Lecture notes
Asymptotic statistics

François Bachoc
University Paul Sabatier

November 29, 2024

Contents

Acknowledgements

Parts of these lecture notes benefited from parts of lecture notes written by Jean-François Dupuy, Elisabeth Gassiat and Thierry Klein.

Introduction

The aim of these lecture notes is to study sequences of random variables and random vectors indexed by $n \to \infty$, where n is most of the cases a number of independent statistical observations. These random variables and vectors will typically stem from estimators of the form $\hat{\theta}_n$ for estimating a vector of parameter θ in a parametric model. This parametric model is for instance $\{\mathcal{L}_\theta; \theta \in \Theta\}$ for a set $\Theta \in \mathbb{R}^p$ and where, for all θ, $\mathcal{L}_\theta$ is a distribution on $\mathbb{R}$. In this case, the statistical observations are $X_1, \ldots, X_n \in \mathbb{R}$ with unknown distribution $\theta_0 \in \Theta$.

An important result that will be proved is the asymptotic normality of the maximum likelihood estimator $\hat{\theta}_n$ based on independent $X_1, \ldots, X_n$ as $n \to \infty$. Under regularity conditions, we will show that

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$$

converges in distribution to a centered Gaussian vector.

For (much) more content on the topic of asymptotic statistics, we refer in particular to the book [VdV07].

General notations

Throughout, $\mathbb{N}$ will be the set of non-zero natural numbers, $\mathbb{N} = \{1, 2, \ldots\}$. For a set A in a metric space E, $\overline{A}$ will be its closure, $\mathring{A}$ will be its interior, $\delta A = \overline{A} \backslash \mathring{A}$ will be its boundary and $A^c = E \backslash A$ will be its complement. Also the diameter of A will be defined as $\mathrm{diam}(A) = \sup\{\mathrm{dist}(u, v) : u, v \in A\}$ where dist is the distance in the space E.

We write $\mathbb{1}\{\text{event}\}$ as the indicator function that an event holds true. For a function $g : E \to F$ and $A \subset F$, we write $g^{-1}(A) = \{x \in E : g(x) \in A\}$. For $c \in \mathbb{R}^k$ and $r \geq 0$ we let $B(c, r) = \{x \in \mathbb{R}^k : \|x - c\| < r\}$. On an Euclidean space, the inner product is written $\langle \cdot, \cdot \rangle$ and the Euclidean norm is written $\| \cdot \|$. The acronym c.d.f. will stand for cumulative distribution function. The acronym i.i.d. will stand for independent and identically distributed. The acronyms l.h.s. and r.h.s. will stand for left-hand side and right-hand side. The acronym w.r.t. will stand for with respect to.

For a random vector X, its covariance matrix is written $\mathrm{cov}(X)$. For two numbers u, v, we write $u \wedge v = \min(u, v)$. The transpose of a matrix M is written $M^\top$. If M is square and invertible, we write $M^{-\top} = (M^{-1})^\top = (M^\top)^{-1}$. For a function $\phi : \mathbb{R}^k \to \mathbb{R}^m$ that is differentiable at x, its $m \times k$ Jacobian matrix at x is written $J\phi(x)$. For a function $\phi : \mathbb{R}^k \to \mathbb{R}$ that is differentiable at x, its $k \times 1$ gradient column vector at x is written $\nabla\phi(x)$.

For $t \in \mathbb{R}$ we write

$$\mathrm{sign}(t) = \begin{cases} -1 & \text{if } t < 0 \\ 0 & \text{if } t = 0 \\ 1 & \text{if } t > 0 \end{cases}.$$

1 Convergence of random vectors

1.1 Definitions

Let $X = (X_1, \ldots, X_k)$ be a random vector of $\mathbb{R}^k$. We can naturally extend the definition of a cumulative distribution function (c.d.f.) of a random variable by defining

$$F_X : \mathbb{R}^k \to [0, 1]$$

by, for $x = (x_1, \ldots, x_k) \in \mathbb{R}^k$,

$$F_X(x) = \mathbb{P}\left(X_1 \leq x_1, \ldots, X_k \leq x_k\right).$$

Definition 1. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random vectors of $\mathbb{R}^k$ and X be a random vector of $\mathbb{R}^k$. Then we say that X_n converges to X*

- **in distribution** *if* $F_{X_n}(x) \to F_X(x)$ *as* $n \to \infty$ *for all* x *such that* F_X *is continuous at* x*. In this case we write*

$$X_n \xrightarrow[n\to\infty]{\mathcal{L}} X;$$

- **in probability** *if for all* $\epsilon > 0$,

$$\mathbb{P}\left(\|X_n - X\| \geq \epsilon\right) \underset{n\to\infty}{\longrightarrow} 0.$$

In this case we write

$$X_n \xrightarrow[n\to\infty]{p} X;$$

- **almost surely** *if*

$$\mathbb{P}\left(\|X_n - X\| \underset{n\to\infty}{\longrightarrow} 0\right) = 1.$$

In this case we write

$$X_n \xrightarrow[n\to\infty]{a.s.} X.$$

In the above definition, we remark that convergence in distribution can hold even if X_n and X are not defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Indeed, this definition actually apply to the distributions of X_n and X on $\mathbb{R}^k$. On the other hand, convergence in probability and almost surely need X_n and X to be defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, for instance for $X_n - X$ to be well-defined.

Remark 2. *Because of the above discussion, the definition of the convergence in distribution, and all the properties presented next, hold, up to obvious changes, if the limit random vector* X *is replaced by a limit distribution* $\mathcal{L}$ *on* $\mathbb{R}^k$.

1.2 Equivalent conditions for convergence in distribution and continuous mapping

Lemma 3 (Portmanteau)**.** *Let* $(X_n)_{n\in\mathbb{N}}$ *be a sequence of random vectors of* $\mathbb{R}^k$ *and* X *be a random vector of* $\mathbb{R}^k$*. The following statements are equivalent.*

1. $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X$.
2. $\mathbb{E}[f(X_n)] \underset{n\to\infty}{\longrightarrow} \mathbb{E}[f(X)]$ *for any bounded continuous function* f.
3. $\mathbb{E}[f(X_n)] \underset{n\to\infty}{\longrightarrow} \mathbb{E}[f(X)]$ *for any bounded* L*-Lipschitz-continuous function* f $(L < \infty)$.
4. $\liminf_{n\to\infty} \mathbb{E}[f(X_n)] \geq \mathbb{E}[f(X)]$ *for any continuous non-negative function.*
5. $\liminf_{n\to\infty} \mathbb{P}(X_n \in O) \geq \mathbb{P}(X \in O)$ *for any open set* O.
6. $\limsup_{n\to\infty} \mathbb{P}(X_n \in F) \leq \mathbb{P}(X \in F)$ *for any closed set* F.
7. $\mathbb{P}(X_n \in B) \underset{n\to\infty}{\longrightarrow} \mathbb{P}(X \in B)$ *for all Borel set* B *such that* $\mathbb{P}(X \in \delta B) = 0$.

Proof. We skip this proof in the lecture notes. □

Let us illustrate some of the statements above with the simple example where $X_n \sim \mathcal{N}(0, \frac{1}{n})$ and $X = 0$ a.s. Then one can check that $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X$ (**exercize**). Let us illustrate the statement 6 with the closed set $\{0\}$. We have

$$\limsup_{n\to\infty} \mathbb{P}(X_n \in \{0\}) = \limsup_{n\to\infty} 0 = 0 \leq 1 = \mathbb{P}(X \in \{0\}).$$

Now let us illustrate the statement 5 with the open set $(-\epsilon, \epsilon)$ for some $\epsilon > 0$. We have

$$\liminf_{n\to\infty} \mathbb{P}(X_n \in (-\epsilon, \epsilon)) = \liminf_{n\to\infty} \mathbb{P}\left(\sqrt{n}X_n \in (-\sqrt{n}\epsilon, \sqrt{n}\epsilon)\right) = \liminf_{n\to\infty} \underbrace{\mathbb{P}\left(Z \in (-\sqrt{n}\epsilon, \sqrt{n}\epsilon)\right)}_{Z\sim\mathcal{N}(0,1)} = 1$$

$$= \mathbb{P}(X \in (-\epsilon, \epsilon)).$$

Theorem 4 (Continuous mapping)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random vectors of $\mathbb{R}^k$ and X be a random vector of $\mathbb{R}^k$. Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous at all points of a set C satisfying $\mathbb{P}(X \in C) = 1$. Then*

1. *If $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X$ then $g(X_n) \xrightarrow[n\to\infty]{\mathcal{L}} g(X)$.*
2. *If $X_n \xrightarrow[n\to\infty]{p} X$ then $g(X_n) \xrightarrow[n\to\infty]{p} g(X)$.*
3. *If $X_n \xrightarrow[n\to\infty]{a.s.} X$ then $g(X_n) \xrightarrow[n\to\infty]{a.s.} g(X)$.*

Proof. **3.** Proving Item 3 is left as an **exercize**.

2. Let $\epsilon > 0$ and $\delta > 0$. We have

$$\mathbb{P}\left(\|g(X_n) - g(X)\| \geq \epsilon\right) \leq \mathbb{P}\left(\|g(X_n) - g(X)\| \geq \epsilon, \|X_n - X\| \leq \delta\right) + \mathbb{P}\left(\|X_n - X\| \geq \delta\right). \quad (1)$$

The quantity $\mathbb{P}\left(\|X_n - X\| \geq \delta\right)$ goes to zero as $n \to \infty$ since $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X$. Let us define

$$B_\delta = \{x \in \mathbb{R}^k : \exists y \in \mathbb{R}^k s.t. \|x - y\| \leq \delta, \|g(x) - g(y)\| \geq \epsilon\}.$$

Then (1) yields

$$\limsup_{n\to\infty} \mathbb{P}\left(\|g(X_n) - g(X)\| \geq \epsilon\right) \leq \mathbb{P}(X \in B_\delta) = \mathbb{P}(X \in B_\delta \cap C).$$

For all $x \in C$, g is continuous at x so there is $\delta > 0$ small enough such that for all y, $\|x - y\| \leq \delta$ implies $\|g(x) - g(y)\| < \epsilon$. Hence, for $\delta > 0$ small enough $\mathbb{1}\{x \in B_\delta \cap C\} = 0$. Hence by dominated convergence, $\mathbb{P}(X \in B_\delta \cap C) \to 0$ as $\delta \to 0$. Hence $\limsup_{n\to\infty} \mathbb{P}\left(\|g(X_n) - g(X)\| \geq \epsilon\right) = 0$ and thus Item 2 is proved.

1. We will apply Item 6 from Lemma 3. Let F be a closed set of $\mathbb{R}^m$. We have $\{g(X_n) \in F\} = \{X_n \in g^{-1}(F)\}$. We have

$$g^{-1}(F) \subset \overline{g^{-1}(F)} \subset g^{-1}(F) \cup C^c.$$

To prove the second inclusion, consider $x \in \overline{g^{-1}(F)}$. There is a sequence x_n such that $x_n \to x$. If $x \in C$, then by continuity of g at x, $g(x_n) \to g(x)$ and thus $g(x) \in F$ and thus $x \in g^{-1}(F)$. Otherwise $x \notin C$.

Hence,

$$\limsup_{n\to\infty} \mathbb{P}\left(g(X_n) \in F\right) \leq \limsup_{n\to\infty} \mathbb{P}\left((X_n \in \overline{g^{-1}(F)}\right)$$

Hence, by Item 6 from Lemma 3,

$$\limsup_{n\to\infty} \mathbb{P}\left(g(X_n) \in F\right) \leq \mathbb{P}\left(X \in \overline{g^{-1}(F)}\right) \leq \mathbb{P}\left(X \in g^{-1}(F)\right) + \mathbb{P}(x \in C^c) = \mathbb{P}(g(X) \in F).$$

Hence, by Item 6 from Lemma 3, $g(X_n) \xrightarrow[n\to\infty]{\mathcal{L}} g(X)$. □

We remark from the theorem statement that if the random variable X is a fixed constant c, we just need the continuity of g at c.

1.3 Uniformly tight random vectors

We observe that for any random vector X and any $\epsilon > 0$, there exists $M > 0$ such that

$$\mathbb{P}(\|X\| \geq M) \leq \epsilon$$

(**exercize**). We thus say that any fixed random vector is **tight**.

Definition 5. *Let $F = \{X_a, a \in A\}$ be a family of random vectors. We say that F is* **uniformly tight** *is*

$$\forall \epsilon > 0,\ \exists M > 0 \ s.t. \ \sup_{a\in A} \mathbb{P}(\|X_a\| \geq M) \leq \epsilon.$$

Equivalently

$$\sup_{a\in A} \mathbb{P}(\|X_a\| \geq M) \underset{M\to\infty}{\longrightarrow} 0.$$

Theorem 6 (Prokhorov)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random vectors bounded in probability.*

1. *If there exists a random vector X such that $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X$ then the family $(X_n)_{n\in\mathbb{N}}$ is uniformly tight.*
2. *If the family $(X_n)_{n\in\mathbb{N}}$ is uniformly tight then there exists a random vector X and a subsequence $(X_{\phi(n)})_{n\in\mathbb{N}}$ such that $X_{\phi(n)} \xrightarrow[n\to\infty]{\mathcal{L}} X$.*

Proof. We skip this proof in the lecture notes. □

We remark that these definitions and results related to tightness actually apply to the distributions of the vectors X_n, not the random vectors themselves.

Also, we can see this theorem as an extension of a well-known deterministic result in finite dimension: any convergent sequence is bounded and from any bounded sequence we can extract a convergent subsequence.

