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)

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

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]

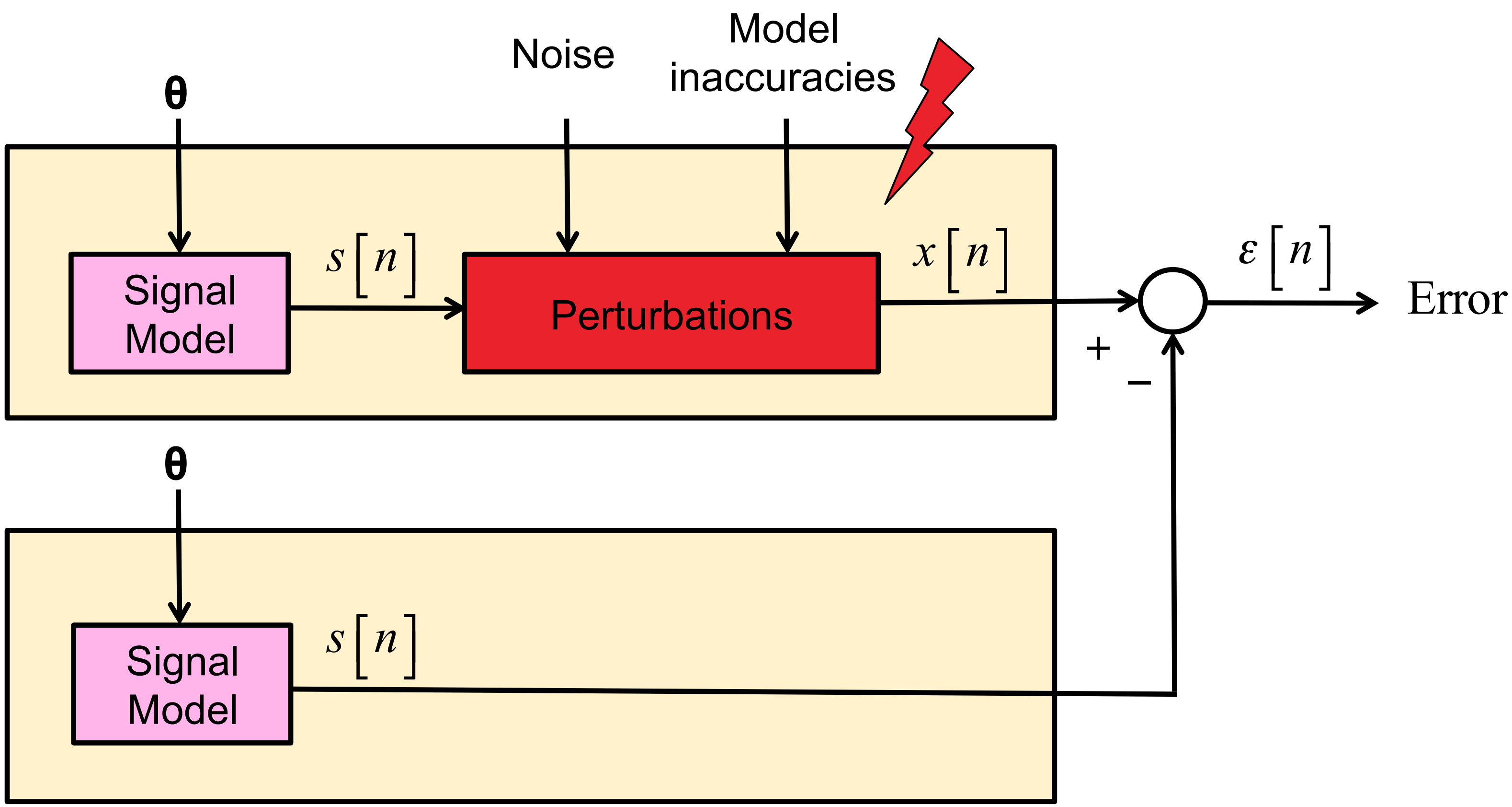
The least squares approach; Linear and nonlinear least squares; Geometric interpretation; Constrained least squares; Examples;

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;

