Asymptotic Statistics

Lecture notes

Anne Sabourin, François Portier

February 1, 2022

Contents

Chapter 1

Definitions and first examples

Let $(X_i)_{i\in\mathbb{N}}$ be an independent and identically distibuted (i.i.d.) sequence of $\mathbb{R}^d$-valued random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The distribution of X_1 is denoted by P. The real composant of $X \in \mathbb{R}^d$ are denoted by $(X^{(1)}, \dots, X^{(d)})$. Statistical estimators are measurable trasformation of a finite number $n \in \mathbb{N}$ of observations $X_1, \dots, X_n$. They are denoted using "hat quantities" such as $\hat{\mu}_n$ or $\hat{\theta}_n$. The "true" quantities depending on the unknown distribution P of the data are denoted with a dot, for instance μ_0 or θ_0. The Euclidean norm is denoted by $\|\cdot\|$.

We start by introducing some vocabulary.

Definition 1. *An estimator $\hat{\theta}_n$ of θ_0,*

(i) is said to be weakly consistent whenever $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$.

(ii) is said to be strongly consistent whenever $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$.

(iii) is said to be asymptotically normal whenever there exists $v > 0$ such that $n^{1/2}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}(0, v)$.

Note that a weakly or strongly consistent estimator might have some bias and that an unbiased estimator might not be weakly or strongly consistent. An asymptotically normal estimator is always weakly consistent. A strongly consistent estimator is always weakly consistent. The following is a consequence of the continuous mapping theorem and shall be useful to prove the weak and strong consistency in some cases (see below).

Proposition 1. *Let $\hat{\theta}_n \in \Theta$ be weakly (resp. strongly) consistent estimating θ_0 and $h : \Theta \to \mathbb{R}^q$ be continuous at θ_0, then $h(\hat{\theta}_n)$ is weakly (resp. strongly) consistent estimating $h(\theta_0)$.*

From the law of large number or the central limit theorem, one can deduce some asymptotic properties of estimators of the form $n^{-1}\sum_{i=1}^n g(X_i)$, where g is a measurable function. In most statistical applications, estimators are not just empirical sums. The convergence properties are then obtained from tools introduced in Part I, such as, the Delta-Method, Sltutsky's lemma and the continuous mapping theorem as stated in Proposition 1. In the following, we give some examples. Each example is independent from the other.

Example 1 (variance of a bernoulli)**.** *Suppose that $X_1^{(1)} \sim \mathcal{B}(p_0)$. Define $\hat{\theta}_n = n^{-1}\sum_{i=1}^n X_i^{(1)}$. An estimator of the variance is*

$$\hat{\sigma}^2 = \hat{\theta}_n(1 - \hat{\theta}_n).$$

We have that

(i) $\hat{\theta}_n$ is strongly consistent estimating p_0.

(ii) $\hat{\sigma}_n^2$ is strongly consistent estimating $\sigma_0^2 = p_0(1 - p_0)$.

Example 2 (exponential law)**.** *Suppose* $X_1^{(1)} \sim \exp(\theta_0)$, $\theta_0 > 0$. *A natural "plug-in" estimator of* θ_0 *is then given by*

$$\hat{\theta}_n = \left(n^{-1} \sum_{i=1}^{n} X_i^{(1)} \right)^{-1}.$$

We have

(i) $\hat{\theta}_n \overset{a.s.}{\to} \theta_0$.

(ii) $n^{1/2}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}\left(0, \theta_0^2\right)$.

Example 3 (covariance estimation, the real case)**.** *The classical estimator of the covariance between* $X_1^{(1)}$ *and* $X_1^{(2)}$, $c_0 = \mathrm{cov}(X_1^{(1)}, X_1^{(2)})$, *is given by*

$$\hat{c}_n = n^{-1} \sum_{i=1}^{n} \left(X_i^{(1)} - \overline{X^{(1)}}^n \right) \left(X_i^{(2)} - \overline{X^{(2)}}^n \right),$$

with $\overline{X^{(1)}}^n = n^{-1} \sum_{i=1}^n X_i^{(1)}$ *and* $\overline{X^{(2)}}^n = n^{-1} \sum_{i=1}^n X_i^{(2)}$. *We have:*

(i) if $\mathbb{E}[|X_1^{(1)}|] < \infty$, $\mathbb{E}[|X_1^{(2)}|] < \infty$ *and* $\mathbb{E}[|X_1^{(1)} X_1^{(2)}|] < \infty$, *then* $\hat{c}_n \overset{a.s.}{\to} c_0$.

(ii) if $\mathbb{E}[X_1^{(1)2}] < \infty$, $\mathbb{E}[X_1^{(2)2}] < \infty$ *and* $\mathbb{E}[(X_1^{(1)} X_1^{(2)})^2] < \infty$, *then*

$$n^{1/2}(\hat{c}_n - c_0) \Rightarrow \mathcal{N}\left(0, \mathrm{var}((X_1^{(1)} - \mathbb{E}[X_1^{(1)}])(X_1^{(2)} - \mathbb{E}[X_1^{(2)}]))\right).$$

A usefull notation to deal with the law of random symmetric matrices vec operator. For any symmetric matrix A, $\mathrm{vec}(A)$ is the vector obtained by stacking the column of the lower triangular part of A, i.e., $\mathrm{vec}(A) = (A_{11}, A_{2,1}, \ldots, A_{d,1}, A_{2,2}, \ldots, A_{d,2}, \ldots, A_{d,d})^T$.

Example 4 (covariance estimation, the multidimensional case)**.** *The classical estimator of the covariance matrix* $\Sigma = \mathrm{cov}(X_1)$ *is given by*

$$\hat{\Sigma}_n = n^{-1} \sum_{i=1}^{n} (X_i - \overline{X}^n)(X_i - \overline{X}^n)^T.$$

We have

(i) if for all $k \in \{1, \ldots, d\}$, $\mathbb{E}[|X_1^{(k)}|^2] < \infty$, *then* $\hat{\Sigma}_n \overset{a.s.}{\to} \Sigma_0$.

