Illustrations (1–2)

In this segment, two examples of discriminant analysis will used to show how to carry out this procedure with R.

Both linear and quadratic discriminant analysis will be illustrated using the data set `Returns.txt` available on `myCourses`.

These financial data form the `Smarket` dataset in the `ISLR` library.

The analysis will follow closely the presentation made in the textbook *Introduction to Statistical Learning* by James et al. (2013).

Illustrations (2–2)

This data consist of percentage returns for the S&P 500 stock index over 1250 days, from the beginning of 2001 until the end of 2005.

For each date, one has the percentage returns for each of the five previous trading days, `Lag1` through `Lag5`, as well

- ✓ `Volume` (number of shares traded on the previous day, in billions);
- ✓ `Today` (the percentage return on the date in question) and
- ✓ `Direction` (whether the market was Up or Down on this date).

We will try to see whether discriminant analysis can predict today's `Direction` as a function of `Lag1` and `Lag2`. The 2001–04 data will be used as the training sample; the 2005 data will be used for validation.

Linear DA (1–6)

```
# Data input
returns <- read.table("Returns.txt", header = T, sep = "\t")

# Linear discriminant analysis
# Definition of the training and validation samples
train <- (1:1250)[returns$Year<2005]
valid <- (1:1250)[returns$Year>=2005]
library(MASS)

# If no prior is specified, R uses by default the
# proportions observed in the data set
returns.lda <- lda(Direction ~ Lag1 + Lag2,
                   prior = rep(1/2,2),
                   subset = train, data = returns)
```

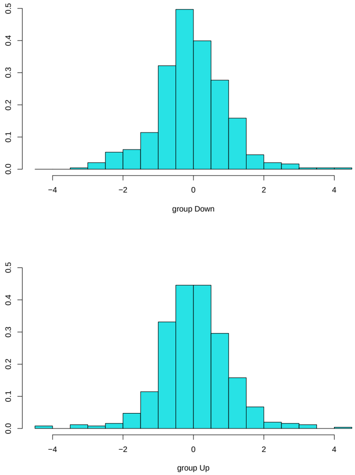
Linear DA (2–6)

```
# Summary of the results
returns.lda

Call:
lda(Direction ~ Lag1 + Lag2, data = returns,
      prior = rep(1/2, 2), subset = train)

Prior probabilities of groups:
Down   Up
 0.5  0.5

Group means:
            Lag1        Lag2
Down  0.04279022  0.03389409
Up   -0.03954635 -0.03132544
```

Linear DA (3–6)

```
# Coefficients of Fisher's discriminant function

returns.lda$scaling

            LD1
Lag1 -0.6420190
Lag2 -0.5135293

# Histogram of the scores based on the discriminant
# function in the two groups...
# The more similar they are, the less likely
# that the predictive power will be good!

plot(returns.lda)
```

Linear DA (4–6)


0.0
0.1
0.2
0.3
0.4
0.5
−4
−2
0
2
4
group Down
0.0
0.1
0.2
0.3
0.4
0.5
−4
−2
0
2
4
group Up

Linear DA (5–6)

```
# Prediction
returns.lda.pred = predict(returns.lda,
                           newdata = returns[valid,])
names(returns.lda.pred)

[1] "class"     "posterior" "x"

returns.lda.group = returns.lda.pred$class
table(returns.lda.group, returns[valid,]$Direction)

returns.lda.groupe Down Up
              Down   64 67
              Up     47 74

mean(returns.lda.group == returns[valid,]$Direction)
[1] 0.547619
```

Linear DA (6–6)

The coefficients of Fisher's discriminant function are

-0.6420190 for `Lag1`;

-0.5135293 for `Lag2`.

The histograms show that the two score distributions are essentially the same, so the predictive power is not expected to be high.

This is confirmed by the confusion table based on the validation data.

With a 55% success rate, the procedure does hardly any better than a random allocation!

