IFAC

INTERNATIONAL FEDERATION
OF AUTOMATIC CONTROL

WARSZAWA 1969

Identification
Methods for Parameters Estimation

Fourth Congress of the International
Federation of Automatic Control
Warszawa 16–21 June 1969

TECHNICAL
SESSION
26

NOT

Organized by
Naczelna Organizacja Techniczna w Polsce

INTERNATIONAL FEDERATION OF AUTOMATIC CONTROL

Identification

Methods for Parameters Estimation

TECHNICAL SESSION No 26

FOURTH CONGRESS OF THE INTERNATIONAL
FEDERATION OF AUTOMATIC CONTROL
WARSZAWA 16 -- 21 JUNE 1969

Organized by
Naczelna Organizacja Techniczna w Polsce

Contents

Wydawnictwa Czasopism Technicznych NOT
Warszawa, ul. Czackiego 3/5 — Polska

Zakład Poligraficzny WCT NOT. Zam. 15/69.

ON-LINE ESTIMATION OF DYNAMIC MODEL PARAMETERS FROM INPUT-OUTPUT DATA

V. Peterka and K. Šmuk
Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
Prague, Czechoslovakia

1. Introduction

The paper deals with the identification of a linear dynamic system by the input and output signal. The output signal of most industrial plants is not determined only by the input signal and the initial state of the system but it is also influenced by unmeasurable disturbances of a stochastic nature. The mode of description of such systems most frequently used at present time (though not the only one posible) is represented in Fig. 1. Block S represents here the ideal deterministic system with an ideal (not identifiable by measurement) output signal v. Upon this ideal output signal v there is superimposed noise ε, and the sum

$$x = v + \varepsilon \qquad (1)$$

represents thus the real and measurable output signal of the system. This mode of description is justified especially in the case of linear systems. Here the law of superimposition holds, and the internal noise of the system can be transformed to the output. Noise ε also includes possible random measuring errors. The problem of the identification of systems with additive noise was studied by many authors from various aspects and for various purposes. Some of them are listed in the enclosed references [1-9]. Other references can be found in survey papers [10-12]. In respect of the stating of the problem closest to this paper are the important work of Åström and Bohlin [1-3] and the paper by Clarke [7]. Differently from these papers, our approach permits the simplification of the computation algorithm, removes iterations and the connected problems of convergence. Neither have we problems with local extremes in minimisation. Moreover, our algorithm provides for the successive reduction of measured data and preserving at the same time all the necessary information. Data reduction is done simultaneously with measurement, there is no necessity of logging, the memorizing of the whole process is not required either, and so the application of the on-line method is made easier. Requirements on storage capacity

are comparatively small and independent of the length of the observation interval. We also use slightly less demanding assumption concerning the statistic characteristics of noise. In difference from [1] the identification of the system is separated from that of noise. This paper deals only with the identification of the system proper.

2. Statement of the Problem

A single-parameter, time-independent system will be considered with the aim of determining its model suited for the purposes of digital or impulse control. It is thus assumed that input signal y is discrete, and that output signal x is sampled within the same sampling period. It is well known that the large class of systems of this type can be described by the difference equation

$$v(t) + \sum_{i=1}^{n} a_i v(t-i) = \sum_{i=0}^{n} b_i y(t-d-i) \qquad (2)$$

where t is the discrete time, n the order of the system, d stands for the possible transport lag, and coefficients a_i $(i = 1,2,\ldots n)$ and b_i $(i = 0,1,2,\ldots n)$ are parameters to be determined by experiment.

Equation (2) contains the ideal output signal v which is not measurable. Available is only output signal x which includes additive noise ε. By substituting (1) into (2) it follows that

$$x(t) + \sum_{i=1}^{n} a_i x(t-i) - \sum_{i=0}^{n} b_i y(t-d-i) = \varepsilon(t) + \sum_{i=1}^{n} a_i \varepsilon(t-i) \qquad (3)$$

The task is now to estimate coefficients a_i $(i = 1,2,\ldots n)$ and b_i $(i = 0,1,2,\ldots n)$ with a given final sequence of the values of input signal $\{y(t)\}$ and the corresponding final sequence of the measured values of output signal $\{x(t)\}$. For the solution of the problem it is necessary to make some assumptions concerning the statistical properties of noise ε which has to be eliminated. In most practical cases (e.g. chemical systems or thermo-technological plants) only very little is known in advance about the statistical properties of the internal noise of the system. Therefore, we shall confine ourselves to the most simple assumptions.

If measurement is made in an open loop system, it can be assumed that noise is statistically independent of the input signal, and thus it holds that

$$E\left[\varepsilon(t)/y(t),y(t-1),y(t-2),\ldots\right] = E\varepsilon(t) \qquad (4)$$

for all values of t within the interval of observation.

In measurements of real industrial plants cases are frequently encountered where a very low frequency drift is superimposed on the output signal. Disregarding the presence of drift could lead to grave errors . We shall therefore assume that the expected mean value of noise depends on time, and that within the observation interval this dependence can be expressed by the polynomial

$$E\varepsilon(t) = \sum_{i=0}^{\nu} c_i t^i \qquad (5)$$

where c_i (i = 0,1,...) are unknown coefficients. In practical cases a polynomial grade of ν = 1 or 2 will normally satisfy.

For finding the asymptotic properties of estimations we shall further need the assumption that the random process $\varepsilon(t) - E\varepsilon(t)$ is ergodic and weakly stationary. No other a priori knowledge of statistic characteristics is required.

As far as input signal y(t) is concerned, it is assumed that the system is "sufficiently excited" by it. This will be specified in more detail in section 4.

Before embarking on the solution of the stated problem let us derive the algorithm for the successive regression analysis with growing data. The algorithm has a general significance and can also be used for other purposes.

3. Algorithm for the successive regression analysis with growing data and limited memory.

Let us consider the classical case of a least square linear regression, and the system of equations

$$\sum_{j=1}^{N} r_j y_{ij} = x_i + e_i \qquad (i = 1,2,\ldots L) \qquad (6)$$

where $L > N$, x_i and y_{ij} are values obtained by observation, and e is an unknown random error. The latter can be interpreted as the deviation from the conditional expectation, i.e.

$$e_i = E\left[x/y_{i1}, y_{i2}, \ldots y_{iN}\right] - x_i \qquad (7)$$

Now the task is to find the estimates $\hat{r}_j$ of regression coefficients r_j

minimising the sum of the squares of errors

$$Q = \sum_{i=1}^{L} e_i^2 \tag{8}$$

Relations (6) and (8) can be written in matrix form as follows

$$Y_{[L\times N]}\, r_{[N\times 1]} - x_{[L\times 1]} = e_{[L\times 1]} \tag{9}$$

$$Q = e^T_{[1\times L]}\, e_{[L\times 1]} \tag{10}$$

The subscript in the brackets indicates the dimensions of the respective matrix, whereas superscript T denotes transposition.

The classical solution of the given problem is (see e.g. [13])

$$\hat{r}_L = (Y^T Y)^{-1} Y^T x \tag{11}$$

provided that

$$\det[Y^T Y] \neq 0 \tag{12}$$

The numerical expression of formula (11) by ordinary matrix calculus (i.e. multiplication and inversion of matrices) is cumbersome when the number L of observations is large (e.g. hundreds or thousands); it requires a large capacity of memory and is numerically less stable [14]. Further on we shall derive an algorithm permitting the solution of the problem with substantially reduced demands on the memory of the computer used. This algorithm is based on orthogonal transformations of the system of linear equations [14,15,16] and is numerically very stable.

Let us arrange the system of equations (9) into the form

$$Z_{[L\times(N+1)]}\, \tilde{r}_{[(N+1)\times 1]} = e_{[L\times 1]} \tag{13}$$

where

$$Z = [Y, -x], \qquad \tilde{r} = \begin{bmatrix} r \\ 1 \end{bmatrix} \tag{14}$$

and multiply from the left by square matrix T

By denoting

$$T_{[L\times L]}\, Z_{[L\times(N+1)]} = \tilde{Z}_{[L\times(N+1)]} \tag{15}$$

$$T_{[L\times L]}\, e_{[L\times 1]} = \tilde{e}_{[L\times 1]} \tag{16}$$

after this transformation we have (instead of (13))

$$\tilde{Z}_{[L\times(N+1)]}\,\tilde{r}_{[(N+1)\times 1]}=\tilde{e}_{[L\times 1]} \tag{17}$$

The sum of squares of the right sides of the transformed system of equations (17) will then be

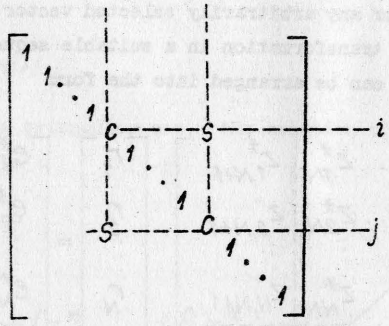
$$\tilde{Q}=\tilde{e}^T\tilde{e}=e^TT^TTe \tag{18}$$

It is obvious that the sum of squares has not been changed by this transformation, i.e. $\tilde{Q}=Q$, provided that matrix T is orthogonal

$$T^TT=I \tag{19}$$

Let us further consider a special type of this orthogonal transformation where matrix T is a so called elementary matrix of rotation [15]. By representing only the non-zero elements this matrix can be written as follows:

$$T_{ij[L\times L]}= \tag{20}$$



It will be easy to discover that matrix (20) will be orthogonal, and the sum of squares Q will not change, provided that condition is fulfilled that

$$c^2+s^2=1 \tag{21}$$

It is obvious that the transformation of the system of equations (13) by multiplication by matrix (20) applies only to the i-th and j-th equations, in other words, only the i-th and j-th rows will change in matrices Z and e.

For $k\neq i \wedge k\neq j$

$$\tilde{z}_{k\nu}=z_{k\nu}\ (\nu=1,2,\ldots N+1)\quad,\quad \tilde{e}_k=e_k$$

for k = i

$$\tilde{z}_{i\nu} = cz_{i\nu} + sz_{j\nu} \quad , \quad \tilde{e}_i = ce_i + se_j \tag{22}$$

for k = j

$$\tilde{z}_{j\nu} = -sz_{i\nu} + cz_{j\nu} \quad , \quad \tilde{e} = -se_i + ce_j \tag{23}$$

Coefficients c and s are bound by the condition of orthogonality (21), however, one of them can be selected. Let the selection make so that it holds that

$$\tilde{z}_{j\mu} = 0$$

i.e. according to (23) we obtain

$$-sz_{i\mu} + cz_{j\mu} = 0 \tag{24}$$

From (24) and (21) it follows that

$$c = \frac{z_{i\mu}}{\sqrt{z_{i\mu}^2 + z_{j\mu}^2}} \quad , \qquad s = \frac{z_{j\mu}}{\sqrt{z_{i\mu}^2 + z_{j\mu}^2}} \tag{25}$$

In this way we can annul any element in matrix Z without changing the sum of squares Q for any arbitrarily selected vector r. By the successive application of this transformation in a suitable sequence the original system of equations (13) can be arranged into the form

$$L \left[\begin{array}{cccc|c} z_{11}^* & z_{12}^* & \cdots & z_{1N}^* & z_{1,N+1}^* \\ & z_{22}^* & \cdots & z_{2N}^* & z_{2,N+1}^* \\ & & & \vdots & \vdots \\ & & & z_{NN}^* & z_{N,N+1}^* \\ \hline & & & & z_{N+1,N+1} \end{array}\right] \left[\begin{array}{c} r_1 \\ r_2 \\ \vdots \\ r_N \\ 1 \end{array}\right] = \left[\begin{array}{c} e_1^* \\ e_2^* \\ \vdots \\ e_N^* \\ e_{N+1}^* \\ 0 \\ 0 \\ \vdots \\ 0 \end{array}\right] \tag{26}$$

while it still holds that

$$Q = \sum_{i=1}^{L} e_i^2 = \sum_{i=1}^{N+1} e_i^{*2} \tag{27}$$

for any arbitrarily selected r_j $(j = 1,2,\ldots N)$.

After this transformation the determination of estimates $\hat{r}$ (j = 1, 2,...N) minimising (27) becomes a simple matter. From equation (26) it is obvious that by selection of r (j = 1,2,...N) it is impossible to influence the last non-zero element of the right side $e^*_{N+1} = z^*_{N+1,N+1}$, however, it is possible to annul all the other elements e^*_j (j = 1,2,...N). The estimates of regression coefficients $\hat{r}$ (j = 1,2,...N) can thus be found by the solution of the system of linear equations

$$Z^*_{[N\times N]}\,\hat{r}_{[N\times 1]} \quad z^*_{[N\times 1]} = 0 \tag{28}$$

where

$$Z^*_{[N\times N]} = \begin{bmatrix} z^*_{11} & z^*_{12} & \cdots & z^*_{1N} \\ & z^*_{22} & \cdots & z^*_{2N} \\ 0 & & & \vdots \\ & & & z^*_{NN} \end{bmatrix} \qquad z^*_{[N\times 1]} = \begin{bmatrix} z^*_{1,N+1} \\ z^*_{2,N+1} \\ \vdots \\ z^*_{N,N+1} \end{bmatrix} \tag{29,30}$$

At the same time we obtain the minimum of the sum of squares

$$Q_{min} = z^{*2}_{N+1,N+1} \tag{31}$$

Since matrix Z^* (29) is a triangular one, the solution of the system of equations (29) is very simple :

$$\begin{aligned} \hat{r}_N &= -\frac{z^*_{N,N+1}}{z^*_{N,N}} \\ \hat{r}_{N-1} &= -\frac{1}{z^*_{N-1,N-1}}\left(z^*_{N-1,N+1} + \hat{r}_N z^*_{N-1,N}\right) \\ &\vdots \\ \hat{r}_{N-k} &= -\frac{1}{z^*_{N-k,N-k}}\left(z^*_{N-k,N+1} - \sum_{i=0}^{k-1} \hat{r}_{N-i}\, z^*_{N-k,N-i}\right) \end{aligned} \tag{32}$$

Let us consider now the situation where the data from L observations have been reduced into a triangular matrix with elements z^*_{ij} (i = 1, 2,...N+1; j = i, i+1,...N+1), and new data are obtained forming a futher L+1 row in the matrix on the left side of equation (26). All elements in this new row can be annuled by the gradual application of the described transformation, and this simultaneously means the correction of the upper triangular matrix. This step is more accurately described by the following procedure in ALGOL-60:

```
procedure REDUCE (matrix) new data:(row) order:(N);
  value N, row; array matrix, row; integer N;
  begin real c, s, de; integer i, k;
    for i:= 1 step 1 until N+1 do
     if row[i] ≠ 0 then
     begin de:= sqrt(row[i]↑ 2 + matrix[i,i]↑ 2);
           c:= matrix[i,i]/de; s:= row[i]/de;
      for k:= i step 1 until N+1 do
       begin de:= c x row[k] - s x matrix[i,k];
         matrix[i,k]:= s x row[k] + c x matrix[i,k];
         row[k]:= de
       end
     end
  end REDUCE;
```

The successive application of this procedure permits the processing of growing data without the necessity to memorize them. All necessary information on observed past history accumulates in the triangular matrix.

Let us now consider how this unified algorithm could also be used for processing the first N+1 rows of data in matrix Z (14). It is obvious that the sum of squares (10) will not change, if the matrix in (13) is extended by a zero matrix of dimensions (N+1)x(N+1)

$$\begin{bmatrix} 0 \\ Z \end{bmatrix}_{[(L+N+1)\times(N+1)]} \cdot \tilde{r}_{[(N+1)\times 1]} = \begin{bmatrix} 0 \\ e \end{bmatrix}_{[(L+N+1)\times 1]}$$

This zero matrix (more exactly, its upper triangular portion) can thus be considered as the initial state of matrix Z^*, when it does not contain yet any information on the process. This approach permits the use of the unified algorithm for all the processed data (i= 1,2,...L).

Triangular matrix (29) is non-singular, if reduction was applied to N linearly indipendent data rows at least. Beginning from this instant equations (32) can be used for computing in any arbitrary step estimates $\hat{r}$ (j = 1,2,...N) which are optimal in the sense of least squares for the whole past history.

4. The solution of the identification problem

Let us now revert to the principal problem stated in section 2. Without abstaining from generalization it is possible to consider only the

case of no transport lag, $d = 0$. Any possible lag can be considered by a simple shifting of the values of input signal by d sampling periods. Let us introduce a new random variable

$$\delta(t) = \varepsilon(t) + \sum_{i=1}^{n} a_i \varepsilon(t-i) \tag{33}$$

appearing on the right hand side of equation (3)

$$x(t) + \sum_{i=1}^{n} a_i x(t-i) - \sum_{i=0}^{n} b_i y(t-i) = \delta(t) \tag{34}$$

What do we know about the statistical properties of this random variable $\delta(t)$? No more and no less than about the statistical properties of the noise $\varepsilon(t)$:

If the noise expectation (drift) can be expressed by polynomial (5), for variable $\delta(t)$ it holds that

$$E\delta(t) = \sum_{i=0}^{\nu} d_i t^i \tag{35}$$

Instead of unknown coefficients c_i ($i = 0,1,2,\ldots,\nu$) we thus have the same number of unknown coefficients d_i ($i = 0,1,2,\ldots,\nu$). The relationship between these coefficients is

$$d_i = \sum_{j=i}^{\nu} \binom{j}{i} c_j \sum_{k=0}^{n} a_k (-k)^{j-i} \tag{36}$$

where $\binom{j}{i}$ are binomial coefficients. Equation (36) holds good for all values of i if we put $\binom{0}{0} = 1$; $0^0 = 1$. For instance, for $\nu = 2$

$$\begin{aligned} d_0 &= c_0 \sum_k a_k - c_1 \sum_k k a_k \sum_k k^2 a_k \\ d_1 &= c_1 \sum_k a_k - 2c_2 \sum_k k a_k \\ d_2 &= c_2 \sum_k a_k \end{aligned} \tag{37}$$

As originally assumed, noise $\varepsilon(t)$ is independent of input signal $y(t)$; the same property is also ascribed to random variable $\delta(t)$ introduced by relation (33). This means that in the linear regression model

$$E\left[\delta(t) \middle| y(t), y(t-1), \ldots, y(t-N+1)\right] = E\delta(t) + \sum_{i=0}^{N-1} r_i y(t-i) \tag{38}$$

all regression coefficients equal zero

$$r_i = 0 \quad , \quad i = 0,1,2,\ldots N-1 \tag{39}$$

Apart from the form of the relationship (35) for noise expectation (drift), this is the only statistical characteristic known in advance which follows from the assumptions made in section 2.

4.1 The principle of the method

Our approach will be based on the a priori knowledge of regression coefficients r_i in the regression model (38). In our case these coefficients equal zero, since noise $\varepsilon(t)$, and thus also random variable $\sigma(t)$, are independent of input signal $y(t)$.*)

Let us divide the solution of the given problem into two steps:

a) In the first step let us find the unbiased estimate $\hat{r}$ of the vector of the regression coefficients. This estimate will be the function of unknown parameters a_i $(i= 1,2,\ldots n)$, b_i $(i= 1,2,\ldots n)$ and of the discrete values of the input and output signals within the considered interval of observation.

$$\hat{r} = f(a, b, x, y) \tag{40}$$

b) The assumption seems to be justified that the experiment will produce the most likely results, i.e. that estimate $\hat{r}$ will be "as close as possible" to the real value of $r = 0$. Were the number of estimated regression coefficients just equal to the number of unknown parameters a_i, b_i, i.e. if $N = 2n + 1$, the condition of $\hat{r} = r = 0$ could be fulfilled without errors. In this way we would obtain $2n + 1$ equations for the estimates of $\hat{a}_i$ and $\hat{b}_i$. Of primary interest will be the case of $N > 2n + 1$. In such a case, however, it is not generally possible to fulfil the equality condition between $\hat{r}$ and the real value $r = 0$ known in advance by the selection of a_i and b_i. We can comply only with the condition of the least possible distance between both vectors. Let us measure this distance by the sum of squares, i.e.

$$Q_r = \sum_{i=0}^{N-1} \hat{r}_i^2 = \hat{r}^T \hat{r} \tag{41}$$

*) However, the same procedure could also be used in the case of non-zero coefficients, provided they were known in advance (say, as the functions of unknown parameters a_i and b_i). This situation would occur, e.g. if the noise were correlated with the input signal and the cross-correlation function were known.

Estimates $\hat{a}$ and $\hat{b}$ are thus determined by the minimisation of (41).

The question to be logically posed now is: How to select N representing the number of the estimated regression coefficients? When the order n of the system is known in advance, the question is not a crutial one, and satisfactory results can be obtained by $N = 2n + 1$ or slightly larger. In practical situation the order n of the dynamic model (2) is normally not known in advance and the best suited approximation must be found. In such case N must be selected sufficiently large, so that regression model (38) should contain, as far as possible, all past values of input y(t-i) which are capable of markedly influencing the output of the system within time t.

The derived algorithm also permits the simultaneous investigation of all cases of $n \leq n_{max}$ where n_{max} is the maximum order of model (2) intended to be considered. As shown by results described in section 6 the values of criterion (41) can be well used in the selection of a suitable approximation.

4.2 The detailed solution

Re a) First let us find the estimates of regression coefficients r_i (i= 0,1,2,...N-1). Let us substitute

$$E\left[\delta(t)/y(t),\ y(t-1),\ \ldots,\ y(t-N+1)\right]$$

into (38) and at the same time also substitute (35)

$$\sum_{i=0}^{\nu} d_i t^i + \sum_{i=0}^{N-1} r_i\, y(t-i) = \delta(t) + e(t)$$

The unbiased estimate of regression coefficients r_i and d_i can be obtained by the minimisation of the sum of squares

$$Q_e = \sum_{t=1}^{L} e^2(t) \qquad (42)$$

where L represents the interval of observation. However, the realisation of random variable $\delta(t)$ is not available. Therefor, let us substitute for $\delta(t)$ the left hand side of equation (34).

$$\sum_{i=0}^{\nu} d_i t^2 - \sum_{i=0}^{N-1} r_i\, y(t-i) - x(t) - \sum_{i=1}^{n} a_i\, x(t-i) + \sum_{i=0}^{n} b_i y(t-i) = e(t) \qquad (43)$$

Consider L to be the number of equations of type (43) which can be compiled from experimental data, and let us assume that $L \geq N + \nu + 1$. The

system of equations (43) for t= 1,2,...L can be written in the matrix form of

$$Z_{[L\times(N+\nu+n+2)]}\,\vartheta_{[(N+\nu+n+2)\times 1]} = e_{[L\times 1]} \tag{44}$$

where

$$Z_{[L\times(N+\nu+n+2)]} = \left[\begin{array}{ccc|ccc|ccc} 1 & 1\dots 1 & & y(1) & y(0)\dots & y(-N+2) & x(1) & x(0)\dots & x(-n+1) \\ 1 & 2\dots 2^{\nu} & & y(2) & y(1)\dots & y(-N+3) & x(2) & x(1)\dots & x(-n+2) \\ 1 & 3\dots 3^{\nu} & & y(3) & y(2)\dots & y(-N+4) & x(3) & x(2)\dots & x(-n+3) \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & L\dots L^{\nu} & & y(L) & y(L-1)\dots & y(L-N+1) & x(L) & x(L-1)\dots & x(L-N) \end{array}\right] \tag{45}$$

is the matrix containing all experimental data and vector ϑ contains the unknown coefficients

$$\vartheta_{[(N+\nu+n+2)\times 1]} = col\left[d_0\, d_1 \dots d_{\nu} \mid r_0+b_0, \dots, r_n+b_n, r_{n+1}, \dots, r_{N-1} \mid -1 \mid -a_1 \dots -a_n\right] \tag{46}$$

By applying the procedure described in section 3 these data can be reduced into the triangular matrix

$$Z^{*}_{[(N+n+\nu+2)\times(N+n+\nu+2)]} = \left[\begin{array}{ccc|ccc|ccc} p^{*}_{0,0} & p^{*}_{0,1}\dots p^{*}_{0,\nu} & & y^{*}_{0,0}\dots & y^{*}_{0,N-1} & & x^{*}_{0,0} & x^{*}_{0,1}\dots & x^{*}_{0,n} \\ & p^{*}_{1,1}\dots p^{*}_{1,\nu} & & y^{*}_{1,0}\dots & y^{*}_{1,N-1} & & x^{*}_{1,0} & x^{*}_{1,1}\dots & x^{*}_{1,n} \\ & \vdots & & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & p^{*}_{\nu,\nu} & & y^{*}_{\nu,0}\dots & y^{*}_{\nu,N-1} & & x^{*}_{\nu,0} & x^{*}_{\nu,1}\dots & x^{*}_{\nu,n} \\ \hline & & & y^{*}_{\nu+1,0}\dots & y^{*}_{\nu+1,N-1} & & x^{*}_{\nu+1,0} & x^{*}_{\nu+1,1}\dots & x^{*}_{\nu+1,n} \\ & 0 & & & \vdots & & \vdots & \vdots & \vdots \\ & & & 0 & y^{*}_{\nu+N,N-1} & & x^{*}_{\nu+N,0} & x^{*}_{\nu+N,1}\dots & x^{*}_{\nu+N,n} \\ \hline & & & & & & x^{*}_{\nu+N+1,0} & x^{*}_{\nu+N+1,1}\dots & x^{*}_{\nu+N+1,n} \\ & & & & & & & x^{*}_{\nu+N+2,1}\dots & x^{*}_{\nu+N+2,n} \\ & 0 & & 0 & & & 0 & & \vdots \\ & & & & & & & & x^{*}_{\nu+N+n+1,n} \\ & & & & & & & 0 & \end{array}\right] \tag{47}$$

As described in section 3 this reduction can be made successively with growing data without the necessity of memorizing past data. Instead of (44) a reduced system of equations is obtained

$$Z^*_{[(N+n+\nu+2)x(N+n+\nu+2)]}\ \vartheta^*_{[(N+n+\nu+2)x1]} = e^*_{[(N+n+\nu+2)x1]} \qquad (48)$$

Let us bear in mind that the sum of squares (42) has not been changed by this reduction, i.e. it holds that

$$Q_e = e^T_{[1xL]}\ e_{[Lx1]} = e^{*T}_{[1x(N+n+\nu+2)]}\ e^*_{[(N+n+\nu+2)x1]} \qquad (49)$$

The first step in the solution of our problem is the estimation of coefficients r_i $(i = 0,1,2,\ldots N-1)$ and d_i $(i = 0,1,..,\nu)$. In other words, these coefficients have to be determined so that the sum of squares (49) is minimum, and assuming for this instant that coefficients a and b are known. The minimisation can be performed in the following way.

Let us divide matrix Z^* and vector ϑ^* as shown in (47) and (46).

$$Z^*_{[(N+n+\nu+2)x(N+n+\nu+2)]} = \begin{bmatrix} P^*_{[(\nu+1)x(\nu+1)]} & Y^*_{d[(\nu+1)xN]} & X^*_{d[(\nu+1)x1]} & X^*_{d[(\nu+1)xn]} \\ 0 & Y^*_{r[NxN]} & X^*_{r[Nx1]} & X^*_{r[Nxn]} \\ 0 & Y & X^*_{e[(n+1)x1]} & X^*_{e[(n+1)xn]} \end{bmatrix} \qquad (50)$$

$$\vartheta^*_{[(N+n+\nu+2)x1]} = \begin{bmatrix} d_{[(\nu+1)x1]} \\ r_{b[Nx1]} \\ -1 \\ -a_{[nx1]} \end{bmatrix} \qquad (51)$$

The definition of newly introduced matrices and vectors can be seen from comparing (50) with (47) and (51) with (46). Let us note that

$$r_{b[Nx1]} = \begin{bmatrix} r_0 + b_0 \\ r_1 + b_1 \\ \vdots \quad \vdots \\ r_n \quad b_n \\ r_{n+1} \\ \vdots \\ r_{N-1} \end{bmatrix} = r_{[Nx1]} + \begin{bmatrix} b_{[(n+1)x1]} \\ 0 \end{bmatrix} \qquad (52)$$

By the substitution of (50) and (51) the system of equations (48) decomposes into three systems of linear equations

$$P^* d + Y_d^* r_b - x_d^* - X_d^* a = e_d^* \tag{53}$$

$$Y_r^* r_b - x_r^* - X_r^* a = e_r^* \tag{54}$$

$$-x_e^* - X_e^* a = e_e^* \tag{55}$$

It is obvious that by the selection of r and d it is impossible to influence e_e^*, however, vectors e_d^* and e_r^* can be completely annuled by this selection. Consequently,

$$Q_{emin} = e_e^{*T} e_e^* \tag{56}$$

is the smallest attainable sum of squares, and estimates $\hat{r}$ and $\hat{d}$ can be found by solving equations (55) and (54) by setting $e_r^* = 0$ and $e_d^* = 0$ in these equations.

Equation (54) will completely suffice for computing the sought for estimate $\hat{r}$. From equation (54) follows that

$$\hat{r}_{[Nx1]} = u_{[Nx1]} + W_{[Nxn]} a_{[nx1]} - \begin{bmatrix} b_{[(n+1)x1]} \\ 0 \end{bmatrix} \tag{57}$$

where vector u and matrix W are defined by relationship

$$\left[u, W\right]_{[Nx(n+1)]} = Y_{r[NxN]}^{*-1} \left[x_r^*, X_r^*\right]_{[Nx(n+1)]} \tag{58}$$

which follows from the solution of equation (54). The fact of Y_r^* being a triangular matrix makes the computation of (58) very easy.

Matrix Y_r^* contains only reduced input data. For the existence of its inversion this matrix must not be a singular one. The input signal complying with this condition was designated in section 2 as "sufficiently exciting". For instance, this condition is not fulfilled by a periodical signal with a period smaller than N.

Formula (57) yields the sought for estimation of the vector of regression coefficients (40). All the necessary information about the input and output of the system is concentrated in vector u and matrix W.

Re b) After finding the estimate of the vector of regression coefficients (57) it is possible to begin the second step of the computation, i.e. the

minimisation of the sum of squares (41). The procedure described in section 3 will be used again. Let us arrange equation (57) into the form

$$\hat{r}_{[Nx1]} = S_{[Nx(2n+2)]} \cdot \begin{bmatrix} -b_{[(n+1)x1]} \\ a_{[nx1]} \\ 1 \end{bmatrix} \tag{59}$$

where

$$S_{[Nx(2n+2)]} = \left[\begin{array}{c|c|c} I_{[(n+1)x(n+1)]} & W_{[Nxn]} & u_{[Nx1]} \\ 0 & & \end{array}\right] \tag{60}$$

After dividing matrix W and vector u into two parts

$$S_{[Nx(2n+2)]} = \left[\begin{array}{c|c|c} I_{[(N+1)x(n+1)]} & W_{b[(n+1)xn]} & u_{b[(n+1)x1]} \\ 0 & W_{a[(N-n-1)xn]} & u_{a[(N-n-1)x1]} \end{array}\right]$$

it can be seen that for reducing matrix S into triangular form it will suffice to reduce into the upper triangular form only matrix W_a together with vector u_a. After this reduction we obtain instead of (59)

$$\hat{r}^*_{[(2n+2)x1]} = \begin{bmatrix} I_{[(n+1)x(n+1)]} & W_{b[(n+1)xn]} & u_{b[(n+1)x1]} \\ 0 & W^*_{a[nxn]} & u^*_{a[nx1]} \\ 0 & 0 & u^*_r \end{bmatrix} \begin{bmatrix} -b_{[(n+1)x1]} \\ a_{[nx1]} \\ 1 \end{bmatrix} \tag{61}$$

The last row (2n+2) contains only the last element u^*_r, and for any arbitrary a and b the sum of squares (41)

$$Q_r = \sum_{i=0}^{N-1} \hat{r}_i^2 = \sum_{i=0}^{2n+1} \hat{r}_i^{*2} \tag{62}$$

remains unchanged. The smallest value of Q_r attainable by the selection of a and b is thus

$$Q_{r\,min} = u_r^{*2} \tag{63}$$

Estimates $\hat{a}$ are obtained from equation

$$W_a^* \hat{a} + u_a^* = 0 \tag{64}$$

the solution of which is very simple because W_a^* is a triangular matrix. For estimate $\hat{b}$ it follows from (61) that

$$\hat{b} = u_b + W_b \hat{a} \tag{65}$$

The solution of our problem is thus completed. The asymptotic properties of the estimates will be given in section 5.

As far as the drift, eliminated in the course of computation, is concerned, it can be determined in the following way. From equations (53) and (54), where we put $e_r^* = e_d^* = 0$, let us compute estimate $\hat{d}$, and, using relations (36) or (37), the estimates of coefficients $c_i (i = 0,1,\ldots,p)$ of polynomial (5).

Owing to the limited scope of this article all numerical details cannnot be discussed here. Let us bear in mind, however, that the described algorithm is arranged so that a small number of additional numerical operations permits the simultaneous computation of all variants for the lower orders of dynamic model (2).

5. The asymptotic properties of the estimate

The representation of the low-frequency drift by polynomial (5) makes sense only for a finite interval of observation. For this reason the asymptotic properties of the estimate were studied only for the case of $E\varepsilon(t) = c_0$ where c_0 is an unknown constant. The following theorem holds:

Theorem 1 When the following assumptions are fulfilled:

[A] Input signal $y(t)$ is persistently and sufficiently exciting, i.e. the following limits exist with probability one

$$\lim_{L\to\infty} \frac{1}{L} \sum_{t=1}^{L} y(t), \qquad \lim_{L\to\infty} \frac{1}{L} \sum_{t=1}^{L} y(t)\, y(t+\tau)$$

and inverse matrix Y_r^{*-1} in (58) also exists,

[B] noise $\varepsilon(t)$ is weakly stationary and ergodic,

[C] the dynamic system can be described by difference equation (2) whose order n is finite and known,

[D] the characteristic polynomials of the left and right sides of equation have no common root,

then estimates $\hat{a}$ and $\hat{b}$ computed from equations (64) and (65) are strongly consistent, i.e. it holds that

$$P\left\{\lim_{L\to\infty} \hat{a} = a;\quad \lim_{L\to\infty} \hat{b} = b\right\} = 1$$

If assumption [D] were not fulfilled, it would be possible to determine only a reduced modal not containing the common root factor. However, for the purposes of automatic control this type of model is entirely satisfactory.

The proof of Theorem 1 is given in [17] for $E\varepsilon(t) = 0$. The validity of the theorem for the case of $E\varepsilon(t) = c_0 \neq 0$ can be proved in a very similar way.

6. Experimental results

The algorithm described in sections 3 and 4 was written in ALGOL-60 and the method tested on a ELLIOTT 4100 computer.*) Input data of the programme: 1) the maximum order n_{max} of the dynamic model considered, 2) order of polynomial (5) respecting the drift of the output, 3) the number N of regression coefficients considered, 4) the sequence of the input-output pairs $y(t), x(t)$ of the investigated system which are gradually read from a punched tape (or can be directly obtained from the system by means of an analog-to-digital converter). The computer prints the estimates $\hat{a}$ and $\hat{b}$ of the coefficients of the dynamic model, the estimates $\hat{c}$ of the coefficients of the polynomial drift, and the respective minimal values Q_e (56) and Q_r (57) of the sum of squares. These results are simultaneously obtained for all orders $n \leq n_{max}$ of the model as shown in the computer print in Fig.2. The comparison of the Q values permits the estimation of the order of the dynamic model, or the selection of the suitable approximation. On request the results are printed by the computer after the processing of each 50 pairs of input and output data of the system. This permits the follow-up of the gradually increasing accuracy of estimates with the growing length of observation. The modified version of the programme was also compiled for the frequently occuring case when it is known in advance that $b_0 = 0$.

For testing the method the punched tape containing the input-output pairs of the system was generated by a special separate programme simulating the real system with noise and drift inclusive. Random and not-random input signals were applied. Pseudorandom binary signal proved to be the best suited. The Gaussian white noise was approximately obtained as the sum of 12 random numbers with a uniform distribution. Correlated noise with a selected auto-correlation function was obtained from uncorrelated noise by means of a discrete filter $F(z)$ of the type of rational fraction function.

Figs. 3a and 3b show the identification curves of a second order system. Samples of input, ideal output and applied noise are shown in Fig.4. The sum of ideal output and noise were used for identification. In the case shown in Fig. 3a uncorrelated noise with linear drift was used the sample of which is designated (a) in Fig. 4. The ratio of the effective value of this noise (without drift) to the effective value of the ideal output is $\sigma_\varepsilon / \sigma_y = 1$.

---

*) The complete program for practical applications is available at the Institute of Information Theory and Automation of Czechoslovak Academy of Sciences, Vysehradska 49, Prague 2.

In the case shown in Fig. 5b autocorrelated noise was used obtained from uncorrelated noise by means of filter $F(z) = z/(z - 0.6)$. In this case the ratio $\sigma_\varepsilon / \sigma_v = 0.5$.

Fig.5 shows the comparison of the impulse responses of the original and of the models obtained as the results of identification in case (a) (Fig.3a) for $L = 4400$, and in case (b) (Fig.3b) for $L = 1000$.

Fig. 6a shows the identification curve of a third order system. In this case the noise was uncorrelated and the ratio $\sigma_\varepsilon / \sigma_v = 0.32$. The drift was also linear. Fig. 6b shows the identification of the same third order system, however, the model was considered to be only of second order. The results shown in Figs. 6a and 6b were simultaneously obtained by a single computation. The impuls response of the third order original are compared in Fig. 7 with the impulse responses of the third and second order models obtained.

References:

[1] Åström K.J., Bohlin T.: Numerical Identification of Linear Dynamic Systems from Normal Operating Records. Proc. of IFAC Symposium on Self-Adaptive Control Systems, Teddington Sept. 1966

[2] Åström K. J., Bohlin T., Wensmark S.: Automatic Construction of Linear Stochastic Dynamic Models for Stationary Processes with Random Disturbance Using Operating Records, IBM Nordic Laboratory, Sweden 1965 (Report TP 18.150)

[3] Åström K.J.: On the Achievable Accuracy in Identification Problems, IFAC Symposium Identification in Automatic Control Systems, Prague, June 1967

[4] Balakrishnan A.V.: Determination of Nonlinear Systems from Input-Output Data, Proc. Princeton Conf. Identification Problems in Communication and Control, 1963, Academic, New York

[5] Balakrishnan A.V.: Identification of Control Systems from Input-Output Data, IFAC Symposium Identification in Automatic Control Systems, Prague, June 1967, Academia, Prague

[6] Briggs P.A.N., Godfrey K.R., Hammond P.H.: Estimation of Process Dynamic Characteristics by Correlation Methods using Pseudo Random Signals, IFAC Symposium Identification in Automatic Control Systems, Prague, June 1967

[7] Clarke D.W.: Generalized-Least-Squares Estimation of the Parameters of a Dynamic Model, IFAC Symposium Identification in Automatic Control Systems, Prague, June 1967

[8] Rajbman N.S.: Statistical Methods Model Construction for Automatic Control Systems (in Russian), Trans. Fourth Prague Conference on Information Theory, Statistical Decision Function, Random Processes, Prague, 1965

[9] Davies W.D.T., Douce J.L.: On-Line System Identification in the Presence of Drift, IFAC Symposium Identification in Automatic Control Systems, Prague, June 1967

[10] Cuenod M., Sage A.P.: Comparison of Some Methods Used for Process Identification, IFAC Symposium Identification in Automatic Control Systems, Prague, June 1967

[11] Eykhoff P., Van der Grinten P.M.E.M., Kwakernaak H., Veltman B.P.Th.: Systems Modeling and Identification, Third Congress IFAC, London, June 1966

[12] Eykhoff P.: Process Parameter and State Estimation, IFAC Symposium Identification in Automatic Control Systems, Prague, June 1967

[13] Kendal M.G., Stuart A.: The Advanced Theory of Statistics, Vol. II, Griffin, London 1961

[14] Golub G.: Numerical Methods for Solving Problems, Numerische Mathematik 7, 1965, No. 3

[15] Faddeev D.K., Faddeeva V. N.: Numerical Methods of Linear Algebra (in Russian), Moscow 1960

[16] Faddeev D.K., Kublanowskaya V.N., Faddeeva V.N.: On the Solution of Linear Algebraic Systems with Rectangular Matrices (in Russian), to be published in Trudy LOMI, 1968, Leningrad

[17] Peterka V.: New Approach to the Identification of Discrete Dynamic Systems (in Czech), to be published in Kybernetika 4 (1968), No.6

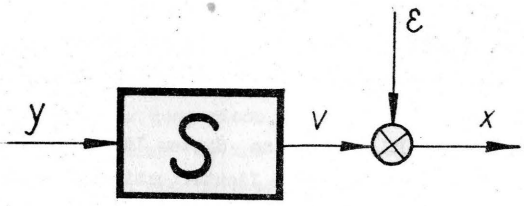


FIG. 1.

LENGTH OF OBSERVATION L = 500

ESTIMATES:

ORDER N = 4 Qr = 2.050Δ-06 Qe = 5.842Δ+00

NUMER:	-0.000492	0.187857	1.961858	3.304158	1.521789
DENOM:	1.000000	-0.722175	-0.839585	0.883417	-0.253558
DRIFT:	25.001206	0.249985			

ORDER N = 3 Qr = 4.300Δ-06 Qe = 3.205Δ+02

NUMER:	-0.000510	0.188462	1.761883	1.432370
DENOM:	1.000000	-1.784605	1.056370	-0.238767
DRIFT:	24.991964	0.250093		

ORDER N = 2 Qr = 2.653Δ+00 Qe = 8.161Δ+02

NUMER:	-0.065485	0.120247	1.692743
DENOM:	1.000000	-1.754764	0.820709
DRIFT:	27.195239	0.216817	

ORDER N = 1 Qr = 1.780Δ+01 Qe = 5.107Δ+01

NUMER:	0.087738	0.280492
DENOM:	1.000000	-0.993374
DRIFT:	-71.183293	0.698681

FIG. 2.

