11 Discriminant Analysis

In MANOVA (or the T^2) have a set of groups (or one group for the T^2) and we perform basic hypothesis tests to determine differences between mean vectors of each group (here we base our discussion on the notes of J. Templin, 2005, and A. Wiesner, 2006).

Now we shift the focus of our analysis. That is if we have groups where multivariate data have been collected, two things we are interested in:

1. **Describe group differences** The way we will do this is by trying to find the lowest number of dimensions that are needed to describe group differences, i.e. *discriminant analysis*.
2. **Prediction and allocation of observations to groups** Define functions that are used to assign a multivariate observation to a group, i.e. *classification*.

11.1 Discrimination Basics

Discriminant analysis seeks to find a set of functions that can be used to separate observations into known groups. To use this procedure several elements must be known prior to the analysis:

- The **number of groups**
- A **training data set** with group membership indicators for each subject must be present

Once such an analysis is completed, we will be able to classify new observations without knowing group membership a priori.

Example 1. Consider again the pottery data sampled from four sites: L) Llanedyrn, C) Caldicot, I) Ilse Thornes, and A) Ashley Rails and the concentrations of the following chemical constituents were measured at a laboratory:

- Al: Aluminum
- Fe: Iron
- Mg: Magnesium
- Ca: Calcium

- Na: Sodium

Now suppose that an archaeologist encounters a pottery specimen of unknown origin. To determine possible trade routes, the archaeologist may wish to classify its site of origin.

Example 2. In the wine data set (Coghlan, 2010), we have 13 chemical concentrations describing wine samples from 3 cultivars. The purpose of linear discriminant analysis (LDA) is to find the linear combinations of the original variables (i.e. 13 chemical concentrations) that gives the best possible separation between the groups (wine cultivars here) of wine. What is the maximum number of useful discriminant functions that can separate the wines by cultivar?

Statistical Likelihoods The primary portion of the discriminant function is the statistical likelihood. This likelihood is represented by using the probability density function of a statistical distribution (the functional form of the distribution).

To demonstrate, consider the plot of the following two distributions:

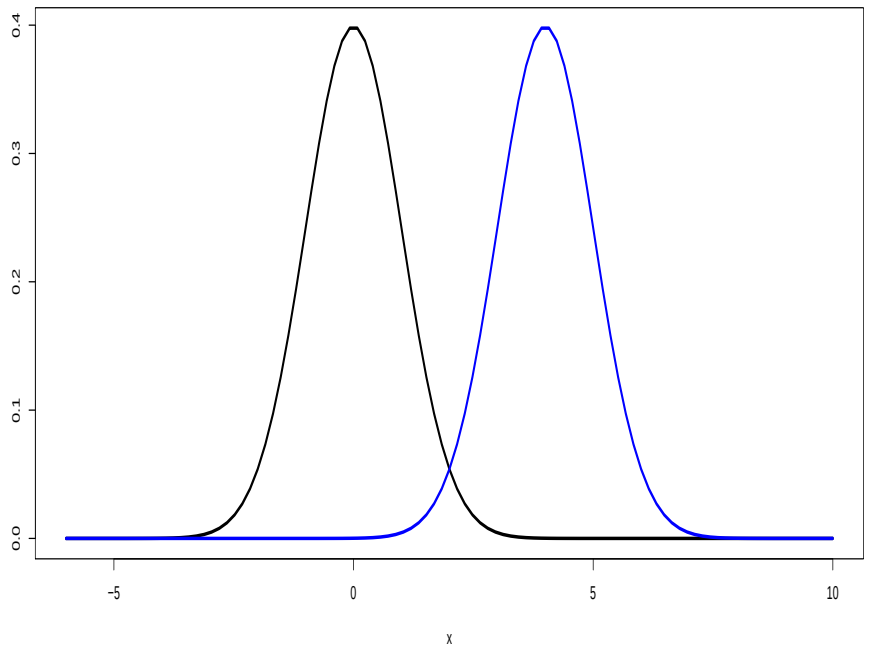

Figure 11.1: The plots of Distribution 1 $N(0, 1)$ (black) and Distribution 2 $N(4, 1)$ (blue).

Functionally, the likelihood for an observation x of being in Distribution 1 is given by the

normal density:

$$f_1(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

and the likelihood for an observation x of being in Distribution 2 is given by the normal density:

$$f_2(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-4)^2/2}.$$

Imagine we encounter an observation, $x = 3$, what would the likelihood be for this observation coming from Distribution 1 or 2?

