Maximum Likelihood Theory

Umberto Triacca

Università dell'Aquila
Department of Computer Engineering, Computer Science and Mathematics, University of L'Aquila, L'Aquila, Italy
umberto.triacca@univaq.it

Maximum Likelihood Theory

- Introduction
- Likelihood and log-likelihood
- Score vector and score function
- Fisher information matrix
- Information matrix equality
- Cramer-Rao inequality
- Maximum Likelihood esimate/estimator
- Invariance
- Consistency
- Large sample distribution
- Asymptotic efficiency

Lesson 1: The Log-likelihood Function

Umberto Triacca

Università dell'Aquila
Department of Computer Engineering, Computer Science and Mathematics, University of L'Aquila, L'Aquila, Italy
umberto.triacca@univaq.it

Probabilistic model

Consider an observable random phenomenon. Suppose that this phenomenon can be appropriately described by random variable X with probability **density** function (*pdf*) (or probability **mass** function (*pmf*)) belonging to the family

$$\Phi = \{f(x;\theta);\ \theta \in \Theta\}$$

where Θ is a subset of the k-dimensional Euclidean space $\mathbb{R}^k$ called the **parametric space**. The family Φ is a **probabilistic model**.

Gaussian model

$$\Phi = \left\{ f(x;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2};\ \theta = (\mu, \sigma^2)' \in \Theta = \mathbb{R} \times \mathbb{R}^+ \right\}$$

Bernoulli model

$$\Phi = \left\{ f(x;\theta) = \theta^x (1-\theta)^{1-x};\ \theta \in \Theta = (0,1) \right\}$$

Poisson model

$$\Phi = \left\{ f(x;\theta) = \frac{e^{-\theta}\theta^x}{x!};\ \theta \in \Theta = (0,\infty) \right\}$$

Probabilistic model

The starting point of any inferential process is a probabilistic model. Since the functional form of the density functions of the probabilistic model is known, we have that all the uncertainty concerning the random phenomenon is that concerning the parameter θ.

In order to get information on θ, we will consider a sample from the population described by our random variable. In particular, we will consider a random sample. What is a random sample?

Random sample

Definition 1. Let $\mathbf{X}_n = (X_1, X_2, ..., X_n)'$ be a vector of n random variables independent, identically distributed (i.i.d.) with *pdf* (or *pmf*) belonging to the family

$$\Phi = \{f(x;\theta);\ \theta \in \Theta\}.$$

We say that $\mathbf{X}_n = (X_1, X_2, ..., X_n)'$ is a **random sample** of size n from $f(x;\theta)$.

The distribution of the random sample $\mathbf{X}_n = (X_1, X_2, ..., X_n)'$ is the joint distribution of the random variables $X_1, X_2, ..., X_n$ denoted by

$$f_{1,2,...,n}(\mathbf{x}_n;\theta) = f_{1,2,...,n}(x_1, x_2, ..., x_n;\theta)$$

We have that

$$f_{1,2,...,n}(x_1, x_2, ..., x_n;\theta) = f(x_1;\theta)f(x_2;\theta)...f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

An example

Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a random sample from an $N(\mu, \sigma^2)$ distribution with μ and σ^2 unknown.
In this case $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$, and the distribution of the random sample is

$$\begin{aligned} f_{1,2,...,n}(\mathbf{x}_n; \theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \left(\frac{x_i - \mu}{\sigma} \right)^2 \right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\} \end{aligned}$$

Likelihood Function

Definition 2.Let $\mathbf{X}_n = (X_1, X_2, ..., X_n)$ be a random sample of size n from $f(x;\theta)$. Given a realization $\mathbf{x}_n = (x_1, x_2, ..., x_n)$ of the random sample $\mathbf{X}_n = (X_1, X_2, ..., X_n)$, the function

$$L : \Theta \to [0, \infty)$$

defined by $L(\theta; \mathbf{x}_n) = f_{1,2,...,n}(\mathbf{x}_n; \theta)$ is called the **likelihood function**.

An example

Let $\mathbf{x} = (x_1, x_2, ..., x_n)$ be a realization of a random random sample from an $N(\mu, \sigma^2)$ distribution with μ and σ^2 unknown.
In this case $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$, and the likelihood function is

$$L(\mu, \sigma^2; \mathbf{x}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}$$

Another example

Let $\mathbf{x}_n = (x_1, x_2, ..., x_n)'$ be a realization of a random sample $\mathbf{X} = (X_1, X_2, ..., X_n)'$ from a Bernoulli distribution with probability mass function

$$f(x;\theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \end{cases}$$

The likelihood function is

$$L(\theta;\mathbf{x}_n) = \theta^{x_1}(1-\theta)^{(1-x_1)}\theta^{x_2}(1-\theta)^{(1-x_2)}...\theta^{x_n}(1-\theta)^{(1-x_n)}$$

$$= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{(n-\sum_{i=1}^n x_i)}$$

Likelihood Function

Let $\mathbf{x}_n = (x_1, x_2, ..., x_n)'$ and $\mathbf{x}_n^* = (x_1^*, x_2^*, ..., x_n^*)'$ be two different realizations of a random sample $\mathbf{X}_n = (X_1, X_2, ..., X_n)'$.

The likelihood function at the point $\mathbf{x}_n$ is (generally) a different function from what it is at the point $\mathbf{x}_n^* = (x_1^*, x_2^*, ..., x_n^*)'$, that is

$$L(\theta; \mathbf{x}_n) \neq L(\theta; \mathbf{x}_n^*)$$

Likelihood Function

Consider a realization $\mathbf{x}_5 = (x_1, x_2, x_3, x_4, x_5)$ of a random sample $\mathbf{X}_5 = (X_1, X_2, X_3, X_4, X_5)'$ from a Bernoulli distribution with parameter θ.

