Model-structure selection by cross-validation

Measurement and Control Group
Department of Electrical Engineering
Eindhoven University of Technology
The Netherlands

MODEL-STRUCTURE SELECTION BY CROSS-VALIDATION

Petre Stoica*
Pieter Eykhoff
Peter Janssen
Torsten Söderström**

Internal Report ER 85/08
August 1985

*) Facultatea de Automatica, Institutul Politehnic Bucuresti, Bucharest; The work of this author was supported by the Ministry of Education and Science of the Netherlands and by the Measurement and Control Group, in the form of a Fellowship grant.

**)Uppsala University, Institute of Technology, Uppsala.

MODEL-STRUCTURE SELECTION BY CROSS-VALIDATION

Petre Stoica, Facultatea de Automatica, Institutul Politehnic Bucuresti, Splaiul Independentei 313, R-77206 Bucharest, Romania*)

Pieter Eykhoff, Peter Janssen, Dept. of Electrical Engineering, Eindhoven University of Technology, EUT, P.O. Box 513, 5600 MB Eindhoven, Netherlands

Torsten Söderström, Uppsala University, Institute of Technology, P.O. Box 534, S-751 21 Uppsala, Sweden.

Abstract: Two criteria for choosing between different model-structures are proposed. Their derivation is within a natural cross-validatory assessment context and is fairly assumption-free. In particular, the two criteria can be used for discriminating between non-nested model structures and, more important, the "true" system is not required to belong to the considered set of models. Should the true system belong to the model set, the two proposed criteria will asymptotically reduce to some well-known structure selection criteria. This is believed to be a desirable feature of our proposals. On the other hand, it provides a nice cross-validation interpretation of some well-known model structure selection rules. Also, the cross-validation interpretation helps to choose which of the criteria to be used in a given application.

The paper also has a second purpose which is somewhat decoupled from that mentioned above. It contains a rather extensive survey of the literature which may be useful in its own right.

*) The work of this author was supported by the Ministry of Education and Science of the Netherlands and by the Measurement and Control Group of the EUT, in the form of a fellowship grant.

1 INTRODUCTION AND REVIEW OF LITERATURE

Let S denote the system that generated the data. We shall generally assume that the data are realizations of stationary ergodic processes but otherwise will not impose any other restrictions on S.

Let $M(\theta)$ denote a model of S, where θ stands for the finite-dimensional ($\dim \theta < \infty$) vector of unknown parameters. When θ spans a set of feasible values, say Θ, then $M(\theta)$ describes a set of models, say M, M can (and will) be called a model structure. We will make some assumptions on the model $M(\theta)$ in section 2. Here let it suffice to say that those assumptions are fairly weak and that throughout this paper we will consider a general M rather than specializing the discussion to specific model structures.

Once a model structure M has been chosen, the problem of estimating the unknown parameter vector θ has a number of well-established solutions, see for example Åström and Eykhoff (1971), Eykhoff (1974), Kashyap and Rao (1976), Goodwin and Payne (1977), Ljung and Söderström (1983), Söderström and Stoica (1983).

However, an essential question is how to choose the model structure. It has been treated by many researchers and has received a number of answers. In Table 1, we present a review of the literature on the model structure selection problem. Needless to say, we do not claim that the table is "complete". We believe, however, that it includes most of the key references.

Clearly a certain familiarity with the topic is necessary in order to understand the various entries and comments of the table. We have to accept this situation since we cannot, in one single paper, give details on every procedure included in Table 1. For the paper, such detailed descriptions will not be needed since our aim is not to compare all the model-structure selection rules given in the table, but rather to introduce some new ones and to show how they are related with some selection-rules in Table 1.

The procedures which belong to the first four columns of the table are sometimes called "subjective selection rules" [see e.g. Chan et al. (1975)], the reason being that their application requires some subjective judgement (most typical, the choice of a significance level). Such procedures will not be discussed in this paper.

The next three columns of Table 1 contain the so-called "modern selection rules", the application of which does not require the choice of significance levels etc. (sometimes, they are also called "objective" but as we shall see their use is not completely free from subjectivity). Such selection rules will naturally occur in the subsequent discussions. Their properties will also be reviewed and extended to some degree.

Most solutions to the problem of model structure selection are tied to specific parameter estimation methods. The prediction error method (PEM) is a typical example of an estimation method for which model structure selection rules have been designed. Also, for most procedures, it is customary to consider a number of competitive model structures, say $\{M_i\}$, and to select the "right" structure by using a certain rule/criterion. It is generally assumed that the model sets M_i are nested and that the "true" system S belongs to one of these sets.

It is clear that in practice the assumption $S \in M_i$ is unlikely to be fulfilled. Then the aim of the model structure selection rule should not be that of choosing a "true" structure, simply because such a structure does not exist. The aim should be, rather, to find the "best" model structure within the considered set $\{M_i\}$, the "best" with respect to a certain criterion expressing the intended use of the model. To this end we may wish to compare non-nested structures as well.

While the above facts appear to be widely recognized, it seems that we still lack a model-structure selection rule incorporating all the desirable features mentioned above. In this paper we will try to fill this gap. The cross-validatory assessment, (Stone, 1974), will be the framework within which we will derive our model-structure selection criteria.

An outline of this paper is as follows. In the next section we state the problem and introduce the basic assumptions. In section 3 we derive a first cross-validation criterion which, in section 4, is shown to be asymptotically equivalent to Akaike's criteria, provided some additional assumptions are made. A second class of cross-validation criteria is proposed in section 5, while in section 6 it is shown that this class asymptotically includes the well-known criteria of Hannan, Kashyap, Rissanen and Schwarz, under certain additional assumptions. Finally, section 7 contains some concluding remarks.

2. PRELIMINARIES AND BASIC ASSUMPTIONS

Let us consider a generic model structure M and let $M(\theta)$ be a model belonging to M. We will assume that the estimate, say $\hat{\theta}$, of the unknown parameter vector of $M(\theta)$ is obtained as

$$\hat{\theta} = \arg \min_{\theta \in \Theta} V(\theta) \qquad V(\theta) = \frac{1}{N} \sum_{t=1}^{N} \varepsilon^2(t,\theta) \tag{2.1}$$

In (2.1) $\varepsilon(t,\theta)$ is the "residual" of $M(\theta)$ at time instant t and N is the number of data points.

Many parameter estimation methods currently in use are of the type (2.1), for example, the least-squares method (LSM), the output error method (OEM), the PEM, and - under the gaussian hypothesis - also the maximum likelihood method (MLM). This is true indeed since residuals $\{\varepsilon(t,\theta)\}$ in (2.1) can be defined in many ways. They can, for instance, be equation errors or output errors. They could also be one-step prediction errors or multi-step prediction errors etc. The above discussion also implies that by using criteria of the type (2.1), we can express a number of possible intended uses for an estimated model. For example, the "quality" of a simulation model, a prediction model or a model to be used for predictive control could be expressed by criteria such as (2.1).

In practice estimated models are quite often used for purposes such as those mentioned above. However, there are certainly intended uses of a model which cannot be expressed by criteria of the type (2.1). For such cases, the theory we shall develop in this paper will be only of a limited interest. There may indeed be little reason to use a good prediction model for spectral estimation (to give only one example). However, the basic ideas of this paper might be useful in also approaching, in a similar way, those other cases where the estimated model is to be used for another purpose than prediction, simulation etc. In our opinion, this, if possible, would be strongly recommended. The reason is simply the obvious (but sometimes neglected) fact that system identification should be done with the final aim of the model in mind.

Some further remarks on (2.1) are in order.

Remark 2.1 The quantities in (2.1) should normally be indexed to show that they correspond to the model structure M, for example Θ_M, $V_M(\theta_M)$, $\varepsilon_M(\cdot,\theta_M)$. However, to simplify the notation we shall omit the index M whenever there is no possibility of confusion. ●

Remark 2.2 In (2.1) we have implicitly assumed that $\{\varepsilon(t,\theta)\}$ are scalars. The extension to the multivariable case is possible but the notations will then be a great deal more complicated. This extension will eventually have to be presented elsewhere. ●

Remark 2.3 The analysis that follows can be directly extended to slightly more general criteria than (2.1) of the form

$$\frac{1}{N}\sum_{t=1}^{N} h\bigl(\varepsilon(t,\theta)\bigr)$$

with h(.) being some suitable function. Note that for a general distribution of the data, the ML criterion is of this type $\bigl[$take $h(\varepsilon) = -\ln f(\varepsilon)$, with $f(\varepsilon)$ the probability density function of $\varepsilon(t,\theta)\bigr]$. However, to keep the notation simple we will concentrate on the analysis of (2.1) and leave the extension to the more general criteria of the above form as a simple exercise for the reader. ●

Now, let us introduce some regularity conditions that will be assumed to hold true throughout the paper.

A1: $\varepsilon(t,\theta)$ is a sufficiently smooth function of θ so that its derivatives with respect to θ exist and are finite for any $\theta\in\Theta$, where Θ is the compact set of feasible values.

A2: The second-order derivative matrix

$$V_{\theta\theta}(\hat{\theta}) \triangleq \frac{\partial^2}{\partial\theta^2} V(\theta)\bigg|_{\theta=\hat{\theta}}$$

is positive definite. [In particular this implies that $\hat{\theta}$ is an isolated minimum point of $V(\theta)$].

A3: The residuals $\{\varepsilon(t,\theta)\}$ and

$$\varepsilon_{\theta}(t,\theta) \triangleq \frac{\partial}{\partial\theta}\varepsilon(t,\theta)$$

$$\varepsilon_{\theta\theta}(t,\theta) \triangleq \frac{\partial^2}{\partial\theta^2}\varepsilon(t,\theta)$$

are stationary and ergodic processes for any $\theta\in\Theta$. Moreover, we assume that the sample moments involving the above processes converge to the theoretical moments, as N tends to infinity, at a rate of order $0(1/\sqrt{N})$.

Assumptions A1 and A2 are fairly weak. Assumption A3 might appear rather technical and in any case difficult to check in a given practical situation. The ergodicity property, however, does not appear to be restrictive. It seems to be necessary in practice, where in general we have access to only one realization of the stochastic process under study. Once ergodicity is accepted, the rate of convergence of the sample moments is under rather general conditions of order $0(1/\sqrt{N})$, see e.g. Bartlett (1966). However, we would like to stress that this point is not essential for the analysis that comes. Should, however, the rate be smaller than $0(1/\sqrt{N})$, the main results of the paper will basically still hold; only the order of some remainder terms will be affected.

We shall also make a general assumption on the experimental conditions under which the data used in (2.1) were obtained. We thus assume that those conditions are the same as (or, more realistic, quite similar to) the experimental conditions under which the model will be operating. This assumption is not added as a fourth condition. It will not be used in the analysis. In fact, it is a general meaningful principle rather than just an assumption. The importance of this principle for identification from real-life data is emphasized, for example, by Ljung and Van Overbeek (1978).

Turn now to the problem of model structure determination which is the main theme of this paper. As is well known, minimizing the values of $V(\hat{\theta})$ obtained in different model structures is not an appropriate method for structure selection. Indeed, consider for example two nested structures M_1, M_2 with $M_1 \subset M_2$. Then we necessarily have

$$V_{M_1}(\hat{\theta}_{M_1}) \geq V_{M_2}(\hat{\theta}_{M_2})$$

even though M_1 may be a "better" structure than M_2. By "M_1 being a better structure than M_2" we mean that on data sets other than those used for estimation, M_1 will lead to smaller residual-sum-of-squares criteria more frequently than M_2.

A conceptually simple solution to the above dilemma is provided by what is called cross-checking or cross-validation. What this may mean is perhaps best illustrated by the following quotation from Stone (1974):

> "In its most primitive but nevertheless useful form [cross-validation] consists in the controlled or uncontrolled division of the data sample into two subsamples, the choice of a statistical predictor, including any necessary estimation, on one subsample and then the assessment of its performance by measuring its predictions against the other subsample".

Some refinements of the above "primitive" form of the cross-validatory assessment have been developed in Stone (1974) which was the main source of inspiration for our study.

In the next sections we shall propose two cross-validation schemata (which, we note in passing, can be seen as generalizations of Stone's scheme) for assessment criteria of the type (2.1). These two schemata will, in turn, lead to our model structure selection criteria.

