Maximum-likelihood Estimates I

Ching-Kang Ing (銀慶剛)

Institute of Statistics, National Tsing Hua University, Taiwan

Maximum Likelihood Estimates

• **Likelihood function.** The likelihood function of n random variables $y_1, \cdots, y_n$ is defined to be the joint density of the n random variables, say $f_{\mathbf{y}}(y_1, \cdots, y_n; \boldsymbol{\theta})$, which is considered to be a function of $\boldsymbol{\theta}$.

In particular, if $y_1, \cdots, y_n$ are i.i.d. (which is called a "random sample") from the p.d.f. $f(y; \boldsymbol{\theta})$, then the likelihood function is $f(y_1; \boldsymbol{\theta}) \times \cdots \times f(y_n; \boldsymbol{\theta})$.

Remark. To remind ourselves to think of the likelihood function as a function of $\boldsymbol{\theta}$, in the following we shall use $L(\boldsymbol{\theta}; y_1, \cdots, y_n), L(\boldsymbol{\theta}; \mathbf{y})$ or $L(\boldsymbol{\theta})$ for the likelihood function $f_{\mathbf{y}}(y_1, \cdots, y_n; \boldsymbol{\theta})$.

- Maximum likelihood Estimate (MLE):

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}),$$

where Θ is the parameter space of interest.

Example 1. If $y_1, \cdots, y_n$ are i.i.d. from the Bernoulli distribution

$$f(y;p) = p^y(1-p)^{1-y} I_{\{0,1\}}(y), \quad 0 \le p \le 1,$$

then the MLE of p is $\hat{p} = \dfrac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$.

Example 2. If $y_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, then the MLE of $(\mu, \sigma^2)'$ is

$$\begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} = \begin{pmatrix} \bar{y} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(y_i - \bar{y})^2 \end{pmatrix}.$$

Example 3. If $y_i \overset{indep.}{\sim} N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$, then the MLE of $(\boldsymbol{\beta}, \sigma^2)'$ is

$$\begin{pmatrix} (X'X)^{-1}X'\mathbf{y} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2 \end{pmatrix}.$$

Here, we denoted $(X'X)^{-1}X'\mathbf{y}$ by $\hat{\boldsymbol{\beta}}$.

Example 4. (A location-dispersion model) If

$$y_i \overset{indep.}{\sim} N\left(\mathbf{x}_i'\boldsymbol{\beta}, e^{\mathbf{x}_i'\boldsymbol{\alpha}}\right),$$

then

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}) = (2\pi)^{-n/2}\prod_{i=1}^{n} e^{-\mathbf{x}_i'\boldsymbol{\alpha}/2} \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 e^{-\mathbf{x}_i'\boldsymbol{\alpha}}\right),$$

hence the log-likelihood function $l(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \log L(\boldsymbol{\beta}, \boldsymbol{\alpha}) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}\boldsymbol{\alpha}'\mathbf{x}_i \\ & \quad -\frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 e^{-\mathbf{x}_i'\boldsymbol{\alpha}}, \end{aligned}$$

here, log is the natural log function.

Define $\boldsymbol{\eta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})'$. The MLE of $\boldsymbol{\eta}$ is the solution to $\dfrac{\partial l(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \mathbf{0}$, which has no closed from. We therefore need to solve the equation numerically.

This example motivates us to explore the behavior of MLE without having its closed form. Let's start with some "non-asymptotic" results before moving on to asymptotic investigations.

Invariance Principle

- **Invariance principle of MLE.**

If $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, then $\boldsymbol{\tau}(\hat{\boldsymbol{\theta}})$ is the MLE of $\boldsymbol{\tau}(\boldsymbol{\theta})$, where $\boldsymbol{\tau}(\cdot)$ is a well-defined function.

proof. Define

$$M(\boldsymbol{\tau}) = \sup_{\{\boldsymbol{\theta}:\boldsymbol{\tau}(\boldsymbol{\theta})=\boldsymbol{\tau}\}} L(\boldsymbol{\theta}) \quad \text{and} \quad \Lambda = \{\boldsymbol{\tau}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\},$$

where $M(\boldsymbol{\tau})$ is the likelihood function induced by $\boldsymbol{\tau}(\cdot)$.

