A METHOD OF STATISTICAL IDENTIFICATION OF DISCRETE TIME PARAMETER LINEAR SYSTEMS*

HIROTUGU AKAIKE

(Received Oct. 8, 1968)

1. Introduction and summary

In this paper the use of least squares method for the identification of time-invariant discrete-time linear systems is discussed. The output of the system is assumed to contain an additive disturbance besides the response to the input. It is well known that there is an inherent difficulty in the least squares method for the application to the case where the correlation between the input and the disturbance exists. The case where a feedback loop connects the output to the input is a typical one of this situation. In a recent paper [1] the present author proposed an identification procedure, which was tentatively called predictive identification, of noisy linear feedback systems. The one-sided moving average representation of the disturbance played a fundamental role for the development of the procedure.

In this paper, the representation of the disturbance as the sum of the effect of its past history and the present innovation is called the predictive representation of the process and its general implication for the identification problem is discussed. For linear systems the autoregressive process is adopted as a basic model of the disturbance and various methods of estimation are developed for the identification of the system with possible feedback from the output to the input. It is observed that the model developed for the consistency of the estimate leads us to an asymptotically efficient estimate, which was first introduced by J. Durbin [2], when the additive disturbance is independent of the input.

Based on the results of discussion of the statistical characteristics of the proposed estimates, a practical computation scheme for the general use of identification is proposed. Numerical examples are given to provide a feeling of the relative efficiencies of the estimates. An example of application of ordinary method of least squares to the case with feed-

* Presented at the U.S.-Japan Joint Seminar on Engineering Application of Stochastic Processes, Washington, D.C., on September 19, 1968.

back is given to show the difference from the result obtained by our identification procedure.

It is further pointed out that the estimate of the frequency response function obtained by taking the numerical Fourier transform of the estimate of the impulse response obtained by the present procedure is consistent even when the feedback exists and also asymptotically efficient if the feedback is absent and thus will generally be superior to that obtained by the conventional cross-spectral method which is not applicable to the case with feedback [3]. This consideration suggests that this new approach will eventually replace the conventional cross-spectral approach in many practical applications. The only possible drawback of the present time domain approach will be the necessity of the strong assumption of independency of the innovation process of the noise for the evaluation of the sampling variabilities of the estimate, while in the spectral approach the basic local Gaussian assumption of the Fourier transforms of the data sequence generally did not mean a serious restriction ([4], p. 197).

2. Basic model

In this paper we consider a linear system with discrete time parameter $n(=0, \pm 1, \pm 2, \cdots)$ and described by the relation

$$x_0(n)=\sum_{j=1}^{K}\sum_{m=0}^{M} a_{jm}x_j(n-m)+u(n)\ ,$$

where $x_0(n)$ is the output and $\{x_j(n);\ j=1, 2, \cdots, K\}$ is the input of the system and $u(n)$ is the additive disturbance. Our problem of identification is to get good estimates of the impulse response functions $\{a_{jm};\ m=0, 1, \cdots, M\}$ $(j=1, 2, \cdots, K)$ (: identification of the system) and also of the statistical characteristics of $u(n)$ (: identification of the noise) from the observation of $\{x_0(n), x_1(n), \cdots, x_K(n)\}$.

For the sake of simplicity of the following discussions and also from its practical plausibility in many practical applications we assume that $\{x_0(n), x_1(n), \cdots, x_K(n), u(n)\}$ forms a strictly stationary ergodic stochastic process with zero mean vector and finite variance matrix. Thus we can expect that the time averages of various statistics over an observed data converge to their expectations, or the ensemble averages, with the exception of probability zero as the data length is increased to infinity.

Now if the ordinary method of least squares is applied to the observed data $\{x_0(n), x_1(n), x_1(n-1), \cdots, x_1(n-M), x_K(n), x_K(n-1), \cdots, x_K(n-M);\ n=1, 2, \cdots, N\}$ the estimate $\{\hat{a}_{jm}\}$, which we shall hereafter call OLS (ordinary least squares) estimate, of $\{a_{jm}\}$ is given as $\{\alpha_{jm}\}$ which minimizes $\sum_{n=1}^{N}\left(x_0(n)-\sum_{j=1}^{K}\sum_{m=0}^{M}a_{jm}x_j(n-m)\right)^2$. $\{\hat{a}_{jm}\}$ is the solution of the normal equation

$$C\hat{a}=b\,,$$

where C is a $(M+1)K\times(M+1)K$ matrix of which $((k-1)(M+1)+l+1,\ (j-1)(M+1)+m+1)$ element is $\frac{1}{N}\sum_{n=1}^{N}x_k(n-l)x_j(n-m)$ and $\hat{a}$ and b are $(M+1)K\times 1$ matrices, or vectors, with $((k-1)(M+1)+l+1,1)$ elements equal to $\hat{a}_{kl}$ and $\frac{1}{N}\sum_{n=1}^{N}x_k(n-l)x_0(n)$, respectively.

From the relation

$$\sum_{n=1}^{K}x_k(n-l)x_0(n)=\sum_{j=1}^{K}\sum_{m=0}^{M}a_{jm}\sum_{n=1}^{N}x_k(n-l)x_j(n-m)+\sum_{n=1}^{N}x_k(n-l)u(n)$$

we get

$$b=Ca+d\,,$$

where a and d are the $(M+1)K\times 1$ matrices with $((k-1)(M+1)+l+1,1)$ elements equal to a_{kl} and $\frac{1}{N}\sum_{n=1}^{K}x_k(n-l)u(n)$, respectively. Assuming the non-singularity of C we get

$$\hat{a}=a+C^{-1}d\,.$$

Thus, if the $(M+1)K\times(M+1)K$ matrix D which is obtained by replacing $\frac{1}{N}\sum_{n=1}^{N}x_k(n-l)x_j(n-m)$ in C by $Ex_k(n-l)x_j(n-m)$ is non-singular, $\hat{a}$ converges with probability one to $a+D^{-1}d_\infty$ as N tends to infinity, where d_∞ is given by replacing $\frac{1}{N}\sum_{n=1}^{N}x_k(n-l)u(n)$ in d by $Ex_k(n-l)u(n)$. From this result we can see that the condition $d=0$ or $Ex_k(n-l)u(n)=0$ $(k=1,2,\cdots,M,\ l=0,1,\cdots,M)$ is essential for the OLS estimate $\{\hat{a}_{jm}\}$ to be a consistent estimate of $\{a_{jm}\}$, i.e., an estimate which converges stochastically to $\{a_{jm}\}$.

