Finite Sample Criteria for Autoregressive Order Selection

Piet M. T. Broersen

Abstract—The quality of selected AR models depends on the true process in the finite sample practice, on the number of observations, on the estimation algorithm, and on the order selection criterion. Samples are considered to be finite if the maximum candidate model order for selection is greater than N/10, where N denotes the number of observations. Finite sample formulae give empirical approximations for the statistical average of the residual energy and of the squared error of prediction for several autoregressive estimation algorithms. This leads to finite sample criteria for order selection that depend on the estimation method. The special finite sample information criterion (FSIC) and combined information criterion (CIC) are necessary because of the increase of the variance of the residual energy for higher model orders that has not been accounted for in other criteria. Only the expectation of the logarithm of the residual energy, as a function of the model order, has been the basis for the previous classes of asymptotical and finite sample criteria. However, the behavior of the variance causes an undesirable tendency to select very high model orders without the special precautions of FSIC or CIC.

Index Terms—Model quality, parameter estimation, spectral estimation, system identification, time series.

I. INTRODUCTION

C EVERAL autoregressive (AR) estimation algorithms have been developed [1], [2]. The asymptotical theory is the same for all these different estimation algorithms. In practice, selection criteria have to be used to select an order p for the best AR(p) model for given data. Asymptotical criteria are a first category for order selection. They are derived from asymptotical expressions for the residual variance and the prediction error or for the variance of estimated parameters. AIC [3] belongs to this class as do several consistent criteria like BIC [4] and similar criteria [5], [6]. The statistics of selected orders and the probability of overfit for asymptotic criteria like AIC have been evaluated theoretically [7]. Application of those results to the penalty function in selection criteria gives the conclusion that the penalty 2 of AIC reduces the probability of underfit at the cost of overfit [8]. On the other hand, consistent criteria may produce infinite costs of underfit when the sample size increases. A compromise between overfit and underfit is found by taking a constant penalty factor 3 [8].

Samples of autoregressive observations are called finite if the model orders concerned are not small in comparison with the sample size N, say, for orders greater than 0.1N. This applies

The author is with the Department of Applied Physics, Delft University of Technology, Delft, The Netherlands (e-mail: broersen@tn.tudelft.nl).

Publisher Item Identifier S 1053-587X(00)10160-6.

to the true order and to orders that are considered to be candidates for order selection. Simulations for finite samples show that estimated parameters, reflection coefficients, and residuals depend on the AR estimation method [9]. These differences are not accounted for in the asymptotical theory; therefore, a finite sample description is necessary. Finite sample formulae are empirical formulae that describe the variances of estimated parameters or reflection coefficients as well as the average variances of residuals or of predictions for different autoregressive estimation algorithms [9]. They also define a second category of order selection criteria for finite samples that are adapted to the method of estimation. An interesting alternative possibility in this second category is the predictive least squares (PLS) criterion [10]. This criterion computes true or "honest" predictions from the data themselves, and hence, it is adapted to all peculiarities of estimation methods and to the actual probability distribution of the data. A lattice implementation of the PLS criterion separates the data in a variable part for parameter estimation and the rest to compute the true predictions for order selection [11]. The selection with PLS is based on the assumption that the true or best order is a constant and independent of the sample size used for estimation. Unfortunately, the best AR order for estimated models of many processes depends on the sample size. Hence, the best order for "honest" predictions depends on the variable number of observations that is used to estimate the parameters in PLS algorithms and is not always a constant. This limits the processes for which the PLS criterion can be used. Moreover, separating data in a group for estimation and a group for determining the best order from "honest" predictions reduces the estimation accuracy. This is not a good idea in finite samples, certainly not if the maximum candidate model order is as high as 0.5N. In general, consistent criteria and PLS perform well in examples with a finite number of very significant true parameters and enough samples available for estimation.

Recently, a criterion of a third category has been defined as the finite sample information criterion (FSIC) [12]. FSIC, if applied to AR estimates with Burg's method [13], is very similar to AIC_C, which is a corrected version of AIC [14]; remarkably, that criterion AIC_C has been derived with asymptotical theory. The performance of FSIC remains good for all estimation methods, but AIC_C loses its good properties when applied to other than Burg results [12]. The penalty factor that is applied to additional parameters in a model is no constant for FSIC, like in the first two categories, but the penalty increases for higher model orders. Using FSIC, the results of order selection turn out to become practically independent of the choice of the maximum candidate order L, even if that is chosen very high. This

Manuscript received December 10, 1998; revised August 22, 1999. The associate editor coordinating the review of this paper and approving it for publication was Prof José R. Casar.

3551

is realized without hampering the ability to select those high model orders if they are really significant.

In the paper, some finite sample definitions are repeated. A new formula is given for the variance of the logarithm of the residual energy as a function of the model order. The increase of that variance for higher orders gives an explanation for the performance of FSIC. Another criterion for the third category [the Combined Information Criterion (CIC)] has been introduced to combine the favorable asymptotical penalty factor 3 with the increased resistance against the selection of too high model orders in finite samples [15]. Simulations show that the expected performance of FSIC and CIC is realized for different AR estimation methods and for various sample sizes.

