Maximum Likelihood Estimation

Assume $\boldsymbol{X} \sim P_\theta, \theta \in \Theta$, with joint pdf (or pmf) $f(\boldsymbol{x} \,|\, \theta)$.

Suppose we observe $\boldsymbol{X} = \boldsymbol{x}$.

The **Likelihood function** is

$$L(\theta \,|\, \boldsymbol{x}) = f(\boldsymbol{x} \,|\, \theta)$$

as a function of θ (with the data $\boldsymbol{x}$ held fixed).

> The likelihood function $L(\theta \,|\, \boldsymbol{x})$ and joint pdf $f(\boldsymbol{x} \,|\, \theta)$ are the same except that $f(\boldsymbol{x} \,|\, \theta)$ is generally viewed as a function of $\boldsymbol{x}$ with θ held fixed, and $L(\theta \,|\, \boldsymbol{x})$ as a function of θ with $\boldsymbol{x}$ held fixed.
>
> $f(\boldsymbol{x} \,|\, \theta)$ is a density in $\boldsymbol{x}$ for each fixed θ.
>
> But $L(\theta \,|\, \boldsymbol{x})$ is **not** a density (or mass function) in θ for fixed $\boldsymbol{x}$ (except by coincidence).

The Maximum Likelihood Estimator (MLE)

A point estimator $\hat{\theta} = \hat{\theta}(\boldsymbol{x})$ is a MLE for θ if

$$L(\hat{\theta} \,|\, \boldsymbol{x}) = \sup_\theta L(\theta \,|\, \boldsymbol{x}),$$

that is, $\hat{\theta}$ maximizes the likelihood.

In most cases, the maximum is achieved at a unique value, and we can refer to "the" MLE, and write

$$\hat{\theta}(\boldsymbol{x}) = \operatorname*{argmax}_\theta L(\theta \,|\, \boldsymbol{x}).$$

(But there are cases where the likelihood has flat spots and the MLE is not unique.)

Motivation for MLE's

Note: We often write $L(\theta \,|\, \boldsymbol{x}) = L(\theta)$, suppressing $\boldsymbol{x}$, which is kept fixed at the observed data.

Suppose $\boldsymbol{x} \in \mathbb{R}^n$.

Discrete Case:

If $f(\cdot \,|\, \theta)$ is a mass function ($\boldsymbol{X}$ is discrete), then

$$L(\theta) = f(\boldsymbol{x} \,|\, \theta) = P_\theta(\boldsymbol{X} = \boldsymbol{x})\,.$$

$L(\theta)$ is the probability of getting the observed data $\boldsymbol{x}$ when the parameter value is θ.

Continuous Case:

When $f(\cdot \,|\, \theta)$ is a continuous density $P_\theta(\boldsymbol{X} = \boldsymbol{x}) = 0$, but if $B \subset \mathbb{R}^n$ is a very, very small ball (or cube) centered at the observed data $\boldsymbol{x}$, then

$$P_\theta(\boldsymbol{X} \in B) \approx f(\boldsymbol{x} \,|\, \theta) \times \text{Volume}(B) \propto L(\theta)\,.$$

$L(\theta)$ is proportional to the probability the random data $\boldsymbol{X}$ will be **close** to the observed data $\boldsymbol{x}$ when the parameter value is θ.

Thus, the MLE $\widehat{\theta}$ is the value of θ which makes the observed data $\boldsymbol{x}$ "most probable".

To find $\widehat{\theta}$, we maximize $L(\theta)$. This is usually done by calculus (finding a stationary point), but **not** always.

If the parameter space Θ contains endpoints or boundary points, the maximum can be achieved at a boundary point without being a stationary point.

If $L(\theta)$ is not "smooth" (continuous and everywhere differentiable), the maximum does **not** have to be achieved at a stationary point.

Cautionary Example:

Suppose $X_1, \ldots, X_n$ are iid Uniform$(0, \theta)$ and $\Theta = (0, \infty)$.

Given data $\boldsymbol{x} = (x_1, \ldots, x_n)$, find the MLE for θ.