1.4 Relationships between the various modes of convergence

Theorem 7. *Let $(X_n)_{n\in\mathbb{N}}$, $(Y_n)_{n\in\mathbb{N}}$, X and Y be random vectors and let c be a constant vector. Then*

1. *If $X_n \xrightarrow[n\to\infty]{a.s.} X$ then $X_n \xrightarrow[n\to\infty]{p} X$,*
2. *If $X_n \xrightarrow[n\to\infty]{p} X$ then $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X$,*
3. *$X_n \xrightarrow[n\to\infty]{p} c$ if and only if $X_n \xrightarrow[n\to\infty]{\mathcal{L}} c$,*
4. *If $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X$ and $\|X_n - Y_n\| \xrightarrow[n\to\infty]{p} 0$ then $Y_n \xrightarrow[n\to\infty]{\mathcal{L}} X$,*
5. **(Slutsky)** *If $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X$ and $Y_n \xrightarrow[n\to\infty]{p} c$ then $(X_n, Y_n) \xrightarrow[n\to\infty]{\mathcal{L}} (X, c)$,*
6. *If $X_n \xrightarrow[n\to\infty]{p} X$ and $Y_n \xrightarrow[n\to\infty]{p} Y$ then $(X_n, Y_n) \xrightarrow[n\to\infty]{p} (X, Y)$.*

Proof. **1.** Let $\epsilon > 0$. Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Consider the function $\omega \mapsto \mathbb{1}\{\|X_n(\omega) - X(\omega)\| \geq \epsilon\}$. For $\mathbb{P}$-a.e. $\omega \in \Omega$, we have $X_n(\omega) \to X(\omega)$ as $n \to \infty$ and thus $\mathbb{1}\{\|X_n(\omega) - X(\omega)\| \geq \epsilon\} \to 0$ as $n \to \infty$. Hence, from the dominated convergence theorem $\int_\Omega \mathbb{1}\{\|X_n(\omega) - X(\omega)\| \geq \epsilon\} \mathrm{d}\mathbb{P} \to 0$ as $n \to \infty$. We conclude by using $\int_\Omega \mathbb{1}\{\|X_n(\omega) - X(\omega)\| \geq \epsilon\} \mathrm{d}\mathbb{P} = \mathbb{E}\left[\mathbb{1}\{\|X_n - X\| \geq \epsilon\}\right] = \mathbb{P}\left(\|X_n - X\| \geq \epsilon\right)$.

2. is a consequence of Item 4.

3. Because of Item 2, only $\Longrightarrow$ needs to be proved. We will use Item 6 from Lemma 3. Let $\epsilon > 0$ and $B = B(c, \epsilon)$, the open Euclidean ball of center c and radius ϵ. We have

$$\limsup_{n\to\infty} \mathbb{P}(\|X_n - c\| \geq \epsilon) = \limsup_{n\to\infty} \mathbb{P}(X_n \in B^c) \leq \mathbb{P}(c \in B^c) = 0.$$

4. We will use Item 3 from Lemma 3. Consider a bounded L-Lipschitz function f. Let M be an upper bound on $|f|$. We have

$$\begin{aligned} |\mathbb{E}[f(Y_n)] - \mathbb{E}[f(X)]| &\leq |\mathbb{E}[f(Y_n)] - \mathbb{E}[f(X_n)]| + |\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]| \\ &\leq \mathbb{E}\left[|f(Y_n) - f(X_n)|\right] + |\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]|. \end{aligned}$$

Above, $\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)] \to 0$ as $n \to \infty$ from Item 3 from Lemma 3. Also

$$\mathbb{E}\left[|f(Y_n) - f(X_n)|\right] \leq L\mathbb{E}\left[\|Y_n - X_n\|\right] \leq L\epsilon\mathbb{P}\left(\|Y_n - X_n\| \leq \epsilon\right) + 2LM\mathbb{P}\left(\|Y_n - X_n\| \geq \epsilon\right).$$

From this we obtain

$$\limsup_{n\to\infty} |\mathbb{E}[f(Y_n)] - \mathbb{E}[f(X)]| \leq L\epsilon.$$

Since this is true for all $\epsilon > 0$ this lim sup is zero and thus we conclude from Item 3 from Lemma 3.

5. We have

$$\limsup_{n\to\infty} \mathbb{P}\left(\|(X_n, Y_n) - (X_n, c)\| \geq \epsilon\right) = \limsup_{n\to\infty} \mathbb{P}\left(\|Y_n - c\| \geq \epsilon\right) = 0$$

since $Y_n \underset{n\to\infty}{\overset{p}{\longrightarrow}} c$. Hence $\|(X_n, Y_n) - (X_n, c)\| \underset{n\to\infty}{\overset{p}{\longrightarrow}} 0$. Hence from Item 4 it suffices to show that $(X_n, c) \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} (X, c)$. Let k be the dimension of X and m be the dimension of c. For any continuous bounded function $f : \mathbb{R}^{k+m} \to \mathbb{R}$, the function $f_c : \mathbb{R}^k \to \mathbb{R}$ defined by $f_c(x) = f(x, c)$ is bounded continuous. Hence $\mathbb{E}[f(X_n, c)] = \mathbb{E}[f_c(X_n)] \underset{n\to\infty}{\longrightarrow} \mathbb{E}[f_c(X)] = \mathbb{E}[f(X, c)]$. Hence $(X_n, c) \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} (X, c)$ from Item 2. in Lemma 3.

6. is left as an **exercize**. □

From the above theorem and Theorem 4, we obtain the following theorem (**exercize**).

Theorem 8 (Slutsky). *Let $(X_n)_{n\in\mathbb{N}}$, X and $(Y_n)_{n\in\mathbb{N}}$ be random vectors and let c be a constant vector. If $X_n \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} X$ and $Y_n \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} c$ then*

1. $X_n + Y_n \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} X + c$, *when* $X_n, Y_n, c \in \mathbb{R}^k$;
2. $Y_n X_n \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} cX$, *when* $X_n \in \mathbb{R}^k$ *and* $Y_n, c \in \mathbb{R}$;
3. $\frac{1}{Y_n} X_n \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \frac{1}{c} X$, *when* $X_n \in \mathbb{R}^k$ *and* $Y_n, c \in \mathbb{R}\backslash\{0\}$.

Lemma 9 (Uniform convergence of the c.d.f. and convergence in distribution). *Let $(X_n)_{n\in\mathbb{N}}$ and X be random vectors on $\mathbb{R}^k$ and assume that $X_n \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} X$ and that F_X is continuous on $\mathbb{R}^k$. Then*

$$\sup_{x\in\mathbb{R}^k} |F_{X_n}(x) - F_X(x)| \underset{n\to\infty}{0}.$$

Proof. We write the proof for $k = 1$ to simplify the notations. The extension to a general k is left as an **exercize**. Let $\epsilon > 0$ and an integer N such that $1/N \leq \epsilon$. Since F_X is continuous, there exist $x_1, \ldots, x_{N-1}$ such that $F_X(x_i) = i/N$ for $i = 1, \ldots, N-1$. Let also by convention $x_0 = -\infty$ and $x_N = +\infty$. Since F_X and F_{X_n} are non-decreasing, we have, for any $i = 1, \ldots, N$ and $x \in [x_{i-1}, x_i]$[1]

$$F_{X_n}(x) - F(x) \leq F_{X_n}(x_i) - F_X(x_{i-1}) \leq F_{X_n}(x_i) - F_X(x_i) + \frac{1}{N}$$

(we use the conventions $F_{X_n}(-\infty) = F_X(-\infty) = 0$ and $F_{X_n}(+\infty) = F_X(+\infty) = 1$) and

$$F_{X_n}(x) - F(x) \geq F_{X_n}(x_{i-1}) - F_X(x_i) \geq F_{X_n}(x_{i-1}) - F_X(x_{i-1}) - \frac{1}{N}.$$

Hence

$$\sup_{x\in\mathbb{R}} |F_{X_n}(x) - F_X(x)| \leq \max_{i=1,\ldots,N} |F_{X_n}(x_i) - F_X(x_i)| + \frac{1}{N}$$

and thus by definition of convergence in distribution,

$$\limsup_{n\to\infty} \sup_{x\in\mathbb{R}} |F_{X_n}(x) - F_X(x)| \leq \frac{1}{N}.$$

This is true for all N which concludes the proof. □

[1] Actually if $i = 1$, $x \leq x_1$ and if $i = N$, $x \geq x_{N-1}$.

1.5 The symbols $o_{\mathbb{P}}$ and $\mathcal{O}_{\mathbb{P}}$

We introduce here two symbols that will be very useful in the sequel. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random vectors.

- $X_n = o_{\mathbb{P}}(1)$ means that $\|X_n\| \xrightarrow[n\to\infty]{p} 0$. More generally, for a sequence $(R_n)_{n\in\mathbb{N}}$ of non-negative random variables, $X_n = o_{\mathbb{P}}(R_n)$ means that there exists a sequence of random vectors $(Y_n)_{n\in\mathbb{N}}$ such that $X_n = R_n Y_n$ and $\|Y_n\| \xrightarrow[n\to\infty]{p} 0$.
- $X_n = \mathcal{O}_{\mathbb{P}}(1)$ means that $(X_n)_{n\in\mathbb{N}}$ is uniformly tight. More generally, for a sequence $(R_n)_{n\in\mathbb{N}}$ of non-negative random variables, $X_n = \mathcal{O}_{\mathbb{P}}(R_n)$ means that there exists a sequence of random vectors $(Y_n)_{n\in\mathbb{N}}$ such that $X_n = R_n Y_n$ and $(Y_n)_{n\in\mathbb{N}}$ is uniformly tight.

The next lemma allows us to replace deterministic quantities by random quantities in the deterministic standard notations o and $\mathcal{O}$.

Lemma 10. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random vectors on $\mathbb{R}^k$ such that $X_n \xrightarrow[n\to\infty]{p} 0$. Then for all $q > 0$ and for all function $R : \mathbb{R}^k \to \mathbb{R}^m$ such that $R(0) = 0$,*

1. *$\|R(h)\| = o(\|h\|^q)$ as $h \to 0$ implies $R(X_n) = o_{\mathbb{P}}(\|X_n\|^q)$;*
2. *$\|R(h)\| = O(\|h\|^q)$ as $h \to 0$ implies $R(X_n) = \mathcal{O}_{\mathbb{P}}(\|X_n\|^q)$.*

Proof. We define $g : \mathbb{R}^k \to \mathbb{R}^m$ by $g(h) = \frac{R(h)}{\|h\|^q}$ if $h \neq 0$ and $g(0) = 0$. Then $R(X_n) = g(X_n)\|X_n\|^q$.

1. In this case the function g is continuous at 0. Hence by Theorem 4 (continuous mapping), since $\|X_n\| \xrightarrow[n\to\infty]{p} 0$, $g(X_n) \xrightarrow[n\to\infty]{p} 0$.

2. Since $R(h) = O(\|h\|^q)$ there exists $\delta > 0$ such that when $\|h\| \leq \delta$ we have $R(h) \leq M\|h\|^q$ and thus $g(h) \leq M$. Hence

$$\limsup_{n\to\infty} \mathbb{P}(\|g(X_n)\| \geq M) \leq \limsup_{n\to\infty} \mathbb{P}(\|X_n\| \geq \delta) = 0$$

since $X_n \xrightarrow[n\to\infty]{p} 0$. Hence $g(X_n)$ is uniformly tight and thus $R(X_n) = \mathcal{O}_{\mathbb{P}}(\|X_n\|^q)$. □

1.6 Characteristic function

Definition 11. *Let X be a random vector of $\mathbb{R}^k$ and $t \in \mathbb{R}^k$ be deterministic. The* **characteristic function** *of X at t is defined by*

$$\phi_X(t) = \mathbb{E}\left[e^{i\langle t,x\rangle}\right]$$

with $i = \sqrt{-1}$.

Theorem 12 (Paul Levy)**.**

1. *Let $(X_n)_{n\in\mathbb{N}}$ and X be random vectors of $\mathbb{R}^k$. Then the two following statements are equivalent.*
   (a) $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X$;
   (b) $\phi_{X_n}(t) \xrightarrow[n\to\infty]{} \phi_X(t)$ for all $t \in \mathbb{R}^k$.
2. *If there is a function $\phi : \mathbb{R}^k \to \mathbb{R}$ such that ϕ is continuous at zero and $\phi_{X_n}(t) \xrightarrow[n\to\infty]{} \phi(t)$ for all $t \in \mathbb{R}^k$, then there is a random vector X such that $\phi = \phi_X$ and $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X$.*

Proof. We skip the proof in these lecture notes. □

Lemma 13. *Two random vectors X and Y have the same distribution if and only if their characteristic functions are equal.*

Proof. We skip the proof in these lecture notes. □

1.7 Strong law of large number and central limit theorem

Proposition 14. *Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of i.i.d. random vectors such that $\mathbb{E}[\|X_1\|] < \infty$. Then*

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}[X_1].$$

Proof. We skip the proof in these lecture notes. □

Proposition 15. *Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of i.i.d. random vectors such that $\mathbb{E}[\|X_1\|^2] < \infty$. Then*

$$\sqrt{n}\left(\frac{X_1 + \cdots + X_n}{n} - \mathbb{E}[X_1]\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \mathrm{cov}(X_1)).$$

Proof. We skip the proof in these lecture notes. □

1.8 Uniform integrability and convergence of moments

Definition 16 (Uniform integrability)**.** *A sequence of random vectors $(X_n)_{n\in\mathbb{N}}$ is* **uniformly integrable (u.i.)** *if*

$$\lim_{M\to\infty} \sup_{n\in\mathbb{N}} \mathbb{E}\left[\|X_n\|\mathbb{1}\{\|X_n\| \geq M\}\right] = 0.$$

Note that convergence in distribution does not necessarily imply convergence of expectation for unbounded functions. The next theorem shows that this occurs under the additional condition of uniform integrability.

Theorem 17. *Consider a function $f : \mathbb{R}^k \to \mathbb{R}$ which is continuous on a set C. Let X be a random vector of $\mathbb{R}^k$ which belongs a.s. to C. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random vectors of $\mathbb{R}^k$. Then if $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X$ and if $(f(X_n))_{n\in\mathbb{N}}$ is u.i., we have*

$$\mathbb{E}[f(X_n)] \xrightarrow[n\to\infty]{} \mathbb{E}[f(X)].$$

Proof. We assume that $f(X_n)$ is non-negative, otherwise (**exercize**) we separate the positive and negative parts.

By continuity, $f(X_n) \xrightarrow[n\to\infty]{\mathcal{L}} f(X)$ from Theorem 4 (continuous mapping). We have for all $M > 0$,

$$\begin{aligned}
&\limsup_{n\to\infty} |\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]| \\
&\leq \limsup_{n\to\infty} |\mathbb{E}[f(X_n)] - \mathbb{E}[f(X_n) \wedge M]| + \limsup_{n\to\infty} |\mathbb{E}[f(X_n) \wedge M] - \mathbb{E}[f(X) \wedge M]| \\
&+ \limsup_{n\to\infty} |\mathbb{E}[f(X) \wedge M] - \mathbb{E}[f(X)]| \,.
\end{aligned} \tag{2}$$

Fix $\epsilon > 0$. Remark that

$$|\mathbb{E}[f(X_n)] - \mathbb{E}[f(X_n) \wedge M]| \leq \mathbb{E}[|f(X_n)|\mathbb{1}\{|f(X_n)| \geq M\}].$$

Since $(f(X_n))_{n\in\mathbb{N}}$ is u.i. we can fix M such that the first lim sup on the r.h.s. of (2) is smaller than ϵ. Similarly, we can increase M such that the third lim sup is smaller than ϵ. The second lim sup is then zero from Theorem 4 (continuous maping), because $f(\cdot) \wedge M$ is bounded and continuous on C. Hence we have

$$\limsup_{n\to\infty} |\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]| \leq 2\epsilon$$

for all $\epsilon > 0$ which concludes the proof. □

2 The Delta method

2.1 The theorem

Let $\theta \in \mathbb{R}^k$ be a parameter in a statistical model and let $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ be a sequence of estimators for it. Consider a function $\phi : \mathbb{R}^k \to \mathbb{R}^m$. It is natural to estimate $\phi(\theta)$ by $\phi(\widehat{\theta}_n)$ and to ask if asymptotic properties of $\widehat{\theta}_n - \theta$ can be transferred to $\phi(\widehat{\theta}_n) - \phi(\theta)$.

