Asymptotic Theorem for Maximum Likelihood Estimates

Jin-Lung Lin

1 A baseline theorem

Let $X_1, X_2, \cdots, X_n$ be *iid* with mean μ and variance $\sigma^2 (< \infty)$. Denote $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then,

- *Strong Law of Large Number (SLLN)*
  $\overline{X}_n \overset{wp1}{\to} \mu$
- *Central Limit Theorem (CLT)*

$$\begin{aligned} \sqrt{n}(\overline{X}_n - \mu) &\overset{D}{\to} N(0, \sigma^2), \quad \text{or} \\ \sqrt{n}\frac{(\overline{X}_n - \mu)}{\hat{\sigma}} &\overset{D}{\to} N(0,1) \end{aligned}$$

  where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$. Note that how the mode of convergence depends upon the normalization rate.

2 Maximum Likelihood Estimates

See Serfling, Robert J. (1980) *Approximation Theorems of Mathematical Statistics*, Section 4.2 for details.

2.1 Definitions

Let $X_1, X_2, \cdots, X_n$ be *iid* with density function $f(x;\theta), \theta \in \Theta$. Then, the *likelihood function* of the sample $X_1, X_2, \cdots, X_n$ is defined as

$$L(\theta; X_1, X_2, \cdots, X_n) = \Pi_{i=1}^{n} f(X_i; \theta)$$

The maximum likelihood estimate $\hat{\theta}$ is the one which maximizes L (or equivalently $\log L$) in Θ.

$$\hat{\theta} = argmax_\theta \log L(\theta; X_1, X_2, \cdots, X_n)$$

The first order conditions are :

$$\frac{\partial \log L}{\partial \theta}\Big|_{\theta=\hat{\theta}} = 0$$

2.2 Regularity conditions

Consider θ to be an open interval in R.

(R1) For each $\theta \in \Theta$, the derivatives

$$\frac{\partial \log f(x;\theta)}{\partial \theta}, \frac{\partial^2 \log f(x;\theta)}{\partial \theta^2}, \frac{\partial^3 \log f(x;\theta)}{\partial \theta^3}$$

exist for all x.

(R2) For each $\theta_0 \in \Theta$, there exists functions $g(x)$, $h(x)$ and $H(x)$ such that for θ in a neighborhood $N(\theta_0)$, the relations

$$|\frac{\partial f(x;\theta)}{\partial \theta}| \leq g(x), \quad |\frac{\partial^2 f(x;\theta)}{\partial \theta^2}| \leq h(x), \quad |\frac{\partial^3 f(x;\theta)}{\partial \theta^3}| \leq H(x),$$

hold for all x and

$$\int g(x)dx < \infty, \quad \int h(x)dx < \infty, \quad E_\theta\{H(X)\} < \infty$$

for $\theta \in N(\theta_0)$.

(R3) For each $\theta \in \Theta$
$0 < E_\theta\{(\frac{\partial \log f(X;\theta)}{\partial \theta})^2\} < \infty$

2.3 Theorem

Assume that $X_1, X_2, \cdots, X_n$ is *iid* with density function $f(x;\theta), \theta \in \Theta$ and the regularity conditions hold. Let $\hat{\theta}_n$ be the MLE for θ for sample $X_1, X_2, \cdots, X_n$. Then,

1. strong consistency
$\hat{\theta}_n \overset{wp1}{\to} \theta$

2. Asymptotic normality
$\sqrt{n}(\hat{\theta}_n - \theta) \overset{D}{\to} N(0, V^{-1}(\theta))$ where $V(\theta) = E_\theta(\frac{\partial \log f(X;\theta)}{\partial \theta})^2$

2.4 Proof

For λ in the neighborhood $N(\theta)$, A Taylor expansion of $\partial \log f(x;\lambda)/\partial\lambda$ about the point $\lambda = \theta$ gives

$$\frac{\partial \log f(x;\lambda)}{\partial \lambda} = \frac{\log f(x;\lambda)}{\partial \lambda}\Big|_{\lambda=\theta} + (\lambda - \theta)\frac{\partial^2 \log f(x;\lambda)}{\partial \lambda^2}\Big|_{\lambda=\theta} + \frac{1}{2}\xi(\lambda - \theta)^2 H(x)$$

where $|\xi| < 1$. Putting,

$$A_n = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(X_i;\lambda)}{\partial \lambda}\Big|_{\lambda=\theta}, \quad B_n = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(X_i;\lambda)}{\partial \lambda^2}\Big|_{\lambda=\theta}, \quad C_n = \frac{1}{n}\sum_{i=1}^{n} H(X_i)$$

we have,

$$\frac{1}{n}\frac{\partial \log L(\lambda)}{\partial \lambda} = A_n + B_n(\lambda-\theta) + \frac{1}{2}\xi^* C_n(\lambda-\theta)^2$$

where $|\xi| < 1$.

By $(R1, R2)$,

$$\int \frac{\partial f(x;\lambda)}{\partial \lambda} dx = \frac{\partial}{\partial \lambda}\int f(x;\lambda)dx = \frac{\partial}{\partial \theta}(1) = 0$$

and

$$\int \frac{\partial^2 f(x;\lambda)}{\partial \lambda^2} dx = 0$$

It follows that

$$\begin{aligned} E_\theta\{\frac{\partial \log f(X;\theta)}{\partial \theta}\} &= \int \frac{1}{f(x;\theta)}\frac{\partial f(x;\theta)}{\partial \theta} f(x;\theta)dx = 0 \\ E_\theta\{\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\} &= \int \left[\frac{1}{f(x;\theta)}\frac{\partial^2 f(x;\theta)}{\partial \theta^2} - (\frac{1}{f(x;\theta)}\frac{\partial f(x;\theta)}{\partial \theta})^2\right] f(x;\theta)dx \\ &= -E_\theta\{(\frac{\partial \log f(X;\theta)}{\partial \theta})^2\} \end{aligned}$$

