POLITECNICO DI MILANO

Maximum likelihood estimation

Marco Lovera

Dipartimento di Scienze e Tecnologie Aerospaziali, Politecnico di Milano

Problem statement

- Assume that N independent identically distributed observations

$$x_i, \quad i = 1, \ldots, N$$

are available.

- The measurements are distributed according to

$$x_i \simeq f(q_i|\theta), \quad i = 1, \quad , N.$$

- Then, the *likelihood function* is defined as the joint probability of the observed data-set:

$$L(x|\theta) = f(x_1|\theta) f(x_2|\theta) \ldots f(x_N|\theta).$$

Maximum likelihood principle

- The ML principle consist in choosing as estimate of the parameter the one which makes the likelihood as large as possibile:

$$\widehat{\theta}_N : \qquad L(x|\widehat{\theta}_N) \geq L(x|\theta).$$

- Intuitive intepretation:
  - the drawn sample was «the most probable» one;
  - so the value of the estimate which makes the probability of the dataset as large as possible must be close to the true value of the parameter.

- This intuitive idea leads to a systematic approach to estimator design, with many useful properties.

Example

We measure samples drawn from

$$f(q) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(q-\mu)^2}{2\sigma^2}}$$

and we want to estimate the expected value from a single observation.

The likelihood in this case is simply

$$L(x_1|\mu) = f(x_1) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}}$$

Example

- Note that now the data point is fixed and the likelihood is a function of the parameter:

$$L(x_1|\mu) = \frac{1}{\sigma\sqrt{2\pi}} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\mu - x_1)^2}{2\sigma^2}}$$

- This function can be interpreted as a Gaussian density with expected value equal to the data point.

- As the expected value of a Gaussian is its maximum, we see that the maximum likelihood estimator is

$$\widehat{\mu}_1 = x_1.$$

Example

- What if we now have *N* samples

$$x_1, x_2, \ldots, x_N$$

drawn independently from the same density:

$$f(q) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(q-\mu)^2}{2\sigma^2}}$$

- and we want to estimate the expected value from the *N* observations.

Example

The likelihood of the data set can then be written as

$$L(x_1, x_2, \ldots, x_K|\theta) = f_1(x_1|\theta) f_2(x_2|\theta) \ldots f_N(x_N|\theta)$$

$$L(x_1, x_2, \ldots, x_N|\theta) = \frac{1}{(\sigma\sqrt{2\pi})^N} e^{-\frac{(x_1-\mu))^2}{2\sigma^2}} \ldots e^{-\frac{(x_N-\mu)^2}{2\sigma^2}}$$

$$L(x_1, x_2, \ldots, x_N|\theta) = \frac{1}{(\sigma\sqrt{2\pi})^N} e^{-\frac{\sum_{i=1}^{N}(x_k-\mu)^2}{2\sigma^2}}$$

Example

- Note now that maximising the likelihood is the same as maximising its logarithm.

- Indeed $L \to \log L$ is a monotonic transformation which does not change the location of maxima.

- The logarithm of the likelihood is

$$\log L(x_1, x_2, \ldots, x_N|\theta) = -N \log(\sigma\sqrt{2\pi}) - \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{2\sigma^2}$$

and its derivative with respect to the parameter:

$$\frac{\partial \log L}{\partial \theta} = \frac{\sum_{i=1}^{N}(x_i - \mu)}{\sigma^2}$$

Example

- Therefore imposing stationarity we have

$$\frac{\partial \log L}{\partial \theta} = 0 \quad \Rightarrow \quad \widehat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

- The sample mean is a maximum likelihood estimator…

Determination of maximum likelihood estimators

- Generally difficult to find closed forms

- Easier if L is twice differentiable with respect to the parameter.

- In this case stationary points are given by

$$L' = \frac{\partial L}{\partial \theta} = 0$$

- And the sufficient condition for local maxima

$$L'' < 0$$

can be used.

Determination of maximum likelihood estimators

- As in the previous example, it is frequently easier to work with the logarithm of L (recall that L > 0 by definition).

- Therefore letting $\dfrac{\partial \log L}{\partial \theta} = \dfrac{L'}{L} = (\log L)'$

  we seek estimators such that

$$(\log L)' = 0 \quad (\log L)'' < 0.$$

Example

We measure a sample of *N* i.i.d. data drawn from

$$f(q) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(q-\mu)^2}{2\sigma^2}}$$

and we want to estimate the expected value AND the variance from the available dataset.

The likelihood is the same as in the previous example:

$$L(x|\theta) = f(x_1|\theta) f(x_2|\theta) \dots f(x_N|\theta), \qquad \theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$$

Example

- The logarithm of the likelihood is

$$\log L(x_1, x_2, \ldots, x_N|\theta) = -\frac{N}{2}\log(\sigma^2 2\pi) - \frac{\sum_{i=1}^N (x_i-\mu)^2}{2\sigma^2}$$

and its derivative with respect to the parameters:

$$\frac{\partial \log L}{\partial \mu} = \frac{\sum_{i=1}^N (x_i-\mu)}{\sigma^2} = 0 \Rightarrow \qquad \hat{\mu}_N = \frac{1}{N}\sum_{i=1}^N x_i$$

$$\frac{\partial \log L}{\partial \sigma^2} = \frac{1}{2}\frac{\sum_{i=1}^N (x_i-\mu)^2}{(\sigma^2)^2} - \frac{N}{2}\frac{1}{\sigma^2} = 0$$

$$\Rightarrow \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^N (x_i-\mu)^2$$

$$\Rightarrow \qquad \hat{\sigma}_N^2 = \frac{1}{N}\sum_{i=1}^N (x_i-\hat{\mu}_N)^2.$$

- So the maximum likelihood solution to the problem is

$$\widehat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\widehat{\sigma}_N^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \widehat{\mu}_N)^2.$$

- Recall that we have seen that this estimator for the variance is biased for finite samples:

$$E[\widehat{\sigma}_N^2] = \frac{N-1}{N}\sigma^2.$$

- Therefore the ML estimate is only asymptotically unbiased.

Properties of ML estimators

- ML estimators have a number of useful properties, which motivate their widespread use in applications.

- ML estimators are asymptotically unbiased:

$$E[\widehat{\theta}_N] \underset{N\to\infty}{\longrightarrow} \theta.$$

- BUT they may be biased for finite N.

- ML estimators are consistent:

$$\underset{N\to\infty}{plim}\, \widehat{\theta}_N = \theta.$$

- ML estimators are efficient:

$$Var[\widehat{\theta}_N] \underset{N\to\infty}{\rightarrow} M^{-1}.$$

- And finally, ML estimators are asymptotically Gaussian:

$$\widehat{\theta}_N \underset{N\to\infty}{\rightarrow} G(\theta, M^{-1}) \quad \text{in distribution.}$$