Classical Non-Linear Methods: Asymptotic Normality of M-Estimators:

Bertel Schjerning

University of Copenhagen, Department of Economics

January 29, 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

Recap: M-Estimation Framework

Recap: M-Estimand

Let $q\left(\mathbf{w}, \boldsymbol{\theta}\right)$ denote loss function, depending on

1. random vector $\mathbf{w}$ [observables, e.g. $\mathbf{w} = \left(\mathbf{y}, \mathbf{x}\right)$],
2. parameters $\boldsymbol{\theta}$.

"True" parameter $\boldsymbol{\theta}_o$ assumed solution to population problem

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

M is for minimization/maximization.

Recap: M-Estimator

Given random (as in i.i.d.) sample $\{\mathbf{w}_i\}_{i=1}^{N}$.

Analogy principle suggests solving sample problem

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

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

Asymptotic Properties of M-Estimators

Recap: Setting

M-estimand solves population problem,

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

M-estimator solves sample problem,

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

Q: Properties of such $\{\widehat{\boldsymbol{\theta}}_N\}_{N=1}^{\infty}$?

Recap: Consistency

M-Estimator Consistency with Compactness

Theorem (W. Theorem 12.2)

If

1. $\boldsymbol{\theta}_0$ *is the unique solution to PP* *(“identification”)*
2. $\Theta \subseteq \mathbb{R}^P$ *compact* *(i.e.* Θ *closed + bounded),*
3. $q(\mathbf{w}, \cdot)$ *continuous (in* $\boldsymbol{\theta}$*),*

(+ technical conditions), then

1. *SP has a solution (i.e.* $\widehat{\boldsymbol{\theta}}_N$ *exists), and*
2. *any selection* $\{\widehat{\boldsymbol{\theta}}_N\}_{N=1}^{\infty}$ *of minimizers is consistent for* $\boldsymbol{\theta}_o$*,* $\widehat{\boldsymbol{\theta}} \rightarrow_p \boldsymbol{\theta}_o$.

M-Estimator Consistency without Compactness

Theorem (Newey and McFadden, 1994)

Let

1. $Q : \mathbb{R}^P \to \mathbb{R}$ *be uniquely minimized at* $\boldsymbol{\theta}_o$; *(ID'n)*
2. *each (random)* $\{\widehat{Q}_N : \mathbb{R}^P \to \mathbb{R}\}_{N=1}^{\infty}$ *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 selection* $\{\widehat{\boldsymbol{\theta}}_N\}_{N=1}^{\infty}$ *of minimizers,* $\widehat{\boldsymbol{\theta}}_N \to_p \boldsymbol{\theta}_o$.

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

Normality

Additional Assumptions

Have for consistency (as in W. Thm. 12.1) invoked:

- $\boldsymbol{\theta}_o$ identified
- Θ compact
- $q(\mathbf{w}, \cdot)$ continuous

(+ technical...)

Asymptotic normality requires *stronger* assumptions.

Additional Assumptions

For asymptotic normality, add:

- $\boldsymbol{\theta}_o$ interior to Θ. [Draw]
- $q(\mathbf{w}, \cdot)$ twice continuously differentiable on $\operatorname{int}\Theta$

Remarks:

- Interiority requires $\operatorname{int}\Theta$ non-empty
- ... used to expand around $\boldsymbol{\theta}_o$
- Twice cont' diff' facilitates second-order expansion.

Additional Assumptions

Abbreviate

$$\text{Score:} \quad \mathbf{s}(\mathbf{w}, \boldsymbol{\theta}) := \frac{\partial}{\partial \boldsymbol{\theta}} q(\mathbf{w}, \boldsymbol{\theta}), \qquad (P \times 1)$$

$$\text{Hessian :} \quad \mathbf{H}(\mathbf{w}, \boldsymbol{\theta}) := \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} q(\mathbf{w}, \boldsymbol{\theta}). \qquad (P \times P)$$

Further add:

- $\mathrm{E}[\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o)] = \mathbf{0}$,
- $\mathrm{E}[\mathbf{H}(\mathbf{w}, \boldsymbol{\theta}_o)]$ positive definite.
  - Essentially FOC/SOC for minimization.

