A New Look at the Statistical Model Identification

HIROTUGU AKAIKE, MEMBER, IEEE

Abstract—The history of the development of statistical hypothesis testing in time series analysis is reviewed briefly and it is pointed out that the hypothesis testing procedure is not adequately defined as the procedure for statistical model identification. The classical maximum likelihood estimation procedure is reviewed and a new estimate minimum information theoretical criterion (AIC) estimate (MAICE) which is designed for the purpose of statistical identification is introduced. When there are several competing models the MAICE is defined by the model and the maximum likelihood estimates of the parameters which give the minimum of AIC defined by

$$\text{AIC} = (-2)\log(\text{maximum likelihood}) + 2(\text{number of independently adjusted parameters within the model}).$$

MAICE provides a versatile procedure for statistical model identification which is free from the ambiguities inherent in the application of conventional hypothesis testing procedure. The practical utility of MAICE in time series analysis is demonstrated with some numerical examples.

I. Introduction

In spite of the recent development of the use of statistical concepts and models in almost every field of engineering and science it seems as if the difficulty of constructing an adequate model based on the information provided by a finite number of observations is not fully recognized. Undoubtedly the subject of statistical model construction or identification is heavily dependent on the results of theoretical analyses of the object under observation. Yet it must be realized that there is usually a big gap between the theoretical results and the practical procedures of identification. A typical example is the gap between the results of the theory of minimal realizations of a linear system and the identification of a Markovian representation of a stochastic process based on a record of finite duration. A minimal realization of a linear system is usually defined through the analysis of the rank or the dependence relation of the rows or columns of some Hankel matrix [1]. In a practical situation, even if the Hankel matrix is theoretically given, the rounding errors will always make the matrix of full rank. If the matrix is obtained from a record of observations of a real object the sampling variabilities of the elements of the matrix will be by far the greater than the rounding errors and also the system will always be infinite dimensional. Thus it can be seen that the subject of statistical identification is essentially concerned with the art of approximation which is a basic element of human intellectual activity.

As was noticed by Lehman [2, p. viii], hypothesis testing procedures are traditionally applied to the situations where actually multiple decision procedures are required. If the statistical identification procedure is considered as a decision procedure the very basic problem is the appropriate choice of the loss function. In the Neyman–Pearson theory of statistical hypothesis testing only the probabilities of rejecting and accepting the correct and incorrect hypotheses, respectively, are considered to define the loss caused by the decision. In practical situations the assumed null hypotheses are only approximations and they are almost always different from the reality. Thus the choice of the loss function in the test theory makes its practical application logically contradictory. The recognition of this point that the hypothesis testing procedure is not adequately formulated as a procedure of approximation is very important for the development of practically useful identification procedures.

A new perspective of the problem of identification is obtained by the analysis of the very practical and successful method of maximum likelihood. The fact that the maximum likelihood estimates are, under certain regularity conditions, asymptotically efficient shows that the likelihood function tends to be a quantity which is most sensitive to the small variations of the parameters around the true values. This observation suggests the use of

$$S(g;f(\cdot|\theta)) = \int g(x)\log f(x|\theta)\,dx$$

as a criterion of "fit" of a model with the probabilistic structure defined by the probability density function $f(x|\theta)$ to the structure defined by the density function $g(x)$. Contrary to the assumption of a single family of density $f(x|\theta)$ in the classical maximum likelihood estimation procedure, several alternative models or families defined by the densities with different forms and/or with one and the same form but with different restrictions on the parameter vector θ are contemplated in the usual situation of identification. A detailed analysis of the maximum likelihood estimate (MLE) leads naturally to a definition of a new estimate which is useful for this type of multiple model situation. The new estimate is called the minimum information theoretic criterion (AIC) estimate (MAICE), where AIC stands for an information theoretic criterion recently introduced by the present author [3] and is an estimate of a measure of fit of the model. MAICE is defined by the model and its parameter values which give the minimum of AIC. By the introduction of MAICE the problem of statistical identification is explicitly formulated as a problem of estimation and the need of the subjective judgement required in the hypothesis testing procedure for the decision on the levels of significance is completely eliminated. To give an explicit definition of MAICE and to discuss its characteristics by comparison with the conventional identification procedure based on estimation

Manuscript received February 12, 1974; revised March 2, 1974.
The author is with the Institute of Statistical Mathematics, Minato-ku, Tokyo, Japan.

and hypothesis testing form the main objectives of the present paper.

Although MAICE provides a versatile method of identification which can be used in every field of statistical model building, its practical utility in time series analysis is quite significant. Some numerical examples are given to show how MAICE can give objectively defined answers to the problems of time series analysis in contrast with the conventional approach by hypothesis testing which can only give subjective and often inconclusive answers.

II. Hypothesis Testing in Time Series Analysis

The study of the testing procedure of time series started with the investigation of the test of a simple hypothesis that a single serial correlation coefficient is equal to 0. The utility of this type of test is certainly too limited to make it a generally useful procedure for model identification. In 1947 Quenouille [4] introduced a test for the goodness of fit of autoregressive (AR) models. The idea of the Quenouille's test was extended by Wold [5] to a test of goodness of fit of moving average (MA) models. Several refinements and generalizations of these test procedures followed [6]-[9] but a most significant contribution to the subject of hypothesis testing in time series analysis was made by Whittle [10], [11] by a systematic application of the Neyman-Pearson likelihood ratio test procedure to the time series situation.

A very basic test of time series is the test of whiteness. In many situations of model identification the whiteness of the residual series after fitting a model is required as a proof of adequacy of the model and the test of whiteness is widely used in practical applications [12]-[15]. For the test of whiteness the analysis of the periodgram provides a general solution.

