Classical Non-Linear Methods: Consistency of M-Estimators

Bertel Schjerning

University of Copenhagen, Department of Economics

February 9, 2026

Plan for Classical Non-Linear Methods

Lecture 4: M-estimation, Intro, Non-linear LS (W.12)

Lecture 5: Asymptotic properties of M-estimators (W.12)

- Consistency, Asymptotic Normality

Lecture 6: M-estimator inference, Variance estimation (W.12)

Lecture 7: Maximum likelihood estimation (W.13)

Outline

M-Estimation Framework

M-Estimand

We now consider more abstract setting.

Let $q(\mathbf{w}, \boldsymbol{\theta})$ denote function of

- random vector $\mathbf{w}$ [observables, typically $\mathbf{w} = (\mathbf{y}, \mathbf{x})$],
- parameters $\boldsymbol{\theta}$.

"True theta" $\boldsymbol{\theta}_o$ framed as a solution to PP

$$\boldsymbol{\theta}_o \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \mathrm{E}\left[q(\mathbf{w}, \boldsymbol{\theta})\right]. \tag{PP}$$

"M" short for "minimization" (/"maximization")

q sometimes called loss function.

M-Estimator

Given i.i.d. observations $\{\mathbf{w}_i\}_{i=1}^N$.

Analogy principle suggests sample problem (SP)

$$\min_{\boldsymbol{\theta} \in \Theta} \frac{1}{N} \sum_{i=1}^{N} q\left(\mathbf{w}_i, \boldsymbol{\theta}\right). \tag{SP}$$

Definition: Any SP solution is an M-estimator of $\boldsymbol{\theta}_o$.

Example M-Estimators

- OLS: $$q(\mathbf{w}, \boldsymbol{\theta}) = (y - \mathbf{x}\boldsymbol{\theta})^2.$$
- NLS: $$q(\mathbf{w}, \boldsymbol{\theta}) = [y - m(\mathbf{x}, \boldsymbol{\theta})]^2.$$
- Maximum likelihood: $$q(\mathbf{w}, \boldsymbol{\theta}) = -\ln f(y|\mathbf{x}; \boldsymbol{\theta}).$$
- Least absolute deviations (LAD): $$q(\mathbf{w}, \boldsymbol{\theta}) = |y - \mathbf{x}\boldsymbol{\theta}|.$$
  - ... used for linear *median* regression $\mathrm{Med}(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\theta}$.
- $\vdots$

Scope of Framework

Observables $\mathbf{w}_i$ allow scalar/vector outcome.

- One equation, one cross section $\Rightarrow$ scalar y_i.
- Multiple equations, one cross section $\Rightarrow$ J-vector $\mathbf{y}_i$.
  - **Ex:** Joint labor supply decision (wife/husband),

$$
\begin{aligned}
y_i^{\mathrm{w}} &= \text{labor supply, wife, family } i, \\
y_i^{\mathrm{h}} &= \text{labor supply, husband, family } i.
\end{aligned}
$$

- One equation, panel data $\Rightarrow$ vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$.
  - FE: $q\left(\mathbf{w}_i, \boldsymbol{\theta}\right) = \sum_{t=1}^{T} \left(\ddot{y}_{it} - \ddot{\mathbf{x}}_{it}\boldsymbol{\theta}\right)^2$
- Multiple equations, panel data $\Rightarrow$ $\mathbf{y}_i$ is a $T \times J$-vector

Formulation very general

Consistency of M-Estimators

Recap: Setting

M-estimand solves population problem (PP),

$$\boldsymbol{\theta}_o \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \mathrm{E}\left[q\left(\mathbf{w}, \boldsymbol{\theta}\right)\right]. \tag{PP}$$

M-estimator solves sample problem (SP),

$$\widehat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \frac{1}{N} \sum_{i=1}^{N} q\left(\mathbf{w}_i, \boldsymbol{\theta}\right). \tag{SP}$$

Q: Properties of M-estimators? Consistency? Normality?

Informal Look at Consistency

Criterion functions (minimands) and minimizers:

$$N^{-1}\sum_{i=1}^{N} q(\mathbf{w}_i, \boldsymbol{\theta}) \qquad\qquad \mathrm{E}\left[q(\mathbf{w}, \boldsymbol{\theta})\right]$$

$$\widehat{\boldsymbol{\theta}} \qquad\qquad \boldsymbol{\theta}_o$$

Q: Relationships?