Quadratic DA (1–4)

```
# Quadratic discriminant analysis
# Note: The function qda is also in the package MASS

returns.qda <- qda(Direction ~ Lag1 + Lag2,
                   prior = rep(1/2,2),
                   subset = train, data = returns)

# Summary of the results

returns.qda
```

```
Call:
qda(Direction ~ Lag1 + Lag2, data = returns,
      prior = rep(1/2, 2), subset = train)

Prior probabilities of groups:
Down   Up
 0.5  0.5

Group means:
            Lag1        Lag2
Down  0.04279022  0.03389409
Up   -0.03954635 -0.03132544
```

Quadratic DA (3–4)

```
names(returns.qda.pred)

[1] "class"     "posterior"

returns.qda.group = returns.qda.pred$class

table(returns.qda.group, returns[valid,]$Direction)

returns.qda.group Down Up
             Down   55 58
             Up     56 83

mean(returns.qda.group == returns[valid,]$Direction)

[1] 0.547619
```

Quadratic DA (4–4)

The success rate turns out to be exactly the same as for linear discriminant analysis: $138/252 = 54.76\%$.

However, the confusion tables are different.

```
returns.lda.groupe Down Up
              Down   64 67
              Up     47 74
```

```
returns.qda.group Down Up
              Down   55 58
              Up     56 83
```

Given q populations with means $\mu_1, \ldots, \mu_q$, there is little hope that discriminant analysis will be successful unless one can reject

$$\mathcal{H}_0 : \mu_1 = \cdots = \mu_q.$$

One such test is based on Roy's largest root, $\xi_1 = \lambda_1/(1-\lambda_1)$, where $\lambda_1 \in (0,1)$ measures the discriminating power of Fisher's discriminant function

$$f(\mathbf{X}) = \mathbf{a}^\top(\mathbf{X} - \bar{\mathbf{X}}),$$

where $\mathbf{a}$ is an eigenvector associated with the largest eigenvalue of $\mathbf{S}^{-1}\mathbf{B}$.

The statistic ξ_1 is called after Samarendra Nath Roy (1906–1964).

S.N. Roy (1906–1964) was an Indian-born American mathematician and an applied statistician.

He grew up in Kolkata and worked at the famous Indian Statistical Institute from 1931 to 1950 under the direction of Mahalanobis. He then joined the University of North Carolina at Chapel Hill in 1950.

Roy died in Jasper, Alberta, in July 1964; he was only 57.

We will now see that ξ_1 is the largest eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$ and that $\mathbf{a}$ is a corresponding normed eigenvector.

First, from the definitions of eigenvalue and eigenvector, one has

$$\mathbf{Ba} = \lambda_1 \mathbf{Sa} = \lambda_1 (\mathbf{B} + \mathbf{W})\mathbf{a}.$$

It follows that

$$\mathbf{Ba} = \xi_1 \mathbf{Wa},$$

where $\xi_1 = \lambda_1/(1-\lambda_1)$.

Assuming that $\mathbf{W}^{-1}$ exists, i.e., that $n - q > p$, one has

$$\mathbf{W}^{-1}\mathbf{Ba} = \xi_1 \mathbf{a}.$$

Tests on Means (4–5)

Upon multiplication on the left by $\mathbf{W}^{1/2}$, one deduces that

$$\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}\mathbf{d} = \xi_1 \mathbf{d},$$

where $\mathbf{d} = \mathbf{W}^{1/2}\mathbf{a}$.

It follows that ξ_1 is an eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$ and also of $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$.

Given that $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$ is symmetric, one has $\xi_1 > 0$.

This leads to the conclusion that

$$\lambda_1 = \xi_1/(1+\xi_1) \in [0, 1],$$

as previously stated.

Tests on Means (5–5)

The sample value of ξ_1 is also used to test the equality of the means of q populations with the same but unknown covariance, viz.

$$\mathcal{H}_0 : \mu_1 = \cdots = \mu_q.$$

This is called Roy's largest root test.

Unless Roy's test rejects $\mathcal{H}_0$, there is little hope that a discriminant analysis can be successful.

There are interesting results concerning the distribution of the largest eigenvalue λ_1 under the assumption that the q populations are Gaussian and have the same covariance matrix.