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \theta^{-1} I(0 \le x_i \le \theta) = \theta^{-n} I(0 \le \min x_i) I(\max x_i \le \theta) \\ &= \begin{cases} \theta^{-n} & \text{for } \theta \ge \max x_i \\ 0 & \text{for } 0 < \theta < \max x_i \end{cases} \qquad \text{(Draw this!)} \end{aligned}$$

which is maximized at $\theta = \max x_i$, which is a point of discontinuity (and certainly **not** a stationary point).

Thus, the MLE is $\widehat{\theta} = \max x_i = x_{(n)}$.

Notes:

$L(\theta) = 0$ for $\theta < \max x_i$ is just saying that these values of θ are absolutely ruled out by the data (which is obvious).

A strange property of the MLE in this example (not typical):

$$P_\theta(\widehat{\theta} < \theta) = 1$$

The MLE is biased; it is always less than the true value.

A Similar Example:

Let $X_1, \ldots, X_n$ be iid Uniform(α, β) and $\Theta = \{(\alpha, \beta) : \alpha < \beta\}$.

Given data $\boldsymbol{x} = (x_1, \ldots, x_n)$, find the MLE for $\theta = (\alpha, \beta)$.

$$\begin{aligned} L(\alpha, \beta) &= \prod_{i=1}^{n} (\beta - \alpha)^{-1} I(\alpha \le x_i \le \beta) \\ &= (\beta - \alpha)^{-n} I(\alpha \le \min x_i) I(\max x_i \le \beta) \\ &= \begin{cases} (\beta - \alpha)^{-n} & \text{for } \alpha \le \min x_i \text{, } \max x_i \le \beta \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

which is maximized by making $\beta - \alpha$ as small as possible without entering "0 otherwise" region.

Clearly, the maximum is achieved at $(\alpha, \beta) = (\min x_i, \max x_i)$. Thus the MLE is $\widehat{\theta} = (\widehat{\alpha}, \widehat{\beta}) = (\min x_i, \max x_i)$.

Again, $P_{\alpha,\beta}(\alpha < \widehat{\alpha}\,,\ \widehat{\beta} < \beta) = 1$.

Maximizing the Likelihood (one parameter)

Basic Result: A <u>continuous</u> function $g(\theta)$ defined on a <u>closed</u>, <u>bounded</u> interval J attains its supremum (but might do so at one of the endpoints).

(That is, there exists a point $\theta_0 \in J$ such that $g(\theta_0) = \sup_{\theta \in J} g(\theta)$.)

Consequence: Suppose $g(\theta)$ is a continuous, non-negative function defined on an open interval $J = (c, d)$ (where perhaps $c = -\infty$ or $d = +\infty$). If

$$\lim_{\theta \to c} g(\theta) = \lim_{\theta \to d} g(\theta) = 0\,,$$

then g attains its supremum.

• Thus, MLE's usually exist when the likelihood function is continuous.

Suppose the function $g(\theta)$ is defined on an interval Θ (which may be open or closed, infinite or finite).

If g is differentiable and attains its supremum at a point θ_0 in the interior of Θ, that point must be a stationary point (that is, $g'(\theta_0) = 0$).

(1) If $g'(\theta_0) = 0$ and $g''(\theta_0) < 0$, then θ_0 is a local maximum (but might not be the global maximum).

(2) If $g'(\theta_0) = 0$ and $g''(\theta) < 0$ for **all** $\theta \in \Theta$, then θ_0 is a global maximum (that is, it attains the supremum).

(1) is necessary (but not sufficient) for θ_0 to be a global maximum. (2) is sufficient (but not necessary).

A function satisfying $g''(\theta) < 0$ for all $\theta \in \Theta$ is called **strictly concave**. It lies below any tangent line.

Maximizing the Likelihood (multi-parameter)

Basic Result: A continuous function $g(\theta)$ defined on a closed, bounded set $J \subset R^k$ attains its supremum (but might do so on the boundary).

Consequence: Suppose $g(\theta)$ is a continuous, non-negative function defined for all $\theta \in R^k$. If $g(\theta) \to 0$ as $\theta \to \infty$, then g attains its supremum.

• Thus, MLE's usually exist when the likelihood function is continuous.

Suppose the function $g(\theta)$ is defined on a **convex** set $\Theta \subset R^k$ (that is, the line segment joining any two points in Θ lies entirely inside Θ).