Suppose $\mathbf{x}_5 = (1, 0, 1, 0, 1)'$. The likelihood function is:

$$L(\theta; (1, 0, 1, 0, 1)') = \theta^3(1 - \theta)^2$$

Suppose $\mathbf{x}_5 = (1, 0, 1, 0, 0)'$. The likelihood function is:

$$L(\theta; (1, 0, 1, 0, 0)') = \theta^2(1 - \theta)^3$$

Suppose $\mathbf{x}_5 = (1, 0, 0, 0, 0)'$. The likelihood function is:

$$L(\theta; (1, 0, 0, 0, 0)') = \theta(1 - \theta)^4$$

Likelihood Function

The likelihood function at the point $\mathbf{x}_n$ is (generally) a different function from what it is at the point $\mathbf{x}_n^* = (x_1^*, x_2^*, ..., x_n^*)'$, that is

$$L(\theta; \mathbf{x}_n) \neq L(\theta; \mathbf{x}_n^*)$$

Generally, but not always!

Consider again two realizations of a random sample $\mathbf{X}_5 = (X_1, X_2, X_3, X_4, X_5)'$ from a Bernoulli distribution, $\mathbf{x}_5 = (1, 0, 0, 0, 0)'$ and $\mathbf{x}_5^* = (0, 0, 0, 0, 1)'$. We have that $\mathbf{x}_5 \neq \mathbf{x}_5^*$ but

$$L(\theta; \mathbf{x}_5) = L(\theta; \mathbf{x}_5^*) = \theta(1-\theta)^4$$

Likelihood Function

The likelihood function expresses the plausiblities of different parameters after we have observed $\mathbf{x}_n$. In particular, for $\theta = \theta^*$, the number $L(\theta^*; \mathbf{x}_n)$ is considered a measure of support that the observation $\mathbf{x}_n$ gives to the parameter θ^*.

Likelihood Function

Consider a realization $\mathbf{x}_5 = (x_1, x_2, x_3, x_4, x_5)$ of a random sample $\mathbf{X}_5 = (X_1, X_2, X_3, X_4, X_5)'$ from a Bernoulli distribution with parameter θ.

Suppose $\mathbf{x}_5 = (1, 1, 1, 1, 1)'$ and consider two possible values of θ: $\theta_1 = 1/3$ and $\theta_2 = 2/3$. The plausibility of θ_1 is:

$$L(\theta_1; (1, 1, 1, 1, 1)') = \left(\frac{1}{3}\right)^5 = 0.004115226$$

The plausibility of θ_2 is:

$$L(\theta_2; (1, 1, 1, 1, 1)') = \left(\frac{2}{3}\right)^5 = 0.1316872$$

Clearly

$$L(\theta_2; (1, 1, 1, 1, 1)') > L(\theta_1; (1, 1, 1, 1, 1)')$$

Likelihood Function

Now, suppose $\mathbf{x}_5 = (0, 0, 0, 0, 0)'$. The plausibility of θ_1 is:

$$L(\theta_1; (0, 0, 0, 0, 0)') = \left(\frac{2}{3}\right)^5 = 0.1316872$$

The plausibility of θ_2 is:

$$L(\theta_2; (0, 0, 0, 0, 0)') = \left(\frac{1}{3}\right)^5 = 0.004115226$$

Clearly

$$L(\theta_1; (0, 0, 0, 0, 0)') > L(\theta_2; (0, 0, 0, 0, 0)').$$

The Log-likelihood Function

Often, we work with the natural logarithm of the likelihood function, the so-called **log-likelihood function**:

$$l(\mathbf{x};\theta) = \ln L(\mathbf{x};\theta) = \sum_{i=1}^{n} \ln f(x_i;\theta)$$

An example

Let $\mathbf{x} = (x_1, x_2, ..., x_n)$ be a realization of a random random sample from an $N(\mu, \sigma^2)$ distribution with μ and σ unknown.
In this case $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$, and the likelihood function is

$$\begin{aligned} L(\mu, \sigma^2; \mathbf{x}) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \\ &= \frac{1}{\sigma^n(2\pi)^{\frac{n}{2}}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right] \end{aligned}$$

and the log-likelihood function is given by

$$l(\mu, \sigma^2; \mathbf{x}) = -n\ln\sigma - \frac{n}{2}\ln 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Lesson 2: Score vector and information matrix

Umberto Triacca

Università dell'Aquila
Department of Computer Engineering, Computer Science and Mathematics, University of L'Aquila, L'Aquila, Italy
umberto.triacca@univaq.it

Score function

Definition 3. If the likelihood function, $L(\theta; \mathbf{x})$, is differentiable, then the gradient of the log-likelihood

$$s(\theta; \mathbf{x}) = \frac{\delta l(\theta; \mathbf{x})}{\delta \theta} = \frac{\delta \ln L(\theta; \mathbf{x})}{\delta \theta}$$

is called **the score function**.

