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.

FIX FIX CHAPTER REFS

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.

ADD comment on high dimensional SPACE IS BIG!

TO DO

Miscellaneous topics Centering matrix, ellipses, lines. Is this useful and where should it go?

PART I: Prerequisites

Much of modern multivariate statistics (and machine learning) relies upon linear algebra. Consequently, we will spend some time reminding you of the basics of linear algebra (vector spaces, matrices etc), and introducing a few additional concepts that you may not have seen before. It is worth spending time familiarizing yourself with these ideas, as we will rely heavily upon this material in later chapters.

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, most of which will be familiar to you. In particular, we will introduce vector spaces and the concept of a basis for a vector space, discuss the column, row and null space of matrices, and discuss inner product spaces and the concept of projections.

In Chapter 3 we recap the eigen or spectral decomposition of square symmetric matrices, and introduce the singular value decomposition (SVD) which generalises the concept of eigenvalues for non-square matrices. We will rely upon this material in later chapters.

Chapter 1

Statistical Preliminaries

In this chapter we will define some notation, and recap some basic statistical properties and results.

NOT YET FILMED VIDEOS.

1.1 Multivariate data

We will think of datasets as consisting of measurements of p different **variables** for n different **cases/subjects**. We organise the data into a $n \times p$ **data matrix**.

Multivariate analysis (MVA) refers data analysis methods where there are two or more **response** variables for each case (you are familiar with situations where there is more than one explanatory variable, e.g., multiple linear regression).

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

$$\mathbf{X} = \begin{bmatrix} x_1^\top \\ x_2^\top \\ \cdot\cdot \\ \cdot\cdot \\ \cdot\cdot \\ x_n^\top \end{bmatrix}.$$

In words: the *rows* of $\mathbf{X}$ are $x_1^\top, \ldots, x_n^\top$.

We will often consider $\mathbf{X}^\top$

$$\mathbf{X}^\top = [x_1, \ldots, x_n]$$

i.e., the *columns* of $X^\top$ are $x_1, \ldots, x_n$.

In this setup, we think of $x_1, \ldots, x_n \in \mathbb{R}^p$ as 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. The `iris` dataset in R contains data on the length and width petal and sepal

Example 1.2. 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 $n = 4$ cases (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

The data vector for the USA is

$$x_1 = (1, 2, 4, 3)$$

Example 1.3. 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}

Example 1.4. The MNIST dataset is a collection of handwritten digits that is widely used in statistics and machine learning to test algorithms. It contains 60,000 images of hand-written digits.

Each digit has been converted to a grid of 28×28 pixels, with a grayscale intensity level specified for each pixel. When we store these on a computer, we flatten each grid to a vector of length 784

```
##  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

“‘

In MVA we attempt to answer questions such as:

- How can we visualise the data?
- What is the joint distribution of marks?
- Can we simplify the data? For example, we rank football teams using $3W + D$ and we rank students by their average module mark. 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

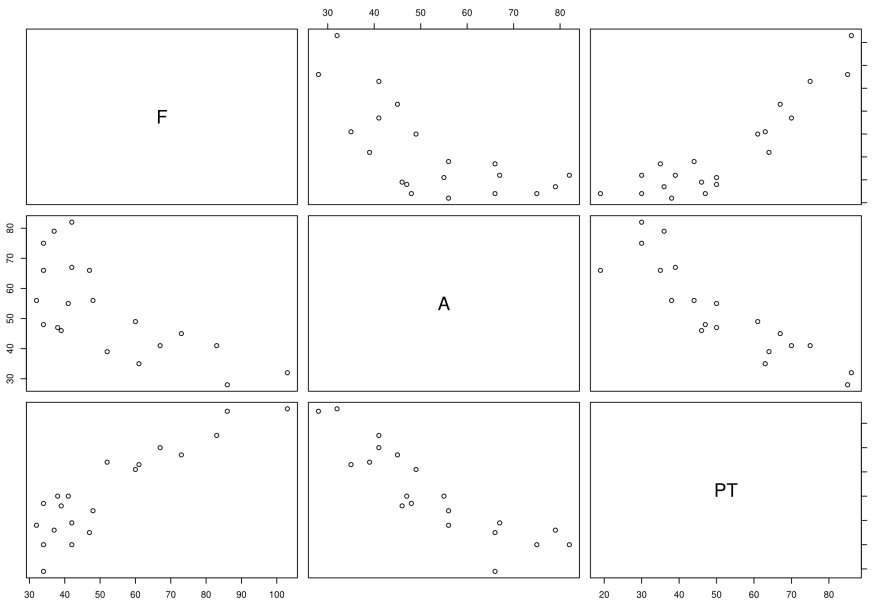
but this ignores possible dependencies between the variables which we must represent to draw valid conclusions.

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 **??**.

1.1.1 Graphical techniques

It is always a good idea to plot your data before analysing it. With multivariate data, this is not always simple, and many different approaches have been proposed.

When $p = 1$ or $p = 2$ we can simply draw histograms and scatter plots (respectively) to view the distribution. 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. Note that it is possible to miss key relationships when looking at *marginals* plots such as these: for example, relationships between three variables will not be visible.



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.

1.2 Summary statistics

In univariate statistics we define the sample mean and sample variance of samples $x_1, \ldots, x_n$ to be

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

and for two samples, $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, we define the sample covariance to be

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

We now define analogous multivariate quantities:

Definition 1.1. For a sample of n points, each containing p variables,

$x_1, x_2, \ldots, x_n \in \mathbb{R}^p$, the **sample mean** and **sample covariance matrix** are

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

where $x_i \in \mathbb{R}^p$ denotes the p variables observed on the ith subject.

Note that

- $\bar{x} \in \mathbb{R}^p$. The jth entry in $\bar{x}$ is simply the (univariate) sample mean of the jth variable.
- $S \in \mathbb{R}^{p\times p}$. Note that the ij^{th} entry of S is s_{ij}, the sample covariance betweenb variable i and variable j. The i^{th} diagonal element is the (univariate) sample variance of the ith variable.

- S is symmetric since $s_{ij} = s_{ji}$.
- an alternative formula for S is

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

- We have divided by n rather than $n-1$ here, which gives the maximum likelihood estimator of the variance, rather than the unbiased variance estimator that is often used.

Definition 1.2. The **sample correlation matrix**, R, is the matrix with ij^{th} entry r_{ij} equal to the sample correlation between variables i and j, that is

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}.$$

Note that

- If $D = \text{diag}(\sqrt{s_{11}}, \ldots, \sqrt{s_{pp}})$, then

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

- R is symmetric
- the diagonal entries of R are exactly 1 (each variable is perfectly correlated with itself)
- $|r_{ij}| \leq 1$ for all i, j

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

Definition 1.3. The **total variation** in a 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 other words, it is the sum of the univariate variances of each of the p variables.

Example 1.5. 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

As an exercise, calculate the sample mean, sample covariance, sample correlation and total variation by hand. Check

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 %>% knitr::kable(booktabs = TRUE) %>% kable_styling(full_width = F)
```

Student	P	S
A	41	63
B	72	82
C	46	38
D	77	57
E	59	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) %>% cov*4/5
```

```
##       P     S
## P 197.2  92.8
## S  92.8 297.2
```

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

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