(ii) if for all $k \in \{1, \ldots, d\}$, $\mathbb{E}[|X_1^{(k)}|^4] < \infty$, *then*

$$\mathrm{vec}\left(n^{1/2}(\hat{\Sigma}_n - \Sigma_0) \right) \Rightarrow \mathcal{N}\left(0, \mathrm{var}(\mathrm{vec}((X_1 - \mathbb{E}[X_1])(X_1 - \mathbb{E}[X_1])^T))\right).$$

To treat examples 3 and 4, it is better to use asymptotic decomposition with $o_{\mathbb{P}}$ and $O_{\mathbb{P}}$ consideration than the Delta-Method, which leads to some algebra. For any non-decreasing function $F : \mathbb{R} \to \mathbb{R}$, we define

$$F^-(u) = \inf\{x \in \mathbb{R} \,:\, F(x) \geq u\}.$$

Let Φ denote the cumulative distribution function of the standard normal distribution.

Example 5 (confidence interval)**.** *Suppose that $\hat{\mu}_n$ is an estimator of μ, such that $n^{1/2}(\hat{\mu}_n - \mu_0) \Rightarrow \mathcal{N}(0, \sigma^2)$ with σ_0 unknown but we have $\hat{\sigma} \xrightarrow{\mathbb{P}} \sigma_0$. The interval*

$$\hat{IC}_n = \left[\hat{\mu}_n - \hat{\sigma}\Phi^-(1-\alpha/2)/n^{1/2},\, \hat{\mu}_n - \hat{\sigma}\Phi^-(\alpha/2)/n^{1/2}\right],$$

is a $(1-\alpha)\%$ asymptotic confidence interval, i.e.,

$$\mathbb{P}(\mu_0 \in \hat{IC}_n) \to 1 - \alpha.$$

Example 6 (estimation under constraints)**.** *Suppose that the true parameter $\theta_0 \in \mathbb{R}^q$ is living on a closed curve defined as the subset $\Theta = \{\theta \in \mathbb{R}^q : g(\theta) = 0\}$ where g is continuous and that $\hat{\theta}_n$ is an estimator of θ_0. The constrained estimator is then defined by*

$$\hat{\theta}_{n,c} = \operatorname*{arg\,min}_{\theta \in \Theta} (\theta - \hat{\theta}_n)^T \Gamma (\theta - \hat{\theta}_n),$$

where Γ is a positive definite matrix. Whenever $\hat{\theta}_n \xrightarrow{a.s.} \theta$, we have $\hat{\theta}_{n,c} \xrightarrow{a.s.} \theta$.

Example 7 (ordinary least squares)**.** *Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be an i.i.d. sequence of random vectors. Each pair (X_i, Y_i) is valued in $\mathbb{R}^p \times \mathbb{R}$. The ordinary least squares estimator is given by*

$$\hat{\theta}_n = \hat{G}_n^+ \left(n^{-1} \sum_{i=1}^n X_i Y_i \right), \qquad n \in \mathbb{N}^*,$$

with $\hat{G}_n = n^{-1} \sum_{i=1}^n X_i X_i^T$ and A^+ denote the Moore-Penrose inverse of A. Suppose that $\mathbb{E}[\|X_1\|^2] < \infty$ and $G = \mathbb{E}[X_1 X_1^T]$ is invertible. Then it holds that

$$n^{1/2}(\hat{\theta}_n - \theta^*) \rightsquigarrow \mathcal{N}(0, \sigma^2 G^{-1}),$$

where $\theta^* = (\mathbb{E}[X_1 X_1^T])^{-1} \mathbb{E}[X_1 Y_1]$.

Chapter 2

M and Z-estimators

Many important statistical estimators are defined as maximizers of empirical sums. Examples include moment estimators, maximum likelihood estimators, least-square estimators, estimators under constraints. In the following the parameter space is denoted by $\Theta \subset \mathbb{R}^q$.

2.1 Definition

Definition 2 (M-estimator)**.** *An estimator $\hat{\theta}_n$ is called an M-estimator whenever, almost surely, $\hat{\theta}_n \in \arg\min_{\theta\in\Theta} M_n(\theta)$, where $M_n : \Theta \to \mathbb{R}$, $\Theta \subset \mathbb{R}^q$.*

In most statistical applications, M_n is given by,

$$M_n(\theta) = \sum_{i=1}^{n} \rho(X_i, \theta), \qquad \text{for every } \theta \in \Theta,$$

where for every $\theta \in \Theta$, $\rho(\cdot, \theta) : \mathbb{R}^d \to \mathbb{R}$ is a measurable function. Least-squares estimators, least-absolute deviations and the Huber regression estimator are such examples. However, the previous general definition is suitable to treat in the mean time the consistency of M-estimator and Z-estimator, that we introduce below.

Definition 3 (Z-estimator)**.** *An estimator $\hat{\theta}_n$ is called a Z-estimator whenever there exists a (measurable) function $\Psi_n : \Theta \to \mathbb{R}^q$, $\Theta \subset \mathbb{R}^q$, such that, almost surely, $\Psi_n(\hat{\theta}_n) = 0$.*

As for M-estimator, a common situation is when

$$\Psi_n(\theta) = \sum_{i=1}^{n} \psi(X_i, \hat{\theta}_n), \qquad \text{for every } \theta \in \Theta, \tag{2.1}$$

where for every $\theta \in \Theta$, $\psi(\cdot, \theta) : \mathbb{R}^d \to \mathbb{R}^{q'}$ is a measurable function.

Proposition 2. *It holds that*

(i) A Z-estimator is always an M-estimator.

