Practical estimators

Raj Thilak Rajan

Overview

1. Recap
2. Maximum Likelihood Estimation (MLE)
3. Best Linear Unbiased Estimator (BLUE)
4. Summary

Overview

Consider estimating θ from stochastic observations

$$p(\mathbf{x}; \theta),$$

i.e., characterized by the pdf, which in turn is parameterized by θ. Let the potential estimator take the form

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

Note that

- $\hat{\theta}$ itself is a random variable, and
- performance of $\hat{\theta}$ should be described *statistically*

Unbiasedness and Optimality criterion

Unbiased estimators: Let $\hat{\theta} = g(\mathbf{x})$ be an estimator of θ, then if $\hat{\theta}$ is an unbiased estimator, then

$$\mathbb{E}(\hat{\theta}) = \int g(\mathbf{x}) p(\mathbf{x};\theta) d\mathbf{x} = \theta \quad \text{for all } \theta,$$

where $p(\mathbf{x};\theta)$ is the probability density function. In other words, for an unbiased estimator

$$bias(\theta) = \mathbb{E}(\hat{\theta}) - \theta = 0.$$

Mean square error (MSE): The MSE of $\hat{\theta}$ is

$$\begin{aligned} mse(\hat{\theta}) &= \mathbb{E}\left[(\hat{\theta}-\theta)^2\right] = \mathbb{E}\left\{\left[(\hat{\theta}-\mathbb{E}(\hat{\theta})) + (\mathbb{E}(\hat{\theta})-\theta)\right]^2\right\} \\ &= \mathbb{E}\left[(\hat{\theta}-\mathbb{E}(\hat{\theta}))^2\right] + (\mathbb{E}(\hat{\theta})-\theta)^2 = var(\hat{\theta}) + (\mathbb{E}(\hat{\theta})-\theta)^2 \end{aligned}$$

Minimum Variance Unbiased Estimator (MVU)

Constrain the bias of the MSE to zero, i.e., consider $\mathbb{E}(\hat{\theta}) = \theta$, then

$$mse(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2\right] + (\mathbb{E}(\hat{\theta}) - \theta)^2 = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2\right] = var(\hat{\theta})$$

where $\hat{\theta}$ is an unbiased estimator, and let

$$var(\hat{\theta}) \leq var(\tilde{\theta})$$

for any other unbiased estimator $\tilde{\theta}$, then $\hat{\theta}$ is the minimum variance unbiased estimator (MVU) for all θ.

Does a MVU exist i.e., an unbiased estimator with minimum variance for all θ ?

Cramér-Rao Lower Bound (CRLB)

- Assume the pdf $p(\mathbf{x};\theta)$ satisfies the regularity condition:

$$\mathbb{E}\left[\frac{\partial \ln\ p(\mathbf{x};\theta)}{\partial\theta}\right] = 0,$$

then the variance of any unbiased estimator $\hat{\theta}$ satisfies

$$var(\hat{\theta}) \geq \frac{1}{-\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta^2}\right]} = \frac{1}{\mathbb{E}\left[\left(\frac{\partial \ln\ p(\mathbf{x};\theta)}{\partial\theta}\right)^2\right]} = \frac{1}{I(\theta)}$$

- An estimator is *efficient* if it meets the CRLB with equality, in which case the estimator is the MVU.
- However, the converse is not necessarily true.

MVUE and CRLB

- An unbiased estimator may be found that attains the bound for all θ iff

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

for some function g and I, then $\hat{\theta} = g(\mathbf{x})$ is an estimator with

$$\text{Mean} : \mathbb{E}(\hat{\theta}) = \theta \qquad \text{Variance} : var(\hat{\theta}) = \frac{1}{I(\theta)}.$$

- If $s(\mathbf{x};\theta) = I(\theta)(g(\mathbf{x}) - \theta)$, then $\hat{\theta}$ is the MVU Estimator (MVUE)

Practical estimators

Motivation:

- Determining the MVU requires to knowledge of the PDF.
- Even when knowing the PDF, finding the MVU is not guaranteed.

Sub-optimal estimators:

- MLE: Maximum Likelihood Estimator
- BLUE: Best Linear Unbiased Estimator
- LS: Least Squares (next lecture)

Under certain conditions, these sub-optimal estimators

- equal the MVU, or
- their variance converges to the variance of the MVU.