3 FIRST CROSS-VALIDATION CRITERION

Let

$$I = \{1,2,\ldots,N\} \tag{3.1a}$$

and

$$I_p = \{(p-1)m+1,\ldots,pm\} \quad p = 1,\ldots,k-1 \tag{3.1b}$$

$$I_k = \{(k-1)m+1,\ldots,N\}$$

for some positive integer m and $k = [\frac{N}{m}]$. ($[x]$ denotes the largest integer not greater that x). Note that in places we shall let m and k depend on N.

For cross-validatory assessment of the model structure M, in this section we shall use the following criterion

$$C_I = \sum_{p=1}^{k} \sum_{t\in I_p} \varepsilon^2(t,\hat{\theta}_p) \tag{3.2}$$

where

$$\hat{\theta}_p = \arg \min_{\theta\in\Theta} \sum_{t\in I-I_p} \varepsilon^2(t,\theta) \qquad p = 1,\ldots,k \tag{3.3}$$

<u>Remark 3.1</u>: It may be worth noting that for dynamic systems, in general, we cannot have a <u>neat</u> division of I into an "estimation" subsample and a "check" subsample. For example, generally we shall need data from I_p to compute the estimation criterion in (3.3). This does not appear, however, to be a serious drawback and we should accept this situation since correcting it, even if possible in principle, would complicate the analysis a great deal. After all, there is a clear separation between the residuals used for checking, (3.2), and those used for estimation, (3.3), and this seems to be what is important. ●

<u>Remark 3.2</u>: We will assume that all the intervals $\{I_p\}_{p=1}^{k}$ have the same length m. This assumption will simplify the notation and also some calculations. However, the length of the last interval I_k will, in general, be larger than m (but, of course, smaller than 2m). It is not difficult to see that the main results derived in the following sections remain

valid also when this fact is recognized. More specifically, we will use the assumption that $\#I_p = m$ (for $p = 1,\ldots,k$) in equations (3.8) and (5.7) below. The corresponding (intermediary) results (3.8) and (5.13), respectively, obtained under this assumption remain unchanged if we let $\#I_k$ belong to the interval (m, 2m). We omit the straightforward calculations showing this. ●

The exact evaluation of the assessment criterion C_I, (3.2), even if clearly possible, may not be advisable since the computing time required will be prohibitive for many applications. In the following, we will derive an asymptotically valid approximation of C_I which is much easier to compute.

<u>Theorem 3.1</u> Let assumptions A1-A3 be true. Then for k large enough we have

$$\frac{1}{N} C_I = C_1 + 0\left(\frac{1}{k^2 m}\right) \tag{3.4}$$

where

$$\boxed{\begin{aligned} C_1 &\triangleq V(\hat{\theta}) + \frac{4}{N^2} \sum_{p=1}^{k} w_p^T(\hat{\theta})\, V_{\theta\theta}^{-1}(\hat{\theta}) w_p(\hat{\theta}) = \\ &= V(\hat{\theta}) + \frac{4}{N^2} \operatorname{tr} V_{\theta\theta}^{-1}(\hat{\theta})\, W(\hat{\theta}) \end{aligned}} \tag{3.5a}$$

with

$$\boxed{\begin{aligned} w_p(\hat{\theta}) &= \sum_{t\ I_p} \varepsilon(t,\hat{\theta})\varepsilon_\theta(t,\hat{\theta}) \qquad p = 1,\ldots,k \\ W(\hat{\theta}) &= \sum_{p=1}^{k} w_p(\hat{\theta})\, w_p^T(\hat{\theta}) \end{aligned}} \tag{3.5b}$$

The above result holds for both "large" and "small" m's.

<u>Proof</u>: For sufficiently large k, $\hat{\theta}_p$ is close to $\hat{\theta}$, (2.1), and then we can write:

$$\varepsilon^2(t,\hat{\theta}_p) = \varepsilon^2(t,\hat{\theta}) + 2\varepsilon(t,\hat{\theta})\varepsilon_\theta^T(t,\hat{\theta})(\hat{\theta}_p-\hat{\theta}) + 0(|\hat{\theta}_p-\hat{\theta}|^2) \tag{3.6}$$

Similarly we have

$$0 = \frac{\partial}{\partial\theta}\left[\frac{1}{N}\sum_{t\in I-I_p}\varepsilon^2(t,\theta)\right]\Bigg|_{\theta=\hat\theta_p} = V_\theta(\hat\theta_p) - \frac{2}{N}\sum_{t\in I_p}\varepsilon(t,\hat\theta_p)\varepsilon_\theta(t,\hat\theta_p) =$$

$$= V_\theta(\hat\theta) - \frac{2}{N}\sum_{t\in I_p}\varepsilon(t,\hat\theta)\varepsilon_\theta(t,\hat\theta) +$$

$$\left\{V_{\theta\theta}(\hat\theta) - \frac{m}{N}\frac{\partial}{\partial\theta}\left[\frac{2}{m}\sum_{t\in I_p}\varepsilon(t,\theta)\varepsilon_\theta(t,\theta)\right]\Bigg|_{\theta=\hat\theta}\right\}(\hat\theta_p-\hat\theta) + 0(|\hat\theta_p-\hat\theta|^2) =$$

$$= -\frac{2}{N}\sum_{t\in I_p}\varepsilon(t,\hat\theta)\varepsilon_\theta(t,\hat\theta) + \left[V_{\theta\theta}(\hat\theta) + 0(\tfrac{1}{k})\right](\hat\theta_p-\hat\theta) + 0(|\hat\theta_p-\hat\theta|^2) \tag{3.7}$$

Since (E denotes expectation)

$$\frac{1}{m}\sum_{t\in I_p}\varepsilon(t,\hat\theta)\varepsilon_\theta(t,\hat\theta) = E\varepsilon(t,\hat\theta)\varepsilon_\theta(t,\hat\theta) + 0(\tfrac{1}{\sqrt{m}}) =$$

$$= \left[\frac{1}{N}\sum_{t=1}^{N}\varepsilon(t,\hat\theta)\varepsilon_\theta(t,\hat\theta) + 0(\tfrac{1}{\sqrt{N}})\right] + 0(\tfrac{1}{\sqrt{m}}) = 0(\tfrac{1}{\sqrt{m}}) \tag{3.8}$$

it follows from (3.7) that $|\hat\theta_p - \hat\theta| = 0(\frac{1}{k\sqrt{m}})$.

[Note that for "small" m, $0(1/\sqrt{m})$ should be interpreted as $0(1)$].

Therefore we get from (3.7) the following asymptotically valid expression for $(\hat\theta_p-\hat\theta)$:

$$\hat\theta_p - \hat\theta = V_{\theta\theta}^{-1}(\hat\theta)\,\frac{2}{N}\sum_{t\in I_p}\varepsilon(t,\hat\theta)\varepsilon_\theta(t,\hat\theta) + 0(\frac{1}{k^2\sqrt{m}}) \tag{3.9}$$

where we have used that

$$\left[V_{\theta\theta}(\hat\theta) + 0(\tfrac{1}{k})\right]^{-1} = V_{\theta\theta}^{-1}(\hat\theta) + 0(\tfrac{1}{k}) \tag{3.10}$$

Finally, from (3.6) and (3.9) we have that

$$\frac{1}{N}C_I = \frac{1}{N}\sum_{p=1}^{k}\sum_{t\in I_p}\varepsilon^2(t,\hat\theta) + \frac{2m}{N}\sum_{p=1}^{k}\left[\frac{1}{m}\sum_{t\in I_p}\varepsilon(t,\hat\theta)\varepsilon_\theta^T(t,\hat\theta)\right].$$

$$\cdot\left[V_{\theta\theta}^{-1}(\hat\theta)\,\frac{2}{N}\sum_{s\in I_p}\varepsilon(s,\hat\theta)\varepsilon_\theta(s,\hat\theta) + 0(\frac{1}{k^2\sqrt{m}})\right] + 0(\frac{1}{k^2 m})$$

$$= C_1 + 0(\frac{1}{k^2 m})$$

and the proof is finished. ●

The (approximate) cross-validation criterion C_{1m} will apparently be much easier to compute than C_{Im}. [Note that for convenience of the following discussion we emphasize by notation the dependence of C_I and C_1 on m]. The calculation of C_{1m} is particularly simple when the minimization in (2.1) is performed by using a Newton-Raphson algorithm (and indeed this may be the case quite often). Then both $V_{\theta\theta}^{-1}(\hat{\theta})$ and $\{w_p(\hat{\theta})\}_{p=1}^{k}$ can be obtained from the last iteration of the minimization algorithm without any additional calculations. Once $V_{\theta\theta}^{-1}(\hat{\theta})$ and $\{w_p(\hat{\theta})\}_{p=1}^{k}$ are given we can use one of the two expressions given in (3.5a) to compute C_{1m}. Note that depending on the values of m, N and dim θ one of these two expressions may be computationally more efficient than the other. The number of arithmetic operations required to evaluate either of these expressions can easily be counted. We do not insist on this aspect since it seems minor.

We now state our first model structure selection rule.

<u>First cross-validation model structure selection rule</u>: Choose the model structure which leads to the smallest value of C_{1m}, where C_{1m} is defined by (3.5). ●

The above selection procedure depends on m, and the choice of this parameter should thus be discussed. We cannot give precise rules on how to choose m. However, the cross-validation interpretation of the selection criterion may, at least, give some ideas about the value m should have in a particular application. Indeed, we can expect that the model (structure) which will minimize C_{1m} will asymptotically minimize C_{Im} as well (see Section 5 for a discussion on this point). Then given the obvious interpretation of C_{Im}, it would appear that the value of m should be chosen so as to indicate, on how many future sampling points we intend to use our model which is estimated from N data points. More specifically, suppose we wish to use the estimated model at some n (say) future time instants. Then we may choose m such that

$$\frac{m}{N-m} \approx \frac{n}{N} \tag{3.11}$$

This choice will assure the desired ratio between the "check" and the "estimation" sample lengths.

Remark 3.3: For convenience of the subsequent discussion we introduce the following terminology. When $n << N$ we say that the estimated model is a "short-term" model, or perhaps, a more suggestive description, a "short term operating" model [a "one-step" model if $n = 1$], and we call it a "long-term" model if $n >> N$. This wording might be somewhat unconventional but should not be confusing. ●

Note that since $m/(N-m)$ must be small enough for C_{1m} to be a good approximation of C_{Im}, it follows from (3.11) that C_{1m} can be used only for "small" n/N ratios (in such cases (3.11) implies $m \approx n$). In the terminology of Remark 3.3 we can therefore say that C_{1m} can be used to select a good "short-term" model structure.

We should also remark that the remainder term in (3.4) depends on m. The smaller m is, the better the approximation order seems to be. For $m = 1$ we apparently get the best approximation order (then the difference $\left(\frac{1}{N} C_{Im}-C_{1m}\right)$ is $0(1/N^2)$ which is quite small indeed). The choice $m = 1$ is advocated by Stone (1974) on heuristical grounds (cf. also the discussion on Stone's paper). As a matter of fact, Stone used exact cross-validation criteria so he could not invoke in favour to the "one-at-a-time omission schema" the improvement in approximation degree that, as mentioned above, may result for $m = 1$.

Even if the difference between the exact cross-validation criterion C_{Im} and its asymptotically valid approximation C_{1m} may increase when m increases we, depending on the type of application, may wish to consider also values $m > 1$. This point will be further discussed in the next sections.

4. ASYMPTOTIC EQUIVALENCE WITH AKAIKE'S CRITERIA

The cross-validation criterion C_{1m} introduced in the previous section appears to have a number of desirable features.

First, for sufficiently large k, C_{1m} has a nice cross-validation interpretation. Second, it can be shown that C_{1m} is invariant to parameter scale changes. [A discussion on the importance of this point may be found, for example, in Rissanen (1976)]. Third, and more important, in order to use C_{1m} for model structure selection we need to assume neither that the structures $\{M_i\}$ under consideration are nested, nor that $S \in M_i$ for some i. Only the fairly weak conditions A1-A3 need to be true.

