PURDUE
UNIVERSITY

Lecture 8: Properties of Maximum Likelihood Estimation (MLE)

(LaTeX prepared by Haiguang Wen)
April 27, 2015

This lecture note is based on ECE 645(Spring 2015) by Prof. Stanley H. Chan in the School of Electrical and Computer Engineering at Purdue University.

1 Efficiency of MLE

Maximum Likelihood Estimation (MLE) is a widely used statistical estimation method. In this lecture, we will study its properties: efficiency, consistency and asymptotic normality.

MLE is a method for estimating parameters of a statistical model. Given the distribution of a statistical model $f(y\,;\theta)$ with unkown deterministic parameter θ, MLE is to estimate the parameter θ by maximizing the probability $f(y\,;\theta)$ with observations y.

$$\widehat{\theta}(y) = \arg\min_{\theta} f(y\,;\theta) \tag{1}$$

Please see the previous lecture note **Lecture 7** for details.

1.1 Cramér–Rao Lower Bound (CRLB)

Cramér–Rao Lower Bound (CRLB) is introduced in **Lecture 7**. Briefly, CRLB describes a lower bound on the variance of estimators of the deterministic parameter θ. That is

$$\text{Var}\left(\widehat{\theta}(Y)\right) \geqslant \frac{(\frac{\partial}{\partial\theta}\mathbb{E}[\widehat{\theta}(Y)])^2}{I(\theta)}, \tag{2}$$

where $I(\theta)$ is the *Fisher information* that measures the information carried by the observable random variable Y about the unknown parameter θ. For unbiased estimator $\widehat{\theta}(Y)$, Equation 2 can be simplified as

$$\text{Var}\left(\widehat{\theta}(Y)\right) \geqslant \frac{1}{I(\theta)}, \tag{3}$$

which means the variance of any unbiased estimator is as least as the inverse of the Fisher information.

1.2 Efficient Estimator

From section 1.1, we know that the variance of estimator $\widehat{\theta}(y)$ cannot be lower than the CRLB. So any estimator whose variance is equal to the lower bound is considered as an efficient estimator.

Definition 1. Efficient Estimator
An estimator $\widehat{\theta}(y)$ is efficient if it achieves equality in CRLB.

Example 1.
Question: $\mathbf{Y} = \{Y_1, Y_2, \cdots, Y_n\}$ are i.i.d. Gaussian random variables with distribution $N(\theta, \sigma^2)$. Determine the maximum likelihood estimator of θ. Is the estimator efficient?

Solution: Let $\boldsymbol{y} = \{y_1, y_2, \cdots, y_n\}$ be the observation, then

$$
\begin{aligned}
f(\boldsymbol{y}\,;\,\theta) &= \prod_{k=1}^{n} f(y_k\,;\,\theta) \\
&= \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_k-\theta)^2}{2\sigma^2}\right\} \\
&= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{-\frac{\sum_{k=1}^{n}(y_k-\theta)^2}{2\sigma^2}\right\}.
\end{aligned}
$$

Take the log of both sides of the above equation, we have

$$
\log f(\boldsymbol{y}\,;\,\theta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{k=1}^{n}(y_k-\theta)^2}{2\sigma^2}.
$$

Since $\log f(\boldsymbol{y}\,;\,\theta)$ is a quadratic concave function of θ, we can obtain the MLE by solving the following equation.

$$
\frac{\partial \log f(\boldsymbol{y}\,;\,\theta)}{\partial\theta} = \frac{2\sum_{k=1}^{n}(y_k-\theta)}{2\sigma^2} = 0.
$$

Therefore, the MLE is $\widehat{\theta}_{MLE}(\boldsymbol{y}) = \frac{1}{n}\sum_{k=1}^{n} y_k$.