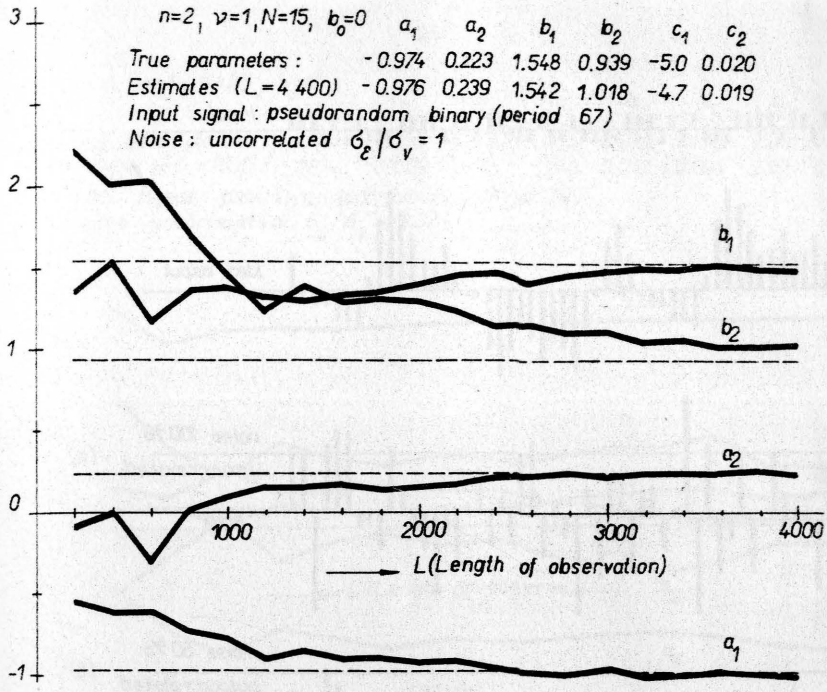
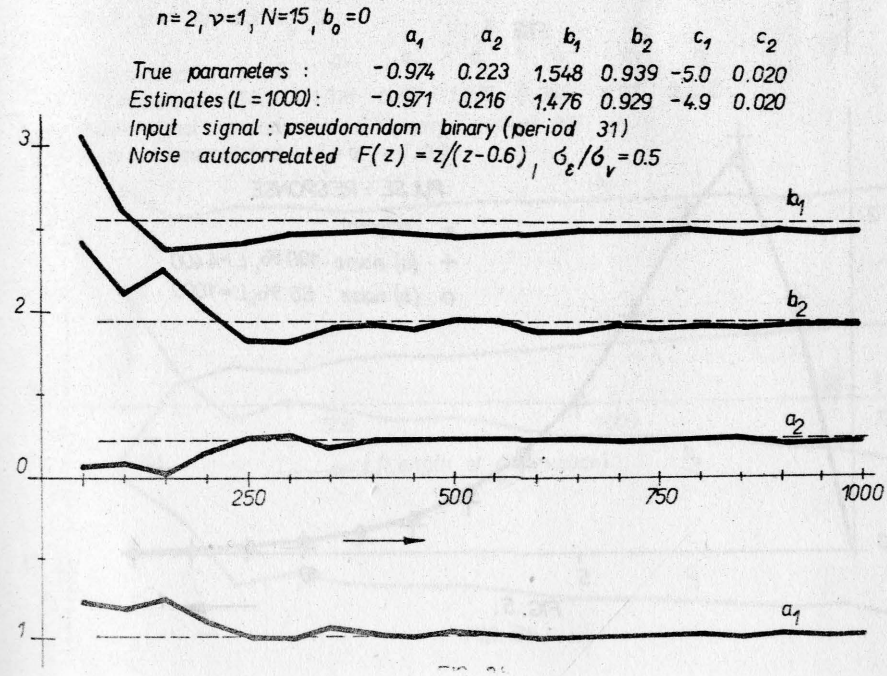

n=2, ν=1, N=15, b0=0
a1 a2 b1 b2 c1 c2
True parameters : -0.974 0.223 1.548 0.939 -5.0 0.020
Estimates (L=4 400) -0.976 0.239 1.542 1.018 -4.7 0.019
Input signal : pseudorandom binary (period 67)
Noise : uncorrelated σε/σν = 1
3
2
1
0
-1
1000
2000
3000
4000
L(Length of observation)
b1
b2
a2
a1

FIG. 3a.


n=2, ν=1, N=15, b0=0
a1 a2 b1 b2 c1 c2
True parameters : -0.974 0.223 1.548 0.939 -5.0 0.020
Estimates (L=1000) : -0.971 0.216 1.476 0.929 -4.9 0.020
Input signal : pseudorandom binary (period 31)
Noise autocorrelated F(z) = z/(z-0.6), σε/σν = 0.5
3
2
1
0
250
500
750
1000
b1
b2
a2
a1

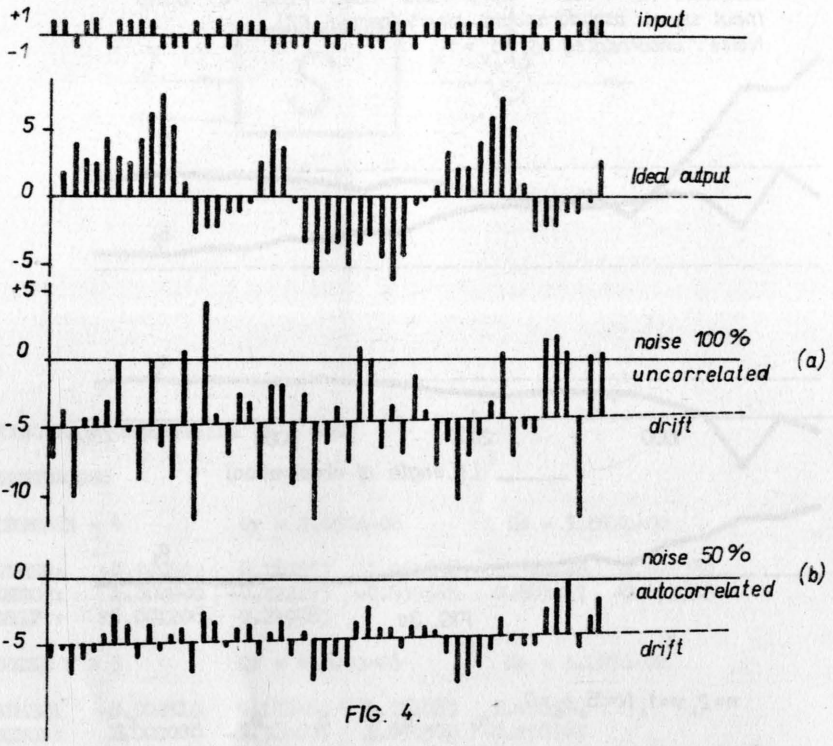

+1
-1
input
5
0
-5
Ideal output
+5
0
-5
-10
noise 100 %
uncorrelated
drift
(a)
0
-5
noise 50 %
autocorrelated
drift
(b)

FIG. 4.

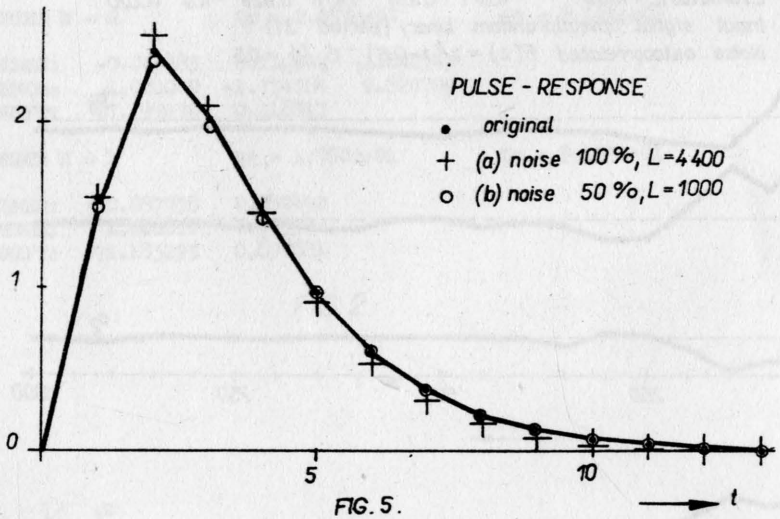

PULSE - RESPONSE
original
(a) noise 100 %, L=4400
(b) noise 50 %, L=1000
2
1
0
5
10
t

FIG. 5.

$n=3,\ \nu=1,\ N=25,\ b_0=0$

	a_1	a_2	a_3	b_1	b_2	b_3	c_0	c_1
True parameters :	-1.110	0.355	-0.030	1.309	-0.092	0.248	2.00	0.050
Estimates (L=1500) :	-0.874	0.009	0.111	1.280	0.213	0.094	2.02	0.050

Input signal : pseudorandom binary (period 67)
Noise uncorrelated $\sigma_\varepsilon/\sigma_v = 0{,}32$.

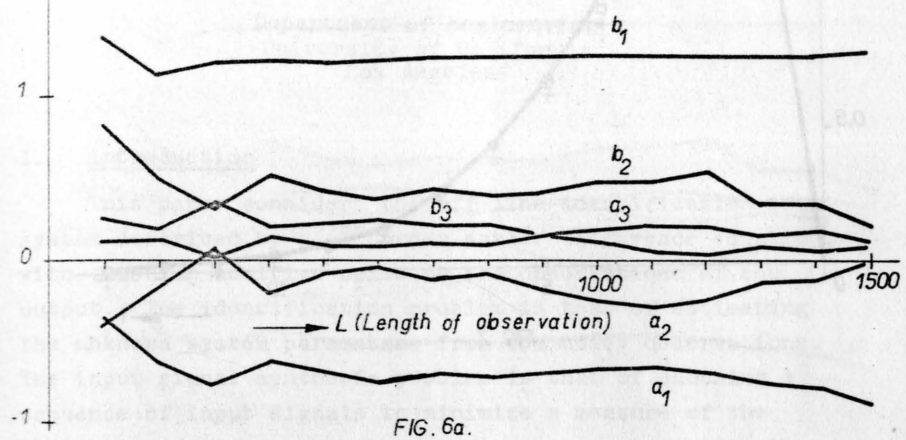


FIG. 6a.

$n=3,\ \nu=1,\ N=25,\ b_0=0$

		a_1	a_2	b_1	b_2	c_0	c_1
Approximation	n = 2						
Estimates (L=1500)		-1.321	0.474	1.276	-0.368	1.97	0.050

Input signal : pseudorandom binary (period 67)
Noise uncorrelated $\sigma_\varepsilon/\sigma_v = 0.32$

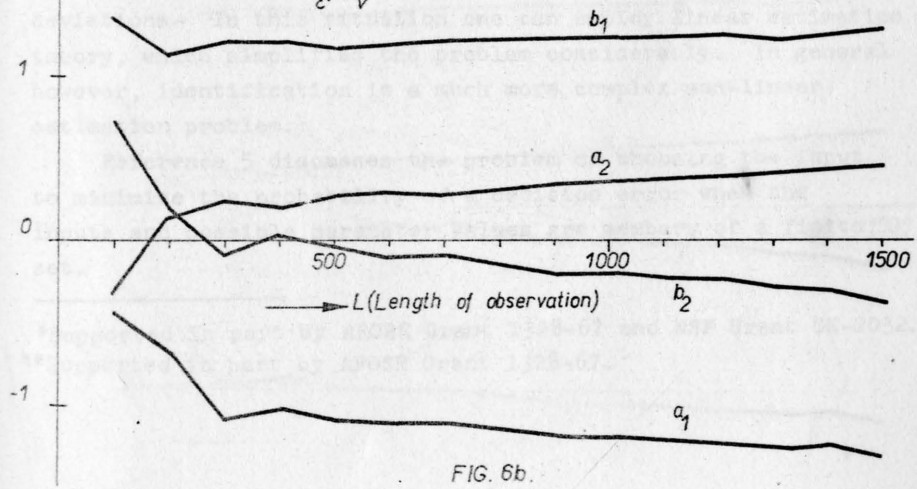


FIG. 6b.

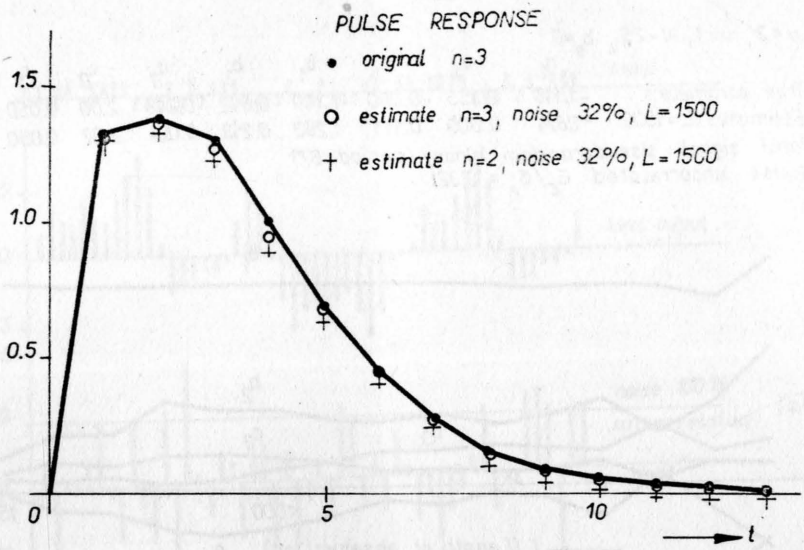

PULSE RESPONSE
original n=3
estimate n=3, noise 32%, L=1500
estimate n=2, noise 32%, L=1500
1.5
1.0
0.5
0
5
10
t

FIG. 7.

ON INPUT SIGNAL SYNTHESIS IN PARAMETER IDENTIFICATION

by

Masanao Aoki* and R. M. Staley**

Department of Engineering
University of California
Los Angeles

1. Introduction

This paper considers the off-line identification of a system described by a k^{th} order scalar difference equation, with Gaussian additive noise on the observations of the output. The identification problem is that of estimating the unknown system parameters from the noisy observations. The input signal synthesis problem is that of choosing a sequence of input signals to minimize a measure of the estimation errors, subject to an energy constraint on the input or output of the system.

Past work in the area of input signal synthesis includes brief discussions of the problem in References 1-3. Reference 4 gives a solution to the problem for the case that the parameter variations are small enough to permit linearization of the output with respect to parameter deviations. In this situation one can employ linear estimation theory, which simplifies the problem considerably. In general however, identification is a much more complex non-linear estimation problem.

Reference 5 discusses the problem of choosing the input to minimize the probability of a decision error when the inputs and possible parameter values are members of a finite set.

---

*Supported in part by AFOSR Grant 1328-67 and NSF Grant GK-2032.
**Supported in part by AFOSR Grant 1328-67.

Reference 6 gives a solution numerically to the one parameter problem in the case that the parameter is treated as a random variable with known statistics. In that paper the sequence of inputs is chosen to minimize the lower bound on estimation error, as given by the trace of the Fisher information matrix which appears in the Cramer-Rao inequality for unbiased estimators[7,8] for the general k-parameter case leading to an open-loop controls. The relevance of Fisher information matrix in the system parameter estimation problem has been noted, for example, in Reference 9 and 10.

In the main body of this paper the unknown parameters are treated as constants rather than random variables. Both the input constrained and output constrained cases are considered. The closed form analytic solutions are presented for the simpler one parameter problem. A useful approximate procedure is developed using the Toeplitz matrix.[12] The type of analysis presented in this paper is a necessary preliminary step to discuss the input signal synthesis problem where the system parameters are random variables with known probability distribution function. Extension to this latter problem including open-loop feedback controls and feedback controls can be made as in Reference 6, and are briefly indicated at the end of this paper.

II. Problem Statement and System Representation

Consider a scalar difference equation

$$x_{i+1} = \sum_{j=1}^{k} a_j x_{i-j+1} + u_i \qquad i=0,1,\ldots \tag{1}$$

where the parameters $a_1,\ldots,a_k$ are assumed to be unknown constants. The system state variable at time i is x_i and the input variable is u_i. Assume for simplicity, that the initial conditions are all zero, i.e.,

$$x_0 = x_{-1} = \ldots = x_{-k+1} = 0$$

It is more convenient for our purpose to re-write (1) as

$$\underline{x}_n = G_n \underline{u}_n \tag{2}$$

where

$$\underline{x}_n^T \triangleq (x_1, x_2, \ldots, x_n), \quad \underline{u}_n^T \triangleq (u_0, u_1, \ldots, u_{n-1})$$

and where the superscript T denotes transposition.

A convenient representation for the matrix G_n in (2) is given in terms of its inverse, i.e.,

$$G_n^{-1} = \begin{bmatrix} 1, & 0 \ldots & & 0 \\ -a_1, & 1, 0, \ldots & & 0 \\ -a_k, & -a_1, & \ldots, 1, 0 \ldots & 0 \\ & & & \\ 0, \ldots, & -a_k, \ldots, & -a_1, & 1 \end{bmatrix} : \text{nxn matrix} \tag{3}$$

namely G_n^{-1} is a lower triangular matrix with non-zero elements $-a_1$, $-a_2$, ... $-a_k$ on the first, second ... and the k^{th} lower co-diagonal lines respectively, the elements on the diagonal line being all ones.

The state variables are observed through noise

$$y_i = x_i + n_i \tag{4}$$

where the noises are assumed to be Gaussian with

$$E(n_i) = 0$$

$$E(n_i n_j) = \sigma^2 \delta_{ij}$$

Define

$$\underline{y}_n^T = (y_1, y_2, \ldots, y_n)$$

and

$$\underline{a}^T = (a_1, \ldots, a_k)$$

Under the above assumption, the Fisher information matrix M_c, defined as

$$M_c = \int p(\underline{y}_n|\underline{a})\, \nabla_{\underline{a}} \ln p(\underline{y}_n|\underline{a})\, [\nabla \underline{a} \ln p(\underline{y}_n|\underline{a})]^T d\, \underline{y}_n \quad (5)$$

where the integration is over the space of all possible observation vector $\underline{y}_n$, and where

$$\nabla_{\underline{a}} \triangleq \begin{pmatrix} \frac{\partial}{\partial a_1} \\ \frac{\partial}{\partial a_2} \\ \vdots \\ \frac{\partial}{\partial a_k} \end{pmatrix}$$

is easily seen to be the (kxk) matrix

$$M_c = \frac{1}{\sigma^2} (\nabla_{\underline{a}}\, \underline{x}_n)^T\, (\nabla_{\underline{a}}\, \underline{x}_n) \quad (6)$$

where

$$\nabla_{\underline{a}} \underline{x}_n = \begin{pmatrix} \frac{\partial x_1}{\partial a_1} & \frac{\partial x_1}{\partial a_k} \\ \vdots & \vdots \\ \frac{\partial x_n}{\partial a_1} & \frac{\partial x_n}{\partial a_k} \end{pmatrix}$$

Given any unbiased conditional estimator $\underline{\gamma}_c(\underline{y}_n)$ of $\underline{a}$, it is known that a lower bound of its covariance matrix is given by the so-called Cramer-Rao inequality[8]*

$$E_{\underline{y}_n|\underline{a}} [(\underline{\gamma}_c(\underline{y}_n) - \underline{a})(\underline{\gamma}_c(\underline{y}_n) - \underline{a})^T] \geq M_c^{-1} \quad (7)$$

where

$$E_{\underline{y}_n|\underline{a}} (\underline{\gamma}_c(\underline{y}_n)) = \underline{a} \quad (8)$$

---

*If B and C are two (LxL) matrices then the notation $B \geq C$ means $\underline{x}^T(B-C)\underline{x} \geq 0$ for all $\underline{x} \in R_L$.

This paper investigates the dependence of M_c on the $\underline{u}_n$ vector. This will be done for the case of known $\underline{a}$, leading to an open-loop control for the system which maximizes a function of the characteristic values of M_c, namely tr M_c.** Thus, the question of constructing unbiased conditional estimators $\underline{\gamma}_c$ such that their error covariance matrices approach M_c^{-1} asymptotically is not discussed.

III. Input Constrained Case

III.1 General Discussions

Define the performance index, J, to be the trace of the Fisher information matrix, which from (2) and (6) is

$$J \triangleq \text{tr}(M_c) = \frac{1}{\sigma^2} \sum_{i=1}^{k} \left(\frac{\partial \underline{x}_n}{\partial a_i}\right)^T \left(\frac{\partial \underline{x}_n}{\partial a_i}\right) \tag{9}$$

where

$$\frac{\partial \underline{x}_n}{\partial a_i} = \frac{\partial G_n}{\partial a_i} \underline{u}_n = - G_n \frac{\partial G_n^{-1}}{\partial a_i} G_n \underline{u}_n \quad 1 \leq i \leq k \tag{10}$$

It is desired to maximize J subject to the constraint $\underline{u}_n^T \underline{u}_n = 1$. It is convenient to introduce an (nχn) shift matrix, S_i, $1 \leq i \leq n-1$, as the (nχn) matrix with elements all zero except on the i-th co-diagonal line below the main diagonal line where the elements are all one's, i.e.,

$$(S_i)_{j,j-1} = 1 \qquad i+1 \leq j \leq n$$

all other elements are zero.
In terms of the shift matrices, G_n^{-1} of (3) can be written as

$$G_n^{-1} = I - \sum_{i=1}^{k} a_i S_i$$

** At the end of this paper, the input signal synthesis problem is briefly discussed under the assumption that $\underline{a}$ is a random vector with known a priori probability density function $p_o(\underline{a})$.

Since

$$S_i S_j = S_j S_i \quad \text{for} \quad i+j \leq n-1$$

(see Appendix 1) G_n^{-1} commutes with S_i, $1 \leq i \leq n-1$. Similarly G_n commutes with S_i.
From (10)

$$\frac{\partial \underline{x}_n}{\partial a_i} = + G_n S_i G_n \underline{u}_n = + S_i G_n^2 \underline{u}_n$$

and from (9)

$$\text{tr } M_c = \frac{1}{\sigma^2} \underline{u}_n^T (G_n^2)^T \left(\sum_{i=1}^{k} S_i^T S_i \right) G_n^2 \underline{u}_n \qquad (11)$$

$$= \frac{1}{\sigma^2} \underline{u}_n^T (G_n^2)^T S_1^T (I + S_1^T S_1 + \ldots + S_{k-1}^T S_{k-1}) S_1 G_n^2 \underline{u}_n$$

It is easy to see that G_n^2 is a lower triangular matrix and that

$$S_1 G_n^2 = \begin{array}{|c:c|} \hline 0 & 0 \\ \hdashline A_1 & 0 \\ \hline \end{array}$$

where

A_1 is the $(n-1)\times(n-1)$ main submatrix of G_n^2

$$G_n^2 = \begin{array}{|c:c|} \hline A_1 & 0 \\ \hdashline A_2 & A_3 \\ \hline \end{array}$$

where A_1 and A_3 are both non-singular lower triangular matrices.
Define

$$D_k = I + S_1^T S_1 + \ldots + S_{k-1}^T S_{k-1}$$

Write

$$D_k = \begin{array}{|c|c|} \hline k & 0 \\ \hline 0 & E_k \\ \hline \end{array} \qquad (12)$$

where E_k is the $(n-1)\chi(n-1)$ matrix $E_k = \text{diag}(k,\ldots,k,k-1, k-2,\ldots,2,1)$

Then

$$\sigma^2 J = \underline{u}_{n-1}^T \, G \underline{u}_{n-1}$$

where

$$G \triangleq A_1^T E_k A_1 : (n-1)x(n-1)$$

Thus the constrained optimization problem will be solved if u_n is set to zero and $\underline{u}_{n-1}$ is chosen to be the eigenvector corresponding to the largest eigenvalue of G.

It is more convenient to work with the inverse of G. Then $\underline{u}_{n-1}$ is to be chosen as the eigenvector corresponding to the smallest eigenvalue of G^{-1}.

Now

$$G^{-1} = A_1^{-1} E_k^{-1} (A_1^{-1})^T \qquad (13)$$

where $E_k^{-1} = \text{diag}\,(\frac{1}{k}, \ldots \frac{1}{k}, \frac{1}{k-1}, \frac{1}{k-2}, \ldots, \frac{1}{2}, \frac{1}{1})$ and where A_1^{-1} is the $(n-1)\chi(n-1)$ main submatrix of G_n^{-2} since G_n^2 is lower triangular. Thus

$$A_1^{-1} = (I_{n-1} - \sum_{i=1}^{k} a_i S'_i)^2 = I_{n-1} + \alpha_1 S'_1 + \ldots + \alpha_{2k} S'_{2k} \qquad (14)$$

where S'_i's are $(n-1)\chi(n-1)$ shift matrices defined analogously to $(n\chi n)$ shift matrices. For examples with $k=3$, $n>6$

$$A_1^{-1} = I_{n-1} - 2a_1 S'_1 + (a_1^2 - 2a_2) S'_2 + (2a_1a_2 - 2a_3) S'_3 + (a_2^2 + 2a_1a_3) S'_4 + 2a_2a_3 S'_5 + a_3^2 S'_6$$

Thus, A_1^{-1} is a lower triangular matrix with one's on the diagonal line, and elements on the co-diagonal lines below the main diagonal line are respectively the scalars multiplying $S'_1, S'_2, \ldots, S'_6$ matrices, or

$$\alpha_1 = -2a_1$$
$$\alpha_2 = (a_1^2 - 2a_2)$$

etc

Some numerical solutions for the optimal $\underline{u}$ vector are shown in Figures (1) through (3) for the cases K = 1, 2, 3.

III.2 Case k=1

The matrix G^{-1} takes a specially simple form when k=1. We take the dimension of G^{-1} to be nχn.

$$G_n^{-1} = A_1^{-1}(A_1^{-1})^T$$

where

$$A_1^{-1} = I_n + \alpha_1 S_1 + \alpha_2 S_2$$

where

$$\alpha_1 = -2a$$
$$\alpha_2 = a^2$$

Thus

$$G_n^{-1} = \begin{array}{|cc|} \hline G_{11} & G_{12} \\ G_{12}^T & G_{22} \\ \hline \end{array} \qquad (15)$$

where

$$G_{11} = \begin{bmatrix} 1 & -2a \\ -2a & 1+4a^2 \end{bmatrix} \quad : \quad 2x2$$

$$G_{12} = \begin{bmatrix} a^2 & 0 & 0, \ldots \\ -2a(1+a^2), & a^2, & 0, \ldots \end{bmatrix} \quad : \quad 2x(n-2)$$

and

$$G_{22} = (c_{i-j}) \; : \; (n-2)\chi(n-2)$$

where

$$c_o = 1 + 4a^2 + a^4$$

$$c_1 = c_{-1} = -2a(1+a^2)$$

$$c_2 = c_{-2} = a^2$$

all other c's are zero.
The matrix G_{22} is therefore $(n-2)\chi(n-2)$ Toeplitz matrix.[10]
The matrix G^{-1} can also be written as

$$G_2^{-1} = C_2 - D_2 D_2^T \qquad (15)'$$

where

C_2 is the $n\chi n$ Toeplitz matrix with c_o, $c_{\pm 1}$, $c_{\pm 2}$ as specified above and where

$$D_2^T = \begin{bmatrix} -2a, & a^2, & 0 \ldots \\ a^2, & 0, & 0 \ldots \end{bmatrix} \; : \; 2xn$$

The analytic solution of this problem is discussed in more detail in Appendix (3).

IV. Output Constrained Case

IV.1 General Discussions

In this section, the output vector $\underline{x}_n$ is obtained which maximizes a function of M_c. This problem may be of interest in its own right in some situations. It is shown in the next section that an approximate method of solving the input constrained case also leads to the study of this case. From (9)

$$\sigma^2 J = \sum_{i=1}^{k} \left(\frac{\partial \underline{x}_n}{\partial a_i}\right)^T \left(\frac{\partial \underline{x}_n}{\partial a_i}\right)$$

where

$$\frac{\partial \underline{x}_n}{\partial a_i} = - G_n \frac{\partial G_n^{-1}}{\partial a_i} \quad \underline{x}_n = + G_n S_i \underline{x}_n$$

Therefore

$$\sigma^2 J = \underline{x}_n^T \left(\sum_{i=1}^{k} S_i^T G_n^T G_n S_i \right) \underline{x}_n$$

Since S_i and G_n commute

$$\sigma^2 J = \underline{x}_n^T G_n^T S_1^T (I + S_1^T S_1 + \ldots + S_{k-1}^T S_{k-1}) \cdot$$

$$S_1 G_n \underline{x}_n = \underline{x}_{n-1}^T S \underline{x}_{n-1}$$

where

$$S \triangleq B_1^T E_k B_1$$

where B_1 is $(n-1)\chi(n-1)$ matrix

$$G_n = \begin{array}{|c|c|} \hline B_1 & 0 \\ \hline B_2 & B_3 \\ \hline \end{array}$$

and where E_k has been defined in connection with (11)'. It is now desired to maximize J subject to the constraint $\underline{x}_n^T \underline{x}_n < 1$. This problem is solved if $x_n = 0$ and $\underline{x}_{n-1}$ is an eigenvector of the matrix S corresponding to its largest eigenvalue. Again it is more convenient to find the maximum eigenvalue of S and its corresponding eigenvector by obtaining the smallest eigenvalue of S^{-1} and its corresponding eigenvector, where

$$S^{-1} = B_1^{-1} E_k^{-1} (B_1^{-1})^T$$

Some numerical solutions for the optimal $\underline{x}$ vector are shown in Figure 4, the cases k=1.

IV.2 Case K=1

When k=1, S^{-1} takes a particularly simple form

$$S^{-1} = B_1^{-1}(B_1^{-1})^T = \begin{pmatrix} S_{11} & S_{12} \\ S_{12}^T & S_{22} \end{pmatrix} \; ; \; (n-1)\chi(n-1)$$

where

$$S_{11} = 1 \qquad S_{12} = [-a \;\; 0 \;\; 0 \;\; ..] \; ; \; 1\chi(n-2)$$

$$S_{22} = (c_{i-j}) \; ; \; (n-2)\chi(n-2)$$

$$c_o = 1 + a^2$$

$$c_1 = c_{-1} = -a$$

all other c's are zero. An alternate expression for S^{-1} is

$$S^{-1} = C - DD^T$$

where

C is the $(n-1)\chi(n-1)$ Toeplitz matrix with only non-zero elements being c_o, c_1 and c_{-1} mentioned above, and where

$$D^T = S_{12}$$

As shown in Appendix (2), it is possible to obtain the characteristic values and vectors of $(n\chi n)\ S^{-1}$ explicitly in the case k=1. The characteristic values are

$$\lambda_k = 1 + a^2 - 2a \cos \theta_k \qquad k=1,2,\ldots,n$$

where θ_k are the solutions of

$$\frac{\sin (n+1)\theta}{\sin n\theta} = a$$

The characteristic vectors are given as

$$x_k = A \sin (n+1-k)\theta_1$$

where A is a constant of normalization, where θ_1 corresponds to the smallest eigenvalue. Thus for k=1, the output constrained problem has been solved exactly. From (2)

$$\underline{u}_n = G_n^{-1}\ \underline{x}_n$$

i.e.,

$$u_i = -ax_{i-1} + x_i$$

$$= A[-a \sin (n+2-i)\theta_1 + \sin(n+1-i)\theta_1\ ] \quad 0 \leq i \leq n-1$$

IV.3 Approximation

Rewrite (10) as

$$\frac{\partial \underline{x}_n}{\partial a_i} = + G_n S_i \underline{x}_n$$

$$M_c = \frac{1}{\sigma^2} X^T G_n^T G_n X$$

where

$$X \triangleq [S_1\ \underline{x}_n,\ S_2\ \underline{x}_n,\ \ldots,\ S_k\ \underline{x}_n] : (n\chi k) \text{ matrix}$$

Denote the smallest and the largest eigenvalues of $G_n^T G_n$ by μ_{min} and μ_{max} respectively.

Then

$$\mu_{min}\, X^T X \leq \sigma^2 M_c \leq \mu_{max}\, X^T X \qquad (16)$$

where

$$\mu_{min} \geq \frac{1}{(1 + \sum_{i=1}^{k} |a_i|)^2}$$

and

$$\mu_{max} \leq \frac{1}{(1 - \sum_{i=1}^{k} |a_i|)^2}$$

provided $\Sigma_{i=1}^{k} |a_i| < 1$*

by arguments similar to those of establishing Gersgorin's disks[9]. Therefore, as an approximation, one might try maximizing trX^TX rather than tr M_c directly with the inputs constraint $\underline{u}_n^T \underline{u}_n = 1$.

Let

$$J' = \text{tr}\, X^T X$$

Then

$$J' = \underline{x}_n^T \left(\sum_{i=1}^{k} S_i^T S_i \right) \underline{x}_n$$

$$= \underline{u}_n^T G_n^T \left(\sum_{i=1}^{k} S_i^T S_i \right) G_n \underline{u}_n = \underline{u}_{n-1}^T S \underline{u}_{n-1}$$

Thus the problem reduces to that of the output constrained case discussed previously in IV.2.

V. Approximation by the Toeplitz Matrix

In the input constrained problem of Section III it has been shown that the optimal $\underline{u}_{n-1}$ is the eigenvector

*In case $\Sigma_{i=1}^{k} |a_i| > 1$ but the system is stable, then $\mu_{max} \leq \left[\sum |g_i|\right]^2 < \infty$ where $\{g_i\}$ is the weighting sequence of the system (1).

corresponding to the smallest eigenvalue of the (n-1)x(n-1) matrix

$$G^{-1} = A_1^{-1} E_k^{-1} (A_1^{-1})^T$$

while in the output constrained problem, it has been shown that the optimal $\underline{x}_{n-1}$ is the eigenvector corresponding to the smallest eigenvalue of the (n-1)x(n-1) matrix

$$S^{-1} = B_1^{-1} E_k^{-1} (B_1^{-1})^T$$

where E_k is the (n-1)x(n-1) matrix defined in connection with (12).

Since both A_1^{-1} and B_1^{-1} are lower triangular matrices of the general form*

$$Q = I_{n-1} + \sum_{i=1}^{\ell} \alpha_i S_i$$

an approximation technique is developed here in a general way which can apply to either the input constrained or output constrained case. Hence, in this section, we consider the problem of finding the eigenvector corresponding to the smallest eigenvalue of the matrix

$$H = Q E_k^{-1} Q^T \tag{17}$$

where Q is either A_1^{-1} or B_1^{-1}.

Note that

$$E_k^{-1} = \frac{1}{k} \left\{ I_{n-1} + \begin{bmatrix} 0 & 0 \\ 0 & F \end{bmatrix} \right\} \tag{18}$$

when F is the (k-1)x(k-1) matrix given by F = diag($\frac{1}{k-1}$, $\frac{2}{k-2}$, ..., $\frac{k-1}{1}$).

---

*$\ell=k$ for $Q=B_1^{-1}$, and $\ell=2k$ for $Q=A_1^{-1}$

Therefore

$$H = \frac{1}{k} [Q\ Q^T + T] \tag{19}$$

where

$$T = \begin{bmatrix} 0 & 0 \\ 0 & Q_3 F Q_3^T \end{bmatrix}$$

and where

$$Q = \begin{bmatrix} Q_1 & 0 \\ Q_2 & Q_3 \end{bmatrix}$$

where Q_1 is $(n-k)\chi(n-k)$ and Q_3 is $(k-1)\chi(k-1)$

The matrix Q_3 is of the form

$$Q_3 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \alpha_1 & 1 & 0 & \cdots & . \\ \alpha_2 & \alpha_1 & 1 & & . \\ . & . & & & . \\ . & . & & & . \\ . & . & & & . \\ \alpha_{k-2} & . & . & . \ . \ . & \alpha_1 \quad 1 \end{bmatrix}$$

As in (15)',

$$QQ^T = C - D_\ell\ D_\ell^T \tag{20}$$

where C is the $(n-1)\chi(n-1)$ Toeplitz matrix with elements $c_{ij} = c_{|i-j|}$ given by

$$c_o = 1 + \sum_{i=1}^{\ell} \alpha_i^2$$

$$c_1 = \alpha_1 + \alpha_1 \alpha_2 + \cdots \alpha_{\ell-1} \alpha_\ell$$

$$c_2 = \alpha_2 + \alpha_1 \alpha_3 + \cdots \alpha_{\ell-2} \alpha_\ell$$

$$\begin{aligned} c_i &= \alpha_i + \alpha_1 \alpha_{i+1} + \cdots \alpha_{\ell-i}\ \alpha_\ell \\ &\vdots \\ c_\ell &= \alpha_\ell \end{aligned}$$

with $c_{|i-j|} = 0$ for $|i-j|>\ell$, and where the $\ell\chi(n-1)$ matrix D_ℓ is described by

$$D_\ell^T = [R_\ell \quad 0]$$

where R_ℓ is the $\ell\chi\ell$ matrix

$$R_\ell = \begin{bmatrix} \alpha_1 & \alpha_2 & & \alpha_\ell \\ \alpha_2 & & & \\ & & & 0 \\ \alpha_\ell & & & \end{bmatrix}$$

The elements $c_o, c_1, \ldots, c_\ell$ are generated by the equality

$$g(z) = p(z)\ p(\tfrac{1}{z}) = c_o + c_1(z+z^{-1}) + \ldots + c_\ell\ (z^\ell + z^{-\ell})$$

where

$$p(z) = 1 + \sum_{i=1}^{\ell} \alpha_i\ z^i$$

It is known that the minimimum eigenvalue of H is given by

$$\lambda_{min} = \min_{\underline{x}} \frac{\underline{x}^T H \underline{x}}{||\underline{x}||^2}$$

and that the minimizing $\underline{x}$ is the corresponding eigenvector. From (19) and (20)

$$\frac{\underline{x}^T H \underline{x}}{||\underline{x}||^2} = \frac{1}{k}\left[\frac{\underline{x}^T C \underline{x}}{||\underline{x}||^2} - \frac{\underline{x}^T D\ D^T \underline{x}}{||\underline{x}||^2} + \frac{\underline{x}^T T^T \underline{x}}{||\underline{x}||^2}\right] \qquad (21)$$

If the (n-1) dimensional vector $\underline{x}$ is partitioned as

$$\underline{x} = \begin{bmatrix} \underline{u} \\ \underline{v} \\ \underline{w} \end{bmatrix}$$

where $\underline{u}$ is ℓx1, $\underline{w}$ is (k-1)x1, and $\underline{v}$ is (n-k-ℓ-1)x1, from (21)

$$\frac{\underline{x}^T H \underline{x}}{||\underline{x}||^2} = \frac{1}{k}\left[\frac{\underline{x}^T C \underline{x}}{||\underline{x}||^2} - \frac{\underline{u}^T R_\ell^T R_\ell \underline{u}}{||\underline{x}||^2} + \frac{\underline{w}^T Q_3 F Q_3^T \underline{w}}{||\underline{x}||^2}\right]$$

where

$$||\underline{x}||^2 = ||\underline{u}||^2 + ||\underline{v}||^2 + ||\underline{w}||^2$$

If the normalization of $\underline{x}$ is chosen as

$$\lim_{n\to\infty} \frac{1}{n} ||\underline{x}||^2 = M \text{ such that } |x_i| \leq L \text{ for all } i$$

where M and L are some constants, then

$$\lim_{n\to\infty} \frac{1}{n} ||\underline{u}||^2 = \lim_{n\to\infty} \frac{1}{n} ||w||^2 = 0$$

Thus, under the above conditions

$$\lim_{n\to\infty} \frac{\underline{x}^T H \underline{x}}{||x||^2} = \frac{1}{k} \frac{\underline{x}^T C \underline{x}}{||x||^2}$$

Therefore, the choice of $\underline{x}$ to be an eigenvector corresponding to the smallest eigenvalue of the Toeplitz matrix C is a good approximation for large n of the eigenvector of H.

The Toeplitz matrix has many interesting properties which make it a desirable approximation. In the case that n is infinite the eigenvalues of C coincide with the set of values g(z) assumes on the unit circle $|z| = 1$, where

$$g(z) = \sum_{-\infty}^{+\infty} c_i z_i$$

and where the coefficients are the numbers constituting the elements of C. If they are all zero for $i>\ell$ then the above function reduces to

$$g(z) = \sum_{-\ell}^{\ell} c_i z^i = c_o + 2 \sum_{i=1}^{\ell} c_i \cos i\theta \qquad (22)$$

If n is finite then the eigenvalues still take values given by (22) but now θ is restricted to take on n discrete values in $[0, 2\pi]$.

VI. Extensions

VI.1 System with Random System Parameters

When $\underline{a}$ is regarded as a random vector with known a priori probability density function $p_o(\underline{a})$, given any unbiased unconditional estimator $\underline{\gamma}(\underline{y}_n)$, it is known that[8]

$$E_{\underline{y}_n,\ \underline{a}}\ [\underline{\gamma}(\underline{y}_n) - \underline{a}]\ [\underline{\gamma}(\underline{y}_n) - \underline{a}]^T \geq M^{-1}$$

where

$$E_{\underline{y}_n,\ \underline{a}}\ [\underline{\gamma}(\underline{y}_n)] = E_{\underline{a}}(\underline{a})$$

and where

$$M \triangleq \int p_o(\underline{a})\ p(\underline{y}_n|\underline{a})[\nabla_{\underline{a}} \ln p_o(\underline{a})\ p(\underline{y}_n|\underline{a})]\cdot$$

$$[\nabla_{\underline{a}} \ln p_o(\underline{a})\ p(\underline{y}_n|\underline{a})]^T\ d\ \underline{y}_n\ d\ \underline{a}$$

In the case of Gaussian noises assumed here, it is easy to see that

$$M = E_{\underline{a}}\ (M_{\underline{a}} + M_{\underline{c}})$$

where

$$M_{\underline{a}} \triangleq [\nabla_{\underline{a}} \ln p_o(\underline{a})]\ [\nabla_{\underline{a}} \ln p_o(\underline{a})]^T$$

Since $M_{\underline{a}}$ depends only on $p_o(\underline{a})$ and not on $\underline{u}_n$, a measure of error may be taken to be again tr $E_{\underline{a}}(M_{\underline{c}})$. With this criterion function, results for the single parameter output signal constrained case are summarized here. Obvious extensions to open-loop feedback control laws can be made as in Reference 6. The detail will be the subject of a separate paper.

From Appendix 2 and IV.2, the smallest eigenvalue of NxN matrix $E_a(S^{-1})$ is given by

$$\lambda_{min} = 1 + m_2 - 2m_1 \cos \bar{\theta}$$

where

$$m_1 \triangleq E_a(a)$$

$$m_2 \triangleq E_a(a^2)$$

and where

$$\bar{\theta} \approx \frac{\pi}{N}\left[1 - \frac{1}{N(1-m_1) + 1}\right]$$

To obtain the corresponding eigenvector, the recursion equation (2.6) of Appendix 2 is now replaced by

$$x_2 = (2\mu - m_2/m_1)\, x_1$$

$$x_{i-1} + x_{i+1} = 2\mu\, x_i \qquad 2 \leq i \leq n-1$$

$$x_{n-1} = 2\mu\, x_n \qquad \text{where } \mu = \cos\bar{\theta}$$

Proceeding analogously as in Appendix 2,

$$x_k = [(2\mu - m_2/m_1)\sin(k-1)\theta - \sin(k-2)\theta]\, x_1 \qquad 2 \leq k \leq n$$

VI.2 Non-zero Initial Conditions

It is straightforward to extend the analysis presented in this paper to the case with non-zero initial conditions. Some of the results in this paper can also be extended to the system with the dynamic equation given by

$$x_{i+1} + \sum_{j=1}^{k} a_j\, x_{i-j} = \sum_{j=0}^{m} b_j\, u_{i-j}$$

where b's as well as a's are unknowns. These extensions will be presented in Reference (11).

VI.3 Identifiability Conditions

Equation (16) can be used to define several types of "identifiability" and/or give the criteria of identifiability. For example, using the concept of complete identifiability[3], its necessary and sufficient conditions under suitably regularity conditions on $p(\underline{a}|\underline{y}_n)$ are that all the eigenvalues of $X^T X \to \infty$ as $n \to \infty$.

The concept of weak identifiability may be defined to be that the Fisher information matrix is positive definite for some n, quite analogously to the concept of weak observability[12] the necessary condition of weak identifiability is that $n > k$ where k is the number of unknown parameters. The detailed account of these topics will be found in Reference 11.

