Lecture notes on Statistics

Lecture notes on Statistics

Stéphane Gaïffas*

January 18, 2021

*

Stéphane Gaïffas

Contents

Statistical models 1

Let us start with the most classical and simplest statistical experiment: the coin toss. We toss a coin n times, and we model each toss by a random variable in $\{0,1\}$, where we decide that 1 means that the toss ended up with heads (so that 0 means tails). To each toss is associated a random variable, leading to random variables $X_1,\dots,X_n$ valued in $\{0,1\}$, where X_i encodes the outcome of the i-th toss. Each X_i has distribution $\mathrm{Bernoulli}(p)$ for $p\in[0,1]$, where p corresponds to the probability that a coin toss gives heads, namely $\mathbb{P}[X_i=1]$.[1] We assume that the X_i are *independent* (since the outcome of the tosses are physically unrelated), and since we are tossing the same coin each time, we assume that these outcomes have the same distribution (meaning that p is constant along the tosses). Therefore, we assume that $X_1,\dots,X_n$ are *iid*.[2]

1: The notation $\mathrm{Bernoulli}(p)$ corresponds to the *Bernoulli* distribution: we will write $X\sim\mathrm{Bernoulli}(p)$ whenever $X\in\{0,1\}$ and $\mathbb{P}[X=1]=p=1-\mathbb{P}[X=0]$. Another way to obtain a Bernoulli distribution is by setting $X_i=\mathbf{1}_{Y_i\in A}$ where $Y_1,\dots,Y_n$ are random variables valued in a probability space $(E,\mathcal{E})$, with $A\in\mathcal{E}$, so that $X_i\sim\mathrm{Bernoulli}(p)$ with $p=\mathbb{P}[Y_i\in A]$.

2: From now on, iid will stand for *independent and identically distributed*. More about this fundamental assumption will follow.

1.1 Probabilities and statistics

Since we assume that the reader is familiar with probability theory, we start this chapter with a comparison between what we do in probabilities and what we do in statistics for the $\mathrm{Bernoulli}(p)$ model described above.

Probabilities. In the field of probabilities, we suppose that $p\in(0,1)$ is known, and we study the properties of the sequence $(X_i)_{i\geq1}$. For instance, we know that the distribution of $S_n=\sum_{i=1}^n X_i$ is $\mathrm{Binomial}(n,p)$, namely that $\mathbb{P}[S_n=k]=\binom{n}{k}p^k(1-p)^{n-k}$ for $k\in\{0,\dots,n\}$.[3] It is easy to see that $\mathbb{E}[S_n]=np$ and that $\mathbb{V}[S_n]=np(1-p)$, where $\mathbb{E}[\cdot]$ and $\mathbb{V}[\cdot]$ stand respectively for the expectation and the variance.[4] We can study the asymptotic properties of S_n: the law of large number tells us that

$$\frac{S_n}{n}\overset{\text{as}}{\to}p$$

as $n\to+\infty$ and the central limit theorem tells us by how much we need to normalize $\frac{S_n}{n}-p$ in order to obtain a non-zero limit

$$\sqrt{n}\Big(\frac{S_n}{n}-p\Big)\rightsquigarrow\mathrm{Normal}(0,p(1-p))$$

as $n\to+\infty$,[5] where $\mathrm{Normal}(\mu,\sigma^2)$ stands for the Gaussian dis-

3: Where $\binom{n}{k}$ is the (n,k) binomial coefficient given by $\frac{n!}{k!(n-k)!}$.

4: The linearity of the expectation gives $\mathbb{E}[S_n]=n\mathbb{E}[X_1]=np$ since the X_i are identically $\mathrm{Bernoulli}(p)$ distributed, and, since the X_i are iid, we know that $\mathbb{V}[S_n]=n\mathbb{V}[X_1]=np(1-p)$.

5: The notation $X_n\overset{\text{as}}{\to}X$ stands for the *almost sure* convergence of X_n towards X while $X_n\rightsquigarrow X$ stands for the convergence of X_n towards X in distribution.

tribution with expectation μ and variance σ^2. In the field of probabilities, the object of interest would be the *random variable* S_n, that we study knowing the value of p. In particular, if we replace the $\mathrm{Bernoulli}(p)$ distribution by the *Rademacher* distribution where $\mathbb{P}[X_i = 1] = 1 - \mathbb{P}[X_i = -1] = p$, the random variable S_n becomes a *random walk* for which many things can be said, depending on the value of p.[6]

6: But such things are way beyond the topic of this book, let us just cite [1] as a reference on the study of the random walk and its importance in the field of probability theory.

Statistics. In statistics, for the $\mathrm{Bernoulli}(p)$ example, we don't really care about S_n, but we do care about p. We assume that p is *unknown*, and we want to find out things about it. This objective is called *statistical inference* of the *parameter* p. For instance, we would like to know if $p = 1/2$ or not, in order to find out if the coin is well-balanced and not rigged. The random variables $X_1, \dots, X_n$ (and S_n) live on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$, but we don't really care about it either in statistics. We will always assume, in statistics, that each *observed* outcome $x_i \in \{0, 1\}$ of a coin toss is a *realization* of the random variable X_i, namely that

$$x_i = X_i(\omega)$$

for some *event* $\omega \in \Omega$. The realizations x_i are also called *data* or *samples* or *observations*. But, actually, we will also refer to the random variables X_i in the same way, as *data*, *samples* or *observations*, since we won't manipulate the x_i mathematically,[7] while we will work a lot with the random variables $X_1, \dots, X_n$. In statistics, we can do whatever we want with $X_1, \dots, X_n$ in order to say things about p, but we will *never* assume p to be known.[8] We will construct measurable functions of $(X_1, \dots, X_n)$ that do not depend on p, these are called *statistics*, in order to tell things about the unknown parameter p. The object of interest in the field of statistics is, therefore, the *distribution of the observations* rather than the observations themselves.

7: nothing can be done with them... it's just deterministic zeros and ones

8: The parameter p will quickly become a mathematical variable that we will use in equations, in order to perform calculus for instance. Therefore, we will use the specific notation p_0 for the ground truth parameter, namely $X_1 \sim \mathrm{Bernoulli}(p_0)$, while p will be used as a generic parameter. A statistical parameter will usually be denoted as θ, while the ground truth parameter will be denoted as θ_0 when necessary.

1.2 Statistical models and experiences

Let us consider another very classical problem: the election poll problem, where a population of size N vote for one of two candidates A and B. There are N_A people voting for A while $N - N_A$ vote for B, and we want to know about $\theta_0 = N_A/N$. We perform of poll including $n \ll N$ voters and obtain observations $x_1, \dots, x_n \in \{0, 1\}$, where $x_i = 1$ (resp. $x_i = 0$) means that voter i votes for A (resp. B). In this problem, both N_A and N are so large that we can suppose that

$$(x_1, \dots, x_n) = (X_1(\omega), \dots, X_n(\omega))$$

for some $\omega \in \Omega$, where all $X_i : (\Omega, \mathcal{A}, \mathbb{P}) \to (\{0, 1\}, \mathcal{P}(\{0, 1\})$ for $i = 1, \dots, , n$ are such that $X_i \sim \mathrm{Bernoulli}(\theta_0)$. Let's look a little bit

For a finite set E, the notation $\mathcal{P}(E)$ stands for the σ-algebra corresponding to the set of all the parts of E.

at all these mathematical objects. In statistics, we are mainly only interested by the fact that the observations are valued in $(\{0,1\}, \mathcal{P}(\{0,1\}))$ and that the distribution is $\mathbb{P}_{X_i} = \mathbb{P} \circ X_i^{-1} = \text{Bernoulli}(\theta_0)$, which is fully described by its parameter $\theta_0 \in (0,1)$. Once again, we don't really care about $(\Omega, \mathcal{A}, \mathbb{P})$.

Statistical model. The *statistical model* for $X = (X_1, \dots, X_n) \in \{0,1\}^n$ is the family of distributions

$$\big\{\mathbb{P}_\theta^{\otimes n} : \theta \in (0,1)\big\} = \big\{\text{Bernoulli}(\theta)^{\otimes n} : \theta \in (0,1)\big\},$$

which is a family indexed by $\theta \in (0,1)$. The notation $\mathbb{P}^{\otimes n} = \mathbb{P} \otimes \cdots \otimes \mathbb{P}$ stands for the tensor product, namely $\mathbb{P}^{\otimes n}[A_1 \times \cdots \times A_n] = \prod_{i=1}^n \mathbb{P}[A_i]$ for any $A_i \in \mathcal{A}$, $i = 1, \dots, n$.[9] When we say that this is a statistical model for X, we assume that there exist $\theta_0 \in (0,1)$ such that $X \sim \text{Bernoulli}(\theta_0)$. Once again, let us insist on the following: we do whatever we want with $X_1, \dots, X_n$ but never with θ_0, which is the unknown parameter.

9: In this book, we will quickly forget to write $\mathbb{P}^{\otimes n}$ and will write simply $\mathbb{P}$ when computations are clear enough, to avoid overloaded notations.

Another (naive) example. Let us consider the problem of the quality of production of screws. The dimensions of the screws must satisfy some strong constraints, for instance their length must match quite accurately some fixed size.[10]

10: The author of this book does not know anything about screws.

Millions of screws come out of the production chain, and we can't test all of them. Therefore, we need to assess the production quality by selecting at random a small set of n screws, and we measure their lengths $x_1, \dots, x_n$. Since these lengths are highly concentrated around the theoretical desired length μ, and since production errors are usually small, we decide to choose a Gaussian model: we assume that $x_i = X_i(\omega)$ for $X_i \sim \text{Normal}(\mu, \sigma^2)$, where σ^2 corresponds to a variance coming from the (hopefully) small production errors.

Figure 1.1: I can't resist the temptation of showing you a screw, so here it is.

A model is a simplification of the reality. For this example, we make the following modeling assumptions.

Distribution choice. We choose the distribution $\text{Normal}(\mu, \sigma^2)$ for the lengths. So, we implicitly assume that the true underlying distribution of the lengths is *symmetrical* and highly *concentrated* around μ. This may or may not be realistic.[11]

The iid assumption. We will assume also that X_i are iid. What this means in practice is that we need to very careful in the way we select the screws coming out of the production lines: for instance, we should pick screws all along a week or a month, at different times, and not all from the same production line, in order to avoid time and machine biases.

11: A real random variable X is *symmetrical* whenever $\mathbb{P}_X = \mathbb{P}_{-X}$. This means that $\mathbb{P}[X \le -x] = \mathbb{P}[X \ge x]$ so that $F(-x) = 1 - F(x_-)$ if F is the distribution function of X. Also, if X has a density f with respect to the Lebesgue measure, then f is an even function.

Once again, a statistical model is always a simplification and an approximation of the truth. By *truth* we mean the true distribution $\mathbb{P}_{(X_1,\dots,X_n)}$ of $(X_1,\dots,X_n)$. For instance, the $\mathrm{Normal}(0,1)$ distribution has density $\phi(x) = e^{-x^2/2}/\sqrt{2\pi}$ supported on $\mathbb{R}$. This means that realizations of a $\mathrm{Normal}(0,1)$ distribution can take any value in $\mathbb{R}$, while real samples are usually bounded. However, we can prove that if $Z \sim \mathrm{Normal}(0,1)$, then $\Phi(x) := \mathbb{P}[Z \le x]$ satisfies

$$\Big(\frac{1}{x} - \frac{1}{x^3}\Big)\frac{1}{\sqrt{2\pi}}e^{-x^2/2} \le 1 - \Phi(x) \le \frac{1}{x\sqrt{2\pi}}e^{-x^2/2} \qquad (1.1)$$

for any $x \ge 1$, which means that the *queue*[12] of the Normal distribution is very tight. For $x = 6$ for instance, we have $\mathbb{P}(Z > x) \le 10^{-9}$: we will actually never see realizations of $\mathrm{Normal}(0,1)$ outside of $[-6,6]$, and rarely outside of $[-3,3]$. A proof of (1.1) is given in Section 1.6 below.

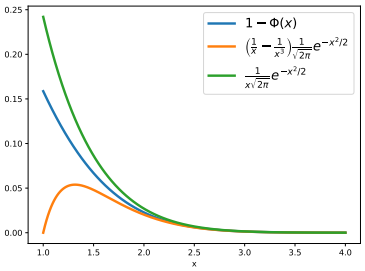


Figure 1.2: Illustration of the lower and upper bounds proposed in Equation (1.1).

12: The *queue* of a real random variable Z is the function $x \mapsto \mathbb{P}(Z > x) =: 1 - F_Z(x)$. Whenever $Z \in [0, +\infty)$ almost surely, this function is called also the *survival* function.

Definition 1.1 A *statistical experiment* consists of the following two things:

- A *random object* X valued in a measurable space $(E, \mathcal{E})$
- A *family of distributions* $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ on $(E, \mathcal{E})$.

We suppose that $\mathbb{P}_X = \mathbb{P} \circ X^{-1} \in \mathcal{P}$, which means that $\mathbb{P}_X = P_{\theta_0}$ for some $\theta_0 \in \Theta$. We say that $\mathcal{P}$ is a *statistical model* for X and we will denote the statistical experiment as $(X, \mathcal{P})$. We call Θ the set of *parameters* of the model.

We call X a *random object* in Definition 1.1 to stress that it can be a real random variable, a random vector, a random matrix, among other things. Also, the assumption $\mathbb{P}_X \in \mathcal{P}$ means that the model is *well-specified*, which is a strong assumption, since it requires that the true distribution belongs to the chosen model $\mathcal{P}$.

The random variable $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (E, \mathcal{E})$ has distribution $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$ which is the probability image of $\mathbb{P}$ by X on $(\Omega, \mathcal{A})$. We will always suppose that there is a family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ on $(\Omega, \mathcal{A})$ that induce $\{P_\theta : \theta \in \Theta\}$ on $(E, \mathcal{E})$ and we will use the notations

$$P_\theta[A] = \mathbb{P}_\theta[X \in A] = \mathbb{P}_\theta[\{\omega \in \Omega : X(\omega) \in A\}] \quad \forall A \in \mathcal{E}\,.$$

Once again, $(\Omega, \mathcal{A}, \mathbb{P})$ is a purely mathematical build of little interest in statistics. We could even assume that X is the identity function and that $(\Omega, \mathcal{A}) = (E, \mathcal{E})$. Because of the transfer formula

$$\int f(X(\omega))\mathbb{P}_\theta(d\omega) = \int f(x)P_\theta(dx),$$

we can even work only with P_θ and forget about $\mathbb{P}_\theta$ and the space $(\Omega, \mathcal{A})$.

We will often work with a set of finite-dimensional parameters $\Theta \subset \mathbb{R}^d$, which corresponds to a *parametric* model, but this space can be more complicated than that (it can be a set of functions with some smoothness properties for instance, such a case is covered by a field called *non-*

parametric statistics [2, 3]). We will use the notations

[2]: Tsybakov (2008), *Introduction to nonparametric estimation*
[3]: Wasserman (2006), *All of nonparametric statistics*

$$\mathbb{E}_Q[f(Y)] = \mathbb{E}_{Y\sim Q}[f(Y)] := \int f(y)Q(dy)$$

where we implicitly assume, when computing this expectation, that $Y \sim Q$, and whenever $Q = P_\theta$, we will shorten this notation as follows:

$$\mathbb{E}_\theta[f(X)] := \mathbb{E}_{\mathbb{P}_\theta}[f(X)] = \int f(x)P_\theta(dx).$$

We will often work with iid data, namely a *sampled* statistical experiment, as explained in the next definition.

Definition 1.2 A n*-sampled* statistical experiment corresponds to data $X = (X_1, \dots, X_n)$ with X_i iid and $\mathcal{P} = \{P_\theta^{\otimes n} : \theta \in \Theta\}$.

Namely, for a n-sampled statistical experiment and $A = \prod_{i=1}^n A_i$, one has the following:

$$\begin{aligned} P_\theta^{\otimes n}[A] &= \mathbb{P}_\theta^{\otimes n}[(X_1, \dots, X_n) \in A_1 \times \cdots \times A_n] \\ &= \prod_{i=1}^n \mathbb{P}_\theta[X_i \in A_i] = \prod_{i=1}^n P_\theta[A_i]. \end{aligned}$$

However, we will quickly forget about n-sampled experiments and simply say that we observe iid data $X_1, \dots, X_n$ from a model $\mathcal{P} = \{P_\theta \in \Theta\}$ (namely $\mathbb{P}_{X_1} \in \mathcal{P}$).

and when there is little doubt about what we are computing, we will simply forget to write the $\otimes n$ exponents.

1.3 Statistics

We really need at this point to tell the reader what a *statistic* is.

Definition 1.3 Given a statistical experiment $(X, \{P_\theta : \theta \in \Theta\})$, we call *statistic* any measurable function of X that does not depend on θ. A statistic is therefore a quantity that we can compute *using data only*.

If X is a random variable in $\mathbb{R}^n$ and S a random variable in $\mathbb{R}$, then we know that S is a statistic, namely $S = f(X)$ for some Borel measurable function f, if an only if S is $\sigma(X)$-measurable (or simply X-measurable). [13]

13: We recall that $\sigma(X)$ is the σ-algebra *generated* by X, namely $\sigma(X) = X^{-1}(\mathcal{B}^n)$ where $\mathcal{B}^n$ is the Borel σ-algebra of $\mathbb{R}^n$ and note that this statement comes from the Doob lemma: for such X and S, we have that S is X-measurable if and only if $S = f(X)$ for some Borel measurable function $\mathbb{R}^n \to \mathbb{R}$.

Now, let us go back to the $\mathrm{Bernoulli}(\theta)$ experiment. Let us first recall that for this experiment we have iid samples $X = (X_1, \dots, X_n)$, that $(E, \mathcal{E}) = (\{0,1\}^n, \mathcal{P}(\{0,1\}^n))$ and that $P_\theta = \mathrm{Bernoulli}(\theta)$ with $\Theta = (0,1)$. Intuitively, an "equivalent" experiment is $S_n := \sum_{i=1}^n X_i$

and $(E, \mathcal{E}) = (\{0, \dots, n\}, \mathcal{P}(\{0, \dots, n\}))$ with $P_\theta = \text{Binomial}(n, \theta)$ and $\theta \in \Theta = (0, 1)$. Let us observe that

$$\begin{aligned} \mathbb{P}_\theta[X_1 = x_1, \dots, X_n = x_n | S_n = k] &= \frac{\theta^k (1-\theta)^{n-k}}{\binom{n}{k} \theta^k (1-\theta)^{n-k}} \\ &= \frac{1}{\binom{n}{k}} \end{aligned} \tag{1.2}$$

whenever $x_i \in \{0, 1\}$ for $i = 1, \dots, n$ and $k = \sum_{i=1}^n x_i$, while $\mathbb{P}_\theta[X_1 = x_1, \dots, X_n = x_n | S_n = k] = 0$ otherwise. This proves that the *conditional distribution* of $(X_1, \dots, X_n)|S_n$ *does not depend* on θ. If we know S_n, then we can, *without knowing* θ, build a "copy" $X' = (X'_1, \dots, X'_n)$ of the original sample X, in the sense that X' has the same distribution as X. This is simply achieved, as indicated by Equation (1.2), by choosing the positions of the $S_n = k$ ones (among n ones and zeros) uniformly at random. This means that X does not bring more information about θ than S_n, and that that S_n is somehow "sufficient" from a statistical point of view. Such a random variable is called a *sufficient statistic*.

For the $\text{Bernoulli}(\theta)$ experiment, we will look for statistics of X allowing to infer the unknown parameter θ. We already now that we can restrict ourselves to statistics of S_n instead of X, since S_n is sufficient.

1.4 Identifiable models

The only source of information about θ available to us about is X, through its distribution P_θ. So, in order to infer θ, we need, at least, to be able to recover θ given P_θ. We will therefore often require that the model is *identifiable*, as defined below.

Definition 1.4 We say that a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is *identifiable* whenever the function $\Theta \to \mathcal{P}$ given by $\theta \mapsto P_\theta$ is *injective*.

Identifiability is a necessary requirement when one wants to perform statistical inference. If $\theta \mapsto P_\theta$ is not injective, then there is no way to find back θ from $X \sim P_\theta$.

Example 1.1 Obviously, $\theta \mapsto \text{Bernoulli}(\theta)$ is injective on $(0, 1)$ and similarly $x \mapsto \text{Bernoulli}(\text{sigmoid}(x))$ is injective on $\mathbb{R}$. A stupid example is $\mu \mapsto \text{Normal}(\mu^2, 1)$ on $\mathbb{R}$, which corresponds to a non-identifiable model.

In Example 1.1, we used the sigmoid function given by $\text{sigmoid}(x) = 1/(1 + e^{-x})$ for any $x \in \mathbb{R}$.[14] Identifiability is generally a property that we ensure by choosing and parametrizing correctly the considered statistical model.

14: The sigmoid function is heavily used in statistics and machine learning, we will come back to it later in the book.

However, not all interesting and useful statistical models are identifiable. An interesting example of *non-identifiable* model is given by *mixture models*, such as the Gaussian mixture model, where we consider a distribution on $\mathbb{R}^d$ with a density with respect to the Lebesgue measure given by [15]

$$f_\theta(x) = \sum_{k=1}^{K} \pi_k \phi_{\mu_k, \mathbf{\Sigma}_k}(x)$$
$$=: \sum_{k=1}^{K} \frac{\pi_k}{\sqrt{(2\pi)^d \det(\mathbf{\Sigma}_k)}} \exp\Big(-\frac{1}{2}(x-\mu_k)^\top \mathbf{\Sigma}_k^{-1}(x-\mu_k)\Big)$$

for any $x \in \mathbb{R}^d$, where $\theta = (\pi_k, \mu_k, \mathbf{\Sigma}_k)_{k=1,\dots,K}$ and $K \geq 1$ is an integer corresponding to the number of "clusters". Such a mixture density is non-identifiable, since we have $f_\theta = f_{\sigma(\theta)}$ for any permutation $\sigma(\theta) = (\pi_{\sigma(k)}, \mu_{\sigma(k)}, \mathbf{\Sigma}_{\sigma(k)})_{k=1,\dots,K}$ of θ where σ is a permutation of $\{1, \dots, K\}$. This simply means that the density distribution f_θ is invariant by a relabeling of the clusters numbers. Despite the fact that such a mixture model is not identifiable, it is often used for *model-based clustering*, which is an instance of *unsupervised learning* [4] . Another family of non-identifiable models is deep neural networks, in which an infinitely large number of parametrizations lead to the same prediction function [5] .

15: The expectations $\mu_k \in \mathbb{R}^d$ correspond to the centroids of the clusters. The covariances $\mathbf{\Sigma}_k \in \mathbb{R}^{d\times d}$ are such that $\mathbf{\Sigma}_k \succ 0$, which means that $\mathbf{\Sigma}_k$ is a positive definite matrix. The matrix $\mathbf{\Sigma}_k$ parametrizes the shape of cluster k around μ_k. Finally, the parameters $\pi_k \geq 0$ are such that $\sum_{k=1}^K \pi_k = 1$ and parametrize the relative population of each cluster.

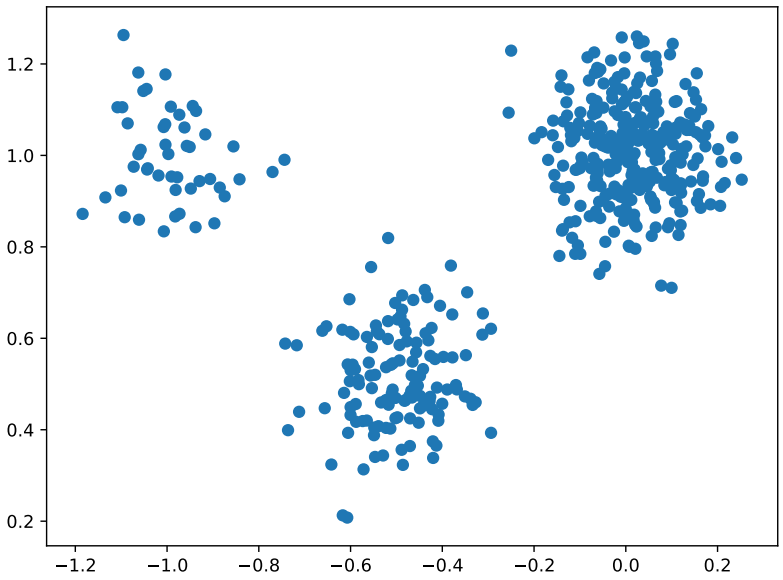


Figure 1.3: 500 realizations of a Gaussian mixture with $d = 2$, $K = 3$ $\pi = [\pi_1, \pi_2, \pi_3] = [0.1, 0.6, 0.3]$, $\mu_1 = [-1, 1]$, $\mu_2 = [0, 1]$, $\mu_3 = [-0.5, 0.5]$ and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3 = 0.01 \times \boldsymbol{I}_2$.

[4]: Murphy (2012), *Machine Learning, A Probabilistic Perspective*

[5]: Goodfellow et al. (2016), *Deep learning*

1.5 Dominated models

Whenever $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^d$, we say that $\mathcal{P}$ is a *parametric* model, since it is parametrized by a finite-dimensional pa-

rameter, trivial instances being $\{\text{Bernoulli}(\theta) : \theta \in (0,1)\}$ for which $d = 1$ and $\{\text{Normal}(\mu, v) : (\mu, v) \in \mathbb{R} \times (0, +\infty)\}$ for which $d = 2$. We say that both models are *dominated*, the first being dominated by the counting measure $\nu = \delta_0 + \delta_1$ on $\{0, 1\}$,[16] and the second by the Lebesgue measure on $\mathbb{R}$.[17]

Definition 1.5 We say that a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is dominated if there is a σ-finite measure μ such that $P_\theta \ll \mu$ for all $\theta \in \Theta$.

16: The notation δ_x will stand for the Dirac mass at x which is the probability measure satisfying $\int f(u)\delta_x(du) = f(x)$ for any measurable function f.

17: We recall that if P and Q are two finite measures on the same probability space, $P \ll Q$ means that the measure P is *absolutely continuous* with respect to Q, namely that $Q(A) = 0 \Rightarrow P(A) = 0$ for any measurable set A.

In Definition 1.5, we require that the dominating measure is σ-finite, so that we can apply the Radon-Nikodym theorem: since $P_\theta \ll \mu$ for all $\theta \in \Theta$, there is a density

$$f_\theta = \frac{dP_\theta}{d\mu}$$

for all $\theta \in \Theta$, with is unique μ-almost surely. This means that $P_\theta[A] = \int f_\theta(x)\mu(dx)$ for any measurable set A. This domination property allows to work with densities instead of distributions: a model can be therefore defined as a family of densities $\{f_\theta : \theta \in \Theta\}$ together with a dominating measure (which is in most cases the Lebesgue measure, a counting measure, or a combination of both.)

Let us recall the Radon-Nikodym theorem. Let P be a probability and Q be a σ-finite measure on a measurable space $(\Omega, \mathcal{A})$ and assume that $P \ll Q$. Then, there is a non-negative random variable L such that $P[A] = \int_\Omega L(\omega)\mathbf{1}_A(\omega)Q(d\omega)$ for any $A \in \mathcal{A}$. We denote $L = dP/dQ$ and L is unique Q-almost surely.

Example 1.2 Let us consider the *zero-inflated Laplace distribution*, which is a distribution on $\mathbb{R}$ given by

$$P_\theta(dx) = \pi_0\delta_0(dx) + (1 - \pi_0)\frac{\lambda}{2}e^{-\lambda|x|}dx$$

for $\theta = (\pi_0, \lambda) \in \Theta = (0, 1) \times (0, +\infty)$, which is dominated by the measure $\mu = \delta_0 + \text{Lebesgue}$, where Lebesgue stands for the Lebesgue measure on $\mathbb{R}$.

The notation $P(dx) = f(x)dx$ means that the distribution P has density f with respect to the Lebesgue measure.

Non-dominated models are usually pathological and uninteresting examples, such as the model

$$P_\theta = \frac{1}{e}\sum_{n\in\mathbb{N}}\frac{1}{n!}\delta_{\theta n}$$

for $\theta \in (0, +\infty)$, which cannot be dominated by a σ-finite measure.

1.6 Proofs

Proof of Inequalities (1.1). The upper bound is just easily obtained using

$$1-\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_x^{+\infty} e^{-t^2/2}dt \leq \frac{1}{\sqrt{2\pi}}\int_x^{+\infty}\frac{t}{x}e^{-t^2/2}dt = \frac{e^{-x^2/2}}{x\sqrt{2\pi}}$$

while the lower bound comes from the fact that

$$\int_x^{+\infty} \Big(1 - \frac{3}{t^4}\Big) e^{-t^2/2} dt = \Big(\frac{1}{x} - \frac{1}{x^3}\Big) e^{-x^2/2}$$

which concludes the proof. □

Statistical inference 2

In this Chapter, we introduce the three main *statistical inference* problems: *estimation*, *confidence intervals* and *tests*. Each problem will be instantiated with the simple Bernoulli model, where we have iid samples $X_1, \dots, X_n$ distributed as $\mathrm{Bernoulli}(\theta)$ with $\theta \in (0,1)$. Let us start with the first inference problem: *estimation*.

2.1 Estimation

We want to *infer* θ, or *estimate* it by finding a statistic which is a measurable function of $(X_1, \dots, X_n)$[1] or a measurable function of $S_n = \sum_{i=1}^n X_i$ thereof, since S_n is sufficient, see Section 1.3. We will denote such a statistic as

$$\widehat{\theta}_n = \widehat{\theta}_n(X_1, \dots, X_n).$$

This function *does not depend* on θ, but of course its distribution does. Ideally, we want $\widehat{\theta}_n$ to be "close" to θ, since we want a good estimator, so that the first thing we need to do is to quantify "closeness". For instance, we could want $|\widehat{\theta}_n - \theta|$ to be close to 0 with a large probability, since we do not forget that $\widehat{\theta}_n$ is a random variable, as a function of the data $(X_1, \dots, X_n)$. The most natural distance is arguably the Euclidean one, in this context the L^2 distance, which leads to the *quadratic risk*.[2]

Definition 2.1 (Quadratic risk) Consider a statistical model with data X and set of parameters $\Theta \subset \mathbb{R}$ and an estimator $\widehat{\theta}(X)$. The quadratic risk of $\widehat{\theta}$ is given by

$$R(\widehat{\theta}, \theta) = \mathbb{E}_\theta[(\widehat{\theta} - \theta)^2] = \int_E (\widehat{\theta}(x) - \theta)^2 P_\theta(dx).$$

We consider the quadratic risk as a function $\Theta \to \mathbb{R}^+$ of the parameter given by $\theta \mapsto R(\widehat{\theta}, \theta)$.

At this point, it's useful to recall some classical inequalities on the queues of random variables. The Markov inequality tells us that if Y is a real random variable such that $\mathbb{E}|Y|^p < +\infty$ for some $p > 0$ then

$$\mathbb{P}[|Y| > t] \leq \frac{\mathbb{E}|Y|^p}{t^p}$$

for any $t > 0$. This tells us that the more Y has moments[3] the more

1: Once again, since we are doing statistics, the only thing we are allowed to use is the data.

2: Although the quadratic risk corresponds to a *squared* L^2 norm.

3: We say that Y as moments up to order p if $\mathbb{E}|Y|^p < +\infty$. Note that this entails $\mathbb{E}|Y|^q < +\infty$ for any $q < p$ since $\mathbb{E}|Y|^p = \mathbb{E}[|Y|^q]^{p/q} \geq (\mathbb{E}|Y|^q)^{p/q}$ using Jensen's inequality.

the queue of Y is tight (it goes faster to 0 with $t \to +\infty$). Markov's inequality with $p = 2$ entails

$$\mathbb{P}[|\widehat{\theta} - \theta| > t] \leq \frac{R(\widehat{\theta}, \theta)}{t^2} \tag{2.1}$$

which tells us that whenever the quadratic risk is small, then $\widehat{\theta}$ is close to θ with a large probability.

Whenever $R(\widehat{\theta}_n, \theta) \to 0$ with $n \to +\infty$, we will write $\widehat{\theta}_n \overset{L^2}{\to} \theta$, which stands for convergence in L^2 norm, which entails, because of Inequality (2.1), that $\widehat{\theta}_n \overset{\mathbb{P}}{\to} \theta$, which stands for convergence in probability.[4]

4: More precisely, in $\mathbb{P}_\theta$-probability, namely $\mathbb{P}_\theta[|\widehat{\theta}_n - \theta| > \varepsilon] \to 0$ as $n \to +\infty$ for any $\varepsilon > 0$, but we will write $\widehat{\theta}_n \overset{\mathbb{P}}{\to} \theta$ in order to keep the notations as simple as possible.

Definition 2.2 We say that $\widehat{\theta}_n$ is *consistent* whenever $\mathbb{P}_\theta[|\widehat{\theta}_n - \theta| > \varepsilon] \to 0$ as $n \to +\infty$ for any $\varepsilon > 0$ and any $\theta \in \Theta$. We say that it is strongly consistent whenever $\mathbb{P}_\theta[\widehat{\theta}_n \to \theta] = 1$ for any $\theta \in \Theta$.

In Definitions 2.1 and 2.2 above, if $\Theta \subset \mathbb{R}^d$, it suffices to replace $|\cdot|$ by the Euclidean norm $\|\cdot\|_2$, where $\|x\|_2 = (x^\top x)^{1/2} = (\sum_{j=1}^d x_j^2)^{1/2}$.

Bias variance decomposition. The *bias-variance decomposition* is the following decomposition of the quadratic risk between two terms: a bias term denoted $b(\widehat{\theta}, \theta)$ (squared in the formula) and a variance term:

$$\begin{aligned} R(\widehat{\theta}, \theta) &= \mathbb{E}_\theta[(\widehat{\theta} - \theta)^2] = (\mathbb{E}_\theta[\widehat{\theta}] - \theta)^2 + \mathbb{V}_\theta[\widehat{\theta}] \\ &= b(\widehat{\theta}, \theta)^2 + \mathbb{V}_\theta[\widehat{\theta}]. \end{aligned} \tag{2.2}$$

When $b(\widehat{\theta}, \theta) = 0$ for all $\theta \in \Theta$ we say that the estimator $\widehat{\theta}$ is *unbiased*. This means that this estimator will not over or under-estimate θ, since its expectation equals θ.

Back to Bernoulli. Going back to the $\mathrm{Bernoulli}(\theta)$ model, we consider the estimator $\widehat{\theta}_n = S_n/n = n^{-1}\sum_{i=1}^n X_i$. We already know many things about this estimator:

1. We have $\mathbb{E}_\theta[\widehat{\theta}_n] = \theta$ which means that $\widehat{\theta}_n$ is unbiased;
2. The bias-variance decomposition gives

$$R(\widehat{\theta}_n, \theta) = \mathbb{V}_\theta[\widehat{\theta}_n] = \frac{\theta(1-\theta)}{n} \leq \frac{1}{4n} \to 0 \tag{2.3}$$

   which means that $\widehat{\theta}_n \overset{L^2}{\to} \theta$ and which entails that $\widehat{\theta}_n$ is consistent;
3. The law of large number tells us that $\widehat{\theta}_n \overset{\text{as}}{\to} \theta$, hence $\widehat{\theta}_n$ is strongly consistent;

4. The central limit theorem tells us that

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow \text{Normal}(0, \theta(1-\theta)). \tag{2.4}$$

The points 2–4 from above are all different ways of saying that when n is large, then $\widehat{\theta}_n$ is close to θ.

In practice, an estimator leads to a value: for the Bernoulli experiment with $n = 100$ and 42 ones you end up with a single estimated value 0.42. But what if we want to include uncertainty in this estimation? Namely how confident are we about this 0.42 value? Moreover, what do we mean by "when n is large enough"? Can we quantify this somehow? These questions can be answered by considering another inference problem: confidence intervals.

2.2 Confidence intervals

Here, we don't only want to build an estimator $\widehat{\theta}_n$ but also to quantify the uncertainty associated to this estimation.

2.2.1 Non-asymptotic coverage

Combining Inequalities (2.1) and (2.3) leads to

$$\mathbb{P}_\theta[|\widehat{\theta}_n - \theta| > t] \leq \frac{1}{4nt^2},$$

so that for $\alpha \in (0,1)$ and the choice $t_\alpha = 1/(2\sqrt{n\alpha})$ we have

$$\mathbb{P}_\theta\big\{\theta \in [\widehat{\theta}_n^L, \widehat{\theta}_n^R]\big\} \geq 1 - \alpha \tag{2.5}$$

for any $\theta \in (0,1)$, where

$$\widehat{\theta}_n^L := \widehat{\theta}_n - \frac{1}{2\sqrt{n\alpha}} \quad \text{and} \quad \widehat{\theta}_n^R := \widehat{\theta}_n + \frac{1}{2\sqrt{n\alpha}}.$$

Therefore, if we choose $\alpha = 0.05 = 5\%$, we know that $\theta \in [\widehat{\theta}_n^L, \widehat{\theta}_n^R]$ with a probability larger than 95%. We say in this case that the interval $[\widehat{\theta}_n^L, \widehat{\theta}_n^R]$ is a *confidence interval* with *coverage* 95%.[5]

If $\alpha = 0$ we have no other choice than using the whole $\mathbb{R}$ as a confidence interval: α provides us with some slack, so that we can build a non-absurdly large confidence interval. We have that $|\widehat{\theta}_n^R - \widehat{\theta}_n^L|$ increases as α decreases, since a smaller α means more confidence, hence a larger interval. On the contrary, $|\widehat{\theta}_n^R - \widehat{\theta}_n^L|$ decreases with the sample size n.

5: If we toss the coin 1000 times and get 420 heads, the realization of this confidence interval at 95% is $[0.35, 0.49]$.

Definition 2.3 (Confidence interval) Consider a statistical model with data X and set of parameters $\Theta \subset \mathbb{R}$. Fix a *confidence level* $\alpha \in (0,1)$ and consider two statistics $\widehat{\theta}^L(X)$ and $\widehat{\theta}^R(X)$. Whenever

$$\mathbb{P}_\theta\big\{\theta \in [\widehat{\theta}^L(X), \widehat{\theta}^R(X)]\big\} \geq 1-\alpha \tag{2.6}$$

for any $\theta \in \Theta$, we say that $[\widehat{\theta}^L(X), \widehat{\theta}^R(X)]$ is a *confidence interval* at *level* or *coverage* $1-\alpha$.

Inequality (2.6) is called the *coverage* property of the confidence interval. More generally, when $\Theta \subset \mathbb{R}^d$, we will say that $S(X)$ is a *confidence set* if it is a statistic satisfying the coverage property $\mathbb{P}_\theta[\theta \in S(X)] \geq 1-\alpha$ for any $\theta \in \Theta$.

Remark 2.1 Whenever we need only an upper or lower bound on θ (for instance, when we need to check statistically that some toxicity level is below some threshold), we build a *unilateral* or *one-sided* confidence interval, where we choose either $\widehat{\theta}^L = -\infty$ (0 for the Bernoulli model) or $\widehat{\theta}^R = +\infty$ (1 for the Bernoulli model). Indeed, at a fixed level $1-\alpha$, the bound provided by a one-sided confidence interval is tighter than the bound of a two-sided interval.

But, we can do better for the Bernoulli model (or any model where samples are bounded almost surely) thanks to the following Hoeffding inequality.

Theorem 2.1 (Hoeffding) Let $X_1, \ldots, X_n$ be independent random variables such that $X_i \in [a_i, b_i]$ almost surely and let $S = \sum_{i=1}^n X_i$. Then,

$$\mathbb{P}[S \geq \mathbb{E}S + t] \leq \exp\Big(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big)$$

holds for any $t > 0$.

Theorem 2.1 is something called a deviation inequality: it provides a control on the probability of deviation of S with respect to its mean. It shows that bounded random variables are *sub-Gaussian*, since it shows that the queue of $S - \mathbb{E}S$ is bounded by $\exp(-ct^2)$ for some constant c (that depends on n). The proof of Theorem 2.1 is provided in Section 2.4.

