# Advanced Control Systems Detection, Estimation, and Filtering

*Graduate Course on the*
*MEng PhD Program*
*Spring 2012/2013*

*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

# Objectives:

- Motivation for estimation, detection, filtering, and identification in stochastic signal processing

- Methodologies on design and synthesis of optimal estimation algorithms

- Characterization of estimators and tools to study their performance

- To provide an overview in all principal estimation approaches and the rationale for choosing a particular technique

**Both for parameter and state estimation,**

**always on the presence of stochastic disturbances**

In RADAR (Radio Detection and Ranging), SONAR (sound navigation and ranging), speech, image, sensor networks, geo-physical sciences,…

# Pre-requisites:

- Random Variables and Stochastic Processes

  Joint, marginal, and conditional probability density functions: Gaussian / normal distributions; Moments of random variables (mean and variance); Wide-sense stationary processes; Correlation and covariance; Power spectral density;

- Linear Algebra

  Vectors: orthogonality, linear independence, inner product; norms;

  Matrices: eigenvectors, rank, inverse, and pseudo-inverse;

- Linear Systems

  LTIS and LTVs; ODEs and solutions; Response of linear systems; Transition matrix; Observability and controlability; Lyapunov stability.

The implementation of solutions for problems require the use of ***MATLAB*** and ***Simulink***.

# Syllabus:

## Classical Estimation Theory

Chap. 1 - ***Motivation for Estimation in Stochastic Signal Processing*** [1/2 week]

Motivating examples of signals and systems in detection and estimation problems;

Chap. 2 - ***Minimum Variance Unbiased Estimation*** [1 week]

Unbiased estimators; Minimum Variance Criterion; Extension to vector parameters; Efficiency of estimators;

Chap. 3 - ***Cramer-Rao Lower Bound*** [1 week]
Estimator accuracy; Cramer-Rao lower bound (CRLB); CRLB for signals in white Gaussian noise; Examples;

continues…

# Syllabus (cont.):

Chap. 4 - ***Linear Models in the Presence of Stochastic Signals*** [1/2 week]

Stationary and transient analysis; White Gaussian noise and linear systems; Examples; Sufficient Statistics; Relation with MVU Estimators;

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

Definition of BLUE estimators; White Gaussian noise and bandlimited systems; Examples; Generalized minimum variance unbiased 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…

# Syllabus (cont.):

## Bayesian Estimation Theory

Chap. 8 – ***Bayesian Estimation*** [1/2 week]

Philosophy and estimator design; Prior knowledge; Bayesian linear model; Bayesian estimation on the presence of Gaussian pdfs; Minimum Mean Square Estimators;

Chap. 9 – ***Wiener Filtering*** [1/2 week]

The Wiener filter problem; Causal and non-causal solutions; Complementary filters;

Chap. 10 – ***Kalman Filtering*** [2 weeks]

Optimal estimator  in the presence of white Gaussian noise – the Kalman filter; Stability, convergence and robustness for LTV and LTI systems; Kalman and Wiener filters; Optimal smoothers; Examples; Extended Kalman Filters;

continues…

# Syllabus (cont.):

## Advanced Estimation Topics

Chap. 11 – ***Multiple Model Adaptive Estimation*** [1 week]

Joint system identification and parameter/state estimation using multiple models.

Chap. 12 – ***Optimal Smoothing*** [1 week]

Fixed point, fixed interval, and fixed lag smoothers.

Chap. 13 – ***Advanced Topics*** [2 weeks]

To be detailed later, e.g. Positioning and navigation systems; Failure detection and isolation; Multiple model adaptive estimation; Discretization; Missing data estimation; Outlier detection and removal; Feature based estimation; Principal component analysis; Nonlinear signal processing; Compressive sensing;…

End.

# Grading:

- Five problem sets (50%)

  Due dates: weeks of 25-03, 08-04, 22-04, 06-05, 06-20 and 03-06 (tentative).

and

- Term paper (50%)

  Topic selected randomly by the student. Worked jointly by the faculty/student.
  To be completed in the final 3-4 weeks, i.e.  week of 5-07.

or

- Final exam (50%)

  Week of 15-07.

## Classes:

| | |
|---|---|
| Tuesdays: | 16h00 – 17h30, room C11 |
| Thurdays: | 16h00 – 17h30, room P9 |

***To discuss issues: i) e-mail, ii) phone, or iii) schedule an interview.***

# Bibliography:

## Main references

- Steven M. Kay, ***Fundamentals of Statistical Signal Processing: Estimation Theory,*** *Vol. I*, Prentice Hall Signal Processing Series, 1993.
- A. Gelb, ***Applied Optimal Estimation,*** MIT Press, 1974.

***Complementary reading***

- Steven M. Kay, ***Fundamentals of Statistical Signal Processing: Detection Theory, Vol. II,*** Prentice Hall Signal Processing Series, 1998.
- Harry L. Van Trees, ***Detection, Estimation, and Modulation Theory, Parts I to IV,*** John Wiley, 2001.
- Athanasios Papoulis and S. Unnikrishna Pillai, ***Probability, Random Variables and Stochastic Processes,*** McGraw Hill, 2001.
- Robert Brown and Patrick Hwang, ***Introduction to Random Signals and Applied Kalman Filtering,*** John Wiley, 1997.
- Gonzalo Arce, ***Nonlinear Signal Processing: A Statistical Approach,*** John Wiley, 2005.

# *Advanced Control Systems Detection, Estimation, and Filtering*

***Graduate Course on the MEng PhD Program Spring 2012/2013***

## Chapter 1
## Motivation

***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)***

# Summary:

- Motivation for estimation, detection, filtering, and identification in stochastic signal processing

- Methodologies on how to design optimal estimation algorithms

- Characterization of estimators and tools to study their performance

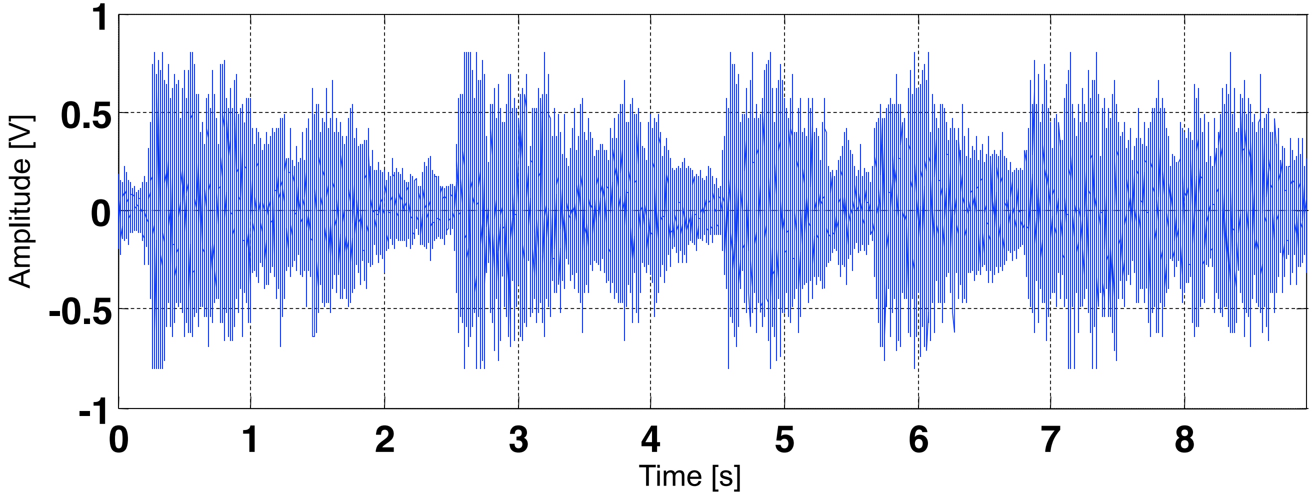
- To provide an overview in all principal estimation approaches and the rationale for choosing a particular technique

**Both for parameter and state estimation,**

**always on the presence of stochastic disturbances**

In RADAR (Radio Detection and Ranging), SONAR(sound navigation and ranging), speech, image, sensor networks, geo-physical sciences,…

# Speech



Signals can be represented by functions (continuous time) or by vectors (where a sampling operation takes place)

Examples of speech/sound processing:

Automatic systems commanded by voice;

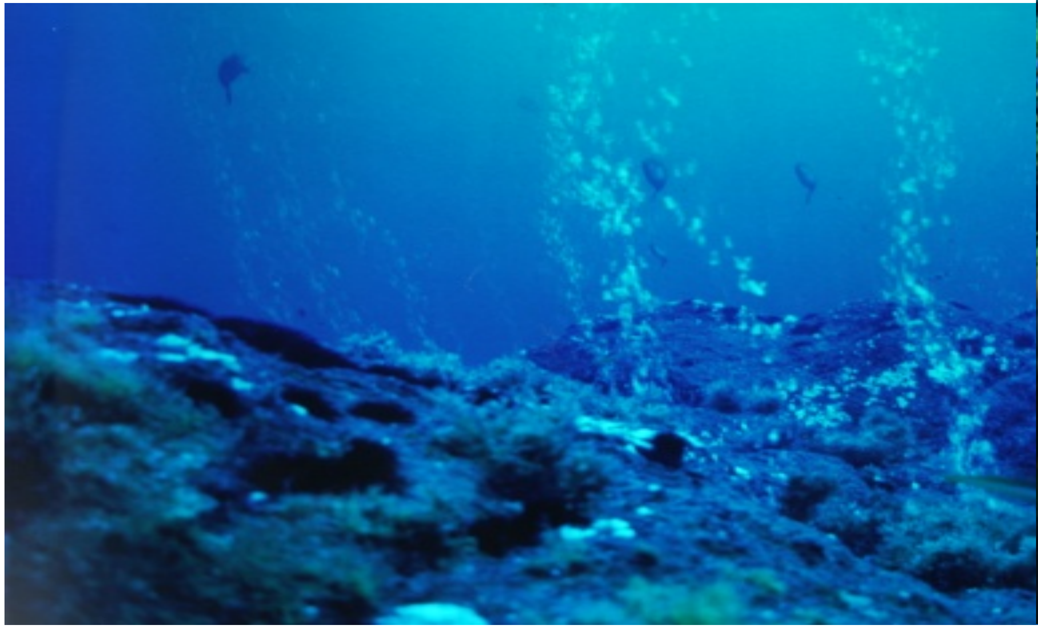
Automatic translation; Voice recognition

Synthesis of voice

# Echograms

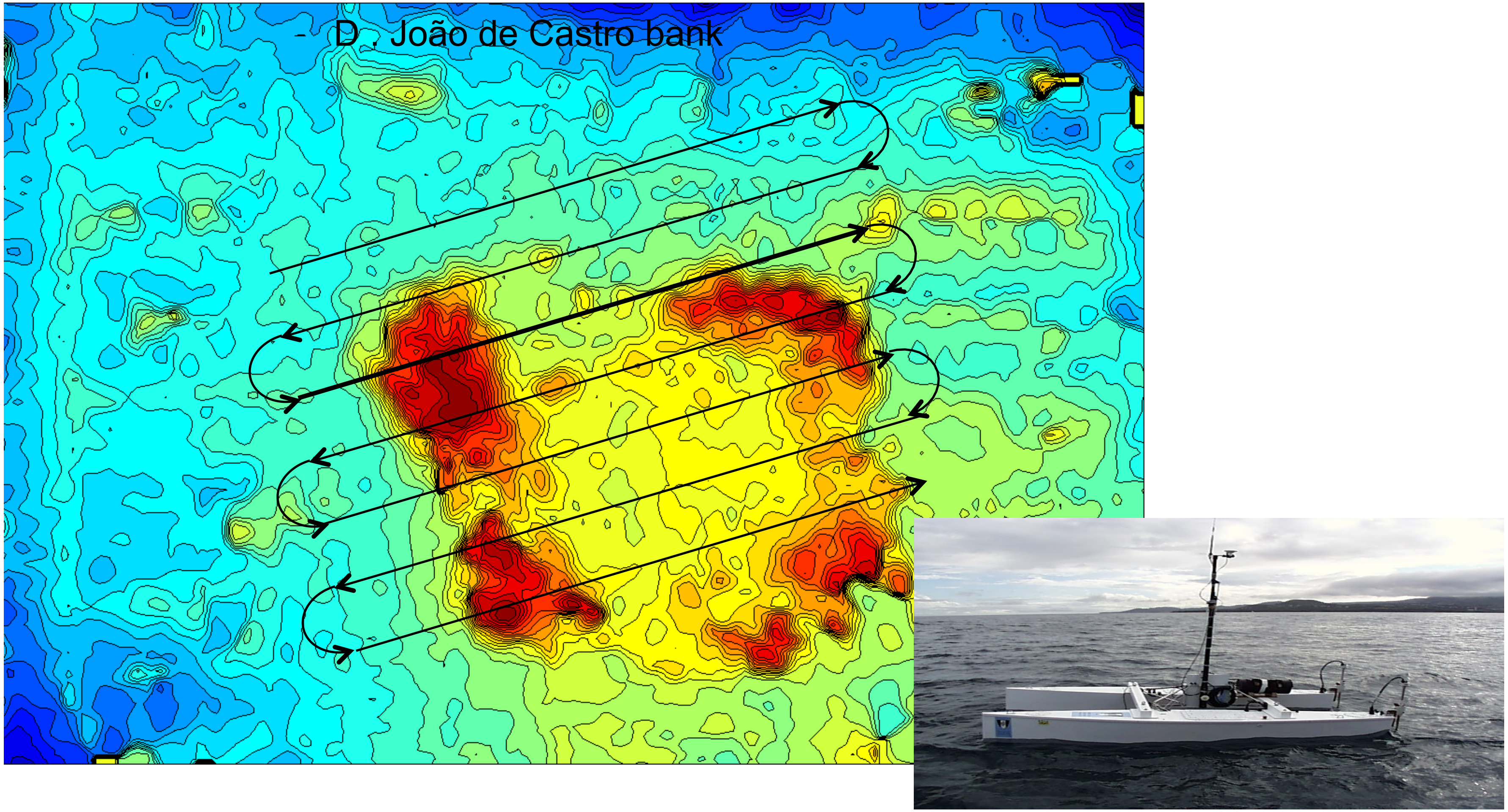
***The quest for hydrothermal vents***

D. João de Castro bank



# Echograms

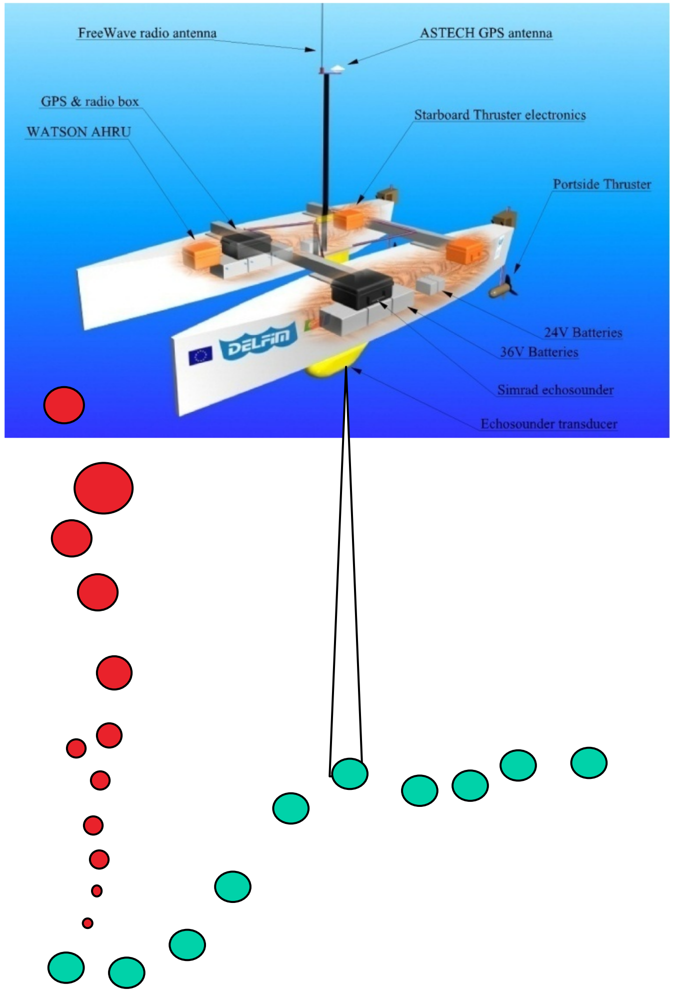
***The quest for hydrothermal vents (cont.)***

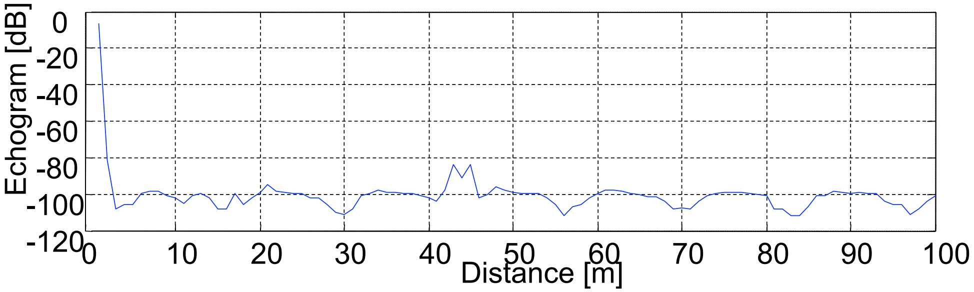


DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

# Echograms

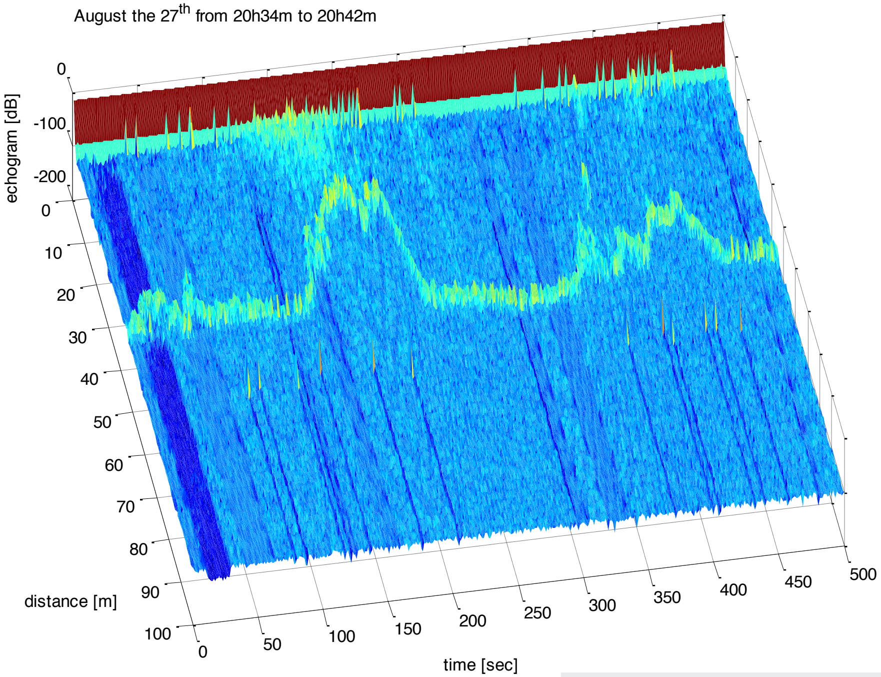
## *The quest for hydrothermal vents (cont.)*





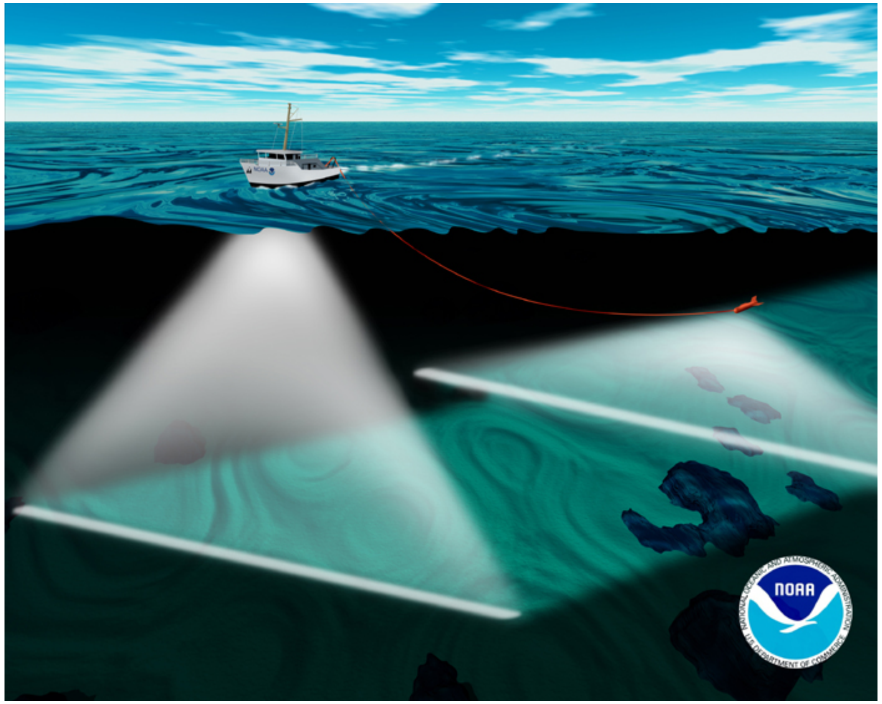
Delfim at D. João de Castro Bank

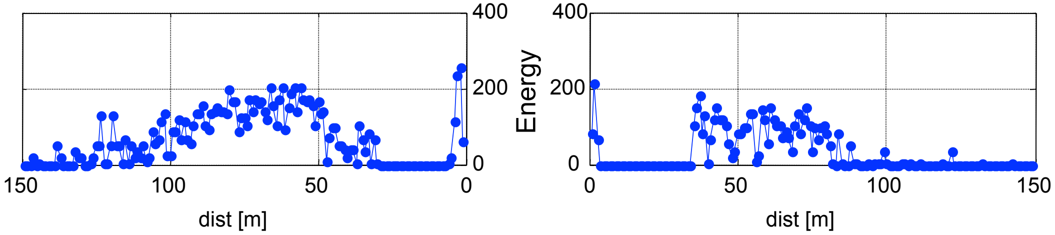
August the 27$^{th}$ from 20h34m to 20h42m

