Maximum Likelihood Estimation (MLE)

① Notation

1) $X \sim f(\cdot|\theta)$

2) $\mathcal{L}(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$ ⟶ likelihood

$$\ell(\theta) = \sum_{i=1}^{n} \ell_i(\theta) = \sum_{i=1}^{n} \log\{f(x_i|\theta)\}$$

3) $\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \mathcal{L}(\theta) = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \ell(\theta)$

$\hat{\theta}_{MLE}$ is RV with respect to $(X_1, \dots, X_n)$

4) Quantity w.r.t log-likelihood $\ell(\theta)$

① Score Function $S(\theta) = \nabla_\theta \ell(\theta)$ ⟶ RV !!!

② Fisher Information $I(\theta) = \mathbb{E}_\theta[S(\theta)S(\theta)^T]$

(Differentiability) Under Some Regularity Condition

$$= -\mathbb{E}_\theta[\nabla_\theta S(\theta)]$$
$$= -\mathbb{E}_\theta[\nabla_\theta^2 \ell(\theta)] = -H(\theta)$$

Note: ① $\mathbb{E}_\theta[S(\theta)S(\theta)^T] = \mathrm{Var}_\theta[S(\theta)]$ since $\mathbb{E}_\theta[S(\theta)] = 0$

↳ this quantity will be used in MLE asymptotic property

② $\mathbb{E}_\theta[\nabla_\theta^2 \ell(\theta)]$ (= $-H(\theta)$) will be used in:

algorithm to solve $\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \ell(\theta)$

③ Under some regularity conditions, $\mathbb{E}_\theta[S(\theta)S(\theta)^T] = -\mathbb{E}_\theta[\nabla_\theta^2 \ell(\theta)]$

$\Rightarrow I(\theta) = -H(\theta)$

④ Summary:

a) $I(\theta) \rightsquigarrow$ Fisher Information $\rightsquigarrow \mathbb{E}[S(\theta)S(\theta)^T]$

$= \text{var}[S(\theta)]$

$\hat{\theta} - \theta_0 \sim N(0, \frac{1}{nI(\theta)})$

$\Rightarrow I(\theta)$ large $\longrightarrow$ variance small $\longrightarrow$ efficiency

b) $H(\theta) \rightsquigarrow$ Hessian of log-likelihood $\ell(\theta)$

$$H(\theta) := \nabla^2_\theta \ell(\theta)$$

Curvature information!

$\Rightarrow H(\theta)$ large $\longrightarrow$ more curved $\longrightarrow$ converge faster

---

② Detailed Discussion

include the Consistency

1. MLE $\Leftrightarrow$ Empirical Risk Minimization (ERM)

Recap: $\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\text{argmax}}\ \ell(\theta) = \underset{\theta \in \Theta}{\text{argmax}} \sum_{i=1}^{n} \log f(x_i|\theta)$

$\Leftrightarrow \hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i|\theta_0)}{f(x_i|\theta)}$

$= \underset{\theta \in \Theta}{\text{argmin}}\ R_n(\theta_0, \theta) := \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i|\theta_0)}{f(x_i|\theta)}$

Therefore, $R(\theta_0, \theta) = \mathbb{E}_{\theta_0}[R_n(\theta_0, \theta)]$

$= \mathbb{E}_{\theta_0}\left[\log \frac{f(X|\theta_0)}{f(X|\theta)}\right]$

$$= \int_{\mathcal{X}} \log \frac{f(x|\theta_0)}{f(x|\theta)} \cdot f(x|\theta_0)\, dx$$

$$= KL\left(f(x|\theta_0) \,\|\, f(x|\theta)\right)$$

2 things:

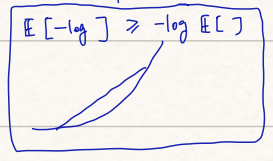
1. From LLN, for fix $\theta \in \Theta$.

$$R_n(\theta_0, \theta) \xrightarrow{P} R(\theta_0, \theta)$$

2. $\theta_0 \in \underset{\theta \in \Theta}{\operatorname{argmin}} \; R(\theta_0, \theta)$

$$= \underset{\theta \in \Theta}{\operatorname{argmin}} \; KL\left(f(x|\theta_0) \,\|\, f(x|\theta)\right)$$