The continuous mapping theorem (Theorem 4) provides a first answer. If $\widehat{\theta}_n \xrightarrow[n\to\infty]{p} \theta$ and ϕ is continuous, then $\phi(\widehat{\theta}_n) \xrightarrow[n\to\infty]{p} \phi(\theta)$.

Consider now that we have a stronger result, a central limit theorem: $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma)$ for some covariance matrix Σ. Then, if ϕ is linear and defined by a $m \times k$ matrix M, we have (continuous mapping, **exercize**) $\sqrt{n}(M\widehat{\theta}_n - M\theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, M\Sigma M^\top)$.

The intuition of the Delta method is that a similar result takes place if ϕ is continuously differentiable, where the role of M will be played by the Jacobian matrix $J\phi$.

Theorem 18 (Delta method)**.** *Let $\theta \in \mathbb{R}^k$ be fixed. Let $\phi : \mathbb{R}^k \to \mathbb{R}^m$ be differentiable at θ. Let $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ be a sequence of random vectors and let X be a random vector such that, for a sequence $(r_n)_{n\in\mathbb{N}}$ that goes to infinity, we have*

$$r_n \left(\widehat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} X.$$

Then

$$r_n \left(\phi(\widehat{\theta}_n) - \phi(\theta)\right) \xrightarrow[n\to\infty]{\mathcal{L}} (J\phi(\theta))X \tag{3}$$

and

$$r_n \left(\phi(\widehat{\theta}_n) - \phi(\theta)\right) - r_n(J\phi(\theta))(\widehat{\theta}_n - \theta) \xrightarrow[n\to\infty]{p} 0. \tag{4}$$

Proof. Observe first that $\widehat{\theta}_n - \theta = \frac{1}{r_n} r_n(\widehat{\theta}_n - \theta)$ goes to zero from Lemma 8 (Slutsky). Observe also that the sequence $r_n(\widehat{\theta}_n - \theta)$ is uniformly tight from Theorem 6 (Prokhorov). Next, write

$$R(h) = \phi(\theta + h) - \phi(\theta) - (J\phi(\theta))h.$$

By definition of differentiability we have $R(h) = o(\|h\|)$ as $h \to 0$. Hence from Lemma 10,

$$r_n \left(\phi(\widehat{\theta}_n) - \phi(\theta)\right) = (J\phi(\theta))r_n(\widehat{\theta}_n - \widehat{\theta}) + r_n R(\widehat{\theta}_n - \widehat{\theta}) = r_n(J\phi(\theta))(\widehat{\theta}_n - \widehat{\theta}) + r_n o_{\mathbb{P}}(\widehat{\theta}_n - \widehat{\theta}).$$

Above, $r_n o_{\mathbb{P}}(\widehat{\theta}_n - \widehat{\theta}) = o_{\mathbb{P}}(r_n(\widehat{\theta}_n - \widehat{\theta})) = o_{\mathbb{P}}(1)$ because $r_n(\widehat{\theta}_n - \widehat{\theta}) = \mathcal{O}_{\mathbb{P}}(1)$ (**exercize**). This proves (4).

From Theorem 4 (continuous mapping) and because $r_n \left(\widehat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} X$, it follows that $r_n(J\phi(\theta))(\widehat{\theta}_n - \theta) = (J\phi(\theta))r_n(\widehat{\theta}_n - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} (J\phi(\theta))X$. Hence (3) holds from Item 4 in Theorem 7. □

2.2 The example of variance estimation

Consider a sequence of i.i.d. random variables $(X_i)_{i\in\mathbb{N}}$ such that $\mathbb{E}[X_1^4] < \infty$. We can thus define the mean $\mathbb{E}[X_1]$ and the 3 **centered moments** μ_2, μ_3, μ_4 with

$$\mu_k = \mathbb{E}\left[(X_1 - \mathbb{E}[X_1])^k\right].$$

We naturally estimate $\mathbb{E}[X_1]$ by $\widehat{\mu}_{1,n} = \frac{1}{n}\sum_{i=1}^n X_i$ and μ_2 is the variance that we naturally estimate by

$$\widehat{\mu}_{2,n} = \frac{1}{n}\sum_{i=1}^n (X_i - \hat{\mu}_{1,n})^2.$$

Giving asymptotic results for $\widehat{\mu}_{2,n}$ is not easy because we may not be able to write it as an average of independent variables, for instance (contrarily to $\widehat{\mu}_{1,n}$). Let us apply the Delta method. We write $\phi : \mathbb{R}^2 \mapsto \mathbb{R}$ defined by $\phi(x,y) = y - x^2$. We have (**exercize**)

$$\widehat{\mu}_{2,n} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_1])^2 - \left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_1])\right)^2.$$

We write

$$Y_i = \begin{pmatrix} X_i - \mathbb{E}[X_1] \\ (X_i - \mathbb{E}[X_1])^2 \end{pmatrix}$$

such that $\widehat{\mu}_{2,n} = \phi\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \frac{1}{n}\sum_{i=1}^{n} Y_i^2\right)$. Also we have, since $(Y_i)_1$ is centered

$$\operatorname{cov}(Y_i) = \begin{pmatrix} \mu_2 & \mu_3 \\ \mu_3 & \mu_4 - \mu_2^2 \end{pmatrix}.$$

Hence from the central limit theorem

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \begin{pmatrix} 0 \\ \mu_2 \end{pmatrix}\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \begin{pmatrix} \mu_2 & \mu_3 \\ \mu_3 & \mu_4 - \mu_2^2 \end{pmatrix}\right).$$

Then from the Delta method

$$\sqrt{n}\left(\widehat{\mu}_{2,n} - \mu_2\right) = \sqrt{n}\left(\phi\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) - \phi(0,\mu_2)\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \begin{pmatrix} 0 & 1 \end{pmatrix}\begin{pmatrix} \mathbb{E}[X_1] & \mu_3 \\ \mu_3 & \mu_4 - \mu_2^2 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = \mathcal{N}(0, \mu_4 - \mu_2^2).$$

3 Statistical model and method of moments

3.1 Statistical model

Consider a sequence $(X_i)_{i\in\mathbb{N}}$ of i.i.d. random vectors of $\mathbb{R}^k$. We call a (parametric) **statistical model** a set of the form

$$\{\mathcal{L}_\theta; \theta \in \Theta\}$$

for $\Theta \subset \mathbb{R}^p$ where each $\mathcal{L}_\theta$ is a distribution on $\mathbb{R}^k$. It is a set of candidate distributions for the law of X_1.

We will make the assumption that the statistical model is well-specified and contains this law. Hence we assume that there is a $\theta_0 \in \mathring{\Theta}$ such that the distribution of X_1 is $\mathcal{L}_{\theta_0}$. The goal is to estimate θ_0 from $X_1, \ldots, X_n$.

We write $\mathbb{E}_\theta$, $\mathbb{P}_\theta$, $\operatorname{cov}_\theta$ for the expectation, probability and covariance computed "as if" we had $\theta_0 = \theta$. For instance

$$\mathbb{E}_\theta[\|X_1\|^2] = \int_{\mathbb{R}^k} \|x\|^2 \mathrm{d}\mathcal{L}_\theta(x)$$

and if $k = 1$ and $\mathcal{L}_\theta = \mathcal{N}(0,\theta)$ with $\Theta = (0,\infty)$, we have

$$\mathbb{E}_3[X_1^2] = \int_{\mathbb{R}} x^2 \mathrm{d}\mathcal{L}_3(x) = \underbrace{\mathbb{E}[Z^2]}_{Z\sim\mathcal{N}(0,3)} = 3.$$

Note that we still write $\mathbb{E}_{\theta_0} = \mathbb{E}$, $\mathbb{P}_{\theta_0} = \mathbb{P}$ and $\operatorname{cov}_{\theta_0} = \operatorname{cov}$ since $\mathcal{L}_{\theta_0}$ is "really" the distribution of $X_1, \ldots, X_n$.

3.2 Method of moments

Consider a sequence $(X_i)_{i\in\mathbb{N}}$ of i.i.d. random vectors of $\mathbb{R}^k$. Consider a statistical model

$$\{\mathcal{L}_\theta; \theta \in \Theta\}$$

for $\Theta \subset \mathbb{R}^p$ where each $\mathcal{L}_\theta$ is a distribution on $\mathbb{R}^k$. Assume that there is a $\theta_0 \in \mathring{\Theta}$ such that the distribution of X_1 is $\mathcal{L}_{\theta_0}$.

The idea of the **method of moments** is to choose p functions $f_1, \ldots, f_p : \mathbb{R}^k \to \mathbb{R}$ and to find a parameter θ such that the empirical moments and the theoretical moments are equal, that is

$$\begin{cases} \frac{1}{n}\sum_{i=1}^n f_1(X_i) = \mathbb{E}_\theta[f_1(X_1)] \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n f_p(X_i) = \mathbb{E}_\theta[f_p(X_1)] \end{cases}. \tag{5}$$

The idea is that as n is large the empirical moments are close to the theoretical ones, and if we have indentifiability from the k moments, that is, for $\theta \neq \theta_0$,

$$\begin{pmatrix} \mathbb{E}_\theta[f_1(X_1)] \\ \vdots \\ \mathbb{E}_\theta[f_p(X_1)] \end{pmatrix} \neq \begin{pmatrix} \mathbb{E}_{\theta_0}[f_1(X_1)] \\ \vdots \\ \mathbb{E}_{\theta_0}[f_p(X_1)] \end{pmatrix}$$

we hope that the θ selected by the method of moments will be close to θ_0.

Example 19. *Let $\Theta = \mathbb{R} \times [0, \infty)$, $\theta = (m, \sigma^2)$ and $\mathcal{L}_\theta = \mathcal{N}(m, \sigma^2)$. Let us consider the method of moments with $f_1(x) = x$ and $f_2(x) = x^2$. We have*

$$\mathbb{E}_\theta[f_1(X_1)] = \mathbb{E}_{Z\sim\mathcal{N}(m,\sigma^2)}[Z] = m$$

and

$$\mathbb{E}_\theta[f_2(X_1)] = \mathbb{E}_{Z\sim\mathcal{N}(m,\sigma^2)}[Z^2] = m^2 + \sigma^2.$$

Also we have

$$\frac{1}{n}\sum_{i=1}^n f_1(X_i) = \frac{\sum_{i=1}^n X_i}{n}$$

and

$$\frac{1}{n}\sum_{i=1}^n f_2(X_i) = \frac{\sum_{i=1}^n X_i^2}{n}.$$

Hence the estimators $\widehat{m}_n$ and $\widehat{\sigma}_n^2$ solve the system of equations

$$\begin{cases} \frac{\sum_{i=1}^n X_i}{n} = \widehat{m}_n \\ \frac{\sum_{i=1}^n X_i^2}{n} = \widehat{m}_n^2 + \widehat{\sigma}_n^2 \end{cases}.$$

We obtain the usual empirical mean and empirical variance estimators

$$\widehat{m}_n^2 = \frac{\sum_{i=1}^n X_i}{n}$$

and

$$\widehat{\sigma}_n^2 = \frac{\sum_{i=1}^n X_i^2}{n} - \left(\frac{\sum_{i=1}^n X_i}{n}\right)^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{m}_n)^2.$$

Theorem 20. *Let us define the function $e : \Theta \to \mathbb{R}^p$ by*

$$e(\theta) = \begin{pmatrix} \mathbb{E}_\theta[f_1(X_1)] \\ \vdots \\ \mathbb{E}_\theta[f_p(X_1)] \end{pmatrix}.$$

Assume that $\theta_0 \in \mathring{\Theta}$ and there is $\epsilon > 0$ such that $B(\theta_0, \epsilon) \subset \Theta$ and such that e is continuously differentiable on $B(\theta_0, \epsilon)$ with an invertible Jacobian matrix $Je(\theta_0)$ at θ_0. Assume also that for $j = 1, \dots, p$, $\mathbb{E}[|f_j(X_1)|^2] < \infty$.

Then, we can define a random vector $\widehat{\theta}_n$ that satisfies (5) with probability going to 1 as $n \to \infty$ and such that

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}\left(0, (Je(\theta_0))^{-1}\Sigma_f(Je(\theta_0))^{-\top}\right),$$

where Σ_f is the $p \times p$ covariance matrix of the random vector $(f_1(X_1), \dots, f_p(X_1))$.

When $p = 1$, we can interpret the asymptotic covariance matrix (here simply a variance) expression as follows. This variance is smaller (thus the method of moments works better) if the two following properties hold. (1) the derivative of $\theta \mapsto \mathbb{E}_\theta[f_1(X_1)]$ at θ_0 is large, which means that f_1 is a good function for **discriminating** between θ_0 and the other candidate parameters θ. (2) the variance of $f_1(X_1)$ is small so that the empirical and theoretical versions of $\mathbb{E}_{\theta_0}[f_1(X_1)]$ have a smaller difference.

