Maximum Likelihood, EM Algorithm and Bootstrap

CUHK Structural Estimation Workshop

Zhongji Wei

September 25, 2023

Thank’s a lot to Chengwang Liao

- MLE
- EM Algorithm
- Bootstrap

- A random sample of size n: $\mathbf{Z}^n = (\mathbf{Z}'_1, \ldots, \mathbf{Z}'_n)'$
  - Or you can say, (k-dimensional) random vector $\mathbf{Z}$, (independently) sample n times
  - Realization: $\mathbf{z}^n = (\mathbf{z}'_1, \ldots, \mathbf{z}'_n)'$
  - Notations: capital letter: random variable, small letter: realization; superscript: history, or product of event until that period, subscript: event at that period
- Joint pdf (not necessarily independent)

$$
\begin{aligned}
f_{\mathbf{Z}^n}(\mathbf{z}^n) &= f_{\mathbf{Z}_n|\mathbf{z}^{n-1}} f_{\mathbf{Z}^{n-1}}(\mathbf{z}^{n-1}) \\
&= \prod_{t=1}^{n} f_{\mathbf{Z}_t|\mathbf{Z}^{t-1}}(\mathbf{z}_t|\mathbf{z}^{t-1}) \qquad (1)
\end{aligned}
$$

- For $\mathbf{Z}_t = (Y_t, \mathbf{X}'_t)'$, let $\Psi_t = (x_t, \mathbf{z}^{t-1})$:
  - For example, Y_t is GDP growth, x_t are other stock variables today (capital stock, labor, policy, etc), Ψ_t is today's variable and variables yesterday

$$
f_{\mathbf{Z}_t|\mathbf{Z}^{t-1}}(\mathbf{z}_t, \mathbf{z}^{t-1}) = f_{Y_t|\Psi_t}(y_t|\psi_t) f_{X_t|Z^{t-1}}(\mathbf{x}_t|\mathbf{z}^{t-1}) \qquad (2)
$$

$$
f_{\mathbf{Z}^n}(\mathbf{z}^n) = \prod_{t=1}^{n} f_{Y_t|\psi_t}(y_t|\psi_t) f_{\mathbf{X}_t|Z^{t-1}}(x_t|z^{t-1}) \qquad (3)
$$

- Likelihood function
  - Given observation $\mathbf{x}^n$, the joint pdf of random sample $\mathbf{X}^n$ as a **function** of parameter(vector) θ

$$\mathcal{L}(\theta|\mathbf{x}^n) = f_{\mathbf{X}^n}(\mathbf{x}^n|\theta) \tag{4}$$

- MLE (Maximum likelihood estimator) of θ

$$\hat{\theta}_n(\mathbf{X}^n) = arg\max_{\theta\in\Theta} \mathcal{L}(\theta|\mathbf{X}^n) \tag{5}$$

  - Θ: parameter space
  - For each sample set $\mathbf{X}^n$, $\mathcal{L}(\theta|\mathbf{X}^n)$ attains its maximum at $\theta = \hat{\theta}(\mathbf{X}^n)$

- What we care about: how Y_t is related to ψ_t, characterized by a parameter, say β
  - How yesterday and today's capital stock, labor, policy, as well as GDP yesterday, affect today's GDP
  - Not how they affect yesterday's capital stock, etc.
- Variation-free parameters assumption
  - Parameters are independent on of each other in the sense that their relationship with r.v.s

$$f_{\mathbf{Z}^n|\mathbf{Z}^{t-1}}(z_t|z^{t-1},\beta,\gamma) = \underbrace{f_{Y_t|\Psi_t}(y_t|\psi_t,\beta)}_{\text{What we're interested in}} \; f_{\mathbf{X_t}|\mathbf{Z}^{t-1}}(x_t|z^{t-1},\gamma) \tag{6}$$

- Under this assumption, conditional MLE and global MLE give the same estimator of β
  - Conditional likelihood: $\prod_{t=1}^{n} f_{Y_t|\Psi_t}(y_t|\psi_t,\beta) f_{\mathbf{X_t}|\mathbf{Z}^{t-1}}(x_t|z^{t-1},\gamma)$
  - Global likelihood: $f_{\mathbf{Z}^n}(\mathbf{z}^n|\beta,\gamma)$

