ST5215 Advanced Statistical Theory, Lecture 9

HUANG Dongming

National University of Singapore

8 Sep 2020

Overview

Last time

- Completeness
- Basu's theorem

Today

- Basic elements of statistical inferences
- Point Estimation
  - Method of Moments Estimators(MM estimator)
  - Maximum Likelihood Estimators (MLE)

Recap: Ancillary statistics, Completeness, Basu's theorem

- A statistic $V(X)$ is said to be *ancillary* if its distribution does not depend on the population P
  $V(X)$ is said to be *first-order ancillary* if $E_P[V(X)]$ does not depend on P
- A statistic $T(X)$ is said to be (boundedly) *complete* for $P \in \mathcal{P}$ if no (bounded) function of $T(X)$ is first-order ancillary
- "Completeness + Sufficiency $\Rightarrow$ Minimal Sufficiency" (provided a minimal sufficient statistic exists)
- The $T(X)$ statistic in a natural exponential family of full rank is complete, sufficient, and minimal sufficient
- Basu's theorem: If $T(X)$ is boundedly complete and sufficient for $P \in \mathcal{P}$, and $V(X)$ is ancillary, then $T(X) \perp\!\!\!\perp V(X)$
  - For i.i.d. sample from a normal distribution, $\bar{X} \perp\!\!\!\perp S^2$

Suppose $X_1, \ldots, X_n \sim P_\theta \in \mathcal{P}$, where $\theta = (\theta_1, \ldots, \theta_k) \in \Theta$.

- An estimator for estimating θ

$$\widehat{\theta} = \widehat{\theta}_n = w\left(X_1, \ldots, X_n\right)$$

  is a function of the data (it is a statistic)
- The parameter is a fixed, unknown constant, while the estimator is a random variable (a realization of an estimator is called an *estimate*)
- Methods of constructing estimators:
  1. The Method of Moments (MM)
  2. Estimating Equation (EE)
  3. Maximum likelihood (MLE)
  4. Bayesian estimators
  5. ...
- For a given parameter, there may be many reasonable estimators
- Methods for evaluating estimators including:
  1. Bias and Variance
  2. Mean squared error (MSE)
  3. Admissibility
  4. Minimax Theory
  5. Large sample theory

The Method of Moments

Suppose X_i's are i.i.d. from P_θ and $E_\theta |X_1|^k < \infty$. (Recall that $E_\theta X$ is defined as $\int X \, \mathrm{d}P_\theta$)

- Let $\mu_j = E_\theta X_1^j$ be the jth moment of P_θ
- Typically,

$$\mu_j = h_j(\theta), \quad j = 1, \dots, k \tag{1}$$

  for some functions h_j on $\mathcal{R}^k$
- The jth sample moment ($j = 1, \dots, k$): $\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^n X_i^j$ is an unbiased estimator of μ_j
- Substitute μ_j 's on the LHS of Eq. (1) by the sample moments $\hat{\mu}_j$, we obtain an estimator $\hat{\theta}$ that satisfies

$$\hat{\mu}_j = h_j(\hat{\theta}), \quad j = 1, \dots, k$$

  which is a sample analogue of Eq. (1) (*the substitution principle*)
- Let $\hat{\mu} = (\hat{\mu}_1, \dots, \hat{\mu}_k)$ and $h = (h_1, \dots, h_k)$. Then $\hat{\mu} = h(\hat{\theta})$
- If h^{-1} exists, the unique moment estimator of θ is $\hat{\theta} = h^{-1}(\hat{\mu})$

Example: Normal models

Suppose $P_\theta = N\left(\beta, \sigma^2\right)$ with $\theta = \left(\beta, \sigma^2\right)$

- $\mu_1 = \beta$ and $\mu_2 = \sigma^2 + \beta^2$
- Equate:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \widehat{\beta}, \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \widehat{\sigma}^2 + \widehat{\beta}^2$$