Back to Bernoulli. Let's apply Theorem 2.1 to the Bernoulli model $X_i \sim \text{Bernoulli}(\theta)$ so that $a_i = 0$, $b_i = 1$ and therefore $\mathbb{P}[S \geq \mathbb{E}S + t] \leq e^{-2t^2/n}$. Using again Theorem 2.1 with X_i replaced by $-X_i$ together with an union bound[6] leads to $\mathbb{P}[|S - \mathbb{E}S| \geq t] \leq 2e^{-2t^2/n}$. So, for some $\alpha \in (0,1)$, we obtain another confidence interval, since

6: Using Theorem 2.1 with X_i replaced by $-X_i$ gives $\mathbb{P}[-S + \mathbb{E}S \geq t] \leq e^{-2t^2/n}$, so that $\mathbb{P}[|S - \mathbb{E}S| \geq t] \leq \mathbb{P}[S - \mathbb{E}S \geq t] + \mathbb{P}[S - \mathbb{E}S \leq -t] \leq 2e^{-2t^2/n}$.

the following coverage property holds:

$$\mathbb{P}\left[\widehat{\theta}_n - \sqrt{\frac{\log(2/\alpha)}{2n}} \leq \theta \leq \widehat{\theta}_n + \sqrt{\frac{\log(2/\alpha)}{2n}}\right] \geq 1 - \alpha.$$

This proves that $[\widehat{\theta}_n \pm \sqrt{\log(2/\alpha)/(2n)}]$ is a confidence interval at level $1 - \alpha$.[7] Let's compare the two confidence intervals we obtained so far for the Bernoulli model. It can bee seen that

$$\frac{1}{2\sqrt{n\alpha}} > \sqrt{\frac{\log(2/\alpha)}{2n}}$$

for $\alpha < 0.23$, although both sides are $O(1/\sqrt{n})$. Only the dependence on the level α is improved with the confidence interval obtained through Hoeffding's inequality, since it exploits the sub-Gaussianity of the Bernoulli distribution, while the first confidence interval (2.5) only used the upper bound (2.1) on the variance.

7: For 1000 tosses and 420 heads, the realization of this interval at level 95% is $[0.37, 0.46]$. It's a bit more precise than the previous one, which was based on Markov's inequality.

There is yet another way to build a confidence interval, called *exact* confidence interval. Let us denote by $F_{n,\theta}$ the distribution function of $\mathrm{Binomial}(n, \theta)$. It is given by

$$F_{n,\theta}(x) = \sum_{k=0}^{[x]} \binom{n}{k} \theta^k (1-\theta)^{n-k} \tag{2.7}$$

for $x \in [0, n]$, where $[x]$ stands for the integer part of x, while $F_{n,\theta}(x) = 0$ if $x < 0$ and $F_{n,\theta}(x) = 1$ if $x \geq n$. We can consider the generalized inverse $F_{n,\theta}^{-1}$ of $F_{n,\theta}$, also called the *quantile function* of $\mathrm{Binomial}(n, \theta)$, for which we know that $F_{n,\theta}^{-1}(\alpha) \leq F_{n,\theta'}^{-1}(\alpha)$ for any $\theta \leq \theta'$ and $\alpha \in (0, 1)$.[8] Because of this, we know that the set $\{\theta \in (0,1) : F_{n,\theta}^{-1}(\alpha/2) \leq n\widehat{\theta}_n \leq F_{n,\theta}^{-1}(1-\alpha/2)\}$ is an interval, so that defining

8: See Proposition 2.5 below and its proof for details on this generalized inverse and its properties, together with Example 2.3.

$$\widehat{\theta}^L = \inf\{\theta \in (0,1) : F_{n,\theta}^{-1}(1-\alpha/2) \geq n\widehat{\theta}_n\}$$

and

$$\widehat{\theta}^R = \sup\{\theta \in (0,1) : F_{n,\theta}^{-1}(\alpha/2) \leq n\widehat{\theta}_n\}$$

leads to the coverage property

$$\begin{aligned}\mathbb{P}_\theta\{\theta \in [\widehat{\theta}^L, \widehat{\theta}^R]\} &= \mathbb{P}_\theta\big[F_{n,\theta}^{-1}(\alpha/2) \leq n\widehat{\theta}_n \leq F_{n,\theta}^{-1}(1-\alpha/2)\big] \\ &= 1 - \alpha/2 - \alpha/2 = 1 - \alpha\end{aligned}$$

since $n\widehat{\theta}_n \sim \mathrm{Binomial}(n, \theta)$. This confidence interval is called "exact" since it uses the exact quantile function of $n\widehat{\theta}_n$. It is therefore even tighter than the previous ones.

2.2.2 Asymptotic coverage

For the previous confidence intervals, we adopted a *non-asymptotic* approach: the coverage properties hold for any value of $n \geq 1$. This was possible since the distribution of S_n is a simple $\mathrm{Binomial}(n, \theta)$ distribution, for which many computations can be made explicit. However, in general, the *exact* distribution of an estimator $\widehat{\theta}_n$ cannot always be exhibited, and in such cases, we often use Gaussian approximations, thanks to the central limit theorem. Let's do this for the Bernoulli model. We know from (2.4) that

$$\mathbb{P}_\theta\left[\sqrt{\frac{n}{\theta(1-\theta)}}(\widehat{\theta}_n - \theta) \in I\right] \to \mathbb{P}[Z \in I] \tag{2.8}$$

where $Z \sim \mathrm{Normal}(0,1)$ for any interval $I \subset \mathbb{R}$.[9] Using $I = [-q_\alpha, q_\alpha]$ with $q_\alpha = \Phi^{-1}(1-\alpha/2)$ we end up[10] with

$$\mathbb{P}_\theta\left\{\theta \in \left[\widehat{\theta}_n \pm q_\alpha\sqrt{\frac{\theta(1-\theta)}{n}}\right]\right\} \to 1-\alpha. \tag{2.9}$$

This is interesting, but not enough to build a confidence interval, since the interval in (2.9) depends on θ through the variance term $\theta(1-\theta)$. Indeed, a confidence interval must be something that does *not* depend on θ. We need to work a little bit more in order to remove the dependence on θ from this interval. We can do the same as before: we use the fact that $\theta(1-\theta) \leq 1/4$ for any $\theta \in [0,1]$, so that

$$\liminf_n \mathbb{P}_\theta\left\{\theta \in \left[\widehat{\theta}_n \pm \frac{q_\alpha}{2\sqrt{n}}\right]\right\} \geq 1-\alpha. \tag{2.10}$$

This is what we call a confidence interval *asymptotically of level* $1-\alpha$ constructed *by excess*.

In the asymptotic confidence interval (2.10), we used the central limit theorem to approximate the distribution of $\sqrt{n}(S_n/n - \theta)$ by a Gaussian distribution. This requires n to be "large enough", but the central limit theorem does not tell us how large. We can quantify this better by assessing how close the distribution function of $\sqrt{n}(S_n/n-\theta)$ is to the one of the Gaussian distribution, using the following theorem.

Theorem 2.2 (Berry-Esséen) Let $X_1, \ldots, X_n$ be i.i.d random variables such that $\mathbb{E}[X_i] = 0$ and $\mathbb{V}[X_i] = \sigma^2$ and introduce the distribution function

$$F_n(x) = \mathbb{P}\left[\frac{\sum_{i=1}^n X_i}{\sqrt{n\sigma^2}} \leq x\right]$$

9: This uses the porte-manteau theorem, which says that $X_n \rightsquigarrow X$ if and only if $\mathbb{P}[X_n \in A] \to \mathbb{P}[X \in A]$ for any Borelian set A such that $\mathbb{P}[X \in \partial A] = 0$, where ∂A stands for the boundary of A.

10: We recall that Φ^{-1} is the *quantile* function of $\mathrm{Normal}(0,1)$, namely the inverse of the distribution function $\Phi(x) = \mathbb{P}[Z \leq x]$ with $Z \sim \mathrm{Normal}(0,1)$.

for any $x \in \mathbb{R}$. Then, the following inequality holds:

$$\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| \leq \frac{c\kappa}{\sigma^3\sqrt{n}},$$

where $\kappa = \mathbb{E}|X_1|^3$ (assumed finite) and where c is a purely numerical constant (the best known one is $c = 0.4748$).

The best known constant $c = 0.4748$ is from [6], which almost matches the lower bound $c \geq 0.4097$ from [7]. Note also that a similar result holds if the X_i are independent but not identically distributed.

A nice proof of this theorem with worse constants, which relies on Fourier analysis and approximation by Schwartz functions, can be found in [8] . For Bernoulli we have $\mathbb{E}|X_1|^3 = \theta$ and $\sigma^3 = (\theta(1-\theta))^{3/2}$ so that

[8]: Tao (2010), *254A, Notes 2: The central limit theorem*

$$|F_n(x) - \Phi(x)| \leq \frac{3}{\sqrt{n\theta(1-\theta)^3}}$$

which shows that the approximation by the Gaussian distribution deteriorates whenever θ is close to 0 or 1, which is expected since in this case the sequence $X_1, \ldots, X_n$ is almost deterministically constant and equal to 0 (when $\theta \approx 0$) or 1 (when $\theta \approx 1$).

Reparametrization. Another tool used in the construction of confidence intervals with asymptotic coverage is the idea of reparametrization. Indeed, given a statistical model $\{P_\theta : \theta \in \Theta\}$ and a bijective function $g : \Theta \to \Lambda$ we can use instead the "reparametrized"model $\{Q_\lambda : \lambda \in \Lambda\}$ where $Q_\lambda = P_{g^{-1}(\lambda)}$ for which the construction of a confidence interval $[\widehat{\lambda}^L, \widehat{\lambda}^R]$ for λ is easier. If g is a monotonic function, we can easily derive from $[\widehat{\lambda}^L, \widehat{\lambda}^R]$ a confidence interval for θ.

In order to use this reparametrization idea, a natural question is to understand if the convergence in distribution (involved in the central limit theorem) is stable under such a reparametrization.

Example 2.1 Consider a iid dataset $X_1, \ldots, X_n$ with distribution $\mathrm{Exponential}(\theta)$ with scale parameter $\theta > 0$, namely the distribution $P_\theta(dx) = \theta e^{-\theta x}\mathbf{1}_{x\geq 0}dx$. We have $\mathbb{E}(X_1) = 1/\theta$ and $\mathbb{V}(X_1) = 1/\theta^2$, so that using the law of large numbers and the central limit theorem we have

We recall that $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$.

$$\bar{X}_n \overset{\text{as}}{\to} \theta^{-1} \quad \text{and} \quad \sqrt{n}(\bar{X}_n - \theta^{-1}) \rightsquigarrow \mathrm{Normal}(0, \theta^{-2})$$

when $n \to +\infty$. Since $x \mapsto 1/x$ is a continuous function on $(0, +\infty)$, we know that $(\bar{X}_n)^{-1} \overset{\text{as}}{\to} \theta$ so that a strongly consistent estimator is given by $\widehat{\theta}_n = (\bar{X}_n)^{-1}$. But what can be said about the convergence in distribution of $\sqrt{n}(\widehat{\theta}_n - \theta)$?

This is answered by so-called Δ-method, described in the next theorem.

Theorem 2.3 (Δ-method) Let $(Z_n)_{n\geq 1}$ be a sequence of real random variables and assume that $a_n(Z_n - z) \rightsquigarrow Z$, where $(a_n)_{n\geq 1}$ is a positive sequence such that $a_n \to +\infty$, where $z \in \mathbb{R}$ and where Z is a real random variable. If g is a function defined on a neighborhood of z and differentiable at z, we have

$$a_n(g(Z_n) - g(z)) \rightsquigarrow g'(z)Z \tag{2.11}$$

as $n \to +\infty$.

The proof of Theorem 2.3 is given in Section 2.4. It holds also for a sequence (Z_n) of random vectors in $\mathbb{R}^d$ and a differentiable function $g : \mathbb{R}^d \to \mathbb{R}^{d'}$, and it reads in this case

$$a_n(g(Z_n) - g(z)) \rightsquigarrow J_g(z)Z, \tag{2.12}$$

where $J_g(z)$ is the Jacobian matrix of g at z. A particularly useful case is when Z is Gaussian. For instance, if $\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow \mathrm{Normal}(0, \sigma(\theta)^2)$, we have

$$\sqrt{n}(g(\widehat{\theta}_n) - g(\theta)) \rightsquigarrow \mathrm{Normal}(0, \sigma(\theta)^2(g'(\theta))^2)$$

whenever g satisfies the conditions of Theorem 2.3. Going back to the $\mathrm{Exponential}(\theta)$ case of Example 2.1, we obtain with $g(x) = 1/x$ and since $\widehat{\theta}_n = g(\bar{X}_n)$ that $\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow \mathrm{Normal}(0, \theta^2)$.

Another result which provides stability for the convergence in distribution under a smooth mapping is the so-called Slutsky theorem.[11]

Theorem 2.4 (Slutsky) Let $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ be sequences of random vectors in $\mathbb{R}^d$ and $\mathbb{R}^{d'}$ respectively, such that $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow y$ where $X \in \mathbb{R}^d$ is some random vector and $y \in \mathbb{R}^{d'}$. Then, we have that $Y_n \xrightarrow{\mathbb{P}} y$ and $(X_n, Y_n) \rightsquigarrow (X, y)$ as $n \to +\infty$. In particular, we have $f(X_n, Y_n) \rightsquigarrow f(X, y)$ for any continuous function f.

The proof of Theorem 2.4 is given in Section 2.4 below. The Δ-method provides stability for the convergence in distribution when a differentiable function is applied to a sequence, while the Slutsky theorem provides "algebraic" stability when combining two sequences converging respectively in distribution and probability. [12]

Back again to Bernoulli. We have $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta$, so that $(\widehat{\theta}_n(1-\widehat{\theta}_n))^{1/2} \xrightarrow{\mathbb{P}} (\theta(1-\theta))^{1/2}$ since $x \mapsto (x(1-x))^{1/2}$ is continuous on $[0,1]$ and let us write

$$\frac{\sqrt{n}(\widehat{\theta}_n - \theta)}{\sqrt{\widehat{\theta}_n(1-\widehat{\theta}_n)}} = \frac{\sqrt{n}(\widehat{\theta}_n - \theta)}{\sqrt{\theta(1-\theta)}} \times \sqrt{\frac{\theta(1-\theta)}{\widehat{\theta}_n(1-\widehat{\theta}_n)}} =: A_n \times B_n.$$

11: This theorem will be *very* useful for the study of limit distributions. For instance, it entails that $Z_n \xrightarrow{\mathbb{P}} z$ whenever $\sqrt{n}(Z_n - z)$ converges in distribution, and particular cases such as $X_n + Y_n \rightsquigarrow X_n + y$ and $X_n Y_n \rightsquigarrow X_n y$ will be used repeatedly, starting with the proof of Theorem 2.3 for instance.

12: Be careful with the convergence in distribution. Please keep in mind that this mode of convergence is about the convergence of the distributions and not the convergence of the random variables (hence its name). The notation $X_n \rightsquigarrow X$ is rather misleading but convenient. In particular, nothing can be said in general about $f(X_n, Y_n)$ when we know that $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y$ (unless X_n and Y_n are independent sequences).

We know that $A_n \rightsquigarrow \mathrm{Normal}(0,1)$ and that $B_n \xrightarrow{\mathbb{P}} 1$. Therefore, using Theorem 2.4 leads[13] to

$$\sqrt{\frac{n}{\widehat{\theta}_n(1-\widehat{\theta}_n)}}(\widehat{\theta}_n - \theta) \rightsquigarrow \mathrm{Normal}(0,1).$$

We just replaced θ by $\widehat{\theta}_n$ in the variance term $\theta(1-\theta)$ of the limit (2.8), but doing so required Slutsky's theorem to prove this rigorously, and this provides us another confidence interval with asymptotic coverage given by

$$\mathbb{P}_\theta\left\{\theta \in \left[\widehat{\theta}_n \pm q_\alpha\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}}\right]\right\} \to 1-\alpha$$

as $n \to +\infty$.[14]

13: With $f(x,y) = xy$.

14: With 1000 tosses and 420 heads, the realization of this confidence interval at level 95% is $[0.38, 0.45]$.

2.3 Tests

Let us consider, again, in this section, a statistical experiment with data X and model $\{P_\theta : \theta \in \Theta\}$. Here, we want to decide between to hypotheses H_0 and H_1, where

$$H_i \quad \text{means that} \quad \theta \in \Theta_i$$

for $i \in \{0,1\}$, where $\{\Theta_0, \Theta_1\}$ is a partition of the set of parameters Θ. In order to understand the concept of statistical testing, let us consider the following unsettling example: imagine that you need to decide if a patient has cancer or not. The patient has cancer if some parameter $\theta \in (0,1)$ about him satisfies $\theta \geq 0.42$. We *choose* $\Theta_0 = [0.42, 1]$ and $\Theta_1 = [0, 0.42)$, namely we decide that H_0 means that the patient has cancer, while H_1 means that the patient has not. We need to construct a testing function $\varphi : E \to \{0,1\}$ that maps $X \mapsto \varphi(X)$, our decision being given by the value of $\varphi(X)$. The convention is to decide that H_0 is true whenever $\varphi(X) = 0$, in this case we say that we *accept* H_0 and we *reject* H_0 whenever $\varphi(X) = 1$. The convention is with the "1" in $\varphi(X) = 1$ and H_1 which always means that we *reject* the *null hypothesis* H_0.

2.3.1 Type I and Type II errors

When $\theta \in \Theta_i$, we are correct if $\varphi(X) = i$ and incorrect if $\varphi(X) = 1-i$. We have two types of errors: the *Type-I error*, also called the *first-order error*, given by

$$\mathbb{P}_\theta[\varphi(X) = 1] = \mathbb{E}_\theta[\varphi(X)] \quad \text{for} \quad \theta \in \Theta_0 \tag{2.13}$$

and the *Type-II error*, also called *second-order error*, given by

$$\mathbb{P}_\theta[\varphi(X) = 0] = 1 - \mathbb{E}_\theta[\varphi(X)] \quad \text{for} \quad \theta \in \Theta_1. \tag{2.14}$$

For the cancer detection problem, the Type I error corresponds to the *probability of saying to the patient that he has not cancer while he has.* The Type II error corresponds to the *probability of saying to the patient that he has cancer while he has not*. Note that these two types of errors are not symmetrical: we consider that the first one is more serious than the second (although this can be debated, the patient could do a depression, or start an invasive treatment for nothing).[15] The important point here is that H_0 and H_1 must be *chosen* depending on the practical application considered. They are not *given* and they correspond to an important modeling choice. We will see below that H_0 and H_1 must be chosen, in practice, so that the corresponding Type I error is *more serious*, for the considered application, than the Type II error.

15: Of course this morbid example is highly unrealistic, and is used only to stress the asymmetry of errors in a statistical testing problem.

Definition 2.4 The function $\beta : \Theta \to [0, 1]$ that maps $\theta \mapsto \beta(\theta) = \mathbb{E}_\theta[\varphi(X)]$ is called the *power function* of the test φ.

Ideally, we would like both the Type I and Type II errors to be small, namely $\beta(\theta) \approx 0$ for $\theta \in \Theta_0$ and $\beta(\theta) \approx 1$ for $\theta \in \Theta_1$. But this is impossible: if Θ is a connected set then Θ_0 and Θ_1 share a common frontier, so that β must be discontinuous on it, while β is in general a continuous function. Therefore, it is hard to make both the Type I and Type II errors small at the same time.

2.3.2 Desymmetrization of statistical tests

The way a statistical test is performed is through the *Neyman-Pearson approach*, where we *desymmetrize* the problem: choose the hypothesis H_0 using common sense, so that the Type I error is more serious than the Type II error. The Type I error is always the rejection of H_0 when it is true, while the Type II error is always the acceptation of H_0 when it is false. The only thing that we choose is what are H_0 and H_1. Let us wrap up what we said before, and introduce some extra things in the next definition.

Definition 2.5 Consider a statistical testing problem with hypotheses

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$$

and a testing function $\varphi : E \to \{0, 1\}$. We call H_0 the *null* hypothesis and H_1 the *alternative* hypothesis. When $\varphi(X) = 0$ we say that the test *accepts* H_0 or simply that it *accepts*. When $\varphi(X) = 1$ the

test *rejects*. The set

$$R = \{x \in E : \varphi(x) = 1\}$$

is called the *rejection set* of the test φ, and we call its complement $R^{\complement}$ the *acceptation region*. The restriction $\beta : \Theta_0 \to [0,1]$ of the power function β from Definition 2.4 is called the *Type I error*, while the restriction $\beta : \Theta_1 \to [0,1]$ is called the *power* of the test. The function $1 - \beta : \Theta_1 \to [0,1]$ is called the *Type II error* or *second order error*. Whenever

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$$

for some fixed $\alpha \in (0,1)$, we say that the test *has level* α.

The random variable X is valued in a measurable space $(E, \mathcal{E})$.

The idea of desymmetrization is as follows: given a level $\alpha \in (0,1)$ (something like 1%, 5% or 10%) we build a test so that it has, *by construction*, level α. Namely, a test is *built so that the Type I error is controlled*, while nothing is done directly about the Type II error. Given two statistical tests with level α (namely Type I error $\leq \alpha$), we can simply compare their Type II error and choose the one that maximizes it.

Back to Bernoulli. Let us go back to the Bernoulli model where $X_1, \ldots, X_n$ are iid and distributed as $\mathrm{Bernoulli}(\theta)$. We consider the problem of statistical testing with hypotheses:

$$H_0 : \theta \leq \theta_0 \quad \text{against} \quad H_1 : \theta > \theta_0 \tag{2.15}$$

so that $\Theta = (0,1)$, $\Theta_0 = (0, \theta_0]$ and $\Theta_1 = (\theta_0, 1)$. We studied in Sections 2.1 and 2.2 the estimator $\widehat{\theta}_n = S_n/n = \bar{X}_n$ and know that it is a good estimator. A natural idea is therefore to reject H_0 if $\widehat{\theta}_n$ is too large.

Recipe

We build a test by defining its rejection set R. The shape of the rejection set can be easily guessed by looking at the alternative hypothesis H_1.

Since we want to reject when $\theta > \theta_0$, we want to consider a rejection set $R = \{\widehat{\theta}_n > c\}$ for some constant c chosen so that the Type I error is controlled by α. Note that choosing $c = \theta_0$ is a bad idea: using the central limit theorem, we see that $\mathbb{P}_{\theta_0}[\widehat{\theta}_n > \theta_0] \to 1/2$. We need to increase c by some amount, so that the Type I error can be indeed smaller than α.[16]

16: It is very easy to build a statistical test with $\alpha = 0$, namely with zero Type I error. For the cancer example from above, we just need to tell to all the patient that they have cancer. By doing so, we never miss any cancer diagnostic, but on the other hand this test has zero power. Arguably, this is not a good strategy, so we need to give some slack in the construction of the test by considering a small but non-zero α.

2.3.3 Stochastic domination

We understand at this point that c will depend on α, θ_0 and the sample size n, among other things, and that in view of Definition 2.5 it needs to be such that $\sup_{\theta \leq \theta_0} \beta(\theta) = \sup_{\theta \leq \theta_0} \mathbb{P}_\theta[\widehat{\theta}_n > c] \leq \alpha$. But, we know that $n\widehat{\theta}_n = S_n \sim \text{Binomial}(n, \theta)$ under $\mathbb{P}_\theta$[17] , so that

$$\beta(\theta) = \mathbb{P}_\theta[S_n > nc] = \mathbb{P}[\text{Binomial}(n, \theta) > nc] \tag{2.16}$$

for any $\theta \in (0, 1)$.[18] In order to control the supremum of β, we need to study its variations: in view of (2.16) and (2.7), we know that

$$\beta(\theta) = 1 - F_{n,\theta}(nc) = 1 - \sum_{k=0}^{[nc]} \binom{n}{k} \theta^k (1-\theta)^{n-k} \tag{2.17}$$

for $nc \in [0, n]$, where $[x]$ stands for the integer part of $x \geq 0$, so that a direct study of the variations of β is somewhat tedious. Intuitively, $\beta(\theta)$ should be increasing with θ, since when θ increases, we get more ones, so that S_n increases. This can be nicely formalized using the notion of *stochastic domination*.

Proposition 2.5 Let P and Q be two probability measures on the same real probability space. We say that Q *stochastically dominates* P, that we denote $P \preceq Q$, whenever one of the following equivalent points is granted:

1. There are two real random variables $X \sim P$ and $Y \sim Q$ (on the same probability space) such that $\mathbb{P}[X \leq Y] = 1$;
2. We have $F_P(x) \geq F_Q(x)$ for any $x \in \mathbb{R}$, where F_P and F_Q are the distribution functions of P and Q, or equivalently, $P[(x, +\infty)] \leq Q[(x, +\infty)]$ for any $x \in \mathbb{R}$;
3. We have $F_P^-(p) \leq F_Q^-(p)$ for any $p \in [0, 1]$ where $F_P^-(p) = \inf\{x \in \mathbb{R} : F_P(x) \geq p\}$ is the *generalized inverse* of F_P or *quantile function* of P;
4. For any non-decreasing and bounded function f we have $\int f dP \leq \int f dQ$.

The proof of Proposition 2.5 is given in Section 2.4 below and follows rather standard arguments. However, the proof of $(3) \Rightarrow (1)$ deserves to be discussed here, since it uses a simple yet beautiful *coupling* argument, which is a very powerful technique often used in probability theory [9] . More precisely, we use something called a "quantile coupling": consider a random variable $U \sim \text{Uniform}([0, 1])$[19] on some probability space and define $X = F_P^-(U)$ and $Y = F_Q^-(U)$. We have by construction[20] that $X \sim P$ and $Y \sim Q$, and that

$$\mathbb{P}[X \leq Y] = \mathbb{P}[F_P^-(U) \leq F_Q^-(U)] = 1$$

since Point 3 tells us that $F_P^-(p) \leq F_Q^-(p)$ for any $p \in [0, 1]$. This

17: We write "under $\mathbb{P}_\theta$" here since Type I error control must be performed under the null assumption $\theta \leq \theta_0$, so that we must specify under which distribution (which parameter θ) we are working at this point.

18: The notation $\mathbb{P}[\text{Binomial}(n, \theta) > nc]$ stands for $\mathbb{P}[B > nc]$ where $B \sim \text{Binomial}(n, \theta)$. Note also that we replaced $\mathbb{P}_\theta$ simply by $\mathbb{P}$ herein, the notation $\mathbb{P}_\theta$ is required when we need to stress that the computation is performed under $\mathbb{P}_\theta$, while in the last equality we consider a generic probability space with probability $\mathbb{P}$ on which B lives, the dependency on θ is now only through the distribution of it. These semantics are important and will prove useful for statistical computations.

We recall that the distribution function of P is $F_P(x) = P[(-\infty, x]]$.

Since a distribution function is non-decreasing and càdlàg, its generalized inverse is well-defined and unique. See the proof of Proposition 2.5 for more details about it.

19: We say that $X \sim \text{Uniform}([a, b])$ for $a < b$ if it has density $x \mapsto (b-a)^{-1}\mathbf{1}_{[a,b]}(x)$ with respect to the Lebesgue measure, namely $\mathbb{P}_X(dx) = (b-a)^{-1}\mathbf{1}_{[a,b]}(x)dx$.

[9]: Hollander (2012), 'Probability theory: The coupling method'

20: This comes from the fact that $\mathbb{P}[F_P^-(U) \leq x] = \mathbb{P}[U \leq F_P(x)] = F_P(x)$ since $U \sim \text{Uniform}([0, 1])$ and since, by construction of the generalized inverse, we have that $F_P^-(u) \leq x$ is equivalent to $u \leq F_P(x)$ for any $u \in [0, 1]$ and $x \in \mathbb{R}$.

proves Point 3 $\Rightarrow$ Point 1.

The really nice feature of Proposition 2.5 is that it allows to reformulate $P \preceq Q$, which is a property regarding the *distributions* P and Q, as a property about *random variables* $X \sim P$ and $Y \sim Q$. Let us provide two examples.

Example 2.2 Whenever $\lambda_1 \leq \lambda_2$, we have $\mathrm{Exponential}(\lambda_2) \preceq \mathrm{Exponential}(\lambda_1)$. This follows very easily from Point 2 of Proposition 2.5.

Example 2.3 Whenever $\theta_1 \leq \theta_2$, we have $\mathrm{Bernoulli}(n, \theta_1) \preceq \mathrm{Bernoulli}(n, \theta_2)$. This is obtained through Point 1 (namely a coupling argument). Consider $U_1, \dots, U_n$ iid $\mathrm{Uniform}([0,1])$ and define $S_{n,i} = \#\{k : U_k \leq \theta_i\}$ for $i \in \{1,2\}$. By construction we have $S_{n,i} \sim \mathrm{Binomial}(n, \theta_i)$, and obviously $\mathbb{P}[S_{n,1} \leq S_{n,2}] = 1$ since $\theta_1 \leq \theta_2$.

The notation $\#E$ stands for the cardinality of a set E.

Thanks to Example 2.3 together with Proposition 2.5, we know now that $F_{n,\theta_2} \leq F_{n,\theta_1}$ whenever $\theta_1 \leq \theta_2$, so that combined with Inequality (2.16) this provides the following control of the Type I error:

$$\sup_{\theta \leq \theta_0} \mathbb{P}_\theta[\widehat{\theta}_n > c] = \sup_{\theta \leq \theta_0} (1 - F_{n,\theta}(nc)) \leq 1 - F_{n,\theta_0}(nc).$$

We can find out, given θ_0, α and n, a constant c as small as possible that satisfies $F_{n,\theta_0}(nc) \geq 1 - \alpha$, like we did in Section 2.2 for the exact confidence interval. Otherwise, we can use Theorem 2.1 (but it leads to a slightly less powerful test) to obtain

$$\mathbb{P}_{\theta_0}[\widehat{\theta}_n > c] = \mathbb{P}_{\theta_0}[S_n - n\theta_0 > c'] \leq e^{-2c'^2/n} = \alpha,$$

so that choosing $c' = \sqrt{n \log(1/\alpha)/2}$ gives $\sup_{\theta \leq \theta_0} \beta(\theta) \leq \alpha$, and proves that the test with rejection set

$$R = \left\{ \widehat{\theta}_n \geq \theta_0 + \sqrt{\frac{\log(1/\alpha)}{2n}} \right\}$$

is a test of level α for the hypotheses (2.15). Note that we managed to quantify exactly by how much we need to increase θ_0 in order to tune the test so that its Type I error is smaller than α.

2.3.4 Asymptotic approach

We can use also an asymptotic approach by considering the test with rejection set

$$R = \left\{ \widehat{\theta}_n > \theta_0 + \delta_n \right\} \quad \text{where} \quad \delta_n := \sqrt{\frac{\theta_0(1-\theta_0)}{n}} \Phi^{-1}(1-\alpha).$$

Indeed, we know by combining Example 2.3 together with (2.8) that for any $\theta \leq \theta_0$ we have

$$\mathbb{P}_\theta[\widehat{\theta}_n > \theta_0 + \delta_n] \leq \mathbb{P}_{\theta_0}[\widehat{\theta}_n > \theta_0 + \delta_n] \to \alpha$$

as $n \to +\infty$, so that $\limsup_n \sup_{\theta \leq \theta_0} \mathbb{P}_\theta[\widehat{\theta}_n > \theta_0 + \delta_n] \leq \alpha$, which provides an asymptotic control of the Type I error of this test: we say that it is *asymptotically of level* α. But what can be said about the *power* of the test ? We know that $\widehat{\theta}_n \overset{\text{as}}{\to} \theta$ under $\mathbb{P}_\theta$ and that $\delta_n \to 0$, so, *under* H_1, namely whenever $\theta > \theta_0$, we have

$$\beta(\theta) = \mathbb{P}_\theta[\widehat{\theta}_n > \theta_0 + \delta_n] \to 1$$

as $n \to +\infty$, which claims that the power of the test goes to 1. In this case, we say that the test is *consistent* or *convergent*.

Remark 2.2 The convergence of $\beta(\theta)$ is not uniform in θ since its limit is discontinuous while $\beta(\theta)$ is continuous (see Equation (2.17)).

2.3.5 Ancillary statistics

An interesting pattern emerges from what we did for confidence intervals and tests. In both cases, for the Bernoulli case, we constructed a statistic $\sqrt{n}(\widehat{\theta}_n - \theta)/\sqrt{\theta(1-\theta)}$ whose asymptotic distribution is $\mathrm{Normal}(0,1)$, namely a distribution that *does not* depend on the parameter θ. This is called an asymptotically *ancillary* statistic.

Definition 2.6 Whenever $X \sim P_\theta$ and the distribution of $f_\theta(X)$ does not depend on θ, we say that $f_\theta(X)$ is an *ancillary* statistic.

The construction of confidence intervals and tests requires such an ancillary or asymptotically ancillary statistic. Indeed, we need to remove the dependence on θ from the distribution in order to compute quantiles allowing to tune the coverage property of a confidence interval, or the level of a test.

2.3.6 Confidence intervals and tests

There is of course a strong connection between confidence intervals and tests, as explained in the following proposition.

Proposition 2.6 If $S(X)$ is a confidence set of level $1-\alpha$, namely $\mathbb{P}_\theta[\theta \in S(X)] \geq 1-\alpha$ for any $\theta \in \Theta$, then the test with rejection set $\{x : S(x) \cap \Theta_0 = \emptyset\}$ is of level α.

This proposition easily follows from the fact that $\mathbb{P}_\theta[S(X) \cap \Theta_0 = \emptyset] \leq \mathbb{P}_\theta[\theta \notin S(X)] \leq \alpha$ for any $\theta \in \Theta_0$. Confidence intervals and

tests are therefore deeply intertwined notions in the sense that when you have built one of the two, you can build easily the other.

Types of hypotheses. For $\Theta \subset \mathbb{R}$, we often consider one of the null hypotheses listed in Table 2.1, where we provide some vocabulary.

$\Theta_0 = \{\theta_0\}$	Simple hypothesis	
$\Theta_0 = [\theta_0, +\infty)$ $\Theta_0 = (-\infty, \Theta_0]$	Multiple hypothesis	One-sided hypothesis
$\Theta_0 = [\theta_0 - \delta, \theta_0 + \delta]$		Two-sided hypothesis

Table 2.1: Some examples of standard null hypotheses.

A test with a one-sided null hypothesis can be obtained using a one-sided confidence interval in the opposite direction of Θ_0. A test with a two-sided null hypothesis can be obtained using a (two-sided) confidence interval. For hypotheses $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ we use $R = \{\widehat{\theta}_n > \theta_0 + c\}$ while for $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ we use $R = \{|\widehat{\theta}_n - \theta_0| > c\}$, where $\widehat{\theta}_n$ is some estimator of θ and where c is a constant to be tuned so that the test has level α. Note that this is a generic recipe, that holds for any statistical model.

In Chapter **??** below, we provide systematic rules to build *optimal* tests[21] in a fairly general setting, but this will require some extra concepts that we will be developed later.

21: tests with maximum power, in some sense

2.3.7 p-values

Consider a statistical model and a test at level α, and keep everything fixed but α. If α is very small, the test has no choice but to accept H_0, since it has almost no slack to eventually be wrong about it.[22] With everything fixed but α, we can expect that for some value $\alpha(X)$ (that depends on the data X), we have that whenever $\alpha < \alpha(X)$ then the test *accepts* H_0 while when $\alpha > \alpha(X)$ the test *rejects* H_0. Such a value $\alpha(X)$ is called the *p-value* of the test.

22: Once again, the only way to build a test with $\alpha = 0$ is to never reject (tell all the patients that they have cancer).

Let R_α be the rejection set of some test at level α, so that it satisfies $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta[R_\alpha] \leq \alpha < \alpha'$ for any $\alpha' > \alpha$, which means that R_α also is a rejection set at level α'. Usually, the family $\{R_\alpha\}_{\alpha \in [0,1]}$ of rejection sets of a test is *increasing* with respect to α, namely $R_\alpha \subset R_{\alpha'}$ for any $\alpha < \alpha'$. In this case, we can define the p-value as follows.

Definition 2.7 Consider a statistical experiment with data X and a statistical test with an increasing family $\{R_\alpha\}_{\alpha \in [0,1]}$ of rejection sets. The p-value of such a test the random variable given by

$$\alpha(X) = \inf\{\alpha \in [0,1] : X \in R_\alpha\}.$$

Let us compute the p-value of one of the tests we built previously for the $\mathrm{Bernoulli}(\theta)$ model and the hypotheses (2.15). The rejection set is given by

$$R_\alpha = \left\{ \widehat{\theta}_n > \theta_0 + \Phi^{-1}(1-\alpha)\sqrt{\frac{\theta_0(1-\theta_0)}{n}} \right\}$$

so that the p-value can be computed as follows:

$$\begin{aligned}\alpha(X) &= \inf\left\{ \alpha \in [0,1] : \widehat{\theta}_n > \theta_0 + \theta_0 + \Phi^{-1}(1-\alpha)\sqrt{\frac{\theta_0(1-\theta_0)}{n}} \right\} \\ &= 1 - \Phi\Big(\sqrt{\frac{n}{\theta_0(1-\theta_0)}}(\widehat{\theta}_n - \theta_0)\Big)\end{aligned}$$

In practice, when performing a statistical testing procedure, we *do not* choose the level α, but we compute the p-value using the definition of the test and the data. A statistical library will never ask you α but will rather give you the value of the p-value. This value quantifies, somehow, *how much we are willing to believe in* H_0. For instance, if $\alpha(x) \leq 10^{-3}$ then we are strongly rejecting H_0, since it would require a level $\alpha < 10^{-3}$ to accept H_0, which is very small. If $\alpha(x) = 3\%$, the result of the test is rather ambiguous while $\alpha(x) = 30\%$ is a strong acceptation of H_0.

x stands for the realization of the random variable X, namely $x = X(\omega)$

In many sciences, in order to publish conclusions based on experimental observations, researchers must exhibit the p-values of the considered statistical tests in order to justify that some effect is indeed observed. However, the reign of the p-value in many fields of science is highly criticized, see for instance [10] .

[10]: Wasserstein et al. (2019), 'Moving to a World Beyond p < 0.05'

2.4 Proofs

Proof of Theorem 2.1. We follow the proof from [11] . First, we can assume without loss of generality that each X_i is centered: it does not change the length $b_i - a_i$ of the interval containing X_i almost surely. We use the Cramér-Chernoff method: because of the Markov's inequality, we have

[11]: Massart (2007), *Concentration inequalities and model selection*

$$\mathbb{P}[S \geq t] = \mathbb{P}[e^{\lambda S} \geq e^{\lambda t}] \leq e^{-\lambda t}\mathbb{E}[e^{\lambda S}]$$

for any $\lambda > 0$. Now, denoting by $\psi_S(\lambda) = \log \mathbb{E}[e^{\lambda S}]$ the log of the moment generating function of $S = \sum_{i=1}^n X_i$, we have thanks to the independence of $X_1, \ldots, X_n$ that

$$\psi_S(\lambda) = \log \mathbb{E}[e^{\lambda \sum_{i=1}^n X_i}] = \sum_{i=1}^n \log \mathbb{E}[e^{\lambda X_i}] = \sum_{i=1}^n \psi_{X_i}(\lambda),$$

so that we need to control $\psi_{X_i}(\lambda)$. Consider a centered random variable X such that $X \in [a,b]$ almost surely and let us prove that

$$\psi_X(\lambda) \leq \frac{(b-a)^2\lambda^2}{8}, \tag{2.18}$$

which is a result known as the *Hoeffding lemma*. Note also that if Y is any random variable such that $Y \in [a,b]$ almost surely, then[23]

$$\mathbb{V}[Y] \leq \frac{(b-a)^2}{4}. \tag{2.19}$$

23: Just remark that $|Y-(a+b)/2| \leq (b-a)/2$ and that $\mathbb{V}[Y] = \mathbb{V}[Y-(a+b)/2] \leq (b-a)^2/4$.

