Large Sample Theory for MLEs

Introduction

Consider the following:

- i.i.d. sample $X_1, X_2, ..., X_n$
- one-dimensional paramter θ with $l(\theta) = \sum_{i=1}^{n} log(f(x_i|\theta)$
- the true value of θ is θ_0

Thm A: Under appropriate smoothness conditions on f, the MLE of θ from an i.i.d. sample is consistent, i.e.,

$$\hat{\theta} \xrightarrow{p} \theta_0, \quad n \to \infty. \tag{1}$$

Proof: (The idea)

First, think of the transformed variable $logf(X|\theta)$ and consider $logf(X_1|\theta), logf(X_2|\theta), ..., logf(X_n|\theta)$ as a random sample. So its mean is

$$\frac{1}{n}\sum_{i=1}^{n} logf(X_i|\theta) \xrightarrow{p} E_{\theta_0}[logf(X_i|\theta)] \quad n \to \infty. \tag{2}$$

The subscript on θ, θ_0, is to identify the true unknown value of θ.

With some imagination (or doing Calculus) we can guess the maximum of the righthand side should be at θ_0 since the convergence says that the lefthand side and the righthand side will be close for large n.

Lemma A: Define $I(\theta)$, Fisher's Information, by

$$I(\theta) = E\left[\frac{d}{d\theta} logf(X|\theta)\right]^2. \tag{3}$$

Under the appropriate smoothness conditions on f, $I(\theta)$ may be represented as

$$I(\theta) = -E\left[\frac{d^2}{d\theta^2} logf(X|\theta)\right]. \tag{4}$$

Proof:

First recall that for any density funtion

$$\int f(x|\theta)dx = 1 \tag{5}$$

and so

$$\frac{d}{d\theta}\int f(x|\theta)dx = 0. \tag{6}$$

Note that since

$$\frac{d}{dx} log(g(x)) = \frac{\frac{d}{dx} g(x)}{g(x)}, \tag{7}$$

then

$$\frac{d}{d\theta} f(x|\theta) = \left[\frac{d}{d\theta} log f(x|\theta)\right] f(x|\theta). \tag{8}$$

So

$$\begin{aligned} 0 &= \frac{d}{d\theta} \int f(x|\theta) dx \\ &= \int \frac{d}{d\theta} f(x|\theta) dx \\ &= \int \left[\frac{d}{d\theta} log f(x|\theta)\right] f(x|\theta) dx \end{aligned}$$

The moving of the derivative inside the integral takes some math. Now take the second derivative.

$$\begin{aligned} 0 &= \frac{d}{d\theta} \int \left[\frac{d}{d\theta} log f(x|\theta)\right] f(x|\theta) dx \\ &= \int \left[\frac{d^2}{d\theta^2} log f(x|\theta)\right] f(x|\theta) dx + \int \left[\frac{d}{d\theta} log f(x|\theta)\right] \frac{d}{d\theta} f(x|\theta) dx \\ &= \int \left[\frac{d^2}{d\theta^2} log f(x|\theta)\right] f(x|\theta) dx + \int \left[\frac{d}{d\theta} log f(x|\theta)\right]^2 f(x|\theta) dx \\ &= E\left[\frac{d^2}{d\theta^2} log f(X|\theta)\right] + E\left[\frac{d}{d\theta} log f(X|\theta)\right]^2. \end{aligned}$$

Remarks about Fisher's Information

For the one-parameter model $f(x|\theta)$ the Fisher Information $I(\theta)$ is the fundamental measure of the "information" an observation X carries about θ.

$$E\left[\frac{d^2}{d\theta^2} log f(X|\theta)\right] = -E\left[\frac{d}{d\theta} log f(X|\theta)\right]^2. \tag{9}$$

1. $I_X(\theta)$ is best thought of in relative terms, i.e., X is more informative about θ than Y if $I_X(\theta) > I_Y(\theta)$.
2. The more information that X contains about θ, the larger $I_X(\theta)$ should be. If $X = \theta$ with probability one, then we would like to have $I_X(\theta) = \infty$.
3. If $X_1, X_2, ..., X_n$ i.i.d. F_θ, then
$$I_{\mathbf{X}}(\theta) = nI_X(\theta) \tag{10}$$
where $\mathbf{X} = (X_1, X_2, ..., X_n)$.

The information of n X's is n times the information from one X and as $n \to \infty$, the information goes to ∞, and is eventually perfect.

Proof:

$$\begin{aligned}
I_{\mathbf{X}}(\theta) &= -E\left[\frac{d^2}{d\theta^2} log f(X_1, X_2, ..., X_n|\theta)\right] \\
&= -E\left[\frac{d^2}{d\theta^2} log \prod_{i=1}^{n} f(X_i|\theta)\right] \\
&= -E\left[\frac{d^2}{d\theta^2} \sum_{i=1}^{n} log f(X_i|\theta)\right] \\
&= \sum_{i=1}^{n}\left[-E\left[\frac{d^2}{d\theta^2} log f(X_i|\theta)\right]\right] \\
&= \sum_{i=1}^{n} I_X(\theta) \\
&= nI_X(\theta)
\end{aligned}$$

The large sample distribution of an MLE is approximately normal with mean θ_0 and variance $\frac{1}{nI(\theta_0)}$. This implies that the MLE is asymptotically unbiased. It also implies that the variance of the limiting normal distribution has the asymptotic variance of the MLE.

Thm B: Under the smoothness conditions of f, the probability distribution of

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \tag{11}$$

tends to a standard normal distribution.

Proof:

Use the W.L.L.N., the C.L.T. and Lemma A.

So we have

$$\hat{\theta} \sim N\left(\theta_0, \frac{1}{nI(\theta_0)}\right) \tag{12}$$

and approximately

$$\hat{\theta} \sim N\left(\theta_0, \frac{1}{nI(\hat{\theta})}\right) \tag{13}$$

from which we can produce an approximate $100(1-\alpha)$ confidence interval for θ

$$\hat{\theta} \pm z_{\alpha/2} \frac{1}{\sqrt{nI(\hat{\theta})}} \tag{14}$$

Notes about the smoothness conditions

The true parameter value, θ_0, is required to be an interior point of the set of all parameter values.

It is also required that the support of the density $f(x|\theta)$ does not depend of θ.

Final notes

The asymptotic variance of the MLE $\hat{\theta}$ is

$$\frac{1}{nI(\theta_0)} = -\frac{1}{E\left[\frac{d^2}{d\theta^2} l(\theta)\right]}. \tag{15}$$