```
Ex1 %>% select_if(is.numeric) %>% cov %>% diag %>% sum*4/5
```

```
## [1] 494.4
```

Note that by default R uses $n-1$ in the denominator, whereas we used n in our calculation, hence the multiple of \$4/5=(n-1)/n introduced in the covariance calculations above.

We will be using the `dplyr` R package to perform basic data manipulation in R. If you are unfamiliar with `dplyr`, you can read about it at https://dplyr.tidyverse.org/. The pipe command `%>%` is particularly useful for chaining together multiple commands.

1.3 Random Vectors and Matrices

Definition 1.4. The **population mean vector** of the random vector x is

$$\mu = \mathbb{E}(x).$$

The **population covariance matrix** of x is

$$\Sigma = \mathbb{V}\,\mathrm{ar}(x) = \mathbb{E}\left((x - \mathbb{E}(x))(x - \mathbb{E}(x))^\top\right).$$

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

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

Let A denote a $q \times p$ constant matrix, and let b a constant vector of size $q \times 1$. Expectation is a linear operator in the sense that

$$\mathbb{E}(Ax + b) = A\mathbb{E}(x) + b = A\mu + b.$$

The following properties follow:

- $\mathbb{V}\,\mathrm{ar}(x) = \mathbb{E}(xx^\top) - \mu\mu^\top$.
- $\mathbb{V}\,\mathrm{ar}(Ax + b) = A\Sigma A^\top$
- $\mathbb{C}\,\mathrm{ov}(x, y) = \mathbb{E}(xy^\top) - \mathbb{E}(x)\mathbb{E}(y)^\top$.
- $\mathbb{C}\,\mathrm{ov}(x, x) = \Sigma$.
- $\mathbb{C}\,\mathrm{ov}(x, y) = \mathbb{C}\,\mathrm{ov}(y, x)^\top$.
- $\mathbb{C}\,\mathrm{ov}(Ax, By) = A\mathbb{C}\,\mathrm{ov}(x, y)B^\top$
- If $p = q$ then

$$\mathbb{V}\,\mathrm{ar}(x + y) = \mathbb{V}\,\mathrm{ar}(x) + \mathbb{V}\,\mathrm{ar}(y) + \mathbb{C}\,\mathrm{ov}(x, y) + \mathbb{C}\,\mathrm{ov}(y, x).$$

Finally, note that if x and y are independent (in which case I will write $x \perp\!\!\!\perp y$) then $\mathbb{C}\,\mathrm{ov}(x, y) = \mathbf{0}_{p,q}$, i.e., a $p \times q$ matrix of zeros.

Chapter 2

Review of linear algebra

Modern statistics and machine learning rely heavily upon linear algebra, nowhere more so than in multivariate statistics. In the first part of this chapter (sections 2.1 and 2.2) we review some concepts from linear algebra that will be needed throughout the module, including vector spaces, row and column spaces, the rank of a matrix, etc. Hopefully most of this will be familiar to you.

We then cover some basic details on inner-product or normed spaces in 2.3, which are vector spaces equipped with a concept of distance and angle.

Section 3 is perhaps the most important section. Here we provide a reminder about eigenvalues and the spectral decomposition of square symmetric matrices, before introducing the singular value decomposition (SVD) in Section 3.4. The SVD is one of the most important concepts in this module, and is the key linear algebra technique behind many of the methods we will study. Finally, in Section 2.4 we will cover some miscellaneous topics that will be needed in later chapters.

I do not provide proofs of all the results stated in this chapter, but instead prove a small selection which I think it is useful to see. For a complete treatment of the linear algebra needed for this module, see the excellent book "Linear algebra and learning from data" by Gilbert Strang.

I have recorded videos on some (but not all) of the topics in these notes:

- Vector spaces
- Matrices
- Inner product spaces
- Orthogonal matrices
- Projection matrices

NOT DONE A VIDEO ON CENTERING MATRIX, OR ELLIPSES, or VECTOR DIFFERENTIATION.

2.1 Basics

In this section, we recap some basic definitions and notation. Hopefully this material will largely be familiar to you.

2.1.1 Notation

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

$$\begin{aligned}
\mathbf{A} = \overset{n\times p}{\mathbf{A}} &= \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{bmatrix} \\
&= [a_{ij} : i = 1, \dots, m; j = 1, \dots, n] \\
&= (a_{ij}) \\
&= \begin{bmatrix} a_1^\top \\ \vdots \\ a_n^\top \end{bmatrix}
\end{aligned}$$

where the a_{ij} are the individual entries, and $a_i^\top = (a_{i1}, a_{i2}, \dots, a_{ip})$ is the i^{th} row.

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 $\text{diag}\{a_1, \dots, a_n\}$.

$$\mathbf{I}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \text{diag}\{1, 2, 3\} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

2.1.2 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 matrices then $AB = \overset{n\times q}{\mathbf{C}} = [c_{ij}]$ where
$$c_{ij} = \sum_{k=1}^{p} a_{ik}b_{kj}, \qquad i = 1, \dots, n, \qquad j = 1, \dots, q.$$

4. *Matrix Transpose.* If $\overset{m\times n}{A} = [a_{ij} : i = 1, \dots, m; j = 1, \dots, n]$, then the transpose of A, written $A^\top$, is given by the $n \times m$ matrix
$$A^\top = [a_{ji} : j = 1, \dots, n; i = 1, \dots, m].$$
Note from the definitions that $(AB)^\top = \mathbf{B}^\top\mathbf{A}^\top$.

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 $\mathbf{A}B = BA = \mathbf{I}_n$. 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 $(m \times n)$,*
$$tr(AB) = tr(BA).$$

7. The *determinant* of a square matrix $\overset{n\times n}{\mathbf{A}}$ is defined as
$$\det(\mathbf{A}) = \sum(-1)^{|\tau|} a_{1\tau(1)} \dots a_{n\tau(n)}$$
where the summation is taken over all permutations τ of $\{1, 2, \dots, n\}$, 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 2.1. *Matrix $\overset{n\times n}{\mathbf{A}}$ is invertible if and only if* $\det(A) \neq 0$. *If A^{-1} exists then*
$$\det(A) = \frac{1}{\det(A^{-1})}$$

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

2.1.3 Special matrices

Definition 2.1. An $n \times n$ matrix A is symmetric if

$$A = A^\top.$$

An $n \times n$ symmetric matrix A is **positive-definite** if

$$x^\top Ax > 0 \text{ for all } x \in \mathbb{R}^n, x \neq 0$$

and is **positive semi-definite** if

$$x^\top Ax \geq 0 \text{ for all } x \in \mathbb{R}^n.$$

A is **idempotent** if $A^2 = A$.

2.1.4 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.

$$\frac{\partial f}{\partial x}(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ .. \\ .. \\ .. \\ \frac{\partial f}{\partial x_p}(x) \end{bmatrix} \tag{2.1}$$

The following examples can be worked out directly from the definition (2.1), 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 symmetric constant $p \times p$ 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 2.2 and g is as in Example 2.3 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

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

Definition 2.2. The **Hessian** matrix of f is the $p \times p$ matrix of second derivatives.

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

The nature of a stationary point is determined by the Hessian

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 and negative eigenvalues at x then the stationary point will be a *saddle point*.

2.2 Vector spaces

It will be useful to talk about **vector spaces**. These are sets of vectors that can be added together, or multiplied by a scalar. You should be familiar with these from your undergraduate degree. We don't provide a formal definition here, but you can think of a real vector space V as a set of vectors such that for any $v_1, v_2 \in V$ and $\alpha_1, \alpha_2 \in \mathbb{R}$, we have

$$\alpha_1 v_1 + \alpha_2 v_2 \in V$$

i.e., vector spaces are closed under addition and scalar multiplication.

Example 2.5. Euclidean space in p dimensions, $\mathbb{R}^p$, is a vector space. If we add any two vectors in $\mathbb{R}^p$, or multiply a vector by a real scalar, then the resulting vector also lies in $\mathbb{R}^p$.

A subset $U \subset V$ of a vector space V is called a vector **subspace** if U is also a vector space.

Example 2.6. Let $V = \mathbb{R}^2$. Then the sets

$$U_1 = \left\{ \begin{pmatrix} a \\ 0 \end{pmatrix} : a \in \mathbb{R} \right\}, \text{ and } \quad U_2 = \left\{ a \begin{pmatrix} 1 \\ 1 \end{pmatrix} : a \in \mathbb{R} \right\}$$

are both subspaces of V.

2.2.1 Linear independence

Definition 2.3. 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**.

Definition 2.4. Given a set of vectors $S = \{s_1, \ldots, s_n\}$, the **span** of S is the smallest vector space containing S or equivalently, is the set of all linear combinations of vectors from S

$$\text{span}(S) = \left\{ \sum_{i=1}^{k} \alpha_i s_i \mid k \in \mathbb{N}, \alpha_i \in \mathbb{R}, s_i \in S \right\}$$

Definition 2.5. A **basis** of a vector space V is a set of linearly independent vectors in V that span V.

Example 2.7. Consider $V = \mathbb{R}^2$. Then the following are both bases for V:

$$B_1 = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\}$$

$$B_2 = \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right\}$$

Definition 2.6. The **dimension** of a vector space is the number of vectors in its basis.

2.2.2 Row and column spaces

We can think about the matrix-vector multiplication Ax in two ways. The usual way is as the inner product between the rows of A and x.

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 + 2x_2 \\ 3x_1 + 4x_2 \\ 5x_1 + 6x_2 \end{pmatrix}$$

But a better way to think of Ax is as a linear combination of the columns of A.

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1 \begin{pmatrix} 1 \\ 3 \\ 5 \end{pmatrix} + x_2 \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix}$$

Definition 2.7. The **column space** of a $n \times p$ matrix A is the set of all linear combinations of the columns of A:

$$\mathcal{C}(A) = \{Ax : x \in \mathbb{R}^p\} \subset \mathbb{R}^n$$

For

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}$$

we can see that the column space is a 2-dimensional plane in $\mathbb{R}^3$. The matrix B has the same column space as A

$$B = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 7 & 10 \\ 5 & 6 & 11 & 16 \end{pmatrix}$$

The number of linearly independent columns of A is called the **column rank** of A, and is equal to the dimension of the column space of $\mathcal{C}(A)$. The **column rank** of A and B is 2.

The **row space** of A is defined to be the column space of $A^\top$, and the **row rank** is the number of linearly independent rows of A.

Theorem 2.1. *The **row rank** of of a matrix equals the **column rank**.*

Thus we can simply refer to the **rank** of the matrix.

Proof. The proof of this theorem is very simple. Let C be an $n \times r$ matrix (where $r = \text{rank}(A)$) with columns chosen to be a set of r linearly independent columns from A. Then we know each column of A can be written as a linear combination of the columns of C, i.e.

$$A = CR.$$

The dimension of R must be $r \times p$. But now we can see that the rows of A are formed by a linear combination of the rows of R. Thus the row rank of A is at most r (=the column rank of A). This holds for any matrix, so is true for $A^\top$: namely row-rank$(A^\top) \leq$ column-rank$(A^\top)$. But the row space of $A^\top$ equals $\mathcal{C}(A)$, thus proving the theorem! □

Corollary 2.1. *The rank of an $n \times p$ matrix is at most* $\min(n, p)$.

Example 2.8.

$$B = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 & 2 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

Example 2.9.

$$D = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}$$

So the rank of D is 1.

2.2.3 Linear transformations