A good exposition of the classical hypothesis testing procedures including the tests based on the periodgrams is given in Hannan [16].

The fitting of AR or MA models is essentially a subject of multiple decision procedure rather than that of hypothesis testing. Anderson [17] discussed the determination of the order of a Gaussian AR process explicitly as a multiple decision procedure. The procedure takes a form of a sequence of tests of the models starting at the highest order and successively down to the lowest order. To apply the procedure to a real problem one has to specify the level of significance of the test for each order of the model. Although the procedure is designed to satisfy certain clearly defined condition of optimality, the essential difficulty of the problem of order determination remains as the difficulty in choosing the levels of significance. Also the loss function of the decision procedure is defined by the probability of making incorrect decisions and thus the procedure is not free from the logical contradiction that in practical applications the order of the true structure will always be infinite. This difficulty can only be avoided by reformulating the problem explicitly as a problem of approximation of the true structure by the model.

III. Direct Approach to Model Error Control

In the field of nontime series regression analysis Mallows introduced a statistic C_p for the selection of variables for regression [18]. C_p is defined by

$$C_p = (\hat{\sigma}^2)^{-1} \text{ (residual sum of squares) } - N + 2p,$$

where $\hat{\sigma}^2$ is a properly chosen estimate of σ^2, the unknown variance of the true residual, N is the number of observations, and p is the number of variables in regression. The expected value of C_p is roughly p if the fitted model is exact and greater otherwise. C_p is an estimate of the expected sum of squares of the prediction, scaled by σ^2, when the estimated regression coefficients are used for prediction and has a clearly defined meaning as a measure of adequacy of the adopted model. Defined with this clearly defined criterion of fit, C_p attracted serious attention of the people who were concerned with the regression analyses of practical data. See the references of [18]. Unfortunately some subjective judgement is required for the choice of $\hat{\sigma}^2$ in the definition of C_p.

At almost the same time when C_p was introduced, Davisson [19] analyzed the mean-square prediction error of stationary Gaussian process when the estimated coefficients of the predictor were used for prediction and discussed the mean-square error of an adaptive smoothing filter [20]. The observed time series x_i is the sum of signal s_i and additive white noise n_i. The filtered output $\hat{s}_i$ is given by

$$\hat{s}_i = \sum_{j=-M}^{L} \beta_j x_{i+j}, \qquad (i = 1,2,\cdots,N)$$

where β_j is determined from the sample x_i $(i = 1,2,\cdots,N)$. The problem is how to define L and M so that the mean-square smoothing error over the N samples $E[(1/N) \sum_{i=1}^{N} (s_i - \hat{s}_i)^2]$ is minimized. Under appropriate assumptions of s_i and n_i Davisson [20] arrived at an estimate of this error which is defined by

$$\hat{\sigma}_N{}^2[M,L] = s^2 + 2\hat{c}(M + L + 1)/N,$$

where s^2 is an estimate of the error variance and $\hat{c}$ is the slope of the curve of s^2 as a function of $(M + L)/N$ at "larger" values of $(L + M)/N$. This result is in close correspondence with Mallows' C_p, and suggests the importance of this type of statistics in the field of model identification for prediction. Like the choice of $\hat{\sigma}^2$ in Mallows' C_p the choice of $\hat{c}$ in the present statistic $\hat{\sigma}_N{}^2[M,L]$ becomes a difficult problem in practical application.

In 1969, without knowing the close relationship with the above two procedures, the present author introduced a fitting procedure of the univariate AR model defined by $y_i = a_1 y_{i-1} + \cdots + a_p y_{i-p} + x_i$, where x_i is a white noise [21]. In this procedure the mean-square error of the one-step-ahead prediction obtained by using the least squares estimates of the coefficients is controlled. The mean-square error is called the final prediction error (FPE) and when the data y_i $(i = 1,2,\cdots,N)$ are given its estimate is defined by

$$\mathrm{FPE}(p) = \{(N + p)/(N - p)\} \cdot (\hat{C}_0 - \hat{a}_{p1}\hat{C}_1 - \cdots - \hat{a}_{pp}\hat{C}_p),$$

where the mean of y_i is assumed to be 0, $\hat{C}_l = (1/N)\sum_{i=1}^{N-l} y_{i+l}y_i$ and $\hat{a}_{pi}$'s are obtained by solving the Yule–Walker equation defined by $\hat{C}_l$'s. By scanning p successively from 0 to some upper limit L the identified model is given by the p and the corresponding $\hat{a}_{pi}$'s which give the minimum of FPE(p) ($p = 0,1,\cdots,L$). In this procedure no subjective element is left in the definition of FPE(p). Only the determination of the upper limit L requires judgement. The characteristics of the procedure was further analyzed [22] and the procedure worked remarkably well with practical data [23], [24]. Gersch and Sharp [25] discussed their experience of the use of the procedure. Bhansali [26] reports very disappointing results, claiming that they were obtained by Akaike's method. Actually the disappointing results are due to his incorrect definition of the related statistic and have nothing to do with the present minimum FPE procedure. The procedure was extended to the case of multivariate AR model fitting [27]. A successful result of implementation of a computer control of cement kiln processes based on the results obtained by this identification procedure was reported by Otomo and others [28].

One common characteristic of the three procedures discussed in this section is that the analysis of the statistics has to be extended to the order of $1/N$ of the main term.

