Advanced Control Systems
Detection, Estimation, and Filtering

Graduate Course on the
MEng PhD Program
Spring 2012/2013

Chapter 6
Maximum Likelihood Estimation

Instructor:
Prof. Paulo Jorge Oliveira
p.oliveira@dem.ist.utl.pt or pjcro @ isr.ist.utl.pt
Phone: 21 8419511 or 21 8418053 (3511 or 2053 inside IST)

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

Syllabus:

Classical Estimation Theory

…

Chap. 5 - ***Best Linear Unbiased Estimators*** [1 week]

Definition of BLUE estimators; White Gaussian noise and bandlimited systems; Examples; Generalized MVU estimation;

Chap. 6 - ***Maximum Likelihood Estimation*** [1 week]

The maximum likelihood estimator; Properties of the ML estimators; Solution for ML estimation; Examples; Monte-Carlo methods;

Chap. 7 - ***Least Squares*** [1 week]

The least squares approach; Linear and nonlinear least squares; Geometric interpretation; Constrained least squares; Examples;

continues…

Motivating example:

Example (DC level in white Gaussian noise modified):

For this example the methods previously introduced will not work…

Signal model: $x[n] = A + w[n], \qquad n = 0,\ldots,N-1$

Where A is the unknown level to be estimated and $w[n]$ is zero mean white Gaussian with unknown variance A.

First, lets try to find the CRLB. The PDF is:

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

The derivative of the log-likelihood function is

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

$$\overset{?}{=} I(A)(g(\mathbf{x}) - A)$$

It appears that it is not possible…
So, an efficient estimator does not exist.

An example:

Example (DC level in white Gaussian noise modified) (cont):

However, from the second derivative, it is possible to compute the CRLB to be

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

Secondly, to find the MVU estimator based on the theory of sufficient statistics, one must factorize (1) in the form

$$p\left(\mathbf{x};\theta\right) = g\left(T\left(\mathbf{x}\right),\theta\right)h\left(\mathbf{x}\right)$$

It is possible, if one considers

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

DEM DEPARTAMENTO DE ENGENHARIA MECÂNICA TÉCNICO LISBOA

An example:

Example (DC level in white Gaussian noise modified) (cont):

So a sufficient statistics is

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

It is required to find a function of the sufficient statistics that produces an unbiased estimator, i.e.

$$E\left[g\left(\sum_{n=0}^{N-1} x^2[n]\right)\right] = A$$

Taking into account the auxiliary result

$$\operatorname{var}(x[n]) = E\left[\left(x[n] - E(x[n])\right)^2\right] = E\left[x^2[n]\right] - 2E(x[n])E(x[n]) + E^2(x[n])$$

We have that

$$E\left[x^2[n]\right] = \operatorname{var}(x[n]) + E^2(x[n]) \qquad \text{(in our case } E\left[x^2[n]\right] = A + A^2\text{)}$$

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

An example:

Example (DC level in white Gaussian noise modified) (cont):

Since

$$E\left[\sum_{n=0}^{N-1} x^2[n]\right] = NE\left[\sum_{n=0}^{N-1} x^2[n]\right] = N\left[\mathrm{var}\left(x\left[n\right]\right) + E^2\left(x\left[n\right]\right)\right] = N\left[A + A^2\right]!$$

It is impossible to find a solution for a generic unknown parameter A, i.e.

$$N\left[A + A^2\right] \neq A!$$

A final alternative is to find the optimal estimator would be to determine

$$E\left[\hat{A} \mid \sum_{n=0}^{N-1} x^2[n]\right] = ???$$

That appears to be a formidable task!

We exhausted the optimal approaches studied… We can propose other estimators, but without any guarantee of optimality.

An example:

Example (DC level in white Gaussian noise modified) (cont):

Those estimators should be at least *approximately optimal,* i.e.

$$E\left[\hat{A}\right] \rightarrow A$$
$$\operatorname{var}\left(\hat{A}\right) \rightarrow \mathrm{CRLB}$$

For instance, lets consider the estimator (why? explanation will be provided next…)

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

This estimator is biased, since

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

$$= -\frac{1}{2} + \sqrt{A + A^2 + \frac{1}{4}} = A!$$

But it can be verified that is is consistent, i.e.