VII. Discussions and Conclusions

The problem of synthesizing input signals in off-line system identification problem has been discussed by maximizing the trace of the Fisher information matrix. It has been shown that the problem reduces to that of finding the largest eigen-value of a certain matrix and its structure has been investigated.

For details of the case with non-zero initial conditions and the problem with more general linear difference equation than (1), see Ref. 13.

Some computational aspects of the input signal synthesis problem has been presented in Ref. 15.

APPENDIX 1
Some Properties of Shift Matrices

The $(n\chi n)$ shift matrix S_i is defined to be

$$(S_i)_{j,j-1} = 1 \qquad i+1 \leq j \leq n$$

all other elements zero.

Another way of defining S_i is in terms of elementary column vectors

$$e_k^T = (0,\ldots,0,\underset{\text{k-th element}}{1},0\ldots0)$$

$$S_i = (e_{i+1},e_{i+2},\ldots,e_n,0,\ldots,0) \text{ or } S_i^T = (0,0,\ldots0,e_i\ldots e_{n-i})$$

For example

$$S_1 = (e_2,e_3,\ldots,e_n,0) \text{ or } S_1^T = (0,e_1,e_2,\ldots,e_{n-1})$$

It is easy to verify that

$$S_iS_j = S_jS_i = \begin{cases} S_{i+j} & ; \ i+j<n \\ 0 & ; \ i+j>n \end{cases}$$

APPENDIX 2

Computation of the Characteristic Values and the Characteristic Vectors of S^{-1} for k=1

Given the nχn matrix S_n^{-1}

$$S_n^{-1} = B_1^{-1}(B_1^{-1})^T$$

where

$$B_1^{-1} = I_n - aS_1$$

consider its characteristic equation

$$\Delta_n(\lambda) \triangleq |\lambda I_n - S_n^{-1}|$$

Expanding $\Delta_n(\lambda)$ by the last row, it satisfies the difference equation

$$\Delta_n(\lambda) = (\lambda-\beta)\Delta_{n-1}(\lambda) - a^2 \Delta_{n-2}(\lambda), \quad n>2 \tag{2.1}$$

where

$$\beta \triangleq 1 + a^2$$

$$\Delta_o(\lambda) = 1$$

$$\Delta_1(\lambda) = \lambda - 1$$

Introduce the notation

$$\lambda - \beta = -2a \cos\theta$$

Then (2.1) becomes

$$\Delta_n(\beta-2a\cos\theta) = -2a\cos\theta\Delta_{n-1}(\beta-2a\cos\theta) - a^2\Delta_{n-2}(\beta - 2a\cos\theta) \tag{2.2}$$

Assuming a solution of the form

$$\Delta_n(\beta-2a\cos\theta) = A\rho^n$$

equation (2.2) is solved by obtaining the roots of

$$\rho^2 + 2a\cos\theta\rho + a^2 = 0$$

or

$$\rho = -ae^{\pm i\theta}$$

Thus

$$\Delta_n(\beta-2\ a\ \cos\ \theta) = (-a)^n\ [Ae^{in\theta}\ + Be^{-in\theta}] \qquad (2.3)$$

where A and B are constants to be determined by the initial conditions in (2.1), or

$$A + B = 1$$

$$-a(Ae^{i\theta} + Be^{-i\theta}) = \beta-2a\ \cos\ \theta - 1$$

i.e.,

$$A = (-a + e^{i\theta})/(e^{i\theta} - e^{-i\theta})$$

$$B = (a - e^{-i\theta})/(e^{i\theta} - e^{-i\theta})$$

Substituting these expressions in (2.3)

$$\Delta_n(\beta-2a\ \cos\ \theta) = (-a)^n\ \frac{\sin(n+1)\theta\ - a\ \sin\ n\theta}{\sin\ \theta} \qquad (2.4)$$

The n eigenvalues of S_n^{-1} are then given as the zeros of Δ_n, i.e.,

$$\lambda_k = \beta - 2a\ \cos\ \theta_k \quad k=1,\ldots,n$$

where θ_k are the values satisfying

$$\sin(n+1)\theta = a\ \sin\ n\theta \qquad (2.5)$$

Next the eigenvector corresponding to the smallest eigenvalue will be derived.

From the above discussion we know that λ_{min} is of the form

$$\lambda_{min}(n) = 1 + a^2 - 2a\mu$$

where

$$\mu \triangleq \cos\ \theta_1\ ,\ \theta_1 \approx \frac{\pi}{n}\ [1 - \frac{1}{(1-a)n-1}]$$

The equation

$$S_n^{-1}\ \underline{x}_n = \lambda_{min}(n)$$

gives a set of simultaneous equations

$$x_2 = (2\mu - a)\, x_1$$

$$x_{i-1} + x_{i+1} = 2\mu x_i \qquad 2 \leq i \leq n-1 \tag{2.6}$$

$$x_{n-1} = 2\mu x_n$$

Assuming a solution of the form $x_k = A\rho^k$, we see that the middle equation will be satisfied if

$$\rho^2 - 2\mu\rho + 1 = 0 \text{ which has the solutions } \rho = e^{\pm i\theta_1}$$

The components of the eigenvector are expressible as

$$x_k = Ae^{ik\theta_1} + Be^{-ik\theta_1}$$

where A and B must satisfy the boundary conditions

$$(2\mu - a)(Ae^{i\theta_1} + Be^{-i\theta_1}) = Ae^{2i\theta_1} + Be^{-2i\theta_1}$$

and

$$Ae^{i(n-1)\theta_1} + Be^{-i(n-1)\theta_1} = 2\mu(Ae^{in\theta_1} + Be^{-in\theta_1})$$

One of the conditions gives the ratio A/B and the other condition reproduces (2.5). The result is

$$x_k = A_1 \sin [(n+1) - k]\theta_1 \qquad 1 \leq k \leq n \tag{2.7}$$

where A_1 is the constant of normalization.
An alternate expression for x_k is

$$x_k = A_2 [\sin k\,\theta_o - a \sin (k-1)\,\theta_o]$$

where A_1 and A_2 is determined from the normalization condition such as

$$\sum_{k=1}^{n} |x_k|^2 = \text{const.}$$

or

$$\sum_{k=1}^{n} |x_k|^2 = \text{proportional to } n$$

This latter normalization corresponds to the constant power (energy per unit time) requirement while the former corresponds to the constant total energy condition.

APPENDIX 3

Computation of the Characteristic Values and the Characteristic Vectors of G^{-1} for k=1

It was shown earlier that for k=1

$$G_n^{-1} = A_1^{-1} (A_1^{-1})^T \quad :nXn$$

where

$$A_1^{-1} = I_n + \alpha_1 S_1 + \alpha_2 S_2$$

and

$$\alpha_1 = -2a$$

$$\alpha_2 = a^2$$

Assuming that λ has the same form as the λ of the Toeplitz sub-matrix G_{22}, we can write that

$$\lambda = c_o + 2c_1 \cos\theta + 2c_2 \cos 2\theta$$

where

$$c_o = 1 + 4a^2 + a^4$$

$$c_1 = -2a\,(1 + a^2)$$

$$c_2 = a^2$$

The matrix equation for the eigenvector $\underline{x}$, i.e., the equation

$$G_N^{-1}\,\underline{x} = \lambda \underline{x}$$

can thus be written as the set of difference equations

$$c_2x_3 - 2ax_2 + x_1 = \lambda x_1 \qquad (3.1)$$

$$c_2x_4 + c_1x_3 + (1+4a^2)x_2 - 2ax_1 = \lambda x_2 \qquad (3.2)$$

$$c_2x_{k+5} + c_1x_{k+4} + c_ox_{k+3} + c_1x_{k+2} + c_2x_{k+1} = \lambda x_{k+3} \qquad (3.3)$$

$$c_2 x_{n-3} + c_1 x_{n-2} + c_0 x_{n-1} + c_1 x_n = \lambda x_{n-1} \qquad (3.4)$$

$$c_2 x_{n-2} + c_1 x_{n-1} + c_0 x_n = \lambda x_n \qquad (3.5)$$

The difference equation (3.3) can be written as

$$c_1(x_{k+4} + x_{k+2}) + c_2(x_{k+5} + x_{k+1}) = 2\,(c_1 \cos\theta + c_2 \cos 2\theta)\cdot x_{k+3}$$

Assuming a solution of the form

$$x_j = A\rho^j$$

we see that ρ must satisfy the equation

$$c_1(\rho + \rho^{-1}) + c_2(\rho^2 + \rho^{-2}) - 2(c_1 \cos\theta + c_2 \cos 2\theta) = 0 \qquad (3.6)$$

Clearly, two of the solutions are

$$\rho = e^{\pm i\theta}$$

After dividing (3.6) by $(\rho - e^{i\theta})(\rho - e^{-i\theta})$, the other two solution is given as the roots of

$$c_2 \rho^2 + (c_1 + 2c_2 \cos\theta)\rho + c_2 = 0$$

or where

$$\rho = e^{\pm\alpha}$$

$$\frac{c_1 + 2c_2 \cos\theta}{c_2} \triangleq -2\cosh\alpha$$

or

$$\cos\theta + \cosh\alpha = \left(\frac{1}{a} + a\right)$$

Hence, the components of the eigenvector will be of the form

$$x_k = A(e^{i\theta})^k + B(e^{-i\theta})^k + C(e^{\alpha})^k + D(e^{-\alpha})^k \qquad (3.6)$$

The coefficients A, B, C, D and θ can be determined by applying the boundary conditions described in equations (3.1)-(3.5).

In the case of nχn Toeplitz matrix,

$$x_i = x_{n-i}$$

and

$$Ae^{in\theta} = B$$

$$C^{\alpha n} = D$$

APPENDIX 4

Asymptotic Distribution of the Characteristic Values of the Toeplitz Matrix

Let

$$g(\theta) = c_o + 2\sum_{j=1}^{\ell} c_j \cos j\theta$$

Then c_o, c_1, ..., c_ℓ are the Fourier coefficients of $g(\theta)$

Then the characteristic values of c_ℓ and

$$g\left(-\pi + \frac{2v\pi}{n+1}\right) \qquad v=1,2,\ldots,n$$

are equally distributed as $n\to\infty$ in the sense of Reference (10).

REFERENCES

1. Turin, G. L., "On the Estimation in the Presence of Noise of the Impulse Response of a Random Linear Filter," I.R.E. Transactions on Information Theory, Vol. IT-3, pp. 5-10, March, 1967.

2. Levin, M. J., "Optimum Estimation of Impulse Response is the Presence of Noise," I.R.E. Transaction on Circuit Theory, Vol. CT-7, March 1960.

3. Astrom, K. J., "Numerical Identification of Linear Dynamic Systems from Normal Operating Records," Proceedings of the 2nd IFAC Symposium, September 14-17, 1965, Teddington, England.

4. Levadi, V. S., "Design of Input Signals for Parameter Estimation," I.E.E.E. Transactions on Automatic Control Vol. AE-11, No. 2, April 1966.

5. Cagliardi, R. M., "Input Selection for Parameter Identification in Discrete Systems," I.E.E.E. Transactions on Automatic Control, Vol. AC-12, No. 5, October 1967.

6. Aoki, M., and R. M. Staley, "On Approximate Input Signal Synthesis in Plant Parameter Identification," Presented at at the 1st Hawaii International Conference on System Sciences, University of Hawaii, January 1968.

7. Middleton, D., "An Introduction to Statistical Communication Theory," McGraw-Hill Book Company, Inc., New York, 1960.

8. Kullback, S., "Information Theory and Statistics," Wiley & Son, New York, 1959.

9. Astrom, K. J., "On the Achievable Accuracy in Identification Problems," Paper 1.8, Proc. IFAC Symposium on Identification in Automatic Control Systems, Prague, Czechoslovakia, June 1967.

10. Spang, H. A., "Optimum Control of an Unknown Linear Plant Using Bayesian Estimation of Error," Report 64-RL-3703E, G.E. Research Lab, Schenectady, New York, 1964.

11. Marcus, M. and H. Minc, A Survey of Matrix Theory and Matrix Inequalities, Allyan & Bacon, Inc., Boston, 1964.

12. Grenander, M., and G. Szego, Toeplitz Forms and Their Applications, University of California Press, Berkeley and Los Angeles, 1958.

13. Staley, R. M., "Input Signal Synthesis in Identification Problems," Ph.D. Dissertation, Department of Engineering, University of California, Los Angeles, Fall, 1968.

14. Aoki, M., "On Observability of Stochastic Discrete-Time Dynamic Systems," to appear in J. Franklin Institute.

15. Aoki, M. and R. M. Staley, "Some Computational Considerations in Input Signal Synthesis Problems" Second International Conference on Computing Methods in Optimization Problems sponsored by SIAM, San Remo, Italy, September 1968.

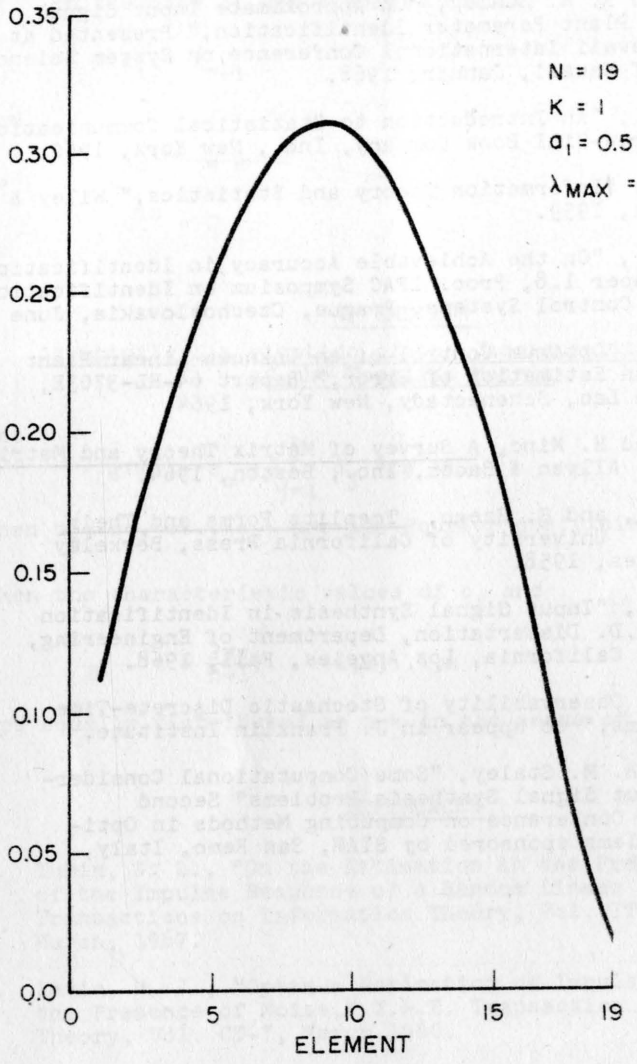


Fig.1 INPUT CONSTRAINT CASE

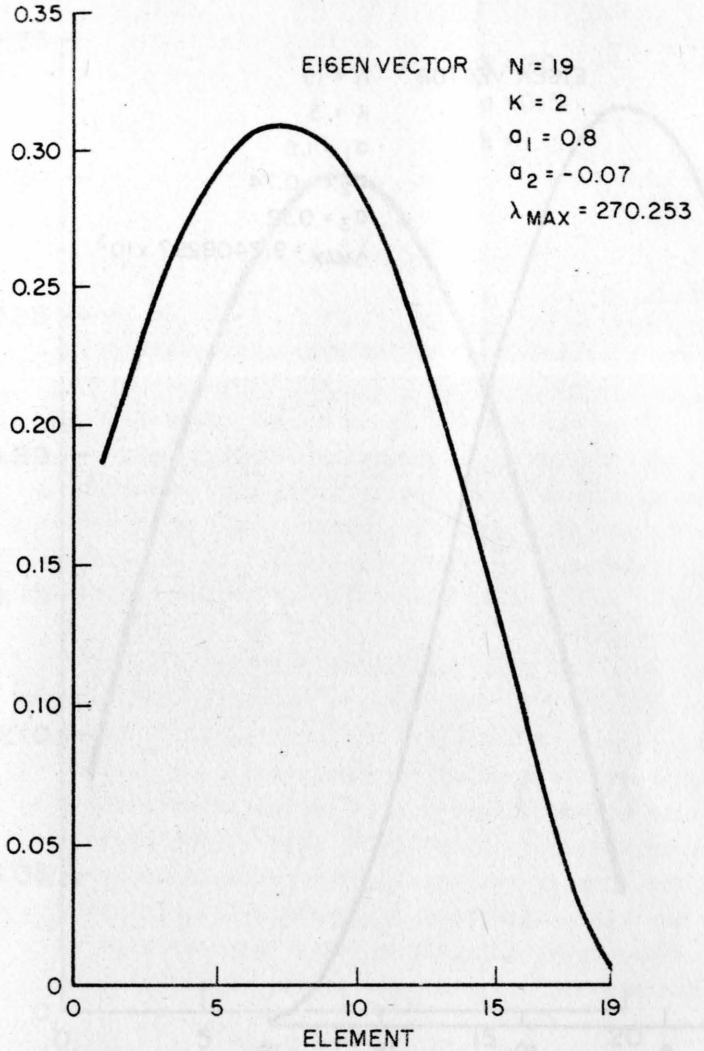

EIGEN VECTOR
N = 19
K = 2
a_1 = 0.8
a_2 = -0.07
λ_{MAX} = 270.253
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0
0
5
10
15
19
ELEMENT
INPUT CONSTRAINT CASE

Fig. 2

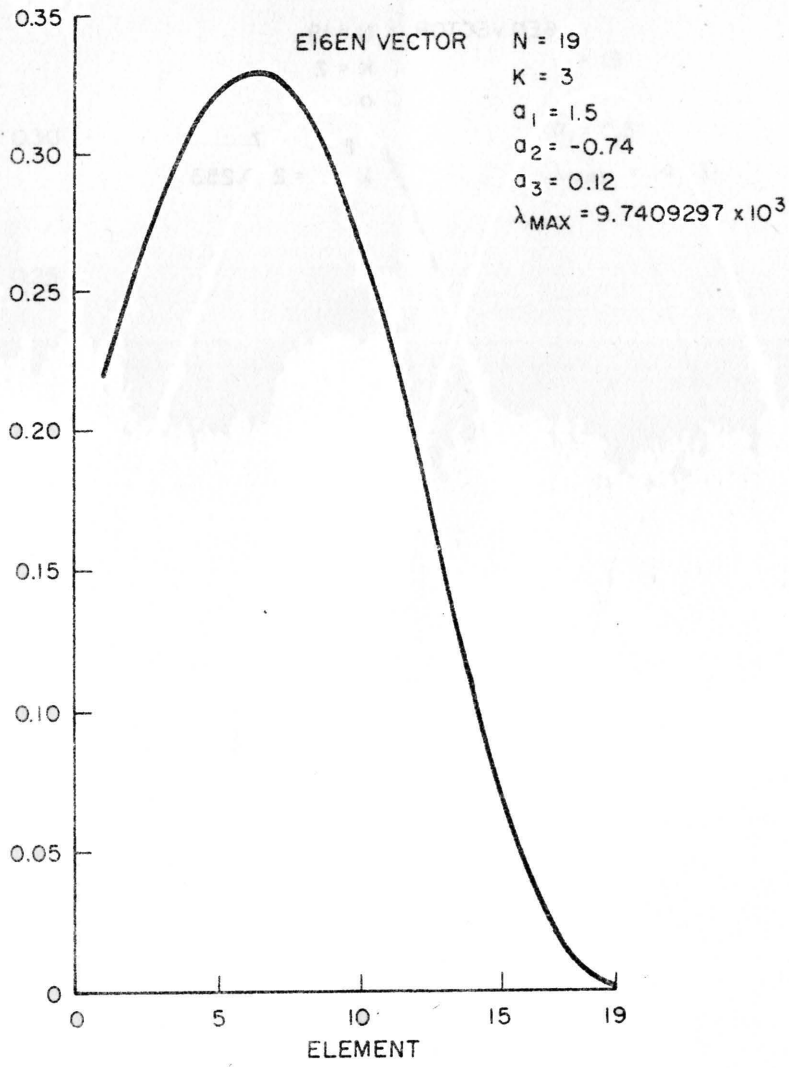


Fig. 3 INPUT CONSTRAINT CASE

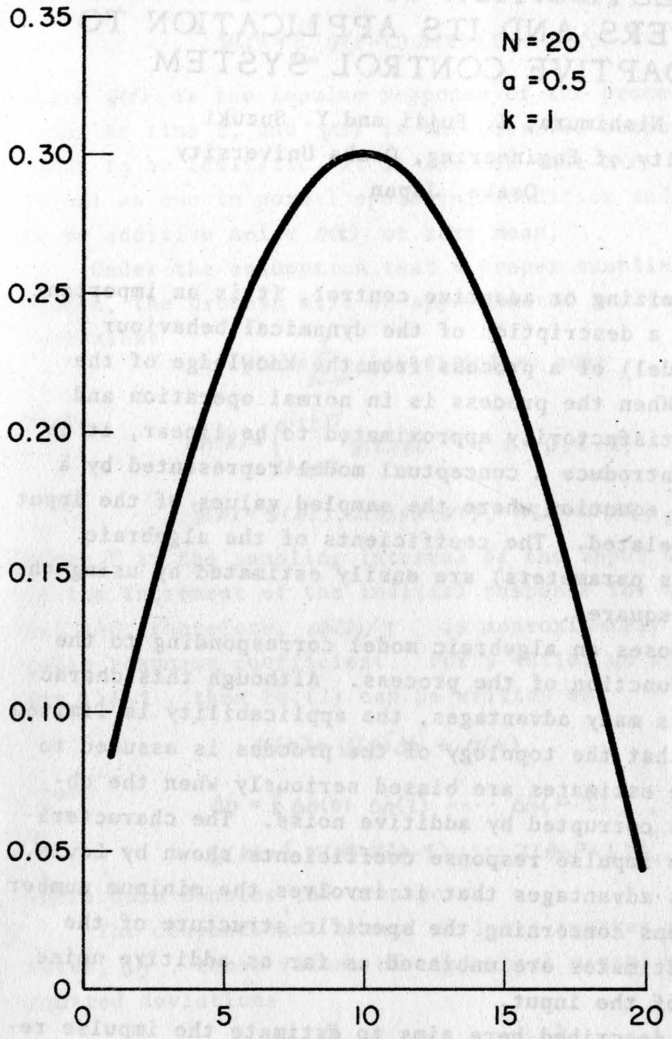


Fig. 4 OUTPUT CONSTRAINED CASE

ON-LINE ESTIMATION OF THE PROCESS PARAMETERS AND ITS APPLICATION TO AN ADAPTIVE CONTROL SYSTEM

M. Nishimura, K. Fujii and Y. Suzuki
Faculty of Engineering, Osaka University
Osaka, Japan

1. Introduction

In the optimizing or adaptive control, it is an important aspect to obtain a description of the dynamical behaviour (mathematical model) of a process from the knowledge of the observed data. When the process is in normal operation and assumed to be satisfactorily approximated to be linear, it is recommended to introduce a conceptual model represented by a linear algebraic equation where the sampled values of the input and output are related. The coefficients of the algebraic equation (process parameters) are easily estimated by using the method of least-squares.

Kalman[1] proposes an algebraic model corresponding to the pulse transfer function of the process. Although this characterization offers many advantages, the applicability is limited by the reasons that the topology of the process is assumed to be known and the estimates are biased seriously when the observed output is corrupted by additive noise[2]. The characterization using the impulse response coefficients shown by Levin[3], by contrast, has advantages that it involves the minimum number of the assumptions concerning the specific structure of the model and the estimates are unbiased as far as additive noise is independent of the input.

The method described here aims to estimate the impulse response coefficients by on-line data processing, where the concept of weighted least-squares suggested by Kalman is introduced. It is pointed out that the method of weighted least-squares poses the estimation process for the low-pass filter. A concept of equivalent data length is introduced and the problem to estimate the slowly time varying parameter is discussed.

2. Principle

A linear and time-invariant process can be expressed in a

convolution integral

$$y(t)=\int_{-\infty}^{t} g(t-\tau)x(\tau)d\tau + n(t) \tag{1}$$

where $g(t)$ is the impulse response of the process, $x(t)$ is the input at time τ, and $y(t)$ is the observed output at time t. In order to be realistic, it is assumed that $x(t)$ is an arbitrary signal as one in normal operating condition and $y(t)$ is corrupted by additive noise $n(t)$ of zero mean.

Under the assumption that a proper sampling period is chosen, the process will be approximated to its sampled data equivalent

$$y(k)=\sum_{\ell=-\infty}^{k} \Delta h(k-\ell)x(\ell) + n(k) \tag{2}$$

where

$$\Delta h(\ell)=\int_{(\ell-\frac{1}{2})T}^{(\ell+\frac{1}{2})T} g(t)\,dt \qquad (\ell = 0, 1, 2, \cdots\cdots),$$

$$y(k)=y(kT),\ x(k)=x(kT),\ n(k)=n(kT),$$

where T is the sampling interval of the input and output. $\Delta h(\ell)$ is the increment of the inditial response for a sampling interval and, therefore, $\Delta h(\ell)/T$ is approximately equal to the impulse response coefficient. For a while, we assume that $\Delta h(\ell)\approx 0$ for $\ell > P-1$, then Eq.(2) can be written as

$$y(k)=\underline{u}(k)'\Delta\underline{h} + n(k) \tag{3}$$

where

$$\Delta\underline{h}=[\,\Delta h(0)\ \Delta h(1)\ \cdots\ \Delta h(P-1)\,]',$$

$$\underline{u}(k)=[\,x(k)\ x(k-1)\ \cdots\ x(k-P+1)\,]',$$

where dash denotes the transpose.

The conventional method of least-squares selects as estimates, $\Delta\underline{h}^{*}$, those values of $\Delta\underline{h}$ which minimize the sum of squared deviations

$$J=\sum_{\ell=k-N}^{k}\left\{\, y(\ell)-\underline{u}(\ell)'\Delta\underline{h}\,\right\}^{2} \tag{4}$$

where $N+1$ is the number of the observations (sampled data) of the output. The estimates are given by the set of simultaneous linear equations (the so-called normal equations)

$$\sum_{\ell=k-N}^{k} \underline{u}(\ell)\underline{u}(\ell)'\Delta\underline{h}^{*} = \sum_{\ell=k-N}^{k} \underline{u}(\ell)\,y(\ell). \tag{5}$$

Usually Eq.(5) is solved directly by using a computer where a batch of the input-output data is processed at a time,

but this approach is out of the present discussion on the on-line estimation. Alternative method to decide $\Delta\underline{h}^{*}$ is to use the recursive equations[4]. However, it is not practical to apply this procedure to the on-line estimation by the reasons that observed data are equally weighted and useless past data are accumulated. Consequently, the estimates never follow the variation of parameters, even if they vary quite slowly with time.

In order to remove the useless past data automatically at the on-line estimation, we assign the different relative weight to the input-output data of different age. For the convenience of the practical computation, it is preferable to select such sum of the weighted squares as

$$J(k)=\sum_{l=-\infty}^{k}\left\{y(l)-\underline{u}(l)'\Delta\underline{h}\right\}^{2}w^{k-l} \tag{6}$$

where w is the weight restricted by $0<w<1$. Here it is noticed that the lower limit of Σ is negative infinite, but the number of considerable data is changed according to the amount of the weight.

The least-squares estimates are given by

$$\partial J(k)/\partial\Delta h(l)=0 \qquad (l=0,1,2,\cdots,P-1). \tag{7}$$

From Eq.(7), we can easily derive

$$\underline{A}(k)\,\widehat{\Delta\underline{h}}=\underline{b}(k) \tag{8}$$

where

$$\underline{A}(k)=\sum_{l=-\infty}^{k}\underline{u}(l)\,\underline{u}(l)'\,w^{k-l} \tag{9}$$

is a $P\times P$ matrix whose $i'j$ th element is given by

$$a_{i'j}(k)=\sum_{l=-\infty}^{k}x(l-i')\,x(l-j)\,w^{k-l} \tag{10}$$

and

$$\underline{b}(k)=\sum_{l=-\infty}^{k}\underline{u}(l)\,y(l)\,w^{k-l} \tag{11}$$

is a P dimensional vector whose i'th element is given by

$$b_{i'}(k)=\sum_{l=-\infty}^{k}x(l-i')\,y(l)\,w^{k-l}. \tag{12}$$

The recursive equations suited for the on-line estimation are derived from Eq.(8).

The similar equations are derived by using a weighting matrix[5]. However, there exists one difficulty that the rank of

the weighting matrix is not determined since its elements are lost in some cases depending on the amount of the weight.

3. Algorithm of the on-line estimation

To derive the recursive equations we notice the latest data, $u(k)$ and $y(k)$, in Eq.(8) and put them out of the summations. Then Eq.(8) becames

$$\{ u(k)u(k)' + w A(k-1) \} \Delta\hat{h} = u(k)y(k) + w b(k-1) . \tag{13}$$

Defining the previous estimates at $t=(k-1)T$ as

$$\Delta\hat{h}(k-1) = A(k-1)^{-1} b(k-1) \tag{14}$$

and writing $\Delta\hat{h}$ in Eq.(13) as $\Delta\hat{h}(k)$ accordingly, Eq.(13) is transformed to a set of recursive equations

$$\Delta\hat{h}(k) = \Delta\hat{h}(k-1) + [w + u(k)'A(k-1)^{-1}u(k)]^{-1} A(k-1)^{-1}u(k)[y(k) - u(k)'\Delta\hat{h}(k-1)] \tag{15}$$

$$A(k)^{-1} = \{A(k-1)^{-1} - [w + u'(k)A(k-1)^{-1}u(k)]^{-1} A(k-1)u(k)u(k)'A(k-1)^{-1}\} w^{-1} . \tag{16}$$

The necessary derivations from Eq.(13) to Eq.(15) and (16) are shown in Appendix. Table 1 gives the computer program to solve Eq.(15) and (16), where further modification is done to avoid the useless duplication of matrix operations and generalized inverse[6,7] is introduced to assure the solution in the redundant case.

4. Statistical properties of the estimates

We investigate some statistical properties of the estimates for the case where the noise is independent of the input. First, the unbiasness of the estimates is shown. The estimates $\Delta\hat{h}$ are given by

$$\Delta\hat{h}(k) = A(k)^{-1} b(k) = A(k)^{-1} \sum_{\ell=-\infty}^{k} u(\ell) y(\ell) w^{k-\ell} . \tag{17}$$

Taking the expectation of Eq.(17), we have

$$E\,\Delta\hat{h} = A(k)^{-1} \sum_{\ell=-\infty}^{k} u(\ell) E\, y(\ell) w^{k-\ell} \tag{18}$$

where E denotes the expectation.
From Eq.(3)

$$E\, y(\ell) = u(\ell)'\Delta h + E\, n(\ell) = u(\ell)'\Delta h . \tag{19}$$

Therefore, it is clear that

$$\mathcal{E}\,\Delta\hat{\underline{h}} = \underline{A}(k)^{-1}\sum_{\ell=-\infty}^{k}\underline{u}(\ell)\underline{u}(\ell)'w^{k-\ell}\Delta\underline{h} = \underline{A}(k)^{-1}\underline{A}(k)\Delta\underline{h} = \Delta\underline{h}. \tag{20}$$

The covariance matrix of $\Delta\hat{\underline{h}}$ is given by

$$\mathrm{Cov}\,\Delta\hat{\underline{h}} = \mathcal{E}\{(\Delta\hat{\underline{h}}-\Delta\underline{h})(\Delta\hat{\underline{h}}-\Delta\underline{h})'\}. \tag{21}$$

Multiplying Eq.(19) by $\underline{u}(\ell)$ and taking the weighted sum, we obtain

$$\sum_{\ell=-\infty}^{k}\underline{u}(\ell)\mathcal{E}\,y(\ell)w^{k-\ell} = \sum_{\ell=-\infty}^{k}\underline{u}(\ell)\underline{u}(\ell)'w^{k-\ell}\Delta\underline{h} = \underline{A}(k)\Delta\underline{h}, \tag{22}$$

from which

$$\Delta\underline{h} = \underline{A}(k)^{-1}\sum_{\ell=-\infty}^{k}\underline{u}(\ell)\mathcal{E}\,y(\ell)w^{k-\ell}. \tag{23}$$

Substituting Eq.(17) and (23) into Eq.(21), we have

$$\begin{aligned}\mathrm{Cov}\,\Delta\hat{\underline{h}} &= \mathcal{E}\left\{\underline{A}(k)^{-1}\sum_{i=-\infty}^{k}\underline{u}(i)n(i)w^{k-i}\sum_{j=-\infty}^{k}\underline{u}(j)'n(j)w^{k-j}\underline{A}(k)^{-1}\right\}\\ &= \underline{A}(k)^{-1}\sum_{i,j=-\infty}^{k}\mathcal{E}\{n(i)n(j)\}\underline{u}(i)\underline{u}(j)'w^{2k-i-j}\underline{A}(k)^{-1}\end{aligned} \tag{24}$$

by virtue of the relation $y(\ell)-\mathcal{E}\,y(\ell) = n(\ell)$. If the noise is white and has the variance of σ_n^2, Eq.(24) is reduced to

$$\mathrm{Cov}\,\Delta\hat{\underline{h}} = \underline{A}(k)^{-1}\sum_{\ell=-\infty}^{k}\sigma_n^2\underline{u}(\ell)\underline{u}(\ell)'w^{2(k-\ell)}\underline{A}(k)^{-1}. \tag{25}$$

Here we consider the physical meaning of the weight w. For an illustrative example, $a_{ij}(k)$, the ij th element of matrix $\underline{A}(k)$, is noticed. Eq.(10) is rewritten in a first order difference equation

$$a_{ij}(k) - w\,a_{ij}(k-1) = x(k-i)x(k-j). \tag{26}$$

Refering to the sampled-data system theory, $a_{ij}(k)$ is regarded as the output of a linear system with the pulse transfer function $1/(1-wz^{-1})$, whose input consists of the product of $x(k-i)$ and $x(k-j)$.

When w is close to unity, an approximate relation is

$$a_{ij}(k) = \frac{1}{1-w}\,\varphi_{xx}(i-j) \tag{27}$$

where $\varphi_{xx}(i-j) = \varphi_{xx}[(i-j)T]$ is the auto-correlation function between $x(\ell-i)$ and $x(\ell-j)$. Then, the matrices in Eq.(25) are written approximately as

$$\underline{A}(k) = \frac{1}{1-w}\left[\varphi_{xx}(i-j)\right] \tag{28}$$

and

$$\sum_{\ell=-\infty}^{k} \underline{u}(\ell)\underline{u}(\ell)' w^{2(k-\ell)} = \frac{1}{1-w^2}\left[\varphi_{xx}(i-j)\right] \tag{29}$$

where $[\varphi_{xx}(i-j)]$ denotes $P \times P$ matrix whose ij th element is $\varphi_{xx}(i-j)$.

Substituting Eq.(28) and (29) into Eq.(25), we have

$$\mathrm{Cov}\,\Delta\hat{\underline{h}} = \sigma_n^2 \frac{1-w}{1+w}\left[\varphi_{xx}(i-j)\right]^{-1}. \tag{30}$$

On the other hand, it is known that the estimates given by Eq. (5) have the covariance matrix

$$\mathrm{Cov}\,\Delta\underline{h}^{*} = \sigma_n^2 \frac{1}{N+1}\left[\varphi_{xx}(i-j)\right]^{-1} \tag{31}$$

under the same situation. Comparing Eq.(30) and (31) we may define

$$\frac{1+w}{1-w} = Ne \tag{32}$$

as "the equivalent data length" in the sense that conventional least-squares estimates are expected to have the same covariance, if the number of the observed data is equal to Ne.

5. Extension for the non-regulatory process

Up to the present, it is assumed that the impulse response of the process tends to zero after enough time elapses. Here we consider the case where the impulse response approaches to a non-zero constant value. In this case, the parameters in Eq.(3) have the property that $\Delta h(\ell) = const.$ for $\ell > P$.

If we take the difference of the succeedingly sampled values of the output,

$$\begin{aligned} y(k) - y(k-1) = \Delta h(0)x(k) + &\left[\Delta h(1) - \Delta h(0)\right]x(k-1) \\ &+ \left[\Delta h(2) - \Delta h(1)\right]x(k-2) + \cdots \end{aligned} \tag{33}$$

the coefficient $\Delta h(\ell) - \Delta h(\ell-1)$ becomes zero for $\ell > P$. Therefore, by rewritting

$$y(k) - y(k-1) = \Delta y(k),\quad \Delta h(\ell) - \Delta h(\ell-1) = \Delta^2 h(\ell),\quad n(k) - n(k-1) = \Delta n(k), \tag{34}$$

we have a new equation

$$\Delta y(k) = \underline{u}'(k)\Delta^2\underline{h} + \Delta n(k) \tag{35}$$

which has the same form as Eq.(3) from a viewpoint of the estmation. Some non-regulatory processes can be simply identifi-

able by taking the difference of the output. Practically sufficient high pass filtering will be necessary at the data processing, since the power of the high-frequence noise is increased considerably.

6. Estimation of process parameters varying slowly with time

In the preceding section we clarified that the method of weighted least-squares poses the estimation process for the low pass filter. If process parameters vary so slowly with time that the significant variation does not take place during the estimation, the time behaviour of the estimates is approximated as $(1-w)/(1-wz^{-1})$ where the coefficient $1-w$ in the numerator is introduced to be coincident with the unbiasness of the estimates. Using this approximation, the amount of the estimation errors (bias) may be analyzed qualitatively.

Here, we study the case where process parameters change linearly with time. Assuming the rate of the change of the parameters as (γ % of nominal value)/(sampling interval), its sampled data form becomes $\gamma z^{-1}/(1-z^{-1})^2$. Then the steady state error of the estimates, d , is estimated as

$$d = \lim_{z\to 1}(z-1)\left(1-\frac{1-w}{1-wz^{-1}}\right)\frac{\gamma z^{-1}}{(1-z^{-1})^2}$$

$$= \gamma w/(1-w) \qquad (36)$$

by applying the final-value theorem.

Several experimental runs are made on the computer to ascertain whether or not the estimates perform as expected. The results shown in Fig.1 are typical examples. The simulated system is a unity feedback system including the process with the transfer function $K/s(2s+1)$ in the forward path. The input signal, $r(t)$, to the system is the low-pass filtered random signal with the bandwidth about 0.2 rad/sec. Assuming the gain is varying as $K=0.1(1+0.01t)$ the inditial response of the process is estimated from the observed data which are obtained by sampling $x(t)$ and $y(t)$ with the interval of $T=1$ sec. It is noticed that the interesting process is a unregulatory one, and therefore, the estimation must be performed by using the procedure mentioned in the section 5. The estimated inditial responses are shown in Fig.1(b), where the weight upon the

data is chosen to be $w = 0.96$. They are easily calculated by using the equation

$$\hat{h}(l) = \sum_{i=1}^{l} \Delta h(i) = \sum_{i=1}^{l} \sum_{j=1}^{i} \Delta^2 h(j) .$$

In Fig.1(c), the estimates of a representative parameter, $\Delta^2 h(2)$, are shown. Using Eq.(36), the predicted bias becomes 24 % and this is shown by the dotted line.

7. An example of the adaptive control system

The observation of the preceding section gives the suggestion that the slowly varying process is identifiable by using the method of weighted least-squares. Many schemes for the realization of the adaptive control have been proposed in the literatures, but few of them are realizable without some priori knowledges about the process.

The adaptive control system described here needs no knowledge about the process except that it is approximated to be linear. The block diagram of the system is shown in Fig.2. A computer is used for the identifier, processor, simulator and optimizer. The identifier of the process estimates its dynamics successively with time. The processor which is not always provided analyzes the input signal and feeds the necessary information to the simulator. This arrangement gives the system the dual characteristics of the signal and process adaptive.

As mentioned above we assume that no priori knowledge is available, the most reliable method to make the system optimum seems to be the trial error approach such as expremalizing an assigned index of performance by using optimization techniques. But this approach is seldom applied to the actual systems, since the convergence is rather slow, mainly because of the process dynamics. To avoid this difficulty the simulator is provided, where the correspondence of the adjustable parameters to the index of performance (hereafter abbreviated IP) is examined in a fast time scale.

For finding the set of parameters which gives the extremum values, it is useful to measure the surface slope of the IP. Although there is no assurance that the value obtained is not relative extremum, the probability of the failure is very small in the case where the parameters are not so far apart from the

optimum values. The optimizer plays the role of the above.

A definite plan of the adaptive control is illustrated on the simplest example. We assume that:

1. The input to the system is the random signal, denoted by $r(k)$.
2. The process is a regulatory one where input $x(k)$ and output $y(k)$ is related as

$$y(k) = \Delta h(0)x(k) + \Delta h(1)x(k-1) + \cdots + \Delta h(P-1)x(k-P+1). \tag{38}$$

3. The compensator is an integrator with gain K.

We take the weighted sum of squared error, denoted by $e(k)$, as IP. Then, IP is written as

$$IP = \sum_{\ell=-\infty}^{k} e(\ell)^2 w^{k-\ell}. \tag{39}$$

If the transfer characteristics from the input to the error is represented by a set of parameters, denoted by $\underline{\alpha}$, where

$$\underline{\alpha} = [\alpha(0)\ \alpha(1)\ \cdots\ \alpha(M-1)]', \tag{40}$$

the error can be predicted as

$$e(k) = \underline{r}(k)'\underline{\alpha} \tag{41}$$

where

$$\underline{r}(k)' = [r(k)\ r(k-1)\ \cdots\ r(k-M+1)]'.$$

Substituting Eq.(41) into Eq.(39), we have

$$IP = \underline{\alpha}'\underline{R}(k)\underline{\alpha} \tag{42}$$

where

$$\underline{R}(k) = \sum_{\ell=-\infty}^{k} \underline{r}(\ell)\underline{r}(\ell)' w^{k-\ell} \tag{43}$$

is a $M \times M$ matrix whose ij th element is given by

$$r_{ij}(k) = \sum_{\ell=-\infty}^{k} r(\ell-i)r(\ell-j)w^{k-\ell}. \tag{44}$$

This element is easily calculated by using the recursive equation similar to Eq.(26). The remaining parameters, $\underline{\alpha}$, are approximately determined as the coefficients of the quotient of the polynominal

$$G_{\tau c}(z) = \frac{1}{1 + T G_{\tau c}(z) G_{\tau p}(z)} \tag{45}$$

where

$$G_{\tau c}(z) = \frac{K}{1 - z^{-1}}$$

is the pulse transfer function of the compensator and

$$G_p(z) = \Delta h(0) + \Delta h(1) z^{-1} + \cdots + \Delta h(P-1) z^{-P+1}. \qquad (46)$$

In Fig.2 it is shown in what place these equations are computed. The parameters $\underline{\alpha}$, which describe the dynamics of the system, vary with every change of the adjustable parameters which follow on the instructions from the optimizer. At this step most of $\underline{\alpha}$ (accordingly assigned values of the adjustable parameters) are rejected by refering to IP about the system such as the settling time, overshoot etc, without computing Eq. (42).

The choice of IP and searching procedure are the problem of the future work together with the comparison with other control systems.

