Maximum Likelihood Estimation

(LaTeX prepared by Shaobo Fang)
April 14, 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 Introduction

For many families besides exponential family, Minimum Variance Unbiased Estimator (MVUE) could be very difficult to find, or it may not even exist. For such models, we need an alternative method to obtain good estimators. With the absence of the prior information, the maximum likelihood estimation might be a viable alternative. (Poor IV.D)

Definition 1. Maximum likelihood estimate (MLE)
The maximum likelihood estimator is defined as:

$$\widehat{\theta}_{ML}(y) \stackrel{\text{def}}{=} \underset{\theta}{\text{argmax}} \;\; f_\theta(y) \tag{1}$$

where $f_\theta(y) = f_Y(y;\theta)$.

Here, the function $f_\theta(y)$ is called the **likelihood function**. We can also take log on $f_\theta(y)$ and yield the same maximizer:

$$\widehat{\theta}_{ML}(y) \stackrel{\text{def}}{=} \underset{\theta}{\text{argmax}} \;\; \log f_\theta(y). \tag{2}$$

The function $\log f_\theta(y)$ is called the **log-likelihood function**.

Example 1.
Let $\boldsymbol{Y} = [Y_1, \ldots, Y_n]$ be a sequence of iid random variables such that

$$Y_k \sim \mathcal{N}(\mu, \sigma^2).$$

Assume that σ^2 is known, find $\widehat{\theta}_{ML}$ for μ.

Solution:
First of all, the likelihood function is

$$f_\theta(\boldsymbol{y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(y_k-\mu)^2\right).$$

Taking the log on both sides we have the log-likelihood function

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

In order to find the maximizer of the log-likelihood function, we take the first order derivative and set it to zero. This yields

$$\widehat{\mu}_{ML}(\boldsymbol{y}) = \frac{1}{n}\sum_{k=1}^{n} y_k.$$

We can also show that

$$\mathbb{E}[\widehat{\mu}_{ML}(\boldsymbol{Y})] = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[Y_k] = \mu,$$

which says that the estimator is **unbiased**.

Example 2.
Now we consider the previous example with both μ and σ unknown. Our goal is to determine $\widehat{\boldsymbol{\theta}}_{ML} \stackrel{\text{def}}{=} [\widehat{\theta}_1, \widehat{\theta}_2]^T$ for $\theta_1 = \mu$ and $\theta_2 = \sigma^2$.

Solution:
Same as the previous problem, the log-likelihood function is

$$\log f_{\boldsymbol{\theta}}(\boldsymbol{y}) = -\frac{1}{2\theta_2}\sum_{k=1}^{n}(y_k - \theta_1)^2 - \frac{n}{2}\log(2\pi\theta_2).$$

Taking the partial derivative wrt to θ_1 yields

$$\frac{\partial}{\partial\theta_1}\log f_{\boldsymbol{\theta}}(\boldsymbol{y}) = \frac{1}{2\theta_2}\sum_{k=1}^{n}2(y_k - \theta_1) = 0,$$

which gives

$$\widehat{\theta}_1(\boldsymbol{y}) = \frac{1}{n}\sum_{k=1}^{n}y_k.$$

Similarly, taking the partial derivative wrt to θ_2 yields

$$\frac{\partial}{\partial\theta_2}\log f_{\boldsymbol{\theta}}(\boldsymbol{y}) = \frac{1}{\theta_2^2}\sum_{k=1}^{n}2(y_k - \widehat{\theta}_1)^2 - \frac{n}{2\theta_2} = 0,$$

which gives

$$\widehat{\theta}_2(\boldsymbol{y}) = \frac{1}{n}\sum_{k=1}^{n}(y_k - \widehat{\theta}_1)^2.$$

Note that $\mathbb{E}[\widehat{\theta}_2(\boldsymbol{Y})] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$. So $\widehat{\theta}_2$ is **biased**.

Remark: In order to obtain an unbiased estimator for the population variance, it is preferred to use the sample variance define as (Zwillinger 1995, p. 603):

$$S_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2,$$

where $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n}Y_i$ is the sample mean. In fact, the function `var` in MATLAB is the sample variance.

Example 3. BERNOULLI
(Statistical Inference: Example 7.2.7, Casella and Berger)

Let $\boldsymbol{Y} = [Y_1, \ldots, Y_n]$ be a sequence of i.i.d. Bernoulli random variables of parameter θ. We would like to find the MLE $\widehat{\theta}_{ML}$ for θ.

