

Outline

- 1) Maximum Likelihood Estimator
- 2) Asymptotic Distribution of MLE
- 3) Consistency of MLE

Maximum Likelihood Estimation

For a generic dominated family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with densities p_θ , a simple estimator for θ is

$$\begin{aligned}\hat{\theta}_{MLE}(X) &= \operatorname{argmax}_{\theta \in \Theta} p_\theta(x) \\ &= \operatorname{argmax}_{\theta \in \Theta} \ell(\theta; X)\end{aligned}$$

Remark 1: argmax may not exist, be unique, or be computable

Remark 2: doesn't depend on parameterization or base measure, MLE for $g(\theta)$ is $g(\hat{\theta}_{MLE})$

$$\underline{E}_X \quad p_\eta(x) = e^{\eta' T(x) - A(\eta)} h(x)$$

$$\ell(\eta; x) = \eta' T(x) - A(\eta) + \log h(x)$$

$$\nabla \ell(\eta; x) = T(x) - \mathbb{E}_\eta T(x)$$

$$\Rightarrow \hat{\eta}_{MLE} \text{ solves } T = \mathbb{E}_{\hat{\eta}} T \quad \text{if such } \eta \text{ exists}$$

Because $\nabla^2 \ell(\eta; x) = -\operatorname{Var}_\eta(T)$ is negative definite unless $\eta' T \stackrel{\text{a.s.}}{=} 0$ (in which case param. redundant)

$\Rightarrow$ at most 1 solution exists

$$\text{Let } \mu = \eta(\eta) = \nabla A(\eta), \quad \hat{\eta} = \eta^{-1}(T)$$

$$\underline{E_x} \quad X_i \stackrel{\text{iid}}{\sim} e^{\eta T(x) - A(\eta)} h(x) \quad \eta \in \Xi \subseteq \mathbb{R}$$

$$\hat{\eta} = \dot{\psi}(\bar{T}), \quad \bar{T} = \frac{1}{n} \sum T(x_i)$$

$$\text{Assume } \eta \in \Xi^\circ. \quad \dot{\psi}(\eta) = \ddot{A}(\eta) > 0 \quad \forall \eta \in \Xi^\circ$$

$$\text{so } \psi^{-1} \text{ cts, } (\dot{\psi}^{-1})(\eta) = \frac{1}{\dot{\psi}(\psi(\eta))} = \frac{1}{\ddot{A}(\eta)}$$

$$\text{Consistency: } \bar{T} \xrightarrow{P_n} \mu$$

$$\text{Cts mapping: } \dot{\psi}(\bar{T}) \xrightarrow{P_n} \dot{\psi}(\mu) = \eta$$

$$\text{Since } \sqrt{n}(\bar{T} - \mu) \Rightarrow N(0, \text{Var}_\eta(T(x_i))) \\ = N(0, \ddot{A}(\eta))$$

Delta method:

$$\text{(Recall } J_1(\eta) = \text{Var}(T)^{-1} \\ = \ddot{A}(\eta)^{-1} \text{)}$$

$$\sqrt{n}(\hat{\eta} - \eta) = \sqrt{n}(\dot{\psi}(\bar{T}) - \eta)$$

$$\Rightarrow N(0, \frac{1}{\ddot{A}(\eta)^2} \cdot \ddot{A}(\eta))$$

$$= N(0, \frac{1}{\ddot{A}(\eta)})$$

$$\text{Recall } J_1(\eta) = \text{Var}_\eta(T(x_i)) = \ddot{A}(\eta)$$

= Fisher info from 1 obs

$$\hat{\eta} \approx N(\eta, \frac{1}{n J_1(\eta)})$$

Asymptotically unbiased, Gaussian, achieves CRLB
($\text{corr}(\bar{T}, \hat{\eta}) \rightarrow 1$)

Ex $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Pois}(\theta)$, $\eta = \log \theta$

$$\hat{\eta} = \log \bar{X}, \quad \sqrt{n}(\bar{X} - \theta) \Rightarrow N(0, \theta)$$

$$\sqrt{n}(\hat{\eta} - \eta) = \sqrt{n}(\log \bar{X} - \log \theta)$$

$$\Rightarrow N(0, \theta \cdot \frac{1}{\theta^2}) \quad (\text{Delta method})$$

$$= N(0, \theta^{-1})$$

But $\forall$ finite n , $\forall \theta > 0$:

$$\begin{aligned} P_\theta(\hat{\eta} = -\infty) &= P_\theta(X_1 = 0)^n \\ &= e^{-\theta n} > 0 \end{aligned}$$

$$\Rightarrow E \hat{\eta} = -\infty \quad \text{Var}(\hat{\eta}) = \infty$$

[MLE can have embarrassing finite-sample performance despite being asy. optimal!]