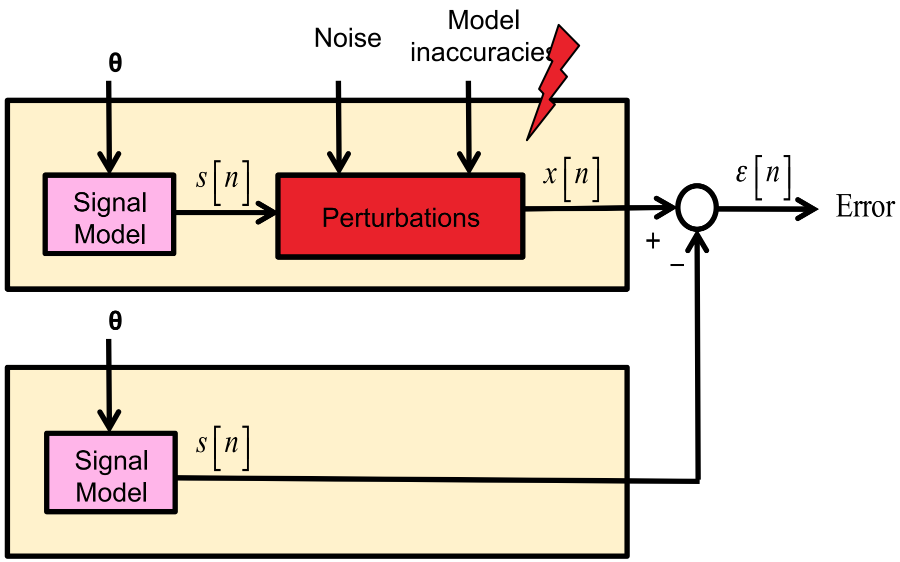
optimal estimator can not be found;

…

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 $\boldsymbol{s} = [s[0]\ s[1]\ \ldots\ s[N-1]]$
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

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

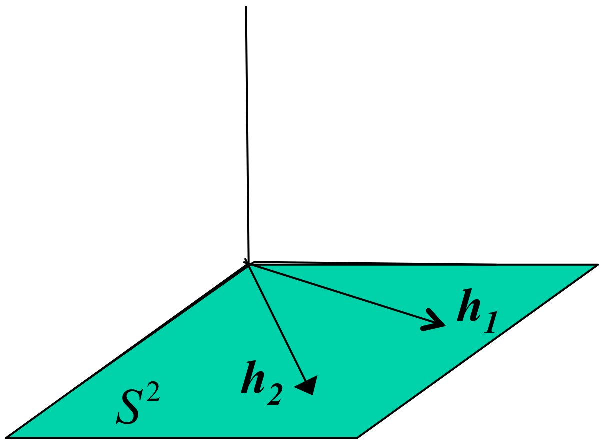
can be rewritten as

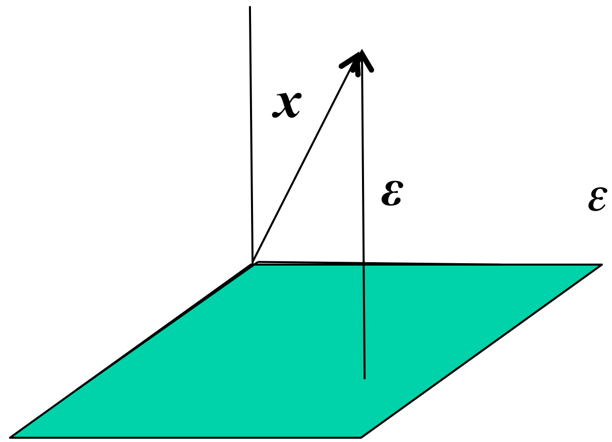
$$\left(\mathbf{H}^T\mathbf{H}\right)\boldsymbol{\theta} = \left(\mathbf{H}^T\mathbf{H}\right)\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}$$

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

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

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





$\varepsilon \perp S^2$

DEM
DEPARTAMENTO DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

Extensions to Least Squares:

Other extensions of the least squares approach are also very popular

Weighted Least Squares:

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

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

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

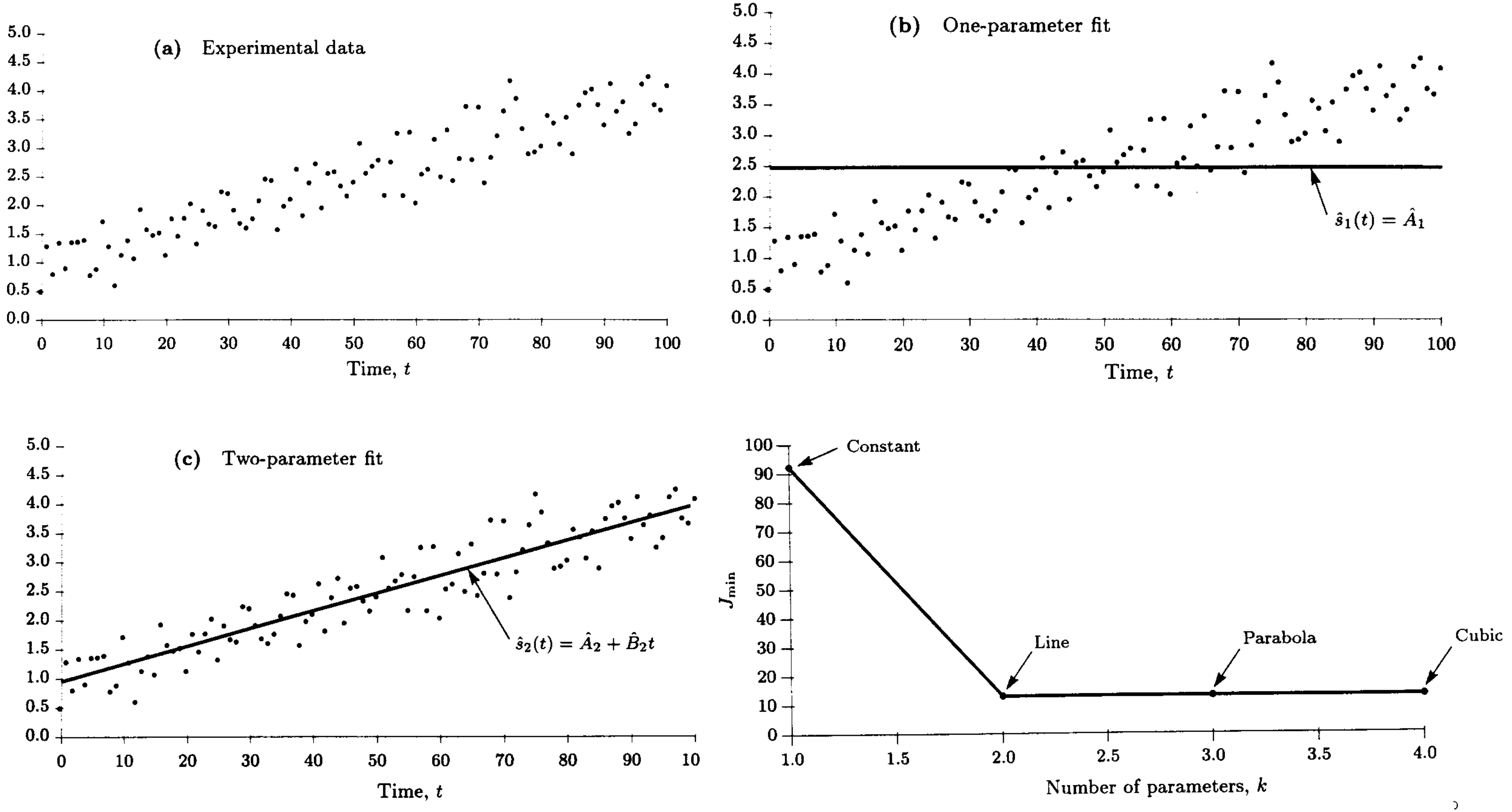
minimum $J_{\min}\left(\theta_{k+1}\right) = J_{\min}\left(\theta_k\right) - \dfrac{\left(\mathbf{h}_{k+1}^T\mathbf{P}_k^{\perp}\mathbf{x}\right)^2}{\mathbf{h}_{k+1}^T\mathbf{P}_k^{\perp}\mathbf{h}_{k+1}}$

Example:
Line fitting

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

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

Example:



(a) Experimental data
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
Constant
Line
Parabola
Cubic
$J_{\min}$
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 $\boldsymbol{x}=[x[0]\ x[1]\ \ldots\ x[N-1]]$ 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…

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

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.