We can view an $n \times p$ matrix A as a linear map between two vector spaces:

$$\begin{aligned} A : \ \mathbb{R}^p &\to \mathbb{R}^n \\ x &\mapsto Ax \end{aligned}$$

The **image** of A is precisely the column space of A:

$$\mathrm{Im}(A) = \{Ax : x \in \mathbb{R}^p\} = \mathcal{C}(A) \subset \mathbb{R}^n$$

The **kernel** of A is the set of vectors mapped to zero:

$$\mathrm{Ker}(A) = \{x : Ax = 0\} \subset \mathbb{R}^p$$

and is sometimes called the **null-space** of A and denoted $\mathcal{N}(A)$.

Theorem 2.2. *The **rank-nullity** theorem says if V and W are vector spaces, and $A : V \to W$ is a linear map, then*

$$\dim \mathrm{Im}(A) + \dim \mathrm{Ker}(A) = \dim V$$

If we're thinking about matrices, then $\dim \mathcal{C}(A) + \dim \mathcal{N}(A) = p$, or equivalently that
$\mathrm{rank}(A) + \dim \mathcal{N}(A) = p$.

We've already said that the row space of A is $\mathcal{C}(A^\top)$. The left-null space is $\{x \in \mathbb{R}^n : x^\top A = 0\}$ or equivalently $\{x \in \mathbb{R}^n : A^\top x = 0\} = \mathcal{N}(A^\top)$. And so by the rank-nullity theorem we must have

$$n = \dim \mathcal{C}(A^\top) + \dim \mathcal{N}(A^\top) = \mathrm{rank}(A) + \dim \mathrm{Ker}(A^\top).$$

Example 2.10. Consider again the matrix $D : \mathbb{R}^3 \to \mathbb{R}^2$

$$D = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}$$

We have already seen that

$$\mathcal{C}(D) = \text{span}\left\{\begin{pmatrix} 1 \\ 2 \end{pmatrix}\right\}$$

and so $\dim \mathcal{C}(D) = \text{rank}(D) = 1$. The kernel, or null-space, of D is the set of vectors for which $Dx = 0$, i.e.,

$$x_1 + 2x_2 + 3x_3 = 0$$

This is a single equation with three unknowns, and so there must be a plane of solutions. We need two linearly independent vectors in this plane to describe it. Convince yourself that

$$\mathcal{N}(D) = \text{span}\left\{\begin{pmatrix} 0 \\ 3 \\ -2 \end{pmatrix}, \begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix}\right\}$$

So we have

$$\dim \mathcal{C}(D) + \dim \mathcal{N}(D) = 1 + 2 = 3$$

as required by the rank-nullity theorem.

If we consider $D^\top$, we already know $\dim \mathcal{C}(D) = 1$ (as row-rank=column rank), and the rank-nullity theorem tells us that the dimension of the null space of $D^\top$ must be $2 - 1 = 1$. This is easy to confirm as $D^\top x = 0$ implies

$$x_1 + 2x_2 = 0$$

which is a line in $\mathbb{R}^2$

$$\mathcal{N}(D^\top) = \text{span}\left\{\begin{pmatrix} -2 \\ 1 \end{pmatrix}\right\}$$

Question: When does a square matrix A have an inverse?

- Precisely when the kernel of A contains only the zero vector, i.e., has dimension 0. In this case the column space of A is the original space, and A is surjective and so must have an inverse. A simpler way to determine if A has an inverse is to consider its determinant.

Question: Suppose we are given a $n \times p$ matrix A, and a n-vector y. When does

$$Ax = y$$

have a solution?

- When y is in the column space of A,

$$y \in \mathcal{C}(A)$$

Question: When is the answer unique?

- Suppose x and x' are both solutions with $x \neq x'$. We can write $x' = x + u$ for some vector u and note that

$$y = Ax' = Ax + Au = y + Au$$

and so $Au = 0$, i.e., $u \in \mathcal{N}(A)$. So there are multiple solutions when the null-space of A contains more than the zero vector. If the dimension of $\mathcal{N}(A)$ is one, there is a line of solutions. If the dimension is two, there is a plane of solutions, etc.

2.3 Inner product spaces

2.3.1 Distances, and angles

Vector spaces are not particularly interesting from a statistical point of view until we equip them with a sense of geometry, i.e. distance and angle.

Definition 2.8. A real **inner product space** $(V, \langle \cdot, \cdot \rangle)$ is a real vector space V equipped with a map

$$\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$$

such that

1. $\langle \cdot, \cdot \rangle$ is a linear map in both arguments:

$$\langle \alpha v_1 + \beta v_2, u \rangle = \alpha \langle v_1, u \rangle + \beta \langle v_2, u \rangle$$

for all $v_1, v_2, u \in V$ and $\alpha, \beta \in \mathbb{R}$. 2. $\langle \cdot, \cdot \rangle$ is symmetric in its arguments: $\langle v, u \rangle = \langle u, v \rangle$ for all $u, v \in V$ 3. $\langle \cdot, \cdot \rangle$ is positive definite: $\langle v, v \rangle \geq 0$ for all $v \in V$ with equality if and only if $v = \mathbf{0}$.

An inner product provides a vector space with the concepts of

- **distance**: for all $v \in V$ define the **norm** of v to be

$$||v|| = \langle v, v \rangle^{\frac{1}{2}}$$

Thus any inner-product space $(V, \langle \cdot, \cdot \rangle)$ is also a normed space $(V, || \cdot ||)$, and a metric space $(V, d(x, y) = ||x - y||)$.

- **angle**: for $u, v \in V$ we define the angle between u and v to be θ where

$$\langle u, v \rangle = ||u||.||v|| \cos\theta$$
$$\implies \theta = \cos^{-1}\left(\frac{\langle u, v \rangle}{||u||\ ||v||}\right)$$

We will primarily be interested in the concept of **orthogonality**. We say $u, v \in V$ are orthogonal if

$$\langle u, v \rangle = 0$$

i.e., the *angle* between them is $\frac{\pi}{2}$.

If you have done any functional analysis, you may recall that a Hilbert space is a *complete* inner-product space, and a Banach space is a complete normed space. This is an applied module, so we will skirt much of the technical detail, but note that some of the proofs formally require us to be working in a Banach or Hilbert space. We will not concern ourselves with such detail.

Example 2.11. We will mostly be working with the Euclidean vector spaces $V = \mathbb{R}^n$, in which we use the *Euclidean* inner product

$$\langle u, v \rangle = u^\top v$$

sometimes called the **scalar** or **dot product** of u and v. Sometimes this gets weighted by a matrix so that

$$\langle u, v \rangle_Q = u^\top Q v.$$

The norm associated with the dot product is the square root of the sum of squared errors, denoted by $||\cdot||_2$. The **length** of u is then

$$||u||_2 = \sqrt{u^\top u} = \left(\sum_{i=1}^{n} u_i^2 \right)^{\frac{1}{2}} \geq 0.$$

Note that $||u||_2 = 0$ if and only if $u = \mathbf{0}_n$ where $\mathbf{0}_n^{n\times 1} = (0, 0, \ldots, 0)^\top$.

We say u is orthogonal to v if $u^\top v = 0$. For example, if

$$u = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \text{ and } v = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$$

then

$$||u||_2 = \sqrt{5} \text{ and } u^\top v = 0.$$

We will write $u \perp v$ if u is orthogonal to v.

Definition 2.9. p-norm: The subscript 2 hints at a wider family of norms. We define the L_p norm to be

$$||v||_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{\frac{1}{p}}.$$

2.3.2 Orthogonal matrices

Definition 2.10. A **unit vector v** is a vector satisfying $||\mathbf{v}|| = 1$, i.e., it is a vector of length 1. Vectors u and v are orthonormal if

$$||u|| = ||v|| = 1 \text{ and } \langle u, v \rangle = 0.$$

An $n \times n$ matrix $\mathbf{Q}$ is an **orthogonal matrix** if

$$\mathbf{Q}Q^\top = \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_n.$$

Equivalently, a matrix $\mathbf{Q}$ is orthogonal if $\mathbf{Q}^{-1} = \mathbf{Q}^\top$.

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

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

Lemma 2.2. *Let Q be a $n \times p$ matrix and suppose $Q^\top Q = \mathbf{I}_p$, where $\mathbf{I}_p$ is the $p \times p$ identity matrix. If Q is a square matrix ($n = p$), then $QQ^\top = \mathbf{I}_p$. If Q is not square ($n \neq p$), then $QQ^\top \neq I_n$.*

Proof. Suppose $n = p$, and think of Q as a linear map""

$$\begin{aligned} Q :& \mathbb{R}^n \to \mathbb{R}^n \\ & v \mapsto Qv \end{aligned}$$

By the rank-nullity theorem,

$$\dim \operatorname{Ker}(Q) + \dim \operatorname{Im}(Q) = n$$

and because Q has a left-inverse, we must have $\dim \operatorname{Ker}(Q) = 0$, as otherwise $Q^\top$ would have to map from a vector space of dimension less than n to $\mathbb{R}^n$. So Q is of full rank, and thus must also have a right inverse, B say, with $QB = \mathbf{I}_n$. If we left multiply by $Q^\top$ we get

$$\begin{aligned} QB &= \mathbf{I}_n \\ Q^\top QB &= Q^\top \\ \mathbf{I}_n B &= Q^\top \\ B &= Q^\top \end{aligned}$$

and so we have that $Q^{-1} = Q^\top$.

Now suppose Q is $n \times p$ with $n \neq p$. Then as $Q^\top Q = \mathbf{I}_{p \times p}$, we must have $\operatorname{tr}(Q^\top Q) = p$. This implies that

$$\operatorname{tr}(QQ^\top) = \operatorname{tr}(Q^\top Q) = m$$

and so we cannot have $QQ^\top = \mathbf{I}_n$ as $\operatorname{tr} \mathbf{I}_n = n$. □

Corollary 2.2. *If $q_1, \ldots, q_n$ are mutually orthogonal $n \times 1$ unit vectors then*

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

Proof. Let Q be the matrix with i^{th} column q_i

$$Q = \begin{pmatrix} | & & | \\ q_1 & \ldots & q_n \\ | & & | \end{pmatrix}.$$

Then $Q^\top Q = \mathbf{I}_n$, and Q is $n \times n$. Thus by Lemma 2.2, we must also have $QQ^\top = \mathbf{I}_n$ and if we think about matrix-matrix multiplication as columns times rows (c.f. section 3.1), we get

$$\mathbf{I}_n = QQ^\top = \begin{pmatrix} | & & | \\ q_1 & \ldots & q_n \\ | & & | \end{pmatrix} \begin{pmatrix} - & q_1^\top & - \\ & \vdots & \\ - & q_n^\top & - \end{pmatrix} = \sum_{i=1}^{n} q_i q_i^\top$$

as required.

□

2.3.3 Projections

Definition 2.11. $\overset{n\times n}{P}$ is a *projection* matrix if

$$P^2 = P$$

i.e., if it is idempotent.

View P as a map from a vector space W to itself. Let $U = \mathrm{Im}(P)$ and $V = \mathrm{Ker}(P)$ be the image and kernel of P.

Proposition 2.3. *We can write $w \in W$ as the sum of $u \in U$ and $v \in V$.*

Proof. Let $w \in W$. Then

$$w = \mathbf{I}_n w = (\mathbf{I} - P)w + Pw$$

Now $Pw \in \mathrm{Im}(P)$ and $(\mathbf{I} - P)w \in \mathrm{Ker}(P)$ as

$$P(\mathbf{I} - P)w = (P - P^2)w = 0.$$

□

Proposition 2.4. *If $\overset{n\times n}{P}$ is a projection matrix then $\mathbf{I}_n - P$ is also a projection matrix.*

The kernel and image of $\mathbf{I} - P$ are the image and kernel (respectively) of P:

$$\mathrm{Ker}(\mathbf{I} - P) = U = \mathrm{Im}(P)$$
$$\mathrm{Im}(\mathbf{I} - P) = V = \mathrm{Ker}(P).$$

2.3.3.1 Orthogonal projection

We are mostly interested in **orthogonal** projections.

Definition 2.12. If W is an inner product space, and U is a subspace of W, then the orthogonal projection of $w \in W$ onto U is the unique element $u \in U$ that minimizes

$$||w - u||.$$

In other words, the orthogonal projection of w onto U is the *best possible approximation* of w in U.

As above, we can split W into U and its orthogonal compliment

$$U^{\perp} = \{x \in W : \langle x, u \rangle = 0\}$$

i.e., $W = U \oplus U^{\perp}$ so that any $w \in W$ can be written as $w = u + v$ with $u \in U$ and $v \in U^{\perp}$.

Proposition 2.5. *If $\{u_1, \dots, u_k\}$ is a basis for U, then the orthogonal projection matrix (i.e., the matrix that projects $w \in W$ onto U) is*

$$P_U = A(A^{\top}A)^{-1}A^{\top}$$

where $A = [u_1 \ \dots \ u_k]$ is the matrix with columns given by the basis vectors.

Proof. We need to find $u = \sum \lambda_i u_i = A\lambda$ that minimizes $||w - u||$.

$$\begin{aligned} ||w-u||^2 &= \langle w-u, w-u \rangle \\ &= w^{\top}w - 2u^{\top}w + u^{\top}u \\ &= w^{\top}w - 2\lambda^{\top}A^{\top}w + \lambda^{\top}A^{\top}A\lambda. \end{aligned}$$

