

MAXIMUM LIKELIHOOD ESTIMATION*

Clayton Scott
Robert Nowak

This work is produced by The Connexions Project and licensed under the Creative Commons Attribution License †

Abstract

This module introduces the maximum likelihood estimator. We show how the MLE implements the likelihood principle. Methods for computing the MLE are covered. Properties of the MLE are discussed including asymptotic efficiency and invariance under reparameterization.

The **maximum likelihood estimator** (MLE) is an alternative to the minimum variance unbiased estimator (MVUE). For many estimation problems, the MVUE does not exist. Moreover, when it does exist, there is no systematic procedure for finding it. In contrast, the MLE does not necessarily satisfy any optimality criterion, but it can almost always be computed, either through exact formulas or numerical techniques. For this reason, the MLE is one of the most common estimation procedures used in practice.

The MLE is an important type of estimator for the following reasons:

1. The MLE implements the likelihood principle.
2. MLEs are often simple and easy to compute.
3. MLEs have asymptotic optimality properties (consistency and efficiency).
4. MLEs are invariant under reparameterization.
5. If an efficient estimator exists, it is the MLE.
6. In signal detection with unknown parameters (composite hypothesis testing), MLEs are used in implementing the generalized likelihood ratio test (GLRT).

This module will discuss these properties in detail, with examples.

1 The Likelihood Principle

Supposed the data \mathbf{X} is distributed according to the density or mass function $p(\mathbf{x}|\theta)$. The **likelihood function** for θ is defined by

$$l(\theta|\mathbf{x}) \equiv p(\mathbf{x}|\theta)$$

At first glance, the likelihood function is nothing new - it is simply a way of rewriting the pdf/pmf of \mathbf{X} . The difference between the likelihood and the pdf or pmf is what is held fixed and what is allowed to vary. When we talk about the likelihood, we view the observation \mathbf{x} as being fixed, and the parameter θ as freely varying.

*Version 1.5: May 12, 2004 11:50 am GMT-5

†<http://creativecommons.org/licenses/by/1.0>

NOTE: It is tempting to view the likelihood function as a probability density for θ , and to think of $l(\theta | \mathbf{x})$ as the conditional density of θ given \mathbf{x} . This approach to parameter estimation is called **fiducial inference**, and is not accepted by most statisticians. One potential problem, for example, is that in many cases $l(\theta | \mathbf{x})$ is not integrable ($\int l(\theta | \mathbf{x}) d\theta \rightarrow \infty$) and thus cannot be normalized. A more fundamental problem is that θ is viewed as a fixed quantity, as opposed to random. Thus, it doesn't make sense to talk about its density. For the likelihood to be properly thought of as a density, a Bayesian approach is required.

The likelihood principle effectively states that all information we have about the unknown parameter θ is contained in the likelihood function.

Rule 1: Likelihood Principle

The information brought by an observation \mathbf{x} about θ is entirely contained in the likelihood function $p(\mathbf{x} | \theta)$. Moreover, if x_1 and x_2 are two observations depending on the same parameter θ , such that there exists a constant c satisfying $p(x_1 | \theta) = cp(x_2 | \theta)$ for every θ , then they bring the same information about θ and must lead to identical estimators.

In the statement of the likelihood principle, it is **not** assumed that the two observations x_1 and x_2 are generated according to the same model, as long as the model is parameterized by θ .

Example 1

Suppose a public health official conducts a survey to estimate $0 \leq \theta \leq 1$, the percentage of the population eating pizza at least once per week. As a result, the official found nine people who had eaten pizza in the last week, and three who had not. If no additional information is available regarding how the survey was implemented, then there are at least two probability models we can adopt.

1. The official surveyed 12 people, and 9 of them had eaten pizza in the last week. In this case, we observe $x_1 = 9$, where

$$x_1 \sim \text{Binomial}(12, \theta)$$

The density for x_1 is

$$f(x_1 | \theta) = \binom{12}{x_1} \theta^{x_1} (1 - \theta)^{12 - x_1}$$

2. Another reasonable model is to assume that the official surveyed people **until** he found 3 non-pizza eaters. In this case, we observe $x_2 = 12$, where

$$x_2 \sim \text{NegativeBinomial}(3, 1 - \theta)$$

The density for x_2 is

$$g(x_2 | \theta) = \binom{x_2 - 1}{3 - 1} \theta^{x_2 - 3} (1 - \theta)^3$$

The likelihoods for these two models are proportional:

$$\left(\ell(\theta | x_1) \propto \ell(\theta | x_2) \propto \theta^9 (1 - \theta)^3 \right)$$

Therefore, any estimator that adheres to the likelihood principle will produce the same estimate for θ , regardless of which of the two data-generation models is assumed.

