MAXIMUM LIKELIHOOD ESTIMATION

Consider the **usual estimation setup** where we observe

$$\underline{X} \sim f_{\underline{\theta}}(\underline{x}) = f(\underline{x}; \underline{\theta}).$$

Viewing $\underline{x}$ as fixed and $\underline{\theta}$ as the variable, we call

$$l(\underline{\theta}; \underline{x}) := f(\underline{x}; \underline{\theta})$$

the *likelihood* of $\underline{\theta}$ (given $\underline{x}$).

Definition The estimator $\hat{\underline{\theta}}$ is called a *maximum likelihood estimator* if $\forall \underline{x}$

$$l(\hat{\underline{\theta}}(\underline{x}); \underline{x}) = \max_{\underline{\theta} \in (H)} l(\underline{\theta}; \underline{x}).$$

Equivalently, $\hat{\underline{\theta}}$ satisfies

$$\hat{\underline{\theta}}(\underline{x}) = \arg\max_{\underline{\theta} \in (H)} l(\underline{\theta}; \underline{x})$$

Intuitively, the MLE selects the value of $\underline{\theta}$ such that, in retrospect, the observed $\underline{x}$ corresponds to the most probable outcome.

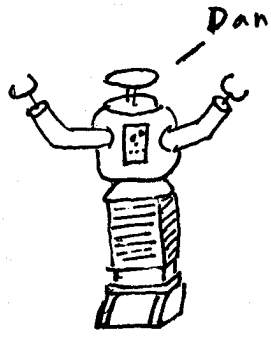
Note: The MLE is not always unique.

Warning

It is tempting to view the likelihood as a density / mass function for $\underline{\theta}$, conditioned on $\underline{X} = \underline{x}$. However, the MLE is a classical estimator that views $\underline{\theta}$ as nonrandom. Furthermore, in some cases

$$\int \ell(\underline{\theta}; \underline{x}) \, d\underline{\theta} = \infty$$

and so the likelihood cannot be normalized.



Likelihood Principle

The information contained in an observation $\underline{x}$ about $\underline{\theta}$ is contained entirely in the likelihood function $\ell(\underline{\theta}; \underline{x})$.

Moreover, if $\underline{x}_1$ and $\underline{x}_2$ are two observations depending on $\underline{\theta}$ (perhaps through different models) such that

$$\ell(\underline{\theta}; \underline{x}_1) = c \cdot \ell(\underline{\theta}; \underline{x}_2) \qquad \forall \underline{\theta}$$

for some constant c, then $\underline{x}_1$ and $\underline{x}_2$ must lead to the same inference about $\underline{\theta}$.

Example Suppose a public health official conducts a survey to estimate $0 \le \theta \le 1$, the percentage of the population eating pizza at least once per week. As a result of the survey, the official found 9 pizza eaters and 3 non-eaters.

If no additional information is available regarding how the survey was implemented, then there are at least two possible probability models

1. $X_1 \sim \text{Bin}(12, \theta)$

$x_1 = 9$ observed

$$p_1(x_1; \theta) = \binom{12}{x_1} \theta^{x_1} (1-\theta)^{12-x_1}$$

2. $X_2 \sim \text{Neg}(3, 1-\theta)$

$x_2 = 12$ observed

$$p_2(x_2; \theta) = \binom{x_2-1}{3-1} (1-\theta)^3 \theta^{x_2-3}$$

In both cases, the likelihood is proportional to

$$l(\theta; x) \propto \theta^9 (1-\theta)^3$$

If we follow the likelihood principle, both models lead to the same inference about θ.

Sufficiency Principle

The MLE also satisfies the sufficiency principle, which states that if $\underline{T} = \tau(\underline{X})$ is sufficient for $\underline{\theta}$ and $\underline{x}_1$ and $\underline{x}_2$ are such that $\tau(\underline{x}_1) = \tau(\underline{x}_2)$, then $\underline{x}_1$ and $\underline{x}_2$ must lead to the same estimate of $\underline{\theta}$.

