Automatica, Vol. 14, pp. 465–471

0005-1098/78/0901-0465 $02.00/0

Modeling By Shortest Data Description*

J. RISSANEN†

Estimates of both integer-valued structure parameters and real-valued system parameters may be obtained from a model based on the shortest data description principle.

Key Word Index—Modeling; parameter estimation; identification; statistics; stochastic systems.

Summary—The number of digits it takes to write down an observed sequence $x_1,\ldots,x_N$ of a time series depends on the model with its parameters that one assumes to have generated the observed data. Accordingly, by finding the model which minimizes the description length one obtains estimates of both the integer-valued structure parameters and the real-valued system parameters.

1. INTRODUCTION

THIS study is an attempt to derive a criterion for estimation of both the integer-valued structure parameters and the real-valued parameters of dynamic systems starting from a single natural and fundamental principle: the least number of digits it takes to write down an observed sample of a time series

$$x_1, x_2, \ldots, x_N.$$

It is known that the usual criteria resulting from such, should we dare to say, *ad hoc* principles as the least squares, the minimum prediction error, and the maximum likelihood principles, cannot be used to estimate the order nor other structure parameters. This very fact seems to suggest that these time-honored principles are successful only insofar as they happen to coincide with one or another facet of some more powerful and comprehensive principle.

The shortest description length of individual recursively definable objects was studied by Kolmogorov[1] and others, and it gives rise to the algorithmic notion of entropy. In statistical estimation, however, the data always includes a random element, and the original notion of entropy of random variables turns out to be the appropriate one to use. In a peripheral way we also invoke a simple combinatorial entropy notion which, however, need not cause any confusion because basically all the different entropies measure one or another aspect of description length, or to put it more bluntly, they all amount to the logarithm of the cardinality of a finite set or its limit. To read this paper no background in information theory is needed.

In Section 2 we argue by Gibbs' theorem that it is not possible to write down the given sequence of the observed data $x_1,\ldots,x_N$ by use of any conceivable coding with as few bits, i.e. binary digits, in the mean as when the statistical mechanism that generates the data is known. This theorem together with its interpretations given in that section provides then the basis for the studied estimation principle.

Perhaps the first attempt to arrive at a criterion which includes a structure dependent term was made by Akaike[2], who proposed the following criterion [MAICE]:

$$N \log r + 2k, \tag{1.1}$$

where N log r is the minimized log-likelihood function (with opposite sign and within a constant) and k the number of parameters in the model. In [3] the criterion (1.1) was derived from the broader principle of maximizing $E_x B(f(\cdot), g(\cdot))$, where $B(f,g)$ is the Kullback-information with opposite sign:

$$B(f,g) = -\int\left[f(z)\log\frac{f(z)}{g(z;x)}\right]dz.$$

Here $f(\cdot)$ is the unknown "true" density function and $g(\cdot\,;x)$ the one given by the model in the light of the observed data x. By taking $\log g(x;x)$ as the estimate of its mean and removing the bias (1.1) results.

This maximum entropy principle, as Akaike calls it, is in our opinion a valuable generalization of the maximum likelihood principle. However, while we have no objection to using Kullback's information as a measure of the distance between the two distributions we remain doubtful of the origin; namely, the bias, of the

*Received July 19, 1976; revised November 10, 1977; revised March 9, 1978. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by associate editor K. J. Åström.

†IBM Research Laboratory, San Jose, California 95193, U.S.A.

crucial structure dependent term in (1.1) which to us seems coincidental. Accordingly, we feel justified to continue our search for a more fundamental and natural estimation principle. As we shall see there is a quite different cause for the structure dependent term resulting naturally from our adherence to the shortest description principle.