Proof of Theorem 20. We will apply the **inverse function theorem** to the function e. This theorem states that there exist two neighborhoods U of θ_0 and V or $e(\theta_0)$ such that $e : U \to V$ is bijective with inverse function e^{-1}. Furthermore, e^{-1} is continuously differentiable on V and for $v = e(u) \in V$, we have

$$(Je^{-1})(v) = (Je(u))^{-1}.$$

Write

$$e_n = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n f_1(X_i) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n f_p(X_i) \end{pmatrix}$$

and note that $e_n \underset{n\to\infty}{\xrightarrow{p}} e(\theta_0)$ from the strong law of large number and Item 1 of Theorem 7. Hence $\mathbb{P}(e_n \in V) \to 1$ as $n \to \infty$ since $e(\theta_0)$ is in the interior of V. We thus define

$$\widehat{\theta}_n = \begin{cases} e^{-1}(e_n) & \text{if } e_n \in V \\ \text{arbitrary value} & \text{if } e_n \notin V \end{cases}$$

and then indeed $\widehat{\theta}_n$ satisfies (5) with probability going to 1 as $n \to \infty$. Let us define

$$\widetilde{e}_n = \begin{cases} e_n & \text{if } e_n \in V \\ e(\theta_0) & \text{if } e_n \notin V \end{cases}$$

and observe that for $\epsilon > 0$

$$\mathbb{P}\left[\left\|\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) - \sqrt{n}\left(e^{-1}(\widetilde{e}_n) - e^{-1}(e(\theta_0))\right)\right\| \geq \epsilon\right] \leq \mathbb{P}\left(e_n \notin V\right) \underset{n\to\infty}{\longrightarrow} 0.$$

Hence, from Item 4 in Theorem 7, it is sufficient to prove that

$$\sqrt{n}\left(e^{-1}(\widetilde{e}_n) - e^{-1}(e(\theta_0))\right) \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}\left(0, (Je(\theta_0))^{-1}\Sigma_f(Je(\theta_0))^{-\top}\right)$$

This is a consequence of the Delta method (Theorem 18). Indeed from the central limit theorem and Item 4 in Theorem 7 we have

$$\sqrt{n}\left(\widetilde{e}_n - e(\theta_0)\right) \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}\left(0, \Sigma_f\right)$$

and we have seen that

$$(Je^{-1})(e(\theta_0)) = (Je(\theta_0))^{-1}.$$

□

4 Consistency of M and Z-estimators

4.1 M-estimator

In general we wish to estimate a parameter θ in a parameter space $\Theta \subset \mathbb{R}^p$. The main example is where θ and Θ come from a statistical model as in Section 3.1, but we also allow for more general settings. Consider a sequence of random functions $(M_n)_{n\in\mathbb{N}}$ where for each $n \in \mathbb{N}$, M_n is a random function from Θ to $\mathbb{R}$. That is for all θ, $M_n(\theta)$ is a random variable and all the random variables $\{M_n(\theta); \theta \in \Theta\}$ are defined on the same probability space.

Then, a **M-estimator** is a sequence of random $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ taking values in Θ and maximizing M_n (hence the name). That is, for all $n \in \mathbb{N}$, a.s.[2]

$$\widehat{\theta}_n \in \underset{\theta\in\Theta}{argmax}\, M_n(\theta).$$

4.2 Maximum likelihood

Maximum likelihood estimators are the most important example of M-estimators in these lecture notes. We consider a statistical model $\{\mathcal{L}_\theta : \theta \in \Theta\}$ as in Section 3.1, where for all θ, $\mathcal{L}_\theta$ is a candidate distribution on $\mathbb{R}^k$ for the common law of $(X_i)_{i\in\mathbb{N}}$. We assume furthermore that for all θ, $\mathcal{L}_\theta$ has a density f_θ w.r.t. Lebesgue measure (this could be straightforwardly extended to a general measure μ). Then, since $X_1, \ldots, X_n$ are i.i.d, if θ was equal to θ_0, that is if $\mathcal{L}_\theta$ was the distribution of X_1, the density of the observation vector $(X_1, \ldots, X_n)$ would be equal to

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

This density, seen now as a function of θ after having observed $(X_1, \ldots, X_n)$ is called the **likelihood**. Taking the log facilitates the theoretical analysis and yields

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

which is called the **log-likelihood**. The maximum likelihood estimator consists in maximizing this log-likelihood (equivalently the likelihood) over Θ. It is thus a M-estimator defined by

$$\widehat{\theta}_n \in \underset{\theta\in\Theta}{argmax}\, M_n(\theta) \tag{6}$$

with

$$M_n(\theta) = \sum_{i=1}^n \log(f_\theta(X_i)). \tag{7}$$

4.3 Consistency of M-estimators

Theorem 21. *Consider a sequence $(M_n)_{n\in\mathbb{N}}$ of random functions from $\Theta \subset \mathbb{R}^p$ to $\mathbb{R}$. Consider a deterministic function $M : \Theta \to \mathbb{R}$. Assume that*

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \underset{n\to\infty}{\overset{p}{\longrightarrow}} 0 \tag{8}$$

and there is $\theta_0 \in \Theta$ such that

$$\forall \epsilon > 0, \qquad \sup_{\substack{\theta\in\Theta:\\ \|\theta-\theta_0\|\geq\epsilon}} M(\theta) < M(\theta_0). \tag{9}$$

[2]As will be seen from the mathematical statements below regarding M-estimators, we can allow for more flexibility that this "almost sure". It will be sufficient that these estimators maximize M_n with probability going to 1 or even up to a $o_{\mathbb{P}}(1)$.

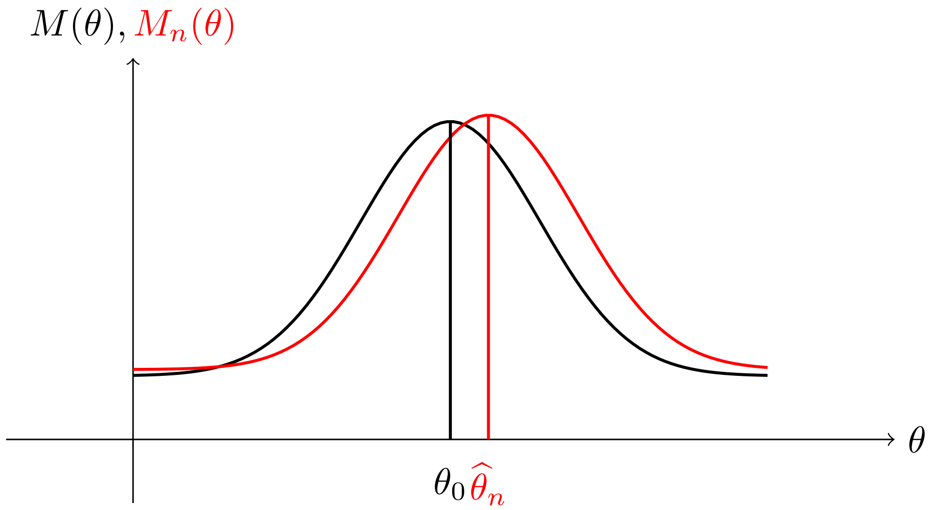

Figure 1: Illustration of Theorem 21. From the condition (8), the curves of M and M_n are uniformly close to each other. From the condition (9) the function M has a global maximum at θ_0 that is well-separated from the values taken away from θ_0. As a result of Theorem 21, the values of θ_0 and $\widehat{\theta}_n$ are close.

Consider a sequence $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ *such that*

$$M_n(\widehat{\theta}_n) \geq \left(\sup_{\theta\in\Theta} M_n(\theta) \right) + o_{\mathbb{P}}(1). \tag{10}$$

Then

$$\widehat{\theta}_n \xrightarrow[n\to\infty]{p} \theta_0.$$

In (8), M is the limit of M_n and the convergence must be uniform over θ and must hold in probability. Often, but not always, M_n is of the form

$$M_n = \sum_{i=1}^{n} m(X_i, \theta)$$

for i.i.d. $(X_i)_{i\in\mathbb{N}}$ and M is taken to be $M(\theta) = \mathbb{E}[m(X_1, \theta)]$. Then (9) means that not only the function M has a global maximum at θ_0 but also this maximum is well-separated from the values taken at parameters θ that are not close to θ_0. These two conditions (8) and (9) are illustrated in Figure 1. Finally, (10) provide the flexibility discussed above: $\widehat{\theta}_n$ needs not exactly maximize M_n, but only up to a margin $o_{\mathbb{P}}(1)$ (that goes to zero in probability as $n \to \infty$).

Proof of Theorem 21. Let $\epsilon > 0$ be fixed. We have

$$\mathbb{P}\left(\|\widehat{\theta}_n - \theta_0\| \geq \epsilon \right) \leq \mathbb{P}\left(M(\widehat{\theta}_n) \leq \sup_{\substack{\theta\in\Theta: \\ \|\theta-\theta_0\|\geq\epsilon}} M(\theta) \right). \tag{11}$$

Note that

$$\begin{aligned} M(\widehat{\theta}_n) &\geq M_n(\widehat{\theta}_n) - \sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \\ \text{(from (10):)} \quad &\geq M_n(\theta_0) + \sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| + o_{\mathbb{P}(1)} \\ &\geq M(\theta_0) - 2\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| + o_{\mathbb{P}(1)}. \end{aligned}$$

Hence back from (11) we obtain

$$\mathbb{P}\left(\|\widehat{\theta}_n - \theta_0\| \geq \epsilon\right) \leq \mathbb{P}\left(M(\theta_0) - 2\sup_{\theta\in\Theta}|M_n(\theta) - M(\theta)| + o_{\mathbb{P}(1)} \leq \sup_{\substack{\theta\in\Theta: \\ \|\theta-\theta_0\|\geq\epsilon}} M(\theta)\right)$$
$$= \mathbb{P}\left(2\sup_{\theta\in\Theta}|M_n(\theta) - M(\theta)| + o_{\mathbb{P}(1)} \leq M(\theta_0) - \sup_{\substack{\theta\in\Theta: \\ \|\theta-\theta_0\|\geq\epsilon}} M(\theta)\right).$$

Above, from (8), $2\sup_{\theta\in\Theta}|M_n(\theta) - M(\theta)| = o_{\mathbb{P}(1)}$ and from (9), $M(\theta_0) - \sup_{\substack{\theta\in\Theta: \\ \|\theta-\theta_0\|\geq\epsilon}} M(\theta) > 0$. Hence by definition of convergence in probability, the above probability goes to zero as $n \to \infty$.

□

4.4 Z-estimator

As for M-estimators, we wish to estimate a parameter θ in a parameter space $\Theta \subset \mathbb{R}^p$. Consider a sequence of random functions $(Z_n)_{n\in\mathbb{N}}$ where for each $n \in \mathbb{N}$, Z_n is a random function from Θ to $\mathbb{R}^q$ for a given $q \in \mathbb{N}$. Then, a **Z-estimator** is a sequence of random $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ taking values in Θ and setting Z_n to zero (hence the name). That is, for all $n \in \mathbb{N}$, a.s.[3]

$$Z_n(\widehat{\theta}_n) = 0.$$

Consider a M-estimator given by the function M_n and assume further that Θ is open and that for all $n \in \mathbb{N}$, and $\theta \in \Theta$, a.s, M_n is differentiable at θ. Then if

$$\widehat{\theta}_n \in \underset{\theta\in\Theta}{argmax}\, M_n(\theta),$$

we have a.s.

$$\nabla M_n(\widehat{\theta}_n) = 0$$

and thus in this case, the M-estimator is also a Z-estimator with Z_n taking values in $\mathbb{R}^p$.

4.5 Consistency of Z-estimators

The next theorem can be interpreted as having similarities with Theorem 21 for M-estimators.

Theorem 22. *Consider a sequence $(Z_n)_{n\in\mathbb{N}}$ of random functions from $\Theta \subset \mathbb{R}^p$ to $\mathbb{R}^q$. Consider a deterministic function $Z : \Theta \to \mathbb{R}^q$. Assume that*

$$\sup_{\theta\in\Theta}\|Z_n(\theta) - Z(\theta)\| \xrightarrow[n\to\infty]{p} 0 \tag{12}$$

and

$$\forall\epsilon > 0, \qquad \inf_{\substack{\theta\in\Theta: \\ \|\theta-\theta_0\|\geq\epsilon}} \|Z(\theta)\| > 0 = Z(\theta_0). \tag{13}$$

Consider a sequence $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ such that

$$Z_n(\widehat{\theta}_n) = o_{\mathbb{P}}(1). \tag{14}$$

Then

$$\widehat{\theta}_n \xrightarrow[n\to\infty]{p} \theta_0.$$

[3] As for M-estimators, we can allow for more flexibility that this "almost sure".

Proof. Let $\epsilon > 0$ be fixed. We have

$$\mathbb{P}\left(\|\widehat{\theta}_n - \theta_0\| \geq \epsilon\right) \leq \mathbb{P}\left(\|Z(\widehat{\theta}_n)\| \geq \inf_{\substack{\theta\in\Theta:\\ \|\theta-\theta_0\|\geq\epsilon}} \|Z(\theta)\|\right). \tag{15}$$

Note that

$$\begin{aligned} \|Z(\widehat{\theta}_n)\| &\leq \|Z_n(\widehat{\theta}_n)\| + \sup_{\theta\in\Theta} |\|Z_n(\theta)\| - \|Z(\theta)\|| \\ \text{(from (14):)} \quad &\leq o_{\mathbb{P}(1)} + \sup_{\theta\in\Theta} \|Z_n(\theta) - Z(\theta)\|. \end{aligned}$$

Hence back from (15) we obtain

$$\mathbb{P}\left(\|\widehat{\theta}_n - \theta_0\| \geq \epsilon\right) \leq \mathbb{P}\left(o_{\mathbb{P}(1)} + \sup_{\theta\in\Theta} \|Z_n(\theta) - Z(\theta)\| \geq \inf_{\substack{\theta\in\Theta:\\ \|\theta-\theta_0\|\geq\epsilon}} \|Z(\theta)\|\right).$$

Above, from (12), $\sup_{\theta\in\Theta} \|Z_n(\theta) - Z(\theta)\| = o_{\mathbb{P}(1)}$ and from (13), $\inf_{\substack{\theta\in\Theta:\\ \|\theta-\theta_0\|\geq\epsilon}} \|Z(\theta)\| > 0$. Hence by definition of convergence in probability, the above probability goes to zero as $n \to \infty$.

□

The next theorem is an example where we can relax the condition (12) of uniform convergence of Z_n to Z, in the one-dimensional case $\Theta \subset \mathbb{R}$.