Then, for any $\boldsymbol{\tau} \in \Lambda$,

$$\begin{aligned} M(\boldsymbol{\tau}) &= \sup_{\{\boldsymbol{\theta}:\boldsymbol{\tau}(\boldsymbol{\theta})=\boldsymbol{\tau}\}} L(\boldsymbol{\theta}) \le \sup_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}) = L(\hat{\boldsymbol{\theta}}) \\ &= \sup_{\{\boldsymbol{\theta}:\boldsymbol{\tau}(\boldsymbol{\theta})=\boldsymbol{\tau}(\hat{\boldsymbol{\theta}})\}} L(\boldsymbol{\theta}) = M(\boldsymbol{\tau}(\hat{\boldsymbol{\theta}})). \end{aligned}$$

Hence

$$M(\boldsymbol{\tau}(\hat{\boldsymbol{\theta}})) = \sup_{\boldsymbol{\tau} \in \Lambda} M(\boldsymbol{\tau}) = M(\hat{\boldsymbol{\tau}}),$$

where $\hat{\boldsymbol{\tau}}$ is the MLE of $\boldsymbol{\tau}$, yielding the desired conclusion.

Some Facts

Fact 1. $E(\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta})) = \mathbf{0}$.

Sketch:

$$\begin{aligned} E(\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta})) &= \int \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}) f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta}) d\mathbf{y} \\ &= \int \frac{1}{f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}} f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta}) f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta}) d\mathbf{y} \\ &\overset{(*)}{=} \frac{\partial}{\partial \boldsymbol{\theta}} \int f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta}) d\mathbf{y} = \frac{\partial}{\partial \boldsymbol{\theta}} 1 = \mathbf{0}. \end{aligned}$$

$(*)$: use the regularity condition: "$\int$" and "$\frac{\partial}{\partial \boldsymbol{\theta}}$" can be exchanged.

Fact 2.

$$E\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}l(\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}l(\boldsymbol{\theta})\right)'\right] = E\left(-\frac{\partial^2}{\partial\boldsymbol{\theta}^2}l(\boldsymbol{\theta})\right).$$

The proof is similar to that of Fact 1.

Cramér-Rao lower bound

※ **Cramér-Rao lower bound (vector version).**

Let $\mathbf{T}$ be an unbiased estimator of $\boldsymbol{\theta}$. Then,

$$Var(\mathbf{T}) \geq I_n^{-1}(\boldsymbol{\theta}),$$

where $I_n(\boldsymbol{\theta}) = E\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta})\right)'\right]$.

proof. Define $\mathbf{T}^* = I_n^{-1}(\boldsymbol{\theta})\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta})$. Then, by Fact 1, $E(\mathbf{T}^*) = \mathbf{0}$.

In addition,

$$Var(\mathbf{T}^*) = E(\mathbf{T}^*\mathbf{T}'^*) = I_n^{-1}(\boldsymbol{\theta}).$$

Now for any $\mathbf{a}$ with $\|\mathbf{a}\| = 1$, we have

$$\begin{aligned}
Var(\mathbf{a}'\mathbf{T}) &= E([\mathbf{a}'(\mathbf{T}-\boldsymbol{\theta})]^2) \qquad (\mathbf{T} \text{ is unbiased}) \\
&= E[\{\mathbf{a}'[(\mathbf{T}-\boldsymbol{\theta}) - \mathbf{T}^*] + \mathbf{a}'\mathbf{T}^*\}^2] \\
&\geq \mathbf{a}' I_n^{-1}(\boldsymbol{\theta})\mathbf{a} + 2\mathbf{a}' E[((\mathbf{T}-\boldsymbol{\theta}) - \mathbf{T}^*)\mathbf{T}^{*'}]\mathbf{a} \\
&= \mathbf{a}' I_n^{-1}(\boldsymbol{\theta})\mathbf{a} + 2\mathbf{a}'[E((\mathbf{T}-\boldsymbol{\theta})\mathbf{T}^{*'}) - I_n^{-1}(\boldsymbol{\theta})]\mathbf{a}.
\end{aligned}$$

Moreover, we have $E(\boldsymbol{\theta}\mathbf{T}^{*'}) = \mathbf{0}$, where $\mathbf{0}$ is the zero matrix, and

$$E(\mathbf{T}\mathbf{T}^{*'}) = E\left(\mathbf{T}\frac{1}{f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})}\left(\frac{\partial}{\partial\boldsymbol{\theta}} f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})\right)'\right) I_n^{-1}(\boldsymbol{\theta})$$

by definition of $\mathbf{T}^*$

$$
\begin{aligned}
&= \left[\int \mathbf{T} \left(\frac{\partial}{\partial \boldsymbol{\theta}} f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta})\right)' d\mathbf{y}\right] I_n^{-1}(\boldsymbol{\theta}) \\
&= \left(\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\theta}\right) I_n^{-1}(\boldsymbol{\theta}) \qquad (\mathbf{T} \text{ is unbiased}) \\
&= I_n^{-1}(\boldsymbol{\theta}) \qquad \left(\text{since } \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\theta}} = I\right).
\end{aligned}
$$

As a result,

$$
Var(\mathbf{a}'\mathbf{T}) \geq \mathbf{a}' I_n^{-1}(\boldsymbol{\theta}) \mathbf{a},
$$

yielding the desired conclusion.