We mention other related works. In [4] Schwarz proposed a variant to Akaike's criterion which is closer to ours than (1.1). Again, there appears to be a degree of arbitrariness in the selection of the premises leading to the result, although the striking similarity of the added term to one in our criterion may suggest another interpretation of the shortest description length principle. In [5] and [6] yet another principle was studied which has provably good properties in certain fairly general cases. But because we used entropies of real-valued random variables, which in reality are no entropies at all, that criterion fails utterly in estimation of the order of an autoregressive model. Although the major defects of this criterion can be fairly easily corrected, an interpretation of the result as a shorted description length seems less natural than in the currently studied case.

2. BASICS OF ESTIMATION

All estimation is based on the idea of comparing the observed data with the result calculated by a proposed model for the data. Here the term "model" is used for any hypothesis that one uses for the purpose of trying to explain or describe the hidden laws that are supposed to govern or constrain the data. Insofar as the estimation is a statistical one an appropriate measure of the fit between the model and the data is provided by Gibbs' theorem[7], whose corollaries are the famous results known as Shannon's noiseless coding theorem[8], and Kullback's information measure[9] as restricted to discrete distributions.

Theorem 1. (Gibbs) Let p_i and q_i for $i=1,\ldots,n$ be non-negative real numbers such that

$$\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i .$$

Then,

$$-\sum_{i=1}^{n} p_i \log p_i \leqq -\sum_{i=1}^{n} p_i \log q_i, \tag{2.1}$$

or equivalently,

$$\sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \geqq 0, \tag{2.2}$$

where the equalities hold iff $p_i = q_i$ for all i. (Here, $0 \log 0$ is taken as 0.)

Proof. The claims follow by minimization of the right hand side of (2.1) over the q_i's subject to the given constraints with the Lagrangian technique.

Remarks. We need this theorem extended to countably many of the p_i's. The same proof is applicable provided that $\Sigma p_i < \infty$. In case $\{p_i\}$ is a probability distribution (2.2) provides a measure of fit between the two distributions, Kullback[9].

In order to get a quick idea of the role of this theorem in estimation let the data x_i be generated by an independent source with an unknown-symbol probability $p_i = P(x_i)$. If a model prescribes the probability q_i for x_i then by (2.1) the true distribution $\{p_i\}$ would come out from minimization of

$$-\sum_i p_i \log q_i . \tag{2.3}$$

Because the expectation would have to be approximated from the data the estimation will involve statistical errors, one of which is the bias, as discussed and utilized by Akaike in his criterion (1.1).

We next describe an interpretation of the basic inequality (2.1) as a size inequality, which provides the basis for the description length principle in estimation. We first describe briefly the standard communication theoretic interpretation, which is useful and simple, but which imposes a restriction on the way the encoding of the data is done.

Suppose again the preceding independent source situation and think of an encoder who uses the model $\{q_i\}$ to assign binary codewords for the data strings $x = x_1 x_2 \ldots x_N$, $x_i \in \{1,\ldots,n\}$. One way to do this is to assign about $-\log_2 q_i$ bits to the symbol x_i, and then build the code of x by concatenating these symbol codes $\bar{x}_i$ thus: $\bar{x} = \bar{x}_1 \bar{x}_2 \ldots \bar{x}_N$. The fact that $-\log_2 q_i$ may not be an integer is not a serious deterrent, for by a new arithmetic coding the code of $\bar{x}$ may be constructed as a sum of the suitably shifted terms $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_N$ so as to permit overlapping, where each codeword $\bar{x}_i$ is now interpreted as a number, [10].

If k_i is the number of times i occurs in x, then the per symbol length of the code $\bar{x}$ is seen to approach (2.3), and by (2.1) the minimum of this length cannot be achieved unless the model coincides with the true source or system. The situation is no different if more generally $\bar{x}_i$ is allotted l_i rather than about $-\log q_i$ bits, because the decoding is possible only if, [10],

$$\sum_i 2^{-l_i} \leqq 1 ,$$

and the equality holds for the best choice of the l_i so that 2^{-l_i} defines a probability q_i. Hence (2.1) applies again.

