M-estimation of Linear Models

Consistency, Asymptotic Normality and Bahadur Representations

Presenter: Ching-Wei Cheng

Department of Statistics
Purdue University

March 23, 2015
(Group Meeting)

M-Estimation of Linear Models with Independent Errors

Consider the linear model

$$y_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} + e_i, \quad 1 \le i \le n, \quad e_i \text{ are iid}$$

- M-estimator

$$\widehat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(y_i - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}) \quad \text{or} \quad \sum_{i=1}^{n} \psi(y_i - \boldsymbol{x}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}_n)\boldsymbol{x}_i = 0,$$

  where ρ is a real-valued function with derivative ψ
- $Q_n = \boldsymbol{X}_n^\mathsf{T}\boldsymbol{X}_n = \sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\mathsf{T}$, where $\boldsymbol{X}_n = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^\mathsf{T}$
- With appropriate regularity conditions, we have

$$\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 = -(\mathsf{E}[\psi'(e_1)]Q_n)^{-1} \sum_{i=1}^{n} \psi(e_i)\boldsymbol{x}_i + o_P(1)$$

  (e.g., van der Vaart (2000, Example 5.28))
- With proper moment conditions on the linear approximation, consistency and asymptotic normality can be obtained in advance

M-Estimation of Linear Models with Independent Errors

Theorem 3.1 in He and Shao (1996)

With some regularity conditions,

$$\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 = -(\mathsf{E}[\psi'(e_1)]Q_n)^{-1} \sum_{i=1}^{n} \psi(e_i)\boldsymbol{x}_i + O_{a.s.}\left(\frac{\log\log n}{n}\right)$$

Why Bahadur representations?

- Typically, an estimator is approximated by a sum of independent variables with a higher-order remainder
- The first-order terms may be used to measure the influence of a single observation or to derive the asymptotic distribution of the estimator
- Bahadur representations could lead to sharp error bound for the high-order remainder, providing a quick guide to how good the linear approximation can be

Outline

M-Estimation of Linear Models with iid Errors

M-Estimation of Linear Models with Dependent Errors
- Consistency and Asymptotic Normality
- Bahadur Representations

Divide-and-Conquer for Big Data

General Model Form and Notations

Consider the linear model

$$y_i = \boldsymbol{x}_i^\mathsf{T} \boldsymbol{\beta} + e_i, \quad 1 \le i \le n$$

- $\boldsymbol{\beta}$ is $p \times 1$ unknown regression coefficient vector
- $\boldsymbol{x}_i$ are $p \times 1$ known design vectors
- $\boldsymbol{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\mathsf{T}$ is $n \times p$ design matrix
- $Q_n = \boldsymbol{X}_n^\mathsf{T} \boldsymbol{X}_n = \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T}$
- M-estimator

$$\widehat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n \rho(y_i - \boldsymbol{x}_i^\mathsf{T} \boldsymbol{\beta}) \quad \text{or} \quad \sum_{i=1}^n \psi(y_i - \boldsymbol{x}_i^\mathsf{T} \widehat{\boldsymbol{\beta}}_n) \boldsymbol{x}_i = 0,$$

  where ρ is a real-valued function with derivative ψ.
- $\nu_{\min}(Q)$ denotes the smallest eigenvalues of a squared matrice Q
- $\lceil a \rceil = \min\{k \in \mathbb{Z} : k \ge a\}$ and $\lfloor a \rfloor = \max\{k \in \mathbb{Z} : k \le a\}$

W.L.O.G., assume the true parameter $\boldsymbol{\beta}_0 = \mathbf{0}$

M-Estimation of Linear Models with iid Errors

He and Shao (1996)
A General Bahadur Representation of M-estimators and Its Application to Linear Regression with Nonstochastic Designs.
The Annals of Statistics, Vol. 24, No. 6, 2608–2630

- Consider the linear model with iid errors

$$y_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} + e_i, \quad 1 \le i \le n, \quad e_i \text{ are iid}$$

- In this paper...
  - $|\boldsymbol{v}| = \max\{|v_1|, \cdots, |v_p|\}$
  - $\lambda_i(\boldsymbol{\beta}) = \mathsf{E}\left[\psi(y_i - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta})\right]\boldsymbol{x}_i$
  - $\Lambda_n(\boldsymbol{\beta}) = \sum_{i=1}^n \lambda_i(\boldsymbol{\beta}) = \sum_{i=1}^n \mathsf{E}\left[\psi(y_i - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta})\right]\boldsymbol{x}_i$
  - f is the density of e_i

M-Estimation of Linear Models with iid Errors

Bahadur representation

(C1) Both ψ and f' are Lipschitz

(C2) $\mathsf{E}[\psi(e_1)] = 0$, $\gamma = \int_{-\infty}^{\infty} \psi(u) f'(u)\, \mathrm{d}u \neq 0$ and $\mathsf{E}[\psi^{2+\epsilon}(e_1)] < \infty$ for some $\epsilon > 0$