(ii) An M-estimator is a Z-estimator whenever $\theta \to M_n(\theta)$ is a continuously differentiable function on Θ and $\hat{\theta}_n$ is an interior point.

Proof. Denote by $B(\theta, \epsilon)$ the open ball centred at θ with radius ϵ. For the first point, take $M_n(\theta) = \|\Psi_n(\theta)\|$. For the second point, there exists $\epsilon > 0$ such that $B(\hat{\theta}_n, \epsilon) \subset \Theta$. As $\hat{\theta}_n$ is a local minimum of a $\mathcal{C}^1$ function and $\hat{\theta}_n$ is contained in an open ball, we necessarily have $\nabla M_n(\hat{\theta}_n) = 0$. Indeed, let $u \in \mathbb{R}^q$, the function $t \mapsto M_n(\hat{\theta}_n + tu)$, defined on $\{t \in \mathbb{R} : \hat{\theta}_n + tu \in B(\hat{\theta}_n, \epsilon)\}$ is $\mathcal{C}^1$. The point $t = 0$ is a local extremum, hence $u^T \nabla M_n(\hat{\theta}_n) = 0$. Take $u = e_k$, $k = 1, \ldots, q$. □

2.2 Consistency

From Proposition 2, it suffices to otain conditions for the consistency of M-estimators. The conditions for Z-estimators will follow directly.

Proposition 3. *Suppose that $\hat{\theta}_n$ is an M-estimator and that for every $\epsilon > 0$,*

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathbb{P}} 0,$$
$$\inf_{\theta \in \Theta \setminus B(\theta_0, \epsilon)} M(\theta) > M(\theta_0),$$

then $\hat{\theta}_n$ is weakly consistent in estimating θ_0.

Proof. By assumption, we have $0 \leq M(\hat{\theta}_n) - M(\theta_0)$ and $M_n(\hat{\theta}_n) - M_n(\theta_0) \leq 0$. It follows,

$$\begin{aligned}
0 &\leq M(\hat{\theta}_n) - M(\theta_0) \\
&= M(\hat{\theta}_n) - M_n(\hat{\theta}_n) + M_n(\hat{\theta}_n) - M_n(\theta_0) + M_n(\theta_0) - M(\theta_0) \\
&\leq M(\hat{\theta}_n) - M_n(\hat{\theta}_n) + M_n(\theta_0) - M(\theta_0) \\
&\leq \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathbb{P}} 0.
\end{aligned}$$

Let $\epsilon > 0$, by the second assumption, there exists $\eta_\epsilon = \inf_{\theta \in \Theta \setminus B(\theta_0, \epsilon)} M(\theta) - M(\theta_0) > 0$ such that

$$|\theta - \theta_0| > \epsilon \quad \Rightarrow \quad M(\theta) - M(\theta_0) > \eta_\epsilon.$$

Consequently,

$$\mathbb{P}(|\hat{\theta}_n - \theta_0| > \epsilon) \leq \mathbb{P}(M(\hat{\theta}_n) - M(\theta_0) > \eta_\epsilon).$$

The previous goes to 0, as demonstrated before. □

A similar propsition is available for Z-estiamtors.

Proposition 4. *Suppose that $\hat{\theta}_n$ is a Z-estimator and that for every $\epsilon > 0$,*

$$\sup_{\theta \in \Theta} \|\Psi_n(\theta) - \Psi(\theta)\| \xrightarrow{\mathbb{P}} 0,$$
$$\inf_{\theta \in \Theta \setminus B(\theta_0, \epsilon)} \|\Psi(\theta)\| > 0 = \|\Psi(\theta_0)\|,$$

then $\hat{\theta}_n$ is weakly consistent in estimating θ_0.

Proof. By proposition 2, $\hat{\theta}_n$ is also an M-estimator with objective function $\|\Psi_n\|$. Because $\|\|a| - |b|\| \leq \|a - b\|$ the first condition of Proposition 3 is verified. The second condition is straightforwardly verified. □

The two previous propositions can be stated with respect to the almost sure convergence (rather that in probability), leading to the strong consistency (rather than the weak consistency). Indeed, we first obtain that $M(\hat{\theta}_n) - M(\theta_0) \xrightarrow{\text{a.s.}} 0$. Then define $A_n(\epsilon) = \{|\hat{\theta}_n - \theta_0| > \epsilon\}$ and $B_n(\epsilon) = \{M(\hat{\theta}_n) - M(\theta_0) > \eta_\epsilon\}$. In the proof of Proposition 3, we obtained that $A_n(\epsilon) \subset B_n(\epsilon)$. Then $\cup_{N \geq 1} \cap_{n \geq N} A_n(\epsilon) \subset \cup_{N \geq 1} \cap_{n \geq N} B_n(\epsilon)$ and conclude using that by almost sure convergence, $\mathbb{P}(\cup_{N \geq 1} \cap_{n \geq N} B_n(\epsilon)) = 0$ for each $\epsilon > 0$.

Propositions 3 and 4 are linked with two types of conditions. The first one is called the uniform law of large number. The second one is non-stochastic and corresponds to an identifiability condition. The following lemmas provide conditions to ensure the validity of these two conditions.

