

430.714

Estimation Theory

Prof. Songhwai Oh
ECE, SNU

Maximum Likelihood Estimation

A procedure for finding the MVU estimator:

Recall

Cramer-Rao lower bound

$\hat{\theta} = \mathbf{g}(\mathbf{x})$ is the MVU estimator if

$$\frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{I}(\boldsymbol{\theta})(\mathbf{g}(\mathbf{x}) - \boldsymbol{\theta})$$

Use Neyman-Fisher factorization theorem to find sufficient statistic

$$p(\mathbf{x}; \theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x}).$$

$T(\mathbf{x})$

Determine if $T(\mathbf{x})$ is complete
See (5.8)

$$\int_{-\infty}^{\infty} v(T)p(T; \theta)dT = 0, \quad \text{for all } \theta$$

is satisfied only by the zero function,
i.e., $v(T) = 0$ for all T .

Find function of $T(\mathbf{x})$ that is unbiased

1. Evaluate $\hat{\theta} = \mathbb{E}(\check{\theta}|T)$, where $\check{\theta}$ is any unbiased estimator.
2. Find a function g so that $\hat{\theta} = g(T)$ is an unbiased estimator of θ .

$\hat{\theta} = g(T(\mathbf{x})) = \text{MVU estimator}$

A Motivating Example: Modified DC Level in WGN

$x[n] = A + w[n]$, $n = 0, 1, \dots, N-1$, where $w[n] \sim \mathcal{N}(0, A)$ and $A > 0$

$$p(\mathbf{x}; A) = \frac{1}{(2\pi A)^{\frac{N}{2}}} \exp \left[-\frac{1}{2A} \sum_{n=0}^{N-1} (x[n] - A)^2 \right]$$

$$\begin{aligned} \frac{\partial \ln p(\mathbf{x}; A)}{\partial A} &= -\frac{N}{2A} + \frac{1}{A} \sum_{n=0}^{N-1} (x[n] - A) + \frac{1}{2A^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \\ &\stackrel{?}{=} I(A)(\hat{A} - A). \end{aligned}$$

No efficient
estimator exists

Neyman-Fisher factorization

$$p(\mathbf{x}; A) = \underbrace{\frac{1}{(2\pi A)^{\frac{N}{2}}} \exp \left[-\frac{1}{2} \left(\frac{1}{A} \sum_{n=0}^{N-1} x^2[n] + NA \right) \right]}_{g \left(\sum_{n=0}^{N-1} x^2[n], A \right)} \underbrace{\exp(N\bar{x})}_{h(\mathbf{x})}$$

$$T(\mathbf{x}) = \sum_{n=0}^{N-1} x^2[n]$$

$$\begin{aligned} E \left[\sum_{n=0}^{N-1} x^2[n] \right] &= NE[x^2[n]] \\ &= N [\text{var}(x[n]) + E^2(x[n])] \\ &= N(A + A^2) \end{aligned}$$

Cannot find $g(T)$ such that $\mathbb{E}(g(T)) = A$.

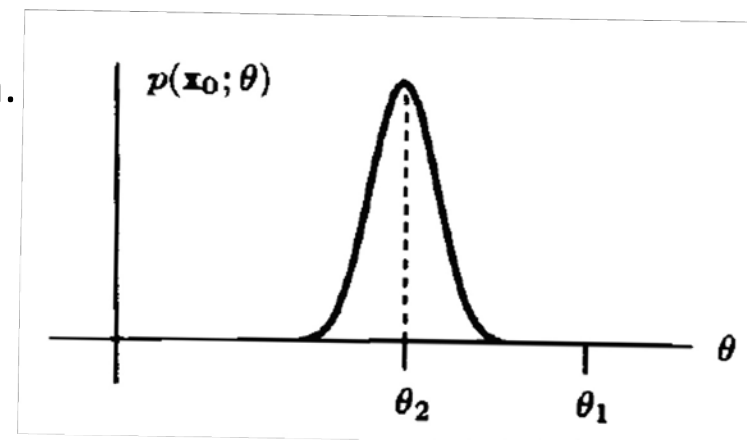
Even worse...

$$E \left(x[0] \left| \sum_{n=0}^{N-1} x^2[n] \right. \right)$$

Can't find the MVU estimator

Maximum Likelihood Estimator (MLE)

- A parameter which maximizes the likelihood of the data.



