# Advanced Statistics and Alternative Data Types

GRA 4153

Adam Lee*

August 26, 2025

*Department of Data Science & Analytics, BI Norwegian Business School; adam.lee@bi.no

# Introduction

## Preface

These are lectures notes for the course "Advanced Statistics and Alternative Data Types".[1]

The course covers four main topics:

2 - Basic probability theory

3 - Statistical inference

4 - (Generalised) Linear models

5 - Time series analysis

Section 1 is a review of some topics in linear algebra in Euclidean spaces which are necessary for the course. This will not be covered in the lectures.

These notes include some results, proofs etc. which are included for the sake of completeness, but are more technical / difficult than the intended level of the course and may be skipped.[2] These are denoted by a preceding asterisk.[3] Nevertheless, there are cases where including a full proof of a result would require too much of a detour. In these cases, a reference to where a proof can be found in the literature is given instead.

Many of the proofs and examples have some details deliberately omitted which should be filled in by the reader. These are denoted by [Exercise] following a statement and correspond to an exercise in the list of practice exercises.

[1] If any typos are found, please let me know at adam.lee@bi.no.

[2] The level of difficulty of the starred results varies from close to the level of this course to *much* more difficult.

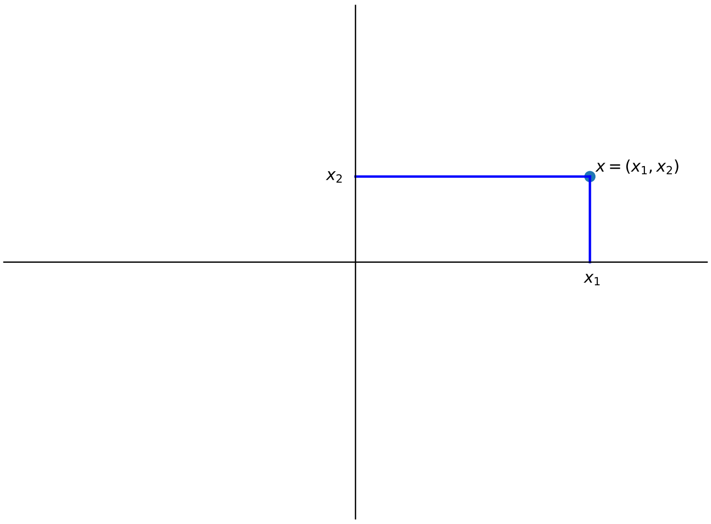
[3] Note that sometimes the statement of a result may not have an asterisk whilst its proof does.

# 1 Linear Algebra in $\mathbb{R}^n$

## 1.1 Euclidean spaces

We are familiar with the real number line, $\mathbb{R}$. Elements in the plane, $\mathbb{R}^2$, take the form of a pair $x = (x_1, x_2)$ with each $x_i \in \mathbb{R}$, i.e. each $x_i$ is a real number. This is an example of a *vector*.

Figure 1: A vector, $x$, in the plane, $\mathbb{R}^2$



The plane is two dimensional. Though it becomes more difficult to draw, we need not stop at 2 copies of $\mathbb{R}$. $\mathbb{R}^n = \mathbb{R} \underbrace{\times \cdots \times}_{n \text{ times}} \mathbb{R}$ is $n$-dimensional and its elements consist of vectors $x = (x_1, \ldots, x_n)$ where each $x_i \in \mathbb{R}$.

We can add vectors (of the same dimension) together: if $x, y \in \mathbb{R}^n$ then $x + y = (x_1 + y_1, \ldots, x_n + y_n)$ is also a vector in $\mathbb{R}^n$. We can also multiply vectors by scalars, that is by elements of $\mathbb{R}$: if $\lambda \in \mathbb{R}$ and $x \in \mathbb{R}^n$, $\lambda x = (\lambda x_1, \ldots, \lambda x_n)$.

We can measure the length of vectors in $\mathbb{R}^n$: the (Euclidean) *norm* of a vector $x \in \mathbb{R}^n$ is $\|x\| = \left(\sum_{i=1}^n x_i^2\right)^{1/2}$. Note that in the case of $n = 1$, $\|x\| = |x|$, the absolute value. Related to this is the concept of the *dot product*: if we have two vectors $x, y \in \mathbb{R}^n$, then their dot product is $x \cdot y = \sum_{i=1}^n x_i y_i$. Note that the norm $\|x\| = \sqrt{x \cdot x}$.

A sequence of vectors $x_1, x_2, \ldots$ is often denoted $(x_m)_{m \in \mathbb{N}}$. Here it is understood that each $x_m \in \mathbb{R}^n$. A sequence $(x_m)_{n \in \mathbb{N}}$ converges to a point (that is, a vector) $x \in \mathbb{R}^n$, if for any $\varepsilon > 0$, there is some $M$ such for all $m \geq M$, $\|x_m - x\| < \varepsilon$. Notationally we write this as $x_m \to x$ or $\lim_{m \to \infty} x_m = x$.

## 1.2 Span, linear independence and bases

A *linear combination* of a list of vectors $x_1, \ldots, x_m$ in $\mathbb{R}^n$ is a vector which has the form $\lambda_1 x_1 + \cdots + \lambda_m x_m$ for some scalars $\lambda_1, \ldots, \lambda_m$, each in $\mathbb{R}$.[4]

[4] Note that here the subscript $i$ in $x_i$ refers to the position of the vector in the list of *vectors*, not the $i$-th element of a vector $x$.

The *span* of a list of vectors is the set of all linear combinations of those vectors:

$$\mathrm{span}(x_1, \ldots, x_m) = \{\lambda_1 x_1 + \cdots + \lambda_m x_m : \lambda_1, \ldots, \lambda_m \in \mathbb{R}\}.$$

Example 1.1: The span of the vectors $e_1 = (1, 0, 0, \ldots, 0, 0)$, $e_2 = (0, 1, 0, \ldots, 0, 0)$, $\ldots$, $e_n = (0, 0, 0, \ldots, 0, 1)$ (in $\mathbb{R}^n$) is $\mathbb{R}^n$. △

Example 1.2: $x_1 = (1, 2)$ and $x_2 = (-1/2, -1)$ have $\mathrm{span}(x_1, x_2) = \{(z, 2z) : z \in \mathbb{R}\} \subset \mathbb{R}^2$. △

A list of vectors $x_1, \ldots, x_m$ is *linearly independent* if

$$\lambda_1 x_1 + \ldots + \lambda_m x_m = 0 \implies \lambda_1 = \cdots = \lambda_m = 0.$$

A list of vectors $x_1, \ldots, x_m$ is *linearly dependent* if it is not linearly independent. That means there is some $\lambda_1, \ldots, \lambda_m \in \mathbb{R}$, not all zero, such that $\lambda_1 x_1 + \ldots + \lambda_m x_m = 0$. Therefore, if $\lambda_k \neq 0$, we can write $x_k = \sum_{i=1, i \neq k}^{m} \frac{-\lambda_i}{\lambda_k} x_i$, so $x_k$ is a linear combination of the other $m - 1$ vectors.

The vector $0 \in \mathbb{R}^n$ (whose elements are all $0 \in \mathbb{R}$) is special: a list of one vector $x$ is linearly independent if and only if $x \neq 0$. Moreover, any list of vectors containing 0 is linearly dependent.

Example 1.3: The vectors $e_1, \ldots, e_n$ in $\mathbb{R}^n$ from example 1.1 are linearly independent. △

Example 1.4: In $\mathbb{R}^2$, the vectors $x_1 = (1, 2)$, $x_2 = (-1/2, -1)$ from example 1.2 are linearly dependent as $x_1 + 2x_2 = 0$. △

There can be no more than $n$ linearly independent vectors in $\mathbb{R}^n$ and to span $\mathbb{R}^n$ we must have at least $n$ vectors. This fact follows from the following fundamental lemmas.

Lemma 1.1: *If $x_1, \ldots, x_m$ is linearly dependent and $x_1 \neq 0$, then there is a $j \in \{1, \ldots, m\}$ such that*

*(i) $x_j \in \mathrm{span}(x_1, \ldots, x_{j-1})$;*

*(ii) If $x_j$ is removed from the list $x_1, \ldots, x_m$, the span of the remaining list is $\mathrm{span}(x_1, \ldots, x_m)$.*

*Proof.* There exist $\lambda_1, \ldots, \lambda_m$, not all zero, such that

$$\lambda_1 x_1 + \cdots + \lambda_m x_m = 0.$$

Since $x_1 \neq 0$, at least one of $\lambda_2, \ldots, \lambda_m$ must be non-zero [why?]. Let $j := \max\{k \in \{2, \ldots, m\} : \lambda_k \neq 0\}$. Then

$$x_j = -\frac{\lambda_1}{\lambda_j} x_1 - \cdots - \frac{\lambda_{j-1}}{\lambda_j} x_{j-1}, \tag{1}$$

hence $x_j \in \mathrm{span}(x_1, \ldots, x_{j-1})$. For the second part, let $v \in \mathrm{span}(x_1, \ldots, x_m)$. Then there are $\gamma_1, \ldots, \gamma_m$ such that

$$v = \gamma_1 x_1 + \cdots + \gamma_m x_m.$$

Substitute (1) into this equation to obtain

$$v = (\gamma_1 - \lambda_1\gamma_j\lambda_j^{-1})x_1 + \cdots + (\gamma_{j-1} - \lambda_{j-1}\gamma_j\lambda_j^{-1})x_{j_1} + \gamma_{j+1}x_{j+1} + \cdots \gamma_m x_m,$$

i.e. $v \in \operatorname{span}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_m)$. □

Lemma 1.2: *The length of every list of linearly independent vectors is less than or equal to the length of every spanning list of vectors.*

*Proof.* First suppose that $x_1, \ldots x_m$ is linearly independent and $w_1, \ldots, w_n$ is a spanning list. It suffices to show that $m \le n$. We will cycle through this in steps:

Step 1: Let $B_1$ be the spanning list $w_1, \ldots, w_n$. Adding any other vector gives us a linearly dependent list, since such a vector is a linear combination of elements of $B_1$. It is therefore possible to replace one of the original elements $e_k$ of $B_1$ with $x_1$ to create $B_2$ which is also a spanning list (by Lemma 1.1). Note that $B_2$ is still of length $n$.

Step $j$: The list $B_{j-1}$ from the previous step is a spanning list of length $n$. It contains $x_1, \ldots, x_{j-1}$ and some elements from $\{w_1, \ldots, w_n\}$. Adding $x_j$ to $B_{j-1}$ makes the list linearly dependent and we can remove one of the remaining $w_k$ vectors to form a new spanning list $B_j$ which is of length $n$.

Once we have reached step $m$ we have added all the $x_k$'s. At each step there was a $w_j$ to remove. Hence $n \ge m$. □

Proposition 1.1: *A list of $m > n$ vectors in $\mathbb{R}^n$ is linearly dependent and a list of $m < n$ vectors in $\mathbb{R}^n$ does not span $\mathbb{R}^n$.*

*Proof.* First suppose that $x_1, \ldots, x_m$ is linearly independent. We know that $e_1, \ldots, e_n$ spans $\mathbb{R}^n$ from Example 1.1. The first claim then follows from Lemma 1.2.

For the second statement, suppose that $x_1, \ldots, x_m$ is a spanning list of vectors in $\mathbb{R}^n$. $e_1, \ldots, e_n$ is linearly independent by Example 1.3. By Lemma 1.2 $n \le m$. □

A *basis* for $\mathbb{R}^n$ is a list of vectors that is linearly independent and spans all of $\mathbb{R}^n$.

Example 1.5: The vectors $e_1, \ldots, e_n$ in $\mathbb{R}^n$ form a basis for $\mathbb{R}^n$ by Examples 1.1 and 1.3. This list of vectors is often called the *canonical* basis of $\mathbb{R}^n$. △

Any basis for $\mathbb{R}^n$ must be comprised of exactly $n$ vectors (as follows from Proposition 1.1). The following proposition provides a useful characterisation of a basis.

Proposition 1.2: *A list $x_1, \ldots, x_n$ of vectors in $\mathbb{R}^n$ is a basis if and only if every $x \in \mathbb{R}^n$ can be written uniquely as*

$$x = \lambda_1 x_1 + \cdots + \lambda_n x_n, \quad \lambda_1, \ldots, \lambda_n \in \mathbb{R}.$$

*Proof.* Suppose $x_1, \ldots, x_n$ is a basis of $\mathbb{R}^n$ and let $x \in \mathbb{R}^n$. Since our list of vectors spans $\mathbb{R}^n$, there are $\lambda_1, \ldots, \lambda_n$ such that $x = \lambda_1 x_1 + \cdots + \lambda_n x_n$. We need to show this representation is unique. Suppose that also $x = \gamma_1 x_1 + \cdots \gamma_n x_n$ with each $\gamma_i \in \mathbb{R}$. Subtract the second equation for $x$ from the first to get

$$0 = (\lambda_1 - \gamma_1)x_1 + \cdots + (\lambda_n - \gamma_n)x_n.$$

By linear independence of $x_1, \ldots, x_n$, this implies that each $\lambda_i - \gamma_i = 0$ and hence each $\lambda_i = \gamma_i$ as required.

Now suppose instead that every $x \in \mathbb{R}^n$ can be written uniquely as $x = \lambda_1 x_1 + \cdots + \lambda_n x_n$. Evidently $x_1, \ldots, x_n$ spans $\mathbb{R}^n$, since we can write each $x \in \mathbb{R}^n$ as a linear combination of $x_1, \ldots, x_n$. For linear independence, suppose the $\lambda_i$ are such that $0 = \lambda_1 x_1 + \cdots + \lambda_n x_n$. Then, since this representation is unique, and clearly $\lambda_i = 0$ for all $i = 1, \ldots, n$ would lead to a linear combination equalling zero, it must be that each $\lambda_i = 0$ and hence the vectors are linearly independent. □

## 1.3 Matrices and linear transformations

A *matrix* is a two – dimensional array of real numbers.[5] An $m \times n$ matrix has $m$ rows and $n$ columns:

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix}.$$

If $m = n$, the matrix is said to be *square*.

### 1.3.1 Linear transformations

A *linear function* or *linear transformation* from $\mathbb{R}^n \to \mathbb{R}^m$ is a function, $T$, which satisfies (i) additivity:

$$T(x + y) = T(x) + T(y) \quad \text{for all } x, y \in \mathbb{R}^n,$$

and (ii) homogeneity:

$$T(\lambda x) = \lambda T(x) \quad \text{for all } x \in \mathbb{R}^n \text{ and all } \lambda \in \mathbb{R}.$$

Given bases $v_1, \ldots, v_n$ of $\mathbb{R}^n$ and $w_1, \ldots, w_m$ of $\mathbb{R}^m$ we can represent a linear transformation $T$ from $\mathbb{R}^n \to \mathbb{R}^m$ as a $m \times n$ matrix. It is the matrix $A$ such that

$$T(v_k) = A_{1k} w_1 + \cdots + A_{mk} w_m. \tag{2}$$

Usually we (implicitly) use the canonical bases of $\mathbb{R}^n$ and $\mathbb{R}^m$ and represent linear transformations as matrices.[6] In this case we can form the matrix as follows: let $a_k = T(e_k)$ for $k = 1, \ldots, n$

[5] At least, a *real* matrix has real entries; we shall not consider any other kinds of matrix in this course.

[6] The converse is also true: given bases $v_1, \ldots, v_n$ of $\mathbb{R}^n$ and $w_1, \ldots, w_m$ of $\mathbb{R}^m$, a $m \times n$ matrix $A$ defines a linear transformation, defined as

$$T(v_k) \coloneqq A_{1k} w_1 + \cdots + A_{mk} w_m.$$

This equation is essentially the same as that given in the main text; the difference is that here it defines the and let

$$A = [a_1, \ldots, a_n] = [T(e_1), \ldots, T(e_n)]\,, \tag{3}$$

i.e. the columns of the matrix $A$ are $a_k = T(e_k)$ for $k = 1, \ldots, n$.

### 1.3.2 Matrix addition, scalar multiplication,transposition and matrix multiplication

Matrices can be added together or multiplied by a scalar. These operations are defined elementwise. For $m \times n$ matrices $A, B$ and a scalar $\lambda \in \mathbb{R}$ we have

$$A+B = \begin{pmatrix} A_{11}+B_{11} & A_{12}+B_{12} & \cdots & A_{1n}+B_{1n} \\ A_{21}+B_{21} & A_{22}+B_{22} & \cdots & A_{2n}+B_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1}+B_{m1} & A_{m2}+B_{m2} & \cdots & A_{mn}+B_{mn} \end{pmatrix}, \quad \lambda A = \begin{pmatrix} \lambda A_{11} & \lambda A_{12} & \cdots & \lambda A_{1n} \\ \lambda A_{21} & \lambda A_{22} & \cdots & \lambda A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda A_{m1} & \lambda A_{m2} & \cdots & \lambda A_{mn} \end{pmatrix}.$$

Note that in order for $A$ and $B$ to be added together they must be of the same size. Matrix addition is *commutative*

$$A + B = B + A,$$

and *associative*:

$$(A + B) + C = A + (B + C).$$

Scalar multiplication is also *associative*:

$$(\lambda\gamma)A = \lambda(\gamma A).$$

Additionally these operations together satisfy two *distributive* properties:

$$\lambda(A + B) = \lambda A + \lambda B, \qquad (\lambda + \gamma)A = \lambda A + \gamma A.$$

Another important operation is the matrix transpose: this "flips" the matrix around the diagonal: if $A$ is a $m \times n$ matrix, $A'$ is a $n \times m$ matrix with

$$A' = \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{m1} \\ A_{12} & A_{22} & \cdots & A_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{mn} \end{pmatrix}.$$

A matrix $A$ is said to be symmetric if $A' = A$. Note that a symmetric matrix must be square, as otherwise $A$ and $A'$ would have different dimensions. The transpose operator satisfies the

*left hand side*, not the values of the $A_{i,k}$. Defining the values of $T(v_k)$ for $k = 1, \ldots, n$ is sufficient for a linear transformation since the $v_k$ form a basis and linear transformations are additive and homogeneous.

following properties:

$$(A')' = A, \qquad (\lambda A)' = \lambda A', \qquad (A+B)' = A' + B'.$$

The fourth operation is matrix multiplication. We can multiply two matrices $A$ and $B$ to form the product $AB$ if the number of columns of $A$ is the same as the number of rows of $B$. Such pairs of matrices are called *conformable.* Then, if $A$ is $m \times n$ and $B$ is $n \times p$, let $A_{i\bullet}$ denote the $i$-th row of matrix $A$ and $B_{\bullet j}$ the $j$-th column of matrix $B$; these are vectors in $\mathbb{R}^n$. $AB$ is $m \times p$ and

$$AB = \begin{pmatrix} A_{1\bullet} \cdot B_{\bullet 1} & A_{1\bullet} \cdot B_{\bullet 2} & \cdots & A_{1\bullet} \cdot B_{\bullet p} \\ A_{2\bullet} \cdot B_{\bullet 1} & A_{2\bullet} \cdot B_{\bullet 2} & \cdots & A_{2\bullet} \cdot B_{\bullet p} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m\bullet} \cdot B_{\bullet 1} & A_{m\bullet} \cdot B_{\bullet 2} & \cdots & A_{m\bullet} \cdot B_{\bullet p} \end{pmatrix},$$

that is, the $i, j$-th element of $AB$ is the dot product of the $i$-th row of $A$ and the $j$-th column of $B$. Note that $AB$ does *not* generally equal $BA$: in fact, unless $p = m$, this product does not even exist.[7] To be explicit, to refer to the product $AB$ we say that we pre-multiply $B$ by $A$ or that we post-multiply $A$ by $B$.

Matrix multiplication in conjunction with matrix addition, scalar multiplication and transposition satisify a number of properties. If $\lambda \in \mathbb{R}$ and the matrices $A, B, C$ are such that the matrix products below exist, then [exercise]

(i) $A(BC) = (AB)C$;

(ii) $A(B + C) = AB + AC$;

(iii) $(B + C)A = BA + CA$;

(iv) $\lambda(AB) = (\lambda A)B = A(\lambda B)$;

(v) $(AB)' = B'A'$.

We can also use the definition of matrix multiplication to express the dot product itself. If $A$ is a $n \times 1$ matrix and $B$ a $n \times 1$ matrix, we can consider them as vectors in $\mathbb{R}^n$ and the dot product of these vectors is equal to

$$A'B = A_{1\bullet} \cdot B_{\bullet 1}.$$

Matrix multiplication also allows us to represent linear transformations as multiplication by the associated matrix. Suppose that $T : \mathbb{R}^n \times \mathbb{R}^m$ and let $A$ be the $m \times n$ representing matrix (from (2)), with respect to the canonical basis. Then, we have that for any $x \in \mathbb{R}^n$, $x = \sum_{k=1}^n x_k e_k$ and so

$$T(x) = \sum_{k=1}^{n} x_k T(e_k) = \sum_{k=1}^{n} x_k a_k.$$

[7] If $A, B$ are matrices such that $AB = BA$, then $A$ and $B$ are said to *commute.*

Using the definition of $A$ in (3) and the definition of matrix multiplication we have

$$Ax = \begin{pmatrix} A_{1\bullet} \cdot x \\ A_{2\bullet} \cdot x \\ \vdots \\ A_{m\bullet} \cdot x \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{n} x_k A_{1k} \\ \sum_{k=1}^{n} x_k A_{2k} \\ \vdots \\ \sum_{k=1}^{n} x_k A_{mk} \end{pmatrix} = \sum_{k=1}^{n} x_k a_k.$$

### 1.3.3 Some important classes of matrices

There are a number of important classes of matrices; here we will consider four of them. The first, the zero matrices, are matrices $O$ of any dimension which have all their entries equal to zero. The product of any (conformable) matrix with a zero matrix is a zero matrix (possibly of a different dimension).

The second, the diagonal matrices, are square matrices, with non-zero elements only on their diagonal. That is, they have the form:

$$D = \begin{pmatrix} D_{11} & 0 & \cdots & 0 \\ 0 & D_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_{nn} \end{pmatrix}.$$

The third is the class of identity matrices. These are special cases of diagonal matrices, with all $D_{kk} = 1$:

$$I = I_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

Multiplying any (conformable) matrix $A$ by the identity matrix (either pre- or post-) results in the original matrix: $IA = A$ and $AI = A$ (whenever these products exist).

Finally we have the lower triangular matrices.[8] A square matrix $L$ is lower triangular if all the elements above the diagonal are equal to 0:

$$L = \begin{pmatrix} L_{11} & 0 & \cdots & 0 \\ L_{21} & L_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ L_{n1} & L_{n2} & \cdots & L_{nn} \end{pmatrix}.$$

### 1.3.4 Matrix inverse, linear systems and rank

A square matrix $A$ has an *inverse* if there exists a matrix, $B$, such that $AB = I = BA$. If such a matrix $B$ exists, $A$ is *invertible* and $B$ is usually written $A^{-1}$. Whenever such a $A^{-1}$ exists it

[8] Upper triangular matrices are defined analogously.

is unique:

PROPOSITION 1.3: *If $A$, $B$ and $C$ are $n \times n$ matrices such that $B$ and $C$ are both inverses of $A$, then $B = C$.*

*Proof.* We have $CA = I = AB$ and so $B = IB = (CA)B = C(AB) = CI = C$. □

If all the matrices below are square and of the same size then [exercise]

(i) If $A$ is invertible, so is $A^{-1}$ and $(A^{-1})^{-1} = A$;

(ii) If $A$ and $B$ are invertible, so is $AB$ and $(AB)^{-1} = B^{-1}A^{-1}$;

(iii) If $A_1, A_2, \ldots, A_k$ are invertible so is $A_1 A_2 \cdots A_k$ and $(A_1 A_2 \cdots A_k)^{-1} = A_k^{-1} \cdots A_2^{-1} A_1^{-1}$;

(iv) If $A$ is invertible and $\lambda \neq 0$ then $(\lambda A)^{-1} = \frac{1}{\lambda} A^{-1}$;

(v) If $A$ is invertible, so is its transpose and $(A')^{-1} = (A^{-1})'$.

A system of linear equations is an equation of the form:

$$Ax = b, \tag{4}$$

for a $m \times n$ matrix $A$, $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$.

THEOREM 1.1: *Suppose that $A$ is $n \times n$. The equation*

$$Ax = b$$

*has a unique solution $x \in \mathbb{R}^n$ for all $b \in \mathbb{R}^n$ if and only if $A$ is invertible.*

*Proof.* If $A$ is invertible, then $x = A^{-1}b$ is a solution. To show uniqueness, suppose there is some $z$ such that $Az = b$. Pre-multiplication by $A^{-1}$ gives

$$z = A^{-1}Az = A^{-1}b = x.$$

Now suppose that there is a unique solution $x$ for each $b$ and call it $x = B(b)$. $B$ is a linear transformation:

$$A(x_1 + x_2) = Ax_1 + Ax_2 = b_1 + b_2,$$

so $B(b_1 + b_2) = x_1 + x_2 = B(b_1) + B(b_2)$. Similarly, $A(\lambda x) = \lambda Ax = \lambda b$ and so $B(\lambda b) = \lambda x = \lambda B(b)$. Therefore we can represent $B$ as a matrix, which we will also call $B$. It remains to show that $B = A^{-1}$. Let $x \in \mathbb{R}^n$ and $b = Ax$. By definition of $B$, we have $B(b) = Bb = x$. So for all $x \in \mathbb{R}^n$, $BAx = Bb = x$, i.e. $BA = I$.[9] For any $b \in \mathbb{R}^n$, let $x = Bb$ so that $b = Ax$. Then, $ABb = Ax = b$, i.e. $AB = I$. □

[9] The identity is unique: if $Ax = x$ then we have $Ax - x = (A - I)x = 0$. Since this holds for all $x \in \mathbb{R}^n$, $A - I$ is the zero matrix or $A = I$.

Proposition 1.4: *Suppose that $A$ is a $n \times n$ matrix and $x_1, \ldots, x_n$ is a basis for $\mathbb{R}^n$. Then, if $A$ is invertible, $Ax_1, \ldots, Ax_n$ is a basis for $\mathbb{R}^n$. Conversely, if $Ax_1, \ldots, Ax_n$ is a basis for $\mathbb{R}^n$, then $A$ is invertible.*

*Proof.* Suppose that $x_1, \ldots, x_n$ is a basis for $\mathbb{R}^n$. Let $y \in \mathbb{R}^n$ be arbitrary. By Theorem 1.1 there is a $x$ such that $y = Ax$. By Proposition 1.2, $x = \sum_{k=1}^n \lambda_k x_k = A^{-1} \sum_{k=1}^n \lambda_k A x_k$ and so

$$y = Ax = AA^{-1} \sum_{k=1}^n \lambda_k A x_k = \sum_{k=1}^n \lambda_k A x_k.$$

Now suppose that for $\gamma_k$ $(k = 1, \ldots, n)$, $y = \sum_{k=1}^n \gamma_k A x_k$. Then,

$$x = A^{-1} y = \sum_{k=1}^n \gamma_k A^{-1} A x_k = \sum_{k=1}^n \gamma_k x_k.$$

The uniqueness statement in Proposition 1.2 implies that $\lambda_k = \gamma_k$. Hence this representation is unique and by Proposition 1.2 $Ax_1, \ldots, Ax_n$ is a basis.

Conversely, suppose that $Ax_1, \ldots, Ax_n$ is a basis. Let $y \in \mathbb{R}^n$ be arbitrary and note that by Proposition 1.2 and linearity we have

$$y = \sum_{k=1}^n \lambda_k A x_k = A \sum_{k=1}^n \lambda_k x_k = Ax,$$

where $x := \sum_{k=1}^n \lambda_k x_k$. This demonstrates that a solution $x$ exists for each $y$ in $Ax = y$. Suppose that also $z = \sum_{k=1}^n \gamma_k x_k$ is a solution. Then, we have

$$y = Az = \sum_{k=1}^n \gamma_k A x_k.$$

By the uniqueness in 1.2 it follows that $\gamma_k = \lambda_k$ and $z = x$. Hence the solution is unique and so by Theorem 1.1, $A$ is invertible. □

We will use this theorem to next show that we only need to check that an inverse satisfies one of $A^{-1}A = I$ *or* $I = AA^{-1}$, with the other then being guaranteed.

Proposition 1.5: *If square matrices $A, B$ exist such that $AB = I$, then also $BA = I$, i.e. $B$ is the inverse of $A$. Similarly, if $BA = I$ then $AB = I$.*

*Proof.* We will prove the first statement, the second follows similarly. Suppose our matrices are $n \times n$. We have that $B = BI = B(AB) = (BA)B$, or $(I - BA)B = 0$. Since the columns of $B$ are equal to $b_k := Be_k$ for $k = 1, \ldots, n$, by Proposition 1.4 the columns of $B$ form a basis for $\mathbb{R}^n$. If $\lambda = (\lambda_1, \ldots, \lambda_n)$ then $B\lambda = \sum_{k=1}^n \lambda_k b_k$ and so for any $x \in \mathbb{R}^n$ we have $x = B\lambda$ for some (unique) choice of $\lambda$ (by Proposition 1.2). Therefore, for any $x \in \mathbb{R}^n$, we have $(I - BA)x = (I - BA)B\lambda = 0$. This can only happen if $(I - BA) = 0$ [Exercise (look at what happens with $x = e_k$)], i.e. $BA = I$. □

An important property of a matrix is its *rank*. The rank of a matrix $A$, $\mathrm{rk}(A)$, is the maximal number of linearly independent columns it has. A fundamental result in linear algebra states that this is equal to the maximal number of linearly independent rows it has. A consequence of this equality is that for a $m \times n$ matrix $A$, $\mathrm{rk}(A) \leq \min\{m, n\}$. If $\mathrm{rk}(A) = \min\{m, n\}$, $A$ is said to have *full rank*.

Example 1.6 [Three full rank matrices]: The identity matrix $I_n$ is full rank, since it has $n$ linearly independent columns by Example 1.3.

Any $n \times n$ diagonal matrix with $D_{ii} \neq 0$ for $i = 1, \ldots, n$ is full rank, as it has $n$ linearly independent columns.

The matrix $A = \begin{bmatrix} 1 & -1/2 \\ 2 & -2 \end{bmatrix}$ has full rank as it has 2 linearly independent columns. △

Example 1.7 [Two rank deficient matrices]: If a $n \times n$ diagonal matrix has $D_{ii} = 0$ for $k < n$ indices in $i = 1, \ldots, n$, it has rank $n - k$.

The matrix $A = \begin{bmatrix} 1 & -1/2 \\ 2 & -1 \end{bmatrix}$ has rank 1 since it has only one linearly independent column (cf. Example 1.4). △

Example 1.8 [The zero matrix]: The zero matrix, $O$, is unique: it is the only matrix with $\mathrm{rk}(O) = 0$. This is easily seen: all of its columns are the zero vector, and any list of vectors containing the zero vector is linearly dependent. △

A square matrix is invertible if and only if it has full rank:

Corollary 1.1: *Suppose that $A$ is $n \times n$. $A$ is invertible if and only if $\mathrm{rk}(A) = n$.*

*Proof.* By Proposition 1.4 if $A$ is invertible, then $Ae_1, \ldots, Ae_n$ is a basis. Since $Ae_k$ is the $k$-th column of $A$, this implies that $A$ has $n$ linearly independent columns and hence $\mathrm{rk}(A) = n$.

Conversely, suppose that the columns of $A$ are linearly independent (i.e. $\mathrm{rk}(A) = n$). Then, $Ae_1, \ldots, Ae_n$ are a linearly independent list of $n$ vectors. Any such list must span $\mathbb{R}^n$: if it did not, there must exist some $a_{n+1} \in \mathbb{R}^n$ with $a_{n+1} \neq \sum_{k=1}^n \lambda_k a_k$ and so $a_1, \ldots, a_{n+1}$ would be linearly independent. But there cannot be $n + 1$ linearly independent vectors in $\mathbb{R}^n$ by Proposition 1.1. As such, $Ae_1, \ldots, Ae_n$ is a basis and so by Proposition 1.4, $A$ is invertible. □

Example 1.9 [Inversion of a $2 \times 2$ matrix]: Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. If this matrix has rank less than 2, then we must be able to write $A_{\bullet 1} = \gamma A_{\bullet 2}$ for some scalar $\gamma$. This gives us the system of simultaneous equations

$$a = \lambda b$$
$$c = \lambda d.$$

If $b = d = 0$ then also $a = c = 0$ and $A$ is the zero matrix, which has rank 0. Otherwise at least one of $b$ or $d$ is non-zero and we can solve for $\lambda$. Suppose that $d$ is non-zero, so that $\lambda = c/d$ (an analogous argument holds if instead $b$ is non-zero). Then, plugging this in we have $a = c/db$ or $ad - bc = 0$. Since this also equals zero if $A$ is the zero matrix, it follows that a sufficient condition that $A$ has rank 2 is that $ad - bc \neq 0$.

In this case $A$ is invertible and

$$A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix},$$

since

$$AA^{-1} = \frac{1}{ad - bc} \begin{pmatrix} ad - bc & ab - ab \\ cd - cd & ad - bc \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I. \qquad \triangle$$

Example 1.10 [Inversion of a lower trianguar matrix]: If $L$ is a $n \times n$ lower triangular matrix it is invertible if and only if all of its diagonal entries are non-zero. This follows from Corollary 1.1 since the columns of a lower triangular matrix are linearly independent if and only if all of the diagonal elements are non-zero [Exercise]. △

Example 1.11 [Inversion of a diagonal matrix]: As a consequence of example 1.10, a diagonal matrix is invertible provided all of its diagonal entries are non-zero. In this case, the inverse is given by

$$D^{-1} = \begin{pmatrix} D_{11}^{-1} & 0 & \cdots & 0 \\ 0 & D_{22}^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_{nn}^{-1} \end{pmatrix}.$$

[Exercise: verify that $DD^{-1} = I$.] As a special case of this, the identity matrix is invertible and equals its own inverse: $I^{-1} = I$. △

General linear systems of equations of the form (4) do not always have solutions and when they do have solutions those solutions may not be unique. In fact, whenever a solution to such a linear system is not unique, there are infinitely many solutions. We showed in Theorem 1.1 that neither of these situations can occur when $A$ is invertible (i.e. if and only if $\mathrm{rk}(A) = n$ by Corollary 1.1).

Rank can also give information about whether a system has no solution or infinite solutions. The *augmented matrix* $M$ (associated to the system (4)) is formed of the columns of $A$, $a_1, \ldots, a_n$, and then $b$ as an additional column:

$$M = [a_1, a_2, \ldots, a_n, b].$$

This allows us to set out the three possibilities:[10]

(i) If $\mathrm{rk}(A) = \mathrm{rk}(M) = n$ then the system (4) has a unique solution;

(ii) If $\mathrm{rk}(A) = \mathrm{rk}(M) < n$ then the system (4) has an infinite number of solutions;

(iii) If $\mathrm{rk}(A) < \mathrm{rk}(M)$ then the system (4) has no solution.

[10] We will not prove this here; see e.g. Section 3.3 of [15] for details.

## 1.4 Eigenvalues & Eigenvectors

Let $A$ be a $n \times n$ matrix. A real number $\lambda$ is an *eigenvalue* of $A$ if there exists a non-zero vector $v \in \mathbb{R}^n$ such that

$$Av = \lambda v. \tag{5}$$

Such a $v$ is called an *eigenvector* (associated to the eigenvalue $\lambda$). Eigenvectors are not unique.

Example 1.12: If $\lambda$ is an eigenvalue of $A$ with eigenvector $v$, then $\lambda$ is also an eigenvalue of $A$ corresponding to the eigenvector $av$ for any non $a \in \mathbb{R}$, $a \neq 0$. This follows since

$$Av = \lambda v \quad \Longleftrightarrow \quad A(av) = \lambda(av). \qquad \triangle$$

Example 1.13: Consider the identity matrix of size $n \times n$, $I_n$. This has $n$ eigenvalues, $\lambda_1 = \cdots = \lambda_n = 1$ with corresponding eigenvalues $e_1, \ldots, e_n$, as

$$I_n e_j = \lambda_j e_j = 1 e_j = e_j. \qquad \triangle$$

Example 1.14: Consider the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}.$$

This has eigenvalues $\lambda_1 = 0$ and $\lambda_2 = 5$ corresponding to eigenvectors $v_1 = (2, -1)'$ and $v_2 = (1, 2)'$. Observe that

$$Av_1 = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \lambda_1 v_1, \qquad Av_2 = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 5 \\ 10 \end{bmatrix} = \lambda_2 v_2. \qquad \triangle$$

The following example demonstrates that not all matrices have eigenvalues.[11]

Example 1.15: The matrix

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

has no eigenvalues. If $A$ did have an eigenvalue there would be $\lambda \in \mathbb{R}$ and $x_1, x_2 \in \mathbb{R}$, not both equal to zero, such that

$$\begin{bmatrix} \lambda x_1 \\ \lambda x_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ -x_1 \end{bmatrix}$$

Suppose that $x_1 \neq 0$. Then, plugging $x_2 = \lambda x_1$ into $\lambda x_2 = -x_1$ we obtain

$$\lambda^2 x_1 = \lambda(\lambda x_1) = -x_1,$$

but this is impossible because dividing by $x_1$ we obtain $\lambda^2 = -1$ which has no solution in the

[11] This is true in the case considered in these notes, i.e. for real matrices working over the field $\mathbb{R}$, i.e. the scalars $a$ are $a \in \mathbb{R}$. If we instead worked over $\mathbb{C}$, then every (complex, hence also real) matrix does have at least one eigenvector. See e.g. pp. 294 – 295 of [15] for an example.

real numbers. Suppose instead that $x_2 \neq 0$ and plug $x_1 = -\lambda x_2$ into $\lambda x_1 = x_2$ to obtain

$$-\lambda^2 x_2 = \lambda(-\lambda x_2) = x_2,$$

i.e. (after dividing by $-x_2$) $\lambda^2 = -1$; again this has no solution in the reals. Hence there is no such $\lambda$. △

The eigenvectors associated to distinct eigenvalues are linearly independent.

PROPOSITION 1.6: *Suppose $A$ is a $n \times n$ matrix with $\lambda_1, \ldots, \lambda_m$ eigenvalues and corresponding (non-zero) eigenvectors $(v_1, \ldots, v_m)$. Then $(v_1, \ldots, v_m)$ is linearly independent.*

*Proof.* Suppose that $(v_1, \ldots, v_m)$ is linearly dependent and let $k$ be the smallest positive integer such that $v_k \in \operatorname{span}(v_1, \ldots, v_{k-1})$. Such a $k$ must exist [Exercise]. Then there are $a_1, \ldots, a_{k-1} \in \mathbb{R}$ such that

$$v_k = a_1 v_1 + \cdots + a_{k-1} v_{k-1}. \tag{6}$$

Multipling by $A$ on both sides of (6) yields

$$\lambda_k v_k = \lambda_1 a_1 v_1 + \cdots + \lambda_{k-1} a_{k-1} v_{k-1}.$$

Multiply both sides of (6) by $\lambda_k$ and subtract the equation above to obtain

$$0 = a_1(\lambda_k - \lambda_1) v_1 + \cdots + a_{k-1}(\lambda_k - \lambda_{k-1}) v_{k-1}.$$

Since we chose $k$ as the *smallest* positive integer such that $v_k \in \operatorname{span}(v_1, \ldots, v_{k-1})$, $(v_1, \ldots, v_{k-1})$ is linearly independent. Since $\lambda_k \neq \lambda_i$ for $i < k$ this implies that $a_1 = \cdots = a_{k-1} = 0$. By (6), $v_k = 0$. But this is a contradiction to our assumption that each $v_i$ is non-zero. Hence our initial supposition that $(v_1, \ldots, v_m)$ is linearly dependent must have been false. □

COROLLARY 1.2: *An $n \times n$ matrix $A$ has at most $n$ distinct eigenvalues.*

*Proof.* Suppose that $\lambda_1, \ldots, \lambda_m$ are distinct eigenvalues and $v_1, \ldots, v_m$ their corresponding eigenvectors. Then by Proposition 1.6, $(v_1, \ldots, v_m)$ are linearly independent. The result then follows from Proposition 1.1. □

Symmetric matrices have $n$ eigenvalues and, moreover, have a decomposition based on their eigenvalues and eigenvectors. This fundamental result (which we will not prove) is known as the (real) Spectral Theorem.

THEOREM 1.2 [Real Spectral Theorem]: *Every symmetric $n \times n$ matrix $A$ has $n$ eigenvalues $\lambda_1, \ldots, \lambda_n$. Moreover, the corresponding eigenvectors $q_1, \ldots, q_n$ such that if $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ and $Q = [q_1, \ldots, q_n]$*

$$A = Q\Lambda Q', \qquad QQ' = I.$$

A $n \times n$ matrix $Q$ with the property $QQ' = I$ is called an *orthogonal matrix.*

Example 1.16: Consider the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}.$$

This matrix has eigenvalues $\lambda_1 = 3$, $\lambda_2 = -1$ and eigenvectors $q_1 = (1, 1)'$ and $q_2 = (1, -1)'$, since

$$Aq_1 = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix} = 3q_1,$$

and

$$Aq_2 = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} = -q_2.$$

Moreover if we scale the eigenvalues by $2^{-1/2}$, [check!]

$$A = 2^{-1/2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} 2^{-1/2}, \quad 2^{-1/2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} 2^{-1/2} = I_2. \qquad \triangle$$

We can use the Spectral Theorem to bound quadratic forms of a matrix $A$, i.e. expressions of the form $x'Ax$.

Lemma 1.3: *Let $A$ be a symmetric $n \times n$ matrix and arrange its eigenvalues such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Then for any $x \in \mathbb{R}^n$,*

$$\lambda_n x'x \leq x'Ax \leq \lambda_1 x'x.$$

*Proof.* Let $A = Q\Lambda Q'$ be the decomposition in Theorem 1.2. Let $u := Q'x$. Then,

$$u'u = uQQ'u = x'x$$

and

$$x'Ax = uQAQ'u = u\Lambda u = \sum_{i=1}^{n} \lambda_i u_i^2.$$

Then,

$$\sum_{i=1}^{n} \lambda_n u_i^2 \leq \sum_{i=1}^{n} \lambda_i u_i^2 \leq \sum_{i=1}^{n} \lambda_1 u_i^2$$

which implies

$$\lambda_n x'x = \lambda_n \sum_{i=1}^{n} u_i^2 \leq \sum_{i=1}^{n} \lambda_i u_i^2 = x'Ax \leq \lambda_1 \sum_{i=1}^{n} u_i^2 \leq \lambda_1 x'x. \qquad \square$$

## 1.5 Positive (semi-)definite matrices

A symmetric $n \times n$ matrix $A$ is *positive semi-definite* if $x'Ax \geq 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$. A symmetric $n \times n$ matrix $A$ is *positive definite* if $x'Ax > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$.

The following properties are easy to verify [Exercise].

(i) If $a > 0$ and $A$ is positive (semi-) definite, $aA$ is positive (semi-) definite;

(ii) If $A, B$ are positive semi-definite, then $A + B$ is positive semi-definite;

(iii) If $A$ is positive definite and $B$ positive semi-definite, then $A + B$ is positive definite.

Example 1.17: A $n \times n$ diagonal matrix $D$ with non-negative diagonal elements is positive semi-definite. If the diagonal entries are all positive, then $D$ is positive definite. Let $x \in \mathbb{R}^n \setminus \{0\}$ and note that

$$x'Dx = \sum_{i=1}^{n} x_i^2 D_{ii} \geq 0.$$

This inequality is strict if all $D_{ii} > 0$ as $x \neq 0$. △

All eigenvalues of a positive semi-definite matrix are non-negative. Positive definite matrices have only positive eigenvalues.

Proposition 1.7: *If $A$ is an $n \times n$ symmetric matrix then its eigenvalues $\lambda_1, \ldots, \lambda_n$ are all non-negative if and only if $A$ is positive semi-definite. $A$ is positive definite if and only if $\lambda_1, \ldots, \lambda_n$ are all positive.*

*Proof.* First suppose $A$ is positive semi-definite. That $A$ has $n$ eigenvalues follows from the Spectral Theorem (Theorem 1.2). Let $q_1, \ldots, q_n$ be the corresponding eigenvaues as in that Theorem. Then,

$$q_i'Aq_i = q_i'\lambda_i q_i = \lambda_i q_i' q_i = \lambda_i e_i' QQ' e_i = \lambda_i e_i' I_n e_i = \lambda_i,$$

where $e_i$ is the $i$-th canonical basis vector of Example 1.1. If $A$ is positive semi-definite, $q_i'Aq_i \geq 0$ hence the same is true of $\lambda_i$; if $A$ is positive definite $q_i'Aq_i > 0$ hence the same is true of $\lambda_i$.

Conversely suppose that each $\lambda_i$ is non-negative and let $\lambda_\star := \min\{\lambda_1, \ldots, \lambda_n\}$.[12] Then for any $x \in \mathbb{R}^n$ with $x \neq 0$ we have

$$\lambda_\star x'x \leq x'Ax,$$

by Lemma 1.3. Since $x'x = \|x\|^2 > 0$ for any $x \neq 0$, this implies that $x'Ax$ is non-negative if $\lambda_\star \geq 0$ and positive if $\lambda_\star > 0$. □

Positive (semi-) definite matrices frequently occur in (statistical) applications.

Lemma 1.4: *If $X$ is a $n \times m$ matrix then $A = XX'$ is a $n \times n$ positive semi-definite matrix.*

*Proof.* The dimension of $A$ is clear from the definition of matrix multiplication. That it is symmetric follows from properties of the transpose: $A' = (XX')' = (X')'X' = XX' = A$. To see that it is positive semi-definite note that $x'Ax = x'XX'x = (X'x)'(X'x) = \|X'x\|^2 \geq 0$. □

A useful set of facts that we will not prove are the following: if $A$ is a positive semi-definite matrix, there is exactly one positive semi-definite matrix $B$ of the same dimension such that $A = BB' = BB$. The matrix $B$ is called the *principal square root* of the matrix $A$. Often this

[12] That $A$ has $n$ eigenvalues follows from Theorem 1.2.

is shortened to just “square root”, though other square roots may exist.[13] The principal square root of $A$ has the same rank as $A$.

Lemma 1.5: *A positive semi-definite matrix $A$ is positive definite if and only if it is of full rank (equivalently, invertible).*

*Proof.* The equivalence of the full rank and invertibility conditions is Corollary 1.1. Suppose that $A$ is not of full rank. Then there is a $\lambda \in \mathbb{R}^n$, $\lambda \neq 0$, such that $A\lambda = \sum_{k=1}^{n} \lambda_k a_k = 0$. Therefore, $\lambda' A \lambda = 0$. Therefore the matrix $A$ is not positive definite.

Conversely suppose that the matrix $A$ is not positive definite. Then there is some $\lambda \in \mathbb{R}^n$ with $\lambda \neq 0$ such that $\lambda' A \lambda = 0$. Since there is a positive semidefinite matrix $B$ such that $A = BB'$, we have that $\|B'\lambda\|^2 = \lambda' BB' \lambda = 0$ and so $B'\lambda = B\lambda = 0$. Since $\lambda \neq 0$, this implies that the columns of $B$ are linearly dependent and hence $B$ is not of full rank. Since $\text{rk}(A) = \text{rk}(B)$, $A$ is not of full rank. □

Corollary 1.3: *A positive semi-definite matrix $A$ is of full rank (equivalently, invertible) if and only if all of its eigenvalues are positive.*

*Proof.* Combine Proposition 1.7 with Lemma 1.5. □

An immediate consequence of Lemma 1.5 is that if $B$ is the principal square root of $A$, then $A$ is positive definite if and only if $B$ is. A second consequence is that the inverse of a positive definite matrix is positive definite.

Another two facts that we will not prove are the following: any positive semi-definite matrix $A$ has a decomposition, its *Cholesky decomposition* into a lower trianguar matrix $L$ with non-negative diagonal entries and its transpose: $A = LL'$. If $A$ is positive definite, the diagonal elements of $L$ are all positive. The Cholesky decomposition is often useful for numerical computations.

Positive semi-definite matrices allow us to define what is called a *partial order* on the symmetric matrices.[14] We write that $A \succeq B$ if $A - B$ is positive semi-definite and $A \succ B$ if $A - B$ is positive definite.

Example 1.18: Let $A$ a positive semi-definite matrix and $O$ a $n \times n$ zero matrix. Then $A \succeq O$ as $A - O = A$ which is positive semi-definite. If $A$ is positive definite then $A \succ O$. △

Example 1.19: Let $A$ be a positive definite matrix and $\sigma > \gamma$ be scalars. Then $\sigma A \succ \gamma A$ since $\sigma A - \gamma A = (\sigma - \gamma)A$ which is positive definite. △

Negative and negative semi-definite matrices can be defined analogously. It is also easy to see that $A$ is positive (semi-) definite if and only if $-A$ is negative (semi-) definite. Definite matrices of all four forms have applications in optimisation as they provide a test for concavity / convexity of twice differentiable functions. In particular, the matrix of second partial derivatives

[13] A matrix square root of the matrix $A$ is *any* matrix $B$ such that $A = BB$; $B$ need not be positive semi-definite even if $A$ is.

[14] It is “partial” because it does not allow us to compare every pair of elements.

of a scalar function $f$ is called the Hessian and denoted $H_f$. If $x$ is a critical point (i.e. the gradient, or vector of first derivatives of $f$, is equal to zero at $x$), then $x$ is a local minimum of $f$ if $H_f(x)$ is positive semi-definite. If $H_f(x)$ is positive definite, $x$ is a *strict* local minimum. Replacing "minimum" with "maximum" and "positive (semi-) definite" with "negative (semi-) definite" the same holds.

### 1.6 The trace

The trace of a square $(n \times n)$ matrix $A$ is defined as

$$\operatorname{tr}(A) = \sum_{i=1}^{n} A_{ii},$$

i.e. it is the sum of the diagonal terms of the matrix. As is easy to check, the trace is a linear function: for any $n \times n$ matrices $A, B$ and $a \in \mathbb{R}$,

$$\operatorname{tr}(A+B) = \operatorname{tr}(A) + \operatorname{tr}(B), \qquad \operatorname{tr}(aA) = a \operatorname{tr}(A).$$

Since transposition of a matrix leaves its diagonal the same

$$\operatorname{tr}(A) = \operatorname{tr}(A').$$

Example 1.20: The trace of the identity matrix $I_n$ is $n$, as $\operatorname{tr}(I_n) = \sum_{i=1}^{n} 1 = n$. △

Example 1.21: The trace of any positive definite $n \times n$ matrix $A$ is positive. This follows since each $A_{ii} = e_i' A e_i > 0$ and hence $\operatorname{tr}(A) = \sum_{i=1}^{n} A_{ii} > 0$. △

The trace is a particularly useful operation due to the following *invariance under cyclic permutations*:[15]

$$\operatorname{tr}(ABC) = \operatorname{tr}(BCA) = \operatorname{tr}(CAB). \tag{7}$$

Proposition 1.8: *If matrices $A, B$ have dimensions such that $AB$ exists and is square, then* $\operatorname{tr}(AB) = \operatorname{tr}(BA)$.

*If matrices $A, B, C$ have dimensions such that $ABC$ exists and is square, then equation* (7) *holds.*

*Proof.* The restriction that $AB$ exists and is square implies that if $A$ is $m \times n$, then $B$ must be $n \times m$. Hence, $BA$ exists and is square. The $i$-th diagonal entry of $AB$ is $\sum_{j=1}^{n} A_{ij}B_{ji}$ and the $j$-th diagonal entry of $BA$ is $\sum_{i=1}^{n} B_{ji}A_{ij}$. Hence

$$\operatorname{tr}(AB) = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}B_{ji} = \sum_{j=1}^{n}\sum_{i=1}^{n} B_{ij}A_{ji} = \operatorname{tr}(BA).$$

The 3-product cylic invariance follows from this. Let $D = BC$, so $ABC = AD$. By the same

[15] This can be extended to $k$ products for any $k \in \mathbb{N}$.

argument as above, then $DA = BCA$ exists (and is square) and $\mathrm{tr}(AD) = \mathrm{tr}(DA)$. Next let $E = CA$. Since $BE = BCA$ is square, the same argument as before implies that $EB = CAB$ exists and is square and $\mathrm{tr}(BE) = \mathrm{tr}(EB)$. We have

$$\mathrm{tr}(ABC) = \mathrm{tr}(AD) = \mathrm{tr}(DA) = \mathrm{tr}(BCA) = \mathrm{tr}(BE) = \mathrm{tr}(EB) = \mathrm{tr}(CAB). \qquad \square$$

## 1.7 Subspaces, orthogonality and projections

A subset $V \subset \mathbb{R}^n$ is a *(linear) subspace* of $\mathbb{R}^n$ if

(i) if $x, y \in V$, then $x + y \in V$

(ii) if $x \in V$ and $\lambda \in \mathbb{R}$, then $\lambda x \in V$.

Example 1.22: Let $V = \mathbb{R}^n$. Then $V$ is a linear subspace of $\mathbb{R}^n$. △

Example 1.23: Let $V = \{(x, 0) : x \in \mathbb{R}\} \subset \mathbb{R}^2$. Then $V$ is a linear subspace of $\mathbb{R}^2$. Taking $x = 0 \in \mathbb{R}$ shows that $0 \in V$. If $v = (x, 0), w = (y, 0)$ are both in $V$ then $u + w = (x + y, 0) \in V$ since $x + y \in \mathbb{R}$. If $\lambda \in \mathbb{R}$ and $v = (x, 0) \in V$, then $\lambda v = (\lambda x, 0) \in V$ as $\lambda x \in \mathbb{R}$. △

There are a number of important subspaces associated with matrices (or the associated linear transformation). Let $A$ be a $m \times n$ matrix. Each of the following is a linear subspace [Exercise]:

(i) The *nullspace* of $A$, $N(A)$ is defined by $N(A) = \{x \in \mathbb{R}^n : Ax = 0\} \subset \mathbb{R}^n$;

(ii) The *range* or *column space* of $A$, $R(A)$ defined by $R(A) := \{Ax \in \mathbb{R}^m : x \in \mathbb{R}^n\}$;

(iii) The *row space* of $A$, which is the column space of $A'$.

The concept of linear independence and span apply to subspaces just as to $\mathbb{R}^n$; the only difference is that the vectors in question must belong to $V$. A basis for $V$ is a linearly independent list of vectors in $V$ which span $V$. Proposition 1.2 is valid for $V$, provided we replace $\mathbb{R}^n$ by $V$ everywhere in the statement and proof:

Proposition 1.9: *A list $x_1, \ldots, x_n$ of vectors in a linear subspace $V$ is a basis if and only if every $x \in V$ can be written uniquely as*

$$x = \lambda_1 x_1 + \cdots + \lambda_n x_n, \quad \lambda_1, \ldots, \lambda_n \in \mathbb{R}.$$

*Proof.* The proof is the same as that of Proposition 1.2 with each occurance of $\mathbb{R}^n$ replaced by $V$. □

Lemma 1.6: *Any two bases for $V$ have the same length.*

*Proof.* Let $B_1 = v_1, \ldots, v_n$ and $B_2 = w_1, \ldots, w_m$ be bases for $V$. Then $B_1$ spans $V$ and $B_2$ is linearly independent. Hence by Lemma 1.2, $m \leq n$. But also $B_2$ spans $V$ and $B_1$ is linearly independent. Hence by Lemma 1.2, $n \leq m$. □

The length of a basis for $V$ is the *dimension* of $V$, $\dim V$.

Lemma 1.7: *The dimension of $R(A)$ is* $\mathrm{rk}(A)$.

*Proof.* Let $r = \mathrm{rk}(A)$ and $a_{i_1}, \dots, a_{i_r}$ be the $r$ linearly independent columns of $A$. It suffices to show that this is a basis for $R(A)$. If $r = n$, then this follows from Proposition 1.4 and Corollary 1.1 and the observation that $R(A) = \mathbb{R}^m$ as any $y \in \mathbb{R}^m$ can be expressed as $y = Ax$ for $x = A^{-1}y$.

Otherwise let $a_{i_{r+1}}, \dots, a_{i_n}$ be the remaning columns in $A$. Let $y \in R(A)$, i.e. $y = Ax$ for some $x \in \mathbb{R}^n$. Equivalently, $y = \sum_{k=1}^n a_k x_k$. Since each of the remaining columns are linear combinations of $a_{i_1}, \dots, a_{i_r}$ we have that $a_{i_l} = \sum_{j=1}^r \lambda_j^l a_{i_j}$ with the $\lambda_i^l$ not all zero for each $l = r+1, \dots, n$. Then, we can write

$$y = Ax = \sum_{i=1}^n x_k a_k = \sum_{j=1}^r x_{i_j} a_{i_j} + \sum_{l=r+1}^n x_{i_l} \sum_{j=1}^r \lambda_j^l a_{i_j} = \sum_{j=1}^r \left( x_{i_j} + \sum_{l=r+1}^n x_{i_l} \lambda_j^l \right) a_{i_j}.$$

Suppose that also $y = \sum_{j=1}^r b_j a_{i_j}$ for some scalars $b_j$. Subtraction of this from the preceding display yields

$$0 = \sum_{j=1}^r \left( x_{i_j} + \sum_{l=r+1}^n x_{i_l} \lambda_j^l - b_j \right) a_{i_j}.$$

Since the $a_{i_1}, \dots, a_{i_r}$ are linearly independent, this implies that each $x_{i_j} + \sum_{l=r+1}^n x_{i_l} \lambda_j^l - b_j = 0$. It then follows from Proposition 1.9 that $a_{i_1}, \dots, a_{i_r}$ is a basis for $V$. □

Theorem 1.3 [Rank - nullity theorem]: *If $A$ is a $m \times n$ matrix, then*

$$n = \dim R(A) + \dim N(A) = \mathrm{rk}(A) + \dim N(A).$$

*Proof.* In view of Lemma 1.7 it suffices to prove the first equality. For this, suppose that $k \leq n$ is the dimension of $N(A)$ and so there is a basis $v_1, \dots, v_k$ of $N(A)$.[16] We can extend this to a basis of $\mathbb{R}^n$. If $v_1, \dots, v_k$ span $\mathbb{R}^n$ then stop. Else there is a $w_1 \in \mathbb{R}^n$ such that $v_1, \dots, v_k, w_1$ are linearly independent. If this list spans $\mathbb{R}^n$ stop. Else there is a $w_2 \in \mathbb{R}^n$ such that $v_1, \dots, v_k, w_1, w_2$ are linearly independent and so forth. This must stop when we have a basis $v_1, \dots, v_k, w_1, \dots, w_l$ with $l + k = n$ as otherwise Proposition 1.1 would be violated. To complete the proof it suffices to show that $\dim R(A) = l$: for this we will show that $Aw_1, \dots, Aw_l$ forms a basis of $R(A)$. Let $x \in V$. Then

$$x = a_1 v_1 + \cdots a_k v_k + b_1 w_1 + \cdots + b_l w_l.$$

[16] Every subspace $V$ of $\mathbb{R}^n$ has a basis (of length $\leq n$): start with an arbitrary vector $x_1 \in V$. If $\mathrm{span}(x_1) = V$ stop. Else there is a $x_2 \in V$ such that $x_1, x_2$ are linearly independent. If $\mathrm{span}(x_1, x_2) = V$ stop. Else there is an $x_3$.... and so forth. This procedure must stop after $m \leq n$ iterations by Proposition 1.1. If that $m = n$, the resulting basis is a list of $n$ linearly independent vectors in $\mathbb{R}^n$ and so spans $\mathbb{R}^n$ (if it didn't we could add another vector and our list would still be linearly independent, but that would violate Proposition 1.1). In this case $V = \mathbb{R}^n$.

Premultiplying by $A$ we have

$$Ax = b_1 Aw_1 + \cdots + b_l Aw_l,$$

since each $v_i \in N(A)$ and so $Av_i = 0$. The above display shows that $Aw_1, \ldots, Aw_l$ spans $R(A)$. For linear independence suppose that $c_1 Aw_1 + \cdots c_l Aw_l = 0$. Using linearity, we have $A(c_1 w_1 + \cdots c_l w_l) = 0$, and so $c_1 w_1 + \cdots c_l w_l \in N(A)$. Since $v_1, \ldots, v_k$ is a basis for $N(A)$ there are scalars $d_1, \ldots, d_k$ such that

$$c_1 w_1 + \cdots c_l w_l + (-d_1)v_1 + \cdots + (-d_k)v_k = 0.$$

But since $v_1, \ldots, v_k, w_1, \ldots, w_l$ is a basis, it is linearly independent and hence each $c_i$ (and each $d_i$) must be 0. Thus $Aw_1, \ldots, Aw_l$ is a basis for $R(A)$ and the proof is complete. □

We may also now strengthen the result we obtained in Corollary 1.2.

Corollary 1.4: *If $A$ is a $n \times n$ matrix it has at most* $\min\{\text{rank}(A)+1, n\}$ *distinct eigenvalues.*

*Proof.* Suppose that $(\lambda_1, \ldots, \lambda_m)$ are distinct eigenvalues of $A$ and $(v_1, \ldots, v_m)$ corresponding (non-zero) eigenvectors. Each $\lambda_i v_i = Av_i \in R(A)$. If each $\lambda_i \neq 0$, then by the definition of a subspace also $v_i \in R(A)$. By Lemma 1.7 and Propositions 1.1, 1.6, $m \leq \text{rank}(A) \leq \text{rank}(A)+1$.

If one $\lambda_i = 0$, then (there can only be one such $\lambda_i$ in our list since they are distinct) repeat the above argument applied to the list $(\lambda_1, \lambda_{i-1}, \lambda_{i+1}, \lambda_m)$ to conclude that $m - 1 \leq \text{rank}(A)$, i.e. $m \leq \text{rank}(A) + 1$. The proof is completed by appealing to Corollary 1.2. □

Two vectors are *orthogonal* if their dot product $x \cdot y = x'y = 0$. A list of vectors $v_1, \ldots, v_k$ is *orthonormal* if $v_i' v_j = 0$ for all $i \neq j$ and each $v_i' v_i = 1$. We will next show that any subspace $V$ has an orthonormal basis $e_1, \ldots, e_k$ (where $k = \dim V$). Such a basis is particularly useful, since it is easy to calculate the coefficients $\lambda_j$ in the basis expansion

$$v = \lambda_1 e_1 + \cdots + \lambda_k e_k.$$

In particular, taking the dot product with $e_j$ on each side of this yields

$$e_j' v = \lambda_j.$$

We start by showing that orthonormal lists are linearly independent.

Proposition 1.10: *Any orthonormal list of vectors is linearly independent.*

*Proof.* Let $v_1, \ldots, v_m$ be an orthonormal list of vectors. Suppose $\lambda_1, \ldots, \lambda_m$ are scalars such that

$$0 = \lambda_1 v_1 + \cdots + \lambda_n v_n.$$

The squared norm of the sum of orthogonal vectors is equal to the sum of their norms since if $x, y$ are orthogonal, $\|x + y\|^2 = (x+y)'(x+y) = x'x + x'y + y'x + y'y = \|x\|^2 + \|y\|^2$. Using this fact repeatedly along with the fact that $\|v_i\| = 1$ for any element of an orthonormal list, we have

$$0 = \|\lambda_1 v_1 + \cdots + \lambda_n v_n\|^2 = |\lambda_1|^2 + \cdots + |\lambda_n|^2.$$

This implies each $\lambda_i = 0$. □

An orthonormal list of vectors can be produced from an initial linearly independent list by the following Gram – Schmidt procedure.