Lemma 5 (strong uniform law of large numbers)**.** *Suppose that Θ is compact and that $\mathbb{E}[|\rho(X_1,\theta)|] < \infty$ for all $\theta \in \Theta$. If moreover there exists $r : \mathcal{X} \to \mathbb{R}_+$ such that $\mathbb{E}r(X_1) < \infty$ and*

$$\forall x \in \mathcal{X},\, \forall (\theta,\theta') \in \Theta \times \Theta, \quad |\rho(x,\theta) - \rho(x,\theta')| \leq r(x)|\theta - \theta'|,$$

then

$$\sup_{\theta\in\Theta} \left| n^{-1} \sum_{i=1}^{n} (\rho(X_i,\theta) - \mathbb{E}[\rho(X_1,\theta)]) \right| \overset{a.s.}{\to} 0.$$

Proof. Let $\epsilon > 0$. As Θ is a compact set, there exists $\theta_1, \ldots, \theta_K \in \Theta$ such that $\Theta \subset \bigcup_{k=1}^{K} B(\theta_k, \epsilon)$. This is because $\Theta \subset \bigcup_{\theta\in\Theta} B(\theta,\epsilon)$, where the union is taken over all the elements in Θ. By definition, any open covers of Θ has a finite subcover. Consequently, for $\theta \in \Theta$, $\theta \in B(\theta_j, \epsilon)$, then

$$\begin{aligned}
&\left| n^{-1} \sum_{i=1}^{n} (\rho(X_i,\theta) - \mathbb{E}[\rho(X_1,\theta)]) \right| \\
&\leq \left| n^{-1} \sum_{i=1}^{n} (\rho(X_i,\theta) - \rho(X_i,\theta_j)) \right| + \left| n^{-1} \sum_{i=1}^{n} (\rho(X_i,\theta_j) - \mathbb{E}[\rho(X_1,\theta_j)]) \right| + |\mathbb{E}[\rho(X_1,\theta_j) - \rho(X_1,\theta)]|.
\end{aligned}$$

Using the Lipschitz condition, we find that

$$\begin{aligned}
&\left| n^{-1} \sum_{i=1}^{n} (\rho(X_i,\theta) - \mathbb{E}[\rho(X_1,\theta)]) \right| \\
&\leq n^{-1} \sum_{i=1}^{n} (r(X_i) + \mathbb{E}[r(X_1)])\epsilon + \left| n^{-1} \sum_{i=1}^{n} (\rho(X_i,\theta_j) - \mathbb{E}[\rho(X_1,\theta_j)]) \right|.
\end{aligned}$$

Hence, using $|v_j| \leq \sum_{k=1}^{K} |v_k|$, we get

$$\begin{aligned}
&\left| n^{-1} \sum_{i=1}^{n} (\rho(X_i,\theta) - \mathbb{E}[\rho(X_1,\theta)]) \right| \\
&\leq n^{-1} \sum_{i=1}^{n} (r(X_i) + \mathbb{E}[r(X_1)])\epsilon + \sum_{k=1}^{K} \left| n^{-1} \sum_{i=1}^{n} (\rho(X_i,\theta_k) - \mathbb{E}[\rho(X_1,\theta_k)]) \right|.
\end{aligned}$$

Taking the supremum on the left side, we get

$$\begin{aligned}
&\sup_{\theta\in\Theta} \left| n^{-1} \sum_{i=1}^{n} (\rho(X_i,\theta) - \mathbb{E}[\rho(X_1,\theta)]) \right| \\
&\leq n^{-1} \sum_{i=1}^{n} (r(X_i) + \mathbb{E}[r(X_1)])\epsilon + \sum_{k=1}^{K} \left| n^{-1} \sum_{i=1}^{n} (\rho(X_i,\theta_k) - \mathbb{E}[\rho(X_1,\theta_k)]) \right|.
\end{aligned}$$

By the strong law of large number, we have that

$$\begin{aligned}
&n^{-1} \sum_{i=1}^{n} r(X_i) \xrightarrow{\text{a.s.}} \mathbb{E}[r(X_1)], \\
&\sum_{k=1}^{K} \left| n^{-1} \sum_{i=1}^{n} (\rho(X_i,\theta_k) - \mathbb{E}[\rho(X_1,\theta_k)]) \right| \xrightarrow{\text{a.s.}} 0,
\end{aligned}$$

as a result, there exists $A_\epsilon \in \mathcal{F}$ such that $\mathbb{P}(A_\epsilon^c) = 0$ and for all $\omega \in A_\epsilon$,

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta} \left| n^{-1} \sum_{i=1}^{n} (\rho(X_i, \theta) - \mathbb{E}[\rho(X_1, \theta)]) \right| \leq 2\mathbb{E}[r(X_1)]\epsilon.$$

Now define $A = \bigcap_{k=1}^{\infty} A_{1/k}$. We have $\mathbb{P}(A^c) = 0$. And for $\omega \in A$, it holds for every $\epsilon > 0$,

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta} \left| n^{-1} \sum_{i=1}^{n} (\rho(X_i, \theta) - \mathbb{E}[\rho(X_1, \theta)]) \right| \leq 2\mathbb{E}[r(X_1)]\epsilon,$$

meaning that the limit is 0. □

Remark 6. In Lemma 5, the Lipschitz condition can be replaced by: there exists $C > 0$ such that

$$\forall(\theta, \theta') \in \Theta \times \Theta, \quad \mathbb{E}[|\rho(X_1, \theta) - \rho(X_1, \theta')|] \leq C|\theta - \theta'|.$$

We now give sufficient conditions guaranteeing the identifiability condition of Proposition 3 (as well as Proposition 4).

Lemma 7. *Suppose that (i) $\Theta \subset \mathbb{R}^q$ is compact, (ii) M is continuous on Θ, and (iii) M is uniquely minimized at θ_0. Then for every $\epsilon > 0$,*

$$\inf_{\theta\in\Theta\setminus B(\theta_0,\epsilon)} M(\theta) > M(\theta_0).$$

Proof. Let $\epsilon > 0$. By (i), $\Theta \cap B(\theta_0, \epsilon)^c$ is closed and bounded, i.e., compact. By (ii), there exists $\theta_\epsilon \in \Theta \cap B(\theta_0, \epsilon)^c$ such that $\inf_{\theta\in\Theta\setminus B(\theta_0,\epsilon)} M(\theta) = M(\theta_\epsilon)$. By (iii), $M(\theta_\epsilon) < M(\theta_0)$.