In the rest of this paper we shall assume the convergence in probability of a consistent estimate.

3. Implication of the predictive representation

As was seen at the end of the preceding section, the condition of orthogonality, or uncorrelatedness, of $u(n)$ to $\{x_j(n-m)\}$ $(j=1,2,\cdots,K,\ m=0,1,\cdots,M)$ is essential to get a consistent estimate of $\{a_{jm}\}$ by the ordinary least squares method. This condition does not hold when there is a feedback which connects the output $x_0(n)$ to the input $\{x_j(n-m)\}$. Thus the ordinary method of least squares can not be applied to this practically important case [3].

Now if the correlation between the additive disturbance and the input is due to the existence of some physically realizable path from $u(n)$ to $\{x_j(n)\}$, which produces some output in $\{x_j(n)\}$ from the present and past values of $u(n)$, the correlation will be minimized by transforming the original model into the form where the disturbance is represented by its innovation, i.e., the residual of $u(n)$ after the deletion of the effect of its past history. Thus if there exists a representation, which we shall call a predictive representation, of $u(n)$

$$u(n)=w(n)+g(u(n-1), u(n-2), \cdots),$$

where the innovation $w(n)$ is independent of $u(n-1), u(n-2), \cdots$ and the effect of infinitely remote past is vanishing in $u(n)$, we shall be able, for a linear g, to develop a transformation of the original model to get the the relation

$$y_0(n)=\sum_{j=1}^{K}\sum_{m=0}^{M} a_{jm}y_j(n-m)+w(n),$$

where $y_j(n)=h\{x_j(s); s\leqq n\}$ is obtained by applying the transformation which produced $w(n)=h\{u(s); s\leqq n\}$ from $\{u(n)\}$ to $\{x_j(n)\}$. In this representation we can expect the minimum possible correlation between $w(n)$ and $\{y_j(n-m)\}$ under the physical realizability assumption of the path from $u(n)$ to $x_j(n)$. The effects of infinitely remote past histories should be vanishing in $y_j(n)$s.

If we are going to limit the path from the output to the input to be linear we shall be able to replace the assumption of independency of $w(n)$ to $u(n-1), u(n-2), \cdots$ in the definition of the predictive representation of $u(n)$ by the weaker assumption of orthogonality to develop the discussion. We shall call this the weak predictive representation of $u(n)$ when the distinction is necessary. For a purely non-deterministic $u(n)$ ([5], p. 75) we have this weak predictive representation in the sense of mean square with linear g, i.e., $g(u(n-1), u(n-2), \cdots)$ is the linear projection of $u(n)$ on the space spanned by $u(n-1), u(n-2), \cdots$, and it quite naturally leads us to the following model of $u(n)$ for practical applications:

$$u(n)=\sum_{l=1}^{L} c_l u(n-l)+w(n),$$

where $w(n)$ is a white noise with $E(w(n))^2=\sigma^2$ (>0) and $Ew(n)w(m)=0$ ($n\neq m$) and the roots of the characteristic equation $1-\sum_{l=1}^{L} c_l z^l=0$ are all lying outside the unit circle [6]. Throughout the rest of this paper we shall assume that $u(n)$ is given by this model.

The practical utility of the present model in relation to the identification of linear noisy feedback systems was discussed in detail in [1] and some of the practical applications of the method given in the paper were discussed in [7]. Our present observation shows that the predictive representation, weak or non-weak, of the disturbance has a definite physical meaning in the identification problem.

E. Parzen ([8], sec. 6) has stressed the importance of the role of white disturbance model to make the least squares estimates efficient ones in synthesizing a linear model for multiple time series. It seems that the existence and use of the white disturbance representation of a noisy system are deeply rooted in the physical realizability of the system and thus the whitening of the disturbance should be realized through its predictive representation.

Our observation also suggests the fact that the identification of the structure of the system will essentially be impossible if $u(n)$ contains an output of some physically realizable path with the input $\{x_j(n)\}$.

4. Consistent estimators

In this section, our problem is to get estimates of $\{a_{jm}\}$, $\{c_l\}$ and $\sigma^2 = Ew^2(n)$ from a record of $\{x_0(n), x_1(n), \cdots, x_K(n)\}$ which follows the model

$$x_0(n) = \sum_{j=1}^{K} \sum_{m=0}^{M} a_{jm} x_j(n-m) + u(n)$$

$$u(n) = \sum_{l=1}^{L} c_l u(n-l) + w(n) ,$$

where the structure of $u(n)$ is as was assumed in the preceding section.

For the purpose of evaluation of sampling variabilities of the estimates we further assume that $w(n)$ is independent of the past of $u(n)$, i.e., $\{u(n-l)\}$ $(l=1, 2, \cdots)$, and of the present and past of the input, i.e., $\{x_j(n-m)\}$ $(j=1, 2, \cdots, K,\ m=0, 1, 2, \cdots)$. Accordingly, $w(n)$ is independent of the past of $x_0(n)$, i.e., $x_0(n-l)$ $(l=1, 2, \cdots)$.

It is also assumed that there does not exist any strict linear relation within $\{x_j(n);\ j=0, 1, \cdots, K,\ n=0, \pm 1, \pm 2, \cdots\}$ so that any variance matrix of finite number of elements of the process is non-singular.

We consider the situation where a record $\{x_0(n), \cdots, x_0(n-L), x_1(n), \cdots, x_1(n-L-M), \cdots, x_K(n), \cdots, x_K(n-L-M);\ n=1, 2, \cdots, N\}$ is given.