Proposition 1.11 [Gram – Schmidt procedure]: *Suppose that $v_1, \ldots, v_k$ is a linearly independent list of vectors in $V$, a linear subspace of $\mathbb{R}^n$. Let $e_1 := v_1/\|v_1\|$ and for $j = 2, \ldots, k$,*

$$e_j := \frac{v_j - (v_j' e_1)e_1 - \cdots - (v_j' e_{j-1})e_{j-1}}{\|v_j - (v_j' e_1)e_1 - \cdots - (v_j' e_{j-1})e_{j-1}\|}.$$

*Then $e_1, \ldots, e_k$ is an orthonormal list of vectors in $V$ with*

$$\operatorname{span}(v_1, \ldots, v_j) = \operatorname{span}(e_1, \ldots, e_j),$$

*for $j = 1, \ldots, k$*

*Proof.* We proceed by induction on $j$. If $j = 1$, then $\operatorname{span}(e_1) = \operatorname{span}(v_1)$ since $e_1 = v_1/\|v_1\|$. Now suppose we have constructed $e_1, \ldots, e_{j-1}$ and verified $\operatorname{span}(v_1, \ldots, v_{j-1}) = \operatorname{span}(e_1, \ldots, e_{j-1})$. Since $v_1, \ldots, v_k$ is linearly independent, this ensures that $v_j \notin \operatorname{span}(e_1, \ldots, e_{j-1})$ and so $\|v_j - (v_j' e_1)e_1 - \cdots - (v_j' e_{j-1})e_{j-1}\| \neq 0$ and $e_j$ is well defined. Clearly $e_j$ has norm 1. For any $1 \leq l < j$,

$$e_l' e_j = \frac{e_l' v_j - (v_j' e_1)e_1 - \cdots - (v_j' e_{j-1})e_l' e_{j-1}}{\|v_j - (v_j' e_1)e_1 - \cdots - (v_j' e_{j-1})e_{j-1}\|} = \frac{e_l' v_j - (v_j' e_l)e_l' e_l}{\|v_j - (v_j' e_1)e_1 - \cdots - (v_j' e_{j-1})e_{j-1}\|} = 0,$$

since $e_l' e_l = \|e_l\|^2 = 1$. Therefore, the $e_1, \ldots, e_l$ so constructed form an orthonormal list. It remains to show that $\operatorname{span}(v_1, \ldots, v_j) = \operatorname{span}(e_1, \ldots, e_j)$. To show this first suppose that $x \in \operatorname{span}(v_1, \ldots, v_j)$. Then, since we have $\operatorname{span}(v_1, \ldots, v_{j-1}) = \operatorname{span}(e_1, \ldots, e_{j-1})$ and $v_j$ can clearly be expressed as a linear combination of $e_1, \ldots, e_j$,

$$x = \sum_{i=1}^{j} \lambda_i v_i = \sum_{i=1}^{j-1} \lambda_i v_i + \lambda_j v_j = \sum_{i=1}^{j-1} \gamma_i e_i + \lambda_j \left( \sum_{i=1}^{j} r_i e_i \right),$$

showing that $x \in \operatorname{span}(e_1, \ldots, e_{j-1}, e_j)$. Conversely suppose that $x \in \operatorname{span}(e_1, \ldots, e_{j-1}, e_j)$. Then by similar arguments, noting that $e_j = \sum_{i=1}^{j-1} \alpha_i e_i + \alpha_i v_j$ for some $\alpha$ coefficients, we have

$$x = \sum_{i=1}^{j-1} \lambda_i e_i + \lambda_j e_j = \sum_{i=1}^{j-1} \gamma_i v_i + \lambda_j \left( \sum_{i=1}^{j-1} \alpha_i e_i + \alpha_j v_j \right) = \sum_{i=1}^{j-1} \gamma_i v_i + \lambda_j \alpha_j v_j + \sum_{i=1}^{j-1} \beta_i v_i,$$

and so $x \in \operatorname{span}(v_1, \ldots, v_j)$. □

Corollary 1.5: *If $V$ is a linear subspace of $\mathbb{R}^n$ it has an orthonormal basis.*

*Proof.* $V$ has a basis by footnote 16. This can be turned into an orthonormal basis by applying the Gram – Schmidt procedure of Proposition 1.11. □

If $V$ is a subset of $\mathbb{R}^n$, then its *orthogonal complement* is $V^\perp = \{u \in \mathbb{R}^n : u'v = 0 \text{ for all } v \in V\}$. This is a subspace of $\mathbb{R}^n$ [Exercise].

If $U$, $V$ are subspaces of $\mathbb{R}^n$ their *sum* is $U + V = \{u + v : u \in U, v \in V\}$. This is a subspace [Exercise]. If each $x \in U + V$ can be written in only one way as $x = u + v$ with $u \in U$, $v \in V$, then $U + V$ is a *direct sum.* This is often notationally indicated by $U \oplus V$.

Lemma 1.8: *If $U$, $V$ are subspaces of $\mathbb{R}^n$, $U + W$ is a direct sum if and only if $U \cap V = \{0\}$.*

*Proof.* Supose $U + V$ is a direct sum. If $x \in U \cap V$ then $0 = x + (-x)$ with $x \in U$ and $-x \in V$. But since the representation is unique, we must have that $x = 0$ and $U \cap V = \{0\}$.[17]

Conversely, suppose $U \cap V = \{0\}$. Now suppose that $u \in U$ and $v \in V$ and $0 = u + v$. We will first show that this implies that $u = v = 0$: we have $u = -v \in V$. So $u \in U \cap V = \{0\}$ and hence $u = -v = 0$ and so $v = 0$. It remains to show that $U + V$ is a direct sum if the only way to write 0 as a sum of $u, v$ in $U, V$ respectively is to take each summand equal to zero. Suppose this is true and let $x \in U + V$. Then, $x = u + v$ for $u \in U$ and $v \in V$. Suppose also that $x = z + y$ with $z \in U$ and $y \in V$. Subtract these to find $0 = (u - z) + (v - y)$. Since $u - z \in U$ and $(v - y) \in V$, by our supposition we must have $u - z = 0 = v - y$. But this means that the representation of $x \in U + V$ is unique, that is, $U + V$ is a direct sum. □

Proposition 1.12: *If $V$ is a linear subspace of $\mathbb{R}^n$, then*

$$\mathbb{R}^n = V \oplus V^\perp.$$

*Proof.* We first show that $\mathbb{R}^n = V + V^\perp$. Suppose that $x \in \mathbb{R}^n$ and let $e_1, \ldots, e_m$ be an orthonormal basis of $V$ (cf. Corollary 1.5). Then

$$x = \left[(x'e_1)e_1 + \cdots + (x'e_m)e_m\right] + \left[x - (x'e_1)e_1 + \cdots + (x'e_m)e_m\right].$$

Let $y$ be the first bracketed term on the right hand side and $z$ the second bracketed term. Clearly $y \in V$. Moreover, since for any $1 \leq i \leq m$,

$$z'e_i = x'e_i - (x'e_1)e_1'e_i + \cdots + (x'e_m)e_m'e_i = x'e_i - x'e_i = 0,$$

$z$ is orthogonal to each $e_i$. It follows immediately that $z$ is orthogonal to all $u \in V$, since these are linear combinations of the basis $e_1, \ldots, e_m$. Thus $z \in V^\perp$. Therefore, $x = y + z$, with $y \in V$ and $z \in V^\perp$, and so $\mathbb{R}^n = V + V^\perp$.

We note that by Lemma 1.8 it suffices to show that $V \cap V^\perp = \{0\}$ to complete the proof. For this note that if $v \in V \cap V^\perp$ then $v$ is orthogonal to itself, i.e. $v'v = \|v\|^2 = 0$, which is possible only if $v = 0$. □

[17] If any other $x$ belonged to the intersection, we would have an alternative representation of $0 \in U + V$ as $x + (-x)$.

Having proven that $\mathbb{R}^n$ can be decomposed as a direct sum of any linear subspace and its orthogonal complement, we can now define the *orthogonal projection* onto any linear subspace. The orthogonal projection $P_V$ onto the linear subspace $V$ is a linear transformation defined as follows: let $x \in \mathbb{R}^n$ and write $x = v + w$ where $v \in V$ and $w \in V^\perp$. Then $P_V x = v$. The orthogonal projection has the following properties:

LEMMA 1.9: *Let $V$ be a linear subspace of $\mathbb{R}^n$ and $P_V$ the orthogonal projection onto $V$. Then*

*(i) $P_V$ is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^n$;*

*(ii) $P_V v = v$ for $v \in V$;*

*(iii) $P_V u = 0$ for $u \in V^\perp$;*

*(iv) $\{P_V v : v \in \mathbb{R}^n\} = V$;*

*(v) $\{v \in \mathbb{R}^n : P_V v = 0\} = V^\perp$;*

*(vi) $x - P_V x \in V^\perp$;*

*(vii) $P_V P_V x = P_V x$ for all $x \in \mathbb{R}^n$;*

*(viii) $\|P_V v\| \leq \|v\|$;*

*(ix) If $e_1, \ldots, e_m$ is an orthonormal basis of $V$, $P_V v = (v'e_1)e_1 + \cdots + (v'e_m)e_m$.*

*Proof.* (i) - (iii) Exercise.

(iv) $\{P_V v : v \in \mathbb{R}^n\}$ is the range of $P_V$, $R(P_V)$. The definition of $P_V$ implies that $R(P_V) \subset V$. The converse direction follows from the fact that if $v \in V$, then $v = P_V v \in R(P_V)$ by (ii).

(v) $\{v \in \mathbb{R}^n : P_V v = 0\}$ is the nullspace of $P_V$, $N(P_V)$. $U^\perp \subset N(P_V)$ follows from (iii). Conversely, if $v$ is such that $P_V v = 0$ then the direct sum decomposition in the definition of $P_V$ must be $v = 0 + v$, with $0 \in V$ and $v \in V^\perp$. Hence $N(P_V) \subset U^\perp$.

(vi) Let $x = v + w$ be the direct sum definition of $x$ in the definition of $P_V$. Then, $x - P_V x = v + w - v = w \in V^\perp$.

(vii) Exercise.

(viii) Let $x = v + w$ be the direct sum definition of $x$ in the definition of $P_V$. Then $P_V(P_V x) = P_V v = v = P_V x$.

(ix) Exercise. □

THEOREM 1.4 [Projection theorem]: *If $V$ is a subspace of $\mathbb{R}^n$, $x \in \mathbb{R}^n$, $v \in V$ then*

$$\|x - P_V x\| \leq \|x - v\|,$$

*with equality if and only if $v = P_V x$.*

*Proof.* By part (vi) of Lemma 1.9, $x - P_V x \in V^\perp$. Moreover, $P_V x - v \in V$ and so we have

$$\|(x-P_V x)+(P_V x-v)\|^2 = (x-P_V x)'(x-P_V x)+(P_V x-v)'(P_V x-v) = \|x-P_V x\|^2+\|P_V x-v\|^2.$$

Therefore,

$$\|x - P_V x\|^2 \leq \|x - P_V x\|^2 + \|P_V x - v\|^2 = \|(x - P_V x) + (P_V x - v)\|^2 = \|x - v\|^2,$$

and taking the square root gives the desired inequality. It will be an equality if and only if $\|P_V x - v\|^2 = 0$, i.e. if only if $P_V x = v$. □

Since the orthogonal projection is a linear transformation, it can be represented by a matrix; see the following example.

Example 1.24 [Projection onto the column space of $X$]: Let $X$ be a $n \times k$ matrix. The column space of $X$, $V = R(X)$, defined by $V = R(X) = \{Xv : v \in\in \mathbb{R}^k\}$ is a linear subspace of $\mathbb{R}^n$. Therefore $P_{R(X)}$ exists. For any $y \in \mathbb{R}^n$, $P_{R(X)}y$ is given by

$$P_{R(X)}y = X(X'X)^{-1}X'y.$$

For this we will form two matrices, $Q = X(X'X)^{-1}X'$ and $M = I - Q$. Clearly any $y \in \mathbb{R}^n$ can be decomposed into

$$y = Qy + My.$$

Moreover, we have that any $Qy \in R(X)$ and for any $Xv \in R(X)$, $My$ satisfies

$$(My)'(Xv) = y'(I - X(X'X)^{-1}X')Xv = y'(Xv - Xv) = 0.$$

Hence we have a decomposition into $R(X)$ and $R(X)^\perp$, and thus $Qy = P_{R(X)}y$. △

# 2 Probability

## 2.1 Probability and random variables

### 2.1.1 Events and probabilities

We start with a set $S$ of possible outcomes; this is called the *sample space*. To define probabilities we need to introduce a class of sets $\mathcal{S}$, called a $\sigma$-algebra. $\mathcal{S}$ is a set of subsets of $S$ which must satisfy certain properties.[18] *Events* are elements of $\mathcal{S}$ and so subsets of $S$. Depending on what $S$ is, $\mathcal{S}$ may contain all subsets of $S$ or it may not; we will not discuss these technicalities in this course.

A (real-valued) function $P$ defined on $\mathcal{S}$ is a *probability* if it satisfies the following conditions

(i) $P(A) \geq 0$ for all $A \in \mathcal{S}$;

(ii) $P(S) = 1$;

(iii) If $A_1, A_2, \ldots \in \mathcal{S}$ are *pairwise disjoint*, i.e. for any $i \neq j$, $A_i \cap A_j = \emptyset$, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

From these conditions we can derive the following properties of a probability.

Proposition 2.1: *If $P$ is a probability defined on $\mathcal{S}$ and $A \in \mathcal{S}$,*

*(i)* $P(\emptyset) = 0$*;*

*(ii)* $P(A) \leq 1$*;*

*(iii)* $P(A^{\complement}) = 1 - P(A)$*;*

*(iv) If the sequence of events $(A_n)_{n\in\mathbb{N}}$ is non-decreasing and $A_n \uparrow A$, then $P(A_n) \uparrow P(A)$;*[19]

*(v) If the sequence of events $(A_n)_{n\in\mathbb{N}}$ is non-increasing, $A_n \downarrow A$, then $P(A_n) \downarrow P(A)$.*[20]

*Proof.* Since $\emptyset$ is the complement of $S$ (in $S$), (i) will follow by combining (iii) and the second defining property of a probability. Similarly, since $P(A^{\complement}) \geq 0$ by the first defining property of a probability, (ii) will also follow from (iii). For (iii), the sets $A$ and $A^{\complement}$ are disjoint and $A \cup A^{\complement} = S$. Hence by the second and third defining properties of a probability, $1 = P(S) = P(A \cup A^{\complement}) = P(A) + P(A^{\complement})$, which proves (iii). For (iv), define $B_1 = A_1$ and then $B_k = E_k \setminus E_{k-1}$ for $k \geq 2$. The sets $(E_n)_{n\in\mathbb{N}}$ are pairwise disjoint and $A_n = \cup_{i=1}^{n} E_i$. Therefore, $\cup_{i=1}^{\infty} E_i = A$ and by the third defining property of a probability

$$P(A) = P\left(\cup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) = \lim_{n\to\infty} \sum_{i=1}^{n} P(E_i) = \lim_{n\to\infty} P\left(\cup_{i=1}^{n} E_i\right) = \lim_{n\to\infty} P(A_n).$$

[18] (i) The empty set $\emptyset \in \mathcal{S}$; (ii) if $A \in \mathcal{S}$, then its complement $A^{\complement} = S \setminus A \in \mathcal{S}$; (iii) if $A_1, A_2, \ldots \in \mathcal{S}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{S}$.

[19] A nondecreasing sequence of sets is one such that $A_n \subset A_{n+1}$ for each $n \in \mathbb{N}$. In this case a limit always exists: $\lim_{n\to\infty} A_n = \cup_{n\in\mathbb{N}} A_n$ and so the statement in the theorem that $A_n \uparrow A$ merely requires that $A = \cup_{n\in\mathbb{N}} A_n$.

[20] A nonincreasing sequence of sets is one such that $A_{n+1} \subset A_n$ for each $n \in \mathbb{N}$. In this case a limit always exists: $\lim_{n\to\infty} A_n = \cap_{n\in\mathbb{N}} A_n$ and so the statement in the theorem that $A_n \downarrow A$ merely requires that $A = \cap_{n\in\mathbb{N}} A_n$.

The monotonicity of the limit follows since probabilities are nonnegative (defining property 1). (v) now follows by combining (iii) and (iv). □

Note that the definition and the preceding result tell us that the range of $P$ is contained in $[0, 1]$. That is, any probability is a function $P : \mathcal{S} \to [0, 1]$.

Proposition 2.2: *If $P : \mathcal{S} \to [0, 1]$ is a probability and $A, B \in \mathcal{S}$ then*

*(i)* $P(B \cap A^{\complement}) = P(B) - P(A \cap B)$;

*(ii)* $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;

*(iii) If $A \subset B$, $P(A) \le P(B)$.*

*(iv) If $(A_n)_{n\in\mathbb{N}}$ is a sequence of events, then for any event $A \subset \cup_{n\in\mathbb{N}} A_n$, $P(A) \le \sum_{n=1}^{\infty} P(A_n)$.*

*Proof.* We can write any set $B$ as $(B \cap A) \cup (B \cap A^{\complement})$. The two sets in parentheses are disjoint and so

$$P(B) = P(B \cap A) + P(B \cap A^{\complement}),$$

which is equivalent to (i). For (ii), we first show that $A \cup B = A \cup (B \cap A^{\complement})$. Suppose that $x \in A \cup B$. If $x \in A$, then also $x \in A \cup (B \cap A^{\complement})$. If $x \notin A$, it is in $A^{\complement}$ and also in $B$, hence again $x \in A \cup (B \cap A^{\complement})$. Conversely suppose that $x \in A \cup (B \cap A^{\complement})$. If $x \in A$, then $x \in A \cup B$. If $x \in (B \cap A^{\complement})$, then $x \in B$ so $x \in A \cup B$. Now, since $A \cup (B \cap A^{\complement})$ are disjoint,

$$P(A \cup B) = P(A) + P(B \cap A^{\complement}) = P(A) + P(B) - P(A \cap B),$$

where the second equality uses (i). For (iii) observe that if $A \subset B$, then $A \cap B = A$ and so

$$0 \le P(B \cap A^{\complement}) = P(B) - P(A),$$

by (i). For (iv), let $A'_n := A_n \cap A$ and $B_n := A'_n \setminus \cup_{m=1}^{n-1} A'_n$. The $B_n$ are disjoint and $\cup_{n\in\mathbb{N}} B_n = A$. Hence, by (iii) $P(A) = \sum_{n=1}^{\infty} P(B_n) \le \sum_{n=1}^{\infty} P(A_n)$. □

### 2.1.2 Random variables

Often the events we work with are defined in terms of *random variables.* A *random variable* is a function $X : S \to \mathbb{R}$.[21] Implicitly this defines a new sample space, $\mathcal{X} \subset \mathbb{R}$, the range of $X$, and a new probability function defined on a new $\sigma$-algebra, $\mathcal{B}$.[22] Probabilities are assigned to events in $\mathcal{B}$ as follows:

$$P_X(B) = P(\{s \in S : X(s) \in B\}) = P(X^{-1}(B)),$$

[21] However it may be that not *every* function from $S \to \mathbb{R}$ is a random variable. Like with subsets of $S$ which may not be events, we will not discuss such technicalities.

[22] We will always assume that $\mathcal{X}$ is either open or closed in $\mathbb{R}$ and take $\mathcal{B}$ to be the *Borel* $\sigma$-algebra on $\mathcal{X}$: this is the smallest $\sigma$-algebra which contains all the open subsets of $\mathcal{X}$.

where $X^{-1}(B)$ is the *pre-image of* $B \in \mathcal{B}$ *under* $X$: $X^{-1}(B) = \{s \in S : X(s) \in B\}$. In probability, the probability $P(\{s \in S : X(s) \in B\})$ is often abbreviated by $P(X \in B)$ or $P(\{X \in B\})$; we will often do the same in these notes.

Example 2.1 [Tossing two dice]: Let

$$S = \{(1,1),(1,2),\dots,(1,6),(2,1),(2,2),\dots,(2,6),\dots,(6,1),(6,2),\dots,(6,6)\}.$$

$\mathcal{S}$ is all subset of $S$, including $S$. Probabilities can all be calculated once we define $P(\{(i,j)\}) = 1/36$ for $i = 1,\dots,6$ and $j = 1,\dots,6$. For example, if $A = \{(1,1),(2,2),\dots,(6,6)\}$, i.e. the probability that doubles appear, we have

$$P(A) = \sum_{i=1}^{6} P(\{(i,i)\}) = \frac{6}{36} = \frac{1}{6},$$

since the sets $\{(i,i)\}$ for $i = 1,\dots,6$ are all disjoint.

Define the function $X : S \to \mathbb{R}$ as $f((i,j)) = i + j$: the sum of the two dice. This is a random variable with sample space $\mathcal{X} = \{2,\dots,12\}$ and $\mathcal{B}$ is all subsets of $\mathcal{X}$. The probability of, for example $\{3\} \in \mathcal{B}$ is

$$P_X(\{3\}) = P(X = 3) = P(X^{-1}(\{3\})).$$

The set $X^{-1}(\{3\}) = \{(1,2),(2,1)\}$ since these are the only two ways that we can get 3 as the sum of two dice rolls. Hence

$$P_X(\{3\}) = P(\{(1,2),(2,1)\}) = \frac{2}{36} = \frac{1}{18}. \qquad \triangle$$

### 2.1.3 Distribution functions

The *(cumulative) distribution function* or *cdf*, $F : \mathbb{R} \to [0,1]$, of a random variable $X$ is defined as

$$F(x) = P(X \le x), \qquad \text{for } x \in \mathbb{R}.$$

Any cdf satisfies the following properties.

Proposition 2.3: *If $F$ is the cdf of a random variable $X$, then*

*(i)* $\lim_{x\to-\infty} F(x) = 0$ *and* $\lim_{x\to\infty} F(x) = 1$*;*

*(ii) $F$ is a nondecreasing function of $x$;*

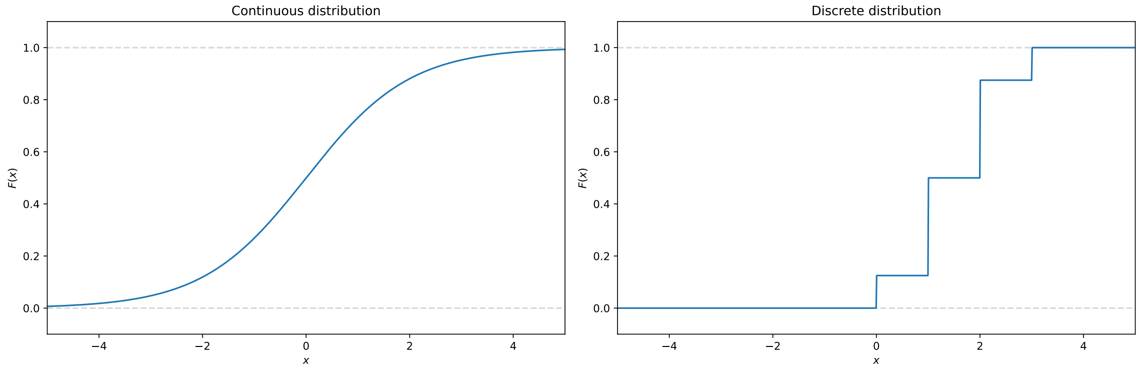
*(iii) $F$ is right-continuous: for every $x \in \mathbb{R}$, $\lim_{z\downarrow x} F(z) = F(x)$.*

*Proof.* For (i), we note that $F(x) = P(X \le x) = P(X \in (-\infty, x])$ and so $\lim_{x\to-\infty} F(x) = \lim_{x_n\to-\infty} P(X \in (-\infty, x_n])$. Since $(-\infty, x_n] \downarrow \emptyset$ as $x_n \to -\infty$, by Proposition 2.1 we have $\lim_{x\to-\infty} F(x) = \lim_{x_n\to-\infty} P(X \in (-\infty, x_n]) \to 0$. The second condition can be shown essentially analogously.[23] (ii) follows from Proposition 2.2 and the fact that $(-\infty, x] \subset (-\infty, y]$ if $x \leq y$. For (iii), note that if $z_n \downarrow x$ then $(-\infty, z_n] \downarrow (-\infty, x]$ and therefore $F(z_n) = P(X \in (-\infty, z_n]) \downarrow P(X \in (-\infty, x]) = F(x)$.[24] □

The converse of the preceding proposition is actually also true: any function with these three properties is the cdf of a random variable; we will not prove this.

There are two particularly important cases of random variables which we will focus on in this course: discrete and continuous random variables. A random variable is *continuous* if its cdf, $F$, is continuous. A random variable is *discrete* if its cdf, $F$, is a step function. We say these random variables have a continuous or discrete distribution, respectively.

Figure 2: cdfs of a continuous and discrete random variable



*Note:* Left panel is cdf of a random variable $X$ with the logistic distribution. Right panel is cdf of $Y$ the number of heads observed from tossing 3 fair coins.

Random variables $X$ and $Y$ are *identically distributed* if for every event $A \in \mathcal{B}$, $P(X \in A) = P(Y \in A)$. Note that this *does not* say that $X = Y$.

Proposition 2.4: *If $X$ and $Y$ are identically distributed their cdfs $F$ and $G$ (respectively) satisfy $F(x) = G(x)$ for all $x \in \mathbb{R}$.*

*Proof.* After possibly extending the probability functions as in footnote 23, for all $x$ we have $F(x) = P(X \in (-\infty, x]) = P(Y \in (-\infty, x]) = G(x)$. □

The converse of the above proposition is also true; that is if $F(x) = G(x)$ for all $x \in \mathbb{R}$ then $X$ and $Y$ are identically distributed. We will not prove this.

### 2.1.4 Probability mass and density functions

If $X$ has a discrete distribution, it has an associated *probability mass function* or *pmf*:

$$f(x) = P(X = x) \quad \text{for all } x \in \mathbb{R}.$$

[23] Technically if $\mathcal{X}$ is a strict subset of $\mathbb{R}$ we may need to extend the definition of $P$ to the Borel subsets of $\mathbb{R}$: for any such $B$ we define $P(B) = P(B \cap \mathcal{X})$, since the latter is in $\mathcal{B}$. This defines a probability function on the Borel subsets of $\mathbb{R}$ which agrees with the original on $\mathcal{B}$.

[24] Left continuity does not follow because if $z_n \uparrow x$ then we have $(-\infty, z_n] \uparrow (-\infty, x) \neq (-\infty, x]$.

The pmf, $f(x)$, is zero whenever $X$ takes on the value $x$ with zero probability. It is positive if when $X = x$ has positive probability: this corresponds to the points $x$ at which the cdf of $X$ jumps up and $f(x)$ is the size of that jump. It can be proven that there are only a countable number of such jumps.[25]

The probability of $X$ taking values in an event $B$ can be calculated by summation in the discrete case:

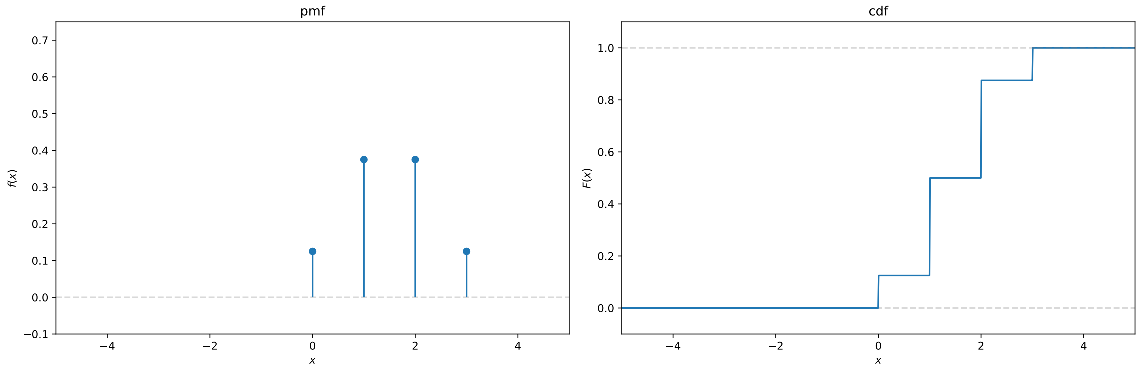
$$P(X \in B) = \sum_{t \in B \cap \mathcal{X}} f(t).$$

In particular,

$$P(X \leq x) = F(x) = \sum_{t \leq x,\ t \in \mathcal{X}} f(t).$$

The following figure shows the pmf of the distribution of the distribution in the right panel of figure 2, next to that same cdf.

Figure 3: pmf and cdf of a discrete random variable



*Note:* Left panel is pmf and right panel cdf of the number of heads observed from tossing 3 fair coins.

We now give 3 examples of commonly used discrete distributions.

Example 2.2 [Bernoulli distribution]: A random variable $X$ has a *Bernoulli(p) distribution* if for some $p \in [0, 1]$

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases},$$

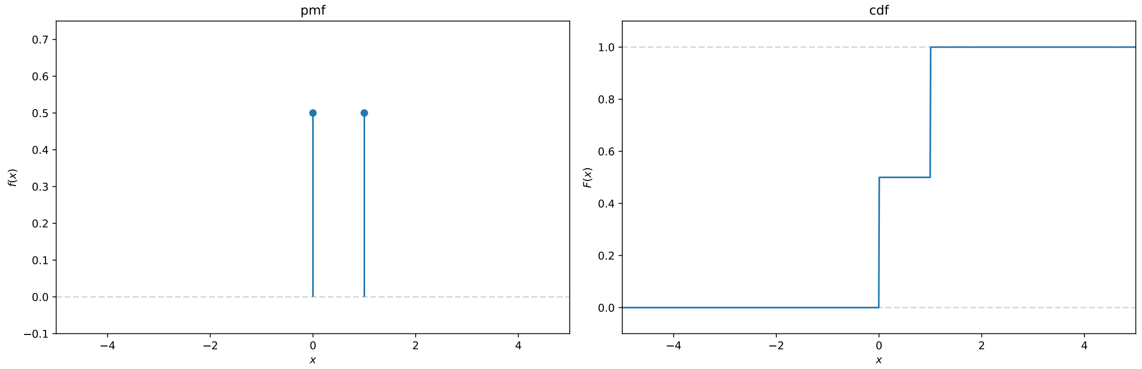
or, equivalently, its pmf is

$$f(x) = \begin{cases} p & \text{x} = 1 \\ 1 - p & \text{x} = 0 \\ 0 & \text{otherwise} \end{cases}.$$

The pmf and cdf of the Bernoulli distribution for $p = 1/2$ is plotted below.

[25] By Proposition 2.3, for any point $x$ where $\lim_{z \uparrow x} F(z) \neq F(x)$ we have that $\lim_{z \uparrow x} F(z) < F(x)$. Therefore there is a rational number $q_x$ such that $\lim_{z \uparrow x} F(z) < q_x < F(x)$. For any other such point $y$ of discontinuity (with $y \neq x$), we must have that $q_y \neq q_x$, since $F$ is nondecreasing (as this implies that we must have that $F(y) > F(x)$; if not $y$ cannot be a point of discontinuity). This gives us an injective mapping between the points of discontinuity and $\mathbb{Q}$, which shows that the former set is countable.

Figure 4: pmf and cdf of a Bernoulli$(1/2)$ random variable



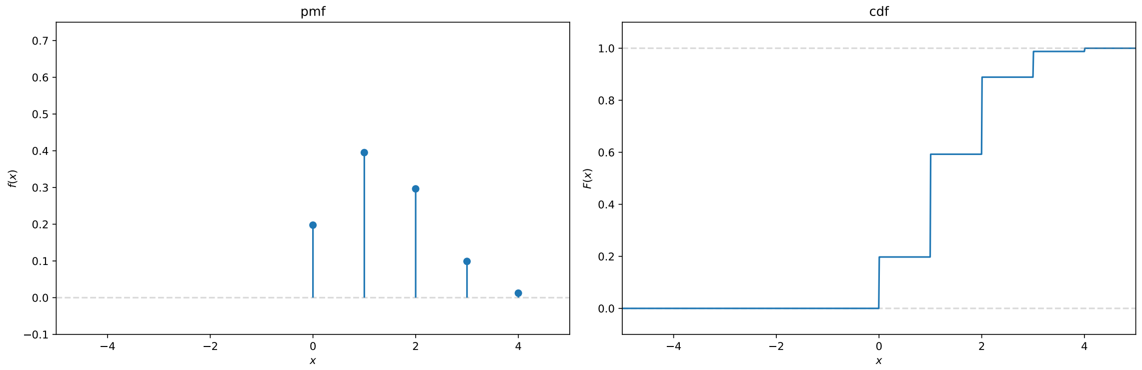
$\triangle$

Example 2.3 [Binomial distribution]: A random variable $X$ has a Binomial$(n, p)$ distribution if it has the pmf

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad \text{for } x = 0, 1, \dots, n,$$

and $f(x) = 0$ otherwise. This is equivalent to defining $X$ as the sum of $n$ independent Bernoulli$(p)$ random variables.[26] The pmf and cdf of a Binomial$(4, 1/3)$ random variable is plotted below.

Figure 5: pmf and cdf of a Binomial$(4, 1/3)$ random variable



$\triangle$

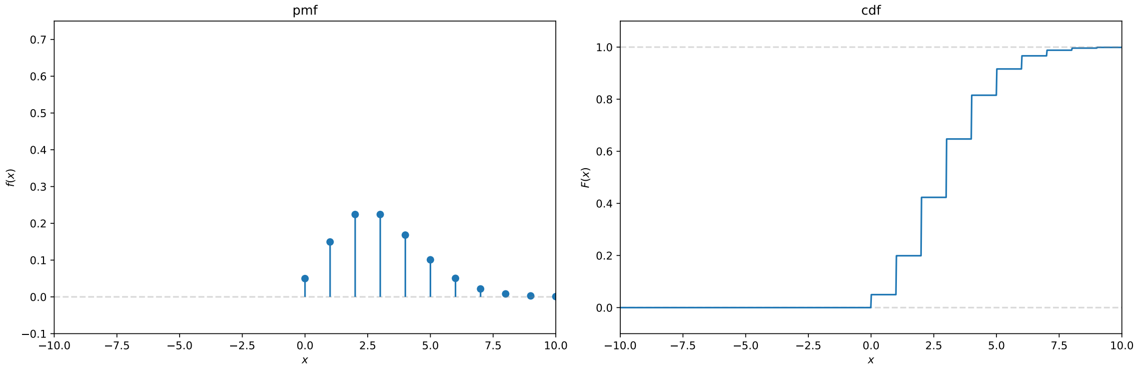
Example 2.4 [Poisson distribution]: A random variable $X$ has a Poisson$(\lambda)$ distribution if it has the pmf

$$f(x) = \frac{\exp(-\lambda)\lambda^x}{x!}, \qquad \text{for } x = 0, 1, \dots$$

and $x$ otherwise. Note this places positive mass on every positive integer. The parameter $\lambda$ is called the *intensity* parameter and the Poisson distribution is often used to model the number of events which occur in a fixed interval. The pmf and cdf of a Poisson$(3)$ random variable is plotted below.

[26] We will discuss independence in section 2.3.

Figure 6: pmf and cdf of a Poisson(3) random variable



△

If $X$ has a continuous distribution and the cdf is absolutely continuous, then it has an associated *probability density function* or *pdf*, $f$.[27] Unlike the discrete case, this does *not* give the probability that $X = x$; in fact for any $x \in \mathbb{R}$, $P(X = x) = 0$ if $X$ is a continuous random variable. Rather, $f$ satisfies

$$P(X \leq x) = F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t.$$

For any event $B$, the probability can be calculated by integration:

$$P(X \in B) = \int_{B} f(t)\,\mathrm{d}t.$$

In the following examples we discuss a few commonly used continuous distributions.

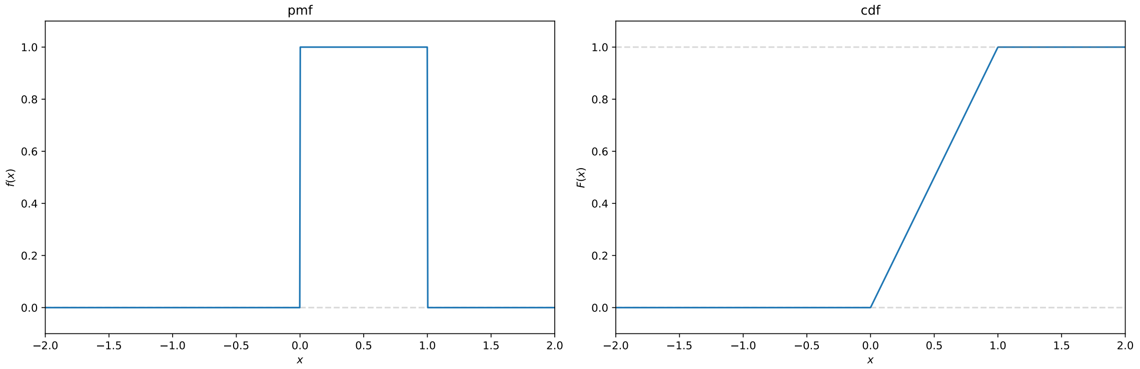
Example 2.5 [Uniform distribution]: A random variable $X$ has the Uniform$(a, b)$ distribution if its pdf is

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{whenever } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}.$$

The pdf and cdf of the Uniform distribution for $a = 0, b = 1$ is plotted below.

[27] Absolute continuity is a smoothness property of functions that is stronger than continuity. This ensures that an $f$ such that the integral representation in the subsequent display holds exists. We will not discuss the details in this course.

Figure 7: pdf and cdf of a Uniform(0, 1) random variable



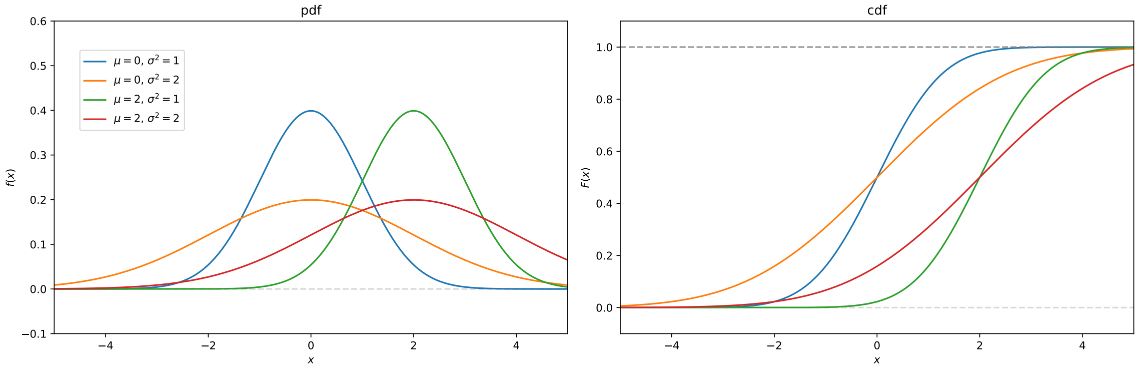
△

Example 2.6 [Normal distribution]: A random variable $X$ has the Normal$(\mu, \sigma)$ distribution if its pdf is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

$\mu$ is the *expected value* of $X$ and $\sigma^2$ the *variance*.[28] The pdf and cdf for a number of different normal distributions are plotted below.

Figure 8: pdf and cdf of a various Normal random variables



△

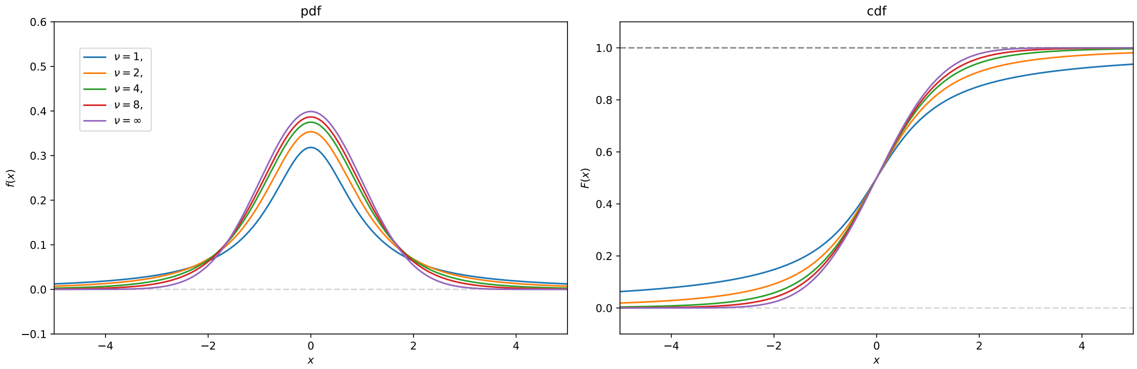
Example 2.7 [Student's t distribution]: A random variable $X$ has the (Student's) t$(\nu)$ distribution if its pdf is

$$f(x) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}.$$

$\nu$ is the *degrees of freedom*. The t(1) distribution is also called the *Cauchy* distribution. The case $\nu = \infty$ is the Normal(0, 1). The pdf and cdf for a number of different t distributions are plotted below.

[28] These concepts will be defined more generally in section 2.2.

Figure 9: pdf and cdf of a various t random variables



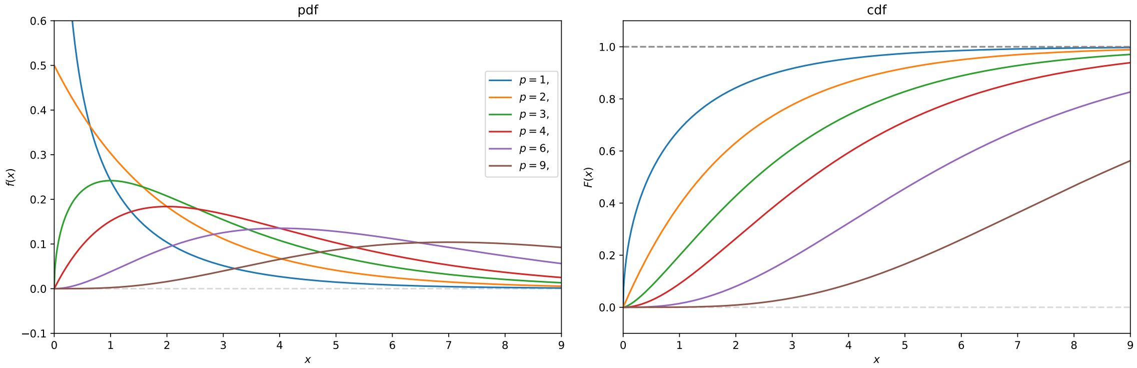
△

Example 2.8 [$\chi^2$ distribution]: $X$ has a $\chi^2(p)$ distribution ($p \in \mathbb{N}$) if its pdf is

$$f(x) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} \exp(-x/2), \qquad \text{for } x \geq 0,$$

and $f(x) = 0$ otherwise. If $X_1, \ldots, X_p$ are independent and each have a Normal(0, 1) distribution, then $X = \sum_{i=1}^{p} X_i^2$ has the $\chi^2(p)$ distribution.

Figure 10: pdf and cdf of a various $\chi^2$ random variables



△

Proposition 2.5: *If $f$ is the pdf or pmf of a random variable then*

*(i)* $f(x) \geq 0$ *for all* $x \in \mathbb{R}$*;*

*(ii)* $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$ *(pdf) or* $\sum_{x \in \mathcal{X}} f(x) = 1$ *(pmf).*

*Proof.* That $f(x) \geq 0$ follows from its definition as a probability in the discete case. In the continuous case, it follows from the fact that $f$ is the derivative of $F$ which is nondecreasing. For the second property, in the discrete case we have

$$\sum_{x \in \mathcal{X}} f(x) = \lim_{t \to \infty} \sum_{x \geq t, x \in \mathcal{X}} f(x) = \lim_{t \to \infty} F(t) = 1,$$

and similarly for the continuous case

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \lim_{t\to\infty}\int_{-\infty}^{t} f(x)\,\mathrm{d}x = \lim_{t\to\infty} F(t) = 1. \qquad \square$$

The converse is also true: if (i) and (ii) are satisfied then $f$ is a pdf / pmf. We will not prove this.

### 2.1.5 Transformations of random variables

Since a random variable $X$ is a function $X : S \to \mathbb{R}$ if we have a function $g : \mathbb{R} \to \mathbb{R}$, then $Y = g(X)$ will also be a random variable.[29] We can express the probabilistic behaviour of $Y$ in terms of $X$:

$$P(Y \in A) = P(g(X) \in A) = P(\{x \in \mathcal{X} : g(x) \in A\}) = P(X \in g^{-1}(A)), \tag{8}$$

just as we express probabilities about $X$ in terms of our original probabilities on $\mathcal{S}$.

In some cases there are simple formulae available for the pmf / pdf / CDF of $Y$.

Theorem 2.1: *If $X$ is a random variable and $Y = g(X)$ for a $g : \mathbb{R} \to \mathbb{R}$ then:*[30]

*(i) If $X$ is discrete then $Y$ is discrete and*

$$f_Y(y) = \sum_{x \in g^{-1}(\{y\})} f_X(x).$$

*(ii) If $X$ is continuous then the cdf of $Y$ is*

$$F_Y(y) = \int_{\{x : g(x) \le y\}} f_X(x)\,\mathrm{d}x.$$

*(iii) If $X$ is continuous and $g$ strictly monotone increasing with $h = g^{-1}$ differentiable then $Y$ is continuous and*

$$F_Y(y) = F_X(h(y)), \qquad f_Y(y) = f_X(h(y))h'(y).$$

*(iv) If $X$ is continuous and $g$ strictly monotone decreasing with $h = g^{-1}$ differentiable then $Y$ is continuous and*

$$F_Y(y) = 1 - F_X(h(y)), \qquad f_Y(y) = -f_X(h(y))h'(y).$$

*Proof.*

[29] Provided the function $f$ is sufficiently "nice"; all continuous functions are "nice". We will not encounter any functions $g$ in this course which are not "nice".

[30] *The stipulation that $h = g^{-1}$ be differentiable is not really needed as the inverse of a strictly monotone function is strictly monotone and any such function is differentiable a.e..

(i) By (8), $f_Y(y) = P(Y = y) = P(X \in g^{-1}(\{y\}))$. Since $X$ is discrete this is equal to $\sum_{x \in g^{-1}(\{y\})} f_X(x)$.

(ii) $F_Y(y) = P(Y \leq y) = P(g(X) \leq y)$ which can be equivalently written as the given integral.

(iii) If $g$ is strictly monotone increasing it is invertible and $\{g(x) \leq y\} = \{x \leq g^{-1}(y)\}$ so,

$$F_Y(y) = \int_{-\infty}^{g^{-1}(y)} f_X(x)\,\mathrm{d}x = F_X(g^{-1}(y)) = F_X(h(y))$$

using (ii). Since $h$ is differentiable, use the chain rule to conclude.

(iv) Argue analogously to in (iii).

□

## 2.2 Expectations

The expected value of a random variable $g(X)$ is denoted by $\mathbb{E}\,g(X)$ and is

$$\mathbb{E}\,g(X) = \int_{-\infty}^{\infty} g(x)f(x)\,\mathrm{d}x, \qquad \mathbb{E}\,g(X) = \sum_{x \in \mathcal{X}} g(x)f(x),$$

provided the integral $\mathbb{E}\,|g(X)| < \infty$ ("exists"), where $f$ denotes the pdf / pmf in the continuous and discrete cases respectively.[31] If $\mathbb{E}\,|g(X)| = \infty$ then the expectation does not exist. Taking $g$ to be the identity function, i.e. $g(x) = x$ for all $x$, yields the expectation of $X$ itself.

Proposition 2.6: *Let $X$ be a random variable, $\alpha$ a constant and $g_1$ and $g_2$ such that $\mathbb{E}\,g_1(X)$ and $\mathbb{E}\,g_2(X)$ exist. Then*

*(i) $\mathbb{E}\,[\alpha g_1(X)] = \alpha\,\mathbb{E}\,g_1(X)$ and $\mathbb{E}\,[g_1(X) + g_2(X)] = \mathbb{E}\,g_1(X) + \mathbb{E}\,g_2(X)$;*

*(ii) If $g_1(x) \geq 0$ for all $x$ with $f(x) > 0$, then $\mathbb{E}\,g_1(X) \geq 0$.*

*Proof.* We give the proof for the continuous case; the proof for the discrete case is left as an exercise. In this proof, if the bounds of an integral are $-\infty, \infty$ they are omitted in the notation. For (i) by Proposition 2.5 and linearity of the integral

$$\mathbb{E}\,[\alpha g_1(X)] = \int \alpha g_1(x)f(x)\,\mathrm{d}x = \alpha \int g_1(x)f(x)\,\mathrm{d}x = \alpha\,\mathbb{E}\,g_1(X),$$

and

$$\begin{aligned}\mathbb{E}\,[g_1(X) + g_2(X)] &= \int (g_1(x) + g_2(x))f(x)\,\mathrm{d}x \\ &= \int g_1(x)f(x)\,\mathrm{d}x + \int g_2(x)f(x)\,\mathrm{d}x \\ &= \mathbb{E}\,g_1(X) + \mathbb{E}\,g_2(X).\end{aligned}$$

[31] Expectations can be defined for other types of random variable, for example if $X$ is continuous but does not have a pdf. Such expectations continue to satisfy all the properties in 2.6, though we will not discuss their construction in this course.

For (ii), we have

$$\mathbb{E}\, g_1(X) = \int g_1(x) f(x)\, \mathrm{d}x = \int_{\{x: f(x)>0\}} g_1(x) f(x)\, \mathrm{d}x \geq 0,$$

since $g_1(x)f(x) \geq 0$ on $\{x : f(x) > 0\}$. □

[Exercise: show that $\mathbb{E}\,\alpha = \alpha$ for any constant $\alpha$.]

### 2.2.1 Moments

Important special cases of expectations are when the function $g$ is a monomial: $x, x^2, x^3$ and so on. These are called moments. The *$k$-th moment* of $X$ is $\mathbb{E}\,X^k$ (if it exists). *Central moments* are also important, and are defined by taking the $k$-th moment of the centered variable $Y = X - \mathbb{E}\,X$. That is, the $k$-th central moment of $X$ is $\mathbb{E}(X - \mathbb{E}\,X)^k$, if it exists.[32]

The first moment $\mathbb{E}\,X$ is often called the *mean* whilst the second central moment $\mathbb{E}(X - \mathbb{E}\,X)^2 = \mathbb{E}\,X^2 - (\mathbb{E}\,X)^2$ is the *variance* and is often denoted by $\mathrm{Var}(X)$ [Exercise: verify the previous equality]. The variance satisfies the following useful property

Lemma 2.1: *If* $\mathrm{Var}(X)$ *exists, then for any constants* $a, b$

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X).$$

*Proof.* Exercise. □

An important class of functions are the indicator functions. If $A$ is an event, its indicator function is $\mathbf{1}_A$:

$$\mathbf{1}_A(x) := \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}.$$

Sometimes we write $\mathbf{1}_A(X)$ as $\mathbf{1}\{X \in A\}$.

Lemma 2.2 [Properties of indicators]: *If* $A, B$ *are events and* $X$ *a random variable:*

*(i)* $\mathbf{1}_A \mathbf{1}_B = \mathbf{1}_{A \cap B}$

*(ii)* $P(X \in A) = \mathbb{E}[\mathbf{1}_A(X)]$

*(iii)* $P(X \in A)[1 - P(X \in A)] = \mathrm{Var}(\mathbf{1}_A(X))$.

*Proof.* Exercise. □

We finish this section with some examples.

[32] If the $k$-th moment exists, the $k$-th central moment also exists and vice versa.

Example 2.9 [Mean and variance of a fair die]: Let $X \in \{1, \ldots, 6\}$ be the value obtained by rolling a fair die. The mean $\mathbb{E}\, X$ is

$$\mathbb{E}\, X = \sum_{i=1}^{6} iP(X = i) = \frac{1}{6} \sum_{i=1}^{6} i = 3.5.$$

The variance $\mathrm{Var}(X)$ is

$$\mathrm{Var}(X) = \sum_{i=1}^{6} (i - \mathbb{E}\, X)^2 P(X = i) = \frac{1}{6} \sum_{i=1}^{6} (i - 3.5)^2 = \frac{35}{12}.$$

$\triangle$

Example 2.10 [Mean and variance of Normal distribution]: Let $X$ have the Normal$(\mu, \sigma)$ distribution. Then $\mathbb{E}\, X = \mu$ and $\mathrm{Var}(X) = \sigma^2$. The calcuations go as follows. First we note that if $Z = (X - \mu)/\sigma$ then $Z$ has a standard normal (i.e. Normal(0, 1)) distribution since

$$\begin{aligned} P(Z \le z) &= P\left(\frac{X - \mu}{\sigma} \le z\right) \\ &= P(X \le z\sigma + \mu) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{z\sigma+\mu} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(-t^2/2)\, \mathrm{d}t, \end{aligned}$$

where we used the substitution $t = (x - \mu)/\sigma$. Therefore by Proposition 2.6 and Lemma 2.1 it suffices to show that $\mathbb{E}\, X = 0$ and $\mathbb{E}\, X^2 = 1$ [Exercise: verify this is sufficient]. For this, we have (with all integrals from $-\infty$ to $\infty$ if not otherwise noted)

$$\begin{aligned} \mathbb{E}\, X &= \int x f(x)\, \mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int x \exp\left(-\frac{x^2}{2}\right)\, \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \left[\int_{-\infty}^{0} x \exp\left(-\frac{x^2}{2}\right)\, \mathrm{d}x + \int_{0}^{\infty} x \exp\left(-\frac{x^2}{2}\right)\, \mathrm{d}x\right] \\ &= \frac{1}{\sqrt{2\pi}} \left[\int_{0}^{\infty} (-x) \exp\left(-\frac{x^2}{2}\right)\, \mathrm{d}x + \int_{0}^{\infty} x \exp\left(-\frac{x^2}{2}\right)\, \mathrm{d}x\right] \\ &= \frac{1}{\sqrt{2\pi}} \left[\int_{0}^{\infty} x \exp\left(-\frac{x^2}{2}\right)\, \mathrm{d}x - \int_{0}^{\infty} x \exp\left(-\frac{x^2}{2}\right)\, \mathrm{d}x\right] \\ &= 0. \end{aligned}$$

For the second moment, integrating by parts yields

$$\begin{aligned}\mathbb{E}\, X^2 = \int x^2 f(x)\,\mathrm{d}x &= \frac{1}{\sqrt{2\pi}} \int x^2 \exp\left(-\frac{x^2}{2}\right)\,\mathrm{d}x \\ &= \frac{2}{\sqrt{2\pi}} \int_0^\infty x^2 \exp\left(-\frac{x^2}{2}\right)\,\mathrm{d}x \\ &= \frac{2}{\sqrt{2\pi}} \left[\lim_{x\to 0} x\exp(-x^2/2) - \lim_{x\to\infty} x\exp(-x^2/2) + \int_0^\infty \exp\left(-\frac{x^2}{2}\right)\,\mathrm{d}x\right] \\ &= 1,\end{aligned}$$

by the fact that $\frac{2}{\sqrt{2\pi}}\int_0^\infty \exp\left(-\frac{x^2}{2}\right)\,\mathrm{d}x = \int \exp\left(-\frac{x^2}{2}\right)\,\mathrm{d}x = 1$, since it is the integral of the pdf of a Normal(0, 1) random variable. △

## 2.3 Conditioning and independence

### 2.3.1 Conditional probability and independence

If $A$ and $B$ are events and $P(B) > 0$, then the *conditional probability* of $A$ given $B$ is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

If we rearrange this we obtain

$$P(A \cap B) = P(A|B)P(B).$$

Since we could also condition on $A$, provided that $P(A) > 0$, then also

$$P(A \cap B) = P(B|A)P(A).$$

Using these we arrive at *Bayes rule*:

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)}.$$

Events $A$ and $B$ are *independent* if

$$P(A \cap B) = P(A)P(B).$$

Note that this implies that if $A$ and $B$ are independent, then $P(A|B) = P(A)$ and $P(B|A) = P(B)$ by Bayes rule (provided each denominator is appropriately non-zero).

Extending independence to more than two events requires some care: events $A_1, \ldots, A_n$ are *mutually independent* if for any subcollection $A_{i_1}, \ldots, A_{i_k}$,

$$P\left(\cap_{j=1}^k A_{i_j}\right) = \prod_{j=1}^k P(A_{i_j}).$$

### 2.3.2 Joint and marginal distributions

A *random vector*, $X$ is a function $X : S \to \mathbb{R}^K$.[33] Just like random variables, random vectors have cdfs. The distribution function of $X$ is defined similarly to the case where $K = 1$.

$$F(x) = P(X_1 \leq x_1, \ldots, X_K \leq x_K), \qquad x \in \mathbb{R}^K.$$

$F$ describes the *joint distribution* of $X$. Each component, $X_k$ ($k = 1, \ldots, K$) is a random variable, and therefore also has a distribution function: $F_k(x) = P(X_k \leq x)$ (defined for $x \in \mathbb{R}$); this describes the *marginal distribution* of $X_k$ and gives no information about the other components in $X$.

In addition to describing the random behaviour of each $X_k$, the joint distribution of $X$ also describes the dependence between them. There is one special case where the marginal distributions fully characterise the joint distribution. Two random variables, $X$ and $Y$ are independent if for any events $A, B$ in $\mathbb{R}$, $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$. This implies that their joint cdf factors into the product of their marginal cdfs:

$$F_{X,Y}(x) = F_X(x_1)F_Y(x_2).$$

More generally, $X_1, \ldots, X_K$ are independent if for any events $A_i \subset \mathbb{R}$ ($i = 1, \ldots, K$), we have

$$P\left(\cap_{i=1}^K \{X_i \in A_i\}\right) = \prod_{i=1}^K P(X_i \in A_i).$$

Again this implies that their joint cdf factors:

$$F_{X_1,\ldots,X_K}(x) = F_{X_1}(x_1) \cdots F_{X_K}(x_K),$$

and so if $X = (X_1, \ldots, X_K)$ is comprised of independent random variables, its cdf is the product of the marginal cdfs of $X_1, \ldots, X_K$.

The idea of being identically distributed extends immediately to random vectors: $X$ and $Y$ are identically distributed if $P(X \in A) = P(Y \in A)$ for all events $A$. Just like in the scalar case, this is satisfied if and only if $F_X(x) = F_Y(x)$ for all $x \in \mathbb{R}^K$. It is *not* sufficient however that $F_{X_k}(x) = F_{Y_k}(x)$ for each $k = 1, \ldots, K$ and each $x \in \mathbb{R}$.[34]

### 2.3.3 Joint, marginal and conditional mass and density functions

If $X$ is a discrete random vector, its *joint pmf* is the function $f : \mathbb{R}^K \to \mathbb{R}$ defined by

$$f(x) = P(X_1 = x_1, \ldots, X_K = x_K), \quad x \in \mathbb{R}^K.$$

[33] Now our $\sigma$-field of events is comprised of subsets of $\mathbb{R}^K$.

[34] This *is* true if $X$ and $Y$ are identically distributed, but it does *not* imply that $X$ and $Y$ are identically distributed.

It follows that for any event $A$,[35]

$$P(X \in A) = \sum_{x \in A} f(x).$$

The *marginal pmf* of any co-ordinate $X_k$ is just the (univariate) pmf of $X_k$ as defined in section 2.1.4. The marginal pmf of a subcollection $X_{i_1}, \ldots, X_{i_k}$ (of $X_1, \ldots, X_K$) is the joint pmf of $X_{i_1}, \ldots, X_{i_k}$.

The *conditional pmf* of $X_{k+1}, \ldots X_K$ given $X_1 = x_1, \ldots, X_k = x_k$ is

$$\begin{aligned} f(x_{k+1}, \ldots, x_K | x_1, \ldots, x_k) &= P(X_{k+1} = x_{k+1}, \ldots, X_K = x_K | X_1 = x_1, \ldots, X_k = x_k) \\ &= \frac{P(X_1 = x_1, \ldots, X_K = x_K)}{P(X_1 = x_1, \ldots, X_k = x_k)} \\ &= \frac{f_X(x_1, \ldots, x_K)}{f_{X_1, \ldots, X_k}(x_1, \ldots, x_k)}, \end{aligned}$$

provided $f_{X_1, \ldots, X_k}(x_1, \ldots, x_k) > 0$, as follows directly from the definition of conditional probability. The joint pmf of $X$ can therefore be calculated as the product of the conditional pmf of $X_{k+1}, \ldots X_K$ given $X_1, \ldots, X_k$, multiplied by the marginal pmf of $X_1, \ldots, X_k$.

Marginal pmfs can be recovered from joint pmfs.

Proposition 2.7: *Suppose that $X = (X_1, \ldots, X_K)$ is discrete. Then the marginal pmf of $X_1, \ldots, X_k$ is given by*

$$f_{X_1, \ldots, X_k}(x_1, \ldots, x_k) = \sum_{(x_{k+1}, \ldots, x_K) \in \mathbb{R}^{K-k}} f_X(x_1, \ldots, x_K).$$

*Proof.* Let $A_{x_1, \ldots, x_k} = \{y \in \mathbb{R}^K : y_i = x_i \text{ for } i = 1, \ldots, k\}$. Then,

$$\begin{aligned} f_{X_1, \ldots, X_k}(x_1, \ldots, x_k) &= P(X_1 = x_1, \ldots, X_k = x_k) \\ &= P(X \in A_{x_1, \ldots, x_k}) \\ &= \sum_{y \in A_{x_1, \ldots, x_k}} f_X(y_1, \ldots, y_n) \\ &= \sum_{(x_{k+1}, \ldots, x_K) \in \mathbb{R}^{K-k}} f_X(x_1, \ldots, x_K). \end{aligned}$$ □

If $X$ is a continuous random vector (i.e. its cdf is continuous) it has a *joint pdf* if there is a (non-negative) function $f : \mathbb{R}^K \to \mathbb{R}$ such that for any event $A \subset \mathbb{R}^K$,

$$P(X \in A) = \int_A f(x) \, \mathrm{d}x.$$

Similarly to the discrete case, we can define the marginal pdf of any co-ordinate $X_k$ simply as the pdf of the random variable $X_k$. The marginal pdf of a subcollection $X_{i_1}, \ldots, X_{i_k}$ (of

[35] If $X$ is discrete, any event $A$ can contain only countably many $x$ with $f(x) \neq 0$; since each $\{x\} \cap \{y\} = \emptyset$ if $x \neq y$ this follows the third defining property of a probability.

$X_1, \dots, X_K)$ is the joint pdf of $X_{i_1}, \dots, X_{i_k}$.

The conditional pdf of $X_{k+1}, \dots X_K$ given $X_1 = x_1, \dots, X_k = x_k$ is

$$f(x_{k+1}, \dots, x_K | x_1, \dots, x_k) = \frac{f_X(x_1, \dots, x_K)}{f_{X_1,\dots,X_k}(x_1, \dots, x_k)},$$

provided $f_{X_1,\dots,X_k}(x_1, \dots, x_k) > 0$. Again, the joint pdf of $X$ can therefore be calculated as the product of the conditional pdf of $X_{k+1}, \dots X_K$ given $X_1, \dots, X_k$, multiplied by the marginal pdf of $X_1, \dots, X_k$.

Marginal pdfs can be recovered from joint pdfs.