By $(R3)$,

$$V(\theta) = E_\theta\{(\frac{\partial \log f(X;\theta)}{\partial \theta})^2 < \infty$$

It follows that

(a) A_n is a mean of *iid*'s with means 0 and variance $V(\theta)$;

(b) B_n is a mean of *iid*'s with means $-V(\theta)$;

(c) C_n is a mean of *iid*'s with means $E_\theta(H(X))$;

By SLLN,

$$A_n \overset{wp1}{\to} 0, \quad B_n \overset{wp1}{\to} -V(\theta), \quad C_n \overset{wp1}{\to} E_\theta(H(X)),$$

and by CLT,

$$\sqrt{n}(A_n - 0) \overset{D}{\to} N(0, V(\theta))$$

The SLLN results follows. To prove CLT, note that

$$0 = \frac{1}{n}\frac{\partial \log L(\lambda)}{\partial \lambda}\Big|_{\lambda=\hat{\theta}_n} = A_n + B_n(\hat{\theta}_n - \theta) + \frac{1}{2}\xi^* C_n(\hat{\theta}_n - \theta)^2$$

As $\hat{\theta}_n - \theta \xrightarrow{wp1} 0$,

$$\sqrt{n}(\hat{\theta}_n - \theta) - \frac{-\sqrt{n}A_n}{B_n + 1/2\xi^* C_n(\hat{\theta}_n - \theta)^2} \xrightarrow{wp1} 0$$

Also, since $\hat{\theta}_n \xrightarrow{wp1} \theta$, we have $B_n + 1/2\xi^* C_n(\hat{\theta} - \theta)^2 \xrightarrow{wp1} -V(\theta)$. Consequently, by Slutsky's Theorem,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} N(0, V^{-1}(\theta)V(\theta)V^{-1}(\theta)) \equiv N(0, V^{-1}(\theta))$$

that $I(\theta) = -E_\theta\left[\frac{\partial^2 \log L(\theta; X_1, X_2, \cdots, X_n)}{\partial \theta^2}\right]$ is called (Fisher) Information matrix. When the model is correctly specified, $I(\theta) = nV(\theta)$.

3 Cramer-Rao Theorem

Theorem Let $X_1, X_2, \cdots, X_n$ be *iid* with density function $f(x; \theta), \theta \in \Theta$. The *likelihood function* of the sample $X_1, X_2, \cdots, X_n$ is

$$L(\theta; X_1, X_2, \cdots, X_n) = \Pi_{i=1}^n f(X_i; \theta)$$

Define the information matrix for the random sample $X' = (X_1, X_2, \cdots, X_n)$ to be

$$I(\theta) = -E\left[\frac{\partial^2 \log L(X; \theta)}{\partial \theta \theta'}\right]$$

where the ijth element of the information matrix is

$$I_{ij} = -E\left[\frac{\partial^2 \log L(X; \theta)}{\partial \theta_i \theta_j}\right], \quad i, j = 1, \cdots, K$$

Let $\hat{\theta}$ be any unbiased estimator of θ with covariance matrix Σ. Then, the matrix $\Sigma - [I(\theta)]^{-1}$ is positive semidefinite and $[I(\theta)]^{-1}$ is known as Cramer-Rao lower bound matrix.

4 Example

Let $X_1, X_2, \cdots, X_n$ be *iid* normal with mean 0 and variance σ^2. Then, the $\theta = (\mu, \sigma^2)$ and the *likelihood function* and *log-likelihood function* of the sample $X_1, X_2, \cdots, X_n$ are respectively

$$\begin{aligned} L(\mu, \sigma^2; X_1, X_2, \cdots, X_n) &= \frac{1}{(2\pi)^n (\sigma^2)^{n/2}} \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2) \\ \log L(\mu, \sigma^2; X_1, X_2, \cdots, X_n) &= C - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2 \end{aligned}$$

$$
\begin{aligned}
\frac{\partial \log L}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu) \\
\frac{\partial \log L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i - \mu)^2 \\
\frac{\partial^2 \log L}{\partial \mu^2} &= -\frac{n}{\sigma^2}; \quad -E(\frac{\partial^2 \log L}{\partial \mu^2}) = \frac{n}{\sigma^2} \\
\frac{\partial^2 \log L}{\partial \sigma^4} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(X_i - \mu)^2; \quad -E(\frac{\partial^2 \log L}{\partial \sigma^4}) = \frac{n}{2\sigma^4} \\
\frac{\partial^2 \log L}{\partial \mu \sigma^2} &= \frac{1}{\sigma^4}\sum_{i=1}^{n}(X_i - \mu); \quad E(\frac{\partial^2 \log L}{\partial \mu \partial \sigma^2}) = 0
\end{aligned}
$$

Thus,

$$
I(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}
$$

Let $\hat{\mu}, \hat{\sigma^2}$ denote the MLE of μ, σ^2 respectively. Obviously,

$$
\begin{aligned}
\hat{\mu} &= \bar{X} \\
\hat{\sigma^2} &= 1/n\sum_{i=1}^{n}(X_i - \bar{X})^2
\end{aligned}
$$

By the asymptotic normality theorem,

$$
\sqrt{n}(\begin{pmatrix} \hat{\mu} \\ \hat{\sigma^2} \end{pmatrix} - \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}) \xrightarrow{D} N\left(0, V^{-1}(\mu, \sigma^2)\right) \equiv N(0, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix})
$$

By Cramer-Roa theorem, MLE $\hat{\mu}, \hat{\sigma^2}$ reach lower bound and is efficient.