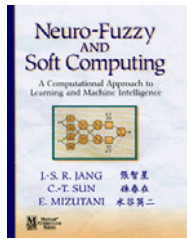
Chapter 5: Least-Square Methods for System Identification

- System Identification: an Introduction (5.1)
- Least-Squares Estimators (5.3)
- Statistical Properties & the Maximum Likelihood Estimator (5.7)
- LSE for Nonlinear Models (5.8)



Jyh-Shing Roger Jang et al., *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*, First Edition, Prentice Hall, 1997

System Identification: Introduction (5.1)

- Goal
  - Determine a mathematical model for an unknown system (or target system) by observing its input-output data pairs
- Purposes
  - To predict a system's behavior, as in time series prediction & weather forecasting
  - To explain the interactions & relationships between inputs & outputs of a system

System Identification: Introduction (5.1) (cont.)

- Purposes (cont.)
  - To design a controller based on the model of a system, as an aircraft or ship control
  - Simulate the system under control once the model is known

System Identification: Introduction (5.1) (cont.)

- There are 2 main steps that are involved
  - Structure identification
  - Parameter identification

System Identification: Introduction (5.1) (cont.)

– Structure identification

Apply a-priori knowledge about the target system to determine a class of models within which the search for the most suitable model is to be conducted; this class of model is denoted by a function y = f(u,θ) where:

- y is the model output
- u is the input vector
- Is the parameter vector

f depends on the problem at hand & on the designer's experience & the laws of nature governing the target system

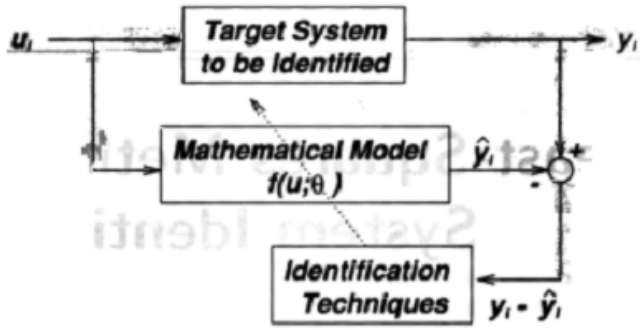
System Identification: Introduction (5.1) (cont.)

– Parameter identification

- The structure of the model is known, however we need to apply optimization techniques in order to determine the parameter vector $\theta = \hat{\theta}$ such that the resulting model $\hat{\mathbf{y}} = \mathbf{f}(\mathbf{u}, \hat{\theta})$ describes the system appropriately:

$$|\mathbf{y_i} - \hat{\mathbf{y}}| \to \mathbf{0} \ \ \textbf{with } \mathbf{y_i} \textbf{assigned to } \mathbf{u_i}$$

System Identification: Introduction (5.1) (cont.)



Block diagram for parameter identification

System Identification: Introduction (5.1) (cont.)

- The data set composed of m desired input-output pairs (ui;yi) (i = 1,…,m) is called the training data
- System identification needs to do both structure & parameter identification repeatedly until satisfactory model is found: it does this as follows:
  - Specify & parameterize a class of mathematical models representing the system to be identified
  - Perform parameter identification to choose the parameters that best fit the training data set
  - Conduct validation set to see if the model identified responds correctly to an unseen data set
  - Terminate the procedure once the results of the validation test are satisfactory. Otherwise, another class of model is selected & repeat step 2 to 4

Least-Squares Estimators (5.3)

◆General form:

$$y = \theta_1 f_1(u) + \theta_2 f_2(u) + \ldots + \theta_n f_n(u) \quad (*)$$

where:

- $u = (u_1, \ldots, u_p)T$ is the model input vector
- $f_1, \ldots, f_n$ are known functions of u
- $\theta_1, \ldots, \theta_n$ are unknown parameters to be estimated

Least-Squares Estimators (5.3) (cont.)

