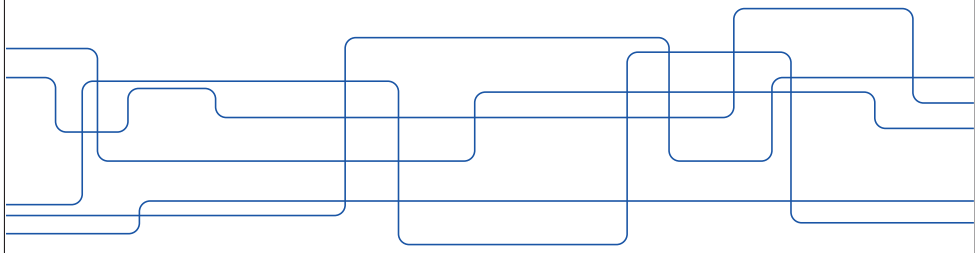
KTH ROYAL INSTITUTE
OF TECHNOLOGY

Estimation theory – Lecture 3

- Ch. 6 Best Linear Unbiased Estimators
- Ch. 7 Maximum Likelihood Estimation

Magnus Jansson



Linear Estimator

It may be difficult to find the general MVU estimator, let us try to restrict the class of estimators to be linear in the data:

$$\hat{\theta} = \sum_{n=0}^{N-1} a_n x(n) = \mathbf{a}^T \mathbf{x}$$

Look for the *best linear unbiased estimator*, the BLUE!

Unbiased constraint

The estimator should be unbiased, i.e.,

$$\mathrm{E}\{\hat{\theta}\} = \mathrm{E}\{\mathbf{a}^T \mathbf{x}\} = \mathbf{a}^T \mathrm{E}\{\mathbf{x}\} = \theta \quad \forall \theta$$

For this to hold it essentially implies that we have a linear data model

$$\mathbf{x} = \mathbf{s}\theta + \mathbf{w}$$

such that $\mathrm{E}\{\mathbf{x}\} = \mathbf{s}\theta$. Hence, amplitude estimation of known signal in zero-mean noise.

The unbiased constraint then becomes $\mathbf{a}^T \mathbf{s} = 1$.

Variance

$$\begin{aligned}\mathrm{var}(\hat{\theta}) &= \mathrm{E}\{(\hat{\theta} - \theta)^2\} = \mathrm{E}\{(\mathbf{a}^T \mathbf{x} - \mathbf{a}^T \mathrm{E}\{\mathbf{x}\})^2\} \\ &= \mathbf{a}^T \mathrm{E}\{(\mathbf{x} - \mathrm{E}\{\mathbf{x}\})(\mathbf{x} - \mathrm{E}\{\mathbf{x}\})^T\}\mathbf{a} = \mathbf{a}^T \mathbf{C} \mathbf{a}\end{aligned}$$

where $\mathbf{C}$ is the covariance matrix of the data $\mathbf{x}$ or equivalently of the noise $\mathbf{w}$.

Scalar BLUE solution

Problem:

$$\min_{\text{s.t. } \mathbf{a}^T\mathbf{s}=1} \mathbf{a}^T\mathbf{C}\mathbf{a}$$

Solution: By the Cauchy-Schwartz inequality

$$1 = |\mathbf{a}^T\mathbf{s}|^2 = |\mathbf{a}^T\mathbf{C}^{1/2} \cdot \mathbf{C}^{-1/2}\mathbf{s}|^2 \leq \mathbf{a}^T\mathbf{C}\mathbf{a} \cdot \mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}$$

or

$$\mathbf{a}^T\mathbf{C}\mathbf{a} \geq \frac{1}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}}$$

We have equality when $\mathbf{C}^{1/2}\mathbf{a} = \mathbf{C}^{-1/2}\mathbf{s}\,\alpha$ for some constant α. To satisfy the constraint we must have $\alpha = 1/\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}$. Hence,

$$\mathbf{a} = \frac{\mathbf{C}^{-1}\mathbf{s}}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}} \quad \text{and} \quad \hat{\theta} = \frac{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{x}}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}}.$$

BLUE: The vector case

The estimator $\hat{\boldsymbol{\theta}} = \mathbf{A}^T\mathbf{x}$ should be unbiased, i.e.,

$$\mathrm{E}\{\hat{\boldsymbol{\theta}}\} = \mathrm{E}\{\mathbf{A}^T\mathbf{x}\} = \mathbf{A}^T\,\mathrm{E}\{\mathbf{x}\} = \boldsymbol{\theta} \quad \forall \boldsymbol{\theta}$$

For this to hold it essentially implies that we have a linear data model

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

such that $\mathrm{E}\{\mathbf{x}\} = \mathbf{H}\boldsymbol{\theta}$. Hence, amplitude estimation of multiple known signals in zero-mean noise (linear regression).