But the model may be used to describe the strings x by a more devious scheme. We show, albeit a bit sketchily, that the same conclusion still holds. If x includes k_i occurrences of i there are

$$S=\frac{N!}{k_1!\ldots k_n!}$$

distinct strings resulting from all possible permutations of the symbols. Now, no matter how the model is used to assign binary codewords for these strings x in one-to-one fashion, at least one codeword must have $\log_2 S$ bits or more, about half of the codewords must have at least $(\log_2 S)-1$ bits, and so on. Therefore, the maximum as well as the average number of bits in the code strings is about $\log_2 S$, or more. If the reader is unfamiliar with these facts he will quickly grasp them by writing down the first 16 integers in binary and counting the number of integers with length 4, 3, etc.

By Stirling's formula,

$$\frac{1}{N}\log_2 S=H\left(\frac{k_1}{N},\ldots,\frac{k_n}{N}\right)+O\left(\frac{\log N}{N}\right),$$

where $H(z_1,\ldots,z_n)=-\sum_i z_i\log_2 z_i$ is the Boltzman-entropy function. If the source is ergodic, too, then $k_i/N\to p_i$ and we get the limiting minimum average per symbol length:

$$L=\lim_N \frac{1}{N}\log_2 S=H(p_1,\ldots,p_n). \qquad (2.4)$$

Now, the encoder must also assign a codeword to every conceivable string x, including the S strings considered above. If he assigns the first S binary numbers to these strings then the average per symbol length will approach the limit (2.4). But he cannot possibly know which S strings out of the n^N possible ones are to receive the first S binary numbers unless he knows the numbers k_i, i.e., in the limit the probabilities p_i. If he picks those S strings that are derived from the model probabilities q_i; namely, $k_i'=q_iN$, then he will waste some of the short first S binary numbers for "wrong" strings; i.e., strings that are not among the ones with the data k_i, and he will have to spend longer codes for some of the "right" strings with an increased mean per symbol length as the result. So, again, we arrive at the conclusion that it is impossible to describe strings in the minimal way unless the model agrees with the true probabilities.

We have taken some space to interpret the basic inequality (2.1) as a limiting mean per symbol length inequality. The main reason for this is that it frees us from the task of interpreting some objects as random variables, which may be awkward, while we can always contemplate a description length of any object of whatever nature. For example, multi-variable linear systems have non-redundantly parameterized canonical forms, which the minimum description length principle automatically forces one to select, without our having to worry about random variable interpretations so that (2.1) could be invoked to reject the redundantly parameterized models. Indeed, if there are redundant parameters then their description costs bits without reducing the right hand side of (2.1).

3. DYNAMIC MODELS

Modeling of time series,

$$x=x_1x_2\ldots x_N$$

where the data indicates a dependency between distinct pairs x_{j+k} and x_j, is particularly important in many applications. On physical grounds there is often reason to expect that the joint distribution $(P(x_1,x_2,\ldots,)$ admits a parameterization with a moderate size parameter vector α, which leads to modeling of the process x by a dynamic system of the general type:

$$x_t=f_\alpha(x_{t-1},\ldots,x_{t-n},u_{t-1},\ldots,u_{t-k},\ e_t,\ldots,e_{t-m}), \qquad (3.1)$$

where the $\{u_t\}$ process is an observed input process, and $\{e_t\}$ is a non-observed "noise" process. The input process adds nothing of interest to our discussion and we drop it, which leaves us with the model

$$x_t=f_\alpha(x_{t-1},\ldots,x_{t-n},e_t,\ldots,e_{t-m}). \qquad (3.2)$$

As initial conditions we put $x_i=e_i=0$ for $i<1$. We also require the function f_α to be invertible in the sense that e_t can be solved from (3.2) in terms of $x_t,\ldots,x_{t-n}$ and $e_{t-1},\ldots,e_{t-m}$. Although all our concrete results will be derived for linear systems, the general approach outlined in this section is not limited to them.

