Finite Sample Properties of ARMA Order Selection

Piet M. T. Broersen and Stijn de Waele

***Abstract*—The cost of order selection is defined as the loss in model quality due to selection. It is the difference between the quality of the best of all available candidate models that have been estimated from a finite sample of N observations and the quality of the model that is actually selected. The order selection criterion itself has an influence on the cost because of the penalty factor for each additionally selected parameter. Also, the number of competitive candidate models for the selection is important. The number of candidates is, of itself, small for the nested and hierarchical autoregressive/moving average (ARMA) models. However, intentionally reducing the number of selection candidates can be beneficial in combined ARMA(p, q) models, where two separate model orders are involved: the AR order p and the MA order q. The selection cost can be diminished by creating a nested sequence of ARMA$(r, r-1)$ models. Moreover, not evaluating every combination (p, q) of the orders considerably reduces the required computation time. The disadvantage may be that the true ARMA(p, q) model is no longer among the nested candidate models. However, in finite samples, this disadvantage is largely compensated for by the reduction in the cost of order selection by considering fewer candidates. Thus, the quality of the selected model remains acceptable with only hierarchically nested ARMA$(\mathrm{r}, \mathrm{r}-1)$ models as candidates.**

***Index Terms*—ARMA process, hierarchical model, order selection, penalty factor, spectral analysis, time series model.**

I. Introduction

THE ESTIMATION of a time series model is a parametric method for the spectral and correlation analysis of the stationary stochastic processes [1]. The nonparametric modified periodogram is a satisfactory solution for periodic and deterministic signals, but periodograms are less appropriate for stochastic observations [2]. The three model types that can be used for time series analysis are autoregressive (AR), moving average (MA), and combined autoregressive/moving average (ARMA) models. Stationary stochastic observations can always be characterized by either AR or by MA models [1]. However, the true theoretical order may be infinite. In practice, most models are estimated from a finite sample of N observations and they can be described adequately by AR(p), MA(q), or combined ARMA(p, q) processes with finite orders for p and/or q. An important problem for the application to practical problems is the choice of the best model order and the best model type for a given set of N measured observations. It is unusual that the true time series model order and type for new stochastic data can be derived from physical modeling principles. However, powerful statistical order selection criteria are available. They are generally based on the statistical significance of the decrease of the residual variance for a growing number of estimated parameters. A successful automatic time series analysis program that includes estimation algorithms and order selection criteria is available [3]. It uses the data as input and computes automatically, without requiring interaction of the experimenter, the time series model with precisely the details that are statistically significant for the data at hand, leaving out all details that are not significant.

The performance of order selection criteria in ARMA selection differs from the usually investigated selection of the AR order [4]. However, so far the theoretical treatment has been only asymptotical for ARMA, while finite sample considerations have produced better selection criteria in AR estimation [5]. Therefore, the cost of selection strategies or criteria in finite sample ARMA estimation has to be analyzed. In time series literature, most attention has been devoted to order selection for AR models. Here, the usual purpose of order selection is an AR model that can accurately predict future observations. Accurate prediction in the time domain leads to the same selection demands as an accurate spectral model in the frequency domain. Akaike's information criterion AIC [6] was based on a powerful mathematical framework, combining the likelihood description from estimation theory and the Kullback–Leibler distance from statistical theory. It became clear that AIC tends to select orders that are too high, even asymptotically. An exact result has been derived for the probability of selecting AR orders that are too high and the cost of overfit as a function of the overfit order [7]. Moreover, the selected AIC order in finite samples turned out to be highly dependent on the highest candidate order that was considered. Solutions suggested to overcome selection problems were consistent criteria [8], higher penalty functions [9], finite sample considerations [5], and some asymptotical corrections to AIC.

A special property of AR and MA models is that the order selection is hierarchical. The nested AR$(p-1)$ model is the only AR model with one parameter less than AR(p). Likewise, the only competitor of one order higher is the AR$(p+1)$ model. The nesting problem is an important reason why order selection in ARMA(p, q) processes is different from AR: two orders are involved. This means that every ARMA(p, q) process has a number of possibly competitive ARMA$(p+k, q-k)$ models with the same number of parameters, but those are biased for $k \neq 0$. Likewise, there are many models with one parameter more or less. This influences the order selection problem and no exact theoretical analysis has been given for models which are not nested. One obvious reason is that the bias and the variance of all ARMA$(p+k, q-k)$ models depend on the true

Manuscript received June 15, 2002; revised November 25, 2003.

P. M. T. Broersen is with the Department of Multi Scale Physics, Delft University of Technology, 2600 Delft, The Netherlands (e-mail: broersen@tn.tudelft.nl).

