Lecture 10
Large Sample Theory of MLE

Dr. Shafayat Abrar

Large Sample Theory for Method of Maximum Likelihood

Theorem A (Rice, 2007)

Under appropriate smoothness conditions on f, the mle from an i.i.d. sample is consistent.

This theorem simply means that the derivative of log likelihood is really zero when the optimizing parameter θ is close to its true but unknown value θ_0. Or in other words, optimization of likelihood does yield a consistent estimator.

$$\frac{\partial}{\partial \theta} \frac{1}{n} \sum_{i=1}^{n} \log[f(x_i \mid \theta)] \rightarrow \frac{\partial}{\partial \theta} \mathbb{E} \log f(x \mid \theta) = 0, \quad \text{when } \hat{\theta}_n \rightarrow \theta$$

Proof of Theorem A

Consider maximizing

$$\frac{1}{n}l(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log[f(X_i \mid \theta)]$$

As n tends to infinity, the law of large numbers implies that

$$\frac{1}{n}l(\theta) \to E \log f(X \mid \theta) = \int \log f(x \mid \theta) f(x \mid \theta_0) dx$$

To maximize $E \log f(X \mid \theta)$, we consider its derivative:

$$\frac{\partial}{\partial\theta}\int \log f(x \mid \theta) f(x \mid \theta_0) dx = \int \frac{\frac{\partial}{\partial\theta} f(x \mid \theta)}{f(x \mid \theta)} f(x \mid \theta_0) dx$$

If $\theta = \theta_0$, this equation becomes

$$\int \frac{\partial}{\partial\theta} f(x \mid \theta_0) dx = \frac{\partial}{\partial\theta}\int f(x \mid \theta_0) dx = \frac{\partial}{\partial\theta}(1) = 0$$

which shows that θ_0 is a stationary point and hopefully a maximum.

Large Sample Theory for Method of Maximum Likelihood

Lemma A (Rice, 2007)

Define $I(\theta)$ by

$$I(\theta) = \mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right]^2$$

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

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

Proof of Lemma A

First, we observe that since $\int f(x \mid \theta)dx = 1$,

$$\frac{\partial}{\partial\theta}\int f(x \mid \theta)dx = 0$$

Combining this with the identity

$$\frac{\partial}{\partial\theta} f(x \mid \theta) = \left[\frac{\partial}{\partial\theta}\log f(x \mid \theta)\right] f(x \mid \theta)$$

$$0 = \frac{\partial}{\partial\theta}\int f(x \mid \theta)dx = \int\left[\frac{\partial}{\partial\theta}\log f(x \mid \theta)\right] f(x \mid \theta)dx$$

Taking second derivative both sides, we obtain

$$0 = \int\left[\frac{\partial^2}{\partial\theta^2}\log f(x \mid \theta)\right] f(x \mid \theta)dx + \int\left[\frac{\partial}{\partial\theta}\log f(x \mid \theta)\right]^2 f(x \mid \theta)dx$$

$$E\left[\frac{\partial^2}{\partial\theta^2}\log f(x \mid \theta)\right] + E\left[\frac{\partial}{\partial\theta}\log f(x \mid \theta)\right]^2 = 0$$

Large Sample Theory for Method of Maximum Likelihood

Theorem B: (Rice, 2007)

Under smoothness conditions on f, the probability distribution of $\sqrt{nI(\theta_0)}\left(\hat{\theta} - \theta_0\right)$ tends to a standard normal distribution.

i.e.,

$$\sqrt{nI(\theta_0)}\left(\hat{\theta} - \theta_0\right) \overset{n\to\infty}{\rightsquigarrow} N(0, 1)$$

or in other words,

$$\tilde{\theta} = \hat{\theta} - \theta_0 \overset{n\to\infty}{\rightsquigarrow} N\left(0, \frac{1}{nI(\theta_0)}\right)$$

$$\hat{\theta} \overset{n\to\infty}{\rightsquigarrow} N\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$$

To prove this theorem, we need to evaluate the variance of estimation error $\tilde{\theta}$.

Proof of Theorem B

From a Taylor series expansion, we may expand $\ell'(\hat{\theta}) = 0$ around the true value θ_0 as follows:

$$0 = \ell'(\hat{\theta}) \approx \ell'(\theta_0) + \left(\hat{\theta} - \theta_0\right)\ell''(\theta_0)$$

$$\left(\hat{\theta} - \theta_0\right) \approx -\frac{\ell'(\theta_0)}{\ell''(\theta_0)}$$

$$n^{1/2}\left(\hat{\theta} - \theta_0\right) \approx -\frac{n^{1/2}\ell'(\theta_0)}{\ell''(\theta_0)} = \frac{-n^{-1/2}\ell'(\theta_0)}{n^{-1}\ell''(\theta_0)}$$

$$\operatorname{var}\left(n^{1/2}\left(\hat{\theta} - \theta_0\right)\right) \approx \operatorname{var}\left(\frac{-n^{-1/2}\ell'(\theta_0)}{n^{-1}\ell''(\theta_0)}\right)$$

