I225E Statistical Signal Processing

9. Maximum Likelihood Estimation

MAWALIM and UNOKI

candylim@jaist.ac.jp and unoki@jaist.ac.jp

School of Information Science

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY,
HOKURIKU 1990

Maximum Likelihood Estimation

What if MVUE (minimum variance unbiased estimator) does not exist or unknown?

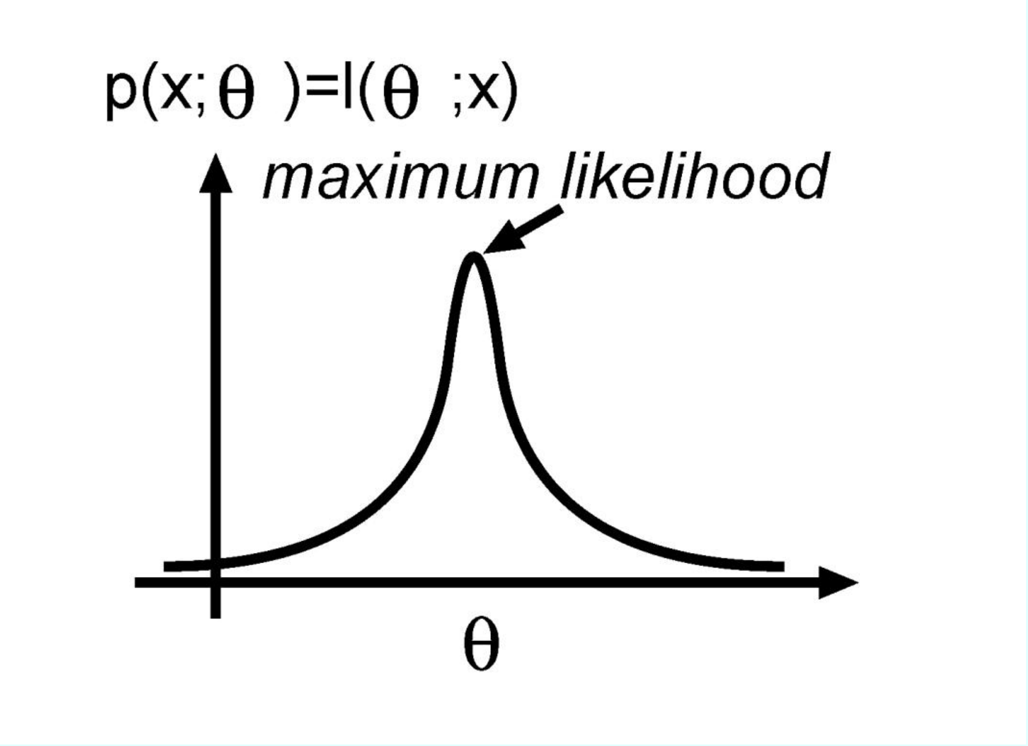
$\Longrightarrow$ **Maximum Likelihood Estimation**

[Features]

1. Easy to implement
2. Optimal for large enough data records
3. Under certain conditions, asymptotically efficient
4. In other words, converges to MVUE

$\Longrightarrow$ Applied to various practical problems.

Random variable $X \sim p(x;\theta)$ is observed. Viewing x as fixed and θ as variable, we call $l(\theta;x) = p(x;\theta)$ as the likelihood of θ (given x).



Maximum Likelihood Estimation

- **Core Idea:** To find the parameter values that make the observed data most probable
- **Steps:**

1. Assume a Model (e.g., based on a probability distribution)
2. Formulate the Likelihood Function ($L(\theta \mid x)$)
3. Maximize the Likelihood
4. Log-likelihood (Often used)
5. Finding the MLE ($\hat{\theta}$)

