Lecture 22: Maximum Likelihood Estimator

March 29, 2016

In the first part of this lecture, we will deal with the consistency and asymptotic distribution of maximum likelihood estimator. The second part of the lecture focuses on signal estimation/tracking.

1 Consistency of Max-Likelihood Estimator

An estimator is said to be consistent if it converges to the quantity being estimated. This section speaks about the consistency of MLE and conditions under which MLE is consistent.

Let $X_1, X_2, ..., X_n \overset{iid}{\sim} f(x|\theta),\ \theta \in \Theta \subseteq \mathbb{R}$. Then, Maximum Likelihood Estimate of θ is given by ,

$$\hat{\theta}_n = \underset{\theta \in \Theta}{argmax} \underbrace{\sum_{i=1}^{n} \log f(X_i|\theta)}_{L(X,\theta)}. \tag{1}$$

The following theorem talks about of the consistency of the maximum likelihood estimator. At first, we state a loose version of the theorem and then a proof sketch is provided.

Theorem 1.1 (Loose version). *Under regularity conditions on* $\{f(x|\theta) : \theta \in \Theta\}$*, we have,* $\hat{\theta}_n \xrightarrow{\mathbb{P}_\theta} \theta,\ \forall \theta \in \Theta$.

Proof. If $L(X, \theta)$ is differentiable in θ, then the derivative at $\theta = \hat{\theta}_n$ should be zero.

$$\frac{d}{d\theta} L(X, \theta)|_{\theta=\hat{\theta}_n} = 0, \tag{2}$$

i.e.,

$$\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,\hat{\theta}_n)=0, \tag{3}$$

where $\psi(X,\theta):=\frac{d}{d\theta'}\log f(X|\theta')|_{\theta'=\theta}$ is the score function at θ. If $\theta'\in\Theta$ is fixed, by weak law of large numbers, we have,

$$\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,\theta')\xrightarrow{n\to\infty}\mathbb{E}_\theta\big[\psi(X_1,\theta')\big], \tag{4}$$

and,

$$\mathbb{E}_\theta\big[\psi(X_1,\theta')\big]=\int\left[\frac{d}{d\theta}\log f(x|\theta)|_{\theta=\theta'}\right]f(x|\theta)dx\triangleq J(\theta,\theta'). \tag{5}$$

Now, consider $J(\theta,\theta)=\mathbb{E}_\theta\big[\psi(X_1,\theta)\big]=0$ (since expectation of score function is zero). So, $\theta'=\theta$ is a root of the equation $J(\theta,\theta')=0$.

Suppose that $\theta'=\theta$ is the unique root of $J(\theta,\theta')=0$. Suppose $J(\theta,\theta')$ and $\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,\theta')$ are smooth functions of θ', almost surely, then,

1. $\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,\theta')\approx J(\theta,\theta')$,
2. $\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,\theta)\approx J(\theta,\theta)=0$,
3. $\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,\hat{\theta}_n)=0$.

From the above equations one can infer that $\hat{\theta}_n\xrightarrow{\mathbb{P}_\theta}\theta$. □

The exact regularity conditions required for consistency of MLE are given in the theorem below:

Theorem 1.2. ***(Consistency of MLEs)*** *Suppose* $X_1,X_2,.....,X_n\overset{iid}{\sim}f(x|\theta)$, $\theta\in\mathbb{R}$, *and suppose*

1. $\forall\theta'$, $\log f(X|\theta')$ *is differentiable over* θ' *almost surely over* $X\sim f(X|\theta)$, *then*

$$J(\theta,\theta'):=\mathbb{E}_\theta\left[\frac{d}{d\tilde{\theta}}\log f(X|\tilde{\theta})|_{\tilde{\theta}=\theta'}\right]$$

   exists and is finite.

2. $J(\theta,\theta')$ *is continuous over* θ' *and has a unique root at* $\theta'=\theta$, *at which it changes sign.*

3. $\psi(X,\theta')$ *is continuous in* θ' *almost surely over* $X\sim f(X|\theta)$.

4. $\forall n\geq 1$, $\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,\theta')$ *has a unique root* $\hat{\theta}_n$.

Then, $\hat{\theta}_n\to\theta$ *in probability.*

2 Asymptotic Distribution of the MLE

In this section, we will talk about the asymptotic distribution of the maximum likelihood estimate. The asymptotic efficiency of the MLE is also shown in this section.

Theorem 2.1. ***(Asymptotic Normality of MLEs)*** *Let* $X_1, X_2, ..., X_n \overset{iid}{\sim} f(x|\theta_0),\ \theta_0 \in \mathbb{R} = \Theta$ *and* $\hat{\theta}_n$ *be an MLE under regularity conditions on* $\{f(x|\theta_0) : \theta_0 \in \Theta\}$. *Then,*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\boldsymbol{d}} \mathcal{N}(0, v(\theta_0)), \tag{6}$$

where,

$$v(\theta_0) = \frac{1}{\mathbb{E}_{\theta_0}\left[\left(\frac{d}{d\theta'}\log f(X|\theta')\big|_{\theta'=\theta_0}\right)^2\right]} = \frac{1}{\textit{Fisher Information at } \theta_0}, \tag{7}$$

is the Cramer-Rao lower bound for unbiased estimation of θ_0.

