STA732
Statistical Inference

Lecture 15: Asymptotic Analysis of MLE

Yuansi Chen

Spring 2023

Duke University

https://www2.stat.duke.edu/courses/Spring23/sta732.01/

Recap from Lecture 14

- Convergence in probability
- Convergence in distribution
- Calculus for functions of converging random variables: continuous mapping theorem, Slutsky's theorem, Delta method

Goal of Lecture 15

1. Maximum likelihood estimator (MLE)
2. Asymptotic efficiency
3. Intuition for the asymptotic distribution of MLE
4. Consistency and Asymptotic normality of MLE

Chap. 8.3, 8.5, 9.2, 9.3 of Keener or Chap. 6.2-6.4 of Lehmann and Casella

Maximum likelihood estimator (MLE)

Def. maximum likelihood estimator

For a dominated family $\mathscr{P} = \{P_\theta, \theta \in \Omega\}$ with densities p_θ, data X drawn from p_θ, the value that maximizes the likelihood function is called the maximum likelihood estimator of θ

$$\begin{aligned}\hat{\theta}_{\mathrm{MLE}}(X) &= \arg\max_{\theta\in\Omega} p_\theta(X) \\ &= \arg\max_{\theta\in\Omega} \ell(\theta; X)\end{aligned}$$

The log-likelihood is $\ell(\theta; X) = \log p_\theta(X)$.

Remark

- In general, argmax may not exist, be unique or be computable
- MLE does not depend on the parametrization: MLE for $g(\theta)$ is $g(\hat{\theta}_{\mathrm{MLE}})$

MLE in exponential family

$$p_\eta(x) = e^{\eta^\top T(x) - A(\eta)} h(x)$$

$$\ell(\eta; x) = \eta^\top T(x) - A(\eta) + \log h(x)$$

Find MLE. Is it unique?

Asymptotic of 1-parameter MLE in exponential family

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} e^{\eta T(x) - A(\eta)} h(x), \eta \in \Xi_0 \subseteq \mathbb{R}$$

Show that

- MLE is consistent
- MLE is asymptotically normal
- MLE achieves the CRLB asympotically

MLE for the natural parameter in Poisson distribution

$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim}$ Poisson$(\theta), \eta = \log(\theta)$

Find MLE for η

- Asymptotic distribution of MLE?
- finite-sample risk under squared error loss?

Asymptotically small modification does not have effect on the convergence in distribution

Prop.

If $\mathbb{P}(B_n) \to 0, X_n \Rightarrow X, Z_n$ arbitrary, then $X_n \mathbf{1}_{B_n^c} + Z_n \mathbf{1}_{B_n} \Rightarrow X$

proof: $\mathbb{P}(\left|Z_n \mathbf{1}_{B_n}\right| > \epsilon) \leq \mathbb{P}(B_n) \to 0. \mathbf{1}_{B_n^c} \xrightarrow{p} 1$, apply Slutsky

Asymptotic efficiency

Asymptotic efficiency

Setting

$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta(x), \theta \in \mathbb{R}^d$. p_θ is twice differentiable and the second derivative is continuous. Let $\ell_1(\theta; X_i) = \log p_\theta(X_i)$, $\ell_n(\theta; X) = \sum_{i=1}^n \ell_1(\theta; X_i)$

$$I_1(\theta) = \mathrm{Var}_\theta(\nabla \ell_1(\theta, X_1))$$
$$I_n(\theta) = \mathrm{Var}_\theta(\nabla \ell_n(\theta, X_1)) = n I_1(\theta)$$

Def. Asymptotic efficiency

We say an estimator $\hat{\theta}_n$ is asympotitcally efficient if

$$\sqrt{n}(\hat{\theta}_n - \theta) \Rightarrow \mathcal{N}(0, I_1(\theta)^{-1})$$

similar for $g(\theta)$

Intuition for the asymptotic distribution of MLE

At θ_0, we know many things about MLE

Denote the true parameter by θ_0.

Derivatives of ℓ_n at θ_0

- $\mathbb{E} \nabla \ell_1(\theta_0; X_i) = 0$
- $\mathrm{Var}\, \nabla \ell_1(\theta_0; X_i) = I_1(\theta_0)$
- $\frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0; X) \Rightarrow \mathcal{N}(0, I_1(\theta_0))$
- $\frac{1}{n} \nabla^2 \ell_n(\theta_0; X) \overset{p_{\theta_0}}{\to} -I_1(\theta_0)$

Informal proof for the asymptotic distribution of MLE

The estimating equation for MLE

$$0 = \nabla \ell_n(\hat{\theta}_n) = \nabla \ell_n(\theta_0) + \nabla^2 \ell_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)$$

Solve $\hat{\theta}_n$, we get

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\left(\frac{1}{n}\nabla^2 \ell_n(\tilde{\theta}_n)\right)^{-1} \frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0)$$

Pointwise convergence seems not enough, need uniform convergence results?

