Maximum likelihood estimation

by Marco Taboga, PhD

Maximum likelihood estimation (MLE) is an estimation method that allows us to use a sample to estimate the parameters of the probability distribution that generated the sample.

This lecture provides an introduction to the theory of maximum likelihood, focusing on its mathematical aspects, in particular on:

- its asymptotic properties;
- the assumptions that are needed to prove the properties.

At the end of the lecture, we provide links to pages that contain examples and that treat practically relevant aspects of the theory, such as numerical optimization and hypothesis testing.

The sample and its likelihood

The main elements of a maximum likelihood estimation problem are the following:

JPMF of sample is roughly "f" of θ.

What does this mean?

- a sample ξ, that we use to make statements about the probability distribution that generated the sample;
- the sample ξ is regarded as the realization of a random vector Ξ, whose distribution is unknown and needs to be estimated;
- there is a set $\Theta \subseteq \mathbb{R}^p$ of real vectors (called the parameter space) whose elements (called parameters) are put into correspondence with the possible distributions of Ξ; in particular:
  - if Ξ is a discrete random vector, we assume that its joint probability mass function $p_\Xi(\xi;\theta_0)$ belongs to a set of joint probability mass functions $p_\Xi(\xi;\theta)$ indexed by the parameter θ; when the joint probability mass function is considered as a function of θ for fixed ξ, it is called **likelihood** (or **likelihood function**) and it is denoted by
  $$L(\theta;\xi) = p_\Xi(\xi;\theta)$$
  Not θ0.
  - if Ξ is a continuous random vector, we assume that its joint probability density function $f_\Xi(\xi;\theta_0)$ belongs to a set of joint probability density functions $f_\Xi(\xi;\theta)$ indexed by the parameter θ; when the joint probability density function is considered as a function of θ for fixed ξ, it is called likelihood and it is denoted by
  $$L(\theta;\xi) = f_\Xi(\xi;\theta)$$
- we need to estimate the true parameter θ_0, which is associated with the unknown distribution that actually generated the sample (we rule out the possibility that several different parameters are put into correspondence with true distribution).

One sample is one discrete random vector!

What is the [?] that explains the sample JPMF.

What is θ?

Maximum likelihood estimator

A maximum likelihood estimator $\widehat{\theta}$ of θ_0 is obtained as a solution of a maximization problem:

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} L(\theta;\xi)$$

In other words, $\widehat{\theta}$ is the parameter that maximizes the likelihood of the sample ξ. $\widehat{\theta}$ is called the **maximum likelihood estimator** of θ.

In what follows, the symbol $\widehat{\theta}$ will be used to denote both a maximum likelihood estimator (a random variable) and a maximum likelihood estimate (a realization of a random variable): the meaning will be clear from the context.

The same estimator $\widehat{\theta}$ is obtained as a solution of

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} \ln[L(\theta;\xi)]$$

i.e., by maximizing the natural logarithm of the likelihood function. Solving this problem is equivalent to solving the original one, because the logarithm is a strictly increasing function. The logarithm of the likelihood is called **log-likelihood** and it is denoted by

$$l(\theta;\xi) = \ln[L(\theta;\xi)]$$

Asymptotic properties

To derive the (asymptotic) properties of maximum likelihood estimators, one needs to specify a set of assumptions about the sample ξ and the parameter space Θ.

when the input approaches ∞.

The next section presents a set of assumptions that allows us to easily derive the asymptotic properties of the maximum likelihood estimator. Some of the assumptions are quite restrictive, while others are very generic. Therefore, the subsequent sections discuss how the most restrictive assumptions can be weakened and how the most generic ones can be made more specific.

Note: the presentation in this section does not aim at being one hundred per cent rigorous. Its aim is rather to introduce the reader to the main steps that are necessary to derive the asymptotic properties of maximum likelihood estimators. Therefore, some technical details are either skipped or de-emphasized. After getting a grasp of the main issues related to the asymptotic properties of MLE, the interested reader can refer to other sources (e.g., Newey and McFadden - 1994, Ruud - 2000) for a fully rigorous presentation of MLE theory.