Diagram Sketch:



$$R(\theta_0, \theta) = \mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X|\theta_0)}{f(X|\theta)}\right)\right]$$

$$= -\mathbb{E}_{\theta_0}\left[\log \frac{f(X|\theta)}{f(X|\theta_0)}\right]$$

$$\geq -\log \mathbb{E}_{\theta_0}\left[\frac{f(x|\theta)}{f(x|\theta_0)}\right]$$

$$= -\log 1 = 0$$

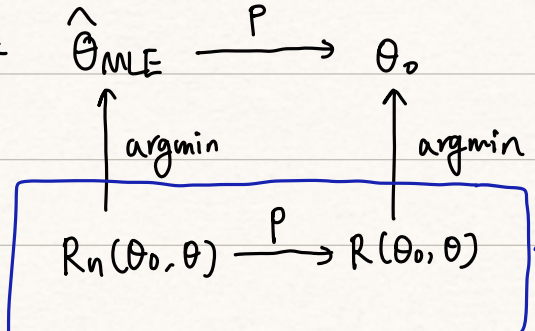
To conclude, $\max_{\theta \in \Theta} \ell(\theta) \Leftrightarrow \min_{\theta \in \Theta} R_n(\theta_0, \theta)$

$$\updownarrow$$

$$R(\theta_0, \theta) := KL\left(f(x|\theta_0) \,\|\, f(x|\theta)\right)$$

$$= \mathbb{E}_{\theta_0}\left[\log \frac{f(x|\theta_0)}{f(x|\theta)}\right]$$

Next is: we want $\hat{\theta}_{MLE} \xrightarrow{P} \theta_0$



$R_n(\theta_0, \theta) \xrightarrow{P} R(\theta_0, \theta)$ ~ too weak (we need uniform converge result)

Thm: Assumption { a) Strong Identifiability → $\inf_{\tilde{\theta}: |\tilde{\theta}-\theta| \geq \varepsilon} KL(p(\theta); p(\tilde{\theta})) > 0$; b) Uniform LLN }

"Continuity assump."

$$\Rightarrow \hat{\theta}_{MLE} \xrightarrow{P} \theta_0 \quad \text{(consistency)}$$

Sketch: Strong Identifiability implies that:

if we want $\hat{\theta}_{MLE} \xrightarrow{P} \theta_0$ uniformly

then we only require $KL(f(X|\theta_0) \,\|\, f(X|\hat{\theta}_{MLE})) \xrightarrow{P} 0$

$= R(\theta_0, \hat{\theta}_{MLE})$ Sufficiently small !!!

In Binary Classification ERM Framework, we try to analyze:

$\tilde{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} R(\theta, \theta_0)$

$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} R_n(\theta, \theta_0)$

$$|R(\hat{\theta}, \theta_0) - R(\tilde{\theta}, \theta_0)|$$
$$= \Big| \underbrace{\big(R(\hat{\theta}, \theta_0) - R_n(\hat{\theta}, \theta_0)\big)}_{\text{uniform LLN}} + \underbrace{\big(R_n(\hat{\theta}, \theta_0) - R_n(\tilde{\theta}, \theta_0)\big)}_{\text{defn of } \hat{\theta}} + \underbrace{\big(R_n(\tilde{\theta}, \theta_0) - R(\tilde{\theta}, \theta_0)\big)}_{\text{pointwise LLN}} \Big|$$

Here, actually we assume $\theta_0 \in \Theta$. Thus,

$$R(\theta_0, \hat{\theta}_{MLE}) = \underbrace{R(\theta_0, \hat{\theta}_{MLE}) - R_n(\theta_0, \hat{\theta}_{MLE})}_{\text{uniform LLN}} + R_n(\theta_0, \hat{\theta}_{MLE})$$

$R_n(\theta_0, \hat{\theta}_{MLE}) \leq R_n(\theta_0, \theta_0) = 0$ (defn of $\hat{\theta}_{MLE}$)