IV. Mean Log-Likelihood as a Measure of Fit

The well known fact that the MLE is, under regularity conditions, asymptotically efficient [29] shows that the likelihood function tends to be a most sensitive criterion of the deviation of the model parameters from the true values. Consider the situation where $x_1,x_2,\cdots,x_N$ are obtained as the results of N independent observations of a random variable with probability density function $g(x)$. If a parametric family of density function is given by $f(x|\theta)$ with a vector parameter θ, the average log-likelihood, or the log-likelihood divided by N, is given by

$$(1/N) \sum_{i=1}^{N} \log f(x_i|\theta), \tag{1}$$

where, as in the sequel of the present paper, log denotes the natural logarithms. As N is increased indefinitely, this average tends, with probability 1, to

$$S(g;f(\cdot|\theta)) = \int g(x) \log f(x|\theta)\, dx,$$

where the existence of the integral is assumed. From the efficiency of MLE it can be seen that the (average) mean log-likelihood $S(g;f(\cdot|\theta))$ must be a most sensitive criterion to the small deviation of $f(x|\theta)$ from $g(x)$. The difference

$$I(g;f(\cdot|\theta)) = S(g;g) - S(g;f(\cdot|\theta))$$

is known as the Kullback–Leibler mean information for discrimination between $g(x)$ and $f(x|\theta)$ and takes positive value, unless $f(x|\theta) = g(x)$ holds almost everywhere [30]. These observations show that $S(g;f(\cdot|\theta))$ will be a reasonable criterion for defining a best fitting model by its maximization or, from the analogy to the concept of entropy, by minimizing $-S(g;f(\cdot|\theta))$. It should be mentioned here that in 1950 this last quantity was adopted as a definition of information function by Bartlett [31]. One of the most important characteristics of $S(g;f(\cdot|\theta))$ is that its natural estimate, the average log-likelihood (1), can be obtained without the knowledge of $g(x)$. When only one family $f(x|\theta)$ is given, maximizing the estimate (1) of $S(g;f(\cdot|\theta))$ with respect to θ leads to the MLE $\hat{\theta}$.

In the case of statistical identification, usually several families of $f(x|\theta)$, with different forms of $f(x|\theta)$ and/or with one and the same form of $f(x|\theta)$ but with different restrictions on the parameter vector θ, are given and it is required to decide on the best choice of $f(x|\theta)$. The classical maximum likelihood principle can not provide useful solution to this type of problems. A solution can be obtained by incorporating the basic idea underlying the statistics discussed in the preceding section with the maximum likelihood principle.

Consider the situation where $g(x) = f(x|\theta_0)$. For this case $I(g;f(\cdot|\theta))$ and $S(g;f(\cdot|\theta))$ will simply be denoted by $I(\theta_0;\theta)$ and $S(\theta_0;\theta)$, respectively. When θ is sufficiently close to θ_0, $I(\theta_0;\theta)$ admits an approximation [30]

$$I(\theta_0;\theta_0 + \Delta\theta) = (\tfrac{1}{2})\|\Delta\theta\|_J^2,$$

where $\|\Delta\theta\|_J^2 = \Delta\theta' J \Delta\theta$ and J is the Fisher information matrix which is positive definite and defined by

$$J_{ij} = E\left\{\frac{\partial \log f(X|\theta)}{\partial\theta_i} \frac{\partial \log f(X|\theta)}{\partial\theta_j}\right\},$$

where J_{ij} denotes the (i,j)th element of J and θ_i the ith component of θ. Thus when the MLE $\hat{\theta}$ of θ_0 lies very close to θ_0 the deviation of the distribution defined by $f(x|\theta)$ from the true distribution $f(x|\theta_0)$ in terms of the variation of $S(g;f(\cdot|\theta))$ will be measured by $(\tfrac{1}{2})\|\theta - \theta_0\|_J^2$. Consider the situation where the variation of θ for maximizing the likelihood is restricted to a lower dimensional subspace Θ of θ which does not include θ_0. For the MLE $\hat{\theta}$ of θ_0 restricted in Θ, if $\boldsymbol{\theta}$ which is in Θ and gives the maximum of $S(\theta_0;\theta)$ is sufficiently close to θ_0, it can be shown that the distribution of $N\|\hat{\theta} - \boldsymbol{\theta}\|_J^2$ for sufficiently large N is approximated under certain regularity conditions by a chi-square distribution the degree of freedom equal to the dimension of the restricted parameter space. See, for example, [32]. Thus it holds that

$$E_\infty 2NI(\theta_0;\hat{\theta}) = N\|\boldsymbol{\theta} - \theta_0\|_J^2 + k, \tag{2}$$

where E_∞ denotes the mean of the approximate distribution and k is the dimension of Θ or the number of parameters independently adjusted for the maximization of the likelihood. Relation (2) is a generalization of the expected prediction error underlying the statistics discussed in the preceding section. When there are several models it will be natural to adopt the one which will give the minimum of $EI(\theta_0;\hat{\theta})$. For this purpose, considering the situation where these models have their $\mathbf{\theta}$'s very close to θ_0, it becomes necessary to develop some estimate of $N\|\mathbf{\theta} - \theta_0\|_J^2$ of (2). The relation (2) is based on the fact that the asymptotic distribution of $\sqrt{N}(\hat{\theta} - \mathbf{\theta})$ is approximated by a Gaussian distribution with mean zero and variance matrix J^{-1}. From this fact if

$$2\left(\sum_{i=1}^{N} \log f(x_i|\theta_0) - \sum_{i=1}^{N} \log f(x_i|\hat{\theta})\right) \qquad (3)$$

is used as an estimate of $N\|\mathbf{\theta} - \theta_0\|_J^2$ it needs a correction for the downward bias introduced by replacing $\mathbf{\theta}$ by $\hat{\theta}$. This correction is simply realized by adding k to (3). For the purpose of identification only the comparison of the values of the estimates of $EI(\theta_0;\hat{\theta})$ for various models is necessary and thus the common term in (3) which includes θ_0 can be discarded.

