

Maximum Likelihood Estimation

Guy Lebanon

May 13, 2006

Maximum likelihood estimation is the most popular general purpose method for obtaining point estimators. It was proposed by Fisher about 100 years ago and has been widely used ever since with a relatively large degree of success.

Definition 1. Let X_1, \dots, X_n be sampled from a distribution with a parameter θ . The maximum likelihood estimator (MLE) $\hat{\theta}_{MLE}$ is the θ that maximizes the likelihood function $L(\theta) = L(X_1, \dots, X_n|\theta)$ (if it exists).

Above, we suppress the dependency of L on X_1, \dots, X_n to emphasize that we are treating the likelihood as a function of θ . Note that the above definition applies to both the one-dimensional θ and vector θ cases. Under some general conditions, MLE can be shown to be consistent (as well as a stronger optimality condition called asymptotic efficiency). Before describing a few examples, we discuss some tools that facilitate the computation of the MLE.

1. Note that monotonic increasing functions g preserve order in the sense that $L(X_1, \dots, X_n|\theta_1) \geq L(X_1, \dots, X_n|\theta_2) \Leftrightarrow g(L(X_1, \dots, X_n|\theta_1)) \geq g(L(X_1, \dots, X_n|\theta_2))$. As a consequence, we can find the MLE by obtaining the maximizer of $g(L(X_1, \dots, X_n|\theta))$ rather than the likelihood itself. A common choice for g is the logarithm function which is particularly helpful since it transforms the multiplicative likelihood into a sum (sums are easier to differentiate than products). A common notation for the log of the likelihood is $\ell(\theta)$.
2. If $L(\theta)$ is differentiable in θ , we can try to find the MLE by solving the equation $\frac{d\ell(\theta)}{d\theta}|_{\hat{\theta}_{MLE}} = 0$. If θ is a vector we solve the system of equations $\frac{\partial \ell(\theta)}{\partial \theta_j}|_{\hat{\theta}_{MLE}} = 0, j = 1, \dots, k$. The obtained solutions are necessarily critical points (maximum, minimum or inflection) of the log-likelihood. To actually prove that the solution is a maximum we need to show (in the single-parameter case) that $\frac{d^2 \ell(\theta)}{d\theta^2}|_{\hat{\theta}_{MLE}} > 0$ or (if θ is a vector) that the Hessian matrix H , defined by $[H]_{ij} = \frac{\partial^2 \ell(\theta)}{\partial \theta_i \partial \theta_j}|_{\hat{\theta}_{MLE}}$, is positive definite¹.
3. Any additive and multiplicative constants in θ (terms that are not functions of θ) to $L(\theta)$ or $\ell(\theta)$ may be ignored. These constants change the value of the likelihood, but not where the maximizer is.
4. The MLE is invariant in the sense that for all 1-1 functions h we have $\widehat{h(\theta)}_{MLE}(X_1, \dots, X_n) = h(\hat{\theta}_{MLE}(X_1, \dots, X_n))$. The key property is that 1-1 functions have an inverse. If we use the parametrization $h(\theta)$ rather than θ , the likelihood function that should be used is $L \circ h^{-1}$ rather than L . Since

$$L \circ h^{-1}(h(\hat{\theta}_{MLE})) = L(\hat{\theta}_{MLE}) \geq L(\theta) = L(h^{-1}(h(\theta))) = L \circ h^{-1}(h(\theta))$$

the parameter $h(\hat{\theta}_{MLE})$ maximizes the re-parameterized likelihood $L \circ h^{-1}$.

Example: Let $X_1, \dots, X_n \sim \text{Ber}(\theta)$. The likelihood is $L(\theta) = \prod \theta^{x_i} (1 - \theta)^{1 - x_i} = \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i}$ and the log-likelihood is $\ell(\theta) = (\sum x_i) \log \theta + (n - \sum x_i) \log(1 - \theta)$. Setting the loglikelihood derivative to zero yields $0 = \sum x_i / \theta - (n - \sum x_i) / (1 - \theta)$ or $0 = (1 - \theta) \sum x_i - (n - \sum x_i) \theta = \sum x_i - n\theta \Rightarrow \hat{\theta}_{MLE} = \frac{1}{n} \sum X_i$.

¹A positive definite matrix H is one that satisfies $v^\top H v > 0$ for all vectors v .