(C3) $n^{-1} Q_n \to Q$ (i.e., $Q_n = O(n)$) for some positive definite matrix Q and $\sum_{i=1}^{n} |\boldsymbol{x}_i|^{4+\epsilon} = O(n)$ for some $\epsilon > 0$

Theorem 3.1 in He and Shao (1996)

With (C1)–(C3),

$$\widehat{\boldsymbol{\beta}}_n = -(\gamma Q_n)^{-1} \sum_{i=1}^{n} \psi(e_i) \boldsymbol{x}_i + O_{a.s.}\left(\frac{\log \log n}{n} \right)$$

Note:

- ρ is not required convex in this work
- $\mathsf{E}[\psi(e_1)] = 0$ implies $\lambda_i(\mathbf{0}) = 0 \; \forall\, i \;\Rightarrow\; \Lambda_n(\mathbf{0}) = 0$

Proof of Theorem 3.1 in He and Shao (1996)

To begin,

$$\left|\widehat{\boldsymbol{\beta}}_n + (\gamma Q_n)^{-1} \sum_{i=1}^{n} \psi(e_i)\boldsymbol{x}_i\right| \leq |\widehat{\boldsymbol{\beta}}_n| + c_n^{-1} \underbrace{\left|\sum_{i=1}^{n} \psi(e_i)\boldsymbol{x}_i\right|}_{\mathrm{I}},$$

with $c_n = \frac{1}{2}\gamma n \nu_{\min}(Q) = O(n^{-1})$, $\nu_{\min}(Q)$ is the smallest eigenvalue of Q

Since f' is Lipschitz and $\sum_{i=1}^{n} |\boldsymbol{x}_i|^3 = O(n)$, we also have

$$\begin{aligned} c_n|\widehat{\boldsymbol{\beta}}_n| &\leq \left|\Lambda_n(\widehat{\boldsymbol{\beta}}_n)\right| = \left|\sum_{i=1}^{n} \lambda_i(\widehat{\boldsymbol{\beta}}_n)\right| \\ &\leq \underbrace{\left|\sum_{i=1}^{n} \psi(e_i)\boldsymbol{x}_i\right|}_{\mathrm{I}} + \underbrace{\left|\sum_{i=1}^{n} \left[\psi(e_i)\boldsymbol{x}_i + \lambda_i(\widehat{\boldsymbol{\beta}}_n)\right]\right|}_{\mathrm{II}} \end{aligned}$$

Proof of Theorem 3.1 in He and Shao (1996)

Further conditions are needed: $\exists$ a constant $C < \infty$ and $d_0 > 0$ s.t.

- (Strong consistency) $\widehat{\boldsymbol{\beta}}_n = o_{a.s.}(1)$. This leads to

$$\limsup_{n\to\infty} |\widehat{\boldsymbol{\beta}}_n| \leq (2C)^{-1} \quad \text{a.s.}$$

- (Some known law of iterated logarithm (LIL)) $s_n = O\left(\sum_{i=1}^n |\boldsymbol{x}_i|^2\right) = O(n)$ s.t.

$$\limsup_{n\to\infty} \frac{|\sum_{i=1}^n \psi(e_i)\boldsymbol{x}_i|}{(s_n \log\log n)^{1/2}} \leq 2.5 \quad \text{a.s.}$$

- (Lemma 4.1, not shown) $A_n = O\left(\sum_{i=1}^n |\boldsymbol{x}_i|^4\right) = O(n)$ s.t. $s_n \leq A_n$ and

$$\limsup_{n\to\infty} \sup_{|\boldsymbol{\beta}|\leq d_0} \frac{\left|\sum_{i=1}^n \left[\psi(e_i - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta})\boldsymbol{x}_i - \psi(e_i)\boldsymbol{x}_i - \lambda_i(\boldsymbol{\beta}) + \lambda_i(\mathbf{0})\right]\right|}{(A_n|\boldsymbol{\beta}|^2+1)^{1/2}(\log\log n)^{1/2}} \leq C \quad \text{a.s.}$$

Note: We omit "a.s." in the following inequalities

Proof of Theorem 3.1 in He and Shao (1996)

By the known LIL,

$$\text{I} \le 2.5(s_n \log\log n)^{1/2} \le 2.5(A_n \log\log n)^{1/2}$$
$$\Rightarrow \quad c_n^{-1}\text{I} = O_{a.s.}\left(\left(\frac{\log\log n}{n}\right)^{1/2}\right)$$