Solution:

First of all, we define the likelihood function:

$$f_\theta(\boldsymbol{y}) = \prod_{k=1}^{n} \theta^{y_k}(1-\theta)^{1-y_k}.$$

Letting $y = \sum_{k=1}^{n} y_k$, we can rewrite the likelihood function as

$$f_\theta(\boldsymbol{y}) = \theta^{y}(1-\theta)^{1-y}.$$

Hence, the log-likelihood function is

$$\log f_\theta(\boldsymbol{y}) = y \log\theta + (n-y)\log(1-\theta).$$

Taking the derivative and setting it to zero yields

$$\frac{\partial}{\partial\theta} \log f_\theta(\boldsymbol{y}) = \frac{y}{\theta} - \frac{n-y}{1-\theta} = 0.$$

Therefore, $\widehat{\theta}_{ML}(\boldsymbol{y})$ is

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

Example 4. Binomial
Let $\boldsymbol{Y} = [Y_1, \ldots, Y_n]$ be a sequence of iid random variables of a Binomial distribution of binomial(k, θ). We would like to find $\widehat{\theta}_{ML}$ for θ.

Solution:
The likelihood function is

$$f_\theta(\boldsymbol{y}) = \prod_{i=1}^{n} \binom{k}{y_i} \theta^{y_i}(1-\theta)^{k-y_i}.$$

By letting $y = \sum_{i=1}^{n} y_i$, we can rewrite the likelihood function as:

$$f_\theta(\boldsymbol{y}) = \theta^{y}(1-\theta)^{1-y} \prod_{i=1}^{n} \binom{k}{y_i}.$$

The log-likelihood function is

$$\log f_\theta(\boldsymbol{y}) = y\log\theta + (k-y)\log(1-\theta) + \underbrace{\sum_{i=1}^{n} \log\binom{k}{y_i}}_{\text{This term does not contain } \theta}$$

Taking the first order derivative and setting to zero yields

$$\widehat{\theta}_{ML}(\boldsymbol{y}) = \frac{y}{k} = \frac{1}{k}\sum_{i=1}^{n} y_i.$$

Example 5. Poisson
Let $\boldsymbol{Y} = [Y_1, \ldots, Y_n]$ be a sequence of i.i.d. Poisson random variables of parameter λ. Recall that Poisson distribution is: $\mathbb{P}(Y_i = y_i) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$. We would like to find $\widehat{\lambda}_{ML}$ for parameter λ.

Solution:
Similarly as in previous examples, first we find the likelihood function:

$$\begin{aligned} f_\lambda(\boldsymbol{y}) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \\ &= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n y_i!}. \end{aligned}$$

Thus the log-likelihood function is

$$\log f_\lambda(\boldsymbol{y}) = -n\lambda + \sum_{i=1}^{n} y_i \log\theta - \underbrace{\log \prod_{i=1}^{n} y_i!}_{\text{This term does not contain } \lambda}$$

Setting the first-order derivative to 0 yields

$$\widehat{\lambda}_{ML}(\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

---

2 Bias v.s. Variance

In general, MLE could be both biased or unbiased. To take a closer look at this property, we write the MLE as a sum of bias and variance terms as below:

$$\begin{aligned} \text{MSE}_\theta &= \mathbb{E}_{\boldsymbol{Y}}[(\widehat{\theta}_{ML}(\boldsymbol{Y}) - \theta)^2] \\ &= \mathbb{E}[(\widehat{\theta}_{ML} - \mathbb{E}[\widehat{\theta}_{ML}] + \mathbb{E}[\widehat{\theta}_{ML}] - \theta)^2] \\ &= \mathbb{E}[(\widehat{\theta}_{ML} - \mathbb{E}[\widehat{\theta}_{ML}])^2] + \mathbb{E}[(\mathbb{E}[\widehat{\theta}_{ML}] - \theta)^2] + 2\mathbb{E}[(\widehat{\theta}_{ML} - \mathbb{E}[\widehat{\theta}_{ML}])(\mathbb{E}[\widehat{\theta}_{ML}] - \theta)] \\ &= \underbrace{\mathbb{E}_{\boldsymbol{Y}}[(\widehat{\theta}_{ML}(\boldsymbol{Y}) - \mathbb{E}[\widehat{\theta}_{ML}(\boldsymbol{Y})])^2]}_{\text{variance}} + \underbrace{(\mathbb{E}[\widehat{\theta}_{ML}(\boldsymbol{Y})] - \theta)^2}_{\text{bias}}. \end{aligned} \tag{3}$$

