Lecture notes on the bootstrap and resampling methods

François Portier

January 5, 2021

Contents

Chapter 1

Efron's bootstrap

1.1 Mathematical background

We recall useful results about the convergence of sequences of random variables. For a complete presentation of those concepts, we refer to Van der Vaart (2000).

Definition 1.1. *Let X be a random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$ valued in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The cumulative distribution function of X, $F_X : \mathbb{R}^d \to [0,1]$, is defined by*

$$F_X(x) = \mathbb{P}(X \leq x),$$

where $\{X \leq x\} = \{X_1 \leq x_1, \ldots, X_d \leq x_d\}$.

The norms $|\cdot|_p$, $p \geq 1$, are defined as follows, for any $u \in \mathbb{R}^d$,

$$|u|_p^p = \sum_{k=1}^{d} u_k^p.$$

Definition 1.2. *Let $X, (X_n)_{n\geq 1}$ be a sequence of random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ valued in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.*

- *$(X_n)_{n\geq 1}$ converges to X, almost surely, if with probability 1, $|X_n - X|_1 \to 0$ as $n \to \infty$.*
- *$(X_n)_{n\geq 1}$ converges to X, in probability, if for all $\epsilon > 0$, $\mathbb{P}(|X_n - X|_1 > \epsilon) \to 0$.*
- *$(X_n)_{n\geq 1}$ converges to X, in L_p, if $\mathbb{E}[|X_n - X|_p^p] \to 0$.*
- *$(X_n)_{n\geq 1}$ converges to X, in distribution, if $F_{X_n}(x) \to F_X(x)$ for any continuity point of F_X. We also say that X_n weakly converges to X and we write $X_n \rightsquigarrow X$.*

Let $C_b(\mathbb{R}^d)$ (resp. $L_b(\mathbb{R}^d)$) be the space of bounded continuous (resp. bounded Lipschitz) real valued functions defined on $\mathbb{R}^d$.

Proposition 1.1. *Let $X, (X_n)_{n\geq 1}$ be a sequence of random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ valued in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The following are equivalent*

- *$(X_n)_{n\geq 1}$ converges to X in distribution*
- *for all $f \in C_b(\mathbb{R}^d)$, $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$*
- *for all $f \in L_b(\mathbb{R}^d)$, $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$*

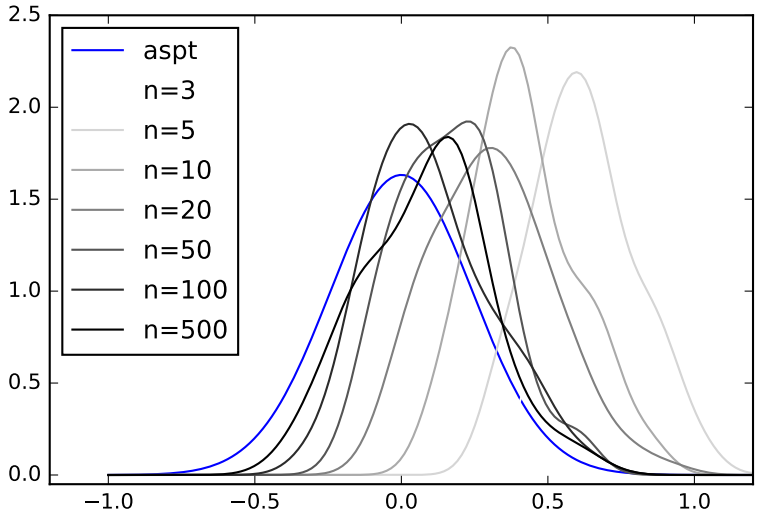

Figure 1.1: Illustration of the central limit theorem. It shows the convergence of the sequence of densities (of empirical means) toward the Gaussian density.

Proposition 1.2 (continuous mapping theorem)**.** *Let $X, (X_n, Y_n)_{n\geq 1}$ be a sequence of random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ valued in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Let $f : \mathbb{R}^d \to \mathbb{R}^q$ be a Borelian function and denote by $C_f \subset \mathbb{R}^d$ the set of continuity points of f. Suppose that that $\mathbb{P}(X \in C_f) = 1$. The following holds:*

- *if $X_n \to X$ almost surely, then $f(X_n) \to f(X)$ almost surely.*
- *if $X_n \to X$ in probability, then $f(X_n) \to f(X)$ in probability.*
- *if $X_n \rightsquigarrow X$, then $f(X_n) \rightsquigarrow f(X)$.*

Lemma 1.3 (Slutsky)**.** *Let $X, (X_n, Y_n)_{n\geq 1}$ be a sequence of random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ valued in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Suppose that $X_n \rightsquigarrow X$ and that $Y_n \to c$ in probability. Then $(X_n, Y_n) \rightsquigarrow (X, c)$.*

Theorem 1.4 (strong law of large numbers)**.** *Let $(X_n)_{n\geq 1}$ be a independent and identically distributed sequence of random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ valued in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ with common probability measure P on $\mathcal{B}(\mathbb{R}^d)$. Suppose that $E|X_1|_1 < \infty$. Then we have*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to E(X_1), \qquad \text{almost surely.}$$

Theorem 1.5 (central limit theorem)**.** *Let $(X_n)_{n\geq 1}$ be a independent and identically distributed sequence of random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ valued in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ with common probability measure P on $\mathcal{B}(\mathbb{R}^d)$. Suppose that $E[|X_1|_2^2] < \infty$ and define $\Sigma = \mathrm{var}(X_1)$. Then we have*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - E(X_1)) \rightsquigarrow \mathcal{N}(0, \Sigma).$$

The next theorem generalizes the original central limit theorem in two directions. First, one is authorized to consider non-identically distributed sequences. Second, each sequence might change with n, the sample size.

Theorem 1.6 (Lindeberg central limit theorem)**.** *For each $n \geq 1$, let $X_{n,1}, X_{n,2}, \dots X_{n,n}$ be a sequence of independent random vectors such that $\mathbb{E}[|X_{n,i}|_2^2] < \infty$ for all $i \geq 1$ and*

$$n^{-1} \sum_{i=1}^{n} \mathbb{E}[|X_{n,i}|^2 \mathbb{1}_{|X_{n,i}|>n^{1/2}\epsilon}] \longrightarrow 0, \quad \textit{for all } \epsilon > 0 \ ,$$
$$n^{-1} \sum_{i=1}^{n} \mathrm{var}(X_{n,i}) \longrightarrow \Sigma \ .$$

Then,

$$n^{-1/2} \sum_{i=1}^{n} (X_{n,i} - \mathbb{E}[X_{n,i}]) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma) \ .$$

1.2 The imitation principle

The bootstrap was initially introduced in the statistical estimation framework where the parameter of interest, say θ_0, depends on an unknown probability distribution. The main goal of the bootstrap is to *measure the accuracy* of an estimate, say $\hat{\theta}_n$, of θ_0 by *reproducing/approximating* the distribution of $(\hat{\theta}_n - \theta_0)$.

1.2.1 The framework

The original bootstrap applies to estimators based on independent and identically distributed sequences of random variables. Let $(X_i)_{i\in\mathbb{N}}$ be an independent and identically distributed sequence of $\mathbb{R}^d$-valued random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The distribution of X_1 on $\mathcal{B}(\mathbb{R}^d)$, the Borel algebra, is denoted by P. Suppose that the parameter of interest θ_0 expresses as

$$\theta_0 = \theta(P),$$

where θ is defined on the space of probability measures and valued in $\mathbb{R}^p$. It could be for instance the theoretical mean of an unknown distribution or the regression vector in a linear regression model. Define the empirical measure

$$P_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i},$$

where for any $x \in \mathbb{R}^d$, δ_x is a probability measure defined on $\mathcal{B}(\mathbb{R}^d)$, called the Dirac measure, and defined by $\delta_x(A) = 1$ if $x \in A$ and $\delta_x(A) = 0$ else for any Borelian set A. The empirical measure $P_n = P_n(w, \cdot)$ is a random (probability) measure, i.e., for any $w \in \Omega$, $P_n(w, \cdot)$ is a (probability) measure and for any $B \in \mathcal{B}(\mathbb{R}^d)$, $w \mapsto P_n(w, B)$ is a real valued random variable (from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. In particular, one might verify that $P_n(w, \cdot)$ is the uniform probability measure on $\{X_1(w), \dots, X_n(w)\}$. In the following development, we shall use the shortcut P_n for $P_n(w, \cdot)$ as soon as possible. The empirical measure is an approximation of the true measure P as illustrated by the following basic property, which can be easily deduced from the strong law of large number. For any Borelian set B, with probability one,

$$|P_n(B) - P(B)| \to 0, \qquad \text{as } n \to \infty.$$

From the previous remark, a natural estimator of θ_0 is given by

$$\hat{\theta}_n = \theta(P_n).$$

This kind of estimators are usually called "plug-in" estimators as the estimated probability measure, the empirical measure, has been "plugged" in the expression in place of the theoretical distribution. In the following examples, we show that this "plug-in" approach permits to recover many classic estimators including the empirical mean, the empirical variance, and the ordinary least-squares estimate.

Example 1.1. *The expectation of g with respect to P is given by $\theta(P) = \int g(x)\,\mathrm{d}P(x)$. The mean corresponds to $g(x) = x$.*

Example 1.2. *The variance corresponds to*

$$\theta(P) = \int x^2\,\mathrm{d}P(x) - \left(\int x\,\mathrm{d}P(x)\right)^2.$$

Example 1.3. *Let $X = (Z_1, Z_2) \in \mathbb{R}^2$. The covariance between Z_1 and Z_2 is given by*

$$\theta(P) = \int z_1 z_2\,\mathrm{d}P(z_1, z_2) - \left(\int z_1\,\mathrm{d}P_1(z_1)\right)\left(\int z_2\,\mathrm{d}P_2(z_2)\right),$$

where P_1 and P_2 stands for the marginals of Z_1 and Z_2.

Example 1.4. *The correlation coefficient between Z_1 and Z_2 is given by*

$$\theta(P) = \frac{\int z_1 z_2\,\mathrm{d}P(z_1, z_2) - \left(\int z_1\,\mathrm{d}P_1(z_1)\right)\left(\int z_2\,\mathrm{d}P_2(z_2)\right)}{\sqrt{\left(\int z_1^2\,\mathrm{d}P_1(z_1) - \left(\int z_1\,\mathrm{d}P_1(z_1)\right)^2\right)\left(\int z_2^2\,\mathrm{d}P_2(z_2) - \left(\int z_2\,\mathrm{d}P_2(z_2)\right)^2\right)}}.$$

Example 1.5. *Suppose that $X = (Z_1, Z_2)$ with $Z_1 \in \mathbb{R}$, $Z_2 \in \mathbb{R}^p$, $p \geq 1$, and that $E\|X\|^2 < \infty$. The regression coefficient $\beta \in \mathbb{R}^p$ defined as*

$$\mathrm{argmin}_{\beta \in \mathbb{R}^p} \int (z_1 - \beta^T z_2)^2\,\mathrm{d}P(z_1, z_2),$$

equivalently, $\theta(P)$ is the solution of

$$\left(\int z_2 z_2^T \mathrm{d}P_2(z_2)\right)\beta = \int z_1 z_2 \mathrm{d}P(z_1, z_2).$$

Example 1.6. *The distribution function F evaluated at y is given by*

$$\theta(P) = \int \mathbb{1}_{\{x \leq y\}} \mathrm{d}P(x).$$

Example 1.7. *The median is given by*

$$\theta(P) = F^-(1/2).$$

where F^- is the generalized inverse of F.

Real world		**Bootstrap world**
$X_1, \dots X_n \sim P$ (**unknown**)		$X^*_{n,1}, \dots X^*_{n,n} \sim P_n$ (**known**)
$\downarrow$		$\downarrow$
Compute $\theta(P_n)$		Compute $\theta(P^*_n)$
$\downarrow$		$\downarrow$
$(\theta(P_n) - \theta(P))$	"$\simeq$"	$(\theta(P^*_n) - \theta(P_n))$

Table 1.1: The bootstrap purpose: to mimic what happen in the real world by a new data generating process

1.2.2 The imitation principle

The bootstrap is based on a simple imitation principle. It mimics the behavior of the original estimate by generating a new sample, called the *bootstrap sample*, according to the empirical measure P_n. With this new sample, the same operations as the one to compute $\hat{\theta}_n$ are carried out to obtain a bootstrap version of $\hat{\theta}_n$. As described before, we consider plug-in estimators $\hat{\theta}_n = \theta(P_n)$ of $\theta_0 = \theta(P)$, where P_n is the empirical measure based on $X_1, \dots, X_n$ with common distribution P. To reproduce this situation, the bootstrap method relies on the generation of a new sample $X^*_{n,1}, \dots, X^*_{n,n}$ according to P_n and then applies the transformation θ to these samples. The bootstrap estimate is then defined by

$$\hat{\theta}^*_n = \theta(P^*_n),$$

where P^*_n is the empirical measure based on the bootstrap sample, i.e.,

$$P^*_n = n^{-1} \sum_{i=1}^{n} \delta_{X^*_{n,i}}.$$

As illustrated in Table 1.1, due to the similarity between P and P_n, we expect that $(\hat{\theta}^*_n - \hat{\theta}_n)$ "behaves in a similar" way as $(\hat{\theta}_n - \theta_0)$.

Given $X_1, \dots, X_n$, P_n is a discrete probability distribution with n atoms. Hence, there are n^n possible bootstrap sample which implies that $(\hat{\theta}^*_n - \hat{\theta}_n)$ is a discrete random variable with n^n state space. In practice, n^n is often too large and computing moments of this distribution is too heavy computationally. In place we shall use Monte Carlo approximation based on generating independent versions of $(\hat{\theta}^*_n - \hat{\theta}_n)$, conditionally on $X_1, \dots, X_n$. Each version is simply computed by generating a new bootstrap samples $X^*_{n,1}, \dots, X^*_{n,n}$. Accordingly, the bootstrap method is made of two steps:

(i) (**definition step**) The bootstrap $(\hat{\theta}^*_n - \hat{\theta}_n)$ must mimic the behaviour of the quantity of interest $(\hat{\theta}_n - \theta_0)$.

(ii) (**simulation step**) For some B, compute $(\hat{\theta}^*_{n,1} - \hat{\theta}_n), \dots, (\hat{\theta}^*_{n,B} - \hat{\theta}_n)$ to approximate the law of $(\hat{\theta}_n - \theta_0)$.

Until now, we left unspecified the meaning of the words "behaves similarly" and "mimics". We shall see in section 1.3 that in general, the bootstrap estimate, properly rescaled, has a similar distribution as the rescaled estimate of interest.

1.2.3 The bootstrap algorithm

The following algorithm is the original bootstrap algorithm as introduced by Efron (1979). To make clear that it corresponds to the plug-in principle described previously, one should

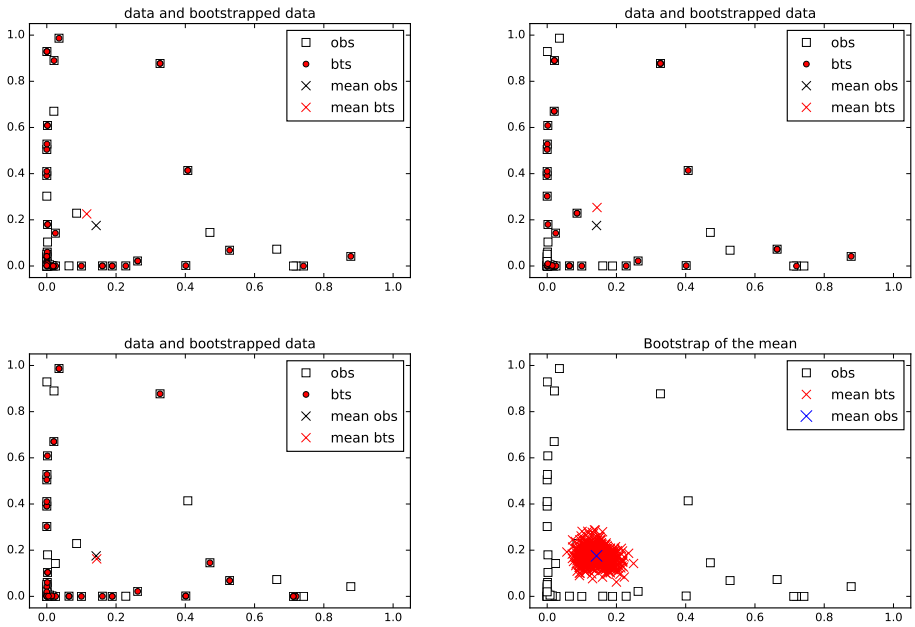

Figure 1.2: The first three graphs provide examples of different bootstrap samples. The last graph shows the true mean value and the different mean values obtained using bootstrap samples.

be aware that when X^* is generated from a uniform draw among $\{X_1, \dots X_n\}$, it has law P_n.

