PURDUE UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

MAXIMUM LIKELIHOOD IDENTIFICATION OF STOCHASTIC LINEAR SYSTEMS*

R. L. Kashyap

Technical Report No. TR-EE 68-28

August 1968

Lafayette, Indiana 47907

MAXIMUM LIKELIHOOD IDENTIFICATION OF STOCHASTIC LINEAR SYSTEMS*

R. L. Kashyap

Technical Report No. TR-EE 68-28

August 1968

*This work was partially supported by NSF under Grant GK-1970 and by NASA under Grant ~~NGR~~ NGL 15-005-021.

MAXIMUM LIKELIHOOD IDENTIFICATION OF STOCHASTIC LINEAR SYSTEMS

R. L. Kashyap

Abstract

The paper deals with the maximum likelihood (ML) estimation of the coefficients of a discrete linear system described by a set of coupled difference equations either from the input-output data or from the output data alone. The input and measurements may be noisy. The methods may be noisy. The methods also estimate the covariances of the disturbing noise. Moreover, the schemes can be modified to allow for real time operation, but the estimates are no longer ML except in the asymptotic sense. Computational results are given for a third order system.

I. INTRODUCTION

This paper deals with the identification of the parameters of a discrete stationary stochastic linear system from noisy input-output measurements $\{v(i), z(i)),\ i=1,2,\ldots\}$ or from output data $\{z(i), i=1,2,\ldots\}$ alone. It is more apt to say that the paper considers the fitting of linear models for the observed data since there may not be such a thing as a linear stochastic system which completely specifies the probabilistic environment under consideration. The criterion function chosen for fitting the linear model to the given data must reflect the ability of the model to perform tasks like prediction for which the model is usually used.

Any model building problem is intimately connected with the volume of the available data. In some problems the number of available measurement pairs $\{v(i), z(i)\}$ is limited. This limitation arises naturally when a cost is attached to the experiment for determining each input-output pair $\{v(i), z(i)\}$. In other problems, the amount of measurements available may be infinite. When the available number of measurements is limited, one is interested in computing the optimal estimates of the unknown parameters including the noise variance like the maximum likelihood estimates. In such circumstances, one is interested only in computing the estimates in an efficient manner and not necessarily in "real time" computation. But when the amount of data available is growing in time or is infinite, there is a need for developing "on line" computing schemes which can

update the estimate every time an additional piece of data comes in. For example, such problems arise naturally in the determination of optimal filters with noises of unknown statistics. In such problems, the estimates are expected to approach their true values as the amount of data handled tends to infinity.

When the number of available measurements is finite, say N, the criterion function $J_N(A)$ used for determining the parameters specified by matrix A, with true value A^*, is given below:

$$J_N(A) = \frac{1}{2N} \sum_{i=1}^{N} \|z(i) - \hat{z}(i;A)\|^2_{R^{-1}(A)} + \frac{1}{2N} \ell n \, |R(A)|$$

where

$\hat{z}(i;A)$ = predicted linear least squares estimate of $z(i)$ based on all previous measurements $z(j), j < i$, the imputs and the parameter A.

$e(i;A) \triangleq z(i) - \hat{z}(i;A)$ = error in prediction

$R(A) = E\,[e(i;A)\, e^T(i;A)]$

= covariance matrix of the prediction error.

The error $e(i;A)$, also known as the innovation, obeys a linear difference equation with $z(i)$ as the forcing function and whose coefficients are functions of A. Thus the estimation problem is reduced to the solution of a standard parameter minimization problem with difference equations as constraints. We refer to Figure 1 for the identification configuration.

When the disturbances are Gaussian, $\exp\,[-J_N(A)]$ is the likelihood function so that the estimate of A obtained by minimizing $J_N(A)$ with respect to A is the maximum likelihood estimate of A.

As a result, as N tends to infinity, the estimate of A tends to A^* almost surely. Moreover, it is possible to get a measure of the variance of the estimate with the aid of cramer-Rao lower bound [12].

When the number of measurements is growing with time, the algorithms mentioned above can be modified to make "on-line" computation of estimates possible.

At this stage, the available results on this problem may be briefly mentioned.

There are three principle methods of identification which are (1) the linear least square (LLS) methods of Kalman [1], Levin [2], Stiglitz and Mcbride [3]; (2) the instrumental variable (IV) methods of Joseph, Lewis, and Tou [4] and Wong and Polak [5]; (3) the stochastic approximation and related techniques (SA) of Ho and Lee [6], Sakrison [7], Oza and Jury [8]. All of them treat only scalar difference equations, although some of them like SA methods can be extended for multiple input-output systems. All of them estimate only the parameters of the difference equation and not the covariance of the associated noises. Moreover, both the LLS and SA techniques need a knowledge of the noise covariances whereas IV methods cannot tolerate input disturbances. All of them require knowledge of both input and output measurements. Except in IV methods, the estimates obtained with limited number of measurements are very poor unless the initial guess is close to the true value.

II. THE MODEL OF THE RANDOM PROCESS

The r-vector output process y(i) is related to the r-vector input process u(i) by the following set of coupled difference equations.

$$y(i) + A_1 y(i-1) + \ldots + A_n y(i-n)$$
$$= C_n u(i-1) + C_{n-1} u(i-2) + \ldots + C_1 u(i-n) \tag{2.1}$$

where A_i, C_i, $i=1,\ldots,n$ are a set of r x r constant matrices. The true values of these matrices have to be estimated. The integers n and r are assumed to be known.

Many a time both y(i) and u(i) cannot be measured exactly for all i. Usually a vector variable z(i) can be measured such that

$$z(i) = y(i) + \eta(i) \tag{2.2}$$

where

$$E[\eta(i)] = 0$$
$$E[\eta(i)\ \eta(j)] = R_\eta\ S_{ij}\ ,\ R_\eta > 0 \tag{2.3}$$
$$E[\eta(i)\ y(j)] = 0$$

The situation regarding the input u(i) is slightly different. In a number of examples such as economic forecasting very little is known about the inputs except that they are completely unpredictable. Moreover, the inputs may have been introduced solely for the purpose of analysis and they may not have any physical significance. In such cases, one can assume u(i) to be a sequence of zero mean random variables. In some other examples with well-defined input-output relationships, the input u(i) may be represented as

$$u(i) = v(i) + \xi(i) \tag{2.4}$$

where u(i) is the actual (unknown) input, v(i) is the (known) nominal input that was planned for the experiment and $\xi(i)$ is the inevitable error in injecting the input.

$$
\begin{aligned}
&E\,[\xi(i)] = 0 \\
&E\,[\xi(i)\ \xi(j)] = R_{\xi}\ \delta_{ij} \\
&E\,[\xi(i)\ \eta(j)] = 0 \\
&E\,[\xi(i)\ v(j)] = 0
\end{aligned}
\tag{2.5}
$$

It is clear that not all sequences $v(i)$ can serve as relevant candidates for the experiment. We shall give later the conditions that should be satisfied by the nominal input sequence $v(i)$ for successful experimentation. Presently the question of the optimal choice of the nominal input sequence $v(i)$ among the various candidates is open and will not be treated here.

