We want to develop “rules” or criteria based on observations, that can be used to classify the observations/subjects. For example, can be develop a rule, based on the morphological measurements, that is able to classify the sparrows into survival groups? There will be some errors in such classification, unless the data are well separated into classes.
Data structure

Assume that we observe data from $G$ groups. Our data matrix $X$ contains the observed values, $x_{ij}$, of the $j$th variable for the $i$th individual/unit; $i = 1, \ldots, n$ and $j = 1, \ldots, p$. It also contains an indicator variable $g_i \in \{1, \ldots, G\}$, $i = 1, \ldots, n$, whose value indicates to which group the observation belongs.

$$ X = \begin{pmatrix} g_1 & x_{11} & x_{12} & \cdots & x_{1p} \\ g_2 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ g_n & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} g_1 & \mathbf{x}_1^T \\ g_2 & \mathbf{x}_2^T \\ \vdots & \vdots \\ g_n & \mathbf{x}_n^T \end{pmatrix} $$
Thus we have $n$ subjects, and on each subject we have measurements on $p$ variables. We also have the information on group membership of the subjects. Each subject falls into only one group, and several subjects will be in a given group. Put $n_g$ as the number of subjects that are in group $g$, $g = 1, 2, \ldots, G$. Put $x_{gi}$ as the $i^{th}$ subject in group $g$, $i = 1, 2, \ldots, n_g$. 
Fisher’s linear discriminant analysis

Idea:
Find a linear projection of the data such that the ratio of between groups variability to within groups variability is maximised.
We first need to define the total, between and within sums of squares. These are defined in an analogous way to ANOVA, except that we now our observations are vectors. Define the total sum of squares $S$ by

$$S = \sum_{i=1}^{n} (x_i - \bar{x})^\top (x_i - \bar{x})$$

Define the between group sums of squares $B$ by

$$B = \sum_{g=1}^{G} n_g (x_g - \bar{x})^\top (x_g - \bar{x})$$
Sums of squares

Finally, define the within group sums of squares $W$ by

$$W = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (x_{gi} - \bar{x}_g)^\top (x_{gi} - \bar{x}_g)$$

Note that $S = B + W$. 
Thus, Fisher’s linear discriminant function is defined as the solution \( \mathbf{a} \in \mathbb{R}^p \mathbf{a} \neq \mathbf{0}_p \) found by maximising

\[
\frac{\mathbf{a}^T \mathbf{B} \mathbf{a}}{\mathbf{a}^T \mathbf{W} \mathbf{a}}
\]

As for principal components analysis, we can take further eigenvectors corresponding to the next largest eigenvalues.

There will be at most \( r = \min(p, G - 1) \) positive eigenvalues.

The corresponding transformed variables are called the \textit{linear discriminants} or \textit{canonical variates}.
Linear discriminant analysis (cont.)

What are the posterior probabilities for the classes?

Remember Bayes’ theorem:
For two events $A$ and $B$, with $P(B) > 0$, we have:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Let $A$ be the event that we observe an individual/unit from group $g$.
Let $B$ be the event that we observe $x$.
If we assume a continuous distribution for $x$, then $P(B) = 0$, but this technical problem can be fixed.
Then

\[ P(g|x) = \frac{P(x|g) P(g)}{P(x)} \propto P(x|g) P(g) \]

\( P(g) \), the probability of observing an individual/unit from group \( g \) is called the **prior** probability.

\( P(g|x) \), the probability that an individual/unit belongs to group \( g \) given that we have observed \( x \) is called the **posterior** probability.

It can be shown that the allocation rule that makes the smallest expected number of errors chooses the class with maximal posterior probability; this is known as the **Bayes rule**.
A linear discriminant analysis can be performed in R using the command `lda`. (Note: This command is provided by the package MASS.)

This command can either be called using a formula as first argument

```
lda(formula, data, ...)
```

or, with a matrix or data frame as first argument

```
lda(x, grouping, ...)
```
Linear discriminant analysis (cont.)

`lda()` returns an object of class “lda” to which we can apply various methods.

`coef(object, ...):`
Returns a matrix with the coefficients of the linear discriminants.

