# Statistical Methods in Finance

Dr Cristopher Salvi

Imperial College London, Autumn 2023

## Contents

# 1 Matrix analysis

For $m, n$ integers, we shall denote by $\mathbb{R}^{m\times n}$ the space of $m \times n$ real matrices. In the sequel, $I_n$ will denote the identity matrix in $\mathbb{R}^{n\times n}$ and $O_n$ will denote the null matrix in $\mathbb{R}^{m\times n}$. Given a matrix $A \in \mathbb{R}^{m\times n}$ we will denote by $A^\top \in \mathbb{R}^{n\times m}$ its transpose. A square matrix $A \in \mathbb{R}^{n\times n}$ is called orthogonal if $AA^\top = A^\top A = I_n$. We will denote by $A^{-1} \in \mathbb{R}^{n\times n}$ its inverse, if it exists, defined by $AA^{-1} = A^{-1}A = I_n$. We will make use of the Euclidean inner product $\langle\cdot,\cdot\rangle$ on $\mathbb{R}^n$, defined by $\langle \mathrm{x}, \mathrm{y}\rangle = \mathrm{x}^\top \mathrm{y}$. Recall that the rank of a matrix $A \in \mathbb{R}^{m\times n}$, denoted by $\mathrm{rank}(A)$, is the maximum number of linearly independent rows, or equivalently the maximum number of linearly independent columns.

**Exercise 1.1.** *Let $A \in \mathbb{R}^{m\times n}$. Prove the following statements.*

- $0 \leq \mathrm{rank}(A) \leq \min\{m, n\}$.
- $\mathrm{rank}(A) = \mathrm{rank}(A^\top) = \mathrm{rank}(AA^\top)$.

For a square $n \times n$ matrix $A$ recall the definitions of trace and determinants:

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

$$\det(A) = \sum_{\sigma\in\mathcal{S}_n} \mathrm{sgn}(\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)},$$

where $\mathcal{S}_n$ is the symmetric group of permutations of size $n$, and $\mathrm{sgn}(\sigma) = +1$ if the permutation $\sigma$ can be written as an even number of transpositions, and $-1$ otherwise.

**Exercise 1.2.** *Let $A, B \in \mathbb{R}^{n\times n}$ be square matrices. Prove the following statements.*

- $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$.
- $\mathrm{Tr}(A^\top) = \mathrm{Tr}(A)$.
- $\det(A) \neq 0$ *iff $A$ is invertible.*
- $\det(\alpha A) = \alpha^n \det(A)$.
- $\det(AB) = \det(BA) = \det(A)\det(B)$.
- *If $A \in \mathbb{R}^{n\times n}$ and $\det(A) \neq 0$, then $\det(A)^{-1} = \det(A^{-1})$.*
- *If $A$ is orthogonal, then $|\det(A)| = 1$.*

**Definition 1.3.** *A square matrix $A \in \mathbb{R}^{n\times n}$ is said positive semi-definite if for all $\mathrm{x} \in \mathbb{R}^n$ one has $\mathrm{x}^\top A\mathrm{x} \geq 0$. It is said positive definite if the inequality is strict. We denote by P(S)D$_n$ the set of all $n \times n$ positive (semi-) definite matrices.*

**Exercise 1.4.** *Prove the following statements.*

- *The identity matrix $I_n$ is positive definite.*
- *For any matrix $A \in \mathbb{R}^{m\times n}$, $A^\top A$ is symmetric and positive semi-definite.*

## 1.1 Spectral theory for matrices

In this section, we consider a square real-valued matrix $A \in \mathbb{R}^{n\times n}$. The spectral theory for matrices is based on the following definition.

**Definition 1.5.** *The characteristic polynomial $\mathcal{P}_A$ of the matrix $A$ is defined as*

$$\mathcal{P}_A(\lambda) := \det(A - \lambda I).$$

It is easy to see that $\deg(\mathcal{P}_A) = n$. The $n$ roots (possibly complex, possibly repeated) of $\mathcal{P}_A$ are called the eigenvalues of $A$. Usually the eigenvalues are ordered

$$\lambda_1(A) \geq \lambda_2(A) \geq ... \geq \lambda_n(A).$$

We will remove the dependency on $A$ and simply write $\lambda_i(A) = \lambda_i$ when clear from the context. We will denote the set of eigenvalues by $\sigma(A)$. For any $\lambda \in \sigma(A)$, a non-null vector $\mathrm{u} \in \mathbb{R}^n$ satisfying $A\mathrm{u} = \lambda\mathrm{u}$ is called the associated eigenvector.

**Exercise 1.6.** *Prove the following statements.*

- *The eigenvalues of a square symmetric matrix are real.*
- *Let $P$ be a polynomial. For any $\lambda \in \sigma(A)$, then $P(\lambda) \in \sigma(P(A))$.*
- $\mathcal{P}_A(A) = 0$.
- $\det(A) = \prod_{\lambda\in\sigma(A)} \lambda$ *and* $\mathrm{Tr}(A) = \sum_{\lambda\in\sigma(A)} \lambda$.

**Theorem 1.7** (Spectral decomposition)**.** *Any symmetric matrix $A \in \mathbb{R}^{n\times n}$ is diagonalisable, that is it admits a decomposition of the form*

$$A = U\Lambda U^\top = \sum_{i=1}^{n} \lambda_i u_i u_i^\top$$

*where $\Lambda = \mathrm{Diag}(\lambda_1, ..., \lambda_n)$ is the diagonal matrix of all eigenvalues of $A$, and $U$ is an orthogonal matrix consisting of unit eigenvectors of $A$.*

For a proof see for instance [HJ12, Section 2.5].

**Exercise 1.8.** *Show that for any symmetric $n \times n$ matrix $A$ with ordered eigenvalues $\lambda_1 \geq ... \geq \lambda_n$ and corresponding eigenvectors $v_1, ..., v_n$ one has*

$$\max_{\|\mathrm{x}\|_2=1} \mathrm{x}^\top A\mathrm{x} = \lambda_1$$

*where the maximum occurs at $\mathrm{x} = \pm v_1$ and*

$$\min_{\|\mathrm{x}\|_2=1} \mathrm{x}^\top A\mathrm{x} = \lambda_n$$

*where the minimum occurs at $\mathrm{x} = \pm v_n$.*

**Exercise 1.9** (Lidskii inequality)**.** *Let $A, B \in \mathbb{R}^{n\times n}$ be two symmetric matrices. Given $p \in \{1, ..., n\}$ and indices $1 \leq i_1 \leq ... \leq i_p \leq n$ show that*

$$\sum_{j=1}^{p} \lambda_{i_j}(A+B) \leq \sum_{j=1}^{p} \lambda_{i_j}(A) + \sum_{i=1}^{p} \lambda_i(B).$$

**Remark 1.10.** *Diagonalisation allows, among many other things, to efficiently compute powers of a matrix. Consider for example the function $P(x) = x^\alpha$, for $\alpha \in \mathbb{R}$, applied to a symmetric matrix $A$. Then $P(A) = U\Lambda^\alpha U^\top$, where $\Lambda^\alpha = \mathrm{Diag}\,(\lambda_1^\alpha, \ldots, \lambda_n^\alpha)$.*

**Exercise 1.11.** *Prove the following statements.*

- *If $A \in \mathbb{R}^{n\times n}$ is symmetric then its rank is equal to the number of its non-zero eigenvalues (counting according to their multiplicities).*
- *$A \in PSD_n \iff \min_{\lambda\in\sigma(A)} \lambda \geq 0$, i.e. all its eigenvalues are non-negative.*
- *If $A \in PSD_n$ then $A$ is invertible.*

**Theorem 1.12** (Gram-Schmidt orthogonalisation)**.** *If $\{v_1, ..., v_m\}$ are linearly independent vectors in $\mathbb{R}^n$ then there exists a set $\{u_1, ..., u_m\}$ of orthonormal vectors defined recursively as $u_1 = \frac{v_1}{\|v_1\|}$ and*

$$u_k = v_k - \sum_{j=1}^{k-1} \frac{\langle v_k, u_j\rangle}{\|u_j\|_2^2} u_j$$

*such that*

$$Span(v_1, ..., v_k) := Span(u_1, ..., u_k), \quad \forall k = 1, ..., m.$$

For a proof see for instance [HJ12, Section 0.6].

## 1.2 Singular value decomposition

Many matrices (for example non-square matrices) cannot be diagonalised. The singular value decomposition (SVD) provides an alternative matrix factorization applicable to all real matrices.

**Theorem 1.13** (Singular value decomposition)**.** *Let $A \in \mathbb{R}^{m\times n}$ be a matrix of rank-$r$. Then $A$ admits the following decomposition as a sum of $r$ rank-1 matrices*

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

*where $u_1 ... u_r \in \mathbb{R}^m$ (called the left singular vectors of $A$) and $v_1, ..., v_r \in \mathbb{R}^n$ (called the right singular vectors of $A$) are orthonormal sets of vectors and $\Sigma \in \mathbb{R}^{r\times r}$ is a diagonal matrix the form $\Sigma = \mathrm{Diag}(\sigma_1, ..., \sigma_r)$ where*

$$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0$$

*are all non-negative real numbers (called the singular values of $A$).*

*Proof.* For any matrix $A \in \mathbb{R}^{m\times n}$, the matrix $A^\top A$ is symmetric, therefore by the spectral decomposition it can be diagonalised as

$$A^\top A = \hat{V}\mathrm{Diag}(\lambda_1, ..., \lambda_n)\hat{V}^\top = \sum_{i=1}^{n} \lambda_i v_i v_i^\top$$

where $\hat{V} = [v_1 ... v_n]$ is an orthonormal matrix of eigenvectors. Because $A^\top A$ is positive semi-definite its eigenvalues $\lambda_1, ..., \lambda_n$ are all non-negative and real. We can order them as $\lambda_1 \geq \lambda_2 \geq \lambda_r > 0$ and $\lambda_{r+1} = ... = \lambda_n = 0$. Set

$$V := \begin{pmatrix} v_1^\top \\ \vdots \\ v_r^\top \end{pmatrix} \in \mathbb{R}^{n\times r}.$$

For any $i \leq r$ define

$$u_i = \frac{1}{\sqrt{\lambda_i}} A v_i.$$

It is easy to verify (exercise) that $u_1, ..., u_r$ are orthonormal eigenvectors of $AA^\top$. With this, it is easy to verify (exercise) that

$$U^\top A V = \mathrm{Diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_r}) := \Sigma$$

as desired. □

**Remark 1.14.** *The full SVD of $A \in \mathbb{R}^{m\times n}$, as presented classically, consists of the slightly different decomposition*

$$A = \tilde{U} \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} \tilde{V}^\top$$

*where the set $u_1, ..., u_r$ is completed into an orthonormal basis of $\mathbb{R}^m$ via Gram-Schmidt and then add rows of zero vectors to $\Sigma$.*

**Exercise 1.15** (Mirsky inequality)**.** *Let $A, B \in \mathbb{R}^{m\times n}$. Given $p \in \{1, ..., \min\{m, n\}\}$ and $1 \leq i_1 \leq ... \leq i_p \leq \min\{m, n\}$ use Lidskii inequality to show that*

$$\sum_{j=1}^{p} |\sigma_{i_j}(A) - \sigma_{i_j}(B)| \leq \sum_{i=1}^{p} \sigma_i(A - B).$$

**Exercise 1.16** (Von Neumann trace inequality)**.** *For any two matrices $A, B \in \mathbb{R}^{m\times n}$*

$$\mathrm{Tr}(A^\top B) \leq \sum_{i=1}^{\min\{m,n\}} \sigma_i(A)\sigma_i(B).$$

**Exercise 1.17.** *Show that for any matrix $A \in \mathbb{R}^{m\times n}$ the following statements hold.*

- $\|A\|_F^2 := \sum_{i=1}^m \sum_{j=1}^n a_{ij}^2 = \mathrm{Tr}(A^\top A) = \sum_{i=1}^{\min\{m,n\}} \sigma_i^2$.

- $\|A\|_2^2 := \sup_{\mathrm{x}\neq 0} \frac{\|A\mathrm{x}\|_2}{\|\mathrm{x}\|_2} = \sup\{\|A\mathrm{x}\|_2 : \|\mathrm{x}\|_2 = 1\} = \sigma_1$.
- $\|A\|_2^2 \leq \|A\|_F^2$.

SVD has many applications. For example it is useful to compute the so-called pseudo-inverse of a matrix. If $A$ is not a square matrix or has not full rank then it can't be inverted. However, we say that $A^+$ is a generalised inverse of $A$ if $AA^+A = A$. One such generalised inverse can be obtained from the SVD by $A^+ = V\Sigma^{-1}U^\top$; this is known as Moore-Penrose pseudo-inverse. Another important application of SVD in this course will be Principal Component Analysis (PCA), a dimensionality reduction technique which we will see in full details later. Another reason why the SVD is so widely used is that it can be used to find the best low rank approximation to a matrix.

## 1.3 The Eckart-Young-Mirsky Theorem

**Definition 1.18** (Ky Fan norms)**.** *For any $1 \leq p \leq \min\{m, n\}$ define the following family of norms*

$$\|A\|_{(p)} := \sum_{i=1}^{p} \sigma_i(A), \quad A \in \mathbb{R}^{m\times n}.$$

**Theorem 1.19** (Eckart-Young-Mirsky Theorem)**.** *Let $\|\cdot\|$ be a Ky Fan norm. Let $A \in \mathbb{R}^{m\times n}$ be of rank-$r$ with singular value decomposition $A = U\mathrm{Diag}(\sigma_1, ..., \sigma_r)V^\top$. Then, for any $k \leq r$, the following rank-$k$ matrix*

$$A_k = U\mathrm{Diag}(\sigma_1, ..., \sigma_k, 0, ..., 0)V^\top = \sum_{i=1}^{k} \sigma_i u_i v_i^\top.$$

*is the best rank-$k$ approximation of $A$ in the sense that it satisfies*

$$\|A - A_k\| \leq \|A - B\|, \quad \forall B \in \mathbb{R}^{m\times n} \text{ with } \mathrm{rank}(B) \leq k.$$

*Proof.* For a $B \in \mathbb{R}^{m\times n}$ of rank-$k$ we apply the Mirsky inequality to the indices

$$\begin{cases} k+1, ..., k+p & \text{if } p \leq r-k \\ k+1, ..., r & \text{if } p > r-k. \end{cases}$$

In the former case ($p \leq r-k$) we have

$$\|A - B\|_{(p)} := \sum_{i=1}^{p} \sigma_i(A - B) \geq \sum_{j=1}^{p} |\sigma_{k+j}(A) - \sigma_{k+j}(B)| = \sum_{j=1}^{p} \sigma_{k+j}(A) = \|A - A_k\|_{(p)},$$

where used the fact that $\sigma_i(B) = 0$ for any $i > k$.

In latter case ($p > r-k$) one has

$$\|A-B\|_{(p)} := \sum_{i=1}^{p} \sigma_i(A-B) \geq \sum_{i=1}^{r-k} \sigma_i(A-B) \geq \sum_{j=1}^{r-k} |\sigma_{k+j}(A) - \sigma_{k+j}(B)|$$
$$= \sum_{j=1}^{r-k} \sigma_{k+j}(A) = \sum_{i=1}^{p} \sigma_{k+i}(A) = \|A-A_k\|_{(p)}.$$

□

**Remark 1.20.** *A simpler argument can be used to prove the EYM Theorem in the case of the Frobenious norm. Namely, given a matrix $B \in \mathbb{R}^{m\times n}$ with rank$(B) \leq k$, the Von-Neumann trace inequality guarantees that*

$$\mathrm{Tr}(A^\top B) \leq \sum_{i=1}^{r} \sigma_i(A)\sigma_i(B) = \sum_{i=1}^{k} \sigma_i(A)\sigma_i(B).$$

*The conclusion then follows from*

$$\|A-B\|_F^2 - \|A-A_k\|_F^2 = \|A\|_F^2 - \|A-A_k\|_F^2 - 2\mathrm{Tr}(A^\top B) + \|B\|_F^2$$
$$\geq \sum_{i=1}^{k} \sigma_i(A)^2 - 2\sum_{i=1}^{k} \sigma_i(A)\sigma_i(B) + \sum_{i=1}^{k} \sigma_i(B)^2$$
$$= \sum_{i=1}^{k} (\sigma_i(A) - \sigma_i(B))^2 \geq 0,$$

*where we used the fact that $\|A\|_F^2 - \|A-A_k\|_F^2 = \sum_{i=1}^{k} \sigma_i(A)^2$.*

**Remark 1.21.** *The Eckart-Young-Mirsky Theorem also holds for the operator norm $\|\cdot\|_2$. In this case we also have that $\|A-A_k\|_2 = \sigma_{k+1}$.*

*In effect, let $B \in \mathbb{R}^{m\times n}$ be of rank $k$. Recall the definition of null space*

$$Ker(B) = \{\mathrm{x} \in \mathbb{R}^n : B\mathrm{x} = 0\}.$$

*By the rank-nullity theorem*

$$dim(Ker(B)) = n-k.$$

*Consider the $n \times (k+1)$ matrix $V_{k+1} = [v_1 ... v_{k+1}]$ with rank $k+1$. Because*

$$dim(Ker(B)) + dim(Span(v_1, ..., v_{k+1})) = n-k+k+1 = n+1$$

*$Ker(B)$ and $Span(v_1, ..., v_{k+1})$ cannot be disjoint. This means that there exists a vector $w \in Ker(B) \cap Span(v_1, ..., v_{k+1})$ chosen so that $\|w\|_2 = 1$ which we write as $w = \sum_{i=1}^{k+1} \alpha_i v_i$. Then*

$$\|A-B\|_2^2 \geq \|(A-B)w\|_2^2 = \|Aw\|_2^2 = w^\top V \Sigma^2 V^\top w = \sum_{i=1}^{k+1} \sigma_i^2 \alpha_i^2$$
$$\geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} \alpha_i^2 = \sigma_{k+1}^2 = \|A-A_k\|_2^2.$$

*where the first inequality follows by the definition of* $\|\cdot\|_2$*, the second equality from the fact that* $w \in Ker(B)$*, the third from the singular value decomposition* $A = U\Sigma V^\top$*, the fourth from substituting* $w = \sum_{i=1}^{k+1} \alpha_i v_i$*, the fifth inequality from the order of the singular values, and the sixth from the fact that* $\|w\|_2 = 1$*.*

# 2 Probability theory

We provide here a brief overview of standard results in probability theory. We shall consider random variables supported on the whole real line. The results below are not restricted to this case, though, but the notation is simpler then.

Given a random variable $X$, we shall denote

$$F_X(x) = \mathbb{P}(X \leq x), \quad f_X(x) = \mathbb{P}(X = x) \quad \text{and} \quad \phi_X = \mathbb{E}[e^{itX}]$$

the corresponding cumulative distribution functions, densities, and characteristic functions respectively. Note that the characteristic function always exists.

**Proposition 2.1** (Taylor expansion of characteristic function)**.** *Let $X$ be a real random variable with finite $k^{th}$ moment for some $k \geq 1$. Then, its characteristic function $\phi_X$ is $k$-times continuously differentiable, and*

$$\phi_X(t) = \sum_{j=1}^{k} \frac{(it)^j}{j!}\mathbb{E}[X^j] + o(|t|^k)$$

*where where $o(|t|^k)$ is a quantity that goes to zero as $t \to 0$, times $|t|^k$. In particular, we get that for all $0 \leq j \leq k$*

$$\frac{d^j}{dt^j}\phi_X(t) = i^j\mathbb{E}[X^j].$$

*Proof.* Just differentiate the characteristic function with respect to $t$ and exchange derivative and integration using the fact that $e^{itX}$ is a bounded function. □

**Proposition 2.2** (Hölder inequality)**.** *Let $p \in (1, \infty)$ and $q$ such that $p^{-1} + q^{-1} = 1$. If $X$ and $Y$ are random variables such that $\mathbb{E}[|X|^p]$ and $\mathbb{E}[|Y|^q]$ are finite, then $\mathbb{E}[|XY|]$ is finite and satisfies the following inequality*

$$\mathbb{E}[|XY|] \leq \mathbb{E}\left[|X|^p\right]^{1/p} \mathbb{E}\left[|Y|^q\right]^{1/q}.$$

*Proof.* Since the logarithm function is concave, the identity

$$\frac{\log(x)}{p} + \frac{\log(y)}{q} \leq \log\left(\frac{x}{p} + \frac{y}{q}\right)$$

holds for all $x, y > 0$. Exponentiating both sides, this is equivalent to $x^{1/p}y^{1/q} \leq \frac{x}{p} + \frac{y}{q}$. Setting $x = |X|^p/\mathbb{E}[|X|^p]$ and $y = |Y|^q/\mathbb{E}[|Y|^p]$ yields the result directly. □

**Remark 2.3** (Lyapunov inequality)**.** *Let $0 < r < q$, and $X$ a random variable such that $\mathbb{E}[|X|^q]$ is finite. The following is an immediate consequence of Hölder inequality*

$$\mathbb{E}[|X|^r]^{1/r} \leq \mathbb{E}[|X|^q]^{1/q}.$$

**Remark 2.4** (Cauchy-Schwarz inequality)**.** *Let $X$ and $Y$ be two square-integrable random variables with $\mathbb{E}[|XY|]$ finite. Applying Hölder's inequality with $p = q = 2$*

$$\mathbb{E}[XY]^2 \leq \mathbb{E}[|XY|]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2].$$

*The inequalities are equalities if $X$ is a scalar multiple of $Y$.*

**Exercise 2.5.** *Let $X$ be a random variable with finite mean $\mu$ and finite and non-zero variance $\sigma^2$. The kurtosis*[1] *of $X$ is defined as*

$$\kappa := \frac{\mathbb{E}\left[(X-\mu)^4\right]}{\sigma^4}.$$

*The excess kurtosis is defined as $\kappa_+ := \kappa - 3$. Using Lyapunov's Inequality, show that the excess kurtosis is always greater than $-2$. Show that this lower bound is attained for the Bernoulli distribution with parameter* 0.5*.*

**Proposition 2.6** (Jensen inequality)**.** *Let $f$ be a convex function (i.e. $f((1-t)x + ty) \geq (1-t)f(x) + tf(y)$ for all $x, y \in \mathbb{R}$ and $0 \leq t \leq 1$) and $X$ a real-valued random variable such that $\mathbb{E}[f(X)]$ is finite. Then $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$, with equality if and only if either $X$ is constant or $f$ coincides with a line on an interval which contains the range of $X$ almost everywhere (i.e. up to sets of measure zero).*

## 2.1 Convergence of random variables

We recall here the different types of convergence for sequences of random variables $(X_n)_{n\geq 1}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

**Definition 2.7.** *A sequence $(X_n)_{n\geq 1}$ of random variables converges to $X$*

1. *in distribution (or weakly or in law) if the sequence $(F_{X_n})_{n\geq 1}$ of CDFs converges pointwise to a function $F_X$*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

   *for all real number $x$ at which $F_X$, the CDF of $X$, is continuous;*

2. *in probability if for all $\varepsilon > 0$*

$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - X| \geq \varepsilon\right) = 0;$$

3. *almost surely if*

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

4. *in $L^r$ ($r \in \mathbb{N}$) if for all $n \geq 1$ the moments $\mathbb{E}\left[|X_n|^r\right]$ and $\mathbb{E}\left[|X|^r\right]$ are finite, and*

$$\lim_{n\to\infty} \mathbb{E}\left[|X_n - X|^r\right] = 0.$$

[1] The kurtosis is a useful quantity to measure the "fatness" of a distribution's tails.