Modified DC Level in WGN Example

$x[n] = A + w[n]$, $n = 0, 1, \dots, N-1$, where $w[n] \sim \mathcal{N}(0, A)$ and $A > 0$

$$p(\mathbf{x}; A) = \frac{1}{(2\pi A)^{\frac{N}{2}}} \exp \left[-\frac{1}{2A} \sum_{n=0}^{N-1} (x[n] - A)^2 \right]$$

$$\frac{\partial \ln p(\mathbf{x}; A)}{\partial A} = -\frac{N}{2A} + \frac{1}{A} \sum_{n=0}^{N-1} (x[n] - A) + \frac{1}{2A^2} \sum_{n=0}^{N-1} (x[n] - A)^2$$

setting it to zero gives: $\hat{A}^2 + \hat{A} - \frac{1}{N} \sum_{n=0}^{N-1} x^2[n] = 0$ $\hat{A} = -\frac{1}{2} \pm \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} x^2[n] + \frac{1}{4}}$

$$\text{MLE: } \hat{A} = -\frac{1}{2} + \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} x^2[n] + \frac{1}{4}}$$

Maximum Likelihood Estimator (MLE)

$$x[n] = A + w[n], \quad n = 0, 1, \dots, N-1, \quad \text{where } w[n] \sim \mathcal{N}(0, A) \text{ and } A > 0$$

$$\text{MLE: } \hat{A} = -\frac{1}{2} + \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} x^2[n] + \frac{1}{4}}$$

Estimator is biased

$$\begin{aligned} E(\hat{A}) &= E\left(-\frac{1}{2} + \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} x^2[n] + \frac{1}{4}}\right) \\ &\neq -\frac{1}{2} + \sqrt{E\left(\frac{1}{N} \sum_{n=0}^{N-1} x^2[n]\right) + \frac{1}{4}} \quad \text{for all } A \\ &= -\frac{1}{2} + \sqrt{A + A^2 + \frac{1}{4}} \\ &= A. \end{aligned}$$

But as $N \rightarrow \infty$, (LLN)

$$\begin{aligned} \frac{1}{N} \sum_{n=0}^{N-1} x^2[n] &\rightarrow E(x^2[n]) = A + A^2 \\ \hat{A} &\rightarrow A \end{aligned}$$

A good estimator for large N.

Asymptotic Properties of the MLE

Theorem 7.1 (Asymptotic Properties of the MLE) *If the PDF $p(\mathbf{x}; \theta)$ of the data $\mathbf{x}$ satisfies some “regularity” conditions, then the MLE of the unknown parameter θ is asymptotically distributed (for large data records) according to*

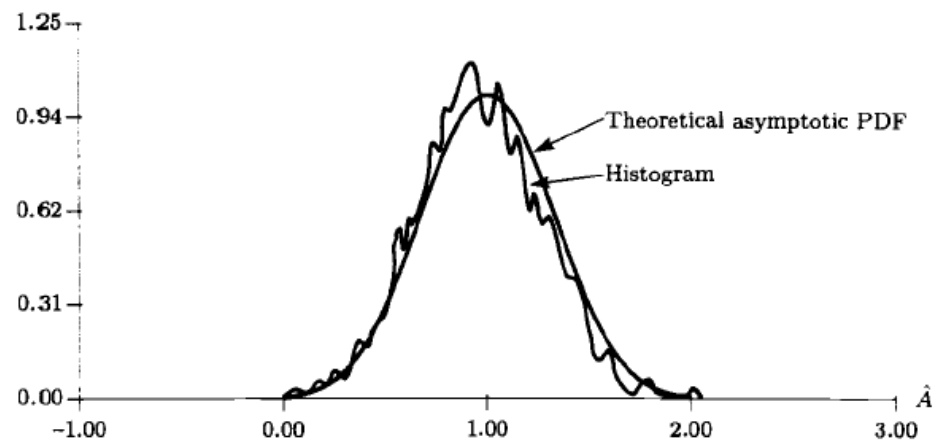
$$\hat{\theta} \stackrel{a}{\sim} \mathcal{N}(\theta, I^{-1}(\theta)) \quad (7.11)$$

where $I(\theta)$ is the Fisher information evaluated at the true value of the unknown parameter.

