Asymptotic Statistics

Lecture 6

M-estimators and Z-estimators

Definition - 1

Let $X_1, \ldots, X_n$ be i.i.d. random vectors in $\mathcal{X}$, with a distribution depending on a parameter $\theta \in \Theta \subset \mathbb{R}^k$.

Definition.
An *M*-estimator for θ is an estimator $\hat{\theta}_n$ that (nearly) **m**aximizes a (random) function of the form

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

where for every $\theta \in \Theta$, $m_\theta : \mathcal{X} \to \bar{\mathbb{R}}$ is a known function.

Definition - 2

Often: M-estimators are computed by setting partial derivatives to zero.

Definition.
A *Z*-estimator for θ is an estimator $\hat{\theta}_n$ that is (nearly) the **z**ero of a (random) function of the form

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

where for every $\theta \in \Theta$, $\psi_\theta : \mathcal{X} \to \mathbb{R}^k$ is a known function.

Examples

- Maximum likelihood estimators: Example 4.1
- Location estimators (i.e. sample mean, median, robust estimators): Example 4.2
- Linear or non-linear least squares estimators in regression: Example 4.3

Consistency - idea

Let $\hat{\theta}_n$ be an M-estimator that maximizes a criterion $\theta \mapsto M_n(\theta)$. When is $\hat{\theta}_n$ consistent?

Idea: The M_n are random functions on Θ. If they converge to some fixed, deterministic function M in an appropriate sense, we can hope that the maximizer $\hat{\theta}_n$ of M_n converges to the maximizer of M. If this is the true parameter θ_0, we obtain consistency.

Consistency - general theorem 1

Theorem.
Let M_n be random functions and M a fixed function on Θ.
Suppose

- The M_n converge uniformly too M, in probability:

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0,$$

- M has a well-separated maximum at θ_0: for all $\varepsilon > 0$

$$\sup_{\theta : \|\theta - \theta_0\| \geq \varepsilon} M(\theta) < M(\theta_0).$$

Then if $M_n(\hat{\theta}_n) \geq \sup_\theta M_n(\theta) - o_P(1)$ (i.e. $\hat{\theta}_n$ nearly maximizes M_n), it holds that $\hat{\theta}_n \xrightarrow{P} \theta_0$.

Consistency - general theorem 2

Theorem.
Let Ψ_n be random functions and Ψ a fixed function on Θ.
Suppose

- The Ψ_n converge uniformly too Ψ, in probability:

$$\sup_{\theta \in \Theta} |\Psi_n(\theta) - \Psi(\theta)| \xrightarrow{P} 0,$$

- Ψ has a well-separated zero at θ_0: for all $\varepsilon > 0$

$$\inf_{\theta : \|\theta - \theta_0\| \geq \varepsilon} \|\Psi(\theta)\| > 0 = \Psi(\theta_0).$$

Then if $\Psi_n(\hat{\theta}_n) = o_P(1)$ (i.e. $\hat{\theta}_n$ is a near zero of Ψ_n), it holds that $\hat{\theta}_n \xrightarrow{P} \theta_0$.

Consistency - connection with uniform LLNs

If $X_1, \ldots, X_n$ are i.i.d. in $\mathcal{X}$ and $M_n(\theta) = n^{-1} \sum_{i=1}^n m_\theta(X_i)$, then by the LLN,

$$M_n(\theta) \xrightarrow{P} M(\theta) = E m_\theta(X_1)$$

for every $\theta \in \Theta$.

The uniform convergence required in the theorem corresponds in this case to a uniform LLN of the form

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n m_\theta(X_i) - E m_\theta(X_1) \right| \xrightarrow{P} 0.$$

Glivenko-Cantelli classes of functions

Let $X_1, \ldots, X_n$ be i.i.d. in $\mathcal{X}$ under P. Notation: for a function $f : \mathcal{X} \to \mathbb{R}$,

$$\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i), \qquad Pf = Ef(X_1).$$

Note: by the LLN, $\mathbb{P}_n f \xrightarrow{P} Pf$ for a fixed $f \in L^1(P)$.

Definition.
A class of functions $\mathcal{F} \subset L^1(P)$ is called Glivenko-Cantelli (in probability) if

$$\sup_{f \in \mathcal{F}} \left| \mathbb{P}_n f - Pf \right| \xrightarrow{P} 0.^1$$

[1] Measurability might be an issue..

GC classes: bracketing numbers

Idea: if a class $\mathcal{F}$ is "not too large", it should be Glivenko-Cantelli.

Example: by the LLN, every finite class of functions in $L^1(P)$ is GC.

Definition.

- For two function l, u, the bracket $[l, u]$ is the collection of all functions f such that $l \leq f \leq u$, pointwise.
- For $\varepsilon > 0$ and $r > 0$, an ε-bracket in $L^r(P)$ is a bracket $[l, u]$ such that $P|u - l|^r \leq \varepsilon^r$.
- For a class of (measurable) functions $\mathcal{F}$, the bracketing number $N_{[]}(\varepsilon, \mathcal{F}, L^r(P))$ is the minimal number of ε-brackets in $L^r(P)$ needed to cover $\mathcal{F}$.

Bracketing numbers: examples

Idea: the bracketing number $N_{[]}(\varepsilon, \mathcal{F}, L^r(P))$ is a measure of the "size" of $\mathcal{F}$.

Examples:

- For a finite class $\mathcal{F}$, $N_{[]}(\varepsilon, \mathcal{F}, L^r(P)) \leq \#\mathcal{F}$ for all $\varepsilon > 0$.
- For the class of indicator functions $\mathcal{F} = \{1_{(-\infty, x]} : x \in \mathbb{R}\}$, $N_{[]}(\varepsilon, \mathcal{F}, L^1(P)) \leq 2/\varepsilon$ for all $\varepsilon > 0$.
- Consider a parametric class $\mathcal{F} = \{f_\theta : \theta \in \Theta\} \subset L^1(P)$, with Θ compact and $\theta \mapsto f_\theta(x)$ continuous for every x, and such that there exists an $F \in L^1(P)$ such that $|f_\theta| \leq F$ for all θ. Then $N_{[]}(\varepsilon, \mathcal{F}, L^1(P)) < \infty$ for every $\varepsilon > 0$. (Check, using some measure theory! See Example 19.8 of Van der Vaart (1998)).

Glivenko-Cantelli theorem under bracketing

Theorem.

If $N_{[]}(\varepsilon, \mathcal{F}, L^1(P)) < \infty$ for every $\varepsilon > 0$, then $\mathcal{F}$ is Glivenko-Cantelli.

Proof.

Let $[l, u]$ be an ε-bracket and $f \in [l, u]$. Then

$$\mathbb{P}_n f - Pf \leq \mathbb{P}_n u - Pu + \varepsilon,$$
$$\mathbb{P}_n f - Pf \geq \mathbb{P}_n l - Pl - \varepsilon.$$

Since there are only finitely many brackets containing all $f \in \mathcal{F}$, the LLN then implies that (check!) a.s.

$$\limsup_{n \to \infty} \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \leq \varepsilon.$$

Now let $\varepsilon \to 0$. □

→ Lemma 4.7

Alternative conditions for consistency

Uniform convergence of the criterion functions can be a (too) strong condition.

In specific cases, there are alternative conditions for consistency that can be less demanding.

Example: Z-estimators for a real-valued parameter with a monotone criterion function. See Lemma 4.9.