Thus, according to the type of data available, one can divide the linear model building problems into 4 groups where $\xi(i)$, $\eta(i)$ indicates sequences of zero mean uncorrelated variables referred to earlier.

(A) $z(t) = y(t)$; $u(t) = \xi(t)$

(B) $z(t) = y(t)$; $u(t) = v(t) + \xi(t)$

(C) $z(t) = y(t) + \eta(t)$; $u(t) = \xi(t)$

(D) $z(t) = y(t) + \eta(t)$; $u(t) = v(t) + \xi(t)$

When $y(\cdot)$ is a scalar process, problems of class (A) occur in obtaining good approximations to spectral density functions of $y(\cdot)$ [16]. In addition, classes (A) and (B) arise very often in many economic forecasting problems where the idea of measurement noise is superfluous. Classes (C) and (D) arise in all problems with well-defined input-output relationships. Classes (A) and (C) can be considered to be special cases of classes (B) and (D), respectively, by setting $v(i) \equiv 0$ for all i. Thus $v(i)$ will be set identically to zero when no information is available on it.

The identification problem is to estimate the unknown matrices among $A_1, \ldots, A_n$, $C_1, \ldots, C_n$, R_ξ and R_η from the available data $\{v(i), z(i)\}$, $i=1, \ldots, N$ where N can be finite or infinite. As before, the collection of all the unknown matrices will be denoted by A.

III. THE INNOVATION EQUATIONS

As mentioned in the introduction, the coefficients that are actually estimated are those of the so-called "innovation" equation relating the successive value of the prediction errors $e(i,A)$ with the measured output and input. This equation has been discussed in great detail in reference [10], and hence we will briefly mention the outlines here.

Let $\hat{z}(i)$ = predicted linear least squares estimate of $z(i)$ given all the previous measurements $z(j)$, $j < i$, previous measured inputs $v(j)$, $j \leq i$, and the coefficients of the model equation (2.1).

$$= \text{Arg} \left[\min_{f_i} E \left\| z(i) - f_i(z(i-1), z(i-2), \ldots) \right\|^2 \right]$$

where $f_i(\cdot)$ is a linear function of its arguments.

$e(i) \triangleq$ innovation at instant i

$$= z(i) - \hat{z}(i) .$$

One can easily demonstrate [9] that $E\,[e(i)\, e^T(j)] = 0$, $\forall\, j \neq i$. Thus, the innovations are nothing more than orthogonalized measurements.

(A) Exact Measurements of the Output Available

By definition $z(i) = y(i)$. In addition, C_n^{-1} is assumed to exist. The system equation (2.1) can be rewritten as (3.1)

$$z(t) = -\sum_{j=1}^{n} A_j z(t-j) + \sum_{j=1}^{n} C_{n-j+1}(v(t-j) + \xi(t-j)) \qquad (3.1)$$

$\hat{z}(t)$ can be obtained from (3.1) by setting $\xi(t-1) = 0$ since one does not have any information on it based on the measurements $z(j), j < t$ alone.

$$\hat{z}(t) = -\sum_{j=1}^{n} A_j z(t-j) + \sum_{j=1}^{n} C_{n-j+1}\, v(t-j) + \sum_{j=2}^{n} C_{n-j+1}\, \xi(t-j) \qquad (3.2)$$

Subtracting (3.2) from (3.1), one gets

$$e(t) \triangleq z(t) - \hat{z}(t) = C_n\, \xi(t) \qquad (3.3)a$$

or

$$\xi(t) = C_n^{-1}\, e(t) \qquad (3.3)b$$

The required innovation equation is obtained by substituting (3.3)b into equation (3.1).

$$e(t) + \sum_{j=2}^{n} C_{n-j+1}\, C_n^{-1}\, e(t-j+1) = z(t) + \sum_{j=1}^{n} A_j\, z(t-j) - \sum_{j=1}^{n} C_{n-j+1}\, v(t-j) \qquad (3.4)$$

Moreover, from (3.3)a, one can show that

$$E\left[e(i)\, e^T(j)\right] = C_n R_\xi C_n^T \delta_{ij} \qquad (3.5)$$

(B) Noisy Measurements of State

$$z(i) = y(i) + \eta(i) \qquad (3.6)$$

The derivation of the innovation equation is different from the case considered earlier and has been considered by the author in reference [10]. Hence only the results will be mentioned here. The innovation equation is given in (3.7).

$$e(t) + \sum_{j=1}^{n} B_{n-j+1}\, e(t-j) + \sum_{j=1}^{n} C_{n-j+1}\, v(t-j) = z(t) + \sum_{j=1}^{n} A_j z(t-j) \tag{3.7}$$

Let $\quad E\left[e(t)\, e^T(t)\right] \triangleq R_e$

The r x r coefficient matrices $B_1, \ldots, B_n$ and R_e are found by solving the following set of algebraic equations.

$$\sum_{k=1}^{i+1} B_{n-i+k} R_e B_k^T = \sum_{k=1}^{i+1} C_{n-i+k} R_\xi C_k^T + \sum_{k=0}^{i} A_{i-k} R_\eta A_{n-k}^T \tag{3.8}$$

$$i = 0, 2, \ldots, n$$

where

$$C_{n+1} \triangleq 0\ , \quad B_{n+1} \triangleq I \quad \text{and} \quad A_0 = I$$

The equations (3.8) can be rewritten in matrix notation.