Proposition 2.8: *Suppose that $X = (X_1, \dots, X_K)$ is continuous. Then the marginal pdf of $X_1, \dots, X_k$ is given by*

$$f_{X_1,\dots,X_k}(x_1, \dots, x_k) = \int_{\mathbb{R}^{K-k}} f_X(x_1, \dots, x_K) \, \mathrm{d}(x_{k+1}, \dots, x_K).$$

*Proof.* Let $A \subset \mathbb{R}^k$ be an event and define the event $B = \{x \in \mathbb{R}^K : (x_1, \dots, x_k) \in A\}$. By properties of the integral,

$$\begin{aligned} P((X_1, \dots, X_k) \in A) &= P(X \in B) \\ &= \int_B f_X((x_1, \dots, x_K) \, \mathrm{d}(x_1, \dots, x_K)) \\ &= \int_A \int_{\mathbb{R}^{K-k}} f_X((x_1, \dots, x_K) \, \mathrm{d}(x_{k+1}, \dots, x_K)) \, \mathrm{d}(x_1, \dots, x_k) \\ &= \int_A f_{X_1,\dots,X_k}(x_1, \dots, x_k) \, \mathrm{d}(x_1, \dots, x_k). \qquad \square \end{aligned}$$

In both the continuous and discrete cases, independence leads to a simplification. In the discrete case, if $X_1, \dots, X_K$ are independent we have

$$f(x_1, \dots, x_K) = P(X_1 = x_1, \dots, X_K = x_k) = P(X_1 = x_1) \cdots P(X_K = x_k) = \prod_{i=1}^K f_{X_i}(x_i).$$

In the continuous case also

$$f(x_1, \dots, x_K) = f_{X_1}(x_1) \cdots f_{X_K}(x_K),$$

since for any event $A = A_1 \times \cdots \times A_K \subset \mathbb{R}^K$, we have

$$\begin{aligned} P(X \in A) &= P(X_1 \in A_1) \times \cdots \times P(X_K \in A_K) \\ &= \int_{A_1} f_{X_1}(x_1) \, \mathrm{d}x_1 \times \cdots \times \int_{A_K} f_{X_K}(x_K) \, \mathrm{d}x_K \\ &= \int_{A_1} \int_{A_2} \cdots \int_{A_K} f_{X_1}(x_1) \times \cdots \times f_{X_K}(x_K) \, \mathrm{d}x_1 \cdots \mathrm{d}x_K \\ &= \int_A f_{X_1}(x_1) \times \cdots \times f_{X_K}(x_K) \, \mathrm{d}x. \end{aligned}$$

In either case this implies that the conditional pmf/pdf is just the marginal pmf/pdf:

$$\begin{aligned} f(x_{k+1}, \dots, x_K | x_1, \dots, x_k) &= \frac{f_X(x_1, \dots, x_K)}{f_{X_1,\dots,X_k}(x_1, \dots, x_k)} \\ &= \frac{\prod_{i=1}^{K} f_{X_i}(x_i)}{\prod_{i=1}^{k} f_{X_i}(x_i)} \\ &= \prod_{i=k+1}^{K} f_{X_i}(x_i) \\ &= f_{X_{k+1},\dots,X_K}(x_{k+1}, \dots, x_K). \end{aligned}$$

Example 2.11: Suppose that $X = (X_1, X_2)$ where $X_1$ and $X_2$ are independent and both have the standard normal (i.e. $N(0,1)$) distribution. Then, their marginal pdfs are

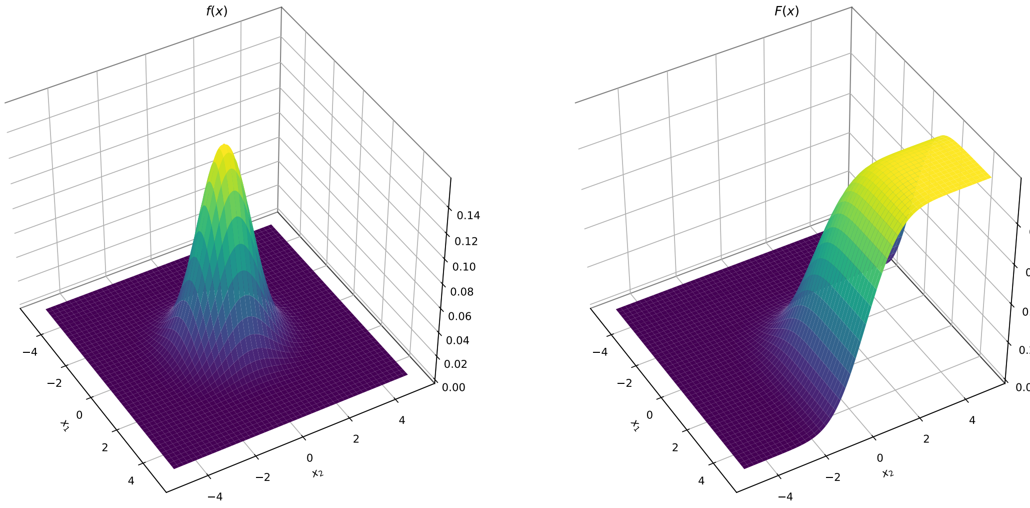
$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2}\right)$$

and the joint pdf of $X$ is

$$f(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\left[x_1^2 + x_2^2\right]\right).$$

The following figure plots the (joint) pdf and cdf of $X$.

Figure 11: pdf and cdf of bivariate standard normal random variables



△

We sometimes write, for example $f_{X|Y}(x|y)$, for the conditional pmf or pdf (of $X$ given $Y = y$) to avoid unnecessary confusion in the notation.

### 2.3.4 Covariance and correlation

The expectation of a function of a random vector, $X$, is defined in the same way as the expectation of a function of a random variable. Specfically, suppose that $X = (X_1, \dots, X_K)$ is a random vector and $g : \mathbb{R}^K \to \mathbb{R}$. Then

$$\mathbb{E}\, g(X) = \int_{\mathbb{R}^K} g(x) f(x) \,\mathrm{d}x,$$

where $f$ is the joint pmf or pdf of $X$, provided again that $\mathbb{E}\,|g(X)| < \infty$. The expectation of functions $g : \mathbb{R}^K \to \mathbb{R}^m$ or $g : \mathbb{R}^K \to \mathbb{R}^{m \times n}$ are defined pointwise, provided all $\mathbb{E}\,|g_k(X)|$ or $\mathbb{E}\,|g_{kl}(X)|$ (respectively) are finite.

A special case of this is the *covariance* between two random variables $X$ and $Y$. If $g : \mathbb{R}^2 \to \mathbb{R}$ is $g(x, y) = (x - \mathbb{E}\, X)(y - \mathbb{E}\, Y)$, the covariance, $\mathrm{Cov}(X, Y)$, is the expectation $\mathbb{E}\, g(X, Y)$. That is:

$$\mathrm{Cov}(X, Y) = \mathbb{E}\,[(X - \mathbb{E}\, X)(Y - \mathbb{E}\, Y)] = \int_{\mathbb{R}^2} (x - \mathbb{E}\, X)(y - \mathbb{E}\, Y) f(x, y) \,\mathrm{d}(x, y).$$

A sufficient condition for the covariance to exist is that $\mathbb{E}\, X^2 < \infty$ and $\mathbb{E}\, Y^2 < \infty$.[36] It is clear from the definition that $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$.

The *correlation* is the covariance normalised by the variances:

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}},$$

this takes values in $[-1, 1]$. Two random variables $X$ and $Y$ are said to be *uncorrelated* if $\mathrm{Corr}(X, Y) = 0$ or equivalently $\mathrm{Cov}(X, Y) = 0$.

If $X = (X_1, \dots, X_K)$ is a random vector, its expectation is

$$\mathbb{E}\, X = \begin{pmatrix} \mathbb{E}\, X_1 \\ \mathbb{E}\, X_2 \\ \vdots \\ \mathbb{E}\, X_K \end{pmatrix},$$

provided all $\mathbb{E}\,|X_k| < \infty$. If $X$ and $Y$ are random vectors in $\mathbb{R}^K$ and $\mathbb{R}^L$ respectively, their *(cross)-covariance matrix*, $\mathrm{Cov}(X, Y)$ is defined as

$$\mathrm{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}\, X)(Y - \mathbb{E}\, Y)'\right] = \begin{pmatrix} \mathbb{E}[\tilde{X}_1 \tilde{Y}_1] & \mathbb{E}[\tilde{X}_1 \tilde{Y}_2] & \cdots & \mathbb{E}[\tilde{X}_1 \tilde{Y}_L] \\ \mathbb{E}[\tilde{X}_2 \tilde{Y}_1] & \mathbb{E}[\tilde{X}_2 \tilde{Y}_2] & \cdots & \mathbb{E}[\tilde{X}_2 \tilde{Y}_L] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[\tilde{X}_K \tilde{Y}_1] & \mathbb{E}[\tilde{X}_K \tilde{Y}_2] & \cdots & \mathbb{E}[\tilde{X}_K \tilde{Y}_L] \end{pmatrix},$$

where $\tilde{Z} = Z - \mathbb{E}\, Z$, provided each $\mathbb{E}\,|\tilde{X}_k \tilde{Y}_l| < \infty$. Note that $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)'$. In the case where $X = Y$, this is the *variance matrix* $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$. As in the scalar case:

[36] This follows from the Cauchy – Schwarz inequality.

Lemma 2.3: *If $X = (X_1, \ldots, X_K)$ is a random vector and $\mathrm{Var}(X)$ exists then for any (constant) vector $b \in \mathbb{R}^K$ and any (constant) matrix $A \in \mathbb{R}^{m \times K}$,*

$$\mathrm{Var}(AX + b) = A\mathrm{Var}(X)A'.$$

*Proof.* Exercise. □

The covariance also satisfies similar properties.

Lemma 2.4: *If $X, Z$ and $Y$ are random vectors in $\mathbb{R}^K, \mathbb{R}^K$ and $\mathbb{R}^L$ respectively, $a \in \mathbb{R}^K, b \in \mathbb{R}^L$ are constant vectors and $A, B$ are constant matrices with $K$ and $L$ columns respectively,*

$$\mathrm{Cov}(X + Z, Y) = \mathrm{Cov}(X, Y) + \mathrm{Cov}(Z, Y),$$

*and*

$$\mathrm{Cov}(AX + a, BY + b) = A\mathrm{Cov}(X, Y)B'.$$

*Proof.* Exercise. □

The variance of sums of random variables/vectors satisfies the following relationship.

Lemma 2.5: *If $X, Y$ are random variables then*

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y).$$

*If $X$ and $Y$ are random vectors of the same dimension then*

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + \mathrm{Cov}(X, Y) + \mathrm{Cov}(X, Y)'.$$

*Proof.* Exercise. □

Calculation of expectations often simplifies under independence.

Proposition 2.9: *If $g_1, \ldots, g_K$ are functions such that $g_i : \mathbb{R} \to \mathbb{R}$ and $X_1, \ldots, X_K$ are independent and each $\mathbb{E}\,|g_i(X_i)| < \infty$, then*

$$\mathbb{E}\left[g_1(X_1) \times \cdots \times g_K(X_K)\right] = \mathbb{E}\left[g_1(X_1)\right] \times \cdots \times \mathbb{E}\left[g_K(X_K)\right].$$

*Proof.* We give the proof for the continuous case; the proof for the discrete case is similar.

$$\begin{aligned}
\mathbb{E}\left[g_1(X_1)\right] \times \cdots \times \mathbb{E}\left[g_K(X_K)\right] &= \int_{\mathbb{R}} g_1(x_1) f_{X_1}(x_1)\,\mathrm{d}x_1 \times \cdots \times \int_{\mathbb{R}} g_K(x_1) f_{X_K}(x_K)\,\mathrm{d}x_K \\
&= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} g_1(x_1) f_{X_1}(x_1) \times \cdots g_K(x_K) f_{X_K}(x_K)\,\mathrm{d}x_1 \cdots \mathrm{d}x_K \\
&= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \left[g_1(x_1) \times \cdots \times g_K(x_K)\right] f(x)\,\mathrm{d}x_1 \cdots \mathrm{d}x_K \\
&= \int_{\mathbb{R}^K} \left[g_1(x_1) \times \cdots \times g_K(x_K)\right] f(x)\,\mathrm{d}x \\
&= \mathbb{E}\left[g_1(X_1) \times \cdots \times g_K(X_K)\right]. \qquad \square
\end{aligned}$$

Corollary 2.1: *If $X$ and $Y$ are independent, then* $\operatorname{Cov}(X, Y) = 0$.

*Proof.* Exercise. □

A distribution of particular importance is the multivariate normal distribution.

Example 2.12 [Multivariate Normal]: A random vector $X = (X_1, \ldots, X_K)$ has the multivariate normal distribution with mean vector $\mathbb{E}\,X = \mu$ and positive definite variance matrix $\mathbb{E}(X - \mu)(X - \mu)' = \Sigma$ if its joint pdf is[37]

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

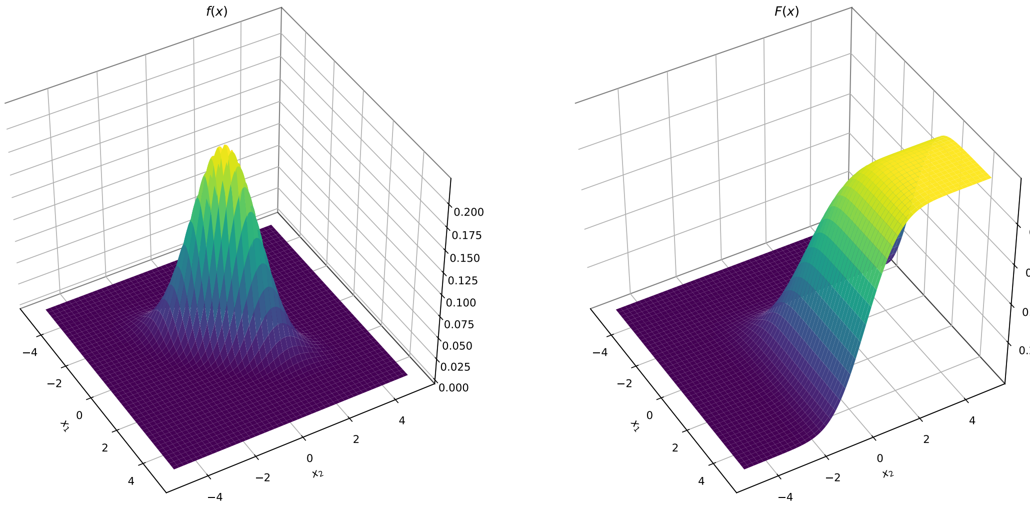
If $X = (X_1, \ldots, X_k)$ is multivariate normally distributed, then each $X_k$ is marginally normally distributed. The converse is not necessarily true. It does hold in the special case that $X_1, \ldots, X_K$ are independent.

If $X$ is multivariate normally distributed and $\operatorname{Cov}(X_i, X_j) = 0$, then $X_i$ and $X_j$ are independent. If $\operatorname{Var}(X_1, \ldots, X_k)$ is diagonal, $X_1, \ldots, X_k$ are independent.

The following figure plots the (joint) pdf and cdf of multivariate normal $X = (X_1, X_2)$, with mean $\mu = 0$ and variance $\Sigma = \left[\begin{smallmatrix} 1 & 0.7 \\ 0.7 & 1.0 \end{smallmatrix}\right]$.

[37] $\det : \mathbb{R}^{m \times m} \to \mathbb{R}$ is a polynomial function of the entries of a square matrix. It is equal to zero if and only if the matrix is not invertible.

FIGURE 12: PDF AND CDF OF BIVARIATE STANDARD NORMAL RANDOM VARIABLES



△

### 2.3.5 Conditional expectation

Given two random vectors, $X = (X_1, \dots, X_K)$ and $Y = (Y_1, \dots, Y_m)$ we can define the conditional expectation of $g(X)$ given $Y = y$:

$$\mathbb{E}\left[g(X)|y\right] = \int_{\mathbb{R}^K} g(x) f(x|y) dx \quad \text{or} \quad \mathbb{E}\left[g(X)|y\right] = \sum_{x \in \mathcal{X}} g(x) f(x|y),$$

where $f(x|y)$ is the conditional pdf or pmf in the continuous and discrete case respectively.[38] The conditional expectation exists when $\mathbb{E}\, g(X)$ does, i.e. when $\mathbb{E}\,|g(X)| < \infty$. Note that $h(y) = \mathbb{E}[g(X)|y]$ is a function of $y$ and hence $h(Y) = \mathbb{E}[g(X)|Y]$ is a random variable / vector (depending on the range of $g$).

The conditional expectation satisfies the same properties as listed in Proposition 2.6 for the usual expectation.

PROPOSITION 2.10: *Let $X$ and $Y$ be random vectors, $\alpha$ a constant and $g_1$ and $g_2$ such that $\mathbb{E}\, g_1(X)$ and $\mathbb{E}\, g_2(X)$ exist. Then*

*(i)* $\mathbb{E}\left[\alpha g_1(X)|Y\right] = \alpha\, \mathbb{E}\left[g_1(X)|Y\right]$ *and* $\mathbb{E}\left[g_1(X) + g_2(X)|Y\right] = \mathbb{E}[g_1(X)|Y] + \mathbb{E}[g_2(X)|Y]$*;*

*(ii) If* $g_1(x) \geq 0$ *for all* $x$ *with* $f(x|y) > 0$*, then* $\mathbb{E}[g_1(X)|Y] \geq 0$.

*Proof.* Exercise. □

Functions of the conditioning variable/vector may be "pulled out":[39]

[38] $\mathcal{X}$ is the range of $X$.

[39] The same conclusion is true if $h$ and $g$ are matrix-valued, provided that the dimensions are such that the product $h(y)g(x)$ is defined.

PROPOSITION 2.11: *Let $X$, $Y$ be random vectors in $\mathbb{R}^K$ and $\mathbb{R}^L$ respectively. If $h : \mathbb{R}^L \to \mathbb{R}$ and $g : \mathbb{R}^K \to \mathbb{R}$ then*

$$\mathbb{E}\left[h(Y)g(X)|Y\right] = h(Y)\,\mathbb{E}\left[g(X)|Y\right].$$

*Proof.* We prove the continuous case; the proof for the discrete case is the same, replacing integrals with sums.

$$\begin{aligned}\mathbb{E}\left[h(Y)g(X)|Y\right] &= \int h(Y)g(x)f_{X|Y}(x|Y)\,\mathrm{d}x \\ &= h(Y)\int g(x)f_{X|Y}(x|Y)\,\mathrm{d}x \\ &= h(Y)\,\mathbb{E}\left[g(X)|Y\right]. \qquad \square\end{aligned}$$

A particularly useful fact is the *law of iterated expectations* or *LIE*.

PROPOSITION 2.12 [Law of iterated expectation]: *Let $X = (X_1, \dots, X_K)$ and $Y = (Y_1, \dots, Y_m)$ be random vectors and $g : \mathbb{R}^K \to \mathbb{R}^n$. Then*

$$\mathbb{E}\left[g(X)\right] = \mathbb{E}\left[\mathbb{E}\left[g(X)|Y\right]\right].$$

*Proof.* We prove the continuous case; the proof for the discrete case is the same, replacing integrals with sums. Let $q(x, y) = g(x)$. Then we have

$$\begin{aligned}\mathbb{E}\left[g(X)\right] &= \mathbb{E}\left[q(X, Y)\right] \\ &= \int_{\mathbb{R}^m}\int_{\mathbb{R}^K} q(x, y)f(x, y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{\mathbb{R}^m}\int_{\mathbb{R}^K} g(x)f(x, y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{\mathbb{R}^m}\left[\int_{\mathbb{R}^K} g(x)f_{X|Y}(x|y)\,\mathrm{d}x\right] f_Y(y)\,\mathrm{d}y.\end{aligned}$$

Now, writing $h(y) = \mathbb{E}[g(X)|y]$ we have that

$$h(y) = \mathbb{E}[g(X)|y] = \int_{\mathbb{R}^K} g(x)f_{X|Y}(x|y)\,\mathrm{d}x,$$

which is the bracketed term in the last line of the first display. As such

$$\int_{\mathbb{R}^m}\left[\int_{\mathbb{R}^K} g(x)f_{X|Y}(x|y)\,\mathrm{d}x\right] f_Y(y)\,\mathrm{d}y = \int_{\mathbb{R}^m} h(y)f_Y(y)\,\mathrm{d}y = \mathbb{E}[h(Y)] = \mathbb{E}\left[\mathbb{E}[g(X)|Y]\right]. \qquad \square$$

### 2.3.6 Transformations of random vectors

Here we record a useful result like Theorem 2.1 for case of continuous random vectors.

THEOREM 2.2: *Suppose that $X$ is a continuous random vector with range $U \subset \mathbb{R}^K$ and $g$ is a continuously differentiable mapping from $U \to \mathbb{R}^K$ which is injective on $U$. Let the Jacobian of* *g be G:*

$$G(x) := \left[\frac{\partial g_i(x)}{\partial x_j}\right]_{i,j=1}^{n}$$

*and suppose that for all $x \in U$, $\det G(x) \neq 0$. Then $Y = g(X)$ is a continuous random variable with*

$$f_Y(y) = |\det G(g^{-1}(y))| f_X(g^{-1}(y)).$$

**Proof:* Apply the change of variables theorem from calculus and combine with the fact that bounded continuous functions form a separating class. □

## 2.4 Some important inequalities and other results

Here we'll record a few useful results. The first is a fundamental inequality which allows us to estimate the *tail probability* using moments.

THEOREM 2.3 [Markov's inequality]: *If $g$ is a non-negative, non-decreasing function satisfying $\mathbb{E}\, g(\|X\|) < \infty$ and $x$ is such that $g(x) > 0$, then*

$$P(\|X\| > x) \leq \frac{\mathbb{E}\, g(\|X\|)}{g(x)}.$$

*Proof.* $\mathbb{E}\, g(\|X\|) \geq \mathbb{E}\, g(\|X\|)\mathbf{1}\{\|X\| > x\} \geq g(x)\,\mathbb{E}\,\mathbf{1}\{\|X\| > x\} = g(x)P(\|X\| > x)$. □

COROLLARY 2.2: *If $t > 0$ then*

$$P(\|X\| > t) \leq t^{-p}\,\mathbb{E}\,\|X\|^p.$$

*Proof.* Exercise. □

Let $p \geq 1$ and let $L_p$ be the collection of random variables (on a given probability space) with finite $p$-th moment: $L_p := \{X : \mathbb{E}\,|X|^p < \infty\}$.[40] For $X \in L_p$ we define the $L_p$ norm as $\|X\|_{L_p} := (\mathbb{E}\,|X|^p)^{1/p}$ if $p < \infty$ and as $\|X\|_{L_\infty} := \inf\{a : P(|X| > a) = 0\}$ for $p = \infty$. This norm turns $L_p$ into a Banach space.

*THEOREM 2.4: *$(L_p, \|\cdot\|_{L_p})$ is a Banach space.*

**Proof:* This is proven on pp. 242 – 243 & Theorem 19.1 in [1]. □

We record here also the following useful results.

THEOREM 2.5 [Hölder's inequality]: *Let $p, q$ be such that either both are in $(1, \infty)$ and $p^{-1} + q^{-1} = 1$ or one is $1$ and the other $\infty$.*

$$\|XY\|_{L_1} \leq \|X\|_{L_p}\|Y\|_{L_q}.$$

[40] *Here we identify random variables which are equal $P$-a.s. and so, formally speaking, $L_p$ is the collection of equivalence classes of random variables which are equal $P$-a.s..

**Proof:* See Theorem 1.5.2 in [5] for the case with $1 < p, q < \infty$. For the other case one has

$$\|XY\|_{L_1} = \int |XY| \,\mathrm{d}P \leq \int \|X\|_{L_\infty} |Y| \,\mathrm{d}P = \|X\|_{L_\infty} \|Y\|_{L_1}. \qquad \square$$

If $X, Y \in L_2$ we can define an inner product as $\langle X, Y \rangle_{L_2} := \mathbb{E}[XY]$. This inner product generates the $L_2$ norm and with this inner product $L_2$ is a Hilbert space.

*THEOREM 2.6: $(L_2, \langle \cdot, \cdot \rangle_{L_2})$ *is a Hilbert space.*

**Proof:* We check the 3 properties of an inner product:

(i) Symmetry: one has $\langle X, Y \rangle_{L_2} = \mathbb{E}[XY] = \mathbb{E}[YX] = \langle Y, X \rangle_{L_2}$

(ii) Linearity in the first argument: for $a, b \in \mathbb{R}$ and $X, Y, Z \in L_2$, $\langle aX + bY, Z \rangle_{L_2} = a\,\mathbb{E}[XZ] + b\,\mathbb{E}[YZ] = a\langle X, Z \rangle_{L_2} + b\langle Y, Z \rangle_{L_2}$.

(iii) Positive-definiteness: if $X \neq 0$ then $\langle X, X \rangle_{L_2} = \mathbb{E}[X^2] > 0$.

Since $\|X\|_{L_2}^2 = \langle X, X \rangle_{L_2}$, that $L_2$ is complete (under the norm induced by $\langle \cdot, \cdot \rangle_{L_2}$) follows from Theorem 2.4. □

COROLLARY 2.3 [Cauchy – Schwarz inequality]: *If $X, Y \in L_2$, then*

$$|\langle X, Y \rangle_{L_2}| = |\,\mathbb{E}[XY]| \leq \mathbb{E}[X^2]^{1/2}\,\mathbb{E}[Y^2]^{1/2} = \|X\|_{L_2} \|Y\|_{L_2}.$$

*Proof.* Apply Hölder's inequality. □

We now consider *Jensen's inequality.*

THEOREM 2.7 [Jensen's inequality]: *Let $X$ be an integrable random variable and $g$ a convex function. Then $g(\mathbb{E}\,X) \leq \mathbb{E}\,g(X)$.*

*Proof.* For each $x_0 \in \mathbb{R}$ there is a $a \in \mathbb{R}$ such that $g(x) \geq g(x_0) + a(x - x_0)$ [e.g. 8, Theorem 3.24]. Let $x_0 = \mathbb{E}\,X$. Then $g(x) \geq g(\mathbb{E}\,X) + a(x - \mathbb{E}\,X)$ for some $a \in \mathbb{R}$. Take expectations to obtain $\mathbb{E}\,g(X) \geq g(\mathbb{E}\,X) + a(\mathbb{E}\,X - \mathbb{E}\,X) = g(\mathbb{E}\,X)$. □

COROLLARY 2.4: *If $X \in L_p$ then for any $1 \leq q \leq p$, $\|X\|_{L_q} \leq \|X\|_{L_p}$.*

*Proof.* $x \mapsto x^{p/q}$ is convex and so, by Jensen's inequality, $\mathbb{E}\,|X|^p \geq (\mathbb{E}\,|X|^q)^{p/q}$. □

The third result is the Borel - Cantelli lemma. For this we need to introduce the concept of the limit supremum of a sequence of sets. If $(E_n)_{n \in \mathbb{N}}$ is a sequence of events, then

$$\limsup_{n \in \mathbb{N}} E_n := \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} E_m.$$

LEMMA 2.6 [Borel - Cantelli]: *Let $(E_n)_{n \in \mathbb{N}}$ be a sequence of events. If $\sum_{n=1}^{\infty} P(E_n) < \infty$ then $P(\limsup_{n \to \infty} E_n) = 0$.*

*Proof.* Let $F_n := \bigcup_{m=n}^{\infty} E_m$. The sequence $(F_n)_{n\in\mathbb{N}}$ is a sequence of events which is non-increasing: $F_n \supset F_{n+1}$. By continuity from above $\lim_{n\to\infty} P(F_n) = P(\limsup_{n\to\infty} E_n)$ (cf. (v) in Proposition 2.1). As probabilities are sub-additive (by (iv) in Proposition 2.2), $P(F_n) \leq \sum_{m=n}^{\infty} P(E_m) \to 0$, since $\sum_{n=1}^{\infty} P(E_n) < \infty$. □

Finally we will discuss a way of describing the distribution of a random vector: via its *characteristic function*. A random vector, $X$, valued in in $\mathbb{R}^K$ has characteristic function $\psi$:

$$\psi(t) := \mathbb{E}\left[\exp(it'X)\right] = \mathbb{E}\left[\cos t'X\right] + i\,\mathbb{E}\left[\sin t'X\right], \qquad t \in \mathbb{R}^K,$$

where $i$ is the imaginary unit, $i := \sqrt{-1} \in \mathbb{C}$. This always exists (exercise).

Proposition 2.13: *For any random vector $X$ its characteristic function $\psi$ has the following properties:*

*(i)* $\psi(0) = 1$;

*(ii)* $\psi(-t) = \overline{\psi(t)}$;

*(iii)* $|\psi(t)| \leq \mathbb{E}\,|\exp(it'X)| = 1$;

*(iv)* $\psi(t+h) - \psi(t)| \leq \mathbb{E}\,|\exp(ih'X) - 1|$;

*(v)* $\mathbb{E}\exp(it'[AX+b]) = \exp(it'b)\psi(A't)$.

*Proof.* Exercise. □

## 2.5 Stochastic convergence

Suppose that $(x_n)_{n\in\mathbb{N}}$ is a sequence of vectors in $\mathbb{R}^n$. This sequence *converges* to a vector $x \in \mathbb{R}^n$, written $x_n \to x$ or $\lim_{n\to\infty} x_n = x$, if for any $\varepsilon > 0$ there is a $N$ (which may depend on $\varepsilon$ and $x$) such that $\|x_n - x\| < \varepsilon$ whenever $n \geq N$.

Lets suppose we have a sequence $(X_n)_{n\in\mathbb{N}}$ of *random* vectors. We will discuss various forms of convergence appropriate for random vectors.

$(X_n)_{n\in\mathbb{N}}$ converges *almost surely* to $X$, written $X_n \xrightarrow{as} X$ if

$$P\left(\left\{\lim_{n\to\infty} X_n = X\right\}\right) = 1.$$

$(X_n)_{n\in\mathbb{N}}$ converges *in probability* to $X$, written $X_n \xrightarrow{P} X$ or $\operatorname{plim}_{n\to\infty} X_n = X$ if for all $\varepsilon > 0$,

$$\lim_{n\to\infty} P(\|X_n - X\| > \varepsilon) = 0.$$

$(X_n)_{n\in\mathbb{N}}$ converges *weakly* or *in distribution* to $X$, written $X_n \rightsquigarrow X$ (often also written as $X_n \xrightarrow{D} X$) if

$$\lim_{n\to\infty} \mathbb{E}\,f(X_n) = \mathbb{E}\,f(X) \qquad \text{for all bounded, continuous functions } f : \mathcal{X} \to \mathbb{R}.$$

Then, if $X \in L_p$, $(X_n)_{n\in\mathbb{N}}$ converges to $X$ in $L_p$, written $X_n \xrightarrow{L_p} X$ if

$$\lim_{n\to\infty} \|X_n - X\|_{L_p} = 0.$$

Convergence almost surely can be given an alternative definition.

Lemma 2.7: *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random vectors and $X$ a random vector. For $\varepsilon > 0$, let $E_{n,\varepsilon} := \{\|X_n - X\| > \varepsilon\}$. Then $X_n \xrightarrow{as} X$ if and only if for each $\varepsilon > 0$,*

$$P(\limsup_{n\to\infty} E_{n,\varepsilon}) = 0.$$

*Proof.* Exercise. □

Weak convergence has many equivalent forms. A set $A$ is a $X$-continuity set if $P(X \in \delta A) = 0$ where $\delta A$ is the boundary of $A$, i.e. its closure less its interior: $\delta A := \operatorname{cl} A \setminus \operatorname{int} A$. For a set $F$, let $\rho(x, F) := \inf_{y\in F} \|x - y\|$.

*Theorem 2.8: *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random vectors and $X$ a random vector. The following are equivalent:*

*(i)* $X_n \rightsquigarrow X$,

*(ii)* $\lim_{n\to\infty} \mathbb{E} f(X_n) = \mathbb{E} f(X)$ *for all bounded, Lipschitz-continuous* $f : \mathcal{X} \to \mathbb{R}$,

*(iii)* $\limsup_{n\to\infty} P(X_n \in F) \le P(X \in F)$ *for all closed sets* $F$,

*(iv)* $\liminf_{n\to\infty} P(X_n \in G) \ge P(X \in G)$ *for all open sets* $G$,

*(v)* $\lim_{n\to\infty} P(X_n \in A) = P(X \in A)$ *for all $X$-continuity sets* $A$,

*(vi) if $F_n$ is the CDF of $X_n$ and $F$ that of $X$, $F_n(x) \to F(x)$ for all $x$ at which $F$ is continuous.*

**Proof: (i) $\implies$ (ii):* obvious since every Lipschitz-continuous function is continuous.

*(ii) $\implies$ (iii):* let $F$ be a closed set, $F^\varepsilon := \{x : \rho(x, F) < \varepsilon\}$ and define $f(x) := \max\{1 - \rho(x, F)/\varepsilon, 0\}$. We note that $\mathbf{1}_F(x) \le f(x) \le \mathbf{1}_{F^\varepsilon}(x)$ and $f$ is (Lipschitz) continuous [Exercise]. Then by (i) or (ii)

$$P(X_n \in F) = \mathbb{E}\,\mathbf{1}_F(X_n) \le \mathbb{E} f(X_n) \to \mathbb{E} f(X) \le \mathbb{E}\,\mathbf{1}_{F^\varepsilon}(X) = P(X \in F^\varepsilon).$$

As $F$ is closed, taking the limit as $\varepsilon \downarrow 0$ yields the required inequality [Exercise].

*(iii) $\implies$ (iv):* Take complements [Exercise].

*(iii) & (iv) $\implies$ (v):* Combining (iii) and (iv):

$$\begin{aligned} P(X \in \operatorname{cl} A) &\ge \limsup_{n\to\infty} P(X_n \in \operatorname{cl} A) \ge \limsup_{n\to\infty} P(X_n \in A) \\ &\ge \liminf P(X_n \in A) \ge \liminf_{n\to\infty} P(X_n \in \operatorname{int} A) \ge P(X \in \operatorname{int} A). \end{aligned}$$

If $P(X \in \delta A) = 0$, i.e. $A$ is an $X$-continuity set the left and right hand side terms coincide (both are equal to $P(X \in A)$), which implies (v).

*(v)* $\implies$ *(vi):* Let $x$ be a continuity point of $F$ and set $A = (-\infty, x]$. Then $A$ is a $X$-continuity set as $\delta A = \{x\}$ and $P(X = x) = 0$ [Exercise]. Then by (v) $F_n(x) = P(X_n \in A) \to P(X \in A) = F(x)$.

*(v)* $\implies$ *(i):* It is enough to consider the case that $0 \le f \le 1$ [Exercise]. Then

$$\mathbb{E}\, f(X) = \int_0^1 P(f(X) > t)\,\mathrm{d}t, \qquad \mathbb{E}\, f(X_n) = \int_0^1 P(f(X_n) > t)\,\mathrm{d}t,$$

by e.g. Lemma 2.2.13 in [5]. Since $f$ is continuous, $\delta\{x : f(x) > t\} \subset \{x : f(x) = t\}$ and so $\{x : f(x) > t\}$ is a $X$-continuity set except for at countably many $t$ [Exercise]. By (v) and the bounded convergence theorem

$$\mathbb{E}\, f(X_n) = \int_0^1 P(f(X_n) > t)\,\mathrm{d}t \to \int_0^1 P(f(X) > t)\,\mathrm{d}t = \mathbb{E}\, f(X).$$

*(vi)* $\implies$ *(iv):* Define $D^i := \{c : P(X \in H_c^i) > 0\}$ for $H_c^i := \{x : x_i = c\}$. Note that $D^i$ is at most countable. Let $A$ be a rectangle $A = \prod_{i=1}^K (a_i, b_i]$ with $a_i, b_i \notin D^i$ for each $i$. By $F_n(x) \to F(x)$, for any such $A$, $P(X_n \in A) \to P(X \in A)$ and therefore for any finite collection of disjoint such rectangles $A_1, \dots, A_k$, their union $B_k$ also satisfies $P(X_n \in B_k) \to P(X \in B_k)$. As any open set $G \in \mathbb{R}^K$ can be written as the increasing limit of such a disjoint union of rectangles with $a_i, b_i \notin D^i$,

$$\liminf_{n\to\infty} P(X_n \in G) \ge \liminf_{n\to\infty} P(X_n \in B_k) = P(X \in B_k).$$

Taking $B_k \uparrow G$ completes the proof. □

Theorem 2.9: *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random vectors, $X$ a random vector and $c$ a constant vector. We have the following relationships between the methods of convergence:*

*(i)* $X_n \xrightarrow{as} X \implies X_n \xrightarrow{P} X$,

*(ii)* *$X_n \xrightarrow{P} X$ *if and only if every subsequence of $(X_n)$ has a further subsequence which converges to $X$ almost surely,*

*(iii)* *For $X \in L_p$,* $X_n \xrightarrow{L_p} X \implies X_n \xrightarrow{P} X$,

*(iv)* $X_n \xrightarrow{P} X \implies X_n \rightsquigarrow X$,

*(v)* $X_n \rightsquigarrow c \implies X_n \xrightarrow{P} c$.

*Proof.* Parts (i) and (iii) - (v) are left as exercises.

*For (ii), let $(X_{n_m})_{m\in\mathbb{N}}$ be an arbitrary subsequence. Since $X_n \xrightarrow{P} X$ there are $k$ such that $P(\|X_{n_{m_k}} - X\| > 2^{-k}) < 2^{-k}$. Letting $E_k := \{\|X_{n_{m_k}} - X\| > 2^{-k}\}$ one has that for all large enough $k$, $2^{-k} < \varepsilon$ and hence $E_k \supset \{\|X_{n_{m_k}} - X\| > \varepsilon\} := F_k$. This implies that $\sum_{k=1}^\infty P(F_k) < \infty$ and so by Borel-Cantelli (Lemma 2.6) $P(\limsup_{k\to\infty} F_k) = 0$. Apply Lemma

2.7. For the converse, let $x_n := P(\|X_n - X\| > \varepsilon)$. Each subsequence of $x_n$ has a further subsequence which converges to 0 by part (i) above. This implies that $x_n \to 0$.[41] □

### 2.5.1 Stochastic order symbols

If $(X_n)_{n\in\mathbb{N}}$ a sequence of random vectors and $(a_n)_{n\in\mathbb{N}}$ a sequence of non-negative numbers, we write $X_n = O_P(a_n)$ if for each $\varepsilon$ there is a finite constant $M$ (which may depend on $\varepsilon$) such that $P(\|X_n\| > Ma_n) < \varepsilon$ for all $n$ larger than some $N$ (which may also depend on $\varepsilon$). If $X_n = O_P(1)$ we call the $X_n$ *stochastically bounded.* If $a_n > 0$, $X_n = O_P(a_n)$ if and only if $X_n/a_n = O_P(1)$ [Exercise].

We write $X_n = o_P(a_n)$ if for each $\varepsilon$, $\lim_{n\to\infty} P(\|X_n\| > \varepsilon a_n) = 0$. If $a_n > 0$, $X_n = o_P(a_n)$ if and only if $X_n/a_n \xrightarrow{P} 0$ [Exercise].

These *stochastic order symbols* often make computations much easier, but some care needs to be taken with their use. They should be read always from left to right and their rules are not those of normal algebraic calculations. For example, if $X_n = O_P(1)$ and $Y_n = O_P(1)$ then $X_n + Y_n = O_P(1)$ (as noted below), hence $O_P(1) + O_P(1) = O_P(1)$ and we *cannot* cancel to conclude that $O_P(1) = 0$.

Lemma 2.8: *Let $(X_n)_{n\in\mathbb{N}}$, $(Y_n)_{n\in\mathbb{N}}$ be sequences of random vectors and $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ sequences of non-negative numbers. Then,*

*(i) If $X_n = o_P(a_n)$, then $X_n = O_P(a_n)$,*

*(ii) If $X_n = O_P(a_n)$, $Y_n = o_P(b_n)$ then for any $k \in \mathbb{R}$, $kX_n = O_P(a_n)$ and $kY_n = o_P(b_n)$,*

*(iii) If $X_n = O_P(a_n)$, $Y_n = O_P(b_n)$, then $X_n + Y_n = O_P(a_n + b_n)$ and $X_nY_n = O_P(a_nb_n)$,*

*(iv) If $X_n = o_P(a_n)$, $Y_n = o_P(b_n)$, then $X_n + Y_n = o_P(a_n + b_n)$ and $X_nY_n = o_P(a_nb_n)$,*

*(v) If $X_n = O_P(a_n)$, $Y_n = o_P(b_n)$, then $X_n + Y_n = O_P(a_n + b_n)$ and $X_nY_n = o_P(a_nb_n)$,*

*(vi) If $X_n = O_P(a_n)$ and $a_n \to 0$ then $X_n = o_P(1)$.*

*Proof.* Exercise. □

The stochastic order symbols are based on the deterministic order symbols $O$ and $o$. These are defined similarly. If $(x_n)_{n\in\mathbb{N}}$ is a sequence of (non-random) vectors then $x_n = O(a_n)$ if there are $N$ and $M$ such that $n \geq N$ implies $\|x_n\| \leq Ma_n$. $x_n = o(a_n)$ if for each $\varepsilon > 0$ there is a $N$ such that $\|x_n\| \leq \varepsilon a_n$ whenever $n \geq N$.

### 2.5.2 Important tools

The following two Theorems are very important.

Theorem 2.10: *If $g : \mathcal{X} \to \mathbb{R}^d$ is continuous except possibly on a set $D$ with $P(X \in D) = 0$,*

*(i) $X_n \xrightarrow{as} X \implies g(X_n) \xrightarrow{as} g(X)$;*

[41] See Exercise 1.5.3.

*(ii)* $X_n \xrightarrow{P} X \implies g(X_n) \xrightarrow{P} g(X)$*;*

*(iii)* $X_n \rightsquigarrow X \implies g(X_n) \rightsquigarrow g(X)$*.*

*Proof.*

(i) Let $E$ be the set on which $X_n \not\to X$. Then $P((D \cup E)^{\complement}) = 1$ and on this set $g(X_n) \to g(X)$.

(ii) Combine part (i) here with part (ii) of Theorem 2.9.

(iii) Let $f$ be any bounded, continuous function from $\mathbb{R}^d \to \mathbb{R}$. Then $h := f \circ g$ is a bounded continuous function from $\mathbb{R}^K \to \mathbb{R}$, hence $\mathbb{E}\, f(g(X_n)) = \mathbb{E}\, h(X_n) \to \mathbb{E}\, h(X) = \mathbb{E}\, f(g(X))$.

□

Lemma 2.9: *If* $X_n \rightsquigarrow X$ *and* $\|X_n - Y_n\| \xrightarrow{P} 0$ *then* $Y_n \rightsquigarrow X$*.*

**Proof:* Let $F$ be a closed set and define $F_\varepsilon := \{x : \inf_{y \in F} \|x - y\| \le \varepsilon\}$. This set is closed and

$$P(Y_n \in F) \le P(\|X_n - Y_n\| \ge \varepsilon) + P(X_n \in F_\varepsilon).$$

As $F_\varepsilon$ is closed, by (iii) of Theorem 2.8

$$\limsup_{n \to \infty} P(Y_n \in F) \le \limsup_{n \to \infty} P(X_n \in F_\varepsilon) \le P(X \in F_\varepsilon).$$

$F_\varepsilon \downarrow F$ as $\varepsilon \downarrow 0$ if $F$ is closed, hence the result follows from (iii) of Theorem 2.8 and (v) of Proposition 2.1. □

Theorem 2.11 [Slutsky]: *If* $X_n \rightsquigarrow X$ *and* $Y_n \xrightarrow{P} c$*, a constant, then*

*(i)* $X_n + Y_n \rightsquigarrow X + c$*,*

*(ii)* $X_n Y_n \rightsquigarrow Xc$*,*

*(iii)* $Y_n^{-1} X_n \rightsquigarrow c^{-1} X_n$*, provided that* $c^{-1}$ *exists.*

*Proof.* Note that $\|(X_n, Y_n) - (X_n, c)\| \xrightarrow{P} 0$, hence by Lemma 2.9 this will follow from Theorem 2.10 provided we show that $(X_n, c) \rightsquigarrow (X, c)$. Let $(x, y) \mapsto f(x, y)$ be a continuous bounded function and let $h(x) := f(x, c)$. This is obviously also a continuous bounded function and hence by $X_n \rightsquigarrow X$, $\mathbb{E}\, f(X_n, c) = \mathbb{E}\, h(X_n) \to \mathbb{E}\, h(X) = \mathbb{E}\, f(X, c)$. □

A further result, called the *delta method* is useful for determining the limiting distribution of smooth functions of sequences of random vectors which have a limit when appropriately scaled.

Theorem 2.12 [Delta method]: *Suppose that* $g : \mathbb{R}^K \to \mathbb{R}^L$ *is differentiable at* $x$ *with Jacobian* $G$*. If* $Z_n := \sqrt{n}\,(X_n - x) \rightsquigarrow Z$*, then* $\sqrt{n}\,(g(X_n) - g(x)) \rightsquigarrow GZ$*.*

**Proof:* By Taylor's theorem

$$g(x + \delta) = g(x) + G\delta + R(\delta),$$

where the remainder $R$ satistifies $|R(\delta_n)| = o(|\delta_n|)$ for any $\delta_n \to 0$. Then replacing $\delta$ with $Z_n/\sqrt{n} = O_P(n^{-1/2})$ [Exercise],

$$\sqrt{n}\,(g(X_n) - g(x)) = GZ_n + \sqrt{n}R(Z_n/\sqrt{n}).$$

$GZ_n \rightsquigarrow GZ$ by Slutsky's theorem, hence it remains to show that $\sqrt{n}R(Z_n/\sqrt{n}) = o_P(1)$ from which the result would follow by another application of Slutsky's theorem. For this note that $|\sqrt{n}R(Z_n/\sqrt{n})| = \sqrt{n}o(\|Z_n\|/\sqrt{n}) = o(\|Z\|_n) = o_P(1)$ [Exercise]. □

### 2.5.3 Convergence of moments

It is useful to establish conditions under which $\mathbb{E}\,X_n \to \mathbb{E}\,X$ if $X_n$ converges to $X$. The first results in this direction are immediate from the definition of weak convergence and convergence in $L_p$.

Lemma 2.10: *If $(X_n)_{n\in\mathbb{N}}$ converges to $X$ weakly and $f : \mathcal{X} \to \mathbb{R}$ is a bounded, continuous function, $\mathbb{E}\,f(X_n) \to \mathbb{E}\,f(X)$.*

*Proof.* Exercise. □

Lemma 2.11: *If $(X_n)_{n\in\mathbb{N}}$ is a sequence of random vectors which converges to $X$ in $L_p$ then for any $1 \le q \le p$, $\mathbb{E}\,\|X_n\|^q \to \mathbb{E}\,\|X\|^q$.*

*Proof.* Exercise. □

In general, other forms of convergence do not ensure this. Consider the following example.

Example 2.13: Let $X_n$ be a random variable with $P(X_n = n) = 1/n$ and $P(X_n = 0) = 1 - 1/n$.Then $X_n \xrightarrow{P} 0$ (hence also weakly) [Exercise] but $\mathbb{E}\,X_n = n \times 1/n = 1$ for each $n$. △

The best we can say in general is the following.

*Theorem 2.13: *If $X_n \rightsquigarrow X$ then $\mathbb{E}\,\|X\| \le \liminf_{n\to\infty} \mathbb{E}\,\|X_n\|$.*

**Proof:* By Theorem 2.10, $\|X_n\| \rightsquigarrow \|X\|$. Since a CDF can have at most countably many discontinuities [Exercise], one has $P(\|X_n\| > t) \to P(\|X\| > t)$ for all but a countable number of $t$. Then by Fatou's lemma [e.g. 5, Theorem 1.5.5], one has[42]

$$\mathbb{E}\,\|X\| = \int_0^\infty P(\|X\| > t)\,\mathrm{d}t \le \liminf_{n\to\infty} \int_0^\infty P(\|X_n\| > t)\,\mathrm{d}t = \liminf_{n\to\infty} \mathbb{E}\,\|X_n\|. \qquad \square$$

What is required is *uniform integrability.* $(X_n)_{n\in\mathbb{N}}$ is uniformly integrable if

$$\lim_{M\to\infty} \sup_{n\in\mathbb{N}} \mathbb{E}\,[\|X_n\|\mathbf{1}\{\|X_n\| > M\}] = 0.$$

[42] For the first equality see e.g. Lemma 2.2.13 in [5].

THEOREM 2.14: *Suppose that $X_n \rightsquigarrow X$ and $(X_n)_{n\in\mathbb{N}}$ is uniformly integrable. Then $X$ is integrable and $\mathbb{E}\, X_n \to \mathbb{E}\, X$.*

**Proof:* That $X$ is integrable follows from Theorem 2.13 since $\sup_{n\in\mathbb{N}} \mathbb{E}\,\|X_n\| < \infty$ [Exercise]. We give the proof for the case where $X_n$ is a non-negative random variable, leaving the extension to the general case as an exercise. We have for $M \in (0,\infty)$,

$$|\,\mathbb{E}\, X_n - \mathbb{E}\, X| \leq |\,\mathbb{E}\, X_n - \mathbb{E}[X_n \wedge M]| + |\,\mathbb{E}[X_n \wedge M] - E[X \wedge M]| + |\,\mathbb{E}[X \wedge M] - \mathbb{E}\, X|,$$

where $a \wedge b := \min\{a, b\}$. The function $x \mapsto x \wedge M$ is a bounded continuous function, so the middle right hand side term converges to zero by $X_n \rightsquigarrow X$. Since $X_n$ is non-negative, the first right hand side term is upper bounded by $\sup_{n\in\mathbb{N}} \mathbb{E}\,[\|X_n\| \mathbf{1}\{\|X_n\| > M\}]$ which can be made arbitrarily small by taking $M$ large enough given the uniform integrability. The third term can also be made arbitrarily small by increasing $M$ by either the monotone convergence theorem or the dominated convergence theorem.[43] □

LEMMA 2.12: *Suppose there is a uniformly integrable sequence $(Y_n)_{n\in\mathbb{N}}$ such that $\|X_n\| \leq \|Y_n\|$ a.s. for each $n \in \mathbb{N}$. Then $(X_n)_{n\in\mathbb{N}}$ is uniformly integrable.*

*Proof.* Exercise. □

LEMMA 2.13: *Suppose that $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$ are uniformly integrable. Then $(X_n + Y_n)_{n\in\mathbb{N}}$ is uniformly integrable.*

*Proof.* Exercise. □

LEMMA 2.14: *Let $1 < p, q < \infty$ with $1/p + 1/q = 1$ and suppose that $(\|X_n\|^p)_{n\in\mathbb{N}}$ and $(\|Y_n\|^q)_{n\in\mathbb{N}}$ are uniformly integrable. Then $(X_n Y_n)_{n\in\mathbb{N}}$ is uniformly integrable.*

*Proof.* Exercise. □

LEMMA 2.15: *Suppose that for some $p > 1$, $\sup_{n\in\mathbb{N}} \mathbb{E}\,\|X_n\|^p < \infty$. Then $(X_n)_{n\in\mathbb{N}}$ is uniformly integrable.*

*Proof.* For any $n \in \mathbb{N}$, $M^{p-1}\,\mathbb{E}[\|X_n\| \mathbf{1}\{\|X_n\| > M\}\}] \leq \mathbb{E}[\|X_n\|^p \mathbf{1}\{\|X_n\| > M\}\}] \leq \mathbb{E}[\|X_n\|^p]$. Hence for $n \geq N$, as $M \to \infty$

$$\mathbb{E}[\|X_n\| \mathbf{1}\{\|X_n\| > M\}\}] \leq M^{1-p}\,\mathbb{E}[\|X_n\|^p] \leq M^{1-p} \sup_{n\geq N} \mathbb{E}[\|X_n\|^p] \to 0. \qquad \square$$

### 2.5.4 Joint vs. marginal convergence

Convergence almost surely, in probability and in $L_p$ are pointwise concepts; weak convergence is not.

[43] See e.g. Theorem 1.5.7 / 1.5.8 in [5].

LEMMA 2.16: *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random vectors and $X$ a random vector. Then for $m \in \{as, P, L_p\}$, $X_n \xrightarrow{m} X$ if and only if $X_{n,k} \xrightarrow{m} X_k$.*

*Proof.* Exercise. □

PROPOSITION 2.14: *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random vectors and $X$ a random vector. Then $X_n \rightsquigarrow X$ implies that $X_{n,k} \rightsquigarrow X_k$ but $X_{n,k} \rightsquigarrow X_k$ does not imply that $X_n \rightsquigarrow X$.*

**Proof:* Let $f$ be a bounded continuous function from $\mathbb{R}$ to $\mathbb{R}$. Then $g := f \circ \pi_k$ is a bounded continuous function from $\mathbb{R}^K \to \mathbb{R}$ where $\pi_k(x) := x_k$. Then $\mathbb{E}\, f(X_{n,k}) = \mathbb{E}\, g(X_n) \to Eg(X) = \mathbb{E}\, f(X_k)$, hence $X_{n,k} \rightsquigarrow X_k$.

For the converse, let $Z \sim \mathcal{N}(0,1)$ and set $X_n = (Z, (-1)^n Z)$. Both $X_{n,1}$ and $X_{n,2}$ are standard normal for each $n$, and hence weakly converge to $Z$ [Why?]. $X_n$ does not weakly converge to a limit. It if did, then we would have $\mathbb{E}[X_{n,1}X_{n,2}] \to \mathbb{E}[X_1 X_2]$ where $X$ is the hypothesised weak limit [Why?]. But $\mathbb{E}[X_{n,1}X_{n,2}] = (-1)^n\, \mathbb{E}[Z^2] = (-1)^n$ which does not converge. □

### 2.5.5 The law of large numbers

Laws of large numbers are results which provide conditions under which averages of random variables $\frac{1}{n}\sum_{i=1}^n X_i$ converge in probability or almost surely to their expectation.[44]

Versions with almost sure convergence are usually called *strong laws of large numbers* or *SLLN*s whilst version with convergence in probability are called *weak laws of large numbers* or *WLLN*s. For statistical applications, WLLNs are generally sufficient so we will consider these.

Here we will consider only the case when the elements $X_1, X_2, \ldots$ of the sequence are independent. A collection of random variables is a *random sample* or *independently and independently distributed* (i.i.d.) if each random variable is independent of the others and all share the same distribution.

THEOREM 2.15 [Weak law of large numbers, i.i.d.]: *If $(X_n)_{n\in\mathbb{N}}$ is an i.i.d. sequence of random vectors with $\mathbb{E}\,\|X_1\| < \infty$,[45] then for $\mu := \mathbb{E}\, X_1$, as $n \to \infty$,*

$$\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{P} \mu.$$

**Proof:* Theorem 2.2.14 in [5] proves this for the case of random variables. The case for random vectors follows from this given Lemma 2.16. □

One often also comes across sequences $(X_n)_{n\in\mathbb{N}}$ which consist of random variables/vectors which are independent but their distribution may change with the index $n$.

THEOREM 2.16 [Markov's weak law of large numbers]: *Suppose that $(X_n)_{n\in\mathbb{N}}$ is a sequence of independent random vectors such that for some $\delta > 0$, $\mathbb{E}\,\|X_n\|^{1+\delta} < C < \infty$. Then for*

[44] Or an average of their expectations.

[45] Note that $\mathbb{E}\,\|X_1\| = \mathbb{E}\,\|X_n\|$ for all $n \in \mathbb{N}$ as the $X_n$ all have the same distribution.

$\mu_n := \mathbb{E}\, X_n$, *as* $n \to \infty$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i) \xrightarrow{P} 0.$$

**Proof:* Given Lemma 2.16, the case for random vectors follows from that for random variables, so we may assume the $X_n$ are random variables.

We verify the conditions in Theorem 2.2.11 of [5]. Let $X_{n,k} = X_k$ for each $n \in \mathbb{N}$ and $b_n = n$. It then suffices to show that (i) $\sum_{k=1}^n P(|X_k| > n)$ and (ii) $\frac{1}{n^2}\sum_{k=1}^n \mathbb{E}\left[X_k^2 \mathbf{1}\{|X_k| \le n\}\right]$ both converge to zero. For (i), by Markov's inequality, $P(|X_k| > n) \le C/n^{1+\delta}$, hence $\sum_{k=1}^n P(|X_k| > n) \le C/n^\delta \to 0$. For (ii) it suffices to show that $n^{-1}\,\mathbb{E}\left[X_k^2\mathbf{1}\{|X_k| \le n\}\right] \le a_n \to 0$. For this, note that by Lemma 2.2.13 in [5]

$$\mathbb{E}\left[X_k^2 \mathbf{1}\{|X_k| \le n\}\right] = 2\int_0^\infty yP(|X_k|\mathbf{1}\{|X_k| \le n\} > y)\,\mathrm{d}y \le 2\int_0^n yP(|X_k| > y)\,\mathrm{d}y,$$

as $P(|X_k|\mathbf{1}\{|X_k| \le n\} > y) = P(|X_k| > y) - P(|X_k| > n)$ if $y \le n$ and 0 otherwise. By Corollary 2.4 we may assume that $\delta \in (0,1)$. Then, using Markov's inequality, one has

$$n^{-1}\,\mathbb{E}\left[X_k^2\mathbf{1}\{|X_k| \le n\}\right] \le \frac{1}{n}\int_0^n yP(|X_k| > y)\,\mathrm{d}y \le \frac{C}{n}\int_0^n y^{-\delta}\,\mathrm{d}y = \frac{Cn^{1-\delta}}{1-\delta}\frac{1}{n} \to 0. \qquad \square$$

COROLLARY 2.5: *In the setting of Theorem 2.16, if* $\frac{1}{n}\sum_{i=1}^n \mu_i \to \mu$ *as as* $n \to \infty$, *then*

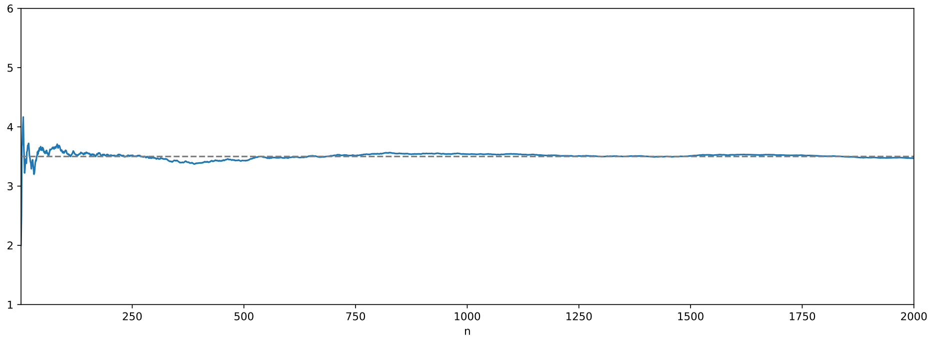
$$\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{P} \mu.$$

*Proof.* Exercise. □

Example 2.14: Consider a fair die and let $X_n \in \{1, \ldots, 6\}$ be the value obtained on the $n$-th roll of the die. Since the die is fair the $X_n$ are i.i.d. with mean $\mathbb{E}\, X_n = 3.5$ and $\mathrm{Var}(X_n) = 35/12$ by Example 2.9. By the weak law of large numbers (e.g. Theorem 2.15) we have $\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{P} 3.5$.

Here we plot this average as $n$ increases for a realisation of $(X_1, X_2, \ldots)$.

FIGURE 13: $\frac{1}{n}\sum_{i=1}^n X_i$



△

### 2.5.6 The central limit theorem

A central limit theorem (CLT) is a result which gives conditions under which centered and scaled averages of random variables converge in distribution to a (zero-mean) Normal distribution.

We start with a result for i.i.d. sequences of random variables.

Theorem 2.17 [Lindeberg – Lévy central limit theorem]: *If $(X_n)_{n\in\mathbb{N}}$ is an i.i.d. sequence of random variables with* $\operatorname{Var}(X_1) = \sigma^2 < \infty$*, then for $\mu := \mathbb{E}\, X_1$, as $n \to \infty$*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu) \rightsquigarrow \mathcal{N}(0, \sigma^2).$$

**Proof:* Theorem 3.4.1 in [5] & Slutsky's Theorem. □

To obtain a CLT for random vectors, we can use the following Theorem.

Theorem 2.18 [Cramér – Wold]: *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of $K$ dimensional random vectors. $X_n \rightsquigarrow X$ if and only if $t'X_n \rightsquigarrow t'X$ for all $t \in \mathbb{R}^K$.*

**Proof:* Theorem 3.10.6 in [5]. □

Corollary 2.6 [Multivariate central limit theorem, i.i.d.]: *If $(X_n)_{n\in\mathbb{N}}$ is an i.i.d. sequence of random vectors with* $\operatorname{Var}(X_1) = \Sigma$ *($\|\Sigma\| < \infty$), then for $\mu := \mathbb{E}\, X_1$, as $n \to \infty$*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu) \rightsquigarrow \mathcal{N}(0, \Sigma).$$

*Proof.* Exercise. □

*Theorem 2.19 [Lindeberg's central limit theorem]: *Suppose that $(X_n)_{n\in\mathbb{N}}$ is a sequence of independent random variables and let $\mu_n := \mathbb{E}\, X_n$, $\sigma_n^2 :=$* $\operatorname{Var}(X_n) < \infty$*. Let $Z_n := X_n - \mu_n$. If $\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2 \to \sigma^2 \neq 0$ and for every $\varepsilon > 0$*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[Z_i^2 \mathbf{1}\left\{|Z_i| > \varepsilon\sqrt{n}\right\}\right] = 0, \tag{9}$$

*then as $n \to \infty$,*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i \rightsquigarrow \mathcal{N}(0, \sigma^2).$$

**Proof:* See e.g. Theorem 27.2 in [1]. □

This CLT can be made to apply to random vectors by the Cramér – Wold Theorem.

Corollary 2.7 [Multivariate Lindeberg's central limit theorem]: *Suppose that $(X_n)_{n\in\mathbb{N}}$ is a sequence of independent random vectors and let $\mu_n := \mathbb{E}\, X_n$, $\Sigma_n :=$* $\operatorname{Var}(X_n) < \infty$*. Let*

$Z_n := X_n - \mu_n$. *If* $\frac{1}{n}\sum_{i=1}^n \Sigma_i \to \Sigma$, *where* $\Sigma$ *is positive definite and for every* $\varepsilon > 0$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|Z_i\|^2 \mathbf{1}\left\{\|Z_i\| > \varepsilon\sqrt{n}\right\}\right] = 0, \tag{10}$$

*then as* $n \to \infty$,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i \rightsquigarrow \mathcal{N}(0,\Sigma).$$

*Proof.* Exercise. □

Condition (9) (or (10)) is called "Lindeberg's condition". A sufficient condition for this is the following "Lyapunov condition"; whilst this is stronger, it is also typically easier to verify in applications. We present the multivariate version which contains the univariate case.

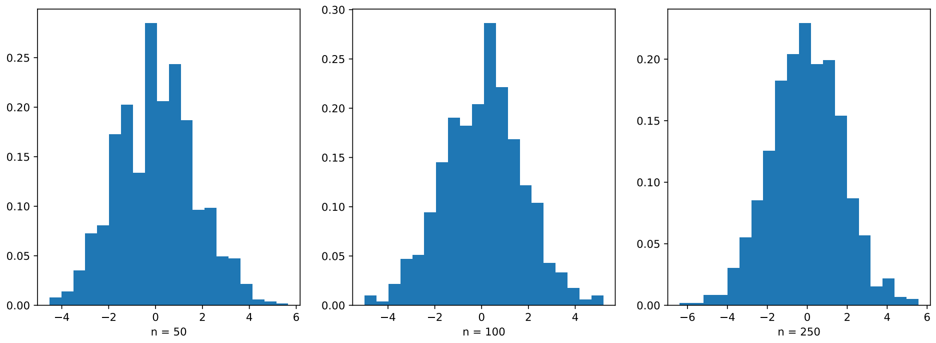
Lemma 2.17 [Lyapunov condition]: *A sufficient condition for* (10) *is that*

$$\lim_{n\to\infty} \frac{1}{n^{1+\delta/2}}\sum_{i=1}^n \mathbb{E}\,\|Z_i\|^{2+\delta} = 0 \qquad \textit{for some } \delta > 0.$$

*Proof.* Exercise. □

Example 2.15: Consider the fair die example. Let $Z_n := X_n - 3.5$ and plot histograms of $\frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i$ based on 1000 repetitions of samples of varying sizes.

Figure 14: $\frac{1}{\sqrt{n}}\sum_{i=1}^n Z_n$



△

# 3 Statistical inference

## 3.1 Statistical models

The starting point for statistical theory is the concept of a *statistical model* (or "model"). Suppose we observe $X = (X_1, \dots, X_n) \in \mathcal{X}$. $X$ is our *sample*. We will consider $X$ is a random vector defined on a sample space $\Omega$, which is equipped with an associated $\sigma$-algebra $\mathcal{F}$. A statistical model is a model for the distribution of $X$. That is, a collection, $\mathcal{P}$, of probability measures $P$, defined on $(\Omega, \mathcal{F})$. Such a collection may also be referred to as a *family.*

Typically we define / describe our statistical model through a *parametrisation.* A parametrisation is a map, $\theta \mapsto P_\theta$, from the *parameter space* $\Theta$, to our statistical model, $\mathcal{P}$:

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}.$$

The whole idea of statistical inference is to use the sample $X$ to learn about the unknown *parameter* $\theta$ (or some function of it).

Example 3.1: Suppose we observe $n$ samples which satisfy

$$X_i = \theta + \epsilon_i \ \ (i = 1, \dots, n), \quad \epsilon_1, \dots, \epsilon_n \text{ are i.i.d. draws from } \mathcal{N}(0,1), \quad \theta \in \Theta \subset \mathbb{R}. \tag{11}$$

Then, our statistical model is

$$\mathcal{P} = \{\mathcal{N}(\theta, 1)^n : \theta \in \Theta\},$$

where $\mathcal{N}(\theta, 1)^n$ denotes the distribution of $n$ i.i.d draws from a $\mathcal{N}(\theta, 1)$ distribution. △

Often we do not explicitly write out the statistical model we are considering in each given statistical problem, but rather specify it implicitly, such as through equation (11).