II. FINITE SAMPLE FORMULAE

An AR(K) process and its AR(p) model are given by

$$y_n + a_1 y_{n-1} + \dots + a_K y_{n-K} = \varepsilon_n \tag{1}$$

$$y_n + \hat{a}_1 y_{n-1} + \dots + \hat{a}_p y_{n-p} = \hat{\varepsilon}_n \tag{2}$$

where ε_n is a stationary purely random process with zero mean and variance σ_{ε}^2 . In the Yule–Walker [1] and Burg [13] estimation methods, the last parameter \hat{a}_p in a model of order p is also known as the reflection coefficient k_p . The Levinson–Durbin algorithm [1] describes the relation between parameters and reflection coefficients. The building blocks of the AR finite sample formulae are finite sample variance coefficients $v(i, \cdot)$; *i* denotes the order and " \cdot " the method of estimation. The $v(i, \cdot)$ s have been introduced as simple expressions that approximate the empirical variance of *reflection* coefficients for the Yule-Walker and Burg method for orders i > K, which is the true process order in (1) [9]. Likewise, the $v(i, \cdot)$ are approximations for the empirical variance of the last *parameter* \hat{a}_i for least squares AR(i) models. The asymptotical expression for the variance of reflection coefficients (or of the last parameter) equals 1/N for all model orders greater than the true order K and for all estimation methods. Formulae that give approximations for the empirical variance for order i from N observations are given by

$$\begin{aligned} v(i, \text{YW}) &= (N-i)/N(N+2) & \text{Yule-Walker} \\ v(i, \text{Burg}) &= 1/(N+1-i) & \text{Burg} \\ v(i, \text{LSFB}) &= 1/(N+1.5-1.5i) & \text{two-sided LS} \\ v(i, \text{LSF}) &= 1/(N+2-2i) & \text{one-sided LS} \end{aligned}$$

for four well-known AR algorithms [9]. Further, $v(0, \cdot) = 1/N$ if the mean is subtracted and 0 otherwise. Those approximations are quite accurate for AR(K) processes for i > K, therefore, for orders that are important for order selection. The formula for Yule–Walker (also denoted autocorrelation method) is known from the statistical literature; in Burg and LSF (or covariance method), the decrease of the degrees of freedom that are available for estimation can be recognized.

The variance of $\hat{\varepsilon}_n$ in (2), $\sigma_{\hat{\varepsilon}}^2$ is called the *residual variance* RES(*p*) if the data y_n in the equation have been used to estimate the parameters. Otherwise, if the estimated parameters \hat{a}_i and the realization y_n of the process are independent, $\sigma_{\hat{\varepsilon}}^2$ is defined to be the *prediction error* PE(*p*). For AR(*p*) models estimated

with the Yule–Walker or Burg method, some theoretical results have been reported for the statistical expectation of the residual variance RES(p) and of the squared prediction errors PE(p) [9]. Those expectations, for model orders p equal to the true process order K or higher, depend on the estimation method and are given by

$$E\{\operatorname{RES}(p)\} \approx \sigma_{\varepsilon}^{2} \prod_{i=0}^{p} \{1 - v(i, \cdot)\}$$

$$p \ge K. \quad (4)$$

$$E\{\operatorname{PE}(p)\} \approx \sigma_{\varepsilon}^{2} \prod_{i=0}^{p} \{1 + v(i, \cdot)\}$$

Many simulations have been carried out on finite sample properties [8], [9], [12]. It turned out that the expectations in (4) are close approximations to the average of RES(p)and PE(p) in simulations with all four estimation methods if the appropriate $v(i, \cdot)$ are substituted, as defined in (3). Therefore, (4) can also be used for the least squares methods, although the derivation is based on reflection coefficients that belong only to the Yule-Walker and Burg methods. When evaluated for different sample sizes N, the outcomes of (4) are approximately nonlinear functions of the ratio p/N. This ratio appears linearly in the well-known asymptotical results, which include $E\{\text{RES}(p)\} = \sigma_{\varepsilon}^2(1 - p/N)$ for the residual variance and $E\{PE(p)\} = \sigma_{\varepsilon}^2(1 + p/N)$ for the squared prediction error. Those asymptotical relations can be retrieved from (4) by substituting 1/N for all coefficients $v(i, \cdot)$ and then deleting the terms of order $1/N^2$ and higher in the product.

Many asymptotical order selection criteria can be written as a single generalized information criterion (GIC) [9] with different values for the penalty factor α :

$$GIC(p, \alpha) = \ln\{RES(p)\} + \alpha p/N.$$
 (5)

To select a model order, $\operatorname{GIC}(p, \alpha)$ is determined for all orders p between 0 and some maximum candidate order L. The order with the minimum value of the criterion is selected. The penalty factor α is 2 for AIC [3], $\ln N$ for consistent criteria BIC [4], [5], $2\ln \ln N$ for the minimally consistent criterion [6], or some other constant. The asymptotical criterion $\operatorname{GIC}(p, \alpha)$ is identical for all AR estimation methods, whereas (4) shows that the behavior of $\operatorname{RES}(p)$ actually is a function of p that depends on the AR estimation method. Therefore, finite sample information criteria $\operatorname{FIC}(p, \alpha)$ have been defined as improvements for the asymptotical criteria $\operatorname{GIC}(p, \alpha)$ as [9]:

$$FIC(p, \alpha) = \ln\{RES(p)\} + \alpha \sum_{i=0}^{p} v(i, \cdot)$$
(6)

with the method dependent sum of $v(i, \cdot)$ instead of p/N in the finite sample criteria. The penalty factor 2 of AIC reduces the probability of underfit at the cost of overfit [8]. Consistent criteria, where the penalty factor depends on N, may produce infinite costs of underfit for increasing sample size. A compromise between overfit and underfit is found with the constant 3 for penalty factor α , both in GIC (p, α) and in FIC (p, α) .

The quotient of the expectations PE(p)/RES(p) in (4) has been used to derive the finite sample information criterion (FSIC) for order selection [12]. It is a finite sample approximation to the Kullback–Leibler discrepancy:

$$FSIC(p) = \ln\{RES(p)\} + \prod_{i=0}^{p} \frac{1 + v(i, \cdot)}{1 - v(i, \cdot)} - 1.$$
(7)

A comparison with (6) reveals that a product containing all previous $v(i, \cdot)$ appears in FSIC(p) instead of a constant penalty factor α times $v(p, \cdot)$ for order p in FIC(p, α). The performance of FSIC(p) is similar to AIC if it is applied to the Yule–Walker estimation method [12] because the product in (7) is very close to 1 + 2p/N for all model orders p if v(i, YW) is substituted. Asymptotical theory has been used to develop the AIC_C criterion, which is defined by replacing the term $\alpha p/N$ in (5) by 2p/(N - p - 1) [14]. The criterion AIC_C is almost identical to FSIC(p) for the Burg method with v(i, Burg) from (3) substituted in (7); for AR estimation methods other than Burg's, FSIC(p) and AIC_C are completely different.

