Lecture 5: Local Asymptotic Normality (LAN) and asymptotic theorems

Lecturer: Yanjun Han

April 12, 2021

Announcements

Project proposal due April 18

- submit via gradescope
- only if you decide to do an original project
- all other students must do a literature review, with paper chosen by the end of week 6

Today's plan

Asymptotic theorems:

- Fisher information, Fisher's program and Hodges' estimator
- three asymptotic theorems
- Gaussian location model, Anderson's lemma
- reduction of LAN models to a Gaussian location model
- examples of asymptotic lower bounds

Score function and Fisher information

Definition

A statistical model $(\mathcal{X}, (P_\theta)_{\theta \in \Theta \subseteq \mathbb{R}^d})$ is quadratic mean differentiable (QMD) at $\theta \in \Theta$ if there exists a score function $\dot{\ell}_\theta : \mathcal{X} \to \mathbb{R}^d$ such that

$$\int_{\mathcal{X}} \left(\sqrt{dP_{\theta+h}} - \sqrt{dP_\theta} - \frac{1}{2} h^\top \dot{\ell}_\theta \sqrt{dP_\theta} \right)^2 = o(\|h\|^2).$$

In this case, $I(\theta) = \mathbb{E}_{P_\theta}[\dot{\ell}_\theta \dot{\ell}_\theta^\top]$ exists and is the Fisher information at θ.

Alternative definitions:

$$\log \frac{p_{\theta+h}}{p_\theta}(x) = h^T \dot{\ell}_\theta(x) - \frac{1}{2} h^T I(\theta) h + o_{P_\theta}(\|h\|^2)$$

$$\ell_\theta(x) = \log p_\theta(x)$$

$$\dot{\ell}_\theta(x) = \frac{d}{d\theta}\left[\log p_\theta(x)\right] = \frac{\dot{p}_\theta(x)}{p_\theta(x)}$$

$$I(\theta) = \mathbb{E}_\theta\left[-\ddot{\ell}_\theta(x)\right] = \mathbb{E}_\theta\left[\dot{\ell}_\theta(x)\, \dot{\ell}_\theta(x)^T\right]$$

Cramér-Rao lower bound

Cramér-Rao lower bound

For any unbiased estimator T, i.e. $\mathbb{E}_\theta[T] = \theta$ for every $\theta \in \Theta$, it holds that

$$\underbrace{\mathrm{Cov}_\theta(T)}_{\text{Covariance matrix}} \succeq I(\theta)^{-1}, \quad \forall \theta \in \Theta.$$

$A \succeq B \Rightarrow A - B$ is positive semi-definite.

Proof via χ^2-divergence:

$$\chi^2(P, Q) = \sup_h \frac{(\mathbb{E}_P[h] - \mathbb{E}_Q[h])^2}{\mathrm{Var}_Q(h)}$$

$h(x) = v^T T(x)$

$P = P_{\theta+h}$

$Q = P_\theta$

$$\mathrm{Var}_{P_\theta}(v^T T(x)) \geq \frac{(v^T(\theta+h) - v^T\theta)^2}{\chi^2(P_{\theta+h}, P_\theta)}$$

$$= \frac{(v^T h)^2}{\underbrace{\chi^2(P_{\theta+h}, P_\theta)}} \approx h^T I(\theta) h + o(\|h\|^2)$$

$$\mathrm{RHS} \approx \frac{(v^T h)^2}{h^T I(\theta) h} \quad \max_h = v^T I(\theta)^{-1} v$$

Fisher's program

Setting throughout this lecture:

- $(P_\theta)_{\theta\in\Theta\subseteq\mathbb{R}^d}$: QMD statistical model with Fisher information $I(\theta)$
- T_n: estimator sequence under product model $(P_\theta^{\otimes n})$
- $\psi(\theta)$: a generic real-valued differentiable function of θ

Fisher's program:

- for any asymptotically normal estimators T_n with

$$\sqrt{n}(T_n-\theta)\xrightarrow{d}\mathcal{N}(0,\Sigma_\theta)$$

  for every $\theta\in\Theta$, then $\Sigma_\theta\succeq I(\theta)^{-1}$; ✗
- the MLE satisfies $\Sigma_\theta = I(\theta)^{-1}$ for every θ. ✓