If g is differentiable and attains its supremum at a point θ_0 in the interior of Θ, that point must be a stationary point:

$$\frac{\partial g(\theta_0)}{\partial \theta_i} = 0 \quad \text{for } i = 1, 2, \ldots, k.$$

Define the gradient vector D and Hessian matrix H:

$$D(\theta) = \left(\frac{\partial g(\theta)}{\partial \theta_i} \right)_{i=1}^{k} \quad (\text{a } k \times 1 \text{ vector}).$$

$$H(\theta) = \left(\frac{\partial^2 g(\theta)}{\partial \theta_i \partial \theta_j} \right)_{i,j=1}^{k} \quad (\text{a } k \times k \text{ matrix}).$$

where $\theta = (\theta_1, \theta_2, \ldots, \theta_k)'$.

(1) If $D(\theta_0) = 0$ and $H(\theta_0)$ is **negative definite**, then θ_0 is a local maximum (but might not be the global maximum).

(2) If $D(\theta_0) = 0$ and $H(\theta)$ is negative definite for **all** $\theta \in \Theta$, then θ_0 is a global maximum (that is, it attains the supremum).

(1) is necessary (but not sufficient) for θ_0 to be a global maximum. (2) is sufficient (but not necessary).

A function for which $H(\theta)$ is negative definite for all $\theta \in \Theta$ is called **strictly concave**. It lies below any tangent plane.

Example:

Observe $X_1, \ldots, X_n$ be iid Gamma(α, β).

Preliminaries:

$$\text{(likelihood)} \quad L(\alpha,\beta) = \prod_{i=1}^{n} \frac{x_i^{\alpha-1} e^{-x_i/\beta}}{\beta^\alpha \Gamma(\alpha)}\,.$$

Maximizing L is same as maximizing $\ell = \log L$ given by

$$\begin{aligned} \ell(\alpha,\beta) &= (\alpha-1)T_1 - T_2/\beta - n\alpha\log\beta - n\log\Gamma(\alpha) \\ &\quad \text{where} \quad T_1 = \textstyle\sum_i \log x_i\,, \quad T_2 = \sum_i x_i\,. \end{aligned}$$

Note that $T = (T_1, T_2)$ is the natural sufficient statistic of this 2pef.

$$\begin{aligned} \frac{\partial \ell}{\partial \alpha} &= T_1 - n\log\beta - n\psi(\alpha) \\ &\quad \text{where} \quad \psi(\alpha) \equiv \frac{d}{d\alpha}\log\Gamma(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \\ \frac{\partial \ell}{\partial \beta} &= \frac{T_2}{\beta^2} - \frac{n\alpha}{\beta} = \frac{1}{\beta^2}(T_2 - n\alpha\beta) \\ \frac{\partial^2 \ell}{\partial \alpha^2} &= -n\psi'(\alpha) \\ \frac{\partial^2 \ell}{\partial \beta^2} &= \frac{-2T_2}{\beta^3} + \frac{n\alpha}{\beta^2} = \frac{-1}{\beta^3}(2T_2 - n\alpha\beta) \\ \frac{\partial^2 \ell}{\partial \alpha \partial \beta} &= \frac{-n}{\beta} \end{aligned}$$

Situation #1: Suppose $\alpha = \alpha_0$ is known. Find MLE for β.

(Drop α from arguments: $\ell(\beta) = \ell(\alpha_0, \beta)$ etc.)

$\ell(\beta)$ is continuous and differentiable.

$\ell(\beta)$ has a unique stationary point:

$$\begin{aligned} \ell'(\beta) &= \frac{\partial \ell}{\partial \beta} = \frac{1}{\beta^2}\left(T_2 - n\alpha_0\beta\right) = 0 \\ &\text{iff} \qquad T_2 = n\alpha_0\beta \quad \text{iff} \quad \beta = \frac{T_2}{n\alpha_0}\,(\equiv \beta^*)\,. \end{aligned}$$

Now we check the second derivative.

$$\ell''(\beta) = \frac{\partial^2 \ell}{\partial \beta^2} = \frac{-1}{\beta^3}\left(2T_2 - n\alpha\beta\right) = \frac{-1}{\beta^3}\left(T_2 + (T_2 - n\alpha\beta)\right).$$

Note $\ell''(\beta^*) < 0$ since $T_2 - n\alpha_0\beta^* = 0$, but $\ell''(\beta) > 0$ for $\beta > 2T_2/(n\alpha_0)$.

