Lectures on the Identification Problem

The Least Squares Method

Åström, Karl Johan

1968

Document Version:
Publisher's PDF, also known as Version of record

LECTURES ON THE IDENTIFICATION PROBLEM - THE LEAST SQUARES METHOD

K. J. ÅSTRÖM

REPORT 6806 SEPTEMBER 1968
LUND INSTITUTE OF TECHNOLOGY
DIVISION OF AUTOMATIC CONTROL

LECTURES ON THE IDENTIFICATION PROBLEM - THE LEAST SQUARES METHOD

by

K. J. Åström

This work was partially supported by the Swedish Board for Technical Development under Contract 68-336-f.

CONTENTS

PREFACE

This report is part of a lecture series on the identification problem. Part of the material is wellknown and collected from various parts of the literature and other parts e.g. the generalization of Mann and Wolds theorem in section 8 and the investigation of the relations between various identification schemes in sections 8, 9, 10 and 11 are believed to be new.

The write-up is aimed both at engineers in industry who are faced with practical identification problem and to graduate students. The material in this report has been presented in lecture form to various audiences such as industry groups and graduate students at Lund Institute of Technology. It was first put together in the present form in a lecture given to the Research Institute of National Defense.

The write-up is not complete, many details of the proofs are omitted, many practical examples should be included etc. As there are no really good surveys available on the identification problem I think it is worth while to distribute these notes although they are not polished.

1. INTRODUCTION

In this chapter we will discuss the least squares method and its application to identification problems such as identification of process dynamics and parametric modelling of random disturbances. The least squares method introduced by Gauss in his famous work on orbit prediction is one of the oldest and most powerful techniques available. It is easy to learn and it gives identification procedures that are easily programmed. It also turns out that many of the identification schemes that are used can be interpreted as least squares procedures. It is thus possible to treat many problems in a unified framework.

In this chapter we will thus present the basic results of the least squares procedure and show how they can be applied to the solution of the identification problem.

Throughout the chapter we will use two simple examples as illustrations.

Example 1.1

Consider a discrete time single input single output system. Assume that a sequence of inputs {y(1), y(2), ..., y(N)} has been applied to the system and that the corresponding sequence of outputs {u(1), u(2), ..., u(N)} has been observed. Consider the following class of models

$$y(t) + a\, y(t-1) = b\, u(t-1)$$

which is characterized by the parameters a and b and let the criterion be to minimize the sum

$$\sum_{t=2}^{N} e^2(t)$$

where

$$e(t) = y(t) - a\, y(t-1) - b\, u(t-1)$$

Example 1.2

Consider a discrete time disturbance. Assume that a sequence of values {y(1), y(2), ..., y(N)} has been observed, and that we would like to describe the series by the model

$$y(t) + a\, y(t-1) = 0$$

in such a way that the criterion

$$\sum_{t=2}^{N} e^2(t)$$

is minimal where

$$e(t) = y(t) + a\, y(t-1)$$

In section 2 we will give a general formulation of the least squares problem which covers these examples as well as many other identification problems. The basic solution to the least squares problem is also given in that section. A geometrical interpretation of the results are given in section 3. In sections 4 and 5 we show how the basic solution can be rewritten for recursive computations. Section 4 covers the situation when we want to fit models of increasing order and section 5 the situation when measurements are obtained recursively and we want to compute the estimate in real-time as the process develops. In section 6 we give a statistical interpretation of the least squares procedure. One important result is that when formulating the identification as a statistical problem, we get a natural loss function. It will also be possible to answer various statistical problems e.g. how accurate are the parameter values that are obtained. The statistical properties are also discussed in section 7 where we among other things discuss tests of order of the models. Section 8 covers an example of identification of linear process dynamics. In sections 9, 10 and 11 we discuss the relations between the least squares procedure and other identification methods. The last sections finally cover exercises and references.

2. GENERAL FORMULATION

Suppose that we are given the model

$$y_i = \theta_1\varphi_1(x_i) + \theta_2\varphi_2(x_i) + \ldots + \theta_n\varphi_n(x_i) + e_i \qquad i = 1,2,\ldots,N \tag{2.1}$$

where x_i are given values, φ are known functions, θ_i unknown coefficients, e_i errors and y_i observed values. Assume that we have N pairs of given values and observations $\{(x_i,y_i), i = 1,2,\ldots,N\}$ and that we want to determine the parameters θ_i in such a way that the error defined as

$$\sum_{i=1}^{N} e_i^2$$

is as small as possible. To solve this problem we introduce a vector formalism

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad \Phi = \begin{bmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \ldots & \varphi_n(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) & \ldots & \varphi_n(x_2) \\ \vdots & & & \\ \varphi_1(x_N) & \varphi_2(x_N) & \ldots & \varphi_n(x_N) \end{bmatrix}, \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_N \end{bmatrix} \tag{2.2}$$

The equation (2.1) can then be written as

$$y = \Phi\theta + e \tag{2.3}$$

and the loss function

$$2V = \sum_{i=1}^{N} e_i^2 = e^Te \tag{2.4}$$

It is now a simple matter to minimize V with respect to θ. We find

$$e^Te = \{y - \Phi\theta\}^T\{y - \Phi\theta\} = y^Ty - y^T\Phi\theta - \theta^T\Phi^Ty + \theta^T\Phi^T\Phi^T\theta$$

Assuming that $\Phi^T\Phi$ is regular we find

$$e^Te = \{\theta - (\Phi^T\Phi)^{-1}\Phi^Ty\}^T\Phi^T\Phi\{\theta - (\Phi^T\Phi)^{-1}\Phi^Ty\}$$
$$+ y^Ty - y^T\Phi(\Phi^T\Phi)^{-1}\Phi^Ty$$
$$\geq y^Ty - y^T\Phi(\Phi^T\Phi)^{-1}\Phi^Ty$$

where the equality occurs for

$$\theta = \hat{\theta} = (\Phi^T\Phi)^{-1}\Phi^Ty \tag{2.5}$$

or

$$\Phi^T\Phi\,\hat{\theta} = \Phi^Ty \tag{2.6}$$

The condition that $\Phi^T\Phi$ is regular will be discussed in detail in sections 3 and 8. Let it suffice at this point to mention that the regularity can be influenced by the choice of x_i.

The solution of the least squares problem is thus very straightforward. We will now give a few examples.

Example 2.1

Consider the problem of example 1.1. In this case we have

$$y(t+1) + a\,y(t) - b\,u(t) = e(t) \qquad t = 0,1,\ldots,N$$

Hence

$$y = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix}, \quad \Phi = \begin{bmatrix} y(0) & -u(0) \\ y(1) & -u(1) \\ & \\ y(N-1) & -u(N-1) \end{bmatrix}, \quad e = \begin{bmatrix} e(1) \\ e(2) \\ \\ e(N) \end{bmatrix}, \quad \theta = \begin{bmatrix} a \\ b \end{bmatrix}$$

$$\Phi^T\Phi = \begin{bmatrix} \sum_{t=0}^{N-1} y^2(t) & \sum_{t=0}^{N-1} y(t)\,u(t) \\ \sum_{t=0}^{N-1} y(t)\,u(t) & \sum_{t=0}^{N-1} u^2(t) \end{bmatrix}$$

$$\Phi^{T}y = \begin{bmatrix} \sum_{t=0}^{N-1} y(t+1)\, y(t) \\ \sum_{t=0}^{N-1} y(t+1)\, u(t) \end{bmatrix}$$

3. GEOMETRICAL INTERPRETATION

The least squares problem discussed in section 2 can be given a simple geometric interpretation. Write the equation (2.1) as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} \varphi_1(x_1) \\ \varphi_1(x_2) \\ \vdots \\ \varphi_1(x_N) \end{bmatrix} \theta_1 - \begin{bmatrix} \varphi_2(x_1) \\ \varphi_2(x_2) \\ \vdots \\ \varphi_2(x_N) \end{bmatrix} \theta_2 - \cdots \begin{bmatrix} \varphi_n(x_1) \\ \varphi_n(x_2) \\ \vdots \\ \varphi_n(x_N) \end{bmatrix} \theta_n = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix} \tag{3.1}$$

or

$$y - \varphi^1\theta_1 - \varphi^2\theta_2 - \ldots - \varphi^n\theta_n = e \tag{3.2}$$

Now consider $y, \varphi^1, \varphi^2, \ldots, \varphi^n$ as vectors in N-dimensional Euclidian space with the norm

$$||x|| = x^T x \tag{3.3}$$

The least squares problem can then be formulated as the problem of approximating the vector y by a linear combination of the vectors $\varphi^1, \varphi^2, \ldots, \varphi^n$ in such a way that the error norm is minimal. It is well known that the solution of this problem is given by the orthogonal projection of y as the linear space spanned by the vectors $\varphi^1, \varphi^2, \ldots, \varphi^n$. Compare fig. 3.1.

Fig. 3.1 - Illustrates the geometrical interpretation of the least squares procedure

Let y^* be the orthogonal projection. The vector $y - y^*$ is then orthogonal to all the vectors $\varphi^1, \varphi^2, \ldots, \varphi^n$. All the scalar products of $y - y^*$ with $\varphi^1, \varphi^2, \ldots, \varphi^n$ must thus be zero. Hence

$$
\begin{aligned}
(y - y^*)^T \varphi^1 &= 0 \\
(y - y^*)^T \varphi^2 &= 0 \\
&\vdots \\
(y - y^*)^T \varphi^n &= 0
\end{aligned}
\qquad (3.4)
$$

must thus equal zero. Now introduce

$$y^* = \theta_1 \varphi^1 + \theta_2 \varphi^2 + \ldots + \theta_n \varphi^n \qquad (3.5)$$

The equation (3.4) then reduces to

$$
\begin{bmatrix}
(\varphi^1)^T \varphi^1 & (\varphi^1)^T \varphi^2 & \cdots & (\varphi^1)^T \varphi^n \\
(\varphi^2)^T \varphi^1 & (\varphi^2)^T \varphi^2 & \cdots & (\varphi^2)^T \varphi^n \\
\vdots & & & \\
(\varphi^n)^T \varphi^1 & (\varphi^n)^T \varphi^2 & \cdots & (\varphi^n)^T \varphi^n
\end{bmatrix}
\theta =
\begin{bmatrix}
y^T \varphi^1 \\
y^T \varphi^2 \\
\vdots \\
y^T \varphi^n
\end{bmatrix}
\qquad (3.6)
$$

which is identical to (2.6). Notice that this argument does not require $\varphi^1, \varphi^2, \ldots, \varphi^n$ to span the n-dimensional space. We have thus removed the condition for $\Phi^T\Phi$ to be regular. Notice, however, that if $\varphi^1, \varphi^2, \ldots, \varphi^n$ are linearly dependent then the elements of θ are not unique. A model of the type (2.1) with fewer parameters can be obtained by projections on the space of linearly independent φ^i.