Example: Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$. To find the MLE for $\theta = (\mu, \sigma)$ we need to set the two partial derivatives of the log-likelihood to zero. The log-likelihood is

$$\ell(\theta) = \log \frac{1}{(2\pi\sigma^2)^{n/2}} \prod e^{-(x_i - \mu)^2 / (2\sigma^2)} = c - \frac{n}{2} \log \sigma^2 + \log e^{-\sum (x_i - \mu)^2 / (2\sigma^2)} = c - \frac{n}{2} \log \sigma^2 - \sum \frac{(x_i - \mu)^2}{2\sigma^2}$$

where c is an inconsequential additive constant. Setting the partial derivative with respect to μ to zero gives

$$\frac{\partial \ell(\theta)}{\partial \mu} = \sum \frac{x_i - \mu}{\sigma^2} = 0 \Rightarrow \hat{\mu}_{MLE} = \bar{x}.$$

Substituting this in the equation resulting from setting the partial derivative with respect to σ^2 to 0

$$\frac{\partial \ell(\theta)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \sum \frac{(x_i - \mu)^2}{2\sigma^4} = 0 \Rightarrow -\frac{n}{2\sigma^2} + \sum \frac{(x_i - \bar{x})^2}{2\sigma^4} = 0 \Rightarrow \sigma^2 n + \sum (x_i - \bar{x})^2 = 0 \Rightarrow \widehat{\sigma^2}_{MLE} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

By property 4 above, the MLE for the standard deviation is $\hat{\sigma}_{MLE} = \sqrt{\widehat{\sigma^2}_{MLE}} = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}$.

Example: $X_1, \dots, X_n \sim U[0, \theta]$. In this case the likelihood is $L(\theta) = \frac{1}{\theta^n}$ if $0 \leq x_1, \dots, x_n \leq \theta$ and 0 otherwise. We need to exercise care in this situation since the definition of the likelihood branches in two options depending on the value of the parameter θ . To treat only one case and not two we write the likelihood as $L(\theta) = \theta^{-n} 1_{\{0 \leq x_1, \dots, x_n \leq \theta\}}$ where $1_{\{A\}}$ is the indicator function which equals 1 if A is true and 0 otherwise. We can't at this point proceed as before since the likelihood is not a differentiable function of θ (and neither will the log-likelihood be). We therefore do not take derivatives and simply examine the function $L(\theta) = \theta^{-n} 1_{\{0 \leq x_1, \dots, x_n \leq \theta\}}$. Recall that we consider the likelihood as a function of θ for fixed x_1, \dots, x_n . In this perspective, the likelihood will be zero for $\theta < \max(x_1, \dots, x_n)$ and for $\theta \geq \max(x_1, \dots, x_n)$ the likelihood function will be non-zero but start decreasing as we increase θ . It follows then that $\hat{\theta}_{MLE} = \max(X_1, \dots, X_n)$.

Example: $X_1, \dots, X_n \sim U(0, \theta)$ (as before but this time the interval is open and not closed). We start as before, but the likelihood at $\max(x_1, \dots, x_n)$ is zero. Since the likelihood increases as we get θ closer to $\max(x_1, \dots, x_n)$ (from the right), and at $\max(x_1, \dots, x_n)$ it is zero - there is no MLE. For any specific value $\hat{\theta}$, we can always come up with $\hat{\theta}'$ that will result in a higher likelihood. Thus there is no value of θ that maximizes the likelihood.

Example: $X_1, \dots, X_n \sim U[\theta, \theta + 1]$. In this case the likelihood is $L(\theta) = 1_{\{\theta \leq x_1, \dots, x_n \leq \theta + 1\}}$. The likelihood is thus either zero or 1. Since it is 1 for many possible values, there are multiple maximizers or MLEs (all $\hat{\theta}_{MLE}$ for which $\hat{\theta}_{MLE} \leq x_1, \dots, x_n \leq \hat{\theta}_{MLE} + 1$) rather than a unique one.

We saw examples where there is a single MLE, there is no MLE or there are multiple MLEs. In general, one has to be careful when using the differentiation method since sometimes the loglikelihood is not differentiable (due perhaps to two branches of definitions). Another potential difficulty with the differentiation method is that it only discovers critical points, and not necessarily maximum. Moreover, it only discovers local maxima and not the global maximum necessarily (which is the MLE).

Another interesting fact is that since we can ignore multiplicative constants (in θ) of the likelihood, the MLE will always be a function of the sufficient statistics. This follows from the factorization theorem $L(\theta) = g(T(X_1, \dots, X_n), \theta)h(X_1, \dots, X_n)$ since $h(X_1, \dots, X_n)$ is constant in θ .