STATISTICAL PREDICTOR IDENTIFICATION

HIROTUGU AKAIKE

(Received Dec. 26, 1969)

1. Introduction and summary

In a recent paper by the present author [1] a simple practical procedure of predictor identification has been proposed. It is the purpose of this paper to provide a theoretical and empirical basis of the procedure.

Our procedure is based on a figure of merit of a predictor, which is called the final prediction error (FPE) and is defined as the mean square prediction error of the predictor. We consider the application of the least squares method for the identification of the predictor when the stochastic process under observation is an autoregressive process generated from a strictly stationary and mutually independent innovations. The identification is realized by fitting autoregressive models of successive orders within a prescribed range, computing estimates of FPE for the models, and adopting the one with the minimum of the estimates.

The statistical characteristics of these estimates of FPE and the overall procedure are discussed to show the practical utility of the procedure. A modified version of this original procedure is proposed, which shows a consistency, as an estimation procedure of the order of a finite order autoregressive process, which is lacking in the original procedure. The notion of FPE is also applied for the determination of the constants of the decision procedure, which was proposed by T. W. Anderson [2] for the decision of the order of a Gaussian autoregressive process, to provide a third procedure.

Performances of the three types of procedures, the original one, a modified version and that of Anderson's type, are compared by using various realizations of artificial time series. The results show that for practical applications, where the true orders of autoregressive processes would generally be infinite, the original procedure would be the most useful.

Implication of the present identification procedure on the estimation of power spectra will be discussed in a subsequent paper [3].

We shall use the convention of denoting by $(u(l))$ the column vector of $u(l)$ $(l=1, 2, \cdots, M)$ and by v or $(v(l, m))$ the matrix with (l, m) ele-

ment $v(l, m)$ $(l, m=1, 2, \cdots, M)$. When the dimension M is of special interest we shall add the subscript M and thus u_M and v_M are used for the above u and v. The symbol $'$ will be used to denote the transpose of a matrix or a vector.

2. Definition of FPE of a predictor and the statement of the problem

Here we first introduce a general definition of a figure of merit of a predictor. This is defined simply as the mean square prediction error and is called the FPE (final prediction error) of the predictor, i.e., for a predictor $\hat{X}(n)$ of $X(n)$

$$\text{FPE of } \hat{X}(n)=E(X(n)-\hat{X}(n))^2 . \tag{2.1}$$

In practical situations $\hat{X}(n)$ is given as a function of the recent values of $X(n)$ and the structure or the parameter of the function is determined, or identified, by using the whole past history of $X(n)$. Assuming the dependency of this identified structure on the recent values of $X(n)$ which are to be used to give $\hat{X}(n)$ to be decreasing as the length of the past history used for the identification is increased, we consider the idealized situation where the dependency is completely vanishing. This is equivalent to the situation where the structure of a predictor is identified by using an observation of a process $X(n)$ and, using the structure, the prediction is made with another process $Y(n)$ which is independent of $X(n)$ but with one and the same statistical property as $X(n)$.

When the process $X(n)$ is stationary and the predictor $\hat{Y}(n)$ of $Y(n)$ is linear and given by

$$\hat{Y}(n)=\sum_{m=1}^{M}\hat{a}_M(m)Y(n-m)+\hat{a}_M(0) , \tag{2.2}$$

where $\hat{a}_M(m)$ is a function of $\{X(n)\}$, we have

$$\text{FPE of } \hat{Y}(n)=\sigma^2(M)+\sum_{l=0}^{M}\sum_{m=0}^{M}E(\Delta a_M(l)\Delta a_M(m))V_{M+1}(l, m) , \tag{2.3}$$

where

$$\begin{aligned}\sigma^2(M)&=E(Y(n)-\sum_{m=1}^{M}a_M(m)Y(n-m)-a_M(0))^2\\ &=\operatorname*{Min}_{\{a(m)\}}E(Y(n)-\sum_{m=1}^{M}a(m)Y(n-m)-a(0))^2 ,\end{aligned} \tag{2.4}$$

$$V_{M+1}(l, m)=EY(n-l)Y(n-m) \qquad l, m=1, 2, \cdots, M , \tag{2.5}$$

$$V_{M+1}(0, m)= V_{M+1}(m, 0)$$
$$=EY(n) \qquad m=1, 2, \cdots, M ,$$
$$V_{M+1}(0, 0) =1 ,$$

and

$$\Delta a_M(m)=\hat{a}_M(m)-a_M(m) \qquad m=0, 1, \cdots, M , \tag{2.6}$$

where $a_M(m)$ is defined by (2.4) and is giving the best (in the sense of mean square) linear predictor. (2.3) shows that the FPE in this case is composed of two components: the first one corresponding to the FPE of the best linear predictor for a given M and the second one due to the statistical deviation of $\hat{a}_M(m)$ from $a_M(m)$. Generally, as the value of M is increased, the first term $\sigma^2(M)$ will decrease but the second term will increase for a finite length of observation of $X(n)$.