4. RECURSIONS IN THE NUMBER OF PARAMETERS

In many cases it is not known a priori how many terms of the expansion (2.1) that should be used. When identifying system dynamics it happens very often that the order of the system is not known beforehand. The least squares problem then has to be solved for several values of the parameter n. The loss function or the error norm will in general decrease with the number of parameters. In section 7 we will put the problem in a statistical framework and use statistical methods to decide whether the decrease in the loss function is significant. In this section we will discuss the computational aspects. When the least squares problem is solved for a model with n parameters and we wich to determine a model with n + 1 parameters, it seems to be a waste of computing effort to start from scratch. Instead it seems to be much preferable to use the results obtained for n parameters to obtain the model of order n + 1. In this section we will show how this should be done.

When introducing extra parameters θ_i we get additional terms in equation (2.1). This implies that additional columns are added to the Φ matrix and that additional elements are added to the vector θ. To fix the ideas we assume that a model of order k has been obtained and that we intend to compute the coefficients of a model of order n, n > k. To do so we partition the matrix Φ and the vector θ as follows

$$\Phi_n = \left[\begin{array}{cccc|ccc} \varphi_1(x_1) & \varphi_2(x_1) & \dots & \varphi_k(x_1) & \varphi_{k+1}(x_1) & \dots & \varphi_n(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) & & \varphi_k(x_2) & \varphi_{k+1}(x_2) & & \varphi_n(x_2) \\ \vdots & & & & & & \\ \varphi_1(x_N) & \varphi_2(x_N) & & \varphi_k(x_N) & \varphi_{k+1}(x_N) & & \varphi_n(x_N) \end{array}\right] = \{\Phi_1 \mid \Phi_2\} \tag{4.1}$$

$$\theta_n = \mathrm{col}\{\theta_1, \theta_2, \dots, \theta_k \mid \theta_{k+1}, \dots, \theta_n\} = \{\theta_1 \mid \theta_2\} \tag{4.2}$$

The least squares equations (2.6) now becomes

$$\left[\begin{array}{c|c} \Phi_1^T\Phi_1 & \Phi_1^T\Phi_2 \\ \hline \Phi_2^T\Phi_1 & \Phi_2^T\Phi_2 \end{array}\right] \left[\begin{array}{c} \theta_1 \\ \hline \theta_2 \end{array}\right] = \left[\begin{array}{c} \Phi_1 y \\ \hline \Phi_2 y \end{array}\right] \tag{4.3}$$

Hence

$$\begin{aligned} \Phi_1^T\Phi_1\theta_1 + \Phi_1^T\Phi_2\theta_2 &= \Phi_1 y \\ \Phi_2^T\Phi_1\theta_1 + \Phi_2^T\Phi_2\theta_2 &= \Phi_2 y \end{aligned} \tag{4.4}$$

Solving these equations we find

$$\begin{aligned} \theta_1 &= \theta_1 + P_3\Phi_2^T(\Phi_1^T\theta_1 - y) \\ \theta_2 &= -P_2\Phi_2^T(\Phi_1^T\theta_1 - y) \end{aligned} \tag{4.5}$$

where

$$P_2 = \{\Phi_2^T\Phi_2 - \Phi_2^T\Phi_1(\Phi_1^T\Phi_1)^{-1}\Phi_1^T\Phi_2\}^{-1} \tag{4.6}$$

$$P_3 = (\Phi_1^T\Phi)^{-1}\Phi_1^T\Phi_2 P_2 \tag{4.7}$$

We also find

$$\left[\begin{array}{cc} \Phi_1^T\Phi_1 & \Phi_1^T\Phi_2 \\ \Phi_2^T\Phi_1 & \Phi_2^T\Phi_2 \end{array}\right]^{-1} = \left[\begin{array}{cc} P_1 - P_3\Phi_1^T\Phi_2 & P_1 \quad P_3 \\ P_3^T & P_2 \end{array}\right] \tag{4.8}$$

where

$$P_1 = (\Phi_1^T\Phi_1)^{-1} \tag{4.9}$$

By using these equations we can compute models of successively higher order recursively. Notice that in each step we only have to invert matrices of order (n - k) x (n - k). Also notice that the quantity $\Phi_1^T\theta_1 - y$ has physical interpretation as the vector of residuals of the model of order k.

5. RECURSIONS IN THE NUMBER OF OBSERVATIONS

We will now consider recursive computations that are different from those of the previous section. It often happens that the observations are obtained sequantially as the process develops. It is then natural to compute the parameter estimates recursively as the data are obtained. This situation is often referred to as on-line identification. Let $\theta(N)$ denote the parameter estimate based on N pairs of data. To derive recursive equations for $\theta(N)$ we will introduce N as a parameter in all quantities involved.

Hence

$$\Phi(N) = \begin{bmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \dots & \varphi_n(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) & & \varphi_n(x_2) \\ \vdots & & & \\ \varphi_1(x_N) & \varphi_2(x_N) & & \varphi_n(x_N) \end{bmatrix} \tag{5.1}$$

$$y(N) = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \tag{5.2}$$

When an additional pair of data (x_{N+1}, y_{N+1}) is obtained a row is added to the matrix Φ and an element to the vector y. To derive recursive equations for the estimate, it is thus natural to partition the matrices as follows.

$$\Phi(N+1) = \begin{bmatrix} \Phi(N) \\ \hline \varphi(N+1) \end{bmatrix} \qquad y(N+1) = \begin{bmatrix} y(N) \\ \hline y_{N+1} \end{bmatrix} \tag{5.3}$$

where

$$\varphi(N+1) = \{\varphi_1(x_{N+1}), \varphi_2(x_{N+1}), \dots, \varphi_n(x_{N+1})\} \tag{5.4}$$

The least squares estimate of the parameters is given by (2.5) i.e.

$$\theta(N) = \{\Phi^T(N)\ \Phi(N)\}^{-1}\ \Phi^T(N)\ y(N) \tag{5.5}$$

Hence

$$\theta(N+1) = \{\Phi^T(N+1)\ \Phi(N+1)\}^{-1}\ \Phi^T(N+1)\ y(N+1)$$

$$= \{\Phi^T(N)\ \Phi(N) + \varphi^T(N+1)\ \varphi(N+1)\}^{-1}\{\Phi^T(N)y(N)+\varphi^T(N+1)y_{N+1}\} \tag{5.6}$$

To simplify the writing in the algebraical manipulations which will follow we will delete the argument N of $\Phi(N)$ and N + 1 of (N + 1). We will thus write (5.6) as

$$\theta(N+1) = \{\Phi^T\Phi + \varphi^T\varphi\}^{-1}\{\Phi^T y + \varphi^T y_{N+1}\} \tag{5.6}$$

In the following we will rewrite the equation (2.7). To do so we will utilize the following well-known matrix lemma.

Lemma

Let A, C and A + BCD be non singular square matrices then the following formula holds

$$\{A + BCD\}^{-1} = A^{-1} - A^{-1}B\{C^{-1} + DA^{-1}B\}^{-1}\ DA^{-1}$$

Proof

By direct substitution we find

$$\{A + BCD\}\{A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}\ DA^{-1}\}$$

$$= I + BCDA^{-1} - B(C^{-1} + DA^{-1}B)^{-1}\ DA^{-1} - BCDA^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$$

$$= I + BCDA^{-1} - B\{I + CDA^{-1}B\}\{C^{-1} + DA^{-1}B\}^{-1}\ DA^{-1}$$

$$= I + BCDA^{-1} - BC\{C^{-1} + DA^{-1}B\}\{C^{-1} + DA^{-1}B\}^{-1}\ DA^{-1}$$

$$= I + BCDA^{-1} - BCDA^{-1} = I$$

Notice in particular that the theorem holds if A and C are positive definite and $D = B^T$.

Using the matrix lemma we get

$$\{\Phi^T\Phi + \varphi^T\varphi\}^{-1} = (\Phi^T\Phi)^{-1} - (\Phi^T\Phi)^{-1}\varphi^T\{1 + \varphi(\Phi^T\Phi)^{-1}\varphi^T\}^{-1}\varphi(\Phi^T\Phi)^{-1} \tag{5.7}$$

Hence

$$\theta(N+1) = (\Phi^T\Phi)^{-1}\Phi^T y - (\Phi^T\Phi)^{-1}\varphi^T\{1 + \varphi(\Phi^T\Phi)^{-1}\varphi^T\}^{-1}\varphi(\Phi^T\Phi)^{-1}\Phi^T y$$
$$+ (\Phi^T\Phi)^{-1}\varphi^T y_{N+1} - (\Phi^T\Phi)^{-1}\varphi^T\{1 + \varphi(\Phi^T\Phi)^{-1}\varphi^T\}^{-1}\varphi(\Phi^T\Phi)^{-1}\varphi^T y_{N+1}$$

We find that the sum of the last two terms is

$$(\Phi^T\Phi)^{-1}\varphi^T\{1 + \varphi(\Phi^T\Phi)^{-1}\varphi^T\}^{-1}\{1 + \varphi(\Phi^T\Phi)^{-1}\varphi^T - \varphi(\Phi^T\Phi)^{-1}\varphi^T\}y_{N+1}$$
$$= (\Phi^T\Phi)^{-1}\varphi^T\{1 + \varphi(\Phi^T\Phi)^{-1}\varphi^T\}^{-1}y_{N+1}$$

But it follows from (5.5)

$$\theta(N) = (\Phi^T\Phi)^{-1}\Phi^T y$$

Further introduce

$$K(N) = (\Phi^T\Phi)^{-1}\varphi^T\{1 + \varphi(\Phi^T\Phi)^{-1}\varphi^T\}^{-1}$$

and we find that the equation (5.6) can be written as

$$\theta(N+1) = \theta(N) + K(N)\{y_{N+1} - \varphi(N+1)\theta(N)\} \tag{5.8}$$

Notice that this equation has a strong intuitive appeal. The next estimate $\theta(N)$ is formed by adding a correction to the previous estimate. The correction is proportional to $y_{N+1} - (N+1)\theta(N)$ The term θ would be the value of y at time N + 1 if the model was perfect and there were no disturbances. The correction term is thus proportional to the difference between the measured value of y_{N+1} and the prediction of y_{N+1} based on the previous modelparameters. The components of the vector K(N) are weighting factors which tells how the corrections and the previous estimate should be weighted.