S. de Waele is with the Philips Research Laboratories, Eindhoven, The Netherlands.

Digital Object Identifier 10.1109/TIM.2004.827058

characteristics of the process. A variance analysis is only representative for unbiased models with both $p' \geq p$ and, at the same time $q' \geq q$; otherwise, the bias should be included in an evaluation of the statistical performance. However, not all biased candidates can be excluded in a realistic finite sample analysis of ARMA selection. The bias of leaving out a parameter may be smaller than the standard deviation of that parameter if it is estimated. Some special order selection methods coming from pattern recognition have been described [4], that do not require nesting. They rely on the difference between biased and unbiased models, which indeed will become clear for an infinite number of observations of an ARMA(p,q) process. Asymptotically, the bias cost of estimation becomes unlimited in comparison with the variance cost. However, the pattern methods do not take into account that the finite sample transition between well fitting models and inadequate models is gradual.

An order selection criterion has been applied to ARMA models in an automatic spectral analysis computer program [2], [3]. To reduce the computing time, the selection has been restricted to nested ARMA$(r,r-1)$ models. By considering only ARMA models with MA orders equal to the AR order minus one, ARMA order selection has artificially been made nested and hierarchical. A special property of this ARMA$(r,r-1)$ selection is that only candidate models with *two* parameters, more or less, are allowed as the closest competitors. This paper compares the cost in terms of prediction accuracy between the selection from unstructured and from hierarchical candidates. Limitation of the number of candidate models gives the possibility that the true ARMA(p,q) model is not a candidate. If $q < p$, the smallest unbiased candidate is the ARMA$(p,p-1)$ model with $p-1-q$ more parameters than the true process with zero as true value. *Estimation* of those additional parameters gives an extra contribution to the minimal possible estimation accuracy. Nonetheless, the cost of *selection* may be reduced because fewer candidates are available for selection in a nested sequence of candidates.

The penalty in the order selection criterion for ARMA models in the automatic order selection program [3] has been based on asymptotical arguments that are valid for AR processes. No special considerations have been given to ARMA models. In this paper, attention is given to the question of the best penalty factor for additional ARMA parameters. The candidates are the penalty 2 of Akaike's AIC and the penalty 3 that follows from a compromise between bias and variance errors in selected AR models [9].

II. ARMA Models

An ARMA(p,q) process is defined as [1]

$$x_n + a_1 x_{n-1} + \cdots a_p x_{n-p} = \varepsilon_n + b_1 \varepsilon_{n-1} + \cdots + b_q \varepsilon_{n-q} \quad (1)$$

where ε_n represents a series of independent, identically distributed, zero mean white noise observations. The process is AR for $q = 0$ and MA for $p = 0$. The ARMA process can also be written with polynomials of AR and of MA parameters as

$$A(z)x_n = B(z)\varepsilon_n, \quad z^{-1}x_n = x_{n-1} \quad (2)$$

with $A(z) = 1 + a_1 z^{-1} + \cdots + a_p z^{-p}$ and $B(z) = 1 + b_1 z^{-1} + \cdots + b_q z^{-q}$. Models have estimated polynomials of arbitrary orders, not necessarily equal to p and q. Models are stationary if the estimated roots of $A(z)$ are inside the unit circle and invertible if the zeros, the roots of $B(z)$ are inside the unit circle. Only models, which are stationary and invertible, are considered as candidates for order selection [2].

Order selection criteria are based on the rate of the decrease of the residual variance as a function of the number of parameters in the estimation from N observations. This residual variance is known in estimation. To evaluate the accuracy of estimated and selected models, they should be compared to the true process. The model error ME is such a measure for the quality of estimated models [10]. This measure can be used in simulations where an omniscient experimenter knows the true process parameters that generated the N observations from which the candidate models for selection are estimated. The ME is a scaled transformation of the expectation of the squared error of prediction PE

$$\mathrm{ME} = N\left(\frac{\mathrm{PE}}{\sigma_\varepsilon^2} - 1\right). \quad (3)$$

The prediction error is an obvious and usual measure for the accuracy in time series analysis, but it is unknown if models are estimated in practice. The PE measures the fit of models with previously estimated parameters on fresh new observations from the same stochastic process. The asymptotical expectation of the ME for an ARMA(p',q') model is independent of the variance σ_ε^2 of the excitation signal and also independent of N for unbiased models. On the contrary, the expectation of a bias contribution of an underfitted model is directly proportional to N in the ME. Only the values of the true and of the estimated parameters are required to compute the ME [10]. The Cramér–Rao lower bound for the spectral accuracy of *unbiased* models, with at least all truly nonzero process parameters included, equals the number of estimated parameters when expressed in the measure ME. The ME denotes the model quality in this paper.