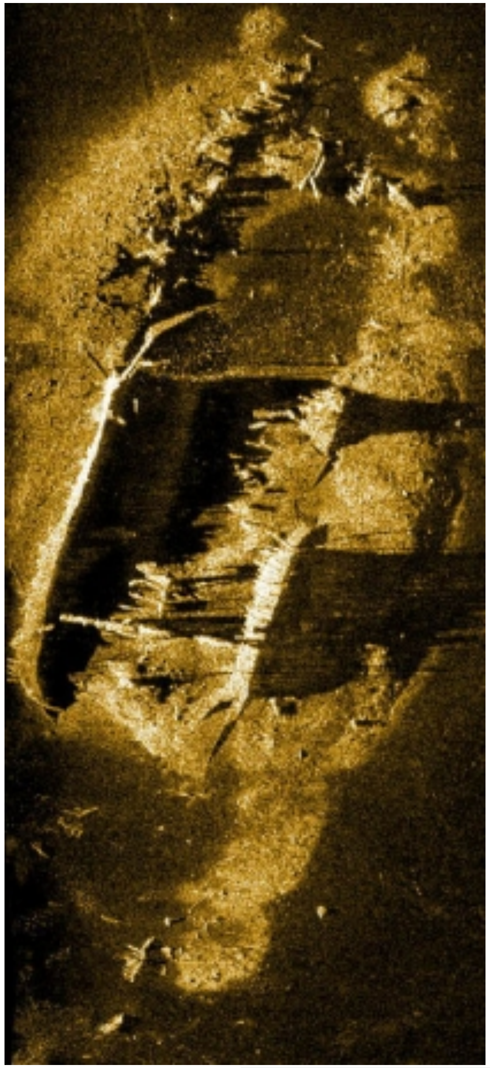


DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

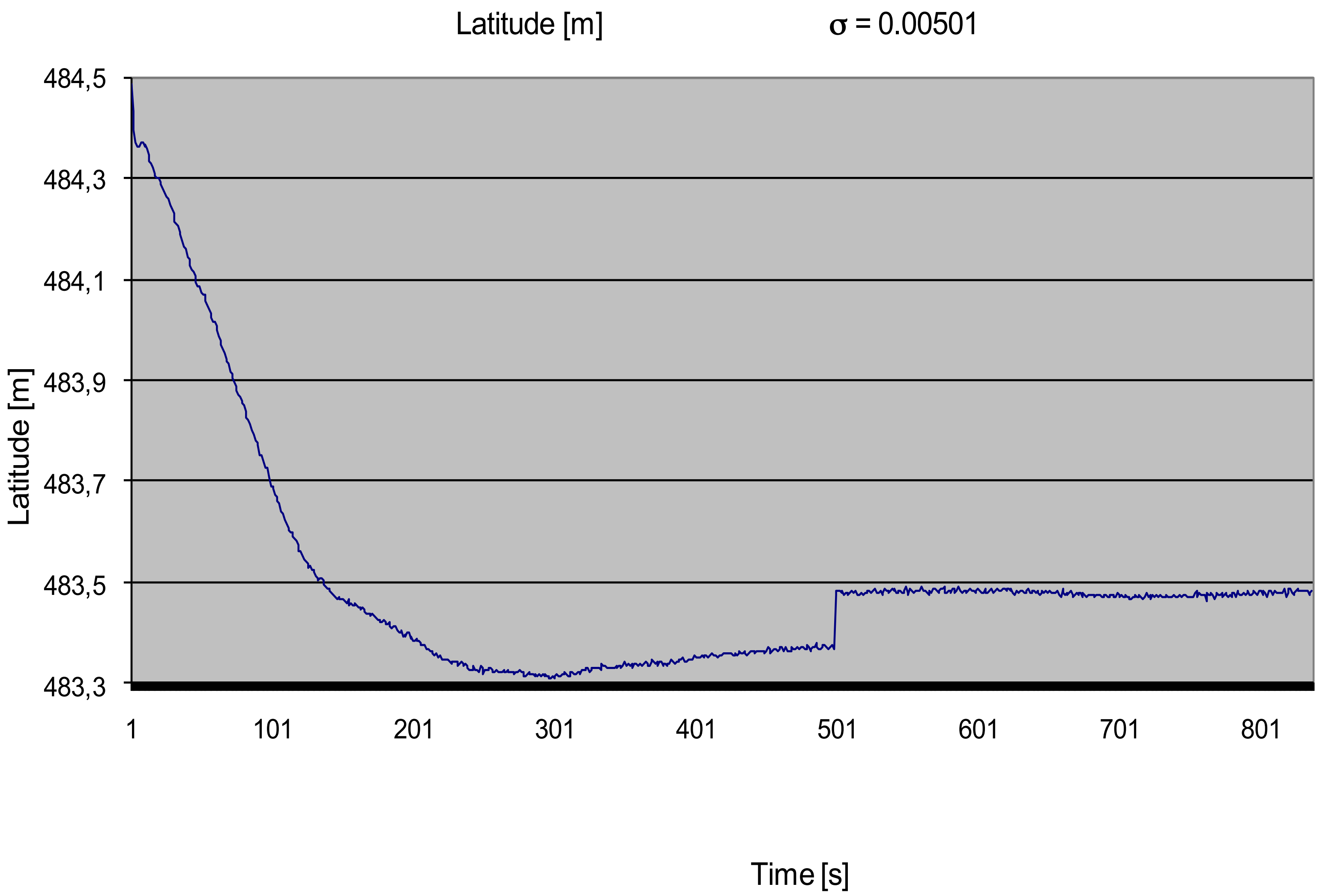
# Sidescan Sonar Imaging





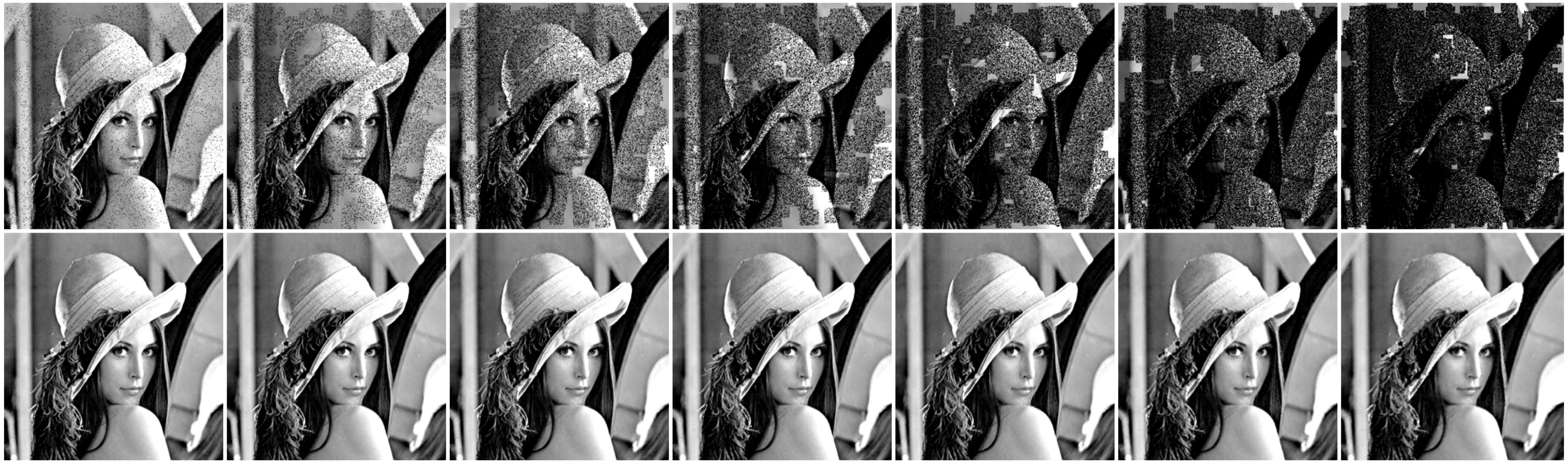


# GPS – Global Positioning System

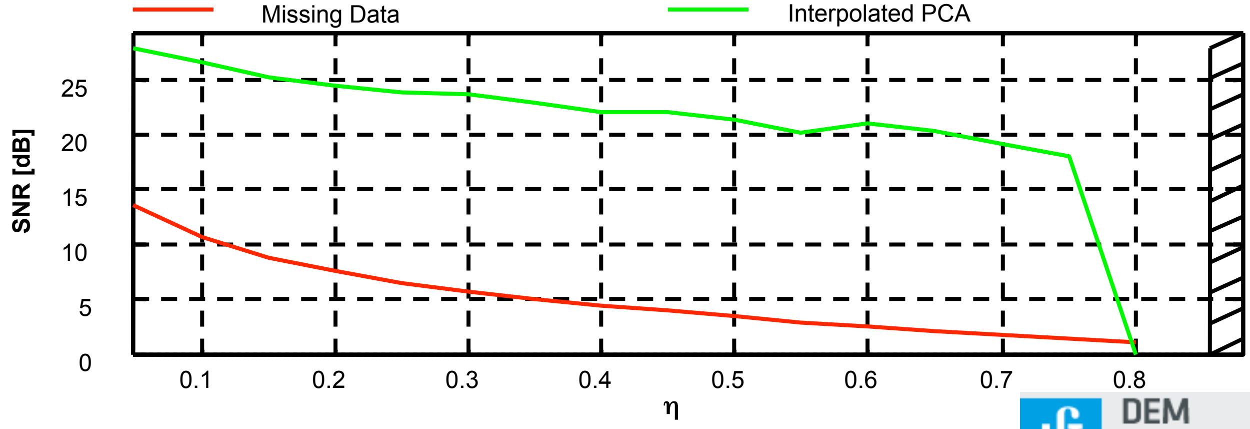


# *Image with missing data*

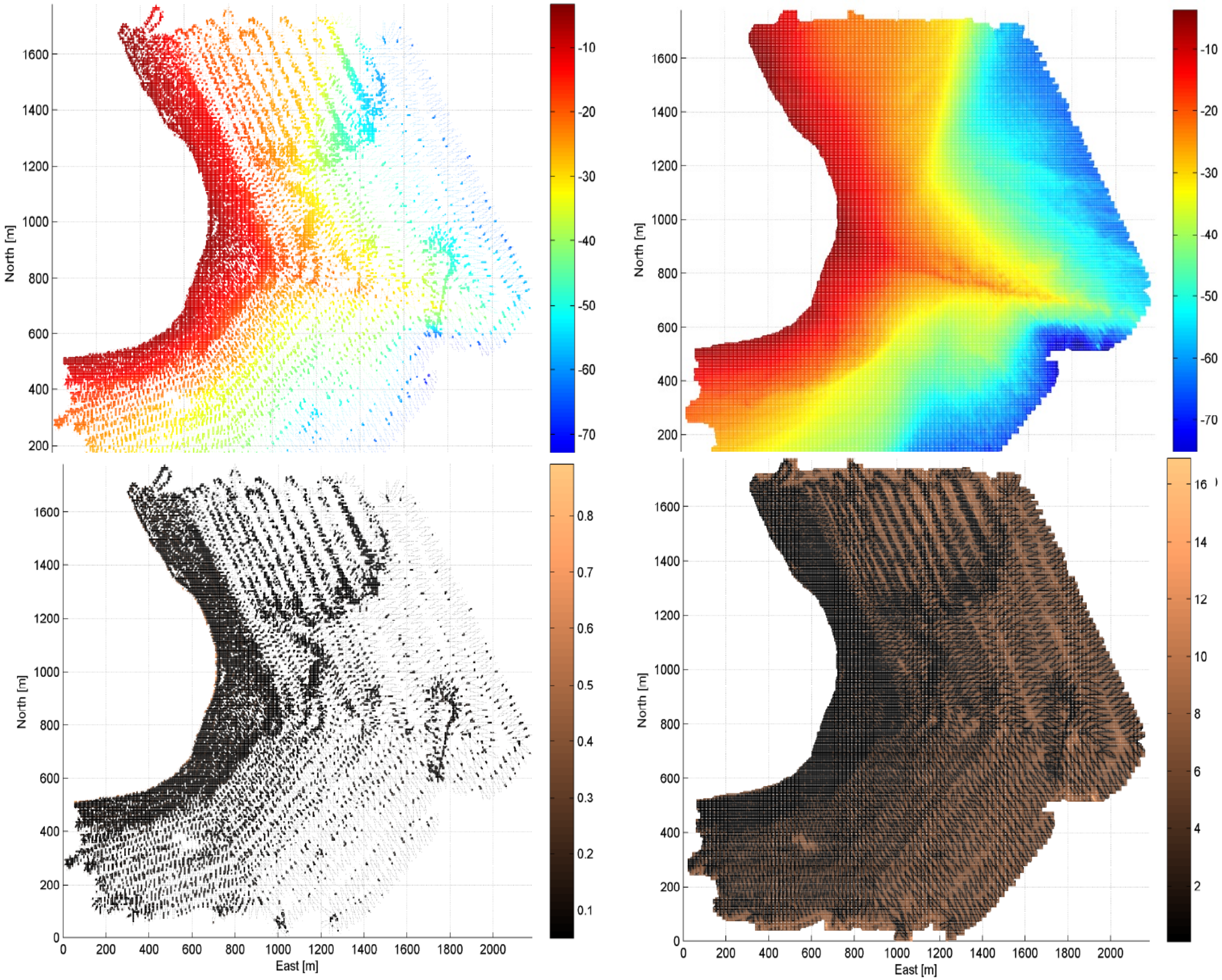
### *Reconstruction*



$\eta$=0.1 $\eta$=0.2 $\eta$=0.3 $\eta$=0.4 $\eta$=0.5 $\eta$=0.6 $\eta$=0.7



# Bathymetric survey



Geo-referenced data

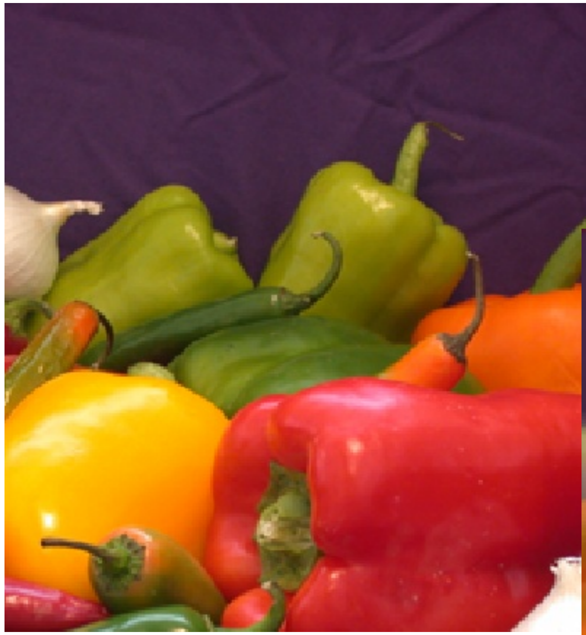
Reconstructed

Uncertainty on data

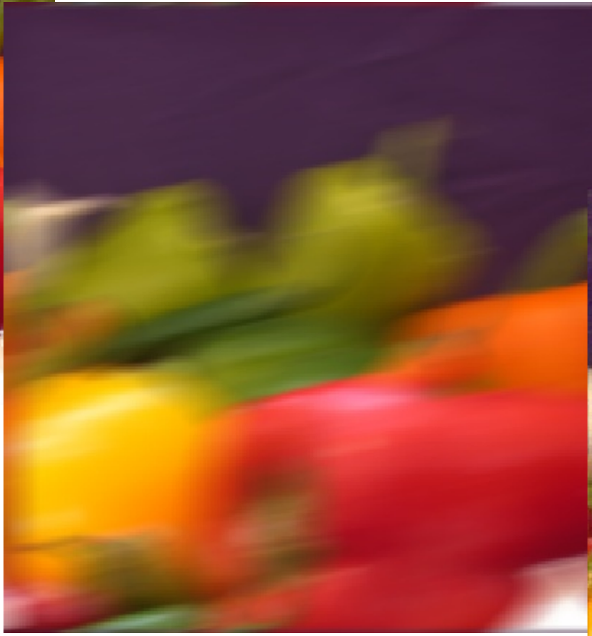
Final uncertainty

# *Deblurring an image*

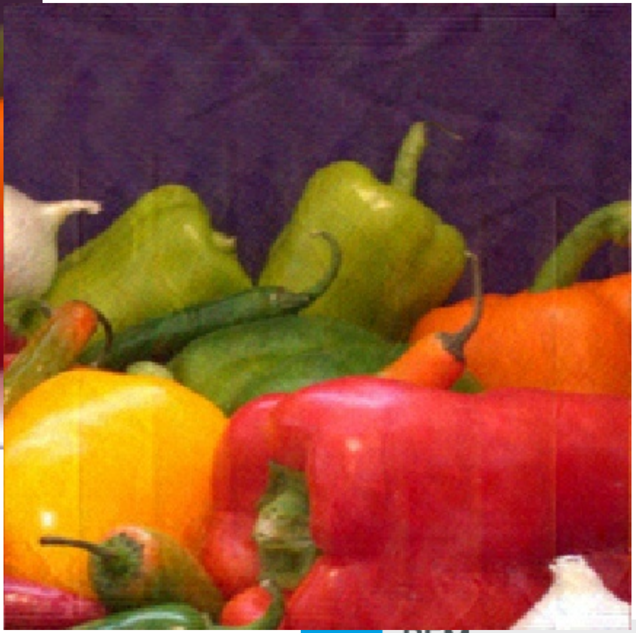
Original Image



Blurred Image



Restored image



Causes:

- Out of focus acquisition
- Camera-object movement
- Shaking
- Shallow field of view …

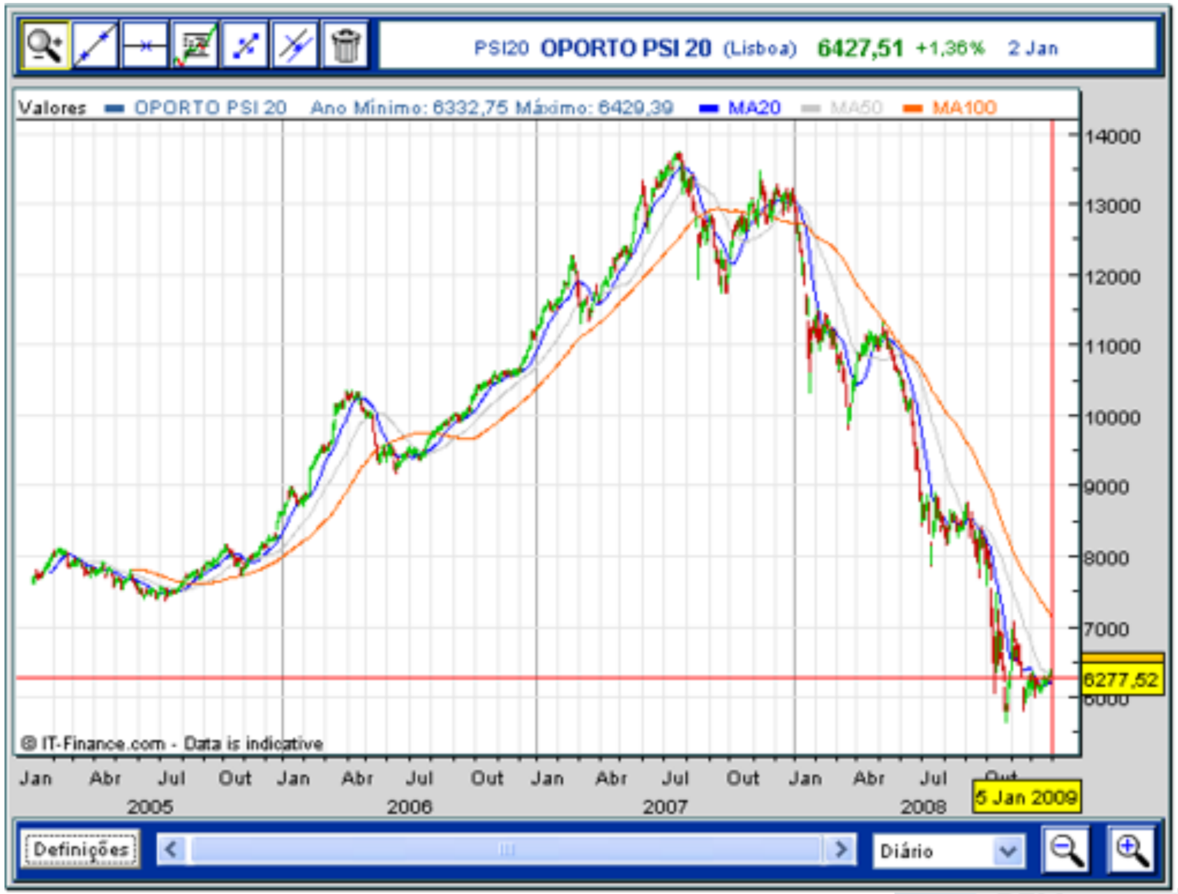
# Stock Exchange

***Models that explain evolution of phenomena***

- Causality
- Number of parameters
- Type of model
- Uncertainty

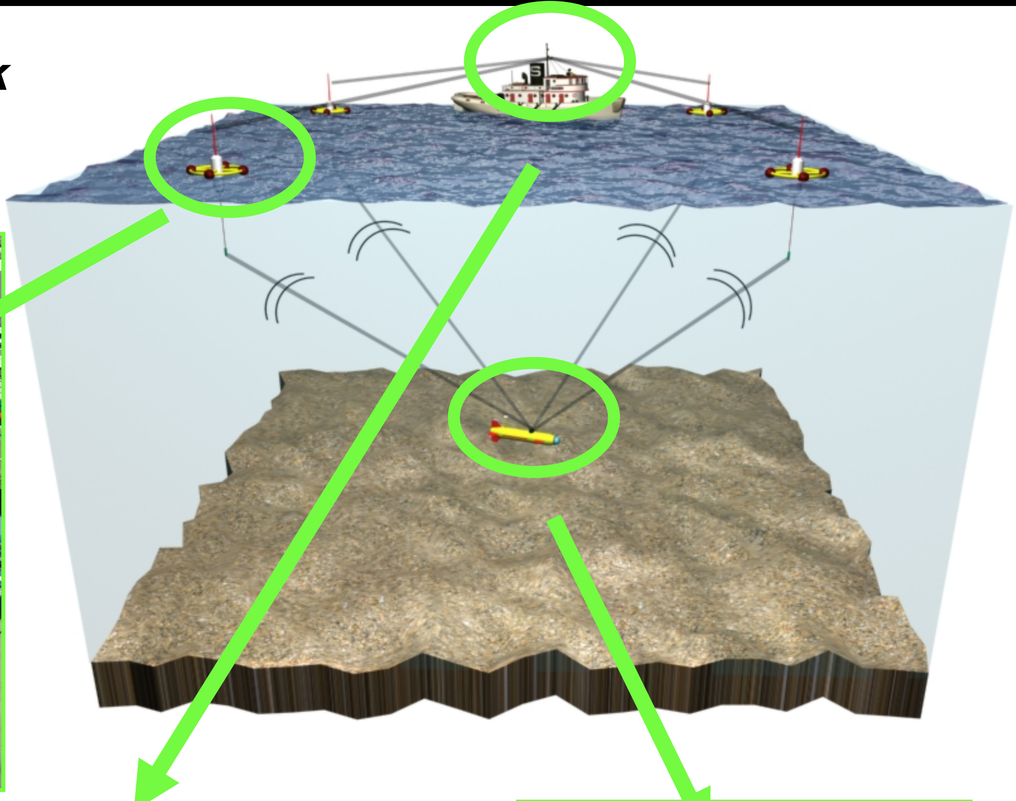
Is it possible to predict the market price tomorrow, next week, next month, next year,?…