Proposition 23. *Let $\Theta = \mathbb{R}$. Consider a sequence $(Z_n)_{n\in\mathbb{N}}$ of random functions from Θ to $\mathbb{R}$. Consider a deterministic function $Z : \Theta \to \mathbb{R}$. Assume that*

1. *For all fixed $\theta \in \Theta$, $Z_n(\theta) \xrightarrow[n\to\infty]{p} Z(\theta)$;*
2. *Z_n is non-decreasing;*
3. *There is a fixed θ_0 such that for all $\epsilon > 0$, $Z(\theta_0 - \epsilon) < 0 < Z(\theta_0 + \epsilon)$.*

Consider a sequence $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ such that

$$Z_n(\widehat{\theta}_n) = o_{\mathbb{P}}(1). \tag{16}$$

Then

$$\widehat{\theta}_n \xrightarrow[n\to\infty]{p} \theta_0.$$

Proof. Let $\epsilon > 0$ be fixed. We have

$$\begin{aligned} \mathbb{P}\left(|\widehat{\theta}_n - \theta_0| \geq \epsilon\right) &= \mathbb{P}\left(\widehat{\theta}_n \leq \theta_0 - \epsilon\right) + \mathbb{P}\left(\widehat{\theta}_n \geq \theta_0 + \epsilon\right) \\ (Z_n \text{ is non-decreasing:}) &\leq \mathbb{P}\left(Z_n(\widehat{\theta}_n) \leq Z_n(\theta_0 - \epsilon)\right) + \mathbb{P}\left(Z_n(\widehat{\theta}_n) \geq Z_n(\theta_0 + \epsilon)\right) \\ \text{(from (16):)} &= \mathbb{P}\left(o_{\mathbb{P}}(1) \leq Z_n(\theta_0 - \epsilon)\right) + \mathbb{P}\left(o_{\mathbb{P}}(1) \geq Z_n(\theta_0 + \epsilon)\right) \\ &= \mathbb{P}\left(o_{\mathbb{P}}(1) \leq Z(\theta_0 - \epsilon) + Z_n(\theta_0 - \epsilon) - Z(\theta_0 - \epsilon)\right) \\ &\quad + \mathbb{P}\left(o_{\mathbb{P}}(1) \geq Z(\theta_0 + \epsilon) + Z_n(\theta_0 + \epsilon) - Z(\theta_0 + \epsilon)\right) \\ \text{(from Item 1:)} &= \mathbb{P}\left(o_{\mathbb{P}}(1) \leq Z(\theta_0 - \epsilon)\right) + \mathbb{P}\left(o_{\mathbb{P}}(1) \geq Z(\theta_0 + \epsilon)\right). \end{aligned}$$

The two above probabilities go to zero by definition of $o_{\mathbb{P}}(1)$ and from Item 3. Hence indeed $\widehat{\theta}_n \xrightarrow[n\to\infty]{p} \theta_0$. □

Let us provide an example to Proposition 23 by considering the empirical median. Consider *i.i.d.* random variables $(X_i)_{i\in\mathbb{N}}$ having a density with respect to Lebesgue measure. Define the **empirical median** as a random variable $\widehat{\theta}_n$ satisfying

$$\sum_{i=1}^{n} \text{sign}(\widehat{\theta}_n - X_i) = 0.$$

Write the order statistic of $X_1, \ldots, X_n$ as $X_{(1)} \leq \cdots \leq X_{(n)}$ with $\{X_1, \ldots, X_n\} = \{X_{(1)}, \ldots, X_{(n)}\}$. Note that a.s. $X_1, \ldots, X_n$ are two-by-two distinct and thus if $n = 2m$ (even number), $\widehat{\theta}_n$ is any number θ satisfying $X_{(m)} < \theta < X_{(m+1)}$ and if $n = 2m+1$ (odd number), then $\widehat{\theta}_n = X_{(m+1)}$. Also, assume that F_{X_1} is strictly increasing on $\mathbb{R}$, such that there is a unique population median such that $F_{X_1}(\theta_0) = 1/2$.

Let us apply Proposition 23 to show that $\widehat{\theta}_n \underset{n\to\infty}{\overset{p}{\longrightarrow}} \theta_0$. We write

$$Z_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \text{sign}(\theta - X_i)$$

and

$$Z(\theta) = F_{X_1}(\theta) - (1 - F_{X_1(\theta)}).$$

For all fixed θ, by the strong law of large number

$$\begin{aligned}
Z_n(\theta) =& \frac{1}{n}\sum_{i=1}^{n} \text{sign}(\theta - X_i) \\
=& \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\theta - X_i > 0\} - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\theta - X_i < 0\} \\
=& \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i < \theta\} - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i > \theta\} \\
\underset{n\to\infty}{\overset{p}{\longrightarrow}}& \mathbb{P}(X_1 < \theta) - \mathbb{P}(X_1 > \theta) \\
\text{(since } \mathbb{P}(X_1 = \theta) = 0\text{:)} \quad =& F_{X_1}(\theta) - (1 - F_{X_1}(\theta)) \\
=& Z(\theta),
\end{aligned}$$

hence Item 1 holds in Proposition 23. Item 2 also holds because $\theta \mapsto \text{sign}(\theta - X_i)$ is non-decreasing. Item 3 also holds because $Z(\theta)$ is strictly increasing on $\mathbb{R}$ because F_{X_1} is strictly increasing. Finally (16) holds because $Z_n(\widehat{\theta}_n) = 0$. Hence from Proposition 23, indeed $\widehat{\theta}_n \underset{n\to\infty}{\overset{p}{\longrightarrow}} \theta_0$.

5 Bracketing number for uniform convergence

5.1 Obtaining uniform convergence

To apply Theorems 21 and 22, a potentially challenging requirement is to obtain **uniform convergence**, that is to show

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \underset{n\to\infty}{\overset{p}{\longrightarrow}} 0 \quad \text{and} \quad \sup_{\theta\in\Theta} \|Z_n(\theta) - Z(\theta)\| \underset{n\to\infty}{\overset{p}{\longrightarrow}} 0.$$

Considering the case of M-estimators, we will provide tools to obtain this uniform convergence in the cases where $(X_i)_{i\in\mathbb{N}}$ are i.i.d., where M_n is of the form

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m(X_i, \theta),$$

for a function $m : \mathbb{R}^k \times \Theta \to \mathbb{R}$, and where

$$M(\theta) = \mathbb{E}[m(X_1, \theta)].$$

.

In this case, we have, with $m_\theta(\cdot) = m(\cdot, \theta)$,

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| = \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n m_\theta(X_i) - \mathbb{E}[m_\theta(X_1)] \right|$$

which is the supremum over a set of functions of differences between the empirical means of these functions and the corresponding theoretical means.

We will address this supremum in a more general abstract setting with a set $\mathcal{F}$ of functions from $\mathbb{R}^k$ to $\mathbb{R}$ such that for all $f \in \mathcal{F}$, $\mathbb{E}[|f(X_1)|] < \infty$. Since the supremum obviously increases with the set $\mathcal{F}$ (with the inclusion relationship), we will define a suitable measure of **size** or **complexity** for $\mathcal{F}$. This measure will be called the **bracketing number**.

Definition 24 (Bracketing number)**.** *Consider ℓ and u two functions from $\mathbb{R}^k$ to $\mathbb{R}$ such that for all $x \in \mathbb{R}^k$ $\ell(x) \leq u(x)$. We define the* **bracket**

$$[\ell, u] = \left\{ f : \mathbb{R}^k \to \mathbb{R} : \quad \forall x \in \mathbb{R}^k, \quad \ell(x) \leq f(x) \leq u(x) \right\}.$$

Then for $\epsilon > 0$, for $q > 0$ and for a measure $\mathcal{L}$ on $\mathbb{R}^k$, we define the **bracketing number** *$\mathcal{N}_{[]}(\mathcal{F}, L^q(\mathcal{L}), \epsilon)$ as the smallest number of brackets that enable to cover $\mathcal{F}$. More precisely*

$$\mathcal{N}_{[]}(\mathcal{F}, L^q(\mathcal{L}), \epsilon) = \min_{N \in \mathbb{N}} \Big\{ \exists [\ell_1, u_1], \ldots, [\ell_N, u_N] : \quad \forall j \in \{1, \ldots, N\}, \left(\int_{\mathbb{R}^k} (u_j - \ell_j)^q \mathrm{d}\mathcal{L} \right)^{1/q} \leq \epsilon, \quad \mathcal{F} \subset \cup_{j=1}^N [\ell_j, u_j] \Big\}. \tag{17}$$

The quantity $\mathcal{N}_{[]}(\mathcal{F}, L^q(\mathcal{L}), \epsilon)$ decreases with ϵ and typically goes to ∞ as $\epsilon \to 0$.

Definition 25. *Consider a set of functions $\mathcal{F} : \mathbb{R}^k \to \mathbb{R}$ and a distribution $\mathcal{L}$ on $\mathbb{R}^k$, we say that $\mathcal{F}$ is* **$\mathcal{L}$-Glivenko-Cantelli** *if for all $f \in \mathcal{F}$, $\int_{\mathbb{R}^k} |f| \mathrm{d}\mathcal{L} < \infty$ and for i.i.d. $(X_i)_{i \in \mathbb{N}}$ with distribution $\mathcal{L}$,*

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_1)] \right| = o_{\mathbb{P}}(1).$$

The next proposition establishes an important relationship between the bracketing number and the $\mathcal{L}$-Glivenko-Cantelli property.

Proposition 26. *Consider a set of functions $\mathcal{F} : \mathbb{R}^k \to \mathbb{R}$ and a distribution $\mathcal{L}$ on $\mathbb{R}^k$, such that for all $f \in \mathcal{F}$, $\int_{\mathbb{R}^k} |f| \mathrm{d}\mathcal{L} < \infty$ and*

$$\forall \epsilon > 0, \quad \mathcal{N}_{[]}(\mathcal{F}, L^1(\mathcal{L}), \epsilon) < \infty.$$

Then $\mathcal{F}$ is $\mathcal{L}$-Glivenko-Cantelli.

Proof. Let $\epsilon > 0$, $N = \mathcal{N}_{[]}(\mathcal{F}, L^1(\mathcal{L}), \epsilon) < \infty$ and $[\ell_1, u_1], \ldots, [\ell_N, u_N]$ some brackets such that for $j \in \{1, \ldots, N\}$, $\int_{\mathbb{R}^k} |u_j - \ell_j| \mathrm{d}\mathcal{L} \leq \epsilon$ and $\mathcal{F} \subset \cup_{j=1}^N [\ell_j, u_j]$. Then, for all $f \in \mathcal{F}$, there is $j \in \{1, \ldots, N\}$ such that

$$\frac{1}{n} \sum_{i=1}^n \ell_j(X_i) \leq \frac{1}{n} \sum_{i=1}^n f(X_i) \leq \frac{1}{n} \sum_{i=1}^n u_j(X_i), \tag{18}$$

and, since $\mathbb{E}[u_j(X_1)] - \mathbb{E}[\ell_j(X_1)] \leq \mathbb{E}[|\ell_j(X_1) - u_j(X_1)|] \leq \epsilon$,

$$\mathbb{E}[\ell_j(X_1)] \leq \mathbb{E}[f(X_1)] \leq \mathbb{E}[u_j(X_1)] \leq \mathbb{E}[\ell_j(X_1)] + \epsilon. \tag{19}$$

From (18) and $\ell_j \leq f \leq u_j$, we have

$$\frac{1}{n} \sum_{i=1}^n \ell_j(X_i) - \mathbb{E}[u_j(X_1)] \leq \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_1)] \leq \frac{1}{n} \sum_{i=1}^n u_j(X_i) - \mathbb{E}[\ell_j(X_1)].$$

Then (19) yields

$$\frac{1}{n}\sum_{i=1}^{n}\ell_j(X_i)-\mathbb{E}[\ell_j(X_1)]-\epsilon\leq\frac{1}{n}\sum_{i=1}^{n}f(X_i)-\mathbb{E}[f(X_1)]\leq\frac{1}{n}\sum_{i=1}^{n}u_j(X_i)-\mathbb{E}[u_j(X_1)]+\epsilon.$$

Hence

$$\left|\frac{1}{n}\sum_{i=1}^{n}f(X_i)-\mathbb{E}[f(X_1)]\right|\leq\max_{j=1,\dots,N}\max\left(\left|\frac{1}{n}\sum_{i=1}^{n}\ell_j(X_i)-\mathbb{E}[\ell_j(X_1)]\right|,\left|\frac{1}{n}\sum_{i=1}^{n}u_j(X_i)-\mathbb{E}[u_j(X_1)]\right|\right)+\epsilon$$

and thus

$$\begin{aligned}&\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(X_i)-\mathbb{E}[f(X_1)]\right|\geq 2\epsilon\right)\\&\leq\mathbb{P}\left(\max_{j=1,\dots,N}\max\left(\left|\frac{1}{n}\sum_{i=1}^{n}\ell_j(X_i)-\mathbb{E}[\ell_j(X_1)]\right|,\left|\frac{1}{n}\sum_{i=1}^{n}u_j(X_i)-\mathbb{E}[u_j(X_1)]\right|\right)\geq\epsilon\right).\end{aligned}$$

Above, there is a finite maximum of terms of the form $\frac{1}{n}\sum_{i=1}^{n}g(X_i)-\mathbb{E}[g(X_1)]$ with $\mathbb{E}[|g(X_1)|]<\infty$. Hence (**exercize**) from the strong law of large number, this probability goes to 0 as $n\to\infty$. Since this holds for all $\epsilon>0$, we indeed have

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(X_i)-\mathbb{E}[f(X_1)]\right|=o_{\mathbb{P}}(1).$$

□

Next is a simple example of application of Proposition 26.

Proposition 27. *Let $\mathcal{L}$ be a distribution on $\mathbb{R}^k$, let $\mathcal{F}=\{g_\theta;\theta\in\Theta\}$ where $g_\theta:\mathbb{R}^k\to\mathbb{R}$ and assume that*

1. *Θ is a compact set of a metric space;*
2. *for all $x\in\mathbb{R}^k$, $\theta\mapsto g_\theta(x)$ is continuous;*
3. $\int_{\mathbb{R}^k}\sup_{\theta\in\Theta}|g_\theta(x)|\mathrm{d}\mathcal{L}(x)<\infty.$

Then $\mathcal{F}$ is $\mathcal{L}$-Glivenko-Cantelli.

Proof. Let us show that for all $\epsilon>0$, $\mathcal{N}_{[]}(\mathcal{F},L^1(\mathcal{L}),\epsilon)<\infty$ in order to apply Proposition 26. Fix $\epsilon>0$. Let $\mathrm{dist}:\Theta^2\to\mathbb{R}^+$ be the distance on Θ. For $\theta\in\Theta$, consider the sequence of sets $(B_{\theta,N})_{N\in\mathbb{N}}$ with $B_{\theta,N}=B(\theta,\frac{1}{N})=\{\widetilde{\theta}\in\Theta:\mathrm{dist}(\theta,\widetilde{\theta})<\frac{1}{N}\}$ (open balls with the metric of Θ).