The likelihood principle is widely accepted among statisticians. In the context of parameter estimation, any reasonable estimator should conform to the likelihood principle. As we will see, the maximum likelihood estimator does.

NOTE: While the likelihood principle itself is a fairly reasonable assumption, it can also be derived from two somewhat more intuitive assumptions known as the **sufficiency principle** and the **conditionality principle**. See *Casella and Berger, Chapter 6*[1].

2 The Maximum Likelihood Estimator

The **maximum likelihood estimator** $\hat{\theta}(\mathbf{x})$ is defined by

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} (l(\theta | \mathbf{x}))$$

Intuitively, we are choosing θ to maximize the probability of occurrence of the observation \mathbf{x} .

NOTE: It is possible that multiple parameter values maximize the likelihood for a given \mathbf{x} . In that case, any of these maximizers can be selected as the MLE. It is also possible that the likelihood may be **unbounded**, in which case the MLE does not exist.

The MLE rule is an implementation of the likelihood principle. If we have two observations whose likelihoods are proportional (they differ by a constant that does not depend on θ), then the value of θ that maximizes one likelihood will also maximize the other. In other words, both likelihood functions lead to the same inference about θ , as required by the likelihood principle.

Understand that maximum likelihood is a **procedure**, not an optimality criterion. From the definition of the MLE, we have no idea how close it comes to the true parameter value relative to other estimators. In contrast, the MVUE is defined as the estimator that satisfies a certain optimality criterion. However, unlike the MLE, we have no clear procedure to follow to compute the MVUE.

3 Computing the MLE

If the likelihood function is differentiable, then $\hat{\theta}$ is found by differentiating the likelihood (or log-likelihood), equating with zero, and solving:

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

If multiple solutions exist, then the MLE is the solution that maximizes $\log(l(\theta | \mathbf{x}))$, that is, the **global** maximizer.

In certain cases, such as pdfs or pmfs with an exponential form, the MLE can be easily solved for. That is,

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

can be solved using calculus and standard linear algebra.

Example 2: DC level in white Gaussian noise

Suppose we observe an unknown amplitude in white Gaussian noise with unknown variance:

$$x_n = A + w_n$$

$n \in \{0, 1, \dots, N-1\}$, where $w_n \sim [\text{U+EF3B}] (0, \sigma^2)$ are independent and identically distributed. We would like to estimate

$$\theta = \begin{pmatrix} A \\ \sigma^2 \end{pmatrix}$$

by computing the MLE. Differentiating the log-likelihood gives

$$\frac{\partial}{\partial A} (\log(p(\mathbf{x} | \theta))) = \frac{1}{\sigma^2} \sum_{n=1}^N (x_n - A)$$

$$\frac{\partial}{\partial \sigma^2} (\log(p(\mathbf{x}|\theta))) = -\left(\frac{N}{\sigma^2}\right) + \frac{1}{2\sigma^4} \sum_{n=1}^N ((x_n - A)^2)$$

Equating with zero and solving gives us our MLEs:

$$\hat{A} = \frac{1}{N} \sum_{n=1}^N x_n$$

and

$$\widehat{\sigma^2} = \frac{1}{N} \sum_{n=1}^N \left((x_n - \hat{A})^2 \right)$$

NOTE: $\widehat{\sigma^2}$ is biased!

As an exercise, try the following problem:

Exercise 1

Suppose we observe a random sample $\mathbf{x} = (x_1, \dots, x_N)^T$ of Poisson measurements with intensity λ : $Pr[x_i = n] = e^{-\lambda} \frac{\lambda^n}{n!}$, $n \in \{0, 1, 2, \dots\}$. Find the MLE for λ .

Unfortunately, this approach is only feasible for the most elementary pdfs and pmfs. In general, we may have to resort to more advanced numerical maximization techniques:

1. **Newton-Raphson** iteration
2. Iteration by the **Scoring Method**
3. **Expectation-Maximization Algorithm**

All of these are iterative techniques which posit some initial guess at the MLE, and then incrementally update that guess. The iteration proceeds until a local maximum of the likelihood is attained, although in the case of the first two methods, such convergence is not guaranteed. The EM algorithm has the advantage that the likelihood is always increased at each iteration, and so convergence to at least a local maximum is guaranteed (assuming a bounded likelihood). For each algorithm, the final estimate is highly dependent on the initial guess, and so it is customary to try several different starting values. For details on these algorithms, see *Kay, Vol. I*[2].