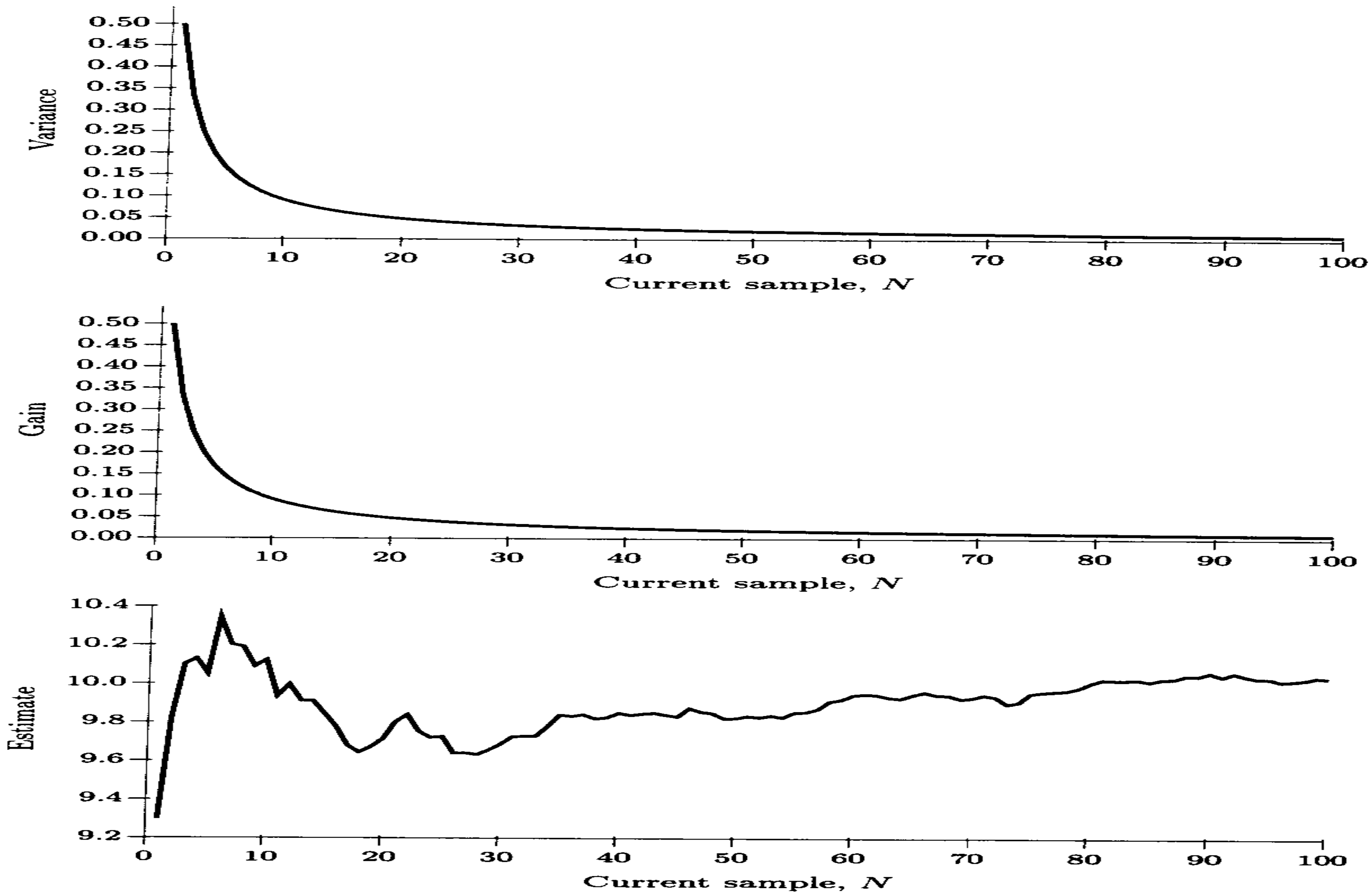
$$J_{\min}(N) = J_{\min}(N-1) + \frac{N}{N+1}\left(x[N] - \hat{A}[N-1]\right)^2$$

