ST5215 Advanced Statistical Theory, Lecture 14

HUANG Dongming

National University of Singapore

6 Oct 2020

Overview

- A quick review
- Cramér-Rao Lower Bound and Fisher's Information

What we have learned

- Measure theory: measurability, integration, Radon-Nikodym derivative, conditional expectation, law of large numbers, CLT
- Extract information from the data generated by experiments and observations
- To capture the uncertainty in data, we need models; $P_\theta \in \mathcal{P}_\Theta$
- Based on the data, we obtain an estimator $\hat{\theta}$
- Construct estimators: method of moments, MLE, Bayes estimators
- Summary of data: sufficiency; minimal sufficiency; completeness
- Evaluate estimators by its risk $R_{\hat{\theta}}(\theta)$
  - Admissible estimators under convex loss: Rao-Blackwell Theorem
  - UMVUE: Lehmann-Scheffé Theorem
  - Minimaxity: sufficient conditions
- Asymptotics: consistency, efficiency, limiting distribution

Today we will learn a powerful tool, *Cramér-Rao Lower Bound*

- Assessing the variance of estimators
- Insight for the theory of asymptotic efficiency

Fisher information

- Suppose $\mathcal{P} = \{f_\theta : \theta \in \Theta\}$ where $f_\theta(x)$ is a p.d.f. with parameter θ w.r.t. ν and Θ is an open subset of $\mathcal{R}$
- Suppose for any $\theta \in \Theta$, $\frac{\partial f_\theta(x)}{\partial \theta}$ exists and is finite, P_θ-a.s.
- Let X be a sample from $P_\theta \in \mathcal{P}$
- To measure the amount of information that an observation X carries about θ, we look at the *Fisher information* defined as

$$\begin{aligned} I(\theta) &= E\left(\frac{\partial}{\partial \theta} \log f_\theta(X)\right)^2 \\ &= \int \left(\frac{\partial}{\partial \theta} \log f_\theta(X)\right)^2 f_\theta(X)\, \mathrm{d}\nu(x). \end{aligned}$$

- The greater $I(\theta)$ is, the easier it is to distinguish θ from neighboring values and, therefore, the more accurately θ can be estimated
- Under some conditions, $I(\theta) = -E\left[\frac{\partial^2}{\partial \theta^2} \log f_\theta(X)\right]$ and $I(\theta) = \mathrm{Var}(\frac{\partial}{\partial \theta} \log f_\theta(X))$

Example: Poisson Families

Suppose $(X_1, \ldots, X_n)$ is a i.i.d. sample from a Poisson distribution $\mathcal{P}(\lambda)$. Then

- The joint p.d.f. w.r.t. the counting measure is

$$f_\lambda(x) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} \exp(-\lambda)$$

- $\log f_\lambda(x) = \sum_i x_i \log(\lambda) - n\lambda - \sum_i \log(x_i!)$
- $\frac{\partial}{\partial \lambda} \log f_\lambda(x) = \frac{\sum_{i=1}^n x_i}{\lambda} - n$
- $I(\lambda) = \mathrm{Var}\left(\frac{\sum_{i=1}^n X_i}{\lambda}\right) = \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda}$

Example: Normal Families with Known Variance

Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with an unknown $\mu \in \mathcal{R}$ and a known σ^2.

- The joint Lebesgue p.d.f. is

$$f_\mu(x) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2\right)$$

- Then

$$\frac{\partial}{\partial\mu} \log f_\mu(x) = \sum_{i=1}^n (x_i - \mu)/\sigma^2 \tag{1}$$

- $I(\mu) = \mathrm{Var}\left(\frac{\sum_{i=1}^n (X_i - \mu)}{\sigma^2}\right) = n\sigma^2/\sigma^4 = n/\sigma^2$.

Property of Fisher Information

1. $I(\theta)$ depends on the particular parameterization:
   - If $\theta = \psi(\eta)$ and ψ is differentiable, then the Fisher information that X contains about η is
   $$\tilde{I}(\eta) = \psi'(\eta)^2 I(\psi(\eta)), \tag{2}$$
   where $I(\theta)$ is the Fisher information about θ.
2. Let X and Y be independent with the Fisher information about θ $I_X(\theta)$ and $I_Y(\theta)$, respectively. Then, the Fisher information about θ contained in (X, Y) is $I_X(\theta) + I_Y(\theta)$.
   - In particular, if $X_1, \ldots, X_n$ are i.i.d. and $I_1(\theta)$ is the Fisher information about θ contained in a single X_i, then the Fisher information about θ contained in $X_1, \ldots, X_n$ is $nI_1(\theta)$
3. Suppose that f_θ is twice differentiable in θ and that
$$\int \frac{\partial^2}{\partial \theta^2} f_\theta(x) I_{f_\theta(x)>0} d\nu = 0 \tag{3}$$
Then $I(\theta) = -E\left[\frac{\partial^2}{\partial \theta^2} \log f_\theta(X)\right]$

