Mathematical Statistics 2 Tutoring1

Seung Bong Jung

Seoul National University

September 30, 2021

Method-of-Moment Estimator (MME)

Let $X_1, \ldots, X_n$ be a random sample from a unknown distribution P. Let η be a unknown characteristic of the distribution P. Assume that it can be expressed as a function of the moments of P, say

$$\eta = g(\mu_1, \ldots, \mu_k)$$

for a given (known) function g, where $\mu_j = EX_1^j$. The method-of-moment estimator of η is defined by

$$\hat{\eta}^{\text{MME}} = g(\hat{\mu}_1, \ldots, \hat{\mu}_k),$$

where $\hat{\mu}_j = n^{-1} \sum_{i=1}^n X_i^j$, that is, $\hat{\eta}^{\text{MME}}$ is obtained by simply replacing the population moments μ_j by the corresponding sample moments $\hat{\mu}_j$.

Consistency of MME

- Consistency: An estimator $\hat{\eta}$ of η is consistent if $\hat{\eta}$ converges in probability to η for all $P \in \mathcal{P}$, where $\mathcal{P}$ is the underlying statistical model.
- Continuous mapping theorem: If g is continuous and if all elements of $\mathcal{P}$ have finite k-th moments, then

$$\hat{\eta}^{\text{MME}} = g(\hat{\mu}_1, \ldots, \hat{\mu}_k) \xrightarrow{P} g(\mu_1, \ldots, \mu_k) = \eta$$

for all $P \in \mathcal{P}$.
Note: The set in $\mathbb{R}^k$ where g is continuous needs to cover the set $\{(\mu_1(P), \ldots, \mu_k(P)) : P \in \mathcal{P}\}$.

Example of MME

2018 Midterm Problem2

Let $X_1, \ldots, X_n$ be i.i.d. with p.d.f. $f_{p,\lambda}(x)$ $(0 < p < 1, \lambda > 0)$, where

$$f_{p,\lambda}(x) = \begin{cases} p + (1-p)e^{-\lambda} & \text{if } x = 0 \\ (1-p)\lambda^x e^{-\lambda}/x! & \text{if } x = 1, 2, 3, \ldots \end{cases}$$

(a) Find a method-of-moment estimator $\hat{\theta}$ of $\theta = (p, \lambda)^t$.
(b) Find a $2-$dimensional vector μ and 2×2 matrix Σ such that $\sqrt{n}(\hat{\theta} - \mu) \xrightarrow{d} N(0, \Sigma)$.

Advantages and Disadvantages of MME

Advantages of MME

- It is fairly simple and satisfies consistency provided that the moments exist.
- Compared to MLE, it is easy to compute.
- More preferred than MLE if the distribution is unknown.

Disadvantages of MME

- It is possible that estimator does not belong to the parameter space even if exists, which is never the case for MLE.
- Do not attain the desirable optimality properties of MLE or LSE.

Idea of Maximizing likelihood

- If $\text{pdf}(x;\theta_1) > \text{pdf}(x;\theta_2)$ for an observation $X = x$, then the distribution $\text{pdf}(\cdot;\theta_1)$ is more likely, than $\text{pdf}(\cdot;\theta_2)$, to be the true distribution that generated the observation x.
- For a set of observations $x_1, \ldots, x_n$ of a random smaple $X_1, \ldots, X_n$ from pdf $f(\cdot;\theta)$, we have the likelihood (function)

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

and the log-likelihood (function)

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta).$$

Maximum Likelihood Estimator (MLE)

- Definition of MLE: The MLE of θ for a given observation x is defined by

$$\hat{\theta}^{\mathsf{MLE}}(x) \equiv \arg\max_{\theta \in \Theta} L(\theta; x)$$

  when it exists.
- MLE may not exist and may not be unique when it exists.