- MM estimator:

$$\widehat{\beta} = \bar{X}, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$$

Example: Gamma models

Suppose $P_\theta = \Gamma(\alpha, \lambda)$, whose density is

$$\frac{1}{\Gamma(\alpha)\gamma^\alpha} x^{\alpha-1} e^{-x/\gamma} I_{(0,\infty)}(x)$$

- $\mu_1 = \alpha/\gamma$ and $\mu_2 - \mu_1^2 = \alpha/\gamma^2$
- Equate:

$$\frac{1}{n}\sum_{i=1}^n X_i = \frac{\widehat{\alpha}}{\widehat{\gamma}}, \quad \frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2 = \frac{\widehat{\alpha}}{\widehat{\gamma}^2}$$

- MM estimator:

$$\widehat{\alpha} = \frac{\bar{X}^2}{S^2}, \quad \widehat{\gamma} = \frac{\bar{X}}{S^2}$$

Exercise: Binomial models with unknown totals

Suppose

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Binomial}(\beta, p)$$

where $\theta = (\beta, p)$ is unknown. Find the MM estimator for θ

-
-
-

Example: Mixture of normals

Suppose

$$X_1, \ldots, X_n \stackrel{i.i.d.}{\sim} \alpha N(m_1, \sigma^2) + (1-\alpha) N(m_2, \sigma^2)$$

where $m_1, m_2 \in \mathcal{R}$, $\alpha \in (0,1)$ are unknown and σ is known.

- This model is not identifiable
- Restrict $\alpha < 1/2$
- The first 3 moments are

$$\begin{aligned} \mu_1 &= \alpha m_1 + (1-\alpha) m_2, \\ \mu_2 &= \alpha(m_1^2 + \sigma^2) + (1-\alpha)(m_2^2 + \sigma^2), \\ \mu_3 &= \alpha(m_1^3 + 3 m_1 \sigma^2) + (1-\alpha)(m_2^3 + 3 m_2 \sigma^2) \end{aligned}$$

- The last two equations can be reduced to

$$\begin{aligned} \mu_2 - \sigma^2 &= \alpha m_1^2 + (1-\alpha) m_2^2, \\ \mu_3 - 3\mu_1 \sigma^2 &= \alpha m_1^3 + (1-\alpha) m_2^3 \end{aligned}$$

- If these equations have a unique solution (m_1, m_2, α) then the MM

Maximum Likelihood

The *maximum likelihood method* is the most popular method for deriving estimators in statistical inference

Definition

Let $X \in \mathcal{X}$ be a sample with a p.d.f. f_θ w.r.t. a σ-finite measure ν, where $\theta \in \Theta \subset \mathcal{R}^k$.

1. The density of X, evaluate at the observed value $X = x \in \mathcal{X}$ and viewed as a function of θ, is called the *likelihood function* and denoted by $\ell(\theta) = f_\theta(X)$
2. A $\hat{\theta} \in \Theta$ satisfying $\ell(\hat{\theta}) = \max_{\theta \in \Theta} \ell(\theta)$ is called a *maximum likelihood estimate* (MLE) of θ. If $\hat{\theta}$ is a Borel function of X a.e. ν, then $\hat{\theta}$ is called a *maximum likelihood estimator* (MLE) of θ
3. Let g be a Borel function from Θ to $\mathcal{R}^p$, $p \leq k$. If $\hat{\theta}$ is an MLE of θ, then $\hat{\vartheta} = g(\hat{\theta})$ is defined to be an MLE of $\vartheta = g(\theta)$

Remark. In JS, the MLE is defined as $\hat{\theta} = \underset{\theta \in \bar{\Theta}}{\operatorname{argmax}} \ell(\theta)$, where $\bar{\Theta}$ is the closure of Θ. We use the above one to be consistent with most textbooks