The likelihood of this observation coming from Distribution 1 is

$$f_1(3) = \frac{1}{\sqrt{2\pi}} e^{-3^2/2} = 0.004$$

and the likelihood for an observation x of being in Distribution 2 is

$$f_2(3) = \frac{1}{\sqrt{2\pi}} e^{-(3-4)^2/2} = 0.242.$$

Hence, we conclude that this observation *most likely* came from Distribution 2.

Notice that the process we just performed is akin to determine the value of the y-axis for the point x.

11.2 The Bayes Rule and Classification Problem

Bayes' Rule Let us consider any two events A and B. To find $P(B|A)$, i.e. the probability that B occurs given that A has happened, we use the Bayes Rule:

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

Now we can apply the Bayes Rule to the Classification Problem. Suppose that we are interested in $P(\pi|\mathbf{x})$, i.e. the conditional probability that an observation came from population π given the observed values of the multivariate vector of variables $\mathbf{x}$. We classify an observation to the population for which the value of $P(\pi|\mathbf{x})$ is greatest. This is the most probable group given the observed values of $\mathbf{x}$.

- Suppose that we have g populations (groups) and that the i-th population is denoted as π_i.

- Let $p_i = P(\pi_i)$ be the probability that a randomly selected observation is in population π_i.
- Let $f(\mathbf{x}|\pi_i)$ be the conditional probability density function of the multivariate set of variables $\mathbf{x}$, given that observation came from population π_i.

Using the notation of the Bayes Rule above, event A is an event of observing the vector $\mathbf{x}$ and event B is an event that the observation belongs to population π_i. Hence, our probability of interest is

$$P(\text{member of } \pi_i|\text{we observed } \mathbf{x}) = \frac{P(\text{ member of } \pi_i \text{ and we observe } \mathbf{x})}{P(\text{we observe } \mathbf{x})}.$$

The numerator of this expression is the likelihood that a randomly selected observation is both from population π_i and has the value $\mathbf{x}$. and this likelihood is $p_i f(x|\pi_i)$. The denominator is the unconditional likelihood (over all populations) that we could observe $\mathbf{x}$. This likelihood is equal to $\sum_{j=1}^{g} p_j f(\mathbf{x}|\pi_j)$.

Hence, the posterior probability that an observation is a member of population π_i is

$$p(\pi_i|\mathbf{x}) = \frac{p_i f(\mathbf{x}|\pi_i)}{\sum_{j=1}^{g} p_j f(\mathbf{x}|\pi_j)} \propto p_i f(\mathbf{x}|\pi_i).$$

The **classification rule** is to assign observation $\mathbf{x}$ to the population for which the posterior probability is the greatest.

Remark Notice that the denominator is the same for all posterior probabilities (across the various populations) and it does not depend on the population since it involves summing over all the populations. Hence, equivalently we can say that we classify an observation to the population for which $p_i f(x|\pi_i)$ is greatest.

Two Populations With only two populations we can express a **classification rule** in terms of the ratio of the two posterior probabilities. In particular, we classify to population 1 if

$$\frac{p_1 f(\mathbf{x}|\pi_1)}{p_2 f(\mathbf{x}|\pi_2)} > 1,$$

i.e. we classify to population 1 when

$$\frac{f(\mathbf{x}|\pi_1)}{f(\mathbf{x}|\pi_2)} > \frac{p_2}{p_1}.$$

General Decision Rule We classify the sample unit or subject into the population π_i that maximizes the posterior probability $p(\pi_i)$, i.e. the population which maximizes

$$f(\mathbf{x}|\pi_\mathbf{i})p_i.$$

We calculate **the posterior probabilities** for each of the populations. Then we assign the subject or sample unit to the population which has the highest posterior probability. Ideally that posterior probability is going to be greater than a half, the closer to 100% the better!

Equivalently we assign $\mathbf{x}$ to the population that maximizes the product:

$$\log f(\mathbf{x}|\pi_i)p_i,$$

due to simplicity of calculating a logarithm in many cases.

