L17 The consistency of MLE

1. Identifiable distributions

   (1) Definition

   pdf/pmf family $f(x;\,\theta)$ is identifiable $\overset{def}{\iff}$ "$f(x;\,\theta)=f(x;\,\beta)\iff\theta=\beta''$
   $\iff$ "$f(x;\,\theta)=f(x;\,\beta)\Longrightarrow\theta=\beta''$

   Ex1: Distribution family $N\sim N(\mu,\,1^2)$ is identifiable.

   Proof: $X\sim N(\mu,\,1^2)$ has pdf $f(x;\,\mu)=\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x-\mu)^2\right]$.

   $$f(x;\,\mu)=f(x;\,\eta)\iff\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x-\mu)^2\right]=\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x-\eta)^2\right]$$
   $$\iff(x-\mu)^2=(x-\eta)^2\overset{*}{\Longrightarrow}\mu=\eta.$$

   $\overset{*}{\Longrightarrow}$: Taking $x=0$, then taking $x=1$ lead to $\mu=\eta$.

   Ex2: Distribution family $N(0,\,\sigma)$ with pdf $f(x;\,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2}\right)$ is not identifiable since $f(x,\,\sigma)=f(x,\,v)\iff\sigma^2=v^2\Longrightarrow\sigma=\pm v$.

   (2) A property

   X has identifiable pdf/pmf $f(x;\,\theta)$ such that $\ln f(X;\,\beta)$ is a family of random variables indexed by β with

   $$E_\theta[\ln f(X;\,\beta)]=\begin{cases}\int_R[\ln f(x;\,\beta)]f(x;\,\theta)dx\\ \sum_x[\ln f(x;\,\beta)]f(x;\,\theta)\end{cases}\overset{def}{=\!=}g(\beta|\theta).$$

   Then $g(\beta|\theta)<g(\theta|\theta)$ for $\beta\neq\theta$.

   Proof: Consider continuous X only.

   $g(\beta|\theta)-g(\theta|\theta)=E_\theta[\ln f(X;\,\beta)]-E_\theta[\ln f(X;\,\theta)]=E_\theta\left[\ln\frac{f(X;\beta)}{f(X;\theta)}\right]$.

   But $\ln(x)$ is a strictly concave function since $(\ln x)'=\frac{1}{x}$ is decreasing on $(0,\,\infty)$.

   By Jensen's inequality $\ln[E(X)]>E[\ln(X)]$. So

   $$\begin{aligned}g(\beta|\theta)-g(\theta|\theta)&=E_\theta\left[\ln\tfrac{f(X;\beta)}{f(X;\theta)}\right]<\ln\left[E_\theta\tfrac{f(X;\beta)}{f(X;\theta)}\right]\\&=\ln\textstyle\int_R\frac{f(x;\beta)}{f(x;\theta)}f(x;\,\theta)dx=\ln\int_R f(x;\,\beta)dx=\ln 1=0.\end{aligned}$$

   Therefore $g(\beta|\theta)<g(\theta|\theta)$. □

2. Log-likelihood function

   (1) $\frac{\ln L(\beta|X_1,\ldots,X_n)}{n}$

   Let $X_1,....,X_n$ is be a random sample from X with identifiable pdf/pmf $f(x;\,\theta)$ and $\ln L(\theta|X_1,...,X_n)$ be the log-likelihood function. When $\beta\neq\theta$, by SLLN and (2) in 1,

   $$\begin{aligned}\tfrac{\ln L(\beta|X_1,..,X_n)}{n}&=\tfrac{\ln f(X_1;\beta)+\cdots+\ln f(X_n;\beta)}{n}\xrightarrow{a.s.}g(\beta|\theta)\\&<g(\theta|\theta)\xleftarrow{a.s.}\tfrac{\ln f(X_1;\theta)+\cdots+\ln f(X_n;\theta)}{n}=\tfrac{\ln L(\theta|X_1,\ldots,X_n)}{n}.\end{aligned}$$