Definition

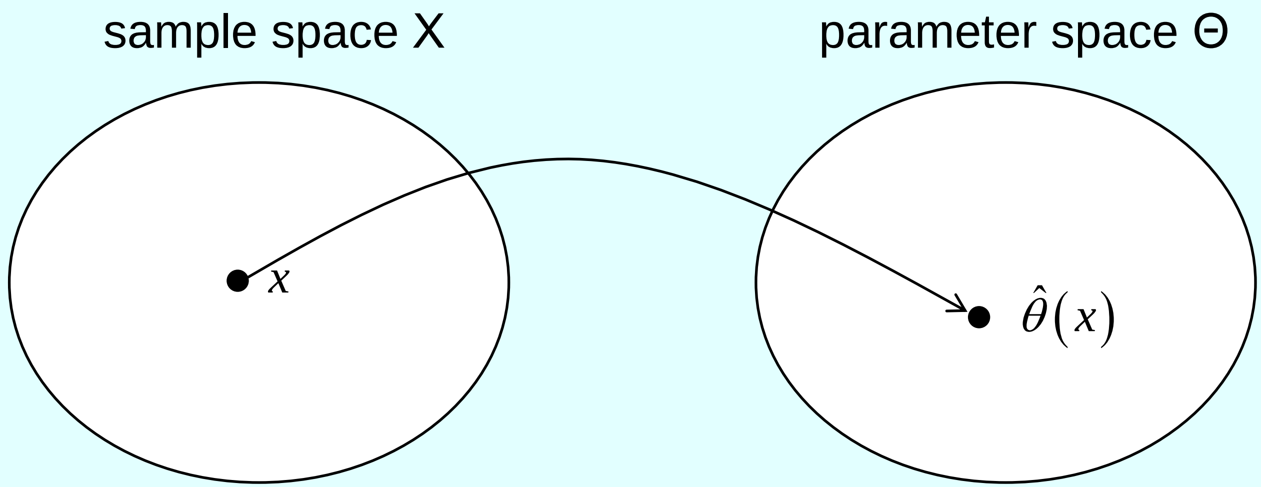
$\hat{\theta}$ is called ***maximum likelihood estimator*** if

$$\forall x, \quad l\left(\hat{\theta}; x\right) = \max_{\theta \in \Theta} l(\theta; x).$$

This is equivalent to $\hat{\theta}(x) = \arg\max_{\theta \in \Theta} l(\theta; x)$

Note:

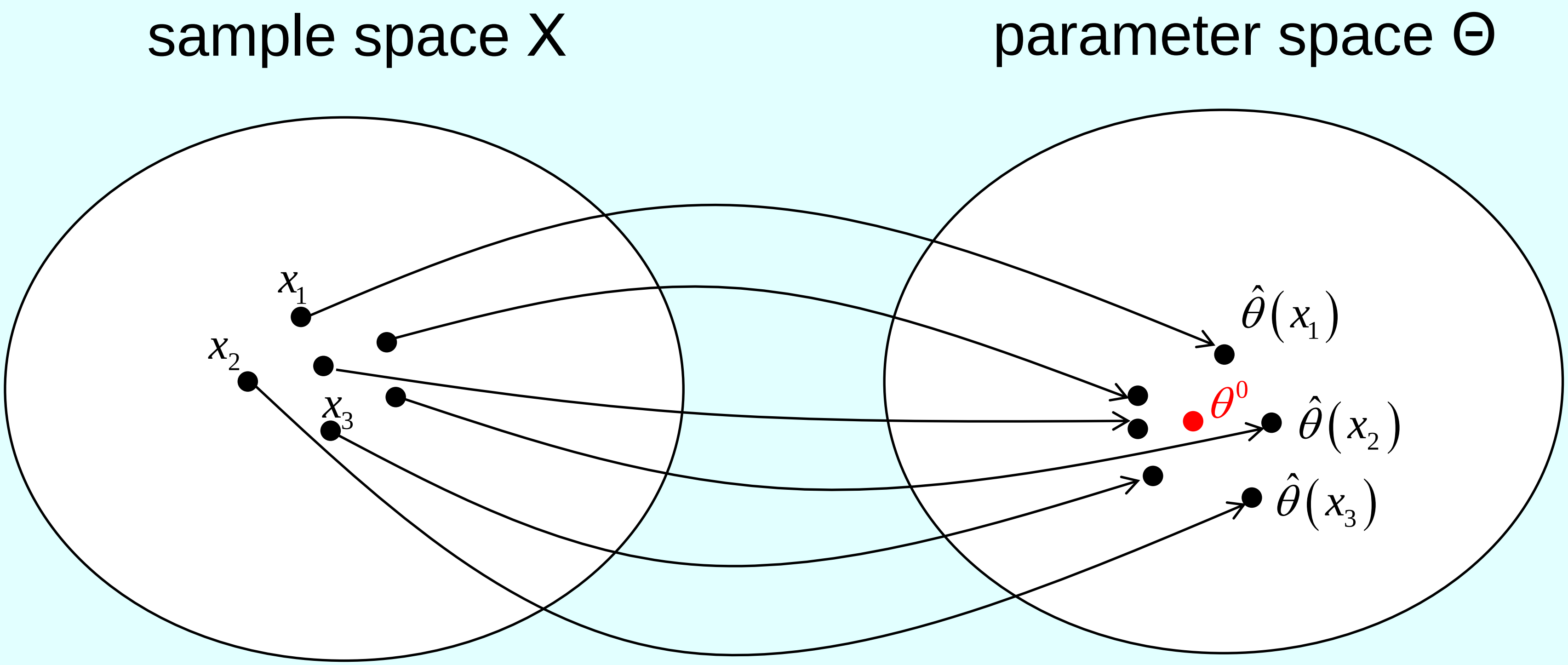
MLE (maximum likelihood estimator) selects the value of θ such that the observed x corresponds to the most probable outcome. Likelihood can be viewed as a density function for θ conditioned on $X = x$. However, classical estimator views θ as nonrandom.



$$\hat{\theta}_{\mathrm{ML}}(x) = \arg\max_{\theta} P(x \mid \theta)$$
$$= \arg\max_{\theta} \log P(x \mid \theta)$$

ML is …

- Asymptotically unbiased (i.e., approaches to a true value).
- Asymptotically efficient (i.e., achieves minimum variance, CRLB).


sample space X
parameter space Θ
x_1
x_2
x_3
$\hat{\theta}(x_1)$
θ^0
$\hat{\theta}(x_2)$
$\hat{\theta}(x_3)$

Exercise 1

Suppose you have observed the following number of successes in 10 independent Bernoulli trials:

Data: [1, 0, 1, 1, 0]

where '1' represents a success and '0' represents a failure. Assume that the probability of success in each trial is p.

a) Find the likelihood function for this data given the parameter p.
b) Find the log-likelihood function.
c) Find the maximum likelihood estimate ($\hat{p}$) of the probability of success p.

   Hint: You should do this by taking the derivative of the log-likelihood function with respect to p, setting it to zero, and solving for p.

Kullback-Leibler (KL) divergence

- The KL divergence, $D_{KL}\,[p(x); q(x \mid \theta)]$, measures the difference between two probability distributions:
  - $p(x)$ (often considered the "true" distribution of the data)
  - $q(x \mid \theta)$ (a model distribution parameterized by θ).

$$D_{\mathrm{KL}}[p(x); q(x|\theta)] = \int dx p(x) \log \frac{p(x)}{q(x|\theta)}$$

$$= \mathrm{E}[\log p\,(x)] - \mathrm{E}[\log q\,(x|\theta)]$$

- In MLE, given a set of observed data $x_1, x_2, \ldots, x_n$ drawn from an unknown distribution $p(x)$, we want to find the parameter θ that makes our model distribution $q(x \mid \theta)$ "closest" to the true distribution $p(x)$ in terms of explaining the observed data.

