Probability and Stochastic Processes II: Estimation Part 2

Zhenyu Liao

School of Electronic Information and Communications, HUST

June, 6, 2024

Outline

1. Maximum Likelihood Estimation

2. Consistency and asymptotic normality

3. Fisher information and the Cramer-Rao bound

Likelihood function and the MLE

- Consider data $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f(x|\theta)$ for a **parameter model** $\{f(x|\theta) \colon \theta \in \Omega\}$
- given observed values $X_1, \ldots, X_n$, we call the function

$$\mathcal{L}(\theta) = f(X_1|\theta) \times \ldots f(X_n|\theta), \tag{1}$$

  the **likelihood function**.
- in the discrete case, $\mathcal{L}(\theta)$ just the probability (function) of observing the values $X_1, \ldots, X_n$ if the true parameter were θ; clearly, $\mathcal{L}(\theta)$ is a function of θ
- The **maximum likelihood estimator (MLE)** of θ is the one that maximizes the function $\mathcal{L}(\theta)$
- intuitively, this is the value of θ that makes the observed data "most probable" or "most likely"

From NP to MLE

- the idea to MLE is related to the use of the likelihood ratio statistic in the Neyman-Pearson lemma. Recall that for testing

$$H_0\colon (X_1, \ldots, X_n) \sim g, \quad H_1\colon (X_1, \ldots, X_n) \sim h, \tag{2}$$

  for g, f joints pdfs of n random variables, the most powerful test in the sense of Neyman-Pearson decides on the likelihood ratio

$$L(X_1, \ldots, X_n) = \frac{g(X_1, \ldots, X_n)}{h(X_1, \ldots, X_n)}. \tag{3}$$

- in the context of parametric model, we test between $f(x|\theta_0)$ and $f(x|\theta_1)$, for two possible different parameter values of $\theta_0, \theta_1 \in \Omega$, and the likelihood ratio is $\mathcal{L}(\theta_0)/\mathcal{L}(\theta_1)$
- the MLE, *if exists and is unique*, is the value of $\theta \in \Omega$ such that

$$\mathcal{L}(\theta)/\mathcal{L}(\theta') > 1 \tag{4}$$

  for **any other** values of $\theta' \in \Omega$.

Examples of MLE: Poisson

- deriving the MLE is an optimization problem, it is in general more convenient to **maximize** the **log likelihood function** as

$$l(\theta) = \log(\mathcal{L}(\theta)) = \sum_{i=1}^{n} \log(f(X_i|\theta)). \tag{5}$$

MLE of Poisson parameter

Let $X_1, \ldots, X_n \sim \text{Poission}(\lambda)$. Then

$$\begin{aligned} l(\lambda) &= \sum_{i=1}^{n} \log\left(\frac{\lambda^{X_i} e^{-\lambda}}{X_i!}\right) = \sum_{i=1}^{n} (X_i \log\lambda - \lambda - \log(X_i!)) \\ &= \log\lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n} \log(X_i!). \end{aligned}$$

Taking the derivation (with respect to λ) to zero, we get

$$0 = l'(\theta) = \frac{1}{\lambda} \sum_{i=1}^{n} X_i - n. \tag{6}$$

And the MLE in this case is $\hat{\lambda}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$.

Examples of MLE: Gamma

MLE of Gamma parameter

Let $X_1, \ldots, X_n \sim \text{Gamma}(\alpha, \beta)$. Then

$$l(\alpha, \beta) = \sum_{i=1}^{n} \log \left(\frac{\beta^\alpha}{\Gamma(\alpha)} X_i^{\alpha-1} e^{-\beta X_i} \right) = \sum_{i=1}^{n} (\alpha \log(\beta) - \log \Gamma(\alpha) + (\alpha - 1) \log X_i - \beta X_i)$$

$$= n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log(X_i) - \beta \sum_{i=1}^{n} X_i.$$

