University of BRISTOL

Computer Science Year 2

Algorithms & Data

Estimation, Regression, Classification

Prof Alin Achim

Last time …

- Minimum Variance Unbiased Estimator (MVUE)
  - MSE criterion sometimes leads to unrealistic estimators
  - MVUE constrains the bias to be zero and minimises the variance
  - No known standard procedure to find the MVUE
- Cramer-Rao Lower Bound (CRLB)
  - Establishes a lower bound for the variance of an estimator
  - Provides a benchmark against which we can compare the performance of any unbiased estimator
  - Rules-out impossible estimators

Objectives

- Maximum Likelihood Estimation (MLE)
- Motivation
- Likelihood principle
- Computing the MLE
- Examples
- Asymptotic properties

DC level in WGN (modified example)

- Consider the multiple observations

$$x[n] = A + w[n],\ \text{where } n = 0,1,\ldots,N-1 \ and \ w[n] \sim N(0,A)$$

- The PDF in this case is

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

- The derivative of the log-likelihood:-

$$\frac{\partial \ln p\,(A|x)}{\partial A} = -\frac{N}{2A} + \frac{1}{A}\sum_{n=0}^{N-1}(x[n]-A) + \frac{1}{2A^2}\sum_{n=0}^{N-1}(x[n]-A)^2$$
$$\stackrel{?}{=} I(A)\left(\hat{A}-A\right)$$

- The CRLB however (without proof):-

$$\mathrm{var}\left(\hat{A}\right) \geq \frac{A^2}{N\left(A+\frac{1}{2}\right)}$$

The likelihood principle

- Definition
  - The likelihood function is defined by

$$L(\theta|x) \equiv P(x|\theta) = \prod_{i=1}^{N} P(x_i|\theta)$$

i.e. $L(\theta|x)$ is the probability that the data x is observed, given that the parameter value is θ. In other words, unlike in the PDF, we view the observation as being fixed and the parameter θ as freely varying.

- Principle
  - The information brought by an observation x about θ is entirely contained in the likelihood function $p(x|\theta)$. Moreover, if x_1 and x_2 are observations depending on the same parameter θ, such that there exists a constant c satisfying $p(x_1|\theta) = cp(x_2|\theta)$ for every θ, then they bring the same information about θ and must lead to identical estimators.

The Maximum Likelihood Estimator

- The MLE is an implementation of the likelihood principle
- The maximum likelihood estimator (MLE) is derived by holding x fixed and maximising L over all possible values of θ

$$\hat{\theta}_{MLE}(x) = \underset{\theta}{\operatorname{argmax}} L(\theta|x)$$

- The maximum likelihood estimate is the value of θ for which the associated distribution (among all distributions parameterised by θ) *is most likely* to have generated the data x.
- MLE is a procedure, NOT an optimality criterion!

Computing the MLE

- If the likelihood function is differentiable, then θ is found by differentiating the likelihood (or log-likelihood), equating with zero and solving:

$$\frac{\partial}{\partial\theta}\left(\log\left(l(\theta|x)\right)\right) = 0$$

- If multiple solutions exist, then the MLE is the solution that maximizes $\log\left(l(\theta|x)\right)$, that is, the *global maximizer*.
- In certain cases, such as PDFs with an exponential form, the MLE can be easily solved for. That is, the above equation can be solved using calculus and standard linear algebra.

DC level in WGN (modified example continued)

- Setting the derivative of the log-likelihood function equal to zero yields

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

Solving produces two solutions

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

- Estimator bias

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

$$\neq -\frac{1}{2} + \sqrt{E\left(\frac{1}{N}\sum_{n=0}^{N-1} x^2[n]\right) + \frac{1}{4}} = -\frac{1}{2} + \sqrt{A + A^2 + \frac{1}{4}} = A$$

DC level in WGN (modified example continued)

- However, as N → ∞

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

And therefore $\hat{A} \rightarrow A$ (*consistent* estimator)

- In addition, it can be shown that

$$E(\hat{A}) \rightarrow A$$
$$\text{var}(\hat{A}) \rightarrow CRLB$$

- The MLE is thus *asymptotically unbiased* and *asymptotically efficient*.