Even though convergence in law is a weak form of convergence, it has a number of fundamental consequences for applications. We list them here without proof and refer the interested reader to [Bil13] for details.

**Theorem 2.8.** *Assume that the sequence $(X_n)_{n\geq 1}$ converges weakly to $X$. Then the following statements hold.*

1. $\lim_{n\to\infty} \mathbb{E}[h(X_n)] = \mathbb{E}[h(X)]$ *for any bounded and continuous function $h$.*
2. $\lim_{n\to\infty} \mathbb{E}[h(X_n)] = \mathbb{E}[h(X)]$ *for any Lipschitz function $h$.*
3. *The sequence $(h(X_n))_{n\geq 1}$ converges weakly to $h(X)$ for every continuous function $h$ (continuous mapping theorem).*

**Remark 2.9.** *The continuous mapping theorem holds under almost sure convergence.*

**Theorem 2.10** (Lévy's continuity theorem)**.** *A sequence $(X_n)_{n\geq 1}$ converges weakly to $X$ if and only if the sequence of characteristic functions $(\phi_{X_n})_{n\geq 1}$ converges pointwise to the characteristic function $\phi_X$ of $X$ (and $\phi$ is continuous at the origin).*

The following theorem makes the link between the different modes of convergence.

**Theorem 2.11.** *The following statements hold.*

- *Almost sure convergence $\implies$ convergence in probability.*
- *Convergence in probability $\implies$ weak convergence.*
- *Convergence in $L^r$ $\implies$ convergence in probability.*
- *For any $r \geq s \geq 1$, convergence in $L^r$ $\implies$ convergence in $L^s$.*

Next we state without proof the two versions of the law of large numbers, a fundamental result in probability theory.

**Theorem 2.12** (Weak law of large numbers)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. random variables with $X_i = X$ and $X$ integrable. Then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\mathcal{P}} \mathbb{E}[X].$$

A stronger result (strong law of large numbers) states that this convergence actually holds almost surely.

We now turn the attention specifically to Gaussian random variables.

**Exercise 2.13** (Moments and characteristic function of standard normal)**.** *Let $X \sim \mathcal{N}(0,1)$. Show that all moments exists and*

$$\mathbb{E}\left[X^p\right] = \begin{cases} 0 & \text{if } p \text{ is odd,} \\ \dfrac{p!}{2^{p/2}(p/2)!} & \text{if } p \text{ is even,} \end{cases}$$

*and*

$$\mathbb{E}\left[|X|^p\right] = 2^{p/2}\frac{\Gamma((p+1)/2)}{\Gamma(1/2)}, \qquad \textit{for all } p \geq 0$$

*where* $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ *is the Gamma-function.*

An immediate consequence of the two previous propositions is that the characteristic function of $X \sim \mathcal{N}(0,1)$ is given by

$$\phi_X(z) := \mathbb{E}\left[\mathrm{e}^{\mathrm{i}zX}\right] = \exp\left(-\frac{z^2}{2}\right), \qquad \text{for all } z \in \mathbb{R}.$$

**Theorem 2.14** (Central Limit Theorem)**.** *Let* $(X_n)_{n\geq 1}$ *be a sequence of i.i.d. random variables with common finite mean* $\mu$ *and common finite variance* $\sigma^2$*. Then*

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0,1).$$

*Proof.* We can assume that $X$ has mean zero and variance one. In this case,

$$Z_n := \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} = \frac{\sum_{i=1}^n X_i}{\sqrt{n}}$$

It's easy to see (exercise) that

$$\phi_{Z_n}(t) = \phi_X(t/\sqrt{n})^n$$

Then, by the Taylor expansion of the characteristic function we have

$$\phi_X(t) = 1 - \frac{t^2}{2} + o(|t|^2)$$

for sufficiently small $t$, or equivalently

$$\phi_X(t) = \exp\left(-\frac{t^2}{2} + o(|t|^2)\right)$$

or sufficiently small $t$. Hence

$$\phi_{Z_n}(t) \to e^{-t^2/2}$$

as $n \to \infty$ for a fixed $t$. But $e^{-t^2/2}$ is the characteristic function of a standard normal. The claim now follows from the Lévy continuity theorem. □

**Exercise 2.15** (Chi-squared distribution)**.** *Let* $X_1, \ldots, X_n$ *be an i.i.d. sequence of centered Gaussian distributions with unit variance. Then the law of* $S_n := \sum_{i=1}^n X_i^2$ *is called the* $\chi^2$ *distribution with* $n$ *degrees of freedom, and we write* $S_n \sim \chi_n^2$*. Prove that* $\mathbb{E}[S_n] = n$ *and* $\mathbb{V}[S_n] = 2n$*, that it admits a density*

$$f_{S_n}(x) = \frac{x^{n/2-1}\mathrm{e}^{-x/2}}{2^{n/2}\Gamma(n/2)}, \qquad \textit{for all } x \geq 0,$$

*and its moment generating function reads*

$$\mathbb{E}\left[\mathrm{e}^{uS_n}\right] = (1-2u)^{-n/2}, \qquad \textit{for all } u < \frac{1}{2}.$$

**Exercise 2.16** (Student Distribution)**.** *If $S_n \sim \chi^2_n$ for some integer $n$ and $Z \in \mathcal{N}(0,1)$ and $S_n$ are independent then the ratio $T_n := \frac{Z}{\sqrt{S_n/n}}$ is called a Student distribution with $n$ degrees of freedom, and we write $T_n \sim \mathcal{T}_n$. Show that*

$$f_{T_n}(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}, \qquad \text{for all } x \in \mathbb{R}.$$

*In addition, show that the expectation is finite if and only if $n > 1$, in which case $\mathbb{E}[T] = 0$. Likewise, show that the variance is finite if and only if $n > 2$, in which case $\mathbb{V}[T_n] = n/(n-2)$.*

## 2.2 Concentration inequalities

We now illustrate classical concentration inequalities that appear frequently in probability theory.

**Proposition 2.17** (Markov inequality)**.** *Let $X$ be an unsigned random variable (i.e. taking non-negative values). Then for any $a > 0$*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

*Proof.* Just take expectation of the obvious inequality $a\mathbf{1}_{X \geq a} \leq X$. □

**Proposition 2.18** (Chebychev inequality)**.** *Let $X$ be a random variable with non-zero and finite variance $\sigma^2$ (and hence finite mean $\mu$). Then, for any $\lambda > 0$*

$$\mathbb{P}(|X-\mu| \geq \lambda\sigma) \leq \frac{1}{\lambda^2}.$$

*Proof.* Just apply Markov inequality to $Y = (X-\mu)^2$ with $a = (\lambda\sigma)^2$ □

**Exercise 2.19.** *Let $S_n = \sum_{i=1}^n X_i$, where $X_i$ are real-valued random variables.*

1. *Show that $\mathbb{P}(|S_n| \geq a) \leq \frac{1}{a}\sum_{i=1}^n \mathbb{E}[|X_i|]$.*
2. *Assume that the $X_i$ are pairwise independent and mean zero. Show that*

$$P(|S_n| \geq a) \leq \frac{1}{a^2}\sum_{i=1}^n \mathbb{V}[|X_i|].$$

**Theorem 2.20** (Hoeffding inequality)**.** *Let $X_1, \ldots, X_n$ be centered independent random variables with $a_i \leq X_i \leq b_i$ almost surely. Then, for any $\varepsilon > 0$ and any $z > 0$*

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq \varepsilon\right) \leq \mathrm{e}^{-z\varepsilon}\prod_{i=1}^n \exp\left(\frac{z^2(b_i-a_i)^2}{8}\right).$$

**Exercise 2.21.** *Let $X_1, \dots, X_n$ be a sequence of iid Bernoulli(p) random variables. Using Hoeffding inequality show that*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - p\right| > \varepsilon\right) \leq 2\mathrm{e}^{-2n\varepsilon^2}, \qquad \text{for any } \varepsilon > 0.$$

**Definition 2.22** (Median)**.** *We say that a (not necessarily unique) real number $\mathbb{M}[X]$ is the median of a scalar random variable $X$ if $\mathbb{P}(X > \mathbb{M}[X]), \mathbb{P}(X < \mathbb{M}[X]) < 1/2$. Equivalently, $F_X(\mathbb{M}[X]) = 1 - F_X(\mathbb{M}[X]) = 0.5$.*

**Exercise 2.23.** *Show that $\mathbb{M}[X] = \arg\min_{c\in\mathbb{R}} \mathbb{E}[|X - c|]$.*

**Lemma 2.24.** *If $X$ has finite second moment then*[2]

$$\mathbb{M}[X] = \mathbb{E}[X] + \mathcal{O}(\mathbb{V}[X]^{1/2}).$$

*Proof.* We have

$$\begin{aligned}
|\mathbb{M}[X] - \mathbb{E}[X]| &= |\mathbb{E}[\mathbb{M}[X] - X]| \\
&\leq \mathbb{E}[|\mathbb{M}[X] - X|] \\
&\leq \mathbb{E}[|\mathbb{E}[X] - X|] \\
&\leq \sqrt{\mathbb{E}[(\mathbb{E}[X] - X)^2]} \\
&= \mathbb{V}[X]^{1/2}
\end{aligned}$$

where the first inequality follows from Jensen's inequality applied to the function $x \mapsto |x|$, the second from the fact that $u : c \mapsto \mathbb{E}[|X - c|]$ is minimised when $c = \mathbb{M}[X]$, and the third from Jensen's inequality applied to the function $x \mapsto x^2$. □

Now we give a powerful concentration inequality of Talagrand.

**Theorem 2.25** (Talagrand concentration inequality)**.** *Let $K > 0$ and let $X_1, ..., X_n$ be i.i.d. scalar random variables such that $|X_i| \leq K$ almost surely. Let $F : \mathbb{R}^n \to \mathbb{R}$ be a 1-Lipschitz (i.e. $|F(x) - F(y)| \leq \|x - y\|$ for any $x, y \in \mathbb{R}^n$) and convex function. Then, for any scalar $\lambda > 0$ one has*

$$\mathbb{P}\left(|F(X) - \mathbb{M}[F(X)]| \geq \lambda K\right) \leq Ce^{-c\lambda^2}$$

*for some absolute constants $C, c > 0$.*

The next lemma is an interesting corollary of Talagrand's inequality. It asserts that the distance between a random vector $X$ and an arbitrary subspace $V$ is typically

$$\sqrt{n - \dim(V)} + \mathcal{O}(1).$$

[2] $f(x) = \mathcal{O}(g(x))$ if there exists $M > 0$ such that $|f(x)| \leq Mg(x)$.

**Lemma 2.26** (Distance between random vector and a subspace)**.** *Let $X_1, ..., X_n$ be independent scalar random variables with mean 0 and variance 1 such that $|X_i| \leq K$ almost surely for some $K > 0$. Let $V$ be a subspace of $\mathbb{R}^n$ of dimension $d$. Then, for any $\lambda > 0$ one has*

$$\mathbb{P}\left(|d(X,V) - \sqrt{n-d}| \geq \lambda K\right) \leq Ce^{-c\lambda^2}$$

*for some absolute constants $C, c > 0$.*

*Proof.* The function $x \mapsto d(x, V)$ is 1-Lipschitz and convex. By Talagrand inequality

$$\mathbb{P}\left(|d(X,V) - \mathbb{M}[d(X,V)]| \geq \lambda K\right) \leq Ce^{-c\lambda^2}.$$

To finish the argument it suffices to show that $\mathbb{M}[d(X,V)] = \sqrt{n-d} + \mathcal{O}(K)$.

We begin with a second moment calculation. Observe that

$$d(X,V)^2 = \|\pi X\|_2^2 = \sum_{1 \leq i,j \leq n} p_{ij} X_i X_j$$

where $\pi$ is the projection on $V^\perp$ and $p_{ij}$ are its entries. Taking expectation, we get

$$\mathbb{E}[d(X,V)^2] = \sum_{i=1}^n p_{ii} = \mathrm{Tr}(\pi) = \mathrm{rank}(\pi) = n - d,$$

where the last equality follows by the fact that $\pi^2 = \pi$ is a projection, so its eigenvalues must satisfy $\lambda(\pi)^2 = \lambda(\pi)$, hence $\lambda(\pi) \in \{0, 1\}$, and the trace is the sum of the eigenvalues. Now

$$d(X,V)^2 - \mathbb{E}[d(X,V)^2] = \sum_{1 \leq i,j \leq n} p_{ij}(X_i X_j - \delta_{i,j}).$$

It is easy to see that the random variables inside the summation are all pairwise uncorrelated (although not independent), therefore

$$\begin{aligned}
\mathbb{V}[d(X,V)^2] &= \mathbb{V}[d(X,V)^2 - \mathbb{E}[d(X,V)^2]] \\
&= \sum_{1 \leq i,j \leq n} \mathbb{V}[p_{ij}(X_i X_j - \delta_{i,j})] \\
&= \sum_{1 \leq i,j \leq n} p_{ij}^2 \mathbb{V}[X_i X_j]
\end{aligned}$$

Because $X_i = \mathcal{O}(K)$ almost surely, we have that $\mathbb{V}[X_i X_j] = \mathcal{O}(K^2)$. Hence,

$$\mathbb{V}[d(X,V)^2] = \mathcal{O}\left(K^2 \sum_{1 \leq i,j \leq n} p_{ij}^2\right).$$

Observe that $\sum_{1\leq i,j\leq n} p_{ij}^2 = \mathrm{Tr}(\pi^2) = \mathrm{Tr}(\pi) = n-d$, thus

$$\mathbb{V}[d(X,V)^2] = \mathcal{O}\left(K^2(n-d)\right).$$

Now, by Lemma 2.24 we know that

$$\begin{aligned}\mathbb{M}[d(X,V)^2] &= \mathbb{E}[d(X,V)^2] + \mathcal{O}(\mathbb{V}[d(X,V)^2]^{1/2}) \\ &= n-d+\mathcal{O}(K\sqrt{(n-d)})\end{aligned}$$

The result follows by taking a square root. □

# 3 The Multivariate Normal Distribution

We have already seen in the previous chapter that given a mean vector $\mu \in \mathbb{R}^p$ and a positive definite covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$, the probability density function of the multivariate normal distribution $\mathcal{N}_p(\mu, \Sigma)$ is given by

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

Let's recall some additional properties about the multivariate normal distribution.

**Proposition 3.1.** *Two vectors $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$ which are jointly multivariate normal are independent if and only if they are uncorrelated, i.e. $Cov(x, y) = 0_{p,q}$.*

*Proof.* Suppose $x \sim \mathcal{N}_p(\mu_x, \Sigma_{x,x})$ and $y \sim \mathcal{N}_p(\mu_y, \Sigma_{y,y})$ are jointly normally distributed and that they are uncorrelated. Thus we can write

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}_{p+q}(\mu, \Sigma)$$

where

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{x,x} & 0_{p,q} \\ 0_{q,p} & \Sigma_{y,y} \end{pmatrix}.$$

We compute

$$\begin{aligned} p(y|x) &= \frac{p(x,y)}{p(x)} \\ &\propto \frac{\exp\left(-\frac{1}{2}(x-\mu_x)^\top \Sigma_{x,x}(x-\mu_x) - \frac{1}{2}(y-\mu_y)^\top \Sigma_{y,y}(y-\mu_y)\right)}{\exp\left(-\frac{1}{2}(x-\mu_x)^\top \Sigma_{x,x}(x-\mu_x)\right)} \\ &\propto \exp\left(-\frac{1}{2}(y-\mu_y)^\top \Sigma_{y,y}(y-\mu_y)\right) \\ &\propto p(y). \end{aligned}$$

□

Note that uncorrelated random variables are not independent in general. For the previous proposition to hold, it is important that $x$ and $y$ are *jointly* multivariate normal. For example, suppose $x \sim N(0, 1)$ and consider an independent r.v.

$$z = \begin{cases} 1 & \text{with probability } \frac{1}{2} \\ -1 & \text{otherwise} \end{cases}$$

Now let $y = zx$. Then $y$ is also a normal random variable $y \sim N(0, 1)$. In addition,

$$\mathrm{Cov}(x, y) = \mathbb{E}(xy) = \mathbb{E}(x^2)\mathbb{E}(z) = 0$$

so that $x$ and $y$ are uncorrelated. However, $x$ and $y$ are clearly not independent!

**Exercise 3.2.** *Let $X \sim \mathcal{N}_p(\mu, \Sigma)$. Show that for any deterministic matrix $A \in \mathbb{R}^{n\times p}$ and deterministic vector $b \in \mathbb{R}^n$ one has*

$$AX + b \sim \mathcal{N}_n(A\mu + b, A\Sigma A^\top).$$

**Exercise 3.3.** *If $x \sim \mathcal{N}_p(\mu, \Sigma)$, show that $(x-\mu)^\top \Sigma^{-1}(x-\mu) \sim \chi^2_p$.*

**Solution 3.4.** *Define $y = \Sigma^{-\frac{1}{2}}(x-\mu)$ so*

$$(x-\mu)^\top \Sigma^{-1}(x-\mu) = \left(\Sigma^{-\frac{1}{2}}(x-\mu)\right)^\top \left(\Sigma^{-\frac{1}{2}}(x-\mu)\right) = y^\top y = \sum_{i=1}^p y_i^2$$

*By the previous exercise $y \sim \mathcal{N}_p(0, I_p)$, and so the components of $y$ have independent univariate normal distributions with mean 0 and variance 1. Recall that if $z \sim \mathcal{N}(0,1)$ then $z^2 \sim \chi^2_1$ and if $z_1, \ldots, z_n$ are i.i.d. $\sim \mathcal{N}(0,1)$ then $\sum_{i=1}^n z_i^2 \sim \chi^2_n$. Thus*

$$\sum_{i=1}^p y_i^2 \sim \chi^2_p.$$

**Proposition 3.5.** *If $x_1, \ldots, x_n$ are i.i.d. random samples from $\mathcal{N}_p(\mu, \Sigma)$, then the sample mean $\bar{x} \in \mathbb{R}^p$ and the sample variance matrix $S \in \mathbb{R}^{p\times p}$ defined as*

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

*are independent.*

*Proof.* Clearly, if $x_1, \ldots, x_n \sim \mathcal{N}_p(\mu, \Sigma)$ then $\bar{x} \sim \mathcal{N}_p(\mu, \frac{1}{n}\Sigma)$. Let $y_i = x_i - \bar{x} \sim \mathcal{N}_p(0, \frac{n-1}{n}\Sigma)$. Clearly $(\bar{x}, y_i)$ are also jointly Gaussian. We compute

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

Thus, by Proposition 3.1 $\bar{x}$ and $y_i$ are independent, and therefore $\bar{x}$ and $S = \frac{1}{n}\sum_{i=1}^n y_i y_i^\top$ are also independent. □

**Theorem 3.6** (Multivariate Central Limit Theorem)**.** *Let $x_1, x_2, \ldots$ be a sample of i.i.d. random vectors with mean $\mu$ and finite variance matrix $\Sigma$. Then*

$$\sqrt{n}(\bar{x} - \mu) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}_p(0, \Sigma).$$

## 3.1 The Wishart distribution

Recall that if $x_1, \ldots, x_n$ are i.i.d. samples from the standard normal $\mathcal{N}_1(0,1)$, then the random variable $S_n := \sum_{i=1}^n x_i^2$ is a $\chi_n^2$ distribution with $n$ degrees of freedom.

The Wishart distribution is a multivariate generalisation of the univariate $\chi^2$ distribution, and it plays an analogous role in multivariate statistics. In this section, we introduce the Wishart distribution and show that for multivariate normal random variables, the sample covariance matrix has a Wishart distribution.

**Definition 7.3** Let $x_1, \ldots, x_n$ be an i.i.d. random sample from $\mathcal{N}_p(0, \Sigma)$. Then

$$M = \sum_{i=1}^n x_i x_i^\top \in \mathbb{R}^{p \times p}$$

is said to have a Wishart distribution with $n$ degrees of freedom and scale matrix $\Sigma$. We write this as

$$M \sim \mathcal{W}_p(\Sigma, n)$$

and refer to $\mathcal{W}_p(I_p, n)$ as a standard Wishart distribution. Note that when $p = 1$, $\mathcal{W}_1(1, n)$ is the $\chi_n^2$ distribution.