To describe the process x amounts now to the description of the sequence e and the function f_α. To describe e, a statistical model is needed, i.e., a distribution of e. This then raises the question of why bother to introduce the sequence e at all, for

we could evidently describe x directly relative to an assumed distribution, which moreover, could be parameterized. Indeed, such approaches have been studied, e.g. by Whittle and by Walker; for an exposition of their work, see Hannan[11]. The reason for studying a model of type (3.2) is that we hope to be able to get a good model with an independent e-process, so that the joint distribution $P(e)$ can be described as the product of the marginal distributions $P(e_t)$. Then the description of the joint distribution is the shortest possible, for, indeed, as the joint distribution determines all of its marginal distributions but the converse is true only with independence, the description length of any joint distribution cannot be shorter than that of its marginal distributions.

These arguments justify the use of models of type (3.2) at least in the case where the e-process can be made independent by some model in this family.

If we parameterize the marginal distributions $P_\beta(e_t)$ by β then the whole model is parameterized by a vector $\theta=(s,\xi)$, where s denotes integer-valued structure parameters such as n and m and possibly others, and ξ consists of the real-valued parameters α and β. With two more conventions we now can write the general form for the length description of x. The real-valued components of the q-component vectors e_t must clearly be written with some precision $\pm\varepsilon/2$; e.g. $\pm 10^{-3}/2$, i.e., with two decimal digits. Similarly, the real-valued parameters ξ_i, the components of ξ, are written with precision $\pm\delta_i/2$. This allows for adjusting the precision of each component according to its contribution to the criterion as will be explained in the next section.

Then the per symbol length of the data $x=x_1,\ldots,x_N$, is of the form:

$$\frac{1}{N}\mathrm{L}(x,\theta)=\frac{1}{N}L(e/\theta)+\frac{1}{N}L(\theta), \tag{3.3}$$

where we take

$$L(e/\theta)=\sum_{t=1}^{N}\log \varepsilon^q p_\beta(e_t)$$

$$L(\theta)=\log(n+1)(m+1)+\sum_{i=1}^{k}\log\frac{|\xi_i|}{\delta_i}, \tag{3.4}$$

p_β denoting the density function of the distribution P_β. The expression for $L(\theta)$ is not entirely accurate; e.g., the signs of certain components should be included, but it is accurate enough to give the features we need. if $|\xi_i|/\delta_i<1$ we drop that term in the sum, although at least one bit would be needed to indicate ξ_i to be as zero. Finally, q and k denote the number of components in e_t and ξ, respectively.

The second term in (3.3) corresponds in a Bayesian interpretation to an assumed *a priori* distribution for the parameters. This distribution, instead of being assumed as uniform is one which for the integer j assigns the probability $1/j$. Other distributions could be contemplated. For instance, because asymptotically efficient estimators in general have a near Gaussian distribution this could be taken as the distribution for the ξ_i as in [5] and [6]. To make the description length interpretation meaningful, however, we would have to assume that the observed data points come in batches of, say, N pieces in each and we calculate the length $L(\theta^k/\theta^{k-1})$ of the parameter vector θ^k for the k'th batch given the previous parameter θ^{k-1} or, alternatively, all the previous parameters. We do not describe this process in detail because we study here the simpler formula (3.4).

It is worth elaborating a bit further the crucial second term in (3.3), which is precisely the difference between the length measure in (3.3) and the maximum likelihood criterion; i.e., the first term. Without any cost in writing down the parameters we could, in principle, bring the first term as near to the entropy in the left hand side of (2.1) as we like by increasing the complexity of the model and the number of parameters in it. This, basically, is why the correct structure cannot be estimated consistently by the maximum likelihood criterion. In the traditional applications of this criterion the number of parameters and their nature: i.e., the "structure", have been agreed upon before the data is being collected. Clearly, then, the minimization of (2.1) is a constrained one, and no problems of over parameterizations arise. In Akaike's criterion (1.1) the structure dependent term resulted from the observation that the bias caused by the asymptotically χ^2-distributed estimates is twice the number of the parameters, and, hence, it could serve as the desired penalty term. As stated in the introduction the bothersome thing about this is the coincidental appearance of the number of parameters in the bias and what seems to be needed to penalize the over parameterization. After all, why should the penalty term be a linear function of the number of the parameters? Certainly not because the bias happens to be such a function of them.

