)
newdata <- data.frame(species = dogs[,1], lda = lda.values$x)
library(ggplot2)
ggplot(newdata) + geom_point(aes(lda.LD1, lda.LD2, colour = species), size = 2.5)
```

I'll try fitting the model on a training set, and predicting species on a test set.

```
> dog.lda2=lda(species~.,data=train)
> aa=predict(object=dog.lda2,newdata=test)
> aa
$class
 [1] modern  modern  modern  prehist jackal  jackal  jackal  modern  jackal  jackal  jackal  cuon    cuon    cuon    cu
[17] cuon    cuon    cuon    indwolf modern  indwolf indwolf modern  prehist prehist prehist
Levels: cuon indwolf jackal modern prehist

$posterior
          cuon       indwolf        jackal        modern       prehist
2  4.995332e-20 2.548051e-09 4.184484e-04 9.978671e-01 1.714481e-03
7  2.598517e-23 5.758137e-10 2.003900e-02 9.779253e-01 2.035685e-03
12 2.963896e-22 1.208921e-12 2.448042e-01 7.551648e-01 3.100589e-05

etc......


> table(aa$class, test$species)

          cuon indwolf jackal modern prehist
  cuon       8       0      0      0       0
  indwolf    0       3      0      0       0
  jackal     0       0      6      0       0
  modern     0       1      1      3       1
  prehist    0       0      0      1       3
```

Other comments: Quadratic Discriminant Analysis, ...