V. Definition of an Information Criterion

Based on the observations of the preceding section an information criterion AIC of θ is defined by

$$\text{AIC}(\hat{\theta}) = (-2) \log (\text{maximum likelihood}) + 2k,$$

where, as is defined before, k is the number of independently adjusted parameters to get $\hat{\theta}$. $(1/N)\text{AIC}(\hat{\theta})$ may be considered as an estimate of $-2ES(\theta_0;\hat{\theta})$. IC stands for information criterion and A is added so that similar statistics, BIC, DIC etc., may follow. When there are several specifications of $f(x|\theta)$ corresponding to several models, the MAICE is defined by the $f(x|\hat{\theta})$ which gives the minimum of $\text{AIC}(\hat{\theta})$. When there is only one unrestricted family of $f(x|\theta)$, the MAICE is defined by $f(x|\hat{\theta})$ with $\hat{\theta}$ identical to the classical MLE. It should be noticed that an arbitrary additive constant can be introduced into the definition of $\text{AIC}(\hat{\theta})$ when the comparison of the results for different sets of observations is not intended. The present definition of MAICE gives a mathematical formulation of the principle of parsimony in model building. When the maximum likelihood is identical for two models the MAICE is the one defined with the smaller number of parameters.

In time series analysis, even under the Gaussian assumption, the exact definition of likelihood is usually too complicated for practical use and some approximation is necessary. For the application of MAICE there is a subtle problem in defining the approximation to the likelihood function. This is due to the fact that for the definition of AIC the log-likelihoods must be defined consistently to the order of magnitude of 1. For the fitting of a stationary Gaussian process model a measure of the deviation of a model from a true structure can be defined as the limit of the average mean log-likelihood when the number of observations N is increased indefinitely. This quantity is identical to the mean log-likelihood of innovation defined by the fitted model. Thus a natural procedure for the fitting of a stationary zero-mean Gaussian process model to the sequence of observations $y_1,y_2,\cdots,y_N$ is to define a primitive stationary Gaussian model with the l-lag covariance matrices $R(l)$, which are defined by

$$\begin{aligned} R(l) &= (1/N) \sum_{n=1}^{N-l} y_{n+l}y_n', \qquad l = 0,1,2,\cdots,N-1 \\ &= 0, \qquad l = N,N+1,\cdots \end{aligned}$$

and fit a model by maximizing the mean log-likelihood of innovation or equivalently, if the elements of the covariance matrix of innovation are within the parameter set, by minimizing the log-determinant of the variance matrix of innovation, N times of which is to be used in place of the log-likelihood in the definition of AIC. The adoption of the divisor N in the definition of $R(l)$ is important to keep the sequence of the covariance matrices positive definite. The present procedure of fitting a Gaussian model through the primitive model is discussed in detail in [33]. It leads naturally to the concept of Gaussian estimate developed by Whittle [34]. When the asymptotic distribution of the normalized correlation coefficients of y_n is identical to that of a Gaussian process the asymptotic distribution of the statistics defined as functions of these coefficients will also be independent of the assumption of Gaussian process. This point and the asymptotic behavior of the related statistics which is required for the justification of the present definition of AIC is discussed in detail in the above paper by Whittle. For the fitting of a univariate Gaussian AR model the MAICE defined with the present definition of AIC is asymptotically identical to the estimate obtained by the minimum FPE procedure.

AIC and a primitive definition of MAICE were first introduced by the present author in 1971 [3]. Some early successful results of applications are reported in [3], [35], [36].

VI. Numerical Examples

Before going into the discussion of the characteristics of MAICE its practical utility is demonstrated in this section.

For the convenience of the readers who might wish to check the results by themselves Gaussian AR models were fitted to the data given in Anderson's book on time series analysis [37]. To the Wold's three series artificially generated by the second-order AR schemes models up to the 50th order were fitted. In two cases the MAICE's were the second-order models. In the case where the MAICE was the first-order model, the second-order coefficient of the generating equation had a very small absolute value compared with its sampling variability and the one-step-ahead prediction error variance was smaller for the MAICE than for the second-order model defined with the MLE's of the coefficients. To the classical series of Wolfer's sunspot numbers with $N = 176$ AR models up to the 35th order were fitted and the MAICE was the eighth-order model. AIC attained a local minimum at the second order. In the case of the series of Beveridge's wheat price index with $N = 370$ the MAICE among the AR model up to the 50th order was again of the eighth order. AIC attained a local minimum at the second order which was adopted by Sargan [38]. In the light of the discussions of these series by Anderson, the choice of eight-order models for these two series looks reasonable.

Two examples of application of the minimum FPE procedure, which produces estimates asymptotically equivalent to MAICE's, are reported in [3]. In the example taken from the book by Jenkins and Watts [39, section 5.4.3] the estimate was identical to the one chosen by the authors of the book after a careful analysis. In the case of the seiche record treated by Whittle [40] the minimum FPE procedure clearly suggested the need of a very high-order AR model. The difficulty of fitting AR models to this set of data was discussed by Whittle [41, p. 38].