◆The task of fitting data using a linear model is referred to as linear regression

◆We collect a training data set

{(ui;yi), i = 1, …, m}

Equation (*) becomes:

$$\begin{cases} f_1(u_1)\theta_1 + f_2(u_1)\theta_2 + \ldots + f_n(u_1)\theta_n = y_1 \\ f_1(u_2)\theta_1 + f_2(u_2)\theta_2 + \ldots + f_n(u_2)\theta_n = y_2 \\ \vdots \\ f_1(u_m)\theta_1 + f_2(u_m)\theta_2 + \ldots + f_n(u_m)\theta_n = y_m \end{cases}$$

Which is equivalent to: $A\,\theta = y$

Least-Squares Estimators (5.3) (cont.)

Where: A is an m*n matrix which is: $\mathbf{A} = \begin{bmatrix} \mathbf{f_1(u_1)} \cdots \mathbf{f_n(u_1)} \\ \vdots \\ \mathbf{f_1(u_m)} \cdots \mathbf{f_n(u_m)} \end{bmatrix}$

θ is n*1 unknown parameter vector: $\theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}$

and y is an m*1 output vector:

$$\mathbf{y} = \begin{bmatrix} \mathbf{y_1} \\ \vdots \\ \mathbf{y_m} \end{bmatrix} \textbf{ and } \mathbf{a_i^T} = \left[\mathbf{f_1(u_i)},...,\mathbf{f_n(u_i)}\right]$$

$A\,\theta = y \Leftrightarrow \theta = A^{-1}y$ (solution)

Least-Squares Estimators (5.3) (cont.)

- We have m outputs & n fitting parameters to find (or m equations & n unknown variables)
- Usually m is greater than n, since the model is just an approximation of the target system & the data observed might be corrupted, therefore an exact solution is not always possible!
- To overcome this inherent conceptual problem, an error vector e is added to compensate

$$A\,\theta + e = y$$

Least-Squares Estimators (5.3) (cont.)

- Our goal consists now of finding $\hat{\theta}$ that reduces the errors between yi & $\hat{y}_i = a_i^T \theta$

$$\min_{\theta} \text{imize} \sum_{i-1}^{m} \left| y_i - a_i^T \theta \right|$$ (not derivable!)

rather

$$\min_{\theta} \text{imize} \sum_{i=1}^{m} \left(y_i - a_i^T \theta \right)^2$$

Least-Squares Estimators (5.3) (cont.)

- If e = y - Aθ then:

$$E(\theta) = \sum_{i=1}^{i=m} (y_i - a_i{}^T \theta)^2 = e^T e = (y - A\theta)^T (y - A\theta)$$

We need to compute:

$$\min_{\theta} (y - A\theta)^T (y - A\theta)$$

Least-Squares Estimators (5.3) (cont.)

◆Theorem [least-squares estimator]

The squared error is minimized when $\theta = \hat{\theta}$ (called the least-squares estimators, LSE) satisfies the normal equation

$$A^T A\hat{\theta} = A^T y,$$

if $A^T A$ is nonsingular, $\hat{\theta}$ is unique & is given by

$$\hat{\theta} = (A^T A)^{-1} A^T y$$

Least-Squares Estimators (5.3) (cont.)

◆Example

– The relationship is between the spring length & the force applied $L = k_1 f + k_0$ (linear model)

– Goal: find $\hat{\theta} = (k_o , k_1)T$ that best fits the data for a given force fo, we need to determine the corresponding spring length L0

- Solution 1: provide 2 pairs (L0,f0) and (L1,f1) and solve a linear system of 2 equations and 2 variables k1 and k0

$\Rightarrow \hat{\mathbf{k}}_1$ & $\hat{\mathbf{k}}_0$. However, because of noisy data, this solution is not reliable!

- Solution 2: use a larger training set (Li,fi)

Least-Squares Estimators (5.3) (cont.)

