

Maximum Likelihood

Nuno Carvalho

July 10, 2025

A lot of this text is adapted from the textbook [Lynch & Walsh, 1998], particularly Appendix 4, as well as some back-and-forth with ChatGPT to help consolidate ideas. The point of this is to re-explain (restricted) maximum likelihood in a way that makes more sense to me, and hopefully to you. Here, I focus more on the theory and abstraction of ML/REML, and in other entries, I plan to apply it to a particular class of problem (e.g. variance component estimation).

1 Maximum Likelihood (ML)

1.1 Likelihood

When we construct a model of some real world phenomenon, we define some process that describes how data, \mathbf{y} , is generated according to some set of parameters, $\boldsymbol{\Theta}$. Typically, our model is not deterministic since we do not understand every single cause of the real world phenomenon with perfect accuracy. Instead, our data is generated probabilistically, which means given some set of parameter values, different possible data values can be generated with differing probabilities. We can also think of the reverse: Given our observed data, there are many different combinations of parameter values that could have generated that data, each with differing likelihoods.

We can define a **likelihood function** that tells us the probability density of some data \mathbf{y} being generated by some set of parameter values $\boldsymbol{\Theta}$:

$$\ell(\boldsymbol{\Theta}|\mathbf{y})$$

Note that because we actually do know what the data is (it is observed), we are asking the question: "How well does an input set of parameter values, $\boldsymbol{\Theta}$, explain the observed data, \mathbf{y} . This does not necessarily tell us the **probability** of the parameter values given our data, since we are not computing the likelihood across all possible parameters (weighted by some prior), as we would in Bayesian statistics.

Because our data is (usually) composed of multiple data points, and our model often describes the probability of a single data point being generated, our likelihood for the

entire data will often be the product of multiple individual likelihoods:

$$\ell(\boldsymbol{\Theta}|\mathbf{y}) = \prod_i^N \ell(\boldsymbol{\Theta}|y_i)$$

Note that this makes the crucial assumption that each data point is independent of each other, such that the likelihood of one data point is not affected by the values of the other datapoints. Because sums are typically easier to work with than products, it's useful to instead compute the natural logarithm of the likelihood, which per the properties of logarithms, turns those products into sums. This defines the **log-likelihood**:

$$L(\boldsymbol{\Theta}|\mathbf{y}) = \sum_i^N \ln(\ell(\boldsymbol{\Theta}|y_i))$$

Because logarithms are a monotonic function, it means that whatever parameter values maximize the likelihood also maximize the log-likelihood, and same for the minimum.

1.1.1 Example of the likelihood of a function

(Adapted from Example 1 of Appendix 4 of [Lynch & Walsh, 1998]). Let's say we have N data points: $\mathbf{y} = [y_1, y_2, \dots, y_N]$, which we believe were generated from some normal distribution with a variance of $\sigma^2 = 1$ but an unknown mean. We know the probability density function (PDF) for a normal distribution, which is a function of the mean μ and variance σ .

$$\Pr(y_i, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_i - \mu}{\sigma}\right)^2\right]$$

In our case, we know $\sigma = 1$ and we also observe the data \mathbf{y} . What we don't know is the mean, μ , which is the only non-fixed parameter. We can plug in our known parameter σ , compute the product of the PDF for all our data points, and take the logarithm. This results in a log-likelihood function that describes how well an inputted μ value explains our observed data:

$$L(\mu|\mathbf{y}) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_i^N (y_i - \mu)^2$$

1.2 Maximum Likelihood Estimation

Once we have described the likelihood of some set of parameter values given our observed data, we can then ask the question: What set of parameter values maximizes the likelihood of our data? The rationale is that if we want to estimate these unknown parameters, our best bet for a single point estimate is the combination of parameter values that are most likely to produce the observed data. The set of parameter values $\hat{\boldsymbol{\Theta}}$ that maximizes the likelihood of our observed data is called the **maximum likelihood estimate (MLE)**.