$$\begin{bmatrix} B_{n+1} & & & \\ B_n & B_{n+1} & & \\ B_{n-1} & B_n & B_{n+1} & \\ \vdots & & & \\ B_1 & B_2 & B_3 \ldots . B_{n+1} \end{bmatrix} R_e \begin{bmatrix} B_1^T \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ B_{n+1}^T \end{bmatrix} = \begin{bmatrix} C_{n+1} & & & \\ C_n & C_{n+1} & & \\ C_{n-1} & C_n & C_{n+1} & \\ \vdots & & & \\ C_1 & C_2 \cdot \cdot \cdot \cdot & & C_{n+1} \end{bmatrix} R_\xi \begin{bmatrix} C_1^T \\ C_2^T \\ \\ \\ \\ C_{n+1}^T \end{bmatrix}$$

$$+ \begin{bmatrix} A_0 & & & \\ A_1 & A_0 & & \\ A_2 & A_1 & A_0 & \\ A_n & A_{n-1} & \cdot \cdot \cdot & A_0 \end{bmatrix} R_\eta \begin{bmatrix} A_n^T \\ A_{n-1}^T \\ \cdot \\ \cdot \\ \cdot \\ A_0^T \end{bmatrix} \qquad (3.9)$$

An iterative scheme will be mentioned below for obtaining B_i's and R_e from A_i, C_i, R_ξ and R_η, θ with k denoting the iteration number.

$$\left.\begin{aligned} B_1^k &= \left[A_n R_\eta\right]\left[R_e^{k-1}\right]^{-1} \\ B_i^k &= \Bigg[\sum_{j=1}^{i-1} C_j R_\xi C_{n-i+1+j}^T + \sum_{j=1}^{i} A_{n-j+1} R_\eta A_{i-j}^T \\ &\quad - \sum_{j=1}^{i-1} B_j^{k-1} R_e^{(k-1)} \left(B_{n-i+j+1}^{k-1}\right)^T\Bigg]\left(R_e^{k-1}\right)^{-1} \\ &\qquad\qquad i=2,\ldots,n \end{aligned}\right\} \qquad (3.10)$$

$$R_e^k = -\sum_{j=1}^{n} B_j^k R_e^{k-1} \left(B_j^k\right)^T + \sum_{j=1}^{n} C_j R_\xi C_j^T + \sum_{j=1}^{n} A_{n+1-j} R_\eta A_{n+1-j}^T \qquad (3.11)$$

The iteration scheme in equations (3.10) and (3.11) is based on a set of difference equations in $B_i(t)$ from which the algebraic equations (3.9) were derived [10].

IV. THE LIKELIHOOD FUNCTION

The conditional probability density $p(z(i),\ldots,z(N)/v(i),\ldots,v(N-1);$ A will be computed assuming the noise sequences $\eta(i)$ and $\xi(i)$ to be Gaussian with the second order properties mentioned earlier. On account of the latter assumption, the linear least squares estimates $\hat{z}(i;A)$ equals the conditional mean of $z(i)$ mentioned below.

$$\hat{z}(i;\theta) = E\left[z(i)/z(j),\ j < i \quad v(j),\ j \leq i;\ A\right]$$

Let $\quad e(i;A) = z(i) - \hat{z}(i;A)$

It has already been mentioned that

$$E\left[e(i,A)\ e^{T}(j,A)\right] = R_e(A)\ \delta_{ij}$$

Hence

$$p(z(i)/z(j),\ j \leq i\ ;\ v(j),\ j \leq i,\ A) \sim N(\hat{z}(i,A), R_e(A))$$

H nce

$$p(z(1),\ldots,z(N)/v(i),\ldots,v(N-1);A)$$

$$= \prod_{j=1}^{N} p(z(j)/z(1),\ldots,z(j-1);\ v(1),\ldots,v(N-1);\ A)$$

$$= \frac{1}{(2\pi)^{rN/2}\ |R_e(A)|^{N/2}}\ \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\|e(i;A)\|^2_{R_e^{-1}(A)}\right]$$

Let

$$J_N(A) = \frac{1}{2N}\sum_{j=1}^{N}\|e(j;A)\|^2_{R_e^{-1}(A)} + \frac{1}{2}\ \ell n\ |R(A)| \qquad (4.1)$$

Thus the estimate of A obtained by minimizing $J_N(A)$ with respect to A with the variable e(j;A) obeying the difference equation (3.4) or (3.7) is the same as the so called "unconditional" maximum likelihood (ML) estimate of A.

By Wald's theorems [13-14] the ML estimate of A will tend to the true value A^* with probability one as the number of measurements N tend to infinity if the conditional density function of the measurements satisfies the assumptions (1) - (8) of Wald's paper [14]. If these, the only assumption of interest in the present context is assumption 4 which states that

$$p(z(1)\ldots,z(N)/A^{(1)}) \neq p(z^{(1)},\ldots,z^{(N)}/A^{(2)}),\ A^{(1)} \neq A^{(2)}$$

for at least one value of $z(1),\ldots,z(N)$. The rest of the assumptions are automatically satisfied here. It is natural to discuss the conditions that need be imposed on the system to assure the validity of the above assumption. These conditions can be obtained by inspection of the innovation equation (3.4) or (3.7). Throughout this discussion $R_\eta > 0$ (positive definite) and/condition $R_\xi > 0$ is imposed only if $v(i) \equiv 0$ for all i.

In systems of class (A), $v(i) \equiv 0$ and $z(i) = y(i)$. The innovation equation (3.4) reveals that one has to know C_n and this should be nonsingular. These conditions permit the remaining coefficients A_1, A_n, $C_1,\ldots,C_{n-1}$ and R_F to be estimated.

In systems of class (B) $C_1,C_2,\ldots,C_n$, $A_1,\ldots,A_n$, R_F

can be estimated as long as the input sequence $v(i)$, $i = 1,2,\ldots$ repeatedly span all the directions of the r-dimensional space. If $v(i)$ is a scalar, this $v(i) \neq 0$ is a sufficient condition. Of course C_n must be nonsingular.

In problems of class (c), $z(i) = y(i) + \eta(i)$ and $v(i) \equiv 0$. In this case, the innovation equation (3.7) does <u>not explicitly</u> involve the coefficients C_i, $i = 1,\ldots,n$. Moreover, there is no unique way of specifying R_ξ. Hence, one can set $R_\xi = I$ without any loss of generality. Since these identification schemes determine only the coefficients of the innovation equation (in this case the A_i, B_i $i = 1,\ldots,n$ and R_e), it is necessary to impose additional conditions so that one can uniquely recover $C_1,\ldots,C_n$ and R_η from the coefficients A_i, B_i, $i = 1,\ldots,n$ and R_e. To do this, one has to consider a scheme for computing the coefficients C_i and R_η recursively from the algebraic equation (3.9). As before, k will stand for the iteration number

$$R_\eta = R_e B_1^T A_n^{-1} \tag{4.2}$$

$$C_i^k = \Big[- \sum_{j=1}^{i+1} A_{n+1-j} R_\eta A_{i+1-j}^T + \sum_{j=1}^{i+1} B_j R_e B_{n-i+j}^T - \sum_{j=0}^{i-1} C_j^{k-1} (C_{n-i+j}^{k-1})^T \Big] ((C_n^{k-1})^{-1})^T \tag{4.3}$$

$$i = 1,\ldots,n$$

with the definitions

$$C_0^k \equiv 0, \quad A_0 = I \quad B_{n+1} = I$$

Hence, in addition to the nonsingularity of C_n, one requires A_n^{-1} to exist. This means that all the individual difference equations in the system (2.1) must be of the same order. If A_n^{-1} does not exist, there is no unique way of recovering C_i R_η from A_i, B_i, and R_e.

