Lecture 13
Properties of Maximum Likelihood Estimator (MLE)

Recall that

- $\hat{\theta}_{MLE}$ is an estimator such that

$$f(x_1, x_2, \ldots, x_N; \hat{\theta}_{MLE}) \geq f(x_1, x_2, \ldots, x_N; \theta), \forall \theta.$$

- If the likelihood function is differentiable with respect to

θ, then $\hat{\theta}_{MLE}$ is given by $\left.\dfrac{\partial f\ (\mathbf{x};\theta)}{\partial \theta}\right|_{\hat{\theta}_{\mathrm{MLE}}} = 0$

equivalently, $\left.\dfrac{\partial \mathrm{L}(\mathbf{x};\theta)}{\partial \theta}\right|_{\hat{\theta}_{\mathrm{MLE}}} = 0$

- The MLE conditions are extended to multiparameter case.

In this lecture, we will establish some important properties of of MLE.

Properties of MLE

- MLEs are backed by elegant mathematical theories. They possess desirable properties of good estimators, particularly for large samples
- In many situations, MLEs turn out to be MVUES
- For some distributions, closed-form solutions of likelihood equations may not exist. In sch cases, MLEs may be constructed numerically through iterative algorithms and their properties may also be studied.

Some properties of MLEs are described next.

Properties of MLE

(1) MLE may be biased or unbiased.

In the Example of iid Gaussian samples,

$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} X_i \quad \text{and}$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \hat{\mu}_{MLE}\right)^2$$

We can show that

$$E\hat{\mu}_{MLE} = \mu \quad \text{and}$$

$$E\hat{\sigma}^2_{MLE} = \frac{N-1}{N}\sigma^2$$

Properties of MLE

(2) If a sufficient statistic $T(\mathbf{x})$ exists for θ, then $\hat{\theta}_{MLE}$ is a function of $T(\mathbf{x})$.

Proof: By the factorization theorem,

$$f(\mathbf{x};\theta) = g(\theta, T(\mathbf{x}))h(\mathbf{x})$$

$$\therefore L(\mathbf{x};\theta) = \ln g(\theta, T(\mathbf{x})) + \ln h(\mathbf{x})$$

$$\left.\frac{\partial \mathrm{L}(\mathbf{x};\theta)}{\partial \theta}\right|_{\hat{\theta}_{\mathrm{MLE}}} = 0$$

$$\Rightarrow \left.\frac{\partial}{\partial \theta}(\ln g(\theta, T(\mathbf{x})) + \ln h(\mathbf{x}))\right|_{\hat{\theta}_{\mathrm{MLE}}} = 0$$

$$\Rightarrow \left.\frac{\partial}{\partial \theta}(\ln g(\theta, T(\mathbf{x}))\right|_{\hat{\theta}_{\mathrm{MLE}}} = 0$$

Therefore, $\hat{\theta}_{MLE}$ is a function of the sufficient statistic $T(\mathbf{x})$.

- ❖ We can find $\hat{\theta}_{MLE}$ by maximizing $f(T(\mathbf{x});\theta)$.

Properties of MLE

(3) If an efficient estimator exists, the MLE estimator is the efficient estimator.

Suppose an efficient estimator $\hat{\theta}$ exists. Then by Cramer Rao theorem,

$$\frac{\partial}{\partial\theta} L(\mathbf{x};\theta) = I(\theta)(\hat{\theta} - \theta)$$

Note that $\hat{\theta}$ is an MVUE

At $\theta = \hat{\theta}_{MLE}$,

$$\left.\frac{\partial L(\mathbf{x};\theta)}{\partial\theta}\right|_{\hat{\theta}\ MLE} = 0$$

$$\Rightarrow I(\theta)(\hat{\theta} - \hat{\theta}_{MLE}) = 0$$

$$\Rightarrow \hat{\theta}_{MLE} = \hat{\theta}$$

(5) Invariance Properties of MLE

It is a remarkable property of the MLE and not shared by other estimators. If $\hat{\theta}_{MLE}$ is the MLE of θ and $h(\theta)$ is a function, then $h(\hat{\theta}_{MLE})$ is the MLE of $h(\theta)$.

Proof- We prove the result when $h(\theta)$ is one-to-one.

Suppose $u = h(\theta)$. Then $\theta = h^{-1}(u)$ is given by

$$\frac{\partial f(\mathbf{x};\theta)}{\partial h(\theta)} = \frac{\partial f(\mathbf{x};\theta)}{\partial \theta} \times \frac{\partial \theta}{\partial h(\theta)}$$

At $\hat{\theta}_{MLE}$, $\frac{\partial f(\mathbf{x};\theta)}{\partial \theta} = 0$, Therefore,

$$\left.\frac{\partial f(\mathbf{x};\theta)}{\partial h(\theta)}\right]_{\hat{\theta}_{MLE}} = 0$$

(5) **Invariance Properties of MLE…**

❖We have proved the invariance property for the simple case when $h(\theta)$ is one to one and differentiable. However, the result is true also when $h(\theta)$ is many-to-one.

❖Invariance to a transformation is a remarkable property, not shared by other estimators.