DC level in WGN (original example)

- For the received data

$$x[n] = A + w[n],\ \text{where } n = 0,1,\ldots,N-1 \ and \ w[n] \sim N(0,\sigma^2)$$

- The PDF is

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

- Taking the first derivative of the log-likelihood

$$\frac{\partial \ln p\,(x;A)}{\partial A} = \frac{\partial}{\partial A}\left[-\ln\left[(2\pi\sigma^2)^{\frac{N}{2}}\right] - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right]$$

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

DC level in WGN (original example)

- First derivative of the log-likelihood again:

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

- Finally, setting to zero yields

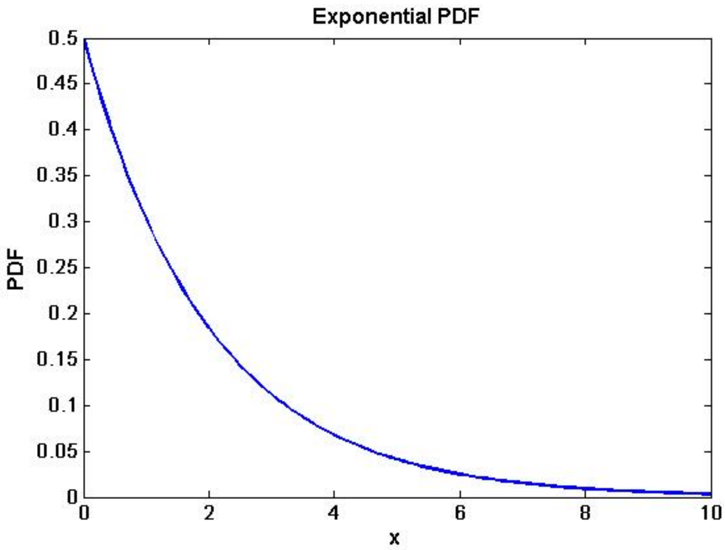
$$\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$$

- *If an efficient estimator exists, the ML procedure will produce it*!!

Example: Exponential distribution

- Assume $x_1, x_2, \ldots, x_N$ is a random sample from an exponential distribution:-

$$f(x|\theta) = \begin{cases} \theta e^{-\theta x}, & x \geq 0 \\ 0 & x < 0 \end{cases}$$



- The likelihood function is

$$L(\theta) = \prod_{i=1}^{N} f(x_i|\theta) = \theta^N e^{-\theta \sum_{i=1}^{N} x_i}$$