Then, denote as P the distribution of X and introduce the distribution

$$P_\lambda(dx) = e^{-\psi_X(\lambda)}e^{\lambda x}P(dx),$$

so that if X_λ is a random variable with distribution P_λ we have $\mathbb{E}[\phi(X_\lambda)] = \mathbb{E}[\phi(X)e^{-\psi_X(\lambda)}e^{\lambda X}]$. An easy computations gives that the second derivative of ψ_X satisfies

$$\psi_X''(\lambda) = e^{-\psi_X(\lambda)}\mathbb{E}[X^2e^{\lambda X}] - e^{-2\psi_X(\lambda)}(\mathbb{E}[Xe^{\lambda X}])^2 = \mathbb{V}[X_\lambda].$$

But, since $X_\lambda \in [a,b]$ almost surely, we have using (2.19) that $\psi_X''(\lambda) \leq (b-a)^2/4$, so that integration proves (2.18).[24] Most of the work is done now, since wrapping up the inequalities from above gives

24: Integration and the facts that $\psi_X(0) = 0$ and that $\psi_X'(0) = 0$ since X is centered.

$$\mathbb{P}[S \geq t] \leq \exp\Big(-\lambda t + \frac{\lambda^2}{8}\sum_{i=1}^n (b_i-a_i)^2\Big)$$

for any $\lambda > 0$: minimizing the right-hand side with respect to λ allows to conclude for the optimal choice $\lambda = 4t/\sum_{i=1}^n (b_i-a_i)^2$. □

Prof of Theorem 2.3. Consider the neighborhood V of z and define $r(h) = (g(z+h)-g(z))/h - g'(z)$ for $h \neq 0$ such that $z+h \in V$ and put $r(0) = 0$. We know that $r(h) \to 0$ as $h \to 0$. By definition of r we have

$$g(z+h) = g(z) + hg'(z) + hr(h),$$

so putting $h = Z_n - z$ gives

$$a_n(g(Z_n)-g(z)) = a_ng'(z)(Z_n-z) + a_n(Z_n-z)r(Z_n-z). \tag{2.20}$$

Now, we need to use Theorem 2.4 (Slutsky) several times. First, we have $Z_n - z = a_n^{-1}a_n(Z_n-z)$, so that $Z_n - z \xrightarrow{\mathbb{P}} 0$ since $a_n^{-1} \to 0$ and $a_n(Z_n-z) \rightsquigarrow Z$, so that $r(Z_n-z) \xrightarrow{\mathbb{P}} 0$. Second, using again Theorem 2.4, we have $a_n(Z_n-z)r(Z_n-z) \xrightarrow{\mathbb{P}} 0$ since $a_n(Z_n-z) \rightsquigarrow Z$ and $r(Z_n-z) \xrightarrow{\mathbb{P}} 0$. Finally, this allows to conclude that $a_n(g(Z_n)-g(z)) \rightsquigarrow g'(z)Z$ because of (2.20) combined with $a_n(Z_n-z)r(Z_n-z) \xrightarrow{\mathbb{P}} 0$ and Theorem 2.4. □

Prof of Theorem 2.4. Let us first prove that since $Y_n \rightsquigarrow y$ with y deterministic, we actually have that $Y_n \xrightarrow{\mathbb{P}} y$. Indeed, since $Y_n \rightsquigarrow y$ we have that $\mathbb{E}[\phi(Y_n)] \to \mathbb{E}[\phi(y)]$ for any continuous and bounded function ϕ, for instance $\phi(x) = \|x-y\|/(\|x-y\|+1)$ so that we know that

$$\mathbb{E}\Big[\frac{\|Y_n - y\|}{\|Y_n - y\| + 1}\Big] \to 0.$$

Now, we can conclude with the Markov's inequality, since $x \mapsto x/(x+1)$ is increasing on $(0, +\infty)$:

$$\begin{aligned}\mathbb{P}\big[\|Y_n - y\| \geq \varepsilon\big] &= \mathbb{P}\Big[\frac{\|Y_n - y\|}{\|Y_n - y\| + 1} \geq \frac{\varepsilon}{1+\varepsilon}\Big] \\ &\leq \frac{1+\varepsilon}{\varepsilon}\mathbb{E}\Big[\frac{\|Y_n - y\|}{\|Y_n - y\| + 1}\Big] \to 0,\end{aligned}$$

which proves $Y_n \xrightarrow{\mathbb{P}} y$. Now, let us prove that $(X_n, Y_n) \rightsquigarrow (X, y)$. Thanks to the Portemanteau theorem, we know that it suffices to prove that $\mathbb{E}[\phi(X_n, Y_n)] \to \mathbb{E}[\phi(X, y)]$ for any function $\phi : \mathbb{R}^d \times \mathbb{R}^{d'} \to \mathbb{R}$ which is Lipschitz and bounded. We have

$$\begin{aligned}\mathbb{E}\big[|\phi(X_n, Y_n) - \phi(X, y)|\big] \leq{}& \mathbb{E}\big[|\phi(X_n, Y_n) - \phi(X_n, y)|\big] \\ &+ \mathbb{E}\big[|\phi(X_n, y) - \phi(X, y)|\big]\end{aligned}$$

and we already know that $\mathbb{E}\big[|\phi(X_n, y) - \phi(X, y)|\big] \to 0$ since $X_n \rightsquigarrow X$. Moreover, we have

$$|\phi(X_n, Y_n) - \phi(X_n, y)| \leq 2b\mathbf{1}_{\|Y_n - y\| > \varepsilon} + L\varepsilon$$

for any $\varepsilon > 0$, where we used the fact that ϕ is bounded by b and where L is the Lipschitz constant of ϕ. This entails

$$\mathbb{E}\big[|\phi(X_n, Y_n) - \phi(X_n, y)|\big] \leq 2b\mathbb{P}[\|Y_n - y\| > \varepsilon] + L\varepsilon,$$

which allows to conclude since $Y_n \xrightarrow{\mathbb{P}} y$, so that $\mathbb{P}[\|Y_n - y\| > \varepsilon] \to 0$. Finally, $f(X_n, Y_n) \rightsquigarrow f(X, y)$ for any continuous function f, since we know that $\mathbb{E}[\phi(f(X_n, Y_n))] \to \mathbb{E}[\phi(f(X, y))]$ for any continuous and bounded function ϕ, since $\phi \circ f$ is itself continuous and bounded, and since $(X_n, Y_n) \rightsquigarrow (X, y)$. □

Proof of Proposition 2.5. We already know that Point (3) ⇒ Point (1). We have Point (2) ⇒ Point (3) since Point (2) entails that $\{x \in \mathbb{R} : F_Q(x) \geq p\} \subset \{x \in \mathbb{R} : F_P(x) \geq p\}$ for any $p \in [0,1]$, so that $F_P^{-1}(p) \leq F_Q^{-1}(p)$ by definition of the generalized inverse. We have easily Point (4) ⇒ Point (2) by choosing $f_x(t) = \mathbf{1}_{(x,+\infty)}(t)$, which is a non-decreasing and bounded function, so that

$$P[(x, +\infty)] = \int f_x dP \leq \int f_x dQ = Q[(x, +\infty)].$$

Finally, Point (1) $\Rightarrow$ Point (4) by taking $X \sim P$ and $Y \sim Q$ such that $X \leq Y$ almost surely, so that

$$\int f dP = \mathbb{E}[f(X)] = \mathbb{E}[f(X)\mathbf{1}_{X \leq Y}] \leq \mathbb{E}[f(Y)] = \int f dQ$$

for any non-decreasing function f. $\square$

Linear regression 3

We observe iid pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ where $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$. We want to learn to *predict* Y_i from X_i, namely we want to *regress* Y_i on X_i. We know that the closest X_1-measurable function which is the closest from Y_1 in L^2 is the conditional expectation $\mathbb{E}[Y_1|X_1] = f(X_1)$ for some measurable function f, but we do not know the joint distribution of (X_1, Y_1), so we don't know f. Therefore, we want to use the observations (X_i, Y_i) for $i = 1, \dots, n$ in order to build some approximation of f.

What kind of functions should we consider ? We can consider the simplest non-constant function $\mathbb{R}^d \to \mathbb{R}$ that we can think of, which is naturally a *linear* function $x \mapsto x^\top w + b$, hence the name *linear regression model*.

Features engineering

Considering only linear functions is of course very limiting. But let us stress that, in practice, we can do whatever we want with the data $(X_1, Y_1), \dots, (X_n, Y_n)$. A linear model is typically *trained* on *mappings* of X_i, that can include non-linear mappings, such as a *polynomial mapping*, that includes all the pairwise products leading to $d(d-1)/2$ extra coordinates, etc. It is uncommon (and suspicious) to work directly with the *raw* vectors of *features* $X_1, \dots, X_n$: a lot of effort is usually put on the construction of a mapping, that requires knowledge about the data itself. The construction of such a *features mapping* is called *feature engineering* in statistics and machine learning, and is more an art than a science. Many industrial large scale problems (such as web-display advertisement) are handled using simple linear models, but on highly tested and engineered feature mappings. We won't discuss this in this chapter, and will assume that X_i are well-crafted vectors of features on which we want to train a linear model.

Training a linear model means *learning* the *model weights* $w \in \mathbb{R}^d$ and the *intercept* or *population bias* $b \in \mathbb{R}$. To simplify notations we will forget about the intercept from now on, since without loss of generality we can simply put $\theta = [1\ w^\top]^\top$ and replace X_i by $[1\ X_i^\top]^\top$ (and d by $d+1$).

The problem considered here is also known as an instance of *supervised learning* with vectors of *features* $X_i \in \mathbb{R}^d$ and *labels* $Y_i \in \mathbb{R}$.

Large scale means that both n and d are large, when n is large and $n \gg d$ we are considering *big data* while when d is large and $d \gg n$ we work with *high-dimensional* data.

3.1 Ordinary least squares estimator

A linear model assumes that

$$Y_i = X_i^\top \theta + \varepsilon_i$$

for $i = 1, \dots, n$, where $\theta \in \mathbb{R}^d$ must be trained using the data $(X_1, Y_1), \dots, (X_n, Y_n)$ and where the random variables ε_i are called *noise*, and are assumed to satisfy $\mathbb{E}[\varepsilon_i | X_i] = 0$. Also, we will assume from now on that $(X_1, Y_1), \dots, (X_n, Y_n)$ is iid. Let us also consider an independent pair (X, Y) with the same distribution. We want to answer to the question: how can we *estimate* or *train* θ? A natural idea is to find $\widehat{\theta}_n \in \mathbb{R}^d$ such that $X_i^\top \widehat{\theta}_n$ is close to Y_i for each $i = 1, \dots, n$. The simplest way to measure this closeness is to use the Euclidean distance on $\mathbb{R}^n$. Let us first introduce the vector of *labels* $\boldsymbol{y} = [Y_1 \cdots Y_n]^\top \in \mathbb{R}^n$ and the *features* matrix

$$\boldsymbol{X} = \begin{bmatrix} X_{1,1} & \cdots & X_{1,d} \\ \vdots & \ddots & \vdots \\ X_{n,1} & \cdots & X_{n,d} \end{bmatrix} = \begin{bmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{bmatrix} = \begin{bmatrix} X^1 \cdots X^d \end{bmatrix} \in \mathbb{R}^{n \times d},$$

A vector in $\mathbb{R}^n$ is written as a column matrix with shape $n \times 1$ and the norm $\| \cdot \|$ stands for the Euclidean norm. We will write inner products between same-shaped vectors as $u^\top v$ or $\langle u, v \rangle$ depending on what is more convenient.

so that $X_i \in \mathbb{R}^d$ is the i-th row of the features matrix while $X^j \in \mathbb{R}^n$ is the j-th column. We introduce also the vector of noise $\boldsymbol{\varepsilon} = [\varepsilon_1 \cdots \varepsilon_n]^\top \in \mathbb{R}^n$. A *least squares* estimator or *ordinary least squares* estimator is defined as

$$\widehat{\theta}_n \in \operatorname*{argmin}_{t \in \mathbb{R}^d} \| \boldsymbol{y} - \boldsymbol{X}\, t \|^2 = \operatorname*{argmin}_{t \in \mathbb{R}^d} \sum_{i=1}^n (Y_i - X_i^\top t)^2, \qquad (3.1)$$

namely, we consider a vector $\widehat{\theta}_n$ that minimizes[1] the function

$$F(t) = \| \boldsymbol{y} - \boldsymbol{X}\, t \|^2. \qquad (3.2)$$

1: Such a minimizer is not necessarily unique, as explained below

How can we characterize $\widehat{\theta}_n$? The definition of $\widehat{\theta}_n$ given by Equation (3.1) entails that

$$\boldsymbol{X}\, \widehat{\theta}_n = \mathrm{proj}_V(\boldsymbol{y}),$$

where proj_V is the orthogonal projection operator onto $V = \{\boldsymbol{X}\, u : u \in \mathbb{R}^d\} = \mathrm{span}(\boldsymbol{X}) = \mathrm{span}(X^1, \dots, X^d)$, the linear space in $\mathbb{R}^n$ which is spanned by the columns of $\boldsymbol{X}$. This means that $\boldsymbol{y} - \boldsymbol{X}\, \widehat{\theta}_n \perp V$, namely

$$\langle \boldsymbol{X}\, u, \boldsymbol{y} - \boldsymbol{X}\, \widehat{\theta}_n \rangle = u^\top \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\, \widehat{\theta}_n) = 0$$

for any $u \in \mathbb{R}^d$, which is equivalent to the so-called *normal equation*

$$\boldsymbol{X}^\top \boldsymbol{X}\, \widehat{\theta}_n = \boldsymbol{X}^\top \boldsymbol{y}\,. \qquad (3.3)$$

This means that $\widehat{\theta}_n$ is a solution to linear system (3.3).[2] Another

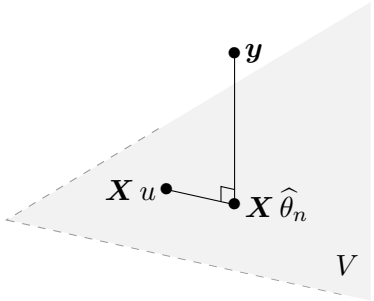


Figure 3.1: Geometric explanation of the normal Equation (3.3) where $V = \mathrm{span}(\boldsymbol{X})$

2: However, from an algorithmic point of view, note that $\widehat{\theta}_n$ is usually *not* computed by solving the linear system, but instead by using an optimization algorithm to minimize the convex function F.

explanation leading to the same characterization is to use the fact F is convex[3] and differentiable on $\mathbb{R}^d$, so that a minimizer must satisfy the first-order condition $\nabla F(t) = 0$ with $\nabla F(t) = 2\,\boldsymbol{X}^\top(\boldsymbol{X}\,t - \boldsymbol{y})$, leading again to Equation (3.3).

3: since its Hessian matrix is positive semidefinite: $\nabla^2 F(t) = \boldsymbol{X}^\top \boldsymbol{X} \succcurlyeq 0$

At this point, let us recall that the covariance matrix between two random vectors U and V (possibly with different dimensions) such that $\mathbb{E}\|U\|^2 < +\infty$ and $\mathbb{E}\|V\|^2 < +\infty$ is given by

$$\mathrm{cov}[U, V] = \mathbb{E}\big[(U - \mathbb{E}U)(V - \mathbb{E}V)^\top\big]$$

The expectation of a vector (or a matrix) is simply the vector (or matrix) containing the expectation of each random entries.

and we will denote $\mathbb{V}[U] = \mathrm{cov}[U, U]$ the covariance matrix of U. Let us also remark that whenever $V = \boldsymbol{A}\,U + b$ for some deterministic matrix $\boldsymbol{A}$ and deterministic vector b, we have that $\mathbb{E}[V] = \boldsymbol{A}\,\mathbb{E}[U] + b$ and $\mathbb{V}[V] = \boldsymbol{A}\,\mathbb{V}[U]\,\boldsymbol{A}^\top$.

From now on, let us assume that $\mathbb{E}\|X\|^2 < +\infty$ and that $\mathbb{E}[Y^2] < +\infty$. This allows to define the $d \times d$ positive semi-definite[4] matrix

$$\mathbb{E}[XX^\top] = (\mathbb{E}[X_j X_k])_{1 \le j,k \le d}. \tag{3.4}$$

4: it is a positive semi-definite matrix since we have $u^\top \mathbb{E}[XX^\top]u = \mathbb{E}[u^\top XX^\top u] = \mathbb{E}[(X^\top u)^2] \ge 0$ for any $u \in \mathbb{R}^d$.

We will assume throughout this section that $\mathbb{E}[XX^\top]$ is invertible, namely that $\mathbb{E}[XX^\top] \succ 0$. The next theorem proves that the uniqueness of the least-squares estimator is equivalent to several equivalent properties on the distribution $\mathbb{P}_X$ of X (the distribution of the features).

Theorem 3.1 Assume that $n \ge d$. The following points about $\mathbb{P}_X$ are all equivalent whenever $X_1, \dots, X_n$ are independent.

1. For any hyperplane $H \subset \mathbb{R}^d$ we have $\mathbb{P}[X \in H] = 0$, namely $\mathbb{P}[X^\top t = 0] = 0$ for any $t \in S^{d-1}$
2. $\boldsymbol{X}^\top \boldsymbol{X} = \sum_{i=1}^n X_i X_i^\top \succ 0$ almost surely
3. The least squares estimator is uniquely defined and given by

$$\widehat{\theta}_n = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\,\boldsymbol{X}^\top \boldsymbol{y}$$

almost surely.

where $S^{d-1} = \{u \in \mathbb{R}^d : \|u\| = 1\}$

Also, whenever $\mathbb{P}_X$ satisfies either of these points, we say that $\mathbb{P}_X$ is *non-degenerate*.

The proof of Theorem 3.1 is given in Section 3.6 below. The non-degenerate assumption stated in Point 1 means that $\mathbb{P}_X$ does not put mass on any hyperplane of $\mathbb{R}^d$. This is a mild assumption: whenever $\mathbb{P}_X \ll \mathrm{Lebesgue}$ then this assumption is satisfied, since $\mathrm{Lebesgue}[H] = 0$ for any hyperplane H. In the next section, we provide some first statistical properties about the least-squares estimator $\widehat{\theta}_n$, under the assumption that $\mathbb{P}_X$ is non-degenerate.

3.2 Properties of the least squares estimator

In this Section we work under the assumption that $\boldsymbol{X}^\top \boldsymbol{X} \succ 0$ almost surely (this is Point 2 of Theorem 3.1), namely under the assumption that $\mathbb{P}_X$ is non-degenerate. So, without loss of generality, and in order to simplify notations, we consider in this section that $\boldsymbol{X}$ is deterministic[5] and such that $\boldsymbol{X}^\top \boldsymbol{X} \succ 0$, so that $\varepsilon_1, \ldots, \varepsilon_n$ are iid and such that $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$. Furthermore, we assume that the noise is *homoscedastic*, namely $\mathbb{V}[\varepsilon_i] = \sigma^2 < +\infty$, which means $\mathbb{V}[\boldsymbol{\varepsilon}] = \sigma^2 \boldsymbol{I}_n$, or equivalently that the covariance of $\boldsymbol{\varepsilon}$ is *isotropic*. In this setting, the least-squares estimator is given by $\widehat{\theta}_n = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ so that

5: If we want to work with random $X_1, \ldots, X_n$ then we just need to replace all expectations by conditional expectation with respect to $X_1, \ldots, X_n$.

$$\mathbb{E}_\theta[\widehat{\theta}_n] = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \mathbb{E}_\theta[\boldsymbol{y}] = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X}\, \theta = \theta \quad (3.5)$$

which means that $\widehat{\theta}_n$ is an *unbiased* estimator. We can write also

$$\mathbb{V}_\theta[\widehat{\theta}_n] = \mathbb{V}_\theta[\boldsymbol{A}\, \boldsymbol{y}] = \boldsymbol{A}\, \mathbb{V}_\theta[\boldsymbol{y}]\, \boldsymbol{A}^\top = \sigma^2\, \boldsymbol{A}\, \boldsymbol{A}^\top = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \quad (3.6)$$

this is because $\mathbb{V}_\theta[\boldsymbol{y}] = \mathbb{V}_\theta[\boldsymbol{X}\, \theta + \boldsymbol{\varepsilon}] = \mathbb{V}[\boldsymbol{\varepsilon}] = \sigma^2 \boldsymbol{I}_n$

where we used $\boldsymbol{A} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$. In particular, this proves that the quadratic risk of $\widehat{\theta}_n$ is given by

$$\mathbb{E}_\theta \|\widehat{\theta}_n - \theta\|^2 = \sigma^2 \operatorname{tr}[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}].$$

Use $\|\widehat{\theta}_n - \theta\|^2 = \|\widehat{\theta}_n - \mathbb{E}_\theta[\widehat{\theta}_n]\|^2$ together with the fact that $\mathbb{E}\|Z - \mathbb{E}Z\|^2 = \operatorname{tr}(\mathbb{V}[Z])$ for a random vector Z such that $\mathbb{E}\|Z\|^2 < +\infty$.

Given $\widehat{\theta}_n$ we can build the vector $\widehat{\boldsymbol{y}}$ of *predictions* and the vector $\widehat{\boldsymbol{\varepsilon}}$ of *residuals* given by

$$\widehat{\boldsymbol{y}} := \boldsymbol{X}\, \widehat{\theta}_n \quad \text{and} \quad \widehat{\boldsymbol{\varepsilon}} := \boldsymbol{y} - \boldsymbol{X}\, \widehat{\theta}_n.$$

Note also that $\widehat{\boldsymbol{y}} = \operatorname{proj}_V(\boldsymbol{y}) = \boldsymbol{H}\, \boldsymbol{y}$ where

$$\boldsymbol{H} := \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$$

is the projection matrix onto V. This matrix is called the *hat matrix* because of the equation $\widehat{\boldsymbol{y}} = \boldsymbol{H}\, \boldsymbol{y}$: it puts a hat on $\boldsymbol{y}$. Also, note that

$$\boldsymbol{y} - \widehat{\boldsymbol{y}} = (\boldsymbol{I}_n - \boldsymbol{H})\, \boldsymbol{y} = (\boldsymbol{I}_n - \boldsymbol{H})(\boldsymbol{X}\, \theta + \boldsymbol{\varepsilon}) = (\boldsymbol{I}_n - \boldsymbol{H})\, \boldsymbol{\varepsilon} \quad (3.7)$$

since $\boldsymbol{I}_n - \boldsymbol{H}$ is the projection matrix onto $V^\perp$ (the orthogonal of V which is of dimension $n - d$ since $\boldsymbol{X}$ is full rank) and since $\boldsymbol{X}\, \theta \in V$ so that

$$\mathbb{E}_\theta \|\, \boldsymbol{y} - \widehat{\boldsymbol{y}}\|^2 = \mathbb{E}\|(\boldsymbol{I}_n - \boldsymbol{H})\, \boldsymbol{\varepsilon}\, \|^2 = \operatorname{tr} \mathbb{V}[(\boldsymbol{I}_n - \boldsymbol{H})\, \boldsymbol{\varepsilon}] = \sigma^2 (n - d),$$

where we used the fact that $(\boldsymbol{I}_n - \boldsymbol{H})(\boldsymbol{I}_n - \boldsymbol{H})^\top = \boldsymbol{I}_n - \boldsymbol{H}$ and $\operatorname{tr}(\boldsymbol{I}_n - \boldsymbol{H}) = n - d$. This proves that the estimator

$$\widehat{\sigma}^2 := \frac{1}{n - d} \|\, \boldsymbol{y} - \widehat{\boldsymbol{y}}\|^2 = \frac{1}{n - d} \|\, \boldsymbol{y} - \boldsymbol{X}\, \widehat{\theta}_n\|^2$$

is an *unbiased* estimator of σ^2 known as the least-squares estimator of the variance.

A quantity often used to quantity the *goodness-of-fit* of a linear model is the R^2, also known as the *coefficient of determination*. Assuming that $\mathbf{1} \in \mathrm{span}(\boldsymbol{X})$, we have by definition of $\widehat{\theta}_n$ that $\boldsymbol{y} - \boldsymbol{X}\,\widehat{\theta}_n \perp \boldsymbol{X}\,\widehat{\theta}_n - \bar{Y}_n\,\mathbf{1}$,[6] so that

$$\|\,\boldsymbol{y} - \bar{Y}_n\,\mathbf{1}\,\|^2 = \|\,\boldsymbol{y} - \boldsymbol{X}\,\widehat{\theta}_n\|^2 + \|\,\boldsymbol{X}\,\widehat{\theta}_n - \bar{Y}_n\,\mathbf{1}\,\|^2$$

and

$$0 \le R^2 := \frac{\|\,\boldsymbol{X}\,\widehat{\theta}_n - \bar{Y}_n\,\mathbf{1}\,\|^2}{\|\,\boldsymbol{y} - \bar{Y}_n\,\mathbf{1}\,\|^2} = 1 - \frac{\|\,\boldsymbol{y} - \boldsymbol{X}\,\widehat{\theta}_n\|^2}{\|\,\boldsymbol{y} - \bar{Y}_n\,\mathbf{1}\,\|^2} \le 1,$$

which corresponds to the proportion of the (empirical) variance of $\boldsymbol{y}$ that is "explained" by the least-squares fit. When R^2 is close to 1, then the linear model fits almost perfectly.[7]

Now, if we want to go further, we need some extra structure, in particular if we want to study the distributions of $\widehat{\theta}_n$ and $\widehat{\sigma}^2$. To do so, we assume in the next section that the noise vector $\boldsymbol{\varepsilon}$ is Gaussian.

6: Where $\bar{Y}_n = n^{-1}\sum_{i=1}^n Y_i$ and $\mathbf{1} \in \mathbb{R}^n$ is the vector with all entries equal to 1

7: This is not necessarily good news, because of the problem of overfitting, that will be discussed later.

3.3 Gaussian linear model

We keep the same setting as in Section 3.2 but furthermore assume that $\varepsilon_1, \ldots, \varepsilon_n$ are iid and that $\varepsilon_i \sim \mathrm{Normal}(0, \sigma^2)$. This means that $\boldsymbol{\varepsilon}$ is a Gaussian vector with multivariate Gaussian distribution $\boldsymbol{\varepsilon} \sim \mathrm{Normal}(0, \sigma^2\,\boldsymbol{I}_n)$. Let us start with some reminders about Gaussian vectors.

Gaussian vectors. We say that a random vector $Z \in \mathbb{R}^n$ is *Gaussian* whenever $\langle u, Z\rangle$ is a Gaussian real random variable for any $u \in \mathbb{R}^d$. In this case, we write $Z \sim \mathrm{Normal}(\mu, \boldsymbol{\Sigma})$ where $\mu = \mathbb{E}[Z]$ and $\boldsymbol{\Sigma} = \mathbb{V}[Z]$. Moreover, if $\boldsymbol{\Sigma} \succ 0$, then Z has density

$$f_Z(z) = \frac{1}{\sqrt{(2\pi)^d \det \boldsymbol{\Sigma}}} \exp\Big(-\frac{1}{2}(z-\mu)^\top \boldsymbol{\Sigma}^{-1}(z-\mu)\Big)$$

with respect to the Lebesgue measure on $\mathbb{R}^n$. If $Z \sim \mathrm{Normal}(0, \boldsymbol{I}_n)$, we say that Z is *standard Gaussian* and note that in this case $\boldsymbol{A}^{1/2}\,Z + b \sim \mathrm{Normal}(b, \boldsymbol{A})$ for any matrix $\boldsymbol{A} \succcurlyeq 0$. Also, if $Z \sim \mathrm{Normal}(\mu, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a diagonal matrix, then the coordinates of Z are independent. Note also that if $Z \sim \mathrm{Normal}(0, \boldsymbol{I}_n)$ and $\boldsymbol{Q}$ is orthonormal then $\boldsymbol{Q}\,Z \sim \mathrm{Normal}(0, \boldsymbol{I}_n)$.

3.3.1 Some classical distributions

In the section, we give some reminders about classical distributions, that will prove useful for the study of the Gaussian linear model.

Gamma distribution. The Gamma distribution $\mathrm{Gamma}(a, \lambda)$, where $a > 0$ is the *shape* and $\lambda > 0$ is the *intensity* has density

$$f_{a,\lambda}(x) = \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x} \mathbf{1}_{x \geq 0}$$

Recall that $\Gamma(a) = \int_0^{+\infty} x^{a-1} e^{-x} dx$ for $a > 0$ and that $\Gamma(a+1) = a\Gamma(a)$.

with respect to the Lebesgue measure on $\mathbb{R}$. If $G \sim \mathrm{Gamma}(a, \lambda)$ then $\mathbb{E}[G] = a/\lambda$ and $\mathbb{V}[G] = a/\lambda^2$ and $\mathrm{mode}(G) = (a-1)/\lambda$ if $a > 1$.[8] Whenever $G_1 \sim \mathrm{Gamma}(a_1, \lambda)$ and $G_2 \sim \mathrm{Gamma}(a_2, \lambda)$ are independent random variable, then $G_1 + G_2 \sim \mathrm{Gamma}(a_1 + a_2, \lambda)$. Also, if $E_1, \ldots, E_n$ are iid distributed as $\exp(\lambda)$ then $\sum_{i=1}^n E_i \sim \mathrm{Gamma}(n, \lambda)$.

8: The mode is defined, whenever it exists, as the argmax of the density. It is therefore a value around which we expect to see most of the observations.

The Chi-squared distribution. If $n \in \mathbb{N} \setminus \{0\}$ then $\mathrm{ChiSq}(n) = \mathrm{Gamma}(n/2, 1/2)$ is called the *Chi-squared distribution with n degrees of freedom*. Although being an instance of the Gamma distribution, the $\mathrm{ChiSq}(n)$ distribution is particularly useful in statistics, in particular since it is the distribution of $\|Z\|^2$ where $Z \sim \mathrm{Normal}(0, \boldsymbol{I}_n)$. This comes from the fact that $Z_i^2 \sim \mathrm{Gamma}(1/2, 1/2)$ so that by independence $\|Z\|^2 = \sum_{i=1}^n Z_i^2 \sim \mathrm{Gamma}(n/2, 1/2) = \mathrm{ChiSq}(n)$. The density of $\mathrm{ChiSq}(n)$ is therefore

$$f_n(x) = \frac{2^{-n/2}}{\Gamma(n/2)} x^{n/2-1} e^{-x/2} \mathbf{1}_{x \geq 0}$$

with respect to the Lebesgue measure on $\mathbb{R}$.

The Student's t distribution. If $U \sim \mathrm{Normal}(0,1)$ and $V \sim \mathrm{ChiSq}(n)$ are independent random variables, then

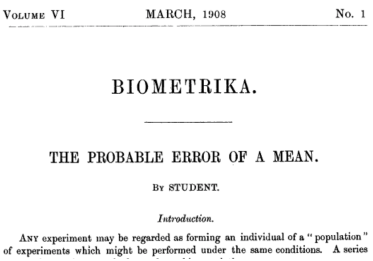
$$\frac{U}{\sqrt{V/n}} \sim \mathrm{Student}(n) \tag{3.8}$$

where $\mathrm{Student}(n)$ is called the *student distribution with n degrees of freedom*[9] which has density

$$f_n(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)} \frac{1}{(1 + x^2/n)^{(n+1)/2}}$$

with respect to the Lebesgue density on $\mathbb{R}$. If $T \sim \mathrm{Student}(n)$ we have $\mathbb{E}[T] = 0$ and $\mathbb{V}(T) = n/(n-2)$ whenever $n > 2$. Also, we have that $\mathrm{Student}(n) \rightsquigarrow \mathrm{Normal}(0,1)$ as $n \to +\infty$ since $V/n \xrightarrow{\mathbb{P}} 1$ (using the law of large numbers and Theorem 2.4).

9: The name "Student" comes from the use of "Student" as a pen name for a research paper by W. William Gosset, a statistician and chemist who worked on stabilizing the taste of the beer at the Guiness factory in Dublin (he used "Student" in order to stay anonymous and keep secret the use of the t-test at the factory).



VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series

The Fisher distribution. Let $p, q \in \mathbb{N} \setminus \{0\}$. If $U \sim \mathrm{ChiSq}(p)$ and $V \sim \mathrm{ChiSq}(q)$ are independent then

$$\frac{U/p}{V/q} \sim \mathrm{Fisher}(p, q) \tag{3.9}$$

where $\mathrm{Fisher}(p, q)$ stands for the *Fisher distribution* with density

$$f_{p,q}(x) = \frac{1}{x\beta(p/2, q/2)} \Big(\frac{px}{px+q}\Big)^{p/2} \Big(1 - \frac{px}{px+q}\Big)^{q/2} \mathbf{1}_{x \geq 0}$$

with respect to the Lebesgue measure on $\mathbb{R}$, where

$$\beta(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 t^{a-1}(1-t)^{b-1} dt. \tag{3.10}$$

The Beta distribution. If $G_1 \sim \mathrm{Gamma}(a, \lambda)$ and $G_2 \sim \mathrm{Gamma}(b, \lambda)$ are independent then

$$\frac{G_1}{G_1 + G_2} \sim \mathrm{Beta}(a, b)$$

where $\mathrm{Beta}(a, b)$ is the Beta distribution with density

$$f_{a,b}(x) = \frac{1}{\beta(a, b)} x^{a-1}(1-x)^{b-1} \mathbf{1}_{[0,1]}(x)$$

with respect to the Lebesgue measure on $\mathbb{R}$, where the β function is given by (3.10). If $B \sim \mathrm{Beta}(a, b)$ then $\mathbb{E}[B] = \frac{a}{a+b}$, $\mathbb{V}[B] = ab/((a+b)^2(a+b+1))$ and $\mathrm{mode}(B) = (a-1)/(a+b-2)$ whenever $a, b > 1$.

3.3.2 Joint distribution of $\widehat{\theta}_n$ and $\widehat{\sigma}^2$ and consequences

In order to study the distribution of $\widehat{\theta}_n$ and $\widehat{\sigma}^2$, we need the following theorem, which proves that the projections onto orthogonal spaces of a Gaussian vector with isometric covariance are independent and Gaussian.

Theorem 3.2 (Cochran theorem) Let $Z \sim \mathrm{Normal}(0, \boldsymbol{I}_n)$ and let $V_1, \ldots, V_k$ be orthogonal linear spaces of $\mathbb{R}^n$. Define the Gaussian vectors $Z_j = \boldsymbol{P}_j Z := \mathrm{proj}_{V_j}(Z)$, where $\boldsymbol{P}_j$ is the orthonormal projection matrix onto V_j. Then, we have that $Z_1, \ldots, Z_k$ are independent Gaussian vectors, and that

$$\|Z_j\|^2 \sim \mathrm{ChiSq}(n_j) \tag{3.11}$$

where $n_j = \dim(V_j)$ (note that $\sum_{j=1}^k n_j \leq n$).

The proof of Theorem 3.2 is given in Section 3.6 below. Let us go back to the Gaussian linear model where $\boldsymbol{y} = \boldsymbol{X}\,\theta + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathrm{Normal}(0, \sigma^2\,\boldsymbol{I}_n)$. We know that $\widehat{\theta}_n$ is a Gaussian vector, as a linear transformation of the Gaussian vector $\boldsymbol{y}$, so that

$$\widehat{\theta}_n \sim \mathrm{Normal}(\theta, \sigma^2(\boldsymbol{X}^\top\,\boldsymbol{X})^{-1})$$

in view of Equations (3.5) and (3.6). Moreover, we know from (3.7) that $\boldsymbol{y} - \boldsymbol{X}\,\widehat{\theta}_n = (\boldsymbol{I}_n - \boldsymbol{H})\,\boldsymbol{\varepsilon} = \mathrm{proj}_{V^\perp}(\boldsymbol{\varepsilon})$ and that $\boldsymbol{X}(\widehat{\theta}_n - \theta) = \mathrm{proj}_V(\boldsymbol{y} - \boldsymbol{X}\,\theta) = \mathrm{proj}_V(\boldsymbol{\varepsilon})$. Since $V \perp V^\perp$, Theorem 3.2 entails that $\boldsymbol{y} - \boldsymbol{X}\,\widehat{\theta}_n$ and $\boldsymbol{X}(\widehat{\theta}_n - \theta)$ are independent, so that $\widehat{\sigma}^2$ and $\widehat{\theta}_n$ are also independent. Moreover, since $\boldsymbol{X}$ is full rank, we have $\dim V = d$ and $\dim V^\perp = n - d$, which entails with Theorem 3.2 that

$$(n-d)\frac{\widehat{\sigma}^2}{\sigma^2} = \|\,\mathrm{proj}_{V^\perp}(\boldsymbol{\varepsilon}\,/\sigma)\|^2 \sim \mathrm{ChiSq}(n-d)$$

and

$$\frac{\|\,\boldsymbol{X}(\widehat{\theta}_n - \theta)\|^2}{\sigma^2} = \|\,\mathrm{proj}_V(\boldsymbol{\varepsilon}\,/\sigma)\|^2 \sim \mathrm{ChiSq}(d).$$

This proves the following theorem.

Theorem 3.3 Assume that $\boldsymbol{X}$ is full rank and that $\boldsymbol{y} = \boldsymbol{X}\,\theta + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathrm{Normal}(0, \sigma^2\,\boldsymbol{I}_n)$. Put $\widehat{\boldsymbol{y}} = \boldsymbol{X}\,\widehat{\theta}_n$ where $\widehat{\theta}_n = (\boldsymbol{X}^\top\,\boldsymbol{X})^{-1}\,\boldsymbol{X}^\top\,\boldsymbol{y}$ and $\widehat{\sigma}^2 = \|\,\boldsymbol{y} - \widehat{\boldsymbol{y}}\|^2/(n-d)$. Then, we have that $\widehat{\theta}_n$ and $\widehat{\sigma}^2$ are *independent* and such that

$$\widehat{\theta}_n \sim \mathrm{Normal}(\theta, \sigma^2(\boldsymbol{X}^\top\,\boldsymbol{X})^{-1}), \quad (n-d)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \mathrm{ChiSq}(n-d)$$

and $\|\,\boldsymbol{X}(\widehat{\theta}_n - \theta)\|^2/\sigma^2 \sim \mathrm{ChiSq}(d)$.

Theorem 3.3 has many consequences for the inference of θ and σ^2 in the Gaussian linear model. If σ^2 is known, the set

$$\mathcal{E} = \Big\{t \in \mathbb{R}^d : \frac{1}{\sigma^2}\|\,\boldsymbol{X}(\widehat{\theta}_n - t)\|^2 \le q_{\mathrm{ChiSq}(d)}(1-\alpha)\Big\}$$

where $q_{\mathrm{ChiSq}(d)}(1-\alpha)$ is the quantile function of the $\mathrm{ChiSq}(d)$ distribution at $1-\alpha$, is a *confidence set*[10] for θ in the Gaussian linear model at level $1-\alpha$, since it satisfies by construction the coverage property $\mathbb{P}_\theta[\theta \in \mathcal{E}] = 1-\alpha$. If σ^2 is unknown (which is always the case), we use the fact[11] that

$$\frac{\|\,\boldsymbol{X}(\widehat{\theta}_n - \theta)\|^2}{d\widehat{\sigma}^2} \sim \mathrm{Fisher}(d, n-d)$$

and consider instead the ellipsoid

$$\Big\{\theta \in \mathbb{R}^d : \frac{1}{d\widehat{\sigma}^2}\|\,\boldsymbol{X}(\widehat{\theta}_n - \theta)\|^2 \le q_{\mathrm{Fisher}(d,n-d)}(1-\alpha)\Big\}, \quad (3.12)$$

which is by construction a confidence set at level $1-\alpha$. Note the

10: we call also $\mathcal{E}$ a confidence *ellipsoid*

11: We proved above that the numerator $\|\,\boldsymbol{X}(\widehat{\theta}_n - \theta)\|^2/\sigma^2$ has $\mathrm{ChiSq}(d)$ distribution while the denominator $(n-d)\widehat{\sigma}^2/\sigma^2$ has $\mathrm{ChiSq}(n-d)$ distribution and that both are independent, so that the definition (3.9) of the Fisher distribution entails the result.

cute trick involved in (3.12): the ratio structure allows to cancel out σ^2, leading to a statistic that does not depend on σ^2, with a *known* distribution.

Confidence intervals. Both previous confidence regions provide coverage for the whole vector $\theta \in \mathbb{R}^d$. We can also build confidence intervals for each coordinate of θ. Indeed, we have $\theta_j = \theta^\top e_j$ where e_j is the canonical basis vector with 1 at coordinate j and 0 elsewhere. More generally, we can build a confidence interval for $a^\top\theta$ for any vector $a \in \mathbb{R}^d$. We know that $a^\top(\widehat{\theta}_n - \theta) \sim \mathrm{Normal}(0, \sigma^2 a^\top(\boldsymbol{X}^\top \boldsymbol{X})^{-1}a)$, so that

$$\frac{a^\top(\widehat{\theta}_n - \theta)}{\sigma\sqrt{a^\top(\boldsymbol{X}^\top \boldsymbol{X})^{-1}a}} \sim \mathrm{Normal}(0,1)$$

and let us recall that $\widehat{\theta}_n$ and $\widehat{\sigma}^2$ are independent and that $(n-d)\widehat{\sigma}^2/\sigma^2 \sim \mathrm{ChiSq}(n-d)$. This entails

$$\frac{a^\top(\widehat{\theta}_n - \theta)}{\sqrt{\widehat{\sigma}^2 a^\top(\boldsymbol{X}^\top \boldsymbol{X})^{-1}a}} \sim \mathrm{Student}(n-d) \tag{3.13}$$

in view of the definition (3.8) of the Student distribution.[12] This proves that the interval

$$I_{a,1-\alpha} = \Big[a^\top\widehat{\theta}_n \pm q_{\mathrm{Student}(n-d)}(1-\alpha/2)\sqrt{\widehat{\sigma}^2 a^\top(\boldsymbol{X}^\top \boldsymbol{X})^{-1}a}\Big],$$

where $q_{\mathrm{Student}(n-d)}$ is the quantile function of the $\mathrm{Student}(n-d)$ distribution, is a confidence interval for $a^\top\theta$ at level $1-\alpha$, since it satisfies $\mathbb{P}_\theta[a^\top\theta \in I_{a,1-\alpha}] = 1-\alpha$ by construction.[13] In particular, for $a = e_j$, we obtain that

$$\Big[(\widehat{\theta}_n)_j \pm q_{\mathrm{Student}(n-d)}(1-\alpha/2)\sqrt{\widehat{\sigma}^2((\boldsymbol{X}^\top \boldsymbol{X})^{-1})_{j,j}}\Big] \tag{3.14}$$

is a confidence interval for θ_j at level $1-\alpha$.