The procedure was also applied to the series E and F given in the book by Box and Jenkins [12]. Second- or third-order AR model was suggested by the authors for the series E which is a part of the Wolfer's sunspot number series with $N = 100$. The MAICE among the AR models up to the 20th order was the second-order model. Among the AR models up to the 10th order fitted to the series F with $N = 70$ the MAICE was the second-order model, which agrees with the suggestion made by the authors of the book.

To test the ability of discriminating between AR and MA models ten series of y_n $(n = 1,\cdots,1600)$ were generated by the relation $y_n = x_n + 0.6x_{n-1} - 0.1x_{n-2}$, where x_n was generated from a physical noise source and was supposed to be a Gaussian white noise. AR models were fitted to the first N points of each series for $N = 50$, 100, 200, 400, 800, 1600. The sample averages of the MAICE AR order were 3.1, 4.1, 6.5, 6.8, 8.2, and 9.3 for the successively increasing values of N. An approximate MAICE procedure which is designed to get an initial estimate of MAICE for the fitting of Markovian models, described in [33], was applied to the data. With only a few exceptions the approximate MAICE's were of the second order. This corresponds to the AR–MA model with a second-order AR and a first-order MA. The second- and third-order MA models were then fitted to the data with $N = 1600$. Among the AR and MA models fitted to the data the second-order MA model was chosen nine times as the MAICE and the third-order MA was chosen once. The average difference of the minimum of AIC between AR and MA models was 7.7, which roughly means that the expected likelihood ratio of a pair of two fitted models will be about 47 for a set of data with $N = 1600$ in favor of MA model.

Another test was made with the example discussed by Gersch and Sharp [25]. Eight series of length $N = 800$ were generated by an AR–MA scheme described in the paper. The average of the MAICE AR orders was 17.9 which is in good agreement with the value reported by Gersch and Sharp. The approximate MAICE procedure was applied to determine the order or the dimension of the Markovian representation of the process. For the eight cases the procedure identically picked the correct order four. AR–MA models of various orders were fitted to one set of data and the corresponding values of AIC(p,q) were computed, where AIC(p,q) is the value of AIC for the model with AR order p and MA order q and was defined by

$$\text{AIC}(p,q) = N \log (\text{MLE of innovation variance}) + 2(p+q).$$

The results are AIC(3,2) = 192.72, AIC(4,3) = 66.54, AIC(4,4) = 67.44, AIC(5,3) = 67.48 AIC(6,3) = 67.65, and AIC(5,4) = 69.43. The minimum is attained at $p = 4$ and $q = 3$ which correspond to the true structure. Fig. 1 illustrates the estimates of the power spectral density obtained by applying various procedures to this set of data. It should be mentioned that in this example the Hessian of the mean log-likelihood function becomes singular at the true values of the parameters for the models with p and q simultaneously greater than 4 and 3, respectively. The detailed discussion of the difficulty connected with this singularity is beyond the scope of the present paper. Fig. 2 shows the results of application of the same type of procedure to a record of brain wave with $N = 1420$. In this case only one AR–MA model with AR order 4 and MA order 3 was fitted. The value of AIC of this model is 1145.6 and that of the MAICE AR model is 1120.9. This suggests that the 13th order MAICE AR model is a better choice, a conclusion which seems in good agreement with the impression obtained from the inspection of Fig. 2.

VII. Discussions

When $f(x|\boldsymbol{\theta})$ is very far from $g(x)$, $S(g;f(\cdot|\boldsymbol{\theta}))$ is only a subjective measure of deviation of $f(x|\boldsymbol{\theta})$ from $g(x)$. Thus the general discussion of the characteristics of MAICE will only be possible under the assumption that for at least one family $f(x|\boldsymbol{\theta})$ is sufficiently closed to $g(x)$ compared with the expected deviation of $f(x|\hat{\theta})$ from $f(x|\boldsymbol{\theta})$. The detailed analysis of the statistical characteristics of MAICE is only necessary when there are several families which satisfy this condition. As a single estimate of $-2NES(g;f(\cdot|\hat{\theta}))$, -2 times the log-maximum likelihood will be sufficient but for the present purpose of "estimating the difference" of $-2NES(g;f(\cdot|\hat{\theta}))$ the introduction of the term $+2k$ into the definition of AIC is crucial. The disappointing results reported by Bhansali [26] were due to his incorrect use of the statistic, equivalent to using $+k$ in place of $+2k$ in AIC.

When the models are specified by a successive increase of restrictions on the parameter θ of $f(x|\theta)$ the MAICE procedure takes a form of repeated applications of conventional log-likelihood ratio test of goodness of fit with automatically adjusted levels of significance defined by the terms $+2k$. When there are different families approximating the true likelihood equally well the situation will at least locally be approximated by the different parametrizations of one and the same family. For these cases the significance of the difference of AIC's between two models will be evaluated by comparing it with the variability of a chi-square variable with the degree of freedom