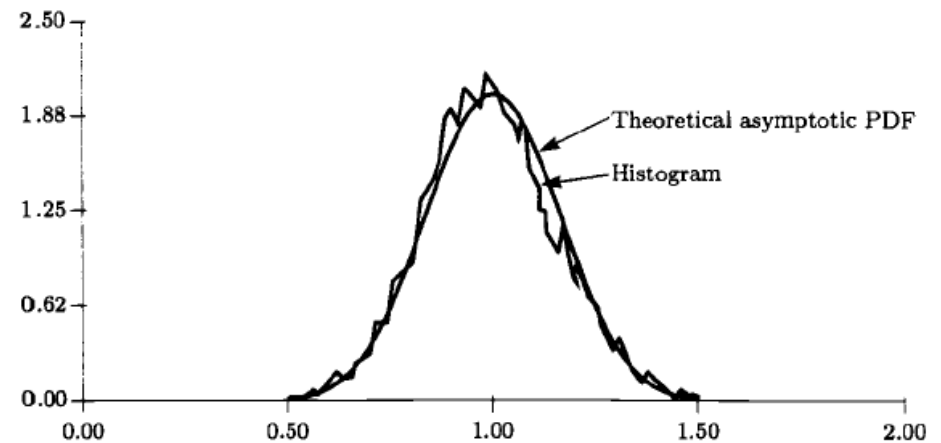
MLE is asymptotically efficient.

Regularity conditions:

- log-likelihood function is differentiable
- Fisher information is nonzero.



(a) $N = 5$



(b) $N = 20$

Notice the scale !!!

Proof Sketch (Asymptotic Properties of the MLE)

$$\begin{aligned}\frac{1}{N} \ln p(\mathbf{x}; \theta) &= \frac{1}{N} \ln \prod_{n=0}^{N-1} p(x[n]; \theta) \\ &= \frac{1}{N} \sum_{n=0}^{N-1} \ln p(x[n]; \theta)\end{aligned}$$

$$\frac{1}{N} \sum_{n=0}^{N-1} \ln p(x[n]; \theta) \rightarrow \int \ln p(x[n]; \theta) p(x[n]; \theta_0) dx[n]$$

$\theta_0 = \text{true value}$
Law of large numbers (LLN)

Kullback-Leibler (KL) divergence:

$$\int \ln \left[\frac{p(x[n]; \theta_1)}{p(x[n]; \theta_2)} \right] p(x[n]; \theta_1) dx[n] \geq 0$$

$$\Rightarrow \int \ln [p(x[n]; \theta_1)] p(x[n]; \theta_1) dx[n] \geq \int \ln [p(x[n]; \theta_2)] p(x[n]; \theta_1) dx[n]$$

$$\int \ln p(x[n]; \theta) p(x[n]; \theta_0) dx[n] \quad \text{maximized when } \theta = \theta_0$$

Hence, as $N \rightarrow \infty$, the MLE is $\hat{\theta} = \theta_0$
(i.e., asymptotically unbiased or consistent).

Proof Sketch (Asymptotic Properties of the MLE)

Take the Taylor expansion about $\theta = \theta_0$. By the mean value theorem,

$$\left. \frac{\partial \ln p(\mathbf{x}; \theta)}{\partial \theta} \right|_{\theta=\hat{\theta}} = \left. \frac{\partial \ln p(\mathbf{x}; \theta)}{\partial \theta} \right|_{\theta=\theta_0} + \left. \frac{\partial^2 \ln p(\mathbf{x}; \theta)}{\partial \theta^2} \right|_{\theta=\tilde{\theta}} (\hat{\theta} - \theta_0)$$

$$\theta_0 < \tilde{\theta} < \hat{\theta}$$

$$0 = \left. \frac{\partial \ln p(\mathbf{x}; \theta)}{\partial \theta} \right|_{\theta=\theta_0} + \left. \frac{\partial^2 \ln p(\mathbf{x}; \theta)}{\partial \theta^2} \right|_{\theta=\tilde{\theta}} (\hat{\theta} - \theta_0)$$