The performance of the criteria can be explained by comparing the corrections to $\ln\{\text{RES}(p)\}$ in the different criteria, as shown in Fig. 1. GIC criteria have a constant slope, which is 2 for AIC and $\ln(100)$ or 4.61 for the consistent asymptotical criterion. GIC(p, 2), FIC(p, 2), and FSIC(p) are almost equal for p/N < 0.1, but the difference increases for greater p. FSIC(p)increases especially strongly there.

For new data with unknown characteristics, the maximum candidate order L for selection of the AR model order must be chosen high enough, certainly higher than the unknown best model order. Taking a high L reduces the possibility of missing the best order. In practice or in simulations, however, selection results for all criteria of the first two categories, including the consistent criteria, are more or less dependent on a proper choice for L. The selected order turns out to depend on the highest candidate order L, especially if that is chosen greater than, say, N/4. If L is taken high, orders close to L are often selected in simulations [9]. FSIC(p) is the only criterion that is not sensitive for the choice of the highest candidate order for selection, independent of the AR estimation method. An exception should be made for the Yule-Walker method, where the influence of L is moderate; this is sufficiently explained by the similarity between the asymptotical criterion AIC and FSIC(p)for Yule-Walker. High AR orders may not always be necessary in AR estimation, but they are certainly required when the AR model is used as an intermediate stage for MA or ARMA modeling [16], [17]. The influence of L is very important for all asymptotical criteria, remains a nuisance for the finite sample criteria, and disappears for FSIC(p) in the third category.

III. VARIANCE OF SELECTION CRITERIA

A fair agreement has been shown between finite sample formulae on the one hand and simulation variances of estimated parameters, reflection coefficients, residuals, and of prediction errors on the other hand [9]. This paper presents a recursive finite sample formula for the variance of $\ln\{\text{RES}(p)\}$ as a function of the model order p. That variance is also the variance of the criteria $\text{GIC}(p, \alpha)$, $\text{FIC}(p, \alpha)$ and of FSIC(p) because $\ln\{\text{RES}(p)\}$ is the only stochastic part of these order selection criteria. For convenience, the analysis is carried out for

Fig. 1. Comparison of order selection criteria. Total penalty of the right-hand side terms of the criteria (5)–(7) is shown for different values of the penalty factor as a function of the relative model order.

white noise processes as data. The asymptotical expression for $var[ln{RES}(p)]$ equals the constant 2/N for all orders p when the AR(p) models are estimated from N observations of a white noise process. A recursive formula for the residual variance for Yule–Walker and Burg estimates of AR processes is given by

$$\operatorname{RES}(p) = \operatorname{RES}(p-1)(1-k_p^2).$$
 (8)

As all reflection coefficients k_i are uncorrelated for the white noise process that is considered here, the covariance between RES(p-1) and k_p equals zero. An elementary result for uncorrelated variables a and b is easily derived as

$$\operatorname{var}\{ab\} = E^2(a)\operatorname{var}(b) + \operatorname{var}(a)E^2(b) + \operatorname{var}(a)\operatorname{var}(b).$$

Applying this to (8) gives

$$\operatorname{RES}(p) = E^{2} \{ \operatorname{RES}(p-1) \} \operatorname{var} \{ k_{p}^{2} \} \\ + E^{2} \{ 1 - k_{p}^{2} \} \operatorname{var} \{ \operatorname{RES}(p-1) \} \\ + \operatorname{var} \{ 1 - k_{p}^{2} \} \operatorname{var} \{ \operatorname{RES}(p-1) \}$$
(9)

which can still be simplified. By using the normal distribution to approximate the unknown finite sample distribution for the k_i for i > K, it follows from the theory and from the definition of the finite sample coefficients that

$$E\{k_i^2\} \approx v(i, \cdot), \operatorname{var}\{k_i^2\} \approx 2v(i, \cdot)^2.$$
(10)

Substitution of those approximations in (9) produces a recursive formula

$$\begin{aligned} \operatorname{RES}(p) &\gtrsim 2E^2 \{ \operatorname{RES}(p-1) \} v(i, \cdot)^2 \\ &+ [1 - 2v(i, \cdot) + 3v(i, \cdot)^2] \operatorname{var} \{ \operatorname{RES}(p-1) \}. \end{aligned}$$



As a starting value, the asymptotical expression for the variance of the estimated white noise variance RES(0) will be used, which is given by var{RES(0)} = $2\sigma_{\varepsilon}^4/N$. In Taylor expansions, higher powers of 1/N are neglected. Finite sample approximations are completely different because products of $v(i, \cdot)$ contain higher powers of 1/N, which are not neglected. Nevertheless, an approximate finite sample formula for the variance of the logarithm of RES(p) will be derived here with a first-order Taylor expansion as

$$\operatorname{var}[\ln\{\operatorname{RES}(p)\}] \approx \frac{1}{E^2\{\operatorname{RES}(p)\}} \operatorname{var}[\operatorname{RES}(p)\}].$$
(12)

Simulations have to show whether the approximations in (10) and (12) are a useful description of this variance in practice because their accuracy cannot be established with theoretical arguments. The distribution of k_i is unknown, and only one Taylor expansion term is used for (12). Moreover, the derivation starts with the use of reflection coefficients in (8) that are only applicable in the Yule–Walker and Burg estimation methods. Nevertheless, the result in (12) will also be tested as an approximation for the one- and two-sided least squares estimation methods.

