M2RI Asymptotic Statistics Lecture 4: M and Z estimators and Uniform Convergence

M-Estimators

Setup $\Theta \subset \mathbb{R}^p$, $\{\mathcal{L}_\theta ; \theta \in \Theta\}$ is a statistical model.

$(M_n)_{n \in \mathbb{N}}$ a sequence of random functions from Θ to $\mathbb{R}^d$: $\forall n, \forall \theta$, $M_n(\theta)$ is a random vector in $\mathbb{R}^d$ and $\{M_n(\theta) ; \theta \in \Theta\}$ are defined on the same space.

Definition A M-Estimator is a sequence of random $(\hat{\theta}_n)$ taking values in Θ such that

$$\forall n, \text{ almost surely, } \hat{\theta}_n \in \underset{\theta \in \Theta}{\operatorname{argmax}}\ M_n(\theta)$$

M = "maximizer".

Exercise: Given random vectors $X_1, \ldots, X_n \in \mathbb{R}^d$, express the empirical mean $\bar{X}_n = \frac{1}{n} \sum_i X_i$ as a M-estimator.

Solution

$$\overline{X_n} \in \underset{x}{\operatorname{argmin}} \underbrace{\sum_{i=1}^{n} \| x - X_i \|^2}_{M_n(x)}$$

Indeed, $\forall x$, $\nabla M_n(x) = 2 \sum_{i=1}^{n} (x - X_i)$.

Furthermore, M_n is C^1 and coercive. So it admits a global minimizer $\hat{\theta}_n$ that satisfies $\nabla M_n(\hat{\theta}_n) = 0$.

Then, $\nabla M_n(\hat{\theta}_n) = 0$ iff $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}_n$.

Example (Maximum Likelihood).

Assumption: $\forall \theta$, $\mathcal{L}_\theta$ has a density f_θ w.r.t. a reference measure μ.

$$L_n(\theta) \equiv \prod_{i=1}^{n} f_\theta(X_i) \quad \text{(Likelihood)}.$$

$$\ell_n(\theta) \equiv \log(L_n(\theta)) \equiv \sum_{i=1}^{n} \log(f_\theta(X_i)) \quad \text{(Log-Likelihood)}.$$

Log-Likelihood estimator:

$$\hat{\theta}_n \in \operatorname*{argmax}_{\theta \in \Theta} M_n(\theta) \text{ with } M_n(\theta) = \ell_n(\theta).$$

Idea: Choose the model that maximizes the "probability" of observing the samples.

Exercice: What is the maximum likelihood estimator in the Gaussian model?

Solution: $\mathcal{L}_{(\mu,\sigma^2)} = \mathcal{N}(\mu, \sigma^2)$ and $f_{(\mu,\sigma^2)} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$.

Let $X_1, \dots, X_n \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$.

$$\ell_n(\mu, \sigma^2) = \sum_{i=1}^{n} \left\{ -\ln(\sigma) - \ln(\sqrt{2\pi}) - \frac{1}{2}\left(\frac{X_i - \mu}{\sigma}\right)^2 \right\}.$$

$$\frac{\partial \ell_n}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu) \text{ and } \frac{\partial \ell_n}{\partial \mu} = 0 \text{ iff } \mu = \overline{X}_n$$

$$\frac{\partial \ell_n}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (X_i - \mu)$$

and $$\frac{\partial \ell_n}{\partial \sigma^2}(\mu = \bar{X}_n) = 0 \text{ iff } \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

Consistency of M-estimators

Theorem: Consider a sequence (M_n) of random functions from $\Theta \subset \mathbb{R}^d$ to $\mathbb{R}$. Consider a deterministic function $M : \Theta \to \mathbb{R}$.