Proof.

$$\hat{\theta}_n = \arg\max_{\theta} \underbrace{\sum_{i=1}^{n} \log f(X_i|\theta)}_{L(X,\theta)}. \tag{8}$$

Let,

$$L^{'}(X,\theta) := \frac{d}{d\theta'} L(X,\theta')|_{\theta'=\theta}. \tag{9}$$

Consider the Taylor series of $L^{'}(X,\theta)$ around the point θ_0,

$$L^{'}(X,\theta) = L^{'}(X,\theta_0) + (\theta - \theta_0)L^{''}(X,\theta_0) + ... \text{ (higher order terms)},$$

$$L^{'}(X,\theta) \approx L^{'}(X,\theta) + (\theta - \theta_0)L^{''}(X,\theta_0). \tag{10}$$

At $\theta = \hat{\theta}_n$, from equation (10) we have,

$$0 = L^{'}(X,\theta_0) + (\hat{\theta}_n - \theta_0)L^{''}(X,\theta_0), \tag{11}$$

which gives,

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta_0) =& \frac{-\sqrt{n}L^{'}(X,\theta_0)}{L^{''}(X,\theta_0)} \\ =& \frac{-\frac{1}{\sqrt{n}}L^{'}(X,\theta_0)}{\frac{1}{n}L^{''}(X,\theta_0)}.\end{aligned} \tag{12}$$

Both the numerator and denominator are random variables.
Numerator:

$$\begin{aligned}\frac{1}{\sqrt{n}}L^{'}(X,\theta_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\underbrace{\frac{d}{d\theta}\log f(X_i|\theta)|_{\theta=\theta_0}}_{\psi(X_i,\theta_0)} \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\psi(X_i,\theta_0) - \underbrace{\mathbb{E}_{\theta_0}(\psi(X_i,\theta_0))}_{=0}] \\ &= \Big\{\frac{1}{\sqrt{n}\sqrt{Var_{\theta_0}(\psi(X_1,\theta_0))}}\sum_{i=1}^{n}\underbrace{\psi(X_i,\theta_0) - \mathbb{E}_{\theta_0}(\psi(X_i,\theta_0))}_{\text{i.i.d r.v}}\Big\}\sqrt{Var_{\theta_0}(\psi(X_1,\theta_0))}.\end{aligned} \tag{13}$$

The summation inside the curly brackets converges in distribution to Gaussian $\mathcal{N}(0,1)$ by the C.L.T. Hence, multiplication with $\sqrt{Var_{\theta_0}(\psi(X_1,\theta_0))}$ gives,

$$\frac{1}{\sqrt{n}}L^{'}(X,\theta_0) \xrightarrow{\mathbf{d}} \mathcal{N}\big(0, Var_{\theta_0}(\psi(X_1,\theta_0)\big), \tag{14}$$

where $Var_{\theta_0}\big[\psi(X_1,\theta_0)\big]$ is the Fisher Info(θ_0).
Denominator:

$$\frac{1}{n}L^{''}(X,\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\theta^2}\log f(X_i|\theta)\Big|_{\theta=\theta_0}. \tag{15}$$

The denominator converges to Fisher Info(θ_0) as $n \to \infty$ by the W.L.L.N.

$$\frac{1}{n}L^{''}(X,\theta_0) \xrightarrow[n\to\infty]{\mathbf{W.L.L.N}} \mathbb{E}_{\theta_0}\left[\frac{d^2}{d\theta^2}\log f(X_i|\theta)\Big|_{\theta=\theta_0}\right] = -\text{Fisher Info}(\theta_0). \tag{16}$$

Therefore,

$$\frac{Numerator}{Denominator} \xrightarrow{\mathbf{d}} \mathcal{N}\Big(0, \frac{1}{\text{Fisher Info}(\theta_0)}\Big) = \mathcal{N}(0, v(\theta_0)). \tag{17}$$

□

Remark 1. For large n,

$$\hat{\theta}_n - \theta \overset{(\approx)}{\sim} \mathcal{N}\left(0, \frac{v(\theta)}{n}\right), \tag{18}$$

and

$$var_\theta[\hat{\theta}_n] = \frac{v(\theta)}{n} = \frac{1}{n \times \text{Fisher Info}(\theta)}. \tag{19}$$

An estimate is said to be asymptotically efficient if it meets the CRLB. Therefore, the MLE $\hat{\theta}_n$ is asymptotically efficient.

3 Signal Estimation/Tracking

Goal: Estimate the evolution of a dynamic system.

Till now we were estimating time invariant parameters only, but now we are interested in parameters that vary with time. Such dynamic parameters are usually called a signal and hence the problem of estimating dynamic parameters is known as signal estimation or tracking. A general dynamic system model can be non-linear in nature but many of these non-linear systems can be approximated as linear systems.

