Universiteit Leiden
Mathematical Institute

M–Estimators

Chapter 3.2

Thomas Nagler <t.w.nagler@math.leidenuniv.nl>
Leiden University

Monday 29th June, 2020

What are M-estimators

Maximum-likelihood (Penalized) Least squares Robust regression Quantile regression Kernel smoothing Support vector machines Neural networks M-estimator

$$\hat{\theta}_n = \arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i).$$

More generally,

$$\hat{\theta}_n = \arg\max_{\theta} \mathbb{M}_n(\theta).$$

Some notation

- Let θ_0 be the 'true' parameter, i.e,

$$\theta_0 = \arg\max_{\theta} \mathbb{M}(\theta),$$

for some limit process $\mathbb{M}$.

- Sometimes it is useful to set $h = r_n(\theta - \theta_0)$ as parameter and reformulate $\mathbb{M}$ accordingly: If

$$\hat{\theta}_n = \arg\max_{\theta} \mathbb{M}_n(\theta),$$

we write

$$\hat{h}_n = \arg\max_{h} \mathbb{M}'_n(h) = \arg\max_{h} \mathbb{M}_n(\theta_0 + h/r_n) - \mathbb{M}_n(\theta_0).$$

Outline

General strategy for dealing with M-estimators:

1. Consistency
2. Rate of convergence
3. Asymptotic distribution

The Argmax theorem

Lemma (3.2.1)

Let $\mathbb{M}_n$, $\mathbb{M}$ *be processes index by a metric space* H*. Suppose:*

- $\mathbb{M}_n \rightsquigarrow \mathbb{M}$ *in* $\ell^\infty(A \cup B)$ *with* $A, B \subset H$ *arbitrary.*
- *There is* $\hat{h}$ *such that for every open set* G *containing* $\hat{h}$*,*

$$\mathbb{M}(\hat{h}) > \sup_{h \notin G, h \in A} \mathbb{M}(h), \quad a.s.$$

- $\hat{h}_n$ *satisfies* $\mathbb{M}_n(\hat{h}_n) \geq \sup_h \mathbb{M}_n(h) - o_P(1)$.

Then, for every closed set F*,*

$$\limsup_{n\to\infty} P^*(\hat{h}_n \in F \cap A) \leq P(\hat{h} \in F \cup B^c).$$

Note: $A = B = H$ would imply $\hat{h}_n \rightsquigarrow \hat{h}$ in H.

The Argmax theorem

Proof. By continuous the mapping theorem,

$$\sup_{h \in F \cap A} \mathbb{M}_n(h) - \sup_{h \in B} \mathbb{M}_n(h) \rightsquigarrow \sup_{h \in F \cap A} \mathbb{M}(h) - \sup_{h \in B} \mathbb{M}(h).$$

Then

$$\begin{aligned}
\limsup_{n\to\infty} P^*(\hat{h}_n \in F \cap A) &\le \limsup_{n\to\infty} P^*\left(\sup_{h \in F \cap A} \mathbb{M}_n(h) \ge \sup_{h \in B} \mathbb{M}_n(h) - o_P(1) \right) \\
\text{[Slutsky, Portmanteau]} \quad &\le P\left(\sup_{h \in F \cap A} \mathbb{M}(h) \ge \sup_{h \in B} \mathbb{M}(h) \right) \\
&\le P\left(\hat{h} \in F \text{ or } \hat{h} \notin B \right). \quad \square
\end{aligned}$$

The Argmax theorem

Theorem (3.2.2, Argmax continuous mapping)

Suppose:

- $\mathbb{M}_n \rightsquigarrow \mathbb{M}$ *in* $\ell^\infty(K)$ *for every compact* $K \subset H$.
- *Almost all paths* $h \mapsto \mathbb{M}(h)$ *are upper semicontinuous and have a unique maximum at* $\hat{h}$, *which is tight as random map in* H.
- $\hat{h}_n$ *is uniformly tight and* $\mathbb{M}_n(\hat{h}_n) \geq \sup_h \mathbb{M}_n(h) - o_P(1)$.

Then $\hat{h}_n \rightsquigarrow \hat{h}$ *in* H.

The Argmax theorem

Proof.