Courtesy Jornal de Negócios



# GPS Intelligent Buoys(GIB)-ACSA/ORCA

***Tracking with a Sensor Network***

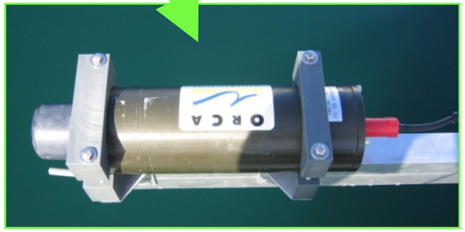


Surface buoys with

- *DGPS receivers*
- *Hydrophones*
- *Radio link*

Control Station

- *DGPS receiver*
- *Radio link*
- *PC with tracking software*

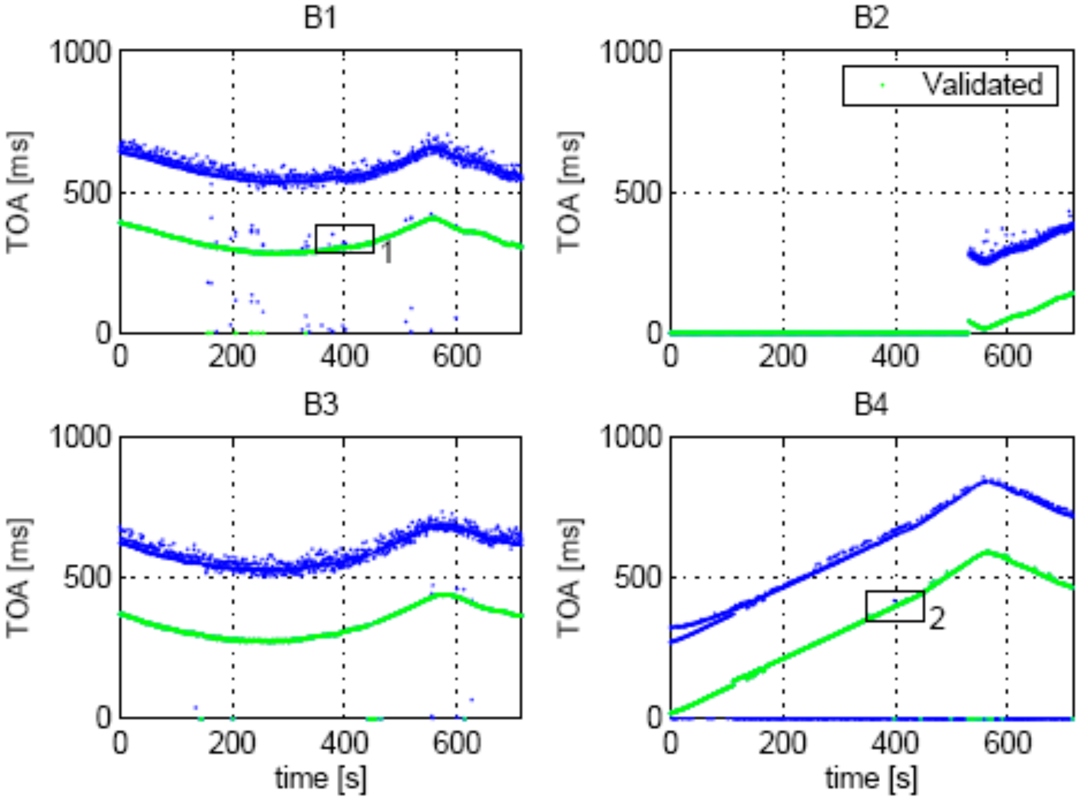


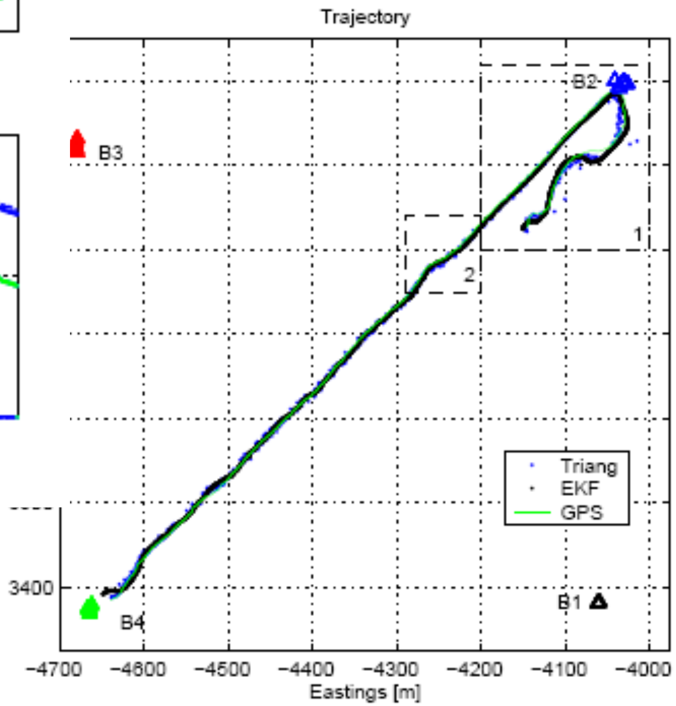
Acoustic pinger

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

# GPS Intelligent Buoys(GIB)-ACSA/ORCA

### *Tracking with a Sensor Network (cont.)*





# The mathematical estimation problem

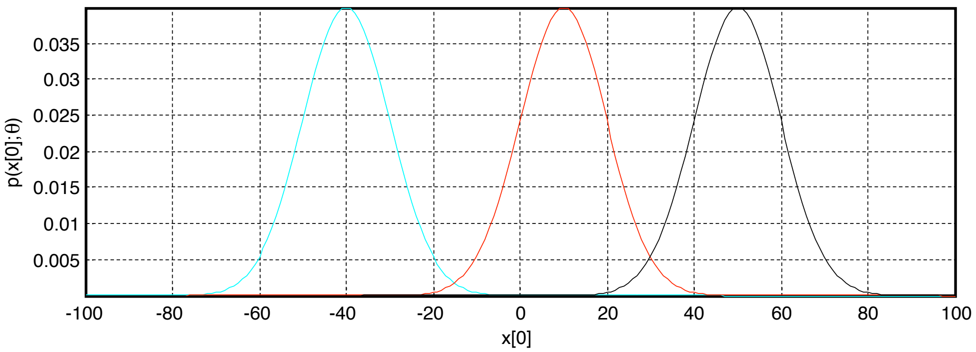
**To be possible to design estimators, first the data must be modeled.**

*Example I:*

Assume that one sample $x$ is available (scalar example, i.e. N=1) with constant **unknown** mean $\theta$.

The probability density function (PDF) is

$$p(x;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}}$$



For instance if x[0]<0 it is doubtful that the unknown parameter is >>0.

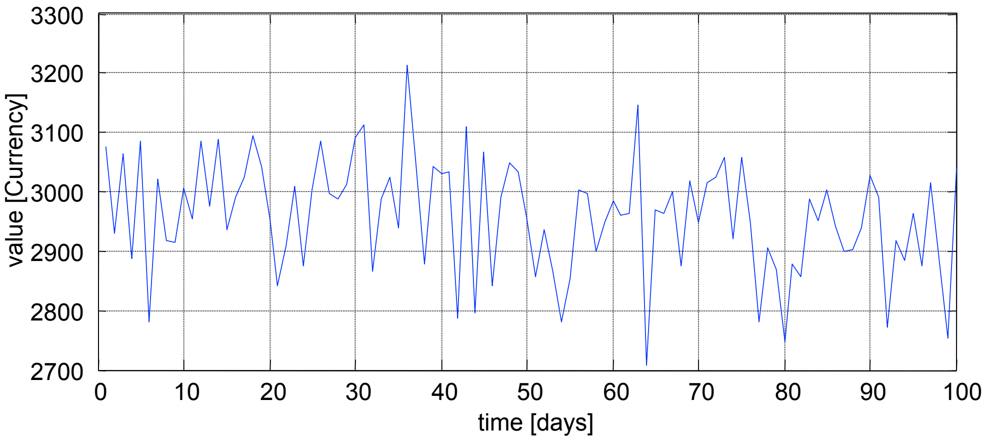
***In a actual problem, we are not given a PDF, but must be chosen to be consistent with the data and with the prior knowledge.***

# The mathematical estimation problem

*Example II:*

Now the following sequence ***x*** is given.

Note that the value along time appears to be decreasing. Lets consider that the phenomena is described by



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

where A and B are constant unknown parameters and w[n] is assumed to be white Gaussian noise, with PDF *N(0,σ 2)*. For ***θ****=[A B]* and **x***=[x[0] x[1] …x[n]] the data PDF is*

$$p(\mathbf{x}\,|\,\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A-Bn)^2}$$

Where the uncertainty in the samples is assumed to be uncorrelated.

**The performance of the estimators is dependent on the models used, so they must be mathematically treatable.**.

# The mathematical estimation problem

**Classical estimation techniques**

**Parameters are assumed deterministic but unknown**

**Bayesian techniques**

**Parameters are used to be unknown but are stochastic also described by a PDF.**

The joint PDF would then be

$$p(\mathbf{x}\,,\,\theta) = \underbrace{p(\mathbf{x} \mid \theta)}_{\text{Dependence of data on the parameters}}\;\underbrace{p(\theta)}_{\text{Prior knowledge}}$$

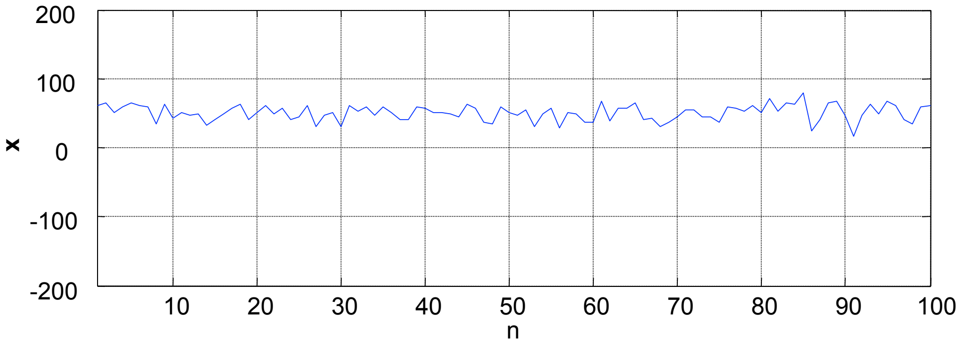
# Exploiting simple estimators

*Example III (Quiz):*

*Given a data sequence from a signal with*

*PDF as described by one of three models*

***Which one is the correct model?***

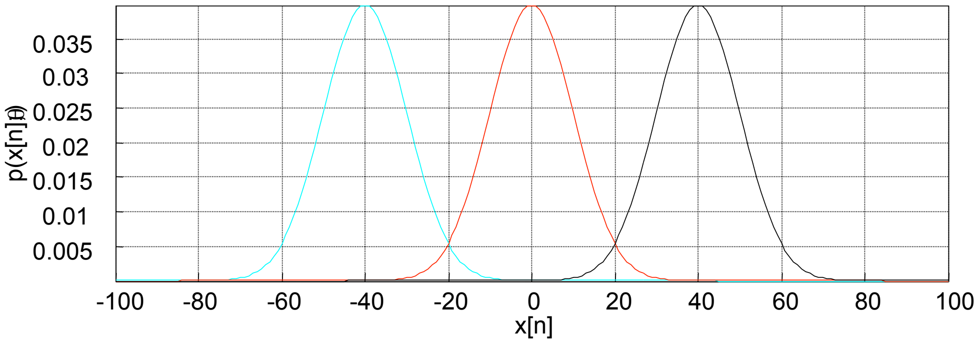


*First scenario:*

$$N = 100$$

$$\theta \in \{-40, 0, 40\}$$

$$\sigma^2 = 10^2$$



*For the signal*

$$x[n] = \theta + w[n] \qquad n = 0, 1, \ldots, N-1$$

*The answer is obvious:*

*$\theta$ = 40!*

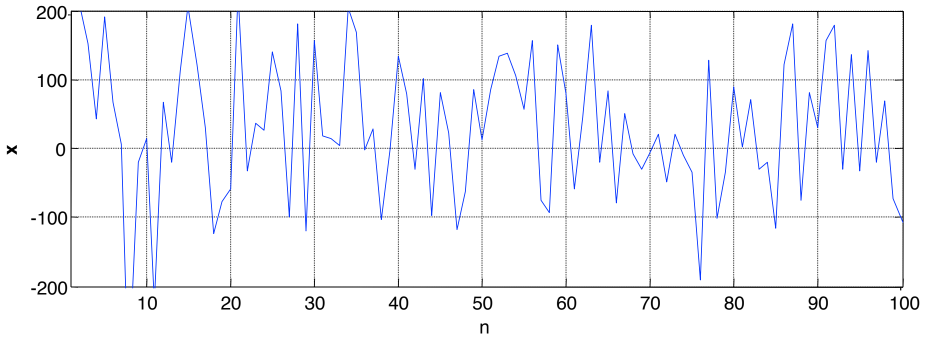
# *Exploiting simple estimators*

Second scenario

(lousy sensor quality or lousy data):

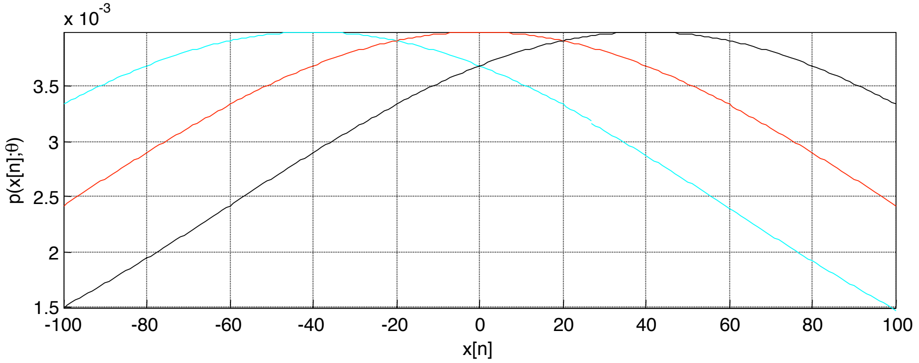
Lets *repeat the problem with*

$$\boldsymbol{\sigma}^2 = \mathbf{100}^2$$



$$N = 100$$

$$\theta \in \{-40, 0, 40\}$$



*The answer is not obvious anymore!*

*Lets propose a couple of estimators*

*and to study them…*

$$\hat{\theta}_1 = x[0] \qquad \hat{\theta}_2 = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$$

# Assessing estimator performance

**Estimators depend only on observed data thus can be viewed as a function**

$$\boldsymbol{\theta} = g\left(x[0], \quad \ldots \quad x[N]\right) = g\left(\boldsymbol{x}\right)$$

**The study of estimator properties must be done resorting to statistic tools.**

**Is it exact?, i.e. Does it return the true value of the unknown parameters?**

**Is this a good estimator? If many experiments can be performed, is it expected that the unknown parameter is achievable? Or are the results expected to be biased?**

$$E\left[\hat{\theta}_1\right] = E\left[x[0]\right] = \theta$$

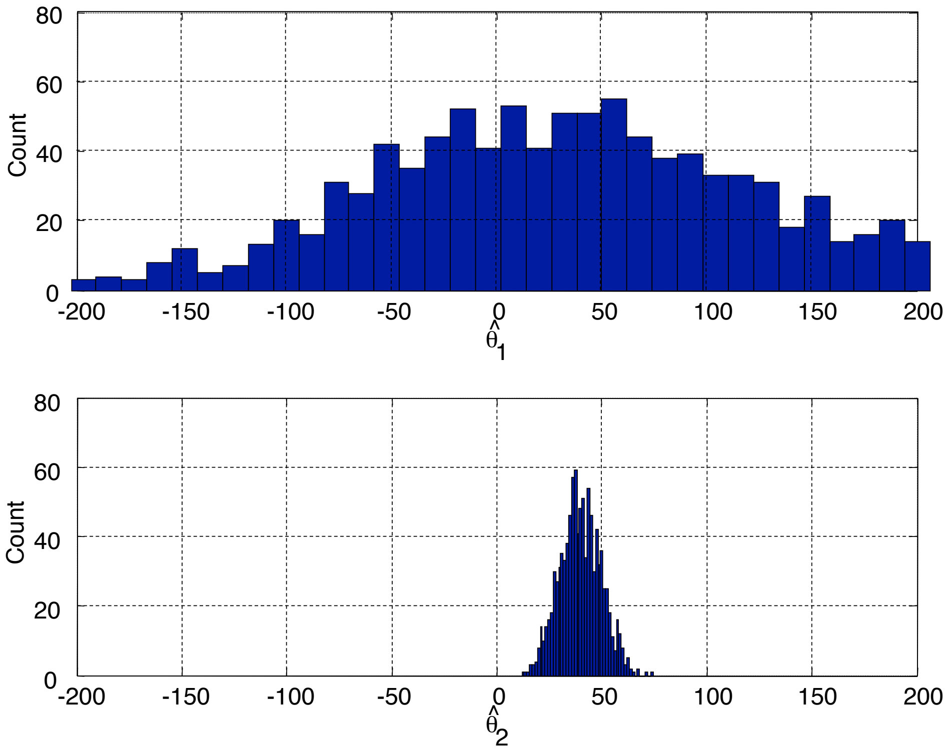
$$E\left[\hat{\theta}_2\right] = E\left[\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right] =$$

$$= \frac{1}{N}\sum_{n=0}^{N-1} E\left[x[n]\right] =$$

$$= \frac{1}{N} N\theta = \theta$$

# Assessing estimator performance

**How good is an estimator? How much uncertainty corresponds to the computed value?**



The use of computational tools is a good idea? No!

Formal methods are required

For our quiz:

$$\mathrm{var}\left(\hat{\theta}_1\right) = \mathrm{var}\left(x[0]\right) = \sigma^2$$

$$\mathrm{var}\left(\hat{\theta}_2\right) = \mathrm{var}\left(\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right) = \frac{1}{N^2}\sum_{n=0}^{N-1}\mathrm{var}\left(x[n]\right) == \frac{1}{N^2}N\sigma^2 = \frac{\sigma^2}{N}!$$

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

# Assessing estimator performance

Questions triggered from this simple example but valid to all our problems:

- **The second estimator is much better than the first estimator.**

- **The quality of the estimate increases with the number of points. Is it reasonable? Is it plausible?**

- **Do we have always data available? How to get data?**

- **Is this the best one can do with N samples?**

- **Are there better estimators that we can exploit?**

***Answers to this questions will be provided along the course…***

# Bibliography:

## Further reading

- Paul Etter, ***Underwater Acoustic Modelling and Simulation,*** Taylor & Francis, 2003.
- François Le Chevalier, ***Principles of Radar and Sonar Signal Processing*****,** Artech House, 2002.
- *Peter Wille,* ***Sound Images of the Ocean in Research and Monitoring,*** Springer, 2005.
- Lawrence Rabiner, Biing Juang, ***Fundamentals of Speech Recognition,*** Prentice Hall, 1993.
- Gilbert Strang, Kai Borre, ***Linear Algebra, Geodesy, and GPS,*** SIAM, 1997.
- Rafael Gonzales, Richard Woods, ***Digital Image Processing,*** Prentice Hall, 2001.
- Joseph Boccuzzi, ***Signal Processing for Wireless Communications,*** McGraw Hill, 2008.
- Venkatesh Saligrama, ***Networked Sensing Information and Control,*** Springer, 2008.
- Ching-Fang Lin, ***Modern Navigation, Guidance, and Control Processing***, Prentice Hall, 1991.

See for instance http://www.**ieee**.org/portal/site

# Advanced Control Systems Detection, Estimation, and Filtering

Graduate Course on the
MEng PhD Program
Spring 2012/2013

## Chapter 2
## Minimum Variance Unbiased 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. 1 - ***Motivation for Estimation in Stochastic Signal Processing*** [1/2 week] Motivating examples of signals and systems in detection and estimation problems;

Chap. 2 - ***Minimum Variance Unbiased Estimation*** [1/2 week]

Unbiased estimators; Minimum Variance Criterion; Extension to vector parameters; Efficiency of estimators;

Chap. 3 - ***Cramer-Rao Lower Bound*** [1 week]
Estimator accuracy; Cramer-Rao lower bound (CRLB); CRLB for signals in white Gaussian noise; Examples;

continues…

# Unbiased estimators:

The search for good estimators for unknown deterministic parameters begins

Example (revisited):

Unbiased estimator for DC level in white Gaussian noise. Signal model is

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

A reasonable estimator is $\hat{A} = k\dfrac{1}{N}\displaystyle\sum_{n=0}^{N-1} x[n]$

Due to the linearity properties of the expectation operator *E[.]*:

$$E\left[\hat{A}\right] = E\left[k\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right] = k\frac{1}{N}\sum_{n=0}^{N-1} E\left[x[n]\right] = k\frac{1}{N}NA = kA$$

***Unbiased estimator iff k=1!***

# Unbiasness:

Let the vector of deterministic unknown parameters, with p components, be described as

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \mathrm{M} \\ \theta_p \end{bmatrix} = \begin{bmatrix} \theta_1 & \theta_2 & \dots & \theta_p \end{bmatrix}^T$$

An estimator must have the same dimensions, i.e.

$$\hat{\theta} = \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \mathrm{M} \\ \hat{\theta}_p \end{bmatrix} = \begin{bmatrix} \hat{\theta}_1 & \hat{\theta}_2 & \dots & \hat{\theta}_p \end{bmatrix}^T$$

The estimator is **unbiased** iff

$$E\left[\hat{\theta}\right] = \theta$$

# Minimum variance criterion:

In searching for estimators some optimality criterion must be adopted.

A natural one is the **mean square error (MSE),** defined as

$$mse\left(\hat{\theta}\right) = E\left[\left(\hat{\theta} - \theta\right)^2\right]$$

Unfortunately the choice of the criterion leads to unrealizable estimators, i.e. not only function of the data

$$mse\left(\hat{\theta}\right) = E\left\{\left[\left(\hat{\theta} - E\left[\hat{\theta}\right]\right) + \left(E\left[\hat{\theta}\right] - \theta\right)\right]^2\right\} = \mathrm{var}\left(\hat{\theta}\right) + \left[\left(E\left[\hat{\theta}\right] - \theta\right)\right]^2 =$$

$$= \mathrm{var}\left(\hat{\theta}\right) + b^2\left(\theta\right), \qquad where \quad b^2\left(\theta\right) = E\left[\hat{\theta}\right] - \theta.$$

**Fortunately, after differentiation an estimator depending on θ will result.**

# Minimum variance criterion:

Example:

Find the value of k such that a realizable mse estimator results

$$x[n] = A + w[n] \qquad n = 0,1,\ldots,N-1, \qquad \hat{A} = k\frac{1}{N}\sum_{n=0}^{N-1} x[n]$$

From the previous page

$$\operatorname{var}\left(\hat{\mathrm{A}}\right) = \frac{k^2\sigma^2}{N}, \qquad and \quad b^2(A) = E\left[\hat{\mathrm{A}}\right] - A = (kA - A)^2$$

$$mse\left(\hat{\mathrm{A}}\right) = \operatorname{var}\left(\hat{\mathrm{A}}\right) + b^2(A) = \frac{k^2\sigma^2}{N} + (k-1)^2 A^2$$

Lets find the minimum

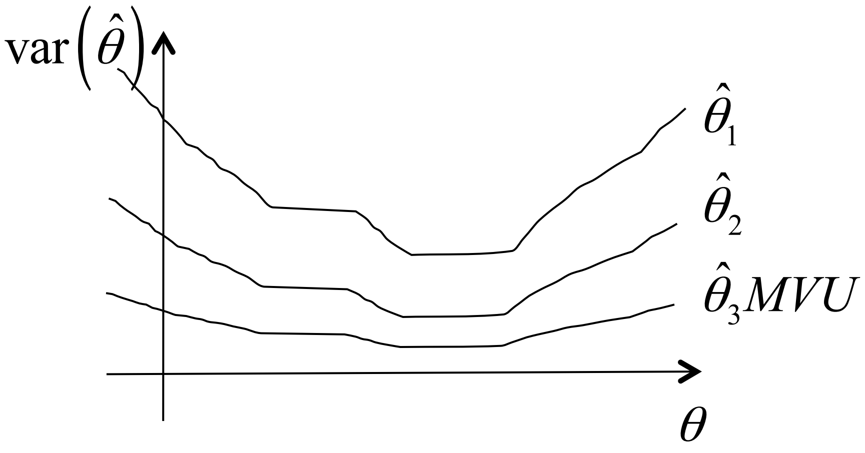
$$\frac{d}{dk} mse\left(\hat{\mathrm{A}}\right) = \frac{2k\sigma^2}{N} + 2(k-1)A^2 = 0, \qquad \text{results} \quad \text{in} \qquad k_{opt=}\frac{A^2}{A^2 + \sigma^2 / N}$$

Unfortunately depends on the unknown parameter A.