In the following we will show that if certain additional assumptions are introduced then C_{1m} can be expected to asymptotically behave like the well-known and frequently used Akaike's criteria (Akaike, 1969, 1973, 1974, 1976, 1981). This is also considered to be a desirable feature of our first cross-validation criterion C_{1m}.

The equivalence of the choice of model structure by cross-validation and Akaike's criteria is not unexpected. Akaike's selection rules can be interpreted as cross-validatory prediction assessments; cf., e.g., Söderström (1977). In fact, the Akaike Information Criterion (AIC) was shown to be asymptotically equal under certain conditions to $\ln \frac{1}{N_1} C_{I1}$, Stone (1977). Here we shall prove the asymptotic equality of $\ln \frac{1}{N} C_{I1}$ and AIC as well as that of $\frac{1}{N} C_{I1}$ and Akaike's FPE (Final Prediction Error) criterion under more general conditions than Stone's. In particular we do not assume that S necessarily belongs to M. We shall also give the order of the difference between $\frac{1}{N} C_{I1}$ $(\ln \frac{1}{N} C_{I1})$ and the FPE criterion (AIC). Furthermore, we shall consider the asymptotic equivalence between $\frac{1}{N} C_{Im}$ or $\ln \frac{1}{N} C_{Im}$ and Akaike's criteria also for $m > 1$.

It is shown in Ljung and Caines (1979) that under weak conditions (implied by our A1-A3), as N tends to infinity

$$\hat{\theta} \to \theta^* = \arg \min_{\theta \in \Theta} E\varepsilon^2(t,\theta) \quad (\text{wp } 1) \tag{4.1a}$$

and

$$\left| \hat{\theta} - \theta^* \right| = 0\left(\frac{1}{\sqrt{N}}\right) \tag{4.1b}$$

Introduce the following assumption

B1 : $E\, \varepsilon(t,\theta^*)\varepsilon_{\theta\theta}(t,\theta^*) = 0$

The above condition is more general than that requiring $S \in M$. For example, in the case of least-squares model structures (for which $\varepsilon(t,\theta)$ is linear in θ) B1 is trivially satisfied under general conditions since for such structures we have $\varepsilon_{\theta\theta}(t,\theta) = 0$ any θ. Furthermore, even if we were to assume that $S \in M$ so that B1 follow, we need not require that $\varepsilon(t,\theta^*)$ is white noise, which seems to be the usual condition imposed in other analyses of Akaike's criteria. Think, for instance, of an OE model for which B1 follows once we accept that $S \in M$, but $\{\varepsilon(t,\theta^*)\}$ may well be a correlated process.

It is now possible to state the result on the asymptotic equivalence between $\frac{1}{N} C_{I1}$ and Akaike's AIC and FPE criteria which for the problem under study are given by [see, e.g., Akaike (1974, 1976, 1981)]:

$$\text{AIC} = \ln V(\hat{\theta}) + \frac{2}{N} \dim \theta \tag{4.2}$$

$$\text{FPE} = V(\hat{\theta}) \frac{N + \dim \theta}{N - \dim \theta} \tag{4.3}$$

Theorem 4.1 Let assumptions A1-A3 and B1 be true. Assume that either $\varepsilon(t,\theta^*)$ and $\varepsilon_\theta(t,\theta^*)$ are gaussian distributed or that they are general linear random processes. Then, for sufficiently large N it holds that

$$\ln \frac{1}{N} C_{I1} = \text{AIC} + 0\left(\frac{1}{N^{3/2}}\right) \tag{4.4}$$

$$\frac{1}{N} C_{I1} = \text{FPE} + 0\left(\frac{1}{N^{3/2}}\right) \tag{4.5}$$

Proof: From Theorem 3.1 we have that (for large N)

$$\frac{1}{N} C_{I1} = V(\hat{\theta}) + \frac{4}{N} \operatorname{tr} V_{\theta\theta}^{-1} (\hat{\theta}) . \frac{1}{N} W(\hat{\theta}) + 0\left(\frac{1}{N^2}\right) \tag{4.6a}$$

where

$$W(\hat{\theta}) = \sum_{t=1}^{N} \varepsilon^2(t,\hat{\theta})\, \varepsilon_\theta(t,\hat{\theta})\, \varepsilon_\theta^T(t,\hat{\theta}) \tag{4.6b}$$

Now, under the assumptions made, we asymptotically have

$$V_{\theta\theta}(\hat{\theta}) = V_{\theta\theta}(\theta^*) + 0\left(\frac{1}{\sqrt{N}}\right) = 2E\left[\varepsilon_\theta(t,\theta^*)\, \varepsilon_\theta^T(t,\theta^*) + \right.$$
$$\left. + \varepsilon(t,\theta^*)\, \varepsilon_{\theta\theta}(t,\theta^*)\right] + 0\left(\frac{1}{\sqrt{N}}\right) = 2E\varepsilon_\theta(t,\theta^*)\, \varepsilon_\theta^T(t,\theta^*) + 0\left(\frac{1}{\sqrt{N}}\right) \tag{4.7}$$

and

$$\frac{1}{N} W(\hat{\theta}) = E\, \varepsilon^2(t,\theta^*)\, \varepsilon_\theta(t,\theta^*)\, \varepsilon_\theta^T(t,\theta^*) + 0\left(\frac{1}{\sqrt{N}}\right) =$$
$$= \left[E\, \varepsilon^2(t,\theta^*)\right]\left[E\varepsilon_\theta(t,\theta^*)\, \varepsilon_\theta^T(t,\theta^*)\right] + 0\left(\frac{1}{\sqrt{N}}\right) =$$
$$= V(\hat{\theta}) . E\varepsilon_\theta(t,\theta^*)\varepsilon_\theta^T(t,\theta^*) + 0\left(\frac{1}{\sqrt{N}}\right) \tag{4.8}$$

In establishing the second equality in (4.8) we have assumed that the well-known formula

$$Ex_1 x_2 x_3 x_4 = (Ex_1 x_2)(Ex_3 x_4) + (Ex_1 x_3)(Ex_2 x_4) +$$
$$+ (Ex_1 x_4)(Ex_2 x_3) \tag{4.9}$$

can be applied to the random variables

$$x_1 = x_2 = \varepsilon(t,\theta^*) , \quad x_3 = x_4^T = \varepsilon_\theta(t,\theta^*).$$

The formula (4.9) is known to hold if $\{x_i\}$ i=1,...4 are either gaussian distributed or general linear random variables (see Bartlett (1966) for example). By using (4.9) the second equality in (4.8) easily follows after noticing that (4.1a) implies

$$E\varepsilon(t,\theta^*)\, \varepsilon_\theta(t,\theta^*) = 0 \tag{4.10}$$

Introducing (4.7) and (4.8) in (4.6) we obtain the following asymptotically valid expression for $\frac{1}{N} C_{I1}$

$$\frac{1}{N} C_{I1} = V(\hat{\theta})\left[1 + \frac{2}{N} \dim \theta\right] + 0\left(\frac{1}{N^{3/2}}\right) \qquad (4.11)$$

The assertions of the theorem readily follow from (4.11). Indeed we have

$$\ln \frac{1}{N} C_{I1} = \ln V(\hat{\theta}) + \ln\left[1 + \frac{2}{N} \dim \theta\right] + 0\left(\frac{1}{N^{3/2}}\right) =$$

$$= \ln V(\hat{\theta}) + \frac{2}{N} \dim \theta + 0\left(\frac{1}{N^2}\right) + 0\left(\frac{1}{N^{3/2}}\right) =$$

$$= \quad \text{AIC} + 0\left(\frac{1}{N^{3/2}}\right) \qquad (4.12)$$

which shows (4.4). To prove (4.5) note that

$$\text{FPE} = V(\hat{\theta})\left[1 + \frac{2 \dim \theta}{N - \dim \theta}\right] = V(\hat{\theta})\left[1 + \frac{2 \dim \theta}{N}\left(1 + \frac{\dim \theta}{N - \dim \theta}\right)\right] =$$

$$= V(\hat{\theta})\left[1 + \frac{2}{N} \dim \theta\right] + 0\left(\frac{1}{N^2}\right) = \frac{1}{N} C_{I1} + 0\left(\frac{1}{N^{3/2}}\right) \qquad (4.13)$$

With this observation the proof is concluded. ●

As a consequence of the above theorem we can expect that for N large enough both AIC and FPE will select the model structure that minimizes the cross-validation criterion C_{I1}. For this to hold, we need to assume neither that the compared structures are nested nor that the system is necessarily included in the structure set under consideration. However, we need assumption B1 to hold. Despite this last remark, the above discussion appears to offer further support to the by now widespread opinion that Akaike's criteria will select model structures with a rather strong intuitive appeal in quite a variety of practical situations. For the present case, given the interpretation of C_{I1}, we can say that under mild conditions, the models selected by AIC or FPE will be good <u>one-step</u> models.

On the other hand, the models selected by using C_{11} will possess the aforementioned feature in more general situations; so if that feature is indeed desirable, then C_{11} might be preferred to AIC or FPE.

Now, let us consider the possible equivalence between $\frac{1}{N} C_{Im}$ and Akaike's criteria for $m > 1$ (eventually $m \to \infty$). It will turn out that AIC or FPE asymptotically behave like $\frac{1}{N} C_{Im}$ also for $m > 1$ provided the following additional assumption holds:

B2 : $\{\varepsilon(t,\theta^*)\}$ is white noise.

This assumption is quite strong. It is essentially equivalent to requiring that $S \in M$ and that $\{\varepsilon(t,\theta)\}$ are one-step ahead prediction errors. Also note that for causal models B2 implies B1.

For $m > 1$ and k large enough we have, cf. (3.5):

$$\frac{1}{N} C_{Im} = V(\hat{\theta}) + \frac{4m^2k}{N^2} \operatorname{tr} V^{-1}_{\theta\theta}(\hat{\theta}) \cdot \frac{1}{m^2k} W(\hat{\theta}) + 0\left(\frac{1}{k^2m}\right) \qquad (4.14)$$

with

$$\frac{1}{m^2k} W(\hat{\theta}) = \frac{1}{k} \sum_{p=1}^{k} \Big[\frac{1}{m} \sum_{t \in I_p} \varepsilon(t,\hat{\theta})\varepsilon_{\theta}(t,\hat{\theta})\Big]\Big[\frac{1}{m} \sum_{s \in I_p} \varepsilon(s,\hat{\theta})\varepsilon^T_{\theta}(s,\hat{\theta})\Big] \qquad (4.15)$$

It follows from (3.8) that $\frac{1}{m^2k} W(\hat{\theta})$ is $0(\frac{1}{m})$ $\left[\text{for "small" } m\ 0(\frac{1}{m}) \text{ should be interpreted as } 0(1)\right]$. Hence the second term in (4.14) is $0(1/N)$ also for $m > 1$ (possibly m very large). However, whether or not it is asymptotically equal to $2V(\hat{\theta}) \frac{\dim \theta}{N}$ seems to be a more technical question than in the case $m = 1$. We can, however, proceed heuristically. Thus we can expect that for large k

$$\frac{1}{mk} W(\theta^*) = \frac{1}{m} \sum_{t \in I_p} \sum_{s \in I_p} E\varepsilon(t,\theta^*)\varepsilon(s,\theta^*)\varepsilon_{\theta}(t,\theta^*)\varepsilon^T_{\theta}(s,\theta^*) +$$

$$+ 0\left(\frac{1}{\sqrt{k}}\right) = V(\hat{\theta}) E\varepsilon_{\theta}(t,\theta^*)\varepsilon^T_{\theta}(t,\theta^*) + 0\left(\frac{1}{\sqrt{k}}\right) \qquad (4.16)$$

The last equality in (4.16) follows from our assumption that $\varepsilon(t,\theta^*)$ is a white process (then, in particular, $E\varepsilon(t,\theta^*)\varepsilon_{\theta}(s,\theta^*) = 0$ for $t > s$), after application of (4.9).