- We apply the Lemma with $A = B = K$.
- Because $\hat{h}$ is unique and $\mathbb{M}$ upper semicontinuous,

$$\mathbb{M}(\hat{h}) > \sup_{h \notin G, h \in K} \mathbb{M}(h), \quad a.s.$$

- Then

$$\begin{aligned} \limsup_{n\to\infty} P^*(\hat{h}_n \in F) &= \limsup_{n\to\infty}\Big(P^*(\hat{h}_n \in F \cap K) + P^*(\hat{h}_n \in F \cap K^c)\Big) \\ \text{[Lemma 3.2.1]} \quad &\leq P(\hat{h} \in F) + P(\hat{h} \notin K) + \limsup_{n\to\infty} P^*(\hat{h}_n \notin K). \end{aligned}$$

- Make 2nd and 3rd term arbitrarily small by taking K large enough.
- Then $\hat{h}_n \rightsquigarrow h$ by the Portmanteau theorem. $\square$

Consistency

Corollary (3.2.3 i, Consistency)

Let $\mathbb{M}_n$ *be indexed by* Θ *and* $\mathbb{M}\colon \Theta \mapsto \mathbb{R}$ *deterministic. Suppose:*

- $\|\mathbb{M}_n - \mathbb{M}\|_\Theta \overset{P^*}{\to} 0.$
- *There is* θ_0 *with* $\mathbb{M}(\theta_0) > \sup_{\theta \notin G}(\theta)$ *for open* G.
- $\mathbb{M}_n(\hat{\theta}_n) \geq \sup_\theta \mathbb{M}_n(\theta) - O_P(1).$

Then $\hat{\theta}_n \overset{P^*}{\to} \theta_0$.

Consistency

Corollary (3.2.3 ii, Consistency)

Let $\mathbb{M}_n$ be indexed by Θ and $\mathbb{M}\colon \Theta \mapsto \mathbb{R}$ deterministic. Suppose:

- $\|\mathbb{M}_n - \mathbb{M}\|_K \overset{P^*}{\to} 0$ *for every compact $K \subset \Theta$.*
- $\theta \mapsto \mathbb{M}(\theta)$ *is upper semicontinuous with unique maximum at θ_0.*
- $\mathbb{M}_n(\hat{\theta}_n) \geq \sup_\theta \mathbb{M}_n(\theta) - O_P(1)$ *and $\hat{\theta}_n$ is uniformly tight.*

Then $\hat{\theta}_n \overset{P^*}{\to} \theta_0$.

Example: MLE

- Let

$$\hat{\theta}_n = \arg\max_{\theta} \mathbb{M}_n(\theta) = \arg\max_{\theta} \mathbb{P}_n m_\theta = \arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ln p_\theta(X_i).$$

- The last corollary gives easy conditions for $\hat{\theta}_n \to_p \theta_0$.
- The Argmax theorem can also be used to derive the asymptotic distribution of $\hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta_0)$.

Example: MLE

- Write $\hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta_0)$ as

$$\begin{aligned}\hat{h}_n = \arg\max_h \mathbb{M}'_n(h) &= \arg\max_h \mathbb{M}_n(\theta_0 + h/\sqrt{n}) \\ &= \arg\max_h \frac{1}{n}\sum_{i=1}^n \ln p_{\theta_0 + h/\sqrt{n}}(X_i).\end{aligned}$$

- If p_θ is sufficiently regular,

$$\frac{1}{n}\sum_{i=1}^n \ln p_{\theta_0+h/\sqrt{n}}(X_i) = h' \frac{1}{\sqrt{n}}\sum_{i=1}^n \nabla_\theta \ln p_{\theta_0}(X_i) - \frac{1}{2} h' I_{\theta_0} h + o_P(1).$$

- Under appropriate conditions (Theorems 2.11.22 or 2.11.23), this converges weakly to

$$h \mapsto h'\Delta_{\theta_0} - \frac{1}{2} h' I_{\theta_0} h, \qquad \Delta_{\theta_0} \sim \mathcal{N}(0, \Sigma_{\theta_0})$$

- Argmax theorem: $\hat{h}_n \rightsquigarrow \hat{h} = I_{\theta_0}^{-1}\Delta_{\theta_0}$.