8. Conclusion

The method of weighted least-squares is applied to the estimation of the impulse response coefficients (precisely, increments of the inditial response for every sampling interval) of the process. There is no necessity for limiting the process to the regulatory one.

The use of weight upon the data makes way for the identification of slowly varying process, and the result of simulation study provides considerable insight into the properties of the estimation procedure. An adaptive control system applied this procedure is suggested.

References

1. *R.E.Kalman, TRANS.ASME, 80 , P.469, 1958.*
2. *M.J.Levin, IEEE Trans., AC-9 , P.229, 1964.*
3. *M.J.Levin, IRE Trans., CT-7 , P.50, 2960.*
4. *R. C. K. Lee, Optimal Estimation, Identification, and Control, P.103, 1964.*
5. *C.T.Leondes, Advances in Control System, Vol.1, P.297, 1964.*
6. *R.Penrose, Cambridge Phil.Soc., 51 , P.406, 1955.*
7. *R.Penrose, Cambridge Phil.Soc., 52 , P.17, 1956.*

Appendix Derivation of the recursive equations

In Eq.(13)

$$\underline{A}(k) = w\,\underline{A}(k-1) + \underline{u}(k)\underline{u}(k)' \tag{A1}$$

Postmultiplying by $\underline{A}(k-1)^{-1}\underline{u}(k)$

$$\underline{A}(k)\underline{A}(k-1)^{-1}\underline{u}(k) = w\,\underline{u}(k) + \underline{u}(k)\underline{u}(k)'\underline{A}(k-1)^{-1}\underline{u}(k).$$

Premultiplying by $\underline{A}(k)^{-1}$

$$\underline{A}(k-1)^{-1}\underline{u}(k) = \underline{A}(k)^{-1}\underline{u}(k)\,[\,w + \underline{u}(k)'\underline{A}(k-1)^{-1}\underline{u}(k)\,]$$

from which, we have

$$\beta\,\underline{A}(k-1)^{-1}\underline{u}(k)\underline{u}(k)' = \underline{A}(k-1)^{-1}\underline{u}(k)\underline{u}(k)' \tag{A2}$$

where

$$\beta = [\,w + \underline{u}(k)'\underline{A}(k-1)^{-1}\underline{u}(k)\,]^{-1}. \tag{A3}$$

Substituting the relation

$$\underline{A}(k)^{-1}\underline{u}(k)\underline{u}(k)' = I - \underline{A}(k)^{-1}[\underline{A}(k) - \underline{u}(k)\underline{u}(k)'] = I - \underline{A}(k)^{-1}w\,\underline{A}(k-1)$$

into Eq.(A2) and post multiplying $\underline{A}(k-1)^{-1}$, we get

$$\beta\,\underline{A}(k-1)^{-1}\underline{u}(k)\underline{u}(k)'\underline{A}(k-1)^{-1} = \underline{A}(k-1)^{-1} - w\,\underline{A}(k)^{-1}.$$

Therefore

$$\underline{A}(k)^{-1} = [\,\underline{A}(k-1)^{-1} - \beta\,\underline{A}(k-1)^{-1}\underline{u}(k)\underline{u}(k)'\underline{A}(k-1)^{-1}\,]\,w^{-1}. \tag{16}$$

From Eq.(13)

$$\Delta\hat{\underline{b}}(k) = \underline{A}(k)^{-1}\,[\,w + \underline{b}(k-1) - \underline{u}(k)\,y(k)\,]. \tag{A4}$$

Substituting Eq.(16) into Eq.(A4)

$$\begin{aligned}\Delta\hat{\underline{b}}(k-1) = {} & \underline{A}(k-1)^{-1}\underline{b}(k-1) + w^{-1}\underline{A}(k-1)^{-1}\underline{u}(k)\,y(k) \\ & - w^{-1}\beta\,\underline{A}(k-1)^{-1}\underline{u}(k)\,[\,w\,\underline{u}(k)'\underline{A}(k-1)^{-1}\underline{b}(k-1) \\ & - w\,y(k) + w\,y(k) - \underline{u}(k)'\underline{A}(k-1)^{-1}\underline{u}(k)\,y(k)\,].\end{aligned} \tag{A5}$$

Substituting Eq.(14) and Eq.(A3) into Eq.(A5), we have

$$\begin{aligned}\Delta\hat{\underline{b}}(k) = {} & \Delta\hat{\underline{b}}(k-1) + w^{-1}\underline{A}(k-1)^{-1}\underline{u}(k)\,y(k) \\ & + \beta\,\underline{A}(k-1)^{-1}\underline{u}(k)\,[\,y(k) - \underline{u}(k)'\Delta\hat{\underline{b}}(k-1)\,] - w^{-1}\underline{A}(k-1)^{-1}\underline{u}(k)\,y(k).\end{aligned}$$

Therefore

$$\Delta\hat{\underline{b}}(k) = \Delta\hat{\underline{b}}(k-1) + \beta\,\underline{A}(k-1)^{-1}\underline{u}(k)\,[\,y(k) - \underline{u}(k)'\Delta\hat{\underline{b}}(k-1)\,]. \tag{15}$$

Table 1 Computer program

```
ESTIMATION OF PROCESS PARAMETERS BY USING THE METHOD OF
WEIGHTED LEAST-SQUARES

USEAGE
 CALL ESTION (N,X,Y,SOL,WE,AA,AB)
DESCRIPTION OF PARAMETERS
 N-NUMBER OF PARAMETERS
 X-VECTOR OF LENGTH N CONTAINING TIME SERIES OF INPUT DATA
 Y-OUTPUT DATA
 SOL-VECTOR OF LENGTH N CONTAINING ESTIMATES
 WE-WEIGHT UPON DATA RESTRICTED BY 0.0,1.0
 AA-WORK MATRIX(NXN)
 AB-WORK MATRIX(NXN)
REMARK
 SOL,AA,AB MUST BE CLEARED AT INITIALIZATION

      SUBROUTINE ESTION
      DO 10 J=1,N
      D(J)=0.0
      D0 10 I=1,N
   10 D(J)=D(J)+AA(J,I)*X(I)
      P=0.0
      DO 20
   20 P=P+X(I)*D(I)
      P=WE+P
      DO 40 J=1,N
      C(J)=0.0
      DO 30 I=1,N
   30 C(J)=C(J)+X(I)*AB(I,J)
   40 C(J)=X(J)-C(J)
      DO 50 I=1,N
      IF(ABS(C(I)).GT.0.001) GO TO 80
   50 CONTINUE
      DO 60 I=1,N
   60 B(I)=D(I)/P
      DO 70 J=1,N
      DO 70 I=J,N
      AA(I,J)=(AA(I,J)-B(I)*D(J))/WE
   70 AA(J,I)=AA(I,J)
      GO TO 130
   80 S=0.0
      DO 90 I=1,N
   90 S=S+C(I)*C(I)
      DO 100 I=1,N
  100 B(I)=C(I)/S
      DO 110 J=1,N
      DO 110 I=J,N
      AB(I,J)=AB(I,J)+B(I)*C(J)
  110 AB(J,I)=AB(I,J)
      DO 120 J=1,N
      DO 120 I=J,N
      AA(I,J)=(AA(I,J)-B(I)*D(J)-D(I)*B(J)+P*B(I)*B(J))/WE
  120 AA(J,I)=AA(I,J)
  130 S=0.0
      DO 140 I=1,N
  140 S=S+X(I)*SOL(I)
      S=Y-S
      DO 150 I=1,N
  150 SOL(I)=SOL(I)+S*B(I)
      RETURN
      END
```

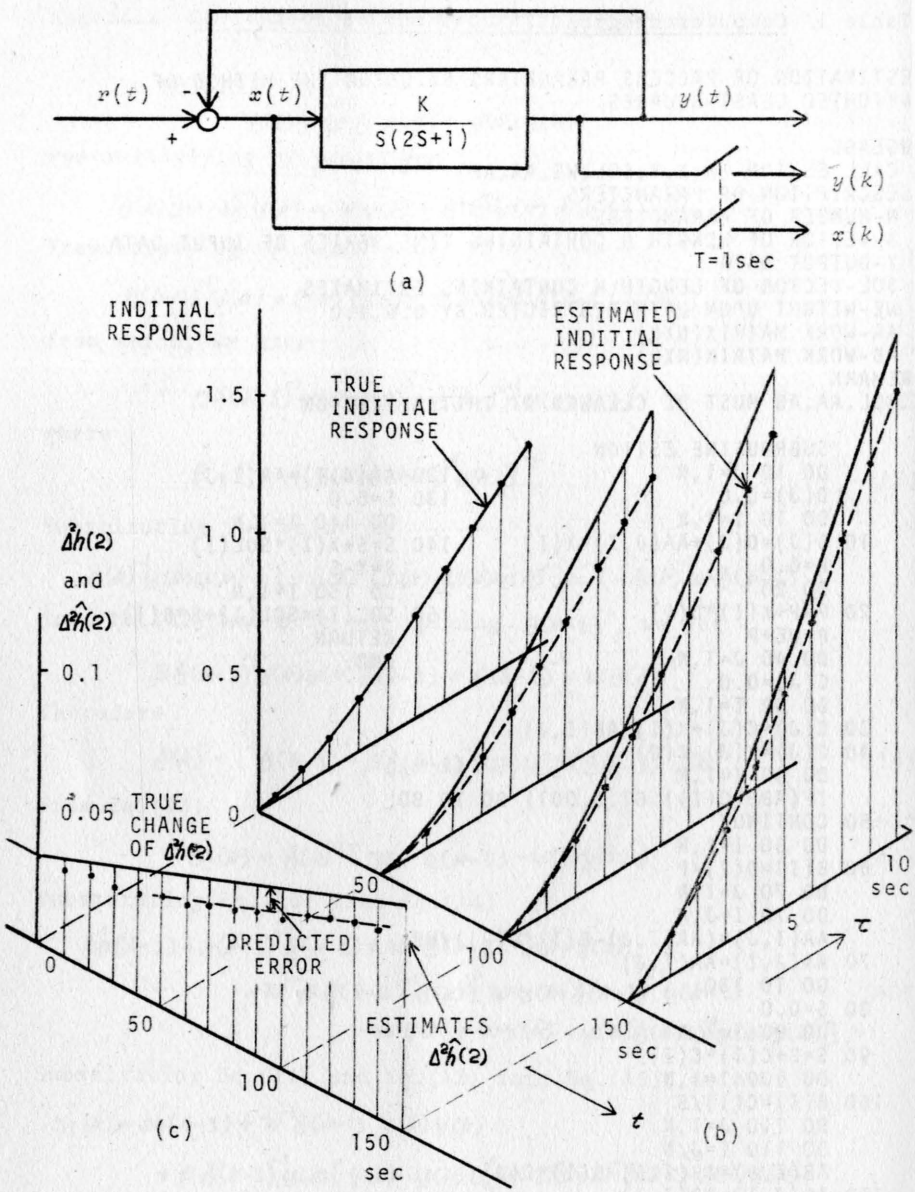


Fig.1 Experimental results.

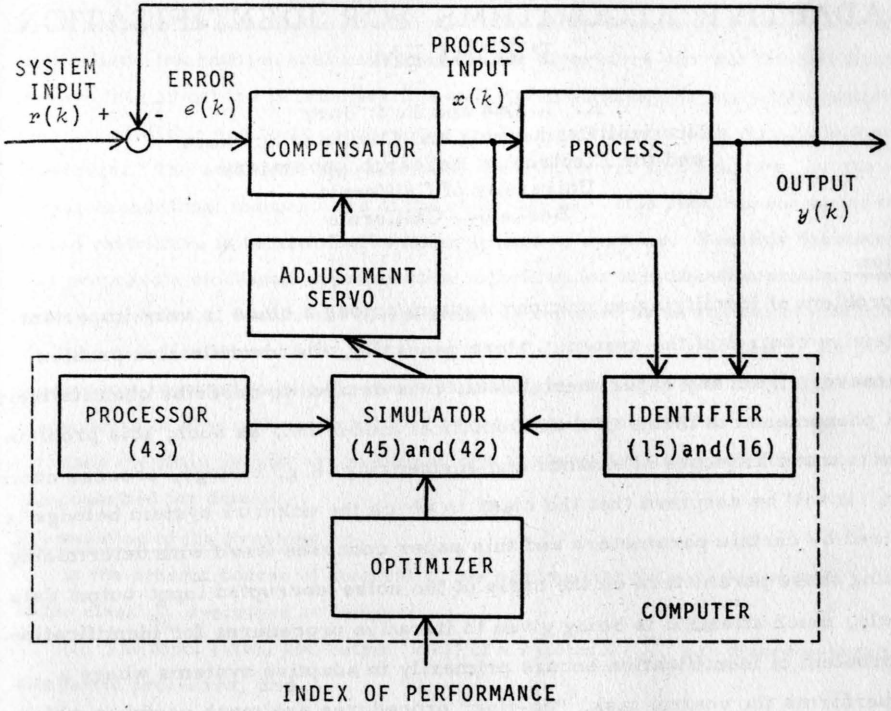


Fig.2 Block diagram of adaptive control system.

ADAPTIVE ALGORITHMS FOR IDENTIFICATION PROBLEM

K. G. Oza and E. I. Jury
Electrical Engineering and Computer Sciences
and the Electronics Research Laboratory
University of California
Berkeley, California

Introduction

The problem of identifying an unknown system among a class is very important for the adaptive control of the system. More generally, the identification problem arises whenever, from any experimental data, it is desired to describe quantitatively a physical phenomenon in terms of a mathematical model and, as such, this problem is common to many branches of science and engineering, e.g. biology, process control, aerospace. It will be assumed that the class to which the unknown system belongs is characterized by certain parameters and this paper concerns itself with determining or estimating these parameters on the basis of the noise-corrupted input-output data.

Recently, much attention is being given to iterative procedures for identification. Since the problem of identification occurs primarily in adaptive systems where a computer performs the control task, "on-line" procedures are much useful in which the estimates are updated as soon as new data becomes available. Moreover, by the very nature of the adaptive systems, any practical on-line procedure is supposed to assume very little or no knowledge about the probabilistic structure of the input, output and noise processes. These requirements have led the control system theorists to methods of stochastic approximation and other iterative procedures which do not require much a priori knowledge. The principle of random contraction mappings has been used to this end.[1, 2] For estimating the impulse response of a discrete system, Kushner[3] describes a simple iterative procedure and Nagumo and Noda[4] have proposed a so-called "learning method" with a random input and noise-free measurements. Recently, Oza[5] has shown that the latter method is identical to Kushner's and also that the convergence can be obtained using the random contraction principle. A procedure based on stochastic approximation was first discussed by Ho and Whalen[6] and later studied in much detail by Ho and Lee.[7] Stochastic approximation is also used for the nonlinear identification problem by Kirvaitis and Fu[8] and for general adaptive systems by Tsypkin.[9]

In the present paper, the authors propose on-line adaptive algorithms for identification of a linear discrete system and extend the same to the case where the linear

system is preceded by a power-series type nonlinearity. The degree to which the term "adaptive" is justified in any iterative procedure depends upon the amount of probabilistic information presumed known. Our algorithms warrant the qualification because they adjust the parameters in a well-defined manner as more information becomes available and their convergence can be established with a very little a priori knowledge. The stochastic approximation algorithms are also adaptive, but these require conditional independence of the observations. This requirement seems to be too restrictive to be satisfied in many dynamical systems. Recently Sakrison[10] has proposed a stochastic approximation algorithm for identification problem where the requirement of conditional independence is replaced by an equivalent condition on the prediction error of the processes involved. In the present paper, the processes are required to be wide-sense stationary up to order four and a finite-time dependence of observations is allowed by the proposed algorithm.[1]

Only the main results are presented due to space limitation and reference 2 should be consulted for details.

Formulation of the Problem

In the present course of discussion, we assume that the unknown system belongs to the class $\mathcal{S}_L$ described as follows:

(1) The input $\{u(k)\}$ and output $\{x(k)\}$ of a system $S \in \mathcal{S}_L$ are discrete-parameter stochastic processes, and

(2) They are related by a stable, linear, constant-coefficient stochastic difference equation

$$x(k) + \sum_{i=1}^{n} \alpha_i x(k - i) = \sum_{j=0}^{n} \alpha_{j+n+1} u(k - j) \quad . \tag{1}$$

Clearly, the system $S \in \mathcal{S}_L$ is characterized by a $(2n + 1)$-dimensional vector

$$\underline{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n, \alpha_{n+1}, \ldots, \alpha_{2n+1}) \tag{2}$$

and so it is appropriate to denote the system by $S(\underline{\alpha})$. Further assumptions will be made in the sequel about the probability structure of the processes $\{u(k)\}$ and $\{x(k)\}$.

If the system under identification is characterized by the equation

$$x(k) + \sum_{i=1}^{n} a_i x(k - i) = \sum_{j=0}^{n} a_{j+n+1} u(k - j) \quad , \tag{3}$$

then the problem of identification consists in determining or estimating the vector

$$\underline{a} = (a_1, a_2, \cdots, a_n, a_{n+1}, \cdots, a_{2n+1})' \quad . \tag{4}$$

Let us define a vector process $\underline{q}(k)$ by

$$\underline{q}(k) = (x(k-1), \ldots, x(k-n), -u(k), \ldots, -u(k-n))' \tag{5}$$

Then (3) can be rewritten as

$$x(k) = -\langle \underline{q}(k), a \rangle \quad . \tag{6}$$

Let $v(k)$ and $y(k)$ be the noise-corrupted input and output data, respectively, i.e.,

$$v(k) = u(k) + n_1(k) \tag{7}$$

$$y(k) = x(k) + n_2(k) \tag{8}$$

where $\{n_1(k)\}$ and $\{n_2(k)\}$ are noise processes. These need not be either independent or Gaussian; nevertheless, in order to simplify certain algebraic manipulations, we shall assume that $\{u(k)\}$, $\{n_1(k)\}$ and $\{n_2(k)\}$ are mutually uncorrelated, zero-mean processes with finite variances. Define a vector process $\underline{p}(k)$ analogous to $\underline{q}(k)$:

$$\underline{p}(k) = (y(k-1), \ldots, y(k-n), -v(k), \ldots, -v(k-n))' \tag{9}$$

Clearly, when observations are noise-free, we have

$$\underline{p}(k) = \underline{q}(k) \, , \tag{10}$$

and

$$y(k) = x(k) \, . \tag{11}$$

Let us assume that the given system is $S(\underline{\alpha})$ with the input $\{v(k)\}$ where $\underline{\alpha}$ is generic at present. The output sequence $\{\eta(k)\}$ of the system $S(\underline{\alpha})$ will be given by the equation

$$\eta(k) + \sum_{i=1}^{n} \alpha_i \eta(k-i) = \sum_{j=0}^{n} \alpha_{j+n+1} v(k-j) \quad . \tag{12}$$

Definition: The explicit error $e_e(k)$ is defined as the difference between the observed output $y(k)$ and the computed output $\eta(k)$, i.e.,

$$e_e(k) = y(k) - \eta(k) \, . \tag{13}$$

Since the system $S(\underline{\alpha})$ uses its own past output to compute the present output $\eta(k)$ and since it looks in the past as far back as n instants of time, we are motivated to define the virtual or implicit error $e_i(k)$ as the weighted sum of the explicit errors $e_e(k - j)$, $j = 0, 1, \ldots, n$. The weights in this sum are chosen such that eventually a positive definite quadratic form $J(\underline{\beta})$ with a unique minimum is obtained, where $\underline{\beta}$ is the identification error defined as the difference between the vectors $\underline{\alpha}$ and $\underline{a}$, i.e.,

$$\underline{\beta} = \underline{\alpha} - \underline{a} \tag{14}$$

Definition: The implicit error $e_i(k)$ is defined as

$$e_i(k) = e_e(k) + \sum_{i=1}^{n} \alpha_i e_e(k - i) \tag{15}$$

where $\alpha_1, \ldots, \alpha_n$ are first n components of $\underline{\alpha}$.

From Equations (7), (8), (12), (13) and (15) and from the definition of the system $S(\underline{a})$, we have

$$e_i(k) = (\underline{\alpha} - \underline{a})'\underline{q}(k) + n_2(k) + \underline{\alpha}'\underline{n}(k) \tag{16}$$

$$= (\underline{\alpha} - \underline{a})'\underline{p}(k) + n_2(k) + \underline{a}'\underline{n}(k) \tag{17}$$

and consequently we get

$$E\{e_i^2(k)\} = (\underline{\alpha} - \underline{a})'R_{\underline{q}}(\underline{\alpha} - \underline{a}) + m_1(\underline{\alpha}) + m_2(\underline{\alpha}) \tag{18}$$

where

$$m_1(\underline{\alpha}) = E\left\{\left[\sum_{j=0}^{n} \alpha_{j+n+1} n_1(k - j)\right]^2\right\} \tag{19}$$

and

$$m_2(\underline{\alpha}) = E\left\{\left[n_2(k) + \sum_{i=1}^{n} \alpha_i n_2(k - i)\right]^2\right\} . \tag{20}$$

Denoting the first quadratic form in (18) by $J(\underline{\beta})$, we have

$$J(\underline{\beta}) = E\{e_i^2(k)\} - m_1(\underline{\alpha}) - m_2(\underline{\alpha}) \tag{21}$$

and the following lemma:

Lemma 1: If the spectral density of the process $\{u(k)\}$ is non-zero at least on a set of positive measure in $(-\pi, \pi)$, the matrix $R_{\underline{q}}$ is positive definite and consequently $J(\underline{\beta})$ has a unique minimum at $\underline{\beta} = \underline{0}$ or $\underline{\alpha} = \underline{a}$.

The minimum of $J(\underline{\beta})$ can be found by taking the gradient of J with respect to $\underline{\alpha}$ and equating it to zero. Assuming that the gradient of $e_i^2(k)$ w.r.t. $\underline{\alpha}$ exists, an interchange of the order of differentiation and expectation becomes possible and we obtain

$$\nabla_{\underline{\alpha}} J(\underline{\beta}) = E\{\text{grad}_{\underline{\alpha}}\, e_i^2(k)\} - \text{grad}_{\underline{\alpha}}\, (m_1(\underline{\alpha}) + m_2(\underline{\alpha})) = \underline{0} \tag{22}$$

Expanding expressions (19) and (20), we get

$$m_1(\underline{\alpha}) + m_2(\underline{\alpha}) = \sum_{j_1}\sum_{j_2} \alpha_{j_1+n+1}\alpha_{j_2+n+1}\, E\{n_1(k-j_1)n_1(k-j_2)\}$$

$$+ \sum_{i_1}\sum_{i_2} \alpha_{i_1}\alpha_{i_2}\, E\{n_2(k-i_1)n_2(k-i_2)\}$$

$$+ 2\sum_{i} \alpha_i\, E\{n_2(k)n_2(k-i)\} + E\{n_2^2(k)\}$$

$$= \underline{\alpha}' R_{\underline{n}} \underline{\alpha} + 2\underline{r}_{\underline{n}}'\underline{\alpha} + R_{n_2}(0) \tag{23}$$ *

and hence

$$\text{grad}_{\underline{\alpha}}(m_1(\underline{\alpha}) + m_2(\underline{\alpha})) = 2R_{\underline{n}}\underline{\alpha} + 2\underline{r}_{\underline{n}} \tag{24}$$

Now using equation (17), we find that

$$\text{grad}_{\underline{\alpha}}\, e_i^2(k) = 2e_i(k) \cdot \nabla_{\underline{\alpha}} e_i(k)$$

$$= 2e_i(k)\underline{p}(k). \tag{25}$$

Now substitution of (24) and (25) in (22) yields

---

* $R_{\underline{n}} = E\{\underline{n}(k)\underline{n}(k)'\} = E\{(\underline{p}(k)-\underline{q}(k))\,(\underline{p}(k)-\underline{q}(k))'\}$

$r_{\underline{n}} = E\{n_2(k)\underline{n}(k)\}$

$$E\{e_i(k)\underline{p}(k)\} - R_{\underline{n}}\underline{\alpha} - \underline{r}_{\underline{n}} = \underline{0} \tag{26}$$

Upon using the definition of $S(\underline{a})$ and carrying out some simplification after substituting (17) into (26), we get

$$(R_{\underline{p}} - R_{\underline{n}})\underline{\alpha} + \underline{r}_{\underline{p}} - \underline{r}_{\underline{n}} = \underline{0}\,. \tag{27}^{\dagger}$$

The solution of equation (27) is clearly $\underline{\alpha} = \underline{a}$. Since the covariance matrix $R_{\underline{p}}$ and the covariance vector $\underline{r}_{\underline{p}}$ in (27) depend on the unknown vector $\underline{a}$ and also on the probability distribution of the process $\{u(k)\}$, solving that equation for $\underline{\alpha}$ is not a routine matter. This motivates us to look for an iterative scheme to solve (27) using sample covariance functions.

Finally we assume that the autocorrelation functions of the noise-processes $\{n_1(k)\}$ and $\{n_2(k)\}$ are known. This implies that $R_{\underline{n}}$ and $\underline{r}_{\underline{n}}$ are known quantities. It is in order to mention here that the assumption about $\{n_1(k)\}$ and $\{n_2(k)\}$ being uncorrelated can be eliminated if their cross-correlation functions are assumed to be known. This will only change the form of the matrix $R_{\underline{n}}$ and the vector $\underline{r}_{\underline{n}}$ and will not affect any of the derived results.

$^{\dagger}R_{\underline{p}} = E\{\underline{p}(k)\underline{p}(k)'\}$, $\underline{r}_{\underline{p}} = E\{y(k)\underline{p}(k)\}$.

Algorithms

Observations on $\{v(k)\}$ and $\{y(k)\}$ for $k = 1, \ldots, N$ will be used to estimate elements of $R_{\underline{p}}$ and $\underline{r}_{\underline{p}}$. Let $\hat{R}_{\underline{p}}(N)$ be the estimate of $R_{\underline{p}}$ at time N whose jth column is

$$\begin{bmatrix} \frac{1}{N}\sum_{k=n}^{N} y(k-i)y(k-1) \\ \vdots \\ \frac{1}{N}\sum_{k=n}^{N} y(k-j)y(k-n) \\ -\frac{1}{N}\sum_{k=n}^{N} y(k-j)v(k) \\ \vdots \\ -\frac{1}{N}\sum_{k=n}^{N} y(k-j)v(k-n) \end{bmatrix} \quad (28)$$

and

$$\begin{bmatrix} -\frac{1}{N}\sum_{k=n}^{N} v(k-j_1)y(k-1) \\ \vdots \\ -\frac{1}{N}\sum_{k=n}^{N} v(k-j_1)y(k-n) \\ \frac{1}{N}\sum_{k=n}^{N} v(k-j_1)v(k) \\ \vdots \\ \frac{1}{N}\sum_{k=n}^{N} v(k-j_1)v(k-n) \end{bmatrix} \quad (29)$$

for $j = 1, 2, \ldots, n$

for $j = n+1, \ldots, 2n+1$ with $j_1 = j - n - 1$. Similarly let

$$\rho_{\underline{p}}(N) = \begin{bmatrix} \frac{1}{N}\sum_{k=n}^{N} y(k)y(k-1) \\ \vdots \\ \frac{1}{N}\sum_{k=n}^{N} y(k)y(k-n) \\ -\frac{1}{N}\sum_{k=n}^{N} y(k)v(k) \\ \vdots \\ -\frac{1}{N}\sum_{k=n}^{N} y(k)v(k-n) \end{bmatrix} \tag{30}$$

be the estimate of $\underline{r}_{\underline{p}}$ at time N. It should be noted that replacing N in the denominator of the estimates in (28), (29) and (30) by (N - n + 1) will yield unbiased estimates and we shall do this when necessary.

Now we state assumptions on the processes $\{u(k)\}$, $\{n_1(k)\}$ and $\{n_2(k)\}$ and on the system $S(\underline{a})$ and some of their immediate consequences. Certain strengthenings will be made as the need arises. Our immediate interest lies in the strong consistency of the estimates just mentioned.

Assumptions:

(1) The input and noise processes are wide sense stationary through order four, i.e., for all $t = 0, \pm 1, \pm 2, \ldots,$

$$R_s(k) = E\,[s(t+k)s(t)] = R_s(-k) \tag{31}$$

and

$$Q_s(k_1, k_2, k_3) = E[s(t)s(t+k_1)s(t+k_2)s(t+k_3)] \tag{32}$$

are all independent of t, where $s = u, n_1, n_2$ and k, k_1, k_2, k_3 range over all positive and negative integers.

(2) The variances $\sigma_s^2 = R_s(0)$, $s = u, n_1, n_2$ are finite.

(3) All fourth-order cumulant and cross-cumulant functions of $\{u(k)\}$, $\{n_1(k)\}$ and $\{n_2(k)\}$ are finite.

(4) The covariance sequences (31) and the fourth-order cumulant function (32) are summable, i.e.,

$$\sum_{k=-\infty}^{\infty} |R_s(k)| < \infty \tag{33}$$

and

$$\sum |Q_s(k_1, k_2, k_3)| < \infty \tag{34}$$

where $s = u, n_1, n_2$ and the latter sum extends over all negative and positive integral values of k_1, k_2, k_3.

(5) Let $\{h_k\}$ be the impulse-response sequence for the system $S(\underline{a})$. There exists a finite positive constant C such that

$$\sum_{k=-\infty}^{\infty} |h_k| \leq C < \infty . \tag{35}$$

The output process $\{x(k)\}$ also satisfies Assumptions (1)-(4) if the input process $\{u(k)\}$ does so and if Assumption (5) holds. Then, under Assumptions (1)-(5), the estimates in (28), (29) and (30) converge to their corresponding correlation functions with probability one[11, 12]. We conclude this property of the strong consistency in the following lemma:

Lemma 2: Under Assumptions (1)-(5), we have

$$P\left[\lim_{N\to\infty} \|\mathcal{R}_{\underline{p}}(N) - R_{\underline{p}}\| = 0\right] = 1 \tag{36}$$

and

$$P\left[\lim_{N\to\infty} \|\rho_{\underline{p}}(N) - \underline{r}_{\underline{p}}\| = 0\right] = 1 . \tag{37}$$

Now we use estimates of correlation functions in connection with the technique of successive substitutions to stochastically approximate the vector $\underline{a}$ defining the system $S(\underline{a})$. To begin, the mapping

$$T_0(\underline{\alpha}) = \underline{\alpha} - \gamma[(R_{\underline{p}} - R_{\underline{n}})\underline{\alpha} + (\underline{r}_{\underline{p}} - \underline{r}_{\underline{n}})] \tag{38}$$

from E^{2n+1} into itself has the unique fixed point $\underline{a}$, for any real number $\gamma \neq 0$, by virtue of Equation (27). If γ is chosen such that

$$\|I - \gamma(R_{\underline{p}} - R_{\underline{n}})\| \leq C < 1 \tag{39}$$

where C is arbitrary otherwise, then T_0 is a contraction mapping. Let λ_1 and λ_{2n+1} be the minimum and maximum eigenvalues of $R_{\underline{q}}$; then the condition

$$0 < |1 - \gamma\lambda_1| < C < 1 , \quad 0 < |1 - \gamma\lambda_{2n+1}| < C < 1 \tag{40}$$

is equivalent to the condition (39).

Since the exact form of the mapping T_0 is not available partly because of the unknown vector $\underline{a}$ and partly because of the unknown probability distribution of the process $\{u(k)\}$, we cannot use the usual contraction algorithm

$$\underline{\alpha}_{N+1} = T_0(\underline{\alpha}_N) ,$$

initialized by an arbitrary point $\underline{\alpha}_1 \in E^{2n+1}$, which converges to $\underline{a}$ as $N \to \infty$. Hence, we are led to consider the following sequence of random mappings:

$$T_N(\underline{\alpha}) = \underline{\alpha} - \gamma[(\mathcal{R}_{\underline{p}}(N) - R_{\underline{n}})\underline{\alpha} + (\underline{\rho}_{\underline{p}}(N) - \underline{r}_{\underline{n}})] \tag{41}$$

for $N = n, n+1, \ldots$. Again with γ satisfying (39) or (40), consider the iterative algorithm

$$\underline{\alpha}_{N+1} = T_N(\underline{\alpha}_N) \tag{42}$$

where the initializing point $\underline{\alpha}_1$ is arbitrary. The sequence $\{\underline{\alpha}_N\}$ generated by (42) is a sequence of random vectors and the following theorem establishes its convergence with probability one.

Theorem 1: If Condition (39) or (40) is satisfied, then, under Assumptions (1)-(5), the sequence of random vectors $\{\underline{\alpha}_N\}$ generated by the algorithm (42), i.e.,

$$\underline{\alpha}_{N+1} = \underline{\alpha}_N - \gamma[(\mathcal{R}_{\underline{p}}(N) - R_{\underline{n}})\underline{\alpha}_N + (\underline{\rho}_{\underline{p}}(N) - \underline{r}_{\underline{n}})] , \tag{43}$$

converges to the vector $\underline{a}$ with probability one.

The proof of Theorem 1 is given in the Appendix.

The algorithm (43) is adaptive in the sense that the estimate $\underline{\alpha}_N$ is being updated to $\underline{\alpha}_{N+1}$ on the basis of new information available at the Nth instant of time. But the gain coefficient γ has to satisfy condition (39) or (40) which calls for some a priori knowledge. The algorithm can be made more adaptive by replacing γ by $\frac{H}{N}$ where H is an arbitrary positive real number:

$$\underline{\alpha}_{N+1} = \underline{\alpha}_N - \frac{H}{N}\left[(\mathcal{R}_{\underline{p}}(N) - R_{\underline{n}})\underline{\alpha}_N + (\underline{\rho}_{\underline{p}}(N) - \underline{r}_{\underline{n}})\right] . \tag{44}$$

We have the following theorem for the algorithm (44).

Theorem 2: If the processes $\{v(k)\}$ and $\{y(k)\}$ possess moments of all orders, then, under Assumptions (1)-(5), the algorithm (44) converges to $\underline{a}$ with probability one.[2]

The algorithms (43) and (44) are both computationally simple. Additional computational simplicity is offered by the following recursive relationships for $\mathcal{R}_{\underline{p}}(N)$ and $\underline{\rho}_{\underline{p}}(N)$[2]:

$$\mathcal{R}_{\underline{p}}(N) = \mathcal{R}_{\underline{p}}(N-1) + \frac{1}{N}\left[\underline{p}(N)\underline{p}(N)' - \mathcal{R}_{\underline{p}}(N-1)\right] \tag{45}$$

$$\underline{\rho}_{\underline{p}}(N) = \underline{\rho}_{\underline{p}}(N-1) + \frac{1}{N}\left[(y(N)\underline{p}(N) - \underline{\rho}_{\underline{p}}(N-1)\right] \tag{46}$$

for $N = n+1, n+2, \ldots$. Thus we can eliminate the need to store all the data up to time N and require to store only the current values of the quantities $y(N)$, $\underline{p}(N)$, $\underline{\rho}_{\underline{p}}(N-1)$ and $\mathcal{R}_{\underline{p}}(N-1)$ for the adaptive algorithms. It is shown in Reference (2) that (45) and (46) represent stochastic approximation procedures under certain alternate, but equivalent conditions on the random processes involved.

Application to A Class of Nonlinear Systems

The adaptive algorithms (43) and (44) are, in fact, applicable to many situations where the parameters to be estimated satisfy a set of regression equations such as (27). In particular, an immediate extension is straight-forward for the problem of identifying a system among a class of nonlinear systems.

We consider a class $\mathcal{S}_N$ of systems each of which consists of a zero-memory nonlinearity $f[\cdot]$ followed by a linear time-invariant system whose impluse response is $h(t)$ [Fig. 1]. Many nonlinear control systems belong to this class. The identification problem among this class has been considered by Narendra and Gallman[12] who propose an interesting iterative method and show the convergence in specific examples, but in their publication do not offer any theoretical proof of convergence. Let $H(z)$ be the z-transform of the impulse response $h(t)$ and suppose that $H(z)$ is of the following form:

$$H(z) = \frac{\sum_{j=0}^{n} a_{j+n+1} z^{-j}}{1 + \sum_{i=1}^{n} a_i z^{-i}} \tag{47}$$

Assuming that the zero-memory nonlinearity f[·] is a power-series type, the output of the nonlinear part is given by

$$f[u(k)] = \sum_{i=1}^{\ell} b_i u^i(k) \tag{48}$$

The identification problem now consists in determining (or estimating) the parameter vectors

$$\underline{a} = (a_1, a_2, \ldots, a_n, a_{n+1}, \ldots, a_{2n+1})' \tag{49}$$

and

$$\underline{b} = (b_1, \ldots, b_\ell) \tag{50}$$

from the observations on the system input and output.

We assume that the input u(k) is a discrete-parameter random process which is observable without being corrupted by any noise. The output x(k) is observable only through an additive noise n(k) [Fig. 2] so that a process

$$y(k) = x(k) + n(k) \tag{51}$$

is available for computation. With reference to Fig. 2, we can write the following relationship:

$$x(k) + \sum_{i=1}^{n} a_i x(k - i) = \sum_{j=0}^{n} a_{j+n+1} \sum_{r=1}^{\ell} b_r u^r(k - j) \quad . \tag{52}$$

Denoting generic values of the estimates of a_i and b_j by α_i and β_j, respectively, for $i = 1, 2, \ldots, 2n+1$ and $j = 1, 2, \ldots, \ell$, the estimated output y(k) will be given by equation (53):

$$\eta(k) + \sum_{i=1}^{n} \alpha_i \eta(k - i) = \sum_{j=0}^{n} \alpha_{j+n+1} \sum_{r=1}^{\ell} \beta_r u^r(k - j) \tag{53}$$

Defining $b_{r,j} = a_{j+n+1} b_r$ and $\beta_{r,j} = \alpha_{j+n+1} \beta_r$, a system $S \in \mathcal{S}_N^{\circ}$ described by the input-output relation (52) is characterized by a vector $\underline{\theta}$ defined as

$$\underline{\theta} = (a_1 \ \cdots \ a_n \ b_{10} \ \cdots \ b_{1n} \ \cdots \ b_{\ell 0} \ \cdots \ b_{\ell n})' \tag{54}$$

and the class $\mathcal{S}_N^{\circ}$ can be identified with a generic vector $\underline{Y}$:

$$\underline{Y} = (\alpha_1 \ \cdots \ \alpha_n \ \beta_{10} \ \cdots \ \beta_{1n} \ \cdots \ \beta_{\ell 0} \ \cdots \ \beta_{\ell n})' \ . \tag{55}$$

Redefine the vectors $\underline{q}(k)$, $\underline{p}(k)$ and $\underline{n}(k)$ as follows:

$$\underline{q}(k) = (x(k-1) \ldots x(k-n) - u(k) \ldots -u(k-n) \ldots -u^{\ell}(k) \ldots -u^{\ell}(k-n)) \tag{56}$$

$$\underline{p}(k) = (y(k-1) \ldots y(k-m) - u(k) \ldots -u(k-n) \ldots -u^{\ell}(k) \ldots -u^{\ell}(k-n)) \tag{57}$$

and

$$\underline{n}(k) = (n(k-1) \ldots n(k-n)\ 0 \ldots 0 \ldots 0 \ldots 0) \tag{58}$$

Then, repeating the steps of deriving (27), we can obtain

$$(R_{\underline{p}} - R_{\underline{n}})\,\underline{\gamma} + (\underline{r}_{\underline{p}} - \underline{r}_{\underline{n}}) = \underline{0} \tag{59}$$

where $R_{\underline{p}}$ and $R_{\underline{n}}$ are correlation matrices of (57) and (58), respectively, and $\underline{r}_{\underline{p}}$ and $\underline{r}_{\underline{n}}$ are corresponding correlation vectors.

Again observations on $\{u(k)\}$ and $\{y(k)\}$ for $k = 1, \ldots, N$ can be used to estimate elements of $R_{\underline{p}}$ and $\underline{r}_{\underline{p}}$, but these estimates will be apparently more complicated than those for the linear case. Specifically, jth column of the estimate $\mathcal{R}_{\underline{p}}(N)$ will be of the following form:

$$\begin{bmatrix}
\frac{1}{N}\sum_{k=n}^{N} y(k-j)y(k-1) \\
\vdots \\
\frac{1}{N}\sum_{k=n}^{N} y(k-j)y(k-n) \\
-\frac{1}{N}\sum_{k=n}^{N} y(k-j)u(k) \\
\vdots \\
-\frac{1}{N}\sum_{k=n}^{N} y(k-j)u(k-n) \\
-\frac{1}{N}\sum_{k=n}^{N} y(k-j)u^2(k) \\
\vdots \\
-\frac{1}{N}\sum_{k=n}^{N} y(k-j)u^2(k-n) \\
\vdots \\
-\frac{1}{N}\sum_{k=n}^{N} y(k-j)u^{\ell}(k) \\
\vdots \\
-\frac{1}{N}\sum_{k=n}^{N} y(k-j)u^{\ell}(k-n)
\end{bmatrix} \tag{60}$$

for $j = 1, 2, \ldots, n$ and

$$\begin{bmatrix} -\frac{1}{N}\sum_{k=n}^{N} u(k-j_1)y(k-1) \\ \vdots \\ -\frac{1}{N}\sum_{k=n}^{N} u(k-j_1)y(k-n) \\ \frac{1}{N}\sum_{k=n}^{N} u(k-j_1)u(k) \\ \vdots \\ \frac{1}{N}\sum_{k=n}^{N} u(k-j_1)u(k-n) \\ \frac{1}{N}\sum_{k=n}^{N} u(k-j_1)u^2(k) \\ \vdots \\ \frac{1}{N}\sum_{k=n}^{N} u(k-j_1)u^2(k-n) \\ \vdots \\ \frac{1}{N}\sum_{k=n}^{N} u(k-j_1)u^{\ell}(k) \\ \vdots \\ \frac{1}{N}\sum_{k=n}^{N} u(k-j_1)u^{\ell}(k-n) \end{bmatrix} \tag{61}$$

for $j = n+1, \ldots, 2n+1$ with $j_1 = j - n - 1$. Let $\underline{\rho}_{\underline{p}}(N)$ be the estimate of $\underline{r}_{\underline{p}}$ and $\underline{\rho}_{\underline{p}}(N)$ can be obtained from (60) by substituting $j = 0$.

