Least Squares

Raj Thilak Rajan

Overview

1. Recap
2. Least Squares estimator
3. Least Squares variants
4. Least Squares properties
5. Summary

Minimum Variance Unbiased Estimator (MVU)

- Consider the estimation of the unknown scalar parameter θ, from the stochastic measurement vector

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

where the PDF is parameterized by θ. A potential estimator $\hat{\theta} = g(\mathbf{x})$ is stochastic, with some statistical properties.

- Let $\hat{\theta}$ 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.

MVU and CRLB

- An unbiased estimator may be found that attains the Cramér-Rao Lower Bound (CRLB) 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 $s(\mathbf{x};\theta) = I(\theta)(g(\mathbf{x}) - \theta)$, for an unbiased estimator $\hat{\theta} = g(\mathbf{x})$ whose Fisher information is given by $\mathbf{I}(\theta)$, then $\hat{\theta}$ is the MVU estimator.

Maximum Likelihood Estimator (MLE)

- Consider the *general linear Gaussian model*, where the likelihood function is

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

  Solve:

$$J = \min_{\theta}\left[(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})\right]$$

  Solution:

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

- Asymptotic property : Let $I(\boldsymbol{\theta})$ be the Fisher information, then the MLE is asymptotically distributed (for large data records) according to

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\theta}, I^{-1}(\boldsymbol{\theta})) \qquad \text{(under some regularity conditions on the PDF)}$$

Best Linear Unbiased Estimator (BLUE)

- Consider the *general linear Gaussian model* for unknown parameter $\boldsymbol{\theta}(p \times 1)$:

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

  where $\mathbf{H}$ $(N \times p)$ is the known observation matrix. Constrain the estimator to have the form $\hat{\theta} = \mathbf{a}^T\mathbf{x}$, which leads to the BLUE

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

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

- To compute the BLUE, we do not need the complete PDF, we only need to know the first two moments

Optimality criterion

- Mean square error (MSE)

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

- MSE can be decomposed into
  - variance of the estimator
  - bias of the estimator, which is a function of the unknown parameter.

Least squares criterion

$$\begin{aligned}\text{Measurement vector} &: \quad \mathbf{x} = x[0], x[1], \ldots, x[N-1] \\ \text{Signal vector} &: \quad \mathbf{s}(\theta) = s[0], s[1], \ldots, s[N-1] \\ \text{Unknown parameter} &: \quad \theta\end{aligned}$$

Least Squares criterion:

$$J(\theta) = \sum_{n=0}^{N-1} (x[n] - s[n])^2$$

Properties:

+ No probabilistic assumptions required
- Estimator may not be statistically efficient

Least squares example

Consider estimating $s[n] = A$ for the following model

$$x[n] = s[n] + w[n]\,, \qquad n = 0, \cdots, N-1 \quad w[n] \text{ is some perturbation}$$

- Solve:

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

- Solution:

$$\frac{\partial J(A)}{\partial A} = 2\sum_{n=0}^{N-1} (A - x[n]) = 0$$

or

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

Least squares estimator (LSE)

For the *linear model*, the LSE solves

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\mathbf{x} - \underbrace{\mathbf{H}\boldsymbol{\theta}}_{\mathbf{s}}\|_2^2$$

Problem:

$$\min_{\boldsymbol{\theta}} \|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|_2^2$$

Solution:

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

Proof: Set the derivative of the cost function

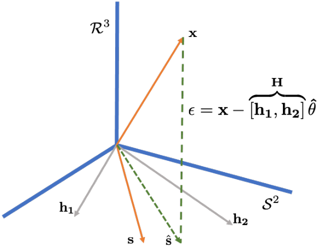
$$J(\boldsymbol{\theta}) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}),$$

with respect to $\boldsymbol{\theta}$ to 0.

Geometrical interpretation

- Euclidean distance: The LS error minimizes the squared distance between the data vector and the signal vector i.e.,

$$J(\boldsymbol{\theta}) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \overbrace{(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})}^{\boldsymbol{\epsilon}} = \|\boldsymbol{\epsilon}\|_2^2.$$



- Orthogonality principle: The error vector $\boldsymbol{\epsilon}$ is orthogonal to the subspace $\mathcal{S}$ spanned by $\mathbf{H}$ i.e.,

$$\underbrace{(\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}})}_{\boldsymbol{\epsilon}} \perp \mathcal{S}$$

- Projections: Let $\mathbf{P}$ be a projection on $\mathcal{S}$, then

$$\hat{\mathbf{s}} = \mathbf{P}\mathbf{x} \quad \text{and} \quad J_{min} = \|\mathbf{P}^{\perp}\mathbf{x}\|_2^2,$$

where $\mathbf{P}^{\perp} = (\mathbf{I} - \mathbf{P})$.