Informal Look at Consistency

By definition of M-estimand and M-estimator:

$$N^{-1}\sum_{i=1}^{N} q(\mathbf{w}_i, \boldsymbol{\theta}) \qquad \mathrm{E}\left[q(\mathbf{w}, \boldsymbol{\theta})\right]$$

$$\rotatebox{90}{$\in \text{argmin}$} \qquad \rotatebox{90}{$\in \text{argmin}$}$$

$$\widehat{\boldsymbol{\theta}} \qquad \boldsymbol{\theta}_o$$

Informal Look at Consistency

By (weak) law of large numbers, for each $\boldsymbol{\theta} \in \Theta$,

$$\begin{array}{ccc}
N^{-1}\sum_{i=1}^{N} q(\mathbf{w}_i, \boldsymbol{\theta}) & \xrightarrow[\text{LLN}]{p} & \mathrm{E}\left[q(\mathbf{w}, \boldsymbol{\theta})\right] \\
\rotatebox{90}{$\in \operatorname{argmin}$} & & \rotatebox{90}{$\in \operatorname{argmin}$} \\
\widehat{\boldsymbol{\theta}} & & \boldsymbol{\theta}_o
\end{array}$$

Informal Look at Consistency

$$\begin{array}{ccc}
N^{-1}\sum_{i=1}^{N} q(\mathbf{w}_i,\boldsymbol{\theta}) & \xrightarrow{p} & \mathrm{E}\left[q(\mathbf{w},\boldsymbol{\theta})\right] \\
\rotatebox{90}{$\in \operatorname{argmin}$} & & \rotatebox{90}{$\in \operatorname{argmin}$} \\
\widehat{\boldsymbol{\theta}} & \xrightarrow[?]{p} & \boldsymbol{\theta}_o
\end{array}$$

Seems reasonable...

Q: When does mini*mand* convergence imply mini*mizer* convergence (in prob)?

Formal Look at Consistency

Q: When is $\widehat{\boldsymbol{\theta}}$ consistent for $\boldsymbol{\theta}_o$?

Suffices (essentially) following two conditions hold:

1. Identification: $\boldsymbol{\theta}_o$ is identified.
2. Uniform Law of Large Numbers: S minimand converges to P equivalent *uniformly in probability*,

$$\max_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{N}\sum_{i=1}^{N} q\left(\mathbf{w}_i,\boldsymbol{\theta}\right)-\mathrm{E}\left[q\left(\mathbf{w},\boldsymbol{\theta}\right)\right]\right|\xrightarrow{p}0.$$

Identification Assumption

Without further structure, *assume* identification, i.e.

$$\mathrm{E}\left[q\left(\mathbf{w},\boldsymbol{\theta}\right)\right] > \mathrm{E}\left[q\left(\mathbf{w},\boldsymbol{\theta}_o\right)\right] \text{ for all } \boldsymbol{\theta}\in\Theta\backslash\left\{\boldsymbol{\theta}_o\right\}.$$

In words: $\boldsymbol{\theta}_o$ *unique* solution to PP,

$$\boldsymbol{\theta}_o = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \mathrm{E}\left[q\left(\mathbf{w},\boldsymbol{\theta}\right)\right].$$

- May reduce/interpret in specific models (later).

Uniform Law of Large Numbers

May *deduce* minimand convergence using:

Theorem (W. Theorem 12.1)

If

1. $\Theta \subseteq \mathbb{R}^P$ *nonempty compact (i.e. closed + bounded),*
2. $q(\mathbf{w}, \cdot)$ *continuous (in* $\boldsymbol{\theta}$*),*

(and additional technical conditions hold), then

$$\max_{\boldsymbol{\theta} \in \Theta} \left| \frac{1}{N} \sum_{i=1}^{N} q(\mathbf{w}_i, \boldsymbol{\theta}) - \mathrm{E}\left[q(\mathbf{w}, \boldsymbol{\theta})\right] \right| \xrightarrow{p} 0.$$

Uniform law of large numbers (ULLN).