Invoking (4.1) we can now write

$$\frac{1}{N}\, C_{Im} = V(\hat{\theta}) + \frac{4mk}{N^2}\ \mathrm{tr}\{[2E\varepsilon_{\theta}(t,\theta^*)\varepsilon_{\theta}^{T}(t,\theta^*)]^{-1} + 0(\frac{1}{\sqrt{N}})\}\cdot$$

$$\cdot\{V(\hat{\theta})[E\varepsilon_{\theta}(t,\theta^*)\varepsilon_{\theta}^{T}(t,\theta^*)] + 0(\frac{1}{\sqrt{k}})\} + 0(\frac{1}{k^2 m}) =$$

$$= V(\hat{\theta})[1 + \frac{2}{N}\dim\theta] + 0(\frac{1}{mk^{3/2}}) \qquad (4.17)$$

The above relation, together with (4.11)-(4.13) shows that if the assumptions of theorem 4.1 hold, and if B2 holds, then for m > 1

$$\frac{1}{N}\, C_{Im} = \mathrm{FPE} + 0(\frac{1}{mk^{3/2}}) \qquad (4.18a)$$

$$\ln\frac{1}{N}\, C_{Im} = \mathrm{AIC} + 0(\frac{1}{mk^{3/2}}) \qquad (4.18b)$$

Thus it follows that for two structures, say $\bar{M}$ and $\tilde{M}$, satisfying B2 we can, for N large enough, expect that

$$\frac{1}{N}\, C_{Im}^{\bar{M}} - \frac{1}{N}\, C_{Im}^{\tilde{M}} \simeq \mathrm{FPE}_{\bar{M}} - \mathrm{FPE}_{\tilde{M}} \qquad (4.19)$$

and similarly for $\ln\frac{1}{N} C_{Im}$ and AIC. Now, assume that (at least) $\bar{M}$ does not satisfy B2. For such an under-parametrized model structure (4.18) does not necessarily hold. However, since in such a case $V_{\bar{M}}(\hat{\theta}_{\bar{M}}) - V_{\tilde{M}}(\hat{\theta}_{\tilde{M}}) = 0(1)$ is the dominant term for both sides of (4.19) we can still conclude that (4.19) holds asymptotically. Hence we have established the asymptotic equivalence between $\frac{1}{N} C_{Im}$ and Akaike's criteria also for the case m > 1. Since we used the quite restrictive assumption B2 in showing this equivalence, the cross-validation interpretation of AIC and FPE given by the above result is more of theoretical than practical interest. The interpretation is that under B2 the models selected by AIC or FPE will not only be "good one-step models" [see the discussion following Theorem 4.1] but also "good short-term models" (recall that the length m of the "check" subsample must be much smaller than N-m, the length of the "estimation" subsample, for the above result to hold).

Needless to say the models selected by minimizing C_{1m} can be interpreted as "good short-term models" (in the sense of minimizing C_{Im}) under (much) more general conditions. The criterion C_{1m} might thus be preferable in some applications to FPE or AIC even if it is computationally more complex.

Now, let us assume for a moment that the assumption S M holds. Furthermore, let M be the smallest model set containing S. The true structure therefore is M. As is well known, the structure minimizing AIC or FPE is not a consistent estimate of M [see, e.g. Shibata (1976), Söderström (1977), Kashyap (1980)]. In particular there exists a non-zero probability, even asymptotically, to over-estimate the true structure. In view of the asymptotic equivalence shown above between C_{1m} and AIC or FPE, the same will be true for C_{1m}. However, this should not be seen as a serious drawback. After all, C_{1m} (like Akaike's criteria) was not designed to provide a consistent estimate of the "true" structure, but rather a "good short-term model structure"; and the two structures just mentioned do not necessarily coincide (!); see, e.g. Stoica and Söderström (1982). It is rather intuitive that the attempt to select a good short-term model structure may lead to overestimating the true structure. The overfitting that may result when using AIC, FPE or C_{1m} on simulated data should be understood in the above light.

With the previous discussion in mind, we may suspect that the simple fact that the check subsample is much shorter than the estimation subsample may be the reason for the inconsistency of the selection rule based on C_{1m}. The consistency might appear for selection rules designed to select "good long-term model structures", therefore for cross-validation rules in which the check subsample is (much) larger than the estimation one. This observation leads us to our second cross-validation criterion which we present in the next section. In section 6 we show that the conjecture made above is valid.

5. SECOND CROSS-VALIDATION CRITERION

We now consider the following criterion for cross-validatory assessment of model structure M

$$C_{II} = \sum_{p=1}^{k} \sum_{t\in I-I_p} \varepsilon^2(t,\hat{\theta}_p) \tag{5.1}$$

where

$$\hat{\theta}_p = \arg \min_{\theta\in\Theta} \sum_{t\ I_p} \varepsilon^2(t,\theta) \qquad p = 1,\dots,k \tag{5.2}$$

All quantities appearing in (5.1) and (5.2) have been previously defined [thus I and $\{I_p\}_{p=1}^{k}$ are given by (3.1)]. The length of the check subsample, N-m, is now (much) larger than the estimation subsample length, m. Otherwise C_{II} is quite similar to C_I and, in fact, both criteria could have been presented in a unified framework. However, as we shall see, the analysis of C_I could not be repeated here. The asymptotic analysis of C_{II} needs more detailed consideration.

In this section our principal concern will be to obtain an asymptotically valid approximation of C_{II} that will be (much) easier to compute than the exact cross-validation criterion [(5.1), (5.2)].

Theorem 5.1 Let assumptions A1-A3 be true. Then for m and k large enough we have the relation:

$$\frac{1}{(k-1)N} C_{II} = C_2 + 0\left(\frac{1}{\min(N,m^{3/2})}\right) \tag{5.3a}$$

where

$$\boxed{\begin{aligned} C_2 &\triangleq V(\hat{\theta}) + \frac{2k}{N^2} \operatorname{tr} V_{\theta\theta}^{-1}(\hat{\theta})\, W(\hat{\theta}) \\ &= V(\hat{\theta}) + \frac{2k}{N^2} \sum_{p=1}^{k} w_p^T(\hat{\theta})\, V_{\theta\theta}^{-1}(\hat{\theta}) w_p(\hat{\theta}) \end{aligned}} \tag{5.3b}$$

and where $w_p(\hat{\theta})$ and $W(\hat{\theta})$ are defined in (3.5b).

Proof: For a sufficiently large m, $\hat{\theta}_p$ is "close" to $\hat{\theta}$, (2.1), and then we can write

$$\frac{1}{(k-1)N} C_{II} = \frac{1}{(k-1)N} \sum_{p=1}^{k} \sum_{t\in I-I_p} \{\varepsilon^2(t,\hat{\theta}) + 2\varepsilon(t,\hat{\theta})\varepsilon_{\theta}^T(t,\hat{\theta})(\hat{\theta}_p-\hat{\theta}) +$$

$$+ \frac{1}{2} (\hat{\theta}_p-\hat{\theta})^T 2[\varepsilon_\theta(t,\hat{\theta})\varepsilon_\theta^T(t,\hat{\theta}) + \varepsilon(t,\hat{\theta})\varepsilon_{\theta\theta}(t,\hat{\theta})](\hat{\theta}_p-\hat{\theta})\} +$$

$$+ 0(|\hat{\theta}_p-\hat{\theta}|^3) \triangleq T_1 + T_2 + T_3 + 0(|\hat{\theta}_p-\hat{\theta}|^3) \tag{5.4}$$

The evaluation of the first term T_1 in (5.4) is readily achieved.

$$T_1 = \frac{1}{(k-1)N} \sum_{p=1}^{k} \sum_{t\in I-I_p} \varepsilon^2(t,\hat{\theta}) = \tag{5.5}$$

$$= \frac{1}{(k-1)N} \sum_{p=1}^{k} [\sum_{t=1}^{N} \varepsilon^2(t,\hat{\theta}) - \sum_{t\in I_p} \varepsilon^2(t,\hat{\theta})] = V(\hat{\theta})$$

To evaluate the second and third term, T_2 and T_3, we need an asymptotically valid expression for the difference $(\hat{\theta}_p-\hat{\theta})$. For m large enough we can write [cf. also (5.2)]

$$0 = \frac{2}{m} \sum_{t\in I_p} \varepsilon(t,\hat{\theta}_p)\varepsilon_\theta(t,\hat{\theta}_p) = \frac{2}{m} \sum_{t\in I_p} \varepsilon(t,\hat{\theta})\, \varepsilon_\theta(t,\hat{\theta}) +$$

$$+ \{\frac{\partial^2}{\partial\theta^2} [\frac{1}{m} \sum_{t\in I_p} \varepsilon^2(t,\theta)]\Big|_{\theta=\hat{\theta}} \} (\hat{\theta}_p-\hat{\theta}) + 0(|\hat{\theta}_p-\hat{\theta}|^2) \tag{5.6}$$

Arguments similar to (3.8) now give

$$\frac{\partial^2}{\partial\theta^2} [\frac{1}{m} \sum_{t\in I_p} \varepsilon^2(t,\theta)] = V_{\theta\theta}(\theta) + 0(\frac{1}{\sqrt{m}}) \tag{5.7}$$

and

$$\frac{1}{m} \sum_{t\in I_p} \varepsilon(t,\hat{\theta})\varepsilon_\theta(t,\hat{\theta}) = 0(\frac{1}{\sqrt{m}}) \tag{5.8}$$

We therefore get from (5.6)-(5.8) that

$$|\hat{\theta}_p-\hat{\theta}| = 0(\frac{1}{\sqrt{m}}) \tag{5.9}$$

and

$$\hat{\theta}_p-\hat{\theta} = -V_{\theta\theta}^{-1}(\hat{\theta}) \frac{2}{m} \sum_{t\in I_p} \varepsilon(t,\hat{\theta})\varepsilon_\theta(t,\hat{\theta}) + 0(|\hat{\theta}_p-\hat{\theta}|^2) \tag{5.10}$$

It is now possible to evaluate the magnitude of T_2. Since, as we shall see, this is a higher order term we do not need an explicit expression for it.

$$T_2 = \frac{2}{(k-1)N} \sum_{p=1}^{k} \Big[\sum_{t\in I-I_p} \varepsilon(t,\hat\theta)\, \varepsilon_\theta^T(t,\hat\theta)\Big]\, (\hat\theta_p - \hat\theta) =$$

$$= \frac{2}{(k-1)N} \sum_{p=1}^{k} \Big[- \sum_{t\in I_p} \varepsilon(t,\hat\theta)\varepsilon_\theta^T(t,\hat\theta)\Big]\, 0\Big(\frac{1}{\sqrt{m}}\Big) =$$

$$= - \frac{2m}{(k-1)N} \sum_{p=1}^{k} \Big[\frac{1}{m} \sum_{t\in I_p} \varepsilon(t,\hat\theta)\varepsilon_\theta^T(t,\hat\theta)\Big]\, 0\Big(\frac{1}{\sqrt{m}}\Big) =$$

$$= \frac{mk}{(k-1)N}\; 0\Big(\frac{1}{m}\Big) = 0\Big(\frac{1}{N}\Big) \qquad (5.11)$$

We now proceed to evaluate the third term T_3. First we note that for k large enough

$$\frac{1}{N} \sum_{t\in I-I_p} 2\big[\varepsilon_\theta(t,\hat\theta)\varepsilon_\theta^T(t,\hat\theta) + \varepsilon(t,\hat\theta)\varepsilon_{\theta\theta}(t,\hat\theta)\big] =$$

$$= V_{\theta\theta}(\hat\theta) + 0\Big(\frac{1}{k}\Big) \qquad (5.12)$$

It follows from (5.10), (5.12) and the definition of T_3 that

$$T_3 = \frac{2}{k-1}\,\mathrm{tr} \sum_{p=1}^{k} \Big[V_{\theta\theta}(\hat\theta) + 0\Big(\frac{1}{k}\Big)\Big]\Big[V_{\theta\theta}^{-1}(\hat\theta).\, \frac{1}{m}\, w_p(\hat\theta).$$

$$.\, \frac{1}{m}\, w_p^T(\hat\theta)\; V_{\theta\theta}^{-1}(\hat\theta) + 0\Big(\frac{1}{m^{3/2}}\Big)\Big] =$$

$$= \frac{2}{m^2(k-1)}\,\mathrm{tr}\, V_{\theta\theta}^{-1}(\hat\theta)W(\hat\theta) + 0\Big(\frac{1}{\min(N,m^{3/2})}\Big) =$$

$$= \frac{2k}{N^2}\,\mathrm{tr}\, V_{\theta\theta}^{-1}(\hat\theta)W(\hat\theta) + 0\Big(\frac{1}{\min(N,m^{3/2})}\Big) \qquad (5.13)$$

The last equality in (5.13) follows after some straightforward calculations. The assertion of the theorem now follows from (5.4), (5.5), (5.11) and (5.13). ●

The expressions for the (approximate) cross-validation critera C_1 and C_2 are strikingly similar. The remarks made in Section 3 on the calculation of C_1 clearly apply to C_2 as well; they will not be repeated here.

