On the penalty factor for AR Order Selection Criteria

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

Signal Processing Group, Department of Applied Physics, Delft University of Technology
P.O. Box 5046, 2600 GA Delft, The Netherlands
e-mail: broersen@duttncb.tn.tudelft.nl
e-mail: wensink@duttncb.tn.tudelft.nl

RÉSUMÉ

La qualité d'une critère pour la sélection d'un ordre se caractérise par sa capacité de sélectionner des modèles dont l'espérance de l'erreur de prédiction *PE(p)* est minimale. Une modèle est sous-optimale si elle contient ou bien trop ou bien pas assez de paramètres. Le coût de la sélection d'un trop grand ordre est strictement statistique, tandis que le coût de la sélection d'un ordre qui ne contient pas assez de paramètres dépend entièrement des valeurs des paramètres du processus qui ait engendré les échantillons. Une critère se sert d'une facteur de décision α qui détermine la raison entre les coûts.

ABSTRACT

The quality of an order selection criterion is characterized by its capacity to select models with minimal expected squared error of prediction, *PE(p)*. The possibilities of selecting a suboptimal order can be divided in overfitting and underfitting. The costs of overfitting are determined by statistics, whereas the costs of underfitting depend on the values of the parameters of the data-generating process, which have been excluded from the model. The penalty factor of a selection criterion influences the ratio between these costs.

INTRODUCTION

In practical autoregressive (AR) model fitting to a data series, the parameters have to be estimated and the order has to be selected. In general the search is for that particular combination of parameters and model order that minimizes the squared error of prediction. This model has optimal forecasting capacities in the time domain and it can be shown that it yields an equally good description in the frequency domain [1]. For the selection of the model order several order selection criteria have been developed [2,3,4,5,6]. Individually all these criteria possess optimal asymptotical properties, but a wrong order is easily selected in practice. A thourough analysis of the behaviour in finite samples has given rise to an improvement, making use of the elements of Finite Sample Theory.

The central idea of the Finite Sample Theory [7,8;9] is to improve upon the asymptotic theory by replacing *1/N*, *N* the number of observations, by the quantity *v(i,.)*, that is based on the actual degrees of freedom which play a role in a given estimation method [10]. Four different AR estimation methods [11] have been investigated for the Finite Sample Theory. The Yule-Walker or autocorrelation method (YW), the method of Burg, the least squares method that minimizes both Forward and Backward residuals, LSFB, and the least squares method that uses Forward residuals only, LSF. The finite sample variance coefficients, *v(i,.)*, have been defined for these four estimation methods as:

$$\begin{aligned} v(i,YW) &= (N-i)/N(N+2) \\ v(i,Burg) &= 1/(N+1-i) \\ v(i,LSFB) &= 1/(N+1.5-1.5i) \\ v(i,LSF) &= 1/(N+2-2i) \end{aligned} \qquad (1)$$

For all methods, $v(0,.) = 1/N$ if the mean of the observations is subtracted before the estimation of the parameters, otherwise $v(0,.) = 0$; all *v(i,.)* approach *1/N* for *N* much greater than *i*.

In the search for accurately predicting AR models the magnitude of the penalty factor in order selection criteria is the subject of investigations in this paper. The asymptotic order selection criteria and their finite sample counterparts have been evaluated with respect to their order selection performances. Models with more as well as models with less parameters than the best model order are suboptimal with respect to their forecasting capacities. The Selection Risk has been used as a meta-criterion to compare the average selection results of these different criteria. Balancing the costs of both sides leads to the order selection criterion with that value of the penalty factor such that the best predicting model will be selected. It is demonstrated that the choice of the penalty factor influences the balance between the costs of overfitting and underfitting.

AR ESTIMATION

An autoregressive process of order *K* is defined as:

$$x_n + \sum_{i=1}^{K} a_i(K)\, x_{n-i} = \varepsilon_n, \qquad (2)$$

where ***a(K)*** is the parameter vector and ε_n is i.i.d., with zero mean, variance σ_ε^2 and finite fourth order moments.