Taking the derivation (with respect to (α, β)) to zero, we get

$$0 = \frac{\partial l(\alpha, \beta)}{\partial \alpha} = n \log \beta - \frac{n\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n} \log(X_i),$$

$$0 = \frac{\partial l(\alpha, \beta)}{\partial \beta} = \frac{n\alpha}{\beta} - \sum_{i=1}^{n} X_i.$$

Examples of MLE: Gamma

- the second equation says that the MLEs $\hat{\alpha}, \hat{\beta}$ should satisfy $\hat{\beta} = \hat{\alpha}/\bar{X}$
- substituting into the first we get

$$0 = \log \hat{\alpha} - \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} - \log(\bar{X}) + \frac{1}{n}\sum_{i=1}^{n} \log(X_i) \tag{7}$$

- the function $f(\alpha) = \log \alpha - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$ decreases from ∞ to 0 as α increases from 0 to ∞
- the value of $-\log(\bar{X}) + \frac{1}{n}\sum_{i=1}^{n} \log(X_i) < 0$ (by Jensen's inequality)
- so the MLE $(\hat{\alpha}, \hat{\beta})$ is unique, and in particular, different from the MoM estimator
- unfortunately, there is no closed-form solution to $\hat{\alpha}$
- can be numerically solved using the **Newton-Raphson method**

Newton-Raphson method

Idea: use linear approximation to iteratively solve a nonlinear equation

$$\boxed{f(x) = 0.} \tag{8}$$

- for $f(x)$ well-behaved function, looking for the root $x = r$ of the equation $f(x) = 0$
- start with an "initial guess" x_0 of r, in each iteration, get a "better" estimate x_{i+1} from previous estimate x_i
- assume x_0 is a good initial guess in the sense that $r = x_0 + h$ for some error h that is "small", use **linear approximation** of the smooth function

$$0 = f(r) = f(x_0 + h) \approx f(x_0) + hf'(x_0), \tag{9}$$

so, for $f'(x_0) \neq 0$ that

$$r = x_0 + h \approx x_0 - \frac{f(x_0)}{f'(x_0)}. \tag{10}$$

- do this iteratively as

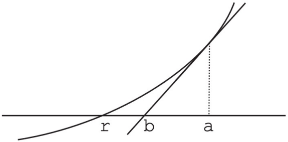
$$\boxed{x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}} \tag{11}$$

Geometric interpretation

- the curve $y = f(x)$ meets the x-axis at $x = r$
- we are current at $x = a$
- the tangent line (i.e., linear approximation) to $y = f(x)$ at the point $(a, f(a))$ is given by

$$y = f(a) + (x - a)f'(a), \tag{12}$$

which meets the x-axis at $b = a - \frac{f(a)}{f'(a)}$, that is the Newton-Raphson estimate 'next' to a



MSE, consistency, and asymptotic normality

- recall from the example of MLE for Poisson distribution above that $\hat{\lambda} = \bar{X}$, and also agrees with the MoM estimator
- we have computed its MSE

$$\mathbb{E}[\hat{\lambda}] = \lambda, \quad \mathrm{Var}[\hat{\lambda}] = \frac{\lambda}{n}, \tag{13}$$

  so that $\hat{\lambda}$ is unbiased and has variance λ/n.
- for n large, we have a precise picture of $\hat{\lambda}$,
  - by LLN, we have $\hat{\lambda} \to \lambda$ in probability or a.s. as $n \to \infty$; and
  - by CLT, $\sqrt{n}(\hat{\lambda} - \lambda) \to \mathcal{N}(0, \lambda)$ in distribution as $n \to \infty$.
- So,

$$\hat{\lambda} \simeq \lambda + \frac{1}{\sqrt{n}} \mathcal{N}(0, \lambda). \tag{14}$$

- This allows to access other measures of error e.g., $\mathbb{E}[|\hat{\lambda} - \lambda|]$ or $\mathbb{P}(|\hat{\lambda} - \lambda| > 0.01)$, as well as obtain a conference interval for $\hat{\lambda}$

Consistency and asymptotic normality

In a parametric model, we say an estimator $\hat{\theta}$ based on $X_1, \ldots, X_n$

- is **consistent** if $\hat{\theta} \to \theta$ in probability as $n \to \infty$; and
- is **asymptotically normal** if $\sqrt{n}(\hat{\theta} - \theta)$ converges in distribution to a normal (or multivariate normal) distribution as $n \to \infty$.