Differentiating with respect to λ and setting equal to zero gives

$$0 = -2A^{\top}w + 2A^{\top}A\lambda$$

and hence

$$\lambda = (A^{\top}A)^{-1}A^{\top}w.$$

The orthogonal projection of w is hence

$$A\lambda = A(A^{\top}A)^{-1}A^{\top}w$$

and the projection matrix is

$$P_U = A(A^{\top}A)^{-1}A^{\top}.$$

□

Notes:

1. If $\{u_1, \ldots, u_k\}$ is an orthonormal basis for U then $A^\top A = \mathbf{I}$ and $P_U = AA^\top$. We can then write
$$P_U w = \sum_i (u_i^\top w) u_i$$
and
$$P_U = \sum_{i=1}^{k} u_i u_i^\top.$$
Note that if $U = W$ (so that P_U is a projection from W onto W, i.e., the identity), then A is a square matrix ($n \times n$) and thus $A^\top A = \mathbf{I}_n \implies AA^\top$ and thus $P_U = \mathbf{I}_n$ as required. The coordinates (with respect to the orthonormal basis $\{u_1, \ldots, u_k\}$) of a point w projected onto U are $A^\top w$.
2. $P_U^2 = P_U$, so P_U is a projection matrix in the sense of definition 2.11.
3. P_U is symmetric ($P_U^\top = P_U$). This is true for orthogonal projection matrices, but not in general for projection matrices.

Example 2.12. Consider the vector space $\mathbb{R}^2$ and let $u = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. The projection of $v \in \mathbb{R}^2$ onto u is given by $(v^\top u)u$. So for example, if $v = (2,\ 1)^\top$, then its projection onto u is
$$P_U v = \frac{3}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$
Alternatively, if we treat u as a basis for U, then the coordinate of $P_U v$ with respect to the basis is 3. To check this, draw a picture!

2.3.3.2 Geometric interpretation of linear regresssion

Consider the linear regression model
$$y = X\beta + e$$
where $y \in \mathbb{R}^n$ is the vector of observations, X is the $n \times p$ design matrix, β is the $p \times 1$ vector of parameters that we wish to estimate, and e is a $n \times 1$ vector of zero-mean errors.

Least-squares regression tries to find the value of $\beta \in \mathbb{R}^p$ that minimizes the sum of squared errors, i.e., we try to find β to minimize
$$||y - X\beta||_2$$

We know that $X\beta$ is in the column space of X, and so we can see that linear regression aims to find the *orthogonal projection* onto $\mathcal{C}(X)$.
$$P_U y = \arg \min_{y' : y' \in \mathcal{C}(X)} ||y - y'||_2.$$

By Proposition 2.5 this is

$$P_U y = X(X^\top X)^{-1} X^\top y = \hat{y}$$

which equals the usual prediction obtained in linear regression ($\hat{y}$ are often called the fitted values). We can also see that the choice of β that specifies this point in $\mathcal{C}(X)$ is

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

which is the usual least-squares estimator.

2.4 Miscellaneous topics

2.4.1 The Centering Matrix

The centering matrix will be play an important role in this module, as we will use it to remove the column means from a matrix (so that each column has mean zero), *centering* the matrix.

Definition 2.13. The **centering matrix** is

$$H = \mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top. \tag{2.2}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix, and $\mathbf{1}_n$ is an $n \times 1$ column vector of ones.

You will be asked to prove the following results about H 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 $\mathbf{1}_n^\top H = \mathbf{0}_n^\top$. In words: the sum of each row and each column of H is 0.
3. If $x = (x_1, \ldots, x_n)^\top$, then $Hx = x - \bar{x}\mathbf{1}_n$ where $\bar{x} = n^{-1} \sum_{i=1}^n x_i$. I.e., H subtracts the mean $\bar{x}$ from x.
4. With x as in 3., 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

$$X = \begin{bmatrix} - & x_1^\top & - \\ & \vdots & \\ - & x_n^\top & - \end{bmatrix} = [x_1, \ldots, x_n]^\top$$

is an $n \times p$ data matrix containing data points $x_1, \ldots, x_n \in \mathbb{R}^p$, then

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

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \in \mathbb{R}^p$$

is the p-dimensional sample mean of $x_1, \ldots, x_n \in \mathbb{R}^p$. In words, H has subtracted the column mean from each column of X.

6. With X as in 5.

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

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 general matrix case.

2.4.2 Quadratic forms and ellipses

POSSIBLY MOVE OR ADD PICTURES - DECIDE ONCE I KNOW WHERE It IS USED.

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

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} \leq 1. \tag{2.3}$$

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 (2.3) 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.

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

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

$$\mathcal{L} = \mathcal{L}(a, b) = \{a + \gamma b : \gamma \in \mathbb{R}\} \tag{2.4}$$

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 $\mathcal{L}(a, b)$.

PICTURE??

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 $\mathcal{H}(a, \gamma)$.

I DON"T KNOW WHY THIS IS HERE? THINK ABOUT

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 (2.4) 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.

Chapter 3

Matrix decompositions

This chapter focusses on two ways to decompose a matrix into smaller parts. We can then think about which are the most important parts of the matrix, and that will be useful when we think about dimension reduction. The highlight of the chapter is the singular value decomposition (SVD), which is one of the most useful mathematical concepts from the past century, and is relied upon throughout statistics and machine learning. The SVD extends the idea of the eigen (or spectral) decomposition of symmetric square matrices to any matrix.

- Matrix-matrix products
- Eigenvalues and the spectral decomposition
- Introduction to the singular value decomposition
- SVD optimization results
- Low-rank approximation

3.1 Matrix-matrix products

Before we get to the SVD, we first need to recap some basic material on matrix multiplication and eigenvalues. We saw in section 2.2.2 that we can think about matrix-vector products in two ways: Ax is rows of A times x; or as a linear combination of the columns of A. We can similarly think about matrix-matrix products in two ways.

The usual way to think about the matrix product AB is as the rows of A times the columns of B:

$$\begin{bmatrix} \cdot & \cdot & \cdot \\ a_{21} & a_{22} & a_{23} \\ \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot & b_{12} & \cdot \\ \cdot & b_{22} & \cdot \\ \cdot & b_{32} & \cdot \end{bmatrix}$$

A better way (for this module) to think of AB is as the columns of A times the rows of B. If we let a_i denote the columns of A, and b_i^* the rows of B then

$$\begin{bmatrix} | & | & | \\ a_1 & a_2 & a_3 \\ | & | & | \end{bmatrix} \begin{bmatrix} - & b_1^* & - \\ - & b_2^* & - \\ - & b_3^* & - \end{bmatrix} = \sum_{i=1}^{3} a_i b_i^*$$

i.e., AB is a sum of the columns of A times the rows of B.

Note that if a is a vector of length n and b is a vector of length p then $ab^\top$ is an $n \times p$ matrix.

Example 3.1.

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix} (2\ 3\ 1) = \begin{pmatrix} 2 & 3 & 1 \\ 4 & 6 & 2 \end{pmatrix}.$$

Note that $ab^\top$ is a rank-1 matrix as its columns are all multiples of a, or in other words, its column space is just multiples of a.

$$\mathcal{C}(ab^\top) = \{\lambda a : \lambda \in \mathbb{R}\}.$$

We sometimes call $ab^\top$ the **outer product** of a with b.

By thinking of matrix-matrix multiplication in this way

$$AB = \sum_{i=1}^{k} a_i b_i^*$$

(where k is the number of columns of A and the number of rows of B) we can see that the product is a sum of rank-1 matrices. We can think of rank-1 matrices as the building blocks of matrices.

This chapter is about ways of decomposing matrices into their most important parts, and we will do this by thinking about the most important rank-1 building blocks.

Firstly though, we need a recap on eigenvectors.

3.2 Eigenvalues and eigenvectors

Consider the $n \times n$ matrix A. We say that vector $x \in \mathbb{R}^n$ is an **eigenvector** corresponding to **eigenvalue** λ of A if

$$\mathbf{Ax} = \lambda \mathbf{x}.$$

To find the eigenvalues of a matrix, we note that if λ is an eigenvalue, then $(A - \lambda \mathbf{I}_n)x = 0$, i.e., the kernel of $A - \lambda \mathbf{I}_n$ has dimension at least 1, so $A - \lambda \mathbf{I}_n$ is not invertible, and so we must have $\det(\mathbf{A} - \lambda \mathbf{I}_n) = 0$.

Let $R(\lambda) = \det(\mathbf{A} - \lambda \mathbf{I}_n)$, which is an n^{th} order polynomial in λ. To find the eigenvalues of A we find the n roots $\lambda_1, \ldots, \lambda_n$ of $R(\lambda)$. We will always consider ordered eigenvalues so that $\lambda_1 \geq \cdots \geq \lambda_n$

Proposition 3.1. *If* $\mathbf{A}$ *is symmetric (i.e.* $\mathbf{A}^\top = \mathbf{A}$*) then the eigenvalues and eigenvectors of* $\mathbf{A}$ *are* real *(in* $\mathbb{R}$*).*

Proposition 3.2. *If* $\overset{n\times n}{\mathbf{A}}$ *is a symmetric matrix then its determinant is the product of its eigenvalues, i.e. det*$(\mathbf{A}) = \lambda_1 \ldots \lambda_n$*.*

Thus,

$$A \text{ is invertible} \iff \det(A) \neq 0 \iff \lambda_i \neq 0 \forall i \iff A \text{ is of full rank}$$

3.3 Spectral/eigen decomposition

The key to much of dimension reduction is finding matrix decompositions. The first decomposition we will consider is the **spectral decomposition** (also called an **eigen-decomposition**).

Proposition 3.3. ***(Spectral decomposition)****. Any symmetric matrix* $\overset{n\times n}{\mathbf{A}}$ *can be written as*

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

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

Because Λ is a diagonal matrix, we sometimes refer to the spectral decomposition as **diagonalizing** the matrix A as $Q^\top AQ = \Lambda$ is a diagonal matrix.

This will be useful at various points throughout the module. Note that it relies upon the fact that the eigenvectors of A can be chosen to be mutually orthogonal, and as there are n of them, they form an orthonormal basis for $\mathbb{R}^n$.

Corollary 3.1. *The rank of a symmetric matrix is equal to the number of non-zero eigenvalues (counting according to their multiplicities).*

Proof. If r is the number of non-zero eigenvalues of A, then we have (after possibly reordering the λ_i)

$$\mathbf{A} = \sum_{i=1}^{r} \lambda_i \mathbf{q}_i \mathbf{q}_i^\top.$$

Each $\mathbf{q}_i \mathbf{q}_i^\top$ is a rank 1 matrix, with column space equal to the span of q_i. As the q_i are orthogonal, the columns spaces $\mathcal{C}(q_i q_i^\top)$ are orthogonal, and their union is a vector space of dimension r. Hence the rank of A is r. □

Lemma 3.1. *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* positive definite *if and only if* $\lambda_n > 0$. *It is positive semi-definite if and only if* $\lambda_n \geq 0$.

Proof. If A is positive definite, and if x is a unit-eigenvalue of A corresponding to λ_n, then

$$0 \leq x^\top Ax = \lambda_n x^\top x = \lambda_n.$$

Conversely, suppose A has positive eigenvalues. Because A is real and symmetric, we can write it as $A = Q\,Q^\top$. Now if x is a non-zero vector, then $y = Q^\top x \neq 0$, (as $Q^\top$ has inverse Q and hence $\dim \mathrm{Ker}(Q) = 0$). Thus