Likelihood function

- The likelihood function $\ell(\theta)$ is a "statistic" of infinite dimension
  - The density $f_\theta(x)$ for any fixed θ gives a pre-experimental summary of our uncertainty about where X will fall
  - The likelihood $\ell(\theta) = f_\theta(X)$ gives a post-experimental summary of how likely it is that model P_θ produced the observed X
  - Sometimes the likelihood function is written as $\ell(\theta; x)$ to indicate the observed value of X
- The likelihood function establishes a preference among the possible parameter values given data $X = x$:
  - A parameter values θ_1 with larger likelihood is better than parameter value θ_2 with smaller likelihood, in the sense that the model P_{θ_1} provides a better fit to the observed data than P_{θ_2}
  - This leads to the introduction of the MLE
- The likelihood function is also of considerable importance in Bayesian analysis

Maximum Likelihood Estimate

- The main theoretical justification for MLE's is provided in the theory of asymptotic efficiency considered later
- According to our definition, an MLE may not exist
- There may be multiple MLE's
- If Θ contains finitely many points, an MLE exists and can always be obtained by comparing finitely many values $\ell(\theta)$, $\theta \in \Theta$
- An MLE may not have an explicit form

Maximum Likelihood Estimate (Cont.)

- The log-likelihood function $\log \ell(\theta)$ is often more convenient to work with
- If $\ell(\theta)$ is differentiable on Θ°, the interior of Θ, then possible candidates for MLE's are the values of $\theta \in \Theta^\circ$ satisfying the *likelihood equation* $\frac{\partial \log \ell(\theta)}{\partial \theta} = 0$
  - A root of the likelihood equation may be local or global minima or maxima, or simply stationary points
  - Extrema may also occur at the boundary of Θ or when $\|\theta\| \to \infty$
- If $\ell(\theta)$ is not differentiable, then extrema may occur at non-differentiable or discontinuity points of $\ell(\theta)$. In this case, it is important to analyze the entire likelihood function to find its maxima.

Example 3.3

Let $X_1, ..., X_n$ be i.i.d. binary random variables with $P(X_1 = 1) = p \in \Theta = (0, 1)$.
When $(X_1, ..., X_n) = (x_1, ..., x_n)$ is observed, the likelihood function is

$$\ell(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{n\bar{x}}(1-p)^{n(1-\bar{x})}, \tag{2}$$

where $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$

- The likelihood equation is

$$\frac{n\bar{x}}{p} - \frac{n(1-\bar{x})}{1-p} = 0. \tag{3}$$

- **If** $0 < \bar{x} < 1$, then this equation has a unique solution $\hat{p} = \bar{x}$
    - The second-order derivative of $\log \ell(p)$ is

$$-\frac{n\bar{x}}{p^2} - \frac{n(1-\bar{x})}{(1-p)^2} < 0 \tag{4}$$

    - When p tends to 0 or 1 (the boundary of Θ), $\ell(p) \to 0$. Thus, $\bar{x}$ is the unique MLE of p

Example 3.3 (Cont.)

- If $\bar{x} = 0$, $\ell(p) = (1-p)^n$ is a strictly decreasing function of p and, therefore, its unique maximum is $\hat{p} = 0$
- If $\bar{x} = 1$, the MLE is $\hat{p} = 1$ similarly
- Combining these results, we conclude that the MLE of p is $\bar{x}$ if $\bar{x} \in (0,1)$; when $\bar{x} = 0$ or 1, a maximum of $\ell(p)$ does not exist on $\Theta = (0,1)$, although $\sup_{p\in(0,1)} \ell(p) = 1$; the MLE does not exist
- However, if $p \in (0,1)$, the probability that $\bar{x} = 0$ or 1 tends to 0 quickly as $n \to \infty$.