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

and $\exists \theta_0 \in \Theta$ s.t. $\forall \varepsilon > 0$, $$\sup_{\substack{\theta \in \Theta \\ \|\theta - \theta_0\| \geq \varepsilon}} M(\theta) < M(\theta_0),$$

and $(\hat{\theta}_n)$ is a sequence s.t. $$M_n(\hat{\theta}_n) \geq \left(\sup_{\theta \in \Theta} M_n(\theta)\right) + o_{\mathbb{P}}(1)$$

Then $$\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0 .$$

__Proof.__ Let $\varepsilon > 0$. We have $$\mathbb{P}(\|\hat{\theta}_n - \theta_0\| \geq \varepsilon) \leq \mathbb{P}\left(M(\hat{\theta}_n) \leq \sup_{\substack{\theta \in \Theta \\ \|\theta - \theta_0\| \geq \varepsilon}} M(\theta)\right).$$

Furthermore, $$M(\hat{\theta}_n) \geq M_n(\hat{\theta}_n) - \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)|$$

$$\geq M_n(\theta_0) - \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| + o_{\mathbb{P}}(1)$$

$$\geq M(\theta_0) - 2 \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| + o_{\mathbb{P}}(1)$$

$$= M(\theta_0) + o_{\mathbb{P}}(1).$$

So, $$\mathbb{P}(\|\hat{\theta}_n - \theta_0\| \geq \varepsilon) \leq \mathbb{P}\left(o_{\mathbb{P}}(1) \leq M(\theta_0) - \sup M(\theta)\right)$$

$$\underbrace{\qquad \underset{\substack{\theta \in \Theta \\ \|\theta-\theta_0\| \geq \varepsilon}}{} \qquad}_{>0}$$

$\leq \varepsilon$ for n big enough $\square$.

Z-Estimators

Setup $\Theta \subset \mathbb{R}^p$, $\{\mathcal{L}_\theta ; \theta \in \Theta\}$ is a statistical model.

$(Z_n)_{n \in \mathbb{N}}$ a sequence of random functions from Θ to $\mathbb{R}^d$: $\forall n, \forall \theta, M_n(\theta)$ is a random vector in $\mathbb{R}^d$ and $\{M_n(\theta) ; \theta \in \Theta\}$ are defined on the same space.

Definition A Z-Estimator is a sequence of random $(\hat{\theta}_n)$ taking values in Θ such that

$$\forall n, \text{ a.s.}, \quad Z_n(\hat{\theta}_n) = 0 .$$

Remark Often, M estimators are defined by $\nabla M_n(\hat{\theta}_n) = 0$, in which case they are also Z-estimators!

Remark The method of moments is a Z-estimator.

Theorem (Z_n) sequence of random functions from $\Theta \subset \mathbb{R}^p$ to $\mathbb{R}^q$ and $Z : \Theta \to \mathbb{R}^q$ a deterministic function.

If $\sup_{\theta \in \Theta} \|Z_n(\theta) - Z(\theta)\| \xrightarrow{\mathbb{P}} 0$

and $\inf_{\substack{\theta \in \Theta \\ \|\theta - \theta_0\| \geq \varepsilon}} \|Z(\theta)\| > 0 = \|Z(\theta_0)\|$

and $(\hat{\theta}_n)$ is a sequence such that $Z_n(\hat{\theta}_n) = o_{\mathbb{P}}(1)$

Then $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$.

Proof Let $\varepsilon > 0$. $\mathbb{P}(\|\hat{\theta}_n - \theta_0\| \geq \varepsilon) \leq \mathbb{P}\left(\|Z(\hat{\theta}_n)\| \geq \inf_{\substack{\theta \in \Theta \\ \|\theta - \theta_0\| \geq \varepsilon}} \|Z(\theta)\|\right)$

Furthermore, $\|Z(\theta_n)\| \leq \|Z_n(\hat{\theta}_n)\| + \sup_{\theta \in \Theta} \big|\|Z_n(\theta)\| - \|Z(\theta)\|\big|$

$$\leq o_{\mathbb{P}}(1) + \sup_{\theta \in \Theta} \|Z_n(\theta) - Z(\theta)\| = o_{\mathbb{P}}(1)$$