First, we consider the denominator:

$$\frac{1}{n}l''(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta_0^2}\log f(x_i \mid \theta_0)$$

By the law of large numbers, the latter expression converges to

$$\mathbb{E}\left[\frac{\partial^2}{\partial\theta_0^2}\log f(X \mid \theta_0)\right] =: -I(\theta_0)$$

Proof of Theorem B

$$\mathrm{var}\left(n^{1/2}\left(\hat{\theta}-\theta_0\right)\right) \approx \mathrm{var}\left(\frac{-n^{-1/2}\ell'(\theta_0)}{-I(\theta_0)}\right) = \frac{1}{I(\theta_0)^2}\mathrm{var}\left(-n^{-1/2}\ell'(\theta_0)\right)$$

$$\begin{aligned}\mathrm{var}\left[n^{-1/2}l'(\theta_0)\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial}{\partial\theta_0}\log f(X_i \mid \theta_0)\right]^2 - \left(\frac{1}{\sqrt{n}}\mathbb{E}\left[\frac{\partial}{\partial\theta_0}\log f(X_i \mid \theta_0)\right]\right)^2 \\ &= I(\theta_0) - 0 \\ &= I(\theta_0)\end{aligned}$$

$$\mathrm{var}\left(n^{1/2}\left(\hat{\theta}-\theta_0\right)\right) \approx \frac{1}{I(\theta_0)^2}\mathrm{var}\left(-n^{-1/2}\ell'(\theta_0)\right) = \frac{1}{I(\theta_0)^2}I(\theta_0) = \frac{1}{I(\theta_0)}$$

$$\mathrm{var}\left(\tilde{\theta}\right) = \mathrm{var}\left(\hat{\theta}-\theta_0\right) \approx \frac{1}{nI(\theta_0)}$$

Q Let $X_1, \ldots, X_n$ be an i.i.d. sample from the following density:

$$f(x \mid a) = ax^{a-1}, \quad \text{for } 0 < x < 1, \text{ and } a > 0$$

(a) Find the maximum likelihood (ML) estimate of a.

(b) Find the approximate variance of ML estimate using large sample theory.

(c) Find the method of moments (MoM) estimate of a.

(d) Find the mean and variance of MoM estimate using approximate method.

$$f(x|a) = a x^{a-1} \quad \text{for } 0 < x < 1 \text{ and } a > 0$$

$$l(a) = \log\left(\prod_i a x_i^{a-1}\right) = \log\left(a^n \prod_{i=1}^{n} x_i^{a-1}\right)$$

$$= n \log(a) + (a-1) \log(x_i)$$

$$l'(a) = \frac{n}{a} + \log(x_i) = 0$$

$$\Rightarrow \quad a = \frac{-1}{\frac{1}{n} \sum_i \log(x_i)}$$

$$\Rightarrow \boxed{\hat{a}_{ML} = \frac{-1}{\frac{1}{n} \sum_{i=1}^{n} \log(x_i)}}$$

$$\operatorname{var}(\hat{a}_{ML}) = \operatorname{var}(\tilde{a}_{ML}) \quad \text{where } \tilde{a}_{ML} = \hat{a}_{ML} - a$$

$$\boxed{\operatorname{var}(\tilde{a}_{ML}) = \frac{1}{n\, I(a)}}$$

$$\boxed{I(a) := -E\left[\frac{\partial^2}{\partial a^2} l(a)\right]}$$

$$\boxed{\operatorname{var}(\tilde{a}_{ML}) = \frac{a^2}{n}}$$

For n-samples $I(a)$ is computed as follows:

$$I(a) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial a^2} \log\left(\prod_{i=1}^{n} a\, x_i^{a-1}\right)$$

$$I(a) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial a}\left(\frac{n}{a} + \log(x_i)\right) = +\frac{1}{a^2}$$

Now let us find the bias of $\hat{a}_{ML}$

$$\hat{a}_{ML} = \frac{-1}{T(x)} \quad \text{where } T(x) = \frac{1}{N} \sum_{i=1}^{n} \log(x_i)$$

or we may write

$$\hat{a}_{ML} = g(T) = -\frac{1}{T}$$

Since $\quad Y = g(x) \approx g(X) + (Y-X)g'(X) + \frac{1}{2!}(Y-X)^2 g''(X)$

Therefore, for $T = \mu_T + \tilde{T}$, we may write

$$\hat{a}_{ML} = g(T) = g(\mu_T + \tilde{T}) \approx g(\mu_T) + \tilde{T} g'(\mu_T) + \frac{1}{2!} \tilde{T}^2 g''(\mu_T)$$

$$= g(\mu_T) + (T - \mu_T) g'(\mu_T) + \frac{1}{2} (T - \mu_T)^2 g''(\mu_T)$$

$$E(\hat{a}_{ML}) = g(\mu_T) + \underbrace{E(T-\mu_T)}_{\text{always zero}} g'(\mu_T) + \frac{1}{2}\underbrace{E\left[(T-\mu_T)^2\right]}_{=\text{variance of } T.} g''(\mu_T)$$