In (3.4) we have not included the lengths of the algorithms needed to calculate the x_t's from the coded values of e_t and to describe the codes themselves. This is because these algorithms do not depend on the data of interest: the observations and the parameters. Their lengths, there-

fore, represent a fixed overhead type of cost and may well be omitted in the present discussion.

Our plan is to work out the length formula (3.4) in the important case of Gaussian distributions and determine the optimum truncation levels δ_i. But before doing that we would like to interpret the criterion (3.4) as one measuring the degree of independence of the process $\{e_t\}$, and hence the minimizing model of x is the one with which e is "most independent." Insofar as the usually dominant part $(1/N)L(e/\theta)$ in (3.3) is in many cases equivalent to the several popular criteria such as the maximum likelihood, least squares, and the prediction error, criteria, [13], it is certainly of interest to support the latent belief that it is the degree of achieved independence that ultimately counts and unifies all the successful criteria. The difficulty in case a complete independence is not achievable, is, of course, which measure of independence one should use. The answer provided here presently is that it is the Kullback-measure, which itself is based on the Gibbs' measure (2.1), and of which the shortest length criterion is a computable and practicable reflection.

For the following discussion we consider scalar processes. The arguments are clearly valid for vector processes as well. For the collection of random variables $e_1,\ldots,e_N$, the numbers truncated to $\pm\varepsilon/2$, we may apply Kullback's information (2.2) to measure the deviation of the joint distribution $\varepsilon^N p(e_1,\ldots,e_N)$ from the product of the marginal ones $\varepsilon^N \prod_{i=1}^N p(e_i)$. The result is given in terms of the entropies of the truncated variables as:

$$R_\varepsilon(e_1,\ldots,e_N)=\sum_{e_1,\ldots,e_N} \varepsilon^N p(e_1,\ldots,e_N)\log\frac{p(e_1,\ldots,e_N)}{\prod_{i=1}^N p(e_i)}$$

$$=-\sum_{e_1}(\log \varepsilon p(e_1))\sum_{e_2,\ldots,e_N}\varepsilon^N p(e_1,\ldots,e_N)$$

$$-\sum_{e_2}(\log \varepsilon p(e_2))\sum_{e_1,e_3,\ldots,e_N}\varepsilon^N p(e_1,\ldots,e_N)$$

$$-\ldots-\sum_{e_1,\ldots,e_N}\varepsilon^N p(e_1,\ldots,e_N)\log \varepsilon^N p(e_1,\ldots,e_N)$$

$$=\sum_{i=1}^N H_\varepsilon(e_i)-H_\varepsilon(e_1,\ldots,e_N). \tag{3.5}$$

Now, without truncations the function f_α is one-to-one. Hence, when x_i is truncated to $\pm\mu/2$ and μ determined so that the ratio of the entropies of the truncated variables

$$\frac{H_\varepsilon(e_1,\ldots,e_N)}{H_\mu(x_1,\ldots,x_N)}=\rho_\varepsilon \tag{3.6}$$

is as near to one as possible, this ratio will go to one as ε goes to zero. From (3.5), then

$$\frac{R_\varepsilon(e_1,\ldots,e_N)}{H_\mu(x_1,\ldots,x_N)}=\frac{\sum_{i=1}^N H_\varepsilon(e_i)}{H_\mu(x_1,\ldots,x_N)}-\rho_\varepsilon.$$

Because $H_\mu(x_1,\ldots,x_N)$ is not dependent on the model we conclude that for small ε, $R_\varepsilon(e_1,\ldots,e_N)$ is minimized with good approximation by the same parameter values as the sum

$$\sum_{i=1}^N H_\varepsilon(e_i).$$

This sum, when divided by N, approximates almost surely as $N\to\infty$ the first and the dominant term in the length expression (3.3), and therefore the minimizing parameter of the length (3.3) also minimizes in the limit $N=\infty$ the measure of independence $(1/N)R_\varepsilon(e_1,\ldots,e_N)$, which is what we wanted to show by this digression.