The accuracy of the variance formula (12) has been studied in simulations for white Gaussian noise and with different AR(K)processes for several values of the sample size N. Fig. 2 gives the accuracy for 100 observations of an AR(10) process with as reflection coefficients $k_i = (-0.6)^i$, $i = 1, \dots, 10$ for the four estimation methods for which empirical variance coefficients have been defined in (3). The asymptotical theory for a white noise process yields the constant 2/N for all model orders. Clearly, the empirical variance in the simulations depends on the model order and on the estimation method. Fig. 2 shows that the white noise approximation with (12) is quite good for the Burg and Yule-Walker results for the orders above 4. This order 4 is the best order to be selected for estimated models from 100 observations of this AR(10) process. The agreement is less perfect for var $[\ln{\{\text{RES}(p)\}}]$ of the least-squares estimation methods LSFB and LSF, but (12) is still a much better approximation than the constant value 0.02 that would follow from the asymptotical theory for all estimation methods and all model orders.

As always in finite samples, the behavior of the Yule-Walker method differs quite from the other methods because it is the only method with $v(i, \cdot)$ less than 1/N. The figures show that the variance of $\ln\{\text{RES}(p)\}$ for the other methods can become more than two times greater than the variance at low models orders. As a consequence, $\ln{\{\text{RES}(p)\}}$ can become much smaller than its expectation in a single simulation run at some high model order p, which in turn produces a very low value of the selection criteria for that order p in that specific realization of the process. This increased variance creates the possibility for the undesirable selection of very high orders because the order selection criteria in the categories one and two have been based only on the average or expectation of $\ln\{\text{RES}(p)\}$ and not on its variance. Only for the Yule-Walker method was this not a serious problem because the increase of the variance at higher model orders is moderate.



IV. COMBINED INFORMATION CRITERION

With asymptotical arguments, a good balance between the possibilities of overfit and underfit requires a constant penalty factor of about 3 in a selection criterion [8]. The same conclusion for the penalty has been found with simulations for finite samples, where the maximum candidate order has been restricted to N/3 [8]. On the other hand, the criterion FSIC(p), as defined in (7), has a desirable resistance against the selection of very high orders because it has an increasing penalty at those higher orders. In agreement with the known theoretical arguments, no examples with a poor performance of FSIC(p) have been found when very high maximum candidate orders are allowed for selection. The asymptotical equivalent of FSIC(p), which describes its behavior at low model orders, is found by substituting 1/N for all coefficients $v(i, \cdot)$ and by removing terms with powers of $1/N^2$ and higher from the result. This leads to the well-known criterion AIC(p) [3], which, here, is also denoted GIC(p, 2) with penalty 2.

Unfortunately, the properties of an optimal asymptotical balance between under- and overfit and a good resistance against selection of high orders are not found in one criterion. It would be interesting if a single criterion had the penalty 3 at low orders and the behavior of FSIC(p) at high orders because that criterion can have favorable properties for all possible choices for the maximum candidate model order for all sample sizes. Therefore, a new criterion is introduced here that combines those properties: good asymptotic performance with penalty 3 and good finite sample performance with the protection of FSIC(p)



against wrongly selecting too high orders. It is the combined information criterion [15]:

$$\operatorname{CIC}(p) = \ln\{\operatorname{RES}(p)\} + \max\left[\prod_{i=0}^{p} \frac{1+v(i,\cdot)}{1-v(i,\cdot)} - 1, \ 3\sum_{i=0}^{p} v(i,\cdot)\right].$$
(13)

The behavior of $\operatorname{CIC}(p)$ can be derived from Fig. 1. For lower model orders, p/N < 0.3, $\operatorname{CIC}(p)$ is equal to $\operatorname{FIC}(p, 3)$ because the sum with penalty 3 is the largest. For higher orders, the total penalty of $\operatorname{FSIC}(p)$ is greater, and $\operatorname{CIC}(p)$ becomes equal to $\operatorname{FSIC}(p)$, thus combining the asymptotical balance between under and overfit and the finite sample resistance against the selection of model orders that are too high. Asymptotically, or for p < 0.1N, $\operatorname{CIC}(p)$ and $\operatorname{FIC}(p, 3)$ are equal and almost the same as $\operatorname{GIC}(p, 3)$ because the average of finite sample results differs only significantly from the asymptotical theory if the ratio p/Nis greater than 0.1.

V. COMPARISON OF ESTIMATION METHODS

This paper considers models of stationary stochastic processes in their capacity to predict future responses or as spectral estimators, demanding a selection criterion that is representative for the overall fit. The true process is known in simulations and can be used to define the quality of estimated models. A measure for the fit is the model error ME, which is defined as a simple function of the expectation of the prediction error [18]

$$ME(p) = N \left[\frac{PE(p)}{\sigma_{\varepsilon}^{2}} - 1 \right] = N \frac{\sigma_{\varepsilon}^{2} - \sigma_{\varepsilon}^{2}}{\sigma_{\varepsilon}^{2}}$$
$$= N \frac{\underline{\hat{a}}_{p}^{T}R(p) \underline{\hat{a}}_{p} - \sigma_{\varepsilon}^{2}}{\sigma_{\varepsilon}^{2}}.$$
(14)

Pre- and post-multiplying the estimated parameter vector $\underline{\hat{a}}_p$ with the true R(p) yields an accurate expectation for the squared error of prediction PE(p) that can easily be computed. R(p) is a $(p+1) \times (p+1)$ submatrix of the infinite-dimensional true covariance matrix of a known AR process, and $\underline{\hat{a}}_p$ is the estimated parameter vector of length p+1 in (2), including the constant 1 for the first parameter. Independent $\underline{\hat{a}}_p$ and y_n in (2) define the prediction error PE(p) as $\sigma_{\hat{z}}^2$, which has the asymptotical expectation $\sigma_{\varepsilon}^2(1+p/N)$ for p equal to or greater than the true process order if the parameters had been estimated from N observations. Hence, the asymptotical expectation of ME(p) equals p, independent of the sample size N. This value p can also be considered as an asymptotical Cramér-Rao lower bound for the achievable accuracy of the relative integrated spectrum of the AR(p) model; see [19]. Finite sample expectations for ME(p)can be computed with (3) and (4). The model error ME(p) in (14) has advantages above the prediction error PE(p) in studying the influence of the sample size on the quality of estimated and selected models because results for quality are made less dependent on N itself and are of a comparable magnitude.