This example indicates that an MLE may not exist on Θ; however, this is unlikely to occur when n is large

Example 3.4: Normal Faimilies

Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$ with unknown $\theta = (\mu, \sigma^2)$, $n \geq 2$. Consider first the case where $\Theta = \mathcal{R} \times (0, \infty)$

- When $(X_1, ..., X_n) = (x_1, ..., x_n)$ is observed, the likelihood function is

$$\log \ell(\theta) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{n}{2} \log \sigma^2 - \frac{n}{2} \log(2\pi). \tag{5}$$

- The likelihood equation is

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0 \qquad \text{and} \qquad \frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{n}{\sigma^2} = 0. \tag{6}$$

- Solving the equations, we obtain $\hat{\theta} = (\bar{x}, \hat{\sigma}^2)$ where

$$\bar{x} = n^{-1} \sum_{i=1}^{n} x_i, \quad \hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Example 3.4 (Cont.)

Show $\hat{\theta} = (\bar{x}, \hat{\sigma}^2)$ is an MLE:

1. Note that Θ is an open set and $\ell(\theta)$ is differentiable everywhere
2. $\ell(\theta)$ is bounded
3. As θ tends to the boundary of Θ or $\|\theta\| \to \infty$, $\ell(\theta)$ tends to 0
4. The maxima of $\ell(\theta)$ exists in Θ and must satisfies the likelihood equation
5. Hence $\hat{\theta}$ is the unique MLE

Remark. We have avoided the calculation of the second-order derivatives above. The Hessian matrix of the log-likelihood

$$\frac{\partial^2 \log \ell(\theta)}{\partial\theta\partial\theta^\top} = -\begin{pmatrix} \frac{n}{\sigma^2} & \frac{1}{\sigma^4}\sum_{i=1}^n (x_i - \mu) \\ \frac{1}{\sigma^4}\sum_{i=1}^n (x_i - \mu) & \frac{1}{\sigma^6}\sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2\sigma^4} \end{pmatrix}$$

is negative definite when $\mu = \bar{x}$ and $\sigma^2 = \hat{\sigma}^2$.

Example 3.4'

Now consider the case where $\Theta = (0, \infty) \times (0, \infty)$, i.e., μ is known to be positive.

- If $\bar{x} > 0$, then the same argument for the previous case can be used to show that $(\bar{x}, \hat{\sigma}^2)$ is the MLE
- If $\bar{x} \leq 0$, then the first equation in the likelihood equation Eq (6) does not have a solution
  - By Eq (5), for any fixed σ^2, $\log \ell(\theta) = \log \ell(\mu, \sigma^2)$ is strictly decreasing in μ ; a maxima of $\log \ell(\mu, \sigma^2)$ with respective to μ is $\mu = 0$, regardless of σ^2
  - The MLE does not exist (in Θ)
- Thus, the MLE is

$$\hat{\theta} \begin{cases} = (\bar{x}, \hat{\sigma}^2) & \bar{x} > 0 \\ \text{does not exist} & \bar{x} \leq 0. \end{cases} \tag{7}$$

Example 3.5

Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on an interval $\mathcal{I}_\theta$ with an unknown θ. Suppose $(X_1, ..., X_n) = (x_1, \ldots, x_n)$ is observed.
Model I: $\mathcal{I}_\theta = (0, \theta)$ and $\theta > 0$, $\Theta^\circ = (0, \infty)$.

- The likelihood function is

$$\ell(\theta) = \prod_{i \le n} \theta^{-1} I_{(0,\theta]}(x_i) = \theta^{-n} I_{[x_{(n)}, \infty)}(\theta), \tag{8}$$

  where $x_{(n)} = \max(x_1, ..., x_n)$
- Note that the density is unique up to "m-a.e."
- $\ell(\theta)$ is not differentiable at $x_{(n)}$ and the method of using the likelihood equation is not applicable
- On $(0, x_{(n)})$, $\ell \equiv 0$
- On $(x_{(n)}, \infty)$, $\ell'(\theta) = -n\theta^{n-1} < 0$ for all θ
- Since $\ell(\theta)$ is strictly decreasing on $(x_{(n)}, \infty)$ and is 0 on $(0, x_{(n)})$, a unique maximum of $\ell(\theta)$ is $x_{(n)}$, which is a discontinuity point of $\ell(\theta)$
- This shows that the MLE of θ is $X_{(n)}$; unreasonable

Example 3.5 (Cont.)