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

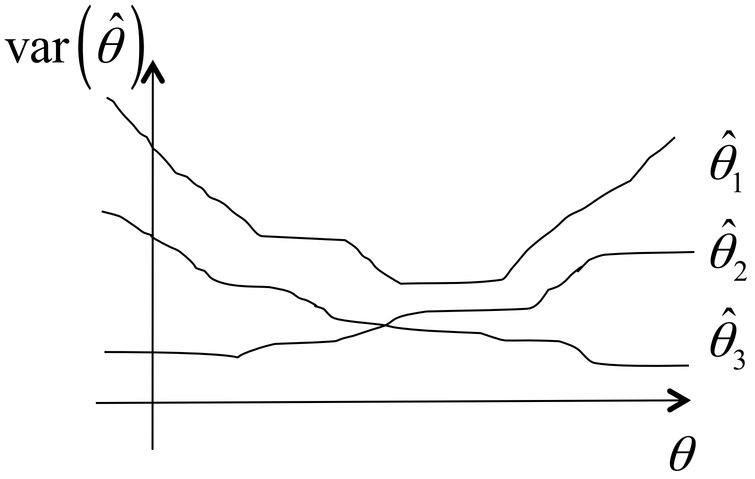
# *Minimum variance unbiased estimator:*

The Minimum Variance Unbiased (MVU) Estimator **must have** smallest variance for all values of θ.



In general, the MVU estimator does not always exist.

There is no “turn-the-crank ” method.



Future approaches:

Chp. 3 – Cramer-Rao lower bound

Chp. 5 – Sufficient statistics

Chp. 6 – Restrict to linear estimators: BLUE

# Bibliography:

## Further reading

- Harry L. Van Trees, ***Detection, Estimation, and Modulation Theory, Parts I to IV,*** John Wiley, 2001.
- J. Bibby, H. Toutenburg, ***Prediction and Improved Estimation in Linear Models,*** John Wiley, 1977.
- C.Rao, ***Linear Statistical Inference and Its Applications,*** John Wiley, 1973.
- P. Stoica, R. Moses, *"On Biased Estimators and the Unbiased Cramer-Rao Lower Bound," Signal Process,* vol.21, pp. 349-350, 1990.

# Advanced Control Systems
# Detection, Estimation, and Filtering

*Graduate Course on the*
*MEng PhD Program*
*Spring 2012/2013*

## Chapter 3
## Cramer-Rao Lower Bounds

*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. 2 - ***Minimum Variance Unbiased Estimation*** [1/2 week]

Unbiased estimators; Minimum Variance Criterion; Extension to vector parameters; Efficiency of estimators;

Chap. 3 - ***Cramer-Rao Lower Bound*** [1 week]

Estimator accuracy; Cramer-Rao lower bound (CRLB); CRLB for signals in white Gaussian noise; Examples;

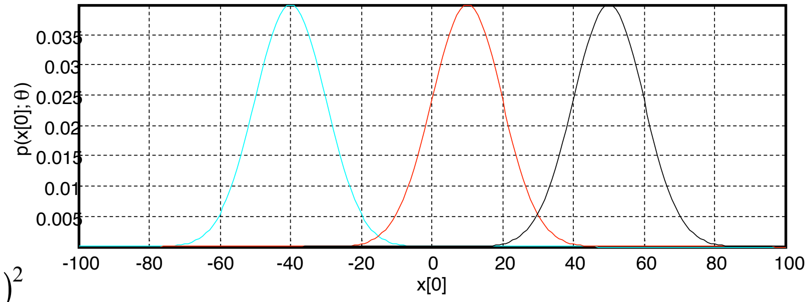
Chap. 4 - ***Linear Models in the Presence of Stochastic Signals*** [1 week]

Stationary and transient analysis; White Gaussian noise and linear systems; Examples; Sufficient Statistics; Relation with MVU Estimators;

continues…

# Estimator accuracy:

The accuracy on the estimates dependents very much on the PDFs

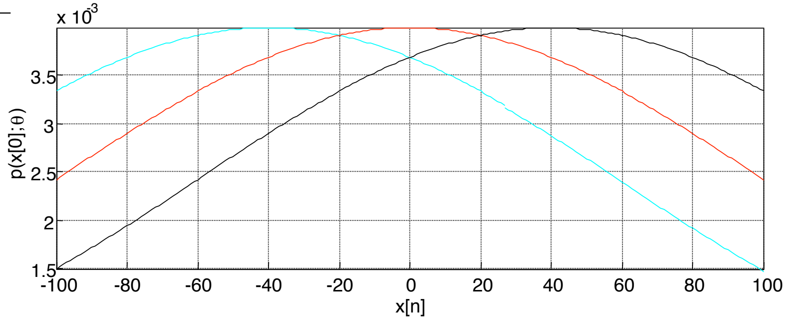


Example (revisited):

Model of signal $x[0]=A+w[0],$

Observation PDF $p\left(x[0];A\right)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x[0]-A)^2}{2\sigma^2}}$

for a disturbance *N(0, σ²)*



*Remarks:*

If $\sigma^2$ is **Large** then the performance of the estimator is **Poor**;

If $\sigma^2$ is **Small** then the performance of the estimator is **Good**; or

If PDF concentration is **High** then the parameter accuracy is **High.**

**How to measure sharpness of PDF (or concentration)?**

# Estimator accuracy:

When PDFs are seen as function of the unknown parameters, for x fixed, they are called as ***Likelihood function.*** To measure the sharpness note that (and ln is monotone…)

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

Its first and second derivatives are respectively:

$$\frac{\partial}{\partial A}\ln p\left(x[0];A\right) = \frac{1}{\sigma^2}\left(x[0]-A\right) \qquad \text{and} \qquad -\frac{\partial^2}{\partial A^2}\ln p\left(x[0];A\right) = \frac{1}{\sigma^2}.$$

As we know that the estimator $\hat{A}$ has variance $\sigma^2$ (at least for this example)

$$\operatorname{var}\left(\hat{A}\right) = \frac{1}{-\frac{\partial^2}{\partial A^2}\ln p\left(x;A\right)} = \frac{1}{curvature}$$

We are now ready to present an important theorem…

# Cramer-Rao lower bound:

***Theorem 3.1 (Cramer-Rao lower bound, scalar parameter)*** *– It is assumed that the PDF p(**x;** θ) satisfies the "regularity" condition*

$$E\left[\frac{\partial}{\partial\theta}\ln p\left(\mathbf{x};\theta\right)\right]=0 \qquad \text{for all} \quad \theta \tag{1}$$

*where the expectation is taken with respect to p(**x**; θ). Then, the variance of any unbiased estimator $\hat{\theta}$ must satisfy*

$$\operatorname{var}\left(\hat{\theta}\right)\geq\frac{1}{-E\left[\frac{\partial^2}{\partial\theta^2}\ln p\left(\mathbf{x};\theta\right)\right]} \tag{2}$$

*where the derivative is evaluated at the true value of $\theta$ and the expectation is taken with respect to p(**x**, $\theta$). Furthermore, an unbiased estimator can be found that attains the bound for all $\theta$ if and only if*

$$\frac{\partial}{\partial\theta}\ln p\left(\mathbf{x};\theta\right)=I\left(\theta\right)\left(g\left(\mathbf{x}\right)-\theta\right) \tag{3}$$

*for some functions g(.) and $I$ (.). The estimator, which is the MVU estimator, is $\hat{\theta}=g\left(\mathbf{x}\right)$, and the minimum variance 1/$I$(θ).*

# Cramer-Rao lower bound:

**Proof outline:**

Lets derive the CRLB for a scalar parameter $\alpha = g(\theta)$. We consider all unbiased estimators

$$E\left[\hat{\alpha}\right] = \alpha = g(\theta) \qquad \text{or} \qquad \int \hat{\alpha}\, p(\boldsymbol{x};\theta)\; d\boldsymbol{x} = g(\theta). \qquad \text{(p.1)}$$

Lets examine the regularity condition (1)

$$E\left[\frac{\partial}{\partial\theta}\ln p(\boldsymbol{x};\theta)\right] = \int \frac{\partial \ln p(\boldsymbol{x};\theta)}{\partial\theta} p(\boldsymbol{x};\theta) d\boldsymbol{x} = \int \frac{\partial p(\boldsymbol{x};\theta)}{\partial\theta} d\boldsymbol{x}$$

$$= \frac{\partial}{\partial\theta}\int p(\boldsymbol{x};\theta) d\boldsymbol{x} = \frac{\partial 1}{\partial\theta} = 0.$$

Remark: differentiation and integration are required to be interchangeable (Leibniz Rule)!

Lets differentiate (p.1) with respect to $\theta$ and use the previous results

$$\int \hat{\alpha} \frac{\partial p(\boldsymbol{x};\theta)}{\partial\theta} d\boldsymbol{x} = \frac{\partial g(\theta)}{\partial\theta} \qquad or \qquad \int \hat{\alpha} \frac{\partial \ln p(\boldsymbol{x};\theta)}{\partial\theta} p(\boldsymbol{x};\theta) d\boldsymbol{x} = \frac{\partial g(\theta)}{\partial\theta} \quad .$$

# Cramer-Rao lower bound:

**Proof outline (cont.):**

This can be modified to

$$\int (\alpha - \hat{\alpha}) \frac{\partial \ln p(\boldsymbol{x};\theta)}{\partial \theta} p(\boldsymbol{x};\theta) d\boldsymbol{x} = \frac{\partial g(\theta)}{\partial \theta},$$

as

$$\int \alpha \frac{\partial \ln p(\boldsymbol{x};\theta)}{\partial \theta} p(\boldsymbol{x};\theta) d\boldsymbol{x} = \alpha E\left[\frac{\partial \ln p(\boldsymbol{x};\theta)}{\partial \theta}\right] = 0.$$

Now applying the Cauchy-Schwarz inequality, i.e.

$$\left[\int w(\boldsymbol{x}) g(\boldsymbol{x}) h(\boldsymbol{x}) d\boldsymbol{x}\right]^2 \leq \int w(\boldsymbol{x}) g^2(\boldsymbol{x}) d\boldsymbol{x} \int w(\boldsymbol{x}) h^2(\boldsymbol{x}) d\boldsymbol{x}$$

considering $w(\boldsymbol{x}) = p(\boldsymbol{x};\theta)$, $g(\boldsymbol{x}) = \hat{\alpha} - \alpha$, and $h(\boldsymbol{x}) = \dfrac{\partial \ln p(\boldsymbol{x};\theta)}{\partial \theta}$

results

$$\left(\frac{\partial g(\theta)}{\partial \theta}\right)^2 \leq \int (\alpha - \hat{\alpha})^2 p(\boldsymbol{x};\theta) d\mathbf{x} \int \left(\frac{\partial \ln p(\boldsymbol{x};\theta)}{\partial \theta}\right)^2 p(\boldsymbol{x};\theta) d\mathbf{x}$$

# Cramer-Rao lower bound:

**Proof outline (cont.):**

It remains to relate this expression with the one in the Theorem $\int\left(\frac{\partial \ln p\left(\boldsymbol{x};\theta\right)}{\partial\theta}\right)^2 p\left(\boldsymbol{x};\theta\right)d\mathbf{x} = ?$

Starting with the previous result

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

thus, this function identically null verifies

$$\frac{\partial}{\partial\theta}\int\frac{\partial}{\partial\theta}\ln p\left(\boldsymbol{x};\theta\right)p\left(\boldsymbol{x};\theta\right)d\boldsymbol{x} = \int\left[\frac{\partial^2 \ln p\left(\boldsymbol{x};\theta\right)}{\partial\theta^2}p\left(\boldsymbol{x};\theta\right) + \frac{\partial \ln p\left(\boldsymbol{x};\theta\right)}{\partial\theta}\frac{\partial p\left(\boldsymbol{x};\theta\right)}{\partial\theta}\right]d\boldsymbol{x} =$$

$$\int\left[\frac{\partial^2 \ln p\left(\boldsymbol{x};\theta\right)}{\partial\theta^2}p\left(\boldsymbol{x};\theta\right) + \frac{\partial \ln p\left(\boldsymbol{x};\theta\right)}{\partial\theta}\frac{\partial \ln p\left(\boldsymbol{x};\theta\right)}{\partial\theta}p\left(\boldsymbol{x};\theta\right)\right]d\boldsymbol{x} = 0$$

And finally

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

# Cramer-Rao lower bound:

**Proof outline (cont.):**

Taking this into consideration, i.e.

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

expression (2) results, in the case where *g(θ)=θ*.

The result (3) will be obtained next…

□

See also appendix 3.B for the derivation in the vector case.

# Cramer-Rao lower bound:

***Summary:***

- ***Being able to place a lower bound on the variance of any unbiased estimator is very useful.***
- ***It allow us to assert that an estimator is the MVU estimator (if it attains the bound for all values of the unknown parameter).***
- ***It provides in all cases a benchmark for the unbiased estimators that we can design.***
- ***It alerts to impossibility of finding unbiased estimators with variance lower than the bound.***
- ***Provides a systematic way of finding the MVU estimator, if it exists and if an extra condition is verified.***

# Example:

Example (DC level in white Gaussian noise):

**Problem:** Find MVU estimator. **Approach:** Compute CRLB, if right form we have it.

Signal model: $x[n] = A + w[n], \qquad n = 0, \dots, N-1, \qquad w[n]: \; N\left(0, \sigma^2\right)$

Likelihood function: $$p(\mathbf{x}; A) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2}$$

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

$$\frac{\partial^2}{\partial A^2}\ln p(\mathbf{x}; A) = -\frac{N}{\sigma^2}$$

**CRLB**: $$\operatorname{var}\left(\hat{A}\right) \geq \frac{1}{N/\sigma^2} = \frac{\sigma^2}{N}$$

The estimator is unbiased and has the same variance, **thus it is a MVU** estimator! And it has the form:

$$\frac{\partial}{\partial A}\ln p(x; A) = I(\theta)\left(g(\boldsymbol{x}) - \theta\right), \qquad \text{for} \quad I(\theta) = \frac{N}{\sigma^2} \qquad g(\boldsymbol{x}) = \bar{\mathbf{x}}.$$

# Cramer-Rao lower bound:

**Proof outline (second part of the theorem):**

Still remains to prove that the CRLB is attained for the estimator $\hat{\theta} = g(\mathbf{x})$

$$\operatorname{var}\left(\hat{\theta}\right) = \frac{1}{I(\theta)}, \qquad \text{for} \quad I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ln p(x;\theta)\right]$$

If

$$\frac{\partial}{\partial\theta}\ln p(x;\theta) = I(\theta)\big(g(\boldsymbol{x}) - \theta\big)$$

differentiation relative to the parameter gives

$$\frac{\partial^2}{\partial\theta^2}\ln p(\boldsymbol{x};\theta) = \frac{\partial I(\theta)}{\partial\theta}\big(g(\mathbf{x}) - \theta\big) - I(\theta)$$

and then

$$-E\left[\frac{\partial^2}{\partial\theta^2}\ln p(\boldsymbol{x};\theta)\right] = -\frac{\partial I(\theta)}{\partial\theta}\big(E\left[g(\mathbf{x})\right] - \theta\big) + I(\theta) = I(\theta)$$

i.e. the bound is attained. ■

# Example:

Example (phase estimation):

Signal model: $x[n] = A\cos(2\pi f_0 n + \phi) + w[n], \qquad n = 0,\dots,N-1$

$A, f_0$ known

Likelihood function: $p(\mathbf{x};\phi) = \dfrac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\left(x[n]-A\cos(2\pi f_0 n+\phi)\right)^2}$

$$\frac{\partial}{\partial\phi}\ln p(\mathbf{x};\phi) = -\frac{A}{\sigma^2}\sum_{n=0}^{N-1}\left(x[n]\sin(2\pi f_0 n + \phi) - \frac{A}{2}\sin(4\pi f_0 n + 2\phi)\right)A$$

$$\frac{\partial^2}{\partial\phi^2}\ln p(\mathbf{x};\phi) = -\frac{A}{\sigma^2}\sum_{n=0}^{N-1}\left(x[n]\cos(2\pi f_0 n + \phi) - A\cos(4\pi f_0 n + 2\phi)\right)A$$

$$-E\left[\frac{\partial^2}{\partial\phi^2}\ln p(\mathbf{x};\phi)\right] = -\frac{A^2}{\sigma^2}\sum_{n=0}^{N-1}\left(\frac{1}{2} - \frac{1}{2}\cos(4\pi f_0 n + 2\phi)\right) \approx \frac{NA^2}{2\sigma^2}$$

# Example:

Example (phase estimation cont.):

$$-E\left[\frac{\partial^2}{\partial\phi^2}\ln p\left(\mathbf{x}\,|\,\phi\right)\right] = -\frac{A^2}{\sigma^2}\sum_{n=0}^{N-1}\left(\frac{1}{2}-\frac{1}{2}\cos\left(4\pi f_0 n+2\phi\right)\right) \approx \frac{NA^2}{2\sigma^2}$$

as $\sum_{n=0}^{N-1}\cos\left(4\pi f_0 n+2\phi\right)\approx 0 \text{ for } f_0 \text{ not near 0 or 1/2.}$

$$\frac{1}{N}\sum_{n=0}^{N-1}\cos\left(\alpha n+\beta\right) = \frac{1}{N}\mathrm{Re}\left\{\sum_{n=0}^{N-1}e^{j(\alpha n+\beta)}\right\} = \ldots = \frac{\sin\left(N\alpha/2\right)}{N\sin\left(\alpha/2\right)}\cos\left(\alpha\frac{N-1}{2}+\beta\right)$$

for large *N*.

$$\mathrm{var}\left(\hat{\phi}\right) \geq \frac{2\sigma^2}{NA^2}$$

- Bound decreases as *SNR=A*$^2$/2*σ*$^2$ increases
- Bound decreases as *N* increases

Does an efficient estimator exists? Does a MVUE estimator exists?

# Fisher information:

*We define the Fisher Information (Matrix) as*

$$I\left(\hat{\theta}\right) = -E\left[\frac{\partial^2}{\partial\theta^2}\ln p\left(x;\theta\right)\right]$$

*Note:*

- *$I$(θ) ≥0*
- *It is additive for independent observations*

$$\ln p\left(\mathbf{x};\theta\right) = \ln \pi_{n=0}^{N-1} p\left(x\left[n\right];\theta\right) = \sum_{n=0}^{N-1}\ln p\left(x\left[n\right];\theta\right)$$

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

- *If identically distributed (same PDF for each x[n])*

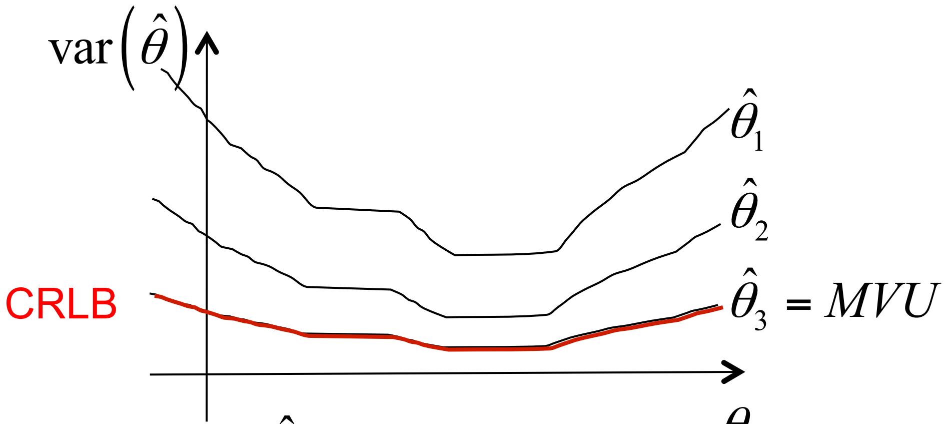
$$I\left(\theta\right) = Ni\left(\theta\right) = -N\left[\frac{\partial^2}{\partial\theta^2}\ln p\left(x\left[.\right];\theta\right)\right]$$

*As N->∞*, for iid => CRLB-> 0

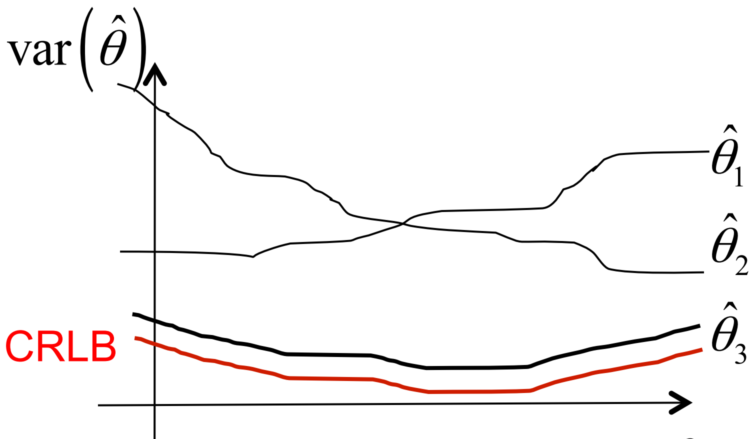
# Other estimator characteristic:

Efficiency:

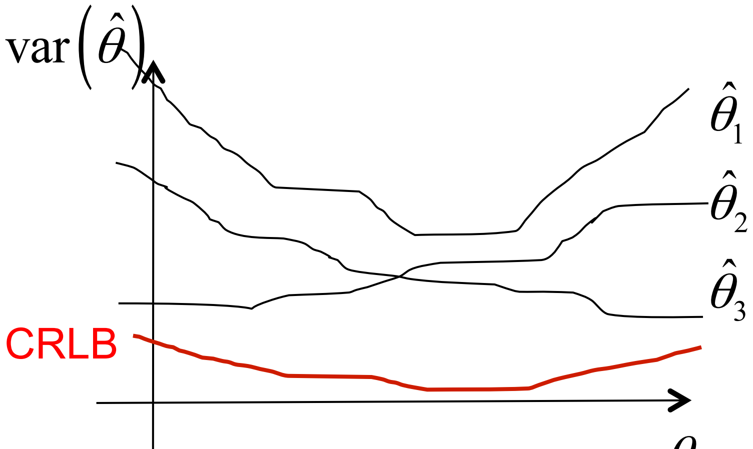
An estimator that is unbiased and attains the CRLB is said to be **efficient.**



$\hat{\theta}_3$ MVU and efficient



$\hat{\theta}_3$ MVU but not efficient



No MVU available.