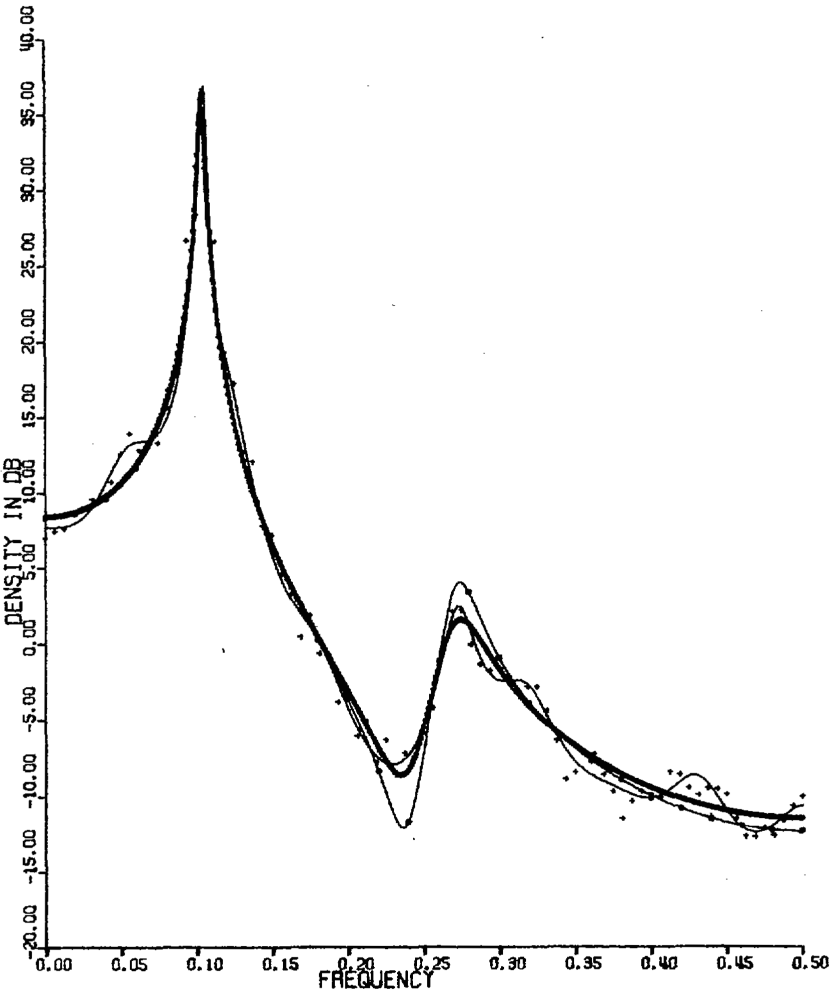

Fig. 1. Estimates of an AR–MA spectrum: theoretical spectrum (solid thin line with dots), AR–MA estimate (thick line), AR estimate (solid thin line), and Hanning windowed estimate with maximum lag 80 (crosses).

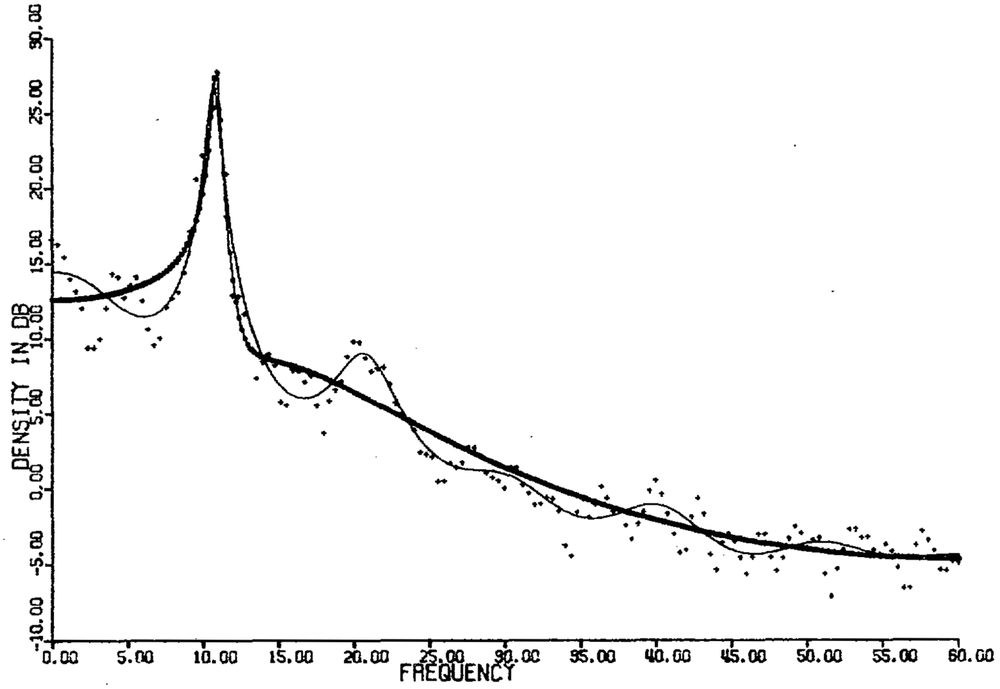

Fig. 2. Estimates of brain wave spectrum: AR–MA estimate (thick line), AR estimate (solid thin line), and Hanning windowed estimate with maximum lag 150 (crosses).

equal to the difference of the k's of the two models. When the two models form separate families in the sense of Cox [42], [43] the procedure developed by Cox and extended by Walker [44] to time series situation may be useful for the detailed evaluation of the difference of AIC.

It must be clearly recognized that MAICE can not be compared with a hypothesis testing procedure unless the latter is defined as a decision procedure with required levels of significance. The use of a fixed level of significance for the comparison of models with various number of parameters is wrong since this does not take into account the increase of the variability of the estimates when the number of parameters is increased. As will be seen by the work of Kennedy and Bancroft [45] the theory of model building based on a sequence of significance tests is not sufficiently developed to provide a practically useful procedure.

Although the present author has no proof of optimality of MAICE it is at present the only procedure applicable to every situation where the likelihood can be properly defined and it is actually producing very reasonable results without very much amount of help of subjective judgement. The successful results of numerical experiments suggest almost unlimited applicability of MAICE in the fields of modeling, prediction, signal detection, pattern recognition, and adaptation. Further improvements of definition and use of AIC and numerical comparisons of MAICE with other procedures in various specific applications will be the subjects of further study.