The results of AR modeling are known to depend on the true process, the sample size, the estimation method, the order selection criterion, and the highest candidate order for selection. The first two influences are unknown and determined by the

TABLE IAverage of ME(p) in Simulations for Models Selected with Various
Criteria, for Yule–Walker, Burg, LSFB and, LSF Algorithms:
 $N = 25, L = 15,500\,000$ Runs. AR(10) Process with Reflection
Coefficients, $k_i = (-0.6)^i$

	Y-W	BURG	LSFB	LSF
AIC	5.30	26.92	125.10	66.18
FIC(p,2)	5.54	14.37	42.86	34.35
GIC(p,3)	5.86	14.39	98.43	55.19
FIC(p,3)	5.61	7.28	13.93	17.16
FSIC(p)	5.25	6.03	6.02	7.03
CIC(p)	5.62	6.44	6.40	7.53
AR(2)	4.33	4.39	4.40	4.47
AR(3)	3.86	3.98	4.04	4.23
AR(10)	9.03	16.58	20.71	39.39
E[ME(10)]	8.23	15.63	18.97	25.41

data. The latter three are chosen by the data analyst. If the selected model depends on those latter choices, any objective statistical analysis of time series becomes questionable. Our aim is maintaining the first two influences of the data and reduction of the last three influences. Preferences will be formulated based on theory and simulations. Three rows in Table I present the average ME(p) of three fixed-order models without order selection: the true order 10 and the best order for prediction with estimated AR models, which is AR(3) for 25 samples of this AR(10) process. The last row gives the finite sample expectation E[ME(10)] for the AR(10) model, which is found by substituting the finite sample formula (4) for PE(p) with the different $v(i, \cdot)$ of (3) in the first representation of (14). This can be compared with the simulation result of the estimated AR(10) model. The asymptotical theory gives as expectation 10 for all methods, and therefore, the finite sample expression for E[ME(10)] gives a much better representation for the actual performance of different estimation methods. In addition, the simulation average ME(p) of models selected with AIC, FIC(p, 2), GIC(p, 3), FIC(p, 3), FSIC(p), and CIC(p) is presented. Many other order selection criteria have been included in this comparison, but those results are not reported here. Consistent order selection criteria, with penalty $\ln(N)$ or $2 \ln \ln(N)$, boil down to a fixed penalty factor α for a given sample size (3.22 and 2.34, respectively) for N = 25, and they are similar to the results of AIC or GIC(p, 3) in Table I. The highest considered candidate order L is 15. It can be taken higher if FSIC(p) or CIC(p) are used, without influencing the average ME of selected models; but higher choices for L will influence the selection results of the other criteria. An additional limitation for candidate orders is that for a certain estimation method no parameters will be estimated for orders i, where $v(i, \cdot)$ of (3) becomes greater than 0.25. This limitation becomes active for LSF estimates in Table I. As $v(i, \cdot)$ is an approximation for the finite sample variance, this value for $v(i, \cdot)$ indicates 0.5 for the standard deviation of the last parameter, whereas the last parameter or the reflection coefficient has to be in the range of -1 to +1 for stationary models. Parameters with that amount of statistical uncertainty (more than 50%) are not useful and cannot contribute significantly to our knowledge of the process. In using

this limitation on $v(i, \cdot)$, the highest candidate order L is completely determined by the capacity of an estimation algorithm to compute useful models. If *a priori* information about an upper boundary for the model order is available, this can be used to choose a lower value of L.

The selection results for the Yule-Walker method are completely different from the other methods. The results of FSIC(p)and AIC(p) in Table I are almost identical for Yule–Walker, like what is expected in Section II. Bias has not been incorporated in the finite sample description of AR modeling. It is known that the Yule–Walker method introduces extra bias terms by using the biased estimates for the covariance [20]. Due to this specific Yule-Walker bias, higher order reflection coefficients are estimated a factor two wrong after a single true reflection coefficient k_p with magnitude |1 - p/N|. More general, the further estimates of higher orders after a large reflection coefficient suffer from a serious bias in the Yule–Walker method, instead of the smaller bias with magnitude 1/N that is present in all other estimation methods. Hence, the application of the Yule-Walker method to unknown data may produce erroneous results, even for a large number of observations. All poles of the AR(10)process of Table I have the absolute value 0.6; therefore, they are not close to the unit circle. Hence, the bias problem is not serious in that example, and under those conditions, the performance of Yule-Walker estimates is good.

Table II shows examples with and without serious bias influence. AR(p) processes with reflection coefficients $k_i = (-\beta)^i$ have the property that all poles have the radius β . With poles closer to the unit circle, the Yule–Walker method produces very high values of ME. ME(10) for Burg also becomes a bit higher for $\beta = 0.9$, which is caused by the fact that 25 observations is much shorter than the correlation length for $\beta > 0.7$, causing still another type of bias. It follows from Table II that the average performance of Yule–Walker over a variety of examples is poor due to bias. Therefore, Yule–Walker cannot be recommended for data of unknown processes, and the discussion in the sequel will be limited to the other three methods.