The unbiased constraint then becomes $\mathbf{A}^T\mathbf{H} = \mathbf{I}$.

Vector BLUE

Problem:

$$\min_{\text{s.t.}\mathbf{A}^T\mathbf{H}=\mathbf{I}} \mathbf{a}_i^T\mathbf{C}\mathbf{a}_i \quad \forall i = 1, 2, \ldots, p$$

where $\mathbf{a}_i$ is the ith column of $\mathbf{A}$.
The p sub-problems are connected by the constraints. However, since the criterion functions involve independent parameters, we might as well study

$$\min_{\text{s.t.}\mathbf{A}^T\mathbf{H}=\mathbf{I}} \sum_{i=1}^{p} \mathbf{a}_i^T\mathbf{C}\mathbf{a}_i = \min_{\text{s.t.}\mathbf{A}^T\mathbf{H}=\mathbf{I}} \mathrm{Tr}\{\mathbf{A}^T\mathbf{C}\mathbf{A}\}$$

Solution: Change variables $\mathbf{B} = \mathbf{C}^{1/2}\mathbf{A}$ and $\mathbf{K} = \mathbf{C}^{-1/2}\mathbf{H}$.

Vector BLUE cont'd

$$\min_{\text{s.t.}\mathbf{B}^T\mathbf{K}=\mathbf{I}} \mathrm{Tr}\{\mathbf{B}^T\mathbf{B}\}$$

Let

- $\boldsymbol{\Pi}_\mathbf{K} = \mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T$ be the orthogonal projection onto the span of $\mathbf{K}$ and
- $\boldsymbol{\Pi}_\mathbf{K}^\perp = \mathbf{I} - \boldsymbol{\Pi}_\mathbf{K}$ the orthogonal projection onto the nullspace of $\mathbf{K}^T$.

The criterion function can then be written as

$$\begin{aligned}\mathrm{Tr}\{\mathbf{A}^T\mathbf{C}\mathbf{A}\} = \mathrm{Tr}\{\mathbf{B}^T\mathbf{B}\} &= \mathrm{Tr}\{\mathbf{B}^T\boldsymbol{\Pi}_\mathbf{K}\mathbf{B}\} + \mathrm{Tr}\{\mathbf{B}^T\boldsymbol{\Pi}_\mathbf{K}^\perp\mathbf{B}\} \\ &= \mathrm{Tr}\{(\mathbf{K}^T\mathbf{K})^{-1}\} + \mathrm{Tr}\{\mathbf{B}^T\boldsymbol{\Pi}_\mathbf{K}^\perp\mathbf{B}\}\end{aligned}$$

where we used the constraint.

Vector BLUE cont'd

Next note that $\mathbf{B}^T\mathbf{\Pi}_{\mathbf{K}}^{\perp}\mathbf{B} = \mathbf{B}^T\mathbf{\Pi}_{\mathbf{K}}^{\perp}\cdot\mathbf{\Pi}_{\mathbf{K}}^{\perp}\mathbf{B}$ and

$$\mathbf{\Pi}_{\mathbf{K}}^{\perp}\mathbf{B} = \mathbf{B} - \mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T\mathbf{B} = \mathbf{B} - \mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}$$

The criterion now reads

$$\mathrm{Tr}\{(\mathbf{K}^T\mathbf{K})^{-1}\} + \mathrm{Tr}\{\left(\mathbf{B} - \mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}\right)^T\left(\mathbf{B} - \mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}\right)\}$$

and is clearly minimized when $\mathbf{B} = \mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}$ or in the original variables

$$\mathbf{A} = \mathbf{C}^{-1}\mathbf{H}(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}$$

and, hence,

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

BLUE summary

Data model:

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

The noise is zero-mean with covariance matrix $\mathbf{C}$ (pdf arbitrary).

BLUE:

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

The covariance matrix of $\hat{\boldsymbol{\theta}}$ is:

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

The variances of $\hat{\theta}_i$ are of course given by the diagonal elements:

$$\mathrm{var}(\hat{\theta}_i) = \left[(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\right]_{ii}$$

Maximum Likelihood (ML) Estimation (Ch. 7)

Given a probabilistic description of the problem, the ML method is very natural and powerful:

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}}(\mathbf{x}) = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x};\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \ln[p(\mathbf{x};\boldsymbol{\theta})]$$

Having observed $\mathbf{x} = \mathbf{x}_0$, $p(\mathbf{x}_0;\boldsymbol{\theta})$ reflects how likely it is that $\mathbf{x}_0$ was generated from the pdf $p(\mathbf{x};\boldsymbol{\theta})$.