To see this, note

$$\begin{aligned}\hat{\underline{\theta}}(\underline{x}) &= \arg\max_{\underline{\theta}} f(\underline{x}; \underline{\theta}) \\ &= \arg\max_{\underline{\theta}} g(\tau(\underline{x}); \underline{\theta}) \cdot h(\underline{x}) \\ &= \arg\max_{\underline{\theta}} g(\tau(\underline{x}); \underline{\theta})\end{aligned}$$

which depends on $\underline{x}$ only through $\underline{t} = \tau(\underline{x})$.

Computing the MLE

Since many of the models we work with have an exponential form, it is often convienient to maximize the log-likelihood

$$\log l(\underline{\theta}; \underline{x})$$

If the likelihood function is differentiable, then $\hat{\underline{\theta}}(\underline{x})$ is a solution of

$$\underbrace{\frac{\partial}{\partial \underline{\theta}}}_{\nabla_{\underline{\theta}}} \log l(\underline{\theta}; \underline{x}) = \underline{0}.$$

We also need to verify that such a solution is in fact a local max and not a local min or a saddle point. This can be accomplished by checking to see that

$$\underbrace{\frac{\partial^2}{\partial \underline{\theta}\, \partial \underline{\theta}^T}}_{\nabla^2_{\underline{\theta}}} \log f(\underline{\theta}; \underline{x})$$

Hessian

is negative semidefinite at $\hat{\underline{\theta}}(\underline{x})$, or by otherwise arguing that $\hat{\underline{\theta}}$ is a local max. If several local maximums exists, the MLE is the one with largest likelihood.

Example Suppose $\underline{x} = [x_1, \ldots, x_N]^T$ where

$$X_i \sim N(\mu, \sigma^2), \quad i = 1, \ldots, N.$$

The log-likelihood of μ is

$$\log l(\mu; \underline{x}) = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2$$

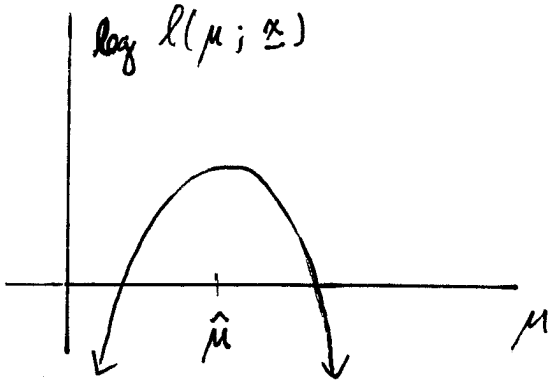
$$\frac{\partial \log l(\mu; \underline{x})}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - \mu)$$

$$= 0$$

$$\Rightarrow \quad \sum (x_i - \mu) = 0$$

$$\Rightarrow \quad \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Note that $\hat{\mu}$ can't be a local min because $\log l(\mu; \underline{x})$ is concave. Therefore the MLE is the sample mean.



Exercise In the previous example, suppose σ^2 is unknown and find the MLE of $\underline{\theta} = [\mu \ \sigma^2]^T$.

Solution

$$\log \ell(\mu, \sigma^2 ; \underline{x}) = -\frac{N}{2} \log (2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2$$

$$\frac{\partial \log \ell(\mu, \sigma^2 ; \underline{x})}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - \mu)$$

$$\frac{\partial \log \ell(\mu, \sigma^2 ; \underline{x})}{\partial (\sigma^2)} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{N} (x_i - \mu)^2 .$$

Therefore, the MLE $[\hat{\mu}, \hat{\sigma^2}]^T$ must solve the system of equations

$$\frac{1}{\hat{\sigma^2}} \sum_{i=1}^{N} (x_i - \hat{\mu}) = 0 \qquad (1)$$

$$\frac{-N}{2\hat{\sigma^2}} + \frac{1}{2(\hat{\sigma^2})^2} \sum_{i=1}^{N} (x_i - \hat{\mu})^2 = 0 . \qquad (2)$$

$$(1) \Rightarrow \quad \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$(2) \Rightarrow \quad \hat{\sigma^2} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2$$

biased

The solution is a maximum because the Hessian is negative semi-definite at $\underline{\hat{\theta}}(\underline{x})$

This is left as an exercise.

