6.434/16.391 Statistics for Engineers and Scientists

Lecture 9 10/02/2013

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology

Lecture 9 10/02/2013

POINT ESTIMATION

Point estimation: concepts

- Observe $X_1, X_2, \ldots, X_n$ that are not necessarily independent or identically distributed. The distribution of $X_1, X_2, \ldots, X_n$ depends on unknown parameter $\theta \in \Theta$, where Θ is a known set
  - e.g., mean of normal distribution can be unknown

- The problem of estimating θ is the problem of point estimation

Point estimation: concepts

- Consider sample space Ω . The experiment essentially chooses an $\omega \in \Omega$ such that the observations are

$$X_1(\omega) = x_1,\, X_2(\omega) = x_2,\, \ldots, X_n(\omega) = x_n$$

with distribution function

$$F_{X_1,\, X_2,\, \ldots,\, X_n}(x_1,\, x_2,\, \ldots,\, x_n \mid \theta)$$

We wish to estimate θ

Estimator and estimate

- Definition: Statistic $T_n(X_1, X_2, \ldots, X_n)$, a function of the random variables that are being observed, is an <u>estimator</u> of θ

- $T_n(x_1, x_2, \ldots, x_n)$ is called <u>estimate</u> of θ

Unbiased estimator

- Definition: $T_n(X_1,\, X_2,\, \ldots,\, X_n)$ is an unbiased estimator of θ if

$$\mathbb{E}\,\{T_n(X_1,\, X_2,\, \ldots,\, X_n)\} = \theta, \quad \forall\theta \in \Theta$$

  – Many estimators are unbiased

- Definition: Bias of T_n is defined as

$$b_n(\theta) = \mathbb{E}\{T_n(X_1,\, X_2,\, \ldots,\, X_n)\} - \theta$$

  – Often T_n will not be unbiased, but some T_n satisfies $\text{bias} \to 0$ as $n \to \infty$, e.g., $T_n = S_n + \frac{1}{n}$, where S_n is unbiased

Consistent sequence of estimator

- Definition: $T_n(X_1, X_2, \ldots, X_n)$ is asymptotically unbiased estimator of θ if $\lim_{n\to\infty} b_n(\theta) = 0, \quad \theta \in \Theta$

- Definition: $T_1, T_2, \ldots$ is called consistent sequence of estimator for θ if and only if

$$T_n(X_1, X_2, \ldots, X_n) \xrightarrow{P} \theta, \quad \forall\theta \in \Theta$$

  – Convergence in probability

Example 1

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with mean μ and variance σ^2, both of which are unknown. Let

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

$$U_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

Example 1

- Since

$$\mathbb{E}\{\bar{X}_n\} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{X_i\}$$
$$= \mu$$

$\bar{X}_n$ is an unbiased estimator of μ

- Next consider

$$\mathbb{E}\{S_n^2\} = \frac{1}{n-1}\mathbb{E}\left\{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right\}$$

and use the equation

$$\sum (X_i - \bar{X}_n)^2 = \sum X_i^2 - 2\sum X_i\bar{X}_n + \sum \bar{X}_n^2$$
$$= \sum X_i^2 - n\bar{X}_n^2$$

Example 1

- We have

$$\mathbb{E}\left\{\sum {X_i}^2\right\} = \sum \mathbb{E}\{X_i^2\}$$
$$= n(\sigma^2 + \mu^2)$$

and

$$\begin{aligned}
\mathbb{E}\left\{\sum \bar{X}_n^2\right\} &= \mathbb{E}\left\{\frac{1}{n^2}\sum_i \sum_j X_i X_j\right\} \\
&= \frac{1}{n^2}\mathbb{E}\left\{\sum_{i=j} X_i^2 + \sum_{i\neq j}\sum_j X_i X_j\right\} \\
&= \frac{1}{n^2}\left[n(\sigma^2+\mu^2) + n(n-1)\mu^2\right] \\
&= \frac{1}{n}(\sigma^2 + n\mu^2) \\
&= \frac{\sigma^2}{n} + \mu^2
\end{aligned}$$

Example 1

- Therefore,

$$\begin{aligned}\mathbb{E}\{S_n^2\} &= \frac{1}{n-1}\left[n(\sigma^2+\mu^2) - n(\frac{\sigma^2}{n}+\mu^2)\right] \\ &= \frac{1}{n-1}(n-1)\sigma^2 \\ &= \sigma^2\end{aligned}$$

i.e., S_n^2 is an unbiased estimator of σ^2

Example 1

- Since

$$\mathbb{E}\{U_n^2\} = \frac{n-1}{n}\mathbb{E}\{S_n^2\}$$
$$= \frac{n-1}{n}\sigma^2$$

U_n is a biased estimator of σ^2. The bias

$$b_n(\theta) = \frac{n-1}{n}\sigma^2 - \sigma^2$$
$$= -\frac{1}{n}\sigma^2 \to 0$$

as $n \to \infty$, so U_n is asymptotically unbiased

Observations

- We use $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ to estimate σ^2. However, if we knew $\mathbb{E}\{X_i\} = \mu$, then we could use

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

  which is an unbiased estimator of σ^2. If we don't know μ, we in effect replace μ by $\bar{X}$. This requires division by $n-1$ instead of n in order to achieve unbiasedness.

