Introduction to (Bayesian) Estimation
MAE 5020

Best linear unbiased estimators

Oklahoma State University

August 15, 2023

Overview

Objective: **Design** an unbiased and efficient *linear* estimator, i.e., best linear unbiased estimator (BLUE)

- WLSE is not always unbiased or efficient.
- Assumptions: 1) $H(k)$ is deterministic; 2) $V(k)$ is zero mean with a PD covariance matrix $R(k)$.
- Design is more complicated but analysis is easier.

Connection

The BLUE of θ is a special case of WLSE where $W(k) = R(k)^{-1}$. $P(k)$ in the recursive WLSE is the covariance matrix for the error between θ and $\hat{\theta}_{BLU}(k)$ in BLUE.

Property

BLUEs are invariant under changes of scales!

Problem

Given $Z(k) = H(k)\theta + V(k)$,

- Assumption 1:
- Assumption 2:
- Linear estimator:

Design $F(k)$ such that

1. $\hat{\theta}_{BLU}(k)$ is unbiased
2. The error variance of each element of $\hat{\theta}_{BLU}(k)$ is minimized.

Solution methodology (sketch)

1. Create constraints on $F(k)$ to ensure unbiasedness of $\hat{\theta}_{BLU}(k)$.

2. Express the variance $E([\theta_i - \hat{\theta}_{i,BLU}(k)]^2)$ in terms of $F(k)$. Here the covariance of $V(k)$ is used.

3. Minimize the variance in Step 2 with the constraints in Step 1 using Lagrange multipliers.

The derived BLUE and its properties

$$\hat{\theta}_{BLU}(k) = \left(H^T(k)R^{-1}(k)H(k)\right)^{-1} H^T(k)R^{-1}(k)Z(k).$$

- The BLUE is a special case of WLSE with $W(k) = R^{-1}(k)$. If $R(k) = \sigma_v^2 I$, $\hat{\theta}_{BLU} = \hat{\theta}_{LS}$.
- Most efficient unbiased estimators that are linear in the measurements $Z(k)$ *given the linear form of the measurement*.

- $cov(\tilde{\theta}_{BLU}(k)) = \left(H^T(k)R^{-1}(k)H(k)\right)^{-1}$, which is the $P(k)$ used in the recursive WLSE.
- The recursive form is the same as recursive WLSE, where $w^{-1}(k+1)$ is replaced with $R(k+1)$.

Invariance to scale changes

$\hat{\theta}_{BLU}(k)$ is invariant under changes of scale.

Proof: Observers A and B read the measurements of the same process in two different scales, related by M.

$$Z_B(k) = H_B(k)\theta + V_B(k) = MZ_A(k) = M(H_A(k)\theta + V_A(k))$$

Proof

Let $\hat{\theta}_{A,BLU}(k)$ and $\hat{\theta}_{B,BLU}(k)$ denote the BLUEs associated with observers A and B. Show $\hat{\theta}_{A,BLU}(k) = \hat{\theta}_{B,BLU}(k)$. The BLUE

algorithm automatically normalizes the data.

Final conclusion

- $R(k)$ is known and $H(k)$ is deterministic, use $\hat{\theta}_{BLUE}(k)$.
- $R(k)$ is known and $H(k)$ is random, use $\hat{\theta}_{WLS}(k)$ with $W(k) = R^{-1}(k)$.
- $R(k)$ is unknown: use $\hat{\theta}_{WLS}(k)$ with heuristic $W(k)$.

Likelihood

- Probability is associated with a forward experiment/model:

- Likelihood is associated with an inverse model:

Hypothesis H

Suppose that θ can be only 0 or 1, then there are two hypotheses associated with θ, $H_0 : \theta = 0$ and $H_1 : \theta = 1$, i.e., *binary hypothesis*.

- Extend to θ taking discrete values, say 10 values.

- Extend to θ taking values within an interval $a \leq \theta \leq b$:

- A vector of parameters, say $\theta \in \mathbb{R}^{n \times 1}$ and each element takes 2 values.

Null hypothesis

All other possibilities that are not already accounted for by the enumerated hypotheses.

Example

Results (of an experiment)

Results are outputs/data of an experiment.

Example

In the linear model $Z(k) = H(k)\theta + V(k)$, results are the data in $Z(k)$ and $H(k)$.

- $P(R|H)$:

- For a fixed H, we can apply the three axioms of probability.

Likelihood

Definition

Likelihood $L(H|R)$ of the hypothesis H given the results R and a specific probability model is proportional to $P(R|H)$ with an arbitrary constant ratio c, i.e.,

$$L(H|R) = cP(R|H) \quad (or \ \propto P(R|H)).$$

- In likelihood, R is fixed where H is variable (or the parameters in the probability model are variables).
- There are no axioms of likelihood.