This confidence interval allows to build a test for the hypotheses $H_{0,j} : \theta_j = 0$ versus $H_{1,j} : \theta_j \neq 0$, which can help to quantify the statistical importance of the j-th feature in the considered dataset. Also, a confidence interval for σ^2 can be easily built using the ancillary statistic $(n-d)\widehat{\sigma}^2/\sigma^2 \sim \mathrm{ChiSq}(n-d)$.

Example 3.1 Consider $Y_1, \dots, Y_n$ iid $\mathrm{Normal}(\mu, \sigma^2)$. This is a Gaussian linear model since $\boldsymbol{y} = \mu\,\mathbf{1} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim \mathrm{Normal}(0, \sigma^2 \boldsymbol{I}_n)$ and $\mathbf{1} = [1 \cdots 1]^\top \in \mathbb{R}^n$. We have $\widehat{\mu}_n = \bar{Y}_n$ together with $\widehat{\sigma}^2 = \frac{1}{n-1}\|\,\boldsymbol{y} - \bar{Y}_n\,\mathbf{1}\,\|^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ and we know from Theorem 3.3 that $\widehat{\mu}_n$ and $\widehat{\sigma}^2$ are independent and such that $\sqrt{n}(\widehat{\mu}_n - \mu)/\sigma \sim \mathrm{Normal}(0,1)$ and $(n-1)\widehat{\sigma}^2/\sigma^2 \sim \mathrm{ChiSq}(n-1)$ so that

12: Note again the fact that the ratio structure in (3.13) cancels out σ^2 and that its *exact* distribution is known, thanks to the assumption that the noise $\boldsymbol{\varepsilon}$ is Gaussian.

13: We use the fact that $q_{\mathrm{Student}(k)}(\alpha) = -q_{\mathrm{Student}(k)}(1-\alpha)$ in this construction, since we know that $\mathrm{Student}(k)$ is a symmetrical distribution in view of (3.8).

using $(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top\boldsymbol{y} = n^{-1}\sum_{i=1}^n Y_i$

by definition of the $\mathrm{Student}(n-1)$ distribution we have

$$\sqrt{\frac{n}{\widehat{\sigma}^2}}(\widehat{\mu}_n - \mu) \sim \mathrm{Student}(n-1)$$

so that we can build, using this ancillary statistic, a confidence interval and tests for μ when σ^2 is unknown.

Example 3.2 Consider the *simple* Gaussian linear regression model where $Y_i = ax_i + b + \varepsilon_i$, with $a, b \in \mathbb{R}$, $x_1, \dots, x_n \in \mathbb{R}$ and $\varepsilon_i \sim \mathrm{Normal}(0, \sigma^2)$ iid. This can be written as a linear model with

$$\boldsymbol{y} = \boldsymbol{X}\,\theta + \boldsymbol{\varepsilon} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \boldsymbol{\varepsilon},$$

where we can compute explicitly $\widehat{\theta}_n$ and $\widehat{\sigma}^2$ and obtain their distributions using Theorem 3.3.

Prediction intervals. In the previous paragraph, we built confidence sets and intervals for the parameter $\theta \in \mathbb{R}^d$. But, let us remind ourselves that one of the main usages of the linear model is to provide *predictions* of the label $Y \in \mathbb{R}$ associated to a vector of features $X \in \mathbb{R}^d$. Once the least-squares estimator $\widehat{\theta}_n$ is computed, we predict the unknown label Y_{new} of a *new*[14] feature vector $X_{\mathrm{new}} \in \mathbb{R}^d$ using $\widehat{Y}_{\mathrm{new}} = X_{\mathrm{new}}^\top \widehat{\theta}_n$. If we are willing to assume that the model is Gaussian, namely $\boldsymbol{\varepsilon} \sim \mathrm{Normal}(0, \sigma^2\, \boldsymbol{I}_n)$, and that the unknown label Y_{new} satisfies the same Gaussian linear model $Y_{\mathrm{new}} = X_{\mathrm{new}}^\top \theta + \varepsilon_{\mathrm{new}}$ where $\varepsilon_{\mathrm{new}}$ is independent of $\boldsymbol{\varepsilon}$ and $\varepsilon_{\mathrm{new}} \sim \mathrm{Normal}(0, \sigma^2)$, then we know that $\widehat{Y}_{\mathrm{new}}$ and Y_{new} are independent Gaussian random variables, so that $\widehat{Y}_{\mathrm{new}} - Y_{\mathrm{new}}$ is also Gaussian, and $\mathbb{E}_\theta[\widehat{Y}_{\mathrm{new}} - Y_{\mathrm{new}}] = X_{\mathrm{new}}^\top \mathbb{E}_\theta[\widehat{\theta}_n] - X_{\mathrm{new}}^\top \theta = 0$ and

$$\begin{aligned} \mathbb{V}[\widehat{Y}_{\mathrm{new}} - Y_{\mathrm{new}}] &= \mathbb{V}[\widehat{Y}_{\mathrm{new}}] + \mathbb{V}[Y_{\mathrm{new}}] \\ &= \sigma^2 (X_{\mathrm{new}}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} X_{\mathrm{new}} + 1), \end{aligned}$$

which means that

$$\widehat{Y}_{\mathrm{new}} - Y_{\mathrm{new}} \sim \mathrm{Normal}\left(0, \sigma^2 (1 + X_{\mathrm{new}}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} X_{\mathrm{new}})\right)$$

and using again the fact that $\widehat{\sigma}^2$ and $\widehat{\theta}_n$ are independent and $(n-d)\widehat{\sigma}^2/\sigma^2 \sim \mathrm{ChiSq}(n-d)$ we obtain

$$\frac{\widehat{Y}_{\mathrm{new}} - Y_{\mathrm{new}}}{\sqrt{\widehat{\sigma}^2 (1 + X_{\mathrm{new}}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} X_{\mathrm{new}})}} \sim \mathrm{Student}(n-d)$$

14: In the sense that X_{new} does not belong to the dataset $(X_1, Y_1), \dots, (X_n, Y_n)$ with which $\widehat{\theta}_n$ is trained

so that the interval

$$I_{\rm new}(X_{\rm new}) = \Big[\widehat{Y}_{\rm new} \pm q_{\rm Student(n-d)}(1-\alpha/2) \times \sqrt{\widehat{\sigma}^2(1+X_{\rm new}^\top(\boldsymbol{X}^\top \boldsymbol{X})^{-1}X_{\rm new})}\Big]$$

is a prediction interval at level $1-\alpha$, since we have by construction $\mathbb{P}[Y_{\rm new} \in I_{\rm new}(X_{\rm new})] = 1-\alpha$.

3.3.3 The Fisher test

Using the confidence interval (3.13) we can test $H_0 = \theta_1 = \theta_2$ by using $a = [1, -1, 0, \ldots, 0]$. But, how can we test $H_0 : \theta_1 = \theta_2 = 0$ or more generally a *multiple* null hypothesis such as

$$H_0 : \theta_1 = \cdots = \theta_k = 0 \tag{3.15}$$

for $k = 2, \ldots, d$? If we fix $j \in \{1, \ldots, k\}$ and consider the simple null hypothesis $H_{0,j} : \theta_j = 0$ versus the alternative $H_{1,j} : \theta_j \neq 0$, we know thanks to the confidence interval (3.14) together with Proposition 2.6 that the test with rejection set

$$R_{j,\alpha} = \Big\{|(\widehat{\theta}_n)_j| > q_{\rm Student(n-d)}(1-\alpha/2)\sqrt{\widehat{\sigma}^2((\boldsymbol{X}^\top \boldsymbol{X})^{-1})_{j,j}}\Big\}$$

has level α, namely $\mathbb{P}_{\theta_j=0}[R_{j,\alpha}] = \alpha$, for any $j = 1, \ldots, k$. So, an approach to test the multiple hypothesis H_0 given by (3.15) would be to consider a rejection set given by the *union* of the individual $R_{j,\alpha}$, with a decreased level α/k, since

$$\mathbb{P}_{H_0}\Big[\bigcup_{j=1}^k R_{j,\alpha/k}\Big] \leq \sum_{j=1}^k \mathbb{P}_{\theta_j=0}[R_{j,\alpha/k}] \leq k \times \alpha/k = \alpha,$$

The notation $\mathbb{P}_{H_0}$ means that we compute the probability assuming that H_0 holds, namely $\theta_1 = \cdots = \theta_k = 0$

so that the test with rejection set $\cup_{j=1}^k R_{j,\alpha/k}$ for the null hypothesis (3.15) has indeed level α. This strategy, which relies on a union bound for the construction of a *multiple test* is called the Bonferroni correction.[15] It is the simplest approach for multiple testing, more about multiple tests will follow later in this book.

15: This is called Bonferroni correction, although this strategy is due to Olive Jean Dunn (1915–2008) who worked on statistical testing for biostatistics.

This Bonferroni correction requires to replace the individual levels α of each test by the decreased α/k, where k is the number of null hypotheses to be tested. If k is large, this is a large decrease, and we expect a large deterioration of the power of each individual test. In the Gaussian linear model, we can do much better than this, thanks to the Fisher test.

Let us continue with the null assumption (3.15) and put $\Theta_0 = \{\theta \in \mathbb{R}^d : \theta_1 = \cdots = \theta_k = 0\}$. Let us that recall that $V = \mathrm{span}(\boldsymbol{X}) = \{\boldsymbol{X}\, u : u \in \mathbb{R}^d\}$ and introduce $W = \{\boldsymbol{X}\, u : u \in \Theta_0\}$. Note that

$\theta \in \Theta_0$ means that $\boldsymbol{A}\,\theta = 0$ with $\boldsymbol{A} = [\boldsymbol{I}_k\, \boldsymbol{O}_{k,d-k}]$ corresponding to the horizontal concatenation of the identity matrix on $\mathbb{R}^k$ and a $k \times (d-k)$ zero matrix. More generally, we can consider a multiple testing problem with null hypothesis

$$H_0 : \theta \in \Theta_0 \quad \text{with} \quad \Theta_0 = \ker(\boldsymbol{A}), \tag{3.16}$$

where $\boldsymbol{A}$ is a $k \times d$ matrix of rank k. The idea of the Fisher test is to use the fact that $\theta \in \Theta_0$ means that $\boldsymbol{X}\,\theta$ lives in a linear subset $W \subset V$, of dimension $d - k < d$, and to detect statistically this fact.

The Fisher test uses a geometric solution to this testing problem: we decompose $\mathbb{R}^n$ as the following direct sums

$$\mathbb{R}^n = V^\perp \oplus V = V^\perp \oplus W \oplus W',$$

where we note that $\dim(V^\perp) = n - d$, $\dim(W) = d - k$ and $\dim(W') = k$, where $W = \{\boldsymbol{X}\,\theta : \theta \in \Theta_0\} \subset V$. Consider now the projections $\mathrm{proj}_V(\boldsymbol{y})$ and $\mathrm{proj}_W(\boldsymbol{y})$ of $\boldsymbol{y}$ onto V and its subspace W. Pythagora's theorem entails that

$$\|\,\boldsymbol{y} - \mathrm{proj}_W(\boldsymbol{y})\|^2 = \|\,\boldsymbol{y} - \mathrm{proj}_V(\boldsymbol{y})\|^2 + \|\,\mathrm{proj}_V(\boldsymbol{y}) - \mathrm{proj}_W(\boldsymbol{y})\|^2$$

since $\boldsymbol{y} - \mathrm{proj}_V(\boldsymbol{y}) \perp \mathrm{proj}_V(\boldsymbol{y}) - \mathrm{proj}_W(\boldsymbol{y}) \in V$, so that

$$\|\,\mathrm{proj}_V(\boldsymbol{y}) - \mathrm{proj}_W(\boldsymbol{y})\|^2 = \|\,\boldsymbol{y} - \mathrm{proj}_W(\boldsymbol{y})\|^2 - \|\,\boldsymbol{y} - \mathrm{proj}_V(\boldsymbol{y})\|^2.$$

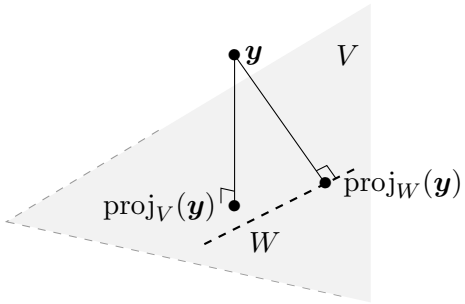


Figure 3.2: Geometric construction of the Fisher test

Recall that $\mathrm{proj}_V(\boldsymbol{y}) = \boldsymbol{X}\,\widehat{\theta}_n$ where $\widehat{\theta}_n$ is the least squares estimator while $\mathrm{proj}_W(\boldsymbol{y}) = \boldsymbol{X}\,\widetilde{\theta}_n$ where $\widetilde{\theta}_n$ is the least squares estimator computed under H_0, namely $\widetilde{\theta}_n = \mathrm{argmin}_{\theta \in \Theta_0} \|\,\boldsymbol{y} - \boldsymbol{X}\,\theta\|^2$. This is where the trick of the test comes into the picture: the quantity $\mathrm{proj}_V(\boldsymbol{y}) - \mathrm{proj}_W(\boldsymbol{y})$ behaves very differently whenever H_0 holds or not. Indeed, *under* H_0, namely when $\boldsymbol{X}\,\theta \in W$, we have

$$\begin{aligned}\mathrm{proj}_V(\boldsymbol{y}) - \mathrm{proj}_W(\boldsymbol{y}) = \mathrm{proj}_{W'}(\boldsymbol{y}) &= \mathrm{proj}_{W'}(\boldsymbol{X}\,\theta) + \mathrm{proj}_{W'}(\boldsymbol{\varepsilon}) \\ &= \mathrm{proj}_{W'}(\boldsymbol{\varepsilon}),\end{aligned}$$

since in this case $\boldsymbol{X}\,\theta \in W \perp W'$, while $\mathrm{proj}_{W'}(\boldsymbol{X}\,\theta) \neq 0$ when $\theta \notin \Theta_0$. So, under H_0, we have that $(\mathrm{proj}_V(\boldsymbol{y}) - \mathrm{proj}_W(\boldsymbol{y}))/\sigma = \mathrm{proj}_{W'}(\boldsymbol{\varepsilon}\,/\sigma)$ while $(\boldsymbol{y} - \mathrm{proj}_V(\boldsymbol{y}))/\sigma = \mathrm{proj}_{V^\perp}(\boldsymbol{\varepsilon}\,/\sigma)$. Therefore, since $V^\perp \perp W'$, Theorem 3.2 together with the definition (3.9) of the Fisher distribution proves that

$$\begin{aligned}\frac{\|\,\mathrm{proj}_V(\boldsymbol{y}) - \mathrm{proj}_W(\boldsymbol{y})\|^2/k}{\|\,\boldsymbol{y} - \mathrm{proj}_V(\boldsymbol{y})\|^2/(n-d)} &= \frac{\|\,\mathrm{proj}_{W'}(\boldsymbol{\varepsilon}\,/\sigma)\|^2/k}{\|\,\mathrm{proj}_{V^\perp}(\boldsymbol{\varepsilon}\,/\sigma)\|^2/(n-d)} \\ &\sim \mathrm{Fisher}(k, n-d)\end{aligned}$$

under the H_0 hypothesis.[16] This can be rewritten, under H_0, as

16: Theorem 3.2 tells us that under H_0, we have that $\|\,\mathrm{proj}_{W'}(\boldsymbol{\varepsilon}\,/\sigma)\|^2 \sim \mathrm{ChiSq}(k)$, that $\|\,\mathrm{proj}_{V^\perp}(\boldsymbol{\varepsilon}\,/\sigma)\|^2 \sim \mathrm{ChiSq}(n-d)$ and that both are independent.

$$\frac{\big(\|\,\boldsymbol{y}-\boldsymbol{X}\,\widetilde{\theta}_n\|^2-\|\,\boldsymbol{y}-\boldsymbol{X}\,\widehat{\theta}_n\|^2\big)/k}{\|\,\boldsymbol{y}-\boldsymbol{X}\,\widehat{\theta}_n\|^2/(n-d)}=\frac{\|\,\boldsymbol{X}(\widehat{\theta}_n-\widetilde{\theta}_n)\|^2}{k\widehat{\sigma}^2}$$
$$\sim \mathrm{Fisher}(k,n-d).$$

Once again, the ratio structure of the ancillary statistic cancels out the unknown σ^2. We can conclude now that the Fisher test with rejection region

$$R_\alpha=\left\{\frac{\|\,\boldsymbol{X}(\widehat{\theta}_n-\widetilde{\theta}_n)\|^2}{k\widehat{\sigma}^2}\geq q_{\mathrm{Fisher}(k,n-d)}(1-\alpha)\right\}$$

has level α for the null hypothesis (3.16), namely that $\sup_{\theta\in\Theta_0}\mathbb{P}_\theta[R_\alpha]=\alpha$. This test is pretty intuitive and can be understood as follows: if $\theta\in\Theta_0$ then both estimators $\widehat{\theta}_n$ and $\widetilde{\theta}_n$ should be close, and $\boldsymbol{X}(\widehat{\theta}_n-\widetilde{\theta}_n)$ should be, consequently, "small", the small miracle being that, in the Gaussian linear model, we can perfectly quantify how small.

Example 3.3 Consider a Gaussian linear model $Y_i=X_i^\top w+b+\varepsilon_i$ where $w\in\mathbb{R}^{d-1}$, where $b\in\mathbb{R}$ is an intercept and the noise $\varepsilon_i\sim\mathrm{Normal}(0,\sigma^2)$ is iid. Using the same notations as before, we can rewrite this as $\boldsymbol{y}=\boldsymbol{X}\,\theta+\boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon}\sim\mathrm{Normal}(0,\sigma^2\,\boldsymbol{I}_n)$, where $\theta=[b,w^\top]^\top\in\mathbb{R}^d$ and where $\boldsymbol{X}=[\mathbf{1}\,X^1\cdots X^{d-1}]$ is assumed to be full-rank. In this model, we wish to test if the features X_i are useful or if a constant intercept is enough to predict Y. Namely, we want to test $H_0:w=0$ versus $H_1:w\neq 0$, namely $H_0:\theta_2=\cdots=\theta_d=0$. This can be done using the Fisher test, putting $W=\mathrm{span}(\mathbf{1})$ so that $\dim(W)=1=d-k$ with $k=d-1$ and $\Theta_0=\{\theta\in\mathbb{R}^d:\theta_2=\cdots=\theta_d=0\}$. Since $\mathrm{proj}_W(\boldsymbol{y})=\bar{Y}_n\,\mathbf{1}_n$, the least-squares estimator under H_0 is $\widetilde{\theta}_n=\bar{Y}_n\,\mathbf{1}_d$, so that the rejection set at level α of the Fisher test writes in this case

$$\left\{\frac{\|\,\boldsymbol{X}\,\widehat{\theta}_n-\bar{Y}_n\,\mathbf{1}_n\,\|^2}{(d-1)\widehat{\sigma}^2}\geq q_{\mathrm{Fisher}(d-1,n-d)}(1-\alpha)\right\},\qquad(3.17)$$

with the same notations as before. The p-value of this test can be therefore used as a quantification of how much the features are informative globally to predict the label using a linear model, versus a constant intercept. This test is known as the *F-test* for linear regression and the statistic used in (3.17) is known as the *F-statistic*.

In practice, you should always include an intercept in a linear model, unless you have a good reason in not doing so.

3.3.4 Analysis of variance

Consider independent random variables $X_{i,j}\sim\mathrm{Normal}(m_i,\sigma^2)$ for $i=1,\ldots,k$ and $j=1,\ldots,n_i$, namely, we observe k Gaussian iid samples with respective sizes $n_1,n_2,\ldots,n_k$, denoted

$$X_{i,\bullet}=[X_{i,1}\cdots X_{i,n_i}]^\top\in\mathbb{R}^{n_i}.$$

The parameters $m = [m_1 \cdots m_k]^\top \in \mathbb{R}^k$ and $\sigma^2 > 0$ are unknown, and we want to build a test for the hypotheses

$$H_0 : m_1 = m_2 = \cdots = m_k \quad \text{against} \quad H_1 : \exists i \neq i' : m_i \neq m_{i'},$$

namely, we want to test if all the samples share the same expectation. We consider the random vector

$$X = [X_{1,\bullet}^\top \cdots X_{k,\bullet}^\top]^\top \in \mathbb{R}^n$$

where $n = \sum_{i=1}^k n_i$, which is the vertical concatenation of the random vectors $X_{1,\bullet}, \ldots, X_{k,\bullet}$. First, we observe that

$$\mu = \mathbb{E}[X] = \big[m_1 \cdots m_1 \, m_2 \cdots m_2 \cdots m_k \cdots m_k\big]^\top \in \mathbb{R}^n$$

belongs to a linear space E of dimension k, since $\mu = \sum_{i=1}^k m_i e_i$ where $e_1 \in \mathbb{R}^n$ is the vector with n_1 first entries equal to 1 and the others equal to 0, e_2 the vector with the first n_1 entries equal to 0, the next n_2 entries equal to 1 and all others equal to 0, up to e_k with n_k last entries equal to 1 and all others 0, so that $E = \operatorname{span}(e_1, \ldots, e_k)$ where the e_i are orthogonal. The orthogonal projection of X onto E is therefore given by

$$X_E := \operatorname{proj}_E(X) = \sum_{i=1}^k \frac{1}{n_i} \langle X, e_i \rangle e_i = \sum_{i=1}^k \bar{X}_{i,\bullet} e_i$$

where $\bar{X}_{i,\bullet} := \frac{1}{n_i}\langle X, e_i\rangle = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{i,j} =$ the average of the i-th sample. The null hypothesis writes $H_0 : \mu \in F$, where $F = \operatorname{span}(\mathbf{1})$ is a linear subspace of E with dimension 1. The orthogonal projection of X onto F is given by

The vector $\mathbf{1} \in \mathbb{R}^n$ has all its entries equal to 1.

$$X_F := \operatorname{proj}_F(X) = \bar{X}\,\mathbf{1}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^k \sum_{j=1}^{n_i} X_{i,j}$ is the average over all the samples.

We can use the Fisher test to test for H_0. We write $X = \mu + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim \operatorname{Normal}(0, \sigma^2 \boldsymbol{I}_n)$ and decompose

$$\mathbb{R}^n = E^\perp \oplus E = E^\perp \oplus F \oplus W.$$

Let us recall that the idea of the Fisher test is to exploit the fact that $X_E - X_F = X_W = \operatorname{proj}_W(\mu + \boldsymbol{\varepsilon}) = \operatorname{proj}_W(\mu) + \operatorname{proj}_W(\boldsymbol{\varepsilon})$ and that *whenever H_0 is true*, namely $\mu \in F$, we have $\operatorname{proj}_W(\mu) = 0$, so that $\frac{1}{\sigma^2}\|X_E - X_F\|^2 = \|\operatorname{proj}_W(\boldsymbol{\varepsilon}/\sigma)\|^2$. Moreover, $X - X_E = \operatorname{proj}_{E^\perp}(X) = \operatorname{proj}_{E^\perp}(\mu + \boldsymbol{\varepsilon}) = \operatorname{proj}_{E^\perp}(\boldsymbol{\varepsilon})$ since $\mu \in E$ so $\frac{1}{\sigma^2}\|X - X_E\|^2 = \|\operatorname{proj}_{E^\perp}(\boldsymbol{\varepsilon}/\sigma)\|^2$. Since $\boldsymbol{\varepsilon}/\sigma \sim \operatorname{Normal}(0, \boldsymbol{I}_n)$ we know from Theorem 3.2 that

$$\|\operatorname{proj}_W(\boldsymbol{\varepsilon}/\sigma)\|^2 \sim \chi^2(k-1) \quad \text{and} \quad \|\operatorname{proj}_{E^\perp}(\boldsymbol{\varepsilon}/\sigma)\|^2 \sim \chi^2(n-k)$$

since $\dim(E^\perp) = n - k$ and $\dim(W) = k - 1$ and that these are independent variables since $E^\perp \perp W$. This proves that

$$T = \frac{\|X_E - X_F\|^2/(k-1)}{\|X - X_E\|^2/(n-k)} \sim \text{Fisher}(k-1, n-k),$$

so we can consider the Fisher test with rejection region

$$R = \{T \geq q_{\text{Fisher}(k-1,n-k)}(1-\alpha)\}.$$

We can rewrite the statistic T in a much more interpretable way. First, we can write

$$\begin{aligned}\|X - X_E\|^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{i,j} - \bar{X}_{i,\bullet})^2 \\ &= \sum_{i=1}^{k} n_i \frac{1}{n_i}\sum_{j=1}^{n_i}(X_{i,j} - \bar{X}_{i,\bullet})^2 =: nV_{\text{intra}},\end{aligned}$$

where V_{intra} is the so-called *intra-class variance*[17] which corresponds to the average of the weighted[18] variances $\frac{1}{n_i}\sum_{j=1}^{n_i}(X_{i,j} - \bar{X}_{i,\bullet})^2$ of each sample $i = 1, \ldots, k$ and second, we have

$$\|X_E - X_F\|^2 = \sum_{i=1}^{k} n_i(\bar{X}_{i,\bullet} - \bar{X})^2 =: nV_{\text{inter}},$$

where V_{inter} is the *inter-class variance* which corresponds to the weighted variance of the averages of each sample. This explains the name ANOVA (ANalysis Of VAriance), since the Fisher test uses here the test statistic

$$T = \frac{V_{\text{inter}}/(k-1)}{V_{\text{intra}}/(n-k)},$$

which is the ratio between the inter and intra-class variances V_{inter} and V_{intra}.

17: A *class* corresponds here to a sample i

18: weighted by the sample proportions n_i/n for $i = 1, \ldots, k$

3.4 Leverages

Let us go back now to the general linear model. We know that the residual vector $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{y}$ is such that $\mathbb{E}[\widehat{\boldsymbol{\varepsilon}}] = 0$ and $\mathbb{V}[\boldsymbol{\varepsilon}] = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$, so that

$$\widehat{\boldsymbol{\varepsilon}}_i \sim \text{Normal}(0, \sigma^2(1 - h_{i,i})) \text{ where } h_{i,i} = \boldsymbol{H}_{i,i} = X_i^\top(\boldsymbol{X}^\top \boldsymbol{X})^{-1}X_i.$$

We know that $h_{i,i} \in [0,1]$ since $\boldsymbol{H}$ is an orthonormal projection matrix.[19] We call $h_{i,i}$ the *leverage score* of sample i. We say that i has small leverage whenever $h_{i,i}$ is close to zero while we say that is as a large leverage when $h_{i,i}$ is close to 1, since in this case the contribution of sample i to the linear model is important, since $\widehat{\boldsymbol{\varepsilon}}_i \approx 0$. Also, we

19: Using $\boldsymbol{H}^2 = \boldsymbol{H}$ and $\boldsymbol{H}^\top\boldsymbol{H}$ we get $h_{i,i} = (\boldsymbol{H}^2)_{i,i} = h_{i,i}^2 + \sum_{i' \neq i} h_{i,i'}^2$ so that $h_{i,i}(1 - h_{i,i}) \geq 0$.

note that

$$h_{i,i} = \frac{\partial \widehat{Y}_i}{\partial Y_i}$$

since $\widehat{\boldsymbol{y}} = \boldsymbol{H}\,\boldsymbol{y}$, namely $\widehat{Y}_i = \sum_{j=1}^{d} \boldsymbol{H}_{i,j} Y_j$. So, the leverage $h_{i,i}$ can be understood as a quantity that measures the "self-sensitivity" to its prediction, namely the influence of Y_i on the computation of $\widehat{Y}_i$. We will see also in the next Section that the leverage score is a very important concept as it is deeply connected to the *theoretical performance* of the least-squares estimation procedure.

3.5 Least squares are minimax optimal

While the previous contents is quite classical and well-known, the results provided in this Section are surprisingly recent and coming from the PhD manuscript of J. Mourtada, see [12, 13] .

[12]: Mourtada (2019), *Contributions to statistical learning: density estimation, expert aggregation and random forests*
[13]: Mourtada (2020), 'Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices'

Let us come back to the general case where $(X_1, Y_1), \dots, (X_n, Y_n)$ are iid with same distribution as (X, Y), with $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ such that $\mathbb{E}\|X\|^2 < +\infty$ and $\mathbb{E}[Y^2] < +\infty$. Also, we assume that $\mathbb{P}_X$ is non-degenerate, as explained in Theorem 3.1 and we assume that

$$\boldsymbol{\Sigma} := \mathbb{E}[XX^\top] \succ 0,$$

namely that $\boldsymbol{\Sigma}$ is invertible. We consider again the linear model (not Gaussian, the results stated here are much more general than that). Given σ^2 and $\mathbb{P}_X$, we consider the following classes of distribution $\mathbb{P}_{X,Y}$ on (X, Y).

Definition 3.1 We consider the set $\mathcal{C}(\mathbb{P}_X, \sigma^2)$ of joint distributions $\mathbb{P}_{X,Y}$ such that $X \sim \mathbb{P}_X$ and

$$Y = X^\top \theta^\star + \varepsilon$$

for some $\theta^\star \in \mathbb{R}^d$, where ε satisfies $\mathbb{E}[\varepsilon|X] = 0$ and $\mathbb{E}[\varepsilon^2|X] \leq \sigma^2$ almost surely. We consider also the set $\mathcal{G}(\mathbb{P}_X, \sigma^2) \subset \mathcal{C}(\mathbb{P}_X, \sigma^2)$ where we assume additionally that $\varepsilon|X \sim \mathrm{Normal}(0, \sigma^2)$.

The set $\mathcal{C}(\mathbb{P}_X, \sigma^2)$ is a general set of joint distributions on (X, Y) with fixed marginal distribution $\mathbb{P}_X$ and such that Y is a linear function of X plus a noise ε which is conditionally centered and with finite variance. The set $\mathcal{G}(\mathbb{P}_X, \sigma^2)$ is the same as $\mathcal{C}(\mathbb{P}_X, \sigma^2)$, but where we assume that ε is centered Gaussian and independent of X.

We consider the quadratic risk

$$R(\theta) := \mathbb{E}[(Y - X^\top \theta)^2] = \int (y - x^\top \theta)^2 \mathbb{P}_{X,Y}(dx, dy).$$

It is easy to see that

$$\theta^\star = \operatorname*{argmin}_{\theta\in\mathbb{R}^d} R(\theta) = \mathbf{\Sigma}^{-1}\,\mathbb{E}[YX].$$

Our aim here is to find an estimator[20] $\widehat{\theta}$ of θ such that the *excess risk*

$$\mathcal{E}(\widehat{\theta}) := R(\widehat{\theta}) - R(\theta^\star)$$

is *minimal*. Note that if $(X,Y)\sim P$ with $P\in\mathcal{C}(\mathbb{P}_X,\sigma^2)$, then

$$\begin{aligned}\mathcal{E}(\theta) &= \mathbb{E}[(Y-X^\top\theta)^2-(Y-X^\top\theta^\star)^2]\\ &= \mathbb{E}[(\theta^\star-\theta)^\top X(2Y-X^\top(\theta+\theta^\star))]\\ &= \mathbb{E}[(\theta^\star-\theta)^\top X(X^\top(\theta^\star-\theta)+2\varepsilon)]\\ &= \mathbb{E}[(\theta^\star-\theta)^\top XX^\top(\theta^\star-\theta)]\\ &= \|\theta^\star-\theta\|_{\mathbf{\Sigma}}^2,\end{aligned}$$

We use here the fact that $Y = X^\top\theta^\star+\varepsilon$ and that $\mathbb{E}[\varepsilon|X]=0$ almost surely.

where we introduced $\|x\|_{\mathbf{\Sigma}}^2 = x^\top\mathbf{\Sigma}\,x$, which is a norm since we assumed $\mathbf{\Sigma}\succ 0$. Whenever $(X,Y)\sim P$ with $P\in\mathcal{C}(\mathbb{P}_X,\sigma^2)$ we will therefore write

$$\mathcal{E}(\theta) = R(\theta)-R(\theta^\star) = \|\theta-\theta^\star\|_{\mathbf{\Sigma}}^2$$

and whenever $\widehat{\theta}$ depends on the data $(X_1,Y_1),\dots,(X_n,Y_n)$, we can consider, since (X,Y) is an independent copy with the same distribution,

$$\mathbb{E}[\mathcal{E}(\widehat{\theta})]$$

where this expectation is with respect to $P^{\otimes n}$, for the randomness coming from the data. We can consider now the *minimax risk* for a set $\mathcal{P}$ of distributions:

$$\inf_{\widehat{\theta}}\sup_{P\in\mathcal{P}}\mathbb{E}[\mathcal{E}(\widehat{\theta})].$$

The infimum is taken over any possible estimator, namely any statistic of the data, while the sup is over all distributions in $\mathcal{P}$. Hence the name minimax, since we look at the worst-case excess risk over the considered set $\mathcal{P}$, but we consider the best possible estimator (with the inf). Since $\mathcal{G}(\mathbb{P}_X,\sigma^2)\subset\mathcal{C}(\mathbb{P}_X,\sigma^2)$, the minimax risk of the former is smaller than the one of the latter.

Some remarks and extra notations are required before we can state the main result of the section.

- The linear model is *well-specified* here, since we assume that $Y = X^\top\theta^\star+\varepsilon$ almost surely with $\mathbb{E}[\varepsilon|X]=0$, so that there is no approximation term of $\mathbb{E}[Y|X]$ by $X^\top\theta^\star$.
- For the class $\mathcal{P}=\mathcal{C}(\mathbb{P}_X,\sigma^2)$, we expect a *minimax estimator* $\widehat{\theta}$ to depend both on $\mathbb{P}_X$ and σ^2. Quite surprisingly, we will see that it is not the case.

A minimax estimator is an estimator achieving the minimax risk.

20: Namely, a measurable function of $(X_1,Y_1),\dots,(X_n,Y_n)$.

Let us introduce

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top = \frac{1}{n}\,\boldsymbol{X}^\top \boldsymbol{X}$$

and let us introduce also the "whitened" random vectors $\widetilde{X}_i = \mathbf{\Sigma}^{-1/2} X_i$ (so that $\mathbb{E}[\widetilde{X}_i(\widetilde{X}_i)^\top] = \boldsymbol{I}_d$) and define

$$\widetilde{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_i \widetilde{X}_i^\top = \mathbf{\Sigma}^{-1/2}\,\widehat{\mathbf{\Sigma}}\,\mathbf{\Sigma}^{-1/2}\,.$$

The following theorem holds.

Theorem 3.4 Assume that $\mathbb{P}_X$ is non-degenerate, that $n \geq d$ and that $\sigma^2 > 0$. Then

$$\begin{aligned}\inf_{\widehat{\theta}} \sup_{P\in\mathcal{C}(\mathbb{P}_X,\sigma^2)} \mathbb{E}[\mathcal{E}(\widehat{\theta})] &= \inf_{\widehat{\theta}} \sup_{P\in\mathcal{G}(\mathbb{P}_X,\sigma^2)} \mathbb{E}[\mathcal{E}(\widehat{\theta})] \\ &= \frac{\sigma^2}{n}\mathbb{E}[\mathrm{tr}(\widetilde{\mathbf{\Sigma}}^{-1})].\end{aligned} \tag{3.18}$$

Furthermore, the infimum in the minimax risk is achieved by the ordinary least squares estimator (3.1).

The proof of Theorem 3.4 is done in two steps. The first step, which proves that the ordinary least squares estimator satisfies the upper bound in (3.18), is given in Section 3.6 below. Since the proof of the lower bound requires extra tools from Bayesian statistics, it will be provided in Section 4.6 of Chapter 4.

This theorem deserves several remarks.

- The theorem proves that the least-squares estimator is, in a fairly general setting, *minimax optimal*: it cannot be improved by another estimator, uniformly over the set of distributions $\mathcal{C}(\mathbb{P}_X, \sigma^2)$.
- The Gaussian noise, namely the class $\mathcal{G}(\mathbb{P}_X, \sigma^2)$ "saturates" the minimax risk, and corresponds to the *least favorable* distribution in the minimax sense.
- The minimax risk is invariant by a linear transformation of the features vectors: it is unchanged if one replaces X_i by $X_i' = \boldsymbol{A}\,X_i$ for some deterministic invertible matrix $\boldsymbol{A}$. Indeed we have in this case $\widehat{\mathbf{\Sigma}}' = \frac{1}{n}\sum_{i=1}^n X_i' X_i'^\top = \boldsymbol{A}\,\widehat{\mathbf{\Sigma}}\,\boldsymbol{A}^\top$ so that $(\widehat{\mathbf{\Sigma}}')^{-1}\,\mathbf{\Sigma}' = (\boldsymbol{A}^\top)^{-1}(\widehat{\mathbf{\Sigma}})^{-1}\,\mathbf{\Sigma}\,\boldsymbol{A}^\top$ which proves that $(\widehat{\mathbf{\Sigma}}')^{-1}\,\mathbf{\Sigma}'$ and $(\widehat{\mathbf{\Sigma}})^{-1}\,\mathbf{\Sigma}$ are congruent matrices, so that they share the same trace, namely $\mathrm{tr}((\widehat{\mathbf{\Sigma}}')^{-1}\,\mathbf{\Sigma}') = \mathrm{tr}((\widehat{\mathbf{\Sigma}})^{-1}\,\mathbf{\Sigma})$, and the minimax risk is indeed invariant when replacing X_i by $\boldsymbol{A}\,X_i$. This is of course expected, since the supremum is over linear functions.

A lower bound for $\sigma^2\mathbb{E}[\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1})]/n$ can be easily obtained thanks to the following proposition.

Proposition 3.5 The function $\boldsymbol{A} \mapsto \mathrm{tr}(\boldsymbol{A}^{-1})$ is convex on the cone of positive definite matrices.

The proof of Proposition 3.5 is given in Section 3.6 below. By combining Proposition 3.5 and Jensen's inequality, we obtain

$$\mathbb{E}[\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1})] \geq \mathrm{tr}(\mathbb{E}[\widetilde{\boldsymbol{\Sigma}}]^{-1}),$$

but $\mathbb{E}[\widetilde{\boldsymbol{\Sigma}}] = \boldsymbol{\Sigma}^{-1/2}\,\mathbb{E}[\widehat{\boldsymbol{\Sigma}}]\,\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{I}_d$, so the lower bound

$$\mathbb{E}[\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1})] \geq d$$

holds, and consequently the minimax risk satisfies

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{C}(\mathbb{P}_X, \sigma^2)} \mathbb{E}[\mathcal{E}(\widehat{\theta})] = \frac{\sigma^2}{n}\mathbb{E}[\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1})] \geq \sigma^2\frac{d}{n}. \tag{3.19}$$

We can also provide another expression for $\mathbb{E}[\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1})]$ using the leverage scores we discussed in Section 3.4. Let us recall at this point that since $\mathbb{P}_X$ is non-degenerate, and if $X_1, \dots X_{n+1}$ are iid and distributed as $\mathbb{P}_X$, we have $\sum_{i=1}^{n+1} X_i X_i^\top \succ 0$.

Theorem 3.6 Under the same assumptions as that of Theorem 3.4, the minimax risk can be written as

$$\frac{1}{n}\mathbb{E}[\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1})] = \mathbb{E}\Big[\frac{\widehat{\ell}_{n+1}}{1-\widehat{\ell}_{n+1}}\Big]$$

where $\widehat{\ell}_{n+1}$ is the leverage of one data point among $n+1$ given by

$$\widehat{\ell}_{n+1} = X_{n+1}^\top\Big(\sum_{i=1}^{n+1} X_i X_i^\top\Big)^{-1} X_{n+1},$$

where $X_1, \dots, X_n, X_{n+1}$ are iid with distribution $\mathbb{P}_X$.

The proof of Theorem 3.6 is given in Section 3.6 below. Let us recall that $\widehat{\ell}_{n+1} = \partial\widehat{Y}_{n+1}/\partial Y_{n+1}$ where $\widehat{Y}_{n+1} = X_{n+1}^\top\widehat{\theta}_{n+1}$ where $\widehat{\theta}_{n+1}$ is the ordinary least squares estimator computed on the $n+1$ samples $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1})$. This theorem entails that the minimax risk, which measures the complexity of the estimation problem, is completely determined by the leverage score. Even more than that, it is the expected value of a convex function of $\widehat{\ell}_{n+1}$, so that the minimax risk is small when $\widehat{\ell}_{n+1}$ is small, and gets larger with a large $\widehat{\ell}_{n+1}$, which is natural since in such a case, the regression problem is more difficult.