**Exercise 3.7.** *Let $M \sim \mathcal{W}_p(\Sigma, n)$. Show that*

$$\mathbb{E}[M] = n\Sigma \quad \textit{and} \quad \mathbb{V}[M_{ij}] = n\left(\Sigma_{ij}^2 + \Sigma_{ii}\Sigma_{jj}\right)$$

Let's look at some properties of Wishart matrices.

**Proposition 3.8.** *If $M \sim \mathcal{W}_p(\Sigma, n)$ and $A$ is a fixed $q \times p$ matrix, then*

$$AMA^\top \sim \mathcal{W}_q(A\Sigma A^\top, n).$$

*Proof.* Let $M = \sum_{i=1}^n x_i x_i^\top$, where $x_i \sim \mathcal{N}_p(0, \Sigma)$. Then

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

where $y_i = Ax_i \sim \mathcal{N}_q(0, A\Sigma A^\top)$. Now we apply the definition of the Wishart distribution to $y_1, \ldots, y_n$, and, hence, $\sum_{i=1}^n y_i y_i^\top \sim \mathcal{W}_q(A\Sigma A^\top, n)$. □

**Proposition 3.9.** *If $M \sim \mathcal{W}_p(\Sigma, n)$ and $a$ is a fixed $p \times 1$ vector then*

$$a^\top M a \sim (a^\top \Sigma a)\chi_n^2.$$

*Proof.* Applying the previous proposition with $A = a^\top$, we see $a^\top M a \sim \mathcal{W}_1(a^\top \Sigma a, n)$. If we let $z_i \sim \mathcal{N}(0,1)$, and $\sigma^2 = a^\top \Sigma a$, then $\sigma z_i \sim \mathcal{N}(0, a^\top \Sigma a)$. Thus

$$a^\top M a \sim \mathcal{W}_1(a^\top \Sigma a, n) \sim \sum_{i=1}^n \sigma^2 z_i^2 = \sigma^2 \sum_{i=1}^n z_i^2 \sim (a^\top \Sigma a)\chi_n^2$$

□

**Exercise 3.10.** *If $M_1 \sim \mathcal{W}_p(\Sigma, n_1)$ and $M_2 \sim \mathcal{W}_p(\Sigma, n_2)$ are independent show that*

$$M_1 + M_2 \sim \mathcal{W}_p(\Sigma, n_1 + n_2).$$

**Solution 3.11.** *From the definition, let*

$$M_1 = \sum_{i=1}^{n_1} x_i x_i^\top$$

*and let*

$$M_2 = \sum_{i=n_1+1}^{n_1+n_2} x_i x_i^\top,$$

*where $x_i \sim \mathcal{N}_p(0, \Sigma)$. Then*

$$M_1 + M_2 = \sum_{i=1}^{n_1+n_2} x_i x_i^\top \sim \mathcal{W}_p(\Sigma, n_1 + n_2)$$

*by the definition of the Wishart distribution.*

The next result is known as Cochran's theorem. We will use Cochran's theorem to show that sample covariance matrices of multivariate normals are Wishart distributed.

**Theorem 3.12** (Cochran's Theorem)**.** *Suppose $P \in \mathbb{R}^{n\times n}$ is a projection matrix of rank $r$. Assume that $X \in \mathbb{R}^{n\times p}$ is a random matrix with i.i.d. rows that have a common $\mathcal{N}_p(0, \Sigma)$ distribution, where $\Sigma$ has full rank $p$. Then*

$$X^\top P X \sim \mathcal{W}_p(\Sigma, r) \quad \textit{and} \quad X^\top (I_n - P) X \sim \mathcal{W}_p(\Sigma, n - r).$$

*Furthermore $X^\top P X$ and $X^\top (I_n - P) X$ are independent.*

*Proof.* We first prove the result for the case $\Sigma = I_p$.

Using the spectral decomposition and noting that the eigenvalues of projection matrices must be either 0 or 1 (exercise), we can write

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

where $v_1, \ldots, v_n \in \mathbb{R}^n$ are mutually orthogonal unit vectors. Then

$$X^\top P X = X^\top \left( \sum_{j=1}^{r} v_j v_j^\top \right) X = \sum_{j=1}^{r} X^\top v_j v_j^\top X = \sum_{j=1}^{r} y_j y_j^\top \tag{3.1}$$

where $\mathbb{R}^{p\times 1} \ni y_j = X^\top v_j$. Similarly,

$$X^\top (I_n - P) X = \sum_{j=r+1}^{n} y_j y_j^\top \tag{3.2}$$

**Claim:** $y_j \sim \mathcal{N}(0_p, I_p)$.

If this claim holds, then the result follows from that (3.1) has a Wishart $\mathcal{W}_p(I_p, r)$ distribution and (3.2) has a Wishart $\mathcal{W}_p(I_p, n-r)$ distribution. Moreover, they are independent because the $y_j$ are all independent.

Let's prove this claim. Clearly $y_j$ must be multivariate Gaussians of dimension $p$, with mean vector $0_p$. Now, note that the $k^{th}$ entry of $y_j$ is

$$(y_j)_k = \sum_{i=1}^{n} X_{ik}(v_j)_i$$

and so the $(k,k')^{th}$ entry of the covariance matrix between $y_j$ and $y_{j'}$ is

$$\begin{aligned}
\mathbb{E}[(y_j)_k(y_{j'})_{k'}] &= \mathbb{E}\left[\sum_{i=1}^{n} X_{ik}(v_j)_i \sum_{i'=1}^{n} X_{i'k'}(v_{j'})_{i'}\right] \\
&= \sum_{i=1}^{n}\sum_{i'=1}^{n}(v_j)_i \mathbb{E}[X_{ik}X_{i'k'}](v_{j'})_{i'} \\
&= \begin{cases} 0 & \text{if } k \neq k' \text{ as } X_{ik} \text{ is independent of } X_{ik'} \\ \sum_{i=1}^{n}(v_j)_i(v_{j'})_i & \text{if } k = k' \text{ as } X_{ik} \text{ is independent of } X_{i'k} \text{ for } i \neq i'. \end{cases}
\end{aligned}$$

Finally, because the $v_j$'s are mutually orthogonal unit vectors

$$\sum_{i=1}^{n}(v_j)_i(v_{j'})_i = v_j^\top v_{j'} = \begin{cases} 1 & \text{if } j = j' \\ 0 & \text{otherwise.} \end{cases}$$

Thus, $\text{Cov}(y_j, y_{j'}) = 0_{p\times p}$ for $j \neq j'$ and $Var(y_j) = I_p$.

To prove the general case with covariance matrix $\Sigma$, note that if each row $x_i \sim \mathcal{N}_p(0, \Sigma)$, then we can write $x_i = \Sigma^{1/2} z_i$, where $z_i \sim \mathcal{N}_p(0, I_p)$. Thus, using the notation $X$ and $Z$ for the matrices with rows given by $x_i$ and $z_i$ respectively, we get

$$\begin{aligned}
X^\top P X &= \Sigma^{1/2} Z^\top P Z \Sigma^{1/2} \\
&\sim \Sigma^{1/2}\mathcal{W}_p(I_p, r)\Sigma^{1/2} \quad \text{by the result above} \\
&\sim \mathcal{W}_p(\Sigma, r)
\end{aligned}$$

where the final line follows by Proposition 3.8. □

The next result is an important consequence of Cochran's Theorem. It states that sample covariance matrices of multivariate normals have a Wishart distribution. This result will be useful in the sequel, as it will allow to compute the sampling distribution of a test statistic that we will then use in hypothesis testing.

**Proposition 3.13.** *If $x_1, \ldots, x_n$ is an i.i.d. sample from $\mathcal{N}_p(\mu, \Sigma)$, then*

$$nS = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top \sim \mathcal{W}_p(\Sigma, n-1).$$

*Proof.* Consider the $n \times n$ matrix $P = I_n - n^{-1}1_n 1_n^\top$, where $1_n$ is the $n \times 1$ vector of ones. It is easy to see that $P$ is a projection matrix, and that $I_n - P = n^{-1}1_n 1_n^\top$ has rank 1. Thus, $P$ must have rank $n-1$. Hence, by Cochran's Theorem we have

$$nS = X^\top PX \sim \mathcal{W}_p(\Sigma, n-1).$$

□

## 3.2 Hotelling's $T^2$ distribution

Recall that if $S_n \sim \chi^2_n$ for some integer $n$ and $z \in \mathcal{N}_1(0,1)$ and $S_n$ are independent, then the ratio $\frac{z}{\sqrt{S_n/n}}$ is called a Student t-distribution with $n$ degrees of freedom.

Hotelling's $T^2$ distribution is the multivariate analogue of Student's t-distribution. It plays an important role in multivariate hypothesis testing and confidence region construction, just as the Student t-distribution does in the univariate setting.

**Definition 3.14.** *Suppose $x \sim \mathcal{N}_p(0, I_p)$ and $M \sim \mathcal{W}_p(I_p, n)$ are independent, then*

$$\tau^2 = nx^\top M^{-1} x$$

*is said to have Hotelling's $T^2$ distribution with parameters $p$ and $n$. We write this as*

$$\tau^2 \sim T^2(p, n).$$

**Proposition 3.15.** *Suppose $x \sim \mathcal{N}_p(\mu, \Sigma)$ and $M \sim \mathcal{W}_p(\Sigma, n)$ are independent and $\Sigma$ has full rank $p$. Then*

$$n(x-\mu)^\top M^{-1}(x-\mu) \sim T^2(p, n).$$

*Proof.* Define $y = \Sigma^{-1/2}(x-\mu)$. Then, $y \sim N_p(0, I_p)$. Furthermore, let

$$Z = \Sigma^{-1/2} M \Sigma^{-1/2} \implies Z \sim \mathcal{W}_p(I_p, n)$$

by applying Proposition 3.8 with $A = \Sigma^{-1/2}$. From the definition,

$$ny^\top Z^{-1} y \sim T^2(p, n)$$

and

$$\begin{aligned} ny^\top Z^{-1} y &= n(x-\mu)^\top \Sigma^{-1/2} \Sigma^{1/2} M^{-1} \Sigma^{1/2} \Sigma^{-1/2} (x-\mu) \\ &= n(x-\mu)^\top M^{-1}(x-\mu). \end{aligned}$$

□

This result gives rise to an important corollary used in hypothesis testing when $\Sigma$ is unknown.

**Corollary 3.16.** *If $\bar{x}$ and $S$ are the sample mean and covariance matrix of a sample of size $n$ from $\mathcal{N}_p(\mu, \Sigma)$ then*

$$(n-1)(\bar{x}-\mu)^\top S^{-1}(\bar{x}-\mu) \sim T^2(p, n-1).$$

*Proof.* We have already seen earlier that $\bar{x} \sim \mathcal{N}_p\left(\mu, \frac{1}{n}\Sigma\right)$. From Proposition 3.13, we know $nS \sim \mathcal{W}_p(\Sigma, n-1)$, and from Proposition 3.5 we know $\bar{x}$ and $S$ are independent. Applying Proposition 3.15 with $x = \sqrt{n}\bar{x}$ and $M = nS$ we obtain

$$\begin{aligned} T^2(p, n-1) &\sim (n-1)(\sqrt{n}\bar{x} - \sqrt{n}\mu)^\top (nS)^{-1}(\sqrt{n}\bar{x} - \sqrt{n}\mu) \\ &= (n-1)(\bar{x}-\mu)^\top S^{-1}(\bar{x}-\mu). \end{aligned}$$

□

# 4 Statistical estimation

Consider a sample $X_1, \ldots, X_n$ of i.i.d. observations from some random variable $X$. The *empirical CDF* of the sample is defined as

$$\widehat{F}_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}, \qquad \text{for all } x \in \mathbb{R}.$$

**Remark 4.1.** *Note that the function $\widehat{F}_n$ is monotonically increasing, piecewise constant, right-continuous, with jump sizes equal to $1/n$ and such that*

$$\lim_{x \downarrow -\infty} \widehat{F}_n(x) = 0 \qquad \textit{and} \qquad \lim_{x \to \infty} \widehat{F}_n(x) = 1.$$

For any $x \in \mathbb{R}$, $\mathbb{1}_{\{X_i \leq x\}}$ is a Bernoulli random variable with mean parameter $p = F(x)$. By the strong LLN we have that the empirical CDF convergence almost surely to the CDF. A stronger (uniform convergence) version of this observation is the following.

**Theorem 4.2** (Glivenko-Cantelli Theorem)**.** *Let $X_1, \ldots, X_n \sim X$ be i.i.d. observations of some random variable $X$ with continuous CDF $F$, then*

$$\lim_{n \to \infty} \left\| \widehat{F}_n - F \right\|_\infty = 0 \quad \textit{almost surely.}$$

*Proof.* Because $F$ is monotonically increasing, for any integer $k$, we can find a sequence $-\infty < x_0 < x_1 < \cdots < x_{k-1} < x_k < +\infty$ such that $F(x_i) - F(x_{i-1}) = 1/k$ for all $i = 1, ..., k$. Now, for any $x \in [x_{i-1}, x_i]$, the monotonicity of both $\widehat{F}_n$ and $F$ implies that

$$\begin{aligned} \widehat{F}_n(x_{i-1}) - F(x_{i-1}) - \frac{1}{k} &= \widehat{F}_n(x_{i-1}) - F(x_i) \\ &\leq \widehat{F}_n(x) - F(x) \\ &\leq \widehat{F}_n(x_i) - F(x_{i-1}) \\ &= \widehat{F}_n(x_i) - F(x_i) + \frac{1}{k}, \end{aligned}$$

Hence, for any $x \in \mathbb{R}$

$$\left| \widehat{F}_n(x) - F(x) \right| \leq \max_{i=1,\ldots,k} \left\{ \left| \widehat{F}_n(x_i) - F(x_i) \right| \right\} + \frac{1}{k}.$$

So, in particular

$$\sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F(x) \right| \leq \max_{i=1,\ldots,k} \left\{ \left| \widehat{F}_n(x_i) - F(x_i) \right| \right\} + \frac{1}{k}.$$

Since, by the strong LLN, $\max_{i=1,\ldots,k} \left\{ \left| \widehat{F}_n(x_i) - F(x_i) \right| \right\} \to 0$ almost surely, for any $\epsilon > 0$ and for any $k$ such that $1/k < \epsilon$, there exists $N$ such that for any $n \geq N$ we have

$\max_{i=1,\dots,k}\left\{\left|\widehat{F}_n(x_i)-F(x_i)\right|\right\}\le\epsilon-1/k$ a.s. Combined with the above inequality we conclude that $\left\|\widehat{F}_n-F\right\|_\infty\le\epsilon$ a.s.. □

Recall that the sample mean and sample variance are defined as follows

$$\overline{X}:=\frac{1}{n}\sum_{i=1}^n X_i,\quad S^2:=\frac{1}{n}\sum_{i=1}^n\left(X_i-\overline{X}\right)^2=\frac{1}{n}\sum_{i=1}^n X_i^2-\overline{X}^2. \tag{4.1}$$

Clearly, $\mathbb{E}[\overline{X}]=\mathbb{E}[X]$; we say that $\overline{X}$ is an *unbiased* estimator of $\mathbb{E}[X]$.

On the other hand

$$\begin{aligned}
\mathbb{E}[S^2]&=\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left(X_i-\frac{1}{n}\sum_{j=1}^n X_j\right)^2\right]\\
&=\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[X_i^2-\frac{2}{n}X_i\sum_{j=1}^n X_j+\frac{1}{n^2}\sum_{j,k=1}^n X_jX_j\right]\\
&=\frac{1}{n}\sum_{i=1}^n\left(\frac{n-2}{n}\mathbb{E}[X_i^2]-\frac{2}{n}\sum_{j\neq i}\mathbb{E}[X_iX_j]+\frac{1}{n^2}\sum_{j\neq k}^n\mathbb{E}[X_jX_k]+\frac{1}{n^2}\sum_{j=1}^n\mathbb{E}[X_j^2]\right)\\
&=\frac{1}{n}\sum_{i=1}^n\left[\frac{n-2}{n}(\sigma^2+\mu^2)-\frac{2}{n}(n-1)\mu^2+\frac{1}{n^2}n(n-1)\mu^2+\frac{1}{n}(\sigma^2+\mu^2)\right]\\
&=\frac{n-1}{n}\sigma^2.
\end{aligned}$$

where we used the notation $\mu=\mathbb{E}[X]$ and $\sigma^2=\mathbb{V}[X]$. This means that $S^2$ is a biased estimator for the variance $\sigma^2$, with bias factor $\frac{n-1}{n}$. Correcting the bias would give us the unbiased estimator $\tilde{S}^2=\frac{n}{n-1}S^2$.

We shall from now on denote by $\widehat{\theta}_n$ an estimator of a real parameter $\theta$, and, remembering that it is a random variable, call it unbiased if $\mathbb{E}_\theta(\widehat{\theta}_n)=\theta$.

**Definition 4.3.** *An estimator $\widehat{\theta}_n$ is said to be (respectively strongly) consistent if it converges to $\theta$ in probability (respectively almost surely) for all $\theta\in\Theta$.*

Note that if $\widehat{\theta}_n$ is a consistent estimator of $\theta$, then so is $\alpha_n\widehat{\theta}_n$ for any sequence $(\alpha_n)$ converging to 1, so that the notion of consistent estimator, though fundamental, is in fact rather weak.

**Definition 4.4.** *The quadratic error of the estimator $\widehat{\theta}_n$ of $\theta$ is defined as*

$$R_n(\widehat{\theta}_n,\theta):=\mathbb{E}_\theta\left[\left(\widehat{\theta}_n-\theta\right)^2\right].$$

**Proposition 4.5.** *If $R_n(\widehat{\theta}_n, \theta)$ converges (pointwise in $\theta \in \Theta$) to zero as $n$ tends to infinity, then $\widehat{\theta}_n$ is a consistent estimator of $\theta$.*

*Proof.* Convergence of $R_n(\widehat{\theta}_n, \theta)$ is the same as $L^2$ convergence, and hence convergence in probability follows directly from Theorem 2.11. □

**Remark 4.6.** *Alternatively, Markov's inequality (Proposition 2.17) implies, for any $\varepsilon > 0$,*

$$\mathbb{P}\left(\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2}\mathbb{E}\left[\left(\widehat{\theta}_n - \theta\right)^2\right] = \frac{R_n(\widehat{\theta}_n, \theta)}{\varepsilon^2},$$

*and the proposition follows by taking limits.*

**Exercise 4.7.** *Let $(X_1, ..., X_n)$ be an i.i.d. sample from a Bernoulli random variable with parameter $\theta \in [0, 1]$, and denote $\widehat{\theta}_n := \frac{1}{n}\sum_{i=1}^n X_i$. Show that $\widehat{\theta}_n$ is a consistent estimator of $\theta$.*

**Solution 4.8.** *Recall that a Bernoulli random variable $X$ with parameter $\theta \in [0, 1]$ takes value 1 with probability $\theta$ and zero with probability $1 - \theta$, and $\mathbb{E}[X] = \theta$ and $\mathbb{V}[X] = \theta(1 - \theta)$, so that $\mathbb{E}[X^2] = \mathbb{V}[X] + \mathbb{E}[X]^2 = \theta$. Therefore,*

$$\begin{aligned}
\mathbb{E}\left[\left(\widehat{\theta}_n - \theta\right)^2\right] = \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n X_i - \theta\right)^2\right] &= \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^n X_i\right)^2\right] - \frac{2\theta}{n}\mathbb{E}\left[\sum_{i=1}^n X_i\right] + \theta^2 \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^n X_i^2 + \sum_{i \neq j} X_i X_j\right] - 2\theta^2 + \theta^2 \\
&= \frac{1}{n^2}\left(\sum_{i=1}^n \mathbb{E}\left[X_i^2\right] + \sum_{i \neq j}\mathbb{E}[X_i X_j]\right) - \theta^2 \\
&= \frac{1}{n^2}\left(n\theta + n(n-1)\theta^2\right) - \theta^2,
\end{aligned}$$

*which clearly converges to zero as $n$ tends to infinity.*

The quadratic risk of an estimator can be decomposed as follows:

$$\begin{aligned}
R_n(\widehat{\theta}_n, \theta) &:= \mathbb{E}_\theta\left[\left(\widehat{\theta}_n - \theta\right)^2\right] \\
&= \left(\mathbb{E}_\theta[\widehat{\theta}_n] - \theta\right)^2 + \mathbb{E}_\theta\left[\left(\widehat{\theta}_n - \mathbb{E}_\theta[\widehat{\theta}_n]\right)^2\right] \\
&=: \beta_n^2(\widehat{\theta}_n, \theta) + \sigma_n^2(\widehat{\theta}_n, \theta), \qquad (4.2)
\end{aligned}$$

where $\beta_n$ is called the bias and $\sigma_n^2$ the variance of the estimator $\widehat{\theta}_n$, so that $\widehat{\theta}_n$ is unbiased if $\beta_n(\widehat{\theta}_n, \theta) = 0$ for all $\theta \in \Theta$.

## 4.1 The method of moments

The moment of order $r$ (if it exists) of a statistical model with pdf $f_\theta$ is defined as

$$\mu_r(\theta) := \mathbb{E}_\theta[X^r] = \int_{\mathbb{R}} x^r f_\theta(\mathrm{d}x), \tag{4.3}$$

Given an i.i.d. sample $X_1, ..., X_n$ from $f_\theta$, the *method of moments estimator* $\widehat{\theta}_n^{\mathrm{MM}}$ for $\theta$ is the solution to the system

$$\mu_r\left(\widehat{\theta}_n^{\mathrm{MM}}\right) = m_r, \quad \text{where } \; m_r := \frac{1}{n}\sum_{i=1}^n X_i^r.$$

For example, given $X_1, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$ we have

$$\mu_1\left(\widehat{\theta}_n^{\mathrm{MM}}\right) = \mathbb{E}[X] = \mu$$
$$\mu_2\left(\widehat{\theta}_n^{\mathrm{MM}}\right) = \mathbb{E}[X^2] = \sigma^2 + \mu^2$$

Hence, the method of moments estimator satisfies

$$\hat{\mu} = \overline{X} \;\text{ and }\; \hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^n X_i^2$$

which yields $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2$.