Cramér-Rao Lower Bound

Theorem

Suppose $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ *satisfies the following conditions*

- Θ *is an open set in* $\mathcal{R}$*;* P_θ *has a p.d.f.* f_θ *w.r.t. a measure* ν *for all* $\theta \in \Theta$
- f_θ *is differentiable as a function of* θ *and satisfies*

$$0 = \frac{\partial}{\partial \theta} \int f_\theta(x) d\nu = \int \frac{\partial}{\partial \theta} f_\theta(x) d\nu, \quad \theta \in \Theta \tag{4}$$

Suppose that $g(\theta)$ *is a differentiable function.*
Let X *be a sample from* $P \in \mathcal{P}$*. Suppose* $T(X)$ *is an unbiased estimator of* $g(\theta)$ *such that*

$$g'(\theta) = \frac{\partial}{\partial \theta} \int T(x) f_\theta(x) d\nu = \int T(x) \frac{\partial}{\partial \theta} f_\theta(x) d\nu, \quad \theta \in \Theta \tag{5}$$

Then $\mathrm{Var}(T(X)) \geq \frac{g'(\theta)^2}{I(\theta)}$ *where* $I(\theta) > 0$ *for any* $\theta \in \Theta$

Proof:

- By the covariance inequality, we have

$$\mathrm{Cov}\left(T(X), \frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2 \leq \mathrm{Var}[T(X)]\mathrm{Var}[\frac{\partial}{\partial\theta}\log f_\theta(X)]$$

- By Eq (4), $E\frac{\partial}{\partial\theta}\log f_\theta(X) = 0$
- $\mathrm{Var}[\frac{\partial}{\partial\theta}\log f_\theta(X)] = E\left([\frac{\partial}{\partial\theta}\log f_\theta(X)]^2\right) = I(\theta)$
- $\mathrm{Cov}(T, \frac{\partial}{\partial\theta}\log f_\theta(X)) = E\left[T\frac{\partial}{\partial\theta}\log f_\theta(X)\right] = \int T(x)\frac{\partial}{\partial\theta}f_\theta(x)\,\mathrm{d}\nu$
- By Eq (5), the last display $= \frac{\partial}{\partial\theta}E[T] = g'(\theta)$
- We conclude that

$$g'(\theta)^2 \leq \mathrm{Var}(T)I(\theta)$$

Remark:

- Equations (4) and (5) are the regularity conditions for the results in Cramér-Rao lower bound and has to be checked
- Typically, they do not hold if the set $\{x : f_\theta(x) > 0\}$ depends on θ

Remarks on C-R Lower Bound

- The theorem is also known as *the Information Inequality*.
- The Cramér-Rao lower bound is not affected by any one-to-one reparameterization.
- If an unbiased estimator $T(X)$ of $g(\theta)$ achieves the C-R lower bound, then it is a UMVUE.
  - However, this is not an effective way to find a UMVUE because the Cramér-Rao lower bound is typically *not sharp*.
- Under some regularity conditions, we can show that (left for exercise) there exists an estimator $T(X)$ that attains the C-R lower bound for all $\theta \Leftrightarrow f_\theta$ is of the form $\exp[\eta(\theta)^\top T(x) - \xi(\theta)]h(x)$

Example: Normal Families

Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with an unknown $\mu \in \mathcal{R}$ and a known σ^2.

- We have showed $I(\mu) = n/\sigma^2$.
- Consider the estimation of μ.
- Note that $\bar{X}$ is unbiased and has variance σ^2/n.
- So $\bar{X}$ attains the Cramér-Rao lower bound and it is a UMVUE.

Example: Normal Families (Cont.)

Model

$X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with an unknown $\mu \in \mathcal{R}$ and a known σ^2. $I(\mu) = n/\sigma^2$.