If the processes $\{u(k)\}$ and $\{y(k)\}$ are ergodic, then each element of $\underline{R}_{\underline{p}}(N)$ and $\underline{\rho}_{\underline{p}}(N)$ converges with probability one to the corresponding element of $R_{\underline{p}}$ and $\underline{r}_{\underline{p}}$. Actually, such convergence requires less stringent conditions than ergodicity: namely, Assumptions (1)-(4) be satisfied for $s = u^2, u^3, \ldots, u^\ell$. Then, (36) and (37) hold for this case also. The algorithms corresponding to (43) and (44) are given by the following:

$$\underline{Y}(N+1) = \underline{Y}(N) - \gamma[(\underline{R}_{\underline{p}}(N) - R_{\underline{n}})\underline{Y}(N) + (\underline{\rho}_{\underline{p}}(N) - \underline{r}_{\underline{n}})] \tag{62}$$

and

$$\underline{Y}(N+1) = \underline{Y}(N) - \frac{H}{N}\left[(\underline{R}_{\underline{p}}(N) - R_{\underline{n}})\underline{Y}(N) + (\underline{\rho}_{\underline{p}}(N) - \underline{r}_{\underline{n}})\right] \tag{63}$$

where γ has to satisfy the condition

$$\|I - \gamma(R_{\underline{p}} - R_{\underline{n}})\| < 1 \quad , \tag{64}$$

H is any positive constant, and $R_{\underline{n}}, \underline{r}_{\underline{n}}$ are given by (65) and (66),

$$R_{\underline{n}} = \begin{bmatrix} \varphi(0) & \varphi(1) \ldots \varphi(n-1) & 0 \ldots 0 \\ \varphi(1) & \varphi(0) \ldots \varphi(n-2) & 0 \ldots 0 \\ \vdots & \vdots & \vdots \\ \varphi(n-1) & \varphi(0) & 0 \ldots 0 \\ 0 & \ldots \ldots 0 & 0 \ldots 0 \\ \vdots & \vdots & \vdots \\ 0 & \ldots \ldots 0 & 0 \ldots 0 \end{bmatrix} \tag{65}$$

and

$$\underline{r}_{\underline{n}} = \begin{bmatrix} \varphi(1) \\ \cdot \\ \cdot \\ \cdot \\ \varphi(n) \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} \tag{66}$$

and $\varphi(k) = E[n(t)n(t + k)]$.

The sequences of random vectors generated by (62) and (63) converge to $\underline{\theta}$ with probability one. The recursive relationships (45) and (46) remain valid for this case also. The rate of convergence of the algorithms (63) and (46) depend on the gain constant H and the fastest rate[2] is 1/N.

Computer Results

The results of computer studies are given in the following examples. The computer studies were made on IBM 7090/7094. The input process $\{u(k)\}$ and the noise processes $\{n_1(k)\}$, $\{n_2(k)\}$ and $\{n(k)\}$ were generated by a subroutine "$IBMAP DEV" which is available in the Computer Center Library of the University of California, Berkeley, and which yields independent, normal deviates with zero mean and σ^2 variance whenever the function subprogram σ^* RANDEV ( ) appears in the main program. The covariance function of the processes involved were determined by varying values σ.

Example 1:

Consider the first-order, linear, time-invariant discrete system

$$x(k) + 0.5x(k - 1) = u(k) . \tag{67}$$

The parameter vector

$$\underline{a} = \begin{pmatrix} -0.5 \\ 1.0 \end{pmatrix} \tag{68}$$

completely characterizes the system (67), if the initial state and the input are specified; so the identification problem consists in estimating the vector $\underline{a}$. In this case, we have

$$\underline{g}(k) = \begin{pmatrix} x(k-1) \\ \cdot \\ u(k) \end{pmatrix} \tag{69}$$

and

$$\underline{p}(k) = \begin{pmatrix} x(k-1) + n_2(k-1) \\ \\ u(k) + n_1(k) \end{pmatrix} = \begin{pmatrix} p_1(k) \\ \\ p_2(k) \end{pmatrix} \tag{70}$$

and

$$y(k) = x(k) + n_2(k) \ . \tag{71}$$

The values of σ for generating $u(k)$, $n_1(k)$ and $n_2(k)$ were, respectively, 1.00, 0.1 and 0.1. Hence the matrix $R_{\underline{n}}$ is given by

$$R_{\underline{n}} = \begin{pmatrix} 0.01 & 0 \\ \\ 0 & 0.01 \end{pmatrix} \tag{72}$$

and the vector $\underline{r}_{\underline{n}}$ is zero.

The algorithm (44) takes the following form in this case:

$$\alpha_1(N+1) = \alpha_1(N) - \frac{H}{N}\left[\rho_{11}(N)\alpha_1(N) + \rho_{12}(N)\alpha_2(N) - 0.01\,\alpha_1(N) + \rho_1(N)\right] \tag{73}$$

$$\alpha_2(N+1) = \alpha_2(N) - \frac{H}{N}\left[\rho_{21}(N)\alpha_1(N) + \rho_{22}(N)\alpha_2(N) - 0.01\,\alpha_2(N) + \rho_2(N)\right] \tag{74}$$

where ρ_{ij}'s are components of $\mathcal{R}_{\underline{p}}(N)$ and ρ_i's are components of $\underline{\rho}_{\underline{p}}(N)$, i.e.,

$$\mathcal{R}_{\underline{p}}(N) = \begin{pmatrix} \rho_{11}(N) & \rho_{12}(N) \\ \\ \rho_{21}(N) & \rho_{22}(N) \end{pmatrix} \tag{75}$$

and

$$\underline{\rho}_{\underline{p}}(N) = \begin{pmatrix} \rho_1(N) \\ \\ \rho_2(N) \end{pmatrix} \tag{76}$$

The recursive relationships for (75) and (76) take the following forms:

$$\begin{pmatrix} \rho_{11}(N+1) & \rho_{12}(N+1) \\ \rho_{21}(N+1) & \rho_{22}(N+1) \end{pmatrix} = \begin{pmatrix} \rho_{11}(N) & \rho_{12}(N) \\ \rho_{21}(N) & \rho_{22}(N) \end{pmatrix}$$

$$+ \frac{1}{N}\left[\begin{pmatrix} p_1^2(N) & p_1(N)p_2(N) \\ p_1(N)p_2(N) & p_2^2(N) \end{pmatrix} - \begin{pmatrix} \rho_{11}(N) & \rho_{12}(N) \\ \rho_{21}(N) & \rho_{22}(N) \end{pmatrix} \right] \tag{77}$$

and

$$\begin{pmatrix} \rho_1(N+1) \\ \rho_2(N+1) \end{pmatrix} = \begin{pmatrix} \rho_1(N) \\ \rho_1(N) \end{pmatrix} + \frac{1}{N}\left[\begin{pmatrix} y(N)p_1(N) \\ y(N)p_2(N) \end{pmatrix} - \begin{pmatrix} \rho_1(N) \\ \rho_2(N) \end{pmatrix} \right] \tag{78}$$

The simulation of the system (67) and of the algorithms (73), (74), (77) and (78) was carried out in the same program, thus imitating the situation of real-time identification. Initial values $\alpha_1(1)$ and $\alpha_2(1)$ are taken quite arbitrary and far from the actual values a_1 and a_2; nevertheless, the distance of $\underline{\alpha}(100)$ from $\underline{a}$ is found to be approximately 0.0022 when H is 3.5. The results for H = 3.5, $\alpha_1(1) = 0.5$ and $\alpha_2(1) = -1.00$ are given graphically in Figs. 3a and 3b.

Fig. 4 shows the results with H = 3.5, $\alpha_1(1) = 0.5$ and $\alpha_2(1) = -1.00$, but the noise is not present. This plot does not considerably differ from the plot in Fig. 3; thus the algorithm (44) has so-called noise-immunity if the noise-variance is known.

Example 2:

Next consider the fourth-order linear difference equation

$$x(k) - x(k-1) + 0.18x(k-2) - 0.784x(k-3) + 0.656x(k-4) = u(k) \tag{79}$$

which represents a sampled-data booster control system having the z-transfer function

$$H(z) = \frac{z^4}{(z^2 - 1.8z + 0.82)(z^2 + 0.8z + 0.8)} \quad . \tag{80}$$

A similar example has been considered in Reference 7 where Kalman filtering theory is used for the identification.

We have

$$\underline{a} = \begin{pmatrix} - 1.000 \\ + 0.180 \\ - 0.784 \\ + 0.656 \\ + 1.000 \end{pmatrix} \tag{81}$$

and

$$\underline{p}(k) = \begin{pmatrix} x(k - 1) \\ x(k - 2) \\ x(k - 3) \\ x(k - 4) \\ - u(k) \end{pmatrix} \tag{82}$$

The algorithm (44) takes the following form in this case:

$$\alpha_i(N+1) = \alpha_i(N) - \frac{H}{N}\left[\rho_{i1}(N)\alpha_1(N) + \rho_{i2}(N)\alpha_2(N)\right. \tag{83}$$

$$\left. + \rho_{i3}(N)\alpha_3(N) + \rho_{i4}(N)\alpha_4(N) + \rho_{i5}(N)\alpha_5(N) + \rho_i(N)\right] , \; i = 1, 2, 3, 4, 5$$

where

$$\rho_{ij}(N+1) = \rho_{ij}(N) + \frac{1}{N}\left[p_i(N)p_j(N) - \rho_{ij}(N)\right] \tag{84}$$

and

$$\rho_k(N+1) = \rho_k(N) + \frac{1}{N}\left[x(N)p_k(N) - \rho_k(N)\right] , \tag{85}$$

i, j, k = 1, 2, 3, 4, 5.

In input process was generated with the value $\sigma = 0.5$. Figures 4 and 5 illustrate the variations in the estimates when $H = 7.0$ and the noise variance is 0.04.

For a comparison with a similar approach[7], we define the normalized error of estimation as follows:

$$\text{error} = \frac{\| \underline{a} - \underline{\alpha}(N) \|^2}{\| \underline{a} - \underline{\alpha}(1) \|^2}$$

With the same initial estimates as in Ref. 7, we found that the error, after 600 iterations, is 0.0003 which is considerably less than the minimum normalized error 0.001 obtained there.

Example 3:

Next consider a system shown in Fig. 2 with $n = 2$, $\ell = 3$ and $a_j = 0$ for $j = 3, 5$ and $b_i = 0$ for $i = 1, 2$. Then, let the input-output relationship of the complete system be described by the following nonlinear difference equation

$$x(k) - 0.272x(k-1) + 0.0185x(k-2) = 0.544u^3(k-1) \qquad (87)$$

Here we have

$$\underline{\theta} = \begin{bmatrix} -0.2720 \\ 0.0185 \\ 0.0000 \\ 0.0000 \\ 0.0000 \\ 0.0000 \\ 0.0000 \\ 0.5440 \end{bmatrix} \qquad (88)$$

$$\underline{p}(k) = \begin{bmatrix} y(k-1) \\ y(k-2) \\ -u(k) \\ -u(k-1) \\ -u^2(k) \\ -u^2(k-1) \\ -u^3(k) \\ -u^3(k-1) \end{bmatrix} \tag{89}$$

Table 1 presents the results of using the algorithm (63) for different values of the gain parameter H when there is no output noise and Table 2 presents the same results when there is putput noise with a variance 0.01. Again a very good convergence is obtained in both cases.

Conclusion

The identification of parameters of a linear discrete system is considered in this paper. The identification problem is shown to be reduced to solving a set of linear regression equations. Iterative procedures (algorithms) are presented to solve these equations by successive approximation. The principle of random construction mapping is used to establish the convergence of the algorithms. Some recursive relationships are established in order to reduce the required computer capacity for these algorithms.

Extension of the identification procedure to certain nonlinear discrete systems are also presented. Several examples studied in the computer are given to illustrate the identification procedure and to show convergence.

REFERENCES

(1) Oza, K. G., and Jury, E. I., "System Identification and the Principle of Random Contraction Mapping," SIAM J. on Control, Vol. 6, No. 2, 1968.

(2) Oza, Kandarp, Identification Problem and Random Contraction Mappings, Ph. D. Dissertation, University of California, Berkeley, California, December 1967.

(3) Kushner, H. J., "A Simple Iterative Procedure for the Identification of the Unknown Parameters of a Linear Time-varying Discrete System," Trans. ASME J. of Basic Engineering, June 1963, pp. 227-235.

(4) Nagumo, J., and Noda, A., "A Learning Method for System Identification," IEEE Trans. on Automatic Control, Vol. AC-12, No. 3, June, 1967, pp. 282-286.

(5) Oza, K. G., "Iterative Procedure of Nagumo-Noda and Random Contraction Principle," private notes.

(6) Ho, Y. C., and Whalen, B., "An Approach to the Identification and Control of Linear Dynamical Systems with Unknown Parameters," IEEE Trans. on Automatic Control, Vol. AC-8, July 1963, pp. 255-256.

(7) Ho, Y. C., and Lee, R. C. K., "Identification of Linear Dynamic Systems," Information and Control, Vol. 8, 1965, pp. 93-110.

(8) Kirvaitis, K., and Fu, K. S., "Identification of Nonlinear Systems by Stochastic Approximation," Proceedings of JACC, June 1966, pp. 255-264.

(9) Tsypkin, Ya. Z., "Adaptation, Learning and Self-Learning in Automatic Systems," Automatika e Telemekhanika, Vol. 26, No. 1, January 1966, pp. 23-61.

(10) Sakrison, D. J., "Use of Stochastic Approximation to Solve the System Identification Problem," IEEE Trans. on Automatic Control, AC-12 (1967), pp. 563-567.

(11) Parzen, E., "An Approach to Time-Series Analysis," Ann. Math. Stat., Vol. 32, 1961, pp. 951-988.

(12) Doob, J. L., Stochastic Processes, John Wiley and Sons, New York, 1953.

(13) Narendra, K. S., and Gallman, P. G., "An Iterative Method for the Identification of Nonlinear Systems Using a Hammerstein Model," Proceedings JACC, June 1966, pp. 634-638.

APPENDIX

Proof of Theorem 1: Since almost uniform convergence and convergence with probability one are equivalent with respect to a probability measure, it suffices to show that, for arbitrary $\epsilon > 0$, $\delta > 0$ there exists a set A with

$$PA < \delta \tag{A-1}$$

such that

$$\|\underline{\hat{a}}_N - \underline{a}\| < \epsilon \tag{A-2}$$

holds for every $\omega \in A_\delta^C$ and for all N exceeding some finite positive integer N_1 which is independent of ω.

Let us choose ϵ so small that $\epsilon + \epsilon^2 < 1 - C$. Condition (39) makes this possible. From Egoroff's theorem and Theorem 1, for any positive numbers ϵ_1, ϵ_2 and δ, there exists a set $A \in \mathcal{A}$ satisfying (A-1) such that

$$\|\underline{\rho}_p(N) - \underline{r}_p\| < \epsilon_1 \quad \|\mathcal{R}_p(N) - R_p\| < \epsilon_2 \tag{A-3}$$

both hold for every $\omega \in A_\delta^C$ and for all N exceeding some finite index N_2 independent of ω. Choosing ϵ_1 and ϵ_2 small such that

$$\max[\gamma\epsilon_2, \gamma(\epsilon_1 + \epsilon_2\|\underline{a}\|)] < \epsilon^2 \quad , \tag{A-4}$$

we can say that

$$\|\gamma(\mathcal{R}_p(N) - R_p)\| < \epsilon^2 \tag{A-5}$$

and

$$\|\gamma[(\mathcal{R}_p(N) - R_p)\underline{a} + (\underline{\rho}_p(N) - \underline{r}_p)]\| < \epsilon^2 \tag{A-6}$$

both hold for every $\omega \in A_\delta^C$ and for all N exceeding N_2.

Now subtracting $\underline{a}$ from both sides of (43) and using the identity

$$\underline{a} - \gamma[(R_{\underline{p}} - R_{\underline{n}})\underline{a} + (\underline{r}_{\underline{p}} - \underline{r}_{\underline{n}})] = \underline{a} \quad , \tag{A-7}$$

we get

$$\hat{\underline{\beta}}_{N+1} = [I - \gamma(\mathcal{R}_{\underline{p}}(N) - R_{\underline{n}})]\hat{\underline{\beta}}_N + \underline{\psi}_N(\underline{a}) \tag{A-8}$$

where

$$\hat{\underline{\beta}}_N = \hat{\underline{\alpha}}_N - \underline{a} \tag{A-9}$$

and

$$\underline{\psi}_N(\underline{a}) = \gamma[(R_{\underline{p}} - \mathcal{R}_{\underline{p}}(N))\underline{a} + (\underline{r}_{\underline{p}} - \underline{\rho}_{\underline{p}}(N))] \quad . \tag{A-10}$$

Assume that, for every integer $N \geq N_2$,

$$\|\hat{\underline{\beta}}_N\| > \epsilon \quad . \tag{A-11}$$

In particular, for some integer $K > N_2$, we have

$$\begin{aligned} \|\hat{\underline{\beta}}_{K+1}\| &\leq \|I - \gamma(\mathcal{R}_{\underline{p}}(K) - R_{\underline{n}})\| \cdot \|\hat{\underline{\beta}}_K\| + \|\underline{\psi}_K(\underline{a})\| \\ &\leq (C + \epsilon^2)\|\hat{\underline{\beta}}_K\| + \epsilon^2 \\ &\leq (C + \epsilon^2 + \epsilon)\|\hat{\underline{\beta}}_K\| \end{aligned} \tag{A-12}$$

where the second inequality is obtained by using (39) and (A-5). Since $\epsilon^2 + \epsilon < 1 - C$, upon iterating (A-12), we can find an integer μ such that

$$\|\hat{\underline{\beta}}_{K+\mu}\| \leq (C + \epsilon^2 + \epsilon)^{\mu}\|\hat{\underline{\beta}}_K\| < \epsilon \quad . \tag{A-13}$$

This contradicts the hypothesis (A-11); hence there must exist at least one integer $N_3 > N_2$ such that

$$\|\hat{\underline{\beta}}_{N_3}\| < \epsilon \tag{A-14}$$

But then using the second inequality of (A-12)

$$\|\hat{\beta}_{N_3+1}\| \leq (C + \varepsilon^2)\|\hat{\beta}_{N_3}\| + \varepsilon^2 \leq (C + \varepsilon^2 + \varepsilon)\varepsilon < \varepsilon \tag{A-15}$$

and similarly

$$\|\hat{\beta}_N\| < \varepsilon \quad \forall N \geq N_3 \tag{A-16}$$

Since (A-3) and consequently (A-5) and (A-6) hold for every $\omega \in A_\delta^C$, (A-16) also holds for every $\omega \in A_\delta^C$. Furthermore, N_3 does not depend on ω because neither N_2 nor C, ϵ do. So $\hat{\beta}_N$ converges to zero almost uniformly.

Q. E. D.

Table 1

IDENTIFICATION OF A NONLINEAR SYSTEM OF CLASS $\mathcal{P}_n$ BY RANDOM CONTRACTION ALGORITHM.....

ACTUAL VALUES OF PARAMETERS

LINEAR PART	A1	A2	A3
	-0.2720	0.0185	0.5440
NONLINEAR PART	B1	B2	B3
	0.0000	0.0000	1.0000

ESTIMATED VALUES OF PARAMETERS

H=5.0

N	A1(N)	A2(N)	B10(N)	B11(N)	B20(N)	B21(N)	B30(N)	B31(N)
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100	-0.2724	0.0185	-0.0269	0.0882	0.0048	-0.0063	0.0082	0.5254
200	-0.2714	0.0182	-0.0110	0.0305	0.0003	-0.0009	0.0025	0.5369
300	-0.2717	0.0182	-0.0067	0.0167	0.0005	-0.0007	0.0016	0.5400
400	-0.2718	0.0184	-0.0047	0.0109	0.0002	-0.0004	0.0011	0.5413
500	-0.2719	0.0185	-0.0034	0.0078	0.0001	-0.0002	0.0008	0.5422

H=7.0

N	A1(N)	A2(N)	B10(N)	B11(N)	B20(N)	B21(N)	B30(N)	B31(N)
1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100	-0.2722	0.0185	-0.0156	0.0444	0.0028	-0.0035	0.0046	0.5346
200	-0.2718	0.0184	-0.0042	0.0099	0.0001	-0.0003	0.0010	0.5417
300	-0.2719	0.0184	-0.0020	0.0042	0.0001	-0.0002	0.0005	0.5430
400	-0.2719	0.0185	-0.0012	0.0023	0.0000	-0.0001	0.0003	0.5434
500	-0.2720	0.0185	-0.0008	0.0015	0.0000	-0.0000	0.0002	0.5437

H=9.0

N	A1(N)	A2(N)	B10(N)	B11(N)	B20(N)	B21(N)	B30(N)	B31(N)
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100	-0.2721	0.0185	-0.0074	0.0176	0.0012	-0.0015	0.0021	0.5403
200	-0.2719	0.0185	-0.0012	0.0025	0.0000	-0.0001	0.0003	0.5434
300	-0.2720	0.0185	-0.0005	0.0008	0.0000	-0.0000	0.0001	0.5438
400	-0.2720	0.0185	-0.0002	0.0004	0.0000	-0.0000	0.0001	0.5439
500	-0.2720	0.0185	-0.0001	0.0002	0.0000	-0.0000	0.0000	0.5440

H=11.0

N	A1(N)	A2(N)	B10(N)	B11(N)	B20(N)	B21(N)	B30(N)	B31(N)
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100	-0.2720	0.0185	-0.0019	0.0050	0.0004	-0.0004	0.0002	0.5432
200	-0.2720	0.0185	-0.0002	0.0004	0.0000	-0.0000	0.0000	0.5439
300	-0.2720	0.0185	-0.0000	0.0001	0.0000	-0.0000	0.0000	0.5440
400	-0.2720	0.0185	-0.0000	0.0000	-0.0000	-0.0000	0.0000	0.5440
500	-0.2720	0.0185	0.0000	0.0000	-0.0000	0.0000	-0.0000	0.5440

H=13.0

N	A1(N)	A2(N)	B10(N)	B11(N)	B20(N)	B21(N)	B30(N)	B31(N)
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100	-0.2770	0.0279	0.0099	-0.0454	0.0009	0.0218	0.0178	0.3961
200	-0.2720	0.0185	0.0003	-0.0005	-0.0000	0.0000	-0.0001	0.5441
300	-0.2720	0.0185	0.0001	-0.0001	-0.0000	0.0000	-0.0000	0.5440
400	-0.2720	0.0185	0.0000	-0.0001	-0.0000	0.0000	-0.0000	0.5440
500	-0.2720	0.0185	0.0000	-0.0000	-0.0000	0.0000	-0.0000	0.5440

Table 2

IDENTIFICATION OF A NONLINEAR SYSTEM OF CLASS S_n BY RANDOM CONTRACTION ALGORITHM IN PRESENCE OF OUTPUT NOISE
WHITE GAUSSIAN NOISE HAS A VARIANCE EQUAL TO 0.01

ACTUAL VALUES OF PARAMETERS

LINEAR PART	A1	A2	A3
	-0.2720	0.0185	0.5440
NONLINEAR PART	B1	B2	B3
	0.0000	0.0000	1.0000

ESTIMATED VALUES OF PARAMETERS

H=5.00

N	A1(N)	A2(N)	B10(N)	B11(N)	B20(N)	B21(N)	B30(N)	B31(N)
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100	-0.2736	0.0197	-0.0430	0.0835	0.0031	-0.0060	0.0115	0.5268
200	-0.2714	0.0210	-0.0201	0.0277	0.0039	-0.0008	0.0035	0.5364
300	-0.2740	0.0191	-0.0170	0.0128	0.0060	-0.0035	0.0048	0.5389
400	-0.2759	0.0188	-0.0137	0.0073	0.0060	-0.0040	0.0051	0.5405
500	-0.2752	0.0181	-0.0116	0.0052	0.0045	-0.0034	0.0037	0.5423

H=7.0

N	A1(N)	A2(N)	B10(N)	B11(N)	B20(N)	B21(N)	B30(N)	B31(N)
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100	-0.2734	0.0197	-0.0367	0.0379	0.0016	-0.0034	0.0088	0.5364
200	-0.2718	0.0211	-0.0129	0.0061	0.0038	-0.0005	0.0018	0.5413
300	-0.2742	0.0195	-0.0124	-0.0005	0.0059	-0.0032	0.0038	0.5421
400	-0.2762	0.0189	-0.0000	-0.0019	0.0058	-0.0038	0.0043	0.5428
500	-0.2752	0.0180	-0.0083	-0.0014	0.0042	-0.0030	0.0029	0.5439

H=9.0

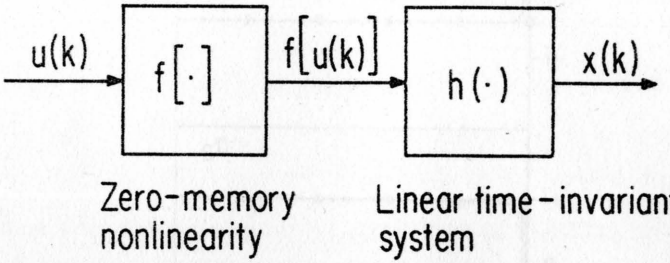
N	A1(N)	A2(N)	B10(N)	B11(N)	B20(N)	B21(N)	B30(N)	B31(N)
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100	-0.2732	0.0198	-0.0303	0.0082	0.0003	-0.0013	0.0064	0.5426
200	-0.2720	0.0212	-0.0088	-0.0020	0.0036	-0.0004	0.0008	0.5431
300	-0.2743	0.0197	-0.0107	-0.0043	0.0059	-0.0031	0.0034	0.5430
400	-0.2763	0.0189	-0.0085	-0.0041	0.0057	-0.0038	0.0040	0.5433
500	-0.2752	0.0180	-0.0071	-0.0025	0.0041	-0.0029	0.0026	0.5442

H=11.0

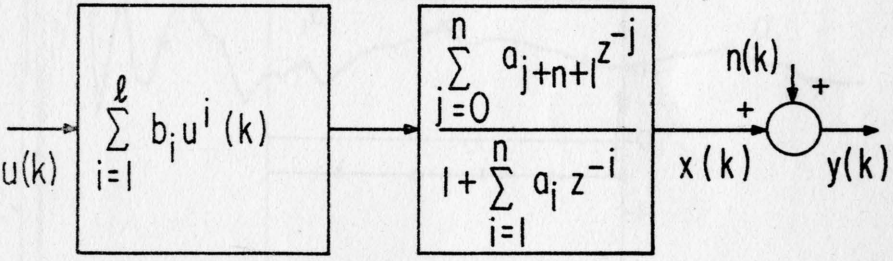
N	A1(N)	A2(N)	B10(N)	B11(N)	B20(N)	B21(N)	B30(N)	B31(N)
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100	-0.2730	0.0196	-0.0248	-0.0065	-0.0003	-0.0006	0.0031	0.5502
200	-0.2721	0.0213	-0.0068	-0.0043	0.0035	-0.0005	0.0003	0.5437
300	-0.2744	0.0198	-0.0103	-0.0052	0.0060	-0.0032	0.0033	0.5432
400	-0.2763	0.0190	-0.0080	-0.0044	0.0056	-0.0038	0.0039	0.5434
500	-0.2751	0.0180	-0.0067	-0.0024	0.0040	-0.0028	0.0025	0.5442

H=13.0

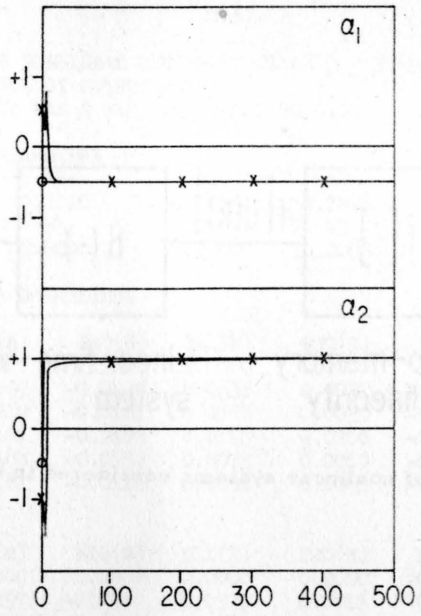
N	A1(N)	A2(N)	B10(N)	B11(N)	B20(N)	B21(N)	B30(N)	B31(N)
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100	-0.2700	0.0132	-0.0264	-0.0012	-0.0009	-0.0137	-0.0215	0.6554
200	-0.2721	0.0213	-0.0059	-0.0056	0.0035	-0.0005	0.0001	0.5439
300	-0.2744	0.0199	-0.0104	-0.0054	0.0061	-0.0032	0.0034	0.5432
400	-0.2763	0.0190	-0.0077	-0.0044	0.0055	-0.0038	0.0038	0.5434
500	-0.2751	0.0180	-0.0064	-0.0022	0.0039	-0.0028	0.0024	0.5442



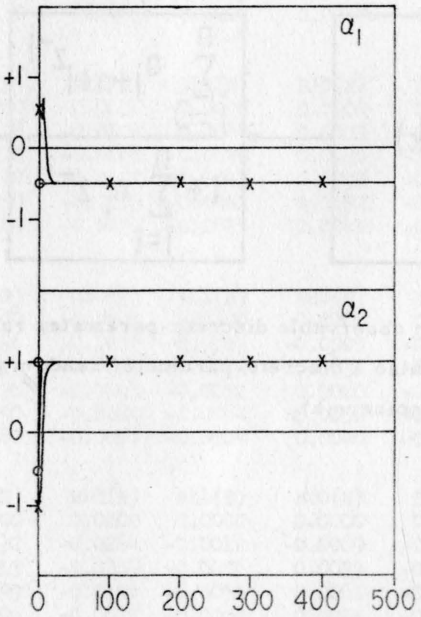
1. The type of nonlinear systems considered in this paper.



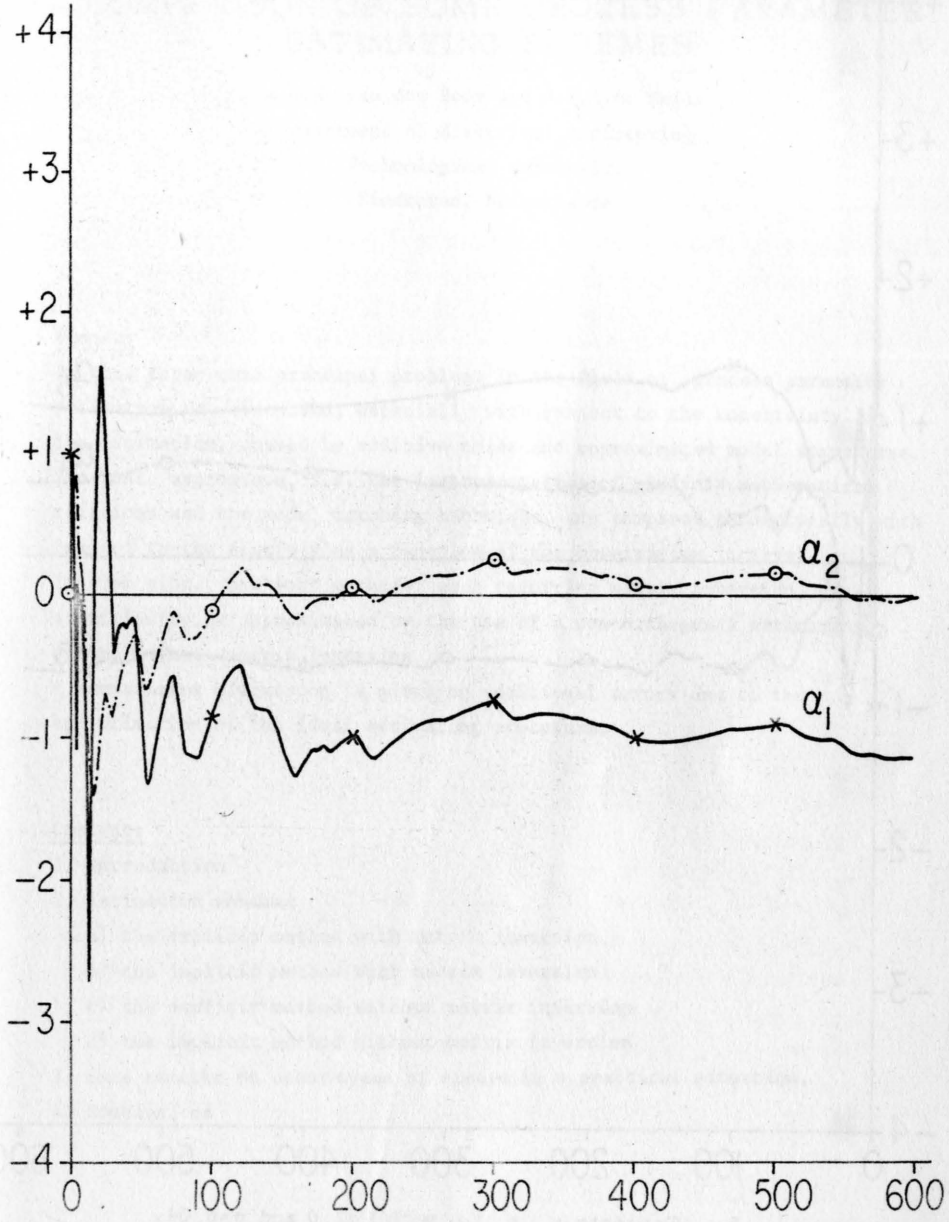
2. Input u(k) is an observable discrete-parameter random process; output y(k) is also a discrete-parameter random process corrupted by an additive noise n(k).



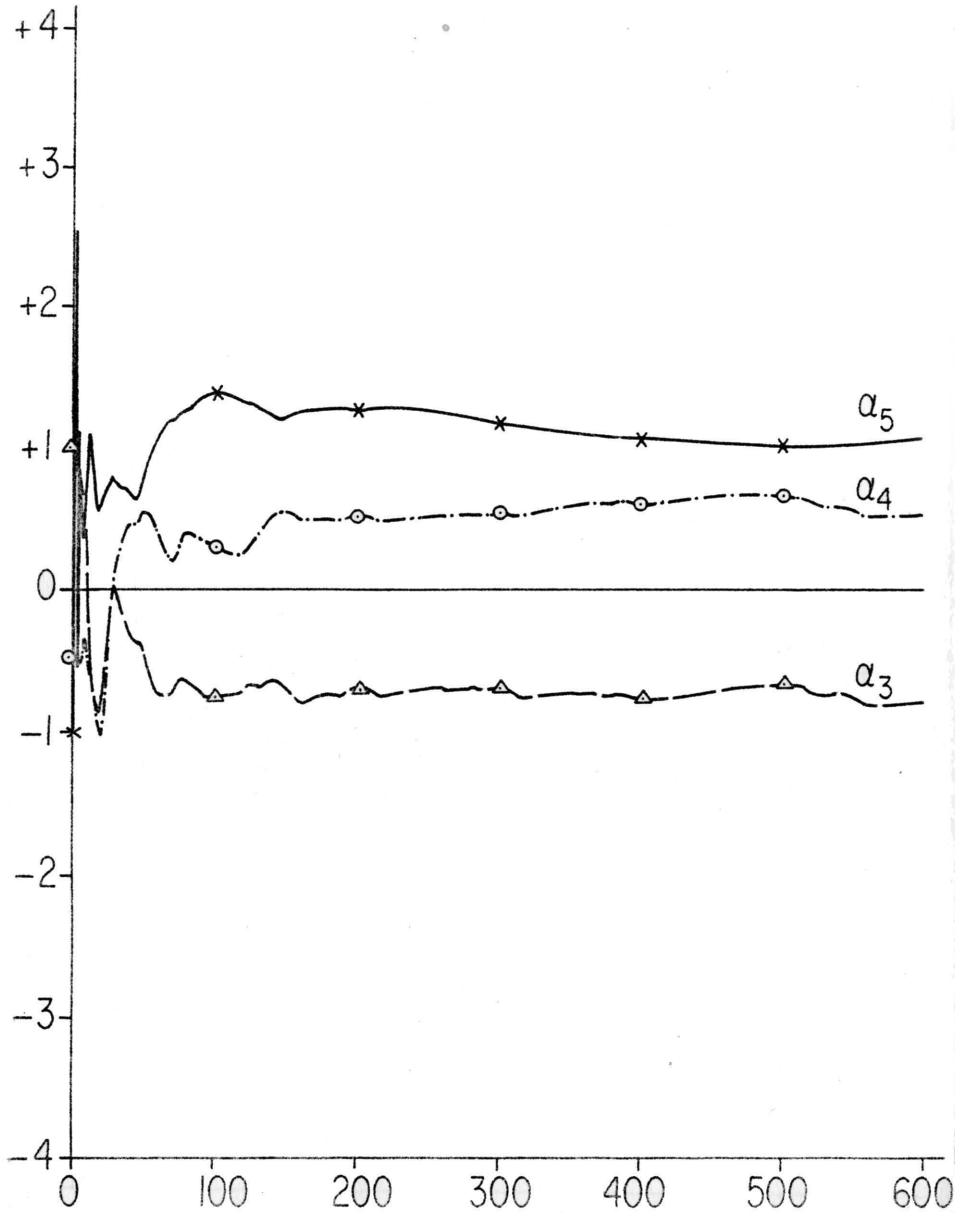
3a. Results with H=3.5, $\alpha_1(1)=.5$, $\alpha_2(1)=1.00$ with noise present.



3b. Results with H=3.5, $\alpha_1(1)=.5$, $\alpha_2(1)=-1.00$, but noise is not present.



4. Estimate α_1, α_2 with gain parameter H=7.0 and noise variance σ=0.04.



5. Estimate α_3, α_4, α_5 with H=7.0 and σ=0.04.

A COMPARISON OF SOME PROCESS PARAMETER ESTIMATING SCHEMES

A.J.W. van den Boom and J.H.A.M. Melis
Department of Electrical Engineering
Technological University
Eindhoven, Netherlands

Summary

In this paper some principal problems in the field of process parameter estimation are discussed, especially with respect to the uncertainty in the estimation, caused by additive noise and approximated model structures. Two basic approaches, i.e. the instrumentation of explicit mathematical relations and the model matching technique, are compared theoretically with respect to the accuracy as a function of the observation interval.
In some situations both methods, each requiring matrix inversion, can sufficiently be approximated by the use of a non-orthogonal estimating scheme without matrix inversion.
A summarizing discussion is given on additional errors due to the approximation of the ideal estimating procedure.

Contents

1. Introduction

In many modern applications of control systems, it is a matter of importance to obtain the required information of the process to be controlled, in order to aim at an optimal action or correct adaptation of the system. In optimal, self-optimizing and adaptive systems the quality of the control greatly depends on the amount of knowledge supplied to the controller. Besides, computers are increasingly used as control elements for complicated processes.

The growing demands concerning speed and quality of the control and consequently the insight into relevant process parameters justify an extension of the computer's task to process input and output data in order to obtain a better knowledge of the proces.

In an effort to determine the characteristics of a process on the basis of input and output data, we are searching for a dynamic operator which, acting on the input signal, results in the "best" estimate of the output signal.

The numerical values of the estimates of the process parameters can be obtained from a model. The difference e(t) between process and model output can be used as a measure for the error in the estimate, cf. fig. 1. Minimization of some function or functional of the error yields as model parameters $\underset{\sim}{\beta}$ estimates which are optimal with respect to the criterion chosen. Minimization of a quadratic type of criterion yields the least squares estimate.

Actually fig. 1 does not represent the most general situation. Bayes and maximum likelihood estimation may take into account more a priori knowledge that may be available [1]. Fig. 1, however, represent by far the most important estimation situation. Moreover it is easily shown that the maximum likelihood estimation method reduces to the (generalized) least squares method if the additive noise is Gaussian. As these least squares methods represent a great majority of parameter estimation cases, this paper is devoted to a closer investigation of their properties.

With respect to the instrumentation we distinct two classes [2] :

1) the explicit method
   the parameters are determined by instrumentation of the mathematical relations resulting from the minimization of the criterion;
2) the implicit method
   the parameters are determined by adjustment of a model.

Many publications deal with different approaches to the problem of parameter estimation, some of them discussing a specific process. A comparison of properties of the different types of instrumentation hardly ever occurs. Therefore, it is desirable to compare explicit and implicit methods for the case of models, which are linear in the parameters.

In principle the estimation of process parameters can be done both on analog and digital computers. Especially with respect to the simulation the digital machine has the following important advantages:

1) the memory of the digital machine is well suited for the simulation of processes with long response times or processes with delay times;
2) the errorless repeating generation of identical test and disturbing signals is apart from being throughout possible also of great importance where the influence of relatively short measuring intervals is studied in relation to noise effects.

With respect to the instrumentation of the estimation mechanism, the digital computer has the following practical advantages:

1) an intermittent measuring and model adjusting procedure can easily be instrumented;
2) fundamental mathematical manipulations such as matrix calculations and statistical computations over a certain amount of estimates can proceed accurately and be easily programmed.

Our investigations mainly concentrated toward an instrumentation using digital techniques, leading to a way of description of the estimation procedure as given below.

If the data are derived from analog processes, one has to count with additional errors caused by imperfect analog-digital conversion, e.g. due to quantization errors. In order to avoid this type of errors we only consider discrete processes.

As the example to be discussed we take a linear system, the impulse response of which is represented by the vector $\underline{h}^T = (h(1),h(2),\ldots,h(p))$ cf. [3]. This can be simulated by means of a delay line, cf. fig. 2.

For the description of the explicit method we define:

$\underline{x}_1^T = (x(1),x(2),\ldots,x(\ell),0\ldots,0)$ vector of inputs

$\underline{x}_2^T = (0,x(1),\ldots,x(\ell),0,\ldots,0)$

$\underline{y}^T = (y(1),y(2),\ldots,y(\ell+p-1))$ vector of outputs

$\underline{n}^T = (n(1),n(2),\ldots,n(\ell+p-1))$ vector of additive noise

$\underline{w}^T = (w(1),w(2),\ldots,w(\ell+p-1))$ vector of model outputs

$\underline{\beta}^T = (\beta(1),\beta(2),\ldots,\beta(p))$ parameter vector of the model

$\underline{h}^T = (h(1),h(2),\ldots,h(p))$ parameter vector of the process

$$X = \begin{bmatrix} x(1) & 0 & \cdots & 0 \\ x(2) & x(1) & & \vdots \\ \vdots & \vdots & & x(1) \\ x(\ell) & \vdots & & \vdots \\ 0 & x(\ell) & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & x(\ell) \end{bmatrix} = (\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_p) \qquad (1)$$

This yields:

$$\begin{aligned} \underline{z} &= \underline{y} + \underline{n} = X\,\underline{h} + \underline{n} \\ \underline{w} &= X\,\underline{\beta} \\ \underline{e} &= \underline{z} - \underline{w} \end{aligned} \qquad (2)$$

The length of the observation sequence is ℓ.

For the description of the implicit method we will consider the j^{th} observation interval; each interval having a length ℓ^*.
During this observation time the model vector is kept constant ($= \underline{\beta}_{j-1}$) prior to the adjustment from $\underline{\beta}_{j-1}$ to $\underline{\beta}_j$ (intermittent adjustment procedure). For the input and output quantities of this interval we write:

$$\begin{aligned} \underline{z}_j^T &= (z\{(j-1)\ell^*+1\},\ldots,z\{j\ell^*\}) \\ \underline{e}_j^T &= (e\{(j-1)\ell^*+1\},\ldots,e\{j\ell^*\}) \\ \underline{\beta}_{j-1}^T &= (\beta_{j-1}(1),\ldots,\beta_{j-1}(p)) \end{aligned} \qquad (3)$$

$$X_j = \begin{bmatrix} x\{(j-1)\,\ell^*+1\} & x\{(j-1)\,\ell^*\} & \cdots\cdots & x\{(j-1)\,\ell^*-p+2\} \\ x\{(j-1)\,\ell^*+2\} & & & \vdots \\ \vdots & & & \\ x\{j\,\ell^*\} & \cdots\cdots\cdots\cdots\cdots\cdots & & x\{j\,\ell^*-p+1\} \end{bmatrix}$$

This yields:

$$\begin{aligned} \underline{z}_j &= X_j\,\underline{h}+\underline{n}_j \\ \underline{w}_j &= X_j\,\underline{\beta}_{j-1} \\ \underline{e}_j &= \underline{z}_j-\underline{w}_j \end{aligned} \qquad (4)$$

In all cases we will suppose that the output noise has zero mean and is independent of the input signal, i.e.

$$\begin{aligned} E[\underline{n}] &= \underline{0} \\ E[X^T\underline{n}] &= \underline{0} \end{aligned} \qquad (5)$$

The criterion R to be minimized is

$$R = \underline{e}^T\,\Phi\,\underline{e} \qquad (6)$$

The choice of Φ is guided by the available a priori knowledge.