Given a set of predictors the definition of FPE naturally suggests the adoption of the predictor with the minimum value of FPE as optimum. The problem we are concerned with in this paper is the realization of a good approximation to the optimum choice of M for the above stated stationary and linear case, using the information obtained by observing $X(n)$ for a finite length of time.

3. FPE of the least squares estimate of an autoregressive model

Hereafter we shall assume that $X(n)$ is a stationary autoregressive process generated by the relation

$$X(n)=\sum_{m=1}^{M} a(m)X(n-m)+a(0)+\varepsilon(n) , \tag{3.1}$$

where $\varepsilon(n)$'s are mutually independently and identically distributed random variables with $E\varepsilon(n)=0$ and $E\varepsilon^2(n)=\sigma^2$.

Given a set of data $\{X(n);\ n=-M+1, -M+2, \cdots, N\}$, the parameter $\hat{a}_M(m)$ of our predictor is defined as the least squares estimate of $a(m)$, i.e., $\hat{a}_M(m)$ is the solution of

$$\sum_{m=1}^{M} C_{xx}(m, l)\hat{a}_M(m)=C_{xx}(0, l) \qquad l=1, 2, \cdots, M , \tag{3.2}$$

and

$$\hat{a}_M(0)=\bar{X}_0-\sum \hat{a}_M(m)\bar{X}_m , \tag{3.3}$$

where

$$\bar{X}_m=N^{-1}\sum_{n=1}^{N} X(n-m) \qquad (m=0, 1, 2, \cdots, M)$$

and

$$C_{xx}(m, l)=N^{-1}\sum_{n=1}^{M}(X(n-m)-\bar{X}_m)(X(n-l)-\bar{X}_l) .$$

Following the definition of $Y(n)$ given in the preceding section, our predictor $\hat{Y}(n)$ of $Y(n)$ is given in this case by

$$\hat{Y}(n)=\sum_{m=1}^{M}\hat{a}_M(m)(Y(n-m)-\bar{X}_m)+\bar{X}_0 . \tag{3.4}$$

We are assuming that $Y(n)$ is generated by the relation $Y(n)=\sum_{m=1}^{M}a_m Y(n-m)+a_0+\delta(n)$, where $\delta(n)$ has one and the same statistical property as $\varepsilon(n)$. We have

$$Y(n)-\hat{Y}(n)=\delta(n)-\sum_{m=1}^{M}\Delta a_M(m)\, y(n-m)-(\Delta\bar{X}_0-\sum_{m=1}^{M}\hat{a}_M(m)\Delta\bar{X}_M) ,$$

where $y(n)=Y(n)-E(Y(n))$ and $\Delta\bar{X}_l=\bar{X}_l-E(X(n))$.

Taking into account the independency of $y(n)$ of Δa_M and $\Delta\bar{X}_l$ we get

$$\begin{aligned}\text{(3.5)}\qquad \text{FPE of } \hat{Y}(n)&=E(Y(n)-\hat{Y}(n))^2\\ &=\sigma^2+\sum_{m=1}^{M}\sum_{l=1}^{M}E(\Delta a_M(m)\Delta a_M(l))R_{xx}(l-m)\\ &\quad+E(\Delta\bar{X}_0-\sum_{m=1}^{M}\hat{a}_M(m)\Delta\bar{X}_m)^2 ,\end{aligned}$$

where

$$R_{xx}(l-m)=EX(n-l)X(n-m)-(EX(n))^2 .$$

For the asymptotic evaluation of this FPE of $\hat{Y}(n)$ we make use of the following basic theorem.

THEOREM 1. *Under the present assumption of $X(n)$, the limit distribution $\sqrt{N}\Delta\bar{X}_0=\sqrt{N}(X_0-E(X(n)))$ and $\sqrt{N}\Delta a_M(m)=\sqrt{N}(\hat{a}_M(m)-a(m))$ $(m=1, 2, \cdots, M)$, when N tends to infinity, is $(M+1)$-dimensional Gaussian with zero mean and the variance matrix*

$$\sigma^2\begin{pmatrix}\delta^{-2} & 0'_M\\ 0_M & R_M^{-1}\end{pmatrix}, \tag{3.6}$$

where $\delta=1-\sum_{m=1}^{M}a(m)$, R_M is the $M\times M$ matrix of $R(l, m)=R_{xx}(l-m)$ and 0 denotes a zero vector.