To an AR series x_n, which has been generated by this *K*-th order process, an AR model of order *p* can be fitted:

$$x_n + \sum_{i=1}^{p} \hat{a}_i(p)\, x_{n-i} = \varepsilon_n. \qquad (3)$$

When fitting this AR(p) model to the data, the elements of the parameter vector ***â(p)*** have to be estimated and the order *p* has to be selected. Two typical quantities, often erroneously interchanged, arise in the estimation procedure: the residual variance, $S^2(p)$, and the Prediction Error, *PE(p)*. The residual variance is defined as the average squared fit of the estimated model to the data x_n from which the parameters have been inferred. The value of $S^2(p)$ will always decrease for increasing model order p, but the capacity of the model to predict the future behaviour of the series will only improve as long as significant parameters are included in it. The predictive capacity of the model, expressed by the squared error of prediction, $PE(p)$, worsens with every extra parameter that is added after all the significant parameters have been included in the model. The $PE(p)$, is found by weighting the estimated parameter vector $\hat{\mathbf{a}}(p)$ with the pxp submatrix of the theoretical covariance structure of the generating AR(K) process, $\mathbf{R}_\infty$ [10]:

$$PE(p) = \hat{\mathbf{a}}^T(p)\ \mathbf{R}_\infty(p)\ \hat{\mathbf{a}}(p). \qquad (4)$$

Small differences between competing model orders determine which order is best to select. These differences are considered to be of magnitude order $1/N$ in the asymptotical Large Sample Theory. Different order selection criteria have been proposed making use of elements of information theory. The criterion $AIC(p)$ [3] has been followed by modifications: a consistent variant of $AIC(p)$ has changed the penalty factor 2 of $AIC(p)$ into $ln(N)$ [4,5], a minimal consistent variant [6] uses $2lnln(N)$. All these criteria use the logarithm of the estimated residual variance. Together they can be described as a generalized information criterion, $GIC(p,\alpha)$:

$$GIC(p,\alpha) = \ln[S^2(p)] + \alpha\frac{p}{N}, \qquad (5)$$

with α the penalty factor, such that with

$\alpha = 2$ it is $AIC(p)$,
$\alpha = ln(N)$ it describes a consistent criterion,
$\alpha = 2lnln(N)$, a minimal consistent criterion,
$\alpha =$ *3, 4, etc.*, any other criterion.

In the Finite sample Information Criterion, $FIC(p,\alpha)$ [7,8,9] the factor p/N has been replaced by a summation over the first p finite sample variance coefficients $v(i,.)$ (1). So the $FIC(p,\alpha)$, besides depending on α, takes the characteristics of the estimation method into account via $v(i,.)$ (1). For each value of α, the finite sample equivalent of the $GIC(p,\alpha)$ is defined as:

$$FIC(p,\alpha) = \ln[S^2(p)] + \alpha\sum_{i=0}^{p} v(i,.). \qquad (6)$$

The performances of $GIC(p,\alpha)$ and $FIC(p,\alpha)$ have been evaluated in so-called fixed order experiments [9], where the average of the criterion value is determined for each model order individually. These simulation results are displayed in Figure 1, where the criterion values are given as a function of the model order for $\alpha = 2$ and $\alpha = ln(N)$. $GIC(p,\alpha)$ has a wrong artificial high order maximum for both values of α; afterwards the criterion values decrease and it will eventually give a deeper minimum than is found at the optimal order. The $FIC(p,\alpha)$, on the contrary, has only one pronounced minimum under the condition that $\alpha > 1$. Above the true model order the criterion value keeps increasing with a speed that depends only on the numerical value of α. The improvement provided by the $FIC(p,\alpha)$ is due to the fact that the influence of the estimation method has been taken into account.

Closely related to the $PE(p)$ and a measure for comparing the performance of the different order selection criteria is the selection error, $SE(p)$, which is defined for a model of order p as:

$$SE(p) = \frac{N[\hat{\mathbf{a}}^T(p,\infty) - \mathbf{a}^T(K,\infty)]\ \mathbf{R}_\infty\ [\hat{\mathbf{a}}(p,\infty) - \mathbf{a}(K,\infty)]}{\sigma_\varepsilon^{\,2}}, \qquad (7)$$

where in the vectors $\hat{\mathbf{a}}(p,\infty)$ and $\mathbf{a}(K,\infty)$ zeros have been added for orders higher than p and K respectively. It is easy to prove that:

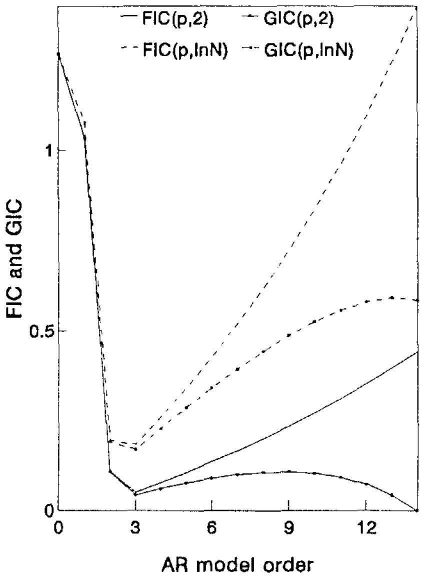

Figure 1: The $GIC(p,\alpha)$ and its finite sample counterpart $FIC(p,\alpha)$ for LSF estimates in 10,000 realisations of an AR(3) process with $N = 40$.