In systems of class (D), $z(i) = y(i) + \eta(i)$ and the innovation equation (3.7) involves all the parameters A_i, C_i, $i = 1,\ldots,n$ explicitly and R_ε, R_η implicitly via B_i's. Thus, the exact identification is possible if the input $v(i)$ can excite all the "modes" of the system. In otherwords, if every set of n consecutive inputs with $v(i)$ as a leading member is rearranged to form a column vector, say $\bar{v}(i)$, then the sequence $\bar{v}(i)$, $\bar{v}(2),\ldots$ must repeatedly span all the directions of the nr dimensional space.

V. ALGORITHMS FOR MAXIMUM LIKELIHOOD ESTIMATION

This section will deal only with the systems represented in equation (2.1). The problems of classes (A),(B),(C) and (D) having different types of input-output information will be treated separately. The noises $\varepsilon(i)$ and $\eta(i)$, when referred to have the properties mentioned in equations (2.3) and (2.5). The matrices to be estimated are among A_i, C_i, $i = 1,\ldots,n$, R_ε and R_η. The available data is $\{z(i), v(i)\}$, $i = 1,\ldots,N$.

(A) Problems of Class (A)

Here $z(i) = y(i)$ and $v(i) \equiv 0$. Since C_n is assumed to be

nonsingular and known, one may as well set $C_n = I$. The unknowns are $A_1,\ldots,A_n$, $C_1,\ldots,C_{n-1}$ and R_ξ. For ease of rotation, relabel C_i and R_ξ as follows:

$$A_{n+i} \triangleq C_i \quad , \quad i = 1,\ldots,n-1$$

$$A_{2n} \triangleq R_\xi$$

The innovation equation is

$$e(t) + \sum_{i=2}^{n} A_{2n-j+1} e(t-j+1) = z(t) + \sum_{j=1}^{n} A_j \, z(1-j) \tag{5.1}$$

$$E\{e(i)e^T(j)\} = A_{2n} \, \delta_{ij}$$

The criterion function is

$$J(A) = \frac{1}{2N} \sum_{j=1}^{N} \|e(j)\|^2_{A_{2n}^{-1}} + \frac{1}{2} \ln |A_{2n}| \tag{5.2}$$

Minimization of $J(A)$ subject to the differential constraint (5-1) is a standard problem in the minimization theory. One can solve it numerically either by a first order gradient method or, preferably, by the conjugate gradient method [15] since with a little increase in computation, one gets a considerable increase in the reate of convergence. The computational scheme is given below involving four steps with the letter k in superscript denoting the iteration number.

(i) With the given A_i^k solve for the difference equation (5.1) for the prediction errors $e(t)$, $t = 1,\ldots,N$. Minimizing $J(A^k)$ with respect to A_{2n} one gets the following estimate for A_{2n}.

$$A_{2n}^k = \frac{1}{N} \sum_{t=1}^{N} e(t)e^T(t)$$

(ii) Let us compute the gradient matrices

$$\frac{\partial J}{\partial (A_i)}_{jp} = \frac{1}{N} \sum_{t=1}^{N} \frac{\partial e^T(t)}{\partial (A_i)}_{jp} A_{2n}^{-1} e(t) \qquad \begin{matrix} i = 1,\ldots,2n-1 \\ j,p = 1,\ldots,r \end{matrix}$$

By differentiating the innovation equation with respect to $(A_i)_{jp}$, one can compute the partial derivatives $\partial e(t)/\partial (A_i)_{jp}$.

Let $$\frac{\partial J^k}{\partial A_i} \triangleq \frac{\partial J}{\partial A_i}\Big|_{A_i = A_i^k}$$

(iii) $$A_i^{k+1} = A_i^k + \rho^{(k)} P_i^k$$

$$P_i^k = -\frac{\partial J^k}{\partial A_i} + \beta^{(k)} P_i^{k-1}, \quad i = 1,\ldots,2n-1$$

where $\beta^{(k)}$ and $\rho^{(k)}$ are scalars.

The gain $\rho^{(k)}$ is chosen to maximize the difference

$$| J(A^{k+1}) - J(A^k) |$$

$$\beta^{(k)} = \frac{\sum_{i=1}^{2n-1} \left\| \frac{\partial J^{(k+1)}}{\partial A_i} \right\|}{\sum_{i=1}^{2n-1} \left\| \frac{\partial J^{(k)}}{\partial A_i} \right\|}, \quad \| A \| \triangleq T\gamma(A^T A)$$

The initial values of P_i are

$$P_i^1 = \frac{-\partial J^1}{\partial A_i}$$

(iv) Increase k to k+1 and go to step (i).

By setting $\beta^{(k)} \equiv 0$ in the above computation one gets the gradient method.

(B) Systems of Class (B)

Here $z(i) = y(i)$ and satisfies the conditions mentioned in section IV. The unknowns are A_i, C_i, $i = 1,\ldots,n$ and R_ξ. The algorithm is very similar to that in section (A) and hence will be skipped.

(C) Systems of Class (C)

Here $z(i) = y(i) + \eta(i)$ and $v(i) \equiv 0$. As mentioned earlier one can set $R_\xi = I$ and assume A_n to be nonsingular. The unknowns are $A_1,\ldots,A_n$ and $C_1,\ldots,C_n$. The innovation equation is

$$e(t) + \sum_{j=1}^{n} B_{n-j+1}\, e(t-j) = z(t) + \sum_{j=1}^{n} A_j\, z(t-j)$$

with $\quad E\left(e(t)e^T(t)\right) = R_e$

The identification procedes in 2 steps.

(i) Estimate the coefficients $A_1,\ldots,A_n$, $B_1,\ldots,B_n$ and R_e by using the method outlined in Part (A) of this section. Label them as $\hat{A}_1,\hat{A}_2,\ldots,\hat{A}_n$, $\hat{B}_1,\ldots,\hat{B}_n$ and $\hat{R}_e$, respectively.