4. LINEAR MODELS

In this section we work out in detail an expression for the description length of the observation sequence when linear models and Gaussian distributions are used. To facilitate the discussion we let the observed process $\{x_t\}$ have just one component, $q=1$. The general case is handled similarly, although the structure to be allowed for the models is more complex as discussed e.g. in [6].

We then consider models of the type:

$$x_t+a_1x_{t-1}+\ldots+a_nx_{t-n}=e_t+b_1e_{t-1}+\ldots+b_me_{t-m}$$
$$x_t=e_t=0 \text{ for } t\leqq 0. \tag{4.1}$$

The parameter vector consists of the data:

$$\theta=(n,m,\xi),\ \xi=(r,a_1,\ldots,a_n,b_1,\ldots,b_m) \tag{4.2}$$

where $r=\xi_0$ is the variance-parameter for the zero-mean normal distribution modeled for e_t.

The numbers e_t are written with precision $\pm\varepsilon/2$, and the parameter ξ_i, the i'th component of ξ, is written with precision $\pm\delta_i/2$. As it turns out we must use (4.1) to express e_t in terms of the observed data x_i, which themselves are truncations. If the purpose is to recover the truncated observations with minimum error from the coded

string of the numbers e_t, then μ and ε should be chosen so that the ratio in (3.6) is as near to one as possible. However, our purpose is parameter estimation, and the choice of the precision ε is less critical; it merely adds a constant to the criterion to be derived.

If h_r denotes the Gaussian distribution modeled for e_t, then by an assignment of a codeword of length $-\ln h_r(e_t)$ for e_t the sequence $e=(e_1,\ldots,e_N)$ can be described with about

$$L(e\,|\,r)=\frac{N}{2}\ln 2\pi\frac{r}{\varepsilon^2}+\frac{1}{2}\sum_{i=1}^{N}e_i^2/r \tag{4.3}$$

"nats"; i.e., natural logarithmic units. From (3.4) and (4.3) we then obtain the total length:

$$L(x,\theta)=\frac{N}{2}\ln 2\pi\frac{r}{\varepsilon^2}+\frac{1}{2}\sum_{i=1}^{N}e_i^2/r + \sum_{i=0}^{m+n}\ln\frac{|\xi_i|}{\delta_i}+\ln(n+1)(m+1). \tag{4.4}$$

By Gibbs' theorem we ought to minimize the mean of $L(x,\theta)$, which, naturally, must be suitably estimated from the data. Whatever the estimate is, the best choice for r is the one which minimizes $L(x,\theta)$; or, the solution to:

$$\frac{N}{2r}+\frac{1}{r}-\frac{1}{2r^2}\sum_{i=1}^{N}e_i^2=0,$$

where, to remind the reader, the second term follows from the convention $\xi_0=r$.
This gives

$$\hat{r}=\frac{1}{N+2}\sum_{i=1}^{N}e_i^2, \tag{4.5}$$

with the corresponding minimum:

$$\hat{L}(x,\theta)=\frac{N+2}{2}(1+\ln\hat{r})+\sum_{i=1}^{n+m}\ln\frac{|\xi_i|}{\delta_i}+\frac{N}{2}\ln\frac{2\pi}{\varepsilon^2} +\ln\frac{1}{\delta_0}+\ln(n+1)(m+1)-\frac{1}{2}. \tag{4.6}$$

For simplicity we take (4.6) as an estimate of its mean, and the other parameters $\xi_i, i=1,\ldots n+m$, will then be determined so as to minimize $\hat{L}(x,\theta)$, where e_i is to satisfy (4.1). Let $\hat{\xi}_i$ be the minimizing parameters for a given pair n and m. We write $L(x,\theta)$ as $L(x,n,m,\xi)$.