1. SLS estimate

First we shall discuss a simple estimation procedure which was originally developed in [1] and will hereafter be called as SLS (simplified least squares) estimation procedure. By a linear transformation of the original model we get a relation

$$x_0(n)=\sum_{l=1}^{L} c_l x_0(n-l)+\sum_{j=1}^{K}\sum_{m=0}^{M+L} A_{jm} x_j(n-m)+w(n)\,,$$

where

$$A_{jm}=a_{jm}-\sum_{l=1}^{L} c_l a_{jm-l}$$

and a_{jm}s with $m<0$ are considered to be zeros. As is obvious from the discussion of section 2 the ordinary method of least squares applied to the data $\{x_0(n), \cdots, x_0(n-L), x_1(n), \cdots, x_1(n-M-L), \cdots, x_K(n), \cdots, x_K(n-M-L);\ n=1, 2, \cdots, N\}$ yields a consistent estimates $\{\hat{c}_l\}$ and $\{\hat{A}_{jm}\}$ of $\{c_l\}$ and $\{A_{jm}\}$, respectively. From the recurrence relation

$$a_{jm}=A_{jm}+\sum_{l=1}^{L} c_l a_{jm-l}$$

a consistent estimate $\{\hat{a}_{jm}\}$ of $\{a_{jm}\}$ is obtained as follows:

$$\hat{a}_{jm}=0 \qquad (m<0 \text{ or } m>M)$$

$$\hat{a}_{j0}=\hat{A}_{j0}$$

$$\hat{a}_{jm}=\hat{A}_{jm}+\sum_{l=1}^{L} \hat{c}_l \hat{a}_{jm-l} \qquad (m=1, 2, \cdots, M)\,,$$

a consistent estimate s^2 of $\sigma^2=Ew^2(n)$ is given by

$$s^2=\frac{1}{N}\sum_{n=1}^{N}\left(x_0(n)-\sum_{l=1}^{L}\hat{c}_l x_0(n-l)-\sum_{j=1}^{K}\sum_{m=0}^{M+L}\hat{A}_{jm}x_j(n-m)\right)^2.$$

To evaluate the asymptotic sampling variabilities of our estimates we assume that the simultaneous distribution of $x_j w(m)=N^{-1/2}\sum_{n=1}^{N} x_j(n-m)\cdot w(n)$ converges to a multi-dimensional Gaussian distribution. Especially this is the case for the linear model treated in the former paper [1], if only the white noise is composed of mutually in dependent random variables.

If we define $\Delta c_l=N^{1/2}(\hat{c}_l-c_l)$ and $\Delta A_{jm}=N^{1/2}(\hat{A}_{jm}-A_{jm})$ it can be shown that

$$\begin{bmatrix} \Delta c_1 \\ \vdots \\ \Delta c_L \\ \Delta A_{10} \\ \vdots \\ \Delta A_{1M+L} \\ \vdots \\ \Delta A_{K0} \\ \vdots \\ \Delta A_{KM+L} \end{bmatrix} \approx R^{-1} \begin{bmatrix} x_0 w(1) \\ \vdots \\ x_0 w(L) \\ x_1 w(0) \\ \vdots \\ x_1 w(M+L) \\ \vdots \\ x_K w(0) \\ \vdots \\ x_K w(M+L) \end{bmatrix},$$

where R is the variance matrix of $(x_0(n-1), \cdots, x_0(n-L), x_1(n), \cdots, x_1(n-M-L), \cdots, x_K(n), \cdots, x_K(n-M-L))$ and the sign $\approx$ means that the both side members of the sign have one and the same limiting distribution. Under some mild assumption, such as the mutual independence of the variables in the white noise of the model of [1], the variance matrix of $R^{-1}[x_0w(1), \cdots, x_Kw(M+L)]'$ converges to that of the limit distribution, where $'$ denotes the tranposed matrix. This type of convergence is hereafter assumed and the variance matrix of this limit distribution is denoted by

$$E_\infty \begin{bmatrix} \Delta c \\ \Delta A \end{bmatrix} [(\Delta c)'(\Delta A)'] .$$

This matrix is called the asymptotic variance matrix of $[\Delta c, \Delta A]$ or simply of $\hat{c}$ and $\hat{A}$. Also the covariance matrix in the limit distribution of $\{\Delta c_l\}$ and $\{\Delta A_{jm}\}$ is denoted by $E_\infty(\Delta c)(\Delta A)'$ and is called the asymptotic covariance matrix of Δc and ΔA. This kind of notation will generally be used for other quantities similar to Δc and ΔA.

As we have the relation $Ex_j w(m) x_k w(l) = \sigma^2 Ex_j(n-m)x_k(n-l)$ we can see that the asymptotic variance matrix of $[\Delta c, \Delta A]$ is equal to $\sigma^2 R^{-1}$. Now if we define $\Delta a_{jm} = N^{1/2}(\hat{a}_{jm} - a_{jm})$ the following relation holds:

$$\Delta a_{jm} \approx \sum_{l=1}^{L} c_l \Delta a_{jm-l} + \sum_{l=1}^{L} \Delta c_l a_{jm-l} + \Delta A_{jm} .$$

Thus we get for $m=0, 1, \cdots, M$

$$\Delta a_{jm} \approx \sum_{l=1}^{m} \Delta c_l (a_j * b)_{m-l} + \sum_{k=0}^{m} \Delta A_{jk} b_{m-k} ,$$

where $\{b_l\}$ is given by the relation

$$\sum_{l=0}^{\infty} b_l z^l = \left(1 - \sum_{l=1}^{L} c_l z^l\right)^{-1} \qquad \text{for } |z| \leqq 1$$

and

$$(a_j * b)_m = \sum_{l=0}^{m} a_{jm-l} b_l \ .$$

This representation allows us to evaluate the asymptotic variability of $\{\hat{a}_{jm}\}$ from that of $\{\hat{c}_l, \hat{A}_{jm}\}$.