Existence of MLE

- When the likelihood is differentiable, an MLE is often found by solving the likelihood equation $\dot{l}(\theta; x) = 0$, where $\dot{l}(\theta; x) = (\partial/\partial\theta) l(\theta; x)$.
- Suppose that the likelihood is twice continuously differentiable and $\Theta = \prod_{i=1}^{k} (a_i, b_i)$. For a given x, if $\dot{l}(\hat{\theta}; x) = 0$ and $\ddot{l}(\theta; x)$ is negative definite for all $\theta \in \Theta$, then $\hat{\theta}$ is the unique MLE.
- Suppose that the likelihood is twice continuously differentiable and $\Theta = \prod_{i=1}^{k} (a_i, b_i)$. For a given x, if $\lim_{\theta \to \partial\Theta} l(\theta) = -\infty$ and the second derivative $\ddot{l}(\theta; x)$ is negative definite for all $\theta \in \Theta$, then the solution of the likelihood equation exists and is the unique MLE.

/

MLE of Function of Parameter

- Let $\theta = (\theta_1, \theta_2)$ and $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ is its MLE. Then, we call $\hat{\theta}_j$ the MLEs of θ_j, respectively.
- For a function g, the MLE of $\eta = g(\theta)$ is given by

$$\hat{\eta}^{\text{MLE}} = g(\hat{\theta}^{\text{MLE}}).$$

Profiling Method of Finding MLE

- Sometimes it is difficult to find the MLE of $\theta = (\theta_1, \theta_2)$ simultaneously, but rather easy to find the MLEs of θ_j with the other being fixed. Let $\hat{\theta}_1(\theta_2)$ denote the MLE of θ_1 when θ_2 is fixed.
- The MLE of θ is given by $\theta = (\hat{\theta}_1(\hat{\theta}_2), \hat{\theta}_2)$, where

$$\hat{\theta}_2 = \underset{\theta_2:(\hat{\theta}_1(\theta_2),\theta_2)\in\Theta}{\arg\max} \; l(\hat{\theta}_1(\theta_2), \theta_2)$$

- Multinomial distribution example: Let $X_1, \ldots, X_n$ be random sample from Multi$(n, (p_1, p_2, p_3)^t)$, where $p_1 + p_2 + p_3 = 1$, $p_j > 0$ for all $j = 1, \ldots, k$. Find a maximum likelihood estimator $\hat{\eta}$ of $\eta = (p_1, p_2)^t$ using profiling method.

Sufficient Conditions for Consistency of MLE

- Suppose that we observe a random sample from P_θ in $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Assume that θ in Θ is identifiable and that P_θ have common support and $\Theta = \prod_{i=1}^{k}(a_i, b_i)$. Assume also that the likelihood is twice continuously differentiable, $\lim_{\theta \to \partial\Theta} l(\theta) = -\infty$, and $E_{\theta_0}(\log f(X_1; \theta))$ exists and is continuous with respect to θ. If $\dot{l}(\hat{\theta}; x) = 0$ and $\ddot{l}(\theta; x)$ is negative definite for all $\theta \in \Theta$, the MLE of θ is consistent.
- Logistic$(\theta, 1)$ example: The support of P_θ equals to $\mathbb{R}$, the likelihood $l(\theta) \to -\infty$ as $\theta \to \pm\infty$ and $\ddot{l}(\theta) < 0$ for all θ. Thus, the MLE is consistent.

Consistency of MLE: Some Examples

- Let $X_1, \ldots, X_n$ be a random sample from $U[0, \theta]$, $\theta \in (0, \infty)$. In this case $\hat{\theta} = X_{(n)}$ and $\hat{\theta} \xrightarrow{P_\theta} \theta$ as $n \to \infty$ for all $\theta > 0$.
- Let $X_1, \ldots, X_n$ be a random sample from DE$(\theta, 1)$, $\theta \in \mathbb{R}$. Then $\hat{\theta} = \text{med}(X_i) \xrightarrow{P_\theta} \theta$ as $n \to \infty$ for all $\theta \in \mathbb{R}$.
- Let $X_1, \ldots, X_n$ be a random sample from Logistic(θ, σ), where $\theta \in \mathbb{R}$ and $\sigma > 0$. Then the MLE of (θ, σ) exists and is consistent.