- Consider the second term in the KL divergence:

$$-\int dx p(x) \log q(x|\theta) = -\mathrm{E}[\log q\,(x|\theta)]$$

Minimization of KL divergence $D_{\mathrm{KL}}\left[p(x);q(x\,|\,\theta)\right]$	⟺ Maximization of $\mathrm{E}\left[\log q(x\,|\,\theta)\right]$

- If we have a dataset of N independent and identically distributed (i.i.d.) samples $\{x_i\}_{i=1}^{N}$ drawn from $p(x)$, the empirical expectation can approximate the true expectation for large N:

$$\mathrm{E}[\log q\,(x|\theta)] \simeq \frac{1}{N}\sum_{i=1}^{N} \log q\,(x_i|\theta)$$

$$\mathrm{E}[\log q\,(x|\theta)] \simeq \frac{1}{N}\sum_{i=1}^{N} \log q\,(x_i|\theta)$$

- Notice that maximizing the likelihood function in MLE is equivalent to maximizing its logarithm (the log-likelihood):

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \prod_{i=1}^{N} q(x_i|\theta) = \arg\max_{\theta} \sum_{i=1}^{N} \log q(x_i|\theta)$$

- Using the empirical approximation, minimizing the KL divergence is approximately equivalent to maximizing $\frac{1}{N}\sum_{i=1}^{N} \log q(x_i|\theta)$, which is the same as maximizing the log-likelihood.

Sampling approximation:

$$\mathrm{E}\left[\log q\left(x|\hat{\theta}\right)\right]-\frac{1}{N}\sum_{i=1}^{N}\log q\left(x_i|\hat{\theta}\right)\approx-\left(\hat{\theta}-\theta^0\right)^{\mathrm{T}}E\left[\frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{T}}}\log q\left(x|\hat{\theta}\right)\right]\left(\hat{\theta}-\theta^0\right)$$

$$=\left(\hat{\theta}-\theta^0\right)^{\mathrm{T}}I\left(\hat{\theta}\right)\left(\hat{\theta}-\theta^0\right)$$

Fisher information:

$$I\left(\hat{\theta}\right)\equiv\mathrm{E}\left[\frac{\partial\log q\left(x|\hat{\theta}\right)}{\partial\theta}\frac{\partial\log q\left(x|\hat{\theta}\right)}{\partial\theta^{\mathrm{T}}}\right]=\mathrm{E}\left[-\frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{T}}}\log q\left(x|\hat{\theta}\right)\right]$$

In the limit of large samples (infinite *N*), the maximum likelihood estimator is unbiased and efficient.

$$\hat{\theta}\sim\mathcal{N}\left(\theta^0,\frac{1}{N}I^{-1}\left(\hat{\theta}\right)\right)$$

Maximum likelihood is …
- Asymptotically unbiased (i.e., approaches to a true value).
- Asymptotically efficient (i.e., achieves minimum variance, CRLB).

Suppose a random variable $X \sim p(x;\theta)$, where θ is fixed but unknown. Assume that $p(x;\theta)$ satisfies the “regularity” condition:

$$\mathrm{E}\left[\frac{\partial}{\partial\theta}\log p\left(x|\theta\right)\right] = 0,$$

where the expectation is with respect to $p(x;\theta)$. Then the variance of any unbiased estimator $\hat{\theta}$ satisfies

$$\mathrm{Var}\left[\hat{\theta}\right] \geq \frac{1}{I(\theta)}$$

Fisher information:

$$I(\theta) \equiv \mathrm{E}\left[\left(\frac{\partial \log p\left(x|\theta\right)}{\partial\theta}\right)^2\right] = \mathrm{E}\left[-\frac{\partial^2 \log p\left(x|\theta\right)}{\partial\theta^2}\right]$$

Suppose a random variable $X \sim p(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is fixed but unknown. Assume that $p(x|\boldsymbol{\theta})$ satisfies the "regularity" condition:

$$\mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log p\,(x|\boldsymbol{\theta})\right] = 0,$$

where the expectation is with respect to $p(x;\theta)$. Then the variance of any unbiased estimator $\widehat{\boldsymbol{\theta}}$ satisfies

$$\mathrm{Cov}\left[\widehat{\boldsymbol{\theta}}\right] \geq \mathbf{I}^{-1}(\boldsymbol{\theta})$$

