Maximum Likelihood

M. Sami Fadali
Electrical Engineering
UNR

What is covered?

- Two steps to obtain estimate:

1. Obtain the likelihood or log likelihood.
2. Maximize to obtain the estimates.

- Relationship between maximum likelihood (ML), BLUE and weighted least-squares estimators.
- Maximum a Posteriori (MAP)

Likelihood & Log-likelihood

- N i.i.d. observations $z(k), k = 1,2,\ldots,N$.
- θ = vector of unknown parameters.

$$\boldsymbol{Z} = [z(1) \quad z(2) \quad \ldots \quad z(N)]^T$$

$$l(\theta|\boldsymbol{Z}) \propto p(\boldsymbol{Z}|\theta) = \prod_{i=1}^{N} p(z(i)|\theta)$$

$$L(\theta|\boldsymbol{Z}) = \ln[l(\theta|\boldsymbol{Z})] = \sum_{i=1}^{N} \ln[p(z(i)|\theta)]$$

Maximum Likelihood

- Optimum parameter estimates maximize the likelihood $l(\theta|\boldsymbol{Z})$ for a particular set of measurements $\boldsymbol{Z}$.
- Log is a monotonic transformation and log-likelihood $L(\theta|\boldsymbol{Z})$ can be used.
- Taylor series gives necessary and sufficient conditions for a maximum.

Likelihood Maximization

$$L(\theta/\boldsymbol{Z}) \approx L\left(\hat{\theta}_{ML}|\boldsymbol{Z}\right) + \left.\frac{\partial L(\theta|\boldsymbol{Z})}{\partial \theta}\right]_{\hat{\theta}_{ML}}^{T} \Delta\theta$$

$$+\frac{1}{2!}\Delta\theta^{T}\left[\left.\frac{\partial^2 L(\theta|\boldsymbol{Z})}{\partial \theta^2}\right]_{\hat{\theta}_{ML}}\right]\Delta\theta, \qquad \Delta\theta = \theta - \hat{\theta}_{ML}$$

- Necessary: $\left.\frac{\partial L(\theta|\boldsymbol{Z})}{\partial \theta}\right]_{\hat{\theta}_{ML}}^{T} = \boldsymbol{0}$
- Sufficient: $\left[\left.\frac{\partial^2 L(\theta|\boldsymbol{Z})}{\partial \theta^2}\right]_{\hat{\theta}_{ML}}\right] < 0$

Example

- N independent normally distributed observations $z(i), i = 1, \ldots, N$.
- Unknown mean μ and variance σ^2.
- Find the maximum likelihood estimators.

$$p\left(z(i)\middle|\mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(z(i)-\mu)^2}{2\sigma^2}\right\}$$

$$\ln\left[p\left(z(i)\middle|\mu, \sigma^2\right)\right] = -\frac{1}{2}\ln\left[2\pi\sigma^2\right] - \frac{[z(i)-\mu]^2}{2\sigma^2}$$

Log-likelihood

$$l(\mu,\sigma^2) = p\left(z(1),\dots,z(N)\middle|\mu,\sigma^2\right) = \prod_{i=1}^{N} p\left(z(i)\middle|\mu,\sigma^2\right)$$

$$L(\mu,\sigma^2) = \sum_{i=1}^{N} \ln\left[p\left(z(i)\middle|\mu,\sigma^2\right)\right]$$

$$= \sum_{i=1}^{N} \left\{-\frac{1}{2}\ln\left[2\pi\sigma^2\right] - \frac{[z(i)-\mu]^2}{2\sigma^2}\right\}$$

$$= -\frac{N}{2}\ln\left[2\pi\sigma^2\right] - \sum_{i=1}^{N} \frac{[z(i)-\mu]^2}{2\sigma^2}$$

Maximum Likelihood: Necessary

$$L(\mu,\sigma^2) = -\frac{N}{2}\left(\ln[2\pi] + \ln[\sigma^2]\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}[z(i)-\mu]^2$$

$$\left.\frac{\partial L(\mu,\sigma^2)}{\partial \theta}\right]_{\hat{\theta}_{ML}} = \begin{bmatrix} \dfrac{\partial L}{\partial \mu} \\ \dfrac{\partial L}{\partial \sigma^2} \end{bmatrix}_{\hat{\theta}_{ML}} = \begin{bmatrix} \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{N}[z(i)-\mu] \\ -\dfrac{N}{2}\times\dfrac{1}{\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{N}[z(i)-\mu]^2 \end{bmatrix}_{\hat{\theta}_{ML}} = \mathbf{0}$$