Despite this similiarity there exists, in fact, an important difference between C_1 and C_2. The second term in C_{1m} is $0(1/N)$ for any m [see, for example, the discussion following (4.15)]. In (5.3b) the second term is $0(1/m)$. This can easily be seen for instance from (5.8), (5.13). Since k is supposed to tend to infinity (as N tends to infinity) the second term in C_2 will take (much) larger values than the corresponding term of C_1.

The assumption that k is "large enough" used in deriving C_2 is perhaps worth discussing. It cannot be removed without affecting the expression (5.3b) of C_2. Indeed for "small" k, T_3 and T_2 are of the same order of magnitude. Hence T_2 can no longer be neglected; but this could be managed. More serious is the fact that for "small" k the second term in (5.12) is $0(1)$ and should therefore be taken into account. This, in turn, will complicate the expression of T_3 and hence of C_2.

The interpretation of C_2 as an approximate cross-validation criterion may help in choosing the value of k and m to be used in a given application. For example, let N = 1000 and suppose we intend to use our model determined from the 1000 data points at hand for other (say) 9000 future time instants. Then we may take k = 10 and m = 100. For this choice the check sample length-to-estimation sample length ratio, (N-m)/m, takes the "desired" value 9000/1000.

We may also choose k and m so as to "minimize" the magnitude of the remainder term in (5.3a). For given N this is clearly achieved for

$$m > N^{2/3} \leftrightarrow k < N^{1/3} \qquad (5.14)$$

Further details on the choice of k and m can be found in the next section.

We now state the model selection rule based on C_2.

<u>Second cross-validation model structure selection rule</u>: Choose the model structure which leads to the smallest value of C_2, where C_2 is defined by (5.3b). ●

We may remark that a sufficient condition asymptotically guaranteeing that both C_2 and C_{II} are minimized by the same model structure is that for any two different structures in the set under consideration, say $\bar{M}$ and $\tilde{M}$, the differences $C_{2\bar{M}} - C_{2\tilde{M}}$ and $C_{II\bar{M}} - C_{II\tilde{M}}$ have for large N the same sign. Since C_2 is an asymptotically valid approximation of $\frac{1}{(k-1)N} C_{II}$, (5.3), the above condition appears to be fairly weak. For example, it certainly holds if the order of magnitude of $C_{2\bar{M}} - C_{2\tilde{M}}$ is greater than $O(1/\min(N,m^{3/2}))$, cf. (5.3a); and we may expect that, in general, $|C_{2\bar{M}} - C_{2\tilde{M}}| > O(\frac{1}{m})$.

6 ASYMPTOTIC EQUIVALENCE WITH SOME CONSISTENT STRUCTURE SELECTION CRITERIA

Let us assume that condition B_2 introduced in section 4 holds true. For an interpretation of B_2 see the discussion preceeding (4.14). Then, parallelling the calculations in (4.14)-(4.17) we can write

$$\frac{1}{(k-1)N} C_{II} = V(\hat{\theta}) + \frac{2mk^2}{N^2} \operatorname{tr}\{[2E\varepsilon_\theta(t,\theta^*)\varepsilon_\theta^T(t,\theta^*)]^{-1} + O(\frac{1}{\sqrt{N}})\}\cdot$$

$$\cdot\{V(\hat{\theta})E\varepsilon_\theta(t,\theta^*)\varepsilon_\theta^T(t,\theta^*) + O(\frac{1}{\sqrt{k}})\} + O(\frac{1}{\min(N,m^{3/2})}) =$$

$$= V(\hat{\theta})[1 + \frac{mk^2}{N^2} \dim\theta] + O(\frac{mk^{3/2}}{N^2}) + O(\frac{1}{\min(N,m^{3/2})}) =$$

$$= V(\hat{\theta})[1 + \frac{k}{N} \dim\theta] + O(\frac{1}{m.\min(k^{1/2},m^{1/2})}) \qquad (6.1)$$

which implies

$$\ln \frac{1}{(k-1)N} C_{II} = \mathrm{GAIC} + O(\frac{1}{m.\min(k^{1/2},m^{1/2})}) \qquad (6.2)$$

where

$$\text{GAIC} = \ln V(\hat{\theta}) + \frac{k_N}{N} \dim \theta \tag{6.3}$$

and where we stressed by notation the dependence of k on N. The conclusion is that under B2 the model selection rules based on C_{II} (or C_2) and GAIC (Generalized AIC) (6.3), will be asymptotically equivalent $\left[\textit{cf.}\right.$ also the discussion immediately following (4.19)$\left.\right]$.

This equivalence is interesting since in the last years there has been a considerable interest in model structure selection criteria of the form (6.3). Kashyap (1977,1982) and Schwarz (1978) have obtained such criteria with

$$k_N = \ln N \tag{6.4}$$

within a Bayesian context. Rissanen (1978) arrived at the same choice of k_N, (6.4), by using the "shortest data description" principle. Hannan (1980, 1981) has considered criteria of the form (6.3) with a general k_N (>0). Assuming that B2 holds and that

$$k_N \to \infty \quad , \quad \frac{k_N}{N} \to 0 \qquad \text{as } N \to \infty \tag{6.5}$$

Hannan proved that for ARMA models the structure minimizing GAIC is a consistent estimate of the true structure $M \ni S$ (in the sense that when N tends to infinity, the probability of selecting a wrong structure by minimizing GAIC goes to zero).

Remark 6.1: Note that in Theorem 5.1 the same condition (6.5) was imposed on k. In the following we shall assume that (6.5) holds true. ●

Hannan also considered the problem of choosing k_N so as to decrease the risk of underfitting (which is clearly more serious than overfitting). Then k_N should increase with N as slowly as possible. A smallest increasing rate that still preserves the consistency property was shown to be, Hannan (1980, 1981),

$$k_N = c \ln \ln N \qquad c > 2 \tag{6.6}$$

Consistency considerations for criteria of the form (6.3) can also be found in Kashyap (1977), Rissanen (1979, 1980), Anděl et al. (1981) etc.

What can be learned from the asymptotic equivalence between C_2 and GAIC shown above On the one hand, the fact that (under B2) our second selection rule asymptotically encompasses a well-established model structure testing procedure (designed to work under B2) should be viewed as a desirable feature of our proposal. On the other hand, the shown equivalence gives a nice cross-validation interpretation to the selection rule based on GAIC. This interpretation may give ideas for choosing k_N. It also suggests that the selection rule based on GAIC, which was mainly used in ARMA model identification, could be applied to other model structures as well. Under B2, the structure selected will be asymptotically optimal in the sense of minimizing the cross-validation criterion C_{II}. Furthermore, it appears that the consistency properties of the rule will also be preserved for more general model structures. As a matter of fact, we show below that a stronger consistency property than that usually stated seems to hold for the model structure estimated by minimizing GAIC. In the rest of this section we shall relax the assumption that S belongs to the considered model set.

Let $\theta^*_{\bar{M}}$ be the parameter vector of the model $\bar{M}(\theta_{\bar{M}})$ given by (4.1a). Let M be a model structure in the class of model structures under consideration, which is such that

$$E\,\varepsilon^2_M\,(t,\theta^*_M) \leq E\,\varepsilon^2_{\bar{M}}\,(t,\theta^*_{\bar{M}}) \qquad (6.7)$$

for any $\bar{M}$ in the class. Furthermore, let M be the "smallest" structure with the above property (i.e. if for some $\bar{M}$ we have equality in (6.7) then $\dim \theta_M < \dim \theta_{\bar{M}}$).

In the following we outline a proof of the fact that, under weak conditions on the class of structures in question (to be specified below), the structure minimizing GAIC asymptotically is M. [This outline may eventually constitute the basis for a more formal proof].

<u>Remark 6.2</u> Note that when the assumption B2 is in force, the above assertion states nothing more than the well-known consistency property of the selection rule based on GAIC. However, as already mentioned, we shall not use such an assumption. ●

First consider a model structure $\bar{M} \subset M$. We have

$$GAIC_{\bar{M}} - GAIC_{M} = \ln\left[V_{\bar{M}}(\hat{\theta}_{\bar{M}})/V_{M}(\hat{\theta}_{M})\right] + \frac{k_N}{N}\,(\dim\,\theta_{\bar{M}} - \dim\,\theta_{M}) \qquad (6.8)$$

Since $\bar{M}$ M the first term in (6.8) is positive

$$\ln\left[V_{\bar{M}}(\hat{\theta}_{\bar{M}})/V_{M}(\hat{\theta}_{M})\right] > 0$$

Moreover it must be of order $O(1)$. Then for N large enough so that the second term in (6.8) can be neglected, we have that

$$GAIC_{\bar{M}} > GAIC_{M} \qquad (6.9)$$

Consider now a model structure $\tilde{M} \supset M$. Since $M \subset \tilde{M}$ there exists a function, say $g(\cdot)$, such that $\tilde{M}\left(g(\theta_M)\right)$ reduces to $M(\theta_M)$ [we assume that $g(\cdot)$ is continuous]. This, in turn, implies that $\theta^*_{\tilde{M}} = g(\theta^*_M)$, under some weak assumptions. Indeed,

$$E\,\varepsilon^2_{\tilde{M}}\left(t, g(\theta^*_M)\right) = E\,\varepsilon^2_M(t,\theta^*_M) < \min_{\theta_{\tilde{M}} \in \Theta_{\tilde{M}}} E\,\varepsilon^2_{\tilde{M}}(t,\theta_{\tilde{M}}) \qquad (6.10)$$

where the inequality follows from (6.7). To conclude from (6.10) that $\theta^*_{\tilde{M}} = g(\theta^*_M)$ we need to assume that the asymptotic loss function $E\,\varepsilon^2_{\tilde{M}}(t,\theta_{\tilde{M}})$ associated with $\tilde{M}$ has a unique (global) minimum in $\Theta_{\tilde{M}}$. We shall make this assumption. Note that it is related to our basic assumption A2. Indeed, if A2 holds for large N, then $\theta^*_{\tilde{M}} = g(\theta^*_M)$ is an isolated (global) minimum and thus a unique minimum in an appropriately chosen (vicinity) set $\Theta_{\tilde{M}}$. We may remark that relaxation of the above assumption appears possible but that would make the analysis more technical (see, e.g., Rissanen (1979) for a discussion relevant to ARMA-models).

According to the above discussion $\hat{\theta}_{\tilde{M}}$ will, under the assumptions made, converge to $g(\theta^*_M)$, as N tends to infinity. Then it follows from (4.1) that for a sufficiently large N we have

$$g(\hat{\theta}_M) - \hat{\theta}_{\tilde{M}} = \left[g(\hat{\theta}_M) - g(\theta_M^*)\right] + \left[g(\theta_M^*) - \hat{\theta}_{\tilde{M}}\right] =$$

$$= 0\left(\frac{1}{\sqrt{N}}\right) \qquad (6.11)$$

Also, we can write

$$V_M(\hat{\theta}_M) = V_{\tilde{M}}\left(g(\hat{\theta}_M)\right) \approx V_{\tilde{M}}(\hat{\theta}_{\tilde{M}}) +$$

$$+ \frac{1}{2}\left[g(\hat{\theta}_M) - \hat{\theta}_{\tilde{M}}\right]^T V_{\theta\theta}^{\tilde{M}}(\hat{\theta}_{\tilde{M}})\left[g(\hat{\theta}_M) - \hat{\theta}_{\tilde{M}}\right] \qquad (6.12)$$

which together with (6.11) implies that

$$V_M(\hat{\theta}_M)/V_{\tilde{M}}(\hat{\theta}_{\tilde{M}}) = 1 + 0\left(\frac{1}{N}\right) \qquad (6.13)$$

Therefore we have

$$\mathrm{GAIC}_{\tilde{M}} - \mathrm{GAIC}_M = 0\left(\frac{1}{N}\right) + \frac{k_N}{N}\left[\dim \theta_{\tilde{M}} - \dim \theta_M\right] \qquad (6.14)$$

For N large enough the first term in (6.14) can be neglected and the second is positive, hence

$$\mathrm{GAIC}_{\tilde{M}} > \mathrm{GAIC}_M \qquad (6.15)$$

From (6.9) and (6.15) we conclude that *if* the class of model structures in question is such that for any $\bar{M}$ ($\tilde{M}$) with dim $\theta_{\bar{M}}$ < dim θ_M (dim θ_M < dim $\theta_{\tilde{M}}$) we have $\bar{M} \subset M$ ($M \supset \tilde{M}$) *then the model structure selected by minimizing GAIC will asymptotically be M*. Neither the structures $\bar{M}$'s nor $\tilde{M}$'s need to be nested. In the cases where we compare nested model structures (as often happens in order testing problems) then it readily follows from the above analysis that the curve GAIC is *unimodal*, at least for large N (this was empirically noticed by Stoica (1979)).