Fisher information matrix:

$$\{\mathbf{I}(\boldsymbol{\theta})\}_{ij} \equiv \mathrm{E}\left[\frac{\partial \log p\,(x|\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \log p\,(x|\boldsymbol{\theta})}{\partial \theta_j}\right] = \mathrm{E}\left[- \frac{\partial^2 \log p\,(x|\theta)}{\partial \theta_i \partial \theta_j}\right]$$

Computing the MLE

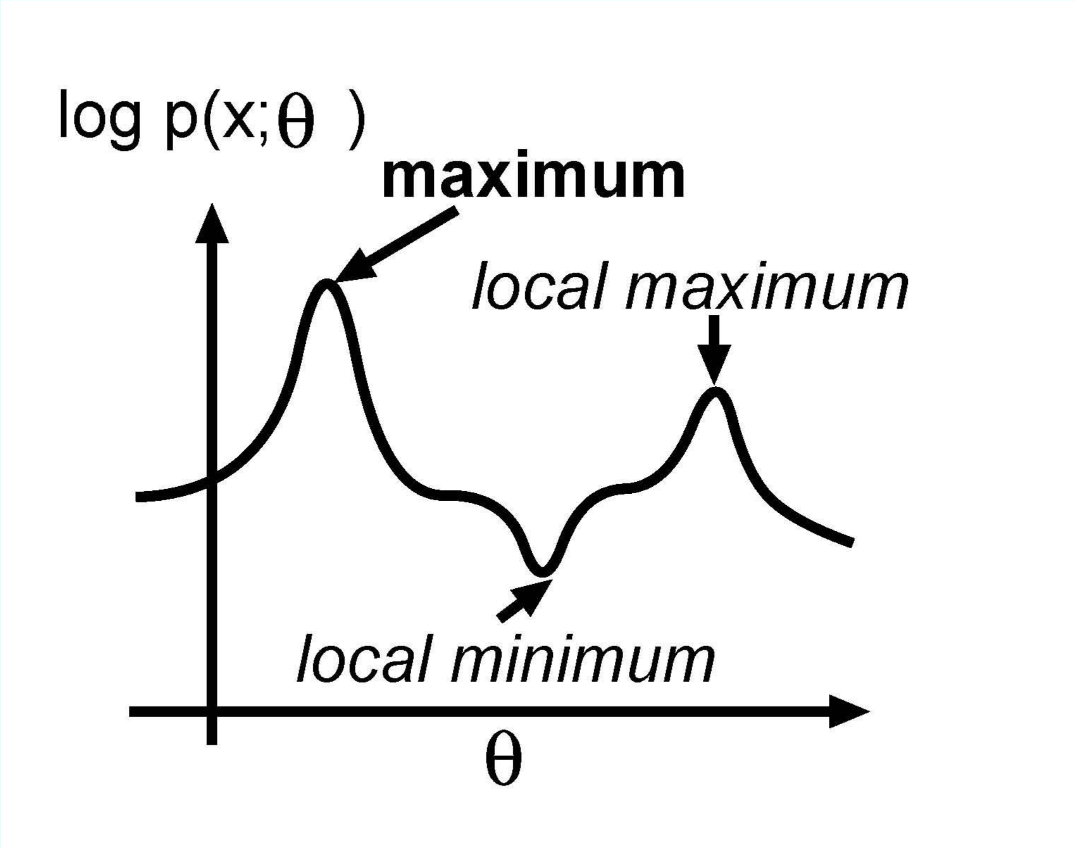
1. Since many models we work with have an exponential form, it is often convenient to maximize the log-likelihood $\ln l(\theta; x)$.

2. If the likelihood function is differentiable, $\hat{\theta}(x)$ is a solution of $\frac{\partial}{\partial\theta}\ln l(\theta; x) = 0$. We need to verify that such a solution is in fact a local max and not a local min or a saddle point.