(ii) The estimate $\hat{R}_{\eta_t}$ of R_{η_t} is given by (4.2)

$$\hat{R}_{\eta_t} = \hat{R}_e \hat{B}_1^T (\hat{A}_n)^{-1}$$

The estimates $\hat{C}_1,\ldots,\hat{C}_n$ of $C_1,\ldots,C_n$ are obtained as the steady state solutions of the recursive (4.3) with R_e, A_i, B_i replaced by $\hat{R}_e$ $\hat{A}_i$, $\hat{B}_i$ respectively.

(D) Systems of Class (D)

Here $z(i) = y(i) + \eta(i)$ and $u(i) = v(i) + \xi(i)$. The input $v(i)$ obeys the conditions mentioned mentioned in section IV. The unknown matrices are A_i, C_i, $i = 1,\ldots,n$ and R_ξ and R_η. The innovation equation involves all the parameters A_i, C_i, $i = 1,\ldots,n$ and is given below

$$e(t) + \sum_{j=1}^{n} B_{n-j+1}\, e(t-j) + \sum_{j=1}^{n} C_{n-j+1} v(t-j)$$

$$= z(t) + \sum_{j=1}^{n} A_j\, z(t-j) \qquad (5.8)$$

where the coefficients B_i, $i = 1,\ldots,n$ and R_e obey the algebraic equations (3.8).

Let

$$A_{n+1} \triangleq C_i \ , \quad i = 1,\ldots,r$$

$$A_{2n+1} \triangleq R_\xi$$

The criterion function is

$$J(A) = \frac{1}{2N} \sum_{j=1}^{N} \| e(t) \|^2_{R_e^{-1}} + \frac{1}{2} \ln | R_e |$$

The scheme is very similar to that of part (A) of this section and consists of the following steps

(i) Using the given values of $A_i^{(k)}$ and $B_i^{(k)}$ evaluate $e(t)$ for all $t = 1,\ldots,N$ from (5.8).

$$R_e^k = \frac{1}{N} \sum_{j=1}^{N} e(t)\, e^T(t)$$

(ii) Evaluate $\frac{\partial e(t)}{\partial (A_i)_{jp}}$ for $i = 1,\ldots,2n+1$; $j,p = 1,\ldots,r$; by differentiating the equation (5.8) with respect to $(A_i)_{jp}$ for all i, j and p. During this process, one needs the partial derivatives of all the elements of B with respect to every element of A. This can be obtained directly by differentiating equation (3.8) or recursively by differentiating equation (3.10). Thus, one can compute the gradient matrices $\partial J/\partial A_i$, $i = 1,\ldots,2n+1$.

(iii) $$A_i^{k+1} = A_i^k + \rho^{(k)} P_i^{(k)} \qquad i = 1,\ldots,2n+1$$

$$P_i^k = - \frac{\partial J^{(k)}}{\partial A_i} + \beta^{(k)} P_i^{k-1}$$

where the scalars $\rho^{(k)}$ and $\beta^{(k)}$ are computed as in part (A) of

this section.

(iv) Using A_i^{k+1}, R_e^k and $\beta_i{}^k$, compute $B_i^{(k+1)}$ using (3.10). Compute R_η^{k+1} by modifying (3.11) as follows

$$R_\eta^{k+1} = - \sum_{j=1}^{n} A_{n+1-j}^{k+1} R_\eta (A_{n+1-j}^{k+1})^T + \sum_{j=1}^{n+1} B_j^{k+1} R_e^k (B_j^{k+1})^T$$

$$- \sum_{j=1}^{n} C_j A_{2n+1}^{k+1} C_j^T$$

(v) Increment k by one and go to step (1).

The computational scheme may be simplified in many respects. When the mean square value of $v(i)$ is much greater than R_ξ then we can neglect the effect of the partial derivatives of the elements of B_i with respect to those of A_i altogether. When the mean square value of $v(i)$ is comparable with that of R_ξ, we may not neglect these partial derivatives, but they may not vary much from iteration to iteration and hence, it is enough if they are computed once in many iterations. It may be also worthwhile to look into the methods of optimization which do not explicitly compute the gradients.

VI. ON-LINE IDENTIFICATION SCHEMES

The maximum likelihood estimation ideas will be used in conjunction with stochastic approximation to develop on-line computing schemes. Instead of handling every measurement individually, batch processing will be performed as it reduces the amount of computation in many problems. Every

batch has a finite number, say m, of measurements. The k^{th} batch contains the measurements $\{z(i),v(i)\}$, $i = (k-1)m+1, \ldots, km$.

Let A denote the matrix of parameters to be estimated and $e(t,A)$ denote the predicted error computed from the innovation equation using the actual measurements $z(t)$ and the parameter A. Assume that $z(t)$ and $e(t,A)$ are Gaussian stationary processes. Consider the criterion function $J(A)$

$$J(A) = E\left[\frac{1}{m} \sum_{t=km+1}^{(k+1)m} \|e(t,A)\|^2_{R_e^{-1}(A)} + \ell n \, | R_e(A) | \right]$$

$$\triangleq E\, f(t,A)$$

where R_e, the covariance matrix of $e(t,A)$, is a function of A.

As demonstrated by Wald [14], one has

$$J(A^o) < J(A) \quad \forall\, A^o \neq A \tag{6.1}$$

On account of the smoothness properties of $J(A)$ one can write

$$\frac{\partial J(A)}{\partial A} = g(A,A^o)\ \varphi(\|A-A^o\|) \tag{6.2}$$

where $g(A,A^o)$ is a positive scalar function and $\varphi(\cdot)$ is a montonically increasing function of the argument. In view of (6.2), the following inequality is valid for all values of A in some neighborhood around the true value A^*.

$$\left\| \frac{\partial J(A)}{\partial A} \right\|^2 \leq k_1(1+ \|A\|^2) \ , \ k_1 > 0 \tag{6.3}$$

One gets the following algorithm by applying the Robbins Munro scheme for finding the zero of the function $E\left(\frac{\partial J}{\partial A}\right)$,

$$A_i^{k+1} = A_i^k - \rho^{(k)} \frac{\partial f(t,A)}{\partial A_i} \Big|_{A=A^k} \tag{6.4}$$

where $\rho^{(k)}$ is a scalar sequence such that

$$\sum_k \rho^{(k)} = \infty \quad \text{and} \quad \sum_k (\rho^{(k)})^2 < \infty \tag{6.5}$$

From Gladyshev's theorem [11], and equations (6.2)-(6.5) it follows that A_i^k tends to the true value A_i^* both in the mean square sense and with probability one. It should be noted that if condition (6.3) is satisfied, we do not need any independence assumptions on the correction terms used in the stochastic approximation scheme of equation (6.4). As in all stochastic approximation procedures, the mean square error of the estimate is inversely proportional to the number of measurements processed.