4 Asymptotic Properties of the MLE

Let $\mathbf{x} = (x_1, \dots, x_N)^T$ denote an IID sample of size N , and each sample is distributed according to $p(\mathbf{x}|\theta)$. Let $\hat{\theta}_N$ denote the MLE based on a sample \mathbf{x} .

Theorem 1: Asymptotic Properties of MLE

If the likelihood $\ell(\theta|\mathbf{x}) = p(\mathbf{x}|\theta)$ satisfies certain "regularity" conditions¹, then the MLE $\hat{\theta}_N$ is **consistent**, and moreover, $\hat{\theta}_N$ converges in probability to $\hat{\theta}$, where

$$\hat{\theta} \sim [\text{U+EF3B}] \left(\theta, (I(\theta))^{-1} \right)$$

where $I(\theta)$ is the **Fisher Information matrix** evaluated at the true value of θ .

Since the mean of the MLE tends to the true parameter value, we say the MLE is **asymptotically unbiased**. Since the covariance tends to the inverse Fisher information matrix, we say the MLE is **asymptotically efficient**.

¹The regularity conditions are essentially the same as those assumed for the Cramer-Rao lower bound (<<http://cnx.org/content/m11429/latest/>>): the log-likelihood must be twice differentiable, and the expected value of the first derivative of the log-likelihood must be zero.

In general, the rate at which the mean-squared error converges to zero is not known. It is possible that for small sample sizes, some other estimator may have a smaller MSE. The proof of consistency is an application of the weak law of large numbers. Derivation of the asymptotic distribution relies on the central limit theorem. The theorem is also true in more general settings (e.g., dependent samples). See, *Kay, Vol. I, Ch. 7*[2] for further discussion.

5 The MLE and Efficiency

In some cases, the MLE is efficient, not just asymptotically efficient. In fact, when an efficient estimator exists, it must be the MLE, as described by the following result:

Theorem 2:

If $\hat{\theta}$ is an efficient estimator, and the Fisher information matrix $I(\theta)$ is positive definite for all θ , then $\hat{\theta}$ maximizes the likelihood.

Proof:

Recall the $\hat{\theta}$ is efficient (meaning it is unbiased and achieves the Cramer-Rao lower bound) if and only if

$$\frac{\partial}{\partial \theta} (\ln(p(\mathbf{x}|\theta))) = I(\theta) (\hat{\theta} - \theta)$$

for all θ and \mathbf{x} . Since $\hat{\theta}$ is assumed to be efficient, this equation holds, and in particular it holds when $\theta = \widehat{\theta(\mathbf{x})}$. But then the derivative of the log-likelihood is zero at $\theta = \widehat{\theta(\mathbf{x})}$. Thus, $\hat{\theta}$ is a critical point of the likelihood. Since the Fisher information matrix, which is the negative of the matrix of second order derivatives of the log-likelihood, is positive definite, $\hat{\theta}$ must be a maximum of the likelihood.

An important case where this happens is described in the following subsection.

5.1 Optimality of MLE for Linear Statistical Model

If the observed data \mathbf{x} are described by

$$\mathbf{x} = H\theta + \mathbf{w}$$

where H is $N \times p$ with full rank, θ is $p \times 1$, and $\mathbf{w} \sim [\mathbf{U}+\text{EF3B}] (\mathbf{0}, C)$, then the MLE of θ is

$$\hat{\theta} = (H^T C^{-1} H)^{-1} H^T C^{-1} \mathbf{x}$$

This can be established in two ways. The first is to compute the CRLB for θ . It turns out that the condition for equality in the bound is satisfied, and $\hat{\theta}$ can be read off from that condition.