VIII. Conclusion

The practical utility of the hypothesis testing procedure as a method of statistical model building or identification must be considered quite limited. To develop useful procedures of identification more direct approach to the control of the error or loss caused by the use of the identified model is necessary. From the success of the classical maximum likelihood procedures the mean log-likelihood seems to be a natural choice as the criterion of fit of a statistical model. The MAICE procedure based on AIC which is an estimate of the mean log-likelihood provides a versatile procedure for the statistical model identification. It also provides a mathematical formulation of the principle of parsimony in the field of model construction. Since a procedure based on MAICE can be implemented without the aid of subjective judgement, the successful numerical results of applications suggest that the implementations of many statistical identification procedures for prediction, signal detection, pattern recognition, and adaptation will be made practical with MAICE.

Acknowledgment

The author is grateful to Prof. T. Kailath, Stanford University, for encouraging him to write the present paper. Thanks are also due to Prof. K. Sato, Nagasaki University, for providing the brain wave data treated in Section V.

References

[1] H. Akaike, "Stochastic theory of minimal realization," this issue, pp. 667–674.

[2] E. L. Lehman, *Testing Statistical Hypothesis.* New York: Wiley, 1959.

[3] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proc. 2nd Int. Symp. Information Theory, Supp. to Problems of Control and Information Theory,* 1972, pp. 267–281.

[4] M. H. Quenouille, "A large-sample test for the goodness of fit of autoregressive schemes," *J. Roy. Statist. Soc.*, vol. 110, pp. 123–129, 1947.

[5] H. Wold, "A large-sample test for moving averages," *J. Roy. Statist. Soc., B,* vol. 11, pp. 297–305, 1949.

[6] M. S. Barlett and P. H. Diananda, "Extensions of Quenouille's test for autoregressive scheme," *J. Roy. Statist. Soc., B,* vol. 12, pp. 108–115, 1950.

[7] M. S. Bartlett and D. V. Rajalakshman, "Goodness of fit test for simultaneous autoregressive series," *J. Roy. Statist. Soc., B,* vol. 15, pp. 107–124, 1953.

[8] A. M. Walker, "Note on a generalization of the large sample goodness of fit test for linear autoregressive schemes," *J. Roy. Statist. Soc., B,* vol. 12, pp. 102–107, 1950.

[9] ——, "The existence of Bartlett–Rajalakshman goodness of fit *G*-tests for multivariate autoregressive processes with finitely dependent residuals," *Proc. Cambridge Phil. Soc.*, vol. 54, pp. 225–232, 1957.

[10] P. Whittle, *Hypothesis Testing in Time-Series Analysis.* Uppsala, Sweden: Almqvist and Wiksell, 1951.

[11] ——, "Some recent contributions to the theory of stationary processes," *A Study in the Analysis of Stationary Time Series.* Uppsala, Sweden: Almqvist and Wiksell, 1954, appendix 2.

[12] G. E. P. Box and G. M. Jenkins, *Time Series, Forecasting and Control.* San Francisco, Calif.: Holden-Day, 1970.

[13] I. Gustavsson, "Comparison of different methods for identification of industrial processes," *Automatica*, vol. 8, pp. 127–142, 1972.

[14] R. K. Mehra, "On the identification of variances and adaptive Kalman filtering," *IEEE Trans. Automat. Contr.*, vol. AC-15, pp. 175–184, Apr. 1970.

[15] R. K. Mehra, "On-line identification of linear dynamic systems with applications to Kalman filtering," *IEEE Trans. Automat. Contr.*, vol. AC-16, pp. 12–21, Feb. 1971.

[16] E. J. Hannan, *Time Series Analysis.* London, England: Methuen, 1960.

[17] T. W. Anderson, "Determination of the order of dependence in normally distributed time series," in *Time Series Analysis*, M. Rosenblatt, Ed. New York: Wiley, 1963, pp. 425–446.

[18] C. L. Mallows, "Some comments on C_p," *Technometrics,* vol. 15, pp. 661–675, 1973.

[19] L. D. Davisson, "The prediction error of stationary Gaussian time series of unknown covariance," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 527–532, Oct. 1965.

[20] ——, "A theory of adaptive filtering," *IEEE Trans. Inform. Theory*, vol. IT-12, pp. 97–102, Apr. 1966.

[21] H. Akaike, "Fitting autoregressive models for prediction," *Ann. Inst. Statist. Math.*, vol. 21, pp. 243–247, 1969.

[22] ——, "Statistical predictor identification," *Ann. Inst. Statist. Math.*, vol. 22, pp. 203–217, 1970.

[23] ——, "On a semiautomatic power spectrum estimation procedure," in *Proc. 3rd Hawaii Int. Conf. System Sciences,* 1970, pp. 974–977.

[24] R. H. Jones, "Autoregressive spectrum estimation," in *3rd Conf. Probability and Statistics in Atmospheric Sciences, Preprints,* Boulder, Colo., June 19–22, 1973.

[25] W. Gersch and D. R. Sharpe, "Estimation of power spectra with finite-order autoregressive models," *IEEE Trans. Automat. Contr.*, vol. AC-18, pp. 367–379, Aug. 1973.

[26] R. J. Bhansali, "A Monte Carlo comparison of the regression method and the spectral methods of prediction," *J. Amer. Statist. Ass.*, vol. 68, pp. 621–625, 1973.

[27] H. Akaike, "Autoregressive model fitting for control," *Ann. Inst. Statist. Math.*, vol. 23, pp. 163–180, 1971.