The details of the computation will be given for problems of the classes (A) and (D). The others are omitted since the methods are similar.

(A) Systems of Class (A)

The unknowns are $A_i, i=1,\ldots,n$. $A_{n+i} \triangleq C_i$, $i=1,\ldots,n-1$ and $A_{2n} \triangleq R_c$. Let $C_n = I$. The algorithm is

$$A_i^{k+1} = A_i^k - \rho^{(k)} \left[\frac{1}{m} \sum_{t=(k-1)m+1}^{km} \frac{\partial}{\partial A_i} \|e(t)\|^2_{A_{2n}^{-1}} + \frac{\partial}{\partial A_i} \ell n |A_{2n}| \right]$$

$$i=1,\ldots,2n-1$$

$$A_{2n}^{k+1} = A_{2n}^k - \rho^{(k)} \left[\frac{1}{m} \sum_t e(t)e^T(t) - A_{2n}^k \right]$$

The auxiliary equations for computing e(t) and its partial derivatives are given below

$$e(t) + \sum_{j=1}^{n} A_{2n-j+1}^{k}\, e(t-j) = z(t) + \sum_{j=1}^{n} A_{j}^{k}\, z(t-j)$$

$$\frac{\partial e(t)}{\partial (A_i)_{pq}} + \sum_{j=1}^{n} A_{2n-j+1} \frac{\partial e(t-j)}{\partial (A_i)_{pq}} = \sum_{j=1}^{n} C \frac{\partial A_j}{\partial (A_i)_{pq}} \cdot z(t-j)$$

$$- \sum_{j=1}^{n} \frac{\partial A_{2n-j+1}}{\partial (A_i)_{pq}} e(t-j) \qquad \begin{matrix} i=1,\ldots,2n \\ p,q=1,\ldots,r \end{matrix}$$

(B) Systems of Class D

The unknowns are $A_1,\ldots,A_n$, $A_i \triangleq A_{n+i}, i=1,\ldots,n$, $R_\xi \triangleq A_{2n+1}$, $R_\xi \triangleq A_{2n+2}$ and R_η. The algorithm is given in the following 3 steps.

(i) From the given values of A_i^k, $i=1,\ldots,2n+2$ and R_η^{k-1} compute B_i^k, $i=1,\ldots,n$ from the equations (3.10). Update R_η^k using the equation given below which is a modification of (3.11).

$$R_\eta^k = - \sum_{p=1}^{n} A_{n+1-\ell}^{k} R_\eta^{k-1} (A_{n+1-p}^{k})^T - \sum_{p=n+1}^{2n} A_p^k A_{2n+1}^k (A_p^k)^T$$

$$+ \sum_{p=1}^{n} B_p^k A_{2n+2}^k (B_p^k)^T + A_{2n+2}^k$$

(ii) The error e(t) is evaluated recursively

$$e(t) + \sum_{j=1}^{n} B_{n-j+1}^{(k)}\, e(t-j) + \sum_{j=1}^{n} A_{2n-j+1}^{k}\, v(t-j)$$

$$= z(t) + \sum_{j=1}^{n} A_j^k\, z(t-j)$$

One can recursively compute the partial derivatives of e(t) with respect to A_i for all i=1,...,2n+1 by differentiating the e(t) equation and using the partial derivatives of all elements of B_i with respect to A_j for all j. The latter partial derivatives can be obtained by differentiation of equations (3.10).

(iii) Update the values of $A_i^{(k)}$ as follows:

$$A_i^{k+1} = A_i^k - \rho^{(k)} \left[\frac{1}{m} \sum_{t=(k-1)m+1}^{km} \frac{\partial}{\partial n_i} \|e(t)\|^2_{A_{2n+2}^{-1}} + \frac{\partial}{\partial A_i} \ell n |A_{2n+2}| \right]$$

$$i=1,\ldots,2n+1$$

$$A_{2n+2}^{k+1} = A_{2n+2}^k - \rho^{(k)} \left[\frac{1}{m} \sum_t e(t)e^T(t) - A_{2n+2}^k \right]$$

The bulk of the computation occurs in step (ii). The batch processing of data helps in the reduction of the partial derivatives to be evaluated. Regarding the partial derivatives of the elements of B_i with respect to those of A_j, all the comments made earlier in part (D) of the section V are also valid here. When r is large, the demands on the computer may be heavy since for fixed n, the amount of computation is proportional to r^3. But there is no way of getting around the problem if there is both process noise and measurement noise and one insists on measuring all the noise covariances and coefficients of the difference equation. Of course, the amount of computation is greatly reduced if the input noise ξ were absent since $B_i = A_{n-i+1}$ and hence the partial derivatives of B_i with respect to A_j can be written down by inspection.

VII. EXAMPLES

(A) Maximum Likelihood Estimation of a Third Order System

A single input-single output system obeying the following difference equation will be considered.

$$y(i) + a_1 y(i-1) + a_2 y(i-2) + a_3 y(i-3) = c_3 \xi(i-1) + c_2 \xi(i-2) + c_3 \xi(i-3)$$

$$z(i) = y(i) + \eta(i)$$

No information is available on the input $\xi(i)$ except that it is zero mean. As before, one can get r_ξ, the covariance of noise $\xi(i)$, to be one. The unknowns are $a_1, a_2, a_3, c_1, c_2, c_3$ and r, the covariance of the uncorrelated noise $\eta(i)$. The only available measurements are $\{z(i),\ i=1,\ldots,N\}$

The computational method mentioned in section V(C) can be used here. However, since the output $z(.)$ is a scalar, one can easily write up a second order gradient method without much effort for carrying out the minimization process. The results are given in figure 2 and table 1. In figure 2 the ML estimates of the seven parameters are graphed against the number of samples N. This figure shows how the ML estimates approach their true value as the number of samples become large. In table 1 the number of samples are fixed at some number, say 400, and the estimates obtained during the minimization process are tabulated against the number of iterations. As can be seen from the table, most of the minimization is performed in the first iteration itself.

(B) On-Line Determination of the Noise Statistics

Consider the following scalar input-scalar output system

Table 1. Results of the Iterative Procedure for Computing the ML Estimates. The number of samples is fixed at the value 400.