Finish the discussion for Wasserman LEC15 Part 4

2. (Extension of 1.) Suppose $\theta_0 \& \Theta$

(Sketch) $\tilde{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} R(\theta_0, \theta)$ $\quad KL(f(X|\theta_0) \,\|\, f(X|\theta))$

$$= \underset{\theta \in \Theta}{\operatorname{argmin}} \; \mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X|\theta_0)}{f(X|\theta)}\right)\right]$$

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmin}} \; R_n(\theta_0, \theta)$$

$$= \underset{\theta \in \Theta}{\operatorname{argmin}} \; \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i|\theta_0)}{f(x_i|\theta)}$$

Interpretation of $\tilde{\theta}$ → $KL(f_{\theta_0} \| f_{\tilde{\theta}}) \leq KL(f_{\theta_0} \| f_\theta) \quad \theta \in \Theta$

⇓ ✡✡

$f_{\tilde{\theta}}$ is the 'KL projection' from f_{θ_0}

to $\{f_\theta : \theta \in \Theta\}$

Similar Analyze $\rightarrow$ $|R(\theta_0, \hat{\theta}_{MLE}) - R(\theta_0, \tilde{\theta})|$

$= |(R(\theta_0; \hat{\theta}_{MLE}) - R_n(\theta_0; \hat{\theta}_{MLE}))$ ← uniform LLN

$+ (R_n(\theta_0; \hat{\theta}_{MLE}) - R_n(\theta_0; \tilde{\theta}))$ ← by defn of $\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmin}} R_n(\theta_0; \theta)$

$+ (R_n(\theta_0; \tilde{\theta}) - R(\theta_0; \tilde{\theta}))|$ ← pointwise LLN

3. Asymptotic Behaviour of $\hat{\theta}_{MLE}$

a) Quantity Recap

score function

① $S(X, \theta) = \nabla_\theta \ell(\theta)$ ← RV wrt $(X_1, \dots, X_n)$

$= \sum_{i=1}^n \nabla_\theta \ell_i(\theta) = \sum_{i=1}^n \nabla_\theta \log f(X_i | \theta)$

↳ $S_i(\theta) := \nabla_\theta \ell_i(\theta)$

$S(\theta)$ for simplicity

$$\mathbb{E}_\theta[S(\theta)] = \mathbb{E}_\theta[S_i(\theta)] = \mathbb{E}_\theta[\nabla_\theta \ell_i(\theta)] \equiv 0$$

② Fisher Information Matrix (FIM)

$(I_1(\theta) := \mathbb{E}_\theta[S_1(\theta) S_1(\theta)^T] = \operatorname{Var}_\theta[S_1(\theta)])$

$I(\theta) = \mathbb{E}_\theta[S(\theta) S(\theta)^T]$ ═══ regular condition (weak) ═══ $\mathbb{E}_\theta[-\nabla_\theta^2 \ell(\theta)] = -H(\theta)$

$= \operatorname{Var}_\theta[S(\theta)]$

$H_1(\theta) = \mathbb{E}_\theta[\nabla_\theta^2 \ell_1(\theta)]$

To conclude, RV $S(\theta) = \sum_{i=1}^n \nabla_\theta \ell_i(\theta)$

→ expectation 0

→ variance $I(\theta)$

③ $\nabla_\theta \ell(\hat{\theta}_{MLE}) = 0$ (roughly speaking)

b) **Result** $\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \sim \mathcal{N}(0, I_1(\theta_0)^{-1})$

Pf: $\nabla_\theta \ell(\hat{\theta}_{MLE}) = \nabla_\theta \ell(\theta_0) + \nabla_\theta^2 \ell(\tilde{\theta})(\hat{\theta}_{MLE} - \theta_0)$

$= 0$

$\tilde{\theta} = \lambda\theta_0 + (1-\lambda)\hat{\theta}_{MLE} \quad 0 < \lambda < 1$

$\Rightarrow \hat{\theta}_{MLE} - \theta_0 = -\nabla_\theta^2 \ell(\tilde{\theta})^{-1} \nabla_\theta \ell(\theta_0)$