Informal asymptotic picture of MLE $(d = 1)$

The log-likelihood is approximated locally by a quadratic function

$$\ell_n(\theta) - \ell_n(\theta_0) \approx \ell_n'(\theta_0)(\theta - \theta_0) + \frac{1}{2}\ell_n''(\theta_0)(\theta - \theta_0)^2$$

Consistency of MLE

Question: assume the model is identifiable $P_\theta \neq P_{\theta_0}$ for $\theta \neq \theta_0$, when do we have

$$\hat{\theta}_n \xrightarrow{p} \theta_0?$$

KL divergence

Def. Kullback-Leibler (KL) divergence

$$\mathcal{D}_{\mathsf{KL}}(\theta_0 \parallel \theta) = \mathbb{E}_{\theta_0} \log \frac{p_{\theta_0}(X_1)}{p_\theta(X_1)}$$

Show that $\mathcal{D}_{\mathsf{KL}}(\theta_0 \parallel \theta) \geq 0$.

Difference in log-likelihood

Define

$$W_i(\theta) = \ell_1(\theta; X_i) - \ell_1(\theta_0; X_i)$$
$$\bar{W}_n = \frac{1}{n}\sum_{i=1}^{n} W_i$$

Note that

- $\hat{\theta}_n \in \arg\max_{\theta\in\Omega} \bar{W}_n(\theta)$
- $\bar{W}_n(\theta) \xrightarrow{p} -\mathcal{D}_{\mathsf{KL}}(\theta_0 \parallel \theta)$

But point-wise convergence is not enough to say about the argmax, need uniform convergence!

Convergence in sup-norm

Def.

For a compact set K, let $C(K) = \{f : K \to \mathbb{R}, \text{continuous}\}$. For $f \in C(K)$, let $\|f\|_\infty = \sup_{t \in K} |f(t)|$. We say $f_n \to f$ in sup-norm if

$$\|f_n - f\|_\infty \to 0.$$

Law of large numbers for random functions

Thm. 9.2 in Keener

Assume K compact, $W_1, W_2, \ldots$, i.i.d. random functions in $C(K)$, each with mean μ and $\mathbb{E}\left\|W_i\right\|_\infty < \infty$, then

$$\left\|\bar{W}_n - \mu\right\|_\infty \xrightarrow{p} 0$$

proof omitted. proof idea: make use of ϵ-cover of K.

This is stronger than the point-wise law of large numbers

Usefulness of uniform convergence

Thm. 9.4 in Keener

Let $G_1, G_2, \ldots,$ random functions in $C(K)$, K compact.
$\|G_n - g\|_\infty \xrightarrow{p} 0$, for some fixed $g \in C(K)$. Then

- If $t_n \xrightarrow{p} t^* \in K$, then $G_n(t_n) \xrightarrow{p} g(t^*)$
- If g is maximized at unique value t^* and $G_n(t_n) = \max_t G_n(t)$, then $t_n \xrightarrow{p} t^*$
- If $g(t) = 0$ has unique solution t^*, and t_n solves $G_n(t) = 0$, then $t_n \xrightarrow{p} t^*$

None of the above is true if one only has point-wise convergence proof:

Consistency of MLE for compact Ω

Thm. 9.9 in Keener

$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta_0}, p_\theta(x)$ continuous in θ for a.e. x. Assume

- Ω is compact
- $\mathbb{E}_{\theta_0} \|W_i\|_\infty < \infty$
- model identifiable

Then $\hat{\theta}_{\mathrm{MLE},n} \xrightarrow{p} \theta_0$

proof: Thm 9.2 + Thm 9.4

Consistency of MLE for general (non-compact) Ω

Thm. 9.11 in Keener

$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta_0}, p_\theta(x)$ continuous in θ for a.e. x. Assume

- For all compact K, $\mathbb{E} \left\| \mathbf{1}_K W_i \right\|_\infty < \infty$
- $\exists r > 0$, $\mathbb{E} \sup_{\|\theta - \theta_0\|_2 > r} W_i(\theta) < 0$
- model identifiable

Then $\hat{\theta}_{\mathrm{MLE},n} \xrightarrow{p} \theta_0$

proof: Let $A = \left\{\theta \mid \|\theta - \theta_0\|_2 > r\right\}, \alpha = \mathbb{E}\sup_{\theta \in A} W_i(\theta) < 0$
Consider $\hat{\theta}_n^A = \hat{\theta}_n \mathbf{1}_{\hat{\theta}_n \in A^c} + \theta_0 \mathbf{1}_{\hat{\theta}_n \in A}$
Show $\mathbb{P}(\hat{\theta}_n \in A) \to 0$

Finally, after proving consistency, we can derive the asymptotic distribution of MLE.

Asymptotic distribution of MLE

Thm. 9.14 in Keener

$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta_0}, \theta_0 \in \Omega$. Assume

- $\hat{\theta}_n \in \arg\max_\theta \ell_n(\theta; X)$, is consistent
- In a neighborhood $\bar{B}_\epsilon(\theta_0) = \left\{\theta : \|\theta - \theta_0\|_2 \le \epsilon\right\}$
    - $\ell_1(\theta; x)$ has two continuous derivatives, $\forall x$
    - $\mathbb{E}\left[\sup_{\theta \in \bar{B}_\epsilon} \left\|\nabla^2 \ell_1(\theta; X_i)\right\|_2\right] < \infty$
- Fisher information $I_1(\theta_0) \succeq 0$

Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}(0, I_1(\theta_0)^{-1})$$

proof sketch:

Summary

Under regular model,

- MLE is consistent
- MLE is asympotically normal, efficient

What is next?

Hypothesis testing

Thank you