Asymptotic Normality of M-Estimators

Theorem (W. Theorem 12.3)

Provided

- $\boldsymbol{\theta}_o$ *unique PP solution + interior to* Θ *compact,*
- $q(\mathbf{w}, \cdot)$ *cont's + twice cont'ly differentiable on* $\operatorname{int}\Theta$,
- $\mathrm{E}[\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o)] = \mathbf{0}$, *and* $\mathrm{E}[\mathbf{H}(\mathbf{w}, \boldsymbol{\theta}_o)]$ *positive definite,*

(+ technical), any selection $\{\widehat{\boldsymbol{\theta}}_N\}_{N=1}^{\infty}$ *of minimizers satisfies*

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_o) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}\right),$$

$$\mathbf{A}_o := \mathrm{E}[\mathbf{H}(\mathbf{w}, \boldsymbol{\theta}_o)],$$

$$\mathbf{B}_o := \mathrm{E}\left[\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o)\,\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o)'\right].$$

Mean Value Theorem

- Normality proof relies on *mean value theorem*.
- Consider *scalar* case ($P = 1$).

Mean Value Theorem (MVT):

- Let $f : [a, b] \to \mathbb{R}$ continuous + differentiable on (a, b).
- Then for some $c \in (a, b)$,

$$f(b) - f(a) = f'(c)(b - a).$$

- Slope of secant attained somewhere in between. [Draw]

Proof Sketch: Mean Value Theorem

In scalar $(P = 1)$ case,

$$s(\mathbf{w}, \theta) = \frac{\partial}{\partial \theta} q(\mathbf{w}, \theta), \quad H(\mathbf{w}, \theta) = \frac{\partial^2}{\partial^2 \theta} q(\mathbf{w}, \theta).$$

We know that $\widehat{\theta}_N \in \operatorname{int} \Theta$ wp $\to 1$. (Why?)

So: Twice cont' diff' + MVT with f = score average yields

$$\frac{1}{N} \sum_{i=1}^{N} s(\mathbf{w}_i, \widehat{\theta}_N) - \frac{1}{N} \sum_{i=1}^{N} s(\mathbf{w}_i, \theta_o) = \frac{1}{N} \sum_{i=1}^{N} H(\mathbf{w}_i, \overline{\theta}_N)(\widehat{\theta}_N - \theta_o).$$

$\widehat{\theta}_N$ solves SP, so LHS vanishes. (FOC.)

Proof Sketch: Rearrange

Have argued:

$$-\frac{1}{N}\sum_{i=1}^{N} s\left(\mathbf{w}_i, \theta_o\right) = \frac{1}{N}\sum_{i=1}^{N} H\left(\mathbf{w}_i, \overline{\theta}_N\right)\left(\widehat{\theta}_N - \theta_o\right).$$

Isolate $\widehat{\theta}_N - \theta_o$ and $\times\sqrt{N}$:

$$\sqrt{N}(\widehat{\theta}_N - \theta_o) = \left[-\frac{1}{\sqrt{N}}\sum_{i=1}^{N} s\left(\mathbf{w}_i, \theta_o\right)\right] \Big/ \left[\frac{1}{N}\sum_{i=1}^{N} H\left(\mathbf{w}_i, \overline{\theta}_N\right)\right].$$

Analyze each RHS factor in turn.

Proof Sketch: Denominator

$$\sqrt{N}(\widehat{\theta}_N - \theta_o) = \left[-\frac{1}{\sqrt{N}} \sum_{i=1}^{N} s\left(\mathbf{w}_i, \theta_o\right) \right] \Big/ \left[\frac{1}{N} \sum_{i=1}^{N} H\left(\mathbf{w}_i, \overline{\theta}_N\right) \right].$$

$\overline{\theta}_N$ trapped between $\widehat{\theta}_N$ and $\theta_o \Rightarrow \overline{\theta}_N \to_p \theta_o$.