$$l(\theta) = N \log \theta - \theta \sum_{i=1}^{N} x_i$$

- Taking the derivative and setting equal to zero:-

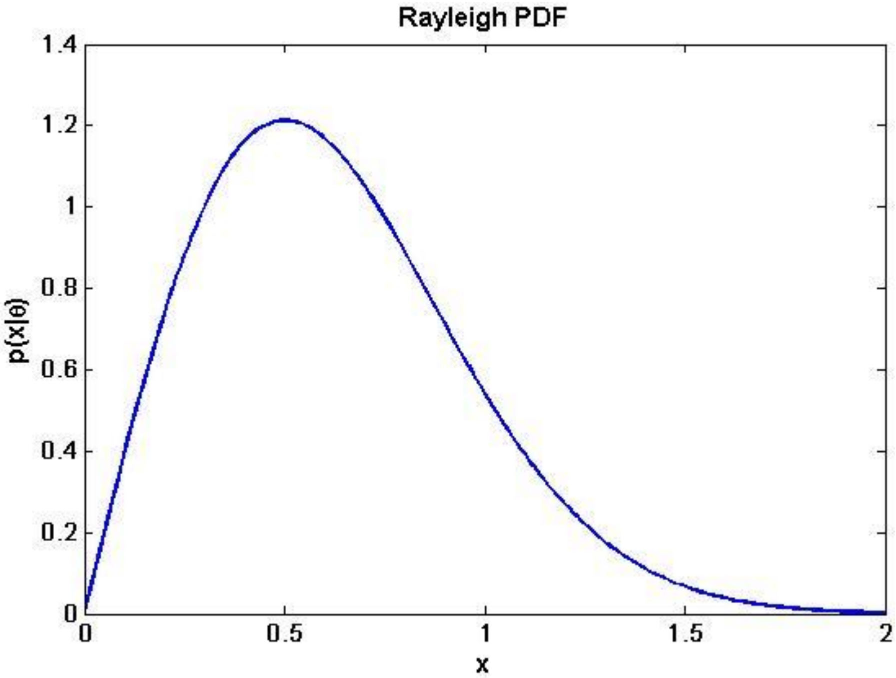
$$\hat{\theta} = \frac{N}{\sum_{i=1}^{N} x_i} = \frac{1}{\bar{x}}$$

Example: Rayleigh distribution

- Assume $x_1, x_2, \dots, x_N$ is a random sample from an exponential distribution:-

$$y = p(x|\theta) = \frac{x}{\theta^2} e^{-\frac{x^2}{2\theta^2}}$$

$$\hat{\theta}_{MLE} = ?$$



Joint MLE for several parameters

- Often in practice, a statistical model has more than one unknown parameter
- If there are k parameters, then we have a vector parameter $\boldsymbol{\theta}=(\theta_1, \theta_2, \ldots, \theta_k)^T$ and the PDF is written as $f(x|\boldsymbol{\theta})$.
- $\widehat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k\right)^T$ are the values in the parameter space that jointly maximise the likelihood function
- If $l(\boldsymbol{\theta})$ is differentiable then the vector estimate satisfy k joint differential equations

$$\frac{\partial}{\partial \theta_j} l(\theta_1, \theta_2, \ldots, \theta_k) = 0 \quad \text{for } j = 1,2,\ldots,k$$

Example: Normal distribution

- Assume $X \sim N(\mu, \sigma^2)$ i.e.

$$p(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$$

$$-\infty < \mu < \infty, \quad \sigma > 0$$

- The log-likelihood is

$$l(\mu,\sigma^2) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\sigma^2 - \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2\sigma^2}$$

- Solving the likelihood equations yields:

$$\frac{\partial l}{\partial \mu} = -\frac{(-2)\sum_{i=1}^{N}(x_i-\mu)}{2\sigma^2} = 0 \Rightarrow \hat{\mu} = \bar{x}$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{\sigma^4}\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2} = 0 \Rightarrow \hat{\sigma}^2 = \frac{\sum_i (x_i-\bar{x})^2}{N}$$

Numerical methods for finding the MLE

- Sometimes there is no explicit solution of the likelihood equation $dl/d\theta=0$. The MLE approach is still applicable but one needs to resort to optimization techniques in order to find the estimate numerically.
- Standard methods:-
  - Newton-Raphson
  - Iteration by Scoring Method
  - Expectation-Maximization Algorithm
- All are based on posing some guess at the MLE and then incrementally updating that guess
- Only the latter is guaranteed to converge to a local (!) maximum.

Summary of MLE

- If a MVU estimator does not exist, or can not be found, the parameters can be obtained from the likelihood function.
- A Maximum Likelihood (ML) parameter estimate is found by maximising the likelihood function $p(x;\theta)$ which is essentially the probability of the data given the parameters;
- ML estimators are asymptotically efficient, as the number of observations increase and the covariance of the estimates tends to CRLB
- A major advantage of the MLE is that we can find an estimate from the given data numerically since it requires only the maximum of a known function. The Newton-Raphson iterative techniques or the Expectation-Maximisation (EM) algorithm can be used for iterative estimation of the parameters.