$$SE(p) = N\left[\frac{PE(p)}{\sigma_\varepsilon^{\,2}} - 1\right], \qquad (8)$$

because all three products with $\mathbf{a}(K,\infty)$ in (7), (8) yield $\sigma_\varepsilon^{\,2}$. The expectation $E[SE(p)]$ equals asymptotically p for $p \geq K$; the finite sample value follows with (7) as [12,13]:

$$E[SE(p)] = N\left[\prod_{i=0}^{p}[1 + v(i,.)] - 1\right], \qquad p \geq K. \qquad (9)$$

The Selection Error $SE(p)$ of (7) is used to compare the quality of different order selection criteria for various sample sizes.

EFFICIENCY OF ORDER SELECTION

The different values of α have been evaluated with respect to their accuracy of selection.The accuracy is expressed in terms of the Selection Risk. The reasoning that underpins the choice for α can be applied to the asymptotic $GIC(p,\alpha)$ as well as to the finite sample $FIC(p,\alpha)$. The possibilities of selecting a wrong order can be separated in overfitting and underfitting. In case of overfitting the selected model M is greater than K and consequently insignificant parameters are included in the model. The expected costs of overfitting can be calculated with statistics. When an order is selected that is too low one or more significant higher order parameters have been excluded from the model. The expected costs of underfitting therefore depend on the true process by the deterministic values of these higher order parameters of the process.

Most order selection criteria minimize either the costs of underfitting or those of overfitting. When $\alpha = 2$ the costs of underfitting are minimized [14] disregarding the statistical costs of overfitting, while $\alpha = ln(N)$ may result in unlimited high costs of underfitting as the sample size increases, since α is a function of N. The "best" criterion, however, balances the costs of overfitting

Table I: Asymptotic value of the Selection Risk due to overfitting, *SR(K,L-K,α)]*, as a function of the penalty α in *E[GIC(p,α)]*, for $K=0$ and $L=100$.

α	*SR*	α	*SR*	α	*SR*
.0	100.000	2.6	1.288	6	.132
.5	98.964	2.8	1.014	7	.080
1.0	56.788	2.915	.915	8	.049
1.5	7.524	3.0	.851	9	.031
2.0	2.568	3.2	.723	10	.019
2.2	1.936	4	.411	12	.007
2.4	1.520	5	.226	20	.000

and underfitting [14,15], which comes down to finding the equilibrium between respectively a stochastic and a deterministic quantity. In this section the two extremes are treated, being the asymptotical $\alpha = 2$ on the one hand and the consistent $\alpha = ln(N)$ on the other.

The Selection Risk, *SR(K,L-K,α)* is a function of the order of the process that originally generated the data, *K*, the maximum order that can be overfitted, *L-K* (*L* is the maximum order considered for selection), and the penalty factor in the order selection criterion used, α. The Risk is determined as follows: In every run of repeated Monte Carlo simulations an AR(K) process generates *N* observations, based on which models are estimated and the criteria select a model order. The Selection Error of (7) is now computed in every run for every criterion and the averages of the simulation runs are calculated for each of the criteria considered. It should be noted that this averaging procedure involves different model orders, because it is highly improbable that a criterion chooses the same model order in every run. The order selection criterion with smallest average *SE(p)* is defined to be the best. This value is called the Selection Risk, *SR(K,L-K,α)*, which expresses the accuracy of the order selection criteria considered. Shibata [14,15] has given a formula for this *SR(K,L-K,α)* of a model of orders selected with *GIC(p,α)*, neglecting the possibility of underfitting. It is given by:

$$SR(K,L-K,\alpha) = K+\sum_{m=1}^{L-K} Prob(\chi^2_{m+2} > \alpha m). \qquad (10)$$

Table I gives the value of *SR(K,L-K,α)*, for $L=100$ and $K=0$. It shows that the costs of overfitting are enormous when α is less than 2. However when α takes values greater than, say, 7 the costs become practically zero. This is the basis of the consistent order selection criteria where α is a function of *N*, e.g. *ln(N)* or *2lnln(N)*. The Selection Error due to overfitting becomes zero, when $N \rightarrow \infty$ in a situation where a deterministic set of process parameters ***a(K)*** is given.