The explicit method yields an estimate $\tilde{\underline{\beta}}$ of $\underline{h}$ according to

$$\tilde{\underline{\beta}} = (X^T\Phi X)^{-1}\,X^T\Phi\underline{z} \qquad (7)$$

under the following conditions:

a) the numerical value of the impulse response h(i) must be constant after a given i;

b) at the beginning of the observation the model must have reached a state identical to that of the process, if no disturbances affect the process; i.e. either $x(i) = 0 \qquad i<0$
or, in the case the test signal constituting a part of a continuous sequence, only those measured outputs $\underline{z}$ appearing after i=p are taken into account;

c) the matrix $X^T\Phi X$ must not be singular and must have an inverse.

Under the same conditions, the implicit method yields an estimate $\underline{\beta}_j$ after j adjustments according to

$$\underline{\beta}_j = \underline{\beta}_{j-1} - \frac{g_j}{2} \left(X_j^T \Phi X_j\right)^{-1} \nabla_{\underline{\beta}_{j-1}} R_j \qquad (8)$$

where g_j is a factor governing the adjusting speed and $\nabla_{\underline{\beta}_{j-1}}$ denotes the gradient with respect to $\underline{\beta}_{j-1}$.

The least squares estimate is that estimate where no a priori knowledge is available, i.e. $\Phi = I$, I being the identity matrix. The input signal X is measurable. Its bandwidth is assumed large compared to the bandwidth of the process ("white" input signal).

Both explicit and implicit techniques request matrix inversion, which of course is a very expensive operation, but which has the advantage that arbitrary stochastic testsignals can be used.

A variant of the above technique is frequently used and can be instrumented in a simple way, because the matrix inversion can be avoided. A restriction inherent to this method is the fact that only a "white" input signal should be used as test signal.
For the <u>explicit</u> method the estimate without matrix inversion follows from

$$\tilde{h}(j) = \frac{\tilde{\psi}_{xy}(j,\ell)}{\psi_{xx}(0,\ell)} \quad \text{with} \quad \tilde{\psi}_{xy}(j,\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} x(i)y(j+i) \qquad (9)$$

A general insight into the estimates (7) and (9) can be obtained by considering that h(i) can be written explicitly in the two following ways if a "white" test signal is used.

$$\text{a)} \quad h(i) = \frac{\psi_{xy}(i)}{\psi_{xx}(0)}$$

$$\text{b)} \quad h(i) = \frac{\tilde{\psi}_{xy}(i,\ell)}{\psi_{xx}(0,\ell)} - \frac{1}{\psi_{xx}(0,\ell)} \underbrace{\left[\{h(i+1)+h(i-1)\}\, \tilde{\psi}_{xx}(1,\ell-1)+\ldots\ldots\right]}_{\text{compensating terms}} \qquad (10)$$

The procedure without matrix inversion neglects the compensating terms.
For the <u>implicit</u> method the estimate without matrix inversion follows from

$$\underline{\beta}_j = \underline{\beta}_{j-1} + g_j \{\ell^* \psi_{xx}(0)\}^{-1} X_j^T \underline{e}_j \qquad (11)$$

The neglection of the compensating terms causes an additional error. This error is called the "truncation error", which means in terms of model structure that the model has a shorter impulse response than that of the process, which means that condition a) is not fulfilled. If the conditions a), b) and c) are fulfilled, the uncertainty in the estimate is only caused by the disturbing noise.

2. Estimation schemes

a) the explicit method with matrix inversion

For the explicit method we found

$$\tilde{\underline{\beta}} = (X^T X)^{-1} X^T \underline{z} \tag{12}$$

Taking the expectation of $\tilde{\underline{\beta}}$ yields

$$E\left[\tilde{\underline{\beta}}\right] = E\left[(X^T X)^{-1} X^T \underline{z}\right] = \underline{h} \tag{13}$$

The estimate $\tilde{\underline{\beta}}$ is unbiased.

The noise error is written as follows

$$\Delta\tilde{\underline{\beta}} = \tilde{\underline{\beta}} - \underline{h} = (X^T X)^{-1} X^T (\underline{y}+\underline{n}) - \underline{h} =$$

$$= (X^T X)^{-1} X^T (X\underline{h}+\underline{n}) - \underline{h} = (X^T X)^{-1} X^T \underline{n} \tag{14}$$

From a statistical point of view the standard deviation of $\Delta\tilde{\underline{\beta}}$ is a useful measure for the error in the estimate. This can be found by calculating the covariance matrix of $\tilde{\underline{\beta}}$

$$\text{cov}\ \tilde{\underline{\beta}} = E\left[\Delta\tilde{\underline{\beta}}.\Delta\tilde{\underline{\beta}}^T\right] = (X^T X)^{-1} X^T E\left[\underline{n}.\underline{n}^T\right] X (X^T X)^{-1} \tag{15}$$

For "white" additive noise $\underline{n}$

$$E\left[\underline{n}.\underline{n}^T\right] = \psi_{nn}(0)\ I \tag{16}$$

$\psi_{nn}(i)$ being the autocorrelation function of $\underline{n}$. This yields for the covariance

$$\text{cov}\ \tilde{\underline{\beta}} = \psi_{nn}(0)\ (X^T X)^{-1} \tag{17}$$

For large ℓ and "white" input noise:

$$\text{cov}\ \tilde{\underline{\beta}} \approx \frac{\psi_{nn}(0)}{\ell \psi_{xx}(0)}\ I \tag{18}$$

For the standard deviation σ in each parameter we find:

$$\sigma = \sqrt{\frac{\psi_{nn}(0)}{\ell \psi_{xx}(0)}} \tag{19}$$

Listing the properties of the explicit method with matrix inversion, we come to the following specifications:

a) the estimate is linear: $\underline{\tilde{\beta}} = Q\,\underline{z}$;

b) the estimate is unbiased: $E\left[\underline{\tilde{\beta}}\right] = \underline{h}$;

c) $\lim_{\ell\to\infty} \operatorname{cov} \underline{\tilde{\beta}} = 0\ I$;

d) the method is suited for all types of input signals, provided all process modes are excited;

e) the method is suited for all types of additive noise;

f) the instrumentation requires matrix inversion, matrix-matrix multiplication and two matrix-vector multiplications.

b) the implicit method with matrix inversion

For the implicit method we found:

$$\underline{\beta}_j = \underline{\beta}_{j-1} - \frac{g_j}{2} (X_j^T X_j)^{-1} \nabla_{\underline{\beta}_{j-1}} R_j \tag{20}$$

or

$$\underline{\beta}_j = \underline{\beta}_{j-1} + g_j (X_j^T X_j)^{-1} X_j^T \underline{e}_j$$

Taking the expectation of $\underline{\beta}_j$ we find:

$$E\left[\underline{\beta}_j\right] = g_j\, \underline{h} + (1-g_j)\, E\left[\underline{\beta}_{j-1}\right] \tag{21}$$

leading to

$$E\left[\underline{\beta}_j\right] = \underline{h} \sum_{i=1}^{j} g_i \prod_{k=i+1}^{j} (1-g_k) + \underline{\beta}_0 \prod_{i=1}^{j} (1-g_i) \tag{22}$$

where $\underline{\beta}_0$ is the initial model guess (a priori knowledge). Where a priori information is lacking, $\underline{\beta}_0$ can be chosen arbitrary e.g. $\underline{\beta}_0 = \underline{0}$.

Considering first

$$\underline{\beta}_j - E\left[\underline{\beta}_j\right] = (1-g_j)\, (\underline{\beta}_{j-1} - E\left[\underline{\beta}_{j-1}\right]) + g_j\, (X_j^T X_j)^{-1} X_j^T\, \underline{n}_j \tag{23}$$

we derive

$$\operatorname{cov} \underline{\beta}_j = (1-g_j)^2 \operatorname{cov} \underline{\beta}_{j-1} + g_j^{\,2} (X_j^T X_j)^{-1} X_j^T\, E\left[\underline{n}\, \underline{n}^T\right] X_j (X_j^T X_j)^{-1} \tag{24}$$

If $\underline{n}$ is "white" noise (24) can be rewritten as a vari-linear difference equation:

$$(1-g_j)^2\, \Delta \operatorname{cov} \underline{\beta}_j - g_j\, (g_j - 2) \operatorname{cov} \underline{\beta}_j = \frac{g_j^{\,2}}{\ell^*}\, \psi_{nn}(0)\, \tilde{\psi}_{X_j}^{-1} \tag{25}$$

where

$$\Delta\text{cov}\ \underline{\beta}_j = \text{cov}\ \underline{\beta}_j - \text{cov}\ \underline{\beta}_{j-1} \tag{26}$$

and

$$\tilde{\Psi}_{X_j} = \frac{1}{\ell^*}\ X_j^T X_j \tag{27}$$

Now we can distinguish the following cases:

1) $g_j = 1/j$ "stochastic approximation", cf. [4].
2) g_j = constant "parameter tracking"

ad 1) $g_j = 1/j$

The algorithm is:

$$\underline{\beta}_j = \underline{\beta}_{j-1} + 1/j\ (X_j^T X_j)^{-1}\ X_j^T \underline{e}_j \tag{28}$$

Considering that in this particular case

$$\sum_{i=1}^{j} g_i \prod_{k=i+1}^{j} (1-g_k) = 1$$

$$\prod_{i=1}^{j} (1-g_i) = 0 \tag{29}$$

$$\lim_{j\to\infty} \prod_{i=a}^{j} (1-g_i) = 0 \qquad a \geq 1$$

it follows from equation (22) that

$$E\left[\underline{\beta}_j\right] = \underline{h} \tag{30}$$

This method of model matching yields an unbiased estimate for all j, irrespective of the initial model guess!

Equation (25) becomes

$$(j-1)^2\ \Delta\text{cov}\ \underline{\beta}_j + (2j-1)\ \text{cov}\ \underline{\beta}_j = \frac{\psi_{nn}(0)}{\ell^*}\ \tilde{\Psi}_{X_j}^{-1} \tag{31}$$

If the observation interval ℓ^* is large enough and if X_j is "white" noise the following approximation can be used for $\tilde{\Psi}_{X_j}^{-1}$

$$\tilde{\Psi}_{X_j}^{-1} \simeq \frac{I}{\psi_{xx}(0)} \tag{32}$$

Equation (26) then becomes

$$(j-1)^2\ \Delta\text{cov}\ \underline{\beta}_j + (2j-1)\ \text{cov}\ \underline{\beta}_j = \frac{\psi_{nn}(0)}{\ell^*\psi_{xx}(0)}\ I \tag{33}$$

with the solution

$$\text{cov}\ \underline{\beta}_j = \frac{\psi_{nn}(0)}{j\ell^*\psi_{xx}(0)}\ I = \frac{\psi_{nn}(0)}{\ell\psi_{xx}(0)}\ I \qquad (34)$$

where ℓ is the equivalent observation length used in the explicit method. In fig. 3 the diagonal elements of cov $\underline{\beta}_j$ are plotted as a function of $j\ell^*$.

For relatively small ℓ^* the approximation of eq. (32) is not longer valid: experimental results show a shift (dotted lines) towards greater variances.

Recapulating the properties of the adjustment procedure with matrix inversion and $g_j = 1/j$, we come to the following specification:

a) the adjustment is unbiased for all j;
b) $\lim_{j\to\infty}$ cov $\underline{\beta}_j = 0\ I$ for ℓ^* large enough
c) the method is suited for all types of input noise;
d) the method is suited for all types of additive noise;
e) the method requires for every interval a matrix inversion, a matrix-matrix multiplication and two matrix-vector multiplications;
f) the method yields intermediate results.
g) the method is not suited for parameter tracking.

ad 2) g_j = constant

The algorithm now is

$$\underline{\beta}_j = \underline{\beta}_{j-1} + c\ (X_j^T X_j)^{-1}\ X_j^T\ \underline{e}_j \qquad (35)$$

Substitution of

$$\sum_{i=1}^{j} g_i \prod_{k=i+1}^{j} (1-g_k) = 1 - (1-c)^j \qquad (36)$$

in eq. (22) and considering

$$\left.\begin{aligned} &\lim_{j\to\infty}\ 1 - (1-c)^j = 1 \\ &\lim_{j\to\infty}\ (1-c)^j = 0 \end{aligned}\right\}\ 0<c<2 \qquad (37)$$

yields

$$E\left[\underline{\beta}_j\right] = \{\ 1-(1-c)^j\ \}\ \underline{h} + (1-c)^j\ \underline{\beta}_0 \qquad (38)$$

For the covariance we find

$$\operatorname{cov} \underline{\beta}_j = (1-c)^2 \operatorname{cov} \underline{\beta}_{j-1} + c^2 \frac{\psi_{nn}(0)}{\ell^*} \tilde{\Psi}_{X_j}^{-1} \tag{39}$$

or written as a difference equation

$$(1-c)^2 \Delta\operatorname{cov} \underline{\beta}_j - c(c-2) \operatorname{cov} \underline{\beta}_j = c^2 \frac{\psi_{nn}(0)}{\ell^*} \tilde{\Psi}_{X_j}^{-1} \tag{40}$$

If ℓ^* is large enough and if a "white" input signal is used:

$$(1-c)^2 \Delta\operatorname{cov} \underline{\beta}_j - c(c-2) \operatorname{cov} \underline{\beta}_j = c^2 \frac{\psi_{nn}(0)}{\ell^* \psi_{xx}(0)} \, I \tag{41}$$

The asymptotic solution is:

$$\lim_{j\to\infty} \operatorname{cov} \underline{\beta}_j = \frac{c}{2-c} \frac{\psi_{nn}(0)}{\ell^* \psi_{xx}(0)} \, I \tag{42}$$

This method of model matching yields even after an infinite observation interval an estimate $\underline{\beta}_\infty$ with a variance greater than zero. This version of the adjustment algorithm is important for the identification of processes with slowly varying parameters (parameter tracking).

c) the explicit method without matrix inversion

As already pointed out above, the algorithm for the explicit estimation without matrix inversion is

$$\tilde{\underline{\beta}} = \frac{1}{\ell \tilde{\psi}_{xx}(0,\ell)} X^T \underline{z} \tag{43}$$

where X is a "white"input signal.

In eq. (10) it can be seen that in this situation a truncation occurs:

$$\tilde{\underline{\beta}} = \underline{h} + \underline{\Delta\beta}_{trunc} + \underline{\Delta\beta}_{noise} \tag{44}$$

$$\underline{\Delta\beta}_{noise} = \frac{1}{\ell \tilde{\psi}_{xx}(0,\ell)} X^T \underline{n} = \underline{\Delta\beta}_n \tag{45}$$

Considering an arbitrary "white" input matrix X, we obtain:

$$E\left[\tilde{\underline{\beta}}\right] = \underline{h} + \underline{\Delta\beta}_{trunc} \tag{46}$$

where $\underline{\Delta\beta}_{trunc}$ depends on the matrix X.

The algorithm (43) yields a biased estimate.

Taking

$$\underline{\Delta\beta}_n \cdot \underline{\Delta\beta}_n^T = \frac{1}{\ell^2 \tilde{\psi}_{xx}^2(0,\ell)} X^T \underline{n}\,\underline{n}^T X \tag{47}$$

we get in the case of "white" additive noise

$$\operatorname{cov} \underline{\beta}_n = \frac{\psi_{nn}(0)}{\ell^2 \tilde{\psi}_{xx}^2(0,\ell)} X^T X \tag{48}$$

For large ℓ we write

$$\operatorname{cov} \underline{\beta}_n \simeq \frac{\psi_{nn}(0)}{\ell \psi_{xx}(0)} \; I \tag{49}$$

In terms of standard deviation for each parameter

$$\sigma = \sqrt{\frac{\psi_{nn}(0)}{\ell \psi_{xx}(0)}} \tag{50}$$

Recapitulating the properties of the explicit method without matrix inversion we find:

a) the estimate is linear $\underline{\tilde{\beta}} = Q\,\underline{z}$;

b) the estimate is biased because $\frac{1}{\ell} X^T X = \tilde{\Psi}_X \neq \Psi_X$
however the following holds

$$\lim_{\ell \to \infty} E\left[\underline{\tilde{\beta}}\right] = \underline{h}$$

$E_X \; E\left[\underline{\tilde{\beta}}\right] = \underline{h}$, where E_X denotes the expectation with respect to X.

c) the method is only suited for "white" input noise;

d) the method is suited for all types of output noise;

e) the instrumentation only requires a multiplication of a $(\ell+p-1) \times p$ matrix by a $\ell+p-1$ vector.

c) <u>the implicit method without matrix inversion</u>

In equation (20) we can approximate the matrix inversion in the case of a "white" input signal:

$$(X_j^T X_j)^{-1} \approx \frac{1}{\ell^* \psi_{xx}(0)} \; I \tag{51}$$

This yields for the algorithm:

$$\underline{\beta}_j = \underline{\beta}_{j-1} + g_j (\ell^* \psi_{xx}(0))^{-1} \; X_j^T \underline{e}_j \tag{52}$$

Equation (52) is an interesting result as the total of four operations per iteration (matrix inversion, matrix-matrix multiplication and two matrix-vector multiplications) is reduced to one single operation (matrix-vector product) only.

Taking the expectation of $\underline{\beta}_j$ yields

$$E\left[\underline{\beta}_j\right] = g_j \left(\ell^* \psi_{xx}(0)\right)^{-1} \underline{X}_j^T \underline{X}_j \underline{h} + \{I - g_j \left(\ell^* \psi_{xx}(0)\right)^{-1} \underline{X}_j^T \underline{X}_j\} E\left[\underline{\beta}_{j-1}\right] \tag{53}$$

The input signal in the j^{th} iteration influences $E\left[\underline{\beta}_j\right]$: the adjustment is biased. For sufficiently large ℓ^* equation (53) leads to

$$E\left[\underline{\beta}_j\right] = g_j \underline{h} + (1-g_j) E\left[\underline{\beta}_{j-1}\right] \tag{54}$$

We can take the expectation of eq. (53) with respect to the input signals in all preceding intervals

$$E_{X_1} E_{X_2} \ldots E_{X_j} E\left[\underline{\beta}_j\right] = \underline{h} \sum_{i=1}^{j} g_i \prod_{k=i+1}^{j} (1-g_k) + \underline{\beta}_0 \prod_{i=1}^{j} (1-g_i) \tag{55}$$

The expression for the covariance is given by

$$\begin{aligned} \operatorname{cov} \underline{\beta}_j &= E\left[\left(\underline{\beta}_j - E\left[\underline{\beta}_j\right]\right)\left(\underline{\beta}_j - E\left[\underline{\beta}_j\right]\right)^T\right] = \\ &= \{I - g_j \left(\ell^* \psi_{xx}(0)\right)^{-1} \underline{X}_j^T \underline{X}_j\} \operatorname{cov} \underline{\beta}_{j-1} \{I - g_j \left(\ell^* \psi_{xx}(0)\right)^{-1} \underline{X}_j^T \underline{X}_j\} \\ &\quad + g_j^2 \left(\ell^* \psi_{xx}(0)\right)^{-2} \psi_{nn}(0) \underline{X}_j^T \underline{X}_j \end{aligned} \tag{56}$$

In order to get some insight, this expression is instrumented. The results are briefly summarized as follows:

1) $g_j = 1/j$

   In fig. 4 $\operatorname{cov} \underline{\beta}_j$ is plotted against $j\ell^*$.

   For large $j\ell^*$ this diagram shows that $\operatorname{cov} \underline{\beta}_j$ has reached approximately the same value as in the adjustment with matrix inversion.

   Only for small $j\ell^*$ a deviation from the line

$$\operatorname{cov} \underline{\beta}_j = \frac{1}{j\ell^*} \frac{\psi_{nn}(0)}{\psi_{xx}(0)} I$$

   can be observed.

2) g_j = constant

   For relatively large ℓ^* (56) can be approximated by

$$\operatorname{cov} \underline{\beta}_j \approx (1-c)^2 \operatorname{cov} \underline{\beta}_{j-1} + c \frac{\psi_{nn}(0)}{\ell^* \psi_{xx}(0)} I \tag{57}$$

yielding as asymptotic solution the same expression as in the case of adjustment with matrix inversion:

$$\operatorname{cov} \underline{\beta}_{\infty} = \frac{c}{2-c} \frac{\psi_{nn}(0)}{\ell^{*}\psi_{xx}(0)} \tag{58}$$

3. Some remarks on other types of errors in a practical situation.

Considering the noise error we assumed that certain conditions were fulfilled. This, however, will mostly not be the case. It is therefore useful to indicate some additional errors that occur when these conditions are not satisfied.

Truncation error.

The delay li[illegible]n our mathematical model will often be too small, which then causes the phenomenon known as truncation error. The existence of this kind of error may be considered in an analogous way as was done with the noise error, namely as if it was caused by additive correlated noise added to the output of the process, cf. fig. 5.

The explicit expression of the truncation error vector $\underline{\Delta\beta}$ is

$$\underline{\Delta\beta} = \underline{\tilde{\beta}} - \underline{h}^{*} = (U^{T}U)^{-1} U^{T}(\underline{y}-\underline{\gamma}) = (U^{T}U)^{-1} U^{T}\underline{r} \tag{59}$$

U is a partial matrix of X satisfying

$$\begin{aligned} X\underline{h} &= \underline{y} \\ X\underline{h}^{*} &= \underline{\gamma} \end{aligned} \tag{60}$$

The truncation error does not affect the unbiasedness of the estimation when the input signal is "white" noise, as appears from:

$$(U^{T}U)\underline{\tilde{\beta}} = U^{T}\underline{y} = U^{T}\underline{\gamma} + U^{T}\underline{r} \tag{61}$$

$$E\left[U^{T}U\right] E\left[\underline{\tilde{\beta}}\right] = E\left[U^{T}\underline{\gamma}\right] + E\left[U^{T}\underline{r}\right] =$$

$$= E\left[U^{T}U\right]\underline{h}^{*} E\left[U^{T}\underline{r}\right] \tag{62}$$

$U^{T}\underline{r}$ appears to be:

$$U^T\underline{r} = \begin{bmatrix} \sum_{v=1}^{k-p} \{ h_{p+v} \sum_{i=1}^{\ell-p-v+1} x_i x_{i+p+v-1} \} \\ \sum_{v=1}^{k-p} \{ h_{p+v} \sum_{i=1}^{\ell-p-v+2} x_i x_{i+p+v-2} \} \\ \vdots \\ \sum_{v=1}^{k-p} \{ h_{p+v} \sum_{i=1}^{\ell-v} x_i x_{i+v} \} \end{bmatrix} \qquad (63)$$

If the input is "white" noise it is easy to see that

$$E\left[U^T\underline{r}\right] = \underline{0}$$

so $$E\left[\underline{\tilde{\beta}}\right] = \underline{h}^*$$

Hence, also $\underline{\Delta\beta} \to 0$ for $\ell \to \infty$.

In contrast with the noise error, the truncation error is independent of the power of the test signal.

Even in processes with "infinite memory" the truncation error need not cause predominantly bad estimates, because always in an actual estimation procedure only a finite number ($\ell+p-1$) of values of the impulse response is calculated with. It is even possible to estimate without errors by joining an integrator to the end of the model's delay line, which is assumed to be sufficiently large.

In some publications, cf. [3], the integrator is placed in front of the delay line but this implies that the estimates will have to be provided with a correction factor afterwards.

In the case of estimating a process of "infinite memory" without the use of an integrator the estimates, through unbiased, will even for $\ell \to \infty$ remain uncertain with a finite variance. This is caused by the fact that the assumptions required for ergodicity do not match. The variance is equal in both cases of instrumentation with or without matrix inversion.

It is reasuring that in a practical situation the truncation errors in the several estimated parameters for large ℓ are as good as equal in value and polarity, since the components in the error vector $\underline{\Delta\beta}$ are strongly correlated. Hence we may conclude that in a single estimation we indeed get a rather good idea of the shape of the impulse response, although it is uncertain to what extent every estimated value deviates from the

corresponding value of the impulse response. So every estimated value can be corrected, if a priori knowledge about at least one value of the impulse response is available, e.g. when the amplification factor is known.

Observation error.

Observation errors appear if the assumption b) turns out to be unjustified. We can avoid this kind of errors in our estimates by neglecting the first and last p measured output values. However, when it is only possible to observe for a very short time, all output values have to be used in the calculations and the errors have to be weighted in a proper way, cf. [5].

4. Conclusions.

The described explicit and implicit methods appear to yield equal results with respect to the uncertainty in the estimate caused by additive noise. Likewise, this uncertainty (noise error) appears to be independent of instrumentation with or without matrix inversion. This offers the experimenter considerable freedom to select a method satisfying his needs and being as simple as possible.

Fundamentally, the instrumentation without matrix inversion causes some uncertainty due to truncation errors. In a practical situation, however, for sufficiently large ℓ the truncation error does not predominantly influence the total uncertainty in the estimate. Besides, such an influence can often be recognized easily.

Acknowledgement.

The authors wish to thank Professor P. Eykhoff for his aid and encouragement and Mr. A.A. van Rede for his helpful suggestions during the course of this work.

Literature.

[1]. P. Eykhoff : "Process Parameter and State Estimation", invited survey paper for the IFAC Symposion, Prague, June 1967; Automatica, vol. 4 (1968), p.p. 205-233.

[2]. P. Eykhoff : "Some Fundamental Aspects of Process-Parameter Estimation", IEEE Trans. on Automatic Control, October 1963, p.p. 347-357.

[3]. M.J. Levin : "Estimation of the Characteristics of Linear Systems in the Presence of Noise", Ph.D. Thesis, Dpt. of El. Eng., Columbia University, April 1959.

[4]. Ya.Z. Tsypkin : "Adaptation, Training and Self-Organisation in Automatic Systems", Automation and Remote Control, vol. 27 (1966), no. 1, p.p. 16-51.

[5]. N.R. Draper and H. Smith: "Applied Regression Analysis", John Wiley and Sons Inc., New York, 1966.

[6]. R. Deutsch : "Estimation Theory", Prentice Hall Inc., Englewood Cliffs, N.J., 1965.

[7]. E. Blandhol : "On the Use of Adjustable Models for Determination of System Dynamics", Division of Automatic Control, Technical University of Norway, Trondheim, Norway, March 1962.

[8]. J.H.A.M. Melis : "The Explicit Estimation of Process Parameters Using a Digital Computer", M.Sc. Thesis, Technological University Eindhoven, Netherlands, June 1967 (in Dutch).

[9]. A.J.W. van den Boom : "The Implicit Estimation of Process Parameters Using a Digital Computer", M.Sc. Thesis, Technological University Eindhoven, Netherlands, June 1967 (in Dutch).

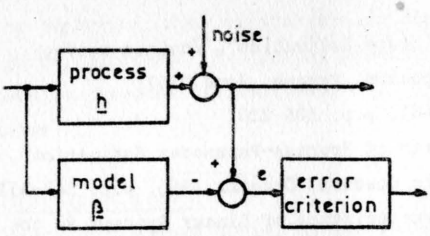


Fig. 1.

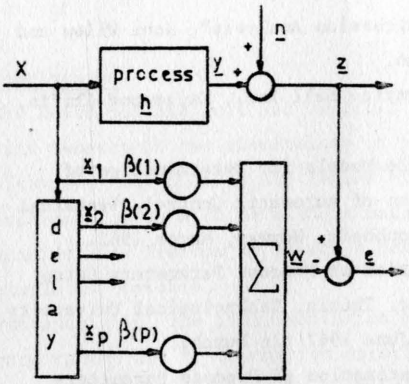


Fig. 2.

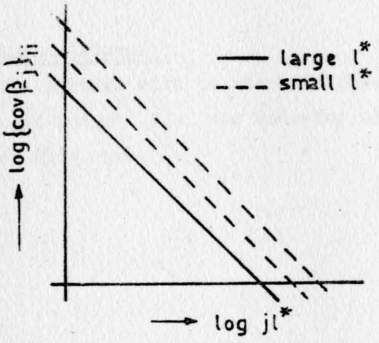


Fig. 3.

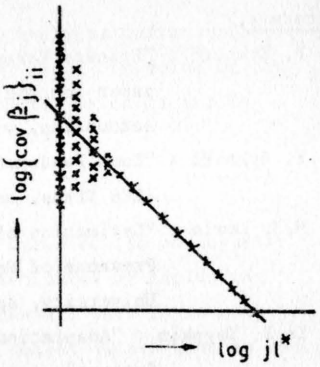


Fig. 4.

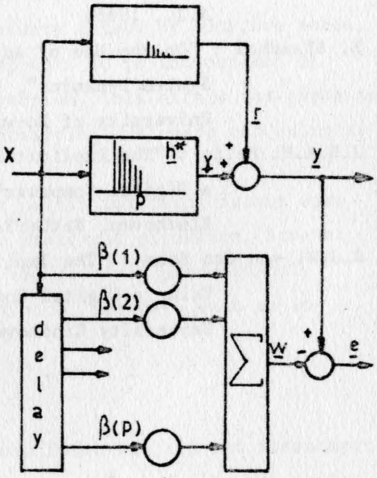


Fig. 5.

AN INSTRUMENTAL VARIABLE METHOD FOR REAL-TIME IDENTIFICATION OF A NOISY PROCESS

Peter C. Young*

Naval Weapons Center
China Lake, California, United States of America

INTRODUCTION

The problem of real-time process parameter estimation from normal operating data has received considerable attention in recent years.[1] The various techniques developed range from largely deterministic procedures[2] to sophisticated statistical methods based on the results of optimal estimation theory.[3] The Instrumental Variable (I.V.) technique outlined in this paper is intended as a compromise between these two extremes; it has a basis in classical statistical estimation theory, but does not require *a priori* information on the signal and noise statistics.

The complete identification scheme requires hybrid instrumentation; the estimation procedure itself takes the form of a computationally efficient recursive algorithm. The basically simple approach represents a logical development of earlier estimation schemes used by the author[4, 5, 6] and is intended to be a widely applicable method for estimating differential equation model parameters of single input—single output processes. However, the algorithm can be extended for use with certain multivariable problems and with difference equation models if desired.[7]

THE PROBLEM

Consider a single input—single output dynamic process that can be described by a linear differential equation model

$$\sum_{n=0}^{N} a_n \frac{d^n x}{dt^n} = u + \sum_{m=1}^{M} b_m \frac{d^m u}{dt^m} \tag{1}$$

where $u = u(t)$ is the input command signal and $x = x(t)$ is the process output response to $u(t)$. The parameters a_n and b_m are a set of $M+N+1$ unknown coefficients which may have time variations that are slow in comparison with the response time of the process. If x_n, $n=0 \rightarrow N$ and u_m, $m=0 \rightarrow M$ are used to denote $\frac{d^n x}{dt^n}$ and $\frac{d^m u}{dt^m}$ respectively, then (1) can be written

$$\sum_{n=0}^{N} a_n x_n = u_0 + \sum_{m=1}^{M} b_m u_m \tag{2}$$

which can be converted into the alternative vector equation

$$x_A^T a = u_0 \tag{3}$$

---

*This paper provides a brief resumé of research carried out between 1965 and 1967 in the Department of Engineering, University of Cambridge, England. The research was supported by the Whitworth Foundation.

where

$$\left.\begin{aligned} x_A^T &= \left[x_0\ x_1 \cdots x_N\ u_1\ u_2 \cdots u_M\right] \\ a^T &= \left[a_0\ a_1 \cdots a_N - b_1 - b_2 \cdots - b_m\right] \end{aligned}\right\} \tag{4}$$

Details of unmeasurable disturbances or measurement noise that may affect the process are not supplied by (3), nor does it provide a picture of the process as seen by an external observer. It will be assumed here that these details can be included by adjoining the observation equations.

$$y = Hx_A + \xi \tag{5}$$

$$v = u_0 + n \tag{6}$$

Here, H is a j x $(M+N+1)$ matrix ($j \leq M+N+1$) with unity or zero elements relating the observed j vector, $y = \left[y_0\ y_1 \cdots y_{j-1}\right]^T$, with the augmented state vector, x_A. The symbol ξ is a j random vector denoting the noise present on the observations. By the principle of superposition for linear systems, it can be assumed that ξ represents the combined effects of all unmeasurable disturbances and measurement noise affecting the process. The observed input signal, v, is contaminated by measurement noise, n, which does not enter the process.

The overall signal topology of the system as described by (3), (5), and (6) is shown in Fig. 1(a). Identifying this process during normal operations is a problem of statistical parameter estimation. To solve such a problem, it is necessary to choose a method for utilizing the observed data, y and v, to derive an estimate $\hat{a}$ of a, whereby the resulting estimated model will adequately describe the dynamic characteristics of the process.

GENERALIZED EQUATION ERROR METHOD

Previous publications by the author describe a limited solution to the estimation problem posed in the last section which can be used with either continuous[4, 5] or discrete[4, 6] data. This approach is based on the definition of an "estimation model" that obeys a similar differential equation law to that of the basic process, but which does not include pure time derivatives of the process input and output signals. This estimation model is obtained by operating on each of the terms appearing in the differential equation (1) with a linear time invariant filter, D_i, where in general

$$D_i(s) = P_i(s)/Q_i(s) \tag{7}$$

and $P_i(s)$; $Q_i(s)$ are constant coefficient polynomials in the Laplace operator, s, with orders I and J. If a_n and b_m are assumed constant, then it is possible to write the following identities:

$$\left.\begin{aligned} D_i\left\{a_n \circ x_n(t)\right\} &= a_n\ D_i\ x_n(t) = a_n\left[(x_n)\right]_{D_i} \\ D_i\left\{b_m \circ u_m(t)\right\} &= b_m\ D_i\ u_m(t) = b_m\left[(u_m)\right]_{D_i} \end{aligned}\right\} \tag{8}$$

where the terms $\left[(x_n)\right]_{D_i}$ and $\left[(u_m)\right]_{D_i}$ can be considered physically as the outputs of the filter $D_i(s)$, whose inputs are x_n and u_m respectively.

Equation (8) states that because the process is linear and the parameters a_n and b_m are time invariant, then the operator, D_i, commutes with the functions $a_n \cdot x_n$ and $b_m \cdot u_m$. If this is the case, then it has been shown[4, 6] that the parameters a_n and b_m can be related by the estimation model

$$\sum_{n=0}^{N} a_n \left[(x_n)\right]_{D_i} = \left[(u_0)\right]_{D_i} + \sum_{m=1}^{M} b_m \left[(u_m)\right]_{D_i} \tag{9}$$

This relationship is valid for all time, t, subsequent to an initial small interval of time, ε, following the initiation of the filtration at $t = t_0$ provided that: (a) the form of $D_i(s)$ is such that any initial conditions on the process variables at $t = t_0$ have insignificant effect on the filter outputs $\left[(x_n)\right]_{D_i}$ and $\left[(u_m)\right]_{D_i}$ for all time, $t > t_0 + \varepsilon$; and (b) the frequency bandwidth of the filter $D_i(s)$ approximately encompasses the frequency band covered by the differential equation model of the process—in other words, the frequency band of interest in the identification.

The significance of (9) is that it provides a relationship between the unknown parameters a_n and b_m that is capable of replacing the original differential equation (2) for identification purposes. The practical utility of the estimation model becomes clearer if the commutation carried out in (8) is taken one step further, i.e.,

$$\left.\begin{aligned} \left[(x_n)\right]_{D_i} &= \left[(x_0)\right]_{D_{in}} \\ \text{and} \qquad \left[(u_m)\right]_{D_i} &= \left[(u_0)\right]_{D_{im}} \end{aligned}\right\} \tag{10}$$

where $D_{in}(s) = s^n D(s)$; $D_{im}(s) = s^m D(s)$. It is now clear that the estimation model can be written in the alternative form

$$\sum_{n=0}^{N} a_n \left[(x_0)\right]_{D_{in}} = \left[(u_0)\right]_{D_{i0}} + \sum_{m=1}^{M} \left[(u_0)\right]_{D_{im}} \tag{11}$$

in which the variables $\left[(x_0)\right]_{D_{in}}$ and $\left[(u_0)\right]_{D_{im}}$ are obtained as the outputs of the "state variable filters" D_{in} and D_{im} (a term used by Hofmann, *et al.*[8]), whose inputs are simply the process output, x_0, and input, u_0, respectively. These filters can be made physically realizable provided the order J of $Q_i(s)$ is selected so pure differentiation, with all its attendant practical limitations, is not specified. This requires that J should be made greater than or equal to the total sum of the order I of $P_i(s)$ and the order of the maximum differential coefficient appearing in the model of the process. Since for most physical processes $N \geq M$, this condition is usually of the form $J \geq I+N$.

Probably the simplest example of a physically realizable set of state

variable filters is obtained by cascading first order low- and high-pass filters as described in Ref. 5. Other related but more subtle approaches are the method of Kohr[9] and the method of multiple filters[4, 6]. A more detailed discussion regarding the choice of particular state variable filter configurations and the general philosophy of the state variable filter approach is given in Ref. 7.

In any practical situation (see (b) of Fig. 1), the process output, x_0, and input, u_0, are not directly measurable and must be replaced by their observed values, y_0 and v, respectively. Consequently, the estimation model (11) has to be modified as shown in (12).

$$\sum_{n=0}^{N} a_n \left[\left(y_0 \right) \right]_{D_{in}} = \left[\left(v \right) \right]_{D_{i0}} + \sum_{m=1}^{M} b_m \left[\left(v \right) \right]_{D_{im}} \tag{12}$$

This can be written in the alternative vector form

$$\mathbf{z}^T \mathbf{a} = w \tag{13}$$

where

$$\mathbf{z}^T = \left[\left[\left(y_0 \right) \right]_{D_{i0}} \left[\left(y_0 \right) \right]_{D_{i1}} \cdots \left[\left(y_0 \right) \right]_{D_{iN}} \left[\left(v \right) \right]_{D_{i1}} \cdots \left[\left(v \right) \right]_{D_{iM}} \right]$$

$$w = \left[\left(v \right) \right]_{D_{i0}}$$

Note that by the principle of superposition, the following relationships are also true (see (b) of Fig. 1):

$$\left.\begin{aligned} \mathbf{z}^T &= \mathbf{x}_{AD}^T + \boldsymbol{\xi}_D^T = \left[\left[\left(x_0 \right) \right]_{D_{i0}} \cdots \left[\left(u_0 \right) \right]_{D_{iM}} \right] + \left[\left[\left(\xi_0 \right) \right]_{D_{i0}} \cdots \left[\left(\eta \right) \right]_{D_{iM}} \right] \\ w &= u_D + n_D = \left[\left(u_0 \right) \right]_{D_{i0}} + \left[\left(n \right) \right]_{D_{i0}} \end{aligned}\right\} \tag{14}$$

Referring to (13), it is now possible to define a generalized equation error function, e, at an arbitrary ith instant of time by the relationship

$$e_i \triangleq \mathbf{z}_i^T \hat{\mathbf{a}} - w_i \tag{15}$$

Here, $\hat{\mathbf{a}}$ represents an estimate of the parameter vector, $\mathbf{a}$, and the suffix, i, denotes values applying at the ith instant. The estimate, $\hat{\mathbf{a}}$, now can be obtained by minimizing some positive definite-criterion function in the generalized equation error.

If the parameters are assumed time invariant, a useful criterion function is the sum of the squares over the observation interval, J_2. For a set of k observations, J_2 takes the form

$$J_2 = \sum_{i=1}^{k} \left[\mathbf{z}_i^T \hat{\mathbf{a}} - w_i \right]^2 \tag{16}$$

The estimates, $\hat{\mathbf{a}}$, that minimize this function can be obtained by differentiating with respect to the $\hat{\mathbf{a}}$ and equating to zero; i.e.,

$$\left[\sum_{i=1}^{k} z_i \, z_i^T\right] \hat{a} - \left[\sum_{i=1}^{k} z_i \, w_i\right] = 0$$

The solution of the $M+N+1$ linear simultaneous algebraic equations, which are similar to the normal equations of linear regression analysis, is given by

$$\hat{a}_k = C_k^{-1} B_k = P_k B_k \tag{17}$$

where

$$P_k^{-1} = C_k = \sum_{i=1}^{k} z_i \, z_i^T$$

and

$$B_k = \sum_{i=1}^{k} z_i \, w_i$$

In (17), $\hat{a}$ can be obtained by direct matrix inversion, or by using a gradient technique, such as the method of *conjugate gradients*.[7, 10] However, an alternative approach more suitable for real-time application is available, which also avoids matrix inversion. This technique uses the well known recursive solution of linear least squares problems of this type, in which the parameter estimate is updated as new data is received. The derivation of this solution is straightforward; it is merely a stepwise solution of the fixed sample length problem.[7, 11] In the resulting recursive algorithm, (17) is replaced by

$$\hat{a}_k = \hat{a}_{k-1} - P_{k-1} z_k \left[z_k^T P_{k-1} z_k + 1\right]^{-1} \left\{z_k^T \hat{a}_{k-1} - w_k\right\} \tag{18}$$

or

$$\hat{a}_k = \hat{a}_{k-1} - P_k \left\{z_k z_k^T \hat{a}_{k-1} - z_k w_k\right\}$$

where P_k is given by a second recursive relationship in the form

$$P_k^{-1} = P_{k-1}^{-1} + z_k z_k^T$$

or

$$P_k = P_{k-1} - P_{k-1} z_k \left[z_k^T P_{k-1} z_k + 1\right]^{-1} z_k^T P_{k-1} \tag{19}$$

Using this approach, the estimate at the kth instant is obtained by the repeated application of (18) and (19) from assumed initial conditions $\hat{a}_0$ and P_0. The selection of these initial conditions is not critical.[7, 11]

NOISE CONSIDERATIONS—THE INSTRUMENTAL VARIABLE (I.V.) ALGORITHM

If there is noise present on the observed signals, the generalized equation error estimate discussed in the previous section is asymptotically biased to a degree which is dependent upon the noise/signal ratio. This unfortunate characteristic arises because the vector, z, is composed of elements subjected to noise contamination. Therefore, (13) is an example of a "structural" rather than a "regression" model.[12]

The reason for the estimation bias becomes apparent if stationary statistical properties are assumed and the expected value of the matrix

$\left[zz^T\right]$ is examined. The following is obtained from (14):

$$E\left[zz^T\right] = E\left[\left(x_{AD} + \xi_D\right)\left(x_{AD} + \xi_D\right)^T\right]$$

However, since x_{AD} and ξ_D are uncorrelated, this becomes

$$E\left[zz^T\right] = E\left[x_{AD}\, x_{AD}^T\right] + E\left[\xi_D\, \xi_D^T\right] \tag{20}$$

where the noise induced term $E\left[\xi_D\, \xi_D^T\right]$ is identically zero only when there is no noise on the observed data. It can be shown[6,7] that the presence of this noise induced term introduces the asymptotic bias on the parameter estimates.