$\Leftrightarrow \sqrt{n}(\hat{\theta}_{MLE} - \theta_0) = \underbrace{(-n^{-1}\nabla_\theta^2 \ell(\tilde{\theta}))^{-1}}_{①} \underbrace{(n^{-\frac{1}{2}} \nabla_\theta \ell(\theta_0))}_{②}$

(since $\tilde{\theta} \xrightarrow{P} \theta_0$)

① (LLN) $\Rightarrow -n^{-1}\nabla_\theta^2 \ell(\tilde{\theta}) \xrightarrow{P} -n^{-1}\nabla_\theta^2 \ell(\theta_0)$

$\xrightarrow{P} -\mathbb{E}_{\theta_0}[\nabla_\theta^2 \ell_1(\theta_0)]$

$\sqrt{n}\left(\frac{\frac{1}{n}\sum \nabla_\theta \ell_i(\theta_0)}{n} - 0\right)$ $= I_1(\theta_0)$

② (CLT) $\Rightarrow n^{-\frac{1}{2}} \nabla_\theta \ell(\theta_0) \xrightarrow{d} \mathcal{N}(0, \text{var}_{\theta_0}[\nabla_\theta \ell_1(\theta_0)])$

$= \mathcal{N}(0, I_1(\theta_0))$

③: ① + ② + Slutsky $\Rightarrow \sqrt{n}(\hat{\theta}_{MLE} - \hat{\theta}_0)$

$= (-n^{-1}\nabla_\theta^2 \ell(\tilde{\theta}))^{-1}(n^{-\frac{1}{2}}\nabla_\theta \ell(\theta_0))$

$\xrightarrow{d} I_1(\theta_0)^{-1}\mathcal{N}(0, I_1(\theta_0))$

$= \mathcal{N}(0, I_1(\theta_0)^{-1})$

$I_1(\theta_0) := \mathbb{E}_{\theta_0}[S_1(\theta_0) S_1(\theta_0)^T]$

$= -\mathbb{E}_{\theta_0}[\nabla_\theta^2 \ell(\theta_0)]$

C) Influence Function Framework

From the previous proof, actually we have shown:

$$\hat{\theta}_{MLE} = \theta_0 - \frac{1}{n}\sum_{i=1}^{n} \frac{\nabla_\theta \log f(x_i \mid \theta_0)}{I_1(\theta_0)} + \text{reminder}$$

$$= \theta_0 + \frac{1}{n}\sum_{i=1}^{n} \psi(x_i) \rightarrow \text{more general expression}$$

11/17

GLM Framework Recap

- Score Function: $S_i(\beta) = \nabla_\beta \ell_i(\beta)$ RV
- Fisher Information Matrix: $I_i(\beta) = \mathbb{E}_\beta[S_i(\beta) S_i(\beta)^T]$

  $= -\mathbb{E}_\beta[\nabla_\beta^2 \ell_i(\beta)]$
- Observation Matrix $J_i(\beta) = -\nabla_\beta^2 \ell_i(\beta)$

⇒ **Briefly go through**

1. **Stochastic Part** : $Y \sim$ exponential family $\quad \theta \in \mathbb{R}^q$

$$f(y|\theta,\phi) = \exp\left\{ \frac{y^T\theta - b(\theta)}{a(\phi)} + c(\phi, y) \right\}$$

Property : ① $\mathbb{E}_\theta[Y] = \nabla_\theta b(\theta)$

② $\mathrm{Var}_\theta[Y] = a(\phi) \cdot \nabla_\theta^2 b(\theta)$

$:= \Sigma(\theta, \phi)$

Model $\longrightarrow$ $\{y^1, \dots, y^n\}$ $\qquad$ **Model** $y_i \longleftrightarrow \theta^i$

$$\mathbb{E}[y^i] = \mu^i = \left[\begin{array}{c} \frac{\partial b}{\partial \theta_1} \\ \vdots \\ \frac{\partial b}{\partial \theta_q} \end{array} \right] \Bigg|_{\theta = \theta^i}$$

2. **Systematic Part** : $\eta^i = X^i \cdot \beta$ $\qquad \begin{cases} X^i \in \mathbb{R}^{q \times p} \\ \beta \in \mathbb{R}^p \end{cases}$