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_{i=1}^{N} z(i) = \bar{z}, \qquad \widehat{\sigma^2}_{ML} = \frac{1}{N}\sum_{i=1}^{N}[z(i)-\bar{z}]^2$$

Sufficient

$$\frac{\partial L(\mu,\sigma^2)}{\partial\theta} = \left[\frac{1}{\sigma^2}\sum_{i=1}^{N}[z(i)-\mu] \quad -\frac{N}{2}\times\frac{1}{\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{N}[z(i)-\mu]^2\right]^T$$

$$\left.\frac{\partial^2 L(\mu,\sigma^2)}{\partial\theta^2}\right|_{\hat{\theta}_{ML}} = \left.\begin{bmatrix} -\frac{N}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{i=1}^{N}[z(i)-\mu] \\ -\frac{1}{\sigma^4}\sum_{i=1}^{N}[z(i)-\mu] & \frac{N}{2}\times\frac{1}{\sigma^4}-\frac{1}{\sigma^6}\sum_{i=1}^{N}[z(i)-\mu]^2 \end{bmatrix}\right|_{\hat{\theta}_{ML}}$$

$$= \begin{bmatrix} -\frac{N}{\widehat{\sigma^2}_{ML}} & 0 \\ 0 & -\frac{N}{2\left(\widehat{\sigma^2}_{ML}\right)^2} \end{bmatrix} < 0$$

Properties of ML Estimates

Theorem: ML estimates are

1. Consistent.
2. Asymptotically Gaussian with mean θ and covariance matrix $J^{-1} = J_i^{-1}/N$.

   $$J_i = -E\left\{\frac{\partial^2 \ln[f(z_i)]}{\partial \theta^2}\right\}, i = 1, \dots, N$$

   Fisher information matrix.
3. Asymptotically efficient.

Example: Exponential Distribution

Given N i.i.d. measurements from an exponentially distributed population.

$$f(z_i) = \theta \exp(-\theta\, z_i)\,, i = 1, \ldots N$$

(i) Obtain the Fisher information

(ii) Obtain the Cramer-Rao lower bound (CRLB)

Solution

$$f(\boldsymbol{z}) = \prod_{i=1}^{N} f(z_i), \qquad f(z_i)\} = \theta \exp(-\theta z_i)$$

$$\ln[f(\boldsymbol{z})] = \sum_{i=1}^{N} \ln[f(z_i)], \qquad \ln[f(z_i)] = \ln[\theta] - \theta z_i$$

$$\frac{\partial(\ln[f(z_i)])}{\partial\theta} = \frac{1}{\theta} - z_i$$

$$J_i = -E\left\{\frac{\partial^2(\ln[f(z_i)])}{\partial\theta^2}\right\} = \frac{1}{\theta^2}, i = 1,2. \dots, N$$

$J = \sum_{i=1}^{N} J_i = N/\theta^2$ CRLB θ^2/N

Maximum-likelihood Estimate

- Necessary Condition

$$\sum_{i=1}^{N} \frac{\partial(\ln[f(z_i)])}{\partial\theta} = \sum_{i=1}^{N} \left[\frac{1}{\theta} - z_i\right] = 0$$

$$\frac{N}{\hat{\theta}_{ML}} = \sum_{i=1}^{N} z_i \,, \hat{\theta}_{ML} = \frac{N}{\sum_{i=1}^{N} z_i} = \frac{1}{\bar{z}}$$

Can show $E\{\hat{\theta}_{ML}\} = \frac{N}{N-1}\theta$

Invariance Property of MLEs

- Continuous function $g(\theta): \mathcal{R}^n \rightarrow \subset \mathcal{R}^r$
- Any continuous function of a consistent estimator is itself a consistent estimator.

$$\widehat{g(\theta)} = g\left(\hat{\theta}\right)$$

- MLEs are consistent.

$$\widehat{g(\theta)}_{ML} = g\left(\hat{\theta}_{ML}\right)$$

Example

- Obtain MLE of the error variance $var(v)$ for the linear model $z(k) = \boldsymbol{h}^T\theta + v(k)$

Approach 1: Obtain the log-likelihood for $var(v)$ and maximize s.t. $var(v) > 0$ (constrained maximization: difficult)

Approach 2: Obtain the log-likelihood for $[var(v)]^{1/2}$ and maximize then square.