Example 6. Image Denoising
Let $\boldsymbol{z}$ be a clean signal and let $\boldsymbol{n}$ be a noise vector such that $\boldsymbol{n} \sim \mathcal{N}(0, \sigma^2 I)$. Suppose that we are given the noisy observation

$$\boldsymbol{y} = \boldsymbol{z} + \boldsymbol{n},$$

our goal is to estimate $\boldsymbol{z}$ from $\boldsymbol{y}$.

In this example, let us consider a *linear* denoising method. We would like to find a $\boldsymbol{W}$ such that the estimator

$$\widehat{\boldsymbol{z}} = \boldsymbol{W}\boldsymbol{y}$$

would be optimal in some sense. We shall call $\boldsymbol{W}$ as a smoothing filter.

To determine what $\boldsymbol{W}$ would be good, we first consider the MSE:

$$\begin{aligned} \text{MSE} &= \mathbb{E}[\|\widehat{\boldsymbol{z}} - \boldsymbol{z}\|^2] \\ &= \mathbb{E}[\|\boldsymbol{W}\boldsymbol{y} - \boldsymbol{z}\|^2] \\ &= \mathbb{E}[\|\boldsymbol{W}(\boldsymbol{z} + \boldsymbol{n}) - \boldsymbol{z}\|^2] \\ &= \mathbb{E}[\|(\boldsymbol{W} - \boldsymbol{I})\boldsymbol{z} + \boldsymbol{W}\boldsymbol{n}\|^2] \\ &= \underbrace{\|(\boldsymbol{W} - \boldsymbol{I})\boldsymbol{z}\|^2}_{\text{bias}} + \underbrace{\mathbb{E}[\|\boldsymbol{W}\boldsymbol{n}\|^2]}_{\text{variance}} \end{aligned}$$

Now, by using eigen-decompostion we can write $\boldsymbol{W}$ as $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$. Then, the bias can be computed as

$$\begin{aligned}
\text{bias} &= \|(\boldsymbol{W} - \boldsymbol{I})\boldsymbol{z}\|^2 \\
&= \mathbb{E}[\|\widehat{\boldsymbol{z}} - \boldsymbol{z}\|^2] \\
&= \|(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T - \boldsymbol{I})\boldsymbol{z}\|^2 \\
&= \|\boldsymbol{U}(\boldsymbol{\Lambda} - \boldsymbol{I})\boldsymbol{U}^T\boldsymbol{z}\|^2 \\
&= \boldsymbol{z}^T\boldsymbol{U}(\boldsymbol{\Lambda} - \boldsymbol{I})^2\boldsymbol{U}^T\boldsymbol{z} \\
&= \sum_{i=1}^{n}(\lambda_i - 1)^2 v_i^2,
\end{aligned}$$

where $\boldsymbol{v} = \boldsymbol{U}^T\boldsymbol{z}$. Similarly, the variance can be computed as

$$\begin{aligned}
\text{variance} &= \mathbb{E}[\|\boldsymbol{W}\boldsymbol{n}\|^2] \\
&= \mathbb{E}[\boldsymbol{n}^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{n}] \\
&= \sigma^2 \,\text{Tr}\left\{\boldsymbol{W}^T\boldsymbol{W}\right\} \\
&= \sigma^2 \sum_{i=1}^{n}\lambda_i^2
\end{aligned}$$

Therefore, the MSE can be written as:

$$\text{MSE} = \sum_{i=1}^{n}(\lambda_i - 1)^2 v_i^2 + \sigma^2 \sum_{i=1}^{n}\lambda_i^2$$

To minimize MSE, λ_i should be chosen such that

$$\frac{\partial}{\partial \lambda_i}\text{MSE} = 2v_i^2(\lambda_i - 1) + 2\sigma^2\lambda_i = 0,$$

which is

$$\widehat{\lambda}_i = \frac{v_i^2}{v_i^2 + \sigma^2}.$$

Thus far we have come across many examples where the estimators are unbiased. So are biased estimators bad? The answer is no. Here is an example.