Recall $\sum_{i=1}^n \psi(e_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_n)\boldsymbol{x}_i = 0$ and so, as n large enough, by Lemma 4.1,

$$\text{II} = \left|\sum_{i=1}^n \left[\psi(e_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_n)\boldsymbol{x}_i - \psi(e_i)\boldsymbol{x}_i - \lambda_i(\widehat{\boldsymbol{\beta}}_n)\right]\right|$$
$$\le C\left(A_n^{1/2}|\widehat{\boldsymbol{\beta}}_n| + 1\right)(\log\log n)^{1/2}$$

By strong consistency of $\widehat{\boldsymbol{\beta}}_n$, II can be further refined by

$$\text{II} \le \frac{1}{2}(A_n \log\log n)^{1/2}$$

Proof of Theorem 3.1 in He and Shao (1996)

A a sharper bound of $|\widehat{\boldsymbol{\beta}}_n|$ can be obtained by

$$|\widehat{\boldsymbol{\beta}}_n| \leq 3c_n^{-1}(A_n \log\log n)^{1/2}$$

and II can be again sharpened by

$$\text{II} \leq 3Cc_n^{-1}(A_n \log\log n) + C(A_n \log\log n)^{1/2}$$

Sharpening $|\widehat{\boldsymbol{\beta}}_n|$ again by

$$\begin{aligned}
|\widehat{\boldsymbol{\beta}}_n| &\leq 3c_n^{-1}(\log\log n)^{1/2}\left[s_n^{1/2} + c_n^{-1}CA_n(\log\log n)^{1/2} + C\right] \\
&= O_{a.s.}\left(\left(\frac{\log\log n}{n}\right)^{1/2}\right) + O_{a.s.}\left(\frac{\log\log n}{n}\right) + O_{a.s.}\left(\frac{(\log\log n)^{1/2}}{n}\right)
\end{aligned}$$

See application to Big Data

Thus finally we have

$$\widehat{\boldsymbol{\beta}}_n + (\gamma Q_n)^{-1}\sum_{i=1}^{n}\psi(e_i)\boldsymbol{x}_i = O_{a.s.}\left(\frac{\log\log n}{n}\right)$$

Discussion of Theorem 3.1 in He and Shao (1996)

- Three additionally required conditions:
  - Strong consistency of $\widehat{\boldsymbol{\beta}}_n$
  - Known LIL of $\sum_{i=1}^n \psi(e_i)\boldsymbol{x}_i$
  - Known LIL of $\sum_{i=1}^n \left[\psi(e_i - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}) - \mathsf{E}[\psi(e_i - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta})]\right]$ in a neighborhood around $\boldsymbol{\beta}_0 = \mathbf{0}$
- The order of the remainder terms are critically determined by
  - Local oscillations of the M-process

$$\sum_{i=1}^n \psi(e_i - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}) - \mathsf{E}\left[\sum_{i=1}^n \psi(e_i - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta})\right]$$

  - The orders of $\sum_{i=1}^n |\boldsymbol{x}_i|^{4+\epsilon}$

M-Estimation of Linear Models with Dependent Errors

Wu (2007)
M-estimation of Linear Models with Dependent Errors.
Annals of Statistics, Vol. 35, No. 2, 495–521

- Consider the linear model with causal errors

$$y_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} + e_i, \quad 1 \le i \le n, \quad e_i = G(\ldots, \varepsilon_i), \quad \varepsilon_i \text{ are iid}$$

- In this paper...
  - For scalar vectors $\boldsymbol{v} \in \mathbb{R}^p$, $|\boldsymbol{v}| = \left(\sum_{i=1}^p v_i^2\right)^{1/2}$
  - For random vectors
    - $V \in \mathcal{L}^q$ if $\mathsf{E}\big[|V|^q\big] < \infty$ for any $q > 0$
    - $\|V\|_q = \left(\mathsf{E}\big[|V|^q\big]\right)^{1/q}$ and $\|V\| = \|V\|_2$
  - For a function g,
    - $g^{(l)}(t) = \partial^l g(t)/\partial^l$
    - $g \in \mathcal{C}^l$ if $g^{(l)}$ exists and is continuous

M-Estimation of Linear Models with Dependent Errors

Notations for dependent errors

- Shift process $\mathcal{F}_k = (\ldots, \varepsilon_{k-1}, \varepsilon_k)$

  $\Rightarrow \ e_k = G(\mathcal{F}_k)$
- Conditional distribution function $F_i(u|\mathcal{F}_0) = P(e_i \leq u|\mathcal{F}_0)$ with conditional density $f_i(u|\mathcal{F}_0)$

  (Recall f denotes the marginal density of e_i)
- $\{\varepsilon_i^*\}$ denote an iid copy of $\{\varepsilon_i\}$
- $\mathcal{F}_k^* = (\ldots, \varepsilon_{-1}, \varepsilon_0^*, \varepsilon_1, \ldots, \varepsilon_k)$

  $\Rightarrow \ \mathcal{F}_j^* = \mathcal{F}_j$ for $j < 0$
- $e_k^* = G(\mathcal{F}_k^*)$

  $\Rightarrow \ e_k \stackrel{D}{=} e_k^*$
- Projection operators $\mathcal{P}_k V = \mathsf{E}[V \mid \mathcal{F}_k] - \mathsf{E}[V \mid \mathcal{F}_{k-1}], V \in \mathcal{L}^1$
- f is the marginal density of e_i

M-Estimation of Linear Models with Dependent Errors

More notations...