The score function can be found through the chain rule:

$$\frac{\delta l(\theta; \mathbf{x})}{\delta \theta} = \frac{1}{L(\theta; \mathbf{x})} \frac{\delta L(\theta; \mathbf{x})}{\delta \theta}$$

Score function: An example

Let $\mathbf{x} = (x_1, x_2, ..., x_n)'$ be a realization of random sample from an $N(\mu, \sigma^2)$ distribution. Here $\theta = (\mu, \sigma^2)$. The score function is given by

$$s(\theta; \mathbf{x}) = \left(\frac{\Sigma(x_i - \mu)}{\sigma^2}, \frac{\Sigma(x_i - \mu)^2}{2\sigma^4} - \frac{n}{2\sigma^2} \right)'$$

Score vector

Definition 4. Evaluating the score function at a specific value of θ and replacing the fixed values $\mathbf{x} = (x_1, x_2, ..., x_n)'$ by their corresponding random variables $\mathbf{X} = (X_1, X_2, ..., X_n)'$, the score function becomes a **random vector**

$$s(\theta; \mathbf{X}) = \frac{\delta l(\theta; \mathbf{X})}{\delta \theta} = \frac{\delta \ln f(\mathbf{X}; \theta)}{\delta \theta}.$$

We call this random vector **score vector**.

Score vector

Which is the expected value of the score vector? **The expected value of the score vector evaluated at the true parameter value equals zero**.

Theorem 1. Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a random sample from a distribution with p.d.f. belonging to the family

$$\Phi = \{f(x; \theta);\ \theta \in \Theta\}$$

and let θ_0 be the true value of the parameter θ, then under suitable regularity conditions

$$E\left[s(\theta_0; \mathbf{X})\right] = \mathbf{0}$$

Proof Theorem 1

In the following the single integral $\int ...d\mathbf{x}$, is used to indicate the multiple integration over all elements of $\mathbf{x}$.

Further, the range of integration is the whole range of $\mathbf{x}$ except for the set of points where $f(\mathbf{x};\theta) = 0$.

In the sequel we use often the phrase 'under suitable regularity conditions'. These conditions mainly relate to differentiability of the density and the ability to interchange differentiation and integration.

Proof Theorem 1

Because $f(\mathbf{x};\theta)\ \ \forall\theta \in \Theta$ is a probability density function, we have that:

$$\int f(\mathbf{x};\theta)d\mathbf{x} = 1\ \ \forall\theta \in \Theta \tag{1}$$

Thus, differentiating (1) w.r.t. θ we get

$$\frac{\delta}{\delta\theta}\left[\int f(\mathbf{x};\theta)d\mathbf{x}\right] = \mathbf{0} \tag{2}$$

The regularity conditions guarantee that operations of differentiation and integration can be interchanged. Thus, we have

$$\frac{\delta}{\delta\theta}\left[\int f(\mathbf{x};\theta)d\mathbf{x}\right] = \int \frac{\delta f(\mathbf{x};\theta)}{\delta\theta}d\mathbf{x}$$

So, (2) can be rewritten as

$$\int \frac{\delta f(\mathbf{x};\theta)}{\delta\theta}d\mathbf{x} = \mathbf{0} \tag{3}$$

Proof Theorem 1

Because

$$\frac{\delta \ln f(\mathbf{x};\theta)}{\delta \theta} = \frac{1}{f(\mathbf{x};\theta)} \frac{\delta f(\mathbf{x};\theta)}{\delta \theta}.$$

we have that

$$\frac{\delta f(\mathbf{x};\theta)}{\delta \theta} = \frac{\delta \ln f(\mathbf{x};\theta)}{\delta \theta} f(\mathbf{x};\theta)$$

and hence

$$\int \frac{\delta \ln f(\mathbf{x};\theta)}{\delta \theta} f(\mathbf{x};\theta) d\mathbf{x} = \mathbf{0} \quad \forall \theta \in \Theta. \tag{4}$$

Proof Theorem 1

On the other hand, we have that

$$\int \frac{\delta \ln f(\mathbf{x};\theta_0)}{\delta\theta} f(\mathbf{x};\theta_0) d\mathbf{x} = E\left[\frac{\delta \ln f(\mathbf{X};\theta_0)}{\delta\theta}\right] = E\left[s(\theta_0;\mathbf{X})\right]. \quad (5)$$

By equation (4) it follows that

$$E\left[s(\theta_0;\mathbf{X})\right] = \mathbf{0} \quad (6)$$

The score vector evaluated at the true parameter value has mean zero.

Score vector

Remark. Consider a vector of parameters $\theta_1 \neq \theta_0$. We have that

$$\int \frac{\delta \ln f(\mathbf{x}; \theta_1)}{\delta \theta} f(\mathbf{x}; \theta_1) d\mathbf{x} = \mathbf{0}.$$

However

$$E\left[s(\theta_1; \mathbf{X})\right]$$

can be different from the null vector $\mathbf{0}$, being, in general,

$$E\left[s(\theta_1; \mathbf{X})\right] = \int \frac{\delta \ln f(\mathbf{x}; \theta_1)}{\delta \theta} f(\mathbf{x}; \theta_0) d\mathbf{x} \neq \int \frac{\delta \ln f(\mathbf{x}; \theta_1)}{\delta \theta} f(\mathbf{x}; \theta_1) d\mathbf{x}.$$

Information matrix

Consider the variance-covariance matrix of the score vector

$$Var\left[s(\theta;\mathbf{X})\right] = Var\left[\frac{\delta \ln f(\mathbf{X};\theta)}{\delta\theta}\right]$$

We note that it is a function of θ. We will denote this function with $I_n(\theta)$.