For all N, we write

$$\widetilde{\ell}_{\theta,N}(x)=\inf_{\widetilde{\theta}\in B_{\theta,N}}g_{\widetilde{\theta}}(x)$$

and

$$\widetilde{u}_{\theta,N}(x)=\sup_{\widetilde{\theta}\in B_{\theta,N}}g_{\widetilde{\theta}}(x).$$

For every fixed $x\in\mathbb{R}^k$, $\widetilde{u}_{\theta,N}(x)-\widetilde{\ell}_{\theta,N}(x)\to 0$ as $N\to\infty$ since $\theta\mapsto g_\theta(x)$ is continuous. Furthermore, for all $N\in\mathbb{N}$

$$\widetilde{u}_{\theta,N}-\widetilde{\ell}_{\theta,N}\leq 2\sup_{\theta\in\Theta}|g_\theta|$$

and thus $\int_{\mathbb{R}^k}\sup_{N\in\mathbb{N}}\left|\widetilde{u}_{\theta,N}-\widetilde{\ell}_{\theta,N}\right|\mathrm{d}\mathcal{L}<\infty$. Hence by dominated convergence

$$\int_{\mathbb{R}^k}\left|\widetilde{u}_{\theta,N}-\widetilde{\ell}_{\theta,N}\right|\mathrm{d}\mathcal{L}\underset{N\to\infty}{\longrightarrow}0.$$

Hence there exists $N_\theta \in \mathbb{N}$ such that $\int_{\mathbb{R}^k} \left| \widetilde{u}_{\theta,N_\theta} - \widetilde{\ell}_{\theta,N_\theta} \right| \mathrm{d}\mathcal{L} \leq \epsilon$. We fix this value N_θ for the rest of the proof.

Now, the set $\{\cup_{\theta\in\Theta} B_{\theta,N_\theta}\}$ is a union of open sets that contains Θ. Now we use the following property of compact spaces (that can also be the definition of compacity)

- *For a compact set K in a metric space E, for every set of open sets of E, $\mathcal{C} = \{E'; E' \in \mathcal{C}\}$ that covers K*

$$K \subset \cup_{E'\in\mathcal{C}} E'$$

there exists a **finite** *subset $\mathcal{C}'$ of $\mathcal{C}$ such that*

$$K \subset \cup_{E'\in\mathcal{C}'} E'.$$

We apply this property to the set $\{\cup_{\theta\in\Theta} B_{\theta,N_\theta}\}$ that covers the compact set Θ. Hence there exist $\theta_1, \ldots, \theta_m$ such that

$$\Theta \subset \cup_{j=1}^m B_{\theta_j,N_{\theta_j}}.$$

We define for $j = 1, \ldots, m$ and $x \in \mathbb{R}^k$

$$\ell_j(x) = \inf_{\widetilde{\theta}\in B_{\theta_j,N_{\theta_j}}} g_{\widetilde{\theta}}(x) = \widetilde{\ell}_{\theta_j,N_{\theta_j}}$$

and

$$u_j(x) = \sup_{\widetilde{\theta}\in B_{\theta_j,N_{\theta_j}}} g_{\widetilde{\theta}}(x) = \widetilde{u}_{\theta_j,N_{\theta_j}}.$$

From the above choice of N_{θ_j}, we have $\ell_j \leq u_j$ and $\int_{\mathbb{R}^k} |u_j - \ell_j| \mathrm{d}\mathcal{L} \leq \epsilon$. For any $\theta \in \Theta$, there is $j = \{1, \ldots, m\}$ such that $\theta \in B_{\theta_j,N_{\theta_j}}$ and thus $g_\theta \in [\ell_j, u_j]$. Hence we have found m brackets such that the property in the min in (17) holds. Hence $\mathcal{N}_{[]}(\mathcal{F}, L^1(\mathcal{L}), \epsilon) < \infty$ and thus we can conclude from Proposition 26. □

5.2 Application to maximum likelihood

We consider the setting of Section 4.2 (maximum likelihood). The following theorem provides the consistency of maximum likelihood, under (quite non-restrictive) regularity conditions.

Theorem 28. *Consider the context of Section 4.2 where there is a set $\{\mathcal{L}_\theta; \theta \in \Theta\}$ of distributions on $\mathbb{R}^k$, with $\mathcal{L}_\theta$ having density f_θ with respect to Lebesgue measure, and where there are $(X_i)_{i\in\mathbb{N}}$ i.i.d. with density f_{θ_0} for $\theta_0 \in \Theta$. Assume that*

1. *Θ is compact in $\mathbb{R}^p$;*
2. *For all $\theta \in \Theta$ and $x \in \mathbb{R}^k$, $f_\theta(x) > 0$;*
3. *For all $x \in \mathbb{R}^k$, $\theta \mapsto f_\theta(x)$ is continuous on Θ;*
4. *$\int_{\mathbb{R}^k} \sup_{\theta\in\Theta} |\log(f_\theta(x))| f_{\theta_0}(x)\mathrm{d}x < \infty$;*
5. *for all $\theta \neq \theta_0$, the distributions $\mathcal{L}_\theta$ and $\mathcal{L}_{\theta_0}$ are different.*

Then $\widehat{\theta}_n$ defined in (6) and (7) satisfies

$$\widehat{\theta}_n \xrightarrow[n\to\infty]{p} \theta_0.$$

Note that Item 5 is called an **identifiability** condition. It is clearly necessary since if $\mathcal{L}_\theta = \mathcal{L}_{\theta_0}$ the observations $(X_i)_{i\in\mathbb{N}}$ are distributed both as $\mathcal{L}_\theta$ and $\mathcal{L}_{\theta_0}$.

Proof. Let us first show that $\{\log(f_\theta); \theta \in \Theta\}$ is $\mathcal{L}_{\theta_0}$-Glivenko-Cantelli using Proposition 27. In this proposition, Item 1 holds by assumption. We let $g_\theta = \log(f_\theta)$ Item 2 in the proposition hold from Items 2 and 3 of the theorem. Finally Item 3 in the proposition holds from Item 4 in the theorem since $\mathrm{d}\mathcal{L}_{\theta_0}(x) = f_{\theta_0}\mathrm{d}x$. Thus Proposition 27 holds and by definition of being $\mathcal{L}_{\theta_0}$-Glivenko-Cantelli, we have

$$\sup_{\theta\in\Theta}\left|\sum_{i=1}^{n}\log(f_\theta(X_i)) - \mathbb{E}[\log(f_\theta(X_1))]\right| \xrightarrow[n\to\infty]{p} 0.$$

The aim now is to apply Theorem 21, and we have just shown that the condition (8) holds, choosing

$$M(\theta) = \mathbb{E}[\log(f_\theta(X_1))].$$

Also the condition (10) holds from (6). It remains to prove (9).

For $\theta \neq \theta_0$,

$$\begin{aligned} M(\theta) - M(\theta_0) =& \mathbb{E}[\log(f_\theta(X_1))] - \mathbb{E}[\log(f_{\theta_0}(X_1))] \\ =& \int_{\mathbb{R}^k} \log(f_\theta(x)) f_{\theta_0}(x)\mathrm{d}x - \int_{\mathbb{R}^k} \log(f_{\theta_0}(x)) f_{\theta_0}(x)\mathrm{d}x \\ =& \int_{\mathbb{R}^k} \log\left(\frac{f_\theta(x)}{f_{\theta_0}(x)}\right) f_{\theta_0}(x)\mathrm{d}x. \end{aligned}$$

Note that all integrals above are well-defined from Item 4 in the theorem statement. We then use the inequality $\log(t) \leq 2(\sqrt{t} - 1)$ for $t > 0$. This yields

$$\begin{aligned} M(\theta) - M(\theta_0) \leq& 2\int_{\mathbb{R}^k}\left(\sqrt{\frac{f_\theta(x)}{f_{\theta_0}(x)}} - 1\right) f_{\theta_0}(x)\mathrm{d}x \\ =& 2\int_{\mathbb{R}^k}\sqrt{f_\theta(x)}\sqrt{f_{\theta_0}(x)}\mathrm{d}x - 2\int_{\mathbb{R}^k} f_{\theta_0}(x)\mathrm{d}x \\ =& 2\int_{\mathbb{R}^k}\sqrt{f_\theta(x)}\sqrt{f_{\theta_0}(x)}\mathrm{d}x - \int_{\mathbb{R}^k} f_{\theta_0}(x)\mathrm{d}x - \int_{\mathbb{R}^k} f_\theta(x)\mathrm{d}x \\ =& -\int_{\mathbb{R}^k}\left(\sqrt{f_\theta(x)} - \sqrt{f_{\theta_0}(x)}\right)^2 \mathrm{d}x \\ <& \, 0 \end{aligned}$$

since the distributions $\mathcal{L}_\theta$ and $\mathcal{L}_{\theta_0}$ are different from Item 5 in the theorem statement.

Next, M is a continuous function on Θ by dominated convergence, because $\theta \mapsto \log(f_\theta(x))$ is continuous for all x from Items 2 and 3 and because Item 4 yields the domination by an integrable function. Hence, by compacity of Θ, (9) holds. Hence we can apply Theorem 21 and conclude. □

6 Asymptotic normality of Z-estimators

6.1 Some intuition

In this section we consider a Z-estimator $\widehat{\theta}_n$ satisfying

$$\frac{1}{n}\sum_{i=1}^{n} z(X_i, \widehat{\theta}_n) = 0$$

for i.i.d. $(X_i)_{i\in\mathbb{N}}$ and for a function $z : \mathbb{R}^k \times \Theta \mapsto \mathbb{R}^p$ with $\Theta \subseteq \mathbb{R}^p$. We assume that there is $\theta_0 \in \Theta$ such that $\mathbb{E}[z(Z_1, \theta_0)] = 0$ and that we have already proved (from Section 4 for instance) that $\widehat{\theta}_n \xrightarrow[n\to\infty]{p} \theta_0$. The aim of this section is to show the asymptotic normality of

$$\sqrt{n}(\widehat{\theta}_n - \theta_0).$$

Assuming enough smoothness, we could write a Taylor expansion of

$$\theta \mapsto Z_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} z(X_i, \widehat{\theta})$$

around θ_0:

$$0 = Z_n(\widehat{\theta}_n) \approx Z_n(\theta_0) + (JZ_n)(\theta_0)\left(\widehat{\theta}_n - \theta_0\right),$$

where JZ_n is the random Jacobian matrix of $\theta \mapsto Z_n(\theta)$ Asymptotically, the $p \times p$ matrix $JZ_n(\theta_0)$ is expected to be close to $\mathbb{E}[J_z(X_1, \theta_0)]$, where for $x \in \mathbb{R}^k$ and $\theta \in \Theta$, $J_z(x, \theta_0)$ is the $p \times p$ matrix defined by $J_z(x,\theta)_{k,\ell} = \dfrac{\partial z(x,\theta)_k}{\partial \theta_\ell}$. If this matrix $\mathbb{E}[J_z(X_1, \theta_0)]$ is invertible, then the matrix $(JZ_n)(\theta_0)$ is invertible with probability going to one and we would have

$$0 = (JZ_n)(\theta_0)^{-1} Z_n(\theta_0) + \left(\widehat{\theta}_n - \theta_0\right)$$

and thus

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = -(JZ_n)(\theta_0)^{-1}\left(\sqrt{n} Z_n(\theta_0)\right).$$

From the central limit theorem and because $\mathbb{E}[z(X_1, \theta_0)] = 0$, $\sqrt{n}Z_n(\theta_0)$ converges in distribution to

$$\mathcal{N}\left(0, \operatorname{cov}\left(z(X_1, \theta_0)\right)\right).$$

Hence from Slutsky lemma we would have

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}\left(0, \mathbb{E}[J_z(X_1,\theta_0)]^{-1} \operatorname{cov}\left(z(X_1, \theta_0)\right) \mathbb{E}[J_z(X_1,\theta_0)]^{-\top}\right).$$

It is possible to obtain a rigorous mathematical statement and proof from this intuition above, but with strong smoothness condition on $z(x, \theta)$ for fixed x. In the next section, we instead present a proof that is more involved, but needs only mild smoothness assumptions. In particular, it will allow us to address the asymptotic normality of the empirical median (Section 4.5), given by $z(x, \theta) = \operatorname{sign}(\theta - x)$, the function z not being differentiable w.r.t. θ for fixed x.

6.2 The main result

We will use the following tool, that enables to bound a quantity of the form

$$\sup_{f\in\mathcal{F}} \frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}\left(f(X_i) - \mathbb{E}[f(X_1)]\right)\right|$$

for i.i.d. $(X_i)_{i\in\mathbb{N}}$ on $\mathbb{R}^k$ and for a set $\mathcal{F}$ of functions from $\mathbb{R}^k$ to $\mathbb{R}$. Note that if $\mathcal{F} = \{f\}$ is a singleton, this quantity is bounded in probability by the central limit theorem. The interest of the next theorem, called a **maximal inequality**, is to allow for infinite sets $\mathcal{F}$.

Theorem 29. *Let $(X_i)_{i\in\mathbb{N}}$ be i.i.d. on $\mathbb{R}^k$ with distribution $\mathcal{L}$. Consider a set $\mathcal{F}$ of functions from $\mathbb{R}^k$ to $\mathbb{R}$ such that there is a function F such that*

$$\text{for all } f \in \mathcal{F}, \text{ for } \mathcal{L}\text{-almost all } x \in \mathbb{R}^k \quad |f(x)| \leq F(x)$$

with

$$\mathbb{E}[F(X_1)^2] < \infty.$$

Then, with the Bracketing number $\mathcal{N}_{[]}(\mathcal{F}, L^2(\mathcal{L}), \epsilon)$ defined in Definition 24,

$$\mathbb{E}^\star\left[\sup_{f\in\mathcal{F}} \frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}\left(f(X_i) - \mathbb{E}[f(X_1)]\right)\right|\right] \leq C_{MI}\int_0^{\sqrt{\mathbb{E}[F(X_1)^2]}} \sqrt{\log\left(\mathcal{N}_{[]}(\mathcal{F}, L^2(\mathcal{L}), \epsilon)\right)}\mathrm{d}\epsilon,$$

for a universal constant C_{MI}.

Proof. We skip this proof in the lecture notes. We refer to Corollary 19.35 in [VdV07]. □

Above, the star in $\mathbb{E}^\star$ means that the sup is allowed to be non-measurable. In this case, we define the expectation as an outer expectation (see Section 18.2 in [VdV07]). We shall not worry about this since this $\mathbb{E}^\star$ will serve to bound expectations or probabilities for measurable quantities.

We can now provide the general asymptotic normality result for Z-estimators.

Theorem 30. *Let $(X_i)_{i\in\mathbb{N}}$ be i.i.d. on $\mathbb{R}^k$ with distribution $\mathcal{L}$.*

1. *Consider consider a Z-estimator $\widehat{\theta}_n$ satisfying*
$$\frac{1}{\sqrt{n}}\sum_{i=1}^n z(X_i,\widehat{\theta}_n) = o_{\mathbb{P}}(1) \tag{20}$$
with $z : \mathbb{R}^k \times \Theta \to \mathbb{R}^p$ satisfying $\mathbb{E}[\|z(X_1,\theta)\|^2] < \infty$ for all $\theta \in \Theta$. Assume that there is $\theta_0 \in \mathring{\Theta}$ such that $\mathbb{E}[z(X_1,\theta_0)] = 0$ and $\widehat{\theta}_n \xrightarrow[n\to\infty]{p} \theta_0$.