Now let us check whether the estimator is efficient or not. It is easy to check that the MLE is an unbiased estimator ($\mathbb{E}[\widehat{\theta}_{MLE}(\boldsymbol{y})] = \theta$). To determine the CRLB, we need to calculate the Fisher information of the model.

$$
I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{y}\,;\,\theta)\right] = \frac{n}{\sigma^2} \tag{4}
$$

According to Equation 3, we have

$$
\operatorname{Var}\left(\widehat{\theta}_{MLE}(\boldsymbol{Y})\right) \geqslant \frac{1}{I(\theta)} = \frac{\sigma^2}{n}. \tag{5}
$$

And the variance of the MLE is

$$
\operatorname{Var}\left(\widehat{\theta}_{MLE}(\boldsymbol{Y})\right) = \operatorname{Var}\left(\frac{1}{n}\sum_{k=1}^{n} Y_k\right) = \frac{\sigma^2}{n}. \tag{6}
$$

So CRLB equality is achieved, thus the MLE is efficient.

1.3 Minimum Variance Unbiased Estimator (MVUE)

Recall that a *Minimum Variance Unbiased Estimator* (MVUE) is an unbiased estimator whose variance is lower than any other unbiased estimator for all possible values of parameter θ. That is

$$
\operatorname{Var}(\widehat{\theta}_{MVUE}(\boldsymbol{Y})) \leqslant \operatorname{Var}(\widehat{\theta}(\boldsymbol{Y})) \tag{7}
$$

for any unbiased $\widehat{\theta}(\boldsymbol{Y})$ of any θ.

Proposition 1. Unbiased and Efficient Estimators
If an estimator $\widehat{\theta}(y)$ is unbiased and efficient, then it must be MVUE.

Proof.
Since $\widehat{\theta}(y)$ is efficient, according to CRLB, we have

$$\mathrm{Var}\left(\widehat{\theta}(Y)\right) \leqslant \mathrm{Var}\left(\widetilde{\theta}(Y)\right) \tag{8}$$

for any $\widetilde{\theta}(Y)$. Therefore, $\widetilde{\theta}(Y)$ must be minimum variance (MV). Since $\widetilde{\theta}(Y)$ is also unbiased, it is a MVUE.

□

Remark: The converse of the proposition is not true in general. That is, MVUE does NOT need to be efficient. Here is a counter example.

Example 2. Counter Example
Suppose that $\boldsymbol{y} = \{Y_1, Y_2, \cdots, Y_n\}$ are i.i.d. exponential random variables with unknown mean $\frac{1}{\theta}$. Find the MLE and MVUE of θ. Are these estimators efficient?

Solution: Let $\boldsymbol{y} = \{y_1, y_2, \cdots, y_n\}$ be the observation, then

$$\begin{aligned} f(\boldsymbol{y}\,;\,\theta) &= \prod_{k=1}^{n} f(y_k\,;\,\theta) \\ &= \prod_{k=1}^{n} \theta \exp\{-\theta y_k\} \\ &= \theta^n \exp\left\{-\theta \sum_{k=1}^{n} y_k\right\}. \end{aligned} \tag{9}$$

Take the log of both sides of the above equation, we have

$$\log f(\boldsymbol{y}\,;\,\theta) = n\log(\theta) - \theta \sum_{k=1}^{n} y_k.$$

Since $\log f(\boldsymbol{y}\,;\,\theta)$ is a concave function of θ, we can obtain the MLE by solving the following equation.

$$\frac{\partial \log f(\boldsymbol{y}\,;\,\theta)}{\partial \theta} = \frac{n}{\theta} - \sum_{k=1}^{n} = 0$$

So the MLE is

$$\widehat{\theta}_{MLE}(\boldsymbol{y}) = \frac{n}{\sum_{k=1}^{n} y_k}. \tag{10}$$