The asymptotic distributions, *Prob(M=p)* have been determined [15] for orders *M* selected at the minimum of *FPE(p)* or *GIC(p,α)* for $\alpha=2$. The probability of selecting an order $M > K$ has been computed together with its Selection Risk, *SR(K,L-K,2)*. The risk of overfitting increases the average Prediction Error. The *FPE(p)*, however, will estimate a decreasing value of the *PE(p)* in the case of overfit, which is logical otherwise no overfitted model would be selected. The result is that the part of the residual variance which seemed to be explained above the true order equals the increase in the Prediction Error. The residual variance can be used to compute the *FPE(p)*, which can be modified like the *PE(p)* in (7), (8). In formula, for asymptotical theory with $\alpha=2$:

$$E[N\{\min[FPE(M),M=K,K+1,\ldots,L]/\sigma_\varepsilon^2-1]]-K = \\ -[SR(K,L-K,2)-K]+2\sum_{i=0}^{L-K} i\ Prob(M=K+i). \qquad (11)$$

Both the theoretically expected increase of the Selection Risk, *SR(K,L-K,2) - K*, and the decrease of its estimator (11) are presented in Figure 2, given without loss of generality for *K=0.* The asymptotical value for $L \rightarrow \infty$ for *SR(K,L-K,2) - K* (again with $\alpha = 2$) is 2.57, which is already achieved for a maximum possible overfit of 20 orders (*L-K*=20). Also the modified *FPE* of (11) has a constant value of -.69 then. Practically speaking, Figure 2 indicates that for a penalty factor $\alpha=2$, the costs of selection remain almost the same, independently whether 10 or 100 possible orders of overfit are considered. The costs are 2.57 times the extra cost of always selecting one order too high, *K+ 1*, in an AR(K) process. Criteria with $\alpha=2$ are asymptotically efficient, but inconsistent which is the price to be paid for minimizing the risks of underfitting.

Given a data series of lenght *N* the values of the parameters ***a(K)*** of the data-generating process determine the loss in a situation of underfitting. When underfitting only one order the selection error depends on the last process parameter a_K and it can become infinite for all consistent methods when its magnitude is near the critical value of [14]:

$$a_K = [(\alpha-1)/N]^{1/2}. \qquad (12)$$

This value depends on the sample size *N*. It can easily be seen that when this parameter value occurs asymptotic criteria have an equal probability for selecting the orders *K-1* and *K*, because

$$E[GIC(K-1,\alpha)] = E[GIC(K,\alpha)]. \qquad (13)$$

Furthermore, using

$$E[PE(K)] \approx E[PE(K-1)]\ [1-a_K^2+1/N] \qquad (14)$$

and (7), (8) and (12) this yields for the critical parameter value:

$$SR(K,L-K,\alpha) = K+\alpha-2. \qquad (15)$$

Hence, the Selection Risk can become *K+ln(N)-2* for order *K-1* in the consistent criterion. Since the expectation of the selection criterion is equal for the orders K-1 and *K*, these will both be selected in 50% of the realisations for $N \rightarrow \infty$. For parameter values larger than the critical one of (12) order *K* will more often

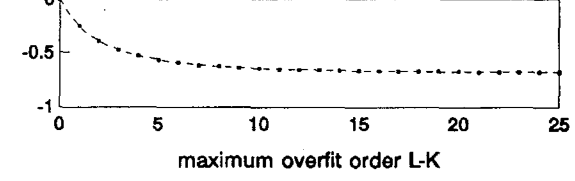

Figure 2: Increase in Selection Risk and decrease in its estimate of equation (11) due to overfit

be selected, whereas for smaller values this will be *K-1*. A similar derivation can be given in the Finite Sample Theory; the results cannot be presented in a formula as simple as (15).

A POSSIBLE CHOICE FOR α

The penalty factor α influences the ratio between the costs of overfitting and the possible costs of underfitting. Each choice of a value of α in *GIC(p,α)* and *FIC(p,α)* results in a different ratio between the two types of errors. Table I shows that the selection error due to overfitting decreases when the value of α increases. However, formula (15) indicates that the maximal error due to underfitting will increase with α for the critical parameter value (12). Consistent methods neglect the underfitting possibility. The Selection Risk may become infinite for $N \rightarrow \infty$ if the last process parameter has the critical value of (12). And $\alpha = 2$, which is asymptotically efficient, is just the limit for which one order underfitting will not give an increase in Selection Error.

We'll now calculate, as an example, the optimal value of α for the case where the *K*-th parameter of the AR(K) process has the critical value of (12). The value of α should be chosen such that the expectation of the selection error due to overfitting is equal to the maximum of the expected loss in underfitting one order.