Algorithm.

Input : The observations $X_1, \dots X_n$, the simulation number for the bootstrap B
Output: Bootstrap estimators $(\hat{\theta}^*_{n,1}, \dots, \hat{\theta}^*_{n,B})$
for $b = 1, \dots, B$ **do**
Draw uniformly with replacement among $\{X_1, \dots, X_n\}$ to obtain the bootstrap sample

$$X^*_{n,1}, \dots, X^*_{n,n}$$

Compute the bootstrap estimator

$$\hat{\theta}^*_{n,b} = \theta(P^*_n)$$

end

The bootstrap estimators $\hat{\theta}^*_{n,b}$, $b = 1, \dots B$ are identically distributed and independent random variables. Taking B large ensures that the law of $\hat{\theta}^*_n$ is well approximated. In the sequel, we suppose that B is large enough so that the simulation error is negligible compared to other approximation.

1.3 Bootstrap approximation

1.3.1 The randomness of bootstrap sequences

Let $X^*_{n,1}, \ldots, X^*_{n,n}$ be identically distributed and independent random variables with common distribution P_n, conditionally to $X_1, \ldots, X_n$ ($X^*_{n,1}, \ldots X^*_{n,n}$ is called the bootstrap sample). By definition, for any collection of positive functions f_i,

$$E[\prod_{i=1}^{n} f_i(X^*_{n,i})|X_1, \ldots, X_n] = \prod_{i=1}^{n} E[f_i(X^*_{n,i})|X_1, \ldots, X_n],$$

and for any i and positive function f

$$E[f(X^*_{n,i})|X_1, \ldots, X_n] = \int f(x)\, \mathrm{d}P_n(x),$$

There is two sources of randomness that must be considered when analyzing bootstrap estimates:

- The randomness of the initial samples $(X_1, \ldots, X_n)$
- The randomness induced by the generation of the bootstrap sample

The sequence of estimators $(\hat{\theta}_n)_{n\geq 1}$ is a random sequence defined on $(\Omega, \mathcal{F}, \mathbb{P})$. The sequence of bootstrap estimators $(\hat{\theta}^*_n)_{n\geq 1}$ depends also on the original $(X_i)_{i\geq 1}$ but another source of randomness has been introduced through the sampling. Let $\{(u^*_{n,i})_{i=1,\ldots,n}, n \geq 1\}$ be an independent triangular array such that each $n \geq 1$, $(u^*_{n,i})_{i=1,\ldots,n}$ is an independent collection of random variable with common distribution $\sum_{i=1}^{n} \delta_i/n$. Denote by $(\Omega^*, \mathcal{F}^*, \mathbb{P}^*)$ the underlying probability space. Hence the sequence of bootstrap estimators $(\hat{\theta}^*_n)$ is defined on the product space $(\Omega \times \Omega^*, \mathcal{F} \otimes \mathcal{F}^*, \mathbb{P} \times \mathbb{P}^*)$. The underlying measure is the product measure because the bootstrap resampling is done independently of the original sequence $(X_i)_{i\geq 1}$.

1.3.2 Convergence of bootstrap estimates

Definitions

We now introduce different notions of convergence for bootstrap estimators. These are basically the same as the one defined for standard estimator except that the convergence is considered with respect to the measure $\mathbb{P}^*$ at fixed $\omega \in \Omega$. A key feature to understand the following definitions is that after integrating a bootstrap estimate with respect to $\mathbb{P}^*$, we obtain a random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Let (Y^*_n) be a sequence defined on $(\Omega \times \Omega^*, \mathcal{F} \otimes \mathcal{F}^*, \mathbb{P} \times \mathbb{P}^*)$. Suppose in addition that for all $\omega \in \Omega$, $Y^*_n(\omega, \cdot)$ is a random variable defined on $(\Omega^*, \mathcal{F}^*, \mathbb{P}^*)$.

Definition 1.3. *We say that (Y^*_n) converges in distribution to F, conditionally on the observations, almost surely, and write $Y^*_n \rightsquigarrow F$, a.s., if with probability one : for every y where F is continuous, $\mathbb{P}^*(Y^*_n \leq y) \to F(y)$ as $n \to \infty$.*

Definition 1.4. *We say that $Y^*_n \xrightarrow{\mathbb{P}^*} 0$ a.s., if with probability 1, for every $\epsilon > 0$, $\mathbb{P}^*(|Y^*_n| > \epsilon) \xrightarrow{a.s.} 0$.*

Note that $Y^*_n \xrightarrow{\mathbb{P}^*} 0$ a.s., if and only if $Y^*_n \rightsquigarrow 0$ a.s. The subsequent propositions might be useful.

Proposition 1.7. *$Y_n^* \xrightarrow{\mathbb{P}^*} 0$ a.s., if and only if for all $k \in \mathbb{N}^*$,*

$$\limsup_n \mathbb{P}^*(|Y_n^*| > 1/k) = 0, \qquad a.s.$$

.

Proposition 1.8. *If $Y_n^* \to 0$ almost-surely, then $Y_n^* \xrightarrow{\mathbb{P}^*} 0$ almost-surely.*

Proof. This is a consequence of the fact that $Y_n^* \to 0$ a.s. implies that for all $\epsilon > 0$, $\mathbb{I}\{Y_n^* > \epsilon\} \to 0$ a.s. □

Preservation properties

We are interested in deriving the continuous mapping theorem and the Slutsky's Lemma for the bootstrap. In the following, we focus on the "almost sure" notion of weak consistency. Nevertheless, the same results hold also with the "in probability statement" by making use of the characterization of the convergence in probability with subsequence : a sequence converges in probability if and only if from any sub-sequence on can extract a further subsequence that converges almost surely. The two following preservation lemmas are derived by applying the (non-bootstrap) continuous mapping theorem and the Slutsky's lemma.

In what follows, X^*, $(X_n^*)_{n\geq 1}$ is a collection of random variables defined on $(\Omega \times \Omega^*, \mathcal{F} \otimes \mathcal{F}^*, \mathbb{P} \times \mathbb{P}^*)$.

Proposition 1.9 (continuous mapping theorem for the bootstrap)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}^q$ be Borelian function and denote by C_f the set of its continuity points. Suppose that $X^* \in C_f$ almost surely.*

(i) If $X_n^* \rightsquigarrow X^*$, almost surely, then $f(X_n^*) \rightsquigarrow f(X^*)$, almost surely.

(ii) If $X_n^* \xrightarrow{\mathbb{P}^*} X^*$, almost surely, then $f(X_n^*) \xrightarrow{\mathbb{P}^*} f(X^*)$, almost surely.

Proof. By assumption there exists $A \in \mathcal{F}$ such that $\mathbb{P}(A) = 1$ and for all $\omega \in A$, $X_n^*(\omega, \cdot) \rightsquigarrow X^*(\omega, \cdot)$ and $\mathbb{P}^*(X^*(\omega, \omega^*) \in C_f) = 1$. The continuous mapping theorem permits to conclude that for all $\omega \in A$, $f(X_n^*(\omega, \cdot) \rightsquigarrow f(X^*(\omega, \cdot))$ which is the conclusion of (i). The other statement can be proved similarly. □

Proposition 1.10 (Slutsky for the bootstrap)**.**

- *If $X_n^* \rightsquigarrow X$, a.s., and $Y_n^* \xrightarrow{\mathbb{P}^*} 0 \in \mathbb{R}$, a.s., then $X_n^* + Y_n^* \rightsquigarrow X$, a.s..*
- *If $X_n^* \rightsquigarrow X$, a.s., and $Y_n^* \xrightarrow{\mathbb{P}^*} c \in \mathbb{R}$, a.s., then $(X_n^*, Y_n^*) \rightsquigarrow (X, c)$, a.s..*

Proof. While the proof is a direct consequence of Slutsky's Lemma (stated but not proved in Lemma 1.3) we provide a proof here. We start with the first statement. Let μ be the cdf of X and D be the discontinuity point of μ. Because D is countable (associate each $x \in D$ to q_x a rational between $F(x-)$ and $F(x+)$), $\mathbb{R}\backslash D$ is dense. Note that, for any x, $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}^*(X_n^* + Y_n^* \leq x) &\leq \mathbb{P}^*(X_n^* + Y_n^* \leq x, |Y_n^*| \leq \epsilon) + \mathbb{P}^*(|Y_n^*| > \epsilon) \\
&\leq \mathbb{P}^*(X_n^* \leq x + \epsilon, |Y_n^*| \leq \epsilon) + \mathbb{P}^*(|Y_n^*| > \epsilon) \\
&\leq \mathbb{P}^*(X_n^* \leq x + \epsilon) + \mathbb{P}^*(|Y_n^*| > \epsilon).
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{P}^*(X_n^* + Y_n^* > x) &\leq \mathbb{P}^*(X_n^* + Y_n^* > x, |Y_n^*| \leq \epsilon) + \mathbb{P}^*(|Y_n^*| > \epsilon) \\
&\leq \mathbb{P}^*(X_n^* > x - \epsilon, |Y_n^*| \leq \epsilon) + \mathbb{P}^*(|Y_n^*| > \epsilon) \\
&\leq \mathbb{P}^*(X_n^* > x - \epsilon) + \mathbb{P}^*(|Y_n^*| > \epsilon).
\end{aligned}$$

It follows that

$$\mathbb{P}^*(X_n^* \leq x - \epsilon) - \mathbb{P}^*(|Y_n^*| > \epsilon) \leq \mathbb{P}^*(X_n^* + Y_n^* \leq x) \leq \mathbb{P}^*(X_n^* \leq x + \epsilon) + \mathbb{P}^*(|Y_n^*| > \epsilon).$$

With probability 1, it holds that, for any $x \notin D$, $x - \epsilon \notin D$ and $x + \epsilon \notin D$,

$$\mu(x - \epsilon) \leq \liminf_{n\to\infty} \mathbb{P}^*(X_n^* + Y_n^* \leq x) \leq \limsup_{n\to\infty} \mathbb{P}^*(X_n^* + Y_n^* \leq x) \leq \mu(x + \epsilon).$$

Since ϵ is arbitrarily small (by density of $\mathbb{R}\backslash D$), we have for any $x \notin D$,

$$\lim_{n\to\infty} \mathbb{P}^*(X_n^* + Y_n^* \leq x) = \mu(x).$$

To prove the second statement, write $(X_n, Y_n) = (0, Y_n - c) + (X_n, c)$ and invoke the first statement to see that the proof reduces to $(X_n, c) \rightsquigarrow (X, c)$, almost surely. For all f continuous note that $x \mapsto f(x, c)$ is continuous as well. By Proposition 1.1, we have, with probability 1, that for all $f \in C_b$, $E^* f(X_n, c) \to E^* f(X, c)$. □

1.3.3 Consistency of the bootstrap

While it makes no doubt that to assess the accuracy of $\hat{\theta}_n$ a central quantity is the statistic $(\hat{\theta}_n - \theta_0)$, we have not yet specified what type of approximation of the statistic is feasible. A pertinent notion is the one of weak convergence (it is helpful to build confidence interval). Consequently, we shall be concerned with regular estimator as defined below.

Definition 1.5. *The map θ is said to be regular with respect to P if $n^{1/2}(\theta(P_n) - \theta(P))$ converges weakly to a certain distribution denoted $L(\theta, P)$.*

The distribution $L(\theta, P)$ is very important in the subsequent development as it is especially this asymptotic distribution that the bootstrap will be able to approximate. In this way, previous notion will be useful to discuss the validity or the failure of the bootstrap.

Definition 1.6. *Given a distribution P and a regular map θ, the bootstrap is said to be consistent if $n^{1/2}(\theta(P_n^*) - \theta(P_n))$ converges weakly, a.s., to $L(\theta, P)$.*

The next proposition shows that the bootstrap is consistent for the empirical mean $n^{-1}\sum_{i=1}^n X_i$.

Theorem 1.11 (Bootstrap central limit theorem)**.** *If $\theta(P) = \int xP(dx)$ then for all distribution P such that $\int x^2 P(dx) < \infty$, the bootstrap is consistent.*

Proof. Let $X_1, X_2, \ldots$ be an independent and identically distributed sequence of random variables defined on $(\Omega, \mathcal{F}, \mathcal{P})$. Verify that

$$\mathbb{E}^* X_{n,i}^* = \int x \, \mathrm{d}P_n(x) = \hat{\theta}_n,$$

so that we can write

$$n^{1/2}(\theta(P_n^*) - \theta(P_n)) = n^{-1/2} \sum_{i=1}^n (X_{n,i}^* - \mathbb{E}^* X_{n,i}^*)$$

Consequently, we are in position to apply the Lindeberg central limit theorem, which is stated in Section 1.1 as Theorem 1.6. More specifically, we will apply this theorem in

an event of probability 1 on which the suitable convergences hold true. The required converges are now given. If we have, with probability 1,

$$\forall \epsilon > 0, \qquad n^{-1} \sum_{i=1}^{n} \mathbb{E}^* \|X_{n,i}^*\|^2 \mathbb{1}_{\{\|X_{n,i}^*\| > \epsilon n^{1/2}\}} \longrightarrow 0, \tag{1.1}$$

$$n^{-1} \sum_{i=1}^{n} \mathbb{E}^* \left(X_{n,i}^* - \mathbb{E}^* X_{n,i}^*\right)^2 \longrightarrow \Sigma, \tag{1.2}$$

then, from the Lindeberg central limit theorem, we have with probability 1, that the sequence

$$n^{-1/2} \sum_{i=1}^{n} (X_i^* - \hat{\theta}_n),$$

weakly converges to a Gaussian random variable with mean 0. In other words, we have that, with probability 1, for any $x \in \mathbb{R}$,

$$\mathbb{P}^* \left(n^{-1/2} \sum_{i=1}^{n} (X_{i,n}^* - \hat{\theta}_n) \leq x \right) \longrightarrow \Phi(x).$$

We now show that (1.1) and (1.2) holds true. Using the union bound, this is equivalent to show that, for all $q \in \mathbb{N}^*$, with probability 1,

$$n^{-1} \sum_{i=1}^{n} \mathbb{E}^* \|X_{n,i}^*\|^2 \mathbb{1}_{\{\|X_{n,i}^*\| > (1/q) n^{1/2}\}} \longrightarrow 0,$$

$$n^{-1} \sum_{i=1}^{n} \mathbb{E}^* \left(X_{n,i}^* - \mathbb{E}^* X_{n,i}^*\right)^2 \longrightarrow \Sigma.$$

By definition,

$$\mathbb{E}^* \left(X_{n,i}^* - \mathbb{E}^* X_{n,i}^*\right)^2 = n^{-1} \sum_{i=1}^{n} (X_i - \hat{\theta}_n)^2 = n^{-1} \sum_{i=1}^{n} X_i^2 - \hat{\theta}_n^2.$$

