Multivariate Statistics

Prof. Richard Wilkinson

Spring 2021

Contents

Introduction

This module is concerned with the analysis of multivariate data, in which the response is a vector of random variables rather than a single random variable.

Part I of the module describes some basic concepts in Multivariate Analysis and gives some examples of multivariate data (in Chapter 1), and also contains a summary of the matrix algebra that will be important in this module (Chapter 2).

A theme running through the module is that of dimension reduction. In Part II we consider three types of dimension reduction: Principal Components Analysis (in Chapter 3), whose purpose is to identify the main modes of variation in a multivariate dataset; Canonical Correlation Analysis (Chapter 4), whose purpose is to describe the association between two sets of variables; and Multidimensional Scaling (Chapter 5), in which the starting point is a set of pairwise distances, suitably defined, between the objects under study.

In Part III, we focus on methods of inference for multivariate data whose distribution is multivariate normal. First, in Chapter 6, we develop relevant distribution theory for the multivariate normal distribution. This includes a study of the Wishart distribution, which is a matrix generalisation of the chi-squared distribution, and Hotelling's T^2, which can be thought of as a multivariate generalisation of the Student t distribution. Then in Chapter 7 we focus on inference in multivariate one-sample and two-sample problems in which the underlying distribution is multivariate normal, making use of the distribution theory developed in Chapter 6. In Chapter 8, we focus on the multivariate linear model in which the dependent variable (or y variable) is a vector and the error distribution is multivariate normal.

Finally, in Part IV, we focus on different methods of classification, i.e. allocating the observations in a sample to different subsets (or groups). In Chapter 9, we focus on an approach called discriminant analysis, in which we have a training sample available, and we use this training sample to set up a suitable classification rule. In Chapter 10, we consider an alternative approach, known as cluster analysis, in which we allocate the observations into clusters (or similar subsets) when a training sample is not available.

PART I: Prerequisites

In Chapter 1 we explain what we mean by multivariate analysis and give some examples of multivariate data. We also introduce basic definitions and concepts such as the sample covariance matrix, the sample correlation matrix and graphical techniques. We also briefly discuss random vectors and random matrices and derive some of their elementary properties.

In Chapter 2 we summarise the definitions, ideas and results from matrix algebra that will be needed later in the module.

Chapter 1

Statistical Preliminaries

In this chapter blah blah

1.1 Multivariate data

What is multivariate analysis (MVA)? Analysis of data where two or more response variables are measured on each object under study. If we measure p variables on n objects then the data can be presented in a $n \times p$ **data matrix**.

We shall often write the data matrix as $\mathbf{X}$ $(n \times p)$ where

$$\mathbf{X} = \begin{bmatrix} x_1^\top \\ x_2^\top \\ .. \\ .. \\ .. \\ x_n^\top \end{bmatrix} = [x_1, \dots, x_n]^\top.$$

In words: the rows of $\mathbf{X}$ are $x_1^\top, \dots, x_n^\top$ and the columns of $X^\top$ are $x_1, \dots, x_n$.

In this setup, we think of the $x_1, \dots, x_n$ are being the observation vectors, and the p columns of $\mathbf{X}$ correspond to the p variables being measured.

Important remark on notation: Throughout the module we shall use non-bold letters, whether upper or lower case, to indicate scalar (i.e. real-valued) quantities; lower-case letters in bold to signify column vectors; and upper case letters in bold to signify matrices. This convention for bold letters will also apply to random quantities. So, in particular, for a random vector we always use (bold) lower case, and for a random matrix we always use bold upper-case, regardless of whether we are referring to (i) the unobserved random quantity or (ii) its observed value. It should always be clear from the context which of these two interpretations (i) or (ii) is appropriate.

Example 1.1. Football league table where W = number of wins, D = number of draws, F = number of goals scored and A = number of goals conceded for four teams. In this example we have $p = 4$ variables $(W, D, F, A)^\top$ measured on $n = 4$ objects (teams).

Team	W	D	F	A
USA	1	2	4	3
England	1	2	2	1
Slovenia	1	1	3	3
Algeria	0	1	0	2

Example 1.2. Exam marks for a set of n students where P = mark in probability and S = mark in statistics. Note that x_{ij} denotes the jth variable measured on the ith subject.

Student	P	S
1	x_{11}	x_{12}
2	x_{21}	x_{22}
$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	x_{n2}

In MVA we attempt to answer questions such as:

- What is the joint distribution of marks?
- How can we visualise the data?
- Can we simplify the data? For example, we rank football teams using $3W + D$ and we rank students by their average module mark but is this fair? Can we reduce the dimension in a better way?
- Can we use the data to discriminate, for example, between male and female students?

We could just apply standard univariate techniques to each variable in turn but this ignores possible dependencies between the variables which we must represent to draw valid conclusions.

Finally, before moving on, we ask the question: what is the difference between MVA and standard linear regression? Answer: in standard linear regression we have a scalar response variable, y say, and a vector of covariates, x, say. The focus of interest is on how knowledge of x influences the distribution of y (in particular, the mean of y). In contrast, with MVA the focus of interest is a response vector y, in which all the components of y are viewed as responses rather than covariates. However, there are also situations where the response is a vector y but we also have covariate information x. This leads to study of the multivariate linear model, which we will investigate later on in Chapter ???.

1.2 Summary statistics

In univariate statistics we define the sample mean and sample variance as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

The analogous multivariate **sample mean** and **sample covariance matrix** are direct extensions:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top,$$

where x_i is the p dimensional vector denoting the p observations on the ith object.

Note that the jth entry in $\bar{x}$ is simply the (univariate) sample mean of the jth variable. Similarly, if s_{ab} is the (a,b)th entry of S, then s_{jj} is the (univariate) sample variance of the jth variable and s_{ab} is the sample covariance of variables a and b. Note that S is symmetric since $s_{ab} = s_{ba}$.

An equivalent alternative formula for S is

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

Similarly, let R be the **sample correlation matrix** where the (a,b)th entry of R is the sample correlation between variables a and b, that is

$$r_{ab} = s_{ab}/\sqrt{s_{aa}s_{bb}}.$$

Note that

$$R = D^{-1}SD^{-1}$$

where $D = \text{diag}(\sqrt{s_{11}}, \ldots, \sqrt{s_{pp}})$. Note that R is symmetric, the diagonal entries are always exactly 1 (each variable is perfectly correlated with itself) and that $|r_{ab}| \leq 1$.

Note that if we change the unit of measurement for the x_i's then S will change but R will not.

The **total variation** in the data set is usually measured by $\text{tr}(S)$ where tr() is the trace function that sums the diagonal elements of the matrix. That is,

$$\text{tr}(S) = s_{11} + s_{22} + \ldots + s_{pp}.$$

In MVA it is much easier to work with vector and matrix notation.

Example 1.3. The table below shows the module marks for 5 students on the modules G11PRB (P) and G11STA (S).

Student	P	S
A	41	63
B	72	82
C	46	38
D	77	57
E	59	85

Calculate the sample mean, sample covariance, sample correlation and total variation.

The sample mean is $\bar{x} = \begin{pmatrix} 59 \\ 65 \end{pmatrix}$.

The sample covariance matrix is $S = \begin{pmatrix} 197.2 & 92.8 \\ 92.8 & 297.2 \end{pmatrix}$.

The sample correlation matrix is

$$\begin{aligned} R &= D^{-1}SD^{-1} \\ &= \begin{pmatrix} 14.0 & 0 \\ 0 & 17.2 \end{pmatrix}^{-1} \begin{pmatrix} 197.2 & 92.8 \\ 92.8 & 297.2 \end{pmatrix} \begin{pmatrix} 14.0 & 0 \\ 0 & 17.2 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} 0.071 & 0 \\ 0 & 0.058 \end{pmatrix} \begin{pmatrix} 197.2 & 92.8 \\ 92.8 & 297.2 \end{pmatrix} \begin{pmatrix} 0.071 & 0 \\ 0 & 0.058 \end{pmatrix} \\ &= \begin{pmatrix} 1.000 & 0.383 \\ 0.383 & 1.000 \end{pmatrix}. \end{aligned}$$

The total variation is $\text{tr}(S) = 197.2 + 297.2 = 494.4$.

To calculate these in R use, 'colMeans', 'cov', and 'cor'. These assume each column is a different variable, and each row a different observation.

```
library(dplyr)
Ex1 <- data.frame(
  Student=LETTERS[1:5],
  P = c(41,72,46,77,59),
  S = c(63,82,38,57,85)
  )

Ex1 %>% select_if(is.numeric) %>% colMeans
```

```
##  P  S
## 59 65
```

```
Ex1 %>% select_if(is.numeric) %>% cov
```

```
##       P     S
## P 246.5 116.0
```

```
## S 116.0 371.5
```

```
Ex1 %>% select_if(is.numeric) %>% cor
```

```
##           P         S
## P 1.0000000 0.3833276
## S 0.3833276 1.0000000
```

NOTE R USES 1/n-1 whereas hand calculation si 1/n…..

1.3 Graphical techniques

We can draw histograms and scatter plots to view the distribution when $p = 1$ and $p = 2$ respectively. For $p \geq 3$ the task is much harder. One solution is a matrix of pair-wise scatter plots using the `pairs` command in R. The graph below shows the relationship between goals scored (F), goals against (A) and points (PT) for 20 teams during a recent Premiership season.

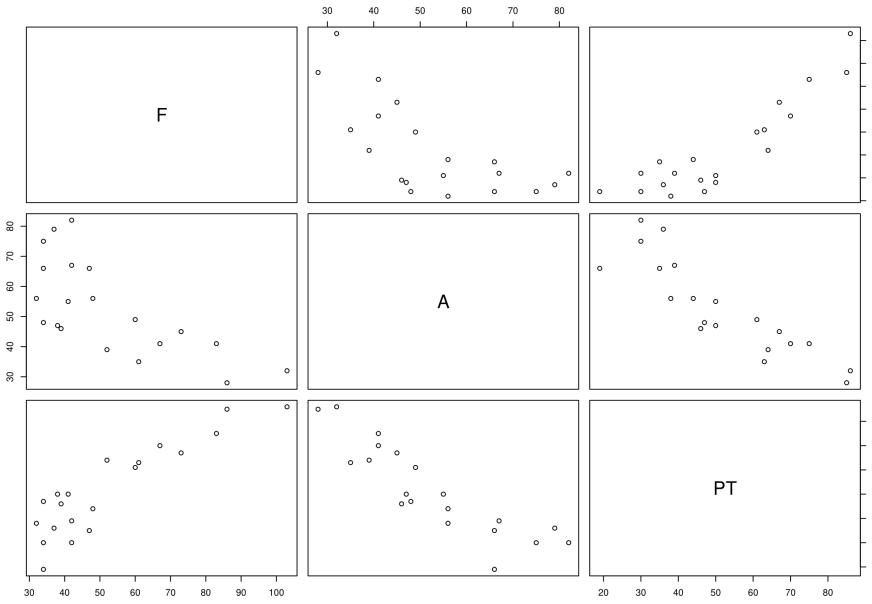


Figure 1.1: Scatter plots

WOULD BE BETTER TO CREATE THIS FIGURE HERE

You can also use the `plot3d` command in the `rgl` library to create an interactive

3D plot of the data. The difficulty of displaying multivariate data is further motivation for developing a method for reducing the number of dimensions in the data.

1.4 Random Vectors and Matrices

The *population mean vector* of the random vector x is

$$\mu = E(x).$$

The *population covariance matrix* of x is

$$\Sigma = \mathrm{Var}(x) = E\left((x-\mu)(x-\mu)^\top\right).$$

The covariance between x $(p \times 1)$ and y $(q \times 1)$ is

$$\mathrm{Cov}(x,y) = E\left((x-E(x))(y-E(y))^\top\right).$$

Let A $(q \times p)$ denote a constant matrix and let b $(q \times 1)$ denote a constant vector. Then the following properties hold:

- $E(Ax+b) = AE(x) + b = A\mu + b$.
- $\Sigma = E(xx^\top) - \mu\mu^\top$.
- $\mathrm{Var}(Ax+b) = A\Sigma A^\top$, where A is a $q \times p$ constant matrix.
- A covariance matrix Σ is always non-negative definite. Moreover, Σ is positive definite if and only if all its eigenvalues are positive, in which case its determinant is positive and Σ is non-singular.
- $\mathrm{Cov}(x,y) = E(xy^\top) - E(x)E(y)^\top$.
- $\mathrm{Cov}(x,x) = \Sigma$. $\mathrm{Cov}(x,y) = \mathrm{Cov}(y,x)^\top$.
- If x and y are independent then $\mathrm{Cov}(x,y) = \mathbf{0}_{p,q}$.
- If $p = q$ then

$$\mathrm{Var}(x+y) = \mathrm{Var}(x) + \mathrm{Var}(y) + \mathrm{Cov}(x,y) + \mathrm{Cov}(y,x).$$

1.5 Unbiased Estimators

Recall from univariate statistics that an estimator $\hat{\theta}$ of a parameter θ is unbiased if $E(\hat{\theta}) = \theta$ for all θ. This concept readily transfers to the multivariate context.

Proposition 1.1. *Let $x_1, \ldots x_n$ be independent and identically distributed (i.i.d.), sampled from a population with mean μ and covariance matrix Σ. If $\bar{x}$ and S are the sample mean and covariance matrix, respectively, then*

1. $E(\bar{x}) = \mu$.
2. $Var(\bar{x}) = \dfrac{1}{n}\Sigma$.

3. $E(S) = \dfrac{n-1}{n}\Sigma$.

Proof. **Part 1** Since expectations behave linearly,

$$E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n} n\mu = \mu.$$

Part 2 We have

$$\begin{aligned}
\mathrm{Var}(\bar{x}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\
&= \sum_{i,j=1}^{n} \mathrm{Cov}\left(\frac{1}{n}x_i, \frac{1}{n}x_j\right) \\
&= \sum_{i=1}^{n} \mathrm{Var}\left(\frac{1}{n}x_i\right) + \sum_{i\neq j} \mathrm{Cov}\left(\frac{1}{n}x_i, \frac{1}{n}x_j\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} \mathrm{Var}(x_i) + \sum_{i\neq j} \mathrm{Cov}(x_i, x_j)\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} \mathrm{Var}(x_i)\right) \\
&= \frac{1}{n^2} n\Sigma \\
&= \frac{1}{n}\Sigma.
\end{aligned}$$

Part 3 From the definition of the sample covariance,

$$\begin{aligned}
S &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu + \mu - \bar{x})(x_i - \mu + \mu - \bar{x})^\top \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big\{(x_i - \mu)(x_i - \mu)^\top + (\bar{x} - \mu)(\bar{x} - \mu)^\top \\
&\qquad -(x_i - \mu)(\bar{x} - \mu)^\top - (\bar{x} - \mu)(x_i - \mu)^\top\Big\} \\
&= \frac{1}{n}\left\{\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^\top\right\} - (\bar{x} - \mu)(\bar{x} - \mu)^\top
\end{aligned}$$

Since $E(\bar{x}) = \mu$ and $\mathrm{Var}(\bar{x}) = n^{-1}\Sigma$, it follows that

$$\begin{aligned} E(S) &= E\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^\top\right\} \\ &\quad -E\left\{(\bar{x}-\mu)(\bar{x}-\mu)^\top\right\} \\ &= \left\{\frac{1}{n}\sum_{i=1}^{n}E\left[(x_i-\mu)(x_i-\mu)^\top\right]\right\} - \mathrm{Var}(\bar{x}) \\ &= E\left\{(x_1-\mu)(x_1-\mu)^\top\right\} - \mathrm{Var}(\bar{x}) \\ &= \mathrm{Var}(x_1) - \mathrm{Var}(\bar{x}) \\ &= \Sigma - \frac{1}{n}\Sigma \\ &= \frac{n-1}{n}\Sigma, \end{aligned}$$

which completes the proof.

□

An implication of this theorem is that $\bar{x}$ is an unbiased estimator for μ but that S is a biased estimator of Σ. Note, however, that $\frac{n}{n-1}S$ is an unbiased estimator of Σ, i.e.

$$\frac{n}{n-1}E[S] = \Sigma.$$

Chapter 2

Review of Matrix algebra

In this chapter blah blah

2.1 Basic definitions

The matrix $\mathbf{A}$ will be referred to in the following equivalent ways:

$$\begin{aligned}\mathbf{A} = \overset{n\times p}{\mathbf{A}} &= \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{bmatrix} \\ &= [a_{ij} : i = 1, \dots, n;\ j = 1, \dots, p\,] = (a_{ij}),\end{aligned}$$

where the a_{ij} are real numbers.

A matrix of order 1×1 is called a *scalar*.

A matrix of order $n \times 1$ is called a *(column) vector*.

A matrix of order $1 \times p$ is called a *(row) vector*.

e.g. $\overset{n\times 1}{\mathbf{a}} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$ is a column vector.

The $n \times n$ *identity matrix* $\mathbf{I}_n$ has diagonal elements equal to 1 and off-diagonal elements equal to zero.

A *diagonal* matrix is an $n \times n$ matrix whose off-diagonal elements are zero. Sometimes we denote a diagonal matrix by $\text{diag}\{a_1, \dots, a_n\}$.

2.2 Elementary matrix operations

1. *Addition/Subtraction.* If $\overset{n\times p}{\mathbf{A}} = [a_{ij}]$ and $\overset{n\times p}{\mathbf{B}} = [b_{ij}]$ are given matrices then

$$\mathbf{A} + \mathbf{B} = [a_{ij} + b_{ij}] \qquad \text{and} \qquad \mathbf{A} - \mathbf{B} = [a_{ij} - b_{ij}].$$

2. *Scalar Multiplication.* If λ is a scalar and $\mathbf{A} = [a_{ij}]$ then

$$\lambda\mathbf{A} = [\lambda a_{ij}].$$

3. *Matrix Multiplication.* If $\overset{n\times p}{\mathbf{A}}$ and $\overset{p\times q}{\mathbf{B}}$ are given matrices then $\mathbf{A}B = \overset{n\times q}{\mathbf{C}} = [c_{ij}]$ where

$$c_{ij} = \sum_{k=1}^{p} a_{ik}b_{kj}, \qquad i = 1, \ldots, n, \qquad j = 1, \ldots, q.$$

4. *Matrix Transpose.* If $\overset{m\times n}{A} = [a_{ij} : i = 1, \ldots, m; j = 1, \ldots, n]$, then the transpose of A, written $A^\top$, is given by the $n \times m$ matrix

$$A^\top = [a_{ji} : j = 1, \ldots, n; i = 1, \ldots, m].$$

Note from the definitions that $(\mathbf{A}B)^\top = \mathbf{B}^\top\mathbf{A}^\top$. In the special case in which $\mathbf{A} = \overset{n\times 1}{\mathbf{a}}$ and $\mathbf{B} = \overset{n\times 1}{\mathbf{b}}$, the quantity $\mathbf{A}^\top\mathbf{B} = \mathbf{a}^\top\mathbf{b}$ is real-valued and is known as the *scalar product of* $\mathbf{a}$ *and* $\mathbf{b}$.

Note that $\mathbf{a}^\top\mathbf{a} = \sum_{i=1}^{n} a_i^2 \geq 0$ *with equality iff* $\mathbf{a} = \mathbf{0}_n$ *where* $\overset{n\times 1}{\mathbf{0}}_n = (0, 0, \ldots, 0)^\top$. *We may interpret* $\mathbf{a}^\top\mathbf{a}$ *as the (length)*^{2} *of the vector* $\mathbf{a}$. *The* norm* of a vector $\mathbf{a}$ is defined by

$$||\mathbf{a}|| = (\mathbf{a}^\top\mathbf{a})^{1/2}.$$

The scalar product has an alternative (but equivalent) representation:

$$a^\top b = ||a|| \cdot ||b|| \cos(\theta),$$

where θ is the angle (in radians) between the vectors a and b.

A *unit vector* $\mathbf{a}$ is a vector satisfying $||\mathbf{a}||^2 = \mathbf{a}^\top\mathbf{a} = 1.\backslash$ A matrix $\overset{n\times n}{\mathbf{Q}}$ satisfying $\mathbf{Q}Q^\top = \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_n$ is called an *orthogonal matrix*. Equivalently, a matrix $\mathbf{Q}$ is orthogonal iff $\mathbf{Q}^{-1} = \mathbf{Q}^\top$.

If $\mathbf{Q} = [q_1, \ldots, q_n]$ is an orthogonal matrix, then the columns $q_1, \ldots, q_n$ are mutually orthogonal unit vectors, i.e.

$$q_j^\top q_k = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases}$$

Proposition 2.1. *If $q_1, \ldots, q_n$ are mutually orthogonal $n \times 1$ unit vectors then*

$$\sum_{i=1}^{n} q_i q_i^\top = \mathbf{I}_n,$$

the $n \times n$ identity matrix.

5. *Matrix Inverse.* The inverse of a matrix $\overset{n\times n}{\mathbf{A}}$ (if it exists) is a matrix $\overset{n\times n}{\mathbf{B}}$ such that $\mathbf{A}B = \mathbf{B}A = \mathbf{I}_n$. We denote the inverse by $\mathbf{A}^{-1}$. Note that if $\mathbf{A}_1$ and $\mathbf{A}_2$ are both invertible, then $(\mathbf{A}_1\mathbf{A}_2)^{-1} = \mathbf{A}_2^{-1}\mathbf{A}_1^{-1}$.

6. *Trace.* The trace of a matrix $\overset{n\times n}{\mathbf{A}}$ is given by

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}.$$

Lemma 2.1. *For any matrices A ($ n $\times m$) and B ($m \times n$),*

$$tr(AB) = tr(BA).$$

2.3 Linear independence and determinants

Vectors $\overset{n\times 1}{\mathbf{x}}_1, \ldots, \overset{n\times 1}{\mathbf{x}}_p$ are said to be *linearly dependent* if there exist scalars $\lambda_1, \ldots, \lambda_p$ not all zero such that

$$\lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2 + \cdots + \lambda_p\mathbf{x}_p = \mathbf{0}.$$

Otherwise, these vectors are said to be *linearly independent*.

The *rank* of a matrix is equal to the maximum number of linearly independent rows (equivalently, columns).

The *determinant* of a square matrix $\overset{n\times n}{\mathbf{A}}$ is defined as

$$\det(\mathbf{A}) = \sum (-1)^{|\tau|} a_{1\tau(1)} \ldots a_{n\tau(n)}$$

where the summation is taken over all permutations τ of $\{1, 2, \ldots, n\}$, and we define $|\tau| = 0$ or 1 depending on whether τ can be written as an even or odd number of transpositions.

E.g. If $\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$, then $\det(\mathbf{A}) = a_{11}a_{22} - a_{12}a_{21}$.

Proposition 2.2. *For any matrices $\overset{n\times n}{\mathbf{A}}$, $\overset{n\times n}{\mathbf{B}}$, $\overset{n\times n}{\mathbf{C}}$ such that $\mathbf{C} = \mathbf{AB}$,*

$$det(\mathbf{C}) = det(\mathbf{A}) \cdot det(\mathbf{B}).$$

2.4 Eigenvalues and eigenvectors

If $\overset{n\times n}{\mathbf{A}}$ is a given matrix then $R(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}_n)$ is an n^{th} order polynomial in λ. The n roots $\lambda_1, \ldots, \lambda_n$ of $R(\lambda)$ (possibly complex numbers) are called *eigenvalues* of $\mathbf{A}$.

For any eigenvalue λ of $\overset{n\times n}{\mathbf{A}}$, there exists a non-zero vector $\overset{n\times 1}{\mathbf{x}}$, called an *eigenvector*, such that $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$.

Proposition 2.3. *If $\mathbf{A}$ is symmetric (i.e. $\mathbf{A}^\top = \mathbf{A}$) then the eigenvalues of $\mathbf{A}$ are all* real *and all the eigenvectors of $\mathbf{A}$ have* real *components.*

Proposition 2.4. *The rank of a symmetric matrix is equal to the number of non-zero eigenvalues.*

Proposition 2.5. ***(Spectral decomposition theorem)**. Any symmetric matrix $\overset{n\times n}{\mathbf{A}}$ can be written*

$$\mathbf{A} = \mathbf{Q}\Lambda Q^\top = \sum_{i=1}^{n} \lambda_i \mathbf{q}_i \mathbf{q}_i^\top,$$

where $\overset{n\times n}{} = diag\{\lambda_1, \ldots, \lambda_n\}$ is a diagonal matrix consisting of the eigenvalues of $\mathbf{A}$ and $\overset{n\times n}{\mathbf{Q}}$ is an orthogonal matrix whose columns are unit eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_n$ of $\mathbf{A}$.

Proposition 2.6. *If $\overset{n\times n}{\mathbf{A}}$ is a symmetric matrix then its determinant is the product of its eigenvalues, i.e. det$(\mathbf{A}) = \lambda_1 \ldots \lambda_n$.*

Let $\overset{n\times n}{\mathbf{A}}$ be a symmetric matrix with (necessarily real) eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Then $\mathbf{A}$ is said to be *positive definite* if and only if $\lambda_n > 0$, and A is said to be *non-negative definite* if and only if $\lambda_n \geq 0$.

For a given symmetric $\overset{n\times n}{\mathbf{A}}$, define the *quadratic form* $\mathcal{Q}(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$.

Proposition 2.7. *In the above notation,*

$$\max_{x: x^\top x = 1} \mathcal{Q}(\mathbf{x}) = \lambda_1,$$

where the maximum occurs at $x = \pm q_1$.

Proposition 2.8. *In the above notation,*

$$\min_{x: x^\top x = 1} \mathcal{Q}(\mathbf{x}) = \lambda_n$$

where the minimum occurs at $x = \pm q_n$.

Proposition 2.9. *We have (i) $\mathcal{Q}(x) > 0$ for all $x \neq \mathbf{0}_n$ if and only if $\mathbf{A}$ is positive definite; and (ii) $\mathcal{Q}(x) \geq 0$ for all x if and only if A is non-negative definite.*

A matrix $\overset{n\times n}{P}$ is a *projection* matrix if it is symmetric, i.e. $P^\top = P$, and

$$P^2 = P.$$

Proposition 2.10. *The eigenvalues of a projection matrix P are all 0 or 1.*

Proposition 2.11. *If $\overset{n\times n}{P}$ is a projection matrix then $\mathbf{I}_n - P$ is also a projection matrix.*

2.5 Matrix square roots

From time to time we shall need to consider square roots of symmetric non-negative definite matrices. From Proposition 2.5, a symmetric matrix $\mathbf{A}$ may be written as $\mathbf{A} = Q\Lambda Q^\top$ where is a diagonal matrix and $\mathbf{Q}$ is an orthogonal matrix. Moreover, A is non-negative definite if and only if the diagonal elements of (the eigenvalues of $\mathbf{A}$) are all non-negative.

For such a matrix we define $A^{1/2}$, a matrix square root of A, by $A^{1/2} = Q\Lambda^{1/2}Q^\top$ where $\Lambda^{1/2} = \text{diag}\{\lambda_1^{1/2}, \ldots, \lambda_n^{1/2}\}$. This definition makes sense because

$$\begin{aligned} A^{1/2}A^{1/2} &= Q\Lambda^{1/2}Q^\top Q\Lambda^{1/2}Q^\top \\ &= Q\Lambda^{1/2}\Lambda^{1/2}Q^\top \\ &= Q\Lambda Q^\top \\ &= A, \end{aligned}$$

where $Q^\top Q = I_n$ and $\Lambda^{1/2}\Lambda^{1/2} = \Lambda$. The matrix $A^{1/2}$ is not the only matrix square root of A, but it *is* the only symmetric, non-negative definite square root of A.

If A is positive definite (as opposed to just non-negative definite), then all the λ_i are positive and so we can also define $A^{-1/2} = Q\Lambda^{-1/2}Q^\top$ where $\Lambda^{-1/2} = \text{diag}\{\lambda_1^{-1/2}, \ldots, \lambda_n^{-1/2}\}$. Note that

$$A^{-1/2}A^{-1/2} = Q\Lambda^{-1/2}Q^\top Q\Lambda^{-1/2}Q^\top = Q\Lambda^{-1}Q^\top = A^{-1},$$

so that, as defined above, $A^{-1/2}$ the matrix square root of A^{-1}. Furthermore, similar calculations show that

$$A^{1/2}A^{-1/2} = A^{-1/2}A^{1/2} = I_n,$$

so that $A^{-1/2}$, as defined above, is the matrix inverse of $A^{1/2}$.

2.6 Singular Value Decomposition

The spectral decomposition theorem (Proposition 2.5) gives a decomposition of any symmetric matrix. We now give a generalisation of this result which applies to *all* matrices. We will need this extra generality in Chapter 4 on Canonical Correlation Analysis.

Proposition 2.12. ***(Singular value decomposition)***. *Let A be a $p \times q$ matrix of rank t, where $1 \leq t \leq \min(p, q)$. Then there exists a $p \times t$ matrix $Q = [q_1, ..., q_t]$, a $q \times t$ matrix $R = [\mathbf{r}_1, ..., \mathbf{r}_t]$, and a $t \times t$ diagonal matrix $= diag\{\xi_1, ..., \xi_t\}$ such that*

$$A = Q\, R^\top = \sum_{i=1}^{t} \xi_i q_i \mathbf{r}_i^\top,$$

where $Q^\top Q = I_t = R^\top R$ and the $\xi_1 \geq ... \geq \xi_t > 0$.

Note that the q_i and the $\mathbf{r}_i$ are necessarily unit vectors.

The scalars $\xi_1, ..., \xi_t$ are called *singular values*.

When A is symmetric, we take $\mathbf{R} = Q$, and the spectral decomposition theorem is recovered, and in this case (but not in general) the singular values of A are in fact eigenvalues of A.

The following result relates $\mathbf{Q}$, and $\mathbf{R}$ to certain eigenvalues and eigenvectors.

Proposition 2.13. *Let A be any matrix of rank t. Then the non-zero eigenvalues of both $AA^\top$ and $A^\top A$ are given by $\xi_1^2, ..., \xi_t^2$; the corresponding unit eigenvectors of $AA^\top$ are given by the columns of $\mathbf{Q}$; and the corresponding unit eigenvectors of $A^\top A$ are given by the columns of $\mathbf{R}$.*

The following result is important in Canonical Correlation Analysis.

Proposition 2.14. *For any matrix A of rank t with singular values $\xi_1 \geq \xi_2 \geq ... \geq \xi_t > 0$, we have*

$$\max_{x,y:\, ||x||=||y||=1} x^\top A y = \xi_1.$$

2.7 The Centering Matrix

From time to time in this module an important role will be played by the *centering matrix*

$$H = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top. \tag{2.1}$$

Note that, in the above, I_n is the $n \times n$ identity matrix, while $\mathbf{1}_n$ is an $n \times 1$ column vector of ones.

The reason for the terminology *centering* will become clear below.

Important properties of the matrix H in Equation (2.1) are now listed. These properties are proved in the example sheets.

1. The matrix H is a projection matrix, i.e. $H^\top = H$ and $H^2 = H$.
2. Writing $\mathbf{0}_n$ for the $n \times 1$ vector of zeros, we have $H\mathbf{1}_n = \mathbf{0}_n$ and $\mathbf{1}_n^\top H = \mathbf{0}_n^\top$.
3. If $x = (x_1, \ldots, x_n)^\top$, then $Hx = x - \bar{x}\mathbf{1}_n$ where $\bar{x} = n^{-1}\sum_{i=1}^n x_i$.
4. With x as in (iii), we have

$$x^\top H x = \sum_{i=1}^n (x_i - \bar{x})^2,$$

and so

$$\frac{1}{n} x^\top H x = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 = \hat{\sigma}^2,$$

where $\hat{\sigma}^2$ is the sample variance.