- Consider now the estimation of $\eta = \mu^2$.
- Since $E\left(\bar{X}^2\right) = \mu^2 + \sigma^2/n$ and $\bar{X}$ is sufficient and complete, the UMVUE of η is $h(\bar{X}) = \bar{X}^2 - \sigma^2/n$ (by Lehmann-Scheffé Theorem).
- A straightforward calculation shows that (left for exercise)

$$\mathrm{Var}(h(\bar{X})) = \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n^2}. \tag{6}$$

- Since $g'(\mu) = 2\mu$, the Cramér-Rao lower bound is $4\mu^2\sigma^2/n$.
- Hence $\mathrm{Var}(h(\bar{X}))$ does not attain the Cramér-Rao lower bound.

Extension to Multi-parameter Case

Let $X = (X_1, ..., X_n)$ be a sample from $P \in \mathcal{P} = \{p(x, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, where Θ is an open set in $\mathcal{R}^k$. Assume similar regularity conditions as before.

- The $k \times k$ matrix

$$I(\theta) = E\left\{ \frac{\partial}{\partial \theta} \log f_\theta(X) \left[\frac{\partial}{\partial \theta} \log f_\theta(X) \right]^\top \right\} \tag{7}$$

  is called the *Fisher information matrix*, where

$$\frac{\partial}{\partial \theta} \log f_\theta(X) = \left(\frac{\partial}{\partial \theta_1} \log f_\theta(X), \ldots, \frac{\partial}{\partial \theta_k} \log f_\theta(X) \right)^\top \tag{8}$$

- Suppose that X has the p.d.f. f_θ that is twice differentiable in θ and that

$$0 = \frac{\partial}{\partial \theta} \int \frac{\partial}{\partial \theta^\tau} f_\theta(x) d\nu = \int \frac{\partial^2}{\partial \theta \partial \theta^\tau} f_\theta(x) d\nu, \quad \theta \in \Theta. \tag{9}$$

  Then

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

Multivariate C-R Lower Bound

When θ is k-dimensional, $g : \Theta \mapsto \mathcal{R}$, the inequality in the Cramér-Rao Lower Bound becomes

$$\mathrm{Var}(T(X)) \geq \left[\frac{\partial}{\partial\theta} g(\theta)\right]^\top [I(\theta)]^{-1} \frac{\partial}{\partial\theta} g(\theta),$$

where the gradient $\frac{\partial}{\partial\theta} g(\theta) = (\frac{\partial}{\partial\theta_1} g(\theta), \dots, \frac{\partial}{\partial\theta_k} g(\theta))^\top$. Again, we assume the regularity conditions hold.

- By the covariance inequality, for any $\mathbf{c} \in \mathcal{R}^k$,

$$\mathrm{Var}(T)\mathrm{Var}(\mathbf{c}^\top \frac{\partial \log f_\theta(X)}{\partial\theta}) \geq \left[\mathrm{Cov}(T, \mathbf{c}^\top \frac{\partial \log f_\theta(X)}{\partial\theta})\right]^2 . \quad (10)$$

- Use $\mathbf{a} = \mathrm{Cov}(\frac{\partial}{\partial\theta} \log f_\theta(X), T(X))$ to simplify notations
- The LHS is $\mathrm{Var}(T) \left(\mathbf{c}^\top I(\theta)\mathbf{c}\right)$
- The RHS is $(\mathbf{c}^\top \mathbf{a})^2$
- Choose $\mathbf{c} = [I(\theta)]^{-1}\mathbf{a}$. Use the regularity to replace $\mathbf{a}$ by $\frac{\partial}{\partial\theta} g(\theta)$

Example: Normal Families

Let $X_1, ..., X_n$ be i.i.d. $\sim N(\mu, \nu)$. Let $\theta = (\mu, \nu)$. Then

$$\log f_\theta(\mathbf{x}) = -\frac{1}{2\nu}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{2}\log(2\pi\nu). \tag{11}$$

It can be calculated that

$$\begin{aligned}
\frac{\partial^2}{\partial\mu^2}\log f_\theta(\mathbf{x}) &= -\frac{n}{\nu}, \\
\frac{\partial^2}{\partial\nu^2}\log f_\theta(\mathbf{x}) &= -\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\nu^3} + \frac{n}{2\nu^2}, \\
\frac{\partial^2}{\partial\nu\partial\mu}\log f_\theta(\mathbf{x}) &= -\frac{\sum_{i=1}^{n}(x_i - \mu)}{\nu^2}.
\end{aligned}$$

Thus, the Fisher information matrix about θ contained in $X_1, ..., X_n$ is

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^\tau}\log f_\theta(X)\right] = \begin{pmatrix} \frac{n}{\nu} & 0 \\ 0 & \frac{n}{2\nu^2} \end{pmatrix}. \tag{12}$$

Exercise

Let $X_1, ..., X_n$ be i.i.d. $\sim N(\mu, \nu)$. Let $\theta = (\mu, \nu)$.
Find the C-R lower bound for $\mu^2 - 2\nu$.