A corollary of Theorem 3.6 is an improved lower bound compared to the previous $\sigma^2 d/n$.

Corollary 3.7 Under the same assumptions as that of Theorem 3.4, we have that the minimax risk satisfies

$$\frac{1}{n}\mathbb{E}[\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1})] = \mathbb{E}\Big[\frac{\widehat{\ell}_{n+1}}{1-\widehat{\ell}_{n+1}}\Big] \geq \sigma^2 \frac{d}{n-d+1}.$$

The proof can be found in Section 3.6. The lower bound $\sigma^2 d/(n-d+1)$ is very sharp since it can be seen that

$$\mathbb{E}[\mathcal{E}(\widehat{\theta}_n)] = \sigma^2 \frac{d}{n-d-1}$$

whenever $\mathbb{P}_X = \mathrm{Normal}(0, \boldsymbol{\Sigma})$, if $\widehat{\theta}_n$ is the ordinary least squares estimator. This comes from the study of the Wishart distribution, which is the distribution of $\boldsymbol{X}^\top \boldsymbol{X}$ when $X \sim \mathrm{Normal}(0, \boldsymbol{\Sigma})$.[21] This result means that the Gaussian *design* $\mathrm{Normal}(0, \boldsymbol{\Sigma})$ is, almost, the most favorable design for linear regression, since for this distribution, the minimax risk is almost minimal (compare the denominators $n-d-1$ and $n-d+1$).[22]

21: We won't pursue further about Wishart distributions.

22: Finding the most favorable distribution is, up to our knowledge, an open problem. We conjecture that it is given by the uniform distribution on the unit sphere of $\mathbb{R}^d$.

We were able to provide a lower bound for $\mathbb{E}[\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1})]$ that is explicit with respect to d, n and σ^2. Now, it remains to provide a similarly explicit upper bound for this quantity. This requires extra assumptions on $\mathbb{P}_X$.

The first assumption is a "quantified" version of the non-degenerate assumption about $\mathbb{P}_X$, see Theorem 3.1. Indeed, we assume that there is $\alpha \in (0, 1]$ and $C \geq 1$ such that

$$\mathbb{P}[|X^\top \theta| \leq t \|\theta\|_{\boldsymbol{\Sigma}}] \leq (Ct)^\alpha \tag{3.20}$$

for any $t > 0$ and any non-zero vector $\theta \in \mathbb{R}^d$. This is equivalent to the assumption that $\mathbb{P}[|\widetilde{X}^\top \theta| \leq t] \leq (Ct)^\alpha$ for any $\theta \in S^{d-1}$ where we recall that $\widetilde{X} = \boldsymbol{\Sigma}^{-1/2} X$. This assumption "quantifies" the assumption $\mathbb{P}[X^\top \theta = 0] = 0$. The second assumption about $\mathbb{P}_X$ requires that

$$\mathbb{E}[\| \boldsymbol{\Sigma}^{-1/2} X\|^4] \leq \kappa d^2. \tag{3.21}$$

This is entailed[23] by the condition $\mathbb{E}[(X^\top \theta)^4]^{1/4} \leq \kappa \mathbb{E}[(X^\top \theta)^2]^{1/2}$ for any $\theta \in \mathbb{R}^d$.

23: Indeed, we have $X^\top \theta = \widetilde{X}_j$ for the choice $\theta = \boldsymbol{\Sigma}^{-1/2} e_j$, so that $\mathbb{E}[\widetilde{X}_j^4] \leq \mathbb{E}[\widetilde{X}_j^2]^2 = \kappa$, where we used $\mathbb{E}[\widetilde{X}\widetilde{X}^\top] = \boldsymbol{I}_d$. This entails that $\mathbb{E}\|\widetilde{X}\|^4 = \sum_{1 \leq j,k \leq d} \mathbb{E}[\widetilde{X}_j^2 \widetilde{X}_k^2] \leq \sum_{j,k} \sqrt{\mathbb{E}[\widetilde{X}_j^4]\mathbb{E}[\widetilde{X}_k^4]} \leq \kappa d^2$.

Theorem 3.8 Assume that X satisfies (3.20) and (3.21) and put $C' = 3C^4 e^{1+9/\alpha}$. Then, if $n \geq \max(6d/\alpha, 12\log(12/\alpha)/\alpha)$, we have

$$\frac{1}{n}\mathbb{E}\,\mathrm{tr}[(\widetilde{\boldsymbol{\Sigma}})^{-1}] \leq \frac{d}{n} + 8C'\kappa\Big(\frac{d}{n}\Big)^2.$$

This entails, together with Theorem 3.4 and the lower bound (3.19),

that

$$\sigma^2 \frac{d}{n} \leq \inf_{\widehat{\theta}} \sup_{P \in \mathcal{C}(\mathbb{P}_X, \sigma^2)} \mathbb{E}[\mathcal{E}(\widehat{\theta})] \leq \sigma^2 \frac{d}{n} \Big(1 + 8C' \kappa\Big(\frac{d}{n}\Big)\Big).$$

The proof of such an explicit upper bound is quite technical and somewhat beyond the scope of this book. It can be found in [13] .

[13]: Mourtada (2020), 'Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices'

Let us wrap up some of the nice things that we learned in this Section.

- Ordinary least-squares are minimax optimal for the well-specified linear regression model. This means that no other statistical procedure can perform (uniformly) better than this simple procedure.
- The Gaussian design is almost the most favorable one in the minimax sense.
- The minimax rate is exactly of order $\sigma^2 d/n$ under mild assumptions on $\mathbb{P}_X$.
- The statistical complexity of the linear regression problem is, when measured by the minimax risk, fully explained by the distribution a leverage scores of one sample among $n+1$.

3.6 Proofs

3.6.1 Proof of Theorem 3.1

Point (2) $\Leftrightarrow$ Point (3) is obvious since $\boldsymbol{X}^\top \boldsymbol{X} \succ 0$ entails that $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible. Point (2) $\Rightarrow$ Point (1) comes from a proof by contradiction. If $0 < p = \mathbb{P}[X^\top u = 0]$ then $X_i^\top u = 0$ for all $i = 1, \ldots, n$ with a probability $p^n > 0$ since $X_1, \ldots, X_n$ are iid, so that $\boldsymbol{X}^\top \boldsymbol{X}\, \theta = \sum_{i=1}^n (X_i^\top \theta) X_i = 0$ and $\boldsymbol{X}^\top \boldsymbol{X}$ cannot be invertible almost surely. The proof of Point (1) $\Rightarrow$ Point (2) can be done by recurrence. We first remark that $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible if and only if $\mathrm{span}(X_1, \ldots X_n) = \mathbb{R}^d$.[24] We will show that $\mathrm{span}(X_1, \ldots, X_d) = \mathbb{R}^d$ almost surely by recurrence. We put $V_k = \mathrm{span}(X_1, \ldots, X_k)$ so that $\dim(V_k) \leq k \leq d$. For $k = 1$ we do have $\dim V_1 = 1$ so it is OK. Now, assume that $\dim(V_{k-1}) = k - 1$. We have that X_k is independent from $V_{k-1} = \mathrm{span}(X_1, \ldots, X_{k-1})$ and $\dim(V_{k-1}) = k - 1 < d$ so that $V_{k-1} \subset H$ where $H \subset \mathbb{R}^d$ is an hyperplane. So, we have again by independence that $\mathbb{P}[X_k \in V_{k-1}] = \mathbb{P}[X_k \in V_{k-1} | X_1, \ldots, X_{k-1}] \leq \mathbb{P}[X_k \in H] = 0$ using Point (1). So, $X_k \notin V_{k-1}$ almost surely, and $\dim(V_k) = k$ almost surely. □

24: Indeed, $\ker(\boldsymbol{X}^\top \boldsymbol{X}) = \ker(\boldsymbol{X})$ so that $\boldsymbol{X}^\top \boldsymbol{X}\, u = 0 \Leftrightarrow \boldsymbol{X}\, u = 0 \Leftrightarrow X_i^\top \theta = 0$ for all $i = 1, \ldots, n$.

3.6.2 Proof of Theorem 3.2

We have $\mathbb{V}[Z_j] = \boldsymbol{P}_j \boldsymbol{P}_j^\top = \boldsymbol{P}_j$ since $\boldsymbol{P}_j$ is an orthogonal projection matrix and $Z_j = \boldsymbol{P}_j\, Z$, which entails that Z_j is a Gaussian vector[25] and

25: as a linear transformation of a Gaussian vector

that $Z_j \sim \mathrm{Normal}(0, \boldsymbol{P}_j)$. Note that Z_j has no density with respect to the Lebesgue measure, it is a random vector on $\mathbb{R}^n$ which belongs to linear subspace of dimension $n_j < n$. Now, we have

$$\mathrm{cov}[Z_j, Z_{j'}] = \mathrm{cov}[\boldsymbol{P}_j\, Z, \boldsymbol{P}_{j'}\, Z] = \boldsymbol{P}_j\, \boldsymbol{P}_{j'}^\top = \boldsymbol{O}$$

The matrix $\boldsymbol{O}$ stands for the matrix with all entries equal to 0.

since $V_j \perp V_{j'}$, so that Z_j and $Z_{j'}$ are independent random vectors, because $[Z_j^\top Z_{j'}^\top]^\top$ is a Gaussian vector with a block diagonal covariance matrix. This proves that the $Z_1, \ldots, Z_k$ are independent random vectors. Finally, since $\boldsymbol{P}_j$ is an orthogonal projection matrix onto a space of dimension n_j, we can decompose it as $\boldsymbol{P}_j = \boldsymbol{Q}\, \boldsymbol{D}_{n_j}\, \boldsymbol{Q}^\top$ where $\boldsymbol{D}_{n_j} = \mathrm{diag}[1, \ldots, 1, 0, \ldots, 0]$ is the diagonal matrix with first n_j diagonal elements equal to 1 and all others equal to 0 and where $\boldsymbol{Q}$ is an orthonormal matrix. We know that $Z' := \boldsymbol{Q}^\top Z \sim \mathrm{Normal}(0, \boldsymbol{I}_n)$, so $\boldsymbol{P}_j\, Z = \boldsymbol{Q}[Z'_1 \cdots Z'_{n_j}]^\top =: \boldsymbol{Q}\, Z'_-$ and $\| \boldsymbol{P}_j\, Z\|^2 = \| \boldsymbol{Q}\, Z''_-\|^2 = \|Z''_-\|^2$ (since $\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}_n$), so that finally, we have $\| \boldsymbol{P}_j\, Z\|^2 = \sum_{j=1}^{n_j} (Z'_j)^2$. This proves that $\| \boldsymbol{P}_j\, Z\|^2 \sim \mathrm{ChiSq}(n_j)$ since $Z' \sim \mathrm{Normal}(0, \boldsymbol{I}_n)$. □

3.6.3 Proof of Proposition 3.5

Put $f(\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{A}^{-1})$ for $\boldsymbol{A} \succ 0$ and consider $\boldsymbol{A}, \boldsymbol{B} \succ 0$ and $\alpha \in [0, 1]$. We write

$$f(\alpha\, \boldsymbol{A} + (1-\alpha)\, \boldsymbol{B}) = f(\boldsymbol{A} + (1-\alpha)\, \boldsymbol{D}) = g(1-\alpha),$$

where we defined $g_{\boldsymbol{A},\boldsymbol{D}}(u) = f(\boldsymbol{A} + u\, \boldsymbol{D})$ for $u \in [0, 1]$ and $\boldsymbol{D} = \boldsymbol{B} - \boldsymbol{A}$. First, let us prove that $g''_{\boldsymbol{A},\boldsymbol{D}}(0) \geq 0$ for any $\boldsymbol{A} \succcurlyeq 0$ and symmetric $\boldsymbol{D}$. Indeed, we have using a Taylor expansion that

$$\begin{aligned}(\boldsymbol{A} + \varepsilon\, \boldsymbol{D})^{-1} &= (\boldsymbol{A}(\boldsymbol{I} + \varepsilon\, \boldsymbol{A}^{-1}\, \boldsymbol{D}))^{-1} \\ &= \boldsymbol{A}^{-1} - \varepsilon\, \boldsymbol{A}^{-1}\, \boldsymbol{D}\, \boldsymbol{A}^{-1} + \varepsilon^2 (\boldsymbol{A}^{-1}\, \boldsymbol{D})^2\, \boldsymbol{A}^{-1} + \cdots\end{aligned}$$

where $\varepsilon > 0$ is small enough and $\cdots$ contains terms of order $O(\varepsilon^3)$. So, we have that

$$g''_{\boldsymbol{A},\boldsymbol{D}}(0) = 2\, \mathrm{tr}((\boldsymbol{A}^{-1}\, \boldsymbol{D})^2\, \boldsymbol{A}^{-1}) = 2\, \mathrm{tr}(\boldsymbol{C}\, \boldsymbol{A}^{-1}\, \boldsymbol{C}^\top)$$

where $\boldsymbol{C} = \boldsymbol{A}^{-1}\, \boldsymbol{D}$. But $\boldsymbol{A}^{-1} \succcurlyeq 0$ so $\boldsymbol{C}\, \boldsymbol{A}^{-1}\, \boldsymbol{C}^\top \succcurlyeq 0$ and $g''_{\boldsymbol{A},\boldsymbol{D}}(0) \geq 0$. But

$$\begin{aligned}g''_{\boldsymbol{A},\boldsymbol{D}}(u) &= \frac{\partial^2}{\partial \varepsilon^2} g_{\boldsymbol{A},\boldsymbol{D}}(u + \varepsilon) = \frac{\partial^2}{\partial \varepsilon^2} f(\boldsymbol{A} + (u + \varepsilon)\, \boldsymbol{D}) \\ &= g''_{\boldsymbol{A} + u\, \boldsymbol{D},\boldsymbol{D}}(0) \geq 0\end{aligned}$$

since $\boldsymbol{A} + u\,\boldsymbol{D} = (1-u)\,\boldsymbol{A} + u\,\boldsymbol{B} \succcurlyeq 0$. This proves that $g_{\boldsymbol{A},\boldsymbol{D}} : [0,1] \to \mathbb{R}^+$ is convex, which allows to conclude since

$$\begin{aligned} f(\alpha\,\boldsymbol{A} + (1-\alpha)\,\boldsymbol{B}) &= g(1-\alpha) = g(\alpha\cdot 0 + (1-\alpha)\cdot 1) \\ &\le \alpha g(0) + (1-\alpha) g(1) \\ &= \alpha f(\boldsymbol{A}) + (1-\alpha) f(\boldsymbol{B}). \end{aligned}$$

3.6.4 Proof of Theorem 3.4: the upper bound

This is actually mainly a computation with no particular tricks. Recall that (X, Y) is such that $Y = X^\top \theta^\star + \varepsilon$ with $\mathbb{E}[\varepsilon | X] = 0$ and $\mathbb{E}[\varepsilon^2 | X] \le \sigma^2$. Thanks to Theorem 3.1, we know that the ordinary least squares estimator $\widehat{\theta}$ satisfies

$$\begin{aligned} \widehat{\theta} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{X}\,\theta^\star + \boldsymbol{\varepsilon}) \\ &= \theta^\star + \widehat{\boldsymbol{\Sigma}}^{-1} \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \end{aligned}$$

so that recalling $\langle u, v \rangle_{\boldsymbol{\Sigma}} = u^\top \boldsymbol{\Sigma}\, v$ and $\|u\|_{\boldsymbol{\Sigma}}^2 = \langle u, u \rangle_{\boldsymbol{\Sigma}}$ we have

$$\begin{aligned} \mathbb{E}[\mathcal{E}(\widehat{\theta})] &= \mathbb{E}\Big\| \widehat{\boldsymbol{\Sigma}}^{-1} \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \Big\|_{\boldsymbol{\Sigma}}^2 = \frac{1}{n^2} \sum_{1 \le i, i' \le n} \mathbb{E} \langle \widehat{\boldsymbol{\Sigma}}^{-1} \varepsilon_i X_i, \widehat{\boldsymbol{\Sigma}}^{-1} \varepsilon_{i'} X_{i'} \rangle \\ &= \frac{1}{n^2} \sum_{1 \le i, i' \le n} \mathbb{E}\Big[\mathbb{E}[\varepsilon_i \varepsilon_{i'} | X_1, \dots, X_n] \langle \widehat{\boldsymbol{\Sigma}}^{-1} X_i, \widehat{\boldsymbol{\Sigma}}^{-1} X_{i'} \rangle \Big]. \end{aligned}$$

But we have $\mathbb{E}[\varepsilon_i \varepsilon_{i'} | X_1, \dots, X_n] = 0$ whenever $i \ne i'$ and $\mathbb{E}[\varepsilon_i \varepsilon_{i'} | X_1, \dots, X_n] \le \sigma^2$ whenever $i = i'$. So, we obtain

$$\begin{aligned} \mathbb{E}[\mathcal{E}(\widehat{\theta})] &\le \frac{\sigma^2}{n^2} \sum_{i=1}^n \mathbb{E} \| \widehat{\boldsymbol{\Sigma}}^{-1} X_i \|_{\boldsymbol{\Sigma}}^2 = \frac{\sigma^2}{n^2} \sum_{i=1}^n \mathbb{E}\big[(\widehat{\boldsymbol{\Sigma}}^{-1} X_i)^\top \boldsymbol{\Sigma}\, \widehat{\boldsymbol{\Sigma}}^{-1} X_i \big] \\ &= \frac{\sigma^2}{n^2} \sum_{i=1}^n \mathbb{E}\big[\operatorname{tr}(X_i^\top \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}\, \widehat{\boldsymbol{\Sigma}}^{-1} X_i) \big] \end{aligned}$$

since $\operatorname{tr}(x) = x$ for $x \in \mathbb{R}$, so that finally, using the cyclic invariance of the trace and linearity, we obtain

$$\begin{aligned} \mathbb{E}[\mathcal{E}(\widehat{\theta})] &\le \frac{\sigma^2}{n^2} \sum_{i=1}^n \mathbb{E}\big[\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}\, \widehat{\boldsymbol{\Sigma}}^{-1} X_i X_i^\top) \big] \\ &= \frac{\sigma^2}{n} \mathbb{E}\big[\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}\, \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\Sigma}}) \big] \\ &= \frac{\sigma^2}{n} \mathbb{E}\big[\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}) \big] \\ &= \frac{\sigma^2}{n} \mathbb{E}\big[\operatorname{tr}((\boldsymbol{\Sigma}^{-1/2}\, \widehat{\boldsymbol{\Sigma}}\, \boldsymbol{\Sigma}^{-1/2})^{-1}) \big] = \frac{\sigma^2}{n} \mathbb{E}[\operatorname{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1})] \end{aligned}$$

which proves the upper bound. □

3.6.5 Proof of Theorem 3.6

Let $X_{n+1} \sim \mathbb{P}_X$ be independent of $X_1, \dots, X_n$ and write

$$\frac{1}{n}\mathbb{E}\operatorname{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1}) = \frac{1}{n}\mathbb{E}\operatorname{tr}((\widehat{\boldsymbol{\Sigma}})^{-1}\boldsymbol{\Sigma}) = \mathbb{E}\operatorname{tr}((n\widehat{\boldsymbol{\Sigma}})^{-1}X_{n+1}X_{n+1}^\top) \\ = \mathbb{E}\langle (n\widehat{\boldsymbol{\Sigma}})^{-1}X_{n+1}, X_{n+1}\rangle.$$

The proof uses the following cute trick based on the Sherman-Morrison lemma.

Lemma 3.9 (Sherman-Morrison) Let $\boldsymbol{A}$ be a $d \times d$ invertible real matrix and $u, v \in \mathbb{R}^d$. Then the following formula holds

$$(\boldsymbol{A} + uv^\top)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}uv^\top\boldsymbol{A}^{-1}}{1 + v^\top\boldsymbol{A}^{-1}u}.$$

This classical lemma allows to inverse the rank-one perturbation of a matrix as a function of its inverse.

Proof. The proof follows from a straightforward computation. Put $q = v^\top\boldsymbol{A}^{-1}u$ and write

$$(\boldsymbol{A} + uv^\top)\Big(\boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}uv^\top\boldsymbol{A}^{-1}}{1 + v^\top\boldsymbol{A}^{-1}u}\Big) \\ = (\boldsymbol{A} + uv^\top)\frac{\boldsymbol{A}^{-1} + q\,\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}uv^\top\boldsymbol{A}^{-1}}{1+q} \\ = \frac{\boldsymbol{I} + q\,\boldsymbol{I}}{1+q} = \boldsymbol{I}$$

We omit obvious computations here, just develop and cancel the terms...

A similar computation shows that

$$\Big(\boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}uv^\top\boldsymbol{A}^{-1}}{1 + v^\top\boldsymbol{A}^{-1}u}\Big)(\boldsymbol{A} + uv^\top) = \boldsymbol{I},$$

which proves the claim. □

Lemma 3.10 For any $\boldsymbol{S} \succ 0$ and any $v \in \mathbb{R}^d$ we have

$$\langle \boldsymbol{S}^{-1}v, v\rangle = \frac{\langle (\boldsymbol{S} + vv^\top)^{-1}v, v\rangle}{1 - \langle (\boldsymbol{S} + vv^\top)^{-1}v, v\rangle}.$$

This lemma gives a nice formula that allows to express a quadratic form as a function of its rank-1 perturbation.

Proof. We have $\boldsymbol{S} + vv^\top \succcurlyeq \boldsymbol{S} \succ 0$ so that $\boldsymbol{S} + vv^\top$ is invertible and using Lemma 3.9 gives

$$(\boldsymbol{S} + vv^\top)^{-1} = \boldsymbol{S}^{-1} - \frac{\boldsymbol{S}^{-1}vv^\top\boldsymbol{S}^{-1}}{1 + v^\top\boldsymbol{S}^{-1}v},$$

so that

$$\langle (\boldsymbol{S} + vv^\top)^{-1} v, v\rangle = v^\top \boldsymbol{S}^{-1} v - \frac{v^\top \boldsymbol{S}^{-1} vv^\top \boldsymbol{S}^{-1} v}{1 + v^\top \boldsymbol{S}^{-1} v}$$
$$= \langle \boldsymbol{S}^{-1} v, v\rangle - \frac{\langle \boldsymbol{S}^{-1} v, v\rangle^2}{1 + \langle \boldsymbol{S}^{-1} v, v\rangle} = \frac{\langle \boldsymbol{S}^{-1} v, v\rangle}{1 + \langle \boldsymbol{S}^{-1} v, v\rangle}$$

which concludes the proof. □

This proves in particular that $\widehat{\ell}_{n+1} \in [0, 1)$ a.s. since $\widehat{\boldsymbol{\Sigma}} \succ 0$ a.s. Now, using Lemma 3.10, we obtain

$$\frac{1}{n}\mathbb{E}[\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1})] = \mathbb{E}\big[\langle (n\widehat{\boldsymbol{\Sigma}})^{-1} X_{n+1}, X_{n+1}\rangle\big]$$
$$= \mathbb{E}\left[\frac{\langle (n\widehat{\boldsymbol{\Sigma}} + X_{n+1}X_{n+1}^\top)^{-1} X_{n+1}, X_{n+1}\rangle}{1 - \langle (n\widehat{\boldsymbol{\Sigma}} + X_{n+1}X_{n+1}^\top)^{-1} X_{n+1}, X_{n+1}\rangle}\right]$$
$$= \mathbb{E}\Big[\frac{\widehat{\ell}_{n+1}}{1 - \widehat{\ell}_{n+1}}\Big]$$

which concludes the proof of Theorem 3.6. □

3.6.6 Proof of Corollary 3.7

Theorem 3.6 and Jensen's inequality applied with the convex function $x \mapsto x/(1-x)$ on $[0, 1)$ gives

$$\mathbb{E}\Big[\frac{\widehat{\ell}_{n+1}}{1 - \widehat{\ell}_{n+1}}\Big] \geq \frac{\mathbb{E}[\widehat{\ell}_{n+1}]}{1 - \mathbb{E}[\widehat{\ell}_{n+1}]}.$$

Now, an exchangeability[26] argument gives

$$\mathbb{E}[\widehat{\ell}_{n+1}] = \mathbb{E}\left[\Big\langle \Big(\sum_{i'=1}^{n+1} X_{i'} X_{i'}^\top\Big)^{-1} X_{n+1}, X_{n+1}\Big\rangle\right]$$
$$= \mathbb{E}\left[\Big\langle \Big(\sum_{i'=1}^{n+1} X_{i'} X_{i'}^\top\Big)^{-1} X_i, X_i\Big\rangle\right]$$

for any $i = 1, \dots, n+1$, so that

$$\mathbb{E}[\widehat{\ell}_{n+1}] = \frac{1}{n+1}\sum_{i=1}^{n+1} \mathbb{E}\left[\Big\langle \Big(\sum_{i'=1}^{n+1} X_{i'} X_{i'}^\top\Big)^{-1} X_i, X_i\Big\rangle\right]$$
$$= \frac{1}{n+1}\mathbb{E}\left[\mathrm{tr}\left(\Big(\sum_{i=1}^{n+1} X_i X_i^\top\Big)^{-1} \sum_{i=1}^{n+1} X_i X_i^\top\right)\right] \quad = \frac{d}{n+1},$$

so that

$$\mathbb{E}\Big[\frac{\widehat{\ell}_{n+1}}{1 - \widehat{\ell}_{n+1}}\Big] \geq \frac{d}{n - d + 1},$$

26: By "exchangeability" we mean that this expectation is unchanged when applying any permutation of $X_1, \dots, X_{n+1}$, since these are iid.

which proves the Corollary □

Bayesian statistics 4

Let us go back to the problems of statistical inference that we considered in Chapter 2. We have data X valued on a measurable space $(E, \mathcal{E})$ and a model $\{P_\theta : \theta \in \Theta\}$ for its distribution, see Definition 1.1 from Chapter 1. For the problems of estimation and testing, we can define a set A of *decisions*: for scalar estimation, it is $A = \Theta \subset \mathbb{R}$, while for testing, we have binary decisions, so that $A = \{0, 1\}$.

4.1 Elements of decision theory

Given a (measurable) statistical procedure $\delta : E \to A$, we *decide* $\delta(X) \in A$. In order to assess a decision, we use a *loss function* $\ell : A \times \Theta \to \mathbb{R}$. This means that if the true parameter is $\theta \in \Theta$ and if we decide $a \in A$, we incur a loss $\ell(a, \theta) \in \mathbb{R}$.

Definition 4.1 Consider a statistical experiment with data $X \in E$ and a set of parameters Θ, a set A of decisions and a loss function $\ell : A \times \Theta \to \mathbb{R}$. The *risk* of a statistical procedure $\delta : E \to A$ is given by

$$R(\delta, \theta) = \mathbb{E}_\theta[\ell(\delta(X), \theta)]$$

for any $\theta \in \Theta$.

For the problem of estimation of a scalar parameter, we have $\Theta = \mathbb{R} = A$ and $\ell(\theta', \theta) = (\theta' - \theta)^2$, so that the risk is, in this case, the quadratic risk introduced in Definition 2.1 from Chapter 2. Note that we could consider other losses, such as $\ell(\theta', \theta) = |\theta' - \theta|^p$ for some $p \geq 1$.

Consider now statistical testing with hypotheses $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ where $\{\Theta_0, \Theta_1\}$ is a partition of Θ. Introduce the loss given by

$$\ell(i, \theta) = 0 \quad \text{if} \quad \theta \in \Theta_i \quad \text{and} \quad \ell(i, \theta) = c_i \quad \text{if} \quad \theta \in \Theta_{1-i} \tag{4.1}$$

for $i \in \{0, 1\}$ and constants $c_0, c_1 > 0$. The risk writes in this case

$$R(\delta, \theta) = c_i \mathbb{P}_\theta[\delta(X) = i] \quad \text{when} \quad \theta \in \Theta_{1-i}$$

for $i \in \{0, 1\}$. The constants $c_0, c_1 > 0$ allow to tune the importance given to the Type I and Type II errors: the approach described here leads to an approach of statistical testing different from the one described in Section 2.3.

4.2 Bayesian risk

Let us go back to the coin flip problem considered in Chapter 2. We observe $X_1, \dots, X_n$ iid distributed as $\mathrm{Bernoulli}(\theta)$ for $\theta \in (0,1)$. The estimator we introduced back then was the empirical mean $\widehat{\theta}_n = \bar{X}_n = n^{-1}\sum_{i=1}^n X_i$, and we know from (2.3) that the quadratic risk is given by $R(\widehat{\theta}_n, \theta) = \theta(1-\theta)/n$ for any $\theta \in (0,1)$. But what about the estimator $\widetilde{\theta}_n = 0$? It is a rather stupid estimator, but if θ is close to 0, it turns out to be a good estimator, and actually, it is easy to see that $R(\widetilde{\theta}_n, \theta) < R(\widehat{\theta}_n, \theta)$ whenever $\theta < 1/(n+1)$. This proves that $\widetilde{\theta}_n = 0$ is better than $\widehat{\theta}_n$, when assessed by the quadratic risk, for θ small enough.[1] This very simple example illustrates the fact that it is not possible to find an estimator with an optimal risk for all $\theta \in \Theta$.

1: A longer story hides beneath this simple example: the Stein effect and the Stein estimator, which provably improves the sample average estimator using thresholding, see [14, 15] for more details on this.

As illustrated above, given two statistical procedures δ, δ' for the same problem of statistical inference, we do not have in general that $R(\delta, \theta) < R(\delta', \theta)$ uniformly for $\theta \in \Theta$. What we can do instead is to consider an *averaged risk*: choose a distribution μ on Θ and use it to integrate the risk over Θ. This distribution is called the *prior distribution* or simply the *prior*.

Definition 4.2 (Bayesian risk) The Bayesian risk of a procedure δ associated to the *prior* μ is given by

$$R_B(\delta, \mu) = \int_\Theta R(\delta, \theta)\mu(d\theta) = \int_\Theta \mu(d\theta) \int_E \ell(\delta(x), \theta) P_\theta(dx).$$

Note that $R_B(\delta, \mu) \leq \sup_{\theta \in \Theta} R(\delta, \theta)$ which means that the Bayes risk is always smaller than the worst-case risk over Θ. We understand this risk as an average of the risk over Θ "weighted" by the prior distribution μ. Given a prior μ, the Bayesian risk is a scalar value: we can compare procedures using it and even try to find a procedure that minimizes it.[2]

2: We will do it in Section 4.4, such an estimator is called a Bayesian estimator.

Example 4.1 For statistical testing with the loss given by (4.1), the Bayesian risk associated to a prior μ writes

$$R_B(\delta, \mu) = \sum_{i \in \{0,1\}} c_i \int_{\Theta_{1-i}} \mathbb{P}_\theta[\delta(X) = i]\mu(d\theta),$$

which is a weighted combination of the Type I and Type II errors averaged by the prior μ.

Another interpretation of the Bayesian risk is of utmost importance in Bayesian statistics. Indeed, we could say that the parameter θ is *itself a random variable* distributed as μ, that we denote T instead of θ, and that P_θ is actually the distribution of X "conditionally" on $T = \theta$.[3]

3: Of course the event $T = \theta$ has zero probability if T is continuous with respect to the Lebesgue measure. We will explain clearly what such a *conditional density* is in Section 4.3 below.

Assuming that the joint distribution $P_{T,X}$ of (T, X) is given by

$$P_{T,X}[B \times C] = \mathbb{P}[T \in B, X \in C] = \int_B \mu(d\theta) \int_C P_\theta(dx),$$

we could write the Bayesian risk as an expectation with respect to $P_{T,X}$, since

$$R_B(\mu, \delta) = \int_\Theta \mu(d\theta) \int_E \ell(\delta(x), \theta) P_\theta(dx) = \mathbb{E}[\ell(\delta(X), T)].$$

What we need to do now, in order to formalize this, is to explain what a conditional density is, and to explain some useful formulas, such as the Bayes formula for conditional densities.

4.3 Conditional densities and the Bayes formula

Let X and Y be random variables on the same probability space and valued in measurable sets $\mathcal{X}$ and $\mathcal{Y}$. Let ϕ be a measurable function such that $\phi(X)$ is integrable. Let us recall that we can define the conditional expectation $\mathbb{E}[\phi(X)|Y]$ as the random variable $r(Y)$ (for some measurable function r, almost surely unique) such that

We assume here that the reader is familiar with the definition of the conditional expectation.

$$\mathbb{E}[\phi(X)\varphi(Y)] = \mathbb{E}[r(Y)\varphi(Y)] \tag{4.2}$$

for any measurable and bounded function φ. The particular value $r(y)$ for some $y \in \mathcal{Y}$ is denoted $\mathbb{E}[\phi(X)|Y = y]$. We know that

$$\mathbb{E}[\phi(X)h(Y)|Y] = h(Y)\mathbb{E}[\phi(X)|Y]$$

almost surely, for any measurable function h such that $h(Y)$ and $\phi(X)h(Y)$ are integrable, and we have also

$$\mathbb{E}[\mathbb{E}[\phi(X)|Y]] = \mathbb{E}[\phi(X)]. \tag{4.3}$$

Finally, we have $\mathbb{E}[\phi(X)|Y] = \mathbb{E}[\phi(X)]$ whenever X and Y are independent. Let us suppose now that the joint distribution $P_{X,Y}$ of (X, Y) has a density $p(x, y)$ with respect to a product of dominating measures $\nu_X \otimes \nu_Y$ on $\mathcal{X} \times \mathcal{Y}$. We can define the marginal densities of X and Y as

$$p_X(x) = \int_{\mathcal{Y}} p(x, y)\nu_Y(dy) \quad \text{and} \quad p_Y(y) = \int_{\mathcal{X}} p(x, y)\nu_X(dx), \tag{4.4}$$

so that we have

$$\mathbb{E}[\phi(X)] = \int_{\mathcal{X}} \phi(x)P_X(dx) = \int_{\mathcal{X}} \phi(x)p_X(x)\nu_X(dx)$$
$$\mathbb{E}[\varphi(Y)] = \int_{\mathcal{Y}} \varphi(y)P_Y(dy) = \int_{\mathcal{Y}} \varphi(y)p_Y(x)\nu_Y(dy)$$

for any ϕ and φ such that $\phi(X)$ and $\varphi(Y)$ are integrable. Let us introduce

$$\mathcal{Y}_0 = \{y \in \mathcal{Y} : p_Y(y) = 0\}$$

and remark that

$$\begin{aligned} P_{X,Y}[\mathcal{X} \times \mathcal{Y}_0] &= \int_{\mathcal{X} \times \mathcal{Y}_0} p(x,y)\nu_X(dx)\nu_Y(dy) \\ &= \int_{\mathcal{Y}_0} \nu_Y(dy) \int_{\mathcal{X}} p(x,y)\nu_X(dx) \\ &= \int_{\mathcal{Y}_0} p_Y(y)\nu_Y(dy) = 0. \end{aligned}$$

Using (4.4)

Given any probability density q on $\mathcal{X}$ with respect to ν_X, we can define

$$p_{X|Y}(x|y) := \frac{p(x,y)}{p_Y(y)}\mathbf{1}_{\mathcal{Y}_0^\complement}(y) + q(x)\mathbf{1}_{\mathcal{Y}_0}(y), \tag{4.5}$$

so that all the versions of $p_{X|Y}$ associated to the choice of q are equal $P_{X,Y}$-almost surely. Moreover, we can check immediately that $\int_{\mathcal{X}} p_{X|Y}(x|y)\nu_X(dx) = 1$, so that it is a probability density with respect to ν_X on $\mathcal{X}$. Now, if we define

$$r'(y) = \int_{\mathcal{X}} \phi(x) p_{X|Y}(x|y)\nu_X(dx),$$

we can write, for any measurable and bounded φ, that

$$\begin{aligned} &\mathbb{E}[r'(Y)\varphi(Y)] \\ &\quad = \int_{\mathcal{Y}} r'(y)\varphi(y)p_Y(y)\nu_Y(dy) \\ &\quad = \int_{\mathcal{Y}_0^\complement} \varphi(y)p_Y(y)\nu_Y(dy) \int_{\mathcal{X}} \phi(x)\frac{p(x,y)}{p_Y(y)}\nu_X(dx) && \text{Using the definition of } r', \text{ Fubini and (4.5)} \\ &\quad = \int_{\mathcal{X} \times \mathcal{Y}} \phi(x)\varphi(y)p(x,y)\nu_X(dx)\nu_Y(dy) && \text{Using the fact that } P_{X,Y}[\mathcal{X} \times \mathcal{Y}_0] = 0 \\ &\quad = \mathbb{E}[\phi(X)\varphi(Y)]. \end{aligned}$$

This corresponds to the definition of the conditional expectation, which is almost surely unique, so that we proved that $r = r'$ almost surely. Now, we know that we can compute a conditional expectation using the formula

$$\mathbb{E}[\phi(X)|Y = y] = r(y) = \int_{\mathcal{X}} \phi(x)p_{X|Y}(x|y)\nu_X(dx).$$

The density $p_{X|Y}$ is called the *conditional density* of X *knowing* Y.

We can define in the exact same way $p_{Y|X}$, the conditional density of Y knowing X, and by construction of $p_{X|Y}$ and $p_{Y|X}$, we have that the following equalities

$$p(x,y) = p_{X|\,Y}(x|\,y)p_Y(y) = p_{Y|X}(y|x)p_X(x) \tag{4.6}$$

hold $P_{X,Y}$-almost surely. From these equalities we can deduce that

$$p_{X|Y}(x|y) = \frac{p(x,y)}{p_Y(y)} = \frac{p_{Y|X}(y|x)p_X(x)}{\int_{\mathcal{X}} p(x',y)\nu_X(dx')}$$

recalling that $P_{X,Y}[\mathcal{X} \times \mathcal{Y}_0] = 0$

holds $P_{X,Y}$-almost surely, which leads, using again (4.6), to the Bayes formula

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{\int_{\mathcal{X}} p_{Y|X}(y|x')p_X(x')\nu_X(dx')}, \tag{4.7}$$

that holds, once again, $P_{X,Y}$-almost surely. This is a remarkable formula, since it allows to *reverse the conditioning*: we can write the conditional density of X knowing Y as a function of the conditional density of Y knowing X. This formula is at the core of Bayesian statistics, as explained in the next Section.

4.4 Posterior distribution and Bayes estimator

Let us go back to the setting introduced in Section 4.2. We have data X and a statistical model $\{P_\theta : \theta \in \Theta\}$. We consider a prior distribution μ on Θ. We assume that μ has a density $p(\cdot)$ with respect to a measure λ on Θ, namely $\mu(d\theta) = p(\theta)\lambda(d\theta)$ and that P_θ has a density that we will denote as $p(\cdot|\theta)$ with respect to a measure ν on E. We want to apply Bayesian reasoning: the density $p(\cdot|\theta)$ of the data is understood as a conditional density of X "knowing the parameter θ".

Using the same letter for both the density of μ (namely $\theta \mapsto p(\theta)$) and the density of P_θ (namely $x \mapsto p(x\,|\theta)$) might look like a bad idea, but it will lead to very nice notations in what follows, and it won't lead to any ambiguity.

The posterior distribution. In order to formalize this, we introduce a random variable T distributed as μ, and we apply (4.6) in order to express the joint density of (X,T) through the product of the conditional density of $X|T$ and the density of T:

$$p_{X,T}(x,\theta) = p_{X|T}(x|\theta)p_T(\theta) = p(x|\theta)p(\theta). \tag{4.8}$$

Which holds $P_{X,T}$-almost surely.

We can only proceed like this to express $p_{X,T}$, since what we are given is the prior density $p(\cdot)$ and the model, namely the density $p(\cdot|\theta)$. We know that the marginal density of X can be computed as

$$p_X(x) = \int_\Theta p_{X,T}(x,\theta)\lambda(d\theta) = \int_\Theta p(x|\theta)p(\theta)\lambda(d\theta).$$

Now, using the Bayes formula (4.7), we can *reverse the conditioning*, and define what we call the *posterior density*

$$p(\theta|x) := p_{T|X}(\theta|x) = \frac{p_{X,T}(x,\theta)}{p_X(x)} = \frac{p(x|\theta)p(\theta)}{\int_\Theta p(x|\theta')p(\theta')\lambda(d\theta')}.$$

This formula expresses the conditional density of the parameter θ knowing the data x (more formally the conditional density of T knowing X)

through the model (the conditional density of X knowing T) and the prior (the density of T) that are both known and chosen beforehand. Let us wrap-up what we constructed in the following definition.