3.1 Kalman-Bucy Filter

Here, first we study *Linear Discrete-time Dynamic System* which can be modeled with the following set of equations

$$\underline{\mathrm{X}}_{n+1} = \mathbf{F}_n \underline{\mathrm{X}}_n + \mathbf{G}_n \underline{\mathrm{U}}_n, \tag{20}$$

$$\underline{\mathrm{Y}}_n = \mathbf{H}_n \underline{\mathrm{X}}_n + \underline{\mathrm{V}}_n. \tag{21}$$

where $\underline{\mathrm{X}}_0, \underline{\mathrm{X}}_1, ...$ are sequence of vectors in $\mathbb{R}^m$ representing state of the system under study and $\underline{\mathrm{Y}}_0, \underline{\mathrm{Y}}_1, ...$ in $\mathbb{R}^k$ are the observation sequence of the system. $\underline{\mathrm{U}}_n$ in $\mathbb{R}^s$ is the Control/Process noise applied to the system, and $\underline{\mathrm{V}}_n$ in $\mathbb{R}^k$ represents the measurement noise. The quantities $\mathbf{F}_n, \mathbf{G}_n, \mathbf{H}_n$ are matrices ($\forall n \geq 0$) of appropriate dimensions ($m \times m$, $m \times s$ and $k \times m$, respectively).

3.2 Applications:

1. Aircraft tracking, navigation: The positional coordinates and attitudinal coordinates are the states of interest in flight control.The inputs may consist of both control and random forces acting on the aircraft.In this case, the state equation describes the dynamics of the aircraft.

2. Chemical process control: The states may be quantities as temperature,pressure and concentration of various chemicals ,and the dynamics of the chemical process is described by the state equation.

3. Radar systems: Estimating the position of the target and predicting the position of the target on the next scan.

4. Missile guidance

5. GPS receivers

Example 3.1 (1-D Kinematics). Consider a particle subjected to a force. Let the state $\underline{\mathrm{X}}_t$ be in $\mathbb{R}^2$,

$$\underline{\mathrm{X}}_t = (P_t, V_t), \tag{22}$$

where P_t represents the position of the particle and V_t represents the velocity of the particle. Let the initial state of the particle be (P_0, V_0). Suppose an acceleration of A_t is applied to the particle, then the system is looked at $t = 0, 1, 2, ...$ with a sampling time interval $\Delta \approx 0$.

$$V_t = \frac{\Delta P_t}{\Delta} = \frac{P_{t+1} - P_t}{\Delta}, \tag{23}$$

$$A_t = \frac{\Delta V_t}{\Delta} = \frac{V_{t+1} - V_t}{\Delta}. \tag{24}$$

From the above equations, we get,

$$P_{t+1} = P_t + \Delta V_t, \tag{25}$$

$$V_{t+1} = V_t + \Delta A_t. \tag{26}$$

The above set of equations can be represented in the matrix form as,

$$\begin{bmatrix} P_{t+1} \\ V_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & \Delta \\ 0 & 1 \end{bmatrix} \begin{bmatrix} P_t \\ V_t \end{bmatrix} + \begin{bmatrix} 0 \\ \Delta \end{bmatrix} A_t. \tag{27}$$

Suppose the particle position is measured with noise.

$$Y_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} P_t \\ V_t \end{bmatrix} + N_t. \tag{28}$$

Now our goal is to estimate $\underline{\mathrm{X}}_n$ using only the observations $[\underline{\mathrm{Y}}_0, \underline{\mathrm{Y}}_1, ...\underline{\mathrm{Y}}_t] \equiv \underline{\mathrm{Y}}_0^t$.

We can classify the signal estimation problem into three types:

1. $n < t$ gives *smoothing problem.*
2. $n = t$ gives *filtering problem.*
3. $n > t$ gives *prediction problem.*

A **very important special case** of signal estimation is "linear dynamical system driven by Gaussian noise/controls".

$$\underline{\mathrm{X}}_{n+1} = \mathbf{F}_n \underline{\mathrm{X}}_n + \mathbf{G}_n \underline{\mathrm{U}}_n, \tag{29}$$

$$\underline{\mathrm{Y}}_n = \mathbf{H}_n \underline{\mathrm{X}}_n + \underline{\mathrm{V}}_n. \tag{30}$$

Here $\underline{\mathrm{X}}_0 \sim \mathcal{N}(\underline{\mathrm{m}}_0, \Sigma_0)$ and $\{\underline{\mathrm{V}}_n\}$, $\{\underline{\mathrm{U}}_n\}$ are independent sequences of independent, zero-mean Gaussian vectors, independent of X_0.

In *filtering*, the goal is to estimate $\underline{\mathrm{X}}_t$ given $\underline{\mathrm{Y}}_0^t$ minimizing the square error. Let the estimate be $\hat{X}_{t|t}$, then our criterion is to minimize,

$$\mathbb{E}\Big[||\hat{X}_{t|t} - X_t||^2\Big]. \tag{31}$$