One approach to the above problem that does not require *a priori* knowledge of the noise statistics has its foundation in statistical estimation theory where it has been termed the Instrumental Variable (I.V.) method.[13] The asymptotic bias is removed by modifying the solution given by (17) in the following manner

$$\hat{a} = \hat{P}_k\, \hat{B}_k \tag{21}$$

where

$$\left.\begin{aligned} \hat{P}_k^{-1} &= \sum_{i=1}^{k} \hat{x}_i\, z_i^T \\ \text{and} \qquad \hat{B}_k &= \sum_{i=1}^{k} \hat{x}_i\, w_i \end{aligned}\right\} \tag{22}$$

Here, $\hat{x}$ is an I.V. vector composed of elements chosen to be highly correlated with unobservable *noise-free* process variables, x_{AD}, but totally uncorrelated with the various additive noise components that corrupt these signals. As a result, the matrix $\left[\hat{x}z^T\right]$, which now replaces the matrix $\left[zz^T\right]$, of the uncompensated algorithm, has the expected value

$$E\left[\hat{x}z^T\right] = E\left[\hat{x}\, x_{AD}^T\right] \neq 0$$

Furthermore, it is clear that

$$E\left[\hat{x}\, x_{AD}^T\right] \to E\left[x_{AD}\, x_{AD}^T\right] \quad \text{as} \quad \hat{x} \to x_{AD} \tag{23}$$

In the long term, therefore, the inclusion of the I.V. vector, $\hat{x}$, eliminates the troublesome noise term, $E\left[\xi_D\, \xi_D^T\right]$, while preserving the basic structure and existence of the solution. In this way, the asymptotic bias is removed from the estimates, ensuring only small bias for finite sample lengths. Unfortunately, the elimination of bias in the above manner is usually accompanied by certain loss of efficiency in the statistical sense.[13] However, as might be expected from (23), the greater the correlation between $\hat{x}$ and the noise-free signals, the smaller the estimation variance. In fact, a simple theoretical analysis shows that the asymptotic variance should approach zero if $\hat{x}$ can be made equal to x_{AD}.[7]

The major problem with the I.V. approach is the generation of the instrumental variables, themselves. The method suggested in Ref. 6 is a de-

velopment of an analog technique used by Levadi.[14] Levadi's technique is a development of earlier procedures for the estimation of difference equation model parameters suggested by Joseph, *et al.*,[15] and Andeen and Shipley.[16] The procedure for the more general case, where the process is contained within a noisy feedback loop, is shown in Fig. 2. The basic philosophy of this approach is that by prefiltering the input command, u^*, by an "auxiliary model" of the process, it is possible to generate an I.V. vector, $\hat{\boldsymbol{x}}$, which is highly correlated with the noise-free process vector, $\boldsymbol{x}_{AD}$. In addition, the elements of $\hat{\boldsymbol{x}}$ will be uncorrelated with any other noise in the system provided the input command is noise-free. In practice, low levels of input noise (< 5%) can be tolerated without introducing noticeable bias.

Two approaches to the problem of selecting the auxiliary model parameters are suggested in Ref. 7. The first approach is an off-line iterative routine. The model parameters are initially selected on the basis of either *a priori* information, or previous uncompensated (i.e., biased) estimates of the process parameters which are then updated by a series of I.V. runs. The second approach (initially outlined by the author[17]) is based on a recursive solution to the problem and can be used in real-time. The method of deriving this recursive algorithm (which is similar to that used in the uncompensated case) is given in Appendix A. The main equations are repeated below for convenience.

$$\hat{a}_k = \hat{a}_{k-1} - \hat{P}_{k-1}\,\hat{x}_k \left[z_k^T \hat{P}_{k-1}\,\hat{x}_k + 1\right]^{-1} \left\{z_k^T \hat{a}_{k-1} - w_k\right\} \tag{24}$$

$$\hat{P}_k = \hat{P}_{k-1} - \hat{P}_{k-1}\,\hat{x}_k \left[z_k^T \hat{P}_{k-1}\,\hat{x}_k + 1\right]^{-1} z_k^T \hat{P}_{k-1} \tag{25}$$

With the help of this algorithm, it is possible to develop a fully adaptive approach; the estimates are low-pass filtered to avoid rapid transients, and then are used to update the auxiliary model on a continuous basis. In both of the above cases, convergence can be argued.[7] However, because of equipment problems, practical verification has been possible for the iterative method only.

IDENTIFIABILITY

It has been noted that the equation error approach is closely related to the procedures of linear regression analysis. One important and often overlooked feature of regression analysis is that the *regressors* should be linearly independent if accurate low variance estimation is to be possible.[12] In the equation error procedures described here, the place of the regressors is taken by the elements of the vector, $\boldsymbol{z}$. Since these elements are dependent upon both the process input signals and the nature of the assumed model, it is clear that both factors have an important bearing on the *identifiability* of the process.[7]

Choice of Input Signals

To ensure satisfactory identification, theoretical considerations indicate that the input signal should be *sufficiently exciting* in that (a)

$$\sum_{i=1}^{k} u_i^2 > 0 \quad ; \text{ where } u_i = \left(u_0\right)_i$$

and (b) the number of discrete frequency components in any periodic signal should exceed d where

$$d \geq \begin{cases} (M+N+1)/2 & ; \quad M+N+1 \text{ even} \\ (M+N+2)/2 & ; \quad M+N+1 \text{ odd} \end{cases}$$

These conditions are a theoretical guide and should merely be considered as minimum requirements. Practical experiments suggest that *if* input signals can be selected, a random noise-type signal usually gives excellent results.[7,8,9]

Nature of the Mathematical Model

The choice of a sufficiently exciting input signal does not guarantee low variance estimation. There usually will be some measure of *partial* linear dependence[7,18] between the elements of the augmented state vector arising because of the model structure. In the present differential equation case, problems of this sort arise when the assumed model has input derivative terms. A simple but interesting example demonstrating the kind of results obtained in such situations is discussed later.

One method of checking the results of a *pure regression* experiment to test for linear dependence is *multiple correlation analysis.*[19] The same approach can be used to good effect in the low-noise, uncompensated equation error case also; however, it must be used with caution since it is well known that correlation coefficients can be biased by errors in observations.[19] The special properties of the adaptive I.V. method are rather useful in this respect since they enable the development of a *pseudo* multiple correlation analysis[7,18] that appears to give good qualitative results and may have quantitative significance.

Controllability

One final point that should be mentioned is the implication of *controllability*[20] on the identification of the process. Theoretically, if the process is uncontrollable in that a pole-zero cancellation is present, it is not identifiable. This is confirmed by experimental results, showing that identification is particularly poor if *exact* cancellation is present. As such a situation rarely occurs in the real world, it will not normally cause problems.

DETECTING PARAMETER VARIATION

Since the accumulated square-type criterion function weights all data equally over the observation interval, it contains an implicit assumption that the parameters are constant over this period. If slow parameter variation is possible, precautions must be taken to ensure that outdated estimates are not obtained. A particularly simple approach to this problem is weighting the data exponentially into the past to gradually remove information as it becomes obsolete. An analog method of exponential weighting by low-pass averaging filters has been described.[5] A discrete data equivalent of this procedure can be developed quite easily,[7] and has been used to detect the nonstationary characteristics of both simulated[7,16] and practical[21] processes.

A more attractive general approach to the problem of parameter variation can be developed by referring to the equivalent pure regression situation.[7,11] Here, a nonstationary version of the recursive least-squares equations is obtained by considering a stochastic interpretation of the problem, then introducing the statistical model of the parameter variations

$$a_k = \Phi(k, k-1)a_{k-1} + q_{k-1} \tag{26}$$

where $\Phi(k, k-1)$ is an assumed known $(M+N+1) \times (M+N+1)$ transition matrix, and the $(M+N+1)$ random disturbance vector, q_{k-1}, has zero mean value and covari-

ance matrix $E\left(q_r q_s^T\right) = Q\delta_{rs}$ (δ is the Kronecker delta function). The vector provides a statistical degree of freedom for the equation. The resulting prediction-correction algorithm is a form of the optimal discrete filter-estimation equations suggested by Kalman.[22]

By using a purely heuristic argument based on the similarity between the equation error method and the linear regression analysis, it is possible to construct a *dynamic* equation error algorithm. The algorithm is of very limited practical utility as it stands because it requires knowledge of the parameter transition matrix, $\Phi(k, k\text{-}1)$. Fortunately, it is a straightforward matter to simplify the algorithm by letting $\Phi(k, k\text{-}1)$ equal I for all k, implying that any parameter variation is due to small random perturbations between samples (a random walk process). In this case, the I.V. algorithm can be written

$$\hat{a}_k = \hat{a}_{k-1} - \hat{P}_{k/k-1}\,\hat{z}_k\left[z_k^T\,\hat{P}_{k/k-1}\,\hat{z}_k + 1\right]^{-1}\left\{z_k^T\,\hat{a}_{k-1} - w_k\right\} \tag{27}$$

$$\hat{P}_{k/k-1} = \hat{P}_{k-1} + D \tag{28}$$

$$\hat{P}_k = \hat{P}_{k/k-1} - \hat{P}_{k/k-1}\,\hat{z}_k\left[z_k^T\,\hat{P}_{k/k-1}\,\hat{z}_k + 1\right]^{-1} z_k^T\,\hat{P}_{k/k-1} \tag{29}$$

The only difference between this procedure and the stationary parameter procedure given by (24) and (25) is the inclusion of (28). According to the heuristic argument, D in this equation is analogous to the covariance matrix of the parameter variations, Q, in the pure regression case. As a result, the choice of D proves to be fairly straightforward, since it can be selected by reference to the expected rates of parameter variation between samples.

The physical effect of introducing the D matrix is to limit the lower bound on the P matrix elements, preventing the elements from becoming too small and allowing for continuous correction of the parameter estimates as time progresses. Since D is a matrix, it is possible to limit individual elements to different degrees. In this way, different expected rates of parameter variation can be specified on the elements of a (see Experimental Results).

The major disadvantage of the type of algorithm described above is its inability to differentiate between actual parameter variations and indicated parameter variations caused by noise on the data. As a result, the approach only proves satisfactory for the estimation of parameter variations larger than the residual estimation variance due to noise.

In general, the dynamic equation error algorithm—although not optimal in any sense—is to be preferred to the alternative data weighting approach. The inclusion of the parameter variation model means that the algorithm has much greater inherent flexibility. For instance, the choice of a random walk model for the parameter variations is arbitrary; in certain situations it may be more realistic to specify other models, such as an exponentially correlated random variation, or a random ramp change. The virtue of the model approach simply is that it allows this type of *a priori* information to be used directly in the estimation algorithm if it is available. It also provides the possibility of a more sophisticated procedure in which the parameter variation model, itself, is updated by a secondary "learning" scheme.

EXPERIMENTAL RESULTS

A number of experiments designed to test the practical utility of the techniques discussed in this paper were made with the experimental equipment

shown in Fig. 3. The process and the auxiliary model were simulated on an analog computer, which also was used to synthesize the track-store devices and filters required by the method of multiple filters.[4, 6]

When the parameter, a_3, of the second order process shown in Fig. 3 is known, the process is clearly identifiable. In this situation, the estimation results obtained with the above equipment were excellent.[23] However, when all four parameters were assumed unknown, the results were not so satisfactory—with particularly high variance on the estimates $\hat{a}_1$ and $\hat{a}_3$. A table of the estimates obtained for various levels of noise contamination is shown below.

Parameter Estimates After 170 Samples

Noise/Signal Ratios, δ	Estimated Parameter Values $\hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
0.269	0.955	1.056	0.677	1.069
0.359	1.055	2.038	1.600	2.233
0.538	0.964	1.091	0.748	1.099
0.897	1.104	1.800	1.212	1.515
Actual Parameter Values	1.000	1.400	1.000	1.500

These estimation results were obtained using the I.V. algorithm given by (24) and (25), with the process activated by a *pseudo* random-step input and filtered white-noise disturbances (Fig. 4). Although they look rather poor at first sight, the results tend to be misleading since the step, initial condition, and frequency responses for the estimated model (as given in Figs. 5 and 6) show good agreement with the actual process responses. Thus, if the object of an identification experiment is to predict process output response to input activation, the estimation results may be considered satisfactory.

The kind of results described above can be explained by the high degree of partial linear dependence existing between the elements of the vector $\boldsymbol{x}_{AD}$ due to the presence of the input derivative term. This fact becomes apparent if a multiple correlation analysis (see Identifiability) is performed on the measurement data: With a_3 assumed known, the multiple correlation coefficients are small compared with the total correlation coefficient; with a_3 unknown, they assume much higher values, indicating strong linear relationships between certain of the elements of $\boldsymbol{x}_{AD}$. A physical explanation of this peculiar phenomenon is that the sensitivity of the process response to changes in certain parameters is small. Consequently, the estimation error does not possess a clearly defined minimum and the estimates tend to "drift," thus producing high variance.

Figure 7 shows the results obtained when the algorithm given by (27), (28), and (29), was used to identify a process in which the parameters, a_1, a_2, and a_3 were time-invariant, and the a_0 parameter was varied sinusoidally. It should be stressed that Fig. 7 illustrates *optimum* performance in the sense that the auxiliary model parameter was varied in accordance with the actual process parameter variation. This approach was necessary because equipment delays prevented full hybrid operation. When two parameters were

varied simultaneously, similar results were obtained (Fig. 8). The results shown in Figs. 7 and 8 were obtained with a_3 assumed known (in other words, with the process clearly identifiable).

Appendix A

DERIVATION OF THE RECURSIVE INSTRUMENTAL VARIABLE ALGORITHM

Matrix Inversion Lemma

The matrix $\hat{P}_k$ is related to the matrix $\hat{P}_{k-1}$ by

$$\hat{P}_k = \left[\hat{P}_{k-1}^{-1} + \hat{x}_k z_k^T\right]^{-1} \tag{A-1}$$

which can be expressed

$$\hat{P}_k = \hat{P}_{k-1} - \hat{P}_{k-1}\hat{x}_k\left(z_k^T\hat{P}_{k-1}\hat{x}_k + 1\right)^{-1} z_k^T\hat{P}_{k-1} \tag{A-2}$$

Proof by Direct Multiplication:

$$\begin{aligned}
\hat{P}_k^{-1}\hat{P}_k &= \left[\hat{P}_{k-1}^{-1} + \hat{x}_k z_k^T\right]\left[\hat{P}_{k-1} - \hat{P}_{k-1}\hat{x}_k\left(z_k^T\hat{P}_{k-1}\hat{x}_k + 1\right)^{-1} z_k^T\hat{P}_{k-1}\right] \\
&= I + \hat{x}_k z_k^T\hat{P}_{k-1} - \hat{x}_k\left(z_k^T\hat{P}_{k-1}\hat{x}_k + 1\right)^{-1} z_k^T\hat{P}_{k-1} \\
&\quad - \hat{x}_k z_k^T\hat{P}_{k-1}\hat{x}_k\left(z_k^T\hat{P}_{k-1}\hat{x}_k + 1\right)^{-1} z_k^T\hat{P}_{k-1} \\
&= I + \hat{x}_k z_k^T\hat{P}_{k-1} - \left(\hat{x}_k z_k^T\hat{P}_{k-1}\hat{x}_k + \hat{x}_k\right)\left(z_k^T\hat{P}_{k-1}\hat{x}_k + 1\right)^{-1} z_k^T\hat{P}_{k-1} \\
&= I + \hat{x}_k z_k^T\hat{P}_{k-1} - \hat{x}_k z_k^T\hat{P}_{k-1} = I \qquad \text{Q.E.D.}
\end{aligned}$$

Estimation Algorithm

From (21), the estimate $\hat{a}_k$ at the kth instant is given by

$$\hat{a}_k = \hat{P}_k\hat{B}_k \tag{A-3}$$

Now B_k is related to B_{k-1} by

$$\hat{B}_k = \hat{B}_{k-1} + \hat{x}_k w_k \tag{A-4}$$

Substituting (A-2) and (A-4) into (A-3) gives

$$\begin{aligned}
\hat{a}_k &= \left[\hat{P}_{k-1} - \hat{P}_{k-1}\hat{x}_k\left(z_k^T\hat{P}_{k-1}\hat{x}_k + 1\right)^{-1} z_k^T\hat{P}_{k-1}\right]\left[\hat{B}_{k-1} + \hat{x}_k w_k\right] \\
&= \hat{a}_{k-1} - \hat{P}_{k-1}\hat{x}_k\left(z_k^T\hat{P}_{k-1}\hat{x}_k + 1\right)^{-1} z_k^T\hat{a}_{k-1} + \hat{P}_{k-1}\hat{x}_k w_k \\
&\quad - \hat{P}_{k-1}\hat{x}_k\left(z_k^T\hat{P}_{k-1}\hat{x}_k + 1\right)^{-1} z_k^T\hat{P}_{k-1}\hat{x}_k w_k \\
&= \hat{a}_{k-1} - \hat{P}_{k-1}\hat{x}_k\left(z_k^T\hat{P}_{k-1}\hat{x}_k + 1\right)^{-1}\left\{z_k^T\hat{a}_{k-1} - \left(z_k^T\hat{P}_{k-1}\hat{x}_k + 1\right)w_k\right. \\
&\quad \left. + z_k^T\hat{P}_{k-1}\hat{x}_k w_k\right\}
\end{aligned}$$

$$\therefore \hat{a}_k = \hat{a}_{k-1} - \hat{P}_{k-1}\,\hat{z}_k \left(z_k^T \hat{P}_{k-1}\,\hat{z}_k + 1\right)^{-1} \left\{ z_k^T \hat{a}_{k-1} - w_k \right\} \cdots \qquad \text{(A-5)}$$

or, more simply

$$\hat{a}_k = \hat{a}_{k-1} - \frac{1}{\Delta} \hat{P}_{k-1} \left\{ \hat{z}_k\, z_k^T \hat{a}_{k-1} - \hat{z}_k\, w_k \right\}$$

where Δ is the scalar quantity defined by $\left(z_k^T \hat{P}_{k-1}\,\hat{z}_k + 1\right)$.

REFERENCES

1. P. C. Young, "Process parameter examination," *Control Magazine* (to be published).
2. P. M. Lion, "Rapid identification of linear and nonlinear systems," *Proc. JACC*, pp. 605-615, 1966.
3. R. E. Kopp and R. J. Orford, "Linear regression applied to system identification for adaptive control systems," *AIAA J.*, vol. 1, no. 10, October 1963.
4. P. C. Young, "In flight dynamic checkout," *IEEE Trans. on Aerospace*, vol. AS-2, no. 3, July 1964.
5. ———, "The determination of the parameters of a dynamic process," *The Radio and Electronic Engineer* (*J. IERE*), vol. 29, no. 6, June 1965.
6. ———, "Process parameter estimation and self adaptive control," *Proc. 1965 IFAC* (Teddington) *Conf.* In *Theory of Self Adaptive Control System*, P. H. Hammond, Ed., pp. 118-139. New York: Plenum Press, 1966.
7. ———, Doctoral thesis, submitted to Dept. Engrg., Univ. of Cambridge, England, 1968 (unpublished).
8. L. G. Hofmann, P. M. Lion, and J. J. Best, "Theoretical and experimental research on parameter tracking systems," NASA Contractor Rept. CR-452, April 1966.
9. L. L. Hoberock and R. H. Kohr, "An experimental determination of differential equations to describe simple nonlinear systems," *Proc. 1966 JACC*, pp. 616-623.
10. P. C. Young, "Parameter estimation and the method of conjugate gradients," *IEEE Proc.*, vol. 54, no. 12, p. 1965, December 1966.
11. R. C. K. Lee, "Optimal estimation identification and control," MIT Press, Research Monograph 28, 1964.
12. M. G. Kendall and A. Stuart, *The advance theory of statistics*. London: Griffin, 1961, vol. 2.
13. J. Durbin, "Errors in variables," *Review of Int. Statist. Inst.*, vol. 22, no. 23, 1954.
14. V. S. Levadi, "Parameter estimation of linear systems in the presence of noise," presented at the *1964 International Conf. on Microwaves, Circuit Theory, and Information Theory*, September 7-11, Tokyo, Japan.

15. P. Joseph, J. Lewis, and J. Tou, "Plant identification in the presence of disturbances and application to digital adaptive systems," *AIEE Trans.*, March 1961.

16. R. E. Andeen and P. P. Shipley, "Digital adaptive flight control system for aerospace vehicles," *AIAA J.*, vol. 1, no. 5, May 1963.

17. P. C. Young, "On a weighted steepest descent method of process parameter estimation," Engrg. Lab., Cambridge Univ., England, Internal Rept., December 1965.

18. ———, "Identification problems associated with the equation error approach to process parameter estimation," submitted for presentation at the *2nd Asilomar Conf. on Circuits and Systems*, 1968.

19. K. A. Brownlee, *Statistical theory and methodology in science and engineering.* New York: Wiley, 1965.

20. P. M. Derusso, R. J. Roy, and C. M. Close, *State variables for engineers.* New York: Wiley, 1965.

21. J. W. Bray, *et al.*, "On line model making for a chemical plant," *Trans. Soc. Inst. Tech.*, vol. 17, no. 8, 1965.

22. R. E. Kalman, "A new approach to linear filtering and prediction theory," *ASME Trans., J. Basic Engrg.*, vol. 82-D, pp. 35-45.

23. P. C. Young, "Regression analysis and process parameter estimation: ... a cautionary message," *Simulation*, vol. 10, no. 3, pp. 125-128, March 1968.

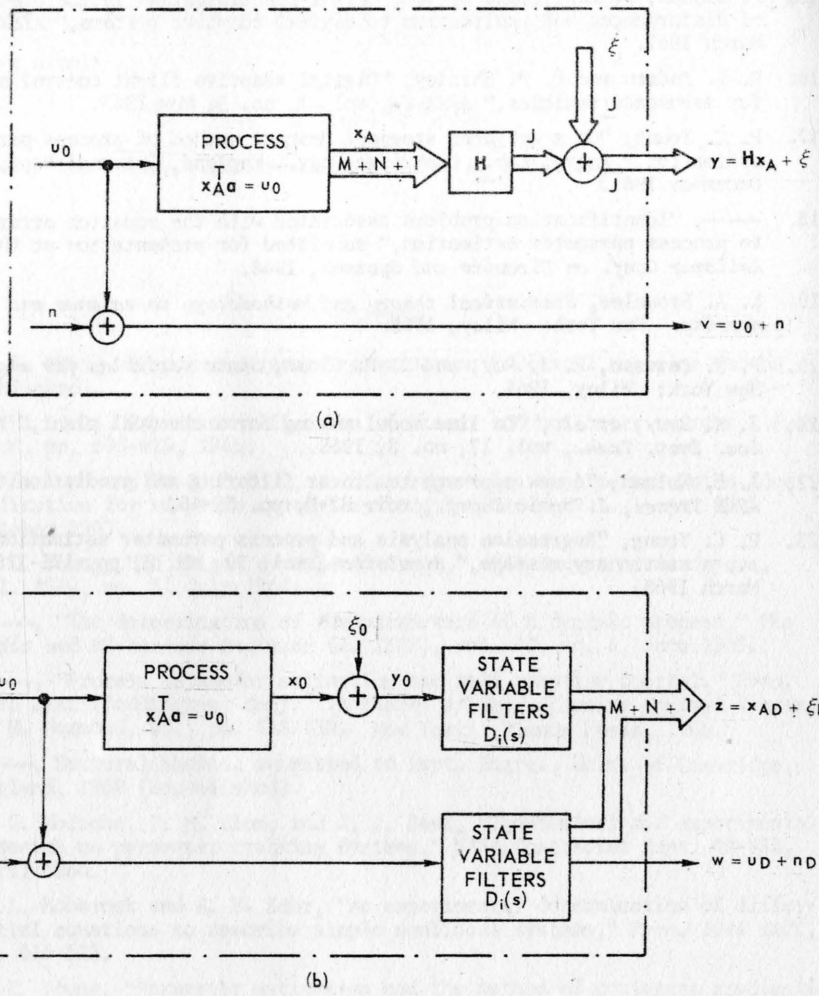


Fig. 1. Signal topology of (a) the process; (b) the process with state variable filters.

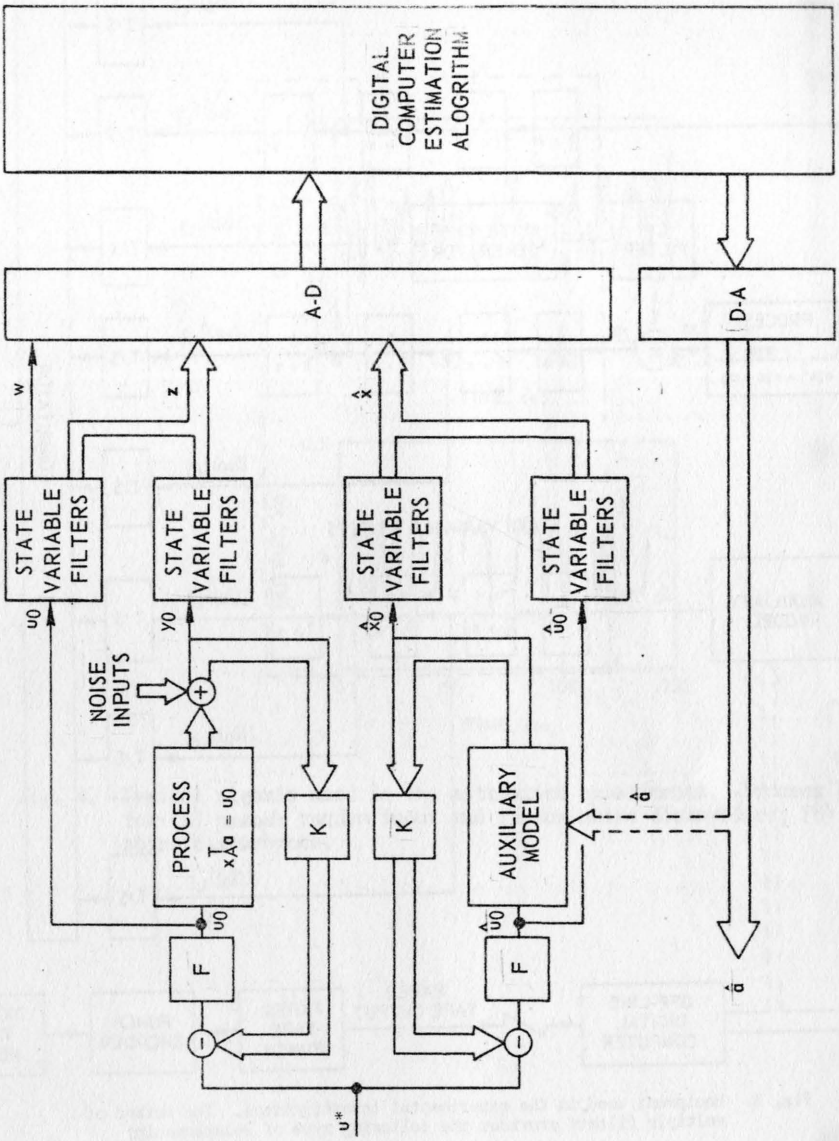


Fig. 2. Instrumental variable method for identifying a process within a closed loop--the auxiliary model approach (F and K are known control system elements).

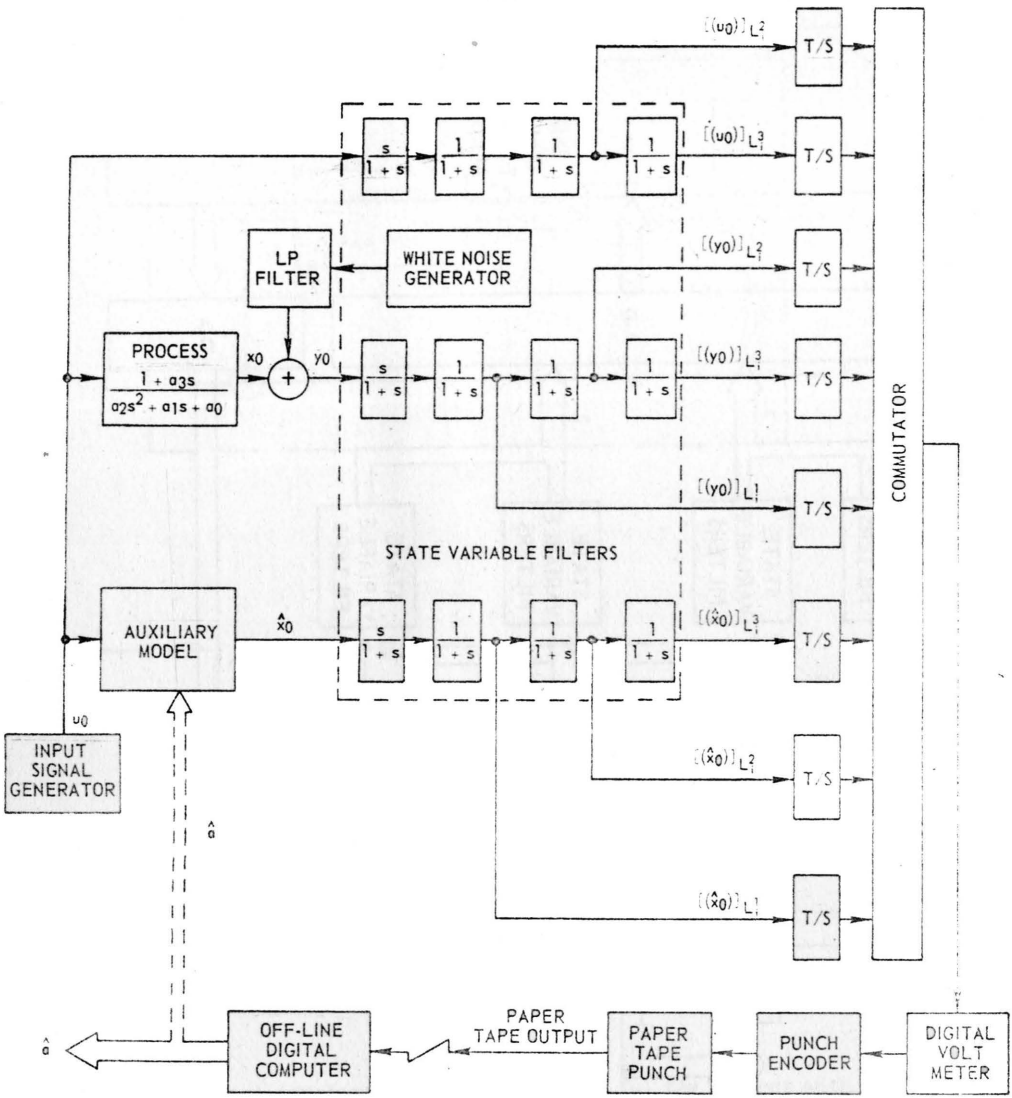


Fig. 3. Equipment used in the experimental investigations. The method of multiple filters provides the following type of relationship:

$$\left[(y_0)\right]_{D_{in}} = (-1)^n \left\{ \left[(y_0)\right]_{L_i^P} - n\left[(y_0)\right]_{L_i^{P-1}} + \cdots \right.$$

$$\left. + (-1)^j \frac{n(n-1)\cdots(n-j+1)}{j!} \left[(y_0)\right]_{L_i^{P-j}} \cdots \right|$$

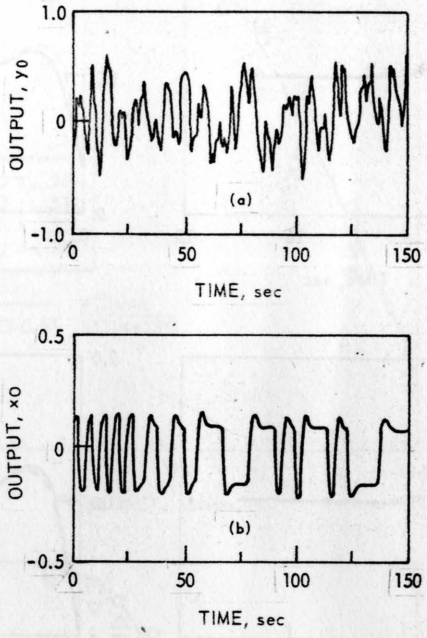


Fig. 4. Typical signals used in the estimation experiments. Process output for: (a) *pseudo* random input and random noise disturbance; (b) without noise disturbance.

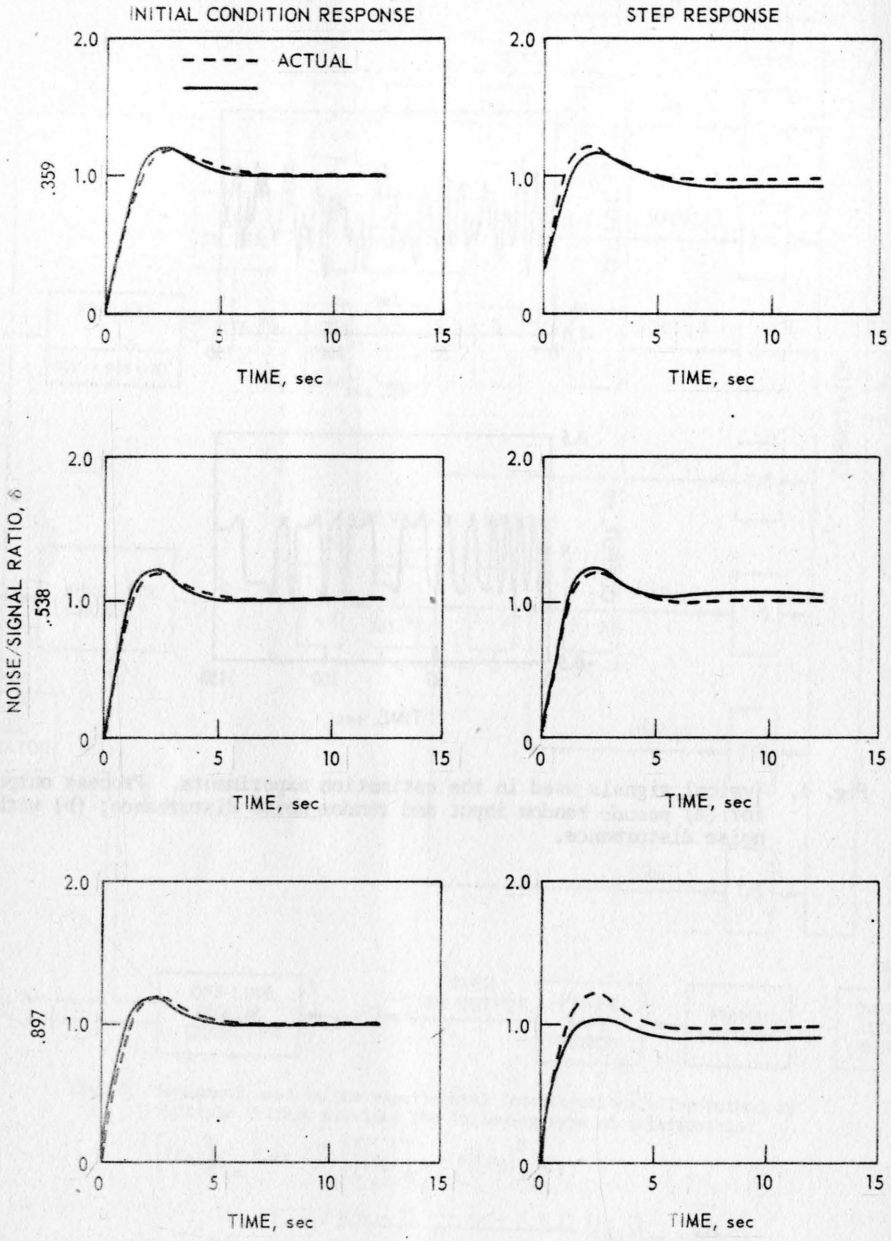


Fig. 5. Comparison of actual and estimated time responses obtained when there is partial linear dependence between the elements composing the augmented state vector.

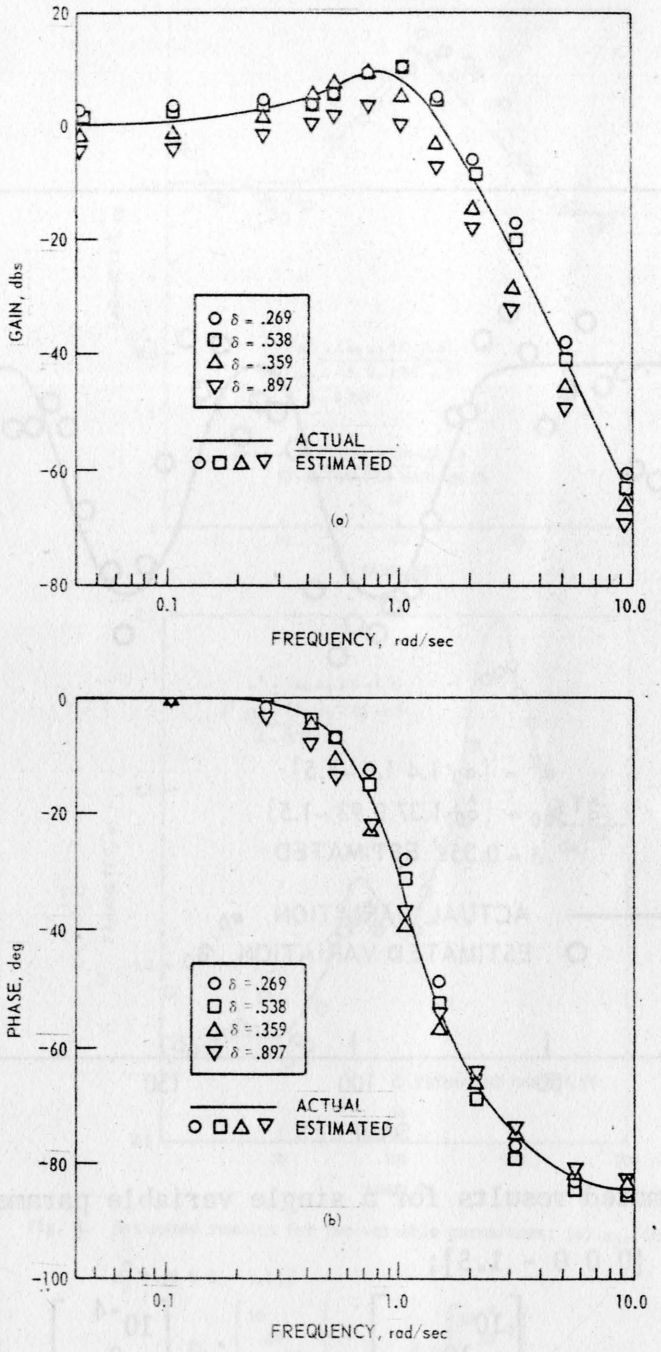


Fig. 6. Comparison of actual and estimated frequency response obtained when there is partial linear dependence between the elements composing the augmented state vector.

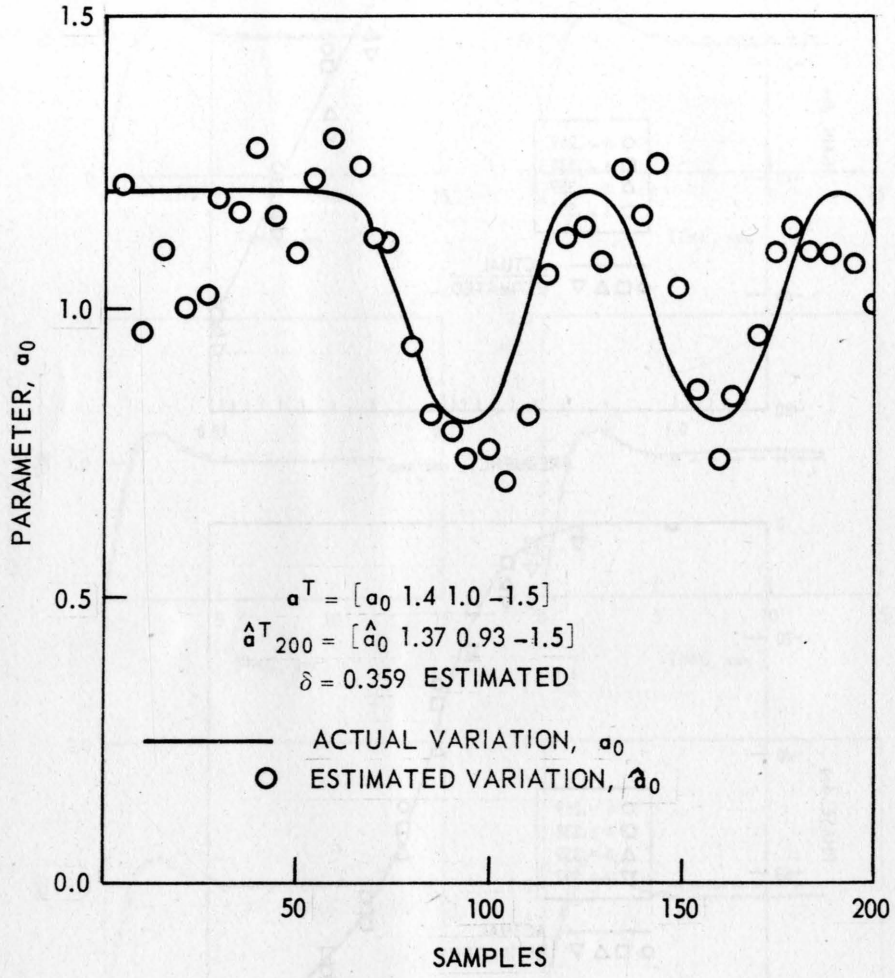


Fig. 7. Estimated results for a single variable parameter: $a_0^T = [0\ 0\ 0\ -\ 1.5]$;

$$P_0 = \begin{bmatrix} 10 & & & \\ & 10 & & \\ & & 10 & \\ & & & 0 \end{bmatrix} \quad ; \quad D = \begin{bmatrix} 10^{-4} & & & \\ & 0 & & \\ & & 0 & \\ & & & 0 \end{bmatrix}$$

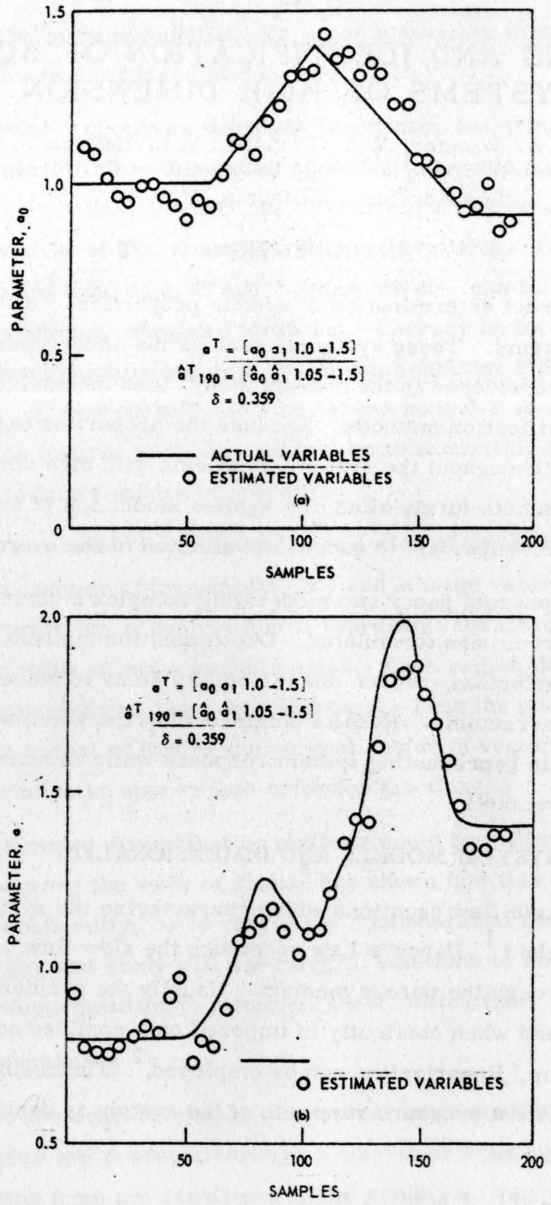


Fig. 8. Estimated results for two variable parameters; (a) a_0, and (b) a_1:

$a_0^T = [0\ 0\ 0\ -1.5]$;

$$P_0 = \begin{bmatrix} 10 & & & \\ & 10 & & \\ & & 10 & \\ & & & 0 \end{bmatrix} \quad ; \quad D = \begin{bmatrix} 10^{-4} & & & \\ & 10^{-4} & & \\ & & 0 & \\ & & & 0 \end{bmatrix}$$

MODELING AND IDENTIFICATION OF AQUIFER SYSTEMS OF HIGH DIMENSION

D. A. Wismer, R. L. Perrine, Y. Y. Haimes
Department of Engineering, University of California
Los Angeles, California, U. S. A.