Outline

1. Consistency
2. Rate of convergence
3. Asymptotic distribution

Rate of convergence

Theorem (3.2.5)

Let $\mathbb{M}_n$ be a stochastic process indexed by Θ and $\mathbb{M}\colon \Theta \mapsto \mathbb{R}$ deterministic. Suppose:

- *for some 'distance' d and every θ in a neighborhood of θ_0,*

$$\mathbb{M}(\theta) - \mathbb{M}(\theta_0) \lesssim -d^2(\theta, \theta_0).$$

- *For every n and δ small,*

$$\mathrm{E}^* \sup_{d(\theta,\theta_0)<\delta} \left|(\mathbb{M}_n - \mathbb{M})(\theta) - (\mathbb{M}_n - \mathbb{M})(\theta_0)\right| \lesssim \frac{\phi_n(\delta)}{\sqrt{n}},$$

  for functions ϕ_n with $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ decreasing for some $\alpha < 2$.

- $r_n^2 \phi_n(r_n^{-1}) \leq \sqrt{n}, \quad \forall n.$
- $\hat{\theta}_n \to_{P^*} \theta_0$ *and* $\mathbb{M}_n(\hat{\theta}_n) \geq \mathbb{M}_n(\theta_0) - O_P(r_n^{-2})$.

Then $d(\hat{\theta}_n, \theta_0) = O_P^*(r_n^{-1})$.

Rate of convergence

Proof.

- Let for simplicity $\mathbb{M}_n(\hat{\theta}_n) \geq \sup_\theta \mathbb{M}_n(\theta)$.

- For each n, partition $\Theta \setminus \theta_0$ into "shells"

$$S_{j,n} = \{\theta \colon 2^{j-1} < r_n d(\theta, \theta_0) \leq 2^j\}, \quad j \in \mathbb{N}.$$

- Suppose $r_n d(\hat{\theta}_n, \theta_0) > 2^M$ for some for some $M \in \mathbb{N}$. Then $\hat{\theta}_n \in S_{j,n}$ for some $j > M$ and $\sup_{\theta \in S_{j,n}} \mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0) \geq 0$.

- $$\begin{aligned} P^*\big(r_n d(\hat{\theta}_n, \theta_0) > 2^M\big) &\leq \sum_{j>M} P^*\Big(\sup_{\theta \in S_{j,n}} \mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0) \geq 0\Big) \\ &\leq \sum_{j>M, 2^j \leq \eta r_n} P^*\Big(\sup_{\theta \in S_{j,n}} \mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0) \geq 0\Big) \\ &\quad + P^*\big(2d(\hat{\theta}_n, \theta_0) \geq \eta\big). \end{aligned}$$

- Because $\hat{\theta}_n \xrightarrow{P^*} \theta_0$, $P^*\big(2d(\hat{\theta}_n, \theta_0) \geq \eta\big) \to 0$ for every $\eta > 0$.

Rate of convergence

Proof (ct'd).

- Choose η small enough for the conditions of the theorem to hold.
- Then for every $\theta \in S_{j,n} = \{\theta\colon 2^{j-1} < r_n d(\theta, \theta_0) \le 2^j\}$,

$$\mathbb{M}(\theta) - \mathbb{M}(\theta_0) \lesssim -d^2(\theta, \theta_0) < -\left(\frac{2^{j-1}}{r_n}\right)^2.$$

- Defining $W_n = \mathbb{M}_n - \mathbb{M}$, we get

$$\begin{aligned}
&\sum_{j>M, 2^j \le \eta r_n} P^*\left(\sup_{\theta \in S_{j,n}} \mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0) \ge 0\right) \\
&= \sum_{j>M, 2^j \le \eta r_n} P^*\left(\sup_{\theta \in S_{j,n}} W_n(\theta) - W_n(\theta_0) \ge -(\mathbb{M}(\theta) - \mathbb{M}(\theta_0))\right) \\
&\le \sum_{j>M, 2^j \le \eta r_n} P^*\left(\sup_{\theta \in S_{j,n}} |W_n(\theta) - W_n(\theta_0)| \gtrsim (2^{j-1}/r_n)^2\right).
\end{aligned}$$

Rate of convergence

Proof (ct'd).

- Recall that $S_{j,n} \subset \{d(\theta, \theta_0) \leq 2^j / r_n\}$.

- Then Markov's inequality gives

$$\sum_{j > M, 2^j \leq \eta r_n} P^* \left(\sup_{\theta \in S_{j,n}} |W_n(\theta) - W_n(\theta_0)| \gtrsim (2^{j-1}/r_n)^2 \right)$$

$$\leq \sum_{j > M} \frac{\phi_n(2^j / r_n)}{\sqrt{n}(2^{j-1}/r_n)^2}.$$

- Observe that $\phi_n(c\delta) \leq c^\alpha \phi_n(\delta)$ for every $c > 1$ and recall $r_n^2 \phi_n(r_n^{-1}) \leq \sqrt{n}$.