# *Transformation of parameters:*

Imagine that the CRLB is known for the parameter $\theta$. Can we compute easily the CRLB for a linear transformation of the form $\alpha = g(\theta) = a\theta + \beta$ ?

$$\hat{\alpha} = a\hat{\theta} + b, \qquad E\left[a\hat{\theta} + b\right] = aE\left[\hat{\theta}\right] + b = \alpha$$

$$\operatorname{var}\left[a\hat{\theta} + b\right] = a^2 \operatorname{var}\left[\hat{\theta}\right] = \frac{\left(\frac{\partial g(\theta)}{\partial \theta}\right)^2}{-E\left[\frac{\partial^2}{\partial \theta^2} \ln p(\boldsymbol{x};\theta)\right]}$$

**Linear transformations preserve biasness and efficiency.**

And for a nonlinear transformation of the form α=g(θ)?

# Transformation of parameters:

**Remark: after a nonlinear transformation, the good properties can be lost.**

$$\operatorname{var}\left(\hat{\theta}\right) \geq \frac{\left(\frac{\partial g(\theta)}{\partial \theta}\right)^2}{-E\left[\frac{\partial^2}{\partial \theta^2} \ln p(\boldsymbol{x};\theta)\right]}$$

Example: Suppose that given a stochastic variable $\overline{x} : N\left(A, \frac{\sigma^2}{N}\right)$ we desire to have an estimator for α=g(A)=A² (power estimator). Note that

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

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

**A bias estimate results. Efficiency is lost.**

# Cramer-Rao lower bound:

***Theorem 3.1 (Cramer-Rao lower bound, Vector parameter)*** *– It is assumed that the PDF p(**x;θ**) satisfies the "regularity" condition*

$$E\left[\frac{\partial}{\partial\theta}\ln p\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right]=\mathbf{0} \qquad \text{for all } \theta$$

*where the expectation is taken with respect to p(**x,** θ). Then, the variance of any unbiased estimator* $\hat{\theta}$ *must satisfy*

$$\mathbf{C}_{\hat{\theta}}-I^{-1}\left(\boldsymbol{\theta}\right)\geq 0,$$

*where ≥ is interpreted as meaning the matrix is positive semi-definite. The Fisher information matrix* $I(\boldsymbol{\theta})$ *is given as*

$$\left[I\left(\boldsymbol{\theta}\right)\right]_{ij}=-E\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln p\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right],$$

*where the derivatives are evaluated at the true value of θ and the expectation is taken with respect to* $p$*(**x**;θ). Furthermore, an unbiased estimator may be found that attains the bound for all **θ** if and only if*

$$\frac{\partial}{\partial\theta}\ln p\left(\mathbf{x};\boldsymbol{\theta}\right)=I\left(\boldsymbol{\theta}\right)\left(\mathbf{g}\left(\mathbf{x}\right)-\boldsymbol{\theta}\right) \qquad (3)$$

*for some functions p dimensional function **g(.)** and some p x p matrix **I (.)**. The estimator, which is the MVU estimator, is* $\hat{\boldsymbol{\theta}}=\mathbf{g}\left(x\right)$, *and its covariance matrix is* $\boldsymbol{I}^{-1}(\boldsymbol{\theta})$.

# Vector Transformation of parameters:

The vector transformation of parameters $\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta})$ impacts on the CRLB computation as

$$\mathbf{C}_{\hat{\alpha}} - \frac{\partial \mathbf{g}(\theta)}{\partial \theta} I^{-1}(\theta) \frac{\partial \mathbf{g}(\theta)}{\partial \theta}^{T} \geq 0$$

where the Jacobian is

$$\frac{\partial \mathbf{g}(\theta)}{\partial \theta} = \begin{bmatrix} \frac{\partial g_1(\theta)}{\partial \theta_1} & \cdots & \frac{\partial g_1(\theta)}{\partial \theta_p} \\ \cdots & \cdots & \cdots \\ \frac{\partial g_r(\theta)}{\partial \theta_1} & & \frac{\partial g_r(\theta)}{\partial \theta_p} \end{bmatrix}$$

In the Gaussian general case for *x[n]=s[n]+w[n]*, where $\mathbf{w} \sim N\left(\mu(\theta), \mathbf{C}_{\theta}\right)$

the Fisher information matrix is

$$\left[I(\boldsymbol{\theta})\right]_{ij} = \left[\frac{\partial \mu(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\right]^{T} \mathbf{C}^{-1}(\boldsymbol{\theta}) \left[\frac{\partial \mu(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\right] \mathbf{C}_{\hat{\alpha}} + \frac{1}{2} tr\left[\mathbf{C}^{-1}(\boldsymbol{\theta}) \frac{\partial C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i} \mathbf{C}^{-1}(\boldsymbol{\theta}) \frac{\partial C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\right].$$

# Example:

Example (line fitting):

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

$A, B$ deterministic unknown quantities

Likelihood function: $p(\mathbf{x};\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A-Bn)^2}, \qquad \text{where} \quad \theta = \begin{bmatrix} A & B \end{bmatrix}^T$

The Fisher Information Matrix is

$$\mathbf{I}(\theta) = \begin{bmatrix} -E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial A^2}\right] & -E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial A \partial B}\right] \\ -E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial B \partial A}\right] & -E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial B^2}\right] \end{bmatrix}$$

where

$$\frac{\partial \ln p(\mathbf{x};\theta)}{\partial A} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A-Bn), \qquad \text{and} \qquad \frac{\partial \ln p(\mathbf{x};\theta)}{\partial B} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A-Bn)n.$$

# Example:

Example (cont.):

Moreover

$$\frac{\partial^2 \ln p\left(\mathbf{x};\theta\right)}{\partial A^2} = -\frac{N}{\sigma^2}, \frac{\partial^2 \ln p\left(\mathbf{x};\theta\right)}{\partial A \partial B} = -\frac{1}{\sigma^2}\sum_{n=0}^{N-1} n, \qquad \text{and} \qquad \frac{\partial^2 \ln p\left(\mathbf{x};\theta\right)}{\partial B^2} = -\frac{1}{\sigma^2}\sum_{n=0}^{N-1} n^2.$$

Since the second order derivatives do not depend on x, we have immediately that

$$\mathbf{I}\left(\theta\right) = \frac{1}{\sigma^2}\begin{bmatrix} N & \frac{N(N-1)}{2} \\ \frac{N(N-1)}{2} & \frac{N(N-1)(2N-1)}{6} \end{bmatrix}$$

And also,

$$\mathbf{I}^{-1}\left(\theta\right) = \sigma^2\begin{bmatrix} \frac{2(2N-1)}{N(N+1)} & -\frac{6}{N(N+1)} \\ -\frac{6}{N(N+1)} & \frac{12}{N(N^2-1)} \end{bmatrix}, \qquad \begin{aligned} \operatorname{var}\left(\hat{A}\right) &\geq \frac{2(2N-1)}{N(N+1)}\sigma^2 \\ \operatorname{var}\left(\hat{B}\right) &\geq \frac{12}{N(N^2-1)}\sigma^2 \end{aligned}$$

# Example:

Example (cont.):

Remarks:

For only one parameter to be determined $\mathrm{var}\left(\hat{A}\right) \geq \frac{\sigma^2}{N}$. Thus a general results was obtained: ***when more parameters are to be estimated the CRLB always degrades.***

*Moreover*

$$\frac{CRLB(\hat{A})}{CRLB(\hat{B})} = \frac{(2N-1)(N-1)}{6} > 1, \qquad \text{for} \quad N \geq 3.$$

The parameter *B* is easier to be determined, as its CRLB decreases with 1/N$^3$. This means that *x[n]* is more sensitive to changes in *B* than changes in *A*.

$$\Delta x\left[n\right] \approx \frac{\partial x\left[n\right]}{\partial A}\Delta A = \Delta A$$

$$\Delta x\left[n\right] \approx \frac{\partial x\left[n\right]}{\partial B}\Delta B = n\Delta B.$$

# Bibliography:

## Further reading

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

- J. Bibby, H. Toutenburg, ***Prediction and Improved Estimation in Linear Models,*** John Wiley, 1977.

- C.Rao, ***Linear Statistical Inference and Its Applications,*** John Wiley, 1973.

- P. Stoica, R. Moses, *“On Biased Estimators and the Unbiased Cramer-Rao Lower Bound,” Signal Process,* vol.21, pp. 349-350, 1990.

# Advanced Control Systems Detection, Estimation, and Filtering

Graduate Course on the
MEng PhD Program
Spring 2012/2013

## Chapter 4
## Linear Models in the Presence of Stochastic Signals

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)

# Syllabus:

## Classical Estimation Theory

…

Chap. 3 - ***Cramer-Rao Lower Bound*** [1 week]
Estimator accuracy; Cramer-Rao lower bound (CRLB); CRLB for signals in white Gaussian noise; Examples;

Chap. 4 - ***Linear Models in the Presence of Stochastic Signals*** [1 week] Stationary and transient analysis; White Gaussian noise and linear systems; Examples; Sufficient Statistics; Relation with MVU Estimators;

Chap. 5 - ***Best Linear Unbiased Estimators*** [1 week]
Definition of BLUE estimators; White Gaussian noise and bandlimited systems; Examples; Generalized minimum variance unbiased estimation;

continues…

# *A very special class of systems:*

FACT:

**The determination of the MVU Estimator is in general a difficult task.**

**A class of systems that allows the determination of this estimator easily…**

***LINEAR SYSTEMS***

**The statistical performance is also easy to compute**

**and an efficient solution is obtained.**

**The key point is on the formulation of a problem as a linear one.**

# MVU Estimator for the Linear Model:

***Theorem 4.1*** *– If the data observed can be modeled as*

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

*where* $\mathbf{x}$ *is a N x 1 vector of observations,* $\mathbf{H}$ *is a known N x p observation matrix (with N>p) and rank p,* $\boldsymbol{\theta}$ *is a p x 1 vector of parameters to be estimated, and* $\mathbf{w}$ *is an N x 1 noise vector with PDF N(0,* $\sigma^2\mathbf{I}$*), then the MVU is*

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

*and the covariance matrix of estimate is*

$$C_{\hat{\theta}} = \sigma^2\left(\mathbf{H}^T\mathbf{H}\right)^{-1}. \qquad (2)$$

# MVU Estimator for the Linear Model:

**Proof outline:**

As discussed in Chapter 3, it is possible to determine the MVU estimator if the equality constraints of the CRLB are satisfied.

From the signal model, it follows that the log-likelihood function is

$$\ln p\left(\mathbf{x};\boldsymbol{\theta}\right) = -\ln\left(2\pi\sigma^2\right)^{N/2} - \frac{\left(\mathbf{x}-\mathbf{H}\boldsymbol{\theta}\right)^T\left(\mathbf{x}-\mathbf{H}\boldsymbol{\theta}\right)}{2\sigma^2}$$

And

$$\frac{\partial \ln p\left(\mathbf{x};\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}} = -\frac{1}{2\sigma^2}\frac{\partial}{\partial\boldsymbol{\theta}}\left[\mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\mathbf{H}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\theta}\right].$$

Using the relations (deduce them, good exercise…)

$$\frac{\partial\mathbf{b}^T\boldsymbol{\theta}}{\partial\boldsymbol{\theta}} = \mathbf{b} \qquad \frac{\partial\boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta}}{\partial\boldsymbol{\theta}} = 2\mathbf{A}\boldsymbol{\theta}$$

# MVU Estimator for the Linear Model:

**Proof outline (cont):**

It follows

$$\frac{\partial \ln p\left(\mathbf{x};\theta\right)}{\partial \theta} = \frac{1}{\sigma^2}\left[\mathbf{H}^T\mathbf{x} - \mathbf{H}^T\mathbf{H}\theta\right].$$

Under the assumptions of the theorem, **H**$^T$**H** is invertible

$$\frac{\partial \ln p\left(\mathbf{x};\theta\right)}{\partial \theta} = \frac{\mathbf{H}^T\mathbf{H}}{\sigma^2}\left[\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x} - \theta\right].$$

$$\left(\frac{\partial \ln p\left(\mathbf{x};\theta\right)}{\partial \theta} = \mathbf{I}\left(\theta\right)\left[\mathbf{g}\left(\mathbf{x}\right) - \theta\right]\right)$$

Note that it is in the format introduced in the previous chapter, from where (1) and (2) follows immediately.

□

Major constraints:

what if **H**$^T$**H** is not invertible?

what if **H**$^T$**H** is ill-conditioned?

# Example - Fourier Analysis:

Cyclic components in white Gaussian noise

Signal model:

$$x[n] = \sum_{k=1}^{M} a_k \cos\left(\frac{2\pi kn}{N}\right) + \sum_{k=1}^{M} b_k \sin\left(\frac{2\pi kn}{N}\right) + w[n], \qquad n = 0,\dots,N-1, \quad w[n] : \ N\left(0,\sigma^2\right)$$

Defining

� $= \begin{bmatrix} a_1 & \dots & a_M & b_1 & \dots & b_M \end{bmatrix}^T, \mathbf{w} = \begin{bmatrix} w_0 & & w_{N-1} \end{bmatrix}^T$, and

$$\mathbf{H} = \begin{bmatrix} 1 & \dots & 1 & 0 & \dots & 0 \\ c\left(\frac{2\pi}{N}\right) & \dots & c\left(\frac{2\pi M}{N}\right) & s\left(\frac{2\pi}{N}\right) & \dots & s\left(\frac{2\pi M}{N}\right) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ c\left(\frac{2\pi(N-1)}{N}\right) & \dots & c\left(\frac{2\pi M(N-1)}{N}\right) & s\left(\frac{2\pi(N-1)}{N}\right) & \dots & s\left(\frac{2\pi M(N-1)}{N}\right) \end{bmatrix}$$

The model can be reformulated as a linear system, with solution if *M* < *N/2*

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

# Example - Fourier Analysis (cont.):

Important fact: The columns of $\mathbf{H}$ are orthogonal.

Define

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \dots & \mathbf{h}_{2M} \end{bmatrix}, \qquad \text{it follows} \quad \mathbf{h}_i^T \mathbf{h}_j = 0, \qquad i \neq j \quad .$$

Moreover, the discrete Fourier Transform (DFT) relations can be applied, i.e.

$$\sum_{n=0}^{N-1} \cos\left(\frac{2\pi in}{N}\right) \cos\left(\frac{2\pi jn}{N}\right) = \frac{N}{2}\delta_{ij}$$

$$\sum_{n=0}^{N-1} \sin\left(\frac{2\pi in}{N}\right) \sin\left(\frac{2\pi jn}{N}\right) = \frac{N}{2}\delta_{ij}$$

$$\sum_{n=0}^{N-1} \cos\left(\frac{2\pi in}{N}\right) \sin\left(\frac{2\pi jn}{N}\right) = 0, \qquad \text{for all} \quad i, j.$$

From where it follows

$$\mathbf{H}^T \mathbf{H} = \frac{N}{2}\mathbf{I}.$$

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

# Example - Fourier Analysis (cont.):

The MVU estimator is

$$\hat{\theta} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x} = \frac{2}{N}\mathbf{H}^T\mathbf{x} = \begin{bmatrix} \frac{2}{N}h_1^T\mathbf{x} \\ \vdots \\ \frac{2}{N}h_{2M}^T\mathbf{x} \end{bmatrix},$$

§

or finally

$$\hat{a}_k = \frac{2}{N}\sum_{n=0}^{N-1} x[n]\cos\left(\frac{2\pi kn}{N}\right),$$

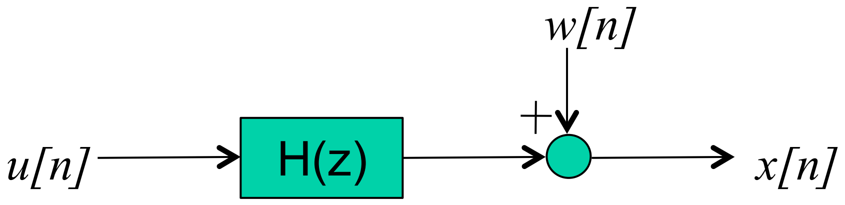
$$\hat{b}_k = \frac{2}{N}\sum_{n=0}^{N-1} x[n]\sin\left(\frac{2\pi kn}{N}\right).$$

with covariance

$$C_{\hat{\boldsymbol{\theta}}} = \frac{2\sigma^2}{N}\mathbf{I}.$$

# Example: System Identification

Signal model, where the Finite Impulse Response (FIR) is to be estimated.



The user can apply the input signal *u:*

$$x[n] = \sum_{k=0}^{p-1} h[k]u[n-k] + w[n], \qquad n = 0,\dots,N-1, \quad w[n] \sim N\left(0,\sigma^2\right) \; .$$

In matrix form, considering $\boldsymbol{x}=[x_0 \; \dots \; x_{N-1}]^T$, the input/output relations of this linear system can be written as

$$\mathbf{x} = \begin{bmatrix} u[0] & 0 & \dots & 0 \\ u[1] & u[0] & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u[N-1] & u[N-2] & \dots & u[N-p] \end{bmatrix} \begin{bmatrix} h[0] \\ h[1] \\ \vdots \\ h[p-1] \end{bmatrix} + \mathbf{w} \qquad \mathbf{w} = \begin{bmatrix} w_0 & & w_{N-1} \end{bmatrix}^T$$

Or in compact form, once again

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

# Example - System Identification (cont.):

The MVU estimator is once again

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}, \quad \text{with covariance} \quad C_{\hat{\boldsymbol{\theta}}} = \sigma^2\left(\mathbf{H}^T\mathbf{H}\right)^{-1}.$$

Note that accuracy depends on the input signal applied. How to choose it?

Problem: Choose u[n] to minimize $\operatorname{var}\left(\hat{\theta}_i\right) = \left[\mathbf{C}_{\hat{\boldsymbol{\theta}}}\right]_{ii}, i = 1,\ldots,p,$ subject to the constraint that $\sum_{n=0}^{N-1} u[n]$ is fixed.

Introducing the crosscorrelation (autocorrelation)

$$r_{ux}[i] = \frac{1}{N}\sum_{n=0}^{N-1-i} u[n]x[n+i] \qquad \mathbf{H}^T\mathbf{H} = \begin{bmatrix} r_{uu}[0] & r_{uu}[1] & \cdots & r_{uu}[p-1] \\ r_{uu}[1] & r_{uu}[0] & \cdots & r_{uu}[p-2] \\ \vdots & \vdots & \ddots & \vdots \\ r_{uu}[p-1] & r_{uu}[p-2] & \cdots & r_{uu}[0] \end{bmatrix}$$

Choosing a Pseudorandom Noise (PRN) makes this last matrix diagonal

$$C_{\hat{\boldsymbol{\theta}}} = \sigma^2\left(r_{uu}[0]\mathbf{I}\right)^{-1} = \frac{\sigma^2}{r_{uu}[0]}\mathbf{I}.$$

Input Signal Energy

# Extension to non-white Gaussian noise:

***Theorem (Generalization of Theorem 4.1)*** *– If the data observed can be modeled as*

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

*where* $\mathbf{x}$ *is a N x 1 vector of observations,* $\mathbf{H}$ *is a known N x p observation matrix (with N>p) and rank p,* $\boldsymbol{\theta}$ *is a p x 1 vector of parameters to be estimated, and* $\mathbf{w}$ *is an N x 1 colored noise vector with PDF* $N(0, C)$ *(***C*** $\neq \sigma^2 \mathrm{I}$), then the MVU is*

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

*(1bis)*

*and the covariance matrix of estimate is*

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

*(2bis)*

# Extension to non-white Gaussian noise:

***Proof:*** *The covariance matrix and its inverse are both positive semi-definite. Thus*

$$\mathbf{C}^{-1} = \mathbf{D}^T\mathbf{D}, \qquad \text{where} \qquad \mathbf{D} \in R^{NxN}$$

*A noise whitening operation can be performed. For that purpose lets compute the covariance of*

$$E\left[\left(\mathbf{D}w\right)\left(\mathbf{D}w\right)^T\right] = \mathbf{DCD}^T = \mathbf{DD}^{-1}\mathbf{D}^{T^{-1}}\mathbf{D}^T = I.$$

*If we define the new variable **x’** as*

$$\mathbf{x}' = \mathbf{Dx} = \mathbf{DH}\theta + \mathbf{Dw} = \mathbf{H}'\theta + \mathbf{w}'.$$

*Applying the usual solution to this linear model (transformed) results in*

$$\hat{\theta} = \left(\mathbf{H}'^T\mathbf{H}'\right)^{-1}\mathbf{H}'^T\mathbf{x}' = \left(\mathbf{H}^T\mathbf{D}^T\mathbf{DH}\right)^{-1}\mathbf{H}^T\mathbf{D}^T\mathbf{Dx} = \left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}$$

*For a covariance*

$$C_{\hat{\theta}} = \left(\mathbf{H}'^T\mathbf{H}'\right)^{-1} = \left(\mathbf{H}^T\mathbf{D}^T\mathbf{DH}'\right)^{-1} = \left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}'\right)^{-1}.$$

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

# *Sufficient Statistics:*

*General MVU Estimation:*

*Assume that the CRLB is not satisfied with equality!*

*There is no efficient estimator.*

*How do we find the MVU estimator (if it exists)?*

***Use the concept of Sufficient Statistics.***

*Example: To compute the value of a DC signal in noise, given n samples, i=0,…,N-1.*

*Consider*

$$S_1 = \{x[0], x[1], \ldots, x[N-1]\}$$

$$S_2 = \{x[0] + x[1], \ldots, x[N-1]\}$$

$$S_3 = \left\{\sum_{n=0}^{N-1} x[n]\right\}$$

*All sets are sufficient since the unknown parameter can be found.* $S_3$ *is the minimal one.*

# Sufficient Statistics:

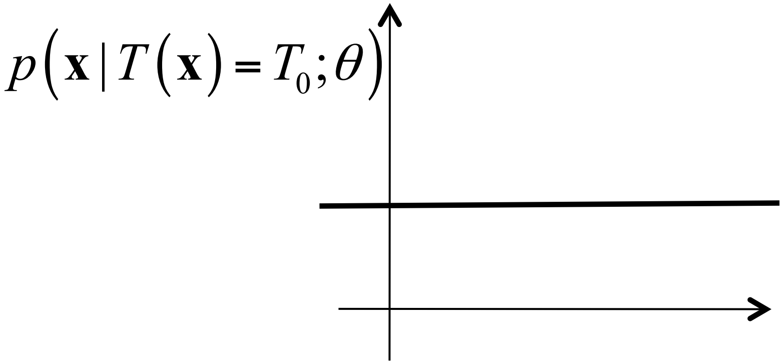
***Theorem 5.1 (Neyman-Fisher Factorization)*** *– If we can factor the PDF p (**x**;θ) as*

$$p(\mathbf{x};\theta) = g(T(\mathbf{x}),\theta)h(\mathbf{x}) \qquad (3)$$

*where $g(.)$ is a function depending on **x** only through $T(\mathbf{x})$ and $h\ (.)$ is a function depending only on **x**, then $T(\mathbf{x})$ is sufficient statistic for $\theta$. Conversely, if $T\ (\mathbf{x})$ is a sufficient statistic for $\theta$ then the PDF can be factored as in (3).*

