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

$$\hat{\theta}_{MLE}(X) = \underset{\theta \in \Theta}{\arg\max}\ p_\theta(x)$$

$$= \underset{\theta \in \Theta}{\arg\max}\ \ell(\theta; X)$$

<u>Remark 1</u>: argmax may not exist, be unique, or be computable

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

<u>Ex</u> $p_\eta(x) = e^{\eta' T(x) - A(\eta)} h(x)$

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

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

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

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

$\Rightarrow$ at most 1 solution exists

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

Ex $X_i \overset{iid}{\sim} e^{\eta T(x) - A(\eta)} h(x) \qquad \eta \in \Xi \subseteq \mathbb{R}$

$$\hat{\eta} = \psi^{-1}(\bar{T}), \quad \bar{T} = \frac{1}{n}\sum T(X_i)$$

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

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

Consistency: $\bar{T} \xrightarrow{P_\eta} \mu$

Cts mapping: $\psi^{-1}(\bar{T}) \xrightarrow{P_\eta} \psi^{-1}(\mu) = \eta$

Since $\sqrt{n}(\bar{T} - \mu) \Rightarrow N(0, \mathrm{Var}_\eta(T(X_1)))$

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

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

Delta method:

$$\sqrt{n}(\hat{\eta} - \eta) = \sqrt{n}(\psi^{-1}(\bar{T}) - \eta)$$

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

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

Recall $J_1(\eta) = \mathrm{Var}_\eta(T(X_i)) = \ddot{A}(\eta)$

$=$ Fisher info from 1 obs

$$\hat{\eta} \approx N\left(\eta, \frac{1}{nJ_1(\eta)}\right)$$

Asymptotically unbiased, Gaussian, achieves CRLB

$(\mathrm{corr}(\bar{T}, \hat{\eta}) \to 1)$

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

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

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

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

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

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

$$\mathbb{P}_\theta(\hat{\eta} = -\infty) = \mathbb{P}_\theta(X_1 = 0)^n$$

$$= e^{-\theta n} > 0$$

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

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

Prop: If $\mathbb{P}(B_n) \to 0$, $X_n \Rightarrow X$, Z_n arbitrary

then $X_n 1_{B_n^c} + Z_n 1_{B_n} \Rightarrow X$

Proof $\mathbb{P}(\|Z_n 1_{B_n}\| > \varepsilon) \le \mathbb{P}(B_n) \to 0$ 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 cvg. 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 $X_1, \ldots, X_n \overset{iid}{\sim} p_\theta(x) \qquad \theta \in \Theta \subseteq \mathbb{R}^d$

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

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

$$J_1(\theta) = \text{Var}_\theta(\nabla \ell_1(\theta; X_i)) = -\mathbb{E}_\theta[\nabla^2 \ell_1(\theta; X_i)]$$

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

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

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

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

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

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

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 $\hat{\ell}(\theta; X)$ as a function of θ

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

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

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

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

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

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 $\theta_0, \hat{\theta}_n$

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

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

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

Asymptotic Picture ($d=1$)

Recall $(\ell_n(\theta)-\ell_n(\theta_0))_{\theta\in\Theta}$ 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