2. TLS estimate

Here we start from the relation

$$y_0(n) = \sum_{j=1}^{K} \sum_{m=0}^{M} a_{jm} y_j(n-m) + w(n) \ ,$$

where $y_j(n) = x_j(n) - \sum_{l=1}^{L} c_l x_j(n-l)$. If the exact values of $\{c_l\}$ were given we would apply the method of ordinary least squares to $\{y_j(n);\ j=0, \cdots, K\}$ to get an estimate of $\{a_{jm}\}$, which is an efficient estimate or a minimum variance unbiased linear estimate ([5], p. 87). As a natural substitute of this approach we use the SLS estimate $\{\hat{c}_l\}$ in place of $\{c_l\}$ and get the TLS (two stage least squares) estimate $\{\hat{a}_{jm}\}$ of $\{a_{jm}\}$ as the values of $\{\alpha_{jm}\}$ which minimizes the sum of squares $\sum_{n=1}^{N} \left\{\hat{y}_0(n) - \sum_{j=1}^{K} \sum_{m=0}^{M} \alpha_{jm} \hat{y}_j(n-m)\right\}^2$, where $\hat{y}_j(n) = x_j(n) - \sum_{l=1}^{L} \hat{c}_l x_j(n-l)$.

From the result of the preceding paragraph it can be shown that

$$\Delta c \approx R_{rx_0}^{-1} rx_0 w \ ,$$

where Δc denotes the vector $[\Delta c_1, \cdots, \Delta c_L]'$ and R_{rx_0} denotes the variance matrix of $[rx_0(n-1), rx_0(n-2), \cdots, rx_0(n-L)]$, where $rx_0(n-l)$ is the residual of $x_0(n-l)$ after the deletion of the projection of $x_0(n-l)$ on $\{x_j(n-m);\ j=1, 2, \cdots, K,\ m=0, 1, \cdots, M+L\}$, and $rx_0 w = [rx_0 w(1), \cdots, rx_0 w(L)]'$, where $rx_0 w(l) = N^{-1/2} \sum_{n=1}^{N} rx_0(n-l) w(n)$. Thus we get

$$E_\infty(\Delta c)(\Delta c)' = \sigma^2 R_{rx_0}^{-1} \ .$$

It also should be noted that $rx_0 = ru$ holds, where ru is the vector of the corresponding residuals of $u(n-l)$s after the deletion of their projection on $\{x_j(n-m)\}$.

Now we get the relation

$$\Delta a \approx R_y^{-1} yw - R_y^{-1} R_{y,u} \Delta c \ ,$$

where Δa denotes the vector $[\Delta \hat{a}_{10}, \cdots, \Delta \hat{a}_{KM}]$ of TLS estimate and $yw = [y_1 w(0), \cdots, y_K w(M)]'$, where $y_j w(m) = N^{-1/2} \sum_{n=1}^{N} y_j(n-m) w(n)$, and $R_{y,u}$ is the covariance matrix $E[y_1(n), \cdots, y_K(n-M)]'[u(n-1), \cdots, u(n-L)]$ and R_y is the variance matrix of $[y_1(n), \cdots, y_K(n-M)]$. By taking into account the relation $Erx_0(n-l)y_j(n-m)=0$ we can see that $E(rx_0w)(y_jw)'=0$ holds and thus $E_\infty(\Delta c)(y_jw)'=0$. Using this relation we can get the equalities

$$E_\infty(\Delta a)(\Delta c)'=-\sigma^2 R_y^{-1}R_{y,u}R_{rx_0}^{-1}$$

$$E_\infty(\Delta a)(\Delta a)'=\sigma^2(R_y^{-1}+R_y^{-1}R_{y,u}R_{rx_0}^{-1}R_{u,y}R_y^{-1}) ,$$

where $R_{u,y}=E[u(n-1), \cdots, u(n-L)]'[y_1(n), \cdots, y_K(n-M)]$. A consistent estimate s^2 of σ^2 is given by

$$s^2=\frac{1}{N}\sum_{n=1}^{N}\left\{\hat{y}_0(n)-\sum_{j=1}^{K}\sum_{m=0}^{M}\hat{a}_{jm}\hat{y}_j(n-m)\right\}^2 .$$

3. ALS estimate

At this point it would be natural to proceed one step further and get a new estimate $\{{}_2\hat{c}_l\}$ of $\{c_l\}$ as $\{\gamma_l\}$ which minimizes $\sum_{n=1}^{N}\left\{\hat{u}(n)-\sum_{l=1}^{L}\gamma_l\hat{u}(n-l)\right\}^2$, where $\hat{u}(n)=x_0(n)-\sum_{m=0}^{M}\hat{a}_{jm}x_j(n-m)$ and $\{\hat{a}_{jm}\}$ is the TLS estimate of $\{a_{jm}\}$. We get

$$\Delta_2 c\approx R_u^{-1}uw-R_u^{-1}R_{u,y}\Delta a ,$$

where R_u is the variance matrix $E[u(n-1), \cdots, u(n-L)]'[u(n-1), \cdots, u(n-L)]$ and uw is the vector of $uw(l)=N^{-1/2}\sum_{n=1}^{N}u(n-l)w(l)$ $(l=1, 2, \cdots, L)$ and, as will be shown in Appendix I, we have

$$E_\infty(\Delta_2 c)(\Delta_2 c)'=R_u^{-1}(R_u+R_{u,y}R_y^{-1}R_{y,u}+R_{u,y}R_y^{-1}R_{y,u}R_{rx_0}^{-1}R_{u,y}R_y^{-1}R_{y,u})R_u^{-1} .$$

It can be shown that $E_\infty(\Delta c)(\Delta c)'-E_\infty(\Delta_2 c)(\Delta_2 c)'$ is non-negative definite and thus any linear combination of the components of $\Delta_2 c$ does not have greater asymptotic variance than the corresponding linear combination of the components of Δc.