Fisher information and exponential families

Proposition

Suppose that the distribution of X is from an exponential family $\{f_\theta : \theta \in \Theta\}$, i.e., the p.d.f. of X w.r.t. a σ-finite measure is

$$f_\theta(x) = exp\big\{[\eta(\theta)]^\top T(x) - \xi(\theta)\big\} h(x), \tag{13}$$

where Θ is an open subset of $\mathcal{R}^k$.

(i) *For any T with $E|T(X)| < \infty$, it holds that*

$$\frac{\partial}{\partial\theta} \int T(x) f_\theta(x) d\nu = \int T(x) \frac{\partial}{\partial\theta} f_\theta(x) d\nu, \quad \theta \in \Theta$$

and

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^\top} \log f_\theta(X)\right]. \tag{14}$$

This is a direct consequence of Theorem 2.1 (of the textbook).

Proposition (Cont.)

(ii) *If $\underline{I}(\eta)$ is the Fisher information matrix for the natural parameter η, then the variance-covariance matrix $\mathrm{Var}(T) = \underline{I}(\eta)$.*

Proof:

(ii) The p.d.f. under the natural parameter η is

$$f_\eta(x) = \exp\left\{\eta^\top T(x) - \zeta(\eta)\right\} h(x). \tag{15}$$

From Theorem 2.1 of (the textbook), $E[T(X)] = \frac{\partial}{\partial\eta}\zeta(\eta)$. The result follows from

$$\frac{\partial}{\partial\eta}\log f_\eta(x) = T(x) - \frac{\partial}{\partial\eta}\zeta(\eta). \tag{16}$$

Proposition (Cont.)

(iii) *Let $\psi = E[T(X)]$. Suppose $\bar{I}(\psi)$ is the Fisher information matrix for the parameter ψ, then $\mathrm{Var}(T) = [\bar{I}(\psi)]^{-1}$.*

(iii)

- Since $\psi = E[T(X)] = \frac{\partial}{\partial \eta}\zeta(\eta)$,

$$\underline{I}(\eta) = \frac{\partial \psi^\top}{\partial \eta}\bar{I}(\psi)\left(\frac{\partial \psi^\top}{\partial \eta}\right)^\top = \frac{\partial^2}{\partial\eta\partial\eta^\top}\zeta(\eta)\bar{I}(\psi)\left[\frac{\partial^2}{\partial\eta\partial\eta^\top}\zeta(\eta)\right]^\top.$$

- By Theorem 2.1 (of the textbook) (see also exercise 1 in Tutorial 9) and the result in (ii),

$$\frac{\partial^2}{\partial\eta\partial\eta^\top}\zeta(\eta) = \mathrm{Var}(T) = \underline{I}(\eta). \tag{17}$$

- Hence

$$\bar{I}(\psi) = [\underline{I}(\eta)]^{-1}\underline{I}(\eta)[\underline{I}(\eta)]^{-1} = [\underline{I}(\eta)]^{-1} = [\mathrm{Var}(T)]^{-1}. \tag{18}$$

Midterm Exam

Source of the questions:

- Q1 (a) is from a definition in Lecture 3; Q1(b) is from an example in Lecture 4
- Q2 is from Lecture 5 (handwritten note)
- Q3(a,b) is a simplified version of Exercise 3 in Tutorial 9; Q3(c) is from Lecture 6 (Page 9)
- Q4 is from an example in Lecture 12 (Page 7)
- Q5(b) is a modified version of an exercise in Lecture 10 (Page 13)

General Suggestions

Ask questions

- To yourself (most important):
  - Do I understand the **definition**? Can I find a simple but nontrivial example that satisfies/dissatisfies the definition?
  - What does the **theorem** say? What are the conditions? Does the result fail to hold if one condition is not satisfied? Where has this theorem been applied?
  - Can I reproduce the **example** or the solution to an **exercise**? What is the key result used in the solution?
  - If I need to design a set of exam questions, what will they be like?
- To instructors: office hours, email
- To your classmates: Forums on LumiNUS

Have some exercises

- Tutorial exercises
- Examples in other textbooks