By the strong law of large number, the previous quantity converges almost surely to Σ. Hence we get the second of the required convergence. Let $\eta > 0$ and choose $M > 0$ large enough such that $\mathbb{E}[\|X_i\|^2 \mathbb{1}_{\{\|X_i\| > M\}}] \leq \eta$. This is possible by virtue of the Lebesgue dominated convergence theorem. We have with probability 1 that

$$n^{-1} \sum_{i=1}^{n} \|X_i\|^2 \mathbb{1}_{\{\|X_i\| > M\}} \to \mathbb{E}[\|X_1\|^2 \mathbb{1}_{\{\|X_1\| > M\}}] \leq \eta.$$

Besides, there exists $N \geq 1$ such that for all $n \geq N$,

$$\mathbb{E}^* \|X_{n,i}^*\|^2 \mathbb{1}_{\{\|X_{n,i}^*\| > (1/q) n^{1/2}\}} = n^{-1} \sum_{i=1}^{n} \|X_i\|^2 \mathbb{1}_{\{\|X_i\| > (1/q) n^{1/2}\}} \leq n^{-1} \sum_{i=1}^{n} \|X_i\|^2 \mathbb{1}_{\{\|X_i\| > M\}}.$$

It follows that with probability 1,

$$\limsup_{n \to +\infty} \mathbb{E}^* \|X_{n,i}^*\|^2 \mathbb{1}_{\{\|X_{n,i}^*\| > (1/q) n^{1/2}\}} \leq \eta,$$

but η is arbitrary. □

1.3.4 Bootstrap confidence intervals

We shall see in the next few lines that the knowledge of the weak limit is enough to assess the accuracy (in some sense) of $\hat{\theta}_n$ estimating θ_0. What is suitable to build confidence intervals and to conduct statistical tests is to control the estimation accuracy under a certain probability level, i.e., to find $\xi_n : (0,1) \to \mathbb{R}$ such that for any $\alpha \in (0,1)$ and any $n \geq 1$,

$$\mathbb{P}\left(n^{-1/2}\xi_n(\alpha/2) \leq \hat{\theta}_n - \theta_0 \leq n^{-1/2}\xi_n(1-\alpha/2)\right) \geq 1-\alpha.$$

Values $\xi_n(1-\alpha/2)$ and $\xi_n(\alpha/2)$ that guarantee the previous inequality to hold for all $n \geq 1$ are often very large and too pessimistic. One can rather weakened the previous condition and ask that

$$\liminf_{n\to\infty} \mathbb{P}\left(n^{-1/2}\xi_n(\alpha/2) \leq \hat{\theta}_n - \theta_0 \leq n^{-1/2}\xi_n(1-\alpha/2)\right) \geq 1-\alpha, \qquad \forall \alpha \in (0,1). \quad (1.3)$$

Definition 1.7. *A (possibly random) sequence of function $\xi_n : (0,1) \to \mathbb{R}$ is called a consistent quantile sequence if (1.3) holds.*

Exercise 1.2 shows that weak convergence is enough to furnish consistent quantile sequences provided that the asymptotic variance is consistently estimated.

As described in Section 1.2.2, the bootstrap technique allows to simulate versions of $n^{1/2}(\hat{\theta}_n^* - \hat{\theta}_n)$ and those simulations enable to approximate accurately the law of $n^{1/2}(\hat{\theta}_n^* - \hat{\theta}_n)$. The question is now to know whether the law of $n^{1/2}(\hat{\theta}_n^* - \hat{\theta}_n)$ can be used to produce consistent quantile sequences. We have seen in Exercise 1.2 that weak convergence was enough. Consequently, for the bootstrap, it is natural to rely on the notion of *conditional* weak convergence defined previously. In fact, another notion of weak convergence of bootstrap estimate, which is weaker than the initial one, is useful.

Definition 1.8. *We say that (Y_n^*) converges in distribution to F, conditionally on the observations, in probability, and write $Y_n^* \rightsquigarrow F$, in probability, if for every y where F is continuous, $\mathbb{P}^*(Y_n^* \leq y) \to F(y)$ in probability as $n \to \infty$.*

The next proposition informs us that this notion of conditional weak convergence, in probability, is enough to provide consistent bootstrap quantile sequences.

Proposition 1.12. *Suppose that $n^{1/2}(\hat{\theta}_n - \theta_0) \rightsquigarrow \mathcal{N}(0,\sigma^2)$ and that $n^{1/2}(\hat{\theta}_n^* - \hat{\theta}_n) \rightsquigarrow \mathcal{N}(0,\sigma^2)$, in probability, then the quantiles of $n^{1/2}(\hat{\theta}_n^* - \hat{\theta}_n)$ are consistent quantile sequences.*

Proof. Define

$$\hat{F}_n(x) = \mathbb{P}(n^{1/2}(\hat{\theta}_n^* - \hat{\theta}_n) \leq x \mid X_1, \dots X_n).$$

The hat over F_n is to remember that this quantity depends on the sample. Denote by $\hat{\xi}_n(\alpha)$ the quantile of level $\alpha \in (0,1)$ of the function $\hat{F}_n$, i.e.,

$$\hat{\xi}_n(\alpha) = \hat{F}_n^-(\alpha) = \inf\{x \in \mathbb{R} : \hat{F}_n(x) \geq \alpha\}.$$

Let Φ denote the cumulative distribution function of $\mathcal{N}(0,\sigma^2)$. By assumption, for every $x \in \mathbb{R}$ and every $\epsilon > 0$, we have, as $n \to \infty$,

$$\mathbb{P}(|\hat{F}_n(x) - \Phi(x)| > \epsilon) \to 0.$$

Let $\alpha \in (0,1)$. Write

$$\begin{aligned}\Phi\left(\hat{F}_n^-(\alpha)\right) &= \int_{-\infty}^{\hat{F}_n^-(\alpha)} \mathrm{d}\Phi(x) \\ &= \int \mathbb{1}_{\{x<\hat{F}_n^-(\alpha)\}} \mathrm{d}\Phi(x) \\ &= \int \mathbb{1}_{\{\hat{F}_n(x)<\alpha\}} \mathrm{d}\Phi(x).\end{aligned}$$

Consequently,

$$\begin{aligned}|\Phi\left(\hat{F}_n^-(\alpha)\right) - \Phi\left(\Phi^-(\alpha)\right)| &\leq \int |\mathbb{1}_{\{\hat{F}_n(x)<\alpha\}} - \mathbb{1}_{\{\Phi(x)<\alpha\}}| \mathrm{d}\Phi(x) \\ &= \int \mathbb{1}_{\{\hat{F}_n(x)<\alpha\leq\Phi(x)\}}\, \mathrm{d}\Phi(x) + \mathbb{1}_{\{\Phi(x)<\alpha\leq\hat{F}_n(x)\}}\, \mathrm{d}\Phi(x). \quad (1.4)\end{aligned}$$

Consider the first term on the left-hand side. Let $\epsilon > 0$, we have

$$\begin{aligned}\int \mathbb{1}_{\{\hat{F}_n(x)<\alpha\leq\Phi(x)\}}\, \mathrm{d}\Phi(x) &\leq \int \mathbb{1}_{\{\hat{F}_n(x)<\alpha\leq\Phi(x),\, |\hat{F}_n(x)-\Phi(x)|\leq\epsilon\}}\, \mathrm{d}\Phi(x) + \int \mathbb{1}_{\{|\hat{F}_n(x)-\Phi(x)|>\epsilon\}}\, \mathrm{d}\Phi(x) \\ &\leq \int \mathbb{1}_{\{\Phi(x)-\epsilon<\alpha\leq\Phi(x)\}}\, \mathrm{d}\Phi(x) + \int \mathbb{1}_{\{|\hat{F}_n(x)-\Phi(x)|>\epsilon\}}\, \mathrm{d}\Phi(x) \\ &= \int \mathbb{1}_{\{\alpha\leq\Phi(x)<\alpha+\epsilon\}}\, \mathrm{d}\Phi(x) + \int \mathbb{1}_{\{|\hat{F}_n(x)-\Phi(x)|>\epsilon\}}\, \mathrm{d}\Phi(x) \\ &\leq \epsilon + \int \mathbb{1}_{\{|\hat{F}_n(x)-\Phi(x)|>\epsilon\}}\, \mathrm{d}\Phi(x).\end{aligned}$$

Taking the expectation in the previous inequality gives, by Fubini's theorem,

$$\mathbb{E}\left[\int \mathbb{1}_{\{\hat{F}_n(x)<\alpha\leq\Phi(x)\}}\, \mathrm{d}\Phi(x)\right] \leq \epsilon + \int P(|\hat{F}_n(x) - \Phi(x)| > \epsilon)\, \mathrm{d}\Phi(x).$$

By assumption, we have for every $x \in \mathbb{R}$ that $\mathbb{P}(|\hat{F}_n(x) - \Phi(x)| > \epsilon) \to 0$. Hence, by the Lebesgue dominated convergence theorem, we get that

$$\limsup_{n\to\infty} \mathbb{E}\left[\int \mathbb{1}_{\{\hat{F}_n(x)<\alpha\leq\Phi(x)\}}\, \mathrm{d}\Phi(x)\right] \leq \epsilon.$$

Because ϵ is arbitrary, the previous limit is 0. The right-hand side term in (1.4) is treated in a similar fashion and we find that

$$\mathbb{E}\left[|\Phi\left(\hat{F}_n^-(\alpha)\right) - \Phi\left(\Phi^-(\alpha)\right)|\right] \longrightarrow 0.$$

Since the L_1 convergence implies the convergence in probability and using the continuous mapping theorem, we have that for every $\alpha \in (0,1)$, $\hat{\xi}_n(\alpha) = \hat{F}_n^-(\alpha) \to \Phi^-(\alpha)$, in probability. As a consequence, invoking Slutsky's Lemma, we have, for every $\alpha \in (0,1)$,

$$n^{1/2}(\hat{\theta}_n - \theta_0) - \hat{\xi}_n(\alpha) \xrightarrow{\mathrm{d}} \mathcal{N}(0,\, \sigma^2) - \Phi^-(\alpha),$$

implying that

$$\begin{aligned}\mathbb{P}\left(\hat{\theta}_n - \theta_0 \leq n^{-1/2}\hat{\xi}_n(1-\alpha/2)\right) &\to 1 - \alpha/2 \\ \mathbb{P}\left(\hat{\theta}_n - \theta_0 \leq n^{-1/2}\hat{\xi}_n(\alpha/2)\right) &\to \alpha/2.\end{aligned}$$

As a consequence, $\hat{\xi}_n$ is a consistent quantile sequence. □

Of course the "in probability" notion of conditional weak convergence, is weaker than the almost sure version. In fact the two previous definitions are related to the Kolmogorov-Smirnov distance when the limit $P(Y \leq y)$ is a continuous function which will be the case in most examples. This is a consequence of this Lemma (which is stated for the "in probability"version only but is valid as well for the almost sure varsion).

Lemma 1.13. *Let F_n be a sequence of (random) cdf defined on $\mathbb{R}$ and suppose that for each $x \in \mathbb{R}$, $\mathbb{P}(|F_n(x) - F(x)| > \epsilon) \to 0$ where F is a continuous (fixed) cdf. Then*

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to 0, \qquad \textit{in probability.}$$

Proof. Let $\epsilon > 0$ and $\Delta_n(x) = F_n(x) - F(x)$. Choose $R > 0$ such that $F(-R) \leq \epsilon/4$. Note that for any $x < -R$,

$$|\Delta_n(x)| \leq F_n(-R) + F(-R) \leq |\Delta_n(-R)| + \epsilon/2.$$

Consequently, $\mathbb{P}(\sup_{x<-R} |\Delta_n(x)| > \epsilon) \leq \mathbb{P}(|\Delta_n(-R)| > \epsilon/2) \to 0$. We have shown that $\sup_{x<-R} |\Delta_n(x)| \to 0$ in probability. Similarly, we obtain that $\sup_{x>R} |\Delta_n(x)| \to 0$ in probability and it only remains to show that $\sup_{x \in [-R,R]} |\Delta_n(x)| \to 0$ in probability. Let $\epsilon > 0$ and take $-R = b_1 < b_2 < \ldots < b_K = R$ such that $F(b_{k+1}) - F(b_k) \leq \epsilon/2$. This is possible in virtue of the Heine-Cantor theorem which asserts that F is uniformly continuous over $[-R, R]$. It follows that, for any $x \in [-R, R]$,

$$|\Delta_n(x)| \leq \max_{k=1,\ldots,K} |\Delta_n(b_k)| + \epsilon/2.$$

Consequently, $\mathbb{P}(\sup_{x \in [-R,R]} |\Delta_n(x)| > \epsilon) \leq \mathbb{P}(\max_{k=1,\ldots,K} |\Delta_n(b_k)| > \epsilon/2) \to 0$. □

1.4 Bootstrapping the covariance estimate

This section investigates the validity of the bootstrap for covariance estimates. Let X and Y be two real valued random variables and denote by Σ_{XY} the covariance between X and Y. We are interested in the bootstrap of the classical covariance estimator constructed from a collection $(X_1, Y_1), \ldots, (X_n, Y_n)$, of independent and identically distributed random variables each having the same distribution as (X, Y). This estimator is given by

$$\hat{\Sigma}_{XY} = n^{-1} \sum_{i=1}^{n} (X_i - \hat{\mu}_X)(Y_i - \hat{\mu}_Y),$$

where

$$\hat{\mu}_X = n^{-1} \sum_{i=1}^{n} X_i, \qquad \hat{\mu}_Y = n^{-1} \sum_{i=1}^{n} Y_i.$$

The associated bootstrap estimator is given by

$$\hat{\Sigma}^*_{XY} = n^{-1} \sum_{i=1}^{n} (X^*_{n,i} - \hat{\mu}^*_X)(Y^*_{n,i} - \hat{\mu}^*_Y),$$

where

$$\hat{\mu}^*_X = n^{-1} \sum_{i=1}^{n} X^*_{n,i}, \qquad \hat{\mu}^*_Y = n^{-1} \sum_{i=1}^{n} Y^*_{n,i}.$$

The next proposition claims that $\sqrt{n}(\hat{\Sigma}^*_{XY} - \hat{\Sigma}_{XY})$ provides a valid bootstrap of the distribution of $\sqrt{n}(\hat{\Sigma}^*_{XY} - \hat{\Sigma}_{XY})$. The proof is based on the bootstrap central limit theorem and some preservation properties.