III. Unconditional Expectations

For AR processes, two different ways to analyze the influence of the order of the estimated model can be used. The first way gives the unconditional expectation of characteristics or accuracy measures as a function of the model order. No order selection is considered. Possible characteristics are: the residual variance, the prediction error, the model error ME of (3), and the numerical outcome of an order selection criterion. The generalized information criterion (GIC) with penalty factor α is defined by [9]

$$\mathrm{GIC}(p,\alpha) = \log\{\mathrm{RES}(p)\} + \alpha\frac{p}{N} \quad (4)$$

where RES(p) denotes the residual variance of a model with p estimated parameters. GIC$(p,2)$ is better known as the AIC criterion [6]. The criterion GIC$(p,3)$ is a compromise between bias and variance costs in AR estimation [9]. If the process is truly AR(p), it has been shown that, for orders p and greater, the behavior of all those unconditional characteristics is independent of the true parameters of the AR(p) process [11]. This is the reason why many important results can be derived from a white noise analysis. Those simple white noise results for the

characteristics apply to all AR(p) processes for the true order p and higher.

The theoretical derivation of most AR order selection criteria is based on unconditional properties. By excluding the possibility of underfit, the analysis of white noise can be applied to the theory of order selection with as candidates only AR models of order p and higher. As an example, the asymptotical unconditional expectation of the ME for models of a true AR(p) process is

$$\mathrm{E}[\mathrm{ME}(k)] = k, \quad k \geq p \tag{5}$$

where ME(k) is the ME for an AR(k) model. For $k < p$, the expectation of the ME has also a bias contribution that is proportional to N. The asymptotical result in (5) also applies to *unbiased* MA or ARMA models with k estimated parameters. Each parameter above the true order adds 1 extra to the expectation of ME. These unconditional derivations, however, do not take into account the fact that an overfit order will be selected if the estimate of the highest order parameter in a specific realization is greater than its unconditional expectation. According to the unconditional expectations, E[AIC(k)] will have the minimum for $k = p$. Using unconditional theory, the penalty factor 2 of AIC would be the best choice.

IV. Conditional Expectations

Conditional expectations describe the expectation after selection. They apply to selected models only and they take the actual estimates of the parameters into account. This shows that the conditional E[AIC(k)|(k is selected)] is generally smaller than the unconditional E[AIC(k)]. That is certainly true for unbiased models, with $k \geq p$. The reason is that an order that is too high is only selected if the estimated parameter of that high order seems to be significant because its estimated value is greater than the true value. Likewise, the conditional expectation of the residual variance of a selected model is smaller than its unconditional expectation. At the same time, the conditional expectations of the model error ME and of the prediction error of selected models are greater than their unconditional *a priori* expectations. Selected models seem to fit better than according to the *a priori* unconditional expectation, but in reality the fit to future data is worse. Using conditional theory, the choice 2 for the penalty factor in AIC has been reconsidered [9].

The cost of selection can be defined as the quality difference expressed in ME between the model of the selected order and the best of all estimated candidate models. That will mostly be the model of the true order, if all true parameters are statistically significant, that is, tacitly assumed in theoretical derivations. The definition of costs deals with selected models. Therefore, the conditional expectations apply and the actual selection costs depend on the data at hand. With conditional theory for the use of AIC in AR order selection, the expectation of the costs of selection shows an asymptotical increase of 2.56 in ME [7]. This is the combined sum of the probabilities of having k orders overfit multiplied with the expected cost if that order is selected. Taking penalty factor 3 has reduced this increase to 0.85 [9], at the cost of a small possible bias contribution. Higher penalties would still give lower values for the cost of overfit, but they give a greater probability of underfit with too much bias. Simulations have shown that penalty 3 is a better compromise than 2 or 4 for the penalty factor [5]. It is essential for the theoretical results that the candidate models are hierarchical. Therefore, all overfitted AR or MA models are unbiased, at least if unbiased parameter estimation algorithms are used. Moreover, there is only one candidate model with a single overfit parameter as well as only one candidate for any arbitrary number of overfit parameters. Only underfitted models have bias, but those models are excluded in this theoretical analysis of the costs by assuming that the parameter of order p of the AR(p) model is statistically very significant, such that it is never missed in order selection. So far, no conditional theory includes the possibility of underfit and the theoretical results are hardly realistic for finite sample AR order selection. The mathematical preference for consistent selection criteria [8] is a purely theoretical and asymptotical concept, for ever-increasing sample sizes. That is not applicable to actual time series analysis where an order has to be selected for a given finite number N of observations.