So $N^{-1} \sum_{i=1}^{N} H\left(\mathbf{w}_i, \overline{\theta}_N\right) \approx N^{-1} \sum_{i=1}^{N} H\left(\mathbf{w}_i, \theta_o\right)$. (sketch)

$N^{-1} \sum_{i=1}^{N} H\left(\mathbf{w}_i, \theta_o\right) \to_p \mathrm{E}\left[H\left(\mathbf{w}, \theta_o\right)\right] = A_o > 0$. (p.d.)

$$\Longrightarrow\ 1 \Big/ \frac{1}{N} \sum_{i=1}^{N} H\left(\mathbf{w}_i, \overline{\theta}\right) \xrightarrow{p} 1/A_o. \qquad \text{(CMT/Slutsky)}$$

Proof Sketch: Numerator

$$\sqrt{N}(\widehat{\theta}_N - \theta_o) = \left[-\frac{1}{\sqrt{N}} \sum_{i=1}^{N} s\left(\mathbf{w}_i, \theta_o\right) \right] \Bigg/ \left[\frac{1}{N} \sum_{i=1}^{N} H\left(\mathbf{w}_i, \overline{\theta}_N\right) \right].$$

I.i.d. + mean-zero scores + CLT combine to yield

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} s\left(\mathbf{w}_i, \theta_o\right) \xrightarrow{d} \mathrm{N}\left(0, B_o\right), \quad B_o = \mathrm{E}[s\left(\mathbf{w}, \theta_o\right)^2].$$

Proof Sketch

Harvesting our results,

$$\sqrt{N}(\widehat{\theta}_N - \theta_o) = \underbrace{\left[-\frac{1}{\sqrt{N}}\sum_{i=1}^{N} s\left(\mathbf{w}_i, \theta_o\right)\right]}_{\to_d \mathrm{N}(0, B_o)} \Bigg/ \underbrace{\left[\frac{1}{N}\sum_{i=1}^{N} H\left(\mathbf{w}_i, \overline{\theta}_N\right)\right]}_{\to_p A_o}$$

$$\xrightarrow{d} \mathrm{N}\left(0, B_o\right)/A_o \qquad \text{(product rule + Slutsky)}$$

$$\stackrel{d}{=} \mathrm{N}\left(0, B_o/A_o^2\right). \qquad \text{(linear(normal)=normal)}$$

Proof in vector-case analogous:

1. Linear approximation of score average (MVT)
2. Convergence of inverse Hessian (ULLN+CMT)
3. Convergence of scaled score average (CLT + product rule)

Discussion

- Thm. gives conditions for *any* M-estimator to be asymptotically normal.
- Implies sandwich form
$$\text{Avar}(\widehat{\boldsymbol{\theta}}) = \mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}/N.$$
- Akin to earlier results (with estimators in closed form).
- Note: $\text{Avar}(\widehat{\boldsymbol{\theta}})$ depends on $\boldsymbol{q}$.
- Ideally: Choose $\boldsymbol{q}$ to get small variance.

Discussion

- $\mathbf{A}_o = \mathrm{E}\left[\mathbf{H}\left(\mathbf{w}, \boldsymbol{\theta}_o\right)\right]$ assumed positive definite.
- Zero on diagonal $\approx$ infinite variance (through $\mathbf{A}_o^{-1}$)
- Failure of p.d $\approx$ P minimand flat around $\boldsymbol{\theta}_o$
- $\approx$ Identification failure.

Role of Interiority

We used $\boldsymbol{\theta}_o \in \operatorname{int} \Theta$ for differentiation (Where?)

Q: What if $\boldsymbol{\theta}_o$ on boundary of parameter space?

A: No reason to expect $\sqrt{N}$-asymptotic normality.

Example: Parameter on Boundary

Let $y_i \sim$ i.i.d. $(\theta_o, 1)$ with θ_o known $\geqslant 0$.

Nonnegativity enforced

$$\widehat{\theta}_N := \max(0, \overline{y}_N) = \underset{\theta \geqslant 0}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} (y_i - \theta)^2 ,$$

If $\theta_o = 0$ (boundary case), then $\sqrt{N}(\widehat{\theta}_N - 0) \geqslant 0$.

$\sqrt{N}(\widehat{\theta}_N - 0)$ does $\rightarrow_d$... but not to normal.