From the ergodicity of the process $X(n)$ we know that $C_{xx}(l, m)$ converges to $R_{xx}(l-m)$, as N tends to infinity, with probability one. Thus $\hat{a}_M$ is a consistent estimate of a_M, in this case with convergence with probability one. From (3.2) we have, for $l=1, 2, \cdots, M$,

$$(3.7)\qquad (\hat{a}_M(l))=(C_{xx}(m,l))^{-1}(C_{xx}(0,l)) \\ =(a_M(l))+(C_{xx}(m,l))^{-1}(C_{\varepsilon x}(l))\ ,$$

where $C_{\varepsilon x}(l)=N^{-1}\sum_{n=1}^{N}\varepsilon(n)(X(n-l)-\bar{X}_l)$. Thus we get

$$(3.8)\qquad (\Delta a_M(l))=(C_{xx}(m,l))^{-1}(C_{\varepsilon x}(l))\ .$$

From the consistency of $C_{xx}(m,l)$ we know that the limit distribution of $\sqrt{N}\,\Delta\bar{X}_0$ and $\sqrt{N}\,\Delta a_M$ is identical to that of $\sqrt{N}\,\Delta\bar{X}_0$ and $\sqrt{N}\,R_M^{-1}C_{\varepsilon x}$. By applying the Diananda's central limit theorem [4] for finitely dependent sequence, as was done by Anderson and Walker [5], we can easily get

LEMMA. *The limit distribution of $\sqrt{N}\,\Delta\bar{X}_0$ and $\sqrt{N}C_{\varepsilon x}$, when N tends to infinity, is $(M+1)$-dimensional Gaussian with zero mean and the variance*

$$\sigma^2\begin{pmatrix}\delta^{-2} & 0_M' \\ 0_M & R_M\end{pmatrix}.$$

It should be noted that as the power spectral density of $X(n)-EX(n)$ at zero frequency is $\sigma^2\delta^{-2}$, where $\delta=1-\sum_{m=1}^{M}a_M(m)$, the variance of the limit distribution of $\sqrt{N}\,\Delta\bar{X}_0$ is equal to $\sigma^2\delta^{-2}$. The assertion of Theorem 1 is a direct consequence of this lemma and the observation following (3.8).

Now we return to the evaluation of FPE of $\hat{Y}(n)$. Instead of taking the expectation of $(Y(n)-\hat{Y}(n))^2$ directly as suggested in (3.5) we first take the conditional expectation of $(Y(n)-\hat{Y}(n))^2$ for a given $X(n)$. This we will denote by $E_x(Y(n)-\hat{Y}(n))^2$. From the independency of $Y(n)$ of $X(n)$ we have

$$(3.9)\qquad E_x(Y(n)-\hat{Y}(n))^2=\sigma^2+\sum_{m=1}^{M}\sum_{l=1}^{M}\Delta a_M(m)\Delta a_M(l)R_{xx}(l-m) \\ +(\Delta\bar{X}_0-\sum_{m=1}^{M}\hat{a}_M(m)\Delta\bar{X}_m)^2\ .$$

Taking into account the fact that in the limit the differences between $\sqrt{N}\,\Delta\bar{X}_0$ and $\sqrt{N}\,\Delta\bar{X}_m$ $(m=1,2,\cdots,M)$ are stochastically vanishing, we can see from the theorem that $N\{E_x(Y(n)-\hat{Y}(n))^2-\sigma^2\}$ has a limit distribution with expectation equal to $(M+1)\sigma^2$. This observation suggests the following definition of $(\mathrm{FPE})_M$ as an asymptotic evaluation of FPE of $\hat{Y}(n)$:

$$\text{(3.10)} \qquad (\text{FPE})_M \text{ of } \hat{Y}(n)=\left(1+\frac{M+1}{N}\right)\sigma^2 .$$

Our identification procedure of the predictor will be based on some estimate of $(\text{FPE})_M$.

4. An estimate of $(\text{FPE})_M$ and the minimum FPE procedure

From the ergodicity of $X(n)$ we know that

$$\text{(4.1)} \qquad S(M)=C_{xx}(0,0)-\sum_{l=1}^{M}\hat{a}_M(l)C_{xx}(0,l)$$

is a consistent estimate of σ^2. By (3.2) we have

$$S(M)=C_{xx}(0,0)-\sum_{l=1}^{M}\sum_{m=1}^{M}\hat{a}_M(l)C_{xx}(m,l)\hat{a}_M(m) ,$$

and by taking into account the relation

$$\sum_{m=1}^{M}\Delta a_M(m)C_{xx}(m,l)=C_{xx}(0,l)-\sum_{m=1}^{M}C_{xx}(m,l)a(m)$$

we get

$$\text{(4.2)} \qquad S(M)=C_{xx}(0,0)-2\sum_{m=1}^{M}a(m)C_{xx}(0,m)+\sum_{l=1}^{M}\sum_{m=1}^{M}a(l)a(m)C_{xx}(m,l)$$
$$-\sum_{l=1}^{M}\sum_{m=1}^{M}\Delta a_M(l)\Delta a_M(m)C_{xx}(m,l) .$$