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

Estimator Update:

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

Where

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

Covariance Update:

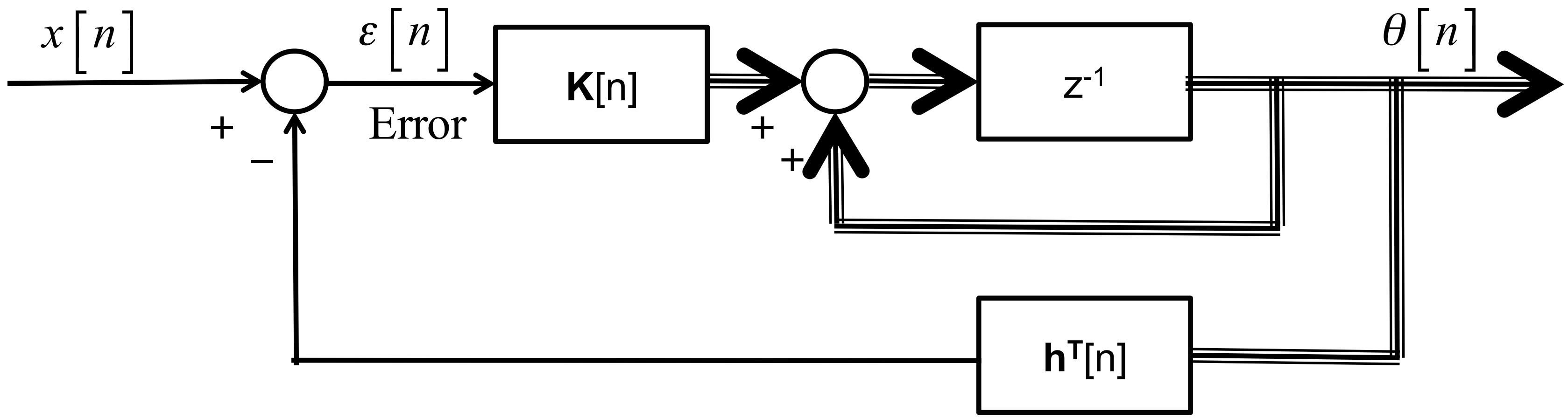
$$\Sigma[n]=\left(\mathbf{I}-\mathbf{K}[n]h^T[n]\right)\Sigma[n-1]$$

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

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



DEM
DEPARTAMENTO DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

Constrained Least Squares:

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

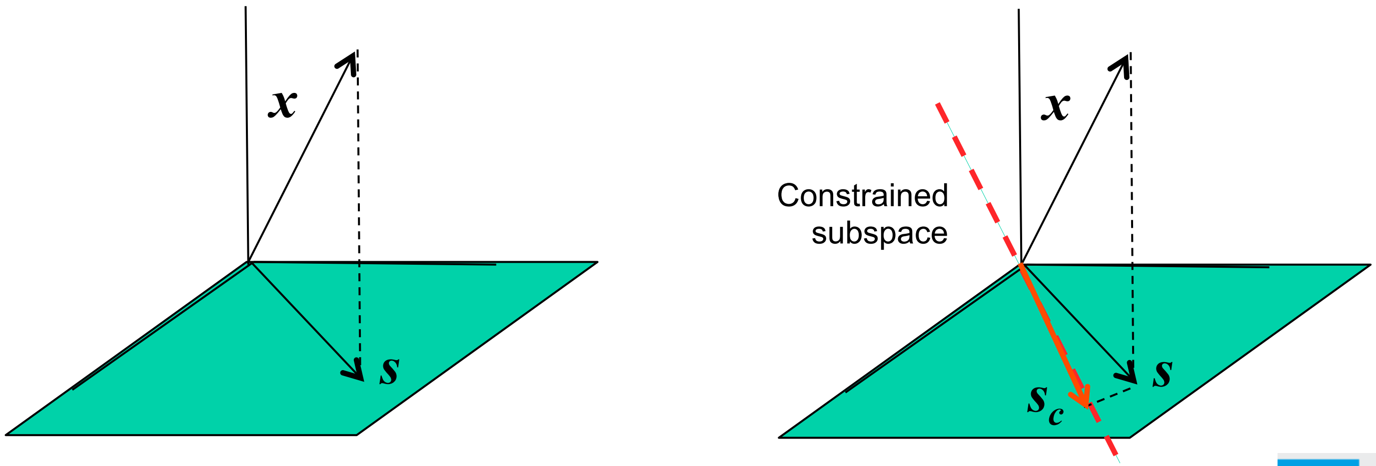
criterion $$J_C(\theta) = (\mathbf{x} - \mathbf{H}\theta)^T (\mathbf{x} - \mathbf{H}\theta)$$