   Ex3: For $\epsilon>0$, let B_ϵ be a ball with center θ and radius ϵ, i.e.,
   $B_\epsilon=\{\beta\in R^k:\,\|\beta-\theta\|\leq\epsilon\}$. Then outside the ball $\beta\in B_\epsilon^c\Longrightarrow\beta\neq\theta$ and hence

   $$\begin{aligned}\tfrac{\ln L(\beta|X_1,..,X_n)}{n}&=\tfrac{\ln f(X_1;\beta)+\cdots+\ln f(X_n;\beta)}{n}\xrightarrow{a.s.}g(\beta|\theta)\\&<g(\theta|\theta)\xleftarrow{a.s.}\tfrac{\ln f(X_1;\theta)+\cdots+\ln f(X_n;\theta)}{n}=\tfrac{\ln L(\theta|X_1,\ldots,X_n)}{n}.\end{aligned}$$

(2) $\ln(\beta|X_1, ..., X_n)$

Suppose for $\epsilon > 0$, $\frac{\ln L(\beta|X_1,...,X_n)}{n} \xrightarrow{a.s.} g(\beta|\theta)$ is uniform with respect to $\beta \in B_\epsilon^c$. Then there exists $N_\epsilon > 0$, $\forall\, n > N_\epsilon$ and all $\beta \in B_\epsilon^c$ $\ln L(\beta|X_1, ..., X_n) < \ln L(\theta|X_1, ..., X_n)$.

Proof: For $\epsilon > 0$, under the given conditions

$$\frac{\ln L(\beta|X_1,..,X_n)}{n} \xrightarrow{a.s.} g(\beta|\theta) < \frac{g(\beta|\theta)+g(\theta|\theta)}{2}$$
$$\iff \exists\, N_{1\epsilon} > 0,\, \forall\, n > N_{1\epsilon}\ \frac{\ln L(\beta|X_1,..,X_n)}{n} < \frac{g(\beta|\theta)+g(\theta|\theta)}{2}.$$

$$\frac{\ln L(\theta|X_1,..,X_n)}{n} \xrightarrow{a.s.} g(\theta|\theta) > \frac{g(\beta|\theta)+g(\theta|\theta)}{2}$$
$$\iff \exists\, N_{2\epsilon} > 0,\, \forall\, n > N_{2\epsilon}\ \frac{\ln L(\theta|X_1,..,X_n)}{n} > \frac{g(\beta|\theta)+g(\theta|\theta)}{2}.$$

Thus with $N_\epsilon = \max[N_{1\epsilon},\, N_{2\epsilon}] > 0$ when $n > N_\epsilon$ for all $\beta \in B_\epsilon^c$
$\ln L(\beta|X_1, .., X_n) < \ln L(\theta|X_1, .., X_n)$. □

3. Consistency of MLE

(1) MLE

$\widehat{\beta}_n$ is MLE for θ if $\ln L(\widehat{\beta}_n|X_1, ..., X_n) \geq \ln L(\beta|X_1, ..., X_n)$ for all $\beta \in R^k$.

(2) Consistency of MLE

By (2) in 2, $L(\beta|X_1, .., X_n) < L(\theta|X_1, .., X_n)$ for all $\beta \in B_\epsilon^c$ for large n.
Thus with these n, the MLE $\widehat{\beta}_n \notin B_\epsilon^c$. So for $\epsilon > 0$, $\exists\, N_\epsilon > 0\ \forall\, n > N_\epsilon$, $\|\widehat{\beta}_n - \theta\| \leq \epsilon$.
Hence for $\epsilon > 0$ and $n > N_\epsilon > 0$, $P(\|\widehat{\beta}n - \theta\| \leq \epsilon) = 1$. Therefore $\widehat{\beta}_n \xrightarrow{p} \theta$.
Thus the MLE for θ, $\widehat{\theta}_n$, is a consistent estimator for θ.

Ex4: $X_1, ..., X_n$ is a random sample from $X \sim N(\mu,\, \sigma^2)$ with parameter $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$.