So $\mathbb{P}(\|\hat{\theta}_n - \theta_0\| \geq \varepsilon) \leq \mathbb{P}\left(o_{\mathbb{P}}(1) \geq \inf_{\substack{\theta \in \Theta \\ \|\theta - \theta_0\| \geq \varepsilon}} \|Z(\theta)\|\right)$

$$\leq \varepsilon \text{ for } n \text{ big enough.}$$ □

Theorem If $\Theta = \mathbb{R}$, (Z_n) is a sequence of random functions from Θ to $\mathbb{R}$ and $Z : \Theta \to \mathbb{R}$ is a deterministic function such that

- $\forall \theta$ fixed $Z_n(\theta) \xrightarrow{\mathbb{P}} Z(\theta)$ (not uniform)

- $\forall n$, Z_n is <u>non-decreasing</u> a.s.
- $\exists \theta_0$ s.t. $\forall \varepsilon > 0$, $Z(\theta_0 - \varepsilon) < 0 < Z(\theta_0 + \varepsilon)$.

Then, if $(\hat{\theta}_n)$ is such that $Z_n(\hat{\theta}_n) = o_{\mathbb{P}}(1)$,

$$\underline{\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta}.$$

<u>Proof</u> Let $\varepsilon > 0$

$$\begin{aligned}
\mathbb{P}(|\hat{\theta}_n - \theta_0| \geq \varepsilon) &= \mathbb{P}(\hat{\theta}_n \leq \theta_0 - \varepsilon) + \mathbb{P}(\hat{\theta}_n \geq \theta_0 + \varepsilon) \\
&\leq \mathbb{P}\big(Z_n(\hat{\theta}_n) \leq Z_n(\theta_0 - \varepsilon)\big) + \mathbb{P}\big(Z_n(\hat{\theta}_n) \geq Z_n(\theta_0 + \varepsilon)\big) \\
&= \mathbb{P}\big(o_{\mathbb{P}}(1) \leq Z_n(\theta_0 - \varepsilon)\big) + \mathbb{P}\big(o_{\mathbb{P}}(1) \geq Z_n(\theta_0 + \varepsilon)\big) \\
&= \mathbb{P}\big(o_{\mathbb{P}}(1) \leq Z(\theta_0 - \varepsilon) + \underbrace{Z_n(\theta_0 - \varepsilon) - Z(\theta_0 - \varepsilon)}_{o_{\mathbb{P}}(1)}\big) \\
&\quad + \mathbb{P}\big(o_{\mathbb{P}}(1) \geq Z(\theta_0 + \varepsilon) + \underbrace{Z_n(\theta_0 + \varepsilon) - Z(\theta_0 + \varepsilon)}_{o_{\mathbb{P}}(1)}\big) \\
&= \underbrace{\mathbb{P}\big(o_{\mathbb{P}}(1) \leq Z(\theta_0 - \varepsilon)\big)}_{\substack{\leq \varepsilon \\ \text{for } n \text{ big enough}}} + \underbrace{\mathbb{P}\big(o_{\mathbb{P}}(1) \geq Z(\theta_0 + \varepsilon)\big)}_{\substack{\leq \varepsilon \\ \text{for } n \text{ big enough.}}}.
\end{aligned}$$

$\square$

<u>Exercice</u> $X_1, \ldots \overset{iid}{\sim} X$. By considering the empirical median $\hat{\theta}_n$ defined as

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

show its consistency. (X has a continuous density w.r.t. Lebesgue bounded > 0

[illegible] by below }.

Solution: $Z_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \operatorname{sgn}(\theta - X_i)$

$$Z(\theta) = F_X(\theta) - (1 - F_X(\theta))$$

$\forall \theta$, $$Z_n(\theta) = \frac{1}{n} \sum_i \operatorname{sgn}(\theta - X_i)$$

$$= \frac{1}{n} \sum_i \mathbb{1}(\theta - X_i > 0) - \frac{1}{n} \sum_i \mathbb{1}(\theta - X_i < 0)$$

$$\xrightarrow[\mathbb{P}]{} \mathbb{P}(X < \theta) - \mathbb{P}(X > \theta) \quad (\text{LLN})$$

$$= Z(\theta)$$

Furthermore, Z_n is non-decreasing and Z is strictly increasing (because of the hypothesis on the density).

Hence we get the result.