Example

Probability of the occurrence of boys and girls in a family of two children (binomial model):

$$P(R|p) = \frac{(m+f)!}{m!f!} p^m (1-p)^f$$

Two data sets:

$$R_1 = \{1 \text{ boy and } 1 \text{ girl}\}$$

$$R_2 = \{2 \text{ boys}\}$$

Two hypotheses:

$$H_1 : \; p = 1/2$$

$$H_2 : \; p = 1/4$$

Calculate $P(R|H)$ and $L(H|R)$

Continuous distributions

- if R is described by a continuous distribution, the probability obtaining a result within $(R, R + dR)$ is given by $P(R|H)dR$, where $P(R|H)$ is the pdf.
- $L(H|R) = cP(R|H)dR$, but $c\ dR$ can be considered another constant
- $L(H|R) = c_1 P(R|H)$, where $P(R|H)$ is the pdf.

Likelihood ratio and test

- On the same dataset, we can form ratios of likelihoods (*likelihood ratio*).

Likelihood-ratio test

$$L(H_1, H_2|R) = \frac{L(H_1|R)}{L(H_2|R)} = \frac{P(R|H_1)}{P(R|H_2)}$$

$$\begin{array}{rl} H_1 & L(H_1, H_2, R) > c \\ H_2 & L(H_1, H_2, R) < c \\ H_1 or H_2 & L(H_1, H_2, R) = c \end{array}$$

Independent data sets

Likelihood of independent data sets:

$$L(H|R_1, \cdots, R_m) = cP(R_1, \cdots, R_m|H)$$

Log likelihood (i.e., $\log L(H|R)$) is used often.

Example: Gaussian random variable generator

Hypothesis:

Results:

Likelihood:

LRT:

Maximum-Likelihood Estimation (MLE)

Find an estimate $\hat{\theta}_{ML}$ that maximizes the data likelihood

- Need the likelihood function:
- Genearlly, mathematical optimization/programming is needed.
- Special case (generic linear model):

Develop MLE

Unknown vector θ in a probability model describing N independent identically distributed (iid) observations $z(k)$, $k = 1, \cdots, N$: $Z = (z(1), \cdots, z(N))$.
Derive the likelihood $\ell(\theta|Z) \propto p(Z|\theta)$

Log-likelihood function

An MLE

$$\hat{\theta}_{ML} = \arg\max \ell(\theta|Z) \quad (or \;\; \arg\max L(\theta|Z)).$$

- If L is differentiable, the partial derivative w.r.t. θ must be zero at the $\hat{\theta}_{ML}$:

$$\frac{\partial L(\theta|Z)}{\partial \theta}\Big|_{\theta=\hat{\theta}_{ML}} = 0.$$

- For maximization, the second order derivative (Hessian) should be negative definite.

$$J_o(\hat{\theta}_{ML}|Z) = \frac{\partial^2 L(\theta|Z)}{\partial \theta_i \partial \theta_j}\Big|_{\theta=\hat{\theta}_{ML}} < 0, \quad i,j = 1,2,\cdots,n.$$

- Recall that the Fisher information matrix is indeed given by $-J_o(\hat{\theta}_{ML}|Z)$, which is positive definite.

Properties

- Very popular and widely used
- Large-sample properties: consistent, asymptotically Gaussian with mean θ and covariance J^{-1}/N, and asymptotically efficient
- *Functions of maximum-likelihood estimates are themselves maximum-likelihood estimates*:

MLE of mean and variance of a Gaussian rv

Observe random samples $z(1), \cdots, z(N)$ of the output of a Gaussian random number generator and would like to compute a ML estimate of its mean μ and variance σ^2.

The linear model: $Z(k) = H(k)\theta + V(k)$

- Common assumptions with BLUE: $V(k) \in \mathbb{R}^N$ is zero mean white noise, with covariance $R(k)$, $H(k)$ is deterministic.
- Likelihood: (Additionally) assume a Gaussian model on $V(k)$
- What about $p(Z(k)|\theta)$?

Show $\hat{\theta}_{ML}(k) = \hat{\theta}_{BLU}(k)$

Maximize $P(Z(k)|\theta)$ is equivalent to

If $R(k) = \sigma_v^2 I$,

A dynamical system example

For any MLE problem, 1) obtain the expression of $L(\theta|Z)$ and 2) maximize $L(\theta|Z)$ w.r.t. θ which typically requires optimization.

Now we look at a LTI system and derive the likelihood function of unknown parameters in the system.

$$
\begin{aligned}
&x(k+1) = \Phi x(k) + \Psi u(k) \\
&z(k+1) = Hx(k+1) + v(k+1) \in \mathbb{R}^m, \quad k = 0, \cdots, N-1.
\end{aligned}
$$

Here $u(k)$ is known, $x(0)$ is deterministic, $v(k)$ is a zero mean Gaussian with $E(v(k)v(j)^T) = R\delta_{kj}$. (iid Gaussian noise).

Say θ contains all the unknown parameters in Φ, Ψ, H and R. Also we assume that θ is identifiable.

The log-likelihood $L(\theta|Z)$

MLE

θ appears in L in a complex nonlinear manner. The only way to do it is to use nonlinear optimization to obtain a local optimal of $\hat{\theta}_{ML}$.