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

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.

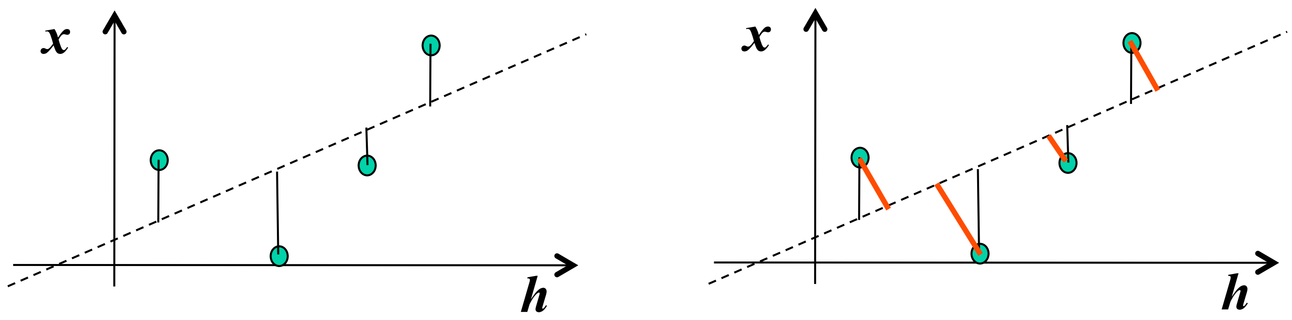
It can also be interpreted as the constrained signal estimate to be the projection of the unconstrained solution onto the constrained subspace.



Extensions to Least Squares:

Other extensions:

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



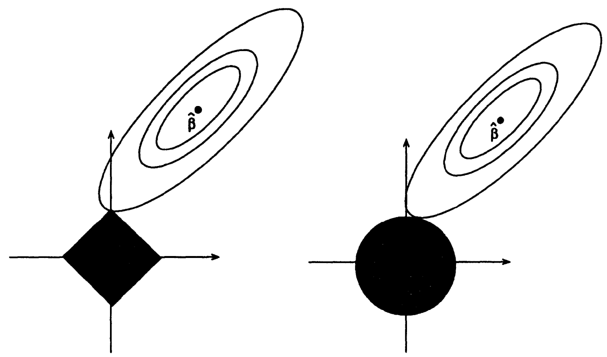
When could also be errors in the independent variables.

Lasso – Least Absolute Shrinkage and Selection Operator

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

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

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.

DEM
DEPARTAMENTO DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

Nonlinear Least Squares:

In general the signal model is

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

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

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 $\mathbf{g}()$ 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 $\mathbf{g}()$, 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 $\boldsymbol{\beta}$ and nonlinear in $\boldsymbol{\alpha}$. For a given $\boldsymbol{\alpha}$ 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 $\boldsymbol{\alpha}$.

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

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

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

The solution can be trivially generalized to the vector case:

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

DEM
DEPARTAMENTO
DE ENGENHARIA MECÂNICA
TÉCNICO LISBOA

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.