Theorem 1.14. *Let $(X,Y),(X_1,Y_1),(X_2,Y_2),\dots$ be independent and identically distributed random variables with distribution P such that $\mathbb{E}[|XY|^2] < \infty$, $\mathbb{E}|X|^2 < \infty$ and $\mathbb{E}|Y|^2 < \infty$. If $\Sigma = \theta(P) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$ where $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$ then the bootstrap $n^{1/2}(\hat{\Sigma}^* - \hat{\Sigma})$ is almost surely consistent.*

Proof. In order to apply the Lindeberg central limit theorem without difficulties the trick is to replace the estimated mean by their expectations and control for the reminder term. First, because $\sum_{i=1}^n (X_i - \hat{\mu}_X) = \sum_{i=1}^n (X^*_{n,i} - \hat{\mu}^*_X) = 0$, it holds that

$$\hat{\Sigma}_{XY} = n^{-1}\sum_{i=1}^n (X_i - \hat{\mu}_X)(Y_i - \mu_Y),$$
$$\hat{\Sigma}^*_{XY} = n^{-1}\sum_{i=1}^n (X^*_{n,i} - \hat{\mu}^*_X)(Y^*_{n,i} - \mu_Y).$$

It follows that

$$\hat{\Sigma}_{XY} = n^{-1}\sum_{i=1}^n (X_i - \mu_X)(Y_i - \mu_Y) + (\mu_X - \hat{\mu}_X)(\hat{\mu}_Y - \mu_Y),$$
$$\hat{\Sigma}^*_{XY} = n^{-1}\sum_{i=1}^n (X^*_{n,i} - \mu_X)(Y^*_{n,i} - \mu_Y) + (\mu_X - \hat{\mu}^*_X)(\hat{\mu}^*_Y - \mu_Y).$$

As a consequence

$$n^{1/2}(\hat{\Sigma}^*_{XY} - \hat{\Sigma}_{XY}) = n^{-1/2}\sum_{i=1}^n \big((X^*_{n,i} - \mu_X)(Y^*_{n,i} - \mu_Y) - (X_i - \mu_X)(Y_i - \mu_X)\big) + R^*_n,$$

with

$$R^*_n = n^{1/2}(\mu_X - \hat{\mu}^*_X)(\hat{\mu}^*_Y - \mu_Y) - n^{1/2}(\mu_X - \hat{\mu}_X)(\hat{\mu}_Y - \mu_Y).$$

Because

$$\begin{aligned}
&(\mu_X - \hat{\mu}^*_X)(\hat{\mu}^*_Y - \mu_Y)\\
&= (\mu_X - \hat{\mu}_X)(\hat{\mu}^*_Y - \mu_Y) + (\hat{\mu}_X - \hat{\mu}^*_X)(\hat{\mu}^*_Y - \mu_Y)\\
&= (\mu_X - \hat{\mu}_X)(\hat{\mu}^*_Y - \hat{\mu}_Y) + (\mu_X - \hat{\mu}_X)(\hat{\mu}_Y - \mu_Y) + (\hat{\mu}_X - \hat{\mu}^*_X)(\hat{\mu}^*_Y - \mu_Y),
\end{aligned}$$

we find that

$$R^*_n = n^{1/2}(\mu_X - \hat{\mu}_X)(\hat{\mu}^*_Y - \hat{\mu}_Y) + n^{1/2}(\hat{\mu}_X - \hat{\mu}^*_X)(\hat{\mu}^*_Y - \mu_Y)$$

The second term is the product between $n^{1/2}(\hat{\mu}^*_X - \hat{\mu}_X)$, that converges in distribution to a Gaussian, almost surely (from Theorem 1.11), and $\hat{\mu}^*_Y - \mu_Y = (\hat{\mu}^*_Y - \hat{\mu}_Y) + (\hat{\mu}_Y - \mu_Y)$, that goes to 0, in $\mathbb{P}^*$-probability, almost surely. Indeed, let $\epsilon > 0$ and $M > 0$, for n large enough,

$$\mathbb{P}^*(|\hat{\mu}_Y - \hat{\mu}^*_Y| > \epsilon) \leq \mathbb{P}^*(|n^{1/2}(\hat{\mu}_Y - \hat{\mu}^*_Y)| > M),$$

hence, using Theorem 1.11, we have with probability 1,

$$\limsup_{n\to\infty} \mathbb{P}^*(|\hat{\mu}_Y - \hat{\mu}^*_Y| > \epsilon) \leq 2(1 - \Phi(M)).$$

As M is arbitrary, we have with probability 1 that $\mathbb{P}^*(|\hat{\mu} - \hat{\mu}^*| > \epsilon) \to 0$. Using Slutsky's Lemma we obtain that the first term converges in distribution to 0, i.e., in probability

to 0. The second term in R_n^* is the product between one term $n^{1/2}(\hat{\mu}_Y^* - \hat{\mu}_Y)$ that converges, almost surely, to a Gaussian distribution (from Theorem 1.11) and one other term $\hat{\mu}_X - \mu_X$ that goes to 0, $\mathbb{P}$-almost surely (implying that it goes to 0 in $\mathbb{P}^*$-probability, $\mathbb{P}$-almost surely).

Invoking again Slutsky's lemma for the bootstrap, it remains to show that the sequence

$$Z_n^* = n^{-1/2} \sum_{i=1}^{n} \big((X_{n,i}^* - \mu_X)(Y_{n,i}^* - \mu_Y) - (X_i - \mu_X)(Y_i - \mu_X) \big)$$

converges in distribution, almost surely, to the same limiting distribution as the one of the non bootstrap covariance estimation limit, i.e. $\mathcal{N}(0, \mathrm{var}((X_1 - \mathbb{E}[X_1])(Y_1 - \mathbb{E}[Y_1])))$. This is a consequence of Theorem 1.11 applied with $(X_1 - \mathbb{E}[X_1])(Y_1 - \mathbb{E}[Y_1])$ in place of X_1.

□

1.5 Edgeworth expansion

1.5.1 Studentization improves the accuracy

Let $(X_i)_{i \geq 1}$ be an independent and identically distributed sequence of random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with common distribution P. Let $(X_{n,i}^*)_{i=1,\ldots,n}$, $n \geq 1$, be the associated bootstrap triangular array. Let $\theta(P) = \int x P(dx)$ and define

$$\theta_0 = \theta(P), \qquad \hat{\theta}_n = \theta(P_n), \qquad \hat{\theta}_n^* = \theta(P_n^*).$$

Consider the problem of approximating the distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ whose cumulative distribution function is

$$F_n(x) = \mathbb{P}(\sqrt{n}(\hat{\theta}_n - \theta_0) \leq x).$$

The asymptotic approximation of F_n is given by

$$F_n^{(a)}(x) = \Phi(x/\hat{\sigma}),$$

where $\hat{\sigma}^2$ is an empirical estimate of the variance. The central limit theorem combined with Slutsky's lemma has shown in exercise that such an approximation leads to asymptotically valid confidence intervals. Concerning the Bootstrap, we have actually 2 different possibilities. The first one is related to the bootstrap central limit theorem. The approximation of F_n is given by

$$F_n^{(b)}(x) = \mathbb{P}(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \leq x).$$

This approach would also lead to valid confidence intervals. The other approach for the bootstrap consists in approximating the distribution of the Studentized-statistics $\tilde{F}_n(x) = \mathbb{P}(\sqrt{n}(\hat{\theta}_n - \theta_0)/\hat{\sigma} \leq x)$ rather than F_n. The bootstrap version is defined as

$$\tilde{F}_n^{(b)}(x) = \mathbb{P}(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma}_n^* \leq x).$$

Working with $\tilde{G}_n$ is better than working with G_n. The appropriate tool to show this is the one of Edgeworth expansion. Under suitable assumption on X_1, that is X_1 admits a density and $\mathbb{E}|X_1|^3 < \infty$, the Edgeworth expansion of the distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)/\sigma$ and $\sqrt{n}(\hat{\theta}_n - \theta_0)/\hat{\sigma}_n$ at second order are

$$\mathbb{P}(\sqrt{n}(\hat{\theta}_n - \theta_0)/\sigma \leq x) = \{\Phi(x) + n^{-1/2} p(x)\phi(x)\} + O(n^{-1}),$$
$$\mathbb{P}(\sqrt{n}(\hat{\theta}_n - \theta_0)/\hat{\sigma} \leq x) = \{\Phi(x) + n^{-1/2} q(x)\phi(x)\} + O(n^{-1}),$$

Replacing the distribution P with empirical one, we obtain the bootstrap version of the previous expansion

$$\mathbb{P}^*(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma} \leq x) = \{\Phi(x) + n^{-1/2}\hat{p}(x)\phi(x)\} + O_{\mathbb{P}}(n^{-1}),$$
$$\mathbb{P}^*(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma}^* \leq x) = \{\Phi(x) + n^{-1/2}\hat{q}(x)\phi(x)\} + O_{\mathbb{P}}(n^{-1}),$$

As a consequence,

$$\begin{aligned}|F_n(x) - F_n^{(a)}(x)| &= |\mathbb{P}(\sqrt{n}(\hat{\theta}_n - \theta_0)/\sigma \leq x/\sigma) - \Phi(x/\hat{\sigma})| \\ &= (\Phi(x/\sigma) - \Phi(x/\hat{\sigma})) + n^{-1/2}p(x/\sigma)\phi(x/\sigma) + O(n^{-1}).\end{aligned}$$

Moreover,

$$\begin{aligned}|F_n(x) - F_n^{(b)}(x)| &= |\mathbb{P}(\sqrt{n}(\hat{\theta}_n - \theta_0)/\sigma \leq x/\sigma) - \mathbb{P}^*(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma} \leq x/\hat{\sigma})| \\ &= (\Phi(x/\sigma) - \Phi(x/\hat{\sigma})) + O_{\mathbb{P}}(n^{-1/2})\end{aligned}$$

Finally for the Studentized option for the bootstrap gives

$$\begin{aligned}|\tilde{F}_n(x) - \tilde{F}_n^{(b)}(x)| &= |\mathbb{P}(\sqrt{n}(\hat{\theta}_n - \theta_0)/\hat{\sigma} \leq x) - \mathbb{P}^*(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma}^* \leq x)| \\ &= (\Phi(x) - \Phi(x)) + n^{-1/2}(q(x) - \hat{q}(x)) + O_{\mathbb{P}}(n^{-1}).\end{aligned}$$

Because $q(x) = q(x, P)$ and $\hat{q}(x) = q(x, \hat{P})$, some smoothness in $Q \mapsto q(x, Q)$ would lead to $q(x) - \hat{q}(x) = O_{\mathbb{P}}(n^{-1/2})$ which would result in

$$|\tilde{F}_n(x) - \tilde{F}_n^{(b)}(x)| = O_{\mathbb{P}}(n^{-1}).$$

The Studentized-bootstrap is then more accurate than the asymptotic approximation. The distribution that is used to approximate $F_n(x) = \tilde{F}_n(x/\hat{\sigma})$ is then $\tilde{F}_n^{(b)}(x/\hat{\sigma})$. The confidence intervals associated to this method are then

$$\left[\hat{\theta}_n - (\hat{\sigma}/\sqrt{n})F_n^{(b)-}(1-\alpha/2), \hat{\theta}_n - (\hat{\sigma}/\sqrt{n})F_n^{(b)-}(\alpha/2)\right].$$

This method is refereed to as the Studentized bootstrap. It is known to improve the accuracy of traditional and basic bootstrap confidence intervals.

1.5.2 Pivotal statistics

The previous phenomenon can be extended to pivotal statistics (Hall, 2013) whose definition is as follows.

Definition 1.9. *Let $(X_i)_{i\geq 1}$ be independent random variables with common distribution P and let $P_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$. A map T defined on the set of probability distributions is said to be pivotal over A whenever $T(P_n) \rightsquigarrow L$, as $n \to \infty$, for all $P \in A$ (L does not depend on P).*

A basic example is the empirical mean where

$$T(P_n) = \frac{1}{\hat{\sigma}_n\sqrt{n}}\sum_{i=1}^n (X_i - \mathbb{E}[X_1]),$$

with $\hat{\sigma}_n^2$ the empirical estimate of the variance of X_1. It is pivotal over the set A of integrable random variables. Any regular map with Gaussian limit admits a pivotal version as using Slutsky, we only need to normalize $\sqrt{n}(\hat{\theta}_n - \theta_0)$ by its empirical variance. This leads a vast framework in which the bootstrap can be implemented using studentized representation of the statistics.

Exercises

Exercise 1.1. *Show that if for any $\alpha \in (0,1)$, $\mathbb{P}(\sqrt{n}(\hat{\theta} - \theta_0) \leq \xi_n(\alpha)) \to \alpha$, then ξ_n is a consistent quantile sequence. Let Φ denote the cumulative distribution function of the standard normal distribution. Show that if $n^{1/2}(\hat{\theta}_n - \theta_0)$ converges in distribution to $\mathcal{N}(0, \sigma^2)$ with $\sigma > 0$ and $\hat{\sigma}_n \xrightarrow{\mathbb{P}} \sigma$, then $\hat{\sigma}_n \Phi^-$ is a consistent quantile sequence.*

Exercise 1.2. *Let $(X_i)_{i \geq 1}$ be an independent and identically distributed sequence with common distribution $\exp(\lambda)$ with $\lambda > 0$. Define*

$$\hat{\lambda} = \left(n^{-1} \sum_{i=1}^{n} X_i \right)^{-1} \qquad \text{and} \qquad \hat{\lambda}^* = \left(n^{-1} \sum_{i=1}^{n} X_{n,i}^* \right)^{-1}.$$

1. *Show that $|\overline{X^*} - \overline{X}| = o_{\mathbb{P}}(1)$ almost-surely.*
2. *Show that $\overline{X^*} - 1/\lambda = o_{\mathbb{P}}(1)$ almost-surely.*
3. *Show that $\hat{\lambda}^* - \lambda = o_{\mathbb{P}}(1)$ almost-surely.*
4. *Show that $\sqrt{n}(\hat{\lambda}^* - \hat{\lambda}) \rightsquigarrow \mathcal{N}(0, \lambda^2)$ almost-surely.*

Exercise 1.3 (Lindeberg condition and LLN)**.** *For each $n \in \mathbb{N}^*$, let $Y_{n,1}, Y_{n,1}, \dots Y_{n,n}$ be a sequence of real-valued independent random variables such that, for every $\epsilon > 0$,*

$$n^{-1} \sum_{i=1}^{n} \mathbb{E}[|Y_{n,i}|^2 1_{\{|Y_{n,i}| > \sqrt{n}\epsilon\}}] \to 0, \qquad \text{when } n \to \infty.$$

We suppose that $\mathbb{E}Y_{n,i}^2 = 1$ for every $i \in \mathbb{N}^*$ and $n \in \mathbb{N}^*$. The aim of this exercise is to show that

$$\hat{v}_n = n^{-1} \sum_{i=1}^{n} Y_{n,i}^2 \to 1, \quad \text{in probability, when } n \to \infty.$$

1. *Define $Z_{n,i}^2 = Y_{n,i}^2 1_{\{|Y_{n,i}| \leq \sqrt{n}\epsilon\}}$. Show that the event $A_n = \{\exists i = 1, \dots, n, \quad Y_{n,i}^2 \neq Z_{n,i}^2\}$ has probability going to 0 as $n \to \infty$ (hint : one might write this event as a union).*
2. *Define $\hat{w}_n = n^{-1} \sum_{i=1}^{n} Z_{n,i}^2$ and show that $\limsup_n \mathbb{P}(|\hat{v}_n - 1| > \eta) \leq \limsup_n \mathbb{P}(|\hat{w}_n - 1| > \eta)$.*
3. *Show that $\hat{w}_n - 1 = n^{-1} \sum_{i=1}^{n} (Z_{n,i}^2 - \mathbb{E}[Z_{n,i}^2]) + o(1)$.*
4. *Show that $\mathbb{E}\left(n^{-1} \sum_{i=1}^{n} (Z_{n,i}^2 - \mathbb{E}[Z_{n,i}^2]) \right)^2 \leq \epsilon^2$.*
5. *Conclude (hint: one might start by showing that for any $\eta > 0$, $\mathbb{P}(|\omega_1 + \omega_2| > \eta) \leq \mathbb{P}(|\omega_1| > \eta/2) + \mathbb{P}(|\omega_2| > \eta/2)$)*

Exercise 1.4 (Lindeberg condition and LLN)**.** *For each $n \in \mathbb{N}^*$, let $Y_{n,1}, Y_{n,1}, \dots Y_{n,n}$ be a sequence of real-valued independent random variables such that, for every $\epsilon > 0$,*

$$n^{-1} \sum_{i=1}^{n} \mathbb{E}[|Y_{n,i}|^2 1_{\{|Y_{n,i}| > \sqrt{n}\epsilon\}}] \to 0, \qquad \text{when } n \to \infty.$$

We suppose that $\mathbb{E}Y_{n,i}^2 = \nu_n^2 \to \nu^2 > 0$ for every $i \in \mathbb{N}^*$ and $n \in \mathbb{N}^*$. Show that

$$(n\nu^2)^{-1} \sum_{i=1}^{n} Y_{n,i}^2 \to 1, \quad \text{in probability, when } n \to \infty.$$

Exercise 1.5. *Let $(X_i)_{i\geq 1}$ be an independent and identically distributed sequence of real valued random variables such that $0 < \mathbb{E}[X_1^2] < \infty$. For each $n \geq 1$, $(X^*_{n,i})_{i=1,\dots,n}$ is independent and identically distributed according to P_n, conditionally on $X_1, X_2, \dots$. Define*

$$\hat{\mu}_n = n^{-1} \sum_{i=1}^{n} X_i,$$
$$\hat{\sigma}_n^2 = n^{-1} \sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2.$$

Define $\hat{\mu}_n^*$ and $\hat{\sigma}_n^*$ accordingly.