$$x^\top Ax = y^\top\, y = \sum_{i=1}^{n} \lambda_i y_i^2 > 0$$

and thus A is positive definite. □

Note: A covariance matrix Σ is always positive semi- definite (and thus always has non-negative eigenvalues). To see this, recall that if x is a random vector with $\mathbb{V}\,\mathrm{ar}(x) = \Sigma$, then for any constant vector a, the random variable $a^\top x$ has variance $\mathbb{V}\,\mathrm{ar}(a^\top x) = a^\top \Sigma a$. Because variances are positive, we must have

$$a^\top \Sigma a \geq 0 \;\forall\; a.$$

Moreover, if Σ is positive definite (so that its eigenvalues are positive), then its determinant will be positive (so that Σ is **non-singular**) and we can find an inverse Σ^{-1} matrix, which is called the **precision** matrix.

Proposition 3.4. *The eigenvalues of a projection matrix* P *are all* 0 *or* 1.

3.3.1 Matrix square roots

From the spectral decomposition theorem, we can see that if A is a symmetric positive semi-definite matrix, then for any integer p

$$A^p = Q\,{}^p Q^\top.$$

If in addition A is positive definite (rather than just semi-definite), then

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

where ${}^{-1} = \mathrm{diag}\{\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_n}\}$.

The spectral decomposition also gives us a way to define a matrix square root. If we assume A is positive semi-definite, then its eigenvalues are non-negative, and the diagonal elements of are all non-negative.

We then define $A^{1/2}$, a matrix square root of A, to be $A^{1/2} = Q\Lambda^{1/2}Q^\top$ 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 $Q^\top Q = \mathbf{I}_n$ 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, positive semi-definite square root of A.

If A is positive definite (as opposed to just positive semi-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, $A^{-1/2}$ is the matrix square root of A^{-1}. Furthermore, similar calculations show that

$$A^{1/2}A^{-1/2} = A^{-1/2}A^{1/2} = \mathbf{I}_n,$$

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

3.4 Singular Value Decomposition (SVD)

The spectral decomposition theorem (Proposition 3.3) gives a decomposition of any symmetric matrix. We now give a generalisation of this result which applies to *all* matrices.

If matrix A is not a square matrix, then it cannot have eigenvectors. Instead, it has **singular vectors** corresponding to **singular values**. Suppose is a $n \times p$ matrix. Then we say σ is a **singular value** with corresponding **left** and **right** singular vectors u and v (respectively) if

$$Av = \sigma u \quad \text{and} \quad A^\top u = \sigma v$$

If A if a symmetric matrix then $u = v$ is a eigenvector and σ is an eigenvalue.

The singular value decomposition (SVD) **diagonalizes** A into a product of a matrix of left singular vectors U, a diagonal matrix of singular values Σ, and a matrix of right singular vectors V.

$$A = U\Sigma V^\top.$$

Proposition 3.5. ***(Singular value decomposition)***. *Let A be a $n \times p$ matrix of rank r, where $1 \leq r \leq \min(n, p)$. Then there exists a $n \times r$ matrix $U = [u_1, ..., u_r]$, a $p \times r$ matrix $V = [\mathbf{v}_1, ..., \mathbf{v}_r]$, and a $r \times r$ diagonal matrix $\Sigma = diag\{\sigma_1, ..., \sigma_r\}$ such that*

$$A = U\, V^\top = \sum_{i=1}^{r} \sigma_i u_i \mathbf{v}_i^\top,$$

where $U^\top U = \mathbf{I}_r = V^\top V$ and the $\sigma_1 \geq ... \geq \sigma_r > 0$.

Note that the u_i and the $\mathbf{v}_i$ are necessarily unit vectors, and that we have ordered the singular values from largest to smallest. The scalars $\sigma_1, ..., \sigma_r$ are called the **singular values** of A, the columns of U are the **left singular vectors**, and the columns of V are the **right singular vectors**.

The form of the SVD given above is called the **compact singular value decomposition**. Sometimes we write it in a non-compact form

$$A = U\Sigma V^\top$$

where U is a $n \times n$ orthogonal matrix ($U^\top U = UU^\top = \mathbf{I}_n$), V is a $p \times p$ orthogonal matrix ($V^\top V = VV^\top = \mathbf{I}_p$), and Σ is a $n \times p$ diagonal matrix

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & ... & & & 0 \\ 0 & \sigma_2 & 0 & ... & & \\ \vdots & & & & & \\ 0 & 0 & & ... & \sigma_r & \\ 0 & 0 & & ... & & 0 \quad ... \\ \vdots & & & & & \\ 0 & 0 & & ... & & 0 \end{pmatrix}. \tag{3.1}$$

The columns of U and V form an orthonormal basis for $\mathbb{R}^n$ and $\mathbb{R}^p$ respectively. We can see that we recover the compact form of the SVD by only using the first r columns of U and V, and truncating Σ to a $r \times r$ matrix with non-zero diagonal elements.

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

Proof. $A^\top A$ is a $p \times p$ symmetric matrix, and so by the spectral decomposition theorem we can write it as

$$A^\top A = V\Lambda V^\top$$

where V is a $p \times p$ orthogonal matrix containing the orthonormal eigenvectors of $A^\top A$, and $\Lambda = \text{diag}(\lambda_1, ..., \lambda_r)$ is a diagonal matrix of eigenvalues with $\lambda_1 \geq ... \geq \lambda_r > 0$ (by Corollary 3.1).

For $i = 1, \dots, r$, let $\sigma_i = \sqrt{\lambda_i}$ and let $u_i = \frac{1}{\sigma_i} A v_i$. Then the vectors u_i are orthonormal:

$$\begin{aligned} u_i^\top u_j &= \frac{1}{\sigma_i \sigma_j} v_i^\top A^\top A v_j \\ &= \frac{\sigma_j^2}{\sigma_i \sigma_j} v_i^\top v_j \quad \text{as } v_j \text{ is an eigenvector of } A^\top A \\ &= \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases} \quad \text{as the } v_i \text{ are orthonormal vectors.} \end{aligned}$$

In addition

$$A^\top u_i = \frac{1}{\sigma_i} A^\top A v_i = \frac{\sigma_i^2}{\sigma_i} v_i = \sigma_i v_i$$

and so u_i and v_i are left and right singular vectors.

Let $U = [u_1 \ u_2 \ \dots \ u_r \ \dots \ u_n]$, where $u_{r+1}, \dots, u_n$ are chosen to complete the orthonormal basis for $\mathbb{R}^n$ given $u_1, \dots, u_r$, and let Σ be the $n \times p$ diagonal matrix in Equation (3.1).

Then we have shown that

$$U = AV\Sigma^{-1}$$

Thus

$$\begin{aligned} U &= AV\Sigma^{-1} \\ U\Sigma &= AV \\ U\Sigma V^\top &= A. \end{aligned}$$

□

Note that by construction we've shown that $A^\top A$ has eigenvalues σ_i^2 with corresponding eigenvectors v_i. We also can also show that $AA^\top$ has eigenvalues σ_i^2, but with corresponding eigenvectors u_i.

$$AA^\top u_i = \sigma_i A v_i = \sigma_i^2 u_i$$

Proposition 3.6. *Let A be any matrix of rank r. Then the non-zero eigenvalues of both $AA^\top$ and $A^\top A$ are $\sigma_1^2, \dots, \sigma_r^2$. The corresponding unit eigenvectors of $AA^\top$ are given by the columns of U, and the corresponding unit eigenvectors of $A^\top A$ are given by the columns of $\mathbf{V}$.*

Notes:

1. The SVD expresses a matrix as a sum of rank-1 matrices

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top.$$

We can think of these as a list of the building blocks of A ordered by their importance ($\sigma_1 \geq \sigma_2 \geq ...$).

2. The singular value decomposition theorem shows that every matrix is diagonal, provided one uses the proper bases for the domain and range spaces. We can **diagonalize** A by

$$U^\top AV = \Sigma.$$

3. The SVD reveals a great deal about a matrix. Firstly, the rank of A is the number of non-zero singular values. The left singular vectors $u_1, ..., u_r$ are an orthonormal basis for the columns space of A, $\mathcal{C}(A)$, and the right singular vectors $v_1, ..., v_r$ are an orthonormal basis for $\mathcal{C}(A^\top)$, the row space of A. The vectors $v_{r+1}, ..., v_p$ from the non-compact SVD are a basis for the kernel of A (sometimes called the null space $\mathcal{N}(A)$), and $u_{r+1}, ..., u_n$ are a basis for $\mathcal{N}(A^\top)$.

4. The SVD has many uses in mathematics. One is as a generalized inverse of a matrix. If A is $n \times p$ with $n \neq p$, or if it is square but not of full rank, then A cannot have an inverse. However, we say A^+ is a generalized inverse if $AA^+A = A$. One such generalized inverse can be obtained from the SVD by $A^+ = V\Sigma^{-1}U^\top$ - this is known as the Moore-Penrose pseudo-inverse.

3.4.1 Examples

In practice, we don't compute SVDs of a matrix by hand: in R you can use the command `SVD(A)` to compute the SVD of matrix `A`. However, it is informative to do the calculation yourself a few times to help fix the ideas.

Example 3.2. Consider the matrix $A = xy^\top$. We can see this is a rank-1 matrix, so it only has one non-zero singular value which is $\sigma_1 = ||x||.||y||$. Its SVD is given by

$$U = \frac{1}{||x||}x, \quad V = \frac{1}{||y||}y, \quad \text{and } \Sigma = ||x||.||y||.$$

Example 3.3. Let

$$A = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{pmatrix}.$$

Let's try to find the SVD of A.

We know the singular values are the square roots of the eigenvalues of $AA^\top$ and $A^\top A$. We'll work with the former as it is only 2×2.

$$AA^\top = \begin{pmatrix} 17 & 8 \\ 8 & 17 \end{pmatrix} \quad \text{and so } \det(AA^\top - \lambda \mathbf{I}) = (17 - \lambda)^2 - 64$$

Solving $\det(AA^\top - \lambda\mathbf{I}) = 0$ gives the eigenvalues to be $\lambda = 25$ or 9. Thus the singular values of A are $\sigma_1 = 5$ and $\sigma_2 = 3$, and

$$\Sigma = \begin{pmatrix} 5 & 0 \\ 0 & 3 \end{pmatrix}.$$

The columns of U are the *unit* eigenvectors of $AA^\top$ which we can find by solving

$$(A - 25\mathbf{I}_2)u = \begin{pmatrix} -8 & 8 \\ 8 & -8 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{and}$$
$$(A - 9\mathbf{I}_2)u = \begin{pmatrix} 8 & 8 \\ 8 & 8 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

And so, remembering that the eigenvectors used to form V need to be *unit* vectors, we can see that

$$U = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Finally, to compute V recall that $\sigma_i v_i = A^\top u_i$ and so

$$V = A^\top U\Sigma^{-1} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & \frac{1}{3} \\ 1 & \frac{-1}{3} \\ 0 & \frac{4}{3} \end{pmatrix}.$$

This completes the calculation, and we can see that we can express A as

$$A = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix}\begin{pmatrix} 5 & 0 \\ 0 & 3 \end{pmatrix}\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{3\sqrt{2}} & \frac{-1}{3\sqrt{2}} & \frac{4}{3\sqrt{2}} \end{pmatrix}$$

or as the sum of rank-1 matrices:

$$A = 5\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{3\sqrt{2}} \end{pmatrix} + 3\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{pmatrix}\begin{pmatrix} \frac{1}{3\sqrt{2}} & \frac{-1}{3\sqrt{2}} & \frac{4}{3\sqrt{2}} \end{pmatrix}$$

This is the compact form of the SVD. To find the non-compact form we need V to be a 3×3 matrix, which requires us to find a 3rd column that is orthogonal to the first two columns (thus completing an orthonormal basis for $\mathbb{R}^3$). We can do that with the vector $v_3 = \frac{1}{\sqrt{17}}(2\ -2\ -3)$ giving the non-compact SVD for A.

$$A = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix}\begin{pmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{pmatrix}\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{3\sqrt{2}} & \frac{2}{\sqrt{17}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{3\sqrt{2}} & \frac{-2}{\sqrt{17}} \\ 0 & \frac{4}{3\sqrt{2}} & \frac{-3}{\sqrt{17}} \end{pmatrix}^\top$$

Let's check our answer in R.