### 3.1.1 Random sampling

Example 3.1 assumes that the error terms, $\epsilon_i$ are i.i.d. ("independent and identically distributed"). That is to say, each pair $\epsilon_i$ and $\epsilon_j$ are independent random variables and each $\epsilon_i$ has the same distribution (in the example, a standard normal). This i.i.d. assumption is also sometimes called "random sampling". In the context of example 3.1 we could say that $(\epsilon_1, \dots, \epsilon_n)$ is a random sample from a $\mathcal{N}(0,1)$ distribution.

The i.i.d. case is the classical case treated in mathematical statistics and it has a number of convenient properties. For example, suppose that $X = (X_1, \dots, X_n)$ are a random sample, each with cdf $F$. Then, with the inequality understood to apply componentwise, we have that

$$F_X(x) \coloneqq P(X \le x) = \prod_{i=1}^{n} P(X_i \le x_i) = \prod_{i=1}^{n} F(x_i).$$

Nevertheless assuming that data form a random sample is not always appropriate; section 5 of this course, on Time Series, will provide many examples where our data consist of *dependent* random variables.

### 3.1.2 Dominated models

It is often technically convenient to work with *dominated models.* The precise definition of this is beyond the scope of the course. All we will need to know is the following: if each $P_\theta \in \{P_\theta : \theta \in \Theta\} = \mathcal{P}$ has a density or mass function, $p_\theta$, then $\mathcal{P}$ is dominated. For a dominated model, $p_\theta$ will always represent the density / mass function corresponding to $P_\theta$.

### 3.1.3 Identifiability

An important consideration in any model is *identifiability.* A parameter $\theta$ is *identifiable* if the map $\theta \mapsto P_\theta$ is *injective.* That is: if $\theta_1 \neq \theta_2$, then $P_{\theta_1} \neq P_{\theta_2}$.

Example 3.2 [Unidentifiable parameter]: Suppose we observe one observation $X$ which satisfies

$$X = \alpha + \epsilon, \quad \epsilon \sim \mathcal{N}(\mu, 1), \quad \theta = (\alpha, \mu) \in \Theta = \mathbb{R}^2.$$

Here our model is $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ where each $P_\theta$ has density

$$p_\theta(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x - \alpha - \mu}{2}\right)^2.$$

$\theta$ is not identifiable. Let $\theta_1 = (0, 0)$ and $\theta_2 = (\alpha, -\alpha)$ for any $\alpha \in \mathbb{R} \setminus \{0\}$. Then, $\theta_1 \neq \theta_2$ but $p_{\theta_1} = p_{\theta_2}$. △

### 3.1.4 Statistics vs. Probability

One of the key differences between statistics and probability is that in statistics we with with many probability functions: all the $P \in \mathcal{P}$. This means that we may need to clarify with which we are performing certain operations. For example, we may write $\mathbb{E}_P X$ to indicate the expectation of $X$ when $X$ has the distribution induced by $P$, i.e. $\mathbb{E}_P X = \int xp(x)\,\mathrm{d}x$.

## 3.2 Statistics, Sufficiency, Completeness & Ancilliarity

Suppose that $X = (X_1, \ldots, X_n)$ comprise our data. A *statistic* is a (measurable) function of the data, $T = T(X)$.

Example 3.3 [Sample mean]: The sample mean of data $X = (X_1, \ldots, X_n)$ is

$$T(X) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

△

### 3.2.1 Sufficiency

Suppose the distribution of $X$ belongs to the model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Then we say that the statistic $T = T(X)$ is *sufficient* for $\mathcal{P}$ (or for $X$ or $\theta$) if the conditional distribution of $X$ given $T(X)$ does not depend on $\theta$ (for all $\theta \in \Theta$).

A equivalent condition for a statistic to be sufficient for a dominated model is given by the *factorisation theorem.*

Theorem 3.1 [Factorisation Theorem]: *Suppose that $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is dominated Then $T(X)$ is sufficient for $\mathcal{P}$ iff there exist functions $g_\theta \geq 0$ and $h \geq 0$ such that*

$$p_\theta(x) = g_\theta(T(x))h(x). \tag{12}$$

*Proof.* A proof for the discrete case only can be found in [4] (pp. 276 – 277). Section 6.4 of [8] contains a full proof. □

We now give some examples of sufficient statistics.

Example 3.4 [A trivial sufficient statistic]: Suppose that the distribution of $X$ belongs to some model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. $T(X) = X$ is a sufficient statistic for $\mathcal{P}$.

To see this note that the conditional distribution of $X$ given $X = t$ is a discrete distribution with all mass at $t$. Clearly this does not depend on $\theta$. △

The above example is a sufficient statistic, but (usually) not a particularly useful one. The benefit of sufficiency is *data reduction.* That is, we want to reduce our data, $X$, to a statistic, $T(X)$, without losing any information (about $\mathcal{P}$) that $X$ contains. This situation is illustrated in the following example.

Example 3.5 [Sufficient statistic for Normal random sample]: Suppose that we observe $X = (X_1, \ldots, X_n)$, which are i.i.d. each with a $\mathcal{N}(\mu, \sigma^2)$ distribution. Our model is $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ with $\theta = (\mu, \sigma^2)$, $\Theta \in \mathbb{R} \times (0, \infty)$ and

$$p_\theta(x) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right).$$

Then $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is sufficient for $\mathcal{P}$.

This follows from Theorem 3.1 applied with $h(x) = 1$ and

$$g_\theta(t) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{n\mu^2}{2\sigma^2}\right) \exp\left(-\frac{1}{2\sigma^2}\left[t_1 - 2\mu t_2\right]\right),$$

since we can re-write

$$p_\theta(x) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{n\mu^2}{2\sigma^2}\right) \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i\right]\right),$$

and so $p_\theta(x) = g_\theta(T(x))h(x)$. △

There are typically many sufficient statistics for any given model. We say that a statistic $T = T(X)$ is *minimally sufficient* for $\mathcal{P}$ if (i) it is sufficient and (ii) for any other sufficient statistic $\tilde{T}$ there is a function $f$ such that $T = f(\tilde{T})$ ($\mathcal{P}$-a.e.).[46]

[46] The qualification $\mathcal{P}$-a.e. means "$\mathcal{P}$-almost everywhere", that is the event $A := \{T \neq f(\tilde{T})\}$ satisfies $P(A) = 0$

Example 3.6 [Minimal sufficient statistic for Normal random sample]: In the setting of Example 3.5, the statistic $T(X) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is minimally sufficient.

Example 3.5 demonstrated that $T(X)$ is sufficient. Let $\tilde{T}(X)$ be another sufficient statistic and note that by Theorem 3.1 and Example 3.5, for some $g_\theta$ and any $\theta, \theta_0 \in \Theta$

$$\frac{p_\theta(x)}{p_{\theta_0}(x)} = \left(\frac{\sigma_0}{\sigma}\right)^n \exp\left(\frac{n}{2}\left[\frac{\mu_0^2}{\sigma_0^2} - \frac{\mu^2}{\sigma^2}\right]\right) \exp\left(\frac{1}{2\sigma_0^2}(t_1 - 2\mu_0 t_2) - \frac{1}{2\sigma^2}(t_1 - 2\mu t_2)\right) = \frac{g_\theta(\tilde{T}(x))}{g_{\theta_0}(\tilde{T}(x))}.$$

Taking logarithms, evaluating at $\theta = (0,1)$ and $\theta_0 = (0, 1/2)$ and re-arranging yields

$$t_1 = T_1(x) = f_1(\tilde{T}(x)) = 2\left[\log\left(g_{(0,1)}(\tilde{T}(x))/g_{(0,1/2)}(\tilde{T}(x))\right) - n\log(1/4)\right].$$

Similarly, evaluating at $\theta = (1/2, 1/2)$ and $\theta_0 = (0, 1/2)$ and re-arranging gives

$$t_2 = T_2(x) = f_2(\tilde{T}(x)) = \left[\log\left(g_{(1/2,1/2)}(\tilde{T}(x))/g_{(0,1/2)}(\tilde{T}(x))\right) + n/4\right]$$

Conclude by taking $f = (f_1, f_2)$. △

Theorem 3.2: *Suppose that $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is dominated and has a sufficient statistic $T = T(X)$. Let $p_\theta$ denote the density or mass functions. If*

$$p_\theta(x) = p_\theta(y)\phi(x, y)$$

*implies $T(x) = T(y)$ where $\phi$ is a (measurable) function, then $T$ is minimally sufficient.*[47]

**Proof:* See Theorem 6.2.13 in [4] for a proof in a simplified special case. Theorem 3.11 in [8] gives the main idea of the full proof without technical details. Theorem 2.3(iii) in [14] has all the details. □

### 3.2.2 Completeness

A statistic $T = T(X)$ is *complete* for $\mathcal{P}$, iff for any (measurable) function $f$, $\mathbb{E}_P f(T) = 0$ for all $P \in \mathcal{P}$ implies that $f(T) = 0$ ($\mathcal{P}$-a.e.).

Theorem 3.3: *Suppose that $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a family of probability distributions and a minimally sufficient statistic for $\mathcal{P}$ exists. If $T = T(X)$ is a complete, sufficient statistic for $\mathcal{P}$ then $T$ is minimally sufficient for $\mathcal{P}$.*

*Proof.* Let $\tilde{T}$ be a minimal sufficient statistic. There is a function $f$ such that $\tilde{T} = f(T)$. Let $g(\tilde{T}) := \mathbb{E}[T|\tilde{T}]$ which does not depend on $\theta$ by sufficiency of $\tilde{T}$. By the law of iterated expectations $\mathbb{E}_{P_\theta} g(\tilde{T}) = \mathbb{E}_{P_\theta} T$ and hence $\mathbb{E}_{P_\theta}[T - g(\tilde{T})] = 0$ (for all $\theta \in \Theta$). Since $T - g(\tilde{T}) = T - g(f(T)) = h(T)$, this is a function of $T$ and hence by completeness, $h(T) = 0$ ($\mathcal{P}$-a.e.). Hence, $T = g(\tilde{T})$ ($\mathcal{P}$-a.e.). Now consider any other sufficient statistic, $\breve{T}$. By minimal sufficiency,

for all $P \in \mathcal{P}$.

[47] Another way to describe the condition on the densities is that $p_\theta(x)/p_\theta(y)$ is constant as a function of $\theta$ (if $p_\theta(y) > 0$).

$\tilde{T} = f(\breve{T})$ for some measurable $f$ ($\mathcal{P}$-a.e.). Thus, $T = g(f(\breve{T}))$ ($\mathcal{P}$-a.e.) and $T$ is minimally sufficient. □

Remark 3.1: *The condition that there* exists *a minimally sufficient statistic for $\mathcal{P}$ in Theorem 3.3 is automatically satisfied if $\mathcal{P}$ is a dominated family of probability distributions on a Euclidean space (i.e. $X \in \mathbb{R}^K$ for some $K \in \mathbb{N}$).*

*This follows from Theorem 6.3 in [11].*

Example 3.7 [Uniform distribution]: Suppose that $X_1, \ldots, X_n$ are i.i.d. from the Uniform distribution on $(0, \theta)$, with $\theta > 0$. Then $T = \max\{X_1, \ldots, X_n\}$ is complete, sufficient statistic for $\theta$ and hence a minimal sufficient statistic.

The density of a single Uniform draw, $X_i$, on $(0, \theta)$ is $1/\theta$ if $X_i \in (0, \infty)$ and 0 otherwise. Hence, the joint density can be written as $\mathbf{1}\{\min_{i=1,\ldots,n} x_i > 0\}\mathbf{1}\{\max_{i=1,\ldots n} x_i < \theta\}/\theta^n$. By (the factorisation) Theorem 12, $T$ is sufficient.

For completeness observe that by independence

$$P_\theta(T \leq t) = P_\theta(X_1 \leq t, \ldots, X_n \leq t) = P_\theta(X_1 \leq t) \times \cdots \times P_\theta(X_n \leq t) = \left[\int_0^t \frac{1}{\theta} \, \mathrm{d}x\right]^n = \frac{t^n}{\theta^n}.$$

Taking the derivative, we obtain that $p_\theta(t) = \frac{nt^{n-1}}{\theta^n}$ (for $t \in (0, \theta)$). Now suppose that $\mathbb{E}_{P_\theta} f(T) = 0$ for all $\theta \in \Theta$. Then, for any $\theta > 0$,

$$\frac{n}{\theta^n} \int_0^\theta f(t) t^{n-1} \, \mathrm{d}t = 0.$$

This implies that $f(t) = 0$ for almost all $t > 0$. Since $T(X) \leq 0$ has probability 0 under any $P \in \mathcal{P}$, this gives us that $f(T(x)) = 0$ $\mathcal{P}$-a.e. and hence $T$ is complete.

That $T$ is minimal sufficient then follows from Theorem 3.3. △

### 3.2.3 Ancilliarity

If $V = V(X)$ is a statistic whose distribution does not depend on $\theta$, then we say that $V$ is *ancillary* (for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$). Such a statistic provides no information about $\theta$.

Theorem 3.4 [Basu's Theorem]: *If $T = T(X)$ is complete and sufficient for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and $V = V(X)$ is ancillary, then $T$ and $V$ are independent under any $P \in \mathcal{P}$.*

*Proof.* Define $q_A(T) := P_\theta(V \in A|T)$ and $p_A := P_\theta(V \in A)$. By sufficiency and ancillarity neither of these depend on $\theta$. By the LIE

$$p_A = P_\theta(V \in A) = \mathbb{E}_{P_\theta} P_\theta(V \in A|T) = \mathbb{E}_{P_\theta} q_A(T).$$

It follows by completeness that $p_A = q_A(T)$ $\mathcal{P}$-a.e.. Using this & the LIE again, for arbitrary events $A, B$,

$$\begin{aligned} P_\theta(T \in B, V \in A) &= \mathbb{E}_{P_\theta}\left[\mathbf{1}_B(T)\, \mathbb{E}_{P_\theta}[\mathbf{1}_A(V)|T]\right] \\ &= \mathbb{E}_{P_\theta}\left[\mathbf{1}_B(T) q_A(T)\right] \\ &= \mathbb{E}_{P_\theta}\left[\mathbf{1}_B(T)\right] p_A \\ &= P_\theta(T \in B) P_\theta(V \in A), \end{aligned}$$

and hence $T, V$ are independent under $P_\theta$. □

### 3.2.4 Exponential families

A large number of (but by no means all) commonly used distributions belong to an *exponential family.*

Let $h : \mathbb{R}^n \to \mathbb{R}$ be a non-negative function and $T_1, \ldots, T_s$ each (measurable) functions from $\mathbb{R}^n \to \mathbb{R}$. Define

$$A(\eta) := \log \int \exp\left[\sum_{i=1}^{s} \eta_i T_i(x)\right] h(x)\, \mathrm{d}x,$$

and let $\Xi := \{\eta \in \mathbb{R}^s : A(\eta) < \infty\}$. The family $\{P_\eta : \eta \in \Xi\}$ is an *$s$-parameter exponential family* in *canonical form* iff the density $p_\eta$ has the form

$$p_\eta(x) := \exp\left[\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right] h(x). \tag{13}$$

$\Xi$ is the *natural parameter space* of our exponential family. We can also consider subfamilies: $\{P_\eta : \eta \in \Gamma\}$ with $\Gamma \subset \Xi$. Other parametrisations are possible: let $R : \Theta \to \Xi$ and define

$$p_\theta(x) := \exp\left[\sum_{i=1}^{s} R_i(\theta) T_i(x) - B(\theta)\right] h(x), \tag{14}$$

for some function $h$, $\theta \in \Theta$, $x \in \mathbb{R}^n$, where $B(\theta) := A(R(\theta))$.[48] Let $\mu$ be some ($\sigma$-finite) measure on $\mathbb{R}^n$. The family $\{P_\theta : \theta \in \Theta\}$ where $p_\theta$ is the density of $P_\theta$ is a *$s$-parameter exponential family.*

The canonical form of an exponential family is not always the most "natural" parametrisation. Nevertheless, it is often convenient to work with the canonical form.

Example 3.8 [Normal distribution]: If $X = (X_1, \ldots, X_n)$ is an i.i.d. sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution then the density of $X$ is

$$p_\theta(x) = \frac{1}{(2\pi)^{n/2}} \exp\left(\frac{1}{2\sigma^2}\left[2\mu \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^2 - n\mu^2\right] - n \log \sigma\right)$$

and $\mathcal{P} = \{p_\theta : \theta = (\mu, \sigma) \in \mathbb{R} \times (0, \infty) = \Theta\}$ is a 2-parameter exponential family. To see

[48] $h$ need not be the same as in equation (13). This choice of $B$ ensures that this is a probability density, i.e. it integrates to 1.

this, put $T(x) = \left(\sum_{i=1}^n x_i, -\sum_{i=1}^n x_i^2\right)$, $R(\theta) = (\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2})$, $\theta = (\mu, \sigma) \in \Theta = \mathbb{R} \times (0, \infty)$, $B(\theta) = \frac{n\mu^2}{2\sigma^2} + n \log \sigma$ and $h(x) = (2\pi)^{-n/2}$.

To obtain the canonical form, let $\eta = (\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2})$ and then $\Xi = \mathbb{R} \times (0, \infty)$ with $A(\eta) = n[\eta_1^2/(4\eta_2) + \log(1/\sqrt{2\eta_2})]$. △

Proposition 3.1 [Sufficient statistics in exponential families]: *If $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a $s$-parameter exponential family, then $T = (T_1(X), \ldots, T_s(X))$ is a sufficient statistic for $\theta$.*

*Proof.* The densities $p_\theta$ of the family may be written as

$$p_\theta(x) = \exp\left[\sum_{i=1}^s R_i(\theta)T_i(x) - B(\theta)\right] h(x) = g_\theta(T(x))h(x),$$

with $g_\theta(T(x)) = \exp\left[\sum_{i=1}^s R_i(\theta)T_i(x) - B(\theta)\right]$. Since $g_\theta(T(x)) \geq 0$, the same must be true of $h$. That $T$ is sufficient then follows from Theorem 3.1. □

Proposition 3.2 [Complete statistics in exponential families]: *Let $\mathcal{P} = \{P_\eta : \eta \in \Gamma\}$ be a subfamily of a $s$-parameter exponential family in canonical form. If the interior of $\Gamma$ is non-empty, then $T = (T_1(X), \ldots, T_s(X))$ is complete for $\mathcal{P}$.*

**Proof:* E.g. Theorem 2.74 in [13]. □

Example 3.9: Suppose that $X = (X_1, \ldots, X_n)$ consists of i.i.d. draws from a $\mathcal{N}(\theta, \sigma^2)$ where $\theta \in \mathbb{R}$ and $\sigma^2 > 0$ is known. Our model is $\mathcal{P} = \{\mathcal{N}(\theta, \sigma^2)^n : \theta \in \mathbb{R}\}$. The sample mean $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ is a complete sufficient statistic for $\theta$ and the sample variance $S_n^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ is ancillary.

Letting $\bar{x} := \frac{1}{n}\sum_{i=1}^n x_i$, the density of $X$ is

$$p_\theta(x) := \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\frac{n\theta}{\sigma^2}\bar{x} - \frac{n\theta^2}{2\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\right)$$

Taking $T(x) = \bar{x}$, $R(\theta) = n\theta/\sigma^2$, $B(\theta) = n\theta^2/(2\sigma^2)$ and $h(x) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\right)$, it is clear this forms a 1-parameter exponential family and hence $T$ is sufficient by Proposition 3.1. Taking $\eta = n\theta/\sigma^2 \in \Xi = \mathbb{R}$ and $A(\eta) = \eta^2\sigma^2/(2n)$ we obtain an 1-parameter exponential family in canonical form. Clearly $\Xi = \mathbb{R}$ has non empty interior and hence $T$ is also complete by Proposition 3.2.

It remains to show that $S_n^2$ is ancillary. For this let $Y_i := X_i - \theta$ for $i = 1, \ldots, n$. Then $Y_i \sim \mathcal{N}(0, \sigma^2)$. Since

$$\bar{Y}_n := \frac{1}{n}\sum_{i=1}^n Y_i = \frac{1}{n}\sum_{i=1}^n X_i - \theta,$$

we have that $X_i - \bar{X}_n = Y_i - \bar{Y}_n$ and so

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2.$$

Hence the distribution of $S_n^2$ is determined by that of $Y = (Y_1, \dots, Y_n)$. As the $Y_i$ are i.i.d. with each $Y_i \sim \mathcal{N}(0, \sigma^2)$, the distribution of $Y$ (and hence $S_n^2$) does not depend on $\theta$.

By Basu's Theorem (Theorem 3.4) it follows that $\bar{X}_n$ and $S_n^2$ are independent. △

## 3.3 Statistical inference and performance criteria

There are a number of common tasks which generally come under the heading of "statistical inference". In this section we will discuss three such tasks: estimation, hypothesis testing and the formation of confidence sets. The latter two tasks are very closely related.

Moreover, we will discuss how "performance" is measured in each of these cases. For example, there are often many estimators available for a certain parameter of interest, but their accuracy may differ.

### 3.3.1 Risk of an estimator

The goal of estimation is to find a statistic, $T = T(X)$, which is "close" in some sense to some function of interest $g(\theta)$. Typically, in this setting, our statistic $T$ is called an *estimator*.

Estimators are evaluated based on a *loss functions*. A loss function is a function $L : \Theta \times \mathcal{T} \to [0, \infty)$, where $\mathcal{T}$ contains the range of possible estimates, such that $L(\theta, g(\theta)) = 0$.

The *loss* of a estimator $T = T(X)$ given data $X$ where $X$ has distribution $P_\theta$ is $L(\theta, T(X))$. As this is a random quantity, we judge estimators by their average loss, or *risk*:

$$R(\theta, T) \coloneqq \mathbb{E}_{P_\theta} L(\theta, T(X)).$$

Example 3.10 [Quadratic loss]: Suppose, given some model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, we want to estimate some $g(\theta) \in \mathcal{G} \subset \mathbb{R}^K$. One commonly used loss function is the *quadratic loss*, defined for $t \in \mathbb{R}^K$ as

$$L(\theta, t) = \|g(\theta) - t\|^2.$$

The risk function of an estimate $T = T(X)$ under the quadratic loss is then

$$R(\theta, T) = \mathbb{E}_{P_\theta} \|g(\theta) - T(X)\|^2.$$

This risk function is often called the *mean squared error* and satisfies a decomposition into bias and variance terms ("the bias-variance tradeoff"):

$$R(\theta, T) = \mathbb{E}_{P_\theta} \|g(\theta) - T(X)\|^2 = \sum_{k=1}^{K} \mathrm{Var}_{P_\theta}(T_k(X)) + \left(\mathbb{E}_{P_\theta} T_k(X) - g_k(\theta)\right)^2. \tag{15}$$

△

An estimator $T = T(X)$ is *inadmissible* if there is another estimator $\tilde{T} = \tilde{T}(X)$ with a better risk function: $R(\theta, \tilde{T}) \leq R(\theta, T)$ for all $\theta \in \Theta$ and $R(\theta^\star, \tilde{T}) < R(\theta^\star, T)$ for (at least) one $\theta^\star \in \Theta$. If there is no such $\tilde{T}$, then $T$ is *admissible*.

The following result demonstrates that when we have a sufficient statistic and a convex loss function, any admissible estimator is based only on that sufficient statistic.

Theorem 3.5 [Rao-Blackwell]: *Suppose that $X$ has a distribution in $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and that $T = T(X)$ is sufficient for $\mathcal{P}$. Let $\delta = \delta(X)$ be an estimator of $g(\theta)$ and put $\gamma(T) := \mathbb{E}[\delta(X)|T]$.*[49] *If $R(\theta, \delta) < \infty$, $a \mapsto L(\theta, a)$ is convex (this includes the requirement that its domain is a convex set) then*

$$R(\theta, \gamma) \leq R(\theta, \delta).$$

*If $a \mapsto L(\theta, a)$ is strictly convex, the preceding inequality will be strict unless $\delta(X) = \gamma(T)$ with probability 1.*

*Proof.* By using (the conditional version of) Jensen's inequality,

$$L(\theta, \gamma(T)) = L(\theta, \mathbb{E}[\delta(X)|T]) \leq \mathbb{E}\left[L(\theta, \delta(X))|T\right].$$

where the expectation is with respect to the conditional distribution of $\delta(X)$, given $T$, under $\theta$. Take expectations on both sides (and use the LIE) to obtain $R(\theta, \gamma) \leq R(\theta, \delta)$.

For the second claim, note that if $\delta(X) \neq \gamma(T)$ a.s., then $\frac{1}{2}\left[\delta(X) + \gamma(T)\right]$ is not a.s. constant.[50] Then by strict convexity

$$L\left(\theta, \delta(X)/2 + \gamma(T)/2\right) < \frac{1}{2} L(\theta, \delta(X)) + \frac{1}{2} L(\theta, \gamma(T)),$$

and so using the strict Jensen inequality (e.g. the second part of Theorem 3.25 in [8]) similarly to as in the first part, we have that

$$L(\theta, \gamma(T)) < \mathbb{E}\left[L(\theta, \delta(X))|T\right].$$

Taking expectations on both sides completes the proof. □

A *randomised estimator* is an estimator that can be written as $\delta(X, U)$ where $U$ is a draw from a uniform distribution on $(0, 1)$, independendent of the data $X$. That is, a randomised estimator is constructed from the data and *auxillary randomisation.* There are often technical advantages from permitting such procedures. Nevertheless, in many cases (as the next corollary demonstrates), these estimators are inadmissible.

Corollary 3.1: *Suppose that $X$ has a distribution in $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Let $\delta(X, U)$ be a randomised estimator of $g(\theta)$. If the loss function $a \mapsto L(\theta, a)$ is convex, then the non-randomised estimator $\gamma(X) := \mathbb{E}[\delta(X, U)|X]$ has a risk function no worse than $\delta$. If $a \mapsto L(\theta, a)$ is strictly convex and $\gamma(X) \neq \delta(X, u)$ with probability 1 then the risk function of the non-randomised estimator is strictly better.*

*Proof.* By example 3.4, $X$ is sufficient. If $a \mapsto L(\theta, a)$ is convex, then apply Theorem 3.5 to conclude that $R(\theta, \gamma) \leq R(\theta, \delta)$, which is the first claim. If $a \mapsto L(\theta, a)$ is strictly convex and $\gamma(X) \neq \delta(X, u)$ with probability 1 then the second part of Theorem 3.5 allows the conclusion

[49] Implicitly here we assume that $\delta(X)$ is $P_\theta$-integrable such that this conditional expectation exists.

[50] Suppose it were constant. Then we have that (a.s.) $\gamma(T) + \delta(X) = c$ for some $c$. Taking conditional expectations we get $2\gamma(T) = c$, or $\gamma(T) = c/2$ and hence $\delta(X) = c - \gamma(T) = c/2$, which gives $\delta(X) = \gamma(T)$ a.s..

that $R(\theta, \gamma) < R(\theta, \delta)$, the second claim. □

In many cases there is no uniformly best estimator.

Example 3.11: Consider the model $\{\mathcal{N}(\theta, 1)^n : \theta \in \mathbb{R}\}$ for (the distribution of) the random sample $X = (X_1, \dots, X_n)$. We will consider two estimators for the mean parameter $\theta$ and examine their risk under quadratic loss. Firstly we consider the sample mean:

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i,$$

and second lets consider an estimator which "shrinks" from the sample mean towards some fixed, known $\theta_0$, based on, for example, prior information of the researcher:

$$\tilde{X}_n := a\theta_0 + (1-a)\bar{X}_n, \quad a \in [0, 1].$$

$\bar{X}_n$ is unbiased since

$$\mathbb{E}_{P_\theta} \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{P_\theta} X_i = \theta,$$

and therefore its risk is (by example 3.10 and independence)

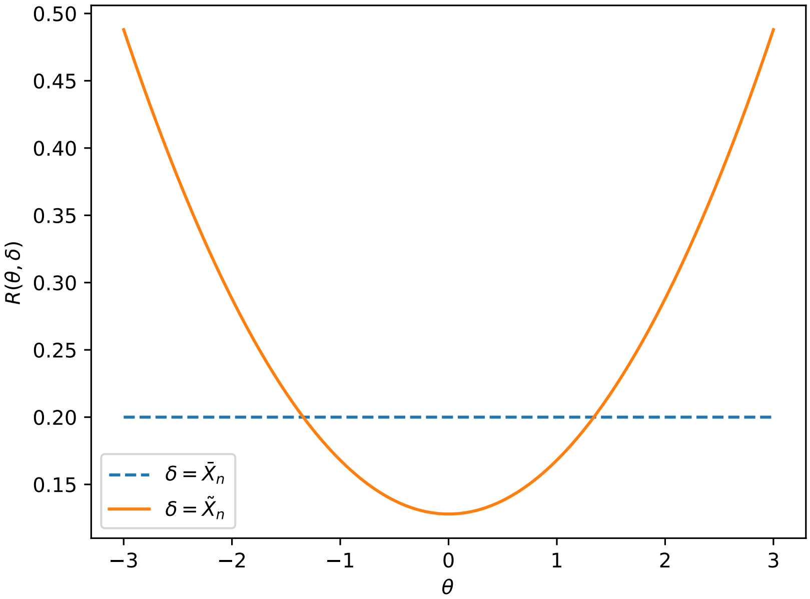
$$R(\theta, \bar{X}_n) = \mathrm{Var}_{P_\theta} \bar{X}_n = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}_{P_\theta}(X_i) + \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{i} \mathrm{Cov}_{P_\theta}(X_i, X_j) = \frac{1}{n} \mathrm{Var}_{P_\theta}(X_i) = \frac{1}{n}.$$

By example 3.10 we can write the risk of $\tilde{X}_n$ as

$$R(\theta, \tilde{X}_n) = (1-a)^2 \mathrm{Var}_{P_\theta} \bar{X}_n + a^2[\theta - \theta_0]^2 = \frac{(1-a)^2}{n} + a^2[\theta - \theta_0]^2.$$

The ranking of these risk functions depends on the values of $n$, $a$, $\theta_0$. Figure 15 below plots the case with $n = 5$, $a = 1/5$ and $\theta_0 = 0$. As can be seen in the figure, the "shrunk" estimator $\tilde{X}_n$ has lower risk than $\bar{X}_n$ when $\theta$ is close to $\theta_0 = 0$ and larger risk otherwise.

FIGURE 15: RISK FUNCTIONS FOR $\bar{X}_n$ AND $\tilde{X}_n$



△

### 3.3.2 Unbiasedness, UMVU estimation and the Cramér - Rao lower bound

One potentially desirable feature of an estimator is *unbiasedness*. An estimator $\delta(X)$ of $g(\theta)$ is *unbiased* iff $\mathbb{E}_{P_\theta}\, \delta(X) = g(\theta)$ for all $\theta \in \Theta$.

Whilst unbiasedness seems like a desirable property, insisting on unbiasedness may not always be desirable. There are many examples in which requiring that estimators are unbiased rules out reasonable estimators. A simple example is given below.

Example 3.12: In example 3.7 we saw that given a random sample $X = (X_1, \dots, X_n)$ from a $U(0, \theta)$ distribution, $T = \max\{X_1, \dots, X_n\}$ is a complete and sufficient statistic.

We can consider as estimators for $\theta$ the family $\delta_a := aT$ for $a \in (0, \infty)$. Let us compare the risk of these estimators under quadratic loss. By example 3.10 we have

$$R(\theta, \delta_a) = \mathrm{Var}_{P_\theta}(\delta_a(X)) + \left[\mathbb{E}_{P_\theta}\, \delta_a(X) - \theta\right]^2.$$

Since $\mathbb{E}_{P_\theta}[\delta_a] = a\, \mathbb{E}_{P_\theta}[T]$ and $\mathrm{Var}_{P_\theta}[\delta_a] = a^2 \mathrm{Var}_{P_\theta}[T]$, we will calculate the first two moments of $T$. Using the density of $T$ derived in example 3.7:

$$\mathbb{E}_{P_\theta}\, T = \int_0^\theta t \frac{n t^{n-1}}{\theta^n}\, \mathrm{d}t = \frac{n\theta^{n+1}}{\theta^n (n+1)} = \frac{n}{n+1}\theta,$$
$$\mathbb{E}_{P_\theta}\, T^2 = \int_0^\theta t^2 \frac{n t^{n-1}}{\theta^n}\, \mathrm{d}t = \frac{n\theta^{n+2}}{\theta^n (n+2)} = \frac{n}{n+2}\theta^2.$$

From this it is clear that $\delta_a$ is unbiased iff $a = a^\star = \frac{n+1}{n}$. The risk of this unbiased estimator

$\delta_{a^\star}$ is

$$R(\theta, \delta_{a^\star}) = \operatorname{Var}_{P_\theta}(\delta_{a^\star}(X)) = \left(\frac{n+1}{n}\right)^2 \left[\frac{n\theta^2}{n+2} - \left(\frac{n\theta}{n+1}\right)^2\right] = \frac{\theta^2}{n^2+2n}.$$

By contrast, the risk for an arbitrary $\delta_a$ may be calculated as

$$R(\theta, \delta_a) = \mathbb{E}_{P_\theta}[aT - \theta]^2 = a^2\,\mathbb{E}_{P_\theta} T^2 + 2a\theta\,\mathbb{E}_{P_\theta} T + \theta^2 = \theta^2 \left(\frac{n}{n+2}a^2 - \frac{2n}{n+1}a + 1\right),$$

which is minimised by $a = a_\star = \frac{n+2}{n+1}$; note that this estimator is *not* unbiased [Exercise]. This choice of $a$ leads to a risk of

$$R(\theta, \delta_{a_\star}) = \frac{\theta^2}{(n+1)^2} < \frac{\theta^2}{n^2+2n} = R(\theta, \delta_{a^\star}).$$

△

Remark 3.2: *Example 3.12 provides a simple example of a case where an unbiased estimator may have larger risk than a biased estimator. In fact, as we shall see below (in example 3.13), the unbiased estimator here has minimum risk of any unbiased estimator in this problem. Hence, we have exhibited a case where the best unbiased estimator is dominated by a biased estimator.*

*In this particular example the difference in risk declines with n and is not particularly large. Nevertheless there are cases where insisting on unbiasedness can lead to absurd choices of estimators; see e.g. Example 4.7 in [8].*

Under quadratic loss, $L$, the risk of an unbiased estimator of $g(\theta) \in \mathcal{G} \subset \mathbb{R}$ is $R(\theta, L) = \operatorname{Var}_{P_\theta}(\delta(X))$ and so we can rank estimators by their variance. An unbiased estimator $\delta$ is *uniformly minimum vairance unbiased* (UMVU) iff $\operatorname{Var}_{P_\theta}(\delta) \leq \operatorname{Var}_{P_\theta}(\delta^*)$ for all $\theta \in \Theta$ for any other unbiased estimator $\delta^*$.

In general, UMVU estimators may not exist.[51] The following Theorem demonstrates that when an unbiased estimator exists and there is a complete, sufficient statistic, a UMVU estimator exists.

Theorem 3.6 [Lehmann - Scheffé]: *Suppose that $X$ has a distribution in $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, $\delta$ is an unbiased estimator of $g(\theta) \in \mathcal{G} \subset \mathbb{R}$ and $T = T(X)$ is complete and sufficient for $\mathcal{P}$. Then, $\eta(T) := \mathbb{E}[\delta|T]$ is UMVU. Moreover any other UMVU estimator $\eta^*(T)$ based on $T$ is equal to $\eta(T)$ $\mathcal{P}$ – almost surely.*

*Proof.* By the LIE, $\eta(T)$ is unbiased since $g(\theta) = \mathbb{E}_{P_\theta} \delta = \mathbb{E}_{P_\theta}[\mathbb{E}[\delta|T]] = \mathbb{E}_{P_\theta} \eta(T)$. Now if $\eta^\star(T)$ is also unbiased (by hypothesis),

$$\mathbb{E}_{P_\theta}[\eta(T) - \eta^\star(T)] = 0 \quad \text{for all} \quad \theta \in \Theta,$$

and so by completeness $\eta(T) = \eta^\star(T) = 0$ $\mathcal{P}$-a.s.. Hence any unbiased estimator based on $T$

[51] Neverthless in many cases there are estimators which are *asymptotically* unbiased; this will be discussed further in section 3.4.

is essentially unique and can differ from $\eta(T)$ only on a $\mathcal{P}$-null set. That $\eta(T)$ has minimum variance follows by combining (the Rao-Blackwell) Theorem 3.5 and Example 3.10. □

The preceding theorem provides two recipes to find UMVU estimators. Firstly, if one has an unbiased estimator $\delta(X)$, taking its conditional expectation given a complete, sufficient statistic $T$ yields a UMVU estimator. Secondly, by the uniqueness, if $\eta(T)$ is unbiased and $T$ complete and sufficient, then any $\delta = \eta(T)$ ($\mathcal{P}$-a.s.) must be UMVU.

Example 3.13 [UMVU estimator for Uniform]: In the setting of Examples 3.7 and 3.12, the UMVU estimator for $\theta$ is $\frac{n+1}{n}T = \frac{n+1}{n}\max\{X_1, \dots, X_n\}$.

First consider estimating $\theta/2$ and note that $\mathbb{E}_{P_\theta} X_1 = \theta/2$. Hence $X_1$ is an unbiased estimator of $\theta/2$. We may thus obtain a UMVU estimate by taking its expectation conditional on the complete sufficient statistic $T = \max\{X_1, \dots, X_n\}$.

Conditional on $T = t$, $X_1 = t$ with probability $1/n$ and is uniformly distributed on $(0, t)$ with probability $(n-1)/n$ [Exercise]. Hence,

$$\mathbb{E}_{P_\theta}[X_1 | T = t] = \frac{1}{n}t + \frac{n-1}{n}\mathbb{E}_{P_\theta}[X_1 | X_1 \neq t] = \frac{1}{n}t + \frac{n-1}{n}\frac{t}{2} = \frac{n+1}{n}\frac{t}{2}.$$

Thus $d = \frac{n+1}{n}\frac{T}{2}$ is UMVU for $\theta/2$. That $\delta = 2d = \frac{n+1}{n}T$ is UMVU for $\theta$ follows as (i) it is unbiased (as $d$ is unbiased by the LIE and $\mathbb{E}_{P_\theta}\delta = 2\,\mathbb{E}_{P_\theta} d = 2\theta/2 = \theta$) and (ii) if there were an alternative unbiased estimator $\delta^\star$ with $\mathrm{Var}_{P_\theta}(\delta^\star) < \mathrm{Var}_{P_\theta}(\delta)$ then $d^\star := \delta^\star/2$ is an unbiased estimator for $\theta/2$ with $\mathrm{Var}_{P_\theta}(d^\star) = \frac{1}{4}\mathrm{Var}_{P_\theta}(\delta^\star) < \frac{1}{4}\mathrm{Var}_{P_\theta}(\delta) = \mathrm{Var}_{P_\theta}(d)$, a contradiction. △

We will now consider the following question: if our model is "well – behaved", what is the lowest possible variance of an unbiased estimator of $g(\theta)$?[52] To start with, we note that if random variables $X$ and $Y$ satisfy $\mathbb{E}\,X^2 < \infty$ and $\mathbb{E}\,Y^2 < \infty$, then the following *covariance inequality* holds:[53]

$$|\mathrm{Cov}(X, Y)| \leq \sqrt{\mathrm{Var}(X)^2}\sqrt{\mathrm{Var}(Y)^2}. \tag{16}$$

If $\delta$ is a statistic and $\psi$ some random variable (both with finite variance) then the covariance inequality (16) yields

$$\mathrm{Var}_{P_\theta}(\delta) \geq \frac{\mathrm{Cov}_{P_\theta}(\delta, \psi)^2}{\mathrm{Var}_{P_\theta}(\psi)}. \tag{17}$$

This can easily be generalised to the case where $\psi \in \mathbb{R}^K$ is a random vector. Let $\gamma := \mathrm{Cov}_{P_\theta}(\delta, \psi)$ and $C = \mathrm{Var}_{P_\theta}(\psi)$ be positive definite. Then for any $a \in \mathbb{R}^K$ we have by (17)

$$\mathrm{Var}_{P_\theta}(\delta) \geq \frac{\mathrm{Cov}_{P_\theta}(\delta, a'\psi)^2}{\mathrm{Var}_{P_\theta}(a'\psi)} = \frac{(a'\gamma)^2}{a'Ca} = \frac{a'\Gamma a}{a'Ca}, \quad \Gamma := \gamma\gamma'.$$

[52] We will actually answer a more general question; the answer to that posed here will follow as a corollary.

[53] This is an immediate consequence of the Cauchy-Schwarz inequality applied to $X - \mathbb{E}\,X$ and $Y - \mathbb{E}\,Y$. Alternatively see (4.11) in [8].

Since this holds for any $a \in \mathbb{R}^K$ we have that[54]

$$\operatorname{Var}_{P_\theta}(\delta) \geq \max_{a \in \mathbb{R}^K} \frac{a'\Gamma a}{a'Ca} = \gamma' C^{-1} \gamma. \tag{18}$$

By itself this is not particularly revealing, but under some conditions and with a clever choice of $\psi$, inequality (18) provides us with a lower bound for the variance of an unbiased estimator $\delta$, based on the *(Fisher) information matrix*. In particular, if the derivatives below exist, the Fisher information matrix $I(\theta)$ has $i, j$-th element

$$[I(\theta)]_{i,j} := \mathbb{E}_{P_\theta}\left[\frac{\partial \log p_\theta(X)}{\partial \theta_i}\frac{\partial \log p_\theta(X)}{\partial \theta_j}\right]. \tag{19}$$

Theorem 3.7 [The information inequality]: *Suppose that $X$ has a distribution in the dominated model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, with $\Theta \subset \mathbb{R}^K$. Let $\ell_\theta := \nabla_\theta \log p_\theta$. If $\delta$ is any (real-valued) statistic with $\mathbb{E}_{P_\theta} \delta(X)^2 < \infty$ and:*

*(i) $\ell_\theta$ exists, $\mathbb{E}_{P_\theta} \ell_\theta(X) = 0$ and $I(\theta) := \mathbb{E}_{P_\theta}[\ell_\theta(X)\ell_\theta(X)']$ exists and is positive definite,*

*(ii) $\nabla_\theta \mathbb{E}_{P_\theta}[\delta(X)]$ exists and satisfies $\nabla_\theta \mathbb{E}_{P_\theta}[\delta(X)] = \int \nabla_\theta \delta(x) p_\theta(x)\,\mathrm{d}x$,*

*then*

$$\operatorname{Var}_{P_\theta}(\delta(X)) \geq (\nabla_\theta \mathbb{E}_{P_\theta}[\delta(X)])' I(\theta)^{-1} (\nabla_\theta \mathbb{E}_{P_\theta}[\delta(X)]).$$

*Proof.* Since $\nabla_\theta \delta p_\theta = \delta \ell_\theta p_\theta$ ($P_\theta$ - a.s.), (ii) implies that $\nabla_\theta \mathbb{E}_{P_\theta}[\delta(X)] = \mathbb{E}_{P_\theta}(\delta(X), \ell_\theta(X))$. Combination with (i) gives $\nabla_\theta \mathbb{E}_{P_\theta}[\delta(X)] = \operatorname{Cov}_{P_\theta}(\delta(X), \ell_\theta(X))$. This combined with (18) completes the proof. □

Remark 3.3: *If $\delta$ is an estimator of $g(\theta)$ with bias $b(\theta) := \mathbb{E}_{P_\theta} \delta(X) - g(\theta)$ and both $\nabla_\theta g$ and $\nabla_\theta b$ exist then under the conditions of Theorem 3.7,*

$$\operatorname{Var}_{P_\theta}(\delta(X)) \geq ([\nabla_\theta g](\theta) + [\nabla_\theta b](\theta))' I(\theta)^{-1} ([\nabla_\theta g](\theta) + [\nabla_\theta b](\theta)).$$

Corollary 3.2 [The Cramér-Rao lower bound]: *Suppose that the conditions of Theorem 3.7 hold and additionally suppose that $\delta$ is an unbiased estimator of $g(\theta)$, where $g$ is differentiable with gradient $\nabla_\theta g$. Then,*

$$\operatorname{Var}_{P_\theta}(\delta) \geq ([\nabla_\theta g](\theta))' I(\theta)^{-1} ([\nabla_\theta g](\theta)).$$

[54] To see that the last equality holds, let $\|x\|_A := \|Ax\|$ for any $K \times K$ matrix $A$ where $x \in \mathbb{R}^K$ and $\|\cdot\|$ is the Euclidean norm. As $C$ is positive definite, one has

$$\frac{a'\Gamma a}{a'Ca} = \frac{a'C^{1/2}(C^{-1/2}\Gamma^{1/2})\Gamma^{1/2}C^{-1/2}(C^{1/2}a)}{(a'C^{1/2})(C^{1/2}a)} = \|C^{-1/2}\Gamma^{1/2}\|^2_{C^{1/2}},$$

where the norm on the right hand side is the operator norm induced by $\|\cdot\|_{C^1/2}$. By e.g. Theorem 5.6.7 in [7], one has

$$\|C^{-1/2}\Gamma^{1/2}\|^2_{C^{1/2}} = \|\Gamma^{1/2}C^{-1/2}\|^2 = \lambda_{\max}(\Gamma^{1/2}C^{-1}\Gamma^{1/2}).$$

Since $\Gamma^{1/2}C^{-1}\Gamma^{1/2}$ is symmetric and rank 1,

$$\lambda_{\max}(\Gamma^{1/2}C^{-1}\Gamma^{1/2}) = \operatorname{tr}(\Gamma^{1/2}C^{-1}\Gamma^{1/2}) = \operatorname{tr}(\Gamma C^{-1}) = \operatorname{tr}(\gamma\gamma' C^{-1}) = \operatorname{tr}(\gamma' C^{-1}\gamma) = \gamma' C^{-1}\gamma.$$

*Proof.* Immediate from Theorem 3.7 and Remark 3.3 since $b(\theta) = \mathbb{E}_{P_\theta}\,\delta(X) - g(\theta) = 0$. □

Remark 3.4: *In the special case where $g(\theta) = \theta \in \mathbb{R}$, the Cramér-Rao lower bound becomes*

$$\mathrm{Var}_{P_\theta}(\delta) \geq I(\theta)^{-1}.$$

A version for estimators of vector parameter can also be given.

Corollary 3.3 [Cramér-Rao lower bound]: *If $\delta(X)$ is an unbiased estimator of $g(\theta)$ where $g$ is differentiable with Jacobian $\nabla_\theta g$ and such that the conditions of Theorem 3.7 hold for each element of $\delta(X)$ then*

$$\mathrm{Var}_{P_\theta}(\delta(X)) - ([\nabla_\theta g](\theta))'\, I(\theta)^{-1}\, ([\nabla_\theta g](\theta))$$

*is positive semi-definite.*

*Proof.* Applying Corollary 3.2 to the unbiased estimators $a'\delta(X)$ of $a'g(\theta)$ implies that

$$\begin{aligned} a'\mathrm{Var}_{P_\theta}(\delta(X))a = \mathrm{Var}_{P_\theta}(a'\delta(X)) &\geq \left([\nabla_\theta a'g](\theta)\right)' I(\theta)^{-1} \left([\nabla_\theta a'g](\theta)\right) \\ &= a' \left([\nabla_\theta g](\theta)\right)' I(\theta)^{-1} \left([\nabla_\theta g](\theta)\right) a. \end{aligned}$$

As this is true for all conformable $a$, the claim follows. □

We now give some simple sufficient conditions on the densities $p_\theta$ which ensure that (i) and (ii) in Theorem 3.7 hold for any statistic $\delta$ such that $\mathbb{E}_{P_\theta}\,\delta(X)^2 < \infty$.

Lemma 3.1: *Suppose that $X$ has a distribution in the dominated model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, with $\Theta \subset \mathbb{R}^K$. Put that $\ell_\theta := \nabla_\theta \log p_\theta$ and suppose that*

*(i) $\Theta$ is open,*

*(ii) The set $A_\theta := \{x : p_\theta > 0\}$ does not depend on $\theta$,*

*(iii) $\delta$ is a real-valued statistic with $\mathbb{E}_{P_\theta}[\delta(X)^2] < \infty$,*

*(iv) $\ell_\theta$ exists and $I(\theta) := \mathbb{E}_{P_\theta}\left[\ell_\theta(X)\ell_\theta(X)'\right]$ is positive definite,*

*(v) There exists a function $d_\theta$ such that $\mathbb{E}_{P_\theta}[d_\theta(X)^2] < \infty$ and*

$$\left|\frac{p_{\theta+e_k\Delta}(x) - p_\theta(x)}{\Delta p_\theta(x)}\right| \leq d_\theta(x) \quad \text{for all } k = 1, \ldots, K \text{ and all } |\Delta| \leq \varepsilon,$$

*for some $\varepsilon > 0$ and where $e_k$ is the $k$-th canonical basis vector in $\mathbb{R}^K$.*

*Then (i) and (ii) in Theorem 3.7 hold.*

**Proof:* We start with (ii) and argue componentwise. As $\nabla_\theta \delta p_\theta = \delta \ell_\theta p_\theta$ ($P_\theta$ - a.s.) we have

$$\int e_k' \left[\nabla_\theta \delta p_\theta\right] \mathrm{d}\nu = \int \delta e_k' \ell_\theta p_\theta \,\mathrm{d}\nu = \int \delta \left[\lim_{\Delta\to 0} \frac{p_{\theta+e_k\Delta} - p_\theta}{\Delta p_\theta}\right] p_\theta \,\mathrm{d}\nu.$$

Since $\mathbb{E}_{P_\theta}\left[|\delta(x)||d_\theta(x)|\right] \leq \left[\mathbb{E}_{P_\theta}[\delta(X)^2]\right]^{1/2}\left[\mathbb{E}_{P_\theta}[d_\theta(X)^2]\right]^{1/2} < \infty$ and

$$\left|\delta(x)\frac{p_{\theta+e_k\Delta}(x)-p_\theta(x)}{\Delta p_\theta(x)}\right| \leq |\delta(x)||d_\theta(x)|,$$

by the Cauchy-Schwarz inequality and our hypotheses, it follows by the dominated convergence theorem that

$$\int \delta\left[\lim_{\Delta\to 0}\frac{p_{\theta+e_k\Delta}-p_\theta}{\Delta p_\theta}\right]p_\theta\,\mathrm{d}\nu = \lim_{\Delta\to 0}\int\delta\left[\frac{p_{\theta+e_k\Delta}-p_\theta}{\Delta p_\theta}\right]p_\theta\,\mathrm{d}\nu = \lim_{\Delta\to 0}\frac{\mathbb{E}_{P_{\theta+e_k\Delta}}\,\delta(X)-\mathbb{E}_{P_\theta}\,\delta(X)}{\Delta},$$

and the last term on the right hand side is $\frac{\partial\mathbb{E}_{P_\theta}\delta(X)}{\partial\theta_k}$. This proves (ii).

For (i), the existence of $\ell_\theta$ is true by assumption. For the second part, apply (ii) with $\delta = 1$. For the third part, firstly note that the $(i,j)$-th component of $I(\theta)$ is $\mathbb{E}\,\theta\,[e_i'\ell_\theta\ell_\theta' e_j]$. Since

$$\left|e_i'\ell_\theta\ell_\theta' e_j\right| = \lim_{\Delta\to 0}\left|\frac{p_{\theta+e_i\Delta}(x)-p_\theta(x)}{\Delta p_\theta(x)}\right|\left|\frac{p_{\theta+e_j\Delta}(x)-p_\theta(x)}{\Delta p_\theta(x)}\right|$$

and

$$\left|\frac{p_{\theta+e_i\Delta}(x)-p_\theta(x)}{\Delta p_\theta(x)}\right|\left|\frac{p_{\theta+e_j\Delta}(x)-p_\theta(x)}{\Delta p_\theta(x)}\right| \leq d_\theta(x)^2$$

with $\mathbb{E}_{P_\theta}\,d_\theta(X)^2 < \infty$, applying the dominated convergence theorem we obtain

$$\mathbb{E}_{P_\theta}\left|e_i'\ell_\theta\ell_\theta' e_j\right| = \lim_{\Delta\to 0}\int\left|\frac{p_{\theta+e_i\Delta}-p_\theta}{\Delta p_\theta}\right|\left|\frac{p_{\theta+e_j\Delta}-p_\theta}{\Delta p_\theta}\right|p_\theta\,\mathrm{d}\nu < \infty,$$

and hence $I(\theta)$ exists. That it is positive definiteness follows by hypothesis. □

### 3.3.3 Other performance criteria for estimators

As noted just preceding Theorem 3.6, in general UMVU estimators may not exist. Moreover, as discussed around Example 3.12, instisting on unbiasedness may not be sensible in a given statistical problem. Nevetheless in many problems estimators which satisfy approximate (i.e. asymptotic) versions of these properties exist and often work well. This situation will be discussed in section 3.4.

There are alternative performance criteria which can be used to evaluate estimators and (sometimes) find estimators which are optimal according to the given criteria. Three typical options are to consider estimators which satisfy an equivariance property ("equivariant"), estimators which minimise average risk (average with respect to some probability distribution on $\Theta$; "Bayes") and estimators which minimise the worst case risk ("minimax"). These will not be discussed in this course.[55]

### 3.3.4 Hypothesis testing

The idea of hypothesis testing is to use the data to discriminate between possible alternatives, usually written as $H_0$ and $H_1$. Tests are used to establish whether or not there is sufficient

[55] These concepts are covered in detail by Chapters 3, 4 and 5 (respectively) of [9].

*statistical* evidence against a hypothesis. These hypotheses are usually called the "null" and "alternative" hypothesis respectively and can be quite general statements. We will restrict our attention to the case where $H_0$ is the hypothesis that $\theta \in \Theta_0$ and $H_1$ is the hypothesis that $\theta \in \Theta_1$ where $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$.

A *test* is a (measurable) function $\varphi : \mathcal{X} \to [0,1]$. The decision between $H_0$ and $H_1$ may be based on auxillary randomisation: given $X = x$, $\varphi(x) \in [0,1]$, is the probability of rejecting $H_0$. A *non-randomised test* is a test with range $\{0,1\}$.

When designing a test, there are two types of possible error we need to consider:

- Type I error: we reject the null hypothesis when it is true
- Type II error: we fail to reject the null hypothesis when it is false

We typically require that tests have a certain *level*: that, when the hypothesis being tested is true, the expected value of the test does not exceed a certain, pre-specified value $\alpha$. To put it another way, we design our test so that the Type I error cannot exceed $\alpha$. The *size* of a test is the greatest probability of rejection of $H_0$, when $H_0$ is true. In symbols, the size of test $\varphi$ is $\sup_{\theta \in \Theta_0} \mathbb{E}_{P_\theta} \varphi(X) = \sup_{\theta \in \Theta_0} \beta(\theta)$.

Provided this condition is satisfied, we want a test with good *power*. Power is 1 - Type II error, i.e. the probability that we *do* reject the null hypothesis when it is false. Mathematically this is equal to the expected value of the test when the hypothesis being tested is false. The *power function* of a test describes its probability of rejection, given $\theta$: $\beta(\theta) := \mathbb{E}_{P_\theta} \varphi(X)$.

Example 3.14: Suppose we observe $X \sim \mathcal{N}(\mu, \sigma^2)$ where $\sigma$ is known and we want to test $H_0 : \mu = a$. Lets design a test of level $\alpha$ and compute its power.

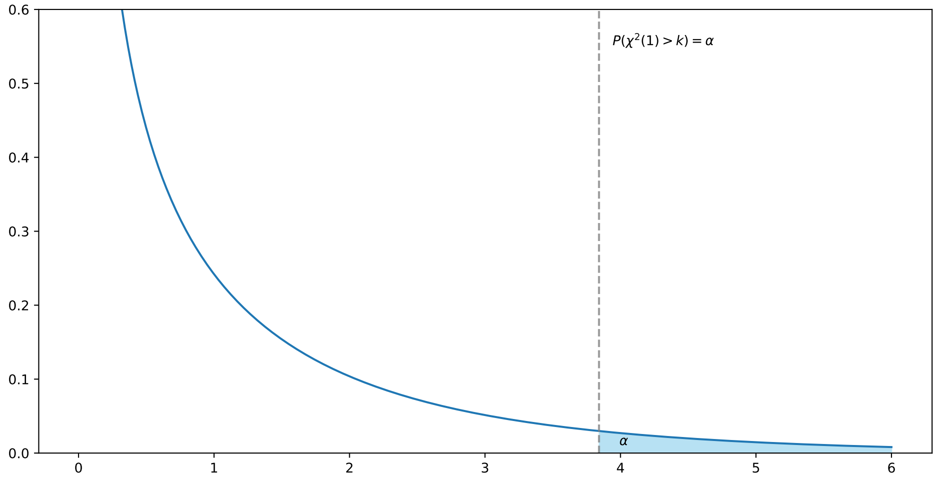
Let $T = (X - a)^2/\sigma^2$. We will set our test equal to $\phi(X) = \mathbf{1}\{T > k_\alpha\}$ for a *critical value* $k_\alpha$ chosen such that the test is of level $\alpha$.

Using properties of the Normal distribution, we have that, under $H_0$ (i.e. if $\mu = a$) then

$$T^{1/2} = (X - a)/\sigma \sim \mathcal{N}(0,1).$$

Hence $T \sim \chi^2(1)$. We will take $k_\alpha$ equal to the $1 - \alpha$ quantile of the $\chi^2(1)$ distribution. This is exactly the number such that $P(T > k_\alpha) = \alpha$ if $T \sim \chi^2(1)$.

Figure 16: $1-\alpha$ quantile of $\chi^2(1)$



If $\mu \neq a$, then we have

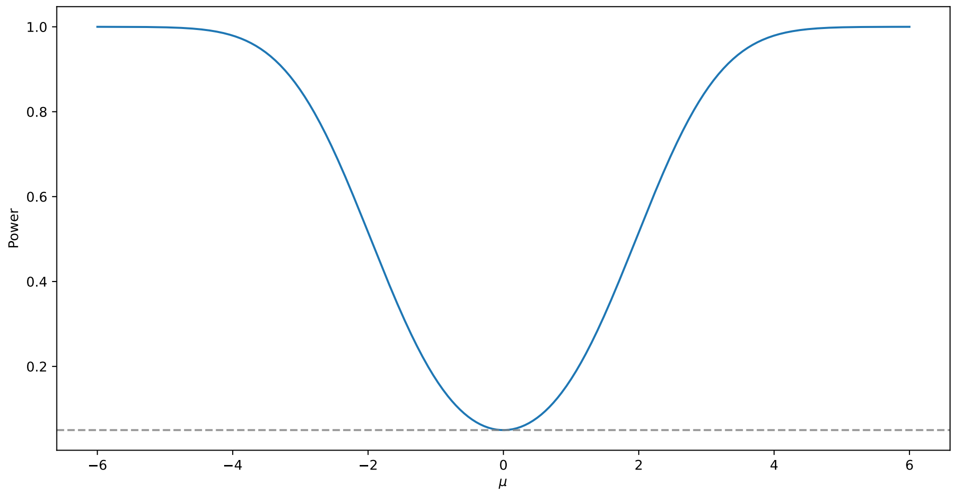
$$T^{1/2} = (X-a)/\sigma \sim \mathcal{N}(d,1), \quad d = (\mu - a)/\sigma,$$

and $T$ has a *non-central* $\chi^2$ *distribution with 1 degree of freedom and non-centrality parameter* $d^2$*:* $T \sim \chi^2(1, d^2)$. This is defined exactly as the distribution of the square of a Normal random variable with variance 1 and mean $d$. Therefore the power of our test $\phi$ is

$$\mathbb{E}_\mu \phi = P_\mu(T > k_\alpha) = P(\chi^2(1, d^2) > k_\alpha).$$

Lets plot this power curve supposing that $a = 0$ and $\sigma = 1$.

Figure 17: Power of $\phi$



The same test $\phi$ is equivalently defined by $\phi(X) = \mathbf{1}\{|T^{1/2}| > c_\alpha\}$ where the critical value $c_\alpha$ is the $1 - \alpha/2$ quantile of the standard normal distribution. △

The preceding example shows a general principle of test design: we base the test on the value of a *test statistic*, $T$, which will be large when $H_0$ is not true and small when it is.

Typically we want to use a test which has the greatest power when $H_1$ is true (i.e. $\theta \in \Theta$) subject to the requirement that the size is bounded above by some (pre-chosen) *significance level* $\alpha \in [0, 1]$; the latter requirement is described as the test having *level* $\alpha$.

If we have a sufficient statistic for our model, we may restrict attention to tests based on this sufficient statistic.

PROPOSITION 3.3: *If $T = T(X)$ is sufficient for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and $\phi = \phi(X)$ is a test, then*

$$\varphi(T) \coloneqq \mathbb{E}\left[\phi(X)|T\right]$$

*is a test with the same power function as $\phi$:*

$$\mathbb{E}_{P_\theta}\, \varphi(T) = \mathbb{E}_{P_\theta}\, \phi(X), \qquad \textit{for all } \theta \in \Theta.$$

*Proof.* Apply the law of iterated expectations:

$$\mathbb{E}_{P_\theta}\, \varphi(T) = \mathbb{E}_{P_\theta}\, \mathbb{E}\left[\phi(X)|T\right] = \mathbb{E}_{P_\theta}\, \phi(X).$$

□

### 3.3.5 Confidence sets

Suppose we are interested in some function $g$ of $\theta$. Rather than testing whether $g(\theta)$ is equal to specific value, we may form a *confidence set* for $g(\theta)$. A $1 - \alpha$ confidence set for $g(\theta)$ is a (random) set $S = S(X)$ such that $P_\theta\left(g(\theta) \in S(X)\right) \geq 1 - \alpha$ for all $\theta \in \Theta$.

There is a tight connection between confidence sets and tests. Suppose that $A(g_0) \coloneqq \{x : \varphi_{g_0}(x) = 1\}$, where $\varphi_{g_0}$ is a non-randomised level $\alpha$ test of $H_0 : g(\theta) = g_0$ against $H_1 : g(\theta) \neq g_0$ and define

$$S(x) \coloneqq \{g : x \notin A(g)\}. \tag{20}$$

It follows from the fact that the test is level $\alpha$ that

$$P_\theta\left(g(\theta) \in S(X)\right) = P_\theta\left(X \notin A(g(\theta))\right) = 1 - P_\theta\left(X \in A(g(\theta))\right) \geq 1 - \alpha.$$

Conversely we can construct level $\alpha$ tests from a confidence set. Define

$$\varphi(x) \coloneqq \begin{cases} 1 & \text{when } g_0 \notin S(x) \\ 0 & \text{otherwise} \end{cases}. \tag{21}$$

Then, for any $\theta$ such that $g(\theta) = g_0$,

$$\mathbb{E}_{P_\theta} \varphi = P_\theta\left(g_0 \notin S(X)\right) = P_\theta\left(g(\theta_0) \notin S(X)\right) \leq \alpha,$$

i.e. the test has level $\alpha$ for testing $H_0 : g(\theta) = g_0$ against $H_1 : g(\theta) \neq g_0$.

Similar constructions can be given in the case of (one-sided) confidence bounds for real valued parameters. See e.g. [8, pp. 229 – 231] or [10, Section 3.5].

Example 3.15: As in Example 3.14 suppose we are interested in the mean of $X \sim \mathcal{N}(\mu, \sigma^2)$ with known $\sigma^2$. In particular, lets suppose we want to construct a 1-$\alpha$ confidence set for $\mu$. Based on Example 3.14 & (20), the set $S(X) = \{a : \phi_a(X) = 0\}$ is a 1-$\alpha$ confidence set for $\mu$. Lets describe this set in a more straightforward way. Note that

$$\phi_a(X) = 0 \quad \iff \quad T \leq k_\alpha \quad \iff \quad T^{1/2} \leq k_\alpha^{1/2} \text{ or } -T^{1/2} \leq k_\alpha^{1/2}$$

Here we note that $c_\alpha := k_\alpha^{1/2}$ is the 1-$\alpha/2$ quantile of a standard normal random variable [Exercise]. Since $T^{1/2} = \frac{(X-a)}{\sigma}$ this gives our set as [Exercise]

$$S(X) = [X - \sigma c_\alpha,\, X + \sigma c_\alpha]. \qquad \triangle$$

## 3.4 Asymptotic statistics

Thus far we have discussed results which are "exact": we have considered examples in which we are able to explicitly derive estimators with certain properties (e.g. unbiasedness) or tests with certain size properties for any given sample size $n$. That this has been possible in these examples is, however, the exception rather than the rule.