$$\frac{1}{N}\sum_{n=0}^{N-1} x^2[n] \rightarrow E\left[x^2[n]\right] = A + A^2 \qquad \text{and therefore} \qquad \hat{A} \rightarrow A$$

DEM DEPARTAMENTO DE ENGENHARIA MECÂNICA TÉCNICO LISBOA

An example:

Example (DC level in white Gaussian noise modified) (cont):

Consider that $\hat{A} = g(u)$, where $g(u) = -\frac{1}{2} + \sqrt{u + \frac{1}{4}}$ and lets linearise this function, near $u_0 = E[u] = A + A^2$.

$$g(u) \approx g(u_0) + \left.\frac{dg(u)}{du}\right|_{u=u_0} (u - u_0) \qquad \text{(using Taylor's series expansion)}$$

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

$$E\left[\hat{A}\right] \approx A.$$

Thus this estimator is asymptotically unbiased.

And what about its variance?…

An example:

Example (DC level in white Gaussian noise modified) (cont):

It is given by

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

But $var(x^2[n])=4A^3+2A^2$, so that

$$\mathrm{var}\left(\hat{A}\right) \approx \frac{\frac{1}{4}}{N\left(A+\frac{1}{2}\right)^2} 4A^2\left(A+\frac{1}{2}\right) \approx \frac{A^2}{N\left(A+\frac{1}{2}\right)}$$

Thus this estimator asymptotically equals the CRLB!!!

Discuss the impact of one such methodology that provides asymptotic results.
The value for science and for engineering

An asymptotically optimal solution:

What to do, if the MVU estimator does not exist or can not be found?

An alternative consists of exploiting the…

Maximum Likelihood Principle.

It can be understood as a “turn the crank” method.

Only suboptimal performance can be achieved.

It is the most popular approach to obtaining practical estimators.

Its optimality is verified for large enough data sets.

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

Motivating example revisited:

Example (DC level in white Gaussian noise modified):

The method consists only on the computation of the maximum of the (log) likelihood function. In our case, it is required to solve:

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

$$= -\frac{N}{2A}+\frac{1}{A}\sum_{n=0}^{N-1}x[n]-\frac{1}{A}NA+\frac{1}{2A^2}\sum_{n=0}^{N-1}\left(x^2[n]-2Ax[n]+A^2\right)=$$

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

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

From where our previous unexplained estimator results

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

Maximum Likelihood Principle:

Theorem 7.1 (Asymptotic Properties of the MLE) *– If the PDF $p(\mathbf{x};\theta)$ of the data* ***x*** *satisfies some regularity conditions, then the MLE of the unknown parameter is asymptotically distributed (for large data records) according to*

$$\hat{\theta} \overset{a}{\sim} N\left(\theta, \boldsymbol{I}^{-1}\left(\theta\right)\right)$$

where $I(\theta)$ is the Fisher information evaluated at the true value of the unknown parameter

In practice it is seldom known in advance how large N must be.

Analytical expression for the PDF of the MLE is usually impossible to derive.

Thus, to assess the MLE performance, computer simulations are usual.

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

Properties of MLE:

Proof outline:

The following regularity conditions are assumed:

1) *The first and second-order derivative of the log-likelihood are well defined.*

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

First, it is required to show that the MLE is consistent. Related with the Kullbak_Leibner information (and also with measure of the difference between two probability distributions)

$$\int \ln\left[\frac{p\left(x\left[n\right];\theta_1\right)}{p\left(x\left[n\right];\theta_2\right)}\right]p\left(x\left[n\right];\theta_1\right)dx[n]\geq 0 \qquad (1)$$

Where equality occurs for $\theta_1=\theta_2$.

Properties of MLE:

Proof outline:

Now, maximizing the log-likelihood

$$\frac{1}{N}\ln p(\mathbf{x};\theta)=\frac{1}{N}\sum_{n=0}^{N-1}\ln p(x[n];\theta)\rightarrow\int \ln p(x[n];\theta)\,p(x[n];\theta_0)\,dx[n]$$

Where the last relation is due to the fact that, by the law of large numbers, it converges to the expected value. The MLE is *consistent* *and is maximized for* $\hat{\theta}=\theta_0$*, i.e.*

$$\int \ln p(x[n];\theta_0)\,p(x[n];\theta_0)\,dx[n]\geq\int \ln p(x[n];\theta_1)\,p(x[n];\theta_0)\,dx[n]$$

Moreover is the maximum, due to suitable continuity argument and the relation (1). Using the Taylor series expansion, one obtains

$$\left.\frac{\partial \ln p(\mathbf{x};\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}}\approx\left.\frac{\partial \ln p(\mathbf{x};\theta)}{\partial\theta}\right|_{\theta=\theta_0}+\left.\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta^2}\right|_{\theta=\theta_0}\left(\hat{\theta}-\theta_0\right)\approx 0$$

Where the last quantity is approx. 0 if near an maximum.