In order to obtain a recursive equation also for the weighting factors K(N) we introduce the quantity P(N) defined by

$$P(N) = \alpha\{\Phi^T(N)\Phi(N)\}^{-1} \tag{5.9}$$

The weighting factors D(N) can then be written as

$$K(N) = P(N)\,\varphi^T\{\alpha + \varphi\,P(N)\varphi^T\}^{-1} \tag{5.10}$$

Introducing the matrix P(N) defined by (5.9) into (5.7) and we get

$$\begin{aligned} P(N+1) &= P(N) - P(N)\varphi^T\{\alpha + \varphi\,P(N)\varphi^T\}^{-1}\varphi\,P(N) \\ &= P(N) - K(N)\,\varphi\,P(N) \\ &= \{I - K(N)\varphi(N+1)\}P(N) \end{aligned} \tag{5.1}$$

Summarizing we thus find that the least squares parameter estimate can be computed using the following recursive equations

$$K(N) = P(N)\varphi^T(N+1)\{\alpha + \varphi(N+1)P(N)\varphi^T(N+1)\}^{-1} \tag{5.11}$$

$$P(N+1) = \{I - K(N)\varphi(N+1)\}P(N) \tag{5.12}$$

$$\theta(N+1) = \theta(N) + K(N)\{y_{N+1} - \varphi(N+1)\theta(N)\} \tag{5.13}$$

Notice that the definition of P(N) given by the equation (5.9) requires that the matrix $\Phi^T\Phi$ is nonsingular. It follows from the equations (2.2) that

$$\Phi^T(N)\,\Phi(N) = \sum_{t=1}^{N} \varphi(t)\,\varphi^T(t) \tag{5.14}$$

where $\varphi(t)$ is the vector defined by the equation (5.4). Hence for small N $\Phi^T\Phi$ can never be positive definite. We must in fact require that N is at least equal to the number of parameters n before $\Phi^T\Phi$ could be positive definite. In order to use the recursive equations (5.11) we thus have to choose an N-value N_o so large that $\Phi^T(N_o)\Phi(N_o)$ is positive definite and compute the initial values from

$$P(N_o) = \alpha\{\Phi^T(N_o)\,\Phi(N_o)\}^{-1} \tag{5.15}$$

$$\hat{\theta}(N_o) = \{\Phi^T(N_o)\,\Phi(N_o)\}^{-1}\,\Phi^T(N_o)\,Y(N_o) \tag{5.16}$$

In some cases it is however very desirable to use recursive equations in all steps. This can be done with an arbitrarily

small error by using the following device. If we use the recursive equation (5.12) for P with the initial condition

$$P(0) = P_o$$

where P_o is positive definite it follows from (5.15) that

$$P(N) = \{P_o^{-1} + \frac{1}{\alpha} \Phi^T(N)\ \Phi(N)\} \tag{5.17}$$

which can be made arbitrarily close to $\Phi^T\Phi$ by choosing P_o sufficiently large e.g. $P_o = \frac{1}{\varepsilon} I$. In section 6 we will also show that it is possible to give a statistical interpretation to this trick, it simply means that we have an apriori estimate of the parameter vector with covariance P_o.

Example

Assume that we would like to identify a model

$$y(t) + a_1 y(t-1) + \ldots + a_2 y(t-n) = b_1 u(t-1) + \ldots + b_n u(t-n)$$

Define the parameter vector as

$$\theta = \text{col}\ \{a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n\}$$

We find that the vector $\varphi(N)$ is given by

$$\varphi(N+1) = \{y(N), y(N-1), \ldots, y(N-n+1), u(N), u(N-1), \ldots, u(N-n+1)\}$$

and the estimate is then given recursively by the equations (5.11), (5.12) and (5.13). The recursive equations are very well suited for numerical computation of the least squares estimate in real time. An ALGOL procedure which evaluates the recursive formulas (5.11) - (5.13) is given below.

```
procedure LSRT (t,p,f,y,n,e,v);
comment The procedure iterates the least squares estimate
        one step. Notations: t, vector of estimated para-
        meters, p covariance matrix, f phi-vector, e resi-
        dual, n number of parameters, v variance of mea-
        surement errors;
integer n; real v,y,e; array t,P,f;
begin
integer i,j; real r; array K[1:10], fl[1:10];
comment Compute gainvector K;
for i:=1 step 1 until n do
begin r:=0;
for j:=1 step 1 until n do
r:=r + P[i,j] x f[j];
fl[i]:=r
end;

r=v;
for i:=1 step 1 until n do
r:=r + f[i] x fl[i];
for i:=1 step 1 until n do
K[i]:=fl[i]/r;
comment Update parameter estimates;
r:=y;
for i:=1 step 1 until n do
r:=r - f[i] x t[i];
for i:=1 step 1 until n do
t[i]:=t[i] + K[i] x r;
e:=r;
comment Update covariances matrix P;
for i:=1 step 1 until n do
for j:=i step 1 until n do
P[j,i]:=P[i,j]:=P[i,j] - K[i] x fl[j];
end procedure LSRT;
```

6. STATISTICAL INTERPRETATION OF THE LEAST SQUARES PROCEDURE

The identification problem discussed in this chapter can be stated as:
Given the linear model (2.1), characterized by the parameters θ_i, a sequence of observations $\{x_i, y_i, i = 1,2, \ldots, N\}$. Find those values of the parameter θ which minimized the criterion (2.4).

There are many additional problems which arise naturally e.g. why choose the criterion (2.4)? Is it worth while to include more parameters in the model? Is it possible to assign accuracy estimates to the parameter values obtained? In order to give partial answers to these questions we will put the problem in a statistical framework. To do so we assume that we have a model given by the equation (2.3) e.g.

$$y = \Phi\theta + e \tag{6.1}$$

where the residuals $\{e_i, i = 1, 2, \ldots, N\}$ are stochastic variables. In some cases we will assume that the probability distributions are known in other cases, it is sufficient to know first and second order moments only. It will be important to distinguish between the true parameter value and the estimate. We introduce the notations θ and $\hat{\theta}$ respectively. We now have

Theorem 6.1

The least squares estimate of θ is unbiased if e has zero mean and if Φ and e are independent.

Proof

The estimate $\hat{\theta}$ is given by (2.5) i.e.

$$\hat{\theta} = \{\Phi^T\Phi\}^{-1}\Phi^T y$$

Introducing y given by (6.1) we find

$$\hat{\theta} = \{\Phi^T\Phi\}^{-1}\Phi^T(\Phi\theta + e) = \theta + \{\Phi^T\Phi\}^{-1}\Phi^T e \tag{6.2}$$

Now if Φ and e are independent we get

$$E\hat{\theta} = \theta + E\{\Phi^T\Phi\}^{-1}\Phi^T e = \theta + E\{\Phi^T\Phi\}^{-1}\Phi^T \cdot Ee = \theta$$

The variance of the estimate is given by

Theorem 6.2

If Φ and e are independent and if e_i are equally distributed with zero mean and variance σ^2 we have

$$E\{\hat{\theta} - \theta\}\{\hat{\theta} - \theta\} = \sigma^2\{\Phi^T\Phi\}^{-1} \tag{6.2}$$

If the variance is not known it can be estimated by

$$s^2 = \frac{1}{N - n}(y - \Phi\hat{\theta})^T(y - \Phi\hat{\theta}) \tag{6.3}$$

Proof

We have

$$E\{\hat{\theta} - \theta\}\{\hat{\theta} - \theta\}^T = E\{\Phi^T\Phi\}\Phi^T ee^T\Phi\{\Phi^T\Phi\}^{-1}$$

$$= \{\Phi^T\Phi\}^{-1}\Phi^T(Eee^T)\Phi\{\Phi^T\Phi\}^{-1} = \{\Phi^T\Phi\}^{-1}\Phi^T\sigma^2 I\Phi\{\Phi^T\Phi\}^{-1}$$

$$= \sigma^2\{\Phi^T\Phi\}^{-1}$$

where the first equality follows by (2.5), the second by Φ and e being independent and the third from $Eee^T = \sigma^2 I$.

To prove the last statement of the theorem we observe that

$$y - \Phi\hat{\theta} = \Phi\theta + e - \Phi\theta + \{\Phi^T\Phi\}^{-1}\Phi^T e = I_N - \Phi\{\Phi^T\Phi\}^{-1}\Phi^T e$$

where I_N is the N x N unit matrix. We also find that $I_N - \Phi\{\Phi^T\Phi\}^{-1}\Phi^T$ is a projection because

$$I_N - \Phi\{\Phi^T\Phi\}^{-1}\Phi^T \; I_N - \Phi\{\Phi^T\Phi\}^{-1}\Phi^T =$$

$$= I_N - \Phi\{\Phi^T\Phi\}^{-1}\Phi^T - \Phi\{\Phi^T\Phi\}^{-1}\Phi^T + \Phi\{\Phi^T\Phi\}^{-1}\Phi^T\Phi\{\Phi^T\Phi\}^{-1}\Phi^T$$

$$= I_N - \Phi\{\Phi^T\Phi\}^{-1}\Phi^T$$

Hence

$$E\{y - \Phi\hat{\theta}\}^T\{y - \Phi\hat{\theta}\} = E\{e^T \left[I_N - \Phi\{\Phi^T\Phi\}^{-1}\Phi^T\right]^2 e\}$$

$$= E \text{ tr } e^T\{I_N - \Phi(\Phi^T\Phi)^{-1}\Phi^T\}e =$$

$$= E \text{ tr } \{I_N - \Phi(\Phi^T\Phi)^{-1}\Phi^T\}ee^T = \sigma^2 \text{ tr}\{I_N - \Phi(\Phi^T\Phi)^{-1}\Phi^T\}$$

$$= \sigma^2(\text{tr } I_N - \text{tr } \Phi(\Phi^T\Phi)^{-1}\Phi^T) =$$

$$= \sigma^2(\text{tr } I_N - \text{tr}(\Phi^T\Phi)^{-1} \Phi^T\Phi) = \sigma^2(\text{tr } I_N - \text{tr } I_n) =$$

$$= (N - n)\sigma^2$$

We thus find

$$E\, s^2 = \frac{1}{N - n} E\{y - \Phi\hat{\theta}\}^T\{y - \Phi\hat{\theta}\} = \sigma^2$$

Equation (2.5) thus gives an unbiased estimate of the variance σ^2.

