TREIZIÈME COLLOQUE GRETSI - JUAN-LES-PINS DU 16 AU 20 SEPTEMBRE 1991

FINITE SAMPLE THEORY FOR AUTOREGRESSIVE MODEL ESTIMATION

P.M.T. Broersen
H.E. Wensink

Faculty of Applied Physics, Delft University of Technology
P.O.Box 5046, 2600 GA Delft, the Netherlands

RÉSUMÉ

La théorie concernant les processus Auto-Regressifs peut être divisée en trois niveaux d'approximation. Un traitement des différents critères pour la sélection d'un ordre est basé sur cette classification, qui est supportée par des résultats de simulations. Les critères mesurent l'exactitude d'un modèle, point de vue de prédiction, au moyen d'une transformation des résidus. La base théorétique rends possible d'arriver à une amélioration des critères existents, comme l'AIC et ses variantes.

ABSTRACT

In autoregressive theory three levels of approximation can be distinguished: plim, asymptotic and finite sample theory. Starting from this classification the different criteria for order selection are considered. The sum of squared prediction errors is a measure for the accuracy of an estimated model. The prediction error is estimated by means of a transformation of the minimized residual variance. The theoretical background provides a possibility for improving the performance of existing criteria like AIC and its consistent variants. Both asymptotic and finite sample theory are supported by simulation results which at the same time indicate the distinction between these levels of approximation.

INTRODUCTION

Autoregressive processes are known in many fields of application like speech analysis, spectroscopy and geophysics. When an AR model is fitted to a time series the parameters have to be estimated and an order has to be selected. The parameters are mostly estimated using one of the estimation methods like Yule-Walker, the method of Burg, the least squares method with forward residuals only (LSF) or the least squares method with forward and backward residuals (LSFB); see Kay and Marple (1981). The order has to be selected by means of an order selection criterion. With each additional parameter above the optimal AR order the residual sum of squares decreases. The sum of squared prediction errors increases at the same time with an equal amount. This is a measure for the accuracy of applying an estimated model to new data. It is the basis for the Final Prediction Error (FPE) of Akaike (1970).

Three levels of approximation to AR estimation will be distinguished: probability limits, asymptotic theory and finite sample theory. The classical theoretical framework is provided by the probability limits of Mann and Wald (1943). In the probability limit the estimates will assume their statistical expectation. Terms of magnitude-order 1/N are neglected, where N is the number of observations.

Accuracy of order 1/N is necessary for order selection, because the essential differences between competing models are of O(1/N). Therefore Akaike (1970,1974) developed his theory. The AIC was introduced in 1974 as an improvement to the previous FPE. Also consistent variants of the AIC were presented, like the criterion of Rissanen (1978) or the criterion of Hannan and Quinn (1979). The theoretical derivations contain some heuristic arguments and were critisized by Bhansali (1986). Lately it has been shown by Broersen and Wensink (1991b) that all results can easily be derived by considering AR estimators as non-linear functions of estimated covariances. A second order Taylor expansion can be used to approximate their expectation with accuracy 1/N.

Broersen (1985, 1990) showed in simulation experiments that the residual variance and the prediction error depend on the method of parameter estimation in finite samples. Finite sample parameter variances were developed to describe the behaviour of each estimation method. The lack of theoretical background of the finite sample results may be an obstacle in their acceptance. Therefore, the FSC is presented as a physically based order selection criterion, being the best possible estimate of the prediction error from the available data. The criterion value itself is an estimate for the prediction error of the selected model. As a natural result evolves from these considerations the Finite Sample Theory. FSC converges to FPE if the sample size increases. The General Information Criterion GIC describes all existing order selection criteria that use the logarithm of the residual variance. A finite sample improvement for this criterion is given with the Finite sample Information Criterion (FIC).

The three levels of approximation in the theory are enlightened by an overview in a Table and by some figures.