**Exercise 4.9.** *Using the first two moments, compute the method of moment estimator for the parameter of the exponential distribution.*

Moments may not exist for some distribution. Instead of moments $\mu_r$, one can consider more general functions $(\phi_r)_r$, and write

$$\mu_r(\theta) := \mathbb{E}_\theta[\phi_r(X)],$$

and replace their estimators by

$$m_r := \frac{1}{n}\sum_{i=1}^n \phi_r(X_i).$$

We denote by $\widehat{\theta}_n^{\mathrm{GM}}$ the corresponding generalised method of moment estimator.

A classical example is the one of the Cauchy distribution. The following Lemma will be useful in the sequel.

**Lemma 4.10.** *Given an i.i.d. sample $X_1, \ldots, X_n$ with finite first two moments and non-zero variance, and a continuously differentiable function $g$ on $\mathbb{R}$ such that $g'(\mathbb{E}[X_1])$ is non-zero, then we have that*

$$\sqrt{n}\frac{g\left(\overline{X}\right) - g\left(\mathbb{E}[X_1]\right)}{g'(\mathbb{E}[X_1])\sqrt{\mathbb{V}[X_1]}} \quad \textit{converges in distribution to } \mathcal{N}(0,1).$$

*Proof.* Let $\mu := \mathbb{E}[X_1]$, $\sigma = \sqrt{\mathbb{V}[X_1]}$, and introduce the function

$$h(x) := \begin{cases} \dfrac{g(x) - g(\mu)}{x - \mu}, & \text{if } x \neq \mu, \\ g'(\mu), & \text{if } x = \mu. \end{cases}$$

Since $h$ is continuous and $\overline{X}$ converges almost surely to $\mu$ by the strong LLN, by the continuous mapping theorem, $h(\overline{X})$ converges almost surely to $h(\mu) = g'(\mu)$. Furthermore

$$\sqrt{n}\frac{g\left(\overline{X}\right) - g\left(\mathbb{E}[X_1]\right)}{\sigma} = \frac{\sqrt{n}}{\sigma} h\left(\overline{X}\right)\left(\overline{X} - \mu\right) = h\left(\overline{X}\right)\eta_n,$$

where $\eta_n := \frac{\sqrt{n}}{\sigma}\left(\overline{X} - \mu\right)$ converges to $\mathcal{N}(0,1)$ by the CLT, and the lemma follows. □

**Exercise 4.11.** *Let $X$ be a random variable with Cauchy distribution with location $x_0 \in \mathbb{R}$ and scale $\gamma > 0$. It admits as density*

$$f_X(x) = \frac{1}{\pi\gamma}\left\{1 + \left(\frac{x - x_0}{\gamma}\right)^2\right\}^{-1}, \qquad \text{for all } x \in \mathbb{R}.$$

*and characteristic function*

$$\phi_X(u) = \exp\left\{\mathtt{i}x_0 u - \gamma|u|\right\}, \qquad \text{for any } u \in \mathbb{R}.$$

- *Show that no moment exists.*
- *Let $x_0 = \theta$ and $\gamma = 1$. With the function $\phi_1(x) := \mathbb{1}_{\{x>0\}} - \mathbb{1}_{\{x\leq 0\}}$, show that the generalised method of moments estimator $\widehat{\theta}_n^{\mathrm{GM}}$, given a sample $X_1, ..., X_n$, is of the form*
$$\widehat{\theta}_n^{\mathrm{GM}} = \tan\left(\frac{\pi}{2n}\sum_{i=1}^{n}\phi_1(X_i)\right).$$
- *Show that $\widehat{\theta}_n^{\mathrm{GM}}$ is strongly consistent and asymptotically Gaussian, i.e. that $\sqrt{n}\left(\widehat{\theta}_n^{\mathrm{GM}} - \theta\right)$ converges in distribution to a centered Gaussian random variable with variance which can be determined explicitly.*

**Solution 4.12.** *Recall that for any integer $n$, $\mathbb{E}[X^n] = \partial_u^n \phi_X(u)|_{u=0}$. However $\phi_X$ is not differentiable at the origin, and hence the moments do not exist.*

*[Another way to argue is to note that the map $x \mapsto x^n f_X(x)$ does not decay fast enough at infinity to make its integral converge.]*

*For any $z \in \mathbb{R}$, its cumulative distribution function then reads*

$$\begin{aligned} F_X(z) &= \int_{-\infty}^{z} f_X(x)\mathrm{d}x \\ &= \frac{1}{\pi\gamma}\int_{-\infty}^{z}\left\{1+\left(\frac{x-x_0}{\gamma}\right)^2\right\}^{-1}\mathrm{d}x \\ &= \frac{1}{\pi}\int_{-\infty}^{(z-x_0)/\gamma}\frac{\mathrm{d}u}{1+u^2} \\ &= \frac{1}{2}+\frac{1}{\pi}\arctan\left(\frac{z-x_0}{\gamma}\right). \end{aligned}$$

*Setting $x_0 = \theta$ and $\gamma = 1$*

$$F_X(0) = \frac{1}{2} - \frac{1}{\pi}\arctan(\theta) \tag{4.4}$$

*where we used the fact that arctan is an odd function. With (4.4), we can write*

$$\begin{aligned} \mu_1(\theta) &:= \mathbb{E}[\phi_1(X)] \\ &= \mathbb{E}\left[\mathbb{1}_{\{X>0\}} - \mathbb{1}_{\{X\leq 0\}}\right] \\ &= \mathbb{P}[X>0] - \mathbb{P}[X\leq 0] \\ &= 1 - 2F_X(0) \\ &= \frac{2}{\pi}\arctan(\theta). \end{aligned}$$

*The equation $m_1 = \mu_1(\theta)$ can be solved explicitly as*

$$\widehat{\theta}_n^{\mathrm{GM}} = \tan\left(\frac{\pi}{2}m_1\right) = \tan\left(\frac{\pi}{2n}\sum_{i=1}^{n}\phi_1(X_i)\right). \tag{4.5}$$

*Denote $Y_i := \phi_1(X_i)$, so that $(Y_i)_{1\leq i\leq n}$ is an i.i.d. sequence of Bernoulli random variables with common finite mean $\mathbb{E}[Y_1] = 1 - 2F_X(0)$. Writing $\overline{Y} := \frac{1}{n}\sum_{i=1}^{n} Y_i$ for the empirical mean, we can rewrite (4.5) as*

$$\widehat{\theta}_n^{\mathrm{GM}} = \tan\left(\frac{\pi}{2}\overline{Y}\right) =: g(\overline{Y}).$$

*By the strong LLN, $\overline{Y}$ converges almost surely to*

$$\mu = \mathbb{E}[Y_1] = 1 - 2F_X(0) = \frac{2}{\pi}\arctan(\theta)$$

*Hence, because $g$ is continuous, by the continuous mapping theorem $g(\overline{Y})$ converges almost surely to $g(\mu) = \theta$.*

*Because $g$ is $C^1$ and $g'(\mu) \neq 0$, Lemma 4.10 yields that*

$$\sqrt{n}\frac{g\left(\overline{Y}\right) - g\left(\mathbb{E}[Y_1]\right)}{g'\left(\mathbb{E}[Y_1]\right)\sqrt{\mathbb{V}[Y_1]}} \quad \textit{converges in distribution to } \mathcal{N}(0,1).$$

## 4.2 Maximum likelihood method

We now move on to one of the most important and widely used estimation method, namely Maximum Likelihood Estimation. In this section we will assume that $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ is an i.i.d. sample from an underlying distribution. We will assume existence of densities $f_\theta$ for all the considered statistical models $\mathcal{F}$.

The density $\theta \mapsto f_\theta(X)$ seen as a map from $\Theta$ to $\mathbb{R}$ is usually referred to as the *likelihood function*. Its logarithm is called the *log-likelihood function*. Because $X_1, ..., X_n$ are i.i.d. we have, by a slight abuse of notation,

$$f_\theta(X) = \prod_{i=1}^{n} f_\theta(X_i).$$

**Definition 4.13.** *The maximum (log)likelihood estimator is defined as*

$$\widehat{\theta}_n^{\mathrm{ML}} := \arg\max_{\theta\in\Theta} f_\theta(X) = \arg\max_{\theta\in\Theta} \log f_\theta(X) \implies \partial_\theta \log f_\theta(X)|_{\theta=\widehat{\theta}_n^{\mathrm{ML}}} = 0.$$

**Example 4.14.** *For the Gaussian statistical model $\mathcal{F} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0\}$, we can write, with $\theta = (\mu, \sigma)$,*

$$\begin{aligned} f_\theta(X) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(X_i-\mu)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2\right) \end{aligned}$$

*Upon taking logarithm we get*

$$-\log f_\theta(X) = n\log(\sqrt{2\pi}) + n\log(\sigma) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2.$$

*The necessary condition $\partial_\theta \log f_\theta(X) = 0$ is therefore equivalent to*

$$\begin{cases} \partial_\mu \log f_\theta(X) &= 0, \\ \partial_\sigma \log f_\theta(X) &= 0 \end{cases} \iff \begin{cases} \displaystyle\sum_{i=1}^{n}(X_i-\mu) &= 0, \\ \displaystyle\sum_{i=1}^{n}(X_i-\mu)^2 &= \sigma^2 n, \end{cases}$$

*so that*

$$\mu = \overline{X} \qquad \text{and} \qquad \sigma = \left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right)^{1/2}.$$

*Direct computations yield that the method of moments estimator are identical to the maximum likelihood estimators above.*

**Example 4.15.** *Consider the Uniform distribution on the closed interval $[0,\theta]$, with density given by $f_\theta(x) = \theta^{-1}\mathbb{1}_{[0,\theta]}(x)$. The likelihood function is then given by*

$$f_\theta(X) := \prod_{i=1}^{n} f_\theta(X_i) = \frac{1}{\theta^n}\prod_{i=1}^{n}\mathbb{1}_{[0,\theta]}(X_i) = \begin{cases} \frac{1}{\theta^n} & \text{if all the } X_i \text{ are in } [0,\theta] \\ 0 & \text{otherwise} \end{cases}$$

*The likelihood function is maximised by making $[0,\theta]$ as small as possible but at the same time ensuring that all the $X_i$ are in $[0,\theta]$. Therefore, the maximum of the function is then attained at the point $\widehat{\theta}_n^{\mathrm{ML}} = \max_{i=1,\dots,n} X_i$.*

**Example 4.16** (Historical volatility under Black-Scholes)**.** *We now present a simple application of maximum likelihood to financial time series. We consider a stock price process, whose dynamics are given by a geometric Brownian motion*

$$\frac{\mathrm{d}S_t}{S_t} = \mu\mathrm{d}t + \sigma\mathrm{d}W_t, \tag{4.6}$$

*for some standard Brownian motion $(W_t)_{t\geq 0}$, over some time period $[0,T]$. We assume that the drift $\mu \in \mathbb{R}$ and the diffusion $\sigma > 0$ are constant. Our goal is to estimate $\mu$ and $\sigma$ given a sample time series of $S$. A direct application of Itô's lemma and the Gaussianity of Brownian increments yield, for any $t \geq 0$ and $\delta > 0$,*

$$X_t^\delta := \log\left(\frac{S_{t+\delta}}{S_t}\right) = \left(\mu - \frac{\sigma^2}{2}\right)\delta + \sigma(W_{t+\delta} - W_t) \sim \mathcal{N}\left(\left(\mu - \frac{\sigma^2}{2}\right)\delta, \sigma^2\delta\right).$$

*Consider the sequence $\mathbf{X}_t^\delta := \left(X_t^\delta, X_{t+\delta}^\delta, \cdots, X_{t+n\delta}^\delta\right)$. Since Brownian increments are independent, $\mathbf{X}_t^\delta$ constitutes an i.i.d. sample with likelihood function*

$$f_{\mu,\sigma}(\mathbf{X}_t^\delta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi\delta}} \exp\left\{-\frac{\left(X_{t+i\delta}^\delta - \left(\mu - \frac{1}{2}\sigma^2\right)\delta\right)^2}{2\sigma^2\delta}\right\}.$$

*The log-likelihood, similar to previously, then reads*

$$\log f_{\mathbf{X}_t^\delta}(\mu,\sigma) = n\log\left(\frac{1}{\sigma\sqrt{2\pi\delta}}\right) - \frac{1}{2\sigma^2\delta}\sum_{i=1}^{n}\left\{\left(X_{t+i\delta}^\delta - \left(\mu - \frac{1}{2}\sigma^2\right)\delta\right)^2\right\}.$$

*We can again solve the maximum likelihood problem to obtain*

$$\mu^{\mathrm{ML}} = \frac{1}{\delta}\left(\overline{X} + \frac{1}{2n}\sum_{i=1}^{n}\left(X_{t+i\delta}^\delta - \overline{X}\right)^2\right) \quad \text{and} \quad \sigma^{\mathrm{ML}} = \sqrt{\frac{1}{n\delta}\sum_{i=1}^{n}\left(X_{t+i\delta}^\delta - \overline{X}\right)^2},$$

*where $\overline{X} := \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_{t+i\delta}^\delta$.*

## 4.3 Maximum likelihood asymptotics

Denote by $\theta^*$ the true value of the parameter to be estimated. For any $\theta \in \Theta$, let

$$\ell_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \log f_\theta(X_i)$$

As $X_1, ..., X_n$ are all i.i.d samples from $f_{\theta^*}(X_1)$, by the weak LLN, the following convergence in probability holds as $n \to +\infty$

$$\ell_n(\theta) \xrightarrow{\mathcal{P}} \ell(\theta) := \int \log f_\theta(x) f_{\theta^*}(x) \mathrm{d}x$$

Note that $\hat{\theta}_n^{ML}$ is also the maximiser of $\ell_n$.

**Lemma 4.17.** *The inequality $\ell(\theta) \leq \ell(\theta^*)$ holds for all $\theta \in \Theta$. Furthermore, the inequality is strict $\ell(\theta) < \ell(\theta^*)$ unless*[3] $\mathbb{P}_\theta = \mathbb{P}_{\theta^*}$.

*Proof.* The difference

$$\begin{aligned}
\ell(\theta) - \ell(\theta^*) &= \int \log f_\theta(x) f_{\theta^*}(x) dx - \int \log f_{\theta^*}(x) f_{\theta^*}(x) dx \\
&= \int f_{\theta^*}(x) \log\left(\frac{f_\theta(x)}{f_{\theta^*}(x)}\right) \mathrm{d}x \\
&= \int f_{\theta^*}(x) \left\{\log\left(\frac{f_\theta(x)}{f_{\theta^*}(x)}\right) - \left(\frac{f_\theta(x)}{f_{\theta^*}(x)} - 1\right)\right\} \mathrm{d}x \\
&= \int f_{\theta^*}(x) \left\{\log\left(1 + \left[\frac{f_\theta(x)}{f_{\theta^*}(x)} - 1\right]\right) - \left(\frac{f_\theta(x)}{f_{\theta^*}(x)} - 1\right)\right\} dx.
\end{aligned}$$

where the penultimate equality follows from

$$\int \left(\frac{f_\theta(x)}{f_{\theta^*}(x)} - 1\right) f_{\theta^*}(x) \mathrm{d}x = \int f_\theta(x) \mathrm{d}x - \int f_{\theta^*}(x) \mathrm{d}x = 0.$$

Now observe that the inequality $\log(1+z) - z \leq 0$ holds for all $z \geq -1$, with equality if and only if $z = 0$. Therefore, the terms inside the curly bracket in the expression above is negative. Hence, $\ell(\theta) - \ell(\theta^*) \leq 0$, proving the first claim. The second claim easily follows. □

This Lemma can be used to prove that the MLE is a consistent estimator. A precise statement and proof would require to enumerate several technical conditions, which are beyond the scope of this course. Therefore, we limit ourselves to provide only a sketch of the proof.

[3] Under the identifiability assumption $\mathbb{P}_\theta = \mathbb{P}_{\theta^*} \iff \theta = \theta^*$.

**Consistency of MLE** Under some regularity conditions on the family of distributions, MLE $\hat{\theta}_n^{ML}$ is consistent, i.e. $\hat{\theta}_n^{ML} \xrightarrow{\mathcal{P}} \theta^*$ as $n \to +\infty$.

The idea of the proof follows from the following facts:

1. $\hat{\theta}_n^{ML}$ is the maximizer of $\ell_n(\theta)$ (by definition).
2. $\theta^*$ is the maximizer of $\ell(\theta)$ (by Lemma above).
3. For every $\theta$ we have $\ell_n(\theta) \to \ell(\theta)$ by LLN.

Therefore, since $\ell_n$ converges $\ell$, the maximum of $\ell_n$ also converges to the one of $\ell$.

**Fisher information** The Fisher information quantifies how much information a random variable carries about its generative parameters.

Let $X = (X_1, \ldots, X_n)$ be an i.i.d. sample with joint density $f_\theta(X)$. The gradient of the log likelihood function w.r.t. to $\theta$, also known as the score, measures how sensitive the log likelihood is to changes in parameter values. The Fisher information is the variance of such score,

$$I_n(\theta) = \mathbb{E}_X \left[ \left( \frac{\partial}{\partial \theta} \log f_\theta(X) \right)^2 \right] = \mathbb{V}_X \left[ \frac{\partial}{\partial \theta} \log f_\theta(X) \right].$$

The second inequality holds because, under certain regularity conditions that essentially boil down to assuming we can interchange integration with respect to $X$ and differentiation with respect to $\theta$ by dominated convergence, we have

$$\mathbb{E} \left[ \frac{\partial}{\partial \theta} \log f_\theta(X) \right] = 0,$$

In effect, under these assumptions, note that

$$\mathbb{E} \left[ \frac{\partial}{\partial \theta} \log f_\theta(X) \right] = \int \left[ \frac{\frac{\partial}{\partial \theta} f_\theta(x)}{f_\theta(x)} \right] f_\theta(x) dx = \int \frac{\partial}{\partial \theta} f_\theta(x) dx = \frac{\partial}{\partial \theta} \int f_\theta(x) dx = 0.$$

**Remark 4.18.** *If $f_\theta$ is twice differentiable and if we can interchange integration and differentiation, then*

$$I_n(\theta) = -\mathbb{E} \left[ \frac{\partial^2}{\partial \theta^2} \log f_\theta(X) \right].$$

*To see this, first note that*

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}\log f_\theta(X) &= \frac{\partial}{\partial\theta}\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)\\ &= \frac{\partial}{\partial\theta}\left(\frac{\frac{\partial}{\partial\theta}f_\theta(X)}{f_\theta(X)}\right)\\ &= \frac{f_\theta(X)\frac{\partial^2}{\partial\theta^2}f_\theta(X) - \frac{\partial}{\partial\theta}f_\theta(X)\frac{\partial}{\partial\theta}f_\theta(X)}{f_\theta(X)^2}\\ &= \frac{\frac{\partial^2}{\partial\theta^2}f_\theta(X)}{f_\theta(X)} - \left(\frac{\frac{\partial}{\partial\theta}f_\theta(X)}{f_\theta(X)}\right)^2.\end{aligned}$$

*Now, notice that*

$$\begin{aligned}\mathbb{E}\left[\frac{\frac{\partial^2}{\partial\theta^2}f_\theta(X)}{f_\theta(X)}\right] &= \int\frac{\frac{\partial^2}{\partial\theta^2}f_\theta(x)}{f_\theta(x)}f_\theta(x)dx\\ &= \int\frac{\partial^2}{\partial\theta^2}f_\theta(x)dx\\ &= \frac{\partial^2}{\partial\theta^2}\int f_\theta(x)dx\\ &= \frac{\partial^2}{\partial\theta^2}1 = 0.\end{aligned}$$

**Cramér–Rao lower bound** The Cramér–Rao lower bound (CRLB) provides a lower bound on the variance of an estimator $\hat{\theta}$. Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with joint density $f_\theta(X)$ where $\theta \in \Theta \subseteq \mathbb{R}$. Let $\hat{\theta}$ be an estimator of $\theta$. Assume the Fisher information is well-defined and that the operations of integration with respect to $X$ and differentiation with respect to $\theta$ can be interchanged. Then

$$\mathbb{V}[\hat{\theta}] \geq \frac{\left(\frac{\partial}{\partial\theta}\mathbb{E}[\hat{\theta}]\right)^2}{I_n(\theta)} := \text{CRLB}(\theta).$$

To see this inequality, by Cauchy-Schwarz we have that

$$V[\hat{\theta}] \geq \frac{\left(\text{Cov}\left[\hat{\theta}, \frac{\partial}{\partial\theta}\log f_\theta(X)\right]\right)^2}{\mathbb{V}\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]}.$$

We have already shown that the denominator is equal to the Fisher information.