To calculate the CRLB, we need to calculate $\mathbb{E}\left[\widehat{\theta}_{MLE}(\boldsymbol{Y})\right]$ and $\mathrm{Var}\left(\widehat{\theta}_{MLE}(\boldsymbol{Y})\right)$. Let $T(\boldsymbol{y}) = \sum_{k=1}^{n} y_k$, then by moment generating function, we can show that the distribution of $T(\boldsymbol{y})$ is the Erlange distribution:

$$f_T(t) = \frac{\theta^n t^{n-1}}{(n-1)!} e^{-\theta t}. \tag{11}$$

So we have

$$\begin{aligned} \mathbb{E}\left[\widehat{\theta}_{MLE}\left(T(\boldsymbol{Y})\right)\right] &= \int_0^\infty \frac{n}{t} \frac{\theta^n t^{n-1}}{(n-1)!} e^{-\theta t} dt \\ &= \frac{n\theta}{n-1} \int_0^\infty \frac{(\theta t)^{n-2}}{(n-2)!} e^{-\theta t} d\theta t \\ &= \frac{n}{n-1}\theta. \end{aligned} \tag{12}$$

Therefore the MLE is a biased estimator of θ.

Similarly, we can calculate the variance of MLE as follows.

$$\begin{aligned}\operatorname{Var}\left(\widehat{\theta}_{MLE}\left(T(\boldsymbol{Y})\right)\right) &= \mathbb{E}\left[\widehat{\theta}^2_{MLE}\left(T(\boldsymbol{Y})\right)\right] - \mathbb{E}\left[\widehat{\theta}_{MLE}\left(T(\boldsymbol{Y})\right)\right]^2 \\ &= \frac{\theta^2 n^2}{(n-1)^2(n-2)}\end{aligned}$$

The Fisher information is

$$I(\theta) = -\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{y}|\theta) = \frac{n}{\theta^2}.$$

So the CRLB is

$$\begin{aligned}\operatorname{Var}\left(\widehat{\theta}_{MLE}\left(T(\boldsymbol{Y})\right)\right) &\geqslant \frac{\left(\frac{\partial}{\partial\theta}\mathbb{E}\left[\widehat{\theta}_{MLE}\left(T(\boldsymbol{Y})\right)\right]\right)^2}{I(\theta)} \\ &= \frac{n^2}{(n-1)^2}\Big/\frac{n}{\theta^2} \\ &= \frac{n}{(n-1)^2}\theta^2.\end{aligned}$$

The CRLB equality does **NOT** hold, so $\widehat{\theta}_{MLE}$ is not efficient.

The distribution in Equation 9 belongs to exponential family and $T(\boldsymbol{y}) = \sum_{k=1}^n y_k$ is a complete sufficient statistic. So the MLE can be expressed as $\widehat{\theta}_{MLE}(T(\boldsymbol{y})) = \frac{n}{T(\boldsymbol{y})}$, which is a function of $T(\boldsymbol{y})$. However, the MLE is a biased estimator (Equation 12). But we can construct an unbiased estimator based on the MLE. That is

$$\begin{aligned}\widetilde{\theta}(T(\boldsymbol{y})) &= \frac{n-1}{n}\widehat{\theta}_{MLE}(T(\boldsymbol{y})) \\ &= \frac{n-1}{T(\boldsymbol{y})}.\end{aligned}$$

It is easy to check $\mathbb{E}\left[\widetilde{\theta}\left(T(\boldsymbol{Y})\right)\right] = \mathbb{E}\left[\frac{n-1}{n}\widehat{\theta}_{MLE}\left(T(\boldsymbol{Y})\right)\right] = \frac{n-1}{n}\frac{n}{n-1}\theta = \theta$. Since $\widetilde{\theta}(T(\boldsymbol{y}))$ is an unbiased estimator and it is a function of complete sufficient statistic, $\widetilde{\theta}(T(\boldsymbol{y}))$ is MVUE. So

$$\widehat{\theta}_{MVUE}\left(T(\boldsymbol{y})\right) = \frac{n-1}{T(\boldsymbol{y})}. \tag{13}$$

The variance of MVUE is

$$\begin{aligned}\operatorname{Var}\left(\widehat{\theta}_{MVUE}\left(T(\boldsymbol{Y})\right)\right) &= \operatorname{Var}\left(\frac{n-1}{n}\widehat{\theta}_{MLE}\left(T(\boldsymbol{Y})\right)\right) \\ &= \frac{(n-1)^2}{n^2}\frac{n^2\theta^2}{(n-1)^2(n-2)} \\ &= \frac{\theta^2}{n-2}.\end{aligned}$$

So the CRLB is

$$\operatorname{Var}\left(\widehat{\theta}_{MVUE}\left(T(\boldsymbol{Y})\right)\right) \geqslant \frac{1}{I(\theta)} = \frac{\theta^2}{n}.$$

Therefore, the MVUE is **NOT** an efficient estimator.