Information matrix

Definition 4. The variance-covariance matrix of the score, evaluated at the true parameter value,

$$I_n(\theta_0) = Var\left[\frac{\delta \ln f(\mathbf{X};\theta_0)}{\delta\theta}\right] = E\left[\frac{\delta \ln f(\mathbf{X};\theta_0)}{\delta\theta}\frac{\delta \ln f(\mathbf{X};\theta_0)}{\delta\theta'}\right] \quad (7)$$

is called **information matrix** for θ_0 (or Fisher's information measure on θ_0 contained in the r.v. $\mathbf{X}$).

It is important to note that the information does not depend on the particular observation $\mathbf{x}$.

$I_n(\theta_0)$ measures the amount of information about θ_0 contained (on average) in a realization $\mathbf{x}$ of the r.v. $\mathbf{X}$.

Information matrix equality

Theorem 2. Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a random sample from a distribution with p.d.f. belonging to the family

$$\Phi = \{f(x;\theta);\ \theta \in \Theta\}$$

and let θ_0 be the true value of the parameter θ, then under some regularity conditions

$$I_n(\theta_0) = -E\left[\frac{\delta^2 \ln f(\mathbf{X};\theta_0)}{\delta\theta\delta\theta'}\right]$$

This is called the **information matrix equality**. It provides an alternative expression for the information matrix. The information matrix equals the negative of the expected value of Hessian (matrix of second partial derivatives) of the log-likelihood evaluated at the true parameter θ_0.

Proof Theorem 2

We note that

$$\int \frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta} f(\mathbf{x};\theta) d\mathbf{x} = \mathbf{0} \quad \forall \theta \in \Theta.$$

Thus, differentiating the above equation w.r.t. θ and evaluating the derivative at θ_0, we get

$$\int \left[\frac{\delta^2 \ln f(\mathbf{x};\theta_0)}{\delta\theta\delta\theta'} f(\mathbf{x};\theta_0) + \frac{\delta \ln f(\mathbf{x};\theta_0)}{\delta\theta} \frac{\delta f(\mathbf{x};\theta_0)}{\delta\theta'} \right] d\mathbf{x} = \mathbf{0} \qquad (8)$$

that is

$$\int \frac{\delta^2 \ln f(\mathbf{x};\theta_0)}{\delta\theta\delta\theta'} f(\mathbf{x};\theta_0) d\mathbf{x} + \int \frac{\delta \ln f(\mathbf{x};\theta_0)}{\delta\theta} \frac{\delta \ln f(\mathbf{x};\theta_0)}{\delta\theta'} f(\mathbf{x};\theta_0) d\mathbf{x} = \mathbf{0} \qquad (9)$$

Proof Theorem 2

Again, because derivatives are computed at θ_0, and $f(\mathbf{x};\theta_0)$ is the probability density of the r. v. $\mathbf{X}$, the two terms in equation (9) are expectations, so

$$E\left[\frac{\delta^2 \ln f(\mathbf{X};\theta_0)}{\delta\theta\delta\theta'}\right] + E\left[\frac{\delta \ln f(\mathbf{X};\theta_0)}{\delta\theta}\frac{\delta \ln f(\mathbf{X};\theta_0)}{\delta\theta'}\right] = \mathbf{0} \qquad (10)$$

Proof Theorem 2

The second term of the sum is the information matrix (7). Thus, from (10) we get an alternative expression for the information matrix

$$I_n(\theta_0) = E\left[-\frac{\delta^2 \ln f(\mathbf{X};\theta_0)}{\delta\theta\delta\theta'}\right] \tag{11}$$

The Hessian of the log-likelihood

The Hessian of the log-likelihood is

$$H(\mathbf{X};\theta) = \frac{\delta^2 \ln f(\mathbf{X};\theta)}{\delta\theta\delta\theta'} = \begin{bmatrix} \frac{\delta^2 \ln f(\mathbf{X};\theta)}{\delta\theta_1^2} & \frac{\delta \ln f(\mathbf{X};\theta)}{\delta\theta_1\delta\theta_2} & \cdots & \frac{\delta \ln f(\mathbf{X};\theta)}{\delta\theta_1\delta\theta_n} \\ \frac{\ln f(\mathbf{X};\theta)}{\delta\theta_2\delta\theta_1} & \frac{\delta^2 \ln f(\mathbf{X})}{\delta\theta_2^2} & \cdots & \frac{\delta \ln f(\mathbf{X};\theta)}{\delta\theta_2\delta\theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\ln f(\mathbf{X};\theta)}{\delta\theta_k\delta\theta_1} & \frac{\delta \ln f(\mathbf{X};\theta)}{\delta\theta_k\delta\theta_2} & \cdots & \frac{\delta^2 \ln f(\mathbf{X};\theta)}{\delta\theta_k^2} \end{bmatrix}$$

By Theorem 2 it follows that if $\mathbf{X} = (X_1, X_2, ..., X_n)'$ be a random sample from a distribution with p.d.f. $f(x;\theta_0)$, then the information matrix at θ_0 is equal to the expected Hessian of the log-likelihood, with the opposite sign,

$$I_n(\theta_0) = -E\left[H(\mathbf{X};\theta_0)\right].$$

Information matrix

It is important to note that the results presented do not depend on the assumption of independence of the random variables $X_1, X_2, ..., X_n$. This assumption can be used in order to get the following result.
Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a random sample from a distribution with p.d.f. $f(x; \theta_0)$. We have that

$$I_n(\theta_0) = nI_1(\theta_0)$$

The information in a random sample of size n is n times that in a sample of size 1.