Counterexample: Hodges' estimator

- Gaussian location model: $X_1, \cdots, X_n \sim \mathcal{N}(\theta, 1)$, with $\theta \in \mathbb{R}$
- Hodges' estimator:

$$\widehat{\theta}_n = \begin{cases} \bar{X} & \text{if } |\bar{X}| \geq n^{-1/4}, \\ 0 & \text{if } |\bar{X}| < n^{-1/4}. \end{cases}$$

- asymptotic normality:

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \begin{cases} 0 & \text{if } \theta = 0, \\ \mathcal{N}(0, 1) & \text{if } \theta \neq 0. \end{cases}$$

if $\theta = 0$:
$|\bar{X}| \leq \tilde{O}(\frac{1}{\sqrt{n}})$ w.h.p.
$\Rightarrow \hat{\theta}_n = 0$ w.h.p.
if $\theta \neq 0$:
$|\bar{X} - \theta| \leq \tilde{O}(\frac{1}{\sqrt{n}})$ w.h.p.
$\Rightarrow |\hat{\theta}_n| > n^{-1/4}$ for large n
$\Rightarrow \hat{\theta}_n = \bar{X}$

- "superefficiency" $I(\theta) = 1 \quad \forall \theta$

First fix: almost everywhere convolution theorem

First fix: show that Fisher's program is true almost everywhere

Almost everywhere convolution theorem

If $\sqrt{n}(T_n - \psi(\theta))$ converges in distribution to some probability measure L_θ for every θ, then there exists some probability measure M_θ such that

$$L_\theta = \mathcal{N}(0, \dot{\psi}(\theta)^\top I(\theta)^{-1} \dot{\psi}(\theta)) * M_\theta$$

for Lebesgue almost every θ, where $*$ denotes convolution.

$P * Q(x) = \int p(y) q(x-y)\, dy.$

meaning: $X \sim P, \quad Y \sim Q$

$X + Y \sim P * Q$

Second fix: convolution theorem

Second fix: restrict to a family of regular estimators

Convolution theorem

If $\sqrt{n}(T_n - \psi(\theta))$ converges in distribution to some probability measure L_θ for every θ, and T_n is regular in the sense that

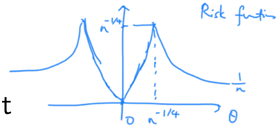
$$\sqrt{n}\left(T_n - \psi\left(\theta + \frac{h}{\sqrt{n}}\right)\right) \xrightarrow{d} L_\theta$$

for any $h \in \mathbb{R}^d$ under $P_{\theta+h/\sqrt{n}}^{\otimes n}$, then there exists some probability measure M_θ such that

$$L_\theta = \mathcal{N}(0, \dot{\psi}(\theta)^\top I(\theta)^{-1}\dot{\psi}(\theta)) * M_\theta$$

for every θ, where $*$ denotes convolution.

Third fix: local asymptotic minimax theorem



Third fix: local minimax instead of a single point

Local asymptotic minimax theorem

For every bowl-shaped loss function $\ell(\cdot)$, i.e. ℓ is symmetric and quasi-convex, then for any sequence of estimators (T_n),

$\ell(x) = \ell(-x)$

$\{x : \ell(x) \le t\}$ convex

$$\lim_{c\to\infty} \liminf_{n\to\infty} \sup_{\|h\|\le c} \mathbb{E}_{\theta+h/\sqrt{n}}\left[\ell\left(\sqrt{n}\left(T_n - \psi\left(\theta + \frac{h}{\sqrt{n}}\right)\right)\right)\right] \ge \mathbb{E}[\ell(Z)],$$

1D: $[\theta - \frac{c}{\sqrt{n}}, \theta + \frac{c}{\sqrt{n}}]$

with $Z \sim \mathcal{N}(0, \dot{\psi}_\theta^\top I(\theta)^{-1}\dot{\psi}_\theta)$.