□

2.3 Application to the consistency of Moment estimators

. We start by giving an example of such an estimator. Suppose that $(X_1, \ldots, X_n)$ are iid with comon distribution $\Gamma(\alpha_0, \beta_0)$. Consequently, $\mathbb{E}[X_1] = \alpha_0/\beta_0$ and $\mathbb{E}[X_1^2] = \alpha_0(\alpha_0 + 1)/\beta_0^2$. To estimate α_0, β_0, one can solve the system

$$\frac{\alpha}{\beta} = n^{-1} \sum_{i=1}^{n} X_i$$
$$\frac{\alpha(1+\alpha)}{\beta^2} = n^{-1} \sum_{i=1}^{n} X_i^2$$

Let $g : \mathbb{R} \to \mathbb{R}^2$, such that $g(x) = (x, x^2)^T$. Then the system is $\mathbb{E}_{(\alpha,\beta)}[g] = n^{-1} \sum_{i=1}^{n} g(X_i)$, $\mathbb{E}_{(\alpha,\beta)}$ stands for the expectation with respect to the distribution $\Gamma(\alpha, \beta)$, i.e., meaning that theoretical moments and empirical moments coincide. Unfortunately and often when too much equations are considered, the system could have no solution. Hence we preferably define the estimator as

$$(\hat{\alpha}_n, \hat{\beta}_n) \in \operatorname*{arg\,min}_{(\alpha,\beta)\in\mathbb{R}_+^2} \|n^{-1} \sum_{i=1}^{n} g(X_i) - \mathbb{E}_{(\alpha,\beta)}[g]\|,$$

where $\|\cdot\|$ is a norm on $\mathbb{R}^2$. More generally, the estimator of moments associated to a family of distributions $\mathcal{P} = \{P_\theta\,, \theta \in \Theta\}$ and "moments" given by $g : \mathcal{X} \to \mathbb{R}^q$, $q \in \mathbb{N}^*$ and $\mathbb{E}\|g\| < \infty$, is defined by

$$\hat{\theta}_n \in \operatorname*{arg\,min}_{\theta\in\Theta} \|n^{-1} \sum_{i=1}^{n} g(X_i) - \mathbb{E}_\theta[g]\|. \tag{2.2}$$

The following proposition provides conditions for the consistency.

Proposition 8. *Suppose that $\mathbb{E}\|g\| < \infty$ and $\hat{\theta}_n$ verifies (2.2). If moreover the parameter space Θ is compact, and $\theta \mapsto \mathbb{E}_\theta[g]$ is injective and continuous, then $\hat{\theta}_n$ is strongly consistent.*

Proof. We apply Theorem 3 with $M_n(\theta) = \|n^{-1}\sum_{i=1}^n g(X_i) - \mathbb{E}_\theta[g]\|$ and $M(\theta) = \|\mathbb{E}[g] - \mathbb{E}_\theta[g]\|$. First, we have $|M_n(\theta) - M(\theta)| \leq \|n^{-1}\sum_{i=1}^n g(X_i) - \mathbb{E}[g]\|$ which is going to 0, almost surely, by the strong law of large number. The function M is continuous on a compact, it attains its minimum at θ_0, uniquely because the function is injective. We can apply Lemma 7.

□

2.4 Asymptotic normality of Z-estimators

We now turn our attention to the asymptotic normality of Z-estimators. As a consequence of Proposition 2, this case includes M-estimators associated with a smooth objective functions, e.g., ordinary least squares and some maximum likelihood estimators. For other M-estimators, based on nonsmooth objective functions, we refer to van der Vaart and Wellner (1996), Part 2.

We consider the case when the function Ψ_n used to defined the Z-estimator is an empirical sum as in (2.1) $\Psi_n(\theta) = \sum_{i=1}^n \psi(X_i, \hat{\theta}_n)$. We assume that θ_n is weakly consistent, that is $\theta_n \to \theta_0$ in probability. This condition is satisfied in particular under the conditions of Proposition 4, see also and Lemmas 5 and 7.

For a function $g : \mathbb{R}^q \to \mathbb{R}$ which is two-times differentiable at x we write

$$\nabla_x g(x) = \left(\frac{\partial g}{\partial x_1}(x), \frac{\partial g}{\partial x_2}(x), \ldots, \frac{\partial g}{\partial x_q}(x)\right)^T, \qquad \nabla_x^2 g(x) = \left(\frac{\partial^2 g}{\partial x_k \partial x_l}(x)\right)_{1\leq k,l\leq q}$$

Theorem 9. *[Asymptotic normality of Z-estimators, classical conditions] Let $\hat{\theta}_n \in \Theta \subset \mathbb{R}^q$ be a Z-estimator with $\Psi_n(\theta) = \sum_{i=1}^n \psi(X_i, \theta)$, where $\psi(x, \cdot)$ is a function $\mathbb{R}^q \to \mathbb{R}$. Suppose that the following assumptions are satisfied.*

1. *$\hat{\theta}_n \xrightarrow{P} \theta_0$ (*i.e. *$\hat{\theta}_n$ is weakly consistent)*
2. *$\mathbb{E}[\|\psi(X_1, \theta_0)\|^2] < \infty$,*
3. *There exists an open neighbourhood $v(\theta_0)$ of θ_0 such that, for every $x \in \mathcal{X}$, the function $\theta \mapsto \psi(x, \theta)$ is two-times continuously differentiable on $v(\theta_0)$. In addition*

$$\mathbb{E}\Big[\sup_{\theta \in v(\theta_0)} \|\nabla_\theta^2 \psi_k(X_1, \theta)\|\Big] < \infty$$

   for all $k = 1, \ldots, q$.
4. *Define the Jacobian matrix $\Phi(x, \theta) = (\nabla_\theta \psi_1(x, \theta), \ldots, \nabla_\theta \psi_q(x, \theta))^T$. Let $\Phi_k(x, \theta)$ denote k^{th} row of $\Phi(x, \theta)$, that is $\Phi_k(x, \theta) = \nabla_\theta \psi_k(x, \theta)$. Then*

$$\mathbb{E}[\|\Phi_k(X_1, \theta_0)\|] < \infty,$$

   and the matrix $\Phi(\theta_0) := \mathbb{E}[\Phi(X_1, \theta_0)] \in \mathbb{R}^{q\times q}$ is invertible.

