Real-Time Identification

Part II

Wieslander, Johan

1971

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

REAL-TIME IDENTIFICATION
PART II

JOHAN WIESLANDER

REPORT 7111 SEPTEMBER 1971
LUND INSTITUTE OF TECHNOLOGY
DIVISION OF AUTOMATIC CONTROL

REAL-TIME IDENTIFICATION.[†]

Part II

J. Wieslander

ABSTRACT.

With real-time identification is understood the estimation from measurements of time-varying parameters in a model of an unknown system. In this report two problems in this field are discussed. The first one is the question of convergence. This has been studied in a simple continuous time case by simulation on a hybrid computer. Results in the form of distributions of the interesting variables are shown. The other problem is that of determining suitable parameters to use in the real-time identification algorithm. A method to do this by off-line calculations using representative plant data has been implemented. Some numerical results are shown.

[†] This work was supported by the Swedish Board for Technical Development under Contract 70-337/U270.

TABLE OF CONTENTS

0. INTRODUCTION.

This report deals with two problems in real-time identification. With real-time identification is meant the estimation from measurements of time-varying parameters in a model of e.g. a physical or chemical process.

In part I of this report (ref. [4]) three different though intimately related schemes for real-time identification were discussed. They were shown, by some examples, to possess good performance even when the parameters varied rapidly with time. The most general of the three was the Kalman filter type identification in that a priori knowledge of the rate of parameter variations is used in the algorithm. This is very favourable if such information is available. Now, this is not a very likely situation, so a method to determine, from measurements, these properties of the unknown parameters is needed. This problem is discussed in Section 2.

Another problem that was not discussed in part I is that of convergence of the algorithms. As is demonstrated in App. A convergence in the usual sense cannot be expected. In an early attempt to get a feeling of what may happen with a real-time identification algorithm, a simple continuous time example was studied. The result is shown in Section 1. The conclusion drawn from the study was that that approach to the problem did not seem profitable. Anyhow, qualitative knowledge was obtained.

1. A STUDY OF A CONTINUOUS TIME REAL-TIME IDENTIFIER WITH ONE UNKNOWN COEFFICIENT.

A question which often arises in connection with identification routines is that of convergence. Of course with real-time identification there is no such thing as convergence (see App. A). What can be hoped for is e.g. that the estimation error covariance is bounded. In the discrete time case the equation for P, the estimation error covariance matrix, is nonlinear with coefficients that are nonstationary stochastic processes. Hence a study of the behaviour of the algorithm in the general case is extremely difficult. Therefore in order to gain some insight in the behaviour of real-time algorithms a simple continuous time example has been studied.

In 1.1 and 1.2 the identified system and the identifier itself are presented. The question of the behaviour of the identifier is then posed in 1.3. To solve the problem analytically, one has to find a solution to the Fokker-Planck equation, 1.4. The method choosen was, however, to simulate the equations on a hybrid computer, 1.5. The results are discussed in 1.6. Using the approximations mentioned in 1.3 an attempt to compute the distribution of the error covariance is then done in 1.7. The validity of the approximations is discussed in 1.8. One of the stochastic differential equations used in the simulations may become unstable for certain parameter values. An estimate of the stability boundary is given in 1.9.

1.1. Formulation of the Estimation Problem.

The following equation is given

$$dy = -(\alpha+x)y\,dt + dv \tag{1.11}$$

where α is a known constant and v is a Wiener process with incremental covariance $R_2 dt$.

The unknown parameter x is given by

$$dx = -\beta x\,dt + de \tag{1.12}$$

where β is a known constant and e is a Wiener process with incremental covariance $R_1 dt$.

The task is now to estimate x(t) when $\{y(s),\ t_0 \leq s \leq t\} = \mathcal{Y}_t$ has been observed.

1.2. Solution of the Estimation Problem.

The estimate that minimizes the expected loss $E\,L(\theta-\hat{\theta})$ where L is a symmetric nondecreasing function is given by the conditional mean $E\{x(t)|\mathcal{Y}_t\}$ provided that the conditional distribution is symmetric around the mean and nonincreasing for $\sigma \geq m$ (see ref. [5]). The conditional distribution of x(t) given $\mathcal{Y}_t$ is given by the Kalman filter, i.e. x(t) is normal $(\hat{x},P)$ where

$$d\hat{x} = -\beta\hat{x}\,dt + K(t)[dy + (\alpha+\hat{x})y\,dt] \tag{1.21}$$

$$\frac{dP}{dt} = -2\beta P + R_1 - P^2y^2/R_2 \tag{1.22}$$

K(t) is the filter gain and depends on P.

These equations are analogous to those of (2.14).