(unconstrained maximization: easier)

Comparison of Estimators

- Compare MLE, WLSE, and BLUE.
- Use linear model with H deterministic.

$$\boldsymbol{z}(k) = H(k)\theta + \boldsymbol{v}(k)$$

- **Assume** zero-mean Gaussian white noise $\boldsymbol{v}(k)$ with known covariance matrix $R(k)$.

$$E\{\boldsymbol{z}(k)|\theta\} = H(k)\theta$$

$$var\{\boldsymbol{z}(k)|\theta\} = E\left\{\boldsymbol{v}(k)\boldsymbol{v}^T(k)\right\} = R(k)$$

Gaussian Density Functions

$$p\big(\boldsymbol{v}(k)\big) = \frac{1}{\sqrt{(2\pi)^N \det\big(R(k)\big)}} \exp\left\{-\frac{1}{2}\boldsymbol{v}^T(k)R^{-1}(k)\boldsymbol{v}(k)\right\}$$

$$p(\boldsymbol{z}(k)|\theta) = \frac{1}{\sqrt{(2\pi)^N \det\big(R(k)\big)}} \exp\left\{-\frac{1}{2}[\boldsymbol{z}(k) - H(k)\theta]^T R^{-1}(k)[\boldsymbol{z}(k) - H(k)\theta]\right\}$$

- Linear transformation of Gaussian is Gaussian.
- Brown & Hwang, Ch.1: linear transformation of Gaussian process, and the Jacobian matrix is the identity for $\boldsymbol{v}$ w.r.t. $\boldsymbol{z}$.

Theorem

- For the linear model with deterministic $H(k)$ and multivariate zero-mean **Gaussian** white noise $\boldsymbol{v}(k)$, MLE and BLUE are identical.

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

- The estimators are

  (i) unbiased, (ii) the most efficient linear estimators, (iii) consistent, and (iv) Gaussian.

Proof

$$p(\mathbf{z}(k)|\theta) = \frac{1}{\sqrt{(2\pi)^N \det\big(R(k)\big)}} \exp\left\{-\frac{1}{2}[\mathbf{z}(k) - H(k)\theta]^T R^{-1}(k)[\mathbf{z}(k) - H(k)\theta]\right\}$$

- Maximizing $p \Leftrightarrow$ minimizing quadratic in its exponent.

$$\frac{d}{d\theta}[\mathbf{z}(k) - H(k)\theta]^T R^{-1}(k)[\mathbf{z}(k) - H(k)\theta]\Big]_{\hat{\theta}_{ML}(k)} = \mathbf{0}$$

- Minimizing the quadratic gives BLUE.
- BLUE is WLSE with $W(k) = R^{-1}(k)$

Properties of Estimators

- Unbiased since BLUE are unbiased.
- Most efficient since BLUE are most efficient.
- Consistent since MLE are consistent.
- Gaussian because they depend linearly on the Gaussian measurement.

Corollary

- For the linear model with (i) deterministic $H(k)$ and (ii) multivariate Gaussian noise $\boldsymbol{v}(k)$ whose variance is $R(k) = \sigma_v^2 I$, MLE, BLUE, and LSE are identical.

$$\hat{\theta}_{ML}(k) = \hat{\theta}_{BLUE}(k) = \hat{\theta}_{LS}(k)$$

- The estimators are

  (i) unbiased, (ii) the most efficient linear estimators, (iii) consistent, and (iv) Gaussian (equally weighted measurements).

Important Application

- Discrete LTI state-space model.
- Obtain MLE of the model parameters.
- Deterministic initial condition $\boldsymbol{x}(0)$

$$\boldsymbol{x}(k+1) = \Phi\boldsymbol{x}(k) + \Psi\boldsymbol{u}(k)$$

$$\boldsymbol{z}(k+1) = H\boldsymbol{x}(k+1) + \boldsymbol{v}(k+1),$$

$$k = 0, 1, \dots, N-1$$

$$E\{\boldsymbol{v}(k)\} = \boldsymbol{0}, \qquad E\{\boldsymbol{v}(k)\boldsymbol{v}^T(j)\} = \bar{R}\delta_{kj}$$

Likelihood

$$L(\theta|\boldsymbol{z}(N)) = \ln\left[\prod_{i=1}^{N} p(z(i)|\theta)\right]$$

$$p(z(i)|\theta)$$
$$= \frac{1}{\sqrt{[2\pi]^N \det(\bar{R})}} \exp\left\{-\frac{1}{2}[\boldsymbol{z}(i) - H\boldsymbol{\theta}(i)]^T \bar{R}^{-1}[\boldsymbol{z}(i) - H\boldsymbol{\theta}(i)]\right\}$$

- Determine the likelihood for N measurements.
- **Terms depend on parameters:** $\boldsymbol{\theta}$, $covariance$, H
- Find the maximum with state equation as a constraint.