The nominator

$$\begin{aligned}\mathrm{Cov}\left[\hat{\theta}, \frac{\partial}{\partial\theta}\log f_\theta(X)\right] &= \mathbb{E}\left[\hat{\theta}\frac{\partial}{\partial\theta}\log f_\theta(X)\right] - \mathbb{E}[\hat{\theta}]\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right] \\ &= \int \hat{\theta}\frac{\partial}{\partial\theta}\log f_\theta(x) f_\theta(x)dx \\ &= \int \hat{\theta}\frac{\partial}{\partial\theta} f_\theta(x)dx \\ &= \frac{\partial}{\partial\theta}\int \hat{\theta} f_\theta(x)dx \\ &= \frac{\partial}{\partial\theta}\mathbb{E}[\hat{\theta}].\end{aligned}$$

**Asymptotic normality of MLE** Next we show that, under suitable conditions, the MLE is asymptotically Gaussian.

**Theorem 4.19** (Asymptotic normality of MLE)**.** *Assuming that $X = (X_1, ..., X_n)$ are i.i.d., and assuming consistency of the MLE, we have,*

$$\sqrt{n}(\hat{\theta}_n^{ML} - \theta^*) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{I(\theta^*)}\right).$$

*Proof.* We will use the following notations for the normalised log-likelihood function and its first and second derivatives with respect to $\theta$:

$$\ell_n(\theta) = \frac{1}{n}\log f_\theta(X), \quad \ell_n'(\theta) = \frac{\partial}{\partial\theta}\left(\frac{1}{n}\log f_\theta(X)\right), \quad \ell_n''(\theta) = \frac{\partial^2}{\partial\theta^2}\left(\frac{1}{n}\log f_\theta(X)\right).$$

Recall the Mean Value Theorem (MVT),

**MVT:** Let $f$ be a continuous function on $[a, b]$ and differentiable on $(a, b)$. Then there exists a point $c \in (a, b)$ such that

$$f'(c) = \frac{f(a) - f(b)}{a - b}$$

Applying the MVT with $f = \ell_n'$, $a = \hat{\theta}_n^{ML}$ and $b = \theta^*$, we have that there exists a point $c = \tilde{\theta} \in (\hat{\theta}_N^{ML}, \theta^*)$ such that

$$\ell_n'(\hat{\theta}_n^{ML}) = \ell_n'(\theta^*) + \ell_n''(\tilde{\theta})(\hat{\theta}_n^{ML} - \theta^*).$$

Now, by definition of MLE, we have $\ell_n'(\hat{\theta}_N^{ML}) = 0$, hence

$$\sqrt{n}(\hat{\theta}_N^{ML} - \theta^*) = -\frac{\sqrt{n}\ell_n'(\theta^*)}{\ell_n''(\tilde{\theta})}.$$

The idea is to show that the numerator converges in distribution to a normal using the CLT, and that the denominator converges in probability to a constant by the weak LLN. Then, we will conclude invoking Slutsky's theorem.

For the numerator

$$\begin{aligned}\sqrt{n}\ell_n'(\theta^*) &= \sqrt{n}\left(\frac{1}{n}\frac{\partial}{\partial\theta}\log f_{\theta^*}(X)\right)\\ &= \sqrt{n}\left(\frac{1}{n}\frac{\partial}{\partial\theta}\log\prod_{i=1}^{n} f_{\theta^*}(X_i)\right)\\ &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f_{\theta^*}(X_i)\right)\\ &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f_{\theta^*}(X_i) - \mathbb{E}\left[\frac{\partial}{\partial\theta}\log f_{\theta^*}(X_1)\right]\right).\end{aligned}$$

In the last line, we use the fact that the expected value of the score function is zero. Hence, by the CLT

$$\sqrt{n}\ell_n'(\theta^*) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \mathbb{V}\left[\frac{\partial}{\partial\theta}\log f_{\theta^*}(X_1)\right]\right) = \mathcal{N}\left(0, I(\theta^*)\right).$$

For the denominator, for any $\theta$, we have

$$\begin{aligned}\ell_n''(\theta) &= \frac{1}{n}\left(\frac{\partial^2}{\partial\theta^2}\log f_{\theta}(X)\right)\\ &= \frac{1}{n}\left(\frac{\partial^2}{\partial\theta^2}\log\prod_{i=1}^{n} f_{\theta}(X_i)\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial^2}{\partial\theta^2}\log f_{\theta}(X_i)\right) \xrightarrow{\mathcal{P}} \mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f_{\theta}(X_1)\right]\end{aligned}$$

by the weak LLN.

Now, note that $\tilde{\theta} \in (\hat{\theta}_n^{ML}, \theta^*)$ by construction, and by consistency assumption of the MLE we have that $\hat{\theta}_n^{ML} \xrightarrow{\mathcal{P}} \theta^*$. Taken together, we have

$$\ell_n''(\tilde{\theta}) \xrightarrow{\mathcal{P}} \mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f_{\theta^*}(X_1)\right] = -I(\theta^*).$$

To summarize, we have shown that

$$\sqrt{n}\ell_n'(\theta^*) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(\theta^*))$$

and

$$\ell_n''(\tilde{\theta}) \xrightarrow{\mathcal{P}} -I(\theta^*).$$

We conclude by Slutsky's theorem

$$\sqrt{n}(\hat{\theta}_n^{ML} - \theta^*) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{I(\theta^*)}\right).$$

□

**Remark 4.20.** *When the MLE is unbiased, in the limit, it achieves the lowest possible variance, the Cramér–Rao lower bound since* $\mathbb{E}[\hat{\theta}] = \theta$.

**Example 4.21.** *Let* $X_1, \ldots, X_n$ *be i.i.d. samples from a Bernoulli*$(p)$ *distribution with true parameter* $p$. *Recall that* $\hat{p}_n^{ML} = \overline{X}$. *The log-density is*

$$\log f_p(X_1) = \log\left(p^{X_1}(1-p)^{1-X_1}\right) = X_1 \log p + (1 - X_1)\log(1-p)$$

*The score, or first derivative is given by*

$$\frac{\partial}{\partial p} \log f_p(X_1) = \frac{X_1}{p} + \frac{X_1 - 1}{1-p}.$$

*The second derivative is*

$$\frac{\partial^2}{\partial p^2} \log f_p(X_1) = -\frac{X_1}{p^2} + \frac{X_1 - 1}{(1-p)^2}.$$

*The Fisher information is the negative expected value of this second derivative or*

$$I(p) = -\mathbb{E}\left[-\frac{X_1}{p^2} + \frac{X_1 - 1}{(1-p)^2}\right] = \frac{1}{p(1-p)}.$$

*Thus,*

$$\sqrt{n}(\hat{p}_n^{ML} - p) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, p(1-p)\right).$$

**Exercise 4.22.** *Show that the MLE of the parameter of the exponential distributions* $E(\alpha)$ *is asymptotically Gaussian and compute the asymptotic variance.*

## 4.4 Bayes estimators

So far we have seen the *frequentist* approach to statistical inference i.e. inferential statements about an unknown parameter $\theta$ are interpreted in terms of repeated sampling. In contrast, the Bayesian approach treats $\theta$ as a *random variable* taking values in the parameter space $\Theta$.

The investigator's information and beliefs about the possible values for $\theta$, before any observation of data, are summarised by a *prior distribution* $\pi(\theta)$. When data $X$ are observed, the extra information about $\theta$ is combined with the prior to obtain the *posterior distribution* $\pi(\theta|X)$ for $\theta$ given $X$.

Let $L(\theta, a)$ be a loss function specified by the user. Common choices include the quadratic loss $L(\theta, a) = (\theta - a)^2$, absolute error loss $L(\theta, a) = |\theta - a|$ etc. When our estimate is $a$, the expected posterior loss is defined as follows

$$h(a) = \int L(\theta, a)\pi(\theta|X)\mathrm{d}\theta.$$

The Bayes estimator $\hat{\theta}_n^B$ minimises the expected posterior loss.

For example, for the quadratic loss

$$h(a) = \int (a-\theta)^2 \pi(\theta|X) \mathrm{d}\theta.$$

we have that

$$h'(\hat{\theta}_n^B) = 0 \quad \text{if} \quad \hat{\theta}_n^B \int \pi(\theta|X)\mathrm{d}\theta = \int \theta\pi(\theta|X)\mathrm{d}\theta.$$

So $\hat{\theta}_n^B = \int \theta\pi(\theta|x)\mathrm{d}\theta = \mathbb{E}[\theta|X]$, the posterior mean, minimises $h(a)$.

For the absolute error loss,

$$h(a) = \int |\theta - a|\pi(\theta|X)\mathrm{d}\theta = \int_{-\infty}^{a} (a-\theta)\pi(\theta|X)\mathrm{d}\theta + \int_a^{\infty} (\theta - a)\pi(\theta|X)\mathrm{d}\theta$$

$$= a\int_{-\infty}^{a} \pi(\theta|X)\mathrm{d}\theta - \int_{-\infty}^{a} \theta\pi(\theta|X)\mathrm{d}\theta + \int_a^{\infty} \theta\pi(\theta|X)\mathrm{d}\theta - a\int_a^{\infty} \pi(\theta|X)\mathrm{d}\theta$$

Now $h'(a) = 0$ if

$$\int_{-\infty}^{a} \pi(\theta|X)\mathrm{d}\theta = \int_a^{\infty} \pi(\theta|X)\mathrm{d}\theta.$$

This occurs when each side is $\frac{1}{2}$ (since the two integrals must sum to 1) so, by definition $\hat{\theta}_n^B = \mathbb{M}[\theta|X]$ must equal to the posterior median.

**Example 4.23.** *Suppose that $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, 1)$, and cosider a prior $\mu \sim N(0, \tau^{-2})$ for some known $\tau^{-2}$. Then, the posterior is given by (exercise)*

$$\pi(\mu|X) \propto f(X|\mu)\pi(\mu) \propto \exp\left(-\frac{1}{2}(n+\tau^2)\left(\mu - \frac{\sum_{i=1}^n X_i}{n+\tau^2}\right)^2\right)$$

*Hence, the posterior distribution of $\mu$ given the data $X$ is a Normal distribution with mean $\sum X_i/(n+\tau^2)$ and variance $1/(n+\tau^2)$. Because the normal density is symmetric, the posterior mean and the posterior median have the same value $\sum X_i/(n+\tau^2)$. This is the optimal Bayes estimate of $\mu$ under both quadratic and absolute error loss.*

**Exercise 4.24.** *Suppose that $X_1, \ldots, X_n$ are i.i.d. Poisson($\lambda$), and cosider a prior $\lambda \sim$ Exp(1) so that $\pi(\lambda) = e^{-\lambda}$ for $\lambda > 0$. Find the optimal Bayes estimator under the quadratic and absolute loss.*

# 5 Hypothesis testing

We now wish to construct a methodology to differentiate two possible scenarios from data. The standard set-up is to consider the null hypothesis $\mathcal{H}_0$ versus the alternative hypothesis $\mathcal{H}_1$, corresponding to two disjoint sets $\Theta_0$ and $\Theta_1$ of the parameter space:

$$\mathcal{H}_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad \mathcal{H}_1 : \theta \in \Theta_1. \tag{5.1}$$

Starting from a given sample $X = (X_1, ..., X_n)$, the rejection region $\mathcal{R}$ allows to retain or reject the hypothesis based on the range of outcomes of $X$, in the sense that, if $X \in \mathcal{R}$, then $\mathcal{H}_0$ is rejected, otherwise $\mathcal{H}_0$ cannot be rejected. Regarding the terminology, a hypothesis of the form $\Theta_0 = \{\theta_0\}$ is called simple, and the corresponding test is one-sided; a hypothesis of the form $\Theta_0 = \{\theta > \theta_0\}$ is called composite, and the test is two-sided. There are two types of errors pertaining to hypothesis testing. Type I errors occur when the test rejects $\mathcal{H}_0$ while it is actually true; Type II errors occur when the test keep $\mathcal{H}_0$ while $\mathcal{H}_1$ is true.

## 5.1 Simple tests

In simple hypothesis testing $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, we can write these two types of errors as

$$\begin{array}{ll} \text{Type I Error:} & \mathbb{P}_{\theta_0}(X \in \mathcal{R}), \\ \text{Type II Error:} & \mathbb{P}_{\theta_1}(X \notin \mathcal{R}). \end{array} \tag{5.2}$$

The goal of any test is obviously to minimise the error. However, in order to minimise the Type I error, one needs to consider a small rejection region $\mathcal{R}$, which in turn is going to yield a large Type II error, so some balance needs to be set between the two. The idea is to set an acceptable threshold for the error, as follows:

**Definition 5.1.** *The power function $\beta(\cdot)$ and the level $\alpha \in (0, 1)$ of a test with rejection region $\mathcal{R}$ are defined as*

$$\mathbb{P}_{\theta_1}(X \in \mathcal{R}) \qquad \textit{and} \qquad \mathbb{P}_{\theta_0}(X \in \mathcal{R}) \leq \alpha.$$

*If $\mathbb{P}_{\theta_0}(X \in \mathcal{R}) = \alpha$, then $\alpha$ is called the size of the test.*

In order to simplify the terminology, we shall call a test with rejection region $\mathcal{R}$ and $\mathcal{R}$-test from now on.

**Definition 5.2.** *For a given level $\alpha$, an $\mathcal{R}^*$-test with level $\alpha$ is called the most powerful test if $\mathbb{P}_{\theta_1}(X \in \mathcal{R}^*) \geq \mathbb{P}_{\theta_1}(X \in \mathcal{R})$ for any $\mathcal{R}$-test of level $\alpha$.*

The higher the power function the lower the Type II error, with a given bound on the Type I error. There are many such hypothesis tests in the literature, including the Neyman-Pearson test, which we present now.

Define a rejection region of the form

$$\mathcal{R} := \left\{ X : \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} > c \right\}, \tag{5.3}$$

for some $c > 0$, where $f_\theta(X) := \prod_{i=1}^n f_\theta(X_i)$ denotes the likelihood function, and the ratio is a random variable called the likelihood ratio.

**Theorem 5.3.** *[Neyman-Pearson Lemma] Let $\alpha \in (0,1)$. If there exists $c^* > 0$ such that $\mathbb{P}_{\theta_0}\left(\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} > c^*\right) = \alpha$, then the test is the most powerful test of level $\alpha$.*

*Proof.* Proving the theorem is equivalent to showing that for all rejection region $\mathcal{R}$ such that $\mathbb{P}_{\theta_0}(X \in \mathcal{R}) \leq \alpha$, then $\mathbb{P}_{\theta_1}(X \in \mathcal{R}) \leq \mathbb{P}_{\theta_1}(X \in \mathcal{R}^*)$, or equivalently $\mathbb{P}_{\theta_1}(X \notin \mathcal{R}) \geq \mathbb{P}_{\theta_1}(X \notin \mathcal{R}^*)$, where $\mathcal{R}^*$ denotes the optimal rejection region corresponding to the optimal value of $c^*$ in the theorem. Now,

$$\begin{aligned}
\mathbb{P}_{\theta_1}(X \in \mathcal{R}^*) - \mathbb{P}_{\theta_1}(X \in \mathcal{R}) &= \int_{\mathcal{R}^*} f_{\theta_1}(\mathrm{x})\mathrm{dx} - \int_{\mathcal{R}} f_{\theta_1}(\mathrm{x})\mathrm{dx} \\
&= \left( \int_{\mathcal{R}^* \setminus \mathcal{R}} - \int_{\mathcal{R} \setminus \mathcal{R}^*} \right) f_{\theta_1}(\mathrm{x})\mathrm{dx}.
\end{aligned}$$

Since $\mathcal{R}^* \setminus \mathcal{R} \subset \mathcal{R}^*$, then $\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} > c^*$ on this set, and obviously $\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} \leq c^*$ on $\mathcal{R} \setminus \mathcal{R}^*$. Therefore,

$$\begin{aligned}
\left( \int_{\mathcal{R}^* \setminus \mathcal{R}} - \int_{\mathcal{R} \setminus \mathcal{R}^*} \right) f_{\theta_1}(\mathrm{x})\mathrm{dx} &\geq c^* \left( \int_{\mathcal{R}^* \setminus \mathcal{R}} - \int_{\mathcal{R} \setminus \mathcal{R}^*} \right) f_{\theta_0}(\mathrm{x})\mathrm{dx} \\
&= c^* \left( \int_{\mathcal{R}^*} - \int_{\mathcal{R}} \right) f_{\theta_0}(\mathrm{x})\mathrm{dx} \\
&= c^* \left[ \mathbb{P}_{\theta_0}(X \in \mathcal{R}^*) - \mathbb{P}_{\theta_0}(X \in \mathcal{R}) \right].
\end{aligned}$$

By assumption, $\mathbb{P}_{\theta_0}(X \in \mathcal{R}) \leq \alpha$ and $\mathbb{P}_{\theta_0}(X \in \mathcal{R}^*) = \alpha$, therefore, the right-hand side of the last inequality is non negative, and the theorem follows. □

We now introduce one of the key concepts in hypothesis testing, called the p-value, which we shall denote by $\pi_0$:

**Definition 5.4.** *Consider a test of size $\alpha \in (0,1)$ and corresponding rejection region $\mathcal{R}_\alpha$. The p-value $\pi_0$ is defined as the smallest level at which the null hypothesis can be rejected, i.e.*

$$\pi_0 := \inf \{ \alpha : X \in \mathcal{R}_\alpha \}.$$

Clearly the possible range of values is $(0,1)$. When the p-value is below 1%, then there is very strong evidence that the null hypothesis should be rejected; the range $(1\%, 5\%)$ represents strong evidence, $(5\%, 10\%)$ weak evidence, and when the p-value is greater than 10%, then the test is inconclusive, in the sense that we cannot decently reject the null hypothesis.

**Example 5.5.** *We consider the statistical model* $\mathcal{F} = \{\mathcal{N}(\theta, \sigma^2), \theta \in \Theta\}$*, where the variance* $\sigma^2$ *is known, and we consider*

$$\mathcal{H}_0 : \theta \in \Theta_0 = \{0\} \qquad \textit{versus} \qquad \mathcal{H}_1 : \theta \in \Theta_1 = \{1\}.$$

*The likelihood function reads (see Example 4.14)*

$$f_\theta(X) = \left(\sigma\sqrt{2\pi}\right)^n \prod_{i=1}^n \exp\left\{-\frac{(X_i-\theta)^2}{2\sigma^2}\right\},$$

*and hence the likelihood ratio can be computed as*

$$\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} = \exp\left\{\frac{n}{2\sigma^2}\left(2\overline{X}-1\right)\right\}.$$

*The rejection region* (5.3) *therefore reads explicitly*

$$\mathcal{R} = \left\{\exp\left\{\frac{n}{2\sigma^2}\left(2\overline{X}-1\right)\right\} \geq \widetilde{c}\right\} = \left\{\overline{X} > c\right\}, \tag{5.4}$$

*for some* $\widetilde{c} > 0$ *with* $c = \frac{1}{2} + \frac{\sigma^2}{n}\log(\widetilde{c})$*. To choose* $c$*, we equate* $\mathbb{P}_{\theta_0}(X \in \mathcal{R}) = \alpha = \mathbb{P}_{\theta_0}(\overline{X} \geq c)$*. Since the sample is Gaussian* $\mathcal{N}(0, \sigma^2)$ *under the null hypothesis, then* $\overline{X} \sim \mathcal{N}(0, \frac{\sigma^2}{n})$*, and therefore*

$$c_\alpha = \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha) = \frac{\sigma}{\sqrt{n}}q_{1-\alpha}. \tag{5.5}$$

*To compute the power of the test, we can write*

$$\mathbb{P}_{\theta_1}\left(X \in \mathcal{R}\right) = \mathbb{P}_{\theta_1}\left(\overline{X} > c\right) = \mathbb{P}_{\theta_1}\left(\mathcal{N}(0,1) > \frac{(c-\theta_1)\sqrt{n}}{\sigma}\right) = 1 - \Phi\left(\frac{(c-\theta_1)\sqrt{n}}{\sigma}\right).$$

*Recall now that the p-value* $\pi_0$ *of a test is, for a given fixed sample, the largest value of* $\alpha$ *such that the null hypothesis* $\mathcal{H}_0$ *is not rejected. From* (5.4)*, for a given level* $\alpha$*, the rejection region is of the form* $\mathcal{R} = \mathcal{R}_\alpha = \left\{\overline{X} > c_\alpha\right\}$*, with* $c_\alpha$ *given in* (5.5)*, or equivalently*

$$1 - \alpha = \Phi\left(\frac{\sqrt{n}c_\alpha}{\sigma}\right).$$

*Therefore, for a given* $\overline{X}$*, the threshold from accepting* $\mathcal{H}_0$ *to rejecting it is* $\alpha^*(\overline{X})$ *such that* $c_\alpha = \overline{X}$*, i.e.*

$$1 - \alpha^*(\overline{X}) = \Phi\left(\frac{\sqrt{n}c_\alpha}{\sigma}\right) = \Phi\left(\frac{\sqrt{n}\overline{X}}{\sigma}\right),$$

*and therefore the critical (p-value) threshold is equal to*

$$\pi_0 = \alpha^*(\overline{X}) = 1 - \Phi\left(\frac{\overline{X}\sqrt{n}}{\sigma}\right).$$

## 5.2 Composite tests

We now consider composite tests, that is tests of the form $\Theta_0 = \{\theta > \theta_0\}$ versus $\Theta_1 = \{\theta \leq \theta_0\}$. We slightly modify the definitions of the error (5.2) in the following form:

$$\text{Type I Error:} \quad \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(X \in \mathcal{R}). \tag{5.6}$$

Note that, in the composite case, we cannot make sense of the notion of Type II errors, but we shall use, similar to the simple case, the notions of level, size and power of a test:

**Definition 5.6.** *The level $\alpha \in (0,1)$ of a test with rejection region $\mathcal{R}$ is such that*

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta \left(X \in \mathcal{R}\right) \leq \alpha.$$

*If the supremum is equal to $\alpha$, then $\alpha$ is called the size of the test. The power function $\beta : \Theta \to [0,1]$ is defined as*

$$\beta(\theta) := \mathbb{P}_\theta \left(X \in \mathcal{R}\right).$$

This is a slight abuse of language as we previously defined the power function as a function of sets, but it should not create any confusion here. In order to extend the Neyman-Pearson lemma to the composite case, we need to introduce the following terminology:

**Definition 5.7.** *A test $\mathcal{R}^*$ with level $\alpha$ is called Uniformly Most Powerful (UMP) if $\mathbb{P}_\theta \left(X \in \mathcal{R}\right) \leq \mathbb{P}_\theta \left(X \in \mathcal{R}^*\right)$ for all $\theta \in \Theta_1$ and any test $\mathcal{R}$ of level $\alpha$.*

A test is called consistent if $\beta(\theta)$ converges to 1 as the sample size tends to infinity, for any $\theta \in \Theta_1$, and is called unbiased if

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \inf_{\theta \in \Theta_1} \beta(\theta).$$