- Recall $Q_n = \boldsymbol{X}_n^\mathsf{T} \boldsymbol{X}_n = \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T}$, where $\boldsymbol{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\mathsf{T}$
- $\boldsymbol{z}_{i,n} = Q_n^{-1/2} \boldsymbol{x}_i$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_n = Q_n^{1/2} \boldsymbol{\beta}$
  $\Rightarrow\ \boldsymbol{x}_i^\mathsf{T} \boldsymbol{\beta} = \boldsymbol{z}_{i,n}^\mathsf{T} \boldsymbol{\theta}$ and $\sum_{i=1}^n \boldsymbol{z}_{i,n} \boldsymbol{z}_{i,n}^\mathsf{T} = \boldsymbol{I}_p$
- Re-parametrize the linear model by

$$y_i = \boldsymbol{x}_i^\mathsf{T} \boldsymbol{\beta} + e_i = \boldsymbol{z}_{i,n}^\mathsf{T} \boldsymbol{\theta} + e_i \tag{1}$$

- Define the k^{th}-step ahead predicted function

$$\psi_k(t; \mathcal{F}_0) = \mathsf{E}\left[\psi(e_k + t) \,|\, \mathcal{F}_0\right], \quad k \geq 0 \tag{2}$$

Standard Regularity Conditions for iid Errors

(A1) ρ is a convex function, $\mathsf{E}[\psi(e_1)] = 0$ and $\|\psi(e_1)\|^2 > 0$

- $\mathsf{Var}(e_i) = \infty$ is allowed, which is actually one of the primary reasons for robust estimation

(A2) $\varphi(t) \equiv \mathsf{E}[\psi(e_1 + t)]$ has a strict positive derivative at $t = 0$

- $\boldsymbol{\theta}$ is estimable and separable

(A3) $m(t) \equiv \|\psi(e_1 + t) - \psi(e_1)\|$ is continuous at $t = 0$

- ψ is nondecreasing and has countably many discontinuous points
- If e_i has a continuous distribution function and $\|\psi(e_1 + t_0)\| + \|\psi(e_1 - t_0)\| < \infty$ for some $t_0 > 0$, then $\lim_{t\to 0} \psi(e_1 + t) = \psi(e_1)$ almost surely and (A3) follows from the Lebeague dominated convergence theorem

(A4) $r_n \equiv \max_{i\le n} |\boldsymbol{z}_{i,n}| = \max_{i\le n}[\boldsymbol{x}_i^{\mathsf{T}} Q_n^{-1} \boldsymbol{x}_i]^{1/2} = o(1)$

- The Lindeberg-Feller type condition, the diagonal elements of the hat matrix $\boldsymbol{X}_n^{\mathsf{T}} Q_n^{-1} \boldsymbol{X}_n$ are uniformly negligible
- $Q_n^{-1} = o(1)$ and $\nu_{\min}(Q_n) \to \infty$; a classical condition for weak consistent of the least squares estimators (Eicker, 1963).
- A necessary and sufficient condition for the least squares estimator $Q_n^{-1} \boldsymbol{X}_n^{\mathsf{T}} \boldsymbol{y}_n$, where $\boldsymbol{y}_n = (y_1, \ldots, y_n)^{\mathsf{T}}$, to be asymptotically normal

Consistency and Asymptotic Normality

Theorem 1 in Wu (2007)

Assume (A1)–(A4) and, for some $\epsilon_0 > 0$,

$$\sum_{i=0}^{\infty} \sup_{|\epsilon| \le \epsilon_0} \left\| \mathsf{E}\left[\psi(e_i + \epsilon) | \mathcal{F}_0\right] - \mathsf{E}\left[\psi(e_i^* + \epsilon) \,|\, \mathcal{F}_0^*\right] \right\| < \infty \tag{3}$$

Then we have

$$\varphi'(0)\widehat{\boldsymbol{\theta}}_n - \sum_{i=1}^{n} \psi(e_i) \boldsymbol{z}_{i,n} = o_P(1) \tag{4}$$

and $\widehat{\boldsymbol{\theta}}_n = O_P(1)$. Additionally, if the limit

$$\lim_{n \to \infty} \sum_{i=1}^{n-|k|} \boldsymbol{z}_{i,n} \boldsymbol{z}_{i+k,n}^{\mathsf{T}} = \Delta_k \tag{5}$$

exists for each $k \in \mathbb{Z}$, then

$$\varphi'(0)\widehat{\boldsymbol{\theta}}_n \xrightarrow{D} \mathcal{N}(0, \Delta), \quad \text{where } \Delta = \sum_{k \in \mathbb{Z}} \mathsf{E}\left[\psi(e_0)\psi(e_k)\right] \Delta_k \tag{6}$$