1. *Show that almost surely, $\hat{\mu}_n^* \overset{\mathbb{P}^*}{\to} E[X_1]$ (hint: use Markov inequality).*
2. *Show that almost surely, $\hat{\sigma}_n^{*2} \overset{\mathbb{P}^*}{\to} \sigma^2 = \mathrm{var}(X_1)$ (hint: use Exercise 1.4).*
3. *Deduce that $\sqrt{n}(\hat{\mu}_n^* - \hat{\mu}_n)/\hat{\sigma}_n^* \rightsquigarrow \mathcal{N}(0,1)$, almost surely.*

Chapter 2

The Wasserstein distance and the bootstrap

The Wasserstein distance is an appropriate tool to derive bootstrap consistency in general metric space (Bickel et al., 1981) and also to handle different resampling schemes as the original bootstrap (e.g., smoothed bootstrap and parametric bootstrap).

2.1 Mathematical background

We start by recalling some basic general definitions dealing with metric spaces and the notion of weak convergence.

Definition 2.1. *A metric space (S, ρ) is called separable if there exists a countable dense subset, i.e., there is $(x_n)_{n\geq 1} \subset S$ such that for any $x \in S$ and $\epsilon > 0$, one can find $i \geq 1$ for which $\rho(x, x_i) < \epsilon$.*

Definition 2.2. *A metric space (S, ρ) is called complete when any Cauchy sequence in S converges in S.*

Definition 2.3. *A metric space (S, ρ) is called Polish when it is complete and separable.*

Definition 2.4. *A metric space (S, ρ) is called totally bounded when for any $\eta > 0$ there exists a finite number of open balls with radius η that covers S.*

Proposition 2.1. *A metric space complete and totally bounded is compact.*

Let $\mathcal{S}$ be the σ-algebra generated by the open sets in S. Denote by $\mathcal{P}(\mathcal{S})$ the set of probability measures defined on $\mathcal{S}$. We will make use of the following notation, for any $\mu \in \mathcal{P}(\mathcal{S})$ and any function $f : S \to \mathbb{R}$ such that $\int f \mathrm{d}\mu$ exists, we write

$$\mu(f) = \int f \mathrm{d}\mu.$$

Let $C_b(S)$ be the space of continuous real-valued functions defined on S.

Proposition 2.2. *Two probability measures α and β on $\mathcal{S}$ coincides if and only if $\alpha(f) = \beta(f)$ for all $f \in C_b(S)$.*

Proof. The proof relies on the regularity of probability measure (see Theorem 1.2 in Billingsley (2013)). □

Definition 2.5. *A sequence $(\mu_n)_{n\geq 1} \subset \mathcal{P}(\mathcal{S})$ is said to converge weakly to μ if*

$$\mu_n(f) \to \mu(f), \qquad \forall f \in C_b(S).$$

This is denoted by $\mu_n \rightsquigarrow \mu$.

Definition 2.6. *A set $\mathcal{M}$ of probability measures defined on $\mathcal{S}$ is called tight if for any $\epsilon > 0$, there exists a compact set $K \subset S$ such that*

$$\forall \mu \in \mathcal{P}, \qquad \mu(K) \geq 1 - \epsilon.$$

Now we can state two useful properties dealing with tightness in Polish spaces.

Proposition 2.3. *If S is a Polish space, any probability measure on $\mathcal{S}$ is tight.*

Proof. The proof follows the one provided in the lecture notes of Jon A. Wellner. Let $m \geq 1$ and $\epsilon > 0$. The open set $\cup_{x\in S} B(x, 1/(2m))$ covers S. Because S is separable, there is $(x_n)_{n\geq 1}$ such that for every $x \in S$ one can find $x_i \in B(x, 1/(2m))$. Consequently $\cup_{n\geq 1} B(x_n, 1/m)$ covers S. Define $A_N(m) = \cup_{1\leq n\leq N} B(x_n, 1/m)$. By the monotone convergence theorem $\mu(A_N(m)) \to 1$ as $N \to \infty$. Choose $N_{m,\epsilon}$ such that the $\mu(A_{N_{m,\epsilon}}(m))$ is greater than $1 - \epsilon/2^m$. Then

$$\mu(\overline{\cap_{m\geq 1} A_{N_{m,\epsilon}}(m)}^c) \leq \mu(\cap_{m\geq 1} A_{N_{m,\epsilon}}(m)^c) \leq \sum_{m\geq 1} \mu(A_{N_{m,\epsilon}}(m)^c) \leq \epsilon.$$

It remains to show that $\overline{\cap_{m\geq 1} A_{N_{m,\epsilon}}(m)}$ is a compact set. It is a closed set in a complete space hence it is complete. It remains to show that it is totally bounded. We have that $\cap_{m\geq 1} A_{N_{m,\epsilon}}(m) \subset A_{N_{m,\epsilon}}(m)$, for each $m \geq 1$ and therefore

$$\overline{\cap_{m\geq 1} A_{N_{m,\epsilon}}(m)} \subset \overline{A_{N_{m,\epsilon}}(m)}$$

The set $\overline{A_{N_{m,\epsilon}}(m)}$ is the union of a finite number of closed balls with radius $1/m$. It is included in the corresponding union of open balls with radius $2/m$. □

A sequence $(\nu_n)_{n\geq 1}$ is called a subsequence of $(\mu_n)_{n\geq 1}$ if there exists $\varphi : \mathbb{N} \to \mathbb{N}$ such that $\nu_n = \mu_{\varphi(n)}$ with $\varphi(n) < \varphi(n+1)$ for any n.

Definition 2.7. *A set $\mathcal{M}$ of probability measures defined on $\mathcal{S}$ is called relatively compact if for any sequence $(\mu_n)_{n\geq 1}$ there exists a further subsequence that converges weakly in $\mathcal{P}(\mathcal{S})$ (not necessarily in $\mathcal{M}$).*

Proposition 2.4 (Prohorov)**.** *Let S be a Polish space and $\mathcal{M}$ be a set of probability measures on $\mathcal{S}$. Then $\mathcal{M}$ is relatively compact if and only if it is tight.*

Proposition 2.5. *Let S be a Polish space and $(\mu_n)_{n\geq 1}$ be a tight set of probability measures on $\mathcal{S}$. Suppose that every weakly convergent subsequence of $(\mu)_{n\geq 1}$ converges weakly to μ. Then $\mu_n \rightsquigarrow \mu$.*

Proof. By contradiction. There exists $\epsilon > 0$, φ_n and $f \in C_b(S)$ such that $|\mu_{\varphi_n}(f) - \mu_{\varphi_n}(f)| > \epsilon$. Because it is relatively compact and using the assumption, there is a further subsequence that converges to μ. □

2.2 The Wasserstein distance

To define the Wasserstein distance, we need to define sets of probability measures with given marginals. Let $\alpha \in \mathcal{P}(\mathcal{S})$ and $\beta \in \mathcal{P}(\mathcal{S})$. Define $\mathcal{A}(\alpha, \beta)$ as the set of measures defined on the product space $\mathcal{S} \otimes \mathcal{S}$ (the product σ-field) having marginals α and β. That is,

$$\mathcal{A}(\alpha, \beta) = \{\mu \in \mathcal{P}(\mathcal{S} \otimes \mathcal{S}) \quad : \quad \mu(A \times S) = \alpha(A), \quad \mu(S \times A) = \beta(A) \quad \forall A \in \mathcal{S}\}.$$

If p_1 and p_2 denote the projection map, i.e., for any $i = 1, 2$, $p_i : S \times S \to S$ with $p_i(x_1, x_2) = x_i$. We have

$$\mathcal{A}(\alpha, \beta) = \{\mu \in \mathcal{P}(\mathcal{S} \otimes \mathcal{S}) \quad : \quad \mu \circ p_1^{-1} = \alpha, \quad \mu \circ p_2^{-1} = \beta\}.$$

Indeed, for all $A \in \mathcal{S}$, we have $p_1^{-1}(A) = \{x \in S \times S : p_1(x) \in A\} = \{x \in S : p_1(x) \in A, p_2(x) \in S\}$.

Proposition 2.6. *Let S is a Polish space and (α, β) be two probability measures on $\mathcal{S}$. The set $\mathcal{A}(\alpha, \beta)$ is tight.*

Proof. Because $A_1 \times A_2 \subset (A_1 \times S) \cup (S \times A_2)$, we have

$$\mu((S \times S) \backslash (K_1 \times K_2)) \leq \mu((S \backslash K_1) \times S) + \mu(S \times (S \backslash K_2)) = \alpha(S \backslash K_1) + \beta(S \backslash K_2).$$

□

Let α and β be two probability measures on $\mathcal{S}$. The p-Wasserstein distance between α and β, denoted $W_p(\alpha, \beta)$, is defined as

$$W_p(\alpha, \beta) = \inf_{\mu \in \mathcal{A}(\alpha, \beta)} \left(\int \rho(x, y)^p \mathrm{d}\mu(x, y) \right)^{1/p},$$

where the space $\mathcal{A}(\alpha, \beta)$ (which has been defined in the previous section) is the space of probability measures on $\mathcal{S} \otimes \mathcal{S}$ with respective marginal α and β. The existence of the $W_p(\alpha, \beta)$ is established in the next proposition.

Proposition 2.7. *Let S be a Polish space and α and β be two probability measures defined on $\mathcal{S}$. There exists μ, a probability measure on $\mathcal{S} \otimes \mathcal{S}$ such that*

$$W_p(\alpha, \beta) = \left(\int \rho(x, y)^p \mathrm{d}\mu(x, y) \right)^{1/p}.$$

The set of such μ is denoted by $\mathrm{opt}(\alpha, \beta)$. *Any such μ is called an optimal plan for (α, β).*

The proof is based on the following Lemma.

Lemma 2.8. *The map $\mu \mapsto \int \rho(x, y)^p \mathrm{d}\mu(x, y)$ defined on the space of probability measure on $\mathcal{S} \otimes \mathcal{S}$ is lower semicontinuous with respect to the topology of weak convergence.*

Proof of Lemma 2.8 Define $K : \mathbb{R}_{\geq 0} \to \mathbb{R}$ as the piece-wise continuous affine function such that $K(x) = 1$ if $0 \leq x \leq 1$ and $K(x) = 0$ if $x \geq 2$. Because the sequence $f_m(x, y) = \rho(x, y)^p K(\rho(x, y)/m)$ is increasing, the monotone convergence theorem gives that

$$\int \rho(x, y)^p \mathrm{d}\mu(x, y) = \limsup_{m \to \infty} \int \rho(x, y)^p K(\rho(x, y)/m) \mathrm{d}\mu(x, y).$$

Let $\mu_n \rightsquigarrow \mu$. It follows that, for any m,

$$\int \rho(x,y)^p \mathrm{d}\mu_n(x,y) \geq \int \rho(x,y)^p K(\rho(x,y)/m)\mathrm{d}\mu_n(x,y).$$

Because $(x,y) \mapsto \rho(x,y)^p K(\rho(x,y)/m)$ is continuous and bounded, the weak convergence implies that

$$\liminf_{n\to\infty} \int \rho(x,y)^p \mathrm{d}\mu_n(x,y) \geq \int \rho(x,y)^p K(\rho(x,y)/m)\mathrm{d}\mu(x,y).$$

As the previous is true for any m, we take the supremum and this gives the lower semi-continuity. □

Proof of proposition 2.7 As for any $\mu \in \mathcal{A}(\alpha,\beta)$, we have that $0 \leq \int \rho(x,y)^p \mathrm{d}\mu(x,y)$, the image of $\mathcal{A}(\alpha,\beta)$ trough the map $\mu \mapsto \int \rho(x,y)^p \mathrm{d}\mu(x,y)$ is contained in $[0,\infty)$. Hence, the infimum exists and is denoted by m. Consequently, $m + 1/n$ is not a lower bound and then there exists $\mu_n \in \mathcal{A}(\alpha,\beta)$ such that

$$m \leq \int \rho(x,y)^p \mathrm{d}\mu_n(x,y) \leq m + 1/n.$$

By proposition 2.6 and Prohorov theorem, $\mathcal{A}(\alpha,\beta)$ is relatively compact. Hence there is a subsequence μ_{φ_n} that converges weakly to a certain measure μ. From the previous inequality and Lemma 2.8, we have

$$m \geq \liminf_{n\to\infty} \int \rho(x,y)^p \mathrm{d}\mu_{\varphi_n}(x,y) \geq \int \rho(x,y)^p \mathrm{d}\mu(x,y),$$

which implies that $m = \int \rho(x,y)^p \mathrm{d}\mu(x,y)$. The proof will be complete if the measure μ belongs to $\mathcal{A}(\alpha,\beta)$. Let $f \in C_b(S)$ and define $g(x,y) = f(x)$ for any $(x,y) \in S \times S$. Clearly, $g \in C_b(S \times S)$ and therefore

$$\mu_{\varphi_n}(g) \to \mu(g).$$

But for any $n \geq 1$, $\mu_{\varphi_n}(g) = \alpha(f)$ and $\mu(g) = \mu \circ p_1^{-1}(g)$. In virtue of Proposition 2.2, $\alpha = \mu \circ p_1^{-1}$.

□

Define the space $\mathcal{L}_p(\mathcal{S})$ as the space of probability measure admitting finite p-moments, that is,

$$\mathcal{L}_p(\mathcal{S}) = \left\{ \mu \in \mathcal{P}(\mathcal{S}) \quad : \quad \int \rho(x,x_0)^p d\mu(x) < \infty, \text{ for some } x_0 \in S \right\}.$$

Note that the previous definition does not depend on x_0. The property that W_p is a distance is established in the following proposition.