$$g(T) = -\frac{1}{T} \Rightarrow g'(T) = \frac{1}{T^2} \Rightarrow g''(T) = -\frac{2}{T^3}.$$

$$\Rightarrow g(\mu_T) = -\frac{1}{\mu_T} \Rightarrow g'(\mu_T) = \frac{1}{\mu_T^2} \Rightarrow g''(T) = -\frac{2}{\mu_T^3}.$$

$$\mu_T = E[T(x)] = E\left[\frac{1}{n}\sum_{i=1}^{n} \log(x_i)\right] = E[\log(x)]$$

$$E[\log(x)] = \int_0^1 \log x \; a x^{a-1} dx = \frac{x^a(a\log(x)-1)}{a}\Big|_0^1$$

not defined for the lower limit

$$\boxed{\mu_T = E(\log(x)) = -\frac{1}{a}}$$

Next we compute $var(T)$

$$var(T(x)) = var\left(\frac{1}{n}\sum_{i=1}^{n} \log(x_i)\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} var(\log(x_i))$$

$$= \frac{1}{n} var(\log(x))$$

$$= \frac{1}{n}\left[E(\log^2(x)) - [E(\log(x))]^2\right]$$

$$E[\log^2(x)] = \int_0^1 \log^2(x)\, a x^{a-1}\, dx$$

$$= \left.\frac{x^a(a^2\log^2(x) - 2a\log(x) + 2)}{a^2}\right|_0^1$$

applying only upper limit (limit not defined for $x=0$)

$$\boxed{E[\log^2(x)] = \frac{2}{a^2}}$$

$$\mathrm{Var}(T(x)) = \frac{1}{n}\left(\frac{2}{a^2} - \frac{1}{a^2}\right) = \frac{1}{na^2}$$

$$\Rightarrow E(\hat{a}_{ML}) = g(\mu_T) + \frac{1}{2} E(T-\mu_T)^2 g''(\mu_T)$$

$$= -\frac{1}{(-1/a)} + \cancel{\frac{1}{2}} \cdot \frac{\cancel{2}}{a^2}\left(-\frac{1}{-1/a^3}\right)$$

$$\boxed{E(\hat{a}_{ML}) = \underbrace{a}_{\text{true value}} + \underbrace{\frac{a}{n}}_{\text{bias}}}$$

as $n \to \infty$

$E(\hat{a}_{ML}) \to a_{true}$ (Asymptotically Consistent.

Similarly show that (As an activity)

$$\hat{a}_{MOM} = \frac{\hat{\mu}_1}{1-\hat{\mu}_1}$$

$$E[\hat{a}_{MOM}] = a + \frac{a(a+1)}{n(a+2)}$$

$$var(\hat{a}_{MOM}) = \frac{a(a+1)^2}{n(a+2)}$$

```
function Lect10_ML_code
clc; drawnow; pause(0.1)
n = 200; % number of samples
a = 1.2; % True (given) value of a
loop =500000;

for ii=1:loop
   X = randQ1(n,a);
   aML(ii) = -1/mean(log(X)) ; % ML estimator
   aMoM(ii) = mean(X)/(1-mean(X)); % MoM estimator
   % disp([a aML(ii) aMoM(ii)]); drawnow
end

format short g; disp('---------------------------------------------')
disp(['ML: Analytically obtained VARIANCE = ' num2str(a^2/n)])
disp(['ML: Empirically obtained VARIANCE = ' num2str(mean((a-aML).^2))])
disp('---------------------------------------------')
disp(['MoM: Analytically obtained VARIANCE = ' num2str(a*(a+1)^2/n/(a+2))])
disp(['MoM: Empirically obtained VARIANCE = ' num2str(mean((a-aMoM).^2))])
disp('---------------------------------------------')
disp(['ML: Analytically obtained BIAS = ' num2str(a/n)])
disp(['ML: Empirically obtained BIAS = ' num2str(mean(aML)-a)])
disp('---------------------------------------------')
disp(['MoM: Analytically obtained BIAS = ' num2str(a*(a+1)/n/(a+2))])
disp(['MoM: Empirically obtained BIAS = ' num2str(mean(aMoM)-a)])
disp('---------------------------------------------’); drawnow

function X = randQ1(n,a)
r = rand(1,n); % Uniformly distributed random numbers
X = r.^(1/a); % Random numbers from pdf f(x|a)=a*x^(a-1)
```

```
---------------------------------------------
ML: Analytically obtained VARIANCE = 0.0072
ML: Empirically obtained VARIANCE = 0.0073878
---------------------------------------------
MoM: Analytically obtained VARIANCE = 0.009075
MoM: Empirically obtained VARIANCE = 0.0092243
---------------------------------------------
ML: Analytically obtained BIAS = 0.006
ML: Empirically obtained BIAS = 0.0058387
---------------------------------------------
MoM: Analytically obtained BIAS = 0.004125
MoM: Empirically obtained BIAS = 0.0039331
---------------------------------------------
```