Then, it holds that

$$\hat{\theta}_n - \theta_0 = -\Phi(\theta_0)^{-1}\left(n^{-1}\sum_{i=1}^n \psi(X_i, \theta_0)\right) + o_{\mathbb{P}}(n^{-1/2}).$$

In particular, we have

$$n^{1/2}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}(0, \Phi(\theta_0)^{-1} \operatorname{var}(\psi(X_1, \theta_0))(\Phi(\theta_0)^{-1})^\top).$$

Proof. Let us first prove a Taylor-Lagrange decomposition. A function $g : \mathbb{R}^q \to \mathbb{R}$, is 2-times continuously differentiable if the partial derivatives with order ≤ 2 exist and are continuous. Let B be an open ball, $a \in B$ and $b \in B$, and define $f(t) = g(a+t(b-a))$, for every t such that $a+t(b-a) \in B$. The function $g : \mathbb{R} \to \mathbb{R}$ is 2-times continuously differentiable on $I = \{t \in \mathbb{R} : a+t(b-a) \in B\}$ with first derivative $(b-a)^T \nabla g(a+t(b-a))$ and second derivative $(b-a)^T \nabla^2 g(a+t(b-a))(b-a)$. By the Taylor-Lagrange formula of order 2 with exact remainder term gives that there exists $c \in [0,1]$ such that $f(1) = f(0) + f'(0)(1-0) + f''(c)(1-0)^2/2$, i.e.,

$$g(b) = g(a) + \nabla g(a)^T (b-a) + \frac{1}{2}(b-a)^T \nabla^2 g(a+c(b-a))(b-a),$$

taking $u = a + c(b-a)$, we have that there exists u in the segment line between a and b such that

$$g(b) = g(a) + \nabla g(a)^T (b-a) + \frac{1}{2}(b-a)^T \nabla^2 g(u)(b-a).$$

By assumption there exists $B(\theta_0, \eta) \subset v(\theta_0)$ such that, for all $x \in \mathcal{X}$, $\theta \mapsto \psi(x, \theta)$ is two times continuously differentiable (Assumption 3 from the statement). Let $\mathcal{E}_n$ denote the event that $\hat{\theta}_n \in B(\theta_0, \eta)$ and let $Z_n = \mathbb{1}\{\mathcal{E}_n\}$ be the corresponding indicator random variable (a Bernoulli). Notice already that using Assumption 2 we have $Z_n \xrightarrow{P} 1$ thus $Z_n = 1 + o_P(1)$. With these notations we may write

$$\hat{\theta}_n = Z_n \hat{\theta}_{1,n} + (1 - Z_n)\hat{\theta}_{2,n}$$

where $\hat{\theta}_{1,n}$ is a random variable which belongs to $B(\theta_0, \eta)$ almost surely. As an example one may take $\hat{\theta}_{1,n} = \hat{\theta}_n Z_n + \theta_0(1 - Z_n)$ and $\hat{\theta}_{2,n} = \hat{\theta}_n(1 - Z_n) + \theta_0(Z_n)$.

Fix $k \leq q$. Applying the previous Taylor-Lagrange decomposition to the function $\Psi_{n,k}$, which stands for the k-th coordinate of Ψ_n, with b equal to $\hat{\theta}_{1,n}$ and a equal to θ_0, we have that there exists $\tilde{\theta}_{n,k}$ in the segment line between $\hat{\theta}_{1,n}$ and θ_0 such that

$$\Psi_{n,k}(\hat{\theta}_{1,n}) = \Psi_{n,k}(\theta_0) + \nabla_\theta \Psi_{n,k}(\theta_0)^T (\hat{\theta}_{1,n} - \theta_0) + (\hat{\theta}_{1,n} - \theta_0)^T \nabla^2_\theta \Psi_{n,k}(\tilde{\theta}_{n,k})(\hat{\theta}_{1,n} - \theta_0).$$

Now on the event $\mathcal{E}_n$ we have that $\hat{\theta}_{1,n} = \hat{\theta}_n$ so that we also have

$$Z_n \Psi_{n,k}(\hat{\theta}_n) = Z_n \left[\Psi_{n,k}(\theta_0) + \nabla_\theta \Psi_{n,k}(\theta_0)^T (\hat{\theta}_n - \theta_0) + (\hat{\theta}_n - \theta_0)^T \nabla^2_\theta \Psi_{n,k}(\tilde{\theta}_{n,k})(\hat{\theta}_n - \theta_0) \right].$$

Using that $\Psi_{n,k}(\hat{\theta}_n) = 0$ we find

$$Z_n \left\{ \nabla_\theta \Psi_{n,k}(\theta_0)^T + (\hat{\theta}_n - \theta_0)^T \nabla^2_\theta \Psi_{n,k}(\tilde{\theta}_{n,k}) \right\} (\hat{\theta}_n - \theta_0) = -Z_n \Psi_{n,k}(\theta_0). \tag{2.3}$$

Now we prove that $\nabla^2_\theta \Psi_{n,k}(\tilde{\theta}_{n,k}) = O_P(1)$. We have by the triangular inequality, and because $\tilde{\theta}_{n,k} \in B(\theta_0, \eta)$,

$$\begin{aligned}
\|\nabla^2_\theta \Psi_{n,k}(\tilde{\theta}_{n,k})\| &\leq n^{-1} \sum_{i=1}^n \left\| \nabla^2_\theta \psi_k(X_i, \tilde{\theta}_{n,k}) \mathbb{1}\left\{ \tilde{\theta}_{n,k} \in B(\theta_0, \eta) \right\} \right\| \\
&\leq n^{-1} \sum_{i=1}^n \sup_{\theta \in B(\theta_0, \eta)} \|\nabla^2_\theta \psi_k(X_i, \theta)\|.
\end{aligned}$$