2 Consistency of MLE

Definition 2. CONSISTENCY
Let $\{Y_1, \cdots, Y_n\}$ be a sequence of observations. Let $\widehat{\theta}_n$ be the estimator using $\{Y_1, \cdots, Y_n\}$. We say that $\widehat{\theta}_n$ is consistent if $\widehat{\theta}_n \xrightarrow{p} \theta$, i.e.,

$$\mathbb{P}\left(|\widehat{\theta}_n - \theta| > \varepsilon\right) \to 0, \text{ as } n \to \infty \tag{14}$$

Remark: A sufficient condition to have Equation 14 is that

$$\mathbb{E}\left[\left(\widehat{\theta}_n - \theta\right)^2\right] \to 0, \quad \text{as } n \to \infty. \tag{15}$$

Proof.
According to Chebyshev's inequality, we have

$$\mathbb{P}\left(|\widehat{\theta}_n - \theta| \geq \varepsilon\right) \leq \frac{\mathbb{E}\left[(\widehat{\theta}_n - \theta)^2\right]}{\varepsilon^2} \tag{16}$$

Since $\mathbb{E}\left[\left(\widehat{\theta}_n - \theta\right)^2\right] \to 0$, the we have

$$0 \leq \mathbb{P}\left(|\widehat{\theta}_n - \theta| \geq \varepsilon\right) \leq \frac{\mathbb{E}\left[(\widehat{\theta}_n - \theta)^2\right]}{\varepsilon^2} \to 0.$$

Therefore, $\mathbb{P}\left(|\widehat{\theta}_n - \theta| > \varepsilon\right) \to 0, \text{ as } n \to \infty.$

□

Example 3.
$\{Y_1, Y_2, \cdots, Y_n\}$ are i.i.d. Gaussian random variables with distribution $N(\theta, \sigma^2)$. Is the MLE using $\{Y_1, Y_2, \cdots, Y_n\}$ consistent?

Solution:
From Example 1., we know that the MLE is

$$\widehat{\theta}_n = \frac{1}{n}\sum_{k=1}^{n} Y_k.$$

Since

$$\mathbb{E}\left[\left(\widehat{\theta}_n - \theta\right)^2\right] = \text{Var}\left(\widehat{\theta}_n\right) = \frac{\sigma^2}{n}, \text{ (see Equation 6)},$$

so $\mathbb{E}\left[\left(\widehat{\theta}_n - \theta\right)^2\right] \to 0$. Therefore $\widehat{\theta}_n \xrightarrow{p} \theta$, i.e. $\widehat{\theta}_n$ is consistent.

In fact, the result of the example above it holds for any distribution. The following proposition states this result:

Proposition 2.
(*MLE of i.i.d observation is consistent*) Let $\{Y_1, \cdots, Y_n\}$ be a sequence of i.i.d. observations where

$$Y_k \overset{iid}{\sim} f_\theta(y).$$

Then the MLE of θ is consistent.

Proof.
(This proof is partially correct. See Levy Chapter 4.5 for complete discussion.)
The MLE of θ is

$$\begin{aligned}
\widehat{\theta}_n &= \arg\max_{\theta} \prod_{k=1}^{n} f_\theta(y_k) \\
&= \arg\max_{\theta}\ \log\left(\prod_{k=1}^{n} f_\theta(y_k)\right) \\
&= \arg\max_{\theta}\ \frac{1}{n}\sum_{k=1}^{n} \log f_\theta(y_k) \\
&= \arg\max_{\theta}\ \varphi_n(\theta),
\end{aligned}$$