For estimation methods other than Yule-Walker, a comparison in Table I between asymptotical and finite sample criteria for different values of the penalty factor shows that the finite sample criteria FIC(p, 2) and FIC(p, 3) have a better performance than their asymptotical equivalents for the same value of α , AIC, and GIC(p, 3), respectively. Furthermore, $\alpha = 3$ is a better value for the penalty factor than $\alpha = 2$. The model error for selection with FSIC(p) and CIC(p) is smaller than with the other criteria. The difference between those two criteria is always small in finite samples. The better performance of finite sample criteria $FIC(p, \alpha)$ in comparison with their asymptotical equivalents $GIC(p, \alpha)$ is sufficiently explained by the fact that $FIC(p, \alpha)$ takes into account the actual expectation of the residual variance and prediction error as given in (4), whereas $GIC(p, \alpha)$ is based on the asymptotical approximations. One aspect had been neglected: the variance of the logarithm of the estimated residual variance. The higher variance of $\ln\{\text{RES}(p)\}$ for higher model orders gives an increased probability that those higher orders accidentally yield the minimum value of the criteria $GIC(p, \alpha)$ and $FIC(p, \alpha)$, thus giving a higher

 TABLE II

 Average of ME(10) of Fixed Order AR(10) Models for Yule–Walker

 AND BURG ALGORITHMS: N = 25, Average of 5000 Simulation Runs of

 AR(10) Processes with 10 Reflection Coefficients, $k_i = (-\beta)^i$

β	Yule-Walker	Burg	
.2	8.83	16.82	
.4	8.78	16.77	
.6	9.03	16.58	
.7	12.53	17.25	
.8	37.82	19.10	
.9	675.55	27.63	

average ME(p) after selection. This effect disappears in Table I if FSIC(p) or CIC(p) are used. LSFB and Burg are almost the same if selected with FSIC(p) or CIC(p).

Substitution of smaller values for $v(i, \cdot)$ in (4) gives a greater residual variance and, at the same time, a smaller value for the expectation of the prediction error. The asymptotical theory makes no distinction between the residual variance of different AR estimation algorithms; in addition, the theoretical prediction error is the same for all methods. The finite sample formulae, however, show that an estimation method, yielding a *smaller* residual variance for an adequate AR(p) model than other estimation algorithms, will give a higher prediction error. This means that the AR method estimating parameters of a model of order p with the lowest RES(p) gives the worst result in terms of an objective quality criterion like PE(p). However, asymptotical selection criteria would give preference to that method with lowest residual variance if they would be used to assess the quality of estimation methods. Hence, those asymptotical selection criteria are not suitable for a useful comparison of models obtained with different estimation algorithms. The finite sample formula for the prediction error PE(p) is for a *fixed* order p determined by the product of $1 + v(i, \cdot)$ in (4). This yields as sequence for decreasing accuracy: Yule-Walker, Burg, two-sided least squares LSFB, and, finally, one-sided LSF. This agrees with the outcomes for ME in Table I for the average of AR(2), AR(3), and AR(10) models. The differences are small for the AR(2)and AR(3) models because those orders are about 0.1N, the value where asymptotical theory and finite sample formulae start to diverge. This divergence is obvious for the AR(10) models estimated with the different methods.

Due to bias, the Yule–Walker method gives, in Table II, a much higher PE(p) than other estimation methods for processes whose reflection coefficients are not close to zero. Furthermore, the LSF method has the additional inconvenience that estimated roots can lie outside the stationarity region, which makes the resulting spectral model quite useless. This applies in a lesser degree to LSFB. According to Table I, to finite sample considerations and to many other simulations, Burg's estimation method is a good choice for estimation of AR parameters if the fit of a selected model is important, e.g., for prediction or as a spectral estimate. CIC(p) or FSIC(p) are good order selection criteria, and the maximum candidate order L can be chosen high enough, e.g., N/2 or even higher. These choices for estimation method, order selection criterion, and maximum candidate order can be

made independent of the observations, which gives advantages in a statistical interpretation.

VI. COMPARISON OF ORDER SELECTION CRITERIA

Comparative studies of selection criteria are mostly limited to a few examples and are therefore difficult for interpretation without some theoretical guideline. The performance of a selection criterion may depend strongly on the true process, the estimation method, the sample size, and the highest candidate order for selection. No single-order selection criterion can perform optimal in all possible circumstances. Consider an example process where M is actually the best model order for a given sample size; therefore, ME has a minimum for the fixedorder M. Now, taking zero as penalty for all orders below Mand ∞ as penalty for all orders above M, one will certainly select order M, which is the best possible solution for this particular example. Evidently, this "selection criterion" with zero and infinite penalties would perform very poorly on data generated by other processes with a different best order. Hence, the practical question is not to find the very best order selection criterion for a given example; that will be the criterion with zero and ∞ penalties, as defined above. The search is for an order selection criterion that never fails to select the order of one of the better fitting AR models that can be computed from the given observations for all types of processes. This problem will be treated in two separate stages: asymptotical and finite sample. A useful criterion will perform well in both conditions. Only results for Burg's estimation method are presented here in order to concentrate the information on the method that is preferable with the arguments in Section V.

Table III gives average ME results for a number of criteria for 10000 observations of four different AR(10) processes. $FIC(p, \alpha)$ and $GIC(p, \alpha)$ are almost identical for all values of α . Moreover, the selected model with FSIC(p) and AIC_C have the same quality as $\alpha = 2$, and CIC(p) gives the same result as penalty $\alpha = 3$. The conclusion is that penalty $\alpha = 3$, or the new criterion CIC(p), is the best in Table III, better than $\alpha = 2$ and $\alpha = 4$, and also better than $\alpha = \ln N$ of the consistent criteria [8]. The different values of β for the simulations in Table III have been chosen such that overfit and underfit are both serious possibilities. Those examples are most critical for the performance of order selection criteria. Taking lower values for β reduces the cost of underfit too much; higher values of β would give all underfitting models of order lower than 10 much higher residual variances than the AR(10) model such that they are almost never selected with any criterion. For $\beta = 0.2$ or $\beta = 0.4$, the average ME(p) values after order selection are less than the asymptotical value for the true order E[ME(10)] ≈ 10 because lower order models are selected, and they have a smaller finite sample expectation than the AR(10) model. In those asymptotical conditions, the differences between the four AR estimation methods are negligible as long as the extra Yule–Walker bias is small.