The above "consistency" property of the selection rule based on GAIC is appealing. However, since it refers to asymptotic models $M(\theta_M^*)$ we feel that for some applications the GAIC procedure may be less attractive than the structure selection rule based on C_2. Furthermore, as already explained, for the last rule the choice of k_N could be tailored to a given application and made on somewhat more precise grounds.

7 CONCLUDING REMARKS

The two cross-validation criteria C_1 and C_2 proposed in this paper are believed to be natural tools for selecting the model structure in those applications of system identification where the parameter estimation problem can be formulated as in (2.1). Under fairly general conditions they will select an optimal structure with respect to a cross-validatory assessment criterion. Furthermore, their cross-validation interpretation gives them an intuitive appeal and makes it possible to tailor them to specific applications by appropriately choosing the criterion parameter k(or m) that is at the user's disposal.

Numerical experience with the structure selection cross-validation criteria introduced here is reported in Van Beek (1985). It is shown there, by means of extensive Monte-Carlo simulations, that the finite sample behaviour of C_1 and C_2 is close to what is predicted by the asymptotic theory developed in this paper.

It is perhaps worth remarking that the cross-validatory assessment schemata used in this paper are only two of a quite large number of possible schemata. Other assessment schemata may exist, leading to model structure selection criteria with interesting features. We were, however, unable to find other "interesting" cross-validation criteria besides those presented.

To conclude, the cross-validatory assessment is an appealing device for model (structure) selection and we hope that this informal paper will stimulate the interest in investigating further possibilities for using this simple but useful concept in system identification and related fields.

REFERENCES

Ahmed, M.S., 1982, In: *Proc. 6th IFAC Symp. on Identification and System Parameter Estimation*, Washington D.C., U.S.A.

Akaike, H., 1969, *Ann. Inst. Statist. Math.*, vol. 21, 243; 1973, In: *Proc. 2nd Intern. Symp. on Information Theory*, edited by B.N. Petrov and F. Czáki (Budapest: Akademiai Kiadó); 1974, *IEEE Trans. Automat. Contr.*, vol. AC-19, 716; 1976, In *System Identification: Advances and Case Studies*, edited by R.K. Mehra and D.G. Lainiotis (New York: Academic Press); 1978, *Int. J. Control*, vol. 27, 323; 1979, *Biometrika*, vol. 66, 237; 1981, In *Trends and Progress in System Identification*, edited by P. Eykhoff (Oxford: Pergamon Press).

Anděl, J., 1982, *Math. Operationsforsch. Statist.*, ser. Statistics, vol. 13, 121.

Anděl, J., Perez, M.G. and Negrao, A.I., 1981, *Kybernetika*, vol. 17, 514.

Åström, K.J. and Eykhoff P., 1971, *Automatica*, vol. 7, 123.

Atkinson, A.C., 1980, *Biometrika*, vol. 67, 413.

Bartlett, M.S., 1966, *An Introduction to Stochastic Processes* (London: Cambridge University Press).

Bednar, J.B. and B.J. Roberts, 1982, In: *Proc. International Conf. ASSP*, 236, IEEE Press.

Beguin, J. M., Gourieroux C. and Monfort, A., 1980, In: *Time Series*, edited by O.D. Anderson (Amsterdam: North-Holland).

Bhansali, R.J. and Downhan, D.Y., 1977, *Biometrika*, vol. 64, 547.

Bohlin, T., 1978, *Automatica*, vol. 14, 137; 1982, *Model validation*. Report TRITA-REG-8203, Department of Automatic Control, The Royal Institute of Technology, Stockholm, Sweden. Also to appear in M. Singh (ed.): Encyclopedia of Systems and Control, Pergamon Press, 1984.

Bonivento, C. and Guidorzi R., 1970, *Linear system canonical models identification in the presence of noise*, Rapporto interno no. 9, Universita di Bologna, Italy.

Bora-Senta, E. and Kounias, S., 1980, In: *Analysing Time Series*, edited by O.D. Anderson (Amsterdam: North-Holland).

Box, G.E.P. and Pierce, D.A., 1970, *J. Am. Statist. Assoc.*, vol. 65, 1509.

Chan, C.W., Harris, C.J. and Wellstead, P.E., 1974, *Int. J. Control*, vol. 20, 817; 1975, in *Prep. 6th IFAC Congr*., paper 18.4, Boston, U.S.A.

Chow, J.C., 1972, *IEEE Trans. Automat. Contr.*, vol. AC-17, 386.

Davies, N., Triggs, C.M. and Newbold, P., 1977, *Biometrika*, vol. 64, 517.

Davies, N. and Newbold, P., 1979, *Biometrika*, vol. 66, 153.

Eykhoff, P., 1974, *System Identification: Parameter and State Estimation* (London: Wiley).

Fine, T.L. and Hwang, W.G., 1979, *IEEE Trans. Automat. Contr*., vol. AC-24, 387.

Godfrey, L.G., 1979, *Biometrika*, vol. 66, 67.

Goodwin, G.C. and Payne, R.L., 1977, *Dynamic System Identification: Experiment Design and Data Analysis* (New York: Academic Press).

Guidorzi, R.P., 1981, *Automatica*, vol. 17, 117.

Guidorzi, R.P., Losito, M.P. and Muratori, T., 1982, *IEEE Trans. Automat. Contr*., vol. AC-27, 1044.

Gupta, N.N., 1979, In: *Proc. International Conf. on Cybernetics and Society*, 725, IEEE Press.

Gustavsson, I., 1972, *Automatica*, vol. 8, 127.

Hajdasinski, A.K., 1980a, *Journal A*, vol. 21, 21; 1980b, *Linear Multivariable Systems: Preliminary Problems in Mathematical Description, Modelling and Identification*, TH-Report 80-E-106, Eindhoven University of Technology, Netherlands.

Hannan, E.J., 1980, *Ann. Statist.*, vol. 8, 1071; 1981, *J. Multivariate Anal.*, vol. 11, 459.

Hannan, E.J. and Quinn, B.G., 1979, *J. R. Statist. Soc. B*, vol. 41, 190.

Hannan, E.J. and Rissanen, J., 1982, *Biometrika*, vol. 69, 81.

Hashimoto, A., Honda, M., Inoue, T. and Taguri, M., 1981, *Rep. Stat. Appl. Res.*, JUSE, vol. 28, 57.

Hipel, K.W., 1981, *IEEE Trans. Automat. Contr.*, vol. AC-26, 358.

Hjorth, U., 1982, *Scand. J. Statist.*, Vol. 9, 95.

Hosking, J.R.M., 1979, *Biometrika*, vol. 66, 156.

Ishii, N. and Suzumura, N., 1977, *Int. J. Systems Sci.*, Vol. 8, 905.

Jategaonkar, R.V., Raol, J.R. and Balakrishna, S., 1982, *IEEE Trans. on Syst., Man and Cybernetics*, vol. SMC-12, 56.

Jones, R.H., 1974, *IEEE Trans. on Autom. Contr.*, vol. AC-19, 894.

Karny, M., 1980, *Problems of Control and Information Theory*, vol.9, 33.

Kashyap, R.L., 1977, *IEEE Trans. Automat. Contr.*, vol. AC-22, 715; 1978, *IEEE Trans. Inform. Theory*, vol. IT-24, 281; 1980, *IEEE Trans. Automat. Contr.*, vol. AC-25, 996; 1982, *IEEE Trans. Pattern Anal. and Machine Intelligence*, vol. PAMI-4, 99.

Kashyap, R.L. and Rao, A.R., 1976, *Dynamic Stochastic Models from Empirical Data* (New York: Academic Press).

Katz, R., 1981, *Technometrics*, vol. 23, 243.

Kaveh, M., 1979, In: *Proc. 17th IEEE Conf. on Decision and Control*, 949.

Kawashima, H., 1981, In: *Prepr. 8th IFAC Congr.*, paper 28.4, Kyoto, Japan.

Kozin, F. and Nakajima, F., 1980, *IEEE Trans. Automat. Control.*, vol. AC-25, 250.

Krolikowski, A., 1982, *Model Structure Selection in Linear System Identification - Survey of Methods with Emphasis on the Information Theory Approach*, EUT Report 82-E-126, Eindhoven University of Technology, Netherlands.

Läuter, H. and Miethe, N., 1979, *Math. Operationsforsch. Statist.*, Ser Statistics, Vol. 10, 395.

Lee, T.-S., 1981, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 503.

Ljung, L., 1982, *Model Validation*, Report LiTH-ISY-I-0534, Linköping University, Sweden.

Ljung, L. and Caines, P.C., 1979, *Stochastics*, vol. 3, 29.

Ljung, L. and van Overbeek, A.J.M., 1978, In: *Proc. 7th IFAC Congr.*, paper 45A.3, Helsinki, Finland.

Ljung, L. and Söderström, T., 1983, *Theory and Practice of Recursive Identification* (Cambridge: MIT Press).

Ljung, G.M. and Box, G.E.P., 1978, *Biometrika*, vol. 65, 297.

Maklad, M.S. and Nichols, S.T., 1980, *IEEE Trans. Syst., Man and Cybernetics.*, vol. SMC-10, 78.

Newbold, P., 1980, *Biometrika*, vol. 67, 463.

Parzen, E., 1974, *IEEE Trans. Automat. Control*, vol. AC-19, 723; 1977, In: *Multivariable Analysis-IV*, edited by P.R. Krishnaiah (North-Holland).

Picci, G., 1982, *Mathematical Programming Study*, vol. 18, 76.

Poskitt, D.S. and Tremayne, A.R., 1981, *Biometrika*, vol. 67, 359; 1981, *Ann. Statist.*, vol. 9, 947; 1982, *Ibid.*, vol. 10, 114.

Rissanen, J., 1976, In: *System Identification: Advances and Case Studies*, edited by R.K. Mehra and D.G. Lainiotis (New York: Academic Press); 1978, *Automatica*, vol. 14, 465; 1979, In: *Proc. Intern Symp. on Systems Optimization and Analysis*, edited by A. Bensoussan and J.L. Lions (Berlin: Springer Verlag); 1980, In *Analysis and Optimization of Stochastic Systems*, edited by O. Jacobs, M. Davis, M. Dempster, C. Harris and P. Parks (New York: Academic Press), 451; 1981, *Methods Oper. Res.*, vol. 44, 143; 1982, *Circuits, Systems and Signal Processing*, vol. 1, 395.

Sagara, S., Gotanda, H. and Wada, K., 1982, *Int. J. Control*, vol. 35, 637.

Sakai, H., 1981, *Int. J. Control*, vol. 33, 175.

Schwarz, G., 1978, *Ann. Statist.*, vol. 6, 461.

Shibata, R., 1976, *Biometrika*, vol. 63, 117; 1980, *Ann. Statist.*, vol. 8, 147; 1983, in *Time Series Analysis: Theory and Practice 4*, edited by O.D. Anderson, 237, North-Holland; 1984, *Biometrika*, Vol. 71, 43.

Söderström, T., 1975, *Automatica*, vol. 11, 537; 1977, *Int. J. Control*, vol. 26, 1 (also Report UPTEC 76 28R, Uppsala University 1976, Sweden); 1981, *Automatica*, vol. 17, 387; 1983, *Model Structure Determination*, Report UPTEC 83 11R, Uppsala University, Sweden.