Consistency and Asymptotic Normality

Remarks for Theorem 1

- Theorem 1 ensures the consistency of $\widehat{\boldsymbol{\beta}}_n$
  - $\widehat{\boldsymbol{\theta}}_n = O_P(1)$ and $Q_n^{-1} = o_P(1)$ implies $\widehat{\boldsymbol{\beta}}_n = o_P(1)$
- Discussion of condition (3)
  - The quantity

$$\left\|\psi_i(\epsilon; \mathcal{F}_0) - \psi_i(\epsilon; \mathcal{F}_0^*)\right\| = \left\|\mathsf{E}\left[\psi(e_i + \epsilon) \,|\, \mathcal{F}_0\right] - \mathsf{E}\left[\psi(e_i^* + \epsilon) \,|\, \mathcal{F}_0^*\right]\right\|$$

    measures the contribution of ε_0 in predicting $\psi(e_i + \epsilon)$
  - (3) suggests short-range dependence (SRD) in the sense that the cumulative distribution of ε_0 in predicting future values is finite.
- To prove Theorem 1, the major tool is to use martingale differences

$$J_k = \sum_{i=1}^{n} \mathcal{P}_{i-k}\psi(e_i)\boldsymbol{z}_{i,n} = \sum_{i=1}^{n} \Big(\mathsf{E}\big[\psi(e_i) \,|\, \mathcal{F}_{i-k}\big] - \mathsf{E}\big[\psi(e_i) \,|\, \mathcal{F}_{i-k-1}\big]\Big)\boldsymbol{z}_{i,n}$$

  to control the higher-order remainder terms

Bahadur Representations

Two more regularity conditions

(A5) There exists an $\epsilon_0 > 0$ such that

$$L_i \equiv \sup_{|s|,|t|\le\epsilon_0, s\neq t} \frac{\|\psi_1(s;\mathcal{F}_i) - \psi_1(t;\mathcal{F}_i)\|}{|s-t|} \in \mathcal{L}^1. \qquad (7)$$

- The function $\psi_1(s;\mathcal{F}_i), |s| \le \epsilon_0$, is stochastically Lipschitz at a neighborhood of $s = 0$, while function ψ itself does not have to be Lipschitz continuous

(A6) Let $\psi_1(\cdot;\mathcal{F}_i) \in \mathcal{C}^l, l \ge 0$. For some $\epsilon_0 > 0$, $\sup_{|\epsilon|\le\epsilon_0} \left\|\psi^{(l)}(\epsilon;\mathcal{F}_i)\right\| < \infty$ and

$$\sum_{i=0}^{\infty} \sup_{|\epsilon|\le\epsilon_0} \left\| \mathsf{E}\left[\psi^{(l)}(\epsilon;\mathcal{F}_i) \,|\, \mathcal{F}_0\right] - \mathsf{E}\left[\psi^{(l)}(\epsilon;\mathcal{F}_i^*) \,|\, \mathcal{F}_0^*\right] \right\| < \infty. \qquad (8)$$

- A generalization of (3), the SRD condition
- Sufficient conditions are provided in Proposition 2
- Satisfied by a very wide range of commonly seen causal, stationary and SRD time series

Bahadur Representations

- Recall what have to be dealt with...

  Discussion of Theorem 3.1 in He and Shao (1996)

- Define M-processes (c.f., Welsh (1989))

$$K_n(\boldsymbol{\theta}) = \Omega_n(\boldsymbol{\theta}) - \mathsf{E}[\Omega_n(\boldsymbol{\theta})] \quad \text{and} \quad \widetilde{K}_n(\boldsymbol{\beta}) = \widetilde{\Omega}_n(\boldsymbol{\beta}) - \mathsf{E}[\widetilde{\Omega}_n(\boldsymbol{\beta})],$$

  where

$$\Omega_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \psi(e_i - \boldsymbol{z}_{i,n}^{\mathsf{T}}\boldsymbol{\theta})\boldsymbol{z}_{i,n}, \quad \boldsymbol{\theta} \in \mathbb{R}^p, \quad \text{and}$$

$$\widetilde{\Omega}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \psi(e_i - \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta})\boldsymbol{x}_i, \quad \boldsymbol{\beta} \in \mathbb{R}^p \tag{9}$$

- For $q > 0$, define

$$\zeta_n(q) = \sum_{i=1}^{n} |\boldsymbol{z}_{i,n}|^q \quad \text{and} \quad \xi_n(q) = \sum_{i=1}^{n} |\boldsymbol{x}_i|^q \tag{10}$$