1.3. Question.

We want to know how good the estimate is, that is what is the expected value of the error variance $E\,\tilde{x}^2$, where the estimation error is defined by

$$\tilde{x} = x - \hat{x}$$

We thus have

$$E\,\tilde{x}^2 = E_{\mathcal{Y}}\left\{E\tilde{x}^2 \,\middle|\, \mathcal{Y}_t\right\} = E\{P\} \qquad (1.31)$$

and we are faced with the problem of finding the distribution of P. It is easily seen that P must be greater than zero and less than $R_1/2\beta$. In this case it is thus possible to give an upper bound on the estimation error covariance, i.e. the estimates cannot diverge in the mean square sense.

Before more thorough discussions, the equations may be normalized, i.e. the time unit is chosen so that $\beta = 1$ and the scale of x is chosen so that $E\,x^2 = 1$ and consequently $R_1 = 2$. Then we have

$$r_x(\tau) = e^{-|\tau|}$$

Hence

$$dx = -x\,dt + de \qquad (1.32a)$$

$$dy = -(\alpha+x)y\,dt + dv \qquad (1.32b)$$

$$\frac{dP}{dt} = -2P + 2 - P^2y^2/R_2 \qquad (1.32c)$$

where $E\,de(de)^T = 2dt$ and $E\,dvdv = R_2dt$.

It was judged unrealistic to try to derive the simultaneous distribution of x, y and P. Therefore it was decided to try a simpler way. If the distribution of y was available the P distribution might be calculated by a "fast" or a "slow" approximation.

Slow approximation: Assume that the process y is so slow that $\frac{dP}{dt}$ is close to zero. The distribution of P could then be computed from that of y using

$$- 2P + 2 - \frac{P^2 y^2}{R_2} = 0 \qquad (1.33)$$

Fast approximation: Assume that the process y is fast compared to P. y could then be replaced by white noise.

1.4. The Fokker-Planck Equation.

The distribution of y may be calculated by solving the so called Fokker-Planck equation for the system (1.32). In the case of two coupled stochastic differential equations driven by white noise the system equations may be written (ref. [3])

$$\begin{aligned} z_1 &= g_1(z_1, z_2) + w_1 \\ z_2 &= g_2(z_1, z_2) + w_2 \end{aligned} \qquad (1.41)$$

where w_i are random variables with $R_{ij}\delta_{ts} = E\, w_i(t)w_j(s)$. The joint probability function $f(z_1, z_2, t)$ then satisfies the partial differential equation

$$\frac{\partial f}{\partial t} = - \frac{\partial}{\partial z_1}(g_1 f) - \frac{\partial}{\partial z_2}(g_2 f) + \frac{1}{2} \sum_{i,j=1}^{2} R_{ij} \frac{\partial^2 f}{\partial z_i \partial z_j} \qquad (1.42)$$

where the probability function f is conditional with respect to the initial values

$$f(z,t) = f(z,t|z_0,t_0)$$

If we let z_1 correspond to x and z_2 to y we have

$$\dot{x} = -x + w_1 \qquad R_{11} = 2 \qquad R_{22} = R_2$$

$$\dot{y} = -(\alpha+x)y + w_2 \qquad R_{12} = R_{21} = 0$$

$$\frac{\partial f}{\partial t} = -\frac{\partial}{\partial x}(-xf) - \frac{\partial}{\partial y}\left(-(\alpha+x)yf\right) + \frac{1}{2}\left\{2\,\frac{\partial^2 f}{\partial x^2} + R_2\,\frac{\partial^2 f}{\partial y^2}\right\}$$

$$\frac{\partial f}{\partial t} = f + x\,\frac{\partial f}{\partial x} + (\alpha+x)f + (\alpha+x)y\,\frac{\partial f}{\partial y} + \frac{\partial^2 f}{\partial x^2} + \frac{R_2}{2}\,\frac{\partial^2 f}{\partial y^2} \qquad (1.43)$$

We are interested in the stationary distribution i.e. $\frac{\partial f}{\partial t} = 0$. The problem thus is to solve the nonlinear partial differential equation

$$(1+\alpha+x)f + x\,\frac{\partial f}{\partial x} + (\alpha+x)y\,\frac{\partial f}{\partial y} + \frac{\partial^2 f}{\partial x^2} + \frac{R_2}{2}\,\frac{\partial^2 f}{\partial y^2} = 0 \qquad (1.44)$$

The author does not know of any analytic solution to equations of this kind. What remains is therefore to seek some numeric solution. There are two possible methods available:

a) One method is to derive a difference approximation of (1.44) and then by an iterative technique compute the values of f in a two-dimensional grid.

b) The other way is to use a Monte-Carlo method, i.e. to simulate the stochastic differential equations

(1.32b) and (1.32c) on a hybrid computer and measure the distributions of x and y.

The Monte-Carlo method was chosen because the next step, calculation of the distribution of P in this way is easily done simultaneously with the simulation of x and y.

1.5. Hybrid Simulation.

The simulations were run at FOA, Stockholm, on an EAI 8945 hybrid computer.

