Statistics 210A Lecture 21 Notes

Daniel Raban

November 4, 2021

1 Maximum Likelihood Estimation and Asymptotic Efficiency

1.1 Recap: Convergence in probability and distribution

Last time we introduced notions of convergence. We had

- Convergence in probability:

$$X_n \xrightarrow{p} c \quad \text{if} \quad \mathbb{P}(\|X_n - c\| > \varepsilon) \to 0 \quad \forall \varepsilon > 0.$$

- Convergence in distribution (sometimes called **weak convergence**[1]):

$$X_n \implies X \quad \text{if} \quad \mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)] \quad \forall \text{ bounded, continuous } f.$$

If $X_n, X \in \mathbb{R}$, then $X_n \implies X$ if and only if $F_n(x) \to F(x)$ for all x where F is continuous at x, where $F_n(x) = \mathbb{P}(X_n \le x)$ and $F(x) = \mathbb{P}(X \le x)$.

We had a few theorems will allow us to extend convergence to more random variables:

Theorem 1.1 (Continuous mapping)**.** *If f is continuous,*

$$X_n \to pX \implies f(X_n) \xrightarrow{p} f(X), \qquad X_n \to DX \implies f(X_n) \xrightarrow{D} f(X).$$

Theorem 1.2 (Slutsky)**.** *If $X_n \implies X$ and $Y_n \implies c$, then*

$$X_n + Y_n \implies X + c, \qquad X_n \cdot Y_n \implies cX, \qquad X_n / Y_n \implies X/c \quad (c \neq 0),$$

Theorem 1.3 (Delta method)**.** *Suppose $g(n)(X_n - \mu) \implies N_d(0, \Sigma)$, where $g(n) \to \infty$. Then for $f : \mathbb{R}^d \to \mathbb{R}^k$, where*

$$Df(x) = \begin{bmatrix} - & \nabla f_1(x)^\top & - \\ & \vdots & \\ - & \nabla f_k(x)^\top & - \end{bmatrix}$$

exists and is continuous at μ, then $g(n)(f(X_n) - f(\mu)) \implies N_k(0, Df(\mu)\Sigma Df(\mu)^\top)$.

[1] The real reason this is called weak convergence is that it corresponds to convergence of the distribution measures in a weak topology on $BC(\mathbb{R}^n)^*$, the dual space of the bounded continuous functions on $\mathbb{R}^n$.

1.2 Maximum likelihood estimators

Definition 1.1. Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a dominated family with densities p_θ for P_θ with respect to μ. The **maximum likelihood estimator (MLE)** is

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

If we are worried about whether this exists, i.e. if the maximum is achieved, we can just take some ε tolerance instead. For now, we won't worry about that.

Remark 1.1. This is invariant to parametrization. If we have two different parameterizations θ and $\eta(\theta)$, then $\widehat{\eta}_{\mathrm{MLE}} = \eta(\widehat{\theta}_{\mathrm{MLE}})$. This is not the case for, for example, the UMVU estimator.

Example 1.1. Let $p_\eta(x) = e^{\eta^\top T(x) - A(\eta)} h(x)$. The log likelihood is

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

so

$$\begin{aligned}\nabla\ell(\eta; X) &= T(X) - \nabla A(\eta) \\ &= T(X) - \mathbb{E}_\eta[T(X)].\end{aligned}$$

Note that $\nabla\ell$ is concave. If we set it equal to 0, we get something like a method of moments estimator.

$$\widehat{\eta}_{\mathrm{MLE}} \quad \text{solves} \quad T(X) = \mathbb{E}_{\widehat{\eta}}[T(X)].$$

Let $\mu = \psi(\eta) = \nabla A$. Then $\widehat{\eta} = \psi^{-1}(T(X))$.

Example 1.2. Let $X_i \overset{\text{iid}}{\sim} e^{\eta T(x) - A(\eta)} h(x)$ with $\eta \in \Xi \subseteq \mathbb{R}$. Then

$$\widehat{\eta} = \psi^{-1}(\overline{T}), \qquad \overline{T} = \frac{1}{n}\sum_{i=1}^n T(X_i).$$

Assume $\eta \in \Xi^o$ and $\dot{\psi}(\eta) = \ddot{A}(\eta) > 0$. Then ψ^{-1} is continuous, so

$$\frac{d}{d\mu}\psi^{-1}(\mu) = \frac{1}{\dot{\psi}(\psi^{-1}(\mu))} = \frac{1}{\ddot{A}(\eta)}.$$

By the law of large numbers, $\overline{T} \xrightarrow{p_\eta} \mu$. Here, we write p_η to emphasize that this convergence depends on η. So the continuous mapping theorem gives consistency: $\psi^{-1}(\overline{T}) \xrightarrow{p_\eta} \eta$.

The central limit theorem gives

$$\sqrt{n}(\overline{T} - \mu) \implies N(0, \operatorname{Var}_\eta T(X_1)) = N(0, \ddot{A}(\eta)),$$

where the Fisher information is $J_1^\eta(\mu) = \ddot{A}(\mu)^{-1}$. The delta method gives

$$\sqrt{n}(\widehat{\eta} - \eta) = \sqrt{(}\psi^{-1}(\overline{T}) - \eta) \implies N(0, \tfrac{1}{\ddot{A}(\eta)^2}\ddot{A}(\eta)) = N(0, \ddot{A}(\eta)^{-1}).$$

Recall that $J_1^\eta(\mu) = \operatorname{Var}_\eta(T(X_1)) = \ddot{A}(\eta)$. Asymptotically, $\widehat{\eta}$ is unbiased and achieves the Cramér-Rao lower bound because $J_n^\eta(\eta) = n\ddot{A}(\eta)$.