*Proof outline (=>):*

- *p(**x**,T(**x**);θ) must have a minimum at **x**=**x₀**, denoted as T(**x₀**)=T₀;*
- *If **y**=g(**x**), for the vector random variable **x**,* $p(y) = \int p(\mathbf{x})\delta(y - g(\mathbf{x}))d\mathbf{x}.$



- *Knowledge of the value of a sufficient statistics makes the conditional PDF not to depend on the parameters*

# Sufficient Statistics:

*Proof outline (cont):*

*Using conditional probability definition:*

$$p\left(\mathbf{x}\,|\,T(\mathbf{x})=T_0;\theta\right)=\frac{p\left(\mathbf{x},T(\mathbf{x})=T_0;\theta\right)}{p\left(T(\mathbf{x})=T_0;\theta\right)}=\frac{p(\mathbf{x};\theta)\delta\left(T(\mathbf{x})-T_0\right)}{p\left(T(\mathbf{x})=T_0;\theta\right)}$$

$$=\frac{g\left(\mathbf{x},T(\mathbf{x})=T_0,\theta\right)h(\mathbf{x})\delta\left(T(\mathbf{x})-T_0\right)}{p\left(T(\mathbf{x})=T_0;\theta\right)}.$$

*Where the factorization was used in the last step. The denominator can be written as*

$$p\left(T(\mathbf{x})=T_0;\theta\right)=\int p(\mathbf{x};\theta)\delta\left(T(\mathbf{x})-T_0\right)d\mathbf{x}=$$

$$=\int g\left(T(\mathbf{x})=T_0,\theta\right)h(\mathbf{x})\delta\left(T(\mathbf{x})-T_0\right)d\mathbf{x}=g\left(T(\mathbf{x})=T_0,\theta\right)\int h(\mathbf{x})\delta\left(T(\mathbf{x})-T_0\right)d\mathbf{x}.$$

*The integral is zero in $R^n$ except over the surface where $T(\boldsymbol{x})=T_0$. where it is constant.*

$$p\left(\mathbf{x}\,|\,T(\mathbf{x})=T_0;\theta\right)=\frac{h(\mathbf{x})\delta\left(T(\mathbf{x})-T_0\right)}{\int h(\mathbf{x})\delta\left(T(\mathbf{x})-T_0\right)d\mathbf{x}},$$

*Which does not depend on $\theta$. Hence, we conclude that $T(\boldsymbol{x})$ is a sufficient statistic.* ■

# Sufficient Statistics:

*Proof outline (<=):*

*Consider the joint PDF*

$$p(\mathbf{x}, T(\mathbf{x}) = T_0;\theta) = p(\mathbf{x} \mid T(\mathbf{x}) = T_0;\theta)\, p(T(\mathbf{x}) = T_0;\theta) = p(\mathbf{x};\theta)\,\delta(T(\mathbf{x}) - T_0).$$

*Because T(**x**) is a sufficient statistic, the conditional PDF does not depend on θ. We can let*

$$p(\mathbf{x} \mid T(\mathbf{x}) = T_0) = w(\mathbf{x})\,\delta(T(\mathbf{x}) - T_0)$$

*Substituting in the previous expression*

$$p(\mathbf{x};\theta)\,\delta(T(\mathbf{x}) - T_0) = w(\mathbf{x})\,\delta(T(\mathbf{x}) - T_0)\, p(T(\mathbf{x}) = T_0;\theta)$$

*Setting w(x) to*

$$w(\mathbf{x}) = \frac{h(\mathbf{x})}{\int h(\mathbf{x})\,\delta(T(\mathbf{x}) - T_0)\,d\mathbf{x}},$$

*Allows one to write*

$$p(\mathbf{x};\theta)\,\delta(T(\mathbf{x}) - T_0) = \frac{h(\mathbf{x})\,\delta(T(\mathbf{x}) - T_0)}{\int h(\mathbf{x})\,\delta(T(\mathbf{x}) - T_0)\,d\mathbf{x}}\, p(T(\mathbf{x}) = T_0;\theta)$$

*Thus based on the factorization a sufficient statistic can be found*

$$p(\mathbf{x};\theta) = g(T(\mathbf{x}) = T_0;\theta)\, h(\mathbf{x})$$

■

DEM DEPARTAMENTO DE ENGENHARIA MECÂNICA TÉCNICO LISBOA

# Motivating Example: illustration

DC Level in WGN: Assuming that we have found the sufficient statistics, the ***Rao - Blackwell-Lehmann-Scheffe Theorem***can be used to find the MVU estimator in two different ways:

1) Find any unbiased estimator of A, say $\breve{A} = x\left[0\right]$, and determine $\hat{A} = E\left[\breve{A} \mid T\right]$ .
The expectation is taken with respect to $p\left(\breve{A} \mid T\right)$.

2) Find some function *g(.)* so that $\hat{A} = g\left(T\right)$is an unbiased estimator of *A.*

First approach:

Let $\breve{A} = x\left[0\right]$ and determine $\hat{A} = E\left[x\left[0\right] \mid \sum_{n=0}^{N-1} x\left[n\right]\right]$

We need auxiliary results for [x y]$^T$ a Gaussian random vector with mean $\mu$=[E[x] E[x]]$^T$

$$E\left[x \mid y\right] = E\left[x\right] + \frac{\operatorname{cov}\left(x, y\right)}{\operatorname{var}\left(y\right)}\left(y - E\left[y\right]\right)$$

(see Appendix 10A for details.)

# Motivating Example:

DC Level in WGN (cont.):

1) Find any unbiased estimator of A, say $\breve{A} = x[0]$ , and determine $\hat{A} = E\left[\breve{A} \mid T\right]$ .

The expectation is taken with respect to $p\left(\breve{A} \mid T\right)$.

Applying the previous results to *x=x[0]* and $y = \sum_{n=0}^{N-1} x[n]$

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x[0] \\ \sum_{n=0}^{N-1} x[n] \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 1 \end{bmatrix}}_{\mathbf{L}} \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix}$$

Hence the PDF of [x y]$^T$ is $N(\mu, C)$, where

$$\boldsymbol{\mu} = \mathbf{L}E[\mathbf{x}] = \mathbf{L}A\mathbf{1} = \begin{bmatrix} A \\ NA \end{bmatrix},$$

$$C = \sigma^2 \mathbf{L}\mathbf{L}^T = \sigma^2 \begin{bmatrix} 1 & 1 \\ 1 & N \end{bmatrix}.$$

# Motivating Example:

DC Level in WGN (cont.):

1) Find any unbiased estimator of A, say $\breve{A} = x[0]$ , and determine $\hat{A} = E[\breve{A} | T]$ .

The expectation is taken with respect to $p(\breve{A} | T)$.

Hence we have finally

$$\hat{A} = E[x | y] = A + \frac{\sigma^2}{N\sigma^2}\left(\sum_{n=0}^{N-1} x[n] - NA\right) = \frac{1}{N}\sum_{n=0}^{N-1} x[n].$$

Which is the MVU estimator. Usually this option is mathematically intractable.

2) Find some function *g(.)* so that $\hat{A} = g(T)$ is an unbiased estimator of *A.*

*We need to find some function* $\hat{A} = g\left(\sum_{n=0}^{N-1} x[n]\right)$ *so that it is an unbiased estimator.*

*That is the case of*

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

# RBLS Theorem:

***Definition:*** a statistic is complete if there is only one function of the statistic that is unbiased.

***Theorem 5.1 (Rao-Blackwell-Lehmann-Scheffe)*** *– If* $\breve{\theta}$ *is an unbiased estimator of* $\theta$ *and T(**x**) is a sufficient statistic for* $\theta$*, then* $\hat{\theta} = E\left[\breve{\theta} \mid T\left(\mathbf{x}\right)\right]$ *is*

*1. A valid estimator for* $\theta$

*2. Unbiased*

*3. Of lesser or equal variance than that of* $\breve{\theta}$ *, for all θ.*

*Additionally, if the sufficient statistic is complete, then* $\hat{\theta}$ *is the MVU estimator.*

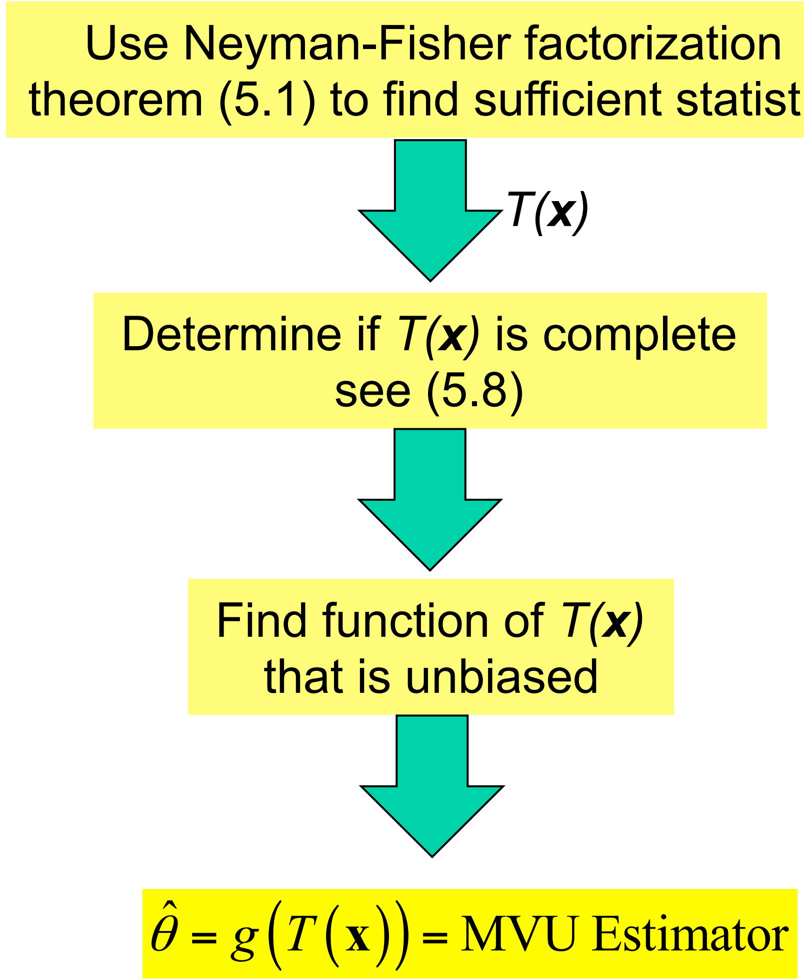
To validate that a statistic is complete is in general very difficult, (see examples 5.6 and 5.7). It must verify

$$\int_{-\infty}^{+\infty} v\left(T\right) p\left(T;\theta\right) dT = 0, \qquad \text{for all} \quad \theta. \qquad (5.8)$$

Only for the zero function and for *v(T).*

*Note:* - For an example of an incomplete statistic check Example 5.7

# *Methodology:*



Use Neyman-Fisher factorization theorem (5.1) to find sufficient statistic

*T*(***x***)

Determine if *T*(***x***) is complete see (5.8)

Find function of *T*(***x***) that is unbiased

$\hat{\theta} = g\left(T\left(\mathbf{x}\right)\right) = \text{MVU Estimator}$

# Example:

Mean of Uniform Noise:

Data model: *x[n]=w[n], n=0,1,…,N-1*

Where *w[n]* is IID noise with *PDF U[0,β], for β>0.*

We wish to find the MVU estimator for the mean *θ=β/2.*

The approach to find the CRLB can not be followed as the PDF does not satisfy the regularity conditions. A natural estimator is

$$\hat{\theta}=\frac{1}{N}\sum_{n=0}^{N-1}x\left[n\right], \qquad with \quad \operatorname{var}\left(\hat{\theta}\right)=\frac{1}{N}\operatorname{var}\left(x\left[n\right]\right)=\frac{\beta^2}{12N}.$$

To determine if the sample mean is the MVU we will follow the methodology previously presented.

# Example:

Lets define the unit step function:

$$u(x) = \begin{cases} 1 \quad for \quad x > 0 \\ 0 \quad for \quad x < 0 \end{cases} .$$

Then,

$$p(x[n];\beta) = \frac{1}{\beta}\left[u(x[n]) - u(x[n]-\beta)\right], \quad where \quad \beta = 2\theta.$$

and the PDF is

$$p(x[n];\beta) = \frac{1}{\beta^N}\prod_{n=0}^{N-1}\left[u(x[n]) - u(x[n]-\beta)\right] = \begin{cases} \frac{1}{\beta^N} & 0 < x[n] < \beta \quad n = 0,1,\ldots,N-1 \\ 0 & otherwise \end{cases} .$$

Alternative, we can write

$$p(x[n];\beta) = \begin{cases} \frac{1}{\beta^N} & \max(x[n]) < \beta, \min(x[n]) > 0 \\ 0 & otherwise \end{cases} ,$$

So that

$$p(x[n];\beta) = \frac{1}{\beta^N} u(\beta - \max(x[n])) u(\min(x[n]))$$

# Example:

Note that can be identified

$$p\left(x[n];\beta\right)=\underbrace{\frac{1}{\beta^{N}}u\left(\beta-\max\left(x[n]\right)\right)}_{g\left(T\left(x\right),\beta\right)}\quad\underbrace{u\left(\min\left(x[n]\right)\right)}_{h\left(x\right)}$$

By the Neyman-Fisher factorization theorem, *T(**x**)=max(x[n])* is a sufficient statistic for θ. Furthermore, it can be shown that the sufficient statistic is complete. We need next to find a function of *T(**x**)* that is not biased (denominated as order statistics). Lets write the cumulative distribution function

$$\Pr\left\{T\le\xi\right\}=\Pr\left\{x[0]\le\xi,x[1]\le\xi,\ldots,x[N-1]\le\xi,\right\}=\prod_{n=0}^{N-1}\Pr\left\{x[n]\le\xi\right\}=\Pr\left\{x[n]\le\xi\right\}^{N}.$$

The PDF follows as

$$p_{T}\left(\xi\right)=\frac{d\Pr\left\{T\le\xi\right\}}{d\xi}=N\Pr\left\{x[n]\le\xi\right\}^{N-1}\frac{d\Pr\left\{x[n]\le\xi\right\}}{d\xi}.$$

# Example:

But
$$\frac{d\Pr\left\{x[n]\le\xi\right\}}{d\xi}=p_{x[n]}(\xi)=\begin{cases}\frac{1}{\beta} & 0<\xi<\beta\\ 0 & otherwise\end{cases},$$

Integrating we obtain

$$p_T(\xi)=\begin{cases}0 & \xi<0\\ N\left(\frac{\xi}{\beta}\right)^{N-1}\frac{1}{\beta} & 0<\xi<\beta\\ 0 & \xi>\beta\end{cases},\quad and\quad E[T]=\int_0^{\beta}\xi N\left(\frac{\xi}{\beta}\right)^{N-1}\frac{1}{\beta}d\xi$$

From where it results

$$E[T]=\frac{N}{N+1}\beta=\frac{2N}{N+1}\theta,\qquad thus\quad \hat{\theta}=\frac{N+1}{2N}T$$ makes the expected value unbiased.

The MVU estimator is
$$\hat{\theta}=\frac{N+1}{2N}\max\left(x[n]\right)$$

with a variance... $\operatorname{var}\left(\hat{\theta}\right)=\frac{\beta^2}{4N(N+2)}<<\frac{\beta^2}{12N}$ (sample mean var) for large N!

# Bibliography:

## Further reading

- Thomas Kailath, ***Linear Systems,*** Prentice Hall, 1980.

- Thomas Kailath, Ali Sayed, and Babak Hassibi, ***Linear Estimation,*** Prentice Hall, 2000.

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

- J. Bibby, H. Toutenburg, ***Prediction and Improved Estimation in Linear Models,*** John Wiley, 1977.

- C.Rao, ***Linear Statistical Inference and Its Applications,*** John Wiley, 1973.

# Advanced Control Systems
# Detection, Estimation, and Filtering

**Graduate Course on the**
**MEng PhD Program**
**Spring 2012/2013**

## Chapter 5
## Best Linear Unbiased Estimators

**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)**

# Syllabus:

Classical Estimation Theory

…

Chap. 4 - ***Linear Models in the Presence of Stochastic Signals*** [1 week]

Stationary and transient analysis; White Gaussian noise and linear systems; Examples;

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;

continues…

# *An alternative strategy:*

FACT:

**It occurs that the MVU estimator, if it exists, can not be found.**

e.g. the PDF for the data is not known, the user would not like to assume a model for the PDF, or the problem can be mathematically untreatable.

**An alternative strategy can be pursued is to study the class of**

***Best Linear Unbiased Estimators***

**Only suboptimal performance can be achieved.**

The performance degradation, relative to the MVU estimator, is unknown but the resulting performance can be enough for the problem at hand.

# BLUE structure:

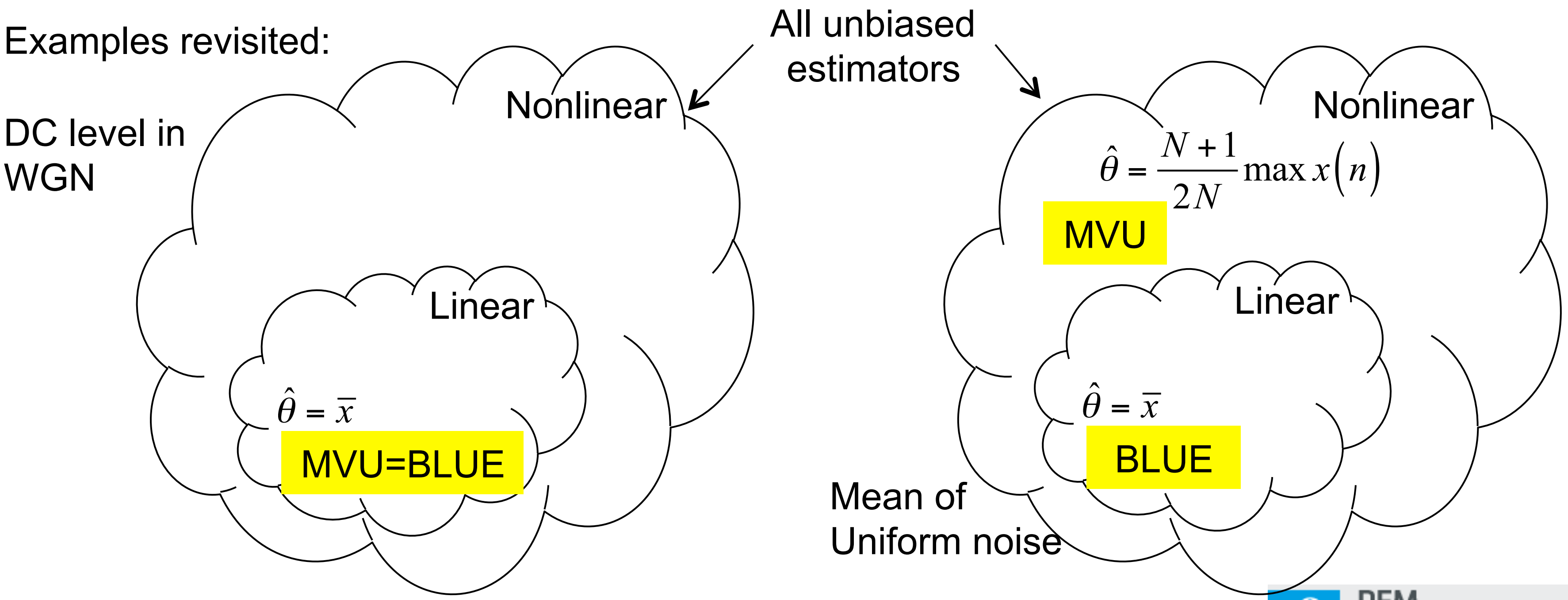
The Best Linear Unbiased Estimator consists of restrict the estimator to be a linear function of the data, i.e.

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

where the $a_n$'s are constants to be determined.

Optimality in general is lost.

Examples revisited:



# Finding the BLUE:

To find the BLUE we constrain the estimator

- to be linear $\hat{\theta} = \sum_{n=0}^{N-1} a_n x[n]$
- to be unbiased $E[\hat{\theta}] = \sum_{n=0}^{N-1} a_n E[x[n]] = \theta \qquad (6.2)$
- to minimize its variance $\operatorname{var}(\hat{\theta}) = E\left[\left(\sum_{n=0}^{N-1} a_n x[n] - E\left[\sum_{n=0}^{N-1} a_n x[n]\right]\right)^2\right]$

Defining ***a***=*[a$_0$ a$_1$ … a$_{N-1}$]$^T$* and ***x***=*[x$_0$ x$_1$ … x$_{N-1}$]$^T$* this last expression can be simplified:

$$\operatorname{var}(\hat{\theta}) = E\left[\left(\mathbf{a}^T\mathbf{x} - \mathbf{a}^T E[\mathbf{x}]\right)^2\right] = E\left[\left(\mathbf{a}^T\left(\mathbf{x} - E[\mathbf{x}]\right)\right)^2\right] =$$

$$= E\left[\mathbf{a}^T\left(\mathbf{x} - E[\mathbf{x}]\right)\left(\mathbf{x} - E[\mathbf{x}]\right)^T\mathbf{a}\right] = \mathbf{a}^T C\mathbf{a}. \qquad (6.3)$$

The problem of finding the BLUE can be stated as, for $\mathbf{a} \in R^N$

$$\begin{array}{ll} \min & \mathbf{a}^T C\mathbf{a} \\ \text{subject to} & \mathbf{a}^T E[\mathbf{x}] = \theta \end{array}$$

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

# Finding the BLUE:

Given the scalar parameter *θ,* the expected value of the samples can be assumed as

*E[x[n]]=s[n]θ,*

where *s[n]* is known.

$$\sum_{n=0}^{N-1} a_n E\left[x\left[n\right]\right] = \sum_{n=0}^{N-1} a_n s\left[n\right]\theta = \theta, \qquad \text{(from 6.2)}$$

Thus the previous problem can be stated as

$$\begin{array}{cc} \min & \mathbf{a}^T\mathbf{C}\mathbf{a} \\ \mathbf{a} \in R^N & \text{s.t. } \mathbf{a}^T\mathbf{s} = 1 \end{array}$$

The method of Lagrangian multipliers can be used to solve this problem. Define the Lagrangian function as

$$J\left(\mathbf{a},\lambda\right) = \mathbf{a}^T\mathbf{C}\mathbf{a} + \lambda\left(\mathbf{a}^T\mathbf{s} - 1\right), \qquad \lambda \in R$$

The gradient of $J$ relative to **a** is

$$\frac{\partial J\left(\mathbf{a},\lambda\right)}{\partial \mathbf{a}} = 2\mathbf{C}\mathbf{a} + \lambda\mathbf{s} = 0$$

# Finding the BLUE:

Solving for **a** produces

$$\mathbf{a} = -\frac{\lambda}{2}\mathbf{C}^{-1}\mathbf{s}$$

Using the constrain as

$$\mathbf{a}^T\mathbf{s} = -\frac{\lambda}{2}\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s} = 1, \quad \text{or} \quad \lambda = -\frac{2}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}} \quad .$$

Finally, the solution is

$$\mathbf{a}_{opt} = -\frac{\mathbf{C}^{-1}\mathbf{s}}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}}, \qquad \text{with a variance} \quad \operatorname{var}\left(\hat{\theta}\right) = \mathbf{a}_{opt}^T\mathbf{C}\mathbf{a}_{opt} = \frac{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{C}\mathbf{C}^{-1}\mathbf{s}}{\left(\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}\right)^2} = \frac{1}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}}.$$

Taking into account that *E[**x**]=**s**θ,* finally the estimator

$$\hat{\theta} = \mathbf{a}_{opt}^T\mathbf{x} = \frac{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{x}}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}}, \qquad \text{Its expected value is} \quad E\left(\hat{\theta}\right) = \frac{\mathbf{s}^T\mathbf{C}^{-1}E\left[\mathbf{x}\right]}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}} = \frac{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}\theta}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}} = \theta!$$

Thus it is unbiased, as required.

