Maximum Likelihood Parameter Estimation

A popular non-random parameter estimation strategy which chooses the estimate to maximize the likelihood of observing what was actually observed.

1. Terminology/Formulae:

   (a) Maximum likelihood estimator

   $$\hat{\theta}_{\text{ML}}(y) \quad = \quad \arg\max_{\theta \in \Lambda} p_\theta(y)$$

   (b) Likelihood equation

   $$\left. \frac{\partial}{\partial \theta} \log p_\theta(y) \right|_{\theta = \hat{\theta}_{\text{ML}}} = 0$$

   - There may be many, one, or no solutions to the likelihood equation.
   - If $\hat{\theta}$ is an ML estimate and $\hat{\theta}$ lies in the interior of Λ, then $\hat{\theta}$ is a solution to the likelihood equation.
   - If $\hat{\theta}$ achieves the CRLB, then $\hat{\theta}$ is a solution to the likelihood equation.
   - If $p_\theta(y)$ is from an exponential family, there is a unique solution to the likelihood equation.

   (c) Consistency

   An estimator is consistent if $\hat{\theta}_n$ converges to θ (as $n \to \infty$) in some probabilistic sense.

   $\hat{\theta}_n \xrightarrow{i.p.} \theta$ *weakly consistent*
   $\hat{\theta}_n \xrightarrow{m.s.} \theta$ *mean-square consistent*
   $\hat{\theta}_n \xrightarrow{a.s.} \theta$ *strongly consistent*

   (d) Asymptotic Efficiency

   An estimator is *asymptotically efficient* if it is asymptotically unbiased *and* if its asymptotic variance approaches the CRLB.

2. Asymptotic Properties of MLE's for IID Observation Sequences:

   (a) *Consistency:*

   Under appropriate regularity conditions, $\hat{\theta}_n\big|_{\text{ML}} \xrightarrow{i.p.} \theta$.

   (b) *Asymptotic normality:*

   Under appropriate regularity conditions, $\sqrt{n}\left(\hat{\theta}_n\big|_{\text{ML}} - \theta\right) \longrightarrow \mathcal{N}\left(0, \frac{1}{i_\theta}\right)$ in distribution, where

   $$\begin{aligned} i_\theta \quad &= \quad \mathbf{E}_\theta \left\{ \left(\frac{\partial}{\partial \theta} \log f_\theta(y_1) \right)^2 \right\} \qquad \text{“information per sample”} \\ &= \quad \frac{I_\theta}{n} \quad \text{(since we assume i.i.d. observations)} \end{aligned}$$

   (c) *Asymptotic efficiency:*

   The consistency and asymptotic normality properties above, with additional regularity conditions, imply that *ML estimates are asymptotically efficient* (hence asymptotically MVUE).

3. Vector Parameter Case ($\Lambda \subset \mathbb{R}^m$):

   (a) Likelihood equation

   A system of m equations:

$$\begin{aligned} \frac{\partial}{\partial \theta_1} \log p_{\underline{\theta}}(y)\Big|_{\underline{\theta}=\hat{\underline{\theta}}_{\text{ML}}} &= 0 \\ \vdots \qquad & \quad \vdots \\ \frac{\partial}{\partial \theta_m} \log p_{\underline{\theta}}(y)\Big|_{\underline{\theta}=\hat{\underline{\theta}}_{\text{ML}}} &= 0 \end{aligned}$$

   (b) Fisher information matrix

$$\begin{aligned} \mathbf{I}_{\underline{\theta}} &= \mathbf{E}_{\underline{\theta}}\Big\{ \big(\tfrac{\partial}{\partial \underline{\theta}} \log p_{\underline{\theta}}(y)\big)\big(\tfrac{\partial}{\partial \underline{\theta}} \log p_{\underline{\theta}}(y)\big)^T \Big\} \\ \text{where } \tfrac{\partial}{\partial \underline{\theta}} \log p_{\underline{\theta}}(y) &= \Big(\tfrac{\partial}{\partial \theta_1} \log p_{\underline{\theta}}(y), \cdots, \tfrac{\partial}{\partial \theta_m} \log p_{\underline{\theta}}(y)\Big)^T \end{aligned}$$

   (c) Cramér-Rao lower bound (CRLB)

   If, for matrices A and B, we use $A \geq B$ to denote that $(A-B)$ is non-negative definite, then

$$\begin{aligned} \mathbf{Cov}_{\underline{\theta}}(\hat{\underline{\theta}}) = \mathbf{E}_{\underline{\theta}}\left\{ (\hat{\underline{\theta}} - \underline{\theta})(\hat{\underline{\theta}} - \underline{\theta})^T \right\} &\geq \mathbf{I}_{\underline{\theta}}^{-1} \\ \Rightarrow \mathbf{Var}_{\underline{\theta}}(\hat{\theta}_i) &\geq \left[\mathbf{I}_{\underline{\theta}}^{-1}\right]_{i,i} \end{aligned}$$

   (d) Asymptotics for i.i.d. observations

   Under similar conditions as in the scalar case, consistency and asymptotic normality still hold:

   - $\|\hat{\underline{\theta}}_n - \underline{\theta}\|_2 \longrightarrow 0$ in probability
   - $\sqrt{n}\left(\hat{\underline{\theta}}_n - \underline{\theta}\right) \longrightarrow \mathcal{N}\left(\underline{0}, \mathbf{i}_{\underline{\theta}}^{-1}\right)$ in distribution

4. Transformation of Parameters:

   (a) CRLB

   Say $\mathbf{I}_{\underline{\theta}}$ is the Fisher information matrix for family $\{P_{\underline{\theta}}(y), \underline{\theta} \in \Lambda\}$, and $\hat{\underline{g}}(y)$ is an estimate of $\underline{g}(\underline{\theta})$ for some arbitrary function $\underline{g} : \mathbb{R}^m \to \mathbb{R}^r$. Then

$$\mathbf{Cov}_{\underline{\theta}}(\hat{\underline{g}}) \geq \frac{\partial \underline{g}(\underline{\theta})}{\partial \underline{\theta}} \mathbf{I}_{\underline{\theta}}^{-1} \frac{\partial \underline{g}(\underline{\theta})}{\partial \underline{\theta}}^T$$

$$\text{where } \frac{\partial \underline{g}(\underline{\theta})}{\partial \underline{\theta}} = \begin{pmatrix} \frac{\partial g_1(\underline{\theta})}{\partial \theta_1} & \frac{\partial g_1(\underline{\theta})}{\partial \theta_2} & \cdots & \frac{\partial g_1(\underline{\theta})}{\partial \theta_m} \\ \frac{\partial g_2(\underline{\theta})}{\partial \theta_1} & \ddots & & \frac{\partial g_2(\underline{\theta})}{\partial \theta_m} \\ \vdots & & \ddots & \vdots \\ \frac{\partial g_r(\underline{\theta})}{\partial \theta_1} & \frac{\partial g_r(\underline{\theta})}{\partial \theta_2} & \cdots & \frac{\partial g_r(\underline{\theta})}{\partial \theta_m} \end{pmatrix}$$

   (b) ML invariance property

   Say $\hat{\underline{\theta}}_{\text{ML}}$ is the ML estimate of $\underline{\theta}$ (given the family of distributions $\{P_{\underline{\theta}}(y), \underline{\theta} \in \Lambda\}$), and $\underline{g}(\cdot)$ is one-to-one. Then $\underline{g}(\hat{\underline{\theta}}_{\text{ML}})$ is the ML estimate of $\underline{g}(\underline{\theta})$.