Bahadur Representations

Theorem 2 in Wu (2007) (Local oscillation rate of the M-process $K_n(\boldsymbol{\theta})$)

Assume (A1)–(A5) and assume (A6) holds with $l = 1, \ldots, p$. Let $\{\delta_n\}_{n\in\mathbb{N}}$ be a sequence of positive numbers such that

$$\delta_n \to \infty \quad \text{and} \quad \delta_n r_n = \delta_n \max_{i\le n} |\boldsymbol{z}_{i,n}| \to 0 \tag{11}$$

then

$$\sup_{|\boldsymbol{\theta}|\le\delta_n} |K_n(\boldsymbol{\theta}) - K_n(\mathbf{0})| = O_P\left(\sqrt{\tau_n(\delta_n)}\log n + \delta_n\sqrt{\zeta_n(4)}\right), \tag{12}$$

where

$$\tau_n(\delta) = \sum_{i=1}^{n} |\boldsymbol{z}_{i,n}|^2 \left[m^2(|\boldsymbol{z}_{i,n}|\,\delta) + m^2(-|\boldsymbol{z}_{i,n}|\,\delta)\right], \quad \delta > 0. \tag{13}$$

Note: In addition, we have $\widehat{\boldsymbol{\theta}}_n = O_P(\delta_n)$

Bahadur Representations

Corollary 1 in Wu (2007) (Weak Bahadur representation for $\widehat{\boldsymbol{\theta}}_n$)

Assume (A1)–(A5) and assume (A6) holds with $l = 1, \ldots, p$, and $\varphi(t) = t\varphi'(t) + O(t^2)$ as $t \to 0$. Further assume $\Omega_n(\widehat{\boldsymbol{\theta}}_n) = O_P(r_n)$. Then for any sequence $c_n \to \infty$,

$$\varphi'(0)\widehat{\boldsymbol{\theta}}_n - \sum_{i=1}^{n} \psi(e_i)\boldsymbol{z}_{i,n} = O_P\left(\sqrt{\tau_n(\delta_n)}\log n + \delta_n r_n\right),$$

$$\text{where } \delta_n = \min(c_n, r_n^{-1/2}). \tag{14}$$

In particular, if $m(t) = O\left(|t|^\lambda\right)$ as $t \to 0$ for some $\lambda > 0$, then

$$\varphi'(0)\widehat{\boldsymbol{\theta}}_n - \sum_{i=1}^{n} \psi(e_i)\boldsymbol{z}_{i,n} = O_P\left(\sqrt{\zeta_n(2+2\lambda)}\log n + r_n\right). \tag{15}$$

Remark: If ψ is continuous, the minimizer solves $\Omega_n(\widehat{\boldsymbol{\theta}}_n) = 0$; otherwise, the condition $\Omega_n(\widehat{\boldsymbol{\theta}}_n) = O_P(r_n)$ is needed
(e.g., in quantile regression, $|\Omega_n(\widehat{\boldsymbol{\theta}}_n)| \leq (p+1)r_n$ a.s.)

Bahadur Representations

Theorem 3 in Wu (2007) (Strong Bahadur representation for $\widehat{\boldsymbol{\beta}}_n$)

(a) Assume (A1)–(A3), (A5) and assume (A6) holds with $l = 1, \ldots, p$.

(b) Assume that

$$\liminf_{n \to \infty} \frac{\nu_{\min}(Q_n)}{n} > 0, \quad \xi_n(2) = O(n)$$

and

$$\widetilde{r}_n \equiv \max_{j \le n} |\boldsymbol{x}_j| = O\left(\frac{\sqrt{n}}{(\log n)^2}\right). \tag{16}$$

Let $b_n = n^{-1/2} (\log n)^{3/2} (\log \log n)^{1/2+\iota}, \iota > 0, \bar{n} = 2^{\lceil \log n / \log 2 \rceil}$ and $q > 3/2$.

Note: $\xi_n(2) = O(n)$ implies $\widetilde{r}_n = O(n^{1/2})$, while the above order of $\widetilde{r}_n$ is stronger for technical reason

Theorem 3 in Wu (2007) (Strong Bahadur representation for $\widehat{\boldsymbol{\beta}}_n$)

Then

(i)
$$\sup_{|\boldsymbol{\beta}|\le b_n} \left|\widetilde{K}_n(\boldsymbol{\beta}) - \widetilde{K}_n(\mathbf{0})\right| = O_{a.s.}\left(L_{\bar{n}} + B_{\bar{n}}\right), \tag{17}$$

where $B_n = b_n\sqrt{\xi_n(4)}\,(\log n)^{3/2}\,(\log\log n)^{(1+\iota)/2}$,
$L_n = \sqrt{\widetilde{\tau}_n(2b_n)}\,(\log n)^q$ and

$$\widetilde{\tau}_n(\delta) = \sum_{i=1}^{n} |\boldsymbol{x}_i|^2\left[m^2(|\boldsymbol{x}_i|\,\delta) + m^2(-|\boldsymbol{x}_i|\,\delta)\right], \quad \delta > 0. \tag{18}$$