From the definition of $C_{xx}(m,l)$ we have

$$\text{(4.3)} \qquad H(M)=C_{xx}(0,0)-2\sum_{m=1}^{M}a(m)C_{xx}(0,m)+\sum_{l=1}^{M}\sum_{m=1}^{M}a(m)a(l)C_{xx}(m,l)$$
$$=N^{-1}\sum_{n=1}^{N}(\varepsilon(n)-\bar{\varepsilon})^2 ,$$

and we get

$$\text{(4.4)} \qquad E(H(M))=(1-N^{-1})\sigma^2 .$$

From Theorem 1 we know that when we assume the model (3.1) the limit distribution of

$$\text{(4.5)} \qquad Q(M)=N\sum_{l=1}^{M}\sum_{m=1}^{M}\Delta a_M(m)\Delta a_M(l)C_{xx}(m,l)$$

has expectation $M\sigma^2$, i.e.,

$$\text{(4.6)} \qquad E_\infty\{Q(M)\}=M\sigma^2 ,$$

where E_∞ denotes the expectation of the limit distribution of the quantily

within the braces, when N tends to infinity. These observations suggest that it would be reasonable to adopt $(1-N^{-1}(M+1))^{-1}S(M)$ as an estimate of σ^2 to define our estimate (FPE)(M) of $(\text{FPE})_M$ by

$$(4.7) \qquad (\text{FPE})(M)=(1+N^{-1}(M+1))(1-N^{-1}(M+1))^{-1}S(M)\ .$$

The discussions in this and the preceding sections naturally lead us to the idea that when there are many predictors obtained by applying the least squares method it would be reasonable for us to pick the one with the minimum value of (FPE)(M). Following this idea, for the identification of the predictor by a single record of $X(n)$, we proceed as follows; we compute (FPE)(M) successively for $M=0, 1, \cdots, L$ (L; preassigned positive integer) and adopt $\hat{a}_M$ with $M=M_0$ to define the predictor, where (FPE)(M_0)=the minimum of (FPE)(M) ($M=0, 1, \cdots, L$). This process which was called by the name of FPE scheme in the former paper [1] will hereafter be called the minimum FPE procedure.

5. Statistical properties of (FPE)(M)

To see the practical utility of the minimum FPE procedure we shall have first to analyze the statistical characteristics of (FPE)(M) ($M=0, 1, \cdots, L$) for a fixed model of $X(n)$. We assume that the order of $X(n)$ is K, i.e., $a_K\neq 0$ and $a_m=0$ for $m>K$ in (3.1). We assume $K\geqq 0$ and exclude the case where $K=-1$, with $a_0=0$, from our discussion. We also assume that the set of data is given in a form $\{X(n);\ n=-L+1, -L+2, \cdots, 1, 2, \cdots, N\}$. From (4.3) we can see that $H(M)$ remains constant for $M\geqq K$ and thus the behavior of $S(M)$ is dependent only on $Q(M)$ of (4.5). From the discussion of Section 3 we know that the limit distribution of $Q(M)$ ($M=K, K+1, \cdots, L$) is identical to that of $NC'_{xM}R_M^{-1}C_{xM}$, where $C_{xM}=(C_{xx}(l))$ ($l=1, 2, \cdots, M$). As was stated in the lemma of Section 3 the covariance matrix of the limit distribution of $\sqrt{N}C_{xM}$ is identical to that of $\{X(n-m)-EX(n-m):\ m=1, 2, \cdots, M\}$ multiplied by σ^2, i.e., σ^2R_M. Thus the successive orthonormalization procedure of $X(n-m)-EX(n-m)$ ($m=1, 2, \cdots, M$) can be applied to $\sqrt{N}C_{xM}$ to give a vector random variable U_M of which limit distribution is the M-dimensional unit normal distribution. The detail of this transformation is already described in [6]. We have $U_M=T_M\sqrt{N}C_{xM}$ with $\sigma^2T_MR_MT'_M=I_M$, where I denotes the identity matrix, and the matrix T_M of the transformation has zeros above the diagonal. From the structure of T_M it is readily seen that

$$T_M(l, m)=T_L(l, m) \qquad (l, m=1, 2, \cdots, M) \quad \text{for} \quad M<L\ ,$$

i.e., the submatrix of the first $M\times M$ elements of T_L is identical to T_M.

Thus we can see that U_M is the vector of the first M elements of U_L. From this observation we can see that the limit distribution of $Q(M)$ is identical to that of $\sigma^2 \sum_{l=1}^{M} U^2(l)$ $(M=K, K+1, \cdots, L)$. The limit distribution of $U^2(l)$ $(l=1, 2, \cdots, L)$ is then the distribution of mutually independent chi-square variables each with d.f.1 [7]. Thus we get

THEOREM 2. *For $M \geqq K$, $\sigma^{-2}Q(M)$ is asymptotically distributed as the partial sum of the first M terms of a sequence of mutually independently distributed chi-square variables each with d.f.* 1.