- Extremum Estimator Lemma
- Suppose
  - $Q(\theta)$ is a nonstochastic (the functional form is unchanged) real-valued function, continuous in $\theta \in \Theta$, Θ is compact, $\theta_0 \in \Theta$ is the unique maximizer of $Q(\theta)$ in Θ
  - $Q_n(\theta)$ is a sequence of random functions continuous in $\theta \in \Theta$ with prob 1 (randomness comes from random variable $\mathbf{Z}^n$)
  - $\lim_{n\to\infty} \sup_{\theta\in\Theta} |Q_n(\theta) - Q(\theta)| = 0$, a.s.
- Then $\hat{\theta}_n = arg \max_{\theta\in\Theta} Q_n(\theta)$ exists and $\hat{\theta}_n \to \theta_0$, a.s.
  - Note that we already supposed θ_0 is the unique maximizer

- Uniform Law of Large Numbers
- $\mathbf{X}^n = (\mathbf{X}_1, ..., \mathbf{X}_n)$ is an $i.i.d$ random sample, if
  - Θ is compact
  - $f(\mathbf{x}, \theta)$ is continuous at each $\theta \in \Theta$ for almost all $\mathbf{x}$
  - There exists a dominating function $d(\mathbf{x})$, s.t.
    (1) $\mathbf{E}[d(\mathbf{x})] < \infty$ and (2) $|f(\mathbf{x}|\theta) \leq d(\mathbf{x}), \forall\theta \in \Theta$
- Then $\mathbf{E}[f(\mathbf{x}|\theta)]$ is continuous in θ, and

$$\sup_{\theta\in\Theta} || \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{X}_i, \theta) - \mathbf{E}[f(\mathbf{X}|\theta)] || \to 0, a.s. \tag{7}$$

MLE Assumptions

- $\mathbf{X}^n$ is an $i.i.d$ random sample from some population distribution (or the "true" distribution) $F(\mathbf{X})$ (**Strong!** problematic when there's serial dependency)
-
  - For each $\theta \in \Theta$, $f(\mathbf{X}|\theta)$ is a pdf, $f(\mathbf{x}|\theta) > 0, \forall \mathbf{x}$
  - $\exists \theta_0 \in \Theta^\circ$, s.t. $f(\mathbf{x}, \theta_0)$ is exactly the population distribution (for interior FOC)
  - θ_0 is the unique maximizer of $\max_{\theta \in \Theta} \mathbf{E}[\ \log f(\mathbf{X}|\theta)\ ]$, where $\mathbf{E}[.]$ is taken on the population distribution
  - Function $\log f(\mathbf{x}|\theta)$ is continuous in $(\mathbf{x}, \theta)$ and its absolute value is bounded by a nonnegative function $b(x)$, s.t. $\mathbf{E}[b(\mathbf{X})] > 0$
- Θ is compact set
- Something else

Under these assumptions,

$$\hat{\theta}_n \xrightarrow{a.s.} \theta_0 \tag{8}$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1}) \tag{9}$$

where $I(\theta) = -\mathbf{E}_\theta[\frac{\partial^2}{\partial\theta\partial\theta'} \log f(\mathbf{X}|\theta)]$

- Let $\mathbf{X}^n$ be a random sample with joint pdf $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta)$, let $W_n = W(\mathbf{X}^n)$ be any estimator of $\tau(\theta)$ (a function of θ, estimated given $\mathbf{x}^n$), s.t.

$$\frac{d}{d\theta}\mathbf{E}_\theta(W_n) = \int W_n \frac{\partial}{\partial\theta} f_{\mathbf{X}^n}(\mathbf{x}^n|\theta) d\mathbf{x} \tag{10}$$

  and $var_\theta(W_n) < \infty$, then

$$Var_\theta(W_n) \geq \frac{[\frac{d}{d\theta}\mathbf{E}_\theta[W_n]]^2}{\mathbf{E}_\theta[(\frac{\partial}{\partial\theta}\log f_{\mathbf{X}^n}(\mathbf{x}^n|\theta))^2]} \tag{11}$$

MLE is consistent, and asympototically more efficient than any unbiased estimators

Quasi MLE (QMLE)

- $\mathbf{X}^n$ is an $i.i.d$ random sample from some population distribution $G(x)$
- θ^* is the unique maximizer of $\max_{\theta\in\Theta} \mathbf{E} \log f(\mathbf{X}|\theta)$
- Others are the same

Note: It's the second that differ, relax to unique maximizer
Then

$$\hat{\theta}_n \xrightarrow{p} \theta^* \tag{12}$$

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} N(0, H(\theta^*)^{-1} I(\theta^*) H(\theta^*)^{-1}) \tag{13}$$