The second way is to maximize the likelihood directly. Equivalently, we must minimize

$$(\mathbf{x} - H\theta)^T C^{-1} (\mathbf{x} - H\theta)$$

with respect to θ . Since C^{-1} is positive definite, we can write $C^{-1} = U^T \Lambda U = D^T D$, where $D = \Lambda^{\frac{1}{2}} U$, where U is an orthogonal matrix whose columns are eigenvectors of C^{-1} , and Λ is a diagonal matrix with positive diagonal entries. Thus, we must minimize

$$(D\mathbf{x} - DH\theta)^T (D\mathbf{x} - DH\theta)$$

But this is a linear least squares problem, so the solution is given by the pseudoinverse of DH :

$$\begin{aligned} \hat{\theta} &= \left((DH)^T (DH) \right)^{-1} (DH)^T (D\mathbf{x}) \\ &= (H^T C^{-1} H)^{-1} H^T C^{-1} \mathbf{x} \end{aligned} \tag{1}$$

Exercise 2

Consider $X_1, \dots, X_N \sim [\text{U+EF3B}] (\mathbf{s}, \sigma^2 I)$, where \mathbf{s} is a $p \times 1$ unknown signal, and σ^2 is known. Express the data in the linear model and find the MLE $\hat{\mathbf{s}}$ for the signal.

6 Invariance of MLE

Suppose we wish to estimate the function $\mathbf{w} = W(\theta)$ and not θ itself. To use the maximum likelihood approach for estimating \mathbf{w} , we need an expression for the likelihood $\ell(\mathbf{w} | \mathbf{x}) = p(\mathbf{x} | \mathbf{w})$. In other words, we would need to be able to parameterize the distribution of the data by \mathbf{w} . If W is not a one-to-one function, however, this may not be possible. Therefore, we define the **induced** likelihood

$$\ell(\mathbf{w} | \mathbf{x}) = \max_{\theta} \{W(\theta) = \mathbf{w}\} \ell(\theta | \mathbf{x})$$

The MLE $\hat{\mathbf{w}}$ is defined to be the value of \mathbf{w} that maximizes the induced likelihood. With this definition, the following invariance principle is immediate.

Theorem 3:

Let $\hat{\theta}$ denote the MLE of θ . Then $\hat{\mathbf{w}} = W(\hat{\theta})$ is the MLE of $\mathbf{w} = W(\theta)$.

Proof:

The proof follows directly from the definitions of $\hat{\theta}$ and $\hat{\mathbf{w}}$. As an exercise, work through the logical steps of the proof on your own.

Example

Let $\mathbf{x} = (x_1, \dots, x_N)^T$ where

$$x_i \sim \text{Poisson}(\lambda)$$

Given \mathbf{x} , find the MLE of the probability that $x \sim \text{Poisson}(\lambda)$ exceeds the mean λ .

$$W(\lambda) = Pr[x > \lambda] = \sum_{n=\lfloor \lambda+1 \rfloor}^{\infty} \left(e^{-\lambda} \frac{\lambda^n}{n!} \right)$$

where $\lfloor z \rfloor =$ largest integer $\leq z$. The MLE of w is

$$\hat{w} = \sum_{n=\lfloor \hat{\lambda}+1 \rfloor}^{\infty} \left(e^{-(\hat{\lambda})} \frac{(\hat{\lambda})^n}{n!} \right)$$

where $\hat{\lambda}$ is the MLE of λ :

$$\hat{\lambda} = \frac{1}{N} \sum_{n=1}^N x_n$$

Be aware that the MLE of a **transformed** parameter does not necessarily satisfy the asymptotic properties discussed earlier.

Exercise 3

Consider observations x_1, \dots, x_N , where x_i is a p -dimensional vector of the form $x_i = \mathbf{s} + w_i$ where \mathbf{s} is an unknown signal and w_i are independent realizations of white Gaussian noise:

$$w_i \sim [\text{U+EF3B}] (\mathbf{0}, \sigma^2 I_{p \times p})$$

Find the maximum likelihood estimate of the energy $E = \mathbf{s}^T \mathbf{s}$ of the unknown signal.

7 Summary of MLE

The likelihood principle states that information brought by an observation \mathbf{x} about θ is entirely contained in the likelihood function $p(\mathbf{x}|\theta)$. The maximum likelihood estimator is **one** effective implementation of the likelihood principle. In some cases, the MLE can be computed exactly, using calculus and linear algebra, but at other times iterative numerical algorithms are needed. The MLE has several desirable properties:

- It is consistent and asymptotically efficient (as $N \rightarrow \infty$ we are doing as well as MVUE).
- When an efficient estimator exists, it is the MLE.
- The MLE is invariant to reparameterization.

References

- [1] Casella and Berger. *Statistical Inference*. Duxbury Press, Belmont, CA, 1990.
- [2] Steven Kay. *Fundamentals of Statistical Signal Processing Volume I: Estimation Theory*. Prentice Hall, 1993.