The equations (1.32) were simulated on the analog part using a time scale factor of 100.

The equations were simulated typically 1000 times and the values of x, y and P were measured at the end of each single run. Between runs the analog computer was left in hold mode so that each run began with a new initial condition. The run length was 10 seconds in problem time, i.e. the equations were simulated at least a couple of time-constants between measurements. The time scale factor of 100 permitted approximately 10 runs a second, thus enabling 1000 runs to be made in about two minutes. Some extra time is allowed for bookkeeping in the digital computer. The two white gaussian noises were generated using a PRBS signal from two 20-bit shift registers in the logical unit of the analog computer. The PRBS signal was fed to a first order filter (an integrator with feedback). The filter is thus summing the pulses of the digital signal. The filter time constant was chosen small compared to the time constant of the process but so long that approximately ten PRBS pulses were summed in the filter, thus producing

an output from the filter with a binomial distribution which sufficiently well approximates the normal distribution. In order that the two noises should be independent the clock signal to one of the shift registers was interrupted for a short period of time each time the analog computer was forced into hold mode. The programming on the digital computer was done in the programming language HOI (Hybrid Operations Interpreter).

The program tested the mode of the analog computer through a sense line and if in hold mode measured the signals y and P. After the sum and sum of squares as well as class interval counts had been updated the analog computer was forced back into operate mode using a control line. The operation of the control line also reinitialized a clock which after 0.1 seconds forced the computer back into hold mode. A block diagram of the set-up is shown in Figure 1. When the predetermined number of runs had been completed the mean and standard deviation was computed and printed out as well as the number of measurements in each class interval. In this way histograms of the variables y and P are obtained.

1.6. Results from the Simulations.

For easy reference, the simulated equations are repeated:

$$dx = - x\,dt + de$$

$$dy = - (\alpha+x)y\,dt + dv \qquad (1.61) =$$

$$\frac{dP}{dt} = - 2P + 2 - \frac{P^2y^2}{R_2} \qquad = (1.32)$$

The distribution of x is known; x is normal (0,1) with covariance function

$$r_x(\tau) = e^{-|\tau|}$$

The parameters α and R_2 where $E(dv)^2 = R_2 dt$ are to be varied between experiments. The distributions of y and P are not known but it is clear that the distribution of y will be close to normal when α is large compared to $E\,x^2 = 1$. If on the other hand α is small the y equation will be unstable (cf. Section 1.9). This is reflected in the choice of R_2; $\alpha = 1.5$ requires a much smaller value of R_2 than $\alpha = 4.0$ in order to avoid overloading of the amplifiers if the amplitude scaling is unchanged. Different values of R_2 does not alter the distribution of P as may be seen from the equations (1.32b) and (1.32c). The distribution of y and P obtained in runs with $\alpha = 4.0$ and different values of R_2 are shown in Diag. 1 for y and in Diag. 2 for P. Notice that the P distribution is unchanged for different values of R_2 as it should. The distribution of y changes with R_2, i.e. the amplitude increases with R_2, and is relatively close to normal as shown in Diag. 3. Two sample functions are shown in Fig. 2. The one depicting the function P shows the typical behaviour: close to one when y is small indicating poor estimates of x, but if y is considerably different from zero permitting better estimates P goes down. y resembles normal random noise.

The next simulation shown is with $\alpha = 1.5$. Due to the decreased stability R_2 must have a much smaller value: $R_2 = 0.1$. The distributions of y and P are shown in Diag. 4 and 5. As may be seen in Diag. 4 the distribution of y has very broad tails, which, of course, depends on the fact that the y equation with this va-

lue of α for certain periods of time becomes unstable due to the variations in x. The broad tails is very clearly demonstrated in Diag. 6. A sample function of y is shown in Fig. 3. Also the distribution of P is different from the previous case. Values considerably smaller than 1, the a priori estimation error variance, are now much more common, indicating better estimates.

1.7. An Attempt to Compute the P Distribution from the y Distribution.

As mentioned earlier a possible method of obtaining the P distribution when the y distribution is known would be to use the slow approximation, i.e. to assume $\frac{dP}{dt} = 0$ in (1.32c). In this case we have

$$0 = -2P + 2 - \frac{P^2 y^2}{R_2} \qquad (1.71) = (1.33)$$

which gives

$$y^* = h(p) = \frac{\sqrt{2R_2(1-P)}}{P} \qquad (1.72)$$

where y^* is an auxiliary variable used to take care of the square in (1.71). y^* is defined as follows:

$$y^* = \begin{cases} y & y \geq 0 \\ -y & y < 0 \end{cases}$$

As the distribution of y should be symmetric the distribution of y^* is readily obtained. Then we have to