AUTOREGRESSIVE THEORY

An AR(K) process is given by:

$$x_n + a_1 x_{n-1} + \ldots + a_K x_{n-K} = \epsilon_n \ ,$$

where ϵ is the generating innovation proces, i.i.d. with zero mean and variance σ_ϵ^2, which is taken to be 1 in our simulations.

The residual variance $S^2(p)$ is the fit of a model to the data and is minimized to estimate the parameters:

$$S^2(p) = \frac{1}{N-p} \sum_{n=p+1}^{N} (x_n + a_{p1} x_{n-1} + \ldots + a_{pp} x_{n-p})^2 \ .$$

The parameter a_{pi} is the i-th estimated parameter in a model of order p.

The prediction error PE(p) is the measure for the model fit to an independent realisation y(n) of the same stochastic proces x(n). It is given by:

$$PE(p) = (y_n + a_{p1} y_{n-1} + \ldots + a_{pp} y_{n-p})^2 \ .$$

Three levels of approximation to Autoregressive estimation are distinghuised in Table 1.

Table 1. Three levels of approximation of AR theory

Probability limits	Asymptotic theory	Finite Sample Theory
$N \to \infty$	$O(1/N)$	$v(i,.)$
Mann and Wald (1943)	Akaike (1970, 1974) Bhansali (1986) , etc. Broersen/Wensink (1991b)	Broersen (1985, 1990) Broersen and Wensink (1991a)
$\underset{N\to\infty}{\text{plim}}\ S^2(p) = \sigma_\epsilon^2 \quad p \geq K$	$E\{S^2(p)\} = \sigma_\epsilon^2\,(1 - p/N) \quad p \geq K$	$E\{S^2(p)\} = \sigma_\epsilon^2 \prod_{i=0}^{p} (1 - v(i,.)) \quad p \geq K$
$\underset{N\to\infty}{\text{plim}}\ PE(p) = \sigma_\epsilon^2 \quad p \geq K$	$E\{PE(p)\} = \sigma_\epsilon^2\,(1 + p/N) \quad p \geq K$	$E\{PE(p)\} = \sigma_\epsilon^2 \prod_{i=0}^{p} (1 + v(i,.))$
No order selection	$FPE(p) = S^2(p)\,\frac{N+p}{N-p}$	$FSC(p) = S^2(p) \prod_{i=0}^{p} \frac{(1 + v(i,.))}{(1 - v(i,.))}$
	$GIC(p,\alpha) = \ln\{S^2(p)\} + \alpha\, p/N$	$FIC(p,\alpha) = \ln\{S^2(p)\} + \alpha \sum_{i=0}^{p} v(i,.)$

⊙ Probability limits of Mann and Wald (1943), using :

$$\underset{N\to\infty}{\text{plim}}\ x_N = a \text{ or } \lim_{N\to\infty} \text{Prob}(|x_N - a| \leq \epsilon) = 1 \text{ for } \epsilon > 0.$$

Thus they prove that autoregression can be dealt with as if it were a classical linear regression problem. The same observations are treated as dependent and as independent variables. In the limit for $N\to\infty$ all estimates will assume their statistical expectations. Parameter estimates become equal to the true parameter values; the bias of order 1/N disappears. Therefore, the probability density functions of predictions and innovations become the same. No difference can be made between prediction error and residual variance. The theoretical Yule-Walker equations describe the exact relations between the true parameters and the expectations of the data covariances at this level.

⊙ Asymptotic theory. Small differences between models of increasing order determine which order is the best to select. It is important to make a clear distinction between PE(p) and $S^2(p)$. Akaike (1970) showed the differences to be of O(1/N). His original derivation contains heuristic arguments. However it is not uncommon, that an essential correct result is obtained by dubious derivations. Broersen and Wensink (1991b) showed, that the expectations of all interesting quantities can be derived in a straight-forward manner. Parameters, residual variance and prediction error are written as non-linear transformations of measured covariances:

$$\hat{R}(p) = \frac{1}{N-p} \sum_{n=1}^{N-p} x_n x_{n+p} .$$

All desired transformations are found by means of the theoretical Yule-Walker relations, substituting the estimated $\hat{R}(p)$ instead of its expectation R(p). A Taylor expansion of those functions will give the desired expectations. They are a function of the variances and covariances of the measured $\hat{R}(p)$, for which the asymptotic expressions of Bartlett can be used; see Priestley (1981, p326).