- This is called "loss of one degree of freedom"

Example 2

- Suppose we observe $X_1, X_2, \ldots, X_n$ which are not necessarily independent or identically distributed, and that their joint distribution depends on the unknown parameter $\theta \in \Theta$
- Suppose we find statistics

$$T_1(X_1, X_2, \ldots, X_n)$$
$$T_2(X_1, X_2, \ldots, X_n)$$
$$\vdots$$
$$T_n(X_1, X_2, \ldots, X_n)$$

such that

$$\mathbb{E}\left\{T_i(X_1, X_2, \ldots, X_n)\right\} = \theta, \quad \forall\theta \in \Theta$$

Example 2

- For fixed constants $\omega_1, \omega_2, \ldots, \omega_k$, $-\infty < \omega_i < \infty$,

$$\mathbb{E}\left\{\sum_{i=1}^{k} \omega_i T_i(X_1, X_2, \ldots, X_n)\right\} = \sum_{i=1}^{k} \omega_i \, \mathbb{E}\left\{T_i(X_1, X_2, \ldots, X_n)\right\}$$
$$= \left(\sum_{i=1}^{k} \omega_i\right) \theta$$

Thus, $\sum_{i=1}^{k} \omega_i T_i(X_1, X_2, \ldots, X_n)$ is an unbiased estimator of θ if and only if

$$\sum_{i=1}^{k} \omega_i = 1$$

Lecture 9   10/02/2013

METHOD OF MOMENTS

Method of moments

- Definition: Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with

$$X_i \sim F(x|\boldsymbol{\theta})$$

for some fixed $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$. Let

$$\mu_1(\boldsymbol{\theta}), \mu_2(\boldsymbol{\theta}), \ldots, \mu_d(\boldsymbol{\theta})$$

be the first d moments, i.e.,

$$\mu_j = \mathbb{E}\left\{X^j\right\}, \quad j = 1, \ldots, d$$

- Let $m_1, m_2, \ldots, m_d$ be

$$m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j, \quad j = 1, \ldots, d$$

Method of moments

- The method of moments (MM) principle prescribes that we estimate θ by the solution of

$$\mu_j = m_j, \;\; j = 1, \ldots, d$$

- Note: MM attempts to equate the first r moments with the first r sample moments

- Justification: If X_i are i.i.d., then X_i^j are also i.i.d. By law of large numbers,

$$m_j \to \mathbb{E}\left\{X^j\right\} = \mu_j, \quad \text{a.s.}$$

  – consistent sequence of estimators

Example 1

- Let $X_1, X_2, \ldots, X_n \; (n \geq 2)$ be i.i.d. with $X_i \sim N(\mu, \sigma^2)$, where μ and σ^2 are unknown, and

$$(\mu, \sigma^2) \in \Theta = \left\{(\mu, \sigma^2) : \; \mu \in \mathbb{R}, \; \sigma^2 \in \mathbb{R}^+\right\}$$

  Find MME for μ and σ^2

- Here $r = 2$, $\mu_1 = m_1$, and $\mu_2 = m_2$, i.e.,

$$\mu_1 = \mathbb{E}\{X_1\} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\mu_2 = \mathbb{E}\{X_1^2\} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Example 1

- We have

$$\hat{\mu} = \bar{X}$$

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \hat{\mu}^2 \qquad \sigma^2 = \mathbb{E}\{X^2\} - \mathbb{E}^2\{X\}$$

$$= \mu_2 - \mu_1^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- Note that in this example, MME for μ and σ^2 are also maximum likelihood estimator (MLE) for μ and σ^2
  - MLE will be introduced in the next section