To get around this, statisticians often rely on "approximations". These are often "asymptotic": the approximations used are ones in which we allow the sample size, $n$, to "go to infinity" and study the limiting behaviour of some statistical procedure in such a thought experiment.

The idea here is that the behaviour of in the limit can often to provide a good approximation to the finite-sample case of interest. The quality of this approximation can sometimes be explicitly quantified by an explicit bound.[56]

More often however, asymptotic analysis of procedures proceeds based on limit theorems: we know that as $n \to \infty$ some sequence of objects has a limit in some precise sense. However, for a given $n$, we do not know how far the $n$-th term in the sequence is from it's limit, only that this distance converges to 0 as $n \to \infty$. In practice, such limit results are combined with *Monte Carlo* simulations for some underlying distributions of the data which are deemed plausible: data are drawn from this distribution many times (independently) and used to evaluate the proposed procedure. If the performance in the Monte Carlo simulation is close enough to that which the limiting thought experiment would suggest, we proceed using the asymptotic approximation. This step is important: limit theorems describe the behaviour of a *sequence* of objects, but say nothing about the behaviour of a single object in that sequence.

[56] Such results are nowadays very common in the statistical analysis of procedures for "high-dimensional data". They are often called "non-asymptotic results" because they hold for all $n$ (or all $n$ beyond a certain point).

### 3.4.1 Consistency

Consider a sequence of estimators $(\hat{\delta}_n)_{n\in\mathbb{N}}$ for a parameter $g(\theta)$. Often one has in mind a situation where the $n$-th element of this sequence is a statistic based on $n$ observations $X_1, \ldots, X_n$: $\hat{\delta}_n = \hat{\delta}_n(X_1, \ldots, X_n)$.

A desirable property of such a sequence of estimators is *consistency*. We say that $(\hat{\delta}_n)_{n\in\mathbb{N}}$ is *consistent* for $g(\theta)$ if $\hat{\delta}_n \xrightarrow{P_\theta} g(\theta)$. In many cases the dependence on $\theta$ is omitted and one writes $\hat{\delta}_n \xrightarrow{P} g(\theta)$ where it is understood that $\theta$ is the parameter corresponding to the distribution $P$ (of the $X$).

Example 3.16 [Consistency of the sample mean]: Suppose that $X_1, X_2, \ldots$ are an i.i.d. sequence of random variables with $\mathbb{E}\,|X_1| < \infty$ and $\mathbb{E}\,X_1 = \mu$. Then by the weak law of large numbers (Theorem 2.15)

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu. \qquad \triangle$$

A sufficient condition for consistency which is often useful is the following.

Proposition 3.4: *If $(\hat{\delta}_n)_{n\in\mathbb{N}}$ is a sequence of estimators such that $b(\theta) := \mathbb{E}_{P_\theta}\,\hat{\delta}_n - g(\theta) \to 0$ and $\operatorname{Var}_{P_\theta}(\hat{\delta}_n) \to 0$ as $n \to \infty$ for each $\theta \in \Theta$, then $\hat{\delta}_n \xrightarrow{P_\theta} g(\theta)$.*

*Proof.* Exercise. □

### 3.4.2 Asymptotic Normality

Often (but not always) sensible estimators are *asymptotically normal* in the following sense: if $P_\theta$ is the distribution of the data, then

$$\sqrt{n}(\hat{\delta}_n - g(\theta)) \rightsquigarrow \mathcal{N}(0, V). \tag{22}$$

This is a desirable property for various reasons: firstly, under conditions we will not go into, for a large class of models a limiting normal distribution of this form, with the smallest possible variance matrix $V$, is – in a certain sense – the best possible limiting distribution. One interpretation of this result is the following: in large samples, the estimator $\hat{\delta}_n$ is a reasonable estimator in that there is no alternative estimator $\hat{\beta}_n$ which is (asymptotically) superior.[57] If we have two estimators which are asymptotically normal with variance matrices $V_1$ and $V_2$, we should prefer the estimator with the smaller variance (in the positive semi-definite sense).

A second reason a property like (22) is desirable is that it makes it straightforward to perform (approximate) *uncertainty quantification*. In practice, what this usually means is: it is easy to provide confidence sets for $g(\theta)$ which are approximately of level $1-\alpha$ (essentially) via the method in Example 3.15. We will see more on this below.

Example 3.17 [Asymptotic normality of the sample mean]: Suppose that $X_1, X_2, \ldots$ are an i.i.d. sequence of random variables with $\operatorname{Var}(X_1) = \sigma^2 < \infty$ and $\mathbb{E}\,X_1 = \mu$. Then by the central limit

[57] See e.g. Chapter 8 in [16] for a formal and detailed discussion of this point.

theorem (Theorem 2.17)

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \rightsquigarrow \mathcal{N}(0,\sigma^2). \qquad \triangle$$

### 3.4.3 Asymptotically valid tests & CS

When we estimate a parameter $g(\theta)$ we would like to have a sense of how accurate our estimator of that parameter is. If (22) holds, one way of doing this is to calculate or estimate the variance $V$. Under regularity conditions $V$ is approximately the variance of $\sqrt{n}(\hat{\delta}_n - g(\theta))$:

$$\sqrt{n}(\hat{\delta}_n - g(\theta)) \approx \mathcal{N}(0,V)$$

and so

$$\hat{\delta}_n - g(\theta) \approx \mathcal{N}(0,V/n).$$

As such we can use (a consistent estimator of) $V/n$ or (in the univariate case) $\sqrt{V}/\sqrt{n}$ as a measure of accuracy of our estimate. An estimator of $\sqrt{V}/\sqrt{n}$ is often called a "standard error".

Example 3.18 [Standard error of the mean]: Consider the setup of Example 3.17. Here $V = \sigma^2$. Whilst this is unknown, we can estimate it. If $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$, then $s_n^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2 \xrightarrow{P} \sigma^2$ [Exercise]. The standard error is then $\sqrt{s_n^2}/\sqrt{n}$. $\triangle$

If (22) holds and we can estimate $V$ consistently we can also perform hypothesis tests which are approximately valid.

We will first consider the "t-test" for a single coefficient $\theta_k$. The hypothesis we will be testing is $H_0 : \theta_k = b$, for some pre-specified $b \in \mathbb{R}$, against $H_1 : \theta_k \neq b$. We suppose that we have estiamtors $\hat{\theta}_n$ and $\hat{V}_n$ such that:

$$\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}(0,V), \qquad \hat{V}_n \xrightarrow{P} V. \tag{23}$$

The t-test is based on the t-statistic:

$$t_n = \frac{\sqrt{n}(\hat{\theta}_{n,k} - b)}{\sqrt{\hat{V}_{n,kk}}}. \tag{24}$$

By (23), our estimate of $\theta_k$, $\hat{\theta}_{n,k}$ and $\hat{V}_{n,kk}$ satisfy

$$\sqrt{n}(\hat{\theta}_{n,k} - \theta_k) \rightsquigarrow \mathcal{N}(0,V_{kk}), \qquad \hat{V}_{n,kk} \xrightarrow{P} V_{kk}. \tag{25}$$

Therefore, under the null hypothesis $H_0 : \theta_k = b$, we have (by Slutsky's Theorem) that

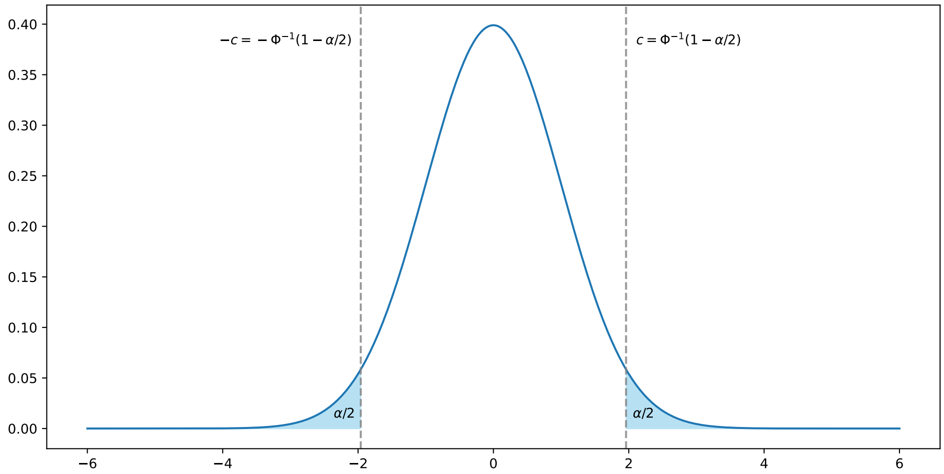
$$t_n \rightsquigarrow \mathcal{N}(0,1). \tag{26}$$

We perform a t-test by comparing $t_n$ to certain *critical values*. Recall from our discussion of hypothesis testing that we want our test to have a certain *level*: that is, we want it to reject $H_0$ no more than a pre-specified amount, $\alpha$, if $H_0$ is indeed true. We choose the critical values to do this (whilst simultaneously attempting to reject as much as possible if $H_0$ is *not* true). In particular to test $H_0$ against $H_1$ we will compare $|t_n|$ to the $1-\alpha/2$ quantile, $c = \Phi^{-1}(1-\alpha/2)$, of the standard normal distribution: we reject if

$$|t_n| > c. \tag{27}$$

The reason we choose this value is easily explained graphically:

Figure 18: $1 - \alpha/2$ quantiles



As depicted in the above figure, the quantile $k = \Phi^{-1}(1-\alpha/2)$ is the value such that for a standard normal random variable $Z$

$$1 - P(Z \leq c) = 1 - \alpha/2.$$

Therefore,

$$P(|Z| > c) = \alpha.$$

Combining the above with (26) and the fact that the standard normal CDF is continuous on $\mathbb{R}$

we have:[58]

$$\begin{aligned}\lim_{n\to\infty} P(|t_n| > c) &= \lim_{n\to\infty} P(t_n < -c) + \lim_{n\to\infty} P(t_n > c)\\ &= \lim_{n\to\infty} P(-t_n > c) + \lim_{n\to\infty} P(t_n > c)\\ &= P(Z > c) + P(Z > c)\\ &= 2\,[1 - P(Z \leq c)]\\ &= 2\,[1 - (1-\alpha/2)]\\ &= 2 \times \alpha/2\\ &= \alpha.\end{aligned}$$

since by (26) and Slutsky's Theorem also $-t_n \rightsquigarrow \mathcal{N}(0,1)$. That is, under $H_0$, asymptotically our test has probability $\alpha$ of rejecting $H_0$ when it is true. Suppose that $H_1$ were instead true: then we would have that $\theta_k = b - d$ for some $d \neq 0$ and so, using (25) we have

$$\sqrt{n}(\hat{\theta}_{n,k} - b) = \sqrt{n}(\hat{\theta}_{n,k} - \theta_k) + \sqrt{n}d \xrightarrow{P} \operatorname{sign}(d) \times \infty.$$

Here the term $\sqrt{n}d \to \operatorname{sign}(d) \times \infty$, which dominates the term $\sqrt{n}(\hat{\theta}_{n,k} - \theta_k) \rightsquigarrow \mathcal{N}(0, V_{kk})$. Using Slutsky's theorem again, in this case we have

$$t_n \xrightarrow{P} \operatorname{sign}(d) \times \infty.$$

As, if $H_1$ is true, our test will – asymptotically – reject $H_0$ with probability one:

$$\lim_{n\to\infty} P(|t_n| > c) = 1.$$

### The Wald test

We now consider a generalisation of the t-test – the Wald test – which allows us to test more general hypotheses.[59] The hypothesis we will consider will be for the null hypothesis: $H_0 : g(\theta) = 0$, for some smooth function $g : \mathbb{R}^K \to \mathbb{R}^p$, against the alternative hypothesis $H_1 : g(\theta) \neq 0$.

There are many possible tests that can be used for such a hypothesis.[60] Here we will focus on the *Wald test*. This is based on the *Wald statistic*:

$$W_n := ng(\hat{\theta}_n)' \left[G(\hat{\theta}_n)\hat{V}_n G(\hat{\theta}_n)'\right]^{-1} g(\hat{\theta}_n), \tag{28}$$

for $G := \nabla_\theta g(\theta)$, the Jacobian of $g$.

[58] Recall that convergence in distribution "$\rightsquigarrow$" is equivalent to the convergence of the CDF at all points at which the latter is continuous.

[59] Proposition 3.5 below allows us to recover all the results we just established for the $t$-test, since the t-statistic squared is a Wald statistic.

[60] E.g. see Chapter 17 of [8] for a brief discussion of the Likelihood ratio, Wald and Score tests in parametric models. Further details can be found in Chapter 12.4 of [10].

The idea behind the Wald test is simple: provided $g$ is continosly differentiable, $G$ is of full row rank and (23) holds, then by the delta method, Slutsky's theorem and the continuous mapping theorem, we have

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \rightsquigarrow \mathcal{N}(0, G(\theta)VG(\theta)'), \qquad G(\hat{\theta}_n)\hat{V}_n G(\hat{\theta}_n)' \xrightarrow{P} G(\theta)VG(\theta)'. \tag{29}$$

Under the null that $g(\theta) = 0$, the first part of this becomes $\sqrt{n}g(\hat{\theta}_n) \rightsquigarrow \mathcal{N}(0, G(\theta)VG(\theta)')$ and by the continuous mapping theorem and Slutsky's theorem we have $\sqrt{n}\left[G(\hat{\theta}_n)\hat{V}_n G(\hat{\theta}_n)'\right]^{-1/2} g(\hat{\theta}_n) \rightsquigarrow \mathcal{N}(0, I)$. The test statistic $W_n$ in (28) is the squared norm of this quantity, and so ought to be asymptotically distributed as a $\chi^2_p$ random variable if the null is true. As such, we form a test by rejecting the null whenever $W$ is "too large".

Proposition 3.5: *Suppose that (23) holds and $g$ is continuously differentiable with full row rank Jacobian $G$. Then, under the null $H_0 : g(\theta) = 0$,*

$$W_n \rightsquigarrow \chi^2_p,$$

*and the sequence of tests which reject when $W_n > k_\alpha$ for $k_\alpha$ the $1 - \alpha$ quantile of the $\chi^2_p$ distribution has asymptotic size $\alpha$:*

$$\lim_{n\to\infty} P(W_n > k_\alpha) = \alpha.$$

*Under any fixed alternative, $H_1 : g(\theta) = \mathfrak{g} \neq 0$, the test is consistent:*

$$\lim_{n\to\infty} P(W_n > k_\alpha) = 1.$$

*Proof.* By (23), the Delta method (Theorem 2.12), continuity of $G(\theta)$, the continuous mapping theorem (Theorem 2.10) and Slutsky's theorem (Theorem 2.11), (29) holds. Since $G(\theta)$ has full row rank and $V$ is full rank, it follows that $G(\theta)VG(\theta)'$ is full rank [Exercise] and hence invertible. Therefore, by the continuous mapping theorem and Slutsky's theorem we have

$$\left[G(\hat{\theta}_n)\hat{V}_n G(\hat{\theta}_n)'\right]^{-1/2} \sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \rightsquigarrow Z \sim \mathcal{N}(0, I_p).$$

First we handle the case under the null $g(\theta) = 0$. This condition implies that $Z_n := \sqrt{n}\left[G(\hat{\theta}_n)\hat{V}_n G(\hat{\theta}_n)'\right]^{-1/2} g(\hat{\theta}_n) \rightsquigarrow Z$. As $x \mapsto x'x$ is continuous, applying the continuous mapping theorem we obtain [Exercise]

$$W_n = Z_n' Z_n \rightsquigarrow Z'Z \sim \chi^2_p.$$

Moreover, since $W_n$ is a random variable, weak convergence is equivalent to convergence of its CDF at all continuity points of the limiting CDF. Here the limit is $\chi^2_p$ which has a continuous CDF on $\mathbb{R}$. Therefore, by the definition of $k_\alpha$,

$$P(W_n > k_\alpha) = 1 - P(W_n \leq k_\alpha) \to 1 - P(Z'Z \leq k_\alpha) = 1 - (1 - \alpha) = \alpha.$$

It remains to consider the case under the fixed alternative $g(\theta) = \mathfrak{g} \neq 0$. Here we note that letting $M_n := G(\hat{\theta}_n)\hat{V}_n G(\hat{\theta}_n)'$ we have

$$W_n = \|M_n^{-1/2}\sqrt{n}(g(\hat{\theta}_n) - \mathfrak{g}) + M_n^{-1/2}\sqrt{n}\mathfrak{g}\|^2.$$

Let $Z_n := M_n^{-1/2}\sqrt{n}(g(\hat{\theta}_n) - \mathfrak{g}) \rightsquigarrow \mathcal{N}(0, I)$. Since $Z_n \rightsquigarrow \mathcal{N}(0, I)$, for any $\varepsilon > 0$ there some $0 < K < \infty$ such that $P(\|Z_n\| \leq K) \geq 1 - \varepsilon$ for each $n \in \mathbb{N}$ [Exercise]. By the spectral decomposition, $\|M_n^{-1/2}v\| \geq \lambda_{\min}(M_n^{-1/2})\|v\|$ for any $v \in \mathbb{R}^K$ [Exercise]. Since $M_n^{-1/2} \xrightarrow{P} M^{-1/2}$ which is positive definite, it follows that for some $\delta > 0$, $P(\|M_n^{-1/2}\sqrt{n}\mathfrak{g}\| \geq \delta\sqrt{n}\|\mathfrak{g}\|) \to 1$.

Using these two observations with the reverse triangle inequality, it follows that

$$W_n^{1/2} = \||Z_n + M_n^{-1/2}\sqrt{n}\mathfrak{g}\| \geq \left|\|M_n^{-1/2}\sqrt{n}\mathfrak{g}\| - \|Z_n\|\right|,$$

and so $W_n \xrightarrow{P} \infty$ in the sense that for any $K > 0$

$$\lim_{n\to\infty} P(W_n > K) = 1.$$

Taking $K = k_\alpha$ completes the proof. □

Example 3.19 [Linear restriction]: Suppose that $g(\theta) = R\theta - r$ for some $p \times K$ matrix $R$ of full row rank and some $r \in \mathbb{R}^p$. Then, $\nabla_\theta g(\theta) = G(\theta) = R$ which is trivially continuous and has full row rank by assumption. The Wald statistic becomes

$$W_n = n(R\hat{\theta}_n - r)' \left[R\hat{V}_n R'\right]^{-1} (R\hat{\theta}_n - r),$$

and provided (23) holds, Proposition 3.5 applies. △

Example 3.20: Suppose that $g(\theta) = \theta_1 - 1$. We can write this as $\theta_1 = R\theta - 1$, for $R = (1, 0, \ldots, 0)$, a $1 \times K$ matrix with full row rank. This is therefore a special case of example 3.19 and the Wald statistic is

$$W_n = n(R\hat{\theta}_n - 1)' \left[R\hat{V}_n R'\right]^{-1} (R\hat{\theta}_n - 1) = \frac{n(\hat{\theta}_{n,1} - 1)^2}{\hat{V}_{n,11}},$$

and provided (3.5) holds, Proposition 3.5 applies. △

Example 3.21: Suppose that $g(\theta) = \theta_1^2 - 1$. This has Jacobian $G(\theta) = (2\theta_1, 0, \ldots, 0)$, which is continuous and of full row rank provided $\theta_1 \neq 0$. The Wald statistic is

$$W_n = n[(\hat{\theta}_{n,1})^2 - 1] \left[(2\hat{\theta}_{n,1}, 0, \ldots, 0)\hat{V}_n(2\hat{\theta}_{n,1}, 0, \ldots, 0)'\right]^{-1} [(\hat{\theta}_{n,1})^2 - 1] = \frac{n[(\hat{\theta}_{n,1})^2 - 1]^2}{4(\hat{\theta}_{n,1})^2\hat{V}_{n,11}},$$

and provided $\theta_1 \neq 0$ and (3.5) holds, Proposition 3.5 applies. △

Remark 3.5: *Note that the hypothesis being tested in Examples 3.20 and 3.21 is the same:*

*$H_0 : \theta_1 = 1$ against $H_1 : \theta_1 \neq 1$. Nevertheless the Wald statistic takes a different value in general and may lead to different conclusions in a given data set. This is refered to as* non-invariance to a nonlinear re-parametrisation of the hypothesis.

*There are many alternatives to the Wald test. The two most common are the likelihood ratio test and score test, both of which* are *invariant to such re-parametrisations.*

As we have already discussed, a *confidence set* for $g(\theta)$ of level $1-\alpha$ is a random set, $S(X)$, such that

$$P(g(\theta) \in S(X)) \geq 1 - \alpha.$$

In pratice, tests satisfying this requirement in finite sample are often difficult to come by and we usually adopt an asymptotic version: a sequence $(S_n)_{n\in\mathbb{N}}$ of confidence sets is asymptotically of level $1-\alpha$ if

$$\lim_{n\to\infty} P(g(\theta) \in S_n) \geq 1 - \alpha. \tag{30}$$

As in the finite $n$ case, we can form an asymptotically valid confidence set for $g(\theta)$ based on a test. Here we use Proposition 3.5.

Corollary 3.4: *Suppose that the conditions of Proposition 3.5 hold and define functions $h_\gamma(\theta) := g(\theta) - \gamma$. Let $W_n^\gamma$ be a Wald statistic for the test $H_0 : h_\gamma(\theta) = 0$ against $H_1 : h_\gamma(\theta) \neq 0$ and let*

$$S_n := \{\gamma \in \mathbb{R}^p : W_n^\gamma \leq k_\alpha\}.$$

*Then (30) holds with equality.*

*Proof.* Let $\gamma = g(\theta)$ and write

$$P(g(\theta) \in S_n) = P(\gamma \in S_n) = P(W_n^\gamma \leq k_\alpha) = 1 - P(W_n^\gamma > k_\alpha).$$

Under $H_0 : h_\gamma(\theta) = 0$, $P(W_n^\gamma > k_\alpha) \to \alpha$ by Proposition 3.5 and so $P(g(\theta) \in S_n) \to 1 - \alpha$. □

The most common frequently encountered type of confidence set is a *confidence interval* for a particular element $\theta_k$, of interest. In this setting the construction of the confidence sets of the preceding Corollary simplifies.

Example 3.22 [Confidence interval for $\theta_k$]: Suppose that $g(\theta) := \theta_k$ and we have

$$\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}(0, V),$$

and $V_n \xrightarrow{P} V$. Then, the Wald statistic for $h_\gamma(\theta) := \theta_k - \gamma = 0$ against $h_\gamma(\theta) \neq 0$ has the form

$$W_n^\gamma = \frac{n(\hat{\theta}_{n,k} - \gamma)^2}{\hat{V}_{n,kk}},$$

(cf. Example 3.20). Then $W_n^\gamma \leq k_\alpha$ iff $|\hat{\theta}_{n,k} - \gamma| \leq \sqrt{k_\alpha}\hat{V}_{n,kk}^{1/2}/\sqrt{n}$ iff $\gamma \leq \hat{\theta}_{n,k} + \sqrt{k_\alpha}\hat{V}_{n,kk}^{1/2}/\sqrt{n}$

*or* $\gamma \geq \hat{\theta}_{n,k} - \sqrt{k_\alpha}\hat{V}_{n,kk}^{1/2}/\sqrt{n}$. This is usually written as the *confidence interval*:

$$S_n = \left[\hat{\theta}_{n,k} - \sqrt{k_\alpha}\hat{V}_{n,kk}^{1/2}/\sqrt{n},\ \ \hat{\theta}_{n,k} + \sqrt{k_\alpha}\hat{V}_{n,kk}^{1/2}/\sqrt{n}\right].$$

The quantity $\hat{V}_{n,kk}^{1/2}/\sqrt{n}$ is the (estimated) *standard error* of $\hat{\theta}_{n,k}$. △

### 3.4.4 Example: Maximum likelihood estimators

One very well known method of contructing estimators is the method of maximum likelihood. Suppose that we have a model $\{P_\theta : \theta \in \Theta\}$ and let $p_\theta$ be the density (or mass) function corresponding to $P_\theta$. If the elements of $X^{(n)} = (X_1, X_2 \ldots, X_n)$ are drawn i.i.d with each $X_i$ having distribution $P_\theta$ then the joint density of $X^{(n)}$ is the product $p_\theta^n(x^{(n)}) = p_\theta(x_1) \times \cdots \times p_\theta(x_n)$. $L(\theta) := p_\theta^n(X^{(n)})$ is called the *likelihood function.*

The maximum likelihood estimator, $\hat{\theta}_n$, is formed by maximising $L(\theta)$ over $\Theta$:

$$\hat{\theta}_n := \operatorname{argmax}_{\theta\in\Theta} L(\theta). \tag{31}$$

It is equivalent, but often convenient, to work instead with the log-likelihood function: $l(\theta) := \log L(\theta)$. We have [Exercise]

$$\hat{\theta}_n := \operatorname{argmax}_{\theta\in\Theta} l(\theta). \tag{32}$$

When the family $\{P_\theta : \theta \in \Theta\}$ is "well-behaved" maximum likelihood estimators are consistent and asymptotically normal. Additionally their asymptotic variance is the Cramér – Rao lower bound, $I(\theta)^{-1}$ (Cf. Corollary 3.2).[61,62]

Theorem 3.8 [Consistency of MLE]: *If $X_1, X_2, \ldots$ is an i.i.d. sequence of random vectors, each with pdf / pmf $p_\theta(x)$ and*

*(i)* $\theta \neq \theta' \implies p_\theta(x) \neq p_{\theta'}(x)$

*(ii)* $\theta \in \Theta \subset \mathbb{R}^K$ *and* $\Theta$ *is compact*

*(iii)* $\log p_\theta(x)$ *is continuous in* $\theta$ *for all* $x$ *in a set* $\mathcal{X}$ *with* $P_\theta(X \in \mathcal{X}) = 1$

*(iv)* $\mathbb{E}_{P_\theta}\left[\sup_{\theta\in\Theta} |\log p_\theta(X)|\right] < \infty$

*Then, for $\hat{\theta}_n$ as defined in* (31) *(or, equivalently,* (32)*)*

$$\hat{\theta}_n \xrightarrow{P_\theta} \theta.$$

**Proof:* E.g. Theorem 2.5 in [12]. □

[61] Note however that (i) this is attained only asymptotically and (ii) maximum likelihood estimators are only asymptotically unbiased and may have a bias in finite samples.

[62] Many of the conditions in Theorems 3.8 and 3.9 can be weakened at the expense of more complex argumentation.

Remark 3.6: *Condition (i) in Theorem 3.8 requires the model to be* identifiable. *Identifiability is a* necesary *condition for consistency [Exercise].*

Theorem 3.9 [Asymptotic normality of MLE]: *Suppose that the hypotheses of Theorem 3.8 are satisfied and additionally,*

*(i)* $\theta$ *is in the interior of* $\Theta$

*(ii)* $p_\theta(x)$ *is twice continuously differentiable in* $\theta$ *and* $p_{\theta'}(x) > 0$ *for all* $\theta'$ *in an open set* $\mathscr{N}$ *around* $\theta$

*(iii)* $\int \sup_{\theta \in \mathscr{N}} \|\nabla_\theta p_\theta(x)\| \, \mathrm{d}x < \infty$, $\int \sup_{\theta \in \mathscr{N}} \|\nabla_{\theta\theta} p_\theta(x)\| \, \mathrm{d}x < \infty$

*(iv) If* $\ell_\theta = \nabla_\theta \log p_\theta$, $I(\theta) = \mathbb{E}_{P_\theta}\left[\ell_\theta(X)\ell_\theta(X)'\right]$ *exists and is positive definite*

*(v)* $\mathbb{E}[\sup_{\theta \in \mathscr{N}} \|\nabla_\theta \ell_\theta(X)\|] < \infty$

*Then,*

$$\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}(0, I(\theta)^{-1}).$$

*Proof.* E.g. Theorem 3.3 in [12]. □

In some cases, MLEs can be derived by hand and consistency and asymptotic normality of MLEs can be established directly via LLNs and CLTs rather than Theorems like 3.8 and 3.9.

Example 3.23 [MLE for Poisson distribution]: Suppose that $X_1, X_2, \ldots$ are i.i.d. with $X_i \sim \text{Poisson}(\lambda)$. The log-likelihood function is

$$l(\lambda) = \sum_{i=1}^n \log(\lambda^{X_i} \exp(-\lambda)/(X_i!)) = -n\lambda - \sum_{i=1}^n \log(X_i!) + \log(\lambda) \sum_{i=1}^n X_i.$$

Setting the first derivative of $l$ to zero gives

$$0 = l'(\lambda) = -n + \frac{1}{\lambda}\sum_{i=1}^n X_i \quad \implies \quad \hat{\lambda}_n = \frac{1}{n}\sum_{i=1}^n X_i.$$

This is a maximum as

$$l''(\lambda) = -\frac{1}{\lambda^2}\sum_{i=1}^n X_i < 0.$$

Since the Poisson distribution has mean $\lambda$ [Exercise] and variance $\lambda < \infty$ one has

$$\sqrt{n}(\hat{\lambda}_n - \lambda) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \lambda) \rightsquigarrow \mathcal{N}(0, \lambda)$$

by the CLT. Note that this implies that $\hat{\lambda}_n$ is consistent for $\lambda$ [Exercise]. △

In many cases however this is not possible (or, at least, not easy) and $\hat{\theta}_n$ must be found via numerical optimisation. See the exercises for an example. We will see further examples of this later in the course.

Our final result in this section concerns what is called the *invariance* property of MLEs. Suppose that our interest is in $g(\theta)$ rather than $\theta$. What is the MLE of $g(\theta)$? To get around technical difficulties which may occur since one value $\eta = g(\theta)$ may correspond to multiple $\theta$ (if $g$ is not injective) we assume all the maxima below exist and define the induced likelihood function as

$$L^*(\eta) := \max_{\theta \in \{\theta : g(\theta) = \eta\}} L(\theta).$$

We define the MLE of $g(\theta)$ as

$$\hat{\eta} = \operatorname*{arg\,max}_{\eta \in \{g(\theta) : \theta \in \Theta\}} L^*(\eta).$$

The following Lemma shows why this is a sensible definition.

LEMMA 3.2: $\max_{\eta \in \{g(\theta):\theta \in \Theta\}} L^*(\eta) = \max_{\theta \in \Theta} L(\theta)$.

*Proof.* Exercise. □

THEOREM 3.10: *If $\hat{\theta}$ is the MLE of $\theta$ then $g(\hat{\theta})$ is the MLE of $g(\theta)$.*

*Proof.* It suffices to show that $L^*(\hat{\eta}) = L^*(g(\hat{\theta}))$. By Lemma 3.2 and the fact that the iterated maximisation is the same as maximisation over $\Theta$ in one step,

$$L^*(\hat{\eta}) = \max_{\eta \in \{g(\theta):\theta \in \Theta\}} \max_{\theta \in \{\theta \in \Theta : g(\theta) = \eta\}} L(\theta) = \max_{\theta \in \Theta} L(\theta) = L(\hat{\theta}).$$

Additionally by the definition of $L^*$,

$$L(\hat{\theta}) = \max_{\theta : g(\hat{\theta}) = g(\theta)} L(\theta) = L^*(g(\hat{\theta})),$$

where the first equality follows because $\hat{\theta}$ is the MLE. □

# 4 Regression

The (linear) regression model is a workhorse model with wide ranging applications in applied statistics. In matrix notation the basic model is

$$y = X\beta + \epsilon, \tag{33}$$

where $y = (y_1, \dots, y_n)' \in \mathbb{R}^n$ is the *response* or *outcome*, $X = (X_1, \dots X_n)' \in \mathbb{R}^{n \times K}$ are *explanatory variables*, *regressors* or *covariates* (with each $X_i \in \mathbb{R}^K$ $[i = 1, \dots, n]$) and $\epsilon = (\epsilon_1, \dots, \epsilon_n)' \in \mathbb{R}^n$ the *error term.* $\beta \in \mathbb{R}^K$ is a vector of (unknown) *coefficients.* We observe $y$ and $X$, but not $\epsilon$ or $\beta$.

Lets consider some examples of how this model can be used to answer various questions.

Example 4.1: The coefficient $\beta_k$ in model (33) describes the effect of a unit increase in the corresponding covariate $X_k$ on the response $y$, holding all other covariates fixed.

For a concrete example: suppose we were interested in the effect of years of education on wages. We might fit the model:

$$\log(\mathit{wage}) = \beta_0 + \beta_1 \times \mathit{education} + \beta_2 \times \mathit{experience} + \beta_3 \times \mathit{experience}^2,$$

where *education* is the number of years of education and *experience* the number of years of work experience.The coefficient $\beta_1$ then measures how much the (log) wage increases for a one year increase in education.

How to *estimate* this parameter (and the other $\beta_j$'s) is the subject of the next section. Later on in the course we will examine how to perform *hypothesis tests* on these parameters. For instance, we might want to statistically test whether the effect of additional years of education on wages is positive. △

Example 4.2: Models such as (33) can also be used for *forecasting* or *prediction.* If we have an estimate $\hat{\beta}$ of the coefficient $\beta$ and a new observation $X_i$, then we can predict the corresponding $y_i$ value as

$$\hat{y}_i = X_i\hat{\beta}.$$

Consider the same model of log wages as in example 4.1. If we have an estimate $\hat{\beta}$ of $\beta$ and data on a new individuals' years of education and experience, we can predict their (log) wage as:

$$\widehat{\log(\mathit{wage})} = \hat{\beta}_0 + \hat{\beta}_1 \times \mathit{education} + \hat{\beta}_2 \times \mathit{experience} + \hat{\beta}_3 \times \mathit{experience}^2. \qquad \triangle$$

We will come back to this example throughout section 4 to illustrate the application of the methods we cover. There are also further examples in the course exercises.

### 4.1 Least squares estimation

The coefficients $\beta$ in (33) are unknown and need to be estimated. The classical estimate of $\beta$ is the "(ordinary) least squares" or "OLS" estimate, $\hat{\beta}$, given by

$$\hat{\beta} := (X'X)^{-1}X'y, \tag{34}$$

provided the inverse in the preceding display exists. This estimator is called least squares for the following reason: it minimises the "sum of the squared residuals". In particular, let $e(\beta) := y - X\beta$ be the residuals given the parameter value $\beta$. Then $\hat{\beta}$ minimises the criterion:

$$S(\beta) := \|y - X\beta\|^2 = \sum_{i=1}^{n} \left(y_i - X_i'\beta_i\right)^2 = \sum_{i=1}^{n} e(\beta)_i^2. \tag{35}$$

Proposition 4.1: *If $X$ has full column rank, then $\hat{\beta}$ in (34) satisfies*

$$\hat{\beta} = \operatorname{argmin}_{\beta \in \mathbb{R}^K} S(\beta),$$

*where $S(\beta)$ is as in (35).*

*Proof.* The assumptions ensure that the matrix $X'X$ is of full rank [Exercise] and hence $(X'X)^{-1}$ exists. The space $C := \{Xb : b \in \mathbb{R}^K\}$ is the linear span of the columns of $X$ and is a linear subspace of $\mathbb{R}^n$. [Exercise]. By the Projection Theorem (Theorem 1.4), there is a unique element $\hat{y} = X\hat{\beta} \in C$ which minimises $Z \mapsto \|y - Z\|$ over $C$ and this element satisfies $(y - X\hat{\beta})'X\beta = 0$ for all $\beta \in \mathbb{R}^K$. This condition implies that $(y - X\hat{\beta})'X = 0$ [Exercise] and hence $y'X = \hat{\beta}X'X$. Transpose and premultiply by $(X'X)^{-1}$ to obtain (34). □

Example 4.3: In the log wage example 4.1, using data from the US current population survey (1985), OLS gives the following estimated model:

$$\widehat{\log(wage)} = 0.5203 + 0.0898 \times \mathit{education} + 0.0349 \times \mathit{experience} - 0.0005 \times \mathit{experience}^2. \quad \triangle$$

#### Projection matrices

The proof of Proposition 4.1 just given relied on the characterisation of $X\hat{\beta}$ as an *orthogonal projection*.[63] In general, given a $n \times K$ matrix $Z$ of full column rank, the $n \times n$ matrices

$$P_Z := Z(Z'Z)^{-1}Z' \quad \text{and} \quad M_Z := I_n - P_Z, \tag{36}$$

project vectors in $\mathbb{R}^n$ to the column space of $X$, $C := \operatorname{col}(Z) := \{Zb : b \in \mathbb{R}^K\}$ and its orthogonal complement, $C^\perp = \ker(Z') := \{c \in \mathbb{R}^n : Z'c = 0\}$, respectively.[64]

These *projection matrices* have a number of useful properties.

[63] Other proofs are of course possible; see the exercises.

[64] $\ker(A) := \{x : Ax = 0\}$ is the *kernel* or *nullspace* of $A$.

Lemma 4.1: *Suppose $Z$ is a $n \times K$ matrix of full column rank and define $P_Z$ and $M_Z$ as in* (36)*. Then* $\ker(Z') = \operatorname{col}(Z)^{\perp}$ *and*

*(i)* *If $x \in \operatorname{col}(Z)$, then $P_Z x = x$ and $M_Z x = 0$,*

*(ii)* *If $x \in \ker(Z')$, then $P_Z x = 0$ and $M_Z x = x$,*

*(iii)* *If $x \in \mathbb{R}^n$, $x = P_Z x + M_Z x$,*

*(iv)* *$P_Z$ and $M_Z$ are symmetric, i.e. $P_Z' = P_Z$ and $M_Z' = M_Z$,*

*(v)* *$P_Z$ and $M_Z$ are idempotent, i.e. $P_Z P_Z = P_Z$ and $M_Z M_Z = M_Z$,*

*(vi)* *$P_Z M_Z = 0$,*

*(vii)* $\operatorname{rank}(P_Z) = K$ *and* $\operatorname{rank}(M_Z) = n - K$.

*Proof.* Exercise. □

Returning to linear regression, applying $P_X$ to $y$ yields the *fitted values*, $\hat{y} := X\hat{\beta}$:

$$P_X y = X(X'X)^{-1}X'y = X\hat{\beta}, \tag{37}$$

whilst applying $M_X$ to $y$ yields the *residuals*, $\hat{\epsilon} := y - X\hat{\beta}$:

$$M_X y = (I_n - P_X)y = y - X\hat{\beta}. \tag{38}$$

**Frisch – Waugh Theorem**

We conclude this section by presenting the *Frisch – Waugh Theorem*. Suppose that we split our regressors into two groups:

$$y = X\beta = [X_1\ X_2]\begin{bmatrix}\beta_1 \\ \beta_2\end{bmatrix} + \epsilon = X_1\beta_1 + X_2\beta_2 + \epsilon, \tag{39}$$

and let $\hat{\beta}$ be the least squares estimate as in (34). The estimate $\hat{\beta}_2$ is just the components of $\hat{\beta}$ corresponding to the covariates in $X_2$. Now consider a different approach to estimate $\beta_2$: first form the matrix $M_{X_1}$ and pre-multiply (39) by $M_{X_1}$ to obtain (cf. Lemma 4.1(i))

$$M_{X_1}y = M_{X_1}X_1\beta_1 + M_{X_1}X_2\beta_2 + M_{X_1}\epsilon = M_{X_1}X_2\beta_2 + u, \quad u := M_{X_1}\epsilon, \tag{40}$$

and let $\tilde{\beta}_2$ be the least-squares estimate of $\beta_2$ based on (40). These two procedures are equivalent and $\hat{\beta}_2 = \tilde{\beta}_2$.

Theorem 4.1 [Frisch – Waugh]: *Suppose that $n > K$ and $X$ has full column rank. Let $\hat{\beta}_2$ be the components of the least squares estimate of $\beta$ in* (39) *corresponding to $X_2$ and let $\tilde{\beta}_2$ be the least squares estimate of $\beta_2$ based on* (40)*. Then (i) $\hat{\beta}_2 = \tilde{\beta}_2$ and (ii) the residuals $\hat{\epsilon} := y - X\hat{\beta}$ and $\tilde{\epsilon} := M_{X_1}y - M_{X_1}X_2\tilde{\beta}_2$ are equal.*

*Proof.* By Lemma 4.1(iii) and the fact that $P_X y = X\hat{\beta}$ we have

$$y = P_X y + M_X y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + M_X y.$$

Pre-multiply by $X_2' M_{X_1}$ and use Lemma 4.1(i) to obtain

$$X_2' M_{X_1} y = X_2' M_{X_1} X_2 \hat{\beta}_2 + X_2' M_{X_1} M_X y = X_2' M_{X_1} X_2 \hat{\beta}_2,$$

where the last equality is due to the fact that $M_X M_{X_1} X_2 = 0$ [Exercise]. The matrix $X_2' M_{X_1} X_2 = X_2' M_{X_1}' M_{X_1} X_2$ is invertible [Exercise], so we can solve the equation in the preceding display to obtain

$$\hat{\beta}_2 = (X_2' M_{X_1} X_2)^{-1} X_2' M_{X_1} y = (X_2' M_{X_1}' M_{X_1} X_2)^{-1} X_2' M_{X_1} y = \tilde{\beta}_2.$$

For (ii) pre-multiply the first equation in the proof by $M_{X_1}$ to obtain

$$M_{X_1} y = M_{X_1} X_2 \hat{\beta}_2 + M_{X_1} M_X y = M_{X_1} X_2 \hat{\beta}_2 + M_X y,$$

since $M_{X_1} M_X = M_X$ [Exercise]. Therefore, using part (i), $M_X y = M_{X_1} y - M_{X_1} X_2 \tilde{\beta}_2$ are the residuals from regression (40) and we have already noted that $M_X y = y - X\hat{\beta}$. □

## 4.2 Statistical properties of the least squares estimator

The previous section dealt with geometric properties of the least squares estimate. We will now consider its *statistical* properties. That is, we investigate properties relating to the distribution of $\hat{\beta}$ that can be shown to hold provided we are willing to make certain assumptions regarding the distribution of the underlying data $(y, X)$.

### 4.2.1 The normal linear regression model

Classically, the linear regression model is coupled with an assumption of normal errors.

THEOREM 4.2: *Suppose that the observations $(y_i, X_i)_{i=1}^n$ are i.i.d. such that $\epsilon_i | X_i$ has (conditional) density $f_\gamma$ and $X_i$ has density/mass function $g$. Then*

*(i) The log-likelihood function is*

$$l(\theta) = \sum_{i=1}^n \log f_\gamma(y_i - X_i'\beta) + \sum_{i=1}^n \log g(X_i), \qquad \theta = (\beta, \gamma) \in \Theta.$$

*(ii) The MLE of $\theta$ and, therefore, of $\beta$ does not depend on $g$.*

*(iii) If $X'X$ is non-singular $\theta$ is identifiable*

*(iv) If $f$ is the density of a $\mathcal{N}(0, \sigma^2)$, then $y_i | X_i \sim \mathcal{N}(X_i'\beta, \sigma^2)$ and if $X'X$ is non-singular, the OLS estimator $\hat{\beta}$ is the MLE.*

*Proof.* Exercise. □

THEOREM 4.3: *Under the conditions of (iv) of Theorem 4.2 $\hat{\beta}$ is unbiased. If we additionally suppose that the $X_i$ are non-random, then $\hat{\beta}$ attains the Cramér – Rao lower bound, $\sigma^2(X'X)^{-1}$.*

*Proof.* By (iii) of Theorem 4.2, $y|X \sim \mathcal{N}(X\beta, \sigma^2 I)$ so $\hat{\beta} = (X'X)^{-1}X'Y$ satisfies

$$\hat{\beta}|X \sim \mathcal{N}\left(\beta, \sigma^2(X'X)^{-1}\right). \tag{41}$$

That $\hat{\beta}$ is unbiased now follows by the LIE. Supposing that the $X_i$ are non-random, to establish that $\hat{\beta}$ attains the CR lower bound, weverify the conditions of Theorem 3.7. $\theta = (\beta, \sigma^2)$ and , [Exercise]

$$\begin{aligned}\ell_\theta(y) = \nabla_\theta \log p_\theta(y) &= \nabla_\theta \sum_{i=1}^{n}\left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y_i - X_i'\beta)^2\right] \\ &= \frac{-1}{2\sigma^2}\begin{pmatrix} 2\sum_{i=1}^{n}(y_i - X_i'\beta)X_i \\ n - \sum_{i=1}^{n}\frac{(y_i - X_i'\beta)^2}{\sigma^2}\end{pmatrix}.\end{aligned}$$

By (iii) of Theorem 4.2 (and the assumed non-randomness of $X_i$), $y_i - X_i'\beta \sim \mathcal{N}(0, \sigma^2)$, hence

$$\mathbb{E}[\ell_\theta(W_i)] = \frac{-1}{2\sigma^2}\begin{pmatrix} 2\sum_{i=1}^{n}\mathbb{E}[(y_i - X_i'\beta)]X_i \\ n - \sum_{i=1}^{n}\frac{\mathbb{E}[(y_i - X_i'\beta)^2]}{\sigma^2}\end{pmatrix} = 0.$$

Moreover, one can check that (by independence) [Exercise]

$$I(\theta) = \mathbb{E}[\ell_\theta(W_i)\ell_\theta(W_i)'] = \frac{1}{4\sigma^4}\begin{pmatrix} 4\sigma^2 X'X & 0 \\ 0 & 2\end{pmatrix} = \begin{pmatrix}\sigma^{-2}X'X & 0 \\ 0 & \frac{n}{2\sigma^4}\end{pmatrix}.$$

In consequence $I(\theta)$ exists, is positive definite and

$$I(\theta)^{-1} = \begin{pmatrix}\sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n}\end{pmatrix}.$$

For the second condition of Theorem 3.7, note that $\mathbb{E}[\hat{\beta}] = \beta$ by (41) and hence $\nabla_\theta \mathbb{E}[\hat{\beta}] = [I, 0]'$ and since $\hat{\beta} = (X'X)^{-1}X'(X\beta + \epsilon) = \beta + (X'X)^{-1}X'\epsilon$ one also has that $\nabla_\theta\hat{\beta} = [I, 0]'$ hence $\mathbb{E}[\nabla_\theta\hat{\beta}] = \nabla_\theta \mathbb{E}[\hat{\beta}]$.

Since here the function $g(\theta) = J\theta$ for $J = \left[\begin{smallmatrix} I & 0 \\ 0 & 0\end{smallmatrix}\right]$, one has $\nabla_\theta g(\theta) = J$ and therefore, the CR lower bound is $\sigma^2(X'X)^{-1}$ by Corollary 3.3 which is attained by $\hat{\beta}$ given (41). □

Equation (41) in the proof of Theorem 4.3 gives the distribution of $\hat{\beta}|X$ when we assume that $\epsilon|X \sim \mathcal{N}(0, \sigma^2 I)$. If we knew $\sigma^2$ we could use this to conduct hypothesis tests about the parameters $\beta$. In practice, we do not know $\sigma^2$; we will discuss its estimation in the next section and subsequently use it to contstruct hypothesis tests in the normal linear regression model.

### 4.2.2 Finite sample properties of least squares estimators

The assumption of $\epsilon_i|X_i \sim \mathcal{N}(0,\sigma^2)$ is strong. We might view this as an approximation which often gives reasonable results. Nevertheless, various properties of the OLS estimator which hold under this normality assumption remain true under weaker conditions.

#### Basic assumptions, unbiasedness and OLS variance

Assumption 4.1 [Linearity]: $(y, X)$ *satisfies* (33).

Assumption 4.2 [Conditional mean independence]: $\mathbb{E}[\epsilon|X] = 0$.

Assumption 4.3 [Linear independence]: $X$ *has full column rank with probability 1.*

These three assumptions are sufficient to demonstrate that $\hat{\beta}$ is a (conditionally) unbiased estimator.

Proposition 4.2 [Least squares estimator is (conditionally) unbiased]: *Under Assumptions 4.1, 4.2, and 4.3, the least squares estimator $\hat{\beta}$ is unbiased conditionally on $X$:*

$$\mathbb{E}\left[\hat{\beta}\middle|X\right] = \beta.$$

*In consequence, $\hat{\beta}$ is unbiased:* $\mathbb{E}[\hat{\beta}] = \beta$.

*Proof.* With probability 1, $(X'X)^{-1}$ exists and therefore the same is true of $\hat{\beta}$. Plugging in $y = X\beta + \epsilon$ from (33) we have

$$\hat{\beta} = (X'X)^{-1}X'y = (X'X)^{-1}X'\left(X\beta + \epsilon\right) = \beta + (X'X)^{-1}X'\epsilon. \tag{42}$$

Taking conditional expectations,

$$\mathbb{E}\left[\hat{\beta}\middle|X\right] = \mathbb{E}\left[\beta + (X'X)^{-1}X'\epsilon\middle|X\right] = \beta + (X'X)^{-1}X'\,\mathbb{E}[\epsilon|X] = \beta.$$

Taking expectations on both sides yields the second conclusion by the law of iterated expectations:

$$\mathbb{E}\left[\hat{\beta}\right] = \mathbb{E}\,\mathbb{E}\left[\hat{\beta}\middle|X\right] = \mathbb{E}\,\beta = \beta. \qquad \square$$

To say something about the variance of the estimator $\hat{\beta}$, we need to assume something about the variance of the error term, $\epsilon$. We first consider the error term to be homoskedastic: that is, each $\epsilon_i$ $(i = 1, \ldots, n)$ is uncorrelated from the others and all have the same variance, $\sigma^2$. We will consider less restrictive assumptions on the error variance matrix later on.

Assumption 4.4 [Homoskedasticity]: $\mathbb{E}[\epsilon\epsilon'|X] = \sigma^2 I_n$.

Proposition 4.3 [Conditional variance of least squares estimator]: *Under Assumptions 4.1, 4.2, 4.3 and 4.4,*

$$\operatorname{Var}\left[\hat{\beta}\middle|X\right] = \sigma^2(X'X)^{-1}.$$

*In consequence,*

$$\operatorname{Var}\left[\hat{\beta}\right] = \sigma^2\,\mathbb{E}\left[(X'X)^{-1}\right].$$

*Proof.* As in the proof of Proposition 4.2, (42) holds and so

$$\begin{aligned}\operatorname{Var}\left[\hat{\beta}\middle|X\right] &= \operatorname{Var}\left[\hat{\beta}-\beta\middle|X\right]\\ &= \operatorname{Var}\left[(X'X)^{-1}X'\epsilon\middle|X\right]\\ &= (X'X)^{-1}X'\,\mathbb{E}\left[\epsilon\epsilon'\middle|X\right]X(X'X)^{-1}\\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1}\\ &= \sigma^2(X'X)^{-1}.\end{aligned}$$

By the preceding display and Proposition 4.2 we have

$$\operatorname{Var}\left[\hat{\beta}\middle|X\right] = \mathbb{E}\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'\middle|X\right] = \sigma^2(X'X)^{-1}.$$

Hence, taking expectations yields the second conclusion by the law of iterated expectations and Proposition 4.2 as:

$$\operatorname{Var}\left[\hat{\beta}\right] = \mathbb{E}\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'\right] = \mathbb{E}\,\mathbb{E}\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'\middle|X\right] = \sigma^2\,\mathbb{E}\left[(X'X)^{-1}\right]. \qquad \square$$

### Gauss – Markov Theorem

By Theorem 4.3, in the special case where $X$ is fixed and the error term is normally distributed, the OLS estimator $\hat{\beta}$ has the smallest variance amongst unbiased estimators. Under Assumptions 4.1 - 4.4 a similar property holds if we further restrict the class of competing estimators by requiring that they be linear in the observations $y$.[65] This result is called the "Gauss – Markov Theorem" and states that the OLS estimator is "BLUE", the "Best Linear Unbiased Estimator".

Theorem 4.4 [Gauss – Markov]: *Under Assumptions 4.1, 4.2, 4.3 and 4.4, the OLS estimator $\hat{\beta}$ is efficient in the class of linear conditionally unbiased estimators in the sense of having smaller conditional variance. That is, for any other estimator $\tilde{\beta}$ which is linear in $y$,* $\operatorname{Var}\left[\hat{\beta}\middle|X\right] \leq \operatorname{Var}\left[\tilde{\beta}\middle|X\right]$.[66]

*Proof.* As $\tilde{\beta}$ is linear in $y$, there is a matrix $C$ (which may be a function of $X$) such that $\tilde{\beta} = Cy$. Let $A := (X'X)^{-1}X'$ and $D := C - A$. We then have

$$\tilde{\beta} = Cy = (D+A)y = Dy + (X'X)^{-1}X'y = DX\beta + D\epsilon + \hat{\beta}.$$

[65] $\hat{\beta}$ is linear in $y$ as is clear from (34).

[66] For two square symmetric matrices $A, B$, $A \geq B$ if and only if $A - B$ is positive semi-definite.

Taking conditional expectations on both sides yields

$$\mathbb{E}\left[\tilde{\beta}\middle|X\right] = \mathbb{E}\left[DX\beta + D\epsilon + \hat{\beta}\middle|X\right] = DX\beta + D\,\mathbb{E}\left[\epsilon|X\right] + \mathbb{E}\left[\hat{\beta}\middle|X\right] = DX\beta + \beta,$$

where the last equality is by Assumption 4.2 and Proposition 4.2. Since $\tilde{\beta}$ is also conditionally unbiased, the left hand side equals $\beta$ and we conclude that $DX\beta = 0$. This argument holds for any $\beta \in \mathbb{R}^K$, and so it must also be that $DX = 0$. Using this, the first display in the proof and the decomposition (42) we obtain $\tilde{\beta} - \beta = D\epsilon + \hat{\beta} - \beta = (D + A)\epsilon$. Then, as $\beta$ is fixed and $D + A$ is a function of $X$

$$\begin{aligned}
\operatorname{Var}\left[\tilde{\beta}\middle|X\right] &= \operatorname{Var}\left[\tilde{\beta} - \beta\middle|X\right] \\
&= \operatorname{Var}\left[(D + A)\epsilon|X\right] \\
&= (D + A)\,\mathbb{E}\left[\epsilon\epsilon'\middle|X\right](D + A)' \\
&= \sigma^2(D + A)(D + A)' \\
&= \sigma^2\left(DD' + AD' + DA' + AA'\right).
\end{aligned}$$

Since $DX = 0$, we have that $AD' = (X'X)^{-1}X'D' = 0$ and $DA' = DX(X'X)^{-1} = 0$. Moreover, $AA' = (X'X)^{-1}X'X(X'X)^{-1} = (X'X)^{-1}$ and so by the fact that $DD'$ is positive semi-definite [Exercise] and Proposition 4.3

$$\operatorname{Var}\left[\tilde{\beta}\middle|X\right] = \sigma^2\left(DD' + (X'X)^{-1}\right) \geq \sigma^2(X'X)^{-1} = \operatorname{Var}\left[\hat{\beta}\middle|X\right]. \qquad \square$$

### Estimating $\sigma^2$

Earlier on we noted that in order to conduct hypothesis tests (as we will discuss in the next section) under our assumptions, we require an estimator of $\sigma^2$. We next show that the choice

$$\hat{\sigma}^2 := \frac{1}{n-K}\sum_{i=1}^{n}\hat{\epsilon}_i^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{n-K}, \qquad \text{with} \qquad \hat{\epsilon} := y - X\hat{\beta}, \tag{43}$$

is unbiased.

Proposition 4.4 [Sample variance estimator is (conditionally) unbiased]: *Under Assumptions 4.1, 4.2, 4.3 and 4.4 and provided $n > K$, $\mathbb{E}\left[\hat{\sigma}^2\middle|X\right] = \sigma^2$.*

*Proof.* By (38) and Lemma 4.1, $\hat{\epsilon}'\hat{\epsilon} = \epsilon' M_X \epsilon$ and so

$$\begin{aligned}
\mathbb{E}\left[\hat{\epsilon}'\hat{\epsilon}|X\right] &= \mathbb{E}\left[\epsilon' M_X \epsilon|X\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}[M_X]_{i,j}\,\mathbb{E}\left[\epsilon_i\epsilon_j|X\right] \\
&= \sum_{i=1}^{n}[M_X]_{i,i}\sigma^2 \\
&= \operatorname{tr} M_X\sigma^2,
\end{aligned} \tag{44}$$

where $\operatorname{tr} M_X$ is the trace of $M_X$. Since the trace is linear, $\operatorname{tr} M_X = \operatorname{tr} I_n - \operatorname{tr} P_X = n - \operatorname{tr} P_X$. Moreover, using the cyclic property of the trace, $\operatorname{tr} PZ = \operatorname{tr} X(X'X)^{-1}X' = \operatorname{tr} X'X(X'X)^{-1} = \operatorname{tr} I_K = K$. Combining these we obtain $\mathbb{E}\left[\hat{\epsilon}'\hat{\epsilon}|X\right] = (n-K)\sigma^2$, which suffices since $n-K > 0$. □

Example 4.4: Consider again the regression model from Example 4.1. Our estimator $\hat{\sigma}^2$ from (43) is $\hat{\sigma}^2 = 0.2134$. △

### Heteroskedasticity

Assumption 4.4, which requires the error term to have a "spherical" variance matrix $\sigma^2 I_n$ may be considered somewhat restrictive and unsuitable for certain applications. As such we analyse the conditionally *heteroskedastic* situation.

Assumption 4.5 [Heteroskedasticity]: $\mathbb{E}\left[\epsilon\epsilon'|X\right] = \sigma^2 V(X)$ *for a positive-definite matrix* $V(X)$.

We note in passing that in this situation the OLS estimator remains conditionally unbiased since Proposition 4.2 does not require Assumption 4.4.

With Assumption 4.5 replacing Assumption 4.15 the conditional variance of the OLS estimator is no longer $\sigma^2(X'X)^{-1}$ but rather [Exercise]

$$\mathbb{E}\left[\hat{\beta}\middle|X\right] = \sigma^2(X'X)^{-1}X'V(X)X(X'X)^{-1}.$$

### Generalised Least Squares (GLS) & Aitken's Theorem

Under heteroskedasticity, the OLS estimator is no longer efficient. To derive an efficient estimator based on our existing results, we will transform the regression equation. Since $V(X)$ is positive definite, there is a positive definite matrix $C(X)$ such that $V(X)^{-1} = C(X)C(X)$.[67]

We will supress the argument $X$ in $C(X)$ and write $C = C(X)$ to keep the notation simple. Transform equation (33) by pre-multiplying by $C$:

$$Cy = CX\beta + C\epsilon.$$

If we let $\tilde{y} := Cy$, $\tilde{X} := CX$ and $\tilde{\epsilon} = C\epsilon$, then this yields a new regresion model

$$\tilde{y} = \tilde{X}\beta + \tilde{\epsilon}. \tag{45}$$

Lemma 4.2: *If Assumptions 4.1, 4.2, 4.3 and 4.5 hold, then* (45) *holds,* $\mathbb{E}[\tilde{\epsilon}|\tilde{X}] = 0$, $\tilde{X}$ *has full column rank with probability one and* $\mathbb{E}\left[\tilde{\epsilon}\tilde{\epsilon}'\middle|\tilde{X}\right] = I_n$.

*Proof.* (45) follows from (33) by pre-multiplying by $C$, which exists since $V(X)$ is positive definite (Assumption 4.5). As the function $x \mapsto Cx$ is bijective (since $C$ is invertible), knowledge of $CX$ is the same as knowledge of $X$, hence

$$\mathbb{E}\left[\tilde{\epsilon}\middle|\tilde{X}\right] = \mathbb{E}\left[C(X)\epsilon\middle|\tilde{X}\right] = C(X)\,\mathbb{E}\left[\epsilon\middle|\tilde{X}\right] = C(X)\,\mathbb{E}\left[\epsilon|X\right] = 0.$$

[67] Since positive (semi-)definite matrices are symmetric, $C(X) = C(X)'$.

Similarly, using Assumption 4.5 and $V(X) = [C(X)C(X)]^{-1} = C(X)^{-1}C(X)^{-1}$,

$$\mathbb{E}\left[\tilde{\epsilon}\tilde{\epsilon}'\middle|\tilde{X}\right] = \mathbb{E}\left[C(X)\epsilon\epsilon' C(X)\middle|\tilde{X}\right] = C(X)\,\mathbb{E}\left[\epsilon\epsilon\middle|X'\right]C(X) = \sigma^2 C(X)V(X)C(X) = \sigma^2 I_n.$$

Finally, to show that (with probability 1) $CX$ has rank $K$, it suffices to show (by the rank - nullity theorem) that $\ker(CX) = \{0\}$. Since $X$ has full column rank, it follows (again by the rank - nullity theorem) that $\ker(X) = \{0\}$. Moreover since $\ker(C) = \{0\}$ since it is also full rank (as it is positive definite). So for any $a \neq 0$, we have $Xa \neq 0$ and hence $CXa \neq 0$. □

The upshot of (the preceding) Lemma 4.2 is that Assumptions 4.1 - 4.4 hold for the transformed data $(\tilde{y}, \tilde{X})$.

The *generalised least squares* (GLS) estimator, $\tilde{\beta}$, is the OLS estimator based on the transformed model (45):

$$\tilde{\beta} := (\tilde{X}'\tilde{X})^{-1}\tilde{X}\tilde{y} = (X'V(X)^{-1}X)^{-1}X'V(X)^{-1}y. \tag{46}$$

Proposition 4.5 [First two (conditional) moments of GLS estimator]: *Suppose that Assumptions 4.1, 4.2, 4.3 and 4.5 hold. Then*

$$\mathbb{E}\left[\tilde{\beta}\middle|X\right] = \beta \qquad \textit{and} \qquad \operatorname{Var}\left[\tilde{\beta}\middle|X\right] = \sigma^2(X'V(X)^{-1}X)^{-1}.$$

*Proof.* Exercise. □

Theorem 4.5 [Aitken]: *Suppose that Assumptions 4.1, 4.2, 4.3 and 4.5 hold. Then the GLS estimator is efficient in the class of linear unbiased estimators in the sense of having smaller conditional variance. That is, for any other estimator $\check{\beta}$ which is linear in $y$,* $\operatorname{Var}\left[\tilde{\beta}\middle|X\right] \leq \operatorname{Var}\left[\check{\beta}|X\right]$.

*Proof.* By Lemma 4.2 is that Assumptions 4.1 - 4.4 hold for the transformed data $(\tilde{y}, \tilde{X})$. $\tilde{\beta}$ is the OLS estimate in the transformed model (45). Hence by Theorem 4.4 and the fact that conditioning on $X$ is the same as on $\tilde{X}$ (cf. the Proof of Lemma 4.2),

$$\operatorname{Var}\left[\tilde{\beta}\middle|X\right] = \operatorname{Var}\left[\tilde{\beta}\middle|\tilde{X}\right] \leq \operatorname{Var}\left[\check{\beta}\middle|\tilde{X}\right] = \operatorname{Var}\left[\check{\beta}|X\right].$$ □

Aitken's Theorem (Theorem 4.5) demonstrates that the GLS estimator is BLUE. However, the form of the GLS estimator in (46) requires the form of $V(X)$ to be *known* and so the estimator is typically infeasible in practice. Moreover, we cannot estimate $V(X)$: it is a symmetric $n \times n$ matrix and hence has $n(n+1)/2$ unique elements, which is larger than our number of samples.

If $V(X)$ is diagonal, with $\operatorname{diag}(V(X)) = v_1(X), \ldots, v_n(X)$, this is called *weighted least squares* (WLS). This gets its name from the following expression:

$$\tilde{\beta} = (X'V(X)^{-1}X)^{-1}X'V(X)^{-1}y\left[\sum_{i=1}^{n} X_i X_i' \frac{1}{v_i(X)}\right]^{-1}\sum_{i=1}^{n} X_i y_i' \frac{1}{v_i(X)}. \tag{47}$$

Each observation is “weighted” by the corresponding $v_i(X)$.

A feasible version may be based on replacing $V(X)$ by an estimator. This usually requires the specification of a functional form for $V(X)$ which is known up to some parameters to be estimated.