Let us consider a random variable $Y \sim \mathcal{N}(0, \sigma^2)$. Now, consider the following two estimators:

- Estimator 1: $\widehat{\theta}_1(Y) = Y^2$. Then $\mathbb{E}[\widehat{\theta}_1(Y)] = \mathbb{E}[Y^2] = \sigma^2$, thus it is unbiased.
- Estimator 2: $\widehat{\theta}_2(Y) = aY^2, a \neq 1$, then $\mathbb{E}[\widehat{\theta}_2(Y)] = a\sigma^2$. Thus it is biased.

Let us now consider the MSE of θ_2. (Note that the MSE of θ_1 can be found by letting $a = 1$.)

$$\begin{aligned}
\text{MSE} &= \mathbb{E}[(\widehat{\theta}_2(Y) - \sigma^2)^2] \\
&= \mathbb{E}[(aY^2 - \sigma^2)^2] \\
&= \mathbb{E}[a^2Y^4] - 2\sigma^2\mathbb{E}[aY^2] + \sigma^4 \\
&= 3a^2\sigma^4 - 2a\sigma^4 + \sigma^4 \\
&= \sigma^4(3a^2 - 2a + 1)
\end{aligned}$$

Therefore, the MSE attains its minimum at:

$$\frac{\partial}{\partial a}\text{MSE} = \sigma^4(6a - 2) = 0,$$

which is

$$a = \frac{1}{3}.$$

This result says: although θ_2 is biased, it actually attains a lower MSE!

3 Fisher Information

3.1 Variance and Curvature of log-likelihood

For unbiased estimators, the variance can provide extremely important information about the performance of the estimators. In order to study the variance more carefully, we first study its relationship with regard to the log-likelihood as demonstrated in the example below.

Example 7.
Let $Y \sim \mathcal{N}(\theta, \sigma^2)$, where σ is known.
Accordingly,

$$f_\theta(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\theta)^2}{2\sigma^2}\right)$$

$$\log f_\theta(y) = -\log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}(y-\theta)^2$$

$$\Rightarrow \frac{\partial \log f_\theta(y)}{\partial\theta} = \frac{1}{\sigma^2}(y-\theta)$$

$$\Rightarrow \underbrace{\frac{\partial^2 \log f_\theta(y)}{\partial\theta^2}}_{\text{curvature of log-likelihood}} = -\frac{1}{\sigma^2}$$

Therefore, as σ^2 increases, we can easily conclude that $-\frac{\partial^2}{\partial\theta^2}\log f_\theta(y)$ will decrease. Thus, we conclude that with the variance increasing, the curvature will be decreasing.

3.2 Fisher-Information

Definition 2. Fisher Information
The Fisher-information is defined as:

$$I(\theta) \stackrel{\text{def}}{=} -\mathbb{E}_{\boldsymbol{Y}}\left[\frac{\partial^2 \log f_\theta(\boldsymbol{Y})}{\partial\theta^2}\right], \tag{4}$$

where

$$\mathbb{E}_{\boldsymbol{Y}}\left[\frac{\partial^2 \log f_\theta(\boldsymbol{Y})}{\partial\theta^2}\right] = \int \frac{\partial^2 \log f_\theta(\boldsymbol{y})}{\partial\theta^2} f_\theta(\boldsymbol{y}) d\boldsymbol{y} \tag{5}$$

We will try to estimate the fisher information in the examples below.

Example 8.
Let $\boldsymbol{Y} = [Y_1, \ldots, Y_n]$ be a sequence of iid random variables such that $Y_i \sim \mathcal{N}(\theta, \sigma^2)$. We would like to determine $I(\theta)$.

First, we know that the log-likelihood is

$$\log f_\theta(\boldsymbol{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n}\frac{(y_i-\theta)}{2\sigma^2}.$$

The first order derivative is

$$\frac{\partial \log f_\theta(\boldsymbol{y})}{\partial\theta} = \frac{n}{\sigma^2}(\overline{y}-\theta),$$

where

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Consequently, the second order derivative is

$$\frac{\partial^2 \log f_\theta(\boldsymbol{y})}{\partial \theta^2} = \frac{-n}{\sigma^2}.$$