calculate the distribution of P using (1.72). We have

$$y^{*} = h(P)$$

and

$$F_{y^{*}}(y^{*}) = P(\eta \leq y^{*})$$

$$F_{p}(P) = P(\pi \leq P)$$

$$F_{p}(P) = P(\pi \leq P) = P\left(h^{-1}(\eta) \leq P\right) =$$

$$= P\left(\eta \leq h(P)\right) = F_{y^{*}}\left(h(P)\right)$$

$$f_{p}(P) = \frac{d}{dP} F_{p}(P) = \frac{d}{dP} F_{y^{*}}\left(h(P)\right) =$$

$$= f_{y^{*}}\left(h(P)\right)h'(P) = f_{y^{*}}(y^{*})h'(P) \qquad (1.73)$$

$$h'(P) = \frac{R_2(2-P)}{P^2\sqrt{2R_2(1-P)}} \qquad (1.74)$$

Thus to calculate the value of the frequency function of the P distribution in a point P using the slow approximation, y^{*} is computed using (1.72) and $f_{y^{*}}(y^{*})$ is read from a diagram, derived from Diagrams 1 or 4 by smoothing and a change of scale to obtain the frequency function, then $f_{p}(P)$ is evaluated using (1.73) and (1.74).

In this way the curves indicated "computed" in Diag. 7 and 8 have been obtained.

As can be seen there is little correspondence, if any, between the measured distribution and the computed one. The conclusion to be drawn is that the slow approximation is not valid.

1.8. Discussion of the Approximations.

Now that measurements have been made some typical values of y and P are available. Some rough estimation of the time behaviour of the equations could then be done through linearization around these points. The time constants could then be compared. Linearizing (1.32b) and (1.32c) around (y_0, P_0) we have

$$\dot{y} = -\alpha y \qquad (E\ x = 0)$$

$$\dot{P} = \left[-2 - \frac{2P_0 y_0^2}{R_2} \right] P - \frac{2P_0^2 y_0}{R_2}\, y \tag{1.81}$$

By assumption we have (cf. (1.72))

$$y_0 = \frac{\sqrt{2R_2(1-P_0)}}{P_0} \tag{1.82}$$

Hence the time constants are

$$T_1 = \frac{1}{\alpha}$$

$$T_2 = \frac{1}{2 + \dfrac{2P_0 y_0^2}{R_2}} = \frac{P_0}{4 - 2P_0}$$

In the case of $\alpha = 4.0$ P is seldom less than 0.8. If we evaluate the time constants for P = 0.8 we find $T_1 = 0.25$ and $T_2 = 0.33$ which contradicts the assumption of the slow approximation. For $\alpha = 1.5$ a value of P corresponding to the one in the example above is P = 0.55. In this case the time constants are $T_1 = 0.67$

and $T_2 = 0.19$. These values are more in agreement with the assumption, also the region $P \lesssim 0.6$ with $\alpha = 1.5$ is where the correspondence between the computed and the measured distribution in Diagram 8 is acceptable. This discussion confirms that the slow approximation is not valid. On the other hand, from inspection of Fig. 2 it does not seem likely that the fast approximation should be valid either. In that case the process y should look like white noise compared to the process P which is not the case. Unfortunately it seems unavoidable to treat the full 3-dimensional problem if the distribution of P is to be computed. This has not been done.

1.9. Stability of y Equation.

As was mentioned in a previous section the parameter α of equation (1.32b) had to be chosen sufficiently large in order that the amplifiers should not be overloaded for a given value of R_2. We are going to establish a stability boundary for α when there is no input, i.e. we are going to study

$$dy = - (\alpha+x)y\, dt \tag{1.91}$$

The following stability criterion is going to be used:

Let the system

$$\dot{x} = \{A + F(t)\}x \tag{1.92}$$

be given where A is a matrix with constant elements and F a matrix whose elements are either zero or stochastic processes which are stationary, ergodic, measurable, and continuous with probability one.

For the constant part

$$\dot{x} = A \cdot x \tag{1.93}$$

there exist numbers a and b so that

$$||x|| \leq b||x_0||e^{-at} \tag{1.94}$$

According to Kozin, ref. [2], we then have the theorem:

> If (1.93) is asymptotically stable and $E||F(t)||$ exists then if
>
> $$E||F(t)|| < a/b \tag{1.95}$$
>
> the system (1.92) is asymptotically stable in the large with probability one.
>
> The proof of the theorem is mainly an application of the Gronwall lemma.

In our case we have $a = \alpha$ and $b = 1$. $F(t)$ is replaced by the gaussian process x which meets the requirements of the theorem. We then have to evaluate

$$E|x| = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |x| e^{-\frac{x^2}{2}} dx =$$

$$= \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} x\, e^{-\frac{x^2}{2}} dx = \sqrt{\frac{2}{\pi}}$$

We thus find that (1.91) is asymptotically stable in the large with probability one if

$$\alpha > \sqrt{\frac{2}{\pi}} = 0.7979$$

This value seems quite reasonable compared to the experience from the simulations. The smallest value of α in the experiments was $\alpha = 1.25$ with $R_2 = 0.1$ which gave overload indication approximately every 10:th run.