2. *Assume that there is a neighborhood A of θ_0 such that $\theta \mapsto \mathbb{E}[z(X_1,\theta)]$ is continuously differentiable on A. We write $J\mathbb{E}[z(X_1,\theta)]$ for its $p\times p$ Jacobian matrix at θ. Assume that $J\mathbb{E}[z(X_1,\theta_0)]$ is invertible.*

3. *For $j = 1,\dots,p$ let $\mathcal{F}_j = \{\mathbb{R}^k \ni x \mapsto z(x,\theta)_j; \theta \in A\}$. Assume that for all $0 < \delta < \infty$,*
$$\int_0^\delta \sqrt{\log\left(\mathcal{N}_{[]}(\mathcal{F}_j, L^2(\mathcal{L}), \epsilon)\right)}\mathrm{d}\epsilon < \infty.$$

4. *Assume that*
$$\mathbb{E}\left[\sup_{\substack{\theta\in A\\ \|\theta-\theta_0\|\le\delta}} \|z(X_1,\theta) - z(X_1,\theta_0)\|^2\right] \xrightarrow[\delta\to 0]{} 0.$$

Then
$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = -\left(J\mathbb{E}[z(X_1,\theta_0)]\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n z(X_i,\theta_0) + o_{\mathbb{P}}(1) \tag{21}$$

and thus
$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \left(J\mathbb{E}[z(X_1,\theta_0)]\right)^{-1}\operatorname{cov}\left(z(X_1,\theta_0)\right)\left(J\mathbb{E}[z(X_1,\theta_0)]\right)^{-1}\right). \tag{22}$$

A main strength of Theorem 30 is that we don't need differentiability of the random function $\theta \mapsto z(X_1,\theta)$, only of its expectation.

Proof of Theorem 30. Write for concision $V = J\mathbb{E}[z(X_1,\theta_0)]$. Let us write a Taylor expansion of $\theta \mapsto \mathbb{E}[z(X_1,\theta)]$ around θ_0:
$$\int_{\mathbb{R}^k} z(x,\theta)\mathrm{d}\mathcal{L}(x) = \int_{\mathbb{R}^k} z(x,\theta_0)\mathrm{d}\mathcal{L}(x) + V(\theta-\theta_0) + o(\|\theta-\theta_0\|).$$

Since we assume $\widehat{\theta}_n - \theta_0 = o_{\mathbb{P}}(1)$, from Lemma 10,
$$\int_{\mathbb{R}^k} z(x,\widehat{\theta}_n)\mathrm{d}\mathcal{L}(x) = \int_{\mathbb{R}^k} z(x,\theta_0)\mathrm{d}\mathcal{L}(x) + V(\widehat{\theta}_n-\theta_0) + o_{\mathbb{P}}(\|\widehat{\theta}_n-\theta_0\|).$$

This can be written (**exercize**)
$$\int_{\mathbb{R}^k} z(x,\widehat{\theta}_n)\mathrm{d}\mathcal{L}(x) = \int_{\mathbb{R}^k} z(x,\theta_0)\mathrm{d}\mathcal{L}(x) + \left(V + o_{\mathbb{P}}(1)\right)(\widehat{\theta}_n-\theta_0),$$

where this last $o_{\mathbb{P}}(1)$ is a sequence of $p \times p$ random matrices Q_n such that $\|Q_n\| = o_{\mathbb{P}}(1)$ (for any norm $\|\cdot\|$ on the space of matrices).

Multiplying the above display by $\sqrt{n}$ and using $\int_{\mathbb{R}^k} z(x,\theta_0)\mathrm{d}\mathcal{L}(x) = 0$ and (20), we obtain

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\int_{\mathbb{R}^k} z(x,\widehat{\theta}_n)\mathrm{d}\mathcal{L}(x) - z(X_i,\widehat{\theta}_n)\right) = (V + o_{\mathbb{P}}(1))\sqrt{n}(\widehat{\theta}_n - \theta_0) + o_{\mathbb{P}}(1).$$

We rewrite this as

$$\begin{aligned}(V + o_{\mathbb{P}}(1))\sqrt{n}(\widehat{\theta}_n - \theta_0) =& o_{\mathbb{P}}(1) - \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(z(X_i,\theta_0) - \int_{\mathbb{R}^k} z(x,\theta_0)\mathrm{d}\mathcal{L}(x)\right) \\ &+ \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\left(z(X_i,\theta_0) - z(X_i,\widehat{\theta}_n)\right) - \int_{\mathbb{R}^k}\left(z(x,\theta_0) - z(x,\widehat{\theta}_n)\right)\mathrm{d}\mathcal{L}(x)\right).\end{aligned}$$

If we prove that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\left(z(X_i,\theta_0) - z(X_i,\widehat{\theta}_n)\right) - \int_{\mathbb{R}^k}\left(z(x,\theta_0) - z(x,\widehat{\theta}_n)\right)\mathrm{d}\mathcal{L}(x)\right) = o_{\mathbb{P}}(1), \tag{23}$$

we can conclude the proof of both (21) and (22) because $V = J\mathbb{E}[m(X_1,\theta_0)]$ is fixed and invertible and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \left(z(X_i,\theta_0) - \int_{\mathbb{R}^k} z(x,\theta_0)\mathrm{d}\mathcal{L}(x)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n z(X_i,\theta_0) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \operatorname{cov}\left(z(X_1,\theta_0)\right)\right).$$

Call r_n the quantity in (23), note that it is a $p \times 1$ vector and write it $(r_{1,n},\ldots,r_{p,n})^\top$. For $j = 1,\ldots,p$, for $\delta > 0$ such that $B(\theta_0,\delta) \subset A$, define

$$\mathcal{F}_{j,\delta} = \left\{\mathbb{R}^k \ni x \mapsto z(x,\theta)_j - z(x,\theta_0)_j; \theta \in B(\theta_0,\delta)\right\}.$$

Note that if $\|\widehat{\theta}_n - \theta_0\| \le \delta$, we have

$$\|r_n\| \le \sqrt{p}\max_{j=1,\ldots,p}\sup_{f\in\mathcal{F}_{j,\delta}}\frac{1}{\sqrt{n}}\left|\sum_{i=1}^n \left(f(X_i) - \mathbb{E}[f(X_1)]\right)\right|.$$

Note that if $[\ell_1,u_1],\ldots,[\ell_N,u_N]$ is a finite set of brackets that covers $\mathcal{F}_j$ (as in (17) with $q = 2$), then $[\ell_1 - z(\cdot,\theta_0)_j, u_1 - z(\cdot,\theta_0)_j],\ldots,[\ell_N - z(\cdot,\theta_0)_j, u_N - z(\cdot,\theta_0)_j]$ is a finite set of brackets that covers $\mathcal{F}_{j,\delta}$ (as in (17) with $q = 2$). Indeed, for all $k \in \{1,\ldots,N\}$, $u_k - z(\cdot,\theta_0)_j - (\ell_k - z(\cdot,\theta_0)_j) = u_k - \ell_k$ and

$$\int_{\mathbb{R}^k}\left(u_k(x) - z(x,\theta_0)_j - (\ell_k(x) - z(x,\theta_0)_j)\right)^2\mathrm{d}\mathcal{L}(x) = \int_{\mathbb{R}^k}\left(u_k(x) - \ell_k(x)\right)^2\mathrm{d}\mathcal{L}(x).$$

Also, if $f \in [\ell_k,u_k]$ then $f - z(\cdot,\theta_0)_j \in [\ell_k - z(\cdot,\theta_0)_j, u_k - z(\cdot,\theta_0)_j]$.

Hence for all $\epsilon > 0$,

$$\mathcal{N}_{[]}(\mathcal{F}_{j,\delta}, L^2(\mathcal{L}), \epsilon) \le \mathcal{N}_{[]}(\mathcal{F}_j, L^2(\mathcal{L}), \epsilon). \tag{24}$$

Next, for all $\delta,\epsilon > 0$, with $B(\theta_0,\delta) \subset A$, we have

$$\mathbb{P}\left(\|r_n\| \ge \epsilon\right) \le \mathbb{P}\left(\|\widehat{\theta}_n - \theta_0\| \ge \delta\right) + \mathbb{P}\left(\sqrt{p}\max_{j=1,\ldots,p}\sup_{f\in\mathcal{F}_{j,\delta}}\frac{1}{\sqrt{n}}\left|\sum_{i=1}^n \left(f(X_i) - \mathbb{E}[f(X_1)]\right)\right| \ge \epsilon\right).$$

Since $\widehat{\theta}_n$ is assumed to converge to θ_0 in probability, applying $\limsup_{n\to\infty}$ yields

$$\limsup_{n\to\infty}\mathbb{P}\left(\|r_n\| \ge \epsilon\right) \le \mathbb{P}\left(\sqrt{p}\max_{j=1,\ldots,p}\sup_{f\in\mathcal{F}_{j,\delta}}\frac{1}{\sqrt{n}}\left|\sum_{i=1}^n \left(f(X_i) - \mathbb{E}[f(X_1)]\right)\right| \ge \epsilon\right).$$

For all $f \in \mathcal{F}_{j,\delta}$ and $x \in \mathbb{R}^k$, we have

$$|f(x)| \leq F_\delta(x),$$

with

$$F_\delta(x) = \sup_{\substack{\theta \in A \\ \|\theta-\theta_0\| \leq \delta}} \|z(X_1, \theta) - z(X_1, \theta_0)\|.$$

Hence from Theorem 29 (maximum inequality) and Markov inequality, we obtain

$$\begin{aligned}
\limsup_{n\to\infty} \mathbb{P}\left(\|r_n\| \geq \epsilon\right) &\leq \sum_{j=1}^{p} \mathbb{P}\left(\sup_{f \in \mathcal{F}_{j,\delta}} \frac{1}{\sqrt{n}} \left| \sum_{i=1}^{n} \left(f(X_i) - \mathbb{E}[f(X_1)]\right) \right| \geq \frac{\epsilon}{\sqrt{p}} \right) \\
&\leq \sum_{j=1}^{p} \frac{\sqrt{p}}{\epsilon} \mathbb{E}\left[\sup_{f \in \mathcal{F}_{j,\delta}} \frac{1}{\sqrt{n}} \left| \sum_{i=1}^{n} \left(f(X_i) - \mathbb{E}[f(X_1)]\right) \right| \right] \\
&\leq \sum_{j=1}^{p} \frac{\sqrt{p}}{\epsilon} C_{\mathrm{MI}} \int_0^{\sqrt{\mathbb{E}[F_\delta(X_1)^2]}} \sqrt{\log\left(\mathcal{N}_{[]}(\mathcal{F}_j, L^2(\mathcal{L}), u)\right)}\, \mathrm{d}u.
\end{aligned}$$

By assumption $\mathbb{E}[F_\delta(X_1)^2] \to 0$ as $\delta \to 0$ and the above function is integrable on any set $[0,t]$, $t < \infty$, and thus the lim sup above can be arbitrarily small by taking $\delta > 0$ small enough. Hence this lim sup is zero and thus (23) holds, which concludes the proof. □

6.3 Application to the empirical median

Let us apply Theorem 30 to the empirical median discussed at the end of Section 4.5. Consider thus i.i.d. random variables $(X_i)_{i\in\mathbb{N}}$, having a c.d.f. F_{X_1} and a density f with respect to Lebesgue measure, and their empirical median $\widehat{\theta}_n$ satisfying

$$\sum_{i=1}^{n} \mathrm{sign}(\widehat{\theta}_n - X_i) = 0.$$

This is as in (20) with $z(x,\theta) = \mathrm{sign}(\theta - x)$. Assume that f is strictly positive on $\mathbb{R}$, and thus F_{X_1} is strictly increasing on $\mathbb{R}$. Hence there is a unique θ_0 (the population median) such that $F_{X_1}(\theta_0) = 1/2$ and $f(\theta_0) > 0$ Hence from the discussion after Proposition 23, Item 1 of Theorem 30 holds.

Also, assume that f is continuous in a neighborhood of θ_0. Then $\mathbb{E}[\mathrm{sign}(\theta - X_1)] = 2F_{X_1}(\theta) - 1$ is continuously differentiable in a neighborhood of θ_0 with positive derivative $2f(\theta_0)$ at θ_0. Hence Item 2 of Theorem 30 holds.

The next lemma shows that Item 3 of Theorem 30 holds.

Lemma 31. *Let*

$$\mathcal{F} = \{\mathbb{R} \ni x \mapsto \mathrm{sign}(\theta - x); \theta \in \mathbb{R}\}$$

and $\mathcal{L}$ be a distribution on $\mathbb{R}$. Then for $\epsilon > 0$,

$$\mathcal{N}_{[]}(\mathcal{F}, L^2(\mathcal{L}), \epsilon) \leq \frac{4}{\epsilon^2} + 1.$$

Proof. Let us start by considering the set

$$\mathcal{F}_+ = \{\mathbb{R} \ni x \mapsto \mathbb{1}\{x < \theta\}; \theta \in \mathbb{R}\}.$$

Let $-\infty < t_1 < \cdots < t_N < +\infty$. Let $t_0 = -\infty$ and $t_{N+1} = \infty$. For $j = 1, \ldots, N$, let $\ell_{+,j}(x) = \mathbb{1}\{x \leq t_j\}$ and $u_{+,j}(x) = \mathbb{1}\{x < t_{j+1}\}$. Let $\ell_{+,0}(x) = 0$ and $u_{+,0}(x) = \mathbb{1}\{x < t_1\}$. Then, for all $\theta \in \mathbb{R}$, there is $j \in \{0, \ldots, N\}$ such that $t_j < \theta \leq t_{j+1}$ and thus for all $x \in \mathbb{R}$

$$\ell_{+,j}(x) \leq \mathbb{1}\{x < \theta\} \leq u_{+,j}(x)$$

and thus $f \in [\ell_{+,j}, u_{+,j}]$.