Definition 4.3 Consider a *prior* $\mu(d\theta) = p(\theta)\lambda(d\theta)$ and a model $P_\theta(dx) = p(x|\theta)\nu(dx)$ for $\theta \in \Theta$, and the corresponding joint distribution $P(dx, d\theta) = p(x|\theta)p(\theta)\nu(dx)\lambda(d\theta)$. The *posterior distribution* is the distribution with density

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int_\Theta p(x|\theta')p(\theta')\lambda(d\theta')}$$

with respect to λ. It is well-defined and unique for P–almost all (x, θ).

The Bayesian reasoning is therefore as follows: choose a prior density $p(\theta)$ and a model $p(x|\theta)$ for the data x knowing the parameter θ. Then, compute[4] the posterior distribution $p(\theta|x)$ of θ knowing the data x. A nice aspect of this approach is that we can quantify uncertainty right out of the box, since instead of an estimator $\widehat{\theta}_n$ (which is, given data, a single value), we obtain a full posterior distribution $p(\theta|x)$, which takes into account the data x that we observed. However, such a reasoning is of course only possible when we know how to choose a prior, and when we are able to compute exactly or to approximate efficiently the posterior. [5]

4: or approximate it using numerical methods, whenever the posterior distribution cannot be computed explicitly

5: This is the main criticism with Bayesian methods: beyond simple models and conjugate distributions (more on this later), the computation of the posterior is not explicit and requires approximation algorithms that can be numerically expensive, or can depart significantly from the original model, see for instance Chapters 9 and 10 in [16].

The Bayes estimator. Let us consider the estimation problem where $A = \Theta \subset \mathbb{R}$ and use the Bayes risk to assess the error of an estimator $\delta : E \to \Theta$. Arguably, an optimal Bayesian estimator should minimize the Bayes risk, and a beautiful aspect of the Bayesian approach is that such a minimizer can be defined precisely. Indeed, we can rewrite the Bayes risk as follows:

$$\begin{aligned} R_B(\delta, \mu) &= \int_\Theta \int_E \ell(\delta(x), \theta)p(x|\theta)p(\theta)\nu(dx)\lambda(d\theta) \\ &= \int_E p_X(x)\nu(dx) \int_\Theta \ell(\delta(x), \theta)p(\theta|x)\lambda(dx). \end{aligned}$$

using (4.8)

Using Fubini and since we know that $p(x|\theta)p(\theta) = p(\theta|x)p_X(x)$ almost surely

What is remarkable with this rewriting is that in order to minimize $R_B(\delta, \mu)$, we need to minimize, for any fixed $x \in E$, the quantity

$$\int_\Theta \ell(\delta(x), \theta)p(\theta|x)\lambda(d\theta) = \mathbb{E}[\ell(\delta(X), T)|X = x],$$

Once again, T is a random variable with distribution $\mu(d\theta) = p(\theta)\lambda(d\theta)$

namely the expectation of the loss with respect to the posterior distribution given by Definition 4.3. This leads to the following definition of a Bayes estimator.

Definition 4.4 Given a prior $\mu(d\theta) = p(\theta)\lambda(d\theta)$, a model $P_\theta(dx) = p(x|\theta)\nu(dx)$ and a loss ℓ, any estimator $\widehat{\theta}(X)$ defined as

$$\widehat{\theta}(x) \in \operatorname*{argmin}_{t\in\Theta} \int_\Theta \ell(t,\theta)p(\theta|x)\lambda(d\theta) = \operatorname*{argmin}_{t\in\Theta} \mathbb{E}[\ell(t,T)|X=x],$$

namely a minimizer of the expectation of the loss with respect to the posterior distribution, is called a *Bayes* or a *Bayesian estimator*.

such an estimator is not necessarily unique

For the square loss $\ell(\theta',\theta) = (\theta'-\theta)^2$, the Bayes estimator is given by the expectation of the posterior distribution. Indeed, it is easy to see that

$$\widehat{\theta}(x) = \operatorname*{argmin}_{t\in\mathbb{R}} \int_\Theta (t-\theta)^2 p(\theta|x)\lambda(d\theta) = \int_\Theta \theta p(\theta|x)\lambda(d\theta). \quad (4.9)$$

If $\ell(\theta',\theta) = |\theta'-\theta|$, we can see that

$$\widehat{\theta}(x) = \operatorname*{argmin}_{t\in\mathbb{R}} \int_\Theta |t-\theta| p(\theta|x)\lambda(d\theta) = F_x^{-1}(1/2),$$

where $F_x^{-1}(1/2)$ is the *median* of the posterior distribution. Here, the notation F_x^{-1} stands for the generalized inverse of the distribution function $F_x(\theta) = \int_{-\infty}^{\theta} p(\theta'|x)\lambda(d\theta')$ of the posterior distribution.[6]

6: This comes from the fact that if X is an integrable real random variable with distribution function F, a minimizer of $t \mapsto \mathbb{E}|X-t|$ is given by the median $F^{-1}(1/2)$ of X.

Recipe

On some examples, we can compute explicitly the posterior distribution. Given the data density $p(x|\theta)$ and the prior density $p(\theta)$, the joint density of the data and the prior is $p(x|\theta)p(\theta)$ and we know from Definition 4.3 that the posterior density is proportional to the joint density, namely

$$p(\theta|x) = \text{constant}(x)p(x|\theta)p(\theta),$$

where $\text{constant}(x) = 1/\int_\Theta p(x|\theta)p(\theta)\lambda(d\theta)$. So, using the fact that the integral of the posterior density with respect to θ equals 1, we can try to identify directly the posterior distribution by having a careful look at $p(x|\theta)p(\theta)$, together with some coffee, and looking for a density with respect to θ.

4.5 Examples

Let us give some standard examples of priors and data distributions where can apply this recipe.

4.5.1 How to choose a restaurant ? (Bayesian coin flip)

The first example is, once again, a coin flip, but this time with a Bayesian flavor. Consider the data distribution $X \sim \text{Binomial}(n, \theta)$, namely the model with density

$$p(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \mathbf{1}_{\{0,\dots,n\}}(x)$$

with respect to the counting measure ν on $\mathbb{N}$ and the *flat prior* on θ with distribution $\text{Uniform}([0,1])$ on θ, namely a density

$$p(\theta) = \mathbf{1}_{[0,1]}(\theta)$$

with respect to the Lebesgue measure λ.

This model can be useful to help us choose a restaurant: given a restaurant, θ is the unknown probability that a customer is happy (1) or unhappy (0) with it and n is the number of customers who gave their (binary) opinion. We have no prior information on θ, so we consider the flat prior. For each restaurant, we observe the percentage of happy customers (rescaled to a 0 to 5 stars rating in Figure 4.1) and the number n of customers who rated it, so X stands here for the number of happy customers among the n who rated it. Assuming that the customers opinions are independent, we have $X \sim \text{Binomial}(n, \theta)$. The question is the following: how can we choose a restaurant? Is a restaurant with a good rating but few rates better than a restaurant with a slightly worse rating but more rates?

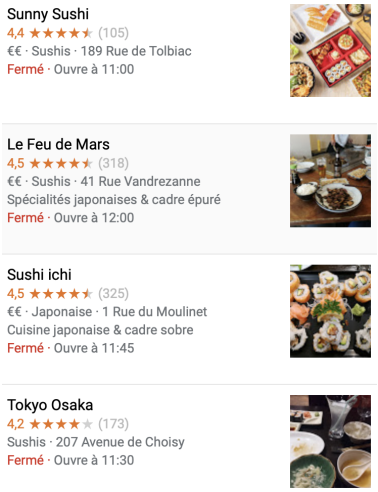


Figure 4.1: How to choose a restaurant? Is a restaurant with a good rating but few rates better than a restaurant with a slightly worse rating but more rates?

Using the previous recipe, we know that the density of the posterior distribution is proportional to the density of the joint distribution of the data and prior

$$p(x, \theta) = p(x|\theta)p(\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \mathbf{1}_{[0,1]}(\theta) \mathbf{1}_{\{0,\dots,n\}}(x)$$

with respect to the product measure $\nu \otimes \lambda$. This means that the posterior density is proportional to $\mapsto \theta^x (1-\theta)^{n-x} \mathbf{1}_{[0,1]}(\theta)$. We recognize the density

$$\theta \mapsto \frac{1}{\beta(a,b)} \theta^{a-1} (1-\theta)^{b-1} \mathbf{1}_{[0,1]}(\theta)$$

of the $\text{Beta}(a, b)$ distribution, that we introduced in Section 3.3.1 of Chapter 3, where we recall that $\beta(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1} dt = \Gamma(a)\Gamma(b)/\Gamma(a+b)$. Therefore, we know that the posterior distribution is $\text{Beta}(x+1, n-x+1)$, namely

$$p(\theta|x) = \frac{1}{\beta(x+1, n-x+1)} \theta^x (1-\theta)^{n-x} \mathbf{1}_{[0,1]}(\theta)$$

If $B \sim \mathrm{Beta}(a, b)$, we know that

$$\mathbb{E}[Z] = \frac{a}{a+b} \quad \text{and} \quad \mathbb{V}[Z] = \frac{ab}{(a+b)^2(a+b+1)}.$$

If we consider the square loss for the estimation of θ, we know from Equation (4.9) that the Bayesian estimator is given by the expectation of the posterior, namely

$$\widehat{\theta} = \frac{X+1}{X+1+n-X+1} = \frac{X+1}{n+2}.$$

Note the difference with the *frequentist* (non-Bayesian) estimator X/n that we considered all along Chapter 2 (it was denoted S_n/n back then). This estimator gives a cute Bayesian answer to the restaurant selection problem. This estimator is also known as the Laplace's *rule of succession* (see [17] and Figure 4.2).

[17]: Wikipedia (2020), *Rule of succession*

Using the bias-variance formula (2.2) from Chapter 2, we can compute the quadratic risk of $\widehat{\theta}$ as follows:

$$\begin{aligned}\mathbb{E}_\theta[(\widehat{\theta}-\theta)^2] &= \mathbb{V}_\theta[\widehat{\theta}] + (\mathbb{E}_\theta[\widehat{\theta}]-\theta)^2 \\ &= \frac{n\theta(1-\theta)}{(n+2)^2} + \Big(\frac{1-2\theta}{n+2}\Big)^2 \\ &= \frac{n\theta - n\theta^2 + 1 - 4\theta + 4\theta^2}{(n+2)^2},\end{aligned}$$

so that the Bayes risk with uniform prior $\mu(d\theta) = p(\theta)d\theta$ where $p(\theta) = \mathbf{1}_{[0,1]}(\theta)$ is given by

$$R_B(\widehat{\theta}, \mu) = \int_0^1 \mathbb{E}_\theta[(\widehat{\theta}-\theta)^2]d\theta = \frac{1}{6(n+2)}.$$

Using the exact same arguments, it is easy to see that if we use the prior distribution $\mathrm{Beta}(a, b)$ instead of $\mathrm{Uniform}([0, 1])$ (which is a particular case, since $\mathrm{Uniform}([0, 1]) = \mathrm{Beta}(1, 1)$) the posterior distribution is given by

$$p(\theta|x) = \mathrm{Beta}(a + x, b + n - x + b),$$

which leads to a Bayesian estimator (for the square loss) given by

$$\widehat{\theta} = \frac{X+a}{n+a+b}.$$

In this example, the prior and the posterior both belong to the family of Beta distributions. In such a case, we say that the $\mathrm{Binomial}$ and Beta distributions are *conjugated*, which corresponds to a situation where the posterior distribution can be *explicitly* computed.

Definition 4.5 (Conjugated distributions) Given a prior $\mu(d\theta) = p(\theta)\lambda(d\theta)$ and a model $P_\theta(dx) = p(x|\theta)\nu(dx)$, we say that the

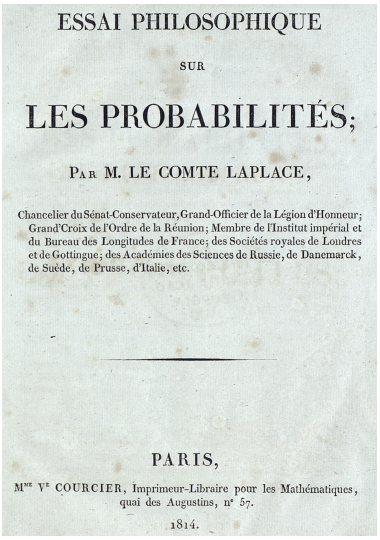

ESSAI PHILOSOPHIQUE
SUR
LES PROBABILITÉS;
PAR M. LE COMTE LAPLACE,
Chancelier du Sénat-Conservateur, Grand-Officier de la Légion d'Honneur; Grand'Croix de l'Ordre de la Réunion; Membre de l'Institut impérial et du Bureau des Longitudes de France; des Sociétés royales de Londres et de Gottingue; des Académies des Sciences de Russie, de Danemarck, de Suède, de Prusse, d'Italie, etc.
PARIS,
Mme Ve COURCIER, Imprimeur-Libraire pour les Mathématiques, quai des Augustins, n° 57.
1814.

Figure 4.2: "Essai philosophique sur les probabilités" by Pierre-Simon Laplace (1814) in which is introduced the *rule of succession* formula in order to "solve" the sunrise problem (What is the probability that the sun will rise tomorrow ?).

distributions of the prior and of the model are *conjugated* whenever the prior and the posterior distribution belong to the same family of distributions.

4.5.2 Gaussian sample with a Gaussian prior

Another classical example is with the Gaussian distribution. Consider data $X_1, \dots, X_n$ iid with $\mathrm{Normal}(\theta, \sigma^2)$ distribution, namely

$$p(x|\theta) = \mathrm{const}(\sigma) \exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2 \Big),$$

where $x = [x_1 \cdots x_n]^\top$ and $\mathrm{const}(\sigma)$ is a constant which depends only on σ and a prior $\mathrm{Normal}(0, \tau^2)$ distribution on θ, namely

$$p(\theta) = \mathrm{const}(\tau) \exp\Big(- \frac{\theta^2}{2\tau^2} \Big).$$

We proceed as previously and write the joint distribution

$$\begin{aligned} p(x|\theta)p(\theta) &= \mathrm{const}(\sigma, \tau) \exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2 - \frac{\theta^2}{2\tau^2} \Big) \\ &= \mathrm{const}(\sigma, \tau, x) \exp\left(- \frac{1}{2} \Big(\frac{n}{\sigma^2} + \frac{1}{\tau^2} \Big) \theta^2 + \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i \theta \right) \\ &= \mathrm{const}(\sigma, \tau, x) \exp\left(- \frac{1}{2\gamma} \Big(\theta - \frac{\gamma}{\sigma^2} \sum_{i=1}^{n} x_i \Big)^2 \right), \end{aligned}$$

where we put $\gamma = \sigma^2/(n + \sigma^2/\tau^2)$. This proves that

$$p(\theta|x) = \mathrm{Normal}\left(\frac{1}{n + \sigma^2/\tau^2} \sum_{i=1}^{n} x_i, \frac{\sigma^2}{n + \sigma^2/\tau^2} \right),$$

and that the Bayes estimator for the square loss is given by

$$\widehat{\theta} = \frac{1}{n + \sigma^2/\tau^2} \sum_{i=1}^{n} X_i.$$

This proves, in particular, that the Gaussian family is *conjugated* with itself.

4.5.3 Bayesian linear regression with a Gaussian prior

Another very interesting example is the Gaussian linear regression model that we considered in Section 3.3 of Chapter 3, where

$$Y_i = X_i^\top \theta + \varepsilon_i$$

with deterministic $X_i \in \mathbb{R}^d$ and iid $\varepsilon_i \sim \mathrm{Normal}(0, \sigma^2)$. This means that $\boldsymbol{y} \sim \mathrm{Normal}(\boldsymbol{X}\,\theta, \sigma^2\,\boldsymbol{I}_n)$, where we recall that $\boldsymbol{y} = [Y_1 \cdots Y_n]^\top$ and that $\boldsymbol{X}$ is the $n \times d$ matrix with rows given by $X_1, \ldots, X_n$. We consider this model in a Bayesian setting, by using a $\mathrm{Normal}(0, \lambda^{-1}\,\boldsymbol{I}_d)$ prior on θ, where $\lambda > 0$. The joint distribution of $(\theta, \boldsymbol{y})$ is given by

$$p(\theta, \boldsymbol{y}) = \mathrm{const}(\sigma, \lambda) \exp\Big(-\frac{1}{2\sigma^2} \| \boldsymbol{y} - \boldsymbol{X}\,\theta\|^2 - \frac{\lambda}{2}\|\theta\|^2 \Big).$$

What is, in this setting, the posterior distribution $p(\theta|\,\boldsymbol{y})$? This is slightly more complicated than what we did in both previous examples, and deserves the next theorem.

Theorem 4.1 Consider the Gaussian linear model, namely the data distribution

$$p(\boldsymbol{y}\,|\theta) = \mathrm{Normal}(\boldsymbol{X}\,\theta, \sigma^2\,\boldsymbol{I}_n)$$

with prior

$$p(\theta) = \mathrm{Normal}(0, \lambda^{-1}\,\boldsymbol{I}_d)$$

for some $\lambda > 0$. Then, we have

$$p(\theta|\,\boldsymbol{y}) \\ = \mathrm{Normal}\Big((\boldsymbol{X}^\top\,\boldsymbol{X} + \lambda\sigma^2\,\boldsymbol{I}_d)^{-1}\,\boldsymbol{X}^\top\,\boldsymbol{y},\ \sigma^2 (\boldsymbol{X}^\top\,\boldsymbol{X} + \lambda\sigma^2\,\boldsymbol{I}_d)^{-1} \Big).$$

The proof of Theorem 4.1 is given in Section 4.6 below. If we consider the square loss $\ell(\theta', \theta) = \|\theta' - \theta\|^2$ where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^d$, we have that the Bayesian estimator is the expectation[7] of the posterior distribution, namely

$$\widehat{\theta} = (\boldsymbol{X}^\top\,\boldsymbol{X} + \lambda\sigma^2\,\boldsymbol{I}_d)^{-1}\,\boldsymbol{X}^\top\,\boldsymbol{y}\,.$$

7: since $L(t) = \mathbb{E}[\|Z - t\|^2]$ for $t \in \mathbb{R}^d$ is minimized at $t^* = \mathbb{E}[Z]$ whenever $\mathbb{E}[\|Z\|^2] < +\infty$

In this example, the Bayes estimator coincides[8] with the so-called MAP estimator (Maximum A Posteriori), which is given, when it exists, by the *mode* of the posterior distribution.

8: since the mode and the expectation of a Gaussian distribution are the same

We could have computed the MAP estimator without computing the posterior distribution. Indeed, we know that the posterior distribution $p(\boldsymbol{y}\,|\theta)$ is proportional to the joint distribution $p(\theta, \boldsymbol{y})$, hence proportional to

$$\exp\ \Big(-\frac{1}{2\sigma^2} \| \boldsymbol{y} - \boldsymbol{X}\,\theta\|^2 - \frac{\lambda}{2}\|\theta\|^2 \Big),$$

so that maximizing this function with respect to θ corresponds to minimizing

$$F(\theta) = \| \boldsymbol{y} - \boldsymbol{X}\,\theta\|^2 + \sigma^2\lambda\|\theta\|^2.$$

The function F is strongly convex on $\mathbb{R}^d$, since its Hessian satisfies $\nabla^2 F(\theta) = 2\,\boldsymbol{X}^\top\,\boldsymbol{X} + 2\sigma^2\lambda\,\boldsymbol{I}_d \succcurlyeq \sigma^2\lambda\,\boldsymbol{I}_d \succ 0$ for any $\theta \in \mathbb{R}^d$. So, its unique global minimizer cancels out the gradient

$$\nabla F(\theta) = 2\,\boldsymbol{X}^\top(\boldsymbol{X}\,\theta - \boldsymbol{y}) + 2\sigma^2\lambda\theta,$$

and is therefore equal to

$$\widehat{\theta} = (\boldsymbol{X}^\top \boldsymbol{X} + 2\sigma^2 \lambda \boldsymbol{I}_d)^{-1} \boldsymbol{X}^\top \boldsymbol{y},$$

as announced above. This estimator corresponds to a *regularized* or *penalized* version of the least-squares estimator. Any minimizer of

$$\widehat{\theta}_{\mathrm{pen}} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \Big\{ \| \boldsymbol{y} - \boldsymbol{X}\, \theta \|^2 + \mathrm{pen}(\theta) \Big\},$$

where $\mathrm{pen} : \mathbb{R}^d \to \mathbb{R}^+$ is a so-called *penalization*, is called a *penalized* least-squares estimator.[9] A penalization function pen typically satisfies $\mathrm{pen}(0) = 0$ and that $\mathrm{pen}(\theta)$ is a non-increasing function with respect to the absolute value of each coordinate of θ, so that pen *penalizes* the fact that θ has large coordinates.

9: it might be unique or not, depending on pen and $\boldsymbol{X}$

Ridge penalization. Whenever $\mathrm{pen}(\theta) = \lambda \|\theta\|^2$, we call pen the *ridge* penalization and the problem is called *ridge regression*. This penalization "forces" the coordinates of $\theta \in \mathbb{R}^d$ to remain "small". It is the most widely used form of penalization in statistics and machine learning and it is used way beyond the setting of least-squares regression. For instance, this penalization is used in deep learning under the name *Weight decay*.

A prior is a form of regularization. Interestingly, we observe in this example that for the model of Gaussian linear regression, an isotropic Gaussian prior $p(\theta) = \mathrm{Normal}(0, \lambda^{-1} \boldsymbol{I}_d)$ acts exactly as a ridge penalization, which forbids the coordinates of θ to be *free*.[10] Given $\lambda > 0$, we define the minimizer of the ridge regression problem as

$$\widehat{\theta}_\lambda = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \Big\{ \| \boldsymbol{y} - \boldsymbol{X}\, \theta \|^2 + \lambda \|\theta\|^2 \Big\}. \tag{4.10}$$

10: and eventually take arbitrary large values, whenever the conditioning of $\boldsymbol{X}$ is bad for instance

Whenever λ is very small, then the prior is almost "flat" which is equivalent to the fact that the ridge penalization term in (4.10) is negligible. In this case, we expect $\widehat{\theta}_\lambda$ to be close to the least-squares estimator $\widehat{\theta}_0$.[11] On the other hand, if λ is large, the prior is highly concentrated around 0, which is equivalent to a very strong ridge penalization in the computation of $\widehat{\theta}_\lambda$.

11: which is $\widehat{\theta}_\lambda$ with $\lambda = 0$

The parameter $\lambda > 0$ used in the ridge penalization correspond to a regularization *strength*. It is also called in machine learning a *hyper-parameter*[12] , which is tuned in practice using cross-validation.

12: since it "parametrizes" the parameters...

Show the regularization path of the ridge estimator on a dataset

4.6 Proofs

4.6.1 Proof of Theorem 4.1

Let us first recall that the prior is given by

$$p(\theta) = \text{const}(\lambda) \exp\Big(-\frac{\lambda}{2}\|\theta\|^2 \Big)$$

and that the model is given by

$$p(\boldsymbol{y}\,|\theta) = \text{const}(\sigma) \exp\Big(-\frac{1}{2\sigma^2}\|\,\boldsymbol{y} - \boldsymbol{X}\,\theta\|^2 \Big),$$

so that the logarithm of the joint density of $(\theta, \boldsymbol{y})$ writes

$$\begin{aligned}\log p(\theta, \boldsymbol{y}) &= \log p(\theta) + \log p(\boldsymbol{y}\,|\theta) \\ &= \text{const}(\sigma^2, \lambda) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\,\theta)^\top(\boldsymbol{y} - \boldsymbol{X}\,\theta) - \frac{\lambda}{2}\theta^\top\theta.\end{aligned}$$

Let us develop and rewrite this expression as a quadratic form with respect to $(\theta, \boldsymbol{y})$ (forgetting about the constant terms):

$$\begin{aligned}&\frac{1}{\sigma^2}y^\top y - \frac{2}{\sigma^2}\,\boldsymbol{y}^\top\,\boldsymbol{X}\,\theta + \frac{1}{\sigma^2}\theta^\top\,\boldsymbol{X}^\top\,\boldsymbol{X}\,\theta + \lambda\theta^\top\theta \\ &\quad= \begin{bmatrix}\theta \\ \boldsymbol{y}\end{bmatrix}^\top \begin{bmatrix}\frac{1}{\sigma^2}\,\boldsymbol{X}^\top\,\boldsymbol{X} + \lambda\,\boldsymbol{I}_d & -\frac{1}{\sigma^2}\,\boldsymbol{X}^\top \\ -\frac{1}{\sigma^2}\,\boldsymbol{X} & \frac{1}{\sigma^2}\,\boldsymbol{I}_n\end{bmatrix}\begin{bmatrix}\theta \\ \boldsymbol{y}\end{bmatrix} \\ &\quad= \begin{bmatrix}\theta \\ \boldsymbol{y}\end{bmatrix}^\top \boldsymbol{K} \begin{bmatrix}\theta \\ \boldsymbol{y}\end{bmatrix}.\end{aligned}$$

This computation proves that the joint distribution of $(\theta, \boldsymbol{y})$ is Gaussian with *precision* matrix $\boldsymbol{K}$,[13] namely

$$p(\theta, \boldsymbol{y}) = \text{Normal}(0, \boldsymbol{K}^{-1}).$$

13: The *precision* matrix is the *inverse* of the covariance matrix

Now, in order to obtain the posterior distribution $p(\theta|\,\boldsymbol{y})$, we need to compute the conditional density of θ knowing $\boldsymbol{y}$ from the joint distribution of $(\theta, \boldsymbol{y})$. Since the joint distribution is Gaussian, it turns out to be particularly easy, as explained in the following proposition.

Proposition 4.2 Let Z be a Gaussian vector $Z \sim \text{Normal}(\mu, \boldsymbol{\Sigma})$ on $\mathbb{R}^m$ with $\boldsymbol{\Sigma} \succ 0$. We consider the decomposition of Z, and of its expectation and covariance matrix, into two blocks X_a and X_b as follows

$$X = \begin{bmatrix}X_a \\ X_b\end{bmatrix}, \quad \mu = \begin{bmatrix}\mu_a \\ \mu_b\end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_{a,a} & \boldsymbol{\Sigma}_{a,b} \\ \boldsymbol{\Sigma}_{a,b}^\top & \boldsymbol{\Sigma}_{b,b}\end{bmatrix}.$$

We decompose in the same way the precision matrix, namely

$$\boldsymbol{K} = \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{K}_{a,a} & \boldsymbol{K}_{a,b} \\ \boldsymbol{K}_{a,b}^\top & \boldsymbol{K}_{b,b} \end{bmatrix}. \tag{4.11}$$

Then, the conditional density of X_a knowing X_b is given by

$$p_{X_a|X_b}(x_a|x_b) = \text{Normal}\Big(\mu_a - \boldsymbol{K}_{a,a}^{-1}\,\boldsymbol{K}_{a,b}(x_b - \mu_b),\ \boldsymbol{K}_{a,a}^{-1}\Big),$$

where we used the precision matrix $\boldsymbol{K}$. We can also compute the conditional density as

$$\begin{aligned} &p_{X_a|X_b}(x_a|x_b) \\ &= \text{Normal}\Big(\mu_a + \boldsymbol{\Sigma}_{a,b}\,\boldsymbol{\Sigma}_{b,b}^{-1}(x_b - \mu_b),\ \boldsymbol{\Sigma}_{a,a} - \boldsymbol{\Sigma}_{a,b}\,\boldsymbol{\Sigma}_{b,b}^{-1}\,\boldsymbol{\Sigma}_{a,b}^\top\Big), \end{aligned}$$

where we used this time the covariance $\boldsymbol{\Sigma}$.

The proof of Proposition 4.2 is given below. In order to compute the posterior $p(\theta|\,\boldsymbol{y})$, we use Proposition 4.2 with $X_a = \theta$, $X_b = \boldsymbol{y}$, $\mu_a = 0$, $\mu_b = 0$ and the precision matrix

$$\boldsymbol{K} = \begin{bmatrix} \boldsymbol{K}_{a,a} & \boldsymbol{K}_{a,b} \\ \boldsymbol{K}_{a,b}^\top & \boldsymbol{K}_{b,b} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2}\,\boldsymbol{X}^\top\boldsymbol{X} + \lambda\,\boldsymbol{I}_d & -\frac{1}{\sigma^2}\,\boldsymbol{X}^\top \\ -\frac{1}{\sigma^2}\,\boldsymbol{X} & \frac{1}{\sigma^2}\,\boldsymbol{I}_n \end{bmatrix}.$$

Namely, we obtain that $p(\theta|\,\boldsymbol{y}) = \text{Normal}(\mu_{\theta|\,\boldsymbol{y}},\ \boldsymbol{\Sigma}_{\theta|\,\boldsymbol{y}})$ with

$$\mu_{\theta|\,\boldsymbol{y}} = \mu_a - \boldsymbol{K}_{a,a}^{-1}\,\boldsymbol{K}_{a,b}(x_b - \mu_b) = \big(\,\boldsymbol{X}^\top\boldsymbol{X} + \lambda\sigma^2\,\boldsymbol{I}_d\,\big)^{-1}\,\boldsymbol{X}^\top\,\boldsymbol{y},$$

and

$$\boldsymbol{\Sigma}_{\theta|\,\boldsymbol{y}} = \boldsymbol{K}_{a,a}^{-1} = \sigma^2\Big(\,\boldsymbol{X}^\top\boldsymbol{X} + \lambda\sigma^2\,\boldsymbol{I}_d\,\Big)^{-1},$$

which concludes the proof of Theorem 4.1.

Proof of Proposition 4.2. We have that

$$\log p_{(X_a,X_b)}(x_a, x_b) = \text{const}(\boldsymbol{K}) - \frac{1}{2}(x-\mu)^\top\,\boldsymbol{K}(x-\mu)$$

and that

$$\begin{aligned} &(x-\mu)^\top\,\boldsymbol{K}(x-\mu) \\ &= (x_a-\mu_a)^\top\,\boldsymbol{K}_{a,a}(x_a-\mu_a) + (x_a-\mu_a)^\top\,\boldsymbol{K}_{a,b}(x_b-\mu_b) \\ &\quad + (x_b-\mu_b)^\top\,\boldsymbol{K}_{a,b}^\top(x_a-\mu_a) + (x_b-\mu_b)^\top\,\boldsymbol{K}_{b,b}(x_b-\mu_b) \\ &= (x_a-\mu_a+\boldsymbol{K}_{a,a}^{-1}\,\boldsymbol{K}_{a,b}(x_b-\mu_b))^\top\,\boldsymbol{K}_{a,a} \\ &\quad \times (x_a-\mu_a+\boldsymbol{K}_{a,a}^{-1}\,\boldsymbol{K}_{a,b}(x_b-\mu_b)) + \text{const}(\mu, \boldsymbol{K}, x_b). \end{aligned}$$

We know that $p_{X_a|X_b}(x_a|x_b)$ is proportional to $p_{(X_a,X_b)}(x_a,x_b)$, so we already know from the previous computation that

$$p_{X_a|X_b}(x_a|x_b) = \mathrm{Normal}\left(\mu_a - \boldsymbol{K}_{a,a}^{-1}\,\boldsymbol{K}_{a,b}(x_b-\mu_b),\ \boldsymbol{K}_{a,a}^{-1}\right).$$

Now, in order to express $p_{X_a|X_b}$ through the covariance $\boldsymbol{\Sigma}$, we need to compute the inverse of $\boldsymbol{\Sigma}$. Let us recall the following classical block inversion formula

$$\begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{bmatrix}^{-1} = \begin{bmatrix}\boldsymbol{S} & -\boldsymbol{S}\,\boldsymbol{B}\,\boldsymbol{D}^{-1}\\ -\boldsymbol{D}^{-1}\,\boldsymbol{C}\,\boldsymbol{S} & \boldsymbol{D}^{-1}+\boldsymbol{D}^{-1}\,\boldsymbol{C}\,\boldsymbol{S}\,\boldsymbol{B}\,\boldsymbol{D}^{-1}\end{bmatrix},$$

where $\boldsymbol{S} = (\boldsymbol{A}-\boldsymbol{B}\,\boldsymbol{D}^{-1}\,\boldsymbol{C})^{-1}$ is called the *Schur complement* with respect to the block $\boldsymbol{D}$. We use this formula to compute

$$\boldsymbol{K} = \begin{bmatrix}\boldsymbol{K}_{a,a} & \boldsymbol{K}_{a,b}\\ \boldsymbol{K}_{a,b}^\top & \boldsymbol{K}_{b,b}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\Sigma}_{a,a} & \boldsymbol{\Sigma}_{a,b}\\ \boldsymbol{\Sigma}_{a,b}^\top & \boldsymbol{\Sigma}_{b,b}\end{bmatrix}^{-1}.$$

This gives us

$$\boldsymbol{K}_{a,a} = \boldsymbol{S} = (\boldsymbol{A}-\boldsymbol{B}\,\boldsymbol{D}^{-1}\,\boldsymbol{C})^{-1} = (\boldsymbol{\Sigma}_{a,a}-\boldsymbol{\Sigma}_{a,b}\,\boldsymbol{\Sigma}_{b,b}^{-1}\,\boldsymbol{\Sigma}_{a,b}^\top)^{-1}$$

and

$$\boldsymbol{K}_{a,a}^{-1}\,\boldsymbol{K}_{a,b} = -\,\boldsymbol{S}^{-1}\,\boldsymbol{S}\,\boldsymbol{B}\,\boldsymbol{D}^{-1} = -\,\boldsymbol{B}\,\boldsymbol{D}^{-1} = -\,\boldsymbol{\Sigma}_{a,b}\,\boldsymbol{\Sigma}_{b,b}^{-1},$$

so that

$$\mu_a - \boldsymbol{K}_{a,a}^{-1}\,\boldsymbol{K}_{a,b}(x_b-\mu_b) = \mu_a + \boldsymbol{\Sigma}_{a,b}\,\boldsymbol{\Sigma}_{b,b}^{-1}(x_b-\mu_b),$$

which concludes the proof of Proposition 4.2. □

4.6.2 Proof of the lower bound from Theorem 3.4

We have now all the tools required to prove the lower bound involved in Theorem 3.4 from Chapter 3, namely that

$$\inf_{\widehat{\theta}}\ \sup_{P\in\mathcal{G}(\mathbb{P}_X,\sigma^2)} \mathbb{E}\|\widehat{\theta}-\theta^*\|_{\boldsymbol{\Sigma}}^2 \geq \frac{\sigma^2}{n}\mathbb{E}[\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1})], \tag{4.12}$$

where we recall that $\boldsymbol{\Sigma} = \mathbb{E}[XX^\top] \succ 0$ and that $\mathcal{G}(\mathbb{P}_X,\sigma^2)$ is the set of joint distributions P for (X,Y) satisfying $X\sim\mathbb{P}_X$, $Y = X^\top\theta^* + \varepsilon$ almost surely and ε independent of X and such that $\varepsilon\sim\mathrm{Normal}(0,\sigma^2)$. Let us recall also that $\widehat{\theta}$ is any estimator, namely any measurable function of $(X_1,Y_1),\dots,(X_n,Y_n)$ iid with the same distribution $P\in\mathcal{G}(\mathbb{P}_X,\sigma^2)$.

First, let us remark that $\sup_{P\in\mathcal{G}(P_X,\sigma^2)}$ corresponds to $\sup_{\theta^*\in\mathbb{R}^d}$, so that denoting $\mathbb{P}_{\theta^*} = P_{X,Y}$ and the corresponding expectation $\mathbb{E}_{\theta^*}$, we

need to lower bound

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\widehat{\theta} - \theta\|_{\boldsymbol{\Sigma}}^2.$$

The first, and certainly most important trick, is to lower bound the minimax risk by the *Bayes* risk. Let us choose some prior distribution $\Pi(d\theta) = p(\theta)d\theta$ for θ and write

$$\begin{aligned} \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\widehat{\theta} - \theta\|_{\boldsymbol{\Sigma}}^2 &\geq \inf_{\widehat{\theta}} \int_{\mathbb{R}^d} \mathbb{E}_\theta \|\widehat{\theta} - \theta\|_{\boldsymbol{\Sigma}}^2 \, p(\theta) d\theta \\ &= \inf_{\widehat{\theta}} \mathbb{E}_{\theta \sim \Pi} \mathbb{E}_\theta \|\widehat{\theta} - \theta\|_{\boldsymbol{\Sigma}}^2. \end{aligned} \tag{4.13}$$

Let us reason conditionally on $X_1, \ldots, X_n$ in what follows to simplify notations and use Bayesian reasoning where the data has density

$$p(\boldsymbol{y} \,|\theta) = \mathrm{Normal}(\boldsymbol{X}\,\theta, \sigma^2\,\boldsymbol{I}_n)$$

and where the prior on θ is given by $\Pi_\lambda(d\theta) = p_\lambda(\theta)d\theta$ with

$$p_\lambda(\theta) = \mathrm{Normal}\left(0, \frac{\sigma^2}{\lambda n}\,\boldsymbol{I}_d\right)$$

for some $\lambda > 0$. Note that this example is exactly the one considered in Section 4.5.3 with $\lambda' = n\,\lambda/\sigma^2$ instead of λ. So, using Theorem 4.1, we have that

$$p(\theta|\,\boldsymbol{y}) = \mathrm{Normal}\left(\widehat{\theta}_\lambda, \frac{\sigma^2}{n}(\boldsymbol{X}^\top\,\boldsymbol{X} + \lambda\,\boldsymbol{I}_d)^{-1}\right)$$

where

$$\begin{aligned} \widehat{\theta}_\lambda &= (\boldsymbol{X}^\top\,\boldsymbol{X} + n\lambda\,\boldsymbol{I}_d)^{-1}\,\boldsymbol{X}^\top\,\boldsymbol{y} \\ &= \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left(\frac{1}{n}\|\,\boldsymbol{y} - \boldsymbol{X}\,\theta\|^2 + \lambda\|\theta\|^2\right) \end{aligned} \tag{4.14}$$

is the ridge-penalized least squares estimator. The second trick is that we *know how to minimize the Bayes risk* (4.13): it can be minimized by looking for

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta' \in \mathbb{R}^d} \int_{\mathbb{R}^d} \|\theta' - \theta\|_{\boldsymbol{\Sigma}}^2 \, p(\theta|\,\boldsymbol{y}) d\theta,$$

as explained in Section 4.4. But, let us remark that if Z is a random vector such that $\mathbb{E}\|Z\|^2 < \infty$, then the function $F : \mathbb{R}^d \to \mathbb{R}^+$ given by $F(t) = \mathbb{E}\|Z - t\|_{\boldsymbol{\Sigma}}^2$ is minimized at $t^* = \mathbb{E}[Z]$ whenever $\boldsymbol{\Sigma} \succ 0$. This entails that the Bayes estimator for the loss $\ell(\theta', \theta) = \|\theta' - \theta\|_{\boldsymbol{\Sigma}}^2$ is indeed $\widehat{\theta}_\lambda$. So, we end up with the lower bound

$$\begin{aligned} \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\widehat{\theta} - \theta\|_{\boldsymbol{\Sigma}}^2 &\geq \int_{\mathbb{R}^d} \mathbb{E}_\theta \|\widehat{\theta}_\lambda - \theta\|_{\boldsymbol{\Sigma}}^2 \Pi_\lambda(d\theta) \\ &= \mathbb{E}_{\theta \sim \Pi_\lambda} \mathbb{E}_\theta[\mathcal{E}(\widehat{\theta}_\lambda)] \end{aligned}$$

for any $\lambda > 0$, that we are able to compute exactly thanks to the next Lemma. Let us recall that $\widehat{\boldsymbol{\Sigma}} = n^{-1}\, \boldsymbol{X}^\top \boldsymbol{X} = n^{-1} \sum_{i=1}^n X_i X_i^\top$ and introduce $\widehat{\boldsymbol{\Sigma}}_\lambda = \widehat{\boldsymbol{\Sigma}} + \lambda\, \boldsymbol{I}_d$.

Lemma 4.3 The excess risk of the ridge estimator $\widehat{\theta}_\lambda$ given by (4.14) is given by

$$\mathbb{E}_\theta[\mathcal{E}(\widehat{\theta}_\lambda)] = \lambda^2 \mathbb{E}\|\theta\|^2_{(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\, \boldsymbol{\Sigma}\, (\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}} + \frac{\sigma^2}{n} \mathbb{E}\,\mathrm{tr}\left((\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\, \boldsymbol{\Sigma}\, (\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1} \widehat{\boldsymbol{\Sigma}} \right)$$

under the assumption that $Y_i = X_i^\top \theta + \varepsilon_i$ for $\varepsilon_i \sim \mathrm{Normal}(0, \sigma^2)$.