Assumptions

Let $\{X_n\}$ be a sequence of $K\times 1$ random vectors. Denote by ξ_n the sample comprising the first n realizations of the sequence

$$\xi_n = \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix}$$

which is a realization of the random vector

$$\Xi_n = \begin{bmatrix} X_1 & \ldots & X_n \end{bmatrix}$$

We assume that:

1. **IID.** $\{X_n\}$ is an IID sequence.
2. **Continuous variables.** A generic term X_j of the sequence $\{X_n\}$ is a continuous random vector, whose joint probability density function
$$f_{X_j}(x_j;\theta_0)$$
belongs to a set of joint probability density functions $f_X(x;\theta)$ indexed by a $K\times 1$ parameter $\theta\in\Theta$ (where we have dropped the subscript j to highlight the fact that the terms of the sequence are identically distributed).
3. **Identification.** If $\theta\neq\theta_0$, then the ratio
$$\frac{f_X(X_j;\theta)}{f_X(X_j;\theta_0)}$$
is not almost surely constant. This also implies that the parametric family is identifiable: there does not exist another parameter $\theta\neq\theta_0$ such that $f_{X_j}(x_j;\theta)$ is the true probability density function of X_j.
4. **Integrable log-likelihood.** The log-likelihood is integrable:
$$\mathrm{E}[|\ln(f_X(X_j;\theta))|] < \infty,\ \forall\theta\in\Theta$$
5. **Maximum.** The density functions $f_X(x;\theta)$ and the parameter space Θ are such that there always exists a unique solution $\widehat{\theta}_n$ of the maximization problem:
$$\widehat{\theta}_n = \arg\max_{\theta\in\Theta} L(\theta;\xi_n) = \arg\max_{\theta\in\Theta} \prod_{j=1}^n f_X(x_j;\theta)$$
where the rightmost equality is a consequence of independence (see the IID assumption above). Of course, this is the same as
$$\widehat{\theta}_n = \arg\max_{\theta\in\Theta} l(\theta;\xi_n) = \arg\max_{\theta\in\Theta} \sum_{j=1}^n l_j(\theta;x_j)$$
where $l(\theta;\xi_n)$ is the log-likelihood and
$$l_j(\theta;x_j) = \ln(f_X(x_j;\theta))$$
are the contributions of the individual observations to the log-likelihood. It is also the same as
$$\widehat{\theta}_n = \arg\max_{\theta\in\Theta} \frac{1}{n} l(\theta;\xi_n) = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{j=1}^n l_j(\theta;x_j)$$
6. **Exchangeability of limit.** The density functions $f_X(x;\theta)$ and the parameter space Θ are such that
$$\operatorname*{plim}_{n\to\infty}\left(\arg\max_{\theta\in\Theta}\frac{1}{n}l(\theta;\Xi_n)\right) = \arg\max_{\theta\in\Theta}\left(\operatorname*{plim}_{n\to\infty}\frac{1}{n}l(\theta;\Xi_n)\right)$$
where plim denotes a limit in probability. Roughly speaking, the probability limit can be brought inside the $\arg\max$ operator.
7. **Differentiability.** The log-likelihood $l(\theta;\xi_n)$ is two times continuously differentiable with respect to θ in a neighborhood of θ_0.
8. **Other technical conditions.** The derivatives of the log-likelihood $l(\theta;\xi_n)$ are well-behaved, so that it is possible to exchange integration and differentiation, compute their first and second moments, and probability limits involving their entries are also well-behaved.

Information inequality

Given the assumptions made above, we can derive an important fact about the expected value of the log-likelihood:

$$\mathrm{E}[l(\theta_0;\Xi_n)] > \mathrm{E}[l(\theta;\Xi_n)],\ \forall\theta\neq\theta_0$$

Proof

This inequality, called **information inequality** by many authors, is essential for proving the consistency of the maximum likelihood estimator.

Consistency

Given the assumptions above, the maximum likelihood estimator $\widehat{\theta}_n$ is a consistent estimator of the true parameter θ_0:

$$\operatorname*{plim}_{n\to\infty}\widehat{\theta}_n = \theta_0$$

where plim denotes a limit in probability.

Proof

Score vector

Denote by $\nabla_\theta l(\theta;\Xi_n)$ the gradient of the log-likelihood, that is, the vector of first derivatives of the log-likelihood, evaluated at the point θ. This vector is often called the **score vector**.

Given the assumptions above, the score has zero expected value:

$$\mathrm{E}[\nabla_\theta l(\theta_0;\Xi_n)] = 0$$

Proof

Information matrix

Given the assumptions above, the covariance matrix of the score (called **information matrix** or Fisher information matrix) is

$$\mathrm{Var}[\nabla_\theta l(\theta_0;\Xi_n)] = -\mathrm{E}[\nabla_{\theta\theta} l(\theta_0;\Xi_n)]$$

where $\nabla_{\theta\theta} l(\theta;\Xi_n)$ is the Hessian of the log-likelihood, that is, the matrix of second derivatives of the log-likelihood, evaluated at the point θ.

Proof

The latter equality is often called **information equality**.