11.3 Discriminant Analysis Procedure

We can summarize the discriminant analysis as a 7-step procedure:

1. **Step 1** Collect ground truth or training data with known group memberships.
2. **Step 2** The prior probability p_i represents the expected portion of the community that belongs to population π_i. There are three common choices:
   (a) **Equal priors:**
   $$\hat{p}^i = \frac{1}{g}.$$
   This would be used if we believe that all of the population sizes are equal.
   (b) **Subjective priors** selected according to the investigators beliefs regarding the relative population sizes. Note that we require:
   $$\hat{p}^1 + \hat{p}^2 + \ldots + \hat{p}^g = 1.$$
   (c) **Estimated priors**:
   $$\hat{p}^i = \frac{n_i}{N},$$

where n_i is the number observations from population π_i in the training data, and $N = n_1 + n_2 + \ldots + n_g$.

3. **Step 3** Use multivariate heteroscedasticity tests to determine if variance-covariance matrices are homogeneous for the two or more populations involved. Result of this test determines whether to use Linear or Quadratic Discriminant Analysis.

   - **Case 1** ***Linear discriminant analysis (LDA)*** is for homogeneous variance-covariance matrices:

     $$\Sigma_1 = \Sigma_2 = \ldots = \Sigma_g = \Sigma.$$

     In this case the variance-covariance matrix does not depend on the population from which the data are obtained.

   - **Case 2** ***Quadratic discriminant analysis (QDA)*** is used for heterogeneous variance-covariance matrices:

     $$\Sigma_i = \Sigma_j, \quad i \neq j$$

     This allows the variance-covariance matrices to depend on which population we are looking at.

4. **Step 4** Estimate the parameters of the conditional probability density functions $f(\mathbf{x}|\pi_i)$. Here, we make the following standard assumptions as in MANOVA:

   - The data from group i has common mean vector $\boldsymbol{\mu}_i$
   - The data from group i has common variance-covariance matrix Σ.
   - The subjects are independently sampled
   - The data are multivariate normally distributed

5. **Step 5** Compute discriminant functions. This is the rule for classification of the new object into one of the known populations.

6. **Step 6** Use cross validation to estimate misclassification probabilities.

   As in all statistical procedures it is helpful to use diagnostic procedures to asses stability of the discriminant analysis, e.g. crossvalidation.

   Typically we have some prior rule as to what is an acceptable misclassification rate and these rules might involve questions like: **"what is the cost of misclassification?"**

7. **Step 7** Classify observations with unknown group memberships.

11.4 Linear, or Canonical, Discriminant Analysis (LDA)

We assume that in population π_i the probability density function of $\mathbf{x}$ is multivariate normal with mean vector $\boldsymbol{\mu}_i$ and variance-covariance matrix Σ (same for all populations), i.e.

$$f(\mathbf{x}|\pi_i) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right].$$

We classify to the population for which $p_i f(\mathbf{x}|\pi_i)$ is largest. Since a logarithm transform is monotonic, this equivalent to classifying an observation to the population for which $\log p_i f(\mathbf{x}|\pi_i)$ is largest.

Linear discriminant analysis (LDA) is used when the variance-covariance matrix does not depend on the population from which the data are obtained. In this case, our decision rule is based on the so-called **Linear Score Function** which is a function of the population means for each of our g populations $\boldsymbol{\mu}_i$, as well as the pooled variance-covariance matrix.

The **Linear Score Function** is:

$$s_i^L(\mathbf{x}) = -\frac{1}{2}\boldsymbol{\mu}_i'\Sigma^{-1}\boldsymbol{\mu}_i + \boldsymbol{\mu}_i'\Sigma^{-1}\mathbf{x} + \log p_i.$$

We define the related **Linear Discriminant Function** as

$$d_i^L(\mathbf{x}) = -\frac{1}{2}\boldsymbol{\mu}_i'\Sigma^{-1}\boldsymbol{\mu}_i + \boldsymbol{\mu}_i'\Sigma^{-1}\mathbf{x}$$

Notice that this expression resembles a linear regression with intercept term $-\frac{1}{2}\boldsymbol{\mu}_i'\Sigma^{-1}\boldsymbol{\mu}_i$ and slope $\boldsymbol{\mu}_i'\Sigma^{-1}$.

Given a sample unit with measurements $x_1, x_2, \ldots, x_p$, we classify the sample unit into the population that has the largest **Linear Score Function**. This is equivalent to classifying to the population for which the posterior probability of membership is largest. The linear score function is computed for each population, then we assign the unit to the population with the largest score.