Under the assumptions of Theorem 6.2 the least squares estimate $\hat{\theta}$ is thus unbiased, its covariance is $\sigma^2(\Phi^T\Phi)^{-1}$. Theorem 6.2 thus makes it possible to estimate the accuracy of the parameter values obtained by the least squares procedure.

So far we have only made assumptions on the first and second order moments of the numbers e_i. We will now assume that all e_i are normal $(0,\sigma)$. In this case we could try to estimate the parameters using the maximum likelihood procedure. To do so we first determine the probability density function of the observations y_i. We have

$$y_i \in N(\theta, \varphi_1(x_i) + \theta_2\varphi_2(x_i) + \ldots + \theta_n\varphi_n(x_i), \sigma) \qquad (6.4)$$

The probability density function of the observations is then

$$F(y|\theta) = (2\pi)^{-N/2}\sigma^{-N} \exp - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \{y_i - \theta_1\varphi_1(x_i) - \ldots - \theta_n\varphi_n(x_i)\}^2 \qquad (6.5)$$

To maximize the likelihood function we thus have to choose θ_i such that the loss function

$$2V = \sum_{i=1}^{N} \{y_i - \theta_1\varphi_1(x_i) - \theta_2\varphi_2(x_i) - \ldots - \theta_n\varphi_n(x_i)\}^2 \qquad (6.6)$$

is minimal. Having found the estimate of the standard deviation is then obtained by

$$\sigma^2 = \frac{2V_o}{N} \tag{6.7}$$

When the residuals are independent and normal the least squares estimate is thus equivalent to the maximum-likelihood estimate.

In order to investigate the efficiency of the estimate, we will now compute the information matrix associated with the estimation problem. This matrix is defined by

$$J = E(\text{grad}_\theta \log f)(\text{grad}_\theta \log f)^T \tag{6.8}$$

Equation (6.5) gives

$$\log f = - N \log \sigma - \frac{1}{2\sigma^2} (y - \Phi\theta)^T(y - \Phi\theta) - N \log 2\pi$$

Hence

$$\text{grad}_\theta \log f = \frac{1}{\sigma^2} \Phi^T(y - \Phi\theta) = \frac{1}{\sigma^2} \Phi^T e^T$$

We thus get

$$J = E \frac{1}{\sigma^4} \Phi^T e e^T \Phi = \frac{1}{\sigma^2} \Phi^T \Phi$$

We have previously shown that under the assumptions of Theorem 6.2 the covariance matrix of the estimate is given by (6.2) and we can thus also conclude that the estimate is efficient.

Summarizing the main results of this chapter, we thus find that by putting the identification problem in a statistical framework we have a natural way of introducing the loss function. We also find it possible to estimate the accuracy of the parameter estimate.

7. ANALYSIS OF THE RESIDUALS AND TEST OF THE ORDER OF THE MODEL

When solving an identification problem in practice, it is very unusual to know the model structure a priori. This means e.g. the we seldom know the order of the model and what terms of (2.1) that are of importance.

In this section we will briefly analyse these matters. It should, however, be pointed out that it is very difficult to answer such problems in a way that is acceptable from practical as well as a theoretical point of view.

It seems quite natural to analyse the residuals ε defined by

$$\varepsilon = Y - \Phi\hat{\theta} \tag{7.1}$$

where $\hat{\theta}$ is the least squares estimate given by (2.5) i.e.

$$\hat{\theta} = (\Phi^T\Phi)^{-1}\Phi^T Y \tag{7.2}$$

From a purely deterministic point of view the residuals will give an indication of how well the measurements are explained by the model. It is in particular often quite revealing just to plot the residuals and analyse for large deviations, jumps or trends.

From a statistical point of view it seems even more important to analyse the residuals. When the problem was stated in statistical terms it was actually assumed that

$$Y = \Phi\theta + e \tag{7.3}$$

where e was a vector of random variables having certain properties. If the estimate $\hat{\theta}$ equals the true parameters θ, we thus find

$$\varepsilon = e \tag{7.4}$$

In this particular case the residuals should thus equals the stochastic variables e and we can thus make statistical tests

to see if the data confirms or contradicts the assumptions made on the statistical character of the errors e. We can thus test that the residuals obtained can be samples from a normal distribution. We can evaluate their covariance function to see if they are dependent.

In practice we have found it convenient to include the following

1. Plot the residuals
2. Check the distribution of the residuals (Histogram, χ^2-test etc.)
3. Compute sample covariance function of the residuals. Test if residuals are uncorrelated.

The order of the model is one structural problem that often is of interest. If the order is not known, we can compute the least squares estimates for models of successively increasing order. The loss function will naturally decrease as the order of the model increases and the problem is thus to judge whether the decrease in the loss function is statistically significant. Such a test can be carried out in many ways. One commonly used test is based on the following result.

Theorem 7.1

Assume that the residuals e_i of the model (2.1) are independent normal $(0,\sigma)$. Let $\hat{\theta}_i$ be the least squares estimate based on a model of order n_i and let V_i be the corresponding loss-function defined by

$$V_i = \{y - \Phi_i\hat{\theta}_i\}^T\{y - \Phi_i\hat{\theta}_i\} \tag{7.5}$$

Let the nullhypothesis be H_o: $n_2 > n_1 \geq n$, where n is the order of the system. If the nullhypothesis is true then the quantities V_2 and $(V_1 - V_2)$ are independent random variables having χ^2 distributions with $N - n_2$ and $n_2 - n_1$ degrees of freedom.

Proof

We have

$$\hat{\theta}_i = (\Phi_i^T \Phi_i)^{-1} \Phi_i^T y \qquad i = 1,2$$

where

$$y = \Phi_n \theta_n + e$$

Partitioning Φ_i as follows

$$\Phi_i = \{\Phi_n \mid \psi_i\}$$

we find

$$\left[\begin{array}{c|c} \Phi_n^T \Phi_n & \Phi_n^T \psi_i \\ \hline \psi_i^T \Phi_n & \psi_1^T \psi_i \end{array}\right] \hat{\theta}_1 = \begin{bmatrix} \Phi_n^T \\ \psi_i^T \end{bmatrix} y = \begin{bmatrix} \Phi_n^T \Phi_n \\ \Phi_i^T \Phi_n \end{bmatrix} \theta_n + \Phi_1^T e$$

Solving this equation e.g. by the principle of superposition we find

$$\hat{\theta}_i = \left[\begin{array}{c} \theta_n \\ \hline 0 \end{array}\right] + (\Phi_i^T \Phi_i)^{-1} \Phi_i^T e$$

We thus find that

$$y - \Phi_i \hat{\theta}_i = \Phi_n \theta_n + e - \Phi_n \theta_n - (\Phi_i^T \Phi_i)^{-1} \Phi_i^T e$$

The loss functions V_1 and V_2 can thus be expressed as follows

$$V_i = e^T \left[I_n - \Phi_i (\Phi_i^T \Phi_i)^{-1} \Phi_i^T\right] e \qquad i = 1,2$$

We have furthermore the identity

$$V_2 + (V_1 - V_2) + (e^T I_n e - V_1) = e^T I_n e$$

where the quantities on the left are nonnegative definite quadratic forms in e_i. We will now investigate the properties of the matrices of the quadratic forms. We have

$$\text{rank } \Phi_i (\Phi_i^T \Phi_i)^{-1} \Phi_i^T = \text{tr } \Phi_i (\Phi_i^T \Phi_i)^{-1} \Phi_i^T = \text{tr}(\Phi_i^T \Phi_i)^{-1} \Phi_i^T \Phi_i =$$

$$= \text{tr } I_{n_i} = n_i$$

because the matrix $\Phi_i(\Phi_i^T\Phi_i)^{-1}\Phi_i^T$ is idependent and thus has eigenvalues zero or one.

Hence

$$\text{rank}\left[I_N - \Phi_2(\Phi_2^T)^{-1}\Phi_2^T\right. = N - n_2$$

$$\text{rank}\left[\Phi_2(\Phi_2^T\Phi_2)^{-1}\Phi_2^T - \Phi_1(\Phi_1^T\Phi_1)^{-1}\Phi_1^T\right. = n_2 - n_1$$

$$\text{rank}\left[\Phi_1(\Phi_1^T\Phi_1)^{-1}\Phi_1^T\right] = n_1$$

It now follows from Cockrans Theorem (see e.g. Kendall Stuart, The Advanced Theory of Statistics, Vol I second edition, London 1962, p. 360-362) that the quantities V_2, $V_1 - V_2$ and $e^T I_N e - V_1$ are independent and χ^2 under the null hypothesis and the theorem is proven.

To test it the lossfunction is significantly reduced when the number of parameters are increased from n_1 to n_2 we can thus use the following test quantity

$$t = \frac{V_1 - V_2}{V_2} \quad \frac{N - n_2}{n_2 - n_1} \tag{7.6}$$

According to Theorem $V_1 - V_2$ and V_2 are independent and χ^2 and t thus has an $F(N - n_2, n_2 - n_1)$ distribution.

It should also be noted that if there are no disturbances the columns of the Φ matrix in equation (2.2) will be dependent if too many parameters are used. This implies that the matrix $\Phi^T\Phi$ will be singular for a model with to many parameters. Hence if the identification is carried out recursively in the number of parameters for noise free data the matrix $\Phi^T\Phi$ will become singular when the order of the model identified is greater than the order of the true model. By analysing the route of $\Phi^T\Phi$ when it is singular, we can, however, detect the linear dependence of its rows and thereby estimate redundant parameters.

8. IDENTIFICATION OF LINEAR SYSTEM DYNAMICS

As an example of the application of the least squares method we will now discuss the identification of linear system dynamics. Consider a linear, time-invariant single-input, single-output system. If the input signal u is piece-wise constant over sampling intervals of constant length the input-output relation of the system can always be represented by the generic model

$$y(t) + a_1 y(t-1) + \ldots + a_n y(t-n) = b_1 u(t-1) + \ldots + b_n u(t-n) + e(t) \tag{8.1}$$

If we perform experiments on the system by changing the input u and observing the output y, we can thus state the identification problem as follows:

Given a sequence of input output pairs {u(t), y(t), t = 1, 2, ..., N + n}, find a model of type (8.1) which fits the observed data as well as possible in the sense of least squares.