$= \nabla\psi(\theta)$

Gaussian location model

- model: $X \sim \mathcal{N}(\theta, \Sigma)$
- unknown mean $\theta \in \mathbb{R}^d$
- known covariance Σ

Theorem (Anderson)

For any bowl-shaped loss function ℓ, it holds that

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}_\theta[\ell(\widehat{\theta} - \theta)] = \mathbb{E}[\ell(Z)]$$

$\widehat{\theta} = X$ achieves the minimax risk.

with $Z \sim \mathcal{N}(0, \Sigma)$.

Proof of theorem

- consider a Gaussian prior $\theta \sim \mathcal{N}(0, \Sigma_0)$
- then

$$\theta \mid X = x \sim \mathcal{N}\left((\Sigma_0^{-1} + \Sigma^{-1})^{-1}\Sigma^{-1}x, (\Sigma_0^{-1} + \Sigma^{-1})^{-1}\right)$$

- Bayes estimator:

$\hat{\theta}(x) =$

$$\widehat{\theta}(x) = \arg\min_{z \in \mathbb{R}^d} \mathbb{E}_{p(\theta|X=x)}[\ell(z - \theta)]$$

- use Anderson's lemma and choose $\Sigma_0 = \lambda I$ with $\lambda \to \infty$ $\hat{\theta}(x) \to x$

Anderson's lemma

Let $Z \sim \mathcal{N}(0, \Sigma)$ and ℓ be bowl-shaped. Then

$$\min_{x \in \mathbb{R}^d} \mathbb{E}[\ell(Z + x)] = \mathbb{E}[\ell(Z)].$$

Proof of Anderson's lemma

Theorem (Prépoka–Leindler)

Let $\lambda \in (0,1)$, and f, g, h be three non-negative real-valued functions on $\mathbb{R}^d$. If

$$h(\lambda x + (1-\lambda)y) \geq f(x)^\lambda g(y)^{1-\lambda}, \quad \forall x, y \in \mathbb{R}^d,$$

Brunn–Minkowski: $h(x) = 1_{\lambda K + (1-\lambda)L}(x)$, $f(x) = 1_K(x)$, $g(x) = 1_L(x)$

then

$$\int_{\mathbb{R}^d} h(x)dx \geq \left(\int_{\mathbb{R}^d} f(x)dx\right)^\lambda \left(\int_{\mathbb{R}^d} g(x)dx\right)^{1-\lambda}.$$

- let K be any symmetric convex set, and ϕ be the pdf of $\mathcal{N}(0, \Sigma)$

  $K = -K$
- apply above theorem to

  $\forall x \in \mathbb{R}^d$

  1) ϕ is log-concave
  2) $K \subseteq \frac{K-x}{2} + \frac{K+x}{2}$

$$f(z) = \phi(\cancel{x}\, z) \cdot \mathbb{1}_{K-x}(z), \quad g(z) = \phi(z) \cdot \mathbb{1}_{K+x}(z), \quad h(z) = \phi(z) \cdot \mathbb{1}_K(z)$$

$\mathbb{P}(Z \in K+x)$ (K symmetric, Z symmetric)

$\shortparallel$

- consequently, $\mathbb{P}(Z \in K - x) \leq \mathbb{P}(Z \in K)$ for any $x \in \mathbb{R}^d$
- final step: $\mathbb{E}[\ell(Z+x)] = \int_0^\infty (1 - \mathbb{P}(Z + x \in \{z : \ell(z) \leq t\}))dt$

  $= \int_0^\infty \mathbb{P}(\ell(Z+x) \geq t)\, dt$ maximized at $x=0$

Local asymptotic normality (LAN)

Definition (Local asymptotic normality)

A sequence of models $(\mathcal{X}_n, (P_{n,\theta})_{\theta \in \Theta_n})$ with $\Theta_n \uparrow \mathbb{R}^d$ is called locally asymptotically normal if

$$\log \frac{dP_{n,h}}{dP_{n,0}} = h^\top Z_n - \frac{1}{2} h^\top \underset{\substack{\uparrow \\ \text{Fisher information}}}{I_0} h + o_{P_{n,0}}(1),$$

with central sequence $Z_n \xrightarrow{d} \mathcal{N}(0, I_0)$ under $P_{n,0}$.