For the finite sample performance, a similar AR(10) process has been used in Table IV; the reflection coefficients are powers of -0.7 here. Table IV demonstrates, only for the Burg method,

TABLE IIIAsymptotical Performance of Criteria: Average of ME(p) inSimulations with Order Selection for the Burg Algorithm: $N = 10\,000, L = 20, 1000$ Runs, AR(10) Process with ReflectionCoefficients, $k_i = (-\beta)^i$

β	0.2	0.4	0.6	0.7
$\overline{AIC(p)}=GIC(p,2)$	4.08	7.54	11.99	12.45
FIC(p,2)	4.08	7.54	11.99	12.45
FSIC(p) and AIC _C	4.08	7.54	12.00	12.44
GIC(p,3) and FIC(p,3)	3.31	6.57	11.47	11.44
CIC(p)	3.31	6.57	11.47	11.44
GIC(p,4)	3.34	6.79	12.02	11.60
FIC(p,4)	3.34	6.80	12.02	11.60
GIC(p,lnN)	4.92	9.04	15.99	15.82
FIC(p,lnN)	4.92	9.04	16.00	15.83

 TABLE IV

 Average of ME(p) of Selected and of Fixed-Order Models,

 Estimated with Burg Algorithm as a Function of the Sample Size.

 Average of 10 000 Simulation Runs of an AR(10) Process with

 Reflection Coefficients $k_i = (-0.7)^i$

N	16	16	64	256	1024
L	15	8	32	128	512
AIC	153.45	17.30	27.20	18.71	12.91
FIC(p,2)	99.15	13.51	12.97	11.94	12.80
GIC(p,3)	145.49	13.90	11.12	11.14	12.30
FIC(p,3)	54.08	10.89	9.92	11.13	12.30
GIC(p,4)	131.11	12.22	10.59	11.81	13.04
FIC(p,4)	23.73	10.63	10.70	11.87	13.06
GIC(p,lnN)	148.88	14.62	10.72	13.47	14.07
FIC(p,lnN)	60.81	11.18	10.86	13.59	14.11
FSIC(p)	8.74	8.74	9.63	11.31	12.65
CIC(p)	9.31	9.31	9.92	11.13	12.30
AICc	xxxx	9.07	9.88	11.47	12.70
AR(3)	6.05	6.05	11.10	34.31	127.48
AR(5)	7.97	7.97	7.10	11.95	32.52
AR(7)	12.89	12.89	8.12	8.59	12.89
AR(9)	21.37	21.37	10.40	9.46	9.86
AR(10)	28.08	28.08	11.78	10.39	10.09
E[ME(10)]	22.68	22.68	11.64	10.37	10.09

the influence of the sample size N (with the value N/2 as maximum candidate order L) on selecting the order with different criteria. The first column gives results for L = N - 1 for N = 16. The difference between the first two columns is only caused by different maximum candidate orders for selection: 15 and 8, respectively. This demonstrates that only FSIC(p) and CIC(p) are insensitive to extremely high values for the highest candidate order L. The results for the AIC_C criterion [14] are close to those of FSIC(p) for all sample sizes; only for orders as high as N - 2 or N - 1, AIC_C is not properly defined.

The five AR(p) rows in Table IV with ME for fixed-order models show that the best order for this AR(10) process depends on the sample size. Therefore, the ability to select the true order 10 would not be a good quality measure for a comparison of order selection criteria in this example. The average ME(10) of the AR(10) model is close to its finite sample expectation E[ME(10)] for all N. The best theoretical value for the order of estimated models as a function of the sample size can be found

Some artifacts in simulations about order selection are clear in Table IV. Comparing the results of the criteria $GIC(p, \alpha)$ and $FIC(p, \alpha)$, the best choice between those would be FIC(p, 3)for N = 64, N = 256 and N = 1024, but penalty 4 is slightly better for N = 16. This somewhat irregular behavior is caused by the interaction between the highest candidate order L, the best order for a given sample size, and the possibility that the second best model order is one order lower or perhaps one order higher than the best order. Consistent criteria with $\alpha = \ln N$ always have a greater ME than FIC(p, 3) in this example for all sample sizes. However, consistent criteria may outperform other criteria in examples where the true process has a few very significant parameters, the number of observations is high, and, at the same time, the highest candidate order L is not too high. Under those circumstances, consistent criteria approximate the performance of the optimal criterion with zero and infinite penalties as described in the beginning of this section.

Tables III and IV show that the differences between $GIC(p, \alpha)$ and $FIC(p, \alpha)$ disappear almost completely for increasing N for every value of α . CIC(p) and FIC(p, 3) give exactly the same selection results for ME if the sample size Nis 64 or more in Table IV. Based on the finite sample behavior in this example and in numerous other simulations, the criteria CIC(p) and FSIC(p) of the third category are preferred. A slight preference for CIC(p) above FSIC(p) is found with Table III because the theoretical properties of CIC(p) agree with the best asymptotical choice for the penalty factor. In finite sample simulations, the comparison between those two criteria depends on the specific process in the example and slightly on the AR estimation method as well. Many other simulations with AR, MA, and ARMA processes of different orders and with small or very significant values for the true parameters have been carried out. Normal, uniform, and other probability density functions have been used for ε_n . The criteria have also been applied to practical data. No single example has been found where the performance of FSIC(p) or CIC(p) was poor. The results of the least squares methods LSFB and LSF are similar and lead to the same conclusions: $GIC(p, \alpha)$ is very sensitive for the choice of L, FIC(p, α) is less sensitive, and FSIC(p) and CIC(p) are almost independent of the choice of L if only L is chosen greater than the best fixed-order p' with the smallest E[ME(p')] for a given sample size. Moreover, penalty 3 is a good asymptotical compromise, which means that CIC(p) is also preferred for the least squares AR methods. The results of these simulations for the Yule-Walker method show that if the extra bias is not small, no good models are available, and order selection cannot solve that problem.