It is well-known that the MLE for θ is $\widehat{\theta}_n = \begin{pmatrix} \overline{X}_n \\ \frac{\sum_i (X_i - \overline{X})^2}{n} \end{pmatrix}$ where
$\frac{\sum_i (X_i - \overline{X})^2}{n} = \frac{1}{n}\left[\sum_i X_i^2 - \frac{1}{n}\left(\sum_i X_i\right)^2\right] = \frac{1}{n}\sum_i X_i^2 - \left(\overline{X}_n\right)^2$.
By SLLN, $\overline{X}_n \xrightarrow{as} \mu$ and $\frac{\sum_{i=1}^n X_i^2}{n} \xrightarrow{as} E(X^2)$.
So $\begin{pmatrix} \overline{X}_n \\ \frac{\sum_{i=1}^n X_i^2}{n} - \overline{X}_n^2 \end{pmatrix} \xrightarrow{as} \begin{pmatrix} \mu \\ E(X^2) - \mu^2 \end{pmatrix}$, i.e., $\begin{pmatrix} \overline{X}_n \\ \frac{\sum_{i=1}^n (X_i - \overline{X}_n)^2}{n} \end{pmatrix} \xrightarrow{as} \theta$.
Hence the MLE of θ is a consistent estimator for θ.

L18: Convergence in distributions

1. Convergence in distributions

(1) The concept

The convergence in distribution also called the convergence in law of X_n to $X \in R^k$ denoted as $X_n \xrightarrow{d} X \in R^k$ or $X_n \xrightarrow{L} X$

$\overset{def}{\iff}$ $P(X_n \in A) \longrightarrow P(X \in A)$ for all A with $P(X \in [\text{the boundary of } A]) = 0$

$\iff$ $F_n(x) \longrightarrow F(x)$ for all continuity points x of $F(x)$.

$F_n(x)$ and $F(x)$ are cdfs for X_n and X. $F(x) = P(X_1 \le x_1, ..., X_k \le x_k)$.

(2) Properties

(i) $X_n \xrightarrow{d} X \in R^k \iff g(X_n) \xrightarrow{d} g(X) \in R^q$ for all continuous $g(\cdot)$.

(ii) $X_n \xrightarrow{d} X \iff X_{n_i} \xrightarrow{d} X$ for all subsequence X_{n_i}.

(iii) $X_n \xrightarrow{d} X \iff \alpha' X_n \xrightarrow{d} \alpha' X$ for all $\alpha \in R^k$.

Comment: (iii) is called the Cramer-Wold device.

By (iii) $X_n \xrightarrow{d} X \Longrightarrow X_{in} \xrightarrow{d} X_i$ for all $i = 1, .., k$.

But going back needs the convergence in all directions in R^k.

2. Convergence of random variables in distributions

(1) CDF $F(x) = P(X \le x)$ for $X \in R$

(i) $F(x)$ is a non-decreasing function with $\lim_{x \to -\infty} F(x) = F(-\infty) = 0$ and $\lim_{x \to \infty} F(x) = F(\infty) = 1$.

(ii) $F(x)$ is a right-continuous, i.e., $\lim_{x > x_0,\, x \to X_0} F(x) = F(x_0+) = F(x_0)$ for all $x_0 \in R$.

(iii) For $x_0 \in R$, $\lim_{x < x_0,\, x \to x_0} F(x) = F(x_0-)$ exist. $P(X = x_0) = F(x_0) - F(x_0-)$. Thus $F(x)$ is continuous at $x_0 \iff P(X = x_0) = 0$.

(iv) Let D be the collection of all discontinuity points for $F(\cdot)$. Then D contains at most countable points. Consequently the continuous points for $F(\cdot)$ is almost everywhere.

(2) $X_n \xrightarrow{p} X \in R \Longrightarrow X_n \xrightarrow{d} X \in R$

Proof: With continuity point x for $F(x)$, we show $F_n(x) \longrightarrow F(x)$.

$$\begin{aligned} F(x-\epsilon) &= P(X \le x-\epsilon,\, |X_n - X| < \epsilon) + P(X \le x-\epsilon,\, |X_n - X| \ge \epsilon) \\ &\le F_n(x) + P(|X_n - X| \ge \epsilon). \\ F_n(x) &= P(X_n \le x,\, |X - X_n| < \epsilon) + P(X_n \le x,\, |X - X_n| \ge \epsilon) \\ &\le F(x+\epsilon) + P(|X - X_n| \ge \epsilon). \end{aligned}$$

So $F(x-\epsilon) - P(|X_n - X| \ge \epsilon) \le F_n(x) \le F(x+\epsilon) + P(|X_n - X| \ge \epsilon)$.