However, this is a function of unknown parameters $\boldsymbol{\mu}_i$ and Σ. Discriminant analysis requires estimates of the following parameters are estimated from training data, in which the population membership is known.

- Prior probabilities

$$p_i = P(\pi_i), \quad i = 1, 2, \ldots, g$$

  are estimated in Step 2 above.

- The Population Means

$$\boldsymbol{\mu}_i = E(\mathbf{x}|\pi_i), \quad i = 1, 2, \ldots, g$$

  can be estimated by the sample mean vectors the sample means $\bar{\mathbf{x}}_i$

- The Variance-covariance matrix

$$\Sigma = var(\mathbf{x}|\pi_i), \quad i = 1, 2, \ldots, g$$

  can be estimated by using the pooled variance-covariance matrix

$$S_p = \frac{\sum_{i=1}^{g}(n_i - 1)S_i}{\sum_{i=1}^{g}(n_i - 1)}.$$

Now substituting these estimates into the Linear Score Function as shown yields:

$$\hat{s}_i^L(\mathbf{x}) = -\frac{1}{2}\bar{\mathbf{x}}_i' S_p^{-1}\bar{\mathbf{x}}_i + \bar{\mathbf{x}}_i' S_p^{-1}\mathbf{x} + \log \hat{p}_i$$

This is a function of the sample mean vectors, the pooled variance-covariance matrix and prior probabilities for g different populations.

Decision Rule Classify the sample unit into the population that has the largest estimated linear score function.

Example Now let us illustrate LDA by considering the wine example (Coghlan, 2010).

```
> library(MASS)

> wine.lda <- lda(wine$V1 ~ wine$V2 + wine$V3 + wine$V4 + wine$V5 + wine$V6 + wine$V7 +
wine$V8 + wine$V9 + wine$V10 + wine$V11 + wine$V12 + wine$V13 + wine$V14)
> wine.lda
Call:
lda(wine$V1 ~ wine$V2 + wine$V3 + wine$V4 + wine$V5 + wine$V6 +
wine$V7 + wine$V8 + wine$V9 + wine$V10 + wine$V11 + wine$V12 +
wine$V13 + wine$V14)

Prior probabilities of groups:
        1         2         3
0.3314607 0.3988764 0.2696629

Group means:
   wine$V2  wine$V3  wine$V4  wine$V5  wine$V6  wine$V7
1 13.74475 2.010678 2.455593 17.03729 106.3390 2.840169
2 12.27873 1.932676 2.244789 20.23803  94.5493 2.258873
3 13.15375 3.333750 2.437083 21.41667  99.3125 1.678750

   wine$V8    wine$V9   wine$V10 wine$V11
1 2.9823729 0.290000 1.899322 5.528305
2 2.0808451 0.363662 1.630282 3.086620
3 0.7814583 0.447500 1.153542 7.396250

   wine$V12 wine$V13  wine$V14
1 1.0620339 3.157797 1115.7119
2 1.0562817 2.785352  519.5070
3 0.6827083 1.683542  629.8958

Coefficients of linear discriminants:
```

```
                  LD1           LD2
wine$V2  -0.403399781  0.8717930699
wine$V3   0.165254596  0.3053797325
wine$V4  -0.369075256  2.3458497486
wine$V5   0.154797889 -0.1463807654
wine$V6  -0.002163496 -0.0004627565
wine$V7   0.618052068 -0.0322128171
wine$V8  -1.661191235 -0.4919980543
wine$V9  -1.495818440 -1.6309537953
wine$V10  0.134092628 -0.3070875776
wine$V11  0.355055710  0.2532306865
wine$V12 -0.818036073 -1.5156344987
wine$V13 -1.157559376  0.0511839665
wine$V14 -0.002691206  0.0028529846

Proportion of trace:
   LD1    LD2
0.6875 0.3125
```

This means that the first discriminant function is a linear combination of the variables: $-0.403V_2 + 0.165V_3 - 0.369V_4 + 0.155V_5 - 0.002V_6 + 0.618V_7 - 1.661V_8 - 1.496V_9 + 0.134V_{10} + 0.355V_{11} - 0.818V_{12} - 1.158V_{13} - 0.003V_{14}$, where $V_2, V_3, \ldots, V_{14}$ are the concentrations of the 13 chemicals found in the wine samples.