We can further continue the process indefinitely, obtaining a new estimate of $\{a_{jm}\}$ or $\{c_l\}$ alternatively by utilizing the last estimate of the other set of parameters. It can be shown (c.f. Appendix I) that this process gives a sequence of estimates of which asymptotic variance matrices of Δa and Δc, which we shall denote by $V[\Delta a]$ and $V[\Delta c]$ respectively, must satisfy the relation

$$V[\Delta c]=R_u^{-1}[R_u+R_{u,y}V[\Delta a]R_{y,u}]R_u^{-1} ,$$

$$V[\Delta a]=R_y^{-1}[R_y+R_{y,u}V[\Delta c]R_{u,y}]R_y^{-1}$$

and we have

$$V[\Delta a]=(R_y-R_{pu,y})^{-1} ,$$

$$V[\Delta c]=(R_u-R_{py,u})^{-1}$$

where pu and py denote the vectors of the projections of $(u(n-1), \cdots, u(n-L))$ and $(y_1(n), \cdots, y_K(n-M))$ on $\{x_j(n-m);\ j=1,2,\cdots,K,\ m=0,1,\cdots,M\}$. It can further be shown that the asymptotic covariance matrix of Δa and Δc is given by $E_\infty(\Delta c)(\Delta a)'=-R_u^{-1}R_{u,y}E_\infty(\Delta a)(\Delta a)'$. Thus the limit of the asymptotic variance matrix of $[\Delta c, \Delta a]$ when the stage proceeds indefinitely is given by

$${}_\infty E_\infty\begin{bmatrix}\Delta c\\ \Delta a\end{bmatrix}[(\Delta c)'(\Delta a)']=\begin{pmatrix}R_u & R_{u,y}\\ R_{y,u} & R_y\end{pmatrix}^{-1}\sigma^2\,.$$

Our final estimates are given as the limits of these estimates and they are the sets of values of $\{\alpha_{jm}\}$ and $\{\gamma_l\}$ which minimize the sum of squares $\sum_{n=1}^{N}\left\{x_0(n)-\sum_{l=1}^{L}\gamma_l x_0(n-l)-\sum_{j=1}^{K}\sum_{m=0}^{M}\alpha_{jm}\left(x_j(n-m)-\sum_{l=1}^{L}\gamma_l x_j(n-m-l)\right)\right\}^2$. We shall call these estimates the ALS (absolute least squares) estimates. A consistent estimate of σ^2 is obtained in the same way as in the case of TLS estimates.

In the simple case where the disturbance $\{u(n)\}$ is independent of the input $\{x_j(n)\}$ the relation $R_{u,y}=O$ (zero matrix) holds and thus the asymptotic variance matrices do not decrease after the TLS estimate. This case was discussed by J. Durbin [2]. As $R_{y,u}=O$ holds in this case we have

$$\Delta a\approx R_y^{-1}yw$$

and this shows that our estimate $\{\hat{a}_{jm}\}$ is asymptotically equivalent to the one obtained by applying the ordinary method of least squares to the model $y_0(n)=\sum_{j=1}^{K}\sum_{m=0}^{M}a_{jm}y_j(n-m)+w(n)$ after the linear transformation by $\{c_l\}$ to make the least squares estimate the minimum variance unbiased linear estimate ([5], p. 87). It is interesting to note that our effort to get a consistent estimate when the feedback exists has led us asymptotically to this efficient estimate in the case without feedback.

It also should be noted that our general result of the limit of the asymptotic variance matrix can be obtained by formally applying the asymptotic variance formula of ordinary maximum likelihood estimate for independent observations assuming $\{w(n)\}$ to be Gaussian.

By using a Schwartz type inequality for matrices ([5], p. 88) we can show (c.f. Appendix II) that the asymptotic variance matrix of the SLS estimates of $\{c_l, a_{jm}\}$ is greater than or equal to that of the ALS estimate. By definition, a matrix A is greater than a matrix B if $A-B$ is positive definite. It can further be shown that when $\{u(n)\}$ is white and independent of the input $\{x_j(n);\ j=1,2,\cdots,K\}$ and the $\{x_j(n)\}$s are mutually independent or $K=1$ the SLS estimate has one and the same asymptotic variance matrix as the ALS estimate for any value of L chosen for computation.

From the foregoing discussions it has become clear that:

a) SLS, TLS, ALS estimates are definitely superior to OLS in the case with feedback,

b) TLS estimate is asymptotically more efficient than OLS in the case without feedback and with non-white $\{u(n)\}$.

It should be mentioned that if some of a_{jm}s are assumed to be vanishing in the original model we have only to disregard the corresponding terms in the foregoing discussion.

When the true values of L and M are unknown we have to adopt some guessed values of them for computation. Under our present assumption, if only the guessed values are not smaller than the corresponding true values, we can always get consistent estimates. In the simple case where the feedback is absent the guess of L does not affect the consistency but affects the efficiency of the estimate of $\{a_{jm}\}$. For the feedback case we have to expect bias when a too small value is selected for L. Generally, taking these values too large merely contributes to increase the sampling variability of the estimates.

5. A practical estimation procedure

The following procedure would be the most useful for practical applications:

1. Compute the OLS estimate of $\{a_{jm}\}$.
2. Compute the SLS estimate of $\{a_{jm}\}$, $\{c_l\}$ and σ^2.
3. If the difference between the two estimates of $\{a_{jm}\}$ is not quite significant and the estimate of $\{c_l\}$ is very nearly equal to the zero vector, adopt the OLS estimate of $\{a_{jm}\}$ as the final estimate of $\{a_{jm}\}$.
4. If the difference between the two estimates of $\{a_{jm}\}$ is not quite significant but the estimate of $\{c_l\}$ is far from the zero vector compute the TLS estimate of $\{a_{jm}\}$ and adopt it as the final estimate.
5. If the difference between the former two estimates of $\{a_{jm}\}$ is quite significant, compute TLS or further ALS estimates of $\{a_{jm}\}$ and $\{c_l\}$ and adopt it as the final estimates. In this case value of L for computation should carefully be checked.