The simplest specification we might consider is that our data contains two groups, $g = 1, 2$. Suppose that our data are ordered such that all the observations from group 1 have indices $i = 1, \dots, n_1$ and those from group 2, $i = n_1 + 1, \dots, n$ and

$$\mathrm{Var}(\epsilon^2|X) = \begin{bmatrix} \sigma_1^2 I & 0 \\ 0 & \sigma_2^2 I \end{bmatrix}. \tag{48}$$

That is, we still have the homoskedastic form of Assumption 4.4 but with different scales across the two groups. We can estimate $\sigma_g^2$ using (43), for example, but only including the observations belonging to group $g$. Specifically, let $\hat{\epsilon} := y - X\hat{\beta}$ and $\hat{\epsilon}_{(1)}$ be the first $n_1$ elements and $\hat{\epsilon}_{(2)}$ the remaining $n_2 = n - n_1$ elements. Then

$$\hat{\sigma}_g^2 := \frac{\hat{\epsilon}_{(g)}{}'\hat{\epsilon}_{(g)}}{n_g - K}.$$

We can then form an estimate of $V = V(X)$ as

$$\hat{V} = \begin{bmatrix} \hat{\sigma}_1^2 I & 0 \\ 0 & \hat{\sigma}_2^2 I \end{bmatrix},$$

and then a feasible WLS (FWLS) estimator of $\beta$ as

$$\breve{\beta} = (X'\hat{V}(X)^{-1}X)^{-1}X'\hat{V}(X)^{-1}y. \tag{49}$$

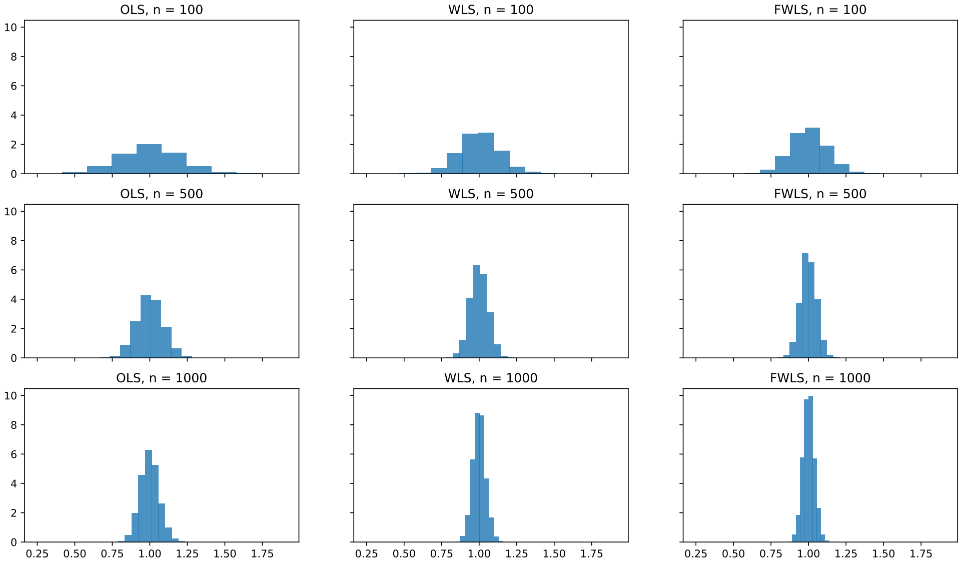
Lets use this (F)WLS estimator to illustrate Aitken’s Theorem.

Example 4.5: Lets simulate data from model (4.1) where $K = 1$, $\beta_1 = 1$, $X_i \sim t(8)$ and $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ is independent of $X_i$ with $\sigma_i^2 = 1$ for $i = 1, \dots, \lfloor n/2 \rfloor$ and $\sigma_i^2 = 3$ for $i = \lfloor n/2 \rfloor, \dots, n$.

In particular lets draw $n \in \{100, 500, 1000\}$ samples each $M = 5000$ times and compare (a) the OLS estimator; (b) the infeasible WLS estimator where we use the true $V$; (c) the feasible WLS estimator using $\hat{V}$ as above.

Plotting a histogram of each of the estimates for each sample size gives:

Figure 19: Aitken's Theorem in simulation



We see here the increase in efficiency from weighting the observations appropriately. △

Another practically useful specification of WLS concerns the case where each $\epsilon_i$ is conditionally uncorrelated with the others, but may depend on $X_i$ (and the dependence is the same for each $i$). Specifically we will suppose that

$$\sigma_i^2(X_i) := \mathbb{E}[\epsilon_i^2|X_i] = v(X_i),$$

where $v$ is a function from $\mathbb{R}^K \to (0,\infty)$. Then we have that

$$\mathbb{E}[\epsilon\epsilon'|X_i] = V(X) = \begin{bmatrix} v(X_1) & 0 & 0 & \cdots & 0 \\ 0 & v(X_2) & 0 & \cdots & 0 \\ 0 & 0 & v(X_3) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & v(X_n) \end{bmatrix}. \tag{50}$$

In practice, this is again infeasible, as we do not know $v(X_i)$. In order to estimate $\hat{V}(X)$ we can use the following procedure. We will suppose that we can write $v(X_i)$ in the following way:

$$v(X_i) = \exp(Z_i'\gamma), \qquad Z_i = g(X_i), \tag{51}$$

for some $\gamma \in \mathbb{R}^K$ and $g : \mathbb{R}^K \to \mathbb{R}^L$ (typically with $L > K$).[68] We can consider forming estimates of $v$ in the following manner: regress $Z_i$ on the logged residuals of our equation formed with

[68] For example, we might want to include e.g. powers of our explanatory variables.

the OLS estimates, i.e. run the regression

$$\log \hat{\epsilon}_i^2 = Z_i'\gamma + u_i,$$

to obtain $\hat{\gamma}$. Then form the estimate

$$\hat{V}(X) = \begin{bmatrix} \exp(Z_1'\hat{\gamma}) & 0 & 0 & \cdots & 0 \\ 0 & \exp(Z_2'\hat{\gamma}) & 0 & \cdots & 0 \\ 0 & 0 & \exp(Z_3'\hat{\gamma}) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \exp(Z_n'\hat{\gamma}) \end{bmatrix}.$$

We can then use our $\hat{V}(X)$ in place of the true (unknown) $V(X)$ to form the *feasible WLS* (FWLS) estimator

$$\breve{\beta} = (X'\hat{V}(X)^{-1}X)^{-1}X'\hat{V}(X)^{-1}y. \tag{52}$$

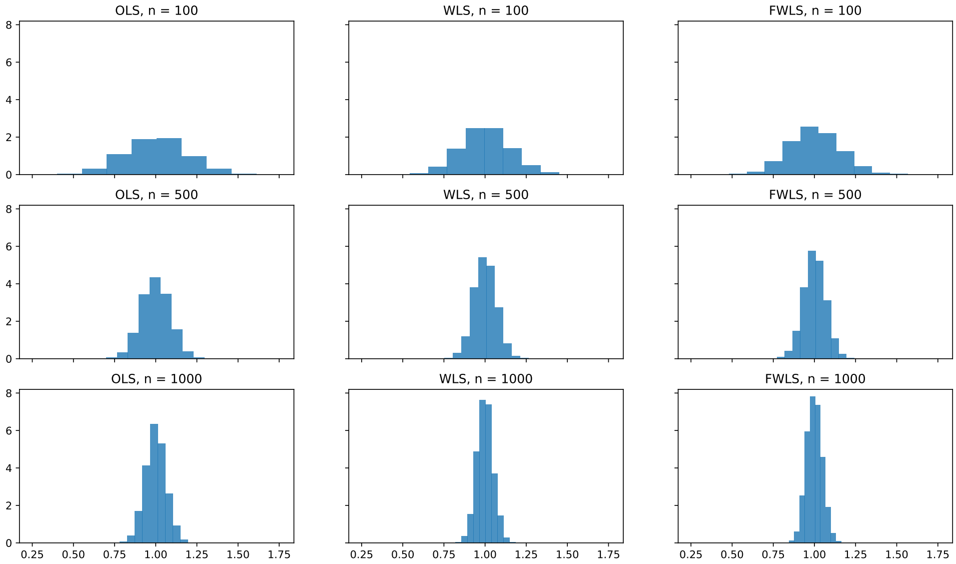
Lets use this (F)WLS estimator to illustrate Aitken's Theorem, again.

Example 4.6: Lets simulate data from model (4.1) where $K = 1$, $\beta_1 = 1$, $X_i \sim t(8)$ and $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ is independent of $X_i$ with $\mathbb{E}[\epsilon\epsilon'|X] = V$ for $V$ as in (50) where each $\sigma^2(X_i) = 0.5 + \log(1 + X_i^2)$.

In particular lets draw $n \in \{100, 500, 1000\}$ samples each $M = 5000$ times and compare (a) the OLS estimator; (b) the infeasible WLS estimator where we use the true $V$; (c) the feasible WLS estimator using $\hat{V}(X)$ as above, using $Z_i = (1, X_i, X_i^2, X_i^3)$.

Plotting a histogram of each of the estimates for each sample size gives:

Figure 20: Aitken's Theorem in simulation



The histograms reveals that all the estimators appear to be unbiased (as they should be, given Proposition 4.2). As $n$ increases each estimator becomes more and more accurate – more tightly distributed around the true value of $\beta = 1$. The WLS estimators, both feasible and infeasible are more tightly concentrated around the true value that then OLS estimator. △

Example 4.7: Consider again the regression model from Example 4.1. The feasible weighted least squares estimator gives the model

$$\widehat{\log(wage)} = 0.4981 + 0.0902 \times \mathit{education} + 0.0378 \times \mathit{experience} - 0.0006 \times \mathit{experience}^2. \quad \triangle$$

### 4.2.3 *Asymptotic properties of least squares estimators

To derive asymptotic properties of the least squares estimator $\hat{\beta}$, we will require similar assumptions than those we considered in section 4.2.2.

#### 4.2.3.1 i.i.d. data

Our first assumption of linearity (Assumption 4.1) remains the same, though we express it in a different form. We additionally need to impose a restriction on the stochastic properties of the data sequence $(y_n, X_n)_{n\in\mathbb{N}}$. We will first consider the case where the data is i.i.d..

Assumption 4.6 [Linearity]: *For each $i \in \mathbb{N}$,*

$$y_i = X_i'\beta + \epsilon_i.$$

Assumption 4.7 [Random sample]: $(y_n, X_n)_{n\in\mathbb{N}}$ *is an i.i.d. sequence.*

We will replace our conditional mean independence assumption with the weaker condition of no correlation between $X_i$ and $\epsilon_i$ (along with a moment existence condition). The full column rank assumption is also weakened slightly.

Assumption 4.8 [No correlation between error and regressors]: $\mathbb{E}\,\|X_i\epsilon_i\| < \infty$ *and* $\mathbb{E}\,[X_i\epsilon_i] = 0$ *for $i \in \mathbb{N}$.*

Assumption 4.9 [Linear independence]: $\mathbb{E}\,X_iX_i'$ *exists (i.e.* $\mathbb{E}\,\|X_iX_i'\| < \infty$*) and is positive definite.*

Under these four assumptions, the OLS estimator $\hat{\beta}_n$ is *consistent*: as $n \to \infty$, it converges in probability to the true value $\beta$.

Proposition 4.6 [Consistency of the OLS estimator]: *Under Assumptions 4.6, 4.7, 4.8 and 4.9, as $n \to \infty$,*

$$\hat{\beta}_n \xrightarrow{P} \beta.$$

*Proof.* We may re-write $X'X$ and $X'y$ as $\sum_{i=1}^n X_iX_i'$ and $\sum_{i=1}^n X_iy_i$ respectively and by dividing and multiplying by $n$, we can re-express $\hat{\beta}_n$ as

$$\hat{\beta}_n = \left[\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} X_i y_i\right],$$

provided the inverse in the preceding display exists. By Assumptions 4.7 and 4.9, each of the sequences $(X_{i,j}X_{i,k})_{i\in\mathbb{N}}$ are i.i.d. and with finite mean. Hence by the weak law of large numbers applied to each component of $M_n := \frac{1}{n}\sum_{i=1}^{n} X_i X_i'$ we have that $M_n \xrightarrow{P} M$. Since $M := \mathbb{E}\, X_i X_i'$ is positive definite, $\det(M) > 0$ and since the determinant is a continuous function, $\det(M_n) \xrightarrow{P} \det(M)$ by the continuous mapping theorem. This implies that $M_n$ has non-zero determinant and hence is positive definite with probability approaching one. As such, with probability approaching one, $M_n^{-1}$ and hence $\hat{\beta}_n$ exists. Moreover, since matrix inversion is continuous on the set of nonsingular matrices, a second application of the continuous mapping theorem gives $M_n^{-1} \xrightarrow{P} M^{-1}$.

Plugging in $y_i = X_i' + \epsilon_i$ from Assumption 4.6, we obtain $\frac{1}{n}\sum_{i=1}^{n} X_i y_i = \frac{1}{n}\sum_{i=1}^{n} X_i X_i'\beta + \frac{1}{n}\sum_{i=1}^{n} X_i \epsilon_i$. Pre-multiplying by $M_n^{-1}$ gives (with probability approaching one)

$$\hat{\beta}_n = \left[\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right]\beta + \left[\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right]^{-1}\frac{1}{n}\sum_{i=1}^{n} X_i\epsilon_i = \beta + M_n^{-1}\frac{1}{n}\sum_{i=1}^{n} X_i\epsilon_i. \tag{53}$$

Under Assumption 4.8, each $(X_{i,k}\epsilon_i)_{n\in\mathbb{N}}$ $(k = 1, \ldots, K)$ has mean zero and so an application of the WLLN gives $\frac{1}{n}\sum_{i=1}^{n} X_i\epsilon_i \xrightarrow{P} 0$. Applying Slutsky's Theorem twice then allows us to conclude that $M_n^{-1}\frac{1}{n}\sum_{i=1}^{n} X_i\epsilon_i \xrightarrow{P} 0$ and in consequence $\hat{\beta}_n \xrightarrow{P} \beta$. □

We now consider the asymptotic distribution of a scaled and recentered version of the OLS estimator. For this we need to strengthen one of our moment conditions.

Assumption 4.10 [Second moments of $X_i\epsilon_i$]: *$\Sigma := \operatorname{Var}(X_i\epsilon_i)$ exists (i.e. $\|\Sigma\| < \infty$) and is positive definite.*

Proposition 4.7 [Asymptotic distribution of the OLS estimator]: *Under Assumptions 4.6, 4.7, 4.8, 4.9 and 4.10, as $n \to \infty$,*

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) \rightsquigarrow \mathcal{N}\left(0, M^{-1}\Sigma M^{-1}\right),$$

*where $M := \mathbb{E}\,[X_1 X_1']$.*

*Proof.* The same argument as in the proof of Proposition 4.6 applies to demonstrate that for $M_n := \frac{1}{n}\sum_{i=1}^{n} X_i X_i'$, $M_n^{-1}$ exists with probability approaching one (and so the same is true of $\hat{\beta}_n$), $M_n^{-1}$ converges in probability to $M^{-1}$ and (53) holds. Rearranging this last equation and

[68] If $g$ is a measurable function and $X, Y$ are (i) independent or (ii) identically distributed then $g(X)$ and $g(Y)$ are also (i) independent or (ii) identically distributed.

multipling by $\sqrt{n}$ we obtain

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) = M_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\epsilon_i.$$

Since $(X_i\epsilon_i)_{i\in\mathbb{N}}$ is i.i.d. (cf. foonote 68), $\mathbb{E}[X_1\epsilon_1] = 0$ (Assumption 4.8) and $\Sigma = \mathrm{Var}(X_1\epsilon_1)$ exists and is positive definite (Assumption 4.10), the multivariate version of the Lindeberg – Lévy CLT yields

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\epsilon_i \rightsquigarrow Z \sim \mathcal{N}(0,\Sigma).$$

Use this along with $M_n^{-1} \xrightarrow{P} M^{-1}$ and Slutsky's Theorem to conclude that

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) = M_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\epsilon_i \rightsquigarrow M^{-1}Z \sim \mathcal{N}(0, M^{-1}\Sigma M^{-1}). \qquad \square$$

The i.i.d. assumption (Assumption 4.7) utilised here rules out heteroskedasticity (in addition to other potential differences in the distributions of the random vectors in our data sequence). However, it does not rule out *conditional heteroskedasticity.*

Example 4.8 [Conditional heteroskedasticity]: Suppose that $\epsilon_i = \xi_i h(X_i)$ where $\xi_i$ is independent of $X_i$, $\mathbb{E}[\xi_i^2] = \varsigma < \infty$, $\mathbb{E}[h(X_i)^2] = \varrho < \infty$ and $\mathbb{E}\,\|X_i\| < \infty$. Then, $(\epsilon_i)_{i\in\mathbb{N}}$ is homoskedastic in that $\mathbb{E}[\epsilon_i^2] = \varsigma\varrho$ for all $i \in \mathbb{N}$, but conditionally heteroskedasticity in that $\mathbb{E}[\epsilon_i^2|X_i] = \varsigma h(X_i)^2$.

Since $\xi_i$ and $X_i$ are independent,

$$\mathbb{E}[\epsilon_i^2|X_i] = \mathbb{E}[\xi_i^2 h(X_i)^2|X_i] = \mathbb{E}[\xi_i^2|X_i]h(X_i)^2 = \mathbb{E}[\xi_i^2]h(X_i)^2 = \varsigma h(X_i)^2.$$

Taking expectations on both sides yields

$$\mathbb{E}[\epsilon_i^2] = \mathbb{E}\,\mathbb{E}[\epsilon_i^2|X_i] = \varsigma\,\mathbb{E}\,h(X_i)^2 = \varsigma\varrho.$$

If also $(\xi_i, X_i)_{i\in\mathbb{N}}$ are i.i.d., with $\mathbb{E}[\xi_i] = 0$, $\mathbb{E}\,\|X_i h(X_i)\|^2 < \infty$ and $\mathrm{Var}(X_1\epsilon_1)$ positive definite then provided Assumption 4.6 holds, Assumptions 4.7, 4.8 and 4.10 are satisfied.

For Assumption 4.7 note that $(\epsilon_i)_{i\in\mathbb{N}} = (\xi_i h(X_i))_{i\in\mathbb{N}}$ is i.i.d. and hence (by Assumption 4.6) so is $(y_i, X_i)_{i\in\mathbb{N}}$.[69]

Assumption 4.10 follows as

$$\mathbb{E}\left[\|X_i\epsilon_i\epsilon_i' X_i'\|\right] \le \mathbb{E}\,\|X_i\epsilon_i\|^2 = \mathbb{E}\,\|X_i\xi_i h(X_i)\|^2 \le \mathbb{E}\,\|X_i h(X_i)\|^2\,\mathbb{E}\,\xi_i^2 < \infty,$$

and the required positive definiteness was assumed. The first part of Assumption 4.8 follows since $\mathbb{E}\,\|X_i\epsilon_i\| < \infty$ is implied by $\mathbb{E}\,\|X_i\epsilon_i\|^2 < \infty$ (by the Cauchy – Schwarz inequality) For the second part, by independence $\mathbb{E}[X_i\epsilon_i] = \mathbb{E}\,\mathbb{E}[X_i h(X_i)\xi_i|X_i] = \mathbb{E}\left[X_i h(X_i)\,\mathbb{E}[\xi_i|X_i]\right] = \mathbb{E}\left[X_i h(X_i)\,\mathbb{E}[\xi_i]\right] = 0.$ △

[69] Cf. footnote 68.

#### 4.2.3.2 i.n.i.d. data

As in the finite sample setting, homoskedasticity or identically distributed random variables in general may be restrictive and/or inappropriate for many applications. As such we will now consider a set of assumptions which permits i.n.i.d. data. Following this we state Propositions analogous to Propositions 4.6 and 4.7 which are applicable in this setting. Their proofs proceed very similarly to those which apply in the i.i.d. case and are therefore left as exercises.

We firstly replace Assumption 4.7 with an assumption of independent observations, removing the requirement that they be identically distributed. The key moment restriction $\mathbb{E}[X_i\epsilon_i] = 0$ remains the same, though we require stronger moment conditions.

Assumption 4.11 [Independent sample]: *$(y_n, X_n)_{n\in\mathbb{N}}$ is a sequence of independent random vectors.*

Assumption 4.12 [No correlation between error and regressors]: *$\mathbb{E}\left\|X_i\epsilon_i\right\|^{1+\delta} < C < \infty$ for some $\delta > 0$ and $\mathbb{E}\left[X_i\epsilon_i\right] = 0$ for $i \in \mathbb{N}$.*

Assumption 4.13 [Linear independence]: *$\mathbb{E}\left\|X_iX_i'\right\|^{1+\delta} < C < \infty$ for some $\delta > 0$ and $M := \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbb{E}\, X_iX_i'$ exists and is positive definite.*

Proposition 4.8 [Consistency of the OLS estimator]: *Under Assumptions 4.6, 4.11, 4.12 and 4.13, as $n \to \infty$,*

$$\hat{\beta}_n \xrightarrow{P} \beta.$$

*Proof.* Exercise. □

As in the i.i.d. case we need to add a further assumption to obtain the asymptotic distribution of the OLS estimator.[70]

Assumption 4.14 [Second moments of $X_i\epsilon_i$]: *$\mathbb{E}\left\|X_i\epsilon_i\right\|^{2+\delta} < C < \infty$ for some $\delta > 0$ and $\Sigma := \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \operatorname{Var}(X_i\epsilon_i)$ exists and is positive definite.*

Proposition 4.9 [Asymptotic distribution of the OLS estimator]: *Under Assumptions 4.6, 4.11, 4.12, 4.13 and 4.14, as $n \to \infty$,*

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) \rightsquigarrow \mathcal{N}\left(0, M^{-1}\Sigma M^{-1}\right).$$

*Proof.* Exercise. □

#### 4.2.3.3 Estimation of asymptotic variance matrices

In order to conduct hypothesis tests or construct confidence intervals based on the asymptotic normality statements in Proposition 4.7 or 4.9 we need to estimate the asymptotic variance, $M^{-1}\Sigma M^{-1}$.

The estimation of $M$ is straightforward: we replace it with its finite-sample analog.

[70] The condition we impose here is not the weakest possible, but allows the use of the Lyapunov condition to check Lindeberg's condition.

Lemma 4.3 [Estimation of $M$]: *(i) Suppose that Assumptions 4.7 and 4.9 hold. Then $M_n := \frac{1}{n}\sum_{i=1}^n X_i X_i' \xrightarrow{P} M := \mathbb{E}\, X_i X_i'$.*

*(ii) Suppose that Assumptions 4.11 and 4.13 hold. Then $M_n := \frac{1}{n}\sum_{i=1}^n X_i X_i' \xrightarrow{P} M := \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbb{E}\, X_i X_i'$.*

*Proof.* (i) Apply the weak law of large numbers to $(X_i X_i')_{i\in\mathbb{N}}$.

(ii) Apply the weak law of large numbersto $(X_i X_i')_{i\in\mathbb{N}}$. □

### Conditional homoskedasticity

For $\Sigma$ the situation is more complicated as we do not observe the residuals. We first consider the simplest special case by adding an assumption of conditional homoskedasticity to Assumptions 4.6, 4.7, 4.8, 4.9 and 4.10.

Assumption 4.15 [Conditional homoskedasticity]: $\mathbb{E}\left[\epsilon_i^2 | X_i\right] = \sigma^2 \in (0, \infty)$.

Under assumptions 4.7, 4.8, 4.9, 4.10 and 4.15 we have that

$$\Sigma = \operatorname{Var}(X_i\epsilon_i) = \mathbb{E}\left[\epsilon_i^2 X_i X_i'\right] = \mathbb{E}\left[\mathbb{E}\left[\epsilon_i^2|X_i\right] X_i X_i'\right] = \sigma^2\,\mathbb{E}\left[X_i X_i'\right] = \sigma^2 M. \tag{54}$$

This provides a simplification of the asmyptotic variance formula:

$$M^{-1}\Sigma M^{-1} = M^{-1}\sigma^2 M M^{-1} = \sigma^2 M^{-1}. \tag{55}$$

As such, in this case we need only a consistent estimate for $\sigma^2$. We form such an estimate as follows:

$$\hat{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2, \qquad \text{with } \hat{\epsilon}_i := y_i - X_i\check{\beta}_n, \tag{56}$$

for any consistent estimator $\check{\beta}_n$ of $\beta$ (i.e. such that $\check{\beta}_n \xrightarrow{P} \beta$; the OLS estimator $\hat{\beta}_n$ is one such estimator).

Note that the estimator $\hat{\sigma}_n^2$ here is *different* to the estimator of $\sigma^2$ we considered in Section 4.2.2, equation (43). In particular, it has $1/n$ as the scaling factor outside the sum rather than $1/(n-K)$. Can we use the unbiased estimator from (43) here? Can we use the estimator $\hat{\sigma}_n^2$ in Proposition 4.4? [Exercise]

Lemma 4.4: *Under Assumptions 4.6, 4.7, 4.8, 4.9, 4.10 and 4.15, $\hat{\sigma}_n^2 \xrightarrow{P} \sigma^2$ and for $M_n := \frac{1}{n}\sum_{i=1}^n X_i X_i'$,*

$$\hat{\sigma}_n^2 M_n^{-1} \xrightarrow{P} \sigma^2 M^{-1} = \Sigma.$$

*Proof.* We can decompose $\hat{\epsilon}_i$ as

$$\hat{\epsilon}_i = \epsilon_i + (\hat{\epsilon}_i - \epsilon_i) = \epsilon_i + (y_i - X_i'\check{\beta}_n - y_i + X_i'\beta) = \epsilon_i + X_i'(\beta - \check{\beta}_n).$$

Therefore,

$$\hat{\epsilon}_i^2 = \epsilon_i^2 + 2\epsilon_i X_i'(\beta - \check{\beta}_n) + (\beta - \check{\beta}_n)' X_i X_i'(\beta - \check{\beta}_n),$$

and so

$$\hat{\sigma}_i^2 = \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 + 2\left[\frac{1}{n}\sum_{i=1}^n \epsilon_i X_i'\right](\beta - \check{\beta}_n) + (\beta - \check{\beta}_n)'\left[\frac{1}{n}\sum_{i=1}^n X_i X_i'\right](\beta - \check{\beta}_n).$$

By the weak law of large numbers, $\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \xrightarrow{P} \sigma^2$, $\frac{1}{n}\sum_{i=1}^n \epsilon_i X_i' \to \mathbb{E}[\epsilon_i X_i'] = 0$ and $M_n := \frac{1}{n}\sum_{i=1}^n X_i X_i' \to M$ [Exercise]. By assumption $\beta - \check{\beta}_n \xrightarrow{P} 0$. Combination of these with Slutsky's Theorem yields that

$$2\left[\frac{1}{n}\sum_{i=1}^n \epsilon_i X_i'\right](\beta - \check{\beta}_n) \xrightarrow{P} 0, \qquad (\beta - \check{\beta}_n)'\left[\frac{1}{n}\sum_{i=1}^n X_i X_i'\right](\beta - \check{\beta}_n) \xrightarrow{P} 0.$$

Therefore, again by Slutsky's Theorem,

$$\hat{\sigma}_n^2 \xrightarrow{P} \sigma^2 + 0 + 0 = \sigma^2.$$

Since $M_n^{-1} \xrightarrow{P} M^{-1}$ by the continuous mapping theorem. A final use of Slutsky's Theorem gives that $\hat{\sigma}_n^2 M_n^{-1} \xrightarrow{P} \sigma^2 M^{-1}$, which equals $\Sigma$ under our Assumptions as in (54). □

### Conditional heteroskedasticity

If we are not satisfied that Assumption 4.15 holds, we instead use an estimator of $\Sigma$. Define

$$\hat{\Sigma}_n := \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2 X_i X_i', \qquad \text{where } \hat{\epsilon}_i := y_i - X_i\check{\beta}_n, \tag{57}$$

for any consistent estimator $\check{\beta}_n$ of $\beta$ (i.e. such that $\check{\beta}_n \xrightarrow{P} \beta$). To show this is consistent, in the i.i.d. setting we need the existence of fourth moments of the $X_i$ variables; in the inid setting bounded $4 + \delta$ moments for some $\delta > 0$ will suffice.

Assumption 4.16 [Fourth moments of $X_i$]: $\mathbb{E}\,\|X_i\|^4 < \infty$.

Assumption 4.17 [Four + $\delta$ moments of $X_i$]: $\mathbb{E}\,\|X_i\|^{4+\delta} < C < \infty$ *for some* $\delta > 0$.

Lemma 4.5: *Suppose that Assumption 4.6 holds and $\check{\beta}_n$ (in (57)) is consistent for $\beta$ (i.e. $\check{\beta}_n \xrightarrow{P} \beta$). If also*

*(i) Assumptions 4.7, 4.8, 4.10 and 4.16 hold; or*

*(ii) Assumptions 4.11, 4.12, 4.14 and 4.17 hold,*

*then $\hat{\Sigma}_n \xrightarrow{P} \Sigma$.*

*Proof.* We can decompose $\hat{\epsilon}_i$ as

$$\hat{\epsilon}_i = \epsilon_i + (\hat{\epsilon}_i - \epsilon_i) = \epsilon_i + (y_i - X_i'\check{\beta}_n - y_i + X_i'\beta) = \epsilon_i + X_i'(\beta - \check{\beta}_n).$$

Squaring the left and right hand sides of this equation yields

$$\hat{\epsilon}_i^2 = \epsilon_i^2 + 2\epsilon_i X_i'(\beta - \check{\beta}_n) + (\beta - \check{\beta}_n)' X_i X_i' (\beta - \check{\beta}_n).$$

Plugging this into equation (57) we obtain

$$\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 X_i X_i' + \frac{2}{n}\sum_{i=1}^n \left[\epsilon_i X_i'(\beta - \check{\beta}_n)\right] X_i X_i' + \frac{1}{n}\sum_{i=1}^n \left[(\beta - \check{\beta}_n)' X_i X_i'(\beta - \check{\beta}_n)\right] X_i X_i'. \tag{58}$$

We next bound the second and third right hand side terms. For the second right hand side term, using properties of norms and the Cauchy – Schwarz inequality we have

$$\begin{aligned} R_{1,n} := \left| \frac{1}{n}\sum_{i=1}^n \left[\epsilon_i X_i'(\beta - \check{\beta}_n)\right] X_i X_i' \right| &\leq \frac{1}{n}\sum_{i=1}^n \left|\epsilon_i X_i'(\beta - \check{\beta}_n)\right| \|X_i X_i'\| \\ &\leq \|\beta - \check{\beta}_n\| \sum_{i=1}^n \frac{\|\epsilon_i X_i\|}{\sqrt{n}} \frac{\|X_i\|^2}{\sqrt{n}} \\ &\leq \|\beta - \check{\beta}_n\| \left(\frac{1}{n}\sum_{i=1}^n \|\epsilon_i X_i\|^2\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^n \|X_i\|^4\right)^{1/2}. \end{aligned} \tag{59}$$

For the third right hand side term, again using properties of norms and the Cauchy – Schwarz inequality,

$$\begin{aligned} R_{2,n} :=& \left| \frac{1}{n}\sum_{i=1}^n \left[(\beta - \check{\beta}_n)' X_i X_i'(\beta - \check{\beta}_n)\right] X_i X_i'. \right| \\ &\leq \sum_{i=1}^n \frac{\left|(\beta - \check{\beta}_n)' X_i X_i'(\beta - \check{\beta}_n)\right|}{\sqrt{n}} \frac{\|X_i\|^2}{\sqrt{n}} \\ &\leq \left(\frac{1}{n}\sum_{i=1}^n \left[(\beta - \check{\beta}_n)' X_i X_i'(\beta - \check{\beta}_n)\right]^2\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^n \|X_i\|^4\right)^{1/2} \\ &\leq \|\beta - \check{\beta}_n\|^2 \frac{1}{n}\sum_{i=1}^n \|X_i\|^4. \end{aligned} \tag{60}$$

(i) By the WLLN , $\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 X_i X_i' \xrightarrow{P} \mathbb{E}\left[\epsilon_i^2 X_i X_i'\right] = \mathrm{Var}(X_1\epsilon_1) = \Sigma$. Additionally, under our assumptions, $\frac{1}{n}\sum_{i=1}^n \|\epsilon_i X_i\|^2 \xrightarrow{P} \mathbb{E}\,\|\epsilon_1 X_1\|^2$ and $\frac{1}{n}\sum_{i=1}^n \|X_i\|^4 \xrightarrow{P} \mathbb{E}\,\|X_1\|^4$ by the WLLN. Combined with the consistency of $\check{\beta}_n$, Slutsky's Theorem and equations (59) and (60), it follows that $2R_{1,n} \xrightarrow{P} 0$ and $R_{2,n} \xrightarrow{P} 0$. These observations combined with (58) yield $\hat{\Sigma}_n \xrightarrow{P} \Sigma$.

(ii) By the WLLN ,

$$\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 X_i X_i' \xrightarrow{P} \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\epsilon_i^2 X_i X_i'\right] = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(X_i\epsilon_i) = \Sigma.$$

Additionally, under our assumptions, $\frac{1}{n}\sum_{i=1}^n \|\epsilon_i X_i\|^2 - \mathbb{E}\,\|\epsilon_i X_i\|^2 \xrightarrow{P} 0$ and $\frac{1}{n}\sum_{i=1}^n \|X_i\|^4 -$

$\mathbb{E}\,\|X_i\|^4 \xrightarrow{P} 0$ by the WLLN. In conjunction with (59) and (60), this implies

$$R_{1,n} \leq \|\beta - \check{\beta}_n\|(E_{1,n} + \tilde{C}) \qquad \text{and} \qquad R_{2,n} \leq \|\beta - \check{\beta}_n\|^2(E_{2,n} + \tilde{C}), \tag{61}$$

for $E_{i,n} \xrightarrow{P} 0$ $(i = 1, 2)$ and some $0 \leq \tilde{C} < \infty$ [Exercise]. Combined with the consistency of $\check{\beta}_n$ and Slutsky's Theorem it follows that $2R_{1,n} \xrightarrow{P} 0$ and $R_{2,n} \xrightarrow{P} 0$. These observations combined with (58) yield $\hat{\Sigma}_n \xrightarrow{P} \Sigma$. □

Corollary 4.1 [Estimator of OLS asymptotic variance]: *Suppose that Assumption 4.6 holds and $\check{\beta}_n$ (in (57)) is consistent for $\beta$ (i.e. $\check{\beta}_n \xrightarrow{P} \beta$). If also*

*(i) Assumptions 4.7, 4.8, 4.9, 4.10 and 4.16 hold; or*

*(ii) Assumptions 4.11, 4.12, 4.13, 4.14 and 4.17 hold,*

*then $M_n^{-1}\hat{\Sigma}_n M_n^{-1} \xrightarrow{P} M^{-1}\Sigma M^{-1}$.*

*Proof.* Combine Lemmas 4.3, 4.5 with the continuous mapping theorem and Slutsky's theorem. □

## 4.3 Hypothesis testing and confidence sets in linear regression

Example 4.9: Lets come back to our log wage equation from Example 4.1:

$$\log(wage) = \beta_0 + \beta_1 \times education + \beta_2 \times experience + \beta_3 \times experience^2.$$

We might want to test, for example, whether experience has an effect on wages. In this case our null hypothesis, $H_0$ would be $\beta_2 = \beta_3 = 0$, i.e. experience does not affect the (log) wage. We can represent this in the form

$$H_0: \quad R\beta = 0 \quad \text{vs. } H_1: \quad R\beta \neq 0 \qquad \text{with } R = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \qquad \triangle$$

### 4.3.1 Asymptotically justified tests and confidence sets

There are two basic approaches to constructing tests in the linear regression model. Firstly, the asymptotic route: based on Proposition 3.5 and the results of Section 4.2.3 we can build asymptotically justified tests for hypotheses like those in Example 4.9.

Example 4.10: Lets use the Wald statistic $W_n$ from Example 3.19 to test the hypothesis in Example 4.9. Using the estimator of $\Omega$ in Lemma (4.4) we obtain $W_n \approx 65.5$. Using the "heteroskedasticity robust" estimator (cf. equation (57)) we get $W_n \approx 61.9$. In either case we would reject $H_0$ at any conventional significance level (e.g. 0.01, 0.05 or 0.1). △

Confidence intervals for regression parameters can be constructed similarly, as in Example 3.22. These are routinely reported in statistical packages.

### 4.3.2 Finite sample tests in the normal linear regression model

Under the assumption of normality we imposed in Section 4.2.1: $y|X \sim \mathcal{N}(X\beta, \sigma^2 I)$, we can construct tests which are valid in finite sample. We will consider testing a hypothesis of the form $R\beta = r$ where $R$ is a *restriction matrix* of dimension $d \times K$, $\text{rank}(R) = d$ and $r \in \mathbb{R}^d$. Our test statistic will be the $F$ statistic:

$$F = \frac{(R\hat{\beta} - r)[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/d}{\hat{\sigma}^2},$$

with $\hat{\sigma}^2$ as in (43).

THEOREM 4.6: *Under the conditions of (iv) of Theorem 4.2, $F \sim \mathcal{F}(d, n-K)$ whenever $R\beta = r$, the $F$ distribution with $d$ and $n - K$ degrees of freedom.*

*In consequence, if $\kappa_\alpha$ is the $1 - \alpha$ quantile of the $\mathcal{F}(d, n - K)$ distribution, the test which rejects when $F > \kappa_\alpha$ is such that if $R\beta = r$, then*

$$P_\theta(F > k_\alpha) = \alpha.$$

*Proof.* Exercise. □

## 4.4 Prediction

To evaluate the quality of an estimator $\delta(X)$ of a parameter $\beta \in \mathbb{R}^K$, we use a *loss function*, $L(\beta, \delta(X))$, which is valued in $\mathbb{R}$ (typically in $[0, \infty)$).

We can also use loss functions to evaluate the quality of predictions. Here our loss function takes a slightly different form: if $y \in \mathbb{R}^n$ is the variable we want to predict and $f(X)$ is our predictor, then our loss function is $L(y, f(X))$. Note that here bowth arguments of $L$ are random. We shall focus on the quadratic loss; it is commmon in this setting to normalise the loss by the number of predictions being made:

$$L(y, f(X)) := \frac{\|y - f(X)\|^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(X)_i)^2 \tag{62}$$

and taking expectation over this loss function yields the mean squared error:

$$\mathbb{E}\, L(y, f(X)) = \frac{\mathbb{E}\, \|y - f(X)\|^2}{n} = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - f(X)_i)^2\right]. \tag{63}$$

### 4.4.1 MSE for prediction

We earlier described $\hat{y} = X\hat{\beta}$ as the *fitted values* based on the observations $X$ and the estimator $\hat{\beta}$; these could also be described as predictions. Specifically, in this context, we used $X$ to construct our estimate $\hat{\beta}$: as such any prediction $\hat{y}_i = X_i'\hat{\beta}$ is an *in-sample* prediction. We can also use our model for *out-of-sample* prediction: given a *new* observation $X_i$ (i.e. $i \notin \{1, \ldots, n\}$) we can predict the value of $y_i$ (based on our linear model) as $\hat{y}_i = X_i'\hat{\beta}$.

The prediction rule $\hat{y} = f(X) = X\hat{\beta}$ is based on a linear regression model and we may well consider other prediction rules. Under MSE, there is a *general* solution to the question "what is the best prediction rule $f(X)$?": it is the conditional expectation.

PROPOSITION 4.10 [Conditional expectation minimises MSE]: *Suppose that* $\mathbb{E}\, y^2 < \infty$*.* $f(X) = \mathbb{E}\,[y|X]$ *minimises* $\mathbb{E}[y - f(X)]^2$ *over all functions* $f(X)$ *with* $\mathbb{E}\, f(X)^2 < \infty$*.*

*Proof.* Add and subtract $\mathbb{E}[y|X]$ and expand the square to obtain

$$\begin{aligned}\mathbb{E}\,[y - f(X)]^2 &= \mathbb{E}\,[y - \mathbb{E}[y|X] + \mathbb{E}[y|X] - f(X)]^2 \\ &= \mathbb{E}\,[y - \mathbb{E}[y|X]]^2 + 2\,\mathbb{E}[(y - \mathbb{E}[y|X])(\mathbb{E}[y|X] - f(X))] + \mathbb{E}\,[\mathbb{E}[y|X] - f(X)]^2 \\ &= \mathbb{E}\,[y - \mathbb{E}[y|X]]^2 + \mathbb{E}\,[\mathbb{E}[y|X] - f(X)]^2\,,\end{aligned}$$

where the last equality follows from the law of iterated expectations [Exercise]. It follows that $\mathbb{E}[y - f(X)]^2$ is minimised when $f(X) = \mathbb{E}[y|X]$. □

As in the case with parameter estimation, the MSE for prediction also satisfies a decomposition into "bias" and "variance" terms.

PROPOSITION 4.11 [Bias – variance tradeoff for prediction]: *Suppose that* $\mathbb{E}\, y^2 < \infty$ *and* $\mathbb{E}\, f(X, \mathcal{D})^2 < \infty$*. Then*

$$\mathbb{E}\left[(f(X,\mathcal{D}) - \mathbb{E}[y|X])^2|X\right] = \mathbb{E}\left[(f(X,\mathcal{D}) - \mathbb{E}\,[f(X,\mathcal{D})|X])^2\middle|X\right] + (\mathbb{E}\,[f(X,\mathcal{D})|X] - \mathbb{E}[y|X])^2\,.$$

*Proof.* By addition and subtraction,

$$f(X,\mathcal{D}) - \mathbb{E}[y|X] = f(X,\mathcal{D}) - \mathbb{E}\,[f(X,\mathcal{D})|X] + \mathbb{E}\,[f(X,\mathcal{D})|X] - \mathbb{E}[y|X]$$

and so

$$\begin{aligned}(f(X,\mathcal{D}) - \mathbb{E}[y|X])^2 = [f(X,\mathcal{D}) &- \mathbb{E}\,[f(X,\mathcal{D})|X]]^2 + [\mathbb{E}\,[f(X,\mathcal{D})|X] - \mathbb{E}[y|X]]^2 \\ &+ 2\,[f(X,\mathcal{D}) - \mathbb{E}\,[f(X,\mathcal{D})|X]]\,[\mathbb{E}\,[f(X,\mathcal{D})|X] - \mathbb{E}[y|X]]\,.\end{aligned}$$

Since the last right hand side term in the preceding display is mean zero conditional on $X$, the result follows by taking conditional expectation. □

REMARK 4.1: *Propositions 4.10 and 4.11 are stated for the case where* $n = 1$ *prediction is being made. Since expectation is linear, analogous results hold for arbitrary* $n \in \mathbb{N}$ *[Exercise].*

Despite the result of Proposition 4.10, in practice computation of the conditional expectation is generally infeasible. As such we base predictions on models for the conditional expectation. The linear regression model (33) is one such model: assume $\mathbb{E}[\epsilon|X] = 0$ and take conditional expectations in (33) to obtain

$$\mathbb{E}[y|X] = \mathbb{E}[X\beta|X] + \mathbb{E}[\epsilon|X] = X\beta. \tag{64}$$

That is, the linear regression model (with the assumption $\mathbb{E}[\epsilon|X] = 0$) is a model for the conditional expectation. Specifically it supposes that the conditional expectation of $y$ on $X$ (also called the *regression function*) is a linear combination of the predictors $X$.

Note that *linearity* here refers to linearity in the parameters. Linear regression needn't be linear in the variables, as we may add non-linear transformations of given variables as covariates.

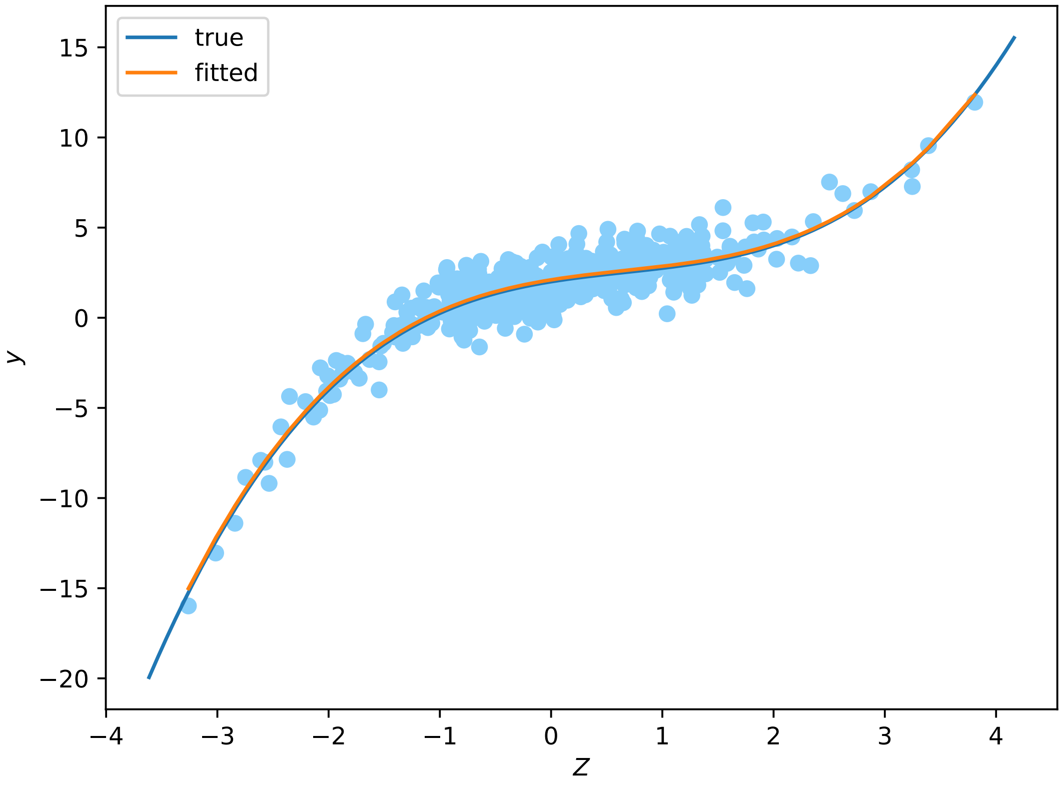
Example 4.11 [Linear in parameters]: Consider the regresion function

$$\mathbb{E}[y|X] = \beta_0 + \beta_1 Z + \beta_2 Z^2 + \beta_3 Z^3.$$

This is clearly linear in parameters (the $\beta_i$'s) but not in the variables.

Figure 21 below shows a linear regression fit to this model, based on 400 data points drawn according to the linear regression model implied by the above regresion function, with an (additive) error term $\epsilon \sim \mathcal{N}(0,1)$ independent of $Z \sim t(10)$. The true parameters are $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)' = (2, 1, -1/2, 1/4)'$.

Figure 21: OLS fit to non-linear regression function



△

**In − sample vs. out − of − sample evaluation**

A key distinction in prediction evaluation is whether our predictions are evauated *in-sample* or *out-of-sample*. The former implies that the prediction rule is evaluated based on the same data that has been used to create that rule; the latter evaluates the rule on new data. To illustrate the difference we will evaluate the in-sample and out-of-sample MSE of the least squares predictions (in the model (33)).

Example 4.12 [In-sample vs. out-of-sample MSE in linear regression]: Suppose that we observe an i.i.d. data sample $\mathcal{D} := (y_i, X_i)_{i=1}^n$ such that assumptions 4.1, 4.2, 4.3 and 4.4 hold. Consider the prediction rule given by $f(X) := X'\hat{\beta}$ where $\hat{\beta}$ is the OLS estimate based on $\mathcal{D}$.

Now, since (using Proposition 4.1)

$$\sum_{i=1}^n (y_i - X_i'\hat{\beta})^2 = y'M_X y = (\epsilon + X\beta)'M_X(\epsilon + X\beta) = \epsilon' M_X \epsilon,$$

by (44) and $\operatorname{tr} M_X = \operatorname{tr} I_n - \operatorname{tr} X'X(X'X)^{-1} = n - \operatorname{tr} I_K = n - K$, the in-sample MSE of this prediction rule is given by

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (y_i - f(X_i))^2\right] = \frac{1}{n}\mathbb{E}\left[\mathbb{E}\left[\epsilon' M_X \epsilon | X\right]\right] = \sigma^2 \frac{n-K}{n} = \sigma^2\left(1 - \frac{K}{n}\right).$$

This equation reveals that the in-sample fit improves (i.e. the in-sample MSE decreases) with the number of covariates, $K$.

The out-of-sample MSE is given by $\mathbb{E}\left[(y_{n+1} - X_{n+1}'\hat{\beta})^2\right]$ where $(y_{n+1}, X_{n+1}')$ is a new data point, not included in $\mathcal{D}$, but with the same distributional properties.[71] Letting $\tilde{X} = (X_1, \ldots, X_n, X_{n+1})'$, if we condition on $X$, we have

$$\mathbb{E}\left[(y_{n+1} - X_{n+1}'\hat{\beta})^2 | \tilde{X}\right] = \mathbb{E}\left[\epsilon_{n+1}^2 - (\hat{\beta} - \beta)' X_{n+1} X_{n+1}' (\hat{\beta} - \beta) - 2\epsilon_{n+1}(\hat{\beta} - \beta)' X_{n+1} | \tilde{X}\right].$$

By (conditional) homoskedasticity,[72]

$$\mathbb{E}\left[\epsilon_{n+1}(\hat{\beta} - \beta)' X_{n+1} | \tilde{X}\right] = \mathbb{E}\left[\epsilon_{n+1}\epsilon' X(X'X)^{-1} X_{n+1} | \tilde{X}\right] = \mathbb{E}\left[\epsilon_{n+1}\epsilon' | \tilde{X}\right] X(X'X)^{-1} = 0,$$

and therefore,

$$\begin{aligned}
\mathbb{E}\left[(y_{n+1} - X_{n+1}'\hat{\beta})^2 | \tilde{X}\right] &= \mathbb{E}\left[\epsilon_{n+1}^2 - (\hat{\beta} - \beta)' X_{n+1} X_{n+1}' (\hat{\beta} - \beta) | \tilde{X}\right] \\
&= \mathbb{E}\left[\epsilon_{n+1}^2 | \tilde{X}\right] + \operatorname{tr}\left(\mathbb{E}\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \middle| \tilde{X}\right] X_{n+1} X_{n+1}'\right) \\
&= \sigma^2 + \operatorname{tr}\left(\mathbb{E}\left[(X'X)^{-1} X' \epsilon \epsilon' X (X'X)^{-1} \middle| \tilde{X}\right] X_{n+1} X_{n+1}'\right) \\
&= \sigma^2 \left(1 + \operatorname{tr}(X'X)^{-1} X_{n+1} X_{n+1}'\right).
\end{aligned}$$

Taking the expectation we then have

$$\mathbb{E}\left[(y_{n+1} - X_{n+1}'\hat{\beta})^2\right] = \sigma^2\left(1 + \frac{1}{n}\operatorname{tr}\left(\mathbb{E}\left[\left(\frac{1}{n}X'X\right)^{-1}\right]\mathbb{E}[X_{n+1}X_{n+1}']\right)\right)$$

Evaluation of the unconditional expectation here is not as straightforward as in the in-sample

[71] That is to say, the same assumptions would hold if the data sample also included $(y_{n+1}, X_{n+1}')$.

[72] Note that here $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$.

case, and we will rely on asymptotics. In our setting we have that $(\frac{1}{n}X'X)^{-1} \xrightarrow{P} \mathbb{E}[X_1X_1']^{-1} = \mathbb{E}[X_{n+1}X_{n+1}']^{-1}$ [Exercise]. Moreover, since the data are i.i.d., the sequence is UI [Exercise] and therefore, we have that $\mathbb{E}(\frac{1}{n}X'X)^{-1} \to \mathbb{E}[X_1X_1']^{-1}$. Therefore, for $r_n \to 0$, and $C := \operatorname{tr}\mathbb{E}[X_1X_1']$

$$\begin{aligned}
&\mathbb{E}\left[(y_{n+1} - X_{n+1}'\hat{\beta})^2\right] \\
&\quad = \sigma^2\left(1 + \frac{1}{n}\operatorname{tr}\left(\left[\mathbb{E}\left[\left(\frac{1}{n}X'X\right)^{-1}\right] - \mathbb{E}[X_1X_1']^{-1} + \mathbb{E}[X_1X_1']^{-1}\right]\mathbb{E}[X_1X_1']\right)\right) \\
&\quad = \sigma^2\left(1 + \frac{1}{n}\operatorname{tr}\left(r_n\,\mathbb{E}[X_1X_1']\right) + \frac{1}{n}\operatorname{tr}\left(\mathbb{E}[X_1X_1']^{-1}\,\mathbb{E}[X_1X_1']\right)\right) \\
&\quad = \sigma^2\left(1 + \frac{1}{n}\operatorname{tr}\left(r_n\,\mathbb{E}[X_1X_1']\right) + \frac{1}{n}\operatorname{tr}\left(I_K\right)\right) \\
&\quad = \sigma^2\left(1 + \frac{K}{n} + \frac{Cr_n}{n}\right).
\end{aligned}$$

This expression demonstrates that, unlike the in-sample MSE, the out-of-sample MSE increases with the number of covariates $K$ (up to a term which is asymptotically smaller than $1/n$). △

In general we *should* measure performance out-of-sample. The reason is hinted at in the previous example: by simply adding enough free parameters we can continually reduce the error in-sample, but this will typically cause out-of-sample performance to deteriorate, as it improves the fit to the *specific sample* ("overfitting") rather than the data generating process. In practice this may mean splitting our data into an estimation subsample and a "test" subsample.[73]

## 4.5 Regularisation

One general approach to try and reduce overfitting is to *regularise* or *penalise*. This entails adding to our objective function (e.g. (35)) a penalty function which penalises complex models. An important example is to add a quadratic penalty term: this is called "ridge regression" and minimises the objective function

$$\check{S}_\lambda(\beta) := S(\beta) + \lambda\|\beta\|^2 = \|y - X\beta\|^2 + \lambda\|\beta\|^2. \tag{65}$$

The solution to this problem is known in closed form, and depends on the value of the *penalty parameter* $\lambda \in [0, \infty)$:

$$\check{\beta}_\lambda := (X'X + \lambda I)^{-1}X'y. \tag{66}$$

Note that if $\lambda = 0$ in (65) then the objective function is identical to (35) and the solution (66) coincides with the OLS estimator $\hat{\beta}$ in (34), provided $X$ has full column rank.

Whenever the penalty parameter $\lambda$ is positive, the matrix inverse in (66) exists [Exercise]. One implication of this is that the ridge regression estimator $\check{\beta}_\lambda$ exists even when $K > n$ (i.e. when we have more covariates than observations).[74]

[73] In statistical (machine) learning these are often refered to as the "training" and "test" sets.

[74] The OLS estimator does not exist when $K > n$, since $X'X$ is not full rank and hence not invertible.

Proposition 4.12: *If $\lambda > 0$, $\check{\beta}_\lambda$ in (66) satisfies*

$$\check{\beta}_\lambda = \arg\min_{\beta \in \mathbb{R}^K} \check{S}_\lambda(\beta).$$

*Proof.* Note that $\check{S}_\lambda(\beta) = (y - X\beta)'(y - X\beta) + \lambda\beta'\beta$ and so

$$\nabla_\beta \check{S}_\lambda(\beta) = -2X'(y - X\beta) + 2\lambda\beta.$$

Therefore the first-order condition $\nabla_\beta \check{S}_\lambda(\beta) = 0$ is satisfied when

$$X'y = (X'X + \lambda I)\beta.$$

Premultiply both sides by $(X'X + \lambda I)^{-1}$ (which always exists for $\lambda > 0$ [Exercise]) to obtain (66). Taking a second derivative yields that the Hessian is $2(X'X + \lambda I)$, which is positive definite (for all $\beta$). Hence the objective is (globally) convex and so $\check{\beta}_\lambda$ is a global minimum. □

Remark 4.2: *The case where $\lambda = 0$ and $X$ is of full rank is covered by Proposition 4.1.*

### 4.5.1 Shrinkage

The ridge regression estimator (66) is an example of a *shrinkage* estimator. The addition of the penalty or regularisation term leads to a solution with a smaller $\|\beta\|$ that would otherwise be the case (i.e. than the OLS estimator).

Example 4.13 [Ridge estimator is a shrinkage estimator]: We will re-express the ridge and OLS estimators using the singular value decomposition (SVD) of $X$. Specifically supposing that $q = \min(n, K)$, (a version of) the SVD allows us to write $X = UDV'$ where $U, V$ are $n \times n$ and $K \times K$ orthogonal matrices respectively and $D$ is a $n \times K$ matrix, with the *singular values* of $X$ on its "diagonal", i.e.

$$D_q = \begin{bmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_q \end{bmatrix} \quad \text{and} \quad \begin{cases} D = D_q & \text{if } n = K \\ D = \begin{bmatrix} D_q & 0 \end{bmatrix} & \text{if } n < K \\ D = \begin{bmatrix} D_q \\ 0 \end{bmatrix} & \text{if } n > K \end{cases},$$

where $d_1 \geq d_2 \geq \cdots \geq d_r \geq d_{r+1} = \cdots d_q = 0$, for $r = \text{rank}(X)$.[75]

Supposing that $n \geq K = \text{rank}(X)$ so that the OLS estimator exists, the SVD and basic

[75] The singular values of a $n \times K$ matrix $X$ are $d_1, \ldots, d_q$, where $d_1, \ldots, d_r$ are the (decreasingly ordered) non-zero eigenvalues of $XX'$ (or, equivalently, those of $X'X$).

manipulations give

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'y \\
&= (VD'U'UDV')^{-1}VD'U'y \\
&= (VD_q^2V')^{-1}VD'U'y \\
&= VD_q^{-2}V'VD'U'y \\
&= V\left[\begin{smallmatrix} D_q^{-1} & 0 \end{smallmatrix}\right]U'y.
\end{aligned}$$

Therefore, the $j$-th component of $\hat{\beta}$ has the form $\sum_{k=1}^{K} V_{jk}d_k^{-1}u_k'y$, where $u_k$ are the $k$-th columns of $U$ and $d_k$ the $k$-th diagonal element of $D_q$ (i.e. the $k$-th singular value of $X$).

We can do similar with the ridge estimator:

$$\begin{aligned}
\check{\beta}_\lambda &= (X'X + \lambda I)^{-1}X'y \\
&= (VD'U'UDV' + \lambda VV')^{-1}VD'U'y \\
&= (V[D_q^2 + \lambda I]V')^{-1}VD'U'y \\
&= V[D_q^2 + \lambda I]^{-1}V'VD'U'y \\
&= V[D_q^2 + \lambda I]^{-1}\left[\begin{smallmatrix} D_q & 0 \end{smallmatrix}\right]U'y.
\end{aligned}$$

Therefore, the $j$-th component of $\check{\beta}_\gamma$ has the form $\sum_{k=1}^{K} V_{jk}\frac{d_k}{d_k^2+\lambda}u_k'y$. Whenever $\lambda > 0$, the factor $\frac{d_k}{d_k^2+\lambda}$ is smaller than the corresponding factor $(d_k^{-1})$ in the OLS estimate.

We can also pre-multiply each estimator by X to obtain their corresponding predictions for $y$. Using the SVD once more, we have

$$\hat{y} = X\hat{\beta} = UDV'V\left[\begin{smallmatrix} D_q^{-1} & 0 \end{smallmatrix}\right]U'y = UU'y,$$

whilst for the ridge estimator

$$\check{y}_\lambda = X\check{\beta}_\lambda = UDV'V[D_q^2 + \lambda I]^{-1}D'U'y = UD[D_q^2 + \lambda I]^{-1}DU'y = \sum_{k=1}^{K} u_k\frac{d_k^2}{d_k^2 + \lambda}u_k'y.$$

The OLS estimator computes its predictions based on the orthonormal basis given by (the columns of) $U$. Ridge regression does the same, but weights each direction by the shrinkage factor $\frac{d_k^2}{d_k^2+\lambda} \leq 1$ (strictly whenever $\lambda > 0$).

The shrinkage factor $\frac{d_k^2}{d_k^2+\lambda}$ depends on the singular values of the matrix $X$ and its relation to the $\lambda$ parameter. As such different scaling of the matrix $X$ will lead to a different amount of shrinkage given the same $\lambda$ parameter. As such, we typically (centre and) scale the data prior to using ridge regression (or other shrinkage methods). △

The next Lemma calculates the bias and variance of $\check{\beta}_\lambda$.

Lemma 4.6 [Moments of ridge estimator]: *If Assumptions 4.1, 4.2 and 4.3 hold,*

$$\mathbb{E}\left[\check{\beta}_\lambda \middle| X\right] = \beta - \lambda(X'X + \lambda I_K)^{-1}\beta.$$

*If additionally* $\mathbb{E}[\epsilon\epsilon']$ *exists, then*

$$\text{Var}\left[\check{\beta}_\lambda \middle| X\right] = (X'X + \lambda I_K)^{-1}X'\,\mathbb{E}\left[\epsilon\epsilon' \middle| X\right] X(X'X + \lambda I_K)^{-1}.$$

*Proof.* By substitution of (33) into (66) we have

$$\check{\beta}_\lambda = (X'X + \lambda I_K)^{-1}X'y = (X'X + \lambda I_K)^{-1}X'X\beta + (X'X + \lambda I_K)^{-1}X'\epsilon.$$

Therefore, by Assumption 4.2

$$\mathbb{E}\left[\check{\beta}_\lambda \middle| X\right] = \mathbb{E}\left[(X'X + \lambda I_K)^{-1}X'y \middle| X\right] = \mathbb{E}\left[(X'X + \lambda I_K)^{-1}X'X\beta \middle| X\right].$$

It can be shown that $I_K - \lambda(X'X + \lambda I_K)^{-1} = (X'X + \lambda I_K)^{-1}X'X$ [Exercise]. As a result

$$\mathbb{E}\left[\check{\beta}_\lambda \middle| X\right] = \mathbb{E}\left[\beta - \lambda(X'X + \lambda I_K)^{-1}\beta \middle| X\right] = \beta - \lambda(X'X + \lambda I_K)^{-1}\beta.$$

For the variance, we have

$$\begin{aligned}
\text{Var}\left[\check{\beta}_\lambda \middle| X\right] &= \text{Var}\left[(X'X + \lambda I_K)^{-1}X'y \middle| X\right] \\
&= (X'X + \lambda I_K)^{-1}X'\text{Var}\left[X\beta + \epsilon | X\right] X(X'X + \lambda I_K)^{-1} \\
&= (X'X + \lambda I_K)^{-1}X'\text{Var}\left[\epsilon | X\right] X(X'X + \lambda I_K)^{-1} \\
&= (X'X + \lambda I_K)^{-1}X'\,\mathbb{E}\left[\epsilon\epsilon' \middle| X\right] X(X'X + \lambda I_K)^{-1}.
\end{aligned}$$

□

Remark 4.3: *In the case of conditional homoskedasticity (i.e. if Assumption 4.4 holds), the expression for the (conditional) variance simplifies to*

$$\text{Var}\left[\check{\beta}_\lambda \middle| X\right] = \sigma^2(X'X + \lambda I_K)^{-1}X'X(X'X + \lambda I_K)^{-1}.$$

*This is smaller than the conditional variance of the OLS estimator. Letting* $W_\lambda := (X'X + \lambda I_K)^{-1}(X'X)$, *we have*

$$\begin{aligned}
\text{Var}\left[\hat{\beta} \middle| X\right] - \text{Var}\left[\check{\beta}_\lambda \middle| X\right] &= \sigma^2\left[(X'X)^{-1} - W_\lambda(X'X)^{-1}W_\lambda'\right] \\
&= \sigma^2 W_\lambda\left[(I + \lambda(X'X)^{-1})(X'X)^{-1}(I + \lambda(X'X)^{-1})' - (X'X)^{-1}\right] W_\lambda' \\
&= \sigma^2 W_\lambda\left[2\lambda(X'X)^{-2} + \lambda^2(X'X)^{-3}\right] W_\lambda' \\
&= \sigma^2(X'X + \lambda I_K)^{-1}\left[2\lambda I_K + \lambda^2(X'X)^{-1}\right](X'X + \lambda I_K)^{-1}.
\end{aligned}$$

*The last right hand side term is the product of 3 positive semi-definite matrices and hence positive semi-definite.*

Corollary 4.2 [MSE of ridge estimator]: *Suppose the required hypotheses of (both parts of) Lemma 4.6 hold. Then, whenever the expectations below exist*

$$\begin{aligned}\mathbb{E}\left[\|\beta-\check{\beta}_\lambda\|^2\right] &= \mathbb{E}\operatorname{tr}(X'X+\lambda I_K)^{-1}X'\,\mathbb{E}\left[\epsilon\epsilon'\middle|X\right]X(X'X+\lambda I_K)^{-1}+\|\lambda\,\mathbb{E}\left[(X'X+\lambda I_K)^{-1}\right]\beta\|^2\\ &= \mathbb{E}\operatorname{tr}X'\,\mathbb{E}\left[\epsilon\epsilon'\middle|X\right]X(X'X+\lambda I_K)^{-2}+\|\lambda\,\mathbb{E}\left[(X'X+\lambda I_K)^{-1}\right]\beta\|^2.\end{aligned}$$

*Proof.* Apply Lemma 4.6 with (15) and the law of iterated expectations. □

Lemma 4.6 shows that the ridge estimator is biased, whilst – in the conditionally homoskedastic case – Remark 4.3 notes that its variance is smaller than that of the OLS estimator. Both the bias and variance depend on the penalty parameter $\lambda$: as $\lambda$ increases the (conditional) variance decreases and the (conditional) bias increases [Exercise]. It is therefore possible that for certain choices of $\lambda$ the MSE of the ridge estimator is lower than that of the OLS estimator. The next proposition demonstrates that this holds under Assumptions 4.1 – 4.4.