The transition region between biased and unbiased AR models will be discussed. Special consideration is given to parameter values that are on the edge of statistical significance. Suppose that an AR(p) process has an AR($p-1$) model that is significant beyond any doubt for the given sample size N. The residual variance RES(p) of the AR(p) model can be derived from the previous residual variance RES($p-1$) with the estimated parameter $\hat{a}_p$ as [12]

$$\mathrm{RES}(p) = \mathrm{RES}(p-1)\left\{1 - \hat{a}_p^2\right\}. \tag{6}$$

It is easily verified that GIC(p, α) of (4) is smaller than GIC($p-1, \alpha$) and that order p is selected above order $p-1$ if $\hat{a}_p^2 > \alpha/N$. The asymptotic theoretical expression for the standard deviation of the last parameter estimated from N AR(p) observations is given by

$$\mathrm{var}[\hat{a}_p] = \frac{(1 - a_p^2)}{N}. \tag{7}$$

For small values of a_p, this can be approximated by $1/N$. For a normally distributed estimate $\hat{a}_p$, the unilateral probability to be further from the expectation a_p than 1.96 times the standard deviation equals 2.5%. In other words, if the true last parameter a_p equals about $3/\sqrt{N}$, there is still a nonvanishing probability that the actual estimate $\hat{a}_p$ is twice the standard deviation under that true value and the estimated $\hat{a}_p^2$ will become $\approx 1/N$. This gives a residual reduction that is too small to be selected with the usual values 2 or 3 for α in GIC(p, α). On the other hand, if the true parameter a_p equals zero, there is almost 5% probability that the absolute value of $\hat{a}_p$ exceeds $2/\sqrt{N}$, which will lead to selection with the criterion GIC(p, α) of (4) if the penalty factor α is less than 4. Final true parameters must be greater than four or five times their standard deviation $1/\sqrt{N}$ to make practically sure that in no single simulation run an estimate that is too small is found that is not selected. A final true AR(p) parameter value $a_p = 0.01$ would require more than 200 000 observations to be almost never missed in the practice of order selection, or in a proper *conditional* theory. However, the *unconditional* limit for the statistical significance of the final AR parameter is $1/\sqrt{N}$

and 10.000 observations are sufficient to let $a_p = 0.01$ be significant in the unconditional theory.

The reasoning for the last parameter a_p can be extended to processes where both a_p and a_{p-1} have true values near $\sqrt{\alpha}/\sqrt{N}$. In those critical cases, a large estimate for the last parameter $\hat{a}_p$ can compensate for a smaller estimate for $\hat{a}_{p-1}$. This situation becomes much more complicated than the study of Shibata [7] with only the probability of overfit. Such a study may be feasible for AR processes, but the bias results depend on the particular AR(p) process, especially on the true values of the last parameters. For a true ARMA(p,q) process the situation becomes still much more complicated because there are many candidate ARMA$(p+i, q-i)$ models with the same number of parameters or with one parameter more or less. It can be expected that estimated parameters for those different models are strongly correlated if they are estimated from the same data. The conditional expectations are required for theoretical results about the costs of selection and it is not attractive to make such a study, even if it would be possible.

V. Hierarchical Versus All Possibilities

A special property of AR and MA models is that the order selection is hierarchical. The nested AR$(p-1)$ model is the only AR model with one parameter less than AR(p). Likewise, the only competitor of one order higher is the AR$(p+1)$ model, no matter what the highest candidate order would be. Arbitrary ARMA(p,q) processes have a number of models ARMA$(p+k, q-k)$ with the same number of parameters and many others with one parameter less or with one parameter more. The unconditional expectation of the quality ME of the estimated ARMA(p,q) model will be ME $= p+q$ if the estimation is unbiased. The unconditional expectation of ME for all other ARMA$(p+k, q-k)$ models gives ME $= p+q$ as contribution of the estimation variance and an additional bias term that depends on the true process characteristics. Some of them may be close competitors for the ARMA(p,q) model in order selection. Not all models with $p+q$ or more parameters are unbiased, only ARMA(p',q') models with both $p' \geq p$ and at the same time $q' \geq q$ are unbiased. Moreover, the bias of some underfit models may become small if a close pole-zero pair is omitted. Hence, an analysis of the statistical cost of selection based only on variance contributions like for AR processes [9] is not possible and the AR theory of selection costs cannot be applied to arbitrary ARMA(p,q) models.

The number of ARMA(p',q') candidate models with $p' \leq L$ and $q' \leq L$ is L^2. If all ARMA(p',q') models would be available for order selection, the true order model has many close competitors and it may be expected that the performance of order selection criteria deteriorates. It would more be related to *subset* selection with arbitrary subsets than to selection in a hierarchically nested class of candidate models, where each higher order model contains all parameters of lower order models. Nested selection has L candidate models if L is the highest order considered, whereas there are 2^L possible subset models. In the subset selection, each possible subset model is a next candidate for selection and each estimated parameter that individually seems statistically significant is included, even if many previous model orders were not significant. Theoretically, the selected subset size will increase with the number of candidate models whereas the selected hierarchical model order is independent of L. The costs of subset selection have been investigated: the selection between al possible subsets selects many more parameters, but leads to a poor model in time series [13]. The explanation is simple. If many closely competing candidates with similar unconditional accuracy are available for selection, generally one of those candidates is an estimated model that seems to fit much better than the others and that one will be selected. The performance of L^2 ARMA(p',q') candidate models will be somewhere between L hierarchical models and 2^L subset models.