**Example 5.8.** *Consider $\mathcal{F} = \{\mathcal{N}(\theta, \sigma^2), \theta \in \mathbb{R}\}$, with $\sigma > 0$ known, and the hypotheses $\Theta_0 = (-\infty, 0]$ and $\Theta_1 = (0, +\infty)$. As in Example 5.5, consider the test with rejection region*

$$\mathcal{R} := \{\overline{X} > c_\alpha\}, \qquad \text{with } c_\alpha := \frac{\sigma}{\sqrt{n}} q_{1-\alpha}. \tag{5.7}$$

*Since, for any $\theta \in \mathbb{R}$, in the model $\mathcal{F}_\theta$, the random variable $\sqrt{n}\frac{\overline{X}-\theta}{\sigma} \sim \mathcal{N}(0,1)$, then*

$$\begin{aligned}\beta(\theta) &:= \mathbb{P}_\theta\left(X \in \mathcal{R}\right)\\ &= \mathbb{P}(\overline{X} > c_\alpha)\\ &= \mathbb{P}_\theta\left(\frac{\sqrt{n}(\overline{X}-\theta)}{\sigma} > \frac{\sqrt{n}(c_\alpha-\theta)}{\sigma}\right)\\ &= 1-\Phi\left(\frac{\sqrt{n}(c_\alpha-\theta)}{\sigma}\right)\\ &= \Phi\left(\frac{\sqrt{n}(\theta-c_\alpha)}{\sigma}\right)\\ &= \Phi\left(\frac{\sqrt{n}\theta}{\sigma}-q_{1-\alpha}\right).\end{aligned}$$

*Note further that $\beta(0) = \alpha$. Since the function $\Phi$ is monotone, then*

$$\sup_{\theta\in\Theta_0} \mathbb{P}_\theta(X \in \mathcal{R}) = \sup_{\theta\in\Theta_0} \beta(\theta) = \beta(0) = \Phi\left(q_{1-\alpha}\right) = \alpha.$$

*Fix now some value $\theta' \in \Theta_1$, and consider the simple hypotheses*

$$\widetilde{\mathcal{H}}_0 : \theta = 0 \qquad vs \qquad \widetilde{\mathcal{H}}_1 : \theta = \theta'.$$

*By Neyman-Pearson's lemma (Theorem 5.3) and Example 5.5, the test $\mathcal{R}$ in (5.7) satisfies*

$$\mathbb{P}_{\theta'}(X \in \mathcal{R}) \geq \mathbb{P}_{\theta'}(X \in \mathcal{R}_\alpha), \tag{5.8}$$

*for any test $\mathcal{R}_\alpha$ of level $\alpha$ (e.g. such that $\mathbb{P}_0(X \in \mathcal{R}_\alpha) \leq \alpha$). Since $0 \in \Theta_0$, any test satisfying*

$$\sup_{\theta\in\Theta_0} \mathbb{P}_\theta(X \in \mathcal{R}_\alpha) \leq \alpha$$

*also satisfies $\mathbb{P}_0(X \in \mathcal{R}_\alpha) \leq \alpha$, and therefore, for any test $\mathcal{R}_\alpha$ of level $\alpha$ for the null hypothesis $\mathcal{H}_0 : \theta \leq 0$ against $\mathcal{H}_1 : \theta > 0$, and for any $\theta' > 0$, the inequality (5.8) holds, so that $\mathcal{R}$ is UMP.*

**Exercise 5.9.** *Show that the test in Example 5.8 is consistent and unbiased.*

**Remark 5.10.** *In our recurring example above, we always assumed that the variance $\sigma^2$ of the Gaussian sample was known. In case it is not, we can however replace it by the unbiased estimator $ns_n^2/(n-1)$, where*

$$s_n^2 := \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

*In this case, the quantiles appearing in the rejection region will not be those of the Gaussian distribution any longer, but those of the Student distribution.*

**Exercise 5.11.** *Consider $\mathcal{F} = \{\mathcal{N}(\mu,\theta^2), \theta > 0\}$, with $\mu \in \mathbb{R}$ known, and the hypotheses $\Theta_0 = (0,\sigma_0]$ and $\Theta_1 = (\sigma_0,+\infty)$, for some $\sigma_0 > 0$. Analyse the test*

$$\mathcal{R} := \left\{\frac{f_\theta(X)}{f_{\sigma_0}(X)} > c\right\},$$

*for some constant $c > 0$ to be determined, where $f$ denotes as usual the likelihood function.*

**Solution 5.12.** *The log-likelihood ratio takes the form*

$$\frac{f_\theta(X)}{f_{\sigma_0}(X)} = \frac{\theta}{\sigma_0} \exp\left\{ \left( \frac{1}{2\sigma_0^2} - \frac{1}{2\theta^2} \right) S_n \right\},$$

*with $S_n := \sum_{i=1}^n (X_i - \mu)^2$, so that the rejection region can be written $\mathcal{R} = \{S_n > \tilde{c}\}$. As before, we choose the constant $\tilde{c}$ such that $\mathbb{P}_{\sigma_0}(X \in \mathcal{R}) = \alpha$. Since the sample is assumed to be Gaussian, the random variable $S_n/\sigma_0^2$ follows, under $\mathbb{P}_{\sigma_0}$, a Chi-Squared $\chi_n^2$ distribution, and hence*

$$\mathbb{P}_{\sigma_0}(X \in \mathcal{R}) = \mathbb{P}_{\sigma_0}(S_n > \tilde{c}) = \mathbb{P}_{\sigma_0}\left( \frac{S_n}{\sigma_0^2} > \frac{\tilde{c}}{\sigma_0^2} \right) = 1 - F_{\chi_n^2}\left( \frac{\tilde{c}}{\sigma_0^2} \right),$$

*and therefore $\tilde{c} = q_{1-\alpha}^{\chi_n^2} \sigma_0^2$. The power of the test can then be computed as*

$$\begin{aligned}
\beta(\theta) &= \mathbb{P}_\theta(X \in \mathcal{R}) \\
&= \mathbb{P}_\theta\left(S_n > \tilde{c}\right) \\
&= \mathbb{P}_\theta\left(S_n > q_{1-\alpha}^{\chi_n^2} \sigma_0^2\right) \\
&= \mathbb{P}_\theta\left(\frac{S_n}{\theta^2} > \frac{\sigma_0^2}{\theta^2} q_{1-\alpha}^{\chi_n^2}\right) \\
&= 1 - F_{\chi_n^2}\left(\frac{\sigma_0^2}{\theta^2} q_{1-\alpha}^{\chi_n^2}\right),
\end{aligned}$$

*and it is then easy to see that the test is of level $\alpha$ since*

$$\beta(\theta) = 1 - F_{\chi_n^2}\left(\frac{\sigma_0^2}{\theta^2} q_{1-\alpha}^{\chi_n^2}\right) \leq 1 - F_{\chi_n^2}\left(q_{1-\alpha}^{\chi_n^2}\right) = \alpha.$$

## 5.3 Confidence intervals

**Definition 5.13.** *Let $\alpha \in (0,1)$. The $1-\alpha$ confidence set $\mathcal{C}_n$ for $\theta$, in general depending on the data, is such that*

$$\mathbb{P}_\theta\left(\theta \in \mathcal{C}_n\right) \geq 1 - \alpha, \qquad \text{for all } \theta \in \Theta.$$

**Example 5.14.** *Consider the statistical model $\mathcal{F} = \{\mathcal{N}(\theta, \sigma^2), \theta \in \mathbb{R}\}$ for $\sigma > 0$ known, and let $\alpha \in (0,1)$. Consider now the (random) interval*

$$\mathcal{C}_n := \left[ \overline{X} - \frac{\sigma}{\sqrt{n}} q_{1-\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} q_{1-\alpha/2} \right].$$

*Then we can compute, for any* $\theta \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}_\theta(\theta \in \mathcal{C}_n) &= \mathbb{P}_\theta\left(\left|\overline{X} - \theta\right| \leq \frac{\sigma}{\sqrt{n}} q_{1-\alpha/2}\right) \\ &= \mathbb{P}_\theta\left(\frac{\left|\overline{X} - \theta\right|}{\sigma/\sqrt{n}} \leq q_{1-\alpha/2}\right) \\ &= \mathbb{P}_\theta\left(|Z| \leq q_{1-\alpha/2}\right) = 1 - \alpha,\end{aligned}$$

*where* $Z \sim \mathcal{N}(0,1)$*, so that* $\mathcal{C}_n$ *is indeed a confidence interval of level* $1-\alpha$*.*

**Exercise 5.15.** *For* $\alpha \in (0,1)$ *and the statistical model* $\mathcal{F} = \{\mathcal{N}(\theta, \sigma^2), \theta \in \mathbb{R}\}$ *for* $\sigma > 0$ *known, show that the (random) interval*

$$\mathcal{C}_n := \left[\overline{X} - \frac{\sigma}{\sqrt{n}} q_{1-3\alpha/4}, \overline{X} + \frac{\sigma}{\sqrt{n}} q_{1-\alpha/4}\right].$$

*is also a confidence interval of level* $1-\alpha$*. How does it compare to the one in Example 5.14?*

It often happens that we can construct confidence intervals based on the Gaussian distribution:

**Theorem 5.16.** *Assume that* $\widehat{\theta}_n \sim \mathcal{N}(\theta, \widehat{\sigma}_n^2)$*, and define the interval*

$$\mathcal{C}_n := \left(\widehat{\theta}_n - z_{\alpha/2}\widehat{\sigma}_n, \widehat{\theta}_n + z_{\alpha/2}\widehat{\sigma}_n\right),$$

*where* $z_{\alpha/2} := \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$*, with* $\Phi$ *the Gaussian cdf. Then* $\lim\limits_{n\uparrow\infty} \mathbb{P}_\theta(\theta \in \mathcal{C}_n) = 1 - \alpha$*.*

*Proof.* Let $Z \in \mathcal{N}(0,1)$. Note first that the definition of $z_{\alpha/2}$ is equivalent to $\mathbb{P}(Z > z_{\alpha/2}) = \alpha/2$, or $\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$. Therefore,

$$\mathbb{P}_\theta(\theta \in \mathcal{C}_n) = \mathbb{P}_\theta\left(\widehat{\theta}_n - z_{\alpha/2}\widehat{\sigma}_n < \theta < \widehat{\theta}_n + z_{\alpha/2}\widehat{\sigma}_n\right) = \mathbb{P}_\theta\left(-z_{\alpha/2} < \frac{\theta - \widehat{\theta}_n}{\widehat{\sigma}_n} < z_{\alpha/2}\right).$$

Since the random sequence $\left(\frac{\theta - \widehat{\theta}_n}{\widehat{\sigma}_n}\right)_{n>0}$ converges in probability to $\mathcal{N}(0,1)$ as $n$ tends to infinity, the theorem follows. □

**Confidence interval for the cumulative distribution function** We built so far estimators and confidence intervals thereof; those were parametric in the sense that we assumed the true distribution to be known up to knowledge of its parameters. We may however challenge this and, getting back to the empirical cdf constructed at the very beginning of the course, try and determine whether the latter is in fact a good estimator for the true cdf.

We already proved that, pointwise for any $x \in \mathbb{R}$, the empirical cdf $\widehat{F}_n(x)$ is distributed as a Binomial distribution, so that the Central Limit Theorem implies that

$$\frac{\sqrt{n}\left(\widehat{F}_n(x) - F(x)\right)}{\sqrt{F(x)(1-F(x))}}$$

converges in distribution to a centered Gaussian distribution with unit variance as $n$ tends to infinity.

Now, $F(x)$ is by definition unknown. However, since $\widehat{F}_n(x)$ converges in probability to $F(x)$, then Slutsky's theorem implies that

$$\frac{\sqrt{n}\left(\widehat{F}_n(x) - F(x)\right)}{\sqrt{\widehat{F}_n(x)(1-\widehat{F}_n(x))}}$$

still converges in distribution to a centered Gaussian distribution with unit variance.

Therefore, for any $\alpha \in (0,1)$, a $(1-\alpha)$ confidence interval for $F(x)$ is given by

$$\mathcal{C}_n(\alpha) = \left[\widehat{F}_n(x) - q_{1-\alpha/2}\sqrt{\frac{\widehat{F}_n(x)\left(1-\widehat{F}_n(x)\right)}{n}}, \widehat{F}_n(x) + q_{1-\alpha/2}\sqrt{\frac{\widehat{F}_n(x)\left(1-\widehat{F}_n(x)\right)}{n}}\right].$$

In terms of hypothesis testing, pointwise again, it makes sense to consider the following test:

$$\mathcal{H}_0 : F(x) = F_0(x) \qquad \text{vs} \qquad \mathcal{H}_1 : F(x) \neq F_0(x), \tag{5.9}$$

for some given $F_0(x)$. Following similar steps to before, we can construct a test of level $\alpha$ based on a rejection region of the form

$$\mathcal{R} = \left\{ \frac{\left|\widehat{F}_n(x) - F_0(x)\right|}{\sqrt{F_0(x)\left(1-F_0(x)\right)}} > \frac{q_{1-\alpha/2}}{\sqrt{n}} \right\}.$$

**Testing for the distribution** The previous paragraph focuses on estimator a cumulative distribution pointwise. We now wish to extend this to a uniform statement. Consider the so-called Kolmogorov-Smirnov statistic

$$D_n := \sup_{x \in \mathbb{R}} \left|\widehat{F}_n(x) - F(x)\right|.$$

Glivenko-Cantelli' result (Theorem 4.2) states that $D_n$ converges almost surely to zero as $n$ tends to infinity. This was refined by Kolmogorov as follows:

**Theorem 5.17** (Kolmogorov-Smirnov)**.** *For any $z \in \mathbb{R}$, the probability $\mathbb{P}\left(\sqrt{n}D_n \leq z\right)$ converges in distribution to $H(z)$ where the function $H$ is the cumulative distribution function of the Kolmogorov-Smirnov distribution and is given explicitly by*

$$H(z) := 1 - 2\sum_{k \geq 1} (-1)^{k-1} \exp\left(-2k^2 z\right).$$

**Remark 5.18.** *In fact, the distribution $H(\cdot)$ appearing in the theorem is exactly that of the supremum of the Brownian bridge on $[0,1]$, so that the theorem can equivalently be stated as the convergence in distribution of $\sqrt{n}D_n$ to $\sup\{|B(F_0(t)|, t \in [0,1]\}$, where $B$ is a Brownian bridge and $F_0$ the hypothesized distribution.*

Consider now the test

$$\mathcal{H}_0 : F = F_0 \qquad \text{vs} \qquad \mathcal{H}_1 : F \neq F_0,$$

for some given cdf $F_0$. Note that this represents a uniform version of the test (5.9). Now, under the null hypothesis $\mathcal{H}_0$, since $F_0$ is given a priori, the distribution of $D_n$, for any $n$ fixed, can be tabulated. If $\mathcal{H}_0$ fails, however, calling $F$ the true cdf, we know by the law of large numbers that $\widehat{F}_n$ converges to $F$, so that, for large enough $n$, $D_n > \delta$, for some $\delta > 0$, and hence $\sqrt{n}D_n > \delta\sqrt{n}$ and $\sqrt{n}D_n$ clearly tends to infinity as $n$ becomes large. We can therefore construct a rejection region of the form

$$\mathcal{R} = \left\{\sqrt{n}D_n > c\right\}$$

Now, the Type-I error reads

$$\mathbb{P}_{\mathcal{H}_0}(X \in \mathcal{R}) = \mathbb{P}_{\mathcal{H}_0}\left(\sqrt{n}D_n > c\right),$$

which converges to $1 - H(c)$ as $n$ tends to infinity. This corresponds to an asymptotic level $\alpha \in (0,1)$ if $1 - H(c) = \alpha$, and we can therefore determine the threshold $c$ as

$$c = H^{-1}(1-\alpha).$$

**Asymptotic tests** The tests discussed above are based on some knowledge of the distribution, which is rarely the case in practice. Suppose that the iid sequence $(X_1, \ldots, X_n)$ has constant mean $\theta \in \mathbb{R}$ and finite strictly positive variance $\sigma^2(\theta) := \mathbb{V}_\theta(X_1)$. The Central Limit Theorem implies that $\sqrt{n}(\overline{X} - \theta)/\sigma(\theta)$ converges in distribution, under $\mathbb{P}_\theta$, to $\mathcal{N}(0,1)$ as $n$ tends to infinity. If the map $\sigma(\cdot)$ is continuous, using the fact that $\overline{X}$ converges to $\theta$ in probability, then Slutsky's theorem yields that $\sqrt{n}(\overline{X} - \theta)/\sigma(\overline{X})$ converges in distribution, under $\mathbb{P}_\theta$, to $\mathcal{N}(0,1)$ as $n$ tends to infinity. Consider therefore the following test:

$$\mathcal{H}_0 : \theta = \theta_0 \qquad \text{vs} \qquad \theta > \theta_0,$$

with rejection region

$$\mathcal{R} := \left\{\overline{X} > \theta_0 + \frac{\sigma(\overline{X})}{\sqrt{n}} q_{1-\alpha}\right\}.$$

Then

$$\begin{aligned}\lim_{n\uparrow\infty} \mathbb{P}_\theta(X \in \mathcal{R}) &= \lim_{n\uparrow\infty} \mathbb{P}_\theta\left(\overline{X} > \theta_0 + \frac{\sigma(\overline{X})}{\sqrt{n}} q_{1-\alpha}\right) \\ &= \lim_{n\uparrow\infty} \mathbb{P}_\theta\left(\frac{\sqrt{n}(\overline{X} - \theta_0)}{\sigma(\overline{X})} > q_{1-\alpha}\right) = \alpha,\end{aligned}$$

This leads us to the following definition:

**Definition 5.19.** *A test $\mathcal{R}$ of the null hypothesis $\mathcal{H}_0 : \theta \in \Theta_0$ vs the alternative $\mathcal{H}_1 : \theta \in \Theta_1$ is called a test of asymptotic level $\alpha$ if*

$$\sup_{\theta \in \Theta_0} \lim_{n \uparrow \infty} \mathbb{P}_\theta(X \in \mathcal{R}) \leq \alpha.$$

**Example 5.20.** *Consider the maximum likelihood estimator $\widehat{\theta}_n^{\mathrm{ML}}$ for some statistical model. Under the regularity hypotheses, we know that it converges in probability to $\theta$ and Theorem* **??** *implies that $\sqrt{nI(\theta)}\left(\widehat{\theta}_n^{\mathrm{ML}} - \theta\right)$ converges in distribution to $\mathcal{N}(0,1)$ as $n$ tends to infinity. Assuming that the Fisher information $I(\cdot)$ is continuous, then, similarly to above, $I(\widehat{\theta}_n^{\mathrm{ML}})$ converges in probability to $I(\theta)$ and $\sqrt{nI(\widehat{\theta}_n^{\mathrm{ML}})}\left(\widehat{\theta}_n^{\mathrm{ML}} - \theta\right)$ converges in distribution to $\mathcal{N}(0,1)$ as $n$ tends to infinity. This provides a natural test of asymptotic level $\alpha$ as*

$$\mathcal{H}_0 : \theta = \theta_0 \qquad vs \qquad \theta \neq \theta_0,$$

*with rejection region*

$$\mathcal{R} := \left\{ \left| \widehat{\theta}_n^{\mathrm{ML}} - \theta_0 \right| > \frac{q_{1-\alpha/2}}{\sqrt{nI(\widehat{\theta}_n^{\mathrm{ML}})}} \right\}.$$

# 6 Linear regression

We start this chapter with the simple case of dataset of scalar input-output pairs $(x_i, y_i)_{i=1,\dots,n}$, from which some dependency can be observed. We would like to determine a relation of the form $y_i \approx f(x_i)$ for some function $f : \mathbb{R} \to \mathbb{R}$. A general formulation can be stated as

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \mathcal{C}\left(y_i - f(x_i)\right),$$

for some given cost function $\mathcal{C}$, where $\mathcal{F}$ is a class of functions of interest.

## 6.1 The univariate case

In linear regression, we assume a dependence of the form

$$y_i = f(x_i) + \varepsilon_i, \qquad \text{for } i = 1, \dots, n$$

where $f(x) \equiv \alpha + \beta x$ is a linear function, and the sequence $(\varepsilon_i)_{i=1,\dots,n}$ are centered independent random noises with constant variance $\sigma^2$. We can rewrite this as

$$\mathbf{Y} = \alpha \mathbb{1} + \beta \mathbf{X} + \boldsymbol{\varepsilon}$$

with $\mathbf{X} = (x_1, \dots, x_n)^\top \in \mathbb{R}^n$, $\mathbf{Y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top \in \mathbb{R}^n$.

Since $\alpha$ and $\beta$ are the only two parameters here, the infinite-dimensional minimisation problem (over $\mathcal{F}$) reduces to one over $\mathbb{R}^2$. In the least-square minimisation problem, we consider the $L^2$ loss function:

$$\mathfrak{L}(\alpha, \beta) := \|\mathbf{Y} - (\alpha \mathbb{1} + \beta \mathbf{X})\|_2^2 = \sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i\right)^2.$$

The least-square estimator $(\widehat{\alpha}, \widehat{\beta})$ is the solution to the minimisation

$$(\widehat{\alpha}, \widehat{\beta}) := \operatorname*{arg\,min}_{(\alpha,\beta)} \mathfrak{L}(\alpha, \beta). \tag{6.1}$$

**Proposition 6.1.** *The solution to* (6.1) *reads*

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x} \qquad \textit{and} \qquad \widehat{\beta} = \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \sum_{i=1}^{n} \left(x_i - \overline{x}\right) y_i.$$

Note that the computation of these estimators is purely deterministic and do not rely on the i.i.d. assumption made about the errors.