Properties of MLE:

Proof outline:

This relation can therefore be approximately written as

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right)=\frac{\left.\frac{1}{\sqrt{N}}\frac{\partial \ln p\left(\mathbf{x};\theta\right)}{\partial\theta}\right|_{\theta=\theta_0}}{\left.-\frac{1}{N}\frac{\partial^2 \ln p\left(\mathbf{x};\theta\right)}{\partial\theta^2}\right|_{\theta=\hat{\theta}}}\rightarrow\frac{\left.\frac{1}{\sqrt{N}}\sum_{n=0}^{N-1}\frac{\partial \ln p\left(x\left[n\right];\theta\right)}{\partial\theta}\right|_{\theta=\theta_0}}{\left.-\frac{1}{N}\sum_{n=0}^{N-1}\frac{\partial^2 \ln p\left(x\left[n\right];\theta\right)}{\partial\theta^2}\right|_{\theta=\hat{\theta}}}\sim N\left(0,i^{-1}\left(\theta_0\right)\right)$$

From where it can be concluded, using the law of large numbers and the IID of the samples, that

$$\hat{\boldsymbol{\theta}}\overset{a}{\sim}N\left(\boldsymbol{\theta}_0,\boldsymbol{I}^{-1}\left(\boldsymbol{\theta}_0\right)\right)$$

MLE PDF:

In general is very difficult (or impossible) to obtain the PDF of the MLE.

How to study its performance?

Use Monte Carlo Method

1 . Simulate the noise characteristics, the signal model, and compute the estimates.

2. Repeat M times these realizations. (How to select M?)

3. Compute the experiments ensemble mean and covariance, using

$$\widehat{E\left[\hat{A}\right]} = \frac{1}{M}\sum_{i=1}^{M}\hat{A}_i$$

$$\widehat{\operatorname{var}\left(\hat{A}\right)} = \frac{1}{M}\sum_{i=1}^{M}\left(\hat{A}_i - \widehat{E\left[\hat{A}\right]}\right)^2$$

Invariance Property:

Theorem 7.2 (Invariance Property of the MLE) *– The MLE of the parameter $\alpha=g(\theta)$, where the PDF $p(\mathbf{x};\theta)$ is parameterized by θ, is given by*

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

Where $\hat{\theta}$ is the MLE of θ. The MLE is obtained by maximization of $p(\mathbf{x};\theta)$, If g is not a one-to-one function, then $\hat{\alpha}$ maximized the modified likelihood fuction

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

Proof outline (simple case: g() one to one WGN, IID, expected value):

The MLE for the transformed parameter can be found minimizing the log-likelihood, i.e.

$$\frac{\partial}{\partial\alpha}\sum_{n=0}^{N-1}\left(x[n]-g^{-1}\left(\alpha\right)\right)^2 = k + k'\sum_{n=0}^{N-1}\left(x[n]-g^{-1}\left(\alpha\right)\right)\frac{\partial}{\partial\alpha}g^{-1}\left(\alpha\right)=0, \qquad k,k'>0 \quad .$$

Thus

$$\sum_{n=0}^{N-1}x[n]-Ng^{-1}\left(\alpha\right)=0, \qquad g^{-1}\left(\alpha\right)=\frac{1}{N}\sum_{n=0}^{N-1}x[n]=\bar{x} \qquad \alpha=g\left(\bar{x}\right).$$

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

Numerical Determination of the MLE:

The MLE in general can not be found in close form.

But it can be found numerically. Grid search, gradient or Newton methods can be used.

Conditions for nonlinear optimization methods are central to that discussion.

For different data-sets, the target function changes and thus also the maximum changes.

In general there is not or maximum, but a number of local maxima.

How to avoid attraction to local maxima? Regions of attraction?...