Finally, the fisher information is

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

Example 9.
Let $\boldsymbol{Y} = [Y_1, \ldots, Y_n]$ be a sequence of iid random variables such that

$$Y_k = A\cos(w_0 k + \theta) + N_k,$$

where $N_k \sim \mathcal{N}(0, \sigma^2)$. Find $I(\theta)$.
The likelihood function is

$$\begin{aligned} f_\theta(\boldsymbol{y}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y_i - A\cos(w_0 k + \theta))^2}{2\sigma^2}) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}^{n/2}} \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - A\cos(w_0 k + \theta))^2) \end{aligned}$$

Then, the first order derivative of the log-likelihood is

$$\begin{aligned} \frac{\partial}{\partial \theta}\log f_\theta(\boldsymbol{y}) &= \frac{\partial}{\partial \theta}\left[\frac{-1}{\sigma^2}\sum_{i=1}^{n}(y_i - A\cos(w_0 k + \theta))^2\right] \\ &= \frac{-1}{\sigma^2}\sum_{i=1}^{n}(y_i - A\cos(w_0 k + \theta))(A\sin(w_0 k + \theta)) \\ &= \frac{-A}{\sigma^2}\sum_{i=1}^{n}(y_i \sin(w_0 k + \theta) - \frac{A}{2}\sin(2w_0 k + 2\theta)) \end{aligned}$$

The second order derivative is

$$\frac{\partial^2}{\partial \theta^2}\log f_\theta(\boldsymbol{y}) = -\frac{A}{\sigma^2}\sum_{i=1}^{n}[y_i \cos(w_0 k + \theta) - A\cos(2w_0 k + 2\theta)]$$

Accordingly, the $\mathbb{E}_{\boldsymbol{Y}}[\frac{\partial^2}{\partial \theta^2}\log f_\theta(\boldsymbol{Y})]$ can be estimated as below:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{Y}}\left[\frac{\partial^2}{\partial \theta^2}\log f_\theta(\boldsymbol{Y})\right] &= \frac{-A}{\sigma^2}\sum_{i=1}^{n}(\mathbb{E}[Y_i]\cos(w_0 k + \theta) - A\cos(2w_0 k + 2\theta)) \\ &= \frac{-A}{2\sigma^2}\sum_{i=1}^{n}[A\cos^2(w_0 k + \theta) - A\cos(2w_0 k + 2\theta)] \\ &= -\frac{A^2}{\sigma^2}\sum_{i=1}^{n}(\frac{1}{2} + \frac{1}{2}\cos(2w_0 k + 2\theta) - \cos(2w_0 k + 2\theta)) \\ &= -\frac{nA^2}{2\sigma^2} + \frac{A^2}{2\sigma^2}\frac{1}{n}\sum_{k=1}^{n}\cos(2w_0 k + 2\theta) \end{aligned}$$

By using the fact that $\frac{1}{n}\sum_{k=1}^{n}\cos(2w_0k+2\theta)\approx 0$, we have:

$$I(\theta)=\frac{nA^2}{2\sigma^2}.$$

3.3 Fisher-Information and KL Divergence

There is an interesting relationship between the Fisher-Information and the KL divergence, which we shall now discuss. To begin with, let us first list out two assumptions.
Assumption:

1. $$\frac{\partial}{\partial\theta}\int f_\theta(y)dy=\int\frac{\partial}{\partial\theta}f_\theta(y)dy$$

2. $$\frac{\partial}{\partial\theta}\int\widehat{\theta}(y)f_\theta(y)dy=\int\frac{\partial}{\partial\theta}f_\theta(y)\widehat{\theta}(y)dy$$

Basically, the two assumptions say that we can interchange the order of integration and the differentiation.

If the assumption holds, we can show the following result:

$$I(\theta)=\mathbb{E}_Y\left[\left(\frac{\partial\log f_\theta(Y)}{\partial\theta}\right)^2\right]. \tag{6}$$

Proof.
By the assumptions and integration by part, we have

$$\begin{aligned}I(\theta)&=\mathbb{E}_Y\left[-\frac{\partial^2\log f_\theta(Y)}{\partial\theta^2}\right]=-\int\left(\frac{f''_\theta(y)f_\theta(y)-(f'_\theta(y))^2}{f^2_\theta(y)}\right)f_\theta(y)dy\\&=-\underbrace{\int f''_\theta(y)dy}_{=0\text{ by (1)}}+\int\frac{1}{f_\theta(y)}(f'_\theta(y))^2dy\\&=\int f_\theta(y)\left(\frac{\partial\log f_\theta(y)}{\partial\theta}\right)^2dy=\mathbb{E}_Y\left[\left(\frac{\partial\log f_\theta(y)}{\partial\theta}\right)^2\right].\end{aligned}$$

□

The following proposition links KL divergence and $I(\theta)$.