5. If $X = [x_1, \ldots, x_n]^\top$ is an $n \times p$ data matrix then

$$HX = \begin{bmatrix} (x_1 - \bar{x})^\top \\ (x_2 - \bar{x})^\top \\ \cdot\cdot \\ \cdot\cdot \\ \cdot\cdot \\ (x_n - \bar{x})^\top \end{bmatrix} = [x_1 - \bar{x}, \ldots, x_n - \bar{x}]^\top$$

6. With X as in (v),

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

where S is the sample covariance matrix.

7. If $A = (a_{ij})_{i,j=1}^n$ is a symmetric $n \times n$ matrix, then

$$B = HAH = A - \mathbf{1}_n \bar{a}_+^\top - \bar{a}_+ \mathbf{1}_n^\top + \bar{a}_{++}\mathbf{1}_n\mathbf{1}_n^\top,$$

or, equivalently,

$$b_{ij} = a_{ij} - \bar{a}_{i+} - \bar{a}_{+j} + \bar{a}_{++}, \qquad i, j = 1, \ldots, n,$$

where

$$\bar{a}_+ \equiv (\bar{a}_{1+}, \ldots, \bar{a}_{n+})^\top = \frac{1}{n} A\mathbf{1}_n,$$

$\bar{a}_{+j} = \bar{a}_{j+}$, for , $j = 1, \ldots, n$,, and , $\bar{a}_{++} = n^{-2}\sum_{i,j=1}^n a_{ij}$.

Note that Property 3. is a special case of Property 5., and Property 4. is a special case of Property 6. However, it is useful to see these results in the simpler scalar case before moving onto the the general matrix case.

2.8 Quadratic forms and ellipses

A standard ellipse in $\mathbb{R}^2$ is given by the equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \quad (a > b > 0).$$

The interior (the shaded region) is given by

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} \leq 1. \tag{2.2}$$

Note that a standard ellipse has axes of symmetry given by the x-axis and y-axis (if $a > b$, the former is the major axis and the latter the minor axis).

If we define $\mathbf{A} = \begin{bmatrix} a^2 & 0 \\ 0 & b^2 \end{bmatrix}$ then Equation (2.2) can be written in the form

$$\begin{pmatrix} x \\ y \end{pmatrix}^\top \mathbf{A}^{-1} \begin{pmatrix} x \\ y \end{pmatrix} \leq 1.$$

If we write $\mathbf{x} = \binom{x_1}{x_2}$ and generalise to an arbitrary symmetric positive definite matrix $\overset{2\times 2}{\mathbf{A}}$, what is the set

$$\{\mathbf{x} \in \mathbb{R}^2 : \mathbf{x}^\top \mathbf{A}^{-1} \mathbf{x} \leq 1\}?$$

We get a rotated ellipse with axes of symmetry given by the eigenvectors of $\mathbf{A}$, with the major axis determined by the eigenvector corresponding to the larger eigenvalue of $\mathbf{A}$, and the minor axis determined by the eigenvector corresponding to the smaller eigenvalue of $\mathbf{A}$.

Note that, for $c > 0$,

$$\mathbf{x}^\top \mathbf{A}^{-1} \mathbf{x} \leq c \qquad \Leftrightarrow \qquad \mathbf{x}^\top (c\mathbf{A})^{-1} \mathbf{x} \leq 1,$$

where $c\mathbf{A}$ is a scalar multiple of $\mathbf{A}$.

If $\mathbf{m}$ is a fixed 2-vector, then what is the set

$$\{\mathbf{x} \in \mathbb{R}^2 : (\mathbf{x} - \mathbf{m})^\top \mathbf{A}^{-1} (\mathbf{x} - \mathbf{m}) \leq 1\}?$$

Since

$$\{\mathbf{x} : (\mathbf{x} - \mathbf{m})^\top \mathbf{A}^{-1} (\mathbf{x} - \mathbf{m}) \leq 1\} = \{\mathbf{z} + \mathbf{m} : \mathbf{z}^\top \mathbf{A}^{-1} \mathbf{z} \leq 1\},$$

it follows that

$$\{\mathbf{x} : (\mathbf{x} - \mathbf{m})^\top \mathbf{A}^{-1} (\mathbf{x} - \mathbf{m}) \leq 1\}$$

is just the ellipse $\{\mathbf{z} : \mathbf{z}^\top \mathbf{A}^{-1} \mathbf{z} \leq 1\}$ translated by $\mathbf{m}$.

Analogous results for ellipsoids and quadratic forms hold in three and higher dimensions.

2.9 Lines and Hyperplanes in $\mathbb{R}^p$

For any $a, b \in \mathbb{R}^p$, the set

$$\mathcal{L} = \mathcal{L}(a, b) = \{a + \gamma b : \gamma \in \mathbb{R}\} \tag{2.3}$$

is a *straight line* in $\mathbb{R}^p$.

If $a^\top b = 0$, i.e. a and b are orthogonal, then a is the perpendicular from the origin $\mathbf{0}_p$ to the line $\mathcal{L}(a, b)$.

For fixed $a \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}$,

$$\mathcal{H} = \mathcal{H}(a, \gamma) = \{x \in \mathbb{R}^p : a^\top x = \gamma\}$$

is a hyperplane of dimension $p-1$ in $\mathbb{R}^p$. The vector a is the perpendicular from the origin $\mathbf{0}_p$ to the hyperplane $\mathcal{H}(a, \gamma)$.

There is an alternative way to define hyperplanes in $\mathbb{R}^P$. Suppose that, for $1 \leq r < p$, $\overset{p\times 1}{a}_1, \ldots, \overset{p\times 1}{a}_r, \overset{p\times 1}{a}_{r+1}$ are linearly independent. Then

$$\mathcal{H} = \left\{ \sum_{j=1}^{r+1} \gamma_j a_j : \sum_{j=1}^{r+1} \gamma_j = 1 \right\}$$

is an r-dimensional hyperplane in $\mathbb{R}^p$.

When $r = 1$, using the fact that $\gamma_1 + \gamma_2 = 1$, we may write

$$\gamma_1 a_1 + \gamma_2 a_2 = (1 - \gamma_2)a_1 + \gamma_2 a_2 = a_1 + \gamma_2(a_2 - a_1),$$

which agrees with $a + \gamma b$ in (2.3) when $a = a_1$, $b = a_2 - a_1$ and $\gamma = \gamma_2$. So we have shown that the two definitions agree in the case of a straight line.

2.10 Vector Differentiation

Consider a real-valued function $f : \mathbb{R}^p \to \mathbb{R}$ of a vector variable $x = (x_1, \ldots, x_p)^\top$. Sometimes we will want to differentiate f. We define the partial derivative of $f(x)$ with respect to x to be the vector of partial derivatives, i.e.

$$\frac{\partial f}{\partial x}(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ .. \\ .. \\ .. \\ \frac{\partial f}{\partial x_p}(x) \end{bmatrix} \tag{2.4}$$

The following examples can be worked out directly from the definition (2.4), using the chain rule in some cases.

Example 2.1. If $f(x) = a^\top x$ where $a \in \mathbb{R}^p$ is a constant vector, then

$$\frac{\partial f}{\partial x}(x) = a.$$

Example 2.2. If $f(x) = (x-a)^\top A(x-a)$ for a fixed vector $a \in \mathbb{R}^p$ and A is a constant symmetry $p \times p$ matrix, then

$$\frac{\partial f}{\partial x}(x) = 2A(x-a).$$

Example 2.3. Suppose that $g : \mathbb{R} \to \mathbb{R}$ is a differentiable function with derivative g'. Then, using the chain rule for partial derivatives,

$$\frac{\partial g(a^\top x)}{\partial x} = g'(a^\top x)\frac{\partial}{\partial x}\left\{a^\top x\right\} = g'(a^\top x)a.$$

Example 2.4. If f is defined as in Example 2 and g is as in Example 3 then, using the chain rule again,

$$\frac{\partial}{\partial x} g\{f(x)\} = g'\{f(x)\}\frac{\partial f}{\partial x}(x) = 2g'\{(x-a)^\top A(x-a)\}A(x-a).$$

If we wish to find a maximum or minimum of $f(x)$ we should search for stationary points of f, i.e. solutions to the system of equations

$$\frac{\partial f}{\partial x}(x) \equiv \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \cdot\cdot \\ \cdot\cdot \\ \cdot\cdot \\ \frac{\partial f}{\partial x_p}(x) \end{bmatrix} = \mathbf{0}_p.$$

The nature of a stationary point is determined by the *Hessian*, i.e. the matrix of second derivatives. The Hessian is the $p \times p$ matrix

$$\frac{\partial^2 f}{\partial x \partial x^\top}(x) = \left\{\partial^2 f(x)/\partial x_j \partial x_k\right\}_{j,k=1}^{p}.$$

If the Hessian is positive (negative) definite at a stationary point x, then the stationary point is a minimum (maximum).

If the Hessian has both positive an negative eigenvalues at x then the stationary point will be a *saddle point*.

PART II: Dimension reduction methods

In applications of statistics in many different fields it is common to measure several (or even a large number) of variables on each experimental unit under study. For example, experimental units could be individual people and variables could be measurements obtained in a general health check-up (e.g. age, blood pressure, cholesterol level, lung function measurements, BMI and other variables).

When analysing data of moderate or high dimension, it is often desirable to seeks ways to restructure the data and reduce its dimension *in such a way that we retain the most important information within the data.* In reduced dimensions it is often much easier to understand and appreciate the most important features of a dataset.

In Part II of this module we investigate three different methods for dimension reduction: Principal Components Analysis (PCA) in Chapter 3; Canonical Correlation Analysis (CCA) in Chapter 4; and Multidimensional Scaling (MDS) in Chapter 5.

Matrix algebra (see Chapter 2) plays a key role in all three of these techniques.

Chapter 3

Principal component analysis

Although it is very common to collect multivariate data, we often want to reduce the dimension of such data *in a sensible way*.

For example, exam marks across different modules are averaged to produce a single overall mark for each student. Similarly, in a football league table we convert the numbers of wins, draws and losses to a single measure of points.

Mathematically, we can express these examples of dimension reduction as a linear combination of the original variables, $y = u^\top x$. For the exam mark example, suppose each student sits $p = 4$ modules with marks, x_1, x_2, x_3, x_4. Then, writing $x = (x_1, x_2, x_3, x_4)^\top$ and choosing $u = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)^\top$ gives an overall average,

$$y = u^\top x = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \frac{x_1}{4} + \frac{x_2}{4} + \frac{x_3}{4} + \frac{x_4}{4}.$$

For the football league table, if w is the number of wins, d is the number of draws and l is the number of losses then, writing $\mathbf{r} = (w, d, l)^\top$, we choose $u = (3, 1, 0)^\top$ to get the points score

$$y = u^\top \mathbf{r} = \begin{pmatrix} 3 & 1 & 0 \end{pmatrix} \begin{pmatrix} w \\ d \\ l \end{pmatrix} = 3w + 1d + 0l = 3w + d.$$

In these examples we use u to convert our original variables, the components of x, to a new variable, y. These choices of u are fairly standard for these types

of data. However,we should ask whether we can do better. In a more general setting, how should we choose u?

A key objective of principal component analysis (PCA): to find the linear combination of the original variables that **maximises the variability** in the new variable. Intuitively, this seems sensible for the exam mark data because a large variance in y would separate out the better students from the weaker students, making it easier to rank them.

3.1 Principal component vectors and scores

Let $x_1, \ldots, x_n$ be $p \times 1$ vectors of measurements on n experimental units with sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and sample covariance matrix\ $S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top$.

We wish to project the data onto a lower-dimensional subspace in which the data displays *maximal variation*, using appropriate scalar products of the observation vectors.

Let u be a unit vector (i.e. $\|u\| = 1$ or $u^\top u = 1$) and define

$$y_i = u^\top (x_i - \bar{x})$$

for $i = 1, \ldots, n$.

Now

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} u^\top (x_i - \bar{x}) = u^\top \sum_{i=1}^{n} (x_i - \bar{x}) = u^\top (n\bar{x} - n\bar{x}) = 0,$$

by the definition of $\bar{x}$, so $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 0$.

The sample variance of the y_i's is

$$\begin{aligned} s^2[u] &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[u^\top (x_i - \bar{x})\right]\left[(x_i - \bar{x})^\top u\right] \\ &= u^\top \left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top\right] u \\ &= u^\top S u. \end{aligned}$$

We would like to find the u which maximises the sample variance, $s^2[u] = u^\top S u$ over unit vectors u.

Since S is symmetric, then by the spectral decomposition theorem we can write

$$S = Q\Lambda Q^\top = \sum_{j=1}^{p} \lambda_j q_j q_j^\top$$

with $Q = [q_1, \ldots, q_p]$ an orthogonal matrix (so $QQ^\top = Q^\top Q = I_p$) and $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_p\}$ where we may assume $\lambda_1 \geq \cdots \geq \lambda_p$ and, since S is a covariance matrix and therefore non-negative definite, $\lambda_p \geq 0$. Note that λ_j and q_j, $j = 1, \ldots, p$, are eigenvalues and eigenvectors, respectively, of S.

Then,

$$\begin{aligned} s^2[u] &= u^\top S u = u^\top Q \Lambda Q^\top u = u^\top \left(\sum_{j=1}^{p} \lambda_j q_j q_j^\top \right) u \\ &= \sum_{j=1}^{p} \lambda_j (u^\top q_j)(q_j^\top u) = \sum_{j=1}^{p} \lambda_j (u^\top q_j)^2 \\ &\leq \sum_{j=1}^{p} \lambda_1 (u^\top q_j)^2 \end{aligned}$$

since $\lambda_1 \geq \lambda_j, j = 1, \ldots, p$. Therefore, using Proposition 2.1,

$$s^2[u] \leq \lambda_1 \sum_{j=1}^{p} (u^\top q_j)^2 = \lambda_1 u^\top \left(\sum_{j=1}^{p} q_j q_j^\top \right) u = \lambda_1 u^\top u = \lambda_1,$$

since, by assumption, $\|u\| = 1$.

Therefore, the maximum $s^2[u]$ is at most λ_1, where λ_1 is the largest eigenvalue of S.

Recall that

$$q_i^\top q_j = \begin{cases} 0 & \text{if } j \neq i, \\ 1 & \text{if } j = i. \end{cases}$$

because eigenvectors are orthogonal to each other, so if we take $u = q_1$ then

$$\begin{aligned} q_1^\top S q_1 &= q_1^\top \left(\sum_{j=1}^{p} \lambda_j q_j q_j^\top \right) q_1 = \sum_{j=1}^{p} \lambda_j (q_1^\top q_j)(q_j^\top q_1) \\ &= \sum_{j=1}^{p} \lambda_j (q_1^\top q_j)^2 = \lambda_1 (q_1^\top q_1)^2 = \lambda_1 \end{aligned}$$

So $s^2[u] = u^\top S u$ is maximised over unit vectors u when $u = q_1$ where q_1 is the unit eigenvector corresponding to the largest eigenvalue, λ_1. By maximising $u^\top S u$ over unit vectors u, we are in effect choosing a projection onto a 1-dimensional subspace which captures as much of the sample variation as possible.

We can repeat this procedure and look for the largest sample variance of the y_i's, when u is chosen to be orthogonal to q_1 (i.e. restrict attention to those u such that $u^\top q_1 = 0$). Similar reasoning shows that this constrained maximum

occurs when $u = q_2$, where q_2 is the eigenvector corresponding to the second largest eigenvalue, λ_2; and the corresponding maximum of $u^\top Su$ is λ_2.

We can repeat the process for $j = 1, \dots, p$ to define p new variables. In general, to find PC j, we solve the following optimisation problem:

$$\max_{u:\, ||u||=1} u^\top Su \tag{3.1}$$

subject to

$$q_k^\top u = 0, \qquad k = 1, \dots, j-1. \tag{3.2}$$

It turns out that the maximum of (3.1) subject to (3.2) is equal to λ_j and is obtained when $u = q_j$.

The 1st PC scores are $y_{i1} = q_1^\top (x_i - \bar{x}), \quad i = 1, \dots, n.$ \ The 2nd PC scores are $y_{i2} = q_2^\top (x_i - \bar{x}), \quad i = 1, \dots, n.$

$$\vdots$$

The pth PC scores are $y_{ip} = q_p^\top (x_i - \bar{x}), \quad i = 1, \dots, n.$

We summarise these findings in the following result.

Proposition 3.1. *Let $x_1, \dots, x_n$ denote a sample of vectors in $\mathbb{R}^p$ with sample mean vector $\bar{x}$ and sample covariance matrix S. Suppose S has spectral decomposition (see Proposition 2.5)*

$$S = Q\Lambda Q^\top = \sum_{j=1}^{p} \lambda_j q_j q_j^\top,$$

where Q is orthogonal, $\Lambda = diag\{\lambda_1, \dots, \lambda_p\}$ and $\lambda_1 \geq \lambda_2 \geq \lambda_p \geq 0$. Then the following holds:

1. *The maximum of* (3.1) *subject to* (3.2) *is equal to λ_j and is obtained when $u = q_j$.*
2. *For $j = 1, \dots, p$, the scores of the jth principal component (PC) are given by*

   $$y_{ij} = q_j^\top (x_i - \bar{x}), \qquad i = 1, \dots, n,$$

   where q_j is the vector of loadings *for the jth PC. Moreover,*

   $$y_i = (y_{i1}, y_{i2}, \dots, y_{ip})^\top = Q^\top (x_i - \bar{x}), \qquad i = 1, \dots, n.$$

3. *In matrix form, the full set of PC scores is given in the matrix*

   $$Y = [y_1, \dots, y_n]^\top = HXQ,$$

   where $\overset{n\times n}{H}$ is the centering matrix and $X = [x_1, \dots, x_n]^\top$ is the original data matrix.

4. *The sample mean vector of* $y_1, \ldots, y_n$ *is the zero vector* $\mathbf{0}_p$ *and the sample covariance matrix is* Λ.

Example 3.1. We consider the marks of $n = 10$ students who studied G11PRB and G11STA.

```
## Warning: package 'kableExtra' was built under R version 3.6.2
```

student	PRB	SMM
1	81	75
2	79	73
3	66	79
4	53	55
5	43	53
6	59	49
7	62	72
8	79	92
9	49	58
10	55	56

The sample mean vector and sample covariance matrix are

$$\bar{x} = \begin{pmatrix} 62.6 \\ 66.2 \end{pmatrix} \qquad \text{and} \qquad S = \begin{pmatrix} 162.04 & 135.38 \\ 135.38 & 175.36 \end{pmatrix}.$$