Exercises

2020 Midterm Problem2

Suppose we observe data (X_i, Y_i), $i = 1, 2, \ldots, n$ where $n \geq 2$. Assume that $Y_i \sim \text{Exp}(\lambda_i)$ and Y_i's are mutually independent, where $\mu_i = E[Y_i] = \lambda_i^{-1} = \exp(\alpha + \beta X_i)$. Further assume that $X_i \neq X_j$ if $i \neq j$. Show that the MLE of (α, β) exists and is unique.

Exercises

2020 Midterm Problem3

Consider the following model.

$$Y_i = \theta + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\epsilon_i = c\eta_{i-1} + \eta_i$, $i = 1, \ldots, n$, $\eta_0 = 0$, $\eta_i \overset{\text{i.i.d.}}{\sim} N(0,1)$, and $0 < c < 1$ is constant.

(a) Find the MLE $\hat{\theta}$ of θ.

(b) Show that $\hat{\theta}$ is consistent.

Exercises

2020 Midterm Problem5

Suppose that $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} Poi(\lambda)$. Suppose we cannot observe X_i but can observe whether $X_i = 0$ or $X_i > 0$.
(a) Find the MLE $\hat{\lambda}$ of λ.
(b) Find the case when $\hat{\lambda}$ does not exist and the probability that $\hat{\lambda}$ does not exist assuming λ_0 is true parameter of λ.

Kullback-Leibler Divergence and MLE

- Kullback-Leibler Divergence: Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a statistical model for an observation X. Let $f(\cdot;\theta)$ denote the density function of P_θ. The Kullback-Leibler divergence (of P_θ from P_{θ_0}) is defined by

$$\mathrm{KL}(\theta, \theta_0) = -\mathrm{E}_{\theta_0}(\log f(X;\theta)/f(X;\theta_0)).$$

- Assume $\theta \in \mathcal{P}$ is identifiable and P_θ have common support, i.e., $\{x : f(x;\theta) > 0\}$ does not depend on $\theta \in \Theta$. Then,

$$\mathrm{KL}(\theta, \theta_0) \geq 0 \quad \text{and} \quad \mathrm{KL}(\theta, \theta_0) = 0 \quad \text{if and only if} \quad \theta = \theta_0.$$

Kullback-Leibler Divergence and MLE

Let $X_1, \ldots, X_n$ be a random sample from P_θ in $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Let $f(\cdot;\theta)$ denote the density function of P_θ. Assume θ in Θ is identifiable and P_θ have common support. For a fixed θ_0, define

$$D_n(\theta) = -n^{-1}\sum_{i=1}^{n} \log f(X_i;\theta)/f(X_i;\theta_0), \quad D_0(\theta) = \mathsf{KL}(\theta, \theta_0).$$

- θ_0 is the unique minimizer of $D_0(\theta)$ over Θ.
- $\hat{\theta}$ is a minimizer of $D_n(\theta)$ when it exists.
- By WLLN, $D_n(\theta) \xrightarrow{P_\theta} D_0(\theta)$ for all $\theta \in \Theta$.
- Does $\hat{\theta}$, the minimizer of $D_n(\theta)$, converges to θ_0, the minimizer of $D(\theta)$, in P_{θ_0} probability?

Kullback-Leibler Divergence and MLE

KL Divergence and Consistency of MLE (2019 Midterm Problem1)

Let G_n be a sequence of random functions, and let G_0 be a function defined on Θ. Assume the following conditions:

- Θ is compact.
- G_0 has the unique minimizer.
- G_0 is continuous on Θ.
- $\sup_{\theta \in \Theta} |G_n(\theta) - G_0(\theta)| \xrightarrow{P} 0$.

Denote the minimizer of G_n by $\hat{\theta}_n$ and the minimizer of G_0 by θ_0 over Θ. Prove that $\hat{\theta}_n$ converges to θ_0 in probability.