5 Discriminant Analysis

5.1 Introduction

In the very simplest type of discrimination problem, we want to allocate an individual or sampling unit to one of two populations; this allocation is to be based on a set of measurements associated with the individual. For e.g., a doctor takes a series of measurements on a patient, on the basis of which, a diagnosis is made.

Consider the following example.
Five diagnostic tests were carried out on two groups of sheep: those suffering from Scrapie, and the other from a more serious disease (possibly CJD). The measurements arising from the tests for each sheep are summarized in the tables below.

Sheep with Scrapie:

<table>
<thead>
<tr>
<th></th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>18</td>
<td>15</td>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>33</td>
<td>27</td>
<td>31</td>
<td>21</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>28</td>
<td>27</td>
<td>23</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>26</td>
<td>18</td>
<td>18</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>22</td>
<td>23</td>
<td>22</td>
<td>16</td>
<td>10</td>
</tr>
</tbody>
</table>

Sheep with more serious disease:

<table>
<thead>
<tr>
<th></th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td>17</td>
<td>20</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>24</td>
<td>31</td>
<td>26</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>16</td>
<td>17</td>
<td>20</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>24</td>
<td>31</td>
<td>26</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>36</td>
<td>28</td>
<td>24</td>
<td>26</td>
<td>29</td>
</tr>
</tbody>
</table>

Can we construct a rule that would allow us to distinguish between the two diseases on the basis of the results of the tests (T1-T5)?

5.2 Problem Formulation and Decision Rule

Suppose that we have two populations, population 1 and population 2.

Let the distributional behaviour of observations from population 1 be characterized by \( f_1(\cdot) \),
and for population 2 by $f_2(\cdot)$.

Let $\pi_i$ be the (prior) probability that a randomly selected individual, from the 'population at large', belongs to population $i$, $i = 1, 2$.

Also, let $C(i|j)$ be the cost of misallocating an individual from population $j$ to population $i$.

A possible allocation rule (based on decision-theoretic considerations) says that

$$\text{if } \frac{f_1(x)}{f_2(x)} \geq \frac{\pi_2 C(1|2)}{\pi_1 C(2|1)}$$

then allocate individual with measurement $x$ to population 1; otherwise, allocate to population 2.

Remarks 5.1
(i) If $\pi_2 C(1|2) = \pi_1 C(2|1)$, then (1) becomes

$$\frac{f_1(x)}{f_2(x)} \geq 1$$

or $f_1(x) \geq f_2(x)$, i.e. allocate to the population with the greatest likelihood.

(ii) When $\pi_2 C(1|2) \neq \pi_1 C(2|1)$ then the decision takes into account:
• the ratio of the misallocation costs;
• the ratio of the prior probabilities.

Example 5.2
A healthcare centre is trying to screen patients in a certain area for a particular disease.

It is widely believed that about 2% of the population have the disease.

Not detecting the disease is considered to be ten times more serious than initially classifying a healthy individual diseased.

Formulate an appropriate allocation rule, based on $x \in \mathbb{R}^p$ taken on an individual, and the above information.

Solution: Let $f_H(\cdot)$ be the p.d.f. associated with the 'healthy' population, and $f_D(\cdot)$ be that associated with the 'diseased'.

Decision rule is:

$$\text{if } \frac{f_H(x)}{f_D(x)} \geq \frac{0.02}{0.98} \times \frac{10}{1} \approx 0.2$$

then the individual with readings given by $x$ is classified as healthy. The individual is thus classified as unhealthy if

$$\frac{f_H(x)}{f_D(x)} \geq \frac{1}{5} \text{ i.e. } \frac{f_D(x)}{f_H(x)} < 5.$$
Thus, we do not recall the individual for further tests unless the likelihood ratio in favour of allocation to the diseased population is more than 5 to 1.

5.3 Form of the Decision Rule under Multivariate Normality

Let us now suppose that readings from the \(i\)-th population have the \(N_p(\mu_i, \Sigma)\) distribution, \(i = 1, 2\). (So the two populations share the same covariance matrix). Then

\[
f_i(x) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2}(x - \mu_i)'\Sigma^{-1}(x - \mu_i) \right\}
\]

Hence

\[
\frac{f_1(x)}{f_2(x)} = \exp \left\{ -\frac{1}{2}(x - \mu_1)'\Sigma^{-1}(x - \mu_1) + \frac{1}{2}(x - \mu_2)'\Sigma^{-1}(x - \mu_2) \right\}
\]

i.e.

\[
\ln \left( \frac{f_1(x)}{f_2(x)} \right) = -\frac{1}{2} [(x - \mu_1)'\Sigma^{-1}(x - \mu_1) - (x - \mu_2)'\Sigma^{-1}(x - \mu_2)] .
\]