1.10. Comparison between digital and hybrid simulation.

One may ask if the use of a hybrid computer gave a significant saving in execution time compared to a simulation on a fast digital computer.

The following estimation can be done: A thousand runs of 10 seconds in problemtime each is to be done. In the integration of the equation the time increment is chosen an order of magnitude less than the estimated time constant (cf. page 12). For simplicity say 0.01 seconds. Thus for a complete simulation 10^6 points in time have to be evaluated.

On UNIVAC 1108 the generation of a gaussian random number takes 350 μs while the evaluation of the expression in (1.32c) takes 320 μs. A conservative estimate of the time for a complete simulation would then be: $2 * 350 + 320 \approx 1000$ seconds, i.e. 16 minutes. This time should be compared to the two minutes for the hybrid simulation.

It is interesting to note that in this simple first order problem the full advantage of the parallell processing capability of the hybrid computer is not utilized. The difference in a more complex problem would be even greater.

2. ON THE ESTIMATION OF THE PARAMETER NOISE.

As has been discussed in a previous report the least squares estimation of the parameters of time varying linear systems may be thought of as operating a Kalman filter with the parameters as state variables. To design a Kalman filter one needs knowledge of the covariance of the parameter noise. This information is not generally available. In this section an implementation of a method described in the literature to compute the characteristics of the parameter noise is discussed. In subsection 1, the Kalman filter equations as used in the identification scheme are displayed. In subsection 2, the loss function used in the calculations is formed and in subsection 3 some numerical aspects are considered. Some examples on simulated data are displayed in subsection 4.

2.1. The Model and the Identifier.

As the model and the identification algorithm already has been discussed in ref. [4], only a brief discussion of the matter is included. The model is the so called equation error model which we write in the form

$$y(t) = \varphi(t-1)\theta(t) + e(t) \qquad (2.11)$$

where $y(t)$ is the output at time t and $\theta(t)$ the parameter vector at time t. $\varphi(t-1)$ is a vector of old input and output values. φ is sometimes called the phase variable vector. We have

$$\begin{cases} \varphi(t-1) = [-y(t-1)\ u(t-1)\ -y(t-2)\ u(t-2)\ \ldots] \\ \theta(t) = [a_1(t)\ b_1(t)\ a_2(t)\ b_2(t)\ \ldots]^T \end{cases} \tag{2.12}$$

If we assume that

$$E\ e(t)e(s) = R_2\delta_{ts}$$

and $\theta(t)$ is governed by the equation

$$\begin{cases} \theta(t+1) = \theta(t) + v(t) \\ \text{with} \\ E\ v(t)v(s)^T = R_1\delta_{ts} \end{cases} \tag{2.13}$$

where e and v are gaussian uncorrelated variables with zero mean a Kalman estimation of θ is given by

$$\begin{cases} \hat{\theta}(t+1) = \hat{\theta}(t) + K(t)\{y(t) - \varphi(t-1)\hat{\theta}(t)\} \\ K(t) = P(t)\varphi(t-1)^T \cdot \\ \qquad \cdot \left\{R_2 + \varphi(t-1)P(t)\varphi(t-1)^T\right\}^{-1} \\ P(t+1) = P(t) + R_1 - K(t) \cdot \\ \qquad \cdot \left\{R_2 + \varphi(t-1)P(t)\varphi(t-1)^T\right\}K(t)^T \end{cases} \tag{2.14}$$

Refer to ref. [4] for a discussion on this as well as a comparison with other algorithms of the same structure. These equations may be scaled so that the disturbance e has unit variance. Scaled quantities are denoted with a prime, e.g. R_1'.

$$\begin{cases} \hat{\theta}(t+1) = \hat{\theta}(t) + K'(t)\{y(t) - \varphi(t-1)\hat{\theta}(t)\} \\ K'(t) = P'(t)\varphi(t-1)^T \cdot \\ \qquad \cdot \left\{1 + \varphi(t-1)P'(t)\varphi(t-1)^T\right\}^{-1} \\ P'(t+1) = P'(t) + R_1' - K'(t) \cdot \\ \qquad \cdot \left\{1 + \varphi(t-1)P'(t)\varphi(t-1)^T\right\}K'(t) \end{cases} \tag{2.15}$$

It is in this form the algorithm is used in a computer program.