A detailed study of the Tables I and IV shows that the ME for selected models is greater than ME for the best fixed order, even if the best criterion is used for selection. The difference is

about 2.5 for all sample sizes. This value is very close to the asymptotical value 2.568 that has been found for selection with AIC(p) if only the possibility of overfit was taken into account [8]. The AR(10) example is a difficult one for order selection because the best order has close competitors of higher and of lower orders with overfit and underfit. The loss in quality due to selection is, statistically seen, a consequence of using the same residual variance for two different purposes: first, for minimization to estimate the parameters and afterwards for order selection with a transformation of the residual variance as a selection criterion. The model error of models selected with FSIC(p) or CIC(p) is about 2.5 greater than E[ME(p')], where p' denotes the best fixed-order model. In other words, even the best selection criterion is not capable to detect the very best order in every simulation run in this example.

VII. CONCLUDING REMARKS

A recursive formula gives an accurate description of the empirical variance of autoregressive order-selection criteria that are based on the logarithm of the residual energy. The empirical variance depends on the estimation method in finite samples. It has influence on selection results if the maximum candidate order for selection is greater than N/4. Such high AR orders are required for subsequent MA or ARMA estimation with Durbin's methods.

A new selection criterion called the combined information criterion CIC(p) that combines the theoretical asymptotical preference for penalty 3 with the good finite sample characteristics of FSIC(p) is introduced. FSIC(p) and CIC(p) are selection criteria with an increasing penalty for higher model orders. This gives protection against abusively selecting those high orders. The performance of these criteria was always good in simulations.

Based on statistical arguments, Burg's AR estimation method with CIC(p) as order selection criterion and with N/2 or higher as maximum candidate order is a good and safe combination for the automatic use of autoregressive modeling for time series in practice.

REFERENCES

- S. M. Kay and S. L. Marple, "Spectrum analysis—A modern perspective," *Proc. IEEE*, vol. 69, pp. 1380–1419, 1981.
- [2] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Upper Saddle River, NJ: Prentice-Hall, 1997.
- [3] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, 1974.
- [4] —, "A Bayesian analysis of the minimum AIC procedure," Ann. Inst. Statist. Math., vol. 30A, pp. 9–14, 1978.
- [5] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 41, pp. 465–471, 1978.
- [6] E. J. Hannan and B. J. Quinn, "The determination of the order of an autoregression," in J. R. Statist. Soc., ser. B-41, 1979, pp. 190–195.
- [7] R. Shibata, "Selection of the order of an autoregressive model by Akaike's information criterion," *Biometrika*, vol. 63, pp. 117–126, 1976.
- [8] P. M. T. Broersen and H. E. Wensink, "On the penalty factor for autoregressive order selection in finite samples," *IEEE Trans. Signal Processing*, vol. 44, pp. 748–752, Mar. 1996.
- [9] —, "On finite sample theory for autoregressive model order selection," *IEEE Trans. Signal Processing*, vol. 41, pp. 194–204, Jan. 1993.
- [10] J. Rissanen, "A predictive least-squares principle," IMA J. Math. Contr. Inform., vol. 3, pp. 211–222, 1986.

- [11] M. Wax, "Order selection for AR models by predictive least squares," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 581–588, 1988.
- [12] P. M. T. Broersen and H. E. Wensink, "Autoregressive model order selection by a finite sample estimator for the Kullback–Leibler discrepancy," *IEEE Trans. Signal Process.*, vol. 46, pp. 2058–2061, July 1998.
- [13] J. P. Burg, "Maximum likelihood spectral analysis," in *Proc. 37th Meet. Soc. Exploration Geophys.*, Oklahoma City, OK, 1967, p. 6.
- [14] C. M. Hurvich and C. L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, pp. 297–307, 1989.
- [15] P. M. T. Broersen, "The ABC of autoregressive order selection criteria," in *Proc. SYSID SICE, Preprints 11th IFAC Symp. Syst. Identification*, 1997, pp. 231–236.
- [16] —, "The best order of long autoregressive models for moving average estimation," in *Proc. Signal Processing VIII Eusipco Conf.*, Trieste, Italy, 1996, pp. 799–802.
- [17] —, "Autoregressive model orders for Durbin's MA and ARMA estimators," *IEEE Trans. on Signal Process.*, vol. 48, pp. 2454–2457, Aug. 2000.

- [18] —, "The quality of models for ARMA processes," *IEEE Trans. Signal Process.*, vol. 46, pp. 1749–1752, June 1998.
- [19] B. Friedlander and B. Porat, "A general lower bound for parametric spectrum estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 728–733, 1984.
- [20] J. S. Erkelens and P. M. T. Broersen, "Bias propagation in the autocorrelation method of linear prediction," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 116–119, Mar. 1997.

Piet M. T. Broersen was born in Zijdewind, The Netherlands, in 1944. He received the M.Sc. degree in applied physics in 1968 and the Ph.D. degree in 1976, both from the Delft University of Technology, Delft, The Netherlands.

He is currently with the Department of Applied Physics, Delft University. His research interests are the selection of order and type of time series models and the application to spectral analysis, model building, and feature extraction.