Now we proceed to the analysis of the statistical behavior of (FPE)(M) $(M=0, 1, \cdots, L)$. For any positive integer M, we define $a_M(m)$, irrespectively of the order K, as the solution of (3.2) when $C_{xx}(m, l)$ is replaced by $R_{xx}(l-m)$ and define $\sigma^2(M)$ by

$$\sigma^2(M)=R_{xx}(0)-\sum_{m=1}^{M} a_M(m)R_{xx}(m) \,. \tag{5.1}$$

We shall denote $\Delta a_M(m)=\hat{a}_M(m)-a_M(m)$, where $\hat{a}_M(m)$ is the solution of (3.2). M is not restricted to be equal or larger than the order K. Corresponding to (4.2) it holds that

$$\begin{aligned} S(M) &= C_{xx}(0, 0)-\sum_{l=1}^{M} \hat{a}_M(l)C_{xx}(0, l) \\ &= C_{xx}(0, 0)-2\sum_{m=1}^{M} a_M(m)C_{xx}(0, m)+\sum_{l=1}^{M}\sum_{m=1}^{M} a_M(l)a_M(m)C_{xx}(m, l) \\ &\quad -\sum_{l=1}^{M}\sum_{m=1}^{M} \Delta a_M(l)\Delta a_M(m)C_{xx}(m, l) \,. \end{aligned} \tag{5.2}$$

Ignoring the terms of order N^{-2}, we have approximately

$$\begin{aligned} (\mathrm{FPE})(M_1)-(\mathrm{FPE})(M_2) &= (1-N^{-1}(M_1+1))^{-1}(1-N^{-1}(M_2+1))^{-1} \\ &\quad \cdot (S(M_1)-S(M_2)-N^{-1}(M_2-M_1)(S(M_1)+S(M_2))) \,, \end{aligned} \tag{5.3}$$

where $0 \leqq M_1$, $M_2 \leqq L$. If we assume the equality (5.3) to be strict, we have for $M_1 < M_2$

$$\begin{aligned} &\mathrm{Prob}\,\{(\mathrm{FPE})(M_1)-(\mathrm{FPE})(M_2)>0\} \\ &\quad \geqq \mathrm{Prob}\,\{S(M_1)-S(M_2)-2N^{-1}(M_2-M_1)S(M_1)>0\} \\ &\quad \geqq \mathrm{Prob}\,\{(S(M_1)-S(M_2))((M_2-M_1)C_{xx}(0))^{-1}>2N^{-1}\} \,. \end{aligned} \tag{5.4}$$

From (5.2) we have

$$\begin{aligned} S(M_1)-S(M_2) &= \sigma^2(M_1)-\sigma^2(M_2)+\Delta(\sigma^2(M_1)-\sigma^2(M_2)) \\ &\quad +N^{-1}(Q(M_2)-Q(M_1)) \,, \end{aligned} \tag{5.5}$$

where $\Delta(\sigma^2(M_1)-\sigma^2(M_2))$ is obtained by replacing $R_{xx}(m, l)$ by $\Delta R_{xx}(m)=$

$C_{xx}(m, l)-R_{xx}(m, l)$ in the definition of $\sigma^2(M_1)-\sigma^2(M_2)$ and $Q(M)$ is as defined in (4.5). For M_1 and M_2 which are very small compared with N and for which the differences $a_{M_1}(m)-a_{M_2}(m)$ $(m=1, 2, \cdots, M_2;\ a_{M_1}(m)=0$ for $m>M_1)$ are of the order of $N^{-1/2}$, $R_{xx}^{-1}(0)(\sigma^2(M_1)-\sigma^2(M_2))$ will be of the order of $N^{-1/2}$ from (5.1), while $C_{xx}^{-1}(0, 0)\Delta(\sigma^2(M_1)-\sigma^2(M_2))$ and $C_{xx}^{-1}(0, 0)(Q(M_2)-Q(M_1))N^{-1}$ are stochastically of the order of N^{-1}. Thus the probability of (5.4) will be very nearly equal to 1 in this case. Generally we have, for $M<K$,

$$\lim_{N\to\infty} \text{Prob}\ \{(\text{FPE})(M)-(\text{FPE})(K)>0\}=1\ . \tag{5.6}$$

By (5.3) and the fact that $S(M)$ is a consistent estimate of σ^2 for $M\geqq K$ we can see that the limit distribution of $N((\text{FPE})(K)-(\text{FPE})(M))$ $(M=K+1, K+2, \cdots, L)$ is identical to that of $N(S(K)-S(M))-2\sigma^2(M-K)$. As it hold that, for $M\geqq K$, $S(K)-S(M)=N^{-1}(Q(M)-Q(K))$, we can see from Theorem 2 that the limit distribution of $N\sigma^{-2}((\text{FPE})(K)-(\text{FPE})(M))+2(M-K)$ is identical to the distribution of the successive sum of $M-K$ chi-square variables $\chi_1^2(i)$ $(i=1, 2, \cdots)$ which are mutually independently distributed each with d.f.1. Thus for this case we have

$$\begin{aligned}\lim_{N\to\infty} &\text{Prob}\ \{(\text{FPE})(K)>(\text{FPE})(M)\} \\ &=\text{Prob}\ \{\sum_{i=1}^{M-K}\chi_1^2(i)>2(M-K)\}\ .\end{aligned} \tag{5.7}$$