Other Statistics for Testing $\mathcal{H}_0$

Over the years, several statistics have been proposed to test

$$\mathcal{H}_0 : \mu_1 = \cdots = \mu_q .$$

The ubiquitous four are:

- ✓ Hotelling–Lawley's trace: $\mathrm{tr}\left(\mathbf{BW}^{-1}\right) = \sum \xi_i$;
- ✓ Pillai's trace: $\mathrm{tr}(\mathbf{BS}^{-1}) = \sum \lambda_i$;
- ✓ Roy's largest root: $\xi_1 = \lambda_1/(1-\lambda_1)$;
- ✓ Wilks' Lambda: $\Lambda = \dfrac{|\mathbf{W}|}{|\mathbf{S}|} = \dfrac{1}{\mathbf{BW}^{-1}+1} = \prod \dfrac{1}{1+\xi_i}$.

In practice, p-values associated with these various tests are often given.

Hotelling's Statistic (1–3)

In the case of $q = 2$ groups, Hotelling's statistic for testing

$$\mathcal{H}_0 : \mu_1 = \mu_2$$

under the assumption that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ (homoscedasticity) leads to an interesting connection with the Student t statistic.

When $q = 2$, it was already seen that

$$\mathbf{B} = \mathbf{C}\mathbf{C}^\top, \quad \mathbf{C} = \sqrt{\frac{n_1 n_2}{n}}\,(\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2)$$

and

$$\mathbf{a} = \mathbf{W}^{-1}\mathbf{C} = \sqrt{\frac{n_1 n_2}{n}}\,\mathbf{W}^{-1}\,(\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2)\,,$$

up to a multiplicative factor.

Hotelling's Statistic (2–3)

The statistic

$$\xi = \mathbf{C}^\top \mathbf{W}^{-1} \mathbf{C} = \frac{n_1 n_2}{n} \left(\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2\right)^\top \mathbf{W}^{-1} \left(\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2\right).$$

is called Hotelling's T^2 statistic, after Harold Hotelling (1895–1973).

The notation highlights the connection between the statistic and Student's t statistic used to compare $q = 2$ means in the univariate case.

To see this connection, recall that when $q = 2$, the group means are

$$m_1 = \mathbf{a}^\top \tilde{\mathbf{x}}_1, \quad m_2 = \mathbf{a}^\top \tilde{\mathbf{x}}_2,$$

so that one has $m_1 - m_2 = \mathbf{a}^\top \mathbf{C}$.

Hotelling's Statistic (3–3)

Furthermore,

$$S_m^2 = (m_1 - \bar{m})^2 + (m_2 - \bar{m})^2 = \mathbf{a}^\top \mathbf{W} \mathbf{a}.$$

Therefore, considering that $\mathbf{a} = \mathbf{W}^{-1}\mathbf{C}$, one gets

$$\begin{aligned}
\xi &= \mathbf{C}^\top \mathbf{W}^{-1} \mathbf{C} \\
&= \frac{(\mathbf{C}^\top \mathbf{W}^{-1} \mathbf{C})^2}{\mathbf{C}^\top \mathbf{W}^{-1} \mathbf{W} \mathbf{W}^{-1} \mathbf{C}} = \frac{(\mathbf{a}^\top \mathbf{C})^2}{\mathbf{a}^\top \mathbf{W} \mathbf{a}} = \\
&= \frac{(m_1 - m_2)^2}{S_m^2} = \left(\frac{m_1 - m_2}{S_m} \right)^2 .
\end{aligned}$$

Extra: Secondary Discriminant Functions

We saw that the principal eigenvalue ξ_1 of the matrix $\mathbf{W}^{-1}\mathbf{B}$ is associated to Fisher's discriminant function.

The other eigenvalues of this matrix, viz. $\xi_2 > \xi_3 > \cdots$, are associated to secondary discriminant functions, viz.

$$\mathbf{a}_2^\top \mathbf{X}, \quad \mathbf{a}_3^\top \mathbf{X}, \quad \ldots$$

These discriminant functions are not correlated. Nevertheless, they are not orthogonal because the matrix $\mathbf{W}^{-1}\mathbf{B}$ is asymmetric.

Extra: Discriminating Power

The discriminating power of a discriminant function is given by the corresponding eigenvalue.

The relative importance of a discriminant function is given by the ratio of its eigenvalue to the sum of the eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$.

Squared canonical correlation coefficients can be computed for each discriminant function.

This is the coefficient of determination between the discriminant function and the variable indicating to which group each individual belongs.