The residual variance and the prediction error are given in Table 1. Also the bias in the parameters can be computed in this way. Starting from the Yule-Walker equations the parameters can be calculated:

$$\hat{R}\,\underline{a} = -\hat{\underline{r}}.$$

A first order model gives $a_1 = -\hat{\rho}(1) = -\hat{R}(1)/\hat{R}(0)$.

Bias is caused by the expectation of the quotient not being equal to the quotient of the expectations. Using the second order Taylor expansion the bias can be computed :

$$E(a_1) = \alpha_1(1-2/N).$$

This result can easily be extented to higher order models. A second contribution to the bias is present if the estimated mean of the time series is subtracted; see Broersen and Wensink (1991b). A third form of bias arises in the Yule-Walker method of parameter estimation, because the estimates of covariances used are divided by N instead of N-i.

⊙ Finite Sample Theory. This provides a theoretical framework for results of simulations. The effect of the estimation method becomes important. The treatment of the first and sometimes the last p data points causes considerable differences between the outcomes of the estimation methods. To discriminate between those different kinds of behaviour the finite sample variances v(i,.) were introduced. They are the basic elements of the Finite Sample Theory. It is shown in the asymptotical theory and in finite sample simulations that the variance of the last estimated parameter p, for p greater than the true proces order K, is independent of the process. The v(i,.) are calculated from a white noise process, so that all estimated parameters are above the true process order. The results are presented for four estimation methods in Fig.1. The points represent the simulation results, the drawn lines are the emperical approximations:

v(i,YW) = (N-i)/(N(N+2))
v(i,Burg) = 1/(N+1-i)
v(i,LSFB) = 1/(N+1.5-1.5i)
v(i,LSF) = 1/(N+2-2i) .

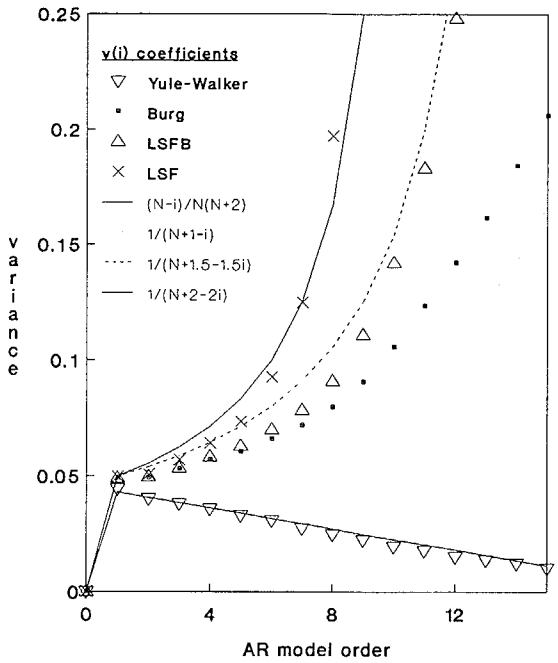

Fig.1 Variance of AR estimates from 20 white noise observations

For all methods $v(0,.) = 0$. It is easily seen that the asymptotical value for the variance of 1/N appears to be only a fair approximation if the model order i is small in comparison with the number of observations

PREDICTION ERROR

The prediction error PE is a measure for the fit of the model to future data. Its value indicates the accuracy of an estimated model. It can be computed in simulations, but in practice it is unknown and must be estimated. In order selection the model with the lowest estimate for PE is in general selected. These models give best descriptions of the data, even if the selected order differs from the order of the process that generated the data. Broersen (1989) illustrated this when he found AR(3) models to be most accurate in describing finite sample sequences from an AR(8) process; the best order depends on the sample size.