The MLE can be computed in closed form for many common distributions, including many members of the exponential family.

Example A coin has $\text{Prob}\{\text{heads}\} = \theta$. To estimate θ, the following experiment is performed N times: The coin is flipped until 10 heads have been observed, and the total number of flips X is recorded. If the values $x_1, \ldots, x_N$ are observed, find the MLE of θ.

X has a negative binomial distribution:

$$p(x;\theta) = \binom{x-1}{9} \theta^{10} (1-\theta)^{x-10}, \quad x \geq 10$$

Assuming independent experiments,

$$\ell(\theta; \underline{x}) = \prod_{i=1}^{N} p(x_i;\theta)$$

$$= \prod_{i=1}^{N} \binom{x_i-1}{9} \theta^{10} (1-\theta)^{x_i-10}$$

$$= \left[\prod_{i=1}^{N} \binom{x_i-1}{9} \right] \theta^{10N} (1-\theta)^{\sum x_i - 10N}$$

$$\log l(\theta; \underline{x}) = 10N \log \theta + (\Sigma x_i - 10N) \log(1-\theta) + C$$

$$\frac{\partial \log l(\theta; \underline{x})}{\partial \theta} = \frac{10N}{\theta} - \frac{\Sigma x_i - 10N}{1-\theta} = 0$$

$$\Rightarrow (1-\theta) 10N = \theta (\Sigma x_i - 10N)$$

$$\Rightarrow \hat{\theta}(\underline{x}) =$$

If you think about it, this makes good sense.

Asymptotic Properties

Theorem Suppose $\underline{X} \sim f(\underline{x}; \underline{\theta})$. Let $\hat{\underline{\theta}}_N$ be the MLE of $\underline{\theta}$ based on n iid realizations $\underline{X}_1, \ldots, \underline{X}_N$ of $\underline{X}$. Under certain regularity conditions,

$$\sqrt{N}(\hat{\underline{\theta}}_N - \underline{\theta}) \xrightarrow{D} N(\underline{0}, I(\underline{\theta})^{-1})$$

where $I(\underline{\theta})$ is the Fisher information matrix evaluated at the true $\underline{\theta}$.

Remarks

- The regularity condition amounts to $f(\underline{x}; \underline{\theta})$ having bounded third derivatives w.r.t. $\underline{\theta}$.
- Proof hinges on central limit theorem

- $E\,\hat{\underline{\theta}}_N \longrightarrow \underline{\theta} \implies$ MLE is asymptotically unbiased
- $\mathrm{Cov}\,\hat{\underline{\theta}}_N \longrightarrow I(\underline{\theta})^{-1} \implies$ MLE is asymptotically efficient
- $\sqrt{N}$ characterizes the rate of convergence.

Example Recall the MLE of $\underline{\theta} = [\mu \;\; \sigma^2]^T$ based on

$$X_i \overset{iid}{\sim} N(\mu, \sigma^2), \qquad i = 1, \ldots, N$$

is $\hat{\underline{\theta}} = [\hat{\mu} \;\; \hat{\sigma}^2]^T$ where

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2$$

Also recall that

$$\frac{1}{\sigma^2} \sum_{i=1}^{N} (X_i - \overline{X})^2 \sim \chi^2_{N-1}$$

and is independent of $\overline{X} = \hat{\mu}$.

This implies

$$\frac{N}{\sigma^2} \cdot \hat{\sigma}^2 \sim \chi^2_{N-1}$$

$$\Rightarrow \quad E\,\hat{\sigma}^2 = \frac{N-1}{N}\sigma^2$$

$$\text{Var}\,\hat{\sigma}^2 = \frac{2(N-1)}{N^2}\sigma^4$$

Therefore, as $N \to \infty$

$$E\,\hat{\underline{\theta}} = \begin{bmatrix} \mu \\ \frac{N-1}{N}\sigma^2 \end{bmatrix} \to \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \underline{\theta}$$

$$\text{Cov}\,\hat{\underline{\theta}} = \begin{bmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2(N-1)}{N^2}\sigma^4 \end{bmatrix} \to \begin{bmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2\sigma^4}{N} \end{bmatrix} = I(\underline{\theta})^{-1}$$

Asymptotic normality can also be verified from the fact

$$\chi^2_r \to \mathcal{N}(r, 2r) \qquad \text{as} \quad r \to \infty$$

which follows from the CLT.

