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.

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}

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 **??**.

1.2 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 $D = \text{diag}(\sqrt{s_{11}}, ..., \sqrt{s_{pp}})$, 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,

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

In other words, it is the sum of the univariate variances of each of the p variables.

Example 1.4. 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
```

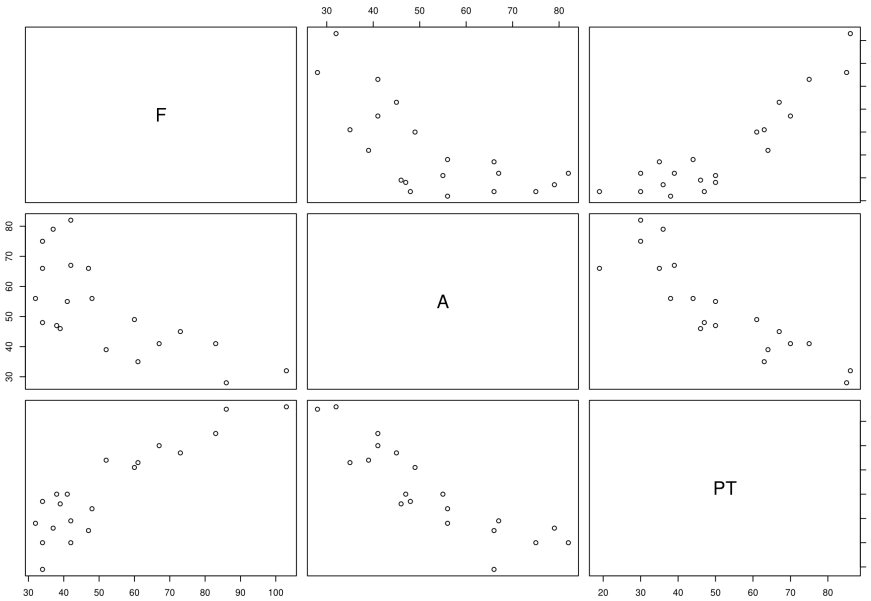
```
## [1] 494.4
```

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 Graphical techniques

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



You can also use the `plot3d` command in the `rgl` library to create an interactive 3D plot of the data. The difficulty of displaying multivariate data is further motivation for developing a method for reducing the number of dimensions in the data.

1.4 Random Vectors and Matrices

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.

1.5 Unbiased Estimators

Definition 1.5. For a staistical model $p(x \mid \theta)$, a statistic $\hat{\theta} \equiv \hat{\theta}(X)$ is said to be an **unbiased estimator** of θ if $\mathbb{E}_{x|\theta}(\hat{\theta}(X)) = \theta$ for all θ.

I.e., if data are generated with parameter value θ, then an estimator $\hat{\theta}$ is an unbiased estimator of θ if its expected value matches the true value (the value used to generate the data). This concept readily transfers to the multivariate context.

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

1. $\mathbb{E}(\bar{x}) = \mu$.
2. $\mathbb{V}\,\mathrm{ar}(\bar{x}) = \frac{1}{n}\Sigma$.
3. $\mathbb{E}(S) = \frac{n-1}{n}\Sigma$.

Proof. **Part 1** By the linearity of expectation,

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

Part 2 We have

$$\begin{aligned}
\mathbb{V}\,\mathrm{ar}(\bar{x}) &= \mathbb{V}\,\mathrm{ar}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\
&= \sum_{i,j=1}^{n} \mathbb{C}\,\mathrm{ov}\left(\frac{1}{n}x_i, \frac{1}{n}x_j\right) \\
&= \sum_{i=1}^{n} \mathbb{V}\,\mathrm{ar}\left(\frac{1}{n}x_i\right) + \sum_{i\neq j} \mathbb{C}\,\mathrm{ov}\left(\frac{1}{n}x_i, \frac{1}{n}x_j\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} \mathbb{V}\,\mathrm{ar}(x_i) + \sum_{i\neq j} \mathbb{C}\,\mathrm{ov}(x_i, x_j)\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} \mathbb{V}\,\mathrm{ar}(x_i)\right) \text{ as } x_i \perp\!\!\!\perp x_j \text{ for } i \neq j \\
&= \frac{1}{n^2} n\Sigma \\
&= \frac{1}{n}\Sigma.
\end{aligned}$$

Part 3 From the definition of the sample covariance,

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

as $\frac{1}{n}\sum(\bar{x}-\mu)(x_i-\mu)^\top = \sum(\bar{x}-\mu)(\bar{x}-\mu)^\top$. Since $\mathbb{E}(\bar{x}) = \mu$ and $\mathbb{V}\,\mathrm{ar}(\bar{x}) = n^{-1}\Sigma$, it follows that

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

which completes the proof.

□

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

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

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 **??** and **??**.

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.