Iteration Number	a_1	a_2	a_3	c_1	c_2	c_3	r
0	0.5	0.5	0.5	0	0	1.0	-
1	0.617	.5388	-0.3493	-.0954	-0.0842	0.9910	0.5700
2	0.6203	.5025	-0.3671	-.1246	-0.0455	.9962	0.5437
3	0.6211	.4926	-0.3700	-.1304	-.0370	1.003	.5420
4	0.6196	.4888	-0.3714	-.1322	-.0367	1.005	.5419

$$y(i) + \sum a_j\, y(t-j) = \sum_{j=1}^{n} c_{n-j+1}\, \xi(t-j)$$

$$z(i) = y(i) + \eta(i)$$

$\xi(i)$ is zero mean, uncorrelated process, not accessable for measurement. Without loss of generality, set $r_\xi = 1$. The coefficients $a_1, \ldots, a_n, c_1, \ldots, c_n$ are known. Only the variance r_η of noise $\eta(i)$ is unknown. This will be estimated on-line from the measurements $z(i)$. The innovation equation is

$$e(t) + \sum_{j=1}^{n} b_j\, e(t-j) = z(t) + \sum_{j=1}^{n} a_j\, z(t-j)$$

Let $E[e^2(t)] = r_e$. The coefficients $b_1, \ldots, b_n$, r_e and r_η and unknown since they depend on r_η. Let $r_e(t)$ and $b_i(t)$ be estimates of r_e, b_i at the t^{th} instant. Then the algorithm can be written as follows using t both as the time and iteration index.

$$e(t) + \sum_{j=1}^{n} b_j(t)\, e(t-j) = z(t) + \sum_{j=1}^{n} a_j\, z(t-j)$$

The recursive equations for b_j can be written down from (3.10)

$$b_1(t) = a_n\, r_\eta(t-1)/r_e(t-1)$$

$$b_i(t) = \frac{1}{r_e(t-1)} \Bigg[\sum_{j=1}^{i-1} c_j\, c_{n-i+1+j} + \sum_{j=1}^{i} a_{n-j+1}\, r_\eta(t-1)\, a_{i-j} - \sum_{j=1}^{i-1} b_j(t-1)\, r_e(t-1)\, b_{n-i+j+1}^{(t-1)} \Bigg]$$

The equation for $r_\eta(t)$ can be obtained by modifying (3.11).

$$r_\eta(t) = -\sum_{j=1}^{n} a_{n+1-j}^2 \, r_\eta(t-1) + \sum_{j=1}^{n+1} b_j^2(t) \, r_e(t-1) - \sum_{j=1}^{n} c_j^2$$

$$r_e(t) = r_e(t-1) + \frac{1}{t} \, [e^2(t) - r_e(t-1)]$$

This completes the identification algorithm.

One can show [10] that the approximate value of the best estimate of $y(t)$ given $z(1), \ldots, z(t)$ can be written down as follows

$$\hat{y}(t/t) \approx z(t) - \frac{r_\eta(t) \, e(t)}{r_e(t)}$$

VIII. DISCUSSION

While comparing the existing methods of identification [1-8] with those of this paper, the following aspects of our algorithms should be mentioned.

(i) With limited number of measurements, the maximum likelihood estimates of the various parameters including the noise variances are computed. One can rarely do better than this. One can get a measure of the variance of the estimate.

(ii) The identification is possible even if there are no input measurements.

(iii) The method handles vector measurements also in a similar manner.

(iv) The noise in the output $\eta(i)$ need not be uncorrelated. It can have finite correlation time (with unknown covariance function).

(v) The algorithms present a convenient method of computing good approximations to the spectral densities of stationary processes on the basis of the observed samples by assuming that they obey a model of equation (2.1). However, a rigorous comparison has not been made with the traditional methods [16].

It has already been mentioned that the on-line identification schemes present a method of optimal filtering with disturbances of unknown statistics. The computational aspects associated with the on-line schemes have already been mentioned. Further, no stability studies are available on our algorithms. But the experimental results are promising.

IX. CONCLUSIONS

The problem of identifying the coefficients and the noise variances of a discrete linear system on the basis of noisy input-output measurements or output measurements has been reduced to a standard parameter minimization problem with difference equations as constraints. The criterion function is the sum of the squares of the optimal prediction error and this can be interpreted as the likelihood function with Gaussian disturbances. Optimal estimates of the parameters are derived when the number of measurements is limited and for various types of input-output data.

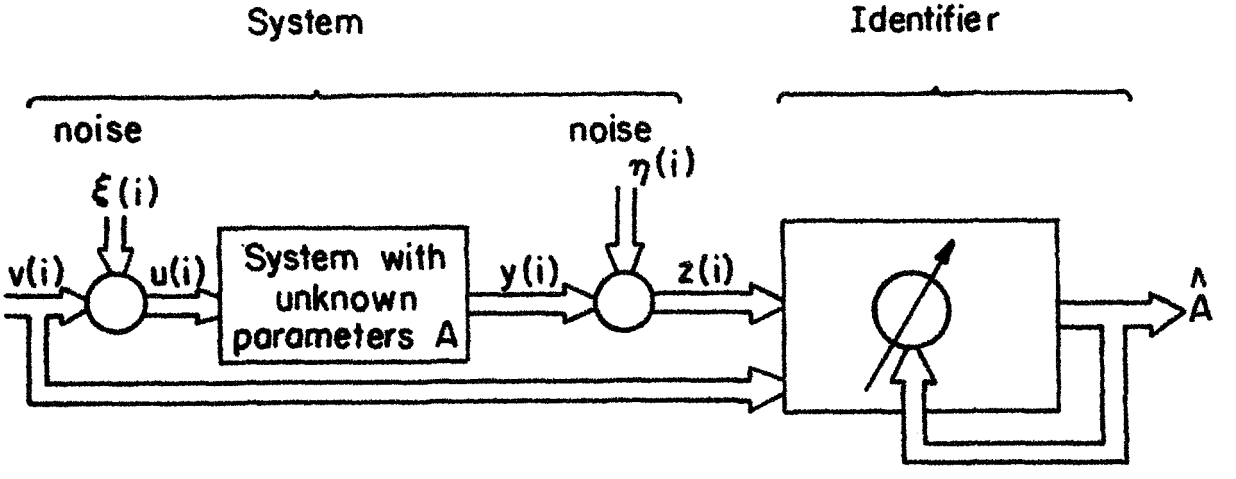

FIGURE 1: IDENTIFICATION CONFIGURATION .