*Proof.* Clearly, the function $\mathfrak{L}$ is smooth, convex and hence admits a unique minimum

$$\nabla \mathfrak{L}(\alpha, \beta) = \begin{pmatrix} \partial_\alpha \mathfrak{L}(\alpha, \beta) \\ \partial_\beta \mathfrak{L}(\alpha, \beta) \end{pmatrix} = -2 \begin{pmatrix} \sum_{i=1}^n (y_i - \alpha - \beta x_i) \\ \sum_{i=1}^n x_i (y_i - \alpha - \beta x_i) \end{pmatrix}.$$

The first element can be written as

$$\sum_{i=1}^n (y_i - \alpha - \beta x_i) = n (\overline{y} - \alpha - \beta \overline{x}),$$

which is equal to zero if and only if $\alpha = \overline{y} - \beta \overline{x}$. Regarding the gradient with respect to $\beta$, we can write, plugging this optimal $\alpha$,

$$\sum_{i=1}^n x_i (y_i - \alpha - \beta x_i) = \sum_{i=1}^n x_i [y_i - (\overline{y} - \beta \overline{x}) - \beta x_i] = \sum_{i=1}^n x_i y_i - \overline{y} \sum_{i=1}^n x_i + \beta \overline{x} \sum_{i=1}^n x_i - \beta \sum_{i=1}^n x_i^2,$$

which is equal to zero if and only if

$$\beta = \frac{\sum_{i=1}^n x_i (y_i - \overline{y})}{\sum_{i=1}^n x_i (x_i - \overline{x})},$$

and the proposition follows noting that

$$\begin{aligned} \|\mathbf{X} - \overline{x} \mathbb{1}\|_2^2 &= \sum_{i=1}^n (x_i - \overline{x})^2 = \sum_{i=1}^n \left(x_i^2 + \overline{x}^2 - 2 x_i \overline{x}\right) = \sum_{i=1}^n x_i^2 + n \overline{x}^2 - 2 \overline{x} \sum_{i=1}^n x_i \\ &= \sum_{i=1}^n x_i^2 - n \overline{x}^2 = \sum_{i=1}^n x_i (x_i - \overline{x}). \end{aligned} \tag{6.2}$$

□

Let us prove some properties of these estimators:

**Theorem 6.2.** *The LSE are unbiased and*

$$\mathbb{V}\left[\widehat{\beta}\right] = \frac{\sigma^2}{\|\mathbf{X} - \overline{x} \mathbb{1}\|_2^2}, \qquad \mathbb{V}\left[\widehat{\alpha}\right] = \frac{\sigma^2 \|\mathbf{X}\|_2^2}{n \|\mathbf{X} - \overline{x} \mathbb{1}\|_2^2}, \qquad \mathrm{Cov}\left(\widehat{\alpha}, \widehat{\beta}\right) = -\frac{\sigma^2 \overline{x}}{\|\mathbf{X} - \overline{x} \mathbb{1}\|_2^2}.$$

*Proof.* Let us first show the following alternative representation for $\widehat{\beta}$:

$$\widehat{\beta} = \beta + \frac{1}{\|\mathbf{X} - \overline{x} \mathbb{1}\|_2^2} \sum_{i=1}^n (x_i - \overline{x}) \varepsilon_i. \tag{6.3}$$

Combining the expression for $\widehat{\beta}$ in Proposition 6.1 and the definition of the linear model $\mathbf{Y} = \alpha\mathbb{1} + \beta\mathbf{X} + \boldsymbol{\varepsilon}$, we can write

$$\begin{aligned}\widehat{\beta} &= \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \sum_{i=1}^{n} (x_i - \overline{x})\, y_i \\ &= \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \sum_{i=1}^{n} (x_i - \overline{x})\, [\alpha + \beta x_i + \varepsilon_i] \\ &= \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \sum_{i=1}^{n} (x_i - \overline{x})\, [\alpha + \beta x_i] + \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \sum_{i=1}^{n} (x_i - \overline{x})\, \varepsilon_i.\end{aligned}$$

Regarding the first term, we can write

$$\frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \sum_{i=1}^{n} (x_i - \overline{x})\, [\alpha + \beta x_i] = \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \left[\alpha \sum_{i=1}^{n} (x_i - \overline{x}) + \beta \sum_{i=1}^{n} x_i\, (x_i - \overline{x})\right] = \beta,$$

since $\displaystyle\sum_{i=1}^{n} (x_i - \overline{x}) = 0$ and using the identity (6.2), so that (6.3) holds. Since the sequence $(\varepsilon_1, \ldots, \varepsilon_n)$ is centered, then clearly $\mathbb{E}[\widehat{\beta}] = \beta$, and furthermore

$$\mathbb{E}[\widehat{\alpha}] = \mathbb{E}\left[\overline{y} - \widehat{\beta}\overline{x}\right] = \overline{y} - \beta\overline{x} = \alpha.$$

It is also easy from (6.3) to show that

$$\begin{aligned}\mathbb{V}\left[\widehat{\beta}\right] &= \mathbb{V}\left[\beta + \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \sum_{i=1}^{n} (x_i - \overline{x})\, \varepsilon_i\right] \\ &= \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^4} \sum_{i=1}^{n} (x_i - \overline{x})^2\, \mathbb{V}[\varepsilon_i] \\ &= \frac{\sigma^2}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^4} \sum_{i=1}^{n} (x_i - \overline{x})^2 \\ &= \frac{\sigma^2}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2},\end{aligned}$$

using (6.2). Now, since $\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x}$ by Proposition 6.1, we can write

$$\begin{aligned}\mathbb{V}\left[\widehat{\alpha}\right] &= \mathbb{V}\left[\overline{y} - \widehat{\beta}\overline{x}\right] \qquad (6.4) \\ &= \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} y_i - \widehat{\beta}\overline{x}\right] \\ &= \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] + \mathbb{V}\left[\widehat{\beta}\overline{x}\right] - \frac{2\overline{x}}{n}\sum_{i=1}^{n} \mathrm{Cov}\left(y_i, \widehat{\beta}\right) \\ &= \frac{\sigma^2}{n} + \overline{x}^2\mathbb{V}\left[\widehat{\beta}\right] - \frac{2\overline{x}}{n}\sum_{i=1}^{n} \mathrm{Cov}\left(y_i, \widehat{\beta}\right). \qquad (6.5)\end{aligned}$$

Now, for any $i = 1, \ldots, n$, we have, using (6.3),

$$\begin{aligned}
\operatorname{Cov}\left(y_i, \widehat{\beta}\right) &= \operatorname{Cov}\left(y_i, \beta + \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \sum_{k=1}^{n} (x_k - \overline{x})\, \varepsilon_k\right) \\
&= \operatorname{Cov}\left(y_i, \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \sum_{k=1}^{n} (x_k - \overline{x})\, \varepsilon_k\right) \\
&= \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \sum_{k=1}^{n} (x_k - \overline{x}) \operatorname{Cov}(y_i, \varepsilon_k) \\
&= \frac{\sigma^2}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} (x_i - \overline{x}),
\end{aligned}$$

since $\operatorname{Cov}(y_i, \varepsilon_k) = 0$ for all $k \neq i$ and is equal to $\sigma^2$ when $k = i$. Now,

$$\operatorname{Cov}\left(\overline{y}, \widehat{\beta}\right) = \frac{1}{n} \sum_{i=1}^{n} \operatorname{Cov}(y_i, \widehat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\sigma^2}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} (x_i - \overline{x}) = 0.$$

Therefore, from (6.4), we obtain the desired variance of $\widehat{\alpha}$. Finally,

$$\operatorname{Cov}\left(\widehat{\alpha}, \widehat{\beta}\right) = \operatorname{Cov}\left(\overline{y} - \widehat{\beta}\overline{x}, \widehat{\beta}\right) = \operatorname{Cov}\left(\overline{y}, \widehat{\beta}\right) - \overline{x}\mathbb{V}\left[\widehat{\beta}\right] = -\overline{x}\mathbb{V}\left[\widehat{\beta}\right] = -\frac{\overline{x}\sigma^2}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2}.$$

□

**Exercise 6.3.** *Suppose we observe a new value, say $x_{n+1}$, and we want to be able to predict the unknown value of $y_{n+1}$. A natural candidate is to consider $\widehat{y}_{n+1} = \widehat{\alpha} + \widehat{\beta}x_{n+1}$. Show that the forecasting error $\widehat{\varepsilon}_{n+1} := y_{n+1} - \widehat{y}_{n+1}$ satisfies*

$$\mathbb{E}\left[\widehat{\varepsilon}_{n+1}\right] = 0 \qquad \textit{and} \qquad \mathbb{V}\left[\widehat{\varepsilon}_{n+1}\right] = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2}\right).$$

We have so far assumed very little about the errors $\boldsymbol{\varepsilon}$. In order to refine the analysis above, we now assume that the sequence of errors $(\varepsilon_1, \ldots, \varepsilon_n)$ is i.i.d. Gaussian with constant variance $\sigma^2$. Then $\mathbf{Y} \sim \mathcal{N}(\alpha + \beta\mathbf{X}, \sigma^2)$.

We can compute the likelihood of the sample as

$$\mathcal{L}_n(\alpha, \beta, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2\right\}.$$

Maximising this likelihood is equivalent to minimising the negative log-likelihood

$$l_n(\alpha, \beta, \sigma^2) := -\frac{1}{n} \log \mathcal{L}_n(\alpha, \beta, \sigma^2) = \frac{1}{2} \log\left(2\pi\sigma^2\right) + \frac{\|\mathbf{Y} - \alpha\mathbb{1} - \beta\mathbf{X}\|_2^2}{2n\sigma^2}.$$

Minimising over $\alpha$ and $\beta$ is analogous to Proposition 6.1 and yields exactly the least-square estimates computed there:

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x} \qquad \text{and} \qquad \widehat{\beta} = \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \sum_{i=1}^{n} (x_i - \overline{x})\, y_i.$$

Now,

$$\partial_{\sigma^2} l_n(\widehat{\alpha}, \widehat{\beta}, \sigma^2) = \frac{1}{2\sigma^2} - \frac{\left\|\mathbf{Y} - \widehat{\alpha}\mathbb{1} - \widehat{\beta}\mathbf{X}\right\|_2^2}{2n\sigma^4} = \frac{1}{2\sigma^2} - \frac{\|\widehat{\boldsymbol{\varepsilon}}\|_2^2}{2n\sigma^4},$$

which is equal to zero if and only if

$$\sigma^2 = \widehat{\sigma}_L^2 := \frac{1}{n} \|\widehat{\boldsymbol{\varepsilon}}\|_2^2 .$$

**Remark 6.4.** *It can be shown (exercise) that*

$$\mathbb{E}\left[\widehat{\sigma}_L^2\right] = \mathbb{E}\left[\frac{\|\widehat{\boldsymbol{\varepsilon}}\|_2^2}{n}\right] = \frac{n-2}{n}\sigma^2,$$

*so that the maximum likelihood estimator of the variance is biased.*

**Exercise 6.5.** *Show that $\widehat{\alpha}$ and $\widehat{\beta}$ are Gaussian with means $\alpha$ and $\beta$ and variances*

$$\sigma_\alpha^2 := \frac{\sigma^2}{n\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \sum_{i=1}^{n} x_i^2 \qquad \textit{and} \qquad \sigma_\beta^2 := \frac{\sigma^2}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2}$$

*Furthermore, show that* $\dfrac{(n-2)}{\sigma^2}\widehat{\sigma}^2 \sim \chi^2_{n-2}$.

The variance $\sigma^2$ of the errors is actually unknown. Replacing it by its estimator, one can show that

$$\frac{\widehat{\alpha} - \alpha}{\widehat{\sigma}_\alpha} \sim t_{n-2} \quad \text{and} \quad \frac{\widehat{\beta} - \beta}{\widehat{\sigma}_\beta} \sim t_{n-2}$$

where $t_{n-2}$ is the student law with $n-2$ degrees of freedom. Hence, The intervals

$$\left[\widehat{\alpha} - t_{n-2}^{1-\eta/2}\widehat{\sigma}_\alpha, \widehat{\alpha} + t_{n-2}^{1-\eta/2}\widehat{\sigma}_\alpha\right] \qquad \text{and} \qquad \left[\widehat{\beta} - t_{n-2}^{1-\eta/2}\widehat{\sigma}_\beta, \widehat{\beta} + t_{n-2}^{1-\eta/2}\widehat{\sigma}_\beta\right]$$

are the confidence intervals for $\alpha$ and $\beta$ with level $1 - \frac{\eta}{2}$.

Regarding forecasting, the prediction error can be shown to satisfy

$$\frac{\varepsilon_{n+1}}{\widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2}}} \sim t_{n-2},$$

and the corresponding confidence interval for $y_{n+1}$ reads

$$\left[\widehat{y}_{n+1} - t_{n-2}^{1-\eta/2}\widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2}}, \widehat{y}_{n+1} + t_{n-2}^{1-\eta/2}\widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2}}\right].$$

## 6.2 The multivariate case

We now consider a multidimensional version of the previous linear regression, so that we are interested in the following problem:

$$\min_{f\in\mathcal{F}} \sum_{i=1}^{n} \mathcal{C}\left(y_i - f(x_i)\right),$$

for some given cost function $\mathcal{C}$, where $\mathcal{F}$ is a class of functions of interest. The main difference here is that, for each $i = 1, \dots, n$, $\mathrm{x}_i$ is a vector in $\mathbb{R}^p$, and the set $\mathcal{F}$ is composed of functions from $\mathbb{R}^p$ to $\mathbb{R}$. A multidimensional linear regression model is therefore a representation of the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{6.6}$$

where $\mathbf{Y} \in \mathbb{R}^n$ is the response/measured/endogenous variable, $\mathbf{X} \in \mathcal{M}_{n,p}(\mathbb{R})$ the exogenous or explanatory variable, $\boldsymbol{\beta} \in \mathbb{R}^p$ the vector of parameters to estimate, and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ the noise vector We shall always assume that the following conditions hold:

**Assumption 6.6.**

$$\mathrm{rank}(\mathbf{X}) = p \qquad \textit{and} \qquad \mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0} \qquad \textit{and} \qquad \mathbb{V}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}_n.$$

Again, the least-square estimator is defined as

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right\|_2^2.$$

We gather in the following proposition several results about $\widehat{\boldsymbol{\beta}}$.

**Proposition 6.7.** *The optimal least-square estimator is*

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y}.$$

*Furthermore*

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta} \qquad \textit{and} \qquad \mathbb{V}[\widehat{\boldsymbol{\beta}}] = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}.$$

Note that the matrix $\mathbf{X}^\top\mathbf{X}$ is invertible by Assumption 6.6.

*Proof.* The proof is a straightforward minimisation problem:

$$\begin{aligned}\nabla_{\boldsymbol{\beta}} \left\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right\|_2^2 &= \nabla_{\boldsymbol{\beta}}\Big(\boldsymbol{\beta}^\top \left(\mathbf{X}^\top\mathbf{X}\right)\boldsymbol{\beta} - 2\mathbf{Y}^\top\mathbf{X}\boldsymbol{\beta} + \left\|\mathbf{Y}\right\|_2^2\Big) \\ &= 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X} - 2\mathbf{Y}^\top\mathbf{X}.\end{aligned}$$

Since the matrix $\mathbf{X}^\top\mathbf{X}$ is invertible, the result for $\widehat{\boldsymbol{\beta}}$ follows directly. Now,

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}] = \mathbb{E}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}\right] = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbb{E}\left[\mathbf{Y}\right] = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbb{E}\left[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right] = \boldsymbol{\beta}.$$

Likewise,

$$\begin{aligned}\mathbb{V}[\widehat{\boldsymbol{\beta}}] &= \mathbb{V}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}\right] \\ &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbb{V}\left[\mathbf{Y}\right]\left\{\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\right\}^\top \\ &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbb{V}\left[\mathbf{Y}\right]\mathbf{X}\left\{\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right\}^\top \\ &= \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\left\{\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right\}^\top,\end{aligned}$$

since $\mathbb{V}[\mathbf{Y}] = \mathbb{V}[\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}] = \mathbb{V}[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}_n$, and the proposition follows. □

Next we show that the least-square estimator is optimal in the following sense:

**Theorem 6.8** (Gauss-Markov Theorem)**.** *Among all unbiased estimators of $\boldsymbol{\beta}$ linear in $\mathbf{Y}$, the least-square estimator $\widehat{\boldsymbol{\beta}}$ has minimal variance.*

*Proof.* Let $\widetilde{\boldsymbol{\beta}} = \mathbf{A}\mathbf{Y}$ be a linear estimator of $\boldsymbol{\beta}$, for some matrix $\mathbf{A} \in \mathcal{M}_{p,n}$. Being unbiased, it satisfies

$$\mathbb{E}\left[\widetilde{\boldsymbol{\beta}}\right] = \boldsymbol{\beta} = \mathbf{A}\mathbb{E}\left[\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\right] = \mathbf{A}\mathbf{X}\boldsymbol{\beta} + \mathbf{A}\mathbb{E}\left[\boldsymbol{\varepsilon}\right] = \mathbf{A}\mathbf{X}\boldsymbol{\beta} \implies \mathbf{A}\mathbf{X} = \mathbf{I}.$$

Furthermore

$$\mathbb{V}\left[\widetilde{\boldsymbol{\beta}}\right] = \mathbb{V}\left[\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\beta}}\right] = \mathbb{V}\left[\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right] + \mathbb{V}\left[\widehat{\boldsymbol{\beta}}\right] + \mathrm{Cov}\left(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}\right) + \mathrm{Cov}\left(\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right).$$

The covariance term reads

$$\begin{aligned}\mathrm{Cov}\left(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}\right) &= \mathrm{Cov}\left(\mathbf{A}\mathbf{Y} - \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}\right) \\ &= \mathrm{Cov}\left(\mathbf{A}\mathbf{Y}, \widehat{\boldsymbol{\beta}}\right) - \mathbb{V}\left[\widehat{\boldsymbol{\beta}}\right] \\ &= \mathrm{Cov}\left(\mathbf{A}\mathbf{Y}, \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}\right) - \mathbb{V}\left[\widehat{\boldsymbol{\beta}}\right] \\ &= \mathbf{A}\mathbb{V}\left[\mathbf{Y}\right]\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} - \mathbb{V}\left[\widehat{\boldsymbol{\beta}}\right] \\ &= \sigma^2\mathbf{A}\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} - \mathbb{V}\left[\widehat{\boldsymbol{\beta}}\right] = 0\end{aligned}$$

The exact same computation yields $\mathrm{Cov}\left(\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right) = 0$ and the theorem follows. □

From a geometric point of view, least-square minimisation is equivalent to determining the projection (in the Euclidean sense) of the random vector $\mathbf{Y}$ onto $\mathcal{M}_{\mathbf{X}}$, the subspace of $\mathbb{R}^n$ spanned by the column vectors of the matrix $\mathbf{X}$. More specifically,

$$\mathcal{M}_{\mathbf{X}} = \left\{\mathbf{X}\mathbf{w} = \sum_{i=1}^{p} w_i\mathbf{X}_i, \mathbf{w} = (w_1, \ldots, w_p) \in \mathbb{R}^p\right\},$$

where we denote by $\mathbf{X}_i$ the $i$-th column of $\mathbf{X}$. Since by assumption, $\mathbf{X}$ has rank $p$, then $\mathcal{M}_{\mathbf{X}}$ is of dimension $p$.

By Proposition 6.7, the projection reads $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} =: \mathbf{P}_{\mathbf{X}}\mathbf{Y}$, where

$$\mathbf{P}_{\mathbf{X}} := \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$$

is called the orthogonal projection matrix on $\mathcal{M}_{\mathbf{X}}$; as a projection matrix, it satisfies indeed $\mathbf{P}_{\mathbf{X}}^2 = \mathbf{P}_{\mathbf{X}}$. Now, the residuals of the least square estimation read

$$\widehat{\boldsymbol{\varepsilon}} := \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{P}_{\mathbf{X}}\mathbf{Y} = \left(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}}\right)\mathbf{Y} =: \mathbf{P}_{\mathbf{X}^\perp}\mathbf{Y} = \mathbf{P}_{\mathbf{X}^\perp}\boldsymbol{\varepsilon}.$$

The matrix $\mathbf{P}_{\mathbf{X}^\perp}$ is the orthogonal projection matrix onto $\mathcal{M}_{\mathbf{X}}^{\perp}$.

**Exercise 6.9.** *Show that the residuals are centered with* $\mathbb{V}\left[\widehat{\boldsymbol{\varepsilon}}\right] = \sigma^2\mathbf{P}_{\mathbf{X}^\perp}$ *and*

$$\mathbb{E}\left[\widehat{\mathbf{Y}}\right] = \mathbf{X}\boldsymbol{\beta}, \qquad \mathbb{V}\left[\widehat{\mathbf{Y}}\right] = \sigma^2\mathbf{P}_{\mathbf{X}}, \qquad \mathrm{Cov}\left(\widehat{\boldsymbol{\varepsilon}}, \widehat{\mathbf{Y}}\right) = 0.$$

**Exercise 6.10.** *Show that* $\widehat{\sigma}^2 := \dfrac{\|\widehat{\boldsymbol{\varepsilon}}\|_2^2}{n-p}$ *is an unbiased estimator of the variance* $\sigma^2$.