The terms in the square brackets can be expanded and shown to be equal to

\[
x'\Sigma^{-1}x - 2x'\Sigma^{-1}\mu_1 + \mu_1'\Sigma^{-1}\mu_1 - x'\Sigma^{-1}x + 2x'\Sigma^{-1}\mu_2 - \mu_2'\Sigma^{-1}\mu_2
\]

which upon simplification gives

\[-2x'\Sigma^{-1}(\mu_1 - \mu_2) + \mu_1'\Sigma^{-1}\mu_1 - \mu_2'\Sigma^{-1}\mu_2
\]

i.e.

\[-2x'\Sigma^{-1}(\mu_1 - \mu_2) + (\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2).
\]

Plugging this expression back into (5) yields:

\[
\ln \left( \frac{f_1(x)}{f_2(x)} \right) = x'\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2).
\]

Setting \(L = \Sigma^{-1}(\mu_1 - \mu_2)\) in the above gives

\[
\ln \left( \frac{f_1(x)}{f_2(x)} \right) = x'L - \frac{1}{2}(\mu_1 + \mu_2)'L
\]

\[= L'x - \frac{1}{2}L'(\mu_1 + \mu_2).
\]

Now (1) holds if, and only if,

\[
\ln \left( \frac{f_1(x)}{f_2(x)} \right) \geq \ln \left( \frac{\pi_2C(1|2)}{\pi_1C(2|1)} \right)
\]

i.e.

\[L'x - \frac{1}{2}L'(\mu_1 + \mu_2) \geq k
\]

where

\[k = \ln \left( \frac{\pi_2C(1|2)}{\pi_1C(2|1)} \right)
\]
Remarks 5.3
(i) Allocate individual with observation $x$ to population 1 if
\[ L'x - \frac{1}{2} L'(\mu_1 + \mu_2) \geq k \]
and to population 2 otherwise.

(ii) The L.H.S. of the above is known as the **linear discriminant function**. The unknown parameters can be replaced by appropriate estimators from the sample, yielding the **sample linear discriminant function** (see later).

(iii) $k = 0 \iff \pi_2 C(1|2) = \pi_1 C(2|1)$.

5.4 Misclassification

We can quantify the probability of *misclassification* in certain special cases.

Suppose $L'\mu_1 > L'\mu_2$, and define
\[ \alpha = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) = L'(\mu_1 - \mu_2). \]

Also, define the scalar r.v.
\[ U_1 = L'X_1 - \frac{1}{2} L'(\mu_1 + \mu_2) \]
where $X_1 \sim N_p(\mu_1, \Sigma)$.

Then
\[ E[U_1] = L'\mu_1 - \frac{1}{2} L'(\mu_1 + \mu_2) \]
\[ = \frac{1}{2} L'(\mu_1 - \mu_2) = \frac{\alpha}{2} \]
and
\[ \text{var}(U_1) = \text{var}(L'X_1 - \frac{1}{2} L'(\mu_1 + \mu_2)) \]
\[ = \text{var}(L'X_1) = L'\Sigma L \]
\[ = (\mu_1 - \mu_2)' \Sigma^{-1} \Sigma \Sigma^{-1} (\mu_1 - \mu_2) \]
\[ = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) = \alpha \]

Since $X_1$ has a MVN distribution, then the linear compound $L'X_1$ is Normally distributed, as is $U_1$.

Hence $U_1 \sim N(\frac{\alpha}{2}, \alpha)$. 

4
Define
\[ U_2 = L'X_2 - \frac{1}{2}L'\left(\mu_1 + \mu_2\right) \]
where \( X_2 \sim N_p(\mu_2, \Sigma) \).

Then
\[ E[U_2] = L'\mu_2 - \frac{1}{2}L'\left(\mu_1 + \mu_2\right) = -\frac{\alpha}{2} \]
and
\[ \text{var}(U_2) = \text{var}(L'X_2) = L'\Sigma L = \alpha. \]

Thus, \( U_2 \sim N\left(-\frac{\alpha}{2}, \alpha\right) \).

---

**DIAGRAM¹**

Let \( p(i|j) \) be the probability that an individual from population \( j \) is misallocated to population \( i \).