In case the economy of computation matters very much, the SLS (simplified least squares) estimate will in many cases be the most practical one.

In some applications, the result of estimation of $\{a_{jm}\}$ may be quite poor due to very low S/N ratio at some frequency band. In this case it would be necessary to analyze the sampling variability of the Fourier transform of $\{a_{jm}\}$ and suppress the contribution of the unreliable frequency band.

6. Numerical examples

Monte Carlo experiments were performed of the following model:

$$x_0(n)=0.12x_1(n)+0.20x_1(n-1)+0.05x_1(n-2)+u(n)$$

$$u(n)=0.9u(n-1)+w(n)$$

$$x_1(n)=0.7x_1(n-1)+w_1(n)$$

where $\{w(n)\}$ and $\{w_1(n)\}$ are sequences of uniformely distributed random numbers with $Ew(n)=Ew_1(n)=0$ and $Ew^2(n)=(0.1)^2$ and $Ew_1^2(n)=(0.5)^2$. In this case there is no feedback from $x_0(n)$ to $x_1(n)$ and we have selected as our values of L and M for computation $L=6$, $M=5$, respectively, and N was 496. Some statistics based on the results of ten repetitions of the experiment are illustrated in Table 1.

Table 1. Sample means of $\frac{1}{6}\sum_{m=0}^{5}(\hat{a}_{1m}-a_{1m})^2$ in ten repetitions.

OLS	SLS	TLS
2.170×10^{-4}	0.845×10^{-4}	0.863×10^{-4}

Corresponding results are given in Table 2 for the case with

$$x_0(n)=0.12x_1(n)+0.20x_1(n-1)+0.05x_1(n-2)+u(n)$$

$$u(n)=0.910u(n-1)-0.181u(n-2)+0.092u(n-3)+0.053u(n-4)$$
$$+0.035u(n-5)-0.108u(n-6)+w(n)$$

$$x_1(n)=0.605x_1(n-1)-0.113x_1(n-2)+0.165x_1(n-3)-0.091x_1(n-4)$$
$$+0.095x_1(n-5)-0.006x_1(n-6)+w_1(n)\,,$$

where the random numbers used in the former case are used in this case too but with modified scales $Ew^2(n)=(0.175)^2$ and $Ew_1^2(n)=(0.5)^2$. The values $L=6$, $M=5$ were used for computation and N was 491.

Table 2. Sample means of $\frac{1}{6}\sum_{m=0}^{5}(\hat{a}_{1m}-a_{1m})^2$ in ten repetitions.

OLS	SLS	TLS
5.259×10^{-4}	2.772×10^{-4}	3.042×10^{-4}

The results clearly show that SLS estimate is superior to OLS estimate in efficiency in these non-white residual cases.

In spite of the additive amount of computations of TLS estimates they are not better than SLS estimates in these examples.

To show the definite drawback of OLS estimate in the case with feedback, experiments were performed on a model

$$x_0(n)=0.12x_1(n-1)+0.20x_1(n-2)+0.05x_1(n-3)+u_0(n)$$

$$u_0(n)=0.9u_0(n-1)+w_0(n)$$

$$x_1(n)=-0.1x_0(n-1)-0.1x_0(n-2)-0.1x_0(n-3)+u_1(n)$$

$$u_1(n)=0.7u_1(n-1)+w_1(n)$$

where $\{w_0(n)\}$ and $\{w_1(n)\}$ are white noises of independently and uniformly distributed random variables with $Ew_0(n)=Ew_1(n)=0$ and $Ew_0^2(n)=Ew_1^2(n)=(0.5)^2$. The coefficients a_{j0} $(j=1, 0)$ were assumed to be vanishing and L and M for the computation were put equal to their true values $L=1$ and $M=3$ for both equations and N was 499. The result is shown in Tables 3 and 4. The definite superiority of SLS and TLS estimates over the OLS estimate can clearly be seen. This example is analogous to the one discussed in a former paper [3].

Table 3. Sample means of $\hat{a}_{1m}$ in ten repetitions.

m	OLS	SLS	TLS	true value of a_{1m}
1	−0.228	0.127	0.128	0.12
2	0.157	0.209	0.206	0.10
3	−0.115	0.041	0.036	0.05

Table 4. Sample means of $\hat{a}_{0m}$ in ten repetitions.

m	OLS	SLS	TLS	true value of a_{0m}
1	0.030	−0.077	−0.081	−0.1
2	−0.103	−0.080	−0.084	−0.1
3	−0.198	−0.144	−0.149	−0.1

7. Comments on the estimation of frequency response functions

By taking the numerical Fourier transform of an estimate of $\{a_{jm}\}$ obtained by the present identification procedure we can get an estimate of the frequency response function. If the estimate of $\{a_{jm}\}$ is efficient, its Fourier transform also has the efficiency. Thus if the selection of the values of L and M is properly performed, a better result can generally be expected by this approach than by the conventional cross-spectral approach. Especially, when there exists a feedback from the output to the input the conventional cross-spectral method cannot give a consistent estimate [3].

These two aspects of consistency and efficiency in feedback case and non-feedback case, respectively, strongly suggest the superiority of this new approach to the conventional cross-spectral approach. The only possible practical difficulty of this approach might lie in the necessary computational complexity when $K \times M$ becomes very large.

Asymptotic confidence band for the estimate could be obtained from the results obtained in the former sections under appropriate assumptions.

It is very important to note the difference of the estimate of a frequency response function and that of the corresponding impulse response function with respect to the sampling variabilities. If there is a local frequency band where the estimate of the frequency response function shows a significant sampling fluctuation, the whole estimate of the impulse response function can be blurred by the error of this frequency band. The frequency components in the band with poor S/N ratio are sometimes to be suppressed to give a practically more useful estimate of the impulse response function with a small sacrifice of unbiasedness. This shows the necessity of observing the sampling variability of the estimate in the frequency domain.