Following a suggestion by Boulton and Wallace (private correspondence) we next determine the optimum truncation levels δ_i. Boulton and Wallace have done an analogous minimization in the context of classification problems which they studied by a description length criterion of the same general type as that of ours[12].

Let $\tilde{\xi}_i$ denote the number resulting when $\hat{\xi}_i$ is truncated to the level $\pm\delta_i/2$. Then the difference $\Delta_i=\xi_i-\tilde{\xi}_i$ lies in the interval $[\hat{\xi}_i-\delta_i/2,\hat{\xi}_i+\delta_i/2]$. It seems reasonable to assume the errors Δ_i to be uniformly distributed within their range.

Expand $L(x,n,m,\xi)$ in Taylor's series about $\hat{\xi}$:

$$L(x,n,m,\xi)=L(x,n,m,\hat{\xi})+\frac{1}{2}\sum_{i,j=0}^{n+m}\times\frac{\partial^2 L}{\partial\hat{\xi}_i\partial\hat{\xi}_j}\Delta_i\Delta_j+O(\Delta^3)$$

where the notation for the double derivatives means that they be evaluated at $\hat{\xi}$, and $\Delta=\max_i(\Delta_i)$. Within an error of order $O(\Delta/N)$ we may replace L by the first two terms in (4.4) when calculating the double derivatives:

$$\frac{\partial^2 L}{\partial\hat{\xi}_0^2}=\frac{N+2}{2\hat{r}^2}+O\left(\frac{\Delta}{N}\right),$$

$$\frac{\partial^2 L}{\partial\hat{\xi}_i\partial\hat{\xi}_j}=\frac{N+2}{2}\cdot\frac{\partial^2\ln\hat{r}}{\partial\hat{\xi}_i\partial\hat{\xi}_j}+O\left(\frac{\Delta}{N}\right),\ i,j=1,\ldots,n+m.$$

$$\frac{\partial^2 L}{\partial\hat{\xi}_0\partial\hat{\xi}_i}=0,\ i=1,\ldots,n+m.$$

Then,

$$L(x,n,m,\xi)=L(x,n,m,\hat{\xi})+\frac{N+2}{4}\times\left[\frac{\Delta_0^2}{\hat{r}^2}+\sum_{i,j=1}^{n+m}\frac{\partial^2\ln\hat{r}}{\partial\hat{\xi}_i\partial\hat{\xi}_j}\Delta_i\Delta_j\right]+O\left(\frac{\Delta}{N}\right)+O(\Delta^3). \tag{4.7}$$

The expected value of $L(x,n,m,\xi)$ with respect to the uniform distribution for Δ_i is:

$$EL(x,n,m,\xi)=L(x,n,m,\hat{\xi})+\frac{N+2}{48}\times\left[\frac{\delta_0^2}{\hat{r}^2}+\sum_{i=1}^{n+m}\frac{\partial^2\ln\hat{r}}{\partial\hat{\xi}_i^2}\cdot\delta_i^2\right]. \tag{4.8}$$

This is minimized for

$$\hat{\delta}_0=\hat{r}\sqrt{24/(N+2)}$$

$$\hat{\delta}_i=\sqrt{24\Big/\left((N+2)\frac{\partial^2\ln\hat{r}}{\partial\hat{\xi}_i^2}\right)},\ i=1,\ldots,n+m. \tag{4.9}$$