Example 2

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. with $X_i \sim \text{gamma}(\alpha, 1/\lambda)$, i.e.,

$$f(x|\alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$$

for $x > 0,\ \alpha > 0,\ \lambda > 0$. Let $\boldsymbol{\theta} = (\alpha, \lambda)$. We have

$$\mathbb{E}\{X\} = \frac{\alpha}{\lambda}$$

$$\mathbb{E}\{X^2\} = \frac{\alpha(1+\alpha)}{\lambda^2}$$

For MME,

$$\frac{\hat{\alpha}}{\hat{\lambda}} = \bar{X}$$

$$\frac{\hat{\alpha}}{\hat{\lambda}^2} + \frac{\hat{\alpha}^2}{\hat{\lambda}^2} = \frac{1}{n} \sum X_i^2$$

Example 2

- Then we have $\hat{\alpha} = \hat{\lambda}\bar{X}$. Thus,

$$\frac{\bar{X}}{\widehat{\lambda}} + \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

$$\begin{aligned} \frac{1}{\widehat{\lambda}} &= \frac{1}{\bar{X}}\left[\frac{1}{n}\sum_i X_i^2 - \bar{X}^2\right] \\ &= \frac{1}{\bar{X}}\left[\frac{1}{n}\sum_i (X_i - \bar{X})^2\right] \\ &= \frac{\widehat{\sigma^2}}{\bar{X}} \end{aligned}$$

Finally, the MME is given by

$$\widehat{\alpha} = \frac{\bar{X}^2}{\widehat{\sigma^2}} \quad \text{and} \quad \widehat{\lambda} = \frac{\bar{X}}{\widehat{\sigma^2}}$$

Further remarks on MME

- Basic MME requires solving as many equations as those are unknowns (the number of unknown parameters d)

- This fails sometimes

Example 3

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. with Laplace distribution, i.e.,

$$f(x|\beta) = \frac{1}{2\beta} e^{-|x|/\beta}$$

for which we have $\mathbb{E}\{X\} = 0$, i.e., the PDF is symmetric

- MME for β is derived as

$$\mathbb{E}\{X\} = \frac{1}{n} \sum_{i=1}^{n} X_i = 0$$

The equation does not yield anything about β

Example 3

- Consider the second moment

$$\mathbb{E}\{X^2\} = 2\beta^2 \qquad 2\hat{\beta}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Therefore,

$$\widehat{\beta}_{\text{MME}} = \left[\frac{1}{2}\frac{1}{n}\sum_{i=1}^{n} X_i^2\right]^{\frac{1}{2}}$$

- We can also derive the MLE as

$$\widehat{\beta}_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n} |X_i|$$

Thus, MME and MLE are not the same in general

Example 4

- Illustrate the dependence of MME on how we transform the data
- Let $X_1, X_2, \ldots, X_n$ be i.i.d. log-normal random variables, i.e.,
$$\ln X_i \sim N(\mu, \sigma^2)$$
Find MME for μ and σ^2
- We have
$$\mathbb{E}\{X\} = e^{\mu + \frac{\sigma^2}{2}}$$
$$\mathbb{E}\{X^2\} = \left[e^{\mu + \frac{\sigma^2}{2}}\right]^2 e^{\sigma^2}$$
Then MME satisfies
$$e^{\widehat{\mu} + \frac{\widehat{\sigma^2}}{2}} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$\left[e^{\widehat{\mu} + \frac{\widehat{\sigma^2}}{2}}\right]^2 e^{\widehat{\sigma^2}} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Example 4

- Thus,

$$\widehat{\sigma^2} = \ln\left[\frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2}{\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2}\right]$$

$$\widehat{\mu} = \ln\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \frac{1}{2}\ln\left[\frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2}{\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2}\right]$$

$$= \ln\left[\frac{\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}}\right]$$

Example 4

- Alternatively, consider $Y_i = \ln X_i \sim N\left(\mu, \sigma^2\right)$. For MME,

$$\mathbb{E}\{Y_i\} = \frac{1}{n}\sum_{i=1}^{n} \ln X_i$$

$$\mathbb{E}\{Y_i^2\} = \frac{1}{n}\sum_{i=1}^{n} \left(\ln X_i\right)^2$$

Then we have

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \ln X_i$$

$$\widehat{\sigma^2} + \widehat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} \left(\ln X_i\right)^2$$

and

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} \left[\left(\ln X_i\right) - \widehat{\mu}\right]^2$$

Lecture 9 10/02/2013

MAXIMUM LIKELIHOOD ESTIMATION

Likelihood function

- Previously we talked about several properties (unbiasedness, consistency) of an estimator. How do we find good estimators?