It follows that for all n,

$$\mathbb{E}\left[\|\nabla^2_\theta \Psi_{n,k}(\tilde{\theta}_{n,k})\| \right] \leq \mathbb{E}\left[\sup_{\theta \in B(\theta_0, \eta)} \|\nabla^2_\theta \psi_k(X_1, \theta)\| \right],$$

where the right-hand side of the last inequality is finite and independent from n (Assumption 3). This implies (using Markov inequality) that the sequence $\nabla^2_\theta \Psi_{n,k}(\tilde{\theta}_{n,k})$ is tight, *i.e.* it is a $O_P(1)$.

Getting back to (2.3), we have that each coordinate of $(\hat{\theta}_n - \theta_0)^T \nabla^2_\theta \Psi_{n,k}(\tilde{\theta}_{n,k})$ is a finite sum of $o_P(1)O_P(1)$ hence it is a $o_P(1)$. Hence, (2.3) becomes

$$Z_n \left\{ \nabla_\theta \Psi_{n,k}(\theta_0)^T + o_P(1) \right\} (\hat{\theta}_n - \theta_0) = -\Psi_{n,k}(\theta_0) Z_n. \tag{2.4}$$

Now we show that $\nabla_\theta \Psi_{n,k}(\theta_0)^T \xrightarrow{P} \Phi_k(\theta_0)^T$, where $\Phi_k(\theta_0)^T$ is the k-th line of the matrix $\Phi(\theta_0)$. Since

$$\nabla_\theta \Psi_{n,k}(\theta_0) = n^{-1} \sum_{i=1}^{n} \nabla_\theta \psi_k(X_i, \theta_0),$$

and since $\mathbb{E}[\|\nabla_\theta \psi_k(X_1, \theta_0)\|] < \infty$ (Assumption 4), the law of large number applies. Hence, (2.4) becomes

$$Z_n \left\{ \Phi_k(\theta_0)^T + o_P(1) \right\} (\hat{\theta}_n - \theta_0) = -\Psi_{n,k}(\theta_0) Z_n.$$

The previous decomposition happens for all $k \in \{1, \dots, q\}$, hence with matrix notation, we have

$$Z_n \left\{ \Phi(\theta_0) + o_P(1) \right\} (\hat{\theta}_n - \theta_0) = -\Psi_n(\theta_0) Z_n. \tag{2.5}$$

Now we show that $\|\hat{\theta}_n - \theta_0\| = O_P(n^{-1/2})$. Notice first that $n^{1/2}\Psi_n(\theta_0)$ converges narrowly to a Gaussian random variable by the CLT. By continuity, the norm variables $\|n^{1/2}\Psi_n(\theta_0)\|$ converge narrowly. Thus from the Prohorov theorem they form a tight sequence. From (2.5) we get

$$\begin{aligned} Z_n \|\hat{\theta}_n - \theta_0\| &\leq \Big(\|\Phi(\theta_0)^{-1} \Psi_n(\theta_0)\| + \|o_P(1)\|_\infty \|\hat{\theta}_n - \theta_0\| \Big) Z_n \\ &= \Big(O_P(n^{-1/2}) + o_P(1) \|\hat{\theta}_n - \theta_0\| \Big) Z_n \end{aligned}$$

Since $Z_n = 1 + o_P(1)$, the above display can be re-written as

$$\|\hat{\theta}_n - \theta_0\|(1 + o_P(1)) = O_P(n^{-1/2})(1 + o_P(1) = O_P(n^{-1/2}).$$

Since $1/(1 + o_P(1)) = 1 + o_P(1)$ (see Part 1, Chapter 3, Proposition 3.5), this implies that

$$\|\hat{\theta}_n - \theta_0\| = O_P(n^{-1/2}).$$

Equipped with this tightness result we obtain from (2.5) again that

$$Z_n \Phi(\theta_0)(\hat{\theta}_n - \theta_0) = \{-\Psi_n(\theta_0) + o_P(n^{-1/2})\} Z_n.$$

Usin again that $Z_n = 1 + o_P(1)$ we obtain

$$\begin{aligned} \sqrt{n}\Phi(\theta_0)(\hat{\theta}_n - \theta_0) &= \{-\sqrt{n}\Psi_n(\theta_0) + o_P(1)\}(1 + o_P(1)) \\ &= -\sqrt{n}\Psi_n(\theta_0) + o_P(1) \end{aligned}$$

Inverting the matrix yields the desired result.

□

2.4.1 The case of the maximum likelihood estimator.

Consider a statistical model dominated by a reference measure μ on $\mathcal{X}$,

$$\mathcal{M} = \{p_\theta, \theta \in \Theta\}$$

where p_θ is the density of the law P_θ with respect to μ. For independent observations $X_i, i \leq n$ the log-likelihood function given a sample $X_i = x_i, i = 1, \ldots, n$ is

$$\ell(\theta, x_{1:n}) := \log p_\theta^{\otimes n}(x_{1:n}) = \sum_{i=1}^{n} \log p_\theta(x_i).$$

The maximum likelihood estimator is defined as the maximizer of the log-likelihood function (which is usually unique), i.e.

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \ell(\theta) = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \rho(X_i, \theta)$$

with $\rho(X_i, \theta) = -\log p_\theta(X_i)$ The maximum likelihood is thus a M-estimator. In fact historically, the theory of maximum likelihood estimation precedes that of M-estimators, the latter being a generalization of the former.