As computation of ARMA models may be time consuming, a program for the automatic analysis of time series [3], [14] considers only hierarchical ARMA$(r, r-1)$ models. This particular choice is also inspired by the fact that models of those orders are good discrete time approximations for many continuous time processes [1, p. 382]. Instead of L^2 ARMA(p',q') models, with $p' \leq L$ and $q' \leq L$, only L ARMA$(r, r-1)$ models have to be computed. Those models are ordered hierarchically. If the true process would be ARMA(p,q), the smallest unbiased ARMA$(r, r-1)$ model contains $|q+1-p|$ parameters with expectation zero. Having fewer candidates and hardly any close competitors, the cost of order selection will become small because it becomes easier to select the best among the estimated models. However, it is no longer certain that the very best model which can possibly be estimated from the data is among the candidates. The closest unbiased ARMA$(r+i, r-1+i)$ model has *two* more parameters. The closest biased model has two parameters less. Asymptotically, the probability of one order overfit is for ARMA$(r, r-1)$ models given by the probability that the chi-squared distribution with two degrees of freedom exceeds 2α, where α is the penalty factor. This is smaller than the probability that the chi-squared distribution with one degree of freedom exceeds α ; that applies to AR or to MA models where only one parameter more is considered. The price to be paid is that the true ARMA(p,q) model may not always be among the candidates for selection and some additional parameters with expectation zero have to be included in the selected model.

It has been shown that it is not realistic to exclude biased models like ARMA$(p+i, q-i)$ in a theoretical analysis of order selection properties if all ARMA models with arbitrary orders are selection candidates. Hence, a theoretical comparison of the selection costs of the hierarchical approach with L candidates and the selection with two free model orders and L^2 candidates is not feasible. Therefore, the performance of ARMA order selection algorithms is studied in simulations.

VI. Simulations

The simulations evaluate the quality of models that are selected from different sets of candidate models estimated from the same N observations. They also give an indication about a good choice for the penalty function. Only ARMA models are considered with an MA order lower than the AR order because

the true process also has this property. The selection among hierarchical ARMA$(r, r-1)$ candidates, $r = 0, \ldots, 100$ is compared with the selection with as candidates all ARMA$(r, < r)$ models with the MA order lower than the AR order. Asymptotically, the probability of m orders overfit for selection with penalty α in ARMA$(r, r-1)$ models is given by the probability that the chi-squared distribution with $2m$ degrees of freedom exceeds $2m\alpha$. Therefore, it almost never occurs in truly finite ARMA$(r, r-1)$ processes that the selected order is five or more higher than the true order for α greater than 2. Choosing a high order as maximum order for selection gives confidence that the best model will be among the candidates for measured data of an unknown type, but it takes more computing time. If some *a priori* information about the orders is available from previous similar experiments, it might be useful to limit the maximum orders of the selection candidates.

Simulations have been made with ARMA(7,2) processes with AR parameters given by the reflection coefficients

$$1, -\beta, (-\beta)^2, (-\beta)^3, (-\beta)^4, (-\beta)^5, (-\beta)^6, (-\beta)^7$$

and with MA parameters given by reflection coefficients

$$1,\ \beta,\ \beta^2.$$

The MA parameters are computed with the Levinson–Durbin recursion that also relates AR parameters with reflection coefficients [12]. This choice for the generating reflection coefficients gives an ARMA process with all poles and all zeros at the same radius $|\beta|$. It gives the convenient possibility to generate different levels of significance for the parameters of the true process and to create examples where biased underfit models are the best as well as processes where only unbiased models are attractive candidates for selection. The ARMA(7,2) process is chosen such that all competitive ARMA models are in the class ARMA$(r, < r)$ if β is large enough. In this way, the conclusions can be extrapolated to the full class of all ARMA$(\leq r, \leq r)$ models. White noise is indicated as $\beta = 0$; $\beta = 0.4$ is an example where only 3 AR parameters are statistically significant for $N = 1000$. The MA(3) model is the best unconditional MA model and the ARMA(2,1) model is the best unconditional model in the ARMA$(r, r-1)$ class of candidates. For $\beta = -0.6$, the best order would be AR(6) if only AR(p) models are candidates for selection, MA(4) for MA candidates and ARMA(3,2) for ARMA$(r, r-1)$ candidates. The bias of the AR(6) model is 0.17, expressed in ME for $N = 1000$, the bias of MA(4) is 0.11 and the bias of ARMA(3,2) is 0.41. Hence, adding 1 for every estimated parameter finds the minimum unconditional ME expectation of those three models, yielding 6.17 for AR, 4.11 for MA, and 5.41 for ARMA. Finally, $\beta = 0.8$ is an example where the best AR order equals 15, MA(15) is the best MA candidate and ARMA(7,6) the best ARMA$(r, r-1)$ candidate. All selections have been made with the values 2 and 3 for the penalty factor α in the order selection criterion GIC(p, α) of (4). The variation of β gives the opportunity to study the performance of the penalty factor in the most difficult selection examples where underfitted biased models are expected to be the best candidates among estimated models.