The corresponding minimum gives our final criterion:

$$U(x,n,m,\hat{\xi}) = N \ln \hat{r} + \sum_{i=1}^{n+m} \ln\left(\hat{\xi}_i^2 \cdot \frac{\partial^2 \ln \hat{r}}{\partial \hat{\xi}_i^2}\right) + (n+m+1)\ln(N+2) + 2\ln(n+1)(m+1), \tag{4.10}$$

which is to be minimized with respect to the structure parameters n and m. Evidently, we may also determine all the parameters by minimization of $U(x,n,m,\xi)$. When deriving this expression we used the length term in (3.4) of type $\ln|\xi_i|/\delta_i$ which also appears in (4.6). This, clearly, is reasonable provided that $|\xi_i|/\delta_i > 1$. However, because the value of $\partial^2 \ln \hat{r}/\hat{\xi}_i^2$ is independent of the value of $|\xi_i|$ (4.9) may give the ratio $|\xi_i|/\delta_i$ a value less than 1 for some i. If this happens $\hat{\xi}_i$ is to be written as 0, and the corresponding term in the sum in (4.10) is to be dropped. Moreover, the number of parameters in the third term is to be reduced by one.

We notice that the minimum description length principle has given the three last terms in (4.10) as the penalty term for using the chosen model with its parameters. Of these the last is unimportant and can be dropped. The first and the third terms together virtually coincide with a criterion derived by Schwarz, [4], starting from an asymptotic expansion of the *a posteriori* probability in Bayes' approach. The main difference, in addition to the approach, between this and Schwarz' criterion is, then, the second term and the underlying convention that each parameter here must be written with a precision (4.9) which depends on the sensitivity of the criterion to that parameter.

CONCLUSIONS

A criterion for estimation of parameters, including the structure parameters, in a model for a random time series has been derived from the single and natural principle: minimize the number of bus it takes to write down the observed sequence. This criterion has the negative of the likelihood function as the dominant term. It also has terms that penalize an over parameterization. These terms differ from the corresponding term in Akaike's criterion, and they penalize the over parameterization more strongly, in particular, for large samples. Moreover, these terms depend on the degree the parameters affect the value of the first dominant maximum likelihood term.

REFERENCES

[1] A. KOLMOGOROV: Three approaches to the quantitative definition of information. *Probl. Inf. Transmission* **1,** (1) (1968).

[2] H. AKAIKE: Information theory and an extension of the maximum likelihood principle. *2nd International Symposium and Information Theory*, B. N. Petrov and F. Caski, Eds., pp. 267–281. Akademiai Kiado, Budapest.

[3] H. AKAIKE: On entropy maximization principle, P.R. Krishnaiah, Ed. *Applications of Statistics.* North-Holland (1977).

[4] G. SCHWARZ: Estimating the dimension of a model. *Ann. Stat.* **6,** (2) (1978).

[5] J. RISSANEN: Minmax entropy estimation for vector processes. *System Identification: Advances and Case Studies*, D. G. Lainiotis and R. K. Mehra, Eds., pp. 97–117. Academic Press (1976).

[6] J. RISSANEN and L. LJUNG: Estimation of optimum structures and parameters for linear systems. *Math. Syst. Theory*, **131,** 76–91 (1976).

[7] S. WATANABE: *Knowing and Guessing.* John Wiley & Sons (1969).

[8] N. ABRAMSON: *Information Theory and Coding.* McGraw-Hill (1963).

[9] S. KULLBACK: *Information Theory and Statistics.* John Wiley & Sons (1959).

[10] J. RISSANEN: Generalized Kraft-inequality and arithmetic coding. *IBM J. Res. Dev.* **20,** (3) 198–203, May (1976).

[11] E. J. HANNAN: *Multiple Time Series.* John Wiley & Sons (1970).

[12] C. S. WALLACE and D. M. BOULTON: An information measure for classification. *Comput. J.* **11,** (2) 185 (1968).

[13] K.-J. ÅSTRÖM and P. E. EYKHOFF: System identification, a survey. *Automatica* **7,** 123–162 (1971).