MLE, consistency, and asymptotic normality

Theorem (MLE is consistent and asymptotically normal)

Let $\{f(x|\theta) \colon \theta \in \Omega\}$ be a parametric model, where $\theta \in \mathbb{R}$ is a singe parameter. Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f(x|\theta_0)$ for some $\theta_0 \in \Omega$, and let $\hat{\theta}$ be the MLE based on $X_1, \ldots, X_n$. Suppose certain regularity conditions hold, including:

- *the log-likelihood $l(\theta)$ is differentiable with respect to θ*
- *$\hat{\theta}$ is the unique value in Ω that solve s $0 = l'(\theta)$.*

Then, $\hat{\theta}$ is consistent and asymptotically normal, with

$$\sqrt{n}(\hat{\theta} - \theta_0) \to \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right), \tag{15}$$

with ***Fisher information***

$$I(\theta) = \mathrm{Var}[z(X,\theta)] = -\mathbb{E}[z'(X,\theta)], \tag{16}$$

for ***score function*** $z(x,\theta) = \frac{\partial}{\partial\theta} \log f(x|\theta)$, *and* $z'(x,\theta) = \frac{\partial^2}{\partial\theta^2} \log f(x|\theta)$.

(Some technical conditions in addition to the ones stated are required to make this theorem rigorously true, but they are beyond the scope of this class.)

- Exercise: check this is true for the Poisson estimation problem.

Fisher information matrix

Asymptotic normality of the MLE extends naturally to the setting of multiple parameters.

Theorem (MLE is consistent and asymptotically normal)

Let $\{f(x|\theta)\colon \theta \in \Omega\}$ be a parametric model, where $\theta \in \mathbb{R}^k$ has k parameters. Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f(x|\theta_0)$ for some $\theta_0 \in \Omega$, and let $\hat{\theta}$ be the MLE based on $X_1, \ldots, X_n$. Define the ***Fisher information matrix*** $\mathbf{I}(\theta) \in \mathbb{R}^{k \times k}$*, with its (i,j) entry given by*

$$[\mathbf{I}(\theta)]_{i,j} = \operatorname{Cov}\left[\frac{\partial}{\partial\theta_i}\log f(X|\theta), \frac{\partial}{\partial\theta_j}\log f(X|\theta)\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(X|\theta)\right]. \tag{17}$$

Then, under the same regularity conditions, we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \to \mathcal{N}\left(0, \mathbf{I}(\theta)^{-1}\right). \tag{18}$$

The Cramer-Rao lower bound

Recall the definition of the Fisher information (matrix):

- for $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f(x|\theta_0)$ for some true parameter $\theta_0 \in \Omega$
- and $l(\theta) = \sum_{i=1}^n \log f(X_i|\theta)$ the log-likelihood function
- then the fishier information (at true θ_0) is given by

$$I(\theta_0) = -\mathbb{E}\left[\frac{\partial^2}{\partial^2}[\log f(X|\theta)]_{\theta=\theta_0}\right] = -\frac{1}{n}\mathbb{E}[l''(\theta_0)]. \tag{19}$$

- $I(\theta_0)$ measures the expected curvature of the log-likelihood function $l(\theta)$ around the true parameter $\theta = \theta_0$
  - if $l(\theta)$ is sharply curved around θ_0, then a small change in θ can lead to a large decrease in the log-likelihood
  - the data/observations provide **rich** information whether θ is close to θ_0
- This Fisher information is an **intrinsic property** of the model (note that its definition is independent of MLE)

The Cramer-Rao lower bound

We have the following Cramer-Rao lower bound result.