An example

Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a random sample from from a distribution with p.d.f. belonging to the family

$$\Phi = \left\{ f(x;\theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left\{ -\frac{1}{2}\left(\frac{x-1}{\sqrt{\theta}} \right)^2 \right\};\ \theta \in \Theta \right\}$$

and let θ_0 be the true value of the parameter θ.
We have that

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

and

$$l(\theta, \mathbf{x}) = \sum_{i=1}^{n} \ln f(x_i;\theta) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\theta - \frac{1}{2\theta}\sum_{i=1}^{n}(x_i - 1)^2$$

An example

The score vector is given by

$$s(\theta; \mathbf{X}) = -\frac{n}{2\theta} + \frac{\sum_{i=1}^{n}(X_i - 1)^2}{2\theta^2}.$$

We have that

$$\begin{aligned}
E\left[s(\theta_0; \mathbf{X})\right] &= -\frac{n}{2\theta_0} + E\left[\frac{\sum_{i=1}^{n}(X_i - 1)^2}{2\theta_0^2}\right] \\
&= -\frac{n}{2\theta_0} + \frac{\sum_{i=1}^{n} E(X_i - 1)^2}{2\theta_0^2} \\
&= -\frac{n}{2\theta_0} + \frac{\sum_{i=1}^{n} \theta_0}{2\theta_0^2} \\
&= -\frac{n}{2\theta_0} + \frac{n\theta_0}{2\theta_0^2} \\
&= 0
\end{aligned}$$

An example

Now, consider the score vector evaluated in $\theta_1 \neq \theta_0$

$$s(\theta_1; \mathbf{X}) = -\frac{n}{2\theta_1} + \frac{\sum_{i=1}^{n}(X_i - 1)^2}{2\theta_1^2}.$$

We have that

$$\begin{aligned} E\left[s(\theta_1; \mathbf{X})\right] &= -\frac{n}{2\theta_1} + E\left[\frac{\sum_{i=1}^{n}(X_i - 1)^2}{2\theta_1^2}\right] \\ &= -\frac{n}{2\theta_1} + \frac{\sum_{i=1}^{n} E(X_i - 1)^2}{2\theta_1^2} \\ &= -\frac{n}{2\theta_1} + \frac{\sum_{i=1}^{n} \theta_0}{2\theta_1^2} \\ &= -\frac{n}{2\theta_1} + \frac{n\theta_0}{2\theta_1^2} \\ &\neq 0 \end{aligned}$$

An example

Now, we calculate the information number $I_n(\theta_0)$. Because $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a random sample, we have that

$$I_n(\theta_0) = nI_1(\theta_0)$$

The information in a random sample of size n is n times that in a sample of size 1.

An example

$$
\begin{aligned}
I_1(\theta_0) &= \mathrm{Var}\left(s(\theta_0; X_1)\right) \\
&= \mathrm{Var}\left(\frac{(X_1-1)^2}{2\theta_0^2}\right) \\
&= \frac{1}{4\theta_0^4}\mathrm{Var}\left((X_1-1)^2\right) \\
&= \frac{1}{4\theta_0^4}E\left[\left((X_1-1)^2-\theta_0\right)^2\right] \\
&= \frac{1}{4\theta_0^4}E\left[(X_1-1)^4-2\theta_0(X_1-1)^2+\theta_0^2\right] \\
&= \frac{1}{4\theta_0^4}\left(E\left[(X_1-1)^4\right]-2\theta_0 E\left[(X_1-1)^2\right]+\theta_0^2\right) \\
&= \frac{1}{4\theta_0^4}\left(3\theta_0^2-2\theta_0^2+\theta_0^2\right) \\
&= 1/(2\theta_0^2)
\end{aligned}
$$

An example

We can conclude that

$$I_n(\theta_0) = \frac{n}{2\theta_0^2}$$

The asymptotic information matrix

The matrix

$$I_a(\theta_0) = \lim_{n\to\infty} I_n(\theta_0)/n$$

if it exists, is the **asymptotic information matrix** for θ_0.

The asymptotic information matrix

$$
\begin{aligned}
I_a(\theta_0) &= \lim_{n\to\infty} \frac{I_n(\theta_0)}{n} \\
&= \lim_{n\to\infty} \frac{nI_1(\theta_0)}{n} \\
&= I_1(\theta_0)
\end{aligned}
$$

The asymptotic information matrix is the Fisher information matrix for one observation.

Lesson 3: Cramer-Rao inequality

Umberto Triacca

Università dell'Aquila
Department of Computer Engineering, Computer Science and Mathematics, University of L'Aquila, L'Aquila, Italy
umberto.triacca@univaq.it

Cramer-Rao inequality

Theorem 3. Let $\mathbf{X} = (x_1, ..., x_n)$ be a random sample of n observations from the distribution with p.d.f. $f(x; \theta)$ depending on a real parameter θ. Let $T(\mathbf{X})$ be an unbiased estimator of θ. Then, subject to certain regularity conditions on $f(x; \theta)$, the variance of $T(\mathbf{X})$ satisfies the inequality

$$\text{Var}[T(\mathbf{X})] \geq \frac{1}{E\left[\left(\frac{\delta \ln f(\mathbf{X};\theta)}{\delta\theta}\right)^2\right]}$$

Proof

$T(\mathbf{X})$ is an unbiased estimator of θ, so

$$E[T(\mathbf{X})] = \int T(\mathbf{x}) f(\mathbf{x};\theta) d\mathbf{x} = \theta \tag{12}$$

Differentiating both sides of equation (1) with respect to θ, and interchanging the order of integration and differentiation, gives

$$\int T(\mathbf{x}) \frac{\delta f(\mathbf{x};\theta)}{\delta\theta} d\mathbf{x} = 1 \tag{13}$$

or

$$\int T(\mathbf{x}) \frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta} f(\mathbf{x};\theta) d\mathbf{x} = 1 \tag{14}$$