$\Longrightarrow$ This can be checked whether the Hessian $\frac{\partial^2}{\partial\theta\partial\theta^T}\ln l(\theta; x)$ is negative semidefinite at $\hat{\theta}(x)$.

Computing the MLE

3. If several local maxima exist, MLE is the one with largest likelihood.



Example 1

Suppose $\boldsymbol{X} = [X[0], X[1], \cdots, X[N-1]]^T$, where $X[n] \sim N(\mu, \sigma^2)$, $n = 0, \cdots, N-1$.

Find the MLE $\hat{\mu}$ for μ.

Example 1

Suppose $\boldsymbol{X} = [X[0], X[1], \cdots, X[N-1]]^T$, where $X[n] \sim \mathcal{N}(\mu, \sigma^2)$, $n = 0, \cdots, N-1$. Find the MLE $\hat{\mu}$ for μ.

$$p(\boldsymbol{x}; \mu) = \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x[n]-\mu)^2\right]$$

$$= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-\mu)^2\right]$$

$$\ln p(\boldsymbol{x}; \mu) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-\mu)^2$$

$$\frac{\partial \ln p(\boldsymbol{x};\mu)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-\mu) = 0$$

$$\rightarrow \sum_{n=0}^{N-1}(x[n]-\mu) = 0$$

Hence, MLE is $\hat{\mu} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$

Example 2

Suppose $\boldsymbol{X} = [X[0], X[1], \cdots, X[N-1]]^T$, where $X[n] \sim N(\mu, \sigma^2)$, $n = 0, \cdots, N-1$. Find the MLE $\hat{\theta}$ for $\theta = [\mu, \sigma^2]$.

Example 2

Suppose $\boldsymbol{X} = [X[0], X[1], \cdots, X[N-1]]^T$, where $X[n] \sim N(\mu, \sigma^2)$, $n = 0, \cdots, N-1$. Find the MLE $\hat{\theta}$ for $\theta = [\mu, \sigma^2]$.

$$\ln p(\boldsymbol{x};\theta) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-\mu)^2$$

$$\frac{\partial \ln p(\boldsymbol{x};\theta)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-\mu)$$

$$\frac{\partial \ln p(\boldsymbol{x};\theta)}{\partial (\sigma^2)} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{n=0}^{N-1}(x[n]-\mu)^2$$

Since $\hat{\theta} = [\hat{\mu}, \hat{\sigma}^2]$ should satisfy local maximal condition,

$$\frac{1}{\hat{\sigma}^2}\sum_{n=0}^{N-1}(x[n] - \hat{\mu}) = 0,$$

$$-\frac{N}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2}\sum_{n=0}^{N-1}(x[n] - \hat{\mu})^2 = 0$$

Therefore,

$$\hat{\mu} = \frac{1}{N}\sum_{n=0}^{N-1} X[n]$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=0}^{N-1}(X[n] - \hat{\mu})^2$$

Asymptotic Property

Suppose $X \sim p(x;\theta)$. Let $\hat{\theta}$ be the MLE of θ based on n i.i.d. (independent and identically distributed) realization $X[0], X[1], \cdots, X[N-1]$ of X. Under certain regularity conditions, distribution of $\hat{\theta}$ asymptotically converges as

$$\hat{\theta} \sim N(\theta, \boldsymbol{I}^{-1}(\theta)) \text{ as } N \to \infty.$$

Here, $\boldsymbol{I}(\theta)$ is the Fisher information matrix evaluated at the true θ.

Hence,

- $E\{\hat{\theta}\} \to \theta \Longrightarrow$ MLE is asymptotically unbiased.
- $Cov(\hat{\theta}) \to I^{-1}(\theta) \Longrightarrow$ MLE is asymptotically efficient.

Note: Regularity conditions are:

- Existence of first and second derivatives of log-likelihood function $\ln l(\theta; x)$.
- $E\left\{\frac{\partial \ln p(x;\theta)}{\partial \theta}\right\} = 0$.