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:kableExtra':
##
##     group_rows
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
secondyr %>% select(2:3) %>% colMeans -> xbar
secondyr %>% select(2:3) %>% cov(use="everything")*9/10 -> S
eigs = eigen(S)
```

DELETE THIS - ASSUME THEY CAN DO IT, OR DO ON A COMPUTER.

To find the eigenvalues we need to solve $|S - \lambda I| = 0$, where

$$\begin{aligned} |S - \lambda I_2| &= (162.04 - \lambda)(175.36 - \lambda) - 135.38^2 \\ &= \lambda^2 - 337.4\lambda + 10887.59. \end{aligned}$$

Using the quadratic equation formula we find,

$$\lambda = \frac{337.4 \pm \sqrt{337.4^2 - 4(10087.59)}}{2} = \frac{337.4 \pm \sqrt{73488.4}}{2}.$$

So $\lambda_1 = 304.24$ and $\lambda_2 = 33.16$.

To find the first eigenvector we solve $(S - \lambda_1 I_2)q_1 = 0$. To simplify, we use row operations:

$$\begin{aligned} S - \lambda_1 I_2 = \begin{pmatrix} -142.20 & 135.38 \\ 135.38 & -128.88 \end{pmatrix} &\rightarrow \begin{pmatrix} 1 & -0.952 \\ 135.38 & -128.88 \end{pmatrix} \\ &\rightarrow \begin{pmatrix} 1 & -0.952 \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

If we let $q_1 = (q_{11}, q_{21})^\top$ then solving $(S - \lambda_1 I_2)q_1 = 0$ is equivalent to solving

$$\begin{pmatrix} 1 & -0.952 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} q_{11} \\ q_{21} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

So $q_{11} = 0.952 q_{21}$ and the eigenvectors are of the form $t \begin{pmatrix} 0.952 \\ 1 \end{pmatrix}$ where $t \neq 0$ is a constant. We choose t such that $\|q\| = 1$, so

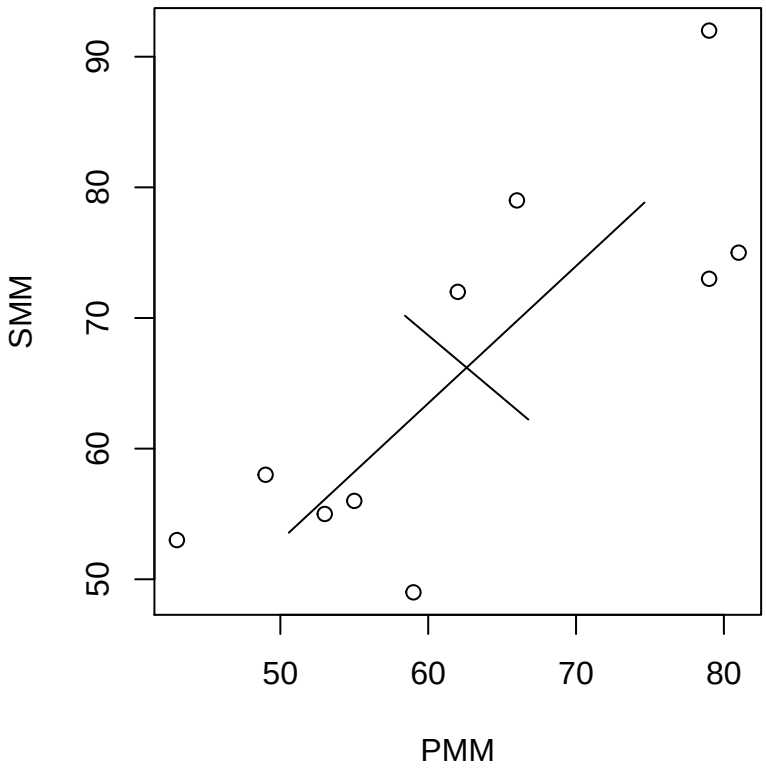
$$t = \pm \frac{1}{\sqrt{0.952^2 + 1^2}} = \pm 0.724.$$

Therefore,

$$q_1 = 0.724 \begin{pmatrix} 0.952 \\ 1 \end{pmatrix} = \begin{pmatrix} 0.690 \\ 0.724 \end{pmatrix}.$$

To find the second eigenvector we use the same method to solve $(S - \lambda_2 I_2)q_2 = 0$ and find that $q_2 = \begin{pmatrix} -0.724 \\ 0.690 \end{pmatrix}$.

The plot below shows the original data. The two lines, centred on $\bar{x}$, have the direction of the eigenvectors, and their lengths are $2\sqrt{\lambda_j}$, $j = 1, 2$.



We can now compute the PC scores using

$$\begin{aligned} y_{i1} &= q_1^\top (x_i - \bar{x}) = 0.690(x_{1i} - \bar{x}_1) + 0.724(x_{2i} - \bar{x}_2) \\ y_{i2} &= q_2^\top (x_i - \bar{x}) = -0.724(x_{1i} - \bar{x}_1) + 0.690(x_{2i} - \bar{x}_2), \end{aligned}$$

which gives

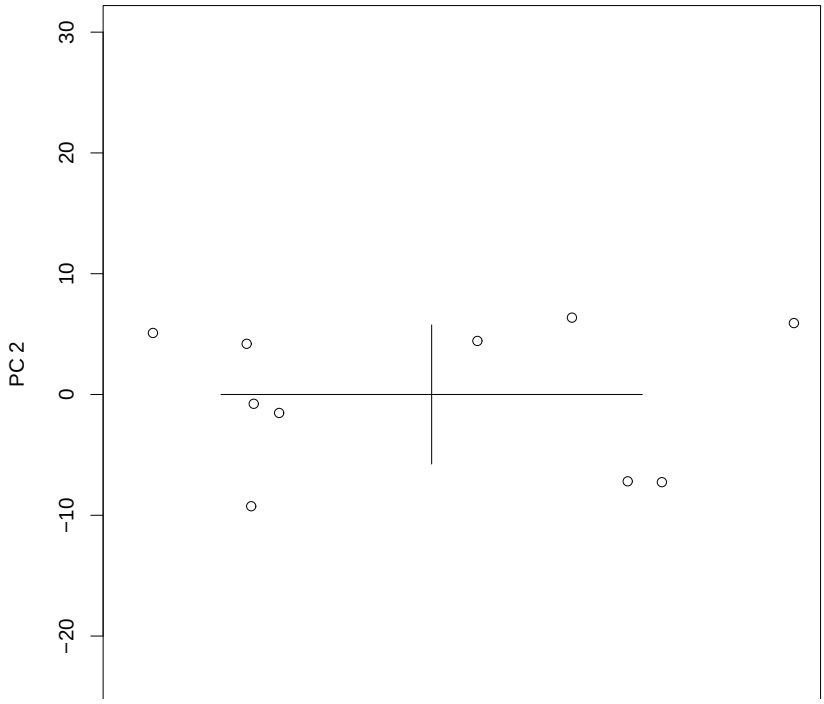
FIX FIX

Student	1	2	3	4	5	6	7	8	9	10
$y_{[1]}^\top$	19.1	16.2	11.6	-14.7	-23.1	-14.9	3.8	30.0	-15.3	-12.6
$y_{[2]}^\top$	-7.3	-7.2	6.4	-0.8	5.1	-9.3	4.4	5.9	4.2	-1.5

Note that these new variables have sample mean $\bar{y} = 0$ and sample covariance matrix (see part 4. of Proposition 3.1)

$$\Lambda = \mathrm{diag}(\lambda_1, \lambda_2) = \begin{pmatrix} 304.24 & 0 \\ 0 & 33.16 \end{pmatrix}.$$

The plot below shows the PC scores $(y_{i1}, y_{i2})^\top$. The two lines shown have lengths $2\sqrt{\lambda_j}$, $j = 1, 2$. Note that $\sqrt{\lambda_j}$ is the standard deviation of the jth PC.



Sometimes the new variables have an obvious interpretation. Note that the first PC gives positive, roughly equal, weight to PRB and STA and thus represents some form of "average" mark. For example, a student that has a high mark on PRB and STA will have a high value for y_1. The second PC, meanwhile, represents a contrast between PRB and STA. For example, a large positive value for y_2 implies the student did much better on STA than PRB, and a large negative value implies the opposite.

Note that we could have chosen $t = -0.724$ instead of $t = +0.724$. The only difference would be that the first eigenvector was $q_1^* = -q_1$. In this case, a student who scored a high mark on PRB and STA would have a low value for y_1. This is perfectly legitimate but makes the interpretation less intuitive. One can always change the sign of the eigenvectors if it makes interpretation easier. "'

3.2 Properties of principal components

Let $x_1, \ldots, x_n$ have sample mean $\bar{x}$ and sample covariance matrix S, with spectral decomposition $S = Q\Lambda Q^\top$ where $Q = [q_1, \ldots, q_p]$ is orthogonal and $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_p\}$. The transformed variables have some important properties.

Proposition 3.2. *For $j, k = 1, \ldots, p$, the following results hold.*

1. $\bar{y}_{+j} = n^{-1}\sum_{i=1}^{n} y_{ij} = n^{-1}\sum_{i=1}^{n} q_j^\top (x_i - \bar{x}) = 0$;
2. $q_j^\top S q_j = \lambda_j$;
3. $q_j^\top S q_k = 0$ *for* $j \neq k$;
4. $q_1^\top S q_1 \geq q_2^\top S q_2 \geq ... \geq q_p^\top S q_p \geq 0$;
5. $\sum_{j=1}^{p} q_j^\top S q_j = \sum_{j=1}^{p} \lambda_j = tr(S)$;
6. $\prod_{j=1}^{p} q_j^\top S q_j = \prod_{j=1}^{p} \lambda_j = |S|$.

In words:

- part 1. tells us that the sample mean of $y_{1j}, ..., y_{nj}$ for each fixed j is 0;
- part 2. tells us that, for each fixed j, the sample variance of the y_{ij}, $i = 1, ..., n$ is λ_j;
- part 3. states that the sample covariance of the pairs (y_{ij}, y_{ik}), $i = 1, ..., n$, is 0 if $j \neq k$;
- part 4. states that the sample variance of y_{ij}, $i = 1, ..., n$, is not less than the sample variance of y_{ik}, $i = 1, ..., n$, if $j \leq k$;
- part 5. states that the sum of the sample variances is equal to the trace of S;
- and part 6. states that the product of the sample variances is equal to the determinant of S.

From these properties we say that a proportion

$$\frac{\lambda_j}{\lambda_1 + ... + \lambda_p}$$

of the variability in the sample is 'explained' by the jth PC.

For the G11PRB and G11STA data above,

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{304.24}{304.24 + 33.16} = 0.90,$$

so 90% of the variability in the sample is explained by the 1st PC.

"'{Example} We can apply PCA to a football league table where W, D, L are the number of matches won, drawn and lost and F and A are the goals scored for and against. An extract of the table for a recent Premiership season is: FIX FIX

Team	W	D	L	F	A
Chelsea	27	5	6	103	32
Manchester United	27	4	7	86	28
Arsenal	23	6	9	83	41
Tottenham Hotspur	21	7	10	67	41
Manchester City	18	13	7	73	45

The sample mean vector is

$$\bar{x} = \begin{pmatrix} 14.2 \\ 9.6 \\ 14.2 \\ 52.6 \\ 52.6 \end{pmatrix}$$

and the sample covariance matrix is

$$S = \begin{pmatrix} 39.4 & -8.27 & -31.1 & 116 & -81.9 \\ -8.27 & 8.14 & 0.13 & -29.4 & 6.01 \\ -31.1 & 0.13 & 31 & -86.3 & 75.9 \\ 116 & -29.4 & -86.3 & 392 & -209 \\ -81.9 & 6.01 & 75.9 & -209 & 231 \end{pmatrix} \tag{3.3}$$

The eigenvalues of S are

$$\Lambda = \text{diag}\begin{pmatrix} 631 & 96.7 & 8.83 & 2.44 & -4.97e-14 \end{pmatrix}$$

Note that we have a zero eigenvalue because one of our variables is a linear combination of the other variables, $L = 38 - W - D$. The corresponding eigenvectors are

$$Q = [q_1 \dots q_5] = \begin{pmatrix} 0.251 & -0.0133 & -0.116 & 0.768 & 0.577 \\ -0.0477 & -0.146 & 0.74 & -0.309 & 0.577 \\ -0.204 & 0.16 & -0.624 & -0.459 & 0.577 \\ 0.776 & 0.582 & 0.0674 & -0.234 & -2e-15 \\ -0.539 & 0.784 & 0.213 & 0.222 & 1.83e-15 \end{pmatrix}$$

The proportion of variability explained by each of the PCs is:

$$\begin{pmatrix} 0.854 & 0.131 & 0.012 & 0.0033 & -6.73e-17 \end{pmatrix}$$

There is no point computing the scores for PC 5 because PC5 does not explain any of the variability in the data. Similarly, there is little value in computing the scores for PCs 3 & 4 because they only account for 1.5% of the variability in the data.

We can, therefore, choose to compute only the first two PC scores. We are reducing the dimension of our data set from $p = 5$ to $p = 2$ while still retaining 98.5% of the variability. The first PC is given by:

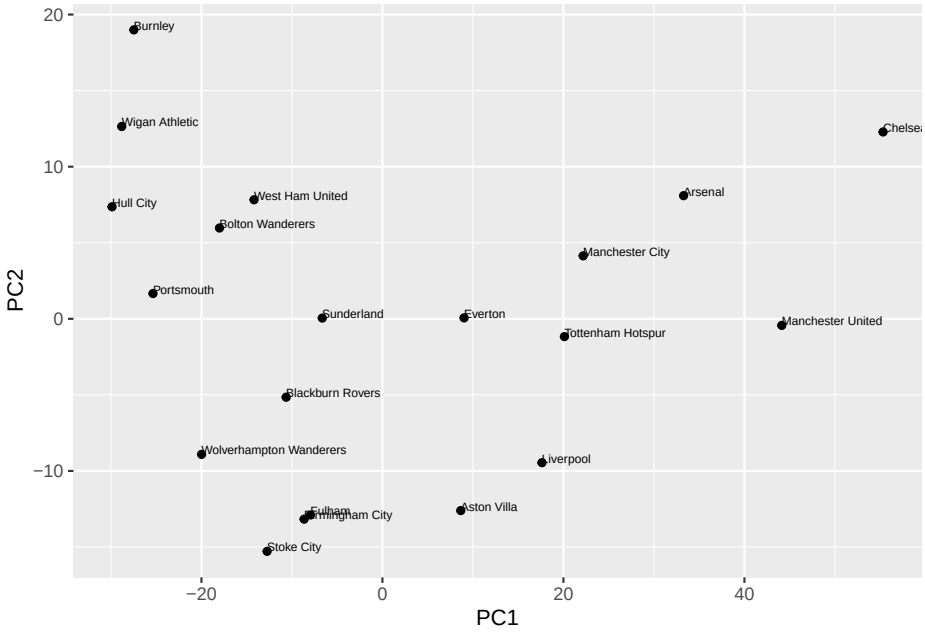
$$\begin{aligned} y_{i1} = 0.25(W_i - \bar{W}) + -0.05(D_i - \bar{D}) + -0.2(L_i - \bar{L}) \\ + 0.78(F_i - \bar{F}) + -0.54(A_i - \bar{A}), \end{aligned}$$

and similarly for PC 2.

The first five rows of our revised "league table" are now

Team	PC1	PC2
Chelsea	55.3	12.3
Manchester United	44.1	-0.4
Arsenal	33.3	8.1
Tottenham Hotspur	20.1	-1.2
Manchester City	22.2	4.1

Now that we have reduced the dimension to $p = 2$, we can visualise the differences between the teams.



We might interpret the PCs as follows. The first PC seems to measure overall performance. It rewards teams with 0.78 for every goal they score and 0.25 for every match they win, while penalising them by 0.54 for every goal they concede, 0.2 for every match they lose and 0.05 for every match they draw.

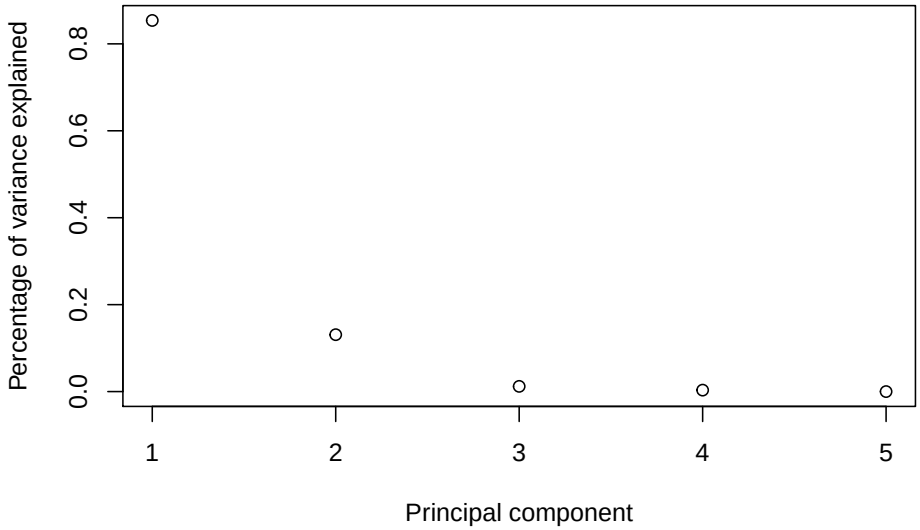
We could, therefore, rank teams by PC 1 and compare this with the rankings using 3 points for a win and 1 point for a draw. The rankings are the same for the top three teams but differ below that. Under our system Wigan would be relegated in place of Portsmouth.

The second PC has a strong negative loading for both goals for and against. A team with a large negative PC 2 score was, therefore, involved in matches

with lots of goals. We could, therefore, interpret PC 2 as an "entertainment" measure, ranking teams according to their involvement in high-scoring games.

The above example raises the question of how many PCs should we use in practice. If we reduce the dimension to $p = 1$ then we can rank observations and analyse our new variable with univariate statistics. If we reduce the dimension to $p = 2$ then it is still easy to visualise the data. However, reducing the dimension to $p = 1$ or $p = 2$ may involve losing lots of information and a sensible answer should depend on the objectives of the analysis and the data itself.

One tool for looking at the contributions of each PC is to look at the **scree graph** which plots the percentage of variance explained by PC j against j. The scree graph for the football example is:



Possible methods for choosing the number of PCs include:

- retain enough PCs to explain, say, 90% of the total variation;
- retain PCs where the eigenvalue is above the average.

For the football example, the first method would retain 2 PCs whereas the second method would only retain 1 PC.

"'

3.3 Population PCA

So far we have considered sample PCA based on the sample covariance matrix

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

We note now that there is a *population* analogue of PCA based on the population covariance matrix Σ. Although the population version of PCA is not of as much direct practical relevance as sample PCA, it is nevertheless of conceptual importance.

Let x denote a $p \times 1$ random vector with $E(x) = \boldsymbol{\mu}$ and $\text{Var}(x) = \boldsymbol{\Sigma}$. As defined, $\boldsymbol{\mu}$ is the population mean vector and $\boldsymbol{\Sigma}$ is the population covariance matrix.

Since $\boldsymbol{\Sigma}$ is symmetric, the spectral decomposition theorem tells us that

$$\boldsymbol{\Sigma} = \sum_{j=1}^{p} \check{\lambda}_j \check{q}_j \check{q}_j^\top = \check{Q} \check{\Lambda} \check{Q}^\top$$

where the 'check' symbol $\check{}$ is used to distinguish population quantities from their sample analogues.

Then:

- the first population PC is defined by $Y_1 = \check{q}_1^\top (x - \boldsymbol{\mu})$; -the second population PC is defined by $Y_2 = \check{q}_2^\top (x - \boldsymbol{\mu})$;
- \$ldots\$
- the pth population PC is defined by $Y_p = \check{q}_p^\top (x - \boldsymbol{\mu})$.

The $Y_1, \ldots, Y_p$ are random variables, unlike the sample PCA case, where the y_{ij} are observed quantities. In the sample PCA case, the y_{ij} can often be regarded as the observed values of random variables.

In matrix form, the above definitions can be summarised by writing

$$y = \begin{pmatrix} Y_1 \\ Y_2 \\ \ldots \\ \ldots \\ Y_p \end{pmatrix} = \check{Q}^\top (x - \boldsymbol{\mu}).$$

The population PCA analogues of the 6 sample PCA properties listed in Proposition 3.2 are now given. Note that the Y_j's are random variables as opposed to observed values of random variables.

Proposition 3.3. *The following results hold for the random variables $Y_1, \ldots, Y_p$ defined above.*

1. $E(Y_j) = 0$ *for* $j = 1, \ldots, p;|$
2. $Var(Y_j) = \check{\lambda}_j$ *for* $j = 1, \ldots, p;|$
3. $Cov(Y_j, Y_k) = 0$ *if* $j \neq k;|$
4. $Var(Y_1) \geq Var(Y_2) \geq \cdots \geq Var(Y_p) \geq 0;|$
5. $\sum_{j=1}^{p} Var(Y_j) = \sum_{j=1}^{p} \check{\lambda}_j = tr(\Sigma);|$

6. $\prod_{j=1}^{p} Var(Y_j) = \prod_{j=1}^{p} \check{\lambda}_j = |\Sigma|$.

Note that, defining $y = (Y_1, \ldots, Y_p)^\top$ as before, part 1. implies that $E(y) = \mathbf{0}_p$ and parts 2. and 3. together imply that

$$\mathrm{Var}(y) = \Lambda \equiv \mathrm{diag}(\check{\lambda}_1, \ldots, \check{\lambda}_p).$$

Example 3.2. Suppose

$$\Sigma = I_p + \delta \mathbf{1}_p \mathbf{1}_p^\top,$$

where $\delta > 0$. What is the proportion of variability explained by the first PC? Since $\delta > 0$, the largest eigenvalue is $\lambda_1 = 1 + p\delta$ which is achieved when $\check{q}_1$ is the unit vector $p^{-1/2}\mathbf{1}_p$. This and related examples are dealt with in more detail in the example sheets.

Consider now a repeated sampling framework in which we assume that $x_1, \ldots, x_n$ are IID random vectors from a population with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.

What is the relationship between the sample PCA based on the sample of observed vectors $x_1, \ldots, x_n$, and the population PCA based on the unobserved random vector x, from the same population?

Assuming n is large, and the elements of Σ are all finite, the elements of the sample covariance matrix S will be close with high probability to the corresponding elements of the population covariance matrix Σ. Justification of this statement comes from the weak law of large numbers applied to the components of Σ (details omitted).

Consequently, when n is large, sample PCA and the corresponding population PCA may be expected to give similar results.

3.4 An Alternative Derivation of PCA

Consider a sample $x_1, \ldots, x_n \in \mathbb{R}^p$.

Recall from §2.8 that any line in $\mathbb{R}^p$ may be written in the form $\{a + ub : u \in \mathbb{R}\}$ where $a, b \in \mathbb{R}^p$ are fixed.

Here we consider the following problem: *find the best-fitting line to the sample* $x_1, \ldots, x_n$.

We first formulate this problem more precisely. Define the function

$$\begin{aligned} F(a, b; u_1, \ldots, u_n) &= \sum_{i=1}^{n} ||x_i - a - u_i b||^2 \\ &= \sum_{i=1}^{n} (x_i - a - u_i b)^\top (x_i - a - u_i b). \end{aligned}$$

We wish to solve the following problem:

$$\begin{aligned} &\textit{minimise } F(a, b; u_1, \ldots, u_n) \textit{ subject to the} \\ &\textit{constraints that } a \textit{ and } b \textit{ are orthogonal, i.e. } a^\top b = 0, \\ &\textit{and } b \textit{ is a unit vector, i.e. } ||b|| = 1. \end{aligned} \tag{3.4}$$

Theorem 3.1. *The solution to optimisation problem* (3.4) *is given by*

$$\hat{a} = \left(\mathbf{I}_p - q_1 q_1^\top\right) \bar{x}, \quad \hat{b} = q_1 \quad \textit{and} \quad \hat{u}_i = x_i^\top q_1, \quad i = 1, \ldots, n, \tag{3.5}$$

where $\bar{x} = n^{-1} \sum_{i=1}^n x_i$ *is the sample mean and the unit vector* q_1 *is the direction of the first sample PC.*

Note that the quantities $u_i - \bar{u} = q_1^\top (x_i - \bar{x})$ are the PC scores associated with the first PC.

Proof. The proof is broken into two steps.

Step 1. In Step 1, we want to minimise $F(a, b; u_1, \ldots, u_n)$ subject to the constraint $a^\top b = 0$, with b an arbitrary *fixed* unit vector in $\mathbb{R}^p$. So we introduce a Lagrangian term for the constraint $a^\top b = 0$ and minimise

$$\bar{F}(a; u_1, \ldots, u_n; \gamma) \equiv \left\{ \sum_{i=1}^n (x_i - a - u_i b)^\top (x_i - a - u_i b) \right\} + \gamma a^\top b$$

over a, $u_1, \ldots, u_n$ and γ. Then, for $i = 1, \ldots, n$,

$$\frac{\partial \bar{F}}{\partial u_i} = -2 b^\top (x_i - a - u_i b); \tag{3.6}$$

$$\begin{aligned} \frac{\partial \bar{F}}{\partial a} &= -2 \left\{ \sum_{i=1}^n (x_i - a - u_i b) \right\} + \gamma b \\ &= -2n \{ \bar{x} - a - (\bar{u} + \gamma/(2n)) b \}, \end{aligned} \tag{3.7}$$

where $\bar{u} = n^{-1} \sum_{i=1}^n u_i$; and

$$\frac{\partial \bar{F}}{\partial \gamma} = a^\top b. \tag{3.8}$$

Setting the partial derivatives (3.6), (3.7) and @ref(eq:barF3}) to zero,

$$\frac{\partial \bar{F}}{\partial \gamma} = 0 \implies \hat{a}^\top b = 0;$$

$$\frac{\partial \bar{F}}{\partial u_i} = 0 \implies \hat{u}_i = b^\top x_i, \tag{3.9}$$

and therefore

$$\bar{\hat{u}} \equiv n^{-1} \sum_{i=1}^{n} \hat{u}_i = b^\top \bar{x}; \tag{3.10}$$

and

$$\frac{\partial \bar{F}}{\partial a} = \mathbf{0}_p \implies \hat{a} = \bar{x} - \{\bar{\hat{u}} + \hat{\gamma}/(2n)\}b.$$

Using (3.10) and the fact that $b^\top \hat{a} = 0$, it follows that

$$0 = b^\top \hat{a} = b^\top [\bar{x} - \{\bar{\hat{u}} + \hat{\gamma}/(2n)\}b] = b^\top \bar{x} - b^\top \bar{x} + \hat{\gamma}/(2n),$$

which implies that $\hat{\gamma} = 0$. Consequently,

$$\hat{a} = \bar{x} - \bar{\hat{u}}b = \bar{x} - bb^\top \bar{x} = \left(I_p - bb^\top\right)\bar{x}; \tag{3.11}$$

and so

$$\begin{aligned}
&\bar{F}(\hat{a}; \hat{u}_1, \ldots, \hat{u}_n; \hat{\gamma}) \qquad (3.12)\\
&= \sum_{i=1}^{n} (x_i - \hat{a} - \hat{u}_i b)^\top (x_i - \hat{a} - \hat{u}_i b)\\
&= \sum_{i=1}^{n} \left\{x_i - (\mathbf{I}_p - bb^\top)\bar{x} - bb^\top x_i\right\}^\top \left\{x_i - (\mathbf{I}_p - bb^\top)\bar{x} - bb^\top x_i\right\}\\
&\sum_{i=1}^{n} (x_i - \bar{x})^\top (\mathbf{I}_p - bb^\top)^2 (x_i - \bar{x})\\
&= n\mathrm{tr}\left\{(\mathbf{I}_p - bb^\top)S\right\}\\
&= n\left\{\mathrm{tr}(S) - b^\top S b\right\}, \qquad (3.13)
\end{aligned}$$

where S is the sample covariance of the x_i.

Step 2. We now minimise (3.13) over unit vectors $b \in \mathbb{R}^p$. But minimising (3.13) is equivalent to maximising $b^\top S b$, so from Proposition 2.7, $\hat{b} = q_1$, and so $\hat{a} = (I_p - q_1 q_1^\top)\bar{x}$ from (3.11), and from (3.9), $\hat{u}_i = q_1^\top x_i$, $i = 1, \ldots, n$, all of which agrees with the expressions in (3.5).

□

3.5 PCA under transformations of variables

Let us return to the example of $n = 10$ students who studied G11PRB and G11STA. Earlier, we calculated the sample mean, sample variance matrix and the eigenvalues/vectors of S,

$$\bar{x} = \begin{pmatrix} 62.6 \\ 66.2 \end{pmatrix}, \qquad S = \begin{pmatrix} 162.04 & 135.38 \\ 135.38 & 175.36 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 304.24 & 0 \\ 0 & 33.16 \end{pmatrix}, \qquad Q = \begin{pmatrix} 0.690 & -0.724 \\ 0.724 & 0.690 \end{pmatrix}$$

with PC 1 scores

$$y_i = q_1^\top (x_i - \bar{x}) = 0.690(x_{1i} - \bar{x}_1) + 0.724(x_{2i} - \bar{x}_2).$$

We now consider what happens to the above quantities under various transformations of the x_i, the 2×1 response vectors.

Addition transformation

Firstly, we consider the transformation of addition where, for example, the G11PRB lecturer decides to add 5 marks for all the students. We can write this transformation as $z_i = x_i + c$, where c is a fixed vector. Under this transformation the sample mean changes, $\bar{z} = \bar{x} + c$, but the sample variance remains S. Consequently, the eigenvalues and eigenvectors of S remain the same and, therefore, so does the PC 1 score,

$$y_i = q_1^\top (z_i - \bar{z}) = q_1^\top (x_i + c - (\bar{x} + c)) = q_1^\top (x_i - \bar{x}).$$

We say that the principal components are **invariant** under the addition transformation. An important special case is to choose $c = -\bar{x}$ so that the PC 1 score is simply $y_i = q_1^\top z_i$.

Scale transformation

Secondly, we consider the scale transformation where, for example, the G11PRB lecturer decides to double the marks for all students. A scale transformation occurs more naturally when we convert units of measurement from, say, metres to kilometres. We can write this transformation as $z_i = Dx_i$, where D is a diagonal matrix with positive elements. Under this transformation the sample mean changes from $\bar{x}$ to $\bar{z} = D\bar{x}$, and the sample covariance matrix changes from S to DSD. Consequently, the principal components also change.

This lack of scale-invariance is undesirable. One solution is to choose

$$D = \text{diag}(s_{11}^{-1/2}, \ldots, s_{pp}^{-1/2}),$$

where s_{ii} is the ith diagonal element of S. In effect, we have standardised all the new variables to have variance 1. In this case the sample covariance matrix of the z_i's is simply the sample correlation matrix of the original variables, x_i. Therefore, we can carry out PCA on the sample correlation matrix, R, which is invariant to changes of scale.

In summary: R is scale-invariant while S is not.

Example 3.3. For the G11PRB/G11STA data, we choose

$$D = \text{diag}(162.04, 175.36)^{-1/2} = \text{diag}(0.079, 0.076)$$

so that $z_i = Dx_i$. The sample correlation matrix is then

$$\begin{aligned} R &= DSD \\ &= \begin{pmatrix} 0.079 & 0 \\ 0 & 0.076 \end{pmatrix} \begin{pmatrix} 162.04 & 135.38 \\ 135.38 & 175.36 \end{pmatrix} \begin{pmatrix} 0.079 & 0 \\ 0 & 0.076 \end{pmatrix} \\ &= \begin{pmatrix} 1.000 & 0.803 \\ 0.803 & 1.000 \end{pmatrix}. \end{aligned}$$

The eigenvalues and eigenvectors of R are then

$$\Lambda = \begin{pmatrix} 1.803 & 0 \\ 0 & 0.197 \end{pmatrix}, \qquad Q = \begin{pmatrix} 0.707 & 0.707 \\ 0.707 & -0.707 \end{pmatrix},$$

and the PC 1 score is

$$\begin{aligned} y_i &= q_1^\top (z_i - \bar{z}) = q_1^\top D(x_i - \bar{x}) \\ &= 0.707 \times 0.079(x_{1i} - \bar{x}_1) + 0.707 \times 0.076(x_{2i} - \bar{x}_2). \end{aligned}$$

In the example above, there is little difference between using S and R for the PCA because the variances for G11PRB and G11STA are similar. In other cases, particularly when the variables are measured on wildly different scales, the difference will be notable. For example, in the football data the sample variances of F and A are much larger than the sample variances of W, D and L.

Orthogonal transformation

Thirdly, we consider a transformation by an orthogonal matrix, $\underset{p\times p}{A}$, such that $AA^\top = A^\top A = I_p$, and write $z_i = Ax_i$. This is equivalent to rotating and/or reflecting the original data.

Let S be the sample covariance matrix of the x_i and let T be the sample covariance matrix of the z_i. Under this transformation the sample mean changes from $\bar{x}$ to $\bar{z} = A\bar{x}$, and the sample covariance matrix S changes from S to $T = ASA^\top$.

However, if we write S in terms of its spectral decomposition $S = Q\Lambda Q^\top$, then $T = AQ\Lambda Q^\top A^\top = B\Lambda B^\top$ where $B = AQ$ is also orthogonal. It is therefore apparent that the eigenvalues of T are the same as those of S; and the eigenvectors of T are given by b_j where $b_j = Aq_j$, $j = 1, \ldots, p$. The PC 1 scores of the transformed variables are

$$y_i = b_1^\top (z_i - \bar{z}) = q_1^\top A^\top A(x_i - \bar{x}) = q_1^\top (x_i - \bar{x}),$$

and so they are identical to the PC 1 scores of the original variables.

Therefore, under an orthogonal transformation the eigenvalues and PC scores are unchanged and the PCs are orthogonal transformations of the original PCs. We say that the principal components are **equivariant** with respect to orthogonal transformations.

Example 3.4. Suppose we rotate the G11PRB/G11STA data by the matrix $A = \begin{pmatrix} 0.866 & -0.500 \\ 0.500 & 0.866 \end{pmatrix}$. The sample covariance matrix of the rotated data is

$$\begin{aligned} T &= ASA^\top \\ &= \begin{pmatrix} 0.866 & -0.500 \\ 0.500 & 0.866 \end{pmatrix} \begin{pmatrix} 162.04 & 135.38 \\ 135.38 & 175.36 \end{pmatrix} \begin{pmatrix} 0.866 & 0.500 \\ -0.500 & 0.866 \end{pmatrix} \\ &= \begin{pmatrix} 48.13 & 61.92 \\ 61.92 & 289.27 \end{pmatrix}. \end{aligned}$$

The eigenvalues of T are 304.24 and 33.16 (same as for S). The eigenvectors of T are then

$$\begin{aligned} B &= AQ = \begin{pmatrix} 0.866 & -0.500 \\ 0.500 & 0.866 \end{pmatrix} \begin{pmatrix} 0.690 & -0.724 \\ 0.724 & 0.690 \end{pmatrix} \\ &= \begin{pmatrix} 0.235 & -0.972 \\ 0.972 & 0.235 \end{pmatrix} \end{aligned}$$

and the PC 1 scores are unchanged.

3.6 PCA based on S versus PCA based on R

Recall the distinction between the sample covariance matrix S and the sample correlation matrix R.

Note that all correlation matrices are also covariance matrices, but not all covariance matrices are correlation matrices.

So in practice we have a choice of using S or R for PCA. As we have seen, PCA based on R is scale invariant, but PCA based on S is not; while PCA based on S is invariant (eigenvalues and PC scores) and equivariant (eigenvectors) under orthogonal transformation, whereas R is not.

This raises the important practical question: for a given dataset, should we use PCA based on S or R?

If the p variables represent very different types of quantity or show marked differences in variances, then it will usually be better to use R rather than S. However, in some circumstances, we may wish to use S, such as when the p variables are measuring similar entities and the sample variances are not too different.

Bearing in mind that the required numerical calculations are so easy to perform in R, we might wish to do it both ways and see if it makes much difference.

Chapter 4

Canonical Correlation Analysis

Suppose we observe a random sample of n bivariate observations

$$z_1 = (x_1, y_1)^\top, \ldots, z_n = (x_n, y_n)^\top.$$

If we are interested in exploring possible dependence between the x_i's and y_i's then among the first things we would do would be to obtain a scatterplot of the x_i's against the y_i's and calculate the correlation coefficient. Recall that the sample correlation coefficient is defined by