8. Conclusion

The results obtained in this paper clarify the fact that the predictive representation of the disturbance process is playing a fundamental role in the identification problem and is forming a physical basis of the existence and use of white disturbance representation of the system.

The procedure described in this paper will be at present the most practical one to assure the consistency and efficiency of the estimates in the general situation of identification of noisy linear systems.

The numerical results suggest that the statistical characteristics of SLS estimates would be worthy of further analysis and evaluation.

Appendix I

The proof of the monotone non-increasing property of the asymptotic variance matrices of the successive estimates given in section 4, under the heading of ALS estimate, is given as follows.

We denote by $\{{}_2\hat{a}_{jm}\}$ the estimate of $\{a_{jm}\}$ which is obtained by replacing $\{\hat{c}_l\}$ by $\{{}_2\hat{c}_l\}$ in the defining formula of $\{\hat{a}_{jm}\}$. We also denote by $\{{}_3\hat{c}_l\}$ the estimate of $\{c_l\}$ which is obtained by replacing $\{\hat{a}_{jm}\}$ by $\{{}_2\hat{a}_{jm}\}$ in the defining formula of $\{{}_2\hat{c}_l\}$. We shall prove that the asymptotic variance matrices of $\{{}_3\hat{c}_l\}$ and $\{{}_2\hat{a}_{jm}\}$ are not greater than those of $\{{}_2\hat{c}_l\}$ and $\{\hat{a}_{jm}\}$, respectively. The proof is such that it can be extended indefinitely to further stages and thus gives the complete proof of the asserted non-increasing property. From the definition we have

$$\Delta c \approx R_{rx_0}^{-1} rx_0 w$$

$$\Delta a \approx R_y^{-1}(yw - R_{y,u}\Delta c)$$

$$\Delta_2 c \approx R_u^{-1}(uw - R_{u,y}\Delta a)$$

$$\Delta_2 a \approx R_y^{-1}(yw - R_{y,u}\Delta_2 c)$$

$$\Delta_3 c \approx R_u^{-1}(uw - R_{u,y}\Delta_2 a)\,,$$

where Δ denotes $N^{1/2}$ times the difference of the estimate from the true value. We have already shown, in the discussion of TLS estimate, that

$$E_\infty \Delta c(yw)' = O \qquad \text{(zero matrix).}$$

It can also be shown that

$$E_\infty \Delta c(uw)' = R_{rx_0}^{-1} E_\infty (rx_0 w)(uw)'$$
$$= \sigma^2 R_{rx_0}^{-1} R_{rx_0,u}\,.$$

As was mentioned in the discussion of TLS estimate $rx_0 = ru$ holds and accordingly $R_{rx_0,u} = R_{rx_0,ru} = R_{rx_0}$. Thus we get the relation

$$E_\infty \Delta c(uw)' = \sigma^2 I\,,$$

where I denotes the identity matrix of respective dimension. From this we can show that

$$E_\infty \Delta a(uw)' = O$$

and from the former relation $E_\infty \Delta c(yw)' = O$

$$E_\infty \Delta a(yw)' = \sigma^2 I\,.$$

It is clear that we can extend the reasoning successively and use the result to get the relation

$$E_\infty(\Delta a)(\Delta a)' = R_y^{-1}(\sigma^2 R_y + R_{y,u} E_\infty(\Delta c)(\Delta c)' R_{u,y}) R_y^{-1}$$

$$E_\infty(\Delta_2 a)(\Delta_2 a)' = R_y^{-1}(\sigma^2 R_y + R_{y,u} E_\infty(\Delta_2 c)(\Delta_2 c)' R_{u,y}) R_y^{-1}$$

$$E_\infty(\Delta_2 c)(\Delta_2 c)' = R_u^{-1}(\sigma^2 R_u + R_{u,y} E_\infty(\Delta a)(\Delta a)' R_{y,u}) R_u^{-1}$$

$$E_\infty(\Delta_3 c)(\Delta_3 c)' = R_u^{-1}(\sigma^2 R_u + R_{u,y} E_\infty(\Delta_2 a)(\Delta_2 a)' R_{y,u}) R_u^{-1}\,.$$

Thus, if we can show that $E_\infty(\Delta_2 c)(\Delta_2 c)' \leqq E_\infty(\Delta c)(\Delta c)'$, where $\leqq$ means that the right-hand side member is greater than or equal to the left-hand side member, we can see that $E_\infty(\Delta_2 a)(\Delta_2 a)' \leqq E_\infty(\Delta a)(\Delta a)'$ and accordingly $E_\infty(\Delta_3 c)(\Delta_3 c)' \leqq E_\infty(\Delta_2 c)(\Delta_2 c)'$. The reasoning can be extended indefinitely and the only thing we have to prove is the relation $E_\infty(\Delta_2 c)\cdot(\Delta_2 c)' \leqq E_\infty(\Delta c)(\Delta c)'$.

From the above relation we have

$$E_\infty(\Delta_2 c)(\Delta_2 c)' = R_u^{-1}(\sigma^2 R_u + \sigma^2 R_{u,y} R_y^{-1} R_{y,u} + R_{u,y} R_y^{-1} R_{y,u} E_\infty(\Delta c)(\Delta c)' R_{u,y} R_y^{-1} R_{y,u}) R_u^{-1} ,$$

and from the former result

$$E_\infty(\Delta c)(\Delta c)' = \sigma^2 R_{rx_0}^{-1} = \sigma^2 R_{ru}^{-1} .$$

Thus we have only to show that

$$R_u R_{ru}^{-1} R_u \geqq R_u + R_{u,y} R_y^{-1} R_{y,u} + R_{u,y} R_y^{-1} R_{y,u} R_{ru}^{-1} R_{u,y} R_y^{-1} R_{y,u} .$$