`plot(object, ..., dimen):`
Plots the linear discriminants. `dimen` controls how many discriminants should be printed.
predict(object, newdata, prior, dimen, ...):
Returns a list with three components:

- **class**: the maximum *a posteriori* (MAP) classification.
- **posterior**: posterior probabilities for the classes.
- **x**: the scores on the linear discriminants.

If `newdata` is not specified, then the original data is used.

`dimen` controls how many linear discriminants are used to calculate the components `posterior` and `x` of the result.

**Note**: The linear discriminants are usually centred to have mean zero on the data set.
Example: Sparrow Data

```r
>library(MASS)
>sparrow<-read.table("sparrow.txt",header=T)
>spar.lda<-lda(Survival~TotalLength+AlarExtent+BeakHead
    +Humerus+ Sternum,data=sparrow)
>spar.lda
```
Sparrow lda output

Call:
lda(Survival ~ TotalLength + AlarExtent + BeakHead + Humerus + Sternum, data = sparrow)

Prior probabilities of groups:

N   Y
0.5714286 0.4285714

Group means:

<table>
<thead>
<tr>
<th>TotalLength</th>
<th>AlarExtent</th>
<th>BeakHead</th>
<th>Humerus</th>
<th>Sternum</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>158.4286</td>
<td>241.5714</td>
<td>31.47857</td>
<td>18.44643</td>
</tr>
<tr>
<td>Y</td>
<td>157.3810</td>
<td>241.0000</td>
<td>31.43333</td>
<td>18.50000</td>
</tr>
</tbody>
</table>
Coefficients of linear discriminants:

LD1
TotalLength  -0.3202026
AlarExtent   -0.0546101
BeakHead     -0.1914251
Humerus      2.1284358
Sternum      0.1429131
> coef(spar.lda)

       LD1
TotalLength    -0.3202026
AlarExtent     -0.0546101
BeakHead       -0.1914251
Humerus        2.1284358
Sternum        0.1429131
lda Output (tcd)

> plot(spar.lda)
> spar.pr<-predict(spar.lda)
> names(spar.pr)
[1] "class"   "posterior" "x"
> table(sparrow$Survival,spar.pr$class)

N  Y
N 24 4
Y 14 7

Interpretation: The rows give the actual survivals and the columns give the predicted or fitted survivals according to the linear discriminant analysis classification rule. We can see that there were 28 non-survivors, of which 24 have been correctly classified and 4 incorrectly. Similarly, there were 21 survivors of which 14 have been incorrectly classified.
Misclassifications

There are 18 (out of 49, that is, 36.7%) misclassifications. In the predicted data, there are 34 non-survivors and 11 survivors. On the basis of the measurements, linear discriminant analysis is reasonably successful at determining the correct groupings. That is, the measurements are useful in determining which birds survive.
We can use the `predict` command to apply the classification rule to any data (the original or new data). But it would be also useful to have a mathematical definition or visual representation of the classification rule.

The classification rule based on one linear discriminant is of the form: If LD1 is greater than \( a \) then classify as Y, otherwise classify as N where \( a \) is the threshold value. The value of LD1 is given by the `predict` command. How do we find the threshold \( a \)? One way is the following:

1. Generate a sequence of closely spaced values of the LD1, covering the range of values of LD1 in the data.

2. Use the `predict` command to compute the classification for each of these possible values, and thus estimate the threshold \( a \).

See laboratory class for details.
It would be good to examine a plot of the classifications against the actual survival. This can be produced by:

```r
> plot(index, spar.pr$x, pch=16, xlim=c(0,52), ylim=c(-2,3))
> chw<-par()$cxy[1]
> chh<-par()$cxy[2]
> text(x=index+chw, y=spar.pr$x+chh, labels=spar.pr$class, col="red", adj=0)
> text(x=index+chw, y=spar.pr$x+chh, labels=sparrow$Survival, adj=1)
> abline(h=0.98)
> title(main = "red=LDA classification, black=actual", adj=0)
```
red=LDA classification, black=actual
Interpretation

This classification is based on only one linear discriminator, as there are only two classes (Y or N) to choose from. For problems with more than two levels to allocate to, more than one linear discriminator will be available. Thus several thresholds will also be required.

In this case, the threshold seems to be a little less than 1. Observations above this threshold are classified as Y, while those below are classified as N. A count indicates that 4 of the N and 14 of the Y have been misclassified, confirming the earlier observation from the table.