2.2. The Loss Function.

From ref. [1] we have the logarithm of the likelihood function of R_1 and R_2 given the measurements:

$$2 \cdot L\left[R_1, R_2; \mathcal{Y}_t, \mathcal{U}_t\right] = - \sum_{t=1}^{N} \log \gamma(t)^2 - \sum_{t=1}^{N} \xi(t)^2 \tag{2.21}$$

where $\gamma(t)^2$ is the covariance of the residual

$$\varepsilon(t) = y(t) - \varphi(t-1)\hat{\theta}(t) \tag{2.22}$$

and $\xi(t)$ is the normalized residual (i.e. standard deviation equal to one)

$$\xi(t) = \varepsilon(t)/\gamma(t) \tag{2.23}$$

$\mathcal{Y}_t(\mathcal{U}_t)$ denotes all available output (input) signals up to and including time t.

Now, we know that the covariance of the residual in the Kalman filter is

$$\gamma(t)^2 = E\{\varepsilon(t)\varepsilon(t)\} = R_2 + \varphi(t-1)P(t)\varphi(t-1)^T \quad (2.24)$$

A scaling so that covariances are measured with R_2 as unit is then introduced (see previous section). Scaled quantities are denoted by a prime. Hence

$$R_1' = \frac{1}{R_2} R_1; \qquad P'(t) = \frac{1}{R_2} P(t)$$

We thus have the scaled covariance of the residual

$$\gamma'(t)^2 = \frac{1}{R_2} \gamma(t)^2 = 1 + \varphi(t-1)P'(t)\varphi(t-1)^T \quad (2.25)$$

Hence the loss function becomes

$$2 \cdot L = - \sum_1^N \log \gamma(t)^2 - \sum_1^N \varepsilon(t)^2 \gamma(t)^{-2} =$$

$$= - \sum_1^N \log R_2 \gamma'(t)^2 - \frac{1}{R_2} \sum_1^N \varepsilon(t)^2 \gamma'(t)^{-2} =$$

$$= - N \log R_2 - \sum_1^N \log \gamma'(t)^2 - \frac{1}{R_2} \sum_1^N \varepsilon(t)^2 \gamma'(t)^{-2}$$

Because $\gamma'(t)$ does not depend on R_2 we can now maximize directly with respect to R_2:

$$\max_{R_2}\left\{- N \log R_2 - \frac{1}{R_2} \sum_1^N \varepsilon(t)^2 \gamma'(t)^{-2}\right\} =$$

$$= N \log \frac{1}{N} \sum_1^N \varepsilon(t)^2 \gamma'(t)^{-2} - N$$

The extremum is achieved for

$$\hat{R}_2 = \frac{1}{N} \sum_1^N \varepsilon(t)^2 \gamma'(t)^{-2} \qquad (2.26)$$

The criterion used to find the extremum with respect to R_1' is then

$$L_{R_1'} = - N \log \frac{1}{N} \sum_1^N \varepsilon(t)^2 \gamma'(t)^{-2} - N - \sum_1^N \log \gamma'(t)^2 =$$

$$= - \sum_1^N \log \gamma'(t)^2 - N\{\log \hat{R}_2 + 1\} \qquad (2.27)$$

Note, that (2.27) gives an estimate of the scaled R_1 matrix (R_1') buth that an estimation of the scaling factor is achieved from (2.26).

An important observation is that in the case of a recursive least squares identification with constant model parameters, i.e. $R_1 = 0$, (2.26) gives an estimate of the least squares loss function that can be computed recursively.

Note, also as has been pointed out in ref [1] that once the optimal value of R_1 and R_2 have been found the normalized residuals $\xi(t)$ from (2.23) are available and may be used in tests of normality and independence.

2.3. Numerical Aspects.

When maximizing the likelihood function only the function values are available. This implies that the maximizing algorithm to be used must not be of a type that requires analytical gradients to be supplied. This limitates the practically possible dimension of the search space. Of this reason in the computer program only diagonal R_1 matrices are considered and furthermore a means by which different diagonal values may be forced to be equal has been implemented. It is felt that this restriction is not a severe one.

The routine used to maximize the likelihood function is a Fletcher-Powell algorithm with numerical calculation of the gradient. The routine calls on another routine that computes the logarithm of the likelihood function by performing a real-time identification on the stored input-output data and then computes the loss function from (2.27). As the time to do a real-time identification for a 4:th order model is approximately 2 ms on a fast computer (UNIVAC 1108) it is clear that the use of a method like this requires that the maximization method be efficient, i.e. it does not require too many function evaluations.

Another problem to have in mind is that typical values of the diagonal elements of R_1 are 10^{-3} to 10^{-7}. It is clear that it would be very difficult for the search algorithm to have to use these values directly. Of this reason the logarithm is used, i.e. the values to be considered are in the range -3 to -7 typically. Yet another problem is when the parameters of the system are constant and consequently the estimated R_1 matrix should equal zero. Unfortunately, because of the logarithm, a zero diagonal element in R_1 corresponds to $-\infty$ in the search variable. On the other hand, input-output sequences of normal length, i.e.

some thousand data pairs maximum, are too short to allow the conclusion that the parameters actually are constant and that R_1 is zero. This is reflected in the behaviour of the P equation. For small values of R_1 the loss function will remain unaffected because the P matrix has not yet converged sufficiently for R_1 to have any effect. This means that the search algorithm in a case with very slow-varying parameters will go to regions where the loss function is close to constant. Of this reason the allowable region in search space has been limited to values corresponding to diagonal elements of R_1 larger than $10^{-8.5}$, this value being chosen by experience. The bounds are introduced using a penalty function technique. Special care is exercised to ensure that the first derivative of the loss function is continuous at the boundary. This is so that no problems are introduced in the numeric differentiation used by the search algorithm to find the gradient.