```
A<- matrix(c(3,2,2,2,3,-2), nr=2, byrow=T)
svd(A)
```

```
## $d
## [1] 5 3
##
## $u
##            [,1]       [,2]
## [1,] -0.7071068 -0.7071068
## [2,] -0.7071068  0.7071068
##
## $v
##                [,1]       [,2]
## [1,] -7.071068e-01 -0.2357023
## [2,] -7.071068e-01  0.2357023
## [3,] -5.551115e-17 -0.9428090
```

The eigenvectors are only defined upto multiplication by -1 and so we can multiply any pair of left and right singular vectors by -1 and it is still a valid SVD.

Note: In practice this is a terrible way to compute the SVD as it is prone to numerical error. In practice an efficient iterative method is used in most software implementations (including R).

3.5 Optimization results

Why are eigenvalues and singular values useful in statistics? It is because they appear as the result of some important optimization problems. We'll see more about this in later chapters, but we'll prove a few preliminary results here.

For example, suppose $x \in \mathbb{R}^n$ is a random variable with $\mathbb{C}\,\mathrm{ov}(x) = \Sigma$ (an $n \times n$ matrix), then can we find a projection of x that has either maximum or minimum variance? I.e., can we find a such that

$$\mathbb{V}\,\mathrm{ar}(a^\top x) = a^\top \Sigma a$$

is maximized or minimized? To make the question interesting we need to constrain the length of a so lets assume that $||a||_2 = \sqrt{a^\top a} = 1$, otherwise we could just take $a = 0$ to obtain a projection with variance zero. So we want to solve the optimization problems involving the quadratic form $a^\top \Sigma a$:

$$\max_{a:\, a^\top a = 1} \mathbf{a}^\top \mathbf{a}, \quad \text{and} \quad \min_{a:\, a^\top a = 1} \mathbf{a}^\top \mathbf{a}. \tag{3.2}$$

Given that Σ is symmetric, we can write it as

$$\Sigma = V \Lambda V^\top$$

where Λ is the diagonal matrix of eigenvalues of Σ, and V is an orthogonal matrix of eigenvectors. If we let $b = V^\top a$ then

$$a^\top \Sigma a = b^\top \Lambda b = \sum_{i=1}^{n} \lambda_i b_i^2$$

and given that the eigenvalues are ordered $\lambda_1 \geq \lambda_2 \geq ...$ and that

$$\sum_{i=1}^{n} b_i^2 = b^\top b = a^\top V V^\top a = a^\top a = 1,$$

we can see that the maximumn is λ_1 obtained by setting $b = (1\ 0\ 0\ ...)^\top$. Then

$$\begin{aligned} V^\top a &= b \\ V V^\top a &= V b \\ a &= v_1 \end{aligned}$$

so we can see that the maximum is obtained when $a = v_1$, the eigenvector of Σ corresponding to the largest eigenvalue λ_1.

Similarly, the minimum is λ_n, which obtained by setting $b = (\ 0\ 0\ ...\ 0\ 1)^\top$ which corresponds to $a = v_n$.

Proposition 3.7. *For any symmetric $n \times n$ matrix Σ,*

$$\max_{a : a^\top a = 1} a^\top \Sigma a = \lambda_1,$$

where the maximum occurs at $a = \pm v_1$, and

$$\min_{a : a^\top a = 1} a^\top \Sigma a = \lambda_n$$

where the minimum occurs at $a = \pm v_n$, where λ_i, v_i are the ordered eigenpairs of Σ.

Note that

$$\frac{a^\top \Sigma a}{a^\top a} = \frac{a^\top \Sigma a}{||a||^\top} = (\frac{a}{||a||})^\top \Sigma (\frac{a}{||a||})$$

and so another way to write the maximization problems (3.2) is as unconstrained optimization problems:

$$\max_a \frac{a^\top \Sigma a}{a^\top a} \quad \text{and} \quad \min_a \frac{a^\top \Sigma a}{a^\top a}.$$

We obtain a similar result for non-square matrices using the singular value decomposition.

Proposition 3.8. *For any matrix A*

$$\max_{x:||x||_2=1} ||Ax||_2 = \max_x \frac{||Ax||_2}{||x||_2} = \sigma_1$$

the first singular value of A, with the maximum achieved at $x = v_1$ (the first right singular vector).

Proof. This is follows from 3.7 as

$$||Ax||_2^2 = x^\top A^\top Ax.$$

□

Finally, we will need the following result when we study canonical correlation analysis:

Proposition 3.9. *For any matrix A, we have*

$$\max_{a,b:||a||=||b||=1} a^\top Ab = \sigma_1.$$

with the maximum obtained at $a = u_1$ and $b = v_1$, the first left and right singular vectors of A.

Proof. □

We'll see much more of this kind of thing in Chapters 4 and 5.

3.6 Best approximating matrices

One of the reasons the SVD is so widely used is that it can be used to find the best low rank approximation to a matrix. Before we discuss this, we need to define what it means for some matrix B to be a good approximation to A. To do that, we need the concept of a matrix norm.

3.6.1 Matrix norms

In Section 2.3.1 we described norms on vectors. Here will extend this idea to include norms on matrices, so that we can discuss the size of a matrix $||A||$, and the distance between two matrices $||A - B||$. There are two particular norms we will focus on. The first is called the Frobenious norm (or sometimes the Hilbert-Schmidt norm).

Definition 3.1. Let $A \in \mathbb{R}^{n \times p}$. The **Frobenius norm** of A is

$$||A||_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{p} |a_{ij}|^2 \right)^{\frac{1}{2}} = (\text{tr}\, A^\top A)^{\frac{1}{2}}$$

where a_{ij} are the individual entries of A.

Note that the Frobenius norm is invariant to rotation by an orthogonal matrix U:

$$\begin{aligned} ||AU||_F^2 &= \text{tr}(U^\top A^\top AU) \\ &= \text{tr}(UU^\top A^\top A) \\ &= \text{tr}(A^\top A) \\ &= ||A||_F^2. \end{aligned}$$

Proposition 3.10.

$$||A||_F = \left(\sum_{i=1}^{r} \sigma_i^2 \right)^{\frac{1}{2}}$$

where σ_i are the singular values of A, and $r = \text{rank}(A)$.

Proof. Using the (non-compact) SVD $A = U\Sigma V^\top$ we have

$$||A||_F = ||U^\top A||_F = ||U^\top AV||_F = ||\Sigma||_F = \text{tr}(\Sigma^\top \Sigma)^{\frac{1}{2}} = \left(\sum \sigma_i^2 \right)^{\frac{1}{2}}.$$

□

We previously defined the p-norms for vectors in $\mathbb{R}^p$ to be

$$||x||_p = \left(\sum |x_i|^p \right)^{\frac{1}{p}}.$$

These vector norms *induce* matrix norms, sometimes also called operator norms:

Definition 3.2. The p-norms for matrices are defined by

$$||A||_p = \sup_{x \neq 0} \frac{||Ax||_p}{||x||_p} = \sup_{x : ||x||_p = 1} ||Ax||_p$$

Proposition 3.11.

$$||A||_2 = \sigma_1$$

where σ_1 is the first singular value of A.

Proof. By Proposition 3.8. □

3.6.2 Eckart-Young-Mirsky Theorem

Now that we have defined a norm (i.e., a distance) on matrices, we can think about approximating a matrix A by a matrix that is easier to work with. We have shown that any matrix can be split into the sum of rank-1 component matrices

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$$

We'll now consider a family of approximations of the form

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^\top \tag{3.3}$$

where $k <= r = \text{rank}(A)$. This is a rank-k matrix, and as we'll now show, it is the best possible rank-k approximation to A.

Theorem 3.1. ***(Eckart-Young-Mirsky)*** *For either the 2-norm $||\cdot||_2$ or the Frobenious norm $||\cdot||_F$*

$$||A - A_k|| \leq ||A - B|| \text{ for all rank-k matrices } B.$$

Moreover,

$$||A - A_k|| = \begin{cases} \sigma_{k+1} & \text{for the } ||\cdot||_2 \text{ norm} \\ \left(\sum_{i=k+1}^{r} \sigma_i^2\right)^{\frac{1}{2}} & \text{for the } ||\cdot||_F \text{ norm.} \end{cases}$$

Proof. The last part follows from Propositions 3.11 and 3.10.

Non-examinable: this is quite a tricky proof, but I've included it as its interesting to see. We'll just prove it for the 2-norm. Let B be an $n \times p$ matrix of rank k. The null space $\mathcal{N}(B) \subset \mathbb{R}^p$ must be of dimension $p - k$ by the rank nullity theorem.

Consider the $p \times (k+1)$ matrix $V_{k+1} = [v_1 \ ... \ v_{k+1}]$. This has rank $k+1$, and has column space $\mathcal{C}(V_{k+1}) \subset \mathbb{R}^p$. Because

$$\dim \mathcal{N}(B) + \dim \mathcal{C}(V_{k+1}) = p - k + k + 1 = p + 1$$

we can see that $\mathcal{N}(B)$ and $\mathcal{C}(V_{k+1})$ cannot be disjoint spaces (as they are both subsets of the p-dimensional space $\mathbb{R}^p$). Thus we can find $w \in \mathcal{N}(B) \cap \mathcal{C}(V_{k+1})$, and moreover we can choose w so that $||w||_2 = 1$.

Because $w \in \mathcal{C}(V_{k+1})$ we can write $w = \sum_{i=1}^{k+1} w_i v_i$ with $\sum_{i=1}^{k+1} w_i^2 = 1$.

Then

$$
\begin{aligned}
||A-B||_2^2 &\geq ||(A-B)w||_2^2 \quad \text{by definition of the matrix 2-norm} \\
&= ||Aw||_2^2 \quad \text{as } w \in \mathcal{N}(B) \\
&= w^\top V \Sigma^2 V^\top w \quad \text{using the SVD} A = U\Sigma V^\top \\
&= \sum_{i=1}^{k+1} \sigma_i^2 w_i^2 \quad \text{by substituting } w = \sum_{i=1}^{k+1} w_i v_i \\
&\geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} w_i^2 \quad \text{as } \sigma_1 \geq \sigma_2 \geq ... \\
&= \sigma_{k+1}^2 \quad \text{as } \sum_{i=1}^{k+1} w_i^2 = 1 \\
&= ||A - A_k||_2^2
\end{aligned}
$$

as required □

This best-approximation property is what makes the SVD so useful in applications.

3.6.3 Example: image compression

As an example, lets consider the image of some peppers from the USC-SIPI image database.

```
library(tiff)
library(rasterImage)
peppers<-readTIFF("figs/Peppers.tiff")
plot(as.raster(peppers))
```

This is a 512×512 colour image, meaning that there are three matrices R, B, G of dimension 512×512) giving the intensity of red, green, and blue for each pixel. Naively storing this matrix requires 5.7Mb.

We can compute the SVD of the three colour intensity matrices, and the view the image that results from using reduced rank versions B_k, G_k, R_k instead (as in Equation (3.3)). The image below is formed using $k = 5, 30, 100$, and 300 basis vectors.

```
svd_image <- function(im,k){
  s <- svd(im)
  Sigma_k <- diag(s$d[1:k])
  U_k <- s$u[,1:k]
  V_k <- s$v[,1:k]
  im_k <- U_k %*% Sigma_k %*% t(V_k)
   ## the reduced rank SVD produces some intensities <0 and >1.
  # Let's truncate these
  im_k[im_k>1]=1
  im_k[im_k<0]=0
  return(im_k)
}

par(mfrow=c(2,2), mar=c(1,1,1,1))

pepprssvd<- peppers
for(k in c(4,30,100,300)){
  svds<-list()
  for(ii in 1:3) {
    pepprssvd[,,ii]<-svd_image(peppers[,,ii],k)
  }
  plot(as.raster(pepprssvd))
}
```

You can see that for $k = 30$ we have a reasonable approximation, but with some errors. With $k = 100$ it is hard to spot the difference with the original. The size of the four compressed images is 45Kb, 345Kb, 1.1Mb and 3.4Mb.

You can see further demonstrations of image compression with the SVD here.

We will see much more of the SVD in later chapters.

PART II: Dimension reduction methods

In many applications, a large number of variables are recorded for each experimental unit under study. For example, if we think of individual people as the *experimental units*, then in a health check-up we might collect data on age, blood pressure, cholesterol level, blood test results, lung function, weight, height, BMI, etc. If you use websites such as Amazon, Facebook, and Google, they store thousands (possibly millions) of pieces of information about you (this article shows you how to download the information Google stores about you, including all the locations you've visited, every search, youtube video, or app you've used and more). They process this data to create an individual profile for each user, which they can then use to create targetted adverts.

When analysing data of moderate or high dimension, it is often desirable to seeks ways to restructure the data and reduce its dimension whilst **retaining the most important information within the data**. There a variety of reasons we might want to do this.

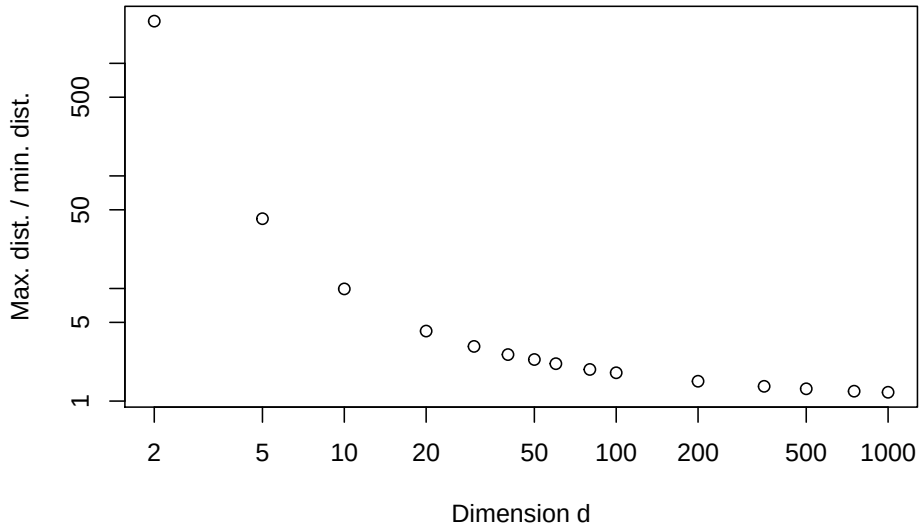
- In reduced dimensions, it is often much easier to understand and appreciate the most important features of a dataset.
- If there is a lot of reduncancy in the data, we might want to reduce the dimension to lower the memory requirements in storing it (e.g. with sound and image compression).
- In high dimensions, it can be difficult to analyse data (e.g. with statistical methods), and so reducing the dimension can be a way to make a dataset amenable to analysis.

In this part of the module we investigate three different methods for dimension reduction: Principal Component Analysis (PCA) in Chapter 4; Canonical Correlation Analysis (CCA) in Chapter 5; and Multidimensional Scaling (MDS) in Chapter 6. Matrix algebra (Chapters 2 and 3) plays a key role in all three of these techniques.

A warning

Beware that high-dimensional data can behave qualitatively differently to low-dimensional data. As an example, lets consider 1000 points uniformly distributed in $[0,1]^d$, and think about how close together or spread out the points are. A simple way to do this is to consider the ratio of the maximum and minimum distance between any two points in our sample.