[28] T. Otomo, T. Nakagawa, and H. Akaike, "Statistical approach to computer control of cement rotary kilns," *Automatica,* vol. 8, pp. 35–48, 1972.

[29] H. Cramer, *Mathematical Methods of Statistics.* Princeton, N. J.: Princeton Univ. Press, 1946.

[30] S. Kullback, *Information Theory and Statistics.* New York: Wiley, 1959.

[31] M. S. Bartlett, "The statistical approach to the analysis of time series," in *Proc. Symp. Information Theory*, London, England, Ministry of Supply, 1950, pp. 81–101.

[32] P. J. Huber, "The behavior of maximum likelihood estimates under nonstandard conditions," in *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability,* vol. 1, pp. 221–233, 1967.

[33] H. Akaike, "Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes," *Ann. Inst. Statist. Math.*, to be published.

[34] P. Whittle, "Gaussian estimation in stationary time series," *Bull. Int. Statist. Inst.*, vol. 39, pp. 105–129, 1962.

[35] H. Akaike, "Use of an information theoretic quantity for statistical model identification," in *Proc. 5th Hawaii Int. Conf. System Sciences*, pp. 249–250, 1972.

[36] H. Akaike, "Automatic data structure search by the maximum likelihood," in *Computer in Biomedicine Suppl. to Proc. 5th Hawaii Int. Conf. on System Sciences*, pp. 99–101, 1972.

[37] T. W. Anderson, *The Statistical Analysis of Time Series*. New York: Wiley, 1971.

[38] J. D. Sargan, "An approximate treatment of the properties of the correlogram and periodgram," *J. Roy. Statist. Soc. B*, vol. 15, pp. 140–152, 1953.

[39] G. M. Jenkins and D. G. Watts, *Spectral Analysis and its Applications*. San Francisco, Calif.: Holden-Day, 1968.

[40] P. Whittle, "The statistical analysis of a seiche record," *J. Marine Res.*, vol. 13, pp. 76–100, 1954.

[41] P. Whittle, *Prediction and Regulation*. London, England: English Univ. Press, 1963.

[42] D. R. Cox, "Tests of separate families of hypotheses," in *Proc. 4th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, 1961, pp. 105–123.

[43] D. R. Cox, "Further results on tests of separate families of hypotheses," *J. Roy. Statist. Soc., B*, vol. 24, pp. 406–425, 1962.

[44] A. M. Walker, "Some tests of separate families of hypotheses in time series analysis," *Biometrika*, vol. 54, pp. 39–68, 1967.

[45] W. J. Kennedy and T. A. Bancroft, "Model building for prediction in regression based upon repeated significance tests," *Ann. Math. Statist.*, vol. 42, pp. 1273–1284, 1971.

Hiortugu Akaike (M'72), for a photograph and biography see page 674 of this issue.

Some Recent Advances in Time Series Modeling

EMANUEL PARZEN

***Abstract*—The aim of this paper is to describe some of the important concepts and techniques which seem to help provide a solution of the stationary time series problem (prediction and model identification). Section I reviews models. Section II reviews prediction theory and develops criteria of closeness of a "fitted" model to a "true" model. The central role of the infinite autoregressive transfer function g_∞ is developed, and the time series modeling problem is defined to be the estimation of g_∞. Section III reviews estimation theory. Section IV describes autoregressive estimators of g_∞. It introduces a criterion for selecting the order of an autoregressive estimator which can be regarded as determining the order of an AR scheme when in fact the time series is generated by an AR scheme of finite order.**

I. Introduction

The aim of this paper is to describe some of the important concepts and techniques which seem to me to help provide realistic models for the processes generating observed time series.

Section II reviews the types of models (model conceptions) which statisticians have developed for time series analysis and indicates the value of signal plus noise decompositions as compared with simply an autoregressive-moving average (ARMA) representation.

Section III reviews prediction theory and develops criteria of closeness of a "fitted" model to a "true" model. The central role of the infinite autoregressive transfer function g_∞ is developed, and the time series modeling problem is defined to be the estimation of g_∞.

Section III reviews the estimation theory of autoregressive (AR) schemes and the basic role of Yule–Walker equations. It develops an analogous theory for moving average (MA) schemes, based on the duality between $f(\omega)$, the spectral density and inverse-spectral density, and $R(v)$ and $Ri(v)$, the covariance and covarinverse. The estimation of $Ri(v)$ is shown to be a consequence of the estimation of g_∞.

Section V describes autoregressive estimators of g_∞. It introduces a criterion for selecting the order of an autoregressive estimator which can be regarded as determining the order of an AR scheme when in fact the time series is generated by an AR scheme of finite order.

II. Time Series Models

Given observed data, statistics is concerned with inference from what *was* observed to what *might have been* observed. More precisely, one postulates a probability model for the process generating the data in which some parameters are unknown and are to be inferred from the data. Statistics is then concerned with parameter inference or parameter identification (determination of parameter values by estimation and hypothesis testing procedures).

A model for data is called *structural* if its parameters have a natural or structural interpretation; such models provide *explanation* and *control* of the process generating the data.

When no models are available for a data set from theory or experience, it is still possible to fit models which suffice for *simulation* (from what has been observed, generate more data similar to that observed), *prediction* (from what has been observed, forecast the data that will be observed), and *pattern recognition* (from what has been observed, infer

Manuscript received January 19, 1974; revised May 2, 1974. This work was supported in part by the Office of Naval Research.

The author is with the Division of Statistical Science, State University of New York, Buffalo, N.Y.