Proposition 2.9. *Let S be a Polish space. The map W_p is a distance on $\mathcal{L}_p(\mathcal{S})$.*

Proof. Let α and β be two probability measures in $\mathcal{L}_p(\mathcal{S})$. Let $\mu \in \mathcal{A}(\alpha,\beta)$. Because of the triangle inequality

$$\int \rho(x,y)^p d\mu(x,y) \leq 2^p \left(\int \rho(x,x_0)^p d\alpha(x) + \int \rho(x_0,x)^p d\beta(x) \right).$$

This implies that W_p is finite on $\mathcal{L}_p(\mathcal{S})$ hence valued in $[0,\infty)$. We now show that $W_p(\alpha,\alpha) = 0$. Consider the joint distribution μ of two random variables X and Y such that X has distribution α and $Y = X$ almost surely. Hence $\mu \in \mathcal{A}(\alpha,\alpha)$ and we have

$$W_p(\alpha,\alpha) \leq \int \rho(x,y)^p d\mu(x,y) = E_\mu[\rho(X,Y)^p] = 0.$$

Now we show that $W_p(\alpha,\beta)=0$ implies $\alpha=\beta$. Let $f\in C_b(S)$. If $W_p(\alpha,\beta)=0$, we have $\rho(x,y)=0$, $(d\mu)$-almost everywhere, hence $f(x)=f(y)$, $(d\mu)$-almost everywhere, implying that

$$\alpha(f)-\beta(f)=\int (f(x)-f(y))\mathrm{d}\mu(x,y)=0.$$

We finish with the most difficult part, the triangle inequality. Let (α,β,γ) be three elements in $\mathcal{L}_p(\mathcal{S})$. Let $\mu\in\mathrm{opt}(\alpha,\beta)$ and $\nu\in\mathrm{opt}(\beta,\gamma)$. Construct δ a measure on $\mathcal{S}\otimes\mathcal{S}\otimes\mathcal{S}$ such that $\delta\circ p_{12}^{-1}=\mu$ and $\delta\circ p_{23}^{-1}=\nu$. This is a consequence of the desintegration theorem: there exists μ_y and ν_y such that $\mathrm{d}\mu(x,y)=\mathrm{d}\beta(y)\mathrm{d}\mu_y(x)$ and $\mathrm{d}\nu(y,z)=\mathrm{d}\beta(y)\mathrm{d}\nu_y(z)$. Then define $\mathrm{d}\delta=\mathrm{d}\beta(y)\mathrm{d}(\mu_y\times\nu_y)(x,z)$. We have

$$\begin{aligned}W_p^p(\alpha,\gamma)&\leq\int\rho(x,z)^p\mathrm{d}\delta(x,y,z)\\&\leq\int(\rho(x,y)+\rho(y,z))^p\mathrm{d}\delta(x,y,z).\end{aligned}$$

Using Minkowsky inequality yields

$$\begin{aligned}W_p(\alpha,\gamma)&\leq\left(\int\rho(x,y)^p\mathrm{d}\delta(x,y,z)\right)^{1/p}+\left(\int\rho(y,z)^p\mathrm{d}\delta(x,y,z)\right)^{1/p}\\&=\left(\int\rho(x,y)^p\mathrm{d}\mu(x,y)\right)^{1/p}+\left(\int\rho(y,z)^p\mathrm{d}\nu(y,z)\right)^{1/p}\\&=W_p(\alpha,\beta)+W_p(\beta,\gamma).\end{aligned}$$

□

2.3 Relation to weak convergence

In this section we start with some useful properties related to W_1. This will allow to provide a first discussion about the links between convergence in W_1 and other convergences before we state the main result of the section that describes fully the relationship between the W_1 and weak convergence.

2.3.1 The case of W_1

An interesting case is when a sequence of measure converges to the dirac measure, i.e., $W_1(\alpha_n,\delta_x)\to 0$. In this particular case, the convergence in Wasserstein distance is equivalent to convergence in L_1.

Proposition 2.10. *Let S be a Polish space and α probability measure on $\mathcal{S}$ and $x\in S$. We have that $W_1(\alpha,\delta_x)=\int d(y,x)\mathrm{d}\alpha(y)$.*

Proof. Let $\alpha\in\mathcal{P}(S)$. The set of admissible measures for α and δ_x contains only one measure which is the product measure $\alpha\times\delta_x$. To show this, let $\mu\in\mathcal{A}(\alpha,\delta_x)$. We have (by definition) $\alpha(A)\geq\mu(A\times\{x\})$. Apply this with $S\backslash A$ to get $1-\alpha(A)\geq\mu(S\times\{x\})-\mu(A\times\{x\})=1-\mu(A\times\{x\})$. Hence $\alpha(A)\leq\mu(A\times\{x\})$. All this together implies that $\mu(A\times\{x\})=\alpha(A)$. It remains to note that $\mu(A\times B)=\mu(A\times B\cap\{x\})$. □

Proposition 2.11. *Suppose that S is a Polish space. We have*

$$W_1(\alpha,\beta)=\sup_{\|f\|_{lip}\leq 1}\left\{\int f\mathrm{d}\alpha-\int f\mathrm{d}\beta\right\},$$

where $\|f\|_{lip}=\sup_{x\neq y}|f(x)-f(y)|/\rho(x,y)$ ($\|f\|_{lip}\leq 1$ is the space of Lipschitz function with Lipschitz constant less than 1 with respect to the metric ρ).

Proof. See Chapter 6 in Villani (2008) or Ambrosio and Gigli (2013). □

The following proposition claims that the Lévy–Prokhorov metric metrizes weak convergence (Dudley, 2018a). The Lévy–Prokhorov metric is defined as

$$d_{LP}(\alpha, \beta) = \sup_{\|f\| \leq 1} \left\{ \int g \mathrm{d}\alpha - \int g \mathrm{d}\beta \right\}.$$

with $\|f\| = \sup_{x \in S} |f(x)| + \|f\|_{lip}$.

Proposition 2.12. *$\alpha_n \rightsquigarrow \alpha$ if and only if $d_{LP}(\alpha_n, \alpha) \rightsquigarrow 0$.*

The previous two propositions imply that convergence in W_1 implies weak convergence. The equivalence is true if S is bounded. The next proposition extends this remark by making a clear link between weak convergence and convergence in Wassertsein distance W_2. In fact we will see that convergence in W_2 implies weak convergence but the converse is false. We need slightly more than weak convergence, namely the convergence of second-order moments, to obtain the convergence of the Wasserstein distance.

2.3.2 Characterizing convergence in W_p using weak convergence

In the rest of the section, we consider the case of vector normed spaces and deal with the case $p = 2$. The metric $\rho(x, y)$ previously used now becomes $\|x - y\|$.

Proposition 2.13. *Let S be a Polish space and $\alpha, (\alpha_n)_{n \geq 1} \subset \mathcal{L}_2(\mathcal{S})$. We have that $W_2(\alpha_n, \alpha) \to 0$ if and only if*

$$\alpha_n \rightsquigarrow \alpha,$$
$$\alpha_n\{\|\cdot\|^2\} \to \alpha\{\|\cdot\|^2\}.$$

Proof. The "only if" part follows from

$$|\alpha_n(f) - \alpha(f)| \leq \int |f(x) - f(y)| \mathrm{d}\mu_n(x, y).$$

where $\mu_n \in \mathrm{opt}(\alpha_n, \alpha)$. The previous is true for any $f \in \mathcal{L}_p(\mathcal{S})$. We apply it taking f a bounded Lipschitz function and use Jensen inequality to get weak convergence. To obtain the second property, use the Minkowski inequality to get

$$|\sqrt{\alpha_n(\|\cdot\|^2)} - \sqrt{\alpha(\|\cdot\|^2)}| = \left| \sqrt{\int \|x\|^2 \, \mathrm{d}\mu_n(x, y)} - \sqrt{\int \|y\|^2 \, \mathrm{d}\mu_n(x, y)} \right| \leq \sqrt{\int \|x - y\|^2 \, \mathrm{d}\mu_n(x, y)}.$$

The "if" part is as follows (here S is a normed vector space). Let $M > 0$ be such that

$$\int_{\|x\| > M} \|x\|^2 \mathrm{d}\alpha \leq \epsilon.$$

Recall that $\mathbb{I}(x > 2M) \leq (1 - K(x/M)) \leq \mathbb{I}(x > M)$ and write

$$\begin{aligned}
\int_{\|x\| > 2M} \|x\|^2 \mathrm{d}\alpha_n &\leq \int \|x\|^2 (1 - K(\|x\|/M)) \mathrm{d}\alpha_n \\
&= \int \|x\|^2 \mathrm{d}\alpha_n - \int \|x\|^2 K(\|x\|/M) \mathrm{d}\alpha_n \\
&\to \int \|x\|^2 \mathrm{d}\alpha - \int \|x\|^2 K(\|x\|/M) \mathrm{d}\alpha \\
&\leq \int_{\|x\| > M} \|x\|^2 \mathrm{d}\alpha
\end{aligned}$$

Define α_M as the probability measure of $XK(\|X\|/M)$ when $X \sim \alpha$. By the triangle inequality, one has

$$W_2(\alpha_n, \alpha) \leq W_2(\alpha_{2M}, \alpha) + W_2(\alpha_{n,2M}, \alpha_n) + W_2(\alpha_{n,2M}, \alpha_{2M}).$$

Let $\epsilon > 0$. Because $\|X - XK(\|x\|/M)\|^2 \leq \|X\|^2 \mathbb{I}(\|X\| > M)$, one may choose M large enough such that

$$W_2(\alpha_{2M}, \alpha) \leq \epsilon$$
$$W_2(\alpha_{n,2M}, \alpha_n) \leq \epsilon.$$

Now we can use that $\alpha_{n,2M}$ and α_{2M} are both supported on a bounded set. Conclude by using the continuous mapping theorem to obtain that $\alpha_{n,2M} \rightsquigarrow \alpha_{2M}$ and then Skorohod representation theorem (combined with the Lebesgue dominated convergence theorem) to get that $W_1(\alpha_{n,K}, \alpha_K) \to 0$. All this together, we obtain that

$$\limsup_n W_2(\alpha_n, \alpha) \leq 2\epsilon.$$

□

2.4 Sums of random variables in Hilbert spaces

In this section S is a Hilbert space. That is S is endowed with a real valued scalar product whose associated norm is $\|x\| = \langle x, x \rangle$.

Proposition 2.14. *Let $\mathcal{S}$ be a separable Hilbert space. Let $(X_i)_{i\geq 1}$ and $(Y_i)_{i\geq 1}$ be two sequence of independent centered random variables valued in S such that $E\|X_i\|^2 < \infty$ and $E\|Y_i\|^2 < \infty$ for any $i = 1, \ldots, n$. Then*

$$W_2^2\left(n^{-1/2}\sum_{i=1}^n X_i, n^{-1/2}\sum_{i=1}^n Y_i\right) \leq n^{-1}\sum_{i=1}^n W_2^2\left(X_i, Y_i\right).$$

Proof. Let $(\tilde{X}_i, \tilde{Y}_i)_{i\geq 1}$ be an independent sequence of random variables such that the probability measure of each $(\tilde{X}_i, \tilde{Y}_i)$ belongs to $\mathrm{opt}(X_i, Y_i)$. Then because the Wassertein distance $W_2^2(n^{-1/2}\sum_{i=1}^n X_i, n^{-1/2}\sum_{i=1}^n Y_i)$ is the infimum over $\mathrm{opt}(n^{-1/2}\sum_{i=1}^n X_i, n^{-1/2}\sum_{i=1}^n Y_i)$, we have

$$W_2^2\left(n^{-1/2}\sum_{i=1}^n X_i, n^{-1/2}\sum_{i=1}^n Y_i\right) \leq E\left[\|n^{-1/2}\sum_{i=1}^n (\tilde{X}_i - \tilde{Y}_i)\|^2\right].$$

$$E\left[\|n^{-1/2}\sum_{i=1}^n (\tilde{X}_i - \tilde{Y}_i)\|^2\right] = n^{-1}\sum_{i=1}^n E\left[\|\tilde{X}_i - \tilde{Y}_i\|^2\right].$$

The result follows. □

Proposition 2.15. *Let $\mathcal{S}$ be a separable Hilbert space. Let $(X_i)_{i\geq 1}$ and $(Y_i)_{i\geq 1}$ be two sequence of random variables valued in S such that $E\|X_i\| < \infty$ and $E\|Y_i\| < \infty$ for any $i = 1, \ldots, n$. Then*

$$W_1\left(n^{-1}\sum_{i=1}^n X_i, n^{-1}\sum_{i=1}^n Y_i\right) \leq n^{-1}\sum_{i=1}^n W_1^2\left(X_i, Y_i\right).$$

Proof. Similar to the previous proof using the triangle inequality. □

Proposition 2.16. *Let $\mathcal{S}$ be a separable Hilbert space and X and Y be random variables in $\mathcal{L}_2(\mathcal{S})$. We have*

$$W_2^2\left(X - EX, Y - EY\right) = W_2^2(X, Y) - \|EX - EY\|^2.$$

Proof. Write

$$E[\|\tilde{X} - EX - (\tilde{Y} - EY)\|^2] = E[\|\tilde{X} - \tilde{Y}\|^2] - \|EX - EY\|^2.$$

Taking $(\tilde{X}, \tilde{Y}) \in \text{opt}(X, Y)$ we obtain the sense $\leq$. Taking $(\tilde{X} - EX, \tilde{Y} - EY) \in \text{opt}(X - EX, Y - EY)$ we obtain the other sense. □

Proposition 2.17. *Let $\mathcal{S}$ be a separable Hilbert space and X and Y be random variables in $\mathcal{L}_2(\mathcal{S})$. We have*

$$W_2^2\left(X, \alpha X\right) = (1 - \alpha)^2 \mathbb{E}\|X\|^2.$$

Proof. Let $Y = \alpha X$. We have $\mathbb{E}[\|X - Y\|^2] = (1 - \alpha)^2 \mathbb{E}\|X\|^2$. This shows the inequality $\leq$. For the other sense, from Cauchy-Schwartz inequality, we get $\mathbb{E}[\langle X, Y\rangle] \leq \alpha \mathbb{E}\|X\|^2$, implying that $\mathbb{E}[\|X - Y\|^2] \geq (1 - \alpha)^2 \mathbb{E}\|X\|^2$.

□

2.5 Resampling schemes

2.5.1 Different resampling schemes

Efron's bootstrap, as introduced in Chapter 1, is based on generating new samples according to the empirical measure P_n defined as the uniform distribution over $\{X_1, \ldots, X_n\}$. We have motivated the use of P_n by the fact that it is potentially a good approximation of the underlying measure P. It seems then natural to think of different sampling schemes such as parametric sampling and smoothed sampling. The probability space is defined as follows.

Let $(\Omega \times \Omega^*, \mathcal{F} \otimes \mathcal{F}^*, \mathbb{P} \times \mathbb{P}^*)$ be a probability space. Let $(X_i)_{i \geq 1}$ be an independent and identically distributed sequence of random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Let Q_n^ω be a random measure. In the example below this measure will depend on $X_1, \ldots, X_n$. Let $(X_{n,i}^*)_{i=1,\ldots,n}$, $n \geq 1$, be an array of random variables defined on $(\Omega \times \Omega^*, \mathcal{F} \otimes \mathcal{F}^*, \mathbb{P} \times \mathbb{P}^*)$ such that

$$\begin{aligned} &\forall \omega \in \Omega,\ \forall n \geq 1,\ \forall f_1, \ldots, f_n \text{ positive functions} \\ &\quad \mathbb{E}^*[f_1(X_{n,1}^*) \times \ldots f_n(X_{n,n}^*)] = Q_n^\omega(f_1) \times \ldots Q_n^\omega(f_n). \end{aligned}$$

In words, each collection $(X_{n,1}^*, \ldots, X_{n,n}^*)$ are independent and identically distributed random variables with law Q_n^ω, conditionally on ω.

2.5.2 Efron's bootstrap

Here we revisit Efron's bootstrap using the Wasserstein distance. Let $(X_i)_{i \geq 1}$ be an independent and identically distributed sequence of random variables valued in the Hilbert space $S = \mathbb{R}^d$. Suppose that $E\|X_1\|^2 < \infty$ and define

$$\begin{aligned} \theta_0 &= EX_1, \\ \hat{\theta}_n &= n^{-1} \sum_{i=1}^n X_i. \end{aligned}$$

Let $(X^*_{i,n})_{i=1,\dots,n}$ be defined as before with $Q^\omega_n = P_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$. Suppose that $E\|X_1\| < \infty$ and define

$$\hat{\theta}^*_n = n^{-1}\sum_{i=1}^n X^*_{i,n}.$$

Proposition 2.18. *We have*

$$W_2^2\left(n^{1/2}(\hat{\theta}_n - \theta_0), n^{1/2}(\hat{\theta}^*_n - \hat{\theta}_n)\right) \leq W_2^2\left(P, P_n\right).$$

Proof. Applying Proposition 2.14 and 2.16 gives that

$$W_2^2\left(n^{1/2}(\hat{\theta}_n - \theta_0), n^{1/2}(\hat{\theta}^*_n - \hat{\theta}_n)\right) \leq W_2^2\left(X_1 - \theta_0, X^*_{1,n} - \hat{\theta}_n\right) = W_2^2\left(X_1, X^*_{1,n}\right) - \|\theta_0 - \hat{\theta}_n\|^2.$$

□

The following theorem provides the consistency of the empirical bootstrap.

Theorem 2.19. *We have, with probability* 1, $n^{1/2}(\hat{\theta}^*_n - \hat{\theta}_n)$ *weakly converges to a centered Gaussian random variable with finite variance* $\mathbb{E}[(X - EX)(X - EX)^T]$.