# Example:

Example (DC level in white ~~Gaussian~~ noise revisited):

Model of signal: $x[n] = A + w[n], \qquad n = 0,\ldots,N-1$

Where $w[n]$ is zero mean white noise with variance σ$^2$ (and an unspecified PDF), the problem is to estimate $A$.

Because $E[x[n]]=A$, we have $\mathbf{s}=\mathbf{1}$, *where* $\mathbf{1}=[1\ 1\ \ldots\ 1]^T$.

The BLUE is $\hat{A} = \dfrac{\mathbf{1}^T \frac{1}{\sigma^2}\mathbf{I}\mathbf{x}}{\mathbf{1}^T \frac{1}{\sigma^2}\mathbf{I}\mathbf{1}} = \dfrac{1}{N}\sum_{N=0}^{N-1} x[n] = \bar{x}$, and the variance is $\operatorname{var}(\hat{A}) = \dfrac{1}{\mathbf{1}^T \frac{1}{\sigma^2}\mathbf{I}\mathbf{1}} = \dfrac{\sigma^2}{N}$.

Hence the sample mean is the BLUE independent of the PDF of data. It is the MVU estimator for the Gaussian case.

And in general: is it optimal?...

# Example:

Example (DC level UNCORRELATED noise):

Model of signal: $x[n] = A + w[n], \qquad n = 0,\ldots,N-1$

Where $w[n]$ is zero mean uncorrelated noise with $var(w[n])=\sigma^2$. Once again, the problem is to estimate $A$.

We have again $\boldsymbol{s=1}$, and $\boldsymbol{C}=diag(\sigma_0^2\ \sigma_1^2 \ldots \sigma_{N-1}^2)$, and $\boldsymbol{C}^{-1}=diag(1/\sigma_0^2\ 1/\sigma_1^2 \ldots 1/\sigma_{N-1}^2)$..

The BLUE is

$$\hat{A} = \frac{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{x}}{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{1}} = \frac{\sum_{N=0}^{N-1}\frac{x[n]}{\sigma_n^2}}{\sum_{N=0}^{N-1}\frac{1}{\sigma_n^2}}$$

and the variance is

$$\operatorname{var}(\hat{A}) = \frac{1}{\sum_{N=0}^{N-1}\frac{1}{\sigma_n^2}}$$

The BLUE weights those samples more heavily with smallest variances, in an attempt to equalize the noise contribution from each sample…

Is it optimal? In what cases?...

# *Extending BLUE to a Vector Parameter:*

To find the BLUE for a $p \times 1$ vector parameter, we constrain the estimator

- to be linear $\hat{\theta}_i = \sum_{n=0}^{N-1} a_{in} x[n] \qquad i = 1,2,\ldots,p \quad \text{or} \quad \hat{\theta} = \mathbf{A}\mathbf{x}, \qquad \mathbf{A} \in R^{p \times N}$
- to be unbiased $E\left[\hat{\boldsymbol{\theta}}\right] = \mathbf{A}E\left[\hat{\boldsymbol{\theta}}\right] = \boldsymbol{\theta}$
- to minimize its variance $\operatorname{var}\left(\hat{\theta}_i\right) = E\left[\left(\sum_{n=0}^{N-1} a_{in} x[n] - E\left[\sum_{n=0}^{N-1} a i_n x[n]\right]\right)^2\right]$

The problem of finding the BLUE can be stated as, for

$$\begin{aligned} &\min \qquad \operatorname{var}\left(\hat{\theta}_i\right) = \mathbf{a}_i^T C \mathbf{a}_i \\ &\text{subject to} \qquad \mathbf{a}_i^T E[\mathbf{x}] = \theta \end{aligned}$$

where

$$E[\mathbf{x}] = \mathbf{H}\boldsymbol{\theta}.$$

# Extension to non-white Gaussian noise:

***Theorem 6.1(Gauss-Markov Theorem)*** *– If the data observed are of the general linear model form*

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

*where* **H** *is a known N x p observation matrix (with N>p) and rank p,* **x** *is a N x 1 vector of observations,* **θ** *is a p x 1 vector of parameters to be estimated, and* **w** *is an N x 1 noise vector with zero mean and covariance* **C** *(for an arbitrary PDF), then the BLUE is*

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

*and the covariance matrix of estimate is*

$$C_{\hat{\theta}} = \left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1}. \qquad \textit{(2bis)}$$

# Bibliography:

## Further reading

- Thomas Kailath, ***Linear Systems,*** Prentice Hall, 1980.

- Thomas Kailath, Ali Sayed, and Babak Hassibi, ***Linear Estimation,*** Prentice Hall, 2000.

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

- J. Bibby, H. Toutenburg, ***Prediction and Improved Estimation in Linear Models,*** John Wiley, 1977.

- C.Rao, ***Linear Statistical Inference and Its Applications,*** John Wiley, 1973.

# 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)

# 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: $x\left[n\right] = A + w\left[n\right], \qquad n = 0,\dots,N-1$

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:

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

The derivative of the log-likelihood function is

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

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

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

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

# An example:

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

So a sufficient statistics is

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

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

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

We have that

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

# 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.

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

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.

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

# 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

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

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

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

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.

# 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.**

# 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 $\theta_1$=$\theta_2$.*

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

# 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.*

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

# Properties of MLE:

*Proof outline:*

*This relation can therefore be approximately written as*

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

*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 $\theta$, 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*

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

*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*

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

# 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?...**

# 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 σ² 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…*

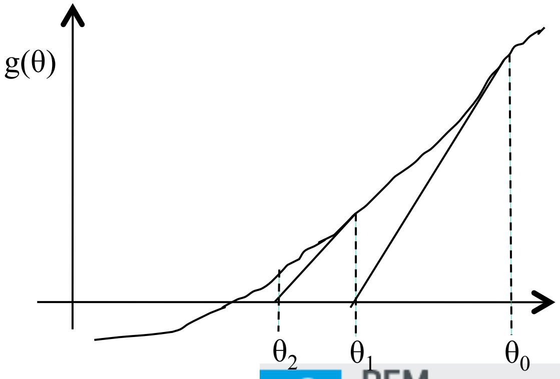
$$\frac{\partial \ln p\left(\mathbf{x};\theta\right)}{\partial \theta} = g\left(\theta\right) = 0$$

*For instance, one of the most basic method, is the Newton-Raphson method. From an initial guess $\Theta_0$, and from a Taylor series expansion results*

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

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

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



DEM DEPARTAMENTO DE ENGENHARIA MECÂNICA TÉCNICO LISBOA

# Motivating example:

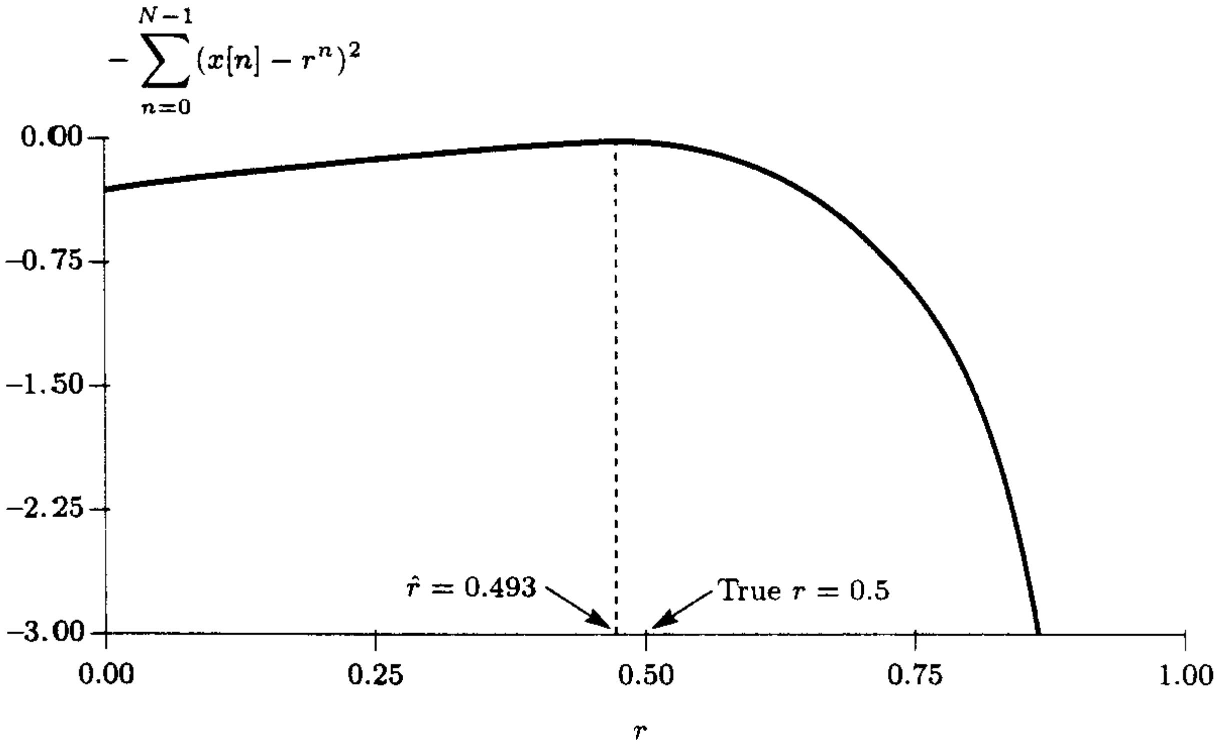
Example (Exponential  in white Gaussian noise):

Computer simulation

N=50, r=0.5, and $\sigma^2$=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 |



# 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)*

# Invariance Property:

***Theorem 7.5 (Optimality of the MLE for the Linear Model)*** *– If the observed data **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*

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

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

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

# 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.*

# Maximum Likelihood Principle:

***Theorem 7.1 (Asymptotic Properties of the MLE – Vector Parameter)*** *– If the PDF $p(\mathbf{x};\boldsymbol{\theta})$ 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.

# Advanced Control Systems Detection, Estimation, and Filtering

Graduate Course on the
MEng PhD Program
Spring 2012/2013

## Chapter 7
## Least Squares

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)

# *Syllabus:*

## Classical Estimation Theory

…

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]

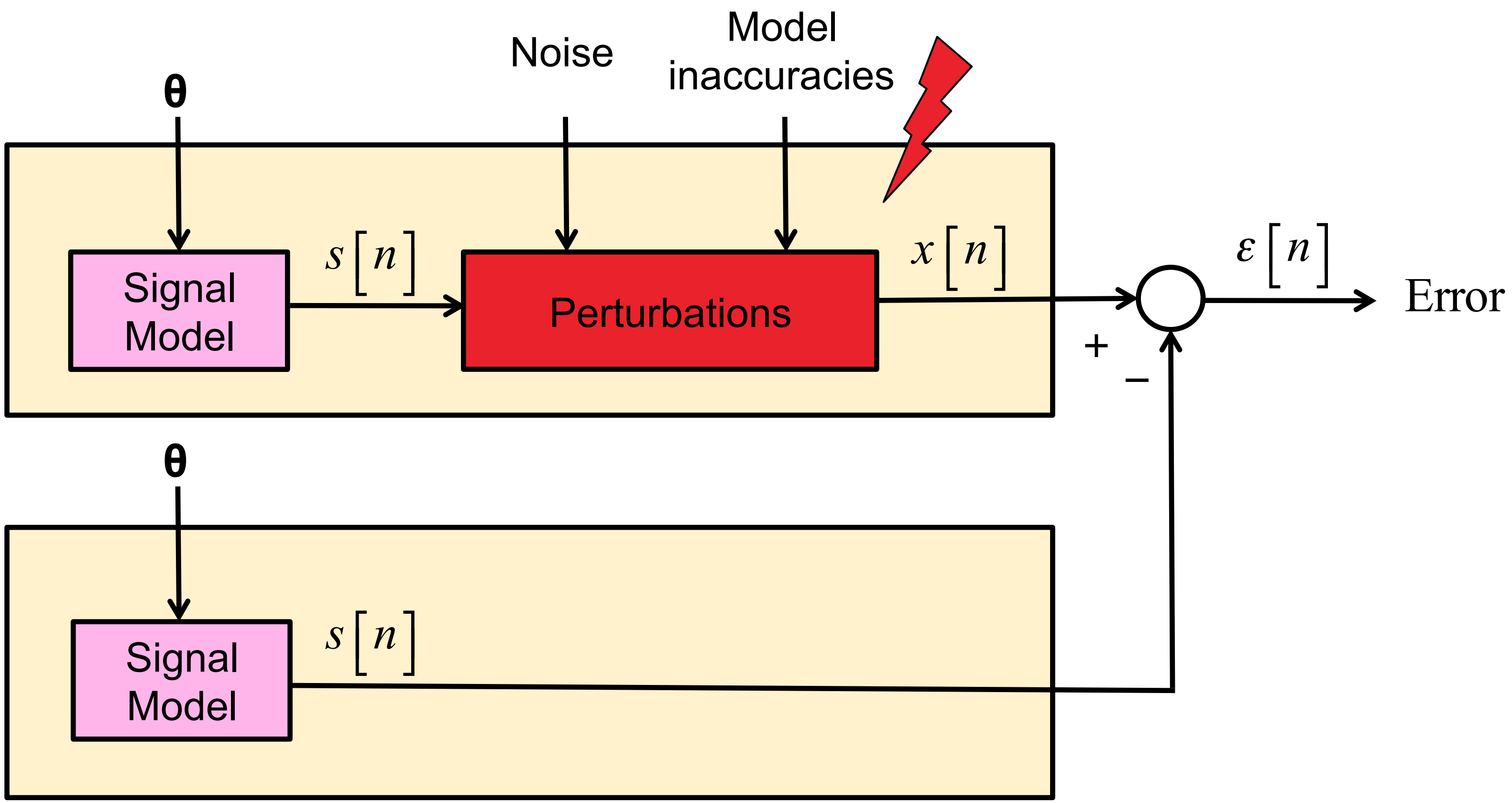
The least squares approach; Linear and nonlinear least squares; Geometric interpretation; Constrained least squares;  Examples;

Chap. 8 – ***Bayesian Estimation*** [1 week]

Philosophy and estimator design; Prior knowledge; Bayesian linear model; Bayesian estimation on the presence of Gaussian pdfs; Minimum Mean Square Estimators;

continues…

# Least Squares Approach:



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

# Least Squares Approach:

The least squares estimator (LSE) is obtained minimizing the LS error criterion

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

where the dependency on θ is via s[n].

Note:

No probabilistic assumptions have been made about the data x[n];

Method valid both for Gaussian and for non-Gaussian disturbances;

Performance optimality of the LSE can not be guaranteed;

Method applied when:

a precise statistical characterization of the data is unknown;

optimal estimator can not be found;

…

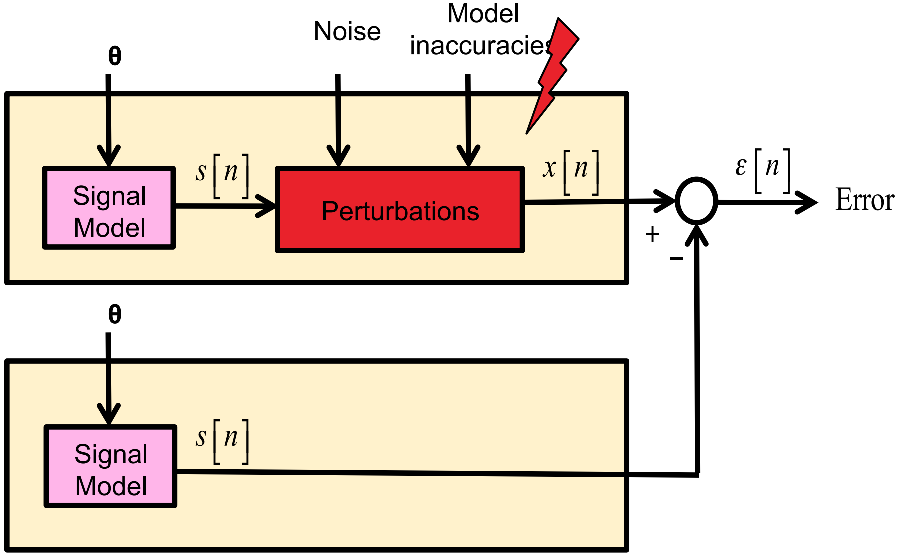
# Linear Least Squares:

The least squares approach for a **scalar parameter**, we must assume

$$s[n] = \theta h[n].$$

The criterion to minimize is

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



It is immediate that

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

With a solution given by

$$\hat{\theta} = \frac{\sum_{n=0}^{N-1} x[n]h[n]}{\sum_{n=0}^{N-1} h^2[n]}.$$

Thus the minimum cost of the criterion verifies

$$0 < J_{\min}(\theta) = \sum_{n=0}^{N-1} x^2[n] - \frac{\left(\sum_{n=0}^{N-1} x[n]h[n]\right)^2}{\sum_{n=0}^{N-1} h^2[n]} < \sum_{n=0}^{N-1} x^2[n].$$

# Linear Least Squares:

The extension of the least squares approach for a **vector parameter** is immediate.

For the signal ***s*** *=[s[0] s[1] … s[N-1]]*
The criterion to minimize is

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1}\left(x[n]-s[n]\right)^2 = \left(\mathbf{x}-\mathbf{H}\boldsymbol{\theta}\right)^T\left(\mathbf{x}-\mathbf{H}\boldsymbol{\theta}\right) = \mathbf{x}^T\mathbf{x}-2\mathbf{x}^T\mathbf{H}\boldsymbol{\theta}+\boldsymbol{\theta}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\theta}.$$

The gradient is

$$\frac{\partial J(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = -2\mathbf{H}^T\mathbf{x}+2\mathbf{H}^T\mathbf{H}\boldsymbol{\theta}.$$

With a solution given by

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

The minimum cost of the criterion verifies

$$0 < J_{\min}(\theta) = \mathbf{x}^T\mathbf{x}-\mathbf{x}^T\mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x} < \mathbf{x}^T\mathbf{x}.$$

# Geometrical Interpretation:

Note that the solution obtained

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

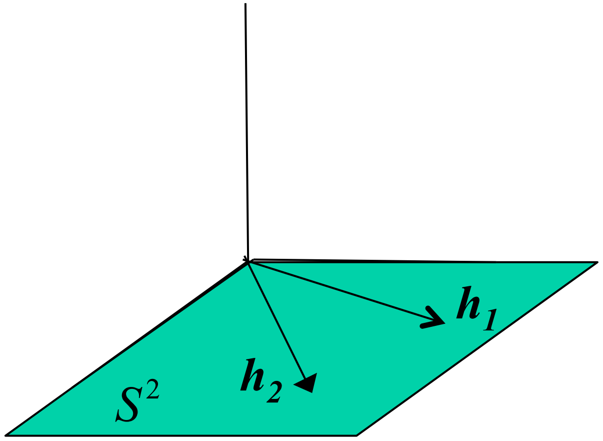
can be rewritten as

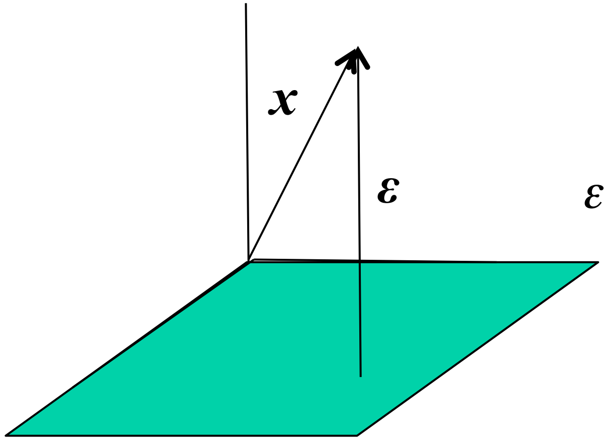
$$\left(\mathbf{H}^T\mathbf{H}\right)\theta = \left(\mathbf{H}^T\mathbf{H}\right)\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}$$

$$\left(\mathbf{H}^T\mathbf{H}\right)\theta = \mathbf{H}^T\mathbf{x}$$

$$\mathbf{H}^T\left(\mathbf{H}\theta - \mathbf{x}\right) = 0$$

Denoting as the error vector $\varepsilon = \mathbf{H}\theta - \mathbf{x}$, the previous expression can be interpreted as that the error vector must be orthogonal to the columns of ***H*.**





$\varepsilon \perp S^2$

# Extensions to Least Squares:

Other extensions of the least squares approach are also very popular

Weighted Least Squares:

criterion $J_W(\theta) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{W}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})$

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

minimum $0 < J_{\min}(\theta) = \mathbf{x}^T\left(\mathbf{W} - \mathbf{W}\mathbf{H}(\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}\right)\mathbf{x} < \mathbf{x}^T\mathbf{W}\mathbf{x}.$

$\mathbf{W}$ can be set as the inverse covariance matrix, leading to an optimal solution in the case of correlated Gaussian noise.