For this purpose we apply the general matrix identity

$$(I-S)^{-1} = I + S + S(I-S)^{-1} S$$

to $S = (\sqrt{R_u})^{-1} R_{pu} (\sqrt{R_u})^{-1}$, where $R_{pu} = R_u - R_{ru}$ and $\sqrt{R_u}$ denotes the positive definite square root matrix of R_u, and we get

$$R_u R_{ru}^{-1} R_u = R_u + R_{pu} + R_{pu} R_{ru}^{-1} R_{pu} .$$

Taking into account of the fact that R_{pu} is the variance matrix of the projections of $u(n-1), u(n-2), \cdots, u(n-L)$ on $\{x_j(n-m);\ j=1, 2, \cdots, K, m=0, 1, \cdots, M+L\}$ and that $R_{u,y} R_y^{-1} R_{y,u}$ is the variance matrix of the projections of $u(n-1), u(n-2), \cdots, u(n-L)$ on $\{y_j(n-m);\ j=1, 2, \cdots, K,\ m=0, 1, \cdots, M\}$ we can see that $R_{pu} \geqq R_{u,y} R_y^{-1} R_{y,u}$ and accordingly $R_u R_{ru}^{-1} R_u \geqq R_u + R_{u,y} R_y^{-1} R_{y,u} + R_{u,y} R_y^{-1} R_{y,u} R_{ru}^{-1} R_{u,y} R_y^{-1} R_{y,u}$. This completes the proof.

Appendix II

Her we will show that the asymptotic variance matrix of the ALS estimate is not greater than that of the TLS estimate.

In the discussion of the SLS estimate we noticed the following relation:

$$\Delta a_{jm} \approx \sum_{l=1}^{m} \Delta c_l (a_j * b)_{m-l} + \sum_{k=0}^{m} \Delta A_{jk} b_{m-k} .$$

From this the asymptotic variance matrix of the TLS estimate is obtained as follows:

$$E \begin{bmatrix} \Delta c \\ \Delta a \end{bmatrix} [(\Delta c)' (\Delta a)'] = \sigma^2 B R^{-1} B' ,$$

where B is the matrix of the linear transformation which transforms $[\Delta c_1, \cdots, \Delta c_L, \Delta A_{10}, \cdots, \Delta A_{1M+L}, \cdots, \Delta A_{K0}, \cdots, \Delta A_{KM+L}]'$ into $[\Delta c_1, \cdots, \Delta c_L,$

$\Delta b_{10}, \cdots, \Delta b_{1M}, \cdots, \Delta b_{K0}, \cdots, \Delta b_{KM}]'$ where $\Delta b_{jm}=\sum_{l=1}^{m} \Delta c_l (a_j * b)_{m-l}+\sum_{k=0}^{m} \Delta A_{jk} b_{m-k}$ and R is the variance matrix of $\{x_0(n-1), \cdots, x_0(n-L), x_1(n), \cdots, x_1(n-M-L), \cdots, x_K(n), \cdots, x_K(n-M-L)\}$. The asymptotic variance matrix of the ALS estimate was given as $\sigma^2 R_{u,y}^{-1}$, where $R_{u,y}$ denotes the variance matrix of $\{u(n-1), \cdots, u(n-L), y_1(n), \cdots, y_1(n-M), \cdots, y_K(n), \cdots, y_K(n-M)\}$. Obviously we have the relation

$$R_{u,y}=CRC' ,$$

where C is the transformation matrix which gives $[u(n-1), \cdots, u(n-L), y_1(n), \cdots, y_1(n-M), \cdots, y_K(n), \cdots, y_K(n-M)]'$ from $[x_0(n-1), \cdots, x_0(n-L), \cdots, x_1(n), \cdots, x_1(n-M-L), \cdots, x_K(n), \cdots, x_K(n-M-L)]'$. It is stated in the book of Grenander and Rosenblatt ([5], p. 88) that for any two $n \times s$ matrices l^* and ϕ, where $*$ denotes the conjugate matrix, we have

$$\phi^* \phi \geqq (l\phi)^*(ll^*)^{-1}(l\phi) .$$

By putting $l=C\sqrt{R}$ and $\phi=(\sqrt{R})^{-1}B'$, where $\sqrt{R}$ is the unique positive definite square root of R, we get

$$BR^{-1}B' \geqq (CB')'(CRC')^{-1}(CB') .$$

By using the relation $\left(\sum_{l=0}^{\infty} b_l z^l\right)\left(1-\sum_{l=1}^{L} c_l z^l\right)=1$ we can show that $CB'=I$ (identity matrix) and thus we have

$$BR^{-1}B' \geqq (CRC')^{-1} .$$

This completes the proof.

Acknowledgements

The present paper is a direct consequence of the cooperative study of a cement rotary kiln by Dr. T. Nakagawa of Chichibu Cement Company and the author. All the necessary programmings for the numerical experiments were prepared by Miss E. Arahata.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] H. Akaike, "On the use of a linear model for the identification of feedback systems," *Research Memorandum*, No. 20, 1968, The Institute of Statistical Mathematics, April, 1968 (To be published in the *Ann. Inst. Statist. Math.*, 20 (1968), 425-439.

[2] J. Durbin, "Estimation of parameters in time-series regression models," *J. R. Statist. Soc.*, B, 22 (1960), 139-153.

[3] H. Akaike, "Some problems in the application of the cross-spectral method," *Spectral Analysis of Time Series* (ed. B. Harris), New York, Wiley, (1967), 81-107.

[4] H. Akaike, "On the statistical estimation of the frequency response function of a system having multiple input," *Ann. Inst. Statist. Math.*, 17 (1965), 185-210.

[5] U. Grenander and M. Rosenblatt, *Statistical Analysis of Stationary Time Series*, New York, Wiley, 1957.

[6] H. Wold, *A Study in the Analysis of Stationary Time Series*, Uppsala: Almqvist & Wiksells, 1938, Second Edition 1954.

[7] T. Otomo, T. Nakagawa and H. Akaike, "Implementation of computer control of a cement rotary kiln through data analysis," (To be presented at the 4th Congress of IFAC in Warszawa, June 16-21, 1969).

[8] E. Parzen, "Analysis and synthesis of linear models for time series," *Technical Report*, No. 4, Department of Statistics, Stanford University, 1966.