We can apparently mould this particular problem in the form of the general least squares formulation discussed in section 2 for example by introducing the quantities

$$y = \begin{bmatrix} y(n+1) \\ y(n+2) \\ \vdots \\ y(n+N) \end{bmatrix} \qquad \Phi = \begin{bmatrix} -y(n) - y(n-1) & \ldots & y(1) & u(n) & u(n-1) & \ldots & u(1) \\ -y(n+1) - y(n) & & -y(2) & u(n+1) & u(n) & \ldots & u(2) \\ \vdots & & & & & & \\ -y(N+n-1) - y(N+n-2) - y(N) & & & u(N+n-1) & & \ldots & u(N) \end{bmatrix}$$

$$\theta = \mathrm{col}\begin{bmatrix} a_1 & a_2 & \ldots & a_n & b_1 & b_2 & \ldots & b_n \end{bmatrix} \tag{8.2}$$

The least squares parameter estimate is then given by the equation (2.5). To evaluate this we need the matrices

$$\phi^T\phi = \begin{bmatrix}
\sum_{t=n}^{N+n-1} y^2(t) & \sum_{t=n}^{N+n-1} y(t)y(t-1) & \cdots & \sum_{t=n}^{N+n-1} y(t)y(t-n+1) & \sum_{t=n}^{N+n-1} y(t)u(t) & \sum_{t=n}^{N+n-1} y(t)u(t-1) & \cdots & \sum_{t=n}^{N+n-1} y(t)u(t-n+1) \\
 & \sum_{t=n-1}^{N+n-2} y^2(t) & & \sum_{t=n-1}^{N+n-2} y(t)y(t-n-2) & \sum_{t=n-1}^{N+n-2} y(t)u(t) & \sum_{t=n-1}^{N+n-2} y(t)u(t-1) & \cdots & \sum_{t=n-1}^{N+n-2} y(t)u(t-n+2) \\
 & & & \sum_{t=1}^{N} y^2(t) & \sum_{t=1}^{N} y(t)u(n-1) & \sum_{t=1}^{N} y(t)u(t+n-2) & & \sum_{t=1}^{N} y(t)u(t) \\
 & & & & \sum_{t=n}^{N+n-1} u^2(t) & \sum_{t=n}^{N+n-1} u(t)u(t-1) & & \sum_{t=n}^{N+n-1} u(t)u(t-n+1) \\
 & & & & & \sum_{t=n-1}^{N+n-2} u^2(t) & & \sum_{t=n-1}^{N+n-2} u(t)u(t-n+2) \\
 & & & & & & & \sum_{t=1}^{N} u^2(t)
\end{bmatrix} \tag{8.3}$$

$$\phi^T y = \begin{bmatrix} -\sum_{t=n+1}^{N+n} y(t)\, y(t-1) \\ -\sum_{t=n+1}^{N+n} y(t)\, y(t-2) \\ \vdots \\ -\sum_{t=n+1}^{N+n} y(t)\, y(t-n) \\ \sum_{t=n+1}^{N+n} y(t)\, n(t-1) \\ \sum_{t=n+1}^{N+n} y(t)\, u(t-2) \\ \vdots \\ \sum_{t=n+1}^{N+n} y(t)\, u(t-n) \end{bmatrix} \qquad (8.4)$$

We have thus no difficulties in formulating the identification of the model (8.1) as a least squares problem. It now remains to show that the least squares estimate has the desired properties of unbiased ness, efficiency etc. Notice that the proofs given in section 6 do not apply directly. The standard theorems from regression theory such as 6.1 and 6.2 requires that Φ and e are independent. This is certainly not the case for the model (8.1) as the elements of the Φ matrix are actually computed from the numbers e(1), e(2), ..., . The problem of unbiasedness of the least squares estimates for the model (8.1) was first studied by Mann and Wold 1943. They assumed $b_1 = b_2 = \ldots b_n = 0$, which is not essential and they showed that the least squares estimate

was asymptotically unbiased. Analyses of the exact distribution of the least squares estimate for short samples have also been given for first order systems by R.L. Andersson and T. Koopmans. It is clear from these analyses that the exact distributions are extremely difficult to handle even for systems of first order. We will now give a result which is a straight-forward generalization of the result of Mann and Wald.

Theorem

Let the residuals e(t) of the model (8.1) be independent, equally distributed with zero mean. Assume that all moments of e(t) exist and are finite. Let all the roots of

$$z^n + a_1 z^{n-1} + \ldots + a_n = 0$$

have magnitudes less than one. Assume that the limits

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} u(t) \quad \text{and} \quad \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} u(t)\, u(t+\tau) = R_u(\tau)$$

exist and let the matrix A defined by

$$A_{ij} = R_u(i-j) \qquad i,j = 1,2, \ldots, n$$

be positive definite. The least squares estimate $\hat{\theta}_N$ then converges to θ in mean square as $N\to\infty$.

Proof

The proof is essentially the same as the one given by Mann and Wald with the obvious modifications required to handle the b-terms. We will therefore here only give an outline and we refer to the paper by Mann and Wald for details. We first observe that the assumptions imply that the limits

$$\lim_{N\to\infty} \frac{1}{N} \Phi_N^T \Phi_N \ , \ \lim \frac{1}{N} \Phi_N^T e_N$$

exist in the sense of mean square. Introduce

$$z_N = \frac{1}{N} (\Phi^T_N \Phi_N)(\hat{\theta}_N - \theta_o) = \frac{1}{N} \Phi^T_N E_N$$

We have

$$E\, z_N = \frac{1}{N} E\, \Phi^T_N E_N = 0$$

Because e(t) is independent of y(t-1), y(t-2), ... and a typical element of $\Phi^T_N E_N$ is

$$y(k-1)\, e(k) + y(k-2)\, e(k) + \ldots + y(k-n) - e(k)$$

Furthermore

$$E\, z_N z_N^T = \frac{1}{N} E\{\Phi^T_N E_N E_N^T \Phi_N\} = \frac{1}{N} E\{\Phi_N^T \Phi_N\}$$

It now follows from ( ) that $\lim z_N$ exist in the sense of mean square convergence. Hence

$$\lim z_N = \lim \frac{1}{N} (\Phi_N^T \Phi_N)\ N\ (\hat{\theta}_N - \theta_o)$$

$$= A \cdot \lim\ N\ (\hat{\theta}_N - \theta_o)$$

and we thus find that $N\ (\hat{\theta}_N - \theta_o)$ converges to a random variable with zero mean and finite covariance. Hence

$$\lim\ (\hat{\theta}_N - \theta_o)^T (\hat{\theta}_N - \theta_o) = \frac{1}{N} \lim z_N^T z_N = 0$$

and we thus find that the least squares estimate is consistent. To find the asymptotic distribution of $\hat{\theta}_N$ we observe that

$$(\hat{\theta}_N - \theta_o) \sim \frac{1}{N} A^{-1} z_N$$

$$E(\hat{\theta}_N - \theta_o)(\hat{\theta}_N - \theta_o)^T = \frac{1}{N} A^{-1} E\{z_N z_N^T\} A \sim \frac{1}{N} A^{-1}$$

In the analysis presented in the previous sections we have seen the importance of the condition $\Phi^T\Phi$ being nonsingular. This condition will depend on the character of the input signal and of the properties of the true model. For the discussion which follows we will assume that the data was actually generated by the mechanism described by

$$y(t) + a_1^o y(t-1) + \ldots + a_{n_o}^o y(t-n_o) = b_1^o u(t-1) + \ldots + b_{n_o}^o u(t-n_o) + e(t) \qquad (8.5)$$

where $e(t)$ is a sequence of incorrelated equally distributed random variables with zero mean and variance σ^2. In the following it will be important to distinguish between the model (8.1) which is fitted to the data and the equation (8.5) which actually generated the data. We will therefore refer to (8.5) as the "true model". "The true model" is thus characterized by the parameters n_o, a_i^o and b_i^o, while the parameters of the fitted model are n, a_i and b_i. Notice in particular that the order of the fitted model n may differ from the order of the true model. We will now give a fundamental result concerning the regularity of the matrix $\Phi^T\Phi$. We have

Theorem 8.1

Let the output signal y be generated by the model (8.5) with $\sigma = 0$ i.e. $e(t) = 0$ for all t. Then the rank of the matrix $\Phi^T\Phi$ can never exceed $2n_o$ and there exist an input signal $\{u(t), t = 1,2,\ldots,3n\}$ such that $\Phi^T\Phi$ has rank $2n_o$.

Proof

Consider the matrix Φ given by the equation (8.2). Let $n > n_o$. If the output y was generated by the model (8.5) with $e(t) = 0$ for all t, we find that the columns of the Φ matrix are linearly dependent. By utilizing the equation (8.5) we find that all columns can be expressed as linear combinations of $2n_o$ columns. The rank of Φ^T is thus at most

equal to $2n_o$. As the rank of a product of two matrices cannot exceed the rank of either factor, we thus find that the rank of $\Phi^T\Phi$ is at most $2n_o$. It then remains to show the existence of a suitable input sequence.

Example 8.1

We will now give a numerical example. To do so we will generate input-output data from a known model.

In Fig. 8.1 and in Table 8.1 we show input output pairs which are generated by the equation

$$y(t) + 1.5y(t-1) + 0.7y(t) = u(t-1) + 0.5u(t-2) + \lambda e(t)$$

where $\{e(t)\}$ is a sequence of independent normal (0.1) random numbers generated by a pseudo random generator. The following values of λ have been used: 0, 0.1, 0.5, 1.0 and 5.0.

In Table 8.2 we show the results obtained when a model having the structure (8.1) is fitted to the generated input-output data using the least squares procedure.

The estimates are calculated recursively for models of increasing order to illustrate the very typical situation in practice when the order of the model is not known. In Table 8.2 we have shown the least squares parameter estimates and their estimated accuracy, the loss function and a conditioning number of the matrix $\Phi^T\Phi$. The conditioning number $\mu = 2n \max\{(A)_{ij}\} \max\{(A^{-1})_{ij}\}$ is chosen rather arbitrarily.