```
N<-1000
averatio <-c()
ii<-1
for(d in c(2,5,10,20,30,40,50,60,80,100, 200, 350, 500, 750, 1000)){
  averatio[ii] <- mean(replicate(10, {
  X<-matrix(runif(N*d), nc=d)
  d <- as.matrix(dist(X))
  # this gives a N x N matrix of the Euclidean distances between the data points.
  maxdist <- max(d)
  mindist <- min(d+diag(10^5, nrow=N))
  # The diagonal elements of the distance matrix are zero,
  # so I've added a big number to the diagonal
  # so that we get the minimum distance between different points
  maxdist/mindist}))
  ii <- ii+1
}
plot(c(2,5,10,20,30,40,50,60,80,100, 200, 350, 500, 750, 1000),
     averatio, ylab='Max. dist. / min. dist.', xlab='Dimension d', log='xy')
```



So we can see that as the dimension increases, the ratio of the maximum and minimum distance between any two random points in our sample tends to 1. In other words, all points are the same distance apart!

3.6.4 Why reduce dimension?

Chapter 4

Principal component analysis

With multivariate data, it is common to 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.

4.1 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 $QQ^\top = Q^\top Q = \mathbf{I}_p$) 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 **??**,

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

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:

$$\max_{u:\, ||u||=1} u^\top Su \tag{4.1}$$

subject to

$$q_k^\top u = 0, \qquad k = 1, \dots, j-1. \tag{4.2}$$

It turns out that the maximum of (4.1) subject to (4.2) 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 4.1. *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 3.3)*

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

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* (4.1) *subject to* (4.2) *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 4.1. 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 $|S - \lambda \mathbf{I}| = 0$, where

$$\begin{aligned} |S - \lambda \mathbf{I}_2| &= (162.04 - \lambda)(175.36 - \lambda) - 135.38^2 \\ &= \lambda^2 - 337.4\lambda + 10887.59. \end{aligned}$$

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 $(S - \lambda_1 \mathbf{I}_2)q_1 = 0$. To simplify, we use row operations:

$$\begin{aligned} S - \lambda_1 \mathbf{I}_2 = \begin{pmatrix} -142.20 & 135.38 \\ 135.38 & -128.88 \end{pmatrix} &\rightarrow \begin{pmatrix} 1 & -0.952 \\ 135.38 & -128.88 \end{pmatrix} \\ &\rightarrow \begin{pmatrix} 1 & -0.952 \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

If we let $q_1 = (q_{11}, q_{21})^\top$ then solving $(S - \lambda_1 \mathbf{I}_2)q_1 = 0$ 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

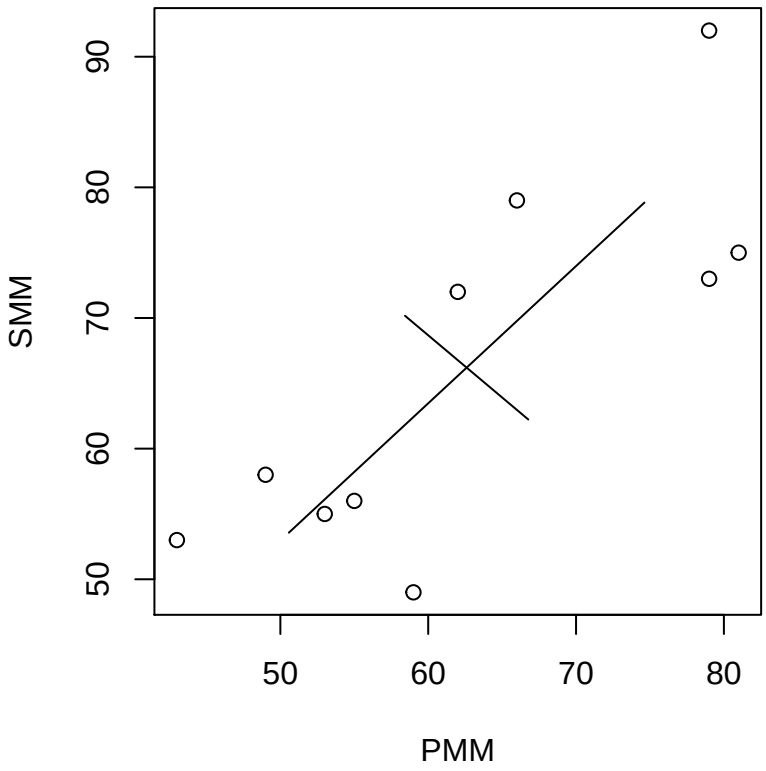
$$t = \pm \frac{1}{\sqrt{0.952^2 + 1^2}} = \pm 0.724.$$

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 $(S - \lambda_2 \mathbf{I}_2)q_2 = 0$ 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

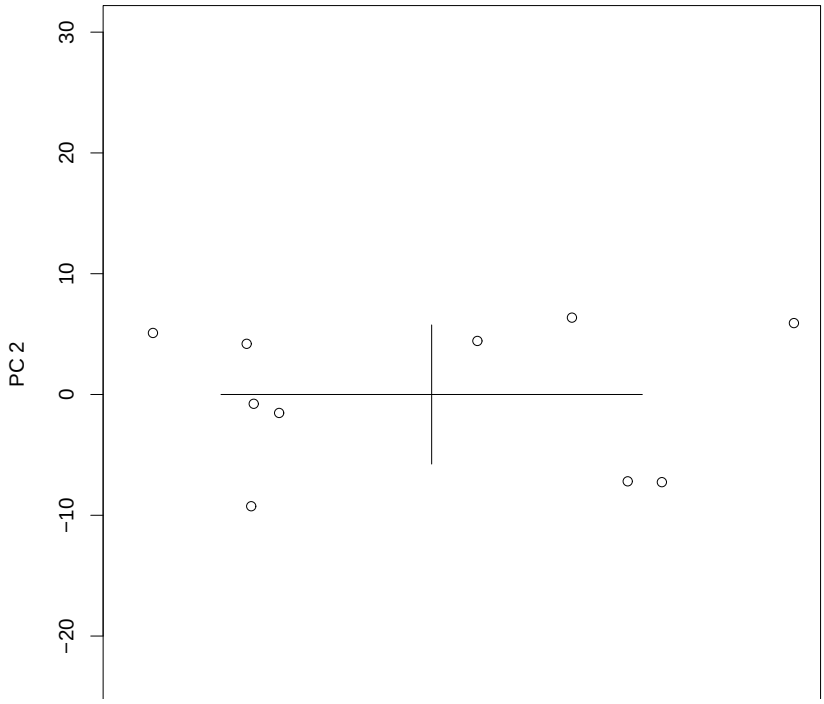
FIX FIX

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 4.1)

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

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. "'

4.2 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 4.2. *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

$$S = \begin{pmatrix} 39.4 & -8.27 & -31.1 & 116 & -81.9 \\ -8.27 & 8.14 & 0.13 & -29.4 & 6.01 \\ -31.1 & 0.13 & 31 & -86.3 & 75.9 \\ 116 & -29.4 & -86.3 & 392 & -209 \\ -81.9 & 6.01 & 75.9 & -209 & 231 \end{pmatrix} \tag{4.3}$$

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:

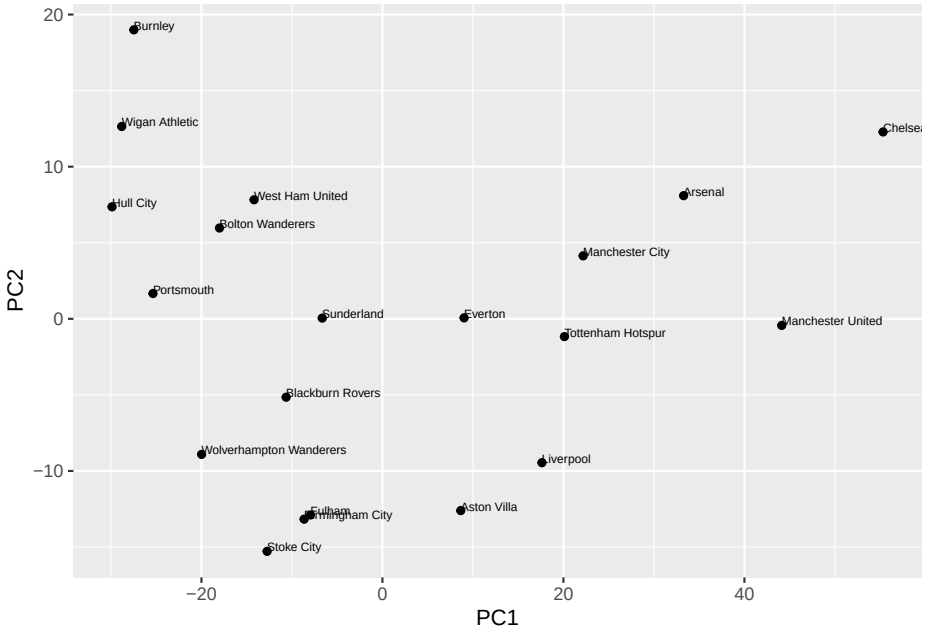
$$\begin{aligned} y_{i1} = 0.25(W_i - \bar{W}) + -0.05(D_i - \bar{D}) + -0.2(L_i - \bar{L}) \\ + 0.78(F_i - \bar{F}) + -0.54(A_i - \bar{A}), \end{aligned}$$

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.

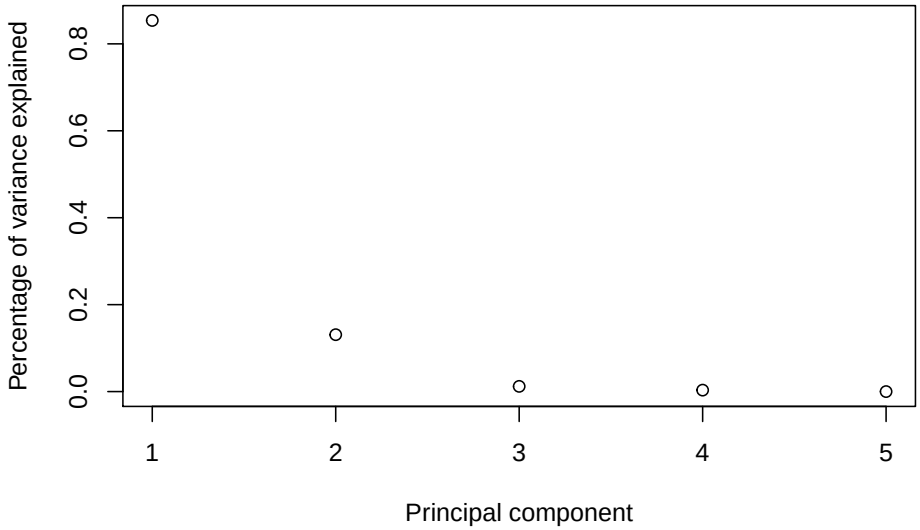
We could, therefore, rank teams by PC 1 and compare this with the rankings using 3 points for a win and 1 point for a draw. The rankings are the same for the top three teams but differ below that. Under our system Wigan would be relegated in place of Portsmouth.

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.

"'

4.3 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 4.2 are now given. Note that the Y_j's are random variables as opposed to observed values of random variables.

Proposition 4.3. *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 4.2. Suppose

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

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.

4.4 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:

$$\begin{aligned} &\textit{minimise } F(a, b; u_1, \ldots, u_n) \textit{ subject to the} \\ &\textit{constraints that } a \textit{ and } b \textit{ are orthogonal, i.e. } a^\top b = 0, \\ &\textit{and } b \textit{ is a unit vector, i.e. } ||b|| = 1. \end{aligned} \tag{4.4}$$

Theorem 4.1. *The solution to optimisation problem* (4.4) *is given by*

$$\hat{a} = \left(\mathbf{I}_p - q_1 q_1^\top\right) \bar{x}, \quad \hat{b} = q_1 \quad \textit{and} \quad \hat{u}_i = x_i^\top q_1, \quad i = 1, \ldots, n, \tag{4.5}$$

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 $i = 1, \ldots, n$,

$$\frac{\partial \bar{F}}{\partial u_i} = -2 b^\top (x_i - a - u_i b); \tag{4.6}$$

$$\begin{aligned} \frac{\partial \bar{F}}{\partial a} &= -2 \left\{ \sum_{i=1}^n (x_i - a - u_i b) \right\} + \gamma b \\ &= -2n \{ \bar{x} - a - (\bar{u} + \gamma/(2n)) b \}, \end{aligned} \tag{4.7}$$

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

$$\frac{\partial \bar{F}}{\partial \gamma} = a^\top b. \tag{4.8}$$

Setting the partial derivatives (4.6), (4.7) and @ref(eq:barF3}) to zero,

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

$$\frac{\partial \bar{F}}{\partial u_i} = 0 \implies \hat{u}_i = b^\top x_i, \tag{4.9}$$

and therefore

$$\bar{\hat{u}} \equiv n^{-1} \sum_{i=1}^{n} \hat{u}_i = b^\top \bar{x}; \tag{4.10}$$

and

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

Using (4.10) 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,

$$\hat{a} = \bar{x} - \bar{\hat{u}}b = \bar{x} - bb^\top \bar{x} = \left(\mathbf{I}_p - bb^\top\right)\bar{x}; \tag{4.11}$$

and so

$$\begin{aligned}
&\bar{F}(\hat{a}; \hat{u}_1, \ldots, \hat{u}_n; \hat{\gamma}) \qquad (4.12)\\
&= \sum_{i=1}^{n} (x_i - \hat{a} - \hat{u}_i b)^\top (x_i - \hat{a} - \hat{u}_i b)\\
&= \sum_{i=1}^{n} \left\{x_i - (\mathbf{I}_p - bb^\top)\bar{x} - bb^\top x_i\right\}^\top \left\{x_i - (\mathbf{I}_p - bb^\top)\bar{x} - bb^\top x_i\right\}\\
&\sum_{i=1}^{n} (x_i - \bar{x})^\top (\mathbf{I}_p - bb^\top)^2 (x_i - \bar{x})\\
&= n\mathrm{tr}\left\{(\mathbf{I}_p - bb^\top)S\right\}\\
&= n\left\{\mathrm{tr}(S) - b^\top S b\right\}, \qquad (4.13)
\end{aligned}$$

where S is the sample covariance of the x_i.

Step 2. We now minimise (4.13) over unit vectors $b \in \mathbb{R}^p$. But minimising (4.13) is equivalent to maximising $b^\top S b$, so from Proposition 3.7, $\hat{b} = q_1$, and so $\hat{a} = (\mathbf{I}_p - q_1 q_1^\top)\bar{x}$ from (4.11), and from (4.9), $\hat{u}_i = q_1^\top x_i$, $i = 1, \ldots, n$, all of which agrees with the expressions in (4.5).

□

4.5 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 4.3. 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 $AA^\top = A^\top A = \mathbf{I}_p$, 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 4.4. 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.

4.6 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 5

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

$$r = r[x, y] = \frac{n^{-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\left(n^{-1} \sum_{i=1}^n (x_i - \bar{x})^2\right)^{1/2} \left(n^{-1} \sum_{i=1}^n (y_i - \bar{y})^2\right)^{1/2}} \tag{5.1}$$

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 5.1. 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?

5.1 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 5.1 (continued). The relevant covariance matrix here is given in (4.3), 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

$$S_{xx} = \begin{pmatrix} 39.4 & -8.3 \\ -8.3 & 8.1 \end{pmatrix}, \qquad S_{yy} = \begin{pmatrix} 392.2 & -208.7 \\ -208.7 & 230.9 \end{pmatrix} \tag{5.2}$$

and

$$S_{xy} = S_{yx}^\top = \begin{pmatrix} 115.7 & -81.9 \\ -29.4 & 6.0 \end{pmatrix}. \tag{5.3}$$

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:

$$\begin{aligned} &\text{Find non-zero vectors} \quad a \ (p \times 1) \text{ and } b \ (q \times 1) \\ &\text{to maximise} \qquad r[a^\top x, b^\top y], \end{aligned} \tag{5.4}$$

where $r[.,.]$ is defined in (5.1). 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 $\delta > 0$,

$$r[\gamma a^\top x, \delta b^\top y] = \frac{\gamma\delta}{\sqrt{\gamma^2\delta^2}} r[a^\top x, b^\top y] = r[a^\top x, b^\top y], \tag{5.5}$$

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 (5.4), 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

$$\max_{a,b} a^\top S_{xy} b \tag{5.6}$$

subject to the constraints

$$a^\top S_{xx} a = 1 \qquad \text{and} \qquad b^\top S_{yy} b = 1. \tag{5.7}$$

Proposition 5.1. *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* (5.4)*, 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 (5.6) *subject to the constraints* (5.7)*. Moreover, if $a = \breve{a}$ and $b = \breve{b}$ maximise* (5.6) *subject to constraints* (5.7) *then, for any $\gamma > 0$ and $\delta > 0$, $a = \gamma\breve{a}$ and $b = \delta\breve{b}$ maximise* (5.4)*.*

2. *The optimum solution to* (5.6) *and* (5.7) *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*

$$A \equiv S_{xx}^{-1/2}S_{xy}S_{yy}^{-1/2} = \sum_{j=1}^{t} \xi_j \mathbf{q}_j \mathbf{r}_j^\top \equiv \mathbf{Q}\Xi\mathbf{R}^\top, \tag{5.8}$$

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 (5.5) 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 (5.4) 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 $a = \breve{a} \equiv \hat{a}/(\hat{a}^\top S_{xx}\hat{a})^{1/2}$ and $b = \breve{b} \equiv \hat{b}/(\hat{b}^\top S_{yy}\hat{b})^{1/2}$. But by definition of $\breve{a}$ and $\breve{b}$, they satisfy the constraints (5.7) because

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

and, similarly,

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

So $a = \breve{a}$ and $b = \breve{b}$ maximises (5.6) subject to the constraints (5.7).

(ii) & (iii) We may write the constraints (5.7) 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 (5.6) subject to (5.7) 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 5.1 (continued) We now want to calculate the matrix A in (5.8) 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

$$\begin{aligned} A &= Q\Xi R^\top \\ &= \begin{pmatrix} -0.997 & 0.082 \\ 0.082 & 0.997 \end{pmatrix} \begin{pmatrix} 0.974 & 0 \\ 0 & 0.508 \end{pmatrix} \begin{pmatrix} -0.790 & -0.613 \\ 0.613 & -0.790 \end{pmatrix}^\top . \end{aligned} \tag{5.9}$$

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.

5.2 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:

$$\max_{a,\, b} a^\top S_{xy} b \tag{5.10}$$

subject to the constraints

$$a^\top S_{xx} a = 1, \qquad b^\top S_{yy} b = 1, \tag{5.11}$$

$$a_1^\top S_{xx} a = 0 \qquad \text{and} \qquad b_1^\top S_{yy} b = 0. \tag{5.12}$$

Note that maximising (5.10) subject to (5.11) is very similar to the optimisation problem (5.6) and (5.7) considered in the previous section. What is new are the constraints (5.12), 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 (5.12) 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 (5.12) as corresponding to orthogonality constraints (cf. PCA) in modified coordinate systems.

We now discuss the optimisation of (5.10), (5.11) and (5.12). 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 (5.10) subject to constraints (5.11) and (5.12) 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 (5.11) and (5.12) by, respectively, (5.13) and (5.14) below, where

$$a^\top S_{xx} a = 1, \qquad b^\top S_{yy} b = 1, \tag{5.13}$$

$$a_j^\top S_{xx} a = 0 \qquad \text{and} \qquad b_j^\top S_{yy} b = 0, \qquad j = 1, \ldots, k-1. \tag{5.14}$$

Then the optimisation problem is

$$\max_{a,\, b} a^\top S_{xy} b \tag{5.15}$$

subject to constraints (5.13) and (5.14). The solution in the general case is as follows.

- The maximum of (5.15) subject to constraints (5.13) and (5.14) 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 5.2. *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 5.1 (continued) From (5.9), 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.

5.3 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 (5.4). However, it does not usually solve the optimisation problem defined by problems (5.6) and (5.7) because typically it will not be the case that $\hat{\boldsymbol{\beta}}^\top S_{xx} \hat{\boldsymbol{\beta}} = 1$, so that (5.7) will not be satisfied.

5.4 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 $\check{\xi}_j$.

5.5 Invariance/equivariance properties of CCA

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

$$\mathbf{h}_i = \mathbf{T}x_i + \boldsymbol{\mu} \qquad \text{and} \qquad \mathbf{k}_i = \mathbf{V}y_i + \boldsymbol{\eta}, \qquad i = 1, \ldots, n, \tag{5.16}$$

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

$$A \equiv A_{xy} = S_{xx}^{-1/2}S_{xy}S_{yy}^{-1/2}. \tag{5.17}$$

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 (5.17) 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 (5.16). 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 (5.16), 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 (5.16) is not an orthogonal matrix then the singular values are not invariant and the cc vectors do not transform in an equivariant manner.

5.6 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 5.1 (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 6

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.

6.1 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

$$d_{ik} \leq d_{ij} + d_{jk}. \tag{6.1}$$

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 (6.1), it follows that Euclidean distance is a metric distance.

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

$$A = \{a_{ij}\}_{i,j=1}^n, \quad \text{where} \qquad a_{ij} = -\frac{1}{2} d_{ij}^2. \tag{6.2}$$

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

Now define the matrix

$$\mathbf{B} = \mathbf{H}A\mathbf{H}, \tag{6.3}$$

where

$$\mathbf{H} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top \tag{6.4}$$

is the $n \times n$ **centering matrix**; see §2.7. For reasons that will soon become clear, B defined by (6.3) 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 6.1. *Let D denote an $n \times n$ distance matrix and suppose A, B and $\mathbf{H}$ be as defined in* (6.2)*,* (6.3) *and* (6.4)*, 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*

$$b_{ij} = (x_i - \bar{x})^\top (x_j - \bar{x}), \qquad i, j = 1, \ldots, n, \tag{6.5}$$

   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* (6.5)*.*

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 6.1 may be useful even if $\mathbf{D}$ 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 6.1. 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 (6.2) first to calculate A, and then using (6.3) 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 6.1, 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.

6.2 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 6.1 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 (6.2).
- **Step 2**: Calculate $B = (b_{ij})_{i,j=1}^n$ in (6.3) 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 6.2. *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.

6.3 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 (6.1), and non-metric distances, or dissimilarities, which need not satisfy (6.1).

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 6.2. 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

$$\frac{a+d}{a+b+c+d}. \tag{6.6}$$

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

Example 6.3. 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 (6.6) 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

$$d_{ij} = \left(f_{ii} + f_{jj} - 2f_{ij}\right)^{1/2}, \qquad i, j = 1, \ldots, n. \tag{6.7}$$

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 6.3. *Suppose that F is a similarity matrix. If, in addition, F is non-negative definite, then D defined in* (6.7) *is Euclidean with centred inner product matrix*

$$B = HFH, \tag{6.8}$$

where $H = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top$ 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 (6.8) holds, where B is given by $B = HAH$ and A is defined as in (6.2), and the d_{ij} are defined by (6.7). 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. □ □