Proof. Use the triangle inequality to obtain that

$$\begin{aligned} W_2(n^{1/2}(\hat{\theta}^*_n - \hat{\theta}_n), \mathcal{N}(0, \sigma^2)) &\leq W_2(n^{1/2}(\hat{\theta}^*_n - \hat{\theta}_n), n^{1/2}(\hat{\theta}_n - \theta_0)) + W_2(n^{1/2}(\hat{\theta}_n - \theta_0), \mathcal{N}(0, \sigma^2)) \\ &\leq W_2\left(P, P_n\right) + W_2(n^{1/2}(\hat{\theta}_n - \theta_0), \mathcal{N}(0, \sigma^2)) \end{aligned}$$

We only have to show that each previous term goes to 0 almost surely. By the law of large number, we have that almost surely, for any $f \in C_b(S)$,

$$P_n(f) \to P(f).$$

This result is established in Varadarajan (1958), see also (Dudley, 2018b, Theorem 11.4.1). Moreover we have

$$P_n\|x\|^2 = n^{-1}\sum_{i=1}^n \|X_i\|^2,$$

which converges almost surely by the strong law of large number. In virtue of Proposition 2.13, we have that $W_2\left(P, P_n\right) \to 0$ almost surely. Let μ_n denote the distribution of $n^{1/2}(\hat{\theta}_n - \theta_0)$, the central limit theorem implies that

$$\mu_n(f) \to P(f).$$

Finally, $\mu_n(\|x\|^2) = E[\|n^{1/2}(\hat{\theta}_n - \theta_0)\|^2] = E[\|X_1 - \theta_0\|^2] = \sigma^2$.

□

2.5.3 The smoothed bootstrap

Suppose that $K : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is such that $\int K = 1$. Define the kernel density estimator, for any $x \in \mathbb{R}^d$,

$$f_n(x) = n^{-1}\sum_{i=1}^n K_{h_n}(x - X_i),$$

with $K_h(u) = K(u/h)/h^d$. Using the results in Devroye et al. (1983) we have that, for all f,

$$\int |f_n - f| \mathrm{d}\lambda \to 0, \qquad \text{almost surely.}$$

if and only if

$$nh_n^d \to \infty \qquad h_n \to 0.$$

From the previous result will follow the next proposition. Define $\hat{\theta}_n^*$ similarly as in the Efron's bootstrap replacing P_n by $f_n \mathrm{d}\lambda$. That is, at each $n \geq 1$, the bootstrap sample is generated according to the mixture distribution between the distributions $K_{h_n}(\cdot - X_i)$, $i = 1, \ldots, n$.

Proposition 2.20. *Let $(X_i)_{i\geq 1}$ be an independent and identically distributed sequence of random variables valued in $\mathbb{R}^d$. Suppose that $E\|X_1\|^2 < \infty$ and that $K : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is such that $\int K = 1$, $\int uK(u)\mathrm{d}u = 0$ and $\int \|u\|^2 K(u)\mathrm{d}u < \infty$. Then*

$$W_2^2\left(n^{1/2}(\hat{\theta}_n - \theta_0), n^{1/2}(\hat{\theta}_n^* - \hat{\theta}_n)\right) \to 0, \qquad \textit{almost surely}$$

Proof. Start by proving that

$$W_2^2\left(n^{1/2}(\hat{\theta}_n - \theta_0), n^{1/2}(\hat{\theta}_n^* - \hat{\theta}_n)\right) \leq W_2^2(f_n, P).$$

This can be done following the proof of Proposition 2.18. Then we can rely on Proposition 2.13 and show that weak convergence holds as well as the convergence of the squared norm. Let $g \in C_b(S)$, we have

$$\left| \int g(f_n - f) \right| \leq C \int |f_n - f|$$

A simple calculation for the squared norm gives

$$\begin{aligned}
\int \|x\|^2 f_n(x) &= n^{-1} \sum_{i=1}^n \int \|x\|^2 K_{h_n}(x - X_i)\, \mathrm{d}x \\
&= n^{-1} \sum_{i=1}^n \int \|X_i + h_n u\|^2 K(u)\, \mathrm{d}u \\
&= h_n^2 \int \|u\|^2 K(u)\mathrm{d}u + n^{-1} \sum_{i=1}^n \|X_i\|^2
\end{aligned}$$

This converges to $E\|X_1\|^2$.

□

Exercises

Exercise 2.1 (the parametric bootstrap)**.** *Let $(X_i)_{i\geq 1}$ be an independent and identically distributed sequence of random variables. Suppose that $X_1 \sim \mathcal{N}(\theta, 1)$ where $\theta \in \mathbb{R}$ is unknown.*

1. *Give the maximum likelihood estimate $\hat{\theta}_n$ of θ. Give the distribution of $\hat{\theta}_n$.*
2. *Define a bootstrap estimate $\hat{\theta}_n^*$ using a sampling according to $Q_n^\omega = \mathcal{N}(\hat{\theta}_n, 1)$.*
3. *Give the distribution of $\hat{\theta}_n^*$ conditionally on the original observations.*
4. *Explain how the the quantiles of $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ can be used to build confidence intervals and show the (asymptotic) validity of those confidence intervals.*

Chapter 3

Cross validation

Cross validation is an approach to assess the performance of a statistical procedure. It is broadly used to compare the accuracy of different methods and in the end to choose the best method among a collection of methods.

3.1 Hyper-parameter tuning

This section provides scenarios where many procedures need to be compared. Most of the statistical procedures or machine learning algorithms require to choose a tuning parameter among a range of possible values or sometimes to choose a model among many different possible models. The accuracy of the chosen candidate depends heavily on the decision process that is employed to make this choice. Let us now give some examples.

Example 3.1 (forward)**.** *In the context of regression with a large number of covariates, forward regression is an algorithm that incorporates at each iteration one additional covariate to build the regression function. When to stop is a parameter that needs to be tuned.*

Example 3.2 (penalized regression)**.** *Consider the following program*

$$\text{argmin}_{g\in\mathcal{G}} \sum_{i=1}^{n} (Y_i - g(X_i))^2 + \lambda pen(g).$$

When $\mathcal{G} = \{g(x) = \beta^T x \,:\, \beta \in \mathbb{R}^d\}$ *and* $pen(g) = \|\beta\|_1$ *(resp.* $pen(g) = \|\beta\|_2^2$*), we recover the Lasso (resp. the Ridge). When* $\mathcal{G} = \{g(x) = \sum_{i=1}^n K(x, X_i)\beta_i \,:\, \beta \in \mathbb{R}^n\}$ *and* $pen(g) = \sum_{1\le i,j\le n} K(X_i, X_j)\beta_i\beta_j$*, we recover support vector regression (including splines as an example). All these methods depend on the choice of* λ *which encodes for the importance of the penalization function with respect to the loss function.*

Example 3.3 (nearest neighbour)**.** *Let* $x \in \mathbb{R}^d$*. Denote by* $(X_i)_{i=1,\dots,k}$ *its* k*-nearest neighbour and by* $(Y_i)_{i=1,\dots,k}$ *the associated output variable valued in* $\{0, 1\}$*. The* k*-nearest neighbour estimate is a majority vote among the* $(Y_i)_{i=1,\dots,k}$*. It equals one if and only if there is a larger number of* 1 *than* 0 *in* $(Y_i)_{i=1,\dots,k}$*. Similar to the bandwidth choice, the performance of such an estimate heavily depends on the choice of* k*.*

Example 3.4 (model choice)**.** *Many models are often available to perform maximum likelihood estimation. For instance, each model can be determined by the set of covariates used in the regression function. Each model, say* $\mathcal{M}$*, is fitted by*

$$\text{argmax}_{f\in\mathcal{M}} \sum_{i=1}^{n} \log(f(X_i)).$$

Example 3.5 (kernel density estimation)**.** *The quantity*

$$n^{-1}\sum_{i=1}^{n} K_h(x - X_i)$$

is the kernel smoothing estimator of the density of X_1 (also known under the name Parzen–Rosenblatt estimate). The performance of such an estimate heavily depends on the choice of $h > 0$, called the bandwidth.

3.2 Statistical framework

Risk minimization. Let P be a probability measure and suppose that we seek to estimate a quantity f living in some space $\mathcal{F}$ that minimizes the expected loss

$$R(f) = E[\ell(Z, f)] = \int \ell(z, f)\,\mathrm{d}P(z),$$

where $Z \sim P$, $\ell : \mathbb{R}^d \times \mathcal{F} \to \mathbb{R}$ is a measurable function called the loss function and $R(f)$ is called the risk associated to f. For instance, in regression the loss function is usually the square loss $z = (y, x) \mapsto (y - f(x))^2$, in classification the 0-1 loss $z = (y, x) \mapsto \mathrm{I}_{y \neq f(x)}$ is often the one of interest, and in density estimation, it is the Kullback-Liebler loss $z \mapsto -\log(f(z))$ that is of interest.

Empirical Risk minimization. Let $(Z_i)_{i=1,\ldots,n}$ be an independent and identically distributed sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with common distribution P. The empirical estimate of f is defined as follows

$$\hat{f} = \mathrm{argmin}_{f \in \mathcal{F}}\, \hat{R}(f),$$
$$\hat{R}(f) = n^{-1}\sum_{i=1}^{n} \ell(Z_i, f).$$

As the model $\hat{f}$ has been "trained" using the sample $\{X_1, \ldots, X_n\}$, it is usually called the "training" sample. The $\hat{f}$ defined above is the minimizer of the empirical risk.

Model evaluation. The performance of $\hat{f}$ is measured through

$$R(\hat{f}) = \int \ell(z, \hat{f})\,\mathrm{d}P(z),$$

the problem being that this quantity is unknown, and, unfortunately, hard to estimate as we shall see in the next few lines. Model assessment or model evaluation means quantifying the performance of a method. Regarding the framework defined above, it correspond to estimating $R(\hat{f})$. Model selection is a generic term for selecting the best method among a collection of methods. The methods in competition can be different regression or classification rules, maximum likelihood estimates obtained from different models, linear regression performed over different set of variables (see the given examples). Model selection is conducted based on different model evaluation. This makes the question of estimating the risk of a predictor $R(\hat{f})$ central to statistics and machine learning.

3.3 The training sample is biased in estimating the risk

The somehow natural plug-in estimate $\hat{R}(\hat{f})$ is actually a poor estimate of $R(\hat{f})$. The problem comes from the learning step in which $\hat{f}$ has been computed minimizing $\hat{R}(f)$ over $\mathcal{F}$. Because for any f,

$$\hat{R}(\hat{f}) \leq \hat{R}(f)$$

it follows that

$$\mathbb{E}[\hat{R}(\hat{f})] \leq R(\hat{f})$$

meaning that it is most likely that $\hat{R}(\hat{f}) \leq R(\hat{f})$ which makes $\hat{R}(\hat{f})$ a poor estimate of the $R(\hat{f})$. Another way to underline the flaws of such an approach is to consider a sequence of nested models $\mathcal{F}_{k-1} \subset \mathcal{F}_k$, $k = 1, \ldots, K$. Suppose we are interested in finding the best model among the collection $\mathcal{F}_k$. Comparing the values of $\hat{R}(\hat{f}_k)$ where $\hat{f}_k$ minimizes $\hat{R}$ over $\mathcal{F}_k$ would always lead to the selection of the largest model $\mathcal{F}_K$.

3.3.1 The Mallow's

This heuristic is easily confirmed when considering *ordinary least squares*. The regression estimate satisfies the normal equation

$$\langle X, X\hat{\beta} - Y \rangle = 0$$

implying that $\langle X, X(\hat{\beta} - \beta) \rangle = \langle X, \epsilon \rangle$. It follows that

$$\begin{aligned} n\hat{R}(\hat{\beta}) &= \|Y - X\hat{\beta}\|_2^2 = \|\epsilon\|_2^2 + 2\langle \epsilon, X(\beta - \hat{\beta}\rangle + \|X(\beta - \hat{\beta})\|_2^2 \\ &= \|\epsilon\|_2^2 - \|X(\hat{\beta} - \beta)\|_2^2. \end{aligned}$$

It is well known that $\sqrt{n}(\hat{\beta} - \beta) \rightsquigarrow \mathcal{N}(0, \Sigma^{-1}\sigma^2)$. Hence,

$$\|X(\hat{\beta} - \beta)\|_2^2 \rightsquigarrow \|\mathcal{N}(0, I_d\sigma^2)\|^2 = \sigma^2\chi_d^2.$$

As a consequence, the asymptotic distribution of $\hat{R}(\hat{\beta})$ is represented as $\sigma^2(1 - \chi_d^2/n)$. Otherwise we have

$$\begin{aligned} R(\hat{\beta}) &= \sigma^2 + (\hat{\beta} - \beta)^T\Sigma(\hat{\beta} - \beta) \\ &\rightsquigarrow \sigma^2(1 + \chi_d^2/n). \end{aligned}$$

It results that $\hat{R}(\hat{\beta})$ is asymptotically biased in estimating $R(\hat{\beta})$. The bias is represented by

$$\begin{aligned} \hat{R}(\hat{\beta}) - R(\hat{\beta}) &= \{\|\epsilon\|_2^2/n - \sigma^2\} - (\hat{\beta} - \beta)^T\Sigma_n(\hat{\beta} - \beta) - (\hat{\beta} - \beta)^T\Sigma(\hat{\beta} - \beta) \\ &= \{\|\epsilon\|_2^2/n - \sigma^2\} - (\hat{\beta} - \beta)^T(\Sigma_n - \Sigma)(\hat{\beta} - \beta) - 2(\hat{\beta} - \beta)^T\Sigma(\hat{\beta} - \beta). \end{aligned}$$

The first term is centered. The second term is negligible. The last term converges to $-2\sigma^2\chi_d^2/n$ whose expectation is $-2d\sigma^2/n$. One of the version of the Mallows's statistic

$$C = \hat{R}(\hat{\beta}) + 2\hat{\sigma}^2\left(\frac{d}{n}\right),$$

where $\hat{\sigma}^2$ is an estimate of the residual variance. The Mallow's statistics is computed for each regression model $\mathcal{M}_k$, $k = 1, \ldots, K$ where typically each model is associated to a covariate subset. Then the selected model is the one having the smallest Mallow's statistic.