Properties of the MLE

A common type of result: If $p(\mathbf{x};\boldsymbol{\theta})$ satisfies some regularity conditions, then

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}) \to \mathcal{N}(0, \mathbf{C})$$

in distribution as $N \to \infty$. Here, $\mathbf{C} = \lim_{N\to\infty} N\cdot\mathbf{I}^{-1}(\boldsymbol{\theta})$.

The regularity conditions require at least existence of derivatives of the log likelihood function and that the Fisher information matrix is positive definite.

One can typically also show that $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is a (strongly) consistent estimator of $\boldsymbol{\theta}$ as $N \to \infty$.

Properties cont'd

In most cases, the MLE is practically (for large samples)

- consistent,
- asymptotically unbiased,
- asymptotically efficient,
- asymptotically Gaussian.

Efficiency

For the linear Gaussian model

$$\hat{\theta}_{\mathrm{ML}} = \hat{\theta}_{\mathrm{MVU}} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}$$

and

$$\hat{\theta}_{\mathrm{ML}} \sim \mathcal{N}(\theta, (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1})$$

so it is unbiased, efficient and Gaussian in *finite* samples.

Efficiency cont'd

If an efficient estimator exists, recall the CRLB identity condition:

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

At a stationary point of the ML criterion

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln\ p(\mathbf{x};\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\mathrm{ML}}} = 0$$

If FIM is positive definite, then we must have $\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \hat{\boldsymbol{\theta}}(\mathbf{x})$.

That is, if an efficient estimator exists, ML will provide it!

Invariance property

- Assume we have data described by $p(\mathbf{x};\boldsymbol{\theta})$ but we are interested in a related quantity $\boldsymbol{\alpha} = g(\boldsymbol{\theta})$.
- What is the MLE of $\boldsymbol{\alpha}$?
- Solution: If $g(\cdot)$ is a one-to-one mapping, then

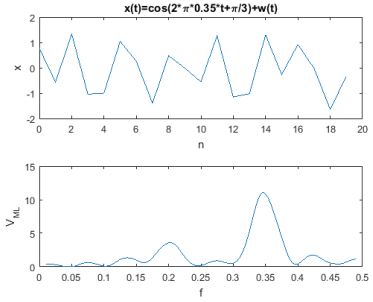
$$\hat{\boldsymbol{\alpha}}_{\mathrm{ML}} = g(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})$$

- If $g(\cdot)$ is a many-to one mapping, then $\hat{\boldsymbol{\alpha}}_{\mathrm{ML}}$ maximizes the transformed likelihood

$$p_T(\mathbf{x};\boldsymbol{\alpha}) = \max_{\boldsymbol{\theta}:\boldsymbol{\alpha}=g(\boldsymbol{\theta})} p(\mathbf{x};\boldsymbol{\theta})$$

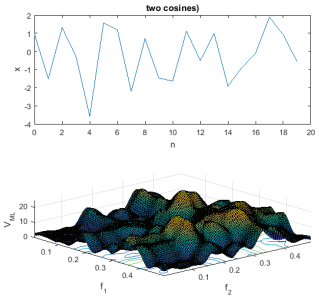
ML example 1

Data: $x(n) = \cos(2\pi n 0.35 + \pi/3) + w(n); n = 0, 1, \ldots, 19$ where $w(n)$ is white Gaussian noise with variance 0.25. $V_{ML}(f)$ is the concentrated likelihood function to be maximized wrt f.



ML example 2

Data: $x(n) = \cos(2\pi n 0.35 + \pi/3) + \cos(2\pi n 0.4 - \pi/3) + w(n)$ where $w(n)$ is white Gaussian noise with variance 1. $V_{ML}(f_1, f_2)$ is the concentrated likelihood function to be maximized.



ML example 3

Consider the data model

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

where $\mathbf{H}$ is a known $N \times p$ real matrix, $\boldsymbol{\theta}$ a $p \times 1$ vector of unknown parameters, and $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_N)$. We want to estimate $\boldsymbol{\theta}$ and σ^2 given $\mathbf{x}$.

- Derive CRB
- Derive ML estimator
- Discuss performance in general as well as when $N \to \infty$
- Let $p = N/2$, and discuss the results