Model II: $\mathcal{I}_\theta = (\theta, \theta + 1)$ with $\theta \in \mathcal{R}$.

- The likelihood function is

$$\ell(\theta) = \prod_{i \leq n} I_{(\theta,\theta+1)}(x_i) = I_{(x_{(n)}-1, x_{(1)})}(\theta), \tag{9}$$

  where $x_{(1)} = \min(x_1, ..., x_n)$
- Again, the method of using the likelihood equation is not applicable
- However, it follows from Definition 4.3 that any statistic $T(X)$ satisfying $x_{(n)} - 1 \leq T(x) \leq x_{(1)}$ is an MLE of θ

This example indicates that MLE's may not be unique and can be unreasonable.

Exercise

Let $X_1, ..., X_n$ be i.i.d. from $N(\theta, \theta^2)$ with unknown $\theta > 0$.
Find an MLE if $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$ is observed.

-
    - Here we have ignored the constant
-
-

Numerical methods

In applications, MLE's typically do not have analytic forms and some numerical methods have to be used to compute MLE's.

- The Newton-Raphson iteration method for solving $\frac{\partial \log \ell(\theta)}{\partial \theta} = \mathbf{0}$ repeatedly computes

$$\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} - \left[\frac{\partial^2 \log \ell(\theta)}{\partial \theta \partial \theta^\top} \bigg|_{\theta = \hat{\theta}^{(t)}} \right]^{-1} \frac{\partial \log \ell(\theta)}{\partial \theta} \bigg|_{\theta = \hat{\theta}^{(t)}}, \qquad (10)$$

  $t = 0, 1, \ldots$

  1. $\hat{\theta}^{(0)}$ is an initial value
  2. The Hessian matrix $\partial^2 \log \ell(\theta)/\partial\theta\partial\theta^\top$ is assumed of full rank for every $\theta \in \Theta$
  3. The rationale: at each time t, we update the current value by, we expand $\frac{\partial \log \ell(\theta)}{\partial \theta}$ around $\hat{\theta}^{(t)}$:

$$\mathbf{0} = \frac{\partial \log \ell(\theta)}{\partial \theta} \approx \frac{\partial \log \ell(\theta)}{\partial \theta} \bigg|_{\theta = \hat{\theta}^{(t)}} + \left[\frac{\partial^2 \log \ell(\theta)}{\partial \theta \partial \theta^\top} \bigg|_{\theta = \hat{\theta}^{(t)}} \right] (\theta - \hat{\theta}^{(t)})$$

- If the iteration converges, then the limit or $\hat{\theta}^{(t)}$ with a sufficiently large t is a numerical approximation to a solution

Numerical methods (Cont.)

- If $\left.\frac{\partial^2 \log \ell(\theta)}{\partial\theta\partial\theta^\top}\right|_{\theta=\hat{\theta}^{(t)}}$ is replaced by by $\left.\left\{E\left(\frac{\partial^2 \log \ell(\theta)}{\partial\theta\partial\theta^\top}\right)\right\}\right|_{\theta=\hat{\theta}^{(t)}}$, where the expectation is taken under P_θ, then the method is known as the *Fisher-scoring method*
- In some applications, ideal observations lead to closed-form MLE but a part of such ideal observations is missing. For such problems, the *EM algorithm* will iteratively
  - compute the **expectation** of the log-likelihood w.r.t. the missing data under the population given by the current $\hat{\theta}^{(t)}$, and
  - compute a new $\hat{\theta}^{(t+1)}$ as the **maxima** of this expectation
- In modern applications, the sample size n is so large that optimizing the likelihood function is intractable. In this case, it is popular to use the *stochastic gradient ascent*, which use a random sub-sample from the data to compute the gradient

MLE in Exponential Families

Suppose that X has a distribution from a natural exponential family so that the likelihood function is

$$\ell(\eta) = \exp\{\eta^\top T(x) - \zeta(\eta)\}h(x), \tag{11}$$

where $\eta \in \Xi$ is a vector of unknown parameters.