Example- In our example of iid Gaussian samples,

$$\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \hat{\mu}_{MLE}\right)^2$$

$$\therefore \hat{\sigma}_{MLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \hat{\mu}_{MLE}\right)^2}$$

Example:

Suppose $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_N$ are iid $\mathrm{B}(1,\theta)$ random variables. Find the MLE for θ and hence the MLE for $\operatorname{var}(X_1)$

$$p(\mathbf{x};\theta) = \theta^{\sum_{i=1}^{N} x_i} (1-\theta)^{N-\sum_{i=1}^{N} x_i}$$

$$\therefore L(\mathbf{x};\theta) = \sum_{i=1}^{N} x_i \ln\theta + \left(N - \sum_{i=1}^{N} x_i \right) ln(1-\theta)$$

Applying the MLE condition, we get

$$\hat{\theta}_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Now, $\operatorname{var}(X_1) = \theta(1-\theta)$

$$\therefore \widehat{\operatorname{var}}(X_1)_{MLE} = \hat{\theta}_{MLE}(1-\hat{\theta}_{MLE})$$

where $\hat{\theta}_{MLE} = \dfrac{1}{N} \sum_{i=1}^{N} x_i$

Large sample properties of MLE

- We saw that if an efficient estimator exists, the MLE is the efficient estimator. Another attractive feature of the MLE is that its behavior becomes better and better with more number of samples.
- The asymptotic properties of MLE holds under the regularity conditions like those applied in deriving the Cramer Rao theorem. These conditions are usually satisfied in practice
- The MLE is asymptotically unbiased and efficient. Thus, for large N, the MLE is approximately efficient.

We will explain this property as follows;

Asymptotic efficiency MLE

- To show that for large N,

$$\frac{\partial}{\partial\theta}L(\mathbf{x};\theta)=I(\theta)(\hat{\theta}_{MLE}-\theta)$$

We have, $\left.\frac{\partial}{\partial\theta}L(\mathbf{x};\theta)\right]_{\hat{\theta}_{MLE}}=0$

$$\therefore 0=\frac{\partial}{\partial\theta}L(\mathbf{x};\hat{\theta}_{MLE})$$

$$=\frac{\partial}{\partial\theta}L(\mathbf{x};\theta+\hat{\theta}_{MLE}-\theta)$$

$$=\frac{\partial L(\mathbf{x};\theta)}{\partial\theta}+(\hat{\theta}_{MLE}-\theta)\frac{\partial^2 L(\mathbf{x};\theta_1)}{\partial\theta^2},\quad \theta<\theta_1<(\hat{\theta}_{MLE}$$

where we have applied the mean-value theorem.

$$\therefore \frac{\partial L(\mathbf{x};\theta)}{\partial\theta}=-\frac{\partial^2 L(\mathbf{x};\theta_1)}{\partial\theta^2}(\hat{\theta}_{MLE}-\theta)$$

Asymptotic efficiency MLE …

We have

$$\frac{\partial L(\mathbf{x};\theta)}{\partial \theta} = -\frac{\partial^2 L(\mathbf{x};\theta_1)}{\partial \theta^2}(\hat{\theta}_{MLE} - \theta)$$

Under the assumption of independence,

$$\frac{\partial^2 L(\mathbf{x};\theta_1)}{\partial \theta^2} = \frac{\partial^2 L(x_1;\theta_1)}{\partial \theta^2} + \frac{\partial^2 L(x_2;\theta_1)}{\partial \theta^2} + ... + \frac{\partial^2 L(x_N;\theta_1)}{\partial \theta^2}$$

=

$$\simeq E\frac{\partial^2 L(\mathbf{x})}{\partial \theta^2} \text{ (Applying WLLN)}$$

$$\therefore \frac{\partial L(\mathbf{x};\theta)}{\partial \theta} \simeq -E\frac{\partial^2 L(\mathbf{x};\theta_1)}{\partial \theta^2}(\hat{\theta}_{MLE} - \theta) = I(\theta)(\hat{\theta}_{MLE} - \theta)$$

Consistency of MLE

Recall that

- An estimator $\hat{\theta}$ is called a consistent estimator of θ if $\hat{\theta}$ converges in probability to θ.

$$\lim_{N\to\infty} \mathrm{P}\left(\left|\hat{\theta}\text{-}\theta\right| \geq \varepsilon\right) = 0 \text{ for any } \varepsilon > 0$$

- Further, if $\hat{\theta}$ is unbiased and $\lim\limits_{N\to\infty} \operatorname{var}(\hat{\theta}) = 0$.

It can be shown that under the regularity conditions, $\hat{\theta}_{MLE}$ is a consistent estimator. We omit the proof.

- The desired properties of $\hat{\theta}_{MLE}$ under large sample conditions make MLE an attractive estimator for the signal processing communities.
- These properties can be easily extended to multi-parameter cases.

Consistency of MLE….

In the Example of iid Gaussian samples,

$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

We can show that

$E\hat{\mu}_{MLE} = \mu$ and

$$\operatorname{var}\hat{\mu}_{MLE} = \frac{\sigma^2}{N}$$

$$\therefore \lim_{N\to\infty} \operatorname{var}\hat{\mu}_{MLE} = 0$$

Thus, $\hat{\mu}_{MLE}$ is a consistent estimator.

Summary

(1)MLE may be biased or unbiased.

(2) If a sufficient statistic $T(\mathbf{x})$ exists for θ, then $\hat{\theta}_{MLE}$ is a function of $T(\mathbf{x})$.

(3) If an efficient estimator exists, the MLE estimator is the efficient estimator. Thus, if

$$\frac{\partial}{\partial\theta} L(\mathbf{x};\theta) = I(\theta)(\hat{\theta} - \theta)$$

then, $\hat{\theta} = \hat{\theta}_{MLE}$

Summary…

(4) The MLE is asymptotically unbiased and efficient. Thus, for large N, the MLE is approximately efficient.

(5) $\hat{\theta}_{MLE}$ is a consistent estimator.

- ❖The desired properties of $\hat{\theta}_{MLE}$ under large sample conditions make MLE an attractive estimator for the signal processing communities.
- ❖These properties can be easily extended to multi-parameter case.

THANK YOU