We now explain how in certain cases $\hat{\theta}_n$ can be seen as a Z-estimator, so that Theorem 9 yields the asymptotic normality of $\hat{\theta}_n$. First we need to relate the empirical contrast defined above with the Kullback-Leibler divergence between P_{θ_0} and P_θ which we define below

Definition 4 (Kullback-Leibler divergence)**.** *Let P_0, P_1 be two probability leasures with densities p_0, p_1 with respect to μ. The Kullback-Leibler divergence between P_0 and P_1 is*

$$K(P_0, P_1) = \int \log\Big(\frac{p_0(x)}{p_1(x)}\Big) p_0(x) d\mu(x) = \mathbb{E}_{P_0}\Big[\frac{p_0(X)}{p_1(X)}\Big].$$

This quantity is always defined, but may be $+\infty$. It is finite only if $P_0 \ll P_1$. In addition, $K(P_0, P_1) \geq 0$ and we have $K(P_0, P_1) = 0$ if and only if $P_0 = P_1$.

We leave the proof of existence and the condition for finiteness as exercises (for existence, consider the negative part of the integrand and use that $\log(y) \leq y$ for $y \geq 1$. For finiteness and non-negativity use the concavity of the logarithm function and Jensen inequality).

Assume that $P = P_{\theta_0}$ for some $\theta_0 \in \Theta$. normalizing the contrast function by the log-likelihood at θ_0 and dividing by n, the estimator $\hat{\theta}_n$ minimizes

$$M_n(X_{1:n}, \theta) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p_{\theta_0}(X_i)}{p_\theta(X_i)}.$$

By the law of large numbers $M_n(X_{1:n}, \theta) \to K(P_{\theta_0}, P_{\theta_1})$, almost surely. By the properties of the Kullback-Leibler divergence, in case the model is identifiable, the unique minimizer of the limit is θ_0. $\hat{\theta}_n$ is thus a M-estimator.

Under regularity conditions on the likelihood function, in particular if $\theta \mapsto \log p_\theta(x)$ is differentiable, then $\hat{\theta}_n$ is a zero of the the criterion

$$\Psi_n(\theta) = \sum_{i=1}^{n} \psi(X_i, \theta)$$

with $\psi(X_i, \theta) = \nabla_\theta \ell(\theta, x)$ is called the *score* function, where we recall that $\ell(\theta, x) = \log p_\theta(x)$. Thus $\hat{\theta}_n$ is also a Z-estimator. Recall from the introductory course in statistics that if the model is *regular* (see below) then for all $\theta_0 \in \Theta$, the expectancy of the score is zero, $\mathbb{E}_{\theta_0}(\nabla_\theta \ell(\theta_0, X)) = 0$. Indeed in such a case

$$\mathbb{E}_{\theta_0}(\nabla_\theta \ell(\theta_0, X)) = \int \nabla_\theta \log p_{\theta_0}(x) p_{\theta_0}(x) \mathrm{d}\mu(x) = \int \nabla_\theta p_{\theta_0}(x) \mathrm{d}\mu(x) = \nabla_\theta \int p_{\theta_0}(x) \mathrm{d}\mu(x) = 0.$$

Definition 5 (regular model)**.** *The model $\mathcal{M}$ is regular if*

1. *Θ is an open subset of $\mathbb{R}^d$ and $\mathcal{A} = \{x : p_\theta(x) > 0\}$ does not depend on θ*

2. *For all θ the score $\nabla_\theta \log p_\theta(x)$ exists and $\mathbb{E}|\nabla_\theta \log p_\theta(X)| < \infty$*

3. *If $T : \mathcal{X} \to \mathbb{R}$ is a measurable function such that*

$$\mathbb{E}_\theta|T(X)| < \infty \quad \text{and} \quad \mathbb{E}_\theta\Big\{|T(X)|\,\|\nabla_\theta \log p_\theta(X)\|\Big\} < \infty$$

then integration and derivation operators may be permuted, i.e. it holds that

$$\nabla_\theta \int T(x)p_\theta(x)\mathrm{d}\mu(x) = \int T(x)\nabla_\theta p_\theta(x)\mathrm{d}\mu(x).$$

We sumarize the discussion.

Theorem 10 (Asymptotic normality of the maximum-likelihood estimator)**.** *If the model is regular and if the score function $\psi(x, \theta) = \nabla_\theta \ell(\theta, x)$ satisfies the assumptions of Theorem 9 then the maximum likelihood estimator is asymptotically normal with asymptotic variance equals to the inverse Fisher information*

$$I(\theta_0) = \mathbb{E}_{\theta_0} \nabla_\theta \ell(\theta_0, X_1) \nabla_\theta \ell(\theta_0, X_1)^\top = -\mathbb{E}_{\theta_0} \nabla^2_\theta \ell(\theta_0, X_1),$$

in other words

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}(0, I(\theta_0)^{-1}).$$

Proof. There only remains to verify the expression for the asymptotic variance. Notice first that an integration by parts and model regularity yield the identity $I(\theta) = -\mathbb{E}_\theta \nabla^2_\theta \ell(\theta, X_1)$. From Theorem 9 the asymptotic variance it is given by

$$\Sigma = \Phi(\theta_0)^{-1} \operatorname{var}(\psi(X_1, \theta_0)\Phi(\theta_0)^{-1}.$$

Since $\psi(X_1, \theta_0) = \nabla_\theta \ell(\theta_0, X_1)$ has expectancy equal to 0 from the assumption that the model is regular, we have $\operatorname{var}(\psi(X_1, \theta_0)) = I(\theta_0)$. Also by definition $\Phi(\theta_0) = \nabla^2_\theta \ell(\theta_0, X_1) = -I(\theta_0)$. We thus obtain

$$\Sigma = -I(\theta_0)^{-1} I(\theta_0)(-I(\theta_0))^{-1} = I(\theta_0)^{-1}.$$

□

Bibliography

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* *7*(1), 1–26.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.