Parameters to Estimate

- θ = col{entries of the system matrices}
- In practice, some of the parameters are known and we estimate a subset.
- Treat state equation as a constraint and numerically obtain the solution (difficult).

$$\boldsymbol{x}(k+1) = \left[\begin{array}{cccc} \multicolumn{2}{c|}{0_{(n-1)\times 1}} & \multicolumn{2}{c}{I_{n-1}} \\ \hline -a_0 & -a_1 & \dots & -a_{n-1} \end{array}\right] \boldsymbol{x}(k) + \left[\begin{array}{c} 0_{(n-1)\times 1} \\ \hline 1 \end{array}\right] u(k)$$

$$y(k) = [b_0 \quad b_1 \quad \dots \quad b_{n-1}]\boldsymbol{x}(k)$$

$$\boldsymbol{\theta} = col\{\boldsymbol{a}, \boldsymbol{b}\}, \boldsymbol{a} = [a_0 \quad a_1 \quad \dots \quad a_{n-1}], \boldsymbol{b} = [b_0 \quad b_1 \quad \dots \quad b_{n-1}]$$

Maximum a Posteriori (MAP)

- A Posteriori: given the data
- Given the distribution of the parameters
- Maximize $p(\theta|\boldsymbol{z})$= a posteriori pdf

$$p(\theta|\boldsymbol{z}) = \frac{p(\boldsymbol{z}|\theta)p(\theta)}{p(\boldsymbol{z})}$$

$$\hat{\theta}_{MAP} = arg\left\{\max_{\boldsymbol{\theta}} p(\boldsymbol{z}|\theta)p(\theta)\right\}$$

Example: Estimate of Mean

- Use N i.i.d. Gaussian measurement

$$z(i) \sim \mathcal{N}(\mu, \sigma^2)$$

- Obtain a map estimate of μ

$$p(\mu) = \left(2\pi\sigma_\mu^2\right)^{-1/2} \exp\left(-\frac{\mu^2}{2\sigma_\mu^2}\right)$$

Solution

$$p(\boldsymbol{z}(N), \mu) = p(\boldsymbol{z}(N)|\mu)p(\mu)$$

$$= p(\mu) \prod_{i=1}^{N} p(z(i)|\mu)$$

$$L\left(\mu, \sigma^2 \middle| \boldsymbol{z}(N)\right) = \ln[p(\mu)] + \sum_{i=1}^{N} \ln[p(z(i)|\mu)]$$

$$= \text{no} - \mu \text{ terms} - \frac{\mu^2}{2\sigma_\mu^2} - \sum_{i=1}^{N} \frac{[z(i) - \mu]^2}{2\sigma^2}$$

Example (Cont.)

$$\max_{\mu} f(\mu) = \max_{\mu}\left\{-\frac{\mu^2}{2\sigma_\mu^2} - \sum_{i=1}^{N}\frac{[z(i)-\mu]^2}{2\sigma^2}\right\}$$

$$\frac{\partial f(\mu)}{\partial \mu} = -\frac{\mu}{\sigma_\mu^2} + \frac{1}{\sigma^2}\sum_{i=1}^{N}[z(i)-\mu] = 0$$

$$\hat{\mu}_{MAP}(N) = \frac{1}{N+\sigma^2/\sigma_\mu^2}\sum_{i=1}^{N} z(i)$$

No info. abt' mean $\left(\sigma_\mu^2 \to \infty\right)$: $\hat{\mu}_{MAP}(N) = \bar{z} = \hat{\mu}_{ML}(N)$

MATLAB

- MLE: gives the maximum likelihood estimate for a given data vector.
- Character string: distribution
- Default=normal, others available.
  - [p,pci]=mle('distribution', data)

p =parameter estimate

pci = 95% confidence interval

MATLAB Example

```
>> x=randn(100,1); % Standard normal.
>> [p,pci]=mle('normal', x) % param., 95% Conf.I
p =
   0.0912    0.9220
pci =
  -0.0927    0.8136
   0.2751    1.0765
```