$$r = r[x, y] = \frac{n^{-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\left(n^{-1} \sum_{i=1}^n (x_i - \bar{x})^2\right)^{1/2} \left(n^{-1} \sum_{i=1}^n (y_i - \bar{y})^2\right)^{1/2}} \tag{4.1}$$

where $\bar{x} = n^{-1} \sum_{i=1}^n x_i$ and $\bar{y} = n^{-1} \sum_{i=1}^n y_i$ are the sample means. Note that the sample correlation is a **scale-free measure** of the strength of **linear dependence** between the x_i's and the y_i's.

In this chapter we investigate the multivariate analogue of this question. Suppose

$$z_i = (x_i^\top, y_i^\top)^\top, \qquad i = 1, \ldots, n,$$

is a random sample of vectors. What is a sensible way to assess and describe the strength of the linear dependence between the x_i vectors and the y_i vectors? That is what this chapter is about. A key role is played by the singular valued decomposition (SVD) introduced in Result 2.13 in Chapter 2.

Example 4.1. From time to time we will return to the Premier League example in this chapter. We shall treat W and D, the number of wins and draws, respectively, as the x-variables; and F and A, the number of goals for and against, will be treated as the y-variables. The number of losses, L, is omitted

as it provides no additional information when we know W and D. A question we shall consider is: how strongly associated are the match outcome variables, W and D, with the goals for and against variables, F and A?

4.1 Canonical Correlation Analysis

Assume we are given a random sample of vectors

$$z_i = (x_i^\top, y_i^\top)^\top : i = 1, \ldots, n,$$

where the x_i are $p \times 1$, the y_i are $q \times 1$ and, consequently, the z_i are $(p+q) \times 1$. We are interested in determining the strength of linear association between the x_i vectors and the y_i vectors.

Write

$$\bar{z} = n^{-1} \sum_{i=1}^{n} z_i, \qquad \bar{x} = n^{-1} \sum_{i=1}^{n} x_i \qquad \text{and} \qquad \bar{y} = n^{-1} \sum_{i=1}^{n} y_i$$

for the sample mean vectors of the z_i, x_i and y_i respectively.

We formulate this task as an optimisation problem (cf. PCA). First, we introduce some notation. Let S_{zz} denote the sample covariance matrix of the z_i, $i = 1, \ldots, n$. Then S_{zz} can be written in block matrix form

$$S_{zz} = \begin{bmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{bmatrix},$$

where S_{xx} $(p \times p)$ is the sample covariance matrix of the x_i, S_{yy} $(q \times q)$ is the sample covariance of the y_i, and the cross-covariance matrices are given by

$$\overset{p \times q}{S}_{xy} = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})^\top \qquad \text{and} \qquad \overset{q \times p}{S}_{yx} = S_{xy}^\top.$$

Example 4.1 (continued). The relevant covariance matrix here is given in (3.3), but we need to delete the middle row and middle column because this relates to the variable L, the number of losses, which we are omitting. So we are left with

$$S_{xx} = \begin{pmatrix} 39.4 & -8.3 \\ -8.3 & 8.1 \end{pmatrix}, \qquad S_{yy} = \begin{pmatrix} 392.2 & -208.7 \\ -208.7 & 230.9 \end{pmatrix} \tag{4.2}$$

and

$$S_{xy} = S_{yx}^\top = \begin{pmatrix} 115.7 & -81.9 \\ -29.4 & 6.0 \end{pmatrix}. \tag{4.3}$$

We shall return to this example in a little while.

We want to find the linear combination of the x-variables and the linear combination of the y-variables which is most highly correlated.

One version of the optimisation problem we want to solve is: find non-zero vectors $\overset{p\times 1}{a}$ and $\overset{q\times 1}{b}$ which maximise the correlation coefficient

$$r[a^\top x, b^\top y] = \frac{a^\top S_{xy} b}{(a^\top S_{xx} a)^{1/2} (b^\top S_{yy} b)^{1/2}}.$$

In other words:

$$\begin{aligned} &\text{Find non-zero vectors} \quad a \ (p \times 1) \text{ and } b \ (q \times 1) \\ &\text{to maximise} \qquad r[a^\top x, b^\top y], \end{aligned} \tag{4.4}$$

where $r[.,.]$ is defined in (4.1). Intuitively, this makes sense, because we want to find the linear combination of the x-variables and the linear combination of the y-variables which are most highly correlated.

However, note that for any $\gamma > 0$ and $\delta > 0$,

$$r[\gamma a^\top x, \delta b^\top y] = \frac{\gamma\delta}{\sqrt{\gamma^2\delta^2}} r[a^\top x, b^\top y] = r[a^\top x, b^\top y], \tag{4.5}$$

i.e. $r[a^\top x, b^\top y]$ is invariant with respect to positive scalar multiplication of a and b. Consequently there will be an infinite number of solutions to this optimisation problem, because if a and b are solutions to optimization problem (4.4), then so are γa and δb, for any $\gamma > 0$ and $\delta > 0$.

A more useful way to formulate this optimisation problem is the following: find

$$\max_{a,b} a^\top S_{xy} b \tag{4.6}$$

subject to the constraints

$$a^\top S_{xx} a = 1 \qquad \text{and} \qquad b^\top S_{yy} b = 1. \tag{4.7}$$

Proposition 4.1. *Assume that S_{xx} and S_{yy} both are non-singular. Then the following holds.*

1. *If $a = \hat{a}$ and $b = \hat{b}$ maximise* (4.4)*, then*

$$a = \breve{a} \equiv \hat{a}/(\hat{a}^\top S_{xx} \hat{a})^{1/2} \qquad \text{and} \qquad b = \breve{b} \equiv \hat{b}/(\hat{b}^\top S_{yy} \hat{b})^{1/2}$$

maximise (4.6) *subject to the constraints* (4.7)*. Moreover, if $a = \breve{a}$ and $b = \breve{b}$ maximise* (4.6) *subject to constraints* (4.7) *then, for any $\gamma > 0$ and $\delta > 0$, $a = \gamma\breve{a}$ and $b = \delta\breve{b}$ maximise* (4.4)*.*

2. *The optimum solution to* (4.6) *and* (4.7) *is obtained when* $a = S_{xx}^{-1/2}\mathbf{q}_1$ *and* $b = S_{yy}^{-1/2}\mathbf{r}_1$, *where* $S_{xx}^{-1/2}S_{xy}S_{yy}^{-1/2}$ *has SVD*

$$A \equiv S_{xx}^{-1/2}S_{xy}S_{yy}^{-1/2} = \sum_{j=1}^{t} \xi_j \mathbf{q}_j \mathbf{r}_j^\top \equiv \mathbf{Q}\Xi\mathbf{R}^\top, \tag{4.8}$$

where A *has rank* t *and* $\xi_1 \geq \cdots \geq \xi_t > 0$.

3. *The maximum value of the correlation coefficient is given by the largest singular value* ξ_1.

Note: the matrix square roots $S_{xx}^{-1/2}$ and $S_{yy}^{-1/2}$ of S_{xx}^{-1} and S_{yy}^{-1}, respectively, are defined using the definition of matrix square roots of symmetric non-negative definite matrices given in Chapter 2.

Proof. (i) In (4.5) it was noted that, for $a \neq \mathbf{0}_p$ and $b \neq \mathbf{0}_q$, the expression for $r[a^\top x, b^\top y]$ is invariant when we change a to γa and change b to δb, where $\gamma > 0$ and $\delta > 0$ are scalars, so the second statement in Result 4.1(i) follows imnmdeiately. Suppose now a solution to problem (4.4) is achieved when $a = \hat{a}$ and $b = \hat{b}$. Then, due to the invariance with respect to rescaling, the optimum is also achieved when $a = \check{a} \equiv \hat{a}/(\hat{a}^\top S_{xx}\hat{a})^{1/2}$ and $b = \check{b} \equiv \hat{b}/(\hat{b}^\top S_{yy}\hat{b})^{1/2}$. But by definition of $\check{a}$ and $\check{b}$, they satisfy the constraints (4.7) because

$$\check{a}^\top S_{xx}\check{a} = \frac{\hat{a}^\top S_{xx}\hat{a}}{\left\{\left(\hat{a}^\top S_{xx}\hat{a}\right)^{1/2}\right\}^2} = \frac{\hat{a}^\top S_{xx}\hat{a}}{\hat{a}^\top S_{xx}\hat{a}} = 1$$

and, similarly,

$$\check{b}^\top S_{yy}\check{b} = \frac{\hat{b}^\top S_{yy}\hat{b}}{\hat{b}^\top S_{yy}\hat{b}} = 1.$$

So $a = \check{a}$ and $b = \check{b}$ maximises (4.6) subject to the constraints (4.7).

(ii) & (iii) We may write the constraints (4.7) as

$$\tilde{a}^\top\tilde{a} = 1 \qquad \text{and} \qquad \tilde{b}^\top\tilde{b} = 1$$

where

$$\tilde{a} = S_{xx}^{1/2}a \qquad \text{and} \qquad \tilde{b} = S_{yy}^{1/2}b.$$

Recall that S_{xx} and S_{yy} are assumed to be non-singular. Then, using results from Chapter 2, $S_{xx}^{1/2}$ and $S_{yy}^{1/2}$ will also be non-singular, and so

$$(S_{xx}^{1/2})^{-1} = S_{xx}^{-1/2} \qquad \text{and} \qquad (S_{yy}^{1/2})^{-1} = S_{yy}^{-1/2}$$

both exist and so we may write

$$a = S_{xx}^{-1/2}\tilde{a} \qquad \text{and} \qquad b = S_{yy}^{-1/2}\tilde{b},$$

and optimisation problem (4.6) subject to (4.7) becomes

$$\max_{\tilde{a},\tilde{b}} \tilde{a}^\top S_{xx}^{-1/2} S_{xy} S_{yy}^{-1/2} \tilde{b}$$

subject to

$$||\tilde{a}|| = 1 \qquad \text{and} \qquad ||\tilde{b}|| = 1.$$

From the properties of the SVD, and in particular Result 2.15 in Chapter 2, we know that the maximum correlation is ξ_1. Moreover, using the SVD again, this is achieved when $\tilde{a} = \mathbf{q}_1$ and $\tilde{b} = \mathbf{r}_1$ or, equivalently, $a = S_{xx}^{-1/2}\mathbf{q}_1$ and $b = S_{yy}^{-1/2}\mathbf{r}_1$.

□

Example 4.1 (continued) We now want to calculate the matrix A in (4.8) and then find its singular valued decomposition. We first need to find $S_{xx}^{-1/2}$ and $S_{yy}^{-1/2}$. Using R to do the calculations, we obtain the following:

$$\begin{aligned} S_{xx} &= Q_x \Lambda_x Q_x^\top \\ &= \begin{pmatrix} -0.970 & -0.241 \\ 0.241 & -0.970 \end{pmatrix} \begin{pmatrix} 41.46 & 0 \\ 0 & 6.04 \end{pmatrix} \begin{pmatrix} -0.970 & -0.241 \\ 0.241 & -0.970 \end{pmatrix}^\top, \end{aligned}$$

and so

$$\begin{aligned} S_{xx}^{-1/2} &= Q_x \Lambda_x^{-1/2} Q_x^\top \\ &= \begin{pmatrix} -0.970 & -0.241 \\ 0.241 & -0.970 \end{pmatrix} \begin{pmatrix} 41.46^{-1/2} & 0 \\ 0 & 6.04^{-1/2} \end{pmatrix} \begin{pmatrix} -0.970 & -0.241 \\ 0.241 & -0.970 \end{pmatrix}^\top \\ &= \begin{pmatrix} 0.170 & 0.059 \\ 0.059 & 0.392 \end{pmatrix}; \end{aligned}$$

and, omitting details of the calculations this time,

$$S_{yy}^{-1/2} = Q_y \Lambda_y^{-1/2} Q_y^\top = \begin{pmatrix} 0.064 & 0.030 \\ 0.030 & 0.086 \end{pmatrix}.$$

Consequently,

$$\begin{aligned} A &= S_{xx}^{-1/2} S_{xy} S_{yy}^{-1/2} \\ &= \begin{pmatrix} 0.170 & 0.059 \\ 0.059 & 0.392 \end{pmatrix} \begin{pmatrix} 115.7 & -81.9 \\ -29.4 & 6.0 \end{pmatrix} \begin{pmatrix} 0.064 & 0.030 \\ 0.030 & 0.086 \end{pmatrix} \\ &= \begin{pmatrix} 0.741 & -0.628 \\ -0.374 & -0.351 \end{pmatrix}. \end{aligned}$$

The SVD of A is given by

$$\begin{aligned} A &= Q\Xi R^\top \\ &= \begin{pmatrix} -0.997 & 0.082 \\ 0.082 & 0.997 \end{pmatrix} \begin{pmatrix} 0.974 & 0 \\ 0 & 0.508 \end{pmatrix} \begin{pmatrix} -0.790 & -0.613 \\ 0.613 & -0.790 \end{pmatrix}^\top . \end{aligned} \tag{4.9}$$

So the 1st CC coefficient is 0.974, which is close to its maximum value of 1. The 1st CC weight vectors are given by

$$\begin{aligned} a_1 &= S_{xx}^{-1/2} q_1 \\ &= \begin{pmatrix} 0.170 & 0.059 \\ 0.059 & 0.392 \end{pmatrix} \begin{pmatrix} -0.997 \\ 0.082 \end{pmatrix} \\ &= \begin{pmatrix} -0.165, \\ -0.027 \end{pmatrix}. \end{aligned}$$

Similar calculations show that

$$b_1 = S_{yy}^{-1/2} r_1 = \begin{pmatrix} -0.032 \\ 0.029 \end{pmatrix}.$$

In order to make interpretation easier:

- We change a_1 to $-a_1$ and b_1 to $-b_1$. [This entails changing q_1 to $-q_1$ and r_1 to $-r_1$; note that, provided we change the sign of **both** q_1 and r_1, we do not change the matrix A.]
- We rescale a_1 and b_1 so that they are unit vectors.

This leads to the standardised 1st CC weight vectors

$$a_1 = \begin{pmatrix} 0.987 \\ 0.160 \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} 0.743 \\ -0.670 \end{pmatrix}$$

and the 1st CC variables, obtained by using these weights, are

$$\eta_1 = 0.987 * (W - \bar{W}) + 0.160 * (D - \bar{D})$$

and

$$\psi_1 = 0.743 * (F - \bar{F}) - 0.670 * (A - \bar{A}),$$

where the bars are used to denote sample means.

We can see that ψ_1 is measuring something similar to goal difference $F - A$, as usually defined, but it gives slightly higher weight to goals scored than goals conceded (0.743 versus 0.670).

It is also seen that η_1 is measuring something similar to number of points $3 * W + D$, as usually defined, but the ratio of points for a win to points for a draw is somewhat higher, at around 6:1, as opposed to the usual ratio 3:1.

4.2 The full set of canonical correlations

Let us first recap what we did in the previous section: we found the choices linear combinations of the x-variables and linear combinations of y-variables which maximise the correlation, and expressed the answer in terms of quantities which arise in the SVD of A, where

$$A \equiv S_{xx}^{-1/2} S_{xy} S_{yy}^{-1/2} = Q \mathbf{\Xi} R^\top = \sum_{j=1}^{t} \xi_j q_j r_j^\top,$$

with t the rank of A, which in most examples is given by $t = \min(p, q)$, and singular values $\xi_1 \geq \xi_2 \geq \cdots \geq \xi_t > 0$. Specifically, the maximum value of the correlation is ξ_1, the optimal weights for the x-variables are given by $a = S_{xx}^{-1/2} q_1 = a_1$, say, and the optimal weights for the y-varables are given by $b = S_{yy}^{-1/2} r_1 = b_1$, say.

Can we repeat this process, as we did with PCA? Yes, we can. To obtain the second canonical correlation coefficient, plus the associated sets of weights, we need to solve the following optimisation problem:

$$\max_{a,\, b} a^\top S_{xy} b \tag{4.10}$$

subject to the constraints

$$a^\top S_{xx} a = 1, \qquad b^\top S_{yy} b = 1, \tag{4.11}$$

$$a_1^\top S_{xx} a = 0 \qquad \text{and} \qquad b_1^\top S_{yy} b = 0. \tag{4.12}$$

Note that maximising (4.10) subject to (4.11) is very similar to the optimisation problem (4.6) and (4.7) considered in the previous section. What is new are the constraints (4.12), which take into account that we have already found the first canonical correlation. If for $j = 1, 2$ we write $\tilde{a}_j = S_{xx}^{1/2} a_j$ and $\tilde{b}_j = S_{yy}^{1/2} b_j$, then it is seen from (4.12) that

$$\tilde{a}_1^\top \tilde{a}_2 = 0 \qquad \text{and} \qquad \tilde{b}_1^\top \tilde{b}_2 = 0.$$

Consequently, we may view constraints (4.12) as corresponding to orthogonality constraints (cf. PCA) in modified coordinate systems.

We now discuss the optimisation of (4.10), (4.11) and (4.12). At first glance it looks complex. However, using arguments very similar to those used to prove Result 2.15 in Chapter 2, we may deduce the following:

- The maximum of (4.10) subject to constraints (4.11) and (4.12) is equal to ξ_2, the second largest singular value of A.
- The optimal weights for the x-variables for the second canonical correlation are given by $a_2 = S_{xx}^{-1/2} q_2$.

- The optimal weights for the y-variables for the second canonical correlation are given by $b_2 = S_{yy}^{-1/2} r_2$.

Consider now the general case of the kth canonical correlation where $2 \leq k \leq t$. In this case we replace (4.11) and (4.12) by, respectively, (4.13) and (4.14) below, where

$$a^\top S_{xx} a = 1, \qquad b^\top S_{yy} b = 1, \tag{4.13}$$

$$a_j^\top S_{xx} a = 0 \qquad \text{and} \qquad b_j^\top S_{yy} b = 0, \qquad j = 1, \ldots, k-1. \tag{4.14}$$

Then the optimisation problem is

$$\max_{a,\, b} a^\top S_{xy} b \tag{4.15}$$

subject to constraints (4.13) and (4.14). The solution in the general case is as follows.

- The maximum of (4.15) subject to constraints (4.13) and (4.14) is equal to ξ_k, the kth largest singular value of A.
- The optimal weights for the x-variables for the kth canonical correlation are given by $a_k = S_{xx}^{-1/2} q_k$.
- The optimal weights for the y-variables for the kth canonical correlation are given by $b_k = S_{yy}^{-1/2} r_k$.

Terminology: we call a_k and b_k the kth cc (weight) vectors for the x-variables and y variables, respectively.

We call $\eta_{ik} = a_k^\top (x_i - \bar{x})$ and $\psi_{ik} = b_k^\top (y_i - \bar{y})$, $i = 1, \ldots, n$, the kth cc scores for the x-variables and the y-variables, respectively.

Define the CC score vectors $\boldsymbol{\eta}_k = (\eta_{1k}, \ldots, \eta_{nk})^\top$ and $\boldsymbol{\psi}_k = (\psi_{1k}, \ldots, \psi_{nk})^\top$. Then we have the following result.

Proposition 4.2. *Assume that S_{xx} and S_{yy} both have full rank. Then for $1 \leq k, \ell \leq t$,*

$$r[\eta_k, \psi_\ell] = \begin{cases} \xi_k & \text{if} \quad k = \ell \\ 0 & \text{if} \quad k \neq \ell, \end{cases}$$

where t is the rank of $A = S_{xx}^{-1/2} S_{xy} S_{yy}^{-1/2}$ and $\xi_1 \geq \xi_2 \geq \cdots \xi_t > 0$ are the strictly positive singular values of A.

Example 4.1 (continued) From (4.9), it is seen that the 2nd CC coefficient is given by $\xi_2 = 0.508$. So the correlation between the second pair of CC variables is a lot smaller than the 1st CC coefficient, though still appreciably different from 0. We now calculate the 2nd CC weight vectors:

$$a_2 = S_{xx}^{-1/2} q_2 = \begin{pmatrix} 0.073 \\ 0.396 \end{pmatrix} \qquad \text{and} \qquad b_2 = S_{yy}^{-1/2} r_2 = -\begin{pmatrix} 0.062 \\ 0.086 \end{pmatrix},$$

with standardised version (without the sign changes this time)

$$a_2 = \begin{pmatrix} 0.181 \\ 0.984 \end{pmatrix} \qquad \text{and} \qquad b_2 = -\begin{pmatrix} 0.589 \\ 0.808 \end{pmatrix},$$

and new variables

$$\eta_2 = 0.181 * (W - \bar{W}) + 0.984 * (D - \bar{D})$$

and

$$\psi_2 = -\{0.589 * (F - \bar{F}) + 0.808 * (A - \bar{A})\}.$$

Note that, to a good approximation, η_2 is measuring something similar to the number of draws and, approximately, ψ_2 is something related to the negative of total number of goals in a team's games. So large ψ_2 means relatively few goals in a team's games, and small (i.e. large negative) ψ_2 means a relatively large number of goals in a team's games.

Interpretation of the 2nd CC: teams that have a lot of draws tend to be in low-scoring games and/or teams that have few draws tend to be in high-scoring games.

4.3 Connection with linear regression when $q = 1$

Although CCA analysis is clearly a different technique to linear regression, it turns out that when either $p = 1$ or $q = 1$, there is a close connection between the two approaches.

Without loss of generality we assume that $q = 1$ and $p > 1$. Hence there is only a single y-variable but we still have $p > 1$ x-variables.

We also make the following assumptions:

1. The x_i have been centred so that $\bar{x} = \mathbf{0}_p$, the zero vector.
2. The covariance matrix for the x-variables, S_{xx}, has full rank p.

Both of these are weak assumptions in the multiple linear regression context.

Since $q = 1$,

$$A = S_{xx}^{-1/2} S_{xy} S_{yy}^{-1/2}$$

is a $p \times 1$ vector. Consequently, in this rather special case, the SVD tells us that

$$A = \xi_1 q_1,$$

where

$$\xi_1 = ||A|| \qquad \text{and} \qquad q_1 = A/||A|| = \tilde{a},$$

and $\tilde{a} = S_{xx}^{1/2} a$.

Consequently,

$$\begin{aligned} a &= S_{xx}^{-1/2} q_1 \\ &= S_{xx}^{-1/2} \frac{1}{||S_{xx}^{-1/2} S_{xy} S_{yy}^{-1/2}||} S_{xx}^{-1/2} S_{xy} S_{yy}^{-1/2} \\ &= \frac{1}{||S_{xx}^{-1/2} S_{xy}||} S_{xx}^{-1/2} S_{xx}^{-1/2} S_{xy} \\ &= \frac{1}{||S_{xx}^{-1/2} S_{xy}||} S_{xx}^{-1} S_{xy}. \end{aligned}$$

But since $\bar{x} = \mathbf{0}_p$ and S has full rank by the assumptions above, it follows that

$$nS_{xx} = \sum_{i=1}^{n} x_i x_i^\top = X^\top X$$

and

$$nS_{xy} = \sum_{i=1}^{n} y_i x_i = X^\top y,$$

where $y = (y_1, \ldots, y_n)^\top$ is the $n \times 1$ data matrix for the y-variable and $X = [x_1, \ldots, x_n]^\top$ is the data matrix for the x-variables. Consequently, the optimal a is a scalar multiple of

$$S_{xx}^{-1} S_{xy} = \left(X^\top X\right)^{-1} X^\top y = \hat{\boldsymbol{\beta}},$$

say, which is the classical expression for least squares estimator. Therefore the least squares estimator $\hat{\boldsymbol{\beta}}$ solves (4.4). However, it does not usually solve the optimisation problem defined by problems (4.6) and (4.7) because typically it will not be the case that $\hat{\boldsymbol{\beta}}^\top S_{xx} \hat{\boldsymbol{\beta}} = 1$, so that (4.7) will not be satisfied.

4.4 Population CCA

So far in this chapter we have based CCA on the sample covariance matrix

$$S_{zz} = \begin{bmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{bmatrix},$$

However, just as there is a population analogue of PCA, so there is a population analogue of CCA.

Given random vectors $\overset{p\times 1}{x}$ and $\overset{q\times 1}{y}$, define the random vector $z = (x^\top, y^\top)^\top$ with population covariance matrix

$$\mathrm{Var}(z) = \Sigma_{zz} = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}.$$

Then, by analogy with what we have seen in the sample CCA, the population CCA is based on the matrix

$$\check{A} = \Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1/2},$$

where, as in §3.4, the check symbol has been used above and below to indicate population quantities. If $\check{A}$ has SVD

$$\check{A} = \sum_{j=1}^{t} \check{\xi}_j \check{\mathbf{q}}_j \check{\mathbf{r}}_j^\top \equiv \check{\mathbf{Q}}\check{\mathbf{\Xi}}\check{\mathbf{R}}^\top,$$

where $\check{\xi}_1 \geq \cdots \geq \check{\xi}_t \geq 0$ and $t = \min(p, q)$, and the $\check{q}_j$ and $\check{\mathbf{r}}_j$ are unit vectors, then the first population CC coefficient is given by $\check{\xi}_1$, and the associated weights are given by

$$\check{a} = \Sigma_{xx}^{-1/2}\check{q}_1 = \check{a}_1 \qquad \text{and} \qquad \check{b} = \Sigma_{yy}^{-1/2}\check{\mathbf{r}}_1 = \check{b}_1.$$

The full set of population CC weight vectors is given by

$$\check{a}_j = \Sigma_{xx}^{-1/2}\check{q}_j \qquad \text{and} \qquad \check{b}_j = \Sigma_{yy}^{-1/2}\check{\mathbf{r}}_1, \qquad , j = 1, \ldots, t,$$

and the jth population CC coefficient is given by $\check{\xi}_j$.

4.5 Invariance/equivariance properties of CCA

Suppose we apply orthogonal transformations and translations to the x_i and the y_i of the form

$$\mathbf{h}_i = \mathbf{T}x_i + \boldsymbol{\mu} \qquad \text{and} \qquad \mathbf{k}_i = \mathbf{V}y_i + \boldsymbol{\eta}, \qquad i = 1, \ldots, n, \tag{4.16}$$

where $\mathbf{T}$ $(p \times p)$ and $\mathbf{V}$ $(q \times q)$ are orthogonal matrices, and $\boldsymbol{\mu}$ $(p \times 1)$ and $\boldsymbol{\eta}$ $(q \times 1)$ are fixed vectors.

How do these transformations affect the CC analysis?

First of all, since the CCA depends only on sample covariance matrices, it follows that the translation vectors $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$ have no effect on the analysis, so we can ignore $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$, and without loss of generality we shall set each to be the zero vector.

As seen in the previous section, the CCA in the original coordinates depends on

$$A \equiv A_{xy} = S_{xx}^{-1/2}S_{xy}S_{yy}^{-1/2}. \tag{4.17}$$

In the new coordinates we have

$$\tilde{S}_{\mathbf{h}h} = \mathbf{T}S_{xx}\mathbf{T}^\top, \qquad \tilde{S}_{\mathbf{k}k} = \mathbf{V}S_{yy}\mathbf{V}^\top,$$

$$\tilde{S}_{\mathbf{h}k} = \mathbf{T}S_{xy}\mathbf{V}^\top \qquad \text{and} \qquad \tilde{S}_{\mathbf{k}h} = \mathbf{V}S_{yx}\mathbf{T}^\top = S_{\mathbf{h}k}^\top,$$

where here and below, a tilde above a symbol is used to indicate that the corresponding term is defined in terms of the new h, k coordinates, rather than the old x, y coordinates. Moreover, due to the fact that $\mathbf{T}$ and $\mathbf{V}$ are orthogonal,

$$\tilde{S}_{\mathbf{h}h}^{1/2} = \mathbf{T}S_{xx}^{1/2}\mathbf{T}^\top, \qquad \tilde{S}_{\mathbf{h}h}^{-1/2} = \mathbf{T}S_{xx}^{-1/2}\mathbf{T}^\top$$
$$\tilde{S}_{\mathbf{k}k}^{1/2} = \mathbf{V}S_{yy}^{1/2}\mathbf{V}^\top \qquad \text{and} \qquad \tilde{S}_{\mathbf{k}k}^{-1/2} = \mathbf{V}S_{yy}^{-1/2}\mathbf{V}^\top.$$

The analogue of (4.17) in the new coordinates is given by

$$\begin{aligned}\tilde{A}_{\mathbf{h}k} &= \tilde{S}_{\mathbf{h}h}^{-1/2}\tilde{S}_{\mathbf{h}k}\tilde{S}_{\mathbf{k}k}^{-1/2} \\ &= \mathbf{T}S_{xx}^{-1/2}\mathbf{T}^\top\mathbf{T}S_{xy}\mathbf{V}^\top\mathbf{V}S_{yy}^{-1/2}\mathbf{V}^\top \\ &= \mathbf{T}S_{xx}^{-1/2}S_{xy}S_{yy}^{-1/2}\mathbf{V}^\top \\ &= \mathbf{T}A_{xy}\mathbf{V}^\top.\end{aligned}$$

So, again using the fact that $\mathbf{T}$ and $\mathbf{V}$ are orthogonal matrices, if A_{xy} has SVD $\sum_{j=1}^{t}\xi_j\mathbf{q}_j\mathbf{r}_j^\top$, then $\tilde{A}_{\mathbf{h}k}$ has SVD

$$\begin{aligned}\tilde{A}_{\mathbf{h}k} = \mathbf{T}A_{xy}\mathbf{V}^\top &= \mathbf{T}\left(\sum_{j=1}^{t}\xi_j\mathbf{q}_j\mathbf{r}_j^\top\right)\mathbf{V}^\top \\ &= \sum_{j=1}^{t}\xi_j\mathbf{T}\mathbf{q}_j\mathbf{r}_j^\top\mathbf{V}^\top = \sum_{j=1}^{t}\xi_j\left(\mathbf{T}\mathbf{q}_j\right)\left(\mathbf{V}\mathbf{r}_j\right)^\top = \sum_{j=1}^{t}\xi_j\tilde{q}_j\tilde{\mathbf{r}}_j^\top,\end{aligned}$$

where, for $j = 1, \ldots, t$, the $\tilde{q}_j = \mathbf{T}q_j$ are mutually orthogonal unit vectors, and the $\tilde{\mathbf{r}}_j = \mathbf{V}\mathbf{r}_j$ are also mutually orthogonal unit vectors.

Consequently, $\tilde{A}_{\mathbf{h}k}$ has the same singular values as A_{xy}, namely $\xi_1, \ldots, \xi_t$ in both cases, and so the canonical correlation coefficients are invariant with respect to the transformations (4.16). Moreover, since the optimal linear combinations for the jth CC in the original coordinates are given by $a_j = S_{xx}^{-1/2}\mathbf{q}_j$ and $b_j = S_{yy}^{-1/2}\mathbf{r}_j$, the optimal linear combinations in the new coordinates are given by

$$\begin{aligned}\tilde{a}_j &= S_{\mathbf{h}h}^{-1/2}\mathbf{T}\mathbf{q}_j \\ &= \mathbf{T}S_{xx}^{-1/2}\mathbf{T}^\top\mathbf{T}\mathbf{q}_j \\ &= \mathbf{T}S_{xx}^{-1/2}\mathbf{q}_j \\ &= \mathbf{T}a_j,\end{aligned}$$

and a similar argument shows that $\tilde{b}_j = \mathbf{V}b_j$. So under transformations (4.16), the optimal vectors a_j and b_j transform in an equivariant manner to $\tilde{a}_j$ and $\tilde{b}_j$, respectively, $j = 1, \ldots, t$.

If either of $\mathbf{T}$ or $\mathbf{V}$ in (4.16) is not an orthogonal matrix then the singular values are not invariant and the cc vectors do not transform in an equivariant manner.

4.6 Testing for zero canonical correlation coefficients

So far in Part II of this module we have not considered formal statistical inference (e.g. hypothesis testing, construction of confidence regions). Inference in various multivariate settings is considered in Part III. However, before moving on, we briefly explain how to perform tests for zero correlations in the CCA setting, under the assumption that the $z_i = (x_i^\top, y_i^\top)^\top$ are IID multivariate normal.