If additionally $\varphi(t) = t\varphi'(0) + O(t^2)$ and $m(t) = O\left(\sqrt{t}\right)$ as $t \to 0$ and $\widetilde{\Omega}_n(\widehat{\boldsymbol{\beta}}_n) = O_{a.s.}\left(\widetilde{r}_n\right)$, then

(ii) $\widehat{\boldsymbol{\beta}}_n = O_{a.s.}(b_n)$ and

(iii) the strong Bahadur representation hold:

$$\varphi'(0)Q_n\widehat{\boldsymbol{\beta}}_n - \sum_{i=1}^{n}\psi(e_i)\boldsymbol{x}_i = O_{a.s.}\left(L_{\bar{n}} + B_{\bar{n}} + \xi_n(3)b_n^2 + \widetilde{r}_n\right). \tag{19}$$

Sketch of Proof of Theorem 3 (iii) in Wu (2007)

To prove (iii) with (i) and (ii), start with

$$-\mathsf{E}[\widetilde{\Omega}_n(\widehat{\boldsymbol{\beta}}_n)] = -\sum_{i=1}^{n} \varphi(-\boldsymbol{x}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}_n)\boldsymbol{x}_i = \sum_{i=1}^{n}\left[\varphi'(0)\boldsymbol{x}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}_n + O\left(|\boldsymbol{x}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}_n|^2\right)\right]\boldsymbol{x}_i$$

$$= \varphi'(0)Q_n\widehat{\boldsymbol{\beta}}_n + O\left(\xi_n(3)b_n^2\right)$$

by Taylor expansion.

Note that $\widehat{\boldsymbol{\beta}}_n = O_{a.s.}(b_n)$ and $\widetilde{\Omega}_n(\widehat{\boldsymbol{\beta}}_n) = O_{a.s.}(\widetilde{r}_n)$. For n large enough, we have

$$\begin{aligned}
&\varphi'(0)Q_n\widehat{\boldsymbol{\beta}}_n - \sum_{i=1}^{n}\psi(e_i)\boldsymbol{x}_i \\
=&-\mathsf{E}[\widetilde{\Omega}_n(\widehat{\boldsymbol{\beta}}_n)] - \widetilde{\Omega}_n(\mathbf{0}) + O\left(\xi_n(3)b_n^2\right) \\
=&\underbrace{\widetilde{\Omega}_n(\widehat{\boldsymbol{\beta}}_n) - \mathsf{E}[\widetilde{\Omega}_n(\widehat{\boldsymbol{\beta}}_n)] - \widetilde{\Omega}_n(\mathbf{0}) + \mathsf{E}[\widetilde{\Omega}_n(\mathbf{0})]}_{\widetilde{K}_n(\widehat{\boldsymbol{\beta}}_n) - \widetilde{K}_n(\mathbf{0})} + O_{a.s.}\left(\xi_n(3)b_n^2 + \widetilde{r}_n\right) \\
=&O_{a.s.}\left(L_{\bar{n}} + B_{\bar{n}} + \xi_n(3)b_n^2 + \widetilde{r}_n\right)
\end{aligned}$$

Sufficient Conditions for the SRD

Lemma 1 in Wu (2007)

Assume the process $X_k = g(\mathcal{F}_k) \in \mathcal{L}^2$. Let $g_k(\mathcal{F}_0) = \mathsf{E}[g(\mathcal{F}_k|\mathcal{F}_0)], k \geq 0$.
Then $\|\mathcal{P}_0 X_k\| \leq \|g(\mathcal{F}_k) - g(\mathcal{F}_k^*)\|$ and
$\|\mathcal{P}_0 X_k\| \leq \|g_k(\mathcal{F}_0) - g_k(\mathcal{F}_0^*)\| \leq 2\|\mathcal{P}_0 X_k\|$

Proposition 2 in Wu (2007)

Let $\psi(u;\epsilon_0) = |\psi(u+\epsilon_0)| + |\psi(u-\epsilon_0)|$. Assume that $f_1(\cdot|\mathcal{F}_i) \in \mathcal{C}^l, l \geq 0$, and

$$\sum_{i=1}^{n} \bar{\omega}_l(i) < \infty, \quad \text{where } \bar{\omega}_l(i) = \int_{\mathbb{R}} \left\| f_1^{(l)}(u|\mathcal{F}_i) - f_1^{(l)}(u|\mathcal{F}_i^*) \right\| \psi(u;\epsilon_0)\, \mathrm{d}u. \tag{20}$$

Then

$$\sum_{i=1}^{\infty} \sup_{|\epsilon|\leq\epsilon_0} \left\| \psi_1^{(l)}(u;\mathcal{F}_i) - \psi_1^{(l)}(u;\mathcal{F}_i^*) \right\| < \infty$$

and the SRD condition (8) holds.