Then
\[ p(1|2) = \mathbb{P}(U_2 > k) \]
\[ p(2|1) = \mathbb{P}(U_1 < k). \]

Hence, the total probability of misclassification is
\[ \pi_1 p(2|1) + \pi_2 p(1|2) \]
i.e.
\[ \pi_1 \mathbb{P}(U_1 < k) + \pi_2 \mathbb{P}(U_2 > k). \]

When \( k = 0 \), this becomes
\[ \pi_1 \mathbb{P}(U_1 < 0) + \pi_2 \mathbb{P}(U_2 > 0) \]
\[ = \pi_1 \mathbb{P}(U_1 < 0) + \pi_2 \mathbb{P}(U_1 < 0) \]
\[ = (\pi_1 + \pi_2)\mathbb{P}(U_1 < 0) = \mathbb{P}(U_1 < 0) \]
\[ = \mathbb{P}\left(\frac{U_1 - \frac{\alpha}{2}}{\sqrt{\alpha}} < -\frac{\alpha/2}{\sqrt{\alpha}}\right) = \Phi\left(\frac{-\sqrt{\alpha}}{2}\right). \]

¹To be copied down in the lecture, or from p.135 of Chatfield and Collins.
5.5 Practicalities

Assumptions regarding the functional form of the distribution from which random samples are drawn were asserted; in particular, multivariate Normality has been assumed in a number of places.

However, in practice, we may know nothing about some or all of the key parameters of the distribution, in which case, these have to be estimated from the available data. Thus, for e.g., a point estimate for \( \mu_i \) could be based on \( x_i \); also \( \Sigma \) could be estimated by \( S \), the pooled 'within-groups' sample covariance matrix. Invoking such estimates would lead to an allocation rule of the form:

\[
\hat{L}'x - \frac{1}{2}\hat{L}'(x_1 + x_2) \geq k
\]

then allocate to population 1, o.w. allocate to population 2, where \( \hat{L} = S^{-1}(x_1 - x_2) \).

Returning to our earlier example, taking \( k = 0 \), our discrimination procedure, calculated on the basis of the available data, can be found as follows:

```r
> s.t1 <- c(11, 33, 20, 18, 22)
> s.t2 <- c(18, 27, 26, 23)
> s.t3 <- c(15, 31, 27, 18, 22)
> s.t4 <- c(18, 21, 23, 18, 16)
> s.t5 <- c(15, 17, 19, 9, 10)
> u.t1 <- c(18, 31, 14, 25, 36)
> u.t2 <- c(17, 24, 16, 24, 28)
> u.t3 <- c(20, 31, 17, 31, 24)
> u.t4 <- c(18, 26, 20, 26, 26)
> u.t5 <- c(18, 20, 17, 18, 29)
> s.dat <- data.frame(s.t1, s.t2, s.t3, s.t4, s.t5)
> u.dat <- data.frame(u.t1, u.t2, u.t3, u.t4, u.t5)
> s.dat
    s.t1 s.t2 s.t3 s.t4 s.t5
  1   11   18   15   18   15
  2   33   27   31   21   17
  3   20   28   27   23   19
  4   18   26   18   18   9
  5   22   23   22   16   10
> u.dat
    u.t1 u.t2 u.t3 u.t4 u.t5
  1   18   17   20   18   18
  2   31   24   31   26   20
  3   14   16   17   20   17
  4   25   24   31   26   18
  5   36   28   24   26   29
```
```r
> no.s <- 5
> no.u <- 5
> s.pooled <- ((no.s - 1) * var(s.dat) + (no.u - 1) * var(u.dat))/(no.s + no.u - 2)
> dimnames(s.pooled) <- NULL
> s.pooled

[1,] 72.700 33.025 41.65 18.675 22.300
[3,] 41.650 21.300 41.30 16.350 9.850
[4,] 18.675 12.725 16.35 11.450 10.200
> sm.s <- apply(s.dat, 2, mean)
> sm.u <- apply(u.dat, 2, mean)
> names(sm.s) <- NULL
> names(sm.u) <- NULL
> sm.s
[1] 20.8 24.4 22.6 19.2 14.0
> sm.u
[1] 24.8 21.8 24.6 23.2 20.4
> L <- solve(s.pooled) %*% (sm.s - sm.u)
> L

[,1]
[1,] -0.7491324
[2,]  2.0307983
[3,]  0.5350933
[4,] -2.3422912
[5,]  0.2175097

Thus

\[ \mathbf{L} = \begin{bmatrix} -0.7491324, & \ldots, & 0.2175097 \end{bmatrix} \]

Question: Would our rule correctly allocate the sheep (in the sample presented in the example) to the correct populations?

> s.mat <- as.matrix.data.frame(s.dat)
> u.mat <- as.matrix.data.frame(u.dat)
> s.ldr <- t(L) %*% t(s.mat) - 0.5 * t(L) %*% (sm.s + sm.u) *rep(1, 5)
> u.ldr <- t(L) %*% t(u.mat) - 0.5 * t(L) %*% (sm.s + sm.u) *rep(1, 5)
> s.ldr

1 2 3 4 5
[1,] 0.8976995  4.663609 10.04319 12.20038 10.15392
> u.ldr

1 2 3 4 5

The components of \( s.ldr \) are all positive, and those of \( u.ldr \) all negative. Thus, the rule has correctly classified all of the sheep.

7