Table 8.1 - Input-output data for example 8.1

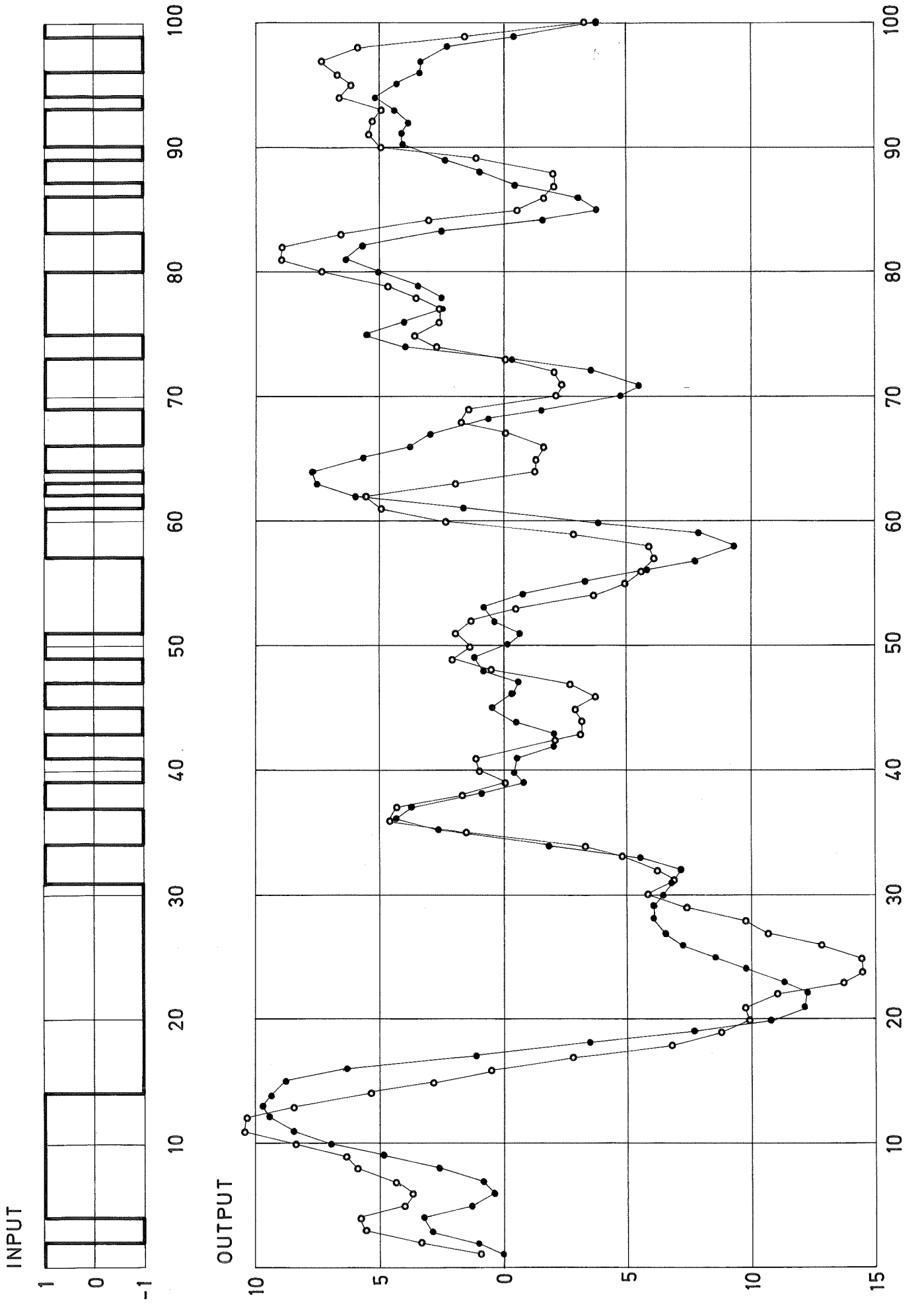
t	u	e	y				
			$\lambda = 0$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 5.0$
1	1	0.857	0.000	0.086	0.429	0.857	4.28
2	1	0.915	1.000	1.220	2.100	3.200	12.00
3	-1	-0.099	3.000	3.260	4.301	5.602	16.01
4	-1	0.246	3.300	3.561	4.604	5.908	16.34
5	1	0.635	1.350	1.623	2.713	4.076	14.98
6	1	1.247	0.215	0.566	1.970	3.726	17.77
7	1	0.172	0.878	1.230	2.642	4.407	18.52
8	1	0.401	2.666	2.990	4.285	5.904	18.85
9	1	-0.847	4.884	5.038	5.654	6.423	12.58
10	1	1.383	6.961	7.103	7.673	8.386	14.09
11	1	0.960	8.522	8.724	9.532	10.54	18.62
12	1	-0.994	9.410	9.514	9.930	10.45	14.60
13	1	-1.336	9.650	9.531	0.054	8.458	3.69
14	1	-1.447	9.388	8.992	7.407	5.426	-10.42
15	-1	-0.826	8.827	8.233	5.860	2.892	-20.85
16	-1	0.541	6.169	5.610	3.375	0.581	-21.77
17	-1	-0.122	1.574	1.139	-0.600	-2.775	-20.17
18	-1	-0.749	-3.457	-3.793	-5.137	-6.818	-20.26
19	-1	1.024	-7.787	-7.884	-8.274	-8.761	-12.66
20	-1	-0.027	-10.76	-10.67	-10.33	-9.895	-6.43
21	-1	0.442	-12.19	-11.95	-10.98	-9.769	-0.08
22	-1	-1.790	-12.25	-12.13	-11.63	-11.02	-6.07
23	-1	-2.487	-11.35	-11.58	-12.51	-13.67	-22.98
24	-1	-0.165	-9.942	-10.39	-12.20	-14.46	-32.55
25	-1	-0.797	-8.471	-9.066	-11.45	-14.42	-38.22
26	-1	0.312	-7.247	-7.792	-9.971	-12.70	-34.49
27	-1	-0.127	-6.441	-6.854	-8.508	-10.57	-27.11
28	-1	-1.196	-6.088	-6.447	-7.880	-9.672	-24.01
29	-1	1.236	-6.124	-6.248	-6.747	-7.369	-12.35
30	-1	0.107	-6.424	-6.349	-6.050	-5.677	-2.69
31	-1	-1.882	-6.849	-6.838	-6.794	-6.738	-6.30
32	1	1.428	-7.277	-7.170	-6.742	-6.206	-1.92
33	1	-0.963	-5.621	-5.565	-5.338	-5.055	-2.79
34	1	-1.566	-1.838	-1.985	-2.571	-3.305	-9.17

Table 8.1 - Input-output data for example 8.1

t	u	e	y				
			$\lambda = 0$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 5.0$
35	-1	1.395	2.678	2.558	2.077	1.477	-3.33
36	-1	0.671	4.804	4.793	4.752	4.699	4.28
37	-1	-0.067	3.831	3.893	4.140	4.448	6.92
38	1	-0.132	0.884	0.970	1.317	1.751	5.22
39	1	-0.050	-0.856	-0.774	-0.447	-0.037	3.24
40	-1	0.855	-0.403	-0.255	0.335	1.073	6.98
41	-1	0.078	-0.505	-0.333	0.355	1.214	8.09
42	1	-0.572	-1.975	-1.878	-1.489	-1.002	2.89
43	1	-1.242	-2.110	-2.208	-2.602	-3.095	-7.04
44	-1	-0.658	-0.282	-0.563	-1.690	-3.099	-14.37
45	-1	0.164	0.554	0.217	-1.132	-2.818	-16.31
46	1	-0.140	-0.472	-0.794	-2.085	-3.698	-16.60
47	1	0.504	-0.595	-0.793	-1.583	-2.570	-10.47
48	-1	0.281	0.937	0.895	0.726	0.515	-1.18
49	-1	0.156	1.322	1.413	1.775	2.227	5.84
50	1	-0.112	-0.172	-0.018	0.598	1.368	7.53
51	1	1.022	-0.684	-0.414	0.666	2.015	12.81
52	-1	-2.157	0.594	0.676	1.001	1.408	4.66
53	-1	-0.732	0.870	0.730	0.170	-0.530	-6.13
54	-1	-0.326	-0.610	-0.910	-2.109	-3.607	-15.59
55	-1	1.644	-3.025	-3.212	-3.960	-4.896	-12.38
56	-1	0.727	-5.610	-5.608	-5.601	-5.591	-5.52
57	-1	0.298	-7.798	-7.634	-6.980	-6.162	0.38
58	1	0.960	-9.269	-8.929	-7.569	-5.869	7.73
59	1	1.340	-7.946	-7.416	-5.298	-2.650	18.53
60	1	0.823	-3.930	-3.291	-0.737	2.456	28.00
61	1	-2.088	1.167	1.545	3.059	4.951	20.09
62	-1	-1.453	6.001	5.977	5.878	5.755	4.77
63	1	-2.715	7.685	7.112	4.818	1.951	-20.99
64	-1	-0.615	7.827	6.923	3.305	-1.217	-37.39
65	1	2.474	5.861	5.153	2.322	-1.217	-29.53
66	1	-1.249	3.812	3.259	1.045	-1.722	-23.86
67	-1	0.094	3.116	2.790	1.489	-0.138	-13.15
68	-1	1.333	1.505	1.538	1.669	1.832	3.14

Table 8.1 - Input-output data for example 8.1

t	u	e	y				
			$\lambda = 0$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 5.0$
69	-1	0.135	-1.423	-1.133	0.028	1.479	13.09
70	1	-1.591	-4.689	-4.435	-3.421	-2.154	7.98
71	1	1.426	-5.537	-5.217	-3.939	-2.341	10.44
72	1	-1.589	-3.523	-3.380	-2.808	-2.093	3.63
73	1	-0.078	0.091	0.074	0.007	-0.078	-0.76
74	-1	0,035	4.103	3.981	3.493	2.883	-2.00
75	-1	-0.144	5.591	5.405	4.663	3.735	-3.69
76	1	0.571	4.014	3.878	3.334	2.655	-2.78
77	1	0.768	2.607	2.610	2.622	2.636	2.75
78	1	0.052	2.601	2.706	3.125	3.648	7.84
79	1	-0.297	3.577	3.702	4.203	4.830	9.84
80	1	1.216	5.044	5.281	6.226	7.407	16.86
81	-1	-0.130	6.563	6.816	7.831	9.099	19.25
82	-1	1.137	5.813	6.142	7.457	9.101	22.25
83	-1	0.851	2.626	3.026	4.630	6.633	22.66
84	1	1.084	-1.631	-1.151	0.766	3.163	22.34
85	1	-1.103	-3.784	-3.456	-2.143	-0.502	12.63
86	1	-0.169	-3.034	-2.895	-2.335	-1.636	3.96
87	-1	-1.443	-0.403	-0.567	-1.224	-2.045	-8.62
88	1	0.439	1.020	0.719	-0.482	-1.984	-14.00
89	1	2.151	2.312	2.191	1.709	1.107	-3.71
90	-1	0.472	4.254	4.330	4.637	5.021	8.09
91	1	-0.691	4.262	4.393	4.914	5.566	10.78
92	1	-0.056	3.916	4.052	4.597	5.278	10.73
93	1	-0.043	4.390	4.499	4.934	5.478	9.83
94	-1	0.733	5.344	5.485	6.050	6.755	12.40
95	1	0.433	4.443	4.622	5.337	6.231	13.38
96	1	1.572	3.424	3.750	5.057	6.690	19.76
97	-1	0.196	3.525	3.910	5.448	7.370	22.75
98	-1	0.051	2.392	2.745	4.157	5.922	20.04
99	-1	-0.580	-0.381	-0.178	0.632	1.645	9.74
100	1	-0.094	-3.745	-3.698	-3.509	-3.273	-1.38


INPUT
1
0
-1
10
20
30
40
50
60
70
80
90
100
OUTPUT
10
5
0
5
10
15
10
20
30
40
50
60
70
80
90
100

We will now analyse the results of Table 8.2. Let us first consider the case of no disturbances $\lambda = 0$. In this case we find that it is only possible to compute models of order 1 and 2. When we try to compute the third order model, we find that the matrix $\Phi^T\Phi$ is singular as would be expected from Theorem 8.1. The conditioning number is 1.3×10^6. (To handle the numerical problems for a model of third order in a case like this, we must use numerical methods which do not require the inversion of $\Phi^T\Phi$ e.g. the reduction of Φ to triangular form using the QR algorithm). We also find that the estimated standard deviation of the coefficients of the second order model are zero.