Proof

Because

$$\int T(\mathbf{x}) \frac{\delta \ln f(\mathbf{x};\theta)}{\delta \theta} f(\mathbf{x};\theta) d\mathbf{x} = E\left[T(\mathbf{X}) \frac{\delta \ln f(\mathbf{X};\theta)}{\delta \theta} \right] \tag{15}$$

by (3) it follows that

$$E\left[T(\mathbf{X}) \frac{\delta \ln f(\mathbf{X};\theta)}{\delta \theta} \right] = 1$$

Proof

On the other hand, since

$$E\left[\frac{\delta \ln f(\mathbf{X};\theta)}{\delta\theta}\right] = 0$$

we have that

$$E\left[T(\mathbf{X})\frac{\delta \ln f(\mathbf{X};\theta)}{\delta\theta}\right] = \text{Cov}\left[T(\mathbf{X}), \frac{\delta \ln f(\mathbf{X};\theta)}{\delta\theta}\right]$$

Hence

$$\text{Cov}\left[T(\mathbf{X}), \frac{\delta \ln f(\mathbf{X};\theta)}{\delta\theta}\right] = 1$$

Proof

Since the squared covariance cannot exceed the product of the two variances, we have

$$1 = \left(\text{Cov} \left[T(\mathbf{X}), \frac{\delta \ln f(\mathbf{X}; \theta)}{\delta \theta} \right] \right)^2 \leq \text{Var}\left[T(\mathbf{X}) \right] \text{Var} \left[\frac{\delta \ln f(\mathbf{X}; \theta)}{\delta \theta} \right]$$

or

$$1 = \left(\text{Cov} \left[T(\mathbf{X}), \frac{\delta \ln f(\mathbf{X}; \theta)}{\delta \theta} \right] \right)^2 \leq \text{Var}\left[T(\mathbf{X}) \right] E \left[\left(\frac{\delta \ln f(\mathbf{X}; \theta)}{\delta \theta} \right)^2 \right]$$

It follows that

$$\text{Var}[T(\mathbf{X})] \geq \frac{1}{E\left[\left(\frac{\delta \ln f(\mathbf{x}; \theta)}{\delta \theta} \right)^2 \right]}$$

Remark

We have seen that the quantity

$$I_n(\theta) = E\left[\left(\frac{\delta \ln f(\mathbf{X};\theta)}{\delta\theta}\right)^2\right]$$

is called information number or Fisher information. Now, we are able to explain the reason of this terminology.

As $I_n(\theta)$ (the information number) gets bigger, we have a smaller bound on the variance of the best unbiased estimator of θ. Therefore, we might expect a smaller variance of the best estimator.
On the other hand, we can obtain a smaller variance of the best estimator if and only if the information about θ provided on average by an observation in $\mathbf{X}$ is augmented. Thus, we can conclude that the information number is a measure of this information.

Efficency

Definition 6. An unbiased estimator is efficient if its variance is the lower bound of the inequality, that is

$$\text{Var}[T(\mathbf{X})] = \frac{1}{E\left[\left(\frac{\delta \ln f(\mathbf{X};\theta)}{\delta\theta}\right)^2\right]}$$

Efficiency: example

Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a random sample from a distribution with p.d.f.

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left\{ -\frac{1}{2}\left(\frac{x-1}{\sqrt{\theta}} \right)^2 \right\}$$

Consider the unbiased estimator for θ

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

Efficiency: example

The variance of $T(\mathbf{X})$ is

$$\begin{aligned} Var\left[T(\mathbf{X})\right] &= \frac{Var\left[(X-1)^2\right]}{n} \\ &= \frac{1}{n}\left\{E\left[(X-1)^4\right] - \left(E\left[(X-1)^2\right]\right)^2\right\} \\ &= \frac{1}{n}\left\{3\theta^2 - \theta^2\right\} \\ &= \frac{2\theta^2}{n} \end{aligned}$$

We have seen that

$$I_n(\theta) = \frac{n}{2\theta^2}$$

We can conclude that $T(\mathbf{X})$ is an efficient estimator for θ.

Multidimensional Cramer-Rao inequality

The Cramer-Rao ineqality (Theorem 3) can be generalized to a vector valued parameter θ

The generalization of the Cramer-Rao inequality states that, again subject to regularity conditions, the variance-covariance matrix of the unbiased estimator $T(\mathbf{X})$, the $k \times k$ matrix $\mathrm{Var}(T(\mathbf{X}))$ is such that $\mathrm{Var}(T(\mathbf{X})) - I_n^{-1}(\theta)$ is positive semi-definite.

Thus $I_n^{-1}(\theta)$ is in a sense a 'lower bound' for the variance matrix of an unbiased estimator of θ.

Lesson 4: Maximum Likelihood estimator

Umberto Triacca

Università dell'Aquila
Department of Computer Engineering, Computer Science and Mathematics, University of L'Aquila, L'Aquila, Italy
umberto.triacca@univaq.it

Maximum Likelihood Estimate

How can we estimate the unknown parameter θ? Given that the likelihood function represents the plausibility of the various $\theta \in \Theta$ given the realization $\mathbf{x}$, it is natural to chose as estimate of θ the most plausible element of Θ.