Theorem (Cramer-Rao lower bound)

Consider a parametric model $\{f(x|\theta) : \theta \in \Omega\}$ (satisfying certain mild regularity assumptions) where $\theta \in \Omega$ is a single parameter. Let T be any ***unbiased*** *estimator of θ based on $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f(x|\theta)$. Then,*

$$\mathrm{Var}[T] \geq \frac{1}{nI(\theta)}. \tag{20}$$

- For two unbiased estimators of θ, the ratio of their variances is called their **relative efficiency**.
- An unbiased estimator is **efficient** if its variance equals the lower bound $\frac{1}{n\mathbf{I}(\theta)}$.
- Since the MLE achieves this lower bound asymptotically, we say it is **asymptotically efficient**.
- Notice: sometimes we can do better with **slightly biased estimators**, check James–Stein estimator for more info.

Proof of Cramer-Rao lower bound

- recall the definition of score function $z(x,\theta) = \frac{\partial}{\partial\theta} \log f(x|\theta) = \frac{1}{f(x|\theta)} \frac{\partial f(x|\theta)}{\partial\theta}$
- let now $Z = \sum_{i=1}^n z(X_i, \theta)$, by the definition of covariance/correlation and the Cauchy-Schwarz inequality that, for any estimator T, we have

$$\mathrm{Cov}[Z,T]^2 \leq \mathrm{Var}[Z] \cdot \mathrm{Var}[T]. \tag{21}$$

- Since the random variables $z(X_i,\theta)$ are i.i.d. and is of zero mean and variance $I(\theta)$ (prove this using the chain rule of differentiation and the definition of Fisher information), we have

$$\mathrm{Var}[Z] = n\mathrm{Var}[z(X_i,\theta)] = nI(\theta). \tag{22}$$

- Since T is **unbiased**, we can write

$$\theta = \mathbb{E}[T] = \int_{\mathbb{R}^n} T(x_1,\ldots,x_n) f(x_1|\theta) \times \ldots \times f(x_n|\theta) dx_1 \ldots dx_n. \tag{23}$$

$$\theta = \mathbb{E}[T] = \int_{\mathbb{R}^n} T(x_1, \dots, x_n) f(x_1|\theta) \times \dots \times f(x_n|\theta) dx_1 \dots dx_n. \tag{24}$$

- differentiating both sides with respect to θ, we get

$$\begin{aligned}
1 &= \int_{\mathbb{R}^n} T(x_1, \dots, x_n) \left(\frac{\partial}{\partial\theta} f(x_1|\theta) \times \dots \times f(x_n|\theta) + \dots \right. \\
&\left. + f(x_1|\theta) \times \dots \times \frac{\partial}{\partial\theta} f(x_n|\theta) \right) dx_1 \dots dx_n. \\
&= \int_{\mathbb{R}^n} T(x_1, \dots, x_n) \times Z(x_1, \dots, x_n) \times f(x_1|\theta) \times \dots \times f(x_n|\theta) dx_1 \dots dx_n. \\
&= \mathbb{E}[TZ]
\end{aligned}$$

- since $\mathbb{E}[Z] = 0$, we must have $\mathrm{Cov}[T, Z] = \mathbb{E}[TZ]^2 = 1$, and thus $\mathrm{Var}[T] \geq \frac{1}{nI(\theta)}$.

Thank you!

Thank you! Q & A?

Exercises

MLE: Gaussian distribution

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, derive the corresponding MLE $\hat{\mu}, \hat{\sigma}^2$ for the mean and variance parameter μ and σ^2, respectively.
(You should check that the obtained results agree with MoM estimates and they are indeed unique minimizer of the likelihood function.)

MLE of Gamma distribution and its asymptotic normality

xxx