2.4. Examples.

To demonstrate the ability of the routine to correctly estimate a suitable R_1 matrix 3 examples will be used. All are simulated digitally. In the first one a 3:rd order system with constant parameters is used (see Table I). In the second example these parameters are varying randomly with starting values as those of the first example. In the third case, finally, the parameters are varying as in case 2 but a constant bias in the y values is added (see Tables II and III). In all cases the input signal amplitude was 3, and the output was corrupted with white noise with a standard deviation of 1.

The results from the first example are shown in Table I. Note the good estimates of the parameters and the poles and that the diagonal value of R_1 is estimated to be 10^{-7} which indicates very slowly varying or constant parameters. The estimated noise RMS value is somewhat below the correct one. Estimation of a 4:th order model from the same data gives no significant decrease in the loss function, the sum of normalized residuals as described in (2.23).

In the second example a time varying system is identified. To facilitate a comparison between actual and estimated parameter values, the system parameters for t = 450 and t = 500 are shown. For an example including a plot of true and estimated parameters see ref. [4]. In this case the R_1 matrix was estimated to be $9 \cdot 10^{-5}$ in good agreement with the true value. The estimated RMS noise value is the correct one: 1.00. An attempt to use a 4:th order model gives in this case a slight increase in the loss. In this case the two complex (and slightly unstable) poles are found with good precision while the real one is estimated not as accurate.

In the third example, when the system parameters are time varying and the output is biased with a constant equal to 5, we observe that a direct identification with a model like the one before gives rather poor estimates of the parameters. It is interesting to observe that this is so because the estimate of the real pole is wrong while the estimates of the complex ones still are good. We also observe that in this case an increase in model order gives a significant decrease in the loss function. When, however, the bias (K) is estimated as the 7:th parameter, the estimates of the parameters are again quite reasonable. A model of order greater than 3 does not give a significant decrease in the loss function. Note also, that the es-

timated R_1 matrix gives a very small ($3 \cdot 10^{-8}$) value for the variance of the constant bias parameter while for the stochastic parameters the estimated variances is the true value 10^{-4}.

APPENDIX A - THE QUESTION OF CONVERGENCE.

We are going to study the estimation error, i.e. the difference between the true parameters $\theta(t)$ and the estimated ones $\hat{\theta}(t)$. Now, the (scaled) estimation error covariance matrix $P'(t)$ is the matrix P' of the real-time identification algorithm (2.15). If the K' is eliminated in (2.15) we have

$$P'(t+1) = P'(t) + R_1' - P'(t)\,\varphi(t-1)^T \cdot \left\{1 + \varphi(t-1)P'(t)\,\varphi(t-1)^T\right\}^{-1} (t-1)P'(t) \qquad (A.1)$$

Recall that (2.12) $\varphi(t-1)$ is a vector of old input and output values and that (2.13) R_1' is the scaled covariance matrix of the parameter noise. Thus in a general case R_1' is a positive definite matrix with rank m. m is the number of unknown parameters in the model.

It is now easy to see that the estimation error never will converge to zero. Viz. assume that

$$\hat{\theta}(t) \to \theta(t) \qquad t \to \infty$$

If the estimation error tends to zero, so does its covariance, i.e. $P(t) \to 0$. On the other hand, it is clear that $P = 0$ is no solution to (A.1) and the assumption cannot be true.

Let us then lessen our ambition and be satisfied if the estimation error in the course of time will have a constant distribution. This would mean that the covariance matrix is constant. Let us then assume that $P'(t) \to P'$. (A.1) then gives

$$R_1' = P_0' \varphi(t-1)^T \left\{1 + \varphi(t-1) P_0' \varphi(t-1)^T\right\}^{-1} \varphi(t-1) P_0'$$

The matrix in the left member is of rank m while the matrix to the right is of rank 1. Clearly this implies that our assumption is wrong and that the estimation error is a non-stationary stochastic process.