Proposition 4.13 [Ridge estimator can MSE dominate OLS]: *Suppose that Assumptions 4.1, 4.2, 4.3 and 4.4 hold. Then, for all* $\lambda\in\Lambda:=(0,2\sigma^2\|\beta\|^{-2})$,

$$\mathbb{E}\left[\|\beta-\check{\beta}_\lambda\|^2\middle|X\right]<\mathbb{E}\left[\|\beta-\hat{\beta}\|^2\middle|X\right],$$

*and hence also* $\mathbb{E}\left[\|\beta-\check{\beta}_\lambda\|^2\right]<\mathbb{E}\left[\|\beta-\hat{\beta}\|^2\right]$.

*Proof.* By Remark 4.3 we have

$$\operatorname{Var}\left[\hat{\beta}\middle|X\right]-\operatorname{Var}\left[\check{\beta}_\lambda\middle|X\right]=\sigma^2(X'X+\lambda I_K)^{-1}\left[2\lambda I_K+\lambda^2(X'X)^{-1}\right](X'X+\lambda I_K)^{-1}.$$

The (conditional) bias of the Ridge estimator is (Lemma 4.6)

$$\mathbb{E}\left[\check{\beta}_\lambda-\beta\middle|X\right]=-\lambda(X'X+\lambda I_K)^{-1}\beta.$$

Hence letting $\hat{M}:=\mathbb{E}\left[(\beta-\hat{\beta})(\beta-\hat{\beta})'\middle|X\right]$ and $\check{M}_\lambda:=\mathbb{E}\left[(\beta-\check{\beta}_\lambda)(\beta-\check{\beta}_\lambda)'\middle|X\right]$ we have by Proposition 4.2 and the displays above,

$$\hat{M}-\check{M}_\lambda=\sigma^2\lambda(X'X+\lambda I_K)^{-1}\left[2I_K+\lambda(X'X)^{-1}-\frac{\lambda}{\sigma^2}\beta\beta'\right](X'X+\lambda I_K)^{-1}.$$

This is positive definite if and only if $\lambda>0$ and $2\sigma^2I_K+\lambda\sigma^2(X'X)^{-1}-\lambda\beta\beta'$ is positive definite. $2\sigma^2I_K-\lambda\beta\beta'$ is positive definite when $\lambda<2\sigma^2\|\beta\|^{-2}$.[76] Since $\sigma\lambda(X'X)^{-1}$ is positive semi-definite, it follows that $\hat{M}-\check{M}_\lambda$ is positive definite for all $\lambda\in\Lambda=(0,2\sigma^2\|\beta\|^{-2})$. To conclude

[76] For a positive definite $m\times m$ matrix $A$, a $b\in\mathbb{R}^m$ with $b\neq0$ and a $c\in\mathbb{R}$, $cA-bb'$ is positive (semi-)definite if and only if $b'A^{-1}b$ is less than (or equal to) $c$. This follows from the fact that for any $x\neq0$

$$x'(cA-bb')x>0\iff\frac{c}{b'A^{-1}b}>\frac{(x'b)^2}{x'Axb'A^{-1}b}.$$

The right hand side attains its maximum value of 1 at $x=\pm A^{-1}b$ [Exercise] and hence the positive definiteness of $cA-bb'$ follows provided $c/(b'A^{-1}b)>1$, which is equivalent to the stated condition.

note that

$$\|\beta - b\|^2 = (\beta - b)'(\beta - b) = \operatorname{tr}(\beta - b)'(\beta - b) = \operatorname{tr}(\beta - b)(\beta - b)'.$$

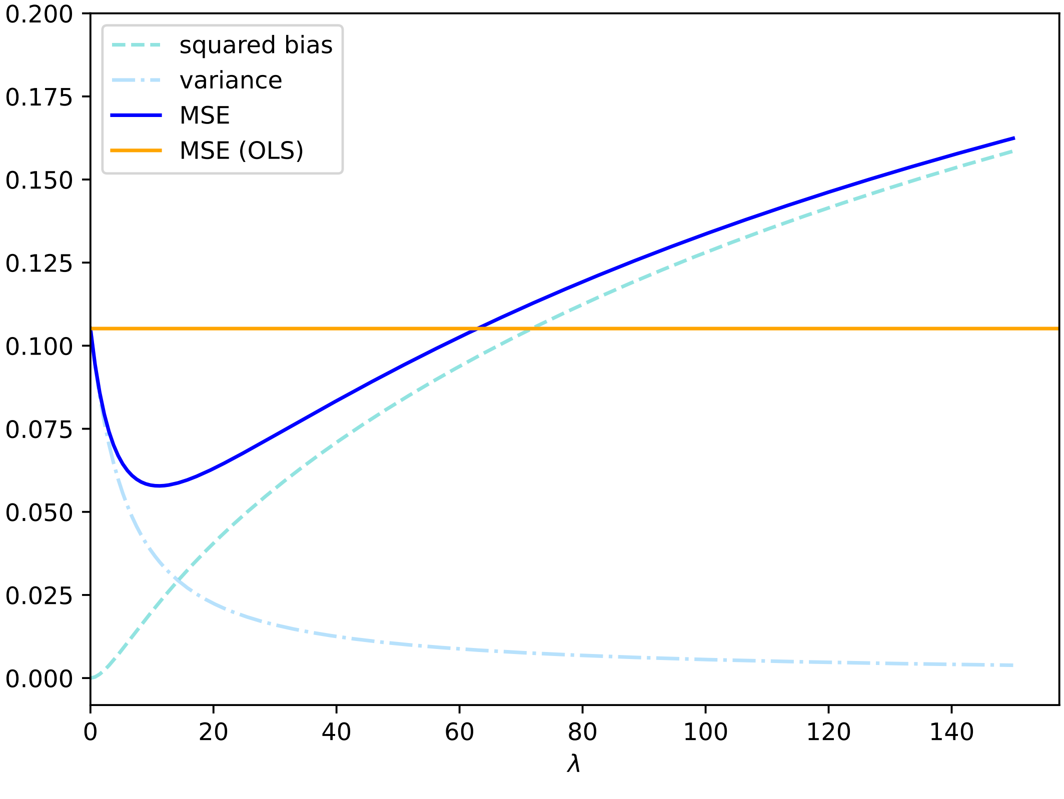
Hence, by the linearity of the trace,

$$\mathbb{E}\left[\|\beta - \hat{\beta}\|^2 \middle| X\right] - \mathbb{E}\left[\|\beta - \check{\beta}_\lambda\|^2 \middle| X\right] = \operatorname{tr} \hat{M} - \check{M}_\lambda > 0,$$

for all $\lambda \in \Lambda$, where the inequality follows since the trace of a matrix is the sum of its eigenvalues. The last claim follows by taking expectations on both sides. □

Figure 22 demonstrates the result of Proposition 4.13 for some artificial data. As can be seen, for small enough $\lambda$, the reduction in variance from using the ridge estimator dominates the contribution to the MSE from the (squared) bias.

Figure 22: (Conditional) MSE for ridge and OLS estimators



*Note:* Blue lines depict the (conditional) squared bias, variance and mean squared error of the ridge estimator $\check{\beta}_\lambda$ as $\lambda$ varies. The orange line is the (conditional) mean squared error of the OLS estimator $\hat{\beta}$.

We can use the same arguments we used to prove Proposition 4.13 to show a similar result for prediction.

Corollary 4.3: *Suppose that $(y_i, X_i)_{i=1}^{n+1}$ is such that assumptions 4.1, 4.2, 4.3 and 4.4 hold and $(y_{n+1}, X_{n+1})$ is independent of $\mathcal{D} := (y_i, X_i)_{i=1}^{n}$. Let $\hat{\beta}$ be the OLS estimator based on $\mathcal{D}$ and $\check{\beta}_\lambda$ the ridge estimator based on $\mathcal{D}$. Then, if $\lambda \in \Lambda := (0, 2\sigma^2\|\beta\|^{-2})$, $\mathbb{E}[(y_{n+1} - X_{n+1}'\hat{\beta})^2] <$*

$\mathbb{E}[(y_{n+1} - X'_{n+1}\check{\beta}_\lambda)^2]$.

*Proof.* Since $\mathbb{E}[y_{n+1}|X_{n+1}] = X'_{n+1}\beta$,

$$\begin{aligned}(y_{n+1} - X'_{n+1}b)^2 &= (y_{n+1} - \mathbb{E}[y_{n+1}|X_{n+1}] + \mathbb{E}[y_{n+1}|X_{n+1}] - X'_{n+1}b)^2 \\ &= \epsilon_{n+1}^2 + 2\epsilon_{n+1}(X'_{n+1}[\beta - b]) + (\beta - b)'X_{n+1}X'_{n+1}(\beta - b).\end{aligned}$$

By the assumed independence and Assumption 4.2,

$$\begin{aligned}\mathbb{E}\left[2\epsilon_{n+1}X'_{n+1}(\hat{\beta} - \check{\beta}_\lambda)\middle|X_{n+1}\right] &= \mathbb{E}\left[2\epsilon_{n+1}(\hat{\beta} - \check{\beta}_\lambda)'\middle|X_{n+1}\right]X_{n+1} \\ &= 2\,\mathbb{E}\left[\epsilon_{n+1}|X_{n+1}\right]\mathbb{E}\left[(\hat{\beta} - \check{\beta}_\lambda)\right]'X_{n+1} \\ &= 0.\end{aligned}$$

Letting $X = (X_i)_{i=1}^n$, it follows that, again using the independence

$$\begin{aligned}&\mathbb{E}[(y_{n+1} - X'_{n+1}\check{\beta}_\lambda)^2|X_{n+1}] - \mathbb{E}[(y_{n+1} - X'_{n+1}\hat{\beta})^2|X_{n+1}] \\ &\quad = \mathbb{E}\left[(\beta - \tilde{\beta}_\lambda)'X_{n+1}X'_{n+1}(\beta - \tilde{\beta}_\lambda)\middle|X_{n+1}\right] - \mathbb{E}\left[(\beta - \hat{\beta})'X_{n+1}X'_{n+1}(\beta - \hat{\beta})\middle|X_{n+1}\right] \\ &\qquad + \mathbb{E}\left[2\epsilon_{n+1}X'_{n+1}(\hat{\beta} - \check{\beta}_\lambda)|X_{n+1}\right] \\ &\quad = \operatorname{tr}\left(X_{n+1}X'_{n+1}\,\mathbb{E}\left[\mathbb{E}\left[(\beta - \check{\beta}_\lambda)(\beta - \check{\beta}_\lambda)'\middle|X\right] - \mathbb{E}\left[(\beta - \hat{\beta})(\beta - \hat{\beta})'\middle|X\right]\right]\right) \\ &\quad = \operatorname{tr}\left(X_{n+1}X'_{n+1}\,\mathbb{E}\left[\check{M}_\lambda - \hat{M}\right]\right)\end{aligned}$$

where $\hat{M} := \mathbb{E}\left[(\beta - \hat{\beta})(\beta - \hat{\beta})'\middle|X\right]$ and $\check{M}_\lambda := \mathbb{E}\left[(\beta - \check{\beta}_\lambda)(\beta - \check{\beta}_\lambda)'\middle|X\right]$ as in the proof of Proposition 4.13. There it was shown that $\check{M}_\lambda - \hat{M}$ is positive definite for $\lambda \in \Lambda$; the same holds for $\mathbb{E}\left[\check{M}_\lambda - \hat{M}\right]$ by taking expectations. Taking expectations once more we obtain

$$\mathbb{E}[(y_{n+1} - X'_{n+1}\check{\beta}_\lambda)^2] - \mathbb{E}[(y_{n+1} - X'_{n+1}\hat{\beta})^2] = \operatorname{tr}\left(\mathbb{E}\left[X_{n+1}X'_{n+1}\right]\mathbb{E}\left[\check{M}_\lambda - \hat{M}\right]\right),$$

which is positive [Exercise]. □

### 4.5.2 Cross validation

As Figure 22 demonstrates, in order for ridge regression to perform well we need a way to choose the penalty parameter $\lambda$.[77] When the goal is prediction, one common approach to the choice of the penalty parameter is the use of *cross validation*, which aims to estimate the prediction error using a certain prediction rule.

Cross validation entails splitting the data into training (i.e. estimation) and test (i.e. evaluation) data sets. The estimator is then "trained" (i.e. the parameters are estimated) on the training data and "tested" (i.e. its performance is measured) on the test data. A common approach is $k$-fold cross validation. Here the data is (possibly randomly) split into $k$ (approximately) equally sized parts, $k - 1$ of these are used to train the model and the last part is used

[77] This is also true for other regularisation methods.

to test it. In particular, if observation $i$ belongs to the $j$-th part, the error is measured according to some loss function $L(y_i, f_\lambda^{-j}(X_i))$ where the prediction rule $f_\lambda^{-j}$ is estimated on the $k-1$ parts of the data which are not the $j$-th part. We then average the loss over all observations and choose $\lambda$ (from some pre-determined set $\Lambda$) as that which yields the smallest (average) loss.

Formally, let $\varrho : \{1, \ldots, n\} \to \{1, \ldots, k\}$ define the partitioning (i.e. so that $\varrho(i) = j$ if observation $i$ belongs to part $j$ of the data). Our estimate of the prediction error is

$$CV(\lambda) := \frac{1}{n} \sum_{i=1}^{n} L(y_i, f_\lambda^{-\varrho(i)}(X_i)),$$

and our chosen value of $\lambda$ is

$$\lambda_\star = \arg\min_{\lambda \in \Lambda} CV(\lambda).$$

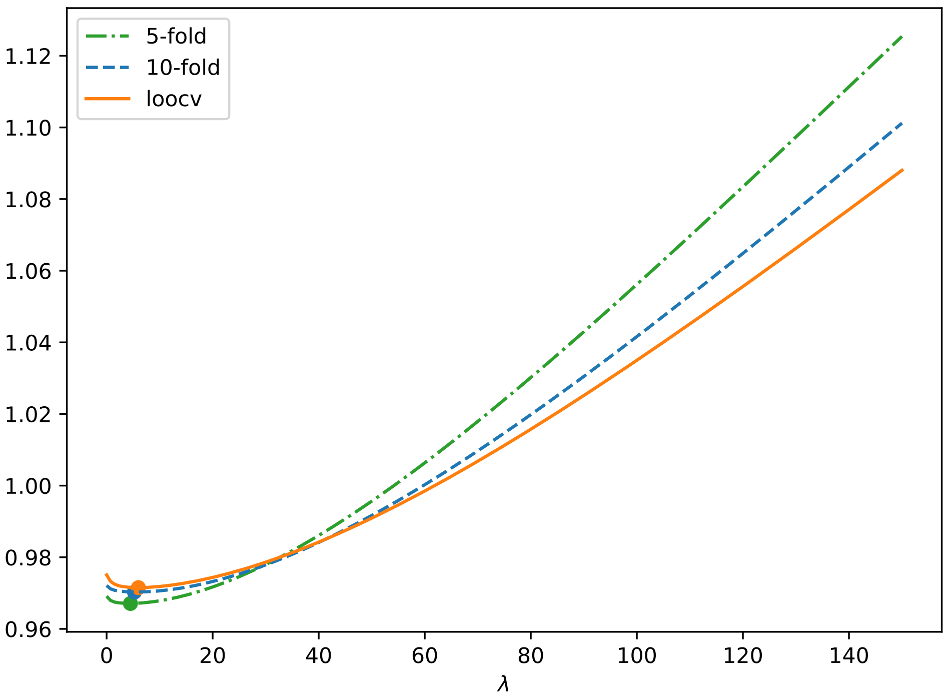
Specialising to the case of ridge regresion and mean squared error we have

$$CV(\lambda) := \frac{1}{n} \sum_{i=1}^{n} \left( y_i - X_i' \check{\beta}_\lambda^{-\varrho(i)} \right)^2.$$

The choice of $k = n$ is called *leave-one-out-cross-validation.* Here for each index $i$ all observations except that with index $i$ is used to estimate the prediction rule.

The figure below plots curves of the 5-fold, 10-fold and LOOCV criterion for different values of $\lambda$, for some artificial data drawn from a linear model, with ridge regression used to form the prediction rule.

FIGURE 23: $k$-FOLD AND LEAVE-ONE-OUT CROSS VALIDATION



*Note:* Dot indicates the minimum value of the criterion along each curve.

### 4.5.3 Other regularisation methods

There are many other regularisation methods which can be used in the linear regression setting. A general class of approaches is based on generalising equation (65) to[78]

$$\check{S}^p_\lambda(\beta) := S(\beta) + \lambda \|\beta\|^p = \|y - X\beta\|^2 + \lambda\|\beta\|_p^p, \quad \text{for } p \geq 1.$$

Taking $p = 2$ corresponds to ridge regression, though other $p \geq 1$ can be used and minimising $\check{S}^p_\lambda(\beta)$ for different $p$ values yields regularised estimators with different properties. For example, $p = 1$ corresponds to *LASSO* regression. This is a particularly special case since the Lasso regularisation not only shrinks the parameter estimates but may also "select" variables, by setting some estimates exactly to zero.[79] Unlike ridge regression, the Lasso estimator (or estimators minimising $\check{S}^p_\lambda(\beta)$ for other $p \neq 2$) do not have closed form solutions. Nevertheless they are computable numerically and high quality implementations exist for many of these estimators in, for example, statsmodels (e.g. `sm.OLS.fit_regularized`) and scikit-learn (e.g. `sklearn.linear_model.Lasso`).

## 4.6 Generalised linear models

The linear modelling approaches we have studied so far in this section are powerful and widely applicable. There are, however, a number of frequently occuring cases where they are not suitable.

Example 4.14 [Binary response]: Suppose you observe data $(y_i, X_i)_{i=1}^n$ where $y_i \in \{0, 1\}$. Typically $y_i$ measures whether an event occured or not.

As an example we shall consider a question related to our log wage equation from above, where instead of being interested in the (log) wage, we shall be interested in how education and experience affects whether the individual is employed in the first place.

Consider using our linear model framework. Then

$$\mathbb{E}[y_i|X_i] = X_i'\beta.$$

But, in this context, $\mathbb{E}[y_i|X_i] = P(y_i = 1|X_i)$, the conditional probability of the event occuring (e.g. the individual being employed). Using our linear modelling framework from before is possible in this scenario but it has some drawbacks.

(i) Predicted (conditional) "probabilities" $P(y_i = 1|X_i)$ need not lie in $[0, 1]$.

- How can we interpret a negative predicted probability of being employed?

(ii) The linear model imposes that a unit change in any variable $X_{i,k}$ has the same effect on the predicted probability for all values of $X_{i,k}$.

[78] Here $\|x\|_p$ is the $p$-norm on $\mathbb{R}^K$: $\|x\|_p := (\sum_{k=1}^K |x_k|^p)^{1/p}$. The case $p = 2$ is the "standard" (Euclidean) norm.

[79] See e.g. Section 3.4.2 in [6] for more details on the Lasso.

- In our employment example this seems unreasonable: we would not expect having 32 years of experience vs. 30 to affect the probability of employment nearly as much as the change from 2 to 0 years of experience.

△

Example 4.15 [Count response]: Suppose you observe data $(y_i, X_i)_{i=1}^n$ where $y_i \in \mathbb{N} \cup \{0\}$, i.e. $y_i$ measures the number of times an event happens.

Using a linear model in this situation is again possible, but suffers from similar drawbacks to the binary case. Here, using a linear model,

$$\mathbb{E}[y_i|X_i] = X_i'\beta$$

may be negative, which is makes little sense when, by the definition of $y_i$, $\mathbb{E}[y_i|X_i] \geq 0$. △

In order to motivate the approach used to handle these kinds of situations, lets recap the normal linear regression model from part (iv) of Theorem 4.2. In this situation we supposed that, conditionally on the $X_i$ variables

$$y_i|X_i \sim \mathcal{N}(X_i'\beta, \sigma^2). \tag{67}$$

Beyond $(y_i, X_i)_{i=1}^n$ forming a random sample, there are three key elements here:

(i) The probability distribution (family) of the response variable conditional on $X_i$: $y_i$ is (conditionally) normally distributed

(ii) The linear combination $\theta(X_i) = X_i'\beta$

(iii) A specification of how $\theta(X_i)$ relates to the (conditional) mean of $y_i$: $\mu(X_i) := \mathbb{E}[y_i|X_i] = \theta(X_i)$.

Generalised linear models allow (i) and (iii) to be changed. The "linear" in the name is because (ii) is retained.

### The distributional specification

The distribution of the response (conditional on the covariates) may come from an exponential family distribution other than the normal. In particular, recall that an exponential family in canonical form has a density / mass function:

$$p_\eta(z) := \exp\left(z'\eta - A(\eta)\right) h(z), \tag{68}$$

for some known functions $A$ and $h \geq 0$.[80]

The moments of this distribution are determined by the function $A$. We prove the following result which applies to the canonical exponential families.

[80] This is actually a subclass of the exponential families we considered in Section 3.2.4.

Lemma 4.7: *Suppose $Z$ is a random variable in $\mathbb{R}^n$ with density / mass function from a s-parameter family in canonical form, i.e.*

$$p_\eta(z) = \exp\left(\sum_{i=1}^{s} \eta_i T_i(z) - A(\eta)\right) h(z).$$

*Then, provided $\eta$ is in the interior of the natural parameter space $\Xi$, the moment generating function of $T(Z)$,*

$$M(t) := \mathbb{E}\left[\exp(t'T(Z))\right]$$

*exists and has the form*

$$M(t) = \exp(A(\eta + t) - A(\eta)),$$

*for all $t$ in a neighbourhood of zero. Moreover,*

$$\mathbb{E}\, T(Z) = \nabla_\eta A(\eta), \qquad \mathrm{Var}(T(Z)) = \nabla_\eta^2 A(\eta).$$

*Proof.* We consider the continuous case, the discrete case is proved the same way with sums replacing integrals.

$$\begin{aligned}
\mathbb{E}\left[\exp(t'T(Z))\right] &= \int \exp\left(\sum_{i=1}^{s} (\eta_i + t_i) T_i(z) - A(\eta)\right) h(z)\,\mathrm{d}z \\
&= \exp(A(\eta+t) - A(\eta)) \int \exp\left(\sum_{i=1}^{s} (\eta_i + t_i) T_i(z) - A(\eta+t)\right) h(z)\,\mathrm{d}z \\
&= \exp(A(\eta+t) - A(\eta)) \int p_{\eta+t}(z)\,\mathrm{d}z \\
&= \exp(A(\eta+t) - A(\eta)).
\end{aligned}$$

The rest follows from the "moment generating" property of the MGF:

$$\mathbb{E}\, T(Z) = M'(0) = \exp(A(\eta+0) - A(\eta)) \nabla_\eta A(\eta) = \nabla_\eta A(\eta),$$

and

$$\begin{aligned}
\mathbb{E}(T(Z)T(Z)') = M''(0) &= \nabla_{t'}[\exp(A(\eta+t) - A(\eta)) \nabla_\eta A(\eta+t)]|_{t=0} \\
&= \exp(A(\eta) - A(\eta)) \left[\nabla_\eta A(\eta) [\nabla_\eta A(\eta)]' + \nabla_\eta^2 A(\eta)\right] \\
&= \nabla_\eta A(\eta) [\nabla_\eta A(\eta)]' + \nabla_\eta^2 A(\eta).
\end{aligned}$$

Combine this with $\mathrm{Var}(T(X)) = \mathbb{E}[T(X)T(X)'] - \mathbb{E}[T(X)]\, \mathbb{E}[T(X)]'$. □

As a consequence, the mean and variance of a random vector with density / mass function (68) are determined by the function $A$. We can allow a little more flexibility by introducing a a new dispersion parameter, $\phi$:

$$p_{\eta,\phi}(z) := \exp\left(\frac{[z'\eta - A(\eta)]}{c(\phi)}\right) h(z,\phi), \tag{69}$$

for some function $c > 0$. This gives a little more flexibility in specifying the variance.

Lemma 4.8: *If $Z$ is a random variable in $\mathbb{R}^n$ with density / mass function (69) then provided $\eta$ is in the interior of $\{\eta \in \mathbb{R}^n : A(\eta) < \infty\}$ then the moment generating function of $Z$ exists for all $t$ in a neighbourhood around zero and*

$$M(t) = \exp\left(\frac{A(\eta + t) - A(\eta)}{c(\phi)}\right).$$

*Moreover,*

$$\mathbb{E}\, Z = \nabla_\eta A(\eta), \qquad \mathrm{Var}(Z) = c(\phi)\nabla^2_\eta A(\eta).$$

*Proof.* Exercise. □

In a GLM the conditional distribution of $y_i|X_i$ is specified to have a exponential family distribution with parameter $\theta = \theta(X_i)$ whose canonical form is of the form (69).

### The link function

In (67), $\mu(X_i) = \theta(X_i)$. More generally we assume that for some one-to-one function $g$, called the *link function*, such that (cf. Lemma 4.8)

$$g(\mu(X_i)) = g(\dot{A}(\eta(X_i))) = \theta(X_i), \qquad \dot{A}(\eta) := \nabla_\eta A(\eta).$$

The most important case is when $\dot{A}$ is an invertible function and $g = \dot{A}^{-1}$. Then $g$ is the *canonical link function* and $\theta(X_i) = \eta(X_i) = g(\mu(X_i))$.

### Inference

These components allow us to fully specify the model conditional on $X_i$. In particular, we have a fully specified (conditional) likelihood function. For this reason, along with the good properties of maximum likelihood estimators discussed previously, maximum likelihood is the most common approach to inference in GLMs.

When $g$ is a canonical link function, the log-likelihood is

$$l(\beta, \phi) = \sum_{i=1}^n \frac{y_i'\eta(X_i) - A(\eta(X_i))}{c(\phi)} + \log h(y_i, \phi), \quad \eta(X_i) = X_i'\beta.$$

As exponential families are very "well-behaved", provided the link function is also "well-behaved", typically GLMs will satisfy the conditions required by Theorems 3.8 and 3.9 and therefore the ML estimator of $\beta$ will be both consistent and asymptotically normal. We will not discuss the details in this class, but you will explore this in simulation in the exercises.

**Examples**

Example 4.16 [Logit model]: One common binary response GLM is the "logistic regression" or "logit" model. This is a GLM where $g$ is the logit function

$$g(p) = \log(p/(1-p)) = \log(p) - \log(1-p).$$

The conditional distribution of $y_i|X_i \sim \text{Ber}(p)$, a Bernoulli distribution with parameter $p \in [0,1]$. Putting together the three ingredients of a GLM we have

$$y_i|X_i \sim \text{Ber}(g^{-1}(\eta(X_i))) = \text{Ber}(g^{-1}(X_i'\beta))$$

As such the log-likelihood based on an i.i.d. sample $(y_i, X_i)_{i=1}^n$ is [Exercise]

$$l(\beta) = \sum_{i=1}^n \left[ y_i \log\left(\frac{p_i(\beta)}{1-p_i(\beta)}\right) + \log(1-p_i(\beta))\right], \qquad p_i(\beta) = g^{-1}(X_i'\beta).$$

As the logit function is the inverse of the logistic function, we have $g^{-1}(z) = [1+\exp(-z)]^{-1}$. Here $g$ is the canonical link function. △

Example 4.17 [Poisson regression]: A common model for count data models is "Poisson regression". Here the (canonical) link function is $g(x) = \log(x)$ and (as given away by the name) the conditional distribution of $y_i|X_i$ is Poisson with mean $g^{-1}(X_i'\beta) = \exp(X_i'\beta)$. Therefore, the log-likelihood based on an i.i.d. sample is [Exercise]

$$l(\beta) = \sum_{i=1}^n \left[ y_i X_i'\beta - \exp(X_i'\beta) - \log(y_i!)\right].$$ △

We will finish our discussion of GLMs by considering two real-data examples.

Example 4.18 [Determinants of employment]: Lets use the logit model to explore how experience and education affects the probability of employment. We will fit a logit model to

$$\begin{aligned} y_i &= \textit{employed} \\ X_i'\beta &= \beta_0 + \beta_1 \textit{education} \times \textit{education} + \beta_2 \times \textit{experience} + \beta_3 \times \textit{experience}^2 + \beta_4 \times \textit{black}, \end{aligned}$$

where *black* is a dummy variable which takes the value 1 if the individual is black and 0 otherwise. The results are

|  |  |  |  |
|---|---|---|---|
| **Dep. Variable:** | employed | **No. Observations:** | 1618 |
| **Model:** | Logit | **Df Residuals:** | 1613 |
| **Method:** | MLE | **Df Model:** | 4 |
| **Date:** | Sun, 29 Sep 2024 | **Pseudo R-squ.:** | 0.2510 |
| **Time:** | 17:38:12 | **Log-Likelihood:** | -615.06 |
| **converged:** | True | **LL-Null:** | -821.16 |
| **Covariance Type:** | nonrobust | **LLR p-value:** | 6.433e-88 |

|  | **coef** | **std err** | **z** | **P> \|z\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **Intercept** | -4.7899 | 0.467 | -10.253 | 0.000 | -5.706 | -3.874 |
| **educ** | 0.3625 | 0.032 | 11.226 | 0.000 | 0.299 | 0.426 |
| **exper** | 0.9583 | 0.097 | 9.868 | 0.000 | 0.768 | 1.149 |
| **np.power(exper, 2)** | -0.0649 | 0.013 | -5.078 | 0.000 | -0.090 | -0.040 |
| **black** | -0.4871 | 0.149 | -3.265 | 0.001 | -0.779 | -0.195 |

△

Example 4.19 [Ship collisions]: Lets use the Poisson regression model to find how the number of months in service affects the number of accidents of ships. We will run a Poisson regression with

$$y_i = accidents$$
$$X_i'\beta = \beta_0 + \beta_1 \log(months) + Z_i'\delta,$$

where *accidents* is the number of accidents for each ship and *months* is the number of months the ship was at sea and $Z_i$ are dummies for when the ship was at sea and when it was constructed. Here $\log(months)$ is used instead of *months* as a mutiplicative effect here seems more plausible than an additive effect. The results are

|  |  |  |  |
|---|---|---|---|
| **Dep. Variable:** | accident | **No. Observations:** | 34 |
| **Model:** | GLM | **Df Residuals:** | 29 |
| **Model Family:** | Poisson | **Df Model:** | 4 |
| **Link Function:** | Log | **Scale:** | 1.0000 |
| **Method:** | IRLS | **Log-Likelihood:** | -84.675 |
| **Date:** | Sun, 29 Sep 2024 | **Deviance:** | 71.484 |
| **Time:** | 17:50:47 | **Pearson chi2:** | 70.4 |
| **No. Iterations:** | 6 | **Pseudo R-squ. (CS):** | 1.000 |
| **Covariance Type:** | nonrobust |  |  |

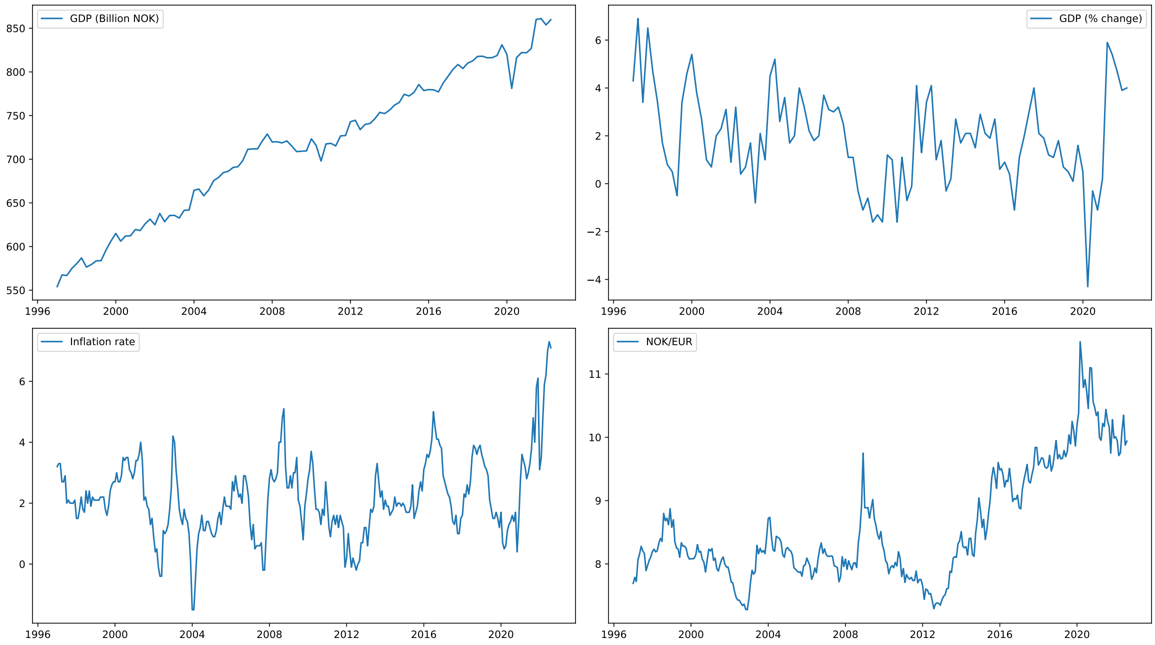
|  | **coef** | **std err** | **z** | **P> \|z\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **Intercept** | -4.5296 | 0.446 | -10.154 | 0.000 | -5.404 | -3.655 |
| **op** | 0.4026 | 0.117 | 3.444 | 0.001 | 0.173 | 0.632 |
| **co74** | 0.3541 | 0.130 | 2.717 | 0.007 | 0.099 | 0.610 |
| **co79** | -0.0695 | 0.203 | -0.342 | 0.732 | -0.468 | 0.329 |
| **np.log(service)** | 0.7974 | 0.044 | 17.963 | 0.000 | 0.710 | 0.884 |

△

# 5 Time Series

A time series is a set of observations $(x_t)_{t=1}^T$ with each $x_t$ recorded at time $t$. We model time series by a family of random variables $(X_t)_{t\in\mathbb{Z}}$, where $x_t$ is thought of as the realisation of $X_t$. We will also refer to $(X_t)_{t\in\mathbb{Z}}$ as a time series. Much data has a time series structure:

Figure 24: Some examples of time series



*Note:* Data obtained from Eurostat.

These series display a number of salient features which depart from the situation we have examined so far. In particular, it is clear that these series are *dependent*: their current value depends on their past values in some manner. This was explicitly ruled out in the first part of the course by our assumption of *independence*.

## 5.1 Stationarity

A time series $(X_t)_{t\in\mathbb{Z}}$ is *strictly stationary* if the joint distribution of $(X_{t_1}, \ldots, X_{t_k})$ is the same as that of $(X_{t_1+h}, \ldots, X_{t_k+h})$ for any $t_1, \ldots, t_k \in \mathbb{Z}$, any $h \in \mathbb{Z}$ and any $k \in \mathbb{N}$.

A time series $(X_t)_{t\in\mathbb{Z}}$ is *weakly stationary*, *covariance stationary* or just *stationary* if

(i) $\mathbb{E}\, X_t^2 < \infty$ for all $t \in \mathbb{Z}$,

(ii) $\mathbb{E}\, X_t = \mu$ for all $t \in \mathbb{Z}$,

(iii) $\operatorname{Cov}(X_t, X_s) = \operatorname{Cov}(X_{t+h}, X_{s+h})$ for all $t, s, h \in \mathbb{Z}$.

Note that this ensures that the variance of the process is also constant [Exercise].

Example 5.1 [i.i.d. sequence with second moment is (strictly) stationary]: Let $(X_t)_{t\in\mathbb{Z}}$ be an i.i.d. sequence. Then $(X_t)_{n\in\mathbb{Z}}$ is strictly stationary [Exercise]. If also $\mathbb{E}\, X_t^2 < \infty$ then the time series is also weakly stationary [Exercise]. △

Example 5.2 [Strict stationarity need not imply weak stationarity]: Suppose that $(X_t)_{t\in\mathbb{Z}}$ is an i.i.d. sequence of Cauchy random variables. Then $(X_t)_{t\in\mathbb{Z}}$ is strictly stationary by example 5.1, but it is not weakly stationary as:[81]

$$\mathbb{E}\,|X_t| = \int_{-\infty}^{\infty} |x| \frac{1}{\pi(1+x^2)}\,\mathrm{d}x = \frac{2}{\pi}\int_0^{\infty} \frac{x}{1+x^2}\,\mathrm{d}x = \frac{1}{\pi}\left[\log(1+x^2)\right]_{x=0}^{\infty} = \infty. \qquad \triangle$$

The existence of second moments is the only roadblock preventing strict stationary implying weak stationarity.

Lemma 5.1: *If $(X_t)_{t\in\mathbb{Z}}$ is a strictly stationary sequence and $\mathbb{E}\, X_t^2 < \infty$ then $(X_t)_{t\in\mathbb{Z}}$ is weakly stationary.*

*Proof.* Exercise. □

There is also an important special case in which the converse is true.

Example 5.3 [In the Gaussian case, weak stationarity implies strict stationarity.]: $(X_t)_{t\in\mathbb{Z}}$ is a Gaussian time series if any $X_{i_1}, \dots, X_{i_k}$ $(i_1, \dots, i_k \in \mathbb{Z},\ k \in \mathbb{N})$ has a multivariate normal distribution. Since the multivariate normal distribution is fully characterised by its mean and covariance matrix, if $(X_t)_{t\in\mathbb{Z}}$ is weakly stationary and Gaussian then it is strictly stationary [Exercise]. △

The autocovariance function of a time series $(X_t)_{t\in\mathbb{Z}}$ with $\mathbb{E}\, X_t^2 < \infty$ for each $t \in \mathbb{Z}$ is a function $\gamma : \mathbb{Z} \times \mathbb{Z} \to \mathbb{R}$, defined by

$$\gamma(t, s) := \operatorname{Cov}(X_t, X_s) = \mathbb{E}\left[(X_t - \mathbb{E}\, X_t)(X_s - \mathbb{E}\, X_s)\right].$$

The third condition that a stationary time series must satisfy can therefore be rephrased as $\gamma(t+h, s+h) = \gamma(t, s)$ for all $t, s, h \in \mathbb{Z}$. Moreover, since in this case we have $\gamma(r, s) = \gamma(r-s, 0)$ for any $r, s \in \mathbb{Z}$, it is convenient to redefine the autocovariance function as a function of only one variable, the *lag* $h \in \mathbb{Z}$:

$$\gamma(h) := \gamma(h, 0) = \operatorname{Cov}(X_{t+h}, X_t).$$

The autocorrelation function is the autocovariance divided by the variance:

$$\rho(h) := \gamma(h)/\gamma(0).$$

The autocovariance function of a stationary time series has a number of useful properties.

[81] For any random variable $X$, we have $\mathbb{E}\,|X| \le \sqrt{\mathbb{E}\, X^2}$ by the Cauchy-Schwarz inequality.

Proposition 5.1: *If $\gamma : \mathbb{Z} \to \mathbb{R}$ is the autocovariance function of a stationary time series, then*

*(i)* $\gamma(0) \geq 0$*;*

*(ii)* $|\gamma(h)| \leq \gamma(0)$ *for all* $h \in \mathbb{Z}$*;*

*(iii)* $\gamma$ *is even:* $\gamma(h) = \gamma(-h)$ *for* $h \in \mathbb{Z}$*.*

*Proof.* (i) follows since $\gamma(0) = \operatorname{Cov}(X_t, X_t) = \operatorname{Var}(X_t) \geq 0$. (ii) is a consequence of stationarity and the Cauchy-Schwarz inequality which together imply

$$|\operatorname{Cov}(X_{t+h}, X_t)| \leq \operatorname{Var}(X_{t+h})^{1/2}\operatorname{Var}(X_t)^{1/2} = \operatorname{Var}(X_t) = \gamma(0).$$

For (iii), note that

$$\gamma(-h) = \operatorname{Cov}(X_{t-h}, X_t) = \operatorname{Cov}(X_t, X_{t+h}) = \gamma(h). \qquad \square$$

## 5.2 Fundamental Processes

### 5.2.1 White noise

A time series $(\epsilon_t)_{t\in\mathbb{Z}}$ is *white noise* if it is a stationary process with $\mathbb{E}\,\epsilon_t = 0$ for all $t \in \mathbb{Z}$ and its autocovariance function satisfies

$$\gamma(h) = \operatorname{Cov}(\epsilon_{t+h}, \epsilon_t) = \begin{cases} \sigma^2 & \text{if } h = 0 \\ 0 & \text{if } h \neq 0 \end{cases}.$$

We write $\epsilon_t \sim \text{WN}(0, \sigma^2)$ if $(\epsilon_t)_{t\in\mathbb{Z}}$ is white noise. White noise processes are *serially uncorrelated.*

Example 5.4 [i.i.d. process with mean zero and finite variance is white noise]: Suppose that $(\epsilon_t)_{t\in\mathbb{Z}}$ is i.i.d. with mean zero and variance $\sigma^2 < \infty$: $\epsilon_t \sim \text{IID}(0, \sigma^2)$. Then $(\epsilon_t)_{t\in\mathbb{Z}}$ is white noise. △

Example 5.5 [White noise processes need not be i.i.d. nor conditionally homoskedastic]: Suppose that $\epsilon_t := u_t u_{t-1}$ where $(u_t)_{t\in\mathbb{Z}}$ is i.i.d. with each $u_t \sim \mathcal{N}(0, 1)$. Then $\epsilon_t \sim \text{WN}(0, 1)$ but $(\epsilon_t)_{t\in\mathbb{Z}}$ is not i.i.d. and $\mathbb{E}[\epsilon_t^2|\epsilon_{t-1}, \epsilon_{t-2}, \ldots] = u_{t-1}^2$.

To see that $\epsilon_t \sim \text{WN}(0, 1)$ note first that by independence

$$\mathbb{E}[\epsilon_t] = \mathbb{E}[u_t u_{t-1}] = \mathbb{E}[u_t]\,\mathbb{E}[u_{t-1}] = 0,$$

$$\mathbb{E}\left[\epsilon_t^2\right] = \mathbb{E}[u_t^2 u_{t-1}^2] = \mathbb{E}[u_t^2]\,\mathbb{E}[u_{t-1}^2] = 1,$$

and for $h > 0$ (the case with $h < 0$ is analogous)

$$\operatorname{Cov}(\epsilon_{t+h}, \epsilon_t) = \mathbb{E}\left[\epsilon_{t+h}\epsilon_t\right] = \mathbb{E}[u_{t+h}u_{t+h-1}u_t u_{t-1}] = \mathbb{E}[u_{t+h}]\,\mathbb{E}\left[u_{t+h-1}u_t u_{t-1}\right] = 0.$$

To see that $(\epsilon_t)_{t\in\mathbb{Z}}$ is not i.i.d. we calculate

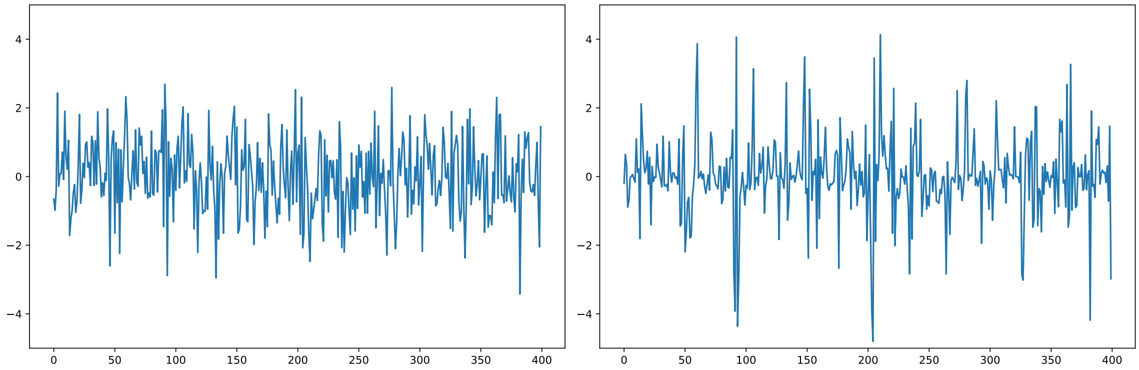
$$\operatorname{Cov}(\epsilon_t^2, \epsilon_{t-1}^2) = \mathbb{E}\left[\epsilon_t^2\epsilon_{t-1}^2\right] - \mathbb{E}[\epsilon_t^2]\,\mathbb{E}[\epsilon_{t-1}^2] = \mathbb{E}[u_t^2 u_{t-1}^4 u_{t-2}^2] - 1 = 3 - 1 = 2 \neq 0.$$

Finally, for the conditional heteroskedasticity, we note that

$$\mathbb{E}\left[\epsilon_t^2|\epsilon_{t-1}, \epsilon_{t-2}, \ldots\right] = \mathbb{E}\left[u_t^2 u_{t-1}^2|\epsilon_{t-1}, \epsilon_{t-2}, \ldots\right] = u_{t-1}^2\,\mathbb{E}\left[u_t^2|\epsilon_{t-1}, \epsilon_{t-2}, \ldots\right] = u_{t-1}\,\mathbb{E}[u_t^2] = u_{t-1}^2.$$

Here the second equality follows since given the past values $\epsilon_{t-1}, \epsilon_{t-2}, \ldots$ we know $u_{t-1}$ and hence can "pull it out"; the third equality follows from the fact that $u_t$ is independent from all the past values. △

Figure 25: Two white noise processes



*Note:* Left hand plot is a white noise process of i.i.d. standard normal random variables $u_t$. The right hand plot is a white noise process $\epsilon_t = u_t u_{t-1}$ (see example 5.5).

### 5.2.2 ARMA processes

*Autoregressive moving average* or *ARMA* processes are built up from autoregressive (AR) and moving average (MA) parts. We will first introduce AR and MA processes separately. For simplicity we will focus on zero-mean AR and MA processes; constants can be added to give these processes non-zero means [see Exercises].

*Moving average* or *MA* processes satisfy an equation of the form

$$X_t = \epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q}, \qquad \epsilon_t \sim \mathrm{WN}(0, \sigma^2),$$

for $q \in \mathbb{N}$. The above process is an MA($q$) process. MA($q$) processes are always stationary:

Lemma 5.2: *If $(X_t)_{t\in\mathbb{Z}}$ is a MA($q$) process, then it is stationary with $\mathbb{E}\,X_t = 0$ and*

$$\gamma(h) = \operatorname{Cov}(X_{t+h}, X_t) = \begin{cases} \sigma^2 \sum_{k=0}^{q-|h|} \theta_k\theta_{k+|h|} & \text{if } |h| \leq q \\ 0 & \text{if } |h| > q \end{cases}.$$

*Proof.* Putting $\theta_0 = 1$, $\mathbb{E}\, X_t = \sum_{j=0}^{q} \theta_j \,\mathbb{E}\, \epsilon_{t-j} = 0$ and since $\epsilon_t \sim \mathrm{WN}(0, \sigma^2)$, for $h \geq 0$,

$$\mathrm{Cov}(X_{t+h}, X_t) = \mathbb{E}\, X_{t+h} X_t = \sum_{j=0}^{q} \sum_{k=0}^{q} \theta_k \theta_j \,\mathbb{E}\, \epsilon_{t+h-j} \epsilon_{t-k} = \sigma^2 \sum_{k=0}^{q-h} \theta_k \theta_{k+h}$$

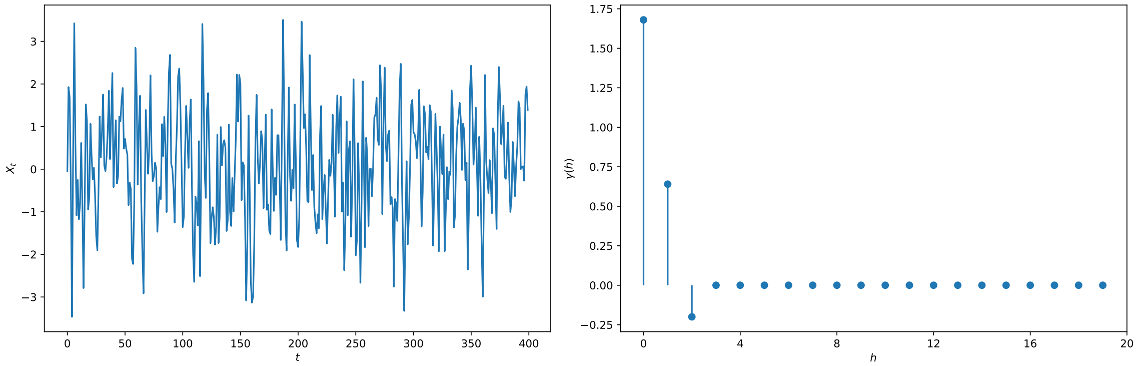
since $\mathbb{E}\, \epsilon_{t+h-j} \epsilon_{t-k} \neq 0$ only if $j = k + h \leq q$. Similarly for $h < 0$,

$$\mathrm{Cov}(X_{t+h}, X_t) = \mathbb{E}\, X_{t-|h|} X_t = \sum_{j=0}^{q} \sum_{k=0}^{q} \theta_k \theta_j \,\mathbb{E}\, \epsilon_{t-|h|-j} \epsilon_{t-k} = \sigma^2 \sum_{j=0}^{q-|h|} \theta_j \theta_{j+|h|},$$

since $\mathbb{E}\, \epsilon_{t+h-j} \epsilon_{t-k} \neq 0$ only if $k = j + |h| \leq q$. □

Example 5.6 [MA(2) process]: The following figure plots the realisation of a MA(2) process with $\theta_1 = 0.8$, $\theta_2 = -0.2$ on the left hand panel and the autocovariance function on the right hand panel.

Figure 26: A stationary MA(2) process and its autovariance function



△

The class of moving average processes can be extended to allow $q = \infty$. We say that a sequence $(x_n)_{n\in\mathbb{Z}}$ is *absolutely summable* if $\sum_{n\in\mathbb{Z}} |x_n| < \infty$. An absolutely summable sequence is square summable: $\sum_{n\in\mathbb{Z}} x_n^2 < \infty$. Any absolutely summable series converges.[82]

A process $(X_t)_{t\in\mathbb{Z}}$ is a MA($\infty$) process if there exists a white noise process $\epsilon_t \sim \mathrm{WN}(0, \sigma^2)$ and a sequence $(\theta_j)_{j=0}^{\infty}$ with $\sum_{j=0}^{\infty} |\theta_j| < \infty$ such that for $t \in \mathbb{Z}$

$$X_t = \sum_{j=0}^{\infty} \theta_j \epsilon_{t-j}.$$

Note that a MA($q$) process is a MA($\infty$) process with $\theta_k = 0$ for $k > q$.

The convergence of the infinite series in the display above is to be understood in the *mean*

[82] This result is valid in any complete metric space, which include all the spaces we will consider in this course.

*square* sense. That is, the above display requires (for each $t \in \mathbb{Z}$)

$$\lim_{n\to\infty} \mathbb{E}\,|X_t - X_{t,n}|^2 = 0, \qquad \text{for } X_{t,n} := \sum_{j=0}^{n} \theta_j \epsilon_{t-j}.$$

We write this as $X_{t,n} \xrightarrow{ms} X_t$. For any two random variables $\mathbb{E}\,|X - Z| \leq \sqrt{\mathbb{E}\,|X - Z|^2}$ (by the Cauchy – Schwarz inequality) and hence if $X_n \xrightarrow{ms} X$, $\mathbb{E}\,|X_n - X| \to 0$ also. Application of the reverse triangle inequality allows us to conclude that also

$$\mathbb{E}\,X_n \to \mathbb{E}\,X \qquad \text{and} \qquad \mathbb{E}\,X_n^2 \to \mathbb{E}\,X^2.$$

Moreover, by Cauchy – Schwarz once more, if $X_n \xrightarrow{ms} X$ and $Y_n \xrightarrow{ms} Y$, then $\mathbb{E}\,X_nY_n \to \mathbb{E}\,XY$.

Lemma 5.3: *If $(X_t)_{t\in\mathbb{Z}}$ is a MA($\infty$) process, then it is stationary with $\mathbb{E}\,X_t = 0$ and*

$$\gamma(h) = \operatorname{Cov}(X_{t+h}, X_t) = \sigma^2 \sum_{j=0}^{\infty} \theta_j \theta_{j+|h|}.$$

*Proof.* Let $X_t = \sum_{j=0}^{\infty} \theta_j \epsilon_{t-j}$ and $X_{t,n} = \sum_{j=0}^{n} \theta_j \epsilon_{t-j}$. Then, where the limits are understood in the mean-square sense

$$\mathbb{E}\,X_t = \mathbb{E}\lim_{n\to\infty} X_{t,n} = \lim_{n\to\infty} 0 = 0.$$

Similarly for $h \geq 0$,

$$\mathbb{E}\,X_tX_{t+h} = \mathbb{E}\lim_{n\to\infty}\sum_{k=0}^{n}\sum_{j=0}^{n}\theta_k\theta_j\epsilon_{t-k}\epsilon_{t+h-j} = \lim_{n\to\infty}\sum_{k=0}^{n}\sum_{j=0}^{n}\theta_k\theta_j\,\mathbb{E}\,\epsilon_{t-k}\epsilon_{t+h-j}.$$

$\mathbb{E}\,\epsilon_{t-k}\epsilon_{t+h-j}$ is non-zero only when $j = k + h$ (and both indices are less than $n$) so the term on the right hand side is $\lim_{n\to\infty}\sigma^2\sum_{k=0}^{n-h}\theta_k\theta_{k+h}$. Since $(\theta_j)_{j=0}^{\infty}$ is absolutely summable it is square summable and by Cauchy – Schwarz we have

$$\sum_{k=0}^{\infty}|\theta_k\theta_{k+h}| \leq \left(\sum_{k=0}^{\infty}\theta_k^2\right)^{1/2}\left(\sum_{k=0}^{\infty}\theta_{k+h}^2\right)^{1/2} = \sum_{k=0}^{\infty}\theta_k^2 < \infty,$$

and hence the limit $\lim_{n\to\infty}\sigma^2\sum_{k=0}^{n-h}\theta_k\theta_{k+h} = \sigma^2\sum_{k=0}^{\infty}\theta_k\theta_{k+h}$ exists. If $h < 0$, the argument is analogous, except that

$$\mathbb{E}\,X_tX_{t+h} = \mathbb{E}\lim_{n\to\infty}\sum_{k=0}^{n}\sum_{j=0}^{n}\theta_k\theta_j\epsilon_{t-k}\epsilon_{t-|h|-j} = \lim_{n\to\infty}\sum_{k=0}^{n}\sum_{j=0}^{n}\theta_k\theta_j\,\mathbb{E}\,\epsilon_{t-k}\epsilon_{t-|h|-j},$$

and so the expectation is non-zero only if $k = j + |h|$ (and both indices are less than $n$). Argue analogously to the case where $h \geq 0$, to obtain the limit $\lim_{n\to\infty}\sigma^2\sum_{k=0}^{n-|h|}\theta_k\theta_{k+|h|} =$

$\sigma^2 \sum_{k=0}^{\infty} \theta_k \theta_{k+|h|}$. Combine these two to obtain

$$\gamma(h) = \text{Cov}(X_{t+h}, X_t) = \sigma^2 \sum_{j=0}^{\infty} \theta_j \theta_{j+|h|}.$$

The proof is completed by observing that the preceeding display does not depend on $t$. □

If $X_t$ is a MA($\infty$) process, its *impulse response function* is

$$\Psi(s) := \frac{\partial X_{t+s}}{\partial \epsilon_t} = \frac{\partial \sum_{j=0}^{\infty} \theta_j \epsilon_{t+s-j}}{\partial \epsilon_t} = \theta_s.$$

This describes the effect on $X_{t+s}$ of a one-time shock in $\epsilon_t$, with all other variables held constant.

*Autoregressive* or *AR* processes satisfy difference equations of the following type:

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \epsilon_t, \quad \epsilon_t \sim \text{WN}(0, \sigma^2).$$

If $(X_t)_{t\in\mathbb{Z}}$ satisfies the preceding display for each $t \in \mathbb{Z}$ it is an AR($p$) process.

Example 5.7 [AR(1)]: The simplest AR process is the AR(1):

$$X_t = \phi X_{t-1} + \epsilon_t, \qquad \epsilon_t \sim \text{WN}(0, \sigma^2).$$

A stationary AR(1) process $(X_t)_{t\in\mathbb{Z}}$ which satisfies the above difference equation exists if and only if $|\phi| \neq 1$. To see that such a solution exists when this condition on $\phi$ holds we will consider the two cases separately. Firstly suppose that $|\phi| < 1$. Then we can "substitute backwards" to obtain

$$\begin{aligned} X_t &= \phi X_{t-1} + \epsilon_t \\ &= \phi \left[\phi X_{t-2} + \epsilon_{t-1}\right] + \epsilon_t \\ &= \phi^2 X_{t-2} + \phi \epsilon_{t-1} + \epsilon_t \\ &= \phi^2 \left[\phi X_{t-3} + \epsilon_{t-2}\right] + \phi \epsilon_{t-1} + \epsilon_t \\ &= \cdots \\ &= \phi^k X_{t-k} + \sum_{j=0}^{k-1} \phi^j \epsilon_{t-j}. \end{aligned}$$

Since $|\phi| < 1$ we may take the mean square limit as $k \to \infty$ to find

$$X_t = \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j}.$$

This equation demonstrates that in the $|\phi| < 1$ case, the stationary solution to the AR(1) equation is backwards looking: we say that such a solution is *causal*: in this case $X_t$ in the preceeding display is a stationary, causal solution.

In the case where $|\phi| > 1$ a stationary solution exists, but not a causal solution. We obtain the stationary solution by "substituting forwards". We can re-arrange the equation (starting from $t+1$) as

$$\begin{aligned}
X_t &= \phi^{-1} X_{t+1} - \phi^{-1}\epsilon_{t+1} \\
&= \phi^{-1}\left[\phi^{-1} X_{t+2} - \phi^{-1}\epsilon_{t+2}\right] - \phi^{-1}\epsilon_{t+1} \\
&= \phi^{-2} X_{t+2} - \phi^{-2}\epsilon_{t+2} - \phi^{-1}\epsilon_{t+1} \\
&= \phi^{-2}\left[\phi^{-1} X_{t+3} - \phi^{-1}\epsilon_{t+3}\right] - \phi^{-2}\epsilon_{t+2} - \phi^{-1}\epsilon_{t+1} \\
&= \cdots \\
&= \phi^{-k} X_{t+k} - \sum_{j=1}^{k} \phi^{-j}\epsilon_{t+j},
\end{aligned}$$

and so taking the limit as $k \to \infty$ we get

$$X_t = -\sum_{j=1}^{\infty} \phi^{-j}\epsilon_{t+j}.$$

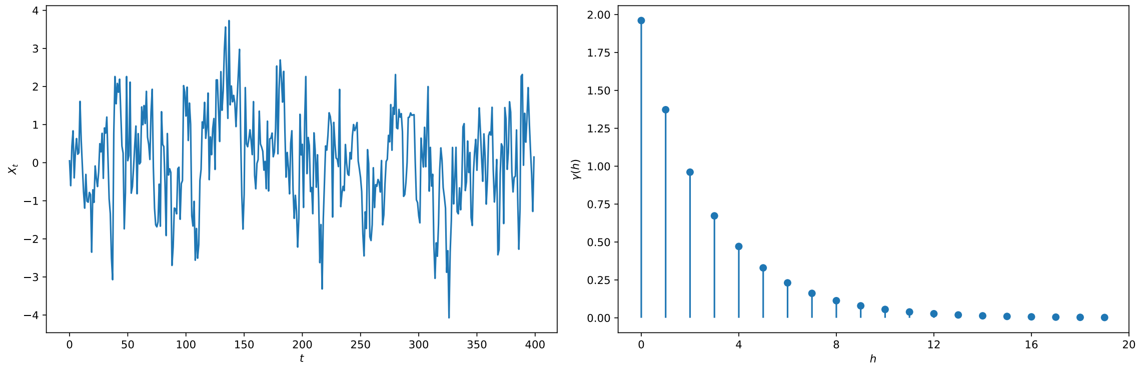
We will formally establish that these solutions exist and that they are stationary in a later section. Additionally, we will also come back to the remaining two cases where $|\phi| = 1$ and show that no stationary solution can exist.

A stationary AR(1) satisfying our defining equation must have mean zero [Exercise]. Moreover, the autocovariance function of the AR(1) process satisfies [Exercise]

$$\gamma(h) = \frac{\sigma^2 \phi^{|h|}}{1-\phi^2}.$$

The next figure shows a realisation of a stationary causal AR(1) process (with $\phi = 0.7$) in the left hand panel and its autocovariance function in the right hand panel.

Figure 27: A stationary, causal AR(1) process and its autovariance function



△

As Example 5.7 shows, unlike white noise processes, AR processes display serial correlation and persistence.

An *autoregressive, moving average* or *ARMA* process has both MA and AR parts. $(X_t)_{t\in\mathbb{Z}}$ is a ARMA$(p,q)$ process if it satisfies

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \cdots \phi_p X_{t-p} = \epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}, \qquad \epsilon_t \sim \mathrm{WN}(0,\sigma^2).$$

We can equivalently write ARMA processes more compactly via the *lag operator* and *lag polynomials*. The lag operator $L$ satisfies:

$$X_{t-j} = L^j X_t \qquad \text{for } j \in \mathbb{Z}.$$

Lag polynomials are polynomial functions with the lag operator as their argument. If we let

$$\begin{aligned}\Phi(z) &\coloneqq 1 - \phi_1 z - \cdots - \phi_p z^p\\ \Theta(z) &\coloneqq 1 + \theta_1 z + \cdots + \theta_q z^q,\end{aligned}$$

then our ARMA$(p,q)$ process can be written more compactly as

$$\Phi(L)X_t = \Theta(L)\epsilon_t, \qquad \epsilon_t \sim \mathrm{WN}(0,\sigma^2). \tag{70}$$

The polynomials $\Theta(z)$ and $\Phi(z)$ are the moving average and autoregressive polynomials respectively. Note that if $(X_t)_{t\in\mathbb{Z}}$ satisfies (70) it also satisfies

$$\Gamma(L)\Phi(L)X_t = \Gamma(L)\Theta(L)\epsilon_t, \qquad \epsilon_t \sim \mathrm{WN}(0,\sigma^2),$$

for any polynomial $\Gamma$. We avoid this redundancy by assuming that the polynomials are always in their simplest form, i.e. they have no *common factors*.

Applying such a lag polynomial (also called a *linear filter*) to a stationary process yields a stationary process. This is true even for infinite filters, provided the coefficients are absolutely summable. This is proven in the Lemma below; subsequently we shall make use of this fact without explicit mention.[83]

Lemma 5.4: *If $(X_t)_{t\in\mathbb{Z}}$ is a time series such that $\sup_{t\in\mathbb{Z}} \mathbb{E}\, X_t^2 < \infty$ and and $(\psi_j)_{j\in\mathbb{Z}}$ is absolutely summable (i.e. $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$) then*

$$Y_t = \Psi(L)X_t = \sum_{j=-\infty}^{\infty} \psi_j X_{t-j},$$

*is convergent in mean square. If, moreover, $(X_t)_{t\in\mathbb{Z}}$ is stationary with autocovariance function $\gamma_X$, then the process $(Y_t)_{t\in\mathbb{Z}}$ is stationary with autocovariance function*

$$\gamma_Y(h) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_X(h-j+k).$$

[83] The proof of this lemma are starred; all such proofs are entirely optional.