Experiment	Force (newtons)	Length of Spring (inches)
1	1.1	1.5
2	1.9	2.1
3	3.2	2.5
4	4.4	3.3
5	5.9	4.1
6	7.4	4.6
7	9.2	5.0

Training data for the spring example

Least-Squares Estimators (5.3) (cont.)

since y = e + Aθ, we can write:

$$\underbrace{\begin{bmatrix} 1 & 1.1 \\ 1 & 1.9 \\ 1 & 3.2 \\ 1 & 4.4 \\ 1 & 5.9 \\ 1 & 7.4 \\ 1 & 9.2 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} k_0 \\ k_1 \end{bmatrix}}_{\theta} + \underbrace{\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ \vdots \\ e_7 \end{bmatrix}}_{e} = \underbrace{\begin{bmatrix} 1.5 \\ 2.1 \\ 2.5 \\ 3.3 \\ 4.1 \\ 4.6 \\ 5.0 \end{bmatrix}}_{y}$$

therefore the LSE of [k0,k1]T which minimizes

$e^T e = \sum_{i=1}^{7} e_i^2$ is equal to : $\begin{bmatrix} \hat{k}_0 \\ \hat{k}_1 \end{bmatrix} = (A^T A)^{-1} A^T y = \begin{bmatrix} 1.20 \\ 0.44 \end{bmatrix}$

Least-Squares Estimators (5.3) (cont.)

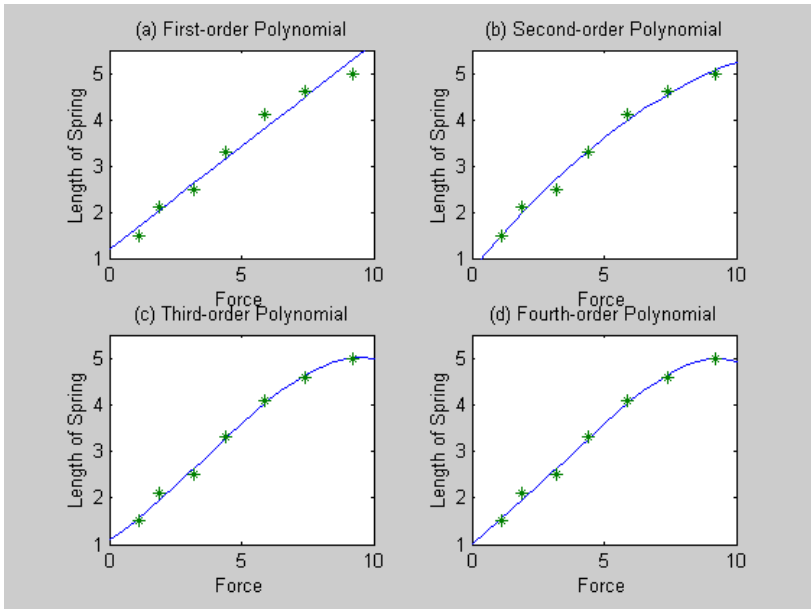
- We rely on this estimation because we have more data
- If we are not happy with the LSE estimators then we can increase the model's degree of freedom such that:

$$L = k_0 + k_1 f + k_2 f_2 + \ldots + k_n f_n$$

(least square polynomial!)

Least-Squares Estimators (5.3) (cont.)

- Higher order models fit better the data but they do not always reflect the inner law that governs the system
- For example, when f is increasing toward 10N, the length is decreasing!



Statistical Properties & the Maximum Likelihood Estimator (5.7)

- Statistical qualities of LSE
  - Definition [unbiased estimator]
    An estimator $\hat{\theta}$ of the parameter θ is unbiased if $E(\hat{\theta}) = \theta$, where E[.] is the statistical expectation
  - Definition [minimal variance]
    An estimator $\hat{\theta}$ is a minimum variance estimator if for any other estimator θ^*: $\text{cov}(\hat{\theta}) \leq \text{cov}(\theta^*)$, where $\text{cov}(\theta)$ is the covariance matrix of the random vector θ.