- Definition: Let $X_1, X_2, \ldots, X_n$ be n random variables (not necessarily independent or identically distributed) with distribution $F(x_1, x_2, \ldots, x_n|\theta)$ where $\theta \in \Theta$ is unknown. The likelihood function is

$$L(\theta) = \begin{cases} f(x_1, x_2, \ldots, x_n|\theta), & \text{if } F \text{ has density } f \\ P(x_1, x_2, \ldots, x_n|\theta), & \text{if } F \text{ has prob mass Fuction } P \end{cases}$$

  - Note: Density f or PMF P is evaluated at the random point provided by the observations $x_1, x_2, \ldots, x_n$ and is considered as a function of θ

Maximum likelihood estimator

- Definition [Casella & Berger, 2002]: For each sample point $\mathbf{x}$, let $\hat{\theta}(\mathbf{x})$ be a parameter value at which $L(\theta|\mathbf{x})$ attains its maximum as a function of θ, with $\mathbf{x}$ held fixed. A maximum likelihood estimator (MLE) of the parameter θ based on a sample $\mathbf{X}$ is $\hat{\theta}(\mathbf{X})$
  - In general there may be none, or many such $\hat{\theta}$ for any given problem. Fortunately, there is often one with good properties
  - In the past, we usually held θ fixed, and let $x_1, x_2, \ldots, x_n$ vary in a density $f(x_1, x_2, \ldots, x_n|\theta)$
  - Here we hold $x_1, x_2, \ldots, x_n$ fixed and let θ vary, in which case the density $f(x_1, x_2, \ldots, x_n|\theta)$ is called likelihood function

Remarks

- $X_1, X_2, \ldots, X_n$ are independent and X_i has density $f_i(x|\theta)$. Then

$$L(\theta) = \prod_{i=1}^{n} f_i(x_i|\theta), \quad \theta \in \Theta$$

- In addition, $f_1 = f_2 = \cdots = f_n = f$, then

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta), \quad \theta \in \Theta$$

- $\widehat{\theta}$ that maximizes $L(\theta)$ also maximizes $\ln(L(\theta))$, since

$$L(\theta_1) < L(\theta_2)$$

if and only if

$$\ln L(\theta_1) < \ln L(\theta_2)$$

Example 1

- $X_1, X_2, \ldots$ are independent random variables with

$$\mathbb{P}\{X_i = 1\} = p$$
$$\mathbb{P}\{X_i = 0\} = g = 1 - p$$

Let N be the number of observations to make to obtain the first success, i.e.,

$$X_1 = X_2 = \ldots = X_{N-1} = 0, \;\; X_N = 1$$

and

$$\mathbb{P}\{N = l\} = g^{l-1} p, \;\; l = 1, 2, \ldots$$

- We observe N, say $p \in \Theta = \{\epsilon : 0 < \epsilon < 1\}$. What is MLE of p?

Example 1

- The likelihood function is

$$L(p) = g^{N-1}p$$

and

$$\begin{aligned}\frac{d}{dp}L(p) &= g^{N-1} + p(N-1)g^{N-2}(-1) \\ &= g^{N-1} - (N-1)g^{N-2}p\end{aligned}$$

further,

$$\frac{d}{dp}L(p) = 0 \Rightarrow g - (N-1)p = 0$$

$$\overbrace{g+p}^{1} - Np = 0$$

$$\Rightarrow \; p = \frac{1}{N}$$

Example 1

- Since

$$\frac{d}{dp}L(p) \geq 0, \quad p \leq \frac{1}{N}$$

  we have

$$\widehat{p} = \widehat{p}(N) = \frac{1}{N}$$

- Intuition: “If you look N trials to observe the first success, estimate the probability of success as $\frac{1}{N}$”

Example 2

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. with $X_i \sim N\left(\mu, \sigma^2\right)$ where σ^2 is known, and $\mu \in \Theta = \{z : -\infty < z < \infty\}$. The likelihood function is

$$\begin{aligned} L(\mu) &= \prod_{i=1}^{n} f(x_i|\mu) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right\} \end{aligned}$$

and

$$\ln L(\mu) = -n\ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Example 2

- Solving $\max_{\mu} \ \ln L(\mu)$ is equivalent to solving

$$\min_{\mu} \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

  and further equivalent to solving

$$\min_{\mu} \sum_{i=1}^{n} (x_i - \mu)^2$$

- Let

$$g(\mu) = \sum_{i=1}^{n} (x_i - \mu)^2$$

  we have

$$g'(\mu) = 2 \sum_{i=1}^{n} (x_i - \mu)(-1) \ \text{ and } \ g''(\mu) = 2 \sum_{i=1}^{n} (+1) = 2n$$

Example 2

- The maximum likelihood estimate is calculated as follows

$$g'(\mu) = \ 0 \ \Rightarrow \ \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$g''(\mu) > \ 0 \ \Rightarrow \ \mu \text{ above is minimum}$$

Therefore, MLE is given by

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$