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

Motivating example:

Example (Exponential in white Gaussian noise):

Signal model: $x\left[n\right] = r^n + w\left[n\right], \qquad n = 0,\ldots,N-1$

Where $w[n]$ is zero mean white Gaussian noise with variance σ^2 and the exponential factor r is to estimated.

For the likelihood function, the MLE is the value of r that maximizes is :

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

Or, equivalently, the value that minimizes

$$J\left(r\right) = \sum_{n=0}^{N-1}\left(x[n]-r^n\right)^2.$$

Differentiating $J(r)$ and setting to zero produces

$$\frac{\partial J\left(r\right)}{\partial r} = 2\sum_{n=0}^{N-1}\left(x[n]-r^n\right)nr^{n-1}.$$

It is a nonlinear equation in r and cannot be solved directly.

Numerical Solution (basics):

The use of iterative methods to maximize the log-likelihood function is an example of application of nonlinear optimization methods. See a good book (or class) on the field…

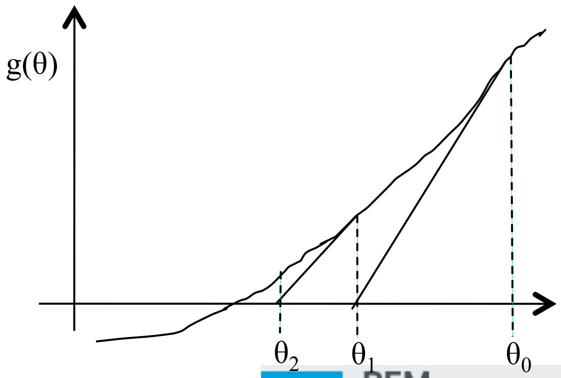
$$\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta} = g(\theta) = 0$$

For instance, one of the most basic method, is the Newton-Raphson method. From an initial guess Θ_0, and from a Taylor series expansion results

$$g(\theta) \approx g(\theta_0) + \left.\frac{\partial g(\theta)}{\partial \theta}\right|_{\theta=\theta_0} (\theta - \theta_0) \approx 0$$

The following recursion results$\boldsymbol{\theta}$

$$\theta_{k+1} = \theta_k - \left[\left.\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right|_{\theta=\theta_k} \right]^{-1} \left.\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta}\right|_{\theta=\theta_k}$$



DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

Motivating example:

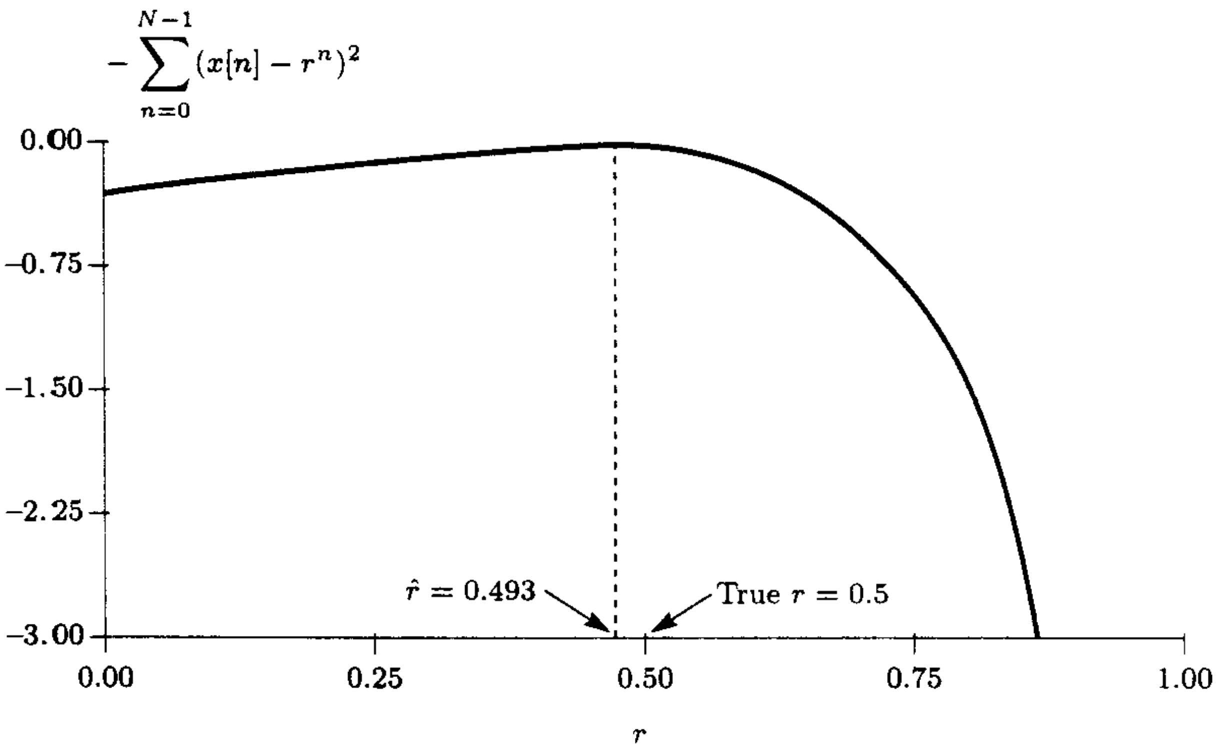
Example (Exponential in white Gaussian noise):