In asymptotical theory this value follows by solving (15), which yields $\alpha = 2.915$, as could also be found approximately with Table I. For the finite sample *FIC(p,α)* the results are generally the same. Table II presents the results of an evaluation of an AR(1) process, where the optimal α in the finite sample *FIC(p,α)* are given for different values of *N*. The best α is shown to depend slightly on the estimation method and on the sample size, but is always close to 2.915, the value of the asymptotical *GIC(p,α)*.

The balance between both kinds of errors can be established by means of formula (15) and Table I. The balance is 0 for $\alpha = 2$, 1 for $\alpha = 2.915$; 1.2 for $\alpha = 3$; 2.6 for $\alpha = 3.5$; 4.9 for $\alpha = 4$; 13.3 for $\alpha = 5$ and 421 for $\alpha = 10$. In consistent methods α is a function of the sample size *N*, resulting in a dependence of the balance on this sample size. However, the formulae for the Selection Risk and the Selection Error yield results that hardly depend on *N*. Therefore, there might be a good reason to take α greater than 2, but there is no reason at all to make it dependent on *N*. An argument to take very high values for α is that the Selection Error due to overfitting is made small, whereas the loss due to underfitting will only occur if the true parameters have a value near the critical of (12). An opposite argument to be careful with higher values of α is that the loss of underfitting becomes much greater than considered if not only the last parameter, but also a number of previous true parameters have about the critical value. The maximum Selection Risk due to underfit becomes *L(α-2)* with (15), if all true parameters have critical values. This illustrates why $\alpha = 2$ is asymptotically efficient: for each value of α greater than 2, the maximum possible increase in the selection error due to underfitting goes to infinity for *N* and *L* going to infinity.

Table II: Optimal coefficients α for *FIC(p,α)* for $K = 1$, $L = N/3$, based on simulations.

N	YW	Burg	LSFB	LSF
20	2.9	3.1	3.3	3.3
30	2.9	3.0	3.2	3.3
50	2.9	3.0	3.2	3.2
100	2.9	3.0	3.2	3.1
250	2.9	2.9	3.1	3.1
1000	2.9	2.9	3.0	3.0
∞	2.915	2.915	2.915	2.915

A priori knowledge about the magnitude of the last parameters is essential in choosing an optimal value for α for a particular application. However, knowledge about the magnitude of the last parameters without knowing the order of the process is exceptional. In the case with a critical value for the last parameter, $\alpha = 3$ seems to be a good compromise. The small variations in Table II allow this advice; if required so, Table II gives refinements for sample size and estimation method.

CONCLUSIONS

The purpose in modeling disturbances is to find the most accurately predicting model from finite samples. A finite sample theory for AR processes shows the method of estimation to have a considerable influence as soon as the magnitude of the model order is not negligible with respect to the sample size. The finite sample theory gives an alternative for almost every existing order selection criterion. The quality of selection criteria can be evaluated with the Selection Risk, that has both an underfitting and an overfitting component. Asymptotic arguments minimize either the underfitting costs or the overfitting costs. The best is a balance between both costs. A good predicting model is found with the criterion *FIC(p,3)* which balances the expected statistical increase of the prediction accuracy due to overfitting with the maximal deterministic cost of underfitting one order, for the critical parameter.

REFERENCES

[1] Parzen, E. (1974). Some Recent Advances in Time Series Modelling. *IEEE Trans. on Automat. Contr.*, vol. AC-19, 723-730.

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

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

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

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

[6] 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.

[7] Broersen, P.M.T. and Wensink, H.E. (1992a). On the theory for autoregressive processes. *Proceedings 1992 International Conference on Acoustics, Speech and Signal Processing*, V497-500.

[8] Broersen, P.M.T. and Wensink, H.E. (1992b). A framework for autoregressive theory. *Proceedings of EUSIPCO-92*, 775-778.

[9] Broersen, P.M.T. and Wensink, H.E. (1993). On finite sample theory for autoregressive model order selection. *IEEE Transactions Signal Processing*, vol. 41, No. 1, 194-204.

[10] Broersen, P.M.T. and Wensink, H.E. (1993). Improved accuracy of asymptotic autoregressive theory. *Proceedings of the 14-ième colloque GRETSI.*

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

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

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

[14] Shibata, R. (1983). A theoretical view of the use of AIC. In O.D. Anderson (Ed.), *Time Series Analysis, Theory and Praxis 4*. Elseviers Science Publishers, North-Holland.

[15] Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, 63, 117-126.