Tables I and II show the average ME of the true ARMA(7,2) model and of five selected models:

TABLE I
AVERAGE UNCONDITIONAL ME OF THE TRUE ARMA(7,2) MODEL AND AVERAGE CONDITIONAL ME OF FIVE MODELS SELECTED WITH PENALTY $\alpha = 2$ FROM 5 DIFFERENT SETS OF CANDIDATES, AS A FUNCTION OF THE RADIUS β. $N = 1000$

$\rightarrow \beta$	0	0.4	-0.6	0.8
ARMA(7,2)	7.8	7.7	7.5	39.6
ARMAsel	2.8	6.6	7.5	21.1
AR	2.6	6.4	8.5	21.4
MA	2.2	6.2	6.3	20.7
ARMA(r,r-1)	1.0	4.5	8.7	17.8
ARMA(r,<r)	2.6	5.9	8.8	20.5

TABLE II
AVERAGE CONDITIONAL ME OF FIVE MODELS SELECTED FROM THE SAME SETS OF CANDIDATES AS IN TABLE I, BUT NOW WITH PENALTY $\alpha = 3$, AS A FUNCTION OF THE RADIUS β OF THE POLES AND ZEROS. $N = 1000$

$\rightarrow \beta$	0	0.4	-0.6	0.8
ARMAsel	0.9	5.4	6.7	20.7
AR	0.8	4.6	7.0	20.8
MA	0.8	4.8	5.6	20.0
ARMA(r,r-1)	0.4	3.9	8.7	18.8
ARMA(r,<r)	0.8	4.3	7.7	16.8

1) estimated model of true ARMA(7,2) structure (only in Table I);
2) ARMAsel selects model order and type, with as candidates AR(p), MA(q) and ARMA$(r, r-1)$ models [2], [3], [14];
3) AR selected from only AR(k) candidates with orders $k = 0, \ldots, 500$;
4) MA has only MA(k) as candidates with orders $k = 0, 1, \ldots, 200$;
5) ARMA$(r, r-1)$ has AR(0) and ARMA$(k, k-1)$ models as candidates with $k = 1, 2, \ldots, 100$;
6) ARMA$(r, < r)$ has AR(0) and all ARMA(r, k) models as candidates with $r = 1, 2, \ldots, 100$ and $k = 0, 1, \ldots, r-1$.

The white noise data with $\beta = 0$ have always the zero model without parameters as a candidate. Therefore, the best model would have ME $= 0$ and is a candidate. Tables I and II give the average ME of selected models, which is here equal to the cost of selection. The costs of selection increase with the number of competitive candidates, which is highest for ARMAsel and smallest for ARMA$(r, r-1)$. Moreover, the costs of pure AR or pure MA selection are close to the asymptotical theoretical ME values 2.56 and 0.85 for the penalties $\alpha = 2$ and $\alpha = 3$, respectively [9]. The ME of ARMA$(r, r-1)$ is the smallest because it has no candidates with 2 or 4 or any even number of estimated parameters. ARMA$(r, < r)$ has more candidates and, hence, the costs of selection are higher. A special property of this white noise example is that *all* estimated and selected parameters are of unbiased models and underfit is impossible. The difference between the first columns of Tables I and II is a clear support for the use of the penalty factor 3 in ARMA selection. The difference is less pronounced in the other columns, but it is mostly in favor of the penalty factor 3.

For $\beta = 0.4$, a plot of the unconditional prediction error as an indication for the model quality is given in Fig. 1. As expected with the unconditional theory, the ARMA(2,1) model is the best