Statistical Properties & the Maximum Likelihood Estimator (5.7) (cont.)

- Statistical qualities of LSE (cont.)
  - Gauss-Markov conditions:
    - The error vector e is a vector of m uncorrelated random variables, each with zero mean & the same variance $\sigma 2$. This means that:

$$E[e] = 0 \;\&\; E[eeT] = \sigma 2I$$

$$\left[E\begin{pmatrix} e_1 \\ e_2 \end{pmatrix}(e_1, e_2) = \begin{pmatrix} E(e_1^2) & E(e_1 e_2) \\ E(e_2 e_1) & E(e_2^2) \end{pmatrix} = \sigma^2 I \right]$$

Statistical Properties & the Maximum Likelihood Estimator (5.7) (cont.)

- *Theorem [Gauss-Markov]:*

LSE is unbiased & has minimum variance.

Proof: $E[\hat{\theta}] = E[(A^TA)^{-1}A^Ty] = (A^TA)^{-1}A^TE[y]$

$= (A^TA)^{-1}A^TA\theta = \theta$

since $E[y] = E[A\theta] + E(e) = A\theta$ and

$(A^TA)^{-1} = A^{-1}(A^T)^{-1}$

Statistical Properties & the Maximum Likelihood Estimator (5.7) (cont.)

◆Maximum likelihood (ML) estimator

– ML is one of the most widely used technique for parameter estimation of a statistical distribution

– *ML definition:*

For a sample of n observations (of a probability density function) $x_1, x_2, \ldots, x_n$, the likelihood function L is defined by:

$$L = f(x_1; \theta)\, f(x_2; \theta) \ldots f(x_n; \theta)$$

Statistical Properties & the Maximum Likelihood Estimator (5.7) (cont.)

◆Maximum likelihood (ML) estimator (cont.)

– The criterion for choosing θ is:
"pick a value of θ that provides a high probability of obtaining the actual observed data $x_1, x_2, \ldots, x_n$"

– Therefore, ML estimator $\hat{\theta}$ is defined as the value of θ which maximizes L: $\frac{\partial L}{\partial \theta}(\hat{\theta}) = 0$

or equivalently: $\frac{\partial \ln L}{\partial \theta}(\hat{\theta}) = 0$

Statistical Properties & the Maximum Likelihood Estimator (5.7) (cont.)

◆Maximum likelihood (ML) estimator (cont.)

– Example 1: ML estimation for exponential distribution

$f(x; \theta) = \theta^{-1} e^{-x/\theta}$

$$L = (\theta^{-1}e^{-x_1/\theta})(\theta^{-1}e^{-x_2/\theta})...(\theta^{-1}e^{-x_m/\theta})$$
$$= \theta^{-m} \exp\left(-\theta^{-1}\sum x_i\right)$$

is the likelihood function for m observations $x_1, x_2, \ldots, x_m$.

$$\ln L = -m \ln \theta - \frac{1}{\theta}\sum_{i=1}^{m} x_i \qquad \frac{\partial \ln L}{\partial \theta} = -\frac{m}{\theta} + \frac{\sum x_i}{\theta^2} = 0$$

$$\text{Therefore: } \hat{\theta} = \frac{\sum x_i}{m}$$

Statistical Properties & the Maximum Likelihood Estimator (5.7) (cont.)

◆Maximum likelihood (ML) estimator (cont.)

– Example 2: ML estimation for normal distribution

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

where μ and $\sigma 2$ are respectively the mean & the variance unknown parameters.

For m observations $x_1, x_2, \ldots, x_m$, we have:

Statistical Properties & the Maximum Likelihood Estimator (5.7) (cont.)