- Example: Logit and probit (especially for binary choice problem)
  - Conditional probability of choosing Y

$$P(Y = 1|\mathbf{X}_t) = \psi(\mathbf{X}_t'\beta) \tag{14}$$

where $\psi(.)$ is logistic function

$$\psi(x) = \frac{1}{1 + e^{-x}}, x \in (0, \infty) \tag{15}$$

- log-likelihood function

$$\begin{aligned} \log \mathcal{L}(\beta|Y_t, \mathbf{X}_t) &= \sum_{t=1}^{n} \log f_{Y_t|\mathbf{X}_t}(y_t|\mathbf{x}_t, \beta) \\ &= \sum_{t=1}^{n} [y_t \log \psi(\mathbf{X}_t'\beta) + (1 - y_t) \log(1 - \psi(\mathbf{X}_t'\beta))] \end{aligned} \tag{16}$$

- A more complex example in application: Bonhomme et.al (2019) See Github file

- MLE
- EM Algorithm
- Bootstrap

- $\mathbf{X}$: observed data
- $\mathbf{Z}$: missing or latent data
- Goal: find $arg\ \max_\theta \mathcal{L}(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$
  - Problem: $\mathcal{L}(\theta|\mathbf{x}) = \int f(\mathbf{x}, \mathbf{z}|\theta) d\mathbf{z}$, hard to integrate
  - Alternative: find $arg\ \max_\theta \mathcal{L}(\theta|\mathbf{x}, \mathbf{z}) = f(\mathbf{x}, \mathbf{z}|\theta)$
  - Can assume there exists a latent variable to simplify, if hard to maximize the original likelihood
- Once you know $f(\mathbf{x}, \mathbf{z}|\theta)$, you can always write down $\mathcal{L}(\theta|\mathbf{x}, \mathbf{z})$
  - But $arg\ \max_\theta \mathcal{L}(\theta|\mathbf{x}, \mathbf{z}) = f(\mathbf{x}, \mathbf{z}|\theta)$ may not have a close form solution like OLS!
  - That's why you need numerical solution, EM algorithm!

- EM algorithm: get maxima of $\mathcal{L}(\theta|\mathbf{x})$ by iteration
- E-step
  - Suppose after n iterations, the estimator of θ is $\theta^{(n)}$
  - Given $\theta^{(n)}$, calculate conditional expectation

$$g^{(n)}(\theta) \equiv \mathbf{E}_{\{\mathbf{Z}|\mathbf{X}=\mathbf{x},\theta^{(n)}\}}[\log \mathcal{L}(\theta|\mathbf{X},\mathbf{Z})] = \int_{\mathbf{z}} [\log f(\mathbf{x},\mathbf{z}|\theta^{(n)}] f(\mathbf{z}|\mathbf{x},\theta^{(n)}) d\mathbf{z} \tag{17}$$

  - In practice, may change integral to sum
- M-step: find

$$\theta^{(n+1)} = arg \max_{\theta} \mathbf{E}_{\{\mathbf{Z}|\mathbf{X}=\mathbf{x},\theta^{(n)}\}}[\log \mathcal{L}(\theta|\mathbf{X},\mathbf{Z})] \tag{18}$$

- Cannot guarantee converging to global maximum for any problem, unless convex optimization problem

- MLE
- EM Algorithm
- Bootstrap

- Sample $\{z_i, i = 1, ..., n\}$ from distribution F_0
- Statistic of interest: $T(z)$
  - Want to know its distribution!!
- CDF of $T(z)$: $G_n(t, F_0) = P(T(z) \leq (t))$
  - G_n may be complicated, can be calculated from F_0
- F_0 may be unknown!!
- Two ways to do
  - Asymptotics
  - Simulation

- Asymptotics
- Create an asymptotic approximation to the G_n: letting $n \to \infty$, use CLT and $\delta-$method
- For example, z_i has unknown distribution, $\mathbf{E}[z_i] = \mu$, $\mathbf{Var}(z_i) = \sigma^2$
- t-stat: $T(z) = \frac{\sqrt{n}}{\hat{\sigma}}(\frac{1}{n}\sum z_i - \mu_0)$
- $T(z) \Longrightarrow N(0,1)$
- Then can do inference

- Instead of using asymptotic distribution to approximate $G_n(., F_0)$, use $G_n(., F_0) \approx G_n(., \hat{F}_n)$
  - $\hat{F}_n(t) - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{x \le t\}}$, empirical CDF

- See Mikusheva’s note
- Mikusheva’s note

Maximum Likelihood, EM Algorithm and Bootstrap

Thank you for listening!

Any questions?