Maximum Likelihood Estimator (MLE)

- MLE (for a scalar θ) is the value of θ that maximizes $p(\mathbf{x};\theta)$ for a fixed $\mathbf{x}$
- Kay-I, Theorem 7.1: Asymptotic properties of the MLE: If the PDF $p(\mathbf{x};\theta)$ of data $\mathbf{x}$ satisfies some "regularity conditions", then the MLE of the unknown parameter θ is asymptotically distributed (for large data records) according to
$$\hat{\theta} \overset{a}{\sim} \mathcal{N}(\theta, I^{-1}(\theta))$$
where $I(\theta)$ is the Fisher information evaluated at the true value of the unknown parameter.
- MLE is asymptotically unbiased and efficient
- If an efficient estimators exists, the ML will (generally) produce it

Example 2 : MLE

Consider estimating A $(A > 0)$ for the following model

$$x[n] = A + w[n]\,, \qquad n = 0, \cdots, N-1 \quad w[n] \sim \mathcal{N}(0, A)$$

- PDF:

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

- Score:

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

Example 2 : MLE

- The MLE is obtained by setting score to zero i.e.,

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

- We then obtain

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

- Solving the above and choosing the positive $\hat{A}$:

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

- It can be shown that $\hat{A}$ has these asymptotic properties
$\mathrm{E}(\hat{A}) \xrightarrow{a} A$ and $\mathrm{var}(\hat{A}) \xrightarrow{a} \dfrac{A^2}{N(A+\frac{1}{2})}$

MLE: Linear Gaussian Model

For the *linear Gaussian model*, the likelihood function is given by

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

It is clear that this function is maximized by solving

$$\hat{\theta} = \arg\min_{\theta}[(\mathbf{x}-\mathbf{h}\theta)^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{h}\theta)]$$

Note that since $\mathbf{x}$ is a stochastic variable that can take many values, so is $\hat{\theta}$.

MLE: Linear Gaussian Model (2)

Solve:

$$J = \min_{\theta}[(\mathbf{x} - \mathbf{h}\theta)^T \mathbf{C}^{-1}(\mathbf{x} - \mathbf{h}\theta)]$$

Solution: Expanding the cost function

$$J = (\mathbf{x} - \mathbf{h}\theta)^T \mathbf{C}^{-1}(\mathbf{x} - \mathbf{h}\theta) = \mathbf{x}^T \mathbf{C}^{-1}\mathbf{x} - 2\mathbf{h}^T \mathbf{C}^{-1}\mathbf{x}\theta + \mathbf{h}^T \mathbf{C}^{-1}\mathbf{h}\theta^2$$

and setting the gradient w.r.t. θ as zero, we have

$$\frac{\partial J}{\partial \theta} = -2\mathbf{h}^T \mathbf{C}^{-1}\mathbf{x} + 2\mathbf{h}^T \mathbf{C}^{-1}\mathbf{h}\theta = 0 \ \rightarrow \ \hat{\theta} = \left(\mathbf{h}^T \mathbf{C}^{-1}\mathbf{h}\right)^{-1} \mathbf{h}^T \mathbf{C}^{-1}\mathbf{x}$$

Note that for the *linear Gaussian model*, the MLE is the MVU estimator.

MLE: Transformed parameters

- The MLE of the parameter $\alpha = g(\theta)$, where the PDF $p(\mathbf{x};\theta)$ is parametrized by θ, is given by

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

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

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

Best Linear Unbiased Estimator (BLUE)

To obtain the BLUE we constrain the estimator to have the form $\hat{\theta} = \mathbf{a}^T\mathbf{x}$.

Requirements:

- Unbiased:

$$\mathbb{E}(\hat{\theta}) = \mathbf{a}^T\mathbb{E}(\mathbf{x}) = \theta \quad \text{for all }, \theta$$

  which is feasible if $\mathbb{E}(\mathbf{x}) = \mathbf{h}\theta$ and $\mathbf{a}^T\mathbf{h} = 1$, for known $\mathbf{h}$.

- Minimum variance:

$$\begin{aligned} var(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2] = \mathbb{E}\{[\mathbf{a}^T(\mathbf{x} - \mathbb{E}(\mathbf{x}))]^2\} \\ &= \mathbf{a}^T \underbrace{\mathbb{E}[(\mathbf{x} - \mathbb{E}(\mathbf{x}))(\mathbf{x} - \mathbb{E}(\mathbf{x}))^T]}_{\mathbf{C}_x} \mathbf{a} \\ &= \mathbf{a}^T\mathbf{C}_x\mathbf{a}, \end{aligned}$$

Solve:

$$\min_{\mathbf{a}} \mathbf{a}^T\mathbf{C}_x\mathbf{a} \quad \text{subject to} \quad \mathbf{a}^T\mathbf{h} = 1$$

Solution to the BLUE

Solve:

$$\min_{\mathbf{a}} \mathbf{a}^T \mathbf{C}_x \mathbf{a} \quad \text{subject to} \quad \mathbf{a}^T \mathbf{h} = 1$$