- Example 2 (cont.)

$$L = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^m \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right]$$

$$\text{and} \quad \ln L = -m * \ln\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2\sigma^2}\sum (x_i - \mu)^2$$

$$\text{Therefore} \quad : \quad \frac{\partial \ln L}{\partial \mu} = 0 \Rightarrow \frac{1}{\sigma^2}\sum (x_i - \mu) = 0 \Rightarrow \hat{\mu} = \frac{\sum x_i}{m}$$

$$\frac{\partial \ln L}{\partial \sigma} = 0 \Rightarrow -\frac{m}{\sigma} - \frac{1}{2}\left(-2\sigma^{-3}\right)\sum (x_i - \mu)^2 = 0$$

$$\Rightarrow \hat{\sigma}^2 = \frac{\sum (x_i - \hat{\mu})^2}{m}$$

Statistical Properties & the Maximum Likelihood Estimator (5.7) (cont.)

- Maximum likelihood (ML) estimator (cont.)
  - Equivalence between LSE & MLE
  - Theorem: Under the Gauss conditions & if each component of the vector e follows a normal distribution then the LSE of θ = MLE of θ

LSE for Nonlinear Models (5.8)

- Nonlinear models are divided into 2 families
  - Intrinsically linear
  - Intrinsically nonlinear
    - Through appropriate transformations of the input-output variables & fitting parameters, an intrinsically linear model can become a linear model
    - By this transformation into linear models, LSE can be used to optimize the unknown parameters

LSE for Nonlinear Models (5.8) (cont.)

- Example: Decay of the radioactivity of chemicals
  $y = aebt$ (a, b < 0 are to be determined)
  (t, y) are the input-output data

given a training data $\{(t_i, y_i), i = 1, \ldots, m\}$

$$\mathbf{E(a,b)} = \sum_{i=1}^{m} \left(\mathbf{y_i} - \mathbf{a e^{bt_i}}\right)^2$$

minimizing the error yields:

$$\begin{cases} \dfrac{\partial \mathbf{E}}{\partial \mathbf{a}} = 2\sum_{i=1}^{m} \left(\mathbf{y_i} - \mathbf{a e^{-bt_i}}\right)\left(-\mathbf{e^{bt_i}}\right) = \mathbf{0} \\ \dfrac{\partial \mathbf{E}}{\partial \mathbf{b}} = 2\sum_{i=1}^{m} \left(\mathbf{y_i} - \mathbf{a e^{-bt_i}}\right)\left(-\mathbf{a t_i e^{-bt_i}}\right) = \mathbf{0} \end{cases}$$

we obtain 2 nonlinear equations with respect to a & b!
Therefore, we are better off to linearize the model before.

LSE for Nonlinear Models (5.8) (cont.)

lny = lna + bt

with lna = a', lny = y' and b = b'

$\Rightarrow$ y' = a' + b't

which is an intrinsically linear model. The linear & nonlinear models are close but can sometimes differ significantly!

The following table shows a family of intrinsically linear models

LSE for Nonlinear Models (5.8) (cont.)

Nonlinear models	Transformation	Linear forms
$y = ae^{bx}$	Natural logarithm	$\ln y = \ln a + bx$
$y = ax^b$	Natural logarithm	$\ln y = \ln a + b \ln x$
$y = \frac{ax}{b+x}$	Reciprocal	$\frac{1}{y} = \frac{1}{a} + \frac{b}{a}\frac{1}{x}$
$y = \frac{a}{b+x}$	Reciprocal	$\frac{1}{y} = \frac{1}{a} + \frac{b}{a}x$

LSE for Nonlinear Models (5.8) (cont.)

Suppose that the underlying model is: $y = ae^{bt}$

t_i	0	0.80	1.84	2.90	4.06	4.81	6.07	7.06	8.15	8.87	9.98
y_i	0.98	0.69	0.47	0.46	0.29	0.16	0.23	0.10	0.03	0.12	0.01

LSE for Nonlinear Models (5.8) (cont.)