Akaike (1970) introduced the Final Prediction Error (FPE) to get an estimate for the PE (see Table 1). It is based on the fact that each extra parameter above the best model order gives a decrease for $S^2(p)$ and an equally valued increase for PE(p).

Fig. 2 presents the average value of PE(p) in 10000 simulation runs for four estimation methods, as a function of the model order p. In finite samples the actual behaviour of PE is apparently dependent on the method of parameter estimation. It has been proved by induction that:

$$E\{PE(p)\} = \sigma_\epsilon^2 \prod_{i=0}^{p} \{ 1 + v(i,.) \}, \quad p \geq K.$$

Likewise the reduction of the residual variance above the best model order is independent of the data:

$$E\{S^2(p)\} = E[\{S^2(p-1)\}\{1 - a_{pp}^2\}]$$

$$= \sigma_\epsilon^2 \prod_{i=0}^{p} \{ 1 - v(i,.) \} , \quad p \geq K.$$

To transform the residual variance of an individual realisation into an estimate of the PE the Finite Sample Criterion (FSC) is suggested. It uses the same approach as FPE, being based on the quotient of expectations of PE(p) and $S^2(p)$:

$$FSC(p) = S^2(p) \prod_{i=0}^{p} \{(1+v(i,.))/(1-v(i,.))\}$$

In Fig.3 the average PE(p) in simulations is compared with its asymptotic estimate FPE(p) and its Finite Sample estimate FSC(p). FPE(p) has a maximum at about order 7 and gets a second minimum at higher orders. FSC(p) appears to be a better approximation. It folllows the curve of PE fairly well up to order 10. Higher order models are not estimated from 20 observations. FSC(p) converges to FPE(p) for large sample sizes with $N \gg p$.

GENERALIZED INFORMATION CRITERION

AIC and its consistent variants can together be described as a Generalized Information Criterion:

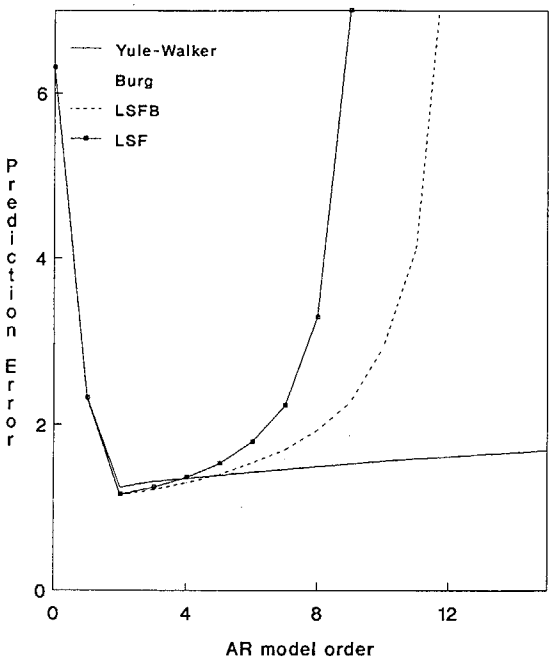

Fig.2 PE for different estimates from 20 AR(2) observations

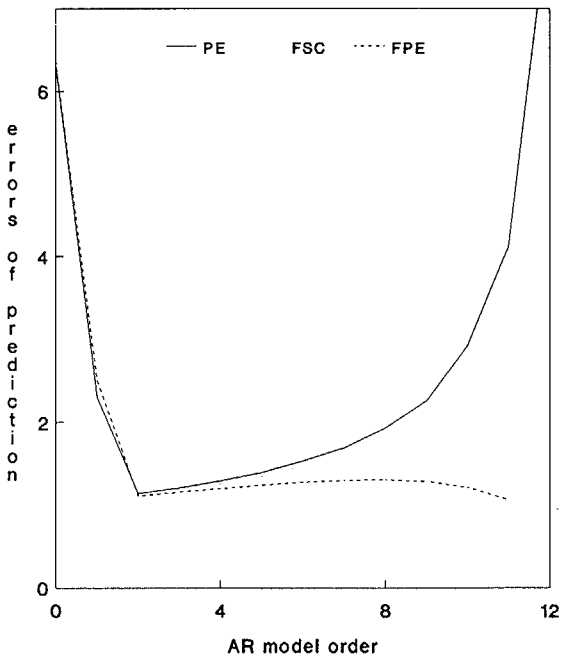

Fig.3 PE, FPE and FSC for LSFB estimates from 20 AR(2) observations

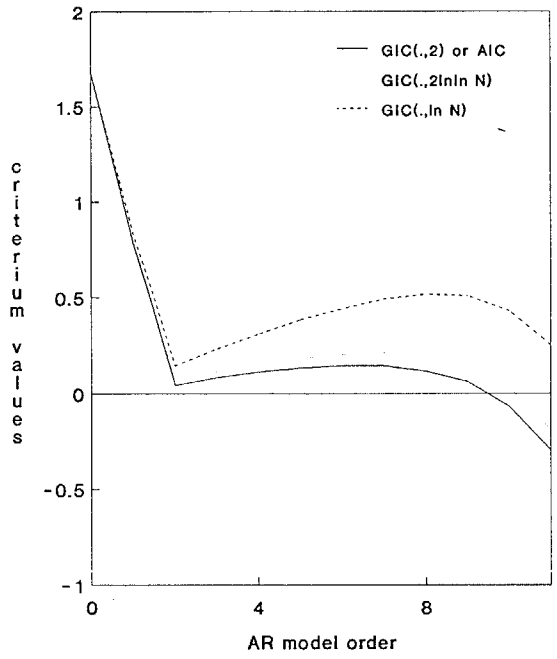

Fig.4 GIC for LSFB estimates from 20 AR(2) observations

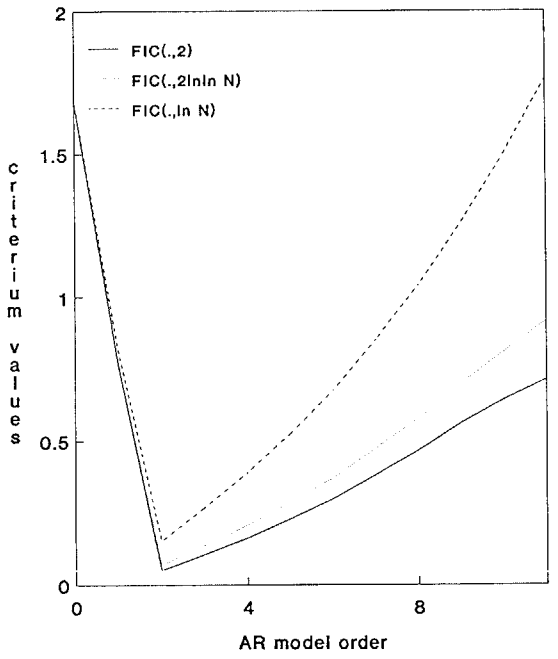

Fig.5 FIC for LSFB estimates from 20 AR(2) observations

$$GIC(p,\alpha) = \ln\{S^2(p)\} + \alpha\, p/N$$

This is AIC for α=2, the criterion of Hannan and Quinn 1979) for α=2ln{ln(N)} and the criterion of Schwarz and Rissanen (1978) for α=ln(N). One of the undesirable properties of the order selection criteria, described by GIC, is their second minimum. The deep high order minimum explains why methods may give poor results. In Fig 4. the curves are shown for LSFB estimates. The same shape is found for LSF and Burg estimates. Only for Yule-Walker no second order minimum is found. The simulation results of Fig.4 are accurately described by substituting the Finite Sample approximation for $E\{S^2(p)\}$ in GIC. This locates the maximum at $v(i,.) = \alpha/N$, which will never occur for Yule-Walker (if $\alpha > 1$), because $v(i,YW) < 1/N.$; see Broersen and Wensink (1991a).