Back to condition (A6)

Sufficient Conditions for the SRD

- The SRD condition (8) is crucial for showing the local oscillations of the M-processes
- (Propositions 3 and 4) Linear processes

$$e_i = \sum_{j=0}^{\infty} a_j \varepsilon_{i-j} \quad \text{with } a_0 = 1 \text{ and } \varepsilon_0 \in \mathcal{L}^q \tag{21}$$

  The condition

$$\sum_{j=0}^{\infty} |a_j|^{\min(2,q)} < \infty \tag{22}$$

- (Proposition 5 and 6) Nonlinear time series

$$e_i = R(r_{i-1}, \varepsilon_i), \quad \text{where } R \text{ is a measurable function} \tag{23}$$

- It is known that the M-estimates behave very differently in long-range dependent case, yet not well studied

Divide-and-Conquer for Big Data

Suppose a evenly partitioned massive data of a total size $N = Sn$ by S subsamples of sizes n

$$\boldsymbol{X}_s = \{X_{s,1}, \ldots, X_{s,n}\}, \quad s = 1, \ldots, S, \quad X_{s,i} \overset{iid}{\sim} P_{\boldsymbol{\vartheta}}, \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$$

- Assume $X_{s,i} = (y_i, \boldsymbol{x}_i)$ and $\boldsymbol{\beta} = \boldsymbol{\vartheta}$ for linear models $y_i = \boldsymbol{x}_{s,i}^\mathsf{T}\boldsymbol{\beta} + e_{s,i}$ with iid errors
- Define oracle M-estimator of $\boldsymbol{\beta}$ as

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OR}} = -(\gamma Q_{\mathrm{OR}})^{-1} \sum_{s=1}^{S} \sum_{i=1}^{n} \psi(e_{s,i})\boldsymbol{x}_{s,i} + R_{\mathrm{OR}},$$

  where $Q_{\mathrm{OR}} = \sum_{s=1}^{S} \sum_{i=1}^{n} \boldsymbol{x}_{s,i}\boldsymbol{x}_{s,i}^\mathsf{T}$
- Define subsample M-estimator of $\boldsymbol{\beta}$ as

$$\widehat{\boldsymbol{\beta}}_s = -(\gamma Q_s)^{-1} \sum_{s=1}^{S} \sum_{i=1}^{n} \psi(e_{s,i})\boldsymbol{x}_{s,i} + R_s,$$

  where $Q_s = \sum_{i=1}^{n} \boldsymbol{x}_{s,i}\boldsymbol{x}_{s,i}^\mathsf{T}$

Divide-and-Conquer for Big Data

Aggregate data estimator

$$\widehat{\boldsymbol{\beta}}_{\mathrm{AD}} = \frac{1}{S}\sum_{s=1}^{S}\widehat{\boldsymbol{\beta}}_s$$

- The quality of $\widehat{\boldsymbol{\beta}}_{\mathrm{AD}}$ can be assessed by the order of

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OR}} - \widehat{\boldsymbol{\beta}}_{\mathrm{AD}} = \gamma\sum_{s=1}^{S}\left[\frac{1}{S}Q_s^{-1} - Q_{\mathrm{OR}}^{-1}\right]\sum_{i=1}^{n}\psi(e_{s,i})\boldsymbol{x}_{s,i} + R_{\mathrm{OR}} - \frac{1}{S}\sum_{s=1}^{S}R_s$$

  so that an upper bound of S could be determined
- The exact forms of the remainder terms need to be known...

▸ Proof of Theorem 3 in He and Shao (1996)

Multivariate confidence distribution inference

- The form is similar to the AD estimator
  (Yet could be too involved to describe here)

References

Eicker, F. (1963) Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Annals of Mathematical Statistics*, **34**, 447–456. URL http://dx.doi.org/10.1214/aoms/1177704156.

He, X. and Shao, Q.-M. (1996) A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *The Annals of Statistics*, **24**, 2608–2630. URL http://dx.doi.org/10.1214/aos/1032181172.

van der Vaart, A. (2000) *Asymptotic Statistics*.

Welsh, A. H. (1989) On M-processes and M-estimation. *The Annals of Statistics*, **17**, 337–361. URL http://dx.doi.org/10.1214/aos/1176347021.

Wu, W. B. (2007) M-estimation of linear models with dependent errors. *The Annals of Statistics*, **35**, 495–521. URL http://dx.doi.org/10.1214/009053606000001406.