It is now possible to sketch some conditions that would ensure that the covariance matrix does not increase in the long run. If we e.g. require that

$$\operatorname{rank}\left\{\sum_{\tau=t}^{t+m-1} P'(\tau) \varphi(\tau-1)^T \varphi(\tau-1) P'(\tau)\right\} = m \qquad (A.2)$$

the effect of adding R_1' in each step as in (A.1) would be eliminated if also P' is sufficiently large. ($\{1 + \varphi(\tau-1)P'(\tau)\varphi(\tau-1)^T\}^{-1}$ is scalar and may be omitted in (A.2).) (A.2) gives restrictions on $\varphi(t-1)$ as a function of time. As φ is the vector of old input and output values this means that from (A.2) both a requirement on the input to be sufficiently exciting as well as an identifiability condition for the system itself might be derived. This is no easy task, however, because of the factors $P(\tau)$ in (A.2) which depend on $\varphi(\tau-2)$ etc.

APPENDIX B - REFERENCES.

[1] T. Bohlin: Information Pattern for Linear Discrete-Time Models with Stochastic Coefficients, IEEE Trans. on Aut. Control, AC-15, No. 1, Febr., 1970.

[2] F. Kozin: On Almost Sure Stability of Linear Systems with Random Coefficients, MIT Jour. Math. Phys., Vol. XIII, 1963.

[3] K.J. Merklinger: Nonlinear Control Systems Disturbed by Gaussian Random Processes, Ph.D. dissertation, Cambridge University, Aug., 1963.

[4] J. Wieslander: Real-Time Identification - Part I, Report 6908, Div. of Aut. Control, Lund Institute of Technology, 1969.

[5] K.J. Åström: Introduction to Stochastic Control Theory, Academic Press, 1970.

APPENDIX C - TABLES, FIGURES AND DIAGRAMS.

Table 1 - System with constant parameters.
Input signal amplitude: 3.

3:rd order model

		True	Estimated
Parameter	a_1	-1.60	-1.59
	a_2	1.61	1.60
	a_3	-0.776	-0.770
	b_1	1.20	1.22
	b_2	-0.950	-0.944
	b_3	0.200	0.198
Poles	p_1	0.400+j0.900	0.397+j0.900
	p_2	0.400-j0.900	0.397-j0.900
	p_3	0.800	0.795
	λ	1.0	0.973
	$R_{1\ ii}$	0	10^{-7}

4:th order model

$\hat{\lambda} = 0.972$ $\qquad$ $\hat{R}_{1\ ii} = 2 \cdot 10^{-7}$

Table 2 - System with stochastic parameters:

$\theta(t+1) = 0.999\ \theta(t) + 0.01\ e(t)$

Input signal amplitude: 3. $E\ ee^T = I$

3:rd order model

		True		Estimated
		t = 450	t = 500	t = 500
Parameter	a_1	-0.658	-0.501	-0.688
	a_2	1.10	1.16	1.18
	a_3	-0.488	-0.384	-0.587
	b_1	0.822	0.765	0.864
	b_2	-0.588	-0.551	-0.659
	b_3	0.037	-0.109	0.021
Poles	p_1	0.0884+j1.003	0.0765+j1.048	0.0767+j1.045
	p_2	0.0884-j1.003	0.0765-j1.048	0.0767-j1.045
	p_3	0.481	0.348	0.535
	λ	1.0		1.00
R_1	ii	10^{-4}		$9 \cdot 10^{-5}$

4:th order model

$\hat{\lambda} = 1.01$ $\qquad R_{1\ ii} = 7 \cdot 10^{-5}$

Table 3 - System with stochastic parameters:

$\theta(t+1) = 0.999\ \theta(t) + 0.01\ e(t)$

Input signal amplitude: 3. $E\ ee^T = I$.
Output signal with bias: $y(t) = y(t) + 5$.

3:rd order model

		True		Estimated	t = 500
		t = 450	t = 500	without K	with K
Parameter	a_1	-0.658	-0.501	-1.14	-0.676
	a_2	1.10	1.16	1.24	1.17
	a_3	-0.488	-0.384	-1.08	-0.574
	b_1	0.822	0.765	0.878	0.864
	b_2	-0.588	-0.551	-1.03	-0.648
	b_3	0.037	-0.109	0.118	0.018
Poles	p_1	0.0884+ +j1.003	0.0765+ +j1.048	0.0752+ +j1.042	0.0797+ +j1.042
	p_2	0.0884- -j1.003	0.0765- -j1.048	0.0752- -j1.042	0.0797- -j1.042
	p_3	0.481	0.348	0.990	0.526
	λ	1.0		1.2	1.02
	$R_{1\ ii}$	10^{-4}		$6\ 10^{-5}$	10^{-4}
	$R_{1\ 77}$	0		-	$3 \cdot 10^{-8}$

4:th order model

without K: $\hat{\lambda} = 1.14$ $\quad \hat{R}_{1\ ii} = 6 \cdot 10^{-5}$

with K: $\hat{\lambda} = 1.02$ $\quad \hat{R}_{1\ ii} = 9 \cdot 10^{-5}$ $\quad \hat{R}_{1\ 77} = 7 \cdot 10^{-7}$