Thus, the stationary point satisfies the necessary condition for a global maximum, but **not** the sufficient condition (i.e., $\ell(\beta)$ is **not** a strictly concave function).

How can we be sure that we have found the global maximum, and not just a local maximum?

In this case, there is a simple argument: The stationary point β^* is unique, and $\ell'(\beta) > 0$ for $\beta < \beta^*$, and $\ell'(\beta) < 0$ for $\beta > \beta^*$. This ensures β^* is the unique global maximizer.

Conclusion: $\widehat{\beta} = \dfrac{T_2}{n\alpha_0}$ is the MLE.

(This is a function of T_2, which is a sufficient statistic for β when α is known.)

Situation #2: Suppose $\beta = \beta_0$ is known. Find MLE for α.

(Drop β from arguments: $\ell(\alpha) = \ell(\alpha, \beta_0)$ etc.)

Note: $\ell'(\alpha)$ and $\ell''(\alpha)$ involve $\psi(\alpha)$ The function ψ is infinitely differentiable on the interval $(0, \infty)$, and satisfies $\psi'(\alpha) > 0$ and $\psi''(\alpha) < 0$ for all $\alpha > 0$. (The function is strictly increasing and strictly concave.) Also

$$\lim_{\alpha \to 0^+} \psi(\alpha) = -\infty \quad \text{and} \quad \lim_{\alpha \to \infty} \psi(\alpha) = \infty. \quad \text{(Draw a Picture.)}$$

Thus $\psi^{-1} : \mathbb{R} \to (0, \infty)$ exists.

$\ell(\alpha)$ is continuous and differentiable.

$\ell(\alpha)$ has a unique stationary point:

$$\begin{aligned} \ell'(\alpha) &= T_1 - n \log \beta_0 - n\psi(\alpha) = 0 \\ &\text{iff} \quad \psi(\alpha) = T_1/n - \log \beta_0 \\ &\text{iff} \quad \alpha = \psi^{-1}(T_1/n - \log \beta_0) \end{aligned}$$

This is the unique global maximizer since

$$\ell''(\alpha) = -n\psi'(\alpha) < 0 \quad \text{for all } \alpha > 0.$$

Thus $\hat{\alpha} = \psi^{-1}(T_1/n - \log \beta_0)$ is the MLE.
(This is a function of T_1, which is a sufficient statistic for α when β is known.)

Situation #3: Find MLE for $\theta = (\alpha, \beta)$

$\ell(\alpha, \beta)$ is continuous and differentiable.

A stationary point must satisfy the system of two equations:

$$\begin{aligned} \frac{\partial \ell}{\partial \alpha} &= T_1 - n \log \beta - n\psi(\alpha) = 0 \\ \frac{\partial \ell}{\partial \beta} &= \frac{1}{\beta^2}(T_2 - n\alpha\beta) = 0 \end{aligned}$$

Solving the second equation for β gives

$$\beta = \frac{T_2}{n\alpha}.$$

Plugging this into the first equation, and rearranging a bit leads to

$$\frac{T_1}{n} - \log\left(\frac{T_2}{n}\right) = \psi(\alpha) - \log \alpha \equiv H(\alpha)$$

The function $H(\alpha)$ is continuous and strictly increasing from $(0, \infty)$ to $(-\infty, 0)$, so that it has an inverse mapping $(-\infty, 0)$ to $(0, \infty)$.

Thus, the solution to the above equation can be written:

$$\alpha = H^{-1}\left(\frac{T_1}{n} - \log\left(\frac{T_2}{n}\right)\right).$$

Thus, the unique stationary point is:

$$\begin{aligned} \widehat{\alpha} &= H^{-1}\left(\frac{T_1}{n} - \log\left(\frac{T_2}{n}\right)\right) \\ \widehat{\beta} &= \frac{T_2}{n\widehat{\alpha}}. \end{aligned}$$

Is this the MLE?

Let us examine the Hessian.