Additional Topics

Nuisance Parameters

Suppose $\underline{X} \sim f(\underline{x}; \underline{\theta})$ and

$\underline{\theta} = \begin{bmatrix} \underline{\theta}_1 \\ \underline{\theta}_2 \end{bmatrix}$, where only $\underline{\theta}_1$ is

of interest. Then the MLE of $\underline{\theta}_1$

is defined to be

$$\tilde{\underline{\theta}}_1(\underline{x}) = \arg\max_{\underline{\theta}_1} \left[\max_{\underline{\theta}_2} \ell(\underline{\theta}_1, \underline{\theta}_2; \underline{x}) \right]$$

Example $X_1, \ldots, X_N \overset{iid}{\sim} N(\mu, \sigma^2)$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

Invariance

Suppose $\underline{\theta} \in \Theta \subseteq \mathbb{R}^p$ and that our objective is to estimate

$$\underline{\varphi} = g(\underline{\theta})$$

If g is invertible, then we may parametrize $f_1(\underline{x}; \underline{\theta})$ in terms of $\underline{\varphi}$:

$$f_2(\underline{x}; \underline{\varphi}) = f_1(\underline{x}; g^{-1}(\underline{\varphi}))$$

We may then estimate $\underline{\varphi}$ via maximum likelihood:

$$\hat{\underline{\varphi}}(\underline{x}) = \arg\max_{\underline{\varphi}} f_2(\underline{x}; \underline{\varphi})$$

Fortunately the MLE is <u>invariant</u> to such transformations:

<u>Theorem</u> Let $\underline{\varphi} = g(\underline{\theta})$ be invertible and let $\hat{\varphi}$ and $\hat{\theta}$ denote the MLEs. Then

$$\hat{\varphi}(\underline{x}) = g(\hat{\underline{\theta}}(\underline{x})).$$

Proof |

$$\hat{\varphi}(\underline{x}) = \arg\max_{\underline{\varphi}} f_2(\underline{x}; \underline{\varphi})$$

$$= \arg\max_{\underline{\varphi}} f_1(\underline{x}; g^{-1}(\underline{\varphi}))$$

$$= g\left(\arg\max_{\underline{\theta}} f_1(\underline{x}; g^{-1}(g(\underline{\theta})))\right)$$

$$= g\left(\arg\max_{\underline{\theta}} f_1(\underline{x}; \underline{\theta})\right)$$

$$= g(\hat{\underline{\theta}}(\underline{x}))$$

Example | Suppose $\underline{X} \sim N(\mu \cdot \underline{1}, \sigma^2 I)$.

Then the MLE of σ is

(a) $\hat{\sigma} =$

If $\underline{\varphi} = g(\underline{\theta})$ and g is many-to-one (i.e. not invertible), then we cannot parametrize the distribution in terms of $\underline{\varphi}$. In this case we define the MLE of $\underline{\varphi}$ to be

$$\hat{\varphi}(\underline{x}) := g(\hat{\underline{\theta}}(\underline{x})).$$

Now the MLE is invariant by definition.

Exercise] Suppose $X_i \overset{iid}{\sim}$ Bernoulli (θ), $i = 1, \ldots, N$.

Find the MLE of the variance of X.

Solution] The MLE of θ is

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

The variance of a Bernoulli trial is

$$\text{Var } X = \theta(1-\theta)$$

Thus the MLE of the variance is

$$\left(\frac{1}{N} \sum_{i=1}^{N} x_i \right) \cdot \left(1 - \frac{1}{N} \sum_{i=1}^{N} x_i \right)$$

MLE of MVG with unknown mean and covariance

Suppose $\underline{X} \sim N(\underline{\mu}, R)$ and a sample $\underline{X}_1, \ldots, \underline{X}_N$ of iid observations is collected.