From calculus, we know that the maximum of a multivariable function can be identified by taking the partial derivatives of the function with respect to each parameter and then

solving for the parameter values that make them equal to zero. We define this first partial derivative of a log-likelihood function to be the **Score** (S) of the likelihood function:

$$S(\boldsymbol{\Theta}) = \frac{\partial L(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}}$$

Note that if $\boldsymbol{\Theta}$ is a vector of M parameters, then $S(\boldsymbol{\Theta})$ is also a vector of length M . The MLE, $\hat{\boldsymbol{\Theta}}$, is therefore the solution to the following:

$$S(\hat{\boldsymbol{\Theta}}) = 0$$

1.2.1 Example of the MLE of a function

Continuing the previous example, we can compute the score of the likelihood function, which in this case is just a single derivative:

$$\begin{aligned} S(\mu) &= \frac{\partial L(\mu)}{\partial \mu} \\ &= \frac{\partial}{\partial \mu} \left[-\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_i^N (y_i - \mu)^2 \right] \\ &= \frac{\partial}{\partial \mu} \left[-\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_i^N (y_i^2 - 2y_i\mu + \mu^2) \right] \\ &= -\frac{1}{2} \sum_i^N (-2y_i + 2\mu) \\ &= N(\bar{y} - \mu) \end{aligned}$$

where $\bar{y} = \frac{1}{N} \sum_i^N (y_i)$ (i.e. the sample mean). We now set $S(\hat{\mu}) = 0$ and solve:

$$\begin{aligned} 0 &= S(\hat{\mu}) \\ &= N(\bar{y} - \hat{\mu}) \end{aligned}$$

The solution is: $\hat{\mu} = \bar{y} = \frac{1}{N} \sum_i^N (y_i)$. Matching our intuitions, the mean parameter that maximizes the likelihood of our data is the arithmetic mean of our data sample.

1.2.2 Asymptotic efficiency of the MLE

What's nice about ML estimation is that as the sample size increases, the MLE converges to the true parameter value θ_0 (if the likelihood function matches that of the true data-generating function). This is called **asymptotic efficiency** and can be written formally as:

$$\hat{\theta}_N \rightarrow^p \theta_0$$

This should make intuitive sense, as the true data-generating function has some true distribution of observed data that it produces, which in most cases, is unique to the specific parameter value of that data-generating function. Consequently, the parameter that maximizes the likelihood of producing the true parameter's distribution of data is the true parameter itself. The more samples we collect, the more our observed data's distribution should match the true distribution of the possible data. Therefore, the MLE should get closer to the true parameter value.

1.3 The Curvature of the Likelihood

While the first derivative describes the rate of change of the likelihood function, the second derivative describes its curvature. At the MLE, a strong curvature (high magnitude) means the likelihood peaks very sharply at the MLE. Therefore, a small change in the parameter value(s) from the MLE leads to a sharp drop in the likelihood. In contrast, if there is weak curvature (low magnitude) at the MLE, then there is a wide region of parameter estimates that yield similar (albeit slightly lower) likelihoods of producing the data.

1.3.1 The Hessian

For M parameters, the second derivative of the log-likelihood function is the **Hessian** (\mathbf{H}), an $M \times M$ matrix where:

$$H_{ij} = \frac{\partial^2 L(\Theta|\mathbf{y})}{\partial \Theta_i \partial \Theta_j}$$

This is also called the gradient and can be evaluated at any set of parameter values, describing the multi-dimensional curvature of the log-likelihood at that set of parameters.

1.3.2 The Fisher information Matrix

Of mathematical interest to us is the negative expected value of the Hessian, which is called the **Fisher information matrix**:

$$\mathcal{I}(\Theta) = -E[\mathbf{H}(\Theta)]$$

Note that the expected value here is taken by averaging across all the data that is potentially generated by the inputted set of parameters, weighted by the probability of generating that data. This means that while the Hessian is a function that depends on both the parameters and the data that is observed, the Fisher information matrix only depends on the parameters. That is why $-\mathbf{H}(\Theta)$ is often called the **observed information matrix**, while $\mathcal{I}(\Theta)$ is the **expected information matrix**.

As sample size increases, the Fisher information matrix converges to the negative of the Hessian evaluated at the MLE:

$$\mathcal{I}(\hat{\Theta}) \approx \mathbf{H}(\hat{\Theta})$$

1.3.3 Asymptotic normality of the MLE

If our likelihood function follows [some regularity conditions](#), as sample size increases, the sampling distribution of the MLE (that is, the distribution of the MLE for different samples of observed data from the same underlying data-generating process) converges to normality. This is called **asymptotic normality** and can be written formally as:

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \rightarrow^d \mathcal{N}(0, \sigma^2)$$

It can be [proven](#) that the variance (σ^2 in the above equation) of the MLE converges to the inverse of the Fisher information at the true parameter value:

$$\hat{\theta}_N \rightarrow^d \mathcal{N}(\theta_0, \mathcal{I}(\theta_0)^{-1})$$

This also applies when there are multiple parameters, such that the inverse of the Fisher information matrix evaluated at the MLE is the covariance matrix of the MLE. The i^{th} diagonal element is the variance of the MLE for the i^{th} parameter. The element (i,j) is the covariance of the MLE for the i^{th} and j^{th} parameters.