For any integer N such that $N+1 \geq \frac{1}{\epsilon^2}$, we can select $t_1, \ldots, t_N$ such that for $j = 0, \ldots, N$, $\mathcal{L}((t_j, t_{j+1})) \leq \frac{1}{\epsilon^2}$ **(exercize)**. With this choice, for $j = 0, \ldots, N$,

$$\int_{\mathbb{R}} (u_{+,j} - \ell_{+,j})^2 \mathrm{d}\mathcal{L} = \int_{\mathbb{R}} \mathbb{1}\{x \in (t_j, t_{j+1})\} \mathrm{d}\mathcal{L}(x) = \mathcal{L}((t_j, t_{j+1})) \leq \epsilon^2.$$

Next considering the set

$$\mathcal{F}_- = \{\mathbb{R} \ni x \mapsto \mathbb{1}\{\theta < x\}; \theta \in \mathbb{R}\}.$$

Keeping the same $t_1, \ldots, t_N$, for $j = 1, \ldots, N$, let $\ell_{-,j}(x) = \mathbb{1}\{t_{j+1} \leq x\}$ and $u_{+,j}(x) = \mathbb{1}\{t_j < x\}$. Let $\ell_{-,0}(x) = \mathbb{1}\{t_1 \leq x\}$ and $u_{-,0}(x) = 1$. Then, for all $\theta \in \mathbb{R}$, there is $j \in \{0, \ldots, N\}$ such that $t_j < \theta \leq t_{j+1}$ and thus for all $x \in \mathbb{R}$

$$\ell_{-,j}(x) \leq \mathbb{1}\{\theta < x\} \leq u_{-,j}(x)$$

and thus $f \in [\ell_{-,j}, u_{-,j}]$.

As before, for $j = 0, \ldots, N$,

$$\int_{\mathbb{R}} (u_{-,j} - \ell_{-,j})^2 \mathrm{d}\mathcal{L} \leq \epsilon^2.$$

Then for any $\theta \in \mathbb{R}$, taking $j \in \{0, \ldots, N\}$ such that $t_j < \theta \leq t_{j+1}$, for all $x \in \mathbb{R}$

$$\mathrm{sign}(\theta - x) = \mathbb{1}\{x < \theta\} - \mathbb{1}\{\theta < x\} \leq u_{+,j}(x) - \ell_{-,j}(x)$$

and also

$$\mathrm{sign}(\theta - x) \geq \ell_{+,j}(x) - u_{-,j}(x).$$

Also, from the triangle inequality

$$\sqrt{\int_{\mathbb{R}} \{u_{+,j}(x) - \ell_{-,j}(x) - (\ell_{+,j}(x) - u_{-,j}(x))\}^2 \mathrm{d}\mathcal{L}} \leq 2\epsilon.$$

Hence we have found the $N+1$ brackets

$$[u_{+,0}(x) - \ell_{-,0}(x), \ell_{+,0}(x) - u_{-,0}(x)], \ldots, [u_{+,N}(x) - \ell_{-,N}(x), \ell_{+,N}(x) - u_{-,N}(x)]$$

that cover $\mathcal{F}$ as in (17) with ϵ there replaced by 2ϵ here. Hence

$$\mathcal{N}_{[]}(\mathcal{F}, L^2(\mathcal{L}), 2\epsilon) \leq N+1.$$

Since we can choose $N+1 \leq \frac{1}{\epsilon^2} + 1$, we obtain

$$\mathcal{N}_{[]}(\mathcal{F}, L^2(\mathcal{L}), 2\epsilon) \leq \frac{1}{\epsilon^2} + 1$$

and thus for all $\epsilon > 0$

$$\mathcal{N}_{[]}(\mathcal{F}, L^2(\mathcal{L}), \epsilon) \leq \frac{4}{\epsilon^2} + 1.$$

□

Finally, for Item 4 of Theorem 30,

$$\mathbb{E}\left[\sup_{\substack{\theta \in \mathbb{R} \\ \|\theta - \theta_0\| \leq \delta}} \left(\mathrm{sign}(\theta - X_1) - \mathrm{sign}(\theta_0 - X_1)\right)^2\right] = 2\mathbb{P}\left(X_1 \in [\theta_0 - \delta, \theta_0 + \delta]\right) \underset{\delta \to 0}{\longrightarrow} 0.$$

Hence Theorem 30 applies to the empirical median and we also have

$$\mathrm{var}(\mathrm{sign}(\theta_0 - X_1)) = \mathbb{E}[\mathrm{sign}(\theta_0 - X_1)^2] = \mathbb{E}[1] = 1$$

and thus

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow[n \to \infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{4f^2(\theta_0)}.\right)$$

6.4 Application to maximum likelihood

We first provide a lemma enabling to bound the bracketing number of general parametric sets of functions.

Lemma 32. *Let $\mathcal{L}$ be a distribution on $\mathbb{R}^k$. Let Θ be a bounded set of $\mathbb{R}^p$ and let $\mathcal{F} = \{f_\theta; \theta \in \Theta\}$ where for each θ, $f_\theta : \mathbb{R}^k \to \mathbb{R}$ and $\int_{\mathbb{R}^k} f_\theta^2 \mathrm{d}\mathcal{L} < \infty$. Assume that there is $h : \mathbb{R}^k \to [0, \infty)$ with $1 \leq \int_{\mathbb{R}^k} h^2 \mathrm{d}\mathcal{L} < \infty$ and for $\theta_1, \theta_2 \in \Theta$ and $x \in \mathbb{R}^k$,*

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq \|\theta_1 - \theta_2\| h(x). \tag{25}$$

Then for each $\epsilon > 0$

$$\mathcal{N}_{[]}(\mathcal{F}, L^2(\mathcal{L}), \epsilon) \leq C_p \mathrm{diam}(\Theta)^p \left(\int_{\mathbb{R}^k} h^2 \mathrm{d}\mathcal{L} \right)^{\frac{p}{2}} \frac{1}{\epsilon^p}$$

for a constant C_p depending only on p.

Proof. One can show (**exercize**) that there is a constant C_p' such that for each $\delta > 0$ there is an integer $N \leq C_p' \mathrm{diam}(\Theta)^p \frac{1}{\delta^p}$ and there are $\theta_1, \dots, \theta_N \in \Theta$ with

$$\sup_{\theta \in \Theta} \min_{j=1,\dots,N} \|\theta - \theta_j\| \leq \delta.$$

For $j = 1, \dots, N$ and $x \in \mathbb{R}^k$ we write $\ell_j(x) = f_{\theta_j}(x) - 2\delta h(x)$ and $u_j(x) = f_{\theta_j}(x) + 2\delta h(x)$. Then we have $\ell_j(x) \leq u_j(x)$ and

$$\int_{\mathbb{R}^k} (u_j(x) - \ell_j(x))^2 \mathrm{d}\mathcal{L}(x) = 16\delta^2 \int_{\mathbb{R}^k} h^2(x) \mathrm{d}\mathcal{L}(x).$$

Also, for each $\theta \in \Theta$, there is j such that $\|\theta - \theta_j\| \leq 2\delta$ and thus from (25)

$$f_\theta(x) \geq f_{\theta_j}(x) - \|\theta - \theta_j\| h(x) \geq f_{\theta_j}(x) - 2\delta h(x) = \ell_j(x).$$

Similarly

$$f_\theta(x) \leq u_j(x).$$

Hence from (17),we have

$$\mathcal{N}_{[]}\left(\mathcal{F}, L^2(\mathcal{L}), 4\delta \sqrt{\int_{\mathbb{R}^k} h^2 \mathrm{d}\mathcal{L}} \right) \leq C_p' \mathrm{diam}(\Theta)^p \frac{1}{\delta^p}.$$

Hence taking $\epsilon = 4\delta \sqrt{\int_{\mathbb{R}^k} h^2 \mathrm{d}\mathcal{L}}$, we obtain that for each $\epsilon > 0$,

$$\mathcal{N}_{[]}\left(\mathcal{F}, L^2(\mathcal{L}), \epsilon \right) \leq 4^p C_p' \mathrm{diam}(\Theta)^p \left(\int_{\mathbb{R}^k} h^2 \mathrm{d}\mathcal{L} \right)^{p/2} \frac{1}{\epsilon^p}.$$

This concludes the proof. □

We now consider the setting of maximum likelihood as in Theorem 28 in Section 5.2. Hence we consider a set $\{\mathcal{L}_\theta; \theta \in \Theta\}$ of distributions on $\mathbb{R}^k$, with $\mathcal{L}_\theta$ having density f_θ with respect to Lebesgue measure, and where there are $(X_i)_{i \in \mathbb{N}}$ i.i.d. with density f_{θ_0} for $\theta_0 \in \mathring{\Theta}$.

For any function $h_\theta(x) \in \mathbb{R}$, we write $\nabla h_{\widetilde{\theta}}(x)$ for its vector of partial derivatives w.r.t. θ at $\theta = \widetilde{\theta}$. We also assume that for each θ and x, $f_\theta(x) > 0$ and $f_\theta(x)$ is twice continuously differentiable w.r.t. θ with gradient $\nabla f_\theta(x)$. We assume that for all θ

$$\mathbb{E}[\|\nabla(\log f_\theta(X_1))\|^2] < \infty.$$

We consider a maximum likelihood estimator $\widehat{\theta}_n$ assumed to be consistent (for instance thanks to Theorem 28) and satisfying

$$\frac{1}{n}\sum_{i=1}^{n} \nabla \log f_\theta(X_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{f_\theta(X_i)} \nabla f_\theta(X_i) = 0.$$

Hence, let

$$z(x,\theta) = \frac{1}{f_\theta(x)} \nabla f_\theta(x).$$

We have

$$\mathbb{E}[z(X_1,\theta_0)] = \mathbb{E}\left[\frac{\nabla f_{\theta_0}(X_1)}{f_{\theta_0}(X_1)}\right] = \int_{\mathbb{R}^k} \nabla f_{\theta_0}(x) \frac{f_{\theta_0}(x)}{f_{\theta_0}(x)} \mathrm{d}x = \int_{\mathbb{R}^k} \nabla f_{\theta_0}(x) \mathrm{d}x.$$

Hence assuming

$$\int_{\mathbb{R}^k} \sup_{\theta\in\Theta} \|\nabla f_\theta(x)\| \mathrm{d}x < \infty, \tag{26}$$

from the dominated convergence theorem

$$\mathbb{E}[z(X_1,\theta_0)] = \nabla\left(\int_{\mathbb{R}^k} f_{\theta_0}(x)\mathrm{d}x\right) = \nabla 1 = 0.$$

Hence Item 1 of Theorem 30 holds.

Next, for any function $h_\theta(x) \in \mathbb{R}^p$, we write $Jh_{\widetilde{\theta}}(x)$ for its $p \times p$ Jacobian matrix with element a, b equal to $\left.\frac{\partial h_\theta(x)_a}{\partial \theta_b}\right|_{\theta=\widetilde{\theta}}$.

We now also assume that for $a, b \in \{1, \ldots, p\}$,

$$\int_{\mathbb{R}^k} \sup_{\theta\in\Theta} \left|\frac{\partial^2(\log f_\theta(x))}{\partial\theta_a \partial\theta_b}\right|^2 f_{\theta_0}(x)\mathrm{d}x < \infty. \tag{27}$$

This implies from dominated convergence that

$$J\mathbb{E}[z(X_1,\theta)] = J\int_{\mathbb{R}^k} \nabla(\log f_\theta(x)) f_{\theta_0}(x) = \int_{\mathbb{R}^k} (J\nabla)(\log f_\theta(x)) f_{\theta_0}(x)$$

is well-defined for all θ. Above, we also notice that $(J\nabla)(\log f_\theta(x))$ is the $p \times p$ Hessian matrix of $\theta \mapsto \log f_\theta(x)$ at θ.

For any $a, b = 1, \ldots, p$, we have

$$\begin{aligned}(J\mathbb{E}[z(X_1,\theta_0)])_{a,b} &= \int_{\mathbb{R}^k} \left(\frac{\frac{\partial^2 f_{\theta_0}(x)}{\partial\theta_a\partial\theta_b} f_{\theta_0}(x) - \frac{\partial f_{\theta_0}(x)}{\partial\theta_a}\frac{\partial f_{\theta_0}(x)}{\partial\theta_b}}{f_{\theta_0}(x)^2}\right) f_{\theta_0}(x)\mathrm{d}x \\ &= \int_{\mathbb{R}^k} \frac{\partial^2 f_{\theta_0}(x)}{\partial\theta_a\partial\theta_b}\mathrm{d}x - \int_{\mathbb{R}^k} \frac{\partial \log f_{\theta_0}(x)}{\partial\theta_a}\frac{\partial \log f_{\theta_0}(x)}{\partial\theta_b} f_{\theta_0}(x). \end{aligned} \tag{28}$$

If we assume that

$$\int_{\mathbb{R}^k} \sup_{\theta\in\Theta} \left|\frac{\partial^2 f_{\theta_0}(x)}{\partial\theta_a\partial\theta_b}\right| \mathrm{d}x < \infty$$

then the two separate integrals in (28) are well-defined and we have

$$\int_{\mathbb{R}^k} \frac{\partial^2 f_{\theta_0}(x)}{\partial\theta_a\partial\theta_b}\mathrm{d}x = \frac{\partial \int_{\mathbb{R}^k} \frac{\partial f_{\theta_0}(x)}{\partial\theta_a}\mathrm{d}x}{\partial\theta_b} = \frac{\partial 0}{\partial\theta_b} = 0.$$

Hence we have

$$J\mathbb{E}[z(X_1,\theta_0)] = -\mathrm{cov}\left(z(X_1,\theta_0)\right) \tag{29}$$

that we can assume to be invertible in order for Item 2 of Theorem 30 to hold.

For Item 3 of Theorem 30, we can use Lemma 32 since for $a = 1, \dots, p$

$$\left| \frac{\partial \log f_{\theta_1}(x)}{\partial \theta_a} - \frac{\partial \log f_{\theta_2}(x)}{\partial \theta_a} \right| \leq \|\theta_1 - \theta_2\| \sqrt{p} \max_{b=1,\dots,p} \sup_{\theta \in \Theta} \left| \frac{\partial^2 \log f_\theta(x)}{\partial \theta_a \partial \theta_b} \right|$$

and we can use (27). Hence Item 3 of Theorem 30 indeed holds.

Finally,

$$\sup_{\substack{\theta \in A \\ \|\theta - \theta_0\| \leq \delta}} \left| \frac{\partial \log f_{\theta_1}(x)}{\partial \theta_a} - \frac{\partial \log f_{\theta_2}(x)}{\partial \theta_a} \right|^2 \leq \delta p \max_{b=1,\dots,p} \sup_{\theta \in \Theta} \left| \frac{\partial^2 \log f_\theta(x)}{\partial \theta_a \partial \theta_b} \right|$$

and thus Item 4 of Theorem 30 holds.

From Theorem 30 and (29), we obtain

$$\sqrt{n} \left(\widehat{\theta}_n - \theta_0 \right) \xrightarrow[n \to \infty]{\mathcal{L}} \mathcal{N} \left(0, \operatorname{cov} \left(z(X_1, \theta_0) \right)^{-1} \right).$$

Note that the matrix $-J\mathbb{E}[z(X_1, \theta_0)] = \operatorname{cov}(z(X_1, \theta_0))$ is called the **Fisher information matrix**.

References

[VdV07] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2007.