Prop: If $P(B_n) \rightarrow 0$, $X_n \Rightarrow X$, Z_n arbitrary
then $X_n 1_{B_n^c} + Z_n 1_{B_n} \Rightarrow X$

Proof $P(\|Z_n 1_{B_n}\| > \varepsilon) \leq P(B_n) \rightarrow 0$ so $Z_n 1_{B_n} \xrightarrow{P} 0$

Also $1_{B_n^c} \xrightarrow{P} 1$, apply Slutsky $\boxtimes$

[So zany behavior has no effect on cug. in dist]

Asymptotic Efficiency

[The nice behavior of MLE we found in the exponential family case generalizes to a much broader class of models]

Setting $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} p_\theta(x)$ $\theta \in \Theta \subseteq \mathbb{R}^d$

p_θ "smooth" in θ , e.g. 2 cts integrable deriv.s
(can be relaxed)

Let $l_1(\theta; X_i) = \log p_\theta(X_i)$, $l_n(\theta; X) = \sum_{i=1}^n l_1(\theta; X_i)$

$$J_1(\theta) = \text{Var}_\theta(\nabla l_1(\theta; X_i)) = -\mathbb{E}_\theta[\nabla^2 l_1(\theta; X_i)]$$

$$J_n(\theta) = \text{Var}_\theta(\nabla l_n(\theta; X)) = n J_1(\theta)$$

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

$$\text{if } \sqrt{n}(\hat{\theta}_n - \theta) \stackrel{P_\theta}{\Rightarrow} \mathcal{N}(0, J_1(\theta)^{-1})$$

($g: \Theta \rightarrow \mathbb{R}$)

Delta method for differentiable estimand $g(\theta)$

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \stackrel{P_\theta}{\Rightarrow} \mathcal{N}(0, \nabla g(\theta)' J_1(\theta)^{-1} \nabla g(\theta))$$

also achieves CRLB if $\hat{\theta}_n$ does; g diff.

Asymptotic Dist. of MLE

Under mild conditions, $\hat{\theta}_{MLE}$ is asy. Gaussian, efficient

We will be interested in $\ell_n(\theta; X)$ as a function of θ

Notate "true" value as θ_0 ($X \sim P_{\theta_0}$)

Derivatives of ℓ_n at θ_0 : ($\theta_0 \in \Theta^0$)

$$\nabla \ell_1(\theta_0; X_i) \stackrel{iid}{\sim} (0, J_1(\theta_0))$$

$$\frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0; X) = \sqrt{n} \cdot \frac{1}{n} \sum \nabla \ell_1(\theta_0; X_i) \xrightarrow{P_{\theta_0}} N(0, J_1(\theta_0))$$

$$\frac{1}{n} \nabla^2 \ell_n(\theta_0; X) \xrightarrow{P_{\theta_0}} \mathbb{E}_{\theta_0} \nabla^2 \ell_1(\theta_0; X_i) = -J_1(\theta_0)$$

Proof sketch:

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

between $\theta_0, \tilde{\theta}_n$

$$\sqrt{n} (\hat{\theta}_n - \theta_0) = - \underbrace{\left(\frac{1}{n} \nabla^2 \ell_n(\tilde{\theta}_n) \right)^{-1}}_{\text{(want)} \xrightarrow{P} J(\theta_0)^{-1}} \underbrace{\frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0)}_{\Rightarrow N_d(0, J(\theta_0))}$$

$$\Rightarrow N_d(0, J(\theta_0)^{-1})$$

More rigorous proof later, but note we need consistency of $\hat{\theta}_n$ first to even justify Taylor expansion

Asymptotic Picture (d=1)

Recall $(\ell_n(\theta) - \ell_n(\theta_0))_{\theta \in \mathcal{U}}$ is minimal suff.

Quadratic approximation near θ_0 :

$$\ell_n(\theta) - \ell_n(\theta_0) \approx \underbrace{\dot{\ell}_n(\theta_0)}_{\approx N(0, nJ_1(\theta_0))} (\theta - \theta_0) + \frac{1}{2} \underbrace{\ddot{\ell}_n(\theta_0)}_{\approx -nJ_1(\theta_0)} (\theta - \theta_0)^2$$

Gaussian linear term Deterministic curvature