Computer simulation

N=50, r=0.5, and σ^2=0.01

Maximum at r=0.493
(a specific realization)

Iteration	Initial Guess, r_0		
	0.8	0.2	1.2
1	0.723	0.799	1.187
2	0.638	0.722	1.174
3	0.561	0.637	1.161
4	0.510	0.560	1.148
5	0.494	0.510	1.136
6	0.493	0.494	1.123
7		0.493	1.111
8			1.098
9			1.086
10			1.074
⋮			⋮
29			0.493



DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

Numerical Solution (basics):

The importance of

stability conditions,

convergence rates, and

domains of attraction

can hardly be overemphasized. Engineering/scientific content…

Other methods mentioned:

Scoring

Expectation / maximization (nice term paper subject)

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

Invariance Property:

Theorem 7.5 (Optimality of the MLE for the Linear Model) *– If the observed data $\mathbf{x}$ are described by the general linear model*

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

where H is a known N x p matrix with N>p and of rank p, θ is a p x 1 parameter vector to be estimated, and w is the noise vector with PDF N(0,C), the the MLE of θ is

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

And is also an efficient estimator in that it attains the CRLB and hence is the MVU estimator. The PDF of θ is

$$\hat{\boldsymbol{\theta}} \sim \boldsymbol{N}\left(\boldsymbol{\theta}, \left(\mathbf{H}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{H}\right)^{-1}\right).$$

Method of Scoring:

The method of scoring is based on the approximation for one element found also in the Newton-Raphson method. Note that for IID samples we have

$$\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta^2} = \sum_{n=0}^{N-1} \frac{\partial^2 \ln p(x[n];\theta)}{\partial\theta^2} = NE\left[\frac{\partial^2 \ln p(x[n];\theta)}{\partial\theta^2}\right] = -Ni(\theta) = -I(\theta).$$

So the iterations on NR method can be transformed in

$$\theta_{k+1} = \theta_k - I^{-1}(\theta)\left.\frac{\partial \ln p(\mathbf{x};\theta)}{\partial\theta}\right|_{\theta=\theta_k}$$

Resulting in a method that is more stable. However it suffers from the same convergence problems as the NR method.

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

Maximum Likelihood Principle:

Theorem 7.1 (Asymptotic Properties of the MLE – Vector Parameter) *– If the PDF $p(\boldsymbol{x};\boldsymbol{\theta})$ of the data **x** satisfies some "regularity" conditions, then the MLE of the unknown parameter **θ** is asymptotically distributed (for large data records) according to*

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} N\left(\boldsymbol{\theta}, \boldsymbol{I}^{-1}\left(\boldsymbol{\theta}\right)\right)$$

where $I(\boldsymbol{\theta})$ is the Fisher information evaluated at the true value of the unknown parameter

In practice it is seldom known in advance how large N must be.

In the cases where the number of parameters increases, relative to the number of samples available, the assumptions fails and the MLE estimator can provide very poor estimates.

Bibliography:

Further reading

• Harry L. Van Trees, ***Detection, Estimation, and Modulation Theory, Parts I to IV,*** John Wiley, 2001.

• Anthony William Fairbank Edwards, **Likelihood,** Cambridge University Press, 1972.

• Jerry M. Mendel, **Lessons in Digital Estimation Theory,** Prentice Hall, 1987.

• Geoffrey J. McLachlan (Author), Thriyambakam Krishnan, **The EM Algorithm and Extensions,** Wiley, 1997.