What do we mean by asymptotically unbiased?

Example 1.3. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Pois}(\theta) = \frac{\theta^x e^{-\theta}}{x!}$, and let $\eta = \log\theta$. Then $\widehat{\eta} = \log\overline{X}$. The central limit theorem says $(\overline{X} - \theta) \implies N(0, \theta)$, so the delta method tells us that

$$\sqrt{n}(\widehat{\eta} - \eta) = \sqrt{n}(\log\overline{X} - \log\theta) \implies N(0, \tfrac{1}{\theta^2}\theta) = N(0, \theta^{-1}).$$

What if $\overline{X} = 0$? In fact, $\widehat{\eta}$ has bias $-\infty$ and infinite variance, so the bias does not converge to 0. What we mean by asymptotically unbiased is that the scaled limiting distribution has no bias.

If you are a glass half-empty person, you might say that we can never use $\widehat{\eta}$, since it will always have infinite mean squared error. But if you are a glass half-full person, you might say that

$$\mathbb{P}_\eta(\overline{X} = 0) = \mathbb{P}_\theta(X_1 = 0)^n = e^{-n\theta},$$

which is an exponentially decaying probability of anything bad happening.

Proposition 1.1. *If $X_n \implies X$, Z_n is arbitrary, and B_n is an event such that $\mathbb{P}(B_n) \to 0$, then*

$$X_n \mathbb{1}_{B_n^c} + Z_n \mathbb{1}_{B_n} \implies X.$$

Proof. Observe that $Z_n \xrightarrow{p} 0$:

$$\mathbb{P}(\|Z_n \mathbb{1}_{B_n}\| > \varepsilon) \le \mathbb{P}(\|\mathbb{1}_{B_n}\| > \varepsilon) \to 0.$$

So $Z_n \mathbb{1}_{B_n} \xrightarrow{p} 0$. Since $\mathbb{1}_{B_n^c} \xrightarrow{p} 1$, use Slutsky's theorem to get the result. □

1.3 Asymptotic efficiency

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_\theta(x)$ with $\theta \in \mathbb{R}^d$. Assume that p_θ is "smooth" in θ, e.g. it has 2 continuous integrable derivatives. Let

$$\ell_1(\theta; X_i) = \log p_\theta(X_i), \qquad \ell_n(\theta, X) = \sum_{i=1}^n \ell_1(\theta; X_i).$$

Recall the Fisher information for a single observation is

$$H_1(\theta) = \mathrm{Var}_\theta(\nabla \ell_1(\theta; X_i)).$$

The likelihood ratio, which captures everything about the data, looks like

$$\frac{\mathrm{Lik}(\theta+\delta; X)}{\mathrm{Lik}(\theta; X)} = \log(\ell_n(\theta+\delta) - \ell_n(\theta)) \approx \log(\delta^\top \nabla \ell_n(\theta)).$$

The Fisher information for n observations is

$$J_n(\theta) = \mathrm{Var}_\theta(\nabla \ell_n(\theta; X)) = nJ_1(\theta).$$

Recall that $\mathbb{E}[\nabla \ell_1(\theta)] = 0$.

Definition 1.2. An estimator $\widehat{\theta}_n$ is **asymptotically efficient** if

$$\sqrt{n}(\widehat{\theta}_n - \theta) \overset{P_\theta}{\implies} N(0, J_1(\theta)^{-1}).$$

Really this is a sequence of estimators converging, but this is usually understood from context. For continuously differentiable $g(\theta) : \mathbb{R}^d \to \mathbb{R}$,

$$\sqrt{n}(g(\widehat{\theta}_n) - g(\theta)) \implies N(0, \nabla g(\theta)^\top J_1(\theta)^{-1} \nabla g(\theta)).$$

You may recognize this as the Cramér-Rao lower bound.

Let θ_0 be the true value, and let θ be a generic value of the parameter. We will maximize $\ell_n(\theta; X)$ by setting $\nabla \ell_n(\widehat{\theta}_{\mathrm{MLE}}) = 0$. We know $\nabla \ell_1(\theta_0; X_i) \overset{\mathrm{iid}}{\sim} (0, J_1(\theta_0))$, so by the central limit theorem,

$$\frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0; X) = \sqrt{n} \frac{1}{n} \sum_{i=1}^n \nabla \ell_1(\theta_0, X_i) \implies N(0, J_1(\theta_0)).$$

Now calculate the second derivative: Using the law of large numbers,

$$\frac{1}{n} \nabla^2 \ell_n(\theta_0, X) \xrightarrow{p} \mathbb{E}_{\theta_0}[\nabla^2 \ell(\theta_0; X_i)] = -J_1(\theta).$$

Here is an informal proof of why the MLE should be asymptotically efficient.

Proof. Assume

$$0 = \nabla \ell_n(\widehat{\theta}_n; X) = \nabla \ell_n(\theta_0) + \nabla^2 \ell_n(\theta_0)(\widehat{\theta}_n - \theta),$$

using a Taylor expansion. Then

$$\sqrt{(}\widehat{\theta}_n - \theta_0) \approx \underbrace{\left(-\frac{1}{n} \nabla^2 \ell_n(\theta_0)\right)^{-1}}_{\xrightarrow{p} J_1(\theta_0)^{-1}} \underbrace{\left(\frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0)\right)}_{\implies N(0, J_1(\theta_0))} \implies N(0, J_1(\theta_0)^{-1}),$$

which gives asymptotic efficiency. □

What's missing from this proof? To do our Taylor expansion, we need to first show that $\widehat{\theta}_n$ is close to θ_0; that is, we want to show consistency: $\widehat{\theta}_n \xrightarrow{p} \theta_0$.