It can be shown that the MLE of $\underline{\theta} = [\underline{\mu}, R]$ is

$$\hat{\underline{\mu}} = \frac{1}{N} \sum_{i=1}^{N} \underline{x}_i$$

$$\hat{R} = \frac{1}{N} \sum_{i=1}^{N} (\underline{x}_i - \hat{\underline{\mu}})(\underline{x}_i - \hat{\underline{\mu}})^T$$

Note: If $\underline{X}$ is n-dimensional, then the dimension of $\underline{\theta}$ is _____

Proof of this result is more involved than the scalar case ($n=1$). You may find the proof online if you are curious.

Computational Issues

In many cases of interest, the MLE cannot be expressed in closed form. Iterative numerical techniques are then necessary to maximize the likelihood. Examples include

- Newton-Raphson iteration
- The "scoring method" of iteration

$$\underline{\hat{\theta}}_{k+1} = \underline{\hat{\theta}}_k + \left[I(\underline{\theta})^{-1} \frac{\partial \log f(\underline{x};\underline{\theta})}{\partial \underline{\theta}} \right]\Bigg|_{\underline{\theta}=\underline{\theta}_k}$$

- An expectation-maximization (EM) algorithm.

Efficiency and the MLE

In previous examples, we have seen that the MLE is sometimes efficient. There is a precise connection:

Theorem Assume that $f(\underline{x};\underline{\theta}) = \ell(\underline{\theta};\underline{x})$ has ≤ 1 local max. If $\hat{\underline{\theta}}$ is efficient, that is, $E\hat{\underline{\theta}} = \underline{\theta}$ and $\operatorname{Cov}\hat{\underline{\theta}} = I(\underline{\theta})^{-1}$ $\forall \underline{\theta}$, then $\hat{\underline{\theta}}$ is an MLE.

Proof From the CRLB, $\hat{\underline{\theta}}$ is efficient iff

$$\frac{\partial \log f(\underline{x};\underline{\theta})}{\partial \underline{\theta}} = I(\underline{\theta})\left(\hat{\underline{\theta}}(\underline{x}) - \underline{\theta}\right) \quad \forall \underline{x}\ \forall \underline{\theta}.$$

Take $\underline{\theta} = \hat{\underline{\theta}}(\underline{x})$. Then

$$\left.\frac{\partial \log f(\underline{x};\underline{\theta})}{\partial \underline{\theta}}\right|_{\underline{\theta}=\hat{\underline{\theta}}(\underline{x})} = I(\hat{\underline{\theta}}(\underline{x})) \cdot \left(\hat{\underline{\theta}}(\underline{x}) - \hat{\underline{\theta}}(\underline{x})\right)$$

$$= \underline{0}.$$

By the product rule

$$\frac{\partial^2}{\partial \underline{\theta}\, \partial \underline{\theta}^T} \log f(\underline{x}; \underline{\theta}) = -I(\underline{\theta}) + (\hat{\underline{\theta}}(\underline{x}) - \underline{\theta}) \frac{\partial}{\partial \underline{\theta}^T} I(\underline{\theta})$$

Again setting $\underline{\theta} = \hat{\underline{\theta}}(\underline{x})$ we have

$$\frac{\partial^2}{\partial \underline{\theta}\, \partial \underline{\theta}^T} \log f(\underline{x}; \underline{\theta}) \Big|_{\underline{\theta} = \hat{\underline{\theta}}(\underline{x})} = -I(\underline{\theta}) \leq 0$$

$\Longrightarrow$ $\hat{\underline{\theta}}(\underline{x})$ is a local max. Since $\log f(\underline{x}; \underline{\theta})$ has at most one local max, $\hat{\underline{\theta}}$ is the MLE.

Summary

1. MLE is one implementation of the likelihood and sufficiency principles.
2. MLE is asymptotically normal and asymptotically efficient. (under certain conditions)
3. MLE is invariant under reparametrization
4. Efficient estimators are usually MLEs

Key

a. $\sqrt{\hat{\sigma^2}} = \sqrt{\frac{1}{N} \sum (x_i - \bar{x})^2}$, b. $n + \frac{n(n+1)}{2}$