We inject the formula given by Lemma 4.3 to end up with the lower bound

$$\mathbb{E}_{\theta \sim \Pi_\lambda} \left[\lambda^2 \mathbb{E}\|\theta\|^2_{(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\, \boldsymbol{\Sigma}\, (\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}} + \frac{\sigma^2}{n} \mathbb{E}\,\mathrm{tr}\left((\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\, \boldsymbol{\Sigma}\, (\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1} \widehat{\boldsymbol{\Sigma}} \right) \right].$$

So, using Fubini, and since $\mathbb{E}_{\theta \sim \Pi_\lambda}[\theta\theta^\top] = \frac{\sigma^2}{\lambda n}\, \boldsymbol{I}_d$ by definition of Π_λ, we end up with

$$\begin{aligned}
&\mathbb{E}_{\theta \sim \Pi_\lambda} \left[\lambda^2 \mathbb{E}\|\theta\|^2_{(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\, \boldsymbol{\Sigma}\, (\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}} \right] \\
&\qquad = \lambda^2\, \mathbb{E}\, \mathbb{E}_{\theta \sim \Pi_\lambda}\, \mathrm{tr}\left[\theta^\top (\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\, \boldsymbol{\Sigma}\, (\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1} \theta \right] && \text{using } \mathrm{tr}[x] = x \text{ for } x \in \mathbb{R} \\
&\qquad = \lambda^2\, \mathbb{E}\, \mathbb{E}_{\theta \sim \Pi_\lambda}\, \mathrm{tr}\left[(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\, \boldsymbol{\Sigma}\, (\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1} \theta\theta^\top \right] \\
&\qquad = \frac{\sigma^2}{n}\, \mathbb{E}\,\mathrm{tr}\left[(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\, \boldsymbol{\Sigma}\, (\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1} \lambda\, \boldsymbol{I}_d \right],
\end{aligned}$$

so that the lower bound becomes now

$$\frac{\sigma^2}{n} \mathbb{E}\,\mathrm{tr}\left[(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\, \boldsymbol{\Sigma}\, (\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1} (\widehat{\boldsymbol{\Sigma}} + \lambda\, \boldsymbol{I}_d) \right] = \frac{\sigma^2}{n} \mathbb{E}\,\mathrm{tr}\left[(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\, \boldsymbol{\Sigma} \right].$$

We proved that the lower bound

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\widehat{\theta} - \theta\|^2_{\boldsymbol{\Sigma}} \;\geq\; \frac{\sigma^2}{n} \mathbb{E}\,\mathrm{tr}\left[(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\, \boldsymbol{\Sigma} \right]$$

holds for any $\lambda > 0$. Since $\mathbb{P}_X$ is non-degenerate, we know from Theorem 3.1 that $\widehat{\boldsymbol{\Sigma}} \succ 0$ almost surely and we have that the function

$$\lambda \mapsto \mathrm{tr}\left[(\widehat{\boldsymbol{\Sigma}} + \lambda\, \boldsymbol{I}_d)^{-1}\, \boldsymbol{\Sigma} \right] = \mathrm{tr}\left[(\boldsymbol{\Sigma}^{-1/2}\, \widehat{\boldsymbol{\Sigma}}\, \boldsymbol{\Sigma}^{-1/2} + \lambda\, \boldsymbol{\Sigma}^{-1})^{-1} \right]$$

is decreasing on $(0, +\infty)$ since $\lambda_2\, \boldsymbol{\Sigma}^{-1} \succ \lambda_1\, \boldsymbol{\Sigma}^{-1}$ whenever $\lambda_2 > \lambda_1$. So, by monotone convergence, we have indeed that

$$\mathbb{E}\,\mathrm{tr}\left[(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\, \boldsymbol{\Sigma} \right] \to \mathbb{E}\,\mathrm{tr}\left[(\widehat{\boldsymbol{\Sigma}})^{-1}\, \boldsymbol{\Sigma} \right] = \mathbb{E}\,\mathrm{tr}\left[(\widetilde{\boldsymbol{\Sigma}})^{-1} \right]$$

as $\lambda \to 0^+$. This proves the desired lower bound of the minimax risk, up to the proof of Lemma 4.3. □

Proof of Lemma 4.3. Let us recall that $Y_i = X_i^\top \theta + \varepsilon_i$ with $\varepsilon_i \sim \mathrm{Normal}(0, \sigma^2)$ and that each ε_i is independent of $X_1, \dots, X_n$. We have

$$\frac{1}{n}\sum_{i=1}^n Y_i X_i = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top \theta + \frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i = \widehat{\boldsymbol{\Sigma}}\theta + \frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i,$$

so that

$$\mathbb{E}_\theta[\mathcal{E}(\widehat{\theta}_\lambda)] = \mathbb{E}_\theta\|\widehat{\theta}_\lambda - \theta\|_{\boldsymbol{\Sigma}}^2 = \mathbb{E}\Big\|(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\Big(\widehat{\boldsymbol{\Sigma}}\theta + \frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i\Big) - \theta\Big\|_{\boldsymbol{\Sigma}}^2,$$

but using $(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}(\widehat{\boldsymbol{\Sigma}} + \lambda\, \boldsymbol{I}_d - \lambda\, \boldsymbol{I}_d) = \boldsymbol{I}_d - \lambda(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}$ we obtain

$$\begin{aligned}
\mathbb{E}_\theta[\mathcal{E}(\widehat{\theta}_\lambda)] &= \mathbb{E}\Big\|(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i - \lambda(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\theta\Big\|_{\boldsymbol{\Sigma}}^2 \\
&= \mathbb{E}\bigg[\mathbb{E}\bigg[\Big\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i - \lambda\theta\Big\|^2_{(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}}\bigg| X_1, \dots, X_n\bigg]\bigg] \\
&= \mathbb{E}\bigg[\mathbb{E}\bigg[\Big\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i\Big\|^2_{(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}}\bigg| X_1, \dots, X_n\bigg]\bigg] + \lambda^2\mathbb{E}\|\theta\|^2_{(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}} \\
&= \frac{\sigma^2}{n^2}\mathbb{E}\bigg[\sum_{i=1}^n \|X_i\|^2_{(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}}\bigg] + \lambda^2\mathbb{E}\|\theta\|^2_{(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}},
\end{aligned}$$

where we used repeatedly that $\mathbb{E}[\varepsilon_i | X_1, \dots, X_n] = 0, \mathbb{E}[\varepsilon_i\varepsilon_j | X_1, \dots, X_n] = 0$ for any $i \neq j$ and $\mathbb{E}[\varepsilon_i^2 | X_1, \dots, X_n] = \sigma^2$. But

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \|X_i\|^2_{(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}} &= \frac{1}{n}\sum_{i=1}^n \mathrm{tr}\left[(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}X_i X_i^\top\right] \\
&= \mathrm{tr}\left[(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\Sigma}}_\lambda)^{-1}\widehat{\boldsymbol{\Sigma}}\right],
\end{aligned}$$

which concludes the proof of Lemma 4.3. □

High dimensional statistics and sparsity

5

This chapter is about high-dimensional statistics, in particular high-dimensional linear regression, which corresponds to a setting where the sample size n is smaller than the number of features d. Let us consider again the Gaussian linear model (see Chapter 3), where we observe labels satisfying

$$Y_i = f(X_i) + \varepsilon_i$$

for $i = 1, \dots, n$, where $X_i \in \mathbb{R}^d$ are vectors of features that we assume deterministic, where $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d $\mathrm{Normal}(0, \sigma^2)$ random variables and where f is the regression function that we want to estimate.

Sparse estimation. We consider a set $\mathcal{F} = \{f_1, \dots, f_M\}$ of functions called a *dictionary*, with M which can be much larger than the sample size n. We want to learn from data an estimator of f of the form

$$f_\theta(x) = \sum_{j=1}^{M} \theta_j f_j(x)$$

with the following properties: the *empirical* estimation error

$$\frac{1}{n} \sum_{i=1}^{n} (f_\theta(X_i) - f(X_i))^2$$

is small and the sparsity of θ, namely

$$\|\theta\|_0 = |J(\theta)| = |\{j = 1, \dots, M : \theta_j \neq 0\}|, \tag{5.1}$$

where $|J|$ stands for the cardinality of a set J, is small compared to M. If we are able to satisfy both points, we say that we can find a *sparse* linear combination of elements of $\mathcal{F}$ to estimate f. This task is called *sparse coding* or *sparse estimation*, since it would allow to select a subset of elements from a typically *redundant* dictionary $\mathcal{F}$ to estimate f. Of course, if $M = d$ and $f_j(x) = x_j$, we recover the standard linear regression model, where $f_\theta(x) = x^\top \theta$.

Let us introduce a bunch of notations before diving into the main matter. Here, the features matrix is a $n \times M$ matrix given by

$$\boldsymbol{X} = \begin{bmatrix} f_1(X_1) & \cdots & f_M(X_1) \\ \vdots & \ddots & \vdots \\ f_1(X_n) & \cdots & f_M(X_n) \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^1 \cdots \boldsymbol{X}^M \end{bmatrix}.$$

Let us introduce also

$$\boldsymbol{y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \boldsymbol{f} = \begin{bmatrix} f(X_1) \\ \vdots \\ f(X_n) \end{bmatrix}, \quad \boldsymbol{f}_\theta = \begin{bmatrix} f_\theta(X_1) \\ \vdots \\ f_\theta(X_n) \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

The problem can be rewritten as a Gaussian linear model

$$\boldsymbol{y} = \boldsymbol{X}\,\theta + \boldsymbol{\varepsilon}$$

from Chapter 3, however this time we can have $M \gg n$, namely the matrix $\boldsymbol{X}$ can be overdetermined: it is not full-rank, in this case we say that the dictionary $\mathcal{F}$ is *redundant*. The notation $\|u\|$ will stand for the Euclidean norm of $u \in \mathbb{R}^n$.

Oracle inequalities. We are looking for an estimator $\widehat{\theta}_n$ such that $\|\widehat{\theta}_n\|_0 \ll M$ and

$$\frac{1}{n}\|\boldsymbol{f}_{\widehat{\theta}_n} - \boldsymbol{f}\|^2 \leq \inf_{\theta \in \mathbb{R}^M} \left\{ \frac{1}{n}\|\boldsymbol{f}_\theta - \boldsymbol{f}\|^2 + \mathrm{remainder}(\theta) \right\} \tag{5.2}$$

where $\mathrm{remainder}$ is, ideally, a small quantity that depends on θ, but might depend also on $n, \mathcal{F}$ and σ^2. If $\mathrm{remainder}$ is small, then such an inequality would prove that the estimator $f_{\widehat{\theta}_n}$ performs almost as well as the best linear combination $f^\star = f_{\theta^\star}$ of elements from $\mathcal{F}$, where $\theta^\star \in \mathrm{argmin}_{\theta \in \mathbb{R}^M} \|\boldsymbol{f}_\theta - \boldsymbol{f}\|^2$. We say that $f^\star$ is an *oracle*, since it depends on f, and an inequality of the form (5.2) is called an *oracle inequality*.

This raises the following questions:

- How can we construct a sparse estimator $\widehat{\theta}_n$?
- What is the value of $\mathrm{remainder}$ in the inequality (5.2) ?

We will deal with this problem using a penalization which incudes sparsity in θ. We already talked about the Ridge penalization in Chapter 4, which corresponds to the estimator

$$\widehat{\theta}_n^{\mathsf{ridge}} = \operatorname*{argmin}_{\theta \in \mathbb{R}^M} \left\{ \frac{1}{n}\sum_{i=1}^n (Y_i - f_\theta(X_i))^2 + \frac{\lambda}{2}\|\theta\|^2 \right\}, \tag{5.3}$$

where $\|\theta\|$ is the Euclidean norm of θ, also called ℓ_2-norm. We proved in Chapter 4 that this penalization can be understood as an isotropic Gaussian prior in the Gaussian linear model, and that $\widehat{\theta}_n^{\mathsf{ridge}}$ is the unique solution to the linear system

$$(\boldsymbol{X}^\top \boldsymbol{X} + n\lambda\, \boldsymbol{I})\theta = \boldsymbol{X}^\top \boldsymbol{y},$$

which has no obvious reasons of being sparse.

In this chapter, we will consider another penalization which involves the ℓ_1 norm, since as explained in what follows, it leads to a simple convex problem which defines a sparse estimator $\widehat{\theta}_n^{\text{lasso}}$, coming from a *convex relaxation* principle. This estimator is called the *Lasso* (Least Absolute Shrinkage and Selection Operator), introduced in [18] and is given by

[18]: Tibshirani (1996), 'Regression shrinkage and selection via the lasso'

$$\widehat{\theta}_n^{\text{lasso}} = \operatorname*{argmin}_{\theta\in\Theta} \Big\{ \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_\theta(X_i))^2 + \lambda\|\theta\|_1 \Big\}, \tag{5.4}$$

where $\|\theta\|_1 = \sum_{j=1}^{M}|\theta_j|$ is the ℓ_1 norm of θ and where $\Theta \subset \mathbb{R}^M$ is a convex set, main examples being

- The whole set $\Theta = \mathbb{R}^M$ (no constraint)
- The set $\Theta = [0, +\infty)^M$ (positivity constraint)
- The set $\theta = [-R, R]^M$ for some $R > 0$ (box constraint)

In order to induce sparsity, it is tempting to use as a penalization the "ℓ_0 norm", but this leads to a problem where we would need to try out all subsets $J \subset \{1, \dots, M\}$ and train a linear model on each subset of coordinates J, which means 2^M problems to solve.

Convex relaxation. The ℓ_1 can be understood as a *convex relaxation* of ℓ_0. Indeed, it is easy to see that the *convex envelope*[1] of the function $g_0(x) = \mathbf{1}_{x\neq 0}$ over the interval $[-1, 1]$ is given by $g_1(x) = |x|$, so that the convex envelope of $x \mapsto \|x\|_0$ over $[-1, 1]^M$ is $x \mapsto \|x\|_1$, see Figure 5.1.

1: The convex envelope of $g : [a, b] \to \mathbb{R}$ is, at each point $x \in [a, b]$, the supremum of all convex functions that lie under g, namely $g^{\text{env}}(x) = \sup\{h(x) : h \text{ convex and } h \leq g \text{ over } [a, b]\}$.

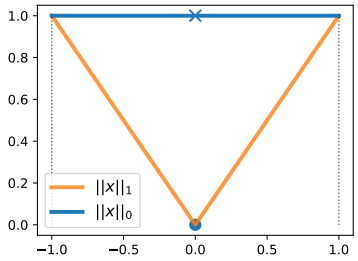


Figure 5.1: The convex envelope of $x \mapsto \|x\|_0$ over $[-1, 1]$ is $x \mapsto \|x\|_1$.

The ℓ_1 norm therefore appears naturally as a convex relaxation of ℓ_0. Another way of understanding it is to consider the following constrained optimization problem

$$\begin{aligned} \min \quad & \|x\|_0 \\ \text{such that} \quad & x \in C \quad \text{and} \quad \|x\|_\infty \leq R, \end{aligned}$$

where C is a convex set[2] which can be reformulated as

$$\begin{aligned} \min \quad & \mathbf{1}^\top u \\ \text{such that} \quad & u \in \{0, 1\}^M, \quad |x_i| \leq R u_i \quad \text{for all} \quad i = 1, \dots, M \\ & x \in C \end{aligned}$$

2: If C is a polyhedron $C : \{x \in \mathbb{R}^d : \boldsymbol{A}\, x \leq b\}$, then we say that the constraints are *linear*.

Such an optimization problem is called a "linear mixed integer program" whenever C is a polyhedron. It is hard to solve exactly, since it requires to try out all possible vectors $u \in \{0, 1\}^M$. A convex relax-

ation of this problem is

$$\begin{aligned} \min \quad & \mathbf{1}^\top u \\ \text{such that} \quad & u \in [0,1]^M, \quad |x_i| \leq R u_i \quad \text{for all} \quad i = 1, \dots, M \\ & x \in C \end{aligned}$$

which can be rewritten as

$$\begin{aligned} \min \quad & \frac{1}{R}\|x\|_1 \\ \text{such that} \quad & x \in C \quad \text{and} \quad \|x\|_\infty \leq R, \end{aligned}$$

where we see that, once again, the ℓ_1 norm naturally appears.

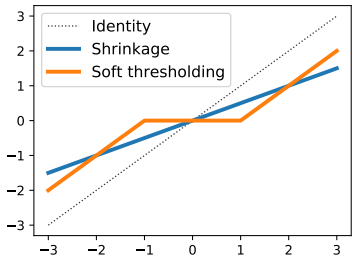


Figure 5.2: Soft-thresholding and shrinkage with $\lambda = 1$ on a single coordinate.

Soft-thresholding. A straightforward computation allows to understand that the ℓ_1 norm induces sparsity. Indeed, we can see easily that

$$\operatorname*{argmin}_{a \in \mathbb{R}} \Big\{ \frac{1}{2}(a-b)^2 + \lambda |a| \Big\} = \operatorname{sign}(b)(|b| - \lambda)_+ \tag{5.5}$$

for any $b \in \mathbb{R}$, where $x_+ = \max(x, 0)$ and $\operatorname{sign}(x) = 1$ if $x > 0$, $\operatorname{sign}(x) = -1$ if $x < 0$ and $\operatorname{sign}(0) = 0$. This proves that

$$\operatorname*{argmin}_{a \in \mathbb{R}^M} \Big\{ \frac{1}{2}\|a-b\|_2^2 + \lambda \|a\|_1 \Big\} = T_\lambda(b) \tag{5.6}$$

for any $b \in \mathbb{R}^M$, where $T_\lambda : \mathbb{R}^M \to \mathbb{R}^M$ is the *soft-thresholding* operator given by

$$(T_\lambda(b))_j = \operatorname{sign}(b_j)(|b_j| - \lambda)_+$$

for $j = 1, \dots, M$, see Figure 5.2. We display also in Figure 5.2 the *shrinkage* operator $(S_\lambda(b))_j = b_j/(1+\lambda)$, which corresponds to the Ridge penalization, since $\operatorname{argmin}_{a \in \mathbb{R}} \{ (a-b)^2 + \lambda a^2 \} = b/(1+\lambda)$.

We observe that the shrinkage operator, which corresponds to the Ridge penalization, does not induce sparsity, while soft-thresholding does. The fact that the ℓ_1 norm induces sparsity (coordinates can be 0) actually comes from the fact that the absolute value is not differentiable at 0. It can be understood geometrically as well, using the fact that a unit ℓ_1 ball has sparse corners at $\pm e_j$ for $j = 1, \dots, M$ (canonical basis vectors) that are *sparse* vectors: when we project a point onto an ℓ_1 ball, we are likely to project onto a corner or an edge, that are sets of sparse points.

The discussion above motivates the use of the ℓ_1 norm to induce sparsity. Let us therefore consider the estimator $\widehat{\theta}_n^{\mathsf{lasso}}$ given by (5.4), that we will simply denote $\widehat{\theta}_n$ in the rest of the chapter. In order to study the statistical properties of this estimator, we need some tools from convex optimization.

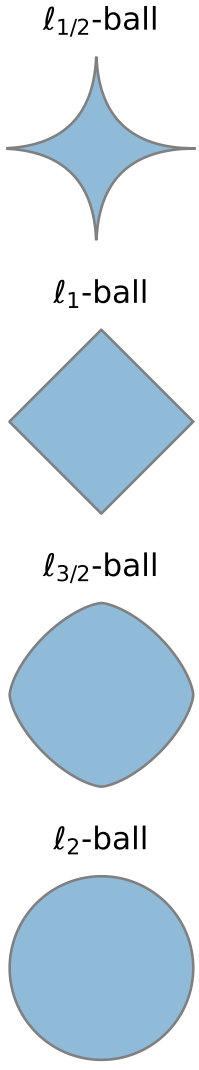


Figure 5.3: Some ℓ_p balls in $\mathbb{R}^2$. The ℓ_1 ball is convex but has spiky corners

5.1 Some tools from convex optimization

Let us consider a convex function $\phi : \mathbb{R}^d \to \mathbb{R}$. A fundamental notion which generalizes the differential to non-differentiable convex functions is the *subdifferential*, see Figure 5.4.

Definition 5.1 We say that $g \in \mathbb{R}^d$ is a *subgradient* of a convex function $\phi : \mathbb{R}^d \to \mathbb{R}$ at $u \in \mathbb{R}^d$ if and only if

$$\phi(v) - \phi(u) \geq g^\top (v - u) \tag{5.7}$$

for any $v \in \mathbb{R}^d$. The set of all subgradients

$$\partial\phi(u) = \big\{ g \in \mathbb{R}^d : \phi(v) - \phi(u) \geq g^\top (v - u) \ \text{ for all } \ v \in \mathbb{R}^d \big\}$$

is called the *subdifferential* of ϕ at u.

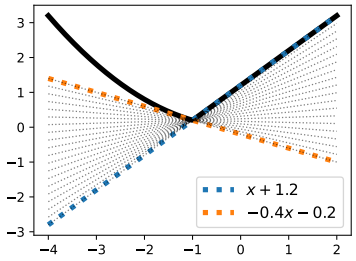


Figure 5.4: An illustration of the subgradients of a convex function at $x = -1$. The subdifferential is equal to the interval $[-0.4, 1]$.

An example is with $\phi(u) = |u|$, where we have $\partial\phi(u) = \{1\}$ if $u > 0$, $\partial\phi(u) = \{-1\}$ if $u < 0$ and $\partial\phi(u) = [-1, 1]$ if $u = 0$. Whenever ϕ is differentiable at u, we have obviously that $\partial\phi(u) = \{\nabla\phi(u)\}$. Another obvious claim is that

$$u^\star \in \operatorname*{argmin}_{u \in \mathbb{R}^d} \phi(u) \quad \text{if and only if} \quad 0 \in \partial\phi(u^\star).$$

Also, it is easy to see that

$$\partial\Big(\sum_{k=1}^K \alpha_k \phi_k(u) \Big) = \sum_{k=1}^K \alpha_k \partial\phi_k(u)$$

whenever $\alpha_k \geq 0$ and ϕ_k are convex functions for all $k = 1, \ldots, K$. Another nice formula allows to express the subdifferential of a maximum of convex functions with the subdifferential of each function. Indeed, if $\phi(u) = \max_{k=1}^K \phi_k(u)$, we have

$$\partial\phi(u) = \operatorname{conv}\Big(\bigcup_{k=1}^K \big\{ \partial\phi_k(u) : \phi_k(u) = \phi(u) \big\} \Big), \tag{5.8}$$

where $\operatorname{conv}(A)$ stands for the convex hull of a set A. For instance, if $\phi_1 : \mathbb{R} \to \mathbb{R}$ and $\phi_2 : \mathbb{R} \to \mathbb{R}$ are convex and differentiable functions, we have

$$\partial \max(\phi_1, \phi_2)(u) = \begin{cases} \{\phi_2'(u)\} & \text{if } \phi_2(u) > \phi_1(u) \\ \{\phi_1'(u)\} & \text{if } \phi_2(u) < \phi_1(u) \\ [\phi_1'(u), \phi_2'(u)] & \text{if } \phi_2(u) = \phi_1(u) \end{cases} \tag{5.9}$$

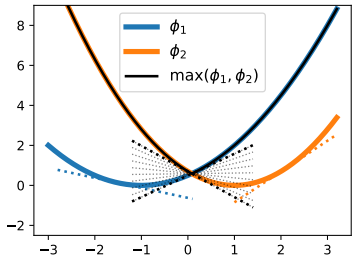


Figure 5.5: An illustration of formula (5.9).

Another useful definition is the indicator function of a convex set $C \subset \mathbb{R}^d$

$$\delta_C(u) = \begin{cases} 0 & \text{if } u \in C \\ +\infty & \text{if } u \notin C. \end{cases}$$

If allows to reformulate a *constrained* problem as an *unconstrained* one, namely to rewrite

$$u^\star \in \operatorname*{argmin}_{u \in C} \phi(u) \quad \text{as} \quad u^\star \in \operatorname*{argmin}_{u \in \mathbb{R}^d} \{\phi(u) + \delta_C(u)\}$$

which means that $0 \in \partial\phi(u^\star) + \partial\delta_C(u^\star)$, namely that there is $g^\star \in \partial\phi(u^\star)$ such that $-g^\star \in \partial\delta_C(u^\star)$. But it is easy to understand what the subdifferential of the indicator function δ_C is, since $g \in \partial\delta_C(u^\star)$ with $u^\star \in C$ means[3] that

$$\delta_C(u) - \delta_C(u^\star) \geq g^\top (u - u^\star) \quad \text{for all} \quad u \in \mathbb{R}^d,$$

but $\delta_C(u^\star) = 0$ so that, for any $u \in C$, we have

$$\partial\delta_C(u) = \big\{ g \in \mathbb{R}^d : g^\top (v - u) \leq 0 \quad \text{for all } v \in C \big\},$$

which is the *normal cone* to C at u, see Figure 5.6.

3: using the definition of the subdifferential

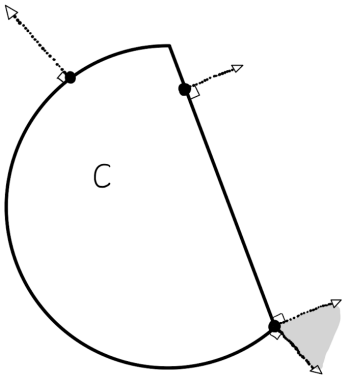


Figure 5.6: Illustration of normal cones

This proves the following proposition.

Proposition 5.1 Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a convex function and $C \subset \mathbb{R}^d$ be a convex set. An optimality criterion for the problem

$$u^\star \in \operatorname*{argmin}_{u \in C} \phi(u)$$

is given by

$$\exists g^\star \in \partial\phi(u^\star) \quad \text{such that} \quad (g^\star)^\top (u - u^\star) \geq 0$$

for all $u \in C$, where $\partial\phi(u^\star)$ is the subdifferential of ϕ at $u^\star$.

Another property about the subdifferential is the following.

Proposition 5.2 (Monotonicity of the subdifferential) Given a convex function $\phi : \mathbb{R}^d \to \mathbb{R}$, any $u_1, u_2 \in \mathbb{R}^d$ and any $g_1 \in \partial\phi(u_1)$ and $g_2 \in \partial\phi(u_2)$, we have

$$(u_1 - u_2)^\top (g_1 - g_2) \geq 0.$$

The proof is straightforward.[4]

4: Just use Definition 5.1 to write that $\phi(u_2) - \phi(u_1) \geq g_1^\top (u_2 - u_1)$ and that $\phi(u_1) - \phi(u_2) \geq g_2^\top (u_1 - u_2)$ and add the two.

Subdifferential of ℓ_1 norm. Let us give as an example the computation of the subdifferential of the ℓ_1 norm. Put $\phi(u) = \|u\|_1$ and note

that it can be rewritten as

$$\|u\|_1 = \max\left\{e^\top u : e \in \{-1,1\}^d\right\} = \max_{i=1,\dots,2^d} \phi_i(u)$$

where we introduced $\phi_i(u) = e_i^\top u$ for $e_1, \dots, e_{2^d}$ the elements of $\{-1,1\}^d$. Note that given $u \in \mathbb{R}^d$, we can choose $e(u) \in \mathbb{R}^d$ such that $e(u)_j = 1$ if $u_j > 0$, $e(u)_j = -1$ if $u_j < 0$ and $e(u)_j = 1$ *or* $e(u)_j = -1$ if $u_j = 0$, in order to obtain that $e(u)^\top u = \|u\|_1$. Let us introduce the set

$$I(u) = \left\{i \in \{1, \dots 2^d\} : e_i^\top u = \|u\|_1\right\}.$$

Each function ϕ_i is differentiable and $\nabla\phi_i(u) = e_i$. So, we can apply Equation (5.8) to obtain that

$$\begin{aligned}\partial\|u\|_1 &= \operatorname{conv}\Big(\bigcup_{i\in I(u)} \{e_i\}\Big) \\ &= \left\{\operatorname{sign}(u) + h : h \in \mathbb{R}^d,\ \|h\|_\infty \le 1,\ h \odot u = 0\right\},\end{aligned} \tag{5.10}$$

where $h \odot u$ is the Hadamard product given by $(h \odot u)_j = h_j u_j$. For instance if $d = 4$ and $u = [17 \quad -42 \quad 0 \quad 3]^\top$ then $\partial\|u\|_1 = \{1\} \times \{-1\} \times [-1,1] \times \{1\}$.

5.2 Oracle inequalities for the Lasso

The material used in this Section is based on [19, 20] . Also, a very nice broader book on the topic of high-dimensional statistics is [21]. Throughout the section, we will assume that the columns are standardized, namely that $\| \boldsymbol{X}^j \|_2 = \sqrt{n}$. This is a rather unrestrictive assumption (we could do without it) that follows good-practice when using linear methods in machine learning. Let us recall that the Lasso estimator is given by

[19]: Bickel et al. (2009), 'Simultaneous analysis of Lasso and Dantzig selector'

[20]: Koltchinskii et al. (2011), 'Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion'

$$\begin{aligned}\widehat{\theta}_n &= \operatorname*{argmin}_{\theta\in\Theta} \left\{\frac{1}{n}\sum_{i=1}^n (Y_i - f_\theta(X_i))^2 + \lambda\|\theta\|_1\right\} \\ &= \operatorname*{argmin}_{\theta\in\Theta} \left\{\frac{1}{n}\|\boldsymbol{y} - \boldsymbol{f}_\theta\|^2 + \lambda\|\theta\|_1\right\},\end{aligned} \tag{5.11}$$

for some convex set $\Theta \subset \mathbb{R}^M$, where

$$\lambda = 2\sigma\sqrt{\frac{2(x + \log M)}{n}}$$

with $x > 0$ which corresponds to a confidence level (see Theorem 5.3 and 5.4 below) and with $\sigma > 0$ the standard-deviation of the noise.

Theorem 5.3 If $\widehat{\theta}_n$ is given by (5.11), we have that

$$\frac{1}{n}\|\boldsymbol{f}_{\widehat{\theta}_n} - \boldsymbol{f}\|^2 \leq \inf_{\theta\in\Theta}\Big\{\frac{1}{n}\|\boldsymbol{f}_\theta - \boldsymbol{f}\|^2 + 2\lambda\|\theta\|_1\Big\}$$

with a probability larger than $1 - 2e^{-x}$.

This inequality is called a *slow oracle inequality* since the remainder is $O(1/\sqrt{n})$. The proof of Theorem 5.3 is given in Section 5.3 below. In order to obtain a faster $O(1/n)$ rate, we need an extra assumption on the matrix of features $\boldsymbol{X}$. Let us introduce the $M \times M$ matrix

$$\boldsymbol{G} = \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} = \Big[\frac{1}{n}\langle \boldsymbol{f}_j, \boldsymbol{f}_{j'}\rangle\Big]_{1\leq j,j'\leq M},$$

where $\boldsymbol{f}_j = [f_j(X_1)\cdots f_j(X_n)] \in \mathbb{R}^n$ and $\langle\cdot,\cdot\rangle$ is the inner product on $\mathbb{R}^n$. Note that if $M > n$, we have that

$$\min_{t\in\mathbb{R}^M\setminus\{0\}}\frac{\sqrt{t^\top\boldsymbol{G}\,t}}{\|t\|} = \min_{t\in\mathbb{R}^M\setminus\{0\}}\frac{\|\boldsymbol{X}\,t\|}{\sqrt{n}\|t\|} = 0$$

since $\boldsymbol{X} : \mathbb{R}^M \to \mathbb{R}^n$ and $\ker(\boldsymbol{X}) \neq \{0\}$ hence the smallest eigenvalue of $\boldsymbol{G}$ is zero.

The assumption we are going to use requires that the smallest eigenvalue *restricted* to *sparse* vectors is positive. For $\theta \in \mathbb{R}^M$ and $c_0 > 0$, let us introduce the cone

$$C_{\theta,c_0} = \big\{t \in \mathbb{R}^M : \|t_{J(\theta)^\complement}\|_1 \leq c_0\|t_{J(\theta)}\|_1\big\}, \tag{5.12}$$

where

- $J(\theta) = \{j \in \{1,\ldots,M\} : \theta_j \neq 0\}$ is the *support* of θ,
- $t_J \in \mathbb{R}^M$ stands for the vector with coordinates $(t_J)_j = t_j$ if $j \in J$ and $(t_J)_j = 0$ for $j \notin J$,
- $J^\complement = \{1,\ldots,M\}\setminus J$.

If $t \in C_{\theta,c_0}$, then both the vectors t and θ almost share the same support, since the coefficients of $t_{J(\theta)}$ dominate those of $t_{J(\theta)^\complement}$. Then, we can introduce

$$\mu_{c_0}(\theta) = \inf\Big\{\mu > 0 : \|t_{J(\theta)}\| \leq \frac{\mu}{\sqrt{n}}\|\boldsymbol{X}\,t\| \ \text{ for all } \ t \in C_{\theta,c_0}\Big\}. \tag{5.13}$$

Note that the function $c_0 \mapsto \mu_{c_0}(\theta)$ is decreasing. If $c_0 = \infty$, then $C_{\theta,c_0} = \mathbb{R}^M$, while if $c_0 = 0$ then $C_{\theta,c_0} = \{t \in \mathbb{R}^M : J(t) = J(\theta)\}$ and in this case

$$\mu_{c_0}(\theta) = \frac{1}{\lambda_{\min}(\boldsymbol{G}_{J(\theta)\times J(\theta)})^{1/2}}$$

the square-root of the inverse of the smallest eigenvalue of the submatrix $(\boldsymbol{G})_{J\times J}$ with $J = J(\theta)$ corresponding to the subset of rows and

columns with index in J.

Theorem 5.4 If $\widehat{\theta}_n$ is given by (5.11) where we replace λ by 2λ, we have that

$$\frac{1}{n}\|\boldsymbol{f}_{\widehat{\theta}_n} - \boldsymbol{f}\|^2 \leq \inf_{\theta\in\Theta}\left\{\frac{1}{n}\|\boldsymbol{f}_\theta - \boldsymbol{f}\|^2 + 18\mu_3(\theta)^2\sigma^2\frac{x+\log M}{n}\|\theta\|_0\right\}$$

with a probability larger than $1-2e^{-x}$, where $\mu_3(\theta)$ is given by (5.13) with $c_0 = 3$ and $\|\theta\|_0$ is the sparsity of θ given by (5.1).

The proof of Theorem 5.4 is given in Section 5.3 below. It proves that the Lasso estimator $\widehat{\theta}_n$ realizes a balance between an *approximation* or *estimation* term $\|\boldsymbol{f}_\theta - \boldsymbol{f}\|^2$ and a *complexity* term which involves the sparsity of θ.

It is a remarkable theorem, since it shows that the Lasso estimator, which is the solution of a simple convex problem, is almost as good as the best *sparse* representation of f using the dictionary $\mathcal{F}$. Indeed, the rate obtained herein is of order $(\log M)\|\theta\|_0/n$, namely the ambient dimension M appears only through $\log M$, while $\|\theta\|_0$ corresponds to the "useful" dimension given by the number of elements of $\mathcal{F}$ that are statistically useful to estimate f.

Definition 5.2 (Restricted eigenvalues) We say that $\boldsymbol{X}$ satisfies the RE(s, c_0) assumption for some $c_0 > 0$ and some $s \in \{1, \dots, M\}$ whenever

$$\kappa(s, c_0) = \min_{\substack{J\subset\{1,\dots,M\}\\ |J|\leq s}} \min_{\substack{t\neq 0\\ \|t_{J^\complement}\|_1\leq c_0\|t_J\|_1}} \frac{\|\boldsymbol{X}\, t\|}{\sqrt{n}\|t_J\|} > 0.$$

Note that we have

$$\kappa(s, c_0) = \inf_{\substack{t\in\mathbb{R}^M\setminus\{0\}\\ \|t\|_0\leq s}} \frac{1}{\mu_{c_0}(t)}.$$

Moreover, whenever $\boldsymbol{X}$ satisfies RE$(s, 1)$, any sub-matrix of $\boldsymbol{X}$ formed by any subset of $2s$ columns from $\boldsymbol{X}$ has full rank.[5]

5: Suppose by contradiction that there is $t \in \mathbb{R}^M$ such that $\|t\|_0 = 2s$ and $\boldsymbol{X}\, t = 0$. Then, we can choose disjoint sets $J_0, J_1 \subset \{1, \dots, M\}$ such that $J(t) = J_0 \cup J_1$ with $|J_0| = s$ and $|J_1| = s$ and such that $\|t_{J_1}\|_1 \leq \|t_{J_0}\|_1$. But obviously $\|t_{J_1}\|_1 = \|t_{J_0^\complement}\|_1$ so $\|t_{J_0^\complement}\|_1 \leq \|t_{J_0}\|_1$, which contradicts the RE$(s, 1)$ assumption.

An immediate corollary of Theorem 5.4 is the following oracle inequality, which holds under the RE$(s, 3)$ assumption:

$$\frac{1}{n}\|\boldsymbol{f}_{\widehat{\theta}_n} - \boldsymbol{f}\|^2 \leq \inf_{\substack{\theta\in\Theta\\ \|\theta\|_0\leq s}}\left\{\frac{1}{n}\|\boldsymbol{f}_\theta - \boldsymbol{f}\|^2 + \frac{18\sigma^2}{\kappa(s,3)^2}\,\frac{s(x+\log M)}{n}\right\}$$

with a probability larger than $1-e^{-x}$. In this inequality, the convergence rate is of order $(s\log M)/n$, where s is the sparsity of the best θ. Let us finish this chapter with several remarks before diving into the proofs.

- The rate of convergence depends on the ambient dimension M only through $\log M$ and depends linearly on the sparsity s of $\theta \in \mathbb{R}^M$. This is a remarkable property called *dimension reduction* or *adaptation to the sparsity* of the Lasso estimator.
- This is not the optimal rate, the minimax optimal rate among s-sparse vector (and in a ℓ_q ball) being $s\log(M/s)/n$, see [22] .
- There are several improvements of these oracle inequalities in literature: beyond Gaussian noise, using the integrated estimator error $\int_{\mathbb{R}^M}(f_{\widehat{\theta}_n}(x) - f(x))^2 P_X(dx)$ instead of the empirical one used here, and we can remove the dependency of λ on the confidence level $x > 0$.
- From Theorem 5.4, we can derive bounds on the estimator error $\|\widehat{\theta}_n - \theta^\star\|_p$ (measured by the ℓ_p norm, $p \geq 1$) of the true parameter $\theta^\star$ and the we can give guarantees on the *signed support recovery* of the parameter, through controls on the probability $\mathbb{P}[\mathrm{sign}(\widehat{\theta}_n) = \mathrm{sign}(\theta^\star)]$, see [23] .