- The likelihood equation is then

$$\frac{\partial \log \ell(\eta)}{\partial \eta} = T(x) - \frac{\partial \zeta(\eta)}{\partial \eta} = 0, \tag{12}$$

  which has a unique solution $T(x) = \partial\zeta(\eta)/\partial\eta$, assuming that $T(x)$ is in the range of $\partial\zeta(\eta)/\partial\eta$.
- Note that

$$\frac{\partial^2 \log \ell(\eta)}{\partial \eta \partial \eta^\top} = -\frac{\partial^2 \zeta(\eta)}{\partial \eta \partial \eta^\top} = -\mathrm{Var}(T) \tag{13}$$

  Since $\mathrm{Var}(T)$ is positive definite, $-\log \ell(\eta)$ is convex in η and $T(x)$ is the unique MLE of the parameter $\mu(\eta) = \partial\zeta(\eta)/\partial\eta$
- The function $\mu(\eta)$ is one-to-one so that μ^{-1} exists.
- By the definition, the MLE of η is $\hat{\eta} = \mu^{-1}(T(x))$.

MLE in Exponential Families (Cont.)

- If the distribution of X is in a general exponential family and the likelihood function is

$$\ell(\theta) = \exp\{[\eta(\theta)]^\top T(x) - \xi(\theta)\} h(x), \tag{14}$$

  then the MLE of θ is $\hat{\theta} = \eta^{-1}(\hat{\eta})$, if η^{-1} exists and $\hat{\eta}$ is in the range of $\eta(\theta)$.
- $\hat{\theta}$ is also the solution of the likelihood equation

$$\frac{\partial \log \ell(\theta)}{\partial \theta} = \frac{\partial \eta(\theta)}{\partial \theta} T(x) - \frac{\partial \xi(\theta)}{\partial \theta} = 0. \tag{15}$$

Tutorial

1. Suppose X has an exponential family distribution with density $p_\theta(x) = h(x)e^{\eta(\theta)T(x)-A(\theta)}$. Derive the mean and variance formulas
$$\mathrm{E}_\theta[T(X)] = \frac{A'(\theta)}{\eta'(\theta)}, \quad \mathrm{V}_\theta[T(X)] = \frac{A''(\theta)}{[\eta'(\theta)]^2} - \frac{\eta''(\theta)A'(\theta)}{[\eta'(\theta)]^3}$$
2. Let X and Y be two random variables such that Y has the binomial distribution $Bi(\pi, N)$ and, given $Y = y, X$ has the binomial distribution $Bi(p, y)$.
(a) Suppose that $p \in (0,1)$ and $\pi \in (0,1)$ are unknown and N is known. Show that (X, Y) is minimal sufficient for (p, π).
(b) Suppose that π and N are known and $p \in (0,1)$ is unknown. Show whether X is sufficient for p and whether Y is sufficient for p
3. Let $X_1, \ldots, X_n$ be i.i.d. random variables having the Lebesgue p.d.f.
$$f_\theta(x) = \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^4 - \xi(\theta)\right\}$$
where $\theta = (\mu, \sigma) \in \Theta = \mathcal{R} \times (0, \infty)$. Show that $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is an exponential family, where P_θ is the joint distribution of $X_1, \ldots, X_n$ and that the statistic below is minimal sufficient for $\theta \in \Theta$:
$$T = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i^3, \sum_{i=1}^n X_i^4\right)$$