Comparison with the Gaussian location model: $\{N(h, I_0^{-1})\}_{h \in \mathbb{R}^d}$

$$\log \frac{f_{(h,I_0^{-1})}(x)}{f_{(0,I_0^{-1})}(x)} = h^\top I_0 x - \frac{1}{2} h^\top I_0 h$$

where $I_0 x \sim \mathcal{N}(0, I_0)$ under $x \sim \mathcal{N}(0, I_0^{-1})$.

Likelihood ratio criterion

Theorem

Let $\mathcal{M}_n = (\mathcal{X}_n, \{P_{n,0}, \cdots, P_{n,m}\})$ and $\mathcal{M} = (\mathcal{X}, \{P_0, \cdots, P_m\})$ be finite statistical models. Define the likelihood ratios

$$L_{n,i}(x_n) = \frac{dP_{n,i}}{dP_{n,0}}(x_n), \quad L_i(x) = \frac{dP_i}{dP_0}(x), \quad i \in [m].$$

$P_i \ll P_j$, $P_i \gg P_j$

If $\mathcal{M}$ is further homogeneous, and $(L_{n,1}, \cdots, L_{n,m})$ under $x_n \sim P_{n,0}$ converges in distribution to $(L_1, \cdots, L_m)$ under $x \sim P_0$, then

$$\lim_{n\to\infty} \Delta(\mathcal{M}_n, \mathcal{M}) = 0.$$

weak convergence of likelihood ratios $\Longrightarrow$ convergence of statistical models

High-level idea: standard model

sufficiency

$$\underset{\text{finite model}}{(\mathcal{X}, \{P_1, \cdots, P_m\})} \Longleftrightarrow \underset{\text{standard model}}{(\mathcal{S}_m, \{Q_1, \cdots, Q_m\})}$$

- $\mathcal{S}_m = \{(t_1, \cdots, t_m) \in \mathbb{R}_+^m : \sum_{i=1}^m t_i = m\}$
- $Q_i(dt) = t_i \mu(dt)$

with the correspondence

- $(t_1, \cdots, t_m) = (\frac{dP_1}{d\bar{P}}(x), \cdots, \frac{dP_m}{d\bar{P}}(x))$
- μ is the distribution of $(\frac{dP_1}{d\bar{P}}(x), \cdots, \frac{dP_m}{d\bar{P}}(x))$ under $x \sim \bar{P}$

Local QMD model is LAN

Theorem

Let $(\mathcal{X}, (P_\theta)_{\theta\in\Theta\subseteq\mathbb{R}^d})$ be QMD with Fisher information $I(\theta)$. Then the localized model $(\mathcal{X}^n, (P^{\otimes n}_{\theta+h/\sqrt{n}})_{h\in\Theta_n})$ with

$$\Theta_n = \left\{h \in \mathbb{R}^d : \theta + \frac{h}{\sqrt{n}} \in \Theta\right\} \uparrow \mathbb{R}^d$$

satisfies the LAN condition with Fisher information $I(\theta)$.

$$\log \frac{P_{\theta+h}}{P_\theta}(x) = h^T \dot{\ell}_\theta(x) - \frac{1}{2} h^T I(\theta) h + o_{P_\theta}(\|h\|^2)$$

$$\Rightarrow \log \frac{P^{\otimes n}_{\theta+h/\sqrt{n}}}{P^{\otimes n}_\theta}(x_1,\dots,x_n) = h^T \underbrace{\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_\theta(x_i)\right)}_{\substack{\downarrow \text{CLT} \\ \xrightarrow{d} N(0, I(\theta))}} - \frac{1}{2} h^T I(\theta) h + o_{P_\theta}(1)$$

LAN condition ✓

Corollary: the above model converges to a Gaussian location model under Le Cam's distance.