**Proof:* Let $m < n$ and define $Y_{n,t} := \sum_{j=-n}^{n} \psi_j X_{t-j}$. Letting $C = \sup_{t\in\mathbb{Z}} \mathbb{E}\, X_t^2 < \infty$ and using the Cauchy-Schwarz inequality we have

$$\begin{aligned}
\mathbb{E}\,[Y_{n,t} - Y_{m,t}]^2 &= \mathbb{E}\left[\sum_{j=m+1}^{n} \psi_j X_{t-j} + \sum_{j=-n}^{-(m+1)} \psi_j X_{t-j}\right]^2 \\
&\le 2\,\mathbb{E}\left[\sum_{j=m+1}^{n} \psi_j X_{t-j}\right]^2 + 2\,\mathbb{E}\left[\sum_{j=-n}^{-(m+1)} \psi_j X_{t-j}\right]^2 \\
&\le 2\sum_{j=m+1}^{n}\sum_{k=m+1}^{n} |\psi_j||\psi_k||\,\mathbb{E}[X_{t-j}X_{t-j}] + 2\sum_{j=-n}^{-(m+1)}\sum_{k=-n}^{-(m+1)} |\psi_j||\psi_k||\,\mathbb{E}[X_{t-j}X_{t-j}] \\
&\le 2C\left[\left(\sum_{j=m+1}^{n} |\psi_j|\right)^2 + \left(\sum_{j=-n}^{-(m+1)} |\psi_j|\right)^2\right].
\end{aligned}$$

Since $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, each of the sums in the last term converge to 0. Hence $(Y_{n,t})_{n\in\mathbb{N}}$ is a Cauchy sequence in $L_2$. Since $L_2$ is complete, $Y_{n,t}$ has a mean square limit in $L_2$.

If $(X_t)_{t\in\mathbb{Z}}$ is stationary, then we have

$$\mathbb{E}\, Y_t = \lim_{n\to\infty} \mathbb{E}\, Y_{n,t} = \lim_{n\to\infty} \sum_{j=-n}^{n} \psi_j\, \mathbb{E}\, X_t = \lim_{n\to\infty} \sum_{j=-n}^{n} \psi_j \mu_X,$$

where $\mu_X = \mathbb{E}\, X_t$. The limit exists by the absolutely summability of the $\psi_j$, and does not depend on $t$. Similarly, we have

$$\begin{aligned}
\mathrm{Cov}(Y_t, Y_{t-h}) &= \lim_{n\to\infty} \mathrm{Cov}(Y_{n,t}, Y_{n,t-h}) \\
&= \lim_{n\to\infty} \sum_{j=-n}^{n}\sum_{k=-n}^{n} \psi_j\psi_k \mathrm{Cov}[X_{t-j}, X_{t-h-k}] \\
&= \lim_{n\to\infty} \sum_{j=-n}^{n}\sum_{k=-n}^{n} \psi_j\psi_k \gamma_X(h-j+k) \\
&= \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} \psi_j\psi_k \gamma_X(h-j+k).
\end{aligned}$$

The limit exists by the absolutely summability of the $\psi_j$ and does not depend on $t$. □

An ARMA$(p,q)$ process $(X_t)_{t\in\mathbb{Z}}$ is *causal* if there are $(\psi_j)_{j=0}^{\infty}$ such that $\sum_{j=0}^{\infty} |\psi_j| < \infty$ and

$$X_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}, \qquad t \in \mathbb{Z}. \tag{71}$$

$(X_t)_{t\in\mathbb{Z}}$ is *invertible* if there are $(\pi_j)_{j=0}^{\infty}$ such that $\sum_{j=0}^{\infty} |\pi_j| < \infty$ and

$$\epsilon_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}, \qquad t \in \mathbb{Z}. \tag{72}$$

There are conditions for causality, invertibility and stationarity in terms of the polynomials $\Phi$ and $\Theta$. These conditions have to do with the behaviour of these polynomials on the unit circle and unit disk in the complex plane.[84] We denote these by (respectively)

$$\mathbb{T} := \{z \in \mathbb{C} : |z| = 1\}, \qquad \mathbb{D} := \{z \in \mathbb{C} : |z| \leq 1\}.$$

Proposition 5.2: *Suppose that $(X_t)_{t\in\mathbb{Z}}$ is an ARMA process for which the polynomials $\Phi$ and $\Theta$ have no common factors. Then $(X_t)_{t\in\mathbb{Z}}$ is causal if and only if $\Phi(z) \neq 0$ for all $z \in \mathbb{D}$. The coefficients $(\psi)_{j=0}^{\infty}$ are determined by*

$$\Psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\Theta(z)}{\Phi(z)}, \quad z \in \mathbb{D}.$$

**Proof:* Suppose that $\Phi(z) \neq 0$ for all $z \in \mathbb{D}$. Since $\Phi(z) \neq 0$ on $\mathbb{D}$ and polynomials are continuous, there is a $\varepsilon \in (0,1)$ such that $\Phi(z) \neq 0$ on $\{z \in \mathbb{C} : |z| < 1+\varepsilon\}$ and hence the function $1/\Phi(z)$ is analytic on $\{z \in \mathbb{C} : |z| < 1+\varepsilon\}$. That is, there exists a (convergent) power series expansion:

$$\frac{1}{\Phi(z)} = \sum_{j=0}^{\infty} \zeta_j z^j := \zeta(z), \quad z \in \{z \in \mathbb{C} : |z| < 1+\varepsilon\}.$$

Taking $z = 1+\varepsilon/2$ we conclude that $\zeta_j(1+\varepsilon/2)^j \to 0$ as $j \to \infty$ and hence there exists a $M \in (0,\infty)$ with

$$|\zeta_j| \leq M(1+\varepsilon/2)^{-j}, \quad j = 0, 1, \ldots$$

Hence $\sum_{j=0}^{\infty} |\zeta_j| < \infty$ and $\Phi(z)/\zeta(z) = 1$ for $z \in \mathbb{D}$. As such

$$X_t = \zeta(L)\Phi(L)X_t = \zeta(L)\Theta(L)\epsilon_t = \frac{\Theta(L)}{\Phi(L)}\epsilon_t = \Psi(L)\epsilon_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$$

where we have that the $(\psi_j)_{j=0}^{\infty}$ are

$$\psi_j = \sum_{k=0}^{\infty} \zeta_k \theta_{j-k},$$

and hence (where all $\theta_j = 0$ for $j < 0$)

$$\sum_{j=0}^{\infty} |\psi_j| \leq \sum_{j=0}^{\infty} \left| \sum_{k=0}^{\infty} \zeta_k \theta_{j-k} \right| \leq \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} |\zeta_k||\theta_{j-k}| \leq \sum_{k=0}^{\infty} |\zeta_k| \sum_{j=0}^{\infty} |\theta_j| < \infty. \tag{73}$$

Next assume that $(X_t)_{t\in\mathbb{Z}}$ is causal, i.e. has representation (71). Then,

$$\Theta(L)\epsilon_t = \Phi(L)X_t = \Phi(L)\Psi(L)\epsilon_t = \Pi(L)\epsilon_t.$$

[84] Many of the remaining proofs in this section are starred: they rely on facts from complex analysis and are entirely optional.

where $\Pi(z) = \Phi(z)\Psi(z) = \sum_{j=0} \pi_j z^j$ for $z \in \mathbb{D}$ (note that $\sum_{j=0}^{\infty} |\pi_j| < \infty$ as can be shown analogously to (73)). That is

$$\sum_{j=0}^{q} \theta_j \epsilon_{t-j} = \sum_{j=0}^{\infty} \pi_j \epsilon_{t-j}.$$

By taking the expectation of the product of each side of this equation with $\epsilon_{t-k}$ we have

$$\theta_k = \sum_{j=0}^{q} \theta_j \, \mathbb{E}[\epsilon_{t-j}\epsilon_{t-k}] = \sum_{j=0}^{\infty} \pi_j \, \mathbb{E}[\epsilon_{t-j}\epsilon_{t-k}] = \pi_k,$$

so $\pi_k = \theta_k$ (including that $\pi_k = 0$ for $k > q$.) It follows that $\Theta(z) = \Pi(z) = \Phi(z)\Psi(z)$ on $\mathbb{D}$, hence $\Psi(z) = \Theta(z)/\Psi(z)$. Since $|\Psi(z)| < \infty$ on $z \in \mathbb{D}$ by absolute summability, if $\Psi(z_0) = 0$ for a $z_0 \in \mathbb{D}$ then $\Theta(z_0) = 0$ also. This implies that $\Phi(z) = (z - z_0)\Phi^*(z)$ and $\Theta(z) = (z - z_0)\Theta^*(z)$ for some lower degree polynomials $\Phi^*$ and $\Theta^*$, i.e. they have common factors, which is a contradiction. □

Proposition 5.3: *Suppose that $(X_t)_{t\in\mathbb{Z}}$ is an ARMA process for which the polynomials $\Phi$ and $\Theta$ have no common factors. Then $(X_t)_{t\in\mathbb{Z}}$ is invertible if and only if $\Theta(z) \neq 0$ for all $z \in \mathbb{D}$. The coefficients $(\pi)_{j=0}^{\infty}$ are determined by*

$$\Pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\Phi(z)}{\Theta(z)}, \quad z \in \mathbb{D}.$$

**Proof:* Argue analogously to as in the proof of Proposition 5.2. □

Theorem 5.1: *Suppose that $(X_t)_{t\in\mathbb{Z}}$ is an ARMA process. such that $\Phi(z) \neq 0$ for all $z \in \mathbb{T}$. Then $(X_t)_{t\in\mathbb{Z}}$ has the unique stationary solution*

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j \epsilon_{t-j}, \tag{74}$$

*where, for some $r > 1$,*

$$\Psi(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j = \frac{\Theta(z)}{\Phi(z)}, \qquad z \in \mathbb{C}, \quad r^{-1} < |z| < r.$$

**Proof:* As polynomials are continuous, there is a $r > 1$ such that $\Psi(z) \neq 0$ for $z$ in the annulus $A := \{z \in \mathbb{C} : |z| \in (r^{-1}, r)\}$. Hence, the function $\Psi(z) := \frac{\Theta(z)}{\Phi(z)}$ is well defined and analytic on this annulus. By a result from complex analysis there are unique coefficients $\psi_j$ such that we have the following *Laurent series* expansion:

$$\Psi(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j, \qquad z \in A.$$

The convergence is absolute and uniform over any compact subset of $A$. Since $\mathbb{T} \subset A$ and compact it follows that $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$. Therefore, by Lemma 5.3, (74) is a stationary process. It is a solution to the ARMA equation as applying the operator $\Phi$ to each side we have

$$\Phi(L)X_t = \Phi(L)\Psi(L)\epsilon_t = \Phi(L)\frac{\Theta(L)}{\Phi(L)}\epsilon_t = \Theta(L)\epsilon_t.$$

For the uniqueness, suppose that $(X_t)_{t\in\mathbb{Z}}$ is a stationary solution to the ARMA equations. Arguing similarly to the first part, there is a $\delta > 1$ such that $\sum_{j=-\infty}^{\infty} \zeta_j z^j = \zeta(z) = 1/\Phi(z)$ is convergent on $B = \{z \in \mathbb{C} : |z| \in (\delta^{-1}, \delta)\}$. As $T \subset B$ and compact we have that $\sum_{j=-\infty}^{\infty} |\zeta_j| < \infty$. Applying the operator $\zeta$ to each size of the ARMA equations gives

$$X_t = \zeta(L)\Phi(L)X_t = \zeta(L)\Theta(L)\epsilon_t = \frac{\Theta(L)}{\Phi(L)}\epsilon_t = \Psi(L)\epsilon_t. \qquad \square$$

The preceding Theorem demonstrates that stationary solutions are unique. Lets check that the result of this theorem is the same as the solutions we have already established for the AR(1) and MA($q$) cases.

Example 5.8 [AR(1)]: Let $(X_t)_{t\in\mathbb{Z}}$ be an AR(1) process with $|\phi_1| < 1$:

$$X_t = \phi_1 X_{t-1} + \epsilon_t, \qquad \epsilon_t \sim \mathrm{WN}(0, \sigma^2).$$

We know from example 5.7 that $X_t = \sum_{j=0}^{\infty} \phi_1^j \epsilon_{t-j}$. From Theorem 5.1, we have the solution $X_t = \frac{1}{\Phi(L)}\epsilon_t$. To verify that these are the same note that we have for any real $z$ with $0 < z < 1$,

$$\frac{1}{1-\phi_1 z} = \sum_{j=0}^{\infty} (\phi_1 z)^j = \sum_{j=0}^{\infty} \phi_1^j z^j.$$

This establishes that $\frac{1}{\Phi(L)} = \Psi(L)$ where $\Psi(z)$ is the polynomial $\Psi(z) = \sum_{j=0}^{\infty} \phi_1^j z^j$ and hence, the two representations are identical. △

Example 5.9 [MA(q)]: Let $(X_t)_{t\in\mathbb{Z}}$ be an MA($q$) process:

$$X_t = \epsilon_t + \theta_1\epsilon_{t-1} + \ldots + \theta_q\epsilon_{t-q}, \qquad \epsilon_t \sim \mathrm{WN}(0, \sigma^2).$$

The expression from Theorem 5.1 yields the same:

$$X_t = \frac{\Theta(L)}{\Phi(L)}\epsilon_t = \frac{\Theta(L)}{1}\epsilon_t = \sum_{j=0}^{q} \theta_j \epsilon_{t-j}, \quad \theta_0 = 1. \qquad \triangle$$

Example 5.10 [ARMA(2, 1)]: Let $(X_t)_{t\in\mathbb{Z}}$ be an ARMA(2, 1) process:

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} = \epsilon_t + \theta_1\epsilon_{t-1}, \qquad \epsilon_t \sim \mathrm{WN}(0, \sigma^2).$$

We will examine conditions for stationarity, causality and invertibility in this model, and – assuming stationarity – derive its autocovariance function.

For the solution to be stationary and causal we need that the roots of the AR polynomial $\Phi(z) = 1 - \phi_1 z - \phi_2 z^2$ are outside the unit circle, i.e. if $\Phi(z_0) = 0$ then $|z| > 1$. Letting $\lambda = z^{-1}$, this is equivalent to $|\lambda| < 1$ and $\lambda_0 = z_0^{-1}$ is a solution to $\lambda^2 - \phi_1 \lambda - \phi_2 = 0$. Recall that whether the solution to a quadratic equation is real or complex depends on the value of the discriminant $\phi_1^2 + 4\phi_2$: if it is non-negative, then we have two real roots,

$$\left| \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{2} \right| < 1.$$

Since the discriminant is positive, the larger of the two roots must satisfy $\phi_1 + \sqrt{\phi_1^2 + 4\phi_2} < 2$, which can re-arranged to obtain that $\phi_2 < 1 - \phi_1$. Similarly the smaller of the two roots must satisfy $\phi_1 - \sqrt{\phi_1^2 + 4\phi_2} > -2$, which can be re-arranged to yield $\phi_2 < 1 + \phi_1$. If, instead, the discriminant is negative then there are two complex roots,

$$\lambda = \frac{\phi_1}{2} \pm i \frac{\sqrt{-(\phi_1^2 + 4\phi_2)}}{2}.$$

We can equivalently write our condition that $|\lambda| < 1$ as $\lambda^2 < 1$, which takes the form

$$\frac{\phi_1^2}{4} - \frac{\phi_1^2 + 4\phi_2}{4} = -\phi_2 < 1,$$

i.e. $\phi_2 > -1$. Hence our conditions are:

(i) $\phi_2 < 1 - \phi_1$,

(ii) $\phi_2 < 1 + \phi_1$,

(iii) $\phi_2 > -1$.

For invertibility of the MA polynomial, we need that the roots of $\Theta(z) = 1 + \theta_1 z$ to be outside the unit circle. This has root $z = 1/\theta_1$ and hence we require that $|\theta_1| < 1$.

Under these conditions, as a consequence of Theorem 5.1, $X_t$ has zero mean. To find its autocovariance function, we start with $h = 0, 1$ and then from these can determine the general expression. Intially,

$$\begin{aligned}
\gamma(0) = \mathbb{E}\, X_t^2 &= \mathbb{E}\, X_t(\phi_1 X_{t-1} + \phi_2 X_{t-2} + \epsilon_t + \theta_1 \epsilon_{t-1}) \\
&= \phi_1 \gamma(1) + \phi_2 \gamma(2) + \mathbb{E}\left[(\phi_1 X_{t-1} + \phi_2 X_{t-2} + \epsilon_t + \theta_1 \epsilon_{t-1})(\epsilon_t + \theta_1 \epsilon_{t-1})\right] \\
&= \phi_1 \gamma(1) + \phi_2 \gamma(2) + \phi_1 \theta_1 \sigma^2 + \sigma^2 + \theta_1^2 \sigma^2;
\end{aligned}$$

for $h = 1$,

$$\begin{aligned}
\gamma(1) = \mathbb{E}\, X_t X_{t-1} &= \mathbb{E}(\phi_1 X_{t-1} + \phi_2 X_{t-2} + \epsilon_t + \theta_1 \epsilon_{t-1}) X_{t-1} \\
&= \phi_1 \gamma(0) + \phi_2 \gamma(1) + \theta_1 \sigma^2;
\end{aligned}$$

and $h \geq 2$,

$$\begin{aligned}\gamma(h) = \mathbb{E}\, X_t X_{t-h} &= \mathbb{E}(\phi_1 X_{t-1} + \phi_2 X_{t-2} + \epsilon_t + \theta_1 \epsilon_{t-1}) X_{t-h} \\ &= \phi_1 \gamma(h-1) + \phi_2 \gamma(h-2).\end{aligned}$$

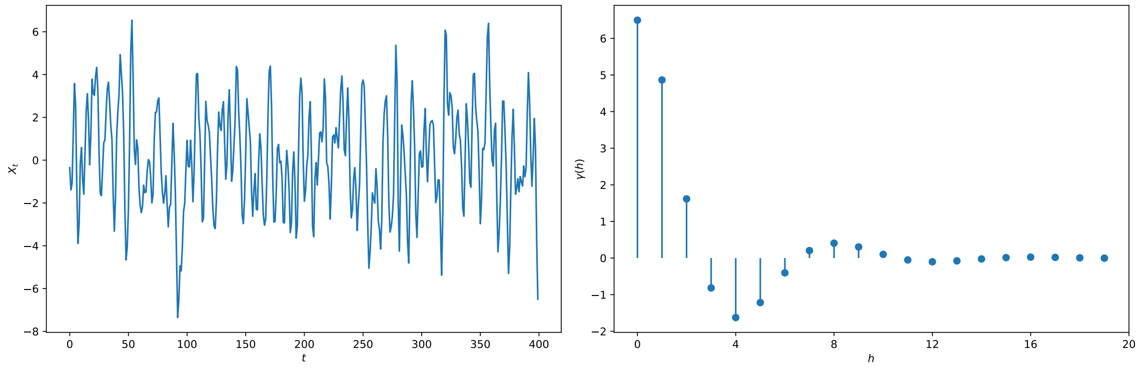
We can re-write these equations as

$$\begin{bmatrix} 1 & -\phi_1 & -\phi_2 \\ -\phi_1 & 1-\phi_2 & 0 \\ -\phi_2 & -\phi_1 & 1 \end{bmatrix} \begin{bmatrix} \gamma(0) \\ \gamma(1) \\ \gamma(2) \end{bmatrix} = \begin{bmatrix} \phi_1\theta_1\sigma^2 + \sigma^2 + \theta_1^2\sigma^2 \\ \theta_1\sigma^2 \\ 0 \end{bmatrix},$$

which can be solved to find $\gamma(0), \gamma(1), \gamma(2)$, from which all $\gamma(h)$ for $h > 2$ can be calculated using the recursive formula $\gamma(h) = \phi_1\gamma(h-1) + \phi_2\gamma(h-2)$.

Below is a plot of a realisation of an ARMA(2, 1) process and is autocovariance function, with $\phi_1 = 1.0, \phi_2 = -0.5, \theta_1 = 0.8$.

Figure 28: A causal, stationary and invertible ARMA(2, 1) process and its autovariance function

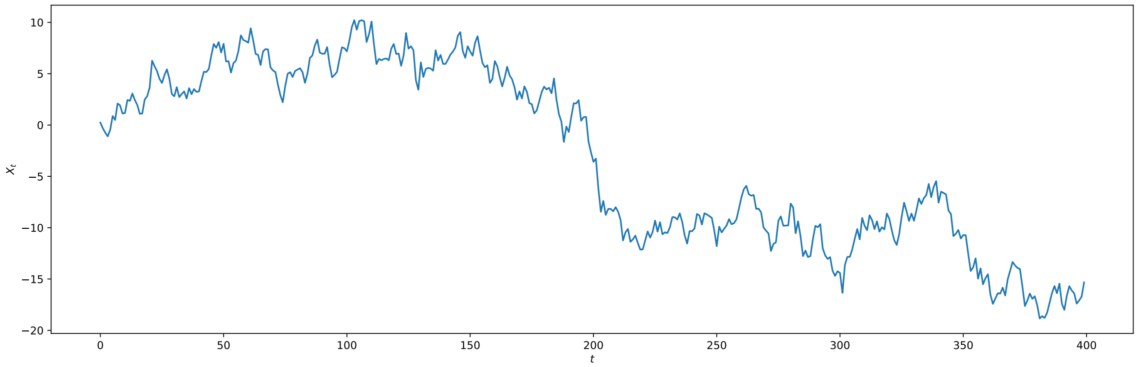


△

### 5.2.3 Unit roots, random walks and ARIMA

An important process is the *random walk*. Let $\epsilon_j \sim \text{WN}(0, \sigma^2)$ and consider the process $Y_t = \sum_{j=1}^{t} \epsilon_j$. A random walk is not stationary [Exercise].

Figure 29: A random walk



Nevertheless, we can take a random walk $(Y_t)_{t\in\mathbb{Z}}$ and convert it into a stationary process by *differencing*. Let $\Delta = I - L$, the *difference operator*. When we apply $\Delta$ to a random walk, we obtain

$$\Delta Y_t = Y_t - LY_t = Y_t - Y_{t-1} = \sum_{j=1}^{t} \epsilon_j - \sum_{j=1}^{t-1} \epsilon_j = \epsilon_t,$$

which is a white noise (and hence stationary) process.

Consider what happens if $\phi_1 = 1$ in an AR(1) model, with initial condition $X_0 = 0$.

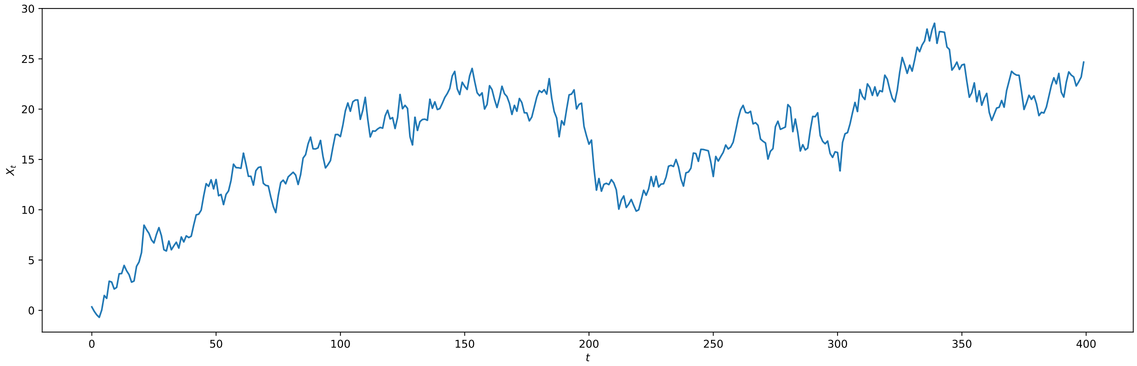
$$X_t = X_{t-1} + \epsilon_t = X_{t-2} + \epsilon_{t-1} + \epsilon_t = \cdots = \sum_{j=1}^{t} \epsilon_j.$$

Our AR(1) model is now a random walk. Note that $\phi_1 = 1$ does *not* satisfy the stationarity condition in Theorem 5.1: it has a root on the unit circle. For this reason, such models are also said to have *unit roots*.

If our AR(1) model had a non-zero intercept term, we would get a *random walk with drift*:

$$X_t = c + X_{t-1} + \epsilon_t = c + c + X_{t-2} + \epsilon_{t-1} + \epsilon_t = \cdots = tc + \sum_{j=1}^{t} \epsilon_j.$$

Figure 30: A random walk with drift



Again, we can convert this into a stationary process by using our differencing filter:

$$\Delta X_t = X_t - LX_t = X_t - X_{t-1} = c + X_{t-1} + \epsilon_t - X_{t-1} = c + \epsilon_t.$$

This approach works much more generally: we can define a whole class of processes whose $d$-th difference is an ARMA($p$, $q$) process. We say that a process $(X_t)_{t\in\mathbb{Z}}$ is an ARIMA($p$, $d$, $q$) process if $\Delta^d X_t$ is a *stationary* ARMA($p$, $q$) process. The "I" in ARIMA stands for "integrated". We will not study ARIMA processes in any detail in this course, but it is important to recognise that raw time series data may not be stationary and may need to be differenced one or more times in order to fit an ARMA model to it.

An important aspect of this is being able to determine whether a time series contains a unit root. This is often done using the Dickey – Fuller test or one of its extensions (e.g. the augmented Dickey – Fuller (ADF) test). We will describe the Dickey – Fuller test and refer to e.g. pp. 118 - 120 in [2] for a discussion of the ADF test.

The idea behind the Dickey – Fuller test is as follows: suppose we had an AR(1) with $|\phi| \leq 1$:

$$X_t = \phi X_{t-1} + \epsilon_t.$$

If we apply the difference operator $\Delta$ then we have

$$\Delta X_t = \mu X_{t-1} + \epsilon_t, \quad \mu = \phi - 1.$$

Then the AR(1) model will be stationary if $\mu < 0$ and non-stationary (i.e. have a unit root) if $\mu = 0$. We want, therefore, to perform a one-sided test of $H_0 : \mu = 0$ against $H_1 : \mu < 0$. We can do this using the standard $t$-statistic:

$$t = \frac{\sqrt{n}(\hat{\phi} - 1)}{\hat{V}^{1/2}},$$

where $\hat{\phi}$ is the OLS estimate of $X_t$ on $X_{t-1}$ and $\hat{V}$ is an estimate of the asymptotic variance of $\hat{\phi}$. Under $H_0$ this has a nonstandard (non-normal) asymptotic distribution and so we cannot use the usual $\mathcal{N}(0, 1)$ critical values. Fortunately, valid critical values for this test have been calculated and are available in standard statistical software packages.[85]

## 5.3 Forecasting

A central goal in time series is to forecast: that is, having observed the series up to time period $t$, we wish to predict the value in period $t+h$ for $h \geq 1$. If we wish our forecast to be optimal in terms of mean squared error, Proposition 4.10 tells us that it should be equal to the conditional expectation of $X_{t+h}$ given our observations $X_t, X_{t-1}, \ldots, X_1$.

In special cases (such as the linear regression model) this conditional expectation is easy to work with. In general it is not. As such we will consider an alternative: the best linear prediction. That is, we will look for the linear combination of $X_t, \ldots, X_1$ which gives the

[85] In statsmodels, the (ADF) test is available as `statsmodels.tsa.stattools.adfuller`.

smallest mean squared error. We will state the result in general terms, since it does not depend on the time structure of the data.

Proposition 5.4: *Let $W = (W_1, \dots, W_K)$ be random variables such that $\mathbb{E}\,\|W\|^2 < \infty$ and $X$ be such that $\mathbb{E}\,X^2 < \infty$. Define*

$$\hat{X} = \sum_{k=1}^{K} \lambda_k W_k,$$

*where the $\lambda_k$ satisfy*

$$\sum_{k=1}^{K} \lambda_k \,\mathbb{E}[W_k W_j] = \mathbb{E}[X W_j] \qquad \text{for } j = 1, \dots, K.$$

*Then $\hat{X}$ is the best linear predictor of $X$ given $W$:*

$$\hat{X} = \underset{Z \in \operatorname{span}(W_1, \dots, W_K)}{\arg\min} \mathbb{E}[X - Z]^2.$$

*Proof.* Let $\lambda = (\lambda_1, \dots, \lambda_K)$ and note that for any other $\alpha \in \mathbb{R}^K$, we have

$$\begin{aligned}\mathbb{E}\left[X - \alpha' W\right]^2 &= \mathbb{E}\left[X - \lambda' W + \lambda' W - \alpha' W\right]^2 \\ &= \mathbb{E}\left[X - \lambda' W\right]^2 + 2\,\mathbb{E}\left[(X - \lambda' W) W'\right](\lambda - \alpha) + \mathbb{E}\left[\lambda' W - \alpha' W\right]^2 \\ &= \mathbb{E}\left[X - \lambda' W\right]^2 + \mathbb{E}\left[\lambda' W - \alpha' W\right]^2,\end{aligned}$$

which is minimised when $\alpha = \lambda$. □

Remark 5.1: *By re-arranging the defining equations, one sees that $\lambda = (\lambda_1, \dots, \lambda_K)$ in Proposition 5.4 can be calculated as $\lambda = \mathbb{E}[WW']^{-1}\,\mathbb{E}[WX]$ provided that $\mathbb{E}[WW']$ is nonsingular.*

Given a collection of random variables $W$ and a random variable $X$ as in Proposition 5.4 we will write the best linear predictor $\hat{X} = \lambda' W$ as $E[X|W]$.[86] We collect a number of useful properties of this operator below:

Lemma 5.5: *Let $X, Y$ be random variables such that $\mathbb{E}\,X^2 < \infty$, $\mathbb{E}\,Y^2 < \infty$, $W = (W_1, \dots, W_K)$ such that $\mathbb{E}\,\|W\|^2 < \infty$ and $\Gamma = \operatorname{Var}(W)$. Also let $\alpha \in \mathbb{R}^K$. Then*

*(i)* $E\left[X|W\right] = \lambda' W$ *where $\lambda$ satisfies* $\mathbb{E}[(X - \lambda' W) W] = 0$,

*(ii)* $E[\alpha_1 X + \alpha_2 Y | W] = \alpha_1 E[X|W] + \alpha_2 E[Y|W]$,

*(iii)* $E\left[\alpha' W | W\right] = \alpha' W$,

*(iv)* $E\left[X|W\right] = E\left[E\left[X|W, V\right], W\right]$ *if* $\mathbb{E}\,\|V\|^2 < \infty$.

*Proof.* (i) Exercise.

[86] Note that this is *not* the same notation as the conditional expectation $\mathbb{E}[X|W]$.

(ii) Let $\lambda_1$ and $\lambda_2$ be such that $\mathbb{E}[(X - \lambda_1' W)W] = 0$ and $\mathbb{E}[(Y - \lambda_2' W)W] = 0$. Then also $\mathbb{E}[(\alpha_1 X - \alpha_1 \lambda_1' W)W] = 0$ and $\mathbb{E}[(\alpha_2 Y - \alpha_2 \lambda_2' W)W] = 0$ and so

$$\mathbb{E}[(\alpha_1 X - \alpha_1 \lambda_1' W + \alpha_2 Y - \alpha_2 \lambda_2' W)W] = 0.$$

(iii) Exercise.

(iv) Suppose $V$ is a random vector of dimension $M$ and let $\lambda_1 \in \mathbb{R}^K$ and $\lambda_2 \in \mathbb{R}^M$ be such that $E[X|W, V] = \lambda_1' W + \lambda_2' V$. By (ii) $E[\lambda_1' W + \lambda_2 V|W] = \lambda_1' W + \beta' W$ where $\beta' W = E[\lambda_2' V|W]$. Since $\beta$ satisfies $\mathbb{E}[(\lambda_2' V - \beta' W)W] = 0$,

$$\begin{aligned}\mathbb{E}\left[(X - (\lambda_1 + \beta)' W)W\right] &= \mathbb{E}\left[(X - \lambda_1' W)W\right] - \mathbb{E}\left[(\beta' W)W\right] \\ &= \mathbb{E}\left[(X - \lambda_1' W - \lambda_2' V)W\right] \\ &= 0,\end{aligned}$$

where the last equality is by definition of $\lambda_1$ and $\lambda_2$.

□

Example 5.11 [AR($p$)]: Suppose that $(X_t)_{t\in\mathbb{Z}}$ is an AR($p$) process. We will calculate the best linear prediction of $X_t$ given $X_{t-1}, \ldots, X_1$. We have that

$$X_t = \sum_{k=1}^{p} \phi_k X_{t-k} + \epsilon_t, \quad \epsilon_t \sim \text{WN}(0, \sigma^2).$$

Using Proposition 5.5,

$$E[X_t|X_{t-1}, \ldots, X_1] = E\left[\sum_{k=1}^{p} \phi_k X_{t-k} + \epsilon_t \middle| X_{t-1}, \ldots, X_1\right] = \sum_{k=1}^{p} \phi_k X_{t-k} + E[\epsilon_t|X_{t-1}, \ldots, X_1],$$

and $E[\epsilon_t|X_{t-1}, \ldots, X_1] = 0$ since $\mathbb{E}\left[\epsilon_t X_l\right] = 0$ for each $l \in \{1, \ldots, t-1\}$. Hence, the best linear predictor of $X_t$ is $\sum_{k=1}^{p} \phi_k X_{t-k}$.[87] △

Calculation of the best linear predictor of an AR process is straightforward. For other processes, such as MA processes, one must find a solution to the equations which define $\lambda$ in Proposition 5.4. Two recursive algorithms, the *Durbin - Levinson* and *innovations* algorithms can be used to find the best linear predictor. For this we will first introduce some notation. Let $X^n = (X_1, \ldots, X_n)'$ and $\lambda_{n,j}$ for $t \in \mathbb{N}$ and $j \in \{1, \ldots, n\}$ be such that

$$E(X_{n+1}|X^n) = \lambda_{n,1} X_n + \cdots + \lambda_{n,n} X_1. \tag{75}$$

Also let $\nu_n$ denote the mean squared prediction error:

$$\nu_n = \mathbb{E}\left[X_{n+1} - E(X_{n+1}|X^n)\right]^2. \tag{76}$$

[87] In practice we do not know $\phi_1, \ldots, \phi_p$ and therefore must replace them by estimates; we will discuss estimation in the following subsection.

PROPOSITION 5.5 [Durbin – Levinson algorithm]: *Suppose that $(X_t)_{t\in\mathbb{Z}}$ is a zero-mean stationary process with autocovariance function $\gamma$. Suppose that $\Gamma = \mathrm{Var}(X^n)$ is positive definite for each $n \in \mathbb{N}$. Then, the coefficients $\lambda_{n,j}$ and mean squared errors $\nu_n$ defined in (75), (76) can be computed as follows: $\lambda_{1,1} = \gamma(1)/\gamma(0)$, $\nu_0 = \gamma(0)$,*

$$\lambda_{n,n} = \left[\gamma(n) - \sum_{j=1}^{n-1} \lambda_{n-1,j}\gamma(n-j)\right] \nu_{n-1}^{-1},$$

$$\begin{bmatrix} \lambda_{n,1} \\ \vdots \\ \lambda_{n,n-1} \end{bmatrix} = \begin{bmatrix} \lambda_{n-1,1} \\ \vdots \\ \lambda_{n-1,n-1} \end{bmatrix} - \lambda_{n,n} \begin{bmatrix} \lambda_{n-1,n-1} \\ \vdots \\ \lambda_{n-1,1} \end{bmatrix},$$

*and $\nu_n = \nu_{n-1}[1 - \lambda_{n,n}^2]$.*

**Proof:* Let $\Pi_S$ denote the orthogonal projection onto any closed subspace $S$ of $L_2$, $S_n := \mathrm{span}(X_1, \ldots, X_n)$, $K_1 := \mathrm{span}(X_2, \ldots, X_n)$ and $K_2 := \mathrm{span}(X_1 - P_{K_1}X_1)$. $K_1$ and $K_2$ are orthogonal and $S_n = K_1 \oplus K_2$. We have

$$\hat{X}_{n+1} = \Pi_{K_1}X_{n+1} + \Pi_{K_2}X_{n+1} = \Pi_{K_1}X_{n+1} + a(X_1 - \Pi_{K_1}X_1),$$

for

$$a = \mathbb{E}[X_{n+1}(X_1 - \Pi_{K_1}X_1)]/\,\mathbb{E}[(X_1 - \Pi_{K_1}X_1)^2].$$

Stationarity ensures that $X^n$, $X^{n:1} := (X_n, \ldots, X_1)'$ and $X^{2:(n+1)} := (X_2, \ldots, X_{n+1})'$ all have the same covariance matrix, $\Gamma_n$, which is positive definite by hypothesis. It follows that

$$\begin{aligned} \lambda_{n-1} &= \mathbb{E}[X^{(n-1):1}(X^{(n-1):1})']^{-1}\,\mathbb{E}[X^{(n-1):1}X_n] \\ &= \mathbb{E}[X^{2:n}(X^{2:n})']^{-1}\,\mathbb{E}[X^{2:n}X_1] \\ &= \mathbb{E}[X^{n:2}(X^{n:2})']^{-1}\,\mathbb{E}[X^{n:2}X_{n+1}]. \end{aligned}$$

where the superscripts are understood in the natural way. As such,

$$\Pi_{K_1}X_1 = E[X_1|X^{2:n}] = \sum_{j=1}^{n-1} \lambda_{n-1,j}X_{j+1},$$

$$\Pi_{K_1}X_{n+1} = E[X_{n+1}|X^{n:2}] = \sum_{j=1}^{n-1} \lambda_{n-1,j}X_{n+1-j},$$

and

$$\mathbb{E}[(X_1 - \Pi_{K_1}X_1)^2] = \mathbb{E}[(X_{n+1} - \Pi_{K_1}X_{n+1})^2] = \mathbb{E}[(X_n - \hat{X}_n)^2] = \nu_{n-1}.$$

Combining these we obtain that

$$\hat{X}_{n+1} = aX_1 + \sum_{j=1}^{n-1}[\lambda_{n-1,j} - a\lambda_{n-1,n-j}]X_{n+1-j},$$

and

$$\begin{aligned} a &= \left( \mathbb{E}[X_{n+1}X_1] - \sum_{j=1}^{n-1} \lambda_{n-1,j}\,\mathbb{E}[X_{n+1}X_{j+1}] \right) \nu_{n-1}^{-1} \\ &= \left( \gamma(n) - \sum_{j=1}^{n-1} \lambda_{n-1,j}\gamma(n-j) \right) \nu_{n-1}^{-1}. \end{aligned}$$

Since $\Gamma_n$ is positive definite, the representation $\hat{X}_{n+1} = \lambda_n' X^{n:1}$ is unique. Comparison of coefficients with the representation in the penultimate display, one has $a = \lambda_{n,n}$ and $\lambda_{n,j} = \lambda_{n-1,j} - a\lambda_{n-1,n-j}$.

For the mean squared error, we have

$$\begin{aligned} \nu_n &= \mathbb{E}[(X_{n+1} - \hat{X}_{n+1})^2] \\ &= \mathbb{E}\left[(X_{n+1} - \Pi_{K_1}X_{n+1} - \Pi_{K_2}X_{n+1})^2\right] \\ &= \mathbb{E}\left[(X_{n+1} - \Pi_{K_1}X_{n+1})^2\right] + \mathbb{E}\left[(\Pi_{K_2}X_{n+1})^2\right] - 2\,\mathbb{E}\left[(X_{n+1} - \Pi_{K_1}X_{n+1})\Pi_{K_2}X_{n+1}\right] \\ &= \nu_{n-1} + a^2\nu_{n-1} - 2a\,\mathbb{E}[X_{n+1}(X_1 - \Pi_{K_1}X_2)], \end{aligned}$$

in the last line using the orthogonality and $\Pi_{K_2}X_{n+1} = a(X_1 - \Pi_{K_1}X_1)$. Using the definition of $a$ and the equivalent representations of $\nu_{n-1}$ one has that

$$2a\,\mathbb{E}[X_{n+1}(X_1 - \Pi_{K_1}X_1)] = 2a^2\nu_{n-1}$$

hence $\nu_n = (1 + a^2 - 2a^2)\nu_{n-1} = (1 - a^2)\nu_{n-1}$. □

The function $\alpha : \mathbb{N} \to \mathbb{R}$ defined by $\alpha(n) := \lambda_{n,n}$ where $\lambda_{n,n}$ is as in (75) is called the *partial autocorrelation function.* This is the correlation between $X_{n+1}$ and $X_1$ after controlling linearly for the intermediate observations $X_2, \dots, X_n$.

PROPOSITION 5.6: *The partial autocorrelation function $\alpha$ satisfies*

$$\alpha(k) := \lambda_{n,n} = \operatorname{Corr}(X_{k+1} - E[X_{k+1}|(X_2, \dots, X_k)], X_1 - E[X_1|(X_2, \dots, X_k)]).$$

**Proof:* Using the notation of the proof of Proposition 5.5,

$$\begin{aligned} \lambda_{n,n} = a &= \mathbb{E}\left[X_{n+1}(X_1 - \Pi_{K_1}X_1)\right] / \,\mathbb{E}\left[(X_1 - \Pi_{K_1}X_1)^2\right] \\ &= \mathbb{E}\left[(X_{n+1} - \Pi_{K_1}X_{n+1})(X_1 - \Pi_{K_1}X_1)\right] / \,\mathbb{E}\left[(X_1 - \Pi_{K_1}X_1)^2\right], \end{aligned}$$

by orthogonality. In the proof of Proposition 5.5 it was shown that $\mathbb{E}\left[(X_1 - \Pi_{K_1}X_1)^2\right] = \mathbb{E}\left[(X_{n+1} - \Pi_{K_1}X_{n+1})^2\right]$ hence the last line in the display above is the claimed correlation. □

The following innovations algorithm is more generally applicable than the above Durbin – Levinson algorithm, since it applies also to non-stationary processes. Firstly define $\hat{X}_1 = 0$

and then let $\hat{X}_{n+1} = E[X_{n+1}|X^n]$. Since each $\hat{X}_l$ is a linear combination of $X_1, \dots, X_{l-1}$, we have $\text{span}(X_1, \dots, X_n) = \text{span}(X_1 - \hat{X}_1, \dots, X_n - \hat{X}_n)$, so that there are $\theta_{n,j}$ for $n \in \mathbb{N}$ and $j \in \{1, \dots, n\}$ such that

$$\hat{X}_{n+1} = \sum_{j=1}^{n} \theta_{n,j}(X_{n+1-j} - \hat{X}_{n+1-j}). \tag{77}$$

Proposition 5.7 [Innovations algorithm]: *Suppose that $(X_t)_{t\in\mathbb{Z}}$ is a zero-mean process with autocovariance function $\gamma$. Suppose that $\Gamma = \text{Var}(X^n)$ is positive definite for each $n \in \mathbb{N}$. Then $\theta_{n,j}$ in (77) and $\nu_n$ in (76) can be found from $\nu_0 = \Gamma(1,1)$,*

$$\theta_{n,n-k} = \nu_k^{-1}\left(\Gamma(n+1,k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{n,n-j}v_j\right) \qquad 0 \le k < n,$$

*and*

$$\nu_n = \Gamma(n+1,n+1) - \sum_{j=0}^{n-1} \theta_{n,n-j}^2 \nu_j.$$

**Proof:* The elements of $B = (X_1 - \hat{X}_1, \dots X_n - \hat{X}_n)'$ are mutually orthogonal in $L_2$ as $X_i - \hat{X}_i \in \text{span}(X_1 - \hat{X}_1, \dots X_{j-1} - \hat{X}_{j-1})$ for $i < j$ and $(X_j - \hat{X}_j) \perp \text{span}(X_1 - \hat{X}_1, \dots X_{j-1} - \hat{X}_{j-1})$ since $\hat{X}_j$ is the orthogonal projection of $X_j$ onto the latter subspace. Given this, taking the inner product in $L_2$ of each side of (77) with $X_{k+1} - \hat{X}_{k+1}$ for $0 \le k < n$ yields

$$\mathbb{E}[\hat{X}_{n+1}(X_{k+1} - \hat{X}_{k+1})] = \sum_{j=1}^{n} \theta_{n,j}\, \mathbb{E}[(X_{n+1-j} - \hat{X}_{n+1-j})(X_{k+1} - \hat{X}_{k+1})] = \theta_{n,n-k}\nu_k.$$

Since $(X_{n+1} - \hat{X}_{n+1}) \perp (X_{k+1} - \hat{X}_{k+1})$, the above implies that

$$\theta_{n,n-k} = \mathbb{E}[X_{n+1}(X_{k+1} - \hat{X}_{k+1})]\nu_k^{-1}, \qquad k = 0, \dots, n-1.$$

So by (77) with $n = k$,

$$\begin{aligned}\theta_{n,n-k} &= \left(\mathbb{E}[X_{n+1}X_{k+1} - \sum_{j=1}^{k} \theta_{k,j}X_{n+1}(X_{k+1-j} - \hat{X}_{k+1-j})]\right)\nu_k^{-1} \\ &= \left(\gamma(n+1,k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j}\, \mathbb{E}[X_{n+1}(X_{j+1} - \hat{X}_{j+1})]\right)\nu_k^{-1} \\ &= \left(\gamma(n+1,k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{n,n-j}\nu_j\right)\nu_k^{-1},\end{aligned}$$

where the last line uses the expression for $\theta_{n,n-k}$ given in the prior display. For the mean squared error, by properties of orthogonal projections

$$\nu_n = \mathbb{E}[(X_{n+1} - \hat{X}_{n+1})^2] = \mathbb{E}[(X_{n+1})^2] - \mathbb{E}[(\hat{X}_{n+1})^2] = \gamma(n+1,n+1) - \sum_{k=0}^{n-1} \theta_{n,n-k}^2 \nu_k,$$

since the elements of $B$ are orthogonal. □

Remark 5.2: *The innovations algorithm can also be used to find h-step ahead predictors. For $h \geq 1$, by Lemma 5.5 and (77),*

$$\begin{aligned} E[X_{n+h}|X^n] &= E\left[E\left[[X_{n+h}|X^{n+h-1}\right]|X^n\right] \\ &= E[\hat{X}_{n+h}|X^n] \\ &= E\left[\sum_{j=1}^{n+h-1} \theta_{n+h-1,j}(X_{n+h-j} - \hat{X}_{n+h-j})\middle| X^n\right] \\ &= \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}(X_{n+h-j} - \hat{X}_{n+h-j}), \end{aligned}$$

*where the $\theta_{n,j}$ are determined by the innovations algorithm. Note the last equality follows from*

$$E\left[X_{n+k} - \hat{X}_{n+k}\middle|X^n\right] = E\left[E\left[X_{n+k} - \hat{X}_{n+k}\middle|X^{n+k}\right]\middle|X^n\right] = E[0|X^n] = 0, \quad k \geq 0.$$

Example 5.12 [MA(1)]: Suppose that $(X_t)_{t\in\mathbb{Z}}$ is an MA(1) process. We will calculate the best linear prediction of $X_t$ given $X_{t-1}, \ldots, X_1$ using the innovations algorithm. We have

$$X_t = \epsilon_t + \theta\epsilon_{t-1}, \quad \epsilon_t \sim \mathrm{WN}(0, \sigma^2).$$

By Lemma 5.2, $\Gamma(i,j) = 0$ for $|i-j| > 1$, $\Gamma(i,i) = \sigma^2(1+\theta^2)$ and $\Gamma(i,i+1) = \sigma^2\theta$. Applying the innovations algorithm we obtain $\nu_0 = \sigma^2(1+\theta^2)$, then $\theta_{1,1} = \sigma^2\theta/\sigma^2(1+\theta^2)$ and $\nu_1 = \sigma^2(1+\theta^2) - \theta^2\sigma^2/(1+\theta^2)$. Letting $n \geq 2$, we have that for all $2 \leq j \leq n$, $\theta_{n,j} = 0$, and

$$\theta_{n,1} = \nu_{n-1}^{-1}\theta\sigma^2, \qquad \nu_n = \sigma^2(1+\theta^2) - \upsilon_{n-1}^{-1}\theta^2\sigma^4.$$ △

The innovations algorithm can be simplified if $(X_t)_{t\in\mathbb{Z}}$ is an ARMA$(p,q)$ process. Rather than applying the innovations algorithm to the process $(X_t)_{t\in\mathbb{Z}}$ directly, we will instead apply it to a transformed process $(W_t)_{t\in\mathbb{Z}}$ where, with $m = \max(p,q)$,

$$W_t = \begin{cases} \sigma^{-1}X_t & \text{if } t = 1, \ldots, m \\ \sigma^{-1}\Phi(L)X_t & \text{if } t > m \end{cases}. \tag{78}$$

To keep the notation as simple as possible suppose that $p, q \geq 1$ (with $\phi_1 = 0$ or $\theta_1 = 0$ permitted) and let $\theta_0 = 1$ and $\theta_j = 0$ for $j > q$. For these new variables, we have that $\mathrm{span}(X_1, \ldots, X_n) = \mathrm{span}(W_1, \ldots, W_n)$ for $n \geq 1$ and we write $\hat{W}_{n+1} = E[W_{n+1}|W^n]$ (with $\hat{W}_1 = 0$). If $\gamma_X$ is the autocovariance function of $(X_t)_{t\in\mathbb{Z}}$ then the autocovariance function of

$(W_t)_{t\in\mathbb{Z}}$ is $\kappa(s,t) = \mathbb{E}[W_s W_t]$ and

$$\kappa(s,t) = \begin{cases} \sigma^{-2}\gamma_X(s-t) & \text{if } 1 \le s,t \le m \\ \sigma^{-2}\left[\gamma_X(s-t) - \sum_{r=1}^{p}\phi_r\gamma_X(r-|s-t|)\right] & \text{if } \min(s,t) \le m < \max(s,t) \le 2m \\ \sum_{r=0}^{q}\theta_r\theta_{r+|s-t|} & \text{if } \min(s,t) > m \\ 0 & \text{otherwise} \end{cases}. \tag{79}$$

Proposition 5.8: *Suppose that $(X_t)_{t\in\mathbb{Z}}$ is an ARMA$(p,q)$ process. Suppose that $\Gamma = \mathrm{Var}(X^n)$ is positive definite for each $n \in \mathbb{N}$. Applying the innovations algorithm (Proposition 5.7) to $W_t$ yields*

$$\hat{W}_{n+1} \begin{cases} \sum_{j=1}^{n}\theta_{n,j}(W_{n+1-j} - \hat{W}_{n+1-j}) & \text{for } 1 \le n < m \\ \sum_{j=1}^{q}\theta_{n,j}(W_{n+1-j} - \hat{W}_{n+1-j}) & \text{for } n \ge m \end{cases}.$$

*Proof.* This will follow from Proposition 5.7 provided we show that $\theta_{n,j} = 0$ when $n \ge m$ and $j > q$. By (79), $\kappa(n,k) = 0$ if $n > m$ and $|n-k| > q$ (which ensures that $\theta_{r+|n-k|} = 0$ for all $r = 0,\dots,q$). Then, using Proposition 5.7, we have that for $n,k$ such that $n \ge m$ and $n-k > q$

$$\theta_{n,n-k} = \nu_k^{-1}\left(\kappa(n+1,k+1) - \sum_{j=0}^{k-1}\theta_{k,k-j}\theta_{n,n-j}v_j\right) = \nu_k^{-1}\left(-\sum_{j=0}^{k-1}\theta_{k,k-j}\theta_{n,n-j}v_j\right).$$

For $k = 0$, this gives zero since the sum is empty. For $k = 1$, the sum only contains $\theta_{k,k}\theta_{n,n}\nu_0$ and $\theta_{n,n} = 0$ by the previous step. Each such $\theta_{n,n-k} = 0$ until $n-k \le q$ and since these terms appear multiplicitatively in the sum, the result follows. □

Proposition 5.9: *Suppose that $(X_t)_{t\in\mathbb{Z}}$ is an ARMA$(p,q)$ process. Suppose that $\Gamma = \mathrm{Var}(X^n)$ is positive definite for each $n \in \mathbb{N}$. Then,*

$$\hat{X}_{n+1} = \begin{cases} \sum_{j=1}^{n}\theta_{n,j}(X_{n+1-j} - \hat{X}_{n+1-j}) & \text{for } 1 \le n < m \\ \sum_{r=1}^{p}\phi_r X_{n+1-r} + \sum_{j=1}^{q}\theta_{n,j}(X_{n+1-j} - \hat{X}_{n+1-j}) & \text{for } n \ge m \end{cases},$$

*and $\mathbb{E}(X_{n+1} - \hat{X}_{n+1})^2 = \sigma^2\,\mathbb{E}(W_{n+1} - \hat{W}_{n+1})^2 = \sigma^2 r_n$, where $r_n := \mathbb{E}(W_{n+1} - \hat{W}_{n+1})^2$ are calculated during the application of the innovations algorithm to the $W_t$.*

*Proof.* By the linearity in Lemma 5.5 and equation (78), we have

$$\begin{aligned} \hat{W}_t = \mathbb{E}[W_t|W^{t-1}] &= \begin{cases} \mathbb{E}[\sigma^{-1}X_t|X^{t-1}] & \text{for } t = 1,\dots,m \\ \mathbb{E}[\sigma^{-1}\Phi(L)X_t|X^{t-1}] & \text{for } t > m \end{cases} \\ &= \begin{cases} \sigma^{-1}\hat{X}_t & \text{for } t = 1,\dots,m \\ \sigma^{-1}[\hat{X}_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p}] & \text{for } t > m \end{cases}. \end{aligned} \tag{80}$$

Using equation (78) again, this yields that

$$X_t - \hat{X}_t = \sigma[W_t - \hat{W}_t], \quad t \geq 1. \tag{81}$$

Then substitute the preceding displays into the conclusion of Proposition 5.8 to arrive at the first conclusion. The second follows directly from the preceding display. □

As before, we can also find $h$-step ahead predictors.

Remark 5.3: *Suppose that $(X_t)_{t\in\mathbb{Z}}$ is an ARMA$(p,q)$ process. By equation (78) and Remark 5.2*

$$\begin{aligned} E[X_{n+h}|X^n] &= \sigma\, \mathbb{E}[W_{n+h}|W^t] \\ &= \begin{cases} E\left[X_{n+h}|X^n\right], & 1 \leq h \leq m-n \\ E\left[\Phi(L)X_{n+h}|X^n\right], & h > m-n \end{cases} \\ &= \begin{cases} \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}(X_{n+h-j} - \hat{X}_{n+h-j}), & 1 \leq h \leq m-n \\ \sum_{r=1}^{p} \phi_r E[X_{n+h-r}|X^n] + \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}(X_{n+h-j} - \hat{X}_{n+h-j}), & h > m-n \end{cases}. \end{aligned}$$

*If $n > m = \max(p,q)$ then for all $h \geq 1$ since $\theta_{n+h-1,j} = 0$ for $j > q$ (cf. the proof of Proposition 5.8) and hence*

$$E[X_{n+h}|X^n] = \sum_{r=1}^{p} \phi_r E[X_{n+h-r}|X^n] + \sum_{j=h}^{q} \theta_{n+h-1,j}(X_{n+h-j} - \hat{X}_{n+h-j}).$$

## 5.4 Estimation & lag order selection

Estimation of the parameters $\gamma := (\phi, \theta, \sigma^2)$ of a causal, invertible, stationary ARMA$(p,q)$ model is usually done by (conditional) maximum likelihood, assuming that the process $(X_t)_{t\in\mathbb{Z}}$ is Gaussian with mean zero and covariance function $\kappa(s,t) := \mathbb{E}[X_s X_t]$. Let $X^n = (X_1, \dots, X_n)'$ and $\Gamma_n = \mathbb{E}[X^n(X^n)']$. Note that $\Gamma_n = \Gamma_n(\gamma)$ is a function of $\gamma$. In this setting the (conditional) likelihood of $X^n$ is

$$L(\gamma) := (2\pi)^{-n/2}(\det \Gamma_n)^{-1/2} \exp\left(-\frac{1}{2}(X^n)'\Gamma_n^{-1}X^n\right). \tag{82}$$

Our parameter estimates are then taken as the values which maximise this *likelihood function*. In practice this is computed *numerically*. Computation of $\det \Gamma_n$ and the inverse $\Gamma_n^{-1}$ is computationally intensive and can be avoided by using the innovations algorithm.

Let $\theta_{i,j}$ be the coefficients and $r_j$ the mean-squared errors calculated from the covariance matrix $\Gamma_n(\gamma)$ using Proposition 5.9. Define the $n \times n$ lower triangular matrix $C$ and $n \times n$

diagonal matrix $D$ as

$$C = \begin{bmatrix} \theta_{0,0} & 0 & \cdots & 0 \\ \theta_{1,1} & \theta_{1,0} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \cdots & \theta_{n,0} \end{bmatrix}, \qquad D = \sigma^2 \begin{bmatrix} r_0 & 0 & \cdots & 0 \\ 0 & r_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{n-1} \end{bmatrix},$$

where each $\theta_{i,0} = 1$. Then, by (77), $\hat{X}^n = (\hat{X}_1, \ldots, \hat{X}_n)$ satisfies

$$\hat{X}^n = (C - I_n)(X^n - \hat{X}^n).$$

Therefore,

$$X^n = C(X^n - \hat{X}^n).$$

As $D$ is the covariance matrix of $X^n - \hat{X}^n$, we have that $\Gamma_n = \mathbb{E}[X^n(X^n)'] = CDC'$.[88] Therefore,

$$\det \Gamma_n = (\det C)^2 (\det D) = \sigma^{2n} \prod_{i=0}^{n-1} r_i,$$

and

$$\begin{aligned} (X^n)'\Gamma_n^{-1}X^n &= (X^n - \hat{X}^n)'C'\Gamma_n^{-1}C(X^n - \hat{X}^n) \\ &= (X^n - \hat{X}^n)'C'(C')^{-1}D^{-1}C^{-1}C(X^n - \hat{X}^n) \\ &= (X^n - \hat{X}^n)'D^{-1}(X^n - \hat{X}^n) \\ &= \sigma^{-2} \sum_{t=1}^{n} \frac{(X_t - \hat{X}_t)^2}{r_{t-1}}. \end{aligned}$$

Plugging these into (82) yields

$$L(\gamma) = (2\pi\sigma^2)^{-n/2} \left( \prod_{i=0}^{n-1} r_i \right)^{-1/2} \exp\left( -\frac{1}{2\sigma^2} \sum_{t=1}^{n} \frac{(X_t - \hat{X}_t)^2}{r_{t-1}} \right).$$

It can be shown that the the maximisers of $L(\gamma)$ satisfy

$$\hat{\sigma}^2 = S(\hat{\phi}, \hat{\theta})/n,$$

where

$$S(\phi, \theta) = \sum_{i=1}^{n} \frac{(X_i - \hat{X}_i)^2}{r_{i-1}},$$

for $\hat{X}_i$ and $r_{i-1}$ which are the outputs of the innovations algorithm applied to the covariance function in (79) given $\phi$ and $\theta$ (which do not depend on $\sigma^2$). $\hat{\phi}$ and $\hat{\theta}$ are the values of $\phi$, $\theta$

[88] Cf. the proof of Proposition 5.9 where it is noted that these elements are orthogonal in $L_2$, i.e. uncorrelated.

which minimise

$$l(\phi, \theta) = \log\left(\frac{1}{n}S(\phi, \theta)\right) + \frac{1}{n}\sum_{i=1}^{n}\log r_{n-i}.$$

In particular we have the following result for the asymptotic distribution of this estimate of $\beta = (\phi', \theta')'$. We omit the proof.

PROPOSITION 5.10: *Suppose that $(X_t)_{t\in\mathbb{Z}}$ is a causal and invertible ARMA$(p, q)$ process where $\Phi$ and $\Theta$ have no common factors and $\epsilon_t \sim IID(0, \sigma^2)$.*

*Let $C$ be the set of values of $\beta = (\phi, \theta)$ such that if $\Phi'$ and $\Theta'$ are the implied lag polynomials, $\Phi'(z)\Theta'(z) \neq 0$ for $|z| \leq 1$, $\phi_p \neq 0$, $\theta_q \neq 0$ and $\Phi'$ and $\Theta'$ have no common factors.*

*Then, if $\hat{\beta} = \arg\min_{\beta\in C} l(\phi, \theta)$,*

$$\sqrt{n}(\hat{\beta} - \beta) \rightsquigarrow \mathcal{N}(0, \Sigma(\beta)), \quad \Sigma(\beta) = \sigma^2 \begin{bmatrix} \mathbb{E}[U^t(U^t)'] & \mathbb{E}[U^t(V^t)'] \\ \mathbb{E}[V^t(U^t)'] & \mathbb{E}[V^t(V^t)'] \end{bmatrix}^{-1},$$

*where $U^t = (U_t, \ldots, U_{t+1-p})'$, $V_t = (V_t, \ldots, V_{t+1-q})'$ are the autoregressive processes*

$$\Phi(L)U_t = \epsilon_t, \qquad \Theta(L)V_t = \epsilon_t.$$

*If $p = 0$, then $\Sigma(\beta) = \sigma^2\,\mathbb{E}[V^t(V^t)']^{-1}$ and if $q = 0$ then $\Sigma(\beta) = \sigma^2\,\mathbb{E}[U^t(U^t)']^{-1}$.*

We now turn to the determination of the lag orders $p, q$. Analogously to the discussion of overfitting in linear regression models, we can typically achieve a better in-sample fit by including more parameters (i.e. large $p$ and/or $q$). However by doing so we run the risk of overfitting and therefore we typically choose the parameters (including $p, q$) with reference to a regularised criterion. In this context (and some others) these are often called "information criteria". We will briefly dicuss 3 such criteria: Akaike information criterion (AIC), AICC and the Bayesian information criterion (BIC).

All have the same basic form:

$$\begin{aligned} AIC(\phi, \theta) &= -2\log L(\phi, \theta, S(\phi, \theta)/n) + 2(p + q + 1) \\ AICC(\phi, \theta) &= -2\log L(\phi, \theta, S(\phi, \theta)/n) + \frac{2(p + q + 1)n}{n - p - q - 2} \\ BIC(\phi, \theta) &= -2\log L(\phi, \theta, S(\phi, \theta)/n) + (p + q + 1)\log n. \end{aligned}$$

The AIC and AICC have very similar forms, but the AICC penalises large order (higher $p$, $q$) models more strongly. The BIC can be shown to have a consistency property: if $\hat{p}_n$ and $\hat{q}_n$ are the chosen orders and $(X_t)_{t\in\mathbb{Z}}$ is an ARMA$(p, q)$ process then as $n \to \infty$, $\hat{p}_n \xrightarrow{P} p$ and $\hat{q}_n \xrightarrow{P} q$. The AIC and AICC do not have this property but they do have an efficiency property which the BIC does not share. See section 5.5 of [3] for further detail and references.

These information criteria are used as follows: $p, q$ and the corresponding $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$ are chosen to minimise $IC(\phi, \theta)$ where $IC$ is one of $AIC, AICC, BIC$.

# References

[1] P. Billingsley. *Probability and Measure.* Wiley, third edition, 1995.

[2] H. C. Bjørnland and L. A. Thorsrud. *Applied time series for macroeconomics.* Gyldendal akademisk, second edition, 2015.

[3] P. J. Brockwell and R. A. Davis. *Introduction to Time Series and Forecasting.* Springer Texts in Statistics. Springer New York, third edition, 2016.

[4] G. Casella and R.L. Berger. *Statistical Inference.* Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002.

[5] R. Durrett. *Probability Theory and Examples.* Cambridge University Press, fifth edition, 2019.

[6] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer series in statistics. Springer, 2009.

[7] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, second edition, 2013.

[8] R. W. Keener. *Theoretical Statistics: Topics for a Core Course.* Springer Texts in Statistics. Springer New York, 2010.

[9] E. L. Lehmann and G. Casella. *Theory of Point Estimation.* Springer Texts in Statistics. Springer New York, second edition, 1998.

[10] E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses.* Springer Texts in Statistics. Springer New York, third edition, 2005.

[11] E. L. Lehmann and Henry Scheffé. Completeness, similar regions, and unbiased estimation: Part i. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 10(4):305–340, 1950.

[12] Whitney K. Newey and Daniel McFadden. Chapter 36 large sample estimation and hypothesis testing. volume 4 of *Handbook of Econometrics*, pages 2111–2245. Elsevier, 1994.

[13] M.J. Schervish. *Theory of Statistics.* Springer Series in Statistics. Springer New York, 1995.

[14] J. Shao. *Mathematical Statistics.* Springer Texts in Statistics. Springer New York, 2003.

[15] Gilbert Strang. *Introduction to Linear Algebra.* Wellesley-Cambridge Press, 5th edition, 2016.

[16] A. W. van der Vaart. *Asymptotic Statistics.* Cambridge University Press, 1998.