The residual variance, expressed as a function of measured covariances, is a stochastic variable. In model order selection criteria described by GIC the logarithm of the residual variance is taken. So a non-linear operation on a stochastic variable is carried out. The Taylor expansion gives the expectation of the logarithm of the residual variance:

$$E[\ln\{S^2(p)\}] = \ln\{E[S^2(p)]\} - \frac{\mathrm{var}\{S^2(p)\}}{2E^2[S^2(p)]} .$$

In many theoretical derivations, the last term is omitted; see Bhansali (1986). The expectation and the logarithm are interchanged as if the probability limit was taken. This is inaccurate in derivations at asymptotic level where all terms of order 1/N should be included (see Table 1).

The finite sample modification of GIC is the Finite sample Information Criterion (FIC) defined as:

$$FIC(p,\alpha) = \ln\{S^2(p)\} + \alpha \sum_{i=0}^{p} v(i,.).$$

This formula follows for α=2 from FSC(p) by taking the logarithm and approximating $\ln(1+\delta)$ by δ. The improvement with respect to GIC is seen when comparing Fig.4 with Fig.5. FIC has only one minimum. A second wrong minimum can only occur if $\alpha < 1$, which never happens in a sensible order selection criterion. It should be noted that, although FIC is improved with respect to GIC, it is by no means a physically based criterion. That remains a privilege of FSC.

CONCLUSIONS

In this paper a theoretical framework for model order selection is proposed. It might be called a physical theory, because the emphasis is on the explanation of observed data rather than on strict mathematics. Three levels of approximation to AR estimation are presented. Thus it is possible to discriminate between different criteria for order selection.

REFERENCES

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Stat. Math.* **22** 203-217.

Akaike, H. (1974). A new look at statistical model identification. *IEEE Trans. Automat Contr.*, **AC-19**, 716-723.

Bhansali, R.J. (1986). A derivation of the information criteria for selecting autoregressive models. *Adv. Appl. Prob.* **18** 360-387.

Broersen. P.M.T. (1985). Selecting the order of autoregressive models from small samples. *IEEE Trans. Acoust, Speech, Signal Proc*, **ASSP-33** 874-879.

Broersen, P.M.T. (1989). A comparison of criteria for autoregressive order selection. *Selected Papers from the eight IFAC/IFORS Symposium on Ident. and System Par. Est., Beijing, PRC, 1988.* **IFAC Proceedings Series 8, Pergamon press**, Oxford, 763-768.

Broersen, P.M.T. (1990). The prediction error of Autoregressive small sample models. *IEEE Trans. Acoust., Speech, Signal Processing* **ASSP-38** 858-860.

Broersen, P.M.T. and Wensink, H.E. (1991a). Finite sample autoregressive order selection criteria. *Proc. IEEE/ProRISC Symp. on Circuits, Systems and Signal Processing, Veldhoven* 135-138.

Broersen, P.M.T. and Wensink, H.E. (1991b). On autoregressive theory. Offered for publication.

Hannan, E.J. and Quinn, B.G. (1979). The determination of the order of an autoregression. *J.R. Statist. Soc. Ser. B.*, B-41 190-195.

Kay, S.M. and Marple, S.L. (1981). Spectrum analysis, a modern perspective. *Proc. IEEE*, **69** 1380-1419.

Mann, H.B. and Wald, A. (1943). On the statistical treatment of linear stochastic difference equations. *Econometrica* 11 173-220.

Priestley, M.B. (1981). *Spectral Analysis and Time Series*. Academic Press, London.

Rissanen, J. (1978). Modelling by shortest data description, *Automatica*, **14** 465-471.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6** 461-464.