Order-recursive Least Squares (see pp. 232)

same criterion but the observation and parameter matrices vary their length

$$\mathbf{H}_{k+1} = \begin{bmatrix} \mathbf{H}_k & h_{k+1} \end{bmatrix} = \begin{bmatrix} N \text{ x } k & N \text{ x } 1 \end{bmatrix}$$

# Extensions to Least Squares:

Order-recursive Least Squares (cont.)

solution

$$\hat{\theta}_{k+1} = \begin{bmatrix} \hat{\theta}_k - \dfrac{\left(\mathbf{H}_k^T\mathbf{H}_k\right)^{-1}\mathbf{H}_k^T\mathbf{h}_{k+1}\mathbf{h}_{k+1}^T\mathbf{P}_k^{\perp}\mathbf{x}}{\mathbf{h}_{k+1}^T\mathbf{P}_k^{\perp}\mathbf{h}_{k+1}} \\ \dfrac{\mathbf{h}_{k+1}^T\mathbf{P}_k^{\perp}\mathbf{x}}{\mathbf{h}_{k+1}^T\mathbf{P}_k^{\perp}\mathbf{h}_{k+1}} \end{bmatrix} = \begin{bmatrix} k \text{ x } 1 \\ 1 \text{ x } 1 \end{bmatrix}$$

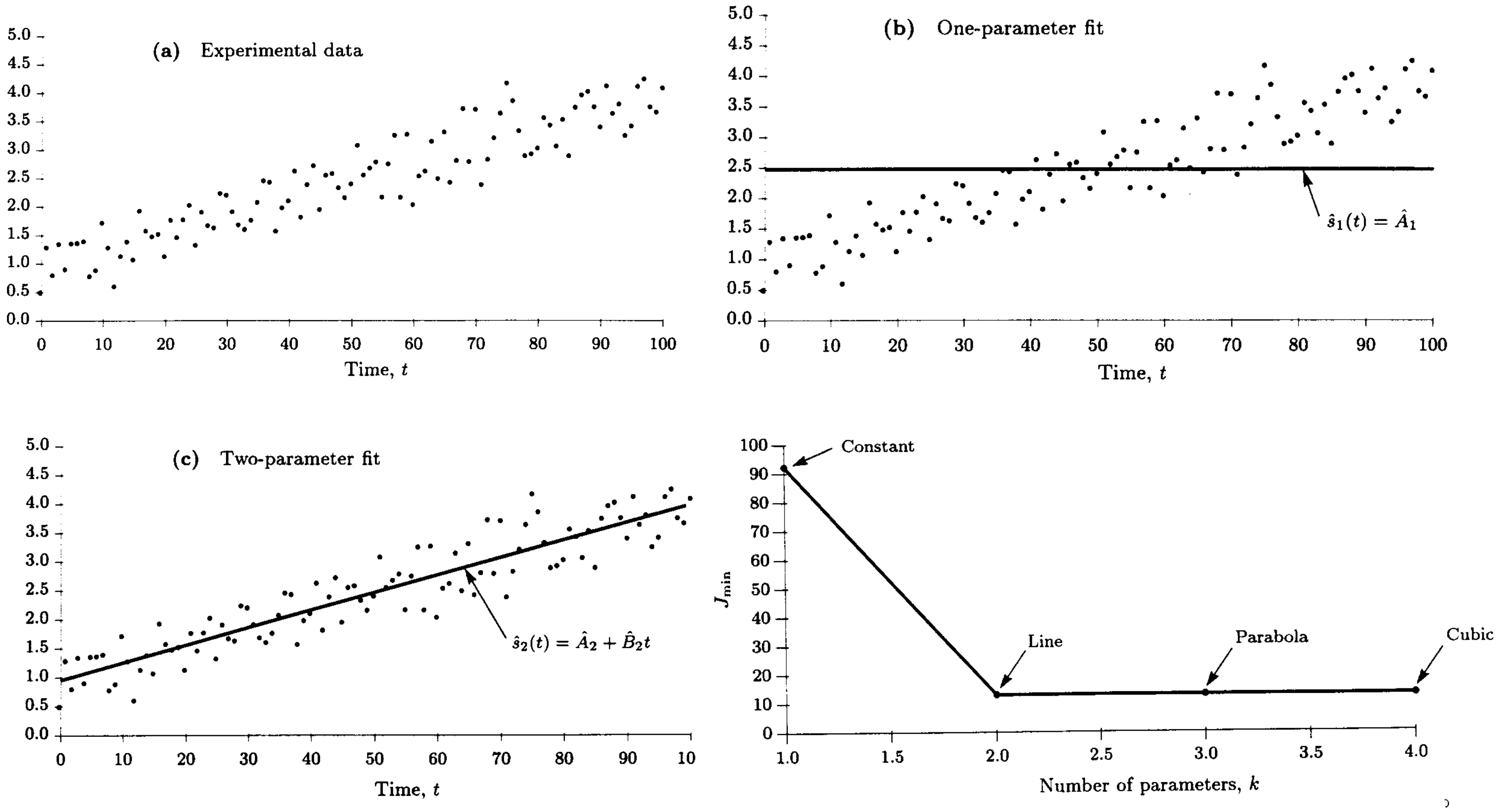
where $\mathbf{P}_k^{\perp} = I - \mathbf{H}_k\left(\mathbf{H}_k^T\mathbf{H}_k\right)^{-1}\mathbf{H}_k^T$

minimum $J_{\min}\left(\theta_{k+1}\right) = J_{\min}\left(\theta_k\right) - \dfrac{\left(\mathbf{h}_{k+1}^T\mathbf{P}_k^{\perp}\mathbf{x}\right)^2}{\mathbf{h}_{k+1}^T\mathbf{P}_k^{\perp}\mathbf{h}_{k+1}}$

Example:
Line fitting

$$s_1\left[n\right] = A_1 \qquad s_2\left[n\right] = A_2 + B_c n \qquad H_1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \qquad H_2 = \begin{bmatrix} H_1 & \begin{matrix} 0 \\ 1 \\ \vdots \\ N-1 \end{matrix} \end{bmatrix}$$

# Example:



(a) Experimental data
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
10
20
30
40
50
60
70
80
90
100
Time, $t$
(b) One-parameter fit
$\hat{s}_1(t) = \hat{A}_1$
Time, $t$
(c) Two-parameter fit
$\hat{s}_2(t) = \hat{A}_2 + \hat{B}_2 t$
Time, $t$
$J_{\min}$
100
90
80
70
60
50
40
30
20
10
0
Constant
Line
Parabola
Cubic
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Number of parameters, $k$

# Sequential Least Squares:

In many estimation, detection, or identification problems data are obtained as samples of the output of a process.

It would be advantageous that the least squares solution could be written as a recursive solution.

Lets revisit our old DC level in Gaussian noise example:

At time N-1, the data set available is $\boldsymbol{x}=[x[0]\ x[1]\ \ldots\ x[N\text{-}1]]$ and the MVU estimator solution is given by

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

If a new sample is obtained, i.e. x[n] is available, the estimator is given by

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

That can be rewritten as

$$\hat{A}[N]=\hat{A}[N-1]+\frac{1}{N+1}\left(x[N]-\hat{A}[N-1]\right).$$

Much remains to be said, see next chapters…

# Sequential Least Squares:

$$\hat{A}[N] = \hat{A}[N-1] + \frac{1}{N+1}\left(x[N] - \hat{A}[N-1]\right)$$

Recursive solution

Correction term, reflecting that with more one sample more is known on the parameter.

The gain is decreasing thus preserving a memory on the past samples.

The value of the criterion can also be written recursively, i.e.

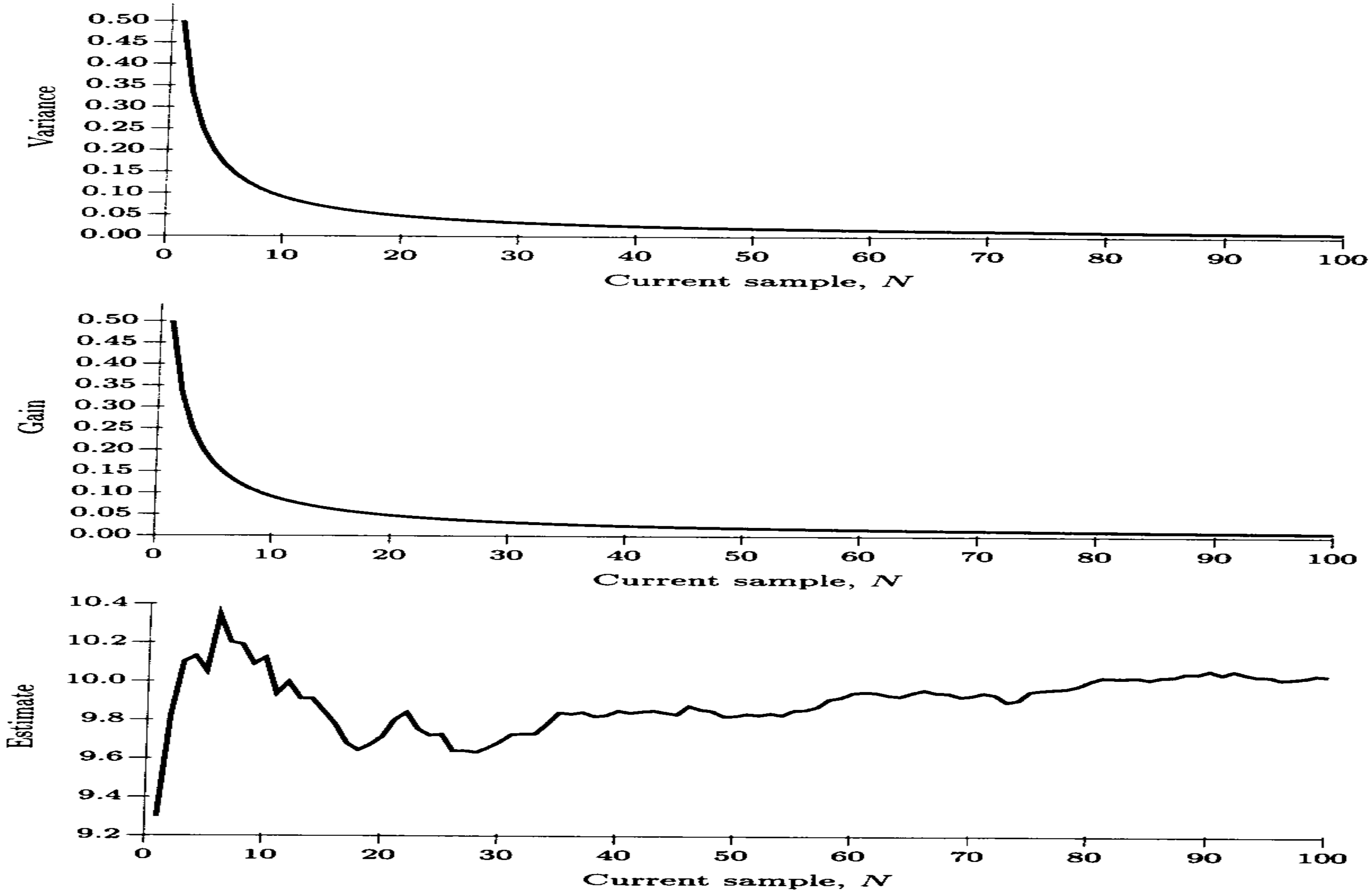
$$J_{\min}(N) = J_{\min}(N-1) + \frac{N}{N+1}\left(x[N] - \hat{A}[N-1]\right)^2$$

Seems a paradox, but if our fitting is parfait does not increases…

More points to be fitted with the same number of parameters.

It is an OPTIMAL solution!

# Sequential Least Squares:



# Sequential Least Squares:

The optimal solution, in the case where a Gaussian noise occurs, with time varying variance

Signal Model $x[n]=\mathbf{h}[n]\boldsymbol{\theta},$ $n=0,\ldots,N\text{-}1,\ldots$

Estimator Update:

$$\hat{\boldsymbol{\theta}}\left[n\right]=\hat{\boldsymbol{\theta}}\left[n-1\right]+\mathbf{K}\left[n\right]\left(x\left[N\right]-h^{T}\left[n\right]\hat{\boldsymbol{\theta}}\left[n-1\right]\right)$$

Where

$$\mathbf{K}\left[n\right]=\frac{\Sigma\left[n-1\right]h\left[n\right]}{\sigma_{n}^{2}+h^{T}\left[n\right]\Sigma\left[n-1\right]h\left[n\right]}$$

Covariance Update:

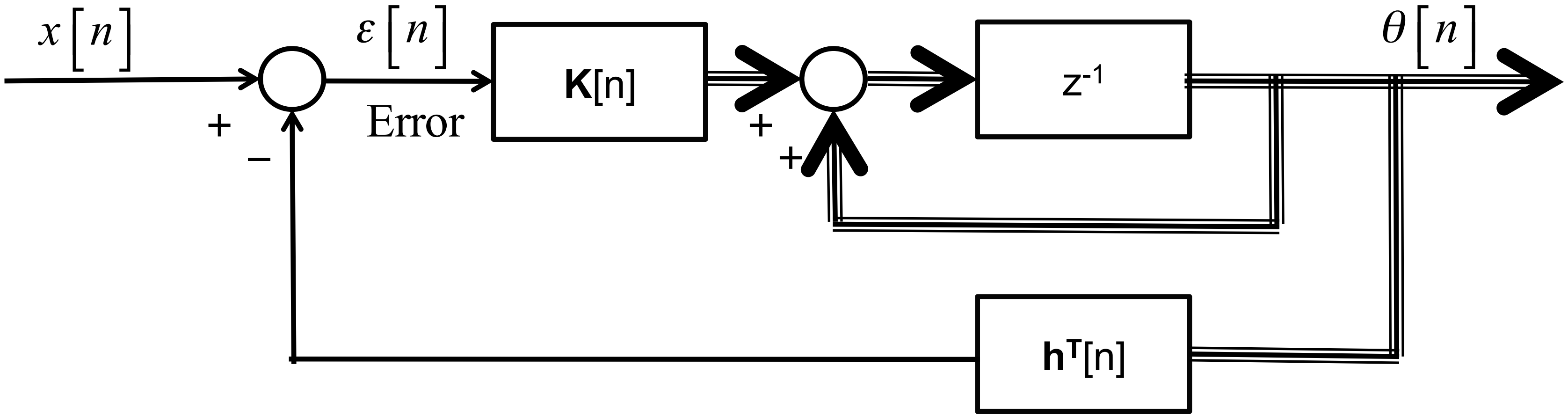
$$\Sigma\left[n\right]=\left(\mathbf{I}-\mathbf{K}\left[n\right]h^{T}\left[n\right]\right)\Sigma\left[n-1\right]$$

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

# Sequential Least Squares:

The signal model and the parameter estimation problem can be interpreted resorting to the dynamic model

$$
\begin{aligned}
\theta[n+1] &= \theta[n] \\
x[n] &= \mathbf{h}^T[n]\theta[n] + w[n]
\end{aligned}
$$



# Constrained Least Squares:

This alternative method can be very useful if the problem at hand verifies some properties.

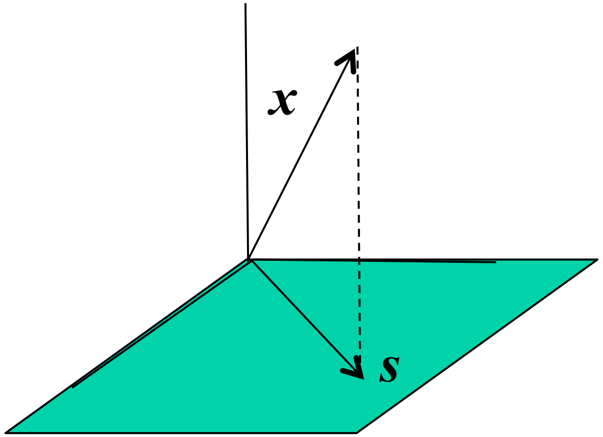
criterion $$J_C(\theta) = (\mathbf{x} - \mathbf{H}\theta)^T (\mathbf{x} - \mathbf{H}\theta)$$

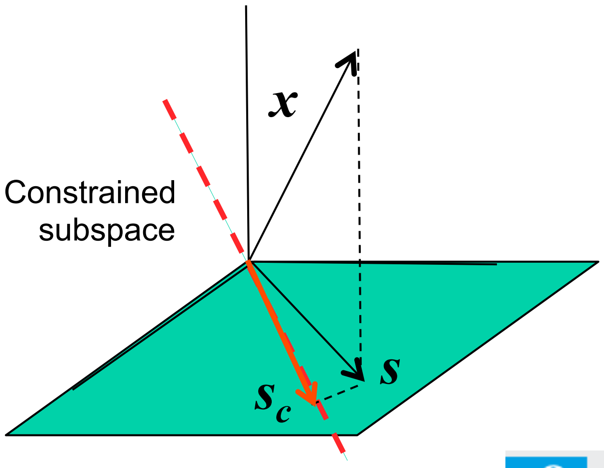
$$s.t. \qquad \mathbf{A}\theta = \mathbf{b}$$

solution $$\hat{\theta}_C = \hat{\theta} - \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{A}^T\left(\mathbf{A}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{A}^T\right)^{-1}\left(\mathbf{A}\hat{\theta} - \mathbf{b}\right)$$

The constrained LSE is a corrected version of the unconstrained LSE.

It can also be interpreted as the constrained signal estimate to be the projection of the unconstrained solution onto the constrained subspace.

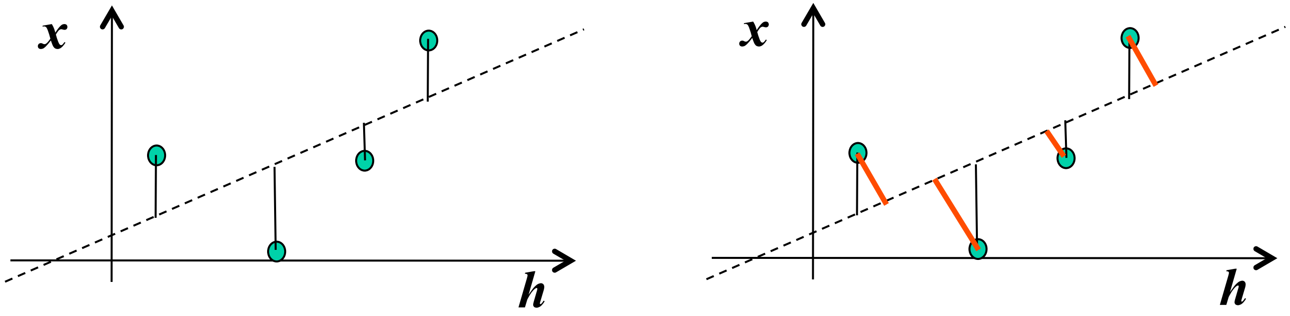




# Extensions to Least Squares:

Other extensions:

Total Least Squares (errors in variables, or orthogonal regression)



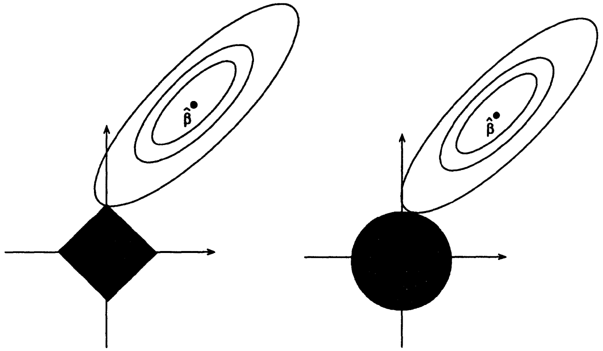
When could also be errors in the independent variables.

Lasso – Least Absolute Shrinkage and Selection Operator

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

$$s.t. \quad \sum_j |\theta| \leq t, \qquad \text{with} \quad t > 0$$

solution $$\hat{\theta} = \left(\mathbf{H}^T\mathbf{H} + \lambda \mathbf{W}^{-}\right)^{-1} \mathbf{H}^T\mathbf{x}$$



$\mathbf{W}$ diagonal matrix with elements $|\hat{\theta}_i|$, and $\mathbf{W}^{-}$ is the generalized inverse.

# Nonlinear Least Squares:

In general the signal model is

model $$\mathbf{x} = s\left(\boldsymbol{\theta}\right)^{T} + \mathbf{w}$$

where s() is in general a nonlinear function of the unknown parameters. The criterion to be minimized can be written as (if a quadratic error is selected)

criterion $$J\left(\boldsymbol{\theta}\right) = \left(\mathbf{x} - \mathbf{s}\left(\boldsymbol{\theta}\right)\right)^{T}\left(\mathbf{x} - \mathbf{s}\left(\boldsymbol{\theta}\right)\right)$$

termed also as nonlinear regression problem, in statistics.

Solution is general is not available, except if resorting to numerical methods.

Two methods than can reduce the complexity can be identified:

1 – Transformation of parameters;

2 – Separability of parameters;

# Nonlinear Least Squares:

## Transformation of parameters

We seek a one-to-one transformation that produces a linear signal model in the new space:

$$\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta})$$

Where **g()** is a p-dimensional function of the unknown parameters, with inverse:

$$\mathbf{s}(\boldsymbol{\theta}(\boldsymbol{\alpha})) = \mathbf{s}(\mathbf{g}^{-1}(\boldsymbol{\alpha})) = \mathbf{H}\boldsymbol{\alpha}.$$

Then the solution is

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

The transformation **g()**, if it exists, is usually quite difficult.

Only a few nonlinear least squares problems may be solved in this manner.

# Nonlinear Least Squares:

## Separability of parameters

Assume that the model is nonlinear but still is linear in some of the parameters. Thus

$$\mathbf{s} = \mathbf{H}(\alpha)\beta$$

Where

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} (p-q)\times 1 \\ q \times 1 \end{bmatrix}$$

The criterion

$$J(\alpha,\beta) = \left(\mathbf{x} - \mathbf{H}(\alpha)\beta\right)^T \left(\mathbf{x} - \mathbf{H}(\alpha)\beta\right)$$

is linear in ***β*** and nonlinear in ***α***. For a given ***α*** can be minimized, with (partial) solution

$$\hat{\beta} = \left(\mathbf{H}^T(\alpha)\mathbf{H}(\alpha)\right)^{-1}\mathbf{H}^T(\alpha)\mathbf{x}$$

The problem now reduces to the maximization of

$$J(\alpha,\hat{\beta}) = \mathbf{x}^T\left( I - \mathbf{H}(\alpha)\left(\mathbf{H}^T(\alpha)\mathbf{H}(\alpha)\right)^{-1}\mathbf{H}^T(\alpha)\right)\mathbf{x}$$

over ***α.***

# Nonlinear Least Squares:

## General case

When all the other methods fail, a Taylor series expansion can be used. The criterion is then approximated…

$$J(\theta) = \sum_{n=0}^{N-1}\left(x[n]-s[n;\theta]\right)^2 \approx \sum_{n=0}^{N-1}\left(x[n]-s[n;\theta_0]-\left.\frac{ds[n;\theta]}{d\theta}\right|_{\theta_0}(\theta-\theta_0)\right)^2$$

If we set up an iterative procedure (as in the Newton-Rawphson case)

$$\theta_{k+1} = \theta_k + \left(\mathbf{H}^T(\theta_k)\,\mathbf{H}(\theta_k)\right)^{-1}\mathbf{H}^T(\theta_k)\left(\mathbf{x}-\mathbf{s}(\theta_k)\right)$$

Where

$$[\mathbf{H}(\boldsymbol{\theta})]_{ij} = \frac{\partial s[i]}{\partial \theta_j}$$

The solution can be trivially generalized to the vector case:

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

# Bibliography:

## Further reading

- Thomas Kailath, ***Linear Systems,*** Prentice Hall, 1980.

- Thomas Kailath, Ali Sayed, and Babak Hassibi, ***Linear Estimation,*** Prentice Hall, 2000.

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

- J. Bibby, H. Toutenburg, ***Prediction and Improved Estimation in Linear Models,*** John Wiley, 1977.

- C.Rao, ***Linear Statistical Inference and Its Applications,*** John Wiley, 1973.