- This gives

$$\sum_{j > M} \frac{\phi_n(2^j / r_n)}{\sqrt{n}(2^{j-1}/r_n)^2} \leq \sum_{j > M} \frac{2^{j\alpha} \phi_n(1/r_n)}{\sqrt{n}(2^{j-1}/r_n)^2} \leq \sum_{j > M} \frac{2^{j\alpha}}{(2^{j-1})^2} \overset{M \to \infty}{\to} 0,$$

because $\alpha < 2$. $\square$

Rate of convergence: iid case

Corollary (3.2.6)

Let $\mathbb{M}_n(\theta) = \mathbb{P}_n m_\theta$, $\mathbb{M}(\theta) = Pm_\theta$, $\sqrt{n}(\mathbb{M}_n - \mathbb{M}) = \mathbb{G}_n m_\theta$. *Suppose:*

- *For every* θ *in a neighborhood of* θ_0,
$$P(m_\theta - m_{\theta_0}) \lesssim -d^2(\theta, \theta_0).$$
- *There is* ϕ *with* $\delta \mapsto \phi(\delta)/\delta^\alpha$ *for some* $\alpha < 2$, *and for* δ *small,*
$$\mathrm{E}^* \sup_{d(\theta,\theta_0)<\delta} |\mathbb{G}_n(m_\theta - m_{\theta_0})| \lesssim \phi(\delta).$$
- $r_n^2 \phi(1/r_n) \leq \sqrt{n}, \quad \forall n.$
- $\hat{\theta}_n \to_{P^*} \theta_0$ *and* $\mathbb{P}_n m_{\hat{\theta}_n} \geq \mathbb{P}_n m_{\theta_0} - O_P(r_n^{-2})$.

Then $d(\hat{\theta}_n, \theta_0) = O_P^*(r_n^{-1})$.

Comments

- $\phi(\delta) = \delta^\alpha$ gives $r_n \geq n^{1/(4-2\alpha)}$.
- $P(m_\theta - m_{\theta_0}) \lesssim -d^2(\theta, \theta_0)$ holds if $\theta \mapsto Pm_\theta$ has two continuous derivatives (2nd nonsingular).
- To find ϕ, we can bound the continuity modulus

$$\mathrm{E}^* \sup_{d(\theta,\theta_0)<\delta} |\mathbb{G}_n(m_\theta - m_{\theta_0})|$$

  by entropy integrals of $\mathcal{M}_\delta = \{m_\theta - m_{\theta_0} \colon d(\theta, \theta_0) < \delta\}$ with respect to an envelope M_δ.
- If $J(1, \mathcal{M}_\delta, L_2)$ or $J_{[]}(1, \mathcal{M}_\delta, L_2(P))$ are bounded as $\delta \swarrow 0$, then we can take $\phi^2(\delta) = P^* M_\delta^2$ (Theorems 2.14.1 and 2.14.15).
- Book gives many detailed examples of applications: Lipschitz in a parameter, location estimation, monotone density estimation, current status distribution.

Outline

1. Consistency
2. Rate of convergence
3. Asymptotic distribution

Linearization

- Idea: Taylor expand the criterion function.

$$\begin{aligned} n\mathbb{P}_n(m_\theta - m_{\theta_0}) &= nP(m_\theta - m_{\theta_0}) + \sqrt{n}\mathbb{G}_n(m_\theta - m_{\theta_0}) \\ &\approx \frac{1}{2}\sqrt{n}(\theta-\theta_0)'V\sqrt{n}(\theta-\theta_0) + \sqrt{n}(\theta-\theta_0)\mathbb{G}_n\dot{m}_{\theta_0} \\ &\quad + o_P(\sqrt{n}\|\theta-\theta_0\|). \end{aligned}$$

- Forgetting about the remainder, this is maximized for

$$\sqrt{n}(\theta-\theta_0) = -V^{-1}\mathbb{G}_n\dot{m}_{\theta_0}.$$

- Thus, we expect the M-estimator $\hat{\theta}_n$ to satisfy

$$\sqrt{n}(\hat{\theta}_n-\theta_0) = -V^{-1}\mathbb{G}_n\dot{m}_{\theta_0} + o_P(1)$$

Linearization

Theorem (3.2.16)

Let $\mathbb{M}_n$ be index by open $\Theta \subset \mathbb{R}^d$ and $\mathbb{M}\colon \Theta \to \mathbb{R}$ deterministic. Suppose:

- $\theta \mapsto \mathbb{M}(\theta)$ *has two continuous derivatives at its maximum* θ_0 *with non-singular Hessian* V.
- *For every* $\tilde{\theta}_n = \theta_0 + o_P^*(1)$, *the 'stochastic differentiability' condition*

$$\begin{aligned} &r_n(\mathbb{M}_n - \mathbb{M})(\tilde{\theta}_n) - r_n(\mathbb{M}_n - \mathbb{M})(\theta_0) \\ &= (\tilde{\theta}_n - \theta_0)' Z_n + o_P^*(\|\tilde{\theta}_n - \theta_0\| + r_n\|\tilde{\theta}_n - \theta_0\| + r_n^{-1}), \end{aligned}$$

  holds with some process Z_n *uniformly tight.*

- $\hat{\theta}_n \overset{P^*}{\to} \theta_0$ *and* $\mathbb{M}_n(\hat{\theta}_n) \geq \sup_\theta \mathbb{M}_n(\theta) - o_P(r_n^{-2})$.

Then $r_n(\hat{\theta}_n - \theta_0) = -V^{-1} Z_n + o_P^*(1)$.

If $r_n(\hat{\theta}_n - \theta_0)$ is uniformly tight, only consider $\tilde{\theta}_n = \theta_0 + O_P^*(r_n^{-1})$.

Linearization

Proof.

- Let for simplicity $\mathbb{M}_n(\hat{\theta}_n) \geq \sup_\theta \mathbb{M}_n(\theta)$.

- For every $\tilde{h}_n = o_P^*(1)$, stochastic differentiability and smoothness of $\theta \mapsto M(\theta)$ gives

$$\mathbb{M}_n(\theta_0 + \tilde{h}_n) - \mathbb{M}_n(\theta_0) = \frac{\tilde{h}_n' V \tilde{h}_n}{2} + r_n^{-1} \tilde{h}_n' Z_n + o_P^*(\|\tilde{h}_n\|^2 + r_n^{-1}\|\tilde{h}_n\| + r_n^{-2}). \quad (*)$$

- Substitute in $\hat{h}_n = \hat{\theta}_n - \theta_0$ for $\tilde{h}_n$. By definition of $\hat{\theta}_n$, LHS ≥ 0.

- Because V is negative definite, $\hat{h}_n' V \hat{h}_n \lesssim -\|\hat{h}_n\|^2$.

- Hence, the first display implies

$$0 \leq -\|\hat{h}_n\|^2 + r_n^{-1}\|\hat{h}_n\| O_P(1) + o_P(\|\hat{h}_n\|^2 + r_n^{-2}).$$

$\Rightarrow \|\hat{h}_n\| = O_P(r_n^{-1})$.

Linearization

Proof (ct'd).

- Invoke $(*)$ with $\tilde{h}_n = \hat{h}_n$ and $\tilde{h}_n = -r_n^{-1} V^{-1} Z_n$:

$$\mathbb{M}_n(\theta_0 + \hat{h}_n) - \mathbb{M}_n(\theta_0) = \frac{\hat{h}_n' V \hat{h}_n}{2} + r_n^{-1} \hat{h}_n' Z_n + o_P^*(r_n^{-2})$$

$$\mathbb{M}_n(\theta_0 - r_n^{-1} V^{-1} Z_n) - \mathbb{M}_n(\theta_0) = -\frac{(V^{-1} Z_n)' V (V^{-1} Z_n)}{2 r_n^2} + o_P^*(r_n^{-2}).$$

- Subtract the second from the first:

$$\frac{(\hat{h}_n + r_n^{-1} V^{-1} Z_n)' V (\hat{h}_n + r_n^{-1} V^{-1} Z_n)}{2} \geq -o_P(r_n^{-2}).$$

- $\Rightarrow \hat{h}_n = -r_n^{-1} V^{-1} Z_n + o_P(r_n^{-1})$.
- $\Rightarrow r_n(\hat{\theta}_n - \theta_0) = V^{-1} Z_n + o_P(1)$. $\square$

More linearization

- Lemma 3.2.19, 3.2.21, and eq. (3.2.22) give easier conditions to verify "differentiability" in the *iid* case.
- Corollary 3.2.23 summarizes easy conditions to establish asymptotic normality for the Euclidean *iid* case.
- More linearization next week with Z-estimators.