We can see from (5.6) that by using $(\text{FPE})(M)$ for our minimum FPE procedure the probability of adopting M smaller than K as M_0 will be made arbitrarily small when N is increased indefinitely, while (5.7) shows that for $M>K$ the probability of observing $(\text{FPE})(M)$ small than $(\text{FPE})(K)$ tends to a non-zero constant. This last observation shows that the value M_0 of M adopted by our minimum FPE procedure as the order of the predictor is not a consistent estimate of K. This does not necessarily mean a serious draw back of the procedure for practical applications. The probability itself, as suggested by (5.7), of adopting M_0 larger than K is not necessarily intolerable for practical applications. Further, it will be more common for us to encounter with the situation where the theoretical value of K is considered to be infinity. For this case the result of the foregoing discussion following (5.5) suggests that the probability of M_0 being equal to an M for which $|a_M(m)-a(m)|$ is larger than $N^{-1/2}$ and the corresponding $\sigma^2(M)$ is differing from σ^2 by a quantity greater than $N^{-1/2}R_{xx}(0)$ would be very small. Also (5.5) suggests that in this case (5.7) will hold approximately when the diffference of $\sigma^2(K)$ from σ^2 is made significantly smaller than N^{-1}.

Admittedly our present analysis is quite rough for the range of $M<K$. We will supplement the discussion with numerical examples in Section 8.

The procedure which is obtained by replacing the definition of (FPE)(M) in the minimum FPE procedure by

$$(5.8) \qquad (\mathrm{FPE})^{\alpha}(M)=(1+N^{-\alpha}(M+1))(1-N^{-1}(M+1))^{-1}S(M)\ ,$$

where $0<\alpha<1$, will be called the minimum (FPE)$^{\alpha}$ procedure. By this modification we shall certainly obtain the consistency of M_0 as an estimate of K of a finite order autoregressive process. But the modification may add much to the tendency of M_0 taking values too small for the minimization of FPE. In Section 8, the performance of the minimum (FPE)$^{1/4}$ procedure will be compared with that of the original procedure.

6. FPE and Anderson's procedure

T. W. Anderson [2] has given a multiple decision procedure for choosing the order of dependence K in normally distributed time series of the type (3.1). The procedure is such that it is completely specified by, and optimum for, a selection of probabilities p_l $(m<l\leqq q)$ for some preassigned m and q, where p_l is the probability of deciding on the order of dependence to be l when the actual order is less than l. Thus in this procedure we are going to keep small the probabilities $q_l=\sum_{\nu=l}^{q}p_\nu$ $(l=m+1, m+2, \cdots, q)$ of errors of choosing a higher order than necessary. On the other hand, we shall have to keep p_l as large as possible within some allowable limit to maintain the sensitivity of the procedure to non-zero autoregression coefficients. If we evaluate the loss, incurred by adopting a higher order than necessary, by FPE, it would be more natural to control the quantities

$$(6.1) \qquad Q_l=\sum_{\nu=l}^{q}(1+N^{-1}(\nu-l+1))\sigma^2p_\nu \qquad (l=m+1, m+2, \cdots, q)\ ,$$

rather than the probabilities q_l. Obviously Q_{m+1} takes the largest value among Q_l and we decide to pay our attention only to this maximum possible loss. We state the allowable limit of this maximum possible loss relatively to the value of FPE for the order m, i.e., we require Q_{m+1} to be less than or equal to $\rho(1+N^{-1}m)\sigma^2$, where ρ is a small positive quantity such as 0.1 and the like. To keep the sensitivity to a possible non-zero $a(l)$ it is necessary to choose p_l as large as possible, but this also contributes to Q_{m+1} with the corresponding amount of $(1+N^{-1}(l-m))\sigma^2p_l$. We introduce here the principle of equal harmfulness which states that these losses $(1+N^{-1}(l-m))\sigma^2p_l$ $(l=m+1, m+2, \cdots, q)$ should all be equal to a positive quantity $\gamma\sigma^2$. By this principle our set of probabilities p_l $(l=m+1, m+2, \cdots, q)$ is determined as follows:

1) Define the allowable relative amount of loss ρ $(<(1+N^{-1}m)^{-1})$.
2) Obtain the value γ by the relation

$$(q-m)\gamma=(1+N^{-1}m)\rho \ . \tag{6.2}$$

3) p_l is given by

$$p_l=(1+N^{-1}(l-m))^{-1}\gamma \qquad (l=m+1, m+2, \cdots, q) \ . \tag{6.3}$$

When we assume that the partial serial correlations, $\hat{a}_M(M)$'s in the formulation of (3.2), are distributed mutually independently and symmetrically around zero when the true order is less than M, the Anderson's procedure is realized by testing the partial serial correlation $\hat{a}_M(M)$ against zero successively for $M=q, q-1, \cdots, m+1$ and taking M_0 equal to the first and the largest M for which $\hat{a}_M(M)$ is decided to be significant. The level of significance β_M and the corresponding critical value δ_M of each test is given by the relations

$$\begin{aligned} &\beta_M=\text{Prob}\,\{|\hat{a}_M(M)|>\delta_M\} \ , \\ &\beta_q=p_q \ , \\ &\beta_M=p_M\prod_{l=M+1}^{q}(1-\beta_l)^{-1} \qquad (M=q-1, q-2, \cdots, m+1) \ . \end{aligned} \tag{6.5}$$