Proceeding to the case of $\lambda = 0.1$, i.e. the standard deviation of the disturbances is one tenth of the magnitude of the input signals, we find that the matrix $\Phi^T\Phi$ is still badly conditioned when a third order model is computed.

Analysing the details we find, however, that the Gauss Jordan method gives a reasonable accurate inverse of $\Phi^T\Phi$. Pre-and postmultiplying the matrix with its computed inverse, we find that the largest off-diagonal element is 0.011 and the largest deviation of diagonal elements from 1.000 is 0.0045. We also find that the estimates $\hat{a}_3$ and $\hat{b}_3$ do not differ significantly from zero.

Proceeding to the cases with more disturbances we find that the matrix $\Phi^T\Phi$ no longer becomes badly conditioned for models of order 3, but that the estimates $\hat{a}_3$ and $\hat{b}_3$ still do not differ significantly from zero. Notice also that the standard deviation of the parameters in all cases increases as the order of the model is increased from 2 to 3.

Table 8.2 - Least squares estimates of the parameters of a model with structure (8.1) computed from data of table 8.1

	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	$\hat{\lambda}$	V	μ
True	-1.50	-0.70	0	1.0	0.5	0.0			
	-0.88±0.03			1.23±0.18			1.71	265.863	59
λ = 0.0	-1.50±0.00	0.70±0.00		1.00±0.00	0.50±0.00		0.00	0.000	203
	-	-	-	-	-	-	-	-	1.3×10^6
	-0.88±0.03			1.19±0.18			1.66	248.447	60
λ = 0.1	-1.50±0.01	0.69±0.01		0.99±0.01	0.49±0.02		0.11	0.987	205
	-1.52±0.11	0.73±0.16	-0.02±0.08	0.99±0.01	0.47±0.11	-0.02±0.06	0.11	0.983	35974
	-0.88±0.03			1.10±0.17			1.59	227.848	63
λ = 0.5	-1.48±0.04	0.67±0.03		0.96±0.06	0.48±0.07		0.53	24.558	206
	-1.54±0.11	0.76±0.16	-0.04±0.08	0.96±0.06	0.42±0.12	-0.04±0.09	0.53	24.451	1518
	-0.88±0.03			1.00±0.20			1.84	308.131	75
λ = 1.0	-1.47±0.06	0.66±0.06		0.93±0.12	0.46±0.14		1.06	99.863	212
	-1.55±0.11	0.82±0.17	-0.09±0.09	0.93±0.13	0.37±0.16	-0.02±0.15	1.07	98.698	476
	-0.86±0.05			0.20±0.83			7.55	5131.905	520
λ = 5.0	-1.48±0.07	0.74±0.08		0.98±0.61	0.41±0.61		5.29	2462.220	1174
	-1.55±0.11	0.87±0.18	-0.09±0.09	0.97±0.64	0.24±0.66	0.13±0.64	5.32	2440.245	2031

We will also discuss some other ways to find the order of the system. We can e.g. consider the variances of the parameters. We find e.g. from table 8.2 that the coefficients $\hat{a}_3$ and $\hat{b}_3$ do not differ significantly from zero in any case. In table 8.3 we also show summarize the values of the lossfunction as well as the values of the F-testvariable when testing the reduction of the lossfunction for a model of order n_1 compared to a model of order n_2 as was discussed in section 8. Compare equation (7.6). We have at the 10% level $F(2,100) = 2.37$ and at the 1% level $F(2,100) = 4.84$. We thus find that by applying the F-test or the χ^2-test in this case we get as a result that the system is of second order for all samples. The actual parametervalues of table 8.2 as well as the estimated accuracies gives an indication of the accuracy that can be obtained in a case like this.

Table 8.3 - Gives the values of the lossfunction V, the conditioning number μ of $\Phi^T\Phi$ and a table of F-test values when identifying models of different order to the samples of Table 8.2.

case 1 $\lambda = 0$

n	V	μ
0	2666.23	59
1	265.86	205
2	0.00	35974

n_1 \ n_2		1	2
0		442.4	∞
1			∞

case 2 $\lambda = 0.1$

n	V	μ
0	2644.15	
1	248.447	60
2	0.987	205
3	0.983	35974

n_1 \ n_2		1	2	3
0		472	57000	42100
1			12000	5900
2				0.191

case 3 $\lambda = 0.5$

n	V	μ
0	2701.35	
1	227.848	63
2	24.558	206
3	24.451	1518
4	24.006	2982

n_1 \ n_2	1	2	3	4
0	532	2616	1715	1283
1		397	195	130
2			0.2	0.1
3				0.8

case 4 $\lambda = 1.0$

n	V	μ
0	3166.78	
1	308.131	212
2	99.863	476
3	98.698	873
4	96.813	1351
5	94.800	1647

n_1 \ n_2	1	2	3	4	5
0	455	734	487	365	292
1		100	50	33	25
2			0.55	0.72	0.80
3				0.88	0.92
4					0.96

case 5 $\lambda = 5.0$

n	V	μ
0	21467.64	
1	5131.905	520
2	2462.220	1174
3	2440.245	2031
4	2375.624	2910
5	2290.730	3847

n_1 \ n_2	1	2	3	4	5
0	156	185	122	92	75
1		52	26	18	14
2			0.21	0.94	1.2
3				1.2	1.5
4					1.6

9. COMPARISON WITH CORRELATION METHODS

It is of interest to compare the least squares procedure with other identification schemes. In this section we will investigate the relationship with the correlation method discussed in chapter 4 . When determining process dynamics for a single input single output system using correlation methods the following quantities are computed.

$$R_u(\tau) = \frac{1}{N-\tau} \sum_{t=1}^{N-\tau} u(t)\, u(t+\tau)$$

$$R_y(\tau) = \frac{1}{N-\tau} \sum_{t=1}^{N-\tau} y(t)\, y(t+\tau)$$

$$R_{yu}(\tau) = \frac{1}{N-\tau} \sum_{t=1}^{N-\tau} y(t)\, u(t+\tau)$$

$$R_{uy}(\tau) = \frac{1}{N-\tau} \sum_{t=1}^{N-\tau} u(t)\, y(t+\tau)$$

Comparing with the least squares identification of process dynamics discussed in section 8 we find that the elements of the matrices (8.3) and (8.4) of the least squares procedure are essentially correlations or crosscorrelations. Neglecting terms in the beginning and end of the series we find

$$\Phi^T\Phi = \begin{bmatrix} R_y(0) & R_y(1) & \dots & R_y(n-1) & R_{yn}(0) & R_{uy}(1) & \dots & R_{uy}(n-1 \\ & R_y(0) & & R_y(n-2) & R_{yu}(1) & R_{yu}(0) & & R_{uy}(n-2) \\ & & & R_y(0) & R_{yu}(n-1) & R_{yu}(n-2) & & R_{yu}(0) \\ & & & & R_u(0) & R_u(1) & & R_u(n-1) \\ & & & & & R_u(0) & & R_u(n-2) \\ & & & & & & & R_u(0) \end{bmatrix}$$

$$\Phi^T y = \begin{bmatrix} -R_y(1) \\ -R_y(2) \\ \\ -R_y(n) \\ R_{uy}(1) \\ R_{uy}(2) \\ \vdots \\ R_{uy}(n) \end{bmatrix}$$

Hence if a correlation analysis is performed, it is a simple matter to calculate the least squares estimate by forming the matrices $\Phi^T\Phi$ and $\Phi^T y$ from the values of the sample covariance functions and solving the equation (2.5). As the order of the system is seldom known apriori it is often convenient to compute the least squares estimate recursively as was discussed in section 4 .

10. RELATIONS TO KALMAN FILTERING THEORY

We will now show that the equations for the least squares estimate are easily obtained from Kalmans results on linear filtering. Apart from being a curiosity this result is of interest in connection with realtime identification. Using Kalmans results it is possible to obtain least squares estimates even in the case when the parameters of the process are random processes.

Kalman considers a dynamical system

$$x(t+1) = \Phi\, x(t) + e(t)$$

$$y(t) = C\, x(t) + v(t) \qquad (10.1)$$

where $\{e(t),\ t = 1,2,\ldots\}$ and $\{v(t),\ t = 1,2,\ldots\}$ are sequences of independent equally distributed random variables with zero means and covariance matrices R_1 and R_2 respectively. Kalman has proven the following theorem.

Theorem (Kalman)

Let the initial condition of (10.1) be a normal random variable (m,R_o). The best estimate of $x(t)$ (in the sense of least squares) given the observed outputs $y(1), y(2), \ldots, y(t)$ is given by the recursive equations

$$\hat{x}(t) = \Phi\, \hat{x}(t-1) + K(t)\,\{y(t) - C\,\Phi\,\hat{x}(t-1)\}$$

$$\hat{x}(0) = m \qquad (10.2)$$

where

$$K(t) = S(t)\, C^T \{C\, S(t)\, C^T + R_2\}^{-1}$$

$$S(t) = \Phi\, P(t-1)\, \Phi^T + R_1$$

$$P(t) = S(t) - K(t)\, C\, S(t)$$

$$S(0) = R_o \qquad (10.3)$$

The matrix S(t) has physical interpretation as the covariance matrix of the apriori estimate of x(t) given y(1), ..., y(t-1) and the matrix P(t) as the covariance of the posterior estimate of x(t) given y(1), ..., y(t).