Consistency Theorem

Theorem (W. Theorem 12.2)

Under the assumptions of W. Theorem 12.1 (ULLN) and assuming identification of $\boldsymbol{\theta}_o$,

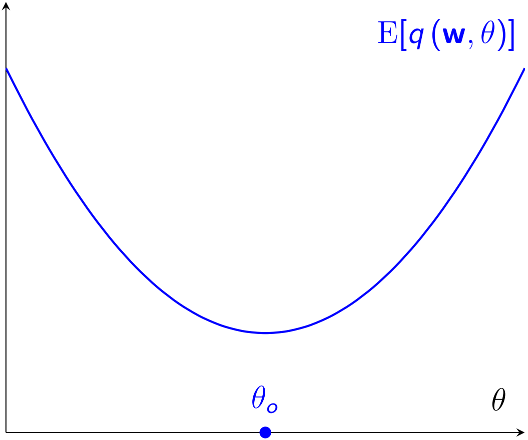
1. *SP has solution, $\widehat{\boldsymbol{\theta}}$, and*
2. *$\widehat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}_o$, $\widehat{\boldsymbol{\theta}} \to_p \boldsymbol{\theta}_o$.*

Sketch of Consistency Argument

Proof Sketch:

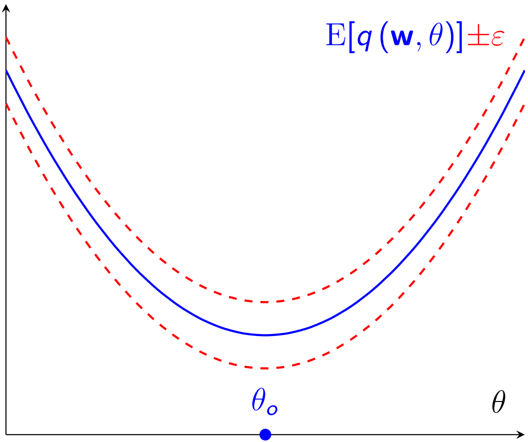
1. Compact Θ + $q(\mathbf{w}, \cdot)$ continuous $\Rightarrow$ SP solution exists.
   - A continuous function defined on a compact set...
2. ULLN $\Rightarrow$ in limit, S/P minimands coincide (in prob).
3. Identification implies unique PP solution, so must have $\widehat{\boldsymbol{\theta}} \to_p \boldsymbol{\theta}_o$.

Graphical Illustration of Consistency

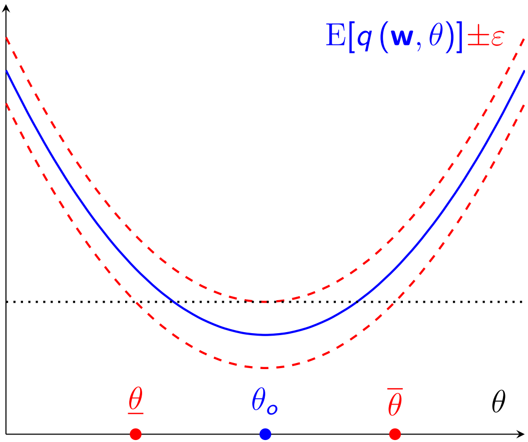


Graphical Illustration of Consistency

When minimand difference $\leqslant \varepsilon$, S minimand in "sleeve"



Graphical Illustration of Consistency



$\widehat{\theta}$ "squeezed in"

Necessity of Uniform Convergence

Uniform convergence sufficient but not necessary.

We use it to *deduce* minimizer existence and convergence.

- Other arguments exist.

Think: Linear model + squared loss

$$q(\mathbf{w}, \boldsymbol{\theta}) = (y - \mathbf{x}\boldsymbol{\theta})^2 .$$

- Natural parameter space entire $\mathbb{R}^P$. Not compact...
- ... but objective convex

If $q(\mathbf{w}, \boldsymbol{\theta})$ convex in $\boldsymbol{\theta}$, so is $N^{-1} \sum_i q(\mathbf{w}_i, \boldsymbol{\theta})$. $[= \widehat{Q}_N(\boldsymbol{\theta})]$

Consistency without Compactness

Theorem (Newey and McFadden, 1994)

Let

1. $Q : \mathbb{R}^P \to \mathbb{R}$ *be uniquely minimized at* $\boldsymbol{\theta}_o$*;*
2. *each (random)* $\widehat{Q}_N : \mathbb{R}^P \to \mathbb{R}, N = 1, 2, \ldots,$ *convex; and,*
3. $\widehat{Q}_N(\boldsymbol{\theta}) \to_p Q(\boldsymbol{\theta})$ *for each* $\boldsymbol{\theta} \in \mathbb{R}^P$*.*

Then

1. *a minimizer* $\widehat{\boldsymbol{\theta}}_N$ *of* $\widehat{Q}_N$ *exists with probability* $\to 1$*; and*
2. *for any minimizer selection,* $\widehat{\boldsymbol{\theta}}_N \to_p \boldsymbol{\theta}_o$*.*