**Solution 6.11.** *First, note that*

$$\begin{aligned}\mathbb{E}\left[\|\widehat{\boldsymbol{\varepsilon}}\|_2^2\right] &= \mathbb{E}\left[\mathrm{Tr}\left[\|\widehat{\boldsymbol{\varepsilon}}\|_2^2\right]\right] = \mathbb{E}\left[\mathrm{Tr}\left[\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}\right]\right] = \mathbb{E}\left[\mathrm{Tr}\left[\widehat{\boldsymbol{\varepsilon}}\,\widehat{\boldsymbol{\varepsilon}}^\top\right]\right] \\ &= \mathrm{Tr}\left[\mathbb{E}\left[\widehat{\boldsymbol{\varepsilon}}\,\widehat{\boldsymbol{\varepsilon}}^\top\right]\right] = \mathrm{Tr}\left[\mathbb{V}\left[\widehat{\boldsymbol{\varepsilon}}\right]\right] = \mathrm{Tr}\left[\sigma^2\mathbf{P}_{\mathbf{X}^\perp}\right] = \sigma^2\mathrm{Tr}\left[\mathbf{P}_{\mathbf{X}^\perp}\right]\end{aligned}$$

*Finally, since*

$$\mathrm{Tr}\left(\mathbf{P}_{\mathbf{X}^\perp}\right) = \mathrm{Tr}\left(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}}\right) = \mathrm{Tr}\left(\mathbf{I}_n\right) - \mathrm{Tr}\left(\mathbf{P}_{\mathbf{X}}\right) = n - p,$$

*the proposition follows.*

**Exercise 6.12.** *Consider a new observation vector* $\mathrm{x}_{n+1}$*, with new response variable* $y_{n+1} = \mathrm{x}_{n+1}^\top\boldsymbol{\beta} + \varepsilon_{n+1}$*, with* $\mathbb{E}[\varepsilon_{n+1}] = 0$*,* $\mathbb{V}[\varepsilon_{n+1}] = \sigma^2$ *and* $\mathrm{Cov}(\varepsilon_{n+1}, \varepsilon_i) = 0$ *for any* $i = 1, \ldots, n$*. Define the new prediction as* $\widehat{y}_{n+1} := \mathrm{x}_{n+1}\widehat{\boldsymbol{\beta}}$*, and the prediction error* $\widehat{\varepsilon}_{n+1} := y_{n+1} - \widehat{y}_{n+1}$*. Show that the following identities hold*

$$\mathbb{E}\left[\widehat{\varepsilon}_{n+1}\right] = 0 \qquad \textit{and} \qquad \mathbb{V}\left[\widehat{\varepsilon}_{n+1}\right] = \sigma^2\left(1 + \mathrm{x}_{n+1}^\top\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathrm{x}_{n+1}\right).$$

**Solution 6.13.** *Note first that*

$$\widehat{\varepsilon}_{n+1} := y_{n+1} - \widehat{y}_{n+1} = \mathrm{x}_{n+1}^\top\boldsymbol{\beta} + \varepsilon_{n+1} - \mathrm{x}_{n+1}^\top\widehat{\boldsymbol{\beta}} = \mathrm{x}_{n+1}^\top\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right) + \varepsilon_{n+1},$$

*so that, since* $\varepsilon_{n+1}$ *is centered and* $\widehat{\boldsymbol{\beta}}$ *is an unbiased estimator of* $\boldsymbol{\beta}$*, we can write*

$$\mathbb{E}\left[\widehat{\varepsilon}_{n+1}\right] = \mathbb{E}\left[\mathrm{x}_{n+1}^\top\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right) + \varepsilon_{n+1}\right] = \mathrm{x}_{n+1}^\top\left(\boldsymbol{\beta} - \mathbb{E}\left[\widehat{\boldsymbol{\beta}}\right]\right) = 0.$$

*Now,*

$$\mathbb{V}\left[\widehat{\varepsilon}_{n+1}\right] = \mathbb{V}\left[\mathrm{x}_{n+1}^\top\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right) + \varepsilon_{n+1}\right] = \mathbb{V}\left[\mathrm{x}_{n+1}^\top\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)\right] + \mathbb{V}\left[\varepsilon_{n+1}\right] = \mathrm{x}_{n+1}^\top\mathbb{V}\left[\widehat{\boldsymbol{\beta}}\right]\mathrm{x}_{n+1} + \sigma^2$$

*since* $\varepsilon_{n+1}$ *is uncorrelated with* $(\varepsilon_i)_{i=1,\ldots,n}$ *and* $\widehat{\boldsymbol{\beta}}$ *only depends on the latter sequence. The result follows using Proposition 6.7.*

# 7 Dimensionality reduction

This section examines two techniques for dimensionality reduction, namely principal component analysis (PCA) and the Johnson—Lindenstrauss lemma.

## 7.1 Principal Component Analysis

Given datapoints $x_1, \ldots, x_m \in \mathbb{R}^d$, it is intuitive that the dimension $d$ could be reduced at least to $m$, since any $m$ points live in a linear space of dimension at most $m$. In fact, it is often realistic to believe that these datapoints live close to a linear space of even smaller dimension $k < m$. The best $k$-dimensional space, in the sense of squared Euclidean distances, is obtained from the singular value decomposition of the data matrix $X = [x_1 \ldots x_m] \in \mathbb{R}^{d \times m}$. The result is stated formally below.

**Theorem 7.1.** *A $k$-dimensional linear subspace $W$ of $\mathbb{R}^d$ that minimizes*

$$\sum_{i=1}^{m} dist_{\ell_2}(x_i, W)^2$$

*is given by span$\{u_1, \ldots, u_k\}$, where $u_1, \ldots, u_m \in \mathbb{R}^d$ are the left singular vectors appearing in the singular value decomposition $X = \sum_{j=1}^{m} \sigma_j u_j v_j^\top$.*

*Proof.* Let $W$ be a $k$-dimensional subspace of $\mathbb{R}^d$ with orthonormal basis $(w_1, \ldots, w_k)$, say. Let $P_W$ be the orthogonal projection matrix from $\mathbb{R}^d$ onto $W$, so that $P_W x = \sum_{j=1}^{k} \langle w_j, x \rangle w_j$ for any $x \in \mathbb{R}^d$. In view of $\mathrm{dist}_{\ell_2}(x, W) = \|x - P_W x\|_2$, the objective function takes the form

$$\sum_{i=1}^{m} \|x_i - P_W x_i\|_2^2 = \|X - P_W X\|_F^2.$$

Since $P_W X = (\sum_{j=1}^{k} w_j w_j^\top) X$ has rank at most $k$ and since, by the Eckart–Young-Mirsky Theorem, the best rank-$k$ approximation of $X$ in Frobenius norm is obtained by truncating the singular value decomposition, one has

$$\|X - P_W X\|_F^2 \geq \|X - \sum_{j=1}^{k} \sigma_j u_j v_j^\top\|_F^2.$$

Summing over $j \in \{1, ..., k\}$ the identity $\sigma_j u_j v_j^\top = u_j u_j^\top \sum_{\ell=1}^{m} \sigma_\ell u_\ell v_\ell^\top = u_j u_j^\top X$, one arrives at $\sum_{j=1}^{k} \sigma_j u_j v_j^\top = P_U X$, where $U := \mathrm{span}\{u_1, \ldots, u_k\}$. Therefore, one concludes that $\|X - P_W X\|_F^2 \geq \|X - P_U X\|_F^2$, i.e., that $U$ is a $k$-dimensional subspace minimizing the objective function under consideration. □

**Remark 7.2.** *The (real) eigenvalues $\lambda_1(A), \ldots, \lambda_n(A)$ of a symmetric matrix $A \in \mathbb{R}^{n \times n}$ are traditionally ordered in a nonincreasing fashion, so that*

$$\lambda_1(A) \geq \lambda_2(A) \geq \ldots \geq \lambda_n(A).$$

*Each of these eigenvalues admits the minimax characterizations*

$$\lambda_i(A) = \max_{\substack{V\subseteq\mathbb{R}^n \\ dim(V)=i}} \min_{\substack{x\in V \\ \|x\|_2=1}} \langle Ax, x\rangle = \min_{\substack{V\subseteq\mathbb{R}^n \\ dim(V)=n-i+1}} \max_{\substack{x\in V \\ \|x\|_2=1}} \langle Ax, x\rangle, \tag{7.1}$$

*often referred to as the Courant–Fischer theorem. This is a special case of the more general Wielandt minimax principle.*

**Theorem 7.3** (Wielandt minimax principle)**.** *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Given $k \in \{1, ..., n\}$ and indices $i_1 < \ldots < i_k \in \{1, ..., n\}$, one has*

$$\sum_{j=1}^{k} \lambda_{i_j}(A) = \max_{\substack{V_1\subseteq\ldots\subseteq V_k\subseteq\mathbb{R}^n \\ dim(V_j)=i_j}} \min_{\substack{(x_1,\ldots,x_k)\ orthonormal \\ x_1\in V_1,\ldots,x_k\in V_k}} \sum_{j=1}^{k} \langle Ax_j, x_j\rangle.$$

*Proof.* To establish the $\leq$-part, one needs to find subspaces $V_1 \subseteq \cdots \subseteq V_k$ of dimensions $i_1, \ldots, i_k$, respectively, such that

$$\sum_{j=1}^{k} \lambda_{i_j}(A) \leq \sum_{j=1}^{k} \langle Ax_j, x_j\rangle$$

whenever $(x_1, \ldots, x_k)$ is an orthonormal system with $x_1 \in V_1, \ldots, x_k \in V_k$.

Picking an orthonormal basis $(v_1, \ldots, v_n)$ of eigenvectors for the eigenvalues $\lambda_1(A) \geq \cdots \geq \lambda_n(A)$, it suffices to take $V_j = \mathrm{span}\{v_1, \ldots, v_{i_j}\}$, for $j \in \{1, ..., k\}$. Indeed, writing $x_j \in V_j$ as $x_j = \sum_{\ell=1}^{i_j} c_{j\ell} v_\ell$, one has $Ax_j = \sum_{\ell=1}^{i_j} c_{j\ell}\lambda_\ell(A) v_\ell$ and $\sum_{\ell=1}^{i_j} c_{j\ell}^2 = \|x_j\|^2 = 1$, so that $\langle Ax_j, x_j\rangle = \sum_{\ell=1}^{i_j} c_{j\ell}^2 \lambda_\ell(A) \geq \lambda_{i_j}(A)$. Summing over $j \in \{1, ..., \}$ gives the required inequality.

For the $\geq$-part, one needs to show that, for all subspaces $V_1 \subseteq \cdots \subseteq V_k$ of dimensions $i_1, \ldots, i_k$, respectively, there is an orthonormal system $(x_1, ..., x_k)$ with $x_1 \in V_1$, ..., $x_k \in V_k$ such that $\sum_{j=1}^{k} \lambda_{i_j}(A) \geq \sum_{j=1}^{k} \langle Ax_j, x_j\rangle$. With $(v_1, \ldots, v_n)$ still denoting an orthonormal basis of eigenvectors for the eigenvalues $\lambda_1(A) \geq \ldots \geq \lambda_n(A)$, one considers the subspaces $W_1, \ldots, W_k$ defined by $W_j = \mathrm{span}\{v_{i_j}, \ldots, v_n\}$, $j \in \{1, ..., k\}$. By Gram-Schmidt, one can find orthonormal systems $(x_1, \ldots, x_k)$ and $(y_1, ..., y_k)$ spanning the same space $S$ in such a way that $x_j \in V_j$ and $x_j \in W_j$ for all $j \in \{1, ..., k\}$. Writing $y_j = \sum_{\ell=i_j}^{n} d_{j\ell} v_\ell$, one has $Ay_j = \sum_{\ell=i_j}^{n} d_{j\ell}\lambda_\ell(A) v_\ell$ and $\sum_{\ell=i_j}^{n} d_{j\ell}^2 = \|y_j\|^2 = 1$, so that $\langle Ay_j, y_j\rangle = \sum_{\ell=i_j}^{n} d_{j\ell}^2 \lambda_\ell(A) \leq \lambda_{i_j}(A)$. Summing over $j \in \{1, ..., k\}$ leads to $\sum_{j=1}^{k} \langle Ay_j, y_j\rangle \leq \sum_{j=1}^{k} \lambda_j(A)$, which in turn implies the required inequality $\sum_{j=1}^{k} \langle Ax_j, x_j\rangle \leq \sum_{j=1}^{k} \lambda_{i_j}(A)$ because $\sum_{j=1}^{k} \langle Ax_j, x_j\rangle$ and $\sum_{j=1}^{k} \langle Ay_j, y_j\rangle$ are both equal to the trace of the compression of $A$ to $S$.

□

**Exercise 7.4.** *Use Wielandt minimax principle to prove the Lidskii inequality.*

Substituting each datapoint $x_i$, living in the high-dimensional space $\mathbb{R}^d$, by its orthogonal projection $\hat{x}_i = P_U(x_i)$, living in the lower-dimensional space $U = \mathrm{span}\{u_1, \ldots, u_k\}$, constitutes a reasonable dimension reduction procedure. The method is commonly called principal component analysis, abbreviated as PCA, due to a statistical interpretation to be elucidated now. For this interpretation, consider a random vector $x \in \mathbb{R}^d$ whose distribution generated the datapoints $x_1, \ldots, x_m$. Its covariance matrix is the symmetric matrix

$$\mathrm{cov}(x) = \mathbb{E}\left[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^\top\right] \in \mathbb{R}^{d\times d}.$$

Making the assumption that the distribution has mean zero, this covariance matrix reduces to $\mathrm{cov}(x) = \mathbb{E}[xx^\top]$. One would like to find the first principal component $u_1$, understood as the direction presenting the maximal variance. Note that the variance of the random vector $x \in \mathbb{R}^d$ projected on a direction $u \in \mathbb{R}^d$ is,

$$\mathbb{V}(\langle x, u\rangle) = \mathbb{E}[\langle x, u\rangle^2] = \mathbb{E}[u^\top xx^\top u] = u^\top \mathrm{cov}(x)u.$$

Hence, by the Courant–Fischer theorem taking $i = 1$ in (7.1) and $A = \mathrm{cov}(x)$, the direction $u_1$ that maximises the variance, called the first principal component, is the leading eigenvector of $\mathrm{cov}(x)$. Next, one would like to find the second principal component $u_2$ orthogonal to $u_1$ presenting the maximal variance. By virtue of

$$\begin{aligned}\max_{\substack{u\perp u_1\\ \|u\|_2=1}} u^\top \mathrm{cov}(x)u &= \max_{\substack{u\perp u_1\\ \|u\|_2=1}} u^\top \left(\mathrm{cov}(x) - \lambda_1(\mathrm{cov}(x))u_1u_1^\top\right) u\\ &\leq \max_{\|u\|_2=1} u^\top \left(\mathrm{cov}(x) - \lambda_1(\mathrm{cov}(x))u_1u_1^\top\right) u\\ &= \lambda_2(\mathrm{cov}(x)),\end{aligned}$$

one easily derives that the second principal component $u_2$ is the second leading eigenvector of $\mathrm{cov}(x)$. In general, the $j$-th principal component $u_j$ is the $j$-th leading eigenvector of $\mathrm{cov}(x)$. Now, since one does not have access to $\mathrm{cov}(x)$, the latter is replaced by the empirical covariance matrix

$$\widehat{C} = \frac{1}{m}\sum_{i=1}^m x_i {x_i}^\top = \frac{1}{m}XX^\top \in \mathbb{R}^{d\times d}.$$

The true principal components are then replaced by the eigenvectors of $XX^\top$, i.e., the left singular vectors of $X$ that appeared in Theorem 7.1.

**Remark 8.2** The above consideration could wrongly suggest obtaining the principal components by performing an eigendecomposition on the matrix $XX^\top \in \mathbb{R}^{d\times d}$. This is not advised if $d \gg m$. Instead, in view of the large size of $d$, it is a better idea, in terms of computational cost, to perform an eigendecomposition on the matrix $X^\top X \in \mathbb{R}^{m\times m}$. The eigenvectors of the latter are the right singular vectors $v_j$ of $X$, while the eigenvectors of the former are the left singular vectors $u_j$ of $X$—the principal components—which can simply be deduced from the $v_j$ via $Xv_j = \sigma_j u_j$.

## 7.2 Johnson-Lindestrauss Lemma

It turns out that often generating substitutes $\hat{x}_1, ..., \hat{x}_m$ for the datapoints $x_1, ..., x_m$ from a low-dimensional linear space $W$ can be carried out by selecting $W$ at random. The so called Johnson-Lindestrauss Lemma below states that it is possible to find a random subspace $W$ such as its dimension is independent of $d$, and scales only logarithmically in the number of observations $m$ and quadratically in the inverse of the desired distortion $\rho$. This result is deduced from the following concentration inequality.

**Lemma 7.5.** *Let $x \in \mathbb{R}^d$ and $\delta \in (0,1)$. If $A \in \mathbb{R}^{k\times d}$ is a random matrix whose entries are independent mean-zero Gaussian variables with variance $1/k$, then*

$$\mathbb{P}\left(\left|\|Ax\|_2^2 - \|x\|_2^2\right| > \delta \|x\|_2^2\right) \leq 2\exp\left(-\left(\frac{\delta^2}{4} - \frac{\delta^3}{6}\right)k\right). \tag{7.2}$$

*Proof.* For any $i \in \{1, ..., k\}$, recalling that a linear combination of independent Gaussian random variables is still Gaussian, one can write

$$(Ax)_i = \sum_{j=1}^{d} A_{ij}x_j = \frac{||x||_2}{\sqrt{k}} g_i,$$

where $g_i$ is a standard normal. Then, since $\|Ax\|_2^2 = \sum_{i=1}^k (Ax)_i^2$, one has

$$\begin{aligned}\mathbb{P}\left(\|Ax\|_2^2 > (1+\delta)\|x\|_2^2\right) &= \mathbb{P}\left(\sum_{i=1}^{k} g_i^2 \geq k(1+\delta)\right) \\ &= \mathbb{P}\left(\exp\left(u\sum_{i=1}^{k} g_i^2\right) > \exp(uk(1+\delta))\right)\end{aligned}$$

for any $u > 0$ to be chosen later. Using the Markov inequality, the independence of the random variables $g_1, \ldots, g_k$, and the expression for the moment generating function $t \mapsto \mathbb{E}[\exp(tg^2)]$ of a squared standard Gaussian random variable, one obtains

$$\begin{aligned}\mathbb{P}\left(\|Ax\|_2^2 > (1+\delta)\|x\|_2^2\right) &\leq \frac{\mathbb{E}\left[\exp\left(u\sum_{i=1}^k g_i^2\right)\right]}{\exp(uk(1+\delta))} \\ &= \frac{\mathbb{E}\left[\prod_{i=1}^k \exp\left(ug_i^2\right)\right]}{\prod_{i=1}^k \exp(u(1+\delta))} \\ &= \prod_{i=1}^{k} \frac{\mathbb{E}\left[\exp\left(ug_i^2\right)\right]}{\exp(u(1+\delta))} \\ &= \left(\frac{1/\sqrt{1-2u}}{\exp(u(1+\delta))}\right)^k\end{aligned}$$

Making the (optimal) choice $u = \delta/(2(1+\delta))$, for which $1 - 2u = 1/(1+\delta)$ and $\exp(u(1+\delta)) = \exp(\delta/2)$, it follows that

$$\begin{aligned}\mathbb{P}\left(\|Ax\|_2^2 > (1+\delta)\|x\|_2^2\right) &\leq \left(\frac{1+\delta}{\exp(\delta)}\right)^{k/2} \\ &\leq \exp\left(-\frac{\delta^2}{2} + \frac{\delta^3}{3}\right)^{k/2} \\ &= \exp\left(-\left(\frac{\delta^2}{4} - \frac{\delta^3}{6}\right)k\right)\end{aligned}$$

where the inequality $\ln(1+\delta) \leq \delta - \delta^2/2 + \delta^3/3$ was used. A similar estimate for $\mathbb{P}\left(\|Ax\|_2^2 < (1-\delta)\|x\|_2^2\right)$ is derived in exactly the same way leading to (7.2) in view of the fact that $|\,\|Ax\|_2^2 - \|x\|_2^2\,| > \delta||x||_2^2$ occurs either when $||Ax||_2^2 > (1+\delta)||x||_2^2$ or when $||Ax||_2^2 < (1-\delta)||x||_2^2$. □

With the concentration inequality (7.2) in place, dimension reduction in the sense of Johnson–Lindenstrauss is now achieved by taking each substitute $\hat{x}_i = Ax_i$ to live in the $k$-dimensional range of a random matrix $A \in \mathbb{R}^{k\times d}$.

**Theorem 7.6** (Johnson-Lindestrauss)**.** *Let $x_1, \ldots, x_m \in \mathbb{R}^d$ and $\delta \in (0, 1/2)$, say. If*

$$k \geq \frac{18\ln(m)}{\delta^2},$$

*and if $A \in \mathbb{R}^{k\times d}$ is a random matrix whose entries are independent mean-zero Gaussian variables with variance $1/k$, then*

$$(1-\delta)\|x_i - x_j\|_2^2 \leq \|Ax_i - Ax_j\|_2^2 \leq (1+\delta)\|x_i - x_j\|_2^2$$

*holds for all $i \neq j$ in $\{1, ..., m\}$ simultaneously with probability at least $1 - \frac{1}{m}$.*

*Proof.* Fixing $i \neq j$ in $\{1, ..., m\}$ and setting $x_{(i,j)} := x_i - x_j$, the previous Lemma ensures that

$$\mathbb{P}\left(\left|\|Ax_{(i,j)}\|_2^2 - \|x_{(i,j)}\|_2^2\right| > \delta\|x_{(i,j)}\|_2^2\right) \leq 2\exp\left(-\frac{\delta^2 k}{6}\right)$$

where the fact that $\delta^3/6 \leq \delta^2/12$ when $\delta \leq 1/2$ was used. Now, unfixing the pair $(i, j)$, one deduces by way of a union bound that

$$\begin{aligned}&\mathbb{P}\left(\left|\|Ax_{(i,j)}\|_2^2 - \|x_{(i,j)}\|_2^2\right| > \delta\|x_{(i,j)}\|_2^2 \text{ for some } i \neq j \in \{1, ..., m\}\right) \\ &\qquad \leq \binom{m}{2} 2\exp\left(-\frac{\delta^2 k}{6}\right) \\ &\qquad \leq m^2 \exp\left(-3\ln(m)\right) = \frac{1}{m}.\end{aligned}$$

The latter means that with probability at least $1 - \frac{1}{m}$, one has

$$\left| \|A(x_i - x_j)\|_2^2 - \|x_i - x_j\|^2 \right| \leq \delta \|x_i - x_j\|_2^2$$

for all $i \neq j$ in $\{1, ..., m\}$ simultaneously, which is the desired conclusion. $\square$

# References

[Bil13] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[HJ12] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.