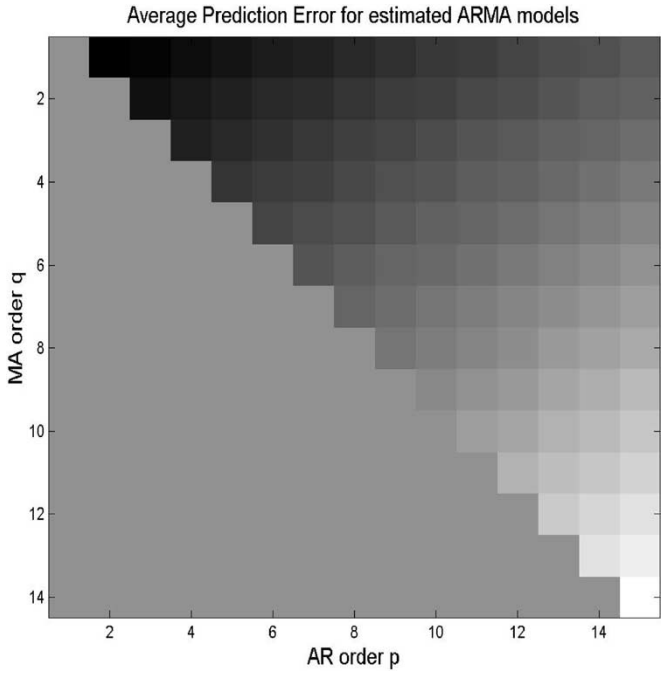

Fig. 1. Average unconditional prediction error for ARMA(p, q) models estimated from an ARMA(7,2) process with $\beta = 0.4$. A darker color indicates a better fit. It is remarkable that slope of lines of equal model quality are 45°, so the estimation variance is predominant, giving the same contribution to all models with the same number of estimated parameters $p + q$.

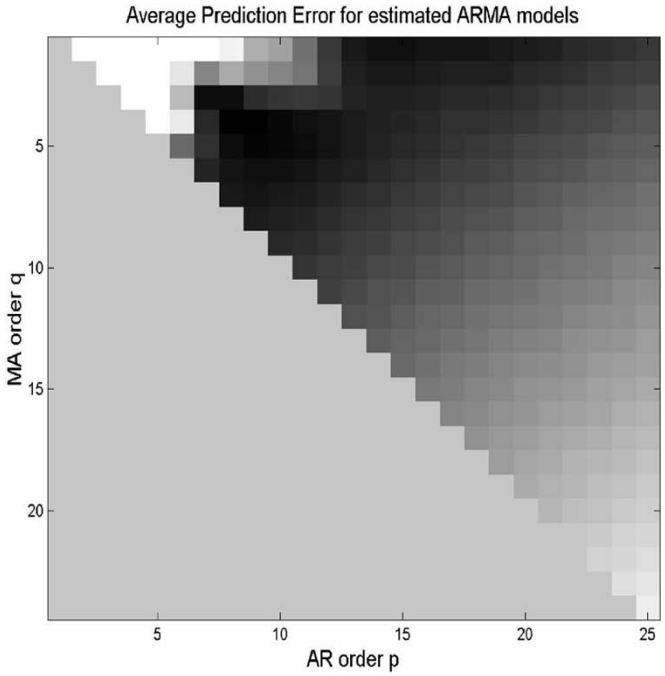

Fig. 2. Average unconditional prediction error for ARMA(p, q) models estimated from an ARMA(7,2) process with $\beta = 0.8$. A darker color represents a better fit. At low orders, the asymptotical bias pattern is visible with poor models for AR orders less than 7 and for MA orders less than 2. For higher model orders, the bias disappears and the finite sample effect of the estimation variance is still recognized in the slope of 45° for lines of equal prediction error.

there, with the darkest color. The finite sample character of the problem can be concluded from the fact that the underfitting bias is much less important than the variance due to the number of estimated parameters. Asymptotically, the quality of all biased models will be poor because the bias contribution to the ME is proportional to N and will grow without limits. Lines with constant $p + q$ are under 45° and the gray intensity in Fig. 1 is more or less the same along those lines. This means that the PE and also the ME are constant there; the expectation for the ME is equal to the number of estimated parameters $p + q$, with an additional very small bias contribution for all models with $p < 7$ or $q < 2$. The conditional numerical results in Tables I and II are less clear because both the bias and the variance play a role in the selection. Moreover, also the best candidate has to be estimated and is subject to statistical variations, unlike in the white noise example. However, the selected result of ARMA$(r, r-1)$ has lower costs than of ARMA$(r, <r)$ for both values of the penalty factor. This shows that reducing the number of candidates for selection may have advantages, even if the true process is no longer among the candidates. Furthermore, a clear preference for penalty 3 follows from the tbles for $\beta = 0.4$ and also for $\beta = -0.6$.