Asymptotic normality

The maximum likelihood estimator is asymptotically normal:

$$\sqrt{n}\left(\widehat{\theta}_n-\theta_0\right)\xrightarrow{d} N\left(0,(\mathrm{Var}[\nabla_\theta\ln(f_X(X;\theta_0))])^{-1}\right)$$

In other words, the distribution of the maximum likelihood estimator $\widehat{\theta}_n$ can be approximated by a multivariate normal distribution with mean θ_0 and covariance matrix

$$\frac{1}{n}(\mathrm{Var}[\nabla_\theta\ln(f_X(X;\theta_0))])^{-1}$$

Proof

By the information equality (see its proof), the asymptotic covariance matrix is equal to the negative of the expected value of the Hessian matrix:

$$\mathrm{Var}[\nabla_\theta\ln(f_X(X;\theta_0))] = -\mathrm{E}[\nabla_{\theta\theta}\ln(f_X(X;\theta_0))]$$

Different assumptions

As previously mentioned, some of the assumptions made above are quite restrictive, while others are very generic. We now discuss how the former can be weakened and how the latter can be made more specific.

Assumption 1 (IID). It is possible to relax the assumption that $\{X_n\}$ is IID and allow for some dependence among the terms of the sequence (see, e.g., Bierens - 2004 for a discussion). In case dependence is present, the formula for the asymptotic covariance matrix of the MLE given above is no longer valid and needs to be replaced by a formula that takes serial correlation into account.

Assumption 2 (continuous variables). It is possible to prove consistency and asymptotic normality also when the terms of the sequence $\{X_n\}$ are extracted from a discrete distribution, or from a distribution that is neither discrete nor continuous (see, e.g., Newey and McFadden - 1994).

Assumption 3 (identification). Typically, different identification conditions are needed when the IID assumption is relaxed (e.g., Bierens - 2004).

Assumption 5 (maximum). To ensure the existence of a maximum, requirements are typically imposed both on the parameter space and on the log-likelihood function. For example, it can be required that the parameter space be compact (closed and bounded) and the log-likelihood function be continuous. Also, the parameter space can be required to be convex and the log-likelihood function strictly concave (e.g.: Newey and McFadden - 1994).

Assumption 6 (exchangeability of limit). To ensure the exchangeability of the limit and the $\arg\max$ operator, the following condition is often imposed:

$$\mathrm{E}\left[\sup_{\theta\in\Theta}|\ln(f_X(X_j;\theta))|\right] < \infty$$

Assumption 8 (other technical conditions). See, for example, Newey and McFadden (1994) for a discussion of these technical conditions.

Numerical optimization

In some cases, the maximum likelihood problem has an analytical solution. That is, it is possible to write the maximum likelihood estimator $\widehat{\theta}$ explicitly as a function of the data.

However, in many cases there is no explicit solution. In these cases, numerical optimization algorithms are used to maximize the log-likelihood. The lecture entitled Maximum likelihood - Algorithm discusses these algorithms.

Examples

The following lectures provide detailed examples of how to derive analytically the maximum likelihood (ML) estimators and their asymptotic variance:

- ML estimation of the parameter of the Poisson distribution
- ML estimation of the parameter of the exponential distribution
- ML estimation of the parameters of the normal distribution
- ML estimation of the parameters of the multivariate normal distribution
- ML estimation of the parameters of a normal linear regression model

The following lectures provides examples of how to perform maximum likelihood estimation numerically:

- ML estimation of the degrees of freedom of a standard t distribution (MATLAB example)
- ML estimation of the coefficients of a logistic classification model
- ML estimation of the coefficients of a probit classification model
- ML estimation of the parameters of a Gaussian mixture

More details

The following sections contain more details about the theory of maximum likelihood estimation.

Estimation of the asymptotic covariance matrix

Methods to estimate the asymptotic covariance matrix of maximum likelihood estimators, including OPG, Hessian and Sandwich estimators, are discussed in the lecture entitled Maximum likelihood - Covariance matrix estimation.

Hypothesis testing

Tests of hypotheses on parameters estimated by maximum likelihood are discussed in the lecture entitled Maximum likelihood - Hypothesis testing, as well as in the lectures on the three classical tests:

1. Wald test;
2. score test;
3. likelihood ratio test.

References

Bierens, H. J. (2004) Introduction to the mathematical and statistical foundations of econometrics, Cambridge University Press.

Newey, W. K. and D. McFadden (1994) "Chapter 35: Large sample estimation and hypothesis testing", in Handbook of Econometrics, Elsevier.

Ruud, P. A. (2000) An introduction to classical econometric theory, Oxford University Press.

How to cite

Please cite as:

Taboga, Marco (2021). "Maximum likelihood estimation", Lectures on probability theory and mathematical statistics. Kindle Direct Publishing. Online appendix. https://www.statlect.com/fundamentals-of-statistics/maximum-likelihood.