3. **Link function** : $g(\mu^i) = \eta^i \iff h(\eta^i) = \mu^i$

$$\left[\mu^i = \nabla_\theta b(\theta) \right]$$

$\xi : \eta^i \longrightarrow \theta^i$

General Architecture

$\theta^i \longrightarrow y^i \longrightarrow \mathbb{E}_{\theta^i}[y^i] = \mu^i = \nabla_\theta b(\theta^i)$ $\quad \overset{\text{Link}}{\longleftrightarrow} \quad \eta^i = X^i \beta$

4. MLE $\Longrightarrow$ $\hat{\beta}_{MLE}$ $\Rightarrow$ argmax $l(\beta)$

$$\ell(\beta) = \sum_{i=1}^{N} \ell_i(\beta)$$

$$= \sum_{i=1}^{N} \left\{ \underbrace{\frac{(y^i)^T \theta^i - b(\theta^i)}{a(\phi)}}_{\text{denote as } \ell_i(\beta)} + \text{const} \right\}$$

Then, $\hat{\beta}_{MLE} = \underset{\beta}{\text{argmax}} \sum_{i=1}^{N} \ell_i(\beta)$

5. One algo is, (Fisher Scoring algo) — Newton Method

$$\hat{\beta}_{t+1} = \hat{\beta}_t + I(\hat{\beta}_t)^{-1} S(\hat{\beta}_t)$$

Remark:
→ in the conventional framework
$\hat{\beta}_{t+1} \leftarrow \hat{\beta}_t + J(\hat{\beta}_t)^{-1} S(\hat{\beta}_t)$
→ Here, replace $J(\hat{\beta}_t)$ with $I(\hat{\beta}_t)$

Tricks

$$S(\beta) = \nabla_\beta \ell(\beta)$$

$$I(\beta) = -\mathbb{E}_\beta[\nabla_\beta^2 \ell(\beta)] = \mathbb{E}_\beta[S(\beta) S(\beta)^T]$$

$$J(\beta) = -\nabla_\beta^2 \ell(\beta)$$

Define $S_i(\beta) = \nabla_\beta \ell_i(\beta)$ $\quad I_i(\beta) = -\mathbb{E}_\beta[\nabla_\beta^2 \ell_i(\beta)] = \mathbb{E}_\beta[J_i(\beta)]$

Then $J_i(\beta) = -\nabla_\beta^2 \ell_i(\beta) = I_i(\beta) - \frac{1}{a(\phi)} R_i(\beta)$

$I_i(\beta)$ → fix; $\frac{1}{a(\phi)} R_i(\beta)$ → random part

Note: when applying Canonical Link Function: $J_i(\beta) = \mathbb{E}_\beta[J_i(\beta)] = I_i(\beta)$

→ that is to say, $J_i(\beta)$ is not a RV!

After Careful Calculation, the Fisher Scoring Algo

turns to

Iterative Reweight Least Square

- for $Y \rightsquigarrow$ 1-dim → check Lecture 3 in ST5213
- for $Y \rightsquigarrow q$-dim → check GLM in ST5213

6. Asymptotic Result for $\hat{\beta}_{MLE}$

① $$\begin{cases} \hat{\beta}_{MLE} - \beta_0 \xrightarrow{d} N(0, I_1(\beta_0)^{-1}) \\ \hat{\beta}_{MLE} - \beta_0 \xrightarrow{d} N(0, J_1(\beta_0)^{-1}) \end{cases}$$

② Hypothesis Testing

a) LRT $\begin{cases} H_0: \beta = \text{given vector} \\ H_A: \text{o/w} \end{cases}$

$$-2\log \Lambda \xrightarrow{d} \chi^2_{df}$$

b) Wald Testing ← asymptotic normality

$I(\beta) = E_\beta [\, S(\beta)\, S(\beta)^T \,]$

- closed form

Standard error $\begin{cases} ① \text{ exact} \quad SE(\hat{\beta}_j) = \text{diag}(I(\hat{\beta}))_j \\ ② \text{ Bootstrap to estimate SE if the sample size is moderate} \end{cases}$

↓

Asymptotic Variance may have large BIAS