Large Sample Theory for MLE

Consider the log likelihood function\(^1\)

\[
l_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} l_i(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f(X_i|\theta)
\]

By large number law we have

\[
l_n(\theta) \xrightarrow{p} \mathbb{E}_X(\log f(X|\theta)|\theta_0), \forall \theta
\]

Therefore, by maximizing \(l_n(\theta)\), we are approximately maximizing

\[
\hat{\theta} = \arg \max_{\theta} l_n(\theta) \approx \arg \max_{\theta} \mathbb{E}_X(\log f(X|\theta)|\theta_0)
\]

Note when \(\theta = \theta_0\),

\[
\frac{\partial \mathbb{E}_X(\log f(X|\theta)|\theta_0)}{\partial \theta} \bigg|_{\theta=\theta_0} = 0
\]

Thus we claim \(l'_n(\hat{\theta}) \approx 0\). We have

\[
0 = l'_n(\hat{\theta}) \approx l'_n(\theta_0) + (\hat{\theta} - \theta_0)l''_n(\theta_0) \Rightarrow \hat{\theta} - \theta = \frac{l'_n(\theta_0)}{-l''_n(\theta_0)}
\]

Definition 1 (Fisher Information\(^1\)).

\[
I(\theta) = \mathbb{E} \left( \frac{\partial f(X|\theta)}{f(X|\theta)} \right)^2 = \mathbb{E} \left( -\partial_{\theta\theta} f(X|\theta) \right)
\]

Remark 1. This is well defined.

\[
\mathbb{E} \left( -\partial_{\theta\theta} f(X|\theta) \right) = \int_X -\frac{\partial}{\partial \theta} \frac{\partial f(X|\theta)}{f(X|\theta)} f(X|\theta) dX
\]

\[
= \int_X \frac{-\partial_{\theta\theta} f(X|\theta) f(X|\theta) + (\partial_{\theta} f(X|\theta))^2}{f(X|\theta)^2} f(X|\theta) dX
\]

\[
= - \int_X \partial_{\theta\theta} f(X|\theta) dX + \mathbb{E} \left( \frac{\partial_{\theta} f(X|\theta)}{f(X|\theta)} \right)^2
\]

\[
= \mathbb{E} \left( \frac{\partial_{\theta} f(X|\theta)}{f(X|\theta)} \right)^2
\]

\(^1\)There are two things, \(l_n(\theta)\), which is called scaled log likelihood and \(l(\theta) = \log f(X|\theta)\)
Note
\[
\mathbb{E}l'_n(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\partial_\theta \log f(X_i|\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\frac{\partial_\theta f(X_i|\theta)}{f(X_i|\theta)}
\]
We have
\[
\mathbb{E}\frac{\partial_\theta f(X_i|\theta)}{f(X_i|\theta)} = \int_X \partial_\theta f(X|\theta) dX = \partial_\theta 1 = 0, \forall i = 1, 2, \ldots, n
\]
thus \(\mathbb{E}l'_n(\theta_0) = 0\). In addition, we have
\[
\text{var}(l'_n(\theta_0)) = \frac{1}{n} \mathbb{E}\left(\frac{\partial_\theta f(X_i|\theta_0)}{f(X_i|\theta_0)}\right)^2 = \frac{I(\theta_0)}{n}
\]
and
\[
\mathbb{E}(-l''_n(\theta_0)) = I(\theta_0)
\]
Thus from (1)(2) we have
\[
l'_n(\theta) \overset{p}{\rightarrow} N(0, \frac{I(\theta_0)}{n})
\]
by Slutsky theorem\(^2\)\([2]\) we have
\[
-l''_n(\theta_0) \overset{p}{\rightarrow} I(\theta_0)
\]
Thus we have
\[
\hat{\theta} - \theta \approx \frac{l'_n(\theta_0)}{-l''_n(\theta_0)} \overset{p}{\rightarrow} N\left(0, \frac{1}{nI(\theta_0)}\right)
\]
in a more familiar form,
\[
\sqrt{n} \left(\hat{\theta} - \theta\right) \sim N\left(0, I(\theta_0)^{-1}\right)
\]
Let’s consider an example. Let \(\{X_i\}_{i=1}^{n}\) be i.i.d. sample from the exponential distribution
\[
f(x|\theta) = \theta e^{-\theta x}, x > 0
\]
\(^2\) Slutsky theorem reads

**Theorem 1** (Slutsky). Let \(X_n\) and \(Y_n\) be two sequences of random vectors such that \(X_n \overset{d}{\rightarrow} X, Y_n \overset{P}{\rightarrow} c\), where \(c\) is a constant. Let \(g(x, y)\) be a continuous function, we have
\[
g(X_n, Y_n) \overset{d}{\rightarrow} g(X, c)
\]
then we have

\[ l_n(\theta) = n \log \theta - \theta \sum_{i=1}^{n} X_i \]

The MLE is \( l'(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} X_i = 0 \Rightarrow \hat{\theta}_n = \frac{n}{\sum X_i} \). Let’s compute the mean and variance of \( \hat{\theta}_n \).

Indeed,

\[ \mathbb{E}(\hat{\theta}_n) = \int_{\mathbb{R}^n} \frac{n}{\sum x_i} e^{-\theta \sum x_i} dx_1 \cdots dx_n = \theta \int_{\mathbb{R}^n} \frac{n}{\sum x_i} e^{-\theta \sum x_i} dx_1 \cdots dx_n \]

this indicates \( \mathbb{E}(\hat{\theta}_n) = c\theta \) for some constant \( c \). Differentiate both sides by \( \theta \), we have

\[ c = nc - n \Rightarrow c = \frac{n}{n-1} \]

therefore we have \( \mathbb{E}(\hat{\theta}_n) = \frac{n}{n-1} \theta \). This indicates the estimator \( \hat{\theta}_n \) is biased but asymptotically consistent. We can also compute

\[ \text{var}(\hat{\theta}_n) = \frac{n^2}{(n-1)^2 (n-2)} \theta^2 \]

note the Fisher information is

\[ I(\theta) = -l''(\theta) = \frac{1}{\theta^2} \]

we have

\[ \text{var}(\hat{\theta}_n) = \frac{n^2}{(n-1)^2 (n-2)} \theta^2 \geq \frac{1}{n} \theta^2 \]

and asymptotically the Crammer-rao bound is achieved.

**References**