Now consider the least squares identification of the system

$$y(t) + a_1 y(t-1) + \ldots + a_n y(t-n) = b_1 u(t-1) + \ldots + b_n u(t-n) + \lambda e(t) \qquad (10.4)$$

where {e(t)} is a sequence of normal (0,1) random variables. Introduce the coefficients of the model as state variables

$$\begin{aligned} x_1(t) &= a_1 \\ x_2(t) &= a_2 \\ &\vdots \\ x_n(t) &= a_n \\ x_{n+1}(t) &= b_1 \\ x_{n+2}(t) &= b_2 \\ &\vdots \\ x_{2n}(t) &= b_n \end{aligned} \qquad (10.5)$$

and define the following vector

$$C(t) = \{-y(t-1), -y(t-2) \ldots -y(t-n)\ u(t-1)\ u(t-2) \ldots u(t-n)\} \qquad (10.6)$$

As the coefficients are constants we have

$$x(t+1) = x(t)$$

The equation (10.4) can now be written as

$$y(t) = C(t)\, x(t) + e(t) \qquad (10.7)$$

and the least squares identification problem can be stated as a Kalman filtering problem with $\Phi = I$, $R_1 = 0$, $R_2 = \lambda^2$. Applying Kalmans theorem we find that the estimate is given by

$$\hat{x}(t) = \hat{x}(t-1) + K(t) \{y(t) - C \hat{x}(t-1)\}$$

$$K(t) = P(t) C(t) \{\lambda^2 + C(t) P(t) C^T(t)\}^{-1}$$

$$P(t) = P(t-1) - K(t) C(t) P(t-1) \qquad (10.8)$$

But these equations are identical to the equations (5.11) - (5.13) for the recursive computations of the least squares estimate. As we have no apriori estimate the initial conditions are $\hat{x}(0) = 0$ and $P(0) = \infty$. To be able to use the recursive equation for P we introduce $P(0) = \frac{1}{\varepsilon} I$ where $\varepsilon > 0$ is a small number and let $\varepsilon \to 0$ in the solution. Compare section 5 for a detailed discussion of the initial conditions.

11. REAL-TIME LEAST SQUARES, CONTRACTION MAPPINGS AND STOCHASTIC APPROXIMATIONS

In this section we will analyse the least squares estimation procedure from still another point of view. To start we will consider the equation (5.8) for recursive computation of the least squares estimate i.e.

$$\theta(t+1) = \theta(t) + K(t)\{y(t+1) - \varphi(t+1)\theta(t)\} \qquad (11.1)$$

We have previously shown that if K(t) was chosen according to the equations (5.11) and (5.12) the sequence $\{\theta(t)\}$ will converge to the time parameter value θ_o. We will now show that there are many other ways to choose K(t) which also ensures convergence. This results are of interest from two points of view.

1. It makes it possible to simplify the computations.
2. The methods used to prove the main results are extremely general and they can be used also in many other problems to construct convergency parameter estimates.

We thus find that if there were no disturbunces there are many ways to choose K(t) in such a way that the equation (11.1) will converge. Notice that by involving the principle of contraction mapping we do not depend heavily on the mapping T being linear. Stochastic versions of the principle of contraction mapping were first discussed by Robbins and Monro under the name of stochastic approximations. The results of Robbins and Monro have later been extended by Dvoretsky who has proven the following theorem.

Theorem 11.1

Let α_n, β_n and γ_n (n = 1,2,...) be non-negative real numbers satisfying

$$\lim_{n\to\alpha} \alpha_n = 0, \quad \sum_{n=1}^{\infty} \beta_n < \infty \quad \text{and} \quad \sum_{n=1}^{\infty} \gamma_n = \infty$$

Let ϑ be a real vector and $T_n (n = 1,2,\ldots)$ be measurable transformations satisfying

$$||T_n(r_1,\ldots,r_n) - \vartheta|| \leq \max \{\alpha_n, (1+\beta_n)||r_n - \vartheta|| - \gamma_n\} \quad (11.2)$$

for all real vectors $r_1,\ldots,r_n$. Let x_n and Y_n be random vectors and define

$$X_{n+1}(\omega) = T_n(X_1(\omega), \ldots, X_n(\omega)) + Y_n(\omega)$$

for $n \geq 1$. Then the conditions $E\, X_1^2 < \infty$

$$\sum_{n=1}^{\infty} E\, Y_n^2 < \infty$$

and $E\{Y_n | x_1,\ldots,x_n\} = 0$ with probability 1 for all n imply

$$\lim E\{(X_n - \vartheta)^2\} = 0$$

and

$$P\{\lim_{n\to\infty} x_n - \vartheta\} = 1$$

We know already that (11.1) will converge if K(t) is chosen according to (10.3). We will now use Dvoretskys theorem to prove that the sequence (11.1) converges to θ_o under very general conditions. In particular we will show that this is true for many choices of K(t). This result is important for practical reasons because it will make it possible to simplify the computations. When doing so it should, however, be observed that even if the sequence θ_n converges for choices of K(t) other than (10.3) the convergence might be slow. The result is also of importance for another reason. Because of the generality of Dvoretskys result it will be possible to set up converging stochastic approximation schemes for much more general identification problems.

We now have

Theorem 11.2

Let

1) $K(t)\,\varphi(t)$ non-negative definite

2) $\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} (K(t)\varphi(t) + \varphi^T(t)\,K^T(t)) \to A$

where A is positive definite.

3) $\sum_{t=1}^{\infty} ||K(t)||^2 < \infty$

and assume that $\{e(t),\ t = 1,2,\ldots\}$ are independent, have zero mean and finite variance then the sequence (11.1) will converge to θ_o with probability one.

Proof

The essence of the proof is to demonstrate that the conditions of the theorem imply that the sequence (11.1) satisfies the conditions of Dvoretskijs theorem. To simplify the writing we introduce

$$A_n = K(n)\,\varphi(n)$$

$$X_n = \theta(n) - \theta_o$$

$$Y_n = K(n)\,e(n+1)$$

The sequence (11.1) can then be written as

$$X_{n+1} = \{I - A_n\}\,X_n + Y_n \qquad (11.3)$$

The numbers K(n) may depend on y(t-1), y(t-2) etc, but according to the assumptions the numbers e(t) were independent (of themselves and of y(t-1), y(t-2), ...). They have zero mean and variance λ^2.

Hence

$$E\{Y_n | x_1, \ldots, x_n\} = 0$$

and

$$E\,Y_n^2 \leq E\,||K(n)||^2 \cdot e^2(n+1) \leq \lambda^2\,||K(n)||^2$$

It now follows from assumption 3) that

$$\sum_{n=1}^{\infty} E\,Y_n^2 \leq \lambda^2 \sum_{n=1}^{\infty} ||K(n)||^2 < \infty$$

It now remains to show that the linear transformation

$$T(x) = (I - A_n)x$$

satisfies the condition (11.2) of Dvoretskijs theorem. To do so we will first assume that all A_n are positive definite. Let a_n be the smallest and b_n the largest eigenvalue of A_n. As A_n is symmetric and positive definite we have $a_n > 0$. The condition 2) also implies $\Sigma\, a_n = \infty$. Now introduce a sequence of positive numbers c having the properties

$$c_n \to 0 \quad \text{and} \quad \sum_{n=1}^{\infty} a_n c_n = \infty$$

We will now estimate $||T_n(x)||$ where $||\cdot||$ denotes the matrix norm which is subordinate to the Euclidian vector norm. We consider two cases seperately

A) $||x|| < c_n$

$$||T_n(x)|| \leq (1 + ||A_n||)||x|| = (1 + b_n)||x||$$
$$\leq (1 + b_n) \cdot c_n$$

B) $||x|| \geq c_n$

$$||T_n(x)||^2 = ||x||^2 + ||A_n x||^2 - 2x^T A_n x$$
$$\leq (1 + b_n^2)||x||^2 - 2x^T A_n x$$

We further have

$$\min_{||x||=1} x^T A_n x = a_n$$

where

$$\sum_{1}^{\infty} a_n = \infty$$

Hence

$$||T_n(x)||^2 \leq (1 + b_n^2)||x||^2 - 2a_n||x||^2$$

$$||T_N(x)|| \leq \sqrt{1 + b_n^2}\ ||x||\ \sqrt{1 - \frac{2a_n}{1 + b_n^2}}$$

For n sufficiently large we have

$$\sqrt{1 - \frac{2a_n}{1 + b_n^2}} \sim 1 - \frac{a_n}{1 + b_n^2}$$

Hence

$$||T_n(x)|| \leq ||x|| \sqrt{1 + b_n^2} - \frac{a_n}{\sqrt{1 + b_n^2}} \cdot ||x||$$

$$\leq ||x|| \sqrt{1 + b_n^2} - \frac{a_n c_n}{\sqrt{1 + b_n^2}}$$

where

$$\sum_{1}^{\infty} a_n c_n = \infty$$

according to the definition of c_n. As $b_n \to 0$ we have also

$$\sum_{1}^{\infty} \frac{a_n c_n}{\sqrt{1 + b_n^2}} = \infty$$

We thus find

$$||T(x)|| \leq \max\{\alpha_n, (1 + \beta_n)||x|| - \gamma_n\}$$

where

$$\alpha_n = (1 + b_n)c_n$$

$$\beta_n = \sqrt{1 + b_n^2} - 1$$

$$\gamma_n = \frac{a_n c_n}{\sqrt{1 + b_n^2}}$$

The numbers α_n, β_n and γ_n obviously satisfy the requirements of Dvoretskijs theorem. We have thus proven the theorem for the special case that att A_n are positive definite. We will now remove this condition. When A_n are not positive definite we start with (11.3) and iterate to twice. Hence

$$x_{n+K} = \left(\prod_{\nu=n}^{n+k-1} A_\nu\right)x_N + \sum_{\mu=n}^{n+k-1} \left(\prod_{\nu=\mu}^{n+k-1} A_\nu\right) K(\mu)\, e(\mu)$$

The product

$$\prod_{\nu=n}^{n+k-1} A_\nu$$

must become positive definite for a finite k because of 2). Now define the sequence

$$z_\nu = x_{n+\nu k}$$

We have

$$z_{\nu+1} = A_\nu^{\times} z_\nu + v_\nu \qquad (11.4)$$

where

$$A^{\times}_{\nu} = \prod_{\mu=n+\nu k}^{n+\nu k+k}$$

and

$$V_\nu = \sum_{\mu=n+\nu k}^{n+\nu k+k} \left(\prod_{\mu} A_\nu \right) K(\mu)\, e(\mu)$$

As all $A^{\times}_{\nu}$ are positive definite the sequence (11.4) converges to zero according to the result just obtained.