Constrained Least squares

Solve:

$$\min_{\boldsymbol{\theta}} \|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|_2^2 \quad \text{s.t. } \mathbf{A}\boldsymbol{\theta} = \mathbf{b}$$

Define:

$$\mathbf{A}^T = \mathbf{Q}\begin{bmatrix}\mathbf{R} \\ \mathbf{0}\end{bmatrix}, \quad \mathbf{H}\mathbf{Q} = \begin{bmatrix}\mathbf{H}_1 & \mathbf{H}_2\end{bmatrix}, \quad \mathbf{Q}^T\boldsymbol{\theta} = \begin{bmatrix}\boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2\end{bmatrix}$$

Solution:

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_1 &: \quad \text{Solve } \mathbf{R}^T\boldsymbol{\theta}_1 = \mathbf{b} \\
\hat{\boldsymbol{\theta}}_2 &: \quad \text{Solve } \min_{\boldsymbol{\theta}_2} \|\mathbf{H}_2\boldsymbol{\theta}_2 - (\mathbf{x} - \mathbf{H}_1\hat{\boldsymbol{\theta}}_1)\|_2^2 \\
\hat{\boldsymbol{\theta}} &: \quad \text{Solve } \mathbf{Q}\begin{bmatrix}\hat{\boldsymbol{\theta}}_1 \\ \hat{\boldsymbol{\theta}}_2\end{bmatrix}
\end{aligned}$$

Non-linear Least squares

For the non-linear model, the LSE solves

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} \|\mathbf{x} - \mathbf{s}(\boldsymbol{\theta})\|_2^2$$

Problem:

$$\min_{\boldsymbol{\theta}} \|\mathbf{x} - \mathbf{s}(\boldsymbol{\theta})\|_2^2$$

Solution:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \left[\mathbf{H}^T(\boldsymbol{\theta}_k)\mathbf{H}(\boldsymbol{\theta}_k)\right]^{-1}\mathbf{H}^T(\boldsymbol{\theta}_k)\left[\mathbf{x} - s(\boldsymbol{\theta}_k)\right]$$

Proof: Linearize the non-linear function $\mathbf{s}(\boldsymbol{\theta})$ around $\boldsymbol{\theta}_0$

$$\mathbf{s}(\boldsymbol{\theta}) \approx \; \mathbf{s}(\boldsymbol{\theta}_0) + \left.\frac{\partial \mathbf{s}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

Statistical properties

Consider a data model with noise, i.e.,

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

where $\mathbf{w}$ is the noise vector. Recollect, the LS estimator yields

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

Discussion:

- Does the LS estimate yield an unbiased estimate ?
- What is the MSE of the LS estimate ?
- When is the LS estimate statistically optimal ?

Statistical properties (2)

Consider a data model with noise, i.e.,

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

where $\mathbf{w}$ is the noise vector. Recollect, the LS estimator yields

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

Summary

Key points:

- In LS approach, we minimize the squared difference between the given data and (noisy) signal model
- No probabilistic assumptions are levied on the measurements
- No claim on statistical optimality of the estimator can be made without more information on the underlying noise.
- Geometrical interpretation of LS, Constrained LS and Weighted LS.
- When is the LS estimator (statistically) optimal ?

Next session:

- Bayesian philosophy

Assignments

Solve:

- Consider the measurement $\mathbf{x} \sim \mathcal{N}(A, A)$. Find the LS and CRLB for A, if they exist. Discuss the properties of the estimator(s).
- 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 LS for μ and discuss properties

Reading:

- Kay-I, Chapter 8: Non-linear Least Squares
- Kay-I, Section 4.5, 8.4: Linear least squares (weighted LS)
- Kay-I, Section 8.6: Order-recursive LS (column update)
- Kay-I, Section 8.7: Sequential LS (column update)
- Kay-I, Section 8.10: Signal processing examples