Maximum Likelihood Estimate

Definition 6. Let $\mathbf{x} = (x_1, ..., x_n)$ be a realization of a random sample from a distribution with p.d.f. $f(x; \theta)$ depending on an unknown parameter $\theta \in \Theta$. A **Maximum Likelihood Estimate** $\hat{\theta} = \hat{\theta}(x_1, ..., x_n)$ is an element of Θ that maximizes the value of $L(\theta; \mathbf{x})$, i.e.,

$$L(\hat{\theta}; \mathbf{x}) = \max_{\theta \in \Theta} L(\theta; \mathbf{x})$$

Maximum Likelihood estimate

There may be one, none or many such MLE's.

Proposition 1 (Sufficient condition for existence). If the parameter space Θ is **compact** and if the likelihood function $L(\theta; \mathbf{x})$ is **continuous** on Θ, then there exists an MLE.

Proposition 2 (Sufficient condition for uniqueness of MLE). If the parameter space Θ is **convex** and if the likelihood function $L(\theta; \mathbf{x})$ is **strictly concave** in Θ, then the MLE is unique when it exists.

Maximum Likelihood estimate

The logarithm is a monotonic function, so the values that maximize $L(\theta; \mathbf{x})$ are the same as those that maximize $\ln L(\theta; \mathbf{x})$.

$$L(\hat{\theta}; \mathbf{x}) = \max_{\theta \in \Theta} L(\theta; \mathbf{x})$$

$$\Updownarrow$$

$$\ln L(\hat{\theta}; \mathbf{x}) = \max_{\theta \in \Theta} \ln L(\theta; \mathbf{x})$$

Maximum Likelihood estimate

In the case where $L(\theta; \mathbf{x})$ is differentiable the MLE can be derived as a solution of the equation

$$\frac{\delta \ln L(\theta; \mathbf{x})}{\delta \theta} = \mathbf{0}$$

called **the likelihood equation**.

Maximum Likelihood estimator

- The likelihood equation represents the **first-order necessary condition** for the maximization of the log-likelihood function.

- The **second-order necessary condition** for a point to be the local maximum of the log-likelihood function is that the Hessian be negative semidefinite at the point.

Five steps for finding MLE

1. Find Likelihood function $L(\theta; \mathbf{x})$.
2. Get natural log of Likelihood function $l(\theta; \mathbf{x}) = ln(L(\theta; \mathbf{x})$.
3. Differentiate log-Likelihood function with respect to θ.
4. Set derivative to zero.
5. Solve for θ.

Maximum Likelihood Estimate: an example

Let $\mathbf{x} = (x_1, x_2, ..., x_n)$ be a realization of a random sample from an $N(\mu, \sigma^2)$ distribution with μ and σ unknown.
In this case $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$, and the likelihood function is

$$L(\mu, \sigma^2; \mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

The log-likelihood function is given by

$$l(\mu, \sigma^2; \mathbf{x}) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Maximum Likelihood Estimate: an example

Taking the first derivative (gradient), we get

$$\frac{\partial l(\theta; \mathbf{x})}{\partial \theta} = \left(\frac{\Sigma(x_i - \mu)}{\sigma^2}, \frac{\Sigma(x_i - \mu)^2}{2\sigma^4} - \frac{n}{2\sigma^2} \right)'.$$

Setting

$$\frac{\partial l(\theta; \mathbf{x})}{\partial \theta} = 0$$

and solve for $\theta = (\mu, \sigma^2)$ we have

$$\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2) = (\overline{x}, \frac{n-1}{n} s^2),$$

where $\overline{x} = \Sigma x_i / n$ is the sample mean and $s^2 = \Sigma(x_i - \overline{x})^2/(n-1)$ is the sample variance.
It is not difficult to verify that these values of μ and σ^2 yield an absolute (not only a local) maximum of the log-likelihood function, so that they are maximum likelihood estimates.

Maximum Likelihood Estimate

Sometimes it is not possible to find an explicit solution of the likelihood equation and so we have to use iterative algorithms to maximize $l(\theta; \mathbf{x})$, as the Newton-Raphson or the Fisher-scoring, which at any iteration update the parameter θ in appropriate way until convergence.

Lesson 5: Properties of the Maximum Likelihood Estimator

Umberto Triacca

Università dell'Aquila
Department of Computer Engineering, Computer Science and Mathematics, University of L'Aquila, L'Aquila, Italy
umberto.triacca@univaq.it

Maximum Likelihood estimator

Definition 7. Let $\mathbf{X} = (X_1, ..., X_n)$ be a random sample from a distribution with p.d.f. $f(x; \theta)$ depending on an unknown parameter $\theta \in \Theta$. An estimator $\hat{\theta}_n(\mathbf{X}) = \hat{\theta}_n(X_1, ..., X_n)$ of θ is a **Maximum Likelihood Estimator** if for any particular realization $\mathbf{x} = (x_1, ..., x_n)$, the resulting estimate $\hat{\theta}_n(\mathbf{x}) = \hat{\theta}_n(x_1, ..., x_n) \in \Theta$ is a Maximum Likelihood estimate i.e.,

$$L(\hat{\theta}_n(\mathbf{x}); \mathbf{x}) = \max_{\theta \in \Theta} L(\theta; \mathbf{x})$$

We will present some properties of MLE's in the context in which θ a single parameter, that is $\Theta \subset \mathbb{R}$.

Maximum Likelihood estimator: Invariance

One of the most attractive properties of MLE's is invariance.

Let $\hat{\theta}_n = \hat{\theta}_n(\mathbf{X})$ be a MLE of θ. If $g : \Theta \to \mathbb{R}$ is a continuous function, then a MLE of $g(\theta)$ exists and is given by $g(\hat{\theta}_n(\mathbf{X}))$.