3.3.2 Akaike's information criterion

There are many approaches to reduce the bias induced when estimating the risk. The most popular one is based on the Akaike's information criterion (AIC) which is defined in the context of maximum likelihood estimation. Let

$$\hat{\theta}_n = \text{argmin}_\theta \sum_{i=1}^n \log(f_\theta(Z_i)).$$

Under regularity condition it holds that (Van der Vaart, 2000, Theorem 5.39)

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{\mathcal{I}^{-1}}{\sqrt{n}} \sum_{i=1}^n \partial \log(f_\theta(Z_i)) + o_P(1),$$

where $\mathcal{I} = E[\partial \log(f_\theta) \partial \log(f_\theta)^T] = -E[\partial^2 \log(f_\theta)]$ is called the Fisher information matrix. The previous identity follows from $\partial^2 \log(f_\theta) = \partial^2 f_\theta / f_\theta - (\partial f_\theta / f_\theta)^{\otimes 2}$ and $\int \partial^2 f_\theta = 0$. We have the following heuristic:

$$\begin{aligned}
&\sum_{i=1}^n \log(f_{\hat{\theta}_n}(Z_i)) - E[\log(f_{\hat{\theta}_n}(Z_1)] \\
&= \sum_{i=1}^n \log(f_\theta(Z_i)) - E[\log(f_\theta(Z_1)] + \sum_{i=1}^n \{\log(f_{\hat{\theta}_n}(Z_i)) - E[\log(f_{\hat{\theta}_n}(Z_1)] - (\log(f_\theta(Z_i)) - E[\log(f_\theta(Z_1)]\} \\
&= \sum_{i=1}^n \log(f_\theta(Z_i)) - E[\log(f_\theta(Z_1)] + \left(\sum_{i=1}^n \partial \log(f_{\tilde{\theta}_n}(Z_i)) - E[\partial \log(f_{\tilde{\theta}_n}(Z_1)] \right)^T (\hat{\theta}_n - \theta) \\
&\simeq \sum_{i=1}^n \log(f_\theta(Z_i)) - E[\log(f_\theta(Z_1)] + \left(\sum_{i=1}^n \partial \log(f_\theta(Z_i)) \right)^T (\hat{\theta}_n - \theta) \\
&\simeq \sum_{i=1}^n \log(f_\theta(Z_i)) - E[\log(f_\theta(Z_1)] + n^{-1} \left(\mathcal{I}^{-1/2} \sum_{i=1}^n \partial \log(f_\theta(Z_i)) \right)^T \left(\mathcal{I}^{-1/2} \sum_{i=1}^n \partial \log(f_\theta(Z_i)) \right).
\end{aligned}$$

The previous development shows that the risk is the sum of 2 terms. One which is centered and another which converges in distribution to a χ^2_p distribution where p stands for the dimension of θ. This leads to the definition of the AIC (Claeskens and Hjort, 2008)

$$\text{AIC} = 2 \sum_{i=1}^n \log(f_{\hat{\theta}_n}(Z_i)) - 2\text{length}(\theta).$$

3.4 Risk estimation with cross validation

3.4.1 The principle

The main goal here is to cancel the bias associated to the estimation of $R(\hat{f})$ by $\hat{R}(\hat{f})$. The main idea is to split the sample $\{1, \ldots, n\}$ in two independent subsamples S and S^c where S^c is the complement of S. The first sample, S^c is used to construct $\hat{f}$ that shall be denoted $\hat{f}_{S^c}$; the second sample, S of size n_S, is used to compute $\hat{R}(\cdot)$ which shall be denoted $\hat{R}_S(\cdot)$. That is

$$\begin{aligned}
\hat{f}_{S^c} &\in \text{argmin}_{f \in \mathcal{F}} \sum_{i \in S^c} \ell(Z_i, f) \\
\hat{R}_S(f) &= n_S^{-1} \sum_{i \in S} \ell(Z_i, f).
\end{aligned}$$

We have

$$\mathbb{E}[\hat{R}_S(\hat{f}_{S^c})|S^c] = n_S^{-1} \sum_{i \in S} \int \ell(z, \hat{f}_{S^c}) P(dz) = R(\hat{f}_{S^c}).$$

Definition 3.1. *The hold-out estimation of the risk associated to sample S is $\hat{R}_S(\hat{f}_{S^c})$. We have the following property.*

Proposition 3.1. *Suppose that for any $f \in \mathcal{F}$, $\mathbb{E}[\ell(Z_1, f)^2] < \infty$. Then*

$$\mathbb{E}[\hat{R}_S(\hat{f}_{S^c}) \mid S^c] = R(\hat{f}_{S^c})$$
$$\operatorname{var}(\hat{R}_S(\hat{f}_{S^c})|S^c) = n_S^{-1} \operatorname{var}(\ell(Z_1, \hat{f}_{S^c})|S^c).$$

Looking at the first equation above, we would like to have $|S^c|$ large enough so that $R(\hat{f}_{S^c})$ is close to $R(\hat{f}_n)$ (where $\hat{f}_n$ is the trained predictor on the whole sample). The the second and third equation, we would like to have n_S as large as possible. As $n = |S^c| + |S|$, there is a common bias-variance trade off that we need to accomplish. The next proposition is obtained as a corollary of the previous one.

Proposition 3.2. *Suppose that for any $f \in \mathcal{F}$, $\mathbb{E}[\ell(Z_1, f)^2] < \infty$. Then*

$$\mathbb{E}[\hat{R}_S(\hat{f}_{S^c})] = \mathbb{E}[R(\hat{f}_{S^c})]$$
$$\operatorname{var}(\hat{R}_S(\hat{f}_{S^c})) = n_S^{-1}\mathbb{E}[\operatorname{var}(\ell(Z_1, f_{S^C})|S^c)] + \operatorname{var}(R(\hat{f}_{S^c}))$$

Training, validation, and test. Evaluating the risk is actually conducted at different stages of most procedures. The common practice is to use three samples[1]: first, one computes many predictors (as many available models or tuning parameters) using a *training set*, second one compares them based on a *validation set*. At the end of this stage, a predictor has been selected. Third (if enough sample is available), one proceed to model assessment by computing the risk of the selected predictor. This is done using a third sample called the *test set*.

In what follows we focus on the problem of estimating the risk $R(\hat{f})$ for a given model $\mathcal{F}$. This task is the one of the second and third stages described before.

3.4.2 Hold-out consistency

In the following proposition, we establish under general assumptions that the hold-out is consistent in estimating the risk. We consider an asymptotic framework where the limiting predictor is denoted f^*. A class $\mathcal{G}$ is called P-Glivenko-Cantelli when

$$\sup_{g \in \mathcal{G}} |P_n(g) - P(g)| \to 0, \qquad \text{almost-surely.}$$

Proposition 3.3. *Suppose that there exists $f^\star$ such that $R(f^\star) \leq R(g)$*

$$\mathcal{L} = \{z \mapsto \ell(z, f) \,:\, f \in \mathcal{F}\} \quad \text{is } P\text{-Glivenko-Cantelli.}$$

Then, if $|S|, |S^c| \to \infty$, we have $\hat{R}_S(\hat{f}_{S^c}) \to R(f^\star)$ almost surely.

Proof. Write

$$0 \leq R(\hat{f}_{S^c}) - R(f^\star) = (R(\hat{f}_{S^c}) - \hat{R}_{S^c}(\hat{f}_{S^c})) + (\hat{R}_{S^c}(\hat{f}_{S^c}) - \hat{R}_{S^c}(f^\star)) + (\hat{R}_{S^c}(f^\star) - R(f^\star))$$
$$\leq 2 \sup_{\ell \in \mathcal{L}} |(P_{|S^c|} - P)(\ell)|.$$

[1] `https://en.wikipedia.org/wiki/Training,_validation,_and_test_sets`

Then decompose $\hat{R}_S(\hat{f}_{S^c}) - R(f^\star) = \hat{R}_S(\hat{f}_{S^c}) - R(\hat{f}_{S^c}) + R(\hat{f}_{S^c}) - R(f^\star)$ to obtain

$$|\hat{R}_S(\hat{f}_{S^c}) - R(f^\star)| \leq 2 \sup_{\ell \in \mathcal{L}} |(P_{|S^c|} - P)(\ell)| + \sup_{\ell \in \mathcal{L}} |(P_{|S|} - P)(\ell)|.$$

Both goes to 0 as $|S|, |S^c| \to \infty$. □

One popular way to obtain the Glivenko-Cantelli property is by using a notion of complexity called the *bracketing numbers*. Let $\mathcal{G} : \mathcal{S} \to \mathbb{R}$ be a space of functions endowed with the $L_1(P)$-norm. Note that $\mathcal{G}$ is a subset of $L_1(P)$. Given two function g_- and g_+ in $L_1(P)$, the set of functions $[g_-, g_+] = \{g : \mathcal{X} \to \mathbb{R} : g_- \leq g \leq g_+\}$ is called an ϵ-bracket whenever $P|g_- - g_+| \leq \epsilon$. The *bracketing number* $\mathcal{N}_{[]}(\epsilon, \mathcal{G}, L_1(P))$ is the minimal number of ϵ-brackets needed to cover $\mathcal{G}$.

Proposition 3.4. *Let $\mathcal{G} \subset L_1(P)$ be such that $\mathcal{N}_{[]}(\epsilon, \mathcal{G}, L_1(P)) < \infty$ for all $\epsilon > 0$. Then, $\mathcal{G}$ is Glivenko-Cantelli.*

Proof. Let $\epsilon > 0$. Let $[g_{k-}, g_{k+}]$, $k = 1, \ldots, N$, be a collection of ϵ-brackets that covers $\mathcal{F}$. Define $A_n(\epsilon) = \max_{k=1,\ldots,N}\{|(P_n - P)(g_{k+})| \vee |(P_n - P)(g_{k-})|\}$. It is easy to see that

$$(P_n - P)(g) \leq \max_{k=1,\ldots,N} |(P_n - P)(g_{k+})| + \epsilon \leq A_n(\epsilon) + \epsilon,$$

and that

$$-(P_n - P)(g) \leq \max_{k=1,\ldots,N} |(P_n - P)(g_{k-})| + \epsilon \leq A_n(\epsilon) + \epsilon.$$

Use that $|x| = \max(x, -x)$ to obtain that

$$\sup_{g \in \mathcal{G}} |(P_n - P)(g)| \leq A_n(\epsilon) + \epsilon$$

With probability 1, $A_n(\epsilon)$ goes to 0. Because $A_n(\epsilon)$ depends on ϵ, the event on which the convergence happens depends on ϵ as well. Some additional work is required. There exists B_ϵ such that $P(B_\epsilon) = 1$ and for all $\omega \in B_\epsilon$ we have $A_n(\epsilon) \to 0$. Let $B = \cap_{k=1}^{\infty} B_{1/k}$ and note that $P(B) = 1$. We have for any $\omega \in B$ and any $k \in \mathbb{N}$,

$$\limsup_n \sup_{g \in \mathcal{G}} |(P_n - P)(g)| \leq 1/k.$$

□

Many classes of functions have finite bracketing numbers. One of the most famous example is the indicators of cells over the real line, $\{z \mapsto \mathbb{I}\{z \leq t\} : t \in \mathbb{R}\}$. Lipschitz functions or monotonic functions also have finite bracketing numbers (Van Der Vaart and Wellner, 1996, Chapter 7).

3.4.3 Cross validation

Cross validation follows from combining, or rather aggregating, several hold-out estimate of the risk. The way we choose to build the hold-out estimates gives rise to many procedure. The first one which is probably the most popular is called K-fold cross validation is described bellow.

K-fold cross validation. Randomize the sample by applying a random permutation $(X_{\sigma(i)})_{i=1,\dots,n}$. This procedure aims at avoiding any dependence structure that would come from data processing (e.g., ordering). Then split this sample into K folds where each fold contains (nearly) the same number of observations. We have K samples, for $k = 1, \dots, K$,

$$S_k = (X_{(k-1)p+1}, \dots, X_{kp}),$$

with $p = n/K$. The cross validation risk is given by

$$\hat{R}_{Kcv} = (1/K) \sum_{k=1}^{K} \hat{R}_{S_k}(\hat{f}_{S_k^c}).$$

Leave-one-out. This corresponds to cross validation when $K = n$. The random permutation is not needed here. The estimated risk is given by

$$\hat{R}_{loo} = (1/n) \sum_{i=1}^{n} \hat{R}_{X_i}(\hat{f}_{S^{(-i)}}).$$

Leave-p-out. This is a generalization of the leave-one-out. Let $\mathcal{S}_p$ be the set of samples made of p observations. There are p among n such sample, the average of which is

$$\hat{R}_{lpo} = \binom{n}{p}^{-1} \sum_{S \in \mathcal{S}_p} \hat{R}_S(\hat{f}_{S^c}).$$

Monte Carlo cross validation. We use a Monte-Carlo additional step consisting in generating a certain number M of samples of size p, uniformly over the set $\mathcal{S}_p$. Denote by $S_1^*, \dots S_M^*$, the generated sample. Define

$$\hat{R}_{MC} = M^{-1} \sum_{m=1}^{M} \hat{R}_{S_m^*}(\hat{f}_{S_m^{*c}}).$$

Bootstrap cross validation. Generate M bootstrap samples by uniform draws among $\{Z_1, \dots, Z_n\}$. For each sample $m = 1, \dots, M$, define $S_m = \{X_{n,1}^*, \dots, X_{n,n}^*\}$. Note that contrary to the previous cross validation procedures, this is a sample of size n. Define S_m^c as the complement of S_m in $\{Z_1, \dots, Z_n\}$. Define

$$\hat{R}_{boot} = M^{-1} \sum_{m=1}^{M} \hat{R}_{S_m^*}(\hat{f}_{S_m^{*c}}).$$

For each bootstrap sample, $P^*(X_i \notin S_m) = P^*(X_{n,1} \neq X_i)^n = (1 - 1/n)^n \simeq e^{-1} = 0.37$.

Proposition 3.5. *Let S and $\tilde{S}$ be subset of $\{1, \dots, n\}$ such that $|S| = |\tilde{S}|$. The estimates $\hat{R}_S(\hat{f}(S^c))$ and $\hat{R}_{\tilde{S}}(\hat{f}(\tilde{S}^c))$ have the same distribution.*

Proof. Use an appropriate permutation of $\{1, \dots, n\}$. □

The following proposition draw a link between cross-validation and aggregation procedure.

Proposition 3.6. *Suppose that $p = n/K$.*

$$var(\hat{R}_{lpo}) \leq \mathrm{var}(\hat{R}_{Kcv}) \leq \mathrm{var}(\hat{R}_{ho}).$$

Proof. We follow Arlot (2018). From Jensen inequality, one has

$$\begin{aligned}\text{var}(\hat{R}_{Kcv}) &= \mathbb{E}\left[\left((1/K)\sum_{k=1}^{K}\{\hat{R}_{S_k}(\hat{f}_{S_k^c}) - \mathbb{E}[\hat{R}_{S_k}(\hat{f}_{S_k^c})]\}\right)^2\right] \\ &\leq \mathbb{E}\left[\left((1/K)\sum_{k=1}^{K}\{\hat{R}_{S_k}(\hat{f}_{S_k^c}) - \mathbb{E}[\hat{R}_{S_k}(\hat{f}_{S_k^c})]\}^2\right)\right] \\ &= (1/K)\sum_{k=1}^{K}\mathbb{E}\{\hat{R}_{S_k}(\hat{f}_{S_k^c}) - \mathbb{E}[\hat{R}_{S_k}(\hat{f}_{S_k^c})]\}^2.\end{aligned}$$

The use Proposition 3.5 to obtain the right-hand side inequality. For the other inequality, one should remark that, for any S of size n/K,

$$\hat{R}_{lpo} = (n!)^{-1}\sum_{\sigma\in\Sigma}\hat{R}_S(\hat{f}_{S^c})$$

Averaging over the S_k of the K folds, we obtain that

$$\hat{R}_{lpo} = (n!)^{-1}\sum_{\sigma\in\Sigma}\hat{R}_{Kcv},$$

and we can apply Jensen inequality. □

Bibliography

Ambrosio, L. and N. Gigli (2013). A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, pp. 1–155. Springer.

Arlot, S. (2018). Validation croisée.

Bickel, P. J., D. A. Freedman, et al. (1981). Some asymptotic theory for the bootstrap. *The annals of statistics 9*(6), 1196–1217.

Billingsley, P. (2013). *Convergence of probability measures*. John Wiley & Sons.

Claeskens, G. and N. L. Hjort (2008). Model selection and model averaging. Technical report, Cambridge University Press.

Devroye, L. et al. (1983). The equivalence of weak, strong and complete convergence in *l*_1 for kernel density estimates. *The Annals of Statistics 11*(3), 896–904.

Dudley, R. M. (2018a). *Real analysis and probability*. CRC Press.

Dudley, R. M. (2018b). *Real analysis and probability*. Chapman and Hall/CRC.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7*(1), 1–26.

Hall, P. (2013). *The bootstrap and Edgeworth expansion*. Springer Science & Business Media.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

Van Der Vaart, A. W. and J. A. Wellner (1996). Weak convergence. In *Weak convergence and empirical processes*, pp. 16–28. Springer.

Varadarajan, V. S. (1958). On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics (1933-1960) 19*(1/2), 23–26.

Villani, C. (2008). *Optimal transport: old and new*, Volume 338. Springer Science & Business Media.