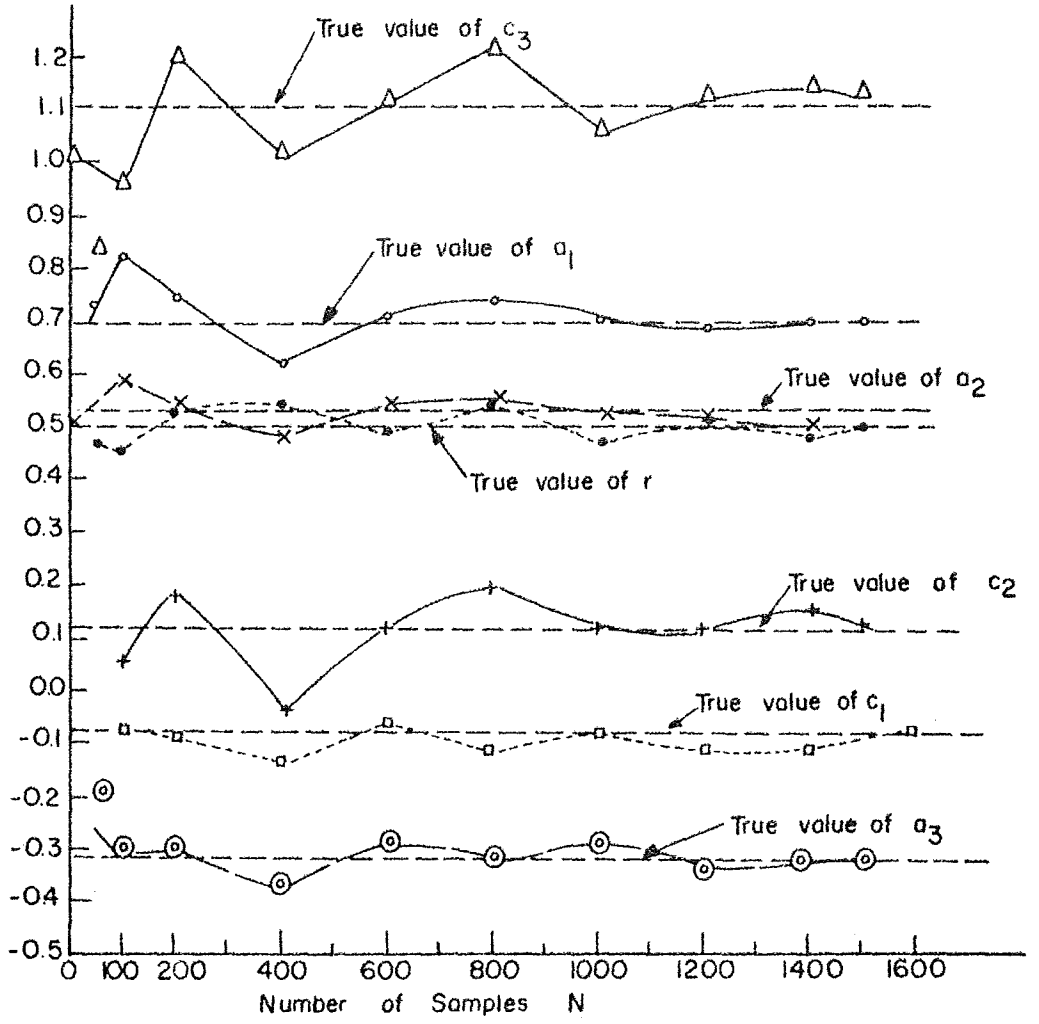

FIGURE 2. GRAPH OF MAXIMUM LIKELIHOOD ESTIMATES OF $a_1, a_2, a_3, c_1, c_2, c_3$, AND r VS NUMBER OF SAMPLES.

REFERENCES

1. R. E. Kalman, "Design of a self-optimizing control system", Trans. ASME Ser. D. J. Basic Engrg., 80 (1958), pp. 468-478.

2. M. J. Levin, "Estimation of a system pulse transfer function in the presence of noise", IEEE Trans. Automatic Control, AC-9 (1964), pp. 229-235.

3. K. Steiglitz and L. E. McBride, "A technique for the identification of linear systems", Ibid, AC-10 (1965), pp. 461-464.

4. P. Joseph, J. Lewis and J. Tou, "Plant identification in the presence of disturbances and application to digital adaptive systems," AIEE. Trans., Part 2, Application and Industry, 80(1961), pp. 18-24.

5. K. Y. Wong and E. L. Polak, "Identification of linear discrete time systems using the instrumental variable method", IEEE Trans. on Automatic Control, Vol. AC-12, Dec. 1967, pp. 707-719.

6. Y. C. Ho and R. C. K. Lee, "Identification of linear dynamic systems", J. Inform. and Control, Vol. 8, pp. 93-110, Feb. 1965.

7. D. J. Sakrison, "Use of stochastic approximation to solve the system identification problem, IEEE Trans. Automatic Control, AC-12(1967), pp. 563-567.

8. K. G. Oza and E. I. Jury, "System identification and the principle of random contraction mapping", SIAM J. on Control, Vol. 6, No. 2, 1968, pp. 244-251.

9. T. J. Kailath, "An Innovation Approach to Linear Least Squares Estimation, Part I: The Filtering Problem", (Submitted to) IEEE Trans. on Automatic Control (to appear).

10. R. L. Kashyap, "A New Method of Recursive Estimation in Discrete Linear Systems," Purdue University, School of Electrical Engineering Tech Rept. TR-EE-68-13, June 1968.

11. E. G. Gladyshev, "On Stochastic Approximation", Theory of Probability and Its Applications, Vol. X, No. 2, (1965), pp. 275-278.

12. C. R. Rao, Linear Statistical Inference and Its Applications, John Wiley, New York, 1965.

13. A. Wald, "Asymptotic Properties of the Maximum Likelihood Estimate of an Unknown Parameter of a Discrete Stochastic Process", Ann. Math. Stat., Vol. 19, (1948), pp. 40-48.

14. A. Wald, "Note on the Consistency of the Maximum Likelihood Estimate", Ann. Math. Stat., Vol. XX, No. 4, Dec. 1949, pp. 595-601.

15. R. Fletcher and C. M. Reeves, "Function Minimization by Conjugate Gradients", The Computer Journal 7, (1964), pp. 149-154.

16. A. M. Walker, "Asymptotic Properties of Least Square Estimates of Parameters of the Spectrum of a Stationary Non-Deterministic Time Series", J. Australian Math. Soc., Vol. 4, (1964), pp. 363-384.