If we adopt the approximation that $N|\hat{a}_M(M)|^2$ is distributed as a chi-square variable with d.f.1, δ_M is very simply obtained by using the table of chi-square or Gaussian distribution.

In the following discussion of numerical results we shall exclusively adopt this chi-square approximation along with the constants $\rho=0.1$, $m=0$ and $q=L$.

7. A practical version of the procedures

For practical applications of the three procedures we propose the following modification. Given a set of data $\{X(n); n=1, 2, \cdots, N\}$ we replace $\bar{X}_l$ and $C_{xx}(l, m)$ in the foregoing description of the procedures by $\bar{X}$ and $C_{xx}(l-m)$, respectively, where by definition

$$\bar{X}=N^{-1}\sum_{n=1}^{N}X(n) \tag{7.1}$$

and

$$C_{xx}(k)=N^{-1}\sum_{n=1}^{N-|k|}(X(n+|k|)-\bar{X})(X(n)-\bar{X}) \ . \tag{7.2}$$

By this modification we lose nothing but the relation (4.4) in the preceding discussions. Above all, the result of discussions in Section 5 re-

mains valid and we can expect that the practical usefulness of the original procedures is not affected by this modification.

The modification introduces a great simplification into the computational procedure, especially when we take into account the fact [8, 9] that the computations of (3.2) and (4.1) for $\hat{a}_M(m)$ and $S(M)$ can most easily be carried out by using the recursive relations

$$\begin{aligned}
&\hat{a}_{M+1}(M+1)=(S(M))^{-1}(C_{xx}(M+1)-\sum_{m=1}^{M}\hat{a}_M(m)C_{xx}(M+1-m)) ,\\
&\hat{a}_{M+1}(m)=\hat{a}_M(M)-\hat{a}_{M+1}(M+1)\hat{a}_M(M+1-m) \qquad m=1, 2, \cdots, M , \qquad (7.3)\\
&S(M+1)=S(M)(1-(\hat{a}_{M+1}(M+1))^2) ,
\end{aligned}$$

with the initial values

$$\begin{aligned}
&\hat{a}_0(m)=0 ,\\
&S(0)=C_{xx}(0) .
\end{aligned} \qquad (7.4)$$

Little difference has been observed between the results obtained by the original and the present versions of the procedures in many applications to artificial time series and the whole numerical results in the following section are obtained by using this practical version.

8. Numerical examples and discussions

Table 1 shows the results of applications of the three procedures, minimum FPE, minimum $(\mathrm{FPE})^{1/4}$ and an Anderson type described in Section 6, with $N=100$ and $L=10$, to ten artificial realizations of the process

$$X(n)=0.3X(n-1)+0.2X(n-2)+0.1X(n-3)+\varepsilon(n) ,$$

where $\varepsilon(n)$'s are mutually independently distributed uniformly over $[-\frac{1}{2}, \frac{1}{2}]$. It can be seen that all the three procedures are showing the tendency of giving M_0 lower than the true order, except the three extreme cases of the Anderson type.

Table 1. Frequency table of adopted order M_0 in ten applications of the three procedures to the process $X(n)=0.3X(n-1)+0.2X(n-2)+0.1X(n-3)+\varepsilon(n)$. $N=100$ and $L=10$.

Type of procedure \ Adopted order M_0	0	1	2	3	10
Anderson with $\rho=0.1$	2	4	1		3
Minimum $(\mathrm{FPE})^{1/4}$	2	3	5		
Minimum (FPE)		2	6	2	

The experiment has exposed the weakness of the Anderson type procedure that, in its present definition, it is not fully protected against adopting extraordinarily large values of M_0. The procedure is also with the difficulty in selecting the value of ρ.

The present results suggest that in spite of its inconsistency, discussed in Section 5, as an estimate of the order of a finite order autoregressive process, M_0 of the minimum FPE procedure will not be giving too large values in practical applications. This point is further backed up by the next example.

We have applied the three procedures with $N=100$ and $L=20$ to the process $X(n)=\varepsilon(n)-0.8\varepsilon(n-1)$, where $\varepsilon(n)$ is as in the former example. In this case $X(n)$ is actually an autoregressive process of infinite order. The orders which gave the estimates $\hat{a}_M$ with the minimum of the one-step prediction error variances in each experiment were identified by numerical computations and are given in the column denoted by "optimum" in Table 2, along with the orderes adopted by the three procedures. The table clearly shows the general tendency of the three procedures giving lower orders than optimum. The differences of the one-step prediction error variances of the optimum predictors and those obtained by the minimum FPE procedure were all relatively small and were at most of the order of 10% of the variance of $\varepsilon(n)$. This shows that for the minimum FPE procedure the present tendency of taking the lower values of orders is not so harmful for prediction.