Söderström, T. and Stoica, P., 1983, *Instrumental Variable Methods for System Identification*, (Berlin: Springer Verlag).

Stoica, P., 1977, *IEEE Trans. Automat. Control*, vol AC-22, 992; 1978, *Rev. Roum. Sci. Techn. Electrotech. et Energ.*, vol. 23, 267; 1979, *IEEE Trans. Automat. Control*, vol. AC-24, 516; 1981a, *Int. J. Control*, vol. 33, 1177; 1981b, *IEEE Trans. Automat. Contr.*, vol. AC-26, 572; 1983, *Int. J. Control*, vol. 37, 1159; 1984, *IEEE Trans. Automat. Control*, vol. AC-28, 379.

Stoica, P. and Söderström, T., 1982, *Int. J. Control*, vol. 36, 409.

Stone, C.J., 1982, *Ann. Inst. Statist. Math.*, vol. 34, A, 123.

Stone, M., 1974, *J.R. Statist. Soc. B.*, vol 36, 111; 1977a, *Ibid.*, vol. 39, 44; 1977b, *Biometrika*, vol. 64, 29; 1979, *J.R. Statist. Soc. B.*, vol. 41, 276.

Tong, H., 1978, *Int. J. Control*, vol. 27, 801; 1979, *Int. J. Control*, vol. 29, 441.

Torrez, W.C., 1983, In: *Time Series Analysis: Theory and Practice 4*, edited by O.D. Anderson, 245 (Amsterdam: North-Holland).

Tse, E. and Weinert, H.L., 1975, *IEEE Trans. Automat. Contr.*, vol. AC-20, 603.

Unbehauen, H. and Gohring, B., 1974, Automatica, vol. 10, 233.

Van Beek, W.J., 1985, Some aspects of system identification: (a) a comparison of order selection rules, (b) a multivariable system parameter estimation technique. Masters' Thesis, Dept. of Elect. Eng., Eindhoven University of Technology, Netherlands.

Van den Boom, A.J.W., 1982, System Identification: On the Variety and Coherence in Parameter and Order Estimation Methods. Doctoral Thesis, Eindhoven University of Technology, Netherlands.

Van den Boom, A.J.W. and Van den Enden A.W.M., 1974, Automatica, vol. 10, 245.

Van Eck, T., 1980, Unpublished manuscript.

Van Overbeek, A.J.M. and Ljung, L., 1982, Automatica, vol. 18, 529.

Wellstead, P.E., 1978, Automatica, vol. 14, 89.

Wellstead, P.E. and Rojas, R.A., 1982, Int. J. Control, vol. 35, 1013.

Wertz, V., Gevers, M. and Hannan, E.J., 1982, IEEE Trans Automat. Control, vol. AC-27, 1200.

Woodside, C.M., 1971, Automatica, vol. 7, 727.

Young, P., Jakeman, A. and McMurtrie, R., 1980, Automatica, vol. 16, 281.

Yuan, Z.D., 1982, In: Proc. 6th IFAC Symp. on Identification and System Parameter Estimation, Washington D.C., U.S.A.

Acknowledgements

To Mrs. B. Cornelissen who skilfully typed this paper from some illegible notes.

TABLE 1 Review of the literature. A modest attempt has been made to include the many existing model-structure selection methods ("Quot homines, tot sententiae" (Terentius)) in a relatively small number of classes.

Continued on following pages.

Reference	F/χ^2 tests (F.T.)	Residual checking tests (R.C.T.)	Rank tests (R.T.)	Pole-zero cancel. like test (P.Z.T.)	Akaike's criteria (A.C.)	General. Akaike criteria (G.A.C.)	Bayesian Criteria (B.C.)	Others (O.)	Numerical Applic. (N.A.)	Remarks
Ahmed (1982)			X						X	The structural indices of a multivariable linear system are estimated by using rank tests on an instrumental product-moment matrix.
Akaike (1969)					X					Introduces the FPE statistic
Akaike (1973)					X					Introduces the AIC statistic
Akaike (1974)					X			X	X	A brief review of the AIC statistic and of its predecesors. The practical utility of minimum AIC procedure is demonstrated by means of some numerical examples
Akaike (1976)					X				X	Use of the AIC and similar statistics for modelling multivariate time series is discussed in detail
Akaike (1978)	X				X					Comments on Söderström (1977) emphasizing that reducing the structure selection problem to discriminating between two nested structures is essential for some of Söderström's results to hold
Akaike (1979)					X	X	X		X	A Bayesian extension of the AIC statistic is discussed

Reference	F.T.	R.C.T.	R.T.	P.Z.T.	A.C.	G.A.C.	B.C.	O.	N.A.	Remarks
Akaike (1981)					X	X	X		X	The minimum AIC procedure and its relation with other model selection techniques are reviewed. A number of numerical examples are included
Anděl et al. (1981)						X			X	A generalized AIC method for consistent estimation of the dimension of a regression model is proposed and tested on simulated data. The consistency proof is particularly simple for the regression model
Anděl (1982)	X				X	X	X	X	X	Survey paper. It reviews the F/χ^2 test and the model selection rules of Akaike, Hannan & Quinn, Parzen, Rissanen and Schwarz
Atkinson (1980)	X				X	X	X		X	A brief review of the literature on generalized Akaike and Bayesian criteria, emphasizing the difficulty of choosing the "best" criterion
Bednar and Roberts (1982)					X			X	X	The estimation of ARMA orders is discussed. An order selection criterion is introduced as an attempt to obtain a computationally simple approximation of AIC. Even if this criterion does not appear to be an approximation of AIC in any well-defined sense, it has some intuitively appealing qualities and was shown to perform relatively well in a (limited) number of simulations

Reference	F.T.	R.C.T.	R.T.	P.Z.T.	A.C.	G.A.C.	B.C.	O.	N.A.	Remarks
Beguin et al (1980)	X		X							The rank properties of the covariance matrix of an ARMA process are used for estimating the ARMA orders. A significance test on the determinants of the ARMA covariance matrix is also briefly discussed. Some relation to the work by Stoica (1981a,b) should be noted
Bhansali and Downham (1977)					X	X			X	Some generalization of the FPE statistic is proposed and the asymptotic distribution of the selected order established. Consistency considerations are implicit in this study
Bohlin (1978)	X	X						X	X	A model structure selection method which does not require overfitting is proposed and its relation to other structure testing procedures discussed. Some relations to the work by Godfrey (1979) and Poskitt and Tremayne (1980) should also be noted
Bohlin (1982)	X	X			X					A general and tutorial discussion on model validation. Specific discussions of tests for the model in accordance with the experimental data and for the model serving its purpose in terms of predicting the output in applications

Reference	F.T.	R.C.T.	R.T.	P.Z.T.	A.C.	G.A.C.	B.C.	O.	N.A.	Remarks
Bonivento and Guidorzi (1970)			X						X	An early reference for structural identification of linear MIMO systems (with white additive noise) by using rank tests on compensated data product-moment matrices
Bora-Senta and Kounias (1980)	X	X			X	X			X	A Monte-Carlo analysis of 5 order selection rules is reported in case they are used to estimate the order of AR processes
Box and Pierce (1970)	X	X							X	Introduces the "portmanteau" statistic for checking the whiteness of the residual sequence of an estimated ARMA model
Chan et al (1974)	X				X	X			X	A criterion which resembles the FPE statistic is proposed for order estimation of ARMAX systems. Its interpretation as FPE type criterion may, however, be questioned since the calculations in the paper contain some flaws see, e.g. Söderström (1977). Nevertheless, the proposed criterion appears to compromise between small residual variance and accurate parameter estimates (like, for example, the criteria of Rissanen (1976), Maklad & Nichols (1980) etc.) and it was reported to behave well in a number of Monte-Carlo simulations. Comparisons with the F-test and Akaike's FPE criterion are included

Reference	F.T.	R.C.T.	R.T.	P.Z.T.	A.C.	G.A.C.	B.C.	O.	N.A.	Remarks
Chan et al (1975)	X				X				X	A discussion is given of some order selection procedures with emphasis on the F-test and Akaike's FPE criterion
Chow (1972)	X									A procedure for estimating the order of a MA process is proposed
Davies et al (1977)	X	X							X	A Monte-Carlo analysis of the "portmanteau" statistic (Box and Pierce (1970))
Davies and Newbold (1979)	X	X							X	A Monte-Carlo simulation analysis of the power of the modified "portmanteau" statistic (Ljung and Box, 1978) is reported
Fine and Hwang (1979)		X						X		A general theory for constructing consistent estimators of system order is developed. Details for MA as well as AR processes are provided. The results are primarily of theoretical interest
Godfrey (1979)	X	X							X	Introduces the Lagrange multiplier test and compares its performance with those of the "portmanteau" test in a number of Monte-Carlo simulations
Guidorzi (1981)		X	X						X	A rank test and a test based on residual variance (presented in a more elaborated form in Guidorzi et al.(1982)) are proposed for estimating the structural indices of MIMO linear systems

Reference	F.T.	R.C.T.	R.T.	P.Z.T.	A.C.	G.A.C.	B.C.	O.	N.A.	Remarks
Guidorzi et al. (1982)		X							X	A computationally efficient way of estimating the variance of LS model residuals is used as a basis for a practical procedure for selecting a "good" structure of a multivariable model
Gupta (1979)			X							An outline of an approach presented in full-length elsewhere. The orders of the system transfer function are determined by testing the rank of an instrumental-variable product moment matrix
Gustavsson (1972)	X	X	X	X					X	Several order selection procedures are critically reviewed
Hajdasinski (1980a)		X	X						X	Three tests for estimation of the "order" of a MIMO system from the sequence of (estimated) Markov parameters are presented
Hajdasinski (1980b)		X	X		X			X	X	Structural identification of MIMO systems is discussed in close connection with parametric identification. Seven structure selection procedures are reviewed. The emphasis is on the rank tests
Hannan and Quinn (1979)						X				Introduces the selection rule analysed in Hannan (1980) in a more general context
Hannan (1980)						X				Consistent selection rules of ARMA process orders are studied. Care is taken to decrease the risk of underfitting

Reference	F.T.	R.C.T.	R.T.	P.Z.T.	A.C.	G.A.C.	B.C.	O.	N.A.	Remarks
Hannan (1981)						X				Extends the results of Hannan (1980) to multivariable ARMA-processes
Hannan and Rissanen (1982)						X			X	A computationally efficient procedure for consistent estimation of ARMA orders is proposed
Hashimoto et al. (1981)					X	X		X	X	Monte-Carlo evaluation of 9 model structure selection criteria
Hipel (1981)		X			X				X	A survey of the applications of minimum AIC procedure, with particular emphasis on geophysical applications
Hjorth (1982)					X		X	X	X	Based on ideas of cross-validation, a model selection and validation procedure is proposed for linear regression analysis and time-series analysis. The procedure is based on weighted sums of recursive prediction errors. The results of the proposed method are compared with those of FPE and BIC for a specific multivariate time series example
Hosking (1979)		X								The asymptotic distribution of the squared multiple correlation coefficient, R^2, for ARMA processes is derived

Reference	F.T.	R.C.T.	R.T.	P.Z.T.	A.C.	G.A.C.	B.C.	O.	N.A.	Remarks
Ishii-Suzumura (1977)					X			X	X	A new order testing procedure for autoregressions is derived from information theoretical grounds (the concept of entroy). In some numerical examples it is superior to the FPE criterion
Jategaonkar et al. (1982)	X	X			X			X	X	Survey paper. Numerical comparison of 12 order testing criteria
Jones (1974)					X				X	Experience with using the minimum AIC procedure for autoregressive spectral estimation is reported
Karny (1980)							X			A general Bayesian approach to model order estimation is proposed. Bayesian estimation of the orders of a multivariable LS/ regression model is discussed in detail. In contrast to the related work by Kashyap (1977,1982) no attempt is made to derive a simplified (approximate) order selection rule. Nevertheless a computationally efficient algorithm is proposed for performing the Bayesian comparison of various LS models of different orders
Kashyap (1977)							X			Introduces the basic ideas used in a more general setting in Kashyap (1982). Consistency properties of the Bayesian selection rule are shown