To extract the loadings for the first discriminant function, we use

```
> wine.lda$scaling[,1]
     wine$V2      wine$V3      wine$V4      wine$V5      wine$V6      wine$V7
-0.403399781  0.165254596 -0.369075256  0.154797889 -0.002163496  0.618052068
     wine$V8
```

```
-1.661191235

  wine$V9     wine$V10     wine$V11     wine$V12     wine$V13     wine$V14
-1.495818440  0.134092628  0.355055710 -0.818036073 -1.157559376 -0.002691206
```

The *proportion of trace* refers to the percentage separation achieved by each discriminant function. For example, for the wine data we get

```
Proportion of trace:
   LD1    LD2
0.6875 0.3125
```

i.e. 0.69% and 0.31%.

Hence, the first LD function does achieve a good separation between the three cultivars, but the second discriminant function does improve the separation of the groups by quite a large amount. Thus, it is worth using the second discriminant function as well.

Visualization of the LDA results A useful way of displaying the results of LDA is a stacked histogram of the values of the discriminant function for the samples from different groups (i.e. different wineries in our example).

Let us first obtain the value of the each discriminant function:

```
> wine.lda.values <- predict(wine.lda, wine[2:14])

#wine.lda.values$x[,1] ## gives us values of each wine for the 1st discriminant function
> length(wine.lda.values$x[,1])
[1] 178

#wine.lda.values$x[,2] ## gives us values of each wine for the 2nd discriminant function
> length(wine.lda.values$x[,2])
[1] 178
```

```
>ldahist(data = wine.lda.values$x[,1], g=wine$V1)
```

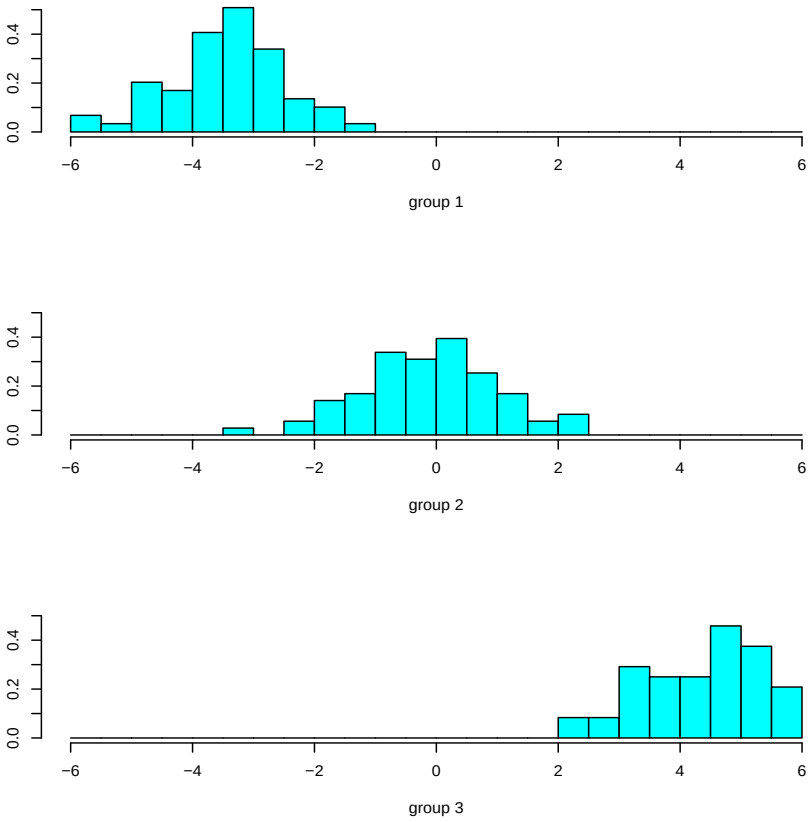

Figure 11.2: Stacked histogram of the 1st discriminant function for the wine data.

We can see from Fig. 11.2 that cultivars 1 and 3 are well separated by the 1st discriminant function, since the values for the first cultivar are between -6 and -1, while the values for cultivar 3 are between 2 and 6, and so there is no overlap in values.

Now let us see how the 2nd discriminant function performs:

```
>ldahist(data = wine.lda.values$x[,2], g=wine$V1)
```

As Fig. 11.3 shows, the 2nd discriminant function separates cultivars 1 and 2 quite well, although there exists a slight overlap in their values. Also, the 2nd discriminant function separates cultivars 2 and 3 quite well, although again there is a some overlap in their values.