As previously, suppose that the x_i are $p \times 1$ vectors and the y_i are $q \times 1$ vectors and the sample size, i.e. the number of z_i vectors, is n. Let $\Sigma_{xy} = \mathrm{Cov}(x, y)$ denote the population cross-covariance matrix as before and consider the null hypothesis

$$H_0 : \Sigma_{xy} = \mathbf{0}_{p,q},$$

i.e. Σ_{xy} is the $p \times q$ matrix of zeros. Let H_A denote the general alternative

$$H_A : \Sigma_{xy} \quad * \textit{unrestricted} * .$$

Then the large-sample log-likelihood ratio test statistic for testing H_0 versus H_A is as follows:

$$W_0 = -\left\{n - \frac{1}{2}(p + q + 3)\right\} \sum_{j=1}^{\min(p,q)} \log(1 - \xi_j^2),$$

where $\xi_1 \geq \xi_2 \cdots \geq \xi_{\min(p,q)} \geq 0$ are the sample canonical correlations. Moreover, when n is large, W_0 is approximately χ^2_{pq} under H_0, and H_0 should be rejected when W_0 is sufficiently large.

We now consider a test concerning the rank of Σ_{xy}. For $0 \leq t < \min(p, q)$, consider the hypothesis:

$$H_t : \text{at most } t \text{ of the CC coefficients are non-zero.}$$

It turns out there is a similar statistic to W_0 above, for testing H_t against H_A, defined by

$$W_t = -\left\{n - \frac{1}{2}(p + q + 3)\right\} \sum_{j=t+1}^{\min(p,q)} \log(1 - \xi_j^2),$$

where, under H_t with n large, W_t is approximately $\chi^2_{(p-t)(q-t)}$. Also, we reject H_t when W_t is sufficiently large.

Example 4.1 (continued). Here $p = q = 2$, $n = 20$ and $\xi_1 = 0.974$ and $\xi_2 = 0.508$. So we should refer W_0 to χ^2_4 and refer W_1 to χ^2_1. Here, $W_0 = 53.92$ and $W_1 = 4.92$. So hypothesis H_0 is strongly rejected, with p-value $<< 0.001$. In contrast, H_1 is rejected at the 0.05 level but is not rejected at the 0.01 level. So there is only moderate evidence that the 2nd CC coefficient is non-zero.

Chapter 5

Multidimensional Scaling

In this chapter our starting point is somewhat different. Suppose we have a sample of n experimental units and we have a way to measure `distance' or`dissimilarity' between any pair of experimental units i and j, leading to a measure of distance or dissimilarity d_{ij}, $i, j = 1, \ldots, n$. The starting point for Multidimensional Scaling (MDS) is a distance matrix $D = (d_{ij} : i, j = 1, \ldots, n)$. A key goal in MDS is to determine coordinates of a set of points in a low-dimensional Euclidean space, e.g. $\mathbb{R}$ or $\mathbb{R}^2$, whose inter-point distances (or dissimilarities) are approximately equal to the d_{ij}. Using this approximate approach we are able to perform a statistical study of the original experimental units in a lower-dimensional space than the original one. We shall also see that there is a close connection between MDS and PCA.

5.1 Multidimensional Scaling

We call an $n \times n$ matrix $D = (d_{ij})_{i,j=1}^n$ a **distance matrix** or, equivalently, a **dissimilarity matrix**, if the following properties are satisfied:

1. For $i = 1, \ldots, n$, $d_{ii} = 0$.
2. Symmetry: $d_{ij} = d_{ji} \geq 0$ for all $i, j = 1, \ldots, n$.
3. Definiteness: $d_{ij} = 0$ implies $i = j$.

A comment on our terminology. We do not require distances necessarily to satisfy the triangle inequality

$$d_{ik} \leq d_{ij} + d_{jk}. \tag{5.1}$$

A distance function which always satisfies the triangle inequality is called a **metric distance** or just a **metric**, and a distance function which does not always satisfy the triangle inequality is called **non-metric** distance.

Suppose $x_1, \ldots, x_n$ are points in $\mathbb{R}^p$. If the d_{ij} are of the form

$$d_{ij} = ||x_i - x_j|| = \sqrt{(x_i - x_j)^\top (x_i - x_j)}.$$

Then each d_{ij} is called a **Euclidean distance** and, in this case, D is called a **Euclidean distance matrix**. Since Euclidean distances satisfy the triangle inequality (5.1), it follows that Euclidean distance is a metric distance.

Given a distance matrix $\mathbf{D} = \{d_{ij}\}_{i,j=1}^n$, define the matrix

$$A = \{a_{ij}\}_{i,j=1}^n, \quad \text{where} \qquad a_{ij} = -\frac{1}{2} d_{ij}^2. \tag{5.2}$$

Note that, for $i = 1, \ldots, n$, $a_{ii} = -d_{ii}^2/2 = 0$.

Now define the matrix

$$\mathbf{B} = \mathbf{H}A\mathbf{H}, \tag{5.3}$$

where

$$\mathbf{H} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top \tag{5.4}$$

is the $n \times n$ **centering matrix**; see §2.7. For reasons that will soon become clear, B defined by (5.3) is known as a centred inner-product matrix.

Let $x_1, \ldots, x_n$ denote n points in $\mathbb{R}^p$. Then the $n \times p$ matrix $\mathbf{X} = [x_1, \ldots, x_n]^\top$ is the data matrix, as before.

We now present the key result for classical MDS.

Proposition 5.1. *Let D denote an $n \times n$ distance matrix and suppose A, B and $\mathbf{H}$ be as defined in* (5.2), (5.3) *and* (5.4), *respectively.*

1. *The matrix D is a Euclidean distance matrix if and only if B is a non-negative definite matrix.*

2. *If D is a Euclidean distance matrix for the sample of n vectors $x_1, \ldots, x_n$, then*

$$b_{ij} = (x_i - \bar{x})^\top (x_j - \bar{x}), \qquad i, j = 1, \ldots, n, \tag{5.5}$$

where $\bar{\mathbf{x}} = n^{-1}\sum_{i=1}^n x_i$ is the sample mean vector. Equivalently, we may write

$$B = (\mathbf{HX})(\mathbf{HX})^\top,$$

where $\mathbf{X} = [x_1, \ldots, x_n]^\top$ is the data matrix, and H is the $n \times n$ centering matrix. Consequently, B is non-negative definite.

3. *Suppose B is non-negative definite with positive eigenvalues $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_k$ and spectral decomposition $B = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$, where $\mathbf{\Lambda} = diag\{\lambda_1 \ldots \lambda_k\}$ and $\mathbf{Q}$ is $n \times k$ and satisfies $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_k$. Then $\mathbf{X} = [x_1, \ldots, x_n]^\top = \mathbf{Q}\mathbf{\Lambda}^{1/2}$ is an $n \times k$ data matrix for points $x_1, \ldots, x_n$ in $\mathbb{R}^k$, which have inter-point distances given by $D = (d_{ij})$. Moreover, for this data matrix $\bar{x} = \mathbf{0}_k$ and B represents the inner product matrix with elements given by* (5.5).

Proof. Part 1. is a direct consequence of parts 2. and 3. Parts 2. and 3. are proved in the example sheets. □

Important Point: Proposition 5.1 may be useful even if **D** is not a Euclidean distance matrix, in which case B has some negative eigenvalues. What we can do is to replace B by its positive part. If B has spectral decomposition $\sum_{j=1}^{p} \lambda_j q_j q_j^\top$, then its positive definite part is defined by

$$B_{\text{pos}} = \sum_{j:\,\lambda_j > 0} \lambda_j q_j q_j^\top.$$

In other words, we sum over those j such that λ_j is positive. Then B_{pos} is non-negative definite and so we can use Theorem 5.1(iii) to determine a Euclidean configuration which has centred inner-product matrix B_{pos}. Then, provided the negative eigenvalues are small in absolute value relative to the positive eigenvalues, the inter-point distances of the new points in Euclidean space should provide a good approximation to the original inter-point distances (d_{ij}).

Example 5.1. Consider the five point in $\mathbb{R}^2$:

$$x_1 = (0,0)^\top, x_2 = (1,0)^\top, \quad x_3 = (0,1)^\top$$
$$x_4 = (-1,0)^\top \quad \text{and} \quad x_5 = (0,-1)^\top.$$

The resulting distance matrix is

$$D = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & \sqrt{2} & 2 & \sqrt{2} \\ 1 & \sqrt{2} & 0 & \sqrt{2} & 2 \\ 1 & 2 & \sqrt{2} & 0 & \sqrt{2} \\ 1 & \sqrt{2} & 2 & \sqrt{2} & 0 \end{bmatrix}.$$

Using (5.2) first to calculate A, and then using (5.3) to calculate B, we find that

$$A = -\begin{bmatrix} 0 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0 & 1 & 2 & 1 \\ 0.5 & 1 & 0 & 1 & 2 \\ 0.5 & 2 & 1 & 0 & 1 \\ 0.5 & 1 & 2 & 1 & 0 \end{bmatrix}$$

and

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}.$$

Further numerical calculations using R show that the eigenvalues of B are

$$\lambda_1 = \lambda_2 = 2 \qquad \text{and} \qquad \lambda_3 = \lambda_4 = \lambda_5 = 0.$$

Note that, as expected from Proposition 5.1, B is non-negative definite because it is a Euclidean distance matrix.

The following mutually orthogonal unit eigenvectors corresponding to the repeated eigenvalue 2 are produced by R:

$$q_1 = \begin{pmatrix} 0 \\ -0.439 \\ -0.554 \\ 0.439 \\ 0.554 \end{pmatrix} \qquad \text{and} \qquad q_2 = \begin{pmatrix} 0 \\ 0.554 \\ -0.439 \\ -0.554 \\ 0.439 \end{pmatrix}.$$

So the coordinates of five points in $\mathbb{R}^2$ which have the same inter-point distance matrix, D, as the original five points in $\mathbb{R}^2$, are given by the rows of the matrix

$$Q\Lambda^{1/2} = \sqrt{2}[q_1, q_2] = \begin{pmatrix} 0 & 0 \\ -0.621 & 0.784 \\ -0.784 & -0.621 \\ 0.621 & -0.784 \\ 0.784 & 0.621 \end{pmatrix}.$$

In the example sheets you asked to verify that there is an orthogonal transformation which maps the original five points onto the new five points.

5.2 Principal Coordinates

Starting with a distance matrix D, and using the matrix B, we now show how to calculate exact or approximate Euclidean coordinates for the n objects under study. We already know from Proposition 5.1 how to do this when the distance matrix D is Euclidean, but we will see now that this construction works more generally. Moreover, there is a very close connection with principal components analysis.

- **Step 1**: Given a distance matrix $\overset{n\times n}{D}$, calculate A according to (5.2).
- **Step 2**: Calculate $B = (b_{ij})_{i,j=1}^n$ in (5.3) using

  $$b_{ij} = a_{ij} - \bar{a}_{i+} - \bar{a}_{+j} + \bar{a}_{++}, \qquad i, j = 1, \ldots, n,$$

  where

  $$\bar{a}_{i+} = n^{-1}\sum_{j=1}^n a_{ij}, \quad \bar{a}_{+j} = n^{-1}\sum_{i=1}^n a_{ij} \quad \text{and} \quad \bar{a}_{++} = n^{-2}\sum_{i,j=1}^n a_{ij}.$$

- **Step 3**: Assume that the k largest eigenvalues of $B = (b_{ij})_{i,j=1}^n$, $\lambda_1 > \lambda_2 > \cdots > \lambda_k$ are all positive and have associated unit eigenvectors $v_1, \ldots, v_k$.

- **Step 4**: Define $V = [v_1, \dots, v_k]$ and

$$X \equiv [x_1, \dots, x_n]^\top = V\Lambda^{1/2} = [\sqrt{\lambda_1}v_1, \dots, \sqrt{\lambda_k}v_k].$$

Then $x_i \in \mathbb{R}^k$, $i = 1, \dots, n$, are the principal coordinates of the n points in k dimensions.

It turns out that there is a very close connection between principal coordinate and principal components.

Proposition 5.2. *Let X be an $n \times p$ data matrix with associated Euclidean distance matrix*

$$d_{ij}^2 = (x_i - x_j)^\top (x_i - x_j),$$

where $x_1^\top, \dots, x_n^\top$ are the rows of X. Then the centred PC scores based on the first k principal components are principal coordinates of the n points in k dimensions based on the distance matrix D.

5.3 Similarity measures

Recap: so far in this chapter we have considered distances matrices $D = (d_{ij})_{i,j=1}^n$ with distances d_{ij}. In this setting, the larger d_{ij} is, the more distant, or dissimilar, object i is from object j.

Recall that we have distinguished between metric distances ("metrics"), which satisfy the triangle inequality (5.1), and non-metric distances, or dissimilarities, which need not satisfy (5.1).

In this section, we now consider the analysis of measures of *similarity* as opposed to measures of dissimilarity.

A *similarity* matrix is defined to be an $n \times n$ matrix $=(f_{ij})_{i,j=1}^n$ with the following properties:

1. Symmetry, i.e. $f_{ij} = f_{ji}$, $i, j = 1, \dots, n$.
2. For all $i, j = 1, \dots, n$, $f_{ij} \leq f_{ii}$.

Note that when working with similarities f_{ij}, the larger f_{ij} is, the more similar objects i and j are.

Condition 1. implies that object i is as similar to object j as object j is to object i (symmetry).

Condition 2. implies that an object is at least as similar to itself as it is to any other object.

One important class of problems is when the similarity between any two objects is measured by the number of common attributes. We illustrate this through two examples.

Example 5.2. Suppose there are 4 attributes we wish to consider.

1. Attribute 1: Carnivore? If yes, put $a_1 = 1$; if no, put $a_1 = 0$.
2. Attribute 2: Mammal? If yes, put $a_2 = 1$; if no, put $a_2 = 0$.
3. Attribute 3: Natural habitat in Africa? If yes, put $a_3 = 1$; if no, put $a_3 = 0$.
4. Attribute 4: Can climb trees? If yes, put $a_4 = 1$; if no, put $a_4 = 0$.

Consider a lion. Each of the attributes is present so $a_1 = a_2 = a_3 = a_4 = 1$.

A tiger? In this case, 3 of the attributes are present (1, 2 and 4) but 3 is absent. So for a tiger, $a_1 = a_2 = a_4 = 1$ and $a_3 = 0$.

How might we measure the similarity of lions and tigers based on the presence or absence of these four attributes?

First form a 2×2 table as follows.

		1	0
	1	a	b
	0	c	d

Here a counts the number of attributes common to both lion and tiger; b counts the number of attributes the lion has but the tiger does not have; c counts the number of attributes the tigher has that the lion does not have; and d counts the number of attributes which neither the lion nor the tiger has.

In the above, $a = 3$, $b = 1$ and $c = d = 0$.

How might we make use of the information in the 2×2 table to construct a measure of similarity?

The simplest measure of similarity is the proportion of the attributes which are shared.

$$\frac{a}{a+b+c+d},$$

which gives 0.75 in this example. A second similarity measure, which gives the same value in this example but not in general, is known as the *similarity matching coefficient* and is given by

$$\frac{a+d}{a+b+c+d}. \tag{5.6}$$

There are many other possibilities, e.g. we could consider weighted versions of the above if we wish to weight different attributes differently.

Example 5.3. Let us now consider a similar but more complex example with 6 unspecified attributes (not the same attributes as in Example 1) and 5 types of living creature, with the following data matrix, consisting of zeros and ones.

	1	2	3	4	5	6
Lion	1	1	0	0	1	1
Giraffe	1	1	1	0	0	1
Cow	1	0	0	1	0	1
Sheep	1	0	0	1	0	1
Human	0	0	0	0	1	0

Suppose we decide to use the similarity matching coefficient (5.6) to measure similarity. Then the following similarity matrix is obtained.

$$F = \begin{array}{c|ccccc} & \text{Lion} & \text{Giraffe} & \text{Cow} & \text{Sheep} & \text{Human} \\ \hline \textit{Lion} & 1 & 2/3 & 1/2 & 1/2 & 1/2 \\ \textit{Giraffe} & 2/3 & 1 & 1/2 & 1/2 & 1/6 \\ \textit{Cow} & 1/2 & 1/2 & 1 & 1 & 1/3 \\ \textit{Sheep} & 1/2 & 1/2 & 1 & 1 & 1/3 \\ \textit{Human} & 1/2 & 1/6 & 1/3 & 1/3 & 1 \end{array}$$

It is easily checked from the definition that $=(f_{ij})_{i,j=1}^{5}$ is a similarity matrix.

We now return to the general case. What should we do once we have calculated a similarity matrix? It turns out there is a nice transformation from a similarity matrix to a distance matrix $D = (d_{ij})_{i,j=1}^{n}$ defined by

$$d_{ij} = \left(f_{ii} + f_{jj} - 2f_{ij}\right)^{1/2}, \qquad i, j = 1, \ldots, n. \tag{5.7}$$

Note that, provided F is a similarity matrix, the d_{ij} are well-defined (i.e. real, not imaginary) because $f_{ii} + f_{jj} - 2f_{ij} \geq 0$ by condition 2., so the bracket is non-negative.

We have the following result.

Proposition 5.3. *Suppose that F is a similarity matrix. If, in addition, F is non-negative definite, then D defined in* (5.7) *is Euclidean with centred inner product matrix*

$$B = HFH, \tag{5.8}$$

where $H = I_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top$ is the centering matrix.

Proof. Since F is non-negative definite by assumption, and $H^\top = H$ by definition of H, it follows that HFH must also be non-negative definite. So by Result 5.1, we just need to show that (5.8) holds, where B is given by $B = HAH$ and A is defined as in (5.2), and the d_{ij} are defined by (5.7). Then

$$a_{ij} = -\frac{1}{2}d_{ij}^2 = f_{ij} - \frac{1}{2}(f_{ii} + f_{jj}).$$

Define

$$t = n^{-1}\sum_{i=1}^{n} f_{ii}.$$

Then, summing over $j = 1, \ldots, n$ for fixed i,

$$\bar{a}_{i+} = n^{-1}\sum_{j=1}^{n} a_{ij} = \bar{f}_{i+} - \frac{1}{2}(f_{ii} + t);$$

similarly,

$$\bar{a}_{+j} = n^{-1} \sum_{i=1}^{n} a_{ij} = \bar{f}_{+j} - \frac{1}{2}(f_{jj} + t),$$

and also

$$\bar{a}_{++} = n^{-2} \sum_{i,j=1}^{n} a_{ij} = \bar{f}_{++} - \frac{1}{2}(t + t).$$

So, using part (vii) of section 7 of Chapter 2 (FIX FIX),

$$\begin{aligned} b_{ij} &= a_{ij} - \bar{a}_{i+} - \bar{a}_{+j} + \bar{a}_{++} \\ &= f_{ij} - \frac{1}{2}(f_{ii} + f_{jj}) - \bar{f}_{i+} + \frac{1}{2}(f_{ii} + t) \\ &\qquad - \bar{f}_{+j} + \frac{1}{2}(f_{jj} + t) + \bar{f}_{++} - t \\ &= \qquad f_{ij} - \bar{f}_{i+} - \bar{f}_{+j} + \bar{f}_{++}. \end{aligned}$$

Consequently, $B = HFH$, using part (vii) of §2.7 again, and the result is proved. □ □

Part III: Inference using the MVN FIX FIGS TABS

Part III of these lecture notes covers statistical inference based on the multivariate normal (MVN) distribution, which is of relevance when there are several measurements per experimental unit and observations consist of random vectors.

Chapter @ref(#multinormal) focuses on classical distribution theory relating to the MVN distribution, including the Wishart distribution, which is defined on the set of symmetric positive definite matrices and is a natural generalisation of the χ^2 distribution. Another important distribution related to the MVN distribution is the Hotelling T^2 distribution, which is a multivariate analogue of the Student t-distribution.

Chapter 7 is concerned with testing hypotheses concerning vector means in 1-sample and 2-sample settings. There is a close connection with the classical 1-sample and 2-sample t-tests in univariate statistics, but here we are dealing with random vectors rather than random variables.

Chapter 8 is concerned with the multivariate linear model, in which the responses consist of random vectors rather than single random variables. Errors in this setting take the form of random vectors.

The results in Part III turn out to be natural but non-trivial generalisations of the results in the univariate case.

Chapter 6

Multivariate Normal Distribution Theory

The multivariate normal distribution (MVN) is important for a number of reasons:

1. It is a straightforward generalisation of the univariate normal distribution.
2. It is entirely defined by its mean vector μ and its covariance matrix Σ.
3. Zero correlation implies independence.
4. Linear functions of multivariate normal vectors are also multivariate normal vectors.
5. There is a multivariate version of the Central Limit Theorem.
6. It has simple geometric properties.

6.1 Definition and Properties of the MVN

Definition 6.1. A random vector $x = (x_1, \ldots, x_p)^\top$ has a p-dimensional MVN distribution if and only if $a^\top x$ is univariate normal for all fixed $p \times 1$ vectors a.

Proposition 6.1. *If* $\overset{p\times 1}{x}$ *is MVN then for each constant matrix* A $(q \times p)$ *and constant vector* c $(q \times 1)$, $y = Ax + c$ *has a* q*-dimensional MVN.*

Proof. Let b $(q \times 1)$ be a fixed vector. Then

$$b^\top y = b^\top Ax + b^\top c = a^\top x + b^\top c$$

where $a^\top = b^\top A$. Now $a^\top x$ is univariate normal for all a since x is MVN. Therefore $b^\top y$ is univariate normal for all b, so y is MVN. □

Corollary 6.1. *Any subset of the components of a MVN vector* x *is also MVN.*

Notation If x $(p \times 1)$ is MVN with mean μ and covariance matrix Σ then we can write

$$x \sim N_p(\mu, \Sigma).$$

This notation is a direct extension of the univariate notation where $p = 1$.

If the population covariance matrix Σ $(p \times p)$ is positive definite, so that Σ^{-1} exists, then the **probability density function** (pdf) of the MVN distribution is given by

$$f(x) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right).$$

If $p = 1$, so that $x = x$, $\mu = \mu$ and $\Sigma = \sigma^2$, say, then the pdf simplifies to

$$\begin{aligned} f(x) &= \frac{1}{|2\pi\sigma^2|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)\right) \\ &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \end{aligned}$$

which is the familiar pdf of the univariate normal distribution $N(\mu, \sigma^2)$.

If $p > 1$ and $\Sigma = I_p$ then

$$\begin{aligned} f(x) &= \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top(x-\mu)\right) \\ &= \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{p}(x_i-\mu_i)^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_1-\mu_1)^2\right)\right) \\ &\qquad \times \ldots \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_p-\mu_p)^2\right)\right) \end{aligned}$$

This means that, by the factorisation theorem for probability densities, the components of x have independent univariate normal distributions.

If $p = 2$ we can plot $f(x)$ using contour plots, where each contour shows values of x for which $f(x) = c$ for some constant, $c > 0$. Four examples are shown below for $c = 0.02, 0.04$.

FIX

6.2 Transformations

Proposition 6.2. *If $x \sim N_p(\mu, \Sigma)$ and $y = Ax + c$, where A $(q \times p)$ and c $(q \times 1)$ are constant, then*

$$y \sim N_q(A\mu + c, A\Sigma A^\top).$$

Proof. We know y is MVN by Proposition 6.1. We also know $E(y)$ and $\mathrm{Var}(y)$ from results (1) and (3) of §1.4. □

The above result implies that a linear transformation of a MVN random variable is also MVN. We can use this result to prove two important corollaries. The first corollary is useful for simulating data from a general MVN distribution.

Corollary 6.2. *If $x \sim N_p(0, I_p)$ and $y = \Sigma^{1/2}x + \mu$ then $y \sim N_p(\mu, \Sigma)$.*

Proof. We apply 6.2 with $A = \Sigma^{1/2}$ and $c = \mu$. Therefore $E(y) = \Sigma^{1/2}0_p + \mu = \mu$ and $\mathrm{Var}(y) = \Sigma^{1/2}I_p\Sigma^{1/2} = \Sigma$. □

The second corollary says that any MVN random variable can be transformed into standard form.

Corollary 6.3. *If $x \sim N_p(\mu, \Sigma)$, Σ has full rank and we define $y = \Sigma^{-1/2}(x-\mu)$ then $y \sim N_p(0, I_p)$.*

Proof. Apply Proposition 6.2 with $A = \Sigma^{-1/2}$ and $c = -\Sigma^{-1/2}\mu$. Then $E(y) = \Sigma^{-1/2}\mu - \Sigma^{-1/2}\mu = 0_p$ and $\mathrm{Var}(y) = \Sigma^{-1/2}\Sigma\Sigma^{-1/2} = I_p$. □

The moment generating function of a random vector x $(p \times 1)$ is given by

$$M(t) = E[e^{t^\top x}],$$

and is defined for all $t \in \mathbb{R}^p$ for which $M(t)$ is finite.

Proposition 6.3. *The moment generating function of $x \sim N_p(\mu, \Sigma)$ is given by*

$$M(t) = \exp\left(\mu^\top t + \frac{1}{2}t^\top \Sigma t\right). \tag{6.1}$$

Proof. For fixed t, define the random variable $Y = x^\top t$. From Proposition 6.2, $Y \sim N(\mu_t, {\sigma_t}^2)$, where $\mu_t = \mu^\top t$ and $\sigma_t^2 = t^\top \Sigma t$.

If $\sigma_t \equiv t^\top \Sigma t = 0$ then $Y = \mu^\top t$ with probability one, and then $M(t) = e^{\mu^\top t}$ which agrees with (6.1). So from now on we assume $\sigma_t > 0$. Then

$$\begin{aligned} M(t) &= E[e^{x^\top t}] \\ &= E[e^Y] = \int_{-\infty}^{\infty} \exp(y) \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{1}{2}\frac{(y-\mu_t)^2}{\sigma_t^2}\right) dy. \end{aligned}$$

The integral above can be evaluated by completing the square in the exponent, using the identity

$$y - \frac{1}{2}\frac{(y-\mu_t)^2}{\sigma_t^2} = \mu_t + \frac{1}{2}\sigma_t^2 - \frac{1}{2}\frac{(y-\mu_t-\sigma_t^2)^2}{\sigma_t^2}.$$

Consequently

$$\begin{aligned} M(t) &= \int_{-\infty}^{\infty} \exp\left\{\mu_t + \frac{1}{2}\sigma_t^2\right\} \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left\{-\frac{1}{2}\frac{(y-\mu_t-\sigma_t^2)^2}{\sigma_t^2}\right\} dy \\ &= \exp\left(\mu_t + \frac{1}{2}\sigma_t^2\right) \\ &= \exp\left(\mu^\top t + \frac{1}{2}t^\top \Sigma t\right), \end{aligned}$$

as required. □

Proposition 6.4. *Two vectors x $(p \times 1)$ and y $(q \times 1)$ which are jointly multivariate normal are independent if and only if they are uncorrelated (i.e. $Cov(x, y) = 0_{p,q}$).*

Proof. We prove this result using the factorisation theorem for moment generating functions (MGFs), which is now stated. Let $t = (t_1^\top, t_2^\top)^\top$ where $t_1 \in \mathbb{R}^p$, $t_2 \in \mathbb{R}^q$ and $t \in \mathbb{R}^{p+q}$. The joint MGF of two arbitrary random vectors $\underset{p\times 1}{x}$ and $\underset{q\times 1}{y}$ is defined by

$$M(t_1, t_2) = E[e^{t_1^\top x + t_2^\top y}],$$

for all $t = (t_1^\top, t_2^\top)^\top$ at which $M(t_1, t_2)$ is finite. The factorisation theorem for MGFs states that x and y are independent if an only if $M(t_1, t_2)$ factorises, i.e.

$$M(t_1, t_2) = M_1(t_1)M_2(t_2)$$

for some functions M_1 and M_2, in which case M_1 and M_2 are the marginal MGFs of x and y. Now we focus on the MVN case. Suppose

$$E[x] = \mu_x, \qquad E[y] = \mu_y, \quad \mathrm{Var}(x) = \Sigma_{xx}, \quad \mathrm{Var}(y) = \Sigma_{yy}, \tag{6.2}$$

and

$$\mathrm{Cov}(x, y) = \Sigma_{xy} = \Sigma_{yx}^\top = \mathrm{Cov}(y, x)^\top. \tag{6.3}$$

Using Proposition 6.3 and definitions (6.2) and (6.3),

$$\begin{aligned} M(t_1, t_2) &= \exp\left(\mu^\top t + \frac{1}{2}t^\top \Sigma t\right) \\ &= \exp\left(\mu_x^\top t_1 + \mu_y^\top t_2 + \frac{1}{2}t_1^\top \Sigma_{xx} t_1 \right. \\ &\qquad \left. + \frac{1}{2}t_2^\top \Sigma_{yy} t_2 + \frac{1}{2}2t_1^\top \Sigma_{xy} t_2\right) \\ &= M_1(t_1)M_2(t_2)M_3(t_1, t_2), \end{aligned}$$

where $M_1(t_1)$ and $M_2(t_2)$ are the marginal MGFs of x and y respectively, and

$$M_3(t_1, t_2) = \exp\left(t_1^\top \Sigma_{xy} t_2\right).$$

The factorisation theorem holds if an only if $M_3(t_1, t_2)$ is constant with respect to t_1 and t_2, which is the case if and only if $\Sigma_{xy} = \mathbf{0}_{p,q}$. □