$$\sqrt{N}(\hat{\theta} - \theta_0) = \frac{\frac{1}{\sqrt{N}} \left. \frac{\partial \ln p(\mathbf{x}; \theta)}{\partial \theta} \right|_{\theta=\theta_0}}{-\frac{1}{N} \left. \frac{\partial^2 \ln p(\mathbf{x}; \theta)}{\partial \theta^2} \right|_{\theta=\tilde{\theta}}} \begin{matrix} \xrightarrow{\text{CLT}} \mathcal{N}(0, i(\theta_0)) \\ \xrightarrow{\text{LLN}} -E \left[\left. \frac{\partial^2 \ln p(x[n]; \theta)}{\partial \theta^2} \right|_{\theta=\theta_0} \right] = i(\theta_0) \end{matrix}$$

$$\sqrt{N}(\hat{\theta} - \theta_0) \overset{a}{\sim} \mathcal{N}(0, i^{-1}(\theta_0))$$

$$\hat{\theta} \overset{a}{\sim} \mathcal{N}(\theta_0, I^{-1}(\theta_0))$$

Invariance Properties of the MLE

Theorem 7.2 (Invariance Property of the MLE) *The MLE of the parameter $\alpha = g(\theta)$, where the PDF $p(\mathbf{x}; \theta)$ is parameterized by θ , is given by*

$$\hat{\alpha} = g(\hat{\theta})$$

where $\hat{\theta}$ is the MLE of θ . The MLE of $\hat{\theta}$ is obtained by maximizing $p(\mathbf{x}; \theta)$. If g is not a one-to-one function, then $\hat{\alpha}$ maximizes the modified likelihood function $\bar{p}_T(\mathbf{x}; \alpha)$, defined as

$$\bar{p}_T(\mathbf{x}; \alpha) = \max_{\{\theta: \alpha = g(\theta)\}} p(\mathbf{x}; \theta).$$

$$x[n] = A + w[n], \quad n = 0, 1, \dots, N-1, \quad \text{where } w[n] \sim \mathcal{N}(0, \sigma^2)$$

When $\alpha = \exp(A)$, find the MLE of α .

$$p(\mathbf{x}; A) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \right] \quad -\infty < A < \infty$$

$$p_T(\mathbf{x}; \alpha) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - \ln \alpha)^2 \right] \quad \alpha > 0$$

$$\dots \sum_{n=0}^{N-1} (x[n] - \ln \hat{\alpha}) \frac{1}{\hat{\alpha}} = 0 \quad \hat{\alpha} = \exp(\bar{x}) \quad \hat{\alpha} = \exp(\hat{A}) = \exp(\hat{\theta})$$

* Check out the book for the case when g is not a one-to-one function.

Linear model: $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$, where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$

$$p(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{(2\pi)^{\frac{N}{2}} \det^{\frac{1}{2}}(\mathbf{C})} \exp \left[-\frac{1}{2}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{C}^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \right]$$

$$J(\boldsymbol{\theta}) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{C}^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})$$

- MLE $\hat{\boldsymbol{\theta}}$ minimizes $J(\boldsymbol{\theta})$.
- MLE is the global minimum of $J(\boldsymbol{\theta})$.

Appendix 3C.5

$$\begin{aligned} \frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_k} = & -\frac{1}{2} \text{tr} \left(\mathbf{C}^{-1}(\boldsymbol{\theta}) \frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \theta_k} \right) + \frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})^T}{\partial \theta_k} \mathbf{C}^{-1}(\boldsymbol{\theta})(\mathbf{x} - \boldsymbol{\mu}(\boldsymbol{\theta})) \\ & -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}(\boldsymbol{\theta}))^T \frac{\partial \mathbf{C}^{-1}(\boldsymbol{\theta})}{\partial \theta_k} (\mathbf{x} - \boldsymbol{\mu}(\boldsymbol{\theta})) \end{aligned}$$

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbf{H}\boldsymbol{\theta}$$

$$\frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial (\mathbf{H}\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}} \mathbf{C}^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})$$

$$\mathbf{H}^T \mathbf{C}^{-1}(\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}) = \mathbf{0}$$

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}^{-1} \mathbf{x}$$

For linear models, the MLE is an efficient estimator and the MVU estimator.
The MLE is optimal for linear models.