We now construct a scatterplot of the best two discriminant functions, with the data points labelled by cultivar.

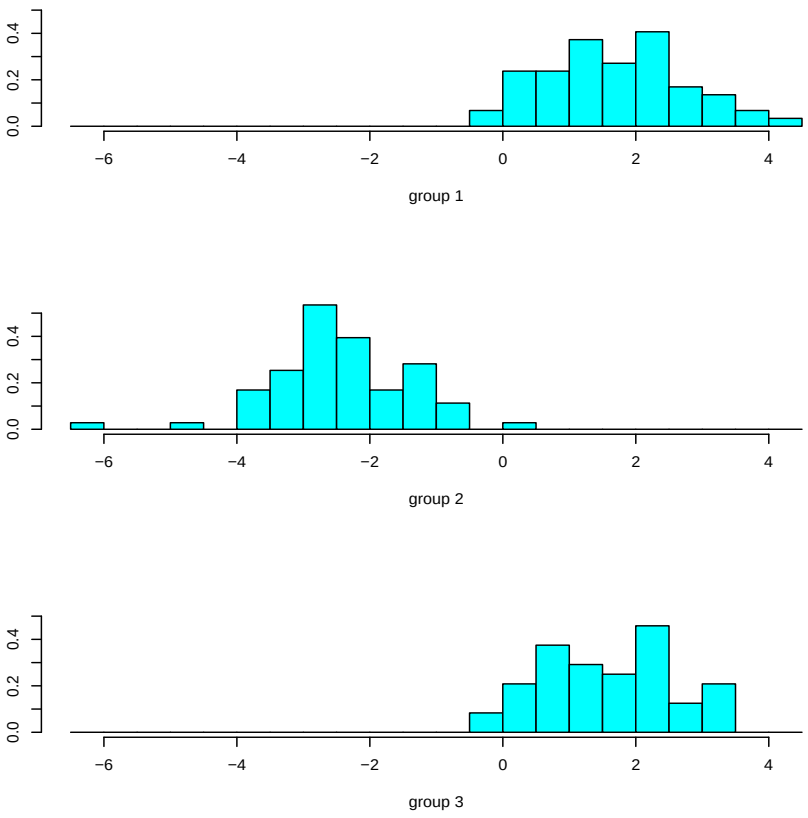

Figure 11.3: Stacked histogram of the 2nd discriminant function for the wine data.

```
> plot(wine.lda, xlab="The 1st discriminant function",
ylab="The 2nd discriminant function")
```

Fig 11.4 indicates that the wines from the three cultivars are well separated. The 1st discriminant function (x-axis) separates cultivars 1 and 3, but does not separate well cultivars 1 and 2, or cultivars 2 and 3. In contrast, the 2nd discriminant function (y-axis) delivers a fairly good separation of cultivars 1 and 3, and cultivars 2 and 3, although the separation is not totally perfect. Hence, to achieve a better separation of the three cultivars, it would be worthwhile to use both the discriminant functions together.

Allocation Rules and Misclassification Rate We can now calculate the mean values of the discriminant functions for each of the three wine groups.

```
> ###### the 1st discriminant function ######
```