Table 2. Orders adopted by the three procedures for $X(n)=\varepsilon(n)-0.8\varepsilon(n-1)$ in nine experiments and the corresponding orders which gave the predictors with the minimum one-step prediction error variance in each experiment. $N=100$ and $L=20$.

Type of procedure / Number of the experiment	Anderson $\rho=0.1$	Minimum $(\text{FPE})^{1/4}$	Minimum FPE	Optimum
1	3	3	5	7
2	2	2	6	8
3	3	3	3	6
4	3	3	3	6
5	12	3	3	9
6	1	2	3	6
7	2	2	2	7
8	2	3	4	8
9	1	7	8	6

The results of Tables 1 and 2 both show the wide variability of M_0 of the present Anderson type procedure. Also they show that the tendency of giving lower estimates of orders than optimum is weakest in the minimum FPE procedure. Furthermore, there is no arbitrariness in the

definition of the minimum FPE procedure, such as ρ in the Anderson's and α in the minimum $(\mathrm{FPE})^{\alpha}$, except the only one constant L which was common to all the three procedures. These observations suggest that the original minimum FPE procedure would be the most useful for practical applications.

To give a feeling of the behavior of (FPE)(M), one example is depicted in Fig. 1. The figure illustrates the behavior of (RFPE)(M)=(FPE)(M)((FPE)(0))$^{-1}$ for one realization of $X(n)=0.8X(n-1)+\varepsilon(n)$ with $N=100$ and $L=20$. $\varepsilon(n)$ was the same as in the former examples. In practical applications of the minimum FPE procedure to real data, (RFPE)(M) ($M=M_0, M_0+1, \cdots, L$) has shown a similar behavior to that of Fig. 1 for $M=1, 2, \cdots, L$.

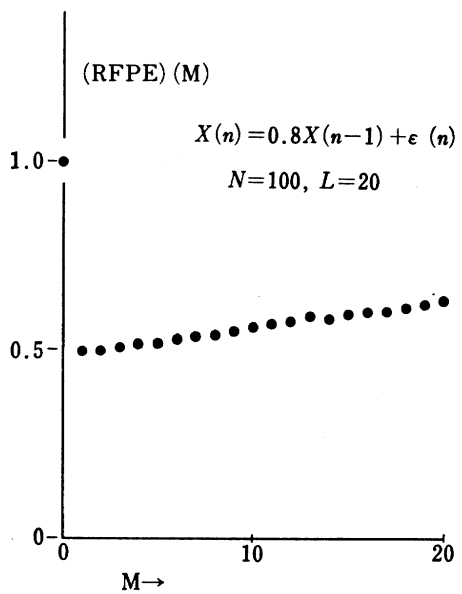

Fig. 1. Behavior of (RFPE)(M)=(FPE)(M)((FPE)(0))$^{-1}$ for a realization of $X(n)=0.8X(n-1)+\varepsilon(n)$.

From the experiences of application to real and artificial data it seems that to set L nearly equal to $0.1N$ would be a reasonable choice for ordinary size of N. Some numerical results of application of the minimum FPE procedure to real data are to be seen elsewhere [3, 6].

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Akaike, H. (1969). Fitting autoregressive models for prediction, *Ann. Inst. Statist. Math.*, **21**, 243-247.

[2] Anderson, T. W. (1963). Determination of the order of dependence in normally distributed time series, *Time Series Analysis* (ed. M. Rosenblatt), New York, John Wiley, 425-446.

[3] Akaike, H. (1970). On a semi-automatic power spectrum estimation procedure, *Proc. 3rd Hawaii International Conference on System Sciences*, 974-977.

[4] Diananda, P. H. (1953). Some probability limit theorems with statistical applications, *Proc. Cambridge Philos. Soc.*, **49**, 239-246.

[5] Anderson, T. W. and Walker, A. M. (1964). On the asymptotic distribution of the autocorrelations of a sample from a linear stochastic process, *Ann. Math. Statist.*, **35**, 1296-1303.

[6] Akaike, H. (1969). Power spectrum estimation through autoregressive model fitting, *Ann. Inst. Statist. Math.*, **21**, 407-419.

[7] Mann, H. B. and Wald, A. (1943). On stochastic limit and order relationships, *Ann. Math. Statist.*, **14**, 217-226.

[8] Durbin, J. (1960). The fitting of time-series models, *Rev. Int. Inst. Stat.*, **28**, 233-244.

[9] Jones, R. H. (1964). Prediction of multivariate time series, *J. of Applied Meteorology*, **3**, 285-289.