Proposition 1.
Let $\theta=\theta_0+\delta$ for some small deviation δ, then

$$D(f_{\theta_0}\|f_\theta)\approx\frac{I(\theta_0)}{2}(\theta-\theta_0)^2+\mathcal{O}(\theta-\theta_0)^3. \tag{7}$$

Interpretation: If $I(\theta)$ is large, then $D(f_{\theta_0}\|f_\theta)$ is large. Accordingly, it would be easier to differentiate θ_0 and θ.

Proof.
First, recall that the KL divergence is defined as

$$D(f_{\theta_0}\|f_\theta)=\int f_{\theta_0}(y)\log\frac{f_{\theta_0}(y)}{f_\theta(y)}dy.$$

Consider Taylor expansion on θ_0, we estimate the first two terms as below.

First-order derivative:

$$\begin{aligned}\frac{\partial}{\partial\theta}D(f_{\theta_0}\|f_\theta)\Big|_{\theta=\theta_0} &= \int f_{\theta_0}(y)\frac{\partial}{\partial\theta}\left[\log f_{\theta_0}(y)-\log f_\theta(y)\right]\Big|_{\theta=\theta_0}dy \\ &= \int f_{\theta_0}(y)\left[\frac{-1}{f_\theta(y)\frac{\partial}{\partial\theta}f_\theta(y)}\right]\Big|_{\theta=\theta_0}dy \\ &= -\int\frac{\partial}{\partial\theta}f_\theta(y)dy \\ &= -\frac{\partial}{\partial\theta}\int f_\theta(y)dy = 0.\end{aligned} \tag{8}$$

Second-order derivative:

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}D(f_{\theta_0}\|f_\theta)|_{\theta=\theta_0} &= \int f_{\theta_0}(y)\frac{\partial^2}{\partial\theta^2}\left[-\log f_\theta(y)\right]dy \\ &= \mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2}\log f_\theta(y)\right]\Big|_{\theta=\theta_0} \\ &= I(\theta_0).\end{aligned} \tag{9}$$

Substitute the above terms into Taylor expansion and ignore the higher order terms:

$$\begin{aligned}D(f_{\theta_0}\|f_\theta) &= D(f_{\theta_0}\|f_{\theta_0})+(\theta-\theta_0)\frac{\partial}{\partial\theta}D(f_{\theta_0}\|f_\theta)+\frac{(\theta-\theta_0)^2}{2}\frac{\partial^2}{\partial\theta^2}D(f_{\theta_0}\|f_\theta)+\mathcal{O}(\theta-\theta)^3 \\ &= \frac{(\theta-\theta_0)^2}{2}I(\theta_0)+\mathcal{O}(\theta-\theta_0)^3.\end{aligned} \tag{10}$$

□

4 Cramer-Rao Lower Bound (CRLB) Theorem

The CRLB is a fundamental result that characterizes the performance of an estimator.

Theorem 1.
Under the assumptions (1) and (2):

$$\mathrm{Var}(\widehat{\theta}(Y)) \geq \frac{(\frac{\partial}{\partial\theta}\mathbb{E}[\widehat{\theta}(Y)])^2}{I(\theta)} \tag{11}$$

for any estimator $\widehat{\theta}(Y)$.

Proof.
To prove the inequality, we first note that

$$\mathrm{Var}(\widehat{\theta}(Y))I(\theta) = \int\left(\widehat{\theta}(y)-\mathbb{E}[\widehat{\theta}(y)]\right)^2 f_\theta(y)dy\cdot\int\left(\frac{\partial}{\partial\theta}\log f_\theta(y)\right)^2 f_\theta(y)dy.$$