In words: Proposition 6.4 means that zero correlation implies independence for the MVN distribution. This is not generally true for other distributions.

Note that Propositions 6.1 - 6.4 each holds regardless of whether the covariance matrix Σ is positive definite or not.

The term $(x-\mu)^\top \Sigma^{-1}(x-\mu)$ appears in the exponent of the pdf and we derive its distribution in Proposition 6.5.

Proposition 6.5. *If $x \sim N_p(\mu, \Sigma)$ and Σ is positive definite then*

$$(x-\mu)^\top \Sigma^{-1}(x-\mu) \sim \chi^2_p.$$

Proof. Define $y = \Sigma^{-1/2}(x-\mu)$ so

$$\begin{aligned}(x-\mu)^\top \Sigma^{-1}(x-\mu) &= \left(\Sigma^{-1/2}(x-\mu)\right)^\top \left(\Sigma^{-1/2}(x-\mu)\right) \\ &= y^\top y = \sum_{i=1}^{p} y_i^2\end{aligned}$$

By Corollary 6.3, $y \sim N_p(0, I_p)$, and so the components of y have independent univariate normal distributions with mean 0 and variance 1. Recall from univariate statistics that if $z \sim N(0,1)$ then $z^2 \sim \chi^2_1$ and if $z_1, \ldots, z_n$ are iid $N(0,1)$ then $\sum_{i=1}^n z_i^2 \sim \chi^2_n$. It therefore follows that $\sum_{i=1}^p y_i^2 \sim \chi^2_p$. □

We saw earlier in this chapter that the MVN distribution in p dimensions has constant density on ellipses or ellipsoids given by $f(x) = c$ for some constant $c > 0$. We can rearrange this equation to be of the form

$$U(x) = (x-\mu)^\top \Sigma^{-1}(x-\mu) = k$$

where $k = -2\log(c) - \log|2\pi\Sigma| > 0$ is a combination of the constant, c, and the normalising constant in the pdf. Proposition 6.5 means we can calculate the probability, $P(U(x) < k)$, which is the probability of x lying within a particular ellipsoid.

6.3 Two important results for the MVN

In this section we present two important results which are natural generalisations of what happens in the univariate case.

Proposition 6.6. *If $x_1, \ldots, x_n$ is an IID random sample from $N_p(\mu, \Sigma)$, then the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and the sample variance matrix $S = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^\top$ are independent.*

Proof. Define $x = \bar{x}$ and $y_i = x_i - \bar{x}$, $i = 1, \ldots, n$. From Proposition 6.1 and Proposition 6.2 we can see that if $x_1, \ldots, x_n$ is a random sample from $N_p(\mu, \Sigma)$ then $\bar{x} \sim N_p(\mu, n^{-1}\Sigma)$. Then

$$\begin{aligned}
\mathrm{Cov}(\bar{x}, y_i) &= \mathrm{Cov}(\bar{x}, x_i - \bar{x}) \\
&= \mathrm{Cov}(\bar{x}, x_i) - \mathrm{Cov}(\bar{x}, \bar{x}) \\
&= n^{-1} \sum_{j=1}^{n} \left\{ E[(x_j - \mu)(x_i - \mu)^\top] \right\} \\
&\qquad - E[(\bar{x} - \mu)(\bar{x} - \mu)^\top] \\
&= n^{-1}\Sigma - n^{-1}\Sigma = \mathbf{0}_{p,p}.
\end{aligned}$$

Now define the $(np) \times 1$ vector $y = (y_1^\top, \ldots, y_n^\top)^\top$. Then $\mathrm{Cov}(x, y) = \mathbf{0}_{p,np}$, so we may apply Proposition 6.4 to conclude that $\bar{x}$ and y are independent. Therefore $\bar{x}$ and

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

are independent, becasue S is a function of y alone.

□

Recall from above that if $x_1, \ldots, x_n$ is a random sample from $N_p(\mu, \Sigma)$ then $\bar{x} \sim N_p(\mu, \frac{1}{n}\Sigma)$. This result is also approximately true for a large random sample from a non-normal population, as is now stated in the multivariate central limit theorem.

Proposition 6.7. ***Central limit theorem*** *Let $x_1, x_2, \ldots$ be a sample of independent and identically distributed random vectors from a distribution with mean vector μ and finite variance matrix Σ. Then asymptotically as $n \to \infty$, $\sqrt{n}(\bar{x} - \mu)$ converges in distribution to $N_p(\mathbf{0}_p, \Sigma)$.*

Proof. Beyond the scope of this module. □

6.4 The Wishart distribution

The Wishart distribution is a multivariate generalisation of the univariate χ^2 distribution. In univariate statistics the χ^2 distribution plays an important role in inference related to the univariate normal, e.g. in the definition of Student's t distribution. An analogous role is played by the Wishart distribution in multivariate statistics.

Definition 6.2. Let $x_1, \ldots, x_n$ be an IID random sample from $N_p(0, \Sigma)$. Then $M = \sum_{i=1}^{n} x_i x_i^\top$ is said to have a Wishart distribution with n degrees of freedom and scale matrix Σ. We write this as $M \sim W_p(\Sigma, n)$.

Note that $W_p(\Sigma, n)$ is a probability distribution on the set of $p \times p$ symmetric non-negative definite random matrices.

We call $W_p(I_p, n)$ a standard Wishart distribution.

When $p = 1$, $W_1(1, n)$ is the χ^2_n distribution and $W_1(\sigma^2, n)$ is the $\sigma^2\chi^2_n$ distribution. This claim follows from 6.9 below.

We now use the definition of $W_p(\Sigma, n)$ to prove some important\ results.

Proposition 6.8. *If $M \sim W_p(\Sigma, n)$ and A is a fixed $q \times p$ matrix, then*

$$AMA^\top \sim W_q\left(A\Sigma A^\top, n\right).$$

Proof. From the definition, let $M = \sum_{i=1}^n x_i x_i^\top$, where the x_i are IID $N_p(0, \Sigma)$. Then

$$\begin{aligned} AMA^\top &= A\left(\sum_{i=1}^n x_i x_i^\top\right) A^\top \\ &= \sum_{i=1}^n (Ax_i)(Ax_i)^\top = \sum_{i=1}^n y_i y_i^\top \end{aligned}$$

where $y_i = Ax_i \sim N_q(0, A\Sigma A^\top)$, by Proposition 6.2. Now we apply the definition of the Wishart distribution to $y_1, \ldots, y_n$ and, hence, $\sum_{i=1}^n y_i y_i^\top \sim W_q\left(A\Sigma A^\top, n\right)$. □

Proposition 6.9. *If $M \sim W_p(\Sigma, n)$ and a is a fixed $p \times 1$ vector then*

$$a^\top Ma \sim (a^\top \Sigma a)\, \chi^2_n.$$

Proof. Apply Proposition 6.8 with $A = a^\top$ then $a^\top Ma \sim W_1(a^\top \Sigma a, n)$. But $W_1(1, n)$ is equal in distribution to $\sum_{i=1}^n z_i^2$ where the z_i are IID $N(0, 1)$, and so has χ^2_n distribution. Moreover, using Proposition 6.8 with $p = q = 1$ and $A = \sigma$, it is seen that $W_1(\sigma^2, n)$ is equal in distribution to $\sigma^2\chi^2_n$. □

Note that an alternative form of the above result is

$$\frac{a^\top Ma}{a^\top \Sigma a} \sim \chi^2_n.$$

Corollary 6.4. *Let m_{ii} and σ_{ii} be the ith diagonal entry for M and Σ respectively, then $m_{ii} \sim \sigma_{ii}\chi^2_n$ for $i = 1, \ldots, p$.*

Proof. Let $a = (a_1, \ldots, a_p)^\top$ where $a_j = 1$ if $j = i$ and $a_j = 0$ otherwise. Then $a^\top Ma = m_{ii}$ and $a^\top \Sigma a = \sigma_{ii}$. Now apply Proposition 6.9

□

Note, however, that the m_{ii}, $i = 1, \ldots, p$, are not, in general, independent.

Proposition 6.10. *If $M_1 \sim W_p(\Sigma, n_1)$ and $M_2 \sim W_p(\Sigma, n_2)$ are independent then*

$$M_1 + M_2 \sim W_p(\Sigma, n_1 + n_2).$$

Proof. From the definition, let $M_1 = \sum_{i=1}^{n_1} x_i x_i^\top$ and let $M_2 = \sum_{i=n_1+1}^{n_1+n_2} x_i x_i^\top$, where $x_i \sim N_p(0, \Sigma)$, then $M_1 + M_2 = \sum_{i=1}^{n_1+n_2} x_i x_i^\top \sim W_p(\Sigma, n_1 + n_2)$ by the definition of the Wishart distribution. □

Our next result is known as Cochran's theorem. Recall the definition of projection matrices at the end of §2.4.

Theorem 6.1. ***(Cochran's Theorem)*** *Suppose $\overset{n\times n}{\mathbf{P}}$ is a projection matrix of rank r. Assume that X is an $n \times p$ data matrix with IID rows that have a common $N_p(\mathbf{0}_p, \Sigma)$ distribution, where Σ has full rank p, and note the identity*

$$X^\top X = X^\top \mathbf{P} X + X^\top (\mathbf{I}_n - \mathbf{P}) X. \tag{6.4}$$

Then

$$X^\top \mathbf{P} X \sim W_p(\Sigma, r), \qquad X^\top (\mathbf{I}_n - \mathbf{P}) X \sim W_p(\Sigma, n - r), \tag{6.5}$$

and $X^\top \mathbf{P} X$ and $X^\top (\mathbf{I}_n - \mathbf{P}) X$ are independent.

Proof. We first of all prove the result in the particular case $\Sigma = \mathbf{I}_p$ and then consider the general case. Using the Spectral Decomposition Theorem 2.5 and noting Proposition 2.1, we may write

$$\mathbf{P} = \sum_{j=1}^{r} q_j q_j^\top \qquad \text{and} \qquad (I_n - \mathbf{P}) = \sum_{j=r+1}^{n} q_j q_j^\top$$

where $q_1, \ldots, q_n$ are mutually orthogonal unit vectors. Then

$$\begin{aligned} X^\top P X &= X^\top \left(\sum_{j=1}^{r} q_j q_j^\top \right) X \\ &= \sum_{j=1}^{r} X^\top q_j q_j^\top X = \sum_{j=1}^{r} y_j y_j^\top, \end{aligned} \tag{6.6}$$

and similarly,

$$\begin{aligned} X^\top (I_n - P) X &= X^\top \left(\sum_{j=r+1}^{n} q_j q_j^\top \right) X \\ &= \sum_{j=r+1}^{n} X^\top q_j q_j^\top X = \sum_{j=r+1}^{n} y_j y_j^\top, \end{aligned} \tag{6.7}$$

where, for $j = 1, \ldots, n$, $y_j = X^\top q_j$ is a $p \times 1$ vector. We shall now prove that the y_j are IID $N_p(\mathbf{0}_p, I_p)$. Write $X = [x_{[1]}, \ldots, x_{[p]}]$, where $x_{[u]}$ is column u of X. Then $x_{[1]}, \ldots, x_{[p]}$ are IID $N_n(\mathbf{0}_n, I_n)$. Moreover,

$$y_j = X^\top q_j = \begin{bmatrix} q_j^\top x_{[1]} \\ q_j^\top x_{[2]} \\ .. \\ .. \\ .. \\ q_j^\top x_{[p]} \end{bmatrix}.$$

But

$$\begin{aligned} E[q_j^\top x_{[u]} q_k^\top x_{[v]}] &= E[q_j^\top x_{[u]} x_{[v]}^\top q_k] \\ &= q_j^\top E[x_{[u]} x_{[v]}^\top] q_k \\ &= q_j^\top (\delta_{uv} I_n) q_k \\ &= q_j^\top q_k \delta_{uv} \\ &= \delta_{jk} \delta_{uv}, \end{aligned} \tag{6.8}$$

where δ is the Kronecker δ defined by

$$\delta_{ab} = \begin{cases} 0 & \text{if} \quad a \neq b \\ 1 & \text{if} \quad a = b \end{cases}.$$

If follows immediately from (6.8) that

$$\mathrm{Var}(y_j) = I_p \qquad \mathrm{Cov}(y_j, y_k) = \mathbf{0}_{pp} \quad \text{if} \quad j \neq k.$$

By Proposition 6.4, the y_j, $j = 1, \ldots, n$, are IID $N_p(\mathbf{0}_p, I_p)$, and therefore it follows from the definition of the Wishart distribution that, when $\Sigma = I_p$, (6.6) has a Wishart $W_p(I_p, r)$ distribiton, (6.7) has a Wishart $W_p(I_p, n-r)$ distrubtion. Moreover, these random Wishart matrices are independent becasue the y_j are all independent.

Finally, we consider the case of a general covariance matrix Σ. We have proved that (6.4) holds when $\Sigma = I_p$, so pre-multiply both sides by the matrix square root $\Sigma^{1/2}$, and post-multiply both sides by $\Sigma^{1/2}$. This corresponds to the case where the x_i are IID $N_p(\mathbf{0}_p, \Sigma)$. Then, using Proposition 6.8,

$$\Sigma^{1/2} W_p(I_p, t) \Sigma^{1/2} \stackrel{d}{=} W_p(\Sigma^{1/2} \Sigma^{1/2}, t) \stackrel{d}{=} W_p(\Sigma, t),$$

when $t = r$ and $t = n - r$. Moreover, since $\Sigma^{1/2}$ is a non-random matrix, independence is preserved when we pre- and post-multiply by $\Sigma^{1/2}$, and the result is proved. □

Proposition 6.11. *If $x_1, \ldots, x_n$ is an IID sample from $N_p(\mu, \Sigma)$, then*

$$nS = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^\top \sim W_p(\Sigma, n-1).$$

Proof. Define $P = \mathbf{H} \equiv I_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top$ where $\mathbf{1}_n$ is the $n \times 1$ vector of ones. Note that H is the $n \times n$ centering matrix and, from Property (i) of §2.7, H is a projection matrix. Clearly, $I_n - P = n^{-1}\mathbf{1}_n\mathbf{1}_n^\top$ has rank 1, so H has rank $n-1$. Therefore, using Theorem 6.1 ,

$$X^\top H X \sim W_p(\Sigma, n-1).$$

But from Property (vi) in §2.7, $X^\top H X = nS$, and consequently, $nS \sim W_p(\Sigma, n-1)$, as required. □

6.5 Hotelling's T^2 distribution

Hotelling's T^2 distribution is a multivariate analogue of the Student t distribution. It plays an important role in multivariate hypothesis testing and confidence region construction, just as the Student t distribution does in the univariate setting.

Definition 6.3. Suppose $x \sim N_p(0, I_p)$ and $M \sim W_p(I_p, n)$ are independent, then the quantity $\tau^2 = nx^\top M^{-1} x$ is said to have Hotelling's T^2 distribution with parameters p and n. We write this as $\tau^2 \sim T^2(p, n)$.

We can generalise the definition with the following result.

Proposition 6.12. *Suppose $x \sim N_p(\mu, \Sigma)$ and $M \sim W_p(\Sigma, n)$ are independent and Σ has full rank p. Then*

$$n(x-\mu)^\top M^{-1}(x-\mu) \sim T^2(p, n).$$

Proof. Define $y = \Sigma^{-1/2}(x-\mu)$. Then, by Corollary 6.3, $y \sim N_p(0, I_p)$. Further, let $Z = \Sigma^{-1/2} M \Sigma^{-1/2}$ then $Z \sim W_p(I_p, n)$ by applying 6.8 with $A = \Sigma^{-1/2}$. From the definition, $ny^\top Z^{-1} y \sim T^2(p, n)$ and

$$\begin{aligned} ny^\top Z^{-1} y &= n(x-\mu)^\top \Sigma^{-1/2} \Sigma^{1/2} M^{-1} \Sigma^{1/2} \Sigma^{-1/2} (x-\mu) \\ &= n(x-\mu)^\top M^{-1}(x-\mu) \end{aligned}$$

so the result is proved. □

This result gives rise to an important corollary used in hypothesis testing when Σ is unknown.

Corollary 6.5. *If $\bar{x}$ and S are the mean and covariance matrix based on a sample of size n from $N_p(\mu, \Sigma)$ then*

$$(n-1)(\bar{x}-\mu)^\top S^{-1}(\bar{x}-\mu) \sim T^2(p, n-1).$$

Proof. We have seen earlier that $\bar{x} \sim N_p(\mu, \frac{1}{n}\Sigma)$. Let $x^* = n^{1/2}\bar{x}$ and let $\mu^* = n^{1/2}\mu$. Then $x^* = n^{1/2}\bar{x} \sim N_p(\mu^*, \Sigma)$.

From Proposition 6.11 we know $nS \sim W_p(\Sigma, n-1)$, and from Theorem 6.6 we know $\bar{x}$ and S are independent. Applying Proposition 6.12 with $x = x^*$ and $M = nS$ we obtain

$$(n-1)(x^*-\mu^*)^\top (nS)^{-1}(x^*-\mu^*) \sim T^2(p, n-1),$$

and given $x^* - \mu^* = n^{1/2}(x-\mu)$ then

$$\begin{aligned}(n-1)&(x^*-\mu^*)^\top (nS)^{-1}(x^*-\mu^*)\\ &= (n-1)n^{1/2}(\bar{x}-\mu)^\top n^{-1}S^{-1}n^{1/2}(\bar{x}-\mu)\\ &= (n-1)(\bar{x}-\mu)^\top S^{-1}(\bar{x}-\mu).\end{aligned}$$

□

Hotelling's T^2 distribution is not often included in statistical tables but the next result tells us that Hotelling's T^2 is a scale transformation of an F distribution.

Proposition 6.13. *If $\tau^2 \sim T^2(p, n)$ then*

$$\gamma^2 = \frac{n-p+1}{np}\tau^2 \sim F_{p,n-p+1}.$$

Proof. Beyond the scope of the module. □

We can apply this result to the previous corollary.

Corollary 6.6. *If $\tau^2 = (n-1)(\bar{x}-\mu)^\top S^{-1}(\bar{x}-\mu)$ then*

$$\gamma^2 = \frac{n-p}{p}(\bar{x}-\mu)^\top S^{-1}(\bar{x}-\mu) \sim F_{p,n-p}.$$

Proof. From Corollary 6.6 we know $\tau^2 \sim T^2(p, n-1)$. Applying Proposition 6.13 we get

$$\begin{aligned}&\gamma^2\\ &= \frac{(n-1)-p+1}{(n-1)p}(n-1)(\bar{x}-\mu)^\top S^{-1}(\bar{x}-\mu) \sim F_{p,(n-1)-p+1}\\ &= \frac{n-p}{p}(\bar{x}-\mu)^\top S^{-1}(\bar{x}-\mu) \sim F_{p,n-p}\end{aligned}$$

□

Chapter 7

Inference in 1 and 2 samples based on MVN

In the univariate case the Student t distribution plays a key role when we are dealing with normal random samples (i) in hypothesis testing for means (especially the 'paired' t-test and the t-test for comparing means in two independent samples), and (ii) the construction of confidence intervals for means. In this chapter we develop the analogous results in the multivariate case, where observations are now assumed to be random vectors. The role of the Student t distribution will be played by Hotelling's T^2, and the role of the χ^2 is played by the Wishart distribution.

Despite there being important technical differences, the results in the multivariate case are seen to be natural generalisations of the univariate results.

7.1 Hypothesis testing: Σ known

Let $x_1, \ldots, x_n$ be a random sample from $N_p(\mu, \Sigma)$ where Σ is assumed known and $\mu = (\mu_1, \ldots, \mu_p)^\top$. We wish to test the null hypothesis $\mu = a$, where a is fixed and pre-specified, against the alternative $\mu \neq a$. We could conduct p separate univariate tests with null hypotheses

$$H_0 : \mu_i = a_i \qquad \text{vs.} \qquad H_1 : \mu_i \neq a_i,$$

$i = 1, \ldots, p$. However, this ignores possible correlations between variables involved in different tests. An alternative approach is to conduct a single hypothesis test using Proposition 6.5.

Recall that $\bar{x} \sim N_p(\mu, \frac{1}{n}\Sigma)$. Therefore $n^{1/2}\bar{x} \sim N_p(n^{1/2}\mu, \Sigma)$. Applying Proposition we see that

$$(n^{1/2}\bar{x} - n^{1/2}\mu)^\top \Sigma^{-1} (n^{1/2}\bar{x} - n^{1/2}\mu) = n(\bar{x} - \mu)^\top \Sigma^{-1} (\bar{x} - \mu) \sim \chi^2_p$$

We use this as our test statistic in a test at the $\alpha\%$ significance level of the following hypotheses:

$$H_0 : \mu = a \qquad \text{vs} \qquad H_1 : \mu \neq a,$$

where a is a fixed, pre-specified vector. The relevant test statistic is

$$\zeta^2 = n(\bar{x} - a)^\top \Sigma^{-1} (\bar{x} - a),$$

and when H_0 is true, then $\zeta^2 \sim \chi^2_p$.

Critical value: We reject H_0 if $\zeta^2 > \chi^2_{p,\alpha}$.

Alternatively, we can state the result as a p-value where $p = P(\chi^2_p > \zeta^2_{\text{obs}})$, and ζ^2_{obs} is the observed value of the statistic ζ^2.

The multivariate equivalent of a confidence interval is a confidence region and the $100(1-\alpha)\%$ confidence region for μ is $\{a : \zeta^2 \leq \chi^2_{p,\alpha}\}$.

This confidence region will be the interior of an ellipse or ellipsoid.

Example 7.1. The scatterplot shows the module marks for $n = 209$ students on probability (PRB, x_1) and statistics (STA, x_2).

FIX FIGURE

The observations $x_1, \ldots, x_{209}$ are assumed to be a random sample from $N_2(\mu, \Sigma)$ where

$$\Sigma = \begin{pmatrix} 200 & 150 \\ 150 & 300 \end{pmatrix},$$

and the sample mean vector is $\bar{x} = (61.957, 62.632)^\top$. The target for the module mean for a large population of students should be exactly 60 for both modules. We now conduct a hypothesis test of H_0 versus H_1 at the 5% level to see if the lecturers have missed their target, where

$$H_0 : \mu = \begin{pmatrix} 60 \\ 60 \end{pmatrix} \qquad \text{and} \qquad H_1 : \mu \neq \begin{pmatrix} 60 \\ 60 \end{pmatrix}.$$

The test statistic is

$$\zeta^2 = 209 \begin{pmatrix} 61.957 - 60 \\ 62.632 - 60 \end{pmatrix}^\top \begin{pmatrix} 200 & 150 \\ 150 & 300 \end{pmatrix}^{-1} \begin{pmatrix} 61.957 - 60 \\ 62.632 - 60 \end{pmatrix}.$$

Now $|\Sigma| = 200 \times 300 - 150^2 = 37500$, so

$$\Sigma^{-1} = \frac{1}{37500} \begin{pmatrix} 300 & -150 \\ -150 & 200 \end{pmatrix} = \begin{pmatrix} 0.008 & -0.004 \\ -0.004 & 0.016/3 \end{pmatrix}$$

and

$$\zeta^2 = 209 \begin{pmatrix} 1.957 \\ 2.632 \end{pmatrix}^\top \begin{pmatrix} 0.008 & -0.004 \\ -0.004 & 0.016/3 \end{pmatrix} \begin{pmatrix} 1.957 \\ 2.632 \end{pmatrix} = 5.512.$$

The critical value is $\chi^2_{2,0.05} = 5.991$ so $\zeta^2 < \chi^2_{p,0.05}$ and we do not reject the null hypothesis at the 5% level.

Note that if we had conducted separate univariate hypothesis tests of $H_0 : \mu_1 = 60$ and $H_0 : \mu_2 = 60$ then the test statistics would have been:

$$\begin{aligned} z_1 &= \frac{\bar{x}_1 - \mu_1}{\sqrt{\sigma_1^2/n}} = \frac{61.957 - 60}{\sqrt{200/209}} = 2.000 \\ z_2 &= \frac{\bar{x}_2 - \mu_2}{\sqrt{\sigma_2^2/n}} = \frac{62.632 - 60}{\sqrt{300/209}} = 2.196. \end{aligned}$$

The critical value would have been $Z_{0.025} = 1.960$ and both null hypotheses would have been rejected. Therefore a multivariate hypothesis can be accepted when each of its univariate components is rejected and vice-versa.

Returning to the multivariate test, the p-value is 0.064 and the 95% confidence region is the interior of an ellipse, centred on $\bar{x}$, with the angle of the major-axis governed by Σ. We can see from the plot below that $(60, 60)^\top$, marked with a cross, lies just inside the confidence region.

FIX FIGURE

7.2 Hypothesis testing - 1 sample case

In §7.2 we considered a hypothesis test of $H_0 : \mu = a$ vs. $H_1 : \mu \neq a$ based on an IID sample from $N_p(\mu, \Sigma)$ when Σ was known. In reality, we rarely know Σ, so we replace it with the sample covariance matrix, S, and the χ^2_p distribution is replaced with the $F_{p,n-p}$ distribution using Corollary 6.6. The details are as follows.

Hypotheses: $H_0 : \mu = a$ vs $H_1 : \mu \neq a$.

Test statistic: $\gamma^2 = \frac{n-p}{p}(\bar{x} - a)^\top S^{-1}(\bar{x} - a)$, where $\gamma^2 \sim F_{p,n-p}$ under the the assumption that H_0 is true.

Critical value: We reject H_0 if $\gamma^2 > F_{p,n-p,\alpha}$, where α is the significance level.

Alternatively, we can state the result as a p-value where $p = P(F_{p,n-p} > \gamma^2)$.

The $100(1-\alpha)\%$ confidence region for μ is $\{a : \gamma^2 \leq F_{p,n-p,\alpha}\}$, which will again be the interior of an ellipse or ellipsoid, but the confidence region is now determined by S rather than Σ.

Example 7.2. We return to the example of §7.2, with the module marks for $n = 209$ students on probability (PRB, x_1) and statistics (STA, x_2).

The observations $x_1, \ldots, x_{209}$ are assumed to be a random sample from $N_2(\mu, \Sigma)$, but now we assume that Σ is unknown. We calculate the sample mean and

sample covariance matrix as,

$$\bar{x} = \begin{pmatrix} 61.957 \\ 62.632 \end{pmatrix} \qquad S = \begin{pmatrix} 215.29 & 157.19 \\ 157.19 & 333.56 \end{pmatrix}.$$

We conduct a hypothesis test at the 5% level of:

$$H_0 : \mu = \begin{pmatrix} 60 \\ 60 \end{pmatrix} \qquad \text{vs.} \qquad H_1 : \mu \neq \begin{pmatrix} 60 \\ 60 \end{pmatrix}.$$

The test statistic is

$$\begin{aligned} \gamma^2 &= \frac{209-2}{2} \begin{pmatrix} 61.957-60 \\ 62.632-60 \end{pmatrix}^\top \begin{pmatrix} 215.29 & 157.19 \\ 157.19 & 333.56 \end{pmatrix}^{-1} \begin{pmatrix} 61.957-60 \\ 62.632-60 \end{pmatrix} \\ &= \frac{207}{2} \begin{pmatrix} 1.957 \\ 2.632 \end{pmatrix}^\top \begin{pmatrix} 0.0071 & -0.0033 \\ -0.0033 & 0.0046 \end{pmatrix} \begin{pmatrix} 1.957 \\ 2.632 \end{pmatrix} = 2.525. \end{aligned}$$

The critical value is $F_{2,207,0.05} = 3.040$ so $\gamma^2 < F_{p,n-p,0.05}$ and we do not reject the null hypothesis at the 5% level.

The p-value is 0.082 and the 95% confidence region is the interior of an ellipse, centred on $\bar{x}$, with the angle of the major-axis governed by S. The confidence region is slightly larger than when Σ was known.

FIX FIGURE

7.3 Hypothesis testing - 2 sample case

As with the univariate case, we may wish to test the difference between two population means. As with univariate statistics, there are two cases to consider:

Paired case If $m = n$ and there exists some experimental link between x_i and y_i then we can look at the differences $z_i = y_i - x_i$ for $i = 1, \ldots, n$. For example, x_i and y_i could be vectors of pre-treatment and post-treatment measurements, respectively, of the same variables. The crucial assumption is that the differences z_i are IID $N_p(\mu, \Sigma)$. To examine the null hypothesis of no difference between the means we would test $H_0 : \mu = \mathbf{0}_p$ against $H_1 : \mu \neq \mathbf{0}_p$.

We then base our inference on $\bar{z} = \frac{1}{n}\sum_{i=1}^n z_i = \bar{y} - \bar{x}$, and proceed exactly as in the 1 sample case, using the test in §7.2 if Σ is known, or the test in §7.3 if Σ if unknown with $S = \frac{1}{n}\sum_{i=1}^n (z_i - \bar{z})(z_i - \bar{z})^\top$.

Unpaired case The unpaired case is where x_i and y_i are independent and not connected to each other. For example, in a clinical trial we may have two separate groups of patients, where one group receives a placebo and the other group receives an active treatment. Let $x_1, \ldots, x_n$ be an IID sample from $N_p(\mu_1, \Sigma)$ and let $y_1, \ldots, y_m$ be an IID sample from $N_p(\mu_2, \Sigma)$. In this case, we can base our inference on the following result.

Proposition 7.1. *Let $x_1, \ldots, x_n$ be a random sample from $N_p(\mu_1, \Sigma_1)$ and let $y_1, \ldots, y_m$ be a random sample from $N_p(\mu_2, \Sigma_2)$. Then when $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$,*

$$\frac{nm}{n+m}(\bar{y} - \bar{x})^\top S_u^{-1}(\bar{y} - \bar{x}) \sim T^2(p, n+m-2),$$

where

$$S_u = \frac{nS_1 + mS_2}{n+m-2}$$

is the pooled unbiased variance matrix estimator and S_j is the sample covariance matrix for group j, $j = 1, 2$.

Proof. From Result 1.1 FIX and Proposition 6.1 we know that $\bar{x} \sim N_p(\mu_1, n^{-1}\Sigma_1)$ and $\bar{y} \sim N_p(\mu_2, m^{-1}\Sigma_2)$, and $\bar{x}$ and $\bar{y}$ are independent, so

$$\bar{y} - \bar{x} \sim N_p\left(\mu_2 - \mu_1, \frac{1}{n}\Sigma_1 + \frac{1}{m}\Sigma_2\right).$$

If $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2 = \Sigma$, then $\bar{y} - \bar{x} \sim N_p\left(0_p, \left(\frac{1}{n} + \frac{1}{m}\right)\Sigma\right)$ and

$$z = \left(\frac{1}{n} + \frac{1}{m}\right)^{-1/2}(\bar{y} - \bar{x}) \sim N_p(0_p, \Sigma).$$

From Proposition 6.11 we know that $nS_1 \sim W_p(\Sigma_1, n-1)$ and $mS_2 \sim W_p(\Sigma_2, m-1)$. Therefore when $\Sigma_1 = \Sigma_2 = \Sigma$,

$$\begin{aligned} M = (n+m-2)S_u &= (n+m-2)\frac{nS_1 + mS_2}{n+m-2} \\ &= nS_1 + mS_2 \sim W_p(\Sigma, n+m-2) \end{aligned}$$

by Proposition 6.10, using the fact that S_1 and S_2 are independent.

Now z is independent of M, since $\bar{x}$ and $\bar{y}$ are independent of S_1 and S_2, respectively, by Proposition 6.6. Therefore, applying Proposition 6.12 with $x = z$ and $M = (n+m-2)S_u$, we have

$$(n+m-2)z^\top((n+m-2)S_u)^{-1}z = z^\top S_u^{-1} z \sim T^2(p, n+m-2)$$

and

$$\begin{aligned} z^\top S_u^{-1} z &= \left(\frac{1}{n} + \frac{1}{m}\right)^{-1/2}(\bar{y} - \bar{x})^\top S_u^{-1}\left(\frac{1}{n} + \frac{1}{m}\right)^{-1/2}(\bar{y} - \bar{x}) \\ &= \left(\frac{1}{n} + \frac{1}{m}\right)^{-1}(\bar{y} - \bar{x})^\top S_u^{-1}(\bar{y} - \bar{x}). \end{aligned}$$

Finally,

$$\left(\frac{1}{n} + \frac{1}{m}\right)^{-1} = \left(\frac{m}{nm} + \frac{n}{nm}\right)^{-1} = \left(\frac{n+m}{nm}\right)^{-1} = \frac{nm}{n+m},$$

so Proposition 7.1 is proved.

□

As in the one sample case, we can convert Hotelling's two-sample T^2 statistic to the F distribution using Proposition 6.13.

Corollary 7.1. *Using the notation of Proposition 7.1, it follows that*

$$\delta^2 = \frac{(n+m-p-1)}{(n+m-2)p}\frac{nm}{(n+m)}(\bar{y}-\bar{x})^\top S_u^{-1}(\bar{y}-\bar{x}) \sim F_{p,n+m-p-1}.$$