Proof of local asymptotic minimax (LAM) theorem

- product model aims to estimate $\psi(\theta + h/\sqrt{n})$ under $(P^{\otimes n}_{\theta+h/\sqrt{n}})_{h\in\mathbb{R}^d}$
- limiting Gaussian model aims to estimate

$$\psi\left(\theta + \frac{h}{\sqrt{n}}\right) = \psi(\theta) + \frac{1}{\sqrt{n}}\dot{\psi}_\theta^\top h + o(n^{-1/2})$$

  under $X \sim \mathcal{N}(h, I(\theta)^{-1})$
- Anderson's lemma gives LAM

Diagram

$$
\begin{tikzcd}
\text{local QMD model} \arrow[d, "\text{local expansion of likelihood ratio}"] \\
\text{LAN condition} \arrow[d, "\text{weak convergence of likelihood ratio}"] \\
\text{Gaussian location model} \arrow[d, "\text{Anderson's lemma}"] \\
\text{local asymptotic minimax theorem}
\end{tikzcd}
$$

Example I: bias of the coin

- Problem: $X_1, \cdots, X_n \sim \text{Bern}(p)$, $p \in [0,1]$, $L(p,T) = (T - \sqrt{p})^2$
- correspondence:

$$I(p) = \frac{1}{p(1-p)}, \quad \psi(p) = \sqrt{p}, \quad \ell(z) = z^2$$

- LAM: for every $p_0 \in [0,1]$,

$$\lim_{h\to\infty} \liminf_{n\to\infty} \sup_{|p-p_0|\le h/\sqrt{n}} n \cdot \mathbb{E}_p(T_n - \sqrt{p})^2 \ge \frac{\overset{1/I(\psi)}{p(1-p)}}{\underset{(\psi'(p_0))^2}{(2\sqrt{p})^2}} = \frac{1-p}{4}$$

- translate into a global minimax lower bound:

$$\inf_{T_n} \sup_{p\in[0,1]} \mathbb{E}_p(T_n - \sqrt{p})^2 \ge \frac{1+o_n(1)}{4n}$$

Example II: entropy estimation

- (differential) entropy:

$$h(f) = \int_{\mathbb{R}^d} -f(x) \log f(x) dx$$

- Problem: $X_1, \cdots, X_n \sim f$, $L(f, T) = (T - h(f))^2$
- one-dimensional submodel: restrict to $f = f_0 + tg$, with $t \in [-\varepsilon, \varepsilon]$
- constraint on g: $\int_{\mathbb{R}^d} g(x) dx = 0$
- correspondence:

$$I(0) = \int_{\mathbb{R}^d} \frac{g(x)^2}{f_0(x)} dx, \quad \psi'(0) = \int_{\mathbb{R}^d} g(x)(1 - \log f_0(x)) dx$$

- LAM:

$$\inf_{\widehat{T}} \sup_{f} \mathbb{E}_f (T - h(f))^2 \geq \frac{1 + o_n(1)}{n} \sup_{g} \frac{\psi'(0)^2}{I(0)}$$

Least favorable one-dimensional submodel

- choose g to maximize

$$V(g) \triangleq \left(\int_{\mathbb{R}^d} \frac{g(x)^2}{f_0(x)} dx \right)^{-1} \left(\int_{\mathbb{R}^d} g(x)(1 - \log f_0(x)) dx \right)^2$$

- by Cauchy–Schwartz:

$$\begin{aligned} V(g) &= \inf_c \left(\int_{\mathbb{R}^d} \frac{g(x)^2}{f_0(x)} dx \right)^{-1} \left(\int_{\mathbb{R}^d} g(x)(c - \log f_0(x)) dx \right)^2 \\ &\le \inf_c \int_{\mathbb{R}^d} f_0(x)(c - \log f_0(x))^2 dx \\ &= \int_{\mathbb{R}^d} f_0(x) \log^2 f_0(x) dx - h(f_0)^2 \end{aligned}$$

$c = -h(f_0)$

Var-entropy

- equality attained at $g(x) = f_0(x)(-\log f_0(x) - h(f_0))$

Limitations of classical asymptotics

- asymptotic vs. non-asymptotic
- parametric vs. non-parametric
- local vs. global

References

- Lucien M. Le Cam, "Limits of experiments." *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1: Theory of Statistics. The Regents of the University of California, 1972.
- Lucien M. Le Cam. "Asymptotic methods in statistical theory." Springer, New York, 1986.
- Aad W. Van der Vaart, "Asymptotic statistics." Vol. 3. Cambridge university press, 2000.
- Thomas B. Berrett, Richard J. Samworth, and Ming Yuan, "Efficient multivariate entropy estimation via k-nearest neighbour distances." *The Annals of Statistics* 47.1 (2019): 288-318.

Next lecture: statistical-computational tradeoff (Jay)