Let $n \longrightarrow \infty$, $\epsilon \longrightarrow 0$. Then $F_n(x) \longrightarrow F(x)$. □

(3) $X_n \xrightarrow{p} c \in R \iff X_n \xrightarrow{d} c \in R$.

Proof: $\Rightarrow$ is by (2). Now only consider $\Leftarrow$. For $\epsilon > 0$,

$$\begin{aligned} P(|X_n - c| \ge \epsilon) &= P(X_n \le c-\epsilon) + P(X_n \ge c+\epsilon) \\ &\le F_n(c-\epsilon) + P(X_n > c + \tfrac{\epsilon}{2}) = F_n(c-\epsilon) + 1 - F_n(c+\tfrac{\epsilon}{2}) \\ &\longrightarrow F(c-\epsilon) + 1 - F(c+\tfrac{\epsilon}{2}) = 0 + 1 - 1 = 0. \end{aligned}$$

So $X_n \xrightarrow{d} c \Longrightarrow X_n \xrightarrow{p} c$. □

Ex1: $X_n \xrightarrow{p} X \in R^k \Longrightarrow X_n \xrightarrow{d} X \in R^k$.

Proof: $X_n \xrightarrow{p} X \Longrightarrow \alpha' X_n \xrightarrow{p} \alpha' X$ for all $\alpha \in R^k \Longrightarrow \alpha' X_n \xrightarrow{d} \alpha' X$ for all $\alpha \in R^k$.
Thus $X_n \xrightarrow{d} X$.

Ex2: $X_n \xrightarrow{p} C \in R^k \Longleftrightarrow X_n \xrightarrow{d} C \in R^k$.

Proof: $\Rightarrow$ is by Ex1. Only consider $\Leftarrow$.
$X_n \xrightarrow{d} C \Longrightarrow X_{in} \xrightarrow{d} C_i$ for all $i = 1, .., k. \Longrightarrow X_{in} \xrightarrow{p} c_i \in R$ for all $i = 1, ..., k$.
Thus $X_n \xrightarrow{p} C$.

3. Review asymptotically efficient estimators

   (1) Efficient estimator

   $\widehat{\theta}_n$ is an efficient estimator for $\theta \iff$ (i) $E(\widehat{\theta}_n) = \theta$; (ii) $\text{Cov}(\widehat{\theta}_n) = \text{CRLB}(\theta)$
   $\Longrightarrow$ $\widehat{\theta}_n$ is MVUE for θ.

   When the observations in sample are iid, $\text{CRLB}(\theta) = [nI(\theta)]^{-1}$.

   $\widehat{\eta}_n$ is an efficient estimator for $\eta(\theta) \iff$ (i) $E(\widehat{\eta}_n) = \eta(\theta)$; (ii) $\text{Cov}(\widehat{\eta}_n = \text{CRLB}(\eta(\theta))$
   $\Longrightarrow$ $\widehat{\eta}_n$ is MVUE for $\eta(\theta)$.

   With iid observations in sample, $\text{CRLB}(\eta(\theta)) = \left[\frac{\partial \eta}{\partial \theta^T}\right] [nI(\theta)]^{-1} \left[\frac{\partial \eta}{\partial \theta^T}\right]^T$

   (2) Asymptotically efficient estimator

   $\widehat{\eta}_n$ is an asymptotically efficient estimator for $\eta(\theta)$
   $\Longleftrightarrow$ (i) $\widehat{\eta}_n \xrightarrow{p} \eta(\theta)$, (ii) $n\left[\text{Cov}(\widehat{\eta}_n) - \text{CRLB}(\eta(\theta))\right] \longrightarrow 0$.

   (3) Asymptotically efficient estimator may not be an efficient estimator.
   Efficient estimator may not be an asymptotically efficient estimator. For example Hodges estimator is efficient but not asymptotically efficient.

   - Le Cam, L. M. (1953). On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes' Estimates. University of California Publications in Statistics, 1(11), 277–329.