```
Simply apply Proposition \@ref(prp:six14) to the statistic in
Proposition \@ref(prp:seven1) (replace $n$ with $n+m-2$).
```

Example 7.3. For the probability and statistics marks in Example 3.5, is there a significant difference between students registered on G100 and G103? The data is shown below, together with the sample means.

FIX FIGURE

Let μ_1 and μ_2 be the population means for G100 and G103 respectively. Our hypotheses are

$$H_0 : \mu_1 = \mu_2 \qquad \text{vs.} \qquad H_1 : \mu_1 \neq \mu_2.$$

Let $x_1, \ldots, x_n$ be the marks for G100 students, which we assume are a random sample from $N_2(\mu_1, \Sigma_1)$. Similarly, let $y_1, \ldots, y_m$ be the marks for G103 students, which we assume are a random sample from $N_2(\mu_2, \Sigma_2)$. The sample summary statistics are:

$$\begin{array}{cc} n = 98 & m = 46 \\ \bar{x} = \begin{pmatrix} 60.582 \\ 62.786 \end{pmatrix} & \bar{y} = \begin{pmatrix} 64.761 \\ 60.457 \end{pmatrix} \\ S_1 = \begin{pmatrix} 201.04 & 129.56 \\ 129.56 & 316.21 \end{pmatrix} & S_2 = \begin{pmatrix} 229.88 & 177.02 \\ 177.02 & 354.16 \end{pmatrix} \end{array}$$

The assumption $\Sigma = \Sigma_1 = \Sigma_2$ does not look unreasonable given the sample covariance matrices, so we compute

$$\begin{aligned} S_u &= \frac{1}{98+46-2}\left(98\begin{pmatrix} 201.04 & 129.56 \\ 129.56 & 316.21 \end{pmatrix} + 46\begin{pmatrix} 229.88 & 177.02 \\ 177.02 & 354.16 \end{pmatrix}\right) \\ &= \begin{pmatrix} 213.21 & 146.76 \\ 146.76 & 332.96 \end{pmatrix} \end{aligned}$$

and, therefore, $S_u^{-1} = \begin{pmatrix} 0.0067 & -0.0030 \\ -0.0030 & 0.0043 \end{pmatrix}$.

The test statistic is

$$\delta^2 = \frac{141}{284} \times \frac{4508}{144}\begin{pmatrix} 4.179 \\ -2.329 \end{pmatrix}^\top \begin{pmatrix} 0.0067 & -0.0030 \\ -0.0030 & 0.0043 \end{pmatrix}\begin{pmatrix} 4.179 \\ -2.329 \end{pmatrix} = 3.089$$

The critical value for $\alpha = 0.05$ is

$$F_{2,98+46-2-1,\alpha} = F_{2,141,0.05} = 3.060.$$

Therefore $\delta^2 > F_{p,n+m-p-1}$, so we reject the null hypothesis at the 5% level. The p-value is 0.049. So there is moderate evidence against H_0, the null hypothesis that $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$.

Chapter 8

The Multivariate Linear Model

In the standard linear model, the response variable, y, is univariate and the mean of y, $\mu = E[y]$, is modelled as a linear function of the elements of a covariate vector $x = (x_1, \ldots, x_q)^\top \in \mathbb{R}^q$, i.e. it is assumed that

$$\mu = \boldsymbol{\beta}^\top x,$$

where $\boldsymbol{\beta} \in \mathbb{R}^q$ is an unknown parameter vector to be estimated. In this chapter we consider a generalisation of the standard linear model in which the response, y, is now a $p \times 1$ vector. In this setting, the linear model takes the form

$$\boldsymbol{\mu} = B^\top x,$$

where the mean vector μ is $p \times 1$, the covariate vector x is $q \times 1$ and the parameter vector B is $q \times p$. The reason for having $B^\top$ above rather than simply B will become clear later in the chapter.

8.1 The standard univariate linear model

In this section we give a brief review of the standard linear model in which the response is univariate. Then, in subsequent sections, we discuss the multivariate linear model, i.e. multiple regression with a vector response which has a general covariance matrix.

Consider the univariate linear model in which

$$y_i = x_i^\top \boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \ldots, n, \tag{8.1}$$

where $\overset{q\times 1}{\boldsymbol{\beta}}$ is a parameter vector and x_i is a $q\times 1$ covariate vector for experimental unit i. We can also write (8.1) in equivalent vector-matrix form

$$y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{8.2}$$

where $\overset{n\times q}{X} = [x_1, \ldots, x_n]^\top$ is the matrix of covariates, $\overset{n\times 1}{y}$ is the vector of univariate responses and $\overset{n\times 1}{\boldsymbol{\epsilon}}$ is the vector of univariate 'error' terms.

It is assumed throughout this chapter that X has full rank p, so that $(X^\top X)^{-1}$ exists.

Most or all of the following assumptions are usually made in the standard linear model:

1. For each $i = 1, \ldots, n$, $E[\epsilon_i] = 0$.
2. The ϵ_i are uncorrelated, i.e. for $i \neq j$, $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$.
3. The ϵ_i have constant variance, i.e. $\mathrm{Var}(\epsilon_i) = \sigma^2$, i.e. σ^2 does not depend on i.
4. The ϵ_i are IID $N(0, \sigma^2)$.

It is clear that assumption 4. implies each of assumptions 1-3. However, note that the least squares approach to be discussed below makes sense under assumptions 1-3 alone. The attraction of assumption 4. is that it enables us to perform exact inference since the relevant distributions are known exactly, and the estimators of $\boldsymbol{\beta}$ and σ^2 are maximum likelihood estimators (MLEs). We shall assume 4. and its multivariate analogue, see (8.12) below, throughout this chapter.

The log-likelihood for models (8.1) and (8.2) under the Gaussian assumption 4. is given by

$$\begin{aligned}\ell(\boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2}\log(\sigma^2) - -\frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^\top\boldsymbol{\beta})^2 \\ &= -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}(y - X\boldsymbol{\beta})^\top(y - X\boldsymbol{\beta}).\end{aligned}$$

Applying the results in §2.10 to the second expression above,

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^2}X^\top(y - X\boldsymbol{\beta})$$

and

$$\frac{\partial \ell}{\partial \sigma^2}(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}(y - X\boldsymbol{\beta})^\top(y - X\boldsymbol{\beta}).$$

Setting $\partial\ell(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2))/\partial\boldsymbol{\beta} = \mathbf{0}_q$, the zero vector, and assuming that $X^\top X$ is invertible, implies that

$$\hat{\boldsymbol{\beta}} = \left(X^\top X\right)^{-1} X^\top y. \tag{8.3}$$

Also, setting $\partial \ell(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2))/\partial \sigma^2 = 0$ gives

$$\hat{\sigma}^2 = \frac{1}{n} y^\top P y, \tag{8.4}$$

where

$$P = I_n - X \left(X^\top X \right)^{-1} X^\top \tag{8.5}$$

is a projection matrix. The maximised log-likelihood is given by

$$\begin{aligned} \ell(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) &= -\frac{n}{2} \log(\hat{\sigma}^2) - \frac{n}{2} \log(2\pi) - \frac{1}{2\hat{\sigma}^2} (y - X\hat{\boldsymbol{\beta}})^\top (y - X\hat{\boldsymbol{\beta}}) \\ &= -\frac{n}{2} \log(\hat{\sigma}^2) - \frac{n}{2} \log(2\pi) - \frac{n}{2y^\top P y} y^\top P y \\ &= -\frac{n}{2} \log(\hat{\sigma}^2) - \frac{n}{2} \log(2\pi) - \frac{n}{2}. \end{aligned}$$

Under the IID Gaussian assumption for the ϵ_i,

$$\hat{\boldsymbol{\beta}} \sim N_q \left\{ \boldsymbol{\beta}, \sigma^2 (X^\top X)^{-1} \right\},$$

and

$$n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-q}.$$

Before moving on, we briefly discuss an important case of the standard linear model - the one-way analysis of variance. Here, there are t groups, say, with n_j observations in group j, and the model takes the form

$$y_{ij} = \mu_j + \epsilon_{ij}, \qquad i = 1, \ldots, n_j; \quad j = 1, \ldots t, \tag{8.6}$$

where the ϵ_{ij} are IID $N(0, \sigma^2)$ random variables, and μ_j is the mean of population j. Note that we are assuming that $\mathrm{Var}(y_{ij}) = \sigma^2$ is constant across populations $j = 1, \ldots, t$.

Define

$$\bar{y}_{+j} = n_j^{-1} \sum_{i=1}^{n_j} y_{ij}, \qquad j = 1, \ldots, t; \qquad \bar{y}_{++} = n^{-1} \sum_{j=1}^{t} \sum_{i=1}^{n_j} y_{ij},$$

where $n = \sum_{j=1}^{t} n_j$. It is easily checked that, under model (8.6), the MLE of μ_j is $\bar{y}_{+j}$. Consider the null hypothesis

$$H_0 : \mu_1 = \cdots = \mu_t.$$

The total sum of squares, T, defined below, has the following decomposition:

$$\begin{aligned} T &\equiv \sum_{j=1}^{t}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_{++})^2 \\ &= \sum_{j=1}^{t}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_{+j}+\bar{y}_{+j}-\bar{y}_{++})^2 \\ &\sum_{j=1}^{t}\sum_{i=1}^{n_j}\{(y_{ij}-\bar{y}_{+j})^2+(\bar{y}_{+j}-\bar{y}_{++})^2+2(y_{ij}-\bar{y}_{+j})(\bar{y}_{+j}-\bar{y}_{++})\} \\ &= \sum_{j=1}^{t}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_{+j})^2+\sum_{j=1}^{t}n_j(\bar{y}_{+j}-\bar{y}_{++})^2 \\ &= W+B, (\#eq: T=B+W) \end{aligned} \tag{8.7}$$

using the fact that sum over i of the product term above is 0. In the above, W stands for `within' sum of squares, also known as the residual sum of squares, and $B$ stands for`between' sum of squares. The MLE of σ^2 is given by W/n. Standard theory tells us that under model (8.6), $W \sim \sigma^2\chi^2_{n-t}$. Moreover, under H_0, $B \sim \sigma^2\chi^2_{t-1}$, W and B are independent, and therefore $T \sim \sigma^2\chi^2_{n-1}$. To test H_0 we use the fact that, under H_0,

$$f = \frac{B/(t-1)}{W/(n-t)} \tag{8.8}$$

has an $F_{t-1,n-t}$ distribution. We reject H_0 when f is 'large' compared with an $F_{t-1,n-t}$ random variable.

In Example Sheet 3 you are asked to show that the 'twice the log-likelihood ratio statistic' for testing H_0 against the general alternative with $\mu_1, \ldots, \mu_t$ is given by

$$n_+ \log\left(1+\frac{B}{W}\right) = n_+ \log(1+W^{-1}B), \tag{8.9}$$

which is an increasing function of the statistic (8.8). Consequently, we can think of the classical F test as being equivalent to a likelihood ratio test.

8.2 Multivariate Linear Model

In the standard linear model the responses y_i are univariate. In the multivariate linear model, the responses are $p \times 1$ vectors y_i. Here, the linear model takes the form

$$y_i = B^\top x_i + \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, n, \tag{8.10}$$

where $\underset{q\times p}{B}$ is a parameter matrix, the x_i are $q \times 1$ covariate vectors as in §8.2, and the $\boldsymbol{\epsilon}_i$ are $p \times 1$ error vectors. The model (8.10) may be written in matrix form as

$$Y = XB + E, \tag{8.11}$$

where $\underset{n\times p}{Y} = [y_1, \dots, y_n]^\top$ is the data matrix for the y-variables, $X = [x_1, \dots, x_n]^\top$ is the $n \times q$ data matrix for the x-variables, defined as in §8.2, and $\underset{n\times p}{E} = [\boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_n]^\top$.

By analogy with assumption 4. in §8.2, it is assumed that

$$\boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_n \quad \text{are IID} \quad N_p(\mathbf{0}_p, \Sigma). \tag{8.12}$$

Under this MVN assumption, the log-likelihood function $\ell(B, \Sigma)$ for the parameter matrices B and Σ is given by

$$\begin{aligned}\ell(B, \Sigma) = &-\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma|) \\ &- \frac{1}{2}\mathrm{tr}\left\{(Y - XB)\Sigma^{-1}(Y - XB)^\top\right\}.\end{aligned} \tag{8.13}$$

Proposition 8.1. *The maximum likelihood estimators of B and Σ in* (8.13) *are given by*

$$\hat{B} = (X^\top X)^{-1}X^\top Y \tag{8.14}$$

and

$$\hat{\Sigma} = \frac{1}{n}Y^\top PY, \tag{8.15}$$

where P is the matrix defined in (8.5)*. The maximised log-likelihood is given by*

$$\ell(\hat{B}, \hat{\Sigma}) = -\frac{n}{2}\log(|\hat{\Sigma}|) - \frac{np}{2}\log(2\pi) - \frac{np}{2}. \tag{8.16}$$

Remark: note how similar (8.14) and (8.15) are to their univariate counterparts (8.3) and (8.4), respectively; the only thing that is different is that Y is now an $n \times p$ matrix rather than an $n \times 1$ vector.

Proof. **of** (8.14)**}.** Recall that P defined in (8.5) is a projection matrix, i.e. $P^\top = P$ and $P^2 = P$. Moreover,

$$PX = X - X(X^\top X)^{-1}X^\top X = X - X = \mathbf{0}_{n,q} \tag{8.17}$$

and

$$X^\top P = X^\top - X^\top X(X^\top X)^{-1}X^\top = X^\top - X^\top = \mathbf{0}_{q,n}. \tag{8.18}$$

Now write

$$\begin{aligned}Y - XB &= Y - X\hat{B} + X\hat{B} - XB \\ &= Y - X(X^\top X)^{-1}X^\top Y + X(\hat{B} - B) \\ &= PY + X(\hat{B} - B).\end{aligned}$$

Then, using (8.17) and (8.18),

$$\begin{aligned}
&(Y - XB)^\top (Y - XB) \\
&\quad = \{PY + X(\hat{B} - B)\}^\top \{PY + X(\hat{B} - B)\} \\
&\quad = Y^\top PY + Y^\top PX(\hat{B} - B) \\
&\qquad\qquad + (\hat{B} - B)^\top X^\top PY + (\hat{B} - B)^\top X^\top X(\hat{B} - B) \\
&\quad = Y^\top PY + (\hat{B} - B)^\top X^\top X(\hat{B} - B). \qquad (8.19)
\end{aligned}$$

As noted in Chapter 2, for any compatible matrices W and Z, we have $\mathrm{tr}(WZ) = \mathrm{tr}(ZW)$. So, using (8.19), it follows that the trace term in (8.13) may be written

$$\begin{aligned}
&\mathrm{tr}\left\{(Y - XB)\Sigma^{-1}(Y - XB)^\top\right\} \\
&= \mathrm{tr}\left\{\Sigma^{-1}(Y - XB)^\top (Y - XB)\right\} \\
&= \mathrm{tr}\left\{\Sigma^{-1}(Y - X\hat{B} + X\hat{B} - XB)^\top (Y - X\hat{B} + X\hat{B} - XB)\right\} \\
&= \mathrm{tr}\left[\Sigma^{-1}\{Y^\top PY + (\hat{B} - B)^\top X^\top X(\hat{B} - B)\}\right] \\
&= \mathrm{tr}\left(\Sigma^{-1}Y^\top PY\right) + \mathrm{tr}\left\{X(\hat{B} - B)\Sigma^{-1}(\hat{B} - B)^\top X^\top\right\} \qquad (8.20)
\end{aligned}$$

The first term on the RHS of (8.20) does not depend on B. In the second term, the matrix inside the trace is non-negative definite for all B, and therefore the second term in (8.20) is non-negative, and has a minimum value of 0, achieved uniquely when $B = \hat{B}$. Therefore, for fixed Σ, (8.13) is maximised when $B = \hat{B}$, as defined in (8.14). □

To prove (8.15), we require the following result, whose proof is omitted because it is a bit too technical. □

Proof. Suppose that A is a fixed symmetric positive definite $p \times p$ matrix and define the function

$$G(\Sigma) \equiv -\frac{n}{2}\log\left(|\Sigma|\right) - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}A\right).$$

Then $G(\Sigma)$ is maximised over symmetric positive definite $p \times p$ matrices Σ at $\hat{\Sigma} = n^{-1}A$, and the maximum value of G is given by

$$G(\hat{\Sigma}) = -\frac{n}{2}\log\left(|n^{-1}A|\right) - np/2.$$

□

Proof. **of** (8.15) **and** (8.16)**}.** Again using the result $\mathrm{tr}(WZ) = \mathrm{tr}(ZW)$ for compatible matrices W and Z, we have

$$\mathrm{tr}\{(Y - X\hat{B})\Sigma^{-1}(Y - X\hat{B})^\top\} = \mathrm{tr}(\Sigma^{-1}(PY)^\top PY) = \mathrm{tr}(\Sigma^{-1}Y^\top PY),$$

because P is a projection matrix. Consequently,

$$\ell(\hat{B}, \Sigma) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}\left\{\Sigma^{-1}Y^\top PY\right\},$$

and we want to maximise $\ell(\hat{B}, \Sigma)$ over symmetric positive definite matrices Σ. So take $A = Y^\top PY$. Then, using Proposition **??**, $\hat{\Sigma} = n^{-1}Y^\top PY$, which agrees with (8.15).

To prove (8.16), note that

$$\begin{aligned}\mathrm{tr}\{(Y - X\hat{B})\hat{\Sigma}^{-1}(Y - X\hat{B})^\top\} &= \mathrm{tr}\{\hat{\Sigma}^{-1}(PY)^\top PY\} \\ &= n\mathrm{tr}\{(Y^\top PY)^{-1}Y^\top PY\} \\ &= n\mathrm{tr}(I_p) = np,\end{aligned}$$

so (8.16) follows after substitution of $B = \hat{B}$ and $\Sigma = \hat{\Sigma}$ into (8.13). □

We now determine the joint distribution of $\hat{B}$ and $\hat{\Sigma}$.

Proposition 8.2. *Assume that* (8.11) *and* (8.12) *both hold and assume that X has full rank q, where $q < n$. Then the following results hold.*

1. *The MLE of B, $\hat{B}$ defined in* (8.14)*, and the MLE of Σ, $\hat{\Sigma}$ defined in* (8.15)*, are independent. Moreover, the elements of $\hat{B}$ are jointly multivariate normal, while $n\hat{\Sigma} \sim W_p(\Sigma, n-q)$.*

2. *The MLE $\hat{B}$ satisfies $E[\hat{B}] = B$, i.e. $\hat{B}$ is unbiased for B.*

3. *Write $\hat{b}_{[j]}$ for column j of $\hat{B}$, $j = 1, \ldots, p$, so that $\hat{B} = [\hat{b}_{[1]}, \ldots, \hat{b}_{[p]}]$. Then for $j, k = 1, \ldots, p$,*

$$Cov(\hat{b}_{[j]}, \hat{b}_{[k]}) = \sigma_{jk}(X^\top X)^{-1}.$$

Proof. of Proposition 8.2 First, observe that $\hat{\Sigma}$ is a function of the elements of PY, namely

$$\hat{\Sigma} = n^{-1}Y^\top P^\top PY = n^{-1}Y^\top PY,$$

because P is a projection matrix. Moreover, under the MVN assumption (8.12), PY and $\hat{B}$ are jointly MVN, so to prove part 1. it will be sufficient to prove that each column of $\hat{B}$ is uncorrelated with each column of PY. Write $y_{[j]}$ for column j of Y. Then

$$Y = [y_{[1]}, \ldots, y_{[p]}], \qquad PY = [Py_{[1]}, \ldots, Py_{[p]}],$$

and

$$\hat{B} \equiv [\hat{b}_{[1]}, \ldots, \hat{b}_{[p]}] = [(X^\top X)^{-1}X^\top y_{[1]}, \ldots, (X^\top X)^{-1}X^\top y_{[p]}],$$

where, as before, $\hat{b}_{[j]}$ is column j of $\hat{B}$. The covariance between column j of $\hat{B}$ and column k of PY is given by

$$\begin{aligned}\mathrm{Cov}(\hat{b}_{[j]}, Py_{[k]}) &= \mathrm{Cov}\{(X^\top X)^{-1}X^\top y_{[j]}, Py_{[k]}\} \\ &= (X^\top X)^{-1}X^\top \mathrm{Cov}(y_{[j]}, y_{[k]})P \\ &= (X^\top X)^{-1}X^\top \sigma_{jk} I_n P \\ &= \sigma_{jk}(X^\top X)^{-1}X^\top (I_n - X(X^\top X)^{-1}X^\top) = \mathbf{0}_{q,n},\end{aligned}$$

where $\Sigma = (\sigma_{jk})_{j,k=1}^p$. In the above we have used the definition of P in (8.5) and the fact that

$$\mathrm{Cov}(y_{[j]}, y_{[k]}) = \sigma_{jk} I_n \tag{8.21}$$

due to the independence of the rows of Y. Therefore $\hat{B}$ and PY are independent by Proposition 6.4, and by Proposition 6.1 (Cochran's theorem), $n\hat{\Sigma} \sim W_p(\Sigma, n-q)$ which concludes the proof of part 1.

Part 2. follows because, taking the expectation of (8.14) we find that

$$\begin{aligned}E[\hat{B}] &= E[(X^\top X)^{-1}X^\top Y] \\ &= (X^\top X)^{-1}X^\top E[Y] \\ &= (X^\top X)^{-1}X^\top XB = B.\end{aligned}$$

For part 3., using (8.21) again,

$$\begin{aligned}\mathrm{Cov}(\hat{b}_{[j]}, \hat{b}_{[k]}) &= \mathrm{Cov}\{(X^\top X)^{-1}X^\top y_{[j]}, (X^\top X)^{-1}X^\top y_{[k]}\} \\ &= (X^\top X)^{-1}X^\top \mathrm{Cov}(y_{[j]}, y_{[k]})X(X^\top X)^{-1} \\ &= (X^\top X)^{-1}X^\top \sigma_{jk} I_n X(X^\top X)^{-1} = \sigma_{jk}(X^\top X)^{-1},\end{aligned}$$

as required.

□

We now investigate the situation where we wish to test (8.11) against a submodel. Specifically, we consider the null hypothesis

$$H_0 : B = \begin{pmatrix} B^* \\ \mathbf{0}_{q-r,p} \end{pmatrix}, \qquad \Sigma \quad \text{unrestricted},$$

where B^* is $r \times p$ with $1 \le r < q$. The alternative hypothesis is

$$H_1 : \overset{q\times p}{B} \quad \text{unrestricted}, \qquad \Sigma \quad \text{unrestricted},$$

i.e. model (8.11) under the MVN assumption (8.12). Note that H_0 is nested within H_1. Let us write $\hat{B}_j$ and $\hat{\Sigma}_j$ for the MLEs of B and Σ under hypothesis H_j, $j = 0, 1$. Then, using (8.16) in Proposition 8.1 to calculate the maximised

likelihood under H_0 and H_1, we obtain the Wilks statistic, $\omega_{0,1}$ (equals "twice the different in maximised log likelihoods"):

$$\begin{aligned}\omega_{0,1} &= 2\{\ell(\hat{B}_1, \hat{\Sigma}_1) - \ell(\hat{B}_0, \hat{\Sigma}_0)\} \\ &= -n\log(|\hat{\Sigma}_1|) + n\log(|\hat{\Sigma}_0|) = n\log\left(\frac{|\hat{\Sigma}_0|}{|\hat{\Sigma}_1|}\right).\end{aligned}$$

In the univariate case there is a simple, explicit transformation from the Wilks statistic $\omega_{0,1}$ to the classical $F_{q-r,n-q}$ distribution for testing H_0 against H_1. In the multivariate case, the exact distribution of $\omega_{0,1}$ under H_0 does not have a simple relationship with an F-distribution, and the situation is a lot more complicated.

When n is large, however, we can use the large sample log-likelihood ratio test, which implies that, under H_0, $\omega_{0,1}$ is approximately χ^2, and we should reject H_0 when $\omega_{0,1}$ is sufficiently large.

The relevant degrees of freedom of the χ^2 are now calculated.

Under H_0, the number of free parameters is

$$rp + \frac{1}{2}p(p+1),$$

while under H_1 there are

$$qp + \frac{1}{2}p(p+1)$$

free parameters. So the difference is $(q-r)p$, and so we should refer $\omega_{0,1}$ to $\chi^2_{(q-r)p}$.

8.3 One-way MANOVA

We now consider the multivariate version of the one-way ANOVA considered at the end of §8.2, known as one-way MANOVA. The model is defined by

$$y_{ij} = \mu_j + \boldsymbol{\epsilon}_{ij}, \qquad i = 1, \ldots, n_j; \quad j = 1, \ldots, t, \tag{8.22}$$

where the $\boldsymbol{\epsilon}_{ij}$ are IID $N_p(\mathbf{0}_p, \Sigma)$, and the y_{ij} and μ_j are also $p \times 1$ vectors. It is assumed that $\mathrm{Var}(\boldsymbol{\epsilon}_i) = \Sigma$ is constant across the t populations.

One possibility is to use the linear model framework developed in §8.3. In this case, X is a matrix with each element equal to zero or one. However, it is also feasible to do the calculations directly. This is what we shall do.

The log-likelihood of model (8.22) is given by

$$\ell(\mu_1, \ldots, \mu_t) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma|) - \frac{1}{2}\sum_{j=1}^{t}\sum_{i=1}^{n_j}(y_{ij}-\mu_j)^\top \Sigma^{-1}(y_{ij}-\mu_j), \tag{8.23}$$

where $n = n_1 + n_2 + \cdots + n_t$.

Using results in §2.10, the partial derivative, or gradient, of (8.23) with respect to the vector μ_k is given by

$$\begin{aligned}\frac{\partial \ell}{\partial \mu_k}(\mu_1, \ldots, \mu_t, \Sigma) &= \sum_{i=1}^{n_j}\Sigma^{-1}(y_{ik}-\mu_k) \\ &= n_k\Sigma^{-1}(\bar{y}_{+k}-\mu_k).\end{aligned}$$

where $\bar{y}_{+k}$ is the sample mean of group k, i.e.

$$\bar{y}_{+k} = n_k^{-1}\sum_{i=1}^{n_k} y_{ik}.$$

So, setting $\partial\ell/\partial\mu_k = \mathbf{0}_p$ implies $\hat{\mu}_k = \bar{y}_{+k}$. Therefore

$$\begin{aligned}\ell(\hat{\mu}_1, \ldots, \hat{\mu}_t, \Sigma) &= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma|) \\ &\qquad - \frac{1}{2}\sum_{j=1}^{t}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_{+j})^\top\Sigma^{-1}(y_{ij}-\bar{y}_{+j}) \\ &= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma|) - \frac{1}{2}\mathrm{tr}(\Sigma^{-1}W),\end{aligned}$$

where W, defined by

$$W = \sum_{j=1}^{t}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_{+j})(y_{ij}-\bar{y}_{+j})^\top,$$

is the matrix version of the 'within' sum of squares considered in §8.2. Using Proposition **??**, we deduce that

$$\hat{\Sigma} = n^{-1}W.$$

Consequently,

$$\ell(\hat{\mu}_1, \ldots, \hat{\mu}_t, \hat{\Sigma}) = -\frac{n}{2}\log(|\hat{\Sigma}|) - \frac{np}{2} - \frac{np}{2}\log(2\pi).$$

Now consider the key null hypothesis for a one-way MANOVA:

$$H_0 : \mu_1 = \cdots = \mu_t, \qquad \Sigma \quad \text{unrestricted}.$$

Under H_0, the y_{ij} are IID $N_p(\mu, \Sigma)$, so the log-likelihood under H_0 is

$$\begin{aligned}\ell_0(\mu, \Sigma) &\equiv \ell(\mu, \dots, \mu, \Sigma) \\ &= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma|) - \frac{1}{2}\sum_{j=1}^{t}\sum_{i=1}^{n_j}(y_{ij}-\mu)^\top\Sigma^{-1}(y_{ij}-\mu),\end{aligned}$$

and so the MLE of μ under H_0 is given by

$$\hat{\mu}_0 = \frac{1}{n}\sum_{j=1}^{t}\sum_{i=1}^{n_j} y_{ij} = \bar{y}_{++},$$

and, using Proposition **??** again, it is seen that the MLE of Σ under H_0 is

$$\hat{\Sigma}_0 = n^{-1}T,$$

where T is the matrix analogue of the total sum of squares, i.e.

$$T = \sum_{j=1}^{t}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_{++})(y_{ij}-\bar{y}_{++})^\top.$$

The Wilks statistic, ω_0, for testing H_0 against the general alternative (8.22) is then

$$\begin{aligned}\omega_0 = 2\{\ell(\hat{\mu}_1, \dots, \hat{\mu}_t, \hat{\Sigma}) - \ell_0(\hat{\mu}_0, \hat{\Sigma}_0)\} &= n\log\left(|\hat{\Sigma}_0|/|\hat{\Sigma}|\right) \\ &= n\log\left(|T|/|W|\right). \qquad (8.24)\end{aligned}$$

The degrees of freedom under H_0 are $p + p(p+1)/2$ and the degrees of freedom under (8.22) are $pt + p(p+1)/2$, so the difference is $p(t-1)$. Consequently, when the n_j are all large, we should refer ω_0 to $\chi^2_{p(t-1)}$ and reject H_0 when ω_0 is sufficiently large.

It is not immediately obvious that (8.24) is a natural generalisation of (8.9). However, in Example Sheet 4 you are asked to prove that

$$T = W + B,$$

where B is the matrix analogue of the 'between' sum of squares B in @ref(eq:T=B+W), i.e.

$$B = \sum_{j=1}^{t} n_j(\bar{y}_{+j}-\bar{y}_{++})(\bar{y}_{+j}-\bar{y}_{++})^\top.$$

Consequently,

$$\begin{aligned}\omega_0 &= n\log(|T|/|W|) = n\log(|W^{-1}||W+B|) \\ &= n\log(|W^{-1}(W+B)|) = n\log(|I_p + W^{-1}B|),\end{aligned}$$

which *is* a natural generalisation of (8.9).

Part IV: Classification and Clustering

In Part IV, we focus on different methods of classification, i.e. allocating the observations in a sample to different subsets (or groups). In Chapter 9, we focus on an approach called discriminant analysis, in which we have a training sample available, and we use this training sample to set up a suitable classification rule. An important type of situation where discriminant analysis is used is in screening tests. Here, several variables may be measured on each of a number of individuals, and we want to decide whether each individual is ''negative'', in which case no further investigations are required, or''positive", in which case further tests are required.

In Chapter 10, we consider an alternative approach, known as cluster analysis, in which we allocate the observations into clusters (or similar subsets). Here, a training sample is not available and typically the number of clusters will not be known in advance. The idea is to form clusters in such a way that experimental units within clusters are as similar as possible, in a suitable sense, and experimental units in different clusters are as dissimilar as possible.

Chapter 9

Discriminant analysis - FIX THE FIGURES

Consider g populations $\Pi_1, \ldots, \Pi_g$. Each population is described by a pdf $f_j(x)$, $x \in \mathbb{R}^p$, $j = 1, \ldots, g$. Let $z \in \mathbb{R}^p$ be a 'new' observation assumed to come from one of $\Pi_1, \ldots, \Pi_g$. The aim of discriminant analysis is to allocate z to one of $\Pi_1, \ldots, \Pi_g$ with 'as small a \ probability of error as possible'.

For example, z might contain numerical measures of a person's \ financial history. A credit rating agency might then want to \ classify this customer as ''safe" or ''risky" based on knowledge of \ previous customers.

A **discriminant rule**, d, corresponds to a division of $\mathbb{R}^p$ into disjoint regions $\mathcal{R}_1, \ldots, \mathcal{R}_g$, where

$$\bigcup_{j=1}^{g} \mathcal{R}_j = \mathbb{R}^p, \qquad \mathcal{R}_j \cap \mathcal{R}_k = \emptyset, j \neq k.$$

The rule d is then defined by

$$d\text{: allocate } z \text{ to } \Pi_j \text{ if and only if } z \in \mathcal{R}_j.$$

Three possible approaches to this problem are considered. We examine the simplest case, where the $f_j(x)$ are known exactly, in Section 9.2. If we have to estimate the parameters of $f_j(x)$ then we use the sample version in Section 9.3. Finally, if we do not know the distribution of the populations, Π_j, then an alternative approach, described in Section 9.4, may be used.

9.1 Maximum likelihood discriminant rule

Suppose initially that $f_1(x), \ldots, f_g(x)$ are **known** pdf's. This is the simplest case but it is an unrealistic assumption in practice unless the sample sizes are

very large.

Definition 9.1. The **maximum likelihood** (ML) discriminant rule allocates z to the population with the largest likelihood at z, i.e. it allocates z to Π_j where

$$f_j(z) = \max_{1 \le k \le g} f_k(z)$$

Example 9.1. Consider the univariate case with $g = 2$ where Π_1 is the $N(\mu_1, \sigma_1^2)$ distribution and Π_2 is the $N(\mu_2, \sigma_2^2)$ distribution. The ML discriminant rule allocates z to Π_1 if and only if

$$f_1(z) > f_2(z),$$

which is equivalent to %\begin{align}

$$\frac{1}{(2\pi\sigma_1^2)^{1/2}} \exp\left(-\frac{1}{2\sigma_1^2}(z-\mu_1)^2\right) > \frac{1}{(2\pi\sigma_2^2)^{1/2}} \exp\left(-\frac{1}{2\sigma_2^2}(z-\mu_2)^2\right).$$

%\end{align} Collecting terms together on the left hand side (LHS) gives

$$\begin{aligned}
& \frac{\sigma_2}{\sigma_1} \exp\left(-\frac{1}{2\sigma_1^2}(z-\mu_1)^2 + \frac{1}{2\sigma_2^2}(z-\mu_2)^2\right) > 1 \\
\iff \quad & \log\left(\frac{\sigma_2}{\sigma_1}\right) - \frac{1}{2\sigma_1^2}(z-\mu_1)^2 + \frac{1}{2\sigma_2^2}(z-\mu_2)^2 > 0 \\
\iff \quad & z^2\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right) + z\left(\frac{2\mu_1}{\sigma_1^2} - \frac{2\mu_2}{\sigma_2^2}\right) + \frac{\mu_2^2}{\sigma_2^2} - \frac{\mu_1^2}{\sigma_1^2} + 2\log\frac{\sigma_2}{\sigma_1} > 0.
\end{aligned}$$

Suppose, for example, that $\mu_1 = \sigma_1 = 1$ and $\mu_2 = \sigma_2 = 2$, then this reduces to the quadratic expression

$$-\frac{3}{4}z^2 + z + 2\log 2 > 0.$$

Suppose that our new observation is $z = 0$, say. Then the LHS is $2\log 2$ which is greater than zero and so we would allocate z to population 1.

For more general values of z we can solve the quadratic equation to find z such that $f_1(z) = f_2(z)$. Using the quadratic equation formula we find that

$$z = \frac{-1 \pm \sqrt{1 + 6\log 2}}{-3/2} = \frac{2}{3} \pm \frac{2}{3}\sqrt{1 + 6\log 2}.$$

Hence the solutions are $z = -0.85$ and $z = 2.18$. Our discriminant rule would then be to allocate z to Π_1 if $-0.85 < z < 2.18$ and allocate it to Π_2 otherwise. The situation is illustrated below.

FIX

Now we consider the case of g multivariate normal populations which, for simplicity, have the same covariance matrix.

Proposition 9.1. *If Π_k is the $N_p(\mu_k, \Sigma)$ population, $k = 1, \ldots, g$, then the ML discriminant rule allocates z to Π_j where j is the value of k which minimises*