Reference	F.T.	R.C.T.	R.T.	P.Z.T.	A.C.	G.A.C.	B.C.	O.	N.A.	Remarks
Kashyap (1978)							X			Same as above. Application to optimal feature selection is emphasized
Kashyap (1980)					X	X				Inconsistency of the AIC rule is shown. A quite related technique of proof is implicit in Söderström (1977)
Kashyap (1982)						X	X			An optimal selection rule which minimizes the average probability of error is derived. It asymptotically reduces to the consistent order selection rules of Rissanen and Schwarz
Katz (1981)					X	X				Use of Akaike's and Schwarz's criteria for estimating the order of a Markov chain is investigated
Kaveh (1979)					X	X			X	A generalized AIC-like criterion for selecting the order of an AR process is introduced on more or less heuristical grounds. It is compared in some numerical simulations with AIC. It is shown that, as expected, the risk of over-estimating the true order is smaller for GAIC but, on the other hand, GAIC may under-estimate the order more often than AIC

Reference	F.T.	R.C.T.	R.T.	P.Z.T.	A.C.	G.A.C.	B.C.	O.	N.A.	Remarks
Kawashima (1981)								X	X	A modification of a cross-validation criterion is used to obtain consistent estimates of the order of autoregressive integrated processes
Kozin and Nakajima (1980)					X					It is shown that the minimum AIC procedure is applicable to a class of time-varying AR processes
Królikowski (1982)	X	X	X		X	X		X		A comprehensive up-to-date review of model structure selection literature with emphasis on structure determination criteria based on information theory
Läuter-Miethe (1979)	X									Analysis of the likelihood ratio applied to ARMA processes and especially the asymptotic distribution of this test
Lee (1981)					X	X				Proposal of generalized criteria, based on Kozin and Nakajima (1980), and applicable for multivariate autoregressions observed with white measurement noise
Ljung (1982)	X	X			X	X		X		A brief review of the basic ideas and techniques for model selection and validation
Ljung and Box (1978)	X	X							X	A refined "portmanteau" statistic test is introduced and its power investigated

Reference	F.T.	R.C.T.	R.T.	P.Z.T.	A.C.	G.A.C.	B.C.	O.	N.A.	Remarks
Maklad and Nichols (1980)	X				X			X		A model structure selection rule compromising between whiteness of model residuals and accuracy of its estimated parameters is obtained from complexity theory arguments. Comparisons with the F-test and Akaike's criteria are included
Newbold (1980)	X	X								The Lagrange multiplier test of Godfrey (1979) and the usual test based on residual autocorrelations are shown to be equivalent in certain cases
Parzen (1974)					X			X		Determination of the order of autoregressive models which "best" approximate stationary time series is discussed and a criterion (CAT) to be used for this purpose is proposed
Parzen (1977)					X			X		Generalization of the CAT criterion to multivariate autoregressions
Picci (1982)			X		X			X		A tutorial paper on model parametrization and model structure selection in multivariable-system identification
Poskitt and Tremayne (1980)	X									Extension of the Lagrangian multiplier test of Godfrey (1979) to more general alternative hypotheses is discussed

Reference	F.T.	R.C.T.	R.T.	P.Z.T.	A.C.	G.A.C.	B.C.	O.	N.A.	Remarks
Poskitt and Tremayne (1981)	X	X								An approach to testing linear time series models by using χ^2 tests on certain quadratic forms is proposed
Poskitt and Tremayne (1982)	X	X								Diagnostic tests of Lagrange and "portmanteau" type for multivariable ARMA models are discussed
Rissanen (1976)						X		X		Information theory tools are used to obtain a "minmax entropy criterion" for estimating multivariable ARMA models. When used for structure selection this criterion may have some advantages over, albeit quite similar, AIC criterion (provided one avoids some scaling problems, see the paper for details)
Rissanen (1978)						X		X		A general model structure selection criterion is obtained from the shortest data description principle. The expression of this criterion is worked out in detail for ARMA models. The criterion belongs to the class of "generalized AIC", at least for sufficiently large sample lengths
Rissanen (1979)						X				Proves the consistency of the ARMA-order selection criterion introduced in Rissanen (1978)

Reference	F.T.	R.C.T.	R.T.	P.Z.T.	A.C.	G.A.C.	B.C.	O.	N.A.	Remarks
Rissanen (1980)						X			X	Paper related to Rissanen (1978 and 1979). The "shortest description length" criterion of Rissanen (1978) is revisited. The consistency of the order of an AR process estimated by using this criterion is proved
Rissanen (1981)						X				An attempt to extend the ARMA order and parameter estimation method of Hannan and Rissanen (1982) to transfer function models (including a control (or exogenous) input). The extended procedure proposed does not seem, however, to be correct since its first step which consists of estimating the system dynamics parameters will in general provide inconsistent estimates. To make the procedure valid we should use in the first step estimates of the output errors instead of the estimated innovations. The estimated output error could be computed by projecting the current output on the space spanned by the past inputs (only)
Rissanen (1982)						X		X		A new structure estimation criterion is derived from the "minimum description length" principle. It represents an improvement over the related "shortest description length" estimation criterion

Reference	F.T.	R.C.T.	R.T.	P.Z.T.	A.C.	G.A.C.	B.C.	O.	N.A.	Remarks
Rissanen (1982) cont.										proposed by the same author, Rissanen (1978). The new criterion includes an additional term which "measures" the (relative) size of the parameters in a model. It is this term which makes it possible to use the new criterion for distinguishing even between input-output equivalent models having the same number of parameters. In particular, it is shown that for linear vector models this additional term will asymptotically take large values whenever the parametrization under consideration is ill-conditioned. A relatively simple algorithm for choosing the "best" linear model for vector processes is proposed
Sagara et al (1982)		X	X						X	The order selection rules proposed are related to the instrumental variable estimation method and are implemented in a computationally efficient way
Sakai (1981)					X				X	The results of Shibata (1976) are extended to multivariable AR models
Schwarz (1978)						X	X			A consistent structure selection rule is derived within a Bayesian context
Shibata (1976)					X					The asymptotic distribution of the order of an autoregression, selected by using the minimum AIC procedure is established

Reference	F.T.	R.C.T.	R.T.	P.Z.T.	A.C.	G.A.C.	B.C.	O.	N.A.	Remarks
Shibata (1980)					X					A certain asymptotic optimality of the AR models selected by using Akaike's criteria is shown
Shibata (1983)					X	X			X	The controversial problem of choosing the factor α weighting the model dimension in a generalized AIC is discussed in the context of AR order estimation. It is shown that the choice α=2 corresponding to the AIC procedure is asymptotically optimal (in the sense of minimizing the one-step prediction error) provided the unknown optimal order is "large enough" (see also Shibata (1980)). When this is not the case, it is shown that neither AIC nor GAIC with a sequence α increasing without bound as the number of data points tends to infinity are necessarily optimal in terms of the mean squared one-step prediction error; moreover a (numerical) minimax procedure for choosing α is outlined
Shibata (1984)					X	X			X	An approximation is given for the mean-squared (prediction)-error in the linear regression case, if the model order is obtained by a generalized final prediction error method FPE . Based on this, an expression is given for the approximate

Reference	F.T.	R.C.T.	R.T.	P.Z.T.	A.C.	G.A.C.	B.C.	O.	N.A.	Remarks
Shibata (1984) cont.										efficiency of the method. This expression is used for proposing a procedure for the choice of the weighting factor α. Also a further generalization of the $(FPE)_\alpha$-criterion is proposed
Söderström (1975)				X					X	A systematic procedure for performing pole-zero cancellation tests is proposed
Söderström (1977)	X	X	X	X	X			X	X	Survey paper containing an analytical comparison of some methods commonly used for model structure selection. Some interesting (asymptotic) equivalences are established. The residual checking and the rank tests are only briefly covered
Söderström (1981)	X									Comments on Bohlin (1978). It is shown that Bohlin's order selection rule can be interpreted as a modified F-test
Söderström (1983)	X	X	X	X	X	X				A tutorial paper briefly dealing with most of the model structure selection procedures currently in use
Stoica (1977)	X	X							X	Some whiteness tests are compared
Stoica (1978)	X								X	A simple way for computing the significance levels needed by the order selection procedure of Chow (1972) is proposed

Reference	F.T.	R.C.T.	R.T.	P.Z.T.	A.C.	G.A.C.	B.C.	O.	N.A.	Remarks
Stoica (1979)	X	X			X		X		X	Use of 5 structure testing procedures for selecting the order of an AR process is discussed and numerically illustrated
Stoica (1981a)			X						X	A significance test on the determinant of a covariance matrix is introduced and an application to testing the orders of ARMA models is presented
Stoica (1981b)			X							Theoretical justification of a rank test used for estimating the orders of ARMA processes is presented
Stoica (1983)			X							The rank properties of the covariance matrix of a multivariable ARMA process are investigated and some implications for testing the orders of time-series models are discussed
Stoica (1984)		X								Comments on Guidorzi et al (1982). It is shown that the Guidorzi's structural identification procedure should be used only for least-squares models
Stone (1977a)					X			X		Asymptotic equivalence between the AIC and a cross-validation criterion is shown

Reference	F.T.	R.C.T.	R.T.	P.Z.T.	A.C.	G.A.C.	B.C.	O.	N.A.	Remarks
Stone (1977b)								X		The asymptotic properties (consistency and efficiency) of the "one-item-out" cross-validatory assessment scheme of Stone (1974) are analysed mainly in the context of some particular applications
Stone (1979)					X	X				A subtle theoretical comparison of Akaike's and Schwarz's criteria. The idea of increasing the number of model parameters as the data number tends to infinity is briefly discussed
Stone (1982)					X	X		X		A general model selection rule of the form min{-log(max likelihood) + (complexity)} is considered
Tong (1978)					X			X		The asymptotic distribution of the estimated coefficients of AR models when the order is also estimated (for example, by using AIC) is derived
Tong (1979)					X			X	X	Theoretical and numerical comparisons between the locally equivalent criteria AIC and Parzen's CAT
Torrez (1983)		X			X				X	The problem of estimating the order of a noisy narrowband autoregressive process is discussed. A procedure essentially based on testing the decrease of residual variance

Reference	F.T.	R.C.T.	R.T.	P.Z.T.	A.C.	G.A.C.	B.C.	O.	N.A.	Remarks
Torrez (1983) cont.										as the model order is increased, is numerically investigated. Alternative procedures based on Akaike's AIC and FPE criteria are briefly discussed.
Tse and Weinert (1975)			X						X	Rank tests on the output covariance matrix are used to select the structural indices of a multivariable linear system
Unbehauen and Gohring (1974)	X	X	X	X					X	Survey paper. Numerical comparisons of 7 order testing procedures
Van den Boom and van den Enden (1974)	X	X	X	X					X	Survey paper. Critical evaluation of 5 order testing criteria
Van den Boom (1982)	X	X	X	X	X	X		X	X	Contains a detailed discussion of many procedures for model order selection. AIC and generalized AIC procedures are only briefly reviewed
Van Eck (1980)					X	X		X	X	Numerical evaluation of some typical order determination methods
Van Overbeek and Ljung (1982)			X					X	X	Use of overlapping model structures to represent a multivariable state-space system of given order is emphasized. A solution to the problem of choosing a well-conditioned model structure is proposed

Reference	F.T.	R.C.T.	R.T.	P.Z.T.	A.C.	G.A.C.	B.C.	O.	N.A.	Remarks
Wellstead (1978)			X						X	An instrumental product moment test for model order estimation is described
Wellstead and Rojas (1982)			X						X	The model order testing procedure of Wellstead (1978) is extended to cover more general model structures and a computationally efficient implementation scheme (in the spirit of Sagara et al (1982)) is briefly discussed
Wertz et al (1982)			X					X	X	Several procedures for selecting a "best" structure of a multivariable linear model are compared
Woodside (1971)		X	X					X	X	An earlier basic reference for rank tests
Young et al (1980)		X	X		X			X	X	Some refinements of the instrumental product moment test for order selection are presented together with a comprehensive set of applications
Yuan (1982)	X									Test of hypothesis on the coefficients of dynamic systems are discussed with model structure determination as an example