Letting

$$\begin{aligned}A &= \widehat{\theta}(y)-\mathbb{E}[\widehat{\theta}(y)], \\ B &= \frac{\partial}{\partial\theta}\log f_\theta(y),\end{aligned}$$

the above equation can be simplified as

$$\begin{aligned}\mathrm{Var}(\widehat{\theta}(Y))I(\theta) &= \mathbb{E}[A^2]\mathbb{E}[B^2] \\ &\geq \mathbb{E}[AB]^2,\end{aligned}$$

where the inequality is due to Cauchy. We can also show that

$$\begin{aligned}\mathbb{E}[AB]^2 &= \left[\int (\widehat{\theta}(y) - \mathbb{E}[\widehat{\theta}(y)])(\frac{\partial}{\partial\theta}\log f_\theta(y)) f_\theta(y)dy\right]^2 \\ &= \left[\int (\widehat{\theta}(y) - \mathbb{E}[\widehat{\theta}(y)])\frac{\partial}{\partial\theta} f_\theta(y)dy\right]^2 \\ &= \left[\int \widehat{\theta}(y)\frac{\partial}{\partial\theta} f_\theta(y)dy - \mathbb{E}[\widehat{\theta}(Y)]\int \frac{\partial}{\partial\theta} f_\theta(y)dy\right]^2 \\ &= \left[\frac{\partial}{\partial\theta}\mathbb{E}[\widehat{\theta}(Y)] - 0\right]^2 = \left(\frac{\partial}{\partial\theta}\mathbb{E}[\widehat{\theta}(Y)]\right)^2.\end{aligned}$$

□

Proposition 2.
An estimator $\widehat{\theta}(Y)$ achieves CRLB equality if and only if $\widehat{\theta}(Y)$ is a sufficient statistic of a one-parameter exponential family.

Proof.
Suppose that CRLB equality holds, then we must have

$$\frac{\partial}{\partial\theta}\log f_\theta(y) = k(\theta)(\widehat{\theta}(y) - \mathbb{E}[\widehat{\theta}(Y)]),$$

for some function $k(\theta)$. This implies that

$$\begin{aligned}\log f_\theta(y) &= \int_a^\theta k(\theta')(\widehat{\theta}(y) - \mathbb{E}[\widehat{\theta}(Y)])d\theta' + H(y) \\ &= \underbrace{-\int_a^\theta k(\theta')\mathbb{E}[\widehat{\theta}(Y)]d\theta'}_{\log C(\theta)} + \underbrace{H(y)}_{\log h(y)} + \underbrace{\widehat{\theta}(y)\int_a^\theta k(\theta')d\theta'}_{\mathrm{Q}(\theta)}.\end{aligned}$$

Thus,

$$f_\theta(y) = C(\theta)\exp(Q(\theta)\widehat{\theta}(y))\cdot h(y),$$

which is a one-parameter exponential family.

Conversely, suppose that $\widehat{\theta}(Y)$ is a sufficient statistic of a one-parameter exponential family, then,

$$f_\theta(y) = C(\theta)\exp(Q(\theta)T(y))\cdot h(y),$$

where $T(y) = \widehat{\theta}(y)$, and

$$C(\theta) = \left(\int \exp(Q(\theta)T(y))\cdot h(y)dy\right)^{-1}.$$

In order to show that $\mathrm{Var}\{T(Y)\}$ attains the CRLB, we need to obtain the Fisher Information:

$$I(\theta) = \mathbb{E}\left\{\left(\frac{\partial}{\partial\theta}\log f_\theta(Y)\right)^2\right\}.$$

Note that since

$$\log f_\theta(y) = Q(\theta)T(y) + \log h(y) - \log \int \exp(Q(\theta)T(y))h(y)dy,$$

we must have

$$\begin{aligned}\frac{\partial}{\partial\theta}\log f_\theta(y) &= Q'(\theta)T(y) - \left(\int T(y)\frac{\exp(Q(\theta)T(y))h(y)}{\int \exp(Q(\theta)T(y))h(y)dy}dy\right)Q'(\theta)\\ &= Q'(\theta)\{T(y) - \mathbb{E}\{T(Y)\}\}.\end{aligned}$$

Therefore,

$$I(\theta) = \mathbb{E}\left(\frac{\partial}{\partial\theta}\log f_\theta(Y)\right)^2 = (Q'(\theta))^2\mathrm{Var}\{T(Y)\}.$$

The Cramer Rao Lower bound is

$$\mathrm{Var}(\widehat{\theta}(Y)) \geq \frac{(\frac{\partial}{\partial\theta}\mathbb{E}[\widehat{\theta}(Y)])^2}{I(\theta)}.$$