Confirmation using Example 2

Suppose $\boldsymbol{X} = [X[0], X[1], \cdots, X[N-1]]^{T}$, where $X[n] \sim N(\mu, \sigma^2)$, $n = 0, \cdots, N-1$. Maximum likelihood estimator $\hat{\theta} = [\hat{\mu}, \hat{\sigma}^2]$ are given by

$$\hat{\mu} = \frac{1}{N} \sum_{n=0}^{N-1} X[n]$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{n=0}^{N-1} (X[n] - \hat{\mu})^2$$

Since random variable $\sum_{n=0}^{N-1} \left(\frac{X[n] - \bar{X}}{\sigma} \right)^2$

(where $\bar{X} = \frac{1}{N} \sum_{n=0}^{N-1} X[n]$) has chi-square distribution with $N-1$ degrees of freedom (χ_{N-1}^2-distribution), its mean and variance are given by $N-1$ and $2(N-1)$. Because of $\frac{N}{\sigma^2} \hat{\sigma}^2 \sim \chi_{N-1}^2$,

$$E[\hat{\sigma}^2] = \frac{N-1}{N}\sigma^2,$$

$$Var(\hat{\sigma}^2) = \left(\frac{\sigma^2}{N^2}\right)^2 \{2(N-1)\}$$

Hence,

$$E\left[\hat{\theta}\right] = \begin{bmatrix} \mu \\ \frac{N-1}{N}\sigma^2 \end{bmatrix} \xrightarrow{(N\to\infty)} \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \theta$$

$$Cov\left(\hat{\theta}\right) = \begin{bmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2(N-1)}{N^2}\sigma^4 \end{bmatrix} \xrightarrow{(N\to\infty)} \begin{bmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2\sigma^4}{N} \end{bmatrix} = I^{-1}(\theta)$$

This shows that $\hat{\theta} = [\hat{\mu}, \hat{\sigma}^2]$ converges asymptotically to an efficient estimator.

Practical Techniques

In practical situations, maximum likelihood estimator cannot be always obtained in explicit form. The likelihood function needs to be maximized via iterative procedure.

- Newton-Raphson method
- EM (Expectation-Maximization) algorithm

Newton-Raphson Method

- **Goal:** Find the parameter value θ^* that maximizes (or minimizes) a function $f(\theta)$. This occurs where the derivative $f'(\theta^*) = 0$.
- **Core Idea:** Iteratively refine an initial guess $\theta(t)$ by using information about the function's first derivative (gradient) and second derivative (Hessian).
- **Update Rule (for maximizing $\boldsymbol{f(\theta)}$):**

$$\theta^{(t+1)} = \theta^{(t)} - \left[Hf\left(\theta^{(t)}\right)\right]^{-1} \nabla f\left(\theta^{(t)}\right)$$

where:

$\theta(t)$ is the parameter estimate at iteration t.

$\nabla f(\theta^{(t)})$ is the gradient of f at $\theta(t)$ (vector of first derivatives).

$Hf\left(\theta^{(t)}\right)$ is the Hessian matrix of f at $\theta(t)$ (matrix of second derivatives).

$\left[Hf\left(\theta^{(t)}\right)\right]^{-1}$ is the inverse of the Hessian matrix.

The EM Algorithm

- **Goal:** Find the Maximum Likelihood Estimates (MLE) of parameters when the model depends on unobserved **latent variables** or has **missing data**.
- **Core Idea:** Iteratively alternate between two steps until convergence:
  - **Expectation (E) Step:**

    Using the current parameter estimates, compute the expectation of the log-likelihood of the complete data (observed + latent/missing).
  - **Maximization (M) Step:**

    Find the parameter values that maximize the expected log-likelihood computed in the E-step.
- **When to Use:** Situations with:
  - Latent variables (e.g., in mixture models, Hidden Markov Models).
  - Missing data.