

Lecture 7 Maximum Likelihood

Maximum likelihood has been the most extensively used frequentist method in point estimation. Unlike empirical frequencies introduced in Lecture 6, maximum likelihood is a fully model-based approach. See Efron's Wald Lecture paper "Maximum Likelihood and Decision Theory" (1981, *Annals of Statistics*) for further discussion.

Given a sample $x \in \mathcal{X}$ from $f(x|\theta)$ with an unknown parameter θ , the *likelihood* $L(\theta|x) \triangleq f(x|\theta)$ is regarded as a function of θ . In the case of exponential families, it is more convenient to consider the (log-)likelihood $L(\theta|x) \triangleq \log f(x|\theta)$. In general, $L(\theta|x)$ is a measure of how likely is to have produced x , i.e. a certain value of θ appears to be more likely than others. $\hat{\theta}$ is called a *maximum likelihood estimate* (MLE) of θ based on x if

$$L(\hat{\theta}|x) = \max_{\theta \in \Theta} L(\theta|x).$$

Proposition 1 (*invariance property of MLE*) *MLE is preserved by parametrization $\lambda = q(\theta)$ with a known function q .*

Proof: Define the likelihood induced by q :

$$L^*(\lambda|x) \triangleq \sup_{\theta: q(\theta)=\lambda} L(\theta|x).$$

Then

$$L^*(\hat{\lambda}|x) = \max_{\lambda} L^*(\lambda|x) = \max_{\lambda} \sup_{\theta: q(\theta)=\lambda} L(\theta|x) = \sup_{\theta} L(\theta|x) = L(\hat{\theta}|x).$$

QED.

Note: Proposition 1 suggests a useful "plug-in" method $\hat{\lambda} = q(\hat{\theta}) = q(\hat{\theta})$ for finding MLEs which will be illustrated in a couple of examples later.

Theorem 1 *By absorbing $h(x)$ into the reference measure $\nu_{\mathcal{X}}$, we simply let $f(x|\theta) = \exp[\langle \theta, T(x) \rangle - \psi(\theta)]$, $x \in \mathcal{X}$, $\theta \in \Theta^{\circ}$ be a minimal exponential family (canonical form).*

- (i) *If the equation $\nabla \psi(\theta) = T(x)$ has a solution in Θ° , then the solution is unique, and is the MLE $\hat{\theta}$ of θ .*
- (ii) *If the natural parametrization $C(\lambda) = \theta$ is a bijection, then $\hat{\lambda} = C^{-1}(\hat{\theta})$ is the MLE of λ in the original form $f(x|\lambda) = \exp[\langle C(\lambda), T(x) \rangle + D(\lambda)]$.*

Proof: For (i), consider the likelihood $L(\theta|x) = \log f(x|\theta) = \langle \theta, T(x) \rangle - \psi(\theta)$. Given x , any θ satisfying $\nabla\psi(\theta) = T(x)$ is a critical point of $L(\theta|x)$, thus a unique maximizer of $L(\theta|x)$, because the Hessian matrix $\nabla^2 L(\theta|x) = -\nabla^2\psi(\theta)$ is strictly negative definite. (ii) follows from Proposition 1. *QED.*

Note:

(a) Theorem 1 assumes $\theta \in \Theta^\circ$. For θ sitting on the boundary $\partial\Theta$, a special care is needed.

(b) In an exponential family, the MLE $\widehat{q}(\theta)$ of $q(\theta) = E_\theta T(X) = \nabla\psi(\theta)$ is just $T(X)$, which agrees with the method-of-moments estimator.

Example 7.1 Let X_1, \dots, X_n be iid $N(\mu, \sigma^2)$ random variables, $\theta = (\mu, \sigma^2)$. A MLE $\hat{\theta}$ could be found using calculus. An alternative “short-cut”, useful in many other similar problems, is to apply Theorem 1 and Proposition 1. Write the density of X^n as

$$f(x^n|\theta) = (2\pi\sigma^2)^{-n/2} \exp\left[\frac{-1}{2\sigma^2} (T_2 - 2\mu T_1 + n\mu^2)\right],$$

where $T_1(x^n) = \sum_{i=1}^n x_i$ and $T_2(x^n) = \sum_{i=1}^n x_i^2$. Note that $E_\theta T_1 = n\mu$ and $E_\theta T_2 = n(\sigma^2 + \mu^2)$. Hence we have the MLEs $\hat{\mu} = \bar{X}$ and $\frac{1}{n} E_\theta T_2 = \widehat{\sigma^2} + (\hat{\mu})^2 = \frac{1}{n} \sum_{i=1}^n X_i^2$, which yield

$$\widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X})^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2,$$

and $\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}$. Furthermore, let $q(\theta) = P_\theta(X_1 > b)$ with a known constant b . Then $q(\theta) = 1 - \Phi\left(\frac{b-\mu}{\sigma}\right)$ where Φ is the cdf of $N(0, 1)$, which leads to $\widehat{q}(\theta) = q(\hat{\theta}) = 1 - \Phi\left(\frac{b-\bar{X}}{\sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}}\right)$.

Example 7.2 Let X_1, \dots, X_n be iid $\mathcal{U}(0, \theta)$ random variables. The likelihood $L(\theta|x^n) = \theta^{-n} I_{\{0 \leq x_{(1)} \leq x_{(n)} \leq \theta\}}$. Note that (a) $L(\theta|x^n) = 0 \forall \theta < x_{(n)}$; (b) $L(\theta|x^n)$ is decreasing for $\theta \geq x_{(n)}$. Hence $\hat{\theta} = X_{(n)}$ is the MLE.

Example 7.3 Let X_1, \dots, X_n be iid $\mathcal{U}(\theta - 1/2, \theta + 1/2)$ random variables. Since $L(\theta|x^n) = I_{\{\theta - 1/2 \leq x_{(1)} \leq x_{(n)} \leq \theta + 1/2\}} = I_{\{x_{(n)} - 1/2 \leq \theta \leq x_{(1)} + 1/2\}}$, any value in the interval $[x_{(n)} - 1/2, x_{(1)} + 1/2]$ is qualified for $\hat{\theta}$, i.e. MLEs are not unique in this case.

Example 7.4 Let X_1, \dots, X_n be iid samples from a Cauchy distribution with location parameter θ . Then

$$L(\theta|x^n) = c \prod_{i=1}^n [1 + (x_i - \theta)^2]^{-1}, \quad \text{with } c = \pi^{-n}.$$

Note that $0 < L(\theta|x^n) < 1 \forall \theta \in \mathbb{R}$; $L(\theta|x^n)$ is continuous in θ ; and $\lim L(\theta|x^n) = 0$ as $|\theta| \rightarrow \infty$. Hence a MLE $\hat{\theta}$ exists, but can only be found via numerical computation except for the case $n \leq 2$.