Thus we need to determine $(\frac{\partial}{\partial\theta}\mathbb{E}[\widehat{\theta}(Y)])^2$. Suppose that $\widehat{\theta}(Y) = T(Y)$,

$$\begin{aligned}&\frac{\partial}{\partial\theta}\{\mathbb{E}\{\widehat{\theta}(Y)\}\}\\ &= \frac{\partial}{\partial\theta}\frac{\int T(y)\exp(Q(\theta)T(y))h(y)dy}{\int \exp(Q(\theta)T(y))h(y)dy}\\ &= Q'(\theta)\frac{\int T(y)^2\exp(Q(\theta)T(y))h(y)dy\int \exp(Q(\theta)T(y))h(y)dy - (\int T(y)\exp(Q(\theta)T(y))h(y)dy)^2}{(\int \exp(Q(\theta)T(y))h(y)dy)^2}\\ &= Q'(\theta)\mathrm{Var}\{T(Y)\}.\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{(\frac{\partial}{\partial\theta}\mathbb{E}[\widehat{\theta}(Y)])^2}{I(\theta)} &= \frac{Q'(\theta)^2\mathrm{Var}\{T(Y)\}^2}{(Q'(\theta))^2\mathrm{Var}\{T(Y)\}}\\ &= \mathrm{Var}(T(Y))\\ &= \mathrm{Var}(\widehat{\theta}(Y)),\end{aligned}$$

which shows that CRLB equality is attained.

□

Example 10.
Let $\boldsymbol{Y} = [Y_1, \ldots, Y_n]$ be a sequence of iid random variables such that $Y_i \sim \mathcal{N}(\theta, \sigma^2)$. Consider the estimator $\widehat{\theta}(\boldsymbol{Y}) = \frac{1}{n}\sum_{i=1}^n Y_i$. Is $\widehat{\theta}(\boldsymbol{Y})$ an MVUE?

Solution:

The CRLB is

$$\mathrm{Var}(\widehat{\theta}) \geq \frac{(\frac{\partial}{\partial\theta}\mathbb{E}[\widehat{\theta}(\boldsymbol{Y})])^2}{I(\theta)},$$

where it is not difficult to show that $I(\theta) = \frac{n}{\sigma^2}$ and $\mathbb{E}[Y_i] = \theta$. Therefore, CRLB becomes

$$\mathrm{Var}(\widehat{\theta}) \geq \frac{1}{I(\theta)} = \frac{\sigma^2}{n}.$$

On the other hand, we can show that

$$\begin{aligned}\mathrm{Var}(\widehat{\theta}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(Y_i)\\ &= \frac{\sigma^2}{n} = \frac{1}{I(\theta)},\end{aligned}$$

which means that CRLB equality is achieved. Therefore, the estimator

$$\widehat{\theta}(\boldsymbol{Y}) = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is an MVUE.

Example 11.
Let $\boldsymbol{Y} = [Y_1, \ldots, Y_n]$ be a sequence of iid random variables such that $Y_k \sim s_k(\theta) + N_k$ where $s_k(\theta)$ is a function of k, and $N_k \sim \mathcal{N}(0, \sigma^2)$. Find CRLB for any unbiased estimator $\widehat{\theta}$.

Solution:

The log-likelihood is

$$\log f_\theta(\boldsymbol{y}) = \frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - s_k(\theta))^2$$

Consequently, we can show that

$$\frac{\partial^2}{\partial\theta^2}\log f_\theta(\boldsymbol{y}) = \frac{1}{\sigma^2}\sum_{k=1}^{n}\left[(y_k - s_k(\theta))\frac{\partial^2 s_k(\theta)}{\partial\theta^2}\right] - \frac{1}{\sigma^2}\sum_{k=1}^{n}\left(\frac{\partial s_k(\theta)}{\partial\theta}\right)^2.$$

Accordingly,

$$\mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2}\widehat{\theta}(\boldsymbol{Y})\right] = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\frac{\partial s_k(\theta)}{\partial\theta}\right)^2.$$

Therefore,

$$\mathrm{Var}(\widehat{\theta}) \geq \frac{1}{I(\theta)} = \frac{\sigma^2}{\sum_{i=1}^{n}(\frac{\partial}{\partial\theta}s_k(\theta))^2}.$$

For example, if $s_k = \theta$, then $\mathrm{Var}(\widehat{\theta}) \geq \frac{\sigma^2}{n}$. If $s_k(\theta) = A\cos(w_0 k + \theta)$, then $\mathrm{Var}(\widehat{\theta}) \geq \frac{2\sigma^2}{nA^2}$.