[22]: Verzelen (2012), 'Minimax risks for sparse regressions: Ultra-high dimensional phenomenons'

[23]: Zhao et al. (2006), 'On model selection consistency of Lasso'

5.3 Proofs

5.3.1 Proof of Theorem 5.3

Let us start with the noise. It is fairly easy, since

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \varepsilon_i f_j(X_i) &\sim \mathrm{Normal}\left(0, \frac{\sigma^2}{n^2}\sum_{i=1}^n f_j(X_i)^2\right)\\ &= \mathrm{Normal}\left(0, \frac{\sigma^2}{n^2}\|\,\boldsymbol{X}_j\,\|^2\right)\\ &= \mathrm{Normal}\left(0, \frac{\sigma^2}{n}\right).\end{aligned}$$

We assumed that $\|\,\boldsymbol{X}_j\,\| = \sqrt{n}$

So, recalling that $\mathbb{P}[|Z| \geq z] \leq 2e^{-z^2/2}$ whenever $Z \sim \mathrm{Normal}(0,1)$ for any $z > 0$,we obtain

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f_j(X_i)\right| \geq \sigma\sqrt{\frac{2x}{n}}\right] \leq 2e^{-x}$$

and using an union bound, we obtain that the event

$$A = \bigcap_{j=1}^M \left\{\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f_j(X_i)\right| \leq \sigma\sqrt{\frac{2(x+\log M)}{n}}\right\}$$

satisfies $\mathbb{P}[A] \geq 1 - 2e^{-x}$. The definition of $\widehat{\theta}$ entails that

$$\frac{1}{n}\|\,\boldsymbol{y} - \boldsymbol{f}_{\widehat{\theta}}\|^2 + \lambda\|\widehat{\theta}\|_1 \leq \frac{1}{n}\|\,\boldsymbol{y} - \boldsymbol{f}_\theta\|^2 + \lambda\|\theta\|_1 \qquad (5.14)$$

for any $\theta \in \Theta$ and an easy computation gives

$$\begin{aligned}
&\frac{1}{n}\|\boldsymbol{y}-\boldsymbol{f}_{\widehat{\theta}}\|^2 - \frac{1}{n}\|\boldsymbol{y}-\boldsymbol{f}_{\theta}\|^2 \\
&= \frac{1}{n}\|\boldsymbol{f}_{\widehat{\theta}}\|^2 + \frac{1}{n}\|\boldsymbol{f}\|^2 + \frac{2}{n}\langle \boldsymbol{y}, \boldsymbol{f}_{\theta} - \boldsymbol{f}_{\widehat{\theta}}\rangle \\
&= \frac{1}{n}\|\boldsymbol{f}_{\widehat{\theta}}\|^2 + \frac{1}{n}\|\boldsymbol{f}\|^2 + \frac{2}{n}\langle \boldsymbol{f}, \boldsymbol{f}_{\theta} - \boldsymbol{f}_{\widehat{\theta}}\rangle + \frac{2}{n}\langle \boldsymbol{\varepsilon}, \boldsymbol{f}_{\theta} - \boldsymbol{f}_{\widehat{\theta}}\rangle \\
&= \frac{1}{n}\|\boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}\|^2 - \frac{1}{n}\|\boldsymbol{f}_{\theta} - \boldsymbol{f}\|^2 + \frac{2}{n}\langle \boldsymbol{\varepsilon}, \boldsymbol{f}_{\theta} - \boldsymbol{f}_{\widehat{\theta}}\rangle.
\end{aligned} \tag{5.15}$$

But on the event A, we have that

$$\begin{aligned}
\frac{2}{n}|\langle \boldsymbol{\varepsilon}, \boldsymbol{f}_{\theta} - \boldsymbol{f}_{\widehat{\theta}}\rangle| &= \Big|\frac{2}{n}\sum_{j=1}^{M}(\widehat{\theta}_j - \theta_j)\sum_{i=1}^{n}\varepsilon_i f_j(X_i)\Big| \\
&\leq \sum_{j=1}^{M}|\widehat{\theta}_j - \theta_j| 2\sigma\sqrt{\frac{2(x+\log M)}{n}} \\
&= \lambda\|\widehat{\theta} - \theta\|_1,
\end{aligned} \tag{5.16}$$

so that, combining Inequalities (5.14), (5.15) and (5.16), we end up with

$$\begin{aligned}
\frac{1}{n}\|\boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}\|^2 &\leq \frac{1}{n}\|\boldsymbol{f}_{\theta} - \boldsymbol{f}\|^2 + \lambda\|\widehat{\theta} - \theta\|_1 + \lambda\|\theta\|_1 - \lambda\|\widehat{\theta}\|_1 \\
&\leq \frac{1}{n}\|\boldsymbol{f}_{\theta} - \boldsymbol{f}\|^2 + 2\lambda\|\theta\|_1,
\end{aligned}$$

which concludes the proof of Theorem 5.3. □

5.3.2 Proof of Theorem 5.4

Let us recall that we study the estimator

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta\in\Theta}\big\{R_n(\theta) + 2\,\mathrm{pen}(\theta)\big\}, \tag{5.17}$$

where $\Theta \subset \Theta$ is a convex set, where

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_{\theta}(X_i))^2 = \frac{1}{n}\|\boldsymbol{y}-\boldsymbol{f}_{\theta}\|^2$$

is convex and differentiable and $\mathrm{pen}(\theta) = \lambda\|\theta\|_1$ is convex. We have $\nabla R_n(\theta) = -\frac{2}{n}\boldsymbol{X}^{\top}(\boldsymbol{y} - \boldsymbol{X}\,\theta)$ so that

$$\partial(R_n + \mathrm{pen})(\theta) = -\frac{2}{n}\boldsymbol{X}^{\top}(\boldsymbol{y} - \boldsymbol{X}\,\theta) + 2\lambda\partial\|\theta\|_1.$$

Using Proposition 5.1, we have that (5.17) is equivalent to the fact that there is $\widehat{g} \in \partial\|\widehat{\theta}\|_1$ such that

$$\Big\langle -\frac{2}{n}\boldsymbol{X}^{\top}(\boldsymbol{y} - \boldsymbol{X}\,\widehat{\theta}) + 2\lambda\widehat{g}, \widehat{\theta} - \theta\Big\rangle \leq 0 \quad \text{for all} \quad \theta \in \Theta. \tag{5.18}$$

Let us choose for now an arbitrary $\theta \in \Theta$ and note $J = J(\theta)$. We can rewrite inequality (5.18) as

$$\frac{2}{n}\big\langle \boldsymbol{X}^\top \boldsymbol{X}\,\widehat{\theta}, \widehat{\theta} - \theta\big\rangle - \frac{2}{n}\big\langle \boldsymbol{X}^\top \boldsymbol{y}, \widehat{\theta} - \theta\big\rangle + 2\lambda\langle \widehat{g}, \widehat{\theta} - \theta\rangle \leq 0,$$

and, recalling that $\boldsymbol{X}\,\theta = \boldsymbol{f}_\theta$ and that $\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\varepsilon}$, we have

$$\begin{aligned}\frac{2}{n}\langle \boldsymbol{f}_{\widehat{\theta}}, \boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}_\theta\rangle - \frac{2}{n}\langle \boldsymbol{f}, \boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}_\theta\rangle \\ - \frac{2}{n}\langle \boldsymbol{\varepsilon}, \boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}_\theta\rangle + 2\lambda\langle \widehat{g}, \widehat{\theta} - \theta\rangle \leq 0,\end{aligned}$$

that we can rewrite as

$$\begin{aligned}\frac{2}{n}\langle \boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}, \boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}_\theta\rangle + 2\lambda\langle \widehat{g} - g, \widehat{\theta} - \theta\rangle \\ \leq -2\lambda\langle g, \widehat{\theta} - \theta\rangle + \frac{2}{n}\langle \boldsymbol{\varepsilon}, \boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}_\theta\rangle\end{aligned}$$

for any $g \in \partial\|\theta\|_1$. But, the monotonicity of the subdifferential entails that $\langle \widehat{g} - g, \widehat{\theta} - \theta\rangle \geq 0$ and and easy computation gives (Al-Kachi)

$$2\langle \boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}, \boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}_\theta\rangle = \|\boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}_\theta\|^2 + \|\boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}\|^2 - \|\boldsymbol{f}_\theta - \boldsymbol{f}\|^2$$

so that we end up with

$$\begin{aligned}\frac{1}{n}\|\boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}\|^2 + \frac{1}{n}\|\boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}_\theta\|^2 \\ \leq \frac{1}{n}\|\boldsymbol{f}_\theta - \boldsymbol{f}\|^2 - 2\lambda\langle g, \widehat{\theta} - \theta\rangle + \frac{2}{n}\langle \boldsymbol{\varepsilon}, \boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}_\theta\rangle.\end{aligned}$$

for any $g \in \partial\|\theta\|_1$. Now, we need to use what the subdifferential of the ℓ_1 norm is. Using (5.10), we can write $g = \operatorname{sign}(\theta) + h$ for any h such that where $h_J = 0$ and $\|h\|_\infty \leq 1$. We have

$$|\langle \operatorname{sign}(\theta), \widehat{\theta} - \theta\rangle| = |\langle \operatorname{sign}(\theta), (\widehat{\theta} - \theta)_J\rangle| \leq \|(\widehat{\theta} - \theta)_J\|_1.$$

and we can choose h such that

$$\langle h, \widehat{\theta} - \theta\rangle = \langle h, \widehat{\theta}_{J^\complement}\rangle = \|\widehat{\theta}_{J^\complement}\|_1 = \|(\widehat{\theta} - \theta)_{J^\complement}\|_1.$$

This leads to

$$\begin{aligned}\frac{1}{n}\|\boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}\|^2 + \frac{1}{n}\|\boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}_\theta\|^2 + 2\lambda\|\widehat{\theta}_{J^\complement}\|_1 \\ \leq \frac{1}{n}\|\boldsymbol{f}_\theta - \boldsymbol{f}\|^2 + 2\lambda\|(\widehat{\theta} - \theta)_J\|_1 + \frac{2}{n}\langle \boldsymbol{\varepsilon}, \boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}_\theta\rangle.\end{aligned}$$

If $\|\boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}\|^2 + \|\boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}_\theta\|^2 - \|\boldsymbol{f}_\theta - \boldsymbol{f}\|^2 \leq 0$, then $\|\boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}\|^2 \leq \|\boldsymbol{f}_\theta - \boldsymbol{f}\|^2$, which concludes the proof, so let us consider from now on the other case, which entails that

$$2\lambda\|\widehat{\theta}_{J^\complement}\|_1 \leq 2\lambda\|(\widehat{\theta} - \theta)_J\|_1 + \frac{2}{n}\langle \boldsymbol{\varepsilon}, \boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}_\theta\rangle. \tag{5.19}$$

On the event A,[6] we have $\frac{1}{n}\| \boldsymbol{X}^\top \boldsymbol{\varepsilon} \|_\infty \leq \lambda/2$ where we recall that $\lambda = 2\sigma\sqrt{2(x+\log M)/n}$. This entails

6: see the proof of Theorem 5.3

$$\begin{aligned}\frac{2}{n}\langle \boldsymbol{\varepsilon}, \boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}_\theta \rangle &= \frac{2}{n}\langle \boldsymbol{X}^\top \boldsymbol{\varepsilon}, \widehat{\theta} - \theta \rangle \\ &= \frac{2}{n}\langle \boldsymbol{X}^\top \boldsymbol{\varepsilon}, (\widehat{\theta} - \theta)_J \rangle + \frac{2}{n}\langle \boldsymbol{X}^\top \boldsymbol{\varepsilon}, \widehat{\theta}_{J^\complement} \rangle \\ &\leq \lambda \|(\widehat{\theta} - \theta)_J\|_1 + \lambda \|\widehat{\theta}_{J^\complement}\|_1,\end{aligned}$$

which combined with (5.19) gives $\|\widehat{\theta}_{J^\complement}\|_1 \leq 3\|(\widehat{\theta} - \theta)_J\|_1$ on A, namely that $\widehat{\theta} - \theta \in C_{\theta,3}$ where we recall that $C_{\theta,3}$ is given by (5.12). This means that, on A, using (5.13), we can write in this case that

$$\|(\widehat{\theta} - \theta)_J\| \leq \frac{\mu(\theta)}{\sqrt{n}} \| \boldsymbol{X}(\widehat{\theta} - \theta)\|$$

with $\mu(\theta) = \mu_3(\theta)$, which entails that

$$\begin{aligned}&\frac{1}{n}\|\boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}\|^2 + \frac{1}{n}\|\boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}_\theta\|^2 + \lambda\|\widehat{\theta}_{J^\complement}\|_1 \\ &\qquad \leq \frac{1}{n}\|\boldsymbol{f}_\theta - \boldsymbol{f}\|^2 + 3\lambda\|(\widehat{\theta} - \theta)_J\|_1 \\ &\qquad \leq \frac{1}{n}\|\boldsymbol{f}_\theta - \boldsymbol{f}\|^2 + 3\lambda|J|^{1/2}\|(\widehat{\theta} - \theta)_J\| \qquad \text{using Cauchy-Schwarz} \\ &\qquad \leq \frac{1}{n}\|\boldsymbol{f}_\theta - \boldsymbol{f}\|^2 + 3\mu(\theta)\lambda|J|^{1/2}\frac{1}{\sqrt{n}}\| \boldsymbol{X}(\widehat{\theta} - \theta)\| \\ &\qquad = \frac{1}{n}\|\boldsymbol{f}_\theta - \boldsymbol{f}\|^2 + 3\mu(\theta)\lambda|J|^{1/2}\frac{1}{\sqrt{n}}\|\boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}_\theta\|,\end{aligned}$$

which concludes the proof since using the fact that $ax - x^2 \leq a^2/4$ for any $x, a > 0$, we have

$$\frac{1}{n}\|\boldsymbol{f}_{\widehat{\theta}} - \boldsymbol{f}\|^2 \leq \frac{1}{n}\|\boldsymbol{f}_\theta - \boldsymbol{f}\|^2 + 9\mu(\theta)^2\lambda^2|J|.$$

□

Maximum likelihood estimation, application to exponential models

6

Maximum likelihood estimation is a fundamental and very general approach for statistical inference of model parameters. In this chapter, we consider a statistical experiment with data $X : \Omega \to \mathcal{X}$ and model $\{P_\theta : \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^d$ dominated by a σ-finite measure μ on $\mathcal{X}$, so that we can define the family of densities

$$f_\theta(x) = \frac{dP_\theta}{d\mu(x)}$$

This uses the Radon-Nikodym theorem

on $\mathcal{X}$, see Definition 1.5 from Chapter 1.

Definition 6.1 The *likelihood* function $L : \Theta \to \mathbb{R}^+$ is defined as

$$L(\theta) := L(\theta; X) = f_\theta(X).$$

We also introduce the *log-likelihood* $\ell : \Theta \to \mathbb{R}$ given by

$$\ell(\theta) := \ell(\theta; X) := \log f_\theta(X)$$

whenever $f_\theta(X) > 0$ almost surely for all $\theta \in \Theta$.

The likelihood and log-likelihood are random functions since they depend on the data X.

Maximum likelihood estimation. We want to infer θ, namely we want to find $\theta_0 \in \Theta$ such that $X \sim P_{\theta_0}$. Given the data X, the likelihood $L(\theta; X)$ is the "probability" to observe X whenever the parameter is θ. So, in order to find θ_0, it is natural to look for θ that maximizes $\theta \mapsto L(\theta)$ (or equivalently $\theta \mapsto \ell(\theta)$) on Θ, since such a θ would *maximize the probability of observing* X (since we do observe it). This simple principle is fundamental and motivates the following definition.

Definition 6.2 (Maximum Likelihood Estimator) We say that $\widehat{\theta} \in \Theta$ is a *maximum likelihood estimator* (MLE) if

$$L(\widehat{\theta}; X) = \sup_{\theta \in \Theta} L(\theta; X),$$

or equivalently that

$$\widehat{\theta} \in \operatorname*{argmax}_{\theta \in \Theta} L(\theta; X).$$

Whenever it exists, a MLE $\widehat{\theta}$ depends on X and on the choice of the model $\{f_\theta : \theta \in \Theta\}$. If $X = (X_1, \dots, X_n)$ with X_i iid and density f_θ then

$$L_n(\theta) = L(\theta; X) = L(\theta; X_1, \dots, X_n) = \prod_{i=1}^{n} f_\theta(X_i)$$

and

$$\ell_n(\theta) = \ell(\theta; X) = \ell(\theta; X_1, \dots, X_n) = \sum_{i=1}^{n} \log f_\theta(X_i).$$

In this case we say that L and ℓ are the likelihood and log-likelihood functions of the *n-sampled experiment*. The existence and uniqueness of the maximum likelihood estimator is not granted in general (even on non-pathological examples).

In the next Chapter, we will see an example where the maximum likelihood estimator of logistic regression (a widely used model for classification) does not exist when data is linearly separable

6.1 A theoretical motivation

Let $X_1, \dots, X_n$ be iid with distribution $P_{\theta_0} = f_{\theta_0} \cdot \mu$. Assuming $\mathbb{E}_{\theta_0} |\log f_\theta(X_1)| < +\infty$ for any $\theta \in \Theta$, we have that

$$\begin{aligned} \frac{1}{n}\ell_n(\theta_0) - \frac{1}{n}\ell_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n} \big(\log f_{\theta_0}(X_i) - \log f_\theta(X_i)\big) \\ &\xrightarrow{\mathbb{P}} \mathbb{E}_{\theta_0}\big[\log f_{\theta_0}(X_1) - \log f_\theta(X_1)\big] \\ &= h(P_{\theta_0}, P_\theta), \end{aligned} \tag{6.1}$$

This convergence holds in P_{θ_0}-probability using the law of large numbers

where we introduce the quantity

$$h(P_{\theta_0}, P_\theta) = \int_{\mathcal{X}} \log\Big(\frac{f_{\theta_0}(x)}{f_\theta(x)}\Big) f_{\theta_0}(x)\mu(dx)$$

which is called the *relative entropy* between P_{θ_0} and P_θ. This is a fundamental quantity which deserves a definition.

Definition 6.3 Let P and Q be two probability measures on a measurable space $(\Omega, \mathcal{A})$. The quantity given by

$$h(P, Q) = \mathbb{E}_P\Big[\log\Big(\frac{dP}{dQ}\Big)\Big] = \int \log\Big(\frac{dP}{dQ}(\omega)\Big) P(d\omega)$$

when $P \ll Q$ and $h(P, Q) = +\infty$ otherwise is called the *relative entropy* between P and Q. It is also called the *Kullback-Liebler divergence* or the *information divergence*.

Let us give some properties about $h(P, Q)$. First of all, if $P \ll Q$ then

$$h(P, Q) = \mathbb{E}_P\Big[\log \frac{dP}{dQ}\Big] = \mathbb{E}_Q\Big[\frac{dP}{dQ} \log \frac{dP}{dQ}\Big].$$

It is always well-defined, eventually it is $+\infty$ since $x \log \geq -e^{-1}$ for any $x \in (0, +\infty)$. Also, if $P \ll Q$ then

$$h(P,Q) = \mathbb{E}_Q\Big[\frac{dP}{dQ}\log\frac{dP}{dQ}\Big] \geq \mathbb{E}_Q\Big[\frac{dP}{dQ}\Big]\log\mathbb{E}_Q\Big[\frac{dP}{dQ}\Big] = 0, \qquad \text{using Jensen's inequality}$$

and note also that $h(P,Q) = 0 \Leftrightarrow P = Q$ since $\phi(x) = x\log x$ is strictly convex.

Using (6.1), we have that $(\ell_n(\theta_0) - \ell_n(\theta))/n \approx h(P_{\theta_0}, P_\theta)$ for n large and as explained above $h(P_{\theta_0}, P_\theta) = 0$ iff $P_{\theta_0} = P_\theta$, namely iff $\theta = \theta_0$ whenever the model is identifiable (see Definition 1.4 from Chapter 1). This motivates the use of the maximum likelihood estimator, since maximizing $\ell_n(\theta)$ means minimizing $(\ell_n(\theta_0) - \ell_n(\theta))/n$, that we expect to be minimal at $\theta \approx \theta_0$ when the model is identifiable.

The MLE is a very general principle that can be used for virtually any statistical model. Its theoretical study requires smoothness assumptions on the family of densities f_θ regarded as functions of the parameter θ. In this Chapter, we study the MLE in the specific family of *exponential models* for two reasons: exponential models contain most parametric models that are of interest in practice, and their study is interesting by itself, since they are the basis of generalized linear models that we will study in the next Chapter. We will see that when $\{f_\theta : \theta \in \Theta\}$ is an exponential model, the MLE is easy to study, since it corresponds to another estimation approach called *method of moments*. But, keep in mind that MLE goes way beyond the setting considered here.

Example 6.1 Let us consider $X \sim \mathrm{Gamma}(a, \lambda)$, namely the likelihood function

$$L(a,\lambda) = \frac{\lambda^a}{\Gamma(a)}x^{a-1}e^{-\lambda x} = \exp(\theta^\top T(x) - \log Z(\theta))$$

with $\theta = [a-1 \ \lambda]^\top$, $T(x) = [\log x \ \ -x]^\top$ and $Z(\theta) = \Gamma(a)/\lambda^a$. This is an example of a so-called *exponential model*. Most parametric distributions can be written in this way (Poisson, Exponential, Binomial, Gaussian, etc.), using if necessary a reparametrization (we defined $\theta = [a-1 \ \lambda]^\top$ in this example).

6.2 Exponential models

Let us first describe the so-called *canonical exponotial model*.

Definition 6.4 (Canonical exponential model) Let μ be a σ-finite measure on a measurable space $\mathcal{X}$ and let $T : \mathcal{X} \to \mathbb{R}^d$ be a measur-

able function. We define

$$\Theta_{\mathrm{dom}} = \left\{\theta \in \mathbb{R}^d : Z(\theta) := \int_{\mathcal{X}} e^{\theta^\top T(x)} \mu(dx) < +\infty\right\}$$

and $\Theta = \mathrm{int}(\Theta_{\mathrm{dom}})$, the interior of Θ_{dom}. We introduce the density

$$f_\theta(x) = \exp\left(\theta^\top T(x) - \log Z(\theta)\right)$$

with respect to μ for $\theta \in \Theta$ and define $P_\theta = f_\theta \cdot \mu$. The family $\{P_\theta : \theta \in \Theta\}$ is called a *canonical exponential model* and the function $\theta \mapsto Z(\theta)$ is called the *partition function* of the model. Also, we call T the *sufficient statistic* of the model.

We discussed sufficient statistics in Section 1.3 of Chapter 1

We consider $\{P_\theta \in \Theta\}$ instead of $\{P_\theta \in \Theta_{\mathrm{dom}}\}$ since we will perform differential calculus and use the inversion theorem on this open domain.

Proposition 6.1 The set Θ_{dom} is convex (if it is not empty) and the function $\Theta_{\mathrm{dom}} \to \mathbb{R}$ defined by $\Theta \mapsto \log Z(\theta)$ is convex.

Proof. Note that if $\theta_1, \theta_2 \in \Theta_{\mathrm{dom}}$ and $\alpha \in [0,1]$ we have

$$\begin{aligned} Z(\alpha\theta_1 + (1-\alpha)\theta_2) &= \int_{\mathcal{X}} \left(e^{\theta_1^\top T(x)}\right)^\alpha \left(e^{\theta_2^\top T(x)}\right)^{1-\alpha} \mu(dx) \\ &\le \left(\int_{\mathcal{X}} e^{\theta_1^\top T(x)} \mu(dx)\right)^\alpha \\ &\quad \times \left(\int_{\mathcal{X}} e^{\theta_2^\top T(x)} \mu(dx)\right)^{1-\alpha} < +\infty \end{aligned}$$

Using Hölder's inequality

which proves that $\alpha\theta_1 + (1-\alpha)\theta_2 \in \Theta_{\mathrm{dom}}$ and also that $\log Z$ is convex since we have

$$\log Z(\alpha\theta_1 + (1-\alpha)\theta_2) \le \alpha \log Z(\theta_1) + (1-\alpha)\log Z(\theta_2). \quad \square$$

Definition 6.5 (Canonical and minimal exponential model) We say that a canonical exponential model is minimal if $T(x)$ does not belong, P_θ-almost surely for all $\theta \in \Theta$, to any hyperplane $H \subset \mathbb{R}^d$, namely if

$$P_\theta[\{x \in \mathcal{X} : T(x) \in H\}] < 1$$

for any hyperplane H and any $\theta \in \Theta$.

Proposition 6.2 If a canonical exponential model is minimal, then it is identifiable.

Proof. Let us consider $\theta_1, \theta_2 \in \Theta$ such that $\theta_1 \neq \theta_2$ and such that $P_{\theta_1} = P_{\theta_2}$. This entails $f_{\theta_1}(x) = f_{\theta_2}(x)$ for any $x \in \mathcal{X}$ and in

particular that

$$(\theta_1 - \theta_2)^\top T(X) - \log(Z(\theta_1)/Z(\theta_2)) = 0$$

P_θ-almost surely for any $\theta \in \Theta$, which contradicts the fact that the model is minimal according to Definition 6.5 with the hyperplane $H = \{x \in \mathcal{X} : (\theta_1 - \theta_2)^\top T(x) = \log(Z(\theta_1)/Z(\theta_2))\}$. □

From now on, we suppose that the model is *minimal*. It is a natural assumption: it means that the coordinates of the sufficient statistic $T(X)$ are not almost-surely linearly redundant. Let us recall also that $\Theta = \mathrm{int}(\Theta_{\mathrm{dom}}) \neq \emptyset$.

Theorem 6.3 Consider a canonical exponential model. Its partition function $\theta \mapsto \log Z(\theta)$ is C^∞ on Θ and we have

$$\mathbb{E}_\theta[|T_j(X)|^k] < +\infty$$

for any $j = 1, \dots, d$, any $k \in \mathbb{N}$ (all the moments of $T(X)$ with $X \sim P_\theta$ are finite) and any $\theta \in \Theta$. Furthermore, the following relations

$$\nabla \log Z(\theta) = \mathbb{E}_\theta[T(X)] \quad \text{and} \quad \nabla^2 \log Z(\theta) = \mathbb{V}_\theta[T(X)]$$

$\nabla F(\theta)$ is the gradient of F at θ while $\nabla^2 F(\theta)$ is the Hessian matrix of F at θ

hold for any $\theta \in \Theta$.

The proof of Theorem 6.3 is left as an exercise, where we just need to use dominated convergence to inverse differentiation and expectation. Let us just do the following computation, which explains why the first moment of the sufficient statistic is equal to the gradient of the partition function:

$$\begin{aligned} \nabla \log Z(\theta) &= \nabla \log \mathbb{E}_{X\sim\mu}\big[\exp(\theta^\top T(X))\big] \\ &= \frac{\mathbb{E}_{X\sim\mu}\big[T(X)\exp(\theta^\top T(X))\big]}{\mathbb{E}_{X\sim\mu}\big[\exp(\theta^\top T(X))\big]} \\ &= \mathbb{E}_\theta[T(X)], \end{aligned}$$

We use the notation $X \sim \mu$ to indicate that we integrate with respect to μ even if μ is not a probability measure

where we just used the definition of P_θ in the last equality.

Corollary 6.4 We have $\nabla^2 \log Z(\theta) \succ 0$ for all $\theta \in \Theta$ iff the model is minimal.

Proof. For any $u \in \mathbb{R}^d$ and $\theta \in \Theta$ we have $u^\top \nabla^2 \log Z(\theta) u = u^\top \mathbb{V}_\theta[T(X)]u = \mathbb{V}_\theta[u^\top T(X)]$ so that $\nabla^2 \log Z(\theta)$ is not positive definite iff $u^\top T(X)$ is constant P_θ almost-surely. □

Note that we recover here the fact that $\theta \mapsto \log Z(\theta)$ is strictly convex when the model is minimal, since $\nabla^2 \log Z(\theta) \succ 0$ for any $\theta \in \Theta$.

A consequence of this is that the differential of $S(\theta) = \nabla \log Z(\theta)$, which is the Hessian matrix $\nabla^2 \log Z(\theta)$, is *invertible* for any $\theta \in \Theta$.

The following computation is insightful:

$$\begin{aligned} h(P_\theta, P_{\theta'}) &= \mathbb{E}_\theta\Big[\log \frac{dP_\theta}{dP_{\theta'}}(X)\Big] \\ &= \mathbb{E}_\theta\Big[(\theta - \theta')^\top T(X) - \log \frac{Z(\theta)}{Z(\theta')}\Big] \\ &= \log Z(\theta') - \log Z(\theta) - (\theta' - \theta)^\top \nabla \log Z(\theta) \end{aligned}$$

which means that $h(P_\theta, P_{\theta'})$ is equivalent to a local "linearization" of $\log Z(\theta)$ and therefore approximately equal to

$$h(P_\theta, P_{\theta'}) \approx \frac{1}{2}(\theta' - \theta)^\top \nabla^2 \log Z(\theta)(\theta' - \theta)$$

for $\theta \approx \theta'$. This makes a connection between the *local curvature* of the model θ and its identifiability. Another proposition goes as follows.

Proposition 6.5 The function $\theta \mapsto \log Z(\theta)$ is injective on Θ if and only if the model is identifiable.

Proof. We have that

$$h(P_\theta, P_{\theta'}) + h(P_{\theta'}, P_\theta) = \langle \nabla \log Z(\theta) - \nabla \log Z(\theta'), \theta - \theta' \rangle$$

which is ≥ 0 by convexity. If $\log Z(\theta') = \log Z(\theta)$ and $\theta \neq \theta'$ then $h(P_\theta, P_{\theta'}) = h(P_{\theta'}, P_\theta) = 0$ so that $P_\theta = P_{\theta'}$. Now, if $P_\theta = P_{\theta'}$ for $\theta \neq \theta'$, we have $\mathbb{E}_\theta[T(X)] = \mathbb{E}_{\theta'}[T(X)]$ and therefore $\log Z(\theta') = \log Z(\theta)$ using Theorem 6.3. □

We proved the following properties about the function $S : \Theta \to \mathbb{R}^d$ given by $S(\theta) = \nabla \log Z(\theta)$:

1. Whenever the model is identifiable, we know that S is injective on Θ using Proposition 6.5;
2. We know that S is C^∞ on Θ using Theorem 6.3;
3. The differential of S is invertible on Θ if the model is minimal using Corollary 6.4.

We can therefore apply the theorem of global inversion to say that S is a diffeomorphism, that $S(\Theta)$ is open and that its inverse S^{-1} is also C^∞. We therefore proved the following theorem.

Theorem 6.6 In a minimal and canonical exponential model, we have that $S : \Theta \to S(\Theta)$ given by $S(\theta) = \nabla \log Z(\theta)$ is a diffeomorphism, that $S(\Theta)$ is open and that S^{-1} is also C^∞.

6.3 Maximum likelihood estimation in an exponential model

Let us consider a iid sample $X_1, \dots, X_n$ with distribution $P_\theta = f_\theta \cdot \mu$ from a canonical and minimal exponential model. The log-likelihood writes

$$\frac{1}{n}\ell_n(\theta) = \theta^\top \overline{T}_n - \log Z(\theta) \tag{6.2}$$

where we introduced $\overline{T}_n = n^{-1}\sum_{i=1}^n T(X_i)$.

Proposition 6.7 In a canonical and minimal exponential model, the log-likelihood is strictly concave. This entails that if the MLE $\widehat{\theta}_n = \operatorname{argmax}_{\theta\in\Theta} \ell_n(\theta)$ exists, it is given by

$$\widehat{\theta}_n = S^{-1}(\overline{T}_n),$$

where we recall that $S(\theta) = \nabla \log Z(\theta)$.

Proof. This is obvious since $\theta \mapsto \theta^\top \overline{T}_n$ is linear, hence concave, and since $\theta \mapsto \log Z(\theta)$ is strictly concave using Theorem 6.3. Hence, any maximizer of ℓ_n must satisfy the first order condition $\nabla \ell_n(\theta) = 0$ which means that $S(\theta) = \nabla \log Z(\theta) = \overline{T}_n$. However, we know using Theorem 6.6 that if it exists, the only solution is $\widehat{\theta}_n = S^{-1}(\overline{T}_n)$. □

Proposition 6.7 proves that, when it exists, the MLE corresponds to the so-called *method of moments estimator*.

Example 6.2 Consider a iid sample $X_1, \dots, X_n$ with $\mathrm{Gamma}(a, \lambda)$ distribution and recall that its density can be written as

$$f_{a,\lambda}(x) = \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x} = \exp(\theta^\top T(x) - \log Z(\theta))$$

for $x \geq 0$, where $\theta = [a-1 \;\; \lambda]^\top$, $T(x) = [\log x \;\; -x]^\top$ and $Z(\theta) = \Gamma(a)/\lambda^a$. We have

$$\nabla \log Z(\theta) = \begin{bmatrix} \frac{\Gamma'(a)}{\Gamma(a)} - \log \lambda \\ -a/\lambda \end{bmatrix}$$

so that finding a closed-form estimator means finding a solution to

$$\frac{\Gamma'(a)}{\Gamma(a)} - \log \lambda = \frac{1}{n}\sum_{i=1}^n \log X_i$$
$$\frac{a}{\lambda} = \frac{1}{n}\sum_{i=1}^n X_i.$$

Such a solution is not explicit, but can be easily approximated using a convex optimization algorithm.

In summary, what we learned so far about the MLE in a minimal and canonical exponential model is the following:

- When the MLE exists, we can express it as $\widehat{\theta}_n = S^{-1}(\overline{T}_n)$, although in general we cannot inverse explicitly S, as observed in the previous example;
- When the MLE exists, we can compute it approximately using a convex optimization algorithm, since the log-likelihood $\ell_n(\theta)$ is strictly concave and smooth.

Note that the MLE does not always exist, as shown in the next simple example.

Remark 6.1 Consider $X_1, \dots, X_n$ iid with geometric distribution, namely a density $f_p(x) = (1-p)^{x-1}p$ with respect to the counting measure, for $x \in \mathbb{N} \setminus \{0\}$ and $p \in (0,1)$. We can write it as an exponential model since

$$f_p(x) = \exp\big((x-1)\log(1-p) + \log p\big),$$

so that the sufficient statistic is $T(x) = 1 - x$ and the canonical parameter is $\theta = -\log(1-p)$ namely $p = 1 - e^{-\theta}$, so that in canonical form, this exponential model writes

$$f_\theta(x) = \exp\big((1-x)\theta + \log(1-e^{-\theta})\big),$$

and the log-likelihood is given by

$$\ell_n(\theta) = \Big(n - \sum_{i=1}^n X_i\Big)\theta + n\log(1-e^{-\theta}).$$

On the event $E = \{X_1 = \cdots = X_n = 1\}$ which has a probability $\mathbb{P}_\theta[E] = (1-e^{-\theta})^n > 0$ (although going towards zero quickly), the MLE does not exist since on E we have $\ell_n(\theta) = n\log(1-e^{-\theta})$ which is concave and strictly increasing.

Definition 6.6 (Score and Fisher information) In a minimal and canonical exponential model, the function $\theta \mapsto \nabla\ell_n(\theta)$ is called the *score* function and the matrix

$$I_n(\theta) = \mathbb{V}_\theta[\nabla\ell_n(\theta)],$$

which is the covariance matrix of the score, is called the *Fisher information*.

This definition goes way beyond the particular case of the exponential models considered here. Let us give some properties of the score and the Fisher information. First of all, let us remark that the score is a

centered random vector, since

$$\mathbb{E}_\theta[\nabla \ell_n(\theta)] = n(\mathbb{E}_\theta[T(X_1)] - \nabla \log Z(\theta)) = 0$$ Using Theorem 6.3

and let us note that the Fisher information satisfies

$$I_n(\theta) = \mathbb{V}_\theta[\nabla \ell_n(\theta)] = n\, \mathbb{V}_\theta[T(X_1)] = nI_1(\theta)$$ Using (6.2) and the definition of Fisher information with $n = 1$

and also that it satisfies

$$I_n(\theta) = n\, \mathbb{V}_\theta[T(X_1)] = n\nabla^2 \log Z(\theta).$$ Using Theorem 6.3

Note that since $T(X_1), \dots, T(X_n)$ are iid and such that $\mathbb{E}_\theta[T(X_1)] = \nabla \log Z(\theta) = S(\theta)$ and $\mathbb{V}_\theta[T(X_1)] = I_1(\theta)$, we can use the multivariate central limit theorem to obtain

$$\sqrt{n}(\overline{T}_n - S(\theta)) \rightsquigarrow \text{Normal}\big(0, I_1(\theta)\big),$$ This a convergence in P_θ distribution

but since $\widehat{\theta}_n = S^{-1}(\overline{T}_n)$, we want to use the multivariate Δ-method (the scalar case was covered in Theorem 2.3 from Chapter 2) with $\varphi(t) = S^{-1}(t)$. The multivariate Δ-method is given by the following theorem.

Theorem 6.8 (Multivariate Δ-method) Let $(a_n)_{n\geq 1}$ be a sequence of positive number such that $a_n \to +\infty$, $(X_n)_{n\geq 1}$ be a sequence of random vectors $X_n \in \mathbb{R}^d$ and φ be measurable function. If $a_n(X_n - x) \rightsquigarrow X$ for some $x \in \mathbb{R}^d$ and some random vector X, and if φ is differentiable at x, we have

$$a_n(\varphi(X_n) - \varphi(x)) \rightsquigarrow J_\varphi(x)X,$$

where $J_\varphi(x)$ is the Jacobian matrix of f at x.

The proof of Theorem 6.8 is omitted since it follows the exact same proof as that of Theorem 2.3. We apply[1] Theorem 6.8 with $\varphi = S^{-1}$, so that $J_\varphi(S(\theta)) = (\nabla^2 \log Z(\theta))^{-1} = I_1(\theta)^{-1}$ and we end up with

1: Thanks to Theorem 6.6

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow \text{Normal}\big(0, I_1(\theta)^{-1}\big)$$ This a convergence in P_θ-distribution

since $\mathbb{V}[I_1(\theta)^{-1}Z] = I_1(\theta)^{-1}\mathbb{V}[Z]I_1(\theta)^{-1} = I_1(\theta)^{-1}$ whenever $Z \sim \text{Normal}(0, I_1(\theta))$. This proves the following theorem.

Theorem 6.9 (Central limit theorem for the MLE) In a statistical experiment where the model is a minimal and canonical exponential model and where the MLE $\widehat{\theta}_n$ exists, we have

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow \text{Normal}\big(0, I_1(\theta)^{-1}\big),$$

which is a convergence in P_θ-distribution, where $I_1(\theta)$ is the Fisher information matrix given in Definition 6.6.

This theorem proves that the MLE is asymptotically normal and that its "asymptotic variance" is equal to the inverse of the Fisher information. In this sense, the Fisher information quantifies the asymptotic performance of the MLE. Moreover, we can prove that the inverse of the Fisher information matrix is the *smallest achievable* asymptotic variance among all asymptotically normal estimators[2] and that the MLE is *efficient*, since it is asymptotically normal with minimal asymptotic variance (given by the inverse Fisher information matrix).

2: An estimator $\widetilde{\theta}_n$ is asymptotically normal if it satisfies $\sqrt{n}(\widetilde{\theta}_n - \theta) \rightsquigarrow \mathrm{Normal}(0, \boldsymbol{V})$ for all $\theta \in \Theta$ with a non-degenerate matrix $\boldsymbol{V}$.

However, we won't go further in this direction, since such results are somewhat "stylized": they hold only for an arbitrarily large n and only for *well-specified* models. Indeed, such asymptotic results hold only when the model is *well-specified*, namely whenever the *true distribution* P_X of the data actually belongs to the model, namely $P_X = P_{\theta^\star}$ for some $\theta^\star \in \Theta$. If the model is miss-specified, namely $P_X \notin \{P_\theta : \theta \in \Theta\}$, then the MLE quickly deteriorates.

Bibliography

[1] Gregory F Lawler and Vlada Limic. *Random walk: a modern introduction*. Vol. 123. Cambridge University Press, 2010 (cited on page 2).

[2] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008 (cited on page 5).

[3] Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006 (cited on page 5).

[4] Kevin P Murphy. *Machine Learning, A Probabilistic Perspective*. MIT Press, 2012 (cited on page 7).

[5] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016 (cited on page 7).

[6] Irina Shevtsova. 'On the absolute constants in the Berry-Esseen type inequalities for identically distributed summands'. In: *arXiv preprint arXiv:1111.6554* (2011) (cited on page 16).

[7] Carl-Gustav Esseen. 'A moment inequality with an application to the central limit theorem'. In: *Scandinavian Actuarial Journal* 1956.2 (1956), pp. 160–170 (cited on page 16).

[8] Terrence Tao. *254A, Notes 2: The central limit theorem*. https://terrytao.wordpress.com/2010/01/05/254a-notes-2-the-central-limit-theorem/. 2010 (cited on page 16).

[9] Frank den Hollander. 'Probability theory: The coupling method'. In: *Leiden University, Lectures Notes-Mathematical Institute* (2012), p. 31 (cited on page 21).

[10] Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar. 'Moving to a World Beyond p < 0.05'. In: *The American Statistician* 73.sup1 (2019), pp. 1–19 (cited on page 25).

[11] Pascal Massart. *Concentration inequalities and model selection*. Vol. 6. Springer, 2007 (cited on page 25).

[12] Jaouad Mourtada. *Contributions to statistical learning: density estimation, expert aggregation and random forests*. 2019 (cited on page 44).

[13] Jaouad Mourtada. 'Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices'. In: *arXiv preprint* (2020) (cited on pages 44, 49).

[14] Charles Stein. *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution*. Tech. rep. Stanford University Stanford United States, 1956 (cited on page 56).

[15] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006 (cited on page 56).

[16] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006 (cited on page 60).

[17] Wikipedia. *Rule of succession*. https://en.wikipedia.org/wiki/Rule_of_succession/. 2020 (cited on page 63).

[18] Robert Tibshirani. 'Regression shrinkage and selection via the lasso'. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288 (cited on page 75).

[19] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. 'Simultaneous analysis of Lasso and Dantzig selector'. In: *Ann. Statist.* 37.4 (Aug. 2009), pp. 1705–1732. DOI: 10.1214/08-AOS620 (cited on page 79).

[20] Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. 'Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion'. In: *Ann. Statist.* 39.5 (Oct. 2011), pp. 2302–2329. DOI: 10.1214/11-AOS894 (cited on page 79).

[21] Christophe Giraud. *Introduction to high-dimensional statistics*. Vol. 138. CRC Press, 2014 (cited on page 79).

[22] Nicolas Verzelen. 'Minimax risks for sparse regressions: Ultra-high dimensional phenomenons'. In: *Electron. J. Statist.* 6 (2012), pp. 38–90. DOI: 10.1214/12-EJS666 (cited on page 82).

[23] Peng Zhao and Bin Yu. 'On model selection consistency of Lasso'. In: *Journal of Machine learning research* 7.Nov (2006), pp. 2541–2563 (cited on page 82).