$$(z - \mu_k)^\top \Sigma^{-1} (z - \mu_k).$$

Proof. The kth likelihood is

$$f_k(z) = |2\pi\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(z - \mu_k)^\top \Sigma^{-1}(z - \mu_k)\right).$$

This is maximised when the exponent is minimised, due to the minus sign in the exponent and the fact that Σ is positive definite. □

Corollary 9.1. *When $g = 2$, the rule allocates z to Π_1 if and only if*

$$a^\top (z - h) > 0,$$

where $a = \Sigma^{-1}(\mu_1 - \mu_2)$ and $h = \frac{1}{2}(\mu_1 + \mu_2)$.

Proof. First, note that

$$\begin{aligned}(z - \mu_k)^\top \Sigma^{-1}(z - \mu_k) &= z^\top \Sigma^{-1} z + \mu_k^\top \Sigma^{-1} \mu_k - \mu_k^\top \Sigma^{-1} z - z^\top \Sigma^{-1} \mu_k \\ &= z^\top \Sigma^{-1} z + \mu_k^\top \Sigma^{-1} \mu_k - 2\mu_k^\top \Sigma^{-1} z.\end{aligned}$$

From Proposition @ref{prp:nine1} we know that $f_1(z) > f_2(z)$ if and only if

$$(z - \mu_1)^\top \Sigma^{-1}(z - \mu_1) < (z - \mu_2)^\top \Sigma^{-1}(z - \mu_2).$$

Expanding both sides we find find that

$$\begin{aligned}& z^\top \Sigma^{-1} z + \mu_1^\top \Sigma^{-1} \mu_1 - 2\mu_1^\top \Sigma^{-1} z \\ &\qquad < z^\top \Sigma^{-1} z + \mu_2^\top \Sigma^{-1} \mu_2 - 2\mu_2^\top \Sigma^{-1} z \\ \iff \quad & -2\mu_1^\top \Sigma^{-1} z + 2\mu_2^\top \Sigma^{-1} z < \mu_2^\top \Sigma^{-1} \mu_2 - \mu_1^\top \Sigma^{-1} \mu_1 \\ \iff \quad & -2\left(\mu_1^\top \Sigma^{-1} z - \mu_2^\top \Sigma^{-1} z\right) < (\mu_2 - \mu_1)^\top \Sigma^{-1}(\mu_1 + \mu_2) \\ \iff \quad & (\mu_1 - \mu_2)^\top \Sigma^{-1} z > \frac{1}{2}\left[(\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_1 + \mu_2)\right] \\ \iff \quad & a^\top z > \frac{1}{2} a^\top (\mu_1 + \mu_2) \\ \iff \quad & a^\top \left(z - \frac{1}{2}(\mu_1 + \mu_2)\right) > 0 \\ \iff \quad & a^\top (z - h) > 0,\end{aligned}$$

where a and h are defined above. □

Note that the discriminant rule is *linear* in z.

Example 9.2. Consider the bivariate case ($p = 2$) with $g = 2$ groups, where Π_1 is the $N_2(\mu_1, I_2)$ distribution and Π_2 is the $N_2(\mu_2, I_2)$ distribution. Suppose $\mu_1 = \begin{pmatrix} c \\ 0 \end{pmatrix}$ and $\mu_2 = \begin{pmatrix} -c \\ 0 \end{pmatrix}$ for some constant $c > 0$. Here, $a = \Sigma^{-1}(\mu_1 - \mu_2) = \begin{pmatrix} 2c \\ 0 \end{pmatrix}$ and $h = \frac{1}{2}(\mu_1 + \mu_2) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

The ML discriminant rule allocates z to Π_1 if $a^\top(z - h) = a^\top z > 0$. If we write $z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$ then $a^\top z = 2cz_1$, which is greater than zero if $z_1 > 0$. Hence we allocate z to Π_1 if $z_1 > 0$ and allocate z to Π_2 if $z_1 \leq 0$.

FIX

Example 9.3. A slightly more complicated version of the previous example: we still assume $\mu_1 = -\mu_2$ but make no assumption about Σ. Write $a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ and $h = \frac{1}{2}(\mu_1 + \mu_2) = 0$. The ML discriminant rule allocates z to Π_1 if $a^\top(z - h) = a^\top z > 0$. If we write $z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$ then the boundary separating $\mathcal{R}_1$ and $\mathcal{R}_2$ is given by $a^\top z = \begin{pmatrix} a_1 & a_2 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = a_1 z_1 + a_2 z_2 = 0$, i.e. $z_2 = -\frac{a_1}{a_2} z_1$. This is a straight line through the origin with gradient $-a_1/a_2$.

Note that when $g > 2$, the boundaries for the ML rule will be piece-wise linear rather than linear.

9.2 The sample ML discriminant rule

To use the ML discriminant rule, above, we need to know the model parameters for each group. In reality, we often do not know these parameters but we can estimate them from "training" data. Training data typically consists of samples $x_{1,j}, \ldots, x_{n_j,j}$ known to be from population Π_j ($j = 1, \ldots, g$). Note that there are n_j observations from population Π_j.

For simplicity, we shall assume that the populations have multivariate normal distributions with different means μ_j, $j = 1, \ldots, g$, and the same covariance matrix, Σ. Let $\bar{x}_j$ and S_j be the sample mean and sample covariance matrix for the jth group. Then $\bar{x}_j$ is an unbiased estimate of μ_j and

$$S_u = \frac{1}{n - g} \sum_{k=1}^{g} n_k S_k$$

is an unbiased estimate of Σ where $n = n_1 + n_2 + \ldots + n_g$. The sample ML discriminant rule is then defined by substituting these estimates into 9.1.

Definition 9.2. If Π_k is the $N_p(\mu_k, \Sigma)$ population, $k = 1, \ldots, g$, then the sample ML discriminant rule allocates z to Π_j where \$j \$ is the value of k which

minimises

$$(z-\bar{x}_k)^\top S_u^{-1}(z-\bar{x}_k).$$

In the case $g=2$, the rule allocates z to Π_1 if and only if

$$\hat{a}^\top(z-\hat{h})>0$$

where $\hat{a}=S_u^{-1}(\bar{x}_1-\bar{x}_2)$, $\hat{h}=\frac{1}{2}(\bar{x}_1+\bar{x}_2)$ and S_u, the pooled estimate of Σ, is given by

$$S_u=\frac{1}{n_1+n_2-2}(n_1S_1+n_2S_2).$$

This is analogous to the Corollary following Proposition 9.1.

Example 9.4. Consider the G11PRB and G11STA module marks for $n_1=98$ students on G100 and $n_2=46$ students on G103. The sample means and variances for each group are given by

$$\bar{x}_1=\begin{pmatrix}60.582\\62.786\end{pmatrix} \qquad \bar{x}_2=\begin{pmatrix}64.761\\60.457\end{pmatrix}$$

$$S_1=\begin{pmatrix}201.04 & 129.56\\129.56 & 316.21\end{pmatrix} \qquad S_2=\begin{pmatrix}229.88 & 177.02\\177.02 & 354.16\end{pmatrix}$$

Hence,

$$\begin{aligned} S_u &= \frac{1}{98+46-2}\,(98S_1+46S_2)=\begin{pmatrix}213.21 & 146.76\\146.76 & 332.96\end{pmatrix},\\ \bar{x}_1-\bar{x}_2 &= \begin{pmatrix}-4.179\\2.329\end{pmatrix},\\ \hat{h} &= \frac{1}{2}(\bar{x}_1+\bar{x}_2)=\begin{pmatrix}62.671\\61.621\end{pmatrix}, \end{aligned}$$

and

$$\hat{a}=S_u^{-1}(\bar{x}_1-\bar{x}_2)=\begin{pmatrix}0.0067 & -0.0030\\-0.0030 & 0.0043\end{pmatrix}\begin{pmatrix}-4.179\\2.329\end{pmatrix}=\begin{pmatrix}-0.035\\0.022\end{pmatrix}.$$

The sample ML discriminant rule allocates a new observation\ $z=(z_1,z_2)^\top$ to Π_1 if and only if

$$\hat{a}^\top(z-\hat{h})=(-0.035 \quad 0.022)\begin{pmatrix}z_1-62.671\\z_2-61.621\end{pmatrix}>0.$$

For example, if a student on this year's course scores 80 on G11PRB and 60 on G11STA then

$$\hat{a}^\top(z-\hat{h})=(-0.035 \quad 0.022)\begin{pmatrix}80-62.671\\60-61.621\end{pmatrix}=-0.644<0,$$

and so we would allocate this student to G103. The boundary, where $\hat{a}^\top(z-\hat{h}) = 0$, is shown below.

FIX

Note that the boundary line passes half-way between the two sample means. In this example it is difficult to discriminate accurately between G100 and G103 because there is a large overlap between the two populations.

We could extend the example to include, say, students on GL11. Here the boundary between the three populations is piece-wise linear and they meet at a common point.

FIX

9.3 Fisher's linear discriminant rule

Recall that when the Π_j are $N_p(\mu_j, \ )$ populations, the ML discriminant rule is linear if $g = 2$ but not when $g > 2$. An alternative approach due to Fisher is to look for a linear discriminant function without assuming that the Π_j are multivariate normal.

Suppose we have a training sample $x_{1,j}, \ldots, x_{n_j,j}$ from Π_j $(j = 1, \ldots, g)$. Calculate the 'within' sum of squares matrix:

$$W = \sum_{j=1}^{g} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)(x_{ij} - \bar{x}_j)^\top = \sum_{j=1}^{g} n_j S_j$$

where $\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}$ is the sample mean of the jth group. Also, calculate the 'between' sum of squares matrix

$$B = \sum_{j=1}^{g} n_j (\bar{x}_j - \bar{x})(\bar{x}_j - \bar{x})^\top,$$

where $\bar{x} = \frac{1}{n} \sum_{j=1}^{g} \sum_{i=1}^{n_j} x_{ij}$ is the overall mean, and $n = \sum_{j=1}^{g} n_j$.

Fisher's Criterion is to choose a unit vector, λ, to maximise

$$\frac{\lambda^\top B \lambda}{\lambda^\top W \lambda},$$

the ratio of the `between' sum of squares to the`within' sum of squares along λ.

The function $L(z) = \lambda^\top z$ is called Fisher's linear discriminant function. Once $L(z)$ has been obtained, we allocate z to the Π_j whose discriminant score $L(\bar{x}_j)$ is closest to $L(z)$, that is, allocate z to Π_j iff

$$|\lambda^\top z - \lambda^\top \bar{x}_j| = \min_{1 \le k \le g} |\lambda^\top z - \lambda^\top \bar{x}_k|.$$

How do we find λ?

Proposition 9.2. *A vector λ that maximises*

$$\frac{\lambda^\top B\lambda}{\lambda^\top W\lambda}$$

is an eigenvector of $W^{-1}B$ corresponding to the largest eigenvalue.

Proof. Assume W is positive definite and note that W is symmetric, so we can use the spectral decomposition theorem and write $W = Q\Lambda Q^\top$.

Define $\gamma = W^{1/2}\lambda$. Then $\lambda = W^{-1/2}\gamma$ where $W^{-1/2} = Q\Lambda^{-1/2}Q^\top$ and

$$\begin{aligned}
\max_{\lambda:\,\lambda^\top\lambda=1}\left\{\frac{\lambda^\top B\lambda}{\lambda^\top W\lambda}\right\} &= \max_{\gamma:\,\gamma\neq 0}\left\{\frac{\gamma^\top W^{-1/2}BW^{-1/2}\gamma}{\gamma^\top W^{-1/2}WW^{-1/2}\gamma}\right\} \\
&= \max_{\gamma:\,\gamma\neq 0}\left\{\frac{\gamma^\top W^{-1/2}BW^{-1/2}\gamma}{\gamma^\top I_p\gamma}\right\} \\
&= \max_{\gamma:\,\gamma^\top\gamma=1}\left\{\gamma^\top W^{-1/2}BW^{-1/2}\gamma\right\}
\end{aligned}$$

This is similar to the PCA situation in §3.2 where we chose u to be the eigenvector corresponding to the largest eigenvalue of S to maximise $u^\top Su$. Hence, we choose γ to be the eigenvector corresponding to the largest eigenvalue of $W^{-1/2}BW^{-1/2}$.

If γ is an eigenvector of $W^{-1/2}BW^{-1/2}$ then, by definition,

$$W^{-1/2}BW^{-1/2}\gamma = \rho\gamma$$

where ρ is the corresponding eigenvalue. Pre-multiplying both sides by $W^{-1/2}$ gives

$$\begin{aligned}
W^{-1}B(W^{-1/2}\gamma) &= \rho W^{-1/2}\gamma \\
W^{-1}B\lambda &= \rho\lambda.
\end{aligned}$$

So, the λ we require is the unit eigenvector corresponding to the largest eigenvalue of $W^{-1}B$. □

When $g = 2$, Fisher's rule and the sample ML rule with $\Sigma_1 = \Sigma_2 = \Sigma$ turn out to be the same. Note that in the sample ML rule we assumed that the two groups are from $N_p(\mu_i, \Sigma)$ populations, but Fisher's rule makes no such assumption.

Proposition 9.3. *If $g = 2$ then Fisher's rule and the sample ML rule described in §9.3 are equivalent.*

Proof. First, note that

$$\begin{aligned}\bar{x}_1 - \bar{x} &= \bar{x}_1 - \left(\frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2}\right) = \frac{(n_1+n_2)\bar{x}_1 - n_1\bar{x}_1 - n_2\bar{x}_2}{n_1+n_2} \\ &= \frac{n_2(\bar{x}_1 - \bar{x}_2)}{n_1+n_2} = \frac{n_2 d}{n_1+n_2}\end{aligned}$$

where $d = \bar{x}_1 - \bar{x}_2$. By analogy $\bar{x}_2 - \bar{x} = \frac{n_1(-d)}{n_1+n_2}$. Therefore,

$$\begin{aligned}B &= n_1(\bar{x}_1 - \bar{x})(\bar{x}_1 - \bar{x})^\top + n_2(\bar{x}_2 - \bar{x})(\bar{x}_2 - \bar{x})^\top \\ &= \frac{n_1 n_2^2}{(n_1+n_2)^2} dd^\top + \frac{n_2 n_1^2}{(n_1+n_2)^2}(-d)(-d)^\top \\ &= \frac{n_1 n_2 (n_1+n_2)}{(n_1+n_2)^2} dd^\top = \frac{n_1 n_2}{n_1+n_2} dd^\top.\end{aligned}$$

Let $c = \frac{n_1 n_2}{n_1+n_2}$. Now λ is an eigenvector of $W^{-1}B = cW^{-1}dd^\top$. Also, the non-zero eigenvalues of $cW^{-1}dd^\top$ are the same as the non-zero eigenvalues of $cd^\top W^{-1}d$, which is scalar and so itself is the only non-zero eigenvalue. The eigenvector, λ, must then satisfy

$$cW^{-1}dd^\top\lambda = cd^\top W^{-1}d\lambda.$$

If we choose $\lambda = W^{-1}d$ then the equation is satisfied. Hence $\lambda = W^{-1}(\bar{x}_1 - \bar{x}_2)$.

Let $r = \lambda^\top z$, $s = \lambda^\top \bar{x}_1$ and $t = \lambda^\top \bar{x}_2$, then Fisher's rule allocates z to Π_1 if and only if

$$\begin{aligned}& |r-s| < |r-t| \\ \iff\ & (r-s)^2 < (r-t)^2 \\ \iff\ & r^2 - 2rs + s^2 < r^2 - 2rt + t^2 \\ \iff\ & 0 < 2r(s-t) + t^2 - s^2 \\ \iff\ & 0 < 2r(s-t) + (t-s)(t+s) \\ \iff\ & 0 < (s-t)(2r-t-s)\end{aligned}$$

Now $s-t = \lambda^\top(\bar{x}_1 - \bar{x}_2) = d^\top W^{-1}d$ which is a quadratic form and must therefore be positive, because W is assumed to be positive definite. Hence Fisher's rule allocates z to Π_1 if

$$\begin{aligned}& (2r-s-t) > 0 \\ \iff\ & r - \frac{1}{2}(s+t) > 0 \\ \iff\ & \lambda^\top\left(z - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)\right) > 0 \\ \iff\ & (\bar{x}_1 - \bar{x}_2)^\top W^{-1}\left(z - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)\right) > 0 \\ \iff\ & (\bar{x}_1 - \bar{x}_2)^\top S_u^{-1}\left(z - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)\right) > 0\end{aligned}$$

where the last line follows since $W = (n_1 + n_2 - 2)S_u$. This is equivalent to the sample ML rule for $g = 2$. □

For $g > 2$, the sample ML rule and Fisher's linear rule will not, in general, be the same. Fisher's rule is linear when $g > 2$ and is easier to implement than ML rules when there are several populations. It is often reasonable to use Fisher's rule for non-normal populations. In particular, Fisher's rule requires fewer assumptions than ML rules. However, the ML rule is 'optimal' in some sense when its assumptions are valid.

9.4 Probability of misclassification

Let p_{jk} denote the probability of allocating an observation to population Π_j, when in fact it comes from Π_k. Therefore p_{kk} is the probability of correctly classifying this observation and $1 - p_{kk}$ is the probability of misclassification.

One way of estimating p_{jk} is to consider the number of observations from the training data that are misclassified. For example, if n_k observations come from population k and n_{jk} is the number of observations from population k classified as from population j, then

$$\hat{p}_{jk} = \frac{n_{jk}}{n_k}$$

is an estimate of p_{jk}.

When $g = 2$, Π_j is $N_p(\mu_j, \Sigma)$ and we use the ML rule, we obtain an explicit expression for p_{12} and p_{21} as follows.

Recall that we allocate z to Π_1 if and only if $U = a^\top(z - h) > 0$ where $a = \Sigma^{-1}(\mu_1 - \mu_2)$ and $h = \frac{1}{2}(\mu_1 + \mu_2)$.

Suppose z is from Π_2. Then $z \sim N_p(\mu_2, \Sigma)$, so

$$\begin{aligned} E[U] &= E[a^\top(z-h)] = a^\top(E[z] - h) = a^\top(\mu_2 - h); \\ \mathrm{Var}(U) &= \mathrm{Var}(a^\top z - a^\top h) = \mathrm{Var}(a^\top z) = a^\top \Sigma a. \end{aligned}$$

Hence, when z is from Π_2, $U \sim N(a^\top(\mu_2 - h), a^\top \Sigma a)$.

Define $\Delta^2 = (\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_1 - \mu_2)$, Then

$$\begin{aligned} a^\top(\mu_2 - h) &= a^\top\left(\mu_2 - \frac{1}{2}\mu_1 - \frac{1}{2}\mu_2\right) = \frac{1}{2}a^\top(\mu_2 - \mu_1) \\ &= \frac{1}{2}(\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_2 - \mu_1) \\ &= -\frac{1}{2}(\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_1 - \mu_2) = -\frac{1}{2}\Delta^2. \end{aligned}$$

and

$$a^\top \Sigma a = (\mu_1 - \mu_2)^\top \Sigma^{-1} \Sigma \Sigma^{-1}(\mu_1 - \mu_2) = \Delta^2.$$

Hence $U \sim N(-\frac{1}{2}\Delta^2, \Delta^2)$ when z is from Π_2.

Now z is allocated to Π_1 if $U > 0$. The probability of this event when z is in fact from Π_2 is

$$\begin{aligned} p_{12} = P(U > 0) &= P\left(\frac{U - (-\Delta^2/2)}{\Delta} > \frac{0 - (-\Delta^2/2)}{\Delta}\right) \\ &= P\left(Z > \frac{\Delta^2}{2\Delta} = \frac{\Delta}{2} \middle| Z \sim N(0,1)\right) \\ &= P\left(Z < -\frac{\Delta}{2}\right) \end{aligned}$$

which can be found from statistical tables. A similar argument shows that $p_{21} = p_{12}$.

If we are using the sample ML rule then we can replace Δ^2 with

$$D^2 = (\bar{x}_1 - \bar{x}_2)^\top S_u^{-1} (\bar{x}_1 - \bar{x}_2)$$

and $\hat{p}_{12} = P(Z < -D/2)$.

Chapter 10

Cluster Analysis

A key difference between Discriminant Analysis, covered in §9, and Cluster Analysis is that in the former, a training sample is available, whereas in the latter, we do not have access to a training sample. In Machine Learning and Data Mining, Cluster Analysis is typically referred to as Unsupervised Learning (''Unsupervised" here refers to the fact that no training sample is available).

In this chapter, we shall limit ourselves to two topics within Cluster Analysis:

- likelihood-based clustering which, as we will see, includes the widely-used method K-means clustering as a special case; and
- hierarchical clustering methods.

10.1 Likelihood-based clustering

Suppose we have a sample of n random vectors $x_1, \ldots, x_n$, assumed independent. Suppose that each x_i comes from one of g sub-populations, where the jth sub-population has probability density function $f_j(x; \theta_j)$, $j = 1, \ldots, g$. For simplicity it is assumed in this section that g is given. However, in many applications, g is unknown, and we do not assume that g is known later in this chapter. The key point is that we do not know which sub-population each x_i comes from, i.e. we do not know how to allocate the observations to sub-populations.

The goal of cluster analysis: to allocate each of n observational vectors to one of g clusters, in such a way that observation vectors within a cluster tend to be more similar, in a suitable sense, than observations in different clusters.

In principle we can estimate the optimal allocation, along with the (usually) unknown parameter vectors $\theta_1, \ldots, \theta_g$, by maximum likelihood, as we shall see.

It will be convenient to introduce two equivalent ways to describe an arbitrary allocation of each x_i to one of the g clusters. Write $\delta = (\delta_1, \ldots, \delta_n)^\top$. Then

consider the equivalence

$$\delta_i = j \iff x_i \in \mathcal{C}_j, \qquad j = 1, \ldots, g, \tag{10.1}$$

where

$$\bigcup_{j=1}^{g} \mathcal{C}_j = \{x_1, \ldots, x_n\}, \qquad \text{and} \qquad \mathcal{C}_j \cap \mathcal{C}_k = \emptyset, \qquad j \neq k.$$

Then the likelihood for $\theta_1, \ldots, \theta_g$ and the allocation δ may be written

$$L(\theta_1, \ldots, \theta_g; \delta) = \left\{ \prod_{x \in \mathcal{C}_1} f(x; \theta_1) \right\} \cdots \left\{ \prod_{x \in \mathcal{C}_g} f(x; \theta_g) \right\}.$$

The sets $\mathcal{C}_1, \ldots, \mathcal{C}_g$ are called **clusters**.

Let $\hat{\theta}_1, \ldots, \hat{\theta}_g$ and $\hat{\delta}$ denote the maximum likelihood estimators of $\theta_1, \ldots, \theta_g$ and the unknown allocation δ. Also, let $\hat{\mathcal{C}}_1, \ldots, \hat{\mathcal{C}}_g$ denote the maximum likelihood clusters, which are determined by $\hat{\delta}$ via (10.1).

Then we have the following result.

Proposition 10.1. *If $x \in \hat{\mathcal{C}}_j$ then*

$$\sup_{1 \leq k \leq g} f(x; \hat{\theta}_k) = f(x; \hat{\theta}_j).$$

Proof. If we move an x from $\hat{\mathcal{C}}_j$ to $\hat{\mathcal{C}}_k$, then the likelihood changes to

$$L(\hat{\theta}_1, \ldots, \hat{\theta}_g; \delta) f(x; \hat{\theta}_k) / f(x; \hat{\theta}_j).$$

But by definition of $L(\hat{\theta}_1, \ldots, \hat{\theta}_g; \hat{\delta})$,

$$L(\hat{\theta}_1, \ldots, \hat{\theta}_g; \hat{\delta}) f(x; \hat{\theta}_k) / f(x; \hat{\theta}_j) \leq L(\hat{\theta}_1, \ldots, \hat{\theta}_g; \hat{\delta}),$$

from which we conclude that Proposition 10.1 holds. □

Note that this result is closely related to the sample ML discriminant rule considered in §9.3.

We now consider the case where the sub-populations are multivariate Gaussian, i.e. $f(x; \theta_j)$ is the density of $N_p(\mu_j, \Sigma_j)$ for $j = 1, \ldots, g$.

In the general case, when the mean vector and covariance matrix are different for each sub-population, we know how to maximise the likelihood when the allocation δ is given. Here, θ_j consists of μ_j and Σ_j for each $j = 1, \ldots, g$, and from §8.4, we know that the maximised log-likelihood for cluster j is

$$\ell(\hat{\mu}_j[\delta], \hat{\Sigma}_j[\delta]) = -\frac{n_j[\delta]}{2} \log(|\hat{\Sigma}_j[\delta]|) - \frac{n_j[\delta]}{2} p(1 + \log 2\pi),$$

where $n_j[\delta]$ is the number of elements in cluster j, $\mathcal{C}_j[\delta]$, for the given allocation δ. Similarly, $\widehat{\mu}_j[\delta]$ and $\widehat{\Sigma}_j[\delta]$ are the MLEs of μ_j and Σ_j for the given allocation δ, i.e. the sample mean and covariance matrix

$$\widehat{\mu}_j[\delta] = \frac{1}{n_j[\delta]} \sum_{x \in \mathcal{C}_j[\delta]} x = \bar{x}_j[\delta]$$

and

$$\widehat{\Sigma}_j[\delta] = \frac{1}{n_j[\delta]} \sum_{x \in \mathcal{C}_j[\delta]} (x - \bar{x}_j[\delta])(x - \bar{x}_j[\delta])^\top.$$

It follows that the MLE of δ is the choice of δ which maximises the log-likelihood

$$-\frac{1}{2} \sum_{j=1}^{g} n_j[\delta] \log(|\widehat{\Sigma}_j[\delta]|) \; + \; \text{constant}$$

over δ.

Under the assumption that the population covariance matrices are the same, i.e. $\Sigma_1 = \cdots = \Sigma_g$, the maximised log-likelihood for a given allocation δ is given by

$$-\frac{n}{2} \log(|W[\delta]|) \; + \; \text{constant}, \tag{10.2}$$

where $n = \sum_{j=1}^{g} n_j$ and

$$W[\delta] = \sum_{j=1}^{g} \sum_{x \in \mathcal{C}_j[\boldsymbol{\delta}]} (x - \widehat{\mu}_j[\delta])(x - \widehat{\mu}_j[\delta])^\top$$

is the ''within" sum of squares and products matrix for the given allocation. So the maximum likelihood allocation $\widehat{\delta}$ is the δ which maximises (10.2).

The final case we consider is where $\Sigma_1 = \cdots = \Sigma_g = \sigma^2 I_p$, i.e. the common covariance matrix is a scalar multiple of the $p \times p$ identity matrix. This is a version of the so-called k-means clustering approach, and here the maximum likelihood allocation $\widehat{\delta}$ is obtained by the δ which minimises

$$\sum_{j=1}^{g} \sum_{x \in \mathcal{C}_j} ||x - \bar{x}_j||^2.$$

Although clustering based on the likelihood function has a certain intuitive appeal, in most practical situation it is not feasible to find the global maximum due to the computational explosion in the number of possible allocations when n is even of moderate size, e.g. $n = 100$. A further problem is that there may be a large number of local maxima of the likelihood function. However, despite these challenges, likelihood-based clustering, such as k-means clustering, is widely used and can lead to useful, even if sub-optimal, solutions to the clustering problem.

10.2 Hierarchical clustering methods

The adjective *hierarchical* applies to clustering methods which have the following property: the arrangement of the experimental units into g clusters and into $g+1$ clusters have the properties that $g-1$ of the clusters are identical, and the remaining single cluster in the g clusters is split into 2 clusters in the $g+1$ clusters.

A hierarchical clustering method is usually of one of two types:

1. an *agglomerative* clustering method progressively combines clusters, usually starting with n singleton clusters; and
2. a *divisive* clustering method progressively splits, or divides, clusters, usually starting with a single cluster containing n elements.

Many clustering methods a based on an $n \times n$ matrix of inter-point distances $D = (d_{ij})_{i,j=1}^{n}$ of the type we considered in §5. However, the goal here is somewhat different to that in Multidimensional Scaling.

Given D, two common clustering procedures are:

1. the *single linkage* method, sometimes called the *nearest neighbour* method; and
2. the *complete linkage* method, sometimes called the *furthest neigbour* method.

We now explain in more detail how these methods are applied. First, we consider single linkage. It is assumed that the set of distances is ordered, so that

$$d_{a_1,b_1} \leq d_{a_2,b_2} \leq \cdots \leq d_{a_{N-1},b_{N-1}} \leq d_{a_N,b_N}, \tag{10.3}$$

where $N = n(n-1)/2$, and we adopt the following conventions: (i) $a_t < b_t$; and (ii) to break a tie such as $d_{a_t,b_t} = d_{a_{t+1},b_{t+1}}$, we write

$$\cdots \leq d_{a_t,b_t} \leq d_{a_{t+1},b_{t+1}} \leq \cdots$$

if $a_t < a_{t+1}$, or $a_t = a_{t+1}$ and $b_t < b_{t+1}$; and we write

$$\cdots \leq d_{a_{t+1},b_{t+1}} \leq d_{a_t,b_t} \leq \cdots$$

if $a_{t+1} < a_t$, or $a_{t+1} = a_t$ and $b_{t+1} < b_t$.

1. Start with singleton clusters $\mathcal{C}_1, \ldots, \mathcal{C}_n$, i.e. $\mathcal{C}_i = \{i\}$.
2. Combine experimental units a_1 and b_1 into a single new cluster, so that we now have one doubleton cluster and $n-2$ singleton clusters.
3. Next, consider a_2 and b_2, where d_{a_2,b_2} is the second smaller distance. If both $a_2 \notin \{a_1, b_1\}$ and $b_2 \notin \{a_1, b_1\}$, then we combine a_2 and b_2 into a second doubleton cluster, $\{a_2, b_2\}$, leading to $n-2$ clusters altogether (2 doubleton clusters and $n-4$ singleton clusters). If, on the other hand,

$a_2 \in \{a_1, b_1\}$, then necessarily $b_2 \notin \{a_1, b_1\}$, and so we form the trippleton cluster $\{a_1, b_1, b_2\}$; while if $b_2 \in \{a_1, b_1\}$, then necessarily $a_2 \notin \{a_1, b_1\}$, and we form the trippleton cluster $\{a_1, a_2, b_1\}$. Either way, in the latter two cases, we are left with $n-2$ clusters altogether (one trippleton and $n-3$ singleton clusters).

4. We continue this process as we pass through the N inter-point distances. However, sometimes we may wish to terminate this process at some threshold, T; i.e. we stop the process at the smallest t such that $d_{a_t,b_t} > T$.

In the single linkage approach, in effect we are defining the distance between two clusters $\mathcal{C}_u$ and $\mathcal{C}_v$ by

$$d_S^*(\mathcal{C}_u, \mathcal{C}_v) = \min_{i \in \mathcal{C}_u,\, j \in \mathcal{C}_v} d_{ij}.$$

In contrast, with the complete linkage method, in effect we are defining the distance between two clusters $\mathcal{C}_u$ and $\mathcal{C}_v$ by

$$d_C^*(\mathcal{C}_u, \mathcal{C}_v) = \max_{i \in \mathcal{C}_u,\, j \in \mathcal{C}_v} d_{ij}.$$

A convenient graphical way to present the output from either a single linkage procedure or a complete linkage procedure is to plot a **dendrogram**.

Example 10.1. The following distance matrix was based on the relative gene frequencies for the four blood-group systems A_1, A_2, B and O for large samples from four human populations: (1) Inuit, (2) African, (3) English and (4) Korean. The inter-point distances were determined using the Mahalanobis distance between μ_i and μ_j defined by

$$d_{ij} = \sqrt{(\mu_i - \mu_j)^\top \Sigma^{-1} (\mu_i - \mu_j)}.$$

The matrix of inter-point distances is given by:

$$\begin{pmatrix} 0 & 23.26 & 16.34 & 16.87 \\ & 0 & 9.85 & 20.43 \\ & & 0 & 19.60 \\ & & & 0 \end{pmatrix}.$$

Hypothesis of interest: *there there is a natural clustering*

$$\{\text{African} = 2, \text{English} = 3\} \qquad \text{and} \qquad \{\text{Inuit} = 1, \text{Korean} = 4\}.$$

Let us first of all look at single linkage. The ordering of the distances is given by

$$d_{23} < d_{13} < d_{14} < d_{34} < d_{24} < d_{12}.$$

Single Linkage

- At Stage 0 the clusters are $\{1\}$, $\{2\}$, $\{3\}$ and $\{4\}$.
- At Stage 1 the clusters are $\{1\}$, $\{2,3\}$ and $\{4\}$.
- At Stage 2 the clusters are $\{1,2,3\}$ and $\{4\}$.
- At Stage 3 we have a single cluster $\{1,2,3,4\}$.

Now let us look at complete linkage. Here, Stage $0'$ and Stage $1'$ are the same at Stage 0 and Stage 1 of single linkage, but Stage $2'$ is different.

Complete Linkage - At Stage $0'$ the clusters are $\{1\}$, $\{2\}$, $\{3\}$ and $\{4\}$. - At Stage $1'$ the clusters are $\{1\}$, $\{2,3\}$ and $\{4\}$. - At Stage $2'$ the clusters are $\{1,4\}$ and $\{2,3\}$. - At Stage $3'$ we have a single cluster $\{1,2,3,4\}$.

The reason that we combine $\{1,4\}$ at Stage $2'$ is because

$$d_{1,4} = 16.87 < d_C^*(\{1\},\{2,3\}) = \max\{d_{12}, d_{13}\} = 23.26$$

and

$$d_{1,4} = 16.87 < d_C^*(\{4\},\{2,3\}) = \max\{d_{24}, d_{34}\} = 20.43.$$

So with complete linkage we should combine $\{1\}$ and $\{4\}$ before combining either $\{1\}$ or $\{4\}$ with $\{2,3\}$.

In this example, single linkage and complete linkage lead to different conclusions: complete linkage supports the hypothesis of interest, while single linkage does not. Some people have argued that single linkage is more appropriate here.

FIX THESE PLOTS

Sketch of Dendrogram: Single Linkage

Sketch of Dendrogram: Complete Linkage

10.3 Further Points

In this chapter we have focused on just two approaches to Cluster Analysis: likelihood-based methods, especially those based on the multivariate Gaussian model; and hierarchical clustering methods, especially single linkage and complete linkage approaches.

There are many algorithms for performing Cluster Analysis, but there is no generally accepted ''best" method. Moreover, different algorithms (or even the same algorithm with a different initialisation) do not necessarily produce the same results on a given dataset, and there is often a fairly large subjective element in the assessment of any particular method.

One way to test a clustering algorithm is to apply it on data with a known group structure. Experience suggests that this will only produce good results when the groups are very distinct. When, on the other hand, there is a lot of overlap between groups, clustering algorithms are not likely to perform particularly well.

However, despite these cautionary remarks, clustering algorithms are often useful in practice, but it is an area where usually the most one can hope for is to find a good, but sub-optimal, solution.