where $\varphi_n(\theta) = \frac{1}{n}\sum_{k=1}^{n} \log f_\theta(y_k)$. Let $\ell_{\widehat{\theta}}(y_k) = \log \frac{f_\theta(y_k)}{f_{\widehat{\theta}}(y_k)}$, then we have

$$\begin{aligned}
\mathbb{E}_{\widehat{\theta}}\left[\ell_{\widehat{\theta}}(Y_k)\right] &\overset{def}{=} \int \log \frac{f_\theta(y_k)}{f_{\widehat{\theta}}(y_k)} f_{\widehat{\theta}}(y_k) dy_k \\
&= D\left(f_\theta \| f_{\widehat{\theta}}\right).
\end{aligned}$$

According to the weak law of large numbers (WLLN), we have

$$\frac{1}{n}\sum_{k=1}^{n} \ell_{\widehat{\theta}}(y_k) \xrightarrow{p} D\left(f_\theta \| f_{\widehat{\theta}}\right). \tag{17}$$

Since $\widehat{\theta}_n$ is the MLE which maximizes $\varphi_n(\theta)$, then

$$\begin{aligned}
0 &\geq \varphi_n(\theta) - \varphi_n(\widehat{\theta}) \\
&= \frac{1}{n}\sum_{k=1}^{n} \log f_\theta(y_k) - \frac{1}{n}\sum_{k=1}^{n} \log f_{\widehat{\theta}}(y_k) \\
&= \frac{1}{n}\sum_{k=1}^{n} \log \frac{f_\theta(y_k)}{f_{\widehat{\theta}}(y_k)} \\
&= \frac{1}{n}\sum_{k=1}^{n} \ell_{\widehat{\theta}}(y_k) \\
&= \frac{1}{n}\sum_{k=1}^{n} \ell_{\widehat{\theta}}(y_k) - D\left(f_\theta \| f_{\widehat{\theta}}\right) + D\left(f_\theta \| f_{\widehat{\theta}}\right).
\end{aligned}$$

Therefore,

$$D\left(f_\theta \| f_{\widehat{\theta}}\right) \leq \left| \frac{1}{n}\sum_{k=1}^{n} \ell_{\widehat{\theta}}(y_k) - D\left(f_\theta \| f_{\widehat{\theta}}\right) \right|.$$

By Equation 17, we have

$$0 \leq D\left(f_\theta \| f_{\widehat{\theta}}\right) \leq \left| \frac{1}{n}\sum_{k=1}^{n} \ell_{\widehat{\theta}}(y_k) - D\left(f_\theta \| f_{\widehat{\theta}}\right) \right| \xrightarrow{p} 0.$$

So we must have $D\left(f_\theta \| f_{\widehat{\theta}}\right) \xrightarrow{p} 0$, and then $\widehat{\theta} \xrightarrow{p} \theta$.

$\square$

3 Asymptotic Normality of MLE

The previous proposition only asserts that MLE of i.i.d. observations is consistent. However, it provides no information about the distribution of the MLE.

Proposition 3. ASYMPTOTIC NORMALITY
Let $\{Y_1, \cdots, Y_n\}$ be a sequence of i.i.d. observations where

$$Y_k \overset{iid}{\sim} f_\theta(y)$$

$\widehat{\theta}$ is the MLE of θ, then

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N\left(0, \frac{1}{I(\theta)}\right).$$

Proof.
See Lehmann, "Elements of Large Sample Theory", Springer, 1999 for proof.

□

Example 4.
$\{Y_1, Y_2, \cdots, Y_n\}$ are i.i.d. Gaussian random variables with distribution $N(\theta, \sigma^2)$. Find the asumptotic distribution of $\widehat{\theta}_{ML}$

Solution:
Similar to Example 2, we can calculate the Fisher information of θ,

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log f_\theta(Y)\right] = \frac{1}{\sigma^2}$$

We know that $\widehat{\theta}_{ML} = \frac{1}{n}\sum_{k=1}^n y_k$. So if $\widehat{\theta}_n = \frac{1}{n}\sum_{k=1}^n y_k$, then we have

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N\left(0, \sigma^2\right).$$