For example, if $g(\theta) = \theta^2$ its MLE is $g(\hat{\theta}_n) = \hat{\theta}_n^2$.

Maximum Likelihood estimator: Unbiasedness and efficiency

It is possible to show that, under some regularity conditions, if $\hat{\theta}_n(\mathbf{X})$ is an unbiased estimator of θ whose variance achieves the Cramer-Rao bound, then the likelihood equation has a unique solution equal to $\hat{\theta}_n(\mathbf{x})$.

In other terms, when there exists an unbiased estimator whose variance attains the lower bound, this estimator is identical to the ML estimator.

Maximum Likelihood estimator: consistency

Definition 8. Let $\mathbf{X} = (X_1, ..., X_n)$ be a random sample from the distribution with p.d.f. $f(x; \theta)$ depending on a real parameter $\theta \in \Theta$. An estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, ..., X_n)$ is said to be **consistent** for θ if

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \epsilon) = 1 \quad \forall \theta \in \Theta$$

and we write $\hat{\theta}_n \xrightarrow{P} \theta$.

Maximum Likelihood estimator: consistency

Theorem 3. Let $\mathbf{X} = (X_1, ..., X_n)$ be a random sample from the distribution with p.d.f. $f(x; \theta)$ depending on a real parameter $\theta \in \Theta$. Under suitable regularity conditions, the ML estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, ..., X_n)$ is a consistent estimator for θ.

Maximum Likelihood estimator: asymptotic normality

Here, we consider θ a vector of parameters

Definition 9. Let $\mathbf{X} = (X_1, ..., X_n)$ be a random sample from the distribution with p.d.f. $f(x; \theta)$ depending on a vector of parameters $\theta \in \Theta \subset \mathbb{R}^k$. An estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, ..., X_n)$ for θ, with covariance matrix $\mathbf{V}_n(\theta)$, is said to be **asymptotically normal** if

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{D} N(\mathbf{0}, \mathbf{V}(\theta))$$

where $\mathbf{V}(\theta) = \lim_{n\to\infty} \mathbf{V}_n(\theta)$

We note that if $\hat{\theta}_n$ is asymptotically normal, then approximately

$$\hat{\theta}_n \sim N(\theta, \frac{1}{n}\mathbf{V}(\theta)).$$

The matrix $\frac{1}{n}\mathbf{V}(\theta)$ is called asymptotic variance.

Maximum Likelihood estimator: asymptotic normality

Theorem 4. Let $\mathbf{X} = (X_1, ..., X_n)$ be a random sample from the distribution with p.d.f. $f(x; \theta)$ depending on a vector of parameters $\theta \in \Theta \subset \mathbb{R}^k$. Under suitable regularity conditions, the ML estimator $\hat{\theta}_n = \hat{\theta}_n(x_1, ..., x_n)$ is asymptotically normal. That is

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{D} N(\mathbf{0}, I_a(\theta_0)^{-1})$$

where

$$I_a(\theta_0) = \lim_{n\to\infty} I_n(\theta_0)/n \quad (\textit{asymptotic information matrix})$$

$$I_n(\theta_0) = -E\left[\left.\frac{\delta^2 \ln f(\mathbf{X}; \theta)}{\delta\theta\delta\theta'}\right|_{\theta=\theta_0}\right]$$

and θ_0 is the true parameter value.

Maximum Likelihood estimator: asymptotic normality

Because

$$I_a(\theta_0) = \lim_{n\to\infty} \frac{I_n(\theta_0)}{n} = I_1(\theta_0),$$

we have that

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{D} N(\mathbf{0}, I_1(\theta_0)^{-1})$$

Maximum Likelihood estimator: asymptotic normality

The practical consequence of this result is that in large samples, when n is large enough, the ML estimator $\hat{\theta}$ has approximately a normal distribution with mean vector θ_0 and variance-covariance matrix $I_1(\theta_0)^{-1}/n$, in symbols

$$\hat{\theta} \;\; \textit{approx.} \;\; \sim N\left[\theta_0, I_1(\theta_0)^{-1}/n\right].$$

Maximum Likelihood estimator: asymptotic efficiency

Definition 10. Let $\mathbf{X} = (X_1, ..., X_n)$ be a random sample from the distribution with p.d.f. $f(x; \theta)$ depending on a vector of parameters $\theta \in \Theta \subset \mathbb{R}^k$. An consistent and asymptotically normal estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, ..., X_n)$ for θ, with asymptotic variance $(1/n)\mathbf{V}(\theta)$. is said to be **asymptotically efficient** if the asymptotic variance of any other consistent, asymptotically normally distributed estimator exceeds $(1/n)\mathbf{V}(\theta)$ by a nonnegative definite matrix.

Maximum Likelihood estimator: asymptotic efficiency

Theorem 5. Let $\mathbf{X} = (X_1, ..., X_n)$ be a random sample from the distribution with p.d.f. $f(x; \theta)$ depending on a vector of parameters $\theta \in \Theta \subset \mathbb{R}^k$. Under suitable regularity conditions, the ML estimator $\hat{\theta}_n = \hat{\theta}_n(x_1, ..., x_n)$ is asymptotically efficient.

Maximum Likelihood estimator: Properties

Under certain regularity conditions, the maximum likelihood estimator possesses many appealing properties:

1. The maximum likelihood estimator is equivariant
2. The maximum likelihood estimator is consistent
3. The maximum likelihood estimator is asymptotically normal
4. The maximum likelihood estimator is asymptotically efficient