INTRODUCTION

The direct determination of system properties is difficult for many physical systems. These systems, such as the underground aquifers or reservoirs considered in the present study, lend themselves to study by system identification methods. Because the properties to be determined vary widely throughout the system, problems with high dimensionality result. Thus both formulation of a system model and of an optimization procedure are important to successful solution of the overall problem.

In the present paper two moderately complex models of an underground reservoir are formulated. Decomposition methods, together with model selection, reduce dimensionality so as to make the identification problem tractable. Results obtained show the solution methods to be effective in representing system response while reducing computational requirements.

SYSTEM MODELS AND DIMENSIONALITY

The basic flow equations which characterize the system model are well established.[1] Darcy's Law describes the slow flow of a compressible fluid through the porous medium. Usually the gradients involved are small, and when continuity is imposed on a confined aquifer or reservoir system, linearization can be employed.[2] The result, then, is that the transient pressure response of the system is described by the diffusion equation.

$$\frac{\partial}{\partial x}\left(T\,\frac{\partial P}{\partial x}\right)+\frac{\partial}{\partial y}\left(T\,\frac{\partial P}{\partial y}\right)=S\,\frac{\partial P}{\partial t}+Q$$

Only two space dimensions are considered because with typical systems vertical flow seldom is important. Pressure is denoted by P, and Q represents a source strength (production rate per unit area). The coefficients in the equation characterize the porous medium. Transmissibility, T(x, y), is a measure of the ease with which fluid moves through

the system. The storage function, $S(x, y)$, is a measure of system capacity. Both are distributed parameters.

At this point appropriate boundary conditions, including production rates, and values for T and S must be specified to be able to predict future system response. A difficulty, of course, is that detailed knowledge of the variation of $T(x, y)$ and $S(x, y)$ is not available. On the other hand, pressure and time data, P and t respectively, can be obtained by observing the system at specified locations. Thereby an inverse problem in reservoir description is created: given some function $F(P_{observed} - P(T, S)_{calculated})$, how must T and S be chosen so that F is minimized? A solution to the inverse problem enables one to accurately predict system response to future modes of operation.

Thus it is assumed that a useful description of the system is given by specifying a number of transmissibility and storage values: $\underline{T}_i$ and $\underline{S}_i$. The interpretation of each is that it represents the effective or average parameter value within a spatial region. Each region then can be treated mathematically as being homogeneous. Thus the problem to be solved reduces to that of finding the optimal choice of vectors $\underline{T}_i$ and $\underline{S}_i$ which will minimize an appropriate criterion function.

Straightforward discretization methods could be utilized for the flow problem. However, the work of Jahns[3] has shown how this leads either to excess dimensionality, or to very large computational requirements, or both. The present study will use integral solutions to the flow problem together with decomposition to minimize these difficulties.

Model One Formulation

Consider the characteristics of a typical large aquifer or reservoir. Quite often fluids are produced through a cluster of wells located near its center. Starting from any stabilized initial condition, over reasonably large periods of time no effect of system boundaries is likely to be felt. The system is very large in areal extent compared to its vertical dimensions, and so can be represented as an infinite two-dimensional system containing a cluster of wells in a region of primary interest.

When each well is producing, "boundaries" with zero potential gradients (no flow) must exist between wells. Fluid on opposite sides of these "boundaries" flows to opposite wells, just as if the line of separation were impermeable. Thus a convenient approximation in modeling the system is to separate one-well regions by straight-line boundaries. A pair of boundary lines on opposite sides of one well will intersect to form a wedge. As a result the model for a cluster of N wells in a large reservoir is obtained by dividing the system into N wedge-shaped homogeneous regions, each radiating out from a single selected origin and enclosing a single well; well i at azimuth θ_i. The regions are separated by N straight, impermeable boundaries also radiating from the origin at azimuths $\alpha_i, (\theta_i < \alpha_i < \theta_{i+1})$. The identification scheme will determine optimal values of the boundaries α_i as well as S_i and T_i. Thus the geometry of the model is actually determined by system behavior. Model geometry is illustrated in Figure 1.

As was noted earlier, the transient flow equation for the system considered leads to a diffusion equation which can be more conveniently written in another coordinate system as follows

$$T\left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial \phi}{\partial r}\right)\right] = S\frac{\partial \phi}{\partial t} \qquad (1)$$

where ϕ denotes the model computed pressure to distinguish it from observed values, and r represents radial distance from the well. The initial and boundary conditions for the partial differential equation are the following[2, 4]

$$\lim_{t\to 0} \phi(r,t) = \lim_{r\to \infty} \phi(r,t) = P_1, \text{ a constant} \qquad (2a)$$

$$\lim_{r\to 0} r\left(\frac{\partial \phi}{\partial r}\right) = \frac{q}{2\pi T}, \text{ a constant} \qquad (2b)$$

The positive constant production rate is given by q. These conditions define slow laminar flow of a slightly compressible fluid in an infinite, horizontal, radial isotropic system of uniform thickness.

Equation (1) is transformed into an ordinary differential equation using the Boltzmann transformation: $\lambda = (Sr^2/4Tt)$. The new equation is:

$$\lambda\frac{d^2\phi}{d\lambda^2} + (1+\lambda)\frac{d\phi}{d\lambda} = 0 \qquad (3)$$

with boundary conditions:

$$\lim_{\lambda \to \infty} \phi(\lambda) = P_1 ; \quad \lim_{\lambda \to 0} 2\lambda \frac{d\phi}{d\lambda} = -\frac{q}{2\pi T} \tag{4a, b}$$

The solution to this equation is obtained in terms of the exponential integral; Ei(u):

$$\phi = P_1 - \frac{q}{4\pi T} \int_{u=\lambda}^{u=\infty} \frac{\exp(-u)}{u} du = P_1 + \frac{q}{4\pi T} \text{Ei}\left(-\frac{Sr^2}{4Tt}\right) \tag{5}$$

Since Ei(-u) is negative, the pressure will decrease as production occurs.

The use of images for extension of the solution to problems in a bounded, wedge-shaped, homogeneous region is straightforward.[1, 4] Suppose the i^{th} actual wedge includes an angle of $2\pi/m_i$ radians, between azimuths α_i and α_{i-1}, where m_i is an even integer. The image system then contains m_i wedges, filling the whole plane. Each wedge has the same origin, and contains a well which is the mirror image of the actual well and/or image wells across each of its adjacent boundaries. The angle between the i^{th} well with azimuth θ_i and its k^{th} image is given by $2\zeta_{ik}$ where

$$\zeta_{ik} = \left[\frac{k}{2}\right](\alpha_i - \theta_i) + \left[\frac{k-1}{2}\right](\theta_i - \alpha_{i-1} + 2\pi\,\delta_{i1}) \tag{6}$$

$$i = 1, 2, \ldots, N \qquad k = 1, 2, \ldots, m_i$$

and δ_{ij} is the Kronecker delta. The notation [k/2] and [(k-1)/2] denotes that these quantities are truncated to integer values. The α_i define wedge boundaries.

Thus the pressure computed at the i^{th} well and j^{th} time is conveniently denoted by ϕ_{ij}, which in the bounded system results from production at a rate q_i. It is computed by superposition of the pressure effects caused by all m_i image wells, for each of N actual wells. The computed pressure response for model one for a given set of $\underline{T}_i$, $\underline{S}_i$ can be determined from simple geometric relationships as:

$$\phi_{ij} = P_1 + \frac{q_i}{4\pi T} \sum_{k=1}^{m_i} \text{Ei}\left(-\frac{S_i r_i^2 \sin^2 \zeta_{ik}}{T_i t_j}\right) \tag{7}$$

where r_i is the radial distance from the origin to the i^{th} well and $r_i \sin \zeta_{i1}$ is defined to be r_{wi}, the bore hole radius for the i^{th} well.

Model Two Formulation

The second basic system model to be utilized takes particular advantage of the fact that pressure response is more strongly influenced by near-well properties than by those further away. In addition, early response is controlled solely by near-source properties, and as time goes on properties further out are reflected with gradually diminishing importance. Thus a useful conceptual model starts by specifying annular regions concentric with each well, such that constant effective parameter values can be used within each region. An external region, outside the last defined ring around any well, must extend to the system boundaries (or to infinity). This external region is common to all wells. Thus at some distance from any one well a set of uniform system properties, common to all subsystems, is assumed. This model geometry is illustrated in Figure 2.

Another important characteristic becomes apparent when fluid is removed at a constant rate. At any time there is a radial distance out from each well beyond which measurable reduction in pressure has not yet occurred. And over most of the region out to this point pressures fall almost uniformly with time, after an initial transient period.

Making the convenient substitution: $R = r^2$, (1) becomes:

$$T \frac{\partial}{\partial R}\left(R \frac{\partial \phi}{\partial R}\right) = \frac{S}{4} \frac{\partial \phi}{\partial t} \tag{8}$$

Initial and boundary conditions correspond to (2). The flow rate at any radial distance corresponding to R, according to Darcy's Law, is given by:

$$q(R,t) = 4\pi T R \frac{\partial \phi}{\partial R} \tag{9}$$

(q is positive for production from the well.) Substituting in (8) gives:

$$\frac{\partial q}{\partial R} = \pi S \frac{\partial \phi}{\partial t} \tag{10}$$

Integrating from the well, $\left(R = r_w^2,\ q = q_w\right)$, to a radius of disturbance (R = D) at which q(D) = 0; that is, beyond which no pressure reduction has been felt and no production of fluid has occurred:

$$\frac{-q_w}{\pi(D - r_w^2)} = \frac{1}{D - r_w^2} \int_{r_w^2}^{D} S \frac{\partial \phi}{\partial t}\, dR \tag{11}$$

Now make the simplifying assumption that the time derivative in (8) can be replaced by its average between the well and the radius of disturbance. Also, note that for all practical times $D(t) \gg r_w^2$. Substituting (11) into (8) gives:

$$T \frac{\partial}{\partial R}\left(R \frac{\partial \phi}{\partial R}\right) = - \frac{q_w}{4\pi D} \tag{12}$$

Integrating twice, since the independent variables in (12) are now separated, yields a model consisting of M annular rings concentric with the well, and with constant effective properties within each annular region. Computed transient pressure response at some radius r is given by:

$$\phi(R,t) = P_1 - \frac{q_w}{4\pi} \int_R^D \frac{dR}{TR} + \frac{q_w}{4\pi D} \int_R^D \frac{dR}{T} \tag{13a}$$

or measured at the well itself:

$$\phi(R_1,t) = P_1 + \frac{q_w}{4\pi D} \sum_{k=1}^{k=n} \frac{R_{k+1} - R_k - D \ln(R_{k+1}/R_k)}{T_k} \tag{13b}$$

where: $R_1 = r_w^2$, $R_{n+1} = D$, $1 \leqq n \leqq M$.

Equation (11) can be written in terms of the pressure drop below the original pressure at any point in the system: $\Delta\phi = (P_1 - \phi)$. Since also:

$$\Delta\phi \to 0 \text{ at } R = D \tag{14}$$

and r_w^2 and S are constants, (11) becomes:

$$\frac{q_w}{\pi} = \frac{d}{dt} \int_{r_w^2}^D (S\Delta\phi) dR \tag{15}$$

Now rewriting (13) in terms of $\Delta\phi$ and substituting into (15) gives:

$$q_w = \frac{d}{dt}\left[\frac{q_w}{4D} \int_{r_w^2}^D S\left(D \int_R^D \frac{dR}{TR} - \int_R^D \frac{dR}{T}\right)\right] dR \tag{16}$$

The parameter D for the M concentric ring model, which is the radius of disturbance, is defined as a function of time by (17), obtained by integration of (16).

$$t(D) = \frac{1}{4D}\left\{\sum_{k=1}^{k=n}\left[\frac{S_k}{2T_k}(R_{k+1} - R_k)(2D - R_{k+1} + R_k) - \frac{DS_kR_k}{T_k} \ln\left(\frac{R_{k+1}}{R_k}\right)\right]\right.$$
$$\left. + \sum_{k=1}^{n-1} S_k(R_{k+1} - R_k) \sum_{m=k+1}^{m=n}\left[\frac{D}{T_m} \ln\left(\frac{R_{m+1}}{R_m}\right) - \frac{1}{T_m}(R_{m+1} - R_m)\right]\right\} \tag{17}$$

For a system of several wells it now is necessary only to combine pressure drops due to the N sources, giving a relation for expected

system behavior of the form:

$$\phi_{ij} = P_1 - \sum_{k=1}^{k=N} \Delta\phi_k \qquad (18)$$

Pressure ϕ_{ij} is computed for the i^{th} well and the j^{th} time due to production from all sources, k = 1, 2, . . . , N. Appropriate limits on the integration represented by $\Delta\phi_k$ must be specified.

One point deserves special mention. Concentric rings around any one well, with defined effective properties, also form a part of the external region surrounding the outer ring of any other one well. Thus, to the extent that each ring contributes a measure to the overall effective average properties of the system, the parameter values obtained are not single-valued. In other words, with this model any one well "sees" the detailed variation in properties immediately surrounding a second well only as a part of the distant, system-average properties. The subsystems represented by individual wells again are related in a simple way. In this instance they are linked by the common external region, and its common set of properties.

One advantage of this model is that it can be reduced to a simple algebraic formulation, for which the number of parameters is reasonable considering the complexity of the overall problem.

IDENTIFICATION VIA DECOMPOSITION

The identification problem posed in the previous sections now will be formulated as a nonlinear programming problem for each of the two models discussed. In order to keep the problem tractable as the number of wells, N, becomes large, decomposition techniques will be employed. These techniques[5,6] allow separation of the original problem into several subproblems which are treated independently. The subproblem solutions then are modified and the subproblems are resolved. The modification of subproblems will be referred to as a second-level control. This method proceeds iteratively and is known to converge under appropriate conditions.[6]

Identification of Model One

The formulation of a decomposed nonlinear programming problem using model one proceeds as follows:

$$\min_{T_i, S_i, \alpha_i} \sum_{i=1}^{N} \sum_{j=1}^{K} (\phi_{ij} - P_{ij})^2 \qquad i=1, 2, \ldots, N \tag{19}$$

subject to the constraints given previously in (6) and (7).

$$\phi_{ij} = P_1 + \frac{q_i}{4\pi T} \sum_{k=1}^{m_i} Ei\left(-\frac{S_i r_i^2 \sin^2 \zeta_{ik}}{T_i t_j}\right) \tag{7}$$

where

$$\zeta_{ik} = \left[\frac{k}{2}\right](\alpha_i - \theta_i) + \left[\frac{k-1}{2}\right](\theta_i - \alpha_{i-1} + 2\pi\,\delta_{i1}) \tag{6}$$

$$i = 1, 2, \ldots, N \quad , \quad k = 1, 2, \ldots, m_i$$

and

$$m_i(\alpha_i - \alpha_{i-1}) = 2\pi, \qquad i = 1, 2, \ldots, N \tag{20}$$

$$\theta_i < \alpha_i < \theta_{i+1}, \qquad i = 1, 2, \ldots, N \tag{21}$$

$$T_i > 0, \qquad i = 1, 2, \ldots, N \tag{22}$$

$$S_i > 0, \qquad i = 1, 2, \ldots, N \tag{23}$$

where:

$$\alpha_o \triangleq \alpha_N - 2\pi \ , \quad \theta_{N+1} \triangleq \theta_1$$

$m_i \triangleq$ even integer.

For ease in applying known nonlinear programming algorithms, (21) to (23) are rewritten as:

$$\alpha_i \leq \theta_{i+1} - \epsilon_i, \qquad i = 1, 2, \ldots, N \tag{24}$$

$$\alpha_i \geq \theta_i + \epsilon_i, \qquad i = 1, 2, \ldots, N \tag{25}$$

$$T_i \geq \beta_i, \qquad i = 1, 2, \ldots, N \tag{26}$$

$$S_i \geq \gamma_i, \qquad i = 1, 2, \ldots, N \tag{27}$$

where ϵ_i, β_i, γ_i are arbitrarily small constants to be specified. On substitution of (7) into (19) a nonlinear programming problem results having 4N inequality constraints and 3N variables. Since at most 3N of the constraints may be active at one time, the problem is not overconstrained.

Because of the very complex nature of the objective function, convexity cannot be guaranteed; hence convergence is not guaranteed by

any of the usual minimizing algorithms. In addition, as N becomes large, the dimensionality of the computation becomes prohibitive. For these practical reasons a decomposition procedure was employed in the solution.

It is convenient to regard each wedge-shaped region containing one well as a subsystem. Since individual subsystems are coupled only by (6), it is possible to regard these subsystems as independent after making a simple change of variables. Thus the decomposed nonlinear programming problem can be written in the following form:

$$\min_{T_i, S_i, \alpha_i, \sigma_i} f = \sum_{i=1}^{N} \sum_{j=1}^{K} \left[P_1 + \frac{q_i}{4\pi T_i} \sum_{k=1}^{m_i} Ei\left(-\frac{S_i r_i^2 \sin^2 \zeta_{ik}}{T_i t_j} \right) - P_{ij} \right]^2 \tag{28}$$

where:

$$\zeta_{ik} = \left[\frac{k}{2} \right] (\alpha_i - \theta_i) + \left[\frac{k-1}{2} \right] (\theta_i - \sigma_i + 2\pi \delta_{i1}) \tag{29}$$

and such that:

$$\underline{G}_i(\alpha_i, T_i, S_i) = \begin{bmatrix} \theta_{i+1} - \epsilon_i - \alpha_i \\ \alpha_i - \theta_i - \epsilon_i \\ T_i - \beta_i \\ S_i - \gamma_i \end{bmatrix}_{i=1,2,\ldots,N} \geq 0 \tag{30}$$

and:

$$\sigma_i = \alpha_{i-1}, \quad i=1,2,\ldots,N \tag{31}$$

Now (28) is separable and can be written:

$$f(\underline{X}; \sigma) = \sum_{i=1}^{N} f_i(\underline{X}_i; \sigma_i), \quad \underline{X}_i = (T_i, S_i, \alpha_i), \quad \underline{X} = [\underline{X}_1, \underline{X}_2, \ldots \underline{X}_N] \tag{32}$$

Regarding σ_i as a known parameter in the i^{th} subsystem, one sees that the subsystems are uncoupled. Hence each subsystem optimization is performed by:

$$\min_{\underline{X}_i} f_i(\underline{X}_i; \sigma_i)$$

subject to the constraints (30). This is a standard problem in nonlinear programming.

In order to assure the minimization of the problem (28) for N wells, it remains only to satisfy the constraints (31). Defining a Lagrangian function:

$$F(\underline{X}, \underline{\lambda}, \underline{\sigma}, \underline{\mu}) = \sum_{i=1}^{N} F_i(\underline{X}_i, \underline{\lambda}_i; \sigma_{i+1}, \sigma_i, \mu_i) \tag{33}$$

where:

$F_i = f_i(\underline{X}_i, \underline{\lambda}_i ; \sigma_i) + \langle \underline{\lambda}_i, \underline{G}_i(\underline{X}_i) \rangle + \mu_i(\sigma_i - \sigma_{i+1})$

$\underline{\lambda} = [\underline{\lambda}_1, \underline{\lambda}_2, \ldots, \underline{\lambda}_N]$

$\underline{\mu}$ = N-dimensional vector of Lagrange multipliers

$\sigma_{N+1} \triangleq \sigma_1$

$\langle \underline{X}, \underline{Y} \rangle$ = inner product of $\underline{X}$ and $\underline{Y}$.

Assuming that the Kuhn-Tucker constraint qualification holds, the necessary conditions for a minimum of each subsystem are:[7]

$$\nabla_{\underline{X}_i} F_i(\underline{X}_i^*, \underline{\lambda}_i^* ; \sigma_i, \sigma_{i+1}, \mu_i) = 0 \tag{34}$$

$$\langle \nabla_{\underline{\lambda}_i} F_i(\underline{X}_i^*, \underline{\lambda}_i^* ; \sigma_i, \sigma_{i+1}, \mu_i), \underline{\lambda}_i^* \rangle = 0 \tag{35}$$

$$\lambda_{ij}^* \leq 0 ; \quad i = 1, 2 \ldots N, \quad j = 1, 2, 3, 4 \tag{36}$$

The solution of (34) to (36) proceeds for given values of the parameters σ_i, σ_{i+1}, and μ_i by any of the standard nonlinear programming algorithms. It remains to determine these parameter values by satisfying the remaining necessary conditions for a minimum of (33); namely:

$$\nabla_{\sigma_i} F = \nabla_{\sigma_i} f_i - \mu_{i-1}^* = 0, \quad i = 1, 2, \ldots, N+1 \tag{37}$$

$$\nabla_{\mu_i} F = \alpha_i - \sigma_{i+1}^* = 0, \quad i = 1, 2, \ldots, N \tag{38}$$

Thus the solution proceeds iteratively by solving (34) to (36) for given values of $\underline{\sigma}^{(k)}$, $\underline{\mu}^{(k)}$ and then using (37) and (38) to determine new values of $\underline{\sigma}^{(k+1)}$ and $\underline{\mu}^{(k+1)}$ for the $(k+1)^{st}$ iteration. The solution continues until:

$$|\underline{\sigma}^{(k+1)} - \underline{\sigma}^{(k)}| \leq \underline{\delta}_1 \tag{39}$$

and

$$|\underline{\mu}^{(k+1)} - \underline{\mu}^{(k)}| \leq \underline{\delta}_2 \tag{40}$$

where $\underline{\delta}_1$, $\underline{\delta}_2$ are specified vectors. This computation procedure is as follows:

1. Estimate α_4 and μ_i, $i = 1, 2, 3, \ldots, N$ and set $j = 1$.

2. Determine σ_j and μ_j from (37) and (38) using latest information.

3. Determine T_j, S_j, α_j, λ_j from (34) - (36) using latest information.

4. Is $j = N$?

   No - Set $j \rightarrow j + 1$ and go to 2.

   Yes - Is convergence attained according to (39), (40)?

No - Set j = 1 and go to 2.

Yes - Stop.

Convergence was found to be rapid for a 4-well example to be described in a a following section.

Identification of Model Two

Consider now the question of formulating the identification of parameters using model two as a decomposed nonlinear programming problem. In this case, the form of the problem is such that we wish to:

$$\min_{\underline{T}_i, \underline{S}_i} \sum_{i=1}^{N} \sum_{j=1}^{K} w_i [t_{ij}(D_{ij}) - t^o_{ij}]^2 \qquad i = 1, 2 \ldots N \tag{41}$$

where $\underline{T}_i$, $\underline{S}_i$ are M_i-dimensional vectors to be determined. (This change in the criterion function is easily justified because variances in P and t^o are of the same order of magnitude.) The model value of t_{ij} is dependent on D_{ij} as well as on $\underline{S}_i$ and $\underline{T}_i$. It is assumed throughout the optimization that the values of D_{ij} are known in terms of values of $\underline{S}_i$ and $\underline{T}_i$ computed previously, and in terms of observed values of pressure, P_{ij}. Since the optimization proceeds iteratively, the calculation of $D_{ij}^{(k)}$ ($\underline{S}_i^{(k-1)}$, $\underline{T}_i^{(k-1)}$) by (13), for which $\phi(R_1, t^o_{ij}) \triangleq P_{ij}$, during the k^{th} iteration imposes an additional step in the computation. This computation is by no means trivial.

It is convenient for this problem to reformulate (17) in vector-matrix notation. Thus (17) can be written:

$$t_{ij} = \langle \underline{Z}_i, C_{ij} \underline{S}_i \rangle \tag{42}$$

where C_{ij} is the $M_i \times M_i$-dimensional lower triangular matrix having the following diagonal elements $\left\{C_{ij}^{kk}\right\}$:

$$C_{ij}^{kk} = \frac{1}{4D_{ij}} \left\{ \frac{1}{2} (R_{i,k+1} - R_{ik})(2D_{ij} - R_{i,k+1} + R_{ik}) - D_{ij} R_{ik} \ell n \left(\frac{R_{i,k+1}}{R_{ik}} \right) \right\} \tag{43a}$$

Off-diagonal elements given by $\left\{C_{ij}^{mk}\right\}$, $k < m$, are:

$$C_{ij}^{mk} = \left[(R_{i,k+1} - R_{ik}) D_{ij} \ell n \frac{R_{i,m+1}}{R_{im}} - (R_{i,k+1} - R_{ik})(R_{i,m+1} - R_{im}) \right] / 4D_{ij} \tag{43b}$$

In addition, a transformation of variables:

$$\underline{Z}_i = [Z_{i1}, Z_{i2}, \ldots Z_{i,m_i}], \quad Z_{ik} = 1/T_{ik} \tag{44}$$

has been performed which does not affect the results.

The following nonlinear program can now be posed:

$$\min_{\underline{Z}_i, \underline{S}_i} \sum_{i=1}^{N} \sum_{j=1}^{K} w_i \left[< \underline{Z}_i, C_{ij} \underline{S}_i > - t_{ij}^{o} \right]^2 \qquad i=1,2 \ldots N \tag{45}$$

such that:

$$\beta_i^{-1} - Z_{ik} \geq 0, \quad S_{ik} - \gamma_i \geq 0 \tag{46}$$

and:

$$S_{i,M_i} = S_{j,M_j}, \quad \text{(all i, j)} \tag{47a}$$

$$Z_{i,M_i} = Z_{j,M_j}, \quad \text{(all i, j)} \tag{47b}$$

In order to satisfy (47) it is necessary and sufficient that:

$$S_{i,M_i} = S_{i-1,M_{i-1}}, \quad i = 1,2 \ldots N \tag{48a}$$

$$Z_{i,M_i} = Z_{i-1,M_{i-1}}, \quad i = 1,2 \ldots N \tag{48b}$$

where:

$$S_o \triangleq S_N, \quad Z_o \triangleq Z_N$$

In the above relations M_i is the number of concentric rings for the i^{th} well and hence the dimension of $\underline{S}_i$ and $\underline{Z}_i$. The number of independent variables is then:

$$Q = 2 \sum_{i=1}^{N} M_i \tag{49}$$

In the problem posed by (45) to (48), the wells considered as subsystems are independent except for (48), a relation which guarantees that all wells will have common average properties in the outermost ring.

If we temporarily consider $\underline{S}_{i-1}$ and $\underline{Z}_{i-1}$ as fixed parameters in the i^{th} subsystem, we can define a Lagrangian:

$$F(\underline{Z}, \underline{S}, \underline{\lambda}, \underline{\nu}, \underline{\mu}) = \sum_{i=1}^{N} F_i(\underline{Z}_i, \underline{S}_i, \underline{\lambda}_i ; \underline{S}_{i-1}, \nu_i, \mu_i) \tag{50}$$

where:

$$F_i = \sum_{j=1}^{K} w_i [< \underline{Z}_i, C_{ij} \underline{S}_i > - t_{ij}^{o}]^2 + < \underline{\lambda}_i, \underline{G}_i > + \nu_i [< \underline{\Gamma}_i, \underline{Z}_i > - < \underline{\Gamma}_{i-1}, \underline{Z}_{i-1} >] + \mu_i [< \underline{\Gamma}_i, \underline{S}_i > - < \underline{\Gamma}_{i-1}, \underline{S}_{i-1} >] \tag{51}$$

and:

$$\underline{\Gamma}_i^T \triangleq [0, 0, \ldots 0, 1], \quad \dim \underline{\Gamma}_i = M_i$$

$$\underline{G}_i(\underline{Z}_i, \underline{S}_i) = \begin{bmatrix} \beta_i^{-1} \underline{u}_i - \underline{Z}_i \\ \underline{S}_i - \underline{\gamma}_i \end{bmatrix} \geq 0$$

$\underline{u}_i$ = M_i-dimensional unit vector

$\underline{Z} = [\underline{Z}_1, \underline{Z}_2, \ldots \underline{Z}_N]$

$\underline{S} = [\underline{S}_1, \underline{S}_2, \ldots \underline{S}_N]$

$\underline{\lambda}_i, \nu_i, \mu_i$ are Lagrange multipliers of dimension M_i, 1, 1, respectively.

Assuming that the Kuhn-Tucker constraint qualification holds, the necessary conditions for each independent subsystem are given by:

$$\nabla_{\underline{Z}_i} F_i(\underline{Z}_i^*, \underline{S}_i^*, \underline{\lambda}_i^*; \underline{S}_{i-1}, \nu_i, \mu_i) = 0 \tag{52}$$

$$\nabla_{\underline{S}_i} F_i(\underline{Z}_i^*, \underline{S}_i^*, \underline{\lambda}_i^*; \underline{S}_{i-1}, \nu_i, \mu_i) = 0 \tag{53}$$

$$< \nabla_{\underline{\lambda}_i} F_i(\underline{Z}_i^*, \underline{S}_i^*, \underline{\lambda}_i^*; \underline{S}_{i-1}, \nu_i, \mu_i), \underline{\lambda}_i^* > = 0 \tag{54}$$

$$\lambda_{ij}^* \leq 0, \quad j = 1, 2 \ldots Q, \quad i = 1, 2 \ldots N \tag{55}$$

For F_i given by (51), conditions (52) and (53) can be solved explicitly for $\underline{Z}_i$ and $\underline{S}_i$, respectively. Thus:

$$\underline{Z}_i = \left\{ \sum_{j=1}^{K} C_{ij} \underline{S}_i > < \underline{S}_i C_{ij}^T \right\}^{-1} \left[\left\{ \sum_{j=1}^{K} t_{ij}^o C_{ij} \right\} \underline{S}_i + \frac{1}{2w_i} (\underline{\lambda}_i^1 - \nu_i \underline{\Gamma}_i) \right] \tag{56}$$

$$\underline{S}_i = \left\{ \sum_{j=1}^{K} C_{ij}^T \underline{Z}_i > < \underline{Z}_i C_{ij} \right\}^{-1} \left[\left\{ \sum_{j=1}^{K} t_{ij}^o C_{ij}^T \right\} \underline{Z}_i - \frac{1}{2w_i} (\underline{\lambda}_i^2 + \mu_i \underline{\Gamma}_i) \right] \tag{57}$$

where

$$\underline{\lambda}_i \triangleq [\underline{\lambda}_i^1, \underline{\lambda}_i^2], \quad \dim \underline{\lambda}_i^1 = \dim \underline{\lambda}_i^2 = M_i$$

or

$$\underline{Z}_i = \phi(\underline{S}_i, \underline{\lambda}_i^1; \nu_i)$$

$$\underline{S}_i = \psi(\underline{Z}_i, \underline{\lambda}_i^2; \mu_i)$$

The remaining necessary conditions to insure the coupling between subsystems are:

$$\nabla_{\mu_i} F = S_{i,M_i} - S_{i-1,M_{i-1}} = 0 \tag{58a}$$

$$\nabla_{\nu_i} F = Z_{i,M_i} - Z_{i-1,M_{i-1}} = 0 \tag{58b}$$

$$\nabla_{\underline{Z}_{i-1}} F = \nabla_{\underline{S}_{i-1}} F = 0 \tag{58c}$$

where (58c) becomes

$$\nu_i = < \underline{\Gamma}_{i-1}, \nabla_{\underline{Z}_{i-1}} F_{i-1} > \tag{59a}$$

$$\mu_i = < \underline{\Gamma}_{i-1}, \nabla_{\underline{S}_{i-1}} F_{i-1} > \tag{59b}$$

The optimization calculation proceeds essentially as with model one except that the subsystem problems are each of dimension $2M_i$ instead of dimension 3. In addition, the D_{ij} calculation is required. The question of convergence for this procedure using model two has not been examined.

The subsystem optimization to determine $\underline{S}_i^{(k)}$ and $\underline{Z}_i^{(k)}$ for the k^{th} iteration may be somewhat easier using model two because each vector can be expressed explicitly in terms of the other. The solution of (56)

and (57), however, still must be obtained iteratively. The iterative procedure is analogous to that discussed previously for model 1. Using model one requires initial estimates of $\alpha_4^{(0)}$ and $\mu_i^{(0)}$, i = 1, 2, 3, 4; however, when model two is used only $\mu_1^{(0)}$ and $\nu_1^{(0)}$ must be estimated initially.

Computational Example

An example problem has been solved using model one with four wells and seven pressure observations for each well. Thus the number of variables in the optimization is 12; 3 for each subsystem. The aquifer data and pressure observations are summarized in Tables 1 and 2. A direct search technique was used for the subsystem optimization. A coarse grid was first employed and then refined until suitable convergence resulted. This procedure was fast, requiring only 8 seconds of IBM 360/75 computer time per iteration, and produced globally optimal results. By comparison, straightforward discretization[3] might require an order of magnitude greater time for similar results.

The overall procedure for all four wells converged in 6 iterations to the values given in Table 3. The rate of convergence for a typical well is illustrated in Figure 3 and the monotonic decrease in the objective function is shown in Figure 4.

CONCLUSIONS

In obtaining optimal solutions to physical problems, one is faced with two challenges; modeling and optimization. These can not be considered independently because the former greatly affects the latter.

For problems of high dimensionality model complexity generally must increase in order to obtain increased accuracy. A highly accurate model may require complexity and dimensionality of the optimization problem such as to make problem solution unrealistic, if not impossible.

In this paper two fairly complex models for a reservoir system have been formulated, in each case anticipating problems which arise with optimization problems of high dimension. A decomposition

procedure was used to limit the magnitude of the computational burden regardless of the number of wells considered. This result requires an iterative solution which may still be formidable. The method was tested by an example and found to have rapid convergence. Computation speed was increased by the fact that no computation of response surface derivatives was required.

REFERENCES

1. De Wiest, Roger J.M., Geohydrology, Wiley, New York (1965).
2. Rowan, G. and M.W. Clegg, "An Approximate Method for Transient Radial Flow," Soc. Pet. Engr. J. 2, 225-256 (1962).
3. Jahns, Hans O., "A Rapid Method for Obtaining a Two-Dimensional Reservoir Description from Field Pressure Data," Soc. Pet. Engr. J. 6, 315-327 (1966).
4. Haimes, Yacov Y., Richard L. Perrine and David A. Wismer, Identification of Aquifer Parameters by Decomposition and Multilevel Optimization, Department of Engineering, University of California, Los Angeles, California, Report No. 67-63 (March 1968).
5. Lasdon, L.S. and James D. Schoeffler, "A Multilevel Technique for Optimization," ISA Trans. 5, 175-183 (April 1966).
6. Wismer, D.A., Ed., Optimization Methods for Large-Scale Systems, McGraw-Hill, New York, (in preparation).
7. Hadley, G., Nonlinear and Dynamic Programming, Addison-Wesley, Reading, Massachusetts (1964), p. 192.

TABLE 1 - FIELD DATA

	Well 1	Well 2	Well 3	Well 4
p_1(psi)	3000	3000	3000	3000
r_i(ft)	1000	2000	2000	3000
r_{wi}(ft)	1	1	1	1
θ_i(degrees)	22.5	112.5	157.5	292.5
q_i(gal/day)	42000	126000	42000	252000

TABLE 2 - PRESSURE OBSERVATION DATA

T_j (days)	P_{1j} (psi)	P_{2j} (psi)	P_{3j} (psi)	P_{4j} (psi)
30	2798	2810	2859	2811
45	2776	2804	2850	2805
60	2758	2798	2843	2802
90	2733	2789	2829	2798
180	2687	2769	2800	2786
360	2640	2743	2763	2774
720	2591	2712	2721	2760

TABLE 3 - OPTIMAL AQUIFER PARAMETERS

	Well 1	Well 2	Well 3	Well 4
T_i^*	42	95	48	366
S_i^*	0.004	0.010	0.002	0.00001
α_i^*	42	132	177	357

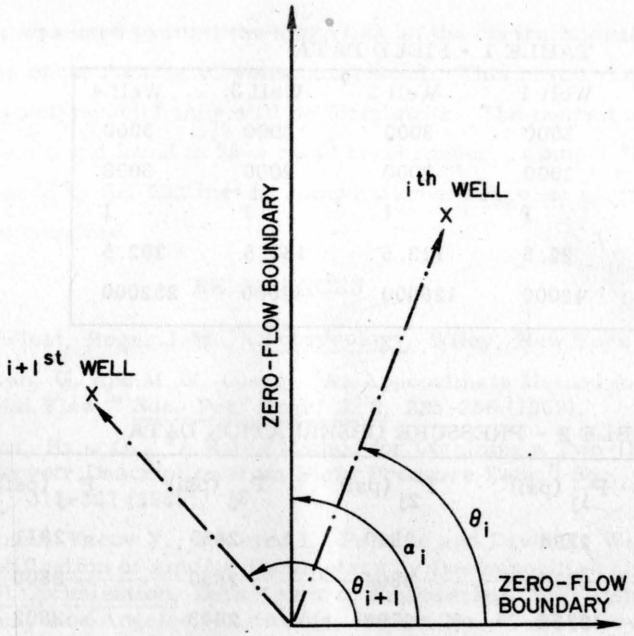


Figure 1 Model One Geometry

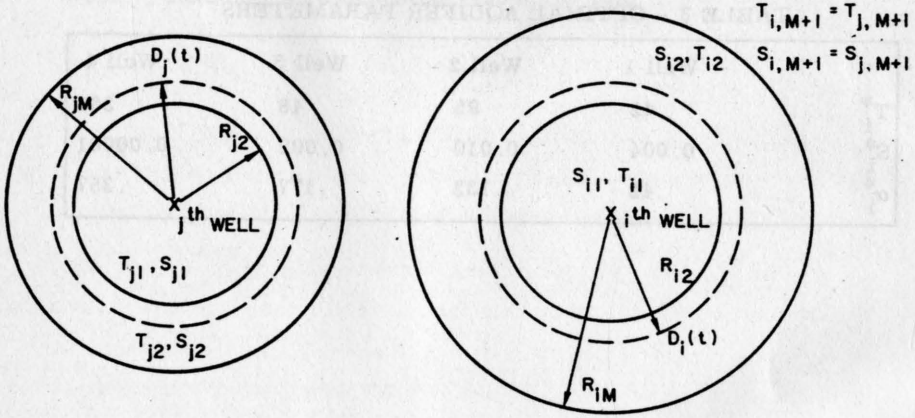


Figure 2 Model Two Geometry

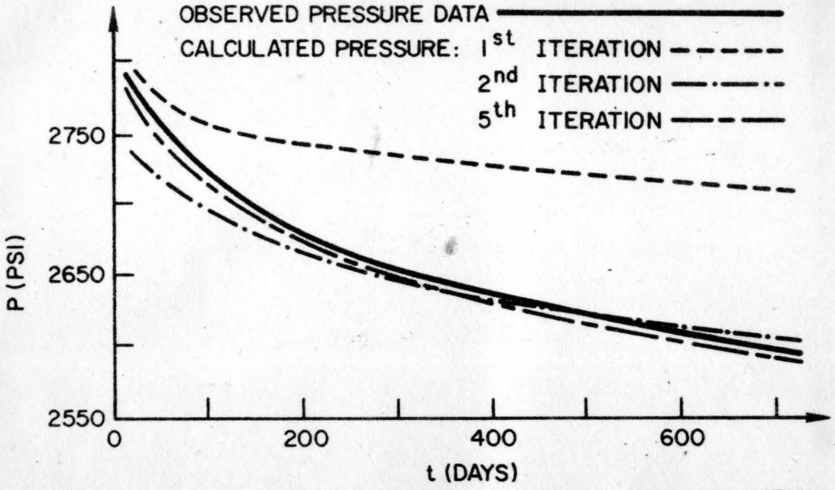


Figure 3 Convergence of Calculated Pressure Toward Observed Pressure

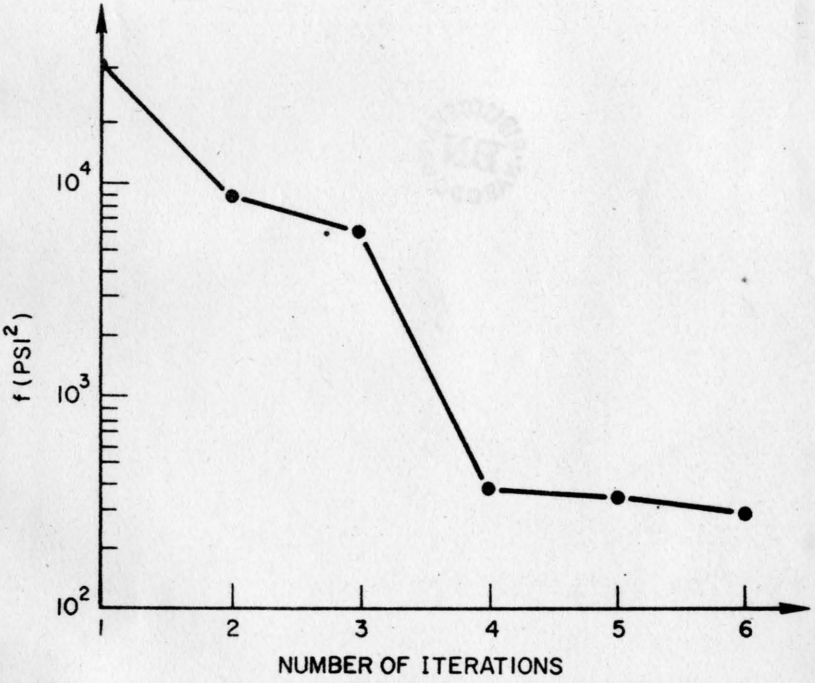


Figure 4 Objective Function Convergence