Solution:

- Use the method of the Lagrange multipliers, we have

$$J = \mathbf{a}^T \mathbf{C}_x \mathbf{a} + \lambda(\mathbf{a}^T \mathbf{h} - 1)$$

- Setting the gradient with respect to $\mathbf{a}$ to zero we get,

$$\frac{\partial J}{\partial \mathbf{a}} = 2\mathbf{C}_x \mathbf{a} + \lambda \mathbf{h} = \mathbf{0} \qquad \Rightarrow \qquad \mathbf{a} = -\frac{\lambda}{\mathbf{2}} \mathbf{C}_{\mathbf{x}}^{\mathbf{-1}} \mathbf{h}$$

Solution to the BLUE (2)

The Lagrange multiplier λ is found using the constraint

$$\mathbf{a}^T\mathbf{h} = -\frac{\lambda}{2}\mathbf{h}^T\mathbf{C}_x^{-1} = 1 \quad \Rightarrow -\frac{\lambda}{2} = \frac{1}{\mathbf{h}^T\mathbf{C}_x^{-1}\mathbf{h}}$$

and the optimal $\mathbf{a}$ is given by

$$\mathbf{a}_{opt} = \frac{\mathbf{C}_x^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{C}_x^{-1}\mathbf{h}}$$

The BLUE estimator is then

$$\hat{\theta} = \mathbf{a}_{opt}^T\mathbf{x} = \frac{\mathbf{h}^T\mathbf{C}_x^{-1}\mathbf{x}}{\mathbf{h}^T\mathbf{C}_x^{-1}\mathbf{h}}$$

with variance

$$var(\hat{\theta}) = \mathbf{a}_{opt}^T\mathbf{C}_x^{-1}\mathbf{a}_{opt} = \frac{1}{\mathbf{h}^T\mathbf{C}_x^{-1}\mathbf{h}}$$

BLUE for linear model

For the linear model

$$\mathbf{x} = \mathbf{h}\theta + \mathbf{w}, \qquad \text{with} \quad \mathbb{E}(\mathbf{w}) = \mathbf{0} \quad \text{and} \quad \mathbf{cov}(\mathbf{w}) = \mathbf{C}$$

For this model, the BLUE is given by

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

Remarks:

- For estimation of the parameters of a *linear* model, the BLUE equals the MVU, if the noise is Gaussian.
- To compute the BLUE, we do not need the complete PDF, we only need to know the mean ($\mathbf{h}$, up to scale) and the covariance matrix ($\mathbf{C}_x$) of $\mathbf{x}$.

BLUE: Gauss-Markov theorem

For the general linear model:

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

- $\mathbf{H}$ $(N \times p)$ is the known observation matrix
- $\boldsymbol{\theta}$ $(p \times 1)$ is the unknown parameter
- $\mathbf{w}$ $(N \times 1)$ is the noise with zero mean and covariance $\mathbf{C}$

BLUE:

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

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

Summary

Key points:

- MVU estimator requires the PDF, and existence is not guaranteed
- MLE maximizes $p(\mathbf{x}; \theta)$ for a fixed $\mathbf{x}$
- MLE is asymptotically unbiased and efficient
- BLUE constraints the estimator to have the form $\hat{\theta} = \mathbf{a}^T \mathbf{x}$
- BLUE does not require the full PDF information, but only the first two moments

Next session:

- Least Squares

Assignments

Solve:

- Example 1 (this lecture): Consider the measurement model $\mathbf{x} \sim \mathcal{N}(A, 0.5A)$. Find the CRLB, BLUE and MLE for A.
- Kay-I, Problem 7.21: For N IID observation from the PDF $\mathcal{N}(A, \sigma^2)$, where A and σ^2 are both unknown, find the MLE of the SNR $\alpha = A^2/\sigma^2$
- Kay-I, Problem 6.4: The observed samples $x[0], x[1], \ldots, x[N-1]$ are IID according to the following PDFs:
  - Laplacian: $p(x[n]; \mu) = 0.5\ exp(-x[n] - \mu)$
  - Gaussian: $p(x[n]; \mu) = (2\pi)^{-0.5} exp(-0.5(x[n] - \mu)^2)$

  Find the BLUE of the mean μ. Discuss the properties of the respective estimators.

Review and derivations:

- Kay-I, Section 6.6, Section 7.10: Signal processing examples
- Kay-I, Theorem 7.3: Asymptotic Properties of the MLE
- Kay-I, 7A, 7B: Monte Carlo methods and Asymptotic PDF of MLE