The first line in Table I for $\beta = 0.8$ with the true model order shows a peculiar property of the ARMA algorithms that have been used. In some simulation runs, they produce a poor estimate for fixed orders while neighboring models with one parameter more or less are estimated without a problem. The poor estimates are never selected. Therefore, the average quality of fixed true order ARMA models is often worse than the average of selected models. The results for $\beta = 0.8$ in Table II show that it is possible to find examples where selection from ARMA$(r, <r)$ is slightly better, at least for penalty 3. This could be expected because all parameters are significant now. Hence, the best unconditional model is ARMA(7,2) in the class ARMA$(r, <r)$ and ARMA(7,6) in the class ARMA$(r, r-1)$. Moreover, AR(9), ARMA(8,1), and ARMA(6,3) are biased and are not close competitors, which is favorable for selection of the ARMA(7,2) model if it is among the candidates. The best ARMA$(r, r-1)$ model for r equal to 7 has four extra MA parameters that have to be estimated. Therefore, the expected ME for the ARMA(7,6) model is four higher than the ME of ARMA(7,2). Fig. 2 shows that models of order that is too low are biased and poor, indicated by the light color. The quality of all models with more than say 15 or 20 parameters depends largely on the number of estimated parameters. The bias is important if less than ten parameters are estimated in this ARMA(7,2) example, but the influence disappears quickly. This is illustrated by the fact that selection between only AR or only MA models as candidates gives about the same ME in the tables. The best unconditional AR model order was 15 with as expectation for the ME 17.8 and MA(15) with ME 15.6. The ME of the selected models is slightly higher than that unconditional minimum. In Fig. 2, the gray color is more or less constant for lines with 45-degree slope with the same number of parameters $p + q$, if that number is greater than 15. This is another indication that it is not necessary to compute all ARMA(p, q) models, but that the limitation to ARMA$(r, r-1)$ models provides enough good candidates for the selection.

More simulations have been studied with different N, different values of β and higher true process orders. No results have been found that contradict the conclusions.

VII. Conclusion

The quality of selected models depends on two factors: the quality of the available estimated candidate models and the capacity to select the very best or one of the best from those candidates. The penalty factor has a strong influence on the selection cost, which is the difference between the quality of the selected model and the quality of the best candidate. Usually, penalty 3 is a good choice for the selection of an ARMA model for an arbitrary number of observations.

It is easy to select a very good model from a small number of good candidates; the selection costs are small. The selection cost increases if many good candidates are available. The hierarchical nesting imposed by ARMA$(r, r-1)$ candidate models gives an enormous reduction in the number of candidates in comparison with the estimation of all combinations of p for AR and q for MA orders. This reduction gives the possibility that the very best model is not estimated, if the true MA order is not equal to the AR order minus one. The best hierarchical candidate model may require that some extra parameters have to be estimated, with true values zero. However, the additional variance inaccuracy of those extra estimated parameters will generally be compensated by the reduction of the selection costs caused by the limitation of the number of good candidates. Generally, limiting the selection to ARMA$(r, r-1)$ candidate models does not lead to a lower quality of the model that is finally selected in finite samples.

References

[1] M. B. Priestley, *Spectral Analysis and Time Series*. London, U.K.: Accademic, 1981.

[2] P. M. T. Broersen, "Facts and fiction in spectral analysis," *IEEE Trans. Instrum. Meas.*, vol. 49, pp. 766–772, Aug. 2000.

[3] ——, ARMASA Matlab Toolbox. [Online]. Available: http://www.tn.tudelft.nl/mmr/downloads

[4] B. S. Choi, *ARMA Model Identification*. New York: Springer-Verlag, 1992.

[5] P. M. T. Broersen, "Finite sample criteria for autoregressive order selection," *IEEE Trans. Signal Processing*, vol. 48, pp. 3550–3558, Dec. 2000.

[6] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, Dec. 1974.

[7] R. Shibata, "Selection of the order of an autoregressive model by Akaike's information criterion," *Biometrika*, vol. 63, pp. 117–126, 1976.

[8] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 41, pp. 465–471, 1978.

[9] P. M. T. Broersen and H. E. Wensink, "On the penalty factor for autoregressive order selection in finite samples," *IEEE Trans. Signal Processing*, vol. 4, pp. 748–752, Mar. 1996.

[10] P. M. T. Broersen, "The quality of models for ARMA processes," *IEEE Trans. Signal Processing*, vol. 46, pp. 1749–1752, June 1998.

[11] ——, "The prediction error of autoregressive small sample models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 858–860, May 1990.

[12] S. M. Kay and S. L. Marple, "Spectrum analysis–modern perspective," *Proc. IEEE*, vol. 69, pp. 1380–1419, 1981.

[13] V. Solo, "Asymptotics for complexity regularized transfer function estimation with orthonormal bases," in *Proc. IEEE/CDC Conf. Decision Control*, 2001, pp. 4766–4769.

[14] P. M. T. Broersen, "Automatic spectral analysis with time series models," *IEEE Trans. Instrum. Meas.*, vol. 51, pp. 211–216, Apr. 2002.

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

He is currently with the Department of Applied Physics, DUT. His main research interest is automatic identification on statistical grounds by letting measured data speak for themselves. He developed a practical solution for the spectral and the autocorrelation analysis of stochastic data by the automatic selection of a suitable order and type for a time series model of the data.

Stijn de Waele was born in Eindhoven, The Netherlands, in 1973. He received the M.Sc. degree in applied physics and the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, in 1998 and 2004, respectively. His thesis was entitled "Automatic model inference from finite time observations of stationary stochastic processes."

He is currently a Research Scientist at Philips Research Laboratories, Eindhoven, where he works in the field of digital video compression.