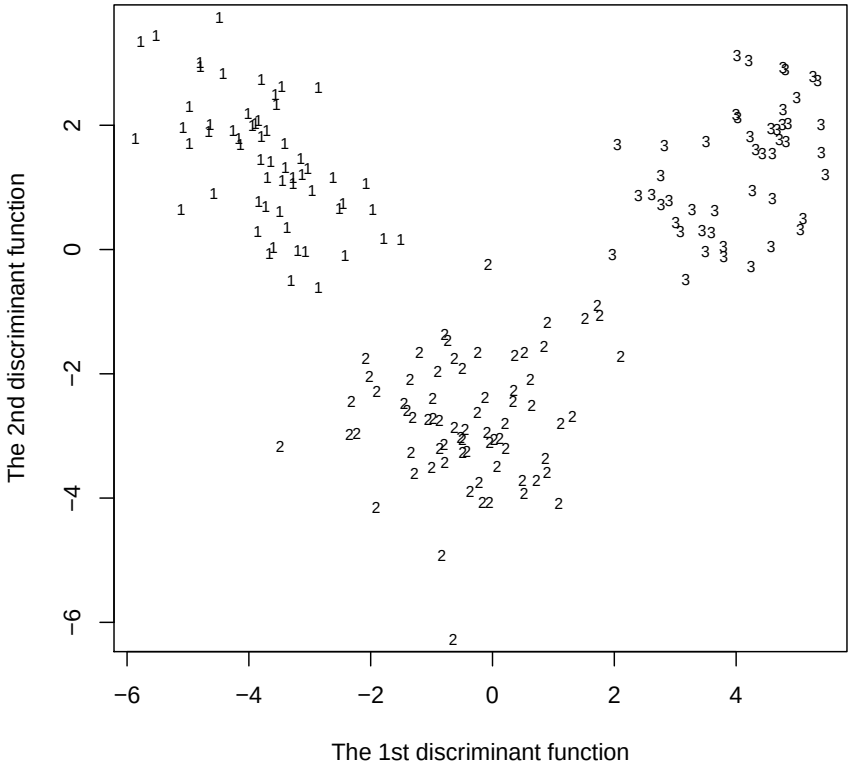

Figure 11.4: Scatter plot of linear discriminant analysis for the wine data.

```
> c(mean(wine.lda.values$x[which(wine[1]==1),1]),
mean(wine.lda.values$x[which(wine[1]==2),1]),
mean(wine.lda.values$x[which(wine[1]==3),1]))
[1] -3.42248851 -0.07972623  4.32473717

> ###### the 2nd discriminant function ######
> c(mean(wine.lda.values$x[which(wine[1]==1),2]),
mean(wine.lda.values$x[which(wine[1]==2),2]),
mean(wine.lda.values$x[which(wine[1]==3),2]))
[1]  1.691674 -2.472656  1.578120
```

I.e., we find that the mean value of the 1st discriminant function is -3.42248851 for cultivar 1, -0.07972623 for cultivar 2, and 4.32473717 for cultivar 3. The mid-way point between the mean

values for cultivars 1 and 2 is $\frac{-3.42248851-0.07972623}{2} = -1.751107$, and the mid-way point between the mean values for cultivars 2 and 3 is $\frac{-0.07972623+4.32473717}{2} = 2.122505$.

Hence, based on the 1st discriminant function, we can use the following allocation rule:

- if the value of the 1st discriminant function $\in (-\infty, -1.751107]$, predict the sample to be from cultivar 1
- if the value of the 1st discriminant function $\in (-1.751107, 2.122505]$, predict the sample to be from cultivar 2
- if the value of the 1st discriminant function $\in (2.122505, \infty)$, predict the sample to be from cultivar 3

We can then allocate all wines to cultivars and construct the respective confusion matrix:

	Our classification		
Truth	Group 1	Group 2	Group 3
Group 1	56	3	0
Group 2	5	65	1
Group 3	0	0	48

There are 9 wine samples that are misclassified, out of 178 wine samples. Hence, the misclassification rate is 9/178, or 5.1%.

However, this is only **in-sample misclassification rate** and ideally we need to repeat the classification procedure for some testing test and calculate **out-of-sample misclassification rate**.

11.5 Quadratic Discriminant Analysis

Recall that the Linear Discriminant Analysis is for **homogeneous variance-covariance matrices**. However, if data exhibit heterogeneity, i.e. $\Sigma_i \neq \Sigma_j, i \neq j$, we can employ the Quadratic Discriminant Analysis (QDA).

Quadratic discriminant analysis calculates a Quadratic Score Function which looks like this:

$$s_i^Q(\mathbf{x}) = -\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \log p_i,$$

i.e., which is a function of population mean vectors and the variance-covariance matrices for i-th group. Similarly as in LDA, we determine a separate quadratic score function for each of the groups, $i = 1, \ldots, g$.

Again, as in LDA, Quadratic Score is a function of unknown population mean vector and the variance-covariance matrix for group i, and we estimate these parameters from the sample training set data. Hence, our estimated quadratic score function take a form:

$$s_i^Q(\mathbf{x}) = -\frac{1}{2}\log|S_i| - \frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}}_i)^T S_i^{-1}(\mathbf{x} - \bar{\mathbf{x}